Análise Preditiva

Predição é poder: para o bem ou para o mal.

Ely Batista do Rêgo Júnior

Resumo

A Análise Preditiva, junto com toda a tecnologia do Big Data, tem chamado muita atenção da mídia e principalmente das empresas públicas e privadas. O crescente aumento do volume de dados gerados a cada dia, com uma grande variedade, o armazenamento desse grande volume de dados, com maior velocidade e a veracidade desses dados, faz com que se extraia valor de tudo isso usando a Mineração de Dados. Analisando as informações armazenadas, podemos, por meio da Análise Preditiva, prever padrões, tendências e comportamentos futuros, proporcionando aos gestores uma tomada de decisão baseada em fatos. Este trabalho tem como intuito, demonstrar as estruturas básicas do Big Data, a Análise Preditiva e sua forma de extrair informação, a importância e as vantagens da utilização da Análise Preditiva nos negócios, mas também, demonstrar os problemas que podem acarretar pelo uso da tecnologia ao se entregar decisões a um algoritmo.

Palavras-chave: Análise Preditiva, Big Data, Mineração de Dados, Data Mining

Introdução

O Big Data já está entre nós há algum tempo e o tema deste artigo é a Análise Preditiva, vamos mostrar o seu poder de extrair informações de uma grande massa de dados e os benefícios e danos que podem ser causados com o uso dessa tecnologia.

A metodologia usada neste artigo constituiu de análises bibliográficas, com o uso de livros físicos e/ou digitais e web sites que definiram e argumentaram o estudo e pesquisa sobre a Análise Preditiva e suas consequências.

A análise preditiva além de nos beneficiar como consumidores, ela ajuda as organizações como uma arma competitiva, deixando as empresas à frente da concorrência. Com a identificação de padrões e tendências do mercado, a

empresa consegue tomar decisões antecipadamente. Como exemplo podemos citar a identificação do aumento de consumidores em uma determinada região e, assim, preparar uma estratégia de expansão para a área identificada. Podemos identificar o interesse por produtos sustentáveis e decidir pelo desenvolvimento de uma linha desses produtos antes dos concorrentes.

Para Braga (2005, p.11)**:** Para atingir estes objetivos não bastam as ferramentas genéricas de CRM (Client Relationship Management), ERP (Enterprise Resources Planning) ou BI (Business Intelligence), mas também da capacidade analítica para identificação de padrões e predição a partir dos dados estratégicos de uma organização. Analistas de "mineração de dados" desenvolvem dois tipos de modelos: preditivo e descritivos. Este artigo se concentra no modelo preditivo e suas consequências.

Desenvolvimento

A principal matéria-prima da Ciência de Dados ou Big Data são: o dado, a informação e o conhecimento. De acordo com Amaral (2016, p.3), "Dados são fatos coletados e normalmente armazenados. Informação é o dado analisado e com algum significado. O conhecimento é a informação interpretada, entendida e aplicada para um fim". Ele define ciência de dados como: "Os processos, modelos e tecnologias que estudam os dados durante todo o seu ciclo de vida: da produção ao descarte", conforme é mostrado na figura 1.

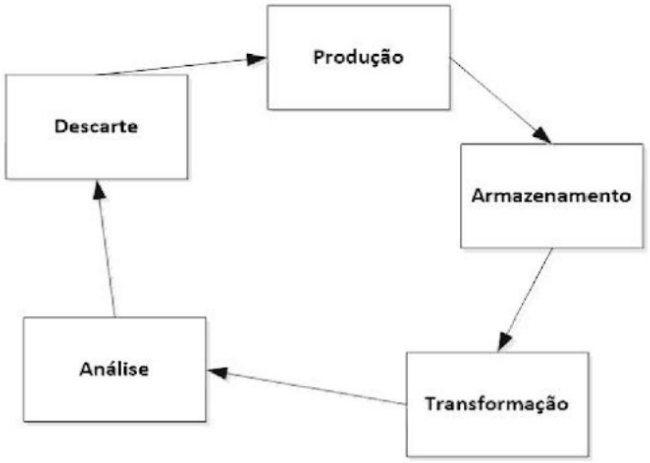

Figura 1 - Fonte: (Amaral, p.6)

Ainda segundo Amaral (2016, p.5), esses dados desde a sua produção até o descarte, pode passar por uma série de etapas. Eles podem não sofrer qualquer tipo de transformação, como podem ser descartados imediatamente. Essas etapas que o dado passa, vai depender de sua natureza e de sua finalidade. Esses dados podem estar em qualquer formato: analógico, digital ou não eletrônico, que é o caso de dados impressos em papel.

Amaral (2016, p.5), diz que esse registro eletrônico de coisas que acontecem no dia-a-dia, que são persistidos, armazenados para reprodução ou análise é chamado de datafication. Esses dados podem estar dispersos, ser relativos às mais diversas situações e podem até parecer irrelevantes, mas podem nos dar alguma informação.

Os dados podem ser gerados por humanos ou por máquinas. Os dados gerados por humanos tem o conteúdo gerado a partir do pensamento de uma pessoa, onde a propriedade intelectual está ligada ao dado. Já os dados gerados por máquinas não precisam da intervenção humana, são gerados por máquinas, dispositivos, processos de computadores, aplicações, etc.

Os dados digitais são gerados de todas as formas e de todas as maneiras, depois disso, ele precisa ser preservado em alguma estrutura específica, como um banco de dados relacional, XML, texto plano, etc., para poder ser utilizado futuramente. Depois disso o dado passa por processos de transformação por causa das diferenças entre as estruturas de dados. Essa transformação faz com que se modifique o modelo que o dado foi armazenado, para um modelo ideal para uso em Big Data. Um exemplo disso são os processos de ETL - Extract, Transform and Load (Extração, transformação e

carga), que trata da sistematização do tratamento e limpeza dos dados oriundos dos diversos sistemas organizacionais e que são usados para a construção de Data Warehouses, que são depósitos de dados usados nas empresas para apoio à decisão.

Uma visão de todo o ambiente da Ciência de Dados é mostrado na figura 2.

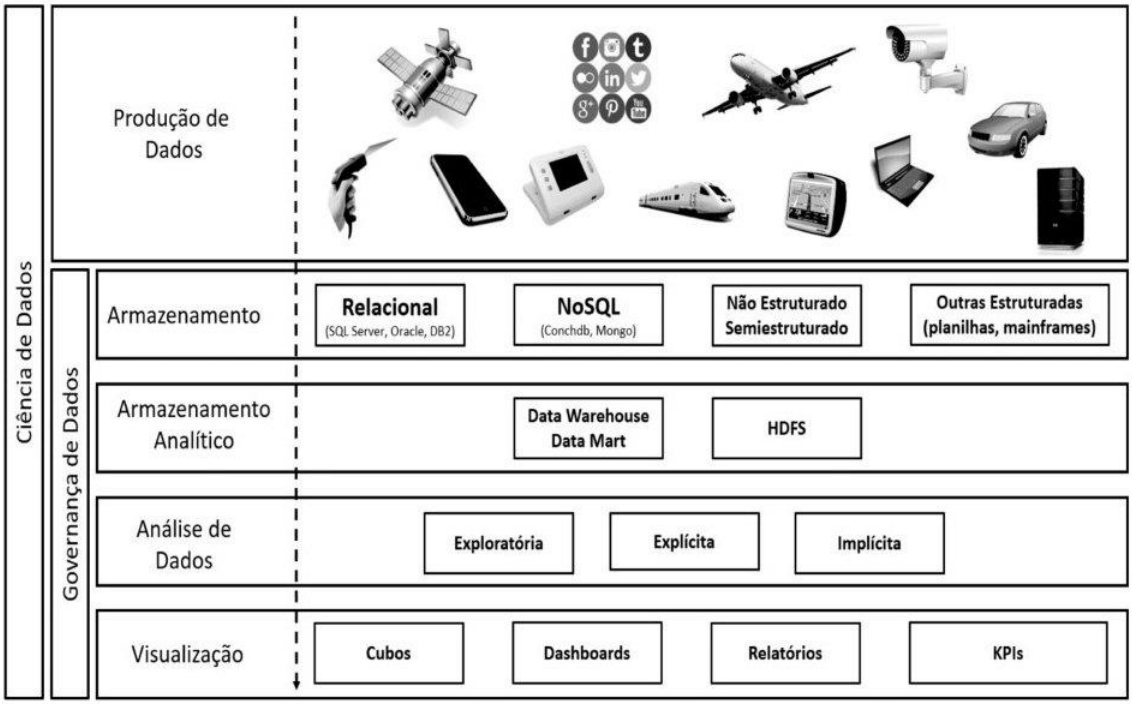

Figura 2 - Fonte: (Amaral, p.7)

Big Data

Segundo Marquesone (2016, p.7), o termo Big Data nos dá a impressão de que se fala de um grande volume de dados, mas isso não é sua única característica. Existem ainda duas propriedades que devemos considerar: a variedade e a velocidade dos dados. Essas três características são conhecidas como os 3 Vs do Big Data. Como é mostrado na figura 3.

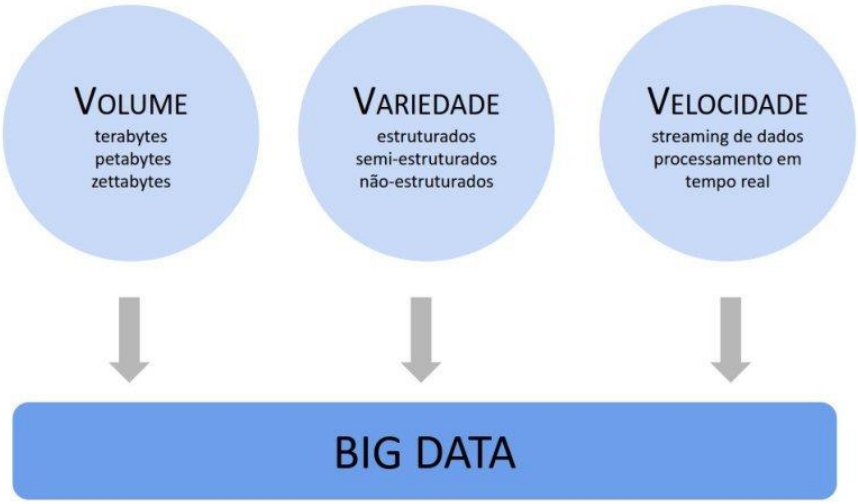

Figura 3 - Fonte: (Marquesone, p.8)

Amaral (2016, p.12) define Big Data como: "o fenômeno da massificação de elementos de produção e armazenamento de dados, bem como os processos e tecnologias, para extraí-los e analisá-los".

Já de acordo com Hurwitz (2013, p.10), grandes dados são definidos como qualquer tipo de fonte de dados que tenha pelo menos três características compartilhadas:

- Volumes de dados extremamente grandes
- Velocidade de dados extremamente alta
- Variedade de dados extremamente ampla

Os autores ainda explanam a importância dos dados da seguinte forma (tradução nossa):

> Os grandes dados são importantes porque permitem que as organizações coletem, armazenem, administram e manipulem grandes quantidades de dados à velocidade certa, no momento certo, para obter os insights adequados. Mas antes de aprofundar os detalhes dos grandes dados, é importante analisar a evolução do gerenciamento de dados e como ele levou a grandes dados. Os grandes dados não são uma tecnologia autônoma; em vez disso, é uma combinação dos últimos 50 anos de evolução tecnológica. (Hurwitz, 2013, p.10)

Marquesone (2016, p15), explica que: “alguns pesquisadores adotam os 5 Vs, onde são acrescentados os atributos valor e veracidade dos dados”. Onde valor está relacionado a importância que um dado pode ter em uma solução e a veracidade está relacionada à confiabilidade dos dados. Sendo Valor, o atributo mais importante do Big Data. Pois não adianta nada ter acesso a uma quantidade abundante de informação a cada segundo, se não puder gerar valor. Ainda pode ser encontrados outros V´s em alguns sites da internet, mas os 3 V´s formam a base do Big Data.

Mineração de Dados

A base da mineração de dados é um processo chamado descoberta de conhecimento em banco de dados, Knowledge Discovery in Database (KDD). Esse processo, composto por cinco etapas, utiliza conceitos de visualização, técnicas de inteligência artificial, métodos estatísticos e base de dados, dividindo-se em diversas etapas. Isso transforma os dados em informação e depois em conhecimento, alcançando depois a sabedoria. Tornando-se um valor imprescindível para as organizações.

De acordo com Júnior (2004, p.161), as etapas do KDD são: Seleção, Processamento, Transformação, Mineração de Dados e Análise/Assimilação. Como podemos ver na figura 4.

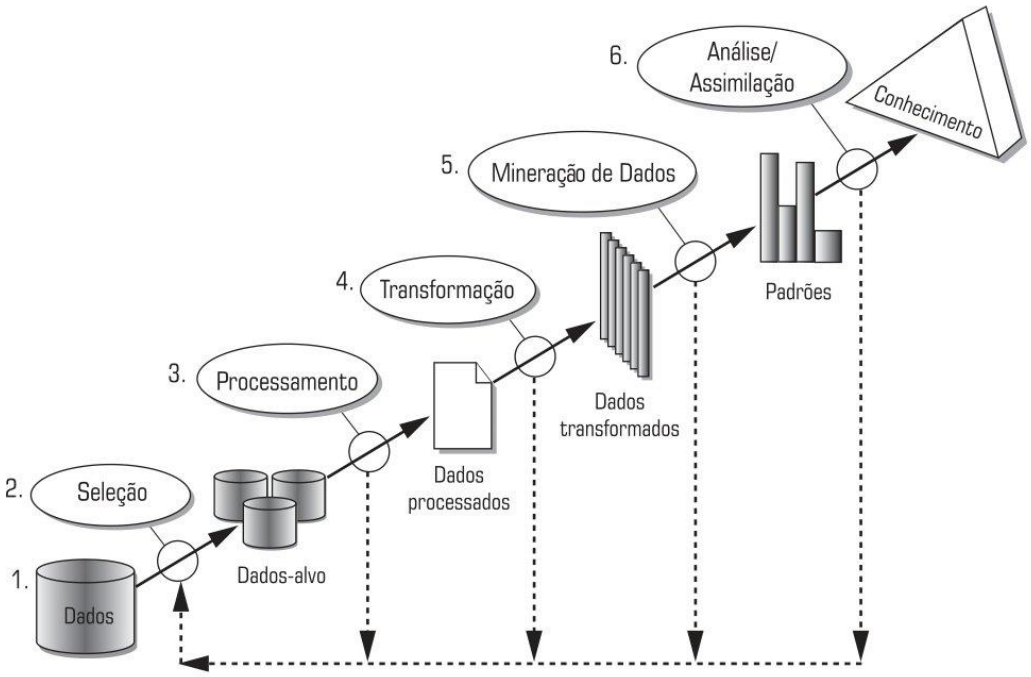

Figura 4 - Fonte: (Júnior, p.161)

Todas essas etapas são importantes no processo de KDD, mas a etapa de Mineração de Dados (data mining), coração do processo, é a que mais se destaca. O conhecimento alcançado pelas técnicas de Mineração de Dados é geralmente expresso na forma de regras e padrões.

Segundo Amaral (2016, p.3): “Minerar dados é a forma mais sofisticada, complexa e difícil de analisar dados. Em consequência, o resultado pode trazer insights sobre o negócio que nenhuma outra técnica seria capaz de produzir”.

A união da inteligência artificial, estatística e banco de dados é o que torna a mineração de dados possível. O objetivo principal do processo é fornecer as empresas informações para montar melhores estratégias de suporte, vendas e marketing.

Segundo Braga (2005, p.15), as etapas para um projeto de "Mineração de dados" são:

- Definição do problema
- Aquisição e Avaliação dos dados
- Extração de características e realce
- Plano de prototipagem, Prototipagem e Desenvolvimento do Modelo
- Avaliação do modelo

- Implementação
- Avaliação do retorno do investimento (pós-projeto)

Para Júnior (2004, p.164), existem vários modelos de descoberta de padrões ou de conhecimento. A escolha de um modelo vai depender do problema em particular. Mas os dois modelos mais difundidos são Classificação e Regras de Associação. Segundo o autor, na classificação, os dados são organizados em classes, baseando-se em propriedades (atributos) comuns entre um conjunto de objetos em uma base de dados. Como exemplo de uso: no diagnóstico médico e na avaliação de risco de crédito.

Júnior (2004), explica o processo da seguinte forma:

> As abordagens de classificação normalmente usam um conjunto de treinamento em que todos os objetos estão já associados a determinadas classes. Um algoritmo de classificação aprende regras de classificação do conjunto de treinamento. Um conjunto de testes analisa se as classificações pelo algoritmo batem com as classes reais dos objetos, o que é denominado classificação supervisionada. O modelo aprovado é então usado para classificar novos objetos. (Júnior, 2004, p.165)

O resultado do algoritmo de classificação normalmente é apresentado sob a forma de árvore de decisão. Um exemplo disso é mostrado na figura 5, onde as pessoas são classificadas em confiáveis ou não confiáveis para permissão de crédito, avaliando pelo grau de escolaridade e sua faixa de renda anual (Ra).

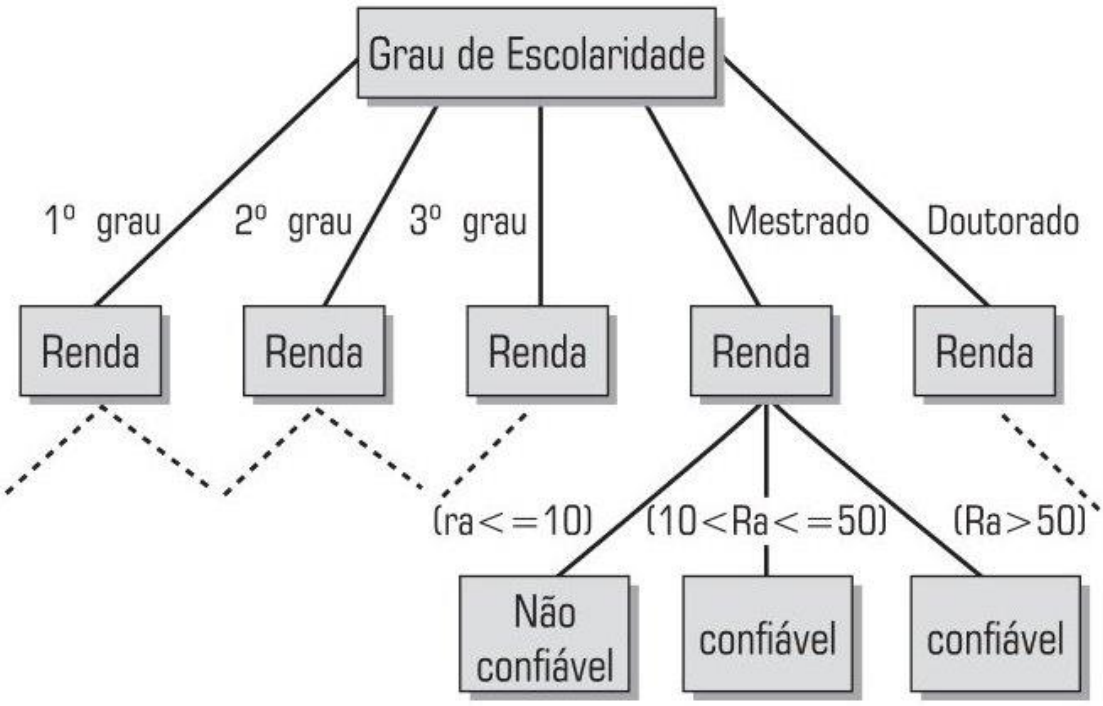

Figura 5 - Fonte: (Júnior, p.166)

O algoritmo também pode mostrar o resultado na forma "Se condição então classe" ou X, Y → Z, que são chamadas de Regras de Classificação. Um exemplo desse tipo de regra baseado na árvore de decisão acima é:

...Se (Grau de Escolaridade = Mestrado), (Ra <= 10) → não confiável
Se (Grau de Escolaridade = Mestrado), (10 < Ra <= 50) → confiável
Se (Grau de Escolaridade = Mestrado), (Ra > 50) → confiável...

Na Regra de Associação, o objetivo principal é realizar associações entre os itens, com o intuito de estabelecer correlações entre eles. Uma de suas típicas aplicações é a análise de transações de compras (market basket analysis), onde se examina padrões de compras de consumidores para determinar produtos que costumam ser adquiridos em conjunto.

Segundo Amaral (2016, p.120):

> A criação de regras de associação tem alto custo computacional, uma cesta de compras com centenas de itens pode gerar milhares de regras, por isso, algumas métricas são utilizadas para selecionar as regras mais valiosas. Destas métricas as mais importantes são o suporte, a confiança e a

força da regra. (Amaral, 2016, p.120)

O suporte representa a porcentagem de transações da base de dados que contêm os itens de A e B, indicando a relevância da mesma. A confiança representa à proporção de vezes que uma transação contendo o item A, também contém B. Já a força da regra é a soma do suporte mais confiança.

Para Gonçalves (2007). O problema da regra de associação consiste em: "Encontrar todas as regras de associação que possuam suporte e confiança maiores ou iguais, respectivamente, a um suporte mínimo (SupMin) e uma confiança mínima (ConfMin), especificados pelo usuário".

A regra de associação é definida como: "Se X então Y". Júnior (2004), explica essa regra:

> Diz-se que X é o antecedente da regra, enquanto Y é o seu consequente. Uma regra pode ter vários itens tanto no antecedente quanto no consequente. Um algoritmo baseado em regras de associação consiste em descobrir regras desse tipo entre os dados preparados para a mineração. (Júnior, 2004, p.167)

Um exemplo é:

{Pão} → {Leite, Manteiga}

A regra pode ser lida da seguinte forma: "Se Pão, então Leite e Manteiga". Indicando que clientes compraram os itens Pão, Leite e Manteiga Juntos.

O autor compara a regra de associação com as regras de classificação:

> Podemos perceber que uma regra de associação é então uma regra de classificação generalizada. A generalização consiste no fato de que Y, o consequente na regra de associação, é uma conjunção de termos com quaisquer atributos, enquanto que, nas regras de classificação, este é só um termo envolvendo unicamente o atributo de classificação. (Júnior, 2004, p.168)

O algoritmo vai gerar muitas regras pelas combinações que são possíveis de itens. Para evitar isso, dois parâmetros são passados para o

algoritmo: suporte mínimo e confiança mínima.

Na tabela abaixo mostramos um exemplo de transações de vendas de uma mercearia.

Transação	Leite	Manteiga	Pão
1	1	1	0
2	1	0	1
3	1	1	1
4	1	1	1
5	0	1	1
6	1	1	1
7	1	1	1
8	1	0	1
9	1	1	1
10	1	1	1

Figura 6 - Fonte: (Júnior, p.169)

Considerando-se a regra:

{Pão, Leite} → {Manteiga} /*Se Pão e Leite então Manteiga*/

Podemos constatar que os itens Pão, Leite e Manteiga foram comprados juntos em 6 das 10 transações. Então a regra tem suporte de 0,60 (6/10) ou 60%. O antecedente {Pão, Leite}, aparece em 8 transações, destas 8, 6 contêm o consequente {Manteiga}. Com isso, a confiança da regra é de 0,75 (6/8) ou 75%.

Júnior (2004, p.169), explica que:

> Os valores de suporte e confiança devem ser passados para o algoritmo de mineração. Desta forma, o usuário poderá aumentar ou diminuir o número de regras geradas. Sintetizando, o objetivo é procurar regras expressivas, aquelas que satisfazem as duas condições: é frequente, ou seja, tem suporte acima de um mínimo considerado aceitável; a

> correlação entre antecedente e consequente, medida pela confiança, é forte ou acima de um mínimo considerado aceitável.

A representação completa da regra é:

{Pão, Leite} → {Manteiga} [0,60 0,75]

Que é interpretada da seguinte forma: 60% dos clientes compraram pão, leite e manteiga e 75% dos clientes que compraram pão e leite, também compraram manteiga.

Com isso podemos pensar em: Quais regras possuem pão como antecedente? Quais regras possuem manteiga como consequente? Essas questões ajudam a descobrir como aumentar as vendas de um produto, fazer vendas casadas a outros produtos, ver qual a melhor organização das prateleiras ou verificar quais itens iriam ter suas vendas influenciadas por causa de um certo item antecedente.

Exemplo de Modelo Preditivo em Python

Python é uma linguagem cada vez mais popular entre os cientistas de dados. Ela é simples e fácil de aprender, portátil e extensível e se integra bem com várias bases de dados e ferramentas. Tem grande intensidade computacional e poderosas bibliotecas de análise de dados. Segundo Jain (2016), “serve para realizar o ciclo de vida completo de qualquer projeto de Data Science, incluindo leitura, análise, visualização e finalmente previsões".

A validação é o processo de testar o código nos dados reais, para depois ser usado em ambiente de produção. Para isso, é necessário sempre dividir os dados históricos em dois grupos: treino e teste. Pegamos os dados de exemplo de transações de vendas de uma mercearia e colocamos em um arquivo no formato .CSV. Para exemplo, junto com os dados de teste, adicionamos o campo {ComprouOsTrês}. Esse campo informa se foi comprado os três produtos (Leite, Manteiga e Pão). Se foi comprado, temos um Y, se não

foi, temos um N. Esses dados são usados para treinar o algoritmo. A baixo o exemplo do conteúdo do arquivo de exemplo datasetSupermercado_treino.csv.

```
Transação,Leite,Manteiga,Pão,ComprouOsTrês
1,1,1,0,N
2,1,0,1,N
3,1,1,1,Y
4,1,1,1,Y
5,0,1,1,N
6,1,1,1,Y
7,1,1,1,Y
8,1,0,1,N
9,1,1,1,Y
10,1,1,1,Y
```

O código em Python mostrado a seguir, foi baseado no código de Jain (2016) e modificado para acessar os dados mostrados no exemplo acima.

```
import pandas as pd
import numpy as np

#Lendo o conjunto de dados em um dataframe do Pandas
df = pd.read_csv("datasetSupermercado_treino.csv")


#Importa os modelos da biblioteca scikit learn:

from sklearn.linear_model import LogisticRegression
from sklearn.cross_validation import KFold #For K-fold cross validation
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn import metrics

#Função genérica para fazer um modelo de classificação para avaliar performanc
e
def classification_model(model, data, predictors, outcome):
  #Acerta o modelo:
  model.fit(data[predictors],data[outcome])

  #Faz previsões nos dados de treino:
  predictions = model.predict(data[predictors])

  #Mostra a acurácia
  accuracy = metrics.accuracy_score(predictions,data[outcome])
  print("Precisão : %s" % "{0:.3%}".format(accuracy))

  #Performa validação cruzada k-fold com 5 folds
  kf = KFold(data.shape[0], n_folds=5)
  error = []
  for train, test in kf:
    # Filtra dados de treino
    train_predictors = (data[predictors].iloc[train,:])

    # O alvo que estamos usando para treinar o algoritmo
    train_target =  data[outcome].iloc[train]

    # Treinando o algoritmo com previsores e alvo
```

```
39.     model.fit(train_predictors, train_target)
40.
41.     #Grava erros de cada loop de validação cruzada
42.     error.append(model.score(data[predictors].iloc[test,:], data[outcome].iloc
    [test]))
43.
44.   print("Pontuação de validação cruzada : %s" % "{0:.3%}".format(np.mean(error
    )))
45.
46.   #Acerta de novo o modelo para que possa se referir fora da função
47.   model.fit(data[predictors],data[outcome])
48.
49.
50. outcome_var = 'ComprouOsTrês'
51. model = LogisticRegression()
52. predictor_var = ['Pão']
53. classification_model(model, df,predictor_var,outcome_var)
```

Colocamos o Pão como antecedente (predictor_var) e o resultado (outcome_var) o campo ComprouOsTrês.

A saída desse programa é mostrada a baixo:

```
Precisão: 70.000%
Pontuação de validação cruzada: 60.000%
```

Ela nos mostra dois valores: Precisão e Pontuação de validação cruzada. A precisão determina o nível de precisão nos dados de teste, já a validação cruzada segundo a Wikipédia significa:

> Uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados. Esta técnica é amplamente empregada em problemas onde o objetivo da modelagem é a predição. Busca-se então estimar o quão preciso é este modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados. (Wikipédia, 2017)

Utilidades do Big Data

Com isso tudo que vimos, vemos que a analise preditiva pode ser usada para muitas coisas, como por exemplo: distribuição de conteúdo personalizado, otimização de campanhas de Marketing, etc. Marquesone (2016, p.27), nos

mostra em uma tabela algumas áreas e exemplos de aplicação do Big Data.

Área	Onde Big Data está sendo aplicado
Cuidados da saúde e medicina	Monitoramento de pacientes em tempo real; Análise de dados de redes sociais para descobertas de pandemias; Análise de padrões de doenças; Extração de informação em imagens médicas; Descoberta e desenvolvimento de medicamentos; Análise de dados genéticos.
Serviços financeiros	Análise de risco; Detecção de fraude; Programas de lealdade; Venda cruzada.
Setor público	Digitalização dos dados; Detecção de fraude e ameaças; Vigilância por vídeo; Manutenção preventiva de veículos públicos; Otimização de rotas no transporte público.
Telecomunicação	Análise de registro de chamadas; Alocação de banda em tempo real; Desenvolvimento de novos produtos; Planejamento da rede; Análise de *churn*; Gerenciamento de fraude; Monitoramento de equipamentos.
Varejo	Análise de sentimento; Segmentação de mercado e cliente; Marketing personalizado; Previsão de demanda; Precificação dinâmica.

Figura 7 - Fonte: (Marquesone, p.27)

Um dos exemplos mais famosos da aplicabilidade do Big Data, segundo Júnior (2004), é o famoso exemplo das fraldas e da cerveja. Onde:

> Utilizando Mineração de Dados, uma rede de supermercados descobriu que a maioria dos pais que iam comprar fraldas para seus filhos levava cerveja. O pessoal de marketing, muito inteligente, colocou a cerveja e as fraldas próximas, com batata fritas entre elas, aumentando consideravelmente a venda dos três produtos. Muitas vezes, o cliente nem pretende levar a cerveja, mas o faz quando vê a tentação do lado das fraldas. (Júnior, 2004, p.8)

Outro bom exemplo (IBM) é o Departamento de Recuperação de Ativos e Cooperação Jurídica Internacional (DRCI) da Secretaria Nacional de Justiça é responsável pela recuperação dos rendimentos provenientes da corrupção, do crime organizado, do tráfico de drogas e da lavagem de dinheiro. Esse departamento, em 2007, criou o Laboratório de Tecnologia contra Lavagem de Dinheiro (LAB-LD) para apoiar as investigações complexas sobre corrupção e

lavagem de dinheiro. Nesses laboratórios, uma grande quantidade de dados é analisada para descobrir e congelar os ativos ilícitos. Ajudando as autoridades a tomar medidas legais contra suspeitos de crime no Brasil. Esses laboratórios foram replicados em outros órgãos estaduais e federais. O conjunto desses laboratórios forma a REDE-LAB.

No passado, as investigações exigiam analistas altamente qualificados que gastavam milhares de horas debruçados em planilhas, e-mails e publicações em redes sociais. Segundo a publicação da IBM:

> A análise de todos esses dados sem as ferramentas adequadas é complexa e demorada. Por exemplo, em uma grande investigação há alguns anos, a análise de centenas de terabytes de dados levou dez meses e milhares de pessoas-hora, pois nossos investigadores tiveram que passar por centenas de unidades de disco rígido manualmente. (IBM Software Information Management, p.2)

Depois que a REDE-LAB começou a trabalhar com o IBM Watson Explore, uma plataforma de Big Data e Analytics da IBM, a REDE-LAB alcançou seu objetivo de automatizar os processos de mineração de dados complexos, permitindo que os investigadores acelerassem seu trabalho de forma significativa. Como resultado, os investigadores podem identificar padrões de atividades financeiras ilícitas mais rapidamente e com muito mais precisão, permitindo a coleta de dados altamente eficiente a partir de dezenas de diferentes fontes, eliminando milhares de horas de investigação manual e esforços de busca.

O Big Data está vigiando você

Harari (2018, p.76) nos mostra que antigamente as pessoas acreditavam que a fonte da autoridade vinha de leis divinas. Depois, em séculos mais recentes, a fonte da autoridade passou das entidades celestiais para os

humanos. Em breve a fonte da autoridade vai mudar dos humanos para os algoritmos. O autor diz:

> Assim como a autoridade divina foi legitimada por mitologias religiosas, e a autoridade humana foi justificada pela narrativa liberal, a futura revolução tecnológica poderia estabelecer a autoridade dos algoritmos de Big Data, ao mesmo tempo que solapa a simples ideia da liberdade individual. (Harari, 2018, p.76)

Hoje vivemos na era da digitalização, praticamente tudo, das pessoas até coisas, estão sendo digitalizadas. Com a Internet das Coisas (IoT) tudo adquire vida própria, a fronteira entre o real e o virtual cada vez faz menos sentido, o real é virtual e o virtual é real e tudo isso gera informação o tempo todo. Com toda essa informação, a meta é encontrar, usando algoritmos, uma norma padrão de comportamento e, então, se prevenir contra a incerteza.

Harari (2018, p.79) diz que, futuramente, com a união da biotecnologia com a tecnologia da informação, os algoritmos de Big Data, alimentados pelo fluxo constante de dados biométricos, vão poder monitorar nossa saúde 24 horas por dia, sete dias por semana. Vão detectar logo no início, gripe, câncer e muitas outras doenças, muito antes de sentimos qualquer coisa em relação a doença. Podendo então já recomendar tratamentos, dietas, etc., sob medida, de acordo com nosso físico, DNA e personalidade. Ele fala que os algoritmos serão capazes de monitorar e compreender sentimentos melhor do que nós mesmos.

Segundo Harari (2018, p.79): "As pessoas usufruirão dos melhores serviços de saúde da história, mas justamente por isso estarão doentes o tempo todo. Existe sempre algo errado em algum lugar do corpo".

Uma questão muito boa colocada pelo autor é que, graças a sensores biométricos e algoritmos de Big Data, as doenças vão ser diagnosticadas e tratadas antecipadamente, e com isso, iremos seguir a recomendação de algum algoritmo. Pelo nosso livre-arbítrio podemos recusar algum tratamento. Mas isso é um bom motivo para que o seguro-saúde seja cancelado. E se os sensores perceberem que não estamos seguindo a recomendação deste ou daquele algoritmo e mandarem informações para a companhia de seguros ou o

plano de saúde? Segundo o autor:

> Se tiverem dados biométricos e capacidade computacional suficientes, sistemas de processamento de dados externos poderão intervir em todos os seus desejos, todas as suas decisões e opiniões. Poderão saber exatamente quem é você. (Harari, 2018, p.81)

Você poderá até se esconder de algumas pessoas mais próximas, mas não irá se esconder de grandes sites da Internet ou da polícia federal. Os algoritmos irão passar seus dados para grandes corporações e qualquer anúncio que você visualizar, vai estar baseado em seu perfil, fazendo com que você compre o produto e se sinta com o poder de compra. Por que a estratégia de negócios é simples: quanto mais personalizadas forem suas informações, mais a chance de você comprar os produtos oferecidos por eles. Mas essas decisões que os algoritmos tomam podem afetar negativamente a sua vida. Pariser (2011, p.13), cita um comentário feito por Tapan Bhat, vice-presidente do Yahoo, que diz: "O futuro da internet é a personalização - a rede agora gira em torno do 'eu'. A ideia é entregar a rede de uma forma inteligente e personalizada para o usuário". Pariser (2011) diz que:

> A fórmula dos gigantes da internet para essa estratégia de negócios é simples: quanto mais personalizadas forem suas ofertas de informação, mais anúncios eles conseguirão vender e maior será a chance de que você compre os produtos oferecidos. (Pariser, 2011, p.12)

Já existem algoritmos que detecta emoções humanas com base nos movimentos dos olhos e dos músculos faciais. Segundo The Wall Street Journal (2015), a matéria: "A tecnologia que desmascara suas emoções ocultas", diz que:

> As empresas estão acumulando um enorme banco de dados de emoções humanas usando tecnologia que depende de algoritmos para analisar os rostos das pessoas e, potencialmente, descobrir seus sentimentos mais profundos. Embora a tecnologia em evolução tenha muitos benefícios

> potenciais, também está aumentando as preocupações com a privacidade. (The Wall Street Journal, 2015)

A matéria do site conta que, Paul Ekman, psicólogo de 80 anos, que foi pioneiro no estudo sobre as emoções e sua relação com as expressões faciais na década de 1970, criando um catálogo de mais de 5.000 movimentos musculares do rosto para revelar emoções ocultas, diz que teme ter criado um mostro. Já que seu catalogo é a base para o software usado pelos anunciantes e varejistas para estudar os clientes.

A tendência é entregarmos aos algoritmos cada vez mais tarefas e aos poucos perderemos nossa aptidão para tomar decisões por nós mesmos. Harari (2018, p.80), diz que já entregamos ao algoritmo do Google uma das tarefas mais importantes: buscar informação relevante e confiável.

De acordo com Pariser (2011):

> A maior parte das pessoas imagina que, ao procurar um termo no Google, todos obtemos os mesmos resultados - aqueles que o PageRank, famoso algoritmo da companhia, classifica como mais relevantes, com base nos links feitos por outras páginas. No entanto, desde dezembro de 2009, isso já não é verdade. Agora, obtemos o resultado que o algoritmo do Google sugere ser melhor para cada usuário específico - e outra pessoa poderá encontrar resultados completamente diferentes. Em outras palavras, já não existe Google único. (Pariser, 2011, p.6)

A tarefa de examinar uma quantidade enorme de resultados e buscar quais são as mais interessantes exige dedicação e tempo. Assim, quando os algoritmos nos entregam informações personalizadas, feitas sob medida para cada um de nós, temos a tendência de aceitá-las.

Os algoritmos só fazem ampliar nosso desejo por coisas conhecidas e nos deixando desatentos às coisas ocultas e desconhecidas. Pariser (2011) chama isso "Bolha dos Filtros". Ele adverte:

> Um mundo construído a partir do que é familiar é um mundo no qual não temos nada a aprender. Se a personalização for excessiva, poderá nos impedir de entrar em contato com

> experiências e ideias estonteantes, destruidoras de preconceitos, que mudam o modo como pensamos sobre o mundo e sobre nós mesmos. (Pariser, 2011, p.21)

A bolha dos filtros sempre vai nos mostrar informações específicas que nos estimulam mais, como: sexo, poder, fofocas, violência, celebridades ou humor. Pois temos uma tendência maior a reagir sobre esses assuntos, e esse tempo gasto em digitar uma pesquisa até o momento de clicar em um dos resultados, revela traços da nossa personalidade e todas essas informações sobre nós são guardadas. Pariser (2011) diz que as notícias moldam a nossa visão do mundo e dá um exemplo do que a bolha dos filtros está fazendo:

> Em 2004, Las Últimas Noticias, um importante jornal chileno, começou a basear todo o seu conteúdo nos cliques dos leitores: as matérias que recebiam muitos cliques ganhavam continuações, e as histórias sem cliques eram eliminadas. Os repórteres já não procuram furos - eles apenas botam lenha na fogueira das matérias que ganham mais cliques. (Pariser, 2011, p.85)

Então, é muito fácil hoje examinar na enxurrada de dados que circulam na rede e verificar quais os termos que as pessoas estão mais pesquisando e baseando-se nisso, gerar conteúdo sobre esse tema que correspondam a essas pesquisas. Isso não afeta apenas o modo como processamos a notícia. Pode afetar o modo como pensamos. Pariser (2011) diz que:

> Os filtros personalizados podem prejudicar de duas maneiras o equilíbrio cognitivo entre o fortalecimento de nossas ideias existentes e a aquisição de novas ideias. Em primeiro lugar, a bolha dos filtros nos cerca de ideias com as quais já estamos familiarizados (e com as quais já concordamos), dando-nos confiança excessiva em nossa estrutura mental. Em segundo lugar, os filtros removem de nosso ambiente alguns dos principais fatores que nos incentivam a querer aprender. (Pariser, 2011, p.99)

Um bom exemplo de erro de algoritmo é citado por Harari (2018, p.101).

Em outubro de 2017, um trabalhador palestino postou na sua conta do Facebook uma foto sua no trabalho, ao lado de uma escavadeira e na imagem escreveu em letras árabes "Ysabechhum" que significa "Bom dia". O algoritmo identificou as letras como sendo "Ydbachhum", que significa "mate-os". As forças de segurança de Israel, imaginando que ele pudesse ser um terrorista e planejava usar a escavadeira para atropelar pessoas, rapidamente o prenderam. Só foi solto depois que comprovaram que o algoritmo tinha cometido um erro.

Outro bom exemplo citado por Harari (2018, p.86) é que em março de 2012, três turistas japoneses na Austrália, decidiram viajar para uma pequena ilha e seguindo as instruções do GPS terminaram caindo com o carro direto no oceano Pacífico.

Os algoritmos vão cometer erros por algum motivo, seja por falta de dados, falhas nos programas, etc. Mas o algoritmo não precisa ser perfeito, como diz Harari (2018, p.84) só precisaria ser em média, melhor que nós humanos. O autor conclui: “E isso não é difícil, porque a maioria das pessoas não conhece a si mesma muito bem, e porque a maioria das pessoas frequentemente comete erros terríveis nas decisões mais importantes da vida”.

Análise dos resultados

O uso da análise preditiva já está revolucionando o mundo e no modo como interagimos com o nosso ambiente. E, à medida que a quantidade de dados aumentarem, a probabilidade de acertos dos algoritmos vão crescer, antecipando certos acontecimentos e ações.

Vimos que com o crescimento do Big Data e da disseminação dos dispositivos e até das Internet das Coisas (IoT). Estamos entregando muitas decisões aos algoritmos. Mas até que ponto, podemos deixar os algoritmos tomarem decisões por nós?

Outra questão é a quantidade de informações pessoais que essas organizações públicas e privadas, estão guardando sobre as pessoas. É de responsabilidade dessas organizações a segurança, o vazamento desses dados, a privacidade, rastreamento e como estão sendo usando esses dados

para uso pessoal e comercial. Existe uma necessidade de implementar proteções éticas para garantir a privacidade dos consumidores, pois a maioria das iniciativas de marketing colocam a busca do lucro acima dos interesses do consumidor.

A ética abrange como as pessoas reagem quando confrontadas com decisões morais. Uma boa questão filosófica que é debatida há muitos anos vem nos servir bem de exemplo aqui. É o dilema do trem. Warburton (2011) explica:

> Um dia, você sai para passear e vê um trem desenfreado indo em direção de cinco trabalhadores. O maquinista está inconsciente, provavelmente por ter sofrido um infarto. Se nada for feito, todos morrerão. O trem passará por cima deles, pois está indo rápido demais e não haverá tempo de saírem do caminho. No entanto, há uma esperança. Há uma bifurcação nos trilhos pouco antes de onde estão os cinco homens, e na outra linha há apenas um trabalhador. Você está bem perto da chave que muda o sentido dos trilhos, de modo que o trem mude de direção e mate apenas um trabalhador em vez de cinco. Matar esse homem inocente é a coisa certa a fazer? (Warburton, 2011, p.206)

Agora, imagine esse dilema ético, só que em vez de ser com um trem, ser em um carro autodirigido. Harari (2018, p.90) coloca da seguinte forma: Dois garotos correm para pegar a bola que foi para o meio da rua e não olham se vem carro. O algoritmo do carro, com base nos seus cálculos, verifica que a melhor maneira de não atingir os dois garotos é passar para a pista oposta, mas, se arriscar a colidir com um caminhão que vem em sentido oposto, tendo a possibilidade de que o dono, que está no banco de trás, morra. O que o algoritmo deveria fazer?

Sabemos que muitos motoristas matam muitas pessoas anualmente. Segundo Gomes (2014), o Brasil é o 4º país do mundo com maior número de mortes no trânsito. Em 2010, foram registradas 42.844 mortes no trânsito do Brasil. Para melhorar isso, os algoritmos teriam que ser perfeitos em suas decisões.

Harari (2018, p.95) diz que a Tesla, empresa americana que desenvolve,

produz e vende automóveis elétricos de alto desempenho, vai deixar essas decisões éticas para o mercado. A empresa vai criar dois modelos de carros autodirigidos: o Tesla Altruísta e o Tesla Egoísta. No caso de uma emergência, o Altruísta sacrifica seu dono, enquanto o Egoísta fará de tudo para salvar seu dono.

Em um estudo pioneiro feito em 2015:

> Apresentou-se a pessoas um cenário hipotético de um carro autodirigido na iminência de atropelar vários pedestres. A maioria disse que nesse caso o carro deveria salvar os pedestres mesmo que custasse a vida de seu proprietário. Quando lhes perguntaram se eles comprariam um carro programado para sacrificar seu proprietário pelo bem maior, a maioria respondeu que não. Para eles mesmos, iam preferir o Tesla Egoísta. (Harari, 2018, p.95)

Essa é uma questão muito difícil. Até que ponto estamos prontos a entregar muitas decisões das nossas vidas aos algoritmos? E que informações as empresas tem o direito de saber sobre as pessoas? As organizações precisam criar um equilíbrio, com ética, do uso dos recursos da internet e a proteção e privacidade de seus consumidores.

O Brasil passou a fazer parte dos países que contam com uma legislação específica para proteção de dados. É a lei nº 13.709/2018, ou Lei Geral de Proteção de Dados Pessoais (LGPD ou LGPDP) (GovBR, 2018). A Lei, que entrará em vigor a partir de agosto de 2020, concede ao cidadão alguma proteção quanto à forma como as empresas e instituições gerem os nossos dados pessoais, inclusive nos meios digitais, e o que podem fazer com eles, e esses dados só podem ser coletados mediante o consentimento do usuário.

Conclusões

Durante o processo de andamento deste trabalho, buscou-se alcançar os objetivos de esclarecimento da Analise Preditiva por meio de revisão bibliográfica, explorando os conceitos de Análise Preditiva, Big Data, Data Mining, ética e leis.

O estudo buscou na literatura e na internet bases para estabelecer uma visão geral das tecnologias apresentadas, mostrando seus principais componentes, desde a parte mais simples, que é o dado, passando pelo processo de construção da informação, a construção de algoritmo para extração de informação, até a parte ética de se usar todos esses dados.

Assim, vemos que a Analise Preditiva é uma ótima ferramenta capaz de nos fornecer informações úteis e confiáveis. O potencial de previsibilidade mais eficaz e eficiente serve como um instrumento decisório para as organizações, que estão em constante evolução e inovação.

Mas não é uma ferramenta perfeita. Se entrar lixo, vai sair lixo. O fator humano é muito importante na criação dos algoritmos e das regras, e tem que haver ética na criação dessas regras. Se não houver leis e fiscalizações, como saberemos se os algoritmos não estarão segregando pessoas, criando uma sociedade desigual?

Cada vez mais instituições, corporações e bancos estão usando algoritmos para analisar dados e tomar decisões a nosso respeito. Eles estão criando seres humanos domesticados que produzem uma quantidade enorme de dados. E toda essa riqueza de informação está se concentrando nas mãos de uma pequena elite. Os dados são o novo petróleo!

Referências

Amaral, Fernando. **Aprenda Mineração de dados:** Teoria e prática. Rio de Janeiro: Alta Books, 2016

Amaral, Fernando. **Introdução à Ciência de Dados:** Mineração de Dados e Big Data. Rio de Janeiro: Alta Books - 2016

Braga, Luis Paulo Vieira. **Introdução a Mineração de Dados**. 2ª ed. Rio de Janeiro: E-papers, 2005

Gomes, Luiz Flávio - **Mortes no trânsito: Brasil é o 4º do mundo,** fev. 2014 – Disponível em: <https://professorlfg.jusbrasil.com.br/artigos/113704460/mortes-no-transito-brasil-e-o-4-do-mundo>, Acesso em: 23 abr. 2019

Gonçalves , Eduardo Corrêa - **Data Mining de Regras de Associação – Parte 1,** 2007 - Disponível em: <https://www.devmedia.com.br/data-mining-de-regras-de-associacao-parte-1/6533>, Acesso em: 23 abr. 2019

GovBR - **Lei nº 13.709/2018 - Lei Geral de Proteção de Dados Pessoais** – Ago. 2018. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm>, Acesso em: 23 abr. 2019

Harari, Yuval Noah. **21 Lições para o século 21.** São Paulo: Companhia das Letras, 2018

Hurwitz, Judith. Et al. **Big Data for Dummies.** Hoboken: John Wiley & Sons, Inc., 2013

IBM - **Ministério da Justiça do Brasil – REDE-LAB do Brasil identifica ativos ilícitos com ajuda do IBM Watson Explorer** (IBM Software Information Management, Agosto de 2014). Disponível em: <http://www.viaapia.com.br/wp-content/uploads/2017/05/caso-de-sucesso.pdf>, Acesso em: 25 Abr. 2019.

JAIN , KUNAL - **Um tutorial completo para aprender Data Science com Python do zero**. Disponível em: <https://www.vooo.pro/insights/um-tutorial-completo-para-aprender-data-science-com-python-do-zero/>, Acesso em: 23 abr. 2019

Júnior, Methanias Colaço. **Projetando Sistemas de Apoio à Decisão Baseados em Data Warehouse.** Rio de Janeiro: Axcel Books, 2004

Marquesone , Rosangela. **Big Data - Técnicas e tecnologias para extração de valor dos dados.** São Paulo: Casa do Código, 2016

Pariser, Eli. **O filtro invisível - O que a internet está escondendo de você.** Rio de Janeiro: Zahar, 2011

The Wall Street Journal – Dwoskin, Elizabeth; Rusli, Evelyn M., **The Technology that Unmasks Your Hidden Emotions** – 28 jan 2015, Disponível em: <https://www.wsj.com/articles/startups-see-your-face-unmask-your-emotions-1422472398>, Acesso em: 23 abr. 2019

Warburton, Nigel. **Uma breve história da filosofia.** Porto Alegre: L&PM, 2011

Wikipédia - **Validação cruzada** – 2017, Disponível em: <https://pt.wikipedia.org/wiki/Valida%C3%A7%C3%A3o_cruzada>, Acesso em: 23 abr. 2019