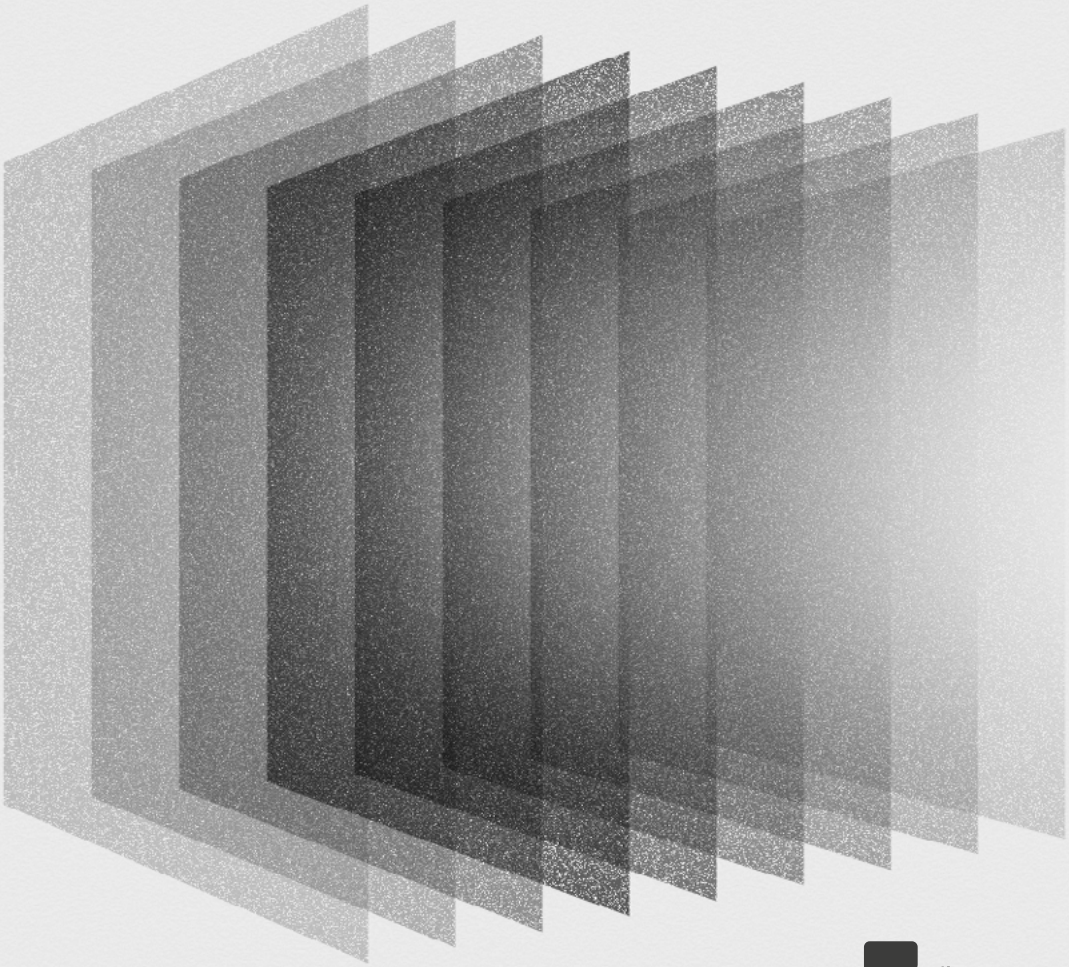
TÓPICOS EM GERENCIAMENTO DE DADOS

Organizadoras
Crishane Azevedo Freire
DamiresYluska de Souza Fernandes



editoraIFPB

TÓPICOS EM GERENCIAMENTO DE DADOS

Organizadoras
Crishane Azevedo Freire
DamiresYluska de Souza Fernandes

editoraIFPB

João Pessoa, 2022

Dados Internacionais de Catalogação na Publicação (CIP)
Departamento de Bibliotecas-DBIBLIO/IFPB/Reitoria

T674 Tópicos em gerenciamento de dados/Organizadores: Crishane Azevedo Freire; Damires Yluska de Souza Fernandes. – João Pessoa/PB: IFPB, 2022.

125f.: Il.
E-book (PDF)
ISBN: 978-65-87572-40-6

I. Gerenciamento de dados 2. Inteligência artificial 3. Tecnologia da informação 4. Tecnologia da Computação I. Instituto Federal de Educação da Paraiba- IFPB. II. Título.

CDU: 004.042

Ficha catalográfica elaborada pelo Departamento de Bibliotecas - DBIBLIO/IFPB

editoraIFPB

CONTATO
Av. João da Mata, 256 - Jaguaribe. CEP: 58015-020, João Pessoa - PB.
Fone: (83) 3612-9722 | E-mail: editora@ifpb.edu.br

SUMÁRIO

Capítulo 4

APRESENTAÇÃO

O livro "Tópicos em Gerenciamento de Dados" é resultado de trabalhos de pesquisa realizados ao longo da disciplina Banco de Dados do Mestrado Profissional em Tecnologia da Informação do IFPB – *Campus* João Pessoa. Os capítulos produzidos abordam temas relevantes da área de Banco de Dados e promovem discussões sobre fundamentos, técnicas, tecnologias, cenários de utilização, desafios e tendências dessa temática. Os quatro capítulos gerados constituem uma excelente oportunidade para familiarização, tanto para acadêmicos quanto para profissionais da área de Tecnologia da Informação, a respeito das questões neles abordadas.

O Capítulo 1, intitulado "Aprendizado de Máquina Supervisionado: Introduzindo Conceitos e Aplicações", discute histórico, conceituação, potencial e descrição de alguns algoritmos, além de casos reais em que estes são utilizados e desafios dessa área, tanto do ponto de vista de empresas quanto da academia. No Capítulo 2, o assunto "Aprendizado de Máquina Não Supervisionado: da Teoria à Aplicabilidade Utilizando Agrupamento" aborda princípios, técnicas, algoritmos, exemplos e desafios direcionados à aplicação de métodos de agrupamento. O Capítulo 3 apresenta conceitos, técnicas, tecnologias e aplicações sobre o tema "Visualização de Dados: Uma Abordagem Introdutória no Contexto de Big Data". Já o Capítulo 4 conclui o livro abordando a "Introdução à Privacidade de Dados e à Lei Geral de Proteção de Dados", por meio de conceitos, técnicas, tendências e perspectivas em relação à nova Lei brasileira.

Gostaríamos de agradecer aos autores pela dedicação aos trabalhos e pela produção dos textos finais. Esperamos que os resultados sejam úteis a toda comunidade de Tecnologia da Informação e a outros interessados nos conteúdos tratados.

As organizadoras

PREFÁCIO

É com grande alegria e satisfação que escrevo o prefácio do livro "Tópicos em Gerenciamento de Dados", fruto da árdua e primorosa organização das professoras do Programa de Pós-Graduação em Tecnologia da Informação (PPGTI) do IFPB Damires Yluska de Souza Fernandes e Crishane Azevedo Freire. O livro contém quatro capítulos e foi resultado de estudos desenvolvidos na disciplina de Banco de Dados do Mestrado Profissional em Tecnologia da Informação, cursada por estudantes da linha de pesquisa Gerenciamento e Desenvolvimento de Sistemas (GDS), no primeiro semestre de 2020.

A obra aborda tópicos relevantes no cenário de Gerenciamento de Dados, apresentando, de forma consistente, revisões bibliográficas enriquecidas com os principais desafios e cenários de uso inerentes a cada temática estudada. Trata-se de uma fonte de pesquisa para estudantes de graduação e pós-graduação, bem como para profissionais da indústria de Computação, em áreas de grande interesse como Big Data, Aprendizado de Máquina, Integração de Dados, Visualização de Dados e Lei Geral de Proteção de Dados.

É de vital importância destacar que a publicação deste livro segue os Referenciais de Formação para os Cursos de Pós-Graduação Stricto Sensu em Computação, publicados em 2019 pela Sociedade Brasileira da Computação. Os referenciais recomendam, como competência geral do eixo Pesquisa, a realização de estudos ou levantamentos do estado da arte de temas relacionados às linhas de pesquisa do programa, aplicando-os a uma problemática de interesse de seu ambiente de exercício profissional. Logo, o resultado alcançado com a obra está alinhado à competência geral do eixo Organização da Informação, que espera dos estudantes um adequado gerenciamento da informação, dos recursos bibliográficos e das fontes de informação tecnológica, de modo a identificar evidências que apoiem suas visões de pesquisa e desenvolvimento tecnológico, sintetizando informação, dados e ideias.

Por fim, não poderia deixar de congratular os onze autores do livro, estudantes do PPGTI, que desenvolveram um trabalho brilhante com o apoio das professoras da disciplina. Além do valor científico bibliográfico da obra entregue, este livro contribui para o processo de avaliação quadrienal da CAPES, à medida que agrega à produção do PPGTI um livro composto por quatro capítulos, todos com autores do programa, além da organização editorial, que é um dos dez produtos técnico-tecnológicos avaliados e pontuados pela coordenação da área de Ciência da Computação.

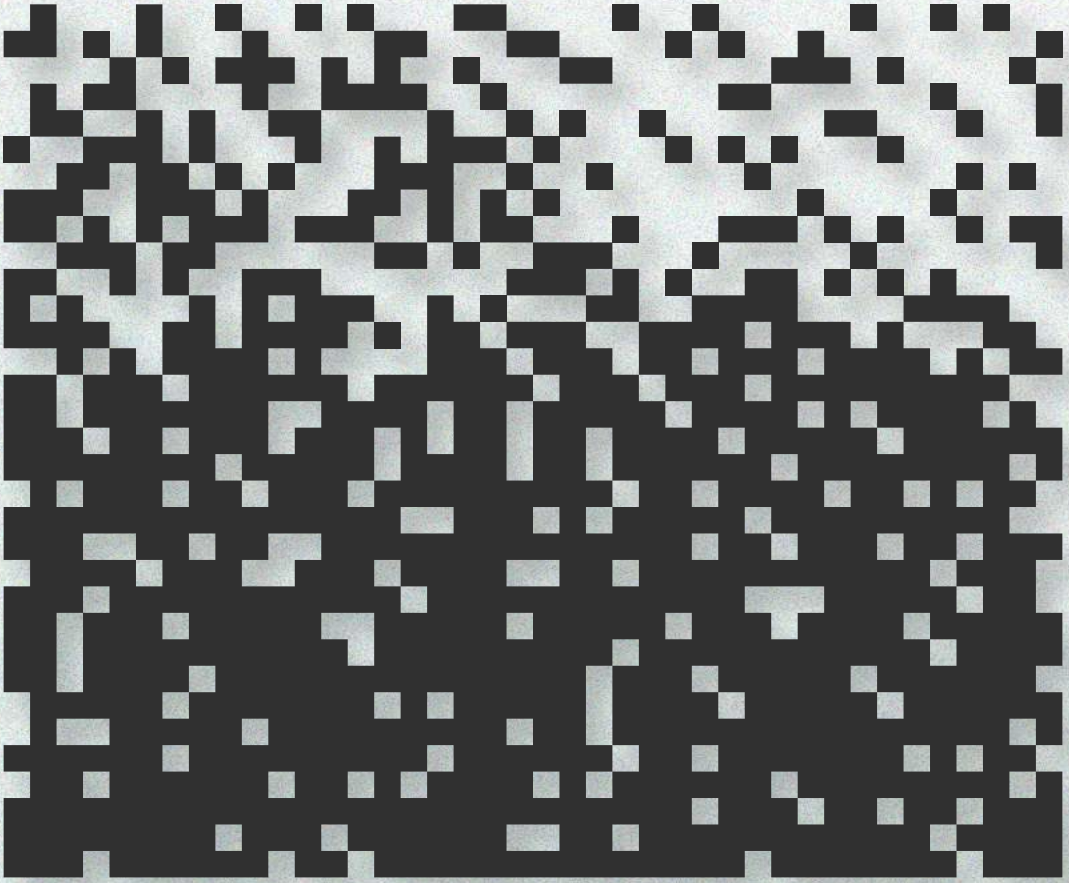
Francisco Petrônio Alencar de Medeiros
Coordenador do Programa de Pós-Graduação
em Tecnologia da Informação do IFPB

Capítulo 1

Aprendizado de Máquina Supervisionado:

Introduzindo Conceitos e Aplicações

Alysson Messias da Silva
Ayrton Douglas Rodrigues Herculano
Helton Souza Lima



1. Introdução

A empresa de consultoria *PricewaterhouseCooper* (PwC) realizou uma pesquisa em 2017 (PWC, 2017) que apontava que a inteligência artificial iria contribuir com mais de US$ 15 trilhões em todas as áreas da economia globalmente. Nesse cenário, a pesquisa mostrava que o aprendizado de máquina, de forma geral, seria uma das principais tecnologias a serem adotadas, além de *chatbots* e assistentes digitais. As áreas finalísticas indicadas que deveriam e deverão usar a tecnologia são as mais diversas, como por exemplo: serviços de saúde; automóveis e mobilidade urbana; serviços financeiros; logística e transporte; tecnologia da informação; companhias de seguros; geração de energia; indústria de manufatura. Nesse contexto, a pesquisa chama atenção para um tema relevante a ser aprofundado, investido e colocado em prática.

Além dessa pesquisa, em outra realizada no ano seguinte (PWC BRASIL, 2018), a PwC apontou que empresas de serviços financeiros estavam obtendo maior valor a partir da utilização de tecnologias como inteligência artificial e análises avançadas, incluindo alguns tipos de métodos de aprendizado de máquina, por exemplo, no combate a fraudes e crimes econômicos. O resultado da pesquisa foi obtido através de entrevistas com 7200 participantes representantes de empresas em 123 países. No relatório da pesquisa, pontuou-se que as empresas de países em desenvolvimento estão investindo em tecnologias avançadas a um ritmo mais rápido do que as nações desenvolvidas. Dessa forma, esse último estudo indica também que pode haver um grande crescimento na procura e na utilização das tecnologias que envolvem a inteligência artificial e, particularmente, o aprendizado de máquina no Brasil, sendo um bom momento para a realização de pesquisas e aplicação dessa ciência em diversas áreas de conhecimento.

Uma das categorias de aprendizado de máquina é o denominado "aprendizado supervisionado". De modo geral, o aprendizado supervisionado trata de problemas cujo objetivo é achar uma função que, para uma dada entrada *X*, resulte em uma saída *Y*, sendo que os valores corretos de saída são disponibilizados por um supervisor (ALPAYDIN, 2010).

Como exemplos de aplicações de aprendizado supervisionado, Alpaydin (2010) cita os sistemas de avaliação de riscos na concessão de créditos existentes nos bancos em que, a partir dos dados do cliente, é possível calcular se existe um alto ou baixo risco para a concessão de crédito. Outro exemplo é o reconhecimento óptico de caracteres em que, a partir de imagens que representam letras, números ou símbolos, o sistema poderá identificar qual caractere a imagem representa. Um terceiro exemplo é o diagnóstico médico de doenças em que, a partir de um conjunto de informações sobre o paciente e seus sintomas, consegue-se indicar um possível diagnóstico para a doença.

Mohri, Rostamizadeh e Talwalkar (2018) e Russell e Norvig (2010) acrescentam mais exemplos à lista citada, que, de fato, não é exaustiva. Alguns são elencados a seguir:

- **Classificação de documentos ou de textos** – por exemplo, determinar automaticamente se um e-mail é um *spam* ou se algum conteúdo na Web é inapropriado.
- **Processamento de linguagem natural –** como o *part-of-speech tagging* (POS), em que as predições para uma sentença consistem em atribuir identificadores morfológicos e sintáticos para cada palavra nela formada.
- **Veículos autônomos –** quando carros equipados com câmeras, sensores e radares podem alimentar sistemas que executam ações de acelerar, parar e manobrar do carro. Nesse cenário, um carro autônomo obteve a primeira colocação no desafio DARPA, cruzando 132 milhas através do deserto de Mojave, 15 anos atrás.
- **Sistemas autônomos de planejamento e agendamento** – como quando a agência espacial da NASA utilizou o primeiro sistema autônomo embarcado de agendamento e controle de operações de uma aeronave espacial.
- **Jogos eletrônicos –** como quando o computador da IBM *Deep Blue* foi o primeiro a derrotar um campeão mundial em xadrez, Garry Kasparov.
- **Robótica –** a exemplo da empresa *iRobot Corporation*, que já vendeu mais de 2 milhões de robôs (chamados de Roomba) de limpeza para residências.
- **Tradução entre línguas** – como quando um programa de computador que traduz automaticamente de Árabe para Inglês foi criado.

Para introduzir conceitos e mostrar algumas aplicações sobre aprendizado de máquina supervisionado, este capítulo está estruturado da seguinte forma: a segunda seção trata de alguns fundamentos; a terceira seção descreve alguns algoritmos utilizados no aprendizado supervisionado; a quarta seção apresenta exemplos de aplicações do aprendizado supervisionado em situações reais; a quinta seção pontua desafios enfrentados pelos engenheiros e cientistas de dados no uso do aprendizado supervisionado em ambiente corporativo e acadêmico; por fim, a sexta seção relata as considerações dos autores deste capítulo acerca desta abordagem e tecnologia.

2. Fundamentos

Esta seção, primeiramente, introduz o conceito de aprendizado de máquina supervisionado e, em seguida, apresenta o processo no qual ele é normalmente utilizado. Pontua, além disso, os principais componentes envolvidos e quais são as tarefas-padrão pelas quais podem ser agrupados os diversos algoritmos de aprendizado supervisionado.

2.1 O que é Aprendizado de Máquina

O aprendizado de máquina, de forma geral, é uma subárea da Inteligência Artificial cujo objetivo é "a construção de sistemas capazes de adquirir conhecimento de forma automática" (REZENDE, 2005). Um sistema baseado em aprendizado é "um programa de computador que toma decisões com base em experiências acumuladas através da solução bem-sucedida de problemas anteriores" (REZENDE, 2005). Os problemas estariam caracterizados sob a forma de uma série de dados que os discrimina.

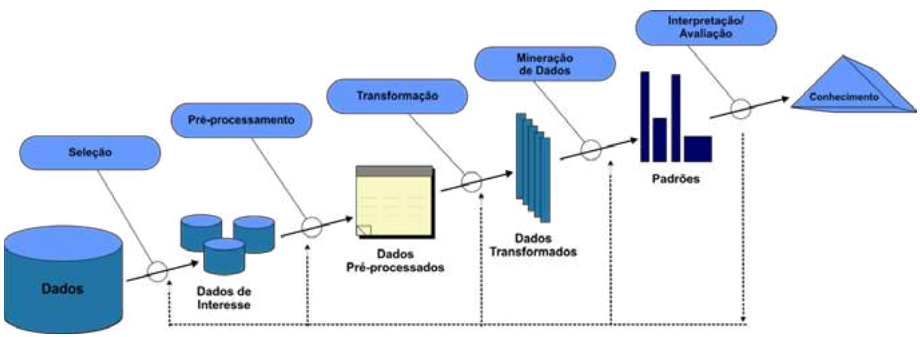
Harrington (2012) comenta que o aprendizado de máquina é uma das formas de realizar a transformação de dados em informação, através do uso de conceitos da interseção entre ciência da computação, engenharias e estatística. As tarefas de aprendizado de máquina incluem dois conjuntos: tarefas de aprendizado supervisionado e tarefas de aprendizado não supervisionado. Métodos chamados de Classificação e Regressão são tipos de aprendizado supervisionado, pois o programador indica a

variável que deve ser alvo de predição do algoritmo. Esses conceitos serão melhor detalhados mais à frente neste capítulo. Em contrapartida, no aprendizado não supervisionado, não há uma variável para ser o alvo de uma predição, e as tarefas consistem em encontrar padrões que descrevem algumas características dos dados (HARRINGTON, 2012).

2.2 Processo de KDD

Fayyad, Piatetsky-Shapiro e Smyth (1996) contextualizam o conceito de aprendizado de máquina inserindo-o dentro de um processo maior chamado *Knowledge Discovery in Databases* (KDD). Os autores definem o KDD como um processo geral usado para extrair conhecimento a partir de um conjunto de dados. Esse processo é estabelecido em algumas etapas que podem ocorrer em sequência, embora seja possível retornar para qualquer uma delas de modo iterativo, como pode ser visto na Figura 1.

Figura 1 | Etapas do processo KDD



Fonte: adaptada de Fayyad, Piatetsky-Shapiro e Smyth (1996)

As etapas do processo KDD são descritas brevemente a seguir (Figura 1):

- **Seleção**: compreende a atividade de criação de um conjunto de dados a ser utilizado, focando-se em um subconjunto de dados de fontes originais, contendo variáveis e amostras dos dados.
- **Pré-processamento**: engloba a realização de tarefas como limpeza nos dados, remoção ou correção de dados incorretos ou ausentes, eliminação de redundâncias, uso de alguma técnica de balanceamento de classes, seleção de atributos, entre outras. As alterações devem ser planejadas levando-se em consideração quais são os dados necessários a serem utilizados no momento de realização do aprendizado.
- **Transformação:** consiste, por exemplo, em converter formatos de dados de acordo com os tipos de dados suportados pelos algoritmos a serem utilizados ou criar atributos derivados que representam o objetivo do problema a ser resolvido.
- **Mineração de Dados**: é nesta etapa que se utilizam os métodos de aprendizado de máquina para a realização de tarefas preditivas (classificação ou regressão) e/ou descritivas (sumarização, agrupamento, associação). Os métodos de aprendizado de máquina supervisionado são apresentados mais adiante neste capítulo.
- **Interpretação/Avaliação:** é a etapa voltada à interpretação dos resultados obtidos com as tarefas de mineração, muitas vezes através de gráficos e técnicas de visualização de resultados. Neste ponto, é provável que aconteça um retorno a quaisquer outras etapas do processo para que se possa corrigir ou melhorar algum passo realizado. Por fim, acontece a divulgação e publicação dos resultados e uma possível incorporação do software resultante em um outro sistema.

2.3 Tarefas-padrão do aprendizado de máquina supervisionado

Os métodos supervisionados utilizam variáveis de saída ou variáveis-alvo, como apontado por Hastie, Tibshirani e Friedman (2008). As variáveis podem ser categóricas ou qualitativas, ou numéricas ou quantitativas. As variáveis categóricas são caracterizadas por um conjunto finito de dados sem uma ordem definida, normalmente chamados de rótulos (*labels*). Por outro lado, as variáveis numéricas ou quantitativas são ca-

racterizadas por um conjunto contínuo ou discreto de dados apresentando uma ordem definida, que podem ser números reais ou até números inteiros inespecíficos.

Essa diferença de variáveis de saída levou a uma separação didática em relação às principais tarefas do aprendizado supervisionado: **Classificação** e **Regressão**. As características de cada tarefa são brevemente introduzidas a seguir:

a. **Classificação:** tarefa que realiza a predição de valores qualitativos ou booleanos a serem atribuídos em categorias predefinidas. Um erro na classificação de um objeto apenas indica que houve um erro; não há um erro maior ou menor. Por exemplo, em uma aplicação que identifica automaticamente um e-mail por SPAM ou NÃO SPAM, essa variável pode ser representada por 0 e 1, respectivamente, ou apenas pelos *labels* SPAM e NÃO SPAM.

b. **Regressão:** tarefa que realiza a predição de valores quantitativos, sendo números inteiros ou reais. É possível medir a "distância" entre o valor correto e o valor que foi alvo de predição do algoritmo. Por exemplo, uma aplicação para estimar o preço de venda de uma casa cujo resultado pode assumir valores bem diversos.

3. Principais algoritmos de aprendizado supervisionado

Fayyad, Piatetsky-Shapiro e Smyth (1996) já comentavam sobre a diversidade de algoritmos de aprendizado supervisionado. Algoritmos comuns nessa linha de aprendizado são: árvores de decisão, métodos baseados em exemplos como o k-NearestNeighbors (vizinho mais próximo), métodos baseados em redes neurais e aqueles baseados em modelos probabilísticos.

De modo mais geral, Harrington (2012) identificou os dez principais algoritmos de aprendizado usados na época e listados em um artigo publicado na IEEE *International Conference on Data Mining*, em 2007. São eles: C4.5 (árvores), k-means, máquinas de vetores de suporte (*Support Vector Machines* – SVM), Apriori, Maximização de Expectativas, PageRank, AdaBoost, k-NearestNeighbors, NaiveBayes e CART. Alguns desses algoritmos são descritos nessa seção.

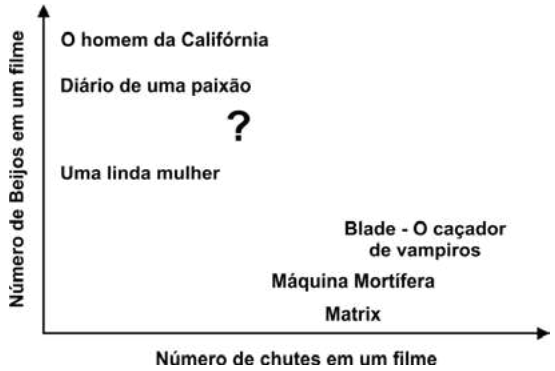
3.1 k-NearestNeighbors

O algoritmo k-NearestNeighbors (k-NN) é considerado o método mais simples de aprendizado supervisionado baseado em instância (MITCHELL, 1997). Por ser um algoritmo de alta precisão e sem suposições sobre os dados, o k-NN é computacionalmente mais caro e normalmente requer mais memória, além de precisar de valores numéricos para calcular a distância/similaridade entre as instâncias (HARRINGTON, 2012).

Todas as instâncias do conjunto de dados possuem rótulos que os enquadram nas suas respectivas classes. Quando um novo dado sem rótulo chega, ele é comparado com todas as instâncias de dados já existentes. Assim, depois de identificadas as instâncias de dados mais semelhantes, ou seja, os vizinhos mais próximos, os seus rótulos são examinados. A partir do conjunto de dados conhecido, são analisados os principais k dados mais parecidos, em que k é um número inteiro, na maioria das vezes, menor que 20. Depois disso, as características são verificadas entre os k dados mais semelhantes, e a maioria será a nova classe da instância de dados solicitada para classificação (HARRINGTON, 2012).

Para melhorar a compreensão deste algoritmo, utilizaremos um exemplo adaptado de Harrington (2012) que classifica filmes em Romance ou Ação, de acordo com a quantidade de chutes e beijos neles contidos. Conforme esses atributos, seis filmes estão representados na Figura 2. Nesse cenário, um novo filme simbolizado pelo ponto de interrogação é o alvo da nova classificação. O número de chutes e de beijos de todos os filmes está listado na Tabela 1.

Figura 2 Representação do número de chutes e beijos em cada filme



Fonte: adaptada de Harrington (2012)

Tabela 1 Classificações dos filmes e as quantidades de chutes e de beijos

Nome do Filme	Nº de chutes	Nº de beijos	Classe
O homem da Califórnia	3	104	Romance
Diário de uma paixão	2	100	Romance
Uma linda mulher	1	81	Romance
Blade – O caçador de vampiros	101	10	Ação
Máquina Mortífera	99	5	Ação
Matrix	98	2	Ação
Filme a ser classificado (?)	18	90	**Desconhecido**

Fonte: adaptada de Harrington (2012)

Primeiramente, é necessário calcular a distância entre o filme a ser classificado e todos os outros existentes no conjunto de dados. Para isso, será utilizada, neste exemplo, a distância euclidiana, a qual calcula a distância entre dois vetores com dois elementos, representados aqui pelos vetores xA e xB, como mostra a Equação 1 (HARRINGTON, 2012):

Equação 1 | Representação da equação da distância euclidiana

$$d = \sqrt{(xA_0 - xB_0)^2 + (xA_1 - xB_1)^2}$$

Fonte: adaptada de Harrington (2012)

Por exemplo, para calcular a distância entre o filme "O homem da Califórnia" e o filme desconhecido, temos os vetores xA = [18,90] e xB = [3,104], em que esses valores são os números de chutes e beijos de cada filme. Dessa forma, aplicando a distância Euclidiana, teremos a Equação 2.

Equação 2 | Aplicação da equação da distância euclidiana

$$d = \sqrt{(18 - 3)^2 + (90 - 104)^2}$$

Fonte: adaptada de Harrington (2012)

Uma vez realizados os cálculos e obtidas as distâncias, conforme a Tabela 2, será necessário descobrir os filmes com as menores distâncias, ou seja, os k-mais próximos, e classificá-los em ordem decrescente. Supondo que k = 3, os três filmes mais próximos são: "O homem da Califórnia", "Uma linda mulher" e "Diário de uma paixão". O k-NN obtém

a maioria dos rótulos desses filmes para definir a classe do filme desconhecido. Assim, como os três filmes são romances, é previsto que o filme desconhecido também seja um romance (HARRINGTON, 2012).

Tabela 2 | Distâncias calculadas entre os filmes classificados e o desconhecido

Nome do filme	Distância para o filme "?"
Diário de uma paixão	18,7
Uma linda mulher	19,2
O homem da Califórnia	20,5
Blade – O caçador de vampiros	115,3
Máquina Mortífera	117,4
Matrix	118,9

Fonte: adaptada de Harrington (2012)

3.2 Árvore de decisão

No aprendizado supervisionado, a árvore de decisão é um algoritmo que pode ser utilizado para tarefas de classificação e regressão. Ele adota um método que divide um problema em problemas menores e possui uma estrutura hierárquica, composta de raiz, nós de decisões internos, nós de decisões e folhas terminais (ALPAYDIN, 2010).

Esse algoritmo aplica uma função de testes baseados no valor dos atributos do objeto recebido como entrada, representando-os em uma estrutura de árvore e produzindo possíveis rótulos que correspondem a uma decisão (RUSSELL; NORVIG, 2010). Uma das vantagens deste algoritmo é que os seres humanos conseguem compreender sua representação mais facilmente; além disso, é geralmente mais barato computacionalmente (HARRINGTON, 2012). Observando essa legibilidade para os

seres humanos, o aprendizado de árvores de decisão pode ser demonstrado através de conjuntos de regras *if-then* (MITCHELL, 1997).

Uma árvore de decisão é representada através de uma função que recebe um vetor com valores de atributos na entrada. Esses valores podem ser discretos, contínuos ou categóricos. Através de uma decisão, é retornado um único valor de saída. O algoritmo executa uma sequência de testes até chegar em uma determinada definição. Esses testes são realizados internamente na árvore através de cada nó e dos ramos do nó. Os ramos equivalem aos rótulos possíveis dos valores do atributo, e o nó refere-se a um valor de um dos atributos de entrada; caso seja um nó folha, indicará um valor retornado pela função (RUSSELL; NORVIG, 2010).

Para ilustrar o aprendizado baseado em árvore de decisão, é utilizada a coleção de dados do Quadro 1, adaptada de Quinlan (1986) e Mitchell (1997), que contém um conjunto de dados denominado de conjunto de treinamento. Para este exemplo, será realizada uma tarefa de classificação que compreende o clima, sendo os atributos descritos e instanciados da seguinte forma:

- Tempo: ensolarado, nublado ou chuvoso;
- Temperatura: quente, fria ou moderada;
- Umidade: alta ou normal;
- Força do vento: forte ou fraca.

Quadro 1 | Conjunto de treinamento

Atributos					**Classe**
Nº	Tempo	Temperatura	Umidade	Força do Vento	Praticar corrida
1	ensolarado	quente	alta	fraca	Não
2	ensolarado	quente	alta	forte	Não
3	nublado	quente	alta	fraca	Sim

Continua

Conclusão

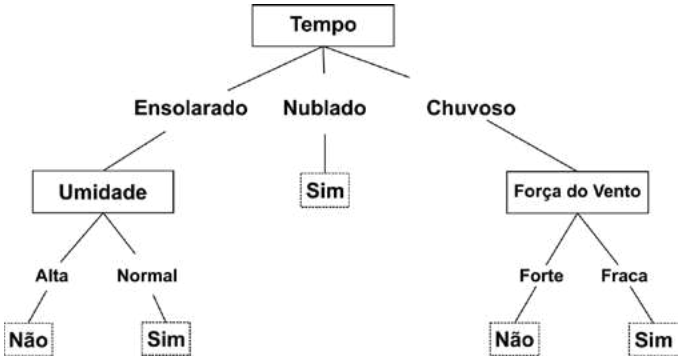
Atributos					Classe
4	chuvoso	moderada	alta	fraca	Sim
5	chuvoso	fria	normal	fraca	Sim
6	chuvoso	fria	normal	forte	Não
7	nublado	fria	normal	forte	Sim
8	ensolarado	moderada	alta	fraca	Não
9	ensolarado	fria	normal	fraca	Sim
10	chuvoso	moderada	normal	fraca	Sim
11	ensolarado	moderada	normal	forte	Sim
12	nublado	moderada	alta	forte	Sim
13	nublado	quente	normal	fraca	Sim
14	chuvoso	moderada	alta	forte	Não

Fonte: adaptado de Quinlan (1986) e Mitchell (1997)

A árvore de decisão realiza a classificação das manhãs de sábado indicando se é pertinente ou não a prática de corrida, conforme a Figura 3, adaptada de Quinlan (1986) e Mitchell (1997). Para simplificar o exemplo, suponha que existem apenas duas classes: Sim – quando for adequado praticar corrida; Não – quando não for adequado realizar essa atividade. Cada atributo refere-se a uma qualidade importante do objeto (MITCHELL, 1997; QUINLAN, 1986). Dessa forma, um novo objeto a ser classificado poderia ter atributos com os seguintes valores:

- Tempo → ensolarado;
- Temperatura → fria;
- Umidade → normal;
- Força do vento → fraca.

Figura 3 Exemplo de árvore de decisão



Fonte: adaptada de Quinlan (1986) e Mitchell (1997)

A classificação inicia-se a partir da raiz, neste caso, Tempo. Os testes são realizados em cada nó da árvore, movimentando-se para baixo. Um teste específico é aplicado no nó, relacionando-o com algum atributo do novo objeto a ser classificado. As ramificações descendentes desse nó representam valores prováveis do atributo. Este procedimento repete-se na subárvore enraizada de cada nó até chegar em um nó folha. O objeto, então, é reconhecido como pertencente à classe "SIM" (MITCHELL, 1997; QUINLAN, 1986), ou seja, "Praticar corrida".

Em uma grande diversidade de problemas, árvores de decisão proporcionam bons resultados; entretanto, em determinadas situações, uma grande árvore poderá ser gerada se nenhum padrão for realmente encontrado (RUSSELL; NORVIG, 2010). A utilização da técnica de poda de árvores de decisão pode remover a excessiva adaptação através da combinação de folhas adjacentes que não possuem grande quantidade de informações relevantes (HARRINGTON, 2012).

O aprendizado baseado em árvore de decisão foi e ainda é utilizado para resolver problemas tais como classificar equipamentos de acordo

com seu mau desempenho, categorizar pacientes médicos de acordo com sua doença ou ainda especificar aspirantes a empréstimos pela probabilidade de não quitar os pagamentos (MITCHELL, 1997). Os algoritmos ID3, ASSISTANT, C4.5 e CART são alguns exemplos de algoritmos para geração de árvores de decisão (HARRINGTON, 2012; MITCHELL, 1997).

3.3 NaiveBayes

O algoritmo NaiveBayes é um método baseado na teoria bayesiana que indica a decisão de acordo com a maior probabilidade. A interpretação dessas probabilidades é chamada de probabilidade bayesiana, em referência ao teólogo do século XVIII, Thomas Bayes (HARRINGTON, 2012).

Baseado no Teorema de Bayes, o algoritmo NaiveBayes estima a classificação de novos objetos através dos cálculos das probabilidades das hipóteses (MITCHELL, 1997). Este modelo, também chamado de classificador bayesiano, é intitulado assim por realizar suposições simplificadas quando as variáveis não são realmente independentes (RUSSELL; NORVIG, 2010).

Para os métodos de aprendizado bayesianos, o Teorema de Bayes é a referência fundamental, permitindo meios para calcular a probabilidade subsequente a partir da probabilidade anterior (MITCHELL, 1997). De forma simplificada, o Teorema de Bayes pode ser representado pela Equação 3 (RASCHKA, 2014):

Equação 3 | Conceito simplificado do Teorema de Bayes

$$probabilidade\ posterior = \frac{probabilidade\ condicional\ x\ probabilidade\ anterior}{evidência}$$

Fonte: adaptada de Raschka (2014)

O objetivo da função é aumentar ao máximo a probabilidade posterior de acordo com os dados de treinamento recebidos e, com isso, desenvolver a regra de decisão (RASCHKA, 2014).

Mesmo sendo um método de suposições simples, o NaiveBayes é eficaz na tarefa de classificação (HARRINGTON, 2012). O algoritmo em pauta pode ser aplicado para, por exemplo, classificar e-mails como *spam*, filtrar postagens maliciosas em um *website*, além de ser bastante utilizado para problemas de classificação de documentos (HARRINGTON, 2012).

O modelo de pensamento bayesiano é importante para o estudo de aprendizado de máquina, pois o classificador NaiveBayes é uma das abordagens mais práticas utilizadas em vários problemas de aprendizado, fornecendo um melhor entendimento para muitos algoritmos que não usam probabilidade explicitamente (MITCHELL, 1997).

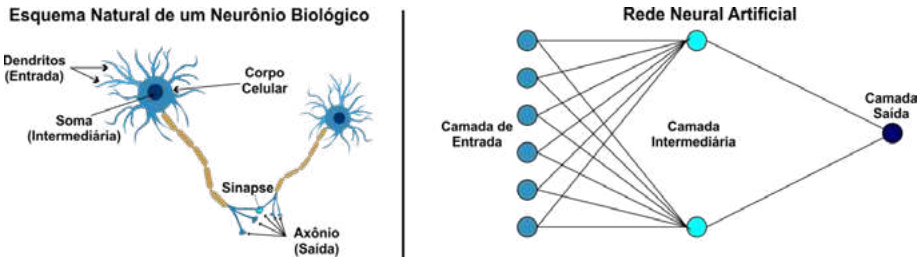
3.4 Redes neurais artificiais

O termo rede neural é utilizado pela motivação de se tentar processar informações de maneira equivalente a um cérebro humano, que o realiza de forma altamente complexa, não linear e em paralelo (HAYKIN, 2007). Dessa forma, as redes neurais artificiais simulam mecanismos de aprendizagem como os organismos biológicos (AGGARWAL, 2018).

Haykin (2007) define uma rede neural artificial como "um processador maciço e paralelamente distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso". Considera-se que uma rede neural se assemelha ao cérebro tanto no conhecimento adquirido pela rede a partir de seu ambiente como quanto à ideia de que as forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

Uma representação de redes neurais é mostrada na Figura 4, adaptada de Manica (2013).

Figura 4 Relação dos neurônios biológicos e artificiais



Fonte: adaptada de Manica (2013)

Na Figura 4, é apresentada, do lado esquerdo, a estrutura de um neurônio biológico, que são células nervosas do sistema nervoso formadas por três partes básicas: os dendritos – prolongamentos do neurônio que garantem os estímulos levando o impulso nervoso em direção ao corpo celular; o axônio – prolongamento que garante a condução do impulso nervoso; e o corpo celular – onde está presente o núcleo da célula e de onde partem os impulsos. A sinapse é a região onde há comunicação entre neurônios, entre neurônios e músculos e entre neurônios e glândulas. A partir de seus dendritos, são produzidos os impulsos elétricos que são conduzidos pelos axônios para os neurônios, estabelecendo uma sinapse. Um cérebro humano é capaz de estabelecer trilhões dessas conexões. Do lado direito da Figura 4, é mostrada uma rede neural artificial, que tenta reproduzir o comportamento e as possibilidades de conexões de uma rede neural biológica.

Um dos grandes benefícios das redes neurais artificiais, além do seu poder computacional, é a capacidade de possibilitar uma programação generalista, como produzir saídas adequadas para entradas que não

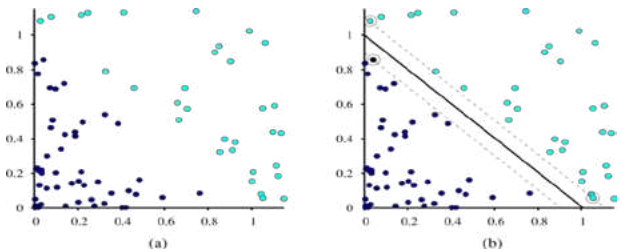
estavam presentes durante o treinamento. Além disso, pode-se aprender através de vários exemplos com determinado comportamento, ou entender e distinguir padrões.

Várias são as aplicações das redes neurais artificiais. No domínio de jogos, um exemplo de utilização de redes neurais profundas é o programa AlphaGo, inspirado no jogo estratégico para tabuleiro Go, muito popular em países asiáticos. Em 2017, foi lançado um documentário sobre o jogo, mostrando que, em 2016, por meio do seu aprendizado de máquina, o programa conseguiu ganhar uma partida do campeão mundial de Go da época por 4 x 1 (DOCUMENTÁRIO..., 2018).

3.5 Máquina de vetores de suporte

O termo máquina de vetores de suporte é a tradução para *Support Vector Machine* (SVM). O algoritmo SVM classifica dados construindo um hiperplano separador para distinguir e identificar dois tipos de classes diferentes, determinando pontos entre dois universos de domínio, normalmente traçando uma linha (ou vetor), diferenciando os dados de ambos os lados (GONZALEZ *et al.*, 2005). Um detalhe importante é que o vetor traçado entre os pontos é realizado de forma simétrica, ou seja, a distância do vetor até os pontos terá o mesmo tamanho, formando as margens como mostrado na Figura 5.

Figura 5 | Exemplo da aplicação do SVM

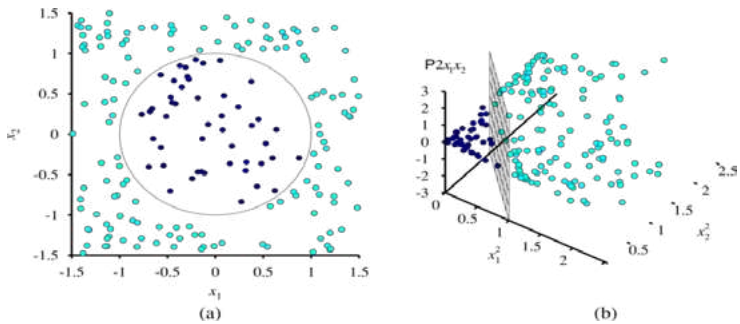


Fonte: adaptada de Russell e Norvig (2010)

Na Figura 5, são demonstrados os passos da aplicação do algoritmo SVM. Do lado (a), as duas classes foram identificadas em pontos com círculos escuros e claros. No lado (b), o separador de margem máxima 1 (linha diagonal) fica no ponto médio da margem (área entre as linhas tracejadas). Os vetores de suporte (pontos com círculos grandes) são os exemplos mais próximos do separador.

Para os casos de dados não lineares, o SVM contém a capacidade de incorporá-los em um espaço de maior dimensão. Com a adição da dimensão, fica facilmente separável, possibilitando a identificação dos pontos desejados (RUSSELL; NORVIG, 2010).

Figura 6 | Aplicação do SVM com adição de uma dimensão



Fonte: Adaptada de Russell e Norvig (2010)

O exemplo da Figura 6, adaptada de Russell e Norvig (2010), mostra como é realizada a criação de uma nova dimensão para possibilitar passar o hiperplano separador. Na Figura 6(a), um treinamento bidimensional é definido com exemplos positivos, como círculos escuros, e exemplos negativos, como círculos claros. Já na 6(b), existem os mesmos dados, porém com a inclusão de um mapeamento em um espaço de entrada tridimensional. Os dados que não podem ser separados no limite de decisão

circular de forma linear são projetados em um plano bidimensional (a) para um plano tridimensional (b).

Devido a sua alta capacidade de generalização, o SVM tem sido empregado com sucesso para resolver problemas de classificação diversos, em alguns deles em combinação com outros algoritmos. Como ilustração, Silva *et al.* (2017) utilizaram um Padrão Binário Local (LBP) – descritor de texturas que realiza uma operação matemática sobre cada pixel, gerando uma nova imagem de padrões binários locais – em conjunto com o SVM. O objetivo foi indicar a presença ou não de Glaucoma, doença óptica que causa dano irreversível ao nervo óptico e é a segunda causa de cegueira no mundo.

4. Exemplos de aplicações

O aprendizado de máquina supervisionado tem ganhado cada vez mais espaço em diversas áreas de aplicação, nas quais se enxergou uma maneira de automatizar tarefas específicas de modo a diminuir o risco e aumentar a eficiência de sua execução. Nessa seção, são descritos exemplos de aplicações desses algoritmos em casos reais. O Quadro 2 mostra um panorama de soluções de aplicações práticas utilizando aprendizado supervisionado.

Quadro 2 – Exemplos de aplicações que utilizam aprendizado supervisionado

Aplicações Práticas Utilizando Aprendizado Supervisionado		
Título do trabalho	**Algoritmo utilizado**	**Solução**
Mineração de dados em triagem de risco de saúde (MACIEL, 2015)	Árvore de decisão	Utiliza classificadores para identificar o risco de vida de pacientes após a aferição de alguns dados

Continua

Conclusão

Título do trabalho	Algoritmo utilizado	Solução
Análise de sentimento em textos escritos na Web (LUNARDI; VITERBO FILHO; BERNARDINI, 2015)	k-NN, Naive Bayes, Árvore de decisão	Analisa textos escritos nas seções de comentários ou em redes sociais para classificá-los em sentimentos positivos ou negativos
Usando o classificador Naive-Bayes para geração de alertas de risco de óbito infantil (SILVA *et al.*, 2017)	NaiveBayes	Utiliza o classificador Naive-Bayes para calcular a probabilidade de uma criança vir a óbito com base nos atributos fornecidos
Uma abordagem utilizando aprendizado de máquina para prever distribuição do carbônico orgânico total (LEE; WOOD; PHRAMPUS, 2019)	k-NN	Utiliza um algoritmo de aprendizado supervisionado para estimar a distribuição do carbono orgânico total no mar baseando-se em propriedades geofísica e geoquímica
Mineração de dados educacionais: uso de redes neurais artificiais na predição do perfil acadêmico do aluno IFAL Campus Maragogi (SILVA; CRUZ, 2019)	Redes neurais	Utiliza redes neurais artificiais como ferramenta na predição de perfis discentes através da análise de dados históricos
Detecção de falhas em redes de sensores sem fio usando SVM (ZIDI; MOULAHI; ALAYA, 2018)	SVM	Utiliza o aprendizado supervisionado através do algoritmo SVM para definir uma função de decisão e detectar sensores irregulares

4.1. Classificação de riscos de pacientes através de árvore de decisão

Maciel *et al.* (2015) apresentam um passo a passo de utilização do processo KDD na área de saúde, através do uso de dados que foram gerados por uma Unidade de Pronto Atendimento (UPA) do Sistema Único de

Saúde (SUS) brasileiro. Os dados são referentes à etapa de triagem de risco de vida, em que os pacientes são recepcionados e seus sinais vitais são mensurados, tais como pressão arterial, frequência cardíaca, temperatura, peso e abertura ocular. O risco pode ser classificado em eletivo, baixo, médio ou alto. O método utilizado é baseado no algoritmo de árvore de decisão, chamado de C4.5[1]. O resultado desse trabalho contribuiu para auxiliar no entendimento de que tipos de características dos pacientes são mais determinantes para a classificação de risco.

4.2. Utilização dos algoritmos k-NN, NaiveBayes e árvore de decisão para análise de sentimentos

Lunardi, Viterbo Filho e Bernardini (2015) realizaram um levantamento de quais são os algoritmos mais usados nos principais trabalhos publicados com a utilização de aprendizado supervisionado na análise de comentários de usuários nas redes sociais, sites sobre filmes e vendas de produtos. Diversos algoritmos foram apontados, como NaiveBayes, SVM, árvores de decisão e k-NN, demonstrando a plena aplicação dessas técnicas para o tipo de solução proposta. Além disso, indicaram diversas aplicações do aprendizado supervisionado em problemas comuns encontrados na área de tecnologia, como sumarização, sistemas de recomendação, direcionamento de marketing e sistemas educacionais.

4.3. Classificação e cálculo de risco de óbito infantil utilizando NaiveBayes

Silva *et al.* (2017) utilizaram o classificador NaiveBayes para calcular a probabilidade de óbito infantil de acordo com atributos da mãe e da criança. Dessa forma, conseguiram prever os casos mais urgentes para priorizar o atendimento. O trabalho reuniu dados das bases públicas Sistema de Informação de Mortalidade (SIM) e Sistema de Informação sobre Nascidos Vivos (SINASC) fornecidos pelo Departamento de Informática do Sistema Único de Saúde (DATASUS). Com a integração dos dados dessas

1 Foi utilizado o software open-source que se chama Waikato Environment for Knowledge Analysis (WEKA).

bases, foi possível diferenciar as crianças que sobreviveram e as que faleceram antes de completarem um ano de idade. Assim, a integração desses dados produziu um *dataset* com vários atributos (idade da mãe, peso ao nascer, local do nascimento, tipo do parto, quantidade de consultas no pré-natal etc.) que foram utilizados para a etapa de análise e testes.

O trabalho realizou experimentos e comparações com outros algoritmos, tais como: k-NN, SVM, redes neurais artificiais, entre outros. A finalidade foi identificar os melhores algoritmos para o escopo da previsão da mortalidade infantil, sendo o classificador NaiveBayes o mais eficiente nesse contexto. Após essa análise, o modelo gerado pelo algoritmo foi aplicado para classificar os riscos de novos pacientes virem a óbito e, em seguida, calcular qual a probabilidade desse fato acontecer.

4.4. Previsão do total de carbono orgânico no fundo do mar global com o algoritmo k-NN

Uma pesquisa desenvolvida por Lee, Wood e Phrampus (2019) utilizou o aprendizado supervisionado aplicando o algoritmo k-NN para realizar uma estimativa da distribuição de carbono no fundo do mar em todo o mundo. Através da observação das quantidades conhecidas ou estimadas de atributos como profundidade da água, distância da costa e temperatura da água no fundo do mar, a abordagem previu semelhanças em pontos do fundo do mar que geograficamente estão distantes, mas muito próximos de acordo com os valores previstos. A utilização do k-NN nesta aplicação conseguiu uma estimativa consistente do total de carbono orgânico, baseando-se em dados geológicos e de fácil atualização em todos os pontos do fundo do mar.

4.5. Utilização de redes neurais artificiais para prever perfil acadêmico do aluno

Um estudo de caso apresentado por Silva e Cruz (2019) utilizou mineração de dados educacionais e implementação de redes neurais artificiais para identificar e predizer os perfis de alunos do Campus Maragogi do Instituto Federal de Alagoas. Com isso, foi possível diminuir o tempo em-

pregado pela instituição para levantar prováveis problemas educacionais e adotar medidas para combater os elevados índices de evasão escolar e reprovação. A pesquisa usou uma abordagem quantitativa e buscou modelos nos dados coletados, o que permitiu elaborar um perfil do aluno por meio da análise, considerando sua performance na vida acadêmica. Como resultados, após a aplicação das redes neurais artificiais, foi possível identificar padrões e predisposições que originariam discentes bem-sucedidos e quais padrões seriam causadores de dificuldades para os alunos.

4.6. Utilização de SVM para detecção de sensores anômalos em redes de sensores sem fio

Zidi, Moulahi e Alaya (2018) aplicaram o aprendizado de máquina supervisionado utilizando a técnica de SVM para identificar falhas em sensores de redes sem fio. Estes dispositivos eletrônicos podem apresentar falhas de hardware (unidade de energia, unidade de detecção etc.), software (problemas nos programas dos sensores) ou de comunicação (devido a falhas no transceptor). Problemas relacionados a falha nos dados também podem ocorrer, causando um mau funcionamento em toda a rede. É possível que esse tipo de falha aconteça de forma simultânea ou separada, além de ocorrer continuamente em um intervalo de tempo ou instantaneamente. Nessa pesquisa, o SVM foi implementado para classificar os dados recebidos do sensor, permitindo a detecção de falhas.

5. Desafios

Os desafios acerca da utilização e evolução dos métodos de aprendizado supervisionado se confundem com os desafios do aprendizado de máquina em geral e são apresentados a seguir através da separação de áreas em que esses desafios estão normalmente se impondo (HALL; PHAN; WHITSON, 2016).

5.1 Desafios organizacionais

A escassez de profissionais capacitados na área continua sendo um dos desafios, devido à necessidade de capacitação demandar um misto de habilidades em ciência da computação, matemática e expertise no domínio do negócio. Associado a isso, é necessário que se realize a ponte entre as demandas dos negócios e as possibilidades das áreas de Tecnologia da Informação das empresas e de seus profissionais. Espera-se que isso alavanque uma mudança cultural da empresa para que decisões em diversas áreas sejam baseadas em dados.

5.2 Desafios com dados

Dados com informações ausentes, conflitantes ou até com erros precisam de análise minuciosa e de correções. Dados muito antigos ou enviesados podem fazer perder o valor dos resultados obtidos. Para que seja possível a utilização plena dos algoritmos de aprendizado supervisionado, grande parte do tempo é gasto com tratamento e pré-processamento dos dados.

Aspectos de segurança da informação e governança devem ser levados em consideração sempre no início de qualquer trabalho envolvendo dados dentro da empresa, para que os conflitos entre a área de ciência de dados e de governança da informação possam ser constantemente gerenciados de forma colaborativa.

Em muitos casos, o maior valor de um processo completo de descoberta de conhecimento a partir de bases de dados acontece quando há integração entre bases de dados distintas. Para isso, entretanto, é preciso realizar tarefas de integração de dados que envolvem seleção de variáveis, alinhamento no tempo e identificação única de entidades.

5.3 Desafios na infraestrutura

As tecnologias de armazenamento de dados vêm avançando na busca de facilitar o gerenciamento de dados não estruturados, semiestruturados e/ou estruturados em tarefas de aprendizado de máquina. Al-

gumas plataformas estão se tornando populares como, por exemplo, o Hadoop, Cassandra, S3 e Redshift. São plataformas voltadas para o processamento de grandes volumes de dados com alta disponibilidade.

O poder de processamento também é um problema a ser cada vez mais enfrentado, pois as tarefas de aprendizado de máquina demandam muito processamento, seja nas etapas de pré-processamento ou de treinamento e teste de modelos. No que diz respeito à infraestrutura, o uso de discos de armazenamento de alta performance tem sido cada vez mais recomendado, assim como arquiteturas especiais ainda sob medida para casos de processamento distribuído.

A utilização de infraestrutura em nuvem tem suprido diversas necessidades devido à sua característica de elasticidade; a infraestrutura é alocada apenas quando demandada pelo usuário, que paga apenas pelo seu uso.

5.4 Desafios na modelagem de aprendizado

A grande diferença entre o aprendizado de máquina e os sistemas de informação baseados em regras tradicionais é que os modelos de aprendizado utilizam muitas variáveis implícitas. A necessidade de interpretação dos resultados dos algoritmos ainda é um problema na adoção da tecnologia, pois a maioria dos algoritmos é do tipo "caixa-preta". As soluções nesse sentido estão adotando abordagens híbridas para encontrar o equilíbrio entre boa acurácia e interpretabilidade dos resultados.

A implantação de técnicas mais modernas de aprendizado de máquina em substituição a sistemas tradicionais é uma decisão desafiadora, pois adiciona complexidade na sua interpretação e documentação e deve ser feita apenas quando o negócio demandar. No caso de modificações em processos de negócio já consolidados, a recomendação é que a implantação de funções preditivas aconteça de forma paralela ou complementar aos processos de negócio já existentes.

5.5 Desafios na operação e produção

Dependendo da realidade da empresa, existe ainda uma grande lacuna entre os ambientes de desenvolvimento, que são geralmente desen-

volvidos em linguagens interpretadas como R ou Python, e o ambiente produtivo, uma vez que a performance exigida, de apenas alguns milissegundos de tempo de resposta, requer ambientes com alta performance. Esses ambientes utilizam, muitas vezes, linguagens como C ou Java, pois vão lidar com grandes volumes de dados que podem, inclusive, ser produzidos em tempo real e não são experimentados em um notebook pessoal.

Quando modelos são avaliados como prontos para estarem em produção, eles vão perdendo sua acurácia à medida que os dados vão crescendo e eles precisam ser retroalimentados e gerados em uma nova versão. É preciso cuidar do processo de monitoramento e gerenciamento de modelos, através do rastreamento de versões, documentação e da evolução do processo decisório dos modelos implantados.

5.6 Desafios na pesquisa

Os desafios na pesquisa envolvendo aprendizado de máquina supervisionado se dão em diversos contextos e áreas de aplicação.

Existem inúmeros desafios nas pesquisas que envolvem o uso de aprendizado de máquina em ambiente de Big Data (ZHOU *et al.*, 2017). Especificamente com aprendizado supervisionado, destacam-se os desafios encontrados na paralelização na execução de algoritmos, com o objetivo de torná-los mais eficientes frente ao grande volume de dados.

Na aplicação de aprendizado supervisionado na área da saúde, por exemplo, um dos maiores desafios é o desenvolvimento de mecanismos de medições de erros, pois as bases de dados da saúde possuem informações que não são corretamente capturadas, por motivos como falhas em instrumentos, respostas humanas a questionários com alto nível de incerteza ou até erros na transformação de dados (JIANG; GRADUS; ROSELLINI, 2020).

De uma forma geral, de acordo com o trabalho apresentado por Patel e Sarvakar (2014), os principais desafios na pesquisa envolvendo algoritmos de classificação são: (i) limpeza de dados, tendo em vista que em muitos conjuntos de dados é necessário realizar, na etapa de pré-processamento dos dados, a identificação, remoção ou alteração de dados incorretos ou ausentes; (ii) seleção de atributos, que refere-se à necessi-

dade de remover as variáveis desnecessárias ou irrelevantes para o modelo – essa ação ajuda a otimizar a performance dos algoritmos, reduzindo o tempo necessário para a etapa de treinamento; (iii) necessidade de normalização para que os dados possam estar em uma faixa menor de dados e abreviações de modo que seja possível resumi-los em conceitos de maior nível de abstração – por exemplo, alterar valores discretos para uma variável apenas com categorias como "baixo", "médio" e "alto" vai fazer com que os algoritmos sejam executados com operações com uma quantidade menor de dados de entrada e saída.

6. Considerações finais

O aprendizado de máquina supervisionado deixou de ser uma promessa e já faz parte da realidade da sociedade, mesmo de quem não o perceba. Consultas realizadas na Internet, pesquisas de itens em *e-commerce*, formulários de satisfação ao cliente, autoatendimentos, eletrodomésticos e os mais diversos serviços virtuais podem estar utilizando técnicas apresentadas neste capítulo para realizar uma tarefa específica.

Nesse contexto, este capítulo apresentou diversos cenários em que o aprendizado supervisionado pode ser aplicado. De acordo com as tendências descritas, será possível ver cada vez mais cenários de uso e uma maior eficácia na utilização dessas técnicas. A tecnologia está ficando cada vez mais popular, e os engenheiros de software e cientistas da computação têm adquirido uma visão ampla da área, entendendo as vantagens que podem ser obtidas com as evoluções que já ocorreram com essa tecnologia. Ao mesmo tempo, esses profissionais enxergam quais avanços ainda estão para acontecer e o que limita o seu uso atualmente.

Nesse sentido, a cautela e o bom senso devem ser levados em consideração nos momentos em que a tecnologia transpuser barreiras morais, podendo colocar em risco a privacidade das pessoas humanas. Vale salientar a importância dessa discussão sobre a ética no uso desses dados, de forma a não criar modelos que continuem a seguir vieses comportamentais presentes em nossa sociedade, como racismo, favorecimentos, injustiças. A interpretabilidade dos modelos deve ser sempre

buscada pela comunidade para entender esses desvios e evitá-los, corrigindo eventuais distorções.

Essas preocupações de forma alguma podem colocar em questionamento o uso da tecnologia, que demonstra ser muito mais benéfica do que qualquer dificuldade encontrada. Assim, espera-se cada vez mais evoluções na área acadêmica e corporativa da tecnologia discutida neste trabalho.

Referências

AGGARWAL, C. C. **Neural Networks and Deep Learning**. Yorktown Heights: Springer, 2018.

ALPAYDIN, E. **Introduction to machine learning**. Cambridge: MIT press, 2010.

DOCUMENTÁRIO sobre inteligência artificial do Google chega ao Netflix. **Estadão,** São Paulo, 3 jan. 2018. Disponível em https://link.estadao.com.br/noticias/cultura-digital,documentario-sobre-computador-do-google-esta-disponivel-na-netflix,70002137222. Acesso em: 18 jul. 2020.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 13, n. 3, p. 37-54, mar. 1996.

GONZALEZ, L.; ANGUL, C.; VELASCO, F.; CATALÀ, A. Unified dual for bi-class SVM approaches. **Pattern Recognition**, v. 38, n. 10, p. 1772-1774, 2005.

HALL, P.; PHAN, W.; WHITSON, K. **The Evolution of Analytics.** 1. ed. Sebastopol: O'Reilly, 2016.

HARRINGTON, P. **Machine Learning in Action.** New York: Manning, 2012.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning:** Data Mining, Inference, and Prediction. 2. ed. New York: Springer, 2008.

HAYKIN, S. **Redes Neurais:** Princípios e Prática. Porto Alegre: Bookman Editora, 2007.

JIANG, T.; GRADUS, J. L.; ROSELLINI, A. J. Supervised machine learning: a brief primer. **Behavior Therapy**, v. 51, n. 5, p. 675-687, 2020.

LEE, T. R.; WOOD, W. T.; PHRAMPUS, B. J. A Machine Learning (kNN) Approach to Predicting Global Seafloor Total Organic Carbon. **Global BiogeochemicalCycles**, v. 33, n. 1, p. 37-46, 2019.

LUNARDI, A. C.; VITERBO FILHO, J.; BERNARDINI, F. C. **Um levantamento do Uso de Algoritmos de Aprendizado Supervisionado em Mineração de Opiniões.** ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 12., Natal. **Anais** [...]. Belo Horizonte: UFMG, 2015. p. 262-269.

MACIEL, T. V.; SEUS, V. R.; MACHADO, K. S.; BORGES, E. N. Mineração de dados em triagem de risco de saúde. **Revista Brasileira de Computação Aplicada**, v. 7, n. 2, p. 26-40, maio 2015.

MANICA, R. Aplicação de uma rede neural artificial simplificada para identificação de gradação de depósitos turbidíticos. **Geociências**, v. 32, n. 3, p. 429-440, 2013.

MITCHELL, T. M. **Machine Learning.** New York: McGraw-Hill Science, 1997.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning.** 2. ed. Massachusetts: MIT Press, 2018.

PATEL, H. G.; SARVAKAR, K. Research Challenges and Comparative Study of Various Classification Technique Using Data Mining. **International Journal of Latest Technology in Engineering, Management & Applied Science,** v. 3, n. 9, p. 170-176, 2014.

PWC. **PwC's Global Artificial Intelligence Study:** Exploiting the AI Revolution. 2017. Disponível em: https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html. Acesso em: 5 abr. 2020.

PWC Brasil. **Tirando a fraude das sombras.** 2018. Disponível em: https://www.pwc.com.br/pt/estudos/assets/2018/gecs-18.pdf. Acesso em: 5 abr. 2020.

QUINLAN, J. R. Induction of decision trees. **Machine Learning,** v. 1, n. 1, p. 81-106, 1986.

RASCHKA, S. **Naive bayes and text classification i-introduction and theory.** arXiv preprint arXiv:1410.5329, 2014.

REZENDE, S. O. **Sistemas inteligentes:** fundamentos e aplicações. 1. ed. Barueri: Manole, 2005.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence:** A Modern Approach. 3. ed. Pearson, 2010.

SILVA, C. L.; BRAGA, O. C.; ANDRADE, L. O. M.; ALVES, J. Q.; PEREIRA JUNIOR, J. W.; OLIVEIRA, A. M. B. Usando o classificador NaiveBayes para a geração de alertas de risco de óbito infantil. **Revista Eletrônica de Sistemas de Informação,** v. 16, n. 2, p. 1-15, 2017.

SILVA, E. H. L.; CRUZ, J. C. Mineração de Dados Educacionais: uso de redes neurais artificiais na predição do Perfil Acadêmico do Aluno. ESCOLA REGIONAL DE COMPUTAÇÃO BAHIA, ALAGOAS E SERGIPE, 19., 2019, Ilhéus. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019. p. 556-564.

ZHOW, L.; PAN, S.; WANG, J.; VASILAKOS, A. V. Machine learning on big data: Opportunities and challenges. **Neurocomputing**, v. 237, p. 350-361, 2017.

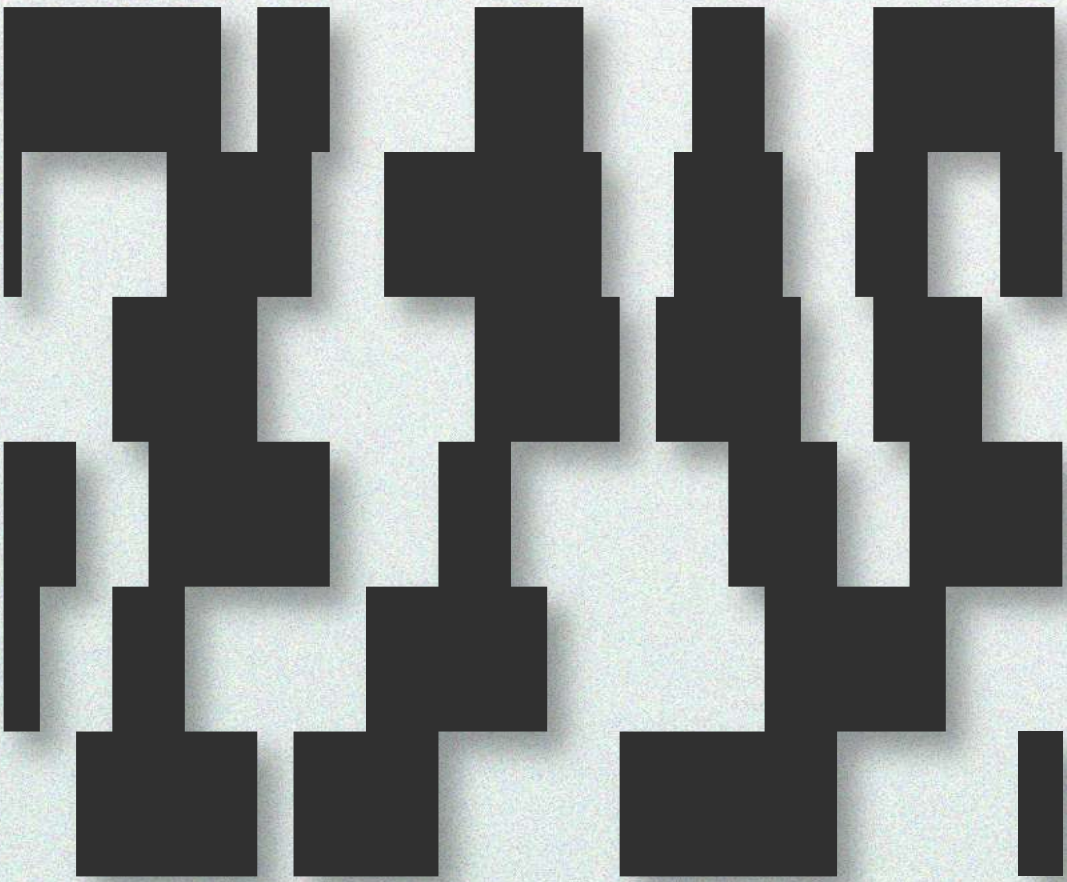
ZIDI, S.; MOULAHI, T.; ALAYA, B. Fault Detection in Wireless Sensor Networks-Through SVM Classifier. **IEEE Sensors Journal**, v. 18, n. 1, p. 340-347, 2018.

Capítulo 2

Aprendizado de Máquina Não Supervisionado:

da Teoria à Aplicabilidade Utilizando Agrupamento

Joerverson Barbosa Santos
Rafael Anderson de Lima Ramos



1. Introdução

A Tecnologia da Informação (TI) vem despontando como uma área de grande importância para diversos setores da sociedade, impactando diretamente nas atividades acadêmicas, comerciais e servindo de molde para a construção de um novo escopo social (RAMOS, 2015).

Desde o surgimento dos primeiros computadores, o homem busca entender como programá-los para que possam "aprender" e melhorar automaticamente com suas próprias experiências ou dados inferidos. Nesse contexto, segundo Mitchell (1997), o processo de aprendizado de máquina é definido formalmente por um sistema computacional que busca realizar uma tarefa T, aprendendo a partir de uma experiência E, procurando melhorar uma performance P.

Ou seja, a partir de um grande volume de dados, um algoritmo pode ter a capacidade de aprender a atingir seus objetivos, constituindo assim experiências focadas em tarefas específicas que tendem a melhorar sua performance perante um maior número de exemplos obtidos.

O Aprendizado de Máquina, do inglês *Machine Learning*, é um dos principais pilares dessa nova era da Indústria, pois permite a extração de informação utilizando dados de forma eficiente e eficaz (FREITAS; SANTANA JUNIOR, 2019). Dessa forma, a eficiência está diretamente associada aos dispositivos de baixo custo voltados para computação de alto desempenho. A eficácia relaciona-se à qualidade dos dados disponíveis e dos modelos de aprendizado obtidos.

No contexto computacional, existem muitos algoritmos eficazes para determinados tipos de aprendizado, entre os quais destacam-se aqueles que têm seu enfoque no aprendizado não supervisionado (LOPEZ, 2010). Neste tipo de aprendizado, a informação dos rótulos históricos é inexistente, ou seja, não se tem as informações das saídas desejadas a serem estimadas e, por esse motivo, dizemos que os dados são não rotulados (ESCOVEDO; KOSHIYAMA, 2020). Sendo assim, o algoritmo não recebe, ao longo do seu treinamento, os resultados esperados; deve descobri-los por si só, por meio da exploração dos dados, os possíveis relacionamentos entre eles. Perante tal cenário, uma das tarefas de

aprendizado não supervisionado visa identificar similaridades nos dados analisados e agrupá-los de acordo com estas. Essa tarefa é chamada de agrupamento ou clusterização.

Segundo Lopez (2010), quando o objetivo é descobrir atividades úteis e desejadas através de tentativa-e-erro e de processos auto-organizáveis, estabelecendo relações de semelhança entre itens, elementos ou perfis, então tem-se uma abordagem prática de agrupamento. Assim, uma das aplicações desse modelo não supervisionado busca categorizar uma dada entrada, tentando responder se ela pertence a um grupo já existente ou se será necessário criar um novo grupo para incluí-la (HAYKIN, 2007).

Vale ressaltar que o aprendizado não supervisionado apresenta algumas tarefas possíveis de serem empregadas além do agrupamento, como, por exemplo, as regras de associação (AGRAWAL; IMIELINSKI; SWAMI, 1993). Este capítulo tem como foco a tarefa de Agrupamento ou Clusterização, que objetiva agrupar as instâncias de dados de interesse ou separar os registros de um conjunto de dados em subconjuntos, fazendo com que tais instâncias contidas em um respectivo grupo possuam alta similaridade (CARVALHO, 2014).

O aprendizado de máquina não supervisionado baseado em agrupamento possui desafios considerados, muitas vezes, mais complexos do que aqueles de outros modelos conhecidos. Tal afirmação é concebida tendo como justificativa o fato de trabalhar-se com dados não rotulados, ou seja, sem a convicção de quantas ou de quais classes existem. Entretanto, mesmo diante dessa dificuldade, sistemas de recomendação de filmes ou músicas, detecção de anomalias e visualização de dados são exemplos de algoritmos muitas vezes baseados nos princípios que permeiam o contexto do aprendizado não supervisionado por agrupamento (HONDA; FACURE; YAOHAO, 2017).

Dessa maneira, inúmeros *softwares* são desenvolvidos utilizando os princípios úteis de algoritmos dessa categoria de aprendizado. Isso faz com que importantes aplicações comerciais modelem suas arquiteturas nas concepções que permeiam o aprendizado de máquina não supervisionado, a exemplo de sistemas de comércio eletrônico e aplicações de análise de crédito.

Considerando a vasta gama de aplicações que utilizam métodos não supervisionados baseados em agrupamento, esta tarefa é o tema central abordado neste capítulo. Assim, este capítulo tem por objetivo prover uma introdução aos princípios que permeiam o aprendizado de máquina não supervisionado, procurando apresentar especificamente técnicas de agrupamento. Para isso, a Seção 2 traz uma introdução aos fundamentos teóricos inerentes ao tema, além de abordar alguns métodos de agrupamento. Na Seção 3, o foco está direcionado para técnicas de agrupamento, demonstrando conceitos, exemplos e trabalhos relacionados. Na Seção 4, explicita-se de forma mais detalhada as diferentes maneiras possíveis de representar-se um *cluster*. A Seção 5 apresenta exemplos de algoritmos aplicados a agrupamento. A Seção 6 descreve os principais desafios ao lidar com o presente tema. Por fim, na Seção 7, são abordadas as considerações finais.

2. Fundamentação teórica

Esta seção introduz os principais conceitos e princípios que compõem e descrevem o aprendizado de máquina não supervisionado, no contexto da tarefa de agrupamento.

2.1. Processo de KDD

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o processo de descoberta de conhecimento, do inglês *Knowledge Discovery in Databases* (KDD), é um conjunto de passos que têm por objetivo a revelação de conhecimento útil a partir de dados. O KDD é um processo não trivial que objetiva identificar padrões válidos, novos, potencialmente úteis e compreensíveis nos dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH,1996). O termo "não trivial" é, na verdade, uma referência à execução de uma série de etapas complexas, que visam a alcançar a identificação de padrões que sejam "úteis" e de fácil compreensão em um contexto de análise de dados (MOURA, 2018).

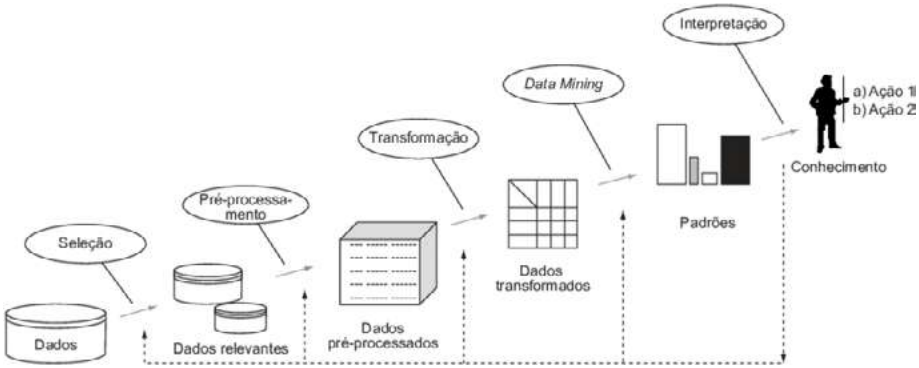
O processo de KDD é constituído de etapas, que são executadas de forma interativa e iterativa. Interativa porque envolvem um conjunto de

atividades em que o resultado de uma etapa depende da outra. Iterativa porque esse processo não é executado de forma sequencial; envolve repetidas seleções de parâmetros e conjunto de dados, possibilitando, inclusive, que a pessoa responsável pela análise dos dados interfira nas demais atividades, orientando a execução do processo. Dessa forma, cada etapa pode ser repetida uma ou mais vezes (MOURA, 2018), dependendo da necessidade de rever ou refinar dados ou resultados obtidos.

As etapas do processo de KDD, mencionadas por Fayyad, Piatetsky-Shapiro e Smyth (1996), podem ser divididas em:

1. Seleção – etapa de coleta e organização dos dados;
2. Pré-processamento – os dados são adequados ou corrigidos, de modo que, ao final desta etapa, eles estejam em formato correto, sem duplicidade, limpos e consistentes;
3. Transformação – etapa que se dedica à conversão de dados ou formatos, caso seja necessária, e ao seu armazenamento, a fim de facilitar a aplicação de técnicas de mineração de dados;
4. Mineração de dados – atividade dedicada à busca pelo conhecimento, a partir do uso de algoritmos para descoberta de padrões;
5. Interpretação e avaliação – etapa que tem por objetivo interpretar os dados obtidos a partir das análises realizadas e verificar sua relevância para o problema proposto.

Figura 1 Etapas do processo de KDD



Fonte: adaptada de Fayyad, Piatetsky-Shapiro e Smyth (1996 apud STEINER *et al.*, 2006)

Segundo Steiner *et al.* (2006), no contexto geral, o KDD refere-se a todo processo de descoberta do conhecimento útil em uma base de dados, sendo a mineração de dados o principal estágio do KDD, referindo-se à utilização das técnicas de aprendizado de máquina visando a extração dos modelos de conhecimento (Figura 1).

2.2. Aprendizado de máquina

O ser humano detém a capacidade de adquirir conhecimento através de experiências vividas. Norteando-se por este princípio, a tentativa de simular tal característica humana em computadores é chamada de aprendizado de máquina. Esta área busca desenvolver métodos de modo a permitir que computadores aprendam com experiências passadas (BISHOP, 2006).

As técnicas de aprendizado de máquina podem ser capazes de gerar modelos que consigam, de forma automática, reconhecer ou imitar formas humanas de se comportar (PORTO FILHO, 2017). De maneira

geral, tais técnicas de aprendizado podem ser divididas entre aprendizado supervisionado e aprendizado não supervisionado (WITTEN; FRANK; HALL, 2011).

No aprendizado supervisionado objetiva-se construir um modelo que possa ser utilizado para predizer qual será a classe/saída para novas instâncias ou registros (WITTEN; FRANK; HALL, 2011). Já para a categoria não supervisionada, não se utilizam informações das variáveis de saída, fazendo com que dados de entrada passem por uma análise e sejam agrupados conforme a proximidade dos seus valores (ROZA, 2016).

2.3. Aprendizado de máquina não supervisionado

Ao operar dados não rotulados, sem se ter a informação de quantos ou de quais grupos realmente existem, observa-se claramente a aplicação prática do aprendizado de máquina não supervisionado com base em agrupamento. O objetivo dessa técnica é classificar as entidades em grupos mutuamente exclusivos baseando-se na similaridade das instâncias (PORTO FILHO, 2017).

Essa estratégia é utilizada para encontrar subgrupos de objetos, sendo que tais subgrupos não são predefinidos e nem criados baseados nas respostas de uma programação realizada por um operador. O agrupamento identifica semelhanças nos dados e reage mediante a presença ou ausência de tais semelhanças em cada novo dado apresentado; daí o nome de tarefa não supervisionada.

Assim, a ideia do agrupamento é encontrar *clusters* (grupos) no conjunto de dados, fazendo com que os itens de um mesmo *cluster* sejam os mais parecidos possíveis, enquanto os itens dos demais *clusters* sejam os mais diferentes possíveis. A semelhança ou diferença entre os dados é obtida de acordo com algoritmos de similaridade adotados (MITCHELL, 1997).

2.4. Agrupamento

Como foi citado anteriormente, o aprendizado de máquina não supervisionado é focado na análise de dados que não possuem um determinado rótulo ou agrupamento predeterminado. O agrupamento consis-

te na separação de dados em grupos (*clusters*). Linden (2009) diz que o agrupamento pode ser o nome dado ao conjunto de técnicas computacionais cujo propósito consiste em separar os elementos em grupos. Tais agrupamentos baseiam-se nas características dos dados e na similaridade entre estes, formando, assim, grupos mutuamente exclusivos.

Analogamente, esta técnica é utilizada também pelos seres humanos, quando, por exemplo, um médico estuda algum fenômeno que intercorre entre um grupo de pessoas, observa suas características e o descreve como uma doença X. Com base nesse conhecimento adquirido e por meio de uma massa de dados (pessoas que possuem os mesmos sintomas), consegue-se categorizar e descrever um catálogo de doenças, gerando novos conhecimentos a partir dessas informações, sendo que um destes conhecimentos pode ser a pesquisa de uma cura.

Com essa premissa, o aprendizado de máquina não supervisionado veio auxiliar o homem na extração de conhecimento em grandes massas de dados. Como apresentam Henning *et al.* (2016), o conhecimento vem crescendo, e uma demonstração deste crescimento é observada na organização e no agrupamento de tudo que foi aprendido, como é o caso do conhecimento da área biológica, em que toda organização acontece em *clusters* hierárquicos. Astrônomos agrupam planetas e galáxias por seus formatos, suas cores e pela curvatura da luz entre eles; empresas analisam seus produtos e usuários baseando-se em seus comportamentos comuns.

Para que um *cluster* seja definido, tal decisão depende do problema e/ou da área de conhecimento que se deseja atingir. Milligan (1996) diz que, primeiramente, é preciso escolher os dados que serão usados. Existem diversos tipos de dados que podem ser utilizados e nem sempre todas as informações em uma massa de dados serão importantes para o conhecimento que se deseja obter.

Outro ponto que vale ressaltar é que dentro da massa de dados é possível encontrar representações de dados de difícil interpretação do ponto de vista computacional, como objetos complexos. Sabendo disso, faz-se necessária uma etapa de ajuste e conformidade dos dados, regulando-os para que sejam legíveis para a máquina. A Seção 2.5 discorre mais sobre o assunto.

Quando se inicia o trabalho com agrupamento, pode-se pensar que quanto mais *clusters* melhor. Tal afirmação, porém, apresenta um risco, pois uma granularização tão numerosa implica uma grande quantidade de *clusters*, que serão, consequentemente, cada vez menores. Isso torna mais difícil de obter conhecimentos, uma vez que há poucas evidências sobre dado fenômeno. Por outro lado, a escolha de poucos *clusters* pode acarretar a impossibilidade de se extrair o conhecimento, já que existe pouco agrupamento.

A definição de quantos *clusters* seriam ideais para uma determinada aplicação pode variar muito com base nos dados e no fenômeno que se deseja estudar. Uma avaliação sobre os dados e o domínio que se deseja estudar deve ser feita; com isso, será possível ter uma noção da quantidade de *clusters* necessários para sua aplicação. Como Hennig e Liao (2013) afirmam, o processo de selecionar um método de identificação do número de *clusters* e a interpretação de seu resultado está diretamente ligado ao entendimento da representação dos dados, sendo importante que a quantidade de *clusters* apropriada seja utilizada.

Existem algumas técnicas que permitem identificar a quantidade ideal de *clusters* que uma aplicação deve ter de uma forma mais genérica, levando em consideração o tamanho do conjunto de dados. Alguns exemplos de algoritmos que auxiliam nessa atividade são o Elbow e o AverageSilhouette (ALVES, 2018). Seus algoritmos são apresentados posteriormente.

2.5. Análise de objetos para uso em *clusters*

Segundo Gordon (1981), existem diversas formas de entradas de dados, tanto dados simples quanto mais complexos. Tais dados podem representar várias informações e possuir diferentes tipos, como (GORDON, 1981):

- vetor de diferentes tipos de dados (incluindo objetos complexos);
- atributos com coordenadas geográficas;
- dados numéricos, como umidade, temperatura, distância;

- dados categóricos (podem assumir um valor dentro de um conjunto de informações predefinidas), como patentes militares (soldados, capitão, cabo), usuários num sistema (administrador, cliente, operário, motorista), valores booleanos que são do tipo se alguém está doente (verdadeiro ou falso).

Tendo em vista a possibilidade de o quantitativo de objetos no conjunto de dados influenciar no processamento dos *clusters*, é necessário saber quais dados são realmente importantes para extraí-los. Linden (2009) comenta que um conjunto de dados extenso é muitas vezes inútil e faz com que o nível de complexidade no processamento dos *clusters* entre em ordem exponencial. Por exemplo, caso se deseje fazer a execução da tarefa de agrupamento definindo 6 *clusters* e possuindo 200 informações a serem processadas e organizadas, existem $6^{200} = 4{,}268^{e+155}$ possíveis comparações – lembrando que cada dado pode ser comparado entre si, em busca da similaridade; ao final, são organizados 6 grandes grupos contendo dados similares. Dessa forma, caso fosse utilizado um computador que consegue calcular 109 dados por segundos, seriam necessários 1.148 anos para finalizar, pois são mais de 154 ordens de grandeza. Por isso a necessidade de filtrar os dados e buscar uma heurística que seja eficiente para o conjunto de dados a ser analisado.

3. Classificação de métodos de agrupamento

Há diferentes abordagens para a análise e o processamento de *clusters*, que se diferem, em suma, pelo modo como os *clusters* se relacionam no momento do agrupamento. A divisão primária clássica de métodos de agrupamento envolve os métodos hierárquicos e os não hierárquicos (JAIN; DUBES, 1988), a serem apresentados com mais detalhes a seguir.

3.1. Métodos não hierárquicos (partitivos)

O método não hierárquico funciona no formato baseado em um único *cluster*, que contém a quantidade total de informações inicialmente conhecidas; no decorrer do processamento, o *cluster* vai sendo dividido em *clusters* de tamanhos iguais sem sobreposição. O particionamento aconte-

ce com base em seus centroides, que funcionam como um ponto central imaginário ou real, no qual o *cluster* é formado, ou seja, quando ocorre a divisão com base em suas similaridades, os dados que são similares agrupam-se em um ponto, ao qual chamamos de centroides. Vale ressaltar que o particionamento é executado até que a quantidade de *clusters* se iguale à quantidade ideal definida anteriormente para o processamento.

Fávero *et al.* (2009) comentam que os procedimentos não hierárquicos são métodos que possuem como objetivo encontrar diretamente uma separação em *n* elementos dentro de K *clusters*, de modo a satisfazer dois requisitos básicos, como similaridade entre os dados e isolamento dos *clusters* formados.

Fávero *et al.* (2009) destacam que os métodos não hierárquicos fornecem uma série de soluções para diferentes agrupamentos de dados. Além da diversidade de aplicações, comentam que a probabilidade de ocorrerem agrupamentos errados é menor em métodos não hierárquicos, mas há uma certa dificuldade em encontrar o número de *clusters* de partida. Pode-se definir o funcionamento do método não hierárquico (partitivo) da seguinte forma:

1. Faz-se uma divisão inicial dos elementos em K *clusters*, definidos pelo analista, com base nos dados que deseja analisar;
2. Executa-se o cálculo a partir da partição do *cluster* inicial. O cálculo é, por exemplo, por meio da distância euclidiana dos centroides de cada elemento na base de dados;
3. Agrupam-se os elementos aos *clusters* que os centroides estão próximos;
4. Executa-se a atualização dos centroides, possibilitando as novas partições. Volta-se para o segundo passo, realizando o cálculo com base nos novos *clusters* criados. Essa ação ocorre até que não haja uma variação tão significativa nas distâncias mínimas de cada elemento da base de dados a cada um dos centroides dos K *clusters*. Também pode ser imposta uma condição a cargo do analista para que a separação seja dada como satisfeita.

3.2. Métodos hierárquicos

O método hierárquico consiste em realizar uma sequência de partições aninhadas objetivando criar grupos e subgrupos (ROKACH; MAIMON 2005). Existem duas categorias desse método: o agrupamento por aglomeração e o agrupamento por divisão.

3.2.1. Agrupamento por aglomeração

Linden (2009) explica que o agrupamento por aglomeração consiste na criação dos *clusters* com base em dados isolados. A distância entre os dados possui importância para o funcionamento do método, pois, mediante tal distância, é possível definir qual o raio de busca, considerando que o método irá agrupar os dados nos *clusters* baseando-se nas similaridades. O comportamento desse tipo de método funciona da seguinte forma:

1. Gera-se um *cluster* para cada dado;
2. Encontram-se os *clusters* mais similares, de acordo com a medida de distância definida;
3. Os *clusters* encontrados juntam-se em um novo *cluster*, e sua distância para todos os outros dados é recalculada;
4. Os passos 2 e 3 são repetidos até sobrar apenas um *cluster*;
5. Quando houver apenas um *cluster* ou ocorrer uma condição de parada, o procedimento é finalizado.

3.2.2. Agrupamento por divisão

O agrupamento por divisão é mais complexo que o agrupamento por aglomeração. Esse método possui duas vantagens principais:

- não requer que as distâncias sejam recalculadas em cada iteração; além disso, pode-se interromper o procedimento antes de chegar no último nó da árvore, o que melhora e muito a sua performance em comparação ao agrupamento por aglomeração;

- como esse algoritmo começa já possuindo todos os dados – diferente do que acontece no agrupamento por aglomeração –, as informações encontram-se mais fiéis quanto à sua real distribuição.

Para que se tenha o método de agrupamento por divisão, não necessariamente o conjunto de dados deve ser dividido ao meio, diversamente do agrupamento por aglomeração. De acordo com Linden (2009), o método por divisão utiliza duas métricas de cortes fundamentais, denominadas de similaridades *intracluster* (que é a distância dos elementos dentro de um *cluster*) e similaridade *extracluster* (que é a proximidade dos dados fora de um *cluster*), buscando aumentar a relação entre as duas, a fim de obter o particionamento dos *clusters* com maior coesão possível.

A execução do método de agrupamento por divisão ocorre da seguinte forma (ROKACH; MAIMON 2005):

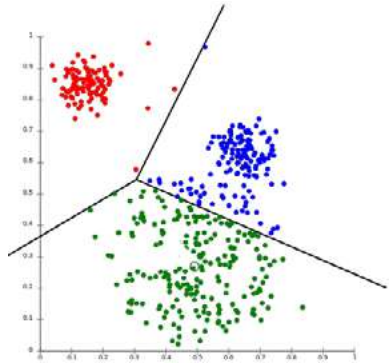
1. Todos os dados são iniciados com um único *cluster*;
2. Calcula-se a similaridade e efetua-se a divisão com base no *intracluster* e no *extracluster,* possuindo, assim, subdivisões do *cluster* inicial;
3. A etapa 2 é repetida até que se alcance a estrutura de *clusters* desejada.

O interessante é que essa estratégia não apenas divide um conjunto de dados inicial que estava na fila ao meio, como pode particioná-lo em tamanhos distintos com base na informação de proximidade dos dados, dependendo do domínio do problema aplicado para tal método.

4. Representação de *clusters*

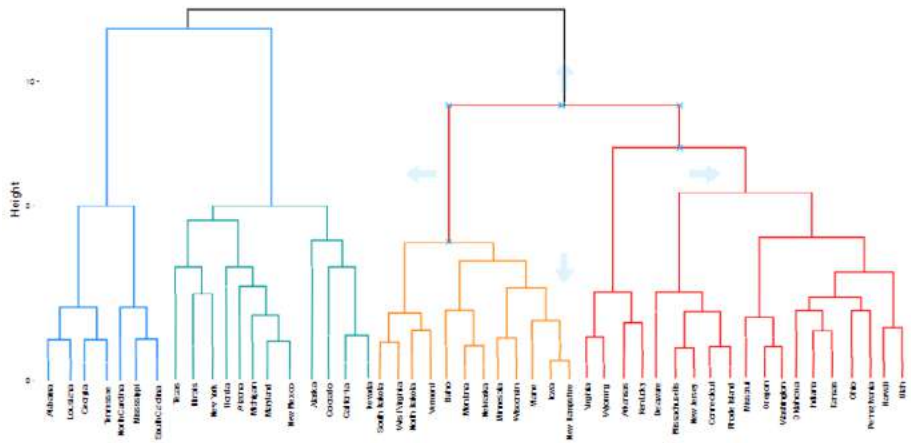
Para que se possa observar o resultado do agrupamento de maneira objetiva, é possível representar os dados extraídos pelos métodos de agrupamento de diversas formas. No geral, no entanto, as representações dos resultados gerados podem ser definidas com base no método utilizado, hierárquico ou não hierárquico. São exemplos de tais representações: o gráfico apresentado na Figura 2, representando grupos de dados não hierárquicos, e o dendograma (Figura 3), que constitui uma forma de representação visual de *clusters* hierárquicos.

Figura 2 Exemplo de representação de dados de *clusters* não hierárquicos



Fonte: adaptada de Porto Filho (2017)

Figura 3 Exemplo de dendograma



Fonte: adaptada de Porto Filho (2017)

5. Algoritmos para agrupamento

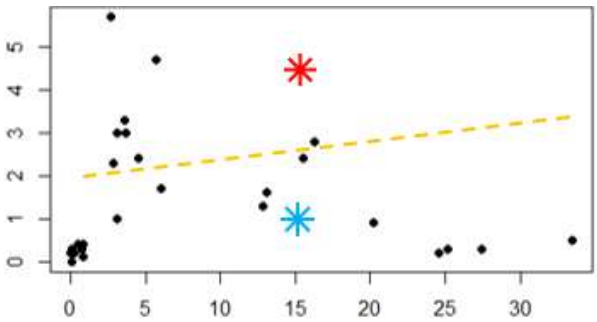
Nesta seção são apresentados alguns dos algoritmos mais utilizados para geração de *clusters*, tanto para métodos hierárquicos quanto para não hierárquicos.

5.1 K-means

Esse é um dos algoritmos mais conhecidos, sendo considerado de rápida e de fácil utilização para gerar *clusters*. Ele faz a separação dos dados em *k* (número predefinido) *clusters* com base na distância dos pontos até chegar nos centros. Ou seja, ao ser definida a quantidade *k*, as massas de dados serão organizadas a partir desse parâmetro e acopladas dentro desses *k clusters* (ALVES, 2018).

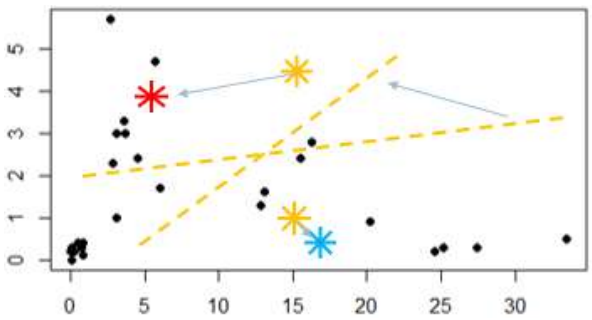
O algoritmo k-means só finaliza seus ciclos quando não existe mais alteração significativa nas distâncias dos centroides – isto é, quando a distância dos centroides está muito próxima uma da outra – ou até alcançar alguma condição de parada. A Figura 4 ilustra a etapa inicial de atuação do k-means, isto é, o início da separação, quando os dados ainda estão bem dispersos. Na Figura 5, percebe-se como estão sendo definidos os grupos, executando a validação das medidas de distância; tal execução prossegue até alcançar uma condição de parada. Por fim, na Figura 6, pode-se observar a etapa de agrupamento em seu estado final, em que a separação dos *clusters* foi finalizada.

Figura 4 Representação do algoritmo k-means no estágio 1



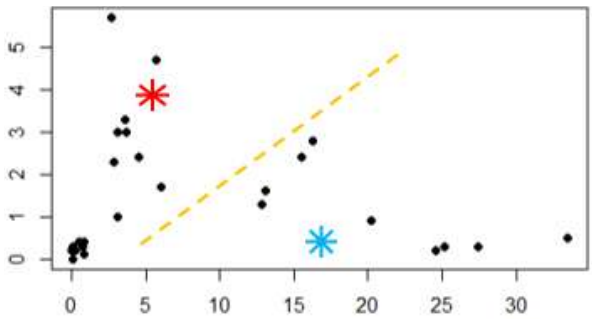
Fonte: Alves (2018)

Figura 5 Representação do algoritmo k-means no estágio 2



Fonte: Alves (2018)

Figura 6 Representação do algoritmo k-means no estágio final



Fonte: Alves (2018)

Como existe a possibilidade de definir quantos *clusters* podem ser utilizados, ao aplicar este algoritmo, é comum deparar-se com o dilema sobre essa quantidade ideal. Como comentado em seções anteriores, é importante definir a quantidade de *clusters* a serem trabalhados, pois cada domínio de conhecimento possui suas peculiaridades e massa de dados. Vale frisar que existem algoritmos para descobrir o número ideal de *clusters* a serem utilizados; alguns destes são citados nos tópicos seguintes.

5.1.1. Algoritmo Elbow (Método do Cotovelo)

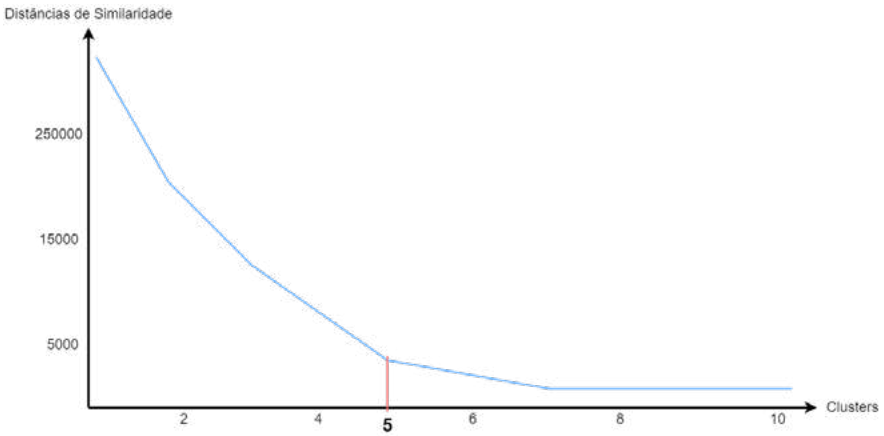
Esse método consiste em executar o algoritmo com um valor aleatório de *clusters* a fim de realizar testes na curva de deslocamento dos dados e, possivelmente, encontrar o número de *clusters* ideal a ser utilizado (ALVES, 2018).

Por exemplo, ao escolher um número de 1 a 15, pode-se, em seguida, calcular a função de custo a partir da soma dos quadrados da distância interna dos *clusters*. Quando posto em um gráfico, pode-se observar

que sua representação começará bem elevada e irá convergindo a 0 na soma dos quadrados das distâncias.

Com essa informação, deve-se procurar o "cotovelo" do gráfico, ou seja, o ponto em que sua curva começa a ficar mais próxima de 0; logo, este é o número ideal de *cluster*s. Observa-se, na Figura 7, que existem de 0 a 10 possíveis *clusters* a serem processados pelo algoritmo. O "cotovelo" é identificado no ponto (5,5000). O número ideal de *clusters* é, portanto, 5 (Figura 7).

Figura 7 | Exemplo de representação do gráfico Elbow



Fonte: adaptada de Alves (2018)

5.1.2. AverageSilhouette

Esse método faz a medição do nível de similaridade do dado dentro de um determinado *cluster* (ALVES, 2018). A técnica busca validar, entre os *clusters* vizinhos, se os dados neles alocados realmente os pertencem ou se tem uma certa parcialidade da posição do dado dentro do *cluster*.

Assim, quanto mais o coeficiente do *silhouette* estiver próximo de 1, mais distante de outro *cluster* o dado está; quanto mais próximo de 0, mais próximos uns dos outros os *clusters* estão, ou estão se interseccionando.

Para calcular o coeficiente de *silhouette*, precisamos fazer a média entre os pontos dentro do *cluster* a(i), definir a distância média do *cluster* mais próximo b(i) e, por fim, dividir o valor da média pelos valores dos dados mais distantes dentro de cada *cluster*, sendo max(b(i), a(i)). Nesse cenário, é realizado o seguinte cálculo do coeficiente (Equação 1):

$$s(i) = (b(i) - a(i)) / max(b(i), a(i)) \quad (1)$$

Para obter a quantidade de *clusters* que seria interessante utilizar, o algoritmo busca usar a média do *score* gerado pelo *silhouette* com todos os dados do *dataset*, como exemplificado na Figura 8. Com o *score,* consegue-se ter visibilidade do espectro de proximidade entre os *clusters*, possibilitando ao analista uma melhor compreensão da massa de dados e de como utilizá-la no seu aprendizado de máquina.

Figura 8 | Resultado da execução do método AverageSilhouette

```
Para n_clusters = 2 O score_silhouette médio é: 0.24604273339845253
Para n_clusters = 3 O score_silhouette médio é : 0.20670607133321856
Para n_clusters = 4 O score_silhouette médio é : 0.188444914764597
Para n_clusters = 5 O score_silhouette médio é : 0.19090629903451375
Para n_clusters = 6 O score_silhouette médio é : 0.18047082769472864
```

Fonte: Alves (2018)

5.2. Bisecting k-means

O algoritmo Bisecting k-means se baseia no k-means, explicado anteriormente. Consiste numa variação hierárquica do k-means, em que,

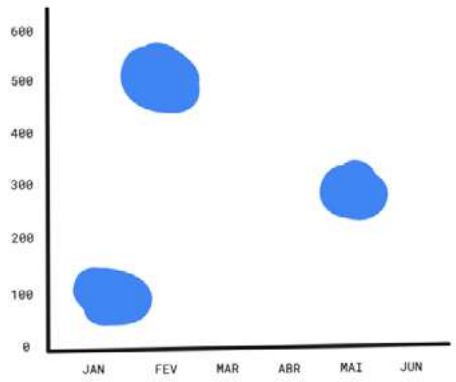
a cada iteração com a base de dados, realiza-se uma divisão desta, seguindo a versão do agrupamento por divisão (FONTANA; NALDI, 2009).

Segundo Steinbach, Karypis e Kumar (2000), o algoritmo se comporta da seguinte forma:

1. Seleciona-se a base de dados a ser dividida;
2. Encontram-se dois s*ubclusters*, utilizando o k-means básico;
3. Repete-se o passo 2, escolhendo o *cluster* que possui maior similaridade global (*cluster* que possui mais similaridade média em todo conjunto de dados);
4. Os passos anteriores são repetidos até que a estrutura de *clusters* desejada seja alcançada.

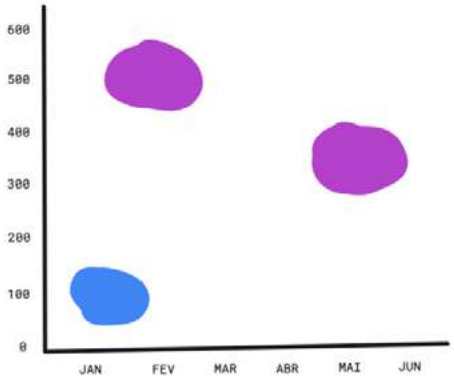
Alguns exemplos gráficos podem ser observados nas Figuras 9, 10 e 11, em que se verifica a separação dos grupos com base no algoritmo Bisecting k-means.

Figura 9 — Início da divisão, quando todos os dados participam de um único *cluster*



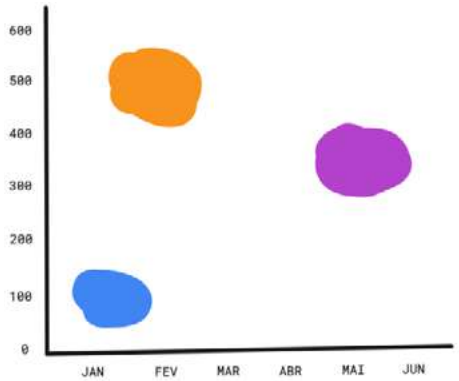
Fonte: adaptada de Linden (2009)

Figura 10 Divisão do *cluster* inicial utilizando o k-means com o parâmetro $k = 2$



Fonte: adaptado de Linden (2009)

Figura 11 Seleção do *cluster* com maior diâmetro e sua divisão em dois novos *clusters* utilizando o k-means com k = 2



Fonte: adaptada de Linden (2009)

6. Desafios

Em 1997, quando o computador IBM venceu o jogador de xadrez Garry Kasparov, um dos maiores desafios do aprendizado de máquina tinha sido vencido (NEWBORN, 1997). O computador mostrava-se melhor que o maior dos homens em uma tarefa tipicamente humana. A partir desse momento, a proposição de desafios mais complexos para as pesquisas fez-se necessária e, juntamente com essa proposição, emergiram novos desafios para suas resoluções (COLOMBINI; SIMÕES, 2019).

De acordo com o parecer dado em 2013 pela União Europeia (UE), cerca de 90% de todos os dados produzidos pela humanidade naquele ano advieram dos dois anos antecedentes (DERVOJEDA *et al.*, 2013). Sendo assim, afirma-se que um dos maiores desafios a ser resolvido pelo aprendizado de máquina não está na capacidade de armazenamento dos registros, mas sim em usá-los de forma prática e eficiente.

É notório que existe uma progressiva busca por processos automáticos que executam o particionamento de dados em grupos. Um exemplo disso é a utilização de sistemas de classificação que não possuem, de maneira já especificada, uma saída desejada, compondo-se, na verdade, de uma rede treinada por padrões de entrada, que, arbitrariamente, organiza tais padrões em categorias. Dessa maneira, como visto em tópicos anteriores, é possível utilizar técnicas de agrupamento para descobrir grupos naturais em conjuntos de dados sem ter qualquer conhecimento prévio das características dos referidos dados (KOGAN, 2007).

Na análise de *clusters*, um dos desafios apontados é encontrar o melhor resultado de agrupamento para determinado conjunto de objetos. Segundo Hruschka e Ebecken (2003), o problema apresentado na busca pelo melhor agrupamento é NP-completo; assim, não é possível computacionalmente encontrá-lo, desde que n (número de objetos) e k (número de *clusters*) sejam extremamente pequenos, dado que o número de partições em que se pode dividir *n* objetos em *k clusters* aumenta aproximadamente como .

Ankerst *et al.* (1999) citam três razões pelas quais a efetividade dos algoritmos de agrupamento mostra-se como um desafio em poten-

cial. Primeiramente, a grande maioria dos algoritmos de agrupamento necessita de valores para os parâmetros de entrada que são difíceis de determinar, especificamente para um conjunto de dados do mundo real contendo objetos de muitos atributos. Em segundo lugar, tais algoritmos mostram-se extremamente sensíveis aos valores de parâmetros, produzindo partições diferentes do conjunto de dados, mesmo para ajustes de parâmetros significativamente pouco diferentes. Em terceiro, os conjuntos de dados reais e de alta dimensão possuem uma distribuição ampla que não consegue ser revelada por um algoritmo de agrupamento, mas somente por um ajuste de parâmetro global.

7. Considerações finais

Mediante as informações apresentadas neste capítulo e as análises previamente realizadas, torna-se possível chegar a algumas considerações acerca do tema discutido.

O aprendizado de máquina notoriamente é uma das maiores tendências tecnológicas da atualidade. A partir dessa perspectiva, toda sua teoria vem sendo aplicada no desenvolvimento de modelos capazes de analisar grandes e complexos conjuntos de dados. Isso permite resposta ágil aos gestores para fins de tomada de decisões e visa, em um futuro próximo, fazer com que tais decisões possam ser tomadas sem nenhum tipo de intervenção humana.

Nesse panorama, o aprendizado de máquina não supervisionado baseado em agrupamento pode ser altamente útil para solução de problemas complexos, tais como a identificação de estruturas adjacentes que visem obter perspectivas sobre os dados e identificação do grau de semelhança entre as formas ou organismos, realizando, assim, uma classificação natural. Trata-se, além disso, de um recurso que pode ser extremamente útil na busca pela compreensão dos métodos para organização dos dados, pois é possível, a partir de um conjunto de dados, obter repostas através de cuidadosas seleções e processamentos.

No contexto geral, é de extrema importância definir o critério a ser utilizado para categorizar se dois elementos de um conjunto são idênti-

cos ou não. Para analisar esse quesito, é necessário considerar medidas que descrevam similaridades entre os elementos e suas respectivas características. Dessa maneira, existe uma grande variedade de medidas que podem ser utilizadas de acordo com a realidade de suas aplicações. Tal variedade mostra-se como um dos desafios em se estabelecer critérios de agrupamentos, o que permite também múltiplas interpretações baseadas na possibilidade de criarem-se várias partições.

Referências

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining Association Rules between Sets of Items in Large Databases. **SIGMOD Rec.,** v. 22, n. 2, p. 207-216, 1993.

ALVES, G. Aprendizado não supervisionado com K-means. **Neuronio.ia**, 11 dez. 2018. Disponível em: https://medium.com/neuronio-br/aprendizado-não-supervisionado-com-k-means-f4272dee98a0. Acesso em: 18 dez. 2020.

ANKERST, M.; BREUNIG, M. M.; KRIEGEL, H.; SANDER J. OPTICS: Ordering Points To Identify the Clustering Structure. **ACM SIGMOD record,** v. 28, n. 2, p. 49-60, 1999.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006.

CARVALHO, H. M. **Aprendizado de Máquina voltado para Mineração de Dados: Árvores de Decisão**. 2014. Monografia (Bacharelado em Engenharia de Software) – Universidade de Brasília, Brasília, 2014.

COLOMBINI, E. L.; SIMÕES, A. S. Robótica e aprendizado de máquina: uma caminhada lado a lado. **Computação Brasil: Revista da Sociedade Brasileira de Computação,** Porto Alegre, n. 39, p. 7-10, jan. 2019.

DERVOJEDA, K.; Verzijl, D.; Nagtegaal, F.; Lengton, M.; Rouwmaat, E.; Netherlands, P. B**ig data - artificial intelligence**. [*S.l.*]: European Union, 2013. Disponível em: https://ec.europa.eu/docsroom/documents/13411/attachments/2/translations/en/renditions/native. Acesso em: 18 dez. 2020.

ESCOVEDO, T.; KOSHIYAMA, K. **Introdução a Data Science: Algoritmos de Machine Learning e métodos de análise**. São Paulo: Ed. Casa do Código, 2020.

FÁVERO, L. P. L.; BELFIORE, P. P.; SILVA, F. L. da; CHAN, B. L. **Análise de dados:** modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37-37, 1996.

FONTANA, A.; NALDI, M. C. **Estudo de Comparação de Métodos para Estimação de Números de Grupos em Problemas de Agrupamento de Dados**. São Carlos: Universidade de São Paulo, 2009. ISSN 0103-2569.

FREITAS, A. L.; SANTANA JUNIOR, O. V. Machine Learning: Desafios para um Brasil competitivo. **Computação Brasil: Revista da Sociedade Brasileira de Computação**, Porto Alegre, n. 39, p. 7-10, jan. 2019.

GORDON, A. D. **Classification**: methods for the exploratory analysis of multivariate data. New York: Chapman and Hall, 1981.

HAYKIN, S. **Redes Neurais:** Princípios e prática. Porto Alegre: Bookman, 2007.

HENNIG, C. M.; MEILĂ, M.; MURTAGH, F.; ROCCI, R. **Handbook of cluster analysis.** Boca Raton: CRC Press, Taylor & Francis Group, 2016.

HENNIG, C.; LIAO, T. F. Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification (with discussion). **Journal of the Royal Statistical Society,** v. 62, p. 309-369, 2013.

HONDA, H.; FACURE, M.; YAOHUAO, P. Os Três Tipos de Aprendizado de Máquina. Brasília, DF: LAMFO/UNB, 2017. Disponível em: https://lamfo-unb.github.io/2017/07/27/tres-tipos-am/. Acesso em: 27 jul. 2021.

HRUSCHKA, E. R.; EBECKEN, N. **A genetic algorithm for cluster analysis. Intelligent Data Analysis**, v. 7, n. 1, p. 15-25, 2003. Disponível em: https://www.researchgate.net/publication/220571471_A_genetic_algorithm_for_cluster_analysis. Acesso em: 5 mar. 2021.

JAIN. A. K; DUBES R. C. **Algorithms for Clusterings Data**. Englewood Cliffs: Prentice Hall, 1988.

KOGAN, J. **Introduction to Clustering Largeand High-Dimensional Data**. NewYork: Cambridge, 2007.

LINDEN, R. Técnicas de Agrupamento. **Revista de Sistemas de Informação da FSMA**, Macaé, v. 4, n. 1, p. 18-36, fev. 2009.

LOPEZ, A. G. T. **Controle Preditivo com Aprendizado por Reforço para Produção de Óleo em Poços Inteligentes**. 2010. Dissertação (Mestrado em Engenharia Elétrica – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2010. Disponível em: http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0812725_10_cap_02.pdf. Acesso em: 1 maio 2020.

MILLIGAN, G. W. Clustering validation: Results and implications for Applied analyses. *In:* ARABIE, P.; HUBERT, L.; DE SOETE, G. (ed.) **Clustering and Classification**. River Edge: World Scientific, 1996. p. 341-375.

MITCHELL, T. M. **Machine Learning**. New York: Mcgraw-hillEducation, 1997.

MOURA, K. V. de. **Data Science:** um estudo dos métodos no mercado e na academia. 2018. Trabalho de Conclusão de Curso (Graduação em Administração) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2018.

NEWBORN, M. **Kasparov versus Deep Blue:** computerchess comes of age. New York: Inger, 1997.

PORTO FILHO, C. H. **Técnicas de Aprendizado Não Supervisionado Baseadas no Algoritmo da Caminhada do Turista**. 2017. Dissertação (Mestrado em Interunidades em Bioengenharia) – Universidade de São Paulo, São Carlos, 2017.

RAMOS, R. A. de L. **Um framework para adaptação de processos de software guiado por TDD em aderência a Norma ISO/IEC e IEEE 12207**. 2015. Monografia (Bacharelado em Ciência da Computação) – Centro Universitário de João Pessoa, João Pessoa, 2015.

ROKACH, L.; MAIMON, O. ClusteringMethods. *In:* MAIMON, O.; ROKACH, L. (ed.) **Data Mining and Knowledge Discovery Handbook**. Boston: Springer, 2005. DOI: https://doi.org/10.1007/0-387-25465-X_15.

ROZA, F. S. da. **Aprendizagem de máquina para apoio à tomada de decisão em vendas do varejo utilizando registros de vendas**. 2016. Trabalho de Conclusão

de Curso (Graduação em Engenharia de Controle e Automação) – Universidade Federal de Santa Catarina, Florianópolis, 2016.

STEINBACH, M.; KARYPIS, G.; KUMAR, V. **A Comparison of Document Clustering Techniques.** Minneapolis: University of Minnesota, 2000. Disponível em: https://hdl.handle.net/11299/215421. Acesso em: 22 jun. 2020.

STEINER, M. T. A. S.; SOMA, N. Y.; SHIMIZU, T.; NIEVOLA, J. C.; STEINER NETO, P. J. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. **Gest. Prod.,** São Carlos, v. 13, n. 2, p. 325-337, 2006. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104530X2006000200013&lng=en&nrm=iso. Acesso em: 5 mar. 2021.

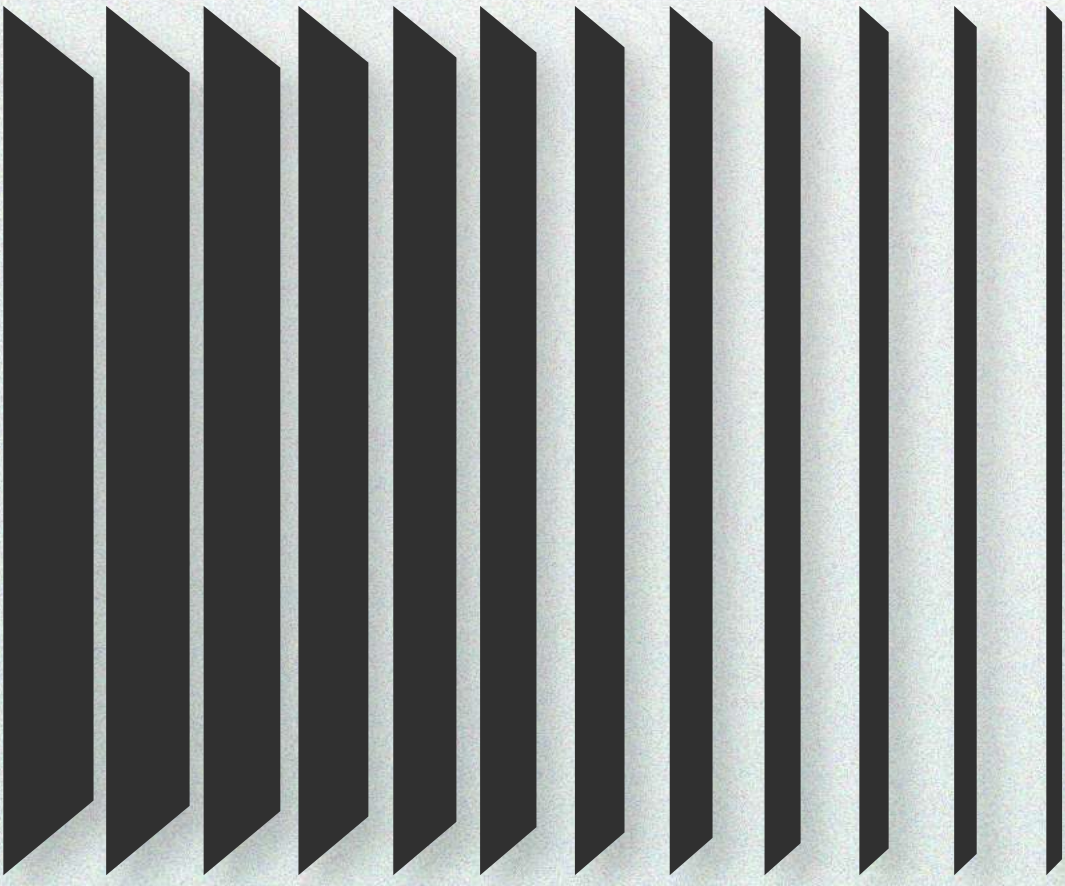
WITTEN, I.; FRANK, E.; HALL, M. A. Data Mining: Practical Machine Learning Tools and Techniques. **SIGSOFT Softw. Eng. Notes,** v. 36, n. 5, p. 51-52, 2011.

Capítulo 3

Visualização de Dados:

Uma Abordagem Introdutória no Contexto de Big Data

Victon Malcolm Rodrigues dos Santos
Wescley Paoli Alcantara de Sousa
Anthony Martins Araújo



1. Introdução

Ao longo dos anos, a humanidade passou por profundas transformações. Foi nesse lapso temporal que invenções como o microprocessador, redes de computadores, fibra óptica e computadores pessoais mudaram completamente a história da humanidade.

Peter Drucker (1995) foi o primeiro a chamar esse momento de "Era da Informação", termo frequentemente utilizado para designar os avanços tecnológicos advindos da Terceira Revolução Industrial e que refletiram na difusão de um ciberespaço, um meio de comunicação instrumentalizado pela Informática e pela Internet.

Uma das particularidades mais notórias da atual Era da Informação é a velocidade dos fluxos econômicos, sociais, culturais, linguísticos, entre outros, que gera um volume imenso de dados. Pode-se observar uma explosão de ferramentas de geração de dados, rastreamento, monitoramento, transações e redes sociais (PENA, 2019). Esse imenso volume de dados disponível tem sido denominado de Big Data. Em princípio, o termo Big Data se refere a um volume de dados extremamente amplo e que, por esse motivo, necessita de ferramentas para lidar com tal volume, de forma que toda e qualquer informação possa ser encontrada, analisada e aproveitada em tempo hábil (MARZ; WARREN, 2015). O aproveitamento desses dados está associado à possível identificação de padrões e de informações que possam agregar valor às empresas e à sociedade. Nesse contexto, a forma em que essas informações e padrões são apresentados pode facilitar seu uso e a tomada de decisão (KHAN; KHAN, 2011).

Esse modo de disponibilizar as informações e padrões ao tomador de decisão pode ser baseado em técnicas de Visualização de Dados. Segundo Romani e Rocha (2001), a Visualização de Dados, de forma geral, é o uso de imagens ou meios para representação de informações de modo significativo. Assim, a ideia é que os tipos de visualização produzidos possam compartilhar um objetivo em comum: transformar dados em algo mais expressivo, ou seja, uma representação visual útil, de forma que o observador humano possa ter um melhor entendimento. Pode-se conceber também a Visualização de Dados como a transmissão de conceitos

universais que permitem ao observador a rápida compreensão do que se é apresentado (MILLER, 2017).

A Visualização de Dados objetiva fornecer ou, ao menos, alavancar a capacidade de percepção dos dados, das informações e mesmo do conhecimento por parte do usuário, exigindo, assim, o uso de novas soluções que forneçam recursos buscando simplificar e, ao mesmo tempo, enriquecer a experiência do usuário (MARQUESONE, 2016). O uso de soluções específicas é recomendado devido à grande complexidade no desenvolvimento de sistemas de análise visual, pois as camadas tradicionais de software e hardware carecem de serviços essenciais, necessários para uma melhor experiência do usuário (FEKETE, 2013). Além disso, a Visualização de Dados tem se tornado cada vez mais pesquisada e entendida como essencial, tendo em vista que a variedade e o volume crescentes dos dados trouxeram o aumento da complexidade de suas análises (MARQUESONE, 2016).

Nesse cenário, este capítulo introduz alguns conceitos acerca do tema de Visualização de Dados considerando o cenário de Big Data e esta' estruturado da seguinte forma: na Seção 2, são introduzidos os conceitos básicos sobre Big Data e Visualização de Dados, assim como os tipos e processos comuns à Visualização de Dados. Na Seção 3, algumas técnicas para Visualização de Dados são apresentadas juntamente com exemplos de abordagens e ferramentas. Um exemplo de aplicação de Visualização de Dados é descrito na Seção 4. Na Seção 5, desafios são apontados. Por fim, a Seção 6 tece considerações, posicionamentos e sugestões de trabalhos futuros sobre o tema.

2. Fundamentação teórica

A Visualização de Dados vem evoluindo com o passar dos anos e, para melhor aplicar e compreender as técnicas associadas a essa área, é preciso ter em mente alguns conceitos básicos. Como o contexto ao qual este tema está inserido é referente a Big Data, seus principais conceitos são também introduzidos.

2.1. Big Data

Miller (2017) define Big Data como sendo "conjuntos de dados que são tão grandes ou complexos que as aplicações tradicionais de processamento de dados são inadequadas".

A globalização permitiu ao mundo conectar-se através da Web. Por meio dela e juntamente com o advento dos *smartphones*, uma grande quantidade de dados foi e esta sendo produzida, sendo ela estruturada ou não. De uma simples mensagem de texto a vídeos com várias horas, por natureza, esses dados são, em sua maioria, desorganizados ou com tipos e formatos diferentes dos convencionais, não podendo ser tratados, processados ou consultados de forma tradicional usando, por exemplo, um Sistema Gerenciador de Banco de Dados Relacional – SGBDR (KHAN; UDDIN; GUPTA, 2014).

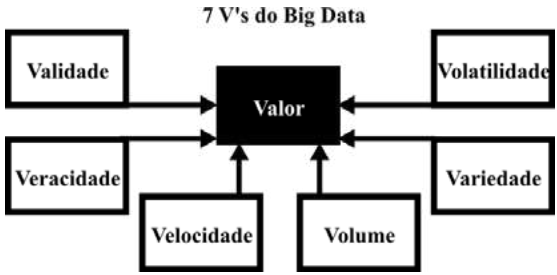
Ou seja, dados que compõem um conjunto Big são coletados a partir de diversas fontes – redes sociais, sensores, dados abertos – e podem ser estruturados em bancos de dados como, por exemplo, os NoSQL[1], que permitem manipular dados tanto semi como não estruturados. Habilidades no manuseio dessas grandes e diferentes massas de dados podem ser uma base para competitividade, aumento de produtividade, inovação e ampliação da gama de clientes em empresas (MANYIKA *et al*., 2011).

As etapas para a construção e utilização de um Big Data podem ser resumidas, segundo Miller (2017), em: Coleta de Dados, Limpeza dos Dados, Integração de Dados, Mineração e Análise de Dados e Visualização dos Dados, sendo, a última, o foco deste capítulo.

Objetivando uma melhor compreensão das características associadas a Big Data, Khan, Uddin e Gupta (2014) descrevem, conforme mostrado na Figura 1, os 7Vs comumente utilizados para sua definição, explicados a seguir:

1 SQL é uma linguagem padrão para acessar e manipular bancos de dados.

Figura 1 | 7Vs do Big Data



Fonte: adaptada de Khan, Uddin e Gupta (2014)

Volume - quantidade de dados produzida e que precisa ser processada;

Velocidade - rapidez com que os dados são produzidos e processados, sendo um dos fatores determinantes para o sucesso das aplicações;

Variedade - está relacionada aos diferentes tipos de dados produzidos e sua complexidade para gerar relacionamentos, sendo eles estruturados, semiestruturados e, em sua maioria, não estruturados;

Veracidade - refere-se a quão corretos estão os dados coletados para que possam ser utilizados em uma aplicação. A confiabilidade nos dados é o ponto principal neste quesito;

Validade - semelhante à veracidade, mas não igual; associa-se ao processo de manipulação dos dados e a quão preciso estes ainda são após atualizados ou transformados;

Volatilidade - diz respeito ao tempo em que os dados devem ser considerados ou descartados. Está relacionada diretamente com Volume, Velocidade e Variedade.

Valor - o principal de todos os Vs, uma vez que a aplicação de todos os Vs anteriores deseja, ao final, a obtenção deste. Quanto melhor for o tratamento dos Vs anteriores, maior será o valor agregado para as organizações que utilizam o Big Data.

A aplicação adequada de cada um dos Vs pode proporcionar o sucesso de uma aplicação de Big Data quanto à sua atuação. O modo em que os dados obtidos são apresentados tem um papel fundamental para o sucesso da aplicação, e cada etapa de preparação dos dados e análises associadas pode contribuir para o êxito do usuário final na tarefa de Visualização de Dados.

2.2. Visualização de Dados

Para Miller (2017), a Visualização de Dados pode ser definida como a representação visual criada a partir do uso de dados, relacionando-os de maneira a retratar informações sobre um cenário ou um momento determinado.

A tarefa de Visualização de Dados almeja dispor as informações representando-as da melhor maneira possível para uma clara compreensão dos usuários. Como exemplo, todo gráfico apresentado deve ter um objetivo claro, no qual uma vasta quantidade de dados é lapidada e organizada para responder às perguntas criadas (FRY, 2008). As perguntas usadas para montagem de visualizações estão associadas ao domínio do negócio e ao entendimento do que se pode obter com os dados.

Assim, definir o propósito da Visualização de Dados é essencial para o sucesso da tarefa e sua consequente aplicação, ja'que possibilita proporcionar a melhor apresentação dos dados, a fim de produzir informações e favorecer que o entendimento dos padrões apresentados seja graficamente para quem vai visualizá-los (MARQUESONE, 2016).

A Visualização de Dados objetiva transformar algo inicialmente de entendimento complexo em algo simples ou, ao menos, de rápida compreensão (MILLER, 2017). Avaliando os resultados e considerando-os positivos, tanto na compreensão quanto na representação, viabiliza-se um tipo de explanação dos dados em que a estrutura pode ser apreciada por grupos de usuários, muitas vezes, diversos (MARQUESONE, 2016).

Compreender o contexto da aplicação e fazer as escolhas corretas, como o modo de prover a análise, o tipo de gráfico mais adequado e a quem sera' destinada a informação, é fundamental para o sucesso

do método de visualização (MILLER, 2017). Um erro comum ao iniciar tarefas de visualização está em se concentrar na montagem de gráficos ou outras formas de visualização sem planejar como contar as histórias associadas aos dados (KNAFLIC, 2019). Knaflic (2019) indica, ainda, que uma Visualização de Dados eficaz pode significar o sucesso ou o fracasso na hora de comunicar resultados com os dados.

Dessa forma, apesar de os usuários serem capazes de processar informações rapidamente, a apresentação de análises cruas nem sempre é satisfatória. Devido a isso, a Visualização de Dados deve enfocar na transmissão de conceitos que favoreçam a melhor compreensão, sob os pontos de vista definidos para um público-alvo, de maneira rápida (MILLER, 2017).

Fry (2008) afirma que "quanto mais específica for a pergunta que suporta a visualização, mais específico e claro sera' o resultado". Ainda que os dados possam ser apresentados em vários formatos de gráficos, nem todo gráfico serve para determinados tipos de problema. Assim, é possível categorizar a Visualização de Dados pelo seu tipo de aplicação: Exploratória e de Apresentação.

2.2.1. Visualização Exploratória

Geralmente empregada em tarefas de Extração, Transformação e Carga (ETL) para verificação e compreensão dos dados, a maioria das visualizações utilizadas na Visualização Exploratória pode ajudar apresentando recursos importantes como, por exemplo, características inerentes aos dados brutos. Outro aspecto desse tipo de visualização é o uso de técnicas quantitativas também baseadas em Estatística, como diagramas, distribuições de dados, entre outros (AMARAL, 2016).

Na Visualização Exploratória, o analista foca em meios que favoreçam novas descobertas, não se preocupando no refinamento visual ou no usuário final (MARQUESONE, 2016). Gráficos gerados neste tipo de visualização geralmente têm um escopo amplo, pelo fato de os dados estarem sendo explorados, e são voltados a quem tem um conhecimento mais aprofundado nos dados (FRY, 2008).

2.2.2. Visualização de Apresentação

A habilidade de fazer boas perguntas é uma das mais importantes e requisitadas quando se trata da compreensão de dados, pois possibilita compartilhar os questionamentos e percepções para captação das visualizações pelo público-alvo (FRY, 2008).

Na Visualização de Apresentação – também chamada de Explanatória –, objetiva-se abordar o que foi descoberto sobre os dados, apresentando para um público preestabelecido os resultados obtidos, de forma eficaz e de fácil compreensão (MARQUESONE, 2016). Uma visualização adequada é aquela que fornece respostas a uma pergunta sem detalhes que fogem ao que foi perguntado; para isso, deve-se abstrair e remover o que não é necessário (FRY, 2008).

Por serem criados pensando em atender às expectativas solicitadas pelo cliente, estes tipos de visualizações são mais focados no usuário final, favorecendo o entendimento geral e podendo, eventualmente, estar adequados à reprodução impressa (CHEN, 2006).

O resultado almejado pela Visualização de Apresentação está relacionado a: i) uma melhor comunicação, possibilitando que os resultados sejam identificados facilmente pelo usuário; ii) um melhor monitoramento, permitindo resumir uma grande quantidade de dados em representações concisas; e iii) um apoio no processo de tomada de decisão, apresentando tendências e desvios que seriam de difícil identificação, proporcionando eficiência e otimização de tempo assim como auxiliando em decisões estratégicas (MARQUESONE, 2016).

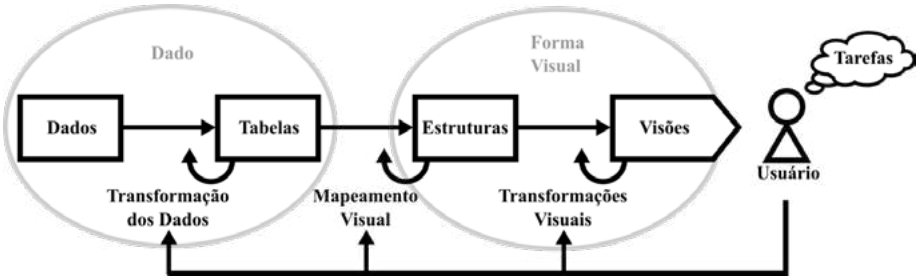
Neste tipo de visualização, que provê representações visuais mais enxutas e de fácil entendimento, são disponibilizados gráficos que geralmente utilizam poucas dimensões e que não exigem do leitor um conhecimento prévio de estatística (CHEN, 2006). Independentemente de flexibilizar ou não a interação, não definindo uma pergunta muito restrita e assim possibilitando diversas visões, é muito importante sempre deixar em destaque as principais conclusões obtidas (FRY, 2008).

Quanto à interação com o usuário final, a construção das visualizações precisa da integração entre as estruturas através de filtros, de forma

a facilitar sua montagem. A Visualização de Dados voltada ao usuário final é tratada como um processo de mapeamento dos dados para a forma visual. Sendo assim, a combinação de dados em estruturas visuais, com propriedades gráficas e organização espacial, cria visões esquematizadas e escaláveis, possibilitando que os usuários interajam de acordo com seu interesse (MENDONÇA NETO; ALMEIDA, 2001).

A parte final do *pipeline*, mostrado na Figura 2, apresenta a ideia de que os dados foram previamente transformados e preparados para serem usados em tarefas de criação de visualizações. Os dados são mapeados para visões que abstraem detalhes dos dados brutos e criam perspectivas mais simples e amigáveis ao usuário final. Na etapa de criação de visões, várias perspectivas ou formas podem ser criadas para os mesmos dados de modo a fornecer ao usuário mais de um formato visual para o mesmo tipo de informação. Logo, é possível que o usuário possa fazer a escolha do formato que considera mais representativo. Ainda na Figura 2, caso sejam realizadas etapas as quais o usuário possa acompanhar, sua interação pode ajudar na configuração da forma em que as visualizações vão ser criadas (FEKETE, 2013).

Figura 2 | Exemplo de etapas para criação de visualizações de dados



Fonte: adaptada de Card (1999 apud FEKETE, 2013) e Romani e Rocha (2001)

A tendência de uso e criação de visualizações de dados com foco e participação do usuário final segue o modelo apresentado na Figura 2. A ideia é que os usuários possam configurar suas preferências quanto às visualizações criadas (FEKETE, 2013).

Como mostra a Figura 3, a Visualização Exploratória é muito usada para prover a compreensão dos dados, principalmente em suas etapas de preparação ou pré-processamento. Já a Visualização Explanatória é voltada ao usuário final, apresentando os conceitos e padrões já descobertos, com informações bem trabalhadas, de forma a serem dispostas em relatórios e apresentações consolidadas.

Figura 3 | Visualização Exploratória e Visualização Explanatória

Fonte: adaptada de Marquesone (2016)

A seguir são apresentadas algumas técnicas para construção de representações visuais mais efetivas.

3. Técnicas para Visualização de Dados

Independente do problema, a forma em que os dados são apresentados pode ser determinante para um negócio, seja para exibição de resultados, para o impacto sobre um determinado setor ou para influenciar na tomada de decisão. Como correlacionar mensagens de uma rede social para ajudar a aprimorar produtos? Como encontrar bandidos procurados

no meio de uma multidão? Como verificar a adesão do confinamento social em tempos de pandemia? A Visualização de Dados, independente de ser exploratória ou explanatória, tem por objetivo, em conjunto com os demais processos de análise de dados, tornar possível que essas e outras perguntas sejam respondidas de maneira eficiente (FRY, 2008).

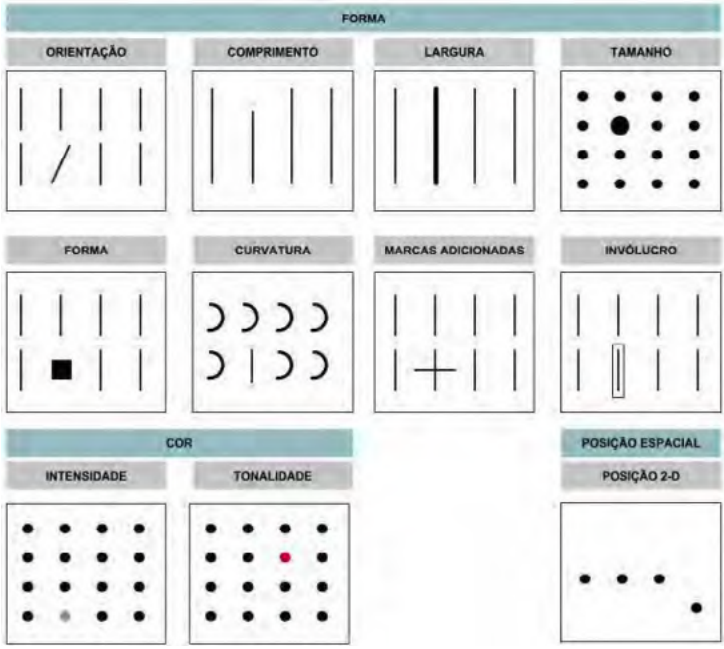
A visualização cria uma nova perspectiva nos dados, favorecendo a identificação de relacionamentos e características que dificilmente seriam percebidas, por exemplo, em uma tabela com milhares de linhas (MARQUESONE, 2016). A apresentação de gráficos de fácil entendimento é fundamental para que a transmissão da informação possa ser bem-sucedida. Para isso, quando a visualização de dados ocorrer por meio desse recurso, a escolha certa do tipo de gráfico pode fazer toda a diferença.

3.1. Interfaces visuais e tipos gráficos

A criação da representação visual dos dados deve lidar com aspectos que permitam ao usuário visualizar pontos específicos desses dados, bem como facilitar a apresentação das relações entre os diversos conjuntos de dados (MARQUESONE, 2016). As técnicas convencionais possuem características genéricas e de entendimento comum, a fim de transmitir os dados eficazmente de forma elegante, descritiva ou interpretável (KHAN; KHAN, 2011).

Os aspectos relacionados a forma, cor e posição são os principais meios de se comunicar quando se quer criar uma representação que facilite a interação, sendo que a escolha de atributos referentes a eles definirá se a visualização sera'ou não eficaz (MARQUESONE, 2016). Alterações na orientação, no tamanho, na área, na saturação, na luminosidade, na transparência, na textura, no rótulo e no movimento são capazes de apresentar uma informação de forma a evidenciar fatos, assim como causar mais ou menos impacto, pois são rapidamente identificadas, como ilustrado na Figura 4.

Figura 4 Atributos para Visualização de Dados



Fonte: adaptado de Marquesone (2016)

Por cada aspecto oferecer uma perspectiva diferente, deve-se escolher qual forma, cor e posição melhor representam a informação que se quer passar na construção de uma representação visual, mantendo o foco em ter uma visão mais clara e objetiva da informação. Para isso, é interessante contar com a ajuda de profissionais com conhecimento de design gráfico. Esse apoio pode ser fundamental para que os aspectos mostrados sejam mais bem percebidos e utilizados (MARQUESONE, 2016).

Embora a experiência seja exigida nessa tarefa, Marquesone (2016) aborda, sem entrar em detalhes, alguns pontos que podem servir de guia para a escolha dos aspectos na montagem de uma representação visual:

- comparação de valores - considera-se o uso de gráficos de colunas, de barras, de áreas circulares, de linhas e de dispersão;
- distribuição dos dados - busca ressaltar anomalias e tendências. Considera-se utilizar gráficos de dispersão, histogramas e gráficos em 3D;
- composição dos dados - considera-se o uso de gráficos de pizza, de área, de barras ou de colunas empilhadas.
- tendências em outros gráficos - considera-se o uso de linha, de linha com dois eixos e de coluna.
- Relação entre dados: considera-se a utilização dos gráficos bolha, linha, dispersão e grafo.

O surgimento de novos métodos e técnicas de visualização desenvolvidos nos últimos anos, mesmo que possuindo limitações quanto à sua implementação e adoção, foi possibilitado pela premissa de que a representação visual deve simplificar a visualização dos dados e seus relacionamentos (KHAN; KHAN, 2011). Nesse panorama, existem outros tipos de gráficos, mais específicos, que podem representar melhor uma informação, dependendo do cenário ao qual são empregados, que são: i) mapas, geralmente utilizados quando ha'dados geográficos envolvidos; ii) *word cloud*, usados para rápida identificação visual quando os dados possuem representações categóricas com indicadores numéricos relacionados; e iii) *layout* circular, utilizado quando se esta'trabalhando com grafos, proporcionando uma melhor visualização dos nós em formato circular (MARQUESONE, 2016).

Vale ressaltar que, independentemente da escolha dos gráficos a serem produzidos, quando se trata de Big Data, o uso de ferramentas adequadas e abordagens específicas também pode fazer a diferença no resultado final.

3.2. Abordagens e ferramentas para Visualização de Big Data

Como comentado anteriormente, quanto ao contexto de Big Data, abordagens simples de visualização geralmente podem não alcançar resultados satisfatórios na apresentação de informações. Ferramentas modestas e seus recursos podem ser inadequados, pois os conceitos e modelos para uma visualização do Big Data eficiente e eficaz exigem aspectos que apenas soluções robustas conseguem entregar (MILLER, 2017).

Existe uma grande variedade de formas convencionais para visualizar dados. Para se transmitir a informação complexa de forma eficaz, no entanto, é necessário, por vezes, mais do que um simples gráfico com os resultados (GOMES, 2011); é preciso realizar a aplicação de filtros ou possibilitar a mudança de perspectivas, tornando o gráfico mais robusto.

Seguindo a tendência das tecnologias de Big Data, a maioria das ferramentas utilizadas atualmente para Visualização de Dados é *open source* e permite a adaptação de seus gráficos em diferentes resoluções e dispositivos, além de possibilitar a interação com as representações visuais geradas (MARQUESONE, 2016).

A seguir, alguns exemplos de ferramentas são descritos considerando o contexto de Visualização de Dados e de Big Data.

3.2.1. Visualização e armazenamento

O problema do armazenamento crescente pode levar à expansão permanente de recursos de máquina, reduzindo a vida útil da solução (MILLER, 2017). Lidar com um grande volume de dados é um desafio complexo, pois encontrar informações relevantes e estabelecer relacionamentos entre os dados pode ser um fator de difícil resolução, dependendo do objetivo da análise (GOMES, 2011).

O uso do Hadoop[2] possibilita a criação de gráficos com indicadores oriundos do resultado de análises, tornando factível a utilização de grandes volumes de dados e oferecendo suporte a diversas outras ferramentas de visualização (MARQUESONE, 2016). O Hadoop pode viabilizar

2 Apache Hadoop. Disponível em: https://hadoop.apache.org.

ainda a análise exploratória dos dados completos de forma facilitada, sem necessitar a retirada de amostras ou coletas, retornando resultados de maneira eficiente em máquinas com poucos recursos (MILLER, 2017).

3.2.2. Visualização e análise de dados

A análise de dados pode apresentar fatos complexos e de difícil compreensão, tornando necessária a identificação do contexto, com base no problema e na aplicação, para a representação visual que será criada (MILLER, 2017). A ligação da análise dos dados com sua visualização está diretamente relacionada aos interesses dos usuários, que desejam entender os resultados obtidos sem ter um conhecimento prévio (KHAN; KHAN, 2011). O mapeamento dos dados em representações gráficas pode favorecer o detalhamento sobre o contexto destes (PEREIRA, 2015), mas é preciso cautela, pois, independentemente de fornecer as métricas pretendidas, a não aplicação de um contexto na utilização dos dados apresentados pode ocasionar distorções nas análises, produzindo resultados aquém do esperado (MILLER, 2017).

É fundamental perceber os tipos de dados que são passíveis de análise para que se possa transformá-los em representações gráficas intuitivas (PEREIRA, 2015). Adicionar contexto na utilização dos dados, com base no problema que está sendo atacado, objetivando uma melhor visualização, requer a manipulação destes, aplicando cálculos, agregações, adição de colunas ou reordenações, com o objetivo de revisar ou mesmo reformatar os dados (MILLER, 2017). Desse modo, não se trata apenas de apresentar os dados numa representação visual, mas também de como esses dados serão entendidos (GOMES, 2011)

O uso do R[3] pode facilitar a análise contextual por fornecer um apanhado de funções estatísticas, de manipulação e limpeza, técnicas gráficas e de modelagem mais sofisticadas. Pode também suprir as necessidades de tarefas mais simples, como produzir resumos para determinar agrupamentos (MILLER, 2017). Como o R armazena tudo em memória, a utilização de amostragem de dados (devido ao volume) é aceitável neste

3 The R Project for Statistical Computing. Disponível em: https://www.r-project.org.

ponto, mesmo quando se trata de Big Data, ja'que a sua aplicação está relacionada à obtenção do contexto (MILLER, 2017).

O resultado da exploração e das análises visuais pode dispor de informações suficientes do ponto de vista do usuário (PEREIRA, 2015). Essa atividade está vinculada diretamente à qualidade dos dados, relacionando-se também com o quanto esses dados atendem aos requisitos mínimos de um projeto em particular ou, pelo menos, ao nível de expectativas impostas a eles (MILLER, 2017).

3.2.3. Qualidade de dados

Mesmo que haja dados complexos conhecidos, qualquer visualização criada com dados deve ser utilizada de maneira a agregar valor a quem os utiliza e visualiza. Isso só ocorre caso as informações estejam em um determinado nível de qualidade, o que é extremamente difícil de alcançar quando se trabalha com uma grande quantidade de dados (MILLER, 2017).

Antes da veracidade ser incluída como característica de um Big Data, a comunidade assumia que os dados recebidos eram limpos e precisos. Hoje, com uma grande quantidade de dados coletados sendo não estruturados e vindos de diversas fontes, isso não pode ser considerado uma verdade absoluta (KHAN; UDDIN; GUPTA, 2014).

Oferecer recursos que permitam limpeza, gerenciamento e disponibilização de dados confiáveis melhora efetivamente a qualidade destes para soluções de Big Data, pois dados desatualizados, mal formatados ou mesmo errados comprometem os resultados apresentados (MILLER, 2017). Fazer bom uso de ferramentas e algoritmos que realizam a limpeza e preparação de um Big Data é de vital importância, uma vez que isso ajuda a garantir sua integridade, proporcionando uma maior confiança nos dados (KHAN; UDDIN; GUPTA, 2014). A título de ilustração, existem pacotes como os dplyr e tidyr, disponíveis no R, que podem ajudar.

3.2.4. Apresentação dos resultados

A apresentação dos resultados é o ponto de chegada almejado, o qual trata da organização e exibição dos dados em si ou de sua exibição gráfica, permitindo a visualização dos resultados do trabalho de análise mais claramente, com a complexidade dos dados sob um contexto simplificada ou com a possibilidade de compreensão de um determinado ponto de vista (MILLER, 2017).

A utilização de representações gráficas para comunicar dados é uma prática antiga que evoluiu com os computadores atuais, capazes de processar e produzir uma grande quantidade de visualizações gráficas rapidamente. Essa evolução permitiu à indústria o estabelecimento de uma grande expectativa quanto à visualização interativa, que possibilitaria a interação com o público em tempo real, tanto na escolha dos dados quanto no processamento e na apresentação dos seus resultados, entregando visualizações mais eficientes e personalizadas (MILLER, 2017).

Independentemente das diferentes maneiras que os usuários podem interagir com as visualizações produzidas, é preciso fornecer representações visuais de dados integrados aos mecanismos de interação, como filtros ou meios para mudanças de perspectiva, favorecendo a implementação da visualização de modo eficaz e sem esforço (KHAN; KHAN, 2011).

As habilidades com as quais o usuário se beneficia ao interagir com as ferramentas de visualização, segundo Marquesone (2016), compreendem:

- a filtragem de itens, permitindo ajustes e controle dos dados visíveis;
- os detalhes em demanda, possibilitando a visualização mais completa e detalhada dos dados;
- a relação entre os dados, que facilita a identificação do relacionamento entre os dados;
- os históricos de ações, que permitem ao usuário retornar às visualizações anteriores;
- a extração de subcoleções e consultas de parâmetros, permitindo navegação entre os diferentes cenários e possibilitando resgatar estados salvos;

- e o zoom, relacionado à redução ou ampliação da complexidade dos dados, além da escala.

Existem diversas técnicas visando à interação com representações gráficas, as quais objetivam o entendimento dos detalhes nos dados abordados, que geralmente seguem as etapas apresentadas de mapeamento e criação de visões (Figura 2), de modo a permitir maior interatividade. No contexto de Big Data, a visualização interativa está em evidência e pode ser notado no fato da adoção de painéis (*dashboards*) ter ganhado cada vez mais espaço, pois uma grande quantidade de dados pode promover diversas informações que precisam estar disponíveis para auxiliar nas análises e em tomadas de decisão (MILLER, 2017).

3.2.5. *Dashboards*

Devido à coleta e ao acúmulo contínuo de dados, as organizações têm confiado cada vez mais em soluções de Big Data, utilizando vários tipos de relatórios e criando diversos painéis, ampliando, assim, a capacidade de visualizar informações (MILLER, 2017).

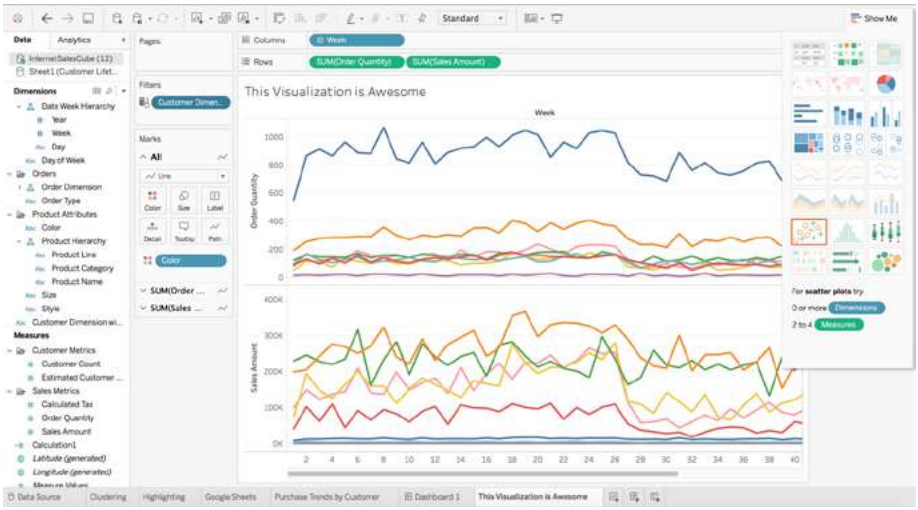
Com a crescente aplicação de painéis interativos, tornou-se ainda mais necessária a criação de representações visuais eficazes visando tirar o máximo proveito de Big Data (GOMES, 2011). Todos os painéis, se bem projetados e construídos, devem fornecer informações importantes e oportunas de maneira relevante e concisa, disponibilizando a capacidade de atualização no painel de controle, caso seja necessário, pois dados desatualizados, incorretos ou obsoletos podem levar organizações ao desastre (MILLER, 2017).

Alguns *frameworks* do pacote R tem chamado a atenção por permitir a criação de aplicações Web para Visualização de Dados, como o Shiny[4] e o Plotly[5], que oferecem diversas funcionalidades de interação e possibilitam a construção de interfaces dinâmicas (MARQUESONE, 2016), além de gerar um *link* direto com análises e perfis criados na ferramenta.

4 Shiny Dashboards. Disponível em: https://shiny.rstudio.com/articles/dashboards.html.
5 Plotly. Disponível em: https://plotly.com/.

Outra ferramenta que pode ser utilizada na criação de *dashboards* para apresentação dos resultados de análises de Big Data é o Tableau, que contém um conjunto de funcionalidades para visualização interativa dos dados, permitindo combinar diversas visualizações em um único painel e fornecendo a possibilidade de trabalhar com vários formatos de dados, sendo eles estruturados ou não (MULLER, 2017). A Figura 5 apresenta um exemplo de *dashboard* feito no Tableau.

Figura 5 | *Dashboard* criado no Tableau



Fonte: Tableau (2020)

Além do Tableau, diversas outras ferramentas, como o Pentaho[6] e o Qlik[7], também permitem a integração com o Hadoop quanto à captura e ao

6 Pentaho. Disponível em: https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-platform.html.
7 Qlik. Disponível em: https://www.qlik.com/pt-br/.

armazenamento dos dados, viabilizando que o usuário tenha diversas possibilidades quanto à criação de ambientes apropriados para qualquer processo de descoberta de dados no contexto de Big Data (MARQUESONE, 2016).

Apresentadas algumas ferramentas e técnicas com diferentes perspectivas, percebe-se que a Visualização de Dados pode ser uma poderosa ferramenta capaz de alinhar conceitos de usabilidade com processos analíticos, exploratórios e descritivos (KHAN; KHAN, 2011).

4. Exemplo de aplicação da Visualização de Dados

Abordados os termos e a importância da Visualização de Dados, vistos nas seções anteriores, serão demonstradas, através de uma aplicação, as capacidades de processamento de informações utilizando algumas ferramentas de Big Data e visualizações ricas que podem ser geradas a partir das práticas. Apesar da existência de diversas linguagens e ferramentas, como ja'mencionado neste trabalho, sera'apresentado um exemplo que, através de duas dessas ferramentas, gerou visualizações representativas e interativas. O objetivo não é detalhar todos os passos realizados, mas demonstrar algumas das práticas existentes no exemplo, que é baseado no trabalho de Silva (2016). Ele utiliza o domínio de investigação criminal hipotética para apresentar conexões e relacionamentos através de um diagrama de grafos.

4.1. Plataforma de dados

Neste exemplo de investigação criminal é criada uma aplicação que trabalha com diversos tipos de dados: áudio, dados bancários, dados biométricos, planilhas com informações financeiras, dados de redes sociais etc. O objetivo geral da aplicação com o Big Data é integrar todas essas informações de modo a ajudar os investigadores na tomada de decisão.

Após o ETL feito sobre os dados coletados, foi utilizado um banco de dados NoSQL orientado a grafos para seu armazenamento. Segundo Angles e Gutierrez (2008), há algumas situações em que bancos orientados a grafos apresentam aspectos positivos para uso em Big Data:

- quando a complexidade dos relacionamentos excede a capacidade de representação de outros modelos;
- em soluções que exigem linguagem de consulta mais poderosa associada à facilidade de uso em dados não estruturados e escalonamento do processamento;
- dependendo do tipo de representação dos dados.

Os insumos informacionais utilizados pelas polícias judiciárias são coletados por inúmeros aplicativos e fontes, estando presentes em sistemas de arquivos e conjuntos de dados diversificados, geralmente de forma dispersa e, muitas vezes, pouco estruturada (SILVA, 2016). Essa massa de dados, então, é submetida ao *Hadoop Distributed File System* (HDFS) a fim de ser armazenada e recuperada, de forma a complementar as informações trabalhadas pelo banco de dados de grafo. O HDFS é o sistema de arquivos distribuídos da plataforma Hadoop.

Com isso, foi possível estruturar o Big Data de forma a abranger todas as fontes de dados policiais no fornecimento de insumos, que foram tratados e mapeados para alimentar o *cluster* Hadoop e o banco de dados, de modo que as representações visuais pudessem ser almejadas. A arquitetura da solução é sintetizada na Figura 6.

Figura 6 Arquitetura da solução para investigação policial proposta

Fonte: adaptada de Silva (2016)

4.2. Inserção de dados no banco orientado a grafos

A escolha do banco de dados orientado a grafos possibilitou o processamento com grandes volumes de dados, assim como a exploração dos relacionamentos entre as entidades. Neste exemplo, Silva (2016) insere vários dados no banco, como investigados, linhas telefônicas, endereços de e-mail, contas bancárias, encontros registrados entre investigados, chamadas telefônicas e mensagens eletrônicas interceptadas e transações financeiras, criando uma teia de conexões que possibilita visualizar de forma direta o quão cada suspeito está ligado a outro.

O banco de dados orientado a grafos escolhido foi o Neo4j[8], por oferecer características de suporte à transação e à clusterização. Todos os dados foram inseridos no banco conforme exemplificado na Figura 7, sendo relacionados entre si por sua popularidade. Para obter os relacionamentos entre os entes, foi utilizada linguagem Cypher[9], uma linguagem declarativa, específica para realizar operações em grafos, através da

8 Neo4j. Disponível em: https://neo4j.com/.
9 Cypher Graph Query Language. Disponível em: https://neo4j.com/cypher-graph-query-language/.

qual é possível caminhar ao longo de todos os elementos do grafo sem que o tamanho deste comprometa o desempenho de consultas e atualizações (OMAND; BARTLETT; MILLER, 2012).

Figura 7 | Comando de inserção de informações no Neo4J de três indivíduos investigados na operação e de três contas bancárias a eles referentes

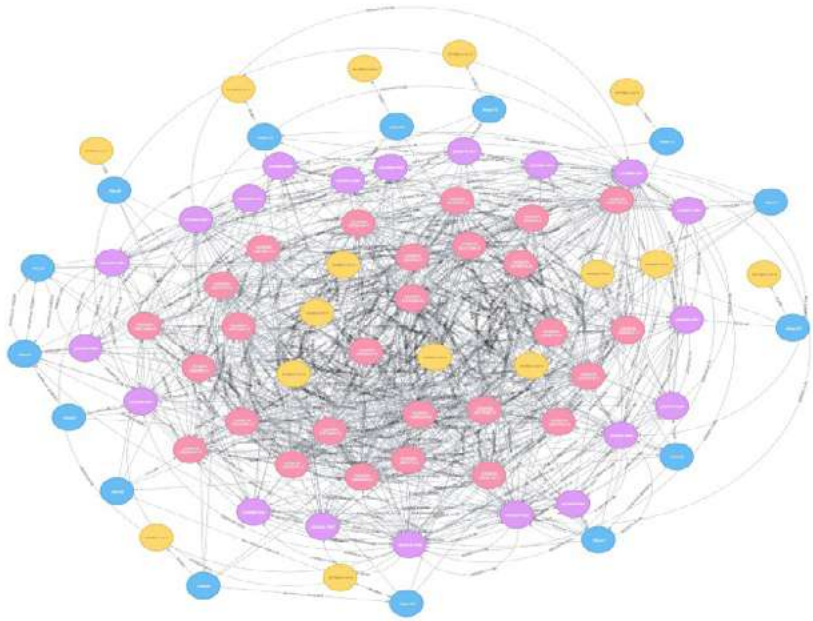
```
CREATE (p1:Pessoa { nome : 'Alvo1' , dataNascimento : '01/06/1956' , CPF : '135.626.530-41'})
CREATE (p2:Pessoa { nome : 'Alvo2' , dataNascimento : '13/09/1941' , CPF : '010.041.791-76'})
CREATE (p3:Pessoa { nome : 'Alvo3' , dataNascimento : '02/08/1966' , CPF : '645.103.203-05'})
```

```
CREATE (c1:Conta { nomeBanco : 'Banco1', agencia : '2912', numero : '514503-0' })
CREATE (c2:Conta { nomeBanco : 'Banco2', agencia : '9954', numero : '591565-3' })
CREATE (c3:Conta { nomeBanco : 'Banco1', agencia : '3455', numero : '205086-5' })
CREATE (p1)-[rt1:TITULAR]->(c1)
CREATE (p2)-[rt2:TITULAR]->(c2)
CREATE (p3)-[rt3:TITULAR]->(c3)
```

Fonte: adaptada de Silva (2016)

O diagrama de grafo gerado utilizando a rede de relacionamentos entre os investigados proporciona a visualização de informações completas, com riqueza de detalhes, por fornecer uma representação visual que atende tanto ao que se propõe na resolução do problema quanto aos anseios dos usuários que o utilizam, como visto na Figura 8.

Figura 8 Diagrama com relacionamentos entre os investigados, contas de e-mail, contas bancárias e telefones



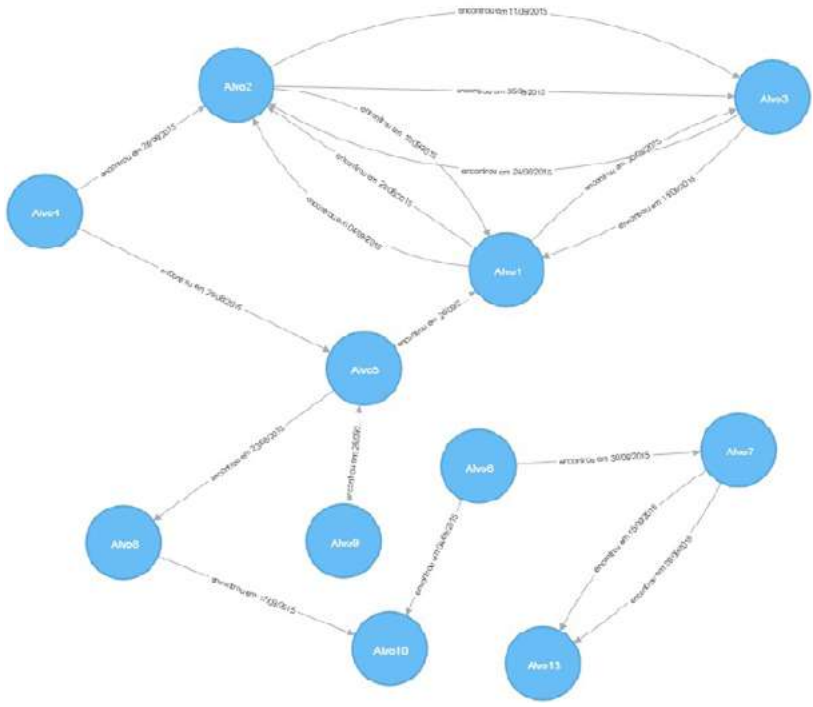
Fonte: adaptada de Silva (2016)

4.3. Exploração dos dados

Após a geração da rede de relacionamentos dos dados, o esforço é concentrado na obtenção de respostas. A partir do subconjunto alcançado através de consultas, fazendo uso dos relacionamentos, é possível analisar a conexão entre os investigados. Isso pode ocorrer separadamente ou em grupo, sendo que cada nó do grafo representa um suspeito, e as conexões representam o envolvimento entre eles. Um exemplo de

análise de relacionamento entre suspeitos pode ser visto por meio do diagrama da Figura 9.

Figura 9 | Análise de relacionamento entre investigados



Fonte: adaptada de Silva (2016)

4.4. Inserção de dados no Hadoop e operações de MapReduce

Para dar suporte a outras análises, Silva (2016) fez uso do Apache Hadoop 2.3.0, carregando esses dados para o sistema de arquivos HDFS.

Com a submissão dos dados a tarefas de MapReduce, foi possível encontrar informações importantes para a investigação.

De acordo com Holmes (2012), Lam (2010), Tao, Lin e Xiao (2013) e Rajaraman e Ullman (2012), conforme citado por Silva (2016), é possível programar várias funcionalidades no MapReduce sob a forma de tarefas, auxiliando a análise dos dados, como:

1. ordenar valores – realiza-se a divisão do grupo em partes, sua ordenação interna e a junção de cada partição ordenada;
2. encontrar palavras-chave e ranquea-las com base em ocorrência numa determinada massa de dados;
3. identificar conteúdo léxicos similares entre milhares de documentos;
4. identificar impressões digitais similares, por meio de comparação de características em fragmentos coletados;
5. contar elementos diferentes num determinado fluxo de dados;
6. identificar pontos geográficos que se localizam próximos uns dos outros.

Assim, como resultado do processamento de uma dessas tarefas, pode-se visualizar o que foi encontrado no processo da investigação. Esse achado é descrito por Silva (2016) quando revela que, ao excluir termos irrelevantes em trabalhos investigativos (preposições, conjunções, verbos e artigos), nomes de investigados aparecem ligados a uma das empresas com uma frequência razoável, além de citar localidades e práticas que poderiam ser associadas a atividades criminosas.

Neste exemplo foram apresentadas possibilidades que algumas ferramentas disponibilizam para trabalhar com grande volumes e variedades de dados. Nota-se que, para chegar ao resultado esperado, mais de uma ferramenta foi utilizada. Os resultados obtidos podem ser usados para facilitar a tomada de decisão, que, no contexto apresentado, seria encontrar os culpados e aplicar as devidas ações legais.

Como o nível de atuação foi mais focado no relacionamento entre os envolvidos na investigação, o uso de ferramentas de Visualização de Dados mais elaboradas não foi necessário, devido ao tipo de apresentação ter um contexto mais exploratório por parte dos investigadores e

analistas. Contudo, pode-se assumir que, caso fosse requerido, seria simples fazer a integração do trabalho por meio de *dashboards*, por exemplo.

5. Considerações

Com a realidade do Big Data, a importância da Visualização de Dados ficou ainda mais evidenciada, tendo em vista que dados obtidos e analisados precisam ser resumidos e organizados para uma melhor compreensão. A análise visual de dados é cada vez mais reconhecida como importante para o mercado e para a sociedade em geral, uma vez que estruturas e meios especiais para a visualização têm sido discutidos e criados. Um exemplo disso é a contínua busca por parte de diversas empresas no mundo, visando manter sua competitividade (FEKETE, 2013).

Este capítulo apresentou uma introdução a conceitos e práticas a respeito de Visualização de Dados no contexto de Big Data. A possibilidade de combinar Big Data com ferramentas de análise e visualização de dados viabiliza o uso de dados não estruturados e de fontes diversas por parte das organizações, proporcionando uma grande vantagem competitiva. A interação dos usuários com as visualizações – como, por exemplo, gráficos –, de acordo com suas necessidades, tem tornado a Visualização de Dados mais efetiva e uma boa experiência para quem a utiliza (MARQUESONE, 2016).

Foram abordadas também as possibilidades e as ferramentas utilizadas para conseguir atingir os objetivos da apresentação visual dos dados. Possibilitar o entendimento da narrativa visual e comunicar comportamentos, padrões e tendências dos dados é um dos resultados almejados quanto à utilização de propriedades gráficas em uma tarefa de Visualização de Dados (SEGEL; HEER, 2010 apud PEREIRA, 2015). Hoje existem diversas ferramentas e soluções para a Visualização de Dados integradas com Big Data, contendo várias formas visuais e compatíveis com os mais diferentes dispositivos (MARQUESONE, 2016).

Um ponto importante a ser destacado é que permitir a realização de alterações no modelo de visualização, através de técnicas interativas de filtragem, é indispensável para obtenção de informações coerentes

por parte do usuário, já que influencia diretamente no aspecto e no conteúdo apresentado, evidenciando contextos e conexões entre os dados (PEREIRA, 2015).

A capacidade de exploração de dados, associada à velocidade nas análises, deve ser simplificada de forma a conseguir apresentar representações gráficas com informações úteis (PEREIRA, 2015).

Apesar das informações apresentadas, muito conteúdo deixou de ser contemplado. Para trabalhos futuros, mediante evolução das práticas sobre o tema, novas técnicas podem ser abordadas. Pretende-se, além disso, atualizar as informações descritas neste conteúdo, melhorando sua estrutura e aprofundando os exemplos.

Referências

AMARAL, F. **Introdução à Ciência de Dados**. Rio de Janeiro: Alta Books, 2016.

ANGLES, R.; GUTIERREZ, C. Survey of graph data base models. **ACM Computing Surveys (CSUR),** v. 40, n. 1, p. 1-39, 2008.

CHEN, C. **Information Visualization:** Beyond the Horizon. 2. ed. Philadelphia: Springer. 2006. ISBN 9781846283406.

DRUCKER, P. F. **Administrando em Tempos de Grandes Mudanças**. São Paulo: Pioneira Thomson Learning. 1995.

FEKETE, J. D. Visual Analytics Infrastructures: From Data Management to Exploration. **Computer,** v. 46, v. 7, p. 22-29, 2013.

FRY, B. **Visualizing Data**. Sebastopol: O'Reilly Media. 2008. ISBN 9780596514556.

GOMES, L. F. O. **Percepção Humana na Visualização de Grandes Volumes de Dados:** Estudo, Aplicação e Avaliação. Dissertação (Mestrado em Multimédia) – Universidade do Porto, Porto, 2011.

HOLMES, A. **Hadoop in practice.** New York: Manning Publications Co., 2012. ISBN 9781617290237.

KHAN, M.; KHAN, S. S. Data and Information Visualization Methods, and Interactive Mechanisms: A Survey. **International Journal of Computer Applications**, v. 34, n. 1, p. 1-14, 2011.

KHAN, M. A.; UDDIN, M. F.; GUPTA, N. Seven V's of Big Data: Understanding Big Data to extract Value. *In:* CONFERENCE OF THE AMERICAN SOCIETY FOR ENGINEERING EDUCATION, 2014, Bridgeport. **Proceedings** [...]. [*S.l.*]: IEEE, 2014. p. 1-5.

KNAFLIC, C. N. **Storytelling com Dados:** um guia sobre visualização de dados para profissionais de negócios. Rio de Janeiro: Alta Books, 2019. ISBN 9788550804682

LAM, C. **Hadoop in action**. New York: Manning Publications Co. 2010. ISBN 9781935182191.

MANYIKA, J.; CHUI, M.; BROWN, B.; BUGHIN, J.; DOBBS, R.; ROXBURGH, C.; BYERS, A. H. **Big data:** The next frontier for innovation, competition, and productivity. [*S.l.*]: McKinsey Global Institute, 2011. Disponível em: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation. Acesso em: 04 ago. 2021

MARQUESONE, R. de F. P. **Big Data:** Técnicas e tecnologias para extração de valor dos dados. São Paulo: Casa do Código, 2016. ISBN 9788555192319.

MARZ, N.; WARREN, J. **Big Data – Principles and Best Practices of Scalable Real time Data Systems**. [*S.l.*]: Manning Publishing, 2015. ISBN 9781617290343

MENDONÇA NETO, M. G. de; ALMEIDA, M. O. Uso de Interfaces Abundantes em Informação para Exploração Visual de Dados. WORKSHOP SOBRE FATORES HUMANOS EM SISTEMAS COMPUTACIONAIS, 4., 2001, Florianópolis. **Anais** [...]. Florianópolis: UFSC: SBC, 2001. p. 256-268. ISBN 8588442094.

MILLER, J. D. **Big Data Visualization**. Birmingham: Packt, 2017. ISBN 9781785281945.

OMAND, D.; BARTLETT, J.; MILLER, C. Introducing social media intelligence (SOCMINT). **Intelligence and National Security**, v. 27, n. 6, p. 801-823, 2012.

PENA, R. F. A. Era da Informação. **Mundo Educação**, Goiânia, 2019. Disponível em: https://mundoeducacao.bol.uol.com.br/geografia/era-informacao.htm. Acesso em: 19 maio 2020.

PEREIRA, F. P. A. **Big Data e Data Analysis:** Visualização da Informação. Dissertação (Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação) – Universidade do Minho, Braga, 2015.

RAJARAMAN, A.; ULLMAN, J. D. **Mining of massive datasets**. Cambridge: Cambridge University Press, 2012.

ROMANI, L. A. S.; ROCHA, H. V. da. O uso de técnicas de Visualização de informação como subsídio à formação de comunidades de aprendizagem em EaD. Anais WORKSHOP SOBRE FATORES HUMANOS EM SISTEMAS COMPUTACIONAIS, 4., 2001, Florianópolis. **Anais** [...]. Florianópolis: UFSC: SBC, 2001. p. 169-182. ISBN 8588442094.

SILVA, GUSTAVO H. M. A. **Um modelo de Visualização de Dados utilizando banco de dados orientado a grafo suportado por big data**. Dissertação (Mestrado em Engenharia Elétrica) – Universidade de Brasília, Distrito Federal, 2016.

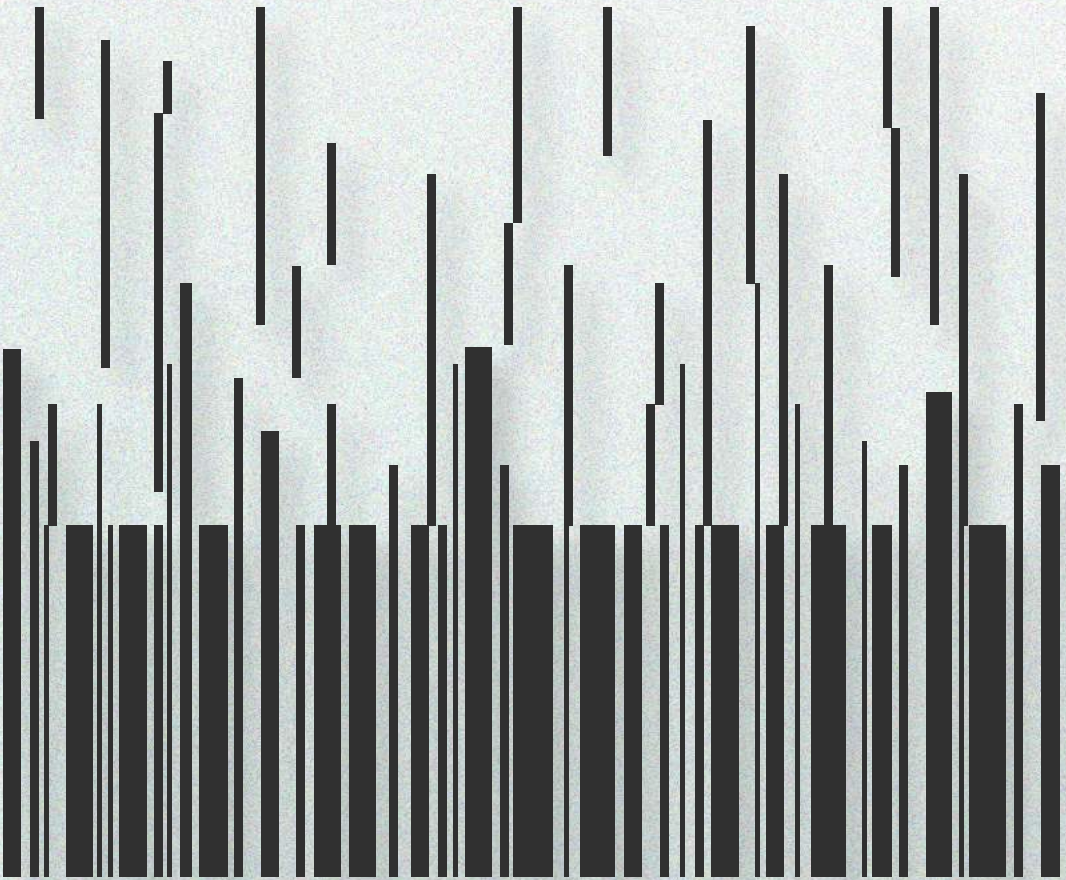
TABLEAU. Disponível em https://www.tableau.com/pt-br. Acesso em: 10 jul. 2020.

TAO, Y.; LIN, W.; XIAO, X. Minimal mapreduce algorithms. *In:* ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2013, New York. **Anais** [...]. New York: Association for Computing Machinery, 2013. p. 529-540.

Capítulo 4

Introdução à Privacidade de Dados e à Lei Geral de Proteção de Dados

Aline Priscila Araújo de Morais
Amanda Days Ramos Novo
Karine Heloise Felix de Sousa



1. Introdução

Em 2015, uma empresa de consultoria britânica chamada *Cambridge Analytica*[1] acessou dados pessoais de 87 milhões de usuários da rede social digital *Facebook*, com o objetivo de analisar e influenciar o comportamento de eleitores dos Estados Unidos da América. A intenção era identificar aqueles usuários que poderiam ser atraídos a votar em *Donald Trump*[2] ou desencorajados a votar em seu oponente (ISAAK; HANNAH, 2018). No Brasil, em 2014, a empresa de telecomunicação Velox foi acusada de, ilegalmente, vender dados pessoais de seus clientes a terceiros, culminando em multa de R$ 3,5 milhões (ZANATTA, 2015).

Até pouco tempo não se dava tanta atenção a questões associadas à privacidade e à segurança de dados. O aumento de notícias vinculadas a roubo de dados e violações de privacidade, entretanto, tornou o cenário tão crítico e preocupante que governos começaram a criar leis a fim de definir os direitos de privacidade dos dados dos usuários e penalidades explícitas para os casos em que as regulamentações não fossem cumpridas (MACHADO *et al*., 2020).

A exemplo disso, em 2018 surgiu a *General Data Protection Regulation* (Regulamento Geral de Proteção de Dados ou GDPR) – lei que regulamenta a segurança de dados no âmbito da União Europeia. Trata-se de um marco legal para a proteção e a privacidade de dados de todos os cidadãos europeus e do espaço econômico, tornando a proteção de dados pessoais um direito fundamental, assim como a liberdade (UE, 2016).

No Brasil, sob a influência da Lei Europeia, surgiu a Lei Geral de Proteção de Dados Pessoais (LGPD), que objetiva regulamentar a atividade de gerenciamento e tratamento de dados pessoais.

Juntamente com o Marco Civil da Internet e o Código de Defesa do Consumidor, a LGPD traz um conjunto de leis e normas que determinam

1 Cambridge Analytica foi uma empresa de consultoria britânica que atuava com serviço de análise de dados para fins comercial e políticos; após escândalo do vazamento de dados do Facebook, registrou pedido de falência em 18 de maio de 2018.

2 Donald Trump é o 45º presidente dos Estados Unidos da América, tendo sido eleito em 2016 pelo partido republicano.

como devem ser tratados os dados pessoais por sistemas de informação que processam e armazenam tais dados.

Tal ordenamento jurídico faz-se necessário diante da revolução tecnológica crescente e do maior número de dados produzidos a cada dia por diversos meios como, por exemplo, as redes sociais, sendo grande parte deles dados pessoais.

O grande volume de informações disponíveis digitalmente é o que se denomina Big Data, e dados pessoais podem estar presentes nesses conjuntos de dados. De acordo com De Mauro, Greco e Grimaldi (2016), Big Data é o ativo de informação caracterizado por um volume, velocidade e variedade tão altos que requer tecnologia específica e métodos analíticos para sua formação em valor. O Big Data representa uma revolução de dados relativamente recente, que tem sua grandeza confirmada pelos números que a acompanham. Trata-se de um fenômeno de rápido crescimento exponencial em todo mundo, com imensas consequências para sociedade, independentemente da classe social.

Quanto maior a capacidade de armazenamento e processamento de dados, maiores as chances de análise e geração de informações "de valor". Entretanto, sem as devidas proteções, o detentor dos dados pode encontrar-se em uma situação em que essas informações são usadas sem seu consentimento, autorização ou mesmo conhecimento. O simples fato de coletar ou tratar dados pessoais sem consentimento pode gerar problemas jurídicos relacionados ao direito à privacidade (CARLONI, 2013).

Este capítulo tem o objetivo de apresentar conceitos sobre Privacidade de Dados no cenário jurídico e no contexto computacional e introduzir aspectos sobre a Lei Geral de Proteção de Dados brasileira. Esses tópicos unidos são essenciais para o entendimento dos desafios a serem enfrentados pelas empresas tanto públicas quanto privadas e para os profissionais que atuam com o tema.

Além desta introdução, este capítulo está organizado da seguinte maneira: a Seção 2 apresenta a fundamentação teórica, conceituando a privacidade de dados no mundo jurídico e no mundo computacional e introduzindo, em seguida, a Lei Geral de Proteção de Dados brasileira, além de alguns conceitos estabelecidos na lei e a proteção aos dados

sensíveis; a Seção 3 discute técnicas de proteção de dados e cita alguns tipos de ataque à privacidade; a Seção 4 trata dos desafios à privacidade de dados; e, finalmente, a Seção 5 tece considerações finais a respeito deste capítulo.

2. Fundamentação teórica

Para melhor análise do tema, é necessário definir, inicialmente, alguns conceitos importantes que são introduzidos a seguir.

2.1. Privacidade de dados no mundo jurídico

De acordo com Castro (2005), o direito à privacidade foi citado pela primeira vez no ano de 1890, nos Estados Unidos da América (EUA), após a publicação do artigo *"The Right to Privacy"*, de autoria de Samuel D. Warren e Louis D. Brandeis, na *Harvard Law Review*[3].

O objetivo do artigo de Warren e Brandeis era estabelecer limites para a intromissão da imprensa na vida privada. Os autores defenderam a importância de que dados pessoais não fossem tornados públicos, sendo, portanto, resguardados. Assim, para Warren e Brandeis (1980), a privacidade pode ser definida amplamente como o "direito de estar só" ou o "direito de ser deixado só".

O conceito de privacidade, assim como de outros direitos, tem sofrido mudanças com o tempo. O que era entendido como privacidade no final do século XIX, quando o assunto começou a ser debatido, já não é suficiente para definir a privacidade na sociedade atual (CARLONI, 2013).

Hoje em dia, a sociedade se depara com um grande desafio: ampliar a proteção à privacidade frente ao desenvolvimento tecnológico. Dessa vez, no entanto, tal proteção ultrapassa o "direito de estar só" – as sensações, as emoções e os pensamentos ganharam a forma de dados pessoais, informações que dizem respeito a um indivíduo ou que o tornam

3 A Harvard Law Review é uma revista criada em 1887 por um grupo de estudantes da Harvard Law School nos Estados Unidos da América.

identificável. Essas informações facilmente circulam pela rede mundial de computadores e por dispositivos digitais (CARLONI, 2013).

Na Constituição Federal Brasileira (CFB), o direito à privacidade é considerado um direito fundamental, sendo a privacidade essencial na formação da pessoa, indispensável à construção do indivíduo e de suas fronteiras com os demais (BRASIL, 1988). A CFB de 1988, no Artigo 5º, inciso X, prescreve que são invioláveis a honra, a intimidade, a vida privada e a imagem das pessoas, assegurando o direito à indenização pelo dano material ou moral decorrente de sua violação (BRASIL, 1988). Não se verifica, no texto desse Artigo, a palavra privacidade, contudo os termos intimidade e vida privada são expressões relacionadas à privacidade do indivíduo.

De acordo com Pezzi (2007), em sua grande maioria, os juristas brasileiros consideram que a intimidade e a vida privada não são semelhantes, mas estão em uma relação de gênero e espécie, constituindo, a intimidade, um âmbito mais restrito da vida privada. Segundo Cancelier (2017), há informações que, mesmo não sendo íntimas, estão inseridas na vida privada, a exemplo do endereço de uma pessoa, que não pode ser classificado como uma informação íntima, mas é parte do contexto da vida privada, e sua exposição, sem a devida autorização, reflete uma violação da vida privada, podendo causar danos à intimidade, portanto, à privacidade.

O direito à privacidade passa, então, a ser compreendido não mais como a "simples ausência do conhecimento alheio sobre os fatos da vida privada do indivíduo, mas sim sobre o controle exercido sobre essas informações e dados pessoais" (LEONARDI, 2011).

2.2. Privacidade de dados no mundo computacional

O uso massivo da Internet por grande parte da população mundial e a evolução da tecnologia da informação e comunicação ocasionaram um crescimento exorbitante no volume e na variedade de dados existentes. Os conjuntos de dados gerados podem ser combinados e utilizados para inferências e análises. Essas tarefas geram riscos à privacidade dos sujeitos quando tais conjuntos de dados criam uma forma de impressão

digital, de modo que os indivíduos podem ser reidentificados mesmo que os dados estejam anonimizados[4] (MOONEY; PEJAVER, 2018).

Da mesma forma, cresceram as ameaças de exposição e de utilização demasiada ou inadequada de dados pessoais de indivíduos. A privacidade das informações pessoais virou fonte de vantagens econômicas; assim, as empresas, as corporações e as instituições devem armazenar as informações dos sujeitos de forma adequada, utilizando, por exemplo, processos e técnicas de anonimização. Independentemente do tempo e do meio onde os indivíduos estão inseridos, existe a necessidade de privacidade. A privacidade encontra uma barreira para a sua existência no mundo virtual, devido à facilidade da transmissão da informação (BRITO; MACHADO, 2017).

Mark Zuckerberg, fundador do Facebook, em evento realizado por uma empresa de tecnologia em 2010, afirmou que a privacidade de dados não é mais considerada uma norma social, pois evoluiu com o tempo na medida em que "as pessoas têm realmente se sentido mais confortáveis não apenas para compartilhar mais informações e de diferentes tipos, mas também de forma mais aberta e com mais pessoas" (LEE, 2013). A visão de Zuckerberg diverge da versão de Warren e Brandeis, contudo, pois o hábito de compartilhar informações e dados pessoais não significa que os indivíduos estão de acordo com o uso indiscriminado desses dados por terceiros, sem seu consentimento ou controle (CARLONI, 2013).

Conforme Vimercati *et al.* (2012) e Camenisch, Fischer-Hübner e Rannenberg (2011), existem algumas maneiras pelas quais é possível identificar ou categorizar um sujeito. As formas de identificação são denominadas de atributos. A classificação dos atributos ligados à privacidade do sujeito está particionada do seguinte modo:

- **atributo identificador** - aquele que torna possível identificar unicamente os indivíduos, como CPF, número da identidade;

4 De acordo com a LGPD, dado anonimizado é um dado que passou por um tratamento e não pode ser vinculado a um titular.

- **atributo semi-identificador** - atributo que, quando combinado com informações externas, torna possível identificar ou supor quem é o indivíduo, como data de nascimento, CEP, cargo, tipo sanguíneo;
- **atributo sensível** - refere-se a informações confidenciais ou sensíveis do indivíduo, como doenças/comorbidades, salário, exames médicos, lançamentos do cartão de crédito.

Vimercati *et al*. (2012) caracterizam os atributos não sensíveis como aqueles que contêm todos os atributos que não estão enquadrados nas três categorias mostradas anteriormente e que a sua publicação não causará nenhum prejuízo à privacidade. Com a percepção da necessidade de classificação e tratamento diferenciado a cada tipo de atributo, verifica-se que a privacidade de dados se tornou um tema que deve ser considerado em diversos ordenamentos jurídicos como um recurso elementar para a proteção da privacidade das pessoas assim como um direito.

2.3. Lei Geral de Proteção de Dados

Em agosto de 2018, foi criada a Lei Geral de Proteção de Dados (LGPD), que regulamenta como os dados pessoais serão tratados no Brasil. Inspirada na Lei Europeia, a LGPD tem entre suas finalidades "proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural" (BRASIL, 2018a). A LGPD surgiu para preencher uma série de lacunas nos ordenamentos jurídicos já existentes no Brasil, que estavam em situação instável, não direcionados a enfrentar uma economia e uma sociedade progressivamente movida por dados.

Pode-se verificar, no ordenamento jurídico brasileiro, um rol de leis que trata sobre a proteção à privacidade, à honra e à intimidade, que estão: na Constituição Federal (1988), no Código de Defesa do Consumidor (1990), na Lei do Habeas Data (1997), no Código Civil (2002), na Lei do Cadastro Positivo (2011), na Lei do Acesso à Informação (2011) e no Marco Civil da Internet (2014). Nesse rol do ordenamento jurídico não existe uma descrição objetiva dos direitos, deveres ou de responsabilidades no que se refere à proteção dos dados pessoais. Essas normas estão delimi-

tadas às suas respectivas finalidades de aplicabilidade;não governam as questões que dizem respeito à nova visão dos métodos de regimes internacionais de proteção dos dados (BRASIL, 2018b).

Na LGPD, o Artigo 5º, inciso X, considera que o tratamento dos dados é toda operação realizada com dados pessoais, ou seja, aquelas que se referem a: coleta, produção, recepção, classificação, utilização, acesso, reprodução, transmissão, distribuição, processamento, arquivamento, armazenamento, eliminação, avaliação ou controle da informação, modificação, comunicação, transferência, difusão ou extração de dados.

De acordo com Alves (2018), a LGPD representa um marco normativo para a sociedade brasileira: é a legislação que mais efetivamente busca solucionar o diálogo necessário entre a preservação e o respeito aos direitos fundamentais da liberdade e da privacidade.

O Artigo 5º da Lei conceitua três tipos de dados, descritos a seguir:

- **dado pessoal** - informação relacionada à pessoa natural identificada ou identificável, a exemplo de nome, endereço residencial, registro geral e cadastro de pessoa física;
- **dado sensível** - dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural, a exemplo de apelidos, fotos, relatórios médicos, salário e geolocalização;
- **dado anonimizado** - dado relativo ao titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento, a exemplo de quando é utilizado um tratamento reverso para descobrir os dados pessoais ou sensíveis dos sujeitos.

Menezes Neto, Morais e Bezerra (2017) afirmam que a classificação cria três níveis de proteção distintos: os dados sensíveis gozarão de maior proteção entre todos, seguidos pelos dados pessoais e, por fim, pelos dados anônimos. Esses últimos gozam de menor privilégio, uma vez que, supostamente, não seriam capazes de identificar os indivíduos aos quais se referem.

Um ponto trazido pela LGPD é a manipulação dos dados pessoais, o meio de consentimento à forma de serem coletados esses dados. O

consentimento dado deve ser de forma abrangente, com a descrição da finalidade e da utilização dos dados. As antigas e novas bases de dados existentes devem se adequar à Lei Geral de Proteção dos Dados.

O Artigo 5º também mostra cada um dos envolvidos no procedimento de manipulação dos dados e suas respectivas condutas, a saber:

- **titular** - pessoa natural a quem se referem os dados pessoais que são objetos de tratamento, a pessoa a quem os dados pertencem ou dizem respeito;
- **controlador** - pessoa física ou jurídica, de direito público ou privado, a quem competem às decisões referentes ao tratamento de dados das pessoas, ou seja, o ente responsável por realizar a coleta dos dados de todas as pessoas ou estabelecimentos;
- **operador** - pessoa física ou jurídica, de direito público ou privado, que realiza o tratamento de dados do titular em nome do controlador;
- **encarregado** - pessoa indicada pelo controlador e operador para atuar como canal de comunicação entre o controlador, os titulares dos dados e a Autoridade Nacional de Proteção de Dados (ANPD).

O direito mais elementar, em termos de proteção de dados, é o de titularidade de seus dados pessoais. A LGPD assegura outros direitos do titular conforme Artigo 5º. A Figura 1 apresenta os direitos assegurados aos titulares dos dados.

Figura 1 Direitos do titular pela LGPD

Fonte: SERPRO (2019)

Na LGPD está previsto um rol de sanções para serem aplicadas a infratores. O ônus da prova acerca do consentimento do titular está em consonância com a referida Lei.

Ao controlador competirá todas as precauções de tratamento, armazenamento e proteção que precisarão ser adotadas para assim se abster do corrompimento dos dados. No entanto, caso ocorra vazamento ou comprometimento de informações, o controlador e os envolvidos ficam sujeitos às sanções administrativas que devem variar em razão das infrações cometidas.

As sanções estão pautadas no Artigo 52º da LGPD e são apresentadas no Quadro 1:

Quadro 1 | Sanções da LGPD

Advertência	Indicação de prazos para adoção de medidas corretivas
Multa Simples	Até 2% do faturamento anual limitada a R$ 50.000.000,00 por infração
Publicização	Após devidamente apurada e confirmada a infração
Bloqueio	Parcial do funcionamento do banco de dados e/ou do tratamento de dados
Eliminação	Eliminação dos dados pessoais a que se refere a infração

Fonte: elaborado pelas autoras

Diante de tantos direitos do titular, como demonstrado na Figura 1, existem também as sanções que podem ser submetidas aos controladores, conforme apresentado no Quadro 1. Além disso, verificou-se a necessidade de criar um órgão que regulamentasse esses direitos. Então, em seu Artigo 55-A, a LGPD determinou a criação de um órgão regulador nomeado pela ANPD, de natureza jurídica, transitória, que poderá ser transformado pelo Poder Executivo em entidade da administração pública federal indireta, submetida a regime autárquico especial e vinculada à Presidência da República (BRASIL, 2019).

A ANPD tem como principal atribuição zelar, implementar e fiscalizar o cumprimento da LGPD em todo o território nacional, de acordo com seu Artigo 55-J.

Sendo assim, Stival (2015) afirma que a aprovação da LGPD equiparou o Brasil a outros países considerados adequados para salvaguardar os dados pessoais, representando um avanço na proteção dos direitos fundamentais de seus cidadãos.

2.4. Proteção aos dados sensíveis

Os dados sensíveis são aqueles associados às opções e às características basilares da *personae*, portanto, aptos a gerar situações de discriminação e desigualdade (MORAES, 2008). É com fundamento na possibilidade de utilização discriminatória, tanto por parte do mercado quanto do Estado, que os dados sensíveis se associam a conjunturas em que podem estar presentes potenciais violações de direitos fundamentais, em razão da sua natureza (MULHOLLAND, 2018).

Acerca da importância da proteção dos dados, Dominguez (2013 apud RAMINELLI; RODEGHERI, 2016) afirma que, além da mera classificação como "informações", deve-se lembrar que a combinação de dados pessoais permite a obtenção de um perfil muito preciso dos interesses e atividades de um indivíduo, sendo que estes dados podem ser utilizados para fins diversos, principalmente comerciais e publicitários.

Ademais, surgem outros riscos, mais preocupantes, como é o caso de roubo de identidade, para fins criminosos, ou até mesmo perda de um possível emprego, devido a buscas prévias acerca do candidato pela empresa que deseja contratar. Sendo assim, proteger dados sensíveis permite a efetivação de diversos direitos, como saúde, liberdades comunicativas, religiosa, de associação, entre outros (MULHOLLAND, 2018).

A LGPD, em seu Artigo 11, institui restrições importantes sobre as estratégias de proteção de dados sensíveis, estabelecendo, ao titular, consentimento de forma específica e destacada. Permite, também, que haja tratamento de dados sensíveis sem a necessidade de fornecimento de consentimento do titular de dados quando estes forem indispensáveis para o tratamento compartilhado de dados necessários à execução, pela administração pública, de políticas públicas previstas em leis ou regulamentos (BRASIL, 2018a), além de outras hipóteses que se referem, em grande medida, a interesses públicos.

De acordo com Negri e Korkmaz (2019), é possível identificar uma estratégia mais rigorosa para os dados pessoais e sensíveis em razão da sua natureza. É relevante considerar que as normas jurídicas, enquanto

geradoras de práticas protetivas, podem avançar para outras esferas, inclusive a interna dos controladores e operadores de tratamento de dados.

3. Técnicas de proteção de dados

Cada vez que um conjunto de dados é viabilizado com intuitos estatísticos, para pesquisas, análises de dados ou testes, processos e técnicas de precaução para a privacidade se fazem necessários para minimizar a descoberta de informações sensíveis por invasor. Um exemplo seria a liberação, pelos hospitais, dos dados sobre seus pacientes, seja para cooperar com pesquisas nas mais diversas áreas ou para encontrarem a cura, fatores, causas, frequência de ocorrência ou biotipo dos afetados por um vírus, bactéria ou doença.

Com o intuito de minimizar os tipos de ataques apresentados anteriormente, foram criados diversos processos de anonimização. No processo de anonimização são propostas formas para mascarar ou maquiar os dados antes de serem publicados ou compartilhados, realizando técnicas para que não seja possível identificar os dados pessoais de maneira prevista (BASSO *et al.*, 2016).

Em um processo de anonimização, um conjunto de dados original é transformado em um novo conjunto, por meio de modificações, tendo como objetivo evitar a descoberta de dados sensíveis por usuários maliciosos. Para executar a anonimização, se faz necessário definir os atributos que irão ser anonimizados e quais as técnicas que vão ser utilizadas em cada um deles.

O Quadro 2, apresentado a seguir, mostra atributos de uma tabela que não sofreram nenhuma técnica de anonimização. É possível identificar a identidade de uma pessoa, através do nome, data de nascimento e CEP. Tais atributos são considerados dados sensíveis e, de acordo com sua sensibilidade, as técnicas de anonimização podem ser utilizadas para proteger a identidade dos indivíduos.

Quadro 2 | Atributos de uma tabela não anonimizados

Nome	Raça	Data de Nascimento	Sexo	CEP	Reclamação
Sean	Negro	20/09/1965	Masculino	02141	Falta de Ar
Daniel	Negro	14/02/1965	Masculino	02141	Dor no Peito
Kate	Negro	23/10/1965	Feminino	02138	Olho dolorido
Marion	Negro	24/08/1965	Feminino	02138	Sibilo
Helen	Negro	07/11/1964	Feminino	02138	Dores nas Articulações
Reese	Negro	01/12/1964	Feminino	02138	Dor no Peito
Forest	Branco	23/10/1964	Masculino	02138	Falta de Ar
Hilary	Branco	15/03/1965	Feminino	02139	Hipertensão
Philip	Branco	13/08/1964	Masculino	02139	Dores nas Articulações
Jamie	Branco	05/05/1964	Masculino	02139	Febre
Sean	Branco	13/02/1967	Masculino	02138	Vômito
Adrien	Branco	21/03/1967	Masculino	02138	Dor nas Costas

Fonte: adaptado de Ohm (2010)

As técnicas encontradas em Brito e Machado (2017), as sugeridas por Ohm (2010) e as definidas por Branco Júnior, Machado e Monteiro (2014) são descritas resumidamente a seguir:

- criptografia - é uma técnica que, através de um algoritmo, embaralha matematicamente os dados, de modo que estes fiquem ilegíveis. Esses dados podem ser transformados de volta para seus valores originais através da utilização de uma chave de acesso;
- distúrbio - é uma técnica que, por meio de um mascaramento, substitui os dados reais por dados fictícios. Como exemplo, pode-se citar a subs-

tituição de sobrenome de família por outro proveniente de uma grande lista randômica de sobrenomes;

- substituição - é uma técnica que substitui os dados originais por outros dados que não se relacionam com esses, através da implementação de uma lista de palavras taticamente estabelecida;
- embaralhamento - é uma técnica em que ocorre a mistura aleatória dos dados semelhantes, porém os dados devem estar localizados na coluna da mesma tabela;
- anulação - esta técnica substitui os dados sensíveis por valores nulos. É utilizada quando os dados existentes na tabela não são requeridos.

Em adição, Brito e Machado (2017) descrevem a técnica de tokenização como aquela que gera aleatoriamente um valor de token, sem nenhuma formatação específica baseada no registro original. A tabela de mapeamento guarda o mapeamento desse token e seu respectivo valor da tabela original.

Ohm (2010), por sua vez, introduz também as seguintes técnicas:

supressão – é uma técnica em que ocorre a remoção completa da coluna que corresponde ao dado a ser anonimizado. A técnica de supressão foi realizada com base no Quadro 1 e apresentada no Quadro 2. Primeiramente, identificam-se os atributos a serem anonimizados – nome, data de nascimento, CEP e sexo, por exemplo; posteriormente, as colunas destes são suprimidas.

Quadro 3 Técnica de supressão

Raça	Reclamação
Negro	Falta de Ar
Negro	Dor no Peito
Negro	Olho dolorido
Negro	Sibilo
Negro	Dores nas Articulações

Continua

Conclusão

Raça	Reclamação
Negro	Dor no Peito
Branco	Falta de Ar
Branco	Hipertensão
Branco	Dores nas Articulações
Branco	Febre
Branco	Vômito
Branco	Dor nas Costas

Fonte: adaptado de Ohm (2010)

- generalização - é uma técnica em que ocorre a verificação dos atributos semi-identificadores – atributos que não são identificadores explícitos, mas podem potencialmente identificar um indivíduo –, e estes são alterados por valores semanticamente equivalentes, mas menos específicos, para, assim, conservar a veracidade dos dados.

A técnica de generalização foi realizada com base no Quadro 1 e apresentada no Quadro 3. Primeiramente, identificam-se os atributos semi-identificadores a serem anonimizados – data de nascimento e CEP, por exemplo; posteriormente, eles são alterados por valores menos específicos.

Quadro 4 Técnica de generalização

Raça	Data de Nascimento	Sexo	CEP	Reclamação
Negro	1965	Masculino	021	Falta de Ar
Negro	1965	Masculino	021	Dor no Peito
Negro	1965	Feminino	021	Olho dolorido
Negro	1965	Feminino	021	Sibilo
Negro	1964	Feminino	021	Dores nas Articulações
Negro	1964	Feminino	021	Dor no Peito

Continua

Conclusão

Raça	Data de Nascimento	Sexo	CEP	Reclamação
Branco	1964	Masculino	021	Falta de Ar
Branco	1965	Feminino	021	Hipertensão
Branco	1964	Masculino	021	Dores nas Articulações
Branco	1964	Masculino	021	Febre
Branco	1967	Masculino	021	Vômito
Branco	1967	Masculino	021	Dor nas Costas

Fonte: adaptado de Ohm (2010)

O propósito das técnicas introduzidas é a preservação da privacidade, tentando ocasionar o anonimato de atributos que tornam um sujeito identificável. É essencial que, para a divulgação dos dados, estes passem por um contexto de anonimização, para que os sujeitos não sejam reidentificados facilmente.

4. Cenários de uso

A quebra da privacidade dos dados ocorre por meio de um ataque de uma pessoa maliciosa ou de um invasor a dados privados que são vulneráveis (BRITO; MACHADO, 2017). Essa pessoa pode associar os registros de dados a uma outra pessoa específica por meio de conhecimentos adquiridos anteriormente de outras fontes.

É possível vislumbrar alguns exemplos de situações capazes de violar a privacidade dos indivíduos no trabalho de Brito e Machado (2017); segundo esses autores, o adversário pode:

- saber que a vítima mora ao lado de sua residência, assim ele pode inferir informações como endereço, CEP, gênero da vítima etc.;
- utilizar dados de serviços baseados em localização, como o "*check-in*" em uma rede social realizado por uma vítima em uma determinada localização;
- ter acesso a dados abertos de uma vítima, caso ela seja funcionária de órgãos públicos, por exemplo.

Ainda de acordo com Brito e Machado (2017), os tipos de ataques podem ser classificados como:

- ataque de ligação ao registro - o objetivo do invasor é identificar novamente o indivíduo através de atributos semi-identificadores no registro, em um pequeno grupo ou em uma pessoa em particular;
- ataque de ligação ao atributo - o invasor é capaz de inferir os valores dos atributos sensíveis pertencentes a uma determinada pessoa, baseando-se no conjunto de dados sensíveis associados ao grupo ao qual a pessoa pertence.
- ataque de ligação à tabela - a partir dos ataques de ligação ao registro e de ligação ao atributo o invasor verifica que o registro da vítima está publicado, mas ele deseja inferir com firmeza a presença ou não dos dados da vítima nos registros publicados.
- ataque probabilístico - seu foco não está nos registros, atributos ou tabelas. O invasor pode inferir as informações sensíveis da pessoa, entretanto pode, também, alterar sua intenção considerando o conhecimento obtido após acessar os registros publicados.

A maioria das vezes em que ocorrem os ataques a dados do usuário, os invasores de maneira empírica conseguem detectar o usuário em específico através de correlação entre dados expostos sem proteção. Assim, deve-se tomar cuidado com os atributos dos usuários que são expostos. Faz-se necessário algum tipo de proteção para esses dados, tendo em vista a prevenção de possíveis ataques. Segundo a LGPD, os dados dos usuários devem ser protegidos e estes devem estar cientes de como os dados serão utilizados. Sendo assim, a disponibilização aberta, sem o consentimento do usuário, acarreta a violação da privacidade deste indivíduo.

5. Desafios à privacidade de dados

Conforme discutido neste trabalho, a LGPD trata a privacidade de dados como regra, sendo este um dos objetivos principais dessa legislação. Pode-se destacar que a privacidade de dados está presente no Sistema de Informação no Mundo Aberto e na Visão Sociotécnica de Sistemas de Informação, que são dois dos quatro temas dos grandes desafios de

pesquisa para Sistemas de Informação no Brasil de 2016 a 2026 (BOSCARIOLI; ARAÚJO; MACIEL, 2017).

A LGPD tem entre seus principais desafios a conscientização da sociedade de que "dado pessoal" é um bem de valor que deve ser protegido. Raminelli e Rodegheri (2016) destacam que o grande desafio que se coloca à frente dos cidadãos diz respeito à privacidade dos dados pessoais a ser proporcionada por empresas ou, até mesmo, pelos governos, que não podem mais se esquivar da obrigação de conhecer e respeitar a LGPD no que tange a coleta, armazenamento, processamento e compartilhamento de dados.

A conformidade com as leis gerais de proteção de dados, portanto, requer tecnologia, infraestrutura e pessoal especializado, para que os dados sejam tratados de forma lícita, justa e responsável em relação aos seus titulares. Prevê, além disso, o princípio da responsabilização através de acompanhamento das atividades de processamento pelas autoridades designadas, que poderão aplicar sanções quando houver descumprimento da lei.

Vale destacar que alguns países possuem centros de dados criados por meio de parcerias entre governo, universidades e institutos de pesquisa para processar e prover acesso a dados anonimizados de forma segura e controlada para pesquisas de interesse público (MILLER, 2020).

6. Considerações finais

Este capítulo trouxe algumas definições associadas à privacidade de dados no cenário jurídico e no contexto computacional, resgatando o conceito histórico sobre privacidade e o entendimento desse conceito na conjuntura atual em que vivemos. Verifica-se que a privacidade, no mundo atual hiperconectado, tem um significado muito mais amplo do que o simples direito de "estar só" definido por Warren e Brandeis. A tecnologia ganhou uma importância maior e se tornou essencial no sistema social e econômico. Hoje em dia, os dados são os protagonistas da Sociedade da Informação.

A LGPD muda a forma de compartilhamento de dados pessoais com relação à privacidade dos indivíduos. Trata-se de um desafio para todos os envolvidos no tema ao tempo que coloca o Brasil em nível regulatório compatível mundialmente.

Observa-se que a proteção e a privacidade das pessoas relativas ao tratamento de dados pessoais são consideradas como um direito fundamental na Constituição Federal Brasileira. Esse direito é complementado pela LGPD. É importante ressaltar que a Lei reforça a confiança no processamento e tratamento de dados pessoais de empresas públicas e privadas, porém a adequação é um desafio tecnológico para tais empresas.

No que se refere à privacidade de dados e à Lei de Proteção de Dados Brasileira, verifica-se que as pesquisas estão ainda no início, tendo sido o tema deste capítulo relevante e atual. Contudo, a literatura ainda carece de mais pesquisas associadas à Lei e resultados concretos relevantes.

Referências

ALVES, F. da M. LGPD ou LPDP: como denominar a lei de proteção de dados brasileira. [*S.l.*]. AB2L, 2018. Disponível em: https://ab2l.org.br/lgpd-ou-lpdp-como-denominiar-a-lei-de-protecao-de-dados-brasileira/. Acesso em: 09 maio 2020.

BAGNOLI, V. The big data relevant market. **Concorrenza e Mercato**, v. 23, 2016. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3064792. Acesso em: 18 jun. 2020.

BASSO, T; MATSUNAGA, R.; MORAES, R; ANTUNES, N. **Challenges on anonymity, privacy, and big data**. *In:* SEVENTH LATIN-AMERICAN SYMPOSIUM ON DEPENDABLE COMPUTING (LADC), 2016, Cali. **Anais** [...]. [*S.l.*]: IEEE, 2016. p. 164-171.

BOSCARIOLI, C.; ARAÚJO, R. M.; MACIEL, R. S. P. (ed.). **I GranDSI-BR – Grand Research Challenges in Information Systems in Brazil 2016-2026**. Porto Alegre: SBC, 2017. 184p. ISBN 978-85-7669-384-0. Disponível em: http://www2.sbc.org.br/ce-si/arquivos/GranDSI-BR_Ebook-Final.pdf. Acesso em: 20 jun. 2020.

BRANCO JUNIOR, E. C.; MACHADO, J. C.; MONTEIRO, J. M. Estratégias para Proteção da Privacidade de Dados Armazenados na Nuvem. *In:* SIMPÓSIO BRASILEIRO DE BANCO DE DADOS. TÓPICOS EM GERENCIAMENTO DE DADOS E INFORMAÇÕES, 29., 2014, Curitiba. **Anais** [...]. Porto Alegre: SBC, 2014. p. 46-75.

BRASIL. Comissão de Assuntos Econômicos. **Parecer Sobre o Projeto de Lei da Câmara nº 53, de 2018**. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014. Brasília, DF: Senado Federal, 2018b. Disponível em: https://legis.senado.leg.br/sdleg-getter/documento?dm=7751566&ts=153479 80 20516&disposition=inline&ts=1534798020516. Acesso em: 9 maio 2020.

BRASIL. **Constituição da República Federativa do Brasil de 1988**. Brasília, DF: Presidência da República, 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 20 maio 2020.

BRASIL. **Lei nº 13.709, 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018a. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm. Acesso em: 1 maio 2020.

BRASIL. **Lei nº 13.853, de 8 de julho de 2019**. Altera a Lei nº 13.709, de 14 de agosto de 2018, para dispor sobre a proteção de dados pessoais e para criar a Autoridade Nacional de Proteção de Dados; e dá outras providências. Brasília, DF: Presidência da República, 2019. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2019/Lei/L13853.htm#art2. Acesso em: 9 maio 2020.

BRITO, F.; MACHADO, J. Preservação de Privacidade de Dados: Fundamentos, Técnicas e Aplicações. *In:* PIRES, P. F.; DELICATO, F. C.; SILVEIRA, I. F. (ed.). **Jornadas de Atualização em Informática 2017.** Porto Alegre: SBC, 2017. Cap. 3. Disponível em: https://www.researchgate.net/publication/318726149. Acesso em: 20 maio 2020.

CAMENISCH, J.; FISCHER-HÜBNER, S.; RANNENBERG, K. **Privacy and identity management for life**. [*S.l.*]: Springer, 2011.

CANCELIER, M. V. de L. O direito à privacidade hoje: perspectiva histórica e o cenário brasileiro. **Sequência**, Florianópolis, n. 76, p. 213-239, 2017. Disponível em:

https://periodicos.ufsc.br/index.php/sequencia/article/view/2177-7055.2017v-38n76p213. Acesso em: 20 maio 2020.

CARLONI, G. **Privacidade e Inovação na Era do Big Data**. 2013. Trabalho de Conclusão de Curso (Graduação em Direito) – Fundação Getúlio Vargas, Rio de Janeiro, 2013. Disponível em: https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/12664/Giovanna%20Louise%20Bodin%20de%20Saint-Ange%20Comn%c3%a8ne%20Carloni.pdf?sequence=1&isAllowed=y. Acesso em: 18 jun. 2020.

CASTRO, C. S. **Direito da Informática, Privacidade e Dados Pessoais**. Coimbra: Editora Almedina, 2005.

DE MAURO, A.; GRECO, M.; GRIMALDI, M. A formal definition of Big Data based on its Essentials Features. **Library Review,** v. 65, n. 3, p. 122-135, 2016. DOI: 10.1108/LR-06-2015-0061. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/LR-06-2015-0061/full/html. Acesso em: 9 set. 2019.

ISAAK, J; HANNA, M. J. **User Data Privacy:** Facebook, Cambridge Analytica, and Privacy Protection, **Computer,** v. 51, n. 8, p. 56-59, 2018. DOI: 10.1109/MC.2018.3191268.

LEE, N. **Facebook Nation**. New York, NY: Springer New York, 2013.

LEONARDI, M. **Tutela e privacidade na internet**. São Paulo: Editora Saraiva, 2011.

MACHADO, R.; KREUTZ, D.; PAZ, G; RODRIGUES, G. Vazamentos de Dados: Histórico, Impacto Socioeconômico e as Novas Leis de Proteção de Dados. *In:* ESCOLA REGIONAL DE REDES DE COMPUTADORES (ERRC), 17., 2019, Alegrete. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 154-159. Disponível em: https://sol.sbc.org.br/index.php/errc/article/view/9230/9133. Acesso em: 18 jun. 2020.

MENEZES NETO, E. J. de; MORAIS, J. L. B. de; BEZERRA, T. J. de S. L. O projeto de lei de proteção de dados pessoais (PL 5276/2016) no mundo do big data: o fenômeno da dataveillance em relação à utilização de metadados e seu impacto nos direitos humanos. **Rev. Bras. Polít. Públicas**, Brasília, v. 7, n. 3, p. 184-198, 2017.

MILLER, M. Coronavirus surveillance concerns ramp up pressure to pass federal privacy law. **The Hill**, Washington DC, 9 Apr. 2020. Disponível em: https://

thehill.com/policy/cybersecurity/492072-coronavirus-surveillance-concerns--ramp-up-pressure-to-pass-federal. Acesso em: 20 jun. 2020.

MOONEY, S. J.; PEJAVER, V. Big data in public health: terminology, machine learning, and privacy. **AnnuRevPublic Health**, v. 39, p. 95-112, 2018.

MORAES, M. C. B. de (org.). Apresentação do autor e da obra. *In:* RODOTÀ, S. **A vida na sociedade de vigilância:** A privacidade hoje. Rio de Janeiro: Renovar, 2008. p. 1-12.

MULHOLLAND, C. Dados pessoais sensíveis e a tutela de direitos fundamentais: uma análise à luz da lei geral de proteção de dados (Lei 13.709/18). **Revista de Direitos e Garantias Fundamentais**, v. 19, p. 159-180, 2018.

NEGRI, S. M. C. de Á.; KORKMAZ, M. R. D. C. R. A normatividade dos dados sensíveis na Lei Geral de Proteção De Dados: ampliação conceitual e proteção da pessoa humana. **Rev. de Direito, Governança e Novas Tecnologias**, Goiânia, v. 5, n. 1, p. 63-85, 2019. e-ISSN 2526-0049.

OHM, P. Broken promises of privacy: Responding to the surprising failure of anonymization. **UCLA Law Review,** v. 57, p. 1701-1777, 2010.

PEZZI, A. P. J. **A necessidade de proteção dos dados pessoais nos arquivos de consumo: em busca da concretização do direito à privacidade**. 2007. Dissertação (Mestrado em Direito) – Universidade do Vale do Rio Sinos, São Leopoldo, 2007.

RAMINELLI, F. P.; RODEGHERI, L. B. A Proteção de Dados Pessoais na Internet no Brasil: Análise de decisões proferidas pelo Supremo tribunal Federal. **Revista Cadernos do Programa de Pós-Graduação em Direito PPGDir./UFRGS**, v. 11, n. 2, 2016. Disponível em: http://seer.ufrgs.br/ppgdir/article/view/61960/39936. Acesso em: 19 jun. 2020.

SERPRO. **Quais são os seus direitos?.** Brasília, DF: SERPRO, 2019. Disponível em: https://www.serpro.gov.br/lgpd/cidadao/quais-sao-os-seus-direitos-lgpd. Acesso em: 9 maio 2020.

STIVAL, J. O anteprojeto brasileiro de lei de proteção dos dados pessoais e os dados dos Trabalhadores na relação laboral. In: FINCATO, D. P. (org.). **Novas tecnologias, processo e relações de trabalho**. Porto Alegre: Sapiens, 2015. p. 129-145.

UE – UNIÃO EUROPEIA. **Regulamento nº 679/2016, de 27 de abril de 2016**. Relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados e que revoga a Diretiva 95/46/CE (Regulamento Geral sobre a Proteção de Dados). Bruxelas: Parlamento Europeu: Conselho da União Europeia, 2016. Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/HTML/?uri=CELEX:32016R0679&from=PT. Acesso em: 18 jun. 2020.

VIMERCATI, S. de C.; FORESTI, S.; LIVRAGA, G.; SAMARATI, P. Data Privacy: Definicitions and Techiniques. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, v. 20, v. 6, p. 793-817, 2012.

WARREN, D. S.; BRANDEIS, L. D. **The Right to Privacy**. **Harvard Law Review**, v. 4, n. 5, p. 193-220, 1890. Disponível em: https://www.jstor.org/stable/1321160?-seq=2#page_scan_tab_contents. Acesso em: 18 maio 2020.

ZANATTA, R. A Proteção de Dados entre Leis, Códigos e Programação: os limites do Marco Civil da Internet. *In:* DE LUCCA, N.; SIMÃO FILHO, A.; LIMA, C. **Direito e Internet III**: Marco Civil da Internet. São Paulo: QuartierLatin, 2015. p. 447-470.

SOBRE AS ORGANIZADORAS

Crishane Azevedo Freire

Doutora em Ciência da Computação pela Universidade Federal de Pernambuco (2014). Professora titular do Instituto Federal da Paraíba, com atuação na linha de pesquisa de Gerenciamento e Desenvolvimento de Software do Programa de Mestrado Profissional em Tecnologia da Informação do IFPB. Tem experiência profissional na área de Ciência da Computação e interesse em temas relacionados a banco de dados, big data e integração de dados.

Damires Yluska de Souza Fernandes

Professora titular do Instituto Federal da Paraíba e permanente do Programa de Mestrado Profissional em Tecnologia da Informação do IFPB – Campus João Pessoa. Doutora em Ciência da Computação pela Universidade Federal de Pernambuco (2009), com ênfase na área de Banco de Dados. Tem experiência profissional em Ciência da Computação e atua principalmente em temas relacionados a gerenciamento de dados em ambientes diversos e distribuídos, big data, análise e mineração de dados, integração de dados e uso de semântica.

SOBRE OS AUTORES

Aline Priscila Araújo de Morais

Bacharel em Sistemas de Informação pela UFPB – Campus IV e discente do Programa de Mestrado Profissional em Tecnologia da Informação pelo IFPB – Campus João Pessoa, com interesse em Engenharia de Software, Banco de Dados, Informática na Educação e Gamificação.

Alysson Messias da Silva

Discente do Programa de Mestrado Profissional em Tecnologia da Informação pelo IFPB – Campus João Pessoa. Funcionário da Dataprev, atuante em Engenharia de Software, com interesse em Inteligência Artificial e Aprendizado de Máquina.

Amanda Days Ramos Novo

Discente do Programa de Mestrado Profissional em Tecnologia da Informação pelo IFPB – Campus João Pessoa. Graduada em Licenciatura em Computação – UFPB (2018). Tecnóloga em Análise e Desenvolvimento de Software (2010) pela Faculdade de Tecnologia Idez. Atuante em Engenharia de Software com interesse em Banco de Dados.

Anthony Martins Araújo

Aluno especial do Programa de Mestrado Profissional em Tecnologia da Informação pelo IFPB – Campus João Pessoa. Servidor público cedido ao Tribunal Regional Eleitoral da Paraíba. Bacharel e pós-graduado pela Universidade Católica de Brasília, com interesse em Ciência de Dados.

Ayrton Douglas Rodrigues Herculano

Discente do Programa de Mestrado Profissional em Tecnologia da Informação pelo IFPB –Campus João Pessoa. Tecnólogo em Sistemas para Internet pelo IFPB – Campus João Pessoa. Desenvolvedor de softwares na E-ticons, com interesse em Inteligência Artificial e Aprendizado de Máquina.

Helton Souza Lima

Discente do Programa de Mestrado Profissional em Tecnologia da Informação pelo IFPB –Campus João Pessoa. Funcionário da Dataprev, com MBA em Gerenciamento de Projetos e Gestão Estratégica de TI.

Joerverson Barbosa Santos

Discente do Programa de Mestrado Profissional em Tecnologia da Informação pelo IFPB –Campus João Pessoa. Funcionário do Polo de Inovação Assert IFPB. Atuante em Desenvolvimento de softwares, com interesse em linguagens de programação, processos e novas tecnologias que possam melhorar a vida e o processo executado.

Karine Heloise Félix de Sousa

Discente do Programa de Mestrado Profissional em Tecnologia da Informação pelo IFPB –Campus João Pessoa. Licenciada em Computação pela Universidade Federal da Paraíba (2019). Especialista em Gestão de Projeto em TI pelo Centro de Inovação VincIT (2020). Bacharelado em Direito pela Faculdade Paraibana (2010) e especialista em Processual Civil pelo Centro Universitário Internacional (2013).

Rafael Anderson de Lima Ramos

Discente do Programa de Mestrado Profissional em Tecnologia da Informação pelo IFPB –Campus João Pessoa. Colaborador externo do Polo de Inovação do IFPB, atuante na área de Qualidade de Software, com interesse em testes ágeis e metodologias de desenvolvimento.

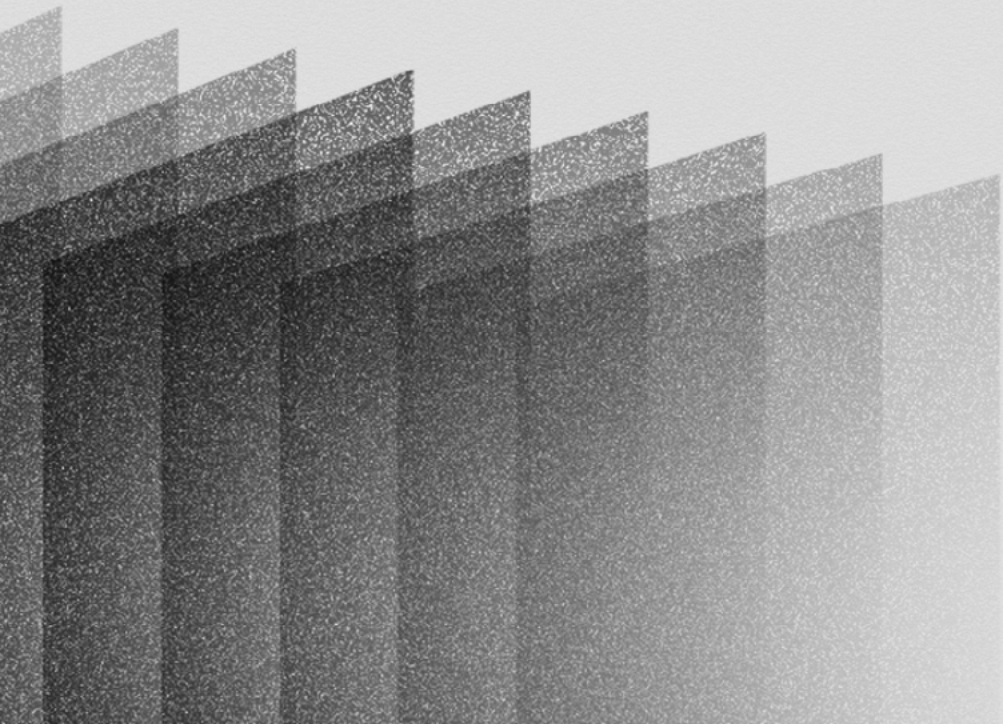
Victon Malcolm Rodrigues dos Santos

Discente do Programa de Mestrado Profissional em Tecnologia da Informação pelo IFPB –Campus João Pessoa, com especialização em Desenvolvimento de Aplicações para Dispositivos Móveis e Testes. Funcionário da Dataprev, atuante em Análise de Sistemas e Testes de Software, com interesse em Análise de Dados e Inteligência Artificial.

Wescley Paoli Alcantara de Sousa

Discente do Programa de Mestrado Profissional em Tecnologia da Informação pelo IFPB –Campus João Pessoa. Servidor público do Tribunal Regional Eleitoral da Paraíba, atuante na área de Banco de Dados, com interesse em Big Data e Ciência de Dados.

A pesquisa abre caminhos e coloca o pesquisador como protagonista da narrativa do seu estudo. As diferentes perspectivas que envolvem a temática sobre gerenciamento de dados vão além do que se costuma aprender sobre o que são dados e de que forma eles podem ser utilizados. A presente coletânea, desenvolvida por estudantes do mestrado profissional em Tecnologia da Informação, campus João Pessoa - IFPB, apresenta princípios e discussões sobre técnicas, tecnologias, cenários de utilização, desafios e tendências relacionados a diversos temas na área de gerenciamento de dados. Os trabalhos constituem uma excelente oportunidade para familiarização a respeito de panoramas apresentados tanto para acadêmicos quanto para profissionais da área de Tecnologia da Informação.



editora
IFPB