Fundação Educacional do Município de Assis
Instituto Municipal de Ensino Superior de Assis
Campus "José Santilli Sobrinho"

MARCELO VICENTE JUNIOR

SOLUÇÕES EM BIG DATA COM DATA SCIENCE: A DESCOBERTA DO CONHECIMENTO POR MEIO DA MINERAÇÃO DE DADOS

Assis/SP
2021

Fundação Educacional do Município de Assis
Instituto Municipal de Ensino Superior de Assis
Campus "José Santilli Sobrinho"

MARCELO VICENTE JUNIOR

SOLUÇÕES EM BIG DATA COM DATA SCIENCE: A DESCOBERTA DO CONHECIMENTO POR MEIO DA MINERAÇÃO DE DADOS

Projeto de pesquisa apresentado ao Curso de Bacharelado em Ciência da Computação do Instituto Municipal de Ensino Superior de Assis – IMESA e da Fundação Educacional do Município de Assis – FEMA como requisito parcial à obtenção do Certificado de Conclusão.

Orientando: Marcelo Vicente Junior
Orientador: Prof. Dr. Alex Sandro Romeo de Souza Poletto

Assis/SP
2021

FICHA CATALOGRÁFICA

V632s VICENTE JUNIOR, Marcelo.

Soluções em Big Data com Data Science: A Descoberta do Conhecimento por meio da Mineração de Dados / Marcelo Vicente Junior. – Assis, 2021.

79 p.

Trabalho de conclusão do curso (Ciência da Computação). – Fundação Educacional do Município de Assis-FEMA

Orientador: Dr. Alex Sandro Romeo de Souza Poletto

1. Big Data. 2. Ciência de dados. 3. Mineração de dados

CDD 005.74

SOLUÇÕES EM BIG DATA COM DATA SCIENCE: A DESCOBERTA DO CONHECIMENTO POR MEIO DA MINERAÇÃO DE DADOS

MARCELO VICENTE JUNIOR

Trabalho de Conclusão de Curso apresentado ao Instituto Municipal de Ensino Superior de Assis, como requisito do Curso de Graduação, avaliado pela seguinte comissão examinadora:

Orientador: _____
Prof. Dr. Alex Sandro Romeo de Souza Poletto

Examinador: _____
Prof. Dr. Luiz Ricardo Begosso

Assis/SP
2021

DEDICATÓRIA

Dedico este trabalho à minha família que esteve sempre comigo em todos os momentos, tanto nos bons quanto nos ruins.

AGRADECIMENTOS

Agradeço, primeiramente a Deus, que em sua infinita bondade me sustentou até aqui. A minha família, pais, irmãos, avó e sobrinho que estiveram sempre comigo, apoiando e encorajando nos momentos difíceis assim como nos momentos felizes, comemorando e compartilhando desses momentos.

Aos amigos que fiz durante a graduação, que contribuíram imensamente em minha formação, no convívio do dia-a-dia, com novos aprendizados, discussões, estudos e, principalmente deixando a rotina da faculdade mais leve.

Aos professores da FEMA/IMESA que nos transmitiram suas experiências e conhecimentos com maestria e, principalmente o Prof. Dr. Alex Sandro Romeo de Souza Poletto, que foi o orientador desse trabalho de conclusão de curso.

E finalmente, agradeço a esta Instituição de Ensino que me permitiu continuar a busca pelo conhecimento ao cursar uma segunda graduação.

“O mais importante é nunca parar de questionar. A curiosidade tem uma razão para existir… Nunca perca sua sagrada curiosidade.”

Albert Einstein

RESUMO

O volume de dados gerado tem crescido a cada dia, oriundo das mais variadas fontes, numa velocidade exponencial, dando origem ao chamado Big Data, termo utilizado para designar a enorme quantidade de informações armazenadas em bancos de dados. Em razão dessas características dos dados gerados, tornaram-se necessárias novas tecnologias para organizar e extrair informações, tendo em vista que os modelos tradicionais não poderiam lidar com tamanho volume, surgindo assim a Ciência de Dados, definida por um conjunto de conceitos e ferramentas, mais voltada em prever comportamentos do que analisar simplesmente, forma em que geralmente é vista. A Mineração de Dados pode ser definida como o processo que permite explorar grande quantidade de informações, a fim de encontrar informações que a princípio não estão disponíveis em um primeiro momento, sendo o elemento central responsável pela parte analítica na Ciência de Dados e intimamente ligada ao Big Data. Ferramentas são desenvolvidas para auxiliar nesse processo de descoberta de conhecimento e, saber como integrá-las é primordial para que esse processo tenha êxito e cumpra seu principal objetivo: auxiliar no processo de tomada de decisão gerencial.

Palavras-chave: Big Data; Ciência de Dados; Mineração de Dados.

ABSTRACT

The volume of data generated has grown every day, coming from the most varied sources and at an exponential speed, giving rise to the so-called Big Data, a term used to designate the enormous amount of information stored by database. Due to these characteristics of the data generated, new technologies became necessary to organize and extract information, considering that traditional models could not deal with such a large volume, thus emerging Data Science, defined by a set of concepts and tools, more focused on predicting behaviors than analyzing simply, the way in which it is usually seen. Data Mining can be defined as the process that allows exploring a large amount of information, in order to find information that at first was not available at first, being the central element responsible for the analytical part in Data Science and closely linked to the Big Data. Tools are developed to assist in this knowledge discovery process, and knowing how to integrate them is essential for this process to be successful and fulfill its main objective: to assist in the management decision-making process.

Keywords: Big Data; Data Science; Data Mining.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BI – *Business Intelligence*

CSV – *Comma-separeted Values*

DBSCAN – *Density Based Spatial Clustering of Aplications*

FP-Growth – *Frequent-pattern growth*

FP-Tree – *Frequent-Pattern Tree*

JSON – *JavaScript Object Notation*

KDD – *Knowledge Discovery in Databases*

k-NN – k-*Nearest Neighbours*

MAE – Erro Médio Absoluto

MLP – *Multilayer Perceptron*

MSE – Erro Médio Quadrático

MST – *Minimal Spanning Tree*

NaN – *Not a Number*

PDF – *Portable Document Format*

RBF – *Radial Basis Function*

SQL – *Structured Query Language*

SVM – *Support Vector Machine*

XML – *eXtensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO

Com a revolução digital, o mundo encontra-se envolto em um volume inédito de dados, estimando-se que 90% dos dados existentes tenham sido criados nos últimos dois anos, sendo que os dados vindos de bancos de dados transacionais constituem uma parcela ínfima, e que a maior parte do volume existente são obtidos de outras fontes, tais como websites, redes sociais, sensores ou câmeras de monitoramentos, por exemplo. A esse grande conjunto de dados exponencial gerado, dá-se o nome de Big Data, cuja dimensão está além da habilidade das ferramentas típicas de capturar, gerenciar e analisar dados (TAURION, 2015, p. 35-36).

Tornou-se necessário transformar esses dados em conhecimentos, que possam ser utilizados em suas próprias atividades, sejam comerciais ou científicas (CÔRTES; PORCARO; LIFSCHITZ, 2002, p. 2). A expressão *"Data Science"* data da década de 60, porém, a ciência de dados é uma ciência nova, que tem por objetivo estudar o dado em todo seu ciclo de vida, desde a produção até o seu descarte (AMARAL, 2016, p. 4). É composta pela combinação de habilidades e áreas diferentes de conhecimento com o objetivo de coleta, preparação, análise, visualização, gerenciamento e preservação de grandes quantidades de informação (PAIXÃO; SILVA; TANAKA, 2015, p. 2). Tem em comum com a Inteligência de Negócios (*Business Intelligence*) suas principais funções, que é converter dados brutos em *insights* a serem usados no processo de tomada de decisões nos negócios ou domínios de aplicações em geral, tendo diferenças em suas abordagens (PAIXÃO; SILVA; TANAKA, 2015, p. 5).

A mineração de dados pode ser definida como um processo analítico, sistemático e, até onde é possível, automatizado, com a finalidade de explorar analiticamente grandes bases de dados, para descobrir padrões relevantes que possam gerar conhecimento (SILVA; PERES; BOSCARIOLLI, 2016, p.10). Ela pode ser considerada como uma parte do processo de Descoberta de Conhecimento em Banco de Dados (KDD - *Knowledge Discovery in Databases*), sendo que o termo KDD é usado para representar o processo de tornar dados de baixo nível em conhecimento de alto nível, enquanto mineração de dados pode ser definida como a extração de padrões ou modelos de dados observados (DIAS, 2002, p. 16-17).

O foco principal tanto da Ciência de Dados, quanto da Mineração de Dados é a extração de conhecimento a partir de grandes volumes de dados, para auxiliar o processo gerencial de tomada de decisão. São inter-relacionados e se misturam e, delimitar a relação entre esses conceitos torna-se importante para a compreensão e, principalmente sua aplicação em situações reais.

1.1. OBJETIVOS

O presente trabalho tem como objetivo apresentar um estudo sobre a Ciência de Dados, suas etapas e como a mineração de dados está inserida dentro deste contexto. Com relação a mineração de dados o estudo apresentará as principais tarefas, técnicas e algoritmos utilizados. Para este fim, pretende-se desenvolver um estudo de caso, utilizando ferramentas já existentes, demonstrando a integração das etapas de armazenamento, pré-processamento, análise estatística, mineração de dados e apresentação dos resultados obtidos.

1.2. JUSTIFICATIVAS

A geração exponencial de dados gera uma demanda de o que fazer com esses dados, já que os dados brutos por si só não representam nada substancial com relação ao conhecimento. Saber o que fazer a partir deles é primordial para agregar valor a esses dados. A ciência de dados tem esse papel, de sistematizar um conjunto de ferramentas e tecnologias que possibilitam a extração de conhecimento a partir de dados brutos. Inserido nesse contexto, está a mineração de dados, cujo objetivo é a descoberta de padrões e tendências de maneira preditiva em grandes volumes de dados.

Um conjunto sistemático de procedimentos, ferramentas e técnicas facilitam a orientação do que deve ser feito desde o início até a apresentação dos resultados. A partir de uma correta seleção, armazenamento e transformação, uma efetiva análise de dados pode ser realizada, aplicando técnicas de mineração de dados e extraindo conhecimento descritivo e preditivo, reconhecendo padrões e comportamentos e, a partir dele, fornecer subsídios para o processo gerencial de tomada de decisão.

1.3. MOTIVAÇÕES

A partir do grande volume de dados existente atualmente, temas como Ciência de Dados (do inglês *Data Science)* e mineração de dados tem se tornado comuns. Saber o que são esses conceitos e, principalmente saber aplicar as técnicas de análise e mineração de dados, é preponderante para que uma organização possa tomar a decisão adequada diante de determinado cenário, prevendo tendências e antecipando comportamentos. Possibilita prever possíveis problemas que possam vir a ocorrer, bem como antecipar soluções, ou efetuar correções para que o eventual problema nem ocorra.

1.4. PERSPECTIVAS DE CONTRIBUIÇÃO

Este trabalho tem a intenção de contribuir para que estes temas recorrentes na atualidade, ciência de dados e mineração de dados sejam melhor compreendidos e, principalmente aplicados em situações reais, não só por grandes empresas, mas também pelos pequenos e médios empresários.

1.5. METODOLOGIA DE PESQUISA

A metodologia escolhida será de, inicialmente, partir de uma revisão bibliográfica sobre os temas propostos, conforme descrito na estrutura do trabalho. Como fonte de pesquisa serão utilizados livros, teses, monografias, artigos em revistas especializadas no assunto, matérias publicadas em sites específicos sobre o tema de modo a agregar conhecimento sobre o tema do trabalho e sua elaboração.

Em seguida, serão estudadas as ferramentas a serem utilizadas, Google BigQuery, Google Data Studio, Jupyter Notebook e bibliotecas da linguagem Python disponíveis para análise de dados e, também será selecionada a base de dados em que será realizado o estudo de caso. Com relação a base de dados a ser utilizada, caso não seja possível utilizar uma base real, pretende-se utilizar uma base de dados disponível em diversos repositórios sobre o tema.

Com ferramentas e dados selecionados, pretende-se realizar o estudo de caso, utilizando as ferramentas selecionadas de modo integrado, para ser obtido o objetivo principal, isto é, a extração de conhecimento em uma base de dados. O estudo de caso será

documentado e integrado ao trabalho, bem como a conclusão final sobre o trabalho realizado.

A figura 6, página 53, ilustra a metodologia descrita com base na proposta de trabalho.

1.6. RECURSOS NECESSÁRIOS

Para a realização do trabalho, será necessário um computador com o sistema operacional Windows, nos quais serão utilizados os softwares da Google BigQuery e Data Studio, além do software *open-source* Jupyter Notebook e bibliotecas existentes para a Linguagem Python. Caso seja necessário, durante a elaboração da pesquisa, poderá ser agregado recursos adicionais visando a melhoria do trabalho.

1.7. ESTRUTURA DO TRABALHO

O trabalho está dividido em 7 capítulos, sendo o Capítulo 1, esta Introdução.

- **Capítulo 2 – Dados:** Abordagem sobre os conceitos de dado, conhecimento e informação; ciclo de vida do dado e Big Data.
- **Capítulo 3 – Ciência de Dados:** Capítulo sobre ciência de dados, conceitos, relações com o Big Data, *Business Intelligence* e as suas etapas.
- **Capítulo 4 – Mineração de Dados:** Neste capítulo, serão discutidas as etapas da pré-processamento dos dados; análise exploratória dos dados; tarefas, técnicas e algoritmos utilizados na mineração de dados.
- **Capítulo 5 – Proposta de Trabalho:** Neste capítulo, será apresentada a proposta de trabalho a ser desenvolvida no estudo de caso, e serão apresentadas as ferramentas a serem utilizadas na análise de dados.
- **Capítulo 6 – Estudo de Caso:** Neste capítulo, será demonstrado a aplicação das ferramentas escolhidas em um caso real de análise de dados, partindo dos dados brutos, integrando as ferramentas apresentadas, extraindo conhecimento e apresentando os resultados.
- **Capítulo 7 – Conclusão:** Capítulo que abordará as considerações finais acerca do trabalho realizado.
- **Referências:** Referências utilizadas no trabalho.

2. DADOS

O homem sempre se deparou com a necessidade de deixar registrados os eventos de sua vida, informações que julgava importante ou que pudessem ser utilizadas futuramente. Essas informações demandam de uma forma de armazenamento e recuperação que seja prática, eficiente e confiável (ALVES, 2021, p. 12).

Segundo Elmasri e Navathe (2011, p. 4), dados podem ser definidos como fatos conhecidos que podem ser registrados e possuem significado implícito. Representam algum aspecto do mundo real, é logicamente coerente e possui algum significado inerente. Também podem ser definidos como os recursos naturais da sociedade da informação, apesar de só ter valor quando tratados, analisados e usados para tomada de decisões (TAURION, 2015, p. 39).

2.1. CICLO DE VIDA DO DADO

Durante sua vida útil, o dado pode passar por várias etapas diferentes ou por nenhuma, de acordo com sua natureza e finalidade, mas de modo geral pode se definir um ciclo padrão, no qual a maioria dos dados são adaptáveis (AMARAL, 2016, p. 17). O ciclo de vida de um dado pode ser definido por cinco etapas: produção, armazenamento, transformação, análise e descarte, conforme ilustrado na Figura 1.

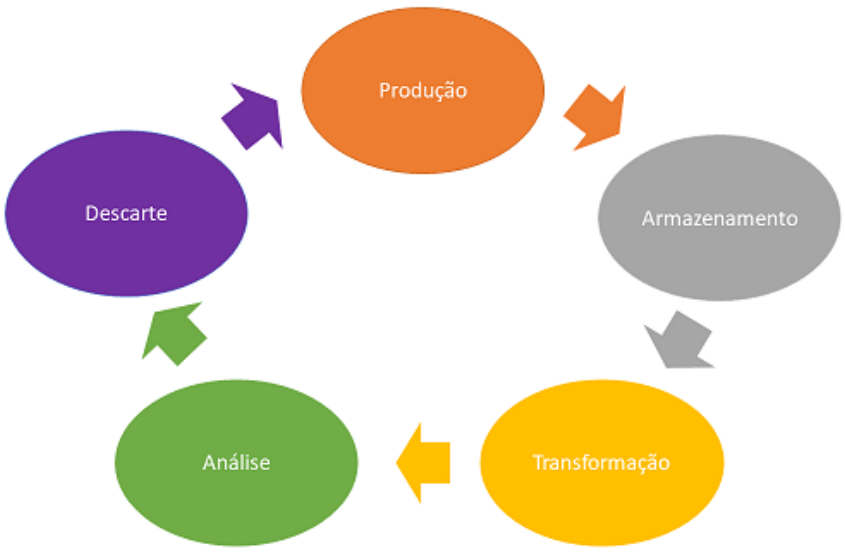

Figura 1: Ciclo de Vida do Dado (SIRQUEIRA; DALPRA, 2018)

A produção de dados abrange todas as formas de geração de dados, tais como sistemas transacionais, pesquisas, dados históricos, arquivos, *Data Warehouse*, por meio de computadores, periféricos, celulares, sensores, câmeras, entre outros. Após serem gerados, os dados devem ser armazenados, para que possam ser recuperados no futuro e, algumas premissas devem ser atendidas: segurança da informação, integridade, minimização de redundância, concorrência, otimização de espaço, etc. (AMARAL, 2016, p. 23). Nesta etapa os dados já podem ser utilizados, porém, para produzir informação, geralmente passam por uma nova etapa chamada transformação, caracterizada pela realização de processos de limpeza, integração e transformação dos dados (FERREIRA et al., 2010, p. 758), otimizando assim os dados para a realização de análise e produção de informação.

Após serem transformados os dados podem ser armazenados em *Data Warehouses,* que são depósitos de dados, que centraliza e consolida grandes quantidades de dados de várias fontes, com informações pré-calculadas e dados não normalizados, destinados exclusivamente a realizar consultas e análises avançadas, contendo grandes quantidades de dados históricos (ORACLE, 2021). Podem também ser utilizados *Data Marts*, que são divisões menores do *Data Warehouse*, divididas por categorias ou setores de uma empresa, por exemplo.

A análise é o processo de produzir informação e conhecimento a partir dos dados, podendo ser exploratória, explícita ou implícita (AMARAL, 2016, p. 61).

Análise Exploratória: tem por objetivo conhecer os dados antes de analisa-los em si, por meio de métodos quantitativos e visuais, para depois aplicar analises explicitas ou implícitas e tirar conclusões.

Análise Explícita: a informação e o conhecimento já estão disponíveis nos dados, necessitando de alguma operação de baixa complexidade para ressaltar e produzir a informação. É muito similar à analise exploratória diferindo desta já que tem um objetivo claro e específico ao se aplicar os métodos.

Análise Implícita: a informação não está clara no conjunto de dados, sendo necessária o uso de uma função mais sofisticada, geralmente com o uso de aprendizado de máquina, para encontrar esta informação oculta. A ciência de dados tem como foco principal realizar análise desse tipo nos dados.

Por fim, quando não tem mais valor para a corporação ou por questões de otimização de recursos, é realizado o descarte dos dados.

2.2. DADO, CONHECIMENTO E INFORMAÇÃO

A ideia de que dado e informação são sinônimos é errônea, assim como informação e conhecimento. Então, é necessário fazer a distinção da diferença entre dado, informação e conhecimento.

Dado é a representação da informação, podendo estar registrada em papel, quadro de aviso ou no disco rígido do computador (ALVES, 2021, p. 14). São os registros soltos, aleatórios e sem qualquer análise (CASTRO; FERRARI, 2016, p.4). Dados são códigos que constituem a matéria prima da informação, ou seja, é a informação não tratada que ainda não apresenta relevância (REZENDE, 2015).

Informação é quando um significado é atribuído ao dado pelo ser humano (SILVA; PERES; BOSCARIOLI, 2016, p. 6), se tratando de qualquer estruturação ou organização desses dados. Um registro, em suporte físico ou intangível, disponível para a produção de conhecimento, derivada dos dados que, sem um sentido ou contexto, significam muito pouco (REZENDE, 2015).

Conhecimento é algo que permite uma tomada de decisão que agregue valor (CASTRO; FERRARI, 2016, p.4). É a informação processada e transformada em experiência pelo indivíduo. Ele vai além de informações, já que, além de ter um significado, tem uma aplicação. Se informação é dado trabalhado, então conhecimento é informação trabalhada (REZENDE, 2015).

2.3. TIPO DE DADOS

Quanto ao tipo, os dados podem ser estruturados, semiestruturados ou não estruturados. Dados estruturados são aqueles que residem em campos fixos em arquivos como, por exemplo, uma planilha, tabela ou banco de dados possuindo um modelo de dados que define as características dos dados a serem armazenados (CASTRO; FERRARI, 2016, p. 28). São resultantes, na maioria dos casos, de processos de geração de dados de sistemas transacionais ou resultado de observações e processos de medição (SILVA; PERES; BOSCARIOLI, 2016, p. 7).

Dados do tipo semiestruturado não possuem uma estrutura completa de um modelo de dados, utilizando em geral marcadores para identificar elementos específicos nos dados, sem possuir uma estrutura rígida (CASTRO; FERRARI, 2016, p. 29). São exemplos de dados semiestruturados e-mails, arquivos de formato JSON (*JavaScript Object Notation*) ou XML (*eXtensible Markup Language*).

Dados não estruturados não possuem um modelo de dados, não está organizado de maneira predefinida e nem locais definidos e são de difícil indexação, acesso e análise (CASTRO; FERRARI, 2016, p. 29). Se referem a textos livres, imagens, sons, páginas web, arquivos pdf entre outros. Grande parte dos dados disponíveis para a análise e extração de informação e conhecimento estão disponíveis nesta forma (SILVA; PERES; BOSCARIOLI, 2016, p. 7).

2.4. BIG DATA

O termo Big Data surgiu no começo dos anos 90, na NASA, para descrever o conceito de conjunto de dados grandes e complexos, onde os sistemas e estruturas computacionais existentes até então, não seria suficiente para a captura, processamento, análise e armazenamento (PASSOS, 2016, p. 394). Pode ser definido como um conjunto de dados cujo crescimento é exponencial e cuja dimensão está além da habilidade das ferramentas típicas de capturar, gerenciar e analisar dados (TAURION, 2015, p. 35). Está relacionado com grandes quantidades de dados, que possuem características distintas, são heterogêneos, providos de diferentes fontes, com controles distribuídos e descentralizados (FAGUNDES; MACEDO; FREUND, 2018, p. 196).

Inicialmente, ao ouvir o termo Big Data, é comum relacioná-lo somente a um grande volume de dados, porém, essa não é sua única propriedade. Velocidade e variedade também são propriedades tipicamente relacionadas com o Big Data. Tais propriedades são popularmente denominadas os 3 Vs de Big Data (MARQUEZONE, 2017, p. 8). Segue abaixo uma breve descrição dessas propriedades:

- **Volume:** é a característica mais significativa do Big Data, relacionada à grande quantidade de dados e informações geradas das mais variadas fontes. Estimativas de grandes consultorias de TI são de que 90% dos dados digitais existentes foram gerados nos últimos dois anos (MARQUEZONE, 2017, p. 9), crescendo de forma exponencial a cada ano. É de tamanha dimensão que novos formatos de

armazenamento e processamento são necessários, de modo a superar desafios quanto à escalabilidade, eficiência, custo e complexidade, não permitindo as antigas abordagens até então utilizadas nos modelos tradicionais.

- **Velocidade:** faz referência a velocidade com que os dados são coletados, analisados e utilizados (MARQUEZONE, 2017, p. 15). Está relacionada primordialmente com a velocidade com que os dados são coletados, mas também com o tempo de resposta para determinada requisição, isto é, a velocidade entre a troca de dados e informações (FAGUNDES; MACEDO; FREUND, 2018, p. 200) e com a velocidade com que estes dados são analisados, já que eles perdem valor com o passar do tempo, em muitos casos.
- **Variedade:** os dados gerados são diversos, e podem ser: estruturados como por exemplo, os próprios bancos de dados relacionais; semiestruturados como por exemplo, arquivos de formato JSON (*JavaScript Object Notation*) ou XML (*eXtensible Markup Language*) ou não estruturados, tais como vídeos, fotos ou alguns formatos de texto. Devido a essa variedade, são necessários modelos que ofereçam flexibilidade quanto a estrutura, sem esquema rígido pré-definido e que sejam adequados para sistemas distribuídos, devido ao grande volume. Também pode ser relacionada a variedade de áreas distintas nas quais o Big Data pode ser aplicado, podendo ser destacada a área governamental, setor financeiro, transporte, varejo, marketing e seguros (MARQUEZONE, 2017, p. 13-14).

Com o tempo, novas propriedades foram incorporadas ao conjunto de aspectos relacionados ao Big Data. Muitos pesquisadores também consideram os atributos veracidade e valor, formando assim um conjunto de 5 Vs, com as principais características de Big Data.

- **Veracidade:** está relacionada à qualidade ou à confiabilidade dos dados. Devido ao grande ao volume e variedade, podem ocorrer dados inconsistentes e, por isso, devem ser avaliados quanto à confiabilidade, podendo demandar atividades específicas para identificar e remover esses dados dos *datasets*. Um dado é considerado de má qualidade quando não pode ser convertido em informação, ocasionando sua falta de valor (FAGUNDES; MACEDO; FREUND, 2018, p. 200).
- **Valor:** faz referência a quão valioso e significativo determinado dado pode ser em uma solução e, sua análise é preponderante para determinar quais dados serão priorizados pela empresa (MARQUEZONE, 2017, p. 17).

Não é apenas um conjunto de hardware ou software, e sim, um conjunto de tecnologias, processos e práticas, que permitem a análise de dados que antes não podiam ser acessados, possibilitando tomar decisões ou gerenciar atividades de forma mais eficiente (TAURION, 2015, p. 39). Conta com diversas tecnologias e algoritmos que são implementados a grandes bancos de dados, com o intuito de efetuar a correta captura, análise, processamento e disseminação das informações, conforme a demanda, tornando as informações úteis ao processo decisório (PASSOS, 2016, p. 394).

Com o Big Data, mudanças significativas na maneira de como a análise de dados é pensada e executada são necessárias já que se tem alteração na percepção dos dados, na população da amostra e nos efeitos de casualidade. Pode-se ver que as associações e análises presentes ao longo do Big Data, não seriam possíveis de se executar somente com os métodos estatísticos e, para que seja possível analisar, sem déficits analíticos causados por desinformação ou má qualidade do dado, temos a ciência de dados (PASSOS, 2016, p. 394-395).

3. CIÊNCIA DE DADOS

Os dados provenientes do Big Data, gerados em tempo real, com grande volume e variedade, acarretou um desafio: o que fazer com eles, utilizando-os de modo significativo e com seu máximo potencial. Assim, surgiu a necessidade de uma abordagem holística e interdisciplinar de análise de dados, que possa utilizar dados provenientes de diferentes fontes (CURTY; SERAFIM, 2016, p.310).

Ciência de dados pode ser entendida como um conjunto de princípios, conceitos e técnicas que estruturam o pensamento e a análise de dados (PROVOST; FAWCETT, 2016, p. 13), extraindo informações valiosas a partir destes. Se trata de um conjunto de princípios fundamentais que suportam e guiam a extração de informações e conhecimento a partir de dados. Se ciência é um método sistemático onde pessoas estudam e explicam fenômenos de um domínio específico que ocorrem no mundo natural, pode-se entender que a ciência de dados é o domínio científico que é dedicado para descobrir conhecimento através da análise de dados (PAIXÃO; SILVA e TANAKA, 2015, p. 3-4). Deste modo, a ciência de dados pode ser entendida não como uma ferramenta, e sim um conjunto de métodos com o objetivo de apoiar decisões de negócio baseada em dados (ESCOVEDO; KOSHIYAMA, 2020, p. 3).

A ciência de dados começou a ganhar corpo a partir de uma nova forma de competição baseada no uso intensivo de análise de dados e tomada de decisões baseada em fatos, no lugar de competir em fatores tradicionais, empregando estatística, análise quantitativa e modelagem preditiva como elementos primários de concorrência. Sua definição continua a ser desenvolvida ao longo dos últimos anos, mas que, em suma, trata justamente desta combinação de habilidades e áreas de conhecimento, visando a coleta, preparação, análise, visualização, gerenciamento e preservação de grandes quantidades de informação (PAIXÃO; SILVA e TANAKA, 2015, p.2).

Normalmente está associada somente aos processos de análise dos dados, com o uso de estatística, aprendizado de máquina, ou simplesmente aplicação de um filtro para a produção de informação e conhecimento. Dessa forma, passa a ser vista somente, como um nome mais elegante para a estatística. Porém, enquanto a estatística, descritiva ou inferencial, está associada a etapa de análise de dados, a ciência de dados abrange todo

o ciclo de vida do dado, desde a produção até o descarte (AMARAL, 2016, p. 4-5), conforme ilustrado na Figura 2 abaixo.

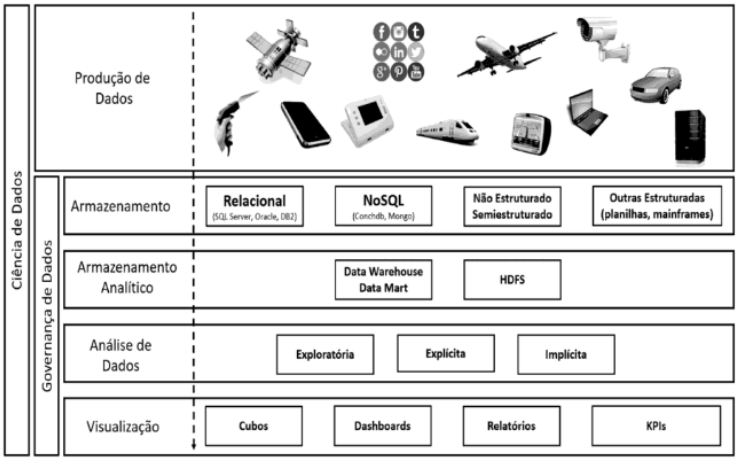

Figura 2: Panorama da Ciência de Dados (AMARAL, 2016)

Se trata de um campo interdisciplinar, que divide definições e áreas de atuação com outros campos já que, ao se falar em ciência de dados, os termos mais comuns utilizados são ligados a outras áreas de conhecimento (PAIXÃO; SILVA e TANAKA, 2015, p. 3-4). Segundo Granville (2014, apud CURTY; SERAFIM, 2016, p. 311), o cientista de dados é descrito como um profissional generalista que conhece negócios, estatística, e ciência da computação, sendo capaz de relacionar alguns conhecimentos específicos, processando grandes volumes de dados e explorando a inteligibilidade em dados a princípio desestruturados e sem sentido. Na Figura 3, temos o Diagrama de Venn que representa essa interdisciplinaridade.

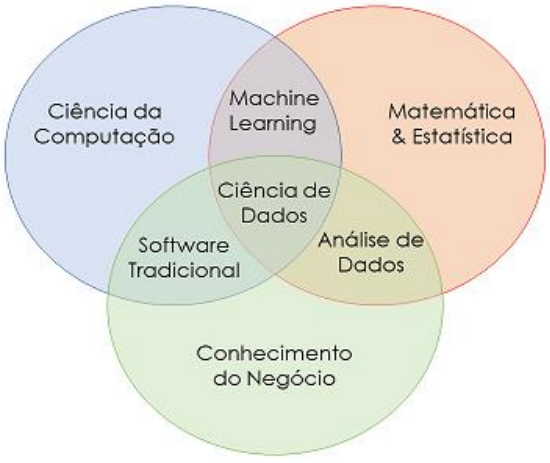

Figura 3: Diagrama de Venn da Ciência de Dados (Baseado em CONWAY, 2010)

3.1. COMPARAÇÕES E RELAÇÕES

Tendo em vista essa interdisciplinaridade e correlação com outras áreas de conhecimento, esta seção tem a intenção de comparar e mostrar a relação existente entre elas e ciência de dados.

3.1.1. Big Data

Big Data por si só não gera valor, devendo ser criados *insights* que provocam resultados tangíveis ao negócio (TAURION, 2015, p. 43). A ciência de dados, que necessita de grandes volumes de dados, encontra no Big Data seu verdadeiro potencial, já que possui técnicas avançadas para o tratamento de grandes volumes de dados e a realização de análises para a extração de conhecimento.

Tecnologias de Big Data podem ser ocasionalmente utilizadas para implementar técnicas de mineração de dados e outras atividades de ciência de dados (PROVOST; FAWCETT, 2016, p. 8).

3.1.2. *Business Intelligence* (BI)

Business Intelligence ou BI pode ser entendida como o processo de transformar dados em informação e então, em conhecimento para ser utilizado como suporte para a tomada de decisão, combinando dados operacionais com ferramentas analíticas para apresentar informações complexas e competitivas (PAIXÃO; SILVA e TANAKA, 2015, p. 2).

Ciência de dados e *Business Intelligence* possuem em comum a principal função que é converter dados brutos em *insights* a serem utilizados no processo de tomada de decisão de negócios, porém diferem nas suas abordagens. No *Business Intelligence* as análises são fundamentadas principalmente na inferência simples sobre dados históricos ou atuais, aplicando métodos estatísticos ou de mineração de dados, com a finalidade de oferecer informações relevantes a curto ou médio prazo para a tomada de decisão. Por outro lado, na ciência de dados procura-se fazer novas descobertas utilizando métodos matemáticos e estatísticos avançados, a partir de grandes quantidades dos dados do negócio, gerando *insights* preditivos para o longo prazo (PAIXÃO; SILVA e TANAKA, 2015, p. 5-6).

3.1.3. Data *Analytics*

Data Analytics é a ciência de examinar dados brutos com o objetivo de encontrar padrões e tirar conclusões sobre eles, com a aplicação de um processo algorítmico ou mecânico para a obtenção de informações, revelando tendências e métricas que, de outra forma, seriam perdidas na massa de informações (CAMPOS, 2018, p. 32).

Enquanto a ciência de dados funciona de modo mais amplo, considerando todo o ciclo de vida do dado e com o principal objetivo de encontrar conhecimento preditivo, a *data analytics* se concentra em formular correlações entre os dados existentes.

3.1.4. Mineração de Dados

Os termos *Data Science* e *Data Mining* são muitas vezes utilizados de forma intercambiável, porém em um nível mais elevado. Ciência de dados é um conjunto de princípios fundamentais que norteiam a extração de conhecimento a partir de dados, enquanto a mineração de dados é a extração de conhecimento a partir destes, por meio de tecnologias que incorporam estes princípios (PROVOST; FAWCETT, 2016, p. 2).

3.2. ESTATÍSTICA

A Estatística é uma área muito vasta que, assim como o *Data Science*, se preocupa em aprender a partir dos dados e transformar dados em informação (ESCOVEDO; KOSHIYAMA, 2020, p. 33). Pode ser entendida como uma coleção de métodos usados para planejar experimentos, coletar, organizar, sumarizar, analisar e interpretar dados de forma a extrair conclusões a partir deles. Pode ser dividida em três áreas complementares: **Estatística Descritiva**, cujo objetivo é descrever e sumarizar os dados obtidos a partir de uma amostra; **Probabilidade**, que é a teoria matemática usada para estudar a incerteza de eventos aleatórios e **Estatística Inferencial**, que é a parte que se ocupa com a generalização da informação e das conclusões obtidas a partir da amostra sobre a população (CASTRO; FERRARI, 2016, p. 331).

A Estatística ajuda a compreender como usar dados para testar hipóteses e para estimar a incerteza das decisões. O teste de hipóteses ajuda a determinar se um padrão observado é uma regularidade geral e válida, ou se é uma ocorrência do acaso em um conjunto de dados particular (PROVOST; FAWCETT, 2016, p. 36). Também é a

ferramenta capaz de descrever ou resumir dados, mostrando aspectos importantes do conjunto de dados, tais como o tipo de distribuição dos dados ou os valores mais representativos do conjunto, permitindo criar visualizações referentes a estes aspectos (SILVA; PERES; BOSCARIOLI, 2016, p. 30). Tem por objetivo, descrever e sumarizar um conjunto de dados, extraindo de variáveis contínuas ou discretas, medidas de locação, dispersão ou associação (ESCOVEDO; KOSHIYAMA, 2020, p. 36).

3.3. MACHINE LEARNING

Um modelo preditivo é uma função matemática que, aplicada a uma grande quantidade de dados, consegue identificar padrões ocultos e prever o que poderá acontecer. Os modelos preditivos podem ser supervisionados ou não-supervisionados. (TAURION, 2015, p. 56).

O aprendizado de máquina ou *machine learning* é uma área de pesquisa que busca desenvolver programas computacionais capazes de melhorar seu desempenho automaticamente por meio da experiência (CASTRO; FERRARI, 2016, p. 14). Em outras palavras, é a aplicação de técnicas computacionais na tentativa de encontrar padrões ocultos nos dados, isto é, padrões que não podem ser observados explicitamente. Está diretamente relacionada com outras ciências, como estatística e inteligência artificial, podendo ser considerada uma área de conhecimento dentro da inteligência artificial e, está diretamente ligada com a mineração de dados, já que enquanto o aprendizado de máquina trata de algoritmos que buscam reconhecer padrões em dados, a mineração de dados é a aplicação desses algoritmos em grandes conjuntos de dados em busca de informação e conhecimento (AMARAL, 2016, p. 81).

Algoritmo de aprendizagem ou algoritmo de treinamento é um procedimento bem definido para treinar uma técnica de aprendizado de máquina, e este aprendizado pode ser realizado de duas formas diferentes:

- **Aprendizado Supervisionado:** baseado em um conjunto de objetos no qual se conhece os valores da classe ou atributo que se deseja descrever ou prever (AMARAL, 2016, p. 87);
- **Aprendizado Não Supervisionado:** baseado em objetos que não se conhece os rótulos que se deseja obter, devendo o algoritmo aprender a categorizar ou rotular os objetos (CASTRO; FERRARI, 2016, p. 16).

3.4. ETAPAS DE UM PROJETO

Um projeto em ciência de dados, por se tratar de uma ciência com métodos e conceitos bem definidos, pode ser divido em algumas etapas, já que existe um processo bem compreendido que coloca uma estrutura no problema, permitindo consistência, repetitividade e objetividade razoáveis (PROVOST; FAWCETT, 2016, p. 28). Na Figura 4, tem-se o esquema básico de um projeto de ciência de dados, dividido em sete etapas e após um breve resumo sobre cada etapa.

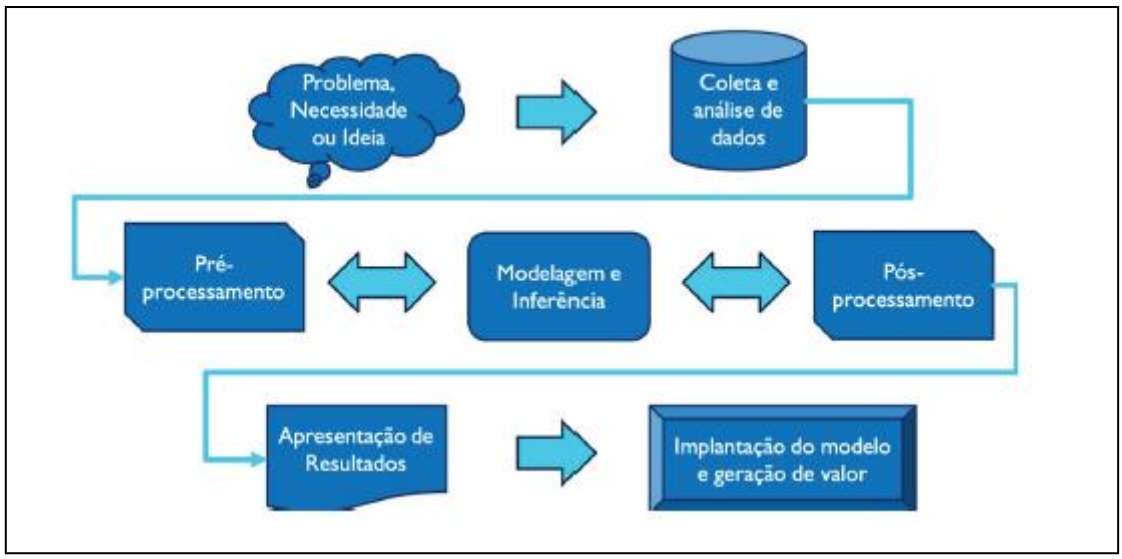

Figura 4: Esquema de um Projeto de *Data Science* (ESCOVEDO; KOSHIYAMA, 2020)

3.4.1. Entendimento do Problema

A partir de uma necessidade ou ideia, deve-se ter em mente o problema que se deseja resolver e, definir os objetivos, bem como as perguntas que se deseja responder (ESCOVEDO; KOSHIYAMA, 2020, p. 10). A formulação inicial pode não ser a ideal ou estar completa, de modo que diversas repetições podem ser necessárias para que uma solução aceitável apareça (PROVOST; FAWCETT, 2016, p. 28).

3.4.2. Coleta e Análise dos Dados

Tendo em mente que a solução do problema do negócio é o objetivo, tem-se que os dados são a matéria-prima a partir da qual a solução será construída (PROVOST; FAWCETT, 2016, p. 28). Esta etapa consiste em levantar informações e coletar os dados para resolver o problema levantado, podendo ser encontrados em um ou mais bancos de

dados, *Data Marts*, *Data Warehouses* ou *Data Lakes* (ESCOVEDO; KOSHIYAMA, 2020, p. 10). Os custos desses dados podem variar, podendo estar disponíveis de graça ou exigir esforços para serem obtidos, podendo serem comprados ou simplesmente não existirem, necessitando de projetos auxiliares para sua obtenção (PROVOST; FAWCETT, 2016, p. 28).

Realiza-se também uma análise inicial, denominada análise exploratória, com o objetivo de conhecer o conjunto de dados, com o uso de ferramentas que possuam a capacidade de mostrar diferentes aspectos dos dados, bem como demonstrar as imprecisões e desvios existentes nos dados. A partir desta análise, estratégias de pré-processamento devem ser definidas e, a informação adquirida na análise exploratória apoia a tomada de decisão sobre o tipo de tarefa de mineração de dados e sobre o algoritmo mais adequado a determinado contexto (SILVA; PERES; BOSCARIOLI, 2016, p. 29-30).

3.4.3. Pré-Processamento

As tecnologias analíticas utilizadas são poderosas e podem exigir determinados requisitos sobre os dados que usam, frequentemente em um formato diferente do que são fornecidos naturalmente, sendo necessária alguma conversão (PROVOST; FAWCETT, 2016, p. 29).

Podem ser realizadas operações de ETL (Extração, Transformação e Carga), com a finalidade de prepará-los para os modelos que serão construídos, podendo remover dados faltantes, corrigir ou amenizar dados discrepantes e selecionar as variáveis e instâncias mais adequadas ao modelo. Esta etapa é a mais demorada e trabalhosa de um projeto de *Data Science*, consumindo cerca de 70% do tempo do projeto (ESCOVEDO; KOSHIYAMA, 2020, p. 11).

3.4.4. Modelagem e Inferência

Na etapa de modelagem é que as técnicas de mineração de dados são aplicadas (PROVOST; FAWCETT, 2016, p. 31). Consiste em elencar os modelos possíveis e passíveis a cada tipo de problema, estimar os parâmetros que compõem os modelos, com base nas instâncias e variáveis pré-processadas e avaliar os resultados de cada modelo (ESCOVEDO; KOSHIYAMA, 2020, p. 11).

3.4.5. Pós-Processamento

Etapa de avaliação, que tem por objetivo estimar os resultados de mineração de dados de formar rigorosa, obtendo a confiança de que são válidos e confiáveis. Deve-se ter a confiança de que os modelos e padrões extraídos dos dados são regularidades verdadeiras e não apenas idiossincrasias ou anomalias da amostra. Igualmente importante, essa fase serve para ajudar a garantir que o modelo satisfaça o objetivo de solucionar o problema levantado sobre o negócio (PROVOST; FAWCETT, 2016, p. 31).

Esta etapa inclui avaliações quantitativas e qualitativas, combinando as heurísticas do negócio com os modelos ajustados, tendo em vista os pontos fortes e dificuldades encontradas em cada modelo (ESCOVEDO; KOSHIYAMA, 2020, p. 11). Pode revelar que os resultados não são bons o suficiente para implantação e, pode ser necessário ajustar a definição do problema ou obter dados diferentes (PROVOST; FAWCETT, 2016, p. 34).

3.4.6. Apresentação dos Resultados

O sucesso do projeto depende da transmissão dos resultados obtidos, em linguagem de negócios, já que darão suporte ao processo de tomada de decisão, e devem ser transmitidos de forma direta, ágil e objetiva (SANTANA, 2019). A visualização dos resultados tem o papel de contar a história de toda a análise realizada. Permite resumir informações, comunicar de forma mais efetiva, compreender, explorar, interpretar e analisar (AMARAL, 2016, p.127).

3.4.7. Implantação do Modelo

Nesta etapa os resultados da mineração de dados e, cada vez mais, as próprias técnicas de mineração de dados são colocadas em uso real. Geralmente requer que o modelo seja recodificado para o ambiente de produção, geralmente para maior velocidade ou compatibilidade com um sistema existente (PROVOST; FAWCETT, 2016, p. 32-33).

4. MINERAÇÃO DE DADOS

Mineração de dados é um ramo da computação que teve início nos anos 80, surgindo a partir da preocupação sobre os grandes volumes de dados estocados e inutilizados nas empresas e organizações, sendo que inicialmente consistia em extrair informação de gigantescas bases de dados da maneira mais automatizada possível (AMO, 2004, p. 2).

A Mineração de dados (do inglês *Data Mining*) pode ser definida como um processo automático ou semiautomático de explorar analiticamente grandes bases de dados, com a finalidade de descobrir padrões relevantes que ocorrem nos dados e que possuam importância para embasar a assimilação de informações importantes, dando suporte à geração de conhecimento para o processo de tomada de decisão (SILVA; PERES; BOSCARIOLI, 2016, p. 10). Utiliza-se de técnicas e reconhecimento de padrões, estatística e outras ferramentas matemáticas, por meio do aprendizado de máquina (ESCOVEDO; KOSHIYAMA, 2020, p. 2).

É uma disciplina interdisciplinar e multidisciplinar que envolve conhecimento de áreas de banco de dados, estatística, aprendizado de máquina, computação de alto desempenho, reconhecimento de padrões, computação natural, visualização de dados, recuperação de informação, processamento de imagens e sinais, analise espacial de dados, inteligência artificial e outras (CASTRO; FERRARI, 2016, p. 7) e não existe uma linha nítida separando a mineração de dados destas outras áreas (ELMASRI; NAVATHE, 2011, p. 700).

Pode ser entendida também como procedimentos aplicados em bases de dados para a descoberta de padrões, aplicando técnicas implementadas por meio de algoritmos computacionais, capazes de receber como valores de entrada fatos ocorridos no mundo real e devolver como saída um padrão de comportamento como, por exemplo uma regra de associação ou a modelagem de um perfil. Dependendo do tipo de dado ou do conhecimento que se deseja descobrir, a mineração de dados oferece diferentes soluções e possibilidades e, por conta disso, é comumente dividida em tarefas, de modo a situar o problema junto aos diferentes algoritmos de análise de dados disponíveis e que tipo de padrão se deseja descobrir (SILVA; PERES; BOSCARIOLI, 2016, p. 11).

As funcionalidades da mineração de dados são usadas para especificar os tipos de informações a serem obtidas nas tarefas de mineração, podendo ser caracterizadas

como: **Preditivas**, que utilizam inferência a partir dos dados para fazer predições de valores futuros ou desconhecidos de outras variáveis de interesse; ou **Descritivas**, que caracterizam as propriedades gerais dos dados, buscando padrões que descrevam os dados ou descobrir um modelo a partir dos dados selecionados (CASTRO; FERRARI, 2016, p. 7). Pode ter como objetivos de sua realização encontrar padrões e a construção de modelos, ou para a utilização dos resultados encontrados, devendo os dois processos serem mantidos distintos (PROVOST; FAWCETT, 2016, p. 25).

4.1. KNOWLEDGE DISCOVERY IN DATABASES (KDD)

A mineração de dados pode ser considerada parte de um processo maior, denominado Descoberta de Conhecimento em Bases de Dados, ou KDD (*Knowledge Discovery in Databases)* (DIAS, 2002, p. 1716), processo que tem por objetivo transformar os dados brutos em informações úteis, gerando conhecimento para a organização (ESCOVEDO, KOSHIYAMA, 2020, p. 7).

A Descoberta de Conhecimento em Bases de Dados foi definida por Fayyad et al. (1996) como o processo não trivial de identificar padrões em dados que sejam válidos, novos, potencialmente de úteis e compreensíveis, de modo a melhorar o entendimento de um problema ou um processo de tomada de decisão. Tem por objetivo encontrar padrões intrínsecos aos dados, apresentando-os de forma a facilitar sua assimilação, sendo associado a um processo analítico, sistemático e, quando possível, automatizado (SILVA, PERES, BOSCARIOLI, 2016, p. 11). Na maioria das vezes, devido ao grande volume de dados, o processamento manual torna-se impraticável e, o KDD é uma tentativa de solucionar o problema causado pela sobrecarga de dados (CAMILO; SILVA, 2009, p. 3).

Segundo Elmasri e Navathe (2011, p. 699), a Descoberta de Conhecimento em Banco de Dados abrange mais do que a Mineração de Dados. Refere-se a todo o processo de extração de conhecimento em base de dados, compreendendo cinco fases: seleção e integração de dados, limpeza da base, seleção e transformação dos dados, mineração de dados e a avaliação dos dados (CASTRO; FERRARI, 2016, p.5). A Figura 5 ilustra o processo de KDD.

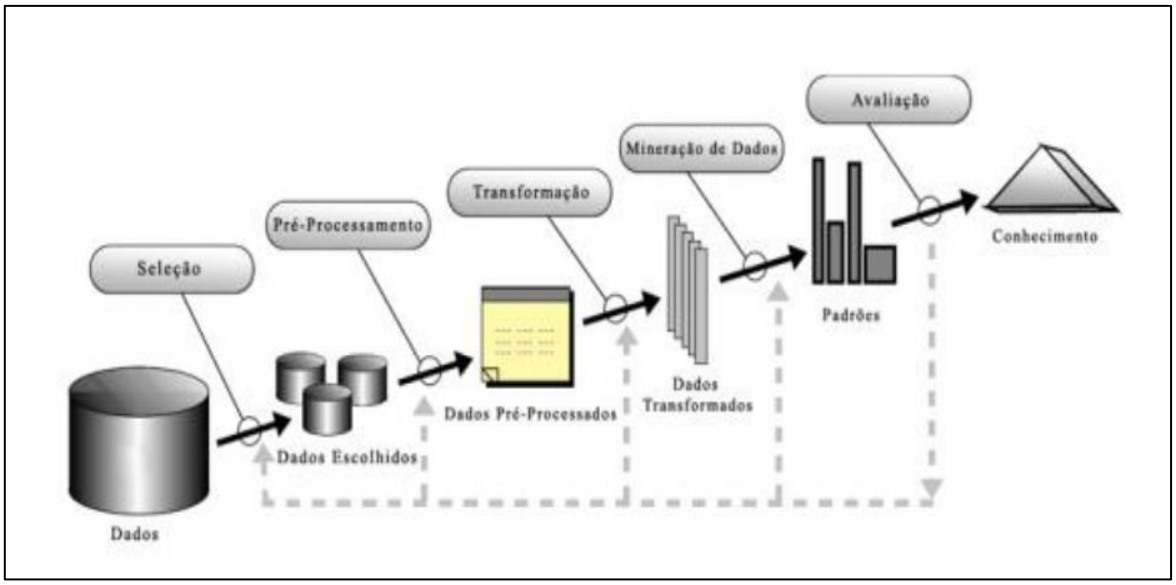

Figura 5: Processo KDD (CAMILO; SILVA, 2009)

Comparando o KDD e a mineração de dados pode-se dizer que o primeiro representa o processo de tornar dados de baixo nível em conhecimento de alto nível, enquanto o segundo pode ser definida como o processo de extrair padrões ou modelos dos dados observados (DIAS, 2002, p. 1716). Por outro lado, ciência de dados e KDD se confundem com relação às suas definições, não possuindo características marcantes que diferenciem as suas áreas (PAIXÃO; SILVA; TANAKA, 2015, p. 9).

4.2. TAREFAS DE MINERAÇÃO DE DADOS

A mineração de dados é comumente classificada de acordo com sua capacidade de realizar determinadas tarefas (CAMILO; SILVA, 2009, p. 8). As principais tarefas dentro desse contexto são: análise descritiva ou sumarização, classificação, estimativa ou regressão, agrupamento e associação. A partir do resultado que se deseja obter, é definida qual técnica será utilizada, sendo que cada uma é utilizada para determinada situação.

4.2.1. Análise Descritiva ou Sumarização

Etapa inicial do processo de mineração de dados, que não requer elevado nível de sofisticação, já que usa ferramentas capazes de medir, explorar e descrever características intrínsecas aos dados, permitindo uma sumarização e compreensão dos

objetos da base e seus atributos (CASTRO; FERRARI, 2016, p. 8). Busca descrever padrões e tendências revelados pelos dados, oferecendo uma possível interpretação para os resultados obtidos sendo utilizada em conjunto com técnicas de análise exploratória de dados, para comprovar a influência de certas variáveis no resultado obtido (CAMILO; SILVA, 2009, p. 8). Aspectos importantes dos conjuntos podem ser obtidos por meio da aplicação de medidas de tendência central, dispersão e correlação, ou pelo uso de recursos gráficos para a visualização dessas e de outras medidas (SILVA; PERES; BOSCARIOLI, 2016, p. 30).

4.2.2. Classificação

É uma das tarefas mais comumente utilizada em mineração de dados, que visa identificar qual classe um determinado registro pertence. Para isso o modelo analisa o conjunto de registros fornecidos os quais é conhecida a indicação de qual classe pertence e, o modelo aprende como classificar um novo registro, a partir dos registros e classificação existente (SILVA, 2014, p. 7). Utiliza-se de treinamento supervisionado já que os rótulos das classes de dados de treinamento são conhecidos *a priori* e usados para ajustar o modelo de predição (CASTRO; FERRARI, 2016, p. 8). A tarefa de classificação pode ser dividida em classificação binária, quando o registro será classificado entre dois rótulos e classificação multiclasse, quando mais de dois rótulos são possíveis para o conjunto de dados (SILVA; PERES; BOSCARIOLI, 2016, p. 79).

4.2.3. Estimativa ou Regressão

Similar à classificação, porém utilizada quando o valor da classe que se deseja determinar é numérico e não categórico. Consiste em utilizar um modelo que permita estimar o valor de determinada variável analisando os valores das demais classes (CAMILO; SILVA, 2009, p. 9). Os modelos de regressão podem ser divididos em: linear simples ou multivariado, ou não linear simples ou multivariado. A diferença entre linear simples ou não linear está no tipo de função a ser utilizada enquanto a diferença entre simples e multivariado está na quantidade de atributos utilizados para estimar o valor da variável alvo (SILVA; PERES; BOSCARIOLI, 2016, p. 117).

4.2.4. Agrupamento

Agrupamento ou *clustering* em inglês é o nome dado ao processo de separar, particionar ou segmentar um conjunto de objetos em grupos de objetos similares. Seu objetivo é formar grupos baseados no princípio de que esses grupos devem ser o mais homogêneo possível entre os membros do grupo e o mais heterogêneo possível entre os grupos (CÔRTES; LIFSCHITZ, 2002, p. 8). Portanto, um agrupamento pode ser definido como uma coleção de objetos similares uns aos outros e dissimilares aos objetos pertencentes a outras classes.

Diferentemente da tarefa de classificação, o agrupamento considera dados de entrada não rotulados, ou seja, não se sabe a qual grupo cada registro pertence *a priori*, sendo este processo utilizado para identificar esses grupos. Devido a esse não conhecimento dos rótulos, esse processo é denominado de treinamento não supervisionado (CASTRO; FERRARI, 2016, p. 9).

4.2.5. Associação

O objetivo é encontrar relações entre os atributos e não entre os objetos, buscando ocorrências frequentes e simultâneas entre os elementos de um contexto (SILVA; PERES; BOSCARIOLI, 2016, p. 14), num formato SE atributo X ENTÃO atributo Y (CAMILO; SILVA, 2009, p. 10).

Também conhecida por mineração de regras de associação, corresponde à descoberta de regras de associação que apresentam valores de atributos que ocorrem concomitantemente em uma base de dados. Esse tipo de análise costuma ser usado em ações de marketing e para estudo de bases de dados transacionais CASTRO; FERRARI, 2016, p. 9).

4.2.6. Detecção de Anomalias (*Outliers*)

Um banco de dados pode conter dados que não apresentam o comportamento geral da maioria e, dados desse formato são denominados anomalias ou *outliers* (AMO, 2004, p. 4). Uma anomalia é um valor discrepante, localizado significativamente longe dos valores considerados normais (CASTRO; FERRARI, 2016, p. 269).

A maioria das ferramentas de mineração de dados descartam essas anomalias como sendo ruído indesejado ou exceções. A detecção de anomalias é usada para detectar e, quando apropriado, executar alguma tomada de decisão sobre objetos anômalos da base de dados (CASTRO; FERRARI, 2016, p. 269). As principais aplicações de detecções de anomalias são: detecção de fraudes em transações de cartões de crédito, telefones celulares, consumo de energia entre outros; análise de crédito detectando clientes potencialmente problemáticos; detecção de intrusão a redes de computadores e ambientes diversos; desempenho de rede identificando gargalos; diagnóstico de falhas em motores, geradores, redes e análise de imagens e vídeos identificando novas características, objetos ou comportamentos.

A detecção de anomalias funciona basicamente como uma classificação binária, na qual se deseja determinar se um ou mais objetos pertencem à classe normal ou à classe anômala. Além das etapas normais ainda devem ser consideradas as etapas de definição de anomalia e do tipo de abordagem a ser utilizada (CASTRO; FERRARI, 2016, p. 273). Os métodos de detecção de anomalias podem ser divididos em: métodos estatísticos que geram um modelo probabilístico dos dados e testam se o objeto foi gerado por tal modelo ou não, e métodos algorítmicos que são baseados em algoritmos de mineração de dados, aplicados à base a procura de objetos anômalos.

4.3. PRÉ-PROCESSAMENTO DE DADOS

Os dados submetidos à mineração de dados podem ser provenientes de diferentes fontes, tais como sistemas transacionais, observações de situações do mundo real ou processos de medição em geral. Devido a isso, não raramente esses dados apresentam algumas falhas ou são organizados de forma inadequada para serem submetidos a determinado algoritmo (SILVA; PERES; BOSCARIOLI, 2016, p. 42). Os dados brutos são aqueles que ainda não foram processados para o uso e neles, podem conter basicamente três problemas: incompletude, inconsistência e ruído, os quais futuramente atrapalharão no processo de análise e mineração desses dados (CASTRO; FERRARI, 2016, p. 26).

Condições como estas geram dificuldade no processo de descoberta de conhecimento e frequentemente levam ao fracasso da análise pretendida. Portanto, torna-se necessário executar procedimentos que pré-processam os dados, corrigindo inconsistências e,

muitas vezes, agregando valor aos dados, dando condições ao processo de descoberta de conhecimento e resultados de qualidade (SILVA; PERES; BOSCARIOLI, 2016, p. 42).

Conhecer e preparar os dados de forma adequada é a etapa chamada de pré-processamento de dados e tem por motivação tornar o processo de mineração muito mais eficaz e eficiente (CASTRO; FERRARI, 2016, p. 27). Este processo depende de três fatores principais: dos problemas encontrados nos dados, de qual problema a ser resolvido e de qual técnica de mineração de dados será empregada (CASTRO; FERRARI, 2016, p. 34). Essa etapa demanda muito tempo e bastante trabalho, mas o sucesso da mineração depende fortemente do cuidado dedicado e ela. Nesta seção serão citadas as principais técnicas utilizadas na etapa de pré-processamento dos dados.

4.3.1. Limpeza

A etapa de limpeza dos dados visa eliminar os problemas de registros incompletos, valores errados e dados inconsistentes, de modo que eles não influam no resultado dos algoritmos usados (SILVA, 2014, p. 23). As técnicas usadas nesta etapa variam de acordo com o problema encontrado: para valores ausentes pode-se usar a remoção da linha com dados ausentes, preenchimento manual dos valores ou preenchimento automático dos valores e, para valores ruídos ou inconsistentes pode-se realizar inspeção e correção manual ou identificação e limpeza automática (SILVA; PERES; BOSCARIOLI, 2016, p. 43).

4.3.2. Integração

Dados obtidos para serem utilizados em processo de mineração de dados comumente são obtidos de diversas fontes, como por exemplo, banco de dados, arquivos textos, planilhas, *Data Warehouses*, vídeos, imagens, entre outras. Surge então, a necessidade da integração destes dados de forma a termos um repositório único e consistente (CAMILO; SILVA, 2009, p. 6). Problemas de redundância, duplicidade e conflitos entre domínios de atributos devem ser observados ao se realizar o processo de integração dos dados (CASTRO; FERRARI, 2016, p. 43-44).

4.3.3. Redução

O volume de dados usado na mineração costuma ser alto, sendo em alguns casos, tão grande que torna o processo de análise dos dados e da própria mineração impraticável. Técnicas de redução são aplicadas para que a massa de dados original seja convertida em uma massa de dados menor, sem perder a representatividade dos dados originais (SILVA, 2014, p. 23). Este processo permite que os algoritmos de mineração sejam executados com mais eficiência, mantendo a qualidade do resultado. As estratégias adotadas nesta etapa são: criação de estruturas otimizadas para os dados (cubos de dados), seleção de um subconjunto dos atributos, redução da dimensionalidade e a discretização. (CAMILO; SILVA, 2009, p. 7).

4.3.4. Transformação

Bases de dados brutas e integradas podem sofrer com valores ausentes, ruídos e inconsistências de dados não ou pouco padronizados, podendo um mesmo atributo ser descrito de formas diferentes ou utilizando unidades distintas (CASTRO; FERRARI, 2016, p. 50). Outro problema encontrado é que alguns algoritmos trabalham apenas com valores numéricos e outros apenas com valores categóricos. Nestes casos, é necessário transformar os valores numéricos em categóricos ou os categóricos em valores numéricos (SILVA, 2014, p.23).

Não existe um critério único para transformação dos dados e diversas técnicas podem ser usadas de acordo com os objetivos pretendidos. Algumas das técnicas empregadas nesta etapa são: padronização (resolver diferenças de unidades e escalas), suavização (remover valores errados dos dados), agrupamento (agrupar valores em faixas sumarizadas), generalização (converter valores muito específicos para valores mais genéricos), normalização (colocar as variáveis em uma mesma escala) e a criação de novos atributos (gerados a partir de outros já existentes).

4.3.5. Discretização

Alguns algoritmos de mineração operam apenas com atributos categóricos e, portanto, não podem ser aplicados a dados numéricos. Em situações assim os atributos numéricos podem ser discretizados, dividindo o domínio do atributo em intervalos e ampliando a

quantidade de métodos de análise disponíveis para aplicação (CASTRO; FERRARI, 2016, 52). A discretização também reduz a quantidade de valores de um dado atributo contínuo, facilitando, em muitos casos, o processo de mineração.

4.4. ALGORITMOS E TÉCNICAS DE MINERAÇÃO

Um modelo é a representação simplificada da realidade criada para servir a um propósito, com base em alguns pressupostos do que é ou não importante para a finalidade específica. Em ciência de dados um modelo preditivo é um modo de estimar o valor desconhecido de interesse, podendo ser uma fórmula matemática ou uma declaração lógica ou regra (PROVOST; FAWCETT, 2016, p. 44-45).

O aprendizado de máquina é a aplicação de técnicas computacionais na tentativa de encontrar padrões ocultos nos dados e este, pode ser supervisionado ou não supervisionado. Sendo o aprendizado supervisionado aquele em que existe uma classe ou atributo conhecido ao qual se quer prever de outros objetos, são os casos das tarefas de mineração de dados da classificação e regressão. No caso do aprendizado não supervisionado não há uma classe específica à qual se quer prever, sendo o caso das tarefas de agrupamento e associação. Análises do tipo descritivo ou sumarização utiliza-se aplicação de ferramentas estatísticas ou de visualização de dados.

Segue na sequencia um resumo sobre as principais técnicas e algoritmos utilizados na mineração de dados.

4.4.1. Análise Descritiva

É utilizada para descrever, simplificar ou sumarizar as principais características de uma base de dados, sendo o princípio de qualquer análise quantitativa de dados. A principal diferença entre a análise descritiva e as outras tarefas de mineração de dados é que esta visa descrever e encontrar o que há nos dados enquanto as demais tarefas buscam conclusões que extrapolam os dados e permitem inferir algo a partir destes (CASTRO; FERRARI, 2016, p. 60). Baseada em técnicas estatísticas, a análise descritiva pode ser desmembrada em três partes principais:

Distribuição de Frequência: resumo dos dados em classes mutuamente exclusivas, ajudando a entender a natureza dos dados. É realizado o agrupamento dos dados em

classes; cálculo de amplitude de classes, limites inferiores e superiores; frequências absoluta, relativa e acumulada.

Visualização de Dados: apresentação dos dados em forma pictórica ou gráfica, com o objetivo de entender a natureza das distribuições das classes e extrair conhecimento mais fácil e rápido. Os dados podem ser visualizados na forma de diagramas de caixa ou *boxplots*, histogramas, polígonos de frequência, gráficos de setores ou de dispersão

Medidas de Resumo: são medidas que resumem a informação contida em uma distribuição de probabilidade ou sumarizam a informação contida em uma base de dados. Podem ser utilizadas medidas de tendência central (média, mediana, ponto médio, moda), de dispersão (amplitude, desvio médio, variância, coeficiente de variação), de distribuição (assimetria, curtose, quartis) e de associação (covariância, coeficiente de correlação, coeficiente de concentração).

4.4.2. *k-Means* (k-Médias)

Esse algoritmo tem por objetivo encontrar partições no conjunto de dados de forma que possua **k** grupos distintos, sendo que **k** é o parâmetro de entrada. A busca pela descoberta das **k** partições é iterativa, sendo iniciada pela escolha de **k** vetores distintos e aleatórios, para representar os centroides dos grupos (SILVA; PERES; BOSCARIOLI, 2016, p. 157). Centroide é o valor médio dos objetos em um grupo (CASTRO; FERRARI, 2016, p. 116). Em seguida o algoritmo procede verificando quais exemplares são mais similares a quais centroides, ajustando os valores de centroides a cada iteração, considerando os exemplares mais similares a eles. O algoritmo k-médias é o mais popular dentre os algoritmos para a tarefa de agrupamento.

4.4.3. *k-Medoids* (k-Medoides)

É uma variação do k-Médias, sendo que o algoritmo ao invés de utilizar o centro do agrupamento como referência, utiliza-se o objeto mais central da partição, denominado medoide (CAMILO; SILVA, 2009, p. 18). Um medoide é o objeto com a menor dissimilaridade média a todos os outros objetos, ou seja, é o objeto mais centralmente localizado no grupo. Uma diferença importante entre os dois métodos é que o k-medoides escolhe objetos da própria base como centros dos grupos, enquanto que o k-médias

calcula o centro dos grupos a partir dos objetos nele contidos. O algoritmo k-medoides é mais robusto a ruído e a valores discrepantes, já que o centro do grupo necessariamente será parte da base, sendo assim ruídos e valores discrepantes não influenciarão tão fortemente na definição do centro (CASTRO; FERRARI, 2016, p. 119).

4.4.4. *Fuzzy* k-Médias

É uma extensão do algoritmo k-médias, sendo que neste caso cada objeto possui um grau de pertinência em relação a todos os grupos da base, podendo cada objeto pertencer a mais de um grupo, com variados graus de pertinência. Cada objeto da base possui um grau de pertinência a cada grupo da base e, a soma dos graus de pertinência de um objeto a todos os grupos da base deve ser igual a 1, valor máximo de pertinência neste algoritmo (CASTRO; FERRARI, 2016, p. 122-123).

4.4.5. Árvore Geradora Mínima (MST)

Este método particional é baseado em teoria dos grafos, mais especificamente no conceito de árvores geradoras mínimas (*minimal spanning tree* – MST) para segmentar a base em diferentes grupos. Inicialmente constrói-se a árvore geradora mínima dos dados de entrada, sendo que os nós correspondem às coordenadas dos objetos e as arestas à distância ou similaridade entre eles. Logo após é definido um critério de inconsistência para as arestas, de modo que as arestas inconsistentes sejam removidas, resultando em subgrafos ou subárvores que são os grupos dos objetos da base. Para a utilização deste método a base de dados deve estar representada na forma numérica (CASTRO; FERRARI, 2016, p. 127).

4.4.6. DBSCAN

Agrupamentos por densidade são especialmente úteis para a aplicação em conjuntos de dados com um grande número de exemplares e também para a descoberta de grupos de formatos arbitrários. Um dos principais algoritmos que utiliza a estratégia de analisar aspectos relacionados com densidade para formar grupos é o DBSCAN, do inglês *Density Based Spatial Clustering of Aplications*. Esse algoritmo é caracterizado por guiar o processo de descoberta de grupos com base na densidade de exemplares existentes na

vizinhança de um exemplar que já pertença ao grupo, considerando que essa densidade nessa vizinhança seja igual ou maior a um limite mínimo que estabelece se um exemplar pode ser parte do grupo associado ao exemplar referencial (SILVA; PERES; BOSCARIOLI, 2016, p. 161-162). O número de grupos é definido automaticamente pelo algoritmo, de acordo com a densidade, podendo não classificar elementos em nenhum grupo, elementos esses definidos com ruídos (AMARAL, 2016, p. 110).

4.4.7. Agrupamento Hierárquico

O agrupamento hierárquico é obtido a partir da criação de uma estrutura em árvore, na qual os exemplares de um conjunto de dados são definidos como nós folhas, e os nós internos mostram uma organização baseada na similaridade entre os exemplares e, em cada nível da árvore temos a organização dos dados em grupos. A árvore é comumente representada graficamente por um dendrograma (SILVA; PERES; BOSCARIOLI, 2016, p. 148).

As estratégias para agrupamento hierárquico podem ser implementadas a partir de uma de duas abordagens: aglomerativa ou divisiva. Na aglomerativa, a cada ciclo os pontos são unidos com os mais próximos e na divisiva inicia-se com todas as instâncias e, a cada etapa, são executadas divisões (AMARAL, 2016, p. 111).

4.4.8. Classificador k-NN

O algoritmo k-Vizinhos Mais Próximos, k-NN do inglês k-*Nearest Neighbours*, funciona muito bem na prática para problemas de classificação e regressão. Sua ideia principal é considerar que os exemplares vizinhos são similares ao exemplo cuja informação se deseja inferir (ESCOVEDO, KOSHIYAMA, 2020, p. 99).

No k-NN o conjunto de dados rotulados é inicialmente armazenado e, quando um novo registro deve ser classificado, ele é comparado a todos os registros do conjunto de treinamento para identificar os **k** vizinhos mais próximos de acordo com alguma métrica de distância. A classe deste novo registro é determinada de acordo com as classes destes **k** vizinhos mais próximos. Na maioria das implementações do k-NN, os atributos são normalizados de modo que tenham a mesma contribuição na predição da classe ou valor (ESCOVEDO, KOSHIYAMA, 2020, p. 100).

4.4.9. Árvores de Decisão

Uma árvore de decisão é uma estrutura em forma de árvore na qual cada nó interno corresponde a um teste de um atributo e cada ramo representa um resultado do teste e, os nós folhas representam classes ou distribuições de classes (CASTRO; FERRARI, 2016, p. 170). Após a árvore de decisão ser montada, basta seguir o fluxo na árvore, do nó raiz até um nó folha, realizando os testes nos nós internos, para classificar um novo registro CAMILO; SILVA, 2009, p. 12).

Árvores de decisão possuem a vantagem de serem normalmente concisas, de fácil visualização e compreensão, sendo fácil de explicar as classificações propostas (CASTRO; FERRARI, 2016, p. 170). Existem diferentes algoritmos que produzem árvores de decisão aos quais podemos citar J48, ADTree, CHAID, CTree, C4.5, CART e Hoeffding Tree, sendo todos bem parecidos entre si. A construção da árvore é realizada de acordo com alguma abordagem recursiva de particionamento dos conjuntos de dados e a principal distinção encontra-se nos processos de seleção de variáveis, critérios de particionamento e critério de parada de crescimento da árvore (ESCOVEDO, KOSHIYAMA, 2020, 94-95).

4.4.10. Classificadores Bayesianos

São algoritmos baseados na técnica estatística de probabilidade condicional baseada na teoria de Thomas Bayes. Segundo o teorema de Bayes, é possível encontrar a probabilidade de um certo evento ocorrer, dada a probabilidade de um outro evento que já ocorreu (CAMILO; SILVA, 2009, p. 13). São algoritmos que apresentam bom desempenho em problemas de classificação tanto com dados categóricos quanto com dados numéricos.

O classificador bayesiano mais utilizado é o Naïve Bayes, que presume que se o valor de um atributo exerce algum efeito sobre a distribuição de classes existente no conjunto, esse efeito é independente dos valores assumidos por outros atributos e de seus respectivos efeitos sobre a mesma distribuição de classes. Em um processo de classificação, o algoritmo Naïve Bayes tomará a decisão sobre qual a classe o exemplar com rótulo desconhecido está associado, por meio do cálculo da probabilidade condicional dele pertencer a cada uma das classes existentes no conjunto de dados usado para treinamento (SILVA; PERES; BOSCARIOLI, 2016, p. 112).

4.4.11. Máquina de Vetores de Suporte (SVM)

Máquina de Vetor de Suporte ou *Support Vector Machine* (SVM) em inglês, é um dos algoritmos mais efetivos para classificação, podendo ser aplicado em dados lineares ou não lineares. Essencialmente o SVM realiza um mapeamento não linear para transformar os dados de treino originais em uma dimensão maior, onde busca pelo hiperplano que separa os dados linearmente de forma ótima. O SVM encontra este hiperplano usando vetores de suporte e margens, definidas pelos vetores de suporte (ESCOVEDO, KOSHIYAMA, 2020, p. 127).

A máquina de vetor de suporte maximiza a margem entre as instâncias mais próximas, criando um vetor otimizado para classificá-las. São bastante eficientes para minimizar superajustes e por suportar muitos atributos, sendo assim menos suscetíveis à erros de dimensionalidade (AMARAL, 2016, p. 102-103)

4.4.12. Regressão Linear, Polinomial e Não Linear

A regressão modela a relação entre uma ou mais varáveis de reposta, chamadas também de variáveis de saída, dependentes, preditas ou explicadas, com os preditores, também chamados de variáveis de controle, independentes, explanatórias ou regressores (CASTRO; FERRARI, 2016, p. 205). Em outras palavras, através das variáveis de entrada é possível prever o valor da variável de saída, por meio de um modelo matemático.

Modelos de regressão linear consideram que o valor da variável de resposta pode ser estimado por uma combinação linear das variáveis preditoras. Sendo assim, deve-se então encontrar valores para os coeficientes de regressão de forma que a reta se ajuste aos valores assumidos pelas variáveis nos exemplares de um conjunto de dados (SILVA; PERES; BOSCARIOLI, 2016, p. 118). A saída do modelo será um valor numérico contínuo, que deve ser o mais próximo possível do valor desejado, sendo que a diferença entre esses valores fornece uma medida de erro do algoritmo (ESCOVEDO, KOSHIYAMA, 2020, p. 192).

Regressão polinomial é um modelo de regressão no qual a relação entre as variáveis independentes e a variável dependente pode ser não linear e tem a forma de um polinômio de grau *n*. Nesses casos, embora o polinômio de aproximação seja não linear, o problema de estimação dos parâmetros do modelo é linear, desse modo o método

também é considerado uma forma de regressão linear (CASTRO; FERRARI, 2016, p. 208).

Quando a relação entre a variável dependente e as variáveis independentes de um conjunto de dados não segue uma relação linear e, sim uma combinação não linear esse modelo é chamado de regressão não linear. Pode ter a forma exponencial, logarítmica e de potência, sendo nesses casos a principal característica é a capacidade de linearização, por meio de um ajuste apropriado dos parâmetros (SILVA; PERES; BOSCARIOLI, 2016, p. 122).

4.4.13. Regressão Logística

Apesar do nome, a regressão logística é um algoritmo utilizado para problemas de classificação. É usado para estimar valores discretos, geralmente binários, com base em um conjunto de variáveis independentes. A regressão logística calcula a probabilidade de ocorrência de um evento, ajustando os dados a uma função *logit*, retornando valores de saída entre 0 (pouco provável) e 1 (muito provável) (ESCOVEDO, KOSHIYAMA, 2020, p. 152).

Similar a regressão linear, utiliza uma equação como representação, onde os valores de entrada são combinados linearmente usando pesos ou valores de coeficiente para prever um valor de saída, modelado em valor binário em vez de valor numérico (ESCOVEDO, KOSHIYAMA, 2020, p. 154). Esse valor de saída binário é o que difere a regressão logística da regressão linear.

4.4.14. Redes Neurais

Redes neurais artificiais é uma área de estudo que busca reproduzir a forma de funcionamento do cérebro dos seres vivos, que dispõe de ampla capacidade de aprender (AMARAL, 2016, p. 101). A simulação do funcionamento das unidades de processamento dos seres vivos, os neurônios, por meio de um computador permite a construção de modelos capazes de resolver tarefas de classificação ou de agrupamento em mineração de dados (SILVA; PERES; BOSCARIOLI, 2016, p. 86).

Neurônios artificiais operam com uma rede de camadas, normalmente uma camada de entrada, uma camada oculta e uma camada de saída, utilizando comunicação

unidirecional da camada de entrada para a de saída e sem comunicação entre os neurônios de uma mesma camada. O neurônio recebe uma entrada alterada por um peso, produz uma resposta que é avaliada quanto ao erro encontrado, sendo este erro utilizado para ajustar o peso de entrada e o processo é repetido várias vezes (AMARAL, 2016, p. 102). São exemplos de redes neurais os tipos Perceptron, Adaline, Multilayer Perceptron (MLP) e Função de Base Radial (RBF), brevemente descritas abaixo:

Perceptron: é a arquitetura mais simples de rede neural sendo capaz de classificar padrões linearmente separáveis. Consiste em uma rede neural com uma única camada de pesos, ou seja, uma camada de entrada e outra de saída, com pesos sinápticos e bias ajustáveis (CASTRO; FERRARI, 2016, p. 209).

Adaline: similar ao Perceptron, porém com os neurônios usando função de ativação linear em vez de função sinal, tendo essa alteração o objetivo de minimizar o erro quadrático médio entre a saída da rede e a saída desejada (CASTRO; FERRARI, 2016, p. 209).

Multilayer Perceptron: é uma rede neural do tipo Perceptron com pelo menos uma camada intermediária. O treinamento desse modelo utiliza um algoritmo denominado Backpropagation, baseado na retropropagação de erro (*error backpropagation*), onde o erro obtido na camada de saída, comparando o resultado obtido e o esperado, é propagado de volta até a camada de entrada, propiciando as alterações necessárias nos pesos das sinapses (SILVA; PERES; BOSCARIOLI, 2016, p. 91-92).

Função de Base Radial: nesse modelo o projeto das redes com múltiplas camadas e propagação positiva do sinal é visto como um problema de aproximação de função em um espaço multidimensional. Desse modo, a aprendizagem é equivalente a encontrar uma hipersuperfície em um espaço multidimensional que forneça a melhor aproximação para os dados de treinamento (CASTRO; FERRARI, 2016, p. 222).

4.4.15. Algoritmo Apriori

A frequência de um item em uma base de dados transacional é o elemento-chave para o desenvolvimento de algoritmos para a descoberta de regras de associação (SILVA; PERES; BOSCARIOLI, 2016, p. 202-203). Comumente a estratégia adotada pelos algoritmos de mineração de regras de associação é decompor o problema em duas

subtarefas: geração do conjunto de itens frequentes e a geração das regras de associação.

O algoritmo Apriori é o método mais conhecido para a mineração de regras de associação, empregando busca em profundidade e gerando conjuntos de itens frequentes percorrendo os conjuntos de itens frequentes iterativamente em ordem crescente de tamanho, gerando e testando cada conjunto até encontrar os frequentes. Ele se utiliza de um princípio importante: se um conjunto de itens é frequente, então todos os seus subconjuntos também são frequentes e, consequentemente, todos os subconjuntos do conjunto frequente também são frequentes. Esse princípio tem o nome de propriedade Apriori (CASTRO; FERRARI, 2016, p. 247-249).

4.4.16. Algoritmo FP-Growth

O algoritmo Apriori, em sua implementação, depende da análise da totalidade da base de dados várias vezes, a fim de encontrar os conjuntos de itens frequentes, tornando o processo computacionalmente caro e torna os conjuntos de itens frequentes números quando o conjunto total de itens é grande. Como alternativa o algoritmo FP-Growth (*Frequent-pattern growth*) permite que os conjuntos de itens frequentes sejam descobertos sem a necessidade da criação dos conjuntos de itens candidatos (SILVA; PERES; BOSCARIOLI, 2016, p. 211).

O algoritmo FP-Growth é baseado em uma estrutura em forma de árvore de prefixos para padrões frequentes, a qual armazena de forma comprimida a informação sobre padrões frequentes. É baseado em três aspectos centrais: a compressão da base de dados em uma estrutura de árvore (FP-Tree), o uso de um algoritmo de mineração da árvore que evite a geração de uma grande quantidade de conjuntos candidatos e o uso de um método particional para decompor a tarefa de mineração em subtarefas menores, de forma a reduzir o espaço de busca significativamente (CASTRO; FERRARI, 2016, p. 253).

4.4.17. Algoritmos Genéticos

Algoritmos genéticos são baseados em processos relacionados com organismos biológicos, onde por meio de muitas gerações, populações de uma mesma espécie evoluem de acordo com os princípios de seleção natural (GOLDSCHMIDT; BEZERRA;

PASSOS, 2015, p. 128). São algoritmos de otimização e busca baseados nos mecanismos de seleção natural e genética, trabalhando com um conjunto de possíveis soluções simultaneamente. Algoritmos genéticos são utilizados para resolver problemas e para agrupar problemas, já que com sua capacidade de resolver problemas em paralelo, se torna uma poderosa ferramenta para a mineração de dados (CÔRTES; PORCARO; LIFSCHITZ, 2002, p. 18-19).

Na modelagem dos algoritmos genéticos cada cromossoma da população representa uma regra e cada gene representa um atributo da base de dados, cada regra gerada é avaliada com relação à sua confiança e à sua abrangência, e o consequente dessa regra apresenta uma conclusão sobre a classe que o objeto alvo pertence. Após a definição das regras por meio de cromossomas, é realizado um cruzamento genético (crossover) entre um par de cromossomas, gerando duas novas regras (GOLDSCHMIDT; BEZERRA; PASSOS, 2015, p. 128).

4.4.18. Mineração de Texto

Textos são formados por dados não estruturados, já que não possuem o tipo de estrutura esperado em dados estruturados, ou seja, tabelas de registros com campos de significados bem definidos. Possuem uma estrutura linguística, destinada aos humanos e não aos computadores e, por conta disso, devem ser submetidos a uma boa quantidade de pré-processamento antes de ser utilizado como entrada de algum algoritmo de mineração de dados (PROVOST; FAWCETT, 2016, p. 252-253).

Um documento é um pedaço de texto, sem importar o tamanho, podendo ter uma única frase ou inúmeras páginas, composto por tokens ou termos individuais. Geralmente todo o texto constante em um documento é considerado como um conjunto, sendo recuperado como item individual quando combinado ou categorizado. O conjunto de documentos é chamado de corpus (PROVOST; FAWCETT, 2016, p. 253).

Um processo de mineração de texto começa com a construção do corpus e, após a criação do corpus, são realizadas operações de remoção de stop words, palavras sem valor semântico para o processo de mineração. Em seguida é produzida uma matriz de termos com suas respectivas frequências, podendo esta ser utilizada para classificar documentos, analisar sentimentos, construir uma nuvem de palavras entre outras (AMARAL, 2016, p. 117-118).

5. PROPOSTA DE TRABALHO

Na Figura 6 é apresentada a proposta do trabalho, de modo a ilustrar as etapas, bem como os passos a serem realizados.

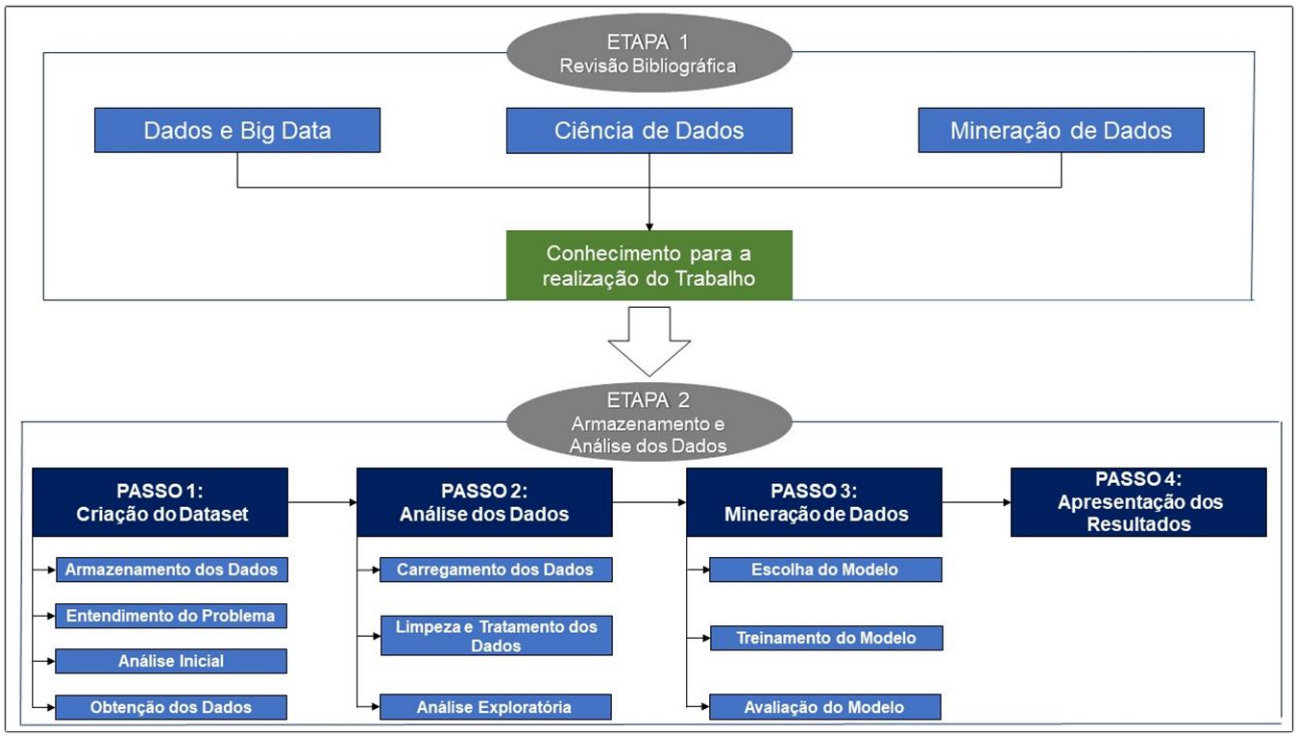

Figura 6: Proposta de Trabalho

5.1. ETAPA 1: REVISÃO BIBLIOGRÁFICA

A primeira etapa do trabalho, já apresentada inicialmente, consiste na revisão bibliográfica dos temas pertinentes para a realização do trabalho, de modo a fornecer subsídios para a compreensão dos temas pertinentes a realização da parte prática, tais como Big Data e ciência de dados, bem como sobre a mineração de dados, abordando as fases de pré-processamento, limpeza e tratamento de dados, analise exploratória e as principais tarefas e técnicas de mineração de dados.

A Etapa 1 tem a sua importância no fato de que, ao realizar a pesquisa destes temas, é desenvolvida uma compreensão de cada etapa e, também da interdependência destas, alicerçando a base de conhecimento para a realização da análise prática proposta na Etapa 2.

5.2. ETAPA 2: ANÁLISE E ARMAZENAMENTO DE DADOS

A segunda etapa é a aplicação do que foi estudado na Etapa 1, consistindo na demonstração de um caso em ciência de dados. Pode-se dividir essa etapa em sete passos que serão realizados, sendo que se pode denominar os mesmos por: criação do *dataset*, limpeza e tratamento dos dados, análise exploratória, mineração de dados, treinamento do modelo gerado, avaliação do modelo gerado e visualização dos resultados. Segue abaixo uma breve descrição do que será realizado nos passos da Etapa 2.

Passo 1 – Criação do *Dataset*: Os dados são gerados, armazenados e coletados para a posterior análise e mineração. Está dividida em 4 subitens:

- **Armazenamento dos dados:** Criação do banco de dados na ferramenta Google BigQuery, sendo que o mesmo carregado com dados brutos.
- **Entendimento do problema:** Levantamento das perguntas a serem respondidas pelas tarefas seguintes.
- **Análise Inicial:** Exploração inicial do conjunto de dados utilizando a ferramenta Google Data Studio, utilizando os dados a partir do BigQuery.
- **Obtenção dos Dados:** Exportação dos dados no formato CSV para utilização na análise de dados.

Passo 2 – Análise dos Dados: Neste passo os dados são carregados, pré-processados e é realizada a análise exploratória. Está dividida em 3 subitens:

- **Carregamento dos Dados:** A partir do arquivo exportado do banco de dados, os dados são carregados na ferramenta Jupyter Notebook, onde será utilizada a linguagem Python e seus módulos.
- **Limpeza e Tratamento dos Dados:** Os dados brutos serão tratados, conforme a necessidade, de modo a remover inconsistências, incompletudes ou ruídos, que podem a vir a atrapalhar as futuras análises.
- **Análise Exploratória:** Nesta etapa, é realizada uma análise utilizando de ferramentas estatísticas de modo a extrair conhecimento da base de dados, conhecimento este que está implícito, necessitando somente desta investigação estatística para ser revelado.

Passo 3 – Mineração de Dados: Será realizada uma análise acerca de qual tarefa será realizada, bem como qual técnica ou algoritmo a ser utilizado, o treinamento e a avaliação do modelo gerado. Está dividida em 3 subitens:

- **Escolha do Modelo:** Definição da tarefa de mineração de dados a ser executada, bem como da técnica e do algoritmo a ser utilizado.
- **Treinamento do Modelo:** Com base na tarefa e técnica ou algoritmo escolhida para a utilização, será criado o modelo utilizando a Linguagem Python, com bibliotecas e ferramentas já existentes, bem como o seu treinamento utilizando os dados reservados para esta finalidade.
- **Avaliação do Modelo:** Após a criação e treinamento do modelo, é realizada a avaliação do modelo criado, utilizando métricas próprias para isto.

Passo 4 – Apresentação dos Resultados: Os resultados obtidos serão apresentados, discutidos e analisados. Neste passo pretende-se são apresentados os *insights* encontrados na base de dados original.

5.3 FERRAMENTAS UTILIZADAS

As ferramentas a ser utilizadas foram definidas conforme a necessidade de armazenamento, análise de dados históricos, exploratória e mineração de dados, utilizando soluções disponíveis no mercado, sendo que algumas necessitam de pagamento, porém com um limite de utilização sem custo e, as demais, a grande maioria são soluções *open-source*. Segue abaixo uma breve descrição de todas as ferramentas utilizadas no trabalho.

5.3.1 BigQuery

O Google BigQuery é um serviço de armazenamento e consulta de grandes conjuntos de dados, acessível por meio do Google Cloud Platform, que utiliza Linguagem SQL (*Structured Query Language*) em conjunto com a capacidade de processamento e infraestrutura da Google para a realização de consultas rápidas em grandes volumes de dados. O armazenamento se dá em várias nuvens, altamente escalonável e sem servidor (GOOGLE, 2021).

5.3.2 Data Studio

O Data Studio é uma ferramenta gratuita da Google para a obtenção de relatórios e painéis informativos a partir de dados que podem ser obtidos de modo simplificado das mais variadas fontes, desde arquivos simples, bancos de dados ou produtos de consumo Google. A partir de uma interface de fácil compreensão, que conta com o recurso de arrastar e soltar, é possível a criação de gráficos de vários tipos e relatórios interativos com filtros de visualização e controles de período, totalmente personalizáveis, de fácil compreensão e compartilhamento (GOOGLE, 2021).

5.3.3 Python

Python é uma linguagem interpretada de alto nível, com sintaxe básica simples e de fácil aprendizado. A Linguagem Python suporta múltiplos paradigmas de programação: imperativo, orientado a objetos e funcional, possui tipagem dinâmica e forte, escopo léxico e gerenciamento automático de memória. Tuplas, listas e dicionários são algumas estruturas de dados embutidas na sintaxe, que aumentam muito a expressividade do código (CRUZ, 2020, p. 3).

Seu surgimento se deu no ano de 1991 e logo se tornou uma das linguagens interpretadas mais populares. Nos últimos anos se tornou uma das linguagens mais importantes em ciência de dados, aprendizado de máquina e desenvolvimento de software em geral, tanto no ambiente acadêmico quanto no mercado. Sua robustez, ampla variedade de bibliotecas, suporte melhorado nos últimos anos e a grande comunidade ativa justificam essa ascensão (MCKINNEY, 2018, p. 20). Neste trabalho foi utilizada a linguagem Python versão 3.9.6.

5.3.4 Anaconda

O Anaconda é uma distribuição que fornece uma plataforma poderosa para a ciência de dados. Além de já possuir a linguagem de programação Python integrada, vem com vários pacotes instalados automaticamente, além de pacotes adicionais que podem ser instalados. Possui também o gerenciador de pacotes, chamado Conda, que monitora e atualiza todos os pacotes e suas versões (COUTINHO, 2020). Todas as bibliotecas

utilizadas nesse projeto, além do próprio Jupyter Notebook já vem de modo nativo no Anaconda.

5.3.5 Jupyter Notebook

Jupyter Notebook é uma ferramenta de web open-source, que oferece um ambiente na forma de notebook, que é um tipo de documento interativo para código, texto (com ou sem marcação), visualizações de dados e outras saídas, organizados de modo a contar uma história. O Jupyter Notebook interage com kernels, que são implementações do protocolo de processamento interativo do Jupyter em várias linguagens de programação, sendo que o kernel do Jupyter para Python utiliza o sistema IPython (MCKINNEY, 2018, p. 40).

5.3.6 NumPy

NumPy é a abreviatura de *Numerical Python* e é um pacote essencial para o processamento de dados com Python. Ele permite o processamento de dados utilizando o formato de *arrays* multidimensionais de modo rápido e eficiente e a realização de funções matemáticas, operações de álgebra linear, transformadas de Fourier e gerações de números aleatórios. Os *arrays* NumPy são eficientes também como um container no qual os dados são passados entre algoritmos e bibliotecas (MCKINNEY, 2018, p. 22-23).

5.3.7 Pandas

O Pandas, cujo o nome é derivado do termo em inglês *panel data*, oferece estruturas de dados de alto nível e funções, de modo que trabalhar com dados estruturados ou tabulares seja rápido, fácil e expressivo. Combina as ideias de processamento de alto desempenho de *arrays* da NumPy com os recursos flexíveis de manipulação de dados de planilhas e bancos de dados relacionais. Como a manipulação, preparação e limpeza de dados são essenciais na análise de dados, o Pandas é essencial em um projeto de ciência de dados (MCKINNEY, 2018, p. 23).

5.3.8 Matplotlib

Matplotlib é a biblioteca Python mais popular para a geração de gráficos e outras visualizações de dados bidimensionais. A biblioteca foi projetada para criar plotagens apropriadas para publicação e, apesar de existirem outras bibliotecas disponíveis para este fim, a Matplotlib é a mais utilizada, tendo boa integração com o restante do ecossistema (MCKINNEY, 2018, p. 25).

5.3.9 Scikit-Learn

O Scikit-Learn é uma biblioteca da linguagem Python de código aberto desenvolvida especificamente para o propósito geral de aprendizado de máquina (MCKINNEY, 2018, p. 27). Dispõe de ferramentas simples e eficientes para análise preditiva de dados, podendo ser reutilizada em diferentes situações. O Scikit-Learn está organizado em muitos módulos, cada um desenvolvido para uma finalidade específica e, em cada módulo, existem funções para as mais diferentes aplicações, tais como pré-processamento, classificação, regressão, clusterização, redução de dimensionalidade e ajuste de parâmetros (DIDÁTICA TECH, 2020).

6. ESTUDO DE CASO

Com a finalidade de simular uma situação real, desde o armazenamento dos dados, tarefas de ciência de dados até a apresentação dos resultados, procurou-se uma base de dados que, mesmo fictícia, reproduzisse uma situação real, de modo que todo o trabalho realizado possa ser utilizado como referência para futuros trabalhos em ciência de dados.

6.1. PASSO 1: CRIAÇÃO DO *DATASET*

A partir de um arquivo no formato de planilha disponibilizado gratuitamente pela Data Science Academy, através do curso gratuito Microsoft Power BI para Data Science 2.0 (DATA SCIENCE ACADEMY, 2021), a base de dados a ser utilizada foi criada com a inclusão de atributos e a adição de linhas, de modo que ampliasse a base de dados a ser utilizada na análise. A tabela criada foi denominada de DadosVendasCarros.xlsx e a Tabela 1, apresentada abaixo, contém os metadados acerca dos atributos da tabela criada.

Atributo	Descrição
DataNotaFiscal	Data em que a venda foi realizada
Fabricante	Nome do fabricante do veículo
Estado	Estado em que foi realizada a venda
ValorVenda	Valor de venda do veículo
ValorCusto	Valor de custo do veículo
TotalDesconto	Desconto concedido na venda do veículo
CustoEntrega	Custo de entrega do veículo
CustoMaoDeObra	Custo de mão-de-obra do veículo
NomeCliente	Nome de cliente que realizou a compra
Modelo	Modelo do veículo vendido
Tipo	Tipo do modelo do veículo
Cor	Cor do veículo
Ano	Ano do modelo do veículo
Pagamento	Tipo de pagamento que foi realizado pelo cliente
TempoEstoque	Tempo em dias em que o veículo permaneceu em estoque

Tabela 1: Metadados da base de dados

6.1.1. Armazenamento dos Dados

O armazenamento dos dados foi realizado por meio da ferramenta Google BigQuery, que possui excelente capacidade de armazenamento, além de possibilitar a utilização de 10 GB gratuitamente por mês. O BigQuery está presente na Google Cloud Plataform podendo ser integrada a outros serviços, podendo ser utilizada tanto diretamente como banco de dados, como pode ser integrada a outros bancos de dados, como por exemplo, PostgreSQL. A estratégia adotada foi de criar um banco de dados a partir dos dados produzidos e exportados no formato CSV (*Comma-separeted Values)*.

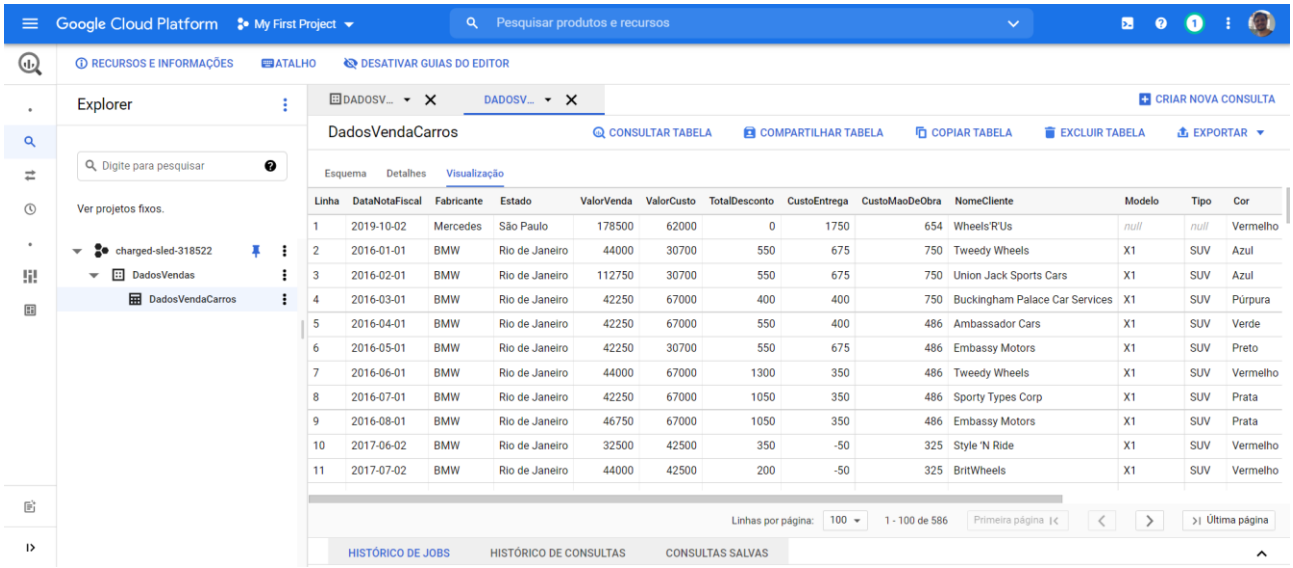

Figura 7: Banco de Dados criado no BigQuery

6.1.2. Entendimento do Problema

Uma das bases de uma boa análise de dados é o entendimento do problema que se deseja resolver, bem como formular as perguntas as quais se deseja responder ao longo da análise dos dados.

Ao se avaliar o conjunto de dados sobre as vendas de carros no período de janeiro de 2016 a dezembro de 2019, as questões que foram levantadas e que pretendem ser respondidas por meio deste estudo de caso foram as seguintes:

- Como foram as vendas no período e qual ano foi o ano que apresentou melhor desempenho?
- Qual o fabricante e modelo foram os mais vendidos?
- Quais outras tendências podem ser observadas no período?

- Qual é a análise descritiva desse *dataset*?
- Existe alguma relação entre o valor de venda e as outras variáveis numéricas? É possível prever o valor de venda de um veículo com base nessas variáveis?

6.1.3. Análise Inicial dos Dados

Para um maior entendimento dos dados e, para uma exploração inicial do conjunto de dados de uma maneira rápida e simples, optou-se pela utilização da ferramenta Data Studio, disponibilizada gratuitamente pela Google. O conjunto de dados foi exportado diretamente do BigQuery através do Data Studio e, a partir disto, realizou-se algumas análises que resultou na criação de gráficos e informações consolidadas, conforme pode-se visualizar na figura abaixo.

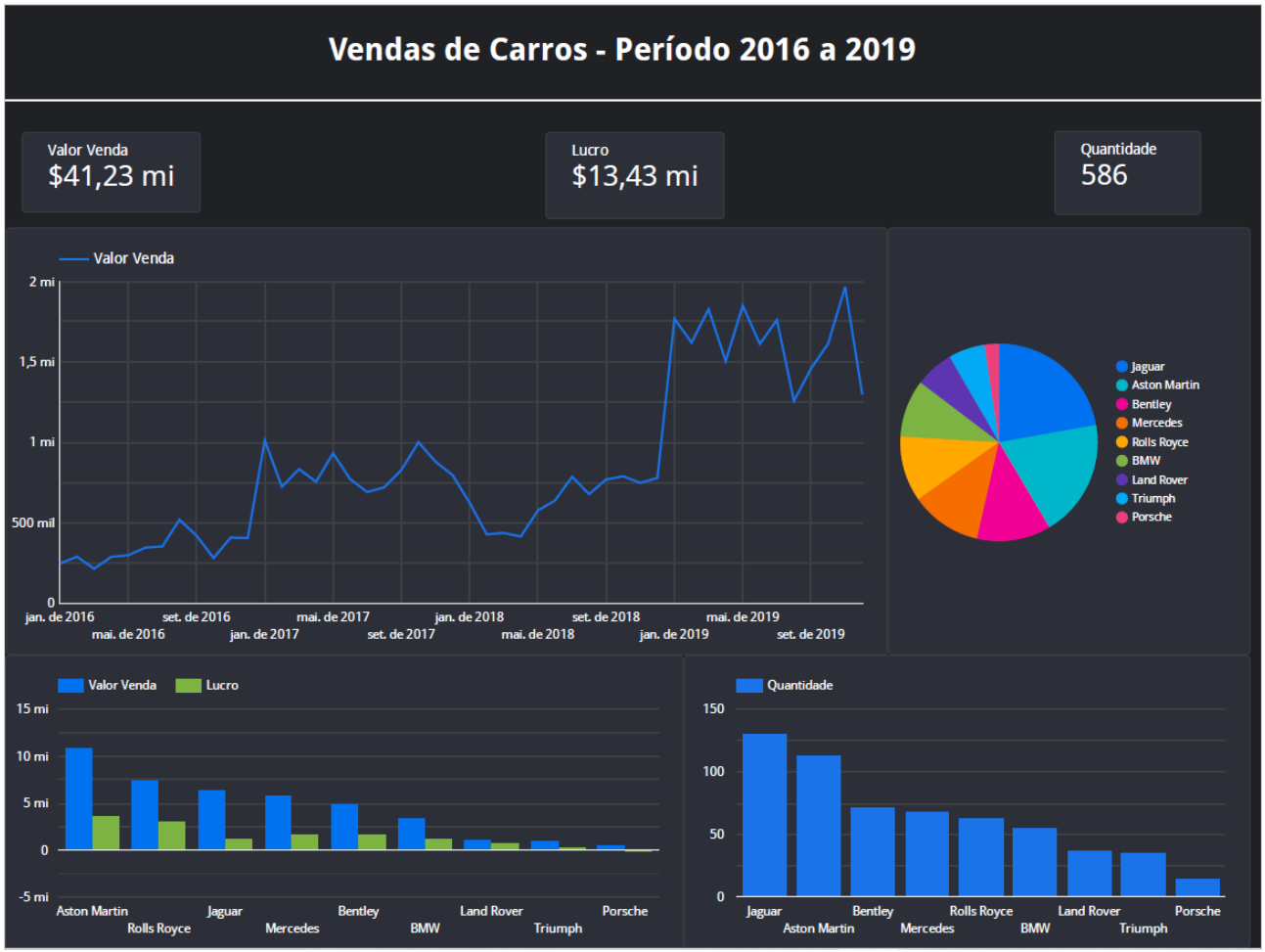

Figura 8: Relatório gerado no Data Studio

O relatório obtido possui no topo um somatório das vendas no período, bem como do lucro e a quantidade de veículos vendidos de 2016 a 2019. Contém uma série temporal das vendas diárias no período, um gráfico de pizza e um gráfico de barras sobre as vendas por fabricante, além de um gráfico comparando as vendas com o lucro por fabricante neste período.

6.1.4. Obtenção dos Dados

Os dados que serão utilizados nas análises que serão feitas Jupyter Notebook serão carregados por meio de um arquivo no formato CSV (*Comma-separeted Values)*. Apesar de já possuir o arquivo nesse formato que originou o banco de dados no BigQuery, quis-se demonstrar como seria possível obter um arquivo CSV diretamente do BigQuery caso fosse necessário. Isso é possível por meio da exportação do arquivo para o Google Cloud do usuário.

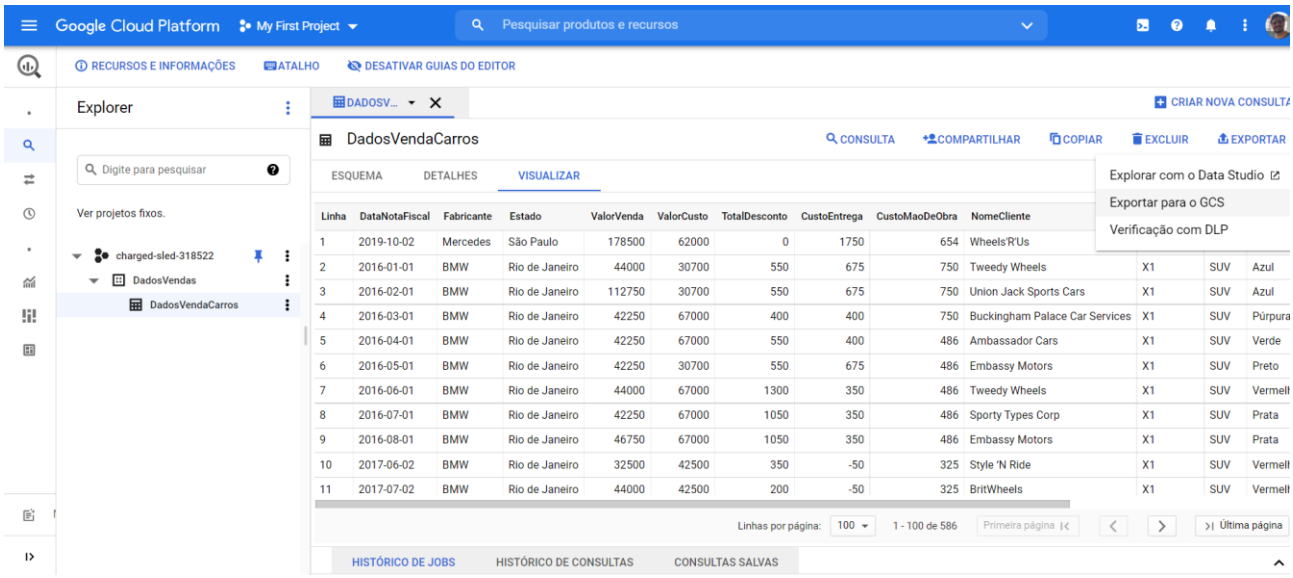

Figura 9: Exportação em CSV

Após a criação do arquivo no formato CSV, ele fica disponível na conta do Google Cloud, de onde é possível fazer o *download* para a máquina em que será realizada a análise.

6.2. PASSO 2: ANÁLISE DOS DADOS

O passo da análise dos dados será feito utilizando a Linguagem Python e suas bibliotecas já existentes por meio da ferramenta Jupyter Notebook, um ambiente no formato de

notebook, no qual as saídas dos comandos se dá logo abaixo da célula em que o código foi digitado.

A primeira ação realizada foi importar as bibliotecas, também chamadas de módulos a serem utilizadas na análise. Num primeiro momento, foram importadas as bibliotecas NumPy, Pandas e o módulo Pyplot da Matplotlib, conforme demostrado logo abaixo.

```
In [1]: #Importando os módulos necessários
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        %matplotlib inline
```

Figura 10: Importando os módulos Python

6.2.1. Carregamento dos Dados

Para dar início na análise propriamente dita é necessário carregar os dados com o auxílio do Pandas, que fornece suporte para carregar dados dos mais variados formatos e visualizá-los em formato de uma planilha. Os dados que utilizaremos foram importados do formato CSV, conforme já citado acima.

```
In [2]: #Carregamento dos dados no formato CSV
        dados = 'C:/Users/marce/OneDrive/TCC/Datasets/DadosVendaCarros.csv'
        df = pd.read_csv(dados, sep=';', parse_dates=['DataNotaFiscal'])
        df
```

Out[2]:

	DataNotaFiscal	Fabricante	Estado	ValorVenda	ValorCusto	TotalDesconto	CustoEntrega	CustoMaoDeObra	NomeCliente	Modelo	Tipo	Cor
0	2016-04-10	Rolls Royce	São Paulo	95000	50000	500	750	750	Aldo Motors	Camargue	Coupe	Vermelho
1	2016-01-01	Aston Martin	São Paulo	120000	75000	0	1500	550	Honest John	DBS	Coupe	Azul
2	2016-02-02	Rolls Royce	São Paulo	88000	75000	750	1000	550	Bright Orange	Prata Ghost	Sedan	Verde
3	2016-03-03	Rolls Royce	São Paulo	89000	88000	0	1000	550	Honest John	Prata Ghost	Sedan	Azul
4	2016-04-04	Rolls Royce	São Paulo	92000	62000	0	1500	550	Wheels'R'Us	Camargue	Coupe	Prata
...	...	...	...	...	...	...	...	...	...	...	...	...
581	2019-07-22	Mercedes	São Paulo	102500	62000	1750	500	987	Aldo Motors	GLA	Crossover	Prata
582	2019-09-23	Mercedes	São Paulo	130000	62000	0	1000	750	Honest John	GLA	Crossover	Prata
583	2019-10-24	Mercedes	São Paulo	178500	75890	0	450	654	Bright Orange	GLA	Crossover	Verde
584	2019-11-25	Mercedes	Rio de Janeiro	46750	42500	800	400	987	BritWheels	GLA	Crossover	Azul
585	2019-11-25	Mercedes	Rio de Janeiro	46750	42500	800	400	987	BritWheels	GLA	Crossover	Azul

586 rows × 16 columns

Figura 11: Visualização dos Dados Carregados

6.2.2. Limpeza e Tratamento dos Dados

Os dados nem sempre vem preparados para serem usados nas análises ou modelagem e, para isso são necessárias algumas etapas de pré-processamento antes da análise propriamente dita. Para o nosso conjunto de dados, algumas etapas foram necessárias para a preparação do *dataset*, conforme descrito abaixo.

O primeiro procedimento realizado foi alteração do tipo da coluna Ano já que ao carregar os dados utilizando o Pandas, ele identifica automaticamente os tipos de cada coluna e, a coluna Ano foi identificada como int64. Apesar de ser uma informação numérica, o atributo diz respeito ao ano de fabricação do veículo, portanto um dado literal, que no Pandas é identificado como *object*.

```
In [4]: #Conversão da Coluna Ano para string
        df['Ano'] = df['Ano'].apply(lambda x: str(x))
        df.dtypes

Out[4]: DataNotaFiscal    datetime64[ns]
        Fabricante                object
        Estado                    object
        ValorVenda                 int64
        ValorCusto                 int64
        TotalDesconto              int64
        CustoEntrega               int64
        CustoMaoDeObra             int64
        NomeCliente               object
        Modelo                    object
        Tipo                      object
        Cor                       object
        Ano                       object
        Pagamento                 object
        TempoEstoque             float64
        Unnamed: 15              float64
        dtype: object
```

Figura 12: Transformação do Tipo da Coluna Ano

O próximo passo foi identificar a existência de células sem valores preenchidos, identificadas como NaN (*Not a Number*) pelo Pandas. Ao realizar essa verificação, constatou a existência de uma coluna sem nenhum valor (Unnamed: 15), ou seja, com todos os valores NaN, e que nas colunas NomeCliente, Modelo, Tipo, Pagamento e TempoEstoque existia um valor do tipo NaN.

Com relação a coluna sem valores, a estratégia utilizada foi a exclusão da coluna, pois a mesma deve ter sido gerada de maneira equivocada, possivelmente devido ao processo de exportação do arquivo para o formato CSV.

```
In [9]: #Exclusão da coluna Unnamed: 15
        df = df.drop(columns=['Unnamed: 15'])
```

Figura 13: Remoção da Coluna Unnamed: 15

Por conta da baixa incidência de valores NaN em todo o *dataset*, 4 no total, a estratégia adotada foi o preenchimento dos valores numéricos com a média e dos valores literais com a moda da coluna.

```
In [12]: #Preenchimento dos valores NaN
         #df = df.dropna()
         df['NomeCliente'].fillna('Bright Orange', inplace=True)
         df['Pagamento'].fillna('Financiamento', inplace=True)
         df['TempoEstoque'].fillna(df['TempoEstoque'].mean(), inplace=True)
         df['Modelo'].fillna('XK', inplace=True)
         df['Tipo'].fillna('Coupe', inplace=True)
         df.count()
```

Figura 14: Preenchimento dos Valores NaN

O próximo passo foi a verificação da existência de células duplicadas, tarefa essa facilitada pela adoção da biblioteca Pandas, que possui comando próprio para a verificação e a remoção de linhas duplicadas.

```
In [13]: df[df.duplicated()]
```

Out[13]:

	DataNotaFiscal	Fabricante	Estado	ValorVenda	ValorCusto	TotalDesconto	CustoEntrega	CustoMaoDeObra	NomeCliente	Modelo	Tipo	Cor	An
443	2019-02-12	Triumph	Rio de Janeiro	28000	22000	200	-50	325	British Luxury Automobile Corp	TR4	Conversível	Verde	201
504	2017-05-18	Mercedes	Rio de Janeiro	46750	42500	50	400	486	Style 'N Ride	GLA	Crossover	Preto	201
585	2019-11-25	Mercedes	Rio de Janeiro	46750	42500	800	400	987	BritWheels	GLA	Crossover	Azul	201

```
In [14]: df = df.drop_duplicates()
         df
```

Figura 15: Remoção de Linhas Duplicadas

Um outro ponto importante na etapa de pré-processamento é a verificação de valores *outliers*, isto é, valores que são muito discrepantes do resto do conjunto de valores do atributo. Ao se realizar essa análise verificou-se que no atributo ValorVenda havia a existência de uma linha cujo valor era muito diferente dos outros valores deste atributo, fato este que foi resolvido com a substituição do valor existente pela média do atributo.

```
In [15]: #Busca por Valores Outliers
         df.describe()
```

Out[15]:

	ValorVenda	ValorCusto	TotalDesconto	CustoEntrega	CustoMaoDeObra	TempoEstoque
count	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000
mean	70517.543739	45760.658662	499.382504	550.145798	678.653516	156.713125
std	46327.363549	30770.818526	644.580004	537.638358	338.126849	127.400679
min	1.000000	124.000000	0.000000	-75.000000	147.000000	1.000000
25%	39500.000000	22500.000000	150.000000	50.000000	325.000000	56.000000
50%	45800.000000	37500.000000	200.000000	425.000000	570.000000	109.000000
75%	110000.000000	67000.000000	750.000000	925.000000	987.000000	250.500000
max	181250.000000	160000.000000	5050.000000	1750.000000	1250.000000	640.000000

```
In [16]: #Localização do Valor Outlier
         df[df['ValorVenda']==1]
```

Out[16]:

	DataNotaFiscal	Fabricante	Estado	ValorVenda	ValorCusto	TotalDesconto	CustoEntrega	CustoMaoDeObra	NomeCliente	Modelo	Tipo	Cor	Ano	
166	2018-04-11	Aston Martin	Minas Gerais	1	104000	50	1475	750	Les Arnaqueurs	DB4	Coupe	Verde	2019	F

```
In [17]: #Substituição do Valor Outlier
         df['ValorVenda'] = df['ValorVenda'].replace([1], df['ValorVenda'].mean())
```

Figura 16: Substituição de Valores *Outliers*

Para auxiliar na análise dos dados foi criada uma coluna, denominada Lucro, que se tratava do valor de venda subtraídas dos valores de custos (custo, desconto, entrega e mão-de-obra). Uma coluna com o percentual de lucro, denominada PerLucro também foi criada com esse mesmo intuito.

```
In [20]: #Criação da Coluna Lucro
         df['Lucro'] = df['ValorVenda'] - df['ValorCusto'] - df['TotalDesconto'] - df['CustoEntrega'] - df['CustoMaoDeObra']
```

Figura 17: Criação da Coluna Lucro

```
In [22]: #Criação da Coluna PerLucro
         df['PerLucro'] = np.round(df['Lucro']/df['ValorVenda'], 6)
         df.head()
```

Out[22]:

	ValorCusto	TotalDesconto	CustoEntrega	CustoMaoDeObra	NomeCliente	Modelo	Tipo	Cor	Ano	Pagamento	TempoEstoque	Lucro	PerLucro
)	50000	500	750	750	Aldo Motors	Camargue	Coupe	Vermelho	2016	Financiamento	125.0	43000.0	0.452632
)	75000	0	1500	550	Honest John	DBS	Coupe	Azul	2016	A Vista	185.0	42950.0	0.357917
)	75000	750	1000	550	Bright Orange	Prata Ghost	Sedan	Verde	2016	Financiamento	32.0	10700.0	0.121591
)	88000	0	1000	550	Honest John	Prata Ghost	Sedan	Azul	2016	Financiamento	62.0	-550.0	-0.006180
)	62000	0	1500	550	Wheels'R'Us	Camargue	Coupe	Prata	2016	Financiamento	94.0	27950.0	0.303804

Figura 18: Criação da Coluna PerLucro

6.2.3. Análise Exploratória

A análise exploratória é a etapa em que os dados são explorados por meio de plotagem de gráficos ou agrupamentos de dados, de modo a procurar por padrões ou informações que, numa primeira análise não está visível no *dataset*.

A primeira análise realizada tinha por objetivo identificar qual fabricante teve melhor desempenho de vendas no período. Para isso criou-se um *dataset* auxiliar agrupando os dados por fabricante e somando as vendas realizadas por fabricante. Além da criação da visualização em formato de *dataframe*, foi plotado um gráfico de barras para melhor compreensão.

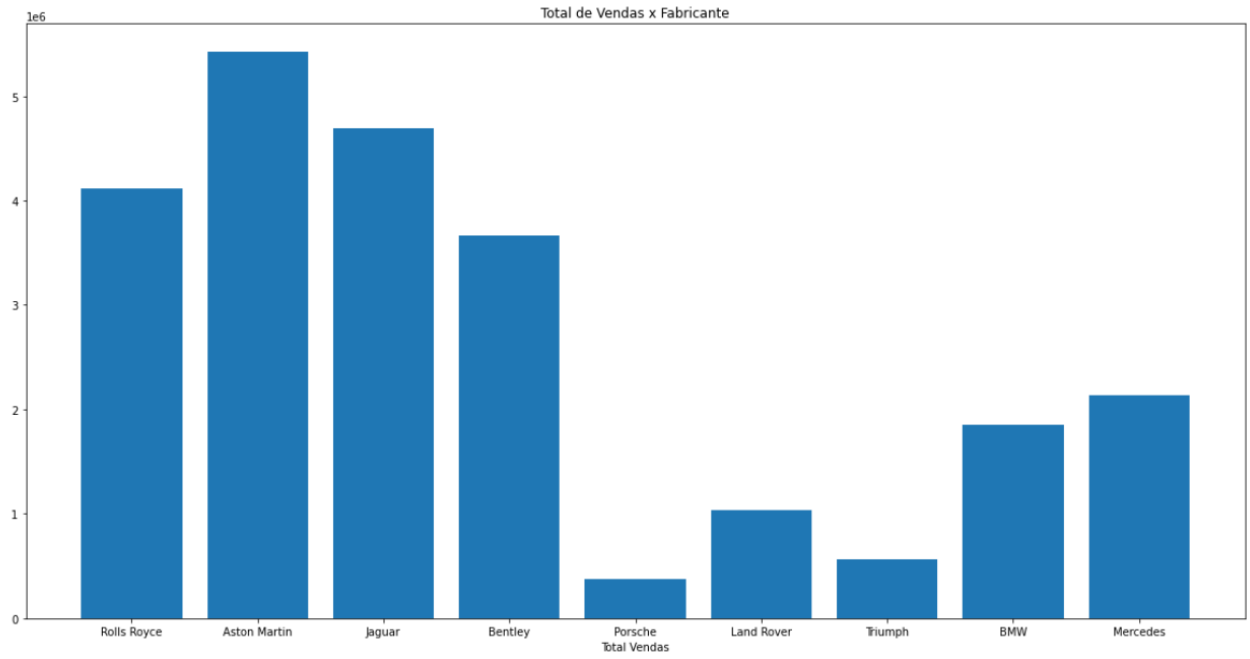

Figura 19: Gráfico Vendas x Fabricantes

Da mesma forma que foi realizado no gráfico acima, as vendas foram agrupadas por ano, também com o auxílio de um *dataframe* auxiliar e o resultado foi plotado no gráfico abaixo.

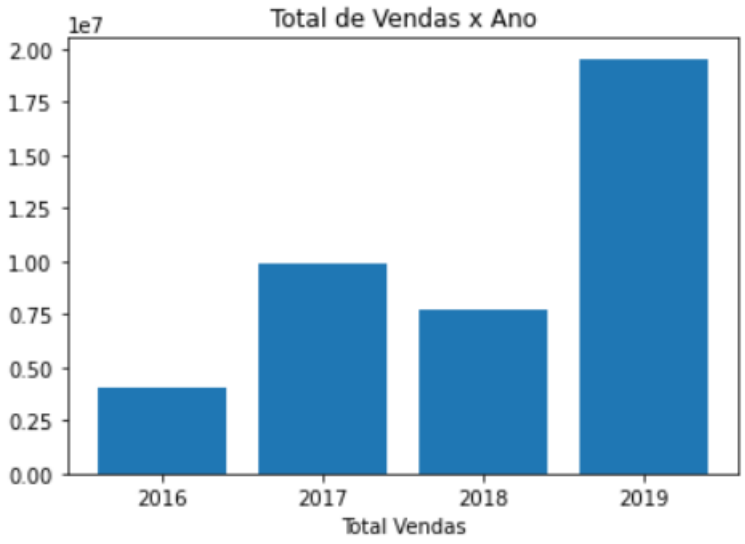

Figura 20: Gráfico Vendas x Anos

Para dar início à investigação da existência de correlação da variável alvo de valor de venda com as outras variáveis numéricas, foram criados gráficos relacionando o valor de venda com cada uma das variáveis numéricas do *dataset*, a saber: valor de custo, total de desconto, custo de entrega, custo de mão-de-obra e tempo de estoque.

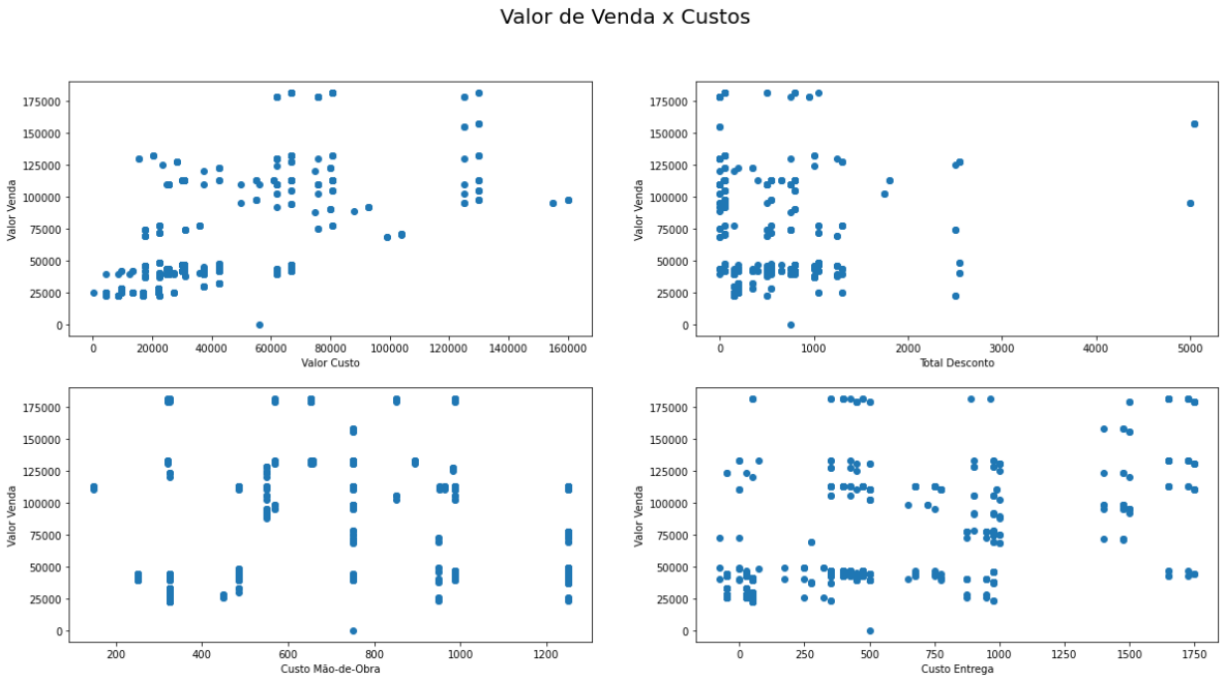

Figura 21: Gráfico Valor Venda x Custos

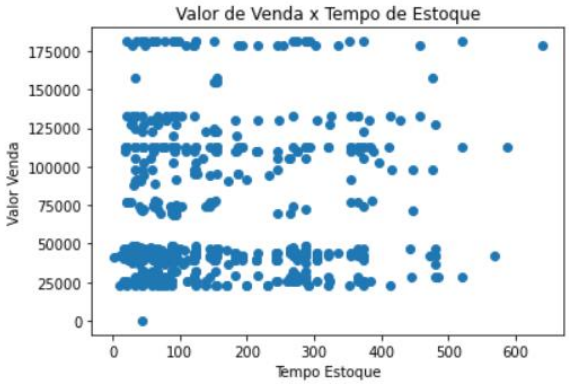

Figura 22: Gráfico Valor Venda x Tempo Estoque

Por fim, foi criado um gráfico de barras verificando a quantidade de vendas por tipo de pagamento, conforme a figura abaixo.

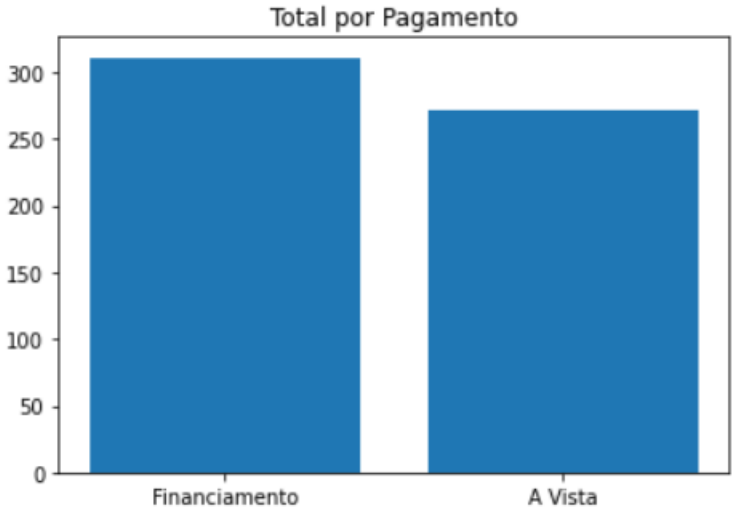

Figura 23: Gráfico Quantidade Vendas x Pagamento

6.3. PASSO 3: MINERAÇÃO DE DADOS

O passo de mineração de dados é onde, a partir dos dados pré-processados, será realizada a análise preditiva. Uma variável alvo, também chamada de variável dependente será prevista por meio de outras variáveis, chamadas de variáveis independentes ou predecessoras.

Conforme descrito na proposta de trabalho, será definido qual a tarefa de mineração será utilizada, bem como qual o algoritmo a ser aplicado. Após a fase de definições, o modelo

será treinado utilizando uma parte do conjunto de dados, denominado conjunto de treinamento e após, será testado utilizando uma outra parte do *dataset*, o conjunto de teste. Para finalizar o modelo gerado será avaliado, verificando se as previsões realizadas para o conjunto de teste são condizentes com os valores existentes e qual a taxa de erro apresentada.

6.3.1. Escolha do Modelo

O problema a ser resolvido com aprendizado de máquina é a tentativa de prever o valor de venda (ValorVenda) de um veículo a partir das outras variáveis numéricas: custo do veículo (ValorCusto), desconto (TotalDesconto), custo de entrega (CustoEntrega), custo de mão de obra (CustoMaoDeObra) e tempo em dias que o veículo ficou em estoque (TempoEstoque).

A tarefa de mineração de dados a ser utilizada é a estimativa ou regressão e, o algoritmo escolhido para ser utilizado é o de regressão linear. Para isto, será utilizada a biblioteca Scikit-Learn, que contém uma série de algoritmos de *machine learning* prontos para a utilização na linguagem Python.

6.3.2. Treinamento do Modelo

O primeiro passo para a construção do modelo é carregar os módulos necessários dentro do *notebook* onde está a análise realizada até o momento.

```
In [33]: #Importar módulo Scikit-Learn
         import sklearn
         from sklearn.linear_model import LinearRegression
         from sklearn.model_selection import train_test_split
```

Figura 24: Carregamento módulos de Aprendizado de Máquina

Em seguida são criados os conjuntos de dados contento as variáveis dependentes (X) e a variável independe (Y).

```
In [34]: #Criação do conjunto de variáveis independentes
         X = df[['ValorCusto', 'TotalDesconto', 'CustoEntrega', 'CustoMaoDeObra', 'TempoEstoque']]

In [35]: #Criação do conjunto de variável dependente
         Y = df[['ValorVenda']]
```

Figura 25: Conjuntos X e Y

Após sua criação, ambos os conjuntos são divididos em conjunto de treino (X_treino e Y_treino) e conjunto de teste (X_teste e Y_teste).

```
In [42]: #Divisão do Conjunto em Treino e Teste
         X_treino, X_teste, Y_treino, Y_teste = train_test_split(X, Y, test_size = 0.2, random_state = 5)

In [43]: print(X_treino.shape, X_teste.shape, Y_treino.shape, Y_teste.shape)
         (466, 5) (117, 5) (466, 1) (117, 1)
```

Figura 26: Conjunto de Treino e Teste

O modelo é criado, instanciando um objeto (regr) da função de regressão linear (LinearRegression), previamente carregada juntamente com o Scikit-Learn.

```
In [43]: #Instanciação do Modelo
         regr = LinearRegression()
```

Figura 27: Criação do Modelo

Por fim, o modelo é treinado, conforme mostra a imagem abaixo.

```
In [44]: #Treinamento do Modelo
         regr.fit(X_treino, Y_treino)
Out[44]: LinearRegression()
```

Figura 28: Treinamento do Modelo

O treinamento do modelo, apesar de aparentar ser somente uma linha de comando, efetua a modelagem de uma função linear na qual, a partir das variáveis preditoras (X_treino) obtém-se como resultado a variável alvo (Y_treino). Conforme são realizadas as iterações durante o treinamento, a função é ajustada de modo a otimizar os resultados obtidos com a aplicação do modelo.

6.3.3. Avaliação do Modelo

Para avaliar o modelo gerado, a princípio foi utilizado a plotagem de dois gráficos para visualizar os resultados obtidos junto ao conjunto teste. O primeiro gráfico compara os valores previstos com os valores esperados para a variável dependente, valor de venda. Uma linha foi inserida representando o resultado esperado, isto é, se o valor predito fosse igual ao valor real do conjunto de teste.

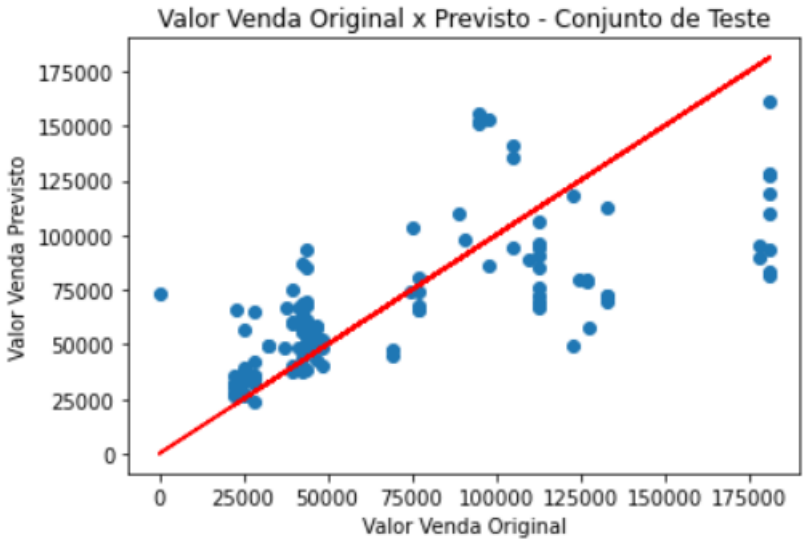

Figura 29: Gráfico Valor Original x Valor Previsto

O segundo gráfico diz respeito ao *Residual Plot*, isto é, ele compara a distribuição de cada valor previsto com a diferença entre o valor previsto e o valor esperado.

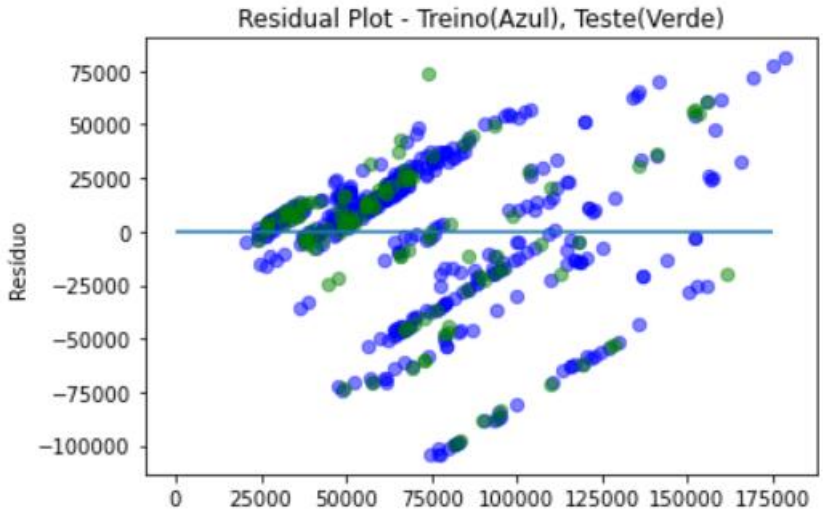

Figura 30: Gráfico Residual

Para avaliar o erro geral foi utilizado duas métricas simples, o erro médio quadrático (MSE) e o erro médio absoluto (MAE). O erro médio quadrático é a média do quadrado dos erros apresentados, sendo o erro a diferença entre o valor previsto pelo modelo e o valor real constante no conjunto de teste. O erro médio absoluto é somente a média entre os erros apresentados pelos valores previstos.

```
In [49]: # Cálculo do MSE (Mean Squared Error)
         mse = np.mean((Y_teste - regr.predict(X_teste)) ** 2)
         print(mse)

         ValorVenda    1.211515e+09
         dtype: float64

In [50]: # Cálculo do MAE (Mean Absolute Error)
         mae = np.mean(Y_teste - regr.predict(X_teste))
         print(mae)

         ValorVenda    3241.525009
         dtype: float64
```

Figura 31: MSE e MAE

Conforme a Figura 31, o MAE foi de aproximadamente 3241,53 e o MSE de 1,21 x 10^9.

6.4. PASSO 4: APRESENTAÇÃO DOS RESULTADOS

Tomando como ponto de partida as perguntas levantadas na etapa de entendimento do problema, temos a apresentação dos resultados obtidos.

Na análise inicial realizada com a ferramenta Data Studio, voltada para a inteligência de negócios, foi possível obter os seguintes *insights*:

- O valor total de vendas no período foi de US$ 41,23 milhões, com o lucro de US$ 13,43 milhões e a venda de 586 unidades.
- As vendas cresceram ao passar do tempo, sendo menores em 2016 e maiores em 2019;
- O fabricante Aston Martin é o que apresentou maior valor de venda acumulado no período, porém o fabricante Jaguar teve maior quantidade de unidades vendidas.

Após essa investigação inicial, já com o auxílio do Jupyter Notebook e após o pré-processamento dos dados, foi realizada uma segunda análise exploratória que, além de ratificar as informações encontradas anteriormente na análise inicial, pode-se levantar novas informações:

- O valor de venda é diretamente proporcional aos custos do veículo, de mão-de-obra, de entrega e desconto, pois conforme estes valores aumentam, o valor de venda também aumenta, apresentando correlação com as mesmas;
- Com relação à variável tempo de estoque, o valor de venda também é proporcional, porém essa dependência não é tão linear quanto a das outras variáveis.

- O tipo de pagamento (À vista ou financiamento) está bem distribuído no conjunto de dados, sendo que 311 linhas são do tipo financiamento e 272 do tipo à vista;
- O atributo lucro pode ser calculado por meio de operações aritméticas entre as outras variáveis, não sendo relevante sua utilização no modelo de regressão linear.

Com relação ao modelo de regressão linear criado para a previsão do valor de venda, após avaliar os resultados obtidos com o conjunto de teste, temos que o erro médio absoluto (MAE) obtido não está tão elevado, porém ao observar o erro médio quadrático (MSE), pode-se concluir que em alguns casos, o modelo vai apresentar valores muito longe dos valores existentes portanto, este modelo pode e deve ser melhorado. Para a otimização do modelo existem algumas possibilidades a serem consideradas como, por exemplo: a não utilização de todas as variáveis numéricas na criação do modelo, a normalização de valores, transformação de atributos categóricos em numéricos ou a utilização de um outro algoritmo de regressão linear.

Outras tarefas de mineração de dados podem ser realizadas nesse conjunto de dados, podendo citar como exemplos: a classificação do tipo de pagamento, tentando predizer se determinado veículo vai ser vendido à vista ou financiado ou a realização do agrupamento do conjunto de dados em grupos de veículos semelhantes entre si, o que pode ser útil em algumas situações de tomadas de decisão.

7. CONCLUSÃO

Este trabalho teve por objetivo uma compreensão maior acerca de temas atuais e relevantes, tais como Big Data, Ciência de Dados e Mineração de Dados e, principalmente demonstrar como estes temas são relacionados e interdependentes.

A Ciência de Dados, definida como um conjunto básico de princípios, conceitos e técnicas que estruturam o pensamento e a análise de dados, se utiliza dos dados como matéria-prima essencial para a extração de conhecimento e, com o exponencial crescimento do volume de dados decorrente do Big Data, ocupa um lugar de destaque no mercado atual. Esta ciência abrange todo o ciclo de vida do dado, desde a sua produção até o seu fim, contemplando fases de coleta, armazenamento, transformação, análise e descarte.

A mineração de dados é a responsável pela exploração de grandes conjuntos de dados a fim de encontrar padrões relevantes, de modo a ser possível realizar tarefas descritivas ou preditivas, sendo possivelmente estes o grande potencial existente na mineração de dados, já que se trata de um grande auxilio no processo de tomada de decisão gerencial.

Por fim o estudo de caso realizado teve o intuito de simular uma situação real, acompanhando diferentes etapas dentro da Ciência de Dados: o armazenamento em nuvem utilizando o Google BigQuery; uma análise inicial, voltada para a inteligência de negócios com o Data Studio e por fim, utilizando a linguagem Python e suas bibliotecas no ambiente do Jupyter Notebook, foi possível a realização da análise dos dados, realizando as etapas de limpeza e tratamento dos dados, análise descritiva do *dataset* e a aplicação de um algoritmo de mineração de dados.

Com relação às ferramentas utilizadas, o BigQuery foi pouco explorado, tendo sido utilizado somente como solução de armazenamento, conforme o escopo do trabalho, porém o mesmo possui várias ferramentas que não foram exploradas. O Data Studio por se tratar de uma ferramenta gratuita, além de ser intuitiva em seu uso, conta com várias possibilidades de plotagem gráficas, o que a torna uma ferramenta interessante para a utilização, tanto de profissionais quanto de usuários menos familiarizados com tecnologia. E quanto a linguagem Python, foi ratificada a importância que ela possui para o cenário de Ciência de Dados, pois além de ser uma linguagem de fácil aprendizado, possui inúmeras bibliotecas e soluções prontas, o que torna o processo de análise de dados mais prático e eficiente.

Pode-se finalmente concluir que, após a realização deste trabalho, um grande volume de dados por si só não possui grande valor agregado, devendo ser aplicadas as técnicas e ferramentas adequadas para que se possa extrair o conhecimento oculto existente nele. E um maior conhecimento sobre estes conceitos, técnicas e ferramentas adequadas, é preponderante para a extração eficiente e otimizada deste conhecimento.

7.1. TRABALHOS FUTUROS

Como já foi observado neste trabalho, o modelo gerado pode ser melhorado, então pode-se realizar um trabalho focado na otimização deste modelo, ou a aplicação de outras técnicas ou algoritmos de regressão nesta mesma base de dados, com um estudo comparativo entre os modelos criados. Outras tarefas de mineração de dados também podem ser aplicadas, tais como agrupamento ou classificação, ampliando assim as possibilidades de trabalhos futuros. Um estudo com a utilização de redes neurais, também seria um tema promissor a se explorar no futuro.

A etapa de pré-processamento dos dados que abrange a limpeza e tratamento dos dados e, em um projeto de ciência de dados ocupa a maior parcela do tempo total, poderia ser abordado em um trabalho futuro, detalhando as técnicas possíveis de serem utilizadas nesta etapa.

Pode-se considerar também que a ferramenta Google BigQuery foi pouco explorada e um trabalho focado na coleta e armazenamento de dados, podendo ou não ser obtidos de diversas fontes de dados e a exploração de suas funcionalidades seria interessante de se realizar.

REFERÊNCIAS

ALVES, William Pereira. **Banco de Dados: teoria e desenvolvimento**. 2ª edição. São Paulo: Erica, 2021.

AMARAL, Fernando. **Introdução à Ciência de Dados, Mineração de Dados e Big Data**. 1ª edição. Rio de Janeiro: Alta Books, 2016.

AMO, Sandra de. **Técnicas de Mineração de Dados**. Jornada de Atualização em Informática. jul. 2004. Disponível em < https://sistemas2012.webnode.com.br/_files/200000095-bf367bfb43/Tecnicas%20de%20Minera%C3%A7%C3%A3o%20de%20Dados.pdf > Acessado em 04.mai.2020.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas.** Relatório Técnico. Instituto de Informática – Universidade Federal de Goiás. ago. 2009. Disponível em < http://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf > Acesso em 04 mai. 2020

CAMPOS, Othon Stuart Ferreira. **Data Analytics Transparente para Descoberta de Padrões e Anomalias na Realização de Convênios e Contratos de Repasse Federais**. 2018. 98p. Dissertação (Mestrado) – Centro de Ciências Exatas e Tecnologia – Universidade Federal de Sergipe, Sergipe, São Cristovão, 2018. Disponível em: < https://ri.ufs.br/bitstream/riufs/10766/2/OTHON_STUART_FERREIRA_CAMPOS.pdf > Acessado em 27.fev.2021

CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. **Introdução à Mineração de Dados: conceitos básicos, algoritmos e aplicações**. 1ª edição. São Paulo: Saraiva, 2016.

CONMAY, Drew. **The Data Science Venn Diagram.** 2010. Disponível em < http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram >. Acessado em 16.mar.2021

CÔRTES, Sergio da Costa; PORCARO, Rosa Maria; LIFSCHITZ, Sérgio. **Mineração de Dados – Funcionalidades, Técnicas e Abordagens.** mai. 2002. Departamento de Informática – PUC RIO. Disponível em < *obaluae.inf.puc-rio.br* › docs › techreports › 02_10_cortes > Acesso em 30 abr. 2020.

COUTINHO, Thiago. **Como o Anaconda IDE pode ajudar na sua programação em Python.** Voitto, 2020. Disponível em < https://www.voitto.com.br/blog/artigo/anaconda-python-ide >. Acessado em 22.jul.2021

CRUZ, Felipe. **Python: Escreva seus primeiros programas.** São Paulo: Casa do Código, 2020.

CURTY, Renata G.; SERAFIM, Jucenir da S. A formação em ciência de dados: uma análise preliminar do panorama estadunidense. **Informação & Informação**. Londrina, v. 21, n. 2, p. 307-331, dez. 2016. Disponível em: < http://www.uel.br/revistas/uel/index.php/informacao/article/view/27945 >. Acessado em: 27.fev.2021

DATA SCIENCE ACADEMY. **Microsoft Power BI para Data Science 2.0.** Data Science Academy, 2021. Disponível em < https://www.datascienceacademy.com.br/path-player?courseid=microsoft-power-bi-para-data-science&unit=5f2ef93fe32fc32103622576Unit > Acessado 10.jun.2021

DIAS, Maria Madalena. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. **Acta Scientiarum. Technology**. Maringá, v.24, n.6, 2002. p. 1715-1725.

DIDÁTICA TECH. **A biblioteca scikit-learn – Python para machine learning.** Didática Tech, 2020. Disponível em < https://didatica.tech/a-biblioteca-scikit-learn-pyhton-para-machine-learning/ >. Acessado em 22.jul.2021

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de Bancos de Dados**. 6ª edição. São Paulo: Pearson Addison Wesley, 2011.

ESCOVEDO, Tatiana; KOSHIYAMA, Adriano. **Introdução a Data Science.** São Paulo: Casa do Código, 2020.

FAGUNDES, Priscila B.; MACEDO, Douglas D. J.; FREUND, Gislaine P. A Produção Científica sobre Qualidade de Dados em Big Data: um estudo na base de dados Web of Science. **Revista Digital de Biblioteconomia e Ciência da Informação**. Campinas, v. 16, n. 1, p. 194-210, jan-abr, 2018. Disponível em: < https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8650412 > Acessado em 20.nov.2018

FERREIRA, João. MIRANDA, Miguel. ABELHA, Antônio. MACHADO, José. O Processo ETL em Sistemas Data Warehouse. In: INForum – II Simpósio de Informática, 2, 2010, Braga, Portugal. **Actas do II Simpósio de Informática**, v. 1, setembro, 2010, p. 757-765.

Disponível em < http://inforum.org.pt/INForum2010/papers/sistemas-inteligentes/Paper080.pdf > Acessado em 23.fev.2021

GOLDSCHMIDT, Ronaldo; BEZERRA, Eduardo; PASSOS, Emmanuel. **Data mining: conceitos, técnicas, algoritmos, orientações e aplicações.** 2ª edição. Rio de Janeiro: Elsevier, 2015.

GOOGLE. **BigQuery.** Google Cloud, 2021. Disponível em < https://cloud.google.com/bigquery?utm_source=google&utm_medium=cpc&utm_campaign=latam-BR-all-pt-dr-SKWS-all-all-trial-p-dr-1009897-LUAC0015653&utm_content=text-ad-none-any-DEV_c-CRE_532282943219-ADGP_Hybrid%20%7C%20SKWS%20-%20PHR%20%7C%20Txt%20~%20Data-Analytics_BigQuery-KWID_43700064907853690-kwd-332753161828&utm_term=KW_bigquery-ST_BigQuery&gclid=CjwKCAjwruSHBhAtEiwA_qCppiZT2HExnbgfSmDCYQlWbdHW88P_j69DoSN6CicMNGXDV8EliJzpkhoCvU0QAvD_BwE&gclsrc=aw.ds >. Acessado em 22.jul.2021

GOOGLE. **Conheça o Data Studio.** Google, 2021. Disponível em < https://support.google.com/datastudio/answer/6283323?hl=pt-BR >. Acessado em 22.jul.2021

GOOGLE. **O que é BigQuery?** Google Cloud, 2021. Disponível em < https://cloud.google.com/bigquery/docs/introduction?hl=pt-br >. Acessado em 22.jul.2021

MARQUESONE, Rosangela. **Big Data: Técnicas e tecnologias para extração de valor dos dados.** São Paulo: Casa do Código, 2017.

MCKINNEY, Wes. **Python para Análise de Dados.** São Paulo: Novatec, 2018.

ORACLE. **O que é um Data Warehouse?** Disponível em < https://www.oracle.com/br/database/what-is-a-data-warehouse/ >. Acessado em 23.fev.2021

PAIXÃO, Alexandre O.; SILVA, Verônica A.; TANAKA, Asterio. De Business Intelligence a Data Science: um estudo comparativo entre áreas de conhecimento relacionadas. In: Congresso Integrado de Tecnologia da Informação, VIII, 2015, Campos dos Goytacazes, **Congresso Integrado de Tecnologia da Informação**, Campos dos Goytacazes: Essentia Editora. Disponível em < http://www.essentiaeditora.iff.edu.br/index.php/citi/article/view/6347 > Acessado em 18.out.2020

PASSOS, Danielle Sandler dos. Big Data, Data Science e seus contributos para o avanço no uso da Open Source Intelligence. **Sistemas & Gestão.** Londrina, v. 11, n. 4, dezembro, 2016. p. 392-396. Disponível em: < https://revistasg.uff.br/sg/article/view/1026/524 >. Acesso em 28.fev.2021.

PROVOST, Foster; FAWCETT, Tom. **Data Science para negócios**. 1ª edição. Rio de Janeiro: Alta Books, 2016.

REZENDE, Eliana. **Dados, Informação e Conhecimento. O que são?** ER Consultoria. Disponível em < https://eliana-rezende.com.br/dados-informacao-e-conhecimento-o-que-sao/#:~:text=Dados%20s%C3%A3o%20c%C3%B3digos%20que%20constituem,(de%20Silva%2C%202007) >. Acesso em 21.fev.2021

SANTANA, Felipe. **Guia passo a passo de como um projeto de Data Science é desenvolvido.** Minerando Dados, 2019. Disponível em < https://minerandodados.com.br/guia-passo-a-passo-de-como-um-projeto-de-data-science-e-desenvolvido/ >. Acessado em 28.fev.2021

SILVA, Leandro Augusto da; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à Mineração de Dados: com aplicações em R.** 1ª edição. Rio de Janeiro: Elsevier, 2016.

SILVA, Michel de Almeida. **O Pré-Processamento em Mineração de Dados como método de suporte à modelagem algorítmica.** 2014. 83p. Dissertação (Mestrado) – Universidade Federal do Tocantins, Palmas, 2014.

SIRQUEIRA, Tassio. DALPRA, Humberto. NoSQL e a Importância da Engenharia de Dados para o Big Data. In: 37ª JAI – Jornadas de Atualização em Informática, 2018. **Anais da 37ª JAI – Jornadas de Atualização em Informática**, julho, 2018. p. 58-98.

TAURION, Cezar. **Big Data**. 1ª edição. Rio de Janeiro: Brasport, 2015.