Enap

Análise de Dados em Linguagem R

Módulo 1

Introdução à análise de dados

Fundação Escola Nacional de Administração Pública

Presidente

Diogo Godinho Ramos Costa

Diretor de Desenvolvimento Profissional

Paulo Marques

Coordenador-Geral de Educação a Distância

Carlos Eduardo dos Santos

Equipe responsável

Ana Carla Gualberto Cardoso (Diagramação, 2020).

Ana Paula Medeiros Araújo (Direção e Produção Gráfica, 2020).

Douglas Gomes Ferreira (Conteudista, 2020).

Guilherme Teles da Mota (Implementação Rise, 2020).

Iara da Paixão Corrêa Teixeira (Designer Instrucional, 2020).

Juliana Bermudez Souto de Oliveira (Revisão Textual, 2020).

Larisse Padua da Silva (Produção Audiovisual, 2020).

Michelli Batista Lopes (Produção Audiovisual e Implementação, 2020).

Patrick Coelho (Implementação Moodle, 2020).

Sheila Rodrigues de Freitas (Coordenação Web, 2020).

Desenvolvimento do curso realizado no âmbito do acordo de Cooperação Técnica FUB / CDT / Laboratório Latitude e Enap.

Curso produzido em Brasília, 2020.

Enap Escola Nacional de Administração Pública

Enap, 2020

Enap Escola Nacional de Administração Pública

Diretoria de Educação Continuada

SAIS - Área 2-A - 70610-900 — Brasília, DF

Sumário

Módulo 1 Introdução à análise de dados

DESTAQUE

Ao final deste módulo, você terá aptidão para listar a importância e os benefícios da análise de dados e reconhecer os principais conceitos relacionados à *data science*, linguagem R e *machine learning* na análise de dados, tendo como foco principal o setor público.

Unidade 1 - Análise de dados no setor público

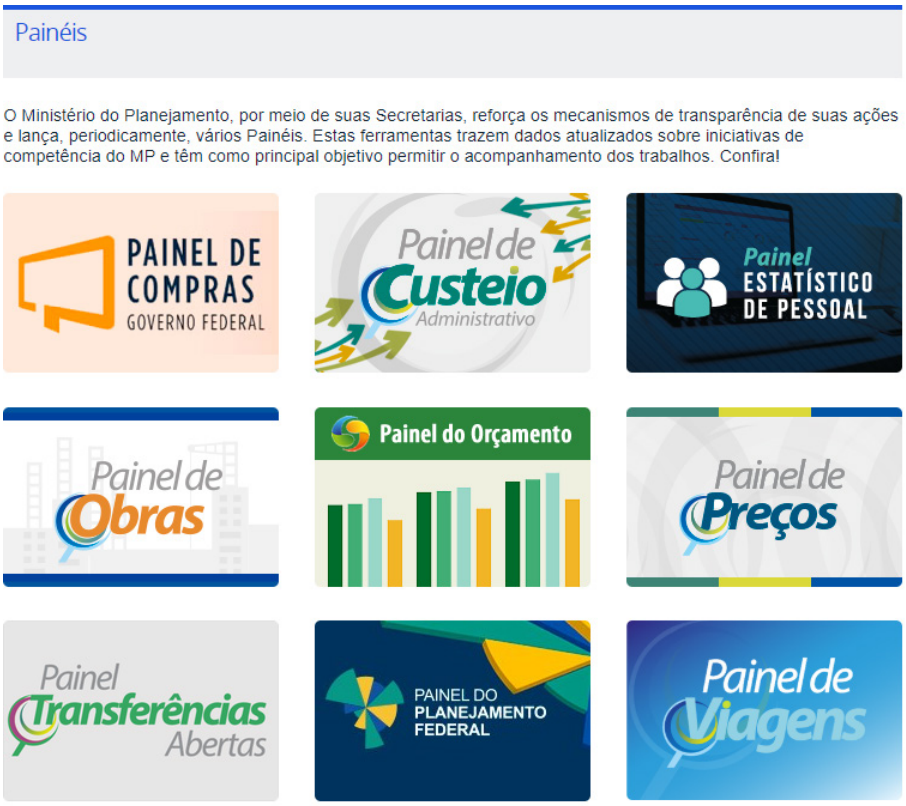
DESTAQUE

Você já parou para pensar na quantidade de dados que são processados a cada minuto por empresas, sejam elas pequenas, médias ou grandes; por bancos, públicos ou privados; e pelos órgãos públicos? E o que essas instituições fazem com esses dados?

Quando bem trabalhados, monitorados e analisados, os dados servem para auxiliar as instituições em muitos aspectos. Cada vez mais os órgãos públicos vêm tomando decisões com base nos dados, seja para detecção de anomalias, monitoramento de indicadores ou melhoria de processos. Muitos órgãos já perceberam a importância de se realizar análise sobre os dados e os ganhos que esta atividade fornece.

O Ministério da Economia, por exemplo, com o objetivo de dar mais transparência em suas ações, disponibiliza diversos painéis públicos em seu portal, conforme imagem a seguir:

Figura 1: Painéis



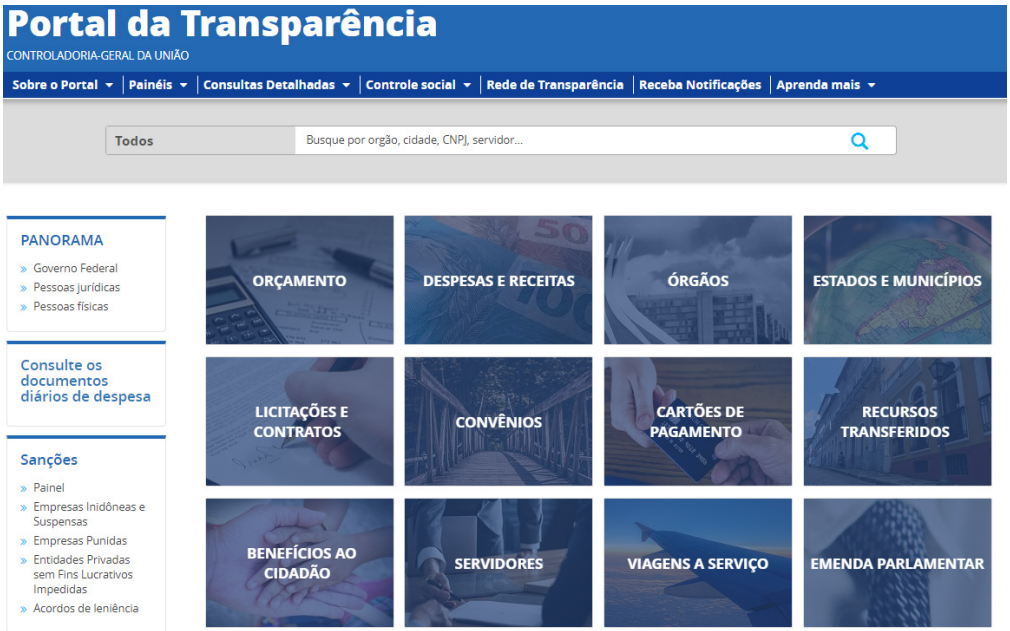
Fonte: Ministério da Economia.

SAIBA MAIS

Para navegar nos painéis, acesse o portal do Ministério da Economia.

Outro exemplo é o Portal da Transparência, mantido pela Controladoria-Geral da União (CGU), que também disponibiliza painéis sobre diversos temas, permitindo que os dados sejam baixados para análise individual. Acompanhe na seguinte imagem:

Figura 2: Portal da Transparência - Painéis



Fonte: Controladoria-Geral da União.

SAIBA MAIS

Para conhecer os painéis disponibilizados pela CGU, acesse o Portal da Transparência.

Diversas áreas governamentais de vários países estão se beneficiando com a análise de dados. Alguns exemplos são:

Saúde
Centros de controle e prevenção de doenças utilizam os dados para prever surtos de gripe e rastrear padrões da doença.

Detecção e prevenção de crimes
O Departamento de Polícia de Durham, no estado da Carolina do Norte (EUA), analisa dados para identificar padrões de atividades criminosas e lugares com alta incidência de criminalidade. Isso ajuda o departamento a decidir onde os policiais devem ser alocados.

Segurança da informação
Nos Estados Unidos, o Departamento de Segurança Interna analisa os dados de tráfego da internet para detectar ameaças e acessos não autorizados.

Desastres naturais
Na Indonésia, a partir de dados históricos coletados por sensores e dados de reclamações dos cidadãos, foi possível identificar áreas propensas a inundações.

Agora que mostramos a importância da análise de dados para os governantes dos países e a sua utilização em incontáveis áreas, vamos saber um pouco sobre como o Brasil vem tratando seus dados na atualidade.

1.1 Análise de dados no Brasil

No Brasil, existem várias iniciativas com foco na análise de dados públicos. Essas iniciativas foram apresentadas na quinta edição do Seminário Internacional sobre Análise de Dados na Administração Pública, que aconteceu em Brasília, em 2019.

Essa edição foi organizada pelo Tribunal de Contas da União, pela Rede de Inovação no Setor Público, pela Controladoria-Geral da União e pela Escola Nacional de Administração Pública, com o apoio da Agência Alemã de Cooperação Internacional. O objetivo principal do evento foi promover a troca de experiências e boas práticas no uso de técnicas de análise e mineração de dados, visando a melhoria da gestão e do controle de órgãos, entidades e políticas públicas.
A seguir lista-se alguns trabalhos que foram apresentados nessa edição:

1. “Análise de dados para localização de vítimas do rompimento da barragem de Brumadinho” pelo Corpo de Bombeiros Militar de Minas Gerais.

2. “Análise de vínculos para detecção de fraudes” pelo Tribunal de Contas do Estado de São Paulo e Conselho Administrativo de Defesa Econômica.

3. “Fiscalização contínua de folhas de pagamento da Administração Pública” pelo Tribunal de Contas da União.

4. “Detecção de anomalias para identificar a prática de conluio em licitações do governo federal” pela Controladoria-Geral da União.

5. “Sinalização de corridas suspeitas do TaxiGov do governo federal com geoprocessamento e detecção de anomalias” pelo Ministério da Economia.

Esses são alguns exemplos de análise utilizando dados de órgãos públicos. Vale ressaltar que nem sempre essas análises são feitas com dados abertos, normalmente devido a questões de sigilo das informações.
Vamos refletir um pouco sobre os dados na sua organização. Você sabe se no seu órgão os dados são extraídos, analisados e monitorados, a fim de subsidiar as decisões estratégicas? Como isso é feito? Alguns desses dados ficam disponíveis para o público externo?

1.2 Dados abertos

Com o objetivo de dar transparência ao cidadão, diversos países disponibilizam na internet bases de dados governamentais classificadas como dados abertos.

Mas o que são dados abertos?

De acordo com a Open Knowledge Internacional citada no Portal Brasileiro de Dados Abertos:

> **Dados são abertos quando qualquer pessoa pode livremente acessá-los, utilizá-los, modificá-los e compartilhá-los para qualquer finalidade, estando sujeito a, no máximo, a exigências que visem preservar sua proveniência e sua abertura. (OPEN KNOWLEDGE INTERNACIONAL apud BRASIL).**

No Brasil, a Lei de Acesso à Informação (LAI) dispõe que os órgãos públicos devem promover a divulgação de informações de interesse coletivo ou geral na internet. A lei também define as hipóteses de sigilo e de informações pessoais, que são as exceções à regra de que os dados devem ser abertos.

SAIBA MAIS

Para saber mais a esse respeito, acesse o Painel Lei de Acesso à Informação, que contém dados sobre número de pedidos e recursos, cumprimento dos prazos, perfil dos solicitantes, transparência ativa e outros. O painel foi desenvolvido pela Controladoria-Geral da União com dados extraídos do Sistema Eletrônico de Informações ao Cidadão (e-SIC).

Vários órgãos no Brasil já disponibilizam dados abertos e o mapeamento dessas iniciativas está consolidado em Catálogos Dados Brasil, que contém os endereços eletrônicos dos dados abertos de órgãos e instituições públicas do país.

Unidade 2 - Conceitos relacionados à análise de dados

Agora, vamos relembrar alguns conceitos fundamentais no estudo da análise de dados, em especial os mais utilizados na Administração Pública, foco do nosso estudo. São conceitos relacionados à *data science*, *machine learning* e linguagem R.

2.1 O que é *Big Data*?

Você sabe o que significa *Big Data*? Pensando na tradução do inglês para o português, *big* significa grande e *data* significa dados, o que nos faz inferir que é um conjunto de palavras que quer dizer grandes dados.

Vamos verificar como alguns renomados autores conceituam *Big Data*. Para tanto, observe o quadro elaborado por Feitas Junior *et al*. (2016, p. 532) para o artigo intitulado *Big Data e Gestão do Conhecimento: Definições e Direcionamentos de Pesquisa*, que faz uma revisão de literatura de artigos publicados em periódicos e congressos científicos nas bases de dados EBSCOhost e Web of Science.

Definições de *Big Data*	
Definições	**Autores**
Trata-se de um termo geral para a enorme quantidade de dados digitais coletados a partir de todos os tipos de fontes.	Kim, Trimi e Ji-Hyong (2014).
Denotam um maior conjunto de dados ao longo do tempo, conjuntos de dados estes que são grandes demais para serem manipulados por infraestruturas de armazenamento e processamento regulares.	Mahrt e Scharkow (2013).
Dados demasiadamente volumosos ou muito desestruturados para serem gerenciados e analisados através de meios tradicionais.	Davenport (2012) e Kwon (2014).
Refere-se ao conjunto de dados cujo tamanho está além da habilidade de ferramentas típicas de banco de dados em capturar, gerenciar e analisar.	Di Martino *et al*. (2014).
São conjuntos de dados que são tão grandes que se tornam difíceis de trabalhar com o uso de ferramentas atualmente disponíveis.	Rajesh (2013).

Fonte: Freitas Junior *et al*. (2016), com adaptações.

O conceito de *Big Data* pode ser caracterizado por quatro pilares, conhecidos por seus 4 Vs:

Volume: refere-se à grande quantidade de dados.

Variedade: refere-se a diversas fontes e diferentes formatos de onde surgem os dados, por exemplo, podemos ter a informação em uma imagem, texto ou vídeo.

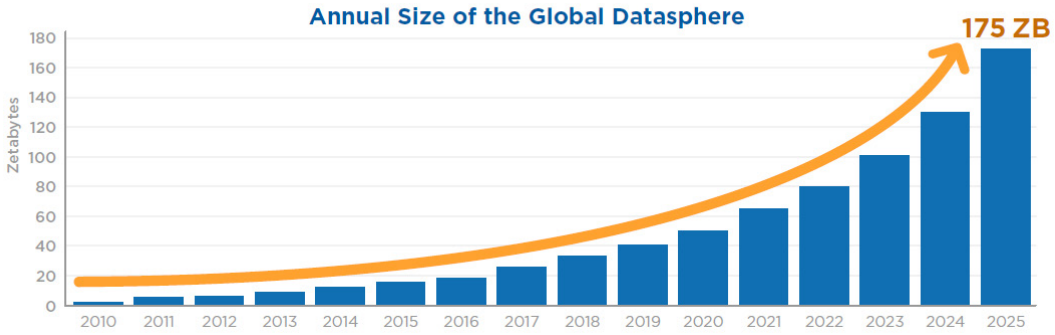
Velocidade: refere-se à velocidade que os dados são gerados, armazenados e recuperados.

Veracidade: refere-se à qualidade, volatilidade e validade dos dados.

De acordo com o International Data Corporation (IDC), estima-se que serão produzidos 175 zettabytes de dados no mundo até 2025, conforme gráfico a seguir.

Gráfico 1: Crescimento dos dados no mundo

Figure 1 – Annual Size of the Global Datasphere



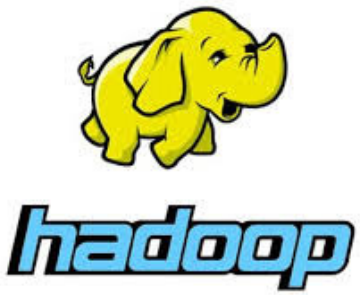
Fonte: International Data Corporation (2018).

Diante do exposto, pode-se afirmar que o termo *Big Data* refere-se a uma grande quantidade de dados que excede a capacidade de processamento de um banco de dados tradicional. Devido a isso, torna-se necessária a utilização de arquiteturas paralelas e distribuídas para armazenar e processar esses grandes conjuntos de dados. Uma das tecnologias que foi desenvolvida para esse propósito é o Hadoop.

2.2 O que é Hadoop?

Figura 1: Hadoop

O Apache Hadoop é um *framework open source* (estrutura de código aberto) para processamento e gerenciamento de grandes volumes de dados (*Big Data*). Também pode ser definido como um ecossistema de ferramentas e métodos para armazenamento, distribuição e análise de dados estruturados e não estruturados.



Fonte: Martech Forum (c2019).

O uso da plataforma Hadoop tem como principais benefícios a sua capacidade de armazenar, gerenciar e analisar grandes quantidades de dados estruturados e não estruturados de forma rápida, confiável, flexível e de baixo custo.

SAIBA MAIS

Para saber mais, acesse a página oficial do Haddop.

2.3 O que são dados estruturados e não estruturados?

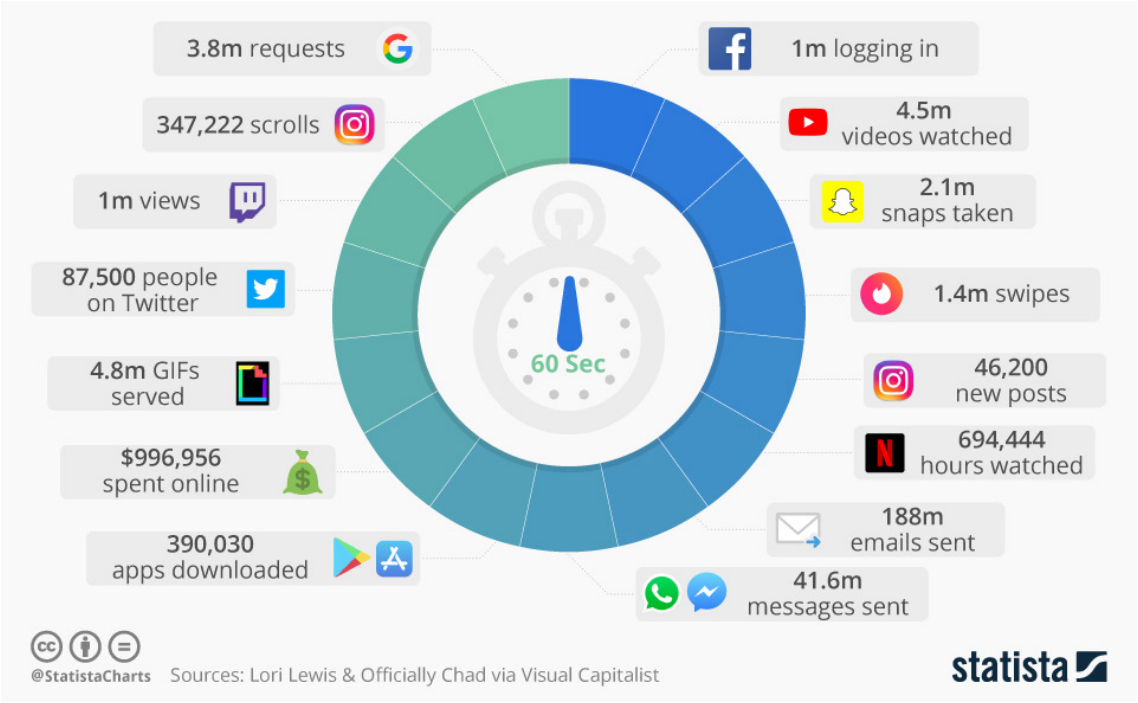
De acordo com a sua estrutura, os dados podem ser separados em duas categorias: dados estruturados e dados não estruturados.

Confira a diferença entre eles no quadro a seguir.

	Dados estruturados	Dados não estruturados
Definição	São organizados em tabelas (linhas e colunas) que representam observações e características.	Os dados não seguem um padrão de organização.
Características	Estrutura rígida e previamente planejada.	Estrutura flexível e dinâmica ou sem estrutura.
Exemplo	Banco de dados, pois são estruturados conforme a definição de um esquema, ou seja, define as tabelas com seus respectivos campos (atributos) e tipos (formato).	Textos, arquivos, documentos, imagens, vídeos, áudios, redes sociais e dados que estão na web.

Estima-se que 80% a 90% dos dados do mundo são não estruturados. Com o crescimento do uso de smartphones e mídias sociais, a produção de dados torna-se cada vez maior. O gráfico a seguir apresenta a estimativa de dados criados na internet em 1 minuto:

Gráfico 2: Um minuto na internet em 2019



Fonte: Feldman (2019).

2.4 O que é *data science*?

Em algum momento, você já deve ter se deparado com o termo *data science* ou ciência de dados. Mas você sabe o que ele significa?

DESTAQUE

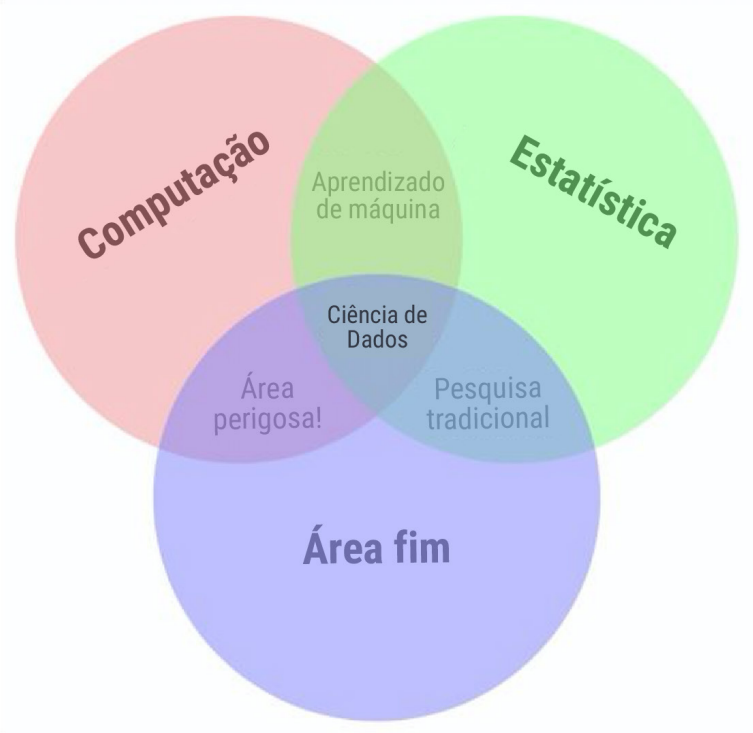
> ***Data science* ou ciência de dados é a arte de extrair conhecimento por meio dos dados para se tomar melhores decisões, realizar previsões e entender o passado.**

Portanto, sem entrar em pormenores, a ciência de dados é encarregada de transformar dados em informação.

A ciência de dados é uma área multidisciplinar baseada em conceitos e áreas bem consolidadas, como a matemática, a estatística e a ciência da computação.

Analisando o Diagrama de Venn, é possível perceber que *data science* é a intersecção de várias áreas de conhecimento. Confira na imagem a seguir.

Figura 4: Diagrama de Venn mostrando habilidades necessárias para um cientista de dados



Fonte: Conway (2013).

Então, vamos relembrar a definição atribuída para cada uma dessas áreas na ciência de dados:

Estatística
Consiste em desenvolver e aplicar métodos para coletar, analisar e interpretar os dados.

Matemática
Em diversas técnicas da matemática, como álgebra linear e cálculo, a análise de dados é usada para a criação de algoritmos inteligentes.

Área específica de conhecimento
A análise de dados pode ser aplicada em diversas áreas de conhecimento, como finanças, administração, negócios, mídias sociais, governo e ciência.

Machine learning
É uma área de estudo que busca dar aos computadores a habilidade de aprender sem serem programados explicitamente.

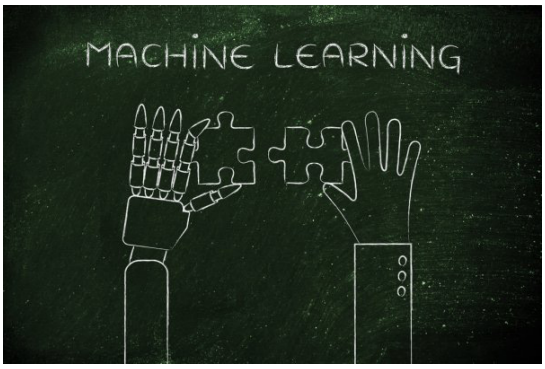
Ainda com base no Diagrama de Venn, podemos inferir que o aprendizado de máquina combina o poder dos computadores com os algoritmos de aprendizagem, baseados em matemática e estatística, para automatizar a descoberta de relacionamentos nos dados.

2.5 *Machine learning*: o que é e como está sendo aplicada?

Uma das maiores tendências da modernidade tecnológica é a *machine learning* ou aprendizagem de máquina.

Segundo artigo da Data Science Academy (2018), "a aprendizagem de máquina é um subconjunto da inteligência artificial (IA), o segmento da ciência da computação que se concentra na criação de computadores que pensam da maneira que os humanos".

Figura 5: Conceito de aprendizado de máquina



Fonte: Faithie (2016).

Esse mesmo artigo ressalta que várias organizações e indústrias já experimentam a *machine learning* em suas atividades ou processos.

Alguns exemplos da aplicação do aprendizado de máquina na atualidade são:

1. Detecção de anomalias e fraudes.
2. Identificação de mensagens spam em e-mails.
3. Segmentação de clientes.
4. Carros e drones autônomos.
5. Mecanismos de busca.
6. Segurança de tecnologia da informação.
7. Logística.

SAIBA MAIS

Para saber um pouco mais sobre machine learning e sua aplicação na atualidade, acesse o vídeo O Que é *Machine Learning* (Inteligência Artificial)? com Marcelo Tas.

2.6 Tipos de aprendizagem de máquina

A aprendizagem de máquina clássica ou *machine learning* é frequentemente dividida em três categorias amplas. Vamos conhecer as características e os exemplos de aplicação para cada uma delas:

Supervisionada

O algoritmo procura associações entre os atributos (variáveis preditoras) e a variável resposta (variável que se quer prever) de um *dataset*. A partir dessas associações, é possível realizar previsões quando o algoritmo for apresentado a novos dados.

Exemplo: com base nos dados históricos de pacientes, podemos prever se um novo paciente irá desenvolver ou não uma determinada doença.

Não supervisionada

O objetivo é agrupar os dados com base em características similares, não sendo necessário apresentar o algoritmo à variável resposta (variável que se quer prever).

Exemplo: pode ser utilizada para identificar anomalias ou agrupar clientes com base em comportamentos similares.

Aprendizagem por reforço

O algoritmo aprende com base nas interações com o ambiente. Não são apresentadas as ações que devem ser tomadas, apenas as consequências das ações.

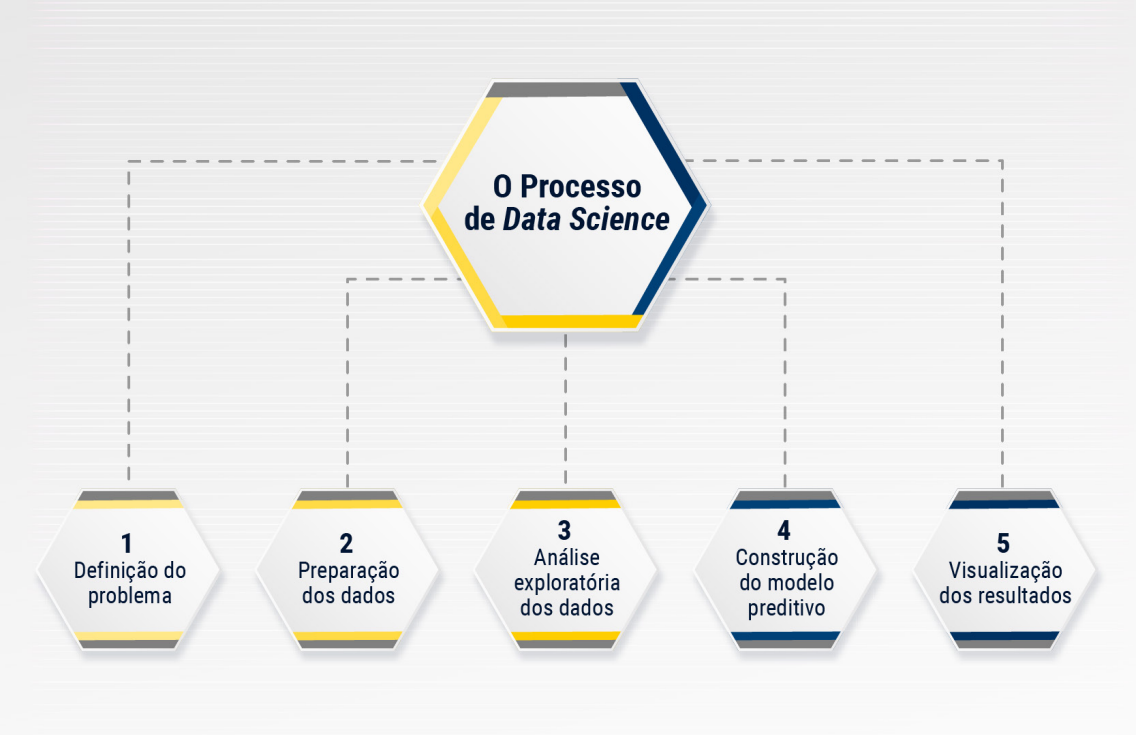
Exemplo: carros autônomos.

Em entrevista ao The Wall Street Journal, Thomas H. Davenport disse que:

> **seres humanos podem, normalmente, criar um ou dois modelos bons por semana; *machine learning* pode criar milhares de modelos por semana.**

2.7 O processo de *data science*

Para ser bem sucedido, o processo de *data science* deve seguir pelo menos 5 etapas básicas.



Definição do problema

Nesta etapa busca-se entender o problema e as questões de negócio que devem ser respondidas. É o momento de se fazer alguns questionamentos, por exemplo:

- O que se quer resolver com a análise?
- Que tipo de análise será feita? Descritiva, diagnóstica, preditiva?
- Quais dados são necessários?

Preparação dos dados

Está relacionada com a obtenção, limpeza, normalização e transformação dos dados.
Podemos ter dados não padronizados entre diferentes bases. Por exemplo, o campo sexo pode ser preenchido de diferentes formas: masculino/feminino, M/F ou 0/1. Antes de partir para as etapas seguintes, é necessário consolidar esses dados.

Análise exploratória dos dados

Nesta etapa busca-se obter um panorama de como os dados estão organizados. A apresentação dos dados é fundamental, pois o objetivo é entender as características e os relacionamentos deles. É uma atividade inicial para entender melhor como estão organizados.

Algumas questões que podem ser respondidas com a análise exploratória:

- Quais são os tipos das variáveis (atributos)?
- Como estão as distribuições dos dados?
- Existem valores *missing* (NA/Null)?
- Existem variáveis redundantes?
- Existem *outliers*?
- Quais variáveis possuem correlação?

Construção do modelo preditivo

Modelagem preditiva é uma técnica estatística para realizar previsões com base em dados históricos por meio da criação de um modelo.
Nem todos os projetos passam por esta etapa, depende da definição do problema de negócio.

Visualização dos resultados

Na última etapa do processo são apresentados os resultados da análise. Existem diversas ferramentas que podem ajudar nesta parte.
Podemos apresentar os resultados de várias formas: por meio de um *dashboard* (painel gerencial), relatório, planilha, arquivo csv e muitos outros. A escolha depende do que se está apresentando, qual o contexto e para quem será apresentado.

2.8 Linguagem R

Para cada etapa do processo de *data science*, existem diversas ferramentas que podem ser utilizadas, dentre elas a linguagem R. Certamente, você ouviu falar ou até já usou essa linguagem de programação.

A linguagem R é uma linguagem de programação usada para análise estatística e produção de gráficos. De acordo com a definição da Wikipedia, trata-se de:

> **uma linguagem de programação multi-paradigma [com ênfase em programação funcional], dinâmica, fracamente tipada, voltada à manipulação, análise e visualização de dados.**

Com ela, é possível preparar e explorar os dados, construir modelos preditivos e apresentar os resultados por meio de gráficos e *dashboards*. Possui diversos pacotes desenvolvidos para facilitar a aplicação de técnicas estatísticas, como: regressão linear, análise de séries temporais, classificação, entre outras.

Outra grande vantagem dessa linguagem é que, atualmente, mais de 15 mil pacotes estão disponíveis no CRAN (Comprehensive R Archive Network) que podem ser acessados gratuitamente no seguinte endereço: cran.r-project.org.

A seguir, acompanhe como ela foi criada.

> **Os criadores da linguagem R, Ross Ihaka e Robert Gentleman, professores na Universidade de Auckland, na Nova Zelândia, sentiram-se incomodados com o fato de que a maioria dos seus alunos não tinham acesso a grande parte dos softwares disponíveis na época, pois eram pagos. Esses alunos, após saírem da universidade, dificilmente tinham acesso a softwares de estatística ou condições financeiras de comprar as licenças destes softwares. Isso se mostrava ainda pior com alunos estrangeiros, já que muitos países sequer tinham representantes comerciais para vender esses softwares.**
>
> **Para resolver esse problema, eles criaram a linguagem R, em 1993, baseando-se nas ideias da linguagem S, que também era uma linguagem de computador voltada para cálculos estatísticos. O nome linguagem R é devido às iniciais de seus idealizadores, Ross e Robert, mas foi somente no ano de 1995 que eles resolveram autorizar a distribuição do R sobre uma licença livre, após seus alunos começarem a distribuir as cópias do recém-criado R, provocando o aumento do interesse sobre a linguagem. (RIBAS, c2020).**

Referências

Unidade 1 - Análise de dados no setor público

BRASIL. Controladoria-Geral da União. **Portal da Transparência**. Brasília: CGU, c2020. Disponível em: http://portaltransparencia.gov.br/. Acesso em: 7 ago. 2020.

BRASIL. Ministério do Planejamento, Desenvolvimento e Gestão. **Painéis**. Brasília: MPDG, [2020]. Ambiente em migração. Disponível em: http://www2.planejamento.gov.br/planejamento/paineis. Acesso em: 7 ago. 2020.

BRASIL. Ministério do Planejamento, Desenvolvimento e Gestão. Secretaria de Tecnologia da Informação. **O que são dados abertos?** Brasília: MPDG, [2020]. Disponível em: http://dados.gov.br/pagina/dados-abertos. Acesso em: 7 ago. 2020.

BRASIL. Ministério do Planejamento, Desenvolvimento e Gestão. Secretaria de Tecnologia da Informação. **Portal Brasileiro de Dados Abertos.** Brasília: MPDG, [2020]. Disponível em: dados.gov.br. Acesso em: 7 ago. 2020.

GOMES, G. L. Análise de dados na prática com R Studio. **DevMedia**, Brasília, 2018. Disponível em: https://www.devmedia.com.br/analise-de-dados-na-pratica-com-r-studio/39279. Acesso em: 7 ago. 2020.

JOSEPH, R. Big Data Analytics in Government: How the Public Sector Leverages Data Insights. **Intellectyx**, Denver, 26 jun. 2019. Disponível em: https://www.intellectyx.com/blog/big-data-analytics-in-government-how-the-public-sector-leverages-data-insights/. Acesso em: 7 ago. 2020.

PERRICOS, C.; KAPUR, V. Anticipatory government: Preempting problems through predictive analytics. **Deloitte Insights**, New York, 24 jun. 2019. Disponível em: https://www2.deloitte.com/us/en/insights/industry/public-sector/government-trends/2020/predictive-analytics-in-government.html. Acesso em: 7 ago. 2020.

SEMINÁRIO INTERNACIONAL SOBRE ANÁLISE DE DADOS NA ADMINISTRAÇÃO PÚBLICA, 5., 2019, Brasília. **Programa** [...]. Brasília: TCU: Enap, 2019. Disponível em: http://www.brasildigital.gov.br/brasil-digital/programa/. Acesso em: 7 ago. 2020.

SIX Big Data Use Cases for the Public Sector. **Ingram Micro**, Irvine, 25 jan. 2017. Disponível em: https://imaginenext.ingrammicro.com/data-center/six-big-data-use-cases-for-the-public-sector. Acesso em: 7 ago. 2020.

Unidade 2 - Conceitos relacionados à análise de dados

17 CASOS de uso de machine learning. **Data Science Academy**, [*s. l.*], 8 ago. 2018. Disponível em: http://datascienceacademy.com.br/blog/17-casos-de-uso-de-machine-learning/. Acesso em: 25 maio 2020.

CÁNEPA, G. **What You Need to Know about Machine Learning**. Birmingham: Packt Publishing, 2016.

CUESTA, H.; KUMAR, S. **Practical Data Analysis**. 2. ed. Birmingham: Packt Publishing, 2016.

FELDMAN, S. A Minute on the Internet in 2019. **Statista**, New York, 29 mar. 2019. Disponível em: https://www.statista.com/chart/17518/internet-use-one-minute/. Acesso em: 7 ago. 2020.

FREITAS JUNIOR, J. C. S.; MAÇADA, A. C. G.; OLIVEIRA, M.; BRINKHUES, R. A. Big Data e Gestão do Conhecimento: Definições e Direcionamentos de Pesquisa. **Revista Alcance**, v. 23, n. 4, p. 529-546, out./dez. 2016. Disponível em: https://www.redalyc.org/jatsRepo/4777/477749961006/477749961006.pdf. Acesso em 22 maio 2020.

FUENTES, A. **Hands-On Predictive Analytics with Python**. Birmingham: Packt Publishing, 2018.

GOLLAPUDI, S. **Practical Machine Learning**. Birmingham: Packt Publishing, 2016.

HADOOP: O que é, conceito e definição. **Cetax**, [*s. l.*]. Disponível em: https://www.cetax.com.br/blog/apache-hadoop/. Acesso em: 25 maio 2020.

LANTZ, B. **Machine Learning with R**. 2. ed. Birmingham: Packt Publishing, 2015.

LIU, Y. H. **Python Machine Learning By Example**. Birmingham: Packt Publishing, 2017.

MACHINE Learning: o que é e qual sua importância? **SAS**, São Paulo, 2019. Disponível em: https://www.sas.com/pt_br/insights/analytics/machine-learning.html. Acesso em: 10 ago. 2020.

MONTEIRO, L. P. Dados Estruturados e Não Estruturados. **Blog Tecnologia da Informação**. São Paulo: Universidade da Tecnologia, 2019. Disponível em: https://universidadedatecnologia.com.br/dados-estruturados-e-nao-estruturados/. Acesso em: 25 maio 2020.

OZDEMIR, S. **Principles of Data Science**. Birmingham: Packt Publishing, 2016.

PENG, R. D. **R Programming for Data Science**. Victoria, BC: Leanpub, 2015.

R (linguagem de programação). *In*: WIKIPÉDIA: a enciclopédia livre. [*S. l.*: *s. n.*]. Disponível em: https://pt.wikipedia.org/wiki/R_%28linguagem_de_programa%C3%A7%C3%A3o%29. Acesso em: 10 ago. 2020.

RAVICHANDIRAN, S. **Hands-On Reinforcement Learning with Python**. Birmingham: Packt Publishing, 2018.

REINSEL, D.; GANTZ, J.; RYDNING, J. **The Digitization of the World:** From Edge to Core. Framingham, MA: IDC, 2018. Disponível em: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. Acesso em: 7 ago. 2020.

REZA, M. **Machine Learning**. Pittsburg, PA: CMUCC, c1995. Disponível em: http://www.contrib.andrew.cmu.edu/~mndarwis/ML.html. Acesso em: 7 ago. 2020.

RIBAS, M. Linguagem de Programação R. **InfoEscola**, [*s. l.*], c2020. Disponível em: https://www.infoescola.com/informatica/linguagem-de-programacao-r/. Acesso em: 10 ago. 2020.

TAS, M. **O que é Machine Learning (Inteligência Artificial)?** [*S. l.*: *s. n.*], 16 abr. 2017. 1 vídeo (5 min). Disponível em: https://www.youtube.com/watch?v=Z1YHbl0lh88. Acesso em: 10 ago. 2020.

WALKOWIAK, S. **Big Data Analytics with R**. Birmingham: Packt Publishing, 2016.