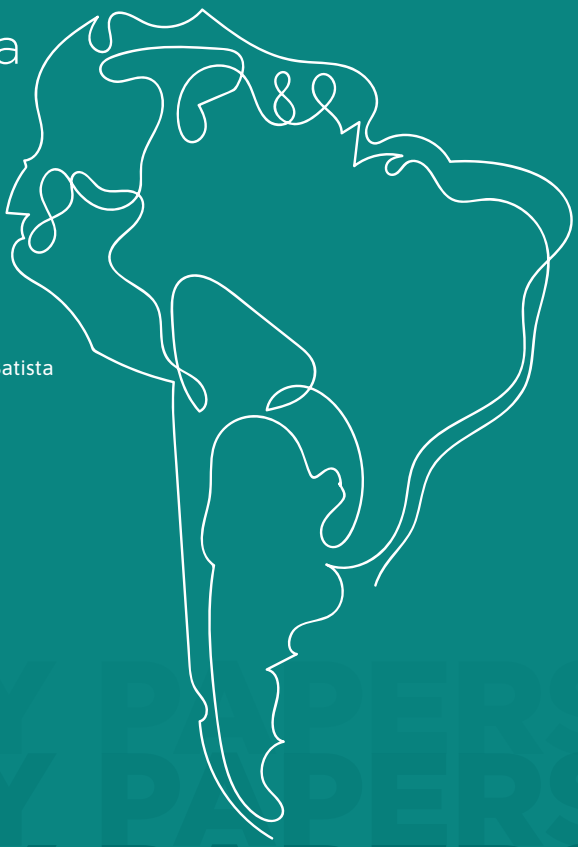
PROGRAMA
CÁTEDRAS
BRASIL

ENAP

Avaliação da Equidade e do Viés Algorítmico na Aplicação de IA na Gestão Pública:

Um Estudo na Saúde Pública Materna



Prof. Dr. André Filipe de Moraes Batista

POLICY PAPERS

Avaliação da Equidade e do Viés Algorítmico na Aplicação de IA na Gestão Pública: Um Estudo na Saúde Pública Materna

Prof. Dr. André Filipe de Moraes Batista

Fundação Escola Nacional de Administração Pública

Presidenta
Betânia Peixoto Lemos

Diretora-Executiva
Natália Teles da Mota

Diretor de Altos Estudos
Alexandre de Ávila Gomide

Diretora de Educação Executiva
Iara Cristina da Silva Alves

Diretor de Desenvolvimento Profissional
Braulio Figueiredo Alves da Silva

Diretora de Inovação
Camila Medeiros

Diretor de Gestão Interna
Lincoln Moreira Jorge Junior

Revisão ortográfica
Adriana Braga
Renata Mourão
Roberto Araújo

Projeto gráfico e editoração eletrônica
Oscar Soler

Ficha catalográfica elaborada pela equipe da Biblioteca Graciliano Ramos da Enap

B333a Batista, André Filipe de Moraes
Avaliação da equidade e do viés algorítmico na aplicação de IA na gestão pública: um estudo na saúde pública maternal / André Filipe de Moraes Batista. -- Brasília: Enap, 2024.
34 p. : il. -- (Policy Paper-7; Coleção: Cátedras 2024)

Inclui bibliografia
ISSN: 0104-7078

1. Inteligência Artificial. 2. Saúde Pública. 3. Políticas Públicas. 4. Administração Pública. I. Título. II. Batista, André Filipe de Moraes

CDD 006.3

Bibliotecária: Kelly Lemos da Silva – CRB1/1880

SUMÁRIO

EDITORIAL – CÁTEDRAS 2023

O **Programa Cátedras Brasil** é um programa de fomento de pesquisas da Escola Nacional de Administração Pública (Enap), coordenado pela Coordenação-Geral de Pesquisa da Diretoria de Altos Estudos (DAE), que busca fomentar o desenvolvimento de pesquisas aplicadas com vistas à construção de novos paradigmas de gestão pública e à reflexão dos desafios de transformação do Estado Brasileiro.

Em 2023, com a chegada de um novo governo e de uma nova gestão na Enap, lançamos um edital do Programa Cátedras Brasil no qual selecionamos projetos em dez áreas temáticas, os quais buscam dar respostas às necessidades de mudanças e de reconstrução de políticas públicas fundamentais para redução de desigualdades da sociedade brasileira e da retomada do papel do Estado para o desenvolvimento sustentável.

A partir desse edital, buscamos trazer duas inovações no âmbito do Programa, quais sejam: i) pesquisas de seis meses capazes de subsidiar com informações e evidências os tomadores de decisão e os gestores públicos envolvidos com as temáticas estudadas; ii) publicação no formato de policy paper, uma publicação mais direta e concisa, a qual tem como principal objetivo o estabelecimento de recomendações para aqueles interessados naquele problema da sociedade.

Ressaltamos que um dos propósitos do Programa Cátedras Brasil consiste em fomentar o elo entre a academia e a comunidade de praticantes (gestores públicos) com vistas à construção coletiva de soluções aos problemas públicos. Essa iniciativa procura absorver contribuições interdisciplinares e inovadoras nos campos de conhecimento correlatos à gestão, à administração e às políticas públicas com o objetivo de agregar valor às atividades de produção e de disseminação de conhecimento aplicado de modo a conceber propostas de soluções para problemas políticos, econômicos, ambientais e sociais.

Em síntese, as pesquisas realizadas no contexto do Edital nº 50/2023 do Programa Cátedras Brasil e apresentadas nesta série de Policy Papers visam ofertar evidências, caminhos e reflexões para se pensar a melhoria e o fortalecimento da administração pública brasileira com vistas à construção de um país mais inclusivo e socialmente justo.

Desejamos a todos e a todas uma ótima leitura!

Alexandre de Ávila Gomide
Diretor de Altos Estudos (DAE)

Rafael Viana
Coordenador-Geral de Pesquisa (CGP).

MINI-BIO DO AUTOR

PROF. DR. ANDRÉ FILIPE DE MORAES BATISTA

André Filipe de Moraes Batista é doutor em Engenharia da Computação pela Escola Politécnica da Universidade de São Paulo. Realizou pós-doutorado na Faculdade de Saúde Pública da Universidade de São Paulo, desenvolvendo modelos preditivos aplicados a políticas públicas de saúde materna e infantil. Pesquisador na área de Ciência de Dados e Inteligência Artificial. É professor do INSPER – Instituto de Ensino e Pesquisa, em São Paulo.

SUMÁRIO EXECUTIVO

A justiça nos sistemas de Inteligência Artificial (IA) é crucial, especialmente nos serviços públicos. Este *policypaper* aborda o desafio da equidade em IA - garantir que modelos de IA tratem todos justamente, sem discriminação de indivíduos ou grupos de indivíduos por raça, gênero, idade ou outros atributos sensíveis. Este desafio é amplificado pelos vieses históricos e culturais presentes nos dados usados para treinar os modelos de IA. Focando na saúde pública materna em São Paulo, Brasil, o estudo buscou determinar as métricas adequadas para avaliar a equidade dos modelos preditivos, essenciais para alocar recursos e melhorar políticas baseadas em dados.

A metodologia envolveu coleta e vinculação de dados de saúde materna, revisão de conceitos de equidade e viés algorítmico, e interações com gestores públicos e sociedade civil. Adaptou-se a ferramenta Aequitas, produzida pela Universidade de Chicago (EUA), para avaliar os modelos preditivos, facilitando a identificação de vieses e disparidades. O estudo enfatiza a importância da auditoria algorítmica e a colaboração entre provedores de IA e gestores públicos para assegurar a responsabilidade ética e social dos modelos de IA.

EVIDÊNCIAS PARA A PRÁTICA

- A incorporação da auditoria algorítmica de modelos de IA desde o início do desenvolvimento de tais modelos, assim como no processo de decisão de incorporação de um modelo a uma política pública, é essencial para garantir que questões de parcialidade e justiça sejam abordadas, assegurando a responsabilidade ética e social;
- Estabelecer uma metodologia para definir métricas de avaliação da equidade dos modelos auxilia na identificação de vieses potenciais em modelos preditivos, permitindo uma análise mais criteriosa das métricas, além das tradicionais métricas de desempenho de modelos preditivos;
- O uso de elementos visuais para identificação de potenciais discriminações em alguns grupos de indivíduos possibilita uma compreensão mais profunda e prática desses conceitos, melhorando a discussão qualificada sobre a adequabilidade dos modelos preditivos em diferentes cenários;
- A aprendizagem contínua a partir das auditorias e o desenvolvimento de práticas para discussão sobre a adequabilidade dos modelos são fundamentais para promover um diálogo eficaz entre as diferentes áreas de especialização, tornando o desenvolvimento de soluções de IA mais humano, ético e responsável.

Destaca-se três lições para profissionais que puderam ser produzidas em decorrência do trabalho, quais sejam:

- Importância da análise de equidade: profissionais devem priorizar a análise de equidade em modelos de IA para evitar decisões discriminatórias, especialmente em contextos de alto impacto social;
- Uso de ferramentas de apoio: ferramentas como Aequitas são essenciais para visualizar e compreender a equidade nos modelos de IA permitindo identificar vieses e avaliar a adequabilidade dos modelos de IA frente aos objetivos de sua aplicação;
- Auditoria algorítmica como prática contínua: a auditoria algorítmica deve ser integrada desde o início do desenvolvimento de modelos de IA assegurando que questões de parcialidade e impacto social sejam continuamente monitoradas e endereçadas.

1 INTRODUÇÃO

A justiça nos sistemas de Inteligência Artificial é essencial em todas as áreas onde essas tecnologias são aplicadas, especialmente nos serviços públicos. A IA tem um grande potencial para tornar a distribuição de recursos mais eficiente e aprimorar as políticas com base em dados. Porém, isso só pode ser alcançado se evitarmos preconceitos nos modelos de IA que possam levar a decisões injustas ou discriminatórias.[1]

Quando falamos em *fairness* ou equidade em IA, nos referimos à capacidade dos modelos de IA de tratar todos de forma justa, sem favorecer ou prejudicar grupos de pessoas por características como raça, gênero ou idade. Isso é um desafio, pois os dados que usamos para "ensinar" esses modelos podem ter preconceitos históricos ou culturais [4].

Na administração pública, é vital usar modelos de IA que promovam igualdade social. Eles podem ajudar a alocar recursos em educação, melhorar a segurança pública através de policiamento preditivo, e planejar cidades de forma mais eficiente. Esses modelos precisam ser transparentes e justos para beneficiar toda a comunidade.

Existem várias ferramentas e métricas que podemos usar para verificar que a IA está sendo justa. Por exemplo, a ferramenta Aequitas, criada pela Universidade de Chicago, ajuda a compreender como os modelos de IA afetam diferentes grupos de pessoas, apontando onde podem estar os preconceitos.

Na área da saúde pública, e mais especificamente na saúde materna, a justiça em IA é ainda mais crítica. Modelos de IA podem prever riscos à saúde, otimizar o uso de recursos em hospitais e ajudar no diagnóstico e tratamento de doenças. É importante que esses modelos sejam justos para que todos tenham acesso igual aos cuidados de saúde.

Este trabalho buscou determinar as métricas mais adequadas para serem incorporadas no processo de avaliação de modelos preditivos, de modo a contribuir com um panorama prático e pragmático sobre a necessidade de avaliação dos modelos que possam ser adotados na gestão pública. A área de concentração delimitada para estudo foi a área de saúde pública, mais especificamente a saúde materna, na avaliação da equidade de modelos preditivos de óbito

[1] Agradecemos à Camila Saeko Kobayashi de Pinho e à Carolina Sayão Lobato Coppetti pela atuação como pareceristas no âmbito desta pesquisa.

materno, construídos com dados provenientes da Secretaria Municipal de Saúde da cidade de São Paulo, no Estado de São Paulo, Brasil.

A metodologia para abordar a equidade no processo de predição de óbito materno envolveu várias etapas inter-relacionadas. Inicialmente, realizou-se a coleta e vinculação de dados de saúde materna, utilizando dados do Sistema de Informações sobre Nascidos Vivos (SINASC), do Sistema de Informações de Mortalidade (SIM) e do Sistema Integrado de Gestão à Assistência à Saúde (SIGA) da Secretaria Municipal de Saúde de São Paulo. Seguiu-se com uma revisão aprofundada dos conceitos de equidade e viés algorítmico em colaboração com gestores públicos da prefeitura de São Paulo e com o apoio de pesquisadores do Laboratório de Big Data e Análise Preditiva em Saúde (LABDAPS) da Faculdade de Saúde Pública da Universidade de São Paulo (FSP-USP), além de outros professores e pesquisadores dessa instituição. Este passo foi crucial para estabelecer uma base teórica sólida para o projeto. Foram promovidas reuniões e workshops com gestores públicos e membros da sociedade civil para sensibilizá-los sobre a importância desses temas na tomada de decisões com o auxílio de modelos de IA. O objetivo dessas interações foi compreender a melhor forma de abordar a temática, as métricas mais relevantes, e o papel de uma ferramenta como a Aequitas na facilitação do processo de avaliação de modelos preditivos.

Realizando essas etapas, endereçou-se o problema principal de equidade algorítmico em saúde pública materna. A integração de dados robustos, a revisão de conceitos críticos, o diálogo com os gestores e com a sociedade, além da adequação da ferramenta Aequitas para as necessidades levantadas, buscaram propiciar instrumentos para que decisões baseadas em IA sejam mais justas e representativas. Este processo não apenas melhora a equidade nas decisões de saúde pública materna, mas também estabelece um modelo replicável para outras áreas da gestão pública, fomentando uma abordagem mais ética e equitativa no uso da IA.

2 DIAGNÓSTICO

2.1. DADOS UTILIZADOS

No âmbito do desenvolvimento deste projeto, foram realizadas várias etapas metodológicas para coletar, processar e analisar dados referentes ao óbito materno, no município de São Paulo, entre os anos de 2012 e 2020. Três sistemas de informação foram utilizados como fonte de dados, fornecidos pela Secretaria Municipal de Saúde da cidade de São Paulo, a saber: o SINASC (Sistema de Informações sobre Nascidos Vivos), o SIM (Sistema de Informações de Mortalidade) e o SIGA (Sistema de Informações Gerenciais e Administrativas da Prefeitura de São Paulo). Do SINASC foram disponibilizados dados detalhados sobre nascidos vivos, incluindo informações como número de declaração de nascido, dados demográficos da mãe, condições do parto, entre outros. Do SIM foram obtidos dados de óbitos de crianças de até 1 ano de idade e gestantes ou parturientes, incluindo tipo de óbito, data do óbito, informações da mãe e causa básica da morte.

No contexto do SIGA, foram obtidos dados de gestantes cadastradas na base entre os anos de 2012 a 2020, com os seguintes dados de atendimento durante a gestação: data de nascimento da mãe, data de atendimento, mês do atendimento, ano do atendimento, data do acolhimento, idade gestacional no momento do atendimento, pressão arterial máxima registrada no momento do atendimento, pressão arterial mínima registrada no momento do atendimento, peso na data do atendimento, estatura na data do atendimento e índice de massa corporal (IMC) na data do atendimento.

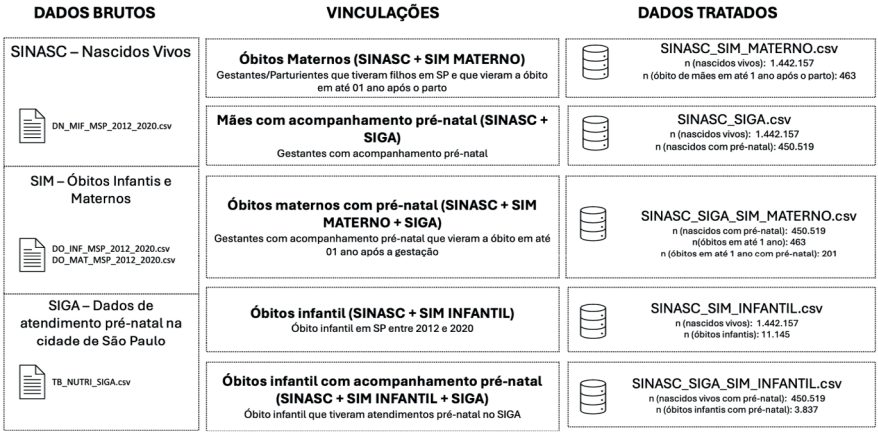
Para cada conjunto de dados foram realizados processos de limpeza e transformação para garantir a precisão e a consistência dos dados. Este processo envolveu a remoção de registros duplicados ou incompletos, a normalização de campos e a conversão de formatos de dados.

Os dados foram vinculados e integrados para permitir uma análise mais abrangente. Esta etapa foi crucial para criar um panorama mais abrangente da saúde materna e no município. Para cada conjunto de dados, foi preparada uma documentação detalhada sobre o processo de limpeza e transformação, juntamente com scripts em linguagem SQL, facilitando a reprodutibilidade dos resultados e análises futuras.

Este processo metodológico robusto permitiu a criação de um banco de dados integrado e detalhado, essencial para investigar as tendências e padrões na saúde materna. O uso desses dados promoverá uma melhor compreensão dos desafios enfrentados no município de São Paulo e apoiará o desenvolvimento de políticas públicas mais eficazes e direcionadas.

A Figura 1 apresenta uma visão geral do processo de coleta de dados brutos, suas vinculações e os arquivos tratados disponíveis. Como resultado desse processo, cinco conjuntos de dados estão disponíveis. Para cada conjunto de dados disponível, há também uma documentação sobre o processo de limpeza e transformação dos dados e scripts em linguagem SQL para reprodutibilidade dos resultados. Foram gerados mais conjuntos de dados do que os analisados nessa pesquisa de modo a possibilitar futuras extensões de escopo e novos modelos preditivos.

Figura 1. visão geral do processo de captura, transformação e integração dos dados brutos para gerar os dados tratados a serem utilizados neste projeto de pesquisa.



2.2. ANÁLISE EXPLORATÓRIA DOS DADOS

A seguir tem-se a apresentação da análise exploratória dos dados. Inicialmente será apresentada a análise dos dados oriundos da vinculação SINASC e SIM MATERNO, onde 463 mães vieram a óbito em até 1 ano após o parto. Ao longo da análise, serão feitas algumas considerações sobre o subconjunto de óbitos maternos de mulheres que fizeram pré-natal e que possuem registros no SIGA (n=201), resultante da vinculação SINASC – SIM – SIGA.

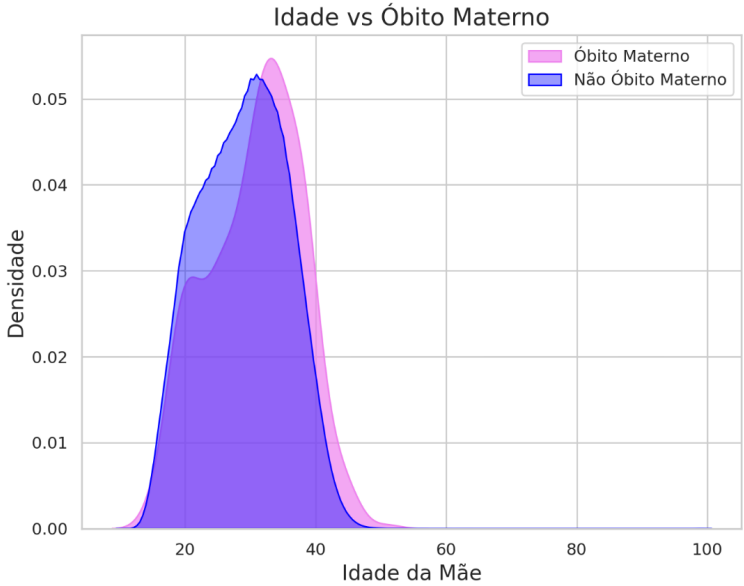
Foram consideradas somente as causas de óbito com código CID (tabela CID10) iniciado por O ou os que se encontram entre B20 e B24 (Referem-se a doenças infecciosas e parasitárias relacionadas ao HIV) ou as causas D39.2(Neoplasia de comportamento incerto ou desconhecido do útero, parte não especificada), E23.0 (Hipopituitarismo), A34 (Sífilis obstétrica),

F53 (transtornos mentais e comportamentais associados ao puerpério, não classificados em outra parte), M83.0 (Osteomalácia do adulto devido à má-absorção). Estes códigos correspondem a condições diretamente associadas a complicações da gravidez, parto e puerpério, bem como a algumas doenças infecciosas e parasitárias, endócrinas, nutricionais e metabólicas, e transtornos mentais e comportamentais que podem estar relacionados ao período gestacional e pós-parto. Esta seleção focou em capturar os eventos mais relevantes e diretamente associados ao óbito materno, permitindo que o modelo seja mais preciso na identificação de fatores de risco específicos para essa população.

As análises foram conduzidas de modo a subsidiar a determinação dos grupos que serão usados para avaliar a equidade do modelo preditivo de risco de óbito materno em até 1 ano, que será apresentado nas seções a seguir.

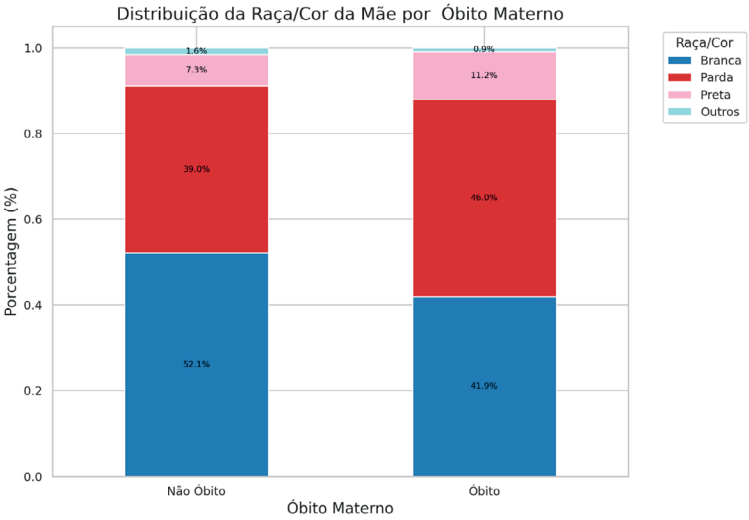
Analisando-se a variável idade da mãe no parto, e comparando a distribuição das idades das mães entre aquelas que vieram a óbito e as que não vieram a óbito, conforme a Figura 02, observa-se que a distribuição das idades das mães que não vieram a óbito apresenta uma média de idade de aproximadamente 28 anos e uma variação desde adolescentes até mulheres na meia-idade. Por outro lado, as mães que sofreram óbito materno têm uma média de idade superior, cerca de 30 anos, e uma distribuição que sugere uma maior prevalência de óbito materno em idades mais avançadas, com a mediana centrada em torno dos 32 anos (nas que não vieram ao óbito o valor da mediana é de 29 anos). As estatísticas descritivas apoiam a visualização de que, embora os óbitos maternos possam ocorrer em qualquer idade, há uma tendência de maior risco associado com idades mais avançadas. A sobreposição das duas distribuições no gráfico também indica que há uma faixa de idade onde o risco de óbito materno é mais pronunciado, particularmente a partir dos 23 anos, aumentando até os 35 anos. Isso reforça a importância de uma atenção especializada e monitoramento para grupos de maior risco etário.

Figura 2. Distribuição da idade das mães no conjunto de dados resultante da vinculação SINASC – SIM MATERNO.



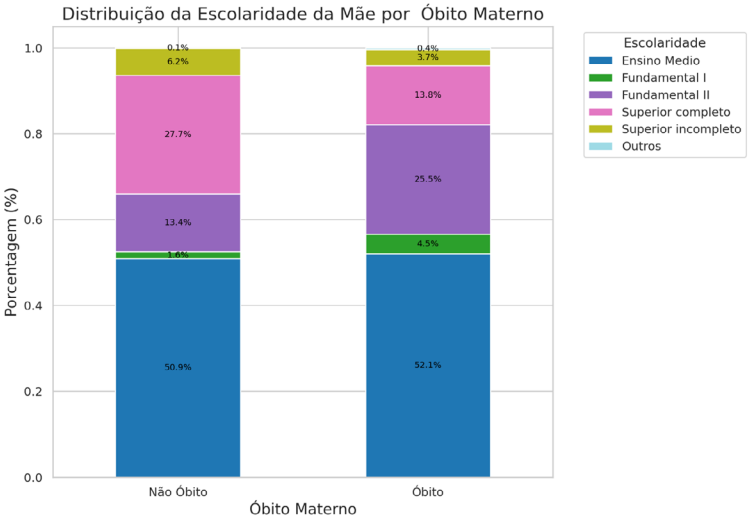
A Figura 03 apresenta a distribuição da raça/cor das mães por óbito materno, comparando os casos em que houve óbito aos que não houve. Observa-se que, para as mães que não vieram a óbito, 52,1% foram identificadas como brancas e 39% como pardas, enquanto para as mães que vieram a óbito, essas proporções mudam para 41,9% e 46%, respectivamente. Isso sugere uma representação maior de mães pardas entre os óbitos maternos em comparação com as mães brancas. Além disso, a categoria "Preta" representa 7.3% entre as que não vieram a óbito e cresce para 11,2% entre as que vieram a óbito, indicando uma maior proporção de óbitos maternos entre mães pretas em relação à sua presença na população de mães não falecidas. A categoria "Outros" mantém uma proporção pequena em ambos os grupos.

Figura 3. Distribuição de Raça/Cor da Mãe (conforme grupos codificados no SINASC), no conjunto de dados decorrente da integração SINASC – SIM MATERNO.



Estes dados podem refletir desigualdades no acesso a cuidados de saúde de qualidade ou na ocorrência de condições socioeconômicas desfavoráveis que afetam desproporcionalmente as mães pardas e pretas.

Figura 4. Distribuição de escolaridade da mãe (conforme grupos codificados no SINASC), no conjunto de dados decorrente da integração SINASC – SIM MATERNO.



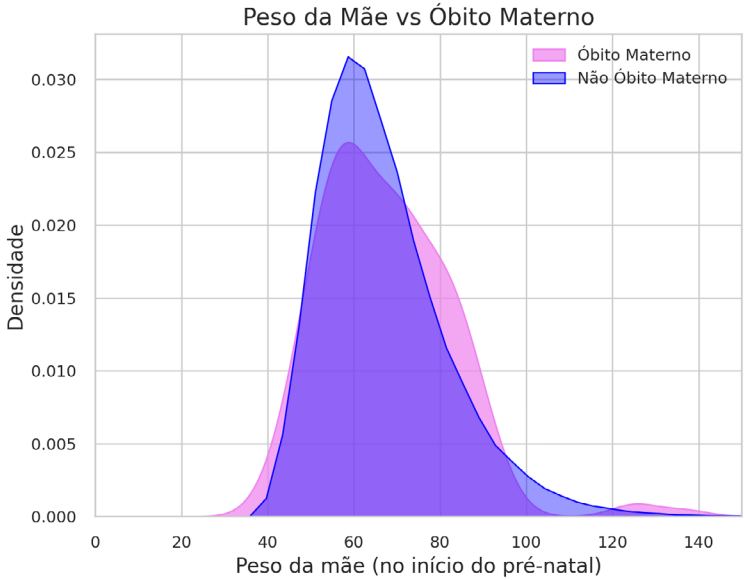
A Figura 04 fornece uma comparação entre os níveis de escolaridade de mães que sofreram óbito materno em contraste com aquelas que não vieram a óbito. Entre as mães que não sofreram óbito materno,50,9 %, concluiu o ensino médio. Observa-se também que 13,4% delas possuem formação no ensino fundamental II. Já entre as mães que foram a óbito, nota-se uma proporção de 52,1% com ensino médio completo, e um aumento significativo, para 25,5, com ensino fundamental II completo.

Interessantemente, a proporção de mães com ensino superior completo é maior entre aquelas que não vieram a óbito, com 27,7% em comparação com 13,8% entre aquelas que sofreram óbito materno. Este dado pode sugerir que um nível de educação mais elevado pode proporcionar um melhor acesso aos cuidados de saúde, juntamente com outros fatores de risco que podem estar influenciando as taxas de mortalidade materna.

O aumento de mães com ensino fundamental II entre os casos de óbito materno pode refletir desigualdades no acesso a cuidados pré e pós-natais, bem como diferenças nas condições socioeconômicas que impactam a saúde materna. Estes achados sublinham a importância de abordagens de saúde pública que priorizem a educação como um meio de melhorar os resultados de saúde materna e reduzir as taxas de óbito.

Com relação ao subconjunto de mães que tiveram assistência pré-natal na cidade de São Paulo/SP, e que tiveram seus atendimentos registrados no SIGA, buscou-se observar se o peso no início da gestação influenciou a taxa de óbitos maternos. O gráfico da Figura 05 compara o peso das mães no início do pré-natal e o desfecho em relação ao óbito materno. A curva para mães que não sofreram óbito materno apresenta um pico pronunciado e é mais estreita, indicando uma concentração de pesos em uma faixa mais específica, enquanto a curva para óbitos maternos é ligeiramente mais alargada e deslocada para a direita, sugerindo uma maior variação no peso das mães e uma média de peso ligeiramente mais alta, fenômeno que corrobora ao encontrado na literatura [5].

Figura 5. Distribuição do peso da mãe no início do pré-natal, obtido das mães que tiveram atendimentos registrados no SIGA.



Este perfil pode indicar uma tendência de maior risco de óbito materno associado a um peso mais elevado no início da gravidez, o que pode estar relacionado a complicações de saúde vinculadas à obesidade. Contudo, é importante considerar que este é apenas um dos fatores de risco, e a causalidade não pode ser estabelecida somente com esta análise. Intervenções voltadas para a promoção de um peso saudável antes e durante a gravidez podem ser benéficas, mas devem ser acompanhadas de outras medidas de saúde pública que considerem a complexidade e a multifatorialidade do óbito materno.

2.3. TERMINOLOGIA BÁSICA DE MÉTRICAS DE EQUIDADE

No campo da IA, quando falamos sobre equidade, nos referimos à justiça e imparcialidade dos sistemas de IA que afetam a vida das pessoas. É crucial para os gestores públicos entender como assegurar que a IA não crie, perpetue ou intensifique discriminações.

Primeiramente, é necessário identificar os chamados "atributos sensíveis", que são características pessoais como sexo, idade ou raça, que, por razões éticas ou legais, merecem atenção especial para evitar discriminação.

Em seguida, temos os "atributos proxy", que são dados que podem não parecer sensíveis à primeira vista, mas que podem indiretamente sugerir informações sobre um atributo sensível. Por exemplo, o código postal pode indicar a etnia de uma pessoa, e, por isso, devemos tratá-lo com cautela.

O conceito de "paridade" se refere à igualdade de tratamento dos grupos pela IA. Quando um sistema é paritário, em teoria, ele não favorece nem prejudica um grupo em relação a outro. Contudo, a paridade por si só não garante a justiça completa, pois não considera mudanças demográficas ao longo do tempo.

Por fim, para medir equidade, usamos ferramentas como a "matriz de confusão", uma tabela que nos ajuda a entender os acertos e erros do modelo de IA. Ela é fundamental para avaliar se a IA está tratando todos os grupos de maneira justa, especialmente, em situações onde há decisões que categorizam as pessoas em diferentes classes, como "alto risco" ou "baixo risco".

2.4. A FERRAMENTA AEQUITAS E O CONCEITO DE GRUPO DE REFERÊNCIA

Em um mundo onde decisões importantes são cada vez mais tomadas por sistemas de inteligência artificial, garantir que esses sistemas sejam justos e imparciais é essencial. A ferramenta Aequitas, criada pela Universidade de Chicago, é um passo nessa direção. A Aequitas simplifica a tarefa de verificar se um modelo de IA trata todos de forma equitativa, sem discriminar com base em características como raça, gênero ou idade.

A Aequitas ajuda a identificar se um modelo de IA tem um viés que pode afetar negativamente certos grupos sociais. Ela faz isso ao dividir os dados de acordo com esses grupos e analisar como o modelo os afeta. Por exemplo, poderia revelar se um modelo prevê incorretamente que mulheres com idade superior em relação à média das idades das gestantes da base de dados têm riscos de saúde mais elevados do que realmente têm.

Para começar a usar a Aequitas, busca-se definir grupos relevantes com base em características como idade, cor da pele, escolaridade etc. Depois, seleciona-se um grupo de referência para comparar como o modelo trata os diferentes grupos. A ferramenta então analisa as previsões do modelo e apresenta gráficos e estatísticas que mostram claramente onde as injustiças podem estar acontecendo.

Vamos considerar um exemplo prático: se estivéssemos usando a IA para prever riscos de saúde materna, poderíamos descobrir que o modelo está superestimando o risco para mulheres acima de 35 anos, quando comparado a um grupo de referência, por exemplo, formado por mulheres com idade inferior a 35 anos. A Aequitas nos ajudaria a ver isso claramente, através de suas métricas e análises gráficas.

2.5. DETERMINAÇÃO DA MÉTRICA MAIS ADEQUADA

Quando usamos a IA para ajudar na tomada de decisões, é fundamental que ela seja justa e não crie, perpetue ou amplifique desigualdades entre diferentes grupos de pessoas. Existem duas maneiras principais de olhar para a justiça em IA: uma se preocupa com a representação equitativa de todos os grupos, e a outra se concentra em garantir que os erros cometidos pelo modelo não afetem desproporcionalmente qualquer grupo [7].

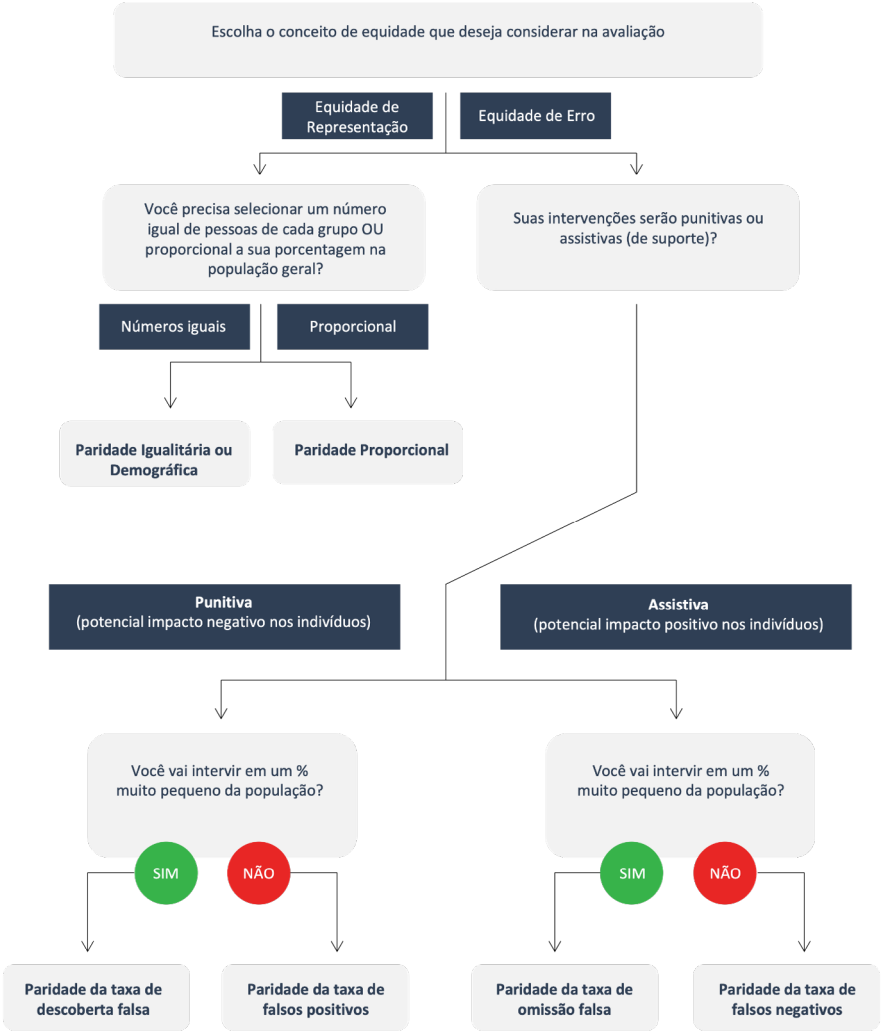
1. **Equidade de representação**: Aqui, o objetivo é assegurar que todos os grupos tenham a mesma chance de serem representados nas decisões do modelo. Por exemplo, se um grupo constitui 50% da população, ele deve receber cerca de 50% das decisões positivas ou negativas. Isso ajuda a garantir que ninguém seja negligenciado ou excessivamente visado pelo modelo.
2. **Equidade de erro**: Esta abordagem se concentra em tornar os erros do modelo equitativos. Se um modelo faz predições erradas, queremos garantir que esses erros não afetem mais um grupo do que outro. Isso é importante, porque erros de predições podem ter consequências sérias para as pessoas envolvidas.

Dependendo do tipo de decisão que o modelo está ajudando a tomar, podemos querer medir a equidade de diferentes maneiras:

- **Para decisões equilibradas**: Se queremos que os grupos sejam tratados igualmente, usamos o que chamamos de "paridade demográfica". Isso significa que se um modelo prevê que 20% de um grupo está em risco, ele deve prever o mesmo para todos os outros grupos, independentemente de suas características.
- **Para decisões proporcionais:** Às vezes, queremos que as decisões reflitam as proporções dos grupos na população. Isso é conhecido como "paridade proporcional" e garante que as taxas de decisões positivas ou negativas correspondam à distribuição de cada grupo na população total.
- **Para decisões punitivas**: Se as decisões do modelo podem levar a penalidades, é importante que erros de previsão positiva, onde pessoas são falsamente identificadas como casos positivos, sejam minimizados e equitativos entre todos os grupos. Isso é o que chamamos de "paridade na taxa de descoberta falsa".
- **Para decisões amplas:** Quando as decisões afetam muitas pessoas e os erros de previsão podem sobrecarregar o sistema com ações desnecessárias, usamos a "paridade na taxa de falsos positivos" para garantir que os grupos sejam tratados de forma equitativa.
- **Para decisões de assistência limitadas**: Quando o modelo intervém raramente e as ações são de assistência, queremos evitar perder aqueles que realmente precisam de ajuda. Aqui, aplicamos a "paridade da taxa de omissão falsa".
- **Para decisões de assistência mais amplas**: Se muitas pessoas precisam de assistência, é crucial que o modelo não ignore aqueles em necessidade. Nesse caso, a "paridade da taxa de falsos negativos" ajuda a assegurar que a ajuda chegue de forma equitativa a todos os grupos.

A Figura 06 apresenta o racional de escolha da melhor métrica com base nesses conceitos. Escolher a métrica certa é crucial para garantir que os sistemas de IA operem de forma justa. Isso não apenas melhora a confiança nas decisões tomadas com o auxílio da IA, mas também assegura que todos sejam tratados de maneira equitativa, sem discriminação ou desigualdade.

Figura 6. Árvore de Decisão para determinar a melhor métrica de avaliação de equidade em modelos preditivos, adaptada de [7].



A Tabela 01 sintetiza as definições das principais métricas de avaliação de equidade, além de apresentar o cálculo que é feito para obtê-las.

Tabela 1. Resumo das principais métricas de avaliação de equidade

Taxa de Descoberta Falsa (False Discovery Rate - FDR)	
Cálculo	**Definição**
$\frac{Falsos\ Positivos}{Verdadeiros\ Positivos + Falsos\ Positivos}$	A Taxa de Descoberta Falsa é a proporção de casos identificados pelo modelo como positivos que são, na verdade, negativos. Ou seja, entre todos os casos que o modelo prevê como óbito materno, a FDR indica quantos destes não resultaram em óbito. A paridade em FDR compara a FDR entre quaisquer grupos definidos para a análise. Por exemplo, considere que a FDR para um modelo preditivo de óbito materno para um grupo de referência seja de 5% e para outro grupo (grupo de interesse) seja de 10%. Isso resulta em uma paridade de 2. Neste contexto, o modelo tem uma probabilidade dobrada de identificar incorretamente não óbitos como óbitos maternos no grupo de interesse em comparação com o grupo de referência. Portanto, quando o modelo prevê que uma mulher do grupo de interesse está em risco de óbito materno, há uma chance duas vezes maior de que essa previsão seja incorreta em comparação com as previsões feitas para mulheres do grupo de referência.
Taxa de Falsos Positivos (False Positive Rate - FPR)	
Cálculo	**Definição**
$\frac{Falsos\ Postivos}{Falsos\ Positivos + Verdadeiros\ Negativos}$	A Taxa de Falsos Positivos (FPR) é a proporção de casos negativos que são incorretamente identificados como positivos pelo modelo, em relação ao total de casos verdadeiramente negativos. A paridade na FPR compara a FPR entre diferentes grupos para garantir equidade na identificação de casos que não necessitam de atenção específica. Por exemplo, em um modelo preditivo de óbito materno, se a FPR para um grupo de referência é de 10% e para um grupo de interesse é de 30%, a paridade seria de 3. Isso indica que o modelo tem uma probabilidade três vezes maior de identificar incorretamente mulheres não em risco de óbito materno como estando em risco no grupo de interesse em comparação com o grupo de referência. Ou seja, quando o modelo prevê que uma mulher do grupo de interesse está em risco de óbito materno, há uma chance 200% maior de essa previsão ser incorreta em relação às previsões feitas para mulheres do grupo de referência. Esta métrica é crucial em contextos como a predição de óbito materno, onde um alto índice de falsos positivos pode levar a intervenções desnecessárias ou estresse adicional para as mulheres incorretamente identificadas como em risco. A paridade da FPR ajuda a assegurar que o modelo não sobrecarregue indevidamente um grupo específico com falsos alarmes, mantendo a equidade no processo de tomada de decisão.
Taxa de Omissão Falsa (False Omission Rate – FOR)	
Cálculo	**Definição**
$\frac{Falsos\ Negativos}{Falsos\ Negativos + Verdadeiros\ Negativos}$	A Taxa de Omissão Falsa é a proporção de casos negativos previstos pelo modelo que são, na verdade, positivos, em relação ao total de casos previstos como negativos. A paridade compara a FOR entre diferentes grupos para assegurar equidade. Por exemplo, em um modelo preditivo de óbito materno, se a FOR para um grupo de referência é de 5% e para um grupo de interesse é de 10%, a paridade é de 2. Este valor sugere que o modelo tem uma probabilidade duas vezes maior de não identificar corretamente mulheres em risco de óbito materno no grupo de interesse em comparação com o grupo de referência. Isto é, quando o modelo prevê que uma mulher do grupo de interesse não está em risco de óbito materno, há uma chance 100% maior de essa previsão ser incorreta em relação às previsões feitas para mulheres do grupo de referência. A métrica de paridade na FOR é crucial em contextos em que é vital fornecer intervenção adequada, pois garante que todos os grupos tenham chances iguais de serem corretamente identificados como em risco e, portanto, recebam a atenção necessária.

Taxa de Falsos Negativos (False Negative Rate – FNR)	
Cálculo	**Definição**
$\frac{Falsos\ Negativos}{Falsos\ Negativos + Verdadeiros\ Positivos}$	A Taxa de Falsos Negativos é a proporção de casos positivos que são incorretamente identificados como negativos pelo modelo, em relação ao total de casos verdadeiramente positivos. A paridade na FNR compara a FNR entre diferentes grupos para assegurar a equidade na identificação de casos que necessitam de atenção. Por exemplo, em um modelo preditivo de óbito materno, se a FNR para um grupo de referência é de 20% e para um grupo de interesse é de 40%, a paridade seria de 2. Isso indica que o modelo tem uma probabilidade duas vezes maior de falhar ao identificar corretamente mulheres em risco de óbito materno no grupo de interesse em comparação com o grupo de referência. Ou seja, quando o modelo prevê que uma mulher do grupo de interesse não está em risco de óbito materno, há uma chance 100% maior de essa previsão ser incorreta em relação às previsões feitas para mulheres do grupo de referência. Esta métrica é extremamente importante em contextos de alto risco, como na predição de óbito materno, onde falhar em identificar corretamente os casos que realmente necessitam de intervenção pode ter consequências graves. A paridade da FNR ajuda a garantir que as mulheres em risco recebam a atenção necessária de maneira justa e equilibrada entre os grupos.

2.6. ADAPTAÇÃO DA FERRAMENTA AEQUITAS

Ao longo das conversas com gestores públicos, pesquisadores e membros da sociedade civil, buscou-se compreender como deveriam ser apresentados os resultados das métricas de equidade e de que maneira a ferramenta Aequitas poderia ser customizada para uma comunicação mais efetiva sobre os resultados.

A ferramenta Aequitas está disponível em código-fonte aberto no GitHub[2], e pode ser importada em programas Python como uma biblioteca. O processo de customização buscou gerar uma página web que invoca as funcionalidades da biblioteca Aequitas e apresenta ao usuário as métricas calculadas, assim como alguns gráficos de apoio. É importante destacar que a avaliação da equidade requer um processo de discussão contextualizada após a análise das métricas, para verificar a pertinência e as ações que devem ser feitas seja para adotar um modelo de inteligência artificial quanto para mitigar os efeitos discriminatórios neste existentes.

Para a adequação da ferramenta Aequitas, buscou-se desenvolver uma aplicação web que faça uso da biblioteca Streamlit, da linguagem Python. No repositório do projeto se encontra o código-fonte dessa adaptação e instruções de como executar a aplicação. A Figura 07 apresenta a interface web construída, que será mais bem explorada na seção que apresenta o modelo preditivo construído e a sua avaliação com o apoio dessa ferramenta.

[2] Disponível em https://github.com/dssg/aequitas, acesso em 18 de jan. 2024.

Figura 7. Interface Web para avaliação de equidade em modelos de IA com o apoio da biblioteca Aequitas, da linguagem Python. Na parte esquerda encontram-se parâmetros de configuração da análise, enquanto na parte direita são apresentados os resultados obtidos.

2.7. MODELO PREDITIVO PARA ÓBITO MATERNO

O objetivo desta etapa foi criar dois modelos preditivos para prever o risco de óbito materno dentro de um ano após o parto. Utilizou-se a vinculação de duas bases de dados distintas: SINASC – SIM MATERNO e SINASC – SIM MATERNO – SIGA. O primeiro modelo, usando apenas dados do SINASC – SIM MATERNO, visava um modelo aplicável em âmbito nacional. O segundo, incluindo dados do SIGA, incorporou o peso no início da gravidez e o ganho de peso ao longo da gravidez para avaliar o impacto dessa variável na equidade do modelo. Este segundo modelo focou em óbitos de mães com pré-natal registrado no sistema municipal (SIGA), explorando o efeito da adição de novos atributos na equidade do modelo.

O processo de pré-processamento e construção do modelo foi conduzido utilizando a linguagem de programação Python, com a utilização da biblioteca PyCaret. Esta biblioteca é uma ferramenta de automação de *machinelearning*, que facilita a construção de modelos preditivos. PyCaret funciona como um agregador para diversas outras bibliotecas e *frameworks* de *machinelearning*, como scikit-learn, XGBoost, LightGBM, entre outros. Sua capacidade de automatizar grande parte do fluxo de trabalho de *machinelearning*, incluindo o pré-processamento de dados, seleção e construção de modelos, otimização de hiperparâmetros e análise de modelos, foi a característica mais considerada para a adoção como ferramenta de construção do modelo.

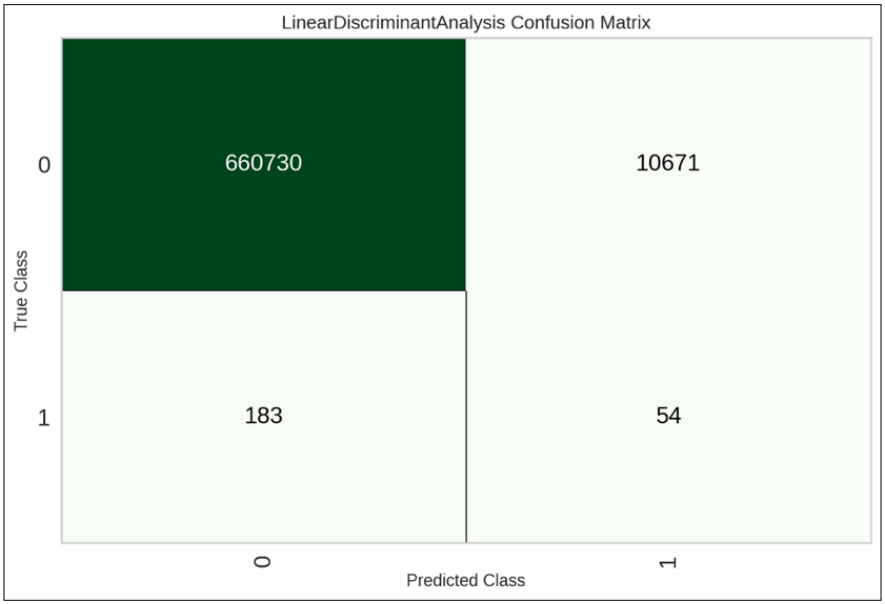
Na fase de construção do modelo, selecionou-se dados de nascimentos ocorridos até o ano de 2016 para o conjunto de treinamento, e os dados a partir de 2016 para o conjunto de testes. A proporção de óbitos maternos ficou razoavelmente balanceada entre os dois conjuntos, com 226 óbitos no treinamento e 237 nos testes. Considerando a raridade do evento de óbito materno, o desbalanceamento dos dados de treinamento representou um desafio significativo, pois poderia levar a um viés de aprendizado no qual o modelo favoreceria a classe majoritária. Para mitigar esse risco, adotou-se a técnica de *undersampling* na classe majoritária (mulheres que não tiveram óbito), ajustando a distribuição das classes no conjunto de treinamento, de modo a propiciar uma melhor capacidade do modelo de captar características relevantes da classe minoritária.

Foram testados diversos algoritmos de *machinelearning* para encontrar o mais adequado para este cenário específico. Entre os algoritmos avaliados estavam: *light gradiente boostingmachine*, *Random forest*, *logisticregression*, *linear discriminantanalysis*, *decisiontree*, *support vector machines*, entre outros. A métrica escolhida para avaliação do modelo foi o F1-score, por representar um equilíbrio entre sensibilidade e especificidade[3]. O algoritmo de análise discriminante linear foi o escolhido, devido à sua superioridade no desempenho, evidenciada por um F1-score de 0.2706 e uma ROC AUC de 0.7644 no conjunto de testes. O pipeline completo de construção do modelo, seu código-fonte e documentação se encontra disponível no repositório do projeto. É importante ressaltar que o foco principal deste projeto não é construir o modelo mais eficiente para a predição de óbito materno, mas avaliar a equidade nos modelos preditivos e investigar as melhores formas de comunicar seus resultados e implicações.

[3] Em um modelo de óbito materno, sensibilidade é a capacidade de detectar corretamente os óbitos reais, enquanto especificidade é a habilidade de identificar corretamente os casos que não são óbitos. O F1-score combina ambas, fornecendo uma visão equilibrada da eficiência do modelo.

A matriz de confusão do modelo, no conjunto de testes, é apresentada na Figura 08. É a partir da matriz que todas as métricas de equidade são calculadas.

Figura 8. matriz de confusão



Foi realizada a avaliação das variáveis mais importantes para a predição e constatou-se que o modelo fez uso de APGAR de 5 minutos, escolaridade da mãe e anomalia congênita. Esse resultado é coerente aos já encontrados em outras pesquisas, e mais detalhes sobre o papel desses atributos no óbito materno podem ser encontrados no trabalho de Batista *et al.* (2021) [6].

Para a análise da equidade do modelo desenvolvido, estabeleceram-se grupos de interesse em idade, escolaridade e raça/cor. Esses grupos foram definidos com base na análise exploratória dos dados resultantes da vinculação dos dados primários do projeto, e seguindo as recomendações de especialistas em saúde pública e gestores públicos que apoiaram o projeto. Os grupos foram categorizados da seguinte forma:

- Idade da mãe: ≤ 27 anos; > 27 anos.
- Escolaridade da mãe: até fundamental II; acima de fundamental II.
- Raça/cor da mãe: brancas; não brancas.

O grupo de referência selecionado é composto por mulheres brancas, com idade até 27 anos e escolaridade acima do fundamental II. Um vídeo no repositório do projeto explica como esses grupos foram configurados na ferramenta web desenvolvida.

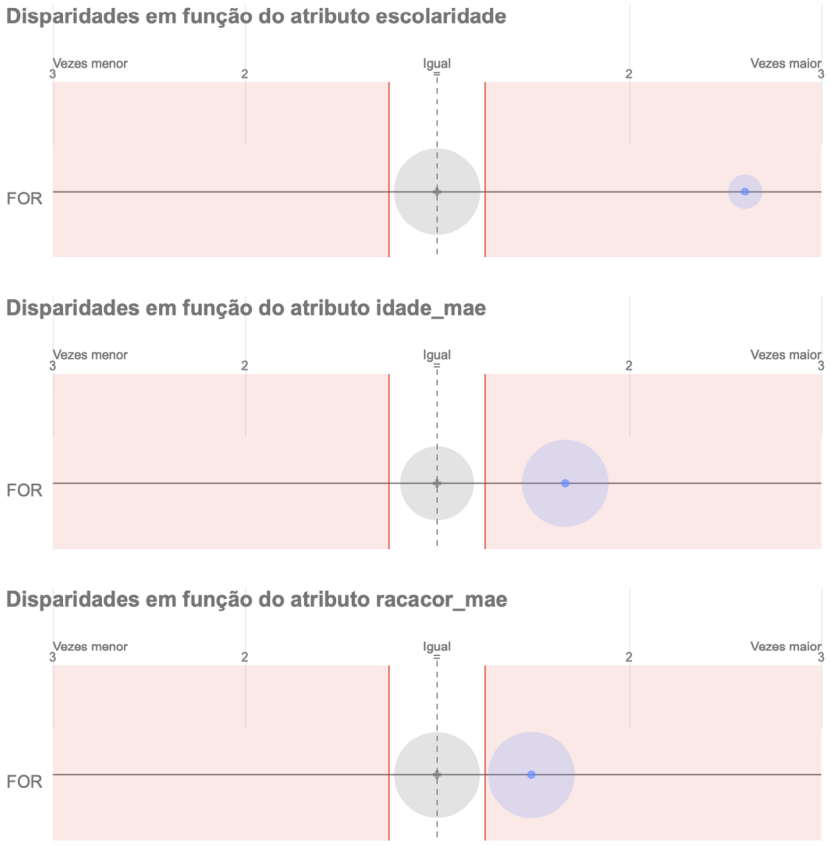
A avaliação da equidade seguiu a orientação da árvore de decisão da Figura 06, optando-se pelo conceito de equidade de erro, considerando uma intervenção assistiva que afeta uma pequena porção da população. Consequentemente, a taxa de omissão falsa (FOR) foi escolhida como a métrica principal. A FOR mede a frequência de erros do modelo ao prever resultados negativos, com uma alta FOR em um grupo indicando que o modelo está omitindo muitos casos positivos (óbitos) neste grupo. A ferramenta web apresenta as principais métricas e seus resultados calculados, conforme Figura 09.

Figura 9. métricas calculadas.

Métricas calculadas:

	attribute_name	attribute_value	Statistical Parity	Impact Parity	fdr	fdr_disparity	fpr	fpr_disparity	for	for_disparity	fnr	fnr_disparity
0	idade_mae	maior_27a	☐	☑	0.9834	0.9929	0.0018	1.237	0.0004	1.6672	0.9277	0.9831
1	idade_mae	menor_igual_27a	☑	☑	0.9905	1	0.0015	1	0.0002	1	0.9437	1
2	escolaridade	Acima de fundamental II	☑	☑	0.9841	1	0.0015	1	0.0003	1	0.9181	1
3	escolaridade	Até fundamental II	☐	☐	0.9925	1.0086	0.0029	1.9617	0.0007	2.6031	0.9697	1.0562
4	racacor_mae	Branca	☑	☑	0.9788	1	0.0014	1	0.0003	1	0.898	1
5	racacor_mae	Não branca	☐	☐	0.9911	1.0126	0.002	1.4329	0.0004	1.4908	0.9568	1.0656

Figura 10. Visualização da disparidade em função da métrica FOR



A Figura 10 apresenta a representação visual de como é a equidade de modelo preditivo em relação a diferentes subgrupos populacionais. O objetivo é identificar e quantificar vieses e injustiças que podem estar presentes nas previsões. Central a este gráfico está a linha que marca o valor "Igual", indicando o ponto de referência onde a taxa de omissão falsa (FOR) é considerada equitativa entre os subgrupos. O intervalo ao redor dessa linha central representa uma zona de paridade aceitável, dentro da qual as disparidades são vistas como não significativas. Este intervalo é muitas vezes estabelecido como uma porcentagem (por exemplo, ±20%) do valor de referência, refletindo a variação que é considerada aceitável por razões práticas e estatísticas.

A partir dos resultados da Figura 09 e com o apoio dos resultados visuais da Figura 10, é possível observar que os valores de FOR variam de 0.0002 a 0.0007, o que indica que o modelo tem uma baixa taxa de omissão de casos positivos em todos os grupos avaliados. No entanto, quando olhamos para a *FOR Disparity*, que compara a FOR entre diferentes grupos com um grupo de referência, observamos algumas discrepâncias. Um valor de 1 indicaria que não há disparidade em relação ao grupo de referência, enquanto valores acima de 1 indicam uma maior taxa de omissões falsas para o grupo em comparação. Por exemplo, o valor de 2.6031 para as mulheres de escolaridade até fundamental II indica que há uma disparidade mais de duas vezes maior na taxa de omissões falsas em comparação com o grupo de referência. Isso sugere que o modelo pode estar sub-representando positivos verdadeiros neste grupo específico.

Para o grupo de mulheres com idade acima de 27 anos, a disparidade da métrica foi, quando comparada ao grupo de referência, de 1.6672. É fundamental que tais disparidades sejam investigadas para compreender as causas subjacentes e garantir que o modelo preditivo não introduza ou exacerbe injustiças. Ações corretivas podem ser necessárias para ajustar o modelo ou o processo de tomada de decisão para abordar essas questões de equidade.

Com relação ao segundo modelo, este buscou incorporar dados de pré-natal, como foco de verificar se o ganho de peso ao longo da gestação pode ser um atributo proxy para uma eventual discriminação algorítmica. Para isso, foi considerado o estado nutricional pré-gestacional (considerando, neste caso, o IMC no primeiro atendimento de pré-natal) e, seguindo as diretrizes do Ministério da Saúde brasileiro, para determinar o ganho de peso total recomendado (em kg) ao final da gestação, conforme Tabela 02. Dessa forma, se uma gestante teve um ganho de peso conforme Tabela 02, ela foi classificada como "ganho de peso dentro da normalidade", caso contrário, recebeu a classificação de "ganho de peso fora da normalidade".

Tabela 2. Ganho de peso total recomendado ao longo da gestação, considerando-se o IMC pré-gestacional [5].

IMC Pré-gestacional (kg/m²)	Classificação	Ganho de peso total recomendado (kg)
Menos de 18,5	Baixo peso	12,5 – 18,0
18,5 – 24,9	Normal	11,5 – 16,0
25,0 – 29,9	Sobrepeso	7,0 – 11,5
30,0 oumais	Obesidade	5,0 – 9,0

O segundo modelo foi construído com metodologia similar ao primeiro modelo e o arquivo de

trabalho para a construção do modelo está disponível no repositório do projeto. Para a avaliação da equidade, estabeleceu-se que o grupo de referência é o de mulheres brancas com idade inferior ou igual a 27 anos, escolaridade acima de fundamental II e com ganho de peso dentro da normalidade.

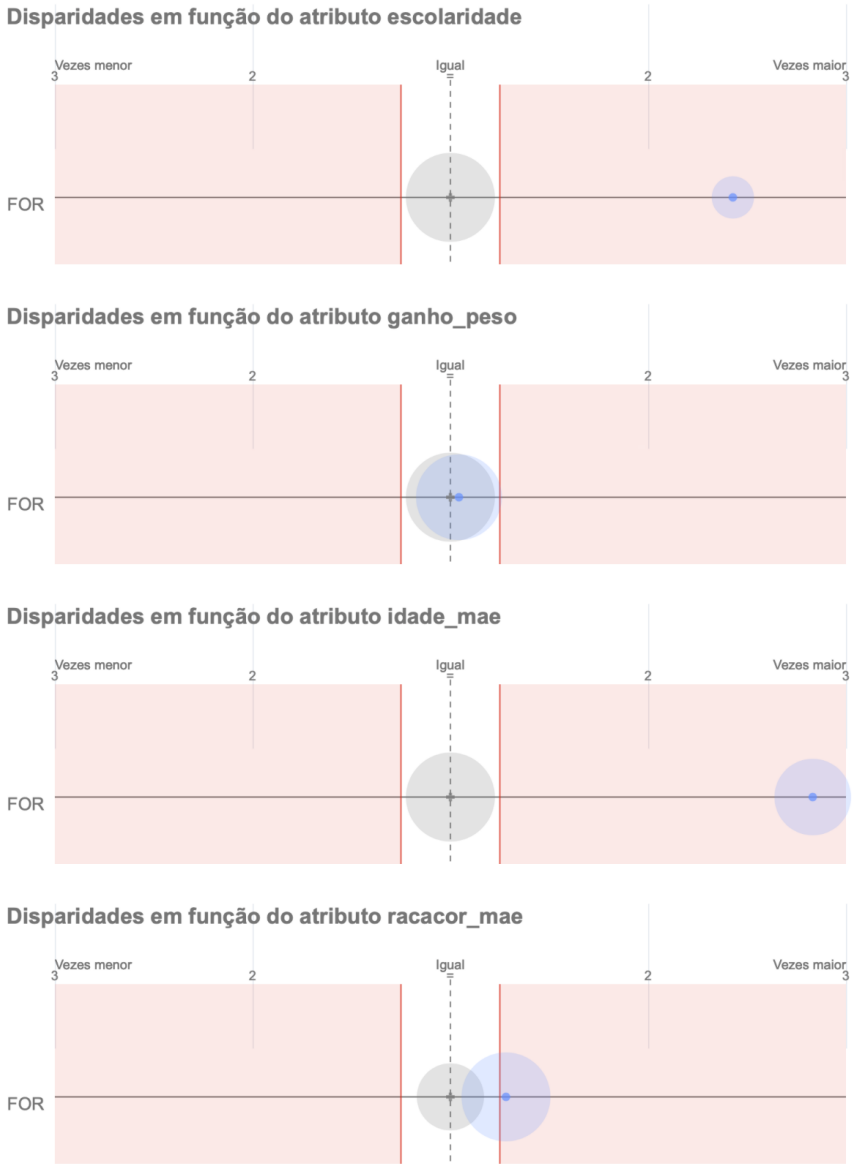
A Figura 11 apresenta os resultados das métricas obtidas, acompanhada pela Figura 12 que apresenta os diagramas visuais para avaliação das disparidades.

Figura 11. Métricas de equidade para o segundo modelo preditivo construído, incorporandodados do pré-natal.

Métricas calculadas:

	attribute_name	attribute_value	Statistical Parity	Impact Parity	fdr	fdr_disparity	fpr	fpr_disparity	for	for_disparity	fnr	fnr_disparity
0	idade_mae	maior_27a	☑	☐	0.9888	0.9998	0.0049	1.3207	0.0006	2.832	0.9167	1.0913
1	idade_mae	menor_igual_27a	☑	☑	0.9891	1	0.0037	1	0.0002	1	0.84	1
2	escolaridade	Acima de fundamental II	☑	☑	0.9873	1	0.0039	1	0.0003	1	0.8571	1
3	escolaridade	Até fundamental II	☐	☐	0.9943	1.0071	0.0055	1.4218	0.0007	2.4283	0.9583	1.1181
4	racacor_mae	Branca	☑	☑	0.9921	1	0.004	1	0.0003	1	0.9091	1
5	racacor_mae	Não branca	☐	☑	0.9873	0.9952	0.0043	1.0611	0.0004	1.281	0.8824	0.9706
6	ganho_peso	dentro da normalidade	☑	☑	0.9887	1	0.004	1	0.0004	1	0.8919	1
7	ganho_peso	fora da normalidade	☑	☑	0.9892	1.0004	0.0044	1.1182	0.0004	1.0432	0.8889	0.9966

Figura 12. Apresentação visual obtida para as disparidades em função dos atributos definidos para representar os grupos.



Os resultados mostram que o atributo 'ganho de peso' foi classificado dentro do intervalo de igualdade, indicando que não é um viés no modelo. Esta descoberta interessou aos gestores públicos, dada a preocupação inicial sobre possíveis preconceitos do modelo contra gestantes obesas. É válido considerar que essa neutralidade pode ser um reflexo das políticas de pré-natal em São Paulo.

Contudo, a análise das outras variáveis revelou que a idade apresenta uma disparidade significativa: mulheres com mais de 27 anos têm quase três vezes mais chances de serem omitidas pelo modelo. Este grupo representa 42.55% das mulheres, descartando a possibilidade de sub-representação. A disparidade relacionada à raça/cor manteve-se semelhante ao modelo anterior.

A implementação do segundo modelo não só permitiu avaliar uma nova variável – o ganho de peso durante a gravidez – mas também proporcionou uma nova perspectiva para o time de parceiros que validou os resultados, refletindo sobre a aplicabilidade do modelo preditivo, assim como possibilitou de maneira efetiva que o grupo de apoio pudesse comparar ambos os modelos. A quantificação numérica, representada pelas métricas calculadas, e a representação visual das disparidades, trazem luz a pontos de discussão vitais para o processo de decisão de incorporar modelos preditos em processos de gestão pública, além de facilitar a comparação de diferentes modelos e de possibilitar que se compreenda a quais grupos de indivíduos o modelo pode beneficiar ou prejudicar, e em qual potencial isso pode acontecer.

3 CONCLUSÃO E RECOMENDAÇÕES

O processo de compreender o conceito de equidade em modelos preditivos, estabelecer um racional para qual métrica deve ser escolhida para avaliar tal equidade e propor uma interface web simplificada para a visualização da equidade, tornou mais evidente esses conceitos e a avaliação da equidade de uma maneira prática, possibilitando uma discussão qualificada sobre a adequabilidade dos modelos preditivos para diversos cenários de atuação pública. Este processo destacou pontos cruciais para a avaliação de modelos preditivos por gestores públicos:

1. A ferramenta estimula a consideração de grupos de referência, permitindo avaliar modelos além de métricas tradicionais como precisão e sensibilidade, e refletir sobre os beneficiários de intervenções baseadas em inteligência artificial.
2. O questionário para definir a métrica de avaliação facilita a identificação de vieses potenciais, exibindo todas as métricas relevantes.
3. A interface visual para cada grupo ajuda na identificação de disparidades. Embora a ferramenta não determine definitivamente se um modelo é justo, ela aponta de forma prática disparidades a serem consideradas.

Na era da inteligência artificial, a equidade em modelos preditivos transcende a mera eficiência técnica, tornando-se uma questão de responsabilidade ética e social. Provedores de soluções de IA e gestores públicos devem, portanto, abraçar a auditoria algorítmica não apenas como um meio de identificar falhas, mas como um pilar fundamental na construção de um futuro mais justo e responsável.

Para os provedores de IA, isso significa ir além da busca por dados representativos e modelos precisos. É crucial incorporar a perspectiva de auditoria desde o início do desenvolvimento dos modelos. Isso envolve uma análise profunda e contínua, garantindo que as questões de parcialidade e impacto social sejam consideradas e endereçadas em todas as etapas. A auditoria algorítmica deve ser vista como um companheiro constante, guiando o desenvolvimento com um olhar crítico e construtivo.

Os gestores públicos, por sua vez, têm um papel decisivo na priorização da auditoria em aplicações de alto impacto, como sistemas usados para determinar a elegibilidade para serviços sociais. Eles devem garantir que as descobertas das auditorias se transformem em ações concretas. Isso envolve não apenas ajustes pontuais, mas uma visão estratégica a longo prazo para melhorar continuamente os sistemas em uso.

A colaboração entre esses dois grupos é vital. Juntos, eles devem estabelecer uma cadência regular para as auditorias, equilibrando a importância dos modelos com as limitações de recursos. Este processo não deve ser episódico, mas um compromisso contínuo, adaptando-se à natureza dinâmica da inteligência artificial.

Além disso, aprender com as auditorias realizadas é essencial. Desenvolver quadros específicos para guiar futuros desenvolvimentos e criar procedimentos para lidar com falhas identificadas são passos cruciais. Isso promove um diálogo mais eficaz entre diferentes áreas de especialização, tornando o desenvolvimento de modelo de IA não apenas mais técnico, mas também mais humano e responsável.

Portanto, encoraja-se provedores de IA e gestores públicos a verem a si mesmos como arquitetos de um futuro em que a tecnologia não apenas avança, mas também protege e valoriza cada indivíduo. Ao integrar auditorias de modelos de IAem seus processos, eles estarão não apenas otimizando sistemas, mas cultivando um ambiente onde a equidade e a responsabilidade social são tão valorizadas quanto a precisão e eficiência.

REFERÊNCIAS BIBLIOGRÁFICAS

[5] ASMUSSEN, K. M.; YAKTINE, A. L., eds. **Weight Gain During Pregnancy**: Reexamining the Guidelines. Washington, DC: Institute of Medicine; National Research Council, [S.l.], [s.d.]. Disponível em: http://www.nap.edu/catalog/12584.html. Acesso em: 19 jan. 2024.

[6] BATISTA, A. F. de M. *et al*.Neonatal mortality prediction with routinely collected data: a machine learning approach. **BMC Pediatrics**, v. 21, p. art. 322, [s.d.]. Disponível em: https://doi.org/10.1186/s12887-021-02788-9. Acesso em: 19 jan. 2024.

[1] DRESSEL, J.; FARID, H. The accuracy, fairness, and limits of predicting recidivism. **Science Advances**, v. 4, n. 1, art. eaao5580, 2018.

[7] EVALUATING MODEL FAIRNESS. **Arize**. [S.l.], [s.d.]. Disponível em: https://arize.com/blog/evaluating-model-fairness/. Acesso em: 19 jan. 2024.

[3] O'NEIL, C. **Weapons of math destruction**: How big data increases inequality and threatens democracy. New York: Broadway Books, 2016.

[2] SALEIRO, P. *et al*. **Aequitas**: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577, 2018.

[4] ZIEWITZ, M. Governing algorithms: Myth, mess, and methods. **Science, Technology& Human Values**, v. 41, n. 1, p. 3-16, 2016.