


# Análise de dados e IA a serviço da inovação em gestão pública

MBA Ciência de Dados e Inteligência Artificial Aplicadas

# Análise de dados e IA a serviço da inovação em gestão pública

*Alexandre Galardinovic Ribeiro, Bruno Gustavo Janoca, Camila Falchetto, Eduardo Melo Moura Filho, Fabiana Ferreira do Nascimento, Fabio Kouri, Francisco Molina, Juliana Pierrobon, Luís César Seixas de Oliveira, Luiz Victor do Espirito Santo, Marcelo Cabreira Bastos, Nelson Oliveira Santos, Rafael Gerardo Rodrigues do Nascimento, Rafael Luis Giacomin, Rosana Gonçalves, Tadeu Gomes, Thiago Mello Venâncio*

Brasília
**2026**

Ficha catalográfica elaborada pela equipe da Biblioteca Graciliano Ramos da Enap

A612 Análise de dados e IA a serviço da inovação em gestão pública / Organizadores: Jackson Silvano De Toni, Rachel Pereira Dorneles. -- Brasília: Enap, 2026.

321 p.: il.

Inclui bibliografia.
MBA Ciência de Dados e Inteligência Artificial Aplicadas

ISBN: 978-65-87791-78-4

1. Inovação no Setor Público. 2. Inteligência Artificial Aplicada. 3. Ciência de Dados Aplicada. 4. Análise de Dados. 5. Políticas Públicas. 6. Governança de Dados. 7. Aprendizado de Máquina. 8. Administração Pública. I. De Toni, Jackson Silvano (org.). II. Dorneles, Rachel Pereira (org.). III. Título.

CDD 352.367

Bibliotecária: Elda Campos Bezerra – CRB 1/1425

**Enap 2026**

Este trabalho está sob a Licença Creative Commons - Atribuição: Não Comercial - Compartilha Igual 4.0 Internacional. As informações e opiniões emitidas nesta publicação são de exclusiva e inteira responsabilidade dos autores, não exprimindo, necessariamente, o ponto de vista da Escola Nacional de Administração Pública (Enap). É permitida a reprodução deste texto e dos dados nele contidos, desde que citada a fonte. Reproduções para fins comerciais são proibidas.

Escola Nacional de Administração Pública (Enap)
Diretoria de Educação Executiva
Coordenação-Geral de Especialização e MBA
SAIS - Área 2-A - 70610-900 — Brasília-DF, Brasil

## Escola Nacional de Administração Pública – Enap

**Presidente**
*Bethania Lemos*

**Diretora de Educação Executiva**
*Iara Alves*

**Coordenadora-Geral de Especialização e MBA**
*Juliana Miranda Leda*

**Coordenadora de Especialização e MBA**
*Patrícia Siqueira de Medeiros*

**Coordenadora do curso**
*Rachel Pereira Dorneles*

**Equipe da Coordenação**
*Bianca de Oliveira Ruskowski*
*Cecília Chachamovitz Baran*
*Felipe Peixoto Ribeiro*
*Fernando Motta Soares*
*Isabela dos Santos Rangel*
*Jana Petaccia de Macedo*
*Jéssica Vieira da Silva*
*Jonathas Ferreira Alexandre Souza*
*Patricia Lucinari*
***(Estagiários)***
*Júlia Martins Ferraris*
*Raabe Noleto*
*José Mateus Silva Montes*
*Thiago Ferreira*

**Organizadores**
*Jackson Silvano De Toni*
*Rachel Pereira Dorneles*

**Colaboradores**
*Natália Massaco Koga*
*Alex Lopes Pereira*
*Hélio Bomfim de Macêdo Filho*

**Autores**
*Alexandre Galardinovic Ribeiro*
*Bruno Gustavo Janoca*
*Camila Falchetto*
*Eduardo Melo Moura Filho*
*Fabiana Ferreira do Nascimento*
*Fabio Kouri*
*Francisco Molina*
*Juliana Pierrobon*
*Luís César Seixas de Oliveira*
*Luiz Victor do Espirito Santo*
*Marcelo Cabreira Bastos*
*Nelson Oliveira Santos*
*Rafael Gerardo Rodrigues do Nascimento*
*Rafael Luis Giacomin*
*Rosana Gonçalves*
*Tadeu Gomes*
*Thiago Mello Venâncio*

**Projeto gráfico e diagramação**
*Patrícia Menezes*

---

A Escola Nacional de Administração Pública (Enap) é uma escola de governo vinculada ao Ministério da Gestão e Inovação no Serviço Público (MGI).

Tem como principal atribuição a formação e o desenvolvimento permanente dos servidores públicos. Atua na oferta de cursos de doutorado e mestrados profissionais, MBA e especialização lato sensu, cursos de aperfeiçoamento para carreiras do setor público, educação executiva e educação continuada.

A instituição também desenvolve pesquisa aplicada e ações de inovação voltadas à melhoria do serviço público.

# SUMÁRIO


AGRADECIMENTOS 5

APRESENTAÇÃO 7

PREFÁCIO 11

INTRODUÇÃO 16

EIXO 1 | Processamento de Linguagem Natural e da Inteligência Artificial Generativa 20

EIXO 2 | Saúde Pública,na Sustentabilidade e no Impacto Social 99

EIXO 3 | Educação, Inteligência Territorial e Inclusão 167

EIXO 4 | Gestão Estratégica, no Orçamento e na infraestrutura 250

# Agradecimento

**Iara Alves**
*Diretora de Educação Executiva da Enap*

Para a Enap, o livro *Análise de Dados e IA a Serviço da Inovação em Gestão Pública* tem um significado especial, pois traduz de maneira concreta uma dimensão central de nossa missão. Como escola de governo, atuamos na formação e no desenvolvimento permanente de servidores e servidoras, na pesquisa aplicada e na inovação em gestão pública. Nosso papel é criar ambientes em que as pessoas possam aprender, experimentar, produzir conhecimento e transformar esse conhecimento em capacidade institucional. Isso exige estar atentos às grandes transformações do nosso tempo e, simultaneamente, apoiar o Estado na tarefa de compreendê-las, incorporá-las e governá-las de forma crítica e responsável.

É uma grande satisfação ver reunidos nesta publicação os resultados de uma experiência formativa do nosso Programa de Pós-Graduação *Lato Sensu*, cujo conjunto de ofertas busca fortalecer as capacidades do Estado para responder aos desafios contemporâneos e construir melhores políticas públicas para a sociedade brasileira.

Quero reconhecer especialmente os autores e autoras deste livro — Alexandre Galardinovic Ribeiro, Bruno Gustavo Janoca, Camila Falchetto Romero, Eduardo Melo Moura Filho, Fabiana Ferreira do Nascimento, Fabio Kouri Paim, Francisco Molina Duarte Júnior, Juliana Pierrobon, Luís César Seixas de Oliveira, Luiz Victor do Espírito Santo, Marcelo Cabreira Bastos, Nelson Oliveira Santos, Rafael Gerardo Rodrigues do Nascimento, Rafael Luis Giacomin, Rosana Gonçalves, Tadeu Gomes e Thiago Mello Venâncio —, nossos alunos e alunas do MBA em Ciência de Dados e Inteligência Artificial Aplicadas.

Cada trabalho apresentado nesta obra representa muitas horas de estudo, investigação e experimentação e, sobretudo, a disposição de olhar para problemas concretos da administração pública e buscar novas maneiras de enfrentá-los. Essa combinação entre conhecimento técnico, capacidade analítica e compromisso com o serviço público é uma das maiores riquezas desta publicação.

Agradeço também aos organizadores, Jakson De Toni e Rachel Dorneles, e aos colaboradores Natália Massaco Koga, Alex Lopes Pereira e Hélio Bomfim de Macêdo Filho, pela dedicação ao projeto e por contribuírem para uma ambição que é muito cara à Enap: ampliar nossa capacidade coletiva de transformar conhecimento em melhores decisões, melhores políticas públicas e melhores serviços para a população.

Meu reconhecimento se estende a todos e a todas as docentes e orientadores do MBA em Ciência de Dados e Inteligência Artificial Aplicadas. Formar profissionais públicos em um campo que se transforma em tão alta velocidade exige atualização permanente, abertura à experimentação e capacidade de rever continuamente práticas de ensino e aprendizagem. Agradeço, em especial, por contribuírem para formar pessoas capazes de pensar criticamente sobre as tecnologias que utilizam.

O verdadeiro valor público de uma formação em ciência de dados e inteligência artificial ultrapassa o domínio de ferramentas e métodos. Ele está também na qualidade das perguntas formuladas durante o processo de aprendizagem, na capacidade de avaliar criticamente os dados disponíveis e na construção de referências éticas para sua coleta, utilização e interpretação. Está, ainda, na compreensão de que os modelos incorporam critérios e escolhas humanas e institucionais e, por isso, precisam ser submetidos a princípios de transparência, explicabilidade, auditabilidade e responsabilização. Essa perspectiva atravessa a própria concepção da obra, que associa capacidade técnica ao julgamento crítico e à compreensão dos impactos das tecnologias sobre a ação do Estado e sobre os direitos dos cidadãos.

Na Enap, acreditamos que formar para a inteligência artificial no setor público significa formar também para a responsabilidade, para a ética, para a análise crítica dos vieses e para a proteção dos direitos das pessoas afetadas pelas decisões públicas. Nosso objetivo é preparar servidores e servidoras capazes de combinar competências tecnológicas com atributos indispensáveis à ação pública: discernimento, conhecimento do contexto, capacidade de formular boas perguntas, compromisso democrático e orientação para o interesse público.

A diversidade dos trabalhos reunidos neste livro demonstra o alcance dessa agenda. As pesquisas transitam por educação, saúde, direitos humanos, sustentabilidade, meio ambiente, inclusão, crédito rural, gestão orçamentária e infraestrutura, mobilizando diferentes métodos de ciência de dados e inteligência artificial para enfrentar desafios concretos do Estado. As propostas de inovação aqui apresentadas mostram que a tecnologia alcança seu sentido mais relevante quando está orientada à solução de problemas públicos e à geração de valor para a sociedade.

Celebramos, portanto, nossos alunos e alunas, que concluem esta pós-graduação mais preparados para contribuir com a construção de um Estado analítico, inovador e orientado por evidências, ao mesmo tempo comprometido com ética, transparência, inclusão e valor público.

Registro, finalmente, meu reconhecimento ao excelente trabalho da coordenadora-geral de Pós-Graduação *Lato Sensu*, Juliana Leda; da coordenadora Patrícia Medeiros; e da coordenadora do curso, Rachel Dorneles, assim como de toda a equipe que tornou possível essa jornada formativa. O trabalho de vocês contribui para fazer da pós-graduação da Enap uma experiência que ultrapassa os limites da sala de aula e aproxima aprendizagem, experimentação, produção de conhecimento e transformação institucional.

Que este livro seja um ponto de partida. Que o conhecimento produzido aqui continue circulando, seja debatido, testado e aperfeiçoado e encontre caminhos para ser incorporado às instituições públicas. E que inspire outros servidores e servidoras a explorar as possibilidades da ciência de dados e da inteligência artificial com curiosidade, rigor, responsabilidade e compromisso com o interesse público.

# Apresentação

**Jackson De Toni**

## *Análise de dados e IA a serviço da inovação na gestão pública*

A incorporação de análise de dados e inteligência artificial na administração pública brasileira não configura um fim em si mesmo, mas um instrumento cuja legitimidade depende da forma como se articula com os arranjos institucionais preexistentes e com as assimetrias de poder que atravessam o território nacional. Um algoritmo de focalização de benefícios sociais, por exemplo, carrega decisões normativas embutidas em sua arquitetura — quais variáveis pondera, quais populações captura adequadamente em seus dados de treinamento, quais vieses históricos reproduz sob a aparência de neutralidade técnica. Eubanks (2018) demonstrou como sistemas automatizados de triagem em programas assistenciais tendem a penalizar desproporcionalmente populações já vulnerabilizadas, quando desenhados sem escrutínio sobre a qualidade e representatividade da base de dados subjacente. A contextualização, portanto, não é um adendo humanista ao rigor técnico: é condição de validade do próprio modelo, na medida em que determina se a inferência estatística corresponde à realidade social que pretende descrever.

Os trabalhos reunidos nesta coletânea partem de uma premissa que merece ser explicitada: a efetividade de uma política pública mediada por dados depende tanto da robustez metodológica do modelo quanto da capacidade burocrática do Estado brasileiro de operá-lo dentro de sua estrutura federativa fragmentada. A literatura sobre capacidades estatais (Cingolani, 2013) já demonstrou que instrumentos tecnicamente sofisticados frequentemente fracassam não por falha de desenho, mas por descolamento entre a complexidade da ferramenta e a capacidade administrativa dos entes subnacionais responsáveis por sua implementação. Nesse sentido, um modelo preditivo aplicado à saúde pública municipal, por exemplo, só produz valor público efetivo se articulado a fluxos de governança que permitam a auditoria de seus resultados pelos próprios cidadãos afetados — o que desloca a discussão da acurácia estatística para a arquitetura de *accountability* que envolve o instrumento.

É nesse ponto que a dimensão democrática e participativa se conecta organicamente à análise de dados, e não como retórica acessória. Mazzucato (2018) argumenta que o valor público de uma tecnologia não reside em sua eficiência isolada, mas na direção coletivamente pactuada que orienta seu desenvolvimento e nos mecanismos de controle social que a acompanham. Os capítulos que seguem, escritos por servidores públicos federais em formação, buscam justamente essa síntese: tratam a ciência de dados não

como caixa-preta técnica delegada a especialistas, mas como campo de disputa distributiva que exige transparência metodológica, auditabilidade dos critérios de decisão e abertura ao escrutínio da sociedade civil organizada. É esse duplo compromisso — com o rigor analítico e com a inteligibilidade pública das escolhas embutidas em cada modelo — que confere a este livro seu valor formativo para a administração pública brasileira.

Inovar na gestão pública não é mais uma questão de escolha, mas uma exigência urgente do nosso tempo. O Estado do século XXI não pode mais ser governado pelas intuições do passado; ele exige a precisão das evidências e a coragem da transformação tecnológica. É com esse espírito provocativo e engajador que apresentamos esta obra, fruto do rigor analítico e da visão de futuro dos alunos da edição de 2025 do MBA em Ciência de Dados e Inteligência Artificial Aplicadas da Escola Nacional de Administração Pública (Enap). Ao longo destas páginas, o leitor não encontrará teorias abstratas ou distantes da realidade, mas sim a aplicação prática e revolucionária da tecnologia para resolver problemas complexos. A Enap, reafirmando seu papel fundamental como vanguarda na capacitação de servidores, entrega à sociedade quatorze trabalhos que provam como os dados podem otimizar processos, promover justiça social e humanizar o atendimento ao cidadão. Para tornar a sua jornada de aprendizado mais fluida e didática, organizamos as pesquisas em quatro grandes eixos temáticos que refletem as fronteiras da inovação governamental.

O primeiro eixo explora o imenso poder do **processamento de linguagem natural e da inteligência artificial generativa** na democratização da informação e automação de processos. Francisco Carlos Molina Duarte Júnior apresenta a “Análise automatizada de avaliações educacionais em larga escala: uma abordagem baseada em processamento de linguagem natural”, na qual utiliza modelos de linguagem avançados para classificar milhares de comentários de alunos da própria Enap, superando as métricas numéricas tradicionais para identificar fragilidades pedagógicas de forma ágil e escalável.

No campo jurídico, Luiz Víctor do Espírito Santo Silva desenvolve o “Benchmark jurídico e arquitetura documental orientada por evidência em casos brasileiros da Corte IDH”, demonstrando com rigor metodológico como arquiteturas de sistemas controladas e orientadas por evidências superam modelos de IA genéricos na triagem complexa de sentenças de direitos humanos.

A acessibilidade aos dados ganha força com Rosana Elisa Gonçalves Pinho em sua “Consulta em linguagem natural a dados de compras centralizadas de medicamentos do Ministério da Saúde”, que implementa um sistema inovador capaz de traduzir perguntas em português para consultas em banco de dados, permitindo que gestores analisem compras públicas sem precisarem dominar linguagens de programação.

Na mesma fronteira tecnológica de facilitar o acesso à informação, Tadeu Gomes Teixeira cria “Um agente inteligente baseado em LLM para consulta em linguagem natural de dados aquaviários da ANTAQ”, desenvolvendo uma solução que interpreta perguntas sobre a movimentação de cargas e as converte em consultas diretas aos microdados da agência, acompanhadas de explicações didáticas que respeitam a rigorosa metodologia oficial.

Nosso segundo eixo mergulha na **saúde pública, na sustentabilidade e no impacto social,** evidenciando como a análise orientada a dados atua como um verdadeiro escudo para as políticas vitais do país. Juliana Pierrobon Lopez traz a pesquisa “Desmatamento nos municípios da Amazônia Legal: uma análise espaço-temporal com dados em painel e

aprendizado de máquina", utilizando algoritmos preditivos para comprovar que a inércia temporal e a fiscalização são cruciais na proteção ambiental, alertando que o combate ao desmatamento deve estar profundamente conectado às redes de assistência social.

Desmistificando o senso comum com dados sólidos, Tadeu Gomes Teixeira, Fabio Kouri Paim, Rafael Luís Giacomin, Thiago Mello Venâncio, Luís César Seixas de Oliveira, Alexandre Galardinovic Ribeiro e Bruno Gustavo Janoca investigam os "Determinantes da cobertura vacinal nos municípios brasileiros: a primazia das condições socioeconômicas sobre o comportamento eleitoral", provando de forma inquestionável que a queda na vacinação nacional responde de forma muito mais sensível a desigualdades de renda e falhas estruturais de acesso do que a simples inclinações políticas do eleitorado.

O cuidado com os efeitos invisíveis da própria tecnologia é o foco de Camila Falchetto Romero em "Avaliação de impacto algorítmico no setor público brasileiro: um modelo complementar ao *framework* de autoavaliação ética", que propõe e simula um modelo prático de auditoria para sistemas de triagem de benefícios sociais, garantindo a identificação prévia de vieses para proteger as populações vulneráveis de discriminações algorítmicas.

No terceiro eixo, voltado à **educação, inteligência territorial e inclusão**, vemos os algoritmos mapearem e respeitarem a profunda diversidade do nosso país. Nelson Oliveira Santos assina a "Inteligência estratégica para a rede federal: perfil econômico e oferta educacional", aplicando inteligência artificial não supervisionada para avaliar o alinhamento entre os cursos oferecidos pelos institutos federais e as reais vocações econômicas de cada município, otimizando o gasto público em prol do desenvolvimento regional.

Focada na jornada de sucesso do estudante, Fabiana Ferreira do Nascimento apresenta a "Governança de dados para permanência estudantil: uma abordagem aplicada com ciência de dados", revelando como o aprendizado de máquina alimentado por dados acadêmicos bem estruturados pode identificar precocemente universitários em risco de evasão, permitindo intervenções assertivas.

O respeito à identidade dos povos originários é o cerne inspirador do trabalho de Luís César Seixas de Oliveira, intitulado "Classificação de línguas e etnias indígenas nos censos demográfico e agropecuário por meio de *machine learning*", que automatiza com alta precisão a codificação das respostas de texto livre do cidadão, agilizando a produção de estatísticas oficiais do IBGE sem perder o rigor antropológico.

Explorando a fronteira da inclusão financeira, Marcelo Cabreira Bastos apresenta a "Estimativa de propensão e exclusão no crédito rural: uma abordagem de *positive-unlabeled learning* e *gradient boosting* com dados do PRONAF e CAF", aplicando técnicas semissupervisionadas para avaliar quem realmente acessa ou é excluído das políticas de crédito voltadas à agricultura familiar.

Por fim, o quarto eixo foca na **gestão estratégica, no orçamento e na infraestrutura,** provando que a própria engrenagem do Estado pode e deve ser otimizada com ciência. Eduardo Melo Moura Filho e Alex Lopes Pereira propõem a "Inteligência orçamentária: um ambiente virtual de apoio à decisão na alocação de emendas parlamentares e na formulação de políticas públicas", desenhando uma plataforma analítica que conecta o comportamento parlamentar às reais lacunas municipais para uma distribuição de recursos mais racional e efetiva.

Lançando luz sobre o valor do capital humano, Rafael Gerardo Rodrigues do Nascimento analisa “Metas e produtividade no serviço público: uma avaliação quantitativa da atividade dos examinadores de marca do INPI”, provando empiricamente que a maturidade profissional e o acúmulo de inteligência cognitiva ao longo dos anos justificam a adoção de metas dinâmicas baseadas na experiência, combatendo a ineficiência de exigências estáticas. Para proteger a competitividade econômica do Brasil, Fábio Kouri Paim desenvolve o “Monitoramento e diagnóstico de eventos disruptivos nos portos brasileiros: detecção de anomalias operacionais e análise de exposição comercial”, interligando inteligência artificial e bases de dados abertos para detectar automaticamente paralisações logísticas e avaliar as vulnerabilidades sistêmicas das nossas cadeias de exportação.

Convidamos todos os gestores públicos, servidores, pesquisadores e cidadãos a mergulharem na leitura destas quatorze obras formidáveis. Elas representam muito mais do que a excelência acadêmica da turma de 2025; são um manual estimulante e um chamado definitivo à ação. Fica claro em cada página que o uso de análise de dados e o engajamento ético com a inteligência artificial não são meros luxos tecnológicos, mas sim a chave e o alicerce para construirmos um serviço público que não apenas reage, mas que antecipa soluções inovadoras, garantindo um futuro mais inteligente, ágil, justo e inegavelmente centrado na humanização e no desenvolvimento do Brasil.

Boa leitura!

## REFERÊNCIAS

CINGOLANI, Luciana. **The State of State Capacity: a review of concepts, evidence and measures**. Maastricht: UNU-MERIT Working Paper, 2013.

EUBANKS, Virginia. **Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor**. New York: St. Martin’s Press, 2018.

MAZZUCATO, Mariana. **The Value of Everything: Making and Taking in the Global Economy**. London: Allen Lane, 2018.

# Prefácio

**Natália Massaco Koga**

É extremamente oportuna esta publicação que retrata como o aprendizado em ciência de dados e inteligência artificial pode contribuir para a produção de conhecimento na construção de políticas públicas mais bem informadas. Vivemos um momento de transformações aceleradas e disruptivas promovidas pelo desenvolvimento de novas tecnologias. Em fevereiro de 2026, duas das principais empresas de IA do mundo lançavam novos modelos de capacidades inéditas que têm causado surpresa e espanto comparáveis aos vividos no prelúdio da pandemia da Covid-19. "Algo grande está acontecendo" na produção mundial de conhecimento (Shumer, 2026; Sampaio e Galdino, 2026). Esta publicação é uma demonstração de que o serviço público brasileiro não está alheio a esse movimento.

Com a inflexão recente introduzida pelo paradigma do agente de Inteligência Artificial — sistemas capazes de escrever código, executar análises, coletar dados e redigir textos de forma autônoma —, a questão deixa de ser de grau ou de velocidade das tarefas destinadas à máquina e passa a ser uma questão de natureza da atividade (Sampaio e Galdino, 2026). A velocidade das transformações não trata apenas do aumento da produtividade, mas essencialmente da alteração das práticas de trabalho e da construção do conhecimento que orienta as políticas públicas. Argumenta-se que a integração da IA ao ciclo de políticas públicas promove uma mudança fundamental na construção do conhecimento, uma reengenharia sociotécnica sem precedentes que instaura um novo regime epistêmico, sustentado por uma *inteligência híbrida* que emerge da interação entre humanos e máquinas (Filgueiras, 2025). Humanos e sistemas de IA coevoluem em ciclos de interação que modificam as próprias condições epistêmicas da decisão pública.

O trabalho de análise e avaliação de políticas públicas que se sintetiza na ideia do *policy advice* (aconselhamento em política pública) é, portanto, fortemente impactado nesse contexto. Função que surge com a própria criação do campo de estudos de políticas públicas, com o intuito de trazer o conhecimento técnico e científico para a tomada de decisão política (Lasswell, 1970), a atividade do analista vive um momento urgente de reflexão e debate sobre a adaptação de suas atividades e capacidades.

A literatura recente começa a discutir as possíveis contribuições da IA para o *policy advice*, como na identificação de problemas à previsão de eventos e à avaliação de alternativas, mas também seus limites. A produção de políticas públicas em democracias não é conformada pelo modelo racionalista de otimização, é permeada por ambiguidade de objetivos e disputas interpretativas e valorativas, não sanáveis pelo mero aumento da capacidade de processamento de informação e inapreensível para o treinamento de máquinas (Koenig e Wenzelburger, 2025).

As potencialidades da IA para o analista de políticas públicas são significativas precisamente porque sua atividade envolve tarefas complementares. Como sugere esse debate, o ganho não está em substituir integralmente o trabalho humano, mas em distribuí-lo de modo a destinar menos tempo às etapas rotineiras e mais tempo às partes do processo em que o julgamento do analista se diferencia da máquina (Sampaio e Galdino, 2026).

Nesse sentido, a IA pode contribuir no aconselhamento de políticas na produção de fontes de evidência via detecção, previsão e simulação, informando decisões sobre *timing* da ação, auxiliando na alocação e focalização de recursos, na escolha de instrumentos, entre outras finalidades. O trabalho do analista pode se beneficiar das novas tecnologias também para atuar não apenas de forma preventiva ou explicativa, mas em temporalidades próximas ao tempo real. Vários dos trabalhos aqui reunidos demonstram como a ciência de dados emerge como resposta possível à necessidade, já apontada na literatura, de um "monitoramento com métodos mistos" capaz de acompanhar ecossistemas digitais em ritmo próximo ao de sua própria dinâmica (Nascimento, Cesarino e Soares, 2026).

Mas as potencialidades só podem ser plenamente usufruídas se conhecidas as fragilidades e os riscos. Os grandes modelos de linguagem são treinados para parecerem convincentes, não para estarem certos. A esses riscos operacionais somam-se questões estruturais. As tecnologias de IA são montagens sociotécnicas situadas numa economia política global profundamente desigual de produção e regulação, e noções como transparência, explicabilidade e confiabilidade são objeto de disputa política, e não atributos técnicos garantidos (Paul, Carmel e Cobbe, 2024: Rikap, 2023).

Ademais, é necessário estar alerta para o fato de que a automação de processos decisórios pode introduzir novas formas de arbitrariedade e opacidade, tanto mais graves quanto mais escassos forem os mecanismos de contestação disponíveis ao cidadão afetado. Vieses podem ser embutidos nos dados, a opacidade dos modelos e as decisões normativas ocultas sob aparência de neutralidade técnica exigem escrutínio permanente. O desafio, portanto, não é apenas dominar as técnicas, mas governá-las, isto é, assegurar transparência metodológica, explicabilidade, auditabilidade dos critérios de decisão e avaliação sistemática de impacto algorítmico, de modo que a tecnologia sirva aos valores republicanos e ao interesse público, e não o contrário (Jannuzzi et al. 2024; 2026).

O que esse conjunto de análises aponta não é uma trajetória de extinção, mas de **redefinição das funções de aconselhamento e análise de políticas públicas**. As capacidades tradicionais do analista, tais como saber formular boas perguntas, buscar e validar informações, contextualizar evidências na dinâmica política, permanecem essenciais. A elas se somam novas competências, como as de elaborar instruções precisas, avaliar os resultados obtidos, detectar possíveis "alucinações" do sistema e saber quando recorrer a métodos tradicionais. E, subjacente à premência pelo desenvolvimento dessas diversas habilidades e conhecimentos, encontra-se a demanda por ações voltadas à construção e à sustentação da credibilidade e da confiança no próprio sistema estatal de *policy advice*.

O que o debate no campo recomenda, portanto, é que necessitamos formar uma nova geração de gestores e analistas equipados não só com domínio técnico, mas com capacidades interpretativas sofisticadas, habilidade de identificar e mitigar vieses, assim como com senso ético de responsabilização, transparência e justiça algorítmica para adaptar a governança de aconselhamento de políticas públicas aos novos tempos (Jannuzzi, 2026).

## A contribuição desta publicação

A dimensão dessa transformação fica evidente quando se compara esta obra com a publicação prévia. Em 2022, a Enap publicava *Ciência de Dados em Políticas Públicas: uma experiência de formação*, reunindo trabalhos da Especialização em Ciência de Dados aplicada a Políticas Públicas — trabalhos concluídos, vale lembrar, poucas semanas antes de a IA generativa irromper no debate público. Naquele volume, o repertório era o da ciência de dados consolidada: estatística descritiva e inferencial, aprendizado de máquina supervisionado e não supervisionado, processamento de linguagem natural em seus moldes então vigentes. A inteligência artificial comparecia, significativamente, ainda como *objeto* de estudo na análise da base normativa nacional de IA à luz de experiências internacionais, e não como instrumento de trabalho do analista.

Quatro anos depois, a mudança se inscreve no próprio nome do curso, agora MBA em Ciência de Dados e Inteligência Artificial Aplicadas, e sobretudo na natureza dos trabalhos: um eixo inteiro dedicado à IA generativa e aos agentes baseados em grandes modelos de linguagem, que deixam de ser tema de fronteira para se tornarem ferramenta cotidiana de consulta a bases públicas, classificação semântica em larga escala e construção de evidências. E a IA segue também como objeto agora de proposição concreta de modelos de avaliação de impacto algorítmico para o setor público brasileiro. O contraste entre os dois volumes é, em si mesmo, um registro histórico, pois documenta como, no curto intervalo entre uma turma e outra, aquilo que era objeto distante de regulação tornou-se instrumento incorporado à prática analítica, por meio da formação de servidores no uso das técnicas, mas também da reflexão crítica sobre seus usos.

O programa de MBA da Enap e esta publicação registram um enfoque específico de incidência neste contexto de transformação, a incidência por meio do desenvolvimento de capacidades de servidores públicos. Os servidores públicos são o elo essencial entre o conhecimento técnico e a tomada de decisão política. São eles que constroem, intermediam e estabilizam a capacidade analítica e a inteligência crítica do Estado. Não é possível prescindir deles, pois são as zonas de validação das escolhas pelos modelos e instrumentos e os fiadores da resiliência epistêmica das instituições públicas diante de um ambiente informacional cada vez mais em disputa.

Esta publicação demonstra, como muito bem retratado na Apresentação e na Introdução do livro, as potencialidades do uso da ciência de dados e da inteligência artificial produzidas por servidores e servidoras públicas, numa notável diversidade não apenas metodológica, mas também de tipos de contribuições e áreas de políticas analisadas.

Os trabalhos reunidos percorrem o repertório analítico contemporâneo, contemplando desde a fronteira mais recente da IA generativa e dos agentes baseados em grandes modelos de linguagem, passando pelo aprendizado de máquina em suas diversas formas, até os métodos estatísticos e econométricos clássicos. Essa convivência é notável, pois demonstra que o analista contemporâneo não substitui um repertório pelo outro, mas aprende a transitar entre eles, exercendo julgamento próprio sobre qual instrumento é adequado a qual problema, atributo nuclear da atividade de análise de políticas públicas.

O conjunto da obra e a trajetória do MBA demonstram, ainda, como a Enap está atenta às transformações em curso e ao compromisso de fortalecer a cultura do uso de evidências qualificadas para orientar as políticas públicas brasileiras, situando esses esforços no contexto epistêmico e institucional do nosso tempo.

É certo que os desafios contemporâneos são de natureza existencial. Inteligência e massa crítica necessitam ser acumuladas e mobilizadas de modo a trazer subsídios aos tomadores de decisão. Para que isso aconteça, diversos desafios para o desenvolvimento dessas capacidades analíticas de ordem sistêmica e organizacional precisam ser enfrentados, tais como as assimetrias e desigualdades entre sistemas internacionais produtores de conhecimento, barreiras de políticas de acesso e assinaturas de ferramentas analíticas, demanda pelo desenvolvimento de modelos e infraestrutura próprias, financiamento apropriado, entre tantos outros. Contudo, nenhum deles poderá ser enfrentado sem que haja analistas públicos detentores não apenas de capacidades de utilização dos métodos e tecnologias analíticas, mas também de julgamento das potencialidades e implicações de seu uso. Podemos dizer que o Estado nunca foi tão demandado a acelerar e sofisticar sua capacidade de produzir evidências e, simultaneamente, a fazê-lo sob condições de auditabilidade, explicabilidade e governança ética.

Esta publicação reflete a preocupação do programa com esse horizonte, assim como a consciência de seus organizadores, autores e autoras de que a construção de um Estado analiticamente capaz, tecnologicamente atualizado e democraticamente legítimo é tarefa coletiva e urgente. Que a leitura destes trabalhos inspire mais servidores e servidoras, assim como organizações e lideranças públicas, a assumi-la.



## REFERÊNCIAS

FILGUEIRAS, Fernando. **Artificial intelligence and public policy**. Cambridge: Cambridge University Press, 2025. (Elements in Public Policy).

JANNUZZI, Paulo de Martino. **Políticas públicas, valores e evidências em tempos de inteligência artificial**. Campinas: Alínea, 2024.

JANNUZZI, Paulo de Martino et al. Informação algorítmica e políticas públicas: potencialidades, vieses e riscos da inteligência artificial generativa. In: KOGA, N. M. et al. (org.). **Políticas públicas informadas por evidências:** conhecimentos, arranjos e capacidades para a governança democrática. Brasília: Ipea, 2026.

KOENIG, Pascal D.; WENZELBURGER, Georg. The potentials and limitations of using artificial intelligence for policymaking. In: HOWLETT, M.; MUKHERJEE, I. (ed.). **Handbook of policy advice**. Cheltenham: Edward Elgar, 2025. p. 622-637.

LASSWELL, H.D. The emerging conception of the policy sciences. **Policy Sci 1**, 3–14 (1970). https://doi.org/10.1007/BF00145189

NASCIMENTO, Leonardo Fernandes; CESARINO, Letícia Maria Costa da Nóbrega; SOARES, Alisson. Ecossistema de desinformação e radicalização: repercussões no policymaking. In: KOGA, N. M. et al. (org.). **Políticas públicas informadas por evidências**: conhecimentos, arranjos e capacidades para a governança democrática. Brasília: Ipea, 2026.

PAUL, Regine; CARMEL, Emma; COBBE, Jennifer (ed.). **Handbook on public policy and artificial intelligence**. Cheltenham: Edward Elgar, 2024.

RIKAP, C. (2023). **Intellectual monopolies as a new pattern of innovation and technological regime. Industrial and Corporate Change**. https://academic.oup.com/icc/advance-article-abstract/doi/10.1093/icc/dtad077/7462137

SAMPAIO, Rafael Cardoso; GALDINO, Manoel. Algo grande está acontecendo na ciência mundial e o Brasil parece estar de fora novamente. **SciELO em Perspectiva**, 24 abr. 2026.

SHUMER, Matt. **Something Big Is Happening.** 2026. Disponível em: https://shumer.dev/something-big-is-happening.

# Introdução

**Alex Lopes Pereira**
**Hélio Bomfim de Macêdo Filho**

A incorporação da Ciência de Dados e da Inteligência Artificial à Administração Pública não se resume à adoção de novas ferramentas computacionais. Ela exige capacidades institucionais para formular problemas, organizar e administrar dados, avaliar evidências e compreender seus efeitos sobre a atuação do Estado e os direitos dos cidadãos. É nesse contexto que se insere esta obra, *Ciência de Dados e IA a Serviço da Gestão Pública*, composta por trabalhos desenvolvidos no âmbito do MBA em Ciência de Dados e Inteligência Artificial Aplicadas da Escola Nacional de Administração Pública (Enap).

A publicação dá continuidade à trajetória iniciada em edições anteriores da especialização em Ciência de Dados Aplicada a Políticas Públicas. Desde sua concepção, a formação articula três grandes eixos - 1) Computação, 2) Matemática e Estatística e 3) Políticas Públicas - por meio de disciplinas, palestras, laboratórios de casos e trabalhos de conclusão de curso.

Nas edições anteriores, os participantes já empregavam estatística descritiva e inferencial, aprendizado de máquina supervisionado e não supervisionado, detecção de anomalias e sistemas de informação geográfica em temas como saúde, educação, meio ambiente, regulação, auditoria e compras públicas.

A presente edição representa um aprofundamento dessa trajetória diante da rápida difusão dos grandes modelos de linguagem, os *Large Language Models* - LLM, e de outras formas de Inteligência Artificial Generativa. A Ciência de Dados permanece como fundamento indispensável para as políticas públicas. Entretanto, a Inteligência Artificial passa a aparecer simultaneamente como instrumento de aprendizagem e ferramenta individual de produtividade. Essa transformação amplia o conjunto de competências exigidas dos agentes públicos e modifica também as formas de ensinar, desenvolver e avaliar soluções computacionais.

A atualização do projeto pedagógico do MBA refletiu esse movimento. A reformulação contemplou a revisão das ementas para incorporar conhecimentos, habilidades e competências relacionados à Inteligência Artificial transversalmente em todas as disciplinas, além da criação da disciplina de Inteligência Artificial Generativa no contexto da Administração Pública.

A Inteligência Artificial Generativa acrescenta uma nova dimensão a esse desafio. Ferramentas capazes de produzir códigos, explicar conceitos e resolver exercícios podem reduzir barreiras de entrada e acelerar determinadas etapas da aprendizagem. Ao mesmo tempo, podem permitir que estudantes obtenham respostas aparentemente corretas sem

compreender o problema, a lógica da solução ou o funcionamento do código apresentado. Para quem vivencia o primeiro contato com programação, a delegação prematura da atividade intelectual à IA pode comprometer a construção dos modelos mentais necessários para avaliar os resultados gerados.

Esse deslocamento - do domínio da ferramenta para o domínio do problema - apareceu na sala de aula. Na disciplina introdutória de programação, o exercício avaliado passou a ser o *prompt*, com o código gerado executado contra gabarito determinístico: erra quem especifica mal o problema, não quem esquece a sintaxe. A inversão é modesta em escala, mas nomeia o que está em jogo nos trabalhos reunidos nesta publicação. Em todos eles, o mérito não está no LLM empregado, e sim na formulação da pergunta, na escolha das informações que fundamentam a resposta e na verificação do resultado.

Formar agentes públicos, atualmente, exige mais do que familiaridade com tecnologias específicas. Exige capacidade para formular bons problemas, o interesse por resolvê-los, bem como reconhecer limitações dos dados e, especialmente, dos LLMs. Exige também que as organizações públicas criem condições para transformar os conhecimentos desenvolvidos nas atividades acadêmicas em capacidades permanentes, sem confundir protótipos, ferramentas auxiliares e processos oficiais de tomada de decisão.

Esta publicação representa uma etapa desse processo. Ao reunir estudos aplicados produzidos por profissionais que vivenciam problemas concretos da Administração Pública, a obra aproxima formação acadêmica, experimentação tecnológica e necessidades do Estado. A tabela presente nas próximas duas páginas resume os trabalhos desta edição do MBA.

Desejamos que os capítulos desta publicação contribuam para a disseminação (inclusive responsável) da Ciência de Dados e da Inteligência Artificial no setor público brasileiro e fortaleçam uma cultura administrativa orientada por evidências. Sobretudo, espera-se consolidar a compreensão de que a inovação pública não decorre apenas da disponibilidade de LLMs mais potentes, mas de competências em formular problemas relevantes, construir sistemas confiáveis e avaliar seus impactos.

| Eixo | Temática | Título e autoria | Breve descrição | Técnica principal |
|---|---|---|---|---|
| **PLN e IA generativa** | **Educação** | *Análise automatizada de avaliações educacionais em larga escala*<br>Francisco Carlos Molina Duarte Júnior | Analisa 3,8 milhões de comentários das avaliações de cursos da Enap e propõe um Índice de Urgência e um Mapa de Risco Educacional para priorizar intervenções. | Classificação semântica por LLM |
| | **Justiça e direitos humanos** | *Benchmark jurídico e arquitetura documental orientada por evidência em casos brasileiros da Corte IDH*<br>Luiz Víctor do Espírito Santo Silva | Constrói um *benchmark* em português com 55 itens auditados e compara quatro regimes de acesso à informação em tarefas jurídicas objetivas. | *Benchmark* auditado e arquitetura documental orientada por evidência |
| | **Saúde — compras públicas** | *Consulta em linguagem natural a dados de compras centralizadas de medicamentos do Ministério da Saúde*<br>Rosana Elisa Gonçalves Pinho | Desenvolve e avalia um portal que converte perguntas em português em consultas SQL sobre a base de compras de medicamentos, testado com usuários. | *Text-to-SQL* com LLM |
| | **Transportes** | *Um agente inteligente baseado em LLM para consulta em linguagem natural de dados aquaviários da ANTAQ*<br>Tadeu Gomes Teixeira | Constrói um agente que interpreta perguntas sobre estatísticas aquaviárias, gera e executa as consultas e devolve os resultados explicados. | Agente *text-to-SQL* com orquestração e recuperação de exemplos |
| **2. Saúde pública, sustentabilidade e impacto social** | **Governança de IA** | *Avaliação de impacto algorítmico no setor público brasileiro*<br>Camila Falchetto Romero | Propõe um modelo de avaliação de impacto algorítmico complementar ao *framework* federal de autoavaliação ética, com matriz de risco em direitos fundamentais. | *Design Science Research* |
| | **Meio ambiente** | *Desmatamento nos municípios da Amazônia Legal*<br>Juliana Pierrobon Lopez | Identifica os fatores associados ao desmatamento municipal, entre 2012 e 2021, com destaque para a inércia temporal e a fiscalização ambiental. | XGBoost com interpretação por SHAP |
| | **Saúde — imunização** | *Determinantes da cobertura vacinal nos municípios brasileiros*<br>Tadeu Gomes Teixeira | Compara o peso das condições socioeconômicas e do comportamento eleitoral na cobertura vacinal dos 5.570 municípios brasileiros. | Regressão multivariada com diagnósticos espaciais |

| Eixo | Temática | Título e autoria | Breve descrição | Técnica principal |
|---|---|---|---|---|
| **3. Educação, inteligência territorial e inclusão** | **Educação superior** | *Governança de dados para permanência estudantil*<br>Fabiana Ferreira do Nascimento | Avalia como a qualidade e a integração dos dados institucionais afetam a identificação precoce de estudantes em risco de evasão. | Modelos supervisionados comparados |
| | **Povos indígenas** | *Classificação de línguas e etnias indígenas nos censos demográfico e agropecuário*<br>Luís César Seixas de Oliveira | Desenvolve modelos que apoiam o processo de codificação das línguas e etnias na atividade censitária do IBGE. | Classificação supervisionada em alta cardinalidade |
| | **Crédito rural** | *Estimativa de propensão e exclusão no crédito rural*<br>Marcelo Cabreira Bastos | Estima a propensão ao crédito entre agricultores sem acesso observado, distinguindo ausência de perfil de exclusão por barreiras. | *Positive-Unlabeled Learning* |
| | **Educação profissional** | *Inteligência estratégica para a Rede Federal*<br>Nelson Oliveira Santos | Confronta a oferta de cursos da Rede Federal com as vocações econômicas dos municípios, medindo aderência e divergência. | Agrupamento por *K*-Means |
| **4. Gestão estratégica, orçamento e infraestrutura** | **Orçamento** | *Inteligência orçamentária*<br>Eduardo Melo Moura Filho | Propõe ambiente virtual que caracteriza o comportamento alocativo dos parlamentares e sugere alocações a partir das necessidades municipais. | Métricas de concentração e regressão logística |
| | **Infraestrutura portuária** | *Monitoramento e diagnóstico de eventos disruptivos nos portos brasileiros*<br>Fábio Kouri Paim | Detecta e classifica anomalias operacionais em 18 portos e mapeia a exposição das cadeias exportadoras a disrupções. | Detecção de anomalias em séries temporais |
| | **Propriedade industrial** | *Metas e produtividade no serviço público*<br>Rafael Gerardo Rodrigues do Nascimento | Avalia como experiência, idade e jornada afetam o tempo de exame de marcas no INPI e propõe metas dinâmicas por faixa de experiência. | Testes não paramétricos e *Random Forest* |

**Fonte:** elaboração própria a partir dos capítulos da coletânea.

# EIXO 01

## *Processamento de Linguagem Natural e da Inteligência Artificial Generativa*

MBA ENAP

# Análise automatizada de avaliações educacionais em larga escala: uma abordagem baseada em processamento de linguagem natural

*Francisco Carlos Molina Duarte Júnior*
*Prof. Dr. Alex Lopes Pereira*

## Resumo

Este estudo propõe uma abordagem automatizada para análise qualitativa das avaliações de reação dos cursos da Escola Nacional de Administração Pública (Enap), cujo volume massivo de *feedbacks* textuais inviabiliza a leitura manual e cujas métricas numéricas tradicionais mascaram fragilidades pedagógicas pelo efeito teto. A metodologia combina um pipeline *Extract, Load, Transform* (ELT) orquestrado pelo Apache Airflow com técnicas de Processamento de Linguagem Natural (PLN), incluindo filtro de densidade cognitiva e classificação semântica por LLM (Gemini), validada com coeficiente Kappa de Cohen de 0,81. Aplicada a 3.859.221 comentários textuais, a solução selecionou 74.285 registros de alta relevância diagnóstica e identificou que 37% das avaliações apontam falhas estruturais, concentradas em cursos de transformação digital, governança de dados e segurança da informação. Conclui-se que o Índice de Urgência e o Mapa de Risco Educacional propostos oferecem à gestão institucional um instrumento escalável e orientado por evidências para priorização objetiva de intervenções pedagógicas e tecnológicas.

**Palavras-chave:** Processamento de Linguagem Natural; Avaliação Educacional; Análise de Sentimentos; Gestão de Risco Educacional; Ciência de Dados.

## INTRODUÇÃO

O crescimento contínuo das plataformas educacionais da Escola Nacional de Administração Pública (Enap) tem ampliado de forma significativa o número de servidores públicos atendidos por suas ações de capacitação. A Escola Virtual de Governo (EVG), em especial, consolidou-se como um dos principais instrumentos de formação em larga escala no âmbito da administração pública federal, ofertando cursos em diferentes modalidades e áreas temáticas para um público cada vez mais amplo e diversificado. Destacam-se também os cursos remotos e presenciais pela qualidade e alta procura pelos servidores públicos.

Nesse contexto, a avaliação sistemática da qualidade dos cursos ofertados torna-se um elemento estratégico para assegurar a efetividade das políticas de capacitação e o aprimoramento contínuo das ações educacionais. O aumento do volume de participantes implica, consequentemente, um crescimento expressivo na quantidade de avaliações de reação registradas, tanto em formatos numéricos quanto em campos de respostas abertas.

Tradicionalmente, a mensuração da satisfação dos participantes baseia-se em indicadores quantitativos, como médias de notas ou índices de recomendação *Net Promoter Score* (NPS). Embora esses indicadores sejam adequados para o acompanhamento global da percepção dos alunos, eles apresentam limitações relevantes em ambientes nos quais as avaliações tendem a se concentrar em valores elevados, fenômeno conhecido como efeito teto (Nunnally & Bernstein, 1994). Nessas situações, a análise exclusivamente numérica torna-se pouco sensível para identificar fragilidades dos cursos e orientar ações corretivas mais direcionadas.

Em complemento às avaliações quantitativas, os instrumentos de avaliação de reação da Enap incluem campos de respostas abertas, nos quais os participantes descrevem livremente percepções, críticas, sugestões e elogios. A literatura sobre avaliação de treinamento no setor público brasileiro destaca que esses comentários constituem uma fonte rica de informação qualitativa sobre aspectos pedagógicos, tecnológicos e organizacionais dos cursos, podendo contribuir para a melhoria contínua das ações educacionais (Abbad, Borges-Andrade & Gama, 2000). Contudo, o volume de dados textuais gerado pelo alcance massivo das plataformas educacionais inviabiliza a leitura manual, isso faz com que grande parte desse conteúdo permaneça subexplorada no processo de melhoria dos cursos.

Diante desse cenário, técnicas de Ciência de Dados e Processamento de Linguagem Natural (PLN) apresentam-se como alternativas viáveis para automatizar a análise de grandes volumes de texto (Jurafsky & Martin, 2023; Devlin et al., 2019), permitindo extrair padrões semânticos, sentimentos e temas recorrentes de forma escalável e reprodutível. A aplicação dessas técnicas possibilita transformar dados textuais não estruturados em informações analíticas capazes de subsidiar o trabalho da gestão educacional baseada em evidências.

Assim, o presente trabalho tem como objetivo desenvolver e aplicar uma abordagem automatizada para a análise qualitativa das avaliações de reação dos cursos ofertados pela Escola Virtual de Governo da Enap. A proposta consiste na integração de um pipeline de engenharia de dados com modelos de linguagem baseados em *deep learning*, visando classificar sentimentos, identificar tópicos críticos e priorizar cursos que demandam intervenção pedagógica ou tecnológica.

Como contribuição principal, o estudo propõe um modelo de gestão de risco educacional orientada por PLN, no qual a análise qualitativa dos comentários dos alunos complementa os indicadores quantitativos tradicionais. Essa abordagem permite não apenas monitorar a satisfação geral, mas também identificar, de forma objetiva e escalável, os fatores que comprometem a experiência do aluno, oferecendo subsídios mais precisos para a tomada de decisão gerencial no âmbito da Enap.

O artigo está organizado da seguinte forma: a Seção 1 descreve a metodologia e a arquitetura da solução; a Seção 2 apresenta os resultados quantitativos e as técnicas de PLN aplicadas; a Seção 3 traz a análise qualitativa das avaliações de reação; a Seção 4 sintetiza o mapeamento do risco educacional; a Seção 5 discute trabalhos futuros; e, por último, a conclusão do trabalho.

# 1. METODOLOGIA

## 1.1 Abordagem Geral

Este estudo adota uma abordagem quantitativa e qualitativa (Han, Kamber & Pei, 2011), com foco no desenvolvimento de uma solução automatizada para a análise de avaliações de reação em larga escala. Do ponto de vista metodológico, o trabalho combina técnicas de engenharia de dados, aprendizado de máquina e PLN, estruturada sobre um fluxo de engenharia de dados analítico contínuo, capaz de transformar dados educacionais brutos em informações estratégicas para a gestão.

A pesquisa caracteriza-se ainda como descritiva e exploratória, uma vez que busca compreender padrões de satisfação e insatisfação a partir de dados históricos de avaliações, ao mesmo tempo em que explora o uso de modelos de linguagem para extração de conhecimento a partir de textos não estruturados. A adoção de métodos mistos — combinando dados quantitativos e análise qualitativa de texto — segue a abordagem consolidada de Creswell (2014) e Creswell e Clark (2013).

## 1.2 Arquitetura e Pipeline de Dados

A solução foi concebida com base no padrão de arquitetura ELT (Reis & Housley, 2022), no qual os dados são inicialmente extraídos e consolidados em um repositório central, para posteriormente serem submetidos ao processamento analítico. Essa abordagem foi escolhida por favorecer escalabilidade, desempenho e desacoplamento entre sistemas transacionais e analíticos. O fluxo de dados opera de forma cíclica e automatizada, com periodicidade diária (D–1), garantindo que as análises reflitam o estado mais recente das avaliações realizadas pelos participantes dos cursos.

A arquitetura lógica da solução é composta por três camadas principais:

- **Camada de Ingestão e Consolidação:** responsável pela extração e unificação dos dados provenientes dos sistemas educacionais da Enap.
- **Camada de Processamento e Inteligência:** onde ocorre a orquestração dos fluxos e a aplicação dos algoritmos de PLN e aprendizado de máquina.
- **Camada de Apresentação:** destinada à visualização interativa dos resultados por meio de painéis analíticos no Power BI.

O Apache Airflow (Apache Software Foundation, 2024) foi utilizado como orquestrador dos fluxos de dados, gerenciando a sequência de tarefas, o agendamento de execuções e o monitoramento de falhas. Essa escolha justifica-se pela maturidade da ferramenta em ambientes de produção, pela facilidade de integração com bancos de dados PostgreSQL e pela capacidade de escalonamento horizontal (Harenslak & Ruiter, 2021).

### 1.2 Fontes de Dados

O ecossistema educacional da Enap é composto por múltiplas plataformas, com diferentes modelos de dados e finalidades operacionais. As avaliações de reação analisadas neste estudo originam-se de dois sistemas principais, ambos baseados em bancos de dados relacionais PostgreSQL (PostgreSQL Global Development Group, 2024):

- **Moodle:** ambiente virtual de aprendizagem utilizado predominantemente para cursos a distância (EAD), responsável pelo maior volume de avaliações.
- **SUAP (Sistema Unificado de Administração Pública):** sistema institucional utilizado para a gestão acadêmica de cursos presenciais e remotos síncronos.

Devido às diferenças estruturais entre os esquemas de dados dessas plataformas, optou-se pela criação de uma *View Materializada* em um banco de dados central. Essa estrutura atua como um *data mart* intermediário, consolidando os registros em um esquema unificado, contendo, entre outros atributos, identificador do curso, data da avaliação, nota numérica e resposta textual aberta. A utilização de uma *view materializada* justifica-se pela necessidade de desempenho e isolamento das cargas analíticas, evitando impactos sobre os bancos de dados transacionais de produção e permitindo que consultas sejam realizadas sobre grandes volumes de dados sem impactar os sistemas que os geram.

## 2. RESULTADOS E ANÁLISE QUANTITATIVA

A etapa inicial da análise concentrou-se na exploração dos dados quantitativos consolidados na *view materializada*. É importante esclarecer a distinção entre os volumes de dados utilizados em cada etapa deste trabalho: a base de avaliações quantitativas – NPS - é composta por 55.814 registros; a base de comentários textuais abertos, provenientes das avaliações de reação, totaliza 3.859.221 registros; após a aplicação do filtro de densidade cognitiva (descrito na Seção 3.5), foram selecionados 74.285 comentários de alta relevância analítica.

### 2.1 Caracterização da Base

A análise estatística descritiva da base consolidada de avaliações revelou um cenário de elevada aprovação institucional. A nota média global observada foi de 9,22, em uma escala de 0 a 10, acompanhada de um baixo desvio padrão de 0,43. Esses resultados indicam não apenas altos níveis de satisfação por parte dos participantes, mas também baixa dispersão entre as avaliações, sugerindo consistência na percepção da qualidade dos cursos ofertados.

A distribuição das notas apresenta forte concentração nos valores superiores da escala, especialmente entre 9 e 10, caracterizando o chamado efeito teto. Esse padrão indica que a maioria dos cursos é percebida como excelente ou muito boa, independentemente da modalidade, temática ou diretoria responsável. Embora esse comportamento seja positivo do ponto de vista institucional, ele impõe limitações à sensibilidade da análise puramente quantitativa para a identificação de fragilidades pontuais, reforçando a necessidade de abordagens complementares.

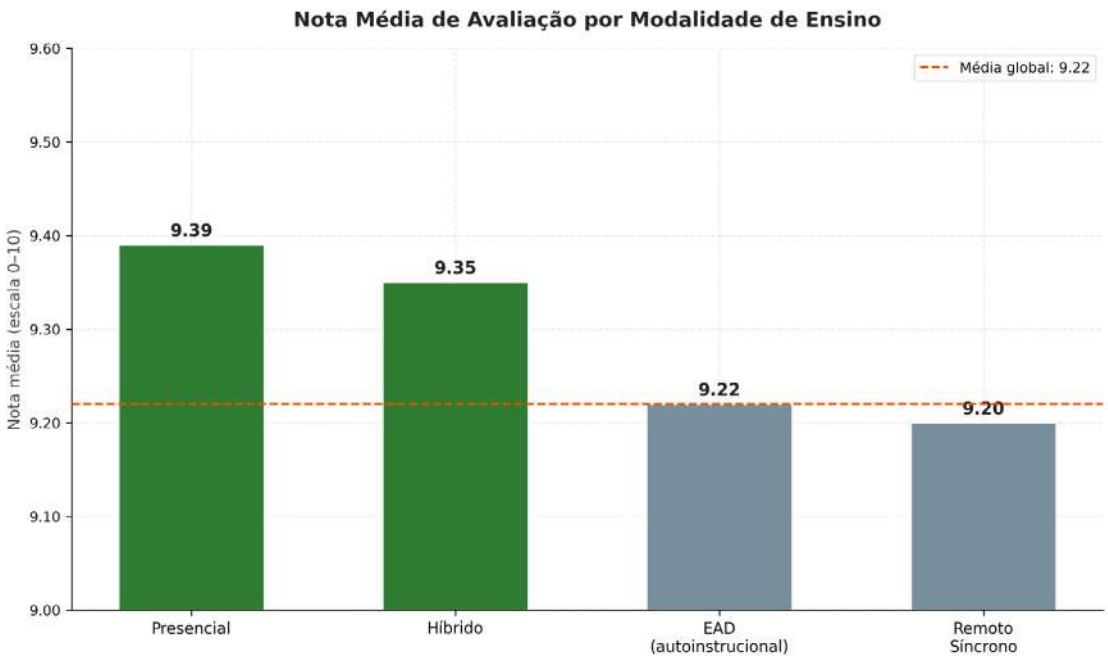
## 2.2 Resultados por Modalidade, Diretoria e Temática

A segmentação das avaliações por modalidade de ensino evidencia uma notável homogeneidade na qualidade percebida, com variações marginais entre os formatos analisados. Os cursos presenciais apresentaram a maior média de avaliação (9,39), possivelmente associada à interação direta entre participantes e instrutores, à troca de experiências e ao *networking* proporcionado por esse formato. Os cursos híbridos mantiveram desempenho próximo (9,35), enquanto a modalidade a distância (EAD) registrou média de 9,22, demonstrando maturidade na oferta de cursos autoinstrucionais em larga escala. Já os cursos remotos síncronos apresentaram média de 9,20, indicando estabilidade na mediação tecnológica das atividades.

A análise por diretoria responsável pelos cursos reforça o padrão de elevada e homogênea satisfação. A Diretoria de Altos Estudos (DAE) apresentou a maior média de avaliação (9,42), enquanto a Diretoria de Desenvolvimento Profissional (DDPRO), responsável pelo maior volume de cursos — especialmente na modalidade EAD —, manteve uma média sólida de 9,22. Em termos temáticos, áreas como Saúde (9,39) e Políticas Sociais (9,37) destacaram-se ligeiramente acima da média geral como mostrado no Gráfico 1.

Destaca-se que a correlação entre o número de respondentes e a nota média mostrou-se praticamente irrelevante, com coeficiente aproximado de 0,03, indicando que o volume de participantes não é determinante para a qualidade percebida. Esse achado reforça a necessidade de métricas complementares para identificar, dentro de cursos com alta nota média, subconjuntos de participantes com percepções críticas.

**Gráfico 1 – Nota média de avaliação por modalidade de ensino**




Fonte: Elaboração própria com base nos dados das avaliações de reação da Enap (2024).

## 2.3 Limitações da Análise Quantitativa

Embora os indicadores numéricos confirmem a excelência geral da Enap, a análise quantitativa apresenta limitações relevantes do ponto de vista da gestão da melhoria contínua. O principal desafio identificado é o chamado efeito teto, caracterizado pela concentração das notas nos valores máximos da escala. Nesse contexto, a média aritmética torna-se uma métrica pouco sensível para detectar falhas pontuais ou oportunidades específicas de aprimoramento. Cursos com nota média elevada podem, simultaneamente, concentrar um volume significativo de comentários críticos relacionados a aspectos como navegação, material desatualizado ou dificuldades operacionais.

Além das limitações estatísticas impostas pelo efeito teto, identifica-se um desafio adicional inerente à natureza das avaliações de reação: o baixo engajamento cognitivo dos respondentes no preenchimento de campos abertos. Observa-se que, com frequência, os usuários tendem a fornecer respostas superficiais ou protocolares com o intuito de concluir rapidamente o processo avaliativo, gerando um volume considerável de ruído informacional. Para mitigar esse viés e isolar as contribuições de maior valor analítico, este estudo implementou um filtro de densidade cognitiva baseado em três dimensões linguísticas:

- **Extensão do Discurso (Comprimento):** utilizado como indicador primário de disposição do respondente em detalhar sua percepção.
- **Diversidade Lexical (Riqueza):** razão entre palavras únicas e o total de termos, visando identificar vocabulário específico e evitar repetições vazias.
- **Articulação Lógica (Conectores):** presença de operadores argumentativos (como “porém”, “consequentemente” ou “além disso”), que sinalizam a elaboração de um raciocínio estruturado sobre a experiência educacional.

Ao cruzar esses critérios, o modelo foi capaz de distinguir as respostas pensadas - aquelas que exigiram maior investimento de tempo e reflexão do aluno - das respostas automáticas. Essa filtragem, aplicada sobre a base de 3.859.221 comentários textuais, resultou na identificação de 74.285 comentários de alta relevância, evidenciando a eficácia do filtro em reduzir ruído informacional e concentrar o esforço analítico nos *feedbacks* com maior potencial diagnóstico.

**Tabela 1 – Exemplos de comentários e classificação segundo a qualidade textual**

| Exemplo de Comentário | Palavras | Riqueza | Conect. | Classificação | Justificativa |
|---|---|---|---|---|---|
| Muito bom o curso, gostei bastante. | 6 | 0,83 | 0 | RUÍDO | Abaixo do limite de palavras e sem conectores. |
| O curso é bom. O curso é fácil. O curso é legal. | 12 | 0,33 | 0 | RUÍDO | Embora longo, possui baixa riqueza (repetitivo) e falta de articulação. |
| O conteúdo é denso, embora a interface seja confusa; portanto, sugiro melhorias no layout. | 13 | 0,92 | 2 | PENSADA | Atende a todos os requisitos: extensão, vocabulário variado e nexo lógico. |

**Fonte:** Elaboração própria com base nos dados das avaliações de reação da Enap (2024).

## 2.4 Índice de Urgência

Para apoiar a tomada de decisão gerencial, foi desenvolvido um indicador sintético denominado Índice de Urgência (IU), cujo objetivo é priorizar cursos que demandam intervenção pedagógica ou tecnológica. O índice que combina a severidade das avaliações negativas com o volume de participantes impactados é definido pela seguinte expressão:

$$IU = (Ncrit / Ntotal) \times ln(1 + Ntotal)$$

*Ncrit* representa o número de avaliações classificadas como críticas, *Ntotal* corresponde ao total de avaliações do curso e *ln* denota o logaritmo natural. Essa abordagem evita distorções associadas ao uso exclusivo de contagens absolutas ou proporções relativas, garantindo equilíbrio entre relevância estatística e gravidade qualitativa. O componente logarítmico funciona como fator de ponderação do impacto institucional: cursos com maior número de participantes recebem peso proporcionalmente maior, sem que cursos menores com alta taxa crítica sejam desconsiderados.

## 2.5 Análise de Linguagem Natural

A análise qualitativa das avaliações de reação fundamenta-se no uso de técnicas de PLN, campo interdisciplinar situado na interseção entre Ciência da Computação, Inteligência Artificial e Linguística, cujo objetivo é capacitar sistemas computacionais a processar e interpretar a linguagem humana em larga escala (Jurafsky & Martin, 2023). Diferentemente de dados estruturados, os comentários textuais produzidos pelos participantes apresentam elevada variabilidade lexical, ambiguidade semântica e dependência de contexto, o que impõe desafios adicionais à extração sistemática de conhecimento.

Para viabilizar a análise automatizada desses textos, foi adotado um pipeline de PLN composto por etapas de pré-processamento, modelagem e interpretação. O pré-processamento incluiu normalização textual, remoção de caracteres especiais, padronização ortográfica e filtragem de termos irrelevantes, com o objetivo de reduzir ruídos e enfatizar o conteúdo semântico informativo. Esse procedimento é essencial para aumentar a robustez dos modelos aplicados, especialmente em cenários de linguagem informal e heterogênea, característicos de avaliações espontâneas de usuários (Jurafsky & Martin, 2023; Liu, 2012).

### 2.5.1 Análise de Sentimentos

A análise de sentimentos, também conhecida como mineração de opinião, tem como objetivo identificar e classificar a polaridade expressa em textos, categorizando-os, de forma geral, como positivos, negativos ou neutros (Liu, 2012). Em contextos educacionais, essa técnica viabiliza a triagem automática de grandes volumes de avaliações, permitindo a identificação célere de cursos ou ações formativas que concentram níveis elevados de insatisfação.

Neste estudo, a análise de sentimentos foi realizada por meio do modelo de linguagem de grande escala LLM (Gemini), baseado na arquitetura Transformer com capacidades generativas avançadas (Google DeepMind, 2023). A escolha por essa tecnologia justifica-se pela sua superioridade em relação aos modelos de representação tradicionais, uma vez que o Gemini demonstra maior robustez na interpretação de relações semânticas

contextuais em textos curtos, informais ou ruidosos. A validação da classificação foi realizada por meio do coeficiente Kappa de Cohen ($\kappa$ = 0,81), indicando concordância substancial entre as classificações automáticas e as revisões humanas independentes, o que confere confiabilidade ao modelo adotado.

### 2.5.2 Análise Baseada em Modelos de Linguagem de Grande Escala (LLMs)

Os avanços recentes em modelos de linguagem de grande escala têm ampliado significativamente as possibilidades de análise textual automatizada. Esses modelos, também denominados *foundation models*, são treinados sobre grandes volumes de dados textuais e aprendem representações contextuais profundas da linguagem, permitindo a realização de múltiplas tarefas linguísticas sem necessidade de treinamento específico para cada uma (Bommasani et al., 2021).

Neste trabalho, utilizou-se a LLM Gemini, desenvolvida pelo Google DeepMind, como componente central da análise qualitativa e classificação semântica dos comentários. O Gemini destaca-se por sua arquitetura multimodal e capacidades avançadas de raciocínio (*reasoning*), o que permite a compreensão, síntese e interpretação de textos complexos sob uma perspectiva densamente contextualizada (Google DeepMind, 2023). A utilização de modelos baseados em Transformers fundamenta-se na arquitetura seminal proposta por Vaswani et al. (2017).

A integração do Gemini ao processamento das avaliações de reação viabilizou a geração de resumos interpretativos de alta fidelidade, o agrupamento semântico por similaridade de intenção e a identificação das causas-raiz subjacentes à insatisfação dos participantes. Tal abordagem foi determinante para converter o volume massivo de dados em informações qualitativas acionáveis, permitindo que a gestão educacional transite de uma visão puramente descritiva para uma análise diagnóstica e prescritiva.

## 2.6 Priorização de Cursos com Índice de Urgência

Com base nos comentários classificados como críticos, o IU foi aplicado para gerar um *ranking* objetivo de cursos que demandam intervenção pedagógica ou tecnológica. O índice equilibra a proporção de avaliações críticas com o volume de participantes impactados, evitando as distorções associadas tanto ao uso de contagens absolutas — que beneficiariam artificialmente os cursos maiores — quanto ao uso exclusivo de proporções relativas — que amplificariam a relevância de cursos com baixa representatividade amostral.

A Tabela 2 apresenta os dez cursos posicionados no quadrante de alta criticidade e alto impacto. Observa-se que os cursos com maiores taxas críticas concentram-se majoritariamente em temáticas de transformação digital, governança de dados e segurança da informação — - áreas de elevada complexidade conceitual e forte dependência de infraestrutura tecnológica. Esse padrão sugere que cursos cujo conteúdo evolui rapidamente e cujo acesso depende de ambientes tecnológicos mais sofisticados tendem a gerar maior volume de comentários críticos, especialmente quando há desalinhamento entre as expectativas dos participantes e a abordagem pedagógica adotada.

**Tabela 2 – Cursos posicionados no quadrante de alta criticidade e alto impacto**

| Curso | Taxa de avaliações críticas (%) | Total de avaliações |
|---|---|---|
| Proteção ao Denunciante e Tratamento de Denúncias em Ouvidoria | 51,83 | 4.445 |
| Transformação Digital no Serviço Público | 49,43 | 4.578 |
| Fundamentos da transformação digital: mapeamento e automação de processos | 48,58 | 3.510 |
| Colaboração com o Microsoft 365 | 48,13 | 2.668 |
| Elaboração de Indicadores de Desempenho Institucional | 47,95 | 2.338 |
| Educação Patrimonial e Inventários Participativos | 47,22 | 2.283 |
| Governança de Dados na Transformação Digital | 46,65 | 1.987 |
| Era Digital, Planejamento Estratégico e Inovação na Gestão Pública | 46,18 | 2.605 |
| Segurança da Informação no contexto da transformação digital | 46,06 | 2.770 |
| Contratações Públicas Sustentáveis | 46,02 | 1.821 |

**Fonte:** Elaboração própria com base nos dados das avaliações de reação da Enap (2024).

# 3. ANÁLISE QUALITATIVA DAS AVALIAÇÕES DE REAÇÃO

A análise qualitativa das avaliações de reação representa o núcleo deste trabalho, pois permite transcender as limitações inerentes às métricas numéricas de satisfação. Enquanto a análise quantitativa fornece uma visão macro da percepção dos alunos, é nos comentários textuais livres que emergem os sinais precoces de insatisfação, falhas operacionais e oportunidades concretas de melhoria pedagógica. Diante do elevado volume de dados textuais — composto por 3.859.221 registros de comentários abertos —, a leitura manual torna-se inviável. Nesse contexto, a aplicação de técnicas de PLN mostrou-se essencial para a extração sistemática e escalável de conhecimento a partir de dados não estruturados (Jurafsky & Martin, 2023).

## 3.1 Classificação Semântica dos Comentários

A base utilizada nesta etapa é composta por 3.859.221 comentários abertos, provenientes das avaliações de reação dos cursos da Enap. Cada registro contém, além do texto livre, metadados relevantes, como curso, critério avaliado, data da avaliação, classificação de sentimento, nível de confiança do modelo e categoria de risco educacional. Os comentários estão associados a diferentes critérios de avaliação, tais como conteúdo e estrutura do curso, interface gráfica, aplicabilidade prática e apoio logístico, possibilitando análises segmentadas e comparativas.

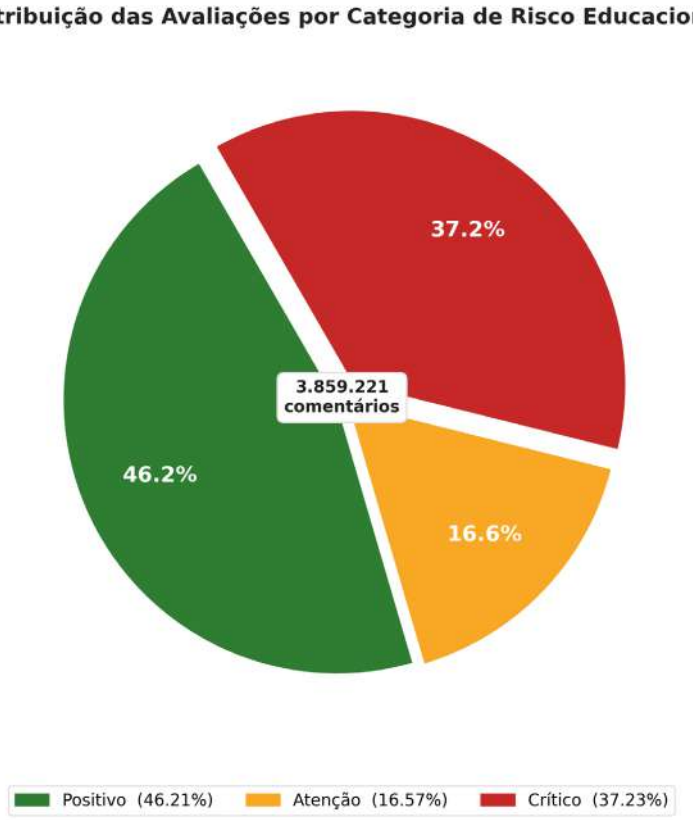
Para a classificação automática dos comentários, utilizou-se o modelo Gemini via técnica de *Few-Shot Prompting*. Essa técnica consiste em fornecer ao modelo exemplos rotulados diretamente no *prompt*, orientando a classificação sem a necessidade de *fine-tuning*, o que reduz custos computacionais e tempo de desenvolvimento. A escolha por uma LLM justifica-se pela sua superioridade em processar linguagem natural complexa, identificar

nuances semânticas e interpretar *feedbacks* qualitativos que modelos tradicionais poderiam classificar erroneamente devido a erros gramaticais ou informalidades. O modelo foi aplicado com o objetivo de classificar cada comentário em três categorias:

- **Positivo:** comentários que expressam satisfação ou elogios explícitos.
- **Atenção:** comentários neutros, ambíguos ou com críticas leves.
- **Crítico:** comentários que indicam insatisfação relevante, falhas recorrentes ou riscos à experiência educacional.

A distribuição resultante, apresentada no Gráfico 2, evidenciou que 46,21% dos comentários foram classificados como positivos, 16,57% como atenção e 37,23% como críticos. Esse último percentual é particularmente relevante, pois contrasta com a nota média de 9,22 obtida pela análise quantitativa, demonstrando empiricamente a existência do efeito teto e a necessidade de abordagens complementares baseadas em texto para uma visão diagnóstica completa da qualidade dos cursos.

**Gráfico 2 – Distribuição das Avaliações por Categoria de Risco Educacional**




*Fonte: Elaboração própria com base nos dados das avaliações de reação da Enap (2024).*

## *3.2 Diagnóstico Qualitativo*

A análise qualitativa das avaliações foi segmentada por critério avaliado, possibilitando identificar onde residem os principais pontos de insatisfação percebidos pelos participantes. Essa segmentação revelou padrões recorrentes de críticas, distribuídos de forma consistente entre diferentes cursos e temáticas. Observou-se a concentração de comentários críticos associados aos seguintes aspectos:

- **Conteúdo e estrutura:** frequentemente relacionados a material desatualizado, excesso de abordagem teórica ou desalinhamento entre os objetivos propostos e o conteúdo apresentado.
- **Interface gráfica:** envolvendo dificuldades de navegação, falhas no salvamento de progresso, inconsistências no controle de conclusão dos módulos e problemas de usabilidade.
- **Aplicabilidade prática:** especialmente em cursos com alta densidade conceitual e escassez de exemplos contextualizados à realidade do serviço público.
- **Apoio logístico:** incluindo ausência de canais claros de suporte, demora no atendimento ou insuficiência de materiais complementares.

Além da classificação de sentimento, foi realizada uma análise semântica qualitativa dos comentários classificados como críticos, com o objetivo de identificar temas recorrentes, padrões discursivos e exemplos representativos de insatisfação. Essa análise revelou que muitos comentários apresentam elevado grau de detalhamento técnico, frequentemente acompanhados de sugestões concretas de melhoria, relatos de falhas específicas e comparações com práticas adotadas no mercado ou em outras instituições.

## 3.3 Síntese Integrada dos Resultados: Mapeamento do Risco Educacional

Este capítulo apresenta uma síntese integrada dos resultados quantitativos e qualitativos obtidos ao longo do estudo, com o objetivo de consolidar as evidências produzidas e traduzi-las em um instrumento analítico de apoio à decisão institucional. O mapeamento do risco educacional adota uma abordagem de métodos mistos, integrando métricas quantitativas e evidências qualitativas extraídas de dados textuais (Creswell, 2014; Creswell & Clark, 2013).

## 3.4 Conceituação do Risco Educacional

Neste trabalho, o risco educacional é definido como a probabilidade de comprometimento da experiência formativa percebida pelos participantes, resultante da combinação entre níveis elevados de insatisfação, recorrência de manifestações críticas e impacto potencial sobre os objetivos educacionais da instituição. Trata-se, portanto, de um construto analítico que se forma a partir da interação entre indicadores quantitativos e padrões qualitativos observados nos comentários textuais.

Esse conceito alinha-se à perspectiva de Kirkpatrick e Kirkpatrick (2006) sobre os quatro níveis de avaliação de treinamento, situando-se primordialmente no Nível 1 (Reação), com potencial de extensão ao Nível 2 (Aprendizagem) à medida que dados de desempenho acadêmico sejam integrados ao modelo.

## 3.5 Dimensões Analíticas do Risco Educacional

As dimensões do risco foram definidas a partir de padrões identificados na análise qualitativa, especialmente por meio do agrupamento semântico e das sínteses interpretativas geradas pelo modelo de linguagem Gemini (Google DeepMind, 2023). Entre as principais dimensões identificadas destacam-se:

- **Risco pedagógico:** relacionado à percepção de inadequação do conteúdo, metodologia ou estrutura didática.
- **Risco tecnológico:** associado a dificuldades de acesso, usabilidade ou estabilidade da plataforma.
- **Risco organizacional:** vinculado a falhas de comunicação, cronograma ou suporte institucional.
- **Risco de engajamento:** expresso por desmotivação, evasão ou baixa percepção de valor formativo.

Essas dimensões não são mutuamente exclusivas e podem coexistir em um mesmo curso, refletindo a complexidade da experiência educacional. A Tabela 3 apresenta os cursos com maior número de ocorrências por dimensão de risco.

**Tabela 3 – Cursos com maior número de ocorrências por dimensão do risco educacional**

| Curso | Dimensão do risco | Ocorrências |
|---|---|---|
| Boas Práticas de Manipulação em Serviços de Alimentação | Pedagógica | 43.727 |
| Linguagem simples aproxima o governo das pessoas | Outros | 37.937 |
| Sistema Eletrônico de Informações – SEI! USAR | Pedagógica | 27.725 |
| Boas Práticas de Manipulação em Serviços de Alimentação | Tecnológica | 20.475 |
| Boas Práticas de Manipulação em Serviços de Alimentação | Organizacional | 19.483 |
| Introdução à Libras | Pedagógica | 19.327 |
| Linguagem simples aproxima o governo das pessoas | Pedagógica | 18.779 |
| SEI! USAR 4.0 | Pedagógica | 18.295 |
| Comunicação não violenta | Pedagógica | 14.862 |
| Sistema Eletrônico de Informações – SEI! USAR | Tecnológica | 12.802 |

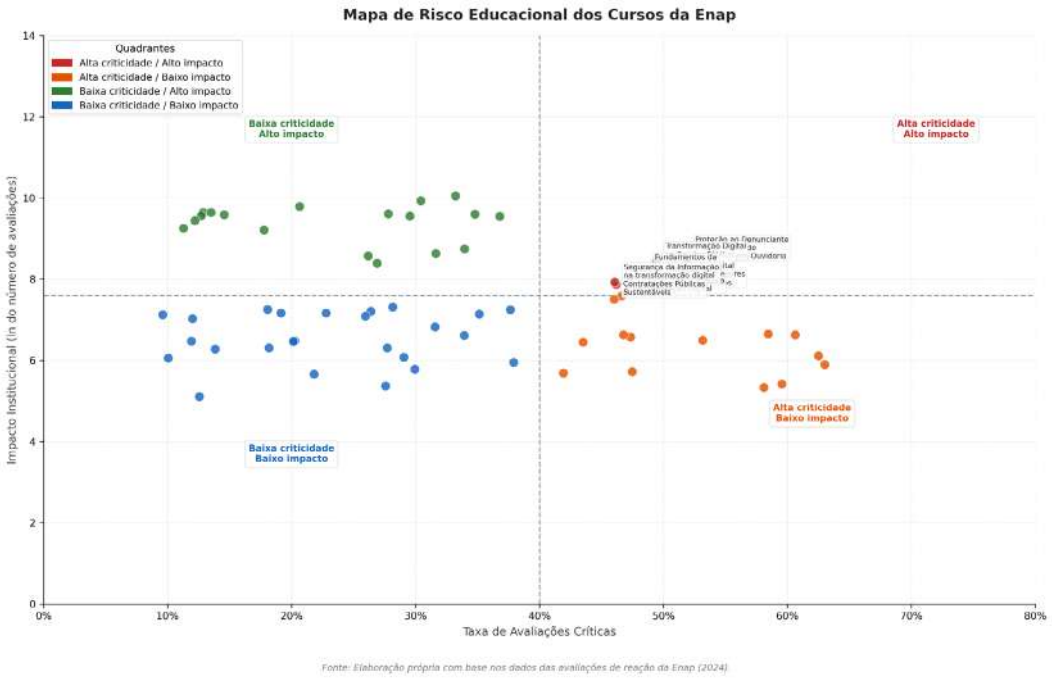
**Fonte:** Elaboração própria com base nos dados das avaliações de reação da Enap (2024).

## 4. INTEGRAÇÃO COM O ÍNDICE DE URGÊNCIA

O mapeamento do risco educacional está diretamente associado ao Índice de Urgência apresentado anteriormente, que sintetiza informações quantitativas relativas à polaridade das avaliações, à frequência de manifestações críticas e ao impacto do curso no conjunto da oferta formativa. Os escores do índice foram utilizados como eixo estruturante para a classificação dos cursos em faixas de risco, permitindo uma leitura comparativa e escalonada.

A integração entre o Índice de Urgência e as dimensões qualitativas do risco possibilita identificar não apenas quais cursos demandam atenção prioritária, mas também quais fatores contribuem para sua classificação em faixas críticas. Assim, o mapa de risco funciona como uma camada interpretativa adicional, que contextualiza os resultados quantitativos à luz das evidências textuais analisadas. O Gráfico 3 apresenta o mapa de risco educacional resultante, com os cursos distribuídos em quatro quadrantes segundo a taxa de avaliações críticas e o impacto institucional.

## Gráfico 3 – Mapa de Risco Educacional dos Cursos da Enap




Fonte: Elaboração própria com base nos dados das avaliações de reação da Enap (2024).

**Fonte:** Elaboração própria com base nos dados das avaliações de reação da Enap (2024).

## *4.1 Evidências Qualitativas Associadas às Faixas de Risco*

Para cada faixa de risco identificada, foram associadas evidências qualitativas representativas, extraídas dos comentários dos participantes. Cursos classificados como de alto risco apresentaram maior recorrência de tópicos relacionados a problemas técnicos persistentes, dificuldades de acompanhamento do conteúdo e insatisfação com o suporte institucional. A Tabela 4 ilustra exemplos representativos dessa evidência qualitativa.

### Tabela 4 – Exemplos de comentários qualitativos em cursos com maior incidência de risco

| Curso | Comentário do participante |
|---|---|
| Sistema Eletrônico de Informações – SEI! USAR | Demora na atualização e, em alguns momentos, o sistema não reconhece adequadamente as funcionalidades. |
| Sistema Eletrônico de Informações – SEI! USAR | Navegação muito ruim. |
| Boas Práticas de Manipulação em Serviços de Alimentação | Imagens pouco nítidas. |
| Boas Práticas de Manipulação em Serviços de Alimentação | Conteúdo muito superficial. |
| Linguagem simples aproxima o governo das pessoas | Conteúdo desatualizado. |
| Linguagem simples aproxima o governo das pessoas | Em muitas situações, a falta de entendimento dificulta a aplicação prática dos conceitos apresentados. |
| Introdução à Libras | Curso desorganizado. |
| SEI! USAR 4.0 | Não foi possível observar na prática como os conteúdos apresentados funcionam no contexto de trabalho. |

**Fonte:** Elaboração própria com base nos dados das avaliações de reação da Enap (2024).

# 5. TRABALHOS FUTUROS

Os resultados obtidos neste trabalho abrem um conjunto de perspectivas de desenvolvimento que podem ampliar significativamente a capacidade diagnóstica e preditiva da abordagem proposta. As principais direções de investigação futura são apresentadas a seguir.

## 5.1 Refinamento Tecnológico e Explicabilidade

Como desdobramento técnico, sugere-se a adaptação ou o ajuste fino (*fine-tuning*) de modelos de linguagem especificamente ao vocabulário e contexto pedagógico da Enap. Esse refinamento permitiria ao modelo capturar nuances linguísticas específicas do serviço público federal, aumentando a precisão das classificações. A incorporação de ferramentas de explicabilidade, como SHAP (SHapley Additive exPlanations) ou LIME, é fundamental para ampliar a transparência das decisões algorítmicas, permitindo que os gestores compreendam exatamente quais termos ou variáveis impulsionaram a classificação de risco de um curso. Essa transparência é especialmente relevante no contexto da administração pública, em que decisões baseadas em algoritmos devem ser auditáveis e justificáveis.

## 5.2 Integração de Indicadores e Monitoramento de Impacto

Para elevar o potencial preditivo da abordagem, propõe-se a integração do Mapeamento de Risco Educacional com indicadores de desempenho acadêmico, taxas de evasão e permanência. Essa visão holística permitiria intervenções ainda mais proativas e direcionadas, alinhando o modelo ao Nível 2 do Modelo de Kirkpatrick. Além disso, torna-se essencial a criação de um painel histórico de indicadores, que permita o acompanhamento temporal das métricas trabalhadas, viabilizando a análise de tendências e a avaliação do impacto das intervenções realizadas.

## 5.3 Interface e Democratização do Acesso

No âmbito da entrega de valor aos envolvidos, a criação de modelos de visualização intuitivos é prioritária para facilitar o consumo desses dados complexos pela coordenação e direção. Complementarmente, sugere-se o desenvolvimento de uma ferramenta de *feedback* estruturado para os professores, que forneça não apenas a média de desempenho, mas também os dados qualitativos e a comparação anônima em relação ao restante da instituição, fechando o ciclo de melhoria contínua entre a gestão, a sala de aula e a inteligência de dados.

## 5.4 Implementação de Alertas Automatizados

Sugere-se o desenvolvimento de um sistema de gatilhos (*triggers*) que envie notificações automáticas à coordenação sempre que um curso atingir um determinado limite de criticidade no Índice de Urgência. Essa funcionalidade transformaria o *dashboard* de uma ferramenta de consulta em um sistema de monitoramento ativo, garantindo que falhas graves sejam tratadas em tempo real e não apenas detectadas em revisões periódicas.

# CONCLUSÃO

O presente trabalho buscou solucionar um dos principais gargalos na gestão acadêmica da Enap: a subutilização da inteligência qualitativa contida nas avaliações de cursos devido à impossibilidade de análise humana em larga escala. Conforme demonstrado, a dependência exclusiva de métricas numéricas tradicionais, como o NPS, cria um efeito teto que mascara falhas estruturais e pedagógicas, uma vez que a nota média elevada de 9,22 não reflete as críticas presentes em 37% das avaliações textuais.

A aplicação de Ciência de Dados e Inteligência Artificial provou ser um caminho viável e eficaz para transformar esse "ouro qualitativo" em ativos de gestão. Por meio do desenvolvimento de um pipeline de análise automatizado — orquestrado pelo Apache Airflow e sustentado pela capacidade semântica do modelo Gemini —, foi possível converter comentários brutos em indicadores de risco e urgência acionáveis. A validação por coeficiente Kappa de Cohen ($\kappa = 0,81$) conferiu robustez metodológica às classificações produzidas.

O IU proposto representa uma contribuição metodológica relevante, ao combinar taxa de criticidade e volume de participantes em uma métrica composta que evita distorções de escala. O Mapa de Risco Educacional resultante oferece à Alta Direção e à Coordenação de Cursos um instrumento robusto de priorização de intervenções, permitindo que a tomada de decisão seja fundamentada em evidências sólidas e não apenas em percepções parciais.

Além do suporte à gestão macro, o estudo destacou a importância de democratizar o acesso a esses dados para o corpo docente. Ao propor uma interface de *feedback* contínuo e comparativo, o trabalho aponta para uma evolução na cultura pedagógica, na qual o professor assume um papel ativo no ciclo de melhoria contínua da instituição.

Os próximos passos, que incluem o ajuste fino de modelos de linguagem, a integração com indicadores de evasão e a implementação de alertas automatizados, prometem elevar o Mapeamento de Risco Educacional a um patamar preditivo ainda mais sofisticado. Com isso, a Enap caminha para uma gestão educacional *data-driven*, capaz de antecipar problemas e garantir a excelência acadêmica que sua missão institucional exige.

## Disponibilidade de Dados e Código

Em conformidade com as diretrizes de Open Science estabelecidas para este trabalho, os dados e o código-fonte estão disponibilizados nos repositórios indicados a seguir.

## Dados

Os dados utilizados nas análises são de caráter institucional e contêm informações sensíveis sobre avaliações de servidores públicos. Em conformidade com a Lei Geral de Proteção de Dados (LGPD — Lei nº 13.709/2018) e com as políticas de segurança da informação da Enap, os dados brutos não podem ser disponibilizados publicamente em sua forma original.

Para garantir a reprodutibilidade e a transparência científica, foi gerado um conjunto de dados sintéticos que preserva as características estatísticas e distribucionais da base original, sem expor informações pessoais ou institucionalmente sensíveis. Esse conjunto sintético, acompanhado de documentação metodológica sobre o processo de geração, está disponível em: *https://doi.org/10.5281/zenodo.20701504*

## Código-Fonte

O código-fonte completo da solução, que inclui os *scripts* Python para pré-processamento, filtro de densidade cognitiva, integração com a API Gemini, cálculo do Índice de Urgência e geração do Mapa de Risco Educacional, está disponível no repositório público: *https://github.com/franciscomduarte/enap-nlp-educational-risk*

O repositório contém um arquivo README.md com instruções detalhadas para instalação de dependências, configuração do ambiente, execução do pipeline de análise e reprodução dos resultados apresentados neste artigo. As variáveis de ambiente necessárias (como chaves de API) estão documentadas sem exposição de valores sensíveis, com uso de arquivo .env.example como referência.

## Declaração de Uso de Inteligência Artificial Generativa

Em conformidade com as diretrizes institucionais para trabalhos acadêmicos com uso de Inteligência Artificial, declara-se que ferramentas de IA Generativa foram utilizadas nas seguintes etapas deste trabalho:

- **Classificação semântica dos comentários textuais:** utilização do modelo Gemini (Google DeepMind, 2023), via API, para classificação de polaridade (positivo, atenção, crítico) e identificação de dimensões de risco educacional. A técnica de Few-Shot Prompting foi empregada para garantir consistência e assertividade das classificações.
- **Geração de resumos interpretativos:** o modelo Gemini foi utilizado para síntese semântica de grupos de comentários de alta relevância, possibilitando a identificação de temas recorrentes e causas-raiz de insatisfação.
- **Agrupamento semântico por similaridade de intenção:** aplicado sobre os 74.285 comentários filtrados, com supervisão humana para validação dos agrupamentos resultantes.

Todas as classificações automatizadas foram validadas por meio do coeficiente Kappa de Cohen ($\kappa = 0{,}81$), com amostra de revisão humana independente. Os dados e o código-fonte estão disponíveis conforme indicado na seção “Disponibilidade de Dados e Código”.

## REFERÊNCIAS


ABBAD, G.; BORGES-ANDRADE, J. E.; GAMA, A. L. G. Treinamento: análise do relacionamento da avaliação nos níveis de reação, aprendizagem e impacto no trabalho. **Revista de Administração Contemporânea**, v. 4, n. 3, p. 25-45, 2000. DOI: 10.1590/S1415-65552000000300003.

ABBAD, G. da S.; BORGES-ANDRADE, J. E.; FREITAS, T. R. de. Avaliação de aprendizagem, reação e impacto de treinamentos corporativos no trabalho. **Psicologia: Teoria e Pesquisa**, v. 35, e35511, 2019. DOI: 10.1590/0102.3772e35511.

APACHE SOFTWARE FOUNDATION. **Apache Airflow Documentation**. 2024. Disponível em: https://airflow.apache.org/. Acesso em: abr. 2026.

BOMMASANI, R. et al. On the opportunities and risks of foundation models. **arXiv preprint arXiv**:2108.07258, 2021. Disponível em: https://arxiv.org/abs/2108.07258. Acesso em: abr. 2026.

CRESWELL, J. W. **Research design**: qualitative, quantitative, and mixed methods approaches. 4. ed. Thousand Oaks: SAGE Publications, 2014.

CRESWELL, J. W.; CLARK, V. L. **Pesquisa de métodos mistos**. Porto Alegre: Penso, 2013.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: pre-training of deep bidirectional transformers for language understanding. In: **Conference of the north american chapter of the association for computational linguistics** (NAACL), 2019, Minneapolis. Proceedings [...]. Minneapolis: ACL, 2019.

GOOGLE DEEPMIND. **Gemini: a family of highly capable multimodal models**. 2023. Disponível em: https://arxiv.org/abs/2312.11805. Acesso em: abr. 2026.

HAN, J.; KAMBER, M.; PEI, J. **Data mining:** concepts and techniques. 3. ed. Waltham: Morgan Kaufmann, 2011.

HARENSLAK, B.; RUITER, J. de. **Data pipelines with Apache Airflow**. Shelter Island: Manning Publications, 2021.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing**. 3. ed. Upper Saddle River: Pearson, 2023.

KIRKPATRICK, D. L.; KIRKPATRICK, J. D. **Evaluating training programs**: the four levels. 3. ed. San Francisco: Berrett-Koehler, 2006.

LIU, B. **Sentiment analysis and opinion mining**. San Rafael: Morgan & Claypool, 2012.

MICROSOFT. **Microsoft Power BI Documentation**. 2024. Disponível em: https://learn.microsoft.com/power-bi/. Acesso em: abr. 2026.

NUNNALLY, J. C.; BERNSTEIN, I. H. **Psychometric theory**. 3. ed. New York: McGraw-Hill, 1994.

POSTGRESQL GLOBAL DEVELOPMENT GROUP. **PostgreSQL Documentation**. 2024. Disponível em: https://www.postgresql.org/docs/. Acesso em: abr. 2026.

REIS, J.; HOUSLEY, M. **Fundamentals of data engineering**: plan and build robust data systems. Sebastopol: O'Reilly Media, 2022.

VASWANI, A. et al. Attention is all you need. **Advances in Neural Information Processing Systems**, v. 30, 2017.

DUARTE JÚNIOR, Francisco Carlos Molina. **Dados sintéticos para análise de avaliações educacionais em larga escala** — Enap/EVG. Zenodo, 2026. DOI: 10.5281/zenodo.20701504. Disponível em: https://doi.org/10.5281/zenodo.20701504. Acesso em: jun. 2026

# *Benchmark* jurídico e arquitetura documental orientada por evidência em casos brasileiros da corte IDH: do conhecimento interno à triagem factual orientada por precedentes

*Luiz Víctor do Espírito Santo Silva*
*Prof. MSc. Clayton Rodrigues da Silva*

## Resumo

Decisões da Corte Interamericana de Direitos Humanos (Corte IDH) são documentos longos e tecnicamente densos, em que tarefas aparentemente simples, como identificar direitos declarados violados e sintetizar medidas de reparação e seu *status* de cumprimento, exigem rastreabilidade, completude e hierarquia de fonte. Este trabalho investiga a aplicação de modelos de linguagem em tarefas jurídicas objetivas no domínio de casos brasileiros julgados pela Corte IDH. O estudo constrói um *benchmark* em português com 55 itens e três tarefas estruturadas, auditadas integralmente, e compara quatro regimes de acesso à informação: duas condições sem consulta a fontes externas, uma condição com busca genérica e uma condição com arquitetura documental orientada por evidência, apoiada em *corpus* local controlado e ferramentas especializadas. Os resultados mostram que parte relevante do desempenho nessas condições sem consulta a fontes externas é mediada pela calibração do *prompt*, mas que o ganho mais consistente surge quando o problema é tratado como arquitetura de sistema. A partir dessa evidência, o trabalho abre uma etapa aplicada de triagem factual orientada por precedentes brasileiros da Corte IDH. Conclui-se que, em domínios jurídicos documentais, confiabilidade depende sobretudo de desenho de sistema com verificabilidade.

**Palavras-chave:** *benchmark* jurídico; Corte Interamericana; modelos de linguagem; arquitetura documental orientada por evidência; triagem factual.

## INTRODUÇÃO

A aplicação de modelos de linguagem de grande porte, Large Language Models (LLMs), a tarefas jurídicas cresceu rapidamente nos últimos anos, tanto em cenários de consulta e síntese quanto em tarefas de pergunta e resposta, extração e apoio à decisão. Entretanto, quando o objeto envolve decisões extensas, linguagem técnico-jurídica e exigências fortes de rastreabilidade, a utilidade prática desses modelos passa a depender menos de "um modelo melhor" e mais de desenho de sistema: acesso controlado à informação, mecanismos de verificação, contratos de saída e critérios de auditabilidade.

Este estudo organiza-se em torno de um recorte específico: casos contenciosos envolvendo o Brasil julgados pela Corte Interamericana de Direitos Humanos (Corte IDH). O trabalho não busca construir um detector geral de violações de direitos humanos nem um sistema universal de análise jurídica. Seu objetivo é mais delimitado e, justamente por isso, mais

auditável: estudar como LLMs se comportam ao responder tarefas objetivas sobre decisões da Corte IDH e como esse comportamento muda quando se altera o regime de acesso à informação, passando de condições *closed-book* para condições com busca genérica e, finalmente, para uma arquitetura documental orientada por evidência.

Nesta pesquisa, usa-se *benchmark* no sentido de um conjunto estruturado de tarefas, itens, métricas e critérios de interpretação, voltado a medir capacidades específicas de modelos, e não como medida absoluta de competência jurídica (Guha et al., 2023; Reuel et al., 2024; Bean et al., 2025). Chama-se *closed-book* a condição em que o modelo responde sem consultar fontes externas durante a inferência. A condição com busca genérica, por sua vez, aproxima o experimento do uso comum de um modelo com *browsing/search* genérico. Já a arquitetura documental orientada por evidência designa, neste trabalho, um arranjo em que a resposta é organizada a partir de *corpus* controlado, navegação orientada por evidência e citação verificável, em diálogo com as literaturas de *document-grounded generation* e *legal RAG*, além de trabalhos que enfatizam verificabilidade e fluxos jurídicos *end-to-end* (Prabhumoye et al., 2021; Pipitone e Houir Alami, 2024; Hou et al., 2024; Stern et al., 2024; Hu et al., 2026; Koenecke et al., 2025).

O problema de pesquisa consiste em compreender em que medida o desempenho e a confiabilidade de saídas de LLMs, em um domínio jurídico específico, dependem de: (i) conhecimento interno do modelo; (ii) calibração do *prompt* em condição *closed-book*; (iii) permissão de busca genérica semelhante ao uso de um usuário comum; e (iv) adoção de uma arquitetura documental orientada por evidência, com *corpus* local controlado, *tools* de navegação e saída estruturada.

Em termos mais diretos, a pergunta que orienta o trabalho é a seguinte: no domínio da Corte IDH, especialmente no eixo de reparações e cumprimento relacionados aos casos brasileiros, o ganho decisivo vem mais de arquitetura de sistema do que de um modelo isolado?

O problema empírico estudado não surge em abstrato. Ele se insere em quadro amplo de baixa permeabilidade da jurisprudência interamericana na jurisdição interna brasileira. O relatório *Comportamento judicial em relação à Convenção Americana sobre Direitos Humanos* (CADH), produzido pelo Conselho Nacional de Justiça - CNJ em parceria com a Pontifícia Universidade Católica do Paraná (PUCPR), identificou déficit relevante de conhecimento e aplicação da CADH, da jurisprudência da Corte IDH e do controle de convencionalidade pelo Poder Judiciário brasileiro (Conselho Nacional de Justiça, 2023). No mesmo estudo, 79,66% dos respondentes afirmaram não conhecer, não ter estudado a fundo ou nunca ter aplicado a jurisprudência da Corte IDH em suas decisões, 79,81% responderam com a mesma afirmação em relação ao controle de convencionalidade, e 84,91% indicaram não considerar a jurisprudência da Corte vinculante ou obrigatória para magistrados brasileiros (Conselho Nacional de Justiça, 2023).

O trabalho foi orientado por quatro hipóteses principais. A primeira é que tarefas dependentes de granularidade documental e completude tendem a sofrer em condição *closed-book*. A segunda é que parte do desempenho observado em *closed-book* é mediada pela formulação do *prompt*, e que *prompts* excessivamente anti-invenção podem induzir abstenção improdutiva. A terceira é que busca genérica melhora a utilidade prática, mas preserva erros relevantes de rastreabilidade e de hierarquia de fonte. A quarta é que uma condição com arquitetura documental orientada por evidência, apoiada em *corpus* controlado e ferramentas especializadas, reduz esses erros dominantes e eleva o desempenho auditado, aproximando o *output* de exigências de confiabilidade operacional.

Do ponto de vista institucional, decisões da Corte IDH são documentos longos, densos e estruturados, nos quais tarefas aparentemente simples (por exemplo, identificar os argumentos que sustentam a outorga de determinada medida de reparação) exigem hierarquia de fontes, completude e leitura contextual. Do ponto de vista técnico, a adoção crescente de LLMs em ambientes jurídicos torna insuficiente qualquer avaliação baseada apenas em exemplos isolados, impressões subjetivas ou respostas plausíveis, porém não verificáveis.

## 1. TRABALHOS RELACIONADOS

O argumento central do trabalho, para além de comparar modelos, reside em mostrar como diferentes regimes de acesso à informação alteram o comportamento de sistemas baseados em modelos de linguagem em um domínio jurídico de alto impacto. Por isso, a revisão de literatura é organizada em cinco eixos conectados ao desenho do estudo: (i) *benchmarks* como problema de validade e como artefato sociotécnico; (ii) *benchmarks* jurídicos e o papel de especialistas de domínio; (iii) validade ecológica, documentos longos e multilinguismo; (iv) confiabilidade, *grounding* e dimensões de *trustworthiness*; e (v) deslocamento de "modelo" para "sistema", incluindo decomposição de tarefas, arquiteturas verificáveis e fluxos de agentes.

### 1.1 Benchmarks como problema de validade e artefato sociotécnico

Uma parte importante da literatura recente sustenta que *benchmarks* de IA não podem ser tratados como instrumentos neutros, autoevidentes e suficientes para fundamentar afirmações sobre capacidade. Eriksson et al. (2025) argumentam que *benchmarks* são moldados por incentivos competitivos, escolhas culturais e interesses comerciais e regulatórios; por isso, não apenas medem sistemas, mas também ajudam a moldar quais capacidades passam a ser perseguidas e valorizadas. Essa leitura é particularmente relevante para um domínio como o jurídico, em que tarefas aparentemente simples podem ocultar exigências fortes de rastreabilidade, hierarquia de fontes e completude.

Bean et al. (2025) reforçam essa crítica ao colocar a validade construtiva no centro da discussão metodológica. Para os autores, um *score* só é informativo na medida em que sustenta uma pretensão inferencial clara sobre o fenômeno que se pretende medir. Em tarefas jurídicas, esse ponto é decisivo: "acertar" pode significar correção factual, correção jurídica, completude, aderência jurisdicional, confiabilidade da fonte ou capacidade de explicitar incerteza residual. Sem essa distinção, há risco de transformar métricas fáceis de pontuar em medidas inadequadas do problema real.

Essa preocupação é aprofundada por Reuel et al. (2024), que tratam *benchmarks* como artefatos de ciclo de vida completo, dependentes de desenho, documentação, implementação, manutenção e clareza de propósito. O argumento é convergente com este trabalho: o *benchmark*» não é o bastante; é necessário justificar por que as tarefas escolhidas são adequadas ao domínio, documentar as limitações do artefato e criar condições de replicabilidade e auditoria. Essa literatura sustenta diretamente escolhas metodológicas adotadas aqui, como *logs* completos de execução, *scoring* offline reproduzível, auditoria manual integral e comparação entre condições experimentais, e não apenas entre modelos.

## 1.2 Benchmarks jurídicos e participação de especialistas de domínio

No domínio jurídico, um problema recorrente é a distância entre tarefas tecnicamente convenientes e tarefas reconhecidas por juristas como juridicamente relevantes. Guha et al. (2023) respondem a esse problema propondo o *Legalbench*, um *benchmark* construído colaborativamente com participação substantiva de profissionais do Direito e organizado em tipologias de raciocínio jurídico. O ponto central não é apenas o tamanho do *benchmark*, mas o fato de que as tarefas foram pensadas para medir capacidades consideradas úteis ou interessantes por especialistas do domínio.

Esse trabalho é especialmente importante por duas razões. Primeiro, ele ajuda a legitimar a ideia de que *benchmarks* jurídicos fortes precisam de curadoria substantiva, e não apenas de coleta automatizada de perguntas. Segundo, ele mostra que a relação entre comunidade técnica e comunidade jurídica melhora quando as tarefas são descritas em um vocabulário reconhecível pelos juristas. No caso deste projeto, isso aparece no esforço de formular as tarefas do nosso *benchmark* em torno de direitos violados, reparações e *status* de cumprimento, e não em torno de categorias abstratas de processamento de linguagem natural (NLP) sem correspondência clara com o objeto jurídico.

Ao mesmo tempo, as limitações reconhecidas pelos próprios autores de Guha et al. (2023) são instrutivas. O *LegalBench* é fortemente marcado pelo contexto anglófono e pelo direito norte-americano, além de cobrir de forma limitada documentos longos, multilinguismo e raciocínio ancorado em precedentes de outros sistemas. Essas limitações reforçam a opção deste trabalho por um *benchmark* menor, mas altamente delimitado: decisões da Corte IDH em casos brasileiros, com tarefas auditáveis e acopladas a um *corpus* controlado.

## 1.3 Validade ecológica: documentos longos, multilinguismo e fluxos de trabalho end-to-end

No estudo, a noção de validade ecológica cumpre uma função argumentativa específica: ligar a discussão de validade construtiva, apresentada nas seções anteriores, às escolhas de desenho experimental do nosso caso. Em outros termos, não basta perguntar se a tarefa mede algo de interesse; é preciso perguntar se ela mede esse algo em condições próximas às de uso real. Essa inflexão, hoje central na avaliação de LLMs, desloca o foco de cenários artificiais para fluxos concretos de aplicação (Li et al., 2025).

Quando esse critério é aplicado ao campo da *computational law*, a consequência é metodologicamente direta: língua, jurisdição e tipo documental deixam de ser contexto e passam a ser parte do próprio objeto avaliado. Küçük e Can (2025) mostram que a IA jurídica opera em ecossistemas sociotécnicos compostos por *datasets*, *benchmarks*, ontologias e infraestrutura semântica; por isso, diferenças entre sistemas jurídicos e comunidades linguísticas impõem limites fortes a generalizações amplas.

A literatura de NLP jurídico reforça esse ponto ao mostrar que documentos longos, linguagem técnica e fragmentação de dados não são ruídos periféricos, mas características estruturais do domínio, Ariai et al. (2025). Daí decorre uma implicação central para este estudo: condições sem consulta a fontes externas podem funcionar como linha de base útil, mas tendem a subestimar o desempenho juridicamente relevante, pois, na prática, a tarefa inclui localizar, conferir e citar evidências.

Esse argumento abre a transição para a literatura de *benchmarks* jurídicos ecológicos. Stern et al. (2024) defendem avaliações *end-to-end* que integrem documentos longos, multilinguismo, *retrieval*, citação e geração. O ganho analítico dessa proposta é tornar visível não apenas a resposta final, mas o percurso que a produz, permitindo avaliar a qualidade das etapas intermediárias.

A arquitetura testada neste estudo dialoga com a família RAG e com abordagens de *document-grounded generation* e *legal RAG*, mas adota um desenho mais restritivo e auditável: *corpus* controlado, *tools* especializadas, navegação por janelas de parágrafo, consultas estruturadas e exigência de citação verificável. Assim, a recuperação não é tratada como adição opcional de contexto, e sim como componente de uma arquitetura documental orientada por evidência, na qual *grounding* e verificabilidade integram o próprio problema metodológico (Prabhumoye et al., 2021; Pipitone e Houir Alami, 2024; Hou et al., 2024; Stern et al., 2024; Hu et al., 2026; Koenecke et al., 2025; Küçük e Can, 2025).

## 1.4 Confiabilidade, grounding e trustworthiness

Quando se observa a literatura mais recente sobre avaliação de LLMs em aplicações jurídicas, torna-se claro que a discussão não se reduz à acurácia superficial. Hu et al. (2026) organizam a análise em dimensões como qualidade do raciocínio jurídico, aderência jurisdicional, *grounding*, consistência e *trustworthiness*. Neste texto, *grounding* designa a ancoragem verificável da resposta em fonte recuperável, enquanto *trustworthiness* remete à confiabilidade global do sistema, incluindo consistência, aderência à fonte e segurança de uso. Em cenários reais, uma resposta plausível, mas mal ancorada em fonte, pode ser mais perigosa do que uma resposta incompleta (Hu et al., 2026).

Outro eixo relevante é o de justiça e sensibilidade normativa. Chalkidis et al. (2022) mostram que disparidades de desempenho em NLP jurídico podem atingir grupos diferentes de forma assimétrica, o que torna inadequado tratar *fairness* (entendida aqui como equidade ou ausência de disparidades sistemáticas de desempenho entre grupos) como preocupação acessória. Em paralelo, Samway et al. (2026) indicam que o comportamento de LLMs frente a princípios de direitos humanos varia por língua, tipo de direito e *framing* do *prompt*, isto é, pelo enquadramento textual da instrução dada ao modelo.

Essa mesma direção aparece em Koenecke et al. (2025), que deslocam a discussão de IA jurídica da centralidade atribuída aos modelos para gargalos concretos de *pipeline*: curadoria de dados, anotação especializada e verificação de saídas. O argumento é decisivo para este trabalho porque o *benchmark*, os *logs*, a auditoria manual e o *corpus* controlado passam a ser lidos como partes do mesmo ecossistema metodológico.

## 1.5 Decomposição de tarefas, multiagentes e raciocínio verificável

A literatura recente investiga como a decomposição de tarefas, a especialização por papéis e as arquiteturas verificáveis podem melhorar o desempenho de sistemas jurídicos baseados em LLMs. Jing et al. (2025) propõem o *MASLegalBench*, um *benchmark* especificamente voltado a sistemas multiagentes em raciocínio jurídico, estruturando o fluxo em etapas como fatos, regras, aplicação e inferências de senso comum. Essa literatura é útil por mostrar que certos problemas jurídicos deixam de ser bem descritos quando tratados como uma única chamada de geração.

Sadowski e Chudziak (2025) avançam nessa direção ao defender arquiteturas com representações intermediárias inspecionáveis, combinando linguagem natural com formalizações e regras verificáveis. O ponto mais importante para este trabalho não é reproduzir integralmente uma abordagem neurossimbólica, como o foi no caso da contribuição desses autores, mas reconhecer que, em domínios de alto impacto, é útil criar pontos de inspeção entre entrada, recuperação de fonte, inferência e saída.

Há ainda uma ponte útil com a literatura de raciocínio argumentativo. Chlapanis et al. (2024) e Habernal et al. (2023) sugerem que decisões de cortes de direitos humanos podem ser analisadas de forma mais rica quando o foco recai sobre cadeias argumentativas e tipologias de argumento, e não apenas sobre classificações rasas. Já Santosh et al. (2023) reforçam que certas tarefas jurídicas são naturalmente dependentes entre si, o que ajuda a justificar a decomposição do *benchmark* em diferentes tarefas. De outra parte, Kennedy et al. (2026) propõem o HumRightsBench, partindo da tradição IRAC e a adaptam para um esquema IRAP, composto por *Issue Identification*, *Rule Recall*, *Rule Application* e *Proposed Remedies*. No presente trabalho, essa referência é usada apenas como inspiração de decomposição analítica para a fase de triagem factual, sem importar integralmente a taxonomia nem o protocolo do *benchmark* original.

### *1.6 Síntese crítica e lacuna ocupada pelo estudo*

Tomada em conjunto, essa literatura sustenta quatro teses que estruturam o presente estudo. Primeiro, *benchmarks* jurídicos precisam ser lidos com atenção à validade construtiva, à validade ecológica e à influência do próprio desenho sobre o comportamento avaliado. Segundo, o domínio jurídico exige participação substantiva de especialistas, delimitação jurisdicional explícita e cautela com generalizações multilíngues ou multijurisdicionais. Terceiro, *trustworthiness* depende de *grounding*, verificabilidade e capacidade de separar evidência de hipótese. Quarto, ganhos robustos tendem a aparecer quando o problema é tratado como sistema, e não apenas como geração textual pontual.

Ao construir um *benchmark* jurídico especializado e compará-lo sob diferentes regimes de acesso à informação, o presente estudo busca mostrar como escolhas de arquitetura alteram a confiabilidade operacional em tarefas jurídicas objetivas, tendo como recorte o domínio dos casos brasileiros julgados pela Corte IDH.

## 2. METODOLOGIA

Esta seção descreve o desenho metodológico do estudo de caso: como o problema foi delimitado, quais artefatos foram observados, quais perguntas orientaram a investigação e quais etapas foram efetivamente executadas. O objeto empírico do trabalho é o comportamento de modelos de linguagem em tarefas objetivas sobre casos brasileiros da Corte IDH, observado sob diferentes regimes de acesso à informação e, posteriormente, desdobrado em um protótipo de triagem factual orientado por precedentes.

## 2.1 Problema investigado e questões de pesquisa

O estudo foi estruturado para responder a uma pergunta central: em tarefas jurídicas objetivas sobre casos brasileiros da Corte IDH, o ganho relevante decorre principalmente de conhecimento interno do modelo, de calibração de *prompt*, de uso genérico de ferramentas de busca, ou de uma arquitetura especializada com *corpus* controlado?

Essa pergunta central foi desdobrada em quatro questões operacionais:

a) qual é o desempenho de modelos generalistas em condição *closed-book*, sem acesso a fontes externas;
b) quanto desse desempenho é mediado pela família de *prompt*, e não apenas por memória ou capacidade inferencial;
c) o que muda quando o modelo passa a operar em uma condição com busca genérica, com *browsing/search* semelhante ao uso de um usuário comum;
d) quanto melhora adicionalmente quando o problema é tratado como uma arquitetura documental orientada por evidência, com *tools* locais sobre um *corpus* jurídico controlado.

Usa-se *prompt* para designar a instrução textual dada ao modelo e *prompting* para o desenho iterativo dessas instruções. Usa-se *fine-tuning* para o ajuste adicional de um modelo com exemplos do domínio. Usa-se *tools* para designar funções externas acionáveis pelo sistema durante a resposta. Por fim, usa-se *scoring* para a pontuação automática offline por comparação com gabaritos estruturados.

## 2.2 Unidade de análise e adaptação do estudo de caso

A formulação deste estudo de caso pressupõe descrição do problema, definição de perguntas, adequação da seleção dos elementos observados e compatibilidade entre métodos de coleta, procedimentos de análise e objeto investigado. Neste trabalho, esse enquadramento exigiu explicitar que não houve participantes humanos tradicionais na etapa principal do estudo. As unidades observadas foram **artefatos digitais e documentais**, organizados em quatro níveis:

- **sentenças dos casos** da Corte IDH envolvendo o Brasil, representados por PDFs, metadados e estruturas derivadas;
- **itens** do *benchmark* v1, os que funcionam como unidade repetida de comparação entre leituras e os que são;
- **runs** experimentais, compostos por prompts, respostas, logs, scoring e auditoria;
- **cenários sintéticos** da fase factual posterior, usados para testar analogia controlada e abertura de interface.

A adequação ao formato de estudo de caso pode ser formulada de forma mais precisa. Na síntese de Silva, em diálogo com Vennesson, um caso é um fenômeno escolhido, conceitualizado e analisado empiricamente como manifestação de uma classe mais ampla de fenômenos; e o estudo de caso é uma estratégia de investigação em profundidade capaz de explorar a configuração do caso e elucidar características de fenômenos similares por meio do desenvolvimento e da avaliação de explicações teóricas (Silva, 2018). Esse enquadramento é compatível com a lógica deste estudo, que não trata o repositório, o *benchmark* ou um *run* isolado como fins em si mesmos.

O caso em estudo é o problema da correção factual e da verificabilidade em IA jurídica aplicada à jurisprudência interamericana, empiricamente delimitado pelo comportamento de LLMs em tarefas objetivas sobre casos brasileiros da Corte IDH sob diferentes regimes de acesso à informação. Essa formulação permite ligar o recorte empírico do trabalho a uma classe mais ampla de problemas sobre *benchmarking* jurídico, uso de modelos em domínios de alto risco e desenho de arquiteturas auditáveis.

## 2.3 Corpus, recorte e justificativa do tamanho

O *corpus* do estudo foi delimitado a **casos contenciosos envolvendo o Brasil julgados pela Corte IDH**. Na versão consolidada do repositório, isso corresponde a 19 PDFs originais organizados no repositório.

A partir dos documentos, o projeto abriu uma camada de extração textual que preserva a numeração dos parágrafos e permite rastrear como cada item do *benchmark* foi derivado. A extração do bloco declarativo da sentença alimentou o detalhamento das violações e outros campos derivados; em paralelo, os parágrafos citados foram gravados por caso em subpastas, com um arquivo *JSONL* para cada sentença.

No eixo das reparações, é importante distinguir duas fontes. As medidas ordenadas foram ancoradas nos pontos resolutivos e no texto da Sentença; já os *status* de cumprimento não foram extraídos dos PDFs das sentenças, mas compilados na camada a partir da página *Casos Contenciosos Brasileiros*, da UMF/CNJ, que declara ter por base as informações da página *Casos em etapa de Supervisão de Cumprimento de Sentença*, da própria Corte IDH (Conselho Nacional de Justiça, 2026; Corte Interamericana de Direitos Humanos, 2026). Essa separação entre *corpus* bruto, texto endereçável por parágrafos e camadas estruturadas foi central para preservar auditabilidade e reuso ao longo das quatro leituras.

Em etapa posterior, já fora do núcleo do *benchmark*, o projeto abriu ainda uma camada derivada de segmentação heurística por seções, acompanhada por metadados. Essa camada não substituiu o pipeline anterior; ela apenas acrescentou um novo modo de navegação documental voltado ao protótipo de triagem factual. A Figura 1 sintetiza esse fluxo.

**Figura 1 – Pipeline documental do estudo: dos 19 PDFs originais às camadas estruturadas reutilizadas no *benchmark* e na etapa factual posterior.**




**Fonte:** elaboração própria.

O *dataset* principal do experimento é **`benchmark/v1/items_closed_book.jsonl`**, com 55 itens distribuídos em três tarefas:

- **Task A (A articles):** 19 itens, cobrindo todos os 19 casos do *corpus*;
- **Task B (B reparations):** 18 itens, cobrindo os casos com camada estruturada de reparações disponível;
- **Task C (C evidence qa):** 18 itens, cobrindo o mesmo subconjunto de casos da Task B.

A diferença entre a cobertura da Task A (19 itens) e das Tasks B e C (18 itens) decorre de uma exclusão substantiva. O caso *Nogueira de Carvalho e Outro vs. Brasil* (Corte IDH, sentença de 28 de novembro de 2006, Série C n. 161), embora integre o *corpus* e permaneça na Task A, não gerou condenação internacional do Estado brasileiro nem imposição de medidas de reparação. Como as Tasks B (**`B_reparations`**) e C (**`C_evidence_qa`**) se ancoram precisamente no núcleo declarativo-reparatório dos casos condenatórios, esse caso foi removido desse subconjunto para preservar coerência entre unidade de análise e variável avaliada.

## *2.4 Tarefas, outputs e comparabilidade*

O *benchmark* v1 foi organizado em três tarefas, todas com *output* estruturado em JSON:

- **Task A:** lista de direitos declarados violados na parte declarativa dos pontos resolutivos;
- **Task B:** lista de medidas de reparação e respectivos estados de cumprimento;
- **Task C:** perguntas objetivas sobre o status de cada medida de reparação.

Nas Tasks B e C, os *status* do gabarito não derivam diretamente das sentenças originais, mas da camada estruturada de supervisão de cumprimento compilada a partir da página do CNJ e, em última instância, das informações publicadas pela Corte IDH sobre supervisão de sentença (Conselho Nacional de Justiça, 2026; Corte Interamericana de Direitos Humanos, 2026).

O desenho comparativo dependeu de manter constante o *dataset*, o contrato de saída e a instrumentação de *scoring*, alterando apenas o regime de acesso à informação. Isso produziu quatro leituras sucessivas:

1. **Leitura 1** — *closed-book* conservadora;
2. **Leitura 2** — *closed-book* recalibrada, com nova família de *prompt*;
3. **Leitura 3** — condição com busca genérica, com *browsing*/*search* genérico;
4. **Leitura 4** — arquitetura documental orientada por evidência, com *tools* locais e *corpus* controlado.

Nas Leituras 1 e 2, três modelos foram comparados:

- OpenAI **`gpt-5.2`**;
- Google **`gemini-3.1-pro-preview`**; e
- Anthropic **`claude-opus-4-6`**.

Nas Leituras 3 e 4, por restrição deliberada de custo e escopo, a comparação foi conduzida apenas com OpenAI **`gpt-5.2`**. Nessa etapa, o objetivo foi comparar regimes de acesso à informação entre as leituras, mantendo fixo o provedor para isolar o efeito da arquitetura e do acesso a fontes.

Também é importante explicitar que a Leitura 4 não foi implementada como um RAG convencional. Em um pipeline de *retrieval-augmented generation*, o núcleo costuma estar na recuperação de trechos para ampliação de contexto antes da geração. Na arquitetura documental orientada por evidência adotada aqui, a recuperação inicial é apenas uma etapa de um fluxo mais amplo, que inclui consultas estruturadas de violações e reparações, abertura contígua de janelas de parágrafos, navegação orientada por evidência e contrato de saída auditável. A distinção importa porque a comparação central desta pesquisa não é entre "sem RAG" e "com RAG", mas entre regimes de acesso à informação com diferentes níveis de controle documental, *grounding* e verificabilidade (Stern et al., 2024; Hu et al., 2026; Koenecke et al., 2025).

## 2.5 Protocolo de scoring e auditoria

O protocolo de avaliação combinou dois níveis complementares:

1. ***scoring* automático**, com métricas derivadas do gabarito estruturado, incluindo F1 e correspondência exata para artigos, além de acurácia e completude para *status*;
2. **auditoria manual integral**, com uma rubrica simples de 0 a 2 aplicada a 100% dos itens (2 = resposta correta no núcleo, sem invenção relevante; 1 = resposta parcialmente correta, com omissões ou pequenos desvios; 0 = resposta incorreta, com erro no núcleo, invenção principal ou falha de formato).

A definição detalhada dessa escala está no repositório do estudo e foi mantida igual ao longo das leituras principais, para que a comparação entre condições usasse o mesmo critério. A auditoria foi feita por um único avaliador e sem mascaramento, o que configura limitação relevante, parcialmente mitigada pela aplicação integral da mesma rubrica a todos os itens e leituras.

Em termos operacionais, o *scoring* automático não substitui a avaliação substantiva do item. Sua função é padronizar a comparação com o gabarito estruturado e produzir um primeiro nível de mensuração reproduzível, que depois é complementado pela auditoria manual.

## 2.6 Passo a passo executado

As Fases 1 a 4 compõem o núcleo empírico do estudo e correspondem ao percurso do *benchmark* propriamente dito. As Fases 5 e 6 registram o desdobramento aplicado posterior aberto a partir dessa evidência, já no eixo do protótipo de triagem factual e de sua interface.

### 2.6.1 Fase 1: construção do corpus e do benchmark

Na primeira fase, foram organizados os 19 PDFs originais e os metadados consolidados por caso. Em seguida, foi executado um pipeline de extração textual orientado à preservação dos parágrafos numerados e ao parseamento da parte declarativa das sentenças. Esse passo permitiu transformar decisões longas em artefatos auditáveis: texto por parágrafo, violações estruturadas, relações com outros artigos da CADH e, no subconjunto pertinente, medidas de reparação identificadas nas próprias sentenças. Os estados de cumprimento dessas medidas foram agregados em etapa distinta, a partir da página *Casos Contenciosos Brasileiros*, do CNJ, que compila dados da página *Casos em etapa de Supervisão de Cumprimento de Sentença*, da Corte IDH (Conselho Nacional de Justiça, 2026; Corte Interamericana de Direitos Humanos, 2026).

Só depois dessa preparação documental o *benchmark* v1 foi efetivamente montado. Os candidatos das tarefas A, B e C foram gerados a partir das camadas estruturadas, revisados manualmente e reduzidos até a configuração final de 55 itens *closed-book*. Essa ordem importa metodologicamente: o *benchmark* não foi escrito "por inspeção" do pesquisador sobre PDFs isolados, mas derivado de um pipeline que separa *corpus* bruto, evidência textual e gabaritos estruturados.

### 2.6.2 Fase 2: Leituras 1 e 2 (closed-book)

Na segunda fase, foram executadas duas famílias de *prompt* sobre o mesmo *benchmark closed-book*. A Leitura 1 usou um *prompt* mais conservador e fortemente anti-invenção. A Leitura 2 manteve o regime *closed-book*, mas recalibrou o *prompt* para reduzir abstenção improdutiva. Em ambas, foram comparados OpenAI, Google e Anthropic com `temperature = 0.0`, JSON estruturado e *scoring* offline padronizado.

### 2.6.3 Fase 3: Leitura 3 (condição com busca genérica)

Na terceira fase, o estudo mudou de pergunta empírica. Em vez de perguntar apenas "o modelo sabe isso de memória?", passou-se a perguntar "o que um usuário comum conseguiria obter com um modelo comercial de ponta dotado de *search*?". Por isso, a Leitura 3 proibiu acesso privilegiado ao *corpus* local e permitiu apenas *browsing/search* genérico. Essa leitura foi executada apenas com OpenAI `gpt-5.2`, em razão do custo adicional envolvido em *browsing/search*.

### 2.6.4 Fase 4: Leitura 4 (arquitetura documental orientada por evidência)

Na quarta fase, o problema deixou de ser tratado como uso genérico de ferramenta e passou a ser tratado como desenho de sistema. A arquitetura documental orientada por evidência operou com um conjunto controlado de *tools* locais sobre o *corpus* do projeto, incluindo consulta estruturada de violações e reparações, busca inicial via FTS (*full-text search*, isto é, busca textual integral) e abertura de janelas de parágrafos. Essa leitura foi suficiente, na versão atual do trabalho, para sustentar empiricamente o argumento de ganho de arquitetura.

### 2.6.5 Fase 5: abertura do protótipo de triagem factual

Depois de consolidada a Leitura 4, o projeto abriu uma etapa aplicada posterior: um protótipo de triagem factual orientado por precedentes brasileiros da Corte IDH. Trata-se de Produto Mínimo Viável (MVP), suficiente para validar o fluxo central da ferramenta antes de qualquer pretensão de uso institucional ampliado. Como passo preparatório, foi executada uma camada derivada de segmentação heurística por seções das sentenças originais. Essa camada passou a funcionar como infraestrutura de apoio ao protótipo factual.

Essa fase já não parte de um caso conhecido, mas de narrativas factuais novas, transformadas em cenários sintéticos controlados. A primeira bateria foi organizada com cenários quase-isomórficos a três casos do *corpus*, para estabilizar recuperação, *ranking*, segmentação e *schema* de saída.

Em seguida, o trabalho abriu uma segunda camada, **`scenarios_v2`**, composta por cenários análogos, mas não idênticos, para medir capacidade de generalização por analogia. Essa etapa introduziu a distinção entre **análogo principal e análogo de apoio**, além de uma rubrica qualitativa simples voltada à utilidade do *output*. A Camada B atingiu resultado agregado de 22/24 e foi considerada suficiente para abrir a primeira interface de entrada factual.

### *2.6.6 Fase 6: interface e arquitetura web do protótipo de triagem factual*

O primeiro meio de uso do protótipo foi aberto em **`Streamlit`**, como camada histórica de validação rápida do fluxo factual. Essa fase não reabre o experimento principal; ela apenas transforma o desdobramento aplicado em superfície observável de uso. Nessa fase, a expressão entrada factual designa precisamente a etapa inicial em que o usuário insere a narrativa de fatos a ser estruturada e analisada pelo sistema. Posteriormente, o projeto preservou essa interface e abriu uma nova arquitetura *web* em duas camadas: **`FastAPI`**, em **`apps/api/`**, como camada HTTP e WebSocket sobre o serviço factual, e **`Next.js`**, em **`apps/web/`**, como interface final do protótipo.

## *2.7 Transparência sobre o uso de IA na pesquisa e na redação*

Além de constituir o próprio objeto do estudo, ferramentas baseadas em IA foram usadas como apoio instrumental no processo de pesquisa e redação. A literatura recente sobre *disclosure* recomenda distinguir usos substanciais (aqueles que afetam conteúdo, análise, redação relevante ou reprodutibilidade) de usos triviais, para os quais a declaração pode ser opcional ou desnecessária (Resnik e Hosseini, 2026).

Entre 2025 e 2026, foram utilizadas ferramentas baseadas em LLM, principalmente em ambiente OpenAI/Codex, como apoio a quatro grupos de tarefas: (i) exploração e depuração de scripts, runners e artefatos do repositório; (ii) organização de saídas experimentais, conversões de formato e preparação de figuras e documentos auxiliares; (iii) revisão estrutural e editorial de trechos do manuscrito; e (iv) apoio à curadoria bibliográfica e à conferência de citações, sempre com retorno aos PDFs originais antes da incorporação de passagens no texto final. Esse uso é compatível com a literatura que defende escrita assistida com curadoria crítica, reescrita autoral, verificação rigorosa de fatos e fontes, e declaração transparente do emprego de modelos de inteligência artificial (Sampaio, 2025; Resnik e Hosseini, 2026).

Esse apoio não implicou delegação de autoria nem aceitação automática de conteúdo. A leitura dos casos da Corte IDH, a definição do recorte, a construção dos gabaritos, a auditoria manual, a seleção final das citações e a interpretação dos resultados permaneceram sob responsabilidade do autor. Em linha com a literatura consultada, a responsabilidade final pelo manuscrito, por suas afirmações e por sua integridade metodológica é integralmente humana (Sampaio, 2025; Resnik e Hosseini, 2026).

## *2.8 Limitações metodológicas*

Algumas limitações permaneceram relevantes ao longo do desenho e precisam ser lidas de forma prática. Primeiro, a auditoria manual foi conduzida por um único avaliador e sem mascaramento, o que amplia o risco de viés de julgamento, mesmo com rubrica fixa e aplicação integral aos itens. Segundo, há uma assimetria entre as leituras: nas Leituras 1 e 2

houve comparação entre provedores, enquanto nas Leituras 3 e 4 a execução ficou restrita a um único provedor nesta versão do estudo. Com isso, o ganho observado nessa etapa deve ser interpretado principalmente como efeito do regime de acesso à informação dentro dessa condição experimental, e não como ranking geral entre os LLMs comerciais.

Terceiro, a Leitura 4 dependeu fortemente de *tools* estruturadas do *corpus*. Esse desenho melhora a confiabilidade operacional, mas ainda não configura, em todos os casos, um contrato pleno de evidência em nível de parágrafo para cada afirmação produzida. Quarto, o recorte empírico ficou limitado aos casos brasileiros da Corte IDH: essa escolha favorece profundidade e auditabilidade, porém reduz a capacidade de generalizar os resultados para outros países, tribunais ou tipos de litígio.

Na fase posterior do protótipo de triagem factual, soma-se uma limitação adicional: os cenários usados até aqui são sintéticos e construídos pelo pesquisador. Eles funcionam bem para validar o pipeline em ambiente controlado, mas não substituem testes com usuários reais nem narrativas oriundas de fluxos institucionais concretos, que tendem a trazer ruído, ambiguidades e restrições próprias do uso em campo.

### 2.9 Disponibilidade de Dados e Código

Os dados brutos e processados utilizados neste estudo, incluindo *corpus* documental, metadados, itens do benchmark, gabaritos estruturados, outputs de avaliação e artefatos de auditoria, estão disponíveis em repositório aberto em https://doi.org/10.5281/zenodo.21011576. O código-fonte empregado para extração, organização, execução dos *runners*, *scoring* automático, auditoria e prototipação está disponível em https://github.com/luizzzvictor/corte-idh-llm-benchmark-code, acompanhado de documentação de reprodução no arquivo README.md.

## 3. RESULTADOS

Esta seção consolida os resultados do *benchmark* v1 em quatro leituras sucessivas do mesmo *dataset* de 55 itens e, em seguida, apresenta o desdobramento aplicado posterior que levou ao protótipo de triagem factual. A Leitura 1 mediu conhecimento interno sob um *prompt* conservador; a Leitura 2 manteve a condição *closed-book* e recalibrou o *prompt*; a Leitura 3 mudou o regime para uma condição com busca genérica, com *browsing/search* genérico; e a Leitura 4 executou uma arquitetura documental orientada por evidência, com *corpus* local controlado e ferramentas especializadas.

### 3.1 Visão geral das quatro leituras

A Tabela 1 resume a média do *score* auditado (0–2) para o OpenAI `gpt-5.2` ao longo das quatro leituras. O comportamento mais importante não é uma simples curva de melhora monotônica por "modelo", mas a mudança de patamar provocada por alterações no regime de acesso à informação. O salto mais substantivo aparece na passagem da Leitura 3 para a Leitura 4 em tarefas fortemente dependentes de fonte e completude

**Tabela 1 – *Score* auditado (0–2) por leitura para OpenAI gpt-5.2 (55 itens).**

| Leitura | Média global | A (artigos) | B (reparações) | C (QA) |
|---|---|---|---|---|
| L1 (*closed-book* cons.) | 200 | 421 | 0 | 167 |
| L2 (*closed-book* rec.) | 855 | 579 | 500 | 1.500 |
| L3 (busca genérica) | 1.327 | 1.421 | 944 | 1.611 |
| L4 (arq. documental) | 1.945 | 1.842 | 2.000 | 2.000 |

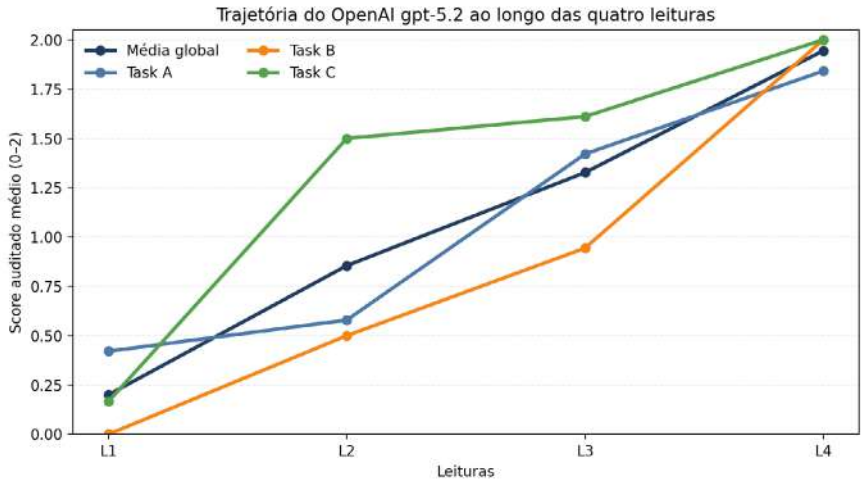
A Tabela 2 apresenta, como complemento, as métricas automáticas do *scorer* estruturado. Em `B_reparations` e `C_evidence_qa`, a Leitura 4 atinge teto automático, alinhando-se ao teto auditado. Em `A_articles`, o *score* automático da Leitura 4 fica artificialmente mais baixo do que a leitura substantiva da auditoria, porque o modelo frequentemente devolveu títulos com prefixos numéricos do artigo, o que o *scorer* não normaliza, e optou-se por não alterar o *scorer* para adaptar à Leitura 4.

**Tabela 2 – Métricas automáticas por leitura para OpenAI `gpt-5.2`.**

| Leitura | A F1 | A EM | B acc | B compl | C acc | C compl |
|---|---|---|---|---|---|---|
| L1 | 342 | 53 | 0 | 0 | 20 | 50 |
| L2 | 335 | 105 | 229 | 371 | 821 | 1.000 |
| L3 | 710 | 421 | 688 | 894 | 890 | 1.000 |
| L4 | 625 | 579 | 1.000 | 1.000 | 1.000 | 1.000 |

A Figura 2 deixa esse padrão mais claro em três degraus. Da Leitura 1 para a 2, há ganho pela recalibração de *prompt*; da 2 para a 3, aparece ganho adicional com *browsing/search* genérico; e da 3 para a 4 ocorre o salto mais forte, quando o problema passa a ser tratado como arquitetura documental orientada por evidência. Esse efeito fica mais nítido nas tarefas **B** e **C**, que dependem mais de fonte hierarquizada e de completude documental.

**Figura 2 – Trajetória do OpenAI gpt-5.2 ao longo das quatro leituras, com base no score auditado médio (0–2).**




**Fonte:** elaboração própria

## 3.2 Leituras 1 e 2: benchmark closed-book

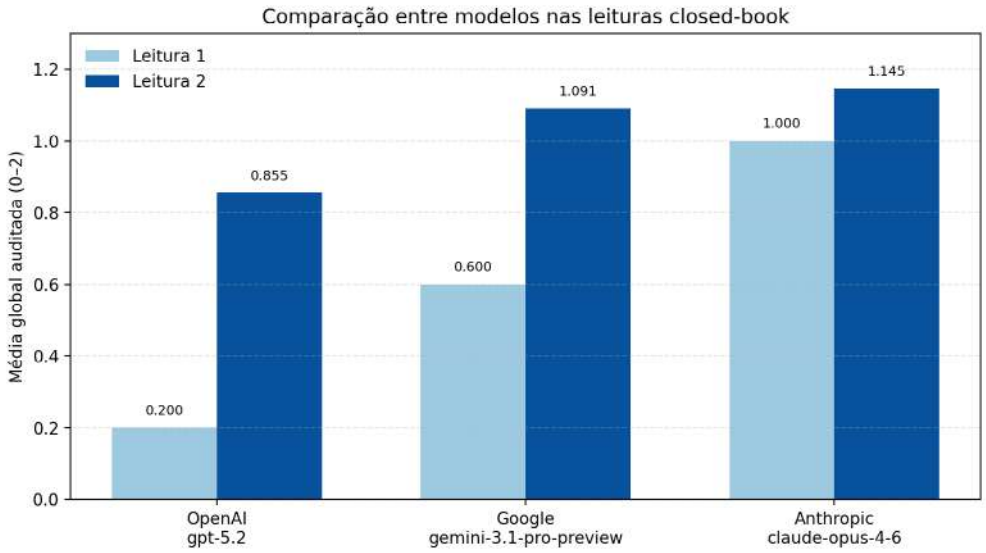
Nas leituras *closed-book*, três modelos foram avaliados: OpenAI **`gpt-5.2`**, Google **`gemini-3.1-pro-preview`** e Anthropic **`claude-opus-4-6`**. O achado principal dessa etapa foi que o *prompt* da Leitura 1, excessivamente orientado a evitar invenção, induzia abstenção improdutiva, sobretudo no **`gpt-5.2`**. A Leitura 2 manteve o mesmo *dataset* e a mesma instrumentação, mudando apenas a família de *prompt*, e isso já foi suficiente para produzir melhora material em **B** e **C**.

**Tabela 3 – Score auditado (0–2) em *closed-book* por modelo (L1 vs. L2).**

| Leitura | Modelo | Média global | A | B | C |
|---|---|---|---|---|---|
| L1 | OpenAI **`gpt-5.2`** | 200 | 421 | 0 | 167 |
| L1 | Google **`gemini-3.1-pro-preview`** | 600 | 1.000 | 444 | 333 |
| L1 | Anthropic **`claude-opus-4-6`** | 1.000 | 947 | 944 | 1.111 |
| L2 | OpenAI **`gpt-5.2`** | 855 | 579 | 500 | 1.500 |
| L2 | Google **`gemini-3.1-pro-preview`** | 1.091 | 1.053 | 778 | 1.444 |
| L2 | Anthropic **`claude-opus-4-6`** | 1.145 | 1.000 | 1.056 | 1.389 |

A Figura 3 ajuda a explicitar a assimetria entre os três modelos nas leituras *closed-book*. Ela mostra que a comparação entre provedores é metodologicamente forte apenas em **`L1`** e **`L2`**, isto é, antes da abertura da condição com busca genérica e da arquitetura documental orientada por evidência. Também fica mais claro, visualmente, que a maior recuperação entre as duas leituras ocorre no OpenAI **`gpt-5.2`**, o que reforça a interpretação de que a Leitura 1 o penalizava por excesso de cautela induzida por *prompt*.

**Figura 3 – Comparação entre modelos nas leituras *closed-book*, usando a média global auditada (0–2).**




**Fonte:** elaboração própria.

Em termos substantivos, a comparação entre Leitura 1 e Leitura 2 permite separar dois tipos de problema. O primeiro é o limite real do conhecimento interno do modelo em temas de alta granularidade documental. O segundo é um limite artificial de desenho de *prompt*, que transforma cautela em silêncio. O caso mais claro é o `gpt-5.2`: o modelo praticamente colapsa em `B` e `C` na Leitura 1, mas recupera fortemente na Leitura 2. Esse comportamento mostra que avaliações *closed-book* podem subestimar a utilidade prática de um modelo quando o *prompt* confunde antialucinação com abstenção global.

## 3.3 Leitura 3: condição com busca genérica

A Leitura 3 alterou a pergunta empírica. Em vez de perguntar apenas "o que o modelo sabe de memória?", passou-se a perguntar "o que um usuário comum consegue obter com *browsing/search* genérico?". Nessa fase, a execução foi deliberadamente restrita à OpenAI `gpt-5.2`, por limitação prática de custo.

Os resultados automáticos da Leitura 3 já indicaram melhora forte sobre a Leitura 2, especialmente em `A_articles` e `B_reparations`. A auditoria manual confirmou esse ganho:

**Tabela 4 – Comparação auditada entre Leitura 2 e Leitura 3 para OpenAI `gpt-5.2`.**

| Condição | Score 0 | Score 1 | Score 2 | Média auditada |
|---|---|---|---|---|
| L2 (*closed-book* recalibrada) | 19 | 25 | 11 | 855 |
| L3 (busca genérica) | 4 | 29 | 22 | 1.327 |

Por tarefa, a melhora foi de 0.579 para 1.421 em `A_articles`, de 0.500 para 0.944 em `B_reparations` e de 1.500 para 1.611 em `C_evidence_qa`. A leitura substantiva é que a busca aberta reduz silenciosamente parte do problema de memória e cobertura, mas não resolve o problema de hierarquia de fontes. O erro residual da Leitura 3 não é mais o vazio; ele passa a ser o uso de fontes abertas incompletas ou não hierarquizadas, o que afeta sobretudo status de cumprimento.

## 3.4 Leitura 4: arquitetura documental orientada por evidência

A Leitura 4 respondeu à pergunta seguinte: quanto melhora quando o problema deixa de ser tratado como uso genérico de *search* e passa a ser tratado como arquitetura jurídica especializada? Nessa condição, o modelo operou com *tools* locais sobre o *corpus* do projeto, incluindo metadados estruturados de violações e reparações, além de ferramentas de busca e abertura de parágrafos.

A comparação mais importante do trabalho é a da Leitura 4 contra a Leitura 3. No *endpoint* primário, — média auditada da Task B — a diferença foi de 0.944 para 2.000. No nível do item, 16 dos 18 itens de `B_reparations` melhoraram, dois empataram e nenhum piorou. O teste de sinais exato, aplicado de forma complementar, foi consistente com a melhoria sistemática nessa transição.

No conjunto inteiro do *benchmark*, a Leitura 4 atingiu média auditada global de 1.945, contra 1.327 da Leitura 3. Em B e C, a arquitetura documental orientada por evidência atingiu teto auditado; em A, a auditoria manual mostrou que o desempenho substantivo também foi

alto, apesar da subcontagem automática causada pelo formato de alguns títulos. Em termos práticos, isso significa que o ganho decisivo apareceu justamente onde a hierarquia de fonte e a completude documental são mais importantes.

## 3.5 Desdobramento aplicado: abertura do protótipo de triagem factual

Depois de consolidada a Leitura 4, o trabalho avançou para uma etapa aplicada posterior: um protótipo de triagem factual orientado por precedentes brasileiros da Corte IDH. Essa fase não substitui o *benchmark*; ela o utiliza como base empírica para justificar a passagem de uma comparação entre regimes de acesso à informação para uma ferramenta de apoio à leitura de narrativas factuais. Por isso, os resultados a seguir devem ser lidos como desdobramento aplicado subsequente, e não como bloco empírico equivalente às quatro leituras do *benchmark*.

O primeiro piloto factual foi organizado em `scenarios_v1`, com três narrativas quase-isomórficas aos casos-âncora do *corpus*. Essa fase teve função de depuração: estabilizar recuperação, *ranking*, segmentação de seções e contrato de saída. A etapa seguinte abriu `scenarios_v2`, com cenários análogos, mas não idênticos, destinados a medir generalização controlada por analogia.

A leitura qualitativa da Camada B mostrou três resultados relevantes. Primeiro, o análogo principal ficou defensável nos três eixos testados. Segundo, a distinção entre análogo principal e análogo de apoio tornou a saída mais explicável. Terceiro, as listas de `information_gaps` se mostraram úteis e acionáveis. Aplicando uma rubrica qualitativa simples de quatro critérios, a Camada B atingiu **22/24**, resultado considerado suficiente para abrir a primeira interface de entrada factual.

**Tabela 5 – Síntese qualitativa da Camada B do *MVP* factual.**

| Cenário | Principal | Apoio | Lacunas | Evid. vs hip. | Total |
|---|---|---|---|---|---|
| `triage_v2_001` | 2 | 1 | 2 | 2 | 7/8 |
| `triage_v2_002` | 2 | 1 | 2 | 2 | 7/8 |
| `triage_v2_003` | 2 | 2 | 2 | 2 | 8/8 |
| Total | | | | | 22/24 |

Na etapa seguinte, o protótipo passou a operar em uma *stack web* própria, preservando a superfície histórica em *Streamlit* e abrindo uma nova arquitetura de aplicação *web*.

A camada de *backend* foi implementada em *FastAPI*, integrada diretamente ao serviço factual em *Python* já validado nos pilotos anteriores, enquanto a interface principal foi implementada em *Next.js*. Essa arquitetura passou a expor *endpoints* dos modelos disponíveis, cenários de amostra e análise factual, além de um canal *WebSocket* para acompanhamento da execução em tempo real. Com isso, o protótipo passou a combinar reaproveitamento do pipeline factual já existente com uma camada de uso mais estável e apresentável para demonstração do fluxo.

**Figura 4 – Interface do MVP de triagem factual.**




**Fonte:** elaboração própria.

Do ponto de vista funcional, a interface atual recebe uma narrativa factual como entrada principal e admite campos contextuais complementares, como local, período e seleção de qual modelo irá fazer a triagem. A execução pode partir de cenários de teste carregados do próprio repositório ou de um relato novo inserido manualmente. Uma vez iniciada a análise, a interface envia a narrativa à API, que a transforma em item factual estruturado e a encaminha à *runner* que pode consultar tools locais sobre o *corpus* controlado, buscar casos análogos, abrir metadados, listar seções, filtrar parágrafos relevantes, recuperar violações declaradas e consultar medidas reparatórias.

**Figura 5 – Exemplo de tela após execução do MVP de triagem factual.**




**Fonte:** elaboração própria.

A saída final do protótipo passou a ser organizada em blocos estáveis de leitura. A interface renderiza um resumo factual estruturado, apresenta hipóteses de direitos com indicação de confiança, distingue precedentes por função analítica, separando análogo principal e análogo de apoio, explicita um bloco de suporte com notas e referências citáveis e encerra a resposta com uma lista de lacunas de informação. Em paralelo, a aplicação exibe um painel lateral de

observabilidade, no qual os eventos da sessão são mostrados em tempo real em linguagem mais legível, preservando ainda a possibilidade de expansão de detalhes técnicos. Esse desenho dá maior visibilidade à separação entre evidência, hipótese, analogia e incerteza, que é uma das exigências centrais do fluxo factual adotado no projeto.

## CONCLUSÃO

Este estudo investigou a aplicação de modelos de linguagem em tarefas jurídicas objetivas no domínio de casos brasileiros julgados pela Corte IDH. O estudo deslocou a comparação de modelos isolados para a comparação entre regimes de acesso à informação. A principal conclusão é que, nesse domínio, o ganho decisivo de confiabilidade decorre do desenho de sistema: *corpus* controlado, *tools* especializadas, contratos de saída e possibilidade de rastrear a base documental da resposta.

Do ponto de vista empírico, o trabalho entregou um *benchmark* jurídico especializado em português, com 55 itens, três tarefas, gabaritos estruturados, *scoring* offline, auditoria manual integral e quatro leituras experimentais. Também entregou uma trilha inicial de triagem factual orientada por precedentes brasileiros da Corte IDH, validada em cenários sintéticos progressivos e com primeira interface aberta.

Os resultados indicam que a condição *closed-book* é útil como linha de base, mas insuficiente para medir confiabilidade operacional em tarefas documentais. A recalibração de prompt melhorou o desempenho e reduziu abstenções. A busca genérica aproximou o experimento de um uso comum de produto comercial e elevou a cobertura. A arquitetura documental orientada por evidência produziu o ganho mais robusto, especialmente em reparações e status de cumprimento, em que completude, atualização e hierarquia de fonte são centrais.

A contribuição metodológica está em tratar avaliação de LLMs jurídicos como problema de sistema. O *benchmark* compara regimes de acesso à informação, registra o percurso experimental, preserva artefatos de execução e combina *scoring* automático com auditoria manual. Essa estrutura permite interpretar erros com mais precisão: ausência de fonte, uso inadequado de fonte, omissão parcial, falha de formato, atualização incorreta ou invenção substantiva.

A contribuição aplicada está na tradução inicial desse achado para uma ferramenta de triagem factual. O protótipo factual apoia organização de narrativas, recuperação de precedentes brasileiros da Corte IDH, indicação de direitos potencialmente implicados e explicitação de lacunas informacionais. Sua utilidade depende de prudência metodológica: a saída trabalha com hipóteses, analogias e evidências, preservando espaço para revisão humana.

As limitações permanecem importantes. A auditoria foi conduzida por um único avaliador. As Leituras 3 e 4 ficaram restritas ao modelo OpenAI *gpt-5.2*. O *corpus* cobre apenas casos brasileiros da Corte IDH. A Leitura 4 depende de ferramentas estruturadas do *corpus* e pode ser fortalecida por evidência em nível de parágrafo para cada afirmação. A etapa factual posterior usou cenários sintéticos, que precisam ser complementados por testes com usuários e narrativas reais.

Essas limitações também indicam como o trabalho pode ser acumulado. A mesma estrutura de *benchmark* permite reexecução com novos provedores, novos prompts, outros recortes nacionais e novas versões da arquitetura. A cada mudança, os mesmos itens podem

funcionar como suíte de regressão, mostrando se houve ganho, perda ou mudança no tipo de erro. Essa característica transforma o *benchmark* em instrumento de pesquisa e também em ferramenta de engenharia do sistema.

Os trabalhos futuros seguem três direções. A primeira é ampliar a avaliação para outros provedores, países e conjuntos documentais. A segunda é fortalecer o contrato de evidência em nível de parágrafo, aproximando o fluxo experimental da arquitetura final do sistema. A terceira é desenvolver a triagem factual com avaliação por usuários e especialistas externos, mantendo a distinção entre evidência, hipótese e conclusão jurídica.

Em síntese, o estudo mostra que a pergunta central em IA jurídica documental não se resolve apenas pela escolha de um modelo. A confiabilidade emerge da combinação entre *corpus* delimitado, ferramentas adequadas, saída estruturada, rastreabilidade e auditoria humana. No recorte dos casos brasileiros da Corte IDH, a arquitetura documental orientada por evidência mostrou-se o caminho mais consistente para transformar respostas plausíveis em respostas verificáveis.



## REFERÊNCIAS


ARIAI, Farid; MACKENZIE, Joel; DEMARTINI, Gianluca. Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models and Challenges. **ACM Computing Surveys**. 2025. arXiv:2410.21306v3. Disponível em: https://arxiv.org/abs/2410.21306. Acesso em: 21 mar. 2026.

BEAN, Andrew M.; et al. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. **arXiv preprint arXiv**:2511.04703. 2025. v1, 3 Nov 2025. Disponível em: https://arxiv.org/abs/2511.04703. Acesso em: 21 mar. 2026.

CHALKIDIS, Ilias; et al. FairLex: A Multilingual Benchmark for Evaluating Fairness in Legal Text Processing. **arXiv preprint arXiv**:2203.07228. 2022. v1, 14 Mar 2022. Disponível em: https://arxiv.org/abs/2203.07228. Acesso em: 21 mar. 2026.

CHLAPANIS, Odysseas S.; GALANIS, Dimitrios; ANDROUTSOPOULOS, Ion. LAR-ECHR: A New Legal Argument Reasoning Task and Dataset for Cases of the European Court of Human Rights. **arXiv preprint arXiv**:2410.13352. 2024. v1, 17 Oct 2024. Disponível em: https://arxiv.org/abs/2410.13352. Acesso em: 21 mar. 2026.

Conselho Nacional de Justiça. **Casos Contenciosos Brasileiros**. 2026. A página informa compilar os dados com base na seção de supervisão de cumprimento da Corte IDH; última atualização indicada na página: 21/02/2025. Disponível em: https://www.cnj.jus.br/poder-judiciario/relacoes-internacionais/monitoramento-e-fiscalizacao-das-decisoes-da-corte-idh/casos-contenciosos-brasileiros/. Acesso em: 22 mar. 2026.

Conselho Nacional de Justiça. Comportamento judicial em relação à Convenção Americana sobre Direitos Humanos: uma análise empírica do Poder Judiciário brasileiro. Conselho Nacional de Justiça. Brasília. 2023. **Relatório produzido em parceria com a Pontifícia Universidade Católica do Paraná** (PUCPR). Disponível em: https://bibliotecadigital.cnj.jus.br/handle/123456789/636?mode=simple. Acesso em: 21 mar. 2026.

Corte Interamericana de Direitos Humanos. **Casos em etapa de Supervisão de Cumprimento de Sentença**. 2026. A página informa que as informações apresentadas se baseiam no que foi declarado em resoluções emitidas pela Corte. Disponível em: https://www.corteidh.or.cr/casos_en_supervision_por_pais.cfm?lang=pt. Acesso em: 22 mar. 2026.

ERIKSSON, Maria; et al. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. **Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society** (AIES 2025). 2025. Disponível em: https://arxiv.org/abs/2502.06559. Acesso em: 21 mar. 2026.

GUHA, Neel; et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. **arXiv preprint arXiv**:2308.11462. 2023. v1, 20 Aug 2023. Disponível em: https://arxiv.org/abs/2308.11462. Acesso em: 21 mar. 2026.

HABERNAL, Ivan; et al. Mining Legal Arguments in Court Decisions. **Artificial Intelligence and Law**. 2023. preprint **arXiv**:2208.06178v2. Disponível em: https://arxiv.org/abs/2208.06178. Acesso em: 21 mar. 2026.

HOU, Abe Bohan; et al. CLERC: A Dataset for Legal Case Retrieval and Retrieval-Augmented Analysis Generation. **arXiv preprint arXiv**:2406.17186. 2024. v2, 27 Jun. 2024. Disponível em: https://arxiv.org/abs/2406.17186. Acesso em: 22 mar. 2026.

HU, Yiran; et al. Evaluation of Large Language Models in Legal Applications: Challenges, Methods, and Future Directions. **arXiv preprint arXiv**:2601.15267. 2026. v1, 21 Jan 2026. Disponível em: https://arxiv.org/abs/2601.15267. Acesso em: 21 mar. 2026.

JING, Huihao; et al. MASLegalBench: Benchmarking Multi-Agent Systems in Deductive Legal Reasoning. **arXiv preprint arXiv**:2509.24922. 2025. v2, 30 Sep 2025. Disponível em: https://arxiv.org/abs/2509.24922. Acesso em: 21 mar. 2026.

KENNEDY, Wm. Matthew; et al. HumRightsBench: A Benchmark for Evaluating Internal Representations of Human Rights Principles in LLMs and LRMs. 2026. **Poster do projeto HumRightsBench**. Disponível em: https://humrightsbench.com/wp-content/uploads/2026/03/CS-Law-26-Poster-HumRightsBench.pdf. Acesso em: 22 mar. 2026.

KOENECKE, Allison; et al. Tasks and Roles in Legal AI: Data Curation, Annotation, and Verification. **arXiv preprint arXiv**:2504.01349. 2025. v1, 2 Apr 2025. Disponível em: https://arxiv.org/abs/2504.01349. Acesso em: 21 mar. 2026.

KÜÇÜK, Dilek; CAN, Fazli. Computational Law: Datasets, Benchmarks, and Ontologies. **arXiv preprint arXiv**:2503.04305. 2025. v2, 13 Mar 2025. Disponível em: https://arxiv.org/abs/2503.04305. Acesso em: 21 mar. 2026.

LI, Charlotte; et al. Towards Ecologically Valid LLM Benchmarks: Understanding and Designing Domain-Centered Evaluations for Journalism Practitioners. **arXiv preprint arXiv**:2511.05501. 2025. Preprint em arXiv. Disponível em: https://arxiv.org/abs/2511.05501. Acesso em: 22 mar. 2026.

PIPITONE, Nicholas; HOUIR ALAMI, Hamza. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. **arXiv preprint arXiv**:2408.10343. 2024. Preprint em arXiv. Disponível em: https://arxiv.org/abs/2408.10343. Acesso em: 22 mar. 2026.

PRABHUMOYE, Shrimai; et al. Focused Attention Improves Document-Grounded Generation. **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Online. 2021. Disponível em: https://aclanthology.org/2021.naacl-main.338/. Acesso em: 22 mar. 2026.

RESNIK, David B.; HOSSEINI, Mohammad. Disclosing Artificial Intelligence Use in Scientific Research and Publication: When Should Disclosure Be Mandatory, Optional, or Unnecessary?. **Accountability in Research**. 2026. Published online 24 Mar. 2025. Disponível em: https://doi.org/10.1080/08989621.2025.2481949. Acesso em: 21 mar. 2026.

REUEL, Anka; et al. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. **arXiv**, 2024. v. 1, 20 nov. 2024. Disponível em: https://arxiv.org/abs/2411.12990. Acesso em: 21 mar. 2026.

SADOWSKI, Albert; CHUDZIAK, Jarosław A. On Verifiable Legal Reasoning: A Multi-Agent Framework with Formalized Knowledge Representations. **arXiv**, 2025. v. 1, 31 ago. 2025. Disponível em: https://arxiv.org/abs/2509.00710. Acesso em: 21 mar. 2026.

SAMPAIO, Rafael Cardoso. Escrita acadêmica ética, responsável e humana com inteligência artificial. **Revista de Sociologia e Política**, 2025. Disponível também em SciELO Preprints. Disponível em: https://doi.org/10.1590/1678-98732433e018. Acesso em: 21 mar. 2026.

SAMWAY, Keenan; et al. When Do Language Models Endorse Limitations on Human Rights Principles?. **arXiv preprint arXiv**:2603.04217. 2026. v1, 4 Mar 2026. Disponível em: https://arxiv.org/abs/2603.04217. Acesso em: 21 mar. 2026.

SANTOSH, T. Y. S. S.; et al. Leveraging Task Dependency and Contrastive Learning for Case Outcome Classification on European Court of Human Rights Cases. **arXiv preprint arXiv**:2302.00768. 2023. v3, 13 Feb 2023. Disponível em: https://arxiv.org/abs/2302.00768. Acesso em: 21 mar. 2026.

SILVA, Glauco Peres da. Desenho de Pesquisa. Enap. Brasília. 2018. **Coleção Metodologias de Pesquisa**. ISBN 978-85-256-0082-0. Disponível em: https://repositorio.enap.gov.br/handle/1/3330. Acesso em: 21 mar. 2026.

STERN, Ronja; et al. One Law, Many Languages: Benchmarking Multilingual Legal Reasoning for Judicial Support. **arXiv preprint arXiv**:2306.09237. 2024. v3, 21 Aug 2024. Disponível em: https://arxiv.org/abs/2306.09237. Acesso em: 21 mar. 2026.

# Consulta em Linguagem Natural a Dados de Compras Centralizadas de Medicamentos do Ministério da Saúde: uma implementação de sistema text-to-SQL com large language model e arquitetura medalhão

*Rosana Elisa Gonçalves Pinho*
*Prof. Dr. Hélio Bonfim de Macêdo Filho*

## Resumo

Este artigo apresenta o desenvolvimento e a avaliação de um sistema Text-to-SQL para ampliar o acesso analítico a dados de compras de medicamentos do Ministério da Saúde. O portal converte perguntas em português em consultas SQL, executadas sobre base analítica estruturada em arquitetura medalhão e alimentada por dados públicos do Compras.gov.br. Na comparação entre modelos, o GPT-4o-mini via API superou o Llama 3.1 8B local (72,2% versus 27,8%) e foi selecionado para o protótipo. No teste beta não supervisionado, registraram-se 172 interações, com taxa geral de sucesso de 73,3% e tempo mediano de resposta de 2,2 segundos. Observou-se satisfação favorável dos usuários (75%). Os resultados indicam melhor desempenho em consultas simples e intermediárias, e menor robustez em perguntas complexas e fora de domínio. Conclui-se que a abordagem é tecnicamente viável e apresenta utilidade prática para apoiar o consumo analítico de dados públicos na gestão pública federal.

**Palavras-chave:** Text-to-SQL. Inteligência Artificial. Compras públicas.

## *INTRODUÇÃO*

A gestão de compras públicas de medicamentos no Brasil movimenta volumes financeiros expressivos: o Ministério da Saúde (MS) acumulou mais de R$ 86,6 bilhões em contratos farmacêuticos entre 2021 e 2025, segundo dados do portal Compras.gov.br. Apesar da disponibilidade pública dessas informações, seu aproveitamento analítico permanece restrito, na prática, a profissionais com domínio técnico em coleta de dados via API (Application Programming Interface) , processamento, modelagem e linguagens de consulta estruturada. Essa dependência de intermediação retarda ciclos de decisão e limita a autonomia analítica dos gestores e equipes técnicas.

Esse cenário evidencia um paradoxo da era dos dados abertos governamentais: o próprio órgão que produz e publica os dados nem sempre dispõe de meios internos para consumi-los analiticamente com agilidade, dependendo de intermediação especializada para transformar informação pública em subsídio à decisão. O Banco Mundial reconhece que governos são frequentemente consumidores importantes de seus próprios dados abertos (World Bank, 2019), e Rogger e Schuster (2023) demonstram que o reaproveitamento de dados administrativos

para *analytics* interno representa uma das maiores oportunidades de modernização da gestão pública. As compras governamentais constituem um setor de particular relevância nesse contexto: Bosio et al. (2022), em estudo abrangendo 187 países, estimam que o volume global de compras públicas equivale a cerca de 12% do PIB mundial, percentual que alcança 13% entre os países membros da OCDE (Organiza ção para a Cooperação e Desenvolvimento Econômico), evidenciando o potencial analítico e fiscal associado à melhoria da capacidade de consumo desses dados pelos próprios órgãos contratantes.

Nesse contexto, este artigo tem por objetivo desenvolver e avaliar o Portal de Análise de Compras de Medicamentos, um sistema Text-to-SQL que permite a gestores, técnicos do MS e demais interessados consultar dados de aquisição de medicamentos em linguagem natural, sem necessidade de conhecimento técnico em SQL. O sistema converte automaticamente perguntas formuladas em português em consultas SQL executadas sobre uma base analítica construída com dados públicos do Compras.gov.br, retornando resultados em formato tabular e gráfico.

O diferencial metodológico do trabalho está na articulação entre engenharia de dados, avaliação comparativa de modelos de linguagem e teste beta com usuários reais. A base analítica foi construída por coleta automatizada via API, integrando três *endpoints* complementares — contratos, itens de contrato e preços praticados —, seguida de tratamento, padronização de variáveis e modelagem em arquitetura medalhão. Sobre essa base, compararam-se dois modelos para geração de SQL: o Llama 3.1 8B, executado localmente, e o GPT-4o-mini, acessado via API. Por fim, o protótipo foi submetido a teste beta não supervisionado, com registro sistemático das interações e da satisfação dos usuários. Os resultados oferecem evidências empíricas sobre a viabilidade técnica e a aceitação de sistemas Text-to-SQL em contextos de gestão pública.

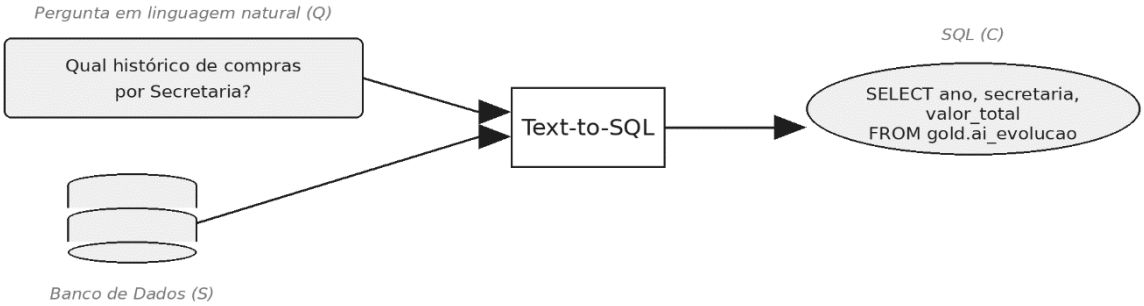
## REFERENCIAL TEÓRICO

### *1.1 Text-to-SQL e Interfaces em Linguagem Natural*

Sistemas Text-to-SQL têm sido utilizados para ampliar o acesso a bases estruturadas por meio da conversão automática de perguntas em linguagem natural em consultas SQL. Essa abordagem busca reduzir a dependência de conhecimento técnico especializado, permitindo que usuários não familiarizados com bancos de dados explorem informações estruturadas de forma mais direta (Katsogiannis-Meimarakis; Koutrika, 2023; Guo et al., 2019).

A tarefa Text-to-SQL pode ser definida como: dado um esquema de banco de dados S e uma pergunta em linguagem natural Q, gerar uma consulta SQL C tal que sua execução retorne a resposta correta para Q (Katsogiannis-Meimarakis; Koutrika, 2023; Guo et al., 2019). A Figura 1 sintetiza essa formulação clássica da tarefa Text-to-SQL.

**Figura 1 – Representação esquemática do Text-to-SQL**




**Fonte:** Elaboração própria, com base em Katsogiannis-Meimarakis e Koutrika (2023).

A evolução do campo é medida por meio de *benchmarks*, isto é, conjuntos padronizados de pares pergunta —SQL utilizados para treinar e comparar modelos. O Seq2SQL (Zhong et al., 2017) inaugurou essa tradição ao propor o primeiro modelo neural de grande escala para a tarefa, acompanhado do WikiSQL, dataset com 80 mil pares de pergunta e consulta. Embora relevante, o WikiSQL é limitado a consultas simples sobre tabelas únicas. O Spider (Yu et al., 2018) avança sobre essas limitações ao reunir 200 bancos de dados em 138 domínios distintos, com consultas complexas envolvendo múltiplas tabelas, tornando-se a referência dominante na avaliação de sistemas Text-to-SQL. Nesse *benchmark*, a acurácia dos melhores modelos evoluiu de 12%, em 2018, para mais de 86%, em 2023, impulsionada pelo advento dos modelos de linguagem pré-treinados, tema aprofundado na seção seguinte.

## *1.2 Large Language Models e Seleção de Modelos*

Large Language Models (LLMs) são modelos de redes neurais baseados na arquitetura Transformer (Vaswani et al., 2017), treinados em grandes volumes de texto e capazes de executar tarefas de geração, interpretação e transformação de linguagem. Em sistemas Text-to-SQL, a escolha do modelo constitui um fator crítico de desempenho, pois influencia a capacidade de compreender a pergunta do usuário, mapear sua intenção ao esquema da base e gerar consultas SQL sintaticamente válidas e semanticamente aderentes. Ao mesmo tempo, essa escolha envolve *trade-offs* relacionados a custo, latência, requisitos de infraestrutura e privacidade dos dados (Gao et al., 2024).

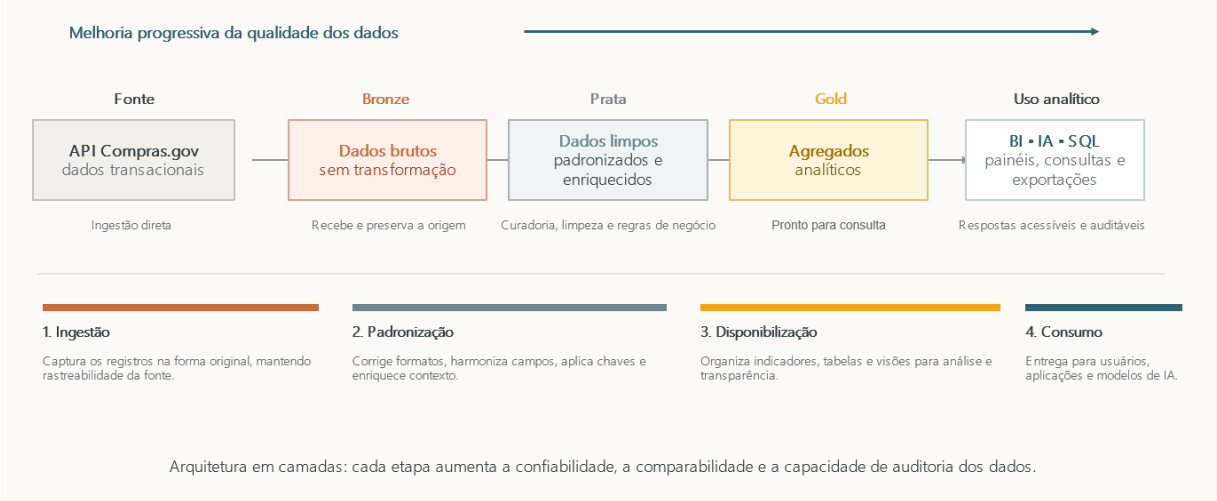
Neste estudo, foram comparados dois modelos com características operacionais distintas. O Llama 3.1 8B (Meta AI, 2024) é um modelo de código aberto com 8 bilhões de parâmetros, passível de execução local, o que o torna relevante para contextos institucionais com restrições quanto ao envio de dados a serviços externos. Já o GPT-4o-mini (OpenAI, 2024) é um modelo proprietário acessado via API, incluído no estudo como alternativa de maior capacidade para geração de consultas em linguagem natural, com custo inferior ao de modelos mais robustos da mesma família. A comparação entre ambos foi realizada empiricamente, com o objetivo de verificar o impacto da escolha do modelo sobre o desempenho do sistema proposto.

## 1.3 Arquitetura Medalhão

Conforme proposto pela Databricks, a Arquitetura Medalhão organiza dados analíticos em três camadas incrementais. A camada Bronze armazena os dados brutos exatamente como extraídos da fonte, sem transformações, preservando o histórico completo e permitindo reprocessamento. A camada Silver aplica limpeza, padronização e enriquecimento, produzindo dados consistentes e confiáveis. A camada Gold disponibiliza visões analíticas otimizadas para consumo, organizadas segundo as necessidades dos usuários finais.

Neste trabalho, a camada Gold foi implementada como um conjunto de *views* semânticas sobre os dados de compras de medicamentos do Compras.gov.br, oferecendo ao modelo de linguagem uma interface estruturada que reduz o espaço de busca e o risco de erros na geração de SQL.

**Figura 2 – Arquitetura medalhão para organização e qualificação dos dados**




**Fonte:** Elaboração própria, adaptado de Databricks.

## 1.4 Dados abertos governamentais

O portal Compras.gov.br, gerenciado pelo Ministério da Gestão e da Inovação em Serviços Públicos, disponibiliza dados públicos sobre contratações da administração pública federal, incluindo informações sobre contratos, itens contratados, fornecedores e preços praticados. Essa abertura amplia as possibilidades de transparência, controle social e reaproveitamento analítico das informações produzidas pelo próprio Estado. No entanto, a mera disponibilização dos dados não assegura sua apropriação efetiva, uma vez que seu uso analítico depende, em muitos casos, de competências técnicas relacionadas à coleta via API, tratamento, integração entre bases e modelagem para consulta.

Nesse contexto, o uso desses dados pelo próprio órgão produtor insere-se na agenda de *government analytics*, entendida como a aplicação sistemática de dados administrativos e ferramentas analíticas para apoiar decisões públicas, aperfeiçoar processos e ampliar a capacidade estatal de monitoramento (Rogger; Schuster, 2023). Em vez de serem vistos apenas como instrumentos de transparência externa, os dados abertos podem também funcionar como ativos estratégicos de gestão, desde que convertidos em estruturas analíticas capazes de responder a perguntas concretas de gestores e técnicos (World Bank, 2019). Transformar dados abertos em inteligência pública requer, portanto, mais do

que acesso aos registros originais: exige processos de padronização, integração semântica e disponibilização por meio de interfaces que reduzam barreiras técnicas e ampliem o potencial de uso das informações. É nessa perspectiva que se insere o presente trabalho, ao propor uma ferramenta de consulta em linguagem natural sobre dados públicos de aquisições de medicamentos no Ministério da Saúde.

## 2. METODOLOGIA

Este trabalho caracteriza-se como pesquisa aplicada de natureza quali-quantitativa, com abordagem de *design science research* (Hevner et al., 2004). O desenvolvimento adotou ciclos iterativos de implementação, teste e refinamento.

### 2.1 Dados e Infraestrutura

Os dados foram obtidos exclusivamente da API pública do Compras.gov.br, abrangendo o período de 2021 a 2025, com filtro pelo código (SIORG) do MS (36000), unidade orçamentária 250005 e classe CATMAT 6505 (Drogas e Medicamentos). Três conjuntos foram coletados: (1) EP1 — Contratos: 2.402 registros após filtro de categoria “Compras”; (2) EP2 — Itens de Contrato: 3.435 itens válidos; e (3) Preços Praticados: 155.534 registros de preços de referência das unidades federativas.

O banco de dados foi estruturado em PostgreSQL 15 (containerizado via Docker) segundo a Arquitetura Medalhão em três *schemas*: raw (Bronze), silver (Silver) e gold (Gold). A camada Gold é composta por 34 visões analíticas (ai_*) organizadas em nove blocos temáticos: visão geral, temporal, secretarias, medicamentos, fornecedores, marcas, modalidades, comparação MS versus UFs e judicialização. A interface do protótipo foi desenvolvida com Streamlit.

### 2.2 Mecanismo de Geração de SQL

O núcleo do sistema é um módulo de geração de SQL baseado em LLM. O *prompt* foi estruturado com três componentes: (1) uma instrução de sistema, que define o papel do modelo; (2) a descrição das 34 visões da camada Gold, incluindo nomes, colunas e descrições; e (3) um conjunto de mais de 40 exemplos de pares pergunta-SQL. A resposta gerada é submetida a um *parser*, que extrai a consulta SQL, verifica a referência ao schema gold e executa a consulta no PostgreSQL via Docker.

### 2.3 Avaliação Comparativa de Modelos

A avaliação técnica utilizou um conjunto de 18 perguntas de referência, distribuídas em três níveis de complexidade: simples (uma visão e um filtro direto), intermediárias (filtros combinados com agrupamento) e complexas (cruzamento de entidades ou séries temporais múltiplas). O mesmo conjunto de perguntas foi aplicado aos dois modelos avaliados (Llama 3.1 8B local e GPT-4o-mini via API), registrando-se, em cada caso, o grau de aderência da resposta produzida pelo sistema em relação à resposta esperada. Para isso, os resultados foram classificados como “Sim”, quando a resposta foi satisfatória; “Parcial”, quando atendeu apenas em parte ao esperado; e “Não”, quando o sistema não respondeu adequadamente à pergunta. O detalhamento encontra-se no Apêndice A.

## 2.4 Teste Beta com usuários e formulário de satisfação

Após a seleção do GPT-4o-mini, o teste beta foi conduzido com usuários reais, com acesso público via link ngrok. Todas as interações foram automaticamente registradas em logs. interacoes (PostgreSQL), com campos: pergunta, SQL gerado, sucesso, erro, tempo de resposta e quantidade de linhas retornadas. Ao final de cada sessão, os usuários tinham a oportunidade de responder a um formulário de satisfação (logs.avaliacoes) com seis dimensões: correspondência do resultado, número de tentativas, dificuldade de formulação, utilidade do gráfico, intenção de uso no trabalho e substituição de intermediação técnica. A análise estatística utilizou o teste qui-quadrado (associação entre tamanho da pergunta e sucesso) e o teste de Mann-Whitney (diferença no tempo de resposta entre sucessos e falhas).

**Figura 3 – Fluxo metodológico de desenvolvimento do protótipo**




**Fonte:** Elaboração própria.

# 3. RESULTADOS E DISCUSSÃO

## 3.1 Avaliação Comparativa de Modelos

A Tabela 1 apresenta os resultados da avaliação com 18 perguntas de referência, distribuídas igualmente entre os três níveis de complexidade, nas duas configurações de modelo testadas. O conjunto completo de perguntas e resultados individuais está disponível no Apêndice A.

**Tabela 1 — Acurácia por nível de complexidade: Llama 3.1 8B vs GPT-4o-mini**

| Nível de Complexidade | Llama 3.1 8B (local) | GPT-4o-mini (API) |
|---|---|---|
| **Simples** | 50,0% (3/6) | 66,7% (4/6) |
| **Intermediário** | 33,3% (2/6) | 100,0% (6/6) |
| **Complexo** | 0,0% (0/6) | 33,3% (2/6) |
| **TOTAL** | **27,8% (5/18)** | **72,2% (13/18)** |

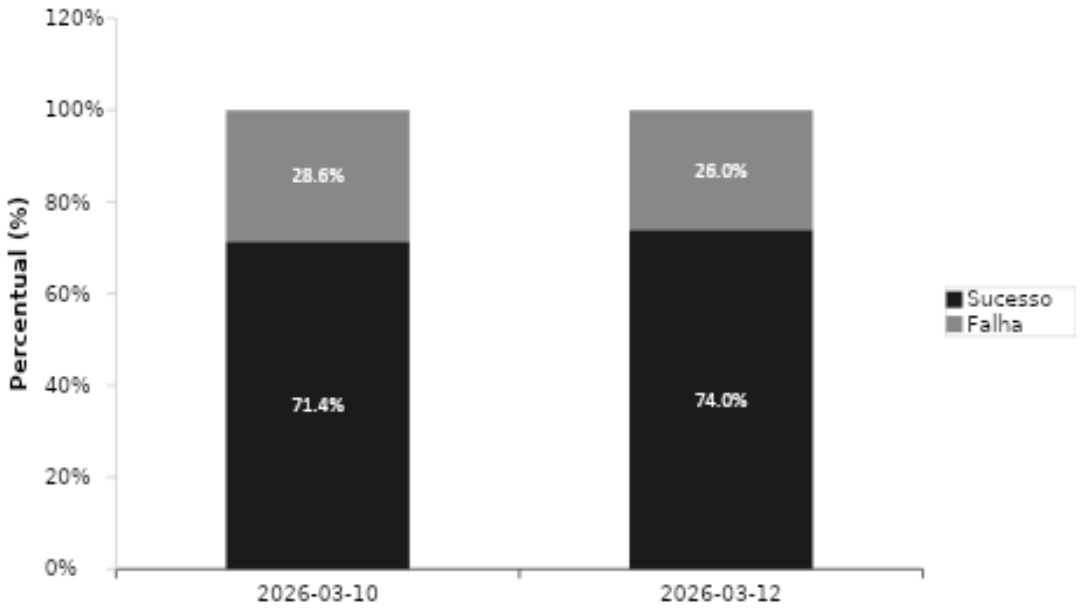
**Fonte:** Elaboração própria.

A migração do Llama 3.1 8B para o GPT-4o-mini resultou em ganho de 44,4 pontos percentuais na acurácia geral (de 27,8% para 72,2%). O salto mais expressivo ocorreu nas perguntas intermediárias, que atingiram 100% de acerto com o GPT-4o-mini. Os principais problemas identificados no Llama 3.1 8B foram: *timeouts* frequentes (tempo superior a 300 segundos), alucinação de nomes de visões inexistentes na camada Gold e truncamento de SQL por limite de tokens no contexto. O GPT-4o-mini eliminou os *timeouts* e reduziu drasticamente as alucinações estruturais; ainda assim, o desempenho em perguntas complexas (33,3%) sugere limitações persistentes em consultas que demandam raciocínio sobre múltiplas dimensões não diretamente representadas no contexto fornecido ao modelo.

## *3.2 Teste Beta - Log de Interações*

O teste beta registrou 172 interações válidas distribuídas em dois dias: 49 no dia 10/03 e 123 no dia 12/03. A taxa de sucesso geral foi de 73,3% (126/172), com leve melhora entre os dias (71,4% → 74,0%) (Figura 4). O tempo mediano de resposta foi de 2,2 segundos e o tempo médio de 2,5 segundos.

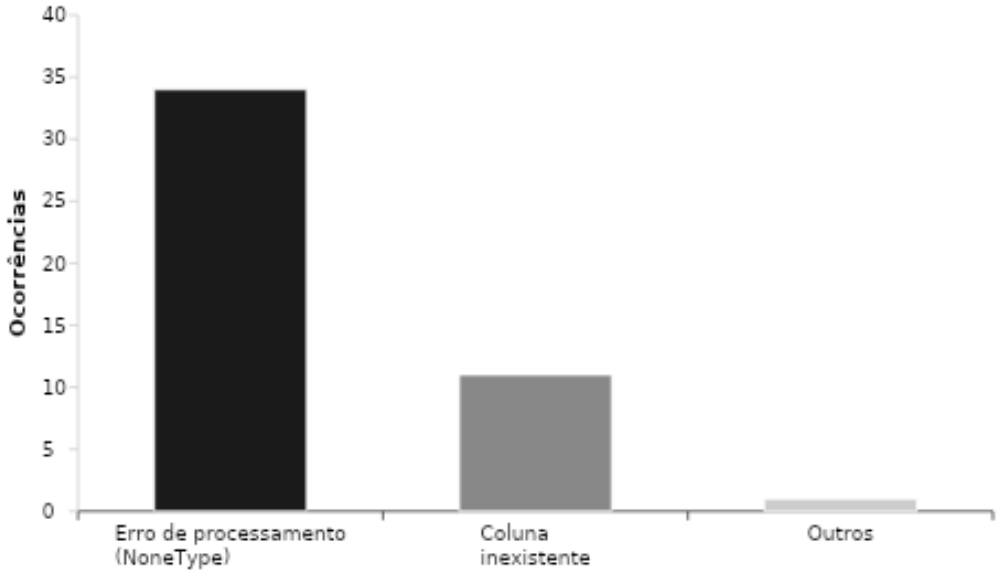
**Figura 4 – Distribuição percentual de sucesso e falha por dia de teste (n = 172)**




**Fonte:** Elaboração própria.

A análise dos 46 erros identificados no teste revelou três categorias principais. A mais frequente foi um erro de processamento da resposta no sistema, registrado em 34 casos (73,9%), com a mensagem "*NoneType object has no attribute 'strip'*". Em seguida, observaram-se 11 casos (23,9%) de referência a colunas inexistentes nas visões analíticas. Houve ainda 1 caso (2,2%) de erro de sintaxe SQL. Esses resultados indicam que as falhas decorreram tanto de problemas de robustez na camada de processamento das respostas quanto de inconsistências entre a consulta gerada e o esquema disponível no banco (Figura 5).

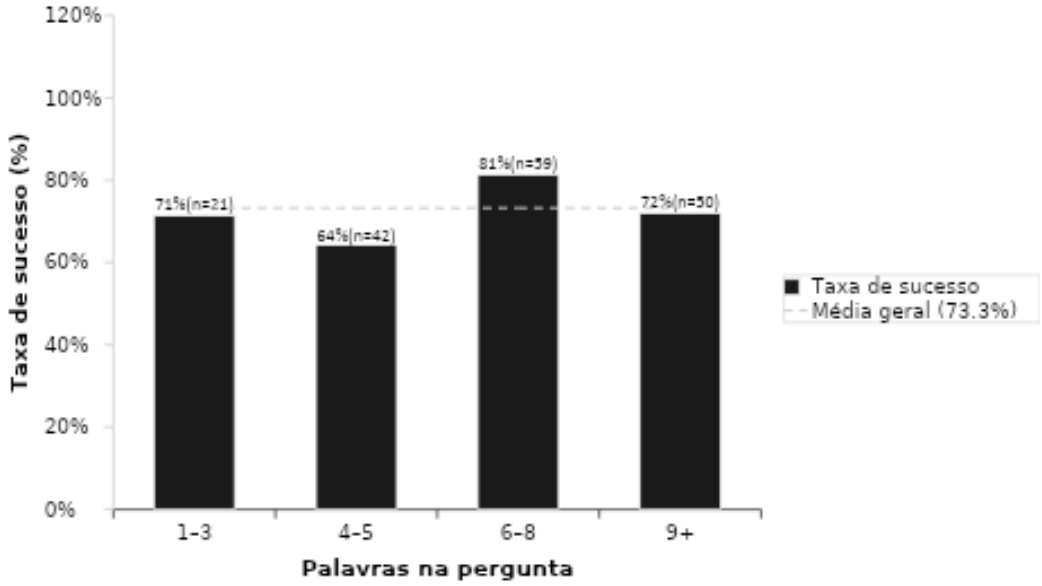
**Figura 5 – Distribuição das categorias de erro no teste beta (n = 46)**




**Fonte:** Elaboração própria.

A Figura 6 apresenta a taxa de sucesso observada em cada faixa de comprimento da pergunta, permitindo comparar descritivamente o desempenho do sistema.

**Figura 6 - Sucesso do sistema segundo o comprimento da pergunta**

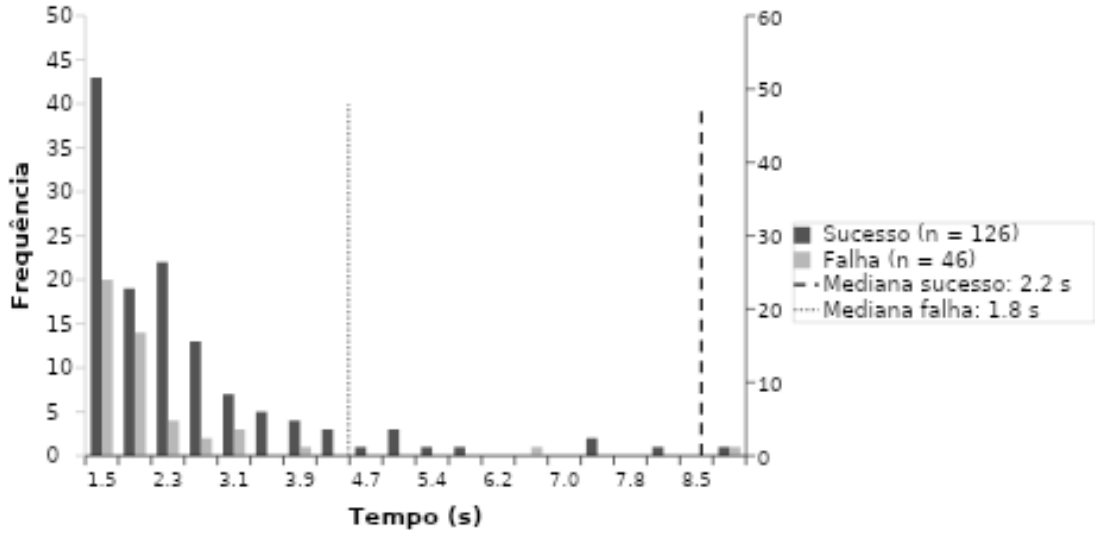



**Fonte:** Elaboração própria.

Embora a análise descritiva do comprimento das perguntas tenha mostrado que a faixa de 6 a 8 palavras apresentou a maior taxa observada de sucesso (81%), o teste qui-quadrado não indicou associação estatisticamente significativa entre tamanho da pergunta e sucesso ($x^2$ = 1,70; p = 0,19), o que recomenda cautela na interpretação desse resultado. O teste qui-quadrado entre tamanho da pergunta e sucesso não indicou associação estatisticamente significativa ($x^2$ = 1,70; p = 0,19), não havendo evidências, nesta amostra, de que variações no tamanho da pergunta estejam associadas ao sucesso do sistema.

O teste de Mann-Whitney aplicado ao tempo de resposta de consultas bem-sucedidas e malsucedidas também não alcançou significância estatística ao nível de 5% (U = 3449; p = 0,057), embora se observe tendência de falhas ocorrerem em menor tempo mediano (1,8 s) do que sucessos (2,2 s) (Figura 7). Esse padrão pode estar relacionado à interrupção precoce do processamento em parte das respostas malsucedidas.

**Figura 7 - Distribuição do tempo de resposta por resultado da interação**



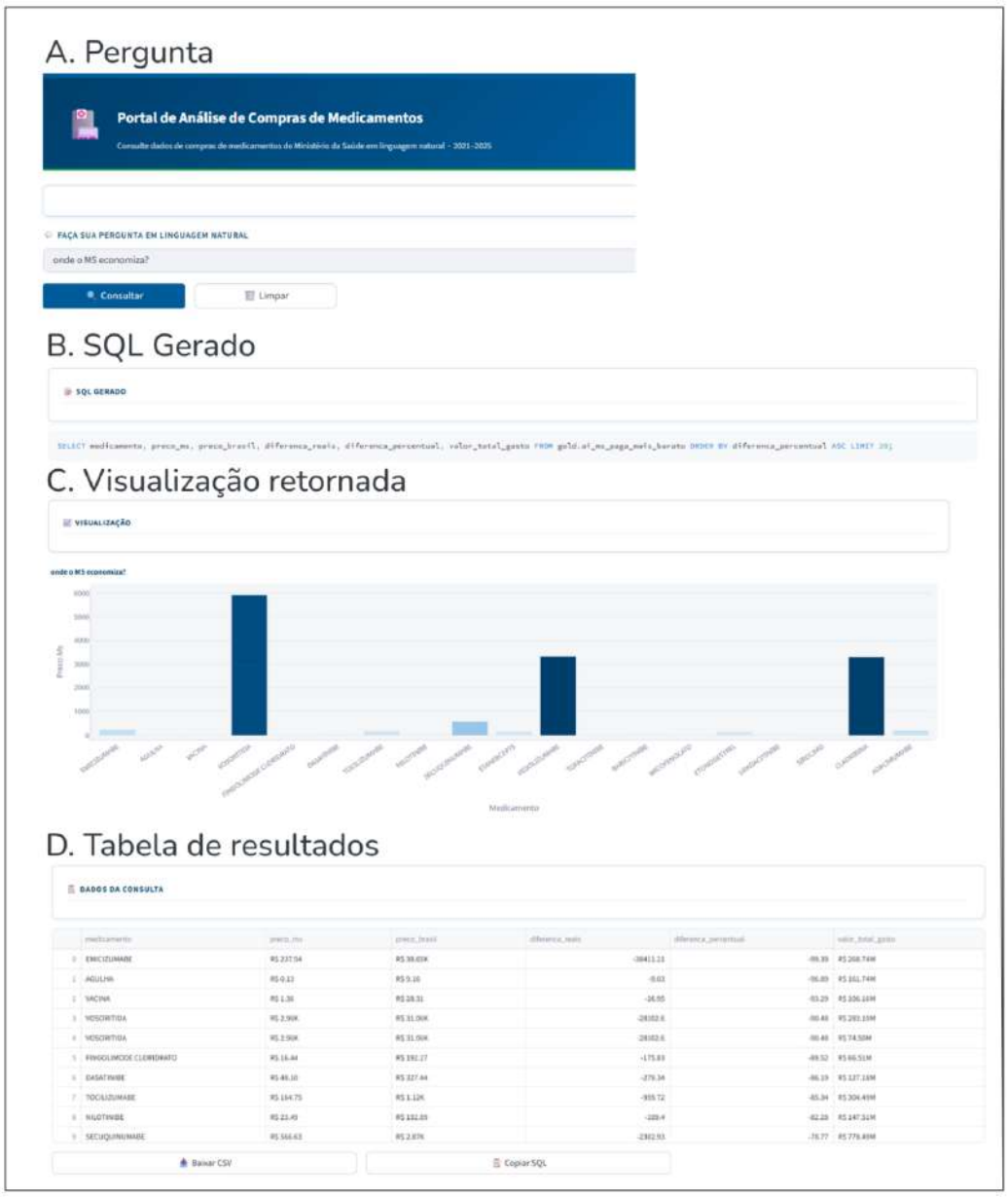

**Fonte:** Elaboração própria.

## 3.3 Teste Beta - Análise das Perguntas Espontâneas

As 172 interações do teste beta geraram 150 formulações distintas, após padronização de variações gráficas superficiais, como diferenças entre maiúsculas e minúsculas e espaços residuais, sem consolidação semântica entre perguntas de conteúdo semelhante. Desse total, 22 ocorrências adicionais distribuíram-se por 15 perguntas repetidas com a mesma formulação padronizada. Em nenhuma dessas repetições houve alteração do resultado entre tentativas: consultas que falharam inicialmente mantiveram a falha nas ocorrências subsequentes, enquanto consultas bem-sucedidas repetiram o sucesso de forma consistente. Esse padrão é compatível com a configuração adotada para geração das respostas, com temperature = 0,0, voltada à redução da variabilidade do modelo. Entre as perguntas repetidas com sucesso, destacaram-se “onde o MS economiza?” (5 ocorrências) e “qual secretaria do MS gasta mais?” (3 ocorrências), ambas disponibilizadas como exemplos na barra lateral da interface, o que sugere uso ativo das sugestões oferecidas pela ferramenta. Também entre as perguntas repetidas com falha, como “MS paga mais caro que outros órgãos?” e “medicamentos com fornecedor único”, observou-se recorrência de erro em todas as tentativas. Como essas consultas também constavam nas sugestões exibidas pela interface, é plausível que as repetições reflitam o uso dos exemplos por diferentes participantes, reforçando que tanto os acertos quanto as limitações do sistema foram sistematicamente reproduzidos durante o teste. A Figura 7 apresenta um exemplo de interação bem-sucedida no sistema, a partir da consulta “Onde o MS economiza?”, ilustrando o fluxo entre pergunta em linguagem natural, SQL gerado, visualização e tabela de resultados.

Para a análise qualitativa das perguntas únicas, as 150 formulações distintas, após exclusão de variações ortográficas, foram classificadas por nível de complexidade, seguindo os mesmos critérios adotados no conjunto de referência. Perguntas simples foram definidas como consultas diretas a uma única dimensão; perguntas intermediárias, como consultas com filtro por entidade específica, recorte temporal ou agregação localizada; e perguntas complexas, como consultas que exigiam comparação entre entes, múltiplas dimensões simultâneas ou medidas derivadas mais elaboradas. Nessa classificação, a complexidade foi definida com base apenas na estrutura da pergunta e nas variáveis disponíveis na própria base analítica, sem recurso a referências externas. A Tabela 2 apresenta a distribuição e a taxa de acerto por nível. O conjunto completo encontra-se no Apêndice B.

**Figura 7 - Exemplo de resposta do Portal de Análise de Compras de Medicamentos a uma consulta em linguagem natural.**




**Fonte:** Print do protótipo desenvolvido

**Tabela 2 — Acurácia por nível de complexidade: Teste Beta**

| Nível | Perguntas (n) | Sucessos | % Resposta |
|---|---|---|---|
| **Simples** | 48 | 40 | 83,3% |
| **Intermediário** | 74 | 34 | 71,6% |
| **Complexo** | 22 | 10 | 60,7% |
| **TOTAL** | **150** | **110** | **73,3%** |

**Fonte:** Elaboração própria

A taxa geral de sucesso sobre as formulações distintas padronizadas foi de 73,3% (110/150), valor idêntico, após arredondamento, à taxa bruta observada no conjunto total de interações (126/172 = 73,3%). Esse resultado sugere que as repetições não distorceram o desempenho global do sistema. Na classificação por nível de complexidade, as perguntas simples apresentaram taxa de sucesso de 83,3% (40/48), as intermediárias de 71,2% (52/73) e as complexas de 60,7% (17/28). Observa-se, assim, melhor desempenho em consultas de menor complexidade analítica, embora as perguntas complexas também tenham alcançado taxa de resposta relevante na base padronizada.

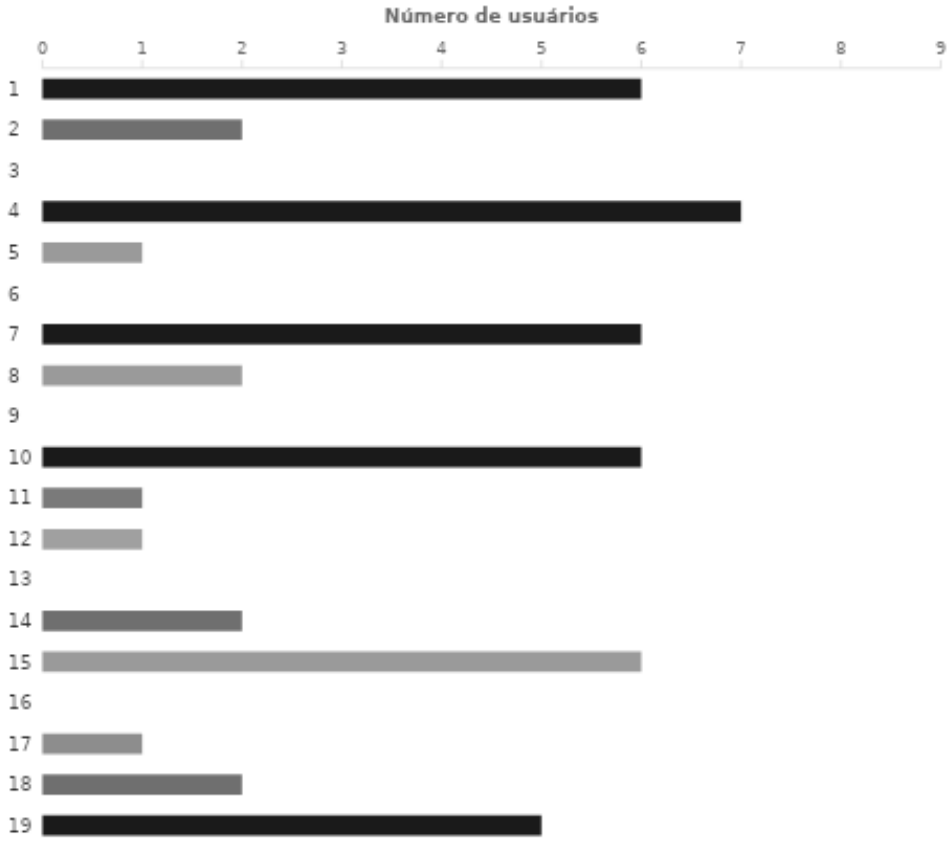
Merece destaque a presença de perguntas semanticamente não aderentes ao escopo da base de dados, como “quais as doenças mais raras detectadas em Sergipe?” e “quais medicamentos fazem mais efeitos para diabetes?”, para as quais o sistema retornou respostas baseadas nos dados disponíveis, sem sinalizar explicitamente a limitação de escopo. Esse comportamento evidencia uma oportunidade de melhoria: a implementação de um mecanismo de detecção de perguntas fora de domínio poderia reduzir respostas potencialmente enganosas e tornar o uso do sistema mais seguro para o usuário.

## *3.4 Teste Beta - Formulário de Satisfação*

O formulário de satisfação foi respondido por 8 usuários. Os resultados foram: 75% consideraram que a resposta correspondeu ao esperado; 25% avaliaram como parcialmente correta; 88% avaliaram o gráfico gerado como útil; 75% afirmaram que utilizariam a ferramenta no trabalho; e 75% indicaram que ela substituiria a necessidade de intermediação técnica. A média de tentativas por consulta foi de 1,5, com 62,5% dos usuários obtendo o resultado na primeira tentativa.

Os comentários qualitativos destacaram como oportunidades de melhoria: geração de respostas também em linguagem natural (além de tabelas e gráficos), seleção dinâmica de tipo de gráfico pelo modelo, e ampliação da cobertura para medicamentos com valores unitários e cruzamentos complexos entre fornecedores.

**Figura 8 - Avaliação dos usuários: Teste Beta (n=8)**




**Fonte:** Elaboração própria.

Os resultados indicam viabilidade técnica e aceitação prática do sistema no contexto avaliado. No teste beta não supervisionado, foram registradas 172 interações, com taxa geral de sucesso de 73,3% e tempo mediano de resposta de 2,2 segundos. A estabilidade desse desempenho também se refletiu na análise das formulações repetidas, nas quais não houve alteração de resultado entre tentativas com a mesma pergunta padronizada. Em termos práticos, esse comportamento sugere consistência do sistema em condições reais de uso.

A comparação entre modelos oferece uma contribuição metodológica relevante para o contexto da gestão pública. Modelos locais menores, como o Llama 3.1 8B, embora atrativos do ponto de vista da soberania dos dados e da ausência de custos diretos de API, apresentaram acurácia insuficiente (27,8%) para uso em produção no domínio avaliado, com problemas recorrentes de *timeout*, alucinação de esquemas e truncamento de SQL. Por outro lado, o GPT-4o-mini demonstrou que modelos acessados via API podem alcançar qualidade satisfatória com custo operacional relativamente reduzido. Esse achado é particularmente relevante para órgãos públicos que enfrentam o dilema entre controle sobre os dados e desempenho do modelo. Nesses contextos, além da escolha entre modelos locais e APIs públicas, pode-se considerar o uso de soluções corporativas já contratadas institucionalmente, quando disponíveis, como forma de conciliar requisitos de privacidade, governança e viabilidade econômica. Em qualquer cenário, a adoção de modelos externos requer avaliação institucional prévia quanto às políticas de proteção de dados, às condições contratuais e à sensibilidade das consultas realizadas.

A arquitetura baseada em camada Gold com visões analíticas pré-construídas foi central para esse desempenho. Ao oferecer ao modelo uma interface semântica mais estável, com nomes descritivos, colunas claramente definidas e escopo analítico delimitado, reduziu-se o espaço de busca e a incidência de alucinações estruturais. Esse resultado é coerente com a literatura que destaca a importância da qualidade do esquema fornecido ao modelo para tarefas Text-to-SQL (Gao et al., 2024). Ao mesmo tempo, essa decisão de projeto impôs um limite objetivo à cobertura do sistema: perguntas que demandavam cruzamentos não previstos nas visões analíticas tenderam a falhar não apenas por limitação do modelo de linguagem, mas também por ausência de estrutura analítica correspondente na camada de dados.

A análise das perguntas espontâneas também mostrou que o sistema tende a responder melhor a consultas mais diretas e semanticamente aderentes à base. Em contrapartida, perguntas fora do escopo informacional da base ou formuladas de modo semanticamente distante da estrutura disponível puderam receber respostas formalmente válidas, mas inadequadas ao conteúdo solicitado. Esse comportamento reforça a necessidade de incorporar mecanismos de verificação de escopo antes da geração da consulta SQL, de modo a reduzir respostas potencialmente enganosas e tornar o uso do sistema mais seguro para usuários não técnicos.

As limitações observadas concentraram-se em três frentes. A primeira foi o desempenho inferior em perguntas mais complexas, especialmente aquelas que exigiam comparação entre entes, múltiplas dimensões simultâneas ou medidas derivadas mais elaboradas. A segunda foi a robustez do pipeline de processamento da resposta, uma vez que parte relevante das falhas decorreu de erros internos de tratamento da saída gerada pelo modelo e de inconsistências entre a consulta produzida e o esquema disponível. A terceira foi a ausência de um mecanismo explícito de detecção de perguntas fora de domínio, o que permitiu ao sistema retornar respostas formalmente válidas, porém semanticamente inadequadas, em consultas não aderentes ao escopo da base.

## CONCLUSÃO

Este artigo apresentou o desenvolvimento e a avaliação de um sistema Text-to-SQL para consulta, em linguagem natural, de dados de compras de medicamentos do Ministério da Saúde. Os resultados indicam que a abordagem é tecnicamente viável e apresenta utilidade prática no contexto da gestão pública, ao permitir que usuários sem domínio de SQL formulem perguntas analíticas e obtenham respostas estruturadas a partir de dados reais de aquisição de medicamentos.

Do ponto de vista técnico, o estudo mostrou que a escolha do modelo de linguagem afeta de forma decisiva o desempenho do sistema. A substituição do Llama 3.1 8B local pelo GPT-4o-mini elevou a acurácia no conjunto de referência de 27,8% para 72,2%, evidenciando a relevância da seleção do modelo em tarefas Text-to-SQL aplicadas a domínios especializados. Também se mostrou central a adoção de uma arquitetura baseada em visões analíticas pré-construídas na camada Gold, que funcionou como mediação entre os dados brutos e o modelo de linguagem, favorecendo maior consistência semântica e melhor desempenho nas consultas simples e intermediárias.

No teste beta não supervisionado, o sistema registrou 172 interações, com taxa geral de sucesso de 73,3% e tempo mediano de resposta de 2,2 segundos, o que sugere desempenho compatível com uso cotidiano. Além disso, a avaliação de satisfação indicou aceitação favorável da ferramenta: 75% dos respondentes afirmaram que a utilizariam em seu trabalho, e 75% consideraram que ela poderia reduzir a necessidade de intermediação técnica no acesso analítico aos dados. Esses achados reforçam o potencial da solução como instrumento de democratização do uso de dados públicos no interior da administração pública.

Como principal contribuição, o estudo demonstra que a combinação entre engenharia de dados, modelagem analítica em arquitetura medalhão e interface em linguagem natural pode ampliar o consumo analítico de dados públicos pelo próprio órgão produtor da informação. Em outras palavras, o trabalho mostra que dados abertos podem deixar de cumprir apenas função de transparência externa e passar a operar também como infraestrutura de inteligência para apoio à decisão.

Persistem, contudo, limitações relevantes, especialmente no tratamento de perguntas complexas, na robustez do pipeline de processamento das respostas e na ausência de mecanismos explícitos de detecção de perguntas fora de domínio. Como agenda futura, destacam-se: o aprimoramento do pipeline de ingestão, com atualização periódica e automatizada da base analítica a partir da API do Compras.gov.br; a integração com outras fontes de dados relevantes à assistência farmacêutica, ampliando o escopo informacional disponível para consulta; e a avaliação de abordagens de geração de SQL diretamente sobre o esquema relacional, sem dependência das visões analíticas pré-construídas da camada Gold, decisão de projeto que, neste trabalho, mostrou-se central para reduzir alucinações e garantir consistência semântica, mas que impõe limites à cobertura das consultas; e a realização de novo teste com usuários, a fim de avaliar o impacto dessas evoluções sobre o desempenho do sistema e a experiência dos gestores.

## Disponibilidade de Dados e Código

Os dados utilizados neste trabalho estão disponíveis em repositório público com identificador persistente: Gonçalves Gonçalves Pinho, Rosana Elisa. Portal de Análise de Compras de Medicamentos do Ministério da Saúde: dados de suporte. Zenodo, 2026. DOI: https://doi.org/10.5281/zenodo.20997853. O código-fonte do sistema — incluindo os *scripts* Python, SQL de configuração e carga da base de dados, e as visões analíticas da camada Gold — está disponível em: https://github.com/rosanagg/portal-analise-medicamentos-ms.

## Declaração de Uso de Inteligência Artificial Generativa

O uso de inteligência artificial generativa foi declarado conforme as diretrizes institucionais. Ferramentas de IA foram utilizadas nas seguintes etapas: (i) refatoração e revisão de código-fonte (scripts Python e SQL); e (ii) revisão textual do manuscrito. A responsabilidade pelo conteúdo final, interpretações e conclusões é integralmente da autora.

## REFERÊNCIAS


BOSIO, E.; DJANKOV, S.; GLAESER, E. L.; SHLEIFER, A. Public procurement in law and practice. **American Economic Review**, 112(4), 1091–1117, 2022.

BRASIL. Ministério da Gestão e da Inovação em Serviços Públicos. **Compras públicas em dados abertos**. Brasília, DF: MGI, [s.d.]. Disponível em: https://www.gov.br/compras/pt-br/cidadao/compras-publicas-dados-abertos. Acesso em: 21 mar. 2026.

DATABRICKS. **What is the medallion lakehouse architecture**? [s.d.]. Disponível em: https://www.databricks.com/blog/what-is-medallion-architecture. Acesso em: 21 mar. 2026.

GAO, D.; WANG, H.; LI, Y.; SUN, X.; QIAN, Y.; DING, B. et al. Text-to-SQL empowered by large language models: a benchmark evaluation. **Proceedings of the VLDB Endowment**, 17(5), 1132–1145, 2024.

GUO, J.; ZHAN, Z.; GAO, Y.; XIAO, Y.; LOU, J. G.; LIU, T. et al. Towards complex text-to-SQL in cross-domain database with intermediate representation. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 57., 2019, Florence. **Proceedings**, Florence: ACL, 2019. p. 4524–4535.

HEVNER, A. R.; MARCH, S. T.; PARK, J.; RAM, S. Design science in information systems research. **MIS Quarterly**, 28(1), 75–105, 2004.

KATSOGIANNIS-MEIMARAKIS, G.; KOUTRIKA, G. A survey on deep learning approaches for text-to-SQL. **The VLDB Journal**, 32, 905–936, 2023.

META AI. Introducing Llama 3.1: our most capable models to date. 2024. Disponível em: https://ai.meta.com/blog/meta-llama-3-1/. Acesso em: 21 mar. 2026.

OCDE. Public procurement performance: a framework for measuring efficiency, compliance and strategic goals. **OECD Public Governance Policy Papers**, (36), 2023.

OPENAI. GPT-4o mini: advancing cost-efficient intelligence. 2024. Disponível em: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Acesso em: 21 mar. 2026.

ROGGER, D.; SCHUSTER, C. (ed.). **The Government Analytics Handbook: Leveraging Data to Strengthen Public Administration**. Washington, DC: World Bank, 2023. Disponível em: https://thedocs.worldbank.org/en/doc/6e1314a7c5ee766fdd531f1aa64324e5-0050042023/original/GAH-Part-1.pdf. Acesso em: 21 mar. 2026.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N. et al. Attention is all you need. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 30., 2017, Long Beach. **Proceedings,** Long Beach: NeurIPS, 2017. p. 6000–6010.

WORLD BANK. **Starting an Open Data Initiative**. Washington, DC: World Bank, 2019. Disponível em: https://opendatatoolkit.worldbank.org/en/data/opendatatoolkit/starting. Acesso em: 21 mar. 2026.

YU, T.; ZHANG, R.; YANG, K.; YASUNAGA, M.; WANG, D.; LI, Z. et al. Spider: a large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2018, Brussels. **Proceedings,** Brussels: ACL, 2018. p. 3911–3921.

ZHONG, V.; XIONG, C.; SOCHER, R. Seq2SQL: generating structured queries from natural language using reinforcement learning. **arXiv**, 2017. Disponível em: https://arxiv.org/abs/1709.00103. Acesso em: 21 mar. 2026.

## APÊNDICE A — CONJUNTO DE REFERÊNCIA PARA AVALIAÇÃO DO SISTEMA TEXT-TO-SQL

As 18 perguntas, a seguir, compõem o conjunto de referência utilizado para avaliação comparativa entre os modelos Llama 3.1 8B (local) e GPT-4o-mini (API). As perguntas foram elaboradas de forma a cobrir os três níveis de complexidade: simples, intermediário e complexo, em proporções iguais (6 perguntas por nível). O resultado indica se a consulta gerada produziu o retorno esperado (Sim), não produziu (Não) ou produziu parcialmente (Parcial).

| Nº | Nível | Pergunta | Llama 3.1 8B | GPT-4o-mini |
|---|---|---|---|---|
| 1 | Simples | Qual o valor total gasto pelo Ministério da Saúde em medicamentos? | **Não** | **Sim** |
| 2 | Simples | Quais são os 10 medicamentos mais comprados por valor? | **Parcial** | **Sim** |
| 3 | Simples | Quais são os 10 fornecedores com maior valor contratado? | **Sim** | **Sim** |
| 4 | Simples | Quantos contratos foram firmados por ano? | **Não** | **Não** |
| 5 | Simples | Qual a distribuição de compras por modalidade (pregão, dispensa, inexigibilidade)? | **Sim** | **Sim** |
| 6 | Simples | Quais secretarias gastaram mais com medicamentos? | **Sim** | **Sim** |
| 7 | Intermediária | Quais os 10 fornecedores que mais vendem para a SCTIE? | **Sim** | **Sim** |
| 8 | Intermediária | Como evoluiu o gasto com medicamentos por ano entre 2021 e 2025? | **Não** | **Sim** |
| 9 | Intermediária | Quais medicamentos têm fornecedor único no Ministério da Saúde? | **Não** | **Sim** |
| 10 | Intermediária | Qual a participação percentual de cada modalidade de compra por secretaria? | **Não** | **Sim** |
| 11 | Intermediária | Quais os 10 medicamentos mais comprados pela SAPS? | **Não** | **Sim** |
| 12 | Intermediária | Como evoluiu o preço médio da insulina ao longo dos anos? | **Sim** | **Sim** |
| 13 | Complexa | O Ministério da Saúde paga mais caro que os estados para quais medicamentos? | **Não** | **Não** |
| 14 | Complexa | Compare a evolução anual do preço mediano da insulina comprada pelo MS versus o preço praticado pelas demais UFs entre 2021 e 2025 | **Parcial** | **Sim** |
| 15 | Complexa | Quais medicamentos têm maior concentração de mercado (poucos fornecedores dominam)? | **Não** | **Não** |
| 16 | Complexa | Como evoluiu a judicialização de medicamentos por ano em comparação com as compras programáticas? | **Parcial** | **Sim** |
| 17 | Complexa | Quais medicamentos tiveram maior variação de preço entre contratos? | **Não** | **Não** |
| 18 | Complexa | Como se distribuem as modalidades de compra da SCTIE ao longo dos anos? | **Não** | **Sim** |

# APÊNDICE B

Este apêndice apresenta a classificação das 150 questões únicas formuladas pelos usuários durante o Teste Beta, organizadas por nível de complexidade. A classificação seguiu os mesmos critérios do conjunto de referência (Apêndice A).

| Pergunta | Ocorrências | Categoria | Gerou resultado |
|---|---|---|---|
| Resumo | 4 | Simples | Sim |
| Qual secretaria do MS gasta mais? | 3 | Simples | Sim |
| Quais os dez medicamentos mais caros em valor unitário? | 2 | Simples | Não |
| Top 5 fornecedores | 2 | Simples | Sim |
| Quais são os top 10 medicamentos em valor? | 2 | Simples | Sim |
| Top 10 fornecedores | 2 | Simples | Sim |
| Qual o histórico de compra por secretaria? | 1 | Simples | Sim |
| Histórico de valor total por ano | 1 | Simples | Sim |
| Quais os dez medicamentos mais caros em valor unitário? | 1 | Simples | Não |
| Quanto foi gasto pelo Ministério da Saúde em compra de medicamentos em 2020? | 1 | Simples | Sim |
| Gastos com medicamentos por ano | 1 | Simples | Sim |
| Resumo geral da base | 1 | Simples | Sim |
| Quais os dez medicamentos mais caros? | 1 | Simples | Sim |
| Quais os medicamentos mais caros em valor unitário? | 1 | Simples | Não |
| Quais as dimensões e medidas da base de dados? | 1 | Simples | Sim |
| Há dados de quais unidades federativas? | 1 | Simples | Sim |
| Quais os 5 medicamentos mais caros? | 2 | Simples | Sim |
| Valor total de gastos mensal | 1 | Simples | Sim |
| Gráfico top 3 fornecedores valor total | 1 | Simples | Sim |
| Gráfico pizza fornecedores valor total | 1 | Simples | Sim |
| Gráfico top 3 fornecedores valor total percentual | 1 | Simples | Sim |
| Gráfico pizza top 3 fornecedores valor total | 1 | Simples | Sim |
| Quantidade comprada por fornecedor | 1 | Simples | Sim |
| Gastos por secretaria | 1 | Simples | Sim |

| Pergunta | Ocorrências | Categoria | Gerou resultado |
|---|---|---|---|
| Quais as secretarias? | 1 | Simples | Sim |
| Qual o medicamento mais caro? | 1 | Simples | Sim |
| Quais são os top 10 medicamentos em valor unitário? | 1 | Simples | Não |
| Quais são os top 10 medicamentos em valor unitário? | 1 | Simples | Não |
| Quais são os top 10 medicamentos em quantidade? | 1 | Simples | Sim |
| Quais são os top 10 medicamentos em valor? percentual | 1 | Simples | Sim |
| Gastos anuais | 1 | Simples | Sim |
| Histórico de compras por secretaria | 1 | Simples | Sim |
| Qual modalidade de compra do eculizumabe? | 1 | Simples | Sim |
| Qual modalidade de compra do abacavir? | 1 | Simples | Sim |
| Qual modalidade de compra da insulina? | 1 | Simples | Sim |
| Quais os medicamentos mais judicializados? | 1 | Simples | Não |
| Quais os medicamentos mais judicializados em 2024? | 1 | Simples | Não |
| Gastos da SCTIE por ano | 1 | Simples | Sim |
| Quanto o Ministério da Saúde gastou em 2024? | 1 | Simples | Sim |
| Quantos medicamentos foram adquiridos em 2025? | 1 | Simples | Sim |
| Quantos medicamentos diferentes foram adquiridos em 2025? | 1 | Simples | Sim |
| Qual secretaria gasta mais? | 1 | Simples | Sim |
| Dez medicamentos mais caros | 1 | Simples | Sim |
| Medicamentos mais judicializados | 1 | Simples | Não |
| Volume total de judicializações por ano | 1 | Simples | Sim |
| Top 10 medicamentos mais judicializados | 1 | Simples | Sim |
| Top 10 medicamentos mais judicializados em quantidade | 1 | Simples | Sim |
| Top 10 medicamentos mais judicializados em valor | 1 | Simples | Sim |
| Qual o top 5 medicamentos adquiridos pela Secretaria executiva em 2021? | 2 | Intermediária | Sim |
| Quanto o MS gastou com atalureno entre 2021 e 2024? | 2 | Intermediária | Sim |
| Quanto foi gasto entre 2022-2023 com medicamento deferasirox? | 2 | Intermediária | Sim |
| Medicamentos com fornecedor único | 2 | Intermediária | Não |

| Pergunta | Ocorrências | Categoria | Gerou resultado |
|---|---|---|---|
| Fornecedor multicare | 2 | Intermediária | Não |
| Quais os fornecedores que mais aparecem para o atalureno? | 1 | Intermediária | Sim |
| Quanto o MS gastou com atalureno 1000mg entre 2021 e 2024? | 1 | Intermediária | Sim |
| Quanto o MS gastou com amoxicilina entre 2021 e 2024? | 1 | Intermediária | Sim |
| Quanto o MS gastou com amoxicilina entre 2022 e 2024? | 1 | Intermediária | Sim |
| Quais os fornecedores que mais aparecem para o medicamento atalureno? | 1 | Intermediária | Sim |
| Quanto o MS gastou com penicilina em 2023? | 1 | Intermediária | Não |
| Quanto o MS gastou com amoxicilina em 2023? | 1 | Intermediária | Sim |
| Quanto o MS gastou com penicilina nos últimos 5 anos? | 1 | Intermediária | Não |
| Quanto o MS gastou com penicilina nos últimos 5 anos? | 1 | Intermediária | Sim |
| Quantidade comprada de etanercepte em 2025 | 1 | Intermediária | Sim |
| Quanto Alagoas gastou em 2023? | 1 | Intermediária | Não |
| Quanto Sergipe gastou em 2023? | 1 | Intermediária | Não |
| Qual o top 5 de medicamentos adquiridos pela Secretaria executiva? | 1 | Intermediária | Sim |
| Quanto de Adalimumabe foi adquirido pelo MS entre 2023 e 2025 | 1 | Intermediária | Sim |
| Golimumabe | 1 | Intermediária | Sim |
| Qual o mês em que o Sergipe gastou mais? | 1 | Intermediária | Sim |
| Qual a quantidade anual de dolutegravir comprada pelo Ministério da Saúde em 2020 ? | 1 | Intermediária | Sim |
| Qual a quantidade anual de dolutegravir comprada pelo Ministério da Saúde em 2021? | 1 | Intermediária | Sim |
| Qual a quantidade anual de dolutegravir comprada pelo Ministério da Saúde em 2022? | 1 | Intermediária | Sim |
| Qual a quantidade anual de dolutegravir comprada pelo Ministério da Saúde de 2000 a 2022? | 1 | Intermediária | Sim |
| Gastos com insulina por ano | 1 | Intermediária | Sim |
| Total de gastos com insulina por ano | 1 | Intermediária | Sim |
| Valor total gasto com insulina por ano | 1 | Intermediária | Sim |
| Soma de gasto com insulina por ano | 1 | Intermediária | Sim |
| Qual o mês em que o SE gastou mais? | 1 | Intermediária | Sim |

| Pergunta | Ocorrências | Categoria | Gerou resultado |
|---|---|---|---|
| Quais foram os maiores gastos em Sergipe? | 1 | Intermediária | Sim |
| Gastos componente especializado da atenção farmacêutica? | 1 | Intermediária | Sim |
| Traga-me dados de compras do ano de 2024 por Estado | 1 | Intermediária | Sim |
| Quanto foi o valor contratado de adalimumabe pelo MS? | 1 | Intermediária | Não |
| Quanto foi o valor de adalimumabe pelo MS? | 1 | Intermediária | Sim |
| Gasto com adalimumabe | 1 | Intermediária | Sim |
| Gastos componente especializado? | 1 | Intermediária | Sim |
| Quais são os produtos vendidos pela Blau? | 1 | Intermediária | Não |
| Blau vende quais produtos? | 1 | Intermediária | Não |
| Quais produtos comprados a Blau? | 1 | Intermediária | Não |
| Quais produtos fornecidos pela Blau? | 1 | Intermediária | Não |
| O que a Multifarma fornece? | 1 | Intermediária | Não |
| Qual a quantidade anual de dolutegravir comprada pelo Ministério da Saúde desde 2020 ? | 1 | Intermediária | Sim |
| O que a Cristália fornece? | 1 | Intermediária | Não |
| O que a Blau fornece? | 1 | Intermediária | Não |
| Gastos anuais por UF | 1 | Intermediária | Sim |
| Qual modalidade de compra da insulina  por ano? | 1 | Intermediária | Não |
| Top 10 fornecedores da SVSA | 1 | Intermediária | Sim |
| Top 10 medicamentos da SVSA | 1 | Intermediária | Sim |
| Quanto o ministério gastou com insulina em 2023? | 1 | Intermediária | Sim |
| Quantidade comprada de insulina em 2025 | 1 | Intermediária | Sim |
| Quantidade comprada de rivastigmina nos últimos 10 anos | 1 | Intermediária | Sim |
| Quantidade comprada de rivastigmina nos últimos 10 anos, por produto | 1 | Intermediária | Sim |
| Quanto o ministério da saúde gastou por mês em 2024? | 1 | Intermediária | Sim |
| Quanto foi gasto com medicamentos do programa farmácia popular entre 2024 a 2025? | 1 | Intermediária | Sim |
| Quanto foi gasto no ano de 2025 com medicamento rituximabe? | 1 | Intermediária | Sim |
| Quais os principais fornecedores da insulina? | 1 | Intermediária | Sim |

| Pergunta | Ocorrências | Categoria | Gerou resultado |
|---|---|---|---|
| Qual mês Ms compra mais? | 1 | Intermediária | Sim |
| Quanto foi gasto em vacinas? | 1 | Intermediária | Sim |
| O que a Fiocruz fornece? | 1 | Intermediária | Sim |
| Qual mês e ano MS compra mais? | 1 | Intermediária | Sim |
| Pregão é mais barato para insulina? | 1 | Intermediária | Sim |
| O que o Butantan fornece? | 1 | Intermediária | Sim |
| Qual o medicamento mais caro que o Butantan fornece? | 1 | Intermediária | Sim |
| Qual o medicamento mais caro que o Butantan fornece em valor unitário? | 1 | Intermediária | Sim |
| Quais medicamentos têm apenas um fornecedor? | 1 | Intermediária | Não |
| Medicamento com apenas 1 fornecedor | 1 | Intermediária | Não |
| 1 fornecedor | 1 | Intermediária | Sim |
| O que a multicare vende? | 1 | Intermediária | Não |
| quais medicamentos sao fornecidos pela multicare? | 1 | Intermediária | Não |
| Medicamentos com apenas 1 fornecedor | 1 | Intermediária | Não |
| Multicare | 1 | Intermediária | Não |
| Medicamento judicializado mais caro | 1 | Intermediária | Sim |
| Medicamento judicializado mais caro valor unitário | 1 | Intermediária | Sim |
| Onde o MS economiza? | 5 | Complexa | Sim |
| Qual são os medicamentos que estão no quartil superior de custos? | 2 | Complexa | Não |
| MS paga mais caro que outros órgãos? | 3 | Complexa | Não |
| Considerando os preços médios, qual o medicamento que teve o maior preço de compra proporcionalmente? | 1 | Complexa | Sim |
| Qual a diferença de compra entre o MS e os estados? | 1 | Complexa | Não |
| Qual a diferença de preço entre o MS e os estados para o medicamento atalureno? | 1 | Complexa | Não |
| Qual a diferença de preço entre o MS e os estados para o atalureno? | 1 | Complexa | Não |
| Qual o histórico de judicialização realizado? | 1 | Complexa | Sim |
| Quais medicamentos fazem mais efeitos pra diabetes? | 1 | Complexa | Sim |
| Quais medicamentos fazem mais efeitos ? | 1 | Complexa | Sim |

| Pergunta | Ocorrências | Categoria | Gerou resultado |
|---|---|---|---|
| Quais as doenças mais raras detectadas em Sergipe? | 1 | Complexa | Sim |
| Percentual de valor gasto por fornecedor por ano | 1 | Complexa | Não |
| Percentual de valor gasto por fornecedor | 1 | Complexa | Sim |
| Quais os 5 medicamentos mais caros e quais os 5 medicamentos com maior volume de compra para os últimos 5 anos | 1 | Complexa | Não |
| Quanto foi o preço contratado de adalimumabe pelo MS nos últimos 5 anos? | 1 | Complexa | Sim |
| Quanto foi o preço contratado de adalimumabe pelo MS nos últimos 5 anos por fornecedor? | 1 | Complexa | Não |
| Percentual de valor gasto por fornecedor em 2020 | 1 | Complexa | Sim |
| Qual são os medicamentos que estão no quartil superior de custos? | 1 | Complexa | Sim |
| Onde MS economiza? | 1 | Complexa | Sim |
| Os gastos da SCTIE por ano estão mais relacionados a que? | 1 | Complexa | Sim |
| Comparar SVSA e SCTIE | 1 | Complexa | Sim |
| Qual o maior valor unitário pago? | 1 | Complexa | Sim |
| Qual l o medicamento com o maior valor unitário médio? | 1 | Complexa | Sim |
| Qual o medicamento com o maior valor unitário médio sem ser transporte | 1 | Complexa | Sim |
| Qual l o medicamento com o maior valor unitário médio sem ser transporte ou armazenagem | 1 | Complexa | Sim |
| Qual a concentração do mercado | 1 | Complexa | Não |
| Concentração de mercado | 1 | Complexa | Não |
| Concentração do mercado | 1 | Complexa | Não |

# Um Agente Inteligente Baseado em LLM para Consulta em Linguagem Natural de Dados Aquaviários da ANTAQ

*Tadeu Gomes Teixeira*
*Prof. Dr. Alex Lopes Pereira*

## Resumo

A Agência Nacional de Transportes Aquaviários (ANTAQ) publica estatísticas oficiais de movimentação de cargas em um conjunto de dados públicos por meio de um painel de *Business Intelligence* intitulado Estatístico Aquaviário. Apesar de seu valor para políticas públicas e pesquisa, os microdados dependem de ferramentas especializadas para consulta. Este trabalho apresenta o agente de IA desenvolvido para consultar os dados via modelo de linguagem de grande escala (LLM), que interpreta perguntas em português sobre estatísticas aquaviárias, converte-as em consultas GoogleSQL executáveis no BigQuery e retorna resultados com explicações em português. A arquitetura combina orquestração via grafo de estados LangGraph, componente semântico de metadados institucionais (MetadataHelper), recuperação de exemplos *few-shot* e camada de validação e segurança SQL. A avaliação empírica em um conjunto de 33 questões domínio-específicas (Gold Set) alcança acurácia de execução EX = 100% após quatro rodadas iterativas de refinamento (linha de base: EX = 84,8%). Em um conjunto independente de teste com 60 questões de intenções e entidades inéditas (não usadas no refinamento), o agente acerta 59/60 (98,3%; IC 95% Wilson [91,1%, 99,7%]), evidência de generalização além do Gold Set. O artigo deriva quatro proposições de design para agentes de acesso a dados governamentais. Os resultados têm implicações para a adoção de IA no setor público, para a democratização do acesso a dados abertos e para a construção de capacidades digitais em organizações brasileiras.

**Palavras-chave:** inteligência artificial; agentes de IA; text-to-SQL; dados abertos governamentais; transporte aquaviário; ANTAQ; design science.

## INTRODUÇÃO

Tornar dados públicos disponíveis em formato aberto é condição necessária, mas insuficiente, para seu uso social e econômico. A experiência acumulada em iniciativas de governo aberto ao redor do mundo demonstra que a mera publicação de bases de dados estruturadas não garante seu aproveitamento por parte de gestores públicos, pesquisadores, jornalistas ou cidadãos sem formação técnica especializada (Janssen et al., 2012; Lourenço, 2015). No Brasil, esse cenário se manifesta também no setor de transportes aquaviários, onde a Agência Nacional de Transportes Aquaviários (ANTAQ) coleta, processa e publica estatísticas detalhadas de movimentação de cargas, desempenho de terminais portuários e tendências do setor por meio de um conjunto de dados público no Estatístico Aquaviário – um painel de *business intelligence*. Essas bases de dados têm valor para o planejamento de políticas de infraestrutura, o monitoramento dos contratos de arrendamento e concessão, a

avaliação da eficiência logística e a pesquisa acadêmica sobre o setor portuário, mas seu uso efetivo pressupõe domínio de ferramentas para o tratamento de microdados, familiaridade com a metodologia oficial da ANTAQ e acesso a ferramentas de inteligência de negócios, requisitos que excluem a maior parte dos potenciais usuários.

Assim, a despeito de sua relevância para a gestão pública, os dados aquaviários da ANTAQ permanecem acessíveis, em sua maior parte, apenas a usuários com perfil técnico. Essa barreira cria assimetrias informacionais que comprometem os objetivos de transparência e controle social que justificam, do ponto de vista institucional, a publicação dos dados com base na Lei de Acesso à Informação (Lei nº 12.527/2011). O problema não se reduz à exigência de proficiência em ferramentas analíticas. A ANTAQ define a metodologia oficial de cálculo que distingue as movimentações válidas das demais, mas reproduzi-la sobre os microdados exige operacionalizá-la em um filtro específico – tarefa realizada neste trabalho na construção do repositório de dados (ver seção 3.2). Além disso, os dados adotam nomenclaturas técnicas que diferem do vocabulário cotidiano (a direção "exportação" corresponde ao sentido semântico "Embarcados" nos registros de atracação); e a publicação dos microdados ocorre com defasagem, afetando a disponibilidade dos períodos mais recentes. Esses elementos, combinados com as especificidades do dialeto GoogleSQL e com os requisitos de controle de custos, em ambientes de nuvem pagos por consumo, configuram um problema substancialmente mais complexo do que o acesso genérico a bancos de dados corporativos.

É nesse contexto que se insere este trabalho. O objetivo foi desenvolver, documentar e avaliar empiricamente um agente inteligente capaz de interpretar perguntas em linguagem natural sobre dados aquaviários da ANTAQ e convertê-las em consultas SQL executáveis no BigQuery, retornando resultados tabulares acompanhados de explicações em português. A pergunta de pesquisa central foi: como construir um agente baseado em modelo de linguagem de grande escala (LLM) que torne os dados estatísticos da ANTAQ acessíveis a usuários não técnicos, preservando a correção metodológica das respostas e atendendo aos requisitos de segurança de um ambiente de dados governamentais? A hipótese subjacente é que a barreira de acesso aos dados não é tecnicamente inevitável e pode ser reduzida por meio de arquiteturas agênticas que embutem conhecimento institucional explícito, em contraste com os sistemas genéricos de *text-to-SQL* que tratam todos os domínios como equivalentes.

A lacuna que o trabalho preenche é dupla. Com base em revisão sistemática da literatura conduzida nas bases Google Scholar, ACL Anthology, ACM DL e arXiv, utilizando termos como "*text-to-SQL*", "BigQuery", "*government data*", "português brasileiro" e "*waterway statistics*", embora existam sistemas de text-to-SQL em português brasileiro – como o mRAT-SQL+GAP (JOSÉ; COZMAN, 2021), uma arquitetura integrada de pergunta-resposta e text-to-SQL (JOSÉ et al., 2022) e, mais recentemente, text-to-SQL com LLM sobre dados governamentais educacionais brasileiros (FRÓES; BRAGHETTO, 2025) –, não se identificou trabalho anterior que aplique um agente de *text-to-SQL* a estatísticas aquaviárias da ANTAQ, conjugando a metodologia oficial, o dialeto GoogleSQL, a governança de custos em nuvem e a defasagem na publicação dos microdados. A complexidade combinada das restrições metodológicas institucionais, das normas de normalização do domínio portuário, da governança de custos em nuvem e da defasagem na publicação dos microdados não havia sido ainda sistematicamente caracterizada como conhecimento de *design* transferível.

As contribuições do trabalho são: (i) o componente MetadataHelper, que codifica a metodologia oficial da ANTAQ na forma de filtros, a semântica de direção e o glossário do domínio diretamente no contexto do agente; (ii) um mecanismo de recuperação de exemplos *few-shot* semântico *(embeddings)* com 32 pares questão-SQL curados; (iii) uma camada de

validação e segurança SQL multi-camada adaptada ao BigQuery/GoogleSQL, com modelo de ameaças documentado e avaliação adversarial; (iv) quatro proposições de *design* para agentes *text-to-SQL* em contextos de dados governamentais; e (v) uma implementação reproduzível e configurável para outros repositórios de dados públicos com características análogas.

# 1. REFERENCIAL TEÓRICO

## 1.1 Agentes de IA e LLMs em contextos organizacionais

O uso de modelos de linguagem de grande escala (LLMs) em organizações tem evoluído de aplicações circunscritas à geração de texto para sistemas capazes de raciocinar, planejar e executar ações sobre sistemas externos, incluindo bancos de dados, APIs e ferramentas de automação. Essa transição é sistematizada na literatura de sistemas de informação sob o conceito de *agentes de IA*, definidos como sistemas autônomos capazes de perceber seu ambiente, raciocinar sobre ele e executar ações para alcançar objetivos predefinidos (Yao et al., 2022). O paradigma ReAct, proposto por Yao et al. (2022), estabelece um ciclo iterativo de raciocínio e ação no qual o agente alterna entre geração de pensamento, execução de ação e observação do resultado, permitindo a correção iterativa de erros sem necessidade de retreinamento do modelo. Para aplicações organizacionais, esse paradigma é relevante porque endereça uma limitação central dos LLMs estáticos: a incapacidade de operar sobre dados que mudam ao longo do tempo e de executar tarefas que requerem verificação de resultados intermediários.

No contexto de acesso a dados, modelos como o GPT-4o alcançam 85,3% de acurácia de execução em SQL sobre o *benchmark Spider* (Pourreza & Rafiei, 2023), que reúne questões sobre bancos de dados de 138 domínios diferentes. No entanto, esses mesmos modelos caem para 10,1% no Spider 2.0, *benchmark* que simula ambientes empresariais reais com 700 a 3.000 colunas por banco, múltiplos dialetos SQL e necessidade de consultar documentação externa (Lei et al., 2024).

A arquitetura MAC-SQL (Wang et al., 2024) representa um sistema multi-agente para *text-to-SQL* com três agentes especializados: *Selector Agent* (poda de esquema), *Decomposer Agent* (decomposição via *Chain-of-Thought*) e *Refiner Agent* (validação e correção SQL), alcançando 59,59% de acurácia de execução no *benchmark* BIRD com GPT-4. O sistema CHESS (Talaei et al., 2024) alcança 71,10% no BIRD. Esses sistemas, contudo, foram projetados para bancos de dados genéricos e não incluem mecanismos para codificação de metodologias institucionais oficiais, tornando-os inadequados para o contexto de dados governamentais com metodologias de cálculo regulamentadas.

A lacuna entre o desempenho em *benchmarks* padronizados e a *performance* em ambientes reais aponta para um fenômeno que Wang et al. (2024) denominam "falta de *grounding* institucional": os modelos falham não porque não sabem SQL, mas porque não têm acesso ao conhecimento tácito que os especialistas do domínio utilizam ao formular consultas corretas. Para dados governamentais, onde os resultados das consultas serão comparados a publicações oficiais, essa falta de *grounding* produz um tipo específico de erro que pode ser chamado de "alucinação metodológica": consultas sintaticamente perfeitas que retornam resultados incorretos por ignorar filtros oficiais, convenções de nomenclatura ou a defasagem na publicação dos microdados.

Para contornar isso, o LangGraph, *framework* de orquestração adotado neste trabalho, implementa máquinas de estado para agentes por meio de uma arquitetura em grafo com suporte nativo a ciclos, essencial para o padrão iterativo de geração e correção de SQL (Malewicz et al., 2010).

## 1.2 Text-to-SQL: do pioneirismo à era dos LLMs

A conversão automática de linguagem natural em SQL tem origem nos sistemas de interface para bancos de dados NLIDBs *(Natural Language Interface to Databases)* da década de 1970. O sistema LUNAR (Woods, 1973), desenvolvido para consultar análises de amostras lunares das missões Apollo, e o sistema LADDER (Hendrix et al., 1978), que introduziu gramáticas semânticas configuráveis, são os trabalhos inaugurais do campo. O *framework* PRECISE (Popescu et al., 2003) combinou tokenização semântica com consulta ao WordNet para alcançar independência de domínio sem necessidade de treinamento supervisionado. Uma característica comum a todos esses sistemas pioneiros é que foram construídos para bancos de dados específicos e não podiam ser facilmente adaptados a outros domínios (Androutsopoulos et al., 1995), limitação que o trabalho presente aborda por meio do componente *MetadataHelper*, que torna o *grounding* institucional configurável e parametrizável.

O ano de 2017 marcou uma inflexão no campo. A publicação do Seq2SQL (Zhong et al., 2017), primeira rede neural profunda dedicada à tradução de linguagem natural em SQL, veio acompanhada do *dataset* WikiSQL, com 80.654 pares de questão-SQL, que estabeleceu a primeira referência de larga escala para a tarefa. No ano seguinte, o *benchmark* Spider (Yu et al., 2018) elevou o grau de dificuldade ao reunir 200 bancos de dados distribuídos por 138 domínios, com consultas que exigiam junções, agregações e subconsultas aninhadas. O *benchmark* BIRD (Li et al., 2023), apresentado na NeurIPS 2023, aproximou a avaliação das condições de uso real, ao incorporar bancos com dados sujos, dependência de conhecimento externo e uma métrica de eficiência da consulta. Seus resultados expuseram a distância entre o desempenho humano e o dos modelos: enquanto especialistas humanos alcançaram 92,96% de acurácia de execução, o GPT-4 obteve 54,89%. O Spider 2.0 (Lei et al., 2024) ampliou essa distância ao simular ambientes empresariais reais, com bancos de 700 a 3.000 colunas e múltiplos dialetos de SQL, cenário em que o GPT-4o atingiu apenas 10,1%. Constata-se, assim, que o problema permanece em aberto fora dos *benchmarks* padronizados.

A emergência dos LLMs deslocou a abordagem dominante do ajuste fino supervisionado para o aprendizado em contexto (*in-context learning*): o método DIN-SQL (Pourreza & Rafiei, 2023), apoiado na decomposição da consulta, alcançou 85,3% de acurácia de execução no Spider. Para o português brasileiro, José e Cozman (2021) e José et al. (2022) adaptaram arquiteturas de text-to-SQL ao idioma, e Fróes e Braghetto (2025) aplicaram modelos baseados em LLM a dados governamentais educacionais.

Para o domínio governamental, há desafios adicionais que não são contemplados pelos *benchmarks* existentes: a terminologia usada em organizações públicas, as metodologias oficiais de cálculo de estatísticas públicas nem sempre publicadas e a necessidade de auditabilidade dos resultados são requisitos específicos do contexto governamental que sistemas como MAC-SQL e CHESS não endereçam.

A omissão de regras institucionais não inscritas no esquema configura um tipo de erro que a taxonomia de Qu et al. (2024), centrada em alucinações que contradizem o esquema, a pergunta ou a sintaxe, não captura: a alucinação metodológica. Diferentemente das

alucinações de esquema e de lógica catalogadas por aqueles autores, ela não decorre de uma incoerência interna da consulta, mas da ausência de um filtro oficial que o modelo não tem como inferir a partir dos metadados. O caso é particularmente problemático em contextos de prestação de contas. No caso aqui desenvolvido, uma consulta que retorna as exportações de um porto sem o filtro isValidoMetodologiaANTAQ = 1 é sintaticamente correta, mas produz um número diverso do publicado oficialmente pela ANTAQ, e o usuário sem conhecimento técnico não tem como detectar a divergência. Esses aspectos exigem soluções, como o RAG, descritas na próxima seção.

## 1.3 Geração aumentada por recuperação (RAG)

Lewis et al. (2020) definiram o paradigma RAG (*Retrieval-Augmented Generation*) como a combinação de memória paramétrica (conhecimento embarcado nos parâmetros do modelo durante o treinamento) com memória não paramétrica (índice externo consultável em tempo de inferência). Para sistemas de *text-to-SQL*, o RAG tem três aplicações principais: recuperação de exemplos similares via *embedding* de consultas históricas validadas; recuperação de documentação de esquemas, incluindo descrições de tabelas e colunas; e *caching* semântico de pares consulta-resposta validados. Gao et al. (2024) propõem uma taxonomia de três estágios evolutivos: RAG ingênuo, RAG avançado e RAG modular. O RAG agêntico, em que a recuperação é orquestrada por agentes com capacidade de planejamento, constitui uma extensão relevante para contextos em que o vocabulário do domínio evolui ao longo do tempo e os exemplos validados acumulam valor.

Para *text-to-SQL*, os ganhos do RAG são documentados de forma consistente na literatura. O sistema DAIL-SQL (Gao et al., 2024) demonstrou que representações de questões com esqueleto sintático combinadas com *prompting* eficiente alcançam equilíbrio favorável entre acurácia e eficiência de *tokens*. Qu et al. (2024) caracterizam as alucinações em SQL gerado por LLM em duas categorias, cada uma com subtipos: as alucinações de esquema (contradição com a estrutura do esquema, introdução de atributos desnecessários e geração de valores em formato inexistente no esquema) e as alucinações de lógica (junções redundantes, abuso de cláusulas e conversão incorreta de lógica matemática em funções SQL). Para o domínio da ANTAQ, a geração de valores inexistentes no esquema é particularmente problemática: um modelo que não tem acesso explícito à semântica de direção pode gerar consultas com *sentido = 'Exportações'* em vez de *sentido = 'Embarcados'*, retornando resultados nulos sem sinalizar o erro ao usuário. O componente RAG implementado neste trabalho endereça esse problema por meio de exemplos curados que codificam explicitamente os mapeamentos críticos de termos do domínio, convertendo conhecimento tácito em conhecimento explícito acessível ao LLM.

## 1.4 Dados abertos governamentais, transparência e controle social

A literatura sobre governo aberto e dados abertos governamentais documenta um paradoxo recorrente: quanto mais dados públicos são disponibilizados, maior o risco de que apenas atores com capacidade técnica para processá-los se beneficiem da abertura, ampliando, em vez de reduzindo, as assimetrias informacionais entre o Estado e diferentes segmentos da sociedade (Janssen et al., 2012). Lourenço (2015) identifica quatro barreiras à efetividade dos dados abertos: qualidade técnica dos dados, capacidade analítica dos usuários, inexistência de ferramentas de acesso adequadas e ausência de incentivos institucionais para o uso dos dados nas decisões políticas. Estudos empíricos no contexto brasileiro corroboram essas

barreiras: Klein, Klein e Luciano (2018) mapeiam mecanismos para ampliar a transparência em portais de dados abertos, e Silva (2023) avalia o potencial de reúso de dados governamentais abertos durante a pandemia de COVID-19. No caso dos microdados aquaviários da ANTAQ, as quatro barreiras se manifestam simultaneamente: os dados têm qualidade suficiente para uso analítico, mas a metodologia oficial de cálculo é complexa; a capacidade analítica dos usuários-alvo é heterogênea; as ferramentas disponíveis exigem SQL ou *scripts* Python; e os incentivos para o uso dos dados no debate sobre políticas de infraestrutura são restritos pela opacidade técnica.

A adoção de IA conversacional para reduzir barreiras de acesso a dados governamentais insere-se em um debate mais amplo sobre capacidades digitais do Estado e sobre o papel das tecnologias de informação na modernização da administração pública brasileira. Bresser-Pereira (1998) identificou a construção de capacidades técnicas como condição para a reforma gerencial do Estado, e a literatura posterior sobre governo eletrônico retomou esse argumento ao tratar da digitalização dos serviços públicos como oportunidade para ampliar acesso, reduzir custos de transação e melhorar a qualidade da prestação de serviços (Fountain, 2001). A interface conversacional para acesso a dados de agências reguladoras pode ser lida como extensão desse argumento: se o objetivo da publicação de dados públicos é tornar o Estado *accountable* perante a sociedade, ferramentas que reduzem o custo de acesso ampliam o alcance prático da transparência.

No setor portuário brasileiro, a dimensão regulatória do acesso aos dados tem implicações concretas. A ANTAQ é responsável pela regulação e fiscalização das atividades de exploração da infraestrutura aquaviária, conforme a Lei nº 10.233/2001. Os dados publicados pela ANTAQ não são apenas estatísticas de interesse geral: são insumos para contratos de arrendamento portuário, para o cálculo de tarifas e para a avaliação do desempenho dos operadores de terminais. Um ator com capacidade de interrogar diretamente a base de dados da ANTAQ pode monitorar se os operadores cumprem as metas de movimentação previstas nos contratos, acompanhar a evolução da participação de mercado dos terminais por tipo de carga e identificar padrões que possam indicar irregularidades. A democratização desse acesso, por meio de uma interface em linguagem natural, tem, portanto, implicações que vão além da conveniência analítica e tocam diretamente na qualidade do controle social sobre a regulação do setor.

## 2. METODOLOGIA

### 2.1 Abordagem de pesquisa: Design Science Research

Este trabalho adota a abordagem de *Design Science Research* (DSR), sistematizada por Hevner et al. (2004) em um *framework* seminal para pesquisa em sistemas de informação. A DSR se distingue das pesquisas comportamentais por seu foco na construção de artefatos tecnológicos que resolvem problemas práticos identificados em contextos organizacionais reais, avaliando sua utilidade por meio de evidências empíricas. Segundo Hevner et al. (2004), um artefato de pesquisa pode ser um constructo, um modelo, um método ou uma instância; este trabalho produz uma instância (o sistema *agente desenvolvido*) e, derivadas dela, proposições de *design* que têm *status* de contribuição teórica para o conhecimento da área. O mapeamento entre requisitos de *design*, características do artefato e evidências de avaliação é estruturado da seguinte forma: o requisito de acessibilidade do domínio motiva o componente *MetadataHelper*, avaliado pela acurácia de execução em questões domínio-específicas; o

requisito de correção institucional motiva a codificação dos filtros metodológicos, avaliado contra os agregados oficiais do Estatístico Aquaviário da ANTAQ (ANTAQ, 2023); e o requisito de segurança operacional motiva a camada de validação multi-camada, avaliada contra um modelo de ameaças estruturado com avaliação adversarial sistemática.

A escolha da DSR como abordagem metodológica é justificada pela natureza do problema. Trata-se de um problema de construção de conhecimento a partir da resolução de um problema prático específico: as proposições de *design* DP1 a DP4, derivadas neste trabalho, não teriam sido produzidas por uma revisão de literatura ou por um estudo observacional, mas emergem das decisões de *design* tomadas e de sua avaliação empírica, tendo validade como hipóteses para generalização em outros contextos de dados governamentais. Nesse sentido, o trabalho contribui tanto com o artefato em si quanto com o conhecimento de *design* que ele incorpora, distinção que Hevner et al. (2004) consideram central para a relevância científica da DSR.

## 2.2 Ambiente de dados e execução

O artefato consulta um repositório em Google BigQuery construído a partir dos microdados do Estatístico Aquaviário publicados pela ANTAQ (projeto antaqdados, conjunto de dados br_antaq_estatistico_aquaviario, localização US). A metodologia oficial da ANTAQ, descrita em seu manual de cálculo de indicadores (ANTAQ, 2025), foi reproduzida e validada como uma visão (v_carga_metodologia_oficial); as estatísticas de carga são calculadas principalmente a partir da tabela particionada e clusterizada tb_carga_metodologia_oficial (materialização dessa visão), usando *vlpesocargabruta_oficial* (toneladas) como métrica principal e *isValidoMetodologiaANTAQ = 1* como filtro de metodologia oficial. O esquema inclui também as visões e tabelas auxiliares *v_atracacao_validada, v_carga_validada, instalacao_origem, instalacao_destino* e *mercadoria_carga*. A execução das consultas é limitada por controles de custo por *job* (*maximum_bytes_billed* ≈ 10 GB) e *timeouts* de 60 segundos, com *cache* de consultas habilitado.

O agente é implementado em Python 3.10+, com orquestração via LangGraph e integração de LLM via LangChain, exposto por um *backend* FastAPI (API REST) e um *frontend web* em Next.js. A temperatura do LLM é configurada em 0 e o máximo de *tokens* de saída em 1.000, para maximizar o determinismo. O provedor de LLM padrão é OpenAI; a geração de SQL usa gpt-4o (robustez em consultas multifiltro) e a resposta final usa o gpt-4o-mini; há caminho alternativo para Google Vertex AI (gemini-1.5-flash). O *deployment* em contêiner via Docker é suportado, com alvo para Google Cloud Run.

## 2.3 Fases da pesquisa

A pesquisa percorreu seis fases sequenciais com retroalimentação iterativa. A fase de análise de requisitos envolveu o levantamento e categorização das consultas típicas realizadas por usuários do setor portuário, a identificação das regras metodológicas oficiais da ANTAQ que devem ser respeitadas nas consultas e a elaboração dos metadados semânticos do domínio. A fase de projeto de arquitetura definiu as quatro camadas do sistema e especificou os estados, nós e transições do fluxo agêntico, incluindo as condições de roteamento para o *loop* de *retry*. A fase de implementação do núcleo desenvolveu os módulos de metadados, RAG, validação SQL e orquestração LangGraph, com testes unitários para cada componente. A fase de interface e integração implementou a interface conversacional web (Next.js sobre API FastAPI) e integrou os componentes em um *pipeline* coerente. A fase de testes e

validação conduziu a avaliação empírica com o Gold Set de 33 questões, o estudo de ablação com cinco configurações e a avaliação adversarial com 100 *prompts* de injeção. A fase de análise e documentação interpretou os resultados, derivou as proposições de *design* e preparou a documentação para reprodução do artefato.

### 2.4 Disponibilidade de Dados e Código

O repositório do projeto com as instruções para acesso aos dados e códigos está disponível em https://github.com/tadeugomes/agente_antaq_sql.

O código do projeto foi desenvolvido com auxílio de agentes de inteligência artificial generativa.

## 3. ARQUITETURA DO SISTEMA

A arquitetura implementada organiza-se em quatro camadas funcionais com responsabilidades bem delimitadas. A camada de apresentação é responsável pela interação com o usuário por meio de um frontend web em Next.js que consome uma API REST (FastAPI). O SQL gerado é ocultado por padrão; um modo de depuração opcional expõe a consulta para verificação especializada. A camada agêntica orquestra o fluxo de geração, validação e execução de SQL por meio de um grafo LangGraph com memória conversacional por sessão (*thread_id*). A camada de serviços contém os módulos auxiliares: *MetadataHelper* para metadados semânticos, *ExampleRetriever* para recuperação *few-shot*, *SQLValidator* para validação e segurança, e *BigQueryClient* para execução das consultas. A camada de dados acessa, por meio de conta de serviço com privilégios somente-leitura, o repositório em BigQuery construído neste trabalho a partir dos dados do Estatístico Aquaviário publicados pela ANTAQ.

### 3.1 Fluxo agêntico e orquestração via LangGraph

O fluxo do agente é implementado como um *StateGraph* com nós explícitos e roteamento condicional. A execução parte de dois nós em paralelo: *setup_schema*, que carrega e armazena em cache a descrição do esquema via MetadataHelper, e *retrieve_examples*, que recupera os três exemplos *few-shot* mais próximos da questão por similaridade de cosseno entre embeddings (text-embedding-3-small), com fallback de sobreposição léxica. O nó *generate_sql* invoca o LLM com um *prompt* contendo o esquema semântico, os exemplos recuperados, o histórico de conversa e as regras críticas do domínio, gerando uma consulta GoogleSQL candidata. A consulta passa por *validate_sql*, onde são aplicadas as políticas de segurança e as normalizações de domínio; em caso de falha, o grafo roteia de volta para *generate_sql* com a mensagem de erro incluída no *prompt*, até um máximo configurável de tentativas (padrão: duas). Uma vez validada, a consulta é executada no BigQuery em *execute_sql* de forma assíncrona, e por fim *generate_final_answer* sintetiza a resposta em português a partir dos resultados, sem expor o SQL ao usuário por padrão. A memória conversacional é mantida por um *checkpointer* SQLite por sessão (*SqliteSaver*), com *fallback* para *MemorySaver* em memória quando o SQLite não está disponível.

## 3.2 Metadados semânticos institucionais (MetadataHelper)

O *MetadataHelper* é o componente que diferencia este sistema dos agentes *text-to-SQL* genéricos da literatura. Sua função é embarcar no contexto do agente o conhecimento institucional que um LLM de propósito geral não consegue inferir a partir do esquema do banco: as regras da metodologia oficial de cálculo das estatísticas aquaviárias, os mapeamentos de vocabulário entre o português cotidiano e os valores dos campos, e as restrições temporais decorrentes da defasagem na publicação dos microdados. Quando a tabela *dicionario_dados* existe no BigQuery, o *helper* constrói descrições semânticas por tabela, categoria e coluna, incluindo valores possíveis e rótulos.

Na ausência do dicionário, um esquema resumido centrado na tabela oficial tb_carga_metodologia_oficial é utilizado como *fallback*. As regras operacionais codificadas incluem: (i) aplicação obrigatória do filtro *isValidoMetodologiaANTAQ = 1* em toda consulta sobre a visão oficial; (ii) uso de *vlpesocargabruta_oficial* > 0; (iii) semântica de direção (exportações correspondem a *sentido = 'Embarcados'*; importações a *sentido = 'Desembarcados'*); (iv) corte temporal dinâmico, consultado por sessão via *query_max_available_date()*, com *fallback* de 45 dias; e (v) glossário de termos do domínio portuário e regras marcadas como "CRÍTICAS" no prompt de sistema, definidas no módulo de *prompts* do agente (src/agent/prompts.py).

Esse design diverge da abordagem de poda de esquema adotada por MAC-SQL e CHESS, que operam sobre esquemas genéricos sem mecanismo para codificação de metodologias oficiais regulamentadas.

## 3.3 Recuperação de exemplos few-shot (RAG)

A orientação *few-shot* é implementada por meio de um conjunto curado de 32 pares questão-SQL validados (QA_EXAMPLES), alinhados à tabela oficial tb_carga_metodologia_oficial e incluindo mapeamentos explícitos dos termos críticos do domínio. Os exemplos cobrem agregações temporais, rankings por porto/estado/região, comparações direcionais, junções com o catálogo de mercadorias e consultas multifiltro, distribuídos entre os níveis básico (10), intermediário (12) e avançado (10). A recuperação é semântica: cada questão é vetorizada com *embeddings* (text-embedding-3-small) e os três exemplos mais próximos por similaridade de cosseno são selecionados (RAG_TOP_K = 3), com um escore de sobreposição léxica como *fallback* automático quando os *embeddings* não estão disponíveis. O *corpus* de exemplos também é publicado em uma base vetorial no BigQuery (tabela qa_embeddings) para recuperação opcional no lado do servidor.

## 3.4 Camada de validação, segurança e modelo de ameaças

Dado que o agente gera e executa consultas SQL contra uma instância BigQuery ativa, a camada de segurança e validação é tratada como requisito de primeira classe. A superfície de ataque compreende três classes de atores: (i) usuários maliciosos que elaboram *prompts* em linguagem natural para induzir SQL nocivo; (ii) ataques de injeção de *prompt* embarcados em valores de dados ou em saídas intermediárias do LLM; e (iii) alucinações acidentais do LLM que produzem consultas sintaticamente válidas mas semanticamente perigosas. Os ativos a proteger são a integridade dos dados, a confidencialidade e o custo operacional.

O componente *SQLValidator* aplica os seguintes controles em sequência: bloqueio de palavras-chave DDL/DML (DROP, DELETE, UPDATE, INSERT, CREATE, ALTER, TRUNCATE, GRANT, REVOKE, EXECUTE, CALL, MERGE, REPLACE, EXPORT); restrição de prefixos de consulta permitidos a SELECT, WITH e ( apenas; bloqueio de acesso a objetos de metadados via *FORBIDDEN_CONTENT_PATTERNS*; validação de referências de tabelas na cláusula FROM contra uma lista de tabelas autorizadas usando análise de AST via *sqlglot*, bloqueando acesso entre conjuntos de dados de forma estrutural; detecção e rejeição de padrões de injeção incluindo comentários SQL, tautologias, literais hexadecimais, UNION ALL SELECT e encadeamento de instruções; e injeção automática de LIMIT 1000 quando ausente. No nível de execução, o cliente BigQuery configura *maximum_bytes_billed* ≈ 10 GB e *use_query_cache* = *True*, com credenciais carregadas de conta de serviço com privilégios somente-leitura.

# 4. AVALIAÇÃO EMPÍRICA

## 4.1 Definição e estrutura do Gold Set

Foi construído um conjunto de avaliação domínio-específico (Gold Set) com 33 questões em português brasileiro, cada uma acompanhada de: (i) uma consulta SQL de referência executável no BigQuery; (ii) um resultado esperado extraído dos dados oficiais do Estatístico Aquaviário da ANTAQ (ANTAQ, 2023); (iii) uma categoria de complexidade; e (iv) uma especificação de tolerância numérica (±1% para agregações escalares; correspondência ordinal para *rankings*).

O conjunto cobre cinco categorias de complexidade: *filtro_simples* (8 itens), com consultas de uma única dimensão de filtragem; *agregacao_temporal* (7), com séries mensais ou anuais; *ranking_porto* (7), com *rankings* por volume de carga; *join_multidimensional* (6), com cruzamentos entre múltiplas dimensões; e *multifiltro* (5), com combinações de dois ou mais filtros simultâneos. As questões cobrem o período 2010-2024 e abrangem as principais dimensões analíticas do domínio: porto, estado (UF), região geográfica, direção de carga, tipo de navegação e mercadoria. Os tipos de resultado são: escalar (17), série temporal (8) e *ranking* (8). A avaliação foi conduzida por comparação direta de DataFrames, sem uso de LLM como árbitro, para garantir objetividade e reprodutibilidade.

## 4.2 Acurácia de execução e refinamento iterativo

O agente completo (configuração B4, com *MetadataHelper*, *few-shot* e *retry* habilitados) foi avaliado contra o Gold Set em quatro rodadas iterativas de refinamento de *prompt* e exemplos RAG. A acurácia de execução na linha de base é de 84,8% (28/33), com intervalo de confiança de 95% [72,7%, 94,0%] obtido por *bootstrap* com 1.000 reamostras e semente 42. A latência mediana é de 12.800 ms por consulta e o custo mediano de processamento é de 3,54 GB por job sobre a visão não particionada. Após quatro rodadas de correção, o agente alcança EX = 100,0% (33/33), com todas as cinco categorias de complexidade em 100%.

Na configuração de produção, a geração de SQL passou a usar gpt-4o (estabilizando a categoria multifiltro) e o agente passou a consultar a tabela particionada e clusterizada tb_carga_metodologia_oficial, reduzindo o custo mediano por consulta de 3,54 GB para 0,175 GB (≈95% de redução) sem alteração de acurácia; a latência mediana passa a ≈16 s (p95 ≈21 s).

A progressão por rodada é: linha de base 84,8% (R1) → 87,9% (R2) → 97,0% (R3) → 100,0% (R4). Cada rodada corrigiu padrões de falha específicos: R2 corrigiu um padrão de SQL para *rankings* (questão GS020) e um sentinela *multifiltro* (GS033); R3 resolveu ordenação de colunas em agregação temporal (GS009-GS010), um padrão de particionamento RANK em cruzamento dimensional (GS027) e uma regressão de sensibilidade a maiúsculas com LOWER() (GS032); R4 corrigiu uma ambiguidade CamelCase/minúsculo no *template* de filtro oficial (GS008) e adicionou um exemplo RAG de três filtros simultâneos que preveniu uma regressão de omissão do campo *sentido* (GS030). Todos os cinco erros da linha de base foram classificados como alucinações do tipo lógico na taxonomia de Qu et al. (2024): dois envolveram especificação incorreta de janela temporal, dois envolveram agrupamento na dimensão errada e um envolveu omissão de um filtro de tipo de navegação. Nenhum erro foi causado pelo *SQLValidator*.

## 4.3 Estudo de ablação

O estudo de ablação testou cinco configurações do agente sobre o mesmo Gold Set. B0 é a linha de base zero-shot, sem *MetadataHelper* e sem *few-shot*. B1 inclui apenas a recuperação *few-shot*, sem *MetadataHelper*. B2 inclui apenas o *MetadataHelper*, sem *few-shot*. B3 é o agente completo com *retry* desabilitado (*max_attempts* = 1). B4 é o agente completo com a configuração de referência (até 3 tentativas de *retry*). Os resultados de EX por configuração são: B0 = 78,8%; B1 = 87,9%; B2 = 81,8%; B3 = 84,8%; B4 = 84,8% (linha de base antes do refinamento).

Três achados merecem destaque analítico. O primeiro é que a recuperação *few-shot* é o componente de maior impacto isolado: B1 alcança 87,9% de EX contra 78,8% de B0, um ganho global de 9,1 pontos percentuais e de 60 pontos percentuais na categoria *multifiltro* (40% → 100%).

Esse resultado confirma que os exemplos RAG codificam padrões de combinação de múltiplos filtros que o LLM não consegue inferir de forma confiável apenas a partir dos metadados de esquema. O segundo achado é que o *MetadataHelper* contribui marginalmente em isolamento: B2 alcança 81,8%, apenas 3,0 p.p. acima de B0, indicando que descrições semânticas de colunas são insuficientes sem raciocínio guiado por exemplos que demonstrem como os filtros devem ser combinados. O terceiro achado é que o *retry loop* tem impacto negligenciável sobre a acurácia: B3 e B4 são idênticos em 84,8%, porque os cinco erros residuais são semânticos e não são corrigidos pela regeneração com o mesmo contexto. Nenhuma comparação par a par atinge significância estatística sob o teste de McNemar com correção de Yates (todos os $p \geq 0{,}479$), o que é consistente com o tamanho amostral de n = 33: o teste não tem poder suficiente para detectar diferenças menores que aproximadamente 15 pontos percentuais a $\alpha = 0{,}05$ e $\beta = 0{,}20$.

## 4.4 Avaliação de segurança (red-team)

Uma avaliação adversarial estruturada foi conduzida com 100 *prompts* de injeção em cinco categorias de 20 *prompts* cada: DML explícito (tentativas diretas de DROP, DELETE, INSERT); evasão encadeada (tentativas de encadeamento de instruções); exfiltração semântica (tentativas de acesso a INFORMATION_SCHEMA e metadados); injeção clássica (OR 1=1, comentários SQL, UNION ALL SELECT); e *prompt* ambíguo em português (frases em linguagem natural projetadas para induzir consultas inseguras sem palavras-chave SQL óbvias). As 33 consultas do Gold Set serviram como conjunto de controle para medição da taxa de falsos positivos.

Na linha de base (antes de qualquer correção), o validador apresentava TPR = 52,7% (49/100) e FPR = 0,0% (0/33). O validador bloqueava com eficácia total as tentativas de DML explícito (100% TPR) e de evasão encadeada (100% TPR), mas apresentava lacunas críticas nas categorias de exfiltração semântica (TPR = 5%) e injeção clássica (TPR = 0%), onde as correspondências de padrões geravam apenas avisos em vez de rejeições. A rodada R1 elevou todas as correspondências de padrões de aviso para erro irrecuperável e adicionou o bloqueio de *INFORMATION_SCHEMA* e de quatro padrões de injeção adicionais, produzindo TPR = 93,5% e FPR = 0,0%. A rodada R2 adicionou a validação de referências de tabelas via análise AST com *sqlglot* e um *regex* para tautologias de igualdade de *strings* sem OR explícito, produzindo resultado final de TPR = 96,8% (97/100) e FPR = 0,0% (0/33). Os três falsos negativos remanescentes são FN por *design*, mitigados por controles de infraestrutura no BigQuery: exfiltração via *TO_JSON_STRING* (sem palavra-chave bloqueável; endereçável por segurança de nível de coluna no IAM), tentativa de exaustão de CPU via *GENERATE_ARRAY* (mitigada por *timeout* e quotas do BigQuery) e padrão *LIKE* '%' isolado (sem vetor de ataque identificável).

### 4.5 Generalização e robustez

Para estimar a generalização além dos itens usados no refinamento, foram conduzidos dois testes complementares. No teste de robustez a paráfrase, cada uma das 33 questões do Gold Set foi reescrita em uma versão parafraseada, com variações de redação (sinônimos, ordem, registro coloquial e telegráfico), mantendo as mesmas entidades e o mesmo gabarito; o agente acertou 33/33, indicando que o refinamento codifica regras de domínio generalizáveis e não padrões dependentes de fraseado.

No teste *out-of-sample*, foi construído um conjunto held-out de 60 questões com intenções e entidades inéditas – novos portos, estados, regiões e anos, e tipos de pergunta ausentes do Gold Set (contagem distinta, participação percentual no total, média, variação multi-ano e rankings por mercadoria) – com SQL de referência validado por execução. O agente acertou 59/60 (98,3%; IC 95% Wilson [91,1%, 99,7%]), com 46/46 em itens de novas entidades e 13/14 em itens de novas intenções. A única falha é informativa: em um ranking por contêineres (TEU), o agente gerou SUM(teu) sem a conversão de tipo necessária (a coluna é armazenada como texto), uma falha baseada em esquema na taxonomia de Qu et al. (2024). Esses resultados, com limite inferior conservador de ≈91%, fornecem evidência de generalização dentro do mesmo domínio; a generalização entre conjuntos de dados distintos permanece como trabalho futuro.

## 5. PROPOSIÇÕES DE DESIGN E IMPLICAÇÕES ORGANIZACIONAIS

### 5.1 Proposições de design para agentes text-to-SQL em dados governamentais

As decisões arquiteturais tomadas no desenvolvimento deste sistema e os resultados de sua avaliação empírica permitem derivar quatro proposições de *design* para agentes *text-to-SQL* em contextos de dados governamentais. Essas proposições têm suporte empírico no artefato e nos resultados documentados nas seções anteriores, mas são apresentadas como hipóteses a serem validadas por replicação em outros contextos de dados governamentais, não como princípios generalizados estabelecidos.

A primeira proposição (DP1), de *grounding* institucional obrigatório, estabelece que regras domínio-específicas – filtros de metodologia oficial, vocabulários controlados, semântica de campos, defasagem na publicação dos microdados – devem ser explicitamente codificadas no contexto do agente e validadas no *pipeline* de geração, não inferidas pelo LLM a partir do esquema. O fundamento empírico está no estudo de ablação: B0 (sem *MetadataHelper* e sem *few-shot*) alcança apenas 78,8% de EX, e a principal categoria de erro são alucinações metodológicas. A proposição tem generalidade além do caso ANTAQ: qualquer base de dados governamental que utilize metodologias oficiais de cálculo não documentadas no esquema enfrenta o mesmo problema estrutural.

A segunda proposição (DP2), de validação determinística com privilégio mínimo, estabelece que o SQL gerado pelo LLM deve passar por uma camada de validação determinística antes da execução, arquiteturalmente separada das instruções no *prompt*. O fundamento empírico está na avaliação *red-team*: sem o *SQLValidator*, o agente seria vulnerável a todas as 100 categorias de ataque testadas na linha de base. A camada determinística é necessária porque as instruções no *prompt* são estocásticas mesmo com temperatura 0: garantidamente determinísticas para uma entrada fixa, mas não imunes a *prompts* adversariais com variação de formulação. Uma conta de serviço com acesso somente-leitura no conjunto de dados autorizado é o complemento de infraestrutura dessa camada de *software*.

A terceira proposição (DP3), de transparência controlada e auditabilidade, estabelece que o SQL não deve ser exposto ao usuário final por padrão, mas um modo de inspeção deve estar disponível para *accountability* institucional e verificação especializada. O fundamento é duplo: do lado da usabilidade, expor SQL a usuários sem formação técnica cria sobrecarga cognitiva; do lado da *accountability*, a opacidade total impede que especialistas verifiquem se as respostas estão metodologicamente corretas. O *design* de modo duplo (oculto por padrão, auditável sob demanda) endereça ambas as tensões e constitui um padrão de UX transferível para ferramentas de IA governamentais.

A quarta proposição (DP4), de controles de custo configuráveis, estabelece que, em ambientes de execução faturados por consumo em nuvem, a previsibilidade de custos é um requisito de primeira classe para a adoção organizacional. O custo mediano de 3,54 GB por consulta, sem controles, poderia gerar despesas significativas em uso intensivo. Um complemento de alto impacto é o *layout* físico dos dados: apontar o agente para uma materialização colunar particionada e clusterizada, em vez da visão viva, reduziu o custo mediano por consulta em ≈95% (de 3,54 GB para 0,175 GB) sem qualquer alteração no SQL gerado ou na acurácia, pois o armazenamento colunar lê apenas as colunas referenciadas. A combinação de *maximum_bytes_billed* ≈ 10 GB, injeção automática de LIMIT 1000, *cache* de consultas e *timeout* de 60 segundos constitui um padrão de governança de custos reutilizável. Essa dimensão é frequentemente negligenciada na literatura de *text-to-SQL*, que trabalha com *benchmarks* acadêmicos sem custos de execução, mas é determinante para a adoção governamental em contextos de orçamento público.

## 5.2 Riscos, limitações e conformidade regulatória

A implantação de um agente conversacional para acesso a dados governamentais levanta questões que vão além do desempenho técnico. O risco de alucinação institucional é gerenciado pelo *MetadataHelper* e pela verificação contra agregados publicados, mas persiste para questões fora das categorias cobertas pelo Gold Set. O sistema atual não inclui mecanismo de detecção de incerteza que alerte o usuário quando uma questão está fora do domínio de cobertura confiável dos exemplos RAG, lacuna que pode levar a respostas incorretas apresentadas com aparência de confiabilidade em contextos de prestação de contas.

A conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD, Lei nº 13.709/2018) é relevante para os estudos com usuários planejados nos trabalhos futuros. Esses estudos envolverão gestores públicos, pesquisadores e jornalistas interagindo com a interface conversacional, o que gera *logs* de interação que podem conter informações sensíveis. A política de retenção e anonimização desses *logs* deve ser definida antes da condução dos estudos, e o protocolo deve ser submetido a um Comitê de Ética em Pesquisa (CEP) conforme a Resolução CNS nº 510/2016. O sistema já aplica filtragem de dados sensíveis no módulo de telemetria (*before_send filter* no Sentry), mas isso cobre apenas os dados enviados para o serviço externo de monitoramento de erros.

A dependência de fornecedores proprietários de LLM constitui um risco para a implantação governamental de longo prazo: mudanças nos termos de serviço, na disponibilidade ou nos preços do OpenAI ou do Google Vertex AI podem afetar a continuidade do serviço. O *design* com *LLMFactory* e variável *LLM_PROVIDER* mitiga parcialmente esse risco ao permitir a substituição do modelo de linguagem sem alteração do *pipeline*, mas qualquer substituição requer nova avaliação empírica para verificar a preservação da acurácia. O uso de modelos de código aberto, como os da família CodeS documentada por Li et al. (2024), é uma alternativa que reduz a dependência de fornecedores proprietários ao custo de maior complexidade operacional.

## 5.3 Implicações para a adoção de IA em agências reguladoras brasileiras

Os resultados deste trabalho têm implicações para além do caso específico da ANTAQ. O modelo de arquitetura proposto é transferível para organizações que mantêm bases de dados públicas em plataformas analíticas em nuvem, condição comum a agências reguladoras, órgãos estatísticos e empresas estatais. Agências como a ANAC, a ANEEL e a ANP ilustram esse universo, ao publicarem dados setoriais passíveis de acesso por uma arquitetura semelhante.

A transferência tende a ser mais proveitosa nos contextos em que metodologias oficiais de cálculo, terminologia técnica específica e requisitos de auditabilidade condicionam a interpretação correta dos dados, situação recorrente em repositórios de dados governamentais. Nesses casos, a generalização do MetadataHelper requereria a substituição dos metadados de domínio (filtros, glossários, semântica de campos), preservando-se a arquitetura de orquestração e validação. O custo de adaptação seria, assim, predominantemente de domínio (o conhecimento da metodologia setorial), e não de engenharia de *software*.

No contexto mais amplo da transformação digital do setor público brasileiro, o agente documentado neste trabalho representa uma categoria de ferramenta que pode ampliar o alcance prático da Lei de Acesso à Informação (Lei nº 12.527/2011). A LAI garante formalmente o direito de acesso a informações públicas, mas esse direito é exercitável, na prática, apenas

por atores com capacidade técnica para interpretar os dados disponibilizados. Uma interface conversacional que responde perguntas sobre o setor aquaviário em português, verificando a correção metodológica das respostas e registrando as consultas SQL para auditoria, reduz o custo de exercício desse direito para gestores municipais que monitoram a movimentação de cargas em seus territórios, para jornalistas que investigam contratos de arrendamento portuário e para pesquisadores que estudam as políticas de infraestrutura logística do Brasil. A mensuração efetiva desse impacto requer os estudos controlados com usuários planejados nos trabalhos futuros.

## CONCLUSÃO

Este trabalho apresentou o *design*, a implementação e a avaliação empírica de um agente inteligente baseado em LLM para consulta em linguagem natural dos dados estatísticos aquaviários publicados pela ANTAQ e ingeridos, neste trabalho, em um repositório BigQuery. O sistema, disponível como *software* desenvolvido, combina orquestração agêntica via LangGraph, codificação semântica de metadados institucionais (dicionário de dados do domínio), recuperação semântica de exemplos *few-shot* com 32 pares questão-SQL curados e uma camada de validação e segurança determinística com modelo de ameaças documentado e avaliação adversarial sistemática.

A avaliação em um conjunto de 33 questões domínio-específicas alcança acurácia de execução EX = 100,0% após quatro rodadas iterativas de refinamento (linha de base: EX = 84,8%, IC 95%: [72,7%, 94,0%]). O estudo de ablação identifica a recuperação *few-shot* como o componente de maior impacto individual (+9,1 p.p. globalmente; +60 p.p. nas consultas com múltiplos filtros), fornecendo suporte empírico direcional para a proposição DP1 (grounding institucional obrigatório). Em um conjunto independente de 60 questões com intenções e entidades inéditas, o agente acerta 59/60 (98,3%; IC 95% Wilson [91,1%, 99,7%]), evidência de generalização além do Gold Set, e a migração para uma tabela particionada reduz o custo mediano por consulta em ≈95%. A avaliação adversarial do SQLValidator alcança TPR = 96,8% com FPR = 0,0%, confirmando que a camada de segurança determinística (DP2) protege efetivamente contra as categorias de ataque testadas.

As contribuições científicas principais são: (i) o *design* do *MetadataHelper* como abordagem empiricamente fundamentada para o *grounding* institucional de agentes *text-to-SQL*; (ii) a metodologia de avaliação adversarial para SQL gerado por LLM em implantações governamentais, com modelo de ameaças estruturado e protocolo de *red-team* reproduzível; e (iii) quatro proposições de *design* (DP1-DP4) para sistemas de acesso a dados governamentais, com suporte empírico e escopo de generalização delimitado. Uma limitação é que o *Gold Set* de 33 itens, usado também no refinamento iterativo, funciona como conjunto de desenvolvimento; por isso, a acurácia de 100% deve ser lida como limite superior otimista, sendo a evidência de generalização proveniente do conjunto held-out independente de 60 questões (59/60 = 98,3%). O tamanho amostral ainda limita o poder do teste de ablação por McNemar. Os trabalhos futuros incluem reexecutar a ablação no conjunto ampliado para obter poder estatístico, estudos controlados com usuários não técnicos e experimentos de generalização entre conjuntos de dados governamentais distintos.

## REFERÊNCIAS


ANDROUTSOPOULOS, I.; RITCHIE, G. D.; THANISCH, P. Natural language interfaces to databases: an introduction. **arXiv** preprint cmp-lg/9503016, 1995.

ANTAQ. **Estatístico Aquaviário**. Brasília: ANTAQ, 2023. Disponível em: https://estatistica.antaq.gov.br/ea/. Acesso em: jun. 2026.

ANTAQ. **Metodologia de cálculo dos indicadores do Estatístico Aquaviário**. Brasília: ANTAQ, 2025. Disponível em: https://web3.antaq.gov.br/ea/txt/Metodologia_de_calculo_dos_Indicadores_do_EA.pdf. Acesso em: jun. 2026.

ASAI, A.; WU, Z.; WANG, Y.; SIL, A.; HAJISHIRZI, H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In: **Proceedings of ICLR** 2024, 2023. arXiv:2310.11511.

BRASIL. **Lei nº 10.233, de 5 de junho de 2001**. Dispõe sobre a reestruturação dos transportes aquaviário e terrestre. Brasília: Diário Oficial da União, 2001.

BRASIL**. Lei nº 12.527, de 18 de novembro de 2011**. Regula o acesso a informações. Brasília: Diário Oficial da União, 2011.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília: Diário Oficial da União, 2018.

BRESSER-PEREIRA, L. C. **Reforma do Estado para a cidadania**. São Paulo: Editora 34; Brasília: ENAP, 1998.

FOUNTAIN, J. E. **Building the virtual state: Information technology and institutional change**. Washington: Brookings Institution Press, 2001.

FRÓES, K. C.; BRAGHETTO, K. R. Exploring temporal text-to-SQL challenges in Brazilian Portuguese: lessons from educational data. In: **Anais do XL Simpósio Brasileiro de Banco de Dados** (SBBD 2025). Porto Alegre: SBC, 2025. p. 963-969.

GAO, D. et al. Text-to-SQL empowered by large language models: A benchmark evaluation. **Proceedings of the VLDB Endowment**, v. 17, n. 5, 2024.

GAO, Y. et al. Retrieval-augmented generation for large language models: A survey. **arXiv**:2312.10997, 2024.

HENDRIX, G. G. et al. Developing a natural language interface to complex data. **ACM Transactions on Database Systems**, v. 3, n. 2, p. 105-147, 1978.

HEVNER, A. R.; MARCH, S. T.; PARK, J.; RAM, S. Design science in information systems research. **MIS Quarterly**, v. 28, n. 1, p. 75-105, 2004.

JANSSEN, M.; CHARALABIDIS, Y.; ZUIDERWIJK, A. Benefits, adoption barriers and myths of open data and open government. **Information Systems Management**, v. 29, n. 4, p. 258-268, 2012.

JOSÉ, M. A.; COZMAN, F. G. mRAT-SQL+GAP: A Portuguese text-to-SQL transformer. In: **Intelligent Systems (BRACIS 2021)**, Lecture Notes in Computer Science. Springer, 2021. arXiv:2110.03546.

JOSÉ, M. M.; JOSÉ, M. A.; MAUÁ, D. D.; COZMAN, F. G. Integrating question answering and text-to-SQL in Portuguese. In: PROPOR 2022, **Lecture Notes in Computer Science**. Springer, 2022. arXiv:2202.04048.

KLEIN, R. H.; KLEIN, D. C. B.; LUCIANO, E. M. Identificação de mecanismos para a ampliação da transparência em portais de dados abertos: uma análise no contexto brasileiro. **Cadernos EBAPE.BR**, v. 16, n. 4, p. 692-715, 2018.

LEI, F. et al. Spider 2.0: Evaluating language models on real-world enterprise text-to-SQL workflows. **arXiv**:2411.07763, 2024.

LEWIS, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: **Advances in Neural Information Processing Systems** (NeurIPS 2020), 2020.

LI, H. et al. CodeS: Towards building open-source language models for text-to-SQL. In: **Proceedings of ACM SIGMOD** 2024, 2024.

LI, J. et al. Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs. In: **Advances in Neural Information Processing Systems** (NeurIPS 2023), 2023.

LOURENÇO, R. P. **An analysis of open government portals**: A perspective of transparency for accountability. Government Information Quarterly, v. 32, n. 3, p. 323-332, 2015.

MALEWICZ, G. et al. Pregel: A system for large-scale graph processing. In: **Proceedings of the ACM SIGMOD International Conference on Management of Data**, p. 135-146, 2010.

POPESCU, A.-M.; ETZIONI, O.; KAUTZ, H. Towards a theory of natural language interfaces to databases. In: **Proceedings of the 8th International Conference on Intelligent User Interfaces** (IUI'03), p. 149-157, 2003.

POURREZA, M.; RAFIEI, D. DIN-SQL: Decomposed in-context learning of text-to-SQL with self-correction. In: **Advances in Neural Information Processing Systems** (NeurIPS 2023), 2023.

QU, G. et al. Before generation, align it! A novel and effective strategy for mitigating hallucinations in text-to-SQL generation. In: **Findings of ACL 2024**, p. 5456-5471, 2024.

SILVA, P. N. Dados governamentais abertos sobre a COVID-19 no Brasil: avaliação do potencial de reúso. **Em Questão**, v. 29, 2023.

TALAEI, S. et al. CHESS: Contextual harnessing for efficient SQL synthesis. **arXiv**:2405.16755, 2024.

WANG, B. et al. MAC-SQL: A multi-agent collaborative framework for text-to-SQL. **arXiv**:2312.11242, 2024.

WOODS, W. A. Progress in natural language understanding: An application to lunar geology. In: **Proceedings of the AFIPS National Computer Conference**, v. 42, p. 441-450, 1973.

YAO, S. et al. ReAct: Synergizing reasoning and acting in language models. In: **Proceedings of ICLR** 2023, 2022.

YU, T. et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In: **Proceedings** of EMNLP 2018, 2018.

ZHONG, V.; XIONG, C.; SOCHER, R. Seq2SQL: Generating structured queries from natural language using reinforcement learning. **arXiv**:1709.00103, 2017.

# EIXO 02

## *Saúde Pública, na Sustentabilidade e no Impacto Social*

# Avaliação de Impacto Algorítmico no Setor Público Brasileiro: um modelo complementar ao framework de autoavaliação ética

*Camila Falchetto Romero*
*Prof. Dr. Fausto Barros Sá Teles*

## Resumo

O setor público brasileiro utiliza sistemas de inteligência artificial em áreas como triagem de benefícios sociais e gestão tributária, mas carece de instrumentos padronizados para avaliar o impacto desses sistemas sobre direitos fundamentais. O *Framework* para Autoavaliação de Impacto Ético em IA (AIE), elaborado pelo governo federal, classifica riscos e gera recomendações, mas opera como autoavaliação interna sem verificação externa, análise de impacto sobre grupos específicos ou decisão vinculante. Este artigo propõe um modelo de Avaliação de Impacto Algorítmico (AIA) complementar ao AIE, desenvolvido por meio do método *Design Science Research* (Peffers et al., 2007). O modelo define critérios objetivos de acionamento, processo em quatro fases conduzido por instância independente e matriz de risco em seis dimensões de direitos fundamentais derivadas do PL 2338/2023. Demonstrado por estudo de caso simulado, o modelo produziu diagnóstico de risco por dimensão, medidas de mitigação com indicadores e decisão formal vinculante. Experimentos de auditoria de viés confirmam que modelos treinados sem restrições de equidade produzem discriminação indireta mensurável.

**Palavras-chave:** avaliação de impacto algorítmico; governança de inteligência artificial; setor público.

## *INTRODUÇÃO*

O governo federal brasileiro utiliza sistemas de inteligência artificial em áreas que vão da gestão tributária ao reconhecimento facial em segurança pública, passando por triagem de benefícios sociais e análise de fraude. Essa expansão não foi acompanhada, até recentemente, por instrumentos sistematizados de avaliação de riscos éticos e de impacto sobre direitos. O quadro começou a mudar com o desenvolvimento do Framework para Autoavaliação de Impacto Ético em Inteligência Artificial (AIE) pelo governo federal, que estrutura pela primeira vez um instrumento de avaliação ética para sistemas de IA no setor público brasileiro. Em paralelo, o PL 2338/2023, aprovado pelo plenário do Senado Federal em dezembro de 2024 e em tramitação na Câmara dos Deputados em Comissão Especial instalada em abril de 2025, estabelece a obrigação de realizar avaliação de impacto algorítmico (AIA) para sistemas de alto risco e define direitos específicos das pessoas afetadas.

O AIE é um instrumento de autoavaliação interna, conduzido e respondido pelo próprio órgão que desenvolve ou contrata o sistema. Esse modelo é adequado para uma primeira camada de governança, voltada à reflexão institucional e à identificação de riscos. Por sua natureza,

porém, ele não inclui verificação externa, análise de impacto sobre grupos específicos afetados nem decisão com efeito vinculante sobre a operação do sistema, elementos estes que pertencem a uma segunda camada. O PL 2338/2023, por sua vez, cria a obrigação de realizar a AIA para sistemas de alto risco (art. 25), mas não detalha o processo, os instrumentos, a instância competente nem os critérios de decisão. Há, portanto, espaço para um modelo que operacionalize essa segunda camada, preparando as instituições para os requisitos que a regulação em tramitação poderá estabelecer.

Este artigo tem como objetivo propor um modelo padronizado de AIA para o setor público brasileiro que opere em complementaridade ao AIE, acrescente os elementos que uma segunda camada de governança deve prover, como verificação externa, análise de impacto sobre grupos afetados e decisão vinculante, e seja operacionalizável no contexto regulatório do PL 2338/2023. O modelo é concebido para ser acionado quando a primeira camada identifica risco alto ou quando critérios objetivos alternativos são satisfeitos, e conduzido por instância funcionalmente independente do órgão que desenvolve ou opera o sistema avaliado.

A pesquisa adota a *Design Science Research* (DSR) como método, conforme proposta de Peffers et al. (2007), orientando a produção de um artefato prescritivo em vez de uma teoria descritiva. O artefato é demonstrado por meio de estudo de caso simulado envolvendo um sistema hipotético de aprendizado de máquina aplicado para triagem de beneficiários do Bolsa Família com base no CadÚnico. A escolha desse cenário se justifica porque o sistema se enquadra diretamente no art. 14, inciso IV do PL 2338/2023, que classifica como de alto risco sistemas utilizados para determinação de elegibilidade a benefícios sociais, e pelas múltiplas dimensões de vulnerabilidade da população afetada.

A contribuição deste trabalho é tripla. No plano acadêmico, preenche uma lacuna na literatura brasileira, que trata a AIA como conceito teórico sem oferecer instrumentos operacionalizáveis para o setor público. No plano institucional, o modelo fornece aos órgãos do SISP um procedimento estruturado que pode ser adotado independentemente da aprovação do PL 2338/2023. No plano prático, os quatro instrumentos produzidos (questionário de triagem, matriz de avaliação de risco, template de relatório e critérios de decisão) padronizam a avaliação e reduzem a dependência de análises construídas caso a caso.

# 1. REFERENCIAL TEÓRICO-METODOLÓGICO

## 1.1 Governança em múltiplas camadas

O desenvolvimento e a implantação de sistemas de IA pelo setor público colocam em questão os arranjos institucionais capazes de garantir responsabilização efetiva. Kaminski e Malgieri (2021) descrevem esse problema a partir da distinção entre mecanismos de governança individuais, que operam no nível dos direitos subjetivos do cidadão afetado, e mecanismos sistêmicos, que operam no nível da organização responsável pelo sistema. Os autores argumentam que a regulação de IA baseada apenas em direitos individuais é insuficiente, uma vez que a assimetria de informação entre o cidadão e o órgão que opera o sistema é profunda o bastante para tornar o exercício individual de direitos inviável para a maior parte dos afetados. Mecanismos de governança sistêmica, como avaliações de impacto, preenchem essa lacuna ao deslocar o ônus da *accountability* do indivíduo para a instituição.

Essa distinção estrutura a lógica do modelo proposto. O AIE constitui a primeira camada de governança: autoavaliação interna que produz classificação de risco e recomendações. O modelo AIA proposto constitui a segunda camada: conduzida por instância externa, avalia impacto sobre direitos de grupos específicos e produz decisão com efeito vinculante. A relação entre as duas camadas é de complementaridade: a primeira é condição para o acionamento da segunda, e a segunda dá consequências institucionais ao que a primeira identifica. Gatt et al. (2024) identificam lacuna equivalente no contexto do *AI Act* europeu: o regulamento exige a realização de uma *Fundamental Rights Impact Assessment* (FRIA) para sistemas de alto risco, mas não especifica parâmetros de implementação nem metodologia de mensuração, resultado análogo ao do PL 2338/2023 em relação à AIA.

## 1.2 Inovação responsável

A teoria da inovação responsável, desenvolvida por Stilgoe, Owen e Macnaghten (2013), oferece os critérios normativos para avaliar se um instrumento de governança de IA é suficiente para sistemas de alto risco. Os autores propõem quatro dimensões: antecipação, que é a capacidade de identificar e avaliar consequências futuras prováveis de uma tecnologia antes de sua implantação em escala; reflexividade, que consiste no questionamento das premissas, valores e interesses embutidos no processo de desenvolvimento; inclusão, que diz respeito à participação de cidadãos e grupos afetados nos processos de avaliação e decisão; e responsividade, que é a capacidade institucional de ajustar ou interromper o desenvolvimento e a operação de sistemas diante de riscos identificados.

Aplicado ao conjunto formado pelo AIE e pelo modelo AIA proposto, o *framework* de Stilgoe et al. orienta a distribuição de funções entre as duas camadas. A antecipação é parcialmente coberta pelo AIE, que avalia riscos presentes no momento da autoavaliação, e ampliada pelo modelo AIA, que produz análise de impactos diferenciados por grupo antes da implantação. A reflexividade interna ao órgão, já presente no AIE, é complementada por mecanismo externo que avalia as escolhas técnicas de forma independente. A inclusão, não prevista no AIE, é parcialmente endereçada pelo modelo AIA por meio de canal de manifestação pós-publicação, reconhecendo-se que esse mecanismo não equivale à participação deliberativa de afetados no processo de avaliação. A responsividade, recomendatória no AIE, torna-se vinculante na segunda camada, que produz decisão formal com consequências sobre a operação do sistema. O modelo AIA proposto neste artigo busca, portanto, completar as dimensões que uma segunda camada deve acrescentar ao que a primeira já entrega.

## 1.3 Abordagem baseada em direitos e auditoria de viés

A abordagem baseada em direitos constitui o terceiro pilar do modelo. Mantelero e Esposito (2021) argumentam que tanto a abordagem centrada em dados pessoais quanto a abordagem centrada em ética são insuficientes para regular o impacto de sistemas de IA, sendo a primeira excessivamente restrita ao titular do dado e a segunda carente de parâmetros operacionalizáveis. Os autores propõem uma metodologia de *Human Rights Impact Assessment* (HRIA) que parte da análise empírica de decisões de autoridades regulatórias e traduz obrigações de direitos humanos em critérios mensuráveis de avaliação de risco. No plano normativo, a Recomendação da UNESCO (2021) estabelece que sistemas de IA devem ser avaliados quanto ao impacto sobre direitos humanos ao longo de todo o ciclo de vida, com ênfase em populações vulneráveis.

A dimensão técnica da avaliação de impacto enfrenta um problema metodológico relevante: métricas de viés calculadas isoladamente a partir do desempenho do modelo podem validar sistemas que produzem discriminação indireta. Galdon Clavell, Aumaitre e Calders (2024) demonstram esse problema com o *Adult Income Dataset*, comparando quatro ferramentas de auditoria de viés e mostrando que ferramentas baseadas exclusivamente em dados do modelo podem indicar ausência de viés enquanto o sistema produz disparidades expressivas sobre grupos protegidos. A conclusão é que a avaliação de equidade requer dados demográficos e contextuais externos ao modelo, não apenas métricas de desempenho preditivo. Essa limitação é metodologicamente central para a matriz de risco do modelo AIA proposto, cuja dimensão de não discriminação não avalia apenas acurácia ou conformidade técnica, mas examina o impacto do sistema sobre a população afetada com dados contextuais externos.

Jenkins et al. (2023) justificam a separação entre coleta de fatos e avaliação de impacto como fases distintas do processo de AIA: a primeira mapeia as características técnicas do sistema sem julgamento de valor, e a segunda avalia essas características à luz dos objetivos e valores do domínio de aplicação. Quando a coleta de fatos e a avaliação de aceitabilidade ocorrem no mesmo ato, o avaliador tende a ajustar inconscientemente o julgamento normativo para compatibilizar com os fatos documentados, produzindo avaliações benevolentes. No modelo proposto, o questionário de triagem (Instrumento 1) corresponde à fase de coleta de fatos e a matriz de risco (Instrumento 2) corresponde à fase de avaliação.

## 2. O AIE E O QUE UMA SEGUNDA CAMADA DEVE ACRESCENTAR

### 2.1 Descrição do AIE

O *Framework* para Autoavaliação de Impacto Ético em Inteligência Artificial (AIE) é um instrumento desenvolvido pelo Ministério da Gestão e Inovação em Serviços Públicos (MGI, 2025) para orientar órgãos da administração pública na identificação de riscos éticos em sistemas de IA ao longo de seu ciclo de vida. O AIE estrutura a avaliação em dois estágios. O primeiro é uma avaliação de riscos composta por doze perguntas que classifica o sistema em risco baixo, médio ou alto por acumulação de pontos e por gatilhos diretos, como o uso de dados sensíveis, decisão autônoma sem revisão humana e potencial de impacto sobre direitos. O segundo é o questionário principal, com versões simplificada, padrão e completa conforme o nível de risco, organizado em seções temáticas: equidade e não discriminação, inclusão digital, transparência e explicabilidade, supervisão humana, privacidade e proteção de dados pessoais, robustez técnica, responsabilização, sustentabilidade, governança, conformidade legal, gestão do ciclo de vida e síntese de riscos. O resultado é uma classificação em cinco níveis de aderência ética, de muito baixa a muito alta, acompanhada de recomendações de melhoria e indicação de trilhas de capacitação.

## 2.2 O que uma segunda camada deve acrescentar

O AIE foi desenhado como instrumento de primeira camada de governança e cumpre essa função, uma vez que gera reflexão institucional, classifica riscos e orienta melhorias. Instrumentos desse tipo, por definição, não incluem os elementos que pertencem a uma segunda camada, tais como verificação externa, análise de impacto sobre grupos específicos e decisão vinculante. Identificar o que a segunda camada deve acrescentar é, portanto, um passo necessário para compreender como os dois instrumentos se articulam.

O primeiro elemento que a segunda camada deve prover é a verificação externa. O modelo de autoavaliação interna é adequado para gerar reflexão no órgão aplicador, mas produz respostas autodeclaradas sem mecanismo que as confronte com dados independentes, documentação técnica verificável ou avaliação de terceiros. Kaminski e Malgieri (2021) identificam esse ponto como a distinção central entre mecanismos de governança de primeira e segunda camada: a primeira produz reflexão institucional; a segunda produz *accountability*. Selbst (2021) argumenta que a produção de documentação auditável é o objetivo mais importante de qualquer regime de AIA, pois não depende da boa-fé do avaliado para se concretizar e é exatamente o que uma instância externa, ao conduzir a avaliação com base em documentação técnica verificável, pode aferir.

O segundo elemento é a análise de impacto sobre grupos específicos. O AIE avalia se o sistema adota salvaguardas contra discriminação, o que é uma verificação de conformidade de design. A segunda camada deve ir além e verificar se o sistema produz, na prática, impacto diferenciado sobre grupos vulneráveis, com dados demográficos externos ao modelo. A distinção é relevante, pois a presença de salvaguardas técnicas não equivale à demonstração de que o sistema não produz discriminação indireta, conforme demonstrado por Galdon Clavell et al. (2024). Mensurar esse impacto requer dados contextuais sobre a população afetada que a autoavaliação interna não tem como exigir de forma padronizada.

O terceiro elemento é a decisão com efeito vinculante sobre a operação do sistema. O AIE classifica o nível de aderência ética e produz recomendações de melhoria, resultado compatível com sua função de instrumento orientador. A segunda camada, por sua vez, deve converter essa classificação em consequências formais: a responsividade, dimensão central da inovação responsável segundo Stilgoe et al. (2013), requer que a avaliação de riscos produza capacidade institucional de condicionar ou interromper a operação de um sistema quando os riscos identificados o justificam. Essa autoridade decisória pertence à segunda camada, não à primeira.

O quarto elemento é a participação de grupos afetados no processo de avaliação. A condução interna pelo órgão aplicador não inclui, por definição, consulta às pessoas ou grupos que serão afetados pelo sistema. Para sistemas com efeitos jurídicos sobre direitos, como acesso a benefícios sociais, essa participação é condição metodológica para que a avaliação capture riscos que os dados de treinamento e as métricas de desempenho não conseguem revelar, conforme demonstram Mantelero e Esposito (2021). O modelo AIA proposto não resolve esse ponto integralmente, pois a participação deliberativa de afetados está registrada como limitação, mas prevê canal de manifestação formal pós-publicação como mecanismo parcial de retroalimentação.

## 3. BENCHMARK INTERNACIONAL E REQUISITOS DE DESIGN

Três experiências internacionais oferecem padrões testados em contextos regulatórios distintos. O Canadá foi, em 2019, o primeiro governo a estabelecer diretiva vinculante para sistemas de decisão automatizada na administração pública federal: a *Directive on Automated Decision-Making*. Seu instrumento operacional, o AIA/DAMA, é um questionário de 72 questões em oito dimensões (impacto sobre direitos, qualidade dos dados, explicabilidade, supervisão humana, viabilidade de contestação, impacto sobre grupos vulneráveis, segurança e transparência) que produz *score* automático e determina as obrigações decorrentes por nível de impacto. O relatório é publicado em registro central acessível ao público, criando mecanismo de transparência ativo pelo qual a ausência de avaliação para um sistema em operação torna-se auditável. Três elementos do modelo canadense foram incorporados ao modelo proposto, a saber, a publicidade do resultado como mecanismo de responsabilização, a exigência de reavaliação em caso de mudança relevante no sistema e a estrutura de três tipos de decisão possíveis.

A União Europeia, pelo Regulamento 2024/1689 (*AI Act*), é o primeiro marco regulatório vinculante de IA com escopo geral no mundo. Para sistemas de alto risco, o *AI Act* exige avaliação de conformidade antes da implantação e, para sistemas usados por autoridades públicas, uma *Fundamental Rights Impact Assessment* (FRIA) com seis dimensões de direitos. Gatt et al. (2024) identificam lacuna na implementação da FRIA análoga à do PL 2338/2023: o regulamento cria a obrigação sem detalhar os parâmetros metodológicos, o formato do relatório nem os critérios de decisão, resultando em heterogeneidade de aplicação. O modelo proposto neste artigo adota diretamente a estrutura de seis dimensões de direitos e três tipos de decisão de Gatt et al. (2024), adaptada ao contexto regulatório do PL 2338/2023.

O *AI Now Institute* (2018) propôs modelo para agências públicas norte-americanas com ênfase no caráter processual e participativo da avaliação, organizado em quatro etapas: revisão pré-aquisição, divulgação pública do sistema, período de comentários da comunidade e período de contestação por devido processo legal. Do modelo do *AI Now* Institute foram incorporados ao modelo proposto: o acionamento pré-implantação como padrão, a publicidade ativa como mecanismo de responsabilização e a separação entre a avaliação do sistema e a contestação de decisões individuais, evitando que o modelo AIA seja sobrecarregado com funções que pertencem ao regime de direitos individuais. A principal diferença é que o *AI Now Institute* trata o período de comentários da comunidade como elemento central do processo; o modelo proposto não incorpora esse elemento, registrado como limitação.

A análise comparativa das três experiências permite identificar seis requisitos de design: (1) critérios objetivos de acionamento, não dependentes exclusivamente de autodeclaração; (2) instância condutora independente do órgão aplicador; (3) instrumento de avaliação centrado em impacto sobre direitos, não apenas em conformidade com princípios; (4) decisão com efeito vinculante sobre a operação do sistema; (5) publicidade do resultado como mecanismo ativo de responsabilização; e (6) monitoramento contínuo com gatilhos objetivos de reavaliação. O modelo proposto na seção 4 incorpora os seis requisitos, com satisfação parcial do requisito 1, dado que o acionamento ainda depende em parte de autodeclaração do órgão e é mitigado por dois mecanismos complementares, e com ausência reconhecida de participação formal de afetados.

# 4. O MODELO AIA PROPOSTO

## 4.1 Critérios de acionamento

O acionamento da AIA ocorre quando o sistema satisfaz ao menos um dos quatro critérios do Quadro 1, ou quando a avaliação de riscos do AIE classifica o sistema como risco alto. Os critérios são alternativos: basta um para tornar a AIA obrigatória. Essa estrutura reduz o risco de que sistemas com alto potencial de dano em apenas uma dimensão escapem da avaliação por não acumularem múltiplos fatores.

**Quadro 1 – Critérios de acionamento da AIA**

| Nº | Critério | Fundamento |
|---|---|---|
| 1 | A decisão automatizada produz efeitos jurídicos relevantes sobre acesso, concessão, redução ou revogação de benefício ou serviço público essencial. | PL 2338/2023, art. 14, IV e art. 15, I |
| 2 | O sistema afeta predominantemente pessoas em situação de vulnerabilidade (condições cognitivas, sociais, étnicas, econômicas ou etárias). | PL 2338/2023, art. 4º, XVII e art. 15, III |
| 3 | Os danos decorrentes de erro ou viés são de baixa reversibilidade: a correção exige ação ativa do cidadão perante o poder público com assimetria de informação ou de poder. | PL 2338/2023, art. 15, IV |
| 4 | O sistema opera em escala com extensão e probabilidade de riscos que justificam avaliação de impacto sistêmico. | PL 2338/2023, art. 15, II e art. 15, VIII |

**Fonte:** elaboração própria, com base no PL 2338/2023.

O modelo não depende exclusivamente da autodeclaração do órgão aplicador para funcionar. Dois mecanismos complementares reduzem o risco de omissão no acionamento: a instância condutora pode determinar a realização de AIA por iniciativa própria, independentemente de provocação do órgão aplicador, conforme previsão análoga no art. 12, § 4º do PL 2338/2023; e, se a autoridade designada mantiver registro público de sistemas de IA de alto risco em operação, a ausência de AIA correspondente será auditável pelos órgãos de controle como irregularidade.

## 4.2 As quatro fases do processo

A Fase 1 é a autoavaliação ética (AIE). O resultado classifica o sistema em risco baixo, médio ou alto, e sistemas com risco alto passam obrigatoriamente para a Fase 2. O modelo não intervém na Fase 1 nem altera o AIE, usando o resultado como filtro de entrada para a segunda camada. Quando o acionamento ocorre diretamente pelos critérios do Quadro 1, sem passagem prévia pelo AIE, o processo começa na Fase 2.

A Fase 2 é o acionamento. O órgão aplicador notifica a instância condutora competente (a SGD/MGI, na condição de órgão central do SISP, ou a autoridade que vier a ser designada pelo SIA) e fornece documentação técnica do sistema: arquitetura do modelo, dados de treinamento, critérios de decisão, mecanismos de supervisão humana existentes, resultado do AIE quando disponível e, quando elaborado, o Relatório de Impacto à Proteção de Dados Pessoais (RIPD). O

acionamento da AIA não suspende o desenvolvimento do sistema, de modo que o órgão pode prosseguir com construção, teste e homologação, mas não pode implantar em ambiente de produção antes da conclusão da avaliação ou, quando houver urgência operacional justificada, antes da emissão de parecer preliminar pela instância condutora.

A Fase 3 é a avaliação de impacto algorítmico, conduzida pela instância condutora em três subfases. Na primeira, a triagem estruturada, a instância aplica o questionário de triagem (Instrumento 1) sobre a documentação recebida, mapeando o sistema, sua finalidade, a população afetada e os mecanismos de supervisão existentes. O instrumento é baseado em análise documental, pela qual o avaliador examina a documentação técnica e pode solicitar complementações ao órgão por escrito, tornando o processo auditável e reproduzível. Na segunda subfase, a avaliação de risco, a instância condutora preenche a matriz de risco (Instrumento 2) com base nos fatos coletados. O órgão aplicador não preenche a matriz, pois quem a preenche é o avaliador externo, condição operacional da independência da avaliação, dado que para que o órgão avaliado possa prestar contas é necessário que a instância avaliadora seja externa a ele (Bovens, 2007). Na terceira subfase, a elaboração do relatório (Instrumento 3), a instância produz o documento com os resultados por dimensão, as medidas de mitigação recomendadas e a proposta de decisão fundamentada. Antes da publicação, o órgão aplicador tem prazo para manifestação em contraditório, podendo contestar fatos ou propor medidas de mitigação adicionais, mas sem poder alterar as pontuações da matriz.

A Fase 4 é a decisão e o monitoramento. A instância emite uma de três decisões: aprovação, quando nenhuma dimensão da matriz apresenta risco alto; aprovação condicionada, quando uma ou mais dimensões apresentam risco alto com mitigação viável, com definição de medidas, responsável, prazo e indicadores de acompanhamento; ou suspensão, quando uma ou mais dimensões apresentam risco alto sem mitigação viável ou com risco à integridade de grupos vulneráveis. O monitoramento posterior é contínuo, com dois gatilhos objetivos de reavaliação. O primeiro é a mudança relevante no sistema, entendida como alteração de algoritmo, base de dados ou critérios de decisão acima de limiar a ser definido em regulamentação. O segundo é a evidência de impacto negativo não previsto, identificada por reclamações, decisões judiciais ou dados de monitoramento que indiquem disparidade nova. O padrão mínimo é reavaliação bienal. A publicação do relatório abre canal formal de manifestação pelo qual cidadãos, organizações da sociedade civil e pesquisadores podem apresentar informações sobre impactos não identificados na avaliação, registradas e consideradas obrigatoriamente na próxima revisão periódica.

## *4.3 A matriz de avaliação de risco*

A matriz organiza seis dimensões de direitos em linhas e quatro critérios de avaliação em colunas. Para cada célula, o avaliador atribui pontuação de 1 a 3, sendo 1 o cenário mais favorável e 3 o mais desfavorável. Na coluna de probabilidade: improvável (1), provável (2), certo ou quase certo (3). Na coluna de gravidade: baixa (1), moderada (2), alta (3). Na coluna de reversibilidade: facilmente reversível (1), parcialmente reversível (2), irreversível (3). Na coluna de mitigação existente: mitigação eficaz (1), parcial (2), inexistente (3). A soma por dimensão gera *score* entre 4 e 12: *score* igual ou superior a 9 classifica a dimensão como risco alto e exige medidas de mitigação específicas no relatório. As seis dimensões são derivadas dos arts. 5º, 6º e 18 do PL 2338/2023.

**Quadro 2 – Matriz de avaliação de risco (template)**

| Dimensão de direitos | Prob. do impacto negativo (1-3) | Gravidade do impacto (1-3) | Reversibilidade do dano (1-3) | Mitigação existente (1-3) |
|---|---|---|---|---|
| Privacidade e proteção de dados pessoais | | | | |
| Não discriminação | | | | |
| Devido processo legal e contestabilidade | | | | |
| Acesso a serviços essenciais | | | | |
| Transparência e explicabilidade | | | | |
| Supervisão humana efetiva | | | | |

**Fonte:** elaboração própria, com base no PL 2338/2023, arts. 5º, 6º e 18.

A separação entre o Instrumento 1, que coleta fatos, e o Instrumento 2, que avalia impacto, é escolha metodológica deliberada. Jenkins et al. (2023) demonstram que, quando a coleta de fatos e a avaliação de aceitabilidade ocorrem no mesmo ato, o avaliador tende a ajustar o julgamento normativo para se compatibilizar com os fatos documentados, produzindo avaliações benevolentes. A separação em dois instrumentos distintos aplicados em subfases sequenciais força o avaliador a documentar primeiro, sem julgamento, e só depois avaliar o que foi documentado.

## *4.4 Ancoragem regulatória e ambiguidade institucional*

O PL 2338/2023 define a AIA como obrigação do aplicador para sistemas de alto risco (art. 25), mas não especifica parâmetros metodológicos, formato do relatório, critérios de decisão e instância competente. Essa lacuna é o problema que o modelo resolve ao operacionalizar as obrigações para o contexto do setor público brasileiro. A SGD/MGI detém competência sobre governança digital como órgão central do SISP. A ANPD possui competência sobre proteção de dados pessoais. Como parte dos riscos de sistemas de IA de alto risco envolve tratamento de dados pessoais, ocorre sobreposição parcial de competências. O modelo opera sob qualquer configuração regulatória, atribuindo a condução à instância competente definida no momento da aplicação.

Um caso-limite emerge quando o próprio MGI é o órgão aplicador, situação em que ele não pode conduzir a AIA de seu próprio sistema, pois isso violaria o requisito de independência e comprometeria a neutralidade da avaliação. Para esses casos, propõe-se governança colaborativa, com transferência da condução para um comitê interministerial independente ou para a ANPD, amparada por sua autonomia como autarquia de regime especial. Essa definição depende de regulamentação infralegal do Sistema Nacional de Regulação e Governança de Inteligência Artificial (SIA), constituindo limitação de implementação até que o marco legal seja sancionado.

# 5. DEMONSTRAÇÃO: APLICAÇÃO AO CENÁRIO CADÚNICO

## 5.1 Descrição do cenário

O cenário simulado é um sistema de aprendizado de máquina aplicado pelo MDS para triagem e priorização de famílias na fila do Bolsa Família com base no CadÚnico. O sistema recebe como entrada variáveis socioeconômicas, a saber, renda *per capita* familiar, composição familiar, situação de moradia, vínculo empregatício do responsável e grau de instrução, e produz um *score* de prioridade que determina a ordem de inclusão das famílias. Decisões de inclusão, exclusão ou rebaixamento de prioridade são tomadas com base nesse *score*, com revisão humana disponível apenas após a decisão ter produzido efeito.

O CadÚnico registrava, em dezembro de 2023, aproximadamente 40 milhões de famílias, das quais cerca de 21 milhões eram beneficiárias ativas do Bolsa Família (MDS, 2023). A população cadastrada é majoritariamente composta por famílias em situação de pobreza e extrema pobreza, com alta concentração de famílias negras, chefiadas por mulheres e localizadas nas regiões Norte e Nordeste. O cenário é hipotético quanto à existência do sistema de IA, mas tecnicamente plausível: a adoção de modelos de aprendizado de máquina para priorização de beneficiários é extensão documentada de práticas adotadas em sistemas análogos internacionais. A escolha se justifica porque o sistema se enquadra diretamente no art. 14, inciso IV do PL 2338/2023, a população afetada satisfaz múltiplas dimensões de vulnerabilidade do art. 4º, XVII, e o CadÚnico possui documentação pública extensa que ancora o cenário em informações verificáveis.

## 5.2 Fases 1 e 2: AIE e acionamento

Dois gatilhos de risco alto são acionados na avaliação de riscos do AIE. O primeiro é que a decisão automatizada produz efeitos jurídicos sobre acesso a benefício social, enquadrando-se no art. 14, inciso IV. O segundo é que o nível de autonomia é alto, com revisão humana disponível apenas após o efeito da decisão. A classificação resultante é risco alto, que aciona a AIA obrigatoriamente. O AIE cumpre aqui exatamente sua função de primeira camada, identificando o nível de risco e sinalizando a necessidade de avaliação adicional. Os elementos que a segunda camada deve prover, a saber, análise de impacto diferenciado por grupo, exame da efetividade dos mecanismos de contestação e decisão vinculante, são acrescentados pelo modelo AIA. O MDS notifica a SGD/MGI e fornece a documentação técnica do sistema, incluindo o RIPD, cabível neste cenário pelo art. 5º, XVII e art. 38 da LGPD combinados com os critérios de tratamento de alto risco do art. 4º da Resolução CD/ANPD nº 2/2022, dado que o sistema opera em larga escala, afeta direitos fundamentais dos titulares e produz decisão com base em tratamento automatizado de dados pessoais.

## 5.3 Avaliação de impacto algorítmico (Fase 3)

A instância condutora aplica o questionário de triagem sobre a documentação recebida e constata quatro pontos. O modelo foi treinado sobre dados históricos do CadÚnico sem auditoria de equidade por subgrupo demográfico, sem documentação de desempenho diferenciado por raça, gênero ou território e sem evidência de que a composição da base representa proporcionalmente os grupos atendidos. O mecanismo de contestação exige que a família identifique e comprove erro na decisão, mas o sistema não fornece explicação

acessível sobre os critérios nem sobre o peso de cada variável. A revisão humana é acionada apenas por iniciativa do beneficiário após o efeito, sem notificação proativa dos afetados sobre a possibilidade de contestação. Por fim, não há registro de auditoria externa prévia.

**Quadro 3 – Aplicação da matriz de risco ao cenário CadÚnico**

| Dimensão de direitos | Prob. (1-3) | Grav. (1-3) | Rev. (1-3) | Mitig. (1-3) | Score | Classific. |
|---|---|---|---|---|---|---|
| Privacidade e proteção de dados pessoais | 2 | 2 | 2 | 2 | 8 | Moderado |
| Não discriminação | 3 | 3 | 3 | 1 | 10 | Alto |
| Devido processo legal e contestabilidade | 3 | 3 | 3 | 1 | 10 | Alto |
| Acesso a serviços essenciais | 3 | 3 | 3 | 1 | 10 | Alto |
| Transparência e explicabilidade | 3 | 2 | 2 | 1 | 8 | Moderado |
| Supervisão humana efetiva | 3 | 3 | 3 | 1 | 10 | Alto |

**Fonte:** elaboração própria.

Na dimensão não discriminação (*score* 10), a probabilidade é alta porque o modelo foi treinado sobre dados históricos sem auditoria de equidade por subgrupo demográfico, e dados históricos de renda e acesso a benefícios no Brasil refletem desigualdades estruturais por raça, gênero e território. Para fundamentar essa pontuação empiricamente, a pesquisa conduziu experimentos de auditoria de viés sobre o *Adult Income Dataset* (*UCI Machine Learning Repository*, 32.561 registros), conjunto de dados de renda com variáveis socioeconômicas análogas ao cenário simulado. O *demographic parity ratio* (DPR) mede a razão entre a taxa de decisão favorável para o grupo menos favorecido e a taxa para o grupo mais favorecido, variando de 0 (desigualdade máxima) a 1 (paridade plena). O limiar regulatório de referência é 0,800, adotado pela regra dos 80% da *U.S. Equal Employment Opportunity Commission*. Sem restrições de equidade, o *Random Forest* atingiu DPR de 0,337 e a rede neural MLP de 0,369, ambos muito abaixo do limiar. Após mitigação via *ThresholdOptimizer* (biblioteca *Fairlearn*), a MLP alcançou DPR de 0,939 com redução de 4,7 pontos percentuais em acurácia; o *Random Forest* permaneceu abaixo do limiar (DPR 0,666) com custo de 7,6 p.p. Os resultados demonstram que a discriminação indireta é mensurável e estatisticamente expressiva mesmo em modelos de alta acurácia geral, e que o *trade-off* entre equidade e desempenho varia segundo o algoritmo, tornando a escolha técnica objeto legítimo de avaliação de impacto. A reversibilidade é baixa porque a contestação exige identificação e prova do erro pela família, com severa assimetria de informação. A mitigação é mínima porque não há auditoria de equidade documentada.

As dimensões devido processo legal e contestabilidade, acesso a serviços essenciais e supervisão humana efetiva recebem *score* 10 pelo mesmo perfil de risco. O sistema não fornece explicação acessível sobre os critérios da decisão, tornando o direito de contestação do art. 6º, IV do PL 2338/2023 inoperante na prática: o beneficiário não sabe o que contestar. A decisão automatizada determina diretamente a inclusão ou exclusão do Bolsa Família, programa que constitui para as famílias afetadas a principal ou única fonte de renda. A revisão humana só está disponível após o efeito da decisão por iniciativa do beneficiário,

sendo insuficiente quando a decisão produz consequência imediata e de difícil reversão, como a exclusão do cadastro. As dimensões privacidade e proteção de dados (*score* 8) e transparência e explicabilidade (*score* 8) são classificadas como risco moderado, já que o CadÚnico tem protocolo estabelecido de proteção de dados e a publicação de critérios de decisão é tecnicamente viável, reduzindo os respectivos *scores*.

### 5.4 Decisão (Fase 4) e diferencial em relação ao AIE

Quatro das seis dimensões apresentam *score* igual ou superior a 9. A instância condutora emite aprovação condicionada com três medidas obrigatórias: auditoria de equidade nos dados de treinamento com recorte por raça, gênero e território, com publicação dos resultados e DPR por subgrupo como indicador; implantação de mecanismo de explicação acessível para beneficiários com prioridade rebaixada, descrevendo em linguagem simples os fatores que influenciaram a decisão, com taxa de utilização do mecanismo como indicador; e implantação de revisão humana disponível antes do efeito definitivo para casos de exclusão do cadastro, com percentual de casos revisados como indicador. O MDS produzirá relatório anual de impacto com esses indicadores, e a reavaliação ocorrerá em 24 meses ou antes se qualquer gatilho de mudança relevante for acionado.

A comparação entre os dois instrumentos no mesmo cenário ilustra a complementaridade entre as duas camadas. O AIE cumpriu sua função, classificou o sistema como risco alto e sinalizou a necessidade de avaliação adicional, sem o que a AIA não seria acionada. O modelo AIA acrescentou o que pertence à segunda camada, a saber, diagnóstico de risco por dimensão com pontuação fundamentada, identificação do risco de discriminação indireta por raça e território, exame da efetividade real do mecanismo de contestação, avaliação do momento da supervisão humana no fluxo decisório e decisão formal vinculante com três medidas de mitigação, prazos e indicadores. Os dois instrumentos fazem coisas diferentes e igualmente necessárias.

## 6. AVALIAÇÃO DO ARTEFATO

A quinta etapa da DSR confronta o artefato com os objetivos definidos na segunda etapa. Na ausência de validação empírica com especialistas e órgãos aplicadores, esta avaliação adota o método de análise estruturada, pelo qual cada requisito de design derivado do *benchmark* é confrontado com os resultados da demonstração, classificado como atendido, parcialmente atendido ou não atendido, com justificativa baseada nos dados da demonstração.

**Quadro 4 – Avaliação do modelo AIA em relação aos requisitos do *benchmark***

| Requisito | Classificação | Evidência na demonstração |
|---|---|---|
| (1) Critérios objetivos de acionamento | Parcialmente atendido | Acionado por dois caminhos independentes (AIE e art. 14, IV do PL). Acionamento por terceiros e controle externo via registro público definidos no modelo, mas não testados por dependerem de infraestrutura institucional inexistente. |
| (2) Instância condutora independente | Atendido | A demonstração simulou a condução pela SGD/MGI como instância externa ao MDS. O conflito quando o MGI é aplicador foi tratado com proposta de governança colaborativa (seção 4.4). |

| Requisito | Classificação | Evidência na demonstração |
|---|---|---|
| (3) Avaliação centrada em impacto sobre direitos | Atendido | A matriz avaliou seis dimensões com pontuação fundamentada por dimensão. A dimensão de não discriminação foi sustentada por experimento empírico com DPR de 0,337 e 0,369 antes da mitigação. |
| (4) Decisão com efeito vinculante | Atendido | A aprovação condicionada foi emitida com três medidas obrigatórias, cada uma com responsável, prazo e indicador de acompanhamento (seção 5.4). |
| (5) Publicidade do resultado | Parcialmente atendido | O modelo prevê publicação conforme art. 28 do PL e canal formal de manifestação de terceiros. A publicação efetiva e a verificação de acesso não puderam ser testadas em cenário simulado. |
| (6) Monitoramento com gatilhos objetivos | Parcialmente atendido | Reavaliação em 24 meses e relatório anual com indicadores definidos (seção 5.4). Os dois gatilhos foram especificados. O funcionamento do ciclo depende de implementação real. |

**Fonte:** elaboração própria.

O modelo atende integralmente a três dos seis requisitos e parcialmente aos outros três. As três satisfações parciais têm a mesma causa, pois decorrem da natureza simulada da demonstração, que permite testar a lógica interna do modelo, mas não os elementos que dependem de infraestrutura institucional real, como registro centralizado de sistemas, publicação efetiva do relatório e ciclo de monitoramento em operação.

A comparação com os quatro elementos que a segunda camada deve acrescentar ao AIE permite avaliar em que medida o modelo resolve o problema que motivou a pesquisa. O primeiro elemento, verificação externa, é atendido pelo requisito 2. O segundo, análise de impacto sobre grupos específicos, é atendido pelo requisito 3, com a dimensão de não discriminação sustentada por auditoria empírica de viés. O terceiro, decisão vinculante, é atendido pelo requisito 4. O quarto, participação de afetados, não é integralmente atendido pelo modelo, que oferece mitigação parcial por canal de manifestação pós-publicação.

Do ponto de vista do referencial de Stilgoe et al. (2013), o modelo satisfaz três das quatro dimensões de inovação responsável. A antecipação é atendida pela avaliação pré-implantação com matriz de risco por dimensão de direitos. A reflexividade é atendida pela separação entre coleta de fatos e avaliação de impacto, que impede o ajuste inconsciente do julgamento normativo. A responsividade é atendida pela decisão vinculante com três tipos possíveis. A inclusão não é atendida pela ausência de participação deliberativa de afetados: a mitigação parcial pelo canal de manifestação não equivale à inclusão tal como definida pelos autores, que requer participação no processo de avaliação, não apenas reação ao resultado.

# CONCLUSÃO

O *Framework* para Autoavaliação de Impacto Ético em Inteligência Artificial é o primeiro instrumento governamental brasileiro a sistematizar a avaliação ética de sistemas de IA no setor público. Como instrumento de primeira camada, ele classifica riscos, gera recomendações e orienta melhorias, cumprindo uma função que nenhum outro instrumento equivalente cobre no contexto brasileiro. A consolidação dessa camada cria as condições para que uma segunda camada seja acrescentada: uma que produza verificação externa, mensure impactos diferenciados sobre grupos com dados demográficos e imponha decisões vinculantes para os sistemas classificados como de alto risco.

O modelo de Avaliação de Impacto Algorítmico proposto opera nessa segunda camada. O artefato define critérios objetivos de acionamento baseados no art. 15 do PL 2338/2023, estabelece processo de avaliação em quatro fases conduzido por instância independente e utiliza matriz de risco focada em seis dimensões de direitos fundamentais. A aplicação ao cenário do CadÚnico demonstra a articulação entre as duas camadas: o AIE identificou o risco alto e acionou a avaliação; o modelo AIA acrescentou diagnóstico de risco por dimensão, três medidas de mitigação com prazos e indicadores, e decisão formal de aprovação condicionada.

Os experimentos de auditoria de viés confirmam que algoritmos com alta acurácia geral produzem discriminação indireta mensurável quando treinados sem restrições de equidade, e que a mitigação reduz o desempenho preditivo entre 4,7 e 7,6 pontos percentuais conforme o algoritmo. Esse *trade-off* exige decisão institucional, não apenas ajuste de código.

O estudo apresenta três limitações. A aplicação a cenário simulado impede verificar a efetividade do modelo em contexto real. A ausência de validação empírica dos instrumentos com servidores públicos e especialistas em direito digital impede verificar a clareza e a aplicabilidade prática dos quatro instrumentos produzidos. A dependência de regulamentação infralegal do PL 2338/2023 impede definir a instância competente, especialmente nos casos em que o próprio MGI atua como desenvolvedor. A pesquisa futura requer validação com gestores públicos, aplicação da matriz a sistemas de reconhecimento facial na segurança pública e replicação dos experimentos com microdados anonimizados brasileiros. O modelo AIA proposto converte a classificação de risco da primeira camada em avaliação de impacto com decisão vinculante, transferindo o ônus da conformidade do cidadão afetado para a instituição operadora do sistema.

## *Disponibilidade de dados e código*

Os dados e o código utilizados nos experimentos de auditoria de viés algorítmico estão disponíveis publicamente. O conjunto de dados *Adult Income Dataset* foi obtido do *UCI Machine Learning Repository*, repositório de acesso aberto. Os *notebooks* Python utilizados nos experimentos com as bibliotecas *Fairlearn* e *scikit-learn* estão depositados no repositório GitHub da pesquisa:

**Repositório de código (GitHub):** https://github.com/camila-fromero/AIA_auditoria_de_vies_algoritmico

O repositório contém arquivo README.md com instruções para execução do código e reprodução dos resultados. Os arquivos são: projeto_vieses_ml.ipynb (Random Forest) e projeto_viese_rede_neural.ipynb (rede neural MLP). Os dados brutos e processados utilizados nos experimentos estão descritos no README.md com instruções de acesso ao *UCI Machine Learning Repository*. Os resultados dos experimentos estão documentados no Apêndice D do TCC completo, disponível mediante solicitação à autora.

O uso de ferramenta de IA generativa (modelos Claude, Anthropic, versões disponíveis entre 2025 e 2026) foi declarado no TCC completo (seção 1.5 e Apêndice E). A ferramenta atuou como suporte técnico e redatorial nas etapas de revisão de texto, geração de trechos de código *Python* para os experimentos e formatação dos instrumentos operacionais. A definição do problema de pesquisa, a seleção do referencial teórico, a concepção do modelo AIA e as conclusões foram realizados exclusivamente pela autora. Todas as citações e referências bibliográficas passaram por conferência manual. Nenhum dado pessoal, sensível ou sigiloso foi inserido na ferramenta.



## REFERÊNCIAS


AI NOW INSTITUTE. **Algorithmic impact assessments: a practical framework for public agency accountability**. New York: AI Now Institute, 2018.

BOVENS, Mark. Analysing and Assessing Accountability: A Conceptual Framework. **European Law Journal**, v. 13, n. 4, p. 447-468, 2007.

BRASIL. Autoridade Nacional de Proteção de Dados. **Resolução CD/ANPD nº 2, de 27 de janeiro de 2022**. Brasília: ANPD, 2022.

BRASIL. Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome. **Painel de Informações Sociais — Cecad 2.0**. Dados de dezembro de 2023. Disponível em: https://cecad.cidadania.gov.br. Acesso em: mar. 2026.

BRASIL. **Projeto de Lei n. 2.338, de 2023**. Dispõe sobre o uso da inteligência artificial. Aprovado pelo Senado Federal em 10 de dezembro de 2024. Em tramitação na Câmara dos Deputados. Brasília: Senado Federal, 2024.

CANADA. Treasury Board Secretariat. **Directive on Automated Decision-Making**. Ottawa: Government of Canada, 2019.

GALDON CLAVELL, Gemma; AUMAITRE, Ariane; CALDERS, Toon. How a sociotechnical approach to AI auditing can change how we understand and measure fairness in machine learning systems. In: **AIMMES 2024 Workshop on AI Bias.** CEUR Workshop Proceedings, v. 3744, 2024.

GATT, L. et al. FRIA Implementation Model According to the AI Act. **IEEE Access**, v. 12, p. 1-15, 2024. DOI: 10.1109/ACCESS.2024.3481234.

GOVERNO FEDERAL BRASILEIRO. **Framework para Autoavaliação de Impacto Ético em Inteligência Artificial** (AIE), versão Beta 2.39. Brasília: MGI, 2025.

JENKINS, R. et al. Separating facts and evaluation: motivation, account, and learnings from a novel approach to evaluating the human impacts of machine learning. **AI and Ethics**, v. 3, p. 1-15, 2023. DOI: 10.1007/s00146-022-01417-y.

KAMINSKI, M. E.; MALGIERI, G. Algorithmic impact assessments under the GDPR: producing multi-layered explanations. **International Data Privacy Law**, v. 11, n. 2, p. 125-144, 2021.

LEMOS, Alessandra et al. **Avaliação de Impacto Algorítmico para a proteção dos direitos fundamentais**. Brasília: Laboratório de Políticas Públicas e Internet, 2023.

MANTELERO, A.; ESPOSITO, M. S. An evidence-based methodology for human rights impact assessment in AI data-intensive systems. **Computer Law and Security Review**, v. 41, 2021.

PEFFERS, K. et al. A design science research methodology for information systems research. **Journal of Management Information Systems**, v. 24, n. 3, p. 45-77, 2007.

SELBST, Andrew D. An Institutional View of Algorithmic Impact Assessments. **Harvard Journal of Law & Technology**, v. 35, n. 1, p. 117-190, 2021.

STILGOE, J.; OWEN, R.; MACNAGHTEN, P. Developing a framework for responsible innovation. **Research Policy**, v. 42, n. 9, p. 1568-1580, 2013.

UNESCO. **Recommendation on the Ethics of Artificial Intelligence**. Paris: UNESCO, 2021.

UNIÃO EUROPEIA. **Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024 (Regulamento IA).** Jornal Oficial da União Europeia, 2024.

# Desmatamento nos Municípios da Amazônia Legal: uma análise espaço-temporal com dados em painel e aprendizado de máquina

*Juliana Pierrobon Lopez*
*Profª MSc. Thaís Salzer Procópio*

## Resumo

Este estudo investiga os fatores associados ao desmatamento nos municípios da Amazônia Legal entre 2012 e 2021, combinando dados de alta dimensionalidade e modelos de aprendizado de máquina, a fim de capturar relações complexas e não-lineares. Por meio do algoritmo XGBoost, que demonstrou superioridade preditiva sobre modelos lineares, e da técnica de interpretação SHAP, o estudo identificou a inércia temporal (desmatamento no ano anterior) como o fator de maior associação com a taxa de desmatamento. Os resultados também destacaram a relevância da fiscalização ambiental (áreas embargadas e autuações) e de indicadores de vulnerabilidade socioeconômica, como o orçamento de assistência social e o atendimento de necessidades humanas básicas. A análise evidencia que o fenômeno é multidimensional e heterogêneo, reforçando a importância de políticas públicas territorialmente diferenciadas e que integrem as agendas ambiental, econômica e social.

**Palavras-chave:** Desmatamento, Amazônia Legal, Aprendizado de máquina.

## *INTRODUÇÃO*

Cerca de 60% da Amazônia, a maior floresta tropical do mundo, encontra-se em território brasileiro. Apesar da megadiversidade e do papel crucial no estoque de gases de efeito estufa (GEE ), o bioma ainda sofre intensa pressão decorrente do desmatamento.

De acordo com dados do Projeto de Monitoramento do Desmatamento da Floresta Amazônica Brasileira por Satélite (Prodes), conduzido desde 1988, o auge do desmatamento se deu em 1995 (29.100 km$^2$), seguido pelos anos de 2004 (27.800 km$^2$) e 2003 (25.400 km$^2$) (INPE, 2026a).

Em 2004, como estratégia para redução do desmatamento, o governo brasileiro elaborou e implementou o Plano de Ação para Prevenção e Controle do Desmatamento na Amazônia Legal (PPCDAm). Durante suas quatro fases iniciais (MMA, 2004, 2009, 2013 e 2018), as taxas de desmatamento foram reduzidas e, em 2012, atingiram o menor nível histórico (4.600 km$^2$).

Em 2019, o PPCDAm foi descontinuado e substituído pelo Plano Nacional de Controle do Desmatamento Ilegal e Recuperação da Vegetação Nativa (MMA, 2020). Em 2021, o desmatamento atingiu 13 mil km$^2$, o maior desde 2006 (14.300 km$^2$). Em 2023, o PPCDAm foi retomado em sua 5ª fase (MMA, 2023), e o desmatamento voltou a recuar (9.100 km$^2$ em 2023 e 6.500 km$^2$ em 2024) (INPE, 2026a).

Os fatores relacionados ao desmatamento na região vêm sendo investigados pela comunidade acadêmica há décadas (e.g., Pfaff, 1999; Margulis, 2003), de forma abrangente e multidisciplinar. Busch & Ferretti-Gallon (2023), em meta-análise a partir de centenas de artigos publicados, identificaram 1.630 variáveis explicativas estudadas (com nomes únicos e, pelo menos, uma análise de correlação com o desmatamento), as quais agruparam em 40 categorias distribuídas em 6 temas: Infraestrutura Construída, Commodities, Demografia e Socioeconomia, Gestão da Terra e Instituições, Políticas, Características Biofísicas.

Grande parte dessa literatura se fundamenta em modelos econométricos tradicionais (regressões em painel com efeitos fixos e modelos de diferenças em diferenças), que possibilitam rigor na identificação causal, mas apresentam limitações na modelagem de relações não-lineares e interações complexas entre variáveis.

Nesse cenário, este trabalho explora os modelos de aprendizado de máquina como uma abordagem complementar, com capacidade de capturar padrões preditivos de alta dimensionalidade. Destaca-se que, apesar de esses modelos serem ferramentas poderosas para previsão e para revelar a existência de estrutura nos dados, são limitados quanto à identificação da contribuição causal de cada variável, uma vez que otimizam o desempenho preditivo fora da amostra e não a recuperação de parâmetros com interpretação causal bem definida (Mullainathan & Spiess, 2017).

Assim, o objetivo desse estudo é identificar fatores associados ao desmatamento nos municípios da Amazônia Legal entre 2012 e 2021, analisando a magnitude, a direção, a heterogeneidade espaço-temporal e a inter-relação dessas associações, sem buscar inferências causais.

A Seção 2 apresenta a metodologia, detalhando a delimitação espacial e temporal, a construção da base de dados, os procedimentos de modelagem e os cenários de análise. A Seção 3 expõe os resultados e a discussão, organizados em análise de inércia do desmatamento, seleção do modelo, contexto geral, contexto temporal, contexto espacial e estudos de caso municipais. A Seção 4 apresenta as considerações finais, com síntese dos achados, limitações e recomendações.

A definição de Amazônia Legal adotada foi a do IBGE, que segue a Lei Complementar nº 124/2007, abrangendo os estados do Acre, Amapá, Amazonas, Mato Grosso, Rondônia, Roraima, Tocantins, Pará e a porção oeste do Maranhão (IBGE, 2026). Destaca-se que essa é uma delimitação político-administrativa criada para fins de planejamento e ação estatal, não correspondendo exatamente à extensão do bioma Amazônia, pois abrange também partes do Cerrado e Pantanal.

A definição de desmatamento utilizada foi a do Prodes, que é a supressão da vegetação nativa primária, não diferenciando o desmatamento legal do ilegal (INPE, 2026b).

# 1. METODOLOGIA

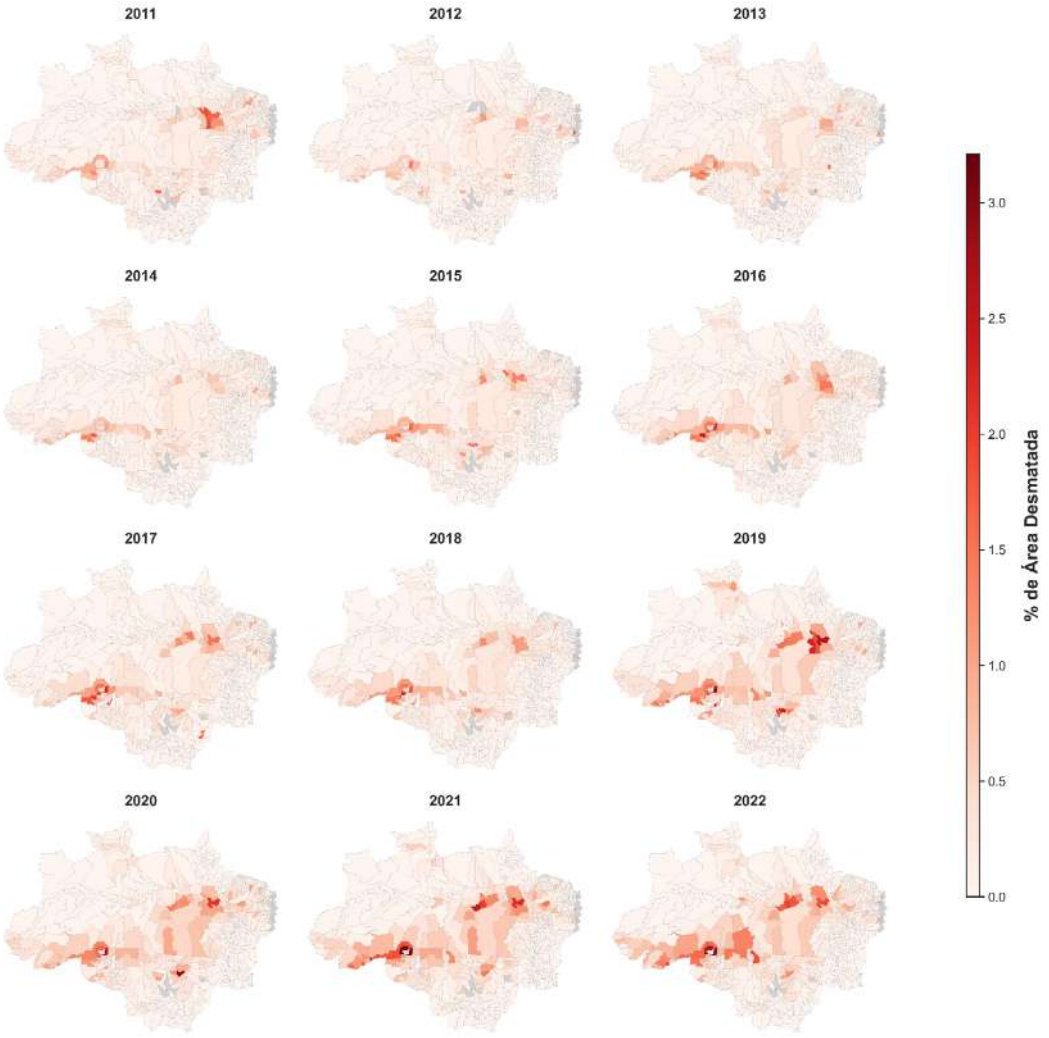
## 1.1. Delimitação Espacial e Temporal

O estudo abrange o período de 2012 a 2021, recorte definido pela maior disponibilidade das bases de dados oficiais, possibilitando, assim, uma análise contínua e consistente da série temporal.

Em razão da granularidade municipal dos dados, a área de estudo abrangeu apenas os municípios integralmente inseridos na Amazônia Legal, visto que a inclusão de municípios localizados parcialmente na região agregaria territórios fora do escopo geográfico da análise.

Por questões cartográficas, foram excluídos também os municípios de Boa Esperança do Norte (MT), Sorriso (MT) e Nova Ubiratã (MT) (o primeiro emancipou-se dos demais em 2025). Como o arquivo vetorial utilizado nas análises (IBGE, 2026) traz os limites atuais - ou seja, com os novos limites de Sorriso (MT) e Nova Ubiratã (MT), ele não é adequado para extração de informações geográficas anteriores a 2025 desses municípios.

**Figura 1 - Evolução do desmatamento nos municípios da Amazônia Legal entre 2011 e 2022.**




**Fonte:** Elaborado pela autora com base nos dados Prodes/INPE (2026).

Assim, foram incluídos 749 dos 773 municípios que compõem o território da Amazônia Legal. Apenas para o ano de 2012, a amostra se reduz a 747 municípios também por compatibilização dos limites cartográficos, já que Mojuí dos Campos (PA) emancipou-se de Santarém (PA) em 2013.

A evolução do desmatamento no período analisado pode ser observada na Figura 1, que apresenta as taxas anuais de supressão da vegetação nativa primária por município entre 2011 e 2022.

## 1.2. Construção da base de dados

Considerando a natureza multifacetada dos fatores relacionados ao desmatamento, a construção da base de dados se deu a partir da integração de diversas fontes públicas oficiais. As variáveis foram selecionadas de modo a contemplar diferentes dimensões: dinâmica de uso e ocupação do solo, áreas protegidas, fiscalização ambiental, infraestrutura, além de indicadores socioeconômicos, fundiários e agropecuários.

Essa estrutura foi concebida para capturar não apenas os reconhecidos vetores diretos do desmatamento (como a pecuária e a soja), mas também possíveis fatores indiretos, incluindo vulnerabilidade social e mecanismos de comando e controle.

Para assegurar a comparabilidade entre unidades territoriais de dimensões heterogêneas e mitigar vieses de escala, as variáveis absolutas foram normalizadas mediante conversão em porções de área do município (*e.g.,* taxa de área desmatada, valores entre 0-1), em densidades (*e.g.*, cabeças de gado por km$^2$) ou em índices *per capita* (*e.g.*, orçamento em assistência social por habitante), conforme a natureza do dado e quando necessário.

Para o tratamento de dados ausentes, presentes nas séries históricas de orçamento e dos índices socioeconômicos, adotou-se a interpolação linear em lacunas internas da série temporal. Para omissões nos extremos da série (anos iniciais ou finais), aplicou-se a técnica de *carry-forward* ou *back-fill,* propagação do último ou primeiro valor conhecido, respectivamente.

A seleção inicial das variáveis foi realizada com o objetivo de reduzir redundâncias e mitigar problemas de multicolinearidade, especialmente relevantes para os modelos lineares. Para tanto, foi realizada uma Análise Exploratória dos Dados (AED), incluindo análises de correlação de Pearson entre variáveis preditoras e a variável-alvo e avaliação de multicolinearidade.

Ressalta-se que essa etapa não teve caráter exaustivo, sendo complementada pela capacidade dos algoritmos de aprendizado de máquina de capturar relações não lineares e interações complexas entre variáveis, mesmo após a redução inicial do conjunto de dados.

Ao final deste processo, foram consolidadas 55 variáveis em 7.478 observações, compondo o conjunto de dados para o treinamento dos algoritmos de aprendizado de máquina. A descrição das variáveis, com respectiva fonte e tratamentos adotados, está detalhada no Anexo 1.

## 1.3. Modelagem e Aprendizado de Máquina

Para a modelagem, a base de dados foi dividida nos subconjuntos treino (80%) e teste (20%), com estratificação por ano e estado, de modo a garantir que a distribuição espaço-temporal do painel fosse mantida em ambas as partições, e evitar que o modelo fosse avaliado em contextos geográficos ou períodos não representados no treinamento.

## Pré-processamento

As variáveis foram classificadas como: Identificadoras, Categóricas ou Numéricas. Com o objetivo de reduzir a assimetria (*skewness*) decorrente da heterogeneidade dos dados municipais, as variáveis numéricas preditoras (ou seja, exceto a variável-alvo) foram submetidas a um protocolo de transformação baseado na distribuição dos dados e na frequência de valores iguais a zero:

- **Alta Assimetria (*skewness* >= 2) com zeros:** transformação de Yeo-Johnson, adequada para estabilizar a variância na presença de zeros;
- **Alta Assimetria (*skewness* >= 2) e sem zeros:** transformação de Box-Cox, robusta para dados estritamente positivos;
- **Assimetria moderada (1 < *skewness* < 2) com proporção de zeros ≥ 20%:** transformação de Yeo-Johnson, indicada para distribuições com cauda longa e alta frequência de zeros;
- **Assimetria moderada (1 < *skewness* < 2) com proporção de zeros < 20%:** transformação Log1p — log(x+1), eficiente para suavização logarítmica com baixa frequência de zeros.

Após as transformações, todas as variáveis numéricas preditoras foram padronizadas via *StandardScaler* (média zero, desvio padrão unitário). A única variável categórica (PPCDAm) foi codificada pelo método *One-Hot Encoding*.

## Seleção e validação de modelos

Para a etapa de modelagem, foram testados quatro algoritmos: Regressão Linear (OLS), utilizada como modelo de referência (*baseline*); ElasticNet; Random Forest; e XGBoost. Os dois primeiros são modelos lineares e os dois últimos, métodos de *ensemble* baseados em árvores de decisão.

A validação cruzada foi realizada por meio do *GroupKFold* com agrupamento por município, para assegurar que as observações de um mesmo município fossem alocadas integralmente em um único subconjunto (treino ou teste) e evitar a contaminação das métricas por autocorrelação intra-município.

Essa abordagem garante que observações de um mesmo município não sejam separadas entre treino e teste, mas não adota validação temporal. Contudo, como o objetivo é identificar padrões associativos, e não prever valores futuros, essa característica não compromete as conclusões principais da pesquisa. A inclusão de defasagens (t-1) permite capturar a dinâmica temporal, enquanto validações estritamente temporais podem ser exploradas em trabalhos futuros.

Os hiperparâmetros de cada algorítimo foram ajustados via *RandomizedSearchCV* com 20 iterações e as seguintes configurações: i) Regressão Linear - nenhuma; ii) ElasticNet - alpha (0.0001, 0.001, 0.01, 0.1, 1.0), l1_ratio (0.1, 0.3, 0.5, 0.7, 0.9); iii) Random Forest - n_estimators (100, 300), max_depth (10, 15), max_features (sqrt, log2); e,iv) XGBoost - n_estimators (100, 300, 500), learning_rate (0.01, 0.05, 0.1) max_depth (3, 6, 9); subsample (0.8, 1.0); colsample_bytree (0.7, 0.9)

Para a seleção do melhor modelo, o desempenho de cada um foi avaliado por meio de quatro métricas: i) Coeficiente de Determinação ($R^2$); ii) Raiz do Erro Quadrático Médio (RMSE); Erro Médio Absoluto (MAE); e, Sobreajuste ($\Delta R^2$).

Como as transformações foram aplicadas exclusivamente nas variáveis preditoras, e não à variável-alvo, as métricas de erro (RMSE e MAE) são interpretáveis diretamente na escala original da taxa de desmatamento (porção da área municipal desmatada, escala 0-1).

### *Interpretabilidade via SHAP*

Para a interpretação do resultado obtido com o modelo de melhor desempenho, aplicou-se o método SHAP (*SHapley Additive exPlanations*), fundamentado na Teoria dos Jogos Cooperativos. O SHAP decompõe a previsão de cada observação (município-ano), quantificando a contribuição marginal de cada variável preditora. Por consequência, a soma dos valores SHAP de todas as variáveis em uma dada observação é igual ao desvio da previsão em relação à média global do modelo.

Foi programada a adoção dos explicadores nativos da biblioteca, que calculam os valores SHAP a partir da estrutura interna do modelo, sem aproximações estocásticas. São eles: *LinearExplainer* (melhor desempenho de modelo linear); e, para melhor desempenho de modelo baseado em árvores, *TreeExplainer* com a perturbação *tree_path_dependent*, na qual o SHAP simula a ausência de cada variável utilizando a distribuição condicional dos dados aprendida internamente pelas árvores.

Os valores SHAP foram calculados sobre o conjunto completo de observações (treino + teste), uma vez que o objetivo do estudo é descritivo e a validade preditiva já havia sido estabelecida sobre o conjunto de teste isolado. Ressalta-se que o SHAP não mensura causalidade, mas associações condicionais, sendo capaz de capturar relações complexas e não-lineares presentes nos dados.

## *1.4. Cenários de Análise*

Com o objetivo de capturar a heterogeneidade espacial e a dinâmica temporal do fenômeno, a análise da interpretabilidade SHAP foi conduzida em três contextos:

1. **Contexto Geral:** série histórica completa para a totalidade dos municípios. O objetivo foi identificar quais variáveis contribuíram no território ao longo da década.
2. **Contexto Temporal:** análise segmentada por ano, com o objetivo de observar a flutuação da contribuição relativa das variáveis ao longo da década.
3. **Contexto Espacial:** análise segmentada por Unidade da Federação (UF), a fim de capturar o comportamento da contribuição relativa das variáveis em cada estado.

Para viabilizar a comparação direta entre estados e entre anos, os valores SHAP foram normalizados em importância relativa, de modo que o somatório da contribuição marginal de cada variável para cada estado/ano fosse igual a 100%. Ressalta-se que as análises por UF e por ano não constituem uma decomposição causal, mas uma descrição da heterogeneidade espacial e temporal nas associações aprendidas pelo modelo.

Para garantir a consistência metodológica e a comparabilidade dos resultados entre contextos, adotou-se em todas as análises o modelo com melhor desempenho no Contexto Geral.

### 1.5. Estudos de caso

Complementarmente à análise de importância global (comportamento agregado nos cenários), foram conduzidos estudos de caso em quatro municípios no ano de 2020 via análise SHAP local. Essa abordagem permite decompor a previsão individual a partir do valor-base do modelo, identificando as contribuições específicas (positivas ou negativas) de cada variável para o resultado observado naquele município-ano.

### 1.6. Disponibilidade de Dados e Código

Dados e códigos disponibilizados em https://github.com/JuPLopez/desmat_Alegal.

### 1.7. Uso de Ferramentas de Inteligência Artificial

No desenvolvimento desta pesquisa, foram empregados Modelos de Linguagem de Grande Escala (LLMs) — especificamente Gemini (Google), DeepSeek e Claude (Anthropic) — como suporte iterativo em fases distintas do projeto. A aplicação dessas ferramentas concentrou-se em: i) Codificação Assistida - suporte à criação, à depuração e à otimização de scripts em linguagem Python; ii) Suporte Bibliográfico - auxílio no levantamento exploratório e na síntese preliminar de literatura especializada, condicionados à conferência e à validação das fontes originais; e iii) Refinamento Textual - revisão estilística e gramatical da redação final.

Em todas as etapas, a autonomia decisória e a supervisão técnica foram exercidas integralmente pela autora. Assim, a IA atuou como facilitadora de processos operacionais, sendo as decisões metodológicas, os resultados finais e a fundamentação teórica de responsabilidade exclusiva da autora.

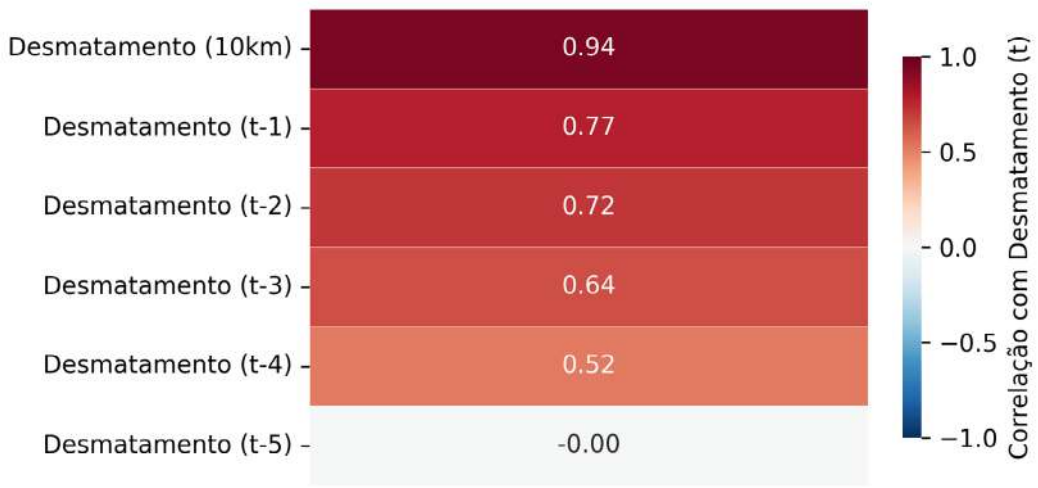
## 2. RESULTADOS E DISCUSSÃO

### 2.1. A inércia do desmatamento

A análise exploratória dos dados evidenciou um padrão de persistência espaço-temporal no desmatamento municipal (Figura 2). A taxa de desmatamento, em determinado ano, apresentou forte correlação positiva com os níveis registrados em períodos anteriores ($r = 0,77$ em $t-1$, decaindo progressivamente até a neutralidade em $t-5$), e de forma ainda mais acentuada no entorno (*buffer* de 10 km, $r = 0,94$).

Esse padrão representou uma autocorrelação temporal do desmatamento, na qual a ocorrência de supressão de vegetação em um determinado ano é um preditor robusto de continuidade do processo no ano subsequente. Uma hipótese para explicar esse comportamento seria a consolidação progressiva de um ambiente favorável à expansão do desmatamento, incluindo a redução de custos marginais de novas supressões e a estruturação de redes de atores locais.

**Figura 2 - Correlação de Pearson entre a taxa de desmatamento municipal (t) na Amazônia Legal no período de 2012 a 2021, as defasagens temporais (t–1 a t–5) e o desmatamento no entorno (10 km).**




**Fonte:** Elaborado pela autora (2026).

Entretanto, dado que este estudo restringe o conceito de desmatamento à supressão de vegetação primária, essa inércia encontraria o limite físico da disponibilidade de florestas remanescentes nessas condições, sugerindo um eventual efeito de saturação (neutralidade da correlação observada em t-5).

Diante desse padrão, dois critérios foram aplicados visando à robustez estatística do modelo: i) exclusão da variável de desmatamento no entorno (10 km) para evitar *data leakage*, por se tratarem de dados do mesmo ano; e ii) inclusão apenas da defasagem de um ano (t-1) para capturar a inércia temporal, descartando-se as defasagens de maior ordem para mitigar a multicolinearidade.

## 2.2. Seleção do modelo

A avaliação de desempenho dos modelos de aprendizado de máquina testados, pautada principalmente pelo poder explicativo do teste ($R^2$ Teste), revelou a superioridade dos algoritmos de *ensemble* baseados em árvores, que explicaram cerca de 65-70% da variância, enquanto os modelos lineares não atingiram 50% (Tabela 1).

Entre os modelos não lineares avaliados, o XGBoost foi selecionado por apresentar a melhor relação entre poder explicativo e capacidade de generalização. Embora o modelo Random Forest tenha atingido um $R^2$ de teste ligeiramente superior (3%), esse ganho foi acompanhado por um aumento de três vezes no nível de sobreajuste ($\Delta R^2$), o que penalizou sua capacidade de generalização.

**Tabela 1. Desempenho comparativo dos algoritmos de aprendizado de máquina testados.**

| Modelo | $R^2$ Treino | $R^2$ Teste | RMSE Teste | MAE Teste | Sobreajuste ($\Delta R^2$) |
|---|---|---|---|---|---|
| Random Forest | 0,9494 | 0,6988 | 0,0014 | 0,0006 | 0,2507 |
| XGBoost | 0,7524 | 0,6659 | 0,0014 | 0,0007 | 0,0865 |
| Linear Regression | 0,5025 | 0,4737 | 0,0018 | 0,0010 | 0,0287 |
| ElasticNet | 0,4890 | 0,4544 | 0,0018 | 0,0010 | 0,0346 |

**Fonte:** elaborado pela autora (2026)

As métricas de erro absoluto (MAE) e quadrático (RMSE) apresentaram valores na casa dos milésimos, refletindo a escala fracionária da variável-alvo. Os modelos lineares apresentaram MAE e RMSE sistematicamente superiores, indicando pior ajuste geral aos dados em comparação aos modelos não lineares.

## *2.3. Contexto Geral*

A análise SHAP aplicada ao modelo XGBoost permitiu avaliar a relação das variáveis com o desmatamento municipal na Amazônia Legal, identificando a magnitude e a direção da contribuição de cada uma sobre as previsões individuais, sem, contudo, indicar causalidade.

No painel superior da Figura 3, observa-se a dominância da variável *Desmatamento (t-1)*, que atingiu o patamar de 0,012 na escala SHAP e representou uma magnitude cerca de cinco vezes superior à de qualquer outra variável isolada. A direção do impacto foi consistente: municípios com alto desmatamento no ano anterior (pontos vermelhos) concentraram-se à direita do zero, indicando contribuição positiva para a previsão de desmatamento no ano corrente, ao passo que municípios com baixo desmatamento no ano anterior (pontos azuis) tenderam a contribuir negativamente. Esse resultado era esperado dado o padrão de inércia evidenciado na análise de correlação entre as duas variáveis.

A segunda variável em importância foi *Áreas Embargadas*, seguida por *Autos de Infração (Flora)*. Ambas apresentaram associação positiva com o desmatamento: municípios com maiores valores dessas variáveis tenderam a registrar maiores taxas de supressão vegetal. Esse resultado reflete a dinâmica da fiscalização ambiental, posto que as ações são guiadas principalmente por sistemas como Prodes e Deter (INPE, 2026c), que identificam áreas desmatadas com base em sensoriamento remoto.

Assim, embargos e autuações concentram-se em focos de maior pressão sobre a floresta, tornando essas variáveis marcadoras de áreas em fronteira ativa de desmatamento. Ressalta-se apenas uma distinção temporal: enquanto *Áreas Embargadas* captura o passivo de infrações com sanções ainda vigentes no início do ano analisado, *Autos de Infração (Flora)* reflete o esforço fiscalizatório do mesmo ano da variável-alvo.

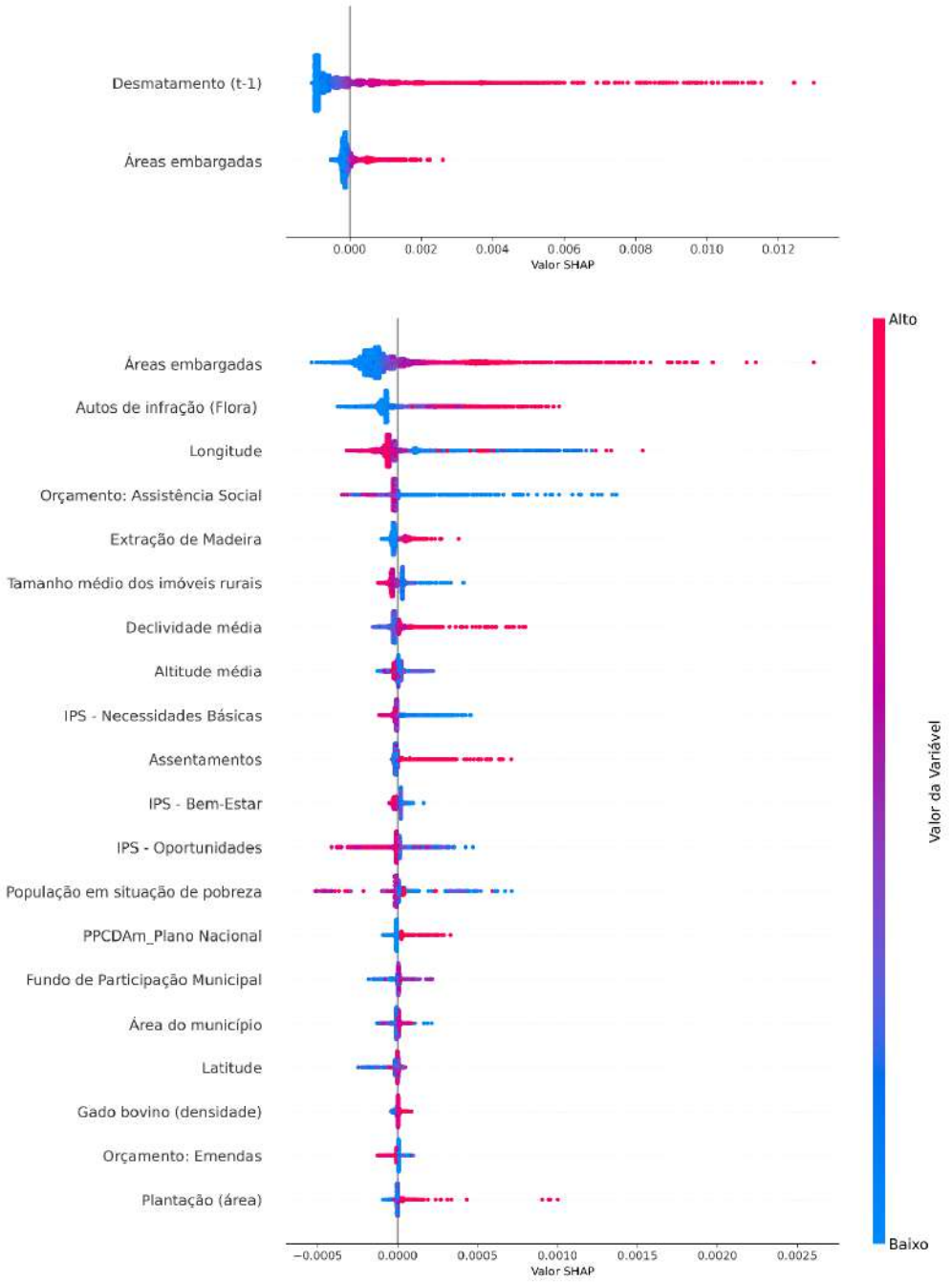
As variáveis *Orçamento: Assistência Social* e *IPS Necessidades Básicas* (abrange Nutrição e Cuidados Médicos Básicos, Água e Saneamento, Moradia e Segurança Pessoal) apresentaram padrão oposto: baixos valores das variáveis (pontos azuis) concentraram-se à direita do zero, indicando que municípios com menor orçamento de assistência social *per capita* e menor atendimento de necessidades básicas tenderam a registrar maiores taxas de desmatamento.

A interpretação da variável *Orçamento: Assistência Social* sugere que o menor aporte de recursos pode refletir duas dinâmicas complementares: i) uma menor densidade de população elegível a programas de transferência de renda nos municípios de maior desmatamento (hipótese corroborada pela tendência observada na variável 'População em situação de pobreza'); e/ou ii) uma estrutura insuficiente da rede de serviços socioassistenciais (CRAS, CREAS, benefícios eventuais) que poderia indicar fragilidade institucional para fazer em face das situações de vulnerabilidade associadas ao desmatamento (hipótese apoiada pela tendência de baixos valores de IPS *Necessidades Básicas, Bem-Estar* e *Oportunidades* a maiores taxas de desmatamento).

De acordo com Busch & Ferretti-Gallon (2023), os estudos analisados apontaram para maior pobreza associada a menores taxas de desmatamento, e os autores sugerem que haveria um suporte empírico misto para a ideia de que o desmatamento primeiro aumentaria e depois diminuiria à medida que a riqueza nacional crescesse.

Em um estudo empírico no estado do Pará, Diniz *et al.* (2025) também encontraram uma correlação negativa entre pobreza e desmatamento, ao mesmo tempo em que indicaram que a renda redistribuída pelo Programa Bolsa Família poderia potencialmente aumentar o envolvimento das famílias em situação de pobreza em atividades relacionadas ao desmatamento. Em função disso, apontam para a necessidade de que a transferência de renda seja complementada por oportunidades de emprego e geração de renda alternativa, desvinculando, assim, os beneficiários de atividades promotoras do desmatamento.

**Figura 3 – Distribuição dos Valores SHAP para as variáveis de maior magnitude - modelo *XGBoost* de desmatamento anual nos municípios da Amazônia Legal entre 2012 e 2021. No painel superior, destaque para *Desmatamento (t-1)*; no painel inferior, omissão de *Desmatamento (t-1)* para redução de escala e melhor visualização das demais variáveis.**




**Fonte:** Elaborado pela autora (2026).

O relatório técnico do Instituto Escolhas (2023) avaliou a hipótese de uma correlação positiva, na qual a diminuição da pobreza e/ou o aumento na disponibilidade de trabalho formal implicaria redução do desmatamento, por aumentar a capacidade de os indivíduos optarem por atividades menos insalubres. Como resultado, os autores avaliaram que a associação estatística identificada não permitiu chegar a uma inferência causal definitiva,

mas forneceu um primeiro grau de validação empírica da hipótese. Esse resultado contrasta com os demais ao colocar a correlação positiva entre pobreza e desmatamento, mas coaduna com Diniz *et al.* (2025) quanto à importância de oportunidades de emprego formal.

A relevância do *IPS - Necessidades Básicas* nas taxas de desmatamento poderia também ser explorada pela avaliação de seu subcomponente Segurança Pessoal, já que a expansão do desmatamento está intrinsecamente ligada à grilagem de terras e ao crime organizado ambiental, dinâmicas que se sustentam pela violência (Lima, 2023). Historicamente, desde 2012, a Amazônia Legal registra índices de violência letal superiores à média nacional. Além disso, a sobreposição do crime transnacional e da violência com ilícitos ambientais vem sendo intensificada e fortalecida pela presença de organizações criminosas na região (Lima, 2023).

A *Extração de Madeira* foi a atividade econômica com maior destaque, seguida por *Gado bovino (densidade)* e *Plantação (área)*, respectivamente. Todas apresentaram uma correlação positiva, ou seja, quanto maiores as atividades, maiores as taxas de desmatamento.

Na dinâmica de uso e ocupação do solo da Amazônia, as atividades agrícolas, em especial, da plantação de soja, costumam se expandir para antigas áreas de pastagem e não sobre áreas recém-desmatadas; assim, pode-se explicar a sua menor importância com relação às demais pela maior distância temporal.

Essa dinâmica de ocupação e o papel principal da pecuária como vetor do desmatamento vêm sendo consistentemente encontrados em estudos (Gomes *et al.*, 2020, Maranhão *et al.*, 2019; Delazeri, 2016; Domingues & Bermann, 2012; Martins & Francisco, 2025). Por outro lado, não há estudos apontando para a *Extração de Madeira* como vetor econômico mais importante do desmatamento. Granger & Andersen (2006) argumentam que muitas regressões municipais não capturam bem a relação entre extração madeireira e expansão agrícola na Amazônia, por conta da seletividade da atividade (madeiras nobres), que não promove o corte raso da vegetação, mas sim a degradação da área. Contudo, os autores apontam que a atividade pode contribuir por meio de i) abertura de estradas para escoamento da madeira, atraindo, e ii) venda de direitos de extração por fazendeiros em áreas já estabelecidas, facilitando a posterior "limpeza do terreno" para uso agropecuário e gerando capital.

Por fim, as variáveis de caracterização municipal — longitude, altitude, declividade, latitude e área — não representam apenas a descrição física, mas podem atuar como *proxies* do custo de oportunidade e da acessibilidade logística na Amazônia Legal. Nesse sentido, a geografia definiria a base de interação das demais variáveis, modulando a probabilidade de desmatamento em função da viabilidade física do acesso.

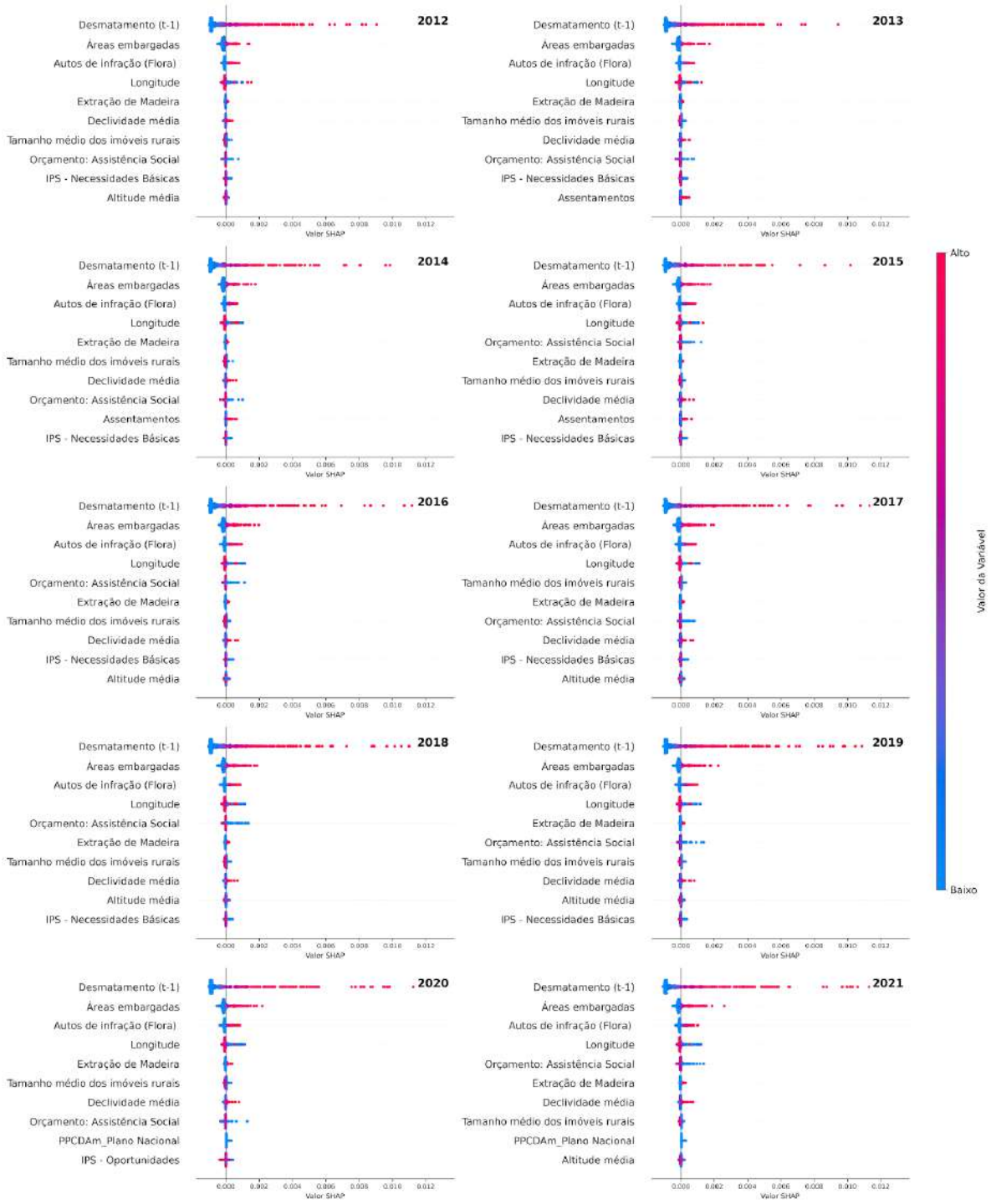
## 2.4. Contexto Temporal

A análise temporal revelou estabilidade ao longo do período 2012-2021, com as quatro variáveis mais importantes — *Desmatamento (t-1), Áreas Embargadas, Autos de Infração* e *Longitude*, se mantendo na mesma posição e com a mesma distribuição de valores do Contexto Geral. *Extração de Madeira, Tamanho Médio dos Imóveis Rurais, Declividade Média* e *Orçamento: Assistência Social* se alternaram nas posições subsequentes, mas, de modo geral, também semelhantes ao comportamento no Contexto Geral. (Figura 4)

Nas demais variáveis, as maiores alterações ocorreram nos anos de 2020 e 2021, quando *IPS Necessidades Básicas, Assentamentos* e *IPS Bem-Estar* tiveram sua importância reduzida ao passo que aumentou a relevância de *IPS Oportunidades* e *PPCDAM_Plano Nacional*. Cabe

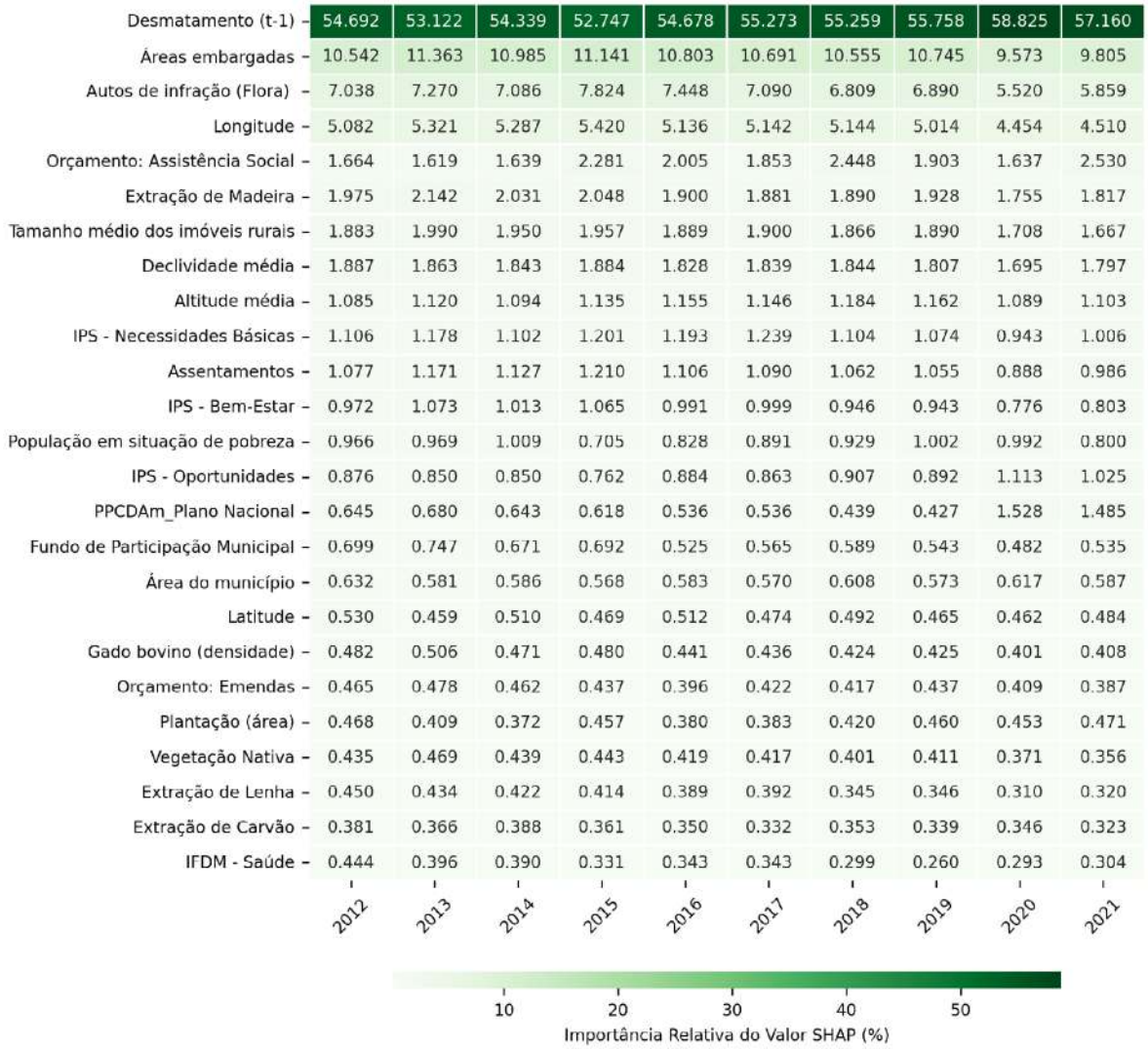
destacar que nesses anos a importância de *Desmatamento (t-1)* apresentou os maiores valores da série (58,8% e 57,1%), sugerindo reforço da inércia histórica — municípios que já desmatavam aceleraram o processo, e as variáveis *Áreas embargadas* e *Autos de Infração (Flora)* tiveram a importância reduzida. Esse período coincide com o período da pandemia de COVID-19 e de enfraquecimento do aparato federal de fiscalização ambiental (Vale *et al.*, 2021; Ziccardi *et al.* 2024). (Figura 5)

**Figura 4 - Distribuição dos Valores SHAP para as variáveis de maior magnitude, por ano - modelo *XGBoost* de desmatamento anual nos municípios da Amazônia Legal (2012 a 2021).**




**Fonte:** Elaborado pela autora (2026).

**Figura 5 - Matriz de evolução temporal da importância relativa SHAP das principais variáveis - modelo *XGBoost* de desmatamento anual nos municípios da Amazônia Legal (2012 a 2021).**




**Fonte:** Elaborado pela autora (2026).

## *2.5. Contexto Espacial*

De modo geral, as variáveis mais relacionadas ao desmatamento se mantiveram em todos os estados; contudo, alguns apresentaram características próprias, indicando maior heterogeneidade no contexto espacial do que no temporal.

Mato Grosso, Pará e Rondônia apresentaram distribuições semelhantes ao Contexto Geral. Em Rondônia, destaca-se a maior importância das variáveis relacionadas à vulnerabilidade socioeconômica (*Orçamento: Assistência Social, População em situação de pobreza, IPS Necessidades Básicas* e *IPS Oportunidades*). (Figura 6)

Amazonas, Acre, Maranhão e Tocantins também apresentaram hierarquia e distribuição dos valores alinhadas ao Contexto Geral; contudo, a escala foi reduzida, indicando menor grandeza do desmatamento. Cabe destacar que *Assentamentos* assumiu uma importância mais elevada no Acre. O estado de Roraima teve comportamento e magnitude semelhante a esse grupo, mas nele as variáveis *Terras Indígenas* e *Fundo de Participação Municipal* assumiram importâncias relativas muito maiores e figuraram entre as principais relacionadas ao desmatamento. (Figuras 6 e 7)

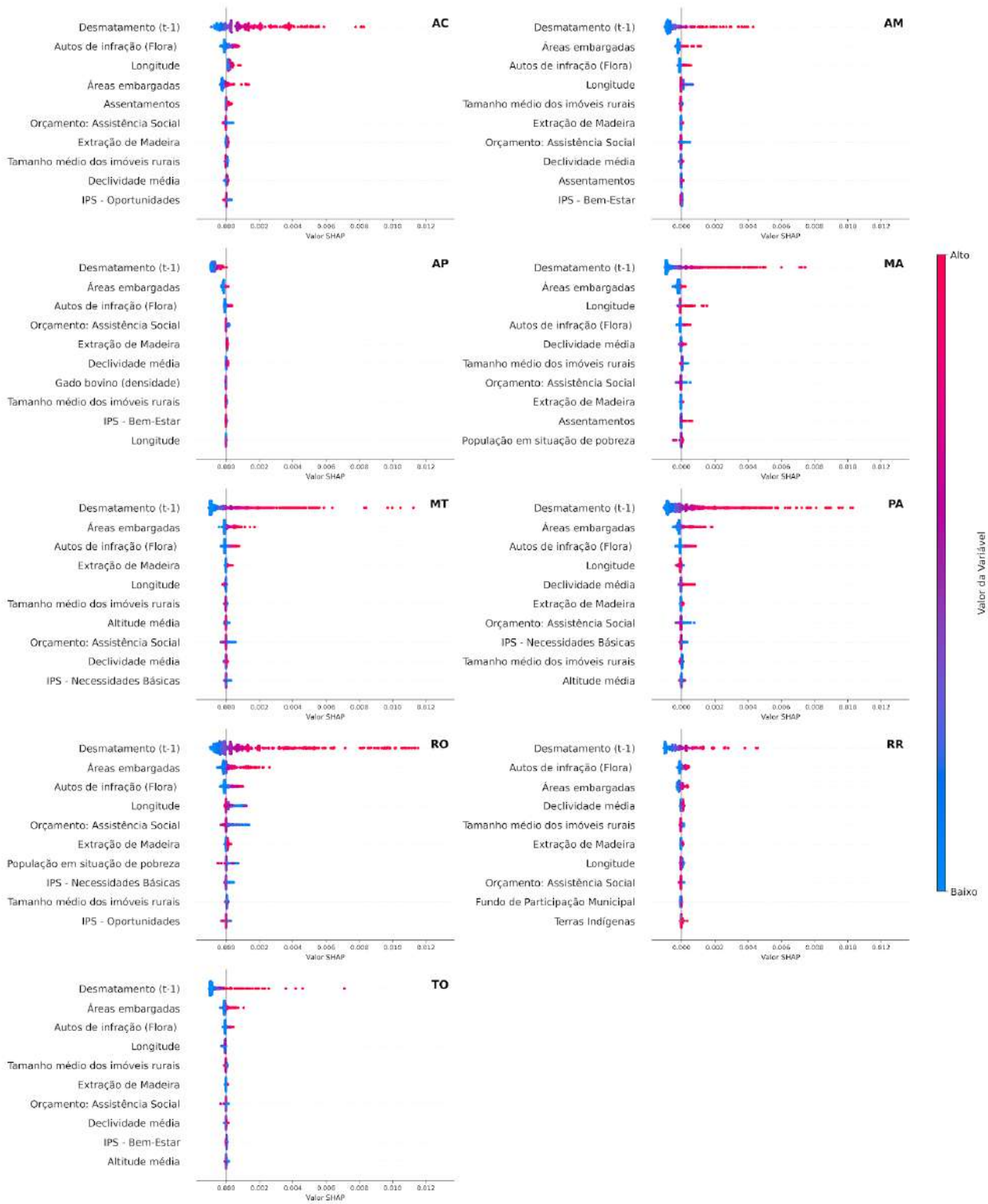
O estado do Amapá, por sua vez, apresentou padrão diferenciado dos demais, inclusive na principal variável, *Desmatamento (t-1)*, para a qual valores altos ou baixos tiveram impacto sempre negativo sobre o desmatamento, assim como *Áreas Embargadas*, provavelmente porque a média da taxa de desmatamento no estado é inferior à média global. Além disso, foi o único estado em que *Gado bovino (densidade)* esteve entre os dez maiores fatores. (Figuras 6 e 7)

Becker (2005) registrou três macrorregiões na Amazônia: i) o Arco do povoamento consolidado, onde estariam as cidades, as maiores densidades demográficas, as estradas e o cerne da economia; ii) a Amazônia central, que seria a porção mais vulnerável da Amazônia, porque é cortada pelos eixos e pelas estradas; e iii) a Amazônia ocidental, com maior área de fronteira política e a mais preservada, porque não cortada por estradas e com povoamento pontual, na Zona Franca de Manaus, que constituiria resistência ao desmatamento.

De acordo com Diniz (2009), o Arco do Povoamento Adensado incluiria os estados do Mato Grosso, Rondônia, Tocantins e parte do Pará (Sudeste e Nordeste); a Amazônia Central englobaria o restante do Pará (Oeste e Norte), o Norte do Amapá e o Vale do Rio Madeira no Amazonas; por fim, a Amazônia Ocidental abarcaria Roraima, o sul do Amapá e o restante do estado do Amazonas. O Maranhão não consta de nenhuma das macrorregiões.

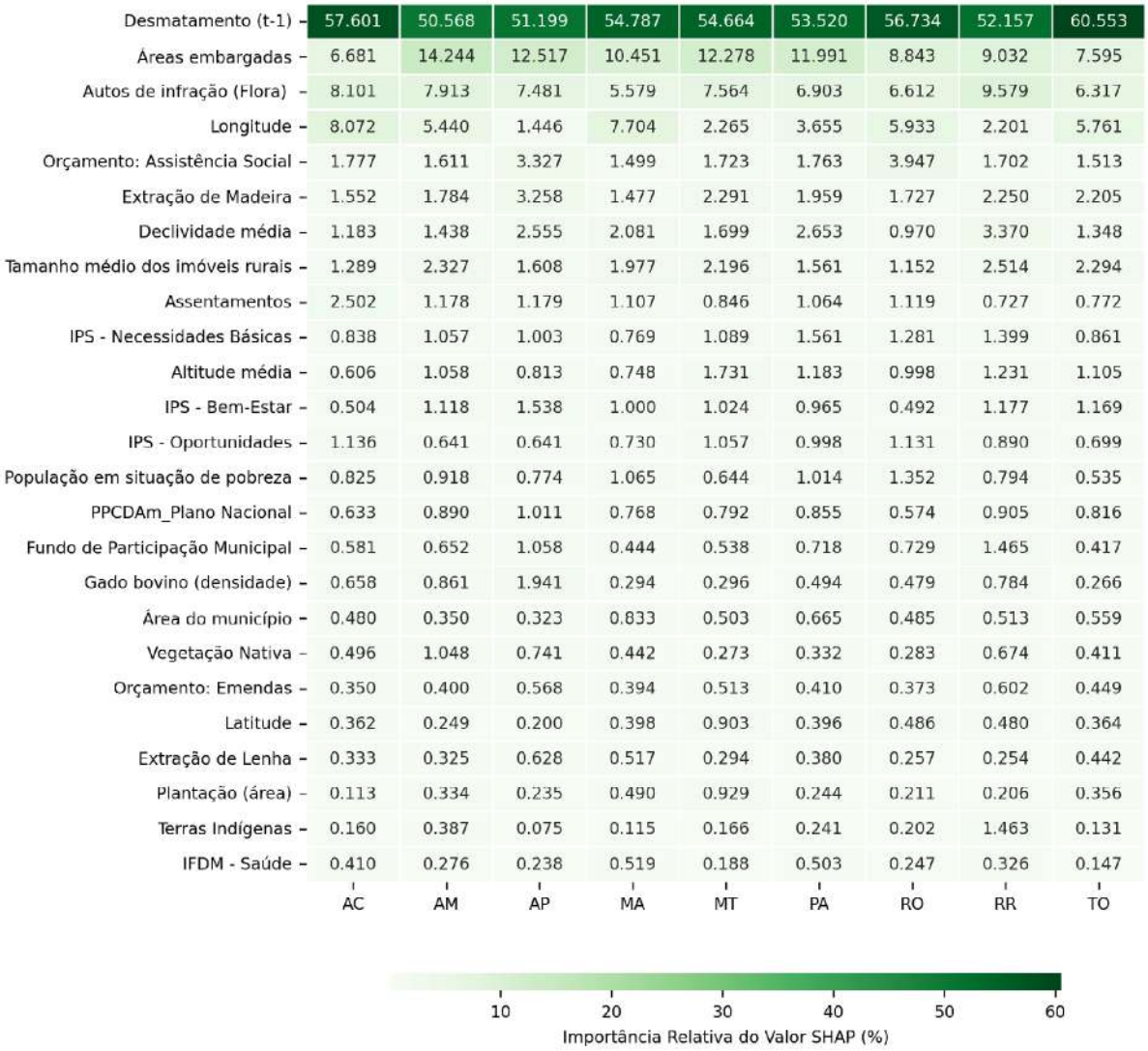
A análise por Estado dificulta a comparação exata com as macrorregiões, mas os resultados apontam para uma divisão semelhante, em especial com relação ao Arco do Povoamento Adensado e da Amazônia Central. Em estudo futuro, seria interessante regionalizar o Estado do Pará, do Amazonas e do Amapá e avaliar os resultados por macrorregião.

## Figura 6 - Distribuição dos Valores SHAP para as variáveis de maior magnitude, por Estado - modelo *XGBoost* de desmatamento anual nos municípios da Amazônia Legal (2012 a 2021).




**Fonte:** Elaborado pela autora (2026).

**Figura 7 - Matriz da importância relativa SHAP das principais variáveis por estado - modelo *XGBoost* de desmatamento anual nos municípios da Amazônia Legal (2012 a 2021).**

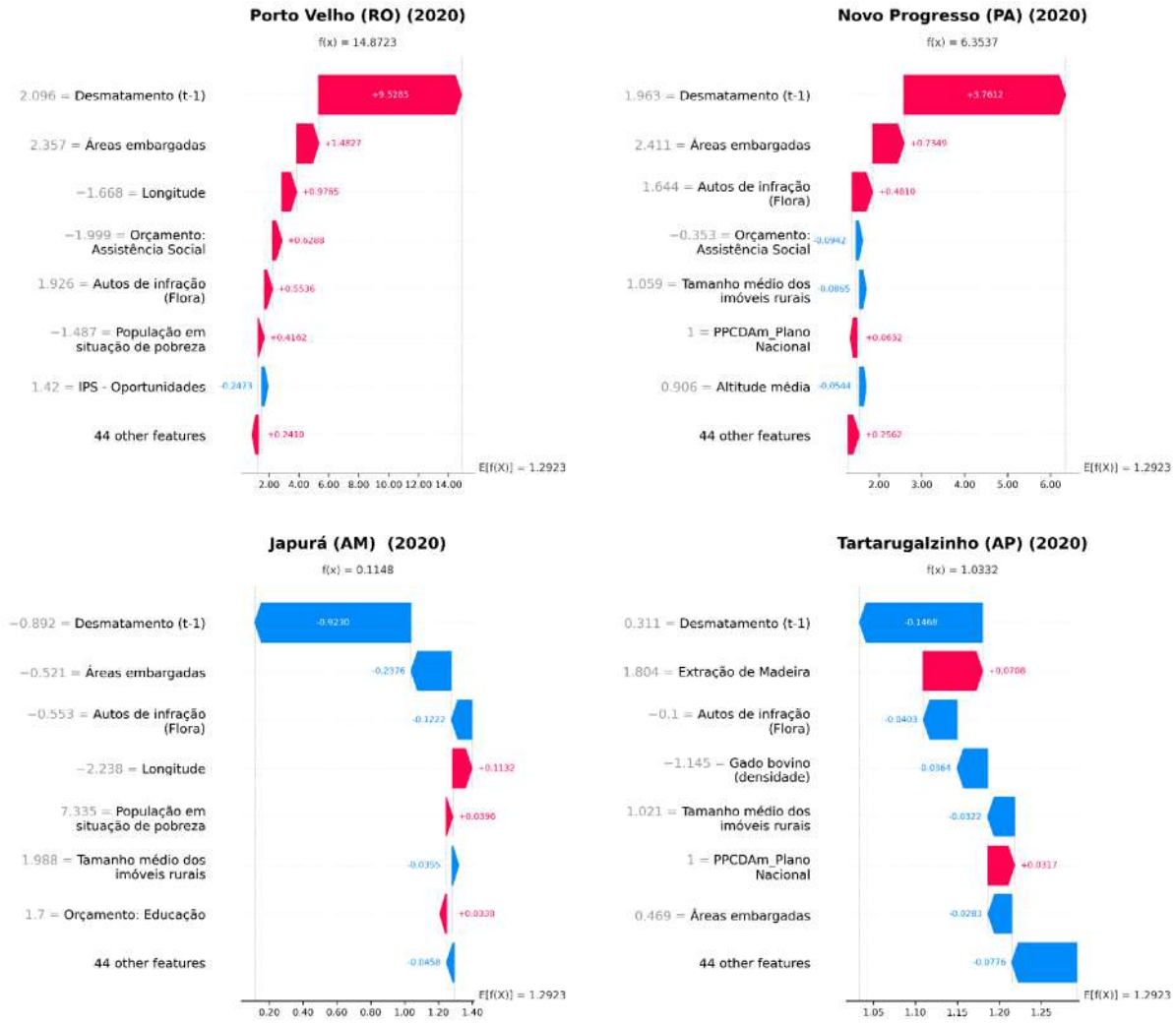



**Fonte:** Elaborado pela autora (2026).

## 2.6. Estudos de caso

Para ilustrar a heterogeneidade do desmatamento na Amazônia Legal, foi avaliada a decomposição SHAP para o ano de 2020 em quatro municípios com perfis contrastantes: Porto Velho (RO) e Novo Progresso (PA), localizados no Arco do Povoamento Adensado, Tartarugalzinho (AP), entre a Amazônia Central e a Amazônia Ocidental, e Japurá (AM), na Amazônia Ocidental.

**Figura 8. Decomposição SHAP da contribuição das variáveis para a descrição do desmatamento nos municípios selecionados em 2020. Modelo *XGBoost* de desmatamento anual nos municípios da Amazônia Legal (2012 a 2021). *Nota: Valores SHAP multiplicados por 1.000 para fins de visualização.***




**Fonte:** Elaborado pela autora (2026).

Porto Velho e Novo Progresso apresentaram altas magnitudes de impacto no desmatamento, com valores para 2020 de 0,0149 e 0,0063, respectivamente — substancialmente acima da média global do modelo (0,0013). Em ambos, a variável *Desmatamento (t-1)* respondeu isoladamente por mais da metade do desvio positivo em relação à média, destacando a inércia do desmatamento nesses municípios. *Áreas Embargadas* e *Autos de Infração* refletem a presença da fiscalização ambiental em resposta à intensidade da pressão sobre a floresta. Já a variável *Orçamento: Assistência Social* apresentou comportamento divergente: seu baixo valor contribuiu positivamente para o desmatamento em Porto Velho e negativamente em Novo Progresso, reforçando o caráter complexo das variáveis socioeconômicas. (Figura 8)

Por sua vez, os municípios de Japurá e Tartarugalzinho ilustraram cenários em que o conjunto de fatores locais contribuiu sistematicamente para manter as taxas de desmatamento abaixo da média global do modelo. Em Tartarugalzinho o valor foi ligeiramente abaixo da média global e em Japurá foi cerca de dez vezes menor. Em ambos, o *Desmatamento (t-1)* contribuiu negativamente, indicando que o histórico de baixo desmatamento pode ter atuado como fator de contenção (Figura 8).

A variabilidade na composição do gráfico de cada município reforça a heterogeneidade espacial e a análise pode apoiar a construção de ações locais para o combate ao desmatamento na Amazônia Legal.

## CONSIDERAÇÕES FINAIS

As considerações finais deste estudo partem da validação da abordagem metodológica adotada, que permitiu investigar a complexa dinâmica do desmatamento na Amazônia Legal sob uma perspectiva de alta dimensionalidade. A performance do modelo *XGBoost*, ao explicar aproximadamente 67% da variância, demonstra a viabilidade do aprendizado de máquina como ferramenta de análise, podendo complementar abordagens econométricas tradicionais sem substituí-las.

Este estudo contribui ao demonstrar o potencial de modelos de aprendizado de máquina para capturar a complexidade do desmatamento na Amazônia Legal, evidenciando o papel central da dependência temporal e a heterogeneidade espacial do fenômeno.

O achado de maior magnitude foi a autocorrelação temporal: a taxa de desmatamento no ano anterior respondeu por mais de 50% da importância relativa em todos os contextos analisados. Esse resultado é consistente com um processo de persistência espacial/temporal, que poderia ser explicado pela consolidação de um ambiente favorável, com a estruturação de uma rede de atores locais e a redução de custos marginais ao longo do tempo.

Em seguida, destacam-se *Áreas Embargadas* e *Autos de Infração (Flora)*, que refletem a orientação da fiscalização ambiental aos indicativos de desmatamento via sensoriamento remoto (Prodes e Deter). Assim, essas variáveis indicam a presença de políticas de comando e controle prioritariamente em municípios sob maior pressão, tornando-se marcadores de fronteira ativa do desmatamento.

A dimensão socioeconômica também foi relevante: municípios com menor orçamento de assistência social *per capita* e piores condições de atendimento de necessidades básicas (*IPS Necessidades Básicas*) tenderam a registrar maiores taxas de desmatamento. A relação entre pobreza e desmatamento, contudo, é complexa e não unidirecional (Busch & Ferretti-Gallon, 2023; Diniz *et al.*, 2025; Instituto Escolhas, 2023), por isso mereceria investigação mais aprofundada em estudos futuros.

Entre os vetores econômicos, *Extração de Madeira* emergiu como o fator mais associado ao desmatamento, acima da pecuária, que tem sido consistentemente reportada como maior fator de pressão direta sobre a floresta na bibliografia (Gomes *et al.*, 2020; Maranhão et al, 2019; Delazeri, 2016; Domingues & Bermann, 2012, Martins & Francisco, 2025). Essa aparente divergência merece uma investigação mais aprofundada em estudos futuros que, como complemento, também podem incorporar variáveis de preços de *commodities* (soja e boi) para enriquecer a avaliação da dinâmica econômica.

A análise temporal revelou estabilidade na hierarquia de importância das variáveis ao longo da maior parte da década, com alterações mais pronunciadas nos anos de 2020 e 2021. Nesses anos, a importância relativa de *Desmatamento (t-1)* atingiu os maiores valores da série (58,8% e 57,6%), ao mesmo tempo em que a relevância das variáveis de fiscalização ambiental se reduziu.

Esse período coincide com o da pandemia de COVID-19 e do enfraquecimento dos mecanismos federais de governança ambiental (Vale *et al.*, 2021; Ziccardi *et al.*, 2024), sugerindo que a aceleração do desmatamento nesse período foi conduzida predominantemente pela inércia de fronteiras já ativas, com menor resistência institucional.

A análise espacial por estado evidenciou que, embora as variáveis centrais se mantenham relevantes, sua importância relativa varia entre as unidades da federação, padrão que pode ser explorado em estudos futuros com base nas macrorregiões amazônicas (Diniz, 2009).

Os estudos de caso em Porto Velho (RO), Novo Progresso (PA), Japurá (AM) e Tartarugalzinho (AP) ilustraram a amplitude da heterogeneidade interna. A diferença de magnitude entre os municípios de dentro e de fora do Arco do Povoamento Adensado demonstra que médias regionais encobrem dinâmicas locais muito distintas, e que políticas uniformes podem ter limitada efetividade num território de tamanha diversidade.

Este estudo apresenta limitações que devem ser consideradas na interpretação dos resultados. O modelo captura associações condicionais nos dados, não relações causais; a identificação de causalidade requereria desenhos específicos.

A variável-alvo restringe-se ao desmatamento de vegetação primária, excluindo a degradação florestal e o desmatamento de vegetação secundária, dimensões também relevantes para o entendimento do fenômeno. Além disso, ela apresenta alta assimetria, mas optou-se por não transformá-la com o mesmo procedimento dado às variáveis independentes.

Também foram sempre considerados os valores das variáveis do mesmo ano do desmatamento, à exceção de *Desmatamento (t-1)*. Entretanto, para variáveis que eventualmente apresentem defasagem temporal, a correlação pode não ter sido capturada adequadamente.

A despeito das limitações, os resultados reforçam que o desmatamento na Amazônia Legal é um fenômeno multidimensional, no qual fatores econômicos, institucionais, geográficos e sociais interagem de forma complexa e heterogênea no espaço e no tempo.

O aprendizado de máquina, aliado à interpretabilidade via SHAP, revelou-se uma ferramenta com potencial para capturar essa complexidade de forma descritiva, oferecendo insumos relevantes para o desenho de políticas públicas territorialmente diferenciadas e empiricamente fundamentadas.

Por fim, os resultados sobre a vulnerabilidade social indicam que a agenda ambiental não pode estar dissociada da agenda social.

## REFERÊNCIAS


BECKER, Bertha K. Geopolítica da Amazônia. **Estudos Avançados**, São Paulo, Brasil, v. 19, n. 53, p. 71–86, 2005. Disponível em: https://revistas.usp.br/eav/article/view/10047. Acesso em: 18 mar. 2026.

BRASIL. Lei Complementar nº 124, de 3 de janeiro de 2007. Institui, na forma do art. 43 da Constituição Federal, a Superintendência do Desenvolvimento da Amazônia – SUDAM; estabelece sua composição, natureza jurídica, objetivos, área de competência e instrumentos de ação; dispõe sobre o Fundo de Desenvolvimento da Amazônia – FDA; altera a Medida Provisória nº 2.157-5, de 24 de agosto de 2001; revoga a Lei Complementar nº 67, de 13 de junho de 1991; e dá outras providências. **Diário Oficial da União**, Brasília, DF, 4 jan. 2007.

BUSCH, J.; FERRETTI-GALLON, K. What Drives and Stops Deforestation, Reforestation, and Forest Degradation? An Updated Meta-analysis. **Review of Environmental Economics and Policy**, v. 17, n. 2, 2023.

GRANGER, C.W.J.; ANDERSEN, L.E.. Modeling Amazon Deforestation for Policy Purposes. **Development Research Working Paper Series**. v. 12, 2006.

DELAZERI, L.M. Determinantes do Desmatamento nos Municípios do Arco Verde - Amazônia Legal: uma abordagem econométrica. **Revista Economia Ensaios**, Uberlândia, Brasil, v. 30, n. 2, 2016.

DINIZ, M.B.; OLIVEIRA JUNIOR, J.N.; TROMPIERI NETO, N.; DINIZ, M.J.T.. Causas do desmatamento da Amazônia: uma aplicação do teste de causalidade de Granger acerca das principais fontes de desmatamento nos municípios da Amazônia Legal brasileira. **Nova Economia**, 19(1), 121-151, 2009.

DINIZ, M.; SIMÕES, J.E.; ALVES, V.; DINIZ, M.; MOIA, G.. Relação entre pobreza e desmatamento na Amazônia brasileira: um estudo empírico para o estado do Pará. **EURE** Santiago, Chile, v. 51, n.154, pp.1-27, 2025. Disponível em: https://www.redalyc.org/journal/196/19682629006/html/. Acesso em: 18 mar. 2026.

DOMINGUES, M.S.D.; BERMANN, C. O arco de desflorestamento na amazônia: Da pecuária à soja. **Ambiente & Sociedade**, v.15 (2), 1-22, 2012.

GOMES, F.B.R.; CANELLA, C.D.M.; BRANCO, O.E.D.A.; CASTRO, M.C.C.S.; CASTRO, S.R. Deforestation in legal amazon: A panel data analysis of potential interferers.. The Journal of Men's Studies, 10 (2), 97-111, 2020.

MARTINS, H.P.L.; FRANCISCO, E.R. Determinantes do desmatamento histórico e recente da Amazônia na perspectiva da estatística espacial. **Cadernos Gestão Pública e Cidadania**, São Paulo, v. 30, p. e91454, 2025.

IBGE. **Amazônia Legal**. 2026. Disponível em: https://www.ibge.gov.br/geociencias/cartas-e-mapas/mapas-regionais/15819-amazonia-legal.html. Acesso em: 18 mar. 2026.

INPE. **Taxas de desmatamento – Amazônia Legal**. 2026a. Disponível em: https://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/legal_amazon/rates. Acesso em: 18 mar. 2026.

INPE. **Prodes: Monitoramento anual da supressão de vegetação nativa**. 2026b. Disponível em: https://data.inpe.br/biomasbr/prodes-monitoramento-anual-da-supressao-de-vegetacao-nativa/. Acesso em: 18 mar. 2026.

INPE. **Deter: Monitoramento Diário da Supressão e Degradação da Vegetação Nativa.** 2026c. Disponível em: https://data.inpe.br/biomasbr/deter-monitoramento-diario-da-supressao-e-degradacao-de-vegetacao-nativa/. Acesso em: 18 mar. 2026.

INSTITUTO ESCOLHAS. **A privação econômica é um determinante do desmatamento no Brasil?** Relatório Técnico. São Paulo, 2023. Disponível em: https://dssbr.ensp.fiocruz.br/wp-content/uploads/2023/04/Relatorio-Tecnico_Desmatamento-e-Pobreza.pdf. Acesso em: 18 mar. 2026.

LIMA, R.S. (coord.). **Segurança pública e crime organizado na Amazônia Legal.** Informe especial. São Paulo: Fórum Brasileiro de Segurança Pública, 2023. Disponível em: https://forumseguranca.org.br/wp-content/uploads/2023/06/informe-especial-seguranca-publica-e-crime-organizado-na-amazonia-legal.pdf. Acesso em: 18 mar. 2026.

MARANHÃO, R.L.A.; JÚNIOR, O.A.D.C.; HERMUCHE, P.M.; GOMES, R.A.T.,; PIMENTEL, C.M.M.; GUIMARÃES, R.F.. The spatiotemporal dynamics of soybean and cattle production in brazil. **Sustainability**, 11 (7), 2019.

MARGULIS, S. **Causas do desmatamento da Amazônia brasileira**. Brasília: Banco Mundial, 2003.

MMA. **Plano de Ação para Prevenção e Controle do Desmatamento na Amazônia Legal (PPCDAm).** 156 p. Brasília, DF: MMA, 2004.

MMA. **Plano de Ação para Prevenção e Controle do Desmatamento na Amazônia Legal (PPCDAm):** 2ª fase (2009-2011) Rumo ao Desmatamento Ilegal Zero. 165p.. Brasília, DF: MMA, 2009.

MMA. **Plano de Ação para prevenção e controle do desmatamento na Amazônia Legal (PPCDAm):** 3ª fase (2012-2015) Pelo Uso Sustentável e Conservação da Floresta. 174p. Brasília, DF: MMA, 2013.

MMA. **Plano de Ação para prevenção e controle do Desmatamento na Amazônia Legal (PPCDAm):** 4.ª fase 2016-2020. Brasília, DF: MMA, 2018.

MMA. **Plano de Ação para Prevenção e Controle do Desmatamento na Amazônia Legal (PPCDAm):** 5ª fase (2023 a 2027). Brasília, DF: MMA, 2023.

MMA. **Plano Nacional para o Controle do Desmatamento Ilegal e Recuperação da Vegetação Nativa (2020-2023).** 29p. Brasília, DF: MMA, 2020.

MULLAINATHAN, S.; SPIESS, J. Machine Learning: An Applied Econometric Approach. **Journal of Economic Perspectives.** 31 (2): 87–106, 2017.

PFAFF, A.S.P. What Drives Deforestation in the Brazilian Amazon: Evidence from Satellite and Socio-economic Data. **Journal of Environmental Economics and Management**, 37:26-43, 1999.

VALE, M.M; BERENGUER, E.; MENEZES, M.A.; CASTRO, E.B.V.; SIQUEIRA, L.P.; PORTELA, R.C.Q. The COVID-19 pandemic as an opportunity to weaken environmental protection in Brazil. **Biological Conservation**, 255, 108994, 2021.

ZICCARDI, L.; YANAI, A.M.; FEARNSIDE, P.M.; STARK, S.C.. Deforestation in the brazilian amazon: Threats and opportunities. https://doi.org/10.21203/rs.3.rs-4591425/v1nt, 2024.

## Quadro descritivo das variáveis presentes no banco de dados utilizados neste trabalho, contendo 7.478 registros. (Códigos no rodapé)

| Variável | Descrição | Tipo | Fonte | Tratamento | Média | Desvio padrão | % zeros | Assimetria | % Outliers IQR | Média | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
| **Identificação e Características Físicas** | | | | | | | | | | | | | | | | | | | |
| Chave de Registro | Código IBGE + Ano de Referência | I | s | | | | | | | | | | | | | | | | |
| Ano de Referência | Ano (2012 a 2021) | I | - | | | | | | | | | | | | | | | | |
| Estado (UF) | Unidade da Federação | I | a | | | | | | | | | | | | | | | | |
| Município | Código IBGE 7 dígitos | I | a | | | | | | | | | | | | | | | | |
| Latitude | Latitude do centroide | N | a | 3 | -7,0504 | 4,7908 | 0 | -0,16 | 0 | -7,0605 | -7,0493 | -7,0493 | -7,0493 | -7,0493 | -7,0493 | -7,0493 | -7,0493 | -7,0493 | -7,0493 |
| Longitude | Longitude do centroide | N | a | 3 | -52,5395 | 7,0087 | 0 | -0,92 | 0,80 | -52,5337 | -52,5401 | -52,5401 | -52,5401 | -52,5401 | -52,5401 | -52,5401 | -52,5401 | -52,5401 | -52,5401 |
| Área do município | Extensão territorial (km2) | N | a | | 6.655,76 | 13.746,04 | 0 | 5,49 | 9,88 | 6.644,21 | 6.657,04 | 6.657,04 | 6.657,04 | 6.657,04 | 6.657,04 | 6.657,04 | 6.657,04 | 6.657,04 | 6.657,04 |
| Densidade Demográfica | Número de habitantes por km² | N | a | 2 | 27,31 | 147,47 | 0 | 12,98 | 9,17 | 25,84 | 26,36 | 26,67 | 26,96 | 27,24 | 27,50 | 27,77 | 28,02 | 28,27 | 28,51 |
| Altitude média | Elevação média (m) | N | t | 3 | 201,69 | 146,73 | 0 | 0,68 | 0,80 | 201,94 | 201,66 | 201,66 | 201,66 | 201,66 | 201,66 | 201,66 | 201,66 | 201,66 | 201,66 |
| Declividade média | Inclinação média (Graus °) | N | u | 3 | 4,24 | 1,61 | 0 | 1,17 | 3,08 | 4,24 | 4,24 | 4,24 | 4,24 | 4,24 | 4,24 | 4,24 | 4,24 | 4,24 | 4,24 |
| Temperatura Média Anual | Temperatura diária média (°C) | N | v | 3 | 26,33 | 0,80 | 0 | -0,55 | 1,22 | 26,15 | 26,05 | 26,02 | 26,62 | 26,67 | 26,49 | 26,11 | 26,50 | 26,57 | 26,16 |
| **Desmatamento** | | | | | | | | | | | | | | | | | | | |
| Taxa de Desmatamento* | Área desmatada. | N | b | 3,2 | 0,127 | 0,257 | 30,37 | 4,45 | 10,31 | 0,092 | 0,103 | 0,098 | 0,114 | 0,122 | 0,131 | 0,113 | 0,167 | 0,157 | 0,176 |
| Desmatamento (t-1)* | Área desmatada no ano anterior. | N | b | 3,2 | 0,122 | 0,243 | 30,17 | 4,36 | 10,16 | 0,123 | 0,092 | 0,103 | 0,098 | 0,114 | 0,122 | 0,131 | 0,113 | 0,167 | 0,157 |
| **Indicadores Socioeconômicos** | | | | | | | | | | | | | | | | | | | |
| IPS - Bem-Estar | IPS: Dimensão Fundamentos para o bem-estar | N | c | 4,6 | 55,053 | 5,449 | 0 | 0,46 | 1,284 | 55,094 | 55,102 | 54,847 | 54,592 | 54,338 | 54,083 | 54,653 | 55,224 | 55,794 | 56,808 |
| IPS - Necessidades Básicas | IPS: Dimensão Necessidades Humanas Básicas | N | c | 4,6 | 64,311 | 6,638 | 0 | 0,05 | 0,749 | 63,408 | 63,424 | 63,493 | 63,562 | 63,631 | 63,700 | 64,564 | 65,429 | 66,294 | 65,607 |
| IPS - Oportunidades | IPS: Dimensão Oportunidades | N | c | 4,6 | 43,852 | 5,918 | 0 | -0,50 | 2,915 | 44,587 | 44,588 | 44,534 | 44,480 | 44,426 | 44,372 | 43,853 | 43,334 | 42,815 | 41,538 |
| IFDM - Emprego & Renda | Índice Firjan de Desenvolvimento Municipal – dimensão Emprego | N | d | 6 | 0,428 | 0,199 | 0 | 0,68 | 0 | 0,421 | 0,421 | 0,428 | 0,430 | 0,415 | 0,416 | 0,427 | 0,432 | 0,433 | 0,461 |

| Variável | Descrição | Tipo | Fonte | Tratamento | Média | Desvio padrão | % zeros | Assimetria | % Outliers IQR | Média | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
| IFDM - Saúde | Índice Firjan de Desenvolvimento Municipal – dimensão Saúde | N | d | 6 | 0,379 | 0,137 | 0 | 0,06 | 0 | 0,342 | 0,342 | 0,353 | 0,365 | 0,364 | 0,376 | 0,398 | 0,404 | 0,414 | 0,428 |
| População em situação de pobreza | População em situação de pobreza | N | e | 1 | 0,489 | 0,204 | 0 | 0,44 | 0,415 | 0,569 | 0,554 | 0,561 | 0,499 | 0,472 | 0,470 | 0,445 | 0,441 | 0,432 | 0,445 |
| **Economia e Infraestrutura** | | | | | | | | | | | | | | | | | | | |
| PIB *per capita* | Produto Interno Bruto**, em R$, *per capita* | N | a | | 19.374,69 | 26.684,35 | 0 | 10,23 | 8,10 | 12.585,44 | 14.009,65 | 14.995,79 | 15.887,78 | 17.470,01 | 18.595,83 | 19.782,43 | 20.708,24 | 25.905,40 | 33.788,19 |
| Impostos | Impostos, líquidos de subsídios, sobre produtos, em R$, *per capita* | N | a | 1 | 0,055 | 0,036 | 0,07 | 1,66 | 3,92 | 0,053 | 0,052 | 0,049 | 0,051 | 0,053 | 0,056 | 0,057 | 0,058 | 0,057 | 0,059 |
| V.A. Agropecuária | Valor adicionado bruto da Agropecuária, em R$, *per capita* | N | a | 1 | 0,267 | 0,169 | 0 | 0,72 | 1,19 | 0,269 | 0,277 | 0,267 | 0,265 | 0,265 | 0,254 | 0,240 | 0,236 | 0,276 | 0,323 |
| V.A. Industrial | Valor adicionado bruto da Indústria, em R$, *per capita* | N | a | 1 | 0,095 | 0,121 | 0,33 | 3,32 | 11,91 | 0,099 | 0,098 | 0,097 | 0,096 | 0,090 | 0,091 | 0,092 | 0,093 | 0,096 | 0,095 |
| V.A. Serviços | Valor adicionado bruto dos Serviços, em R$, *per capita* | N | a | 1 | 0,246 | 0,121 | 0 | 1,24 | 5,54 | 0,236 | 0,234 | 0,246 | 0,248 | 0,253 | 0,257 | 0,262 | 0,259 | 0,238 | 0,222 |
| V.A. Administração | Valor adicionado bruto da Administração, defesa, educação e saúde públicas e seguridade social, em R$, *per capita* | N | a | 1 | 0,393 | 0,164 | 0 | -0,05 | 0,01 | 0,396 | 0,391 | 0,390 | 0,392 | 0,392 | 0,399 | 0,405 | 0,412 | 0,390 | 0,359 |
| Exportação *per capita* | Valor exportado, em US$, (declarante, não produtor) *per capita* | N | g | 1 | 1.096,41 | 5.925,43 | 69,16 | 18,73 | 23,00 | 938,04 | 944,80 | 880,93 | 852,47 | 810,11 | 1.061,03 | 882,64 | 1.332,70 | 1.445,33 | 1.815,62 |
| Rodovias pavimentadas | Densidade de rodovias pavimentadas (km/km2) | N | f | 2, 3 | 0,028 | 0,032 | 13,11 | 2,25 | 5,62 | 0,028 | 0,028 | 0,028 | 0,028 | 0,028 | 0,028 | 0,028 | 0,028 | 0,028 | 0,028 |
| Tamanho médio dos imóveis rurais | Tamanho médio dos imóveis rurais, em hectares | N | h | | 2,55 | 3,76 | 0,13 | 4,27 | 7,89 | 2,55 | 2,55 | 2,55 | 2,55 | 2,55 | 2,55 | 2,55 | 2,55 | 2,55 | 2,55 |

| Variável | Descrição | Tipo | Fonte | Tratamento | Média | Desvio padrão | % zeros | Assimetria | % Outliers IQR | Média | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
| **Uso do Solo e Agropecuária** | | | | | | | | | | | | | | | | | | | |
| Vegetação Nativa* | Área coberta por vegetação nativa | N | i | 2 | 61,86 | 23,62 | 0 | -0,18 | 0 | 63,59 | 63,20 | 62,78 | 62,38 | 61,99 | 61,64 | 61,34 | 60,96 | 60,60 | 60,15 |
| Superfície de Água* | Área coberta por cursos d'água e massas d'água | N | i | 2 | 2,67 | 5,48 | 0,05 | 4,15 | 13,45 | 2,65 | 2,66 | 2,70 | 2,65 | 2,60 | 2,65 | 2,69 | 2,71 | 2,67 | 2,69 |
| Mineração (area)* | Área coberta por atividades de mineração | N | i | 2 | 0,047 | 0,22 | 80,72 | 7,42 | 19,28 | 0,035 | 0,037 | 0,039 | 0,041 | 0,043 | 0,045 | 0,047 | 0,050 | 0,065 | 0,073 |
| Plantação (área)* | Área destinada a lavouras | N | j | 2 | 5,88 | 12,32 | 0,41 | 4,57 | 10,91 | 4,94 | 5,21 | 5,59 | 5,77 | 5,67 | 5,97 | 6,00 | 6,23 | 6,49 | 6,88 |
| Soja (área)* | Área destinada ao plantio de soja | N | j | 2 | 21,66 | 31,13 | 60,60 | 0,99 | 0 | 15,19 | 17,68 | 19,77 | 21,74 | 22,58 | 22,52 | 23,33 | 23,86 | 24,49 | 25,45 |
| Galináceos (densidade) | Densidade de galináceos (cabeças/km²) | N | k | 2 | 96,29 | 560,23 | 0,07 | 14,82 | 12,74 | 60,52 | 68,21 | 86,78 | 108,40 | 110,45 | 92,69 | 102,86 | 108,08 | 110,22 | 114,60 |
| Suínos (densidade) | Densidade de suínos (cabeças/km²) | N | k | 2 | 2,544 | 6,42 | 0,12 | 8,93 | 11,26 | 2,662 | 2,351 | 2,371 | 2,700 | 2,564 | 2,542 | 2,582 | 2,565 | 2,518 | 2,588 |
| Gado bovino (densidade) | Densidade de gado bovino (cabeças/km²) | N | k | 2 | 39,142 | 40,62 | 0 | 1,58 | 4,32 | 36,669 | 36,814 | 37,582 | 38,328 | 38,977 | 38,942 | 39,093 | 40,039 | 41,571 | 43,399 |
| Assentamentos* | Área destinada a Assentamentos | N | l | 2, 3 | 10,104 | 15,20 | 23,01 | 2,28 | 9,76 | 10,036 | 10,064 | 10,093 | 10,112 | 10,113 | 10,117 | 10,125 | 10,125 | 10,125 | 10,125 |
| Terras Indígenas* | Área destinada a Terras Indígenas | N | m | 2, 3 | 8,396 | 18,20 | 62,30 | 2,58 | 21,13 | 8,413 | 8,394 | 8,394 | 8,394 | 8,394 | 8,394 | 8,394 | 8,394 | 8,394 | 8,394 |
| Unidades de Conservação* | Área destinada a Unidades de Conservação | N | n | 2, 3 | 15,819 | 27,84 | 50,11 | 1,83 | 13,68 | 15,609 | 15,603 | 15,603 | 15,757 | 15,757 | 15,898 | 15,906 | 16,002 | 16,003 | 16,049 |
| UC Proteção Integral* | Área destinada a Unidades de Conservação de Proteção Integral | N | n | 2, 3 | 3,051 | 9,30 | 77,15 | 3,81 | 22,85 | 2,997 | 2,990 | 2,990 | 3,012 | 3,012 | 3,051 | 3,081 | 3,126 | 3,126 | 3,126 |
| **Extrativismo** | | | | | | | | | | | | | | | | | | | |
| Extração de Madeira | Volume de madeira em tora extraída (m³/km²) | N | o | 2 | 2,150 | 6,482 | 44,49 | 6,11 | 18,23 | 2,506 | 2,508 | 2,260 | 2,102 | 1,800 | 1,850 | 1,884 | 1,988 | 1,911 | 2,694 |
| Extração de Carvão | Quantidade de carvão vegetal extraído (ton/km²) | N | o | 2 | 0,188 | 0,584 | 41,31 | 6,45 | 19,82 | 0,256 | 0,238 | 0,287 | 0,239 | 0,172 | 0,154 | 0,126 | 0,134 | 0,131 | 0,143 |
| Extração de Lenha | Volume de lenha extraída (m³/km²) | N | o | 2 | 4,340 | 16,113 | 21,69 | 14,86 | 10,28 | 5,970 | 5,339 | 5,304 | 5,007 | 4,404 | 3,997 | 3,546 | 3,195 | 3,205 | 3,440 |
| Extração de Oleaginosas | Quantidade de óleos vegetais extraídos (ton/km²) | N | o | 2 | 0,109 | 0,744 | 74,57 | 15,92 | 24,15 | 0,161 | 0,152 | 0,150 | 0,141 | 0,105 | 0,091 | 0,086 | 0,084 | 0,063 | 0,053 |
| Extração de Alimentícios | Quantidade de alimentíceos extraídos (ton/km²) | N | o | 2 | 0,176 | 1,238 | 48,34 | 14,60 | 17,38 | 0,148 | 0,152 | 0,152 | 0,163 | 0,167 | 0,175 | 0,186 | 0,192 | 0,222 | 0,208 |
| **Institucionais e Orçamento** | | | | | | | | | | | | | | | | | | | |

| Variável | Descrição | Tipo | Fonte | Tratamento | Média | Desvio padrão | % zeros | Assimetria | % Outliers IQR | Média | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
| Áreas embargadas* | Área com embargo ambiental | N | p | 2 | 0,071 | 0,184 | 21,22 | 5,82 | 11,50 | 0,056 | 0,060 | 0,063 | 0,066 | 0,071 | 0,072 | 0,080 | 0,081 | 0,080 | 0,080 |
| Autos de infração (Flora)* | Densidade de autuações contra a flora aplicadas (número de autos/km2) | N | p, q | 2 | 0,117 | 0,514 | 59,15 | 13,67 | 13,96 | 0,097 | 0,116 | 0,117 | 0,175 | 0,182 | 0,128 | 0,132 | 0,094 | 0,064 | 0,059 |
| Orçamento: Emendas | Emenda (R$) per capita | N | r | 1,5 | 755,95 | 2.045,15 | 50,98 | 10,55 | 11,27 | 1.231,06 | 1.229,03 | 1.229,03 | 256,66 | 347,51 | 1.019,05 | 237,92 | 889,40 | 504,22 | 616,84 |
| Orçamento: Educação | Repasse para educação (R$) per capita | N | r | 1,5 | 5.157,33 | 5.030,50 | 0,21 | 1,38 | 0,56 | 4.741,40 | 4.743,92 | 4.743,92 | 4.945,10 | 5.328,92 | 5.010,84 | 5.158,60 | 5.368,27 | 4.577,21 | 6.954,00 |
| Orçamento: Saúde | Repasse para saúde (R$) per capita | N | r | 1,5 | 6.342,81 | 3.631,18 | 0 | 4,04 | 4,48 | 4.473,57 | 4.469,80 | 4.469,80 | 4.620,47 | 5.114,45 | 5.161,63 | 7.803,29 | 7.463,31 | 10.227,58 | 9.619,20 |
| Orçamento: Assistência Social | Repasse para assistência social (R$) per capita | N | r | 1,5 | 732,47 | 551,48 | 0,03 | 2,82 | 6,49 | 684,69 | 683,32 | 683,32 | 641,80 | 816,85 | 809,69 | 704,06 | 824,85 | 1.004,78 | 471,17 |
| Orçamento: Outros | Repasse para outros (R$) per capita | N | r | 1,5 | 1.237,19 | 2.728,63 | 38,91 | 7,54 | 9,49 | 0 | 0 | 1.038,68 | 1.334,19 | 1.184,70 | 937,95 | 1.304,28 | 1.815,48 | 2.604,11 | 2.149,20 |
| Fundo de Participação Municipal | Fundo de Participação dos Municípios (R$) per capita | N | r | 1 | 108,55 | 87,51 | 0 | 3,43 | 9,98 | 81,48 | 82,43 | 88,42 | 90,29 | 142,68 | 102,95 | 110,99 | 119,88 | 116,53 | 149,74 |
| Programa Aquisição de Alimentos | Recurso destinado ao Programa de Aquisição de Alimentos (R$) per capita | N | r | 1 | 5,51 | 13,00 | 43,34 | 5,98 | 11,13 | 7,85 | 4,75 | 5,84 | 6,73 | 5,99 | 7,01 | 4,81 | 2,93 | 5,42 | 3,77 |
| PPCDAm | Plano de combate ao desmatamento vigente | C | s | | | | | | | PPCDAm 3 | | | | PPCDAM 4 | | | | Plano Nacional | |

* Valor da variável multiplicado por 100 para facilitar a visualização.

** PIB = Valor Adicionado Bruto total (VAB total ) + Impostos // VAB total = VAB Agropecuária + VAB Indústria + VAB Serviços + VAB da Administração

**CÓDIGOS**

**Tipo: I** = Identificadora, **N** = Numérica, **C** = Categórica.

**Fonte: a** = IBGE (https://www.ibge.gov.br/ e https://sidra.ibge.gov.br/); **b** = Prodes/INPE (https://terrabrasilis.dpi.inpe.br/downloads/); **c** = IPS Amazônia (https://ipsamazonia.org.br/);

**d** = Firjan (https://www.firjan.com.br/ifdm/downloads/); **e** = CadÚnico (https://aplicacoes.cidadania.gov.br/vis/data3/data-explorer.php);

**f** = BC250 – IBGE, 2025 (https://geoftp.ibge.gov.br/cartas_e_mapas/bases_cartograficas_continuas/bc250/); **g** = Comexstat (SECEX/MDIC - https://comexstat.mdic.gov.br);

**h** = SFB/SICAR (https://consultapublica.car.gov.br); **i** = MapBiomas (https://brasil.mapbiomas.org/); **j** = IBGE - PAM (https://sidra.ibge.gov.br/tabela/5457);

**k** = IBGE – PPM (https://sidra.ibge.gov.br/tabela/3939); **l** = INCRA (https://certificacao.incra.gov.br/csv_shp/export_shp.py);

**m** = FUNAI (https://www.gov.br/funai/pt-br/atuacao/terras-indigenas/geoprocessamento-e-mapas); **n** = CNUC/MMA (https://cnuc.mma.gov.br/);

**o** = IBGE/PEVS (https://sidra.ibge.gov.br/pesquisa/pevs/tabelas); **p** = Ibama (https://dados.gov.br/dados/conjuntos-dados/fiscalizacao-auto-de-infracao e /fiscalizacao-termo-de-embargo );

**q** = ICMBio (https://www.gov.br/icmbio/pt-br/assuntos/dados_geoespaciais/mapa-tematico-e-dados-geoestatisticos-das-unidades-de-conservacao-federais);

**r** = Tesouro Nacional (https://portaldatransparencia.gov.br/download-de-dados); **s** = construído; **t** = NASADEM HGT v001 (provedor original NASA JPL);

**u** = NASADEM (via ee.Terrain) (provedor original NASA JPL); **v** = ERA5-Land Hourly (provedor original ECMWF / Copernicus).

**Tratamento: 1** = normalizado pela população; **2** = normalizado pela área (densidade ou fração da área municipal); **3** = calculadado por geoprocessamento (arquivo vetorial da Amazônia Legal IBGE em https://www.ibge.gov.br/geociencias/cartas-e-mapas/mapas-regionais/15819-amazonia-legal.html); **4** = interpolação linear; **5** = carry-forward; **6** = back-fill.

# Determinantes da cobertura vacinal nos municípios brasileiros: a primazia das condições socioeconômicas sobre o comportamento eleitoral

*Tadeu Gomes Teixeira*
*Fabio Kouri Paim*
*Rafael Luís Giacomin*
*Thiago Mello Venâncio*
*Luís César Seixas de Oliveira*
*Alexandre Galardinovic Ribeiro*
*Bruno Gustavo Janoca*

## Resumo

A cobertura vacinal no Brasil declinou na última década, e atribui-se com frequência esse recuo ao comportamento político do eleitorado. Este estudo investiga os determinantes da cobertura vacinal nos 5.570 municípios brasileiros, ponderando a contribuição relativa das condições socioeconômicas e do comportamento político-eleitoral. A metodologia integra os resultados do primeiro turno presidencial de 2018 e 2022 (TSE), a cobertura de BCG, Tríplice Viral e COVID-19 com duas doses (SI-PNI) e a população e a renda municipais (IBGE), estimando as associações com partição pelo percentual de votos, ponderação populacional, erros-padrão robustos e diagnósticos espaciais e de resíduos, num desenho integralmente reprodutível. Os resultados mostram que as condições socioeconômicas, sobretudo a renda, organizam a cobertura, ao passo que a associação com a votação em Bolsonaro é fraca, instável entre vacinas e anos, e desaparece quando os municípios são ponderados pela população. O declínio da cobertura antecede 2018 e a sua distribuição é fortemente estruturada no espaço. Conclui-se pela primazia dos determinantes socioeconômicos e estruturais sobre a explicação político-eleitoral, com implicações para as políticas de recuperação da cobertura vacinal.

**Palavras-chave:** cobertura vacinal; determinantes socioeconômicos; comportamento eleitoral; políticas públicas de saúde; reprodutibilidade.

## *INTRODUÇÃO*

A cobertura vacinal no Brasil, depois de décadas de expansão sustentada que fizeram do Programa Nacional de Imunizações uma das políticas públicas de saúde mais reconhecidas do país (Domingues et al., 2020), passou a apresentar declínios sucessivos a partir de meados da década de 2010 (Sato, 2018; Arroyo et al., 2020). A reversão de uma trajetória ascendente, num programa historicamente bem avaliado, suscitou um conjunto amplo de explicações que vão desde fatores organizacionais do sistema de saúde, como a fragmentação dos registros e as dificuldades de logística e de definição de denominadores populacionais, até

fenômenos de natureza social e política, entre os quais a hesitação vacinal, fenômeno que a Organização Mundial da Saúde elencou entre as dez maiores ameaças à saúde global (World Health Organization, 2019). A pandemia de COVID-19 deu a essa discussão uma dimensão nova, ao colocar a vacinação no centro de uma disputa política que opôs autoridades sanitárias e a Presidência da República, e ao associar a defesa ou a recusa de medidas de proteção a identidades políticas determinadas.

Nesse contexto, formou-se a hipótese, recorrente no debate público e em parte da literatura (Gramacho; Turgeon, 2021; Seara-Morais et al., 2023), de que o eleitorado alinhado a Jair Bolsonaro seria também mais permeável ao discurso contrário à vacinação, de modo que os municípios com maior votação no então presidente exibiriam menor adesão aos programas de imunização. Trata-se de hipótese atraente pela sua plausibilidade aparente, mas de verificação delicada, porque os dados disponíveis são agregados por município, ao passo que a proposição se refere, em última instância, ao comportamento de indivíduos. A passagem de um nível ao outro é exatamente o terreno em que se incorre na falácia ecológica, isto é, a inferência indevida sobre indivíduos a partir de correlações observadas entre agregados (Robinson, 1950; King, 1997).

A investigação dos determinantes da cobertura vacinal enfrenta, porém, armadilhas de mensuração capazes de inverter conclusões. A mais relevante diz respeito à própria medida da preferência política: comparar municípios pelo número absoluto de votos, e não pelo percentual, confunde a adesão a um candidato com o porte populacional, pois o volume de votos é governado, antes de tudo, pelo tamanho do eleitorado. Como se demonstrará, essa escolha aparentemente trivial é capaz de gerar, sozinha, a aparência de uma associação política forte onde ela não existe. O presente artigo enfrenta essas armadilhas adotando, desde o desenho, a mensuração adequada de cada variável, a ponderação pela população e um procedimento integralmente reprodutível, de modo a identificar os determinantes reais da cobertura.

O objetivo é identificar os fatores que organizam a cobertura vacinal entre os municípios brasileiros e, em particular, ponderar a contribuição relativa das condições socioeconômicas e do comportamento político-eleitoral. Para tanto, o estudo reconstrói as bases a partir de fontes públicas, mede a preferência política pela variável adequada, o percentual de votos válidos, pondera as estimativas pela população, adota erros-padrão robustos e examina a estrutura espacial dos dados, num desenho cujo código e cujos dados derivados são disponibilizados para verificação independente. A reprodutibilidade não é, aqui, um adorno, mas a condição que permite distinguir determinantes substantivos de artefatos de mensuração, problema central quando bases administrativas de grande porte tornam trivial a obtenção de significância estatística.

A relevância do problema não é apenas acadêmica. O declínio da cobertura vacinal teve consequências sanitárias concretas, entre as quais a reintrodução de doenças que se julgavam controladas e a perda, em 2019, da certificação de eliminação do sarampo que o país havia conquistado (Organização Pan-Americana da Saúde, 2019). Compreender os determinantes desse declínio é condição para a formulação de políticas de recuperação da cobertura, e a atribuição de peso indevido a fatores políticos, em detrimento dos estruturais e socioeconômicos, pode orientar mal a ação pública. Se a queda da cobertura fosse primariamente um fenômeno de adesão política, as respostas seriam de comunicação e enfrentamento do discurso anti-vacina; se fosse primariamente um fenômeno de acesso e de condições socioeconômicas, as respostas seriam de fortalecimento da atenção primária e de redução de desigualdades. A correta identificação dos determinantes tem, assim, implicações diretas sobre o desenho das políticas.

Este estudo inscreve-se no campo da ciência de dados aplicada às políticas públicas e mobiliza técnicas de análise exploratória, testes de hipótese, correlação e regressão sobre bases administrativas de grande porte. Essa filiação traz vantagens, como a cobertura censitária dos municípios e a objetividade das fontes, mas também riscos específicos, associados à facilidade com que grandes bases produzem significância estatística e à multiplicidade de decisões de tratamento que se interpõem entre o dado bruto e o resultado. A atenção a essas decisões (qual variável mede a preferência, como tratar denominadores instáveis, quando ponderar as estimativas) é o que separa a identificação de determinantes reais da produção de correlações espúrias, e organiza a análise que se segue.

## 1. REFERENCIAL TEÓRICO E ANTECEDENTES

A discussão sobre os determinantes da cobertura vacinal articula escalas distintas. No plano macroestrutural, a cobertura depende da capacidade do sistema de saúde de adquirir, distribuir e registrar imunobiológicos e da qualidade das estimativas populacionais que servem de denominador aos indicadores. No plano individual e comunitário, depende da confiança nas instituições de saúde, do acesso aos serviços e de fatores socioeconômicos que condicionam tanto a oferta quanto a procura. A literatura sobre hesitação vacinal tem insistido em que a recusa ou o adiamento da vacinação, ainda que haja serviços disponíveis, é fenômeno multifatorial, no qual confiança, complacência e conveniência se combinam de maneiras específicas a cada contexto (Macdonald, 2015; World Health Organization, 2019). Reduzir esse fenômeno a uma única dimensão, política, é simplificação que a própria literatura desautoriza.

No caso brasileiro, a interpretação dos indicadores de cobertura exige cautela adicional. Os sistemas de informação do Programa Nacional de Imunizações registram doses aplicadas, e a cobertura é calculada pela razão entre essas doses e uma população-alvo estimada. Quando o numerador e o denominador provêm de fontes distintas, ou quando há registro de doses aplicadas em residentes de outros municípios, a razão resultante pode ultrapassar 100%, e por vezes assumir valores manifestamente impossíveis. Esse fenômeno, longe de ser exceção, afeta parcela substancial dos municípios brasileiros, sobretudo os de menor porte, nos quais pequenas flutuações no numerador ou no denominador produzem grandes variações relativas (Sato, 2015). Qualquer análise que utilize a cobertura municipal como variável precisa, portanto, reportar e tratar esses valores, sob pena de confundir ruído de registro com variação substantiva.

A relação entre orientação política e comportamento de saúde durante a pandemia de COVID-19 foi objeto de atenção crescente. Estudos sobre diferentes países sugeriram que a adesão a medidas não farmacológicas e, posteriormente, à vacinação, correlacionou-se com a filiação política (Allcott et al., 2020), num quadro de politização sem precedentes de uma emergência sanitária. A transposição desses achados para o nível municipal brasileiro, contudo, enfrenta o já mencionado problema da inferência ecológica, agravado pela enorme heterogeneidade dos municípios brasileiros, que variam em população de algumas centenas a mais de dez milhões de habitantes. Essa heterogeneidade torna especialmente perigosas as variáveis que confundem porte e comportamento, como se verá ser o caso do número absoluto de votos.

Há, por fim, um antecedente de ordem metodológica que perpassa todo o trabalho. A distinção entre significância estatística e magnitude do efeito é antiga na literatura, e a advertência de que, em amostras grandes, diferenças triviais alcançam significância a níveis arbitrariamente baixos é recorrente (Cohen, 1994). Com 5.570 unidades, praticamente qualquer diferença sistemática, mesmo desprovida de relevância prática ou produzida por um confundidor, atravessa os limiares convencionais de significância. A leitura de p-valores, nesse regime, precisa subordinar-se ao exame das magnitudes, dos tamanhos de efeito e da robustez, e não o contrário. Esse princípio orienta a análise que se segue.

A inferência ecológica, problema central deste trabalho, é tema clássico das ciências sociais quantitativas desde a demonstração de que correlações entre agregados podem divergir, em magnitude e até em sinal, das correlações entre os indivíduos que os compõem (Robinson, 1950). A literatura posterior sistematizou as condições sob as quais a inferência do agregado para o indivíduo é possível, e mostrou que, em geral, ela exige suposições fortes e raramente verificáveis (King, 1997). A esse problema soma-se o do efeito da unidade de área modificável, segundo o qual os resultados de análises espaciais dependem da escala e do recorte das unidades adotadas, de modo que conclusões obtidas no nível municipal podem não se sustentar em outras agregações (Openshaw, 1984). Esses dois problemas, conjugados à heterogeneidade dos municípios brasileiros, recomendam prudência extrema na leitura de correlações municipais como evidência sobre comportamento político individual.

A questão do denominador nos indicadores de cobertura merece desenvolvimento, por ser determinante para a interpretação dos resultados. A cobertura vacinal é, por definição, uma razão entre doses aplicadas e população-alvo, e a qualidade dessa razão depende da qualidade de ambos os termos. As estimativas populacionais municipais, produzidas entre os censos por métodos de projeção, acumulam erro à medida que se distanciam do último censo, e esse erro é proporcionalmente maior nos municípios pequenos. Quando a população estimada subestima a população real, ou quando se registram doses aplicadas em residentes de outros municípios, a cobertura calculada ultrapassa 100%, e em municípios de poucos milhares de habitantes pode atingir valores absurdos. Esse não é um defeito ocasional dos dados, mas uma característica estrutural que qualquer análise séria precisa enfrentar, seja pela limitação superior dos valores, seja pela ponderação, seja pelo uso de unidades de maior porte (Sato, 2015).

Por fim, a relação entre significância estatística e tamanho de efeito perpassa toda a análise. A crítica ao uso acrítico de p-valores, e à confusão entre significância e importância, é consolidada na literatura estatística (Cohen, 1994; Wasserstein; Lazar, 2016). Em bases com milhares de observações, a hipótese nula de ausência exata de associação é quase sempre rejeitada, porque diferenças minúsculas tornam-se detectáveis. O que importa, nesse regime, é a magnitude do efeito, medida por coeficientes de determinação e tamanhos de efeito padronizados, e a sua robustez a especificações alternativas. Um coeficiente de determinação da ordem de poucos milésimos, ainda que acompanhado de p-valor ínfimo, indica que a variável explicativa praticamente não explica a variação da resposta.

## 2. METODOLOGIA

A análise reconstruiu as bases de dados a partir de fontes oficiais públicas, integrando, no nível municipal, resultados eleitorais, cobertura vacinal, população e renda. A opção por construir toda a cadeia a partir das fontes primárias e por disponibilizar o código e os dados derivados faz da reprodutibilidade um atributo do próprio desenho: cada número apresentado pode ser regenerado de forma independente, condição que, num tema socialmente sensível, distingue determinantes substantivos de artefatos de tratamento dos dados.

Os resultados eleitorais foram extraídos do repositório de dados abertos do Tribunal Superior Eleitoral, nos arquivos de votação por candidato, município e zona do primeiro turno das eleições presidenciais de 2018 e 2022 (Tribunal Superior Eleitoral, 2024). A agregação por município produziu, para cada unidade, o total de votos válidos e o total de votos de Jair Bolsonaro, dos quais se derivou o percentual de votos válidos, variável que mede a intensidade da preferência política. A consistência da extração foi verificada pela aderência aos resultados oficiais: os percentuais nacionais de votos válidos, 46,01% em 2018 e 43,20% em 2022, reproduzem os números do Tribunal Superior Eleitoral. Registrou-se, desde já, que a média municipal do percentual de votos, 38,7% em 2018, difere substancialmente do percentual nacional, porque a média simples entre municípios atribui peso igual a unidades de população muito desigual, ponto que a análise tratou pela via da ponderação.

A cobertura vacinal de BCG e de Tríplice Viral, primeira dose, foi obtida do Sistema de Informações do Programa Nacional de Imunizações, por meio do aplicativo de tabulação do DataSUS, para os 5.570 municípios e para a série de 1994 a 2022 (Brasil, 2023). A cobertura, na definição do próprio sistema, é a razão entre as doses aplicadas e a população-alvo, multiplicada por cem. A cobertura da vacina contra a COVID-19 com duas doses, que não integra a base de cobertura do Programa Nacional de Imunizações, foi reconstruída a partir dos microdados de vacinação contra a COVID-19, originários do mesmo sistema e da Rede Nacional de Dados em Saúde, agregando-se, por município de residência e até 31 de dezembro de 2022, as aplicações de segunda dose, divididas pela população municipal de 2021 e multiplicadas por cem. A mediana resultante foi de 81,1%, e 8,5% dos municípios excederam 100%, artefato de denominador comum a essas bases e retomado na discussão.

A população municipal por ano proveio das estimativas do Instituto Brasileiro de Geografia e Estatística, e o controle socioeconômico foi construído a partir do Produto Interno Bruto *per capita* municipal, da mesma fonte, para 2018 e 2021 (Instituto Brasileiro de Geografia e Estatística, 2023). O Produto Interno Bruto *per capita* é tomado como indicador socioeconômico (proxy das condições de renda, de obtenção pública e reprodutível), e os resultados a ele relativos devem ser lidos sob essa qualificação. O casamento entre as bases foi feito pelo código de município do IBGE, com a conversão do código do Tribunal Superior Eleitoral por tabela de correspondência.

O procedimento de análise observa cinco cuidados metodológicos. Em primeiro lugar, a comparação de grupos particiona os municípios pela mediana do percentual de votos válidos, e não pela mediana dos votos absolutos, de modo que os grupos representem maior e menor adesão política, e não maior e menor porte populacional. Em segundo lugar, todas as medidas de associação são calculadas também em versão ponderada pela população municipal, para que municípios com poucos habitantes não pesem o mesmo que capitais na estimativa. Em terceiro lugar, a fração de municípios com cobertura superior a 100% é reportada, e as análises são replicadas com a cobertura limitada a 100% como

verificação de robustez. Em quarto lugar, as regressões adotaram erros-padrão robustos à heterocedasticidade, na forma do estimador consistente de White (White, 1980), e foram reestimadas por mínimos quadrados ponderados pela população. Em quinto lugar, o crescimento geométrico da cobertura da Tríplice Viral foi recalculado com proteção contra denominadores instáveis, descartando-se os municípios com cobertura inicial inferior a 50% e winsorizando-se os valores extremos antes de qualquer interpretação.

Para isolar o comportamento estatístico da partição por número absoluto de votos, conduziu-se ainda uma simulação controlada. Geraram-se 5.570 unidades sintéticas em que o eleitorado variava em ordens de grandeza, como entre os municípios reais, a preferência política era independente da cobertura por construção, e a cobertura dependia apenas do porte da unidade. Sobre esses dados, replicaram-se os dois procedimentos de partição, o por votos absolutos e o por percentual de votos, de modo a observar qual deles produziria significância na ausência, deliberada, de qualquer relação política. Todo o tratamento foi implementado em linguagem Python, com as bibliotecas usuais de manipulação de dados e de modelagem estatística, e o código encontra-se disponível para reexecução.

A extração dos resultados eleitorais partiu dos arquivos de votação por candidato, município e zona, distribuídos por unidade da federação, em formato de valores separados por ponto e vírgula e codificação latina. Para cada município, somaram-se os votos nominais válidos atribuídos a cada candidato no primeiro turno do cargo de presidente, restringindo-se a contagem aos votos classificados como válidos pela própria base, de modo a excluir registros anulados. O percentual de votos válidos de Bolsonaro resultou da razão entre os seus votos e o total de votos válidos do município. A conversão entre o código municipal do Tribunal Superior Eleitoral e o código do Instituto Brasileiro de Geografia e Estatística, necessária para o casamento com as bases de saúde e de população, foi feita por tabela de correspondência de domínio público, e a integração entre as bases de cobertura, voto, população e renda utilizou o código de sete dígitos do IBGE como chave.

A obtenção da cobertura de BCG e Tríplice Viral exigiu a reprodução do mecanismo de tabulação do DataSUS. O aplicativo de tabulação em linha gera as tabelas por meio de uma requisição cujos parâmetros descrevem a linha, a coluna e o conteúdo desejados; configurou-se a linha como município, a coluna como ano e o conteúdo como a cobertura de cada imunizante, e a requisição devolveu, para os 5.570 municípios, a série anual de 1994 a 2022. A definição de cobertura empregada pelo sistema, para esses imunizantes, é a razão entre doses aplicadas e população-alvo, expressa em pontos percentuais, e os valores recuperados reproduzem, inclusive, os artefatos de denominador discutidos adiante. Esse procedimento de extração, por ser integralmente automatizável, assegura que a base de cobertura possa ser regenerada a qualquer momento a partir da fonte oficial.

A reconstrução da cobertura de COVID-19 partiu dos microdados de vacinação, distribuídos por unidade da federação em arquivos comprimidos que registram, para cada combinação de data, município de residência, dose, imunizante, sexo e faixa etária, o número de aplicações e a população municipal de referência. Os vinte e sete arquivos estaduais foram processados sequencialmente, somando-se, por município de residência e até 31 de dezembro de 2022, as aplicações classificadas como segunda dose, que foram então divididas pela população municipal de 2021 para produzir a cobertura em pontos percentuais. A escolha da segunda dose corresponde ao esquema primário de duas doses; a definição alternativa, por esquema completo, é examinada como verificação de robustez. O recorte temporal até o fim de 2022 alinha a medida ao período eleitoral em exame, evitando contaminar a cobertura com doses tardias registradas em 2023 e 2024.

As técnicas estatísticas combinaram análise bivariada e multivariada, com os cuidados já indicados. As associações bivariadas foram medidas pela correlação de Pearson, nas versões não ponderada e ponderada pela população, e por comparações de médias entre grupos definidos pela mediana do percentual de votos, com o teste t de Welch, apropriado a variâncias desiguais. As análises multivariadas empregaram regressão linear com as variáveis explicativas padronizadas, de modo que os coeficientes exprimam a variação da cobertura por desvio-padrão da explicativa, com erros-padrão robustos à heterocedasticidade na forma do estimador de White e, alternativamente, com ponderação pela população por mínimos quadrados ponderados. O crescimento geométrico da cobertura da Tríplice Viral foi calculado como a taxa anual equivalente entre 2018 e 2022, com descarte dos municípios de cobertura inicial inferior a 50% e winsorização dos extremos nos percentis um e noventa e nove. Toda a cadeia foi implementada de modo a ser reexecutável a partir das fontes públicas.

### *2.1 Disponibilidade de dados e código*

O repositório do projeto com as instruções para acesso aos dados e códigos está disponível em https://github.com/tadeugomes/reanalise-cobertura-vacinal-voto.

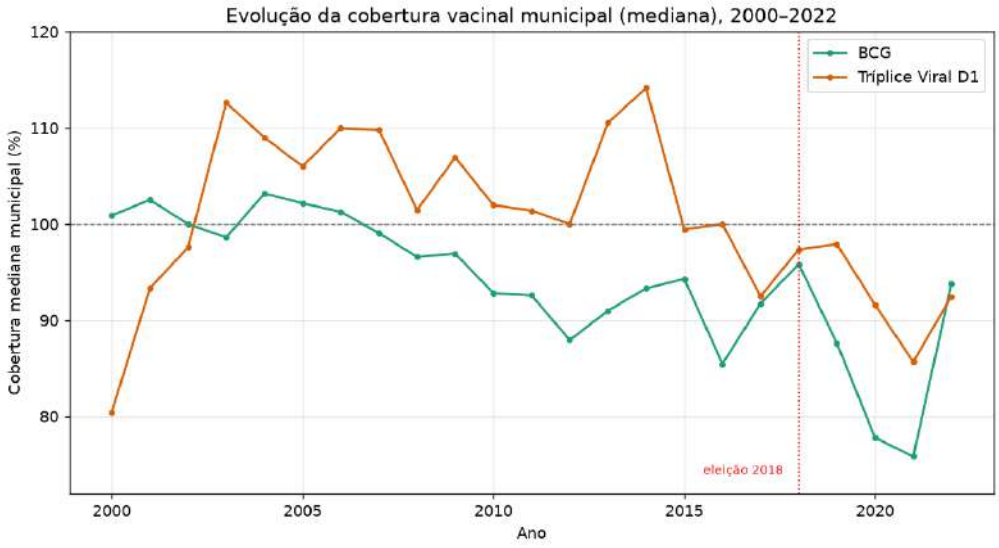
O código do projeto foi desenvolvido com auxílio de agentes de inteligência artificial generativa.

## *3. RESULTADOS*

A apresentação dos resultados parte do retrato descritivo da cobertura e de sua estrutura territorial, passa à mensuração correta do fator político e às associações entre cobertura, voto e renda, detém-se sobre a vacina contra a COVID-19, central à hipótese da politização, e encerra com o exame da qualidade dos dados e da estrutura espacial que condiciona toda a inferência.

O exame descritivo da série histórica situa o problema. A cobertura municipal de BCG e de Tríplice Viral, observada pela mediana entre os municípios, descreve uma trajetória de oscilação seguida de declínio que se manifesta a partir de meados da década de 2010, portanto antes da eleição de 2018, e que atinge tanto a vacina neonatal quanto a do primeiro ano de vida. Esse padrão, exibido na Figura 1, é incompatível com uma explicação centrada em um evento político posterior, e sugere que fatores estruturais do sistema de imunização antecedem e enquadram qualquer variação de natureza política. A distribuição regional da cobertura, sintetizada na Tabela 1, reforça a heterogeneidade territorial do fenômeno, com coberturas medianas que variam de modo expressivo entre as cinco regiões e entre os imunizantes, sem desenho que se alinhe de forma simples a qualquer clivagem política.

## Figura 1 - Evolução da cobertura vacinal municipal (mediana), 2000–2022




**Fonte:** elaboração própria a partir do SI-PNI/DataSUS. O declínio inicia-se antes da eleição de 2018.

## Tabela 1 - Cobertura vacinal mediana por região (em %)

| Região | BCG 2018 | BCG 2022 | Tríplice V. 2018 | Tríplice V. 2022 | COVID 2ª dose 2022 |
|---|---|---|---|---|---|
| Norte | 95,4 | 95,3 | 89,5 | 81,2 | 61,9 |
| Nordeste | 86,5 | 99,6 | 96,1 | 90,8 | 77,5 |
| Centro-Oeste | 99,0 | 93,1 | 96,2 | 93,9 | 77,1 |
| Sudeste | 99,6 | 91,3 | 101,0 | 92,0 | 84,5 |
| Sul | 97,9 | 89,9 | 96,9 | 98,4 | 85,7 |
| Brasil | 95,8 | 93,8 | 97,4 | 92,5 | 81,1 |

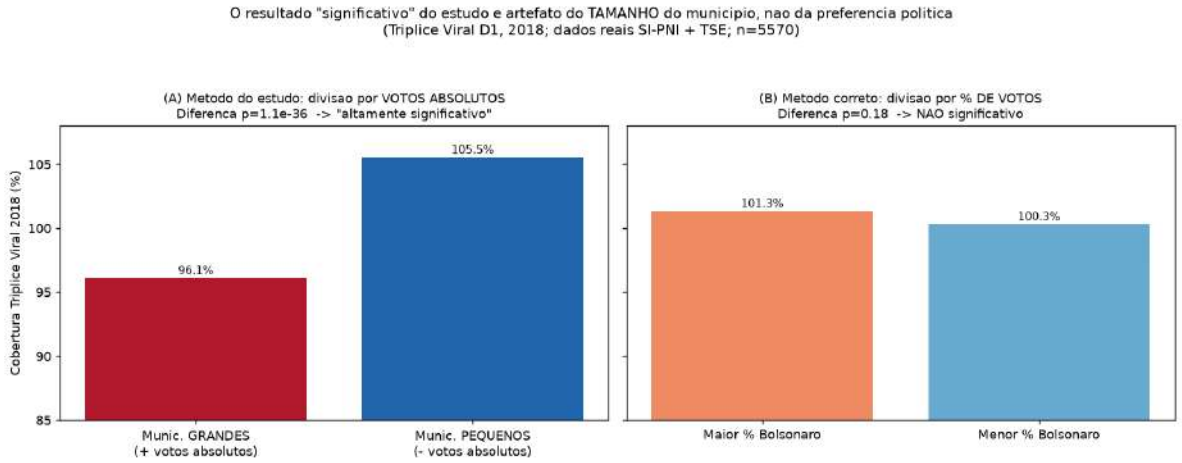
**Fonte:** elaboração própria (SI-PNI/DataSUS). Mediana municipal, em pontos percentuais.

A mensuração do fator político é o primeiro ponto que exige cuidado, pois dele depende todo o sentido da análise. Uma forma intuitiva, porém incorreta, de comparar municípios é dividi-los entre os de mais e os de menos votos em Bolsonaro pelo número absoluto de votos, procedimento que, aplicado a este conjunto, produz diferenças de cobertura significativas a níveis extremos, com p-valores da ordem de 10 elevado a menos 44 em 2018 e de 10 elevado a menos 81 em 2022. O número absoluto de votos, contudo, é determinado antes de tudo pelo tamanho do eleitorado, que varia entre os municípios brasileiros em ordens de grandeza, ao passo que o percentual de votos válidos varia numa faixa estreita. A mediana dos votos absolutos separa, por conseguinte, municípios populosos de municípios pequenos, e não municípios de maior e menor adesão política: o grupo rotulado como de mais votos é, na prática, o conjunto dos municípios grandes. A intensidade da preferência política só é corretamente captada pelo percentual de votos válidos, variável adotada em todas as estimativas que se seguem.

A diferença entre os dois procedimentos é decisiva e mensurável sobre os dados reais. Tomando a cobertura da Tríplice Viral em 2018, a partição pela mediana dos votos absolutos produz diferença entre os grupos com p-valor de 1,1 vezes 10 elevado a menos 36, com os municípios grandes exibindo cobertura de 96,1% contra 105,5% dos pequenos. A mesma comparação, feita

pela mediana do percentual de votos válidos, devolve médias de 101,3% e 100,3% e p-valor de 0,175, isto é, ausência de diferença estatisticamente significativa. O contraste é inequívoco: a divisão por votos absolutos mede porte populacional, e a significância desaparece quando se mede a preferência política pela variável adequada. Em 2022 o padrão se repete, com p-valor de 10 elevado a menos 95 na partição por votos absolutos.

**Figura 2 - Cobertura vacinal (Tríplice Viral D1, 2018) segundo o critério de divisão dos municípios**




**Fonte:** elaboração própria a partir de TSE e SI-PNI/DataSUS (n=5.570). À esquerda, divisão por votos absolutos, que mede o porte do município; à direita, divisão por percentual de votos, que mede a preferência.

**Tabela 2 - Resultados das análises bivariadas corrigidas, por imunizante e ano**

| Imunizante | Ano do voto | n | Cob. >100% (%) | Pearson (% votos) | Pearson (ponderado) | p (divisão por % votos) | p (divisão por votos absolutos) |
|---|---|---|---|---|---|---|---|
| BCG | 2018 | 5.570 | 42,6 | 0,133 | 0,106 | $1{,}3\times10^{-18}$ | 0,11 (n.s.) |
| BCG | 2022 | 5.570 | 41,4 | -0,083 | -0,071 | $9{,}3\times10^{-9}$ | $1{,}4\times10^{-5}$ |
| COVID-19 (2ª dose) | 2018 | 5.570 | 8,5 | 0,117 | 0,148 | $2{,}4\times10^{-27}$ | $3{,}6\times10^{-64}$ |
| COVID-19 (2ª dose) | 2022 | 5.570 | 8,5 | 0,080 | 0,047 | $5{,}5\times10^{-20}$ | $9{,}4\times10^{-66}$ |
| Tríplice Viral D1 | 2018 | 5.570 | 43,5 | 0,007 | -0,061 | 0,175 (n.s.) | $1{,}1\times10^{-36}$ |
| Tríplice Viral D1 | 2022 | 5.570 | 34,5 | 0,065 | 0,031 | $1{,}2\times10^{-6}$ | $1{,}0\times10^{-95}$ |

**Fonte:** elaboração própria a partir de TSE, SI-PNI/DataSUS e IBGE. n.s. = não significativo.

A simulação controlada reforça a interpretação. Construída de modo que a cobertura tivesse correlação nula com a preferência política, e dependesse apenas do porte da unidade, ela produziu, sob a partição por votos absolutos, p-valor da ordem de 10 elevado a menos 45, da mesma ordem extrema observada nos dados reais, ao passo que a partição por percentual de votos devolveu p-valor de 0,53, não significativo, como deveria ser na ausência de relação. Demonstra-se, assim, que a significância de tal partição não depende de qualquer relação entre voto e vacinação, mas apenas do confundimento entre votos absolutos e tamanho do município.

**Tabela 3 - Simulação contrafactual: significância sob ausência de relação política (n = 5.570 unidades sintéticas)**

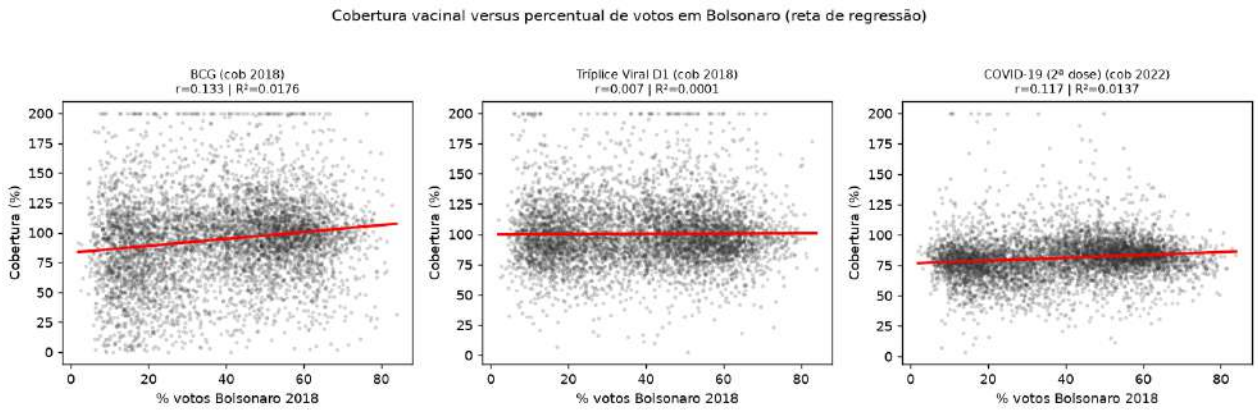
| Procedimento | Estatística | p-valor | Resultado |
|---|---|---|---|
| Correlação real cobertura × preferência | r = 0.025 | — | nula por construção |
| Divisão por votos absolutos (método do estudo) | t de Welch | 2.4e-45 | "altamente significativo" (artefato) |
| Divisão por % de votos (método correto) | t de Welch | 0.53 | não significativo (correto) |

**Fonte:** elaboração própria. Dados sintéticos com correlação política nula por construção.

Quando se emprega a variável adequada, o percentual de votos válidos, as associações revelam-se fracas e sem direção consistente. Para a BCG, vacina aplicada a recém-nascidos e historicamente de cobertura quase universal, a correlação de Pearson com a votação em Bolsonaro é de 0,13 em 2018 e de menos 0,08 em 2022, ou seja, troca de sinal entre as duas eleições, e a versão ponderada pela população atenua ambas. A inversão de sinal entre anos próximos é incompatível com um efeito político estável e sugere que se está diante de flutuações sem significado substantivo. A escolha da BCG como indicador, ademais, é questionável, pois se trata de imunizante neonatal, cuja adesão depende de decisão dos responsáveis no momento do nascimento, pouco sensível ao tipo de hesitação associada ao discurso político sobre a COVID-19; os resultados nulos são coerentes com essa inadequação.

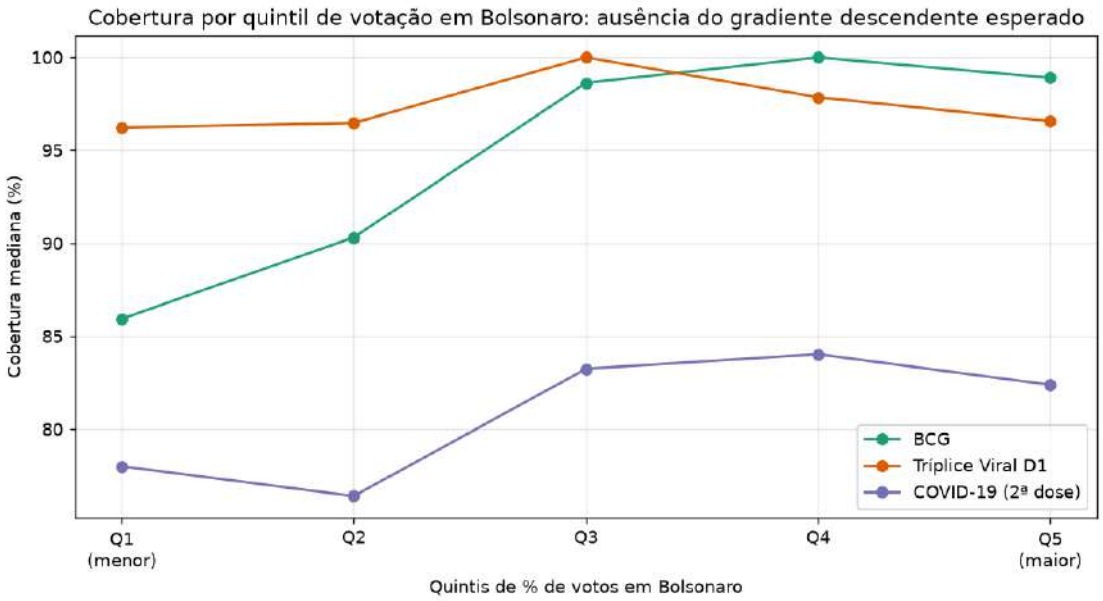
Para a Tríplice Viral, a correlação com o percentual de votos é praticamente nula em 2018, da ordem de 0,007, e de 0,06 em 2022. Nas regressões multivariadas que incluem o controle de renda, os coeficientes de determinação situam-se entre 0,0001 e 0,018, valores que indicam ausência de capacidade explicativa, e os coeficientes da votação invertem de sinal entre 2018 e 2022, repetindo a instabilidade já observada. Sob ponderação populacional, a maioria dos coeficientes perde significância. O quadro que emerge é o de associações que existem apenas no sentido estatístico estrito, sem magnitude nem estabilidade que autorizem interpretação substantiva.

**Figura 3 - Cobertura vacinal *versus* percentual de votos em Bolsonaro, por imunizante**




**Fonte:** elaboração própria (TSE e SI-PNI/DataSUS). Cada ponto é um município; reta de regressão em vermelho; eixo vertical truncado em 200%.

**Figura 4 - Cobertura por quintil de votação em Bolsonaro: ausência do gradiente descendente esperado**




**Fonte:** elaboração própria (TSE e SI-PNI/DataSUS). A hipótese previa cobertura menor nos quintis de maior votação; observa-se o contrário ou a estabilidade, descartando a relação dose-resposta esperada.

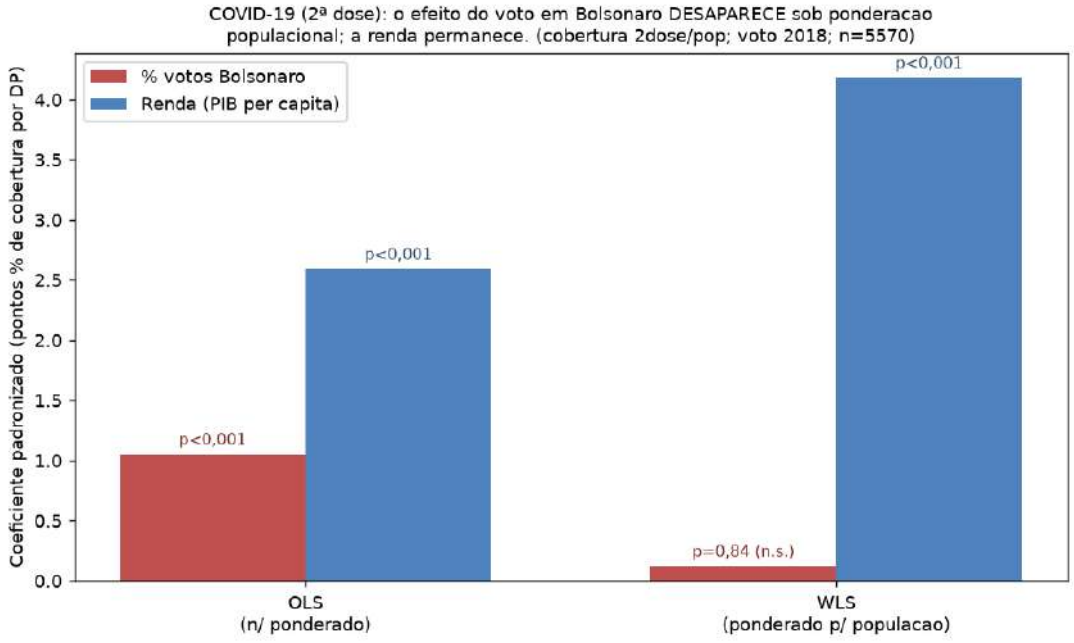
O recálculo do crescimento geométrico da cobertura da Tríplice Viral entre 2018 e 2022 ilustra o efeito das correções de robustez. A versão ingênua do indicador produz valores entre menos 100% e mais 145%, contaminados por municípios cuja cobertura inicial se aproximava de zero, situação em que a razão entre coberturas explode. Os diagnósticos da regressão correspondente exibem curtose superior a vinte e estatística de Jarque-Bera na casa das dezenas de milhares, sinais inequívocos de domínio por valores extremos. Depois de descartar os oitenta e cinco municípios com cobertura inicial abaixo de 50% e de winsorizar os extremos, a regressão devolve coeficiente de determinação de 0,001, o que equivale a ausência de relação. O coeficiente positivo da votação, que poderia parecer contraintuitivo, não sobrevive como achado interpretável: com poder explicativo nulo, não há relação a interpretar.

A análise da cobertura de COVID-19 com duas doses é a mais sensível à hipótese de politização da pandemia. Medida como segunda dose sobre a população e por município de residência, a cobertura apresenta correlação fraca e de sinal positivo com a votação em Bolsonaro: o coeficiente de Pearson com o voto de 2018 é de 0,12, e a partição pela mediana do percentual de votos mostra cobertura maior, e não menor, nos municípios de maior votação, com médias de 83,7% contra 78,5%. No nível municipal, portanto, não há sinal de menor adesão à vacinação onde Bolsonaro foi mais votado.

O sentido positivo dessa associação decorre, em boa medida, do mesmo confundimento de porte que vicia a comparação por votos absolutos. Os municípios pequenos, nos quais a votação em Bolsonaro é proporcionalmente maior, registram cobertura mais alta, ao passo que os grandes centros, de menor votação relativa, sofrem com os problemas de denominador e com a vacinação de residentes fora de seu município. O teste decisivo está na ponderação. Na regressão multivariada por mínimos quadrados ordinários, a votação em Bolsonaro aparece com coeficiente positivo e significativo, de 1,05, e a renda com coeficiente positivo maior, de 2,59, num modelo de baixo poder explicativo, com

coeficiente de determinação de 0,031. Quando a regressão é ponderada pela população municipal, o coeficiente da votação perde a significância, caindo a 0,12 com p-valor de 0,84, e apenas a renda permanece relevante, com coeficiente de 4,18, enquanto o coeficiente de determinação sobe para 0,114.

**Figura 5 - Coeficientes da regressão da cobertura de COVID-19 (2ª dose): o efeito do voto desaparece sob ponderação**




**Fonte:** elaboração própria a partir de TSE, SI-PNI e IBGE (voto de 2018; n=5.570). Coeficientes padronizados.

**Tabela 4 - Cobertura de COVID-19 (2ª dose): regressão multivariada, MQO vs. ponderada pela população (voto de 2018)**

| Modelo | $R^2$ | Coef. % votos Bolsonaro | p | Coef. renda (PIB p.c.) | p |
|---|---|---|---|---|---|
| MQO (erros-padrão robustos HC3) | 0,031 | 1,05 | <0,001 | 2,59 | <0,001 |
| MQ ponderados pela população | 0,114 | 0,12 | 0,84 (n.s.) | 4,18 | <0,001 |

**Fonte:** elaboração própria (TSE, SI-PNI e IBGE). Coeficientes padronizados; erros-padrão robustos à heterocedasticidade.

Em outras palavras, sob o controle pela renda medida como PIB *per capita*, a associação política desaparece quando os municípios são ponderados pela sua população, e o determinante que resta é socioeconômico. Essa leitura, como mostram as análises de robustez adiante, é sensível ao modo de medir a renda: com um vetor socioeconômico mais completo, o coeficiente da votação deixa de ser nulo e assume sinal negativo, ainda que subordinado aos fatores socioeconômicos e restrito à vacina contra a COVID-19. A própria mudança de sinal do coeficiente conforme a especificação, que vai do positivo na estimativa não ponderada ao negativo sob controles mais completos, é em si um alerta, pois um efeito tão dependente de escolhas técnicas não sustenta, isoladamente, conclusão substantiva.

A robustez dessa leitura à definição da cobertura foi verificada com uma especificação alternativa, que adota o esquema primário completo (a segunda dose dos esquemas de duas doses somada à dose única da vacina Janssen) em lugar de apenas a segunda dose. A cobertura mediana sobe de 81,1% para 83,5%, e a fração de municípios acima de 100% de 8,5% para 10,6%, mas o quadro qualitativo permanece idêntico, como mostra a Tabela 5. A correlação com a votação continua fraca e de sinal positivo, apenas ligeiramente maior; o coeficiente da votação na regressão segue significativo na especificação não ponderada e perde a significância quando os municípios são ponderados pela população, restando a renda como o determinante relevante. A conclusão central da análise para a COVID-19, portanto, não depende da definição adotada para a cobertura, o que afasta a hipótese de que o resultado seja um artefato da escolha entre segunda dose e esquema completo.

**Tabela 5 - Robustez da cobertura de COVID-19 à definição: 2ª dose vs. esquema completo (voto de 2018)**

| Definição da cobertura | Cob. mediana (%) | Cob. >100% (%) | Pearson (% votos) | Pearson ponderado | MQO: coef. % votos | Ponderado: coef. % votos | Ponderado: coef. renda |
|---|---|---|---|---|---|---|---|
| 2ª dose (definição principal) | 81,1 | 8,5 | 0,117 | 0,148 | +1,05 (p<0,001) | +0,12 (n.s.) | 4,18 |
| Esquema completo (2ª dose + Janssen) | 83,5 | 10,6 | 0,138 | 0,167 | +1,40 (p<0,001) | +0,37 (n.s.) | 4,33 |

**Fonte:** elaboração própria (microdados SI-PNI/RNDS, TSE e IBGE).
Coeficientes padronizados; n.s. = não significativo.

A análise torna visível, por fim, um problema de qualidade de dados que qualquer estudo dessas séries precisa enfrentar. Entre 34% e 44% dos municípios apresentam cobertura registrada superior a 100% em cada combinação de vacina e ano das séries do Programa Nacional de Imunizações, com máximos manifestamente espúrios, superiores a mil por cento e, em casos isolados, a treze mil por cento; para a COVID-19, a fração acima de 100% é de 8,5%, em razão do denominador adotado. Esses valores decorrem da divergência entre doses aplicadas e população-alvo estimada e, sem tratamento, injetam ruído nas correlações e dominam as transformações de razão, o que torna o seu relato e a replicação com a cobertura limitada a 100% parte necessária de qualquer análise honesta dessas bases.

O exame conjunto das partições por porte e por preferência, para todos os imunizantes e anos, consolida o diagnóstico. Para a BCG, a partição por percentual de votos devolve diferenças significativas pelo tamanho amostral, mas de magnitude pequena e sinal inconstante, enquanto a partição por votos absolutos sequer alcança a significância convencional em 2018 (p de 0,11); para a Tríplice Viral, ao contrário, é a partição por votos absolutos que produz p-valores extremos, de 10 elevado a menos 36 em 2018 e a menos 95 em 2022, ao passo que a partição por percentual ora não é significativa, ora o é com diferença pequena. A variação do padrão entre vacinas confirma que não há um efeito político coerente, e sim associações governadas por características estruturais.

As correlações ponderadas pela população oferecem leitura mais conservadora. Para a BCG, a correlação de Pearson de 0,13 em 2018 cai a 0,11 sob ponderação, e a de menos 0,08 em 2022 atenua-se a menos 0,07; para a Tríplice Viral, a correlação quase nula de 2018 torna-

se ligeiramente negativa, e a de 0,06 em 2022 cai a 0,03. Em nenhum caso a magnitude ultrapassa o limiar da associação fraca. A ponderação, ao dar aos grandes centros o peso de sua população, aproxima as estimativas de uma média populacional e desfaz o confundimento entre porte e preferência que invalida a comparação por votos absolutos.

As regressões multivariadas com controle de renda reforçam a baixa capacidade explicativa das variáveis políticas. Para a BCG, o coeficiente de determinação é de 0,018 com o voto de 2018 e cai a 0,007 com o de 2022, com inversão de sinal da votação, e sob ponderação ambos perdem significância ou magnitude. Para a Tríplice Viral, os coeficientes de determinação são praticamente nulos, de 0,0001 a 0,004, e os sinais oscilam entre especificações. A renda, medida pelo PIB *per capita*, ora é significativa, ora não, conforme a vacina e a ponderação. O conjunto é consistente com a ausência de relação substantiva entre preferência política e cobertura nas vacinas do calendário infantil.

**Tabela 6 - Coeficientes de determinação ($R^2$) das análises**

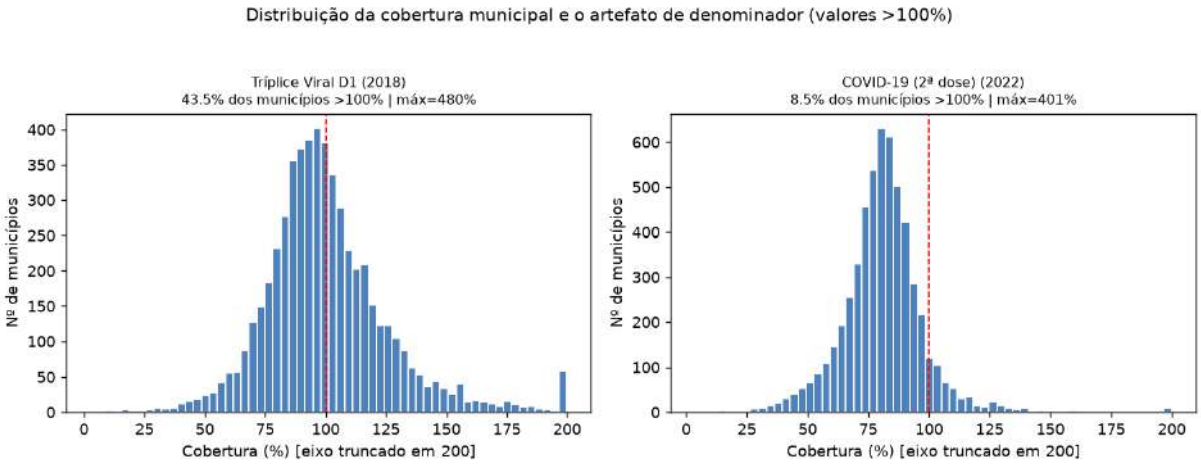
| Imunizante | Ano do voto | R2 bivariado | R2 multivar. (MQO) | R2 multivar. (ponderado) |
|---|---|---|---|---|
| BCG | 2018 | 0,0176 | 0,0177 | 0,0118 |
| BCG | 2022 | 0,0069 | 0,0070 | 0,0052 |
| COVID-19 (2ª dose) | 2018 | 0,0137 | 0,0307 | 0,1141 |
| COVID-19 (2ª dose) | 2022 | 0,0064 | 0,0271 | 0,0800 |
| Tríplice Viral D1 | 2018 | 0,0001 | 0,0001 | 0,0038 |
| Tríplice Viral D1 | 2022 | 0,0042 | 0,0043 | 0,0013 |
| Tríplice Viral (cresc. geom.) | 2018 → 22 | — | 0,0010 | 0,0002 |

Fonte: elaboração própria. $R^2$ bivariado = quadrado da correlação com o percentual de votos.

A vacina contra a COVID-19 destoa desse quadro de nulidade, mas não na direção que a hipótese da politização sugeria. A correlação positiva entre votação em Bolsonaro e cobertura, de 0,12 com o voto de 2018 e de 0,08 com o de 2022, e mais ainda o seu comportamento sob ponderação, indicam que a associação aparente é mediada pelo porte e pela renda dos municípios. O contraste entre a regressão ordinária, em que a votação é significativa, e a ponderada, em que deixa de ser, ao passo que a renda se fortalece, é o resultado mais nítido da análise no que toca à hipótese central. Ele indica que a leitura do efeito político é contingente à mensuração da renda: controlada a renda apenas pelo PIB, não resta efeito político detectável, ao passo que sob um vetor socioeconômico mais completo, examinado adiante, emerge associação negativa modesta e específica da COVID-19; em qualquer especificação, a variável socioeconômica é a que organiza a maior parte da variação da cobertura.

O exame da qualidade dos dados condiciona a leitura de todos os resultados anteriores. A cobertura superior a 100% em mais de um terço dos municípios das séries do Programa Nacional de Imunizações, com valores que chegam às dezenas de milhares por cento, significa que parte da variação da variável de resposta é ruído de registro, e não fenômeno epidemiológico. A replicação com a cobertura limitada a 100% não altera as conclusões qualitativas, o que indica robustez dos achados a esse tratamento, mas a própria necessidade de realizá-lo, rara na literatura aplicada, mede o cuidado que essas bases exigem.

**Figura 6 - Distribuição da cobertura municipal e o artefato de denominador (valores acima de 100%)**




**Fonte:** elaboração própria (SI-PNI/DataSUS). Eixo horizontal truncado em 200%; os máximos reais são muito superiores.

**Tabela 7 - Análise de sensibilidade: correlação com a cobertura bruta vs. limitada a 100%**

| Imunizante | Ano | Pearson (cob. bruta) | Pearson (cob. ≤ 100%) | Cob. >100% (%) |
|---|---|---|---|---|
| BCG | 2018 | 0,1326 | 0,2112 | 42,6 |
| BCG | 2022 | -0,0831 | -0,0182 | 41,4 |
| COVID-19 (2ª dose) | 2018 | 0,1170 | 0,1698 | 8,5 |
| COVID-19 (2ª dose) | 2022 | 0,0801 | 0,1206 | 8,5 |
| Tríplice Viral D1 | 2018 | 0,0074 | 0,0524 | 43,5 |
| Tríplice Viral D1 | 2022 | 0,0649 | 0,0317 | 34,5 |

**Fonte:** elaboração própria (TSE e SI-PNI/DataSUS).

## 4. DISCUSSÃO

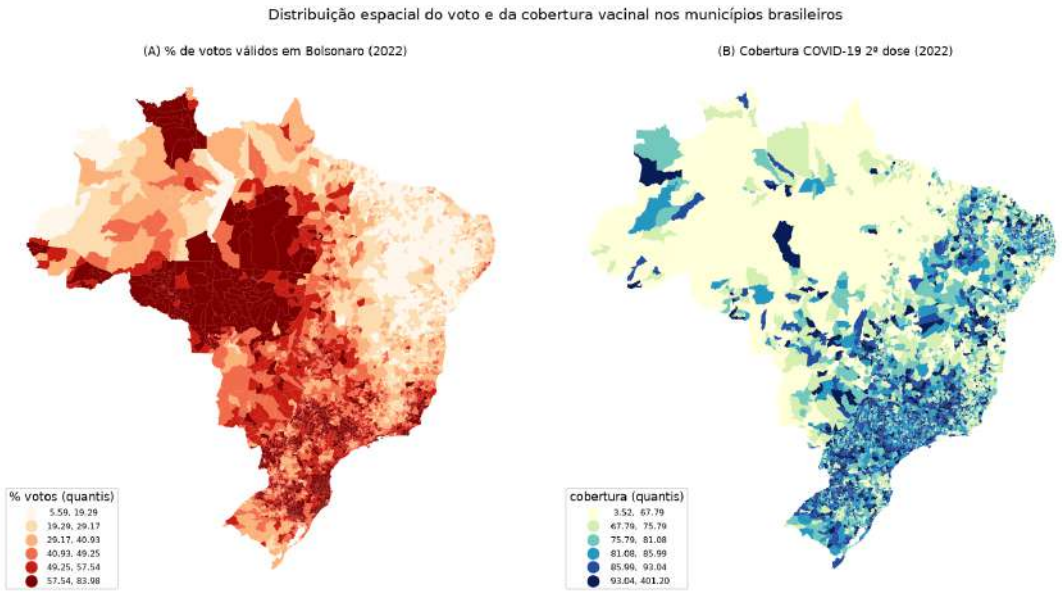
Os resultados convergem para uma conclusão: a cobertura vacinal municipal no Brasil é organizada primordialmente por fatores socioeconômicos e estruturais e a contribuição do comportamento político-eleitoral, quando medida de modo apropriado, é fraca, instável entre anos e vacinas, e sensível à ponderação populacional e ao tratamento dos dados. A renda municipal aparece de forma consistente como o determinante relevante e o declínio da cobertura, anterior a 2018 e fortemente estruturado no espaço, reforça o peso dos fatores estruturais. A esse achado soma-se uma lição de método: a aparência de uma relação política forte, recorrente no debate público, decorre em boa parte de uma mensuração inadequada da preferência, a divisão dos municípios pelo número absoluto de votos, que capta o porte populacional, e não a adesão política.

O caso da COVID-19 acrescenta uma lição sobre proveniência e definição de dados. A cobertura de segunda dose pode ser construída de diferentes maneiras (denominadores distintos, recortes temporais diversos, município de residência ou de aplicação), e essas escolhas afetam a magnitude e, em casos limítrofes, o sinal da associação. A medida aqui adotada, segunda dose sobre a população total por município de residência até o fim de 2022, é explícita e reprodutível. O resultado não deve ser lido como prova de maior vacinação onde Bolsonaro venceu, mas como evidência de que a relação entre voto e cobertura é refém de escolhas de mensuração que precisam ser transparentes, e a primazia que dela emerge, a da renda e das condições de acesso, aponta na direção que a literatura sobre determinantes da cobertura já indicava.

Convém situar esses achados ante o limite intransponível do desenho. Todas as análises operam sobre agregados municipais, e nenhuma das associações observadas, qualquer que seja o seu sinal ou magnitude, autoriza inferência sobre o comportamento de eleitores individuais. A hipótese formulada, de que o eleitorado de Bolsonaro seria mais aderente ao discurso contrário à vacinação, é uma proposição sobre pessoas, e a falácia ecológica adverte que correlações entre médias municipais podem inverter-se no nível individual (Robinson, 1950). Por isso, as associações aqui estimadas, qualquer que seja o seu sinal ou magnitude, devem ser lidas como propriedades de municípios, e não de eleitores, e nenhuma delas confirma ou refuta, no plano individual, a hipótese de partida; o que elas informam é sobre os determinantes territoriais da cobertura.

A estrutura espacial dos dados reforça essa advertência e acrescenta um problema estatístico próprio. Tanto o voto quanto a cobertura distribuem-se no território de modo fortemente agrupado, como mostra a Figura 7: os municípios de alta votação em Bolsonaro concentram-se no Centro-Sul e na fronteira agrícola, ao passo que a cobertura segue padrões regionais distintos. O índice I de Moran (Moran, 1950), na Tabela 8, confirma autocorrelação espacial intensa e altamente significativa, sobretudo para a votação, com índice em torno de 0,88, e os resíduos da regressão da COVID-19 também a exibem, com índice de 0,33. Sob essa violação do pressuposto de independência, mesmo os erros-padrão robustos subestimam a incerteza e os p-valores ficam artificialmente reduzidos, problema que as verificações de robustez adiante enfrentam diretamente.

## Figura 7 - Distribuição espacial do voto em Bolsonaro e da cobertura vacinal nos municípios




**Fonte:** elaboração própria a partir de TSE, SI-PNI/DataSUS e malha municipal do IBGE. Classes por quantis.

## Tabela 8 - Autocorrelação espacial (índice I de Moran) das variáveis e dos resíduos do modelo

| Variável | I de Moran | p-valor |
|---|---|---|
| % votos Bolsonaro (2018) | 0,883 | 0,001 |
| % votos Bolsonaro (2022) | 0,880 | 0,001 |
| Cobertura COVID-19 2ª dose (2022) | 0,351 | 0,001 |
| Cobertura BCG (2022) | 0,208 | 0,001 |
| Cobertura Tríplice Viral (2018) | 0,097 | 0,001 |
| Resíduos da regressão COVID-19 (MQO) | 0,334 | 0,001 |

**Fonte:** elaboração própria. Pesos espaciais por k-vizinhos mais próximos (k=6); significância por 999 permutações.

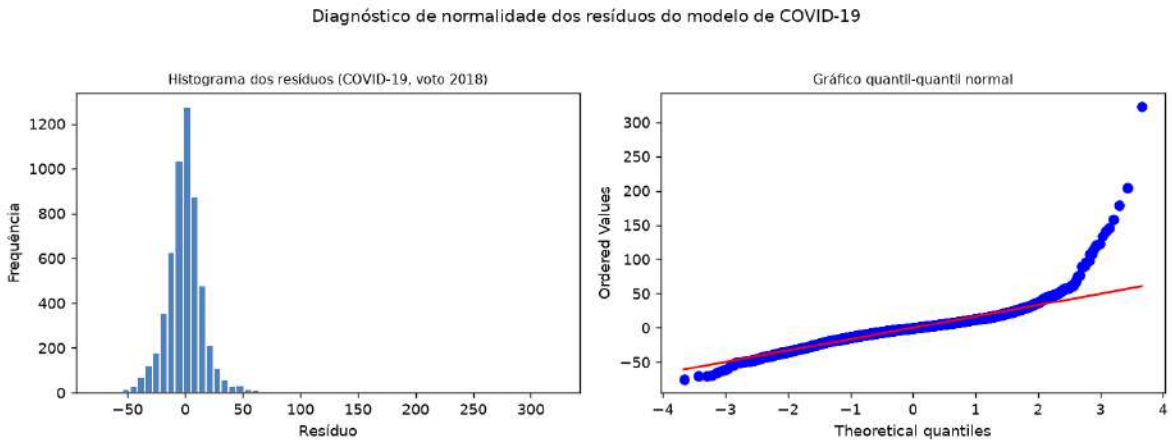
Os diagnósticos dos resíduos completam o quadro de cautela. A Tabela 9 reúne, para os seis modelos multivariados, os indicadores de ajuste e de normalidade, e a Figura 8 ilustra, para o modelo de COVID-19, a distribuição dos resíduos. Em todos os casos, os resíduos afastam-se da normalidade, com assimetria positiva entre 1,3 e 2,8 e curtose entre quinze e trinta e oito, e estatísticas de Jarque-Bera nas dezenas ou centenas de milhares. Esse afastamento justifica o uso de erros-padrão robustos e reforça que as conclusões devem repousar nas magnitudes dos efeitos, consistentemente baixas, e não na significância dos testes.

## Tabela 9 - Diagnóstico dos resíduos dos modelos multivariados

| Modelo (imunizante / ano do voto) | R2 | Assimetria | Curtose | Jarque-Bera | Durbin-Watson |
|---|---|---|---|---|---|
| BCG (voto 2018) | 0,0177 | 2,82 | 37,7 | 287504 | 1,89 |
| BCG (voto 2022) | 0,0070 | 1,34 | 14,7 | 33541 | 1,79 |
| Tríplice Viral (voto 2018) | 0,0001 | 1,98 | 16,2 | 44208 | 1,90 |
| Tríplice Viral (voto 2022) | 0,0043 | 2,07 | 20,7 | 76511 | 1,86 |
| COVID-19 2ª dose (voto 2018) | 0,0307 | 2,44 | 33,6 | 222335 | 1,53 |
| COVID-19 2ª dose (voto 2022) | 0,0271 | 2,37 | 33,0 | 213709 | 1,50 |

**Fonte:** elaboração própria. Sob normalidade dos resíduos, espera-se assimetria ≈ 0 e curtose ≈ 3.

## Figura 8 - Diagnóstico de normalidade dos resíduos do modelo de COVID-19 (histograma e gráfico quantil-quantil)




**Fonte:** elaboração própria. O afastamento da reta no gráfico quantil-quantil indica não normalidade.

As duas principais reservas que pesam sobre a leitura anterior, a limitação do controle socioeconômico a uma proxy única e a dependência espacial dos resíduos, motivaram duas verificações de robustez que, em conjunto, sustentam e qualificam a conclusão. A primeira substituiu o PIB *per capita*, proxy imperfeita por ser distorcida por municípios de grande produção e pequena população, por um vetor mais completo e diretamente medido, com a renda domiciliar *per capita*, a taxa de alfabetização e o grau de urbanização do Censo Demográfico de 2022 (IBGE, 2023) e a cobertura potencial de atenção primária da Estratégia Saúde da Família, do e-Gestor Atenção Básica para dezembro de 2022 (BRASIL, 2022). A Tabela 10 reúne quatro especificações encadeadas do modelo ponderado da cobertura de COVID-19. A simples troca do PIB pela renda domiciliar eleva o coeficiente de determinação de 0,114 para 0,230, e a inclusão de escolaridade, urbanização e atenção primária o leva a 0,288: o

vetor socioeconômico, melhor medido, dobra a variação explicada. Nele, renda, alfabetização e atenção primária aparecem com coeficientes positivos, altos e altamente significativos, e a renda não perde força quando a atenção primária entra, o que sugere que recursos materiais e capacidade de oferta do SUS são canais distintos e complementares da cobertura.

**Tabela 10 - Cobertura de COVID-19 (2ª dose) sob vetor de controles socioeconômicos ampliado (modelo ponderado, voto de 2018)**

| Especificação | Voto | Renda | Escolaridade | Urbanização | Atenção primária | $R^2$ |
|---|---|---|---|---|---|---|
| M1: renda como PIB per capita (original) | +0,12 n.s. | +4,18 *** | — | — | — | 0,114 |
| M2: renda domiciliar (Censo 2022) | −2,73 *** | +5,12 *** | — | — | — | 0,230 |
| M3: + escolaridade e urbanização | −3,80 *** | +4,70 *** | +2,53 ** | −0,82 * | — | 0,237 |
| M4: + atenção primária (ESF) | −4,24 *** | +4,79 *** | +4,55 *** | +1,07 ** | +4,94 *** | 0,288 |

**Fonte:** elaboração própria (TSE, SI-PNI, IBGE/Censo 2022 e e-Gestor AB/SISAB). Modelo ponderado pela população; coeficientes padronizados (pontos percentuais de cobertura por desvio-padrão), erros-padrão robustos (HC3). Em M1 a renda é o PIB per capita; em M2 a M4, a renda domiciliar per capita do Censo. *** p<0,001; ** p<0,01; * p<0,05; n.s. = não significativo.

O vetor ampliado modifica, contudo, a leitura do fator político na COVID-19. Sob controle apenas pelo PIB, o coeficiente da votação era nulo (0,12, sem significância); medida a renda de forma mais adequada, torna-se negativo e significativo, da ordem de quatro pontos percentuais de cobertura por desvio-padrão de votação, sinal compatível com menor adesão à vacinação contra a COVID-19 nos municípios de maior votação em Bolsonaro. O resultado não reabilita a tese de uma determinação política da cobertura, por três razões. A magnitude do coeficiente da votação é da mesma ordem da de cada fator socioeconômico isolado, enquanto o bloco socioeconômico, em conjunto, organiza a maior parte da variação explicada. A associação é específica da COVID-19 e não se reproduz nas vacinas do calendário, para as quais o coeficiente da votação permanece nulo em todas as especificações, o que afasta a leitura de recusa generalizada e remete o achado à politização singular da pandemia (Gramacho; Turgeon, 2021; Seara-Morais et al., 2023). E a própria sensibilidade do coeficiente ao modo de medir a renda, que o desloca de nulo a negativo, confirma a tese metodológica central do trabalho, a de que a leitura política da cobertura é frágil e contingente às escolhas de mensuração.

A segunda verificação enfrenta a dependência espacial dos resíduos sinalizada pelo índice de Moran. Como erros espacialmente autocorrelacionados subestimam os erros-padrão e inflam a significância, reestimaram-se os modelos com erros-padrão de Conley, que corrigem a autocorrelação até um raio definido (Conley, 1999), e com modelos espaciais de erro (SEM) e de defasagem (SAR), sobre a mesma malha de seis vizinhos mais próximos (Anselin, 1988). A Tabela 11 reúne os resultados para a COVID-19. A dependência espacial é forte, com parâmetro próximo de 0,6 no termo de erro e na defasagem, e a sua incorporação altera de modo informativo a leitura da significância: o coeficiente positivo da votação na regressão ordinária,

aparentemente robusto sob erros de White, perde a significância sob erros de Conley a partir de duzentos quilômetros, o que indica que parte dela decorria da autocorrelação espacial, e não de um efeito político. O coeficiente da renda permanece elevado e altamente significativo sob todas as correções, e a associação negativa do voto no modelo ampliado também sobrevive aos modelos espaciais. A primazia dos determinantes socioeconômicos não é, portanto, artefato de inferência inflada pela estrutura espacial dos dados.

**Tabela 11 - Robustez da cobertura de COVID-19 (2ª dose) à autocorrelação espacial (voto de 2018)**

| Especificação | Coef. voto (p) | Coef. renda (p) | λ ou ρ |
|---|---|---|---|
| MQO, erros-padrão de White (HC3) | +1,05 (<0,001) | +2,59 (<0,001) | — |
| MQO, erros-padrão de Conley (200 km) | +1,05 (0,12 n.s.) | +2,59 (<0,001) | — |
| MQ ponderados, White (HC3) | +0,12 (0,84 n.s.) | +4,18 (<0,001) | — |
| MQ ponderados, Conley (200 km) | +0,12 (0,86 n.s.) | +4,18 (<0,001) | — |
| Modelo de erro espacial (SEM) | −0,93 (0,02) | +2,34 (<0,001) | λ = 0,59 |
| Modelo de defasagem espacial (SAR) | −0,26 (0,25 n.s.) | +2,17 (<0,001) | ρ = 0,58 |

**Fonte:** elaboração própria (TSE, SI-PNI e IBGE). Coeficientes padronizados; renda medida pelo PIB per capita, como na Tabela 4. Erros-padrão de Conley com núcleo de Bartlett; SEM e SAR estimados por máxima verossimilhança sobre malha de seis vizinhos mais próximos. λ e ρ são os parâmetros de dependência espacial.

Há também implicações para a prática de pesquisa quantitativa em políticas públicas. A reprodutibilidade é o instrumento que distingue determinantes substantivos de artefatos, pois é ao tornar explícitas a escolha da variável de preferência, a definição do denominador e a opção pela ponderação que se evita confundir tamanho com adesão. O depósito do código e dos dados deveria ser condição de aceitação de qualquer estudo quantitativo.

As limitações deste estudo devem ser explicitadas. A cobertura de COVID-19 foi construída com denominador (população total) e recorte temporal (até 2022) próprios, escolhas que condicionam as magnitudes e devem ser consideradas na leitura dos resultados a ela relativos. O controle socioeconômico da análise principal, o PIB *per capita*, é *proxy* imperfeita das condições de renda, razão pela qual as verificações de robustez o substituíram por um vetor mais completo e diretamente medido, com renda domiciliar, escolaridade, urbanização e cobertura de atenção primária. As vacinas analisadas pela via do Programa Nacional de Imunizações restringiram-se à BCG e à Tríplice Viral, por serem as de cobertura municipal diretamente disponível na base; a extensão a outros imunizantes do calendário poderia enriquecer o quadro, mas não há razão para esperar que altere a conclusão central sobre a primazia dos fatores socioeconômicos.

A multiplicidade de decisões que separam o dado bruto do resultado final merece reflexão. A escolha da variável de partição, do denominador, do recorte temporal, da ponderação e do tratamento dos valores extremos abre um leque de caminhos analíticos, cada qual com resultado possivelmente distinto, o que a literatura descreveu como o jardim dos caminhos que

se bifurcam (Gelman; Loken, 2014). Sem pré-registro ou reprodutibilidade, pode-se convergir, mesmo sem má intenção, para a especificação que confirma a hipótese. No caso em exame, a especificação que a sustentava era também a estatisticamente inadequada; a robustez vem da coerência entre múltiplas especificações corretas, não da significância de uma só.

A inserção dos achados no debate sobre a politização da saúde durante a pandemia exige cautela. A existência, em diversos países, de correlação entre orientação política e comportamento sanitário não implica que essa correlação se manifeste, no Brasil, no nível municipal e na direção suposta, nem que sobreviva ao controle por condições socioeconômicas. Para a cobertura de COVID-19 medida de forma reprodutível, o que se observa no nível municipal é dominado pela renda e pelo porte. Esse resultado não nega a politização da pandemia como fenômeno político e discursivo, mas adverte contra a sua tradução automática em correlações ecológicas interpretadas como evidência de comportamento individual. Um estudo ecológico recente (Seara-Morais et al., 2023) reportou, para as eleições de 2018 e 2022, associação negativa entre a votação em Bolsonaro e um índice de vacinação contra a COVID-19, persistente após ajuste sociodemográfico. A divergência em relação ao presente achado decorre sobretudo de diferenças de mensuração e de método, em particular o uso de um índice composto, a ausência de ponderação pela população e a escolha do denominador.

Convém reconhecer os limites da própria estratégia de ponderação. Ponderar pela população aproxima as estimativas de uma média populacional e reduz o peso desproporcional das pequenas unidades, mas não resolve a inferência ecológica, pois continua a operar sobre agregados, e não é isenta de escolhas, já que ponderar pela população total, pela população-alvo ou pelo eleitorado conduziria a estimativas distintas. A análise optou pela população total, o denominador mais transparente, e afere a robustez pela persistência ao longo de especificações defensáveis, não pela significância de uma escolha conveniente.

## CONSIDERAÇÕES FINAIS

Este artigo investigou os determinantes da cobertura vacinal nos municípios brasileiros e estabeleceu, sobre dados reais e reprodutíveis, a primazia das condições socioeconômicas sobre o comportamento eleitoral. Três resultados a sustentam. A cobertura é organizada sobretudo pela renda, determinante que sobrevive à ponderação, ao controle por vetor socioeconômico ampliado e à correção espacial, ao passo que a associação com a votação é fraca e instável entre anos e vacinas. O declínio antecede a eleição de 2018 e é fortemente estruturado no espaço, o que aponta para fatores estruturais, e não para um evento político. E a aparência de um efeito político forte depende de como se mede a preferência: medi-la pelo número absoluto de votos confunde-a com o porte e produz significância espúria, que se dissolve sob o percentual de votos e a ponderação.

Convém precisar o sentido dessa primazia, para que não se confunda cobertura com hesitação. A associação positiva entre renda e cobertura, mais nítida na COVID-19 e ausente de forma robusta nas vacinas de rotina, indica que o déficit se concentra nos municípios mais pobres, padrão compatível com desigualdades de acesso (oferta de serviços, logística e capacidade da atenção primária), e não com recusa deliberada. A hesitação vacinal, definida como o adiamento ou a recusa da vacinação apesar da disponibilidade do serviço (Macdonald, 2015),

é fenômeno de demanda, distinto do acesso, que a literatura não associa de modo simples à pobreza, sendo por vezes mais visível em grupos de maior renda e escolaridade. Como o estudo mede cobertura, e não hesitação, a leitura mais defensável de seus resultados é a de desigualdade de acesso, e não a de que a população mais pobre hesite mais.

A primazia dos determinantes socioeconômicos tem implicações diretas para a ação pública. Estratégias de recuperação da cobertura centradas apenas no enfrentamento do discurso anti-vacina dificilmente alcançarão o cerne do problema sem o fortalecimento da atenção primária, a qualificação dos sistemas de registro e a redução das desigualdades de acesso. Para pesquisas futuras, recomenda-se a adoção de unidades menos heterogêneas, como as regiões de saúde, a incorporação de indicadores socioeconômicos e de acesso como controles, o uso sistemático de ponderação populacional e de erros-padrão robustos, e o depósito público de código e dados, sem os quais a análise da relação entre comportamento político e adesão vacinal continuará a confundir tamanho com preferência.

A contribuição do artigo é dupla. No plano substantivo, oferece evidência reprodutível de que a cobertura é organizada sobretudo pela renda e pelas condições de acesso, ao passo que a associação com a votação é fraca e instável, nula nas vacinas de rotina e, na COVID-19, sensível à mensuração da renda, assumindo sinal negativo modesto sob controles mais completos. No plano metodológico, documenta como a escolha da variável de preferência é decisiva, pois a partição pelo número absoluto de votos gera a aparência de uma relação forte que os dados, lidos corretamente, não comportam.

Conclui-se, portanto, que a pergunta original, sobre se houve menor adesão à vacinação nos municípios de maior votação em Bolsonaro, não encontra sustentação ampla e robusta nos dados tratados de modo adequado: nas vacinas de rotina não há efeito político detectável e, mesmo na COVID-19, único caso em que emerge associação negativa, ela é modesta, subordinada aos determinantes socioeconômicos e contingente à mensuração da renda. O que os dados sustentam é mais modesto e mais sólido: a cobertura vacinal municipal é organizada sobretudo por fatores socioeconômicos e estruturais, e a leitura política dessa cobertura, feita sobre agregados e sem os devidos cuidados, confunde o tamanho dos municípios com a preferência de seus eleitores.

## REFERÊNCIAS


ALLCOTT, H. et al. Polarization and public health: partisan differences in social distancing during the coronavirus pandemic. **Journal of Public Economics**, v. 191, 104254, 2020.

ANSELIN, L. **Spatial econometrics: methods and models.** Dordrecht: Kluwer Academic Publishers, 1988.

ARROYO, L. H. et al. Áreas com queda da cobertura vacinal para BCG, poliomielite e tríplice viral no Brasil (2006-2016): mapas da heterogeneidade regional. **Cadernos de Saúde Pública**, v. 36, n. 4, e00015619, 2020.

BRASIL. Ministério da Saúde. **e-Gestor Atenção Básica: cobertura da Atenção Básica e da Estratégia Saúde da Família por município**. Brasília: Ministério da Saúde, 2022.

BRASIL. Ministério da Saúde. **Sistema de Informações do Programa Nacional de Imunizações (SI-PNI): coberturas vacinais**. Brasília: DATASUS, 2023. Disponível em: http://tabnet.datasus.gov.br. Acesso em: 23 jun. 2026.

COHEN, J. The earth is round (p < .05). **American Psychologist**, v. 49, n. 12, p. 997-1003, 1994.

CONLEY, T. G. GMM estimation with cross sectional dependence. **Journal of Econometrics**, v. 92, n. 1, p. 1-45, 1999.

DOMINGUES, C. M. A. S. et al. 46 anos do Programa Nacional de Imunizações: uma história repleta de conquistas e desafios a serem superados. **Cadernos de Saúde Pública**, v. 36, supl. 2, e00222919, 2020.

GELMAN, A.; LOKEN, E. The statistical crisis in science. **American Scientist**, v. 102, n. 6, p. 460-465, 2014.

GRAMACHO, W. G.; TURGEON, M. When politics collides with public health: COVID-19 vaccine country of origin and vaccination acceptance in Brazil. **Vaccine**, v. 39, n. 19, p. 2608-2612, 2021.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Censo Demográfico 2022: rendimento domiciliar, alfabetização e situação dos domicílios**. Rio de Janeiro: IBGE, 2023.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Estimativas da população residente e Produto Interno Bruto dos Municípios**. Rio de Janeiro: IBGE, 2023. Disponível em: https://servicodados.ibge.gov.br. Acesso em: 23 jun. 2026.

KING, G. **A solution to the ecological inference problem: reconstructing individual behavior from aggregate data**. Princeton: Princeton University Press, 1997.

MACDONALD, N. E. Vaccine hesitancy: definition, scope and determinants. **Vaccine**, v. 33, n. 34, p. 4161-4164, 2015.

MORAN, P. A. P. Notes on continuous stochastic phenomena. **Biometrika**, v. 37, n. 1-2, p. 17-23, 1950.

OPENSHAW, S. **The modifiable areal unit problem**. Norwich: Geo Books, 1984. (Concepts and Techniques in Modern Geography, 38).

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE (OPAS). Brasil perde o status de país livre do sarampo. Washington, D.C.: OPAS, 2019.

ROBINSON, W. S. Ecological correlations and the behavior of individuals. **American Sociological Review**, v. 15, n. 3, p. 351-357, 1950.

SATO, A. P. S. Programa Nacional de Imunização: Sistema Informatizado como opção a novos desafios. **Revista de Saúde Pública**, v. 49, n. 39, 2015.

SATO, A. P. S. Qual a importância da hesitação vacinal na queda das coberturas vacinais no Brasil? **Revista de Saúde Pública**, v. 52, 96, 2018.

SEARA-MORAIS, G. J.; AVELINO-SILVA, T. J.; COUTO, M.; AVELINO-SILVA, V. I. The pervasive association between political ideology and COVID-19 vaccine uptake in Brazil: an ecologic study. **BMC Public Health**, v. 23, 1597, 2023.

TRIBUNAL SUPERIOR ELEITORAL (TSE). **Repositório de dados eleitorais: votação por candidato, município e zona, 2018 e 2022**. Brasília: TSE, 2024. Disponível em: https://dadosabertos.tse.jus.br. Acesso em: 23 jun. 2026.

WASSERSTEIN, R. L.; LAZAR, N. A. The ASA statement on p-values: context, process, and purpose. **The American Statistician**, v. 70, n. 2, p. 129-133, 2016.

WHITE, H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. **Econometrica**, v. 48, n. 4, p. 817-838, 1980.

WORLD HEALTH ORGANIZATION (WHO). **Ten threats to global health in 2019**. Genebra: WHO, 2019.



# EIXO 03

## *Educação, Inteligência Territorial e Inclusão*

# Governança de Dados para Permanência Estudantil: uma abordagem aplicada com ciência de dados

*Fabiana Ferreira do Nascimento*
*Profª MSc. Thaís Salzer Procópio*

## Resumo

A evasão acadêmica no ensino superior público brasileiro representa um desafio para a eficiência do gasto público e para a democratização do acesso à educação. Este estudo investiga como a qualidade e a integração de dados institucionais influenciam a identificação precoce de estudantes em risco de evasão, utilizando uma abordagem aplicada de Ciência de Dados e Governança de Dados. Foram analisados microdados acadêmicos de uma universidade pública brasileira em três cenários progressivos de integração de dados: acadêmicos, demográficos e socioeconômico. Três algoritmos supervisionados foram aplicados em comparação: Regressão Logística, Random Forest e XGBoost. A avaliação adotou segmentação temporal e métricas adequadas a bases desbalanceadas. Os resultados demonstram que dados acadêmicos básicos já permitem modelos com desempenho elevado (ROC-AUC de 0,855 e Recall de 0,863), viabilizando sistemas de alerta precoce. A incorporação de variáveis demográficas, socioeconômicas e assistenciais ampliou a capacidade de identificação de estudantes vulneráveis, evidenciando ganhos de equidade analítica. O diagnóstico institucional indicou maturidade intermediária em governança de dados, com limitações de integração e padronização. Conclui-se que a governança de dados constitui elemento estratégico para ampliar a efetividade de iniciativas de permanência estudantil baseadas em inteligência analítica.

**Palavras-chave:** Governança de Dados; Permanência Estudantil; Evasão Acadêmica; Aprendizado de Máquina; Learning Analytics; IFES.

## INTRODUÇÃO

A evasão acadêmica na educação superior pública brasileira produz impactos sociais, econômicos e institucionais ao reduzir a efetividade das políticas de acesso e permanência, ampliar desigualdades e comprometer o retorno social dos investimentos públicos em educação.

Embora a Política Nacional de Assistência Estudantil (PNAES) constitua o principal instrumento federal de promoção da permanência (Brasil, 2024b), indicadores recentes mostram que a evasão permanece elevada (INEP, 2025), evidenciando a necessidade de mecanismos capazes de identificar precocemente estudantes em risco.

É nesse contexto que as abordagens de *Learning Analytics* e de aprendizado de máquina supervisionado têm ganhado espaço como instrumentos de apoio à permanência estudantil. Modelos preditivos de evasão permitem identificar padrões de risco antes que o afastamento se consolide, orientando intervenções preventivas e o uso mais eficiente dos recursos de assistência (Romero; Ventura, 2010; Oliveira et al., 2021).

Contudo, a utilidade prática desses modelos depende da qualidade e da integridade dos dados institucionais que os alimentam. A fragmentação de dados entre sistemas heterogêneos compromete o desempenho e a equidade dos modelos preditivos. Nesse sentido, a governança de dados deixa de ser apenas um requisito técnico e passa a constituir condição estruturante da capacidade analítica institucional (Jiang; Zhu; Pan, 2024; Pahune et al., 2025).

No setor público brasileiro, o Tribunal de Contas da União (TCU) identificou que a maioria das instituições públicas ainda se encontra em estágios iniciais no uso estratégico de dados, com lacunas em políticas claras, estruturas definidas e processos sistemáticos de qualidade da informação (Tribunal de Contas da União, 2026). Esse diagnóstico é particulamente relevante para as Instituições Federais de Ensino Superior (IFES), caracterizadas pela descentralização administrativa, pela coexistência de sistemas informacionais heterogêneos e pela dependência de integração manual entre fontes de dados acadêmicas, socioeconômicas e administrativas.

Este estudo investiga como a qualidade e a integração progressiva de dados institucionais influenciam a identificação precoce de estudantes em risco de evasão, analisando três cenários de disponibilidade informacional e comparando modelos de aprendizado de máquina supervisionado.

Além desta introdução, o artigo apresenta o referencial teórico, os procedimentos metodológicos, os resultados e, por fim, as conclusões.

## 1. REFERENCIAL TEÓRICO

Esta seção apresenta os fundamentos teóricos que sustentam a análise proposta, abordando a governança de dados no contexto das IFES, os desafios relacionados à qualidade e integração de dados acadêmicos e o uso de modelos de aprendizado de máquina para predição da evasão.

### 1.1 Governança de Dados nas IFES

A governança de dados refere-se ao conjunto de estruturas decisórias, papéis, processos e mecanismos que asseguram a gestão dos dados como ativos estratégicos, com qualidade, segurança e alinhamento aos objetivos organizacionais (Fraga, 2022).

Modelos de referência consolidados, como o DAMA - Data Management Body of Knowledge (Dama-DMBOK), posicionam a governança como eixo transversal da gestão de dados, orientando domínios como qualidade, integração, arquitetura, metadados e segurança (Dama International, 2017). Nessa perspectiva, a governança estrutura as condições para o uso estratégico da informação, especialmente em contextos corporativos complexos e intensivos em dados.

A revisão sistemática da literatura conduzida por Al-Ruithe et al. (2019) oferece uma taxonomia analítica útil para a compreensão das dimensões que estruturam a governança de dados: função, estrutura, aspectos organizacionais, técnicos, ambientais e de monitoramento e mensuração. Os autores identificam que iniciativas de governança falham quando conduzidas de forma isolada pela área de tecnologia da informação, sem suporte executivo amplo, operando em silos por sistema e com responsabilidades ambiguamente definidas. Esse padrão é particularmente relevante nas IFES, caracterizadas por forte descentralização administrativa e pela coexistência de sistemas informacionais heterogêneos, o que intensifica riscos de inconsistência e dificulta o uso estratégico dos dados.

No setor público brasileiro, a governança de dados tem sido incorporada à agenda de transformação digital do Estado. A Estratégia Federal de Governo Digital 2024-2027 e a Infraestrutura Nacional de Dados estabelecem diretrizes para o tratamento de dados como ativos estratégicos, com ênfase em interoperabilidade, padronização semântica, segurança da informação e transparência (Brasil, 2024a).

Contudo, evidências empíricas indicam um baixo nível de desenvolvimento nesse campo. O Tribunal de Contas da União (TCU), por meio do Acórdão nº 457/2026 - Plenário, identificou que a maioria das instituições públicas avaliadas ainda se encontra em estágios iniciais no uso estratégico de dados, com lacunas notáveis em políticas claras, estruturas definidas e processos sistemáticos para garantir a qualidade e segurança das informações (Tribunal de Contas da União, 2026). Essa carência fundamental limita a capacidade dos órgãos de utilizar seus dados de forma eficiente e estratégica, impactando negativamente a tomada de decisões e a efetividade das políticas públicas, além de aumentar os riscos de inconsistências e dificultar a integração de informações. A relevância da governança de dados na administração pública é sublinhada pelas conclusões do TCU, que apontam que a ausência de uma governança robusta impede a plena utilização do potencial analítico dos vastos acervos de dados governamentais.

Como resposta a esse cenário, o Guia de Implementação do Programa de Governança de Dados (PGDADOS) propõe diretrizes operacionais para institucionalização da governança no setor público, incluindo definição de papéis, estruturação de políticas, gestão da qualidade e mecanismos de catalogação e integração de dados (Dweck; Mascarenhas, 2025). O modelo organiza a evolução da governança em níveis de maturidade alinhados à Infraestrutura Nacional de Dados.

A literatura reforça que os principais entraves à governança de dados são de natureza organizacional. Visão parcial dos dados, ausência de padrões semânticos e baixa coordenação comprometem a consistência e a confiabilidade das informações (Al-Ruithe et al., 2019; Jiang; Zhu; Pan, 2024). Nesse contexto, a governança de dados deve ser compreendida como um processo evolutivo, cuja robustez se reflete na capacidade institucional de produzir, integrar e utilizar dados de forma consistente e orientada à tomada de decisão. Tal desafio é particularmente relevante nas IFES, marcadas pela multiplicidade de sistemas e unidades administrativas, o que intensifica a complexidade da gestão da informação.

## 1.2 Integração de Bases de Gestão Educacional

A evasão no ensino superior é compreendida como um fenômeno multifatorial, resultante da interação entre características individuais dos estudantes e condições administrativas, como suporte acadêmico, engajamento e integração ao ambiente universitário (Tinto, 2012). Abordagens contemporâneas de *Learning Analytics* reforçam esse entendimento ao explorar dados educacionais para identificação de padrões de risco, evidenciando o papel ativo da instituição na promoção da permanência (Siemens; Baker, 2012).

A mensuração da evasão, entretanto, não é trivial: depende da definição operacional adotada e da disponibilidade de dados consistentes e integrados sobre a trajetória acadêmica. Oliveira et al. (2021) destacam que a ausência de padronização conceitual e de interoperabilidade entre sistemas compromete a construção de indicadores confiáveis e a comparabilidade entre estudos, limitando o potencial analítico das abordagens baseadas em dados educacionais.

Segundo o DAMA International (2017), dimensões como completude, consistência, acurácia e atualidade são essenciais para garantir a confiabilidade das análises. Contudo, em ambientes educacionais, os dados frequentemente encontram-se distribuídos em sistemas acadêmicos, administrativos e socioeconômicos, sem padrões semânticos compartilhados. Essa fragmentação eleva o esforço de preparação analítica e introduz riscos de inconsistência que afetam diretamente a qualidade das variáveis disponíveis para modelagem.

A integração dessas fontes, quando adequadamente gerida por práticas de governança, expande significativamente a capacidade explicativa e preditiva dos modelos, proporcionando uma visão longitudinal mais abrangente da trajetória estudantil (Oliveira et al., 2021). A qualidade dos dados empregados é um fator determinante para o desempenho e a confiabilidade de modelos preditivos, conforme reforçado por Pahune et al. (2025), o que posiciona a governança como elemento estruturante da analítica no contexto educacional.

É nesse enquadramento que o presente estudo estrutura seus experimentos em três cenários progressivos de integração de dados: acadêmico isolado, com integração parcial de dados de ingresso e perfil demográfico, e com integração plena de dados socioeconômicos e de assistência estudantil. Essa estruturação permite investigar empiricamente em que medida a ampliação progressiva da base informacional influencia o desempenho preditivo e a utilidade dos modelos para a tomada de decisão organizacional.

## 1.3 Modelos de Aprendizado de Máquina para Predição de Evasão

A aplicação de técnicas de aprendizado de máquina na predição da evasão no ensino superior tem sido amplamente explorada no campo de *Learning Analytics* e *Edu-cational Data Mining* (Romero; Ventura, 2010; Siemens; Baker, 2012; Oliveira et al., 2021). Modelos supervisionados permitem identificar padrões associados ao risco de evasão a partir da trajetória acadêmica dos estudantes, contribuindo para a construção de sistemas de alerta precoce e para o suporte à tomada de decisão (Parra et al., 2024; Dake; Buabeng-Andoh, 2022).

A literatura converge para o uso de abordagens complementares que equilibram desempenho preditivo e interpretabilidade, aspecto particularmente relevante em contextos institucionais nos quais os modelos subsidiam políticas públicas e ações de assistência estudantil. Entre os métodos mais recorrentes destacam-se modelos baseados em *ensemble*, especialmente o Random Forest, amplamente reportado como uma das abordagens mais

eficazes para predição de evasão estudantil. Estudos empíricos indicam que o algo-ritmo apresenta desempenho consistente ao capturar padrões complexos e não lineares ao longo da trajetória acadêmica (Solis et al., 2018).

Complementarmente, métodos de *gradient boosting*, como o XGBoost, têm demonstrado desempenho superior em tarefas de classificação com dados tabulares, sendo frequentemente empregados em aplicações educacionais recentes (Parra et al., 2024; Dake; Buabeng-Andoh, 2022). Esses algoritmos são particularmente eficazes na modelagem de interações complexas entre variáveis e na otimização do desempenho preditivo.

Por outro lado, a literatura enfatiza a importância de manter modelos estatísticos interpretáveis como referência, especialmente em cenários que demandam transparência, auditabilidade e explicabilidade das decisões automatizadas. Nesse contexto, a Regressão Logística permanece como abordagem de base amplamente adotada em estudos de evasão estudantil, permitindo a interpretação direta da influência das variáveis sobre a probabilidade de evasão (Oliveira et al., 2021).

A adoção conjunta desses modelos permite analisar o equilíbrio entre desempenho, interpretabilidade e aplicabilidade, contribuindo para a discussão sobre o uso responsável de modelos preditivos no contexto da permanência estudantil.

## 1.4 Avaliação de Modelos em Cenários Desbalanceados

A predição de evasão no ensino superior configura-se como um problema típico de classificação desbalanceada, no qual a proporção de estudantes que abandonam é significativamente inferior à daqueles que permanecem. Nesse âmbito, métricas tradicionais como acurácia tornam-se inadequadas, pois podem indicar bom desempenho mesmo quando o modelo apresenta baixa capacidade de identificar a classe de maior interesse analítico.

A literatura recomenda, portanto, o uso de métricas mais sensíveis ao desempenho sobre a classe minoritária, como precisão, recall e F1-score (Venkatesan et al., 2024). A precisão indica a proporção de predições positivas corretas, enquanto o recall mede a capacidade do modelo de identificar efetivamente os casos de evasão. O F1-score sintetiza essas duas dimensões, sendo especialmente útil quando há necessidade de equilibrar detecção de risco e redução de falsos positivos.

Além disso, a análise da curva Precision-Recall e da curva ROC permite avaliar o desempenho dos modelos ao longo de diferentes limiares de decisão. Nessas abordagens, os modelos produzem probabilidades de evasão, que precisam ser convertidas em classificações binárias por meio da definição de um limiar (*threshold*).

A escolha do limiar assume papel central em aplicações institucionais. Limiares mais baixos tendem a aumentar o *recall*, ampliando a identificação de estudantes em risco e favorecendo estratégias de intervenção preventiva. Por outro lado, limiares mais elevados priorizam a precisão, reduzindo o volume de intervenções potencialmente desnecessárias.

Assim, a definição do *threshold* deve ser compreendida como uma decisão sociotécnica, que envolve não apenas critérios estatísticos, mas também a capacidade de resposta, a disponibilidade de recursos e os objetivos das políticas de permanência estudantil.

Dessa forma, a avaliação dos modelos neste estudo considera múltiplas métricas e diferentes limiares de decisão, buscando aproximar a análise preditiva das condições reais de uso dos modelos no contexto da gestão universitária.

Nesse contexto, Venkatesan et al. (2024) ressaltam que a definição do limiar não deve ser tratada como otimização puramente estatística, mas como escolha orientada pelos objetivos institucionais e pela capacidade operacional de resposta, argumento que orienta diretamente a análise desenvolvida na Seção 4.6.

## 2. METODOLOGIA

Este estudo adota uma abordagem integrada com componente interpretativo, conduzido como uma pesquisa aplicada baseada em estudo de caso, estruturada a partir do desenvolvimento de um pipeline de ciência de dados para investigar a predição de evasão no ensino superior sob diferentes níveis de disponibilidade e enriquecimento de dados. O desenho metodológico foi estruturado para refletir condições realistas de uso de modelos preditivos na gestão universitária, incorporando decisões relacionadas à preparação dos dados, definição de cenários informacionais, segmentação temporal e avaliação de desempenho.

### 2.1 Delineamento da pesquisa e contexto

O recorte empírico concentra-se na Universidade Federal da Paraíba (UFPB), considerando os cursos de graduação presencial. A escolha do contexto justifica-se pela complexidade dos sistemas institucionais e pela relevância das políticas de assistência estudantil no apoio à permanência discente.

A base utilizada é composta por microdados acadêmicos de estudantes ingressantes entre os períodos letivos de 2018.1 e 2025.2, organizados em registros semestrais que representam a trajetória acadêmica ao longo do tempo. Assim, cada registro corresponde a um estudante em um determinado semestre de referência.

A variável alvo foi definida de forma binária, indicando a ocorrência de evasão em períodos subsequentes. Nessa perspectiva, a evasão é operacionalmente definida como a ausência de renovação de matrícula por dois semestres consecutivos, excluindo-se trancamentos oficialmente registrados e transferências institucionais. Essa abordagem permite a construção de modelos preditivos voltados à identificação precoce de estudantes em risco. O componente qualitativo tem caráter complementar e interpretativo, sendo baseado em análise documental e em relatos institucionais não estruturados. Esses relatos foram obtidos a partir de interações profissionais da autora, incluindo reuniões técnicas, comunicações institucionais e registros administrativos.

Seu uso não visa à produção de evidência generalizável, mas à contextualização dos achados quantitativos, especialmente no que se refere a limitações organizacionais, práticas institucionais e desafios relacionados à governança de dados.

Os relatos foram anonimizados e utilizados em conformidade com princípios éticos de pesquisa e com a Lei Geral de Proteção de Dados Pessoais - LGPD (Lei nº 13.709/2018) (Brasil, 2018), garantindo confidencialidade e uso exclusivo para fins acadêmicos.

## 2.2 Integração e preparação das bases de dados

As bases institucionais consideradas na pesquisa estão dispostas em três grupos: informações do sistema acadêmico, incluindo histórico escolar, desempenho e situação de matrícula; registros socioeconômicos do cadastro da assistência estudantil; e dados socioeconômicos sobre concessão de auxílios institucionais e origem escolar. A integração das bases foi realizada por meio de identificadores institucionais pseudonimizados, permitindo a vinculação dos dados sem exposição de informações pessoais.

Foram realizadas etapas de limpeza e preparação visando garantir consistência analítica e adequação ao processo de modelagem. As principais etapas incluíram:

- remoção de registros inconsistentes e tratamento de valores ausentes;
- padronização de variáveis categóricas e numéricas;
- criação de variáveis derivadas relacionadas ao histórico acadêmico;
- exclusão dos semestres mais recentes, para os quais ainda não há confirmação do desfecho de evasão.

Adicionalmente, foram criadas variáveis temporais e indicadores agregados capazes de representar o histórico acadêmico acumulado dos estudantes. Após as etapas de preparação e engenharia de atributos, foi obtido um conjunto final de variáveis representando dimensões temporais, acadêmicas, administrativas e sociodemográficas da trajetória estudantil. A Tabela 1 apresenta a descrição das variáveis utilizadas na modelagem.

**Tabela 1: Descrição das variáveis utilizadas na modelagem preditiva**

| Variável | Definição | Tipo |
|---|---|---|
| **Variável Alvo** | | |
| **evasao_t2** | **Indica evasão formal em até dois semestres à frente** | **Binária** |
| **Variáveis temporais e de trajetória acadêmica** | | |
| semestre_referencia | Semestre de observação do registro | Inteiro |
| semestre_ingresso | Semestre de ingresso no curso | Inteiro |
| ano_ingresso | Ano de ingresso | Inteiro |
| periodo_ingresso | Período letivo de ingresso | Categórica |
| tempo_no_curso | Tempo acumulado desde o ingresso (semestres) | Numérica |
| semestres_com_atividade | Nº de semestres com matrícula ativa | Numérica |
| semestre_sem_matricula | Nº de semestres sem matrícula | Numérica |
| densidade_matricula | Proporção de semestres com matrícula ativa | Numérica |

| Variável | Definição | Tipo |
|---|---|---|
| | **Variável Alvo** | |
| **evasao_t2** | **Indica evasão formal em até dois semestres à frente** | **Binária** |
| | **Desempenho acadêmico** | |
| cra | Coeficiente de rendimento acadêmico | Numérica |
| taxa_aprovacao_semestre | Proporção de disciplinas aprovadas no semestre | Numérica |
| taxa_aprovacao_acumulada | Proporção acumulada de aprovações | Numérica |
| reprovacoes_consecutivas | Sequência de reprovações | Numérica |
| reprovadas | Nº de reprovações no semestre | Numérica |
| queda_desempenho | Indicador de redução recente de desempenho | Numérica |
| indice_instabilidade | Índice sintético de irregularidade acadêmica | Numérica |
| taxa_atividade | Proporção de semestres com atividades acadêmicas | Numérica |
| taxa_sem_matricula | Proporção de semestres sem matrícula | Numérica |
| | **Histórico administrativo e acadêmico** | |
| qtd_trancamentos | Nº de trancamentos de matrícula | Numérica |
| qtd_cancelamentos | Nº de cancelamentos de disciplinas | Numérica |
| periodo_pandemia | Indicador de semestre durante a pandemia | Binária |
| Assistência estudantil | | |
| semestres_com_bolsa | Nº de semestres com bolsa ativa | Numérica |
| bolsas_concedidas | Quantidade total de bolsas concedidas | Numérica |
| bolsas_indeferidas | Solicitações de bolsa indeferidas | Numérica |
| bolsas_suspensas | Nº de bolsas suspensas | Numérica |
| tipos_bolsa | Nº de modalidades de bolsa recebidas | Numérica |

| Variável | Definição | Tipo |
|---|---|---|
| **tentou_bolsa** | **Indicador de solicitação de bolsa** | **Binária** |
| | **Perfil sociodemográfico e educacional** | |
| sexo | Sexo do estudante | Categórica |
| idade_ingresso | Idade no ingresso | Numérica |
| id_raca | Raça/cor declarada | Categórica |
| id_tipo_rede_ensino | Origem escolar (rede pública/privada) | Categórica |
| id_modalidade_educacao | Modalidade do curso | Categórica |
| id_area_curso | Área de conhecimento do curso | Categórica |
| id_grau_academico | Grau acadêmico (bacharelado/licenciatura/ tecnólogo) | Categórica |
| regiao_naturalidade | Região de origem do estudante | Categórica |
| grupo_forma_ingresso | Forma de ingresso | Categórica |
| cota_macro | Categoria de ingresso e ação afirmativa | Categórica |

## 2.3 Segmentação temporal de treino e teste

Considerando a natureza longitudinal do problema, adotou-se uma estratégia de segmentação temporal dos dados. Diferentemente de divisões aleatórias tradicionais, essa abordagem preserva a ordem cronológica dos registros, evitando vazamento de informação entre os conjuntos de treino e teste.

Os modelos foram treinados com dados de períodos anteriores e avaliados em períodos posteriores, simulando o uso prospectivo dos modelos para identificação de estudantes em risco. Essa estratégia aproxima o experimento das condições reais de aplicação

## 2.4 Construção dos cenários de disponibilidade de dados

Com base na literatura de governança e *learning analytics*, foram definidos três cenários progressivos de disponibilidade informacional:

- **Cenário 1 (C1)** — Dados acadêmicos básicos: variáveis tradicionalmente disponíveis em sistemas acadêmicos, como desempenho e matrícula.
- **Cenário 2 (C2)** — Dados ampliados: inclusão de variáveis adicionais relacionadas ao ingresso e de perfil demográfico básico.
- **Cenário 3 (C3)** — Dados enriquecidos: incorporação de variáveis socioeconômicas e contextuais.

Essa estrutura permite avaliar empiricamente o impacto da ampliação da base informacional sobre o desempenho preditivo dos modelos.

## 2.5 Modelagem preditiva e Métricas de avaliação

A modelagem foi conduzida como um problema de classificação binária supervisionada, tendo como variável resposta a evasão (*evasao_t2*). Para isso, foram adotados três algoritmos de aprendizado de máquina, selecionados com base em sua recorrência e desempenho documentado na literatura aplicada à predição de evasão:

- A **Regressão Logística** compõe o conjunto como *baseline* interpretável visto que seu desempenho competitivo e sua auditabilidade são requisitos relevantes para adoção institucional responsável (Pahune et al., 2025);
- O **Random Forest** é o algoritmo mais frequentemente recomendado para esse problema, por sua capacidade de capturar relações não lineares ao longo da trajetória acadêmica, robustez a dados ruidosos e suporte nativo ao desbalanceamento de classes (Solis et al., 2018; Romero; Ventura, 2010; Oliveira et al., 2021);
- O **XGBoost** foi incluído por seu desempenho superior em experimentos preliminares e por sua adoção crescente em problemas de classificação desbalanceada em contextos educacionais (Parra et al., 2024; Dake; Buabeng-Andoh, 2022).

Os algoritmos foram implementados utilizando a biblioteca SCIKIT-LEARN (Pedregosa et al., 2011), especificamente LogisticRegression, RandomForestClassifier e XGBClassifier, com validação cruzada estratificada para preservar o desbalanceamento da classe minoritária (evasão ≈ 5%) e posteriormente avaliados no conjunto de teste temporal[1].

1 A biblioteca SCIKIT-LEARN está disponível em <https://scikit-learn.org/>. Último acesso: 05 abr. 2026.

A avaliação dos modelos considerou métricas adequadas a problemas de natureza desbalanceada, com ênfase no F1-score, por equilibrar precisão e recall na classe de evasão. Adicionalmente, são utilizadas curvas Precision-Recall e AUC-ROC para análise do desempenho sob diferentes limiares de decisão, permitindo avaliar o comportamento dos modelos sob distintas estratégias de intervenção.

## 2.6 Integração analítica

A integração entre os componentes qualitativo e quantitativo ocorre de forma articulada ao longo da análise, relacionando os resultados obtidos nos modelos preditivos com os desafios institucionais associados à governança dos dados.

Essa abordagem permite compreender o desempenho dos modelos em conjunto com as condições organizacionais e informacionais que influenciam sua construção e aplicação, evidenciando como limitações estruturais dos dados impactam diretamente o uso analítico na tomada de decisão.

O tratamento dos dados observou os princípios estabelecidos pela LGPD, com adoção de medidas de pseudonimização, minimização e controle de acesso. As análises foram conduzidas em ambiente controlado, sem compartilhamento de dados individualizados, sendo as informações utilizadas exclusivamente para fins acadêmicos.

## 2.7 Uso de Ferramentas de Inteligência Artificial

Ferramentas de Inteligência Artificial Generativa foram utilizadas como apoio à redação científica, revisão textual, organização do artigo e desenvolvimento de *scripts* em Python. Seu uso restringiu-se ao suporte técnico e à melhoria da qualidade da apresentação do trabalho.

As definições metodológicas, análises, interpretação dos resultados, elaboração das conclusões e validação do código permaneceram sob responsabilidade da autora. Os dados, códigos e resultados disponibilizados nos repositórios públicos correspondem às versões efetivamente utilizadas na pesquisa.

**Tabela 2: Modelo de maturidade em governança de dados (adaptado de DAMA-DMBOK e PGDados)**

| Nível | Descrição |
|---|---|
| 1. Inexistente | Ausência de práticas formais de governança de dados |
| 2. Inicial | Processos informais, não padronizados e dependentes de indivíduos |
| 3. Repetível | Existência de práticas parcialmente documentadas e replicáveis |
| 4. Definido | Processos formalizados, com papéis e responsabilidades definidos |
| 5. Gerenciado | Monitoramento sistemático da qualidade e uso dos dados |
| 6. Otimizado | Governança integrada à estratégia institucional e orientada a dados |

# 3. RESULTADOS E DISCUSSÃO

## 3.1 Diagnóstico institucional de governança de dados

No âmbito deste estudo, foi realizado um diagnóstico exploratório da maturidade em governança de dados na Universidade Federal da Paraíba (UFPB), conduzido a partir de interações com partes interessadas institucionais e aplicação adaptada de diretrizes do DAMA-DMBOK (DAMA International, 2017) e do Guia de Implementação do Programa de Governança de Dados (PGDados) (Dweck; Mascarenhas, 2025).

A análise indica que a instituição se encontra em um nível intermediário de maturidade (nível 2 – Repetível), com avanços pontuais em direção ao nível 3 (Definido), conforme apresentado na Tabela 2. Observam-se práticas parcialmente estruturadas, como integração entre sistemas acadêmicos e administrativos, definição de papéis relacionados à gestão de dados e iniciativas de transparência ativa, como o Sistema de Inteligência e Governança de Dados Institucionais (SIGOV) (Universidade Federal da Paraíba, 2025).

Apesar desses avanços, não há uma estrutura formal de governança de dados institucionalizada. Persistem limitações relacionadas à padronização de processos, integração sistêmica e gestão de metadados, com impactos diretos na consistência e confiabilidade das informações.

As evidências levantadas incluem inconsistências entre sistemas estruturantes; divergências semânticas na definição de variáveis; duplicidade de registros decorrente de replicação em diferentes bases; e ausência de padronização na coleta de dados por unidades descentralizadas (Tabela 3). Também foi identificada dependência de procedimentos manuais para integração com bases externas, indicando limitações de interoperabilidade.

**Tabela 3: Comparação entre referenciais de governança e prática institucional observada**

| Dimensão | Referencial (DAMA/PGDados) | Situação observada |
|---|---|---|
| Qualidade de dados | Monitoramento contínuo com métricas e processos de melhoria | Métricas em bases específicas, porém com coleta descentralizada e heterogeneidade entre unidades |
| Metadados | Catálogo institucional padronizado e integrado | Iniciativas pontuais de glossários; ausência de catálogo corporativo unificado |
| Integração | Arquitetura orientada à interoperabilidade interna e externa | Sistemas estruturantes integrados, com replicação em bases distintas e integração externa manual |
| Papéis e responsabilidades | Estrutura formal com *data stewards* e *accountability* | Papéis técnicos e legais existentes (LGPD, DBA, curadoria), porém sem estrutura formal de governança instituída |
| Uso analítico | Dados como ativo estratégico na tomada de decisão | Uso crescente para transparência, porém com elevado esforço de integração para respostas analíticas |

A literatura indica que os principais entraves à plena institucionalização da governança de dados são de natureza organizacional, manifestando-se tipicamente na operação fragmentada por sistema e na ausência de padrões semânticos compartilhados entre unidades (Al-Ruithe et al., 2019). O padrão observado na instituição investigada — embora existam papéis definidos e iniciativas em curso, a integração sistêmica e a gestão de metadados ainda operam de forma descentralizada — reflete um estágio de maturidade compatível com o encontrado em organizações públicas de complexidade similar (Tribunal de Contas da União, 2026), e delimita o contexto informacional no qual os experimentos deste estudo foram conduzidos.

As fragilidades identificadas afetam diretamente a produção de indicadores institucionais e ampliam o esforço de consolidação de bases analíticas, tornando as etapas de preparação mais complexas e intensivas. A qualidade dos dados, portanto, não se apresenta como atributo técnico isolado, mas como resultado das estruturas de produção, integração e gestão da informação prevalecentes na organização.

## *3.2 Preparação dos dados e desafios analíticos*

A etapa de preparação dos dados constituiu um dos componentes mais críticos deste estudo, materializando, na prática, os impactos das limitações de governança identificadas no diagnóstico institucional.

A construção da base analítica exigiu a integração de múltiplas fontes institucionais (acadêmicas, socioeconômicas e administrativas), com definição de regras explícitas de junção, uma vez que inconsistências nos identificadores e divergências estruturais entre sistemas resultavam em duplicidades quando não tratadas adequadamente. Foram também identificados problemas semânticos, como diferentes interpretações para variáveis equivalentes entre sistemas distintos, exigindo padronização conceitual antes da consolidação, evidência direta da ausência de catálogo institucional de metadados.

Uma decisão metodológica relevante concerne ao tratamento dos valores ausentes nas variáveis reprovações consecutivas e reprovadas: optou-se por substituí-los por zero, e não pela exclusão dos registros. Essa escolha fundamenta-se em regra semântica institucional documentada de que a ausência de registro nessas variáveis indica que o estudante não apresentou reprovações, e não que o dado está faltante. Esse caso exemplifica como a interpretação correta de nulos depende de conhecimento de domínio, não de critérios puramente técnicos, competência que pressupõe maturidade em metadados.

Outro achado relevante refere-se ao tratamento de valores ausentes em id-cota-ingresso, com aproximadamente 40% de nulos. A investigação da semântica institucional revelou que esses registros correspondiam majoritariamente à modalidade de ampla concorrência, simplesmente não codificada no sistema. Assim, os nulos foram mapeados para a categoria ampla concorrência, em vez de tratados como categoria desconhecida. Trata-se de ausência de metadado, não de dado faltante, distinção fundamental que somente a governança de metadados permite realizar de forma sistemática. Essa evidência é particularmente relevante para gestores institucionais: variáveis com elevada proporção de nulos em sistemas estruturantes devem ser investigadas semanticamente antes de qualquer tratamento técnico.

No que se refere à consistência temporal, foram excluídos os dois semestres mais recentes do conjunto de dados (semestre_referência = 20251 e 20252), pois, como a variável alvo considera evasão nos dois semestres subsequentes (evasao_t2), esses registros ainda não

possuem desfecho observável. A manutenção desses períodos introduziria censura temporal sistêmica, com taxas de evasão artificialmente próximas de zero. Esse tipo de ajuste, aparentemente técnico, é condicionado pela qualidade da documentação e do ciclo de atualização dos dados institucionais, outro indicador de maturidade em governança.

Adicionalmente, a análise temporal revelou descontinuidade nas taxas de evasão entre os semestres 2019.1 e 2021.2, correspondentes ao período da pandemia de COVID-19. Durante esse intervalo, a instituição adotou regimes acadêmicos especiais que suspenderam temporariamente os procedimentos regulares de evasão. Para evitar que o modelo aprendesse esse padrão como comportamento estrutural, foi criada a variável binária período-pandemia, incorporando explicitamente o contexto institucional ao pipeline de modelagem. Essa decisão reflete que evasão não é um fenômeno puramente individual — é também condicionada por políticas institucionais e contextuais (Tinto, 2012).

Por fim, a construção de variáveis derivadas como índice_instabilidade, queda_desempenho, taxa_atividade e tentou_bolsa é, em si mesma, um indicador do custo da baixa maturidade em governança: tais indicadores de risco não existem como atributos nativos nos sistemas institucionais, precisando ser reconstruídos manualmente a cada projeto analítico. A institucionalização dessas métricas como variáveis computadas e versionadas em plataformas de governança de dados, como o SIGOV, reduziria substancialmente o esforço de preparação em estudos futuros e viabilizaria o monitoramento contínuo da permanência estudantil.

## 3.3 Resultados da modelagem preditiva por cenário

### 3.3.1 Configuração experimental

O experimento foi conduzido sobre um conjunto de 284.511 registros semestrais, referentes a trajetória de estudantes ingressantes entre 2018.2 e 2025.2. Após as etapas de preparação e exclusão por critério temporal, adotou-se divisão temporal entre treino e teste: os dez semestres de 2018.2 a 2023.1 compõem o conjunto de treino (192.168 observações; taxa de evasão = 4,97%), enquanto os três semestres de 2023.2 a 2024.2 formam o conjunto de teste (92.343 observações; taxa de evasão = 5,10%). A estabilidade da taxa de evasão entre treino e teste confirma a adequação do corte temporal adotado.

Os modelos foram avaliados por três métricas complementares: ROC-AUC (capacidade discriminativa geral), Recall (sensibilidade à classe de evasão) e F1-score (equilíbrio entre precisão e *recall*). O desbalanceamento da variável alvo (aproximadamente 5% de evasão) foi tratado explicitamente em todos os algoritmos.

### 3.3.2 Desempenho por cenário e algoritmo

A Tabela 4 apresenta os resultados dos nove experimentos (três cenários × três algoritmos), comparando o desempenho na validação cruzada (CV) e no conjunto de teste temporal *holdout*.

A comparação dos valores de AUC entre validação cruzada e teste temporal indica ausência de sobreajuste significativo para Random Forest e Regressão Logística, cujos resultados permanecem próximos nos dois conjuntos de avaliação.

Para o XGBoost, observa-se uma discrepância mais pronunciada entre CV e teste nos cenários C2 e C3 (CV ≈ 0,72; teste ≈ 0,83). Esse comportamento pode estar associado ao uso do *StratifiedKFold* no conjunto de treino, que tende a misturar períodos temporalmente distantes, incluindo semestres do regime pandêmico, produzindo folds de validação com distribuição de evasão atípica.

O desempenho superior no *holdout* temporal, composto por semestres pós-pandêmicos com comportamento mais estável, sugere boa capacidade de generalização do modelo fora do período de regime especial.

A Figura 1 ilustra a comparação do ROC-AUC entre cenários e algoritmos, tanto na validação cruzada quanto no teste temporal.

### 3.3.3 Análise do trade-off Recall × Precisão

O comportamento das métricas de Recall e Precisão é estrutural no problema de evasão com alto desbalanceamento: todos os modelos apresentam Recall elevado (0,78–0,88) e Precisão baixa (0,09–0,15), consequência da prevalência reduzida da classe positiva (≈ 5%) e do uso de pesos de classe para favorecer a identificação de evasões. Esse padrão é esperado e desejável no contexto de alerta precoce, no qual identificar estudantes em risco tem prioridade sobre evitar falsos positivos (Venkatesan et al., 2024).

A Figura 2 apresenta o *scatter* de Recall × Precisão no conjunto de teste, por cenário e algoritmo. O Random Forest concentra o melhor equilíbrio relativo, enquanto a Regressão Logística tende a ampliar o Recall ao custo de maior queda na Precisão. O F1-score, síntese das duas métricas, situa-se entre 0,17 e 0,25 em todos os experimentos, refletindo o desafio intrínseco do problema.

## 3.4 Influência da Governança de Dados no Desempenho dos Modelos

A comparação entre os três cenários experimentais é o principal meio de demonstrar como a disponibilidade e a qualidade dos dados influenciam a capacidade analítica institucional.

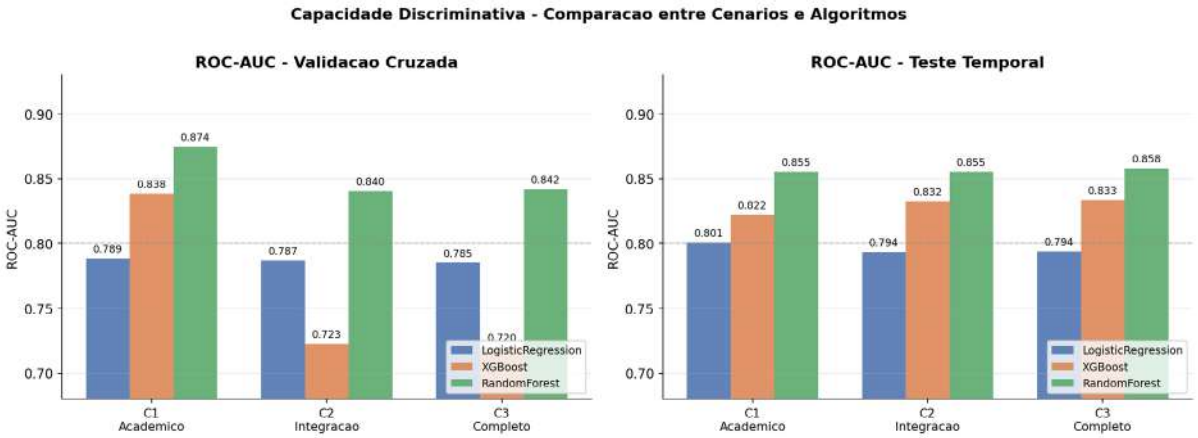
### 3.4.1 Do cenário acadêmico ao cenário integrado: o que os dados revelam

O Cenário C1, restrito a dados do sistema acadêmico, já apresenta desempenho expressivo: ROC-AUC de 0,855 (*Random Forest*) e *Recall* de 0,863 no conjunto de teste. Esse resultado tem implicação direta para gestores: a ausência de integração avançada não impede a implantação de um sistema funcional de alerta precoce. As IFES que dispõem de dados acadêmicos estruturados, tais como desempenho semestral, taxas de aprovação, progressão no curso e trancamentos, podem operacionalizar modelos preditivos sem depender previamente de integração com sistemas socioeconômicos externos. O custo de governança é baixo; o sinal preditivo já existe.

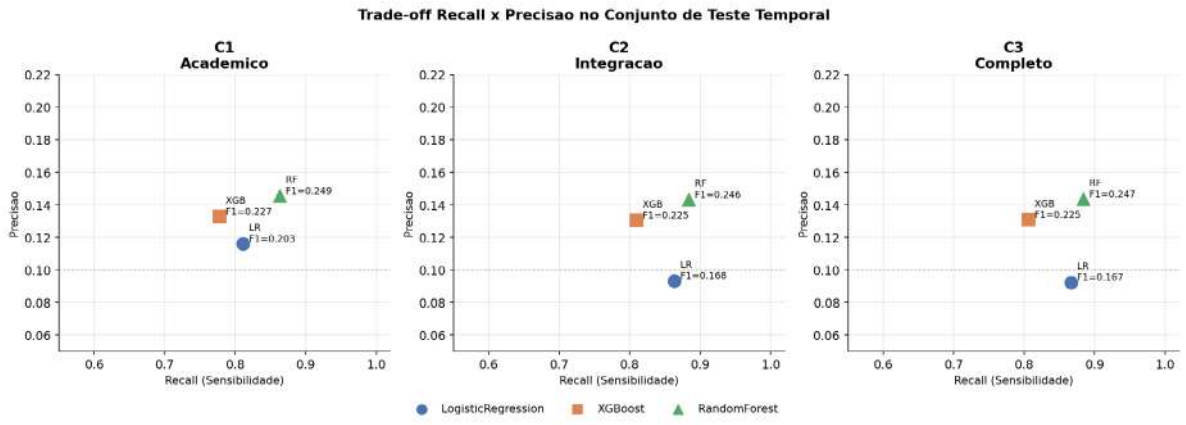
## Tabela 4: Desempenho dos modelos por cenário — validação cruzada e teste temporal

| Cenário | Modelo | AUC-CV | AUC-Teste | Recall | Precisão | F1 |
|---|---|---|---|---|---|---|
| C1 — Acadêmico | Random Forest | 0,874 | 0,855 | 0,863 | 0,146 | 0,249 |
| | XGBoost | 0,838 | 0,822 | 0,778 | 0,133 | 0,227 |
| | Reg. Logística | 0,789 | 0,801 | 0,811 | 0,116 | 0,203 |
| C2 — Ampliado | Random Forest | 0,840 | 0,855 | 0,884 | 0,143 | 0,246 |
| | XGBoost | 0,723 | 0,832 | 0,809 | 0,131 | 0,225 |
| | Reg. Logística | 0,787 | 0,794 | 0,864 | 0,093 | 0,168 |
| C3 — Completo | Random Forest | 0,842 | 0,858 | 0,884 | 0,143 | 0,247 |
| | XGBoost | 0,720 | 0,833 | 0,806 | 0,131 | 0,225 |
| | Reg. Logística | 0,785 | 0,794 | 0,866 | 0,092 | 0,167 |

## Figura 1: ROC-AUC por cenário e algoritmo — validação cruzada vs. teste temporal




## Figura 2: Trade-off Recall × Precisão no conjunto de teste temporal, por cenário




A principal limitação de C1 é a invisibilidade sobre vulnerabilidades não acadêmicas: estudantes em risco por razões financeiras ou geográficas, mas com desempenho ainda preservado, não são identificados. Esse é precisamente o perfil que mais se beneficiaria de intervenção preventiva.

A inclusão de dados de ingresso e perfil demográfico em C2 (forma de acesso, modalidade de cota, região de origem, raça/cor e sexo) mantém o ROC-AUC em 0,855 para o Random Forest, com ganho notável no Recall: 0,884 (+2,1 pontos percentuais sobre C1). O modelo passa a capturar padrões associados à trajetória de acesso, ou seja, dimensões estruturais de vulnerabilidade que antecedem o desempenho acadêmico. Esse ganho em sensibilidade tem significado prático: em um conjunto de teste com mais de 4.700 evasões, cada ponto percentual de Recall equivale a aproximadamente 47 estudantes adicionalmente identificados.

Adicionalmente, a presença de aproximadamente 40% de nulos na variável de cota de ingresso, tratados como *ampla concorrência* por regra semântica, é evidência direta de limitação de governança de metadados. Sua correta codificação nos sistemas de origem produziria variável de maior precisão e potencialmente melhoraria o desempenho de C2.

O Cenário C3, com inclusão de dados de assistência estudantil, alcança o melhor desempenho geral: ROC-AUC de 0,858 e Recall de 0,884 (Random Forest). O ganho sobre C1 em AUC é modesto (+0,003), mas o ganho qualitativo é substancial: o modelo passa a enxergar padrões associados ao acesso, ou à ausência de acesso, ao sistema de assistência estudantil. Estudantes que tentaram bolsas e foram indeferidos ou que nunca interagiram com o sistema tornam-se analiticamente visíveis.

Há, entretanto, uma limitação crítica: estudantes vulneráveis, que nunca acessaram o programa de assistência estudantil, apresentam perfil assistencial nulo, podendo ser subestimados pelo modelo. Paradoxalmente, os mais invisíveis operacionalmente continuam invisíveis analiticamente. Esse achado reforça que o modelo preditivo deve ser compreendido como complemento a mecanismos de busca ativa, não como substituto.

### 3.4.2 Custo operacional da baixa maturidade em governança

Um achado transversal aos três cenários diz respeito ao custo da fragmentação de dados. A engenharia de *features* necessária para este estudo, definida pela construção de variáveis como índice_instabilidade, queda_desempenho, taxa-atividade e tentou-bolsa, é uma medida do esforço analítico adicional imposto pela ausência de governança. Em contextos de maior maturidade, esses indicadores já existiriam como ativos computados e documentados nos sistemas institucionais, disponíveis para consumo imediato por qualquer análise subsequente.

Esse padrão é empiricamente consistente com o observado por Pahune et al. (2025), para quem a maturidade em governança de dados condiciona o desempenho preditivo ao longo de todo o ciclo de vida do modelo, desde a preparação dos dados à operação contínua. A fragmentação e a ausência de padronização semântica introduzem riscos de inconsistência que comprometem tanto o desenvolvimento quanto a sustentação de modelos em contextos de uso real.

## 3.5 Importância das variáveis: valor analítico da integração de dados

A análise de importância das variáveis fornece o argumento mais direto em favor do investimento progressivo em governança e integração de dados. A Figura 3 apresenta as variáveis de maior importância por cenário (Random Forest) e a participação relativa de cada grupo de variáveis no poder preditivo total.

Em todos os cenários, variáveis de desempenho acadêmico, como cra, taxa_aprovacao_semestre, indice_instabilidade e queda_desempenho, concentram alto poder preditivo, confirmando achados da literatura sobre a centralidade do desempenho na trajetória de evasão (Solis et al., 2018; Oliveira et al., 2021). Contudo, a progressão dos cenários revela padrões relevantes.

Em C2, variáveis de ingresso e perfil demográfico emergem como preditores complementares: ano_ingresso, idade_ingresso e regiao_naturalidade apresentam importância não nula, evidenciando que vulnerabilidades estruturais de acesso produzem sinal preditivo independente do desempenho acadêmico. Em C3, apenas a origem escolar de domínio Não Informado, mapeada pelo *OneHot* figura na lista de variáveis com valor preditivo. As variáveis de assistência estudantil, tais como tentou_bolsa, bolsas_indeferidas e semestres_com_bolsa, não constam entre os preditores relevantes, indicando que o padrão de acesso ao sistema assistencial pode conter informação sobre risco de evasão, mas que não é capturada pelas demais fontes.

A análise de importância pelo Random Forest no cenário C3 (Figura 4) confirma os mesmos padrões, com ligeiras diferenças na ordenação relativa, reforçando a estabilidade dos achados entre algoritmos.

**Figura 3: Importância das variáveis (Random Forest) por cenário — as cores indicam a origem da integração**

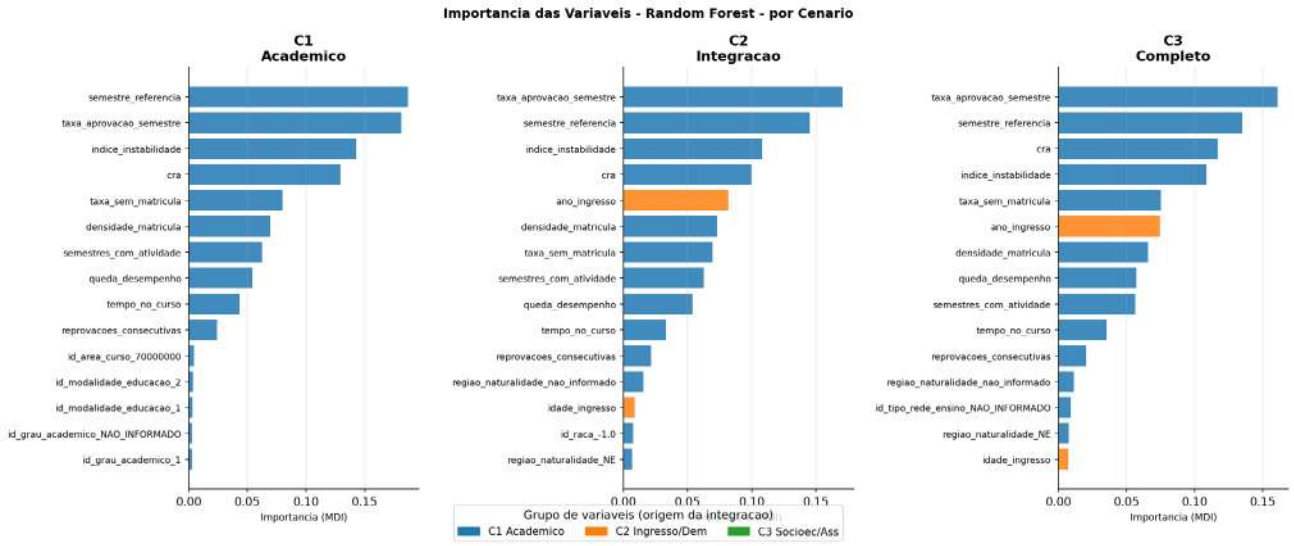
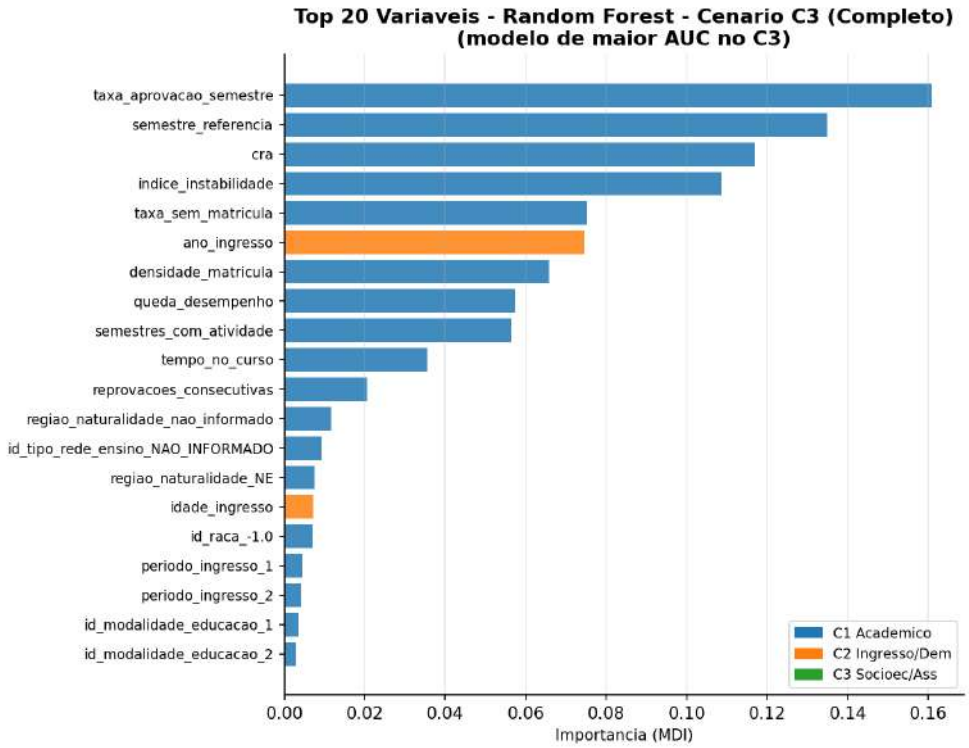


**Figura 4: Importância das variáveis — Random Forest, Cenário C3 (Completo)**




Para gestores institucionais, o achado central desta análise é o seguinte: a cobertura assistencial baixa (aproximadamente 12% dos registros com interação com o sistema de assistência estudantil) limita o sinal preditivo de C3. Isso significa que estudantes vulneráveis que nunca acessaram o sistema de bolsas estão sub-representados nos dados e, consequentemente, subidentificados pelo modelo. A expansão da cobertura assistencial, além de diretamente relevante para a permanência, melhora a qualidade analítica dos modelos preditivos.

## *3.6 Limiar de decisão e implicações operacionais*

### *3.6.1 O threshold como decisão de política pública*

Para o alerta precoce da evasão estudantil, a escolha do limiar de decisão (*th-reshold*) é frequentemente tratada como um parâmetro de otimização de métricas. No contexto de políticas de permanência, essa perspectiva é insuficiente. O *threshold* determina quantos estudantes serão sinalizados por semestre para acompanhamento institucional e, portanto, sua definição é uma decisão de gestão, não de estatística: ela deve ser calibrada à capacidade de resposta institucional disponível.

A Figura 5 apresenta a curva Precision-Recall em função do limiar, para os três algoritmos no cenário C3. A Tabela 5 detalha o *trade-off* operacional para o Random Forest, modelo de melhor desempenho global.

### 3.6.2 Custo assimétrico dos erros e recomendação operacional

No contexto educacional, os erros de classificação possuem custo assimétrico. Não identificar um estudante que irá evadir, representado por um falso negativo, implica na perda de oportunidade de retenção, com impactos diretos sobre o estudante e sobre a eficiência institucional. Um falso positivo, que sinaliza um estudante que não irá evadir, tem custo relativo baixo: se a intervenção prevista é um contato de acompanhamento ou oferta de serviço, o estudante incorretamente sinalizado pode, inclusive, se beneficiar do contato proativo.

Essa assimetria sugere que limiares mais baixos, com maior Recall e mais falsos positivos, são justificáveis institucionalmente, desde que a capacidade de resposta comporte o volume de sinalizações gerado.

**Figura 5: Curva Precision-Recall × Limiar de decisão — Cenário C3 (três algoritmos)**

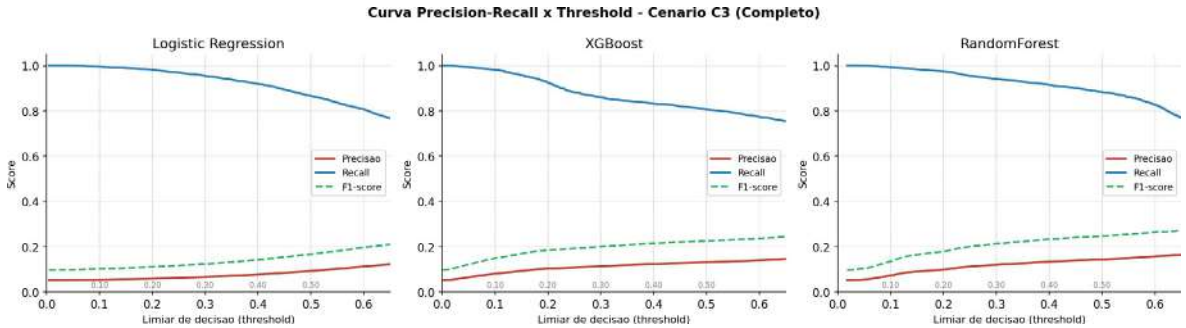



**Tabela 5: Trade-off operacional por limiar — Random Forest, Cenário C3, conjunto de teste (92.343 obs.)**

| Limiar | Recall | Precisão | F1 | Sinalizados | Evasões identificadas |
|---|---|---|---|---|---|
| 0,10 | 0,900 | 0,117 | 0,208 | 35.620 | 4.152 |
| 0,15 | 0,883 | 0,134 | 0,232 | 30.546 | 4.075 |
| 0,20 | 0,858 | 0,143 | 0,247 | 27.810 | 3.958 |
| 0,25 | 0,795 | 0,162 | 0,266 | 22.740 | 3.670 |
| 0,30 | 0,721 | 0,183 | 0,289 | 18.250 | 3.328 |
| 0,40 | 0,534 | 0,225 | 0,316 | 10.987 | 2.465 |
| 0,50 | 0,358 | 0,274 | 0,311 | 6.030 | 1.655 |

A recomendação para implantação institucional é a seguinte: (i) levantar a capacidade de resposta semestral da equipe de assistência estudantil, indicando quantos contatos ou triagens é possível processar por semestre; (ii) mapear o tipo de intervenção prevista, observando que intervenções mais leves (contato automatizado, alerta no sistema do coordenador de curso) suportam volumes maiores; e (iii) selecionar o limiar correspondente ao volume operacionalizável, priorizando Recall quando a capacidade for maior e Precisão quando a capacidade for restrita. Um exemplo aplicado: se a equipe de assistência consegue processar até 20.000 contatos por semestre, o limiar 0,25 (22.740 sinalizações, identificando 3.670 evasões com 16,1% de precisão) é o mais adequado; se o mecanismo for um alerta automático no sistema do coordenador de curso, o limiar 0,20 é viável.

Quatro cautelas são pertinentes para uma implantação responsável. Primeiramente, o modelo foi treinado incluindo o período pandêmico (*concept drift*), durante o qual a evasão apresentou padrão atípico. Recomenda-se reavaliação semestral do desempenho e retreinamento periódico com dados mais recentes. Segundo, que a invisibilidade da vulnerabilidade não declarada por estudantes em situação de risco que nunca acessaram o sistema de assistência estudantil tende a ter *score* preditivo subestimado no Cenário C3; o modelo não substitui mecanismos de busca ativa. Terceiro, a sinalização do modelo deve ser acompanhada de ferramenta de explicação local (como SHAP *values*) com interpretabilidade para que gestores compreendam as razões de uma sinalização específica, condição necessária para adoção institucional responsável e alinhada aos princípios de transparência da LGPD. Quarto, a sinalização gerada pelo modelo deve ser insumo para decisão humana, nunca automação da intervenção, conforme recomendado por Pahune et al. (2025) no contexto da governança de IA em políticas públicas.

### *3.7 Síntese dos achados e resposta às perguntas de pesquisa*

Os experimentos demonstram que dados acadêmicos básicos já possibilitam modelos com ROC-AUC de 0,855, indicando que o sinal preditivo fundamental reside no desempenho e na trajetória de matrícula. A integração progressiva com dados de ingresso, perfil demográfico e assistência estudantil amplia o Recall em 2,1 pontos percentuais e o AUC em 0,003, ganhos que, no volume de estudantes analisados, têm significado prático. Mais importante do que os ganhos em métricas globais, a integração amplia a *equidade analítica do modelo*, tornando visíveis dimensões de vulnerabilidade que dados acadêmicos isolados não capturam.

Os achados evidenciam, ainda, que a governança não é apenas condição para modelos melhores, mas condição para modelos mais justos, mais estáveis e mais operacionalizáveis. O esforço de preparação dos dados, estimado em variáveis que precisam ser reconstruídas manualmente por ausência de documentação, integração semântica e padronização, é a medida direta do custo da baixa maturidade. Governança mais robusta não apenas reduziria esse custo, mas criaria as condições para monitoramento contínuo da permanência estudantil sem dependência de projetos analíticos pontuais.

Por fim, a análise demonstra que não há uma resposta estatisticamente ótima, mas uma resposta operacionalmente adequada: o limiar deve ser derivado da capacidade de resposta institucional, do tipo de intervenção prevista e dos objetivos da política de permanência. Esse enquadramento transforma o *threshold* de parâmetro técnico em instrumento de política pública, colocando o gestor no centro da decisão.

## CONCLUSÃO

Este estudo investigou, de forma empírica e aplicada, como a qualidade e a integração de dados institucionais condicionam a capacidade analítica para identificação precoce de estudantes em risco de evasão no ensino superior público. Os resultados contribuem para a literatura e para a prática de gestão universitária em três dimensões complementares.

No plano teórico, o estudo demonstra que modelos preditivos de evasão não podem ser avaliados isoladamente pela ótica do desempenho estatístico: eles são fortemente condicionados pela maturidade da governança de dados da instituição que os alimenta. A comparação entre cenários progressivos de integração evidencia que governança de dados

não é apenas pressuposto normativo, mas determinante mensurável da capacidade analítica institucional, achado que avança sobre a literatura de *learning analytics*, que frequentemente trata a qualidade dos dados como dada, e não como objeto de análise (Al-Ruithe et al., 2019; Pahune et al., 2025).

No plano metodológico, propõe-se e valida-se um *pipeline* integrado que articula, de forma replicável, quatro componentes: (i) diagnóstico de maturidade em governança de dados; (ii) preparação e qualificação de bases institucionais com decisões metodológicas documentadas; (iii) modelagem preditiva em cenários progressivos de integração; e (iv) análise do limiar de decisão como instrumento de política pública de permanência. Esse *pipeline* é transferível a outras IFES e pode ser adaptado para contextos com diferentes níveis de disponibilidade informacional, aspecto relevante dado o diagnóstico do TCU sobre o estágio ainda inicial da maioria das instituições públicas brasileiras em governança de dados (Tribunal de Contas da União, 2026).

No plano prático, os achados fornecem recomendações diretas para gestores institucionais. De imediato, é possível começar com dados acadêmicos básicos: o Cenário C1 demonstra que um sistema de alerta precoce funcional pode ser implantado com as informações já disponíveis em sistemas acadêmicos estruturados, sem aguardar integração plena. A espera pela governança ideal não é condição necessária para iniciar; é condição para qualificar progressivamente.

Adicionalmente, a integração de dados tem valor analítico real: cada camada adicional de integração acrescenta sinal preditivo sobre dimensões de vulnerabilidade distintas e complementares, ampliando a equidade na identificação precoce de evasão e tornando visíveis estudantes que dados acadêmicos isolados não alcançam.

Finalmente, o limiar de decisão é uma política, não um parâmetro, e sua calibração deve ser orientada pela capacidade de resposta institucional e pelo tipo de intervenção prevista, colocando o gestor no centro da decisão sobre quem recebe atenção e sob qual modelo de acompanhamento.

A análise de importância das variáveis reforça um achado de natureza mais ampla: a baixa cobertura assistencial (aproximadamente 12% dos estudantes com interação documentada com o sistema de bolsas) não é apenas um problema de política de permanência, como também é um problema de qualidade de dados. Estudantes vulneráveis que nunca acessaram o sistema de assistência são invisíveis tanto para as políticas quanto para os modelos. Esse é o ponto de convergência mais relevante entre governança de dados e permanência estudantil: as lacunas de dados refletem lacunas de cobertura real, e melhorar a governança aperfeiçoa simultaneamente a efetividade das políticas e a qualidade das análises que as orientam.

O estudo apresenta limitações que devem ser consideradas. A análise de maturidade em governança, de caráter exploratório e baseada em análise documental e interações com partes interessadas, não segue protocolo de auditoria formal e não permite comparação sistemática com outras instituições. A questão do *concept drift*, decorrente da inclusão do período pandêmico no treino, representa uma limitação importante para a generalização dos resultados ao contexto atual e deve ser endereçada com retreinamento periódico e monitoramento contínuo do desempenho. Adicionalmente, a interpretabilidade local dos modelos, condição necessária para adoção institucional responsável, não foi implementada neste estudo, sendo recomendada como extensão prioritária por meio de técnicas como SHAP *values*.

Como agenda de pesquisa futura, sugere-se: (i) a replicação do experimento em outras IFES com diferentes perfis de maturidade em governança, para verificar a generalização dos achados; (ii) a investigação do impacto do *concept drift* pós-pandêmico no desempenho dos modelos e a comparação de estratégias de atualização contínua; (iii) o desenvolvimento de instrumentos de explicabilidade local integrados ao *pipeline*, visando à adoção institucional responsável; e (iv) a avaliação de estratégias de busca ativa como complemento ao sistema preditivo, especialmente para estudantes com perfil assistencial nulo.

A sinalização gerada pelo modelo deve ser compreendida como insumo qualificado para a decisão humana e não como substituto do julgamento técnico das equipes de assistência estudantil. A responsabilidade pela intervenção permanece com as pessoas: os dados e os modelos ampliam sua capacidade de agir; a governança garante que essa capacidade seja justa, confiável e sustentável.

Os resultados reforçam que políticas de permanência orientadas por dados dependem menos de novos sistemas e mais da governança, integração e uso estratégico dos dados já disponíveis nas instituições.

## *Disponibilidade de Dados e Código*

Os dados processados, *scripts* de preparação, modelagem e avaliação utilizados neste estudo encontram-se disponíveis em repositório público para fins de reprodutibilidade científica.

Dados:

- **Zenodo:** https://doi.org/10.5281/zenodo.21001025

Código-fonte:

- **GitHub:** https://github.com/fabianan/governanca-dados-permanencia-estudantil

## REFERÊNCIAS


AL-RUITHE, M. et al. A systematic literature review of data governance and cloud data governance. **Personal and Ubiquitous Computing**, v. 23, n. 5–6, p. 839–859, 2019.

Brasil. **Lei nº 13.709, de 14 de agosto de 2018** (Lei Geral de Proteção de Dados Pessoais). 2018.

Brasil. **Decreto nº 12.198, de 24 de setembro de 2024**. 2024. Estratégia Federal de Governo Digital.

Brasil. **Lei nº 14.914, de 3 de julho de 2024**. 2024. Política Nacional de Assistência Estudantil.

DAKE, D. K.; BUABENG-ANDOH, C. Using machine learning techniques to predict learner drop-out rate. **Mobile Information Systems**, p. 2670562, 2022.

DAMA International. **The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK2)**. [S.l.]: Technics Publications, 2017.

DWECK, E.; MASCARENHAS, R. S. **Guia de implementação do programa de governança de dados**. [S.l.: s.n.], 2025.

FRAGA, V. d. S. **Governança de dados na transformação digital**. [S.l.]: ENAP, 2022. INEP. Censo da Educação Superior 2024: divulgação dos resultados. [S.l.], 2025.

JIANG, Z.; ZHU, Z.; PAN, S. Data governance in multimodal behavioral research. **Multimodal Technologies and Interaction**, v. 8, n. 7, p. 55, 2024.

OLIVEIRA, C. F. de et al. How does learning analytics contribute to prevent students' dropout. **Big Data and Cognitive Computing**, v. 5, n. 4, p. 64, 2021.

PAHUNE, S. et al. The importance of ai data governance in large language models. **Big Data and Cognitive Computing**, v. 9, n. 6, p. 147, 2025.

PARRA, P. M. et al. Investigating student dropout risk through machine learning. In:**SBIE**. [S.l.: s.n.], 2024.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

ROMERO, C.; VENTURA, S. Educational data mining: a review. **IEEE Transactions**, v. 40, n. 6, p. 601–618, 2010.

SIEMENS, G.; BAKER, R. Learning analytics and educational data mining. In: **Learning Analytics Conference**. [S.l.: s.n.], 2012.

SOLIS, M. et al. Perspectives to predict dropout in university students with machine learning. In: **Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)**. San Carlos: IEEE, 2018.

TINTO, V. **Leaving College: Rethinking the Causes and Cures of Student Attrition**. Chicago: University of Chicago Press, 2012.

Tribunal de Contas da União. **Acórdão nº 457/2026 – Plenário**. 2026. Relatório de auditoria sobre governança de dados na Administração Pública Federal.

Universidade Federal da Paraíba. **SIGOV/UFPB: Sistema de Inteligência e Governança de Dados Institucionais**. João Pessoa: [s.n.], 2025. Acesso em: jan. 2026. Disponível em: <https://sigov.ufpb.br>.

VENKATESAN, R. G. et al. Exploring statistical approaches for predicting student dropout. **Journal of Computational Social Science**, v. 7, n. 1, p. 171–196, 2024.

# Classificação de Línguas e Etnias Indígenas nos Censos Demográfico e Agropecuário por meio de Machine Learning

*Luís César Seixas de Oliveira*
*Prof. Dr. Hélio Bonfim de Macêdo Filho*

## Resumo

A captação da autoidentificação de povos indígenas no Brasil, no que tange às informações de etnias e línguas faladas, representou avanço metodológico importante no Censo Demográfico de 2022. A fim de possibilitar a autodeclaração plena, uma parcela volumosa da coleta se deu por meio de dados textuais não estruturados, exigindo intensa revisão de especialistas na codificação dessas informações e dilatando o prazo de divulgação dessas estatísticas. Este estudo tem como objetivo desenvolver e validar modelos de aprendizado de máquina que reproduzem a classificação especializada desses registros textuais. O desenvolvimento metodológico abrange a análise de desempenho dos modelos e técnicas de classificação capazes de lidar com cenários de alta cardinalidade, utilizando como "verdade fundamental" os microdados já criticados e validados do Censo Demográfico de 2022, possibilitando a implementação de um *pipeline* de desempenho satisfatório e factível de aprimorar a eficiência do processo de codificação em operações futuras.

**Palavras-chave:** aprendizado de máquina, classificação, etnias e línguas indígenas.

## INTRODUÇÃO

O Censo Demográfico de 2022 representou um avanço significativo na produção de estatísticas oficiais sobre os povos originários, alcançado por meio de um refinamento na metodologia de captação da autoidentificação indígena. O Instituto Brasileiro de Geografia e Estatística (IBGE) aprimorou o questionário de coleta de modo a possibilitar a ampliação da captação de etnias e línguas faladas para além dos residentes em terras indígenas, contemplando, também, aqueles indivíduos que se autodeclararam indígenas fora dessas localidades (IBGE, 2025). A ampliação da abordagem permitiu um retrato mais fiel da diversidade sociocultural, mitigando a subenumeração histórica e reconhecendo a pluralidade de denominações existentes no território nacional.

A realização de um censo demográfico em um país de dimensão e heterogeneidade cultural imensas como o Brasil impõe desafios metodológicos e operacionais de grande magnitude. Um ponto de suma importância na cadeia de produção da informação estatística reside na etapa de crítica e codificação das informações coletadas. Ao contrário da maior parte do questionário de coleta estruturada em perguntas para seleção, a captação das etnias e línguas indígenas exige, necessariamente, a abertura de campos para a autodeclaração em função da diversidade de etnônimos e da complexidade linguística dos povos indígenas.

No Censo Demográfico de 2022, o IBGE implementou inovações tecnológicas, como o uso de Bancos Descritores (BDD) carregados nos Dispositivos Móveis de Coleta (DMC) que permitiam ao recenseador selecionar opções de etnias e línguas indígenas pré-cadastradas. Contudo, a realidade de campo frequentemente supera a taxonomia pré-estabelecida. Sendo assim, quando a etnia ou língua declarada pelo informante não estava presente na lista do aplicativo ou o recenseador entendia não haver correspondência, as respostas do informante eram registradas como entradas textuais livres. Embora essencial para garantir o direito à autoidentificação, esse procedimento gerou um considerável volume de dados não estruturados que demandou tratamento posterior complexo (IBGE, 2025).

Nesse contexto, a complexidade metodológica inerente ao tratamento das informações coletadas em campo demandou uma curadoria rigorosa na interpretação da heterogeneidade das autodeclarações e suas variações linguísticas que resultou na consolidação de uma base criticada e validada por especialistas e que padroniza as diversas grafias e denominações locais em bancos descritores atualizados e mais consistentes. A consolidação dessa base de referência é estratégica, pois lança os fundamentos para a possível automatização da atividade de codificação em operações futuras, como o Censo Agropecuário 2027 e o Censo Demográfico de 2030, com foco na celeridade e na redução de esforços, sem sacrifício à sensibilidade antropológica necessária à correta identificação das etnias e línguas.

O problema central que motiva esta pesquisa é o esforço envolvido no processo de codificação realizado no Censo Demográfico 2022, que envolveu etapas exaustivas de pareamento determinístico (busca exata) e probabilístico básico (similaridade de texto), seguidas de intensa revisão manual por especialistas para resolver ambiguidades, homônimos e novas grafias não previstas nos dicionários de dados. A análise manual, realizada pelo Grupo de Trabalho de Povos e Comunidades Tradicionais, envolveu a verificação caso a caso, consultas a entidades externas e a aplicação de regras complexas de desambiguação. Esse fluxo, embora qualificado, é humanamente oneroso, além de gerar impacto no cronograma de divulgação dos resultados censitários e, consequentemente, nas políticas públicas que se utilizam dessas informações.

A problemática desdobra-se em questões específicas:

a) **ambiguidades:**

- nomes de etnias frequentemente apresentam grafias similares para grupos antropologicamente distintos, cuja diferenciação só é possível por meio de variáveis de contextualização;

b) **variabilidade da escrita:**

- campos de texto livre admitem erros de digitação, variações fonéticas na grafia de nomes indígenas e o uso de termos genéricos que dificultam a padronização via regras de negócio rígidas (“se-então”);

c) **volume dos dados:**

- o tratamento manual de milhares de registros textuais cria um gargalo que impede a escalabilidade do processo para pesquisas contínuas ou censos futuros;

d) **dependência de especialistas:**

- a codificação manual depende de um conhecimento tácito de especialista, o que é difícil de transferir ou reproduzir sistematicamente sem o auxílio de modelos que identifiquem esses padrões decisórios.

Diante desse cenário, formula-se a seguinte pergunta de pesquisa: Em que medida modelos de aprendizado de máquina treinados com o histórico de dados criticados, codificados e validados no Censo Demográfico 2022 podem reproduzir a classificação de etnias e línguas indígenas, garantindo a consistência dos dados e a redução do tempo de processamento em operações futuras?

A relevância deste trabalho pode ser analisada sob três dimensões principais: acadêmica, institucional e social.

a) **acadêmica:**

- o tema dialoga diretamente com a formação em ciência de dados e inteligência artificial aplicadas. A disponibilidade de uma base extensa de registros de pessoas indígenas, com cerca de 1,6 milhão de observações, e de bancos de descritores consolidados oferece um cenário favorável à aplicação de técnicas de aprendizado de máquina. A exploração inicial desenvolvida na disciplina "*Machine Learning*" deste MBA demonstrou que é possível integrar múltiplas fontes (descritores de etnias e línguas, registros de coleta, registros criticados e codificados) e construir pipelines de pré-processamento e classificação com desempenho inicial adequado;

b) **institucional:**

- o IBGE tem buscado, em parceria com outros órgãos e por meio de grupos de trabalho sobre povos e comunidades tradicionais, aprimorar os instrumentos de coleta e de classificação de etnias e línguas. A experiência do Censo Demográfico de 2022 resultou em um novo banco de descritores mais aderente à realidade em campo, mas também expôs o volume de trabalho envolvido e a necessidade de sistematização de regras e procedimentos. Uma solução baseada em aprendizado de máquina, ainda que não substitua o olhar especializado, pode servir como ferramenta de apoio à codificação, à detecção de incoerências e à identificação de novos etnônimos;

c) **social:**

- o reconhecimento da diversidade étnica e linguística dos povos indígenas é fundamental para a formulação de políticas públicas de proteção territorial, promoção de direitos e valorização cultural. A qualidade das estatísticas oficiais sobre etnias e línguas indígenas impacta diretamente a visibilidade desses povos em diagnósticos, planos e programas governamentais. Processos de codificação inconsistentes, pouco transparentes ou excessivamente onerosos podem comprometer a capacidade de resposta do Estado às demandas dessas populações.

Este estudo apresenta os experimentos realizados para a predição de etnias e línguas indígenas por meio de modelos desenvolvidos utilizando os algoritmos *K-Nearest Neighbours*, Árvore de Decisão, Floresta Aleatória, Support Vector Classifier e Redes Neurais Profundas.

### Desenvolvimento

O Censo Demográfico 2022 identificou 1.694.836 indígenas, distribuídos em 391 etnias e 295 línguas faladas (IBGE, 2025). A natureza declaratória do censo permitiu múltiplas afiliações étnicas e linguísticas, sugerindo a utilização de algoritmos que superem a classificação binária ou multiclasse tradicional para o desenvolvimento de um modelo de aprendizado de máquina que emule o processo de crítica e codificação do Censo Demográfico 2022.

## 1. REFERENCIAL TEÓRICO

### 1.1 Aprendizado de Máquina

A definição clássica de Samuel (1959) conceitua o Aprendizado de Máquina como a capacidade de dar aos computadores a habilidade de aprender, sem que eles sejam explicitamente programados para funções específicas. Sob uma perspectiva técnica, Géron (2023) reforça a definição de Mitchell (1997), na qual um programa de computador aprende em relação a uma tarefa T e uma medida de desempenho P, se seu desempenho na tarefa T, medido por P, melhora com a experiência E.

O Aprendizado de Máquina substitui a programação tradicional, baseada em regras manuais e específicas de "se-então", por sistemas capazes de realizar a mineração dos dados e de se adaptar a problemas de alta complexidade. Algoritmos de aprendizado supervisionado podem ser treinados para identificar padrões em dados históricos e generalizar esse conhecimento para novos casos, reduzindo, assim, a necessidade de intervenção manual.

### 1.2 K-Nearest Neighbours (KNN)

Originalmente, o conceito de vizinhos mais próximos foi introduzido como uma solução para problemas de análise discriminatória em que a distribuição de probabilidade das populações não era conhecida. Conforme o trabalho pioneiro de Fix e Hodges (1951), a premissa central é a de que observações de uma mesma classe tendem a estar próximas umas das outras no espaço de atributos. O KNN estima a densidade de probabilidade local, ou seja, para classificar um novo ponto, o algoritmo identifica as $K$ observações mais próximas no conjunto de dados e a decisão de classificação é tomada por maioria, onde o novo ponto é atribuído à classe mais frequente entre seus vizinhos. Essa proximidade é geralmente calculada por meio de métricas de distância, sendo a mais comum a distância euclidiana (FIX; HODGES, 1951).

### 1.3 Árvores de Decisão

As árvores de decisão são modelos preditivos que utilizam uma estratégia de "dividir para conquistar" para classificar objetos a partir de um conjunto de atributos. De acordo com Quinlan (1986), essa estrutura consiste em uma hierarquia de nós onde cada nó interno representa um teste sobre um atributo específico, cada ramo corresponde ao resultado desse teste e cada folha (nó terminal) atribui uma classe ou categoria aos dados. O processo de indução busca encontrar uma árvore que descreva corretamente os dados

de treinamento, utilizando critérios de seleção de atributos fundamentados na teoria da informação para minimizar a complexidade da estrutura e maximizar a eficiência da classificação (Quinlan, 1986).

## 1.4 Floresta Aleatória (Random Forest)

As florestas aleatórias são um método de aprendizado de máquina fundamentado na combinação de múltiplos classificadores de árvore de decisão. Segundo o trabalho de Breiman (2001), uma floresta aleatória consiste em um classificador composto por uma coleção de árvores estruturadas, onde cada árvore individual é gerada a partir de um vetor aleatório amostrado de forma independente e com a mesma distribuição para todas as árvores da floresta. O modelo opera por meio do princípio de *ensemble learning*, onde cada árvore elege a classe mais popular para uma determinada entrada. A principal inovação proposta por Breiman (2001) reside no fato de que o erro de generalização do modelo depende da força das árvores individuais e da correlação entre elas. Ao introduzir a seleção aleatória de atributos em cada nó da árvore, o algoritmo reduz essa correlação, aumentando significativamente a precisão e a resistência ao sobreajuste (*overfitting*).

## 1.5 Support Vector Classifier (SVC)

Segundo o trabalho de Cortes e Vapnik (1995), o objetivo primordial de um classificador de vetores de suporte é construir um hiperplano ótimo que separe duas classes de dados, maximizando a distância entre as observações mais próximas de cada classe e o próprio hiperplano.

A inovação apresentada é a introdução de variáveis de folga que permitem que o algoritmo lide com dados que não são perfeitamente separáveis. Em vez de buscar uma separação rígida (*hard margin*), o SVC busca um equilíbrio entre maximizar a margem de separação e minimizar o erro de classificação.

Os pontos que residem exatamente na fronteira da margem ou que a violam são chamados de vetores de suporte (*support vectors*). Eles são os únicos elementos do conjunto de dados que determinam a posição e orientação do hiperplano de decisão. Matematicamente, a busca pelo hiperplano ótimo é formulada como um problema de otimização quadrática utilizando o vetor normal ao hiperplano e o parâmetro de regularização que controla a penalidade nas classificações incorretas (Cortes; Vapnik, 1995).

## 1.6 Distâncias de Edição

De acordo com Levenshtein (1966), a distância entre duas sequências de caracteres é o número mínimo de operações elementares necessárias para transformar uma cadeia de caracteres em outra. As operações permitidas nesse modelo seriam a inserção, a deleção ou exclusão e a substituição de um único caractere. Seu trabalho estabeleceu a base teórica para o campo de busca por aproximação e processamento de linguagem natural, permitindo que sistemas computacionais quantifiquem a similaridade entre textos e lidem com erros de digitação ou variações estruturais.

Como forma de avaliar a correspondência entre a entrada não estruturada e a codificação padronizada adotada, as seguintes técnicas computacionais clássicas de distância entre cadeias de caracteres foram consideradas:

a) **Distância de Levenshtein (Levenshtein, 1966):** utilizada para quantificar o número mínimo de edições de caracteres únicos (inserções, exclusões ou substituições) necessárias para transformar o texto de entrada na nomenclatura padronizada.
b) **Distância de Jaro-Winkler (Jaro, 1989; Winkler, 1990):** adotada como métrica complementar por atribuir maior peso a correspondências nos prefixos das palavras, sendo especialmente útil para identificar erros de digitação comuns no início de termos e variações morfológicas.
c) **Distância de Damerau-Levenshtein (Damerau, 1964; Levenshtein, 1966):** expande a métrica de Levenshtein ao incluir a transposição de dois caracteres adjacentes como uma operação fundamental. É utilizada para calcular a variação mínima entre duas cadeias de caracteres por meio de inserções, exclusões, substituições e inversões, sendo altamente eficaz na detecção de erros de digitação humanos.

## *1.7 Avaliação de Desempenho*

Métricas como acurácia, precisão, revocação e F1-*score* são fundamentais para a avaliação do desempenho de um modelo nas etapas de treinamento, avaliação e teste (Géron, 2023).

a) **Acurácia:** representa a fração de previsões corretas (tanto positivos quanto negativos) em relação ao total de instâncias avaliadas.
b) **Precisão:** indica a exatidão das previsões positivas, calculada como a razão entre os verdadeiros positivos e o total de instâncias classificadas como positivas pelo modelo.
c) **Revocação (ou Sensibilidade):** mede a capacidade do modelo em encontrar todas as instâncias positivas, sendo a razão entre os verdadeiros positivos e o total de instâncias que são realmente positivas.
d) **F1-*score*:** consiste na média harmônica entre a precisão e a revocação, fornecendo um equilíbrio entre ambas, sendo especialmente útil quando há uma distribuição de classes desigual.

## *1.8 Validação e Ajuste de Hiperparâmetros (Tuning)*

De modo a assegurar a capacidade de generalização e otimizar o desempenho de um modelo, aplicam-se os processos de validação e de ajuste de hiperparâmetros (Géron, 2023).

A validação é uma técnica que utiliza um subconjunto dos dados (conjunto de validação) para comparar diferentes modelos e selecionar o que apresenta melhor desempenho, evitando que o modelo final seja escolhido com base apenas no erro de treinamento e sofra de sobreajuste (*overfitting*).

O Ajuste de Hiperparâmetros é o processo de busca sistemática pela combinação ideal de configurações de parâmetros de um algoritmo, como a profundidade de uma árvore ou a taxa de aprendizado, com o objetivo de maximizar o desempenho do modelo utilizando os dados de validação.

### 1.9 Interpretabilidade

A interpretabilidade dos modelos de aprendizado de máquina é de extrema importância em aplicações institucionais, em que decisões automatizadas precisam ser auditáveis e explicáveis (Lundberg; Lee, 2017). O método SHAP (*SHapley Additive exPlanations*) oferece uma abordagem unificada para interpretar as previsões de qualquer modelo, baseando-se na teoria dos jogos cooperativos para atribuir valores de importância a cada característica (Lundberg; Lee, 2017).

De forma complementar, o LIME (*Local Interpretable Model-agnostic Explanations*) permite explicar previsões individuais (Ribeiro; Singh; Guestrin, 2016) por meio de modelos simplificados locais.

### 1.10 Redes Neurais Artificiais e Aprendizado Profundo

Redes Neurais Artificiais são modelos de aprendizado de máquina inspirados na estrutura e no funcionamento dos neurônios biológicos, destacando-se por sua versatilidade e escalabilidade no processamento de tarefas complexas. Quando essas arquiteturas são compostas por uma sequência de múltiplas camadas de neurônios ocultas e empilhadas, elas são classificadas como Redes Neurais Profundas. Essa profundidade estrutural é a propriedade que caracteriza o Aprendizado Profundo (*Deep Learning*) no campo mais amplo do Aprendizado de Máquina, e que permite que a rede aprenda representações de dados em níveis progressivos de abstração, sendo capaz de lidar com bilhões de parâmetros para solucionar problemas de alta dimensionalidade (Géron, 2023).

### 1.11 Embedding

Segundo Chollet (2021), *embeddings* capturam relações entre categorias de forma mais sofisticada que métodos de codificação tradicionais, sendo particularmente úteis quando se lida com grandes volumes de dados e alta cardinalidade. Enquanto o *one-hot encoding* gera matrizes excessivamente esparsas que podem exaurir recursos de memória e inviabilizar o processamento de grandes volumes de dados, os *embeddings* mapeiam categorias para um espaço vetorial denso e de baixa dimensão. Além da eficiência, os *embeddings* resolvem o problema de hierarquia introduzido pelo *ordinal encoding,* que ao atribuir números arbitrários a categorias que não possuem ordem natural, acaba por gerar relações lineares inexistentes, prejudicando a convergência do modelo. Ainda que a semântica original dos dados não importe, o *embedding* permite que a rede aprenda uma proximidade estatística baseada no comportamento do alvo. Dessa forma, categorias que geram resultados similares acabam posicionadas próximas no espaço vetorial, facilitando a identificação de padrões que seriam invisíveis em outras codificações.

## 2. METODOLOGIA

A presente pesquisa adotou uma abordagem quantitativa e aplicada, estruturada pelo método de estudo de caso. O caso em análise foi o próprio processo de classificação de etnias e línguas indígenas realizado no Censo Demográfico 2022, cujos dados finais codificados e revisados serviram como a “verdade fundamental” (*ground truth*) para o treinamento e para a validação dos modelos propostos.

## 2.1 Instrumentos e Recursos

Os dados utilizados nesta pesquisa foram extraídos por meio de consultas SQL às bases de dados de coleta e de pós-crítica dispostas em servidores *Microsoft SQL Server* do centro de dados do IBGE, complementados com as tabelas originais e finais dos descritores de etnias e línguas indígenas do Censo Demográfico de 2022.

Os modelos foram desenvolvidos em linguagem Python, utilizando o ambiente de desenvolvimento integrado Visual Studio Code, com apoio das bibliotecas pandas, numpy, scikit-learn, matplotlib, seaborn, SHAP, LIME, tensorflow, keras, entre outras.

Em observância ao sigilo estatístico e à natureza sensível das informações, a pesquisa foi conduzida estritamente no ambiente de infraestrutura do IBGE. O trabalho contou, ainda, com o suporte consultivo de especialistas e etnólogos da instituição.

Em consonância com as boas práticas científicas, a transparência metodológica e visando ao resguardo ético e institucional, declaro que ferramentas de Inteligência Artificial (IA) Generativa foram utilizadas como suporte instrumental ao longo do desenvolvimento deste estudo. A autoria intelectual, a análise crítica e a validação dos resultados permanecem, no entanto, sob inteira responsabilidade do pesquisador.

## 2.2 Disponibilidade de Dados e Código

Os códigos-fonte foram disponibilizados em repositório público do *GitHub* na *URL* https://github.com/luiscesargit/TCC.

Em conformidade com a Lei Geral de Proteção de Dados (LGPD), a Lei do Sigilo Estatístico e com as diretrizes éticas de pesquisa envolvendo populações vulneráveis, a base de dados original e a base sintética contendo sobrenomes reais não foram disponibilizadas publicamente, mitigando o risco de reidentificação.

Para fins de reprodutibilidade metodológica, uma base de amostra fictícia dos microdados foi depositada junto aos descritores no repositório público *Zenodo* na URL https://doi.org/10.5281/zenodo.21009299.

## 2.3 Análise de Conteúdo e Entrevistas não Estruturadas

O objetivo desta etapa foi compreender o processo de codificação manual, conhecer as regras de negócio, identificar padrões de decisão e caracterizar os dados disponíveis. Entrevistas com os especialistas foram conduzidas, além da análise dos documentos produzidos durante o Censo Demográfico 2022, como relatórios sobre critérios de decisão, notas técnicas, definições e alterações de etnônimos. Por meio da análise documental, foi possível identificar as características (*features*) mais promissoras para a utilização nos modelos de predição:

1. Características numéricas: “Latitude”, “Longitude” e “Setor censitário”.
2. Características categóricas: “Texto de entrada da etnia”, “Texto de entrada da língua”, “Etnia após codificação”, “Língua após codificação”, “Código do tipo do setor”, “Código da terra indígena” e “Sobrenome”.

### 2.4 Preparação e Limpeza de Dados

O universo da pesquisa compreendeu a totalidade dos registros de pessoas que se declararam indígenas no Censo 2022 e que responderam a ao menos um quesito de etnia ou língua. Foram utilizados microdados de uso restrito do IBGE, sendo a base principal composta pela junção das informações extraídas das tabelas de coleta, contendo as descrições textuais originais digitadas pelo recenseador, das tabelas com os códigos finais atribuídos após todo o processo de crítica e codificação, além das tabelas descritoras de coleta e pós-crítica de etnias e línguas indígenas. Como os dados de etnia e língua poderiam ser preenchidos a partir da seleção de um código ou da digitação em texto livre, foi aplicada a lógica de que, se o texto livre estivesse vazio, o campo de entrada receberia o texto do descritor do código correspondente.

Dois *datasets* foram gerados, sendo o primeiro para uso no modelo de predição de etnias e o segundo para uso no modelo de predição de línguas. A diferença entre eles é que o primeiro manteve em suas características o texto original da língua indígena declarada, sendo o rótulo desejado o texto do descritor final da etnia. Já o segundo manteve em suas características o texto original da etnia indígena declarada, sendo o rótulo desejado o texto do descritor final da língua.

Ambos os *datasets* passaram por filtragem e descarte dos registros onde não havia informação para o rótulo. Os valores “Não sabe”, “Mal definida” e “Não determinada” foram padronizados para o termo “Não informado”.

As variáveis de coordenadas geográficas (latitude e longitude), originalmente em formato de texto (graus, minutos e segundos - DMS), foram extraídas por meio de expressões regulares e convertidas matematicamente para o formato de graus decimais. Atribuiu-se sinal negativo às coordenadas correspondentes aos quadrantes Sul (S) e Oeste (O).

Por fim, visando à padronização e à mitigação de ruídos causados pela sensibilidade a caracteres maiúsculos e minúsculos (*case sensitivity*), todas as variáveis do tipo *string* (texto) foram convertidas para letras maiúsculas. Os registros passaram, então, pelo processo de deduplicação para a remoção de repetições.

## 3 ANÁLISE DE DADOS E ENGENHARIA DE FEATURES

A análise de dados focou na distribuição de frequência das etnias e línguas e na avaliação de diferentes algoritmos de distância entre cadeias de caracteres para a comparação entre os nomes de etnias e línguas indígenas coletados em campo e as bases oficiais de descritores.

Essa análise resultou nas métricas a seguir, que serviram como *baseline* para a avaliação do desempenho dos modelos desenvolvidos.

**Quadro 1 – Desempenho dos algoritmos de distância de edição**

| | Classificação de Etnias | | | |
|---|---|---|---|---|
| | Acurácia (%) | Precisão (%) | Revocação (%) | F1-Score (%) |
| **Distância de Levenshtein** | 64,38 | 60,51 | 77,32 | 60,72 |
| **Distância de Jaro-Winkler** | 63,61 | 60,06 | 79,12 | 60,27 |
| **Distância de Damerau-Levenshtein** | 64,38 | 60,57 | 77,32 | 60,76 |
| | **Classificação de Línguas** | | | |
| | Acurácia (%) | Precisão (%) | Revocação (%) | F1-Score (%) |
| **Distância de Levenshtein** | 80,42 | 58,70 | 61,21 | 51,85 |
| **Distância de Jaro-Winkler** | 78,81 | 55,70 | 64,84 | 51,45 |
| **Distância de Damerau-Levenshtein** | 80,41 | 58,63 | 61,23 | 51,82 |

**Fonte:** Elaborado pelo autor (2026).

A partir da avaliação desses algoritmos, duas novas características foram adicionadas aos *datasets* de etnias e de línguas indígenas:

1. **Distância de Damerau-Levenshtein entrada-descritor:** corresponde ao valor mínimo da distância de Damerau-Levenshtein calculado em cada registro a partir do texto da etnia ou da língua coletada e das classificações constantes do banco de descritores final.
2. **Descritor mais similar por Damerau-Levenshtein:** corresponde à classe descritora que mais se aproxima do texto da etnia ou da língua coletada, segundo a distância de Damerau-Levenshtein.

A composição dos *datasets* seguiu, portanto, o dicionário de dados a seguir:

**Quadro 2 – Dicionário de dados**

| Variável | Descrição | Dataset Etnias | Dataset Línguas |
|---|---|---|---|
| Texto de entrada da etnia | Texto original da etnia indígena declarada | Característica | - |
| Texto de entrada da língua | Texto original da língua indígena declarada | - | Característica |
| Sobrenome | Sobrenome da pessoa recenseada | Característica | Característica |
| Código do setor | Setor censitário do domicílio | Característica | Característica |
| Código do tipo do setor | Tipo de setor censitário do domicílio | Característica | Característica |
| Código da terra indígena | Código da terra indígena do domicílio | Característica | Característica |
| Latitude | Latitude do domicílio | Característica | Característica |
| Longitude | Longitude do domicílio | Característica | Característica |

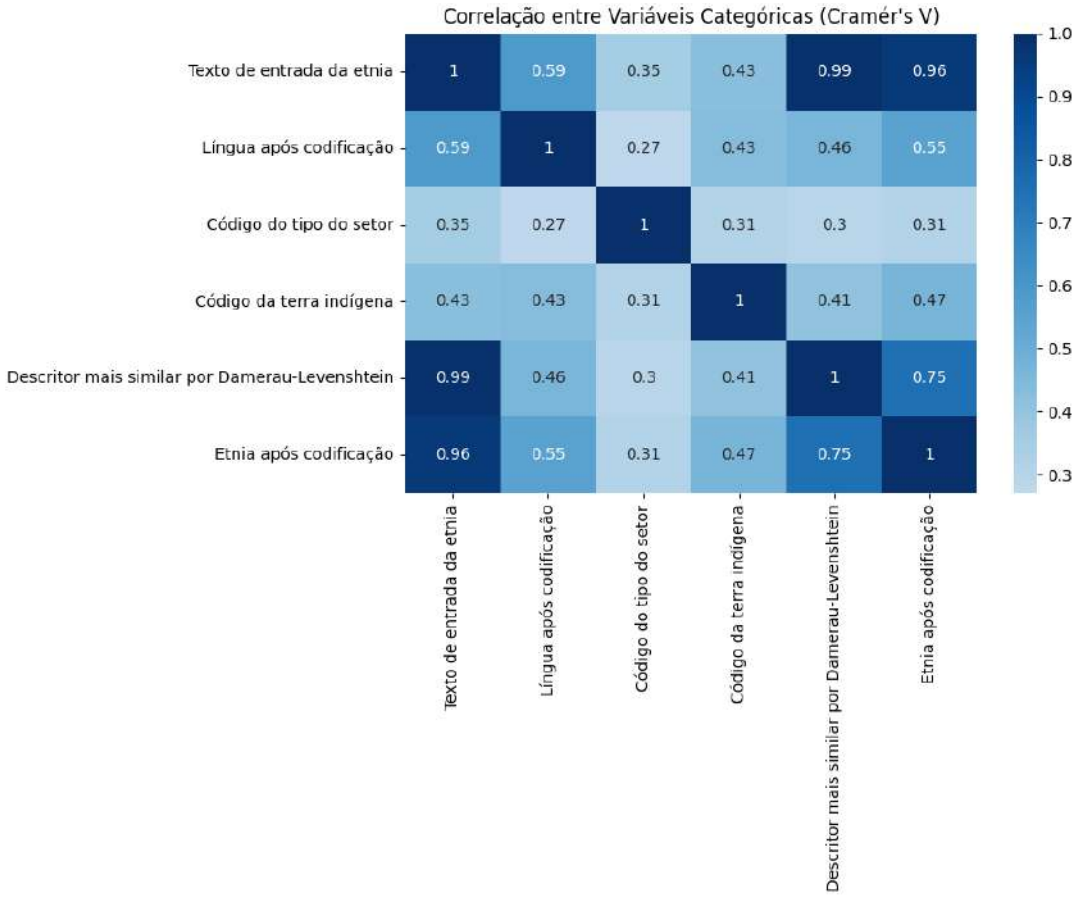
| Variável | Descrição | Dataset Etnias | Dataset Línguas |
|---|---|---|---|
| Distância de Damerau-Levenshtein entrada-descritor | Valor mínimo da distância de Damerau-Levenshtein calculado em cada registro a partir do texto da etnia ou da língua coletada e as classificações constantes do banco de descritores final. | Característica | Característica |
| Descritor mais similar por Damerau-Levenshtein | Classe descritora que mais se aproxima do texto da etnia ou da língua coletada, segundo a distância de Damerau-Levenshtein. | Característica | Característica |
| Etnia após codificação | Texto do descritor da etnia indígena declarada após codificação | Rótulo | Característica |
| Língua após codificação | Texto do descritor da língua indígena declarada após codificação | Característica | Rótulo |

**Fonte:** Elaborado pelo autor (2026).

## 3.1 Estatística Descritiva

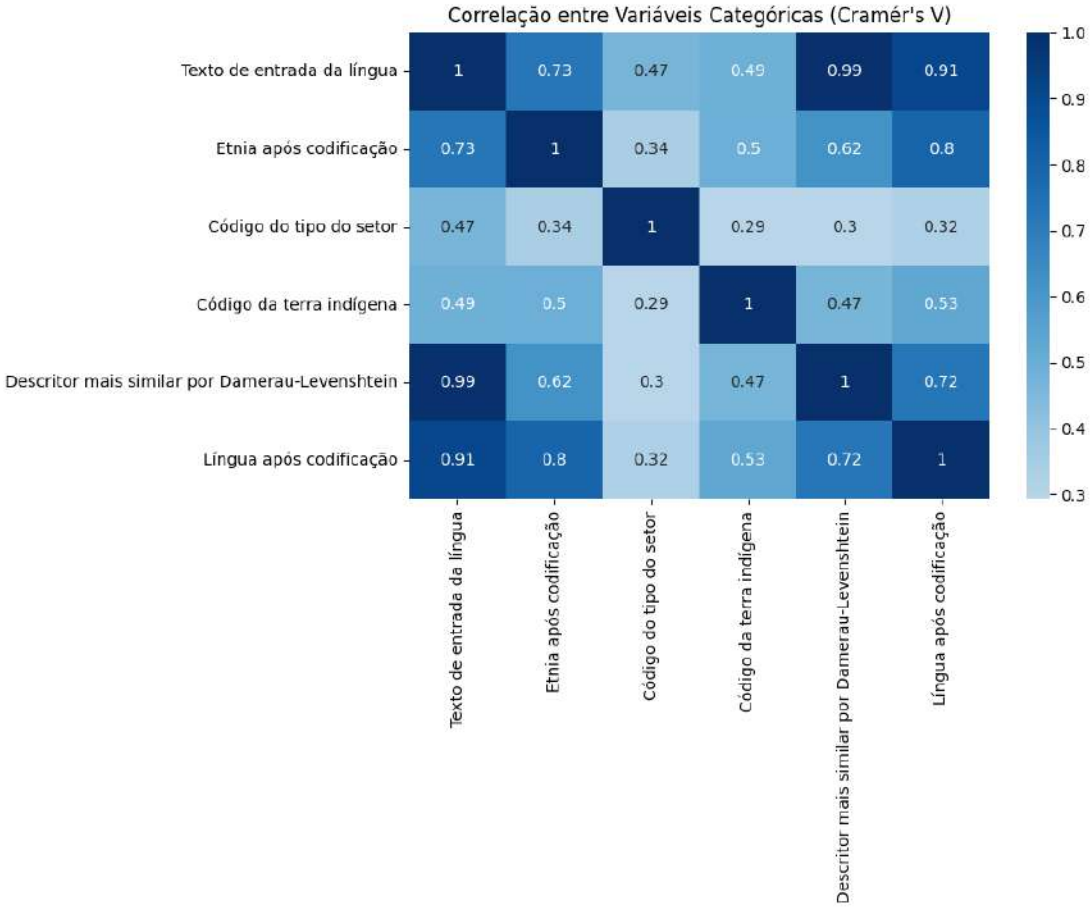
As correlações entre as variáveis categóricas foram investigadas a partir do teste estatístico de Cramér. Os resultados incluíram matrizes de correlação para as variáveis e rótulos categóricos dos *datasets*, além de histogramas de distribuição para as variáveis numéricas.

**Figura 1 – Correlação entre as características categóricas e o rótulo – *dataset* de etnias**



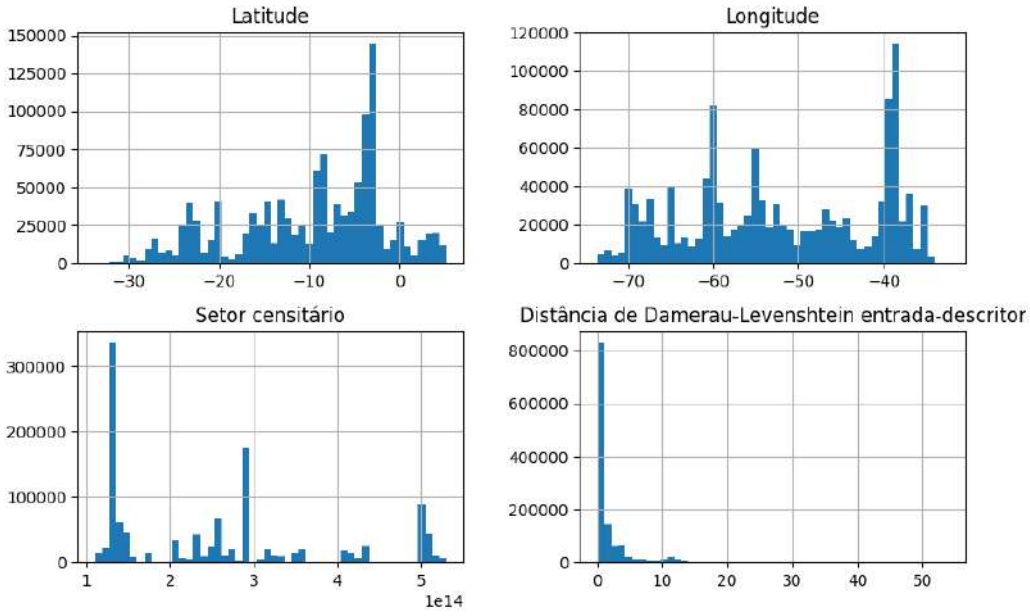

**Fonte:** Elaborada pelo autor (2026).

**Figura 2 — Correlação entre as características categóricas e o rótulo – *dataset* de línguas**



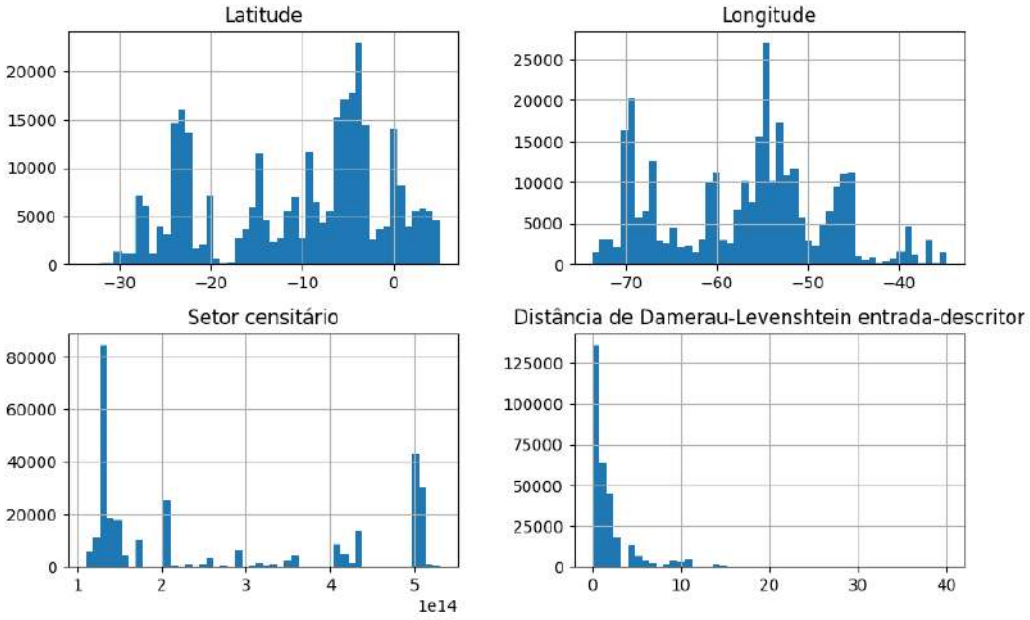

**Fonte:** Elaborada pelo autor (2026).

**Figura 3 — Histogramas de distribuição das características numéricas – *dataset* de etnias**




**Fonte:** Elaborada pelo autor (2026).

**Figura 4 – Histogramas de distribuição das características numéricas – *dataset* de línguas**




**Fonte:** Elaborada pelo autor (2026).

## *3.2 Pré-processamento e Integração em Pipelines*

O pré-processamento foi estruturado para tratar simultaneamente diferentes tipos de dados por meio do *ColumnTransformer*. As transformações foram divididas da seguinte forma:

1. Tratamento de Variáveis Numéricas: As variáveis "Latitude", "Longitude", "Setor censitário" e "Damerau-Levenshtein entrada-descritor" passaram por padronização utilizando o *StandardScaler* de modo a garantir que média fosse zero e variância unitária para cada variável, impedindo, assim, possíveis vieses gerados por variáveis de maior escala.
2. Tratamento de Variáveis Categóricas: As variáveis "Texto de entrada da etnia", "Texto de entrada da língua", "Etnia após codificação", "Língua após codificação", "Descritor mais similar por Damerau-Levenshtein", "Código do tipo do setor", "Código da terra indígena" e "Sobrenome" foram transformadas em representações numéricas utilizando o *one-hot encoder* nos modelos *K-Nearest Neighbors* e *Support Vector Classifier* (SVC) e o *ordinal encoder* nos modelos Árvore de Decisão e Floresta Aleatória. O parâmetro *handle_unknown do ordinal encoder* foi configurado com o valor *"use_encoded_value"*, de modo a assegurar que os modelos não falhassem ao encontrar dados novos não codificados previamente.

O pré-processamento foi encapsulado em objetos *pipeline* onde o fluxo consistiu primeiramente na aplicação das transformações das colunas (*preprocessor*) e, em seguida, na alimentação do estimador (*classifier*), facilitando, dessa forma, a automação de novos modelos de predição.

## 3.3 Treinamento e Avaliação dos Modelos

Os dois *datasets* finais (línguas e etnias) foram particionados de forma estratificada, resultando, cada um, em três subconjuntos independentes que alimentaram os modelos nos processos de treinamento (60% dos dados), validação (20% dos dados) e teste (20% dos dados). Os particionamentos foram adequados de modo a garantir que os *datasets* de treinamento contivessem, ao menos, um registro de cada classe descritora final.

Os *pipelines* de experimentação para os algoritmos *K-Nearest Neighbours*, Árvore de Decisão, Floresta Aleatória e *Support Vector Classifier* foram executados e as métricas de desempenho acurácia, precisão, revocação, F1-*score* calculadas para cada algoritmo, em treinamento e validação, resultando nos seguintes desempenhos:

**Quadro 3 – Desempenho dos modelos com o algoritmo *K-Nearest Neighbours***

| | Classificação de Etnias | | | |
|---|---|---|---|---|
| | Acurácia (%) | Precisão (%) | Revocação (%) | F1-Score (%) |
| Treinamento | 98,81 | 94,64 | 90,15 | 91,86 |
| Validação | 98,45 | 94,17 | 89,53 | 91,10 |

| | Classificação de Línguas | | | |
|---|---|---|---|---|
| | Acurácia (%) | Precisão (%) | Revocação (%) | F1-Score (%) |
| Treinamento | 99,00 | 87,09 | 81,78 | 83,57 |
| Validação | 98,64 | 87,67 | 83,09 | 84,30 |

**Fonte:** Elaborado pelo autor (2026).

**Quadro 4 – Desempenho dos modelos com o algoritmo Árvore de Decisão**

| | Classificação de Etnias | | | |
|---|---|---|---|---|
| | Acurácia (%) | Precisão (%) | Revocação (%) | F1-Score (%) |
| Treinamento | 100,00 | 99,99 | 99,94 | 99,96 |
| Validação | 97,86 | 91,67 | 90,63 | 90,75 |

| | Classificação de Línguas | | | |
|---|---|---|---|---|
| | Acurácia (%) | Precisão (%) | Revocação (%) | F1-Score (%) |
| Treinamento | 100,00 | 99,99 | 99,99 | 99,99 |
| Validação | 98,08 | 84,86 | 83,95 | 83,45 |

**Fonte:** Elaborado pelo autor (2026).

**Quadro 5 – Desempenho dos modelos com o algoritmo Floresta Aleatória**

| | Classificação de Etnias | | | |
|---|---|---|---|---|
| | Acurácia (%) | Precisão (%) | Revocação (%) | F1-Score (%) |
| Treinamento | 100,00 | 99,99 | 99,94 | 99,96 |
| Validação | 98,68 | 96,62 | 91,13 | 93,29 |

| | Classificação de Línguas | | | |
|---|---|---|---|---|
| | Acurácia (%) | Precisão (%) | Revocação (%) | F1-Score (%) |
| Treinamento | 100,00 | 99,99 | 99,99 | 99,99 |
| Validação | 98,96 | 92,76 | 88,80 | 89,95 |

**Fonte:** Elaborado pelo autor (2026).

**Quadro 6 – Desempenho dos modelos com o algoritmo *Support Vector Classifier***

| | Classificação de Etnias | | | |
|---|---|---|---|---|
| | Acurácia (%) | Precisão (%) | Revocação (%) | F1-Score (%) |
| Treinamento | 98,86 | 95,98 | 90,16 | 92,53 |
| Validação | 98,52 | 96,47 | 89,50 | 92,26 |

| | Classificação de Línguas | | | |
|---|---|---|---|---|
| | Acurácia (%) | Precisão (%) | Revocação (%) | F1-Score (%) |
| Treinamento | 99,60 | 89,68 | 86,42 | 87,70 |
| Validação | 99,14 | 89,56 | 85,38 | 86,77 |

**Fonte:** Elaborado pelo autor (2026).

## *3.4 Modelagem e Calibração (Tuning) de Hiperparâmetros*

Conforme as métricas obtidas em treinamento e validação, verificou-se que o modelo de melhor desempenho foi o do algoritmo Floresta Aleatória. Para encontrar a melhor configuração desse modelo, foi utilizado o *RandomizedSearchCV*, que testou valores distintos para os parâmetros "número de árvores", "profundidade máxima", "divisão de nós" e "utilização de *bootstrap*", por meio de uma validação cruzada com 5 dobras (*5-fold cross-validation*).

Os melhores parâmetros identificados foram:

**Quadro 7 – Parâmetros sugeridos pelo *RandomizedSearchCV***

| | Classificação de Etnias | Classificação de Línguas |
|---|---|---|
| **n_estimators** | 50 | 200 |
| **min_samples_split** | 5 | 2 |
| **max_depth** | None | 30,00 |
| **bootstrap** | False | False |

**Fonte:** Elaborado pelo autor (2026).

Finalmente, após a calibração dos parâmetros, os modelos passaram novamente pelas etapas de validação e testes, apresentando, então, os resultados, a seguir.

**Quadro 8 – Desempenho dos modelos com o algoritmo Floresta Aleatória após a calibração dos parâmetros**

| | Classificação de Etnias | | | |
|---|---|---|---|---|
| | **Acurácia (%)** | **Precisão (%)** | **Revocação (%)** | **F1-Score (%)** |
| **Treinamento** | 100,00 | 99,99 | 99,94 | 99,96 |
| **Validação** | 98,68 | 96,62 | 91,13 | 93,29 |
| **Validação após calibração** | 98,71 | 95,99 | 90,68 | 92,81 |
| **Teste** | 98,70 | 96,24 | 91,63 | 93,36 |

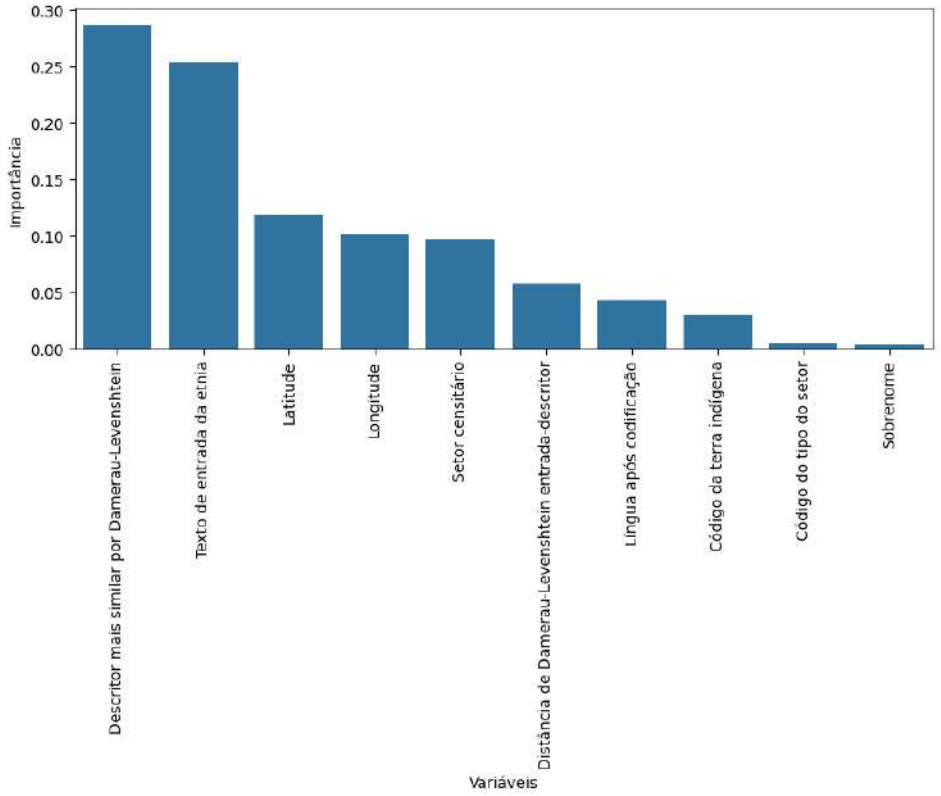
| | Classificação de Línguas | | | |
|---|---|---|---|---|
| | **Acurácia (%)** | **Precisão (%)** | **Revocação (%)** | **F1-Score (%)** |
| **Treinamento** | 100,00 | 99,99 | 99,99 | 99,99 |
| **Validação** | 98,96 | 92,76 | 88,80 | 89,95 |
| **Validação após calibração** | 99,12 | 94,42 | 90,38 | 91,56 |
| **Teste** | 99,01 | 91,47 | 89,28 | 89,88 |

**Fonte:** Elaborado pelo autor (2026).

## *3.5 Análise de Importância de Variáveis (Feature Importance)*

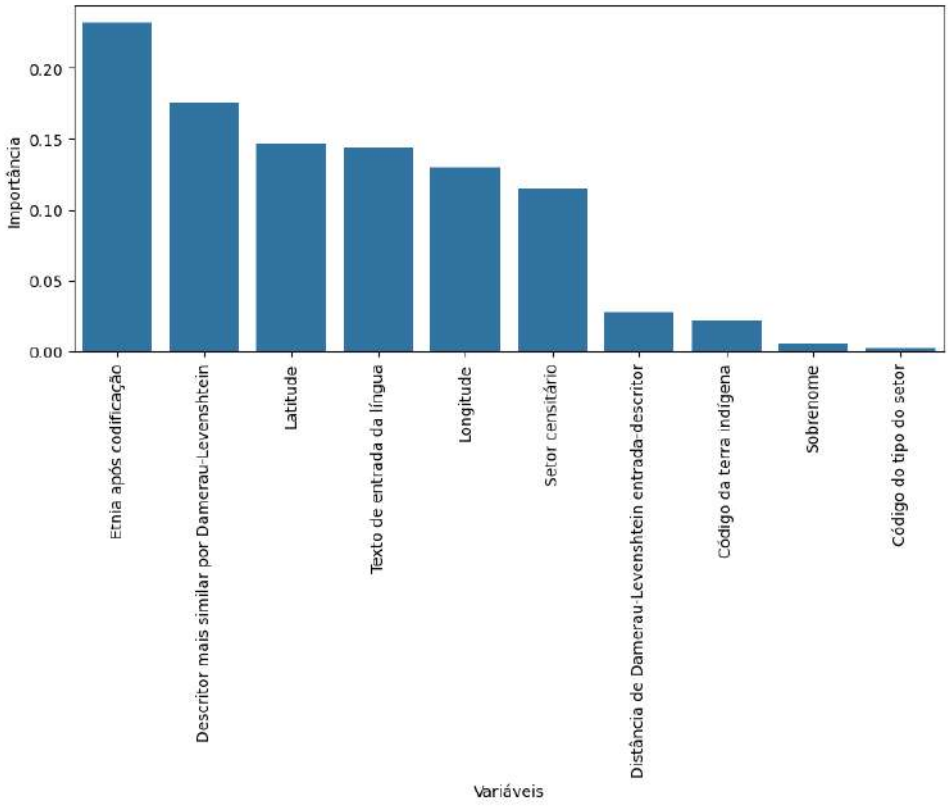
Nesta etapa, foram extraídas as importâncias globais de cada atributo para os modelos de classificação. O algoritmo calculou o quanto cada variável contribuiu para os fatores de decisão dos modelos. Curiosamente, os *rankings* de importância do modelo para classificação de etnias e do modelo para classificação de línguas foram bem distintos, conforme pode ser observado, a seguir.

**Figura 5 – Importância das variáveis – Classificação de Etnias com Floresta Aleatória**




**Fonte:** Elaborada pelo autor (2026).

**Figura 6 – Importância das variáveis – Classificação de Línguas com Floresta Aleatória**

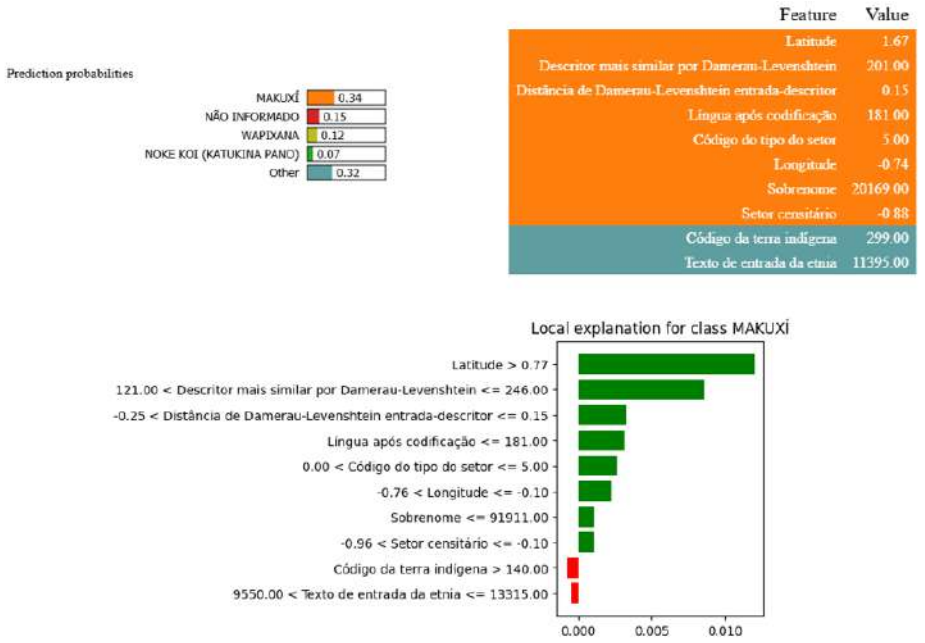



**Fonte:** Elaborada pelo autor (2026).

## 3.6 Interpretabilidade Local com LIME

O método LIME (*Local Interpretable Model-agnostic Explanations*) é aplicado para validar predições de forma individual. A abordagem cria modelos locais simplificados em torno de uma instância específica para explicar o comportamento do classificador. A partir de uma instância aleatória (5585), o modelo de etnias foi testado com um registro referente à entrada "MACUXI", classificada corretamente como "MAKUXÍ", com 100% de confiança. O resultado do LIME apresentou, assim, os principais impulsionadores para essa classificação específica.

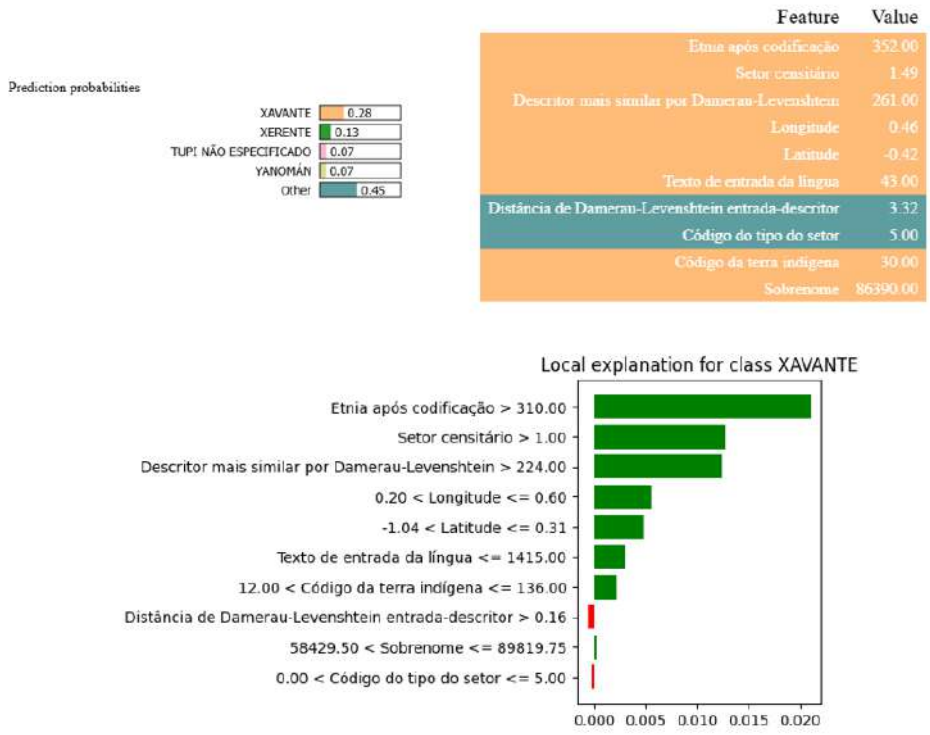
**Figura 7 — Interpretabilidade local LIME – Classificação de Etnias – Instância 5585**




**Fonte:** Elaborada pelo autor (2026).

A técnica foi aplicada, novamente, em uma instância aleatória (5585) referente ao registro de entrada de língua "A'UWE ( XAVANTE )". A língua foi classificada corretamente como "XAVANTE", com 100% de confiança. O LIME apresentou, também, os principais impulsionadores para a classificação.

**Figura 8 — Interpretabilidade local LIME – Classificação de Línguas – Instância 5585**

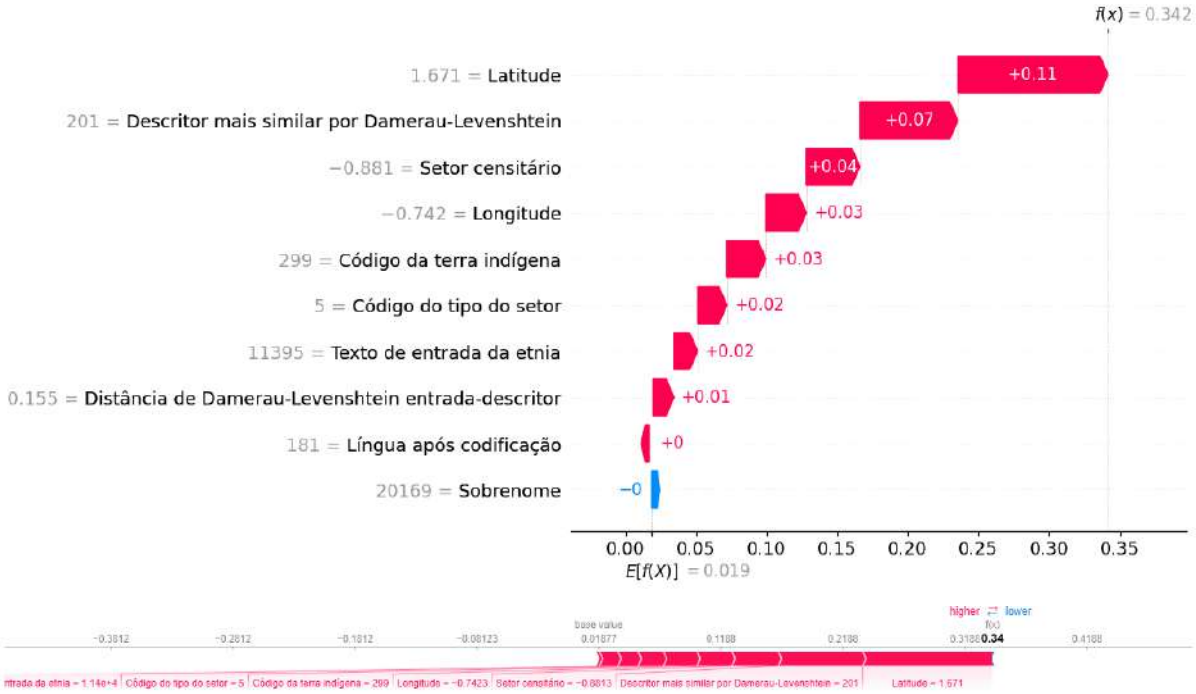



**Fonte:** Elaborada pelo autor (2026).

## *3.7 Interpretabilidade Local e Global com SHAP*

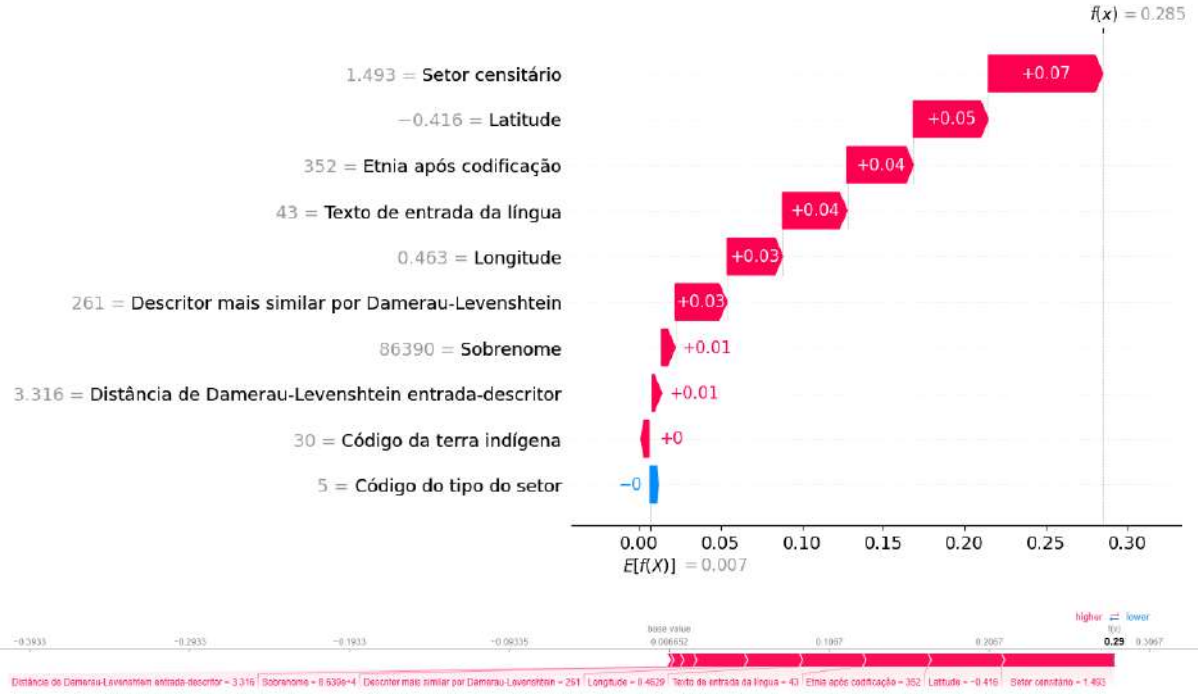
Em complementação aos resultados do LIME, a abordagem SHAP (*SHapley Additive exPlanations*) foi utilizada na mesma instância 5585 nos modelos de classificação de línguas e etnias. Foram gerados, então, gráficos de contribuição onde se observou o impacto aditivo de cada variável na função de probabilidade da classe vencedora. A análise SHAP confirmou os achados do LIME, evidenciando como a combinação das variáveis desloca a probabilidade da predição a partir de um valor base até a decisão final do modelo.

## Figura 9 — Interpretabilidade local SHAP – Gráficos de Força e Cascata – Classificação de Etnias – Instância 5585




**Fonte:** Elaborada pelo autor (2026).

## Figura 10 — Interpretabilidade local SHAP – Gráficos de Força e Cascata – Classificação de Línguas – Instância 5585



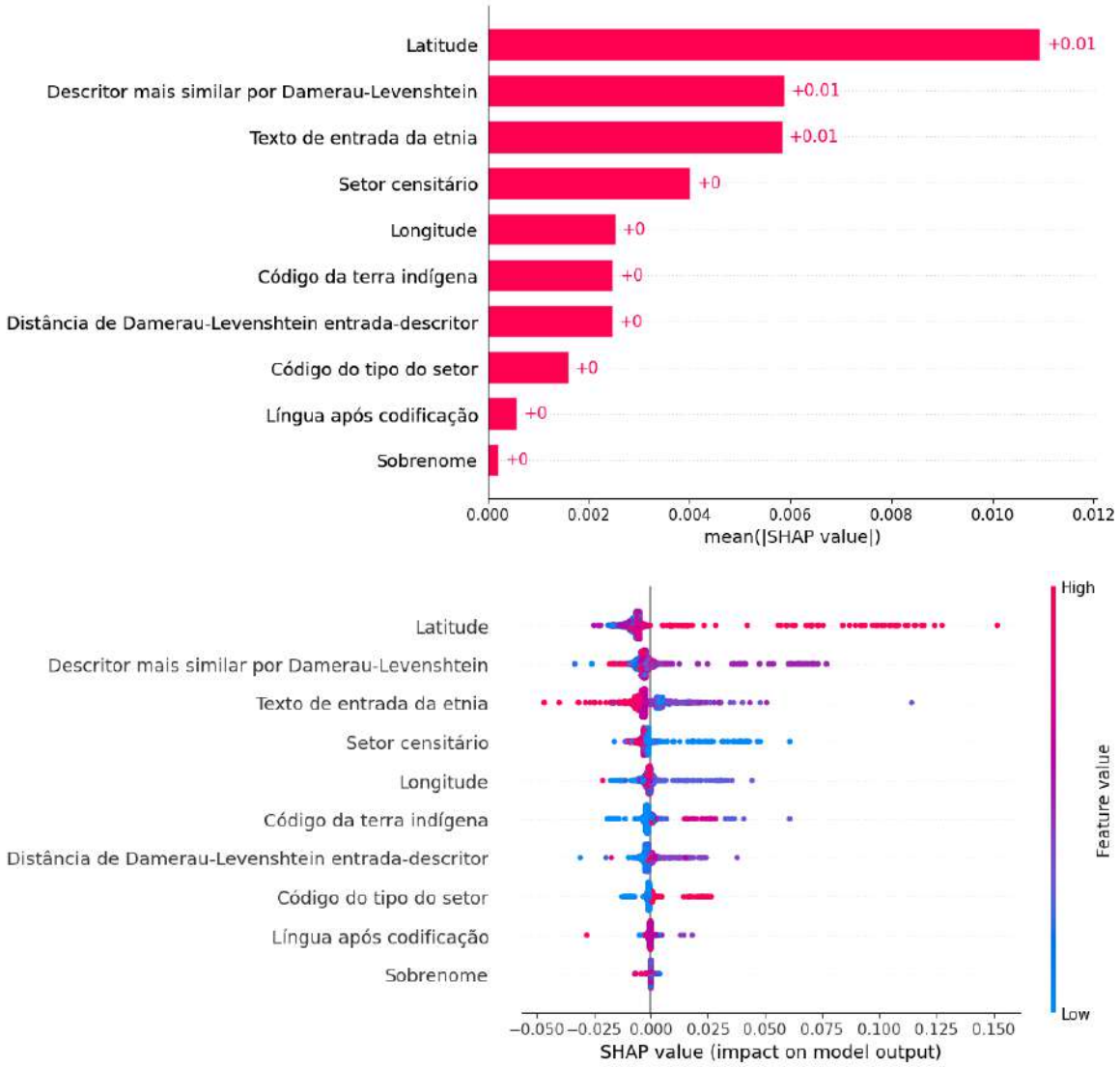

**Fonte:** Elaborada pelo autor (2026).

Para a avaliação global, foi selecionada uma amostra aleatória de 1000 observações de modo a reduzir o custo computacional exigido em modelos complexos como o Floresta Aleatória. No modelo de etnias, a classe vencedora “MAKUXÍ” foi utilizada como referência e no modelo de línguas, a análise concentrou-se na classe de maior probabilidade “XAVANTE”.

Foram gerados, então, gráficos ordenados pelo impacto médio de cada característica na previsão da classificação, de modo a demonstrar como seus valores mais altos ou mais baixos influenciaram a probabilidade de predição da classe correta.

Os rankings apresentam as médias dos impactos das instâncias de cada variável e os gráficos de enxame (*beeswarm plot*) apresentam a concentração do impacto de acordo com o valor de cada instância, que é representada por um ponto colorido conforme a escala, em cada linha de característica.

**Figura 11 — Interpretabilidade global SHAP – Ranking e Gráfico de Enxame – Classificação de Etnias – Classe “MAKUXÍ”**




**Fonte:** Elaborada pelo autor (2026).

## 3.8 Validação da pesquisa

Simultaneamente à validação e calibração dos modelos, foram disponibilizados aos especialistas etnólogos e aos analistas da Gerência de Povos e Comunidades Tradicionais do IBGE relatórios para análise de como as predições dos modelos de classificação estavam se comportando e para apoio ao processo de identificação de padrões de erros que poderiam ser objeto de ajuste.

As seguintes listagens foram extraídas em cada rodada desse processo:

1. Predições x Real – *dataset* completo de etnias.
2. Predições x Real – *dataset* completo de línguas.
3. Erros de predição na classificação de etnias.
4. Erros de predição na classificação de línguas.
5. Erros de predição na classificação de etnias onde o rótulo real é igual a "NÃO INFORMADO".
6. Erros de predição na classificação de línguas onde o rótulo real é igual a "NÃO INFORMADO".
7. Erros de predição na classificação de etnias onde o rótulo predito é igual a "NÃO INFORMADO".
8. Erros de predição na classificação de línguas onde o rótulo predito é igual a "NÃO INFORMADO".
9. Erros de predição na classificação de etnias desconsiderando casos em que o rótulo predito ou o rótulo real é igual a "NÃO INFORMADO".
10. Erros de predição na classificação de línguas desconsiderando casos em que o rótulo predito ou o rótulo real é igual a "NÃO INFORMADO".
11. Quantidades de erros de predição na classificação de etnias, agrupadas pelo texto da etnia coletada, pelo descritor com texto mais próximo, pela etnia predita e pela etnia real.
12. Quantidades de erros de predição na classificação de línguas, agrupadas pelo texto da língua coletada, pelo descritor com texto mais próximo, pela língua predita e pela língua real.

Essa dinâmica de revisão periódica dos resultados possibilitou a identificação de omissões e de oportunidades de otimização dos modelos. No último ciclo de avaliação dos relatórios, verificou-se que alguns registros classificados originalmente como "NÃO INFORMADO" teriam, hoje, alta probabilidade de classificação definida.

Foi avaliado, também, que os registros onde os valores de confiança apresentaram valores mais baixos condizem com registros atípicos, contribuindo para que seja estipulado um valor de corte que maximize a classificação automática e minimize a supervisão manual futuramente.

Ainda que as análises requeiram maior aprofundamento e um estudo espacializado, o retorno da área especialista é de que os modelos generalizaram de forma satisfatória, tanto para etnias, quanto para línguas indígenas.

## 3.9 Experimentos com Redes Neurais Profundas

Além dos algoritmos apresentados, foi proposta a experimentação de redes neurais profundas para a avaliação da possibilidade de alcance de métricas ainda superiores às do modelo Floresta Aleatória. Dessa forma, foram desenvolvidos modelos com arquitetura de rede neural do tipo sequencial, com foco na transformação de entradas categóricas em representações vetoriais densas.

Os modelos foram desenvolvidos a partir de uma camada de *embedding* de 64 dimensões, que buscou aprender relações entre as variáveis codificadas, seguida por uma camada de *flatten* para a conversão desses mapas de características em um vetor único.

O processamento profundo foi realizado por duas camadas densas com 1024 e 512 neurônios, respectivamente, ambas utilizando a função de ativação *ReLU*. De modo a estabilizar o treinamento, a função de *Batch Normalization* foi adicionada após cada camada densa, além de camadas de *dropout* com taxas de 0.3 e 0.2 para mitigar o risco de sobreajuste.

A saída final consistiu em uma camada densa com 311 neurônios no modelo de classificação de línguas e de 396 neurônios no modelo de classificação de etnias. A função de ativação para ambos foi a Softmax, que forneceu a probabilidade de cada amostra pertencer a uma das línguas ou etnias mapeadas.

No que diz respeito ao treinamento, os modelos foram configurados com o otimizador *Adam* e a função de perda *sparse_categorical_crossentropy*. De modo a compensar o forte desequilíbrio entre línguas ou etnias muito frequentes e aquelas com poucos registros, optou-se pela utilização de pesos de classe (*class_weight*). O treinamento contou, também, com a técnica de *early stopping*, monitorando a perda de validação para interromper o processo após 10 épocas que não resultassem em otimização da função de perda.

As redes neurais profundas apresentaram desempenhos bastante satisfatórios, em especial a do modelo de classificação de línguas, que obteve desempenho superior ao modelo com o algoritmo Floresta Aleatória.

Cabe, ainda, a execução aprofundada de ajustes de hiperparâmetros com a pretensão de que os modelos de redes neurais profundas superem com margem mais significativa os modelos treinados com o algoritmo Floresta Aleatória.

**Quadro 9 – Desempenho dos modelos com Redes Neurais Profundas**

| | Classificação de Etnias | | | |
|---|---|---|---|---|
| | **Acurácia (%)** | **Precisão (%)** | **Revocação (%)** | **F1-Score (%)** |
| **Treinamento** | 98,10 | 89,98 | 97,33 | 92,77 |
| **Validação** | 97,01 | 85,70 | 94,92 | 89,04 |
| **Teste** | 97,03 | 85,12 | 94,37 | 88,44 |

| | Classificação de Línguas | | | |
|---|---|---|---|---|
| | **Acurácia (%)** | **Precisão (%)** | **Revocação (%)** | **F1-Score (%)** |
| **Treinamento** | 99,79 | 96,08 | 96,30 | 96,07 |
| **Validação** | 99,38 | 91,82 | 91,68 | 91,31 |
| **Teste** | 99,35 | 93,92 | 94,91 | 94,04 |

**Fonte:** Elaborado pelo autor (2026).

## CONCLUSÃO

Este estudo buscou avaliar até que ponto modelos de Aprendizado de Máquina treinados com o histórico dos dados criticados, codificados e validados podem reproduzir a classificação de etnias e línguas indígenas realizada no Censo Demográfico 2022.

A partir de técnicas de Aprendizado de Máquina, combinadas a algoritmos de medidas de similaridade entre cadeias de caracteres, foi possível uma composição de abordagem que apresentou bons resultados em todos os modelos treinados.

**Quadro 10 – Comparação do desempenho entre os modelos de distância de edição, os resultados de teste dos modelos Floresta Aleatória após a calibração dos parâmetros e os modelos de Redes Neurais Profundas**

| | Classificação de Etnias | | | |
|---|---|---|---|---|
| | Acurácia (%) | Precisão (%) | Revocação (%) | F1-Score (%) |
| Distância de Damerau-Levenshtein | 64,38 | 60,57 | 77,32 | 60,76 |
| Floresta Aleatória após calibração | 98,70 | 96,24 | 91,63 | 93,36 |
| Rede Neural Profunda | 97,03 | 85,12 | 94,37 | 88,44 |

| | Classificação de Línguas | | | |
|---|---|---|---|---|
| | Acurácia (%) | Precisão (%) | Revocação (%) | F1-Score (%) |
| Distância de Damerau-Levenshtein | 80,41 | 58,63 | 61,23 | 51,82 |
| Floresta Aleatória após calibração | 99,01 | 91,47 | 89,28 | 89,88 |
| Rede Neural Profunda | 99,35 | 93,92 | 94,91 | 94,04 |

**Fonte:** Elaborado pelo autor (2026).

Quando comparados à simples busca pela classe de maior similaridade entre os descritores por meio de algoritmos de distância de edição entre cadeias de caracteres, os modelos de Aprendizado de Máquina, com a incorporação das características de contexto, apresentaram desempenho muito superior em todas as métricas.

A convergência entre os resultados obtidos, somada à revisão técnica produzida pelos especialistas, evidencia a consistência dos modelos e das predições obtidas, sustentando, assim, a viabilidade de sua integração à atividade de supervisão com foco em maior eficiência operacional do processo de codificação nas operações censitárias futuras.

## REFERÊNCIAS


BREIMAN, L. **Random forests**. Machine Learning, [s. l.], v. 45, n. 1, p. 5-32, 2001.

CORTES, Corinna; VAPNIK, Vladimir. **Support-vector networks**. Machine Learning, [s. l.], v. 20, n. 3, p. 273-297, 1995.

CHOLLET, François. **Deep learning with Python**. 2. ed. Shelter Island: Manning Publications, 2021.

DAMERAU, F. J. A technique for computer detection and correction of spelling errors. **Communications of the ACM**, New York, v. 7, n. 3, p. 171-176, mar. 1964.

FIX, Evelyn; HODGES, Joseph Lawson. **Discriminatory Analysis**: Nonparametric Discrimination: Consistency Properties. Randolph Field, Texas: USAF School of Aviation Medicine, 1951. (Project 21-49-004, Report 4).

GÉRON, Aurélien. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**: concepts, tools, and techniques to build intelligent systems. 3. ed. Sebastopol: O'Reilly Media, 2023.

IBGE. **Censo Demográfico 2022**: etnias e línguas indígenas: principais características sociodemográficas: resultados do universo. Rio de Janeiro: IBGE, 2025. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv102223.pdf. Acesso em: 12 jan. 2025.

JARO, Matthew A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. **Journal of the American Statistical Association**, [s. l.], v. 84, n. 406, p. 414-420, jun. 1989.

LEVENSHTEIN, Vladimir I. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, [s. l.], v. 10, n. 8, p. 707-710, fev. 1966.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. **Advances in Neural Information Processing Systems**, [S. l.], v. 30, p. 4765-4774, 2017. Disponível em: https://github.com/shap/shap. Acesso em: 12 jan. 2025.

MITCHELL, Tom M. **Machine Learning**. New York: McGraw-Hill, 1997.

OLIVEIRA, Luís César Seixas de. **Arquivos descritores e amostra sintética e anonimizada de microdados**. Zenodo, 2026. Conjunto de dados. DOI: 10.5281/zenodo.21009299. Disponível em: https://doi.org/10.5281/zenodo.21009299. Acesso em: 28 jun. 2026.

OLIVEIRA, Luís César Seixas de. **TCC-Classificação de Línguas e Etnias Indígenas nos Censos Demográfico e Agropecuário por meio de Machine Learning**: Repositório contendo os códigos do Trabalho de Conclusão de Curso. GitHub, 2026. Disponível em: https://github.com/luiscesargit/TCC. Acesso em: 28 jun. 2026.

QUINLAN, J. R. **Induction of decision trees**. Machine Learning, [s. l.], v. 1, n. 1, p. 81-106, 1986.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why Should I Trust You?": explaining the predictions of any classifier. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, [S. l.], p. 1135–1144, 2016. Disponível em: https://github.com/marcotcr/lime. Acesso em: 12 jan. 2025.

SAMUEL, Arthur L. Some Studies in Machine Learning Using the Game of Checkers. **IBM Journal of Research and Development**, [S. l.], v. 3, n. 3, p. 210-229, 1959.

WINKLER, William E. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Washington, D.C.: U.S. Bureau of the Census, 1990.

# Estimativa de Propensão e Exclusão no Crédito Rural: uma abordagem de positive-unlabeled learning e gradient boosting com dados do PRONAF e CAF

*Marcelo Cabreira Bastos*
*Prof. Dr. Hélio Bonfim de Macêdo Filho*

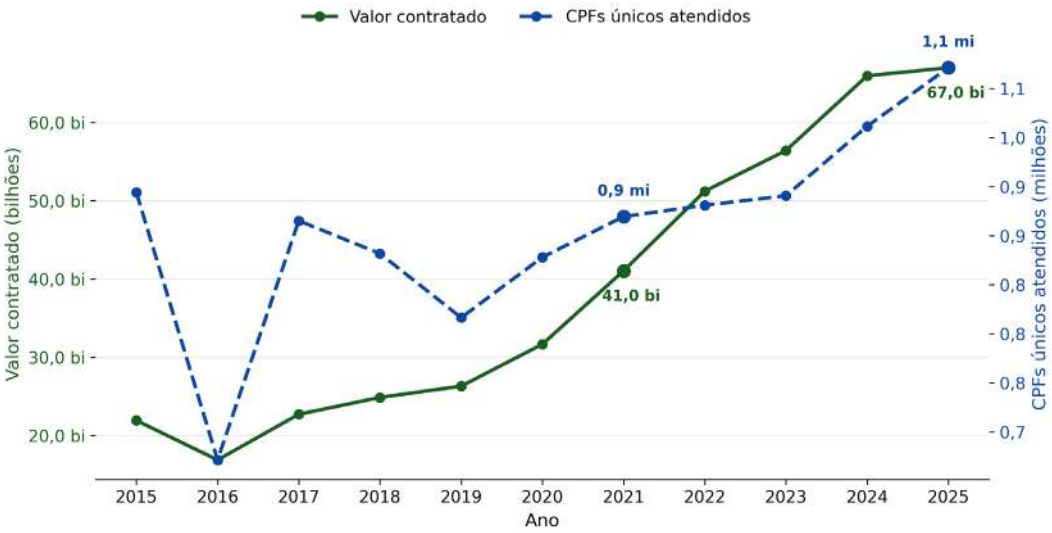
## Resumo

O Programa Nacional de Fortalecimento da Agricultura Familiar (PRONAF) enfrenta um desafio de focalização: a ausência de acesso observado ao crédito não equivale à ausência de propensão. As bases administrativas registram quem contratou, mas não as propostas recusadas ou as demandas não formalizadas, de modo que o conjunto sem acesso observado mistura, de forma indistinta, agricultores sem perfil compatível e *positivos ocultos* — agricultores elegíveis mantidos à margem por barreiras. Esse cenário de rotulagem parcial, em que a classe latente de interesse não é observável, é formulado como um problema de *Positive-Unlabeled* (PU) *Learning*. Integrando os microdados do Cadastro Nacional da Agricultura Familiar (CAF) e do PRONAF, o estudo compara três estratégias — um *baseline* supervisionado, o estimador de Elkan e Noto e o nnPU — para estimar a propensão de acesso e ordenar o universo sem acesso observado. Os resultados revelam um universo heterogêneo, com convergência entre as abordagens na identificação de um núcleo consistente de casos prioritários. Os escores constituem um ranking diagnóstico — e não uma classificação definitiva de elegibilidade —, útil ao monitoramento da cobertura do programa e a estratégias de busca ativa.

**Palavras-chave:** PRONAF. Crédito rural. *Positive-Unlabeled Learning*. Políticas públicas. Dados administrativos.

## INTRODUÇÃO

O Programa Nacional de Fortalecimento da Agricultura Familiar (PRONAF), criado em 1995, constitui a principal política pública de crédito direcionada às Unidades Familiares de Produção Agrária (UFPAs) no Brasil. Ao prover capital para modernização tecnológica e aumento da produtividade, o programa atua como vetor de desenvolvimento rural e segurança alimentar (Araujo et al. 2020; Souza et al. 2015). Em 2025, o PRONAF atingiu R$ 66,5 bilhões concedidos em mais de 2,07 milhões de operações de crédito (BANCO CENTRAL DO BRASIL 2026) — trajetória de expansão ilustrada na Figura 1. Essa escala revela tanto o alcance da política quanto a dimensão do desafio de focalização: garantir que os recursos alcancem o segmento mais vulnerável e com maior potencial de resposta à intervenção estatal (Rodrigues 2019; Silva 2012).




**Fonte:** Banco Central do Brasil (BANCO CENTRAL DO BRASIL 2026). Elaboração própria.

Para viabilizar o acesso ao crédito e aprimorar a focalização da política, o Cadastro Nacional da Agricultura Familiar (CAF) funciona como o instrumento oficial de identificação das UFPAs elegíveis, substituindo a antiga Declaração de Aptidão ao PRONAF (DAP) (Brasil 2025). O CAF consolida os dados de todos os agricultores formalmente cadastrados; o Sistema de Operações do Crédito Rural e do Proagro (SICOR) registra apenas as contratações efetivamente realizadas. Essa distinção é estruturalmente relevante: a presença no CAF não implica contratação, e a ausência no SICOR não implica inelegibilidade. Apenas uma fração das UFPAs cadastradas efetiva a contratação do crédito, enquanto um contingente expressivo permanece à margem do sistema formal. Essa exclusão não decorre necessariamente de incapacidade ou desinteresse: barreiras institucionais, territoriais e informacionais (a exemplo de exigências burocráticas, distância de agências bancárias, escassez de assistência técnica e extensão rural - ATER) atuam como filtros excludentes que afastam do programa agricultores formalmente elegíveis (Souza et al. 2015), com histórico de contratações concentradas nas regiões Sul e Sudeste em detrimento do Norte e Nordeste (Aquino et al. 2018).

A mensuração empírica desse fenômeno esbarra em um obstáculo estrutural dos dados governamentais. Os registros do PRONAF documentam apenas as contratações realizadas; não existe registro sistemático de propostas recusadas, demandas não formalizadas ou desistências no processo. Sabe-se quem acessou o crédito — os **positivos observados** ($S = 1$) —, mas não quem tentou e não conseguiu. Os demais registros do CAF formam um **conjunto não rotulado** ($S = 0$) intrinsecamente heterogêneo: reúnem, de forma indistinta, agricultores com perfil estruturalmente incompatível com o programa, agricultores sem demanda pelo crédito e agricultores com perfil compatível ao dos contratantes que permaneceram à margem por razões institucionais ou territoriais — os **positivos ocultos** (Bekker and Davis 2020). Essa heterogeneidade impossibilita tratar todo caso sem acesso observado como negativo. Métodos supervisionados tradicionais aplicados diretamente incorporariam os vieses de seleção dos dados administrativos, interpretando a não observação como evidência de ausência de perfil e obscurecendo a distinção substantiva entre exclusão estrutural e demanda reprimida por crédito (Bekker and Davis 2020; Elkan and Noto 2008).

Para lidar com esse cenário de rotulagem parcial, em que apenas uma fração dos positivos reais é observada enquanto os demais casos permanecem sem rótulo e podem pertencer a qualquer classe, este estudo adota o arcabouço de *Positive-Unlabeled Learning* (PU Learning) (Bekker and Davis 2020; Elkan and Noto 2008). Sob a hipótese de rotulagem *Selected Completely at Random* (SCAR), em que os positivos rotulados constituem uma amostra aleatória dos positivos reais, a propensão latente de acesso ao crédito pode ser recuperada a partir da probabilidade observada de contratação, sendo a constante de rotulagem interpretável como a taxa de cobertura efetiva do PRONAF. São comparadas três estratégias de modelagem: um modelo *baseline* supervisionado (referência comparativa); o estimador de Elkan e Noto (Elkan and Noto 2008), que corrige a probabilidade observada pela constante de rotulagem; e o estimador nnPU (Kiryo et al. 2017), que reformula o risco de aprendizado diretamente a partir de positivos e não rotulados. O classificador de base do *baseline* e da correção de Elkan e Noto é um modelo de *Gradient Boosting* (Chen and Guestrin 2016), escolhido pela capacidade de capturar relações não lineares em dados tabulares; o nnPU é implementado com uma rede neural.

Este estudo integra os microdados do CAF e do SICOR/PRONAF para estimar a propensão de acesso ao crédito rural entre as UFPAs ativas no cadastro, comparando as três estratégias de modelagem e produzindo um *ranking* diagnóstico. A contribuição empírica está na identificação de dois estratos analiticamente distintos: o de **alta propensão estimada sem acesso observado** composto por perfil semelhante ao dos contratantes, compatível com a presença de barreiras e prioritário para estratégias de busca ativa, e o de **baixa propensão estimada sem acesso observado**, cuja heterogeneidade estrutural demanda instrumentos distintos de política pública. O ranking é um instrumento diagnóstico e de priorização; não propõe decisão automatizada de concessão individual.

O artigo está organizado em quatro seções adicionais: a Seção 2 fundamenta o problema de rotulagem parcial e as estratégias de modelagem; a Seção 3 apresenta as bases, a população e os procedimentos analíticos; a Seção 4 expõe os resultados empíricos; e a Seção 5 discute as implicações e apresenta as conclusões.

# 1. REFERENCIAL E ESTRATÉGIA ANALÍTICA

## 1.1 Rotulagem parcial e identificação sob SCAR

O problema central deste estudo é de rotulagem parcial: observam-se apenas positivos rotulados e uma grande massa sem rótulo. A geração dos dados envolve três variáveis: (1) o vetor de atributos $X$; (2) a **classe latente** $Y \in \{0, 1\}$, que denota a adequação real ao crédito e não é observável nas bases administrativas; (3) e a variável de rotulagem observada $S \in \{0, 1\}$, segundo a estrutura hierárquica $X \rightarrow Y \rightarrow S$ (Bekker and Davis 2020; Elkan and Noto 2008). Nos registros do CAF integrados ao PRONAF, observa-se apenas o par $(X, S)$: $S = 1$ identifica os positivos observados, com acesso registrado, e $S = 0$ reúne o **conjunto não rotulado**.

A formulação assume que o mecanismo de rotulagem não produz falsos positivos, isto é, $P(S = 1 \mid Y = 0, X) = 0$: apenas agricultores estruturalmente elegíveis são registrados como beneficiários. No PRONAF, isso equivale a supor que os critérios de enquadramento e a análise de conformidade bancária são suficientemente rigorosos para tornar desprezível a concessão a perfis de agricultores que seriam inelegíveis. Sob essa hipótese, $S = 0$ não indica inelegibilidade: agrega negativos reais e os **positivos ocultos**, ou seja, agricultores com perfil compatível que não acessam o crédito por barreiras institucionais, informacionais ou territoriais.

Para recuperar $Y$ a partir de $S$, adota-se a hipótese *Selected Completely at Random* (SCAR), segundo a qual a probabilidade de um positivo ser rotulado é constante e independe dos atributos, $P(S = 1 \mid Y = 1, X) = c$, onde $c$ é a frequência de rotulagem entre os positivos. Disso decorre a relação fundamental do PU Learning clássico:

$$P(S = 1 \mid X) = c \cdot P(Y = 1 \mid X) \leftrightarrow P(Y = 1 \mid X) = \frac{P(S = 1 \mid X)}{c}.$$

Como $c$ é uma constante positiva, a probabilidade observada é proporcional à latente e o ordenamento dos agricultores por $P(S = 1 \mid X)$ preserva o ordenamento por $P(Y = 1 \mid X)$. Essa propriedade sustenta o uso dos escores como ranking diagnóstico. A hipótese SCAR é reconhecidamente forte para uma política como o PRONAF, em que o acesso ao crédito dificilmente é homogêneo entre perfis distintos; mantém-se, contudo, como base canônica da formulação clássica, remetendo mecanismos de rotulagem mais flexíveis à agenda futura.

## 1.2 Estratégia de modelagem

A comparação empírica contrasta três estratégias. A primeira é um *baseline* supervisionado, que treina um classificador sobre $S$ tratando todo o conjunto não rotulado como negativo. Por incorporar diretamente o viés de seleção dos dados administrativos, esse modelo subestima a propensão dos positivos ocultos; serve como referência comparativa, como o ponto de partida que as abordagens PU buscam superar.

A segunda é a correção clássica de Elkan e Noto (Elkan and Noto 2008). Em vez de tratar $S$ como classe verdadeira, ela estima $P(S = 1 \mid X)$ e a corrige pela constante de rotulagem $c$, recuperando $P(Y = 1 \mid X)$ sob SCAR conforme a Equação [eq:elkan noto]. Como $c$ apenas reescala os escores, a abordagem preserva o ordenamento do baseline, mas o ancora explicitamente na propensão latente e torna a constante interpretável como a taxa de cobertura efetiva da política.

A terceira é o estimador de risco não negativo (nnPU) (Kiryo et al. 2017), que reescreve o risco de classificação diretamente em termos de positivos e não rotulados, sem tratar $S$ como substituto da classe verdadeira. Depende da prevalência $\varpi = P(Y = 1)$ e impõe um truncamento inferior em zero ao termo associado à classe negativa, o que reduz o sobreajuste a que o estimador não enviesado está sujeito em modelos flexíveis. Por sua sensibilidade a $\varpi$, esse parâmetro é submetido a análise de sensibilidade.

Em ambas as estratégias supervisionadas em $S$ — *baseline* e correção de Elkan e Noto —, o classificador de base é um modelo de *Gradient Boosting* (Chen and Guestrin 2016), adequado à captura de não linearidades em dados tabulares heterogêneos; o nnPU é implementado com uma rede neural, com escores calibrados monotonicamente para garantir comparabilidade entre as abordagens.

## 1.3 Avaliação sem a classe latente observada

A avaliação esbarra na não observação de $Y$: métricas calculadas sobre $S$ não medem diretamente o desempenho sobre a classe verdadeira, pois um positivo oculto corretamente ranqueado permanece como $S = 0$ e é contabilizado como erro. Por isso a avaliação se apoia em métricas de ordenamento. A $AUC_S$, calculada sobre o rótulo observado, permanece informativa porque, sob SCAR, a ordenação induzida por $P(S = 1 \mid X)$ preserva a de $P(S = 1 \mid$

*X*), funcionando como *proxy* da capacidade global de separação. O *Recall@k*, nos cortes de 1%, 5% e 10% do topo, mede a fração de positivos observados concentrada nos estratos prioritários, quantidade diretamente ligada às estratégias de busca ativa. A *Precision@k* complementa, medindo a pureza observacional desse topo. Nenhuma dessas métricas valida *Y* diretamente: a precisão sobre a classe latente não é estimável a partir de dados PU (Bekker and Davis 2020), de modo que o conjunto fornece uma base comparativa transparente, e não uma mensuração do desempenho real sobre os positivos ocultos.

## 2. METODOLOGIA

Esta seção descreve como a estrutura institucional das bases foi convertida em um *dataset* analítico compatível com o PU Learning, detalhando as fontes e sua integração, as variáveis preditivas, a implementação das três estratégias de modelagem definidas na Seção 2 e o desenho de validação.

### 2.1 Bases, população e integração

A operacionalização empírica integra duas fontes administrativas complementares. O Cadastro Nacional da Agricultura Familiar (CAF), mantido pelo Ministério do Desenvolvimento Agrário e Agricultura Familiar (MDA), descreve a população potencial de agricultores familiares e fornece os atributos observáveis (*X*) *X* — demográficos, produtivos, territoriais e fundiários. A base de operações do PRONAF no Sistema de Operações do Crédito Rural e do Proagro (SICOR), gerida pelo Banco Central, registra a evidência observada de contratação de crédito, sem contemplar propostas recusadas ou demandas não formalizadas.

Adotou-se uma extração do CAF com data de referência em 31 de dezembro de 2025, marco também usado para atributos derivados, como a idade. Os dados demográficos foram tratados no nível do indivíduo, identificado por cpf_agricultor; os atributos produtivos, territoriais e fundiários, no nível familiar (id_familia), sendo replicados aos membros da mesma Unidade Familiar de Produção Agrária (UFPA). Do SICOR, utilizaram-se os registros públicos do PRONAF de 2021 a 2025, agregados por CPF para identificar beneficiários únicos com ao menos uma operação no período.

A integração final adotou o CAF como matriz de referência e o CPF como chave, preservando integralmente a população cadastrada em 6.851.002 registros individualizados, sendo essa a unidade de análise do estudo. O diagnóstico de cobertura da chave revelou que, entre os 2.632.960 CPFs únicos do PRONAF, 574.741 (21,83%) não foram localizados no CAF[2], permanecendo o universo analítico restrito aos indivíduos presentes no *snapshot* do cadastro.

Os indivíduos com correspondência no SICOR receberam $S = 1$ (acesso observado ao PRONAF no período); os demais, $S = 0$. A base final reúne 2.058.219 agricultores com acesso observado (30,04%) e 4.792.783 sem acesso observado (69,96%). Reforça-se que $S = 0$ indica estritamente a ausência de contratação observada no intervalo, e não inelegibilidade ou ausência de demanda — interpretação central para a leitura dos resultados. Embora a unidade de análise seja o indivíduo, parte das covariáveis é compartilhada entre membros da mesma UFPA, dependência que exige análises e tratamentos agrupamento familiar na validação.

2 A não correspondência pode refletir beneficiários identificados, no momento da contratação, pela Declaração de Aptidão ao PRONAF (DAP) que era o instrumento anterior ao CAF, cuja emissão foi encerrada em novembro de 2022.

## 2.2 Variáveis e preparação

O conjunto de preditores foi derivado da base integrada. Das 47 colunas originais, excluíram-se a variável de rotulagem (S) e os identificadores (cpf_agricultor e id_familia); o código municipal (ibge) foi removido do vetor de treino para reduzir a memorização espacial por identificadores finos, sendo mantido apenas como metadado para análises territoriais posteriores. Restaram 43 variáveis candidatas, organizadas em quatro grupos (Tabela 1). A engenharia de atributos incluiu o cálculo da idade a partir da data de nascimento (referência em 31/12/2025), a codificação do enquadramento de renda por variáveis indicadoras e a criação de marcadores de ausência (*missing indicators*) onde a própria falta do dado pode ser informativa. O dicionário completo de variáveis e o pré-processamento detalhado constam do repositório do projeto (https://github.com/marcelo7bastos/artigo-pu-learning).

| Grupo | Exemplos de variáveis | Nível |
|---|---|---|
| Demográficas e cadastrais | idade, sexo, etnia, estado civil, tipo de membro familiar, indicador de assentamento | Individual |
| Produtivas e econômicas | enquadramento de renda, renda total, nº de produtos, diversidade de fontes, agroecologia, pecuária | Familiar (UFPA) |
| Territoriais | UF, região, coordenadas do imóvel, validade espacial | Territorial |
| Fundiárias e estruturais | módulo fiscal, nº de imóveis, área agregada, vínculo com o INCRA | Familiar (UFPA) |

**Fonte:** elaboração própria a partir do CAF. O dicionário completo de variáveis consta do repositório.

## 2.3 Modelos e implementação

As três estratégias definidas na Seção 2 foram implementadas sobre a mesma base analítica. O **baseline supervisionado** estima diretamente $P(S = 1 \mid X)$, tratando o conjunto não rotulado como negativo no treino. Foi implementado com modelos de árvores (*HistGradientBoosting* e *XGBoost*), adequados a dados tabulares de grande escala por capturarem interações e não linearidades sem especificação manual de termos. A seleção entre eles baseou-se no desempenho em validação e no custo computacional, com objetivo binário probabilístico e *early stopping* monitorado pela $AUC_S$.

A **correção de Elkan e Noto** estima $P(S = 1 \mid X)$ com um classificador probabilístico e a reescala pela constante de rotulagem $c = P(S = 1 \mid Y = 1)$, recuperando $P(Y = 1 \mid X)$ sob SCAR. A correção foi instanciada com diferentes estimadores-base (regressão logística, SGD e *Gradient Boosting*); o representante selecionado por desempenho em validação foi o baseado em *Gradient Boosting* (XGBoost), levado às análises subsequentes, com calibração monotônica aplicada como pós-processamento dos escores. O alvo de treino permanece *SS*: a aderência ao problema PU vem da correção por *c*, e não de uma nova função de perda.

O **estimador de risco não negativo (nnPU)** reescreve o risco empírico em termos de positivos rotulados e casos não rotulados, usando *S* apenas para definir os conjuntos observados de treino, sem tratar $S = 0$ como classe negativa verdadeira. O risco é ponderado pela prevalência latente $\varpi = P(Y = 1)$, definida como hipótese operacional. Foi implementado

com uma rede neural *Multi-Layer Perceptron* rasa, em PyTorch, otimizada por Adam, com truncamento inferior em zero do termo de risco negativo para conter o sobreajuste a que o estimador PU não enviesado está sujeito.

O *early stopping* foi guiado pela $AUC_S$ na validação, isto é, por uma métrica proxy calculada sobre o rótulo observado *S*, separando a função objetivo otimizada do critério empírico de seleção. A prevalência $\varpi$ não foi otimizada para maximizar desempenho: foi fixada na execução principal e submetida a análise de sensibilidade em uma grade curta, tratada como hipótese operacional cuja robustez precisa ser avaliada.

Em todas as abordagens, a imputação por mediana e, quando aplicável, a padronização foram ajustadas apenas no conjunto de treino e propagadas aos demais subconjuntos, e os particionamentos respeitaram o bloqueio por id_familia. As arquiteturas detalhadas, funções de perda, grades de hiperparâmetros e sementes de aleatorização constam do repositório.

## 2.4 Validação e métricas

Como *Y* não é observável, a seleção de modelos, a avaliação final e os testes de robustez apoiaram-se em métricas *proxy* sobre o rótulo observado *S*; os resultados devem ser lidos como evidência de separabilidade observável e utilidade para ranqueamento, e não como validação direta do desempenho sobre *Y* (Bekker and Davis 2020).

A base foi particionada em treino, validação e teste, em duas etapas, com proporção aproximada de 80%/10%/10%, usando *GroupShuffleSplit* com bloqueio por id_familia. Como a divisão é feita no nível dos grupos familiares, as proporções efetivas de registros podem variar ligeiramente conforme o tamanho das famílias. Esse procedimento impede que membros da mesma família apareçam em subconjuntos distintos, reduzindo o vazamento de informação decorrente do compartilhamento de atributos produtivos, territoriais e estruturais replicados no nível pessoa. Assim, a avaliação passa a refletir melhor a generalização para famílias não observadas no treino, sendo mais conservadora do que uma divisão aleatória por linha. Como o particionamento por grupos não foi estratificado por *S*, pequenas diferenças na proporção de positivos observados ($S = 1$) entre treino, validação e teste são esperadas e foram monitoradas.

Contudo, a análise foi conduzida no nível do agricultor, identificado por cpf_agricultor, enquanto o particionamento em treino, validação e teste foi bloqueado por id_familia. Assim, o modelo estima scores para indivíduos, mas evita que agricultores pertencentes à mesma família apareçam em subconjuntos distintos.

As métricas centrais foram a $AUC_S$, que resume a separação global entre $S = 1$ e $S = 0$ $S = 0$ *Recall@k* nos cortes de 1%, 5% e 10% do topo, que mede a fração de positivos observados recuperada nos estratos prioritários; a *Precision@k* complementa como medida auxiliar da pureza do topo. A seleção ocorreu na validação e a avaliação final, no conjunto de teste.

A escolha do modelo de referência não seguiu uma regra única de maximização: cada família fixou previamente seu ranking operacional por critérios próprios de validação, robustez e interpretabilidade, e só então os rankings foram comparados entre si — sobre os conjuntos Top-100 mil, descritos pelo percentual de interseção e pelo índice de Jaccard. A decisão combinou desempenho *proxy* em *S*, estabilidade do ranking e coerência com a estrutura de PU Learning, evitando eleger automaticamente a abordagem com melhor resultado isolado em uma única métrica.

# 3. RESULTADOS

Esta seção avalia, primeiro, o desempenho dos modelos no espaço observável de validação e, em seguida, traduz os escores em uma leitura de política pública sobre acesso potencial e não inclusão no PRONAF. Como a classe latente *Y* não é observável nas bases administrativas, toda a avaliação recai sobre o rótulo observado *S*.

## 3.1 Comparação dos modelos e escolha da referência

As métricas comparativas são calculadas sobre *S*, e não sobre a classe latente *Y* (Bekker and Davis 2020). Avaliam, portanto, a aderência de cada modelo ao padrão observado de contratação como mecanismo institucional de concessão, e não a recuperação direta da positividade latente. Essa distinção é decisiva porque as estratégias não otimizam o mesmo alvo. O baseline e a correção de Elkan e Noto produzem rankings governados por $P(S = 1 \mid X)$ (a correção por *c*, sendo monotônica, não altera o ordenamento; ver Seção 2.2), ao passo que o nnPU ordena os casos por uma estimativa de $P(Y = 1 \mid X)$ . A Tabela 2 resume o desempenho nas métricas de separação e ranqueamento.

Métricas calculadas sobre o rótulo observado *S*. Recall@k: fração de positivos observados recuperada nos *k*% do topo. Precision@k: proporção de positivos observados nesse topo. Fonte: elaboração própria.

| Modelo | $AUC_s$ | *Recall@k* | | | *Precision@k* | | |
|---|---|---|---|---|---|---|---|
| 3-5(lr)6-8 | | 1% | 5% | 10% | 1% | 5% | 10% |
| Baseline supervisionado | 0,844 | 0,031 | 0,140 | 0,259 | 0,941 | 0,843 | 0,779 |
| PU clássico (Elkan e Noto) | 0,825 | 0,031 | 0,135 | 0,248 | 0,929 | 0,814 | 0,747 |
| nnPU | 0,710 | 0,020 | 0,096 | 0,185 | 0,602 | 0,581 | 0,557 |

Observa-se um ordenamento consistente. O *baseline* supervisionado obtém o melhor desempenho agregado ($AUC_S$ = 0,844), seguido de perto pela correção de Elkan e Noto (0,825); o nnPU fica substancialmente abaixo (0,710). No topo extremo do ranking, *baseline* e Elkan e Noto são praticamente idênticos — ambos recuperam cerca de 3,1% dos positivos observados no 1% superior, com precisão acima de 0,92 —, divergindo apenas marginalmente nos cortes de 5% e 10%. O nnPU apresenta concentração mais difusa, com recuperação e precisão menores em todos os estratos.

Essa diferença tem leitura operacional direta. Sob restrições, de capacidade e recursos, orçamentárias, operacionais ou institucionais, importa concentrar os casos mais aderentes nos primeiros decis. Nesse critério, *baseline* e Elkan e Noto são igualmente eficazes na focalização do topo, enquanto o nnPU dispersa os casos ao longo da ordenação, reduzindo sua utilidade para priorização de curto prazo.

O desempenho inferior do nnPU em *S* não deve ser lido como menor capacidade preditiva sobre *Y*. Por modelar explicitamente a classe latente, esse estimador atribui escores altos a casos não rotulados (*S* = 0) plausivelmente positivos; avaliados contra *S* , tais casos são

contabilizados como erro, o que penaliza sistematicamente suas métricas. O resultado é, assim, compatível com o comportamento esperado de um modelo que se afasta do padrão observado para capturar estrutura latente — e não evidência de inferioridade substantiva.

Para fins de reprodutibilidade, registram-se os parâmetros estruturais efetivamente adotados. A constante de rotulagem estimada pela correção de Elkan e Noto foi $\hat{c} \approx 0{,}48$ (interpretável como a taxa de cobertura efetiva do PRONAF, da ordem de 48%). A prevalência assumida no nnPU foi $\varpi = 0{,}73$, submetida a análise de sensibilidade em $\varpi \in \{0{,}65; 0{,}70; 0{,}73; 0{,}75\}$.

Embora o *baseline* lidere as métricas observáveis, adota-se a correção de Elkan e Noto como referência substantiva, por duas razões. Primeiro, sua maior aderência conceitual ao problema de rotulagem parcial: ela ancora os escores na propensão latente, em vez de tratar o conjunto não rotulado como negativo. Segundo, a elevada convergência entre os dois rankings (no recorte dos 100 mil maiores escores, a interseção alcança 80.769 casos (80,8%), com índice de Jaccard de 0,677), o que indica que a referência preserva o núcleo priorizado pelo modelo de melhor desempenho observável, acrescentando a correção PU sem perda relevante de utilidade. Os escores são usados como instrumento de ranqueamento diagnóstico, e não como classificação definitiva de elegibilidade.

## 3.2 Alta propensão estimada sem acesso observado

A análise desloca o foco da comparação entre modelos para o que os escores revelam sobre a inclusão potencial e a não inclusão no PRONAF. Os recortes Top 100 mil e Bottom 100 mil foram definidos exclusivamente no universo de agricultores sem acesso observado ($S = 0$): não representam acesso efetivo, mas posições relativas no ranking de propensão estimada entre os não atendidos. O Top 100 mil reúne os casos mais próximos do padrão observado de contratação — alta propensão estimada sem acesso observado —; o Bottom 100 mil, os mais distantes.

"No estrato (%)": participação da UF no Top 100 mil (Painel A) ou no Bottom 100 mil (Painel B), definidos sobre o universo $S = 0$.

- UF
- agricultores
- nacional (%)
- $S = 1$ (%)
- (%)

| Painel A — Top 100 mil (alta propensão estimada) | | | | |
|---|---|---|---|---|
| Rio Grande do Sul | 358.130 | 5,23 | 7,04 | 17,19 |
| Minas Gerais | 513.155 | 7,49 | 8,93 | 14,73 |
| Santa Catarina | 156.121 | 2,28 | 3,19 | 9,61 |
| Paraná | 212.118 | 3,10 | 3,81 | 9,50 |
| Bahia | 1.233.532 | 18,01 | 18,78 | 6,53 |
| Pernambuco | 585.074 | 8,54 | 8,46 | 5,21 |
| Alagoas | 164.207 | 2,40 | 3,08 | 5,05 |
| Piauí | 466.811 | 6,81 | 8,28 | 4,93 |
| Paraíba | 359.537 | 5,25 | 6,27 | 4,76 |
| Maranhão | 530.393 | 7,74 | 7,01 | 4,31 |

| Painel B — Bottom 100 mil (baixa propensão estimada) | | | | |
|---|---|---|---|---|
| Bahia | 1.233.532 | 18,01 | 18,78 | 15,25 |
| Minas Gerais | 513.155 | 7,49 | 8,93 | 14,32 |
| Amazonas | 114.172 | 1,67 | 0,33 | 11,41 |
| Ceará | 789.470 | 11,52 | 9,42 | 10,77 |
| Pernambuco | 585.074 | 8,54 | 8,46 | 8,42 |
| Pará | 354.575 | 5,18 | 2,59 | 8,17 |
| Maranhão | 530.393 | 7,74 | 7,01 | 5,45 |
| Espírito Santo | 116.558 | 1,70 | 1,33 | 4,50 |
| Rondônia | 102.426 | 1,50 | 1,28 | 2,86 |
| São Paulo | 69.322 | 1,01 | 0,67 | 2,55 |

**Fonte:** elaboração própria.

O Painel A da Tabela 3 lista as unidades da federação com maior participação no Top 100 mil. O estrato concentra-se no Sul e Sudeste: o Rio Grande do Sul responde por 17,19% do Top 100 mil ante 5,23% da base nacional, padrão de sobre-representação que se repete em Minas Gerais, Santa Catarina e Paraná; estados do Nordeste (Bahia, Pernambuco, Piauí, Paraíba, Maranhão) também aparecem, em proporção mais próxima de seu peso na base. Esse é o subconjunto mais aderente ao padrão observado de contratação entre os não atendidos. A leitura exige cautela: alta propensão estimada é compatível com a presença de barreiras, e não prova de demanda reprimida.

Convém distinguir as duas lógicas regionais. No Sul e Sudeste, a infraestrutura institucional, a organização produtiva e a capilaridade do sistema financeiro permitem reconhecer os perfis elegíveis, de modo que a não contratação aproxima-se de uma lacuna de intermediação. No Nordeste, o contingente expressivo de agricultores com perfil aderente convive com limitações na capacidade do sistema de identificar, orientar e possibilitar acesso à política de crédito. Assim, uma abordagem ativa, combinada a ações de informação e simplificação de procedimentos, tende a ter impacto elevado. Em termos de política pública, esse estrato delimita a fronteira mais imediata de expansão do PRONAF entre os não atendidos, com maior retorno esperado de intervenções de curto prazo voltadas à redução de barreiras de acesso.

Esses casos devem ser entendidos como uma fronteira de investigação, e não como beneficiários excluídos comprovados: o escore expressa proximidade ao perfil administrativo de contratação, não um direito ou uma decisão de concessão. Sua utilidade está em orientar onde concentrar esforços de verificação e busca ativa.

A desagregação por posição familiar e sexo (Tabela 4) revela perfil marcadamente masculino no topo. Entre as pessoas responsáveis pela UFPA, os homens passam de 30,06% da população total para 46,44% do acesso observado e para 83,81% do Top 100 mil; as mulheres responsáveis, de 25,05% para 29,51% e 15,05%. O padrão é consistente com um ranking que reproduz e intensifica hierarquias de gênero já presentes no padrão histórico de acesso. Essa concentração não permite concluir que o modelo aprendeu de forma equivocada: ela é compatível com a reprodução de assimetrias preexistentes nos próprios dados de treinamento. Assim, um eventual uso operacional do modelo (ou de outros modelos que aprendem com dados passados) como instrumento de priorização recomendaria,

fortemente, uma etapa explícita de auditoria de equidade por sexo, com monitoramento de disparidades e, se necessário, mecanismos de mitigação ou calibração, para evitar a sub-representação sistemática de mulheres responsáveis pela UFPA.

Top 100k e Bottom 100k definidos sobre o universo *S* = 0. M: masculino; F: feminino.

| Grupo familiar | Total | | *S* = 1 | | Top 100k | | Bottom 100k | |
|---|---|---|---|---|---|---|---|---|
| 2-3(Ir)4-5(Ir)6-7(Ir)8-9 | M | F | M | F | M | F | M | F |
| Responsável pela UFPA | 30,06 | 25,05 | 46,44 | 29,51 | 83,81 | 15,05 | 0,00 | 0,00 |
| Cônjuge/ companheiro(a) | 11,11 | 17,32 | 8,55 | 13,53 | 0,12 | 1,01 | 0,00 | 0,00 |
| Filho(a) | 7,96 | 6,70 | 0,92 | 0,56 | 0,00 | 0,00 | 48,58 | 46,64 |
| Outros (<1%) | 0,84 | 0,94 | 0,25 | 0,23 | 0,00 | 0,00 | 2,35 | 2,42 |

**Fonte:** elaboração própria.

Do ponto de vista da política pública, os casos de alta propensão sem acesso observado não se distribuem de modo homogêneo: concentram-se entre os responsáveis pela UFPA — sobretudo homens —, o que situa a fronteira imediata de expansão entre agricultores já fortemente aderentes ao padrão institucional de acesso. A participação menor de mulheres responsáveis, cônjuges e demais membros sugere barreiras adicionais de natureza institucional, informacional ou relacional, que demandam estratégias complementares de inclusão.

## *3.3 Baixa propensão estimada sem acesso observado*

O Painel B da Tabela 3 apresenta as unidades da federação com maior participação no Bottom 100 mil — os casos mais distantes do padrão observado de contratação. O estrato concentra-se no Norte e Nordeste: Bahia, Ceará, Pernambuco, Pará e Amazonas combinam peso elevado na base com sobre-representação na cauda inferior. O caso do Amazonas é o mais marcante: 1,67% da base nacional, mas 11,41% do Bottom 100 mil, associado a uma taxa de acesso de apenas 0,33% — indicativo de exclusão estrutural possivelmente ligada a isolamento geográfico, baixa capilaridade institucional e oferta limitada de assistência técnica e serviços financeiros. A presença de Minas Gerais no estrato mostra que a baixa propensão não é exclusivamente regional: reflete também segmentação interna, com perfis que permanecem à margem mesmo em contextos institucionalmente mais desenvolvidos.

O Bottom 100 mil não deve ser lido como um conjunto homogêneo de inelegíveis. Combina baixa aderência ao padrão administrativo de contratação, possíveis restrições de elegibilidade e vulnerabilidades estruturais mais amplas. Baixa propensão estimada, portanto, não é sinônimo de inelegibilidade — é distância do perfil típico de beneficiário, cujas causas variam entre territórios e segmentos.

A desagregação por posição familiar (Tabela 4) reforça essa heterogeneidade. Diferentemente do topo, a cauda inferior é fortemente concentrada em filhos(as) — 48,58% homens e 46,64% mulheres —, enquanto responsáveis pela UFPA e cônjuges praticamente desaparecem do subconjunto. Aqui a posição familiar parece mais decisiva que o sexo: homens e mulheres distribuem-se de forma semelhante dentro do grupo de filhos(as), sugerindo que a baixa aderência associa-se menos a uma hierarquia de gênero isolada e mais à posição periférica

do indivíduo na estrutura familiar e cadastral. Observa-se, então, que o modelo identifica sobretudo pessoas que pertencem ao universo da agricultura familiar, mas que não ocupam a posição mais plausível de interação formal com o sistema de crédito.

Caso o objetivo da política seja a incorporação desse estrato ao sistema financeiro, não se resolve pelo crédito isolado. Requer a articulação de políticas complementares — assistência técnica e extensão rural (ATER), programas de inclusão produtiva, fortalecimento de organizações locais e ampliação da infraestrutura institucional. O valor diagnóstico do Bottom 100 mil está em revelar segmentos para os quais a resposta adequada é mais estruturada do que a simples oferta de crédito. O diagnóstico territorial, assim, orienta intervenções diferenciadas, alinhadas à diversidade de contextos da agricultura familiar.

# 4. DISCUSSÃO E CONCLUSÃO

## 4.1 Síntese e contribuição

A questão central deste estudo decorre da própria realidade da política pública: no PRONAF, a ausência de acesso observado ao crédito não significa, necessariamente, ausência de propensão, ou mesmo interesse, para acessá-lo. As bases administrativas identificam com precisão quem contratou, mas não registram propostas recusadas, demandas não formalizadas ou casos sem acesso apesar de perfil semelhante ao dos beneficiários. O conjunto sem acesso observado ($S = 0$ $S = 0$ reúne, assim, situações distintas e não constitui um conjunto de negativos confiáveis e, por isso, interpretá-lo como tal incorporaria ao modelo o próprio viés de seleção dos registros.

A contribuição do trabalho está em tratar esse desafio de forma explícita, integrando os microdados do CAF e do PRONAF em um arcabouço de *Positive-Unlabeled Learning* e construindo um ranking de propensão estimada que amplia a capacidade diagnóstica sobre a cobertura do programa. A comparação entre estratégias mostrou o *baseline* supervisionado à frente nas métricas calculadas sobre $S$, com a correção de Elkan e Noto muito próxima e o nnPU abaixo. O achado central, porém, não é essa hierarquia, mas a convergência suficiente entre as abordagens para sustentar um núcleo consistente de casos priorizados — o que reforça a utilidade do ranking como instrumento de ordenação do universo sem acesso observado.

A análise dos extremos revelou dois estratos analiticamente distintos. No topo, há agricultores com características próximas às dos beneficiários registrados, cuja ausência de acesso é compatível com fricções de inclusão e que constituem candidatos imediatos à busca ativa. Na base, situações de baixa aderência ao padrão histórico, associadas a heterogeneidade produtiva e restrições territoriais, para as quais o crédito isolado é insuficiente. A ordenação, contudo, não é distributivamente neutra: a concentração dos maiores escores entre homens e responsáveis pela UFPA é compatível com a reprodução de assimetrias já presentes no acesso observado. A principal contribuição empírica, portanto, não está em determinar quem deveria acessar o crédito, mas em diferenciar, no universo sem acesso observado, casos próximos da fronteira de expansão e casos marcados por restrições mais profundas.

## 4.2 Implicações para política pública

As implicações para a política pública organizam-se em três níveis.

No plano do **diagnóstico**, o *ranking* ordena o universo sem acesso observado segundo a propensão estimada, ampliando a capacidade analítica do Estado para distinguir, dentro do conjunto $S = 0$, observações próximas do padrão empírico dos beneficiários e observações cuja não inclusão se associa a restrições mais amplas. É um recurso para mapear lacunas de cobertura e orientar a leitura territorial da política.

No plano da **intervenção**, os dois estratos demandam respostas distintas. Os casos de alta propensão sem acesso observado constituem a fronteira mais imediata de expansão, com maior retorno esperado de ações de busca ativa, qualificação cadastral e redução de barreiras operacionais. Os casos de baixa propensão exigem articulação com instrumentos complementares como: assistência técnica e extensão rural (ATER), inclusão produtiva e fortalecimento de organizações locais; pois o crédito isolado tende a ser insuficiente.

No plano da **governança e das salvaguardas**, escore não é decisão. O *ranking* expressa uma ordenação estatística a partir de padrões administrativos incompletos, e não uma verificação conclusiva sobre quem deve acessar o crédito. Seu uso é compatível com priorização e triagem, mas não com automatização decisória: tratar os maiores escores como autorização automática de inclusão, ou os menores como justificativa de exclusão, excederia o alcance da modelagem. Aqui chegamos à conclusão mais relevante para os formuladores de política pública: como o modelo aprende de padrões historicamente registrados, sua ordenação pode reproduzir assimetrias e não ser capaz de mudar o status quo desejado. Assim, evidencia-se a necessidade de auditoria periódica da distribuição dos casos priorizados por sexo, posição familiar e território, e de articulação com os critérios normativos do programa e com a validação humana. A contribuição do modelo está em qualificar a atenção estatal sobre casos potencialmente subatendidos, não em converter inferência estatística em decisão administrativa.

## 4.3 Limitações

As limitações do estudo delimitam o alcance de suas conclusões. A primeira, e mais importante, é que a classe de interesse $Y$ não é observável nas bases administrativas nem validável *ex post*. A avaliação recaiu sobre a variável de seleção $S$, de modo que $AUC_S$, *Recall@k* e *Precision@kPrecision@k* medem aderência ao padrão de contratação registrado, e não a recuperação direta de $Y$. O melhor desempenho em métricas baseadas em $S$ não equivale, por si só, à identificação mais precisa dos casos substantivamente relevantes.

A segunda limitação é a hipótese de rotulagem. A SCAR é útil como ponto de partida formal, mas tem plausibilidade empírica restrita no crédito rural, em que fatores territoriais, institucionais e operacionais (representados em $X$) podem influenciar a probabilidade de observar $S = 1$. O ranking deve, portanto, ser lido como inferência condicional às hipóteses de modelagem, não como medida direta da composição verdadeira do conjunto de interesse. Adicionalmente, a constante estimada pela correção de Elkan e Noto ($\hat{c} \approx 0{,}48$) implica uma prevalência latente da ordem de 0,62, abaixo do valor $\varpi = 0{,}73$ assumido no nnPU; a diferença é coerente com o fato de as abordagens repousarem sobre hipóteses distintas sobre a estrutura latente.

A terceira refere-se à cobertura e à vinculação das bases. O pareamento entre CAF e PRONAF não recupera todo o universo do crédito rural: 574.741 CPFs do PRONAF (21,83% dos registros do Banco Central) não foram localizados no CAF utilizado, possivelmente em razão de diferenças de cobertura cadastral e do uso anterior da DAP. Isso restringe a abrangência do universo analisado e exige cautela na generalização.

A quarta é a heterogeneidade persistente de $S = 0$, que reúne ausência de demanda, demanda não formalizada, propostas não observadas e baixa propensão real, ou seja, situações que os dados atuais não distinguem completamente. Por isso o ranking é instrumento diagnóstico de priorização, e não classificação definitiva das formas de não inclusão.

A essas somam-se três ressalvas: a ausência de validação externa dos casos priorizados; o caráter associativo, e não causal, das relações estimadas; e o risco, já discutido, de reprodução de desigualdades históricas presentes nos dados. Em conjunto, essas limitações não anulam a contribuição do estudo, mas precisam os termos sob os quais seus resultados podem ser interpretados.

## CONCLUSÃO

Este estudo estimou a propensão de acesso ao crédito do PRONAF entre os agricultores familiares com CAF ativo, sob rotulagem parcial, comparando três estratégias de modelagem. Mostrou que o conjunto sem acesso observado ($S = 0$) não é um grupo de negativos confiáveis, mas um universo heterogêneo, e que o arcabouço de *Positive-Unlabeled Learning* permite construir um ranking de propensão capaz de diferenciar, nesse universo, casos próximos do padrão empírico dos beneficiários e casos marcados por restrições produtivas, territoriais, institucionais e familiares mais amplas. A contribuição metodológica está em formular explicitamente o problema de focalização como aprendizado sob rotulagem incompleta, evitando converter a não observação em evidência de inelegibilidade; a contribuição para a política pública está em oferecer um instrumento de diagnóstico e priorização da cobertura do programa. Reitera-se que esse ranking é um exercício acadêmico, condicionado à variável observada $S$ e às hipóteses adotadas, e não pode ser interpretado como uma classificação definitiva de quem deve acessar o crédito.

A agenda futura organiza-se em quatro prioridades: (i) relaxar a hipótese SCAR, explorando formulações sob SAR que modelem a heterogeneidade da probabilidade de rotulagem; (ii) incorporar validação externa dos casos priorizados, por auditoria amostral ou rótulos auxiliares; (iii) integrar informações sobre recusas e demanda não formalizada, hoje ausentes das bases; e (iv) auditar a equidade distributiva do ranking e acompanhar longitudinalmente se os casos priorizados se convertem em acesso observado.

### *Declaração de uso de inteligência artificial generativa*

Na elaboração deste artigo, foram utilizadas ferramentas de inteligência artificial generativa (NotebookLM, Overleaf AI, Codex, ChatGPT e Claude) como apoio à organização e ao levantamento da bibliografia, à elaboração e à revisão de códigos, à formatação do manuscrito e à revisão, reorganização e aprimoramento da redação.

Todos os conteúdos, códigos e sugestões produzidos com a assistência dessas ferramentas foram submetidos à revisão e ao julgamento crítico do autor. A concepção do problema, as decisões metodológicas, a execução e a validação das análises, a interpretação final dos resultados e a formulação das conclusões não foram delegadas a sistemas automatizados e permanecem sob a responsabilidade integral do autor. As ferramentas de inteligência artificial utilizadas não se qualificam como autoras nem respondem por qualquer aspecto do trabalho.

## *Disponibilidade de dados e código*

O código-fonte utilizado na integração das bases, na engenharia de atributos, no treinamento e na avaliação dos modelos e na geração das tabelas, acompanhado de um arquivo README.md com instruções de reprodução, está disponível em repositório público: https://github.com/marcelo7bastos/artigo-pu-learning.

Os dados do Sistema de Operações do Crédito Rural e do Proagro (SICOR), utilizados neste estudo exclusivamente para identificar a ocorrência de contratação de crédito pelo PRONAF, estão disponíveis no portal do Banco Central do Brasil: https://www.bcb.gov.br/estabilidadefinanceira/tabelas-credito-rural-proagro.

Os microdados individualizados do Cadastro Nacional da Agricultura Familiar (CAF) não são disponibilizados publicamente, por conterem dados pessoais e informações que, isoladamente ou combinadas, podem permitir a identificação dos agricultores, nos termos da Lei Geral de Proteção de Dados (LGPD). O acesso a essa base deve ser solicitado ao Ministério do Desenvolvimento Agrário e Agricultura Familiar (MDA) e está sujeito aos procedimentos de autorização e às condições legais e institucionais aplicáveis.

Dados públicos agregados do CAF e da DAP, por município e unidade da Federação, podem ser consultados em: https://www.gov.br/mda/pt-br/acesso-a-informacao/acoes-e-programas/programas-projetos-acoes-obras-e-atividades/cadastro-nacional-da-agricultura-familiar/transparencia.

A base individual resultante do cruzamento entre o CAF e o SICOR também não é disponibilizada, devido ao risco de identificação ou reidentificação dos agricultores. Dessa forma, o código e os dados públicos permitem examinar a metodologia e reproduzir parte do fluxo computacional, mas a reprodução exata dos resultados depende de acesso autorizado aos microdados do CAF.

## REFERÊNCIAS

AQUINO, Joacir Rufino de, Marcio Gazolla, and Sergio Schneider. 2018. "Dualismo No Campo e Desigualdades Internas Na Agricultura Familiar Brasileira." **Revista de Economia e Sociologia Rural** 56 (1): 123–42.

Araujo, Jair Andrade de, Matheus Oliveira de Alencar, and José Eustáquio Ribeiro Vieira Filho. 2020. **Crédito Rural e Agricultura Familiar No Brasil: Uma Avaliação Do Programa Nacional de Fortalecimento Da Agricultura Familiar**, 2020.

BANCO CENTRAL DO BRASIL. 2026. **Matriz de Dados Do Crédito Rural**. \url{https://www.bcb.gov.br/estabilidadefinanceira/tabelas-credito-rural-proagro}.

Bekker, Jessa, and Jesse Davis. 2020. "Learning from Positive and Unlabeled Data: A Survey." **Machine Learning** 109 (4): 719–60.

BRASIL. 2025. **Portaria MDA no 20, de 21 de Março de 2025. Estabelece as Condições e Os Procedimentos Gerais Para Ingresso Na Rede Das Entidades Credenciadas Para Realizar Inscrição No Cadastro Nacional Da Agricultura Familiar**. Diário Oficial da União.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 785–94.

Elkan, Charles, and Keith Noto. 2008. "Learning Classifiers from Only Positive and Unlabeled Data." **Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 213–20.

Kiryo, Ryuichi, Gang Niu, Marthinus C. du Plessis, and Masashi Sugiyama. 2017. "Positive-Unlabeled Learning with Non-Negative Risk Estimator." **Advances in Neural Information Processing Systems 30**. 2017,

Rodrigues, Gabriel Magalhães. 2019. "O Pronaf Na Zona Da Mata Mineira: Efeitos Nos PIBs Total e Setorial Dos Municípios." **Revista de Economia e Sociologia Rural** 57 (1): 29–48.

Silva, Sandro Pereira. 2012. **Políticas Públicas, Agricultura Familiar e Desenvolvimento Territorial: Uma Análise Dos Impactos Socioeconômicos Do PRONAF No Território Médio Jequitinhonha - MG**. Texto Para Discussão No. 1693. Instituto de Pesquisa Econômica Aplicada (Ipea).

Souza, Paulo Marcelo de, Marlon Gomes Ney, and Niraldo José Ponciano. 2015. "Análise Da Distribuição Dos Financiamentos Rurais Entre Os Estabelecimentos Agropecuários Brasileiros." **Revista de Economia e Sociologia Rural** 53 (2): 251–70.

# Inteligência Estratégica para a Rede Federal: perfil econômico e oferta educacional

*Nelson Oliveira Santos*
*Prof. Dr. Allan V . Castro Quadros*

## Resumo

A presente pesquisa aborda o desafio estratégico do alinhamento entre a oferta de cursos da Rede Federal de Educação Profissional, Científica e Tecnológica e as vocações econômicas territoriais brasileiras. O objetivo central foi desenvolver um modelo de inteligência territorial, fundamentado em Ciência de Dados, para orientar a tomada de decisão na oferta e expansão de cursos. A metodologia consistiu no processamento de dados socioeconômicos dos 5.570 municípios brasileiros (PIB e emprego formal), seguido da aplicação do algoritmo de aprendizado de máquina não supervisionado *K-Means* para agrupamento (clustering) e de técnicas para explicabilidade do modelo. Os resultados revelaram uma aderência da oferta à vocação econômica dos municípios de apenas 13,7% e uma divergência total de 44,4%, evidenciando a necessidade de reorientação para eixos tecnológicos mais aderentes às demandas locais. O trabalho conclui que a utilização de modelos preditivos e a análise de agrupamentos permitem uma gestão da educação profissional baseada em evidências, otimizando o gasto público e fortalecendo o papel dos Institutos Federais como indutores do desenvolvimento regional sustentável.

**Palavras-chave**: Rede Federal. Machine Learning. Eixos Tecnológicos

## *INTRODUÇÃO*

A Rede Federal de Educação Profissional, Científica e Tecnológica desempenha um papel fundamental no desenvolvimento socioeconômico brasileiro, fundamentando-se na premissa de que a oferta educacional deve estar intrinsecamente vinculada às vocações produtivas locais. No entanto, observa-se um desafio estrutural persistente: o descompasso entre os cursos ofertados e as reais demandas do mercado de trabalho regional, o que compromete a eficácia dos Institutos Federais (IFs) como indutores de empregabilidade e inovação. Tradicionalmente, a definição de novos cursos e a expansão de *campi* baseiam-se em análises documentais estáticas ou percepções subjetivas, métodos que carecem de agilidade para acompanhar as transformações dinâmicas da economia.

O problema central desta pesquisa reside na dificuldade de identificar, de forma automatizada e baseada em evidências, o alinhamento entre os eixos tecnológicos ofertados pelas unidades da Rede Federal e os Arranjos Produtivos Locais (APLs) onde estão instaladas. Diante desse cenário, este trabalho propõe a utilização de algoritmos de *Machine Learning*, especificamente técnicas de agrupamento (*clustering*), para mapear as vocações econômicas municipais e sugerir eixos tecnológicos pertinentes.

O objetivo geral é desenvolver um modelo de inteligência territorial que forneça aos gestores da Rede Federal uma ferramenta robusta para orientar a tomada de decisão na oferta e expansão de cursos. Especificamente, busca-se processar bases de dados socioeconômicas, agrupar municípios com perfis produtivos similares e diagnosticar o nível de aderência da rede frente ao potencial identificado pelos modelos de aprendizado de máquina. Esta abordagem visa transpor a gestão da educação profissional do campo das percepções para o campo da ciência de dados, garantindo que cada vaga ofertada represente, de fato, uma oportunidade de transformação regional.

### *Desenvolvimento*

A presente seção detalha a trajetória analítica e metodológica percorrida para fundamentar a estratégia de expansão e alinhamento da Educação Profissional e Tecnológica (EPT). Inicialmente, estabelece-se a base teórica sobre a função social dos Institutos Federais, explorando o papel dos eixos tecnológicos como organizadores do conhecimento e sua correlação vital com os Arranjos Produtivos Locais (APLs). Compreendida a necessidade de sintonizar a oferta educacional às vocações territoriais, o texto avança para a implementação técnica, descrevendo a aplicação do algoritmo *K-Means* e realizando a explicabilidade em *Machine Learning* como ferramenta de transparência para converter abstrações algorítmicas em critérios auditáveis de decisão, culminando em um diagnóstico de aderência que confronta a matriz de vagas atual com os perfis socioeconômicos identificados, permitindo o mapeamento de demandas reprimidas e oportunidades de governança.

Cabe salientar que, no desenvolvimento técnico da pesquisa, ferramentas de Inteligência Artificial (IA) generativa foram empregadas de forma secundária e consultiva, atuando como suporte na automação de rotinas em linguagem *Python*, no refino gramatical e na organização estrutural do texto. Ressalta-se que a condução estratégica, a escolha das premissas metodológicas e a interpretação crítica dos dados refletem o trabalho intelectual autoral, em estrita observância às diretrizes éticas e metodológicas da pesquisa científica. Informações detalhadas sobre a natureza e o escopo do uso dessas ferramentas constam na declaração apresentada no Apêndice A.

## 1. REFERENCIAL TEÓRICO

A fundamentação teórica desta pesquisa articula a concepção político-pedagógica dos Institutos Federais, a organização curricular por eixos tecnológicos e a teoria dos arranjos produtivos locais, sustentando a hipótese de que a implementação de um sistema inteligente, alimentado por documentos institucionais e bases públicas como IBGE e CAGED, tem capacidade de combater a carência de mecanismos de planejamento e a limitada capacidade analítica das instituições, solucionando a baixa articulação existente entre os dados educacionais e as demandas de mercado.

**A função social dos Institutos Federais e o desenvolvimento local**

A criação dos Institutos Federais (IFs) pela Lei nº 11.892/2008 não representou apenas uma mudança de nomenclatura, mas a consolidação de um modelo institucional absolutamente inovador (Pacheco, 2011). Segundo Pacheco (2011), os IFs foram concebidos para firmar um compromisso de intervenção em suas respectivas regiões, devendo explorar as potencialidades de desenvolvimento, a vocação produtiva de seu *lócus*, bem como atuar na geração de tecnologias e conhecimentos aplicados a esse espaço.

Essa missão exige que a oferta de cursos não siga uma lógica de submissão automática às demandas imediatas das empresas. Frigotto (2018) alerta para o risco de um projeto educacional onde o fim é o mercado e o capital; e o meio são os seres humanos, buscando apenas adestrar a maioria dos jovens para uma função simples na divisão internacional do trabalho. Portanto, o alinhamento que se busca neste trabalho não é meramente utilitário, mas estratégico: visa garantir que a infraestrutura pública de ensino esteja onde a demanda social e produtiva é mais latente, superando a visão de formação instrumental em favor do desenvolvimento regional.

**Eixos tecnológicos como princípio de agrupamento**

A transição da organização curricular de grades isoladas para eixos tecnológicos é central para este trabalho, pois justifica a aplicação de algoritmos de agrupamento (*clustering*). Machado (2010) traduz o eixo tecnológico por uma linha imaginária tendo a sua volta tecnologias que por ela são atraídas, possuindo fundamental importância na sustentação da organização curricular e da identidade dos cursos. Ainda, segundo Machado (2010), para cada uma das categorias de tecnologias de um eixo tecnológico pode-se evoluir na definição de matrizes tecnológicas que não são meras listas, mas agrupamentos ordenados de informações tecnológicas que originam os arranjos de conteúdo articulados em seus aspectos lógicos. Essa afinidade estrutural permite agrupar cursos distintos (ex: Agronomia e Zootecnia) sob uma mesma lógica produtiva e de infraestrutura.

Essa estruturação teórica valida a metodologia proposta no trabalho: ao utilizar *Machine Learning* para identificar vocações territoriais, o sistema não está apenas cruzando nomes de cursos, mas identificando "famílias tecnológicas" compatíveis com a realidade econômica local, em consonância com as Diretrizes Curriculares Nacionais Gerais para a Educação Profissional e Tecnológica (Brasil, 2021).

**Arranjos Produtivos Locais (APLs) e alinhamento**

A materialidade da demanda por cursos ocorre nos Arranjos Produtivos Locais (APLs). No âmbito das pesquisas da RedeSist, Cassiolato e Lastres (2003) conceituam APLs como aglomerações territoriais de agentes econômicos, políticos e sociais – com foco em um conjunto específico de atividades econômicas – que apresentam vínculos mesmo que incipientes. Para a Rede Federal, identificar um APL não significa apenas olhar para onde há indústrias isoladas, mas perceber essa dinâmica sistêmica do território.

Nesse sentido, a pesquisa propõe superar a análise manual e intuitiva dessas aglomerações. Adota-se a perspectiva da *Knowledge Discovery in Databases* (KDD) — ou Descoberta de Conhecimento em Bases de Dados —definida por Fayyad, Piatetsky-Shapiro e Smyth (1996) como o processo não trivial de identificar padrões válidos, novos, potencialmente úteis e compreensíveis nos dados. Ao aplicar essa abordagem sobre indicadores massivos do IBGE e CAGED, busca-se revelar a "vocação oculta" dos territórios — aglomerações produtivas que demandam formação técnica, mas que ainda não foram percebidas pelos métodos tradicionais de planejamento educacional.

## *1.1 Implementação*

A implementação da solução seguiu o fluxo metodológico de um projeto de *Machine Learning*, estruturado em cinco etapas principais: preparação, pré-processamento, modelagem, explicabilidade e aplicação estratégica.

**Preparação e fontes de dados:**

A base de dados foi construída a partir da integração de fontes oficiais, utilizando indicadores socioeconômicos dos 5.570 municípios brasileiros. Do IBGE, extraíram-se o Produto Interno Bruto (PIB) e o Valor Adicionado Bruto (VAB) — total e segmentado por setores (agropecuária, indústria, serviços e administração pública). Do Ministério do Trabalho e Emprego (MTE), via CAGED, obteve-se o perfil do emprego formal, incluindo contratações, desligamentos, estoque, saldo do período, duração média dos vínculos e taxas de variação. Adicionalmente, esses indicadores foram cruzados com dados da Plataforma Nilo Peçanha, permitindo identificar a oferta de vagas da Rede Federal por eixo tecnológico em cada localidade.

**Pré-processamento:**

Dada a disparidade de escalas entre municípios (como capitais versus cidades de pequeno porte), os dados foram submetidos à normalização e padronização. Com base na análise diagnóstica da distribuição das variáveis, estabeleceu-se um plano de pré-processamento estruturado para otimizar o desempenho do modelo (Quadro 1). Para as variáveis de riqueza e escala, aplicou-se a transformação *Log1p* seguida do *StandardScaler* com o objetivo de reduzir a assimetria e ajustar a escala. Já para os saldos, utilizou-se o método *Yeo-Johnson (PowerTransformer)*, permitindo a normalização de dados que contêm valores negativos. A variação relativa, por apresentar *outliers* extremos, foi tratada com o *QuantileTransformer*, enquanto para o tempo de emprego optou-se por uma padronização simples via StandardScaler.

**Quadro 1 - Resumo do Plano de Pré-processamento**

| Grupo de Variáveis | Característica dos Dados | Tratamento (Pipeline) |
|---|---|---|
| Riqueza (pib, vl_*) | Apenas positivos, Cauda Longa extrema | 1. np.log1p (Logaritmo) |
| | | 2. StandardScaler |
| Escala (est_mes_*) | Apenas positivos, Cauda Longa extrema | 1. np.log1p (Logaritmo) |
| | | 2. StandardScaler |
| Saldos (saldo_*) | Positivos e Negativos, Pico no Zero | 1. PowerTransformer(method='yeo-johnson') |
| Qualidade (tp_emp) | Leve assimetria, positivos | 1. StandardScaler |
| Variação (vr_rel) | Outliers extremos, percentuais | 1. QuantileTransformer |

vl_*: Valor adicionado bruto dos setores
est_mes_*: estoque de emprego total e dos setores
saldo_*: saldo de emprego total e dos setores
tp_emp: tempo de emprego
vr_rel: taxa de crescimento ou queda de emprego

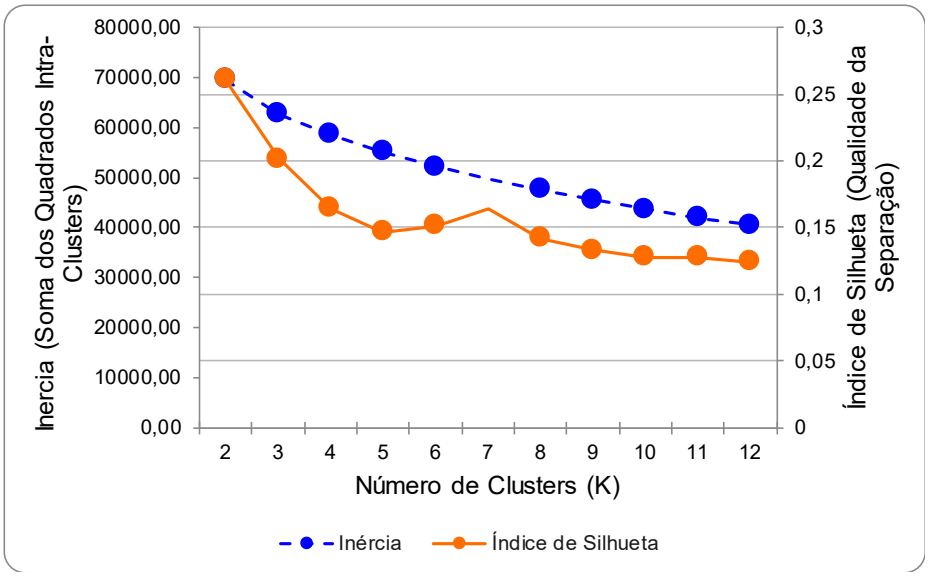
**Fonte:** Elaboração própria

**Modelagem**

Para a determinação da quantidade ideal de *clusters* (k) em algoritmos de aprendizado não supervisionado, aplicou-se uma abordagem matemática fundamentada em dois testes de validação. Primeiramente, utilizou-se o Método do Cotovelo (*Elbow Method*) para avaliar a inércia do modelo e identificar o ponto de inflexão na curva onde o ganho de compactação se tornava marginal. Complementarmente, aplicou-se o Índice de Silhueta (*Silhouette Score*) para mensurar a qualidade da separação e a coesão dos grupos, quantificando a

semelhança de cada município ao seu próprio cluster em comparação aos demais. A partir da execução desses testes e da análise conjunta dos indicadores gráficos, definiu-se o número de 7 clusters como a configuração ideal, uma vez que apresentou o melhor equilíbrio entre baixa inércia e alta silhueta, permitindo uma segmentação econômica detalhada sem gerar fragmentação excessiva dos dados.

**Gráfico 1: Definição do número de *clusters***




**Fonte:** Elaboração própria

Para garantir a robustez estatística dos resultados e mitigar eventuais vieses inerentes ao algoritmo *K-Means* — o qual pressupõe clusters com variância esférica e tamanhos similares — realizou-se uma análise comparativa com outras famílias de algoritmos sob premissas matemáticas distintas. Além do modelo principal, foram avaliados a Clusterização Hierárquica (Aglomerativa), selecionada por sua eficácia em identificar a aglutinação natural de pequenos municípios via dendrograma e o *Gaussian Mixture Models* (GMM), uma abordagem probabilística que flexibiliza a geometria dos grupos ao permitir formatos elípticos, característica crucial para dados socioeconômicos correlacionados como PIB e estoque de emprego. Como o trabalho trata de aprendizado não supervisionado, a comparação entre os modelos fundamentou-se em métricas de validação interna: o *Silhouette Score*, para medir a coesão e separação; o *Índice Calinski-Harabasz*, para avaliar a razão de dispersão entre os grupos; e o *Índice Davies-Bouldin*, utilizado para quantificar a similaridade média entre clusters vizinhos. Após o confronto dos indicadores, o *K-Means* (com k=7) foi selecionado como o algoritmo definitivo, uma vez que apresentou o melhor desempenho comparativo e a estrutura de agrupamento mais consistente para a geração dos resultados finais (Quadro 2).

**Quadro 2 - Desempenho comparativo dos algoritmos de agrupamento.**

| Modelo | Silhouette (Maior é melhor) | Calinski-Harabasz (Maior é melhor) | Davies-Bouldin (Menor é melhor) |
|---|---|---|---|
| K-Means | 0.164405 | 737.082821 | 1.316446 |
| Hierarchical (Ward) | 0.082936 | 608.980218 | 1.567087 |
| GMM (Gaussian Mixture) | 0.052930 | 401.014291 | 3.875657 |

**Fonte:** Elaboração própria

**Explicabilidade e aplicação estratégica**

A etapa de explicabilidade foi conduzida com o objetivo de conferir transparência ao funcionamento interno do modelo, transformando abstrações algorítmicas em critérios auditáveis para o suporte a políticas públicas. Para interpretar a lógica de segmentação do *K-Means*, empregaram-se três técnicas complementares: a análise de centróides (via medianas), para identificar o perfil típico de cada grupo; a construção de um modelo substituto global (Surrogate Model), utilizando uma Árvore de Decisão para extrair regras de negócio explícitas; e a análise de distribuição por meio de boxplots, visando validar a separação das faixas de valores entre os perfis econômicos.

A interpretação conjunta dessas técnicas revelou que a segmentação obedeceu a uma hierarquia de três critérios estruturantes:

**Critério de escala (magnitude econômica)**: O PIB, apoiado pelo estoque total de empregos, configurou-se como a variável determinante para a primeira grande divisão do modelo. O algoritmo estabeleceu um limiar crítico (PIB de 605 mil) que separou as elites econômicas da massa predominante de municípios. Nesse processo, isolou-se o Cluster 3 como uma entidade de "super escala", cuja magnitude de PIB e emprego distorce comparações diretas com os demais grupos.

**Critério de vocação e estabilidade (qualidade do vínculo)**: Para os municípios situados abaixo do limiar de alta escala, o fator discriminante deslocou-se do volume financeiro para a qualidade e duração do vínculo empregatício. O Cluster 6 foi definido pela retenção industrial consolidada (média de 15,5 anos), enquanto o Cluster 1 caracterizou-se por um agronegócio profissionalizado com estabilidade de mão de obra rural superior a 4 anos.

**Critério de dinamismo e dependência setorial**: Nos grupos remanescentes, a estratificação baseou-se nos valores a preços correntes e saldos setoriais. O Cluster 4 consolidou-se como uma "classe média alta" municipal, com indicadores elevados em indústria e administração. Na base da pirâmide, o Cluster 0 demonstrou alta dependência da administração pública e baixa atividade privada, enquanto o Cluster 2 posicionou-se como uma zona de transição entre a base e o dinamismo do Cluster 4.

O Quadro 3 apresenta a síntese da lógica de agrupamento e critérios de segmentação do modelo.

**Quadro 3 - Síntese da lógica de agrupamento e critérios de segmentação do modelo.**

| Variável | Papel no Modelo | Cluster que ela definiu |
|---|---|---|
| **pib / est_mes_total** | **Filtro macro:** Separou a elite da massa. | **3, 5** (Alta Escala) |
| **tp_emp_ind** | **Detector de especialidade:** Identificou estabilidade fabril. | **6** (Exclusivamente) |
| **tp_emp_agro** | **Detector de especialidade:** Identificou estabilidade rural. | **1** (Exclusivamente) |
| **vl_ind / vl_adm** | **Classificador de nível:** Separou a massa restante por "riqueza". | **4** (Alto), **2** (Médio), **0** (Baixo) |

**Fonte:** Elaboração própria

A etapa seguinte consistiu na conversão das saídas estatísticas em diretrizes para a política pública, mediante a integração do modelo a uma lógica de recomendação. A interpretabilidade dos grupos foi alcançada pela atribuição de significado semântico aos sete *clusters* gerados, fundamentada na análise de medianas, nas regras da árvore de decisão e na distribuição espacial. Cada perfil foi denominado conforme sua vocação territorial predominante e associado aos eixos tecnológicos que melhor atendiam às competências demandadas por cada realidade econômica.

Os perfis identificados e suas respectivas estratégias de oferta educacional foram estruturados da seguinte forma:

**Cluster 3 (Metrópole global de serviços e tecnologia):** Caracterizou-se como um *outlier* de magnitude econômica, com PIB e estoque de empregos altíssimos. O alinhamento focou nos eixos de Gestão e Negócios (finanças e inovação) e Informação e Comunicação (infraestrutura de TIC e economia digital).

**Cluster 5 (Polos regionais de gestão e serviços especializados):** Composto por capitais e grandes centros que centralizam a administração pública e serviços de saúde. A recomendação priorizou os eixos de Ambiente e Saúde e Desenvolvimento Educacional e Social, visando suportar a complexa rede de infraestrutura pública regional.

**Cluster 4 (Centros econômicos diversificados em expansão):** Representou municípios de médio e grande porte com crescimento urbano acelerado. A oferta estratégica concentrou-se em Infraestrutura (construção civil e logística) e Produção Cultural e Design, acompanhando o dinamismo do comércio e da economia criativa.

**Cluster 6 (Polos industriais consolidados):** Definido pela maturidade e alta retenção da mão de obra fabril. O direcionamento educacional priorizou Controle e Processos Industriais (automação e manutenção complexa) e Produção Industrial específica para os parques instalados.

**Cluster 1 (Polos de agronegócio profissionalizado):** Agrupou municípios com produção tecnificada e vínculos rurais estáveis. O alinhamento deu-se nos eixos de Recursos Naturais e Produção Alimentícia, focando na gestão tecnológica e na agregação de valor via agroindústria.

**Cluster 2 (Municípios de economia local mista):** Perfil mais frequente do país, com economia baseada no comércio e serviços locais. A estratégia pedagógica enfatizou o Empreendedorismo (Gestão e Negócios) para profissionalizar o mercado local e a Infraestrutura para a construção civil.

**Cluster 0 (Economias de subsistência e dependência pública):** Identificado como a base da pirâmide, com alta vulnerabilidade e dependência da administração pública. A recomendação centrou-se em Desenvolvimento Educacional e Social e na Produção Alimentícia (Agricultura Familiar), visando o fomento à geração de renda autônoma.

O cruzamento analítico entre a vocação econômica e as diretrizes dos Catálogos Nacionais dos Cursos Técnicos e de Tecnologia permitiu que o modelo transcendesse a segmentação estatística, oferecendo um roteiro objetivo para a expansão da Rede EPT em consonância com as necessidades territoriais.

O Quadro 4 apresenta a indicação dos eixos prioritários 1 e 2 por perfil econômico.

**Quadro 4 - Perfil econômico e eixos prioritários**

| Cluster | Perfil econômico | Eixo prioritário 1 | Eixo prioritário 2 |
|---|---|---|---|
| **Grupo 3** | Metrópole Global (Serviços e Tecnologia) | Gestão e Negócios | Informação e Comunicação |
| **Grupo 5** | Polo Regional de Gestão e Serviços | Ambiente e Saúde | Desenvolvimento Educacional e Social |
| **Grupo 4** | Centro Econômico Diversificado | Infraestrutura | Produção Cultural e Design |
| **Grupo 6** | Polo Industrial Consolidado | Controle e Processos Industriais | Produção Industrial |
| **Grupo 1** | Polo Agropecuário Profissionalizado | Recursos Naturais e | Produção Alimentícia |
| **Grupo 2** | Economia Local Mista (Pequeno Porte) | Gestão e Negócios | Infraestrutura |
| **Grupo 0** | Economia de Subsistência / Alta Dependência | Desenvolvimento Educacional e Social | Produção Alimentícia |

**Fonte:** Elaboração própria

## 2. RESULTADOS

A modelagem analítica desenvolvida neste estudo transcende a etapa exploratória, convertendo o processamento de dados em inteligência estratégica aplicada à gestão pública. Os agrupamentos (*clusters*) identificados e validados na etapa de modelagem são confrontados com os dados operacionais da Rede Federal de Educação Profissional, Científica e Tecnológica. Essa abordagem permite a avaliação da eficiência na alocação de recursos sob uma ótica orientada a dados.

A análise dos resultados está estruturada em três eixos:

**Diagnóstico de aderência:** Mensura o grau de alinhamento entre a oferta educacional predominante e a vocação econômica do território.

**Análise de divergência total:** Identifica o desalinhamento estrutural e mapeia localidades onde a oferta acadêmica não guarda correlação com os motores produtivos locais.

**Planejamento de expansão:** Projeta oportunidades para novas unidades em municípios desassistidos, fundamentado no perfil econômico latente para a definição prévia dos eixos tecnológicos recomendados.

Diagnóstico de aderência:

Esta aplicação tem por objetivo mensurar o grau de alinhamento estratégico entre a oferta educacional vigente nos Institutos Federais e as vocações econômicas reais identificadas pelo modelo. A metodologia consiste em um teste lógico de correspondência (match flexível): considera-se que há aderência quando o eixo tecnológico com maior volume de oferta de vagas no município coincide com o eixo prioritário 1 ou com o eixo prioritário 2 sugeridos pelo algoritmo de *clustering*. Esta métrica binária permite calcular a taxa de aderência, servindo como indicador de eficiência na alocação de vagas.

$$Tx\ de\ aderência\ (\%) = \left( \frac{\sum_{i=1}^{n} A_i}{n} \times 100 \right)$$

Em que:

- **n:** representa número total de municípios que ofertam vagas pela Rede EPT
- **Ai:** é uma variável indicadora (binária) para o município i, definida pela seguinte lógica de aderência:

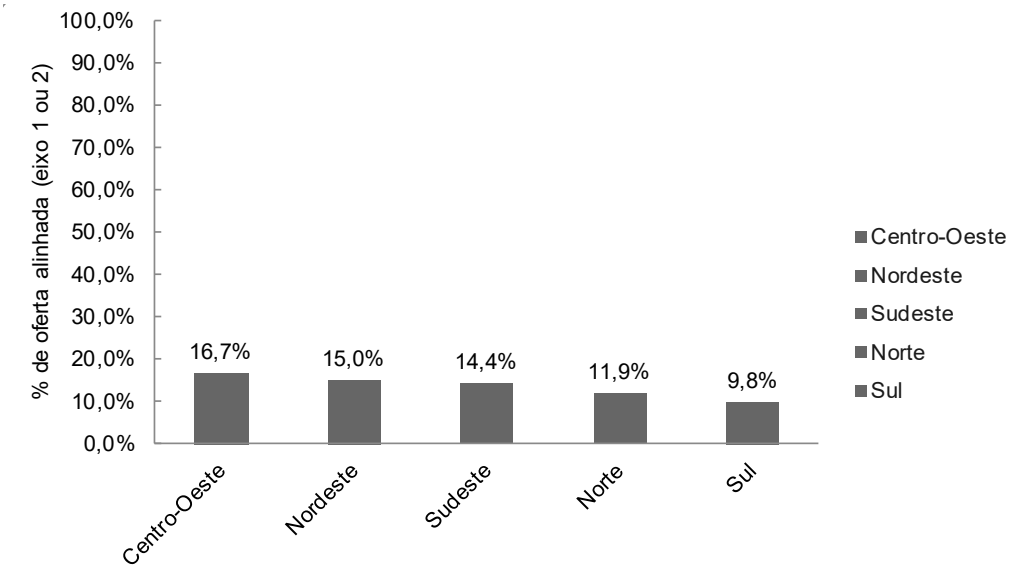
$$A_i = \begin{cases} 1, se\ E_{max} \in \{P_1, P_2\} \\ 1, se\ E_{max} \notin \{P_1, P_2\} \end{cases}$$

Sendo:

- **Emax:** O eixo tecnológico com o maior volume de vagas ofertadas no município.
- **P1, P2:** conjunto formado pelos eixos prioritários (1 e 2) identificados pelo modelo de *clustering* para o perfil econômico daquele município.

O Gráfico 2 apresenta a taxa de aderência calculada por região.

**Gráfico 2 - Taxa de aderência da oferta ao perfil econômico**




**Fonte:** Elaboração própria

O confronto entre a base de dados de vagas ofertadas e os perfis econômicos delineados pelos *clusters* revela um cenário de desalinhamento estrutural, validando a necessidade de uma gestão orientada por dados. A taxa global de aderência apurada foi de apenas 13,7%, indicando que em 86,3% dos municípios analisados o eixo principal de oferta não coincidiu com as vocações prioritárias (eixo prioritário 1 ou 2) identificadas pelo modelo. É possível inferir que a abertura de cursos, historicamente, tendeu a ser orientada por critérios de conveniência infraestrutural ou demandas pregressas, descolando-se da realidade produtiva medida pelo PIB e pelo estoque de empregos.

A análise estratificada por região expôs assimetrias significativas na eficiência da oferta:

**Região Centro-Oeste:** apresentou a maior taxa de aderência (16,7%), desempenho atribuído à clareza da vocação regional voltada ao agronegócio, o que facilitou o alinhamento natural com o eixo de Recursos Naturais. Embora o desempenho da região Centro-Oeste tenha superado o indicador nacional (13,7%), a magnitude do resultado é insuficiente para configurar uma aderência substancial, evidenciando que a região ainda enfrenta desafios estruturais de alinhamento..

**Região Sul:** registrou a menor taxa (9,8%). A hipótese levantada para este hiato é a complexidade econômica da região, onde a demanda por formação industrial especializada possivelmente foi preterida por ofertas genéricas de menor custo, resultando em baixa pontuação na métrica de aderência.

Três fatores principais podem ser considerados como causas do desalinhamento:

1. **Custo de implementação:** Eixos tecnológicos voltados a “Gestão e Negócios” e “Informação e Comunicação” demandam menor aporte financeiro e estrutural para implantação. Em contrapartida, áreas como “Controle e Processos Industriais” ou “Saúde” exigem laboratórios de alta complexidade e insumos onerosos. Sob a ótica da gestão pública, infere-se uma tendência histórica de expansão orientada pela viabilidade orçamentária imediata em detrimento das vocações locais de maior custo de instalação.
2. **Inércia institucional:** Verifica-se uma latência temporal significativa entre as transformações da dinâmica econômica e a efetiva atualização do catálogo de cursos das instituições. Unidades de ensino instaladas há décadas tendem a perpetuar matrizes curriculares e ofertas tradicionais (como a Agropecuária), negligenciando a transição de determinados municípios para perfis socioeconômicos voltados ao Comércio ou ao setor de Serviços. Esse fenômeno gera um anacronismo estrutural na oferta educacional.
3. **Metodologia de concentração de volume:** O indicador proposto avalia a oferta principal da instituição. Desse modo, é possível que o *campus* oferte cursos vinculados aos eixos tecnológicos prioritários mapeados pelo modelo (prioritários 1 ou 2), porém com baixa densidade de vagas. Paralelamente, a concentração massiva de matrículas em cursos genéricos penaliza o desempenho do município nessa métrica específica, mascarando pequenas ações de alinhamento regional.

Em síntese, o diagnóstico de 13,7% de aderência evidencia que uma parcela expressiva dos recursos públicos foi alocada em formações que não atacam as principais alavancas econômicas locais, representando uma oportunidade crítica para o ganho de eficiência institucional.

### Análise de divergência total

Diferente da aplicação anterior, que analisou a predominância da oferta, este indicador adota um critério de avaliação mais abrangente e "generoso". A metodologia verifica se algum dos eixos tecnológicos presentes no portfólio do *campus* (independentemente do número de vagas) corresponde ao perfil econômico identificado pelo modelo. O objetivo é identificar os casos de divergência total: municípios onde a instituição oferta cursos técnicos, mas nenhum deles dialoga com os motores econômicos locais. Esse indicador revela a existência de "ilhas acadêmicas", onde a formação técnica ocorre de costas para a realidade produtiva do território.

A taxa de divergência total é calculada da seguinte forma:

$$Tx\ de\ divergência\ total\ (\%) = \left( \frac{\sum_{i=1}^{n} A_i}{n} \times 100 \right)$$

- **$n$:** representa número total de municípios que ofertam vagas pela Rede EPT
- **$D_i$:** é uma variável indicadora (binária) para o município *i*, definida pela seguinte lógica de aderência:

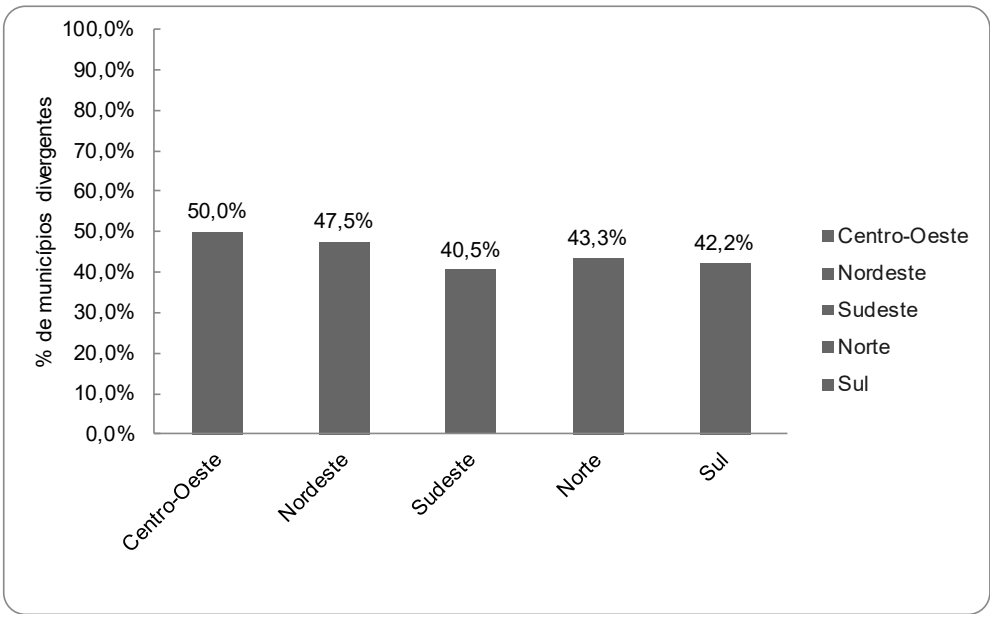
$$D_i = \begin{cases} 0, se \{E_{ofertados}\} \cap \{P_1, P_2\} \neq \varnothing \\ 1, se \{E_{ofertados}\} \cap \{P_1, P_2\} = \varnothing \end{cases}$$

Sendo:

- **$\{E_{ofertados}\}$:** Conjunto de todos os eixos tecnológicos com vagas ofertadas no município.
- **$\{P_1, P_2\}$:** conjunto formado pelos eixos prioritários (1 e 2) identificados pelo modelo de clustering para o perfil econômico daquele município.

Gráfico 3 apresenta a taxa de divergência total calculada por região.

**Gráfico 3 - Taxa de divergência total da oferta ao perfil econômico**




**Fonte:** Elaboração própria

Os resultados obtidos a partir do indicador de divergência expuseram o cenário mais crítico do estudo, evidenciando um descolamento estrutural que transcende ajustes pontuais de oferta. A taxa de divergência nacional apurada foi de 44,4%, o que indica que, em quase metade dos municípios brasileiros atendidos, a oferta de cursos ignora integralmente o perfil econômico. Este fenômeno caracteriza a existência de "ilhas acadêmicas", onde a formação profissional não dialoga com a realidade produtiva local, induzindo potencialmente à evasão de talentos ou à migração forçada de egressos.

Regionalmente, observou-se que todas as regiões apresentam índices de divergência superiores a 40%, com destaque para o Centro-Oeste, que registrou a maior taxa de divergência (50,0%).

O contraste desse dado com o indicador de aderência exige cautela analítica: embora o Centro-Oeste apresente a maior taxa de aderência em relação à oferta principal (16,7%), metade dos *campi* analisados falha integralmente em disponibilizar qualquer curso correlacionado às vocações produtivas locais.

Nos 50,0% restantes dos casos, onde não se observa divergência absoluta, o alinhamento não se mostra necessariamente robusto. Isso indica, tão somente, a existência de pelo menos uma vaga ou curso compatível com a vocação do município, a qual frequentemente coexiste com um portfólio de cursos majoritariamente genéricos, corroborando a baixa aderência do curso principal.

Nas demais regiões, os índices mantiveram-se em patamares elevados, variando entre 47,5% no Nordeste e 40,5% no Sudeste, o que ratifica a sistematicidade do desalinhamento.

A interpretação conjunta desses dados sugere que, mesmo nos casos onde não há divergência total (ou seja, onde existe ao menos uma oferta compatível), o alinhamento não é necessariamente robusto. A coexistência de altas taxas de divergência com baixas taxas de aderência do curso principal reforça a hipótese de que as matrizes de oferta são majoritariamente compostas por portfólios genéricos. Este cenário exige uma revisão dos Planos de Desenvolvimento Institucional (PDIs), priorizando a correção de trajetórias formativas que atualmente operam de forma isolada das demandas do arranjo produtivo local.

**Planejamento de expansão e interiorização**

Esta aplicação volta-se para o planejamento prospectivo da expansão da Rede Federal. Ao isolar o conjunto de municípios que não registram oferta de cursos técnicos por parte dos Institutos Federais (o "vazio assistencial"), o modelo projeta um mapa de oportunidades. O objetivo foi identificar o perfil econômico predominante nestes territórios e definir, *a priori*, qual eixo tecnológico deve fundamentar a criação de novas unidades ou polos de ensino a distância. Esta abordagem inverte a lógica tradicional: em vez de construir a unidade para depois definir os eixos tecnológicos de oferta dos cursos, define-se a vocação do território para dimensionar a infraestrutura necessária.

A análise da expansão fundamentou-se na identificação do subconjunto de municípios em situação de vazio assistencial (V). Para este grupo, quantificou-se a distribuição dos perfis econômicos predominantes e, subsequentemente mapeou-se a demanda reprimida por meio da função de contagem dos Eixos Prioritários (P1) recomendados pelo modelo de *machine learning*.

*Definição do Conjunto de "Vazio de Oferta" (V)*

Primeiro, define-se o conjunto de municípios que não possuem oferta ativa de cursos técnicos.

$$V = \{m \in M \mid O_m = \varnothing\}$$

Onde:

- **$M$:** É o conjunto de todos os municípios do país ($n$ = 5570).
- **$O_m$:** Representa o conjunto de eixos ofertados no município $m$. Se o conjunto é vazio $\varnothing$, o município é classificado como "Sem Oferta".

Matematicamente, isso é o resultado de uma operação de diferença de conjuntos no cruzamento de dados:

Distribuição por perfil econômico (Gráfico 4)

O gráfico 4 representa a frequência absoluta de municípios sem oferta, agrupados por sua tipologia econômica

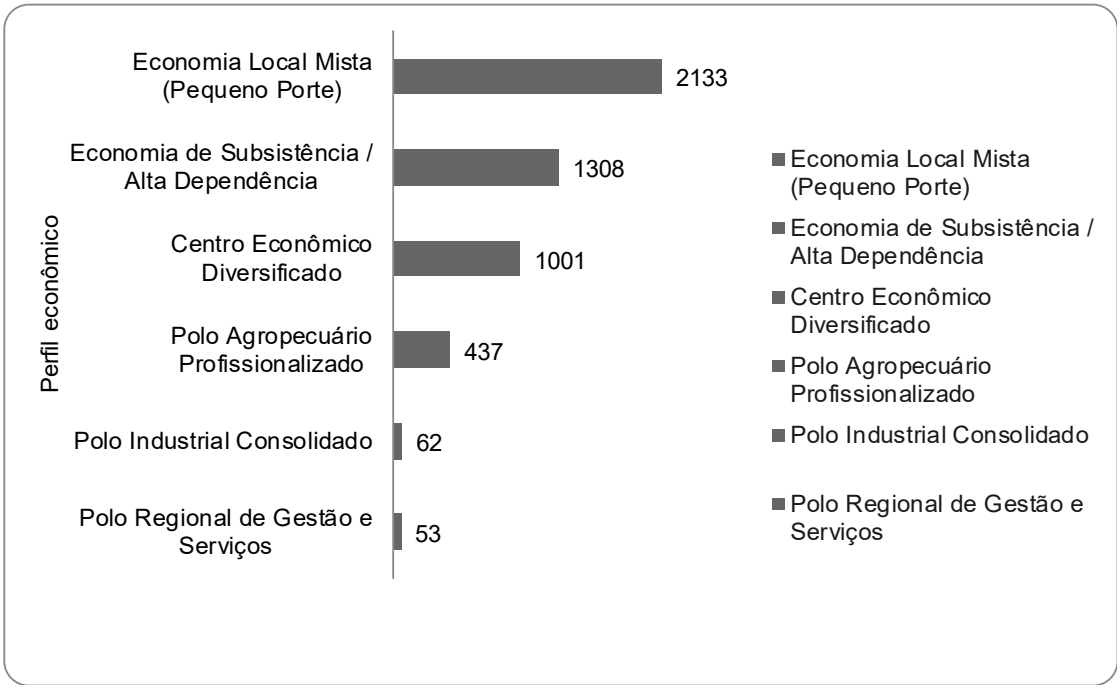
$$f(C_j) = \sum_{m \in V} 1\,(Cluster_m = C_j)$$

Onde:

- **$Cluster_m$:** É o perfil econômico atribuído ao município $m$ pelo modelo de agrupamento.
- **$1\,(Cluster_m = C_j)$:** É a função indicadora que assume valor 1 se a condição for verdadeira e 0 caso contrário.
- **$f(C_j)$:** É a quantidade de cidades de perfil econômico $j$ sem cursos.

(C). A função de contagem para cada perfil econômico j é:

**Gráfico 4 - Perfil econômico dos municípios sem oferta de cursos**




**Fonte:** Elaboração própria

Estimativa de demanda reprimida por eixo (Gráfico 5)

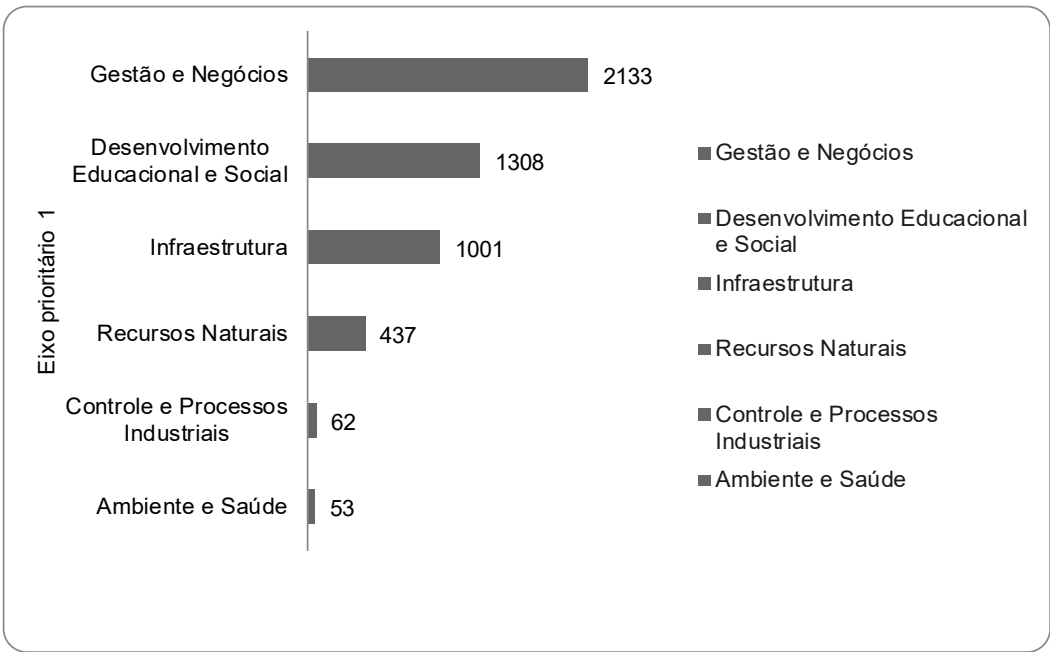
O Gráfico 5 projeta o potencial de abertura de novos cursos, baseando-se na recomendação do modelo para os municípios que estão descobertos. A contagem de eixos a serem ofertados é dada por:

$$D(P_k) = \sum_{m \in V} 1\,(P_{1,m} = E_K)$$

Onde:

- $P_{1,m}$: É o Eixo Prioritário 1 (recomendação principal) gerado pela inteligência do modelo para o município *m*.
- $E_k$: Representa um dos eixos tecnológicos ofertados pela Rede.
- $D(P_k)$: Representa a demanda reprimida, ou seja, o volume de cidades onde o eixo $E_k$ deveria ser a prioridade de expansão institucional.

**Gráfico 5: Demanda reprimida para oferta de eixos tecnológicos**




**Fonte:** Elaboração própria

A análise dos municípios desassistidos, caracterizados pela ausência de oferta direta de cursos, revela um panorama substancialmente distinto daquele observado nos grandes centros urbanos e industriais. Esse cenário sinaliza a necessidade de uma estratégia de expansão institucional pautada na capilaridade e na adequação a economias de menor escala. A distribuição dos perfis econômicos nessas localidades evidencia que o aprofundamento do processo de interiorização deve priorizar o alcance de estruturas produtivas locais que historicamente permanecem desatendidas.

O perfil mais frequente nesse grupo, abrangendo mais de 2.100 municípios, é classificado como Economia Local Mista, que compreende pequenas cidades estruturadas em torno do comércio varejista, de serviços básicos e da administração pública, sem a presença de grandes complexos industriais. O segundo maior segmento, composto por aproximadamente 1.300 municípios, enquadra-se no perfil de Economia de Subsistência, territórios nos quais a

inserção de um *campus* institucional assume o papel de indutor primário de desenvolvimento socioeconômico e cidadania, transcendendo a função puramente técnica. Adicionalmente, foram identificados cerca de 450 municípios caracterizados como Polos Agropecuários que ainda carecem de atendimento, representando oportunidades de retorno socioeconômico mais imediato, uma vez que a implantação de cursos agrícolas encontraria uma demanda produtiva previamente instalada.

Diante desse cenário, a projeção dos eixos tecnológicos recomendados — especificamente o Eixo Prioritário 1 — sugere uma reconfiguração do modelo tradicional de expansão focado no setor industrial secundário. Constata-se a hegemonia do eixo de Gestão e Negócios como a demanda reprimida mais expressiva em termos de volume. Para atender a essas cidades de economia mista, a prioridade analítica desloca-se da formação voltada à mecânica pesada ou à automação para a profissionalização do setor terciário local, abrangendo áreas como logística, contabilidade e empreendedorismo, de modo que a instituição de ensino atue como catalisadora da eficiência dos pequenos negócios.

Paralelamente, a expressiva demanda pelo eixo de Desenvolvimento Educacional e Social reflete as necessidades estruturais das regiões de subsistência, em que a capacitação de agentes comunitários, gestores públicos e educadores sociais constitui um pré-requisito para a elevação dos indicadores de qualidade de vida. O eixo de Infraestrutura manifesta-se como a terceira prioridade do modelo, concentrando-se na qualificação voltada à construção civil, setor fundamental para a absorção de força de trabalho em municípios em estágio de crescimento urbano.

Em síntese, o modelo indica que a estratégia de expansão para as aproximadamente 5.000 localidades desassistidas demanda uma reavaliação dos padrões arquitetônicos e pedagógicos vigentes, demonstrando a inviabilidade de se replicar estruturas industriais de alto custo de implantação nesses territórios. As evidências empíricas apontam para a necessidade de estruturação de unidades com matrizes curriculares direcionadas aos setores de serviços, gestão e agronegócio. Tais áreas apresentam menor ônus orçamentário para instalação e elevado potencial de dinamização das economias locais, garantindo maior eficiência na alocação de recursos públicos.

## CONCLUSÃO

O presente trabalho demonstrou que a integração de técnicas de *Machine Learning* à gestão da Rede Federal de Educação Profissional Científica e Tecnológica oferece uma solução robusta para o histórico desafio de alinhamento entre oferta educacional e vocações econômicas regionais. A utilização do algoritmo *K-Means*, validada por técnicas de explicabilidade, permitiu a transição de um modelo de planejamento reativo para uma abordagem preditiva e territorializada.

Os principais achados indicam que, embora a Rede Federal possua uma capilaridade expressiva, a aplicação dessa inteligência aos dados reais da Rede Federal revelou cenários críticos. O indicador de aderência (13,7%) e a taxa de divergência total (44,4%) evidenciam que, em grande parte dos municípios, a oferta educacional opera desconectada dos motores produtivos locais. Isso demonstra que, historicamente, a expansão e a criação de cursos podem ter sido guiadas por conveniência infraestrutural ou inércia institucional, resultando em "ilhas acadêmicas" que não dialogam com a vocação do território, o que impacta diretamente a eficiência do gasto público e na empregabilidade dos egressos.

A pesquisa revela também que a expansão da Rede não deve ser uniforme; ao contrário, deve priorizar modelos de unidades adaptáveis aos perfis socioeconômicos identificados, especialmente em regiões de economia mista onde o papel social da instituição é tão vital quanto sua função técnica.

Como contribuição prática, o modelo desenvolvido entrega um roteiro baseado em evidências para a tomada de decisão governamental, permitindo otimizar a aplicação de recursos públicos e maximizar a empregabilidade dos egressos. Para trabalhos futuros, sugere-se a incorporação de dados dinâmicos de mercado em tempo real e a criação de mecanismos de *feedback* institucional para refinar as recomendações dos clusters. Em última análise, a inteligência territorial aqui proposta reafirma a missão dos Institutos Federais como indutores de desenvolvimento regional sustentável e soberania tecnológica.

## *Disponibilidade de dados e código*

Os conjuntos de dados brutos e processados que fundamentam os diagnósticos de aderência e análises de divergência territorial apresentados nesta pesquisa estão publicamente disponíveis no repositório Figshare e podem ser acessados diretamente por meio do DOI/URL https://doi.org/10.6084/m9.figshare.32812733. Adicionalmente, todos os scripts de programação em linguagem Python desenvolvidos para o fluxo metodológico de *Machine Learning* (pré-processamento, aplicação do algoritmo K-Means, modelos comparativos e técnicas de explicabilidade algorítmica) foram depositados em repositório público no GitHub, acessível pelo link https://github.com/nosantos/Inteligencia_estrategica_para_RFEPCT

## REFERÊNCIAS

BRASIL. Ministério da Educação. Conselho Nacional de Educação. Conselho Pleno. Resolução CNE/CP nº 1, de 5 de janeiro de 2021. Define as Diretrizes Curriculares Nacionais Gerais para a Educação Profissional e Tecnológica. **Diário Oficial da União**: seção 1, Brasília, DF, p. 19-23, 6 jan. 2021. Disponível em: https://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=167931-rcp001-21&category_slug=janeiro-2021-pdf&Itemid=30192. Acesso em: 2 jan. 2026.

CASSIOLATO, José Eduardo; LASTRES, Helena Maria Martins. O foco em arranjos produtivos e inovativos locais de micro e pequenas empresas. *In*: LASTRES, Helena Maria Martins; CASSIOLATO, José Eduardo; MACIEL, Maria Lucia (Org.). **Pequena empresa**: cooperação e desenvolvimento local. Rio de Janeiro: Relume Dumará, 2003. p. 21-34. Disponível em: https://www.researchgate.net/profile/Jose-Cassiolato/publication/274699819_O_foco_em_arranjos_produtivos_locais_de_micro_e_pequenas_empresas/links/55259e740cf24b822b4056df/O-foco-em-arranjos-produtivos-locais-de-micro-e-pequenas-empresas.pdf?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19. Acesso em: 2 jan. 2026.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, Menlo Park, v. 17, n. 3, p. 37-54, 1996. DOI: 10.1609/aimag.v17i3.1230. Disponível em: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230. Acesso em: 2 jan. 2026.

FRIGOTTO, Gaudêncio (Org.). **Institutos Federais de Educação, Ciência e Tecnologia**: relação com o ensino médio integrado e o projeto societário de desenvolvimento. Rio de Janeiro: UERJ, LPP, 2018. Disponível em: https://proen.ifes.edu.br/images/stories/Institutos_Federais_de_Educa%C3%A7%C3%A3o_Ci%C3%AAncia_e_Tecnologia_-_Rela%C3%A7%C3%A3o_com_o_Ensino_M%C3%A9dio_Integrado_e_o_Projeto_Societ%C3%A1rio_de_Desenvolvimento.pdf. Acesso em: 2 jan. 2026.

PACHECO, Eliezer (Org.). **Institutos Federais**: uma revolução na educação profissional e tecnológica. São Paulo: Fundação Santillana: Moderna, 2011. Disponível em: https://www.fundacaosantillana.org.br/wp-content/uploads/2019/12/67_Institutosfederais.pdf. Acesso em: 2 jan. 2026.

MACHADO, Lucília Regina de Souza. Organização da Educação Profissional e Tecnológica por Eixos Tecnológicos. **Linhas Críticas**, Brasília, v. 16, n. 30, p. 89–108, 2010. DOI: 10.26512/lc.v16i30.3571. Disponível em: https://periodicos.unb.br/index.php/linhascriticas/article/view/3571. Acesso em: 2 jan. 2026.

## APÊNDICE A

### Declaração de uso de Ferramentas de Inteligência Artificial

Declaro, para os devidos fins de transparência acadêmica, que ferramentas de Inteligência Artificial (IA) generativa foram utilizadas como suporte auxiliar no desenvolvimento deste trabalho. O uso da tecnologia ocorreu de forma estratégica e pontual, abrangendo as seguintes frentes:

**Desenvolvimento Tecnológico:** Apoio na escrita e refinamento de *scripts* em linguagem Python para a implementação do modelo de agrupamento (*clustering*) e geração de visualizações de dados.

**Aprimoramento metodológico:** Obtenção de orientações consultivas sobre as melhores técnicas de ciência de dados aplicáveis ao contexto do problema estudado, servindo como suporte à tomada de decisão técnica.

**Refino textual e editorial:** Uso de ferramentas para correção ortográfica, gramatical e auxílio na organização lógica e estrutural das ideias apresentadas no corpo do artigo.

Ressalto que o uso da referida tecnologia deu-se estritamente como apoio secundário. A fundamentação teórica, a seleção das variáveis, a interpretação crítica dos resultados e a condução estratégica da análise refletem o trabalho intelectual autoral, em estrita observância às diretrizes éticas e metodológicas da pesquisa científica.

Ademais, esclareço que todos os dados utilizados nesta pesquisa são de domínio público, extraídos de bases governamentais oficiais (como IBGE, CAGED e bases de oferta educacional), inexistindo qualquer impedimento ético, legal ou contratual no processamento dessas informações por meio de ferramentas externas. O autor assume plena responsabilidade pelo conteúdo final, garantindo a veracidade e a integridade das informações aqui apresentadas.



# EIXO 04

## *Gestão Estratégica, no Orçamento e na infraestrutura*

MBA ENAP

# Inteligência Orçamentária: um ambiente virtual de apoio à decisão na alocação de emendas parlamentares e na formulação de políticas públicas.

*Eduardo Melo Moura Filho*
*Prof. Dr. Alex Lopes Pereira*

## Resumo

A expansão das emendas parlamentares — de R$ 15,3 bilhões em 2014 para R$ 50,5 bilhões em 2025, equivalentes a 22% das despesas discricionárias da União — ampliou o poder alocativo do Congresso Nacional sem que surgissem instrumentos capazes de orientar essas decisões. Este artigo propõe e avalia um ambiente virtual de inteligência orçamentária destinado a apoiar a alocação estratégica de emendas pelos parlamentares e a formulação de políticas indutoras pelo governo federal. A metodologia integra três estudos: análise do comportamento alocativo de 640 parlamentares (SIOP, 2020–2023), com métricas originais e quatro perfis tipológicos; estudo de validação cruzando alocações e resultados eleitorais de 441 parlamentares (TSE, 2022); e formulação de 25 propostas de políticas públicas pelo Modelo MIRE, a partir dos índices de necessidade de 5.571 municípios e do PPA 2024–2027. Os resultados indicam que dispersão geográfica, cobertura da base eleitoral e diversificação temática são as variáveis mais associadas ao sucesso eleitoral (AUC-ROC = 0,673), com potencial de indução estimado em R$ 10,09 bilhões anuais. Conclui-se que dados públicos integrados podem conectar a alocação de emendas às necessidades municipais e às prioridades governamentais, oferecendo instrumento replicável de governança orçamentária.

**Palavras-chave:** emendas parlamentares; inteligência orçamentária; comportamento orçamentário; políticas públicas; apoio à decisão.

## *INTRODUÇÃO*

O orçamento federal brasileiro passou por uma reconfiguração estrutural na última década: a transferência progressiva do controle das despesas discricionárias do Poder Executivo para o Poder Legislativo, por meio da expansão das emendas parlamentares. Em termos reais, as despesas discricionárias sob gestão do Executivo caíram de R$ 380,1 bilhões em 2013 para R$ 179,4 bilhões em 2025 — redução de 47% —, enquanto as emendas parlamentares cresceram de R$ 15,3 bilhões (2014) para R$ 50,5 bilhões (2025), saltando de 4% para 22% das despesas discricionárias (SIOP/SOF, 2025). O processo foi impulsionado pela Emenda Constitucional n. 86/2015, que tornou impositivas as emendas individuais, e consolidado pela Lei Complementar n. 210/2024. O efeito agregado é um novo paradigma: o Executivo formula o plano e executa as despesas obrigatórias, mas uma parcela crescente e legalmente protegida dos recursos discricionários é decidida unilateralmente por parlamentares, substituindo políticas desenhadas com base em planejamento técnico e nas metas do Plano

Plurianual (PPA) por alocações dispersas em dezenas de milhares de objetos.

A literatura aponta que essas destinações tendem a ser difusas, erráticas e desconectadas das necessidades reais dos municípios beneficiados (Ames, 2001; Samuels, 2002), com quatro padrões problemáticos: alocação guiada por interesses eleitorais, em que redutos recebem recursos desproporcionais à necessidade objetiva (Limongi; Mesquita, 2008; Sodré; Alves, 2010); ausência de diagnóstico prévio das lacunas municipais; desconexão com o planejamento federal; e concentração geográfica — entre os 640 parlamentares analisados neste trabalho, o coeficiente de Gini geográfico médio das emendas é de 0,6039, caracterizando alta desigualdade na distribuição territorial. Esse padrão representa um duplo desperdício: do recurso público, que não chega onde mais faz falta, e da oportunidade política, já que o parlamentar poderia alocar de forma mais estratégica tanto para o impacto social quanto para seu retorno eleitoral.

Dessa dupla oportunidade decorre o problema de pesquisa: **como um ambiente analítico baseado em dados pode apoiar decisões mais racionais e efetivas de alocação de emendas parlamentares, conectando o comportamento orçamentário dos parlamentares às necessidades reais dos municípios e ao planejamento do governo federal?**

O objetivo geral é desenvolver e avaliar empiricamente um ambiente virtual integrado — fundamentado em dados de emendas parlamentares (SIOP), votações (TSE) e indicadores municipais (IBGE, DATASUS, SISAB) — capaz de apoiar, simultaneamente, decisões de alocação de emendas por parlamentares e a formulação de políticas públicas indutoras pelo governo federal. São cinco os objetivos específicos: (1) mapear o padrão de alocação de 640 parlamentares (2020–2023), identificando perfis tipológicos; (2) investigar correlações entre comportamento alocativo e desempenho eleitoral nos ciclos de 2018 e 2022; (3) diagnosticar as necessidades dos 5.571 municípios brasileiros em cinco áreas estratégicas (saúde, educação, segurança pública, desenvolvimento regional e assistência social); (4) propor 25 políticas públicas de indução de emendas pelo Modelo MIRE, alinhadas ao PPA 2024–2027; e (5) construir um ambiente virtual com cinco *dashboards* interativos integrando diagnóstico, recomendação e articulação.

A pesquisa tem por objeto as emendas individuais impositivas — que totalizaram R$ 24,6 bilhões na Lei Orçamentária Anual (LOA) 2025 (IFI, 2024) — apresentadas por deputados federais e senadores do Congresso Nacional, registradas no Sistema Integrado de Planejamento e Orçamento (SIOP), da Secretaria de Orçamento Federal do Ministério do Planejamento e Orçamento, cruzadas com os resultados eleitorais do Tribunal Superior Eleitoral (TSE). A justificativa é tripla. Empírica: os estudos disponíveis analisam o comportamento parlamentar de forma agregada, sem conectar sistematicamente alocação orçamentária, desempenho eleitoral individual e necessidade municipal — lacuna confirmada por revisão sistemática do Ipea (Almeida; Dominguez, 2025). Institucional: a LC 210/2024 consolidou a impositividade e criou uma realidade orçamentária que demanda instrumentos de governança ainda inexistentes. Prática: o ambiente oferece ao parlamentar um espelho do próprio comportamento com benchmarks e recomendações, e ao governo federal um sistema de sugestão de políticas com identificação dos parlamentares mais propensos a patrociná-las.

O artigo está organizado em mais cinco seções: referencial teórico, metodologia, resultados, discussão e conclusão.

# 1. REFERENCIAL TEÓRICO

## 1.1 Emendas parlamentares no Brasil: da discricionariedade à impositividade

Nos primeiros anos da Constituição de 1988, as emendas operavam sob discricionariedade plena do Executivo: a execução seletiva funcionava como moeda de troca do presidencialismo de coalizão (Abranches, 1988), e a filiação à coalizão governista era o principal determinante da execução das emendas individuais (Limongi; Figueiredo, 2005; Baião; Couto; Jucá, 2018). A ruptura veio com a EC 86/2015, que tornou obrigatória a execução das emendas individuais; após sua vigência, o pertencimento à coalizão perdeu relevância estatística como variável explicativa da execução (Bonfim; Luz; Vasquez, 2023). Reformas subsequentes ampliaram o escopo — impositividade das emendas de bancada (EC 100/2019), transferências especiais (EC 105/2019), elevação do teto para 2% da receita corrente líquida (EC 126/2022) —, e a LC 210/2024 promoveu a síntese normativa, mantendo a vinculação mínima de 50% das emendas individuais à saúde. Santos e Gasparini (2020) enquadram o movimento como alteração estrutural na natureza jurídica da lei orçamentária, convertida em arranjo de cogestão; Pereira e Mueller (2004) ponderam que o Executivo mantém instrumentos alternativos de coordenação. Este trabalho adota a primeira premissa: a contração de 47% do espaço discricionário em uma década é mudança de magnitude qualitativa.

As consequências são documentadas pela literatura recente: a Instituição Fiscal Independente (IFI, 2024) alerta que as emendas se tornaram o maior bloco das despesas discricionárias da União; Fleury et al. (2025) denominam o fenômeno de "sequestro da política", com repercussões sobre o pacto federativo; Ten Cate (2023) demonstra que a distribuição das emendas não acompanha indicadores de necessidade objetiva; e os recursos do SUS oriundos de emendas saltaram de R$ 5,1 bilhões (2014) para R$ 24,8 bilhões (2024) (Vieira, 2024). Almeida e Dominguez (2025) concluem que instrumentos de inteligência orçamentária capazes de informar as decisões de alocação constituem lacuna relevante — precisamente a que este trabalho se propõe a preencher.

## 1.2 Avaliação de políticas públicas e o Modelo MIRE

O orçamento impositivo criou uma responsabilidade nova para o governo federal: formular políticas que possam ser financiadas por emendas — ou seja, o governo não controla os recursos, ele os *induz*. O Modelo MIRE (Modelo Integrado de Resultados e Evidências), proposto neste trabalho, articula três marcos clássicos da avaliação de políticas: a cadeia de resultados da gestão por resultados (RBM ), paradigma dominante de formulação (Mayne, 2007; PNUD, 2009); a Matriz Lógica (LFA), que organiza objetivos, indicadores verificáveis, meios de verificação e premissas (Norad, 1999; Cassiolato; Gueresi, 2010); e a Teoria da Mudança, que explicita o caminho causal e gradua a confiança nas premissas (Weiss, 1995). A inovação do MIRE é tratar a decisão parlamentar de adesão como premissa graduada (alta/média/baixa), quantificada por scores de afinidade calculados a partir dos dados — dimensão ausente nos marcos originais, que pressupõem formulador e executor como o mesmo agente. O modelo não substitui os marcos dos quais deriva: integra-os e os contextualiza para um cenário em que a execução depende de agente externo e autônomo, tornando a viabilidade política dimensão estrutural da formulação.

### 1.3 Comportamento orçamentário e conexão eleitoral

A literatura de conexão eleitoral explica as escolhas alocativas a partir dos incentivos do sistema eleitoral. Mayhew (1974) identificou que parlamentares são maximizadores de reeleição, com preferência por gastos territorialmente concentrados (*pork barrel*) e de visibilidade imediata (*credit claiming*). No Brasil, Ames (2001) demonstrou a lógica territorialista do comportamento parlamentar, amplificada pela representação proporcional em lista aberta, e Samuels (2002) mostrou que a sobreposição entre a lógica distributiva das emendas e a lógica financeira do ciclo eleitoral incentiva a concentração de recursos em redutos. Essa tese foi matizada por estudos posteriores: Carvalho (2003) mostrou que parcela significativa dos deputados adota estratégia de voto disperso, e Pereira e Mueller (2004) argumentam que a conexão eleitoral opera mais por redes locais do que pela alocação direta. Este trabalho incorpora a crítica ao não pressupor que o padrão concentrador é universal — heterogeneidade refletida nos quatro perfis tipológicos adiante.

A quantificação do comportamento alocativo combina três métricas complementares (fórmulas no Apêndice A): o Índice de Herfindahl-Hirschman (HHI) geográfico (Herfindahl, 1950; Hirschman, 1964), sensível à dominância de poucos municípios — ideal para identificar o *pork barrel* clássico; o coeficiente de Gini geográfico (Gini, 1912), que captura a desigualdade geral da distribuição entre todos os municípios atendidos; e a entropia de Shannon normalizada (Shannon, 1948), que mede a diversidade temática da alocação entre funções orçamentárias. Enquanto HHI e Gini respondem “onde?”, a entropia responde “em quê?”. Da combinação desses eixos derivam quatro perfis, fundamentados na tipologia de Ames (2001) e Carvalho (2003): o **Territorialista Concentrado** (poucos municípios, poucos temas — o *pork barrel* clássico); o **Universalista Concentrado** (base territorial restrita, agenda temática diversificada); o **Temático-Setorial** (disperso no estado, focado em um setor — o voto de opinião); e o **Universalista Difuso** (ampla base territorial e múltiplas funções — típico de líderes partidários e senadores).

Até onde se pôde verificar nas bases Scopus, Web of Science e Google Scholar, não foram identificados estudos que integrem simultaneamente dados de alocação orçamentária (SIOP), resultados eleitorais individuais (TSE) e indicadores de necessidade municipal em uma plataforma analítica unificada — lacuna que a revisão do Ipea confirma ao recomendar o desenvolvimento de instrumentos de inteligência orçamentária como agenda prioritária (Almeida; Dominguez, 2025).

## 2. METODOLOGIA

### 2.1 Fontes de dados e bases analíticas

O ambiente virtual integra cinco fontes de dados oficiais (Tabela 1). Os dados de emendas foram extraídos diretamente do SIOP mediante credenciamento próprio; dados similares podem ser obtidos publicamente no Portal da Transparência e no painel Siga Brasil, do Senado Federal.

**Tabela 1 — Síntese das fontes de dados**

| Fonte | Período | Volume / abrangência | Função |
|---|---|---|---|
| SIOP/SOF | 2015–2025 (foco: 2020–2023) | 122.936 registros; 640 parlamentares | Comportamento orçamentário |
| TSE | 2018 e 2022 | Votação por município | Coorte de reeleição (441 parl.) |
| IBGE | Censo 2022; SIDRA 2022–2024 | 5.571 municípios; malha geográfica | Indicadores socioeconômicos |
| DATASUS | 2022–2024 | Indicadores de saúde municipais | Necessidade em saúde |
| SISAB | 2023–2024 | Atenção primária | Cobertura e desempenho da APS |

**Fonte:** elaboração própria.

Optou-se pelo valor **indicado** — e não pelo empenhado ou pago — porque ele reflete a decisão do parlamentar, enquanto o valor executado sofre influência de fatores exógenos (impedimentos técnicos, contingências, atrasos). A escolha introduz um viés reconhecido: superestima a presença territorial em municípios onde a emenda foi indicada mas não executada (ver seção 5.4).

Quatro bases analíticas foram construídas. A **base de emendas consolidada** padroniza os registros do SIOP (2015–2025); a normalização de nomes reduziu o conjunto a 640 parlamentares únicos (548 deputados federais e 92 senadores). A **base eleitoral integrada** conecta esses parlamentares aos resultados do TSE por nome normalizado e UF, com verificação manual dos casos ambíguos: 441 concorreram ao mesmo cargo em 2022 — 265 reeleitos (60,1%) e 176 não reeleitos (39,9%). A **base municipal de necessidades** integra IBGE, DATASUS e SISAB em índices compostos por área (escala 0–1); municípios com índice acima de 0,55 (0,45 em segurança pública, área de menor dispersão) foram classificados como críticos — 4.327 municípios em pelo menos uma área. O limiar é critério operacional, não fronteira natural: análise de sensibilidade com limiares alternativos (0,50 e 0,60) mostrou que cerca de 87% dos municípios críticos permanecem no grupo em ambos os cenários. A **base de diagnóstico parlamentar** consolida 35 variáveis por parlamentar da 57ª Legislatura, em quatro blocos: identificação, comportamento orçamentário, concentração geográfica e conexão eleitoral.

## *2.2 Métricas desenvolvidas*

Quatro grupos de métricas transformam as bases em informação interpretável (definições formais no Apêndice A; benchmarks completos no Apêndice C). O primeiro grupo mede a **concentração geográfica**: HHI geográfico (mediana 0,071 na base 2024–2025), Gini geográfico (mediana 0,576) e percentual alocado nos cinco principais municípios. O HHI alimenta o sistema de semáforo dos *dashboards*: abaixo de 0,08, dispersão adequada; acima de 0,30, concentração elevada.

O segundo grupo conecta alocação e **base eleitoral**: o percentual de sobreposição emendas-votos (mediana 12,9% — metade dos parlamentares cobre com emendas menos de 13% dos municípios onde tem votos); a cobertura eleitoral, ponderada por votos (mediana 62,9% — indicando que os parlamentares cobrem os municípios onde concentram votos, mas deixam

descobertos muitos municípios menores); o índice de adesão geográfica, calculado como correlação de Spearman entre os vetores municipais de emendas e de votos (Spearman, 1904); e o valor por voto (R$/voto), métrica de eficiência da presença territorial.

O terceiro grupo é o **perfil tipológico**: dois eixos contínuos (0–100) — score de concentração geográfica, derivado do HHI combinado de emendas e votos, e score de diversificação temática, derivado da entropia de Shannon sobre funções orçamentárias —, com limiar de 50 definindo os quatro quadrantes da seção 2.3. O cálculo da entropia desconta o componente obrigatório de 50% em saúde, de modo que apenas a parcela voluntária da alocação contribua para o score — sem esse ajuste, praticamente todos os parlamentares seriam classificados como setoriais em saúde. A classificação final pondera dados históricos (60%) e um questionário de autopercepção de 10 itens (40%), reconhecendo que a intenção estratégica declarada complementa o comportamento observado; uma análise de convergência sinaliza divergências entre ambos.

O quarto grupo é o **score de afinidade** parlamentar × política pública (0–100), composto por percentual de emendas na área temática (35%), valor indicado em 2024–2025 (25%), municípios atendidos na área (20%) e alinhamento de perfil (20%), classificando os parlamentares em alto (≥ 65), médio (35–64) e baixo (< 35) potencial de adesão.

Como referencial interpretativo, os *dashboards* comparam cada parlamentar às medianas dos 265 reeleitos em 2022 — os *benchmarks dos reeleitos* (por exemplo: HHI 0,058; cobertura eleitoral 77,3%; 9 funções orçamentárias distintas; R$ 811 por voto). Os benchmarks não são metas normativas, mas padrões empíricos associados a um resultado observável; sua validação estatística é objeto da seção seguinte.

## 2.3 Estudo de validação — coorte de reeleição

Dos 640 parlamentares com emendas em 2020–2023, 441 concorreram à reeleição ao mesmo cargo em 2022, dos quais 265 (60,1%) tiveram sucesso — coorte equilibrada (razão 1,5:1) que confere confiabilidade aos testes. A subcoorte de variação de votos compreende 394 deputados federais com dados válidos em 2018 e 2022. Foram utilizadas 24 variáveis de comportamento orçamentário, além de controles distritais e institucionais.

Os testes univariados de Mann-Whitney (Tabela 2) revelam diferenças significativas em 10 das 24 variáveis.

**Tabela 2 — Comparação univariada entre reeleitos e não reeleitos (N = 441)**

| Variável | Mediana reeleitos | Mediana não reeleitos | p-valor | Cohen's d |
|---|---|---|---|---|
| % Top 5 municípios | 41,8% | 51,1% | <0,0001*** | -0,525 |
| R$/voto | R$ 811 | R$ 1.116 | <0,0001*** | -0,494 |
| N. beneficiários/ano | 44,8 | 33,8 | <0,0001*** | 0,434 |
| Score de concentração | 6,3 | 10,6 | <0,0001*** | -0,326 |
| N. municípios/ano | 34,3 | 26,5 | <0,0001*** | 0,395 |
| HHI geográfico | 0,058 | 0,089 | <0,0001*** | -0,250 |
| Gap emendas-votos | 0,0030 | 0,0050 | 0,0002*** | -0,356 |

| Variável | Mediana reeleitos | Mediana não reeleitos | p-valor | Cohen's d |
|---|---|---|---|---|
| Valor/beneficiário | R$ 667 mil | R$ 871 mil | 0,0018** | -0,291 |
| Gini geográfico | 0,598 | 0,630 | 0,0093** | -0,247 |
| Valor total indicado | R$ 82,7 mi | R$ 82,7 mi | 0,0133* | 0,412 |

**Fonte:** elaboração própria. Significância: * **p<0,001**; p<0,01; * p<0,05.

A dispersão geográfica (HHI, Gini, % Top 5) e a cobertura eleitoral (gap emendas-votos) são as dimensões que melhor diferenciam os parlamentares de sucesso no plano univariado. Como as variáveis são correlacionadas e o volume financeiro pode atuar como fator de confusão, estimaram-se modelos multivariados complementares. O **modelo de regressão logística penalizada de Firth** (N = 395; erros-padrão por bootstrap agrupado por UF, 1.000 replicações), adequado a dados observacionais de amostra moderada, identifica os preditores independentes da reeleição (Tabela 3).

**Tabela 3 — Modelo Firth parcimonioso (N = 395, bootstrap agrupado por UF)**

| Variável | Coeficiente | p-valor | Odds ratio | Significância |
|---|---|---|---|---|
| Valor total indicado | 0,529 | 0,007 | 1,698 | ** |
| N. ações distintas | 0,780 | 0,026 | 2,180 | * |
| % Top 5 municípios | -0,156 | 0,687 | 0,856 | n.s. |
| N. municípios/ano | -0,208 | 0,669 | 0,812 | n.s. |
| R$/voto | -0,209 | 0,423 | 0,812 | n.s. |
| Magnitude da UF | -0,371 | 0,276 | 0,690 | n.s. |
| Senador (controle) | -0,179 | 0,237 | 0,837 | n.s. |

**Fonte:** elaboração própria. Significância: * *p<0,01;* p<0,05; n.s. = não significativo.

Isolado o efeito das demais variáveis, apenas o volume total indicado (OR = 1,698) e o número de ações orçamentárias distintas (OR = 2,180) permanecem preditores independentes: a dispersão geográfica e a cobertura eleitoral — altamente significativas no plano univariado — são mediadas pelo volume, enquanto a diversificação temática tem impacto independente na sobrevivência eleitoral. Três modelos complementares delimitam o alcance preditivo: a regressão logística Ridge com as 24 variáveis (validação cruzada estratificada em 5 dobras) obteve AUC-ROC de 0,673 ± 0,059 — capacidade discriminativa moderada; os modelos lineares Lasso e Ridge para a variação contínua de votos obtiveram $R^2$ negativo, com o Lasso zerando todos os coeficientes — nenhuma variável orçamentária prevê a variação marginal de votos; e a Random Forest (AUC-ROC 0,668), interpretada por SHAP values (Lundberg; Lee, 2017), consolidou o ranking de importância, liderado pela diversificação temática, pelo volume total e pelo alinhamento emendas-votos (Apêndice C). Na subcoorte de variação de votos (N = 394), os deputados que ganharam votos exibem maior dispersão geográfica (HHI mediano 0,097 vs. 0,129; p = 0,004) e maior alinhamento emendas-votos (p = 0,015), e a correlação positiva entre crescimento eleitoral e R$/voto (rho = +0,223; p < 0,0001) sinaliza investimento de expansão em novas franjas territoriais.

### 2.4 Elaboração das propostas de políticas públicas

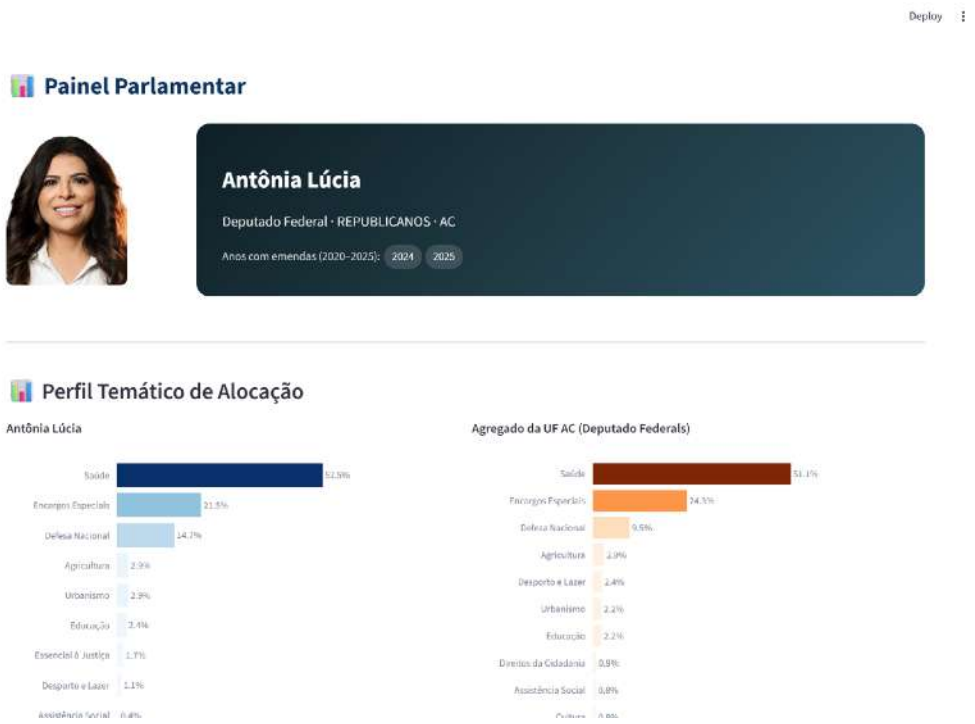
A formulação das 25 propostas seguiu processo sequencial orientado por dados em cinco etapas (detalhamento no Apêndice B): (1) alinhamento ao PPA 2024–2027, com extração de 32 programas finalísticos e 159 objetivos específicos nas cinco áreas; (2) quantificação das necessidades municipais (4.327 municípios críticos); (3) identificação de patrocinadores potenciais (626 perfis com score de afinidade por área); (4) seleção do modelo de construção, após avaliação comparativa de dez alternativas em que RBM (4,60), LFA (4,25) e ToC (4,20) lideraram — combinados no Modelo Híbrido MIRE; e (5) formulação das fichas. Cada proposta é estruturada em ficha-padrão de 13 seções que operacionaliza as três camadas do MIRE, incluindo o mecanismo de indução e a lista de parlamentares patrocinadores. O limiar de score de afinidade para patrocinadores foi fixado em $\geq 20$ após análise de sensibilidade — o limiar original de 40 era excessivamente restritivo (em educação, classificaria apenas dois parlamentares). A priorização das políticas usa score composto por magnitude do problema (30%), instrumentalidade das emendas (25%), potencial de mobilização parlamentar (25%) e alinhamento ao PPA (20%).

## 3. RESULTADOS

### 3.1 O artefato: ambiente virtual de inteligência orçamentária

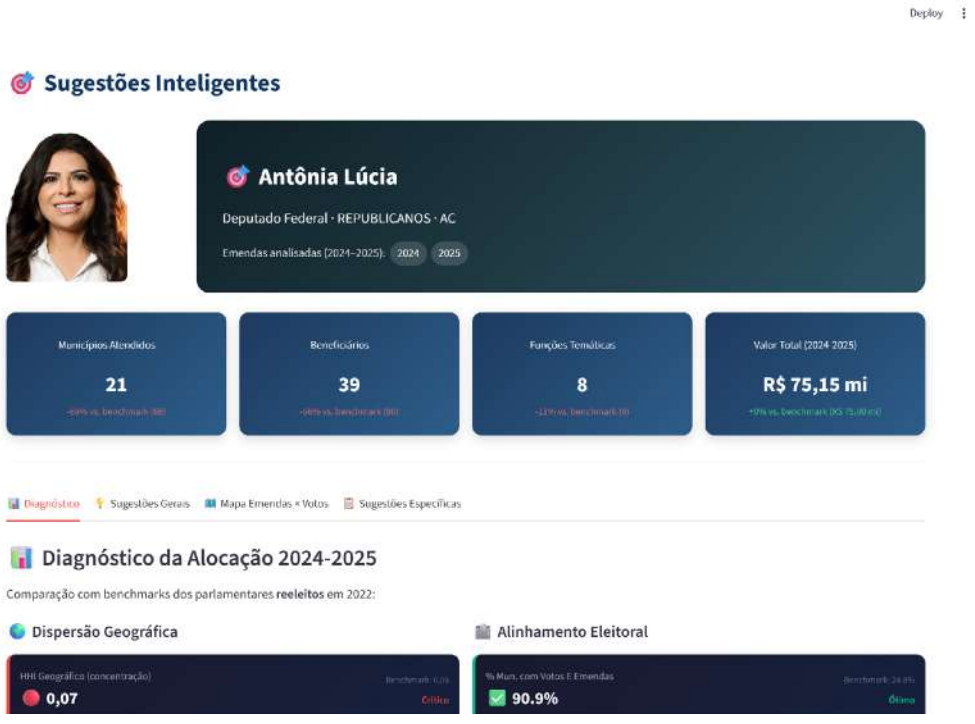
O artefato é um sistema interativo em Python/Streamlit com cinco módulos em dois ambientes — o Ambiente Parlamentar e o Ambiente de Governo Federal —, alimentados pelas mesmas bases analíticas. Três princípios orientam o design: acessibilidade analítica (métricas complexas sempre acompanhadas de semáforo e linguagem natural), rastreabilidade metodológica (todo valor é calculável a partir das bases oficiais documentadas) e dualidade de público-alvo. No Ambiente Parlamentar, o **Painel Parlamentar** (Figura 1) apresenta o diagnóstico do comportamento orçamentário — perfil temático comparado à UF, mapas coropléticos de alocação e votação, indicadores comparados; a **Anamnese Parlamentar** combina questionário de autopercepção e dados históricos para a classificação tipológica; e as **Sugestões Inteligentes** (Figura 2) entregam o diagnóstico por semáforo em três grupos (dispersão geográfica, diversificação temática, alinhamento eleitoral), com 14 métricas comparadas aos benchmarks dos reeleitos, lista de municípios prioritários (com votos e sem emendas) e mapa comparativo emendas × votos. No Ambiente de Governo, o **Painel Governo Federal** oferece a visão macro nacional (filtros por função, partido, UF e ano; rankings; distribuição de perfis por bancada estadual); e as **Sugestões de Políticas Públicas** (Figura 3) conectam cada uma das 25 políticas aos patrocinadores potenciais ranqueados por score de afinidade, com municípios contemplados, teoria da mudança e potencial de recursos.

## Figura 1 — Painel Parlamentar: perfil temático de alocação comparado ao agregado da UF



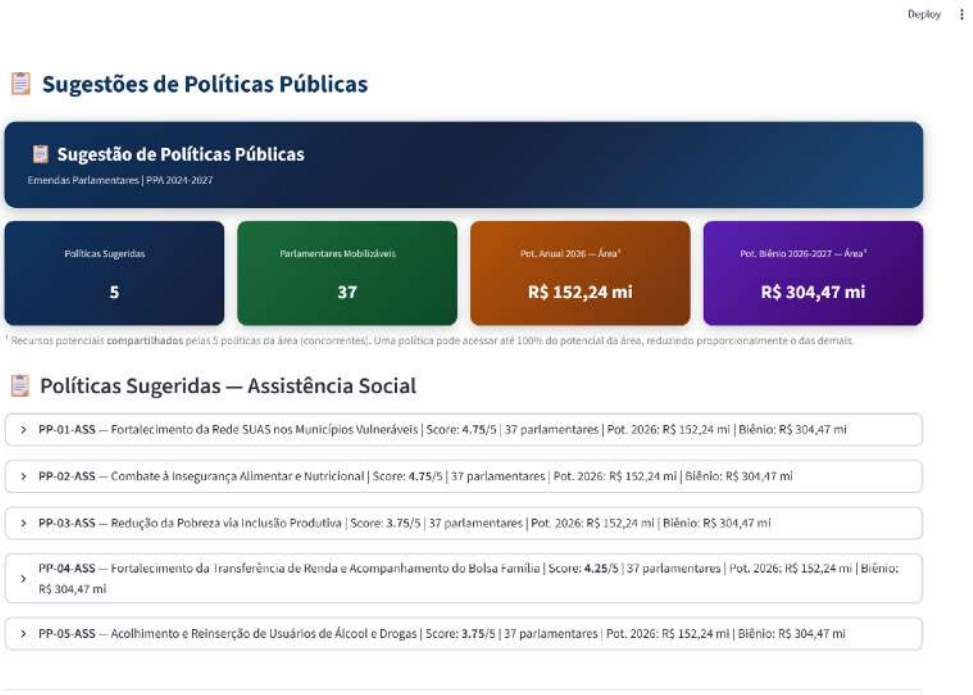

**Fonte:** elaboração própria (Ambiente Virtual de Inteligência Orçamentária, dados SIOP/SOF, 2025).

## Figura 2 — Sugestões Inteligentes: diagnóstico da alocação comparado aos *benchmarks* dos reeleitos




**Fonte:** elaboração própria (Ambiente Virtual de Inteligência Orçamentária, dados SIOP/SOF e TSE, 2025).

**Figura 3 — Sugestões de Políticas Públicas: propostas, patrocinadores e potencial de recursos**




**Fonte:** elaboração própria (Ambiente Virtual de Inteligência Orçamentária, dados SIOP/SOF e PPA 2024–2027).

## *3.2 Perfis parlamentares*

A classificação dos 640 parlamentares revela distribuição assimétrica (Tabela 4): o Universalista Difuso predomina (38,3%), refletindo o incentivo do sistema proporcional de lista aberta à presença territorial ampla — proporção maior entre senadores (~52%) do que entre deputados (~34%), coerente com o mandato pan-estadual. O Territorialista Concentrado (26,1%) confirma que o *pork barrel* geográfico permanece relevante.

**Tabela 4 — Distribuição dos perfis tipológicos (n = 640)**

| Perfil | N | % | Score concentração (mediana) | Score temático (mediana) |
|---|---|---|---|---|
| Universalista Difuso | 245 | 38,3% | 8,2 | 61,4 |
| Territorialista Concentrado | 167 | 26,1% | 63,7 | 28,3 |
| Temático-Setorial | 124 | 19,4% | 7,5 | 31,8 |
| Universalista Concentrado | 104 | 16,2% | 58,9 | 62,1 |

**Fonte:** elaboração própria.

O resultado mais contraintuitivo é que o perfil categórico não diferencia reeleitos de não reeleitos ($x^2$ = 0,75; p = 0,69), enquanto os scores contínuos diferenciam (score de concentração dos que ganharam votos é significativamente inferior; p = 0,039). A discretização em quadrantes descarta informação relevante — o que orientou a decisão de design de priorizar métricas contínuas e semáforo sobre rótulos tipológicos nos *dashboards*.

## 3.3 Fatores associados ao desempenho eleitoral

Os testes da seção 3.3 convergem para cinco achados centrais. Primeiro, a dispersão geográfica é a associação univariada mais robusta: não reeleitos concentram em média 54,2% dos recursos nos cinco principais municípios, contra 44,0% dos reeleitos ($p < 0,0001$; $d = -0,525$). Segundo, o alinhamento espacial emendas-votos é altamente significativo ($p = 0,0002$): parlamentares de sucesso direcionam recursos onde possuem eleitores. Terceiro, no modelo multivariado, apenas o volume total indicado (OR = 1,698) e a diversificação temática (OR = 2,180) são preditores independentes — a amplitude temática tem impacto que independe da escala de recursos. Quarto, reeleitos apresentam menor custo por voto (R$ 811 vs. R$ 1.116; $p < 0,0001$), sinalizando eficiência alocativa. Quinto, a capacidade preditiva geral é moderada (AUC-ROC = 0,673) e a variação contínua de votos não é explicada por modelos lineares ($R^2$ negativo) — as variáveis orçamentárias apoiam o diagnóstico, mas não determinam o desfecho eleitoral.

## 3.4 Diagnóstico municipal e propostas de políticas públicas

Dos 5.571 municípios brasileiros, 4.327 (77,7%) são críticos em pelo menos uma das cinco áreas: assistência social (1.618), desenvolvimento regional (1.584), educação (928), saúde (133) e segurança pública (64). A tensão central do diagnóstico é que as áreas de maior urgência individual (saúde, segurança) têm menor abrangência territorial, enquanto as de maior abrangência (assistência social, desenvolvimento regional) têm menor viabilidade parlamentar — como sintetiza a Tabela 5.

**Tabela 5 — Síntese das 25 propostas de políticas públicas**

| Área | Políticas | Parlamentares mobilizáveis | Potencial anual | Municípios críticos |
|---|---|---|---|---|
| Saúde | 5 | 405 | R$ 9,21 bi | 133 |
| Educação | 5 | 39 | R$ 129 mi | 928 |
| Segurança pública | 5 | 23 | R$ 60 mi | 64 |
| Desenvolvimento regional | 5 | 111 | R$ 538 mi | 1.584 |
| Assistência social | 5 | 37 | R$ 152 mi | 1.618 |
| **Total** | **25** | **615** | **R$ 10,09 bi** | **4.327** |

**Fonte:** elaboração própria.

O domínio da saúde (89,8% do potencial) é inflado pela vinculação obrigatória de 50% — o potencial de expansão voluntária é significativamente inferior. A educação exibe a maior lacuna entre necessidade (928 municípios) e mobilização (39 patrocinadores); a assistência social, o paradoxo mais agudo: 44 municípios críticos por parlamentar mobilizável. São precisamente essas as áreas em que a articulação governamental — apoiada pelo *score* de afinidade — é mais necessária.

## 3.5 Recomendações de alocação baseadas em evidências

As recomendações derivam de padrões associativos, não de relações causais validadas experimentalmente: indicam o que os reeleitos fazem *diferente*, não o que *causa* a reeleição, e devem ser lidas como hipóteses informadas para reflexão estratégica (Tabela 6).

**Tabela 6 — Síntese das seis recomendações e *benchmarks* dos reeleitos**

| # | Recomendação | Benchmark principal (mediana) | Evidência |
|---|---|---|---|
| R1 | Dispersar geograficamente sem fragmentar | HHI ≤ 0,058; % Top 5 ≤ 41,8% | □□□□□ |
| R2 | Alinhar a alocação à base eleitoral | Cobertura eleitoral ≥ 77,3% | □□□□□ |
| R3 | Diversificar temas além da saúde obrigatória | ≥ 9 funções; ≥ 16 ações | □□□□□ |
| R4 | Manter consistência anual | ≥ 44,8 beneficiários/ano | □□□□□ |
| R5 | Preferir investimento a custeio | GND 4 ≥ 46,5% | □□□□□ |
| R6 | Monitorar o custo por voto | R$/voto ≤ R$ 811 | □□□□□ |

**Fonte:** elaboração própria.

Em síntese: dispersar sem pulverizar (o valor médio por município deve sustentar ações identificáveis); cobrir a base eleitoral (o mapa emendas × votos e a lista de municípios descobertos operacionalizam a recomendação); alocar o piso obrigatório em saúde e distribuir estrategicamente a parcela voluntária por sete a oito outras funções — a diversificação é o preditor independente de maior odds ratio; evitar oscilações bruscas de volume entre os anos do mandato; priorizar investimentos (GND 4), que geram marcos tangíveis de entrega; e tratar o R$/voto como indicador contextual — valores 50% acima do benchmark acionam alerta, mas podem refletir expansão territorial deliberada.

# 4. DISCUSSÃO

## 4.1 Implicações teóricas

Três achados têm implicação teórica. Primeiro, o conceito de **dispersão alinhada**: a distinção entre a geometria da distribuição (dispersão simples) e sua sobreposição com a base eleitoral resolve uma contradição da literatura — por que nem todos os parlamentares dispersos têm sucesso. A dimensão relevante não é quanto se dispersa, mas o grau de sobreposição entre emendas e eleitores, capturado pelo índice de adesão geográfica. Segundo, o **paradoxo do R$/voto**: o mesmo indicador tem significado oposto conforme o contexto — ineficiência em base restrita para não reeleitos, investimento em novas bases para reeleitos em expansão —, assimetria consistente com a coordenação estratégica de eleitores e candidatos descrita por Cox (1997); o R$/voto não é, portanto, indicador monotônico de eficiência. Terceiro, as **variáveis orçamentárias operam como condição necessária, não suficiente**: o $R^2$ negativo dos modelos lineares e a AUC-ROC de 0,673 delimitam seu espaço de influência como pré-requisito de sobrevivência eleitoral, não motor de crescimento de votos — interpretação coerente com a distinção de Samuels (2002) entre *credit claiming* (retenção) e *advertising* (conquista). A hierarquia entre evidência univariada e multivariada acrescenta nuance: os fatores geográficos são mediados pelo volume orçamentário, enquanto a diversificação temática tem efeito que independe da escala.

### 4.2 Implicações para a gestão pública

Para o parlamentar, o artefato oferece diagnóstico quantificado do próprio comportamento, análise de convergência entre estratégia declarada e observada, e lista de municípios descobertos como insumo direto de alocação — permitindo planejar proativamente em vez de responder reativamente às demandas que chegam ao gabinete. Para o gestor federal, oferece inteligência de articulação: o *score* de afinidade funciona como pré-filtro das negociações (o gestor com 30 minutos de agenda investe esse tempo no parlamentar do topo da lista), complementado pelo diagnóstico tipológico das bancadas estaduais e pela estimativa de potencial por política. Para a governança, os dados documentam que a lógica de alocação não converge com a necessidade territorial; as 25 políticas estruturadas oferecem alternativa à fragmentação — convergência voluntária de parlamentares para municípios prioritários —, e recomenda-se a criação de índices públicos de desempenho territorial das emendas, publicados anualmente pela Secretaria de Orçamento Federal, sintetizando concentração geográfica, alinhamento às necessidades socioeconômicas e cobertura da base eleitoral.

### 4.3 Limitações

Sete limitações delimitam o alcance dos resultados. (i) *Leakage* temporal: as métricas de alinhamento espacial usam a votação de 2022, contemporânea ao desfecho, circularidade que pode inflar a capacidade discriminativa dos classificadores; a validação preditiva pura exigiria métricas construídas apenas com a votação de 2018. (ii) Valor indicado vs. executado: a opção pelo valor indicado superestima a presença territorial onde a emenda não foi executada. (iii) Validação no sentido de Design Science Research, não causal: os modelos identificam associações robustas, não efeitos de tratamento. (iv) Variáveis não observadas: carisma, escândalos, alianças locais, conjuntura macroeconômica e recursos de campanha não estão nas bases integradas. (v) Especificidade temporal: o período 2020–2023 foi marcado pela pandemia de Covid-19 e pelo regime do orçamento secreto, limitando a generalização para ciclos sob a LC 210/2024. (vi) Heterogeneidade distrital: *benchmarks* nacionais simplificam realidades muito distintas (de 70 a 8 cadeiras por UF), atenuadas pela comparação com médias regionais e da UF. (vii) Risco de adaptação adversa (lei de Goodhart): se as métricas guiarem a conduta de forma excessivamente explícita, parlamentares podem otimizar os indicadores formais, esvaziando sua utilidade como medida de representação real.

## CONCLUSÃO

Os cinco objetivos foram cumpridos: 640 parlamentares classificados em quatro perfis (com a constatação de que os scores contínuos são mais informativos que os rótulos categóricos); a associação sistemática entre dispersão geográfica, alinhamento emendas-votos, diversificação temática e sucesso eleitoral, com delimitação honesta do alcance preditivo; 4.327 municípios críticos diagnosticados, com a assimetria entre necessidade e mobilização mais aguda em assistência social e educação; 25 políticas formuladas pelo Modelo MIRE, com potencial de R$ 10,09 bilhões anuais e 615 patrocinadores identificados; e o ambiente virtual operacional com cinco módulos sobre 122.936 registros de emendas.

As contribuições são de três ordens. Teóricas: a dispersão alinhada como conceito operacional; a demonstração de que a discretização tipológica perde poder discriminativo; e a caracterização das variáveis orçamentárias como condição necessária, não suficiente,

do sucesso eleitoral. Metodológicas: o Modelo MIRE como adaptação replicável de RBM, LFA e ToC a contextos em que a execução depende de agente externo; a normalização do viés de saúde na diversificação temática; o índice de adesão geográfica como métrica original; e a integração inédita das bases do SIOP, TSE, IBGE, DATASUS e SISAB. Práticas: o primeiro sistema público de diagnóstico individualizado do comportamento orçamentário parlamentar e de inteligência de articulação baseada em score de afinidade. A agenda futura inclui a validação com o ciclo 2024–2026, benchmarks estratificados por porte de UF, desenhos quase-experimentais para inferência causal, a incorporação de dados de execução financeira e a mitigação do *leakage* temporal com métricas estritamente pré-tratamento.

O sistema de emendas parlamentares não é, por natureza, um problema de dados — é um problema de incentivos, de governança e de relações entre os Poderes. Mas é também um problema em que a ausência de informação estruturada alimenta decisões subótimas de ambos os lados. Este trabalho demonstrou que é possível preencher essa lacuna com dados que já existem, ferramentas abertas e metodologia replicável. O achado mais fundamental é que os parlamentares que alocam onde a necessidade existe — cobrindo mais municípios, diversificando temas, alinhando emendas à base eleitoral — são os mesmos que, em mediana, têm mais sucesso eleitoral: no sistema representativo, fazer bem o trabalho de representar o eleitorado e construir uma base eleitoral sustentável podem ser, mais do que se imagina, a mesma coisa.

## Disponibilidade de Dados e Código

Em conformidade com os princípios de Ciência Aberta, o material necessário à reprodução dos resultados está disponível publicamente.

**Código-fonte (Open Code).** Todo o código do ambiente virtual — *scripts* de processamento das bases, cálculo das métricas e aplicação interativa em Python/Streamlit — está depositado no repositório público: https://github.com/EdudardoM/intel-orca-open. O repositório contém um arquivo *README.md* com instruções de instalação, execução local e reprodução dos resultados.

**Aplicação em produção.** A versão de demonstração do ambiente está disponível em https://inteligencia-orcamentaria.onrender.com (URL legada: https://intel-orca.streamlit.app).

**Dados (Open Data).** As bases analíticas consolidadas — base de emendas (SIOP, 2020–2025), base eleitoral integrada (TSE, 2018 e 2022), base municipal de necessidades (IBGE, DATASUS, SISAB) e base de diagnóstico parlamentar —, em formato processado e documentado, estão depositadas em repositório de acesso aberto com DOI: https://doi.org/10.5281/zenodo.20856157 — registro direto em https://zenodo.org/records/20856157 (Zenodo). Os dados de emendas foram extraídos do SIOP/SOF mediante credenciamento próprio; dados equivalentes são públicos no Portal da Transparência (https://portaldatransparencia.gov.br) e no painel Siga Brasil, do Senado Federal. As demais bases (TSE, IBGE, DATASUS, SISAB) são públicas e de acesso aberto. Os dados não contêm informações pessoais sensíveis na acepção da Lei Geral de Proteção de Dados (LGPD — Lei n. 13.709/2018): referem-se a agentes públicos no exercício do mandato e a valores orçamentários e resultados eleitorais agregados por município, todos de natureza pública.

## *Declaração de Uso de Inteligência Artificial*

Em atendimento às diretrizes para autores em trabalhos com uso de IA, declara-se que ferramentas de inteligência artificial generativa (Claude, da Anthropic) foram utilizadas, com finalidade descrita por etapa, como apoio a: (i) condensação editorial do texto a partir do trabalho de conclusão de curso original dos autores, para adequação ao formato de artigo; (ii) formatação do manuscrito segundo as normas editoriais do livro (Times New Roman 12, espaçamento 1,5, recuo de 1,25 cm, padrão ABNT); (iii) revisão de consistência terminológica e de coerência interna do texto; e (iv) apoio à redação da documentação técnica do repositório (*README.md*). O desenho da pesquisa, a coleta e o processamento dos dados, a construção das métricas, os modelos estatísticos, os resultados, as interpretações e as conclusões são de autoria e responsabilidade exclusivas dos autores, que revisaram e validaram integralmente todo o conteúdo final. Nenhum dado, referência bibliográfica ou resultado empírico foi gerado por inteligência artificial.



## REFERÊNCIAS


ABRANCHES, S. H. H. de. Presidencialismo de coalizão: o dilema institucional brasileiro. **Dados: Revista de Ciências Sociais**, Rio de Janeiro, v. 31, n. 1, p. 5-34, 1988.

ALMEIDA, A.; DOMINGUEZ, M. **Gastos públicos por emendas orçamentárias: uma revisão de literatura.** Brasília: Ipea, 2025. (Relatório de Pesquisa, RP-236239). DOI: 10.38116/RP-236239.

AMES, B. **The Deadlock of Democracy in Brazil**. Ann Arbor: University of Michigan Press, 2001.

BAIÃO, A. L.; COUTO, C. G.; JUCÁ, I. C. A execução das emendas orçamentárias individuais: papel de ministros, cargos de liderança e normas fiscais. **Revista Brasileira de Ciência Política**, Brasília, n. 25, p. 47-86, 2018. DOI: 10.1590/0103-335220182502.

BONFIM, R.; LUZ, J. H.; VASQUEZ, V. Mandatory individual amendments: a change in the pattern of executive dominance in the Brazilian budgetary and financial cycle. **Brazilian Political Science Review**, v. 17, n. 2, e0001, 2023. DOI: 10.1590/1981-3821202300020001.

BRASIL. **Emenda Constitucional n. 86, de 17 de março de 2015**. Altera os arts. 165 e 166 da Constituição Federal para tornar obrigatória a execução das emendas parlamentares individuais. Brasília: Congresso Nacional, 2015.

BRASIL. **Lei Complementar n. 210, de 22 de novembro de 2024**. Regulamenta as emendas parlamentares ao orçamento federal. Brasília: Congresso Nacional, 2024.

BRASIL. **Lei n. 14.802, de 10 de janeiro de 2024**. Institui o Plano Plurianual da União para o período de 2024 a 2027. Brasília: Presidência da República, 2024.

BRASIL. Ministério da Saúde. *DATASUS* — **Departamento de Informática do SUS**. Disponível em: https://datasus.saude.gov.br. Acesso em: fev. 2026.

BRASIL. Ministério da Saúde. **SISAB — Sistema de Informação em Saúde para a Atenção Básica**. Disponível em: https://sisab.saude.gov.br. Acesso em: fev. 2026.

CARVALHO, N. R. **E no início eram as bases: geografia política do voto e comportamento legislativo no Brasil**. Rio de Janeiro: Revan, 2003.

CASSIOLATO, M.; GUERESI, S. **Como elaborar modelo lógico: roteiro para formular programas e organizar avaliação**. Brasília: Ipea, 2010. (Nota Técnica, n. 6).

COX, G. W. **Making Votes Count: Strategic Coordination in the World's Electoral Systems**. Cambridge: Cambridge University Press, 1997.

FLEURY, S. et al. O sequestro da política: implicações das emendas parlamentares sobre a democracia, o pacto federativo e a saúde pública no Brasil. **Saúde em Debate**, Rio de Janeiro, v. 49, n. 147, 2025.

GINI, C. Variabilità e mutabilità. **Studi Economico-Giuridici della Regia Università di Cagliari**, v. 3, n. 2, p. 3-159, 1912.

HERFINDAHL, O. C. **Concentration in the U.S. Steel Industry**. 1950. Tese (Doutorado) — Columbia University, New York, 1950.

HIRSCHMAN, A. O. The paternity of an index. **American Economic Review**, v. 54, n. 5, p. 761-762, 1964.

IBGE — INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Censo Demográfico 2022: resultados do universo**. Rio de Janeiro: IBGE, 2023.

IFI — INSTITUIÇÃO FISCAL INDEPENDENTE. **Nota Técnica n. 57: despesas discricionárias e emendas parlamentares (2014-2024)**. Brasília: Senado Federal, nov. 2024.

LIMONGI, F.; FIGUEIREDO, A. C. Processo orçamentário e comportamento legislativo: emendas individuais, apoio ao Executivo e programas de governo. **Dados: Revista de Ciências Sociais**, Rio de Janeiro, v. 48, n. 4, p. 737-776, 2005. DOI: 10.1590/S0011-52582005000400002.

LIMONGI, F.; MESQUITA, L. Estratégia partidária e preferência dos eleitores. **Novos Estudos Cebrap**, n. 81, p. 57-74, 2008.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 30., 2017, Long Beach. **Proceedings...** Red Hook: Curran Associates, 2017. p. 4765-4774.

MAYHEW, D. R. **Congress: The Electoral Connection**. New Haven: Yale University Press, 1974.

MAYNE, J. Challenges and lessons in implementing results-based management. **Evaluation**, v. 13, n. 1, p. 87-109, 2007.

NORAD. **The Logical Framework Approach (LFA): handbook for objectives-oriented planning**. 4. ed. Oslo: Norwegian Agency for Development Cooperation, 1999.

PEREIRA, C.; MUELLER, B. The cost of governing: strategic behavior of the president and legislators in Brazil's budgetary process. **Comparative Political Studies**, v. 37, n. 7, p. 781-815, 2004.

PNUD. **Handbook on Planning, Monitoring and Evaluating for Development Results**. Nova York: UNDP, 2009.

SAMUELS, D. Pork barreling is not credit claiming or advertising: campaign finance and the sources of the personal vote in Brazil. **Journal of Politics**, v. 64, n. 3, p. 845-863, 2002.

SANTOS, N. C. B.; GASPARINI, C. E. Orçamento impositivo e relação entre poderes no Brasil. **Revista Brasileira de Ciência Política**, Brasília, n. 31, p. 339-396, jan./abr. 2020. DOI: 10.1590/0103-335220203109.

SHANNON, C. E. A mathematical theory of communication. **Bell System Technical Journal**, v. 27, n. 3, p. 379-423, 1948.

SIOP/SOF. **Sistema Integrado de Planejamento e Orçamento: dados de emendas parlamentares e despesas discricionárias 2013-2025**. Brasília: Secretaria de Orçamento Federal, 2025.

SODRÉ, A. C. A.; ALVES, M. F. C. Relação entre emendas parlamentares e corrupção municipal no Brasil. **Revista de Administração Contemporânea**, v. 14, n. 3, p. 414-433, 2010.

SPEARMAN, C. The proof and measurement of association between two things. **American Journal of Psychology**, v. 15, n. 1, p. 72-101, 1904.

TEN CATE, L. Quem recebe e que diferença fazem: emendas orçamentárias nos municípios brasileiros pós EC 86/2015 (2015-2019). **Opinião Pública**, Campinas, v. 29, n. 3, p. 792-826, set./dez. 2023. DOI: 10.1590/1807-01912023293792.

TSE — TRIBUNAL SUPERIOR ELEITORAL. **Portal de Dados Abertos do TSE: resultados eleitorais 2018 e 2022**. Disponível em: https://dadosabertos.tse.jus.br. Acesso em: fev. 2026.

VIEIRA, F. S. **Emendas parlamentares ao orçamento federal do Sistema Único de Saúde: 2014-2024**. Brasília: Ipea, 2024. (Relatório de Pesquisa).

WEISS, C. H. Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives. In: CONNELL, J. P. et al. (Eds.). **New Approaches to Evaluating Community Initiatives**. Washington: Aspen Institute, 1995.

## APÊNDICE A — DEFINIÇÕES FORMAIS DAS MÉTRICAS

**Índice de Herfindahl-Hirschman geográfico.** Para cada parlamentar *p*: HHI_p = Σ s2_pm, onde s_pm é a participação do município *m* no valor total indicado pelo parlamentar e a soma percorre os M municípios beneficiados. Varia de 1/M (distribuição uniforme) a 1 (concentração total). Sistema de semáforo: < 0,08 dispersão adequada; 0,08–0,15 atenção; 0,15–0,30 concentração moderada; > 0,30 concentração elevada.

**Coeficiente de Gini geográfico.** Gini_p = [2 Σ i·V(i) − (n+1) Σ V(i)] / [n Σ V(i)], onde V(i) são os valores alocados por município em ordem crescente e n o número de municípios.

**Entropia de Shannon normalizada (diversificação temática).** H_p = −Σ s_pf log□(s_pf) / log□(F), onde s_pf é a participação da função orçamentária *f* no valor total do parlamentar e F o número de funções com alocação positiva. O cálculo desconta o componente obrigatório de saúde: o valor da função saúde é reduzido ao excedente voluntário acima do piso de 50%, de modo que apenas a parcela de escolha deliberada contribua para a entropia.

**Perfil tipológico.** Eixo 1 (concentração geográfica, 0–100): min(100, HHI_combinado × 100), onde HHI_combinado é a média do HHI de emendas (50%) e do HHI de votos (50%). Eixo 2 (diversificação temática, 0–100): entropia normalizada × 100. Limiar de 50 em cada eixo define os quatro quadrantes. Ponderação da classificação final: 60% dados históricos, 40% questionário de autopercepção (Likert, 10 itens); 100% questionário para parlamentares sem histórico (confiança "baixa"); 100% dados quando o questionário não é preenchido (confiança "alta").

**Índice de adesão geográfica.** Correlação de Spearman entre o vetor de valores de emendas por município e o vetor de votos por município; varia de −1 a +1. Preferida à correlação de Pearson por não pressupor linearidade nem normalidade.

**Score de afinidade parlamentar × política.** Score (0–100) = 35% × percentual de emendas na área temática + 25% × valor indicado em 2024–2025 (normalizado) + 20% × municípios atendidos na área (normalizado) + 20% × alinhamento de perfil tipológico. Classificação: alto ≥ 65; médio 35–64; baixo < 35.

## APÊNDICE B — O MODELO MIRE: PROCESSO DE FORMULAÇÃO E FICHA-PADRÃO

**Processo em cinco etapas:** (1) Alinhamento PPA — análise dos Anexos III e VI do PPA 2024–2027 (352 páginas), extraindo 32 programas finalísticos e 159 objetivos específicos nas cinco áreas; (2) Necessidades municipais — índices compostos por área para os 5.571 municípios, identificando 4.327 críticos; (3) Perfis parlamentares — 626 perfis com score de afinidade por área e UF; (4) Seleção do modelo — avaliação comparativa de 10 modelos com cinco critérios ponderados (três mais bem classificados: RBM 4,60; LFA 4,25; ToC 4,20), combinados no Modelo Híbrido MIRE; (5) Formulação — 25 fichas MIRE com base de patrocinadores e potencial estimado.

**Ficha-padrão de 13 seções:** (1) Identificação; (2) Diagnóstico (indicadores de necessidade, municípios críticos, linha de base); (3) Alinhamento PPA; (4) Cadeia de Resultados (insumos → atividades → produtos → resultados → impacto); (5) Matriz Lógica; (6) Teoria da Mudança (premissas graduadas alta/média/baixa); (7) Mecanismo de Indução; (8) Parlamentares Patrocinadores (score ≥ 20, por UF); (9) Alcance Territorial; (10) Potencial de Recursos; (11) Indicadores de Monitoramento; (12) Score de Priorização; (13) Plano de Implementação.

**Score de priorização (0–100):** magnitude do problema público (30%) — índice de necessidade médio e número de municípios críticos; instrumentalidade das emendas (25%) — compatibilidade do tipo de gasto com as modalidades de emendas individuais (investimento GND 4 e transferências a municípios); potencial de mobilização parlamentar (25%) — número de parlamentares com score ≥ 20 e valor médio indicado na função; alinhamento normativo (20%) — objetivos do PPA 2024–2027 atendidos.

## *APÊNDICE C — BENCHMARKS DOS REELEITOS E RANKING DE IMPORTÂNCIA DAS VARIÁVEIS*

### Tabela C1 — *Benchmarks* dos parlamentares reeleitos em 2022 (medianas, N = 265)

| Métrica | *Benchmark* | Direção desejável |
|---|---|---|
| HHI geográfico | 0,058 | Menor é melhor |
| Gini geográfico | 0,60 | Menor é melhor |
| % Top 5 municípios | 41,8% | Menor é melhor |
| N. funções distintas | 9,0 | Maior é melhor |
| N. ações distintas | 16,0 | Maior é melhor |
| % Saúde | 52,1% | Contextual |
| % GND 4 (investimento) | 46,5% | Maior é melhor |
| N. beneficiários/ano | 44,8 | Maior é melhor |
| N. municípios/ano | 34,3 | Maior é melhor |
| % sobreposição emendas-votos | 24,8% | Maior é melhor |
| Cobertura eleitoral | 77,3% | Maior é melhor |
| R$/voto | R$ 811 | Menor é melhor |
| Índice de adesão geográfica | 0,65 | Maior é melhor |

**Fonte:** elaboração própria.

### Tabela C2 — *Ranking* consolidado de importância das variáveis (SHAP values, N = 395)

| Posição | Variável | Importância | Interpretação |
|---|---|---|---|
| 1 | N. ações distintas | Alta | Diversidade temática |
| 2 | Valor total indicado | Alta | Volume financeiro |
| 3 | Gap emendas-votos | Moderada | Alinhamento eleitoral |
| 4 | % Top 5 municípios | Moderada | Concentração geográfica (efeito negativo) |
| 5 | N. municípios/ano | Moderada | Dispersão territorial |
| 6 | Cobertura eleitoral | Baixa | Cobertura da base |
| 7 | Magnitude da UF | Baixa | Controle distrital |
| 8 | Senador | Baixa | Controle de cargo |

**Fonte:** elaboração própria.

# Monitoramento e Diagnóstico de Eventos Disruptivos nos Portos Brasileiros: detecção de anomalias operacionais e análise de exposição comercial

*Fábio Kouri Paim*
*Prof. Dr. Alex Lopes Pereira*

## Resumo

Este artigo propõe um sistema integrado de detecção e classificação de anomalias operacionais em portos brasileiros, articulando três bases públicas: o Anuário Estatístico Aquaviário (ANTAQ), o PortWatch/FMI e o ComexStat/MDIC. Esse cruzamento liga a atividade portuária no nível da atracação individual aos fluxos de exportação detalhados por produto, UF de origem e país de destino, conectando a operação física à cadeia produtiva exportadora. A partir de séries temporais semanais de 18 portos, modelos XGBoost com correção AR(1) aprendem o padrão operacional de cada porto, e um ensemble de três métodos estatísticos identifica desvios significativos. Um sistema de classificação dual categoriza cada anomalia quanto à sua abrangência (isolada, nacional ou global), distinguindo choques domésticos de choques no comércio marítimo global. Validado contra eventos históricos, o sistema identifica 1.609 anomalias no período 2015–2026. A integração com dados de exportação por cadeia produtiva mapeia pares porto-cadeia de fragilidade estrutural e estados com dependência elevada de um único porto. Os resultados sugerem que a articulação de dados públicos já disponíveis pode oferecer a gestores públicos uma base analítica reprodutível para avaliar a exposição de cadeias exportadoras a disrupções. Como entrega prática, o trabalho disponibiliza um painel interativo de monitoramento com atualização semanal.

**Palavras-chave:** detecção de anomalias; portos brasileiros; machine learning; cadeia de suprimentos; vulnerabilidade comercial.

## INTRODUÇÃO

O transporte marítimo responde por cerca de 87% do total exportado pelo Brasil em valor FOB na última década, proporção que se elevou consistentemente em função da expansão das commodities (soja, minério de ferro e petróleo bruto), cujo escoamento depende quase exclusivamente da via portuária (cálculo próprio a partir de dados do ComexStat (MDIC, 2026)). A concentração da pauta exportadora em poucos produtos de grande volume e a dependência de um número limitado de portos de navegação de longo curso tornam o sistema logístico brasileiro particularmente sensível a eventos disruptivos, sejam eles de origem global (pandemias, conflitos geopolíticos), nacional (greves, choques regulatórios) ou local (eventos climáticos, falhas operacionais).

Eventos recentes ilustram essa vulnerabilidade. A pandemia de COVID-19 provocou oscilações simultâneas em praticamente todos os portos do país entre março e junho de 2020. A greve dos caminhoneiros de maio de 2018 interrompeu o fluxo de carga terrestre para os terminais marítimos, gerando queda abrupta na tonelagem embarcada. As enchentes no Rio Grande do Sul em 2024 paralisaram parcialmente as operações nos portos de Rio Grande e Porto Alegre. Em cada caso, o impacto não se limitou ao porto afetado: propagou-se pelas cadeias de exportação que dele dependem, afetando estados produtores e setores econômicos a montante.

Apesar da relevância estratégica, o monitoramento sistemático de disrupções portuárias no Brasil ainda carece de uma abordagem que integre três dimensões complementares: (i) a detecção automatizada de anomalias operacionais a partir de dados de alta frequência, (ii) a classificação do escopo e da natureza dos eventos detectados, e (iii) a avaliação da exposição econômica das cadeias produtivas afetadas. Iniciativas como o PortWatch do Fundo Monetário Internacional (IMF; University of Oxford, 2024) fornecem indicadores globais de atividade portuária, mas não operam sobre microdados operacionais nacionais. Ferramentas nacionais, como o painel estatístico da ANTAQ, disponibilizam dados operacionais detalhados, porém sem camada analítica de detecção de anomalias e sem vínculo com as cadeias produtivas por estado de origem.

Essa fragmentação analítica tem consequências práticas. Sem uma ferramenta que articule detecção de anomalias, classificação de escopo e mapeamento de exposição comercial, autoridades portuárias e órgãos como o MDIC  e a ANTAQ dispõem de visibilidade limitada sobre quais produtos, estados e parceiros comerciais são mais afetados por uma disrupção em curso. Da mesma forma, discussões sobre investimentos em infraestrutura portuária e diversificação de rotas de escoamento carecem de insumos analíticos que considerem a concentração de cadeias produtivas, e não apenas o volume movimentado.

Este trabalho propõe um sistema analítico que preenche essa lacuna a partir de uma premissa simples: as três bases de dados necessárias já existem e são de acesso público. O Anuário Estatístico Aquaviário da ANTAQ fornece microdados de atracação (2010–2026), o PortWatch/FMI oferece atividade portuária global diária (2019–2026) e o ComexStat/MDIC detalha exportações por produto, unidade aduaneira e destino (base pública disponível desde 1997). Neste estudo, o recorte analítico do ComexStat vai de 2014 a janeiro de 2026. Nenhuma dessas fontes, isoladamente, permite detectar uma disrupção, classificar seu escopo e estimar sua exposição econômica. Combinadas, sem coleta primária nem infraestrutura dedicada, elas viabilizam um fluxo integrado de detecção, classificação e análise de vulnerabilidade. Um painel interativo, produto complementar deste trabalho, consolida os principais resultados.

O objetivo geral é desenvolver e validar um sistema de monitoramento e diagnóstico de eventos disruptivos nos portos brasileiros de navegação de longo curso. Os objetivos específicos são: (a) construir modelos preditivos do comportamento operacional esperado de cada porto em quatro dimensões (atracações, tonelagem de exportação, tempo de espera para atracar e tempo atracado) e identificar desvios significativos via ensemble de métodos de detecção; (b) classificar cada anomalia detectada quanto à sua abrangência espacial (isolada, nacional ou global) a partir de scores de co-ocorrência e de choque global; (c) mapear a exposição das cadeias produtivas exportadoras brasileiras a disrupções portuárias, identificando pontos de fragilidade pela concentração de fluxos comerciais; e (d) disponibilizar os resultados em um painel interativo que permita consulta e acompanhamento operacional dos portos monitorados.

O artigo está organizado da seguinte forma. A Seção 1 apresenta o referencial teórico que fundamenta a pesquisa. A Seção 2 descreve os dados, o pipeline metodológico e os critérios de validação. A Seção 3 expõe os resultados em quatro blocos: modelagem, detecção e classificação, vulnerabilidade econômica e sistema de alerta. A Seção 4 discute os achados à luz da literatura e das limitações do trabalho. A conclusão sintetiza as contribuições e aponta direções futuras.

# 1. REFERENCIAL TEÓRICO

## 1.1 Resiliência de cadeias de suprimentos e disrupções portuárias

A literatura sobre resiliência de cadeias de suprimentos (*supply chain resilience*) cresceu substancialmente após os choques globais da década de 2020. Christopher; Peck (2004) definem resiliência como a capacidade de um sistema de retornar ao estado original ou a um novo estado desejável após uma perturbação. No contexto portuário, essa definição se traduz na capacidade de um porto de manter níveis operacionais aceitáveis durante e após eventos adversos. Chopra; Sodhi (2004) sistematizam as fontes de risco em cadeias de suprimentos, distinguindo riscos de oferta, demanda, e ambientais, categoria na qual se enquadram desastres naturais, pandemias e conflitos geopolíticos que afetam a infraestrutura portuária.

Portos são nós críticos (*chokepoints*) nas cadeias logísticas globais. Notteboom; Pallis; Rodrigue (2021) argumentam que a concentração de fluxos em poucos hubs amplia o risco sistêmico: a paralisação de um único terminal pode gerar efeitos em cascata sobre rotas marítimas, armazéns e modais terrestres conectados. Verschuur et al. (2022) propõem um framework de risco sistêmico para portos que integra três camadas (infraestrutura local, rede de transporte hinterland e rede marítima global), demonstrando que os riscos indiretos (propagação por rede) são frequentemente maiores que os danos diretos à infraestrutura portuária.

Essas fragilidades foram evidenciadas em eventos recentes. O bloqueio do Canal de Suez em março de 2021 atrasou 432 navios com carga estimada em US$ 92,7 bilhões, gerando perdas globais em cascata (Qu et al., 2024). Verschuur; Koks; Hall (2023) estimam que US$ 81 bilhões em comércio global e pelo menos US$ 122 bilhões em atividade econômica estão em risco anualmente devido a disrupções portuárias induzidas por eventos climáticos, com efeitos que se propagam por setores e geografias. Verschuur; Koks; Hall (2021) demonstraram, a partir de dados AIS, que a pandemia de COVID-19 reduziu o comércio marítimo global entre 7,0% e 9,6% nos primeiros oito meses de 2020, com impactos desproporcionais sobre economias dependentes de poucos parceiros comerciais.

No Brasil, a dependência portuária é agravada pela composição da pauta exportadora: commodities agrícolas e minerais representam parcela dominante do valor exportado, e sua movimentação ocorre por portos específicos (Santos para soja e açúcar, Itaguaí para petróleo bruto e minério de ferro, Paranaguá para grãos e carnes), criando pontos de concentração de risco que não encontram alternativa logística imediata. Wendler-Bosco; Nicholson (2020) revisam sistematicamente a literatura sobre impactos de disrupções portuárias nas cadeias de suprimentos marítimas, identificando lacunas significativas na quantificação de efeitos indiretos e na capacidade de monitoramento em tempo real. O presente trabalho busca preencher essas lacunas no contexto brasileiro.

## 1.2 Detecção de anomalias em séries temporais

A detecção de anomalias em séries temporais é um campo consolidado da estatística e do aprendizado de máquina, com aplicações em domínios que vão de monitoramento industrial a detecção de fraudes. Chandola; Banerjee; Kumar (2009) oferecem uma taxonomia abrangente dos métodos de detecção, distinguindo abordagens baseadas em classificação, *nearest-neighbor, clustering* e estatística. Boniol et al. (2024), em revisão recente, categorizam a literatura da última década sob taxonomia orientada a processos, destacando a crescente adoção de métodos híbridos que combinam modelos preditivos com detectores estatísticos.

No contexto de transportes e logística, abordagens baseadas em modelos preditivos, que aprendem o comportamento "normal" e sinalizam desvios significativos, têm se mostrado eficazes por capturarem padrões sazonais e tendências sem supervisão humana (Ahmad et al., 2017). Métodos de ensemble, que combinam múltiplos detectores para reduzir falsos positivos, ganharam destaque na literatura. Aggarwal (2017) demonstra que a concordância entre detectores independentes (votação majoritária) é mais robusta do que qualquer método isolado, especialmente em séries com sazonalidade complexa e mudanças de regime. A abordagem de ensemble $\geq 2/3$ adotada neste trabalho, combinando MAD, decomposição STL (Cleveland et al., 1990) e Isolation Forest (Liu; Ting; Zhou, 2008), segue essa lógica de diversidade metodológica.

O uso de gradient boosting (XGBoost) como modelo-base para captura do comportamento normal é respaldado por Chen; Guestrin (2016) e por aplicações recentes em previsão de demanda logística (Punia; Singh; Madaan, 2020). A vantagem do XGBoost sobre modelos puramente autorregressivos reside na capacidade de incorporar features exógenas (calendário, médias móveis, variáveis de tendência) sem especificação paramétrica rígida da estrutura temporal. A combinação de XGBoost como modelo preditivo com um filtro AR(1) para remoção de autocorrelação residual, seguida por ensemble de detecção, constitui uma arquitetura não encontrada na literatura aplicada a portos, embora cada componente individualmente seja bem documentado.

## 1.3 Monitoramento portuário e iniciativas existentes

O PortWatch, desenvolvido pelo FMI em parceria com a Universidade de Oxford, é atualmente a principal iniciativa de monitoramento portuário global de acesso aberto (Arslanalp et al., 2024; IMF; University of Oxford, 2024). Baseado em dados AIS (*Automatic Identification System*), o sistema acompanha chamadas de navios (*port calls*) em cerca de 2.000 portos em 180 países, oferecendo indicadores diários de atividade e um simulador de impacto de disrupções. O PortWatch permite identificar choques globais com rapidez, como demonstrado durante os ataques Houthi no Mar Vermelho (nov/2023–jun/2024), quando a plataforma documentou em tempo quase real o desvio de tráfego pelo Cabo da Boa Esperança.

Contudo, o PortWatch opera exclusivamente sobre dados AIS, não capturando microdados operacionais como tempos de espera e tempos de operação no berço. Embora forneça estimativas de tonelagem desagregadas por tipo de embarcação (*contêiner*, granéis sólidos, carga geral, ro-ro e tanker) e por direção (importação e exportação), essa classificação reflete o tipo de navio, não a natureza específica da mercadoria — distinção disponível apenas nos microdados ANTAQ. Essa granularidade é essencial para o diagnóstico operacional, como distinguir se uma queda de atividade resulta de falta de carga (problema logístico terrestre) ou de congestionamento portuário (problema de capacidade). Verschuur; Koks; Hall (2020)

destacam que a integração de fontes complementares de dados (AIS global combinado com dados operacionais locais) é uma fronteira importante para a análise de resiliência portuária.

No Brasil, a ANTAQ disponibiliza microdados do Anuário Estatístico Aquaviário com registros individuais de atracação, incluindo timestamps detalhados e informações de carga. Contudo, esses dados são disponibilizados com defasagem e sem detecção automatizada de anomalias ou cruzamento com cadeias produtivas por estado de origem. O presente trabalho busca preencher essa lacuna ao combinar a granularidade operacional da ANTAQ com o sinal global do PortWatch, integrando ainda dados de comércio exterior do ComexStat para contextualização econômica.

## *1.4 Concentração comercial e vulnerabilidade*

O Índice Herfindahl-Hirschman (HHI) é amplamente utilizado como medida de concentração de mercado (Rhoades, 1993). No contexto do comércio exterior, o HHI pode ser aplicado em duas direções complementares: (i) HHI da pauta do porto, que mede o grau de dependência de um porto em relação a poucos produtos exportados, e (ii) HHI da cadeia produtiva, que mede o grau de concentração da exportação de um produto em poucos portos. Pares produto-porto no quadrante de dupla concentração (ambos os HHIs elevados) representam pontos de fragilidade sistêmica: se o porto para, não há alternativa portuária para o produto, e se o produto desaparece, o porto perde parcela significativa de sua movimentação.

Essa perspectiva de dupla concentração conecta-se ao conceito de *criticality* portuária desenvolvido por Verschuur; Koks; Hall (2022), que quantificam a importância de cada porto para as cadeias globais de suprimentos pela alocação de fluxos comerciais na rede de transporte marítimo. No contexto brasileiro, onde a via marítima responde por cerca de 87% das exportações em valor, a concentração de cadeias produtivas inteiras em um ou dois portos, como o petróleo bruto em São João da Barra e Itaguaí ou o níquel em Santos, configura vulnerabilidades que transcendem o domínio logístico e adquirem relevância macroeconômica. A identificação sistemática desses pontos de fragilidade pode contribuir para que discussões sobre investimento em infraestrutura portuária e diversificação de rotas de escoamento incorporem critérios de concentração de risco como insumo complementar ao volume movimentado.

A Classificação Uniforme para o Comércio Internacional (CUCI), adotada pela SECEX nas divulgações semanais da balança comercial, fornece a taxonomia de produtos utilizada neste trabalho no nível de Grupo (3 dígitos), oferecendo granularidade adequada para distinguir cadeias produtivas relevantes (soja ≠ café ≠ minério de ferro) sem excessiva fragmentação.

# 2. METODOLOGIA

## 2.1 Dados

O pipeline analítico integra três bases de dados públicas, cada uma contribuindo com uma camada específica de informação.

**ANTAQ: Anuário Estatístico Aquaviário.** Microdados de atracação extraídos do portal de dados abertos da Agência Nacional de Transportes Aquaviários. O conjunto bruto contém 1,3 milhão de registros de atracação (2010–2026). Após filtragem para navegação de longo curso e seleção dos 18 principais portos exportadores, restaram 319.414 registros cobrindo o período de janeiro de 2014 a janeiro de 2026. Cada registro corresponde a uma atracação individual de navio, com informações de timestamps operacionais (chegada, atracação, início e fim de operação, desatracação), tonelagem movimentada e tipo de carga. A partir desses microdados, foram construídas séries temporais semanais em quatro dimensões operacionais (Tabela 1).

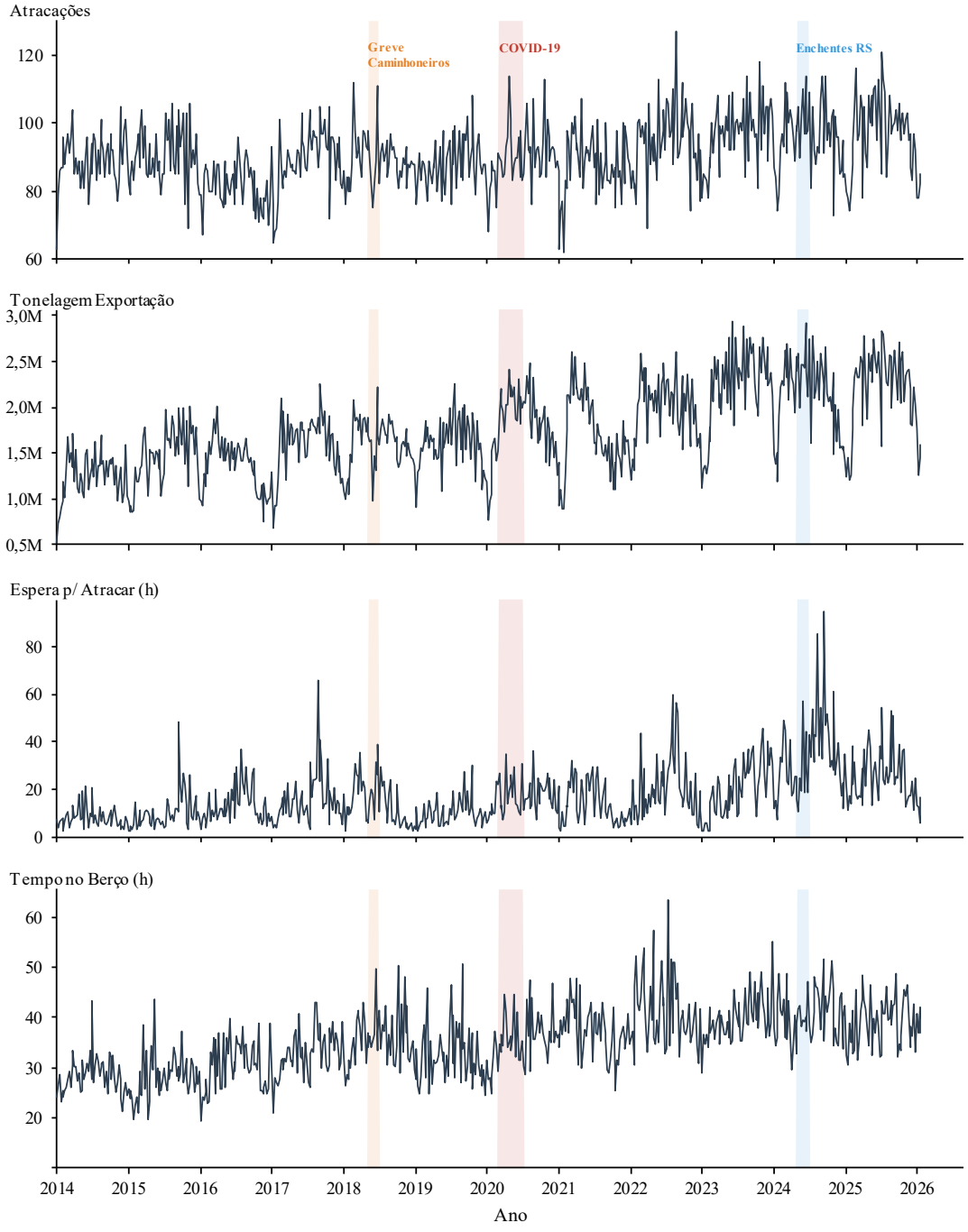
**Tabela 1: Dimensões operacionais monitoradas por porto.**

| Dimensão | O que mede | Exemplo de anomalia |
|---|---|---|
| Atracações/semana | Volume operacional (contagem) | Queda abrupta por greve ou clima |
| Tonelagem de exportação | Throughput de carga (toneladas) | Desvio de rota, safra atípica |
| Tempo de espera para atracar (h) | Congestionamento externo (T1) | Fila de navios por excesso de demanda |
| Tempo atracado (h) | Eficiência de berço (T2+T3+T4) | Operação prolongada, paralisação |

A granularidade semanal foi adotada como compromisso entre resolução temporal e estabilidade estatística: séries diárias apresentam excesso de zeros (portos com poucas atracações por dia), enquanto séries mensais diluem eventos de curta duração como greves e bloqueios logísticos.

A Figura 1 ilustra as quatro dimensões para o Porto de Santos, principal *hub* exportador do país, ao longo de toda a série disponível (2014–2026). Os painéis permitem visualizar tanto os padrões sazonais regulares (ciclos de safra, variações de fim de ano) quanto as perturbações provocadas por eventos disruptivos conhecidos: a Greve dos Caminhoneiros (maio–junho de 2018), a pandemia de COVID-19 (março–junho de 2020) e as Enchentes do Rio Grande do Sul (abril–junho de 2024). É esse tipo de variabilidade que o pipeline de detecção descrito na seção seguinte buscará capturar automaticamente.

**Figura 1: Séries temporais semanais do Porto de Santos nas quatro dimensões operacionais (2014–2026). Faixas verticais indicam períodos de eventos de referência: Greve dos Caminhoneiros (mai–jun/2018, laranja), COVID-19 (mar–jun/2020, vermelho) e Enchentes do RS (abr–jun/2024, azul).**




**Fonte:** Elaboração própria com dados ANTAQ (2026).

**PortWatch/FMI.** Dados diários de atividade portuária global, cobrindo ~2.000 portos em 180 países, disponibilizados tanto em *bulk download* quanto via API REST — esta última utilizada especialmente no painel de alertas para atualização com os dados mais recentes. Para a construção do índice global, foram selecionados os 50 maiores portos não-brasileiros por volume médio de chamadas de navios. A partir das médias semanais de *port calls*, calculou-

se um *z-score* com janela móvel de 12 semanas (com mínimo de 6 observações para início do cálculo). Os dados diários do PortWatch iniciam em janeiro de 2019, o que impõe uma limitação temporal ao *Score B* de classificação global.

**ComexStat/MDIC.** Estatísticas de comércio exterior brasileiro em dados abertos, extraídas da plataforma do Ministério do Desenvolvimento, Indústria, Comércio e Serviços, com agregação mensal no nível Produto (Grupo CUCI) × UF de origem × URF de exportação × País de destino. Neste trabalho, o conjunto analítico cobre o período de 2014 a janeiro de 2026, totalizando 16,3 milhões de registros brutos. Após filtro por via de transporte marítimo e mapeamento da Unidade da Receita Federal (URF) de despacho para o porto correspondente, restaram 7,5 milhões de registros cobrindo 16 dos 18 portos monitorados (Porto Alegre e Barra do Riacho foram excluídos por não possuírem URF dedicada no ComexStat). No manifesto de proveniência, a série de arquivos brutos do ComexStat utilizada no pipeline inicia em 2011.

## 2.2 Pipeline de detecção de anomalias

O sistema de detecção opera em quatro etapas sequenciais, ilustradas na Figura 2. Os scripts de coleta, processamento e análise estão disponíveis em repositório público no GitHub, descrito na seção de Disponibilidade de Dados e Código.

**Figura 2: Pipeline de detecção e classificação de anomalias portuárias.**




**Fonte:** Elaboração própria (2026).

**Etapa 1: Modelagem preditiva (XGBoost).** Para cada porto e cada dimensão operacional, um modelo XGBoost aprende o padrão normal de operação a partir de 15 features temporais restritas à própria dimensão-alvo: *lags* semanais (1, 4, 13, 52 semanas), médias móveis (4, 8, 12 semanas), variações percentuais, desvio em relação à média móvel, volatilidade de curto prazo e variáveis calendário (seno e cosseno do mês, semana do ano). A restrição de features à dimensão própria (*own-only*) evita contaminação cruzada entre dimensões. Os hiperparâmetros (max_depth=4, n_estimators=200, learning_rate=0.05) foram selecionados por validação cruzada com janela expansível (*expanding window*).

**Etapa 2: Correção AR(1) e inovações.** Os resíduos do XGBoost (rt=yt-yt) apresentam autocorrelação remanescente (mediana do ACF(1) ≈ 0,21 nos dados atualizados). Um filtro autorregressivo de primeira ordem é aplicado individualmente a cada par porto×dimensão:

$$\varepsilon_t = r_t - \hat{\rho} \cdot r_{t-1}$$

$\hat{\rho}$

onde é estimado pela autocorrelação amostral dos resíduos. As inovações t resultantes apresentam ACF(1) mediana de -0,03 e 96% dos pares porto×dimensão com |ACF(1)| < 0,15, permitindo que os métodos de detecção operem sobre desvios aproximadamente independentes.

**Etapa 3: Ensemble de detecção (≥ 2/3).** Três métodos independentes avaliam as inovações:

- *MAD (Median Absolute Deviation)* (Leys et al., 2013): threshold adaptativo por porto×dimensão, calibrado sobre a distribuição local das inovações;
- *STL (Seasonal-Trend decomposition using Loess):* detecção sobre o componente residual da decomposição sazonal;
- *Isolation Forest:* detecção multivariada baseada em isolamento de observações anômalas.

Uma semana é marcada como anomalia se ao menos dois dos três métodos concordam (regra de votação ≥ 2/3). O threshold adaptativo do MAD utiliza kbase=3,0, ajustado para cima em portos com variabilidade naturalmente elevada, evitando excesso de falsos positivos.

**Etapa 4: Walk-forward.** Todo o pipeline é executado em regime *walk-forward*: o modelo é re-estimado periodicamente em janela expansível, acumulando todo o histórico disponível a cada retreinamento. Isso garante que apenas informação passada é utilizada na previsão, eliminando *data leakage*. As primeiras 78 semanas (~1,5 ano) de cada série são descartadas como *burn-in* para estabilização do modelo.

## 2.3 Classificação das anomalias

Cada anomalia detectada é classificada em três categorias (isolada, nacional ou global) por meio de dois *scores* independentes, calibrados contra 10 eventos históricos conhecidos (Tabela 3).

**Score A: co-ocorrência espacial.** Conta quantos portos brasileiros apresentaram anomalia na mesma dimensão dentro de uma janela de ±4 semanas:

$$S_A = |\{p \in P : \exists\ anomalia_p\ em\ [t-4, t+4]\ sem\}|$$

Score A alto indica evento sistêmico (muitos portos afetados simultaneamente).

**Score B: índice de choque global.** Construído a partir do z-score da atividade portuária global (PortWatch), tomando o máximo |z| numa janela de ±2 semanas em torno de cada anomalia:

$$S_B = \max_{|w| \leq 2} \left| Z_{GI}(t+w) \right|$$

*Score B* alto indica choque externo severo no comércio marítimo global.

A regra de classificação combina ambos os scores com limiares ta=3 e tb=0,8, encontrados por *grid search* exaustivo sobre o espaço ta□{3,4,5,6} e tb□{0,8; 1,0; 1,2; 1,5}, maximizando o recall ponderado contra os 10 eventos de calibração (Tabela 3). Os limites inferiores foram definidos a priori: ta3 porque, com 18 portos monitorados, co-ocorrência de apenas 2 representa ruído estatístico; tb0,8 (percentil 79 do z-score global) para evitar classificação espúria de flutuações normais como choques globais. O par (ta=3, tb=0,8) detectou todos os 10 eventos de referência, com precisão de classificação de 90%:

$$S_A = |\{p \in P : \exists\ anomalia_p\ em\ [t-4, t+4]\ sem\}|$$

*Score A* alto indica evento sistêmico (muitos portos afetados simultaneamente).

**Score B: índice de choque global.** Construído a partir do *z-score* da atividade portuária global (PortWatch), tomando o máximo |z| numa janela de ±2 semanas em torno de cada anomalia:

$$S_B = \max_{|w| \leq 2} \left| Z_{GI}(t+w) \right|$$

**Tabela 2: Regra de classificação das anomalias.**

| Tipo | Condição | Interpretação |
|---|---|---|
| Global | Score B ≥ 0,8 | Choque externo detectado |
| Nacional | Score A ≥ 3 e Score B < 0,8 | Coocorrência sem *trigger* global |
| Isolado | Score A < 3 e Score B < 0,8 | Evento localizado |

A regra de promoção por *Score B*, que classifica como global mesmo anomalias com poucos portos brasileiros afetados, foi introduzida para capturar eventos como o bloqueio do Canal de Suez, que afetou o comércio mundial mas impactou poucos portos brasileiros diretamente.

Cabe notar que as janelas simétricas do Score A (±4 semanas) e do *Score B* (±2 semanas) incorporam observações posteriores à semana da anomalia, tornando a classificação intrinsecamente retrospectiva. Essa é uma escolha deliberada: para diagnóstico histórico, janelas simétricas produzem classificações mais precisas por capturarem tanto o *onset* quanto a propagação de cada evento. Em aplicação operacional, a detecção de anomalias (etapas 1–3 do pipeline) opera de forma causal, utilizando apenas dados passados no regime *walk-forward*. A classificação do tipo de evento, porém, requer que a janela se feche para ser aplicada, o que introduz uma latência de 2–4 semanas entre a detecção e a classificação definitiva. O painel de alertas do *dashboard* contorna essa latência ao utilizar *z-scores* PortWatch com atualização semanal como mecanismo de vigilância independente do pipeline de classificação.

**Tabela 3 - Eventos utilizados para calibração (10 eventos, 135 pares porto×evento)**

| # | Evento | Período | Classificação | Portos |
|---|---|---|---|---|
| 1 | Greve dos Caminhoneiros | mai–jun/2018 | Nacional | 4 |
| 2 | COVID-19 | mar–jun/2020 | Global | 18 |
| 3 | COVID recuperação | set/2020–jun/2021 | Global | 18 |
| 4 | Bloqueio Cana de Suez | mar–abr/2021 | Global | 18 |
| 5 | Guerra na Ucrânia | fev–dez/2022 | Global | 18 |
| 6 | *Lockdowns* China | mar–jun/2022 | Global | 18 |
| 7 | Ataques Houthi | nov/2023–jun/2024 | Global | 18 |
| 8 | Enchentes RS | abr–jun/2024 | Isolado | 2 |
| 9 | Greve Receita Federal | dez/2024-jan/2025 | Nacional | 3 |
| 10 | Tarifação EUA | Fev-jun/2025 | Global | 18 |

### 2.4 Análise de vulnerabilidade econômica

A camada de exposição comercial cruza os dados de anomalias com os fluxos de exportação por cadeia produtiva. Duas métricas de concentração são calculadas:

- **HHI do porto** (HHIp): grau de concentração da pauta exportadora de cada porto (quanto maior, mais o porto depende de poucos produtos);
- **HHI da cadeia** (HHIc): grau de concentração portuária de cada cadeia produtiva (quanto maior, mais a exportação do produto depende de poucos portos).

Pares p,c com ambos os HHIs elevados (> 0,20) representam **pontos de fragilidade dupla:** uma disrupção no porto afeta severamente a cadeia, e a cadeia não dispõe de rota alternativa.

## 3. RESULTADOS

### 3.1 Desempenho da modelagem preditiva

**Tabela 4: Redução do MAE em relação ao modelo naïve (lag-52 semanal), com IC 95% por dimensão no regime walk-forward.**

| Dimensão | MAE Naïve | MAE XGBoost | Redução (%) | IC 95% | Anomalias |
|---|---|---|---|---|---|
| Atracações | 3,96 | 0,73 | 82 | [79; 84] | [79; 84] |
| Tonelagem Exp. | 172.760 | 41.329 | 76 | [73; 80] | [73; 80] |
| Espera p/ Atracar (h) | 37,29 | 10,08 | 73 | [67; 81] | [67; 81] |
| Tempo no Berço (h) | 15,13 | 3,28 | 78 | [73; 84] | [73; 84] |

**Fonte:** Elaboração própria com dados ANTAQ (2026).

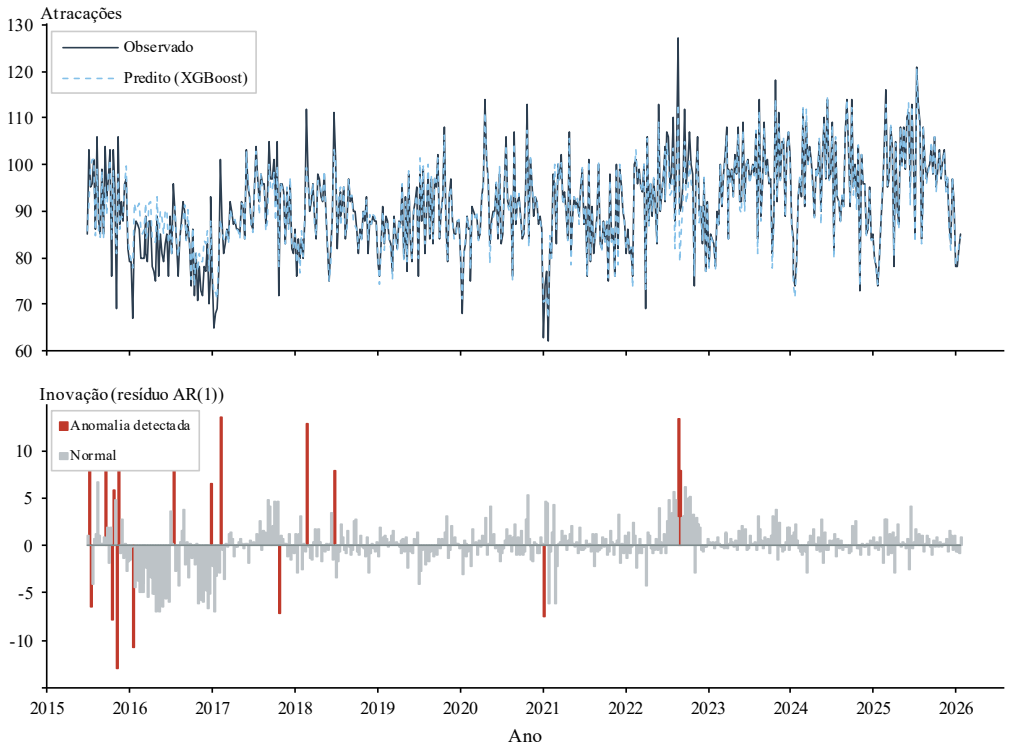
O modelo XGBoost com correção AR(1) demonstrou capacidade consistente de capturar a variabilidade operacional semanal dos portos. Conforme ilustrado para Santos (Figura 1), as séries apresentam sazonalidade marcada e perturbações pontuais; o modelo foi capaz de separar ambos os componentes. Em validação cruzada com janela expansível, as reduções de MAE em relação ao baseline naïve (lag-52 semanal, o valor da mesma semana no ano anterior) variaram de 79% a 84% conforme a dimensão. Na avaliação walk-forward completa, as reduções pontuais de MAE variam de 73% a 82% conforme a dimensão (Tabela 4), refletindo o cenário mais conservador em que o modelo opera estritamente sobre dados passados. Os intervalos de confiança bootstrap (IC 95%, 10.000 iterações reamostrando portos) confirmam a robustez: mesmo o limite inferior mais conservador (67%), observado no tempo de espera, permanece elevado.

As features mais relevantes para o XGBoost, avaliadas por ganho médio entre portos, são consistentes entre dimensões: o desvio em relação à média móvel de 8 semanas (dev_mm8) e a média móvel de 4 semanas (mm4) dominam em todas as quatro dimensões, indicando que o modelo captura o padrão normal principalmente pela tendência de curto prazo e seus desvios recentes.

A correção AR(1) reduziu a autocorrelação mediana dos resíduos de 0,21 para -0,03. Embora o teste de Ljung-Box rejeite ruído branco em todos os portos ($p < 0,05$), resultado esperado com séries de ~550 observações em que mesmo autocorrelações de magnitude negligível são estatisticamente significativas, a métrica operacionalmente relevante é a magnitude: 96% dos pares porto×dimensão apresentam $|ACF(1)| < 0,15$ após o filtro, indicando que a dependência serial residual é insuficiente para comprometer a detecção de anomalias. Adicionalmente, a modelagem absorveu a co-variação esperada entre dimensões: a correlação entre atracações e tonelagem, o par mais forte, caiu de $r=0,57$ nas séries brutas para $r=0,24$ nos resíduos, confirmando que os resíduos capturam desvios genuinamente independentes e justificando o tratamento separado de cada dimensão na etapa de detecção.

A Figura 3 ilustra o pipeline de detecção aplicado à dimensão Atracações do Porto de Santos. No painel superior, a série observada é comparada com a previsão do XGBoost; no painel inferior, as inovações (resíduos corrigidos pelo AR(1)) permitem identificar os desvios que excedem o limiar do ensemble, marcados como anomalias detectadas.

**Figura 3: Exemplo do pipeline de detecção para Santos, Atracações. Painel superior: série observada (escuro) e previsão XGBoost (azul tracejado). Painel inferior: inovações semanais (resíduos após AR(1)); barras vermelhas indicam semanas com anomalia detectada pelo ensemble.**




**Fonte:** Elaboração própria com dados ANTAQ (2026).

## 3.2 Anomalias detectadas e classificação

As classificações apresentadas a seguir resultam da análise retrospectiva com janelas simétricas descrita na Seção 2.3.

O ensemble de detecção identificou **1.609 anomalias** no período de junho de 2015 a janeiro de 2026, correspondendo a aproximadamente 4,0% das observações pós-*burn-in* (39.744 semanas-porto-dimensão). A distribuição por classificação é: 636 globais (40%), 912 nacionais (57%) e 61 isoladas (4%). A Figura 4 apresenta a classificação no espaço bidimensional *Score A* × *Score B*.

A classificação global depende do z-score PortWatch (*Score B*), cujos dados diários iniciam em janeiro de 2019; como o cálculo usa janela móvel de 12 semanas com mínimo de 6 observações, o Score B passa a assumir valores positivos a partir de fevereiro de 2019 (com estabilização da janela completa ao longo de março). As 862 anomalias detectadas antes de **01/02/2019** (53,6% do total) são classificadas como nacionais ou isoladas por construção, o que infla a primeira categoria e deflaciona a segunda na distribuição agregada. A primeira anomalia classificada como global no arquivo final ocorre em 18/02/2019. Quando se restringe a análise ao período com cobertura PortWatch efetiva (fevereiro de 2019 a janeiro de 2026, 747 anomalias), a distribuição se inverte substancialmente, como mostra a Tabela 5. O total de 636 anomalias globais permanece idêntico no período completo e no recorte fev/2019+ porque todas as anomalias globais estão dentro dessa janela; não houve reclassificação adicional no corte.

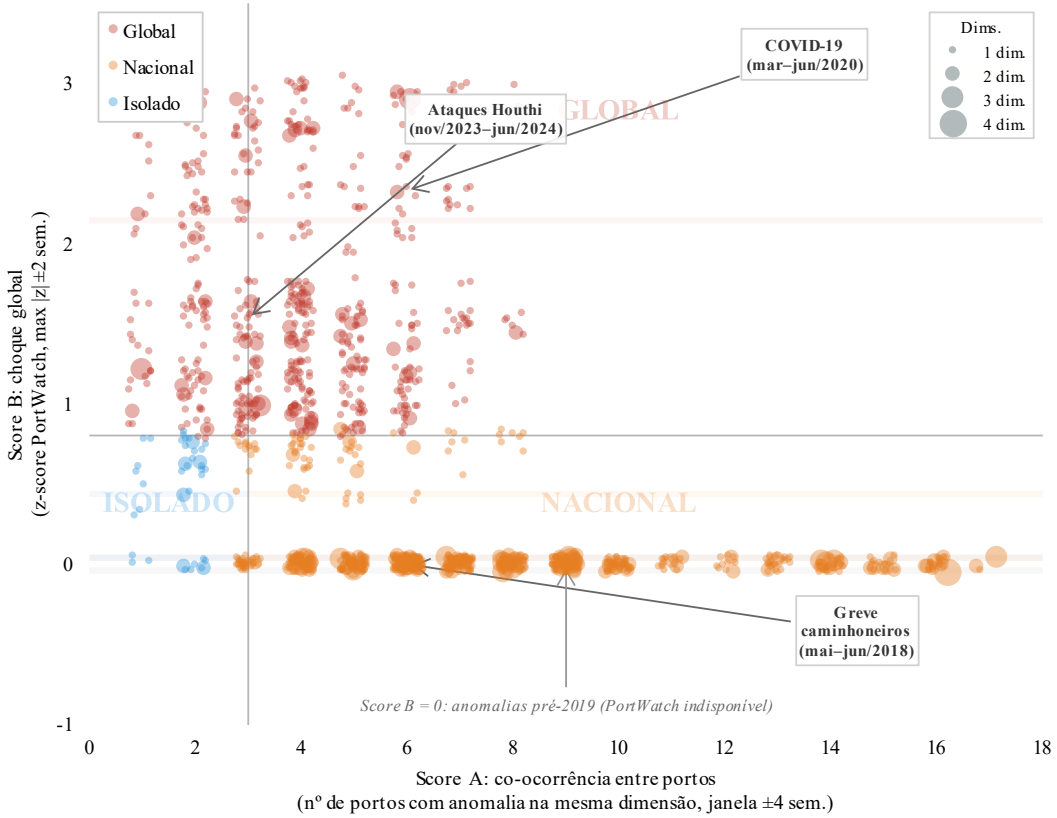
**Tabela 5: Distribuição das anomalias por classificação e período de cobertura. No período completo, 862 anomalias pré-2019 não dispõem de Score B e são classificadas como nacionais ou isoladas por construção.**

| Classificação | Período completo | Com PortWatch (fev/2019+) |
|---|---|---|
| Global | 636 (40%) | 636 (85%) |
| Nacional | 912 (57%) | 80 (11%) |
| Isolado | 61 (4%) | 31 (4%) |
| **Total** | **1.609** | **747** |

A inversão é coerente com a sequência de choques sistêmicos do período 2019–2026 (COVID-19, crise de contêineres, ataques Houthi).

A predominância de anomalias nacionais na distribuição agregada reflete tanto a sensibilidade do Score A à co-ocorrência doméstica — ajustes cambiais, safras atípicas e mudanças regulatórias afetam múltiplos portos sem configurar choque global — quanto o viés de janela descrito acima. A proporção reduzida de eventos isolados (4%) decorre em parte do limiar conservador (ta=3): um evento que afeta dois portos é classificado como isolado, mas a maioria dos choques, mesmo os aparentemente locais, tende a reverberar em pelo menos três terminais pela interconexão das cadeias logísticas. Os 61 eventos isolados concentram-se em São João da Barra (14) e Imbituba (10), são predominantemente unidimensionais, e em quase metade dos casos afetam apenas o tempo de espera. O Score A desses eventos varia entre 1 e 2, confirmando que se trata de perturbações genuinamente localizadas e não artefatos do limiar.

**Figura 4: Classificação das anomalias detectadas por Score A (co-ocorrência entre portos) e Score B (índice de choque global PortWatch). Cada ponto representa um par porto×semana; tamanho proporcional ao número de dimensões com anomalia simultânea (1–4). Limiares ótimos: ta=3, tb=0,8. Anotações indicam a localização dos principais eventos de referência. Pontos com Score B = 0 correspondem a anomalias anteriores a 2019, período sem cobertura PortWatch.**




**Fonte:** Elaboração própria com dados ANTAQ (2026) e PortWatch/FMI (2026).

A validação contra os 10 eventos históricos de referência, expandidos em 135 pares porto×evento, produziu as seguintes métricas:

- **Recall de detecção:** 71% (IC 95%: [64%, 79%]; 96/135 pares com ao menos uma anomalia detectada na janela do evento)
- **Precisão da classificação:** 90% (IC 95%: [83%, 95%]; 86/96 anomalias detectadas classificadas corretamente)
- **Recall da classificação:** 64% (IC 95%: [56%, 72%]; 86/135 pares com classificação correta)

Cabe distinguir os dois indicadores de recall. O primeiro (71%) mede se o sistema *detecta* que algo anômalo ocorreu. O segundo (64%) mede se, além de detectar, o sistema *classifica corretamente* a natureza do evento. A diferença (7 pontos percentuais) resulta de anomalias detectadas na janela do evento mas classificadas no tipo errado, tipicamente eventos globais classificados como nacionais quando o *Score B* ficou abaixo do limiar.

A estabilidade dos limiares foi verificada por validação *leave-one-event-out* (LOEO): em cada um dos 10 folds, os limiares foram re-otimizados sobre 9 eventos e avaliados no evento excluído. O par (ta=3, tb=0,8) foi selecionado em 10 dos 10 folds, com recall médio de detecção de 68%

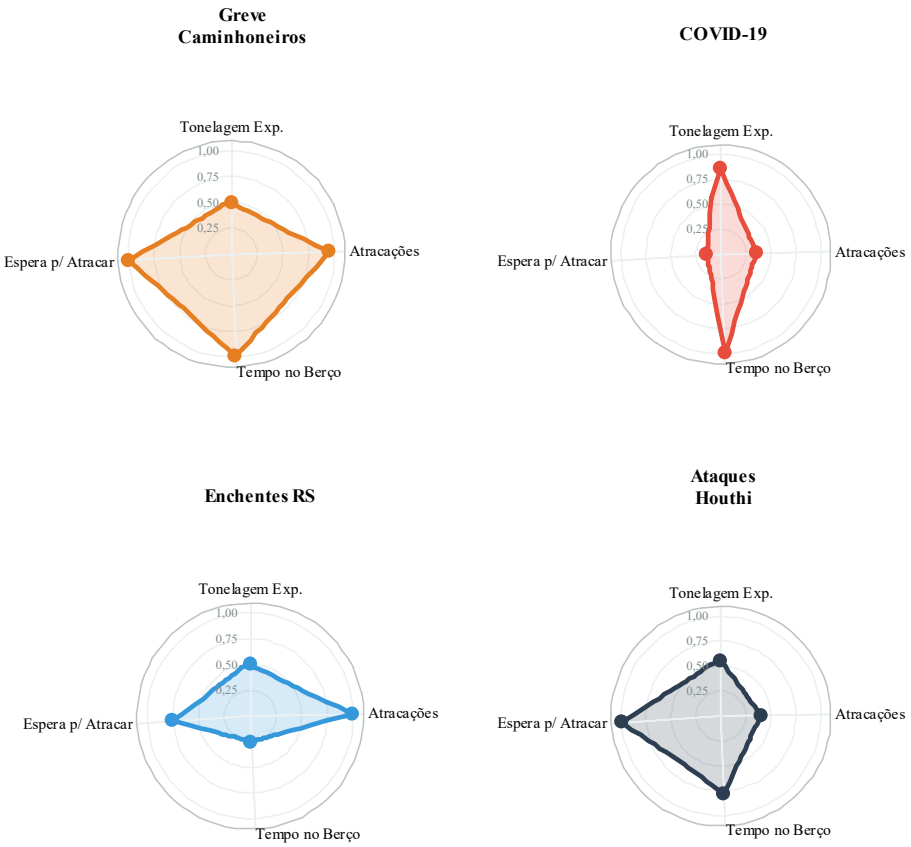
± 26% e precisão média de classificação de 83% ± 32%, confirmando que a calibração não apresenta overfitting ao conjunto de eventos de referência. A maior dispersão das métricas LOEO em relação às globais decorre do tamanho variável dos eventos: folds que excluem eventos de poucos pares (e.g., Enchentes RS, 2 pares) produzem estimativas mais voláteis.

Dois eventos de referência apresentaram detecção nula: a Greve da Receita Federal (dez/2024–jan/2025) nos três portos testados, e a Tarifação EUA (fev/2025 em diante), cujo período cai no fim da série disponível. Ambos são limitações conhecidas: o primeiro por tratar-se de evento de impacto difuso sobre tempos administrativos não capturados pelas quatro dimensões operacionais; o segundo por insuficiência de dados pós-evento na série atual.

### *3.2.1 Fingerprints operacionais*

Cada evento disruptivo produz uma assinatura operacional distinta, o *fingerprint*, que revela quais dimensões do porto são mais afetadas. O *fingerprint* é o vetor de intensidade relativa do z-score residual nas quatro dimensões, calculado como a média dos desvios normalizados nas semanas de anomalia do evento (Figura 5). Para a visualização dos *fingerprints*, utilizam-se janelas temporais mais amplas que as da calibração (Tabela 3), incluindo todos os portos com anomalia no período, o que permite capturar o perfil agregado do evento sem restringi-lo aos portos originalmente selecionados para validação.

**Figura 5: Perfil operacional (*fingerprint*) de quatro eventos de referência. Barras representam o *z-score* médio do resíduo normalizado em cada dimensão, calculado sobre as semanas de anomalia nos portos afetados.**




**Fonte:** Elaboração própria com dados ANTAQ (2026).

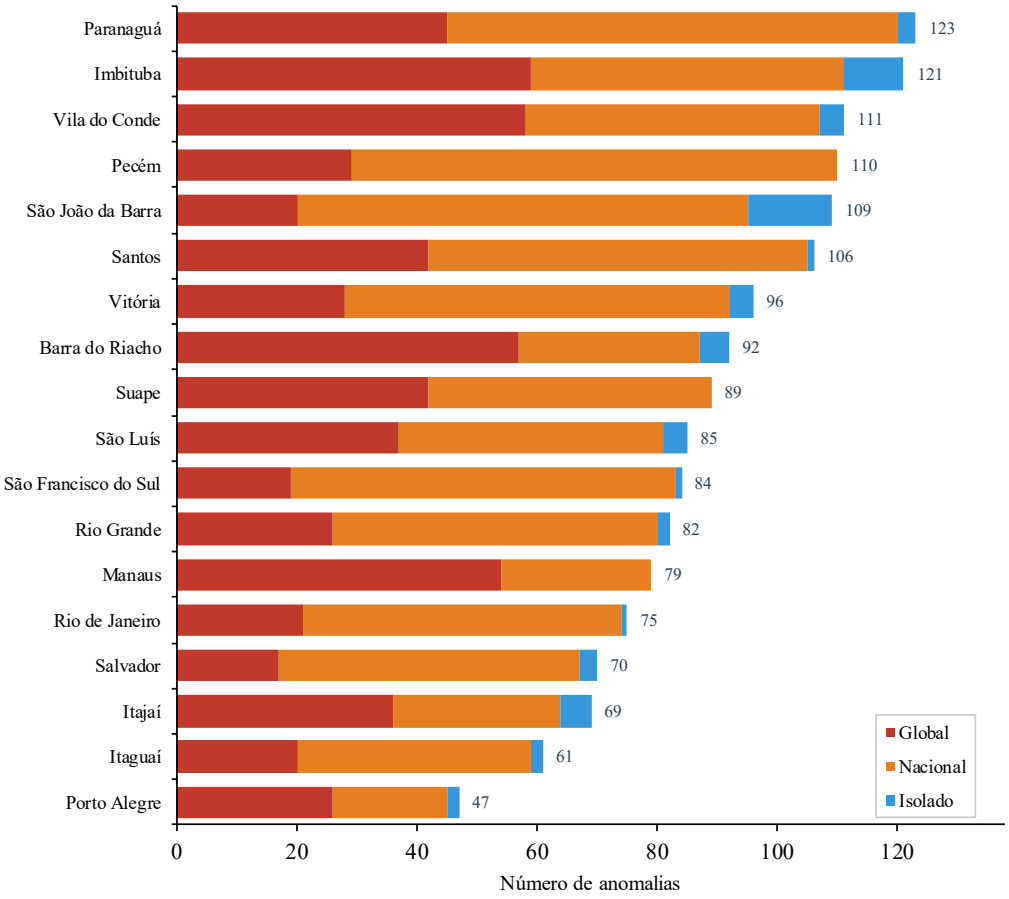
Os padrões observados são qualitativamente distintos. A Greve dos Caminhoneiros concentrou anomalias em tempo no berço (predominantemente elevado) e atracações, enquanto a tonelagem, menos frequente como dimensão anômala, registrou quedas intensas nos portos afetados — perfil consistente com navios que chegaram ao terminal mas enfrentaram operações prolongadas por escassez de carga terrestre. A COVID-19 afetou sobretudo tempo no berço e tonelagem de exportação, com efeito mínimo sobre o tempo de espera para atracar, o que aponta para redução de atividade operacional e não para congestionamento. As Enchentes do RS tiveram na tonelagem a dimensão mais atingida, porém com direções opostas entre portos — quedas nos terminais gaúchos e possível redirecionamento de carga para outros[3].

### *3.2.2 Panorama temporal e perfis de vulnerabilidade*

O panorama de anomalias por porto (Figura 6) revela padrões marcantes na distribuição por classificação. Portos como Paranaguá (123 anomalias), Imbituba (121), Vila do Conde (111), Pecém (110) e São João da Barra (109) acumularam anomalias recorrentes ao longo da série, sugerindo vulnerabilidade estrutural. A proporção entre tipos varia entre portos: Manaus (68,4%) e Barra do Riacho (62,0%) apresentam predominância de anomalias globais, enquanto São João da Barra e São Francisco do Sul são dominados por anomalias nacionais — proporção que deve ser interpretada à luz do viés de janela descrito acima, já que 53,6% das anomalias ocorrem no período sem cobertura PortWatch e são classificadas como nacionais ou isoladas por construção. O ano de 2016 concentrou o maior número absoluto de anomalias no sistema, período associado ao forte ajuste econômico e cambial brasileiro. A COVID-19 em 2020, embora com menor número absoluto, foi o evento com maior dispersão geográfica simultânea: 17 dos 18 portos apresentaram anomalias.

3 Parte dos portos com anomalia no período das enchentes (Manaus, Santos, Suape) provavelmente reflete perturbações concorrentes não ligadas ao evento gaúcho. Os Ataques Houthi elevaram o tempo no berço de forma consistente (11 de 13 anomalias positivas), enquanto o tempo de espera, embora frequente, apresentou direções mistas entre portos, indicando que o desvio de rotas não impactou os terminais de forma homogênea. O fingerprint permite levantar hipóteses sobre a natureza de um evento anômalo, mas a interpretação causal requer cautela quando a janela temporal coincide com perturbações de origens distintas.

**Figura 6: Distribuição de anomalias por porto e classificação (global, nacional, isolado). Portos ordenados por total de anomalias. A predominância de anomalias nacionais em alguns portos é parcialmente influenciada pela ausência de Score B no período pré-2019 (53,6% do total); ver discussão sobre viés de janela no texto.**



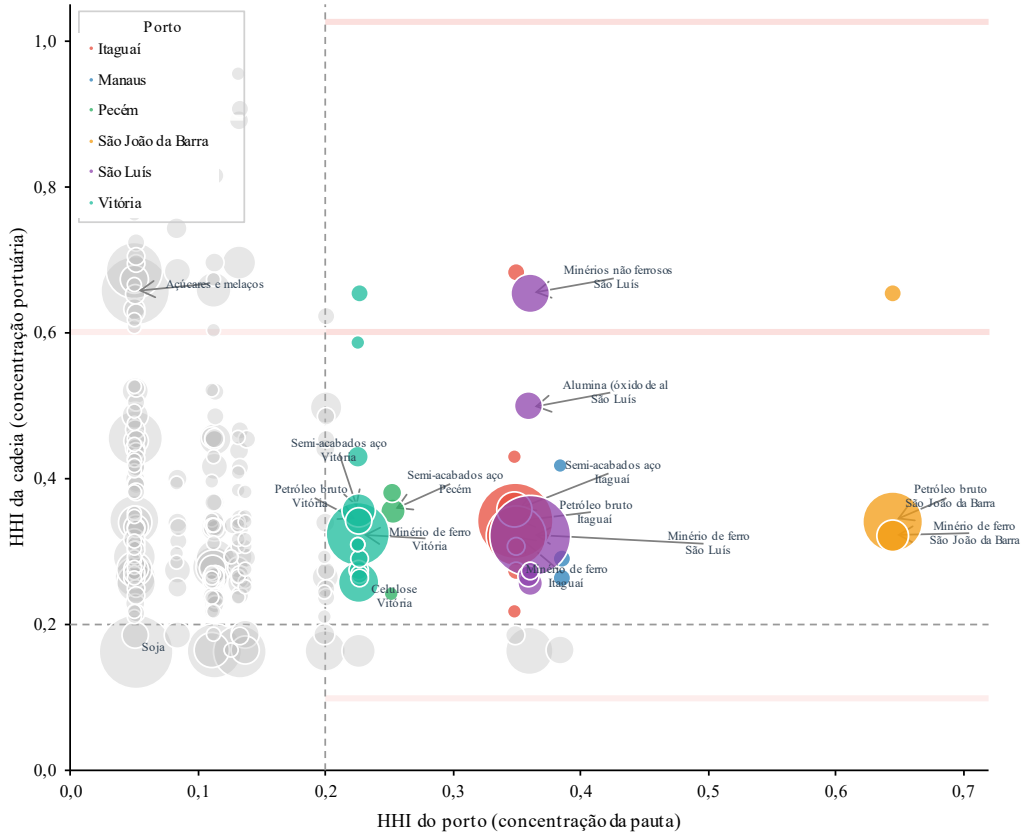

**Fonte:** Elaboração própria com dados ANTAQ (2026) e PortWatch/FMI (2026).

A análise de clusterização multidimensional identificou quatro perfis de disrupção entre as semanas anômalas: (i) picos de volume, com aumento simultâneo de atracações e tonelagem, típico de choques de demanda; (ii) gargalos extremos, com tempo atracado e tempo de espera muito elevados, concentrados em portos menores; (iii) eficiência extrema, com tempos operacionais anormalmente baixos, possivelmente indicando navios menores ou desembarques parciais; e (iv) quedas de volume, com redução de atracações e tonelagem, o padrão mais frequente, associado a greves, bloqueios logísticos e crises econômicas. A análise de componentes principais (PCA) sobre os resíduos das quatro dimensões revela que esses perfis se organizam em dois eixos ortogonais: um modo de volume (atracações e tonelagem com cargas de 0,69 no PC1, explicando 31,6% da variância) e um modo de tempo (T1 e TAtracado com cargas de 0,67–0,74 no PC2, explicando 26,7%). Essa estrutura fatorial dá sustentação conceitual aos *fingerprints* dimensionais: as quatro dimensões não variam de forma aleatória nas anomalias, mas se agrupam em dois modos de disrupção distintos.

## 3.3 Vulnerabilidade econômica cruzada

A Figura 7 cruza as duas dimensões de concentração num espaço bidimensional onde cada ponto representa um par (porto, cadeia CUCI 3 dígitos) com FOB acumulado superior a US$ 1 bilhão. Nesta seção, adota-se deliberadamente a janela 2015–2025 (10 anos completos) para comparabilidade interanual dos indicadores de concentração, distinta do recorte geral 2014 – jan/2026, usado em outras etapas do pipeline. O eixo vertical identifica as cadeias produtivas com maior concentração portuária, ou seja, aquelas para as quais uma disrupção no porto dominante não encontra rota alternativa viável. O eixo horizontal revela o grau de especialização do porto: valores elevados indicam terminais cuja pauta se concentra em poucos produtos, sinalizando tanto uma vulnerabilidade setorial quanto uma oportunidade de política pública — portos altamente especializados são candidatos prioritários a estratégias de diversificação e adensamento da malha logística. Pares situados no quadrante superior-direito (ambos os HHIs acima de 0,20) combinam dependência da cadeia com especialização do porto, configurando os nós onde investimentos em redundância tenderiam a oferecer maior ganho de resiliência, hipótese que demandaria avaliação específica de custo-benefício.

**Figura 7: Concentração cruzada porto–cadeia produtiva. Cada ponto representa um par (porto, cadeia CUCI 3 dígitos) com FOB acumulado superior a US$ 1 bilhão (2015–2025). Eixo horizontal: HHI do porto (concentração da pauta exportadora). Eixo vertical: HHI da cadeia (concentração portuária). Tamanho das bolhas proporcional ao FOB acumulado do par. Linhas tracejadas indicam o limiar de concentração elevada (HHI = 0,20). Pares no quadrante superior-direito representam pontos de fragilidade dupla.**




**Fonte:** Elaboração própria com dados ComexStat/MDIC (2015–2025) e NB3 (HHI).

O gráfico identifica 36 pares na zona de fragilidade dupla. Os casos mais críticos incluem minério de ferro exportado por São Luís (US$ 135 bilhões, **HHI da cadeia** no eixo vertical = 0,32), petróleo bruto (cadeia CUCI 333, **HHI da cadeia** = 0,34, com porto principal em Itaguaí) e minério de ferro por Vitória (US$ 81 bilhões). Em todos esses casos, uma paralisação prolongada do porto principal não encontra rota alternativa viável no curto prazo. São João da Barra aparece com posicionamento extremo no eixo horizontal (**HHI do porto** = 0,64, pauta dominada por petróleo bruto e minério de ferro), confirmando que sua diversificação é mínima.

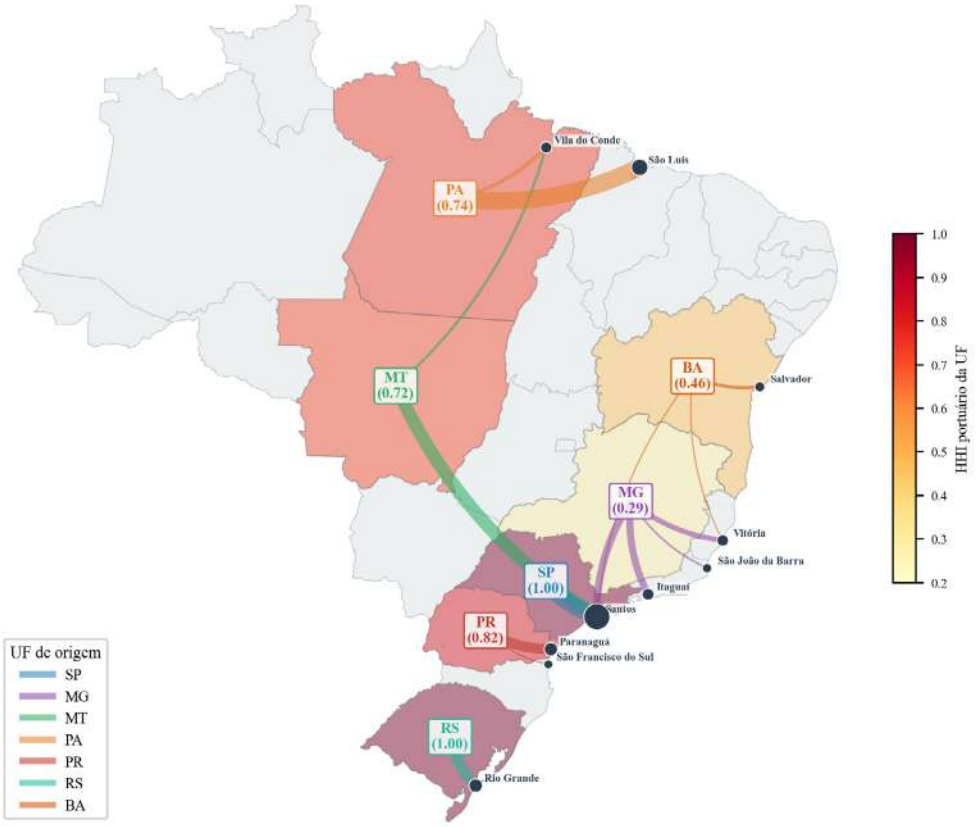
No quadrante inferior-esquerdo concentram-se pares de baixo risco estrutural: cadeias dispersas por vários portos e portos com pauta diversificada. Santos aparece tanto no quadrante de baixo risco (soja, carnes, açúcar) como fora do eixo de fragilidade, refletindo a dualidade de seu papel como *hub* generalista: seu HHI de pauta é o mais baixo da amostra (0,05), o que o protege estruturalmente, embora seja o principal porto para centenas de cadeias.

A análise retrospectiva de FOB exposto complementa essa visão estrutural com a dimensão temporal. No período 2015–2025, portos menores como Imbituba operaram com até 47,0% de seu fluxo em condições anômalas, enquanto Santos ficou em 37,7% e São João da Barra em apenas 12,9%. Essa assimetria indica que vulnerabilidade estrutural (concentração) e vulnerabilidade operacional (frequência de anomalias) são dimensões distintas que se reforçam nos nós mais críticos da rede. Cabe ressalvar que FOB exposto não equivale a perda efetiva, já que nem toda anomalia resulta em prejuízo; ainda assim, o volume de valor em trânsito durante disrupções indica a escala do risco a que cada porto está sujeito.

Essa análise retrospectiva fundamenta a lógica do painel de alerta apresentado na seção seguinte: quando o painel de vigilância sinaliza atividade anormal num porto, o gestor pode consultar o perfil histórico de exposição daquele porto para estimar rapidamente quanto valor de exportação está potencialmente em risco, segmentado por tipo de evento, produto e UF de origem.

A análise por região de destino, no recorte acumulado 2015–2025 dos arquivos de exposição do pipeline, mostrou que a Ásia absorve 57,3% do FOB exportado pelos 16 portos mapeados, seguida pelas Américas (20,3%) e Europa (18,7%). A Figura 8 ilustra a estrutura de dependência UF–porto para os sete principais estados exportadores por via marítima. Santos domina os fluxos de São Paulo (US$ 153 bilhões, HHI = 1,00) e Mato Grosso (US$ 118 bilhões, HHI = 0,72), mas Minas Gerais distribui seus fluxos entre Santos, Itaguaí e Vitória (HHI = 0,29). O Paraná apresenta dependência extrema de Paranaguá (90% do FOB, HHI = 0,82), configurando um dos pontos de maior fragilidade UF–porto do sistema. O Rio Grande do Sul depende integralmente de Rio Grande (HHI = 1,00), porto que demonstrou vulnerabilidade a eventos climáticos em 2024.

**Figura 8: Fluxos de exportação UF → Porto para os sete principais estados exportadores. Largura dos arcos proporcional ao FOB acumulado (2015–2025). Intensidade do preenchimento das UFs por concentração portuária (HHI). Tamanho dos marcadores portuários proporcional ao FOB total recebido.**




**Fonte:** Elaboração própria com dados ComexStat/MDIC (2026).

## *3.4 Sistema de alerta*

As seções anteriores construíram três camadas analíticas complementares: a detecção e classificação de anomalias operacionais a partir das séries ANTAQ (Seção 3.1–3.2), o mapeamento de cadeias produtivas e UFs expostas a fragilidades portuárias via ComexStat (Seção 3.3), e a integração com o índice global PortWatch para distinguir choques domésticos de externos. Isoladamente, cada camada responde a uma pergunta distinta; combinadas, elas permitem uma quarta pergunta operacional: *quando uma disrupção está em curso, quais cadeias, produtos e estados são afetados, e qual a magnitude da exposição?*

Hoje, essa pergunta não encontra resposta numa ferramenta única. O painel estatístico da ANTAQ publica dados operacionais sem camada de detecção; o PortWatch monitora atividade global sem vínculo com cadeias exportadoras nacionais; o ComexStat disponibiliza fluxos comerciais sem conexão com anomalias portuárias. O dashboard desenvolvido neste trabalho reúne as três fontes numa plataforma interativa Quarto com visualizações D3.js, estruturada em duas camadas de vigilância.

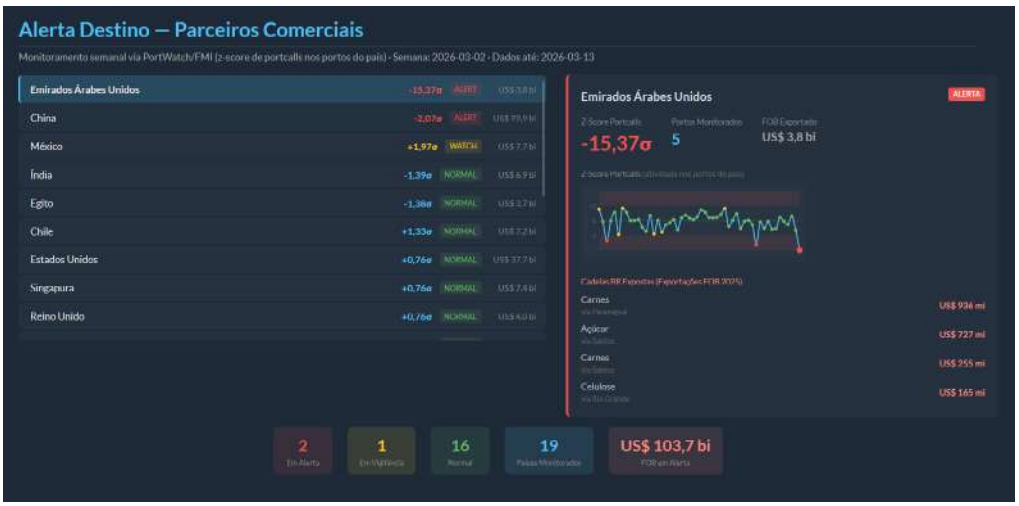
O **Painel de Alerta Origem** (Figura 9) oferece vigilância operacional dos portos brasileiros utilizando *z-scores* semanais do PortWatch contra uma *baseline rolling* de 52 semanas. Para cada porto, o painel exibe o *z-score* atual, a contagem histórica de anomalias com barra de classificação (isolado/nacional/global), e o *fingerprint* dimensional como radar chart. O *status* é categorizado em três níveis: Alerta (|z| > 1,5), Vigilância (|z| > 1,0) e Normal.

**Figura 9: Painel de Alerta Origem: z-scores semanais PortWatch e classificação de anomalias por porto. Fonte: Elaboração própria com dados ANTAQ (2026) e PortWatch/FMI (2026).**




O **Painel de Alerta Destino** (Figura 10) monitora disrupções nos portos de destino das exportações brasileiras. A partir dos dados PortWatch de ~2.000 portos globais e do mapeamento ComexStat de fluxos por produto e UF de origem, o painel identifica quais cadeias exportadoras estão expostas quando um porto de destino apresenta atividade anormal. Isso permite antecipar impactos sobre exportadores brasileiros mesmo quando a disrupção ocorre no destino.

**Figura 10: Painel de Alerta Destino: monitoramento de disrupções nos portos de destino e exposição das cadeias exportadoras por UF. Fonte: Elaboração própria com dados ANTAQ (2026) e PortWatch/FMI (2026).**



Para ilustrar um fluxo hipotético de uso (independente do estado instantâneo exibido na captura da Figura 9): suponha que Paranaguá entre em *status* de Alerta no painel de origem (z-score > 1,5). O *fingerprint* dimensional indica concentração em tonelagem de exportação com tempo no berço normal, o que aponta para interrupção de carga terrestre, não congestionamento portuário. A barra de classificação histórica revela que anomalias anteriores nesse porto foram predominantemente nacionais. No painel de destino, o analista verifica se há disrupções simultâneas nos portos asiáticos que absorvem cerca de 57% das exportações brasileiras no recorte de referência. Com o perfil de concentração à mão (Paraná depende 90% de Paranaguá, HHI = 0,82), a estimativa de FOB mensal em risco é da ordem de US$ 8 bilhões, concentrados em soja, carnes e farelo. Esse percurso, que hoje exigiria consultas separadas à ANTAQ, ao PortWatch e ao ComexStat, é realizado integralmente no *dashboard*.

## 4. DISCUSSÃO

Os resultados apresentados devem ser interpretados à luz da literatura sobre resiliência portuária e das limitações metodológicas do trabalho.

**Detecção e classificação: desempenho em contexto.** O *recall* de detecção obtido (Seção 3.2) situa o sistema numa faixa operacionalmente útil. A maioria dos eventos disruptivos conhecidos é identificada automaticamente. Os 29% não detectados concentram-se em dois casos: eventos de impacto difuso, como a Greve da Receita Federal, cujo efeito recai sobre tempos administrativos de liberação de carga que a ANTAQ não captura; e eventos no fim da série, com dados insuficientes pós-evento. A taxa de detecção de 4,0% é coerente com a fração de semanas genuinamente anômalas esperada para o período e calibra o trade-off entre recall e precisão reportado acima. Num eventual regime de monitoramento contínuo (o sistema de alerta da Seção 3.4), essa taxa também teria implicação operacional: taxas mais altas tenderiam a gerar fadiga de alerta, enquanto taxas mais baixas comprometeriam a cobertura.

A precisão da classificação indica que, uma vez detectada a anomalia, o sistema *dual-score* acerta a natureza do evento na grande maioria dos casos. Os erros residuais concentram-se na fronteira entre nacional e global, em anomalias com *Score B* próximo ao limiar de 0,8. Somam-se a eles os erros do período pré-2019, em que a classificação opera só com o *Score A* e, por construção, superestima a categoria nacional. Essa limitação é documentada e pode ser atenuada caso o PortWatch retroalimente dados históricos.

**Por que 57% das anomalias são nacionais?** Parte da resposta é metodológica. O *Score A* capta coocorrência entre portos brasileiros, e eventos como safras atípicas, mudanças regulatórias e flutuações cambiais afetam vários portos ao mesmo tempo sem configurar choque global. Mas há também uma explicação estrutural, coerente com Marchetti; Ferreira (2012): os elevados custos logísticos do sistema portuário brasileiro, estimados em 11% do PIB, revelam gargalos crônicos de acesso terrestre que reduzem a folga operacional do sistema. Com pouca folga, perturbações que cadeias mais redundantes absorveriam acabam gerando anomalia detectável.

**Vulnerabilidade cruzada e implicações para política pública.** Os pares de dupla concentração identificados na Seção 3.3 têm implicação direta para a gestão de riscos da cadeia exportadora. Os casos mais críticos, como o petróleo bruto dependente de três portos (HHI da cadeia = 0,34), configuram o que Verschuur et al. (2022) denominam risco sistêmico portuário: disrupções nesses nós propagam-se por setores e geografias sem rota alternativa viável no curto prazo.

O níquel, com 99% do FOB concentrado em Santos, ilustra um caso complementar que não aparece na Figura 7. Como Santos tem pauta diversificada (HHI = 0,05), a vulnerabilidade é unilateral: a cadeia depende do porto, mas não o contrário. Esse risco soma-se à concentração geográfica de destinos. No mesmo recorte acumulado 2015–2025, a Ásia absorve 57,3% do FOB total e Itaguaí direciona 80,6% do seu FOB à Ásia, o que cria uma dupla concentração — no destino agregado e no nó de Itaguaí — com exposição a eventos como os ataques Houthi (nov/2023–jun/2024), que o Painel de Alerta Destino foi concebido para monitorar.

***Fingerprints* como ferramenta de diagnóstico.** As assinaturas dimensionais distinguem os eventos de referência de forma visível e são coerentes com o conceito de resiliência diferenciada de Notteboom; Pallis; Rodrigue (2021): cada tipo de choque ativa mecanismos de transmissão distintos na cadeia logística. O radar tem limites. Mede *frequência* de detecção por dimensão, não magnitude ou direção, e janelas amplas podem sobrepor perturbações independentes. Foi o que ocorreu nas Enchentes do RS, quando portos distantes do evento também flagraram anomalias no mesmo período. A interpretação causal exige, portanto, validação com informações contextuais.

**Limitações.** O sistema tem quatro limitações estruturais. Primeira, opera apenas sobre variáveis operacionais portuárias: choques financeiros, comerciais ou geopolíticos manifestam-se com defasagem e magnitude incertas, e o pipeline detecta a manifestação, não a causa. Segunda, a validação *walk-forward* com janela expansível incorpora eventos passados ao treino, o que pode reduzir a sensibilidade a eventos similares subsequentes, como em Paranaguá durante a recuperação da COVID-19. Terceira, os limiares de classificação (ta=3, tb=0,8) foram calibrados contra 10 eventos específicos e não são universais, devendo ser revistos periodicamente em uso operacional. Quarta, Porto Alegre e Barra do Riacho não possuem URF dedicada no ComexStat, o que impede o cruzamento com dados de exportação. Soma-se a isso a defasagem de publicação da ANTAQ, que inviabiliza detecção em tempo real por essa via. Esse papel cabe ao PortWatch, de atualização semanal, enquanto a classificação retrospectiva exige o fechamento das janelas de co-ocorrência (4 semanas) e de choque global (2 semanas).

## CONCLUSÃO

Este trabalho desenvolveu e validou um sistema de monitoramento e diagnóstico de eventos disruptivos nos portos brasileiros de navegação de longo curso, integrando microdados operacionais da ANTAQ, indicadores globais do PortWatch/FMI e fluxos de exportação do ComexStat/MDIC.

O pipeline proposto, composto por modelagem preditiva XGBoost com correção AR(1), detecção por ensemble $\geq 2/3$ e classificação dual-score, demonstrou capacidade de identificar e classificar anomalias operacionais em 18 portos ao longo de uma década, com desempenho validado contra eventos históricos conhecidos (Seção 3.2). A integração com dados de comércio exterior revelou dezenas de pares críticos produto-porto de dupla concentração, onde disrupções representam risco econômico sistêmico para as cadeias exportadoras brasileiras. A métrica de FOB exposto, o valor de exportação que transitou por cada porto durante meses com anomalia detectada, revelou que portos menores operam com parcela proporcionalmente elevada do seu fluxo em condições anômalas, invertendo a hierarquia de risco sugerida pelo volume absoluto e demonstrando que a vulnerabilidade relativa é frequentemente maior onde a escala é menor.

A principal contribuição é demonstrar que dados públicos já disponíveis podem ser integrados num sistema de diagnóstico portuário operacional sem coleta primária. A combinação de ANTAQ, PortWatch e ComexStat gera uma capacidade que nenhuma dessas fontes oferece isoladamente: detectar uma anomalia, classificar se ela é local ou sistêmica, e estimar quanto valor de exportação está exposto. O *framework* de classificação dual (Score A + Score B) automatiza a primeira distinção; a métrica de FOB exposto traduz a segunda em termos econômicos concretos, revelando assimetrias que o volume absoluto esconde. O *dashboard* interativo, um dos produtos deste trabalho, reúne essas camadas numa ferramenta que pode ser operada por gestores sem formação em ciência de dados.

Os achados sinalizam pontos que merecem atenção na formulação de política de infraestrutura logística. A existência de dezenas de pares produto-porto com concentração elevada em ambos os eixos, associada à dependência de estados como Paraná (90% do FOB via Paranaguá) e Rio Grande do Sul (100% via Rio Grande) de um único terminal, sugere que critérios de concentração de cadeias podem ser um complemento relevante ao volume movimentado na avaliação de prioridades de investimento. Portos altamente especializados, como São João da Barra, representam ao mesmo tempo pontos de fragilidade e potenciais candidatos a estratégias de diversificação, hipótese que demandaria estudos específicos de viabilidade logística e econômica.

O código-fonte completo do pipeline está disponível no repositório já citado, permitindo a replicação e extensão dos resultados.

Como direções futuras, destacam-se: a extensão do sistema para monitoramento contínuo com atualização automática dos dados ANTAQ; a incorporação de controles sintéticos como camada adicional de validação; a calibração de limiares por dimensão e por porto, em vez de limiares globais; e a integração de variáveis exógenas (câmbio, preço de commodities, sazonalidade agrícola) como features do modelo preditivo, ampliando a capacidade de antecipação do sistema.

## Disponibilidade de Dados e Código

Este trabalho adere às práticas de Ciência Aberta. Todo o código-fonte do pipeline — coleta, tratamento, modelagem preditiva, detecção e classificação de anomalias, análise de vulnerabilidade e construção do dashboard — está disponível em repositório público no GitHub: https://github.com/fabiokp/tcc-mba-portos. O repositório inclui um arquivo README.md com instruções de instalação de dependências, execução dos scripts e reprodução dos resultados e figuras apresentados neste artigo.

Os dados processados gerados pelo trabalho (séries temporais semanais por porto e dimensão, resíduos dos modelos, anomalias detectadas e classificadas, e métricas de concentração porto-cadeia) estão depositados em repositório de acesso aberto com DOI: https://doi.org/10.5281/zenodo.20946715 (registro: https://zenodo.org/record/20946715). Esses conjuntos permitem reproduzir as análises a partir das etapas intermediárias do pipeline, sem necessidade de reprocessar as bases brutas.

As bases brutas não são redistribuídas, por serem de autoria e governança de terceiros e estarem disponíveis publicamente em suas fontes oficiais: o Anuário Estatístico Aquaviário da ANTAQ (https://web3.antaq.gov.br/ea/sense/download.html), o ComexStat do MDIC (http://comexstat.mdic.gov.br/) e o PortWatch do FMI e da Universidade de Oxford (https://portwatch.imf.org/).

Para assegurar a reprodutibilidade, o repositório inclui um manifesto de proveniência (raw_file_provenance.csv) que registra, para cada arquivo bruto utilizado, a fonte, o nome do arquivo, a data de obtenção e o hash criptográfico SHA-256, permitindo verificar a integridade e a versão exata dos dados originais. Ao todo, o manifesto documenta 44 arquivos, somando aproximadamente 2,5 GB, com cobertura de 2010 a 2026 (ANTAQ), 2011 a 2026 (ComexStat) e 2019 a 2026 (PortWatch). No caso do ComexStat, a base pública é disponibilizada desde 1997, mas o recorte analítico adotado no artigo é 2014–jan/2026.

Os dados utilizados são de natureza operacional e estatística agregada, não contendo dados pessoais na acepção da Lei Geral de Proteção de Dados (Lei nº 13.709/2018).

## REFERÊNCIAS


AGGARWAL, Charu C. **Outlier Analysis**. 2nd. ed. *[S.l.]*: Springer, 2017.

AHMAD, Subutai *et al.* Unsupervised Real-time Anomaly Detection for Streaming Data. **Neurocomputing**, v. 262, p. 134–147, 2017.

ARSLANALP, Serkan *et al.* PortWatch: Monitoring Trade Disruptions from Space. *In*: 2024. Disponível em: <https://www.bis.org/ifc/publ/ifcb62_20.pdf>

BONIOL, Paul *et al.* Dive into Time-Series Anomaly Detection: A Decade Review. **arXiv preprint arXiv:2412.20512**, 2024.

CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. Anomaly Detection: A Survey. **ACM Computing Surveys**, v. 41, n. 3, p. 1–58, 2009.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. *In*: 2016.

CHOPRA, Sunil; SODHI, ManMohan S. Managing Risk to Avoid Supply-Chain Breakdown. **MIT Sloan Management Review**, v. 46, n. 1, p. 53–61, 2004.

CHRISTOPHER, Martin; PECK, Helen. Building the Resilient Supply Chain. **The International Journal of Logistics Management**, v. 15, n. 2, p. 1–14, 2004.

CLEVELAND, Robert B. *et al.* STL: A Seasonal-Trend Decomposition Procedure Based on Loess. **Journal of Official Statistics**, v. 6, n. 1, p. 3–73, 1990.

IMF; UNIVERSITY OF OXFORD. **PortWatch: Monitoring Trade Disruptions from Space**., 2024. Disponível em: <https://portwatch.imf.org/>

LEYS, Christophe *et al.* Detecting Outliers: Do Not Use Standard Deviation around the Mean, Use Absolute Deviation around the Median **Journal of Experimental Social Psychology**, v. 49, n. 4, p. 764–766, 2013.

LIU, Fei Tony; TING, Kai Ming; ZHOU, Zhi-Hua. Isolation Forest. *In*: 2008.

MARCHETTI, Dalmo dos Santos; FERREIRA, Tiago Toledo. **Situação atual e perspectivas da infraestrutura de transportes e da logística no Brasil**. Rio de Janeiro: Banco Nacional de Desenvolvimento Econômico e Social (BNDES), 2012. Disponível em: <https://web.bndes.gov.br/bib/jspui/handle/1408/1981>.

MDIC. **ComexStat: Sistema de Análise das Informações de Comércio Exterior**., 2026. Disponível em: <http://comexstat.mdic.gov.br/>

NOTTEBOOM, Theo; PALLIS, Athanasios; RODRIGUE, Jean-Paul. Disruptions and Resilience in Global Container Shipping and Ports: The COVID-19 Pandemic versus the 2008–2009 Financial Crisis. **Maritime Economics & Logistics**, v. 23, p. 179–210, 2021.

PUNIA, Sushil; SINGH, Surya Prakash; MADAAN, Jitendra K. From Predictive to Prescriptive Analytics: A Data-driven Multi-item Newsvendor Model. **Decision Support Systems**, v. 136, p. 113340, 2020.

QU, Shen *et al.* Modeling the Dynamic Impacts of Maritime Network Blockage on Global Supply Chains. **The Innovation**, v. 5, n. 4, p. 100653, 2024.

RHOADES, Stephen A. The Herfindahl-Hirschman Index. **Federal Reserve Bulletin**, v. 79, p. 188–189, 1993.

VERSCHUUR, Jasper *et al.* A Systemic Risk Framework to Improve the Resilience of Port and Supply-Chain Networks to Natural Hazards. **Maritime Economics & Logistics**, v. 24, n. 3, p. 489–506, 2022.

VERSCHUUR, Jasper; KOKS, Elco E.; HALL, Jim W. Port Disruptions Due to Natural Disasters: Insights into Port and Logistics Resilience. **Transportation Research Part D: Transport and Environment**, v. 85, p. 102393, 2020.

VERSCHUUR, Jasper; KOKS, Elco E.; HALL, Jim W. Global Economic Impacts of COVID-19 Lockdown Measures Stand Out in High-Frequency Shipping Data. **PLOS ONE**, v. 16, n. 4, p. e0248818, 2021.

VERSCHUUR, Jasper; KOKS, Elco E.; HALL, Jim W. Ports' Criticality in International Trade and Global Supply-Chains. **Nature Communications**, v. 13, p. 4351, 2022.

VERSCHUUR, Jasper; KOKS, Elco E.; HALL, Jim W. Systemic Risks from Climate-Related Disruptions at Ports. **Nature Climate Change**, v. 13, p. 804–806, 2023.

WENDLER-BOSCO, Valeria; NICHOLSON, Charles. Port Disruption Impact on the Maritime Supply Chain: A Literature Review. **Sustainable and Resilient Infrastructure**, v. 5, n. 6, p. 378–394, 2020.

## DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL

Utilizei o modelo **Claude Opus 4.6** (Anthropic) como ferramenta de apoio em duas modalidades: via interface web, para revisão de literatura, desenvolvimento metodológico e revisão editorial do texto; e via GitHub Copilot no VSCode, para implementação do código Python de coleta, tratamento, modelagem e visualização, além da construção do *dashboard* Quarto. Todas as saídas do modelo foram verificadas contra as fontes primárias (artigos, documentação das APIs, dados brutos) e os resultados computacionais foram validados por inspeção direta dos notebooks e confronto com estatísticas descritivas independentes. Todas as decisões metodológicas, interpretações e conclusões são de minha responsabilidade.

## APÊNDICE A: TIPOLOGIA DOS PORTOS MONITORADOS

A clusterização dos 18 portos brasileiros monitorados, realizada via K-Means (k=4) sobre 12 features operacionais normalizadas (volume médio, coeficiente de variação, tempo de espera e atracação, taxa de anomalias), identificou quatro perfis estruturais descritos a seguir.

**Cluster 0: Hubs exportadores** (Itaguaí, Santos, São Luís, Vila do Conde, Vitória): média de 36 atracações de longo curso por semana e tonelagem de exportação elevada (1,9 milhão t/semana). Tempos de espera acima da média (82 h), refletindo congestionamento crônico. Das anomalias nesses portos, 41,2% são classificadas como globais e 56,6% como nacionais, indicando alta sensibilidade tanto a choques nas rotas internacionais quanto a disrupções domésticas.

**Cluster 1: Nichos voláteis** (Imbituba, Manaus, São João da Barra): volume baixo (5 atracações/semana), mas alta variabilidade (CVatracações=0,54; CVtonelagem=0,73). São os portos com maior participação relativa de anomalias isoladas (5,7%), coerente com operações especializadas sujeitas a oscilações idiossincráticas. Os nichos concentram a maior proporção de anomalias globais entre os clusters (45,3%). Há heterogeneidade intracluster: Manaus concentra 68,4% de anomalias globais, enquanto São João da Barra registra 18,3%; o valor do cluster reflete essa composição.

**Cluster 2: Generalistas** (Barra do Riacho, Itajaí, Paranaguá, Pecém, Rio de Janeiro, Rio Grande, Salvador, São Francisco do Sul, Suape): grupo mais numeroso (9 portos), com volume intermediário (20 atracações/semana) e diversificação de cargas. Tempos operacionais mais baixos (T1 = 21 h). Predominam anomalias nacionais (60,5%) no agregado do cluster, sugerindo maior sensibilidade a fatores domésticos (greves, safras atípicas, gargalos rodoviários) do que a choques nas rotas marítimas globais. Barra do Riacho é exceção dentro do grupo, com predominância de anomalias globais; trata-se de heterogeneidade intracluster, e não de contradição do padrão agregado.

**Cluster 3: Microporto** (Porto Alegre): operação de longo curso residual (2 atracações/semana, tonelagem média de 1,4 mil t/semana), com variabilidade extrema (CV=0,77 para atracações, CV=3,54 para tonelagem). Taxa de anomalias mais baixa (7,6%), mas com predominância de anomalias globais (57,4%) sobre nacionais (38,3%), o inverso do padrão dos demais clusters. Tratando-se de um único porto de volume residual e número reduzido de anomalias, essa proporção é estatisticamente instável e parcialmente influenciada pelo viés de cobertura pré-2019; sua leitura como maior exposição ao comércio internacional é, portanto, apenas uma hipótese a ser interpretada com cautela.

O cruzamento entre perfis de porto e classificação de anomalias confirma a utilidade da tipologia para gestão de risco diferenciada: hubs e generalistas demandam monitoramento prioritário de eventos nacionais (greves, bloqueios logísticos), enquanto nichos e microportos requerem atenção especial a choques nas cadeias globais que alimentam suas operações especializadas. Cabe registrar que os rótulos dos clusters derivam exclusivamente das features operacionais utilizadas na clusterização (volume, variabilidade, tempos e taxa de anomalias); as interpretações comerciais e causais de cada perfil, como especialização de pauta, congestionamento e sensibilidade a fatores domésticos ou externos, são hipóteses plausíveis à luz do conhecimento do setor, não inferências diretas do K-Means

# Metas e Produtividade no Serviço Público: uma avaliação quantitativa da atividade dos examinadores de marca do INPI

*Rafael Gerardo Rodrigues do Nascimento*
*Prof. Dr. Alex Lopes Pereira*

## Resumo

A transição para a Nova Gestão Pública na Administração Federal brasileira se consolidou com a institucionalização do Programa de Gestão e Desempenho (PGD), substituindo o controle burocrático de assiduidade pela gestão orientada a resultados e entregas. Entretanto, a adoção de metas de produtividade estáticas e uniformes para atividades de alta complexidade intelectual ignora as variações naturais da capacidade laborativa humana ao longo do ciclo de vida profissional. Este artigo avalia empiricamente a produtividade dos examinadores de marcas do Instituto Nacional da Propriedade Industrial (INPI), investigando como a experiência prévia, o envelhecimento cognitivo e a jornada intradiária afetam o tempo de execução do exame de pedidos de marcas. Por meio da extração e do processamento de uma base de dados contendo mais de 1 milhão de decisões publicadas entre 2022 e 2026, foram aplicados testes não paramétricos (Kruskal-Wallis, Mann-Whitney, Spearman) e modelos de aprendizado de máquina (Random Forest) com validação cruzada estratificada. Os resultados evidenciam que a identidade e a maturidade profissional do examinador consistem nas variáveis dominantes para a predição do tempo de análise, possuindo um impacto superior às características intrínsecas do processo examinado. Observou-se que o tempo de exame decresce substancialmente nos primeiros anos de carreira, estabilizando-se nas faixas de maior senioridade. Paralelamente, a análise etária demonstrou que o acúmulo de inteligência cristalizada compensa com êxito o declínio natural da inteligência fluida até a faixa dos 65 anos. Identificou-se, ainda, um fenômeno de aceleração intradiária, sugerindo a presença de adaptações heurísticas decorrentes da fadiga decisória. Conclui-se que metas de desempenho universais são ineficazes, propondo-se uma metodologia de metas dinâmicas baseadas em níveis de experiência, visando uma gestão pública mais justa e baseada em evidências.

**Palavras-chave:** produtividade; metas de desempenho; gestão pública

## INTRODUÇÃO

A otimização da força de trabalho é tema recorrente na administração pública brasileira. Por um lado, novas tecnologias e melhorias de processos são incorporadas às rotinas laborais; por outro, pode ser observado um aumento na complexidade técnica e intelectual das atividades. Diante desse cenário, é fundamental que as organizações desenvolvam métodos justos, eficientes e baseados em evidências para a mensuração e o acompanhamento do desempenho dos servidores. Essa necessidade é premente no serviço público, considerando o dever das organizações em assegurar o uso eficiente da força de trabalho.

Antes da criação do Programa de Gestão e Desempenho (PGD) do Governo Federal, diversos órgãos como o Serpro, a Receita Federal do Brasil e o Ministério da Justiça, já mantinham programas de teletrabalho bem estabelecidos. O próprio Instituto Nacional da Propriedade Industrial (INPI) iniciou seu programa-piloto em 2016.

Ao revisitar os marcos legais que fundamentam o modelo atual, destaca-se o Decreto nº 1.590 (Brasil, 1995) que, embora tratasse da jornada de trabalho dos servidores, continha o embrião do PGD ao prever que, em situações especiais com resultados efetivamente mensuráveis, o Ministro de Estado poderia autorizar programas de gestão com dispensa do controle de assiduidade. Mesmo que o foco em metas fosse então incipiente, já havia a previsão de resultados mensuráveis e acompanhamento regular.

Em 2018, a Instrução Normativa nº 1/2018 da Secretaria de Gestão de Pessoas estabeleceu critérios formais para aferição de resultados das unidades e do desempenho dos servidores. A pandemia de COVID-19 acelerou a adoção do teletrabalho, culminando na Instrução Normativa nº 65, de 2020, que introduziu a gestão baseada em tabelas de atividades e incluiu, de forma pioneira, a preocupação com a qualidade de vida dos participantes, como um dos objetivos do programa. Finalmente, em 2022, o Decreto nº 11.072 instituiu formalmente o PGD, definindo-o como instrumento para disciplinar a mensuração de atividades com foco na entrega de resultados e na qualidade do serviço público.

Nesse contexto, cabe questionar se, apesar dos sistemas de indicadores existentes, os órgãos dispõem de ferramentas metodológicas adequadas para monitorar mudanças na produtividade ao longo do tempo, garantindo o equilíbrio entre o uso otimizado de recursos e a qualidade de vida dos servidores. Metas superdimensionadas podem prejudicar a saúde dos servidores e a qualidade das entregas, enquanto metas subdimensionadas podem reduzir o engajamento e elevar custos para o Estado. Com a crescente incorporação de Inteligência Artificial nos processos, torna-se ainda mais crítico possuir métricas adequadas para mensurar os efeitos de sua adoção sobre a produtividade.

Este artigo propõe uma avaliação quantitativa da métrica relativa ao exame de pedidos de registro de marca, atividade central dos Tecnologistas em Propriedade Industrial do INPI. O objetivo principal é verificar se a experiência profissional se associa a uma redução significativa no tempo de execução e, caso positivo, se essa evidência fundamenta a possibilidade de metas progressivas baseadas na senioridade. Como objetivos secundários, investiga-se o efeito da faixa etária e a dinâmica de produtividade ao longo da jornada de trabalho.

## 1. REFERENCIAL TEÓRICO

### 1.1 A atividade de exame de marcas

O exame de pedidos de registro de marcas é uma atividade de cunho predominantemente intelectual, balizada pela Lei de Propriedade Industrial (Lei nº 9.279/96) e pelo Manual de Marcas do INPI. O examinador avalia se o sinal solicitado não incorre em proibições legais ou infringe direitos anteriores de terceiros. Caso identifique proibições, o pedido é indeferido, cabendo recurso à segunda instância administrativa, área que não integra o escopo deste estudo.

Dados da Organização Mundial da Propriedade Intelectual indicam que, em 2024, o INPI foi o quinto maior escritório de marcas do mundo em volume de depósitos, com 468.667 pedidos, atrás de China, Estados Unidos, Rússia e Índia. Esses cinco escritórios responderam por 61,5% do total global. O Brasil registrou crescimento de 9,7%, o maior entre os cinco maiores escritórios por volume (WIPO, 2025). Essa elevada demanda reforça a importância de métricas que reflitam a real capacidade laborativa dos servidores.

## 1.2 Produtividade, experiência e saúde no trabalho

Trigueiro-Fernandes, Hipólito e Añez (2018) argumentam que a eficiência nas organizações públicas depende do redesenho estratégico de processos e da compreensão das dimensões que afetam a produtividade, e não apenas do aumento do quadro de pessoal. A literatura sobre curvas de aprendizagem sugere que tarefas cognitivas repetitivas apresentam ganho de eficiência ao longo do tempo, com taxa de melhoria mais acentuada nos primeiros anos e tendência à estabilização (Newell & Rosenbloom, 1981).

No que se refere à relação entre idade e produtividade, Aubert e Crépon (2006), analisando dados de firmas francesas, encontraram que a produtividade cresce com a idade até aproximadamente 40 anos e se mantém estável ou superior em faixas maduras, especialmente em atividades de complexidade cognitiva. Essa dinâmica pode ser compreendida pela distinção proposta por Horn e Cattell (1967) entre inteligência cristalizada — que se acumula com a experiência — e inteligência fluida — associada à velocidade de processamento de novas informações.

Quanto aos efeitos da carga de trabalho sobre a saúde, Guedes e Rocha-de-Oliveira (2023) alertam que, no teletrabalho, a pressão constante por resultados pode gerar processos de subjetivação voltados à responsabilização individual pela própria saúde, mascarando efeitos prejudiciais em prol da produtividade.

No INPI, o exame de pedidos de registro de marcas é realizado pelos ocupantes do cargo de Tecnologista em Propriedade Industrial, que após o ingresso, via concurso público, passam por período de aproximadamente três meses em treinamento, seguido por fase de revisão entre 6 e 12 meses. Como não há formação específica prévia, todos ingressam em nível similar de conhecimento. Espera-se que, com o acúmulo de experiência, a tomada de decisão torne-se mais ágil. Entretanto, o modelo de metas atual não prevê progressão quantitativa: um examinador com seis meses de experiência possui a mesma meta de um servidor com mais de 20 anos de carreira.

## 1.3 Dimensionamento da Força de Trabalho e o ITP

A gestão de pessoas na Administração Pública Federal tem passado por uma transformação, migrando de modelos focados no controle direto, para abordagens orientadas a resultados e entregas. Nesse cenário, o Dimensionamento da Força de Trabalho (DFT) é instrumento de gestão que visa estimar, de maneira fundamentada, o quantitativo ideal de servidores para a realização de um conjunto de entregas. Segundo Serrano et al. (2018), os benefícios do DFT vão além da economicidade, proporcionando diagnóstico para decisões baseadas em evidências, favorecendo o bem-estar dos servidores e a melhoria da eficiência do Estado.

Uma variável central, para que o cálculo da produtividade no DFT seja aderente à realidade, é o Índice de Tempo Produtivo (ITP). O ITP reconhece que o servidor não dedica 100% de sua jornada formal às entregas, destinando parte da sua carga horária para outras atividades. Esse mecanismo busca evitar a sobrecarga de trabalho, dimensionando a produtividade com base no tempo líquido real de produção, e não apenas a carga horária total bruta. Conforme Serrano, Mendes e Menezes (2022), após cinco anos de testes, o ITP de 70% foi identificado como a taxa mais próxima da realidade do trabalho na Administração Pública Federal. Isso significa que 30% da jornada é absorvida por atividades como auxílio a colegas, alinhamentos e pausas fisiológicas. De modo a simplificar as análises realizadas ao longo deste artigo, optou-se por adotar o ITP referencial de 70%, sem prejuízo de que futuras avaliações utilizem padrões distintos ou busquem encontrar um índice mais preciso para a atividade específica objeto deste estudo.

### *1.4 Sistema de metas atual*

A meta dos examinadores de marca baseia-se em um sistema de pontos estabelecido há mais de uma década. O exame de pedido sem oposição vale 1 ponto e representa mais de 90% do tempo produtivo da maioria dos examinadores. A meta diária era de 16 pontos em 2014, elevada para 17 em 2017. Para participantes do PGD, a meta é acrescida de 20% na modalidade parcial (20,4 pontos) e 30% na integral (22,1 pontos). A produtividade é calculada dividindo-se a soma dos pontos das decisões e outras atividades (reuniões, capacitações) pelo número de dias efetivamente trabalhados. Por exemplo: se em uma semana um servidor tem 70 pontos de decisões em pedidos sem oposição, 4,25 pontos de reuniões e a semana possui 5 dias úteis, com 1,5 dias de licença médica, temos (70 + 4,25) / (5-1,5) = 21,21. Nesse exemplo, o servidor apresenta uma produtividade de 21,21 pontos no período.

## 2. METODOLOGIA

### *2.1 Coleta e tratamento de dados*

Foram extraídos, via SQL, dados de decisões publicadas entre 2022 e 20 de fevereiro de 2026, abrangendo inicialmente 161 tecnologistas em propriedade industrial ativos e mais de 1,24 milhão de registros. A base foi enriquecida com informações de faixa etária, tempo de experiência, posição na carreira e participação no PGD.

A variável-alvo, denominada TEMPO_AJUSTADO, foi construída em duas etapas. Primeiro, calculou-se o TEMPO_PROCESSO_MINUTOS como a diferença entre o timestamp do despacho atual e o anterior do mesmo examinador no mesmo dia, removendo-se intervalos que cruzavam a virada do dia. Em seguida, aplicaram-se fatores de ajuste conforme a modalidade do PGD: 1,0 para presencial, 1,2 para parcial e 1,3 para integral, de modo a aproximar o esforço equivalente em regime presencial. Essa métrica deve ser interpretada como *proxy* do tempo dedicado ao exame.

Foram excluídos da análise os despachos de gestores e tecnologistas cuja atividade principal não é o exame de marcas. O dataset final resultou em 1.088.220 linhas e 29 colunas. Para garantir privacidade, os *notebooks* foram executados localmente e os nomes anonimizados.

## 2.2 Filtros e ajustes

Para reduzir ruídos, foram excluídas decisões simples que não exigem análise intelectual profunda como arquivamentos, extinções e republicações de pedidos. Realizou-se tratamento para o "grupo de trâmite", funcionalidade do sistema, pouco usada por examinadores no exame de pedidos, mas que gera despachos com timestamp idêntico (tempo 0). Como o uso dessa prática poderia contaminar o tempo de outros processos no mesmo dia, optou-se por remover todos os despachos do examinador no mesmo dia em que ele usou essa funcionalidade, o que removeu 4,66% da base original (50.715 linhas).

## 2.3 Tratamento da distribuição assimétrica

Uma observação importante sobre a metodologia empregada para calcular o tempo de exame é que os processos examinados após qualquer tipo de intervalo absorvem o tempo desse intervalo, gerando ruídos na variável.

**Tabela 1 – Ilustração do efeito de absorção de intervalos pela variável de tempo.**

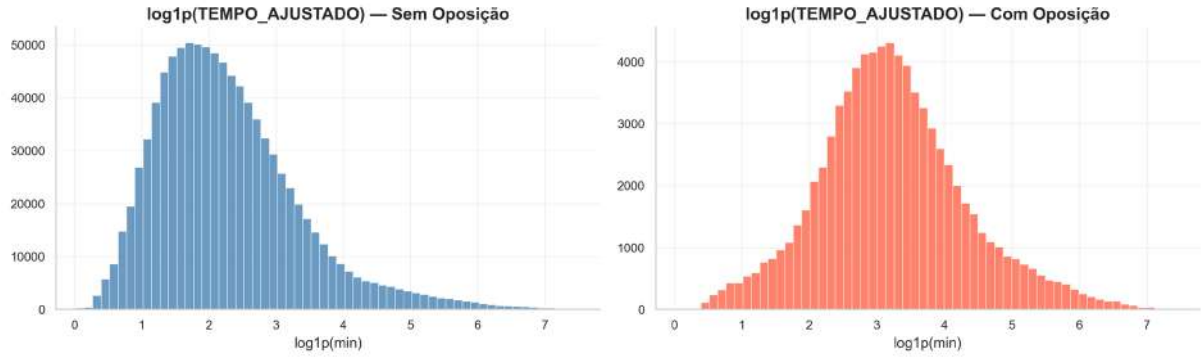
| POSIÇÃO | DATA | HORA | TEMPO AJUSTADO |
|---|---|---|---|
| 1 | xx/xx/xxxx | 09:09:30 | |
| 2 | xx/xx/xxxx | 09:23:41 | 14,18 |
| 3 | xx/xx/xxxx | 09:35:33 | 11,87 |
| 4 | xx/xx/xxxx | 09:44:54 | 9,35 |
| 5 | xx/xx/xxxx | 10:09:01 | 24,12 |
| 6 | xx/xx/xxxx | 10:20:21 | 11,33 |
| 7 | xx/xx/xxxx | 10:30:17 | 9,93 |
| 8 | xx/xx/xxxx | 11:20:00 | 49,72 |
| 9 | xx/xx/xxxx | 11:30:38 | 10,63 |
| 10 | xx/xx/xxxx | 11:56:04 | 25,43 |
| 11 | xx/xx/xxxx | 12:05:10 | 9,10 |
| 12 | xx/xx/xxxx | 12:17:02 | 11,87 |
| 13 | xx/xx/xxxx | 12:30:58 | 13,93 |
| 14 | xx/xx/xxxx | 12:41:54 | 10,93 |
| 15 | xx/xx/xxxx | 12:53:15 | 11,35 |
| 16 | xx/xx/xxxx | 15:02:21 | 129,10 |
| 17 | xx/xx/xxxx | 15:17:16 | 14,92 |
| 18 | xx/xx/xxxx | 15:23:39 | 6,38 |

**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

Ao longo do estudo foram consideradas algumas opções de lidar com essa distorção, como o uso de modelos de limpezas de *outliers*; uma análise para tentar identificar possíveis intervalos intrajornada; ou mesmo um corte em um tempo específico, que poderia ser considerado acima do normal.

Optou-se pela utilização da mediana e dos quartis como principais métricas de tendência central, dado que essas medidas são robustas aos efeitos da cauda longa. Complementarmente, para normalizar a distribuição e mitigar o efeito dessa assimetria à direita — onde poucos examinadores com tempos muito elevados deslocariam artificialmente a média para cima — aplicou-se a função log(1+x) aos dados de tempo, resultando em uma distribuição mais simétrica e adequada para as análises comparativas, como pode ser observado na Figura 1.

**Figura 1 — Distribuição do TEMPO_AJUSTADO antes e após transformação logarítmica.**




**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

## *2.4 Faixas de análise e testes estatísticos*

**Tabela 2 – Classificação das faixas de experiência profissional.**

| FAIXA_ANOS_XP | Linhas (pedidos sem oposição) |
|---|---|
| <= 1 ano | 72.332 |
| > 1 e <= 5 anos | 49.516 |
| > 5 e <= 10 anos | 245.231 |
| > 10 e <= 15 anos | 111.274 |
| > 15 e <= 20 anos | 340.232 |
| > 20 anos | 136.018 |

**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

**Tabela 3 – Classificação das faixas etárias.**

| FAIXA_ETARIA | Linhas (pedidos sem oposição) |
|---|---|
| Até 45 anos | 298.063 |
| Entre 46 e 50 anos | 170.803 |
| Entre 51 e 55 anos | 117.074 |
| Entre 56 e 60 anos | 113.323 |
| Entre 61 e 65 anos | 92.993 |
| mais de 66 anos | 45.182 |

**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

Como estamos trabalhando com uma amostra temporal superior a quatro anos (2022-2026 parcial), é importante mencionar que estas classificações se referem ao tempo de experiência e faixa etária no momento da decisão. Por exemplo, todos os despachos do examinador X, que ingressou em agosto de 2024, realizados até agosto de 2025, estarão na faixa de >= 1 ano. Consequentemente todos os despachos posteriores do mesmo examinador serão classificados na faixa >1 e <=5 anos.

Dada a natureza não paramétrica dos dados, utilizaram-se: teste de Kruskal-Wallis para comparações entre múltiplos grupos; correlação de Spearman para associação monotônica; teste de Mann-Whitney para dois grupos; qui-quadrado para composição de despachos; regressão *log-linear* para evolução intradiária; e Random Forest com validação cruzada para modelagem preditiva.

# 3. RESULTADOS E DISCUSSÃO

## 3.1 Análise descritiva

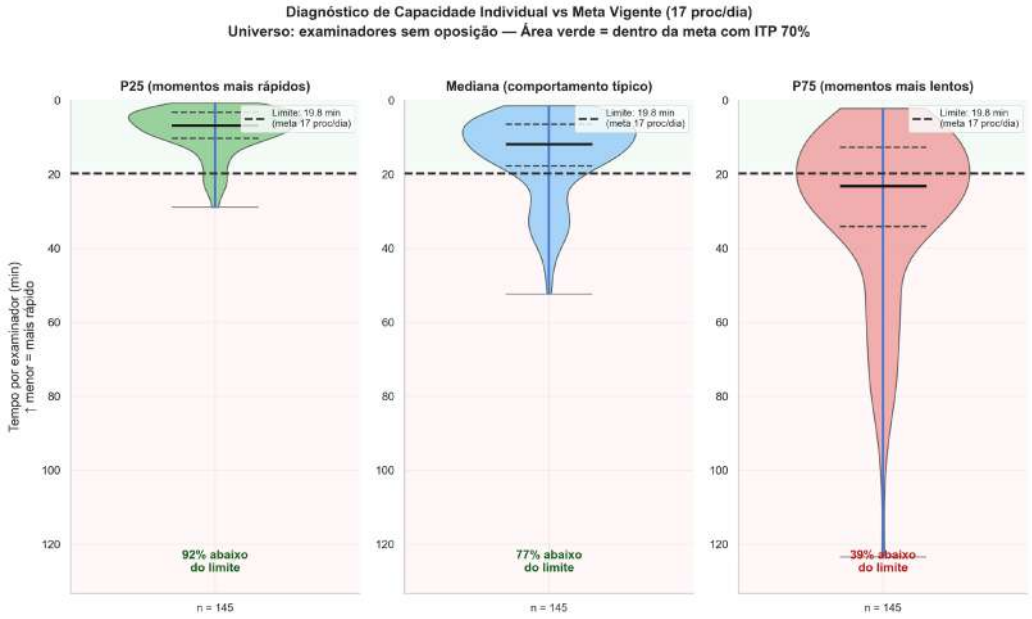
O *dataset* final compreende 954.603 decisões sem oposição (92%), 81.649 com oposição (7,9%) e 1.253 de outros tipos (0,12%). As análises concentram-se nas decisões sem oposição, processos onde não há uma contestação de terceiros. Tendo em vista que a meta atualmente em vigor é de 17 pontos e, com o ITP de 70% adotado, podemos que concluir que, em uma jornada de trabalho ordinária de 8 horas, o tempo necessário esperado para uma atividade de 1 ponto é de 19,76 minutos, aproximadamente 19 minutos e 45 segundos [(8 × 60) × 0,70 / 17].

## 3.2 Diagnóstico inicial: aderência à meta vigente

Para caracterizar a distribuição da variável 'TEMPO_AJUSTADO' no exame de pedidos de registro sem oposição, extraíram-se as seguintes medidas separatrizes para cada examinador dedicado a esse tipo de exame, durante o período avaliado: o primeiro quartil (P25), a mediana e o terceiro quartil (P75), com o objetivo de entender a dispersão e o esforço típico de cada examinador.

A Figura 2 apresenta a distribuição da densidade dos tempos nestes três cenários: P25 (momentos mais rápidos), mediana (comportamento típico) e P75 (momentos mais lentos). Nos despachos mais rápidos, 92% dos examinadores apresentaram tempo inferior ao esperado (19,76 min). Na mediana, 77% foram mais rápidos. Nos momentos mais lentos, apenas 39% mantiveram tempo inferior ao patamar de referência.

**Figura 2 — Diagnóstico da capacidade individual dos examinadores (n=145) em comparação à meta vigente de 17 pontos/dia, considerando ITP de 70%.**




**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

A partir desse resultado, podemos tirar algumas conclusões preliminares. A primeira é que, mesmo nos momentos mais rápidos, alguns examinadores ainda não possuem um tempo de exame que lhes permita alcançar a meta de maneira confortável. Por outro lado, temos mais de um terço dos examinadores (39%) que, mesmo nos momentos mais lentos, ainda apresentam um tempo de exame inferior ao tempo médio esperado, considerada a meta de 17 pontos atualmente estabelecida.

Esse diagnóstico evidencia que, no cenário atual, já existe descolamento entre a meta estabelecida e a real capacidade de exame. Entretanto, a dispersão indica que uma parcela dos examinadores opera próximo ou acima do limite mesmo nos momentos mais produtivos, reforçando a necessidade de cautela em eventuais revisões de meta.

## *3.3 Efeitos da experiência no tempo de exame.*

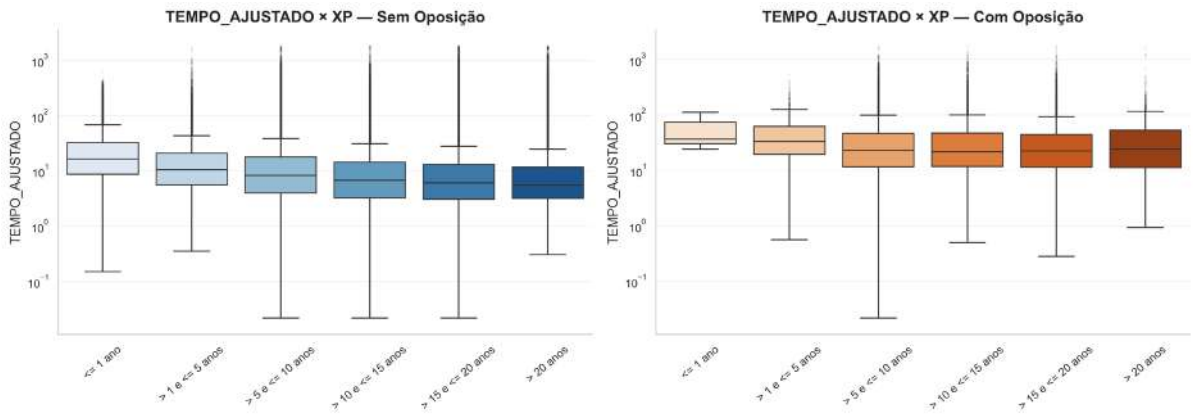
**Tabela 4 – Distribuição dos examinadores por ano de ingresso no INPI.**

| Período de Ingresso | Percentual de Examinadores |
|---|---|
| Antes de 1999 | 2,5% |
| Entre 2000 e 2010 | 28,6% |
| 2013/14 | 13,7% |
| 2017 | 17,4% |
| 2024 | 21,7% |
| 2025 | 16,1% |

**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

Pouco mais da metade dos examinadores analisados ingressou após 2017 (55,2%). Cabe ressaltar que os servidores de 2025 ingressaram ao final do ano, tendo seus despachos computados referentes aos primeiros dias de exame, estando ainda em processo de revisão e curva de aprendizagem.

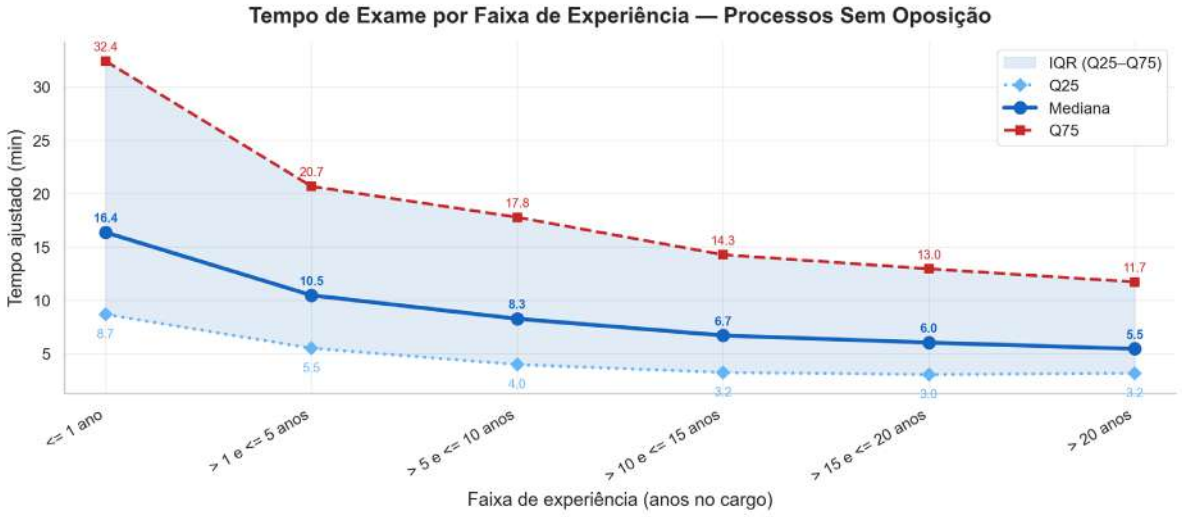
**Figura 3 — Relação entre faixa de experiência e TEMPO_AJUSTADO (escala logarítmica) para pedidos sem oposição.**




**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

A Figura 3 revela tendência clara de redução no tempo de exame conforme a senioridade aumenta. Examinadores com até 1 ano de experiência apresentam a mediana de tempo mais elevada e a maior dispersão no corpo do intervalo interquartil, o que é esperado para um período inicial de treinamento. Em contrapartida, o grupo com mais de 20 anos de experiência demonstra a maior celeridade, com a mediana situada no patamar mais baixo da escala. Nota-se que a redução no tempo de exame é mais acentuada nos primeiros 10 anos, seguida por uma estabilização nas faixas de maior senioridade.

**Figura 4 — Evolução do tempo médio de exame por faixa de experiência, com intervalo interquartílico (IQR).**




**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

A Figura 4 detalha a evolução: examinadores com até 1 ano apresentam mediana de 16,4 minutos, enquanto os mais experientes atingem 5,5 minutos — redução de aproximadamente 66%. O ganho mais expressivo ocorre na transição para a segunda faixa (16,4 para 10,5 min). Além disso, o intervalo interquartílico cai de 23,7 minutos (até 1 ano) para 8,5 minutos (acima de 20 anos), demonstrando maior homogeneidade e previsibilidade nas entregas dos servidores experientes.

O teste de Kruskal-Wallis confirmou diferença estatisticamente significativa ($H = 53.728,81$; $p < 0,001$). A correlação de Spearman entre as medianas por faixa e a ordem das faixas resultou em rho = -1,000 ($p < 0,001$). É necessário ponderar que essa correlação perfeita é esperada quando se trabalha com seis faixas ordinais cujas medianas são monotonicamente decrescentes por construção; ela confirma a tendência agregada, mas não substitui análises de tamanho de efeito no nível individual. Fica evidente um efeito de aprendizagem robusto: cada incremento na senioridade do examinador correlaciona-se diretamente com uma redução sistemática no tempo de execução.

Esses resultados indicam que a experiência não apenas amplia a celeridade, mas também reduz a variabilidade da entrega. Confrontando com a meta vigente de 17 pontos (19,76 min/ponto com ITP 70%), observa-se que todas as medianas, independentemente da faixa, situam-se abaixo desse patamar, atingindo o máximo de apenas 16,4 minutos no primeiro ano de exercício. Mesmo ao analisarmos o terceiro quartil (Q75), que representa os exames mais lentos, o tempo de meta é excedido apenas nas duas faixas iniciais de senioridade. Em suma, a vasta maioria dos exames é concluída em tempo inferior ao previsto pela meta estabelecida, sugerindo que a meta vigente pode apresentar um caráter conservador diante da capacidade produtiva real.

Contudo, a análise da capacidade produtiva não deve se restringir exclusivamente à métrica temporal. Embora o tempo de execução ofereça uma observação clara do potencial produtivo, o estabelecimento de metas deve considerar variáveis qualitativas e humanas. É necessário questionar: qual o efeito de tempos de exame reduzidos sobre o rigor técnico das decisões? Tempos mais elevados se traduzem em decisões mais acertadas ou melhor embasadas? Além disso, é imperativo avaliar o impacto de metas cada vez mais desafiadoras sobre o bem-estar e a saúde mental dos servidores.

O desafio reside em encontrar o ponto de equilíbrio entre a celeridade dos exames e a preservação da qualidade das entregas, sem negligenciar a integridade física e mental dos examinadores. O aprofundamento sobre como as metas influenciam esses dois campos — qualidade e saúde — constitui um campo fértil para trabalhos futuros.

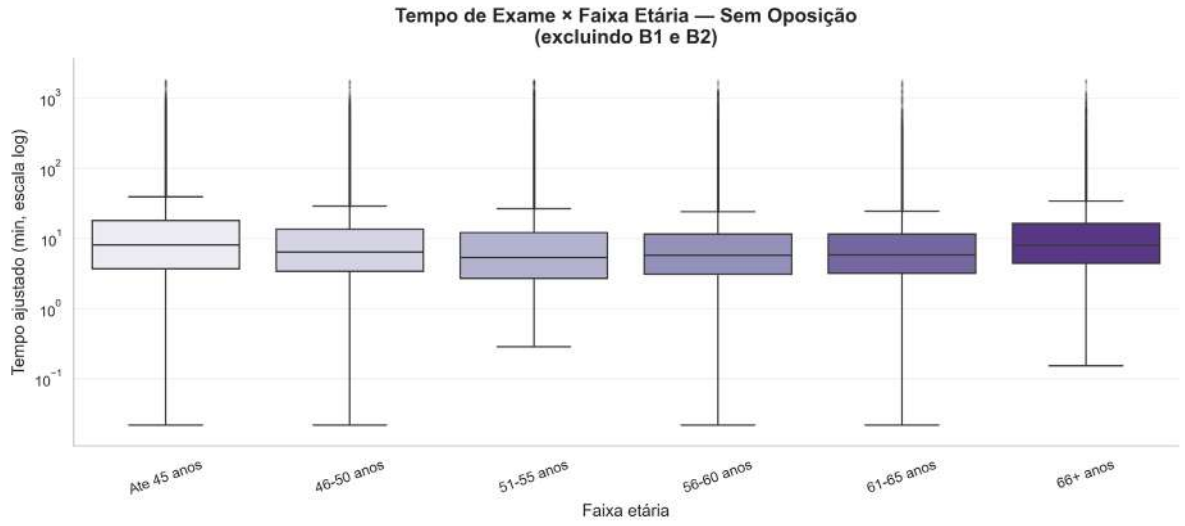
## *3.4 Efeitos da faixa etária no tempo de exame*

Embora experiência e idade sejam variáveis frequentemente correlacionadas, a ciência cognitiva sugere que elas operam por mecanismos distintos: enquanto a experiência no cargo acumula o que se chama de inteligência cristalizada, o avanço da idade pode impactar a inteligência fluida, que diz respeito à velocidade de processamento de novas informações e à agilidade mental.

Uma primeira possibilidade é que, apesar do possível impacto da idade na inteligência fluida, esse efeito pode não se refletir em um aumento no tempo de exame, graças ao ganho de experiência, que eleva a inteligência cristalizada. Contudo, caso essa hipótese não se confirme, é de grande relevância identificar em qual momento essa compensação tem seus efeitos atenuados.

Nesta etapa do estudo, foram removidos da análise os despachos dados por examinadores nos seus primeiros anos de experiência, tendo em vista que são examinadores recém-ingressos, ainda em processo de aprendizagem. Como esses servidores ainda não passaram pela curva completa de ganho de produtividade, incluí-los na análise do efeito da idade poderia distorcer o padrão, tendo em vista que nesses casos, um exame mais longo poderia refletir inexperiência, não envelhecimento.

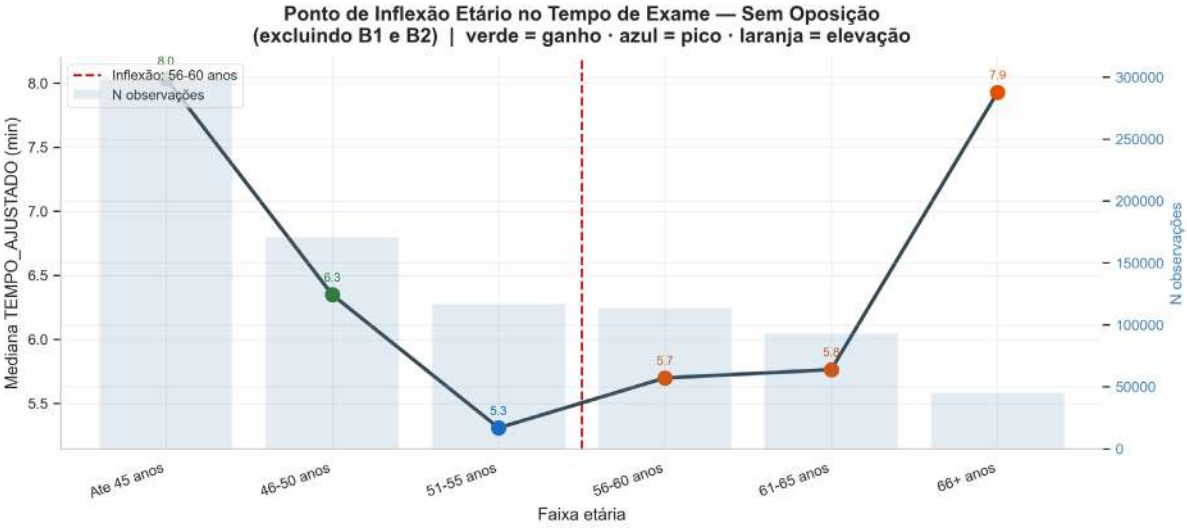
**Figura 5 — Impacto da faixa etária no TEMPO_AJUSTADO (escala logarítmica), excluindo servidores em início de carreira.**




**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

A Figura 5 mostra tendência de redução nas medianas conforme a idade avança, com pico de eficiência entre 51 e 65 anos. O grupo mais jovem (até 45 anos) apresenta a maior mediana, enquanto as faixas de 51 a 65 anos demonstram desempenho mais célere e consistente. Esse resultado corrobora Aubert e Crépon (2006) e sugere que o acúmulo de inteligência cristalizada compensa possíveis declínios na inteligência fluida.

**Figura 6 — Análise estratificada por classe funcional (C e S) e faixa etária.**




**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

Para refinar a compreensão sobre o fator etário, a análise foi estratificada por classe funcional (Classes C e S) e faixa de experiência, visando isolar o efeito da idade cronológica de outras variáveis intervenientes. Testes de Kruskal-Wallis estratificados por classe funcional foram invariavelmente significativos ($p < 0,001$), confirmando que a idade biológica exerce influência própria sobre o tempo de exame, que persiste mesmo quando comparamos indivíduos com idêntico tempo de exercício no cargo. Na Classe C, nota-se um ganho de performance expressivo com a maturidade. Examinadores entre 51 e 60 anos apresentam mediana de ~4,1 minutos, contra 8,0 minutos do grupo até 45 anos. Entretanto, observa-se ponto de inflexão na faixa acima de 66 anos, onde a mediana sobe para 7,6 minutos (Classe C) e 8,8 minutos (Classe S).

Esse fenômeno reforça a hipótese de que o acúmulo de inteligência cristalizada atua como um diferencial competitivo, permitindo que o servidor maduro execute o exame com maior economia cognitiva e celeridade. Entretanto, os resultados da correlação de Spearman para as classes não confirmam uma relação monotônica ($p > 0,05$), indicando que essa relação não é estritamente linear ao longo de toda a vida funcional. Observa-se um ponto de inflexão na faixa acima de 66 anos, onde a mediana de tempo volta a subir para 7,6 minutos na Classe C e 8,8 minutos na Classe S. Esse leve declínio na celeridade em idades mais avançadas corrobora, em estratos específicos relacionados à progressão na carreira e experiência, aquilo que observamos na amostra completa dos dados. Contudo, é fundamental destacar que, mesmo no grupo mais longevo, a performance permanece competitiva ou superior à dos ingressantes mais jovens em certos estratos.

Em termos de gestão, esses dados reforçam que o servidor maduro é um ativo de alta performance para a administração pública, possuindo um ritmo de entrega elevado, fruto da maturidade profissional aliada ao domínio do processo. Apenas na faixa acima de 66 anos nota-se um incremento no tempo, o que pode sinalizar o início de um declínio natural das funções executivas. Isso sugere que políticas de gestão de desempenho devem considerar tanto o estágio de aprendizagem quanto possíveis efeitos do envelhecimento, sobretudo nas faixas etárias mais avançadas, nas quais se observa discreta elevação do tempo de exame.

**Tabela 5 — Medianas de tempo por faixa etária e classe funcional.**

| CLASSE C H=16187.54 p < 0,001 \| Spearman rho = -0.406 p=0.4247 | | |
|---|---|---|
| FAIXA_ETARIA | Linhas (pedidos sem oposição) | mediana |
| Até 45 anos | 215.693 | 8 minutos |
| Entre 46 e 50 anos | 69.082 | 8 minutos |
| Entre 51 e 55 anos | 34.621 | 4,1 minutos |
| Entre 56 e 60 anos | 34.924 | 4,2 minutos |
| Entre 61 e 65 anos | 41.597 | 5,4 minutos |
| mais de 66 anos | 33.284 | 7,6 minutos |

| CLASSE S H=2931.63 p < 0,001 \| Spearman rho = 0.371 p=0.4685 | | |
|---|---|---|
| FAIXA_ETARIA | Linhas (pedidos sem oposição) | mediana |
| Até 45 anos | 4.664 | 7,9 minutos |
| Entre 46 e 50 anos | 70.280 | 5,4 minutos |
| Entre 51 e 55 anos | 78.663 | 5,9 minutos |
| Entre 56 e 60 anos | 78.397 | 6,4 minutos |
| Entre 61 e 65 anos | 51.376 | 6,2 minutos |
| mais de 66 anos | 11.898 | 8,8 minutos |

**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

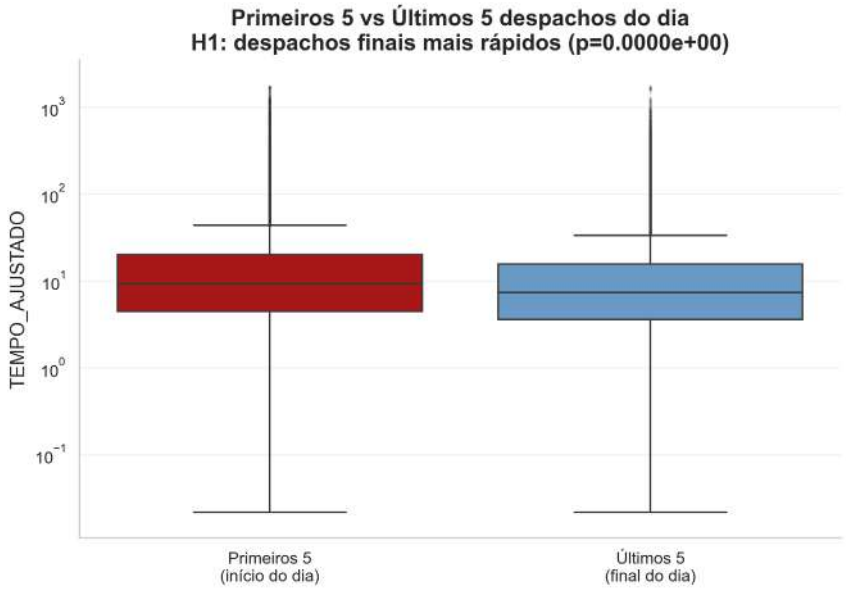
## *3.5 Dinâmica de produtividade intradiária*

A produtividade em tarefas que exigem alta carga cognitiva, como o exame de pedidos de marcas, não deve ser vista como uma constante mecânica, mas como um processo dinâmico influenciado por fatores biológicos e psicológicos ao longo da jornada de trabalho.

Para investigar a evolução do tempo ao longo da jornada, os despachos foram numerados de acordo com sua posição no dia para cada um dos examinadores, permitindo isolar a evolução do ritmo de cada servidor dentro de seu próprio ciclo diário. Também foi criada uma coluna adicional para classificar os 5 primeiros e 5 últimos processos de cada dia de cada examinador. Para garantir o rigor estatístico e evitar que atividades iniciais e finais se sobrepusessem, foram selecionadas apenas as jornadas de trabalho que registraram o despacho de pelo menos 10 processos.

Essa metodologia permite analisar coletivamente se o início do trabalho é sistematicamente mais lento ou mais rápido que o seu encerramento, fornecendo uma base sólida para os testes de hipótese apresentados a seguir. Para a comparação entre os cinco primeiros e cinco últimos despachos, foram utilizados 159.454 despachos classificados como os primeiros do dia, contra 207.133 despachos classificados como os últimos do dia. Essa diferença se dá porque a definição da posição na fila é feita antes dos filtros relacionados às decisões em pedidos sem oposição que foram utilizados nas análises.

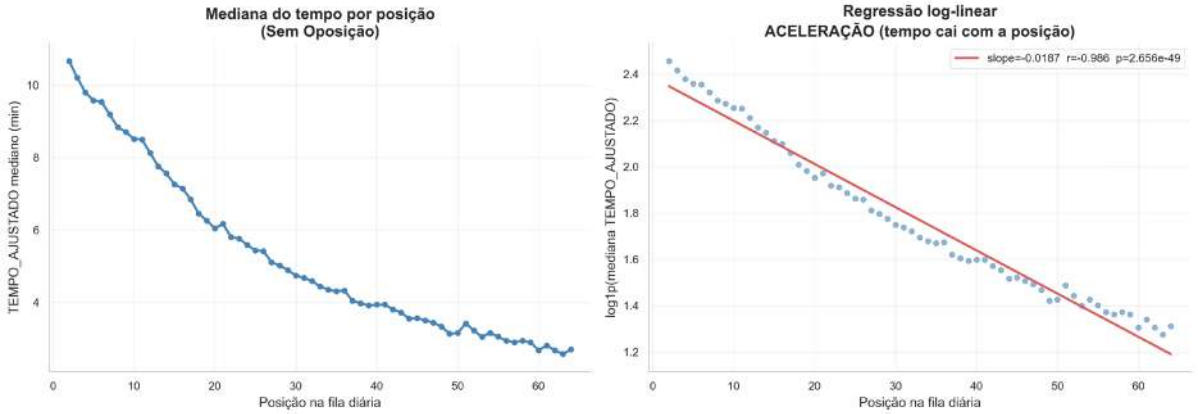
**Figura 7 — Comparação entre os 5 primeiros e 5 últimos despachos do dia (Mann-Whitney).**




**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

A mediana das cinco primeiras decisões do dia foi 9,29 minutos, contra 7,35 minutos das últimas cinco decisões — diferença de 1,95 minutos (1 minuto e 57 segundos). O teste de Mann-Whitney confirmou significância (U = 18.433.619.751; p < 0,001). Esses dados permitem afirmar, com elevada confiança, que a redução observada no tempo de exame ao longo da jornada não é um evento aleatório, mas um padrão sistêmico.

**Figura 8 — Regressão log-linear do TEMPO_AJUSTADO em função da posição na fila diária.**

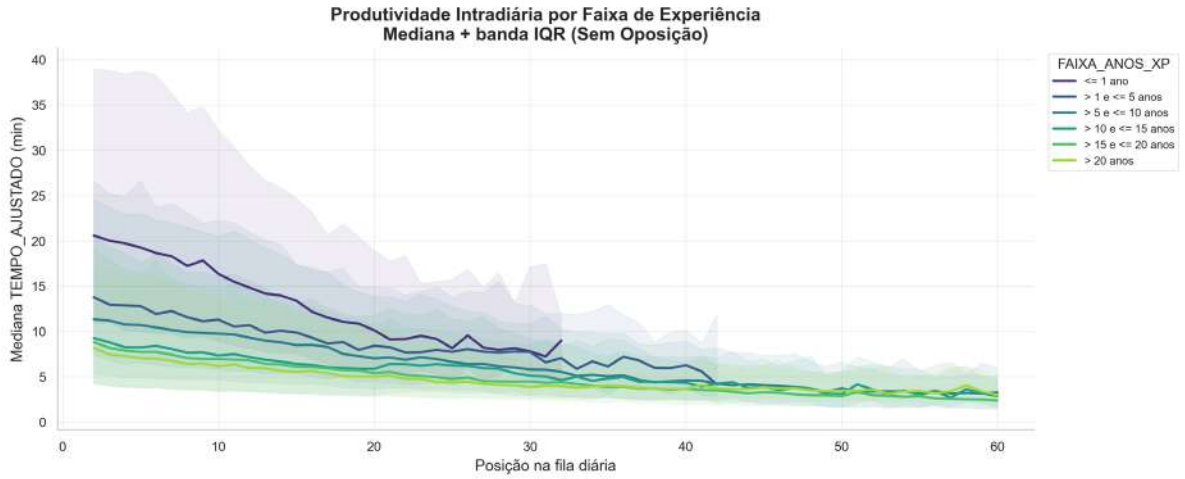



**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

A regressão *log-linear* apresentou slope de -0,01866, significando que a cada novo processo, o tempo se reduz em taxa percentual constante (r = -0,9861; p < 0,001). Isso demonstra que essa aceleração não acontece apenas ao final do dia, mas é um processo de aprimoramento que ocorre a cada despacho realizado. Esse padrão é consistente com a hipótese de aquecimento cognitivo.

Considerando as observações feitas ao longo deste estudo, referentes ao comportamento do tempo em relação a determinados arranjos, procedeu-se à geração de gráficos com objetivo de avaliar a evolução do tempo ao longo da jornada, dentro dos seguintes grupos: faixa de experiência e faixa etária.

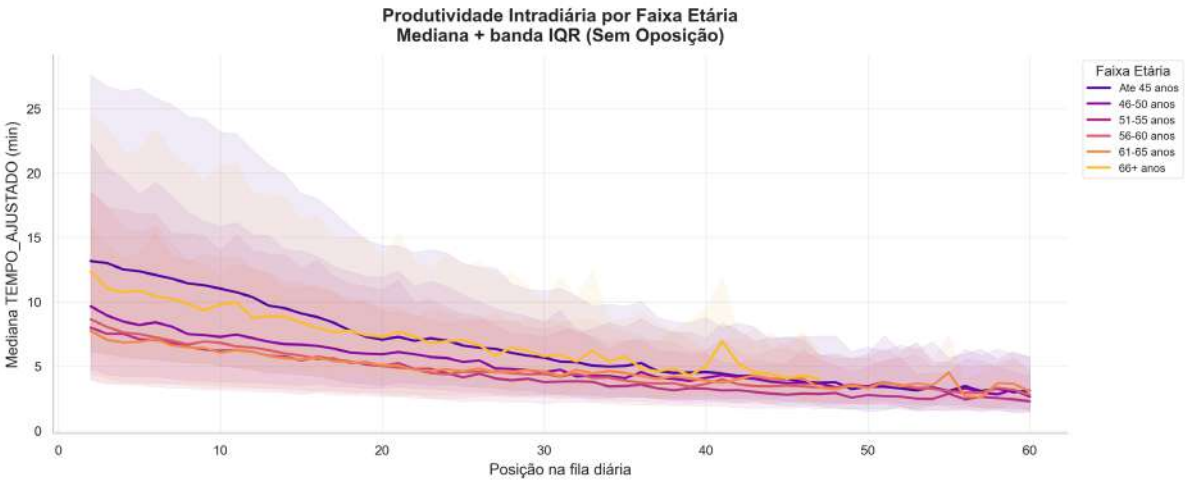
**Figura 9 — Evolução intradiária do TEMPO_AJUSTADO por faixa de experiência.**




**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

Na distribuição por experiência (Figura 9), iniciantes começam a jornada com mediana superior a 20 minutos, enquanto veteranos iniciam com ~8 minutos. O teto produtivo dos menos experientes situa-se em torno de 30 decisões diárias; servidores experientes podem, em alguns momentos, superar 60 decisões em um dia, mantendo o mesmo ritmo. Os veteranos apresentam trajetória linear e previsível, enquanto os iniciantes mostram instabilidades.

**Figura 10 — Evolução intradiária do TEMPO_AJUSTADO por faixa etária.**



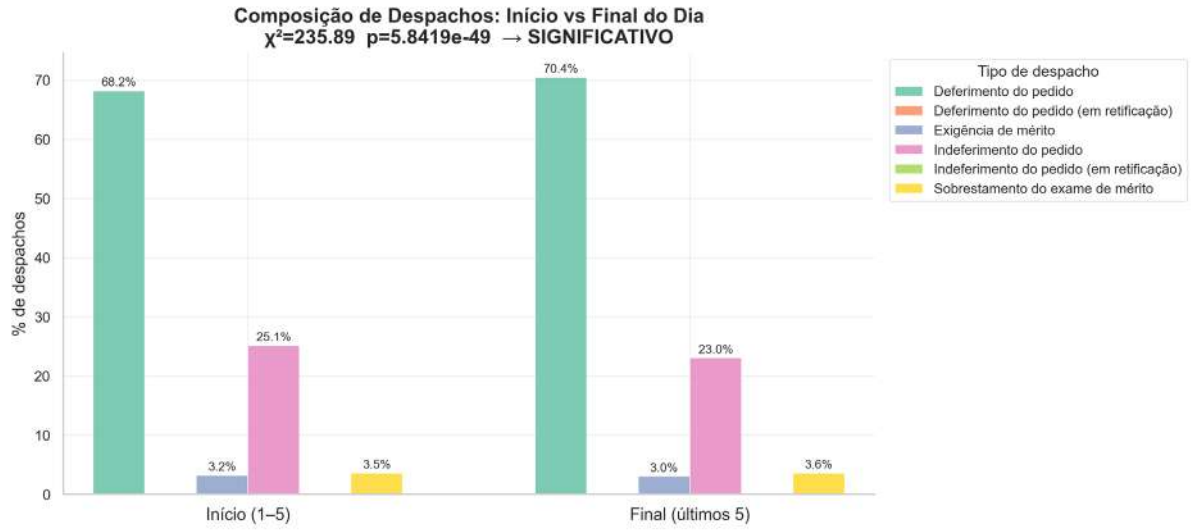

**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

No que tange à faixa etária (Figura 10), os dados corroboram a tese de que o acúmulo de inteligência cristalizada compensa o declínio da inteligência fluida em grande parte da vida funcional. As faixas entre 51 e 65 anos iniciam o dia muito mais rápidas que os examinadores mais jovens. Contudo, o grupo acima de 65 anos apresenta redução no limite de processos diários (abaixo de 50) e instabilidade na curva a partir da 40ª decisão, o que pode sinalizar onde a fadiga de decisão sobrepõe-se ao ganho de experiência.

Embora os resultados apontem para uma aceleração do ritmo de trabalho ao final da jornada, é fundamental interpretar esse fenômeno sob uma perspectiva multidimensional. A tendência de redução no tempo de exame em relação à posição na fila sugere que a imersão

na atividade intelectual reduz a inércia cognitiva inicial e permite que o servidor atinja níveis elevados de execução técnica. Esse aquecimento cognitivo pode suprimir erros e aumentar a fluidez do processamento de informações.

**Figura 11 — Análise de composição dos despachos ao longo do dia (qui-quadrado).**




**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

Entretanto, a celeridade observada não deve ser dissociada da análise qualitativa das entregas. O teste qui-quadrado revelou alteração significativa na composição dos despachos ao longo do dia ($x^2$ = 235,89; p < 0,001), com incremento de deferimentos (+2,2%) e redução de indeferimentos (-2,1%) nos processos finais. Esse achado levanta questões sobre se a aceleração reflete exclusivamente perícia técnica ou envolve adaptação heurística à fadiga decisória. Danziger, Levav e Avnaim-Pesso (2011) afirmam que o esgotamento de recursos cognitivos pode levar o decisor a optar por escolhas que exijam menor esforço de justificação. No contexto de escritórios de propriedade intelectual, Frakes e Wasserman (2020) identificaram que examinadores de patentes do escritório norte-americano (USPTO), sob pressão de metas de produtividade, tendiam a emitir rejeições fracamente fundamentadas (shotgun rejections), facilmente reversíveis em instâncias posteriores, como estratégia para cumprir metas dentro do prazo. Em estudo subsequente, Frakes e Wasserman (2024) demonstraram que a adoção de incentivos contínuos diários, em substituição ao sistema de cotas com deadlines rígidos, reduziu substancialmente esses efeitos sem comprometer a acurácia das decisões. Embora esses estudos tratem do exame de patentes — atividade com características distintas do exame de marcas —, os mecanismos de fadiga decisória e pressão por metas são análogos ao contexto aqui investigado.

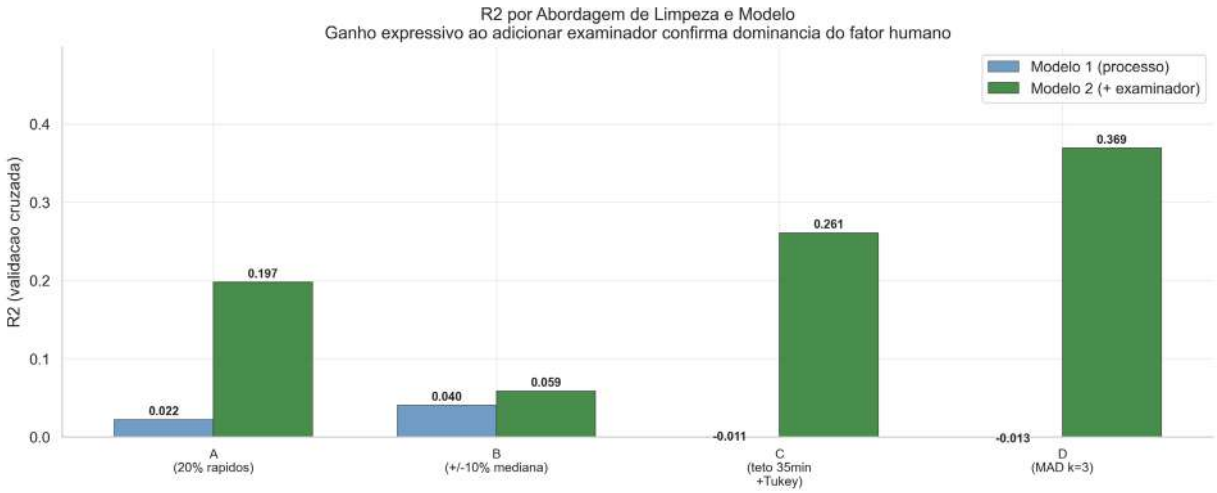
## *3.6 Modelagem preditiva e variância de desempenho*

Para consolidar a hipótese de que o examinador é a variável dominante, treinaram-se dois modelos Random Forest em quatro abordagens de amostragem: (A) examinadores 20% mais rápidos com Tukey; (B) examinadores ±10% da mediana com Tukey; (C) corte de decisões com tempo superior a 35 min + Tukey; (D) MAD k=3. O Modelo 1 utilizou apenas variáveis do pedido (apresentação, tipo de especificação, adequação, volume da classe, marca repetida, palavras da marca) com GroupKFold. O Modelo 2 adicionou a identificação do examinador com KFold simples.

Para as amostras A e B, o objetivo é verificar o comportamento dos modelos em grupos com características de tempos de exame similares. Já na amostra C, a opção pelo corte dos exames que superam o tempo de 35 minutos, entendendo que tempos superiores seriam incompatíveis com tempo de atenção contínua para essa atividade. Nesses três casos foi aplicado o método de Tukey para a remoção de *outliers*.

Para a amostra D, não foi feito nenhum ajuste específico nos dados, apenas foi aplicado o critério MAD - Median Absolute Deviation (k=3) para a detecção e remoção de outliers.

**Figura 12 — Resultados comparativos dos modelos preditivos por amostra.**




**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

A Figura 12 resume os resultados. Na amostra A (n = 348.881), composta pelos examinadores mais rápidos, o modelo apenas com as variáveis dos processos apresentou $R^2$ = 0,022, o que significa que nessa amostra essas variáveis praticamente não explicam a variação do tempo da análise. O modelo 2, que recebe a informação sobre o examinador, apresenta $R^2$ = 0,197, um aumento expressivo na capacidade do modelo em prever o tempo do exame.

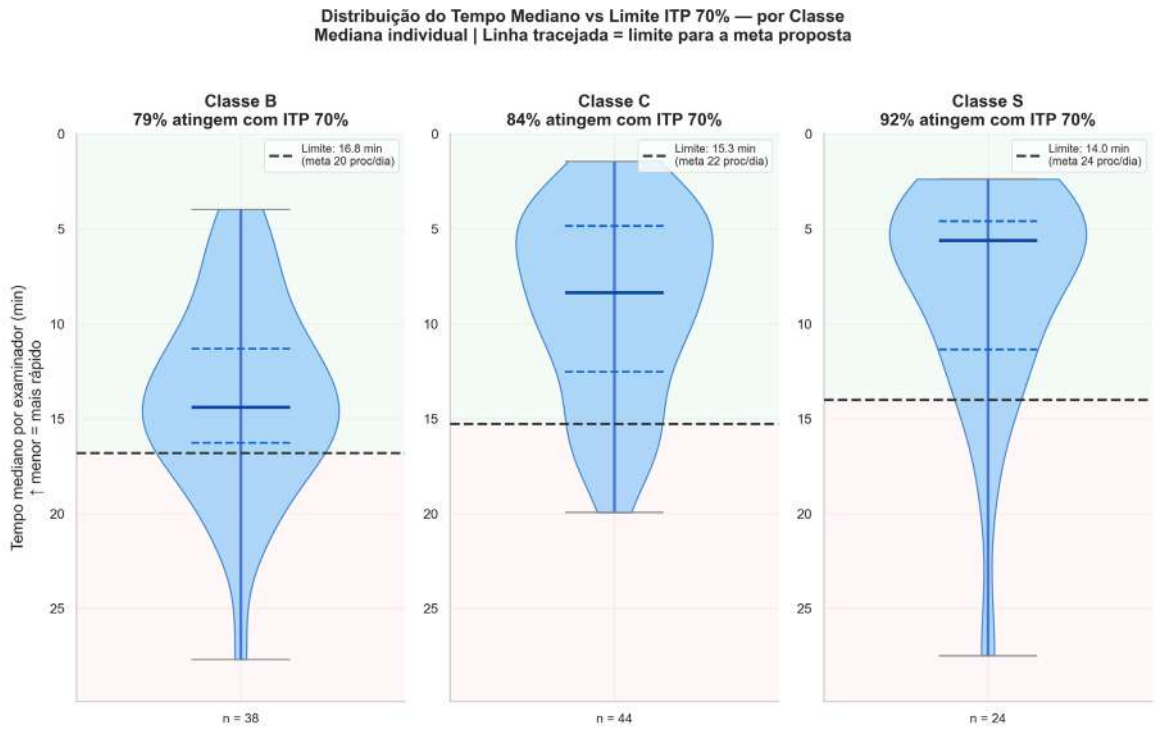
Na amostra B (n = 79.814), composta pelos examinadores com tempos de exame próximos da mediana, o $R^2$ foi de 0,040 no modelo 1, subindo para 0,059 no modelo 2, valor igualmente baixo. Nesse caso, a variação acaba sendo baixa, porque foram selecionados examinadores com padrão de tempo homogêneo. Ao selecionar apenas 15 examinadores com tempos muito parecidos, a variação entre eles quase desaparece. Por isso, adicionar o examinador ao modelo quase não ajuda — não há diferença para o modelo capturar.

No entanto, nas amostras C (n = 845.688) e D (n = 790.544) que contam uma limpeza de valores extremos mais robusta, os resultados do modelo 2 apresentam melhoras significativas, frente ao modelo 1. A amostra C tem um ganho de +0,27, enquanto a amostra D apresenta um ganho de +0,38. Esses resultados, embora possam parecer tímidos para um modelo preditivo, confirmam que o fator humano domina a explicação da variação do tempo de exame.

## 3.7 Estimativa de metas ajustadas

Evidenciada a dominância do fator humano, procedeu-se à simulação exploratória de metas escalonadas: Classe B com 20 pontos, Classe C com 22, Classe S com 24, além de redução de 10% para examinadores acima de 65 anos. Os tempos necessários seriam 16,8, 15,3 e 14 minutos por processo, respectivamente. Esses novos patamares poderiam aumentar o número de decisões adicionais em quase 10 mil por mês, representando um aumento no volume de decisões de cerca de 20% e uma expansão massiva da capacidade operacional sem a necessidade de novas contratações, corroborando a tese de que a gestão estratégica do desempenho é um instrumento efetivo para fundamentar políticas de desenvolvimento pessoal. Conforme discutido por Trigueiro-Fernandes, Hipólito e Añez (2018), a otimização dos resultados em organizações públicas deve focar na plena utilização dos recursos existentes por meio da compreensão das dimensões que afetam a produtividade.

**Figura 13 — Distribuição do TEMPO_AJUSTADO mediano por classe funcional em comparação com as metas propostas.**




**Fonte:** elaboração própria com base em dados administrativos do INPI (2026).

A Figura 13 mostra que a maioria dos examinadores já apresenta tempo mediano compatível com as metas propostas. Foram excluídos dessa análise os examinadores que não atuam mais no exame, ou que deixaram o instituto, examinadores cuja principal atividade não é o exame de pedidos sem oposição, e os novos examinadores recém ingressantes, já que os dados do estudo contam com um número inicial e muito pouco representativo dos seus exames (entre 40 e 90 de cada examinador, que representa menos de 5% da produção média anual esperada). A dispersão revela, contudo, que alguns estariam distantes desses patamares. É fundamental ressalvar que esta simulação é exploratória e que a transição deve ser mediada por diálogo, incorporando dimensões qualitativas e de saúde ocupacional.

## CONCLUSÃO

Este estudo oferece evidência empírica robusta de que a experiência profissional é a variável dominante no tempo de exame de pedidos de registro de marca no INPI. A mediana do tempo reduz-se em aproximadamente 66% entre iniciantes e servidores com mais de 20 anos, com ganho mais expressivo nos primeiros cinco anos. A faixa etária, controlada pelo nível de experiência, exerce influência própria: o acúmulo de inteligência cristalizada compensa possíveis declínios na inteligência fluida até os 65 anos, após o que se observa discreta elevação. A produtividade intradiária apresenta padrão de aceleração, embora acompanhada de alteração na composição decisória que merece investigação qualitativa.

Os achados fundamentam a discussão sobre a inadequação do sistema de metas uniformes vigente, que não reflete a heterogeneidade da capacidade produtiva. A possibilidade de metas progressivas é sustentada pelos dados, mas sua implementação deve considerar dimensões que este estudo não mensurou: qualidade das decisões e impactos sobre a saúde física e mental dos servidores, bem como sobre a qualidade de vida.

Como limitações, reconhece-se que: (a) o TEMPO_AJUSTADO é proxy que pode incluir pausas não identificadas; (b) os fatores de ajuste do PGD não foram validados empiricamente; (c) a ausência de métricas de qualidade impede conclusões sobre a relação entre celeridade e acerto decisório; e (d) os resultados refletem o contexto específico do exame de marcas do INPI.

Para futuras investigações, sugere-se: (i) incorporação de indicadores de qualidade das decisões; (ii) avaliações de impactos na saúde vinculadas ao regime de metas; (iii) identificação do ITP específico para o exame de marcas; (iv) análise comparativa com escritórios internacionais de propriedade industrial; e (v) aplicação de modelos causais para isolar os efeitos de confundidores como tipo de marca e complexidade da classe.

## DECLARAÇÃO DE INTERESSE

O autor é Tecnologista em Propriedade Industrial do INPI e, portanto, diretamente vinculado à atividade e ao contexto institucional objeto deste estudo. A pesquisa foi realizada de forma independente, no âmbito do Trabalho de Conclusão de Curso do MBA em Ciência de Dados e IA Aplicadas da ENAP, e não representa posição institucional do INPI. Os dados administrativos utilizados foram anonimizados e processados localmente, sem inserção em ferramentas externas de IA, em conformidade com a LGPD e as diretrizes da CGMBA/ENAP sobre privacidade e sigilo.

## DISPONIBILIDADE DE DADOS E CÓDIGO

Em atendimento aos requisitos de Ciência Aberta (Open Science), os materiais deste estudo estão disponibilizados publicamente nos seguintes repositórios:

**Dados (Open Data):** Os dados administrativos brutos do INPI contêm informações funcionais de servidores públicos e não podem ser disponibilizados integralmente em conformidade com a LGPD (Lei nº 13.709/2018). Para fins de reprodutibilidade, disponibiliza-se uma versão anonimizada do dataset, na qual identificadores pessoais foram removidos e nomes substituídos por códigos aleatórios. O dataset anonimizado está depositado no Zenodo. DOI: https://doi.org/10.5281/zenodo.20878000.

**Código-fonte (Open Code):** Os três notebooks Python utilizados nas etapas de consolidação, análise exploratória e modelagem preditiva estão depositados no GitHub, acompanhados de arquivo requirements.txt com as dependências e README.md com instruções para reprodução dos resultados. Disponível em: https://github.com/rafaelgrn/metas-produtividade-inpi

**Documentação:** O repositório contém dicionário de variáveis e descrição completa dos procedimentos de anonimização no arquivo data/README_DATA.md.

## REFERÊNCIAS

AUBERT, P.; CRÉPON, B. **Age, wage and productivity: firm-level evidence**. Paris: INSEE, 2006. (Discussion Paper).

BRASIL. **Decreto nº 1.590, de 10 de agosto de 1995**. Dispõe sobre a jornada de trabalho dos servidores da Administração Pública Federal. Diário Oficial da União: Brasília, DF, 11 ago. 1995.

BRASIL. **Decreto nº 11.072, de 17 de maio de 2022**. Dispõe sobre o Programa de Gestão e Desempenho – PGD. Diário Oficial da União: Brasília, DF, 18 maio 2022.

BRASIL. **Instrução Normativa nº 65, de 30 de julho de 2020**. Estabelece orientações sobre o Programa de Gestão. Diário Oficial da União: Brasília, DF, 31 jul. 2020.

DANZIGER, S.; LEVAV, J.; AVNAIM-PESSO, L. Extraneous factors in judicial decisions. **Proceedings of the National Academy of Sciences**, v. 108, n. 17, p. 6889–6892, 2011. DOI: https://doi.org/10.1073/pnas.1018033108

FRAKES, M. D.; WASSERMAN, M. F. **Deadlines Versus Continuous Incentives**: Evidence from the Patent Office. Cambridge: National Bureau of Economic Research, 2024. (NBER Working Paper, n. 32066). https://doi.org/10.3386/w32066.

FRAKES, M. D.; WASSERMAN, M. F. Procrastination in the workplace: evidence from the U.S. Patent Office. **Journal of Public Economics**, v. 183, 104137, 2020. https://doi.org/10.3386/w22987

GUEDES, T. F.; ROCHA-DE-OLIVEIRA, S. Problematizações foucaultianas do teletrabalho: governamentalidade e cuidado de si na experiência de servidores do Judiciário. **Research, Society and Development**, v. 12, n. 2, e28612240268, 2023. https://doi.org/10.33448/rsd-v12i2.40268.

HORN, J. L.; CATTELL, R. B. Age differences in fluid and crystallized intelligence. **Acta Psychologica**, v. 26, p. 107-129, 1967. https://doi.org/10.1016/0001-6918(67)90011-X.

NEWELL, A.; ROSENBLOOM, P. S. Mechanisms of skill acquisition and the law of practice. In: ANDERSON, J. R. (Ed.). **Cognitive skills and their acquisition**. Hillsdale: Lawrence Erlbaum, 1981. p. 1-55.

SERRANO, A. L. M. et al. (Orgs.). **Dimensionamento na administração pública federal**: uma ferramenta do planejamento da força de trabalho. Brasília: Enap, 2018. 113 p.

SERRANO, A. L. M.; MENDES, N. C. F.; MENESES, P. P. M. (Orgs.). **Dimensionamento na administração pública federal**: avanços e resultados alcançados. Brasília: UnB, 2022. 133 p.

TRIGUEIRO-FERNANDES, L.; HIPÓLITO, J. A. M.; AÑEZ, M. E. M. **Dimensões intervenientes na produtividade e no planejamento da força de trabalho**. ReCaPe, São Paulo, v. 8, n. 1, p. 5-18, 2018.

WORLD INTELLECTUAL PROPERTY ORGANIZATION. **World Intellectual Property Indicators** 2025. Geneva: WIPO, 2025. 192 p.

# Sobre as Colaboradoras e Colaboradores, Autoras e Autores MBA em Ciência de Dados e Inteligência Artificial - MBA CDIA

**Iara Alves:** Diretora de Educação Executiva da Enap. Especialista em Políticas Públicas e Gestão Governamental. Autora do texto de agradecimento aos envolvidos neste livro e no MBA CDIA. http://lattes.cnpq.br/9789401887402414

**Jackson De Toni:** Organizador deste livro, autor do texto de apresentação, professor da Disciplina Integradora Aplicada e dos Laboratórios de Casos de Políticas Públicas e orientador de trabalhos de conclusão do MBA CDIA. Analista de Inovação e Produtividade da Agência Brasileira de Desenvolvimento Industrial (ABDI), http://lattes.cnpq.br/9181587534442584

**Natália Massaco Koga:** Autora do prefácio deste livro. Especialista em Políticas Públicas e Gestão Governamental. http://lattes.cnpq.br/6578107008461968.

**Alex Pereira Lopes:** Coautor do texto de Introdução deste livro, professor de Introdução à Programação em Python, Introdução à Ciência de Dados, orientador de trabalhos de conclusão e curador do MBA CDIA. Analista de Planejamento e Orçamento. http://lattes.cnpq.br/2314870317460514

**Hélio Bomfim de Macêdo Filho:** Coautor do texto de Introdução deste livro, professor de IA Generativa, Machine Learning, Laboratórios de Casos de Políticas Públicas, Monitorias em Computação e Matemática e orientador de trabalhos de conclusão do MBA CDIA. Especialista em Políticas Públicas e Gestão Governamental. http://lattes.cnpq.br/8977496078251007

**Rachel Pereira Dorneles:** Coordenadora do MBA CDIA, coorganizadora deste livro, membro da equipe da Coordenação-Geral de Especialização e MBA, da Enap. Analista Técnica do Poder Executivo. http://lattes.cnpq.br/4575557071935634

**Allan Vieira Castro Quadros:** professor de Estatística Inferencial e orientador de trabalhos de conclusão do MBA CDIA. http://lattes.cnpq.br/2241286150446069

**Clayton Rodrigues da Silva:** professor auxiliar de Estatística Descritiva, Introdução à Programação em Python, Machine Learning e orientador de trabalhos de conclusão do MBA CDIA. http://lattes.cnpq.br/2394655663002214

**Thaís Salzer Procopio:** professora auxiliar de Machine Learning e orientadora de trabalhos de conclusão do MBA CDIA. http://lattes.cnpq.br/3512176436618377

**Fausto Barros de Sá Telles:** professor de Aspectos Éticos e Legais do uso da IA na Administração Pública e orientador de trabalhos de conclusão do MBA CDIA. http://lattes.cnpq.br/4055359821244666

## AUTORAS E AUTORES DO LIVRO

**Alexandre Galardinovic Ribeiro:** autor do trabalho em grupo. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Auditor Fiscal da Receita Federal do Brasil. alexandre.ribeiro@rfb.gov.br

**Bruno Gustavo Janoca:** autor do trabalho em grupo aluno do MBA em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Auditor de Finanças e Controle da Controladoria-Geral da União. brunojro@hotmail.com

**Camila Falchetto Romero:** autora do artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Coordenadora-Geral da Secretaria de Governo Digital do Ministério da Gestão e Inovação em Serviços Públicos. camila.falchetto@gestao.gov.br

**Eduardo Melo Moura Filho:** autor do artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Analista de Planejamento e Orçamento do Ministério do Planejamento e Orçamento. dudammf@gmail.com

**Fabiana Ferreira do Nascimento:** autora do artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Analista de Tecnologia da Informação da Universidade Federal da Paraíba - UFPB. fabianan@gmail.com

**Fabio Kouri Paim:** autor do trabalho em grupo e de artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Analista de Comércio Exterior do Ministério do Desenvolvimento, Indústria, Comércio e Serviços. fabiokpaim@gmail.com

**Francisco Carlos Molina Duarte Júnior:** autor do artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Analista de Tecnologia da Informação, atua na Enap. francisco.m.duarte@gmail.com

**Juliana Pierrobon Lopez:** autora do artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Analista Ambiental, atua no Instituto Chico Mendes de Conservação da Biodiversidade. juliana.lopez@icmbio.gov.br

**Luís César Seixas de Oliveira:** autor do trabalho em grupo e de artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Analista de Planejamento, Gestão e Infraestrutura em Informações Geográficas e Estatísticas do Instituto Brasileiro de Geografia e Estatística. luis.s.oliveira@ibge.gov.br.

**Luiz Victor do Espirito Santo:** autor do artigo individual, aluno do MBA em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Técnico Judiciário do Conselho Nacional de Justiça. luizzzvictor@gmail.com

**Marcelo Cabreira Bastos:** autor do artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Analista Técnico de Políticas Sociais, atua no Ministério do Desenvolvimento Agrário e Agricultura Familiar, marceloc.bastos@hotmail.com

**Nelson Oliveira Santos:** autor do artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Engenheiro Civil do Instituto Federal de Ciência e Tecnologia da Bahia. nelson.santos@ifba.edu.br

**Rafael Gerardo Rodrigues do Nascimento:** autor do artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Tecnologista em Propriedade Industrial do Instituto Nacional da Propriedade Industrial. rafael.nascimento@inpi.gov.br

**Rafael Luis Giacomin:** autor do trabalho em grupo e de artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Analista de Planejamento e Orçamento, atua no Ministério da Integração e do Desenvolvimento Regional. rafaellgiacomin@gmail.com

**Rosana Elisa Gonçalves Gonçalves Pinho:** autora do artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Tecnologista Sênior do Ministério da Saúde. rosanagg@gmail.com.

**Tadeu Gomes Teixeira:** autor do trabalho em grupo e de artigo individual. Especialista em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. professor na Universidade Federal do Maranhão. tadeu.teixeira@ufma.br

**Thiago Mello Venâncio:** autor do trabalho em grupo, aluno do MBA em Ciência de Dados e Inteligência Artificial Aplicadas - Enap. Especialista em Indigenismo da Fundação Nacional do Índio. thiagomello@id.uff.br

