UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
COORDENAÇÃO DE ENGENHARIA DE PRODUÇÃO
CURSO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

JOEL DOS SANTOS BOQUIMPANI JUNIOR

APLICAÇÃO DA MINERAÇÃO DE DADOS NA EXTRAÇÃO DE INFORMAÇÃO SOBRE OS FATORES INFLUENCIADORES DA SATISFAÇÃO DO CLIENTE

TRABALHO DE CONCLUSÃO DE CURSO

MEDIANEIRA
2020

JOEL DOS SANTOS BOQUIMPANI JUNIOR

APLICAÇÃO DA MINERAÇÃO DE DADOS NA EXTRAÇÃO DE INFORMAÇÃO SOBRE OS FATORES INFLUENCIADORES DA SATISFAÇÃO DO CLIENTE

Trabalho de conclusão de curso apresentado ao Curso de Graduação, em Engenharia de Produção, da Universidade Tecnológica Federal do Paraná, como requisito parcial à disciplina de TCC2.

Orientador: Prof. MSc. Ricardo Sobjak
Coorientador: Prof. Dr. Sérgio Adelar Brun

MEDIANEIRA
2020

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
Câmpus Medianeira
DIRETORIA DE GRADUAÇÃO E EDUCAÇÃO PROFISSIONAL
Departamento Acadêmico de Produção e Administração
Curso de Graduação em Engenharia de Produção

TERMO DE APROVAÇÃO

APLICAÇÃO DA MINERAÇÃO DE DADOS NA EXTRAÇÃO DE INFORMAÇÃO SOBRE OS FATORES INFLUENCIADORES DA SATISFAÇÃO DO CLIENTE

Por

JOEL DOS SANTOS BOQUIMPANI JUNIOR

Este trabalho de conclusão de curso foi apresentado às 15h do dia 10 de dezembro de 2020 como requisito parcial para aprovação na disciplina de TCC2, da Universidade Tecnológica Federal do Paraná, *Câmpus* Medianeira. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o projeto para realização de trabalho de diplomação aprovado.

_____
Prof. MSc. Ricardo Sobjak
Universidade Tecnológica Federal do Paraná
(Orientador)

_____
Prof. Dr. Sérgio Adelar Brun
Universidade Tecnológica Federal do Paraná
(Coorientador)

_____
Prof. MSc. Neron A. C. Berghauser
Universidade Tecnológica Federal do Paraná
(Convidado)

_____
Prof. MSc. Lidiana Zocche
Universidade Tecnológica Federal do Paraná
(Convidada)

Aos meus pais e aos meus amigos,

Companheiros de todas as horas.

AGRADECIMENTOS

Primeiramente aos meus familiares, pelo amor, incentivo e apoio incondicional. Agradeço a minha mãe, Mônica, que me escutou e incentivou nas horas difíceis, de desanimo e cansaço, um exemplo de perseverança. Ao meu pai, Joel, amigo que sempre buscou me fornecer o que podia para eu estar aqui hoje. Meu irmão, Gabriel que sempre me alegrava quando conversava comigo.

Minha Avó, que sempre me incentivou a estudar, tentando mostrar o caminho com maiores recompensas, uma sábia com as suas perspectivas, e até hoje muito higiênica.

A minha companheira e amiga, Clarissa, impossível descrever a sua ajuda, esteve comigo em todos os momentos desse trabalho, me mostrava uma força extraordinária, meu exemplo de comprometimento, carinho e determinação.

Aos meus amigos, Bruno, Pedro, João, Kuki, Leão, Macolin, Flávio, Tarta, Smargiassi, Baranda, Cabeça, Patrick, Murilo, Lucaas, Elis, Marcos

A Universidade Tecnológica Federal do Paraná, pelos recursos oportunidade de fazer o curso.

Ao professor orientador Ricardo Sobjak, acreditou na proposta desde início, e forneceu o conhecimento que me faltou

A todos que direta ou indiretamente fizeram parte da minha formação.

Muito obrigado a todos.

“Estamos nos afogando em informações
mas com sede de conhecimento.”

Naisbitt, John

RESUMO

JUNIOR, Joel dos Santos Boquimpani. **Aplicação da mineração de dados na extração de informação sobre os fatores influenciadores da satisfação**. 2019. Monografia (Bacharel em Engenharia de Produção) - Universidade Tecnológica Federal do Paraná.

O crescente número de usuários da internet fez com que fossem criados canais de comercialização. A comercialização tradicional evoluiu para o ambiente eletrônico e assim, com o seu desenvolvimento, muitas empresas precisaram elaborar novas estratégias para conquistar clientes e fidelizar os já existentes e a satisfação dos usuários do *e-commerce* passou a ser fator determinantes de sucesso para as organizações do setor. As atitudes dos clientes mudaram, estão mais conectados e com maior poder de decisão de compra, por isso avaliam a qualidade do serviço por diferentes fatores antes de efetuar a transação. O presente trabalho utiliza o processo de Descoberta de Conhecimento em Base de Dados, com o objetivo de analisar os fatores com maior influência na satisfação dos consumidores. As avaliações feitas pelos consumidores de um *e-commerce* serão associadas a características da compra identificando a relevância de cada uma. Na sua execução a pesquisa aborda, de maneira detalhada, a realização das etapas da Mineração de Dados. Foi verificado que os fatores que mais influenciaram na avaliação por parte dos clientes foram relacionados a entrega e econômicos. A análise dos dados obtidos busca evidenciar as características necessárias para satisfazer os consumidores virtuais, com o intuito de revelar pontos com necessidade de melhorias e assim diferenciar a empresa.

Palavras-chave: Mineração de Dados; Satisfação; *E-commerce*.

ABSTRACT

JUNIOR, Joel dos Santos Boquimpani. **Application of data mining in the extraction of information on factors influencing satisfaction**. 2019. Monografia (Bacharel em Engenharia de Produção) - Universidade Tecnológica Federal do Paraná.

The growing number of internet users has created new marketing channels. Traditional distribution has evolved into the electronic environment so, with its development, many companies have had to develop new strategies to win customers and retain existing ones and the satisfaction of e-commerce users has become a determining factor of success for industry associations . Customer attitudes have changed, they are more connected and have greater purchasing decision power, which is why the quality of service is assessed by different factors before making a transaction. The present work uses the Knowledge Discovery process in the Database, in order to analyze the factors with the greatest influence on consumer satisfaction. The evaluations by consumers of an electronic commerce linked to the characteristics of the purchase identifying each one of them. In its execution, a research addresses, in a detailed way, the completion of the steps of Data Mining. It was found that the factors that most led to the assessment by customers were related to delivery and savings. Data analysis seeks to highlight the requirements necessary to satisfy virtual consumers, in order to reveal points with the need for improvements and thus differentiate the company.

Keywords: Data Mining; Satisfaction; E-commerce.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

ARFF	Atribute-Relation File Format
BD	Banco de Dados
CE	Comércio Eletrônico
CSV	Comma Separated Values
DM	Data Mining
EUA	Estados Unidos da América
KDD	Knowledge Discovery in Database
MSE	Erro Médio Quadrático
TIC	Tecnologias de Informação e Comunicação

SUMÁRIO

1 INTRODUÇÃO

Com base na pesquisa sobre o uso das Tecnologias de Informação e Comunicação no Brasil (TIC Empresas 2017), que analisou empresas de diversos setores e portes, foi possível observar que 97% das empresas com 10 a 49 funcionários utilizam o recurso da Internet. O valor chega a 100% das empresas com mais de 250 funcionários.

A era digital fez surgir novos canais de transações comerciais, transformando o conceito de atividade econômica. Refere-se as operações realizadas no âmbito eletrônico a evolução do varejo tradicional. O comércio *on-line* está evoluindo e muitas empresas estão investindo nesta nova modalidade de varejo, proporcionando economia financeira e comodidade aos consumidores que utilizam a internet para pesquisar e adquirir bens e serviços (MANSANO; GORNI, 2014).

Segundo Albertin (2010) o comércio eletrônico é um comércio tradicional que ocorre num ambiente eletrônico, com tecnologia de comunicação e informação, buscando alcançar os objetivos da organização, sendo considerado de fácil acesso e baixo custo.

Assim, a ampla expansão do comércio eletrônico aconselha que as empresas de varejo deverão buscar um diferencial significativo para atrair e fidelizar seus consumidores dentro do seu mercado, visto que os custos de mudanças são menores na Internet, do que os canais convencionais (HERNANDEZ, 2002). Observando o aumento do número de transações em lojas virtuais, as empresas estão cada vez mais usando esse canal como fator estratégico e alavancar seus negócios. Em 2018, o *e-commerce* apresentou 24% de crescimento no mundo todo, atingindo a marca de 2,9 trilhões de dólares em vendas. Esse valor representa 12% da porção das vendas mundiais para o varejo, (E-BIT, 2019)

Além disso, as características do novo ambiente empresarial, tais como globalização, integração interna e externa das empresas, entre outras, têm confirmado as tendências da criação e utilização de mercado e comércio eletrônico (ALBERTIN, 2007). O autor ainda afirma que devido aos impactos sociais e empresariais as organizações passaram a criar estratégias e fazer os planejamentos com base nesse mercado.

Nesse contexto, a qualidade tem um papel determinante no sucesso dos mercados varejistas *on-line,* pois está relacionada com a satisfação, retenção e lealdade dos clientes (WOLFINBARGER; GILLY, 2003). Anderson, Fornell e Lehmann (1994) acrescentam que organizações que alcançam elevada satisfação do consumidor resultam em retornos econômicos superiores.

O comportamento do consumidor já não é mais o mesmo, hoje é conectado, contestador e consciente de uma decisão de compra mais racional (E-BIT, 2019). Mesmo medindo a satisfação num determinado momento, como se fosse algo imutável, a satisfação é na verdade uma característica dinâmica, que evolui com o tempo influenciada por diversos fatores (ZEITHAML; BITNER; GREMLER, 2014).

Desta forma, é essencial que a organização busque compreender quais são as expectativas dos seus consumidores. Identificar quais fatores são mais proeminentes para o *site* e quais são as repercussões dessa qualidade nas compras futuras. Um maior conhecimento desses aspectos pode trazer um negócio com maiores chances de sucesso nas vendas no comércio eletrônico.

Os autores Zeithaml, Bitner e Gremler (2014) explicam que os conceitos dos termos ‘satisfação’ e ‘qualidade’ em serviço geralmente são confundidos, pois mesmo tendo pontos em comum, a satisfação é vista como um conceito mais amplo, enquanto a qualidade se concentra intrinsecamente nas dimensões do serviço. Sendo esta, um dos componentes da satisfação do cliente. Assim, a qualidade tem um grande impacto na satisfação do cliente, principalmente se levarmos em conta que a qualidade do serviço atual influencia mais do que se comparado com as experiências passadas anteriormente (ANDERSON; FORNELL; LEHMANN, 1996).

Zeithaml, Bitner e Gremler (2014) acrescentam que os clientes avaliam a qualidade com base em diferentes fatores levando em consideração o contexto em que estão inseridos. É necessário então compreender os fatores relevantes para os consumidores virtuais pois esta análise facilitará a coordenação dos processos gerenciais na obtenção da satisfação dos consumidores eletrônicos.

Devido à grande quantidade de dados fornecidos por meio das transações *on-line* torna-se difícil obter alguma informação útil em um tempo hábil para tomada de decisão. A análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas (GOLDSCHMIDT; PASSOS, 2005).

Neste contexto, este trabalho pretende analisar os fatores que geram maiores satisfações para o cliente quando este efetua uma compra *on-line*. Os fatores buscados terão como base as dimensões de serviço do *e-commerce* de Parasuraman, Zeithaml e Malhotra (2005) e serão analisados a partir da associação entre o ponto crítico do serviço prestado, e a avaliação do consumidor na plataforma da loja virtual. Devido ao enorme volume de dados e por oferecer confiabilidade nos resultados futuros, será aplicado o processo de Descoberta de Conhecimento na Base de Dados para mensurar o relacionamento dos fatores relevantes com a satisfação do cliente.

2 REVISÃO DE LITERATURA

Neste capítulo será apresentado a base teórica deste trabalho. É dividido em seis assuntos principais: descoberta de conhecimento em base de dados, etapa de mineração de dados, *e-commerce*, satisfação e insatisfação, qualidade do serviço e qualidade do serviço do *e-commerce*.

2.1 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

A capacidade de gerar, coletar e armazenar dados digitalmente tornou-se uma tarefa acessível diante dos avanços na área de Tecnologia da Informação. Devido ao maior acesso à equipamentos eletrônicos, como computadores pessoais e outros dispositivos o número de usuários conectados na rede aumentou significativamente. O universo digital está dobrando de tamanho a cada dois anos e será multiplicado por dez entre 2013 e 2020, com valores de 4.4 trilhões de gigabytes para 44 trilhões de gigabytes (EMC, 2012).

Com essa perspectiva de crescimento no volume de dados que transitam pelas organizações e o consequente armazenamento desses dados, é necessário o uso de tecnologias que auxiliem relacionar e interpretar esses dados para traçar planos estratégicos aplicados em cada contexto. De acordo com Han, Kamber. e Pei (2012, p.5), a enorme quantidade de dados, de rápido crescimento, coletados e armazenados em grandes e numerosos repositórios de dados, excedeu em muito, nossa capacidade humana de compreensão sem o uso de ferramentas poderosas. Goldschmidt e Passos (2005), reforçam esse pensamento, afirmando que a análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas.

A grande preocupação das empresas está na descoberta da informação útil para auxiliar no processo de tomada de decisão e elaboração de estratégias. Nesse contexto, as informações são os atributos chave para a obtenção de conhecimento. Normalmente elas estão implícitas em grandes bancos de dados e com pouca acessibilidade sem auxílio das ferramentas adequadas.

Nesse contexto está inserido o processo denominado Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases - KDD)* que abrange todo o uso dessas ferramentas com o objetivo de gerar conhecimento a partir de grandes bases de dados. Goldschmidt e Passos (2005) explicitam que descobrir conhecimento é extrair dos dados o que eles implicam em termos de riscos – a evitar – e oportunidades – a serem aproveitadas.

O termo KDD é descrito como uma análise e modelagem automática e exploratória de grandes repositórios de dados. Assim, o KDD é o processo organizado de identificar padrões válidos, novos, úteis e compreensíveis a partir de conjuntos de dados grandes e complexos (MAIMON; ROKACH, 2010).

A mineração de dados tenta transformar muita desinformação (na forma de dados espalhados) em informações úteis, criando modelos e regras. Sua meta é usar os modelos e regras para prever um comportamento futuro, melhorar seu negócio, ou apenas explicar coisas que caso contrário não seria possível explicar. Estes modelos confirmam o que já era esperado, ou ainda melhor, podem encontrar coisas novas em nossos dados que nem sabíamos que existiam (ABERNETHY, 2010).

Para entender esse conceito, faz-se necessário entender a diferença e o processo de transformação entre dados, informação e em seguida conhecimento ilustrado na Figura 1.

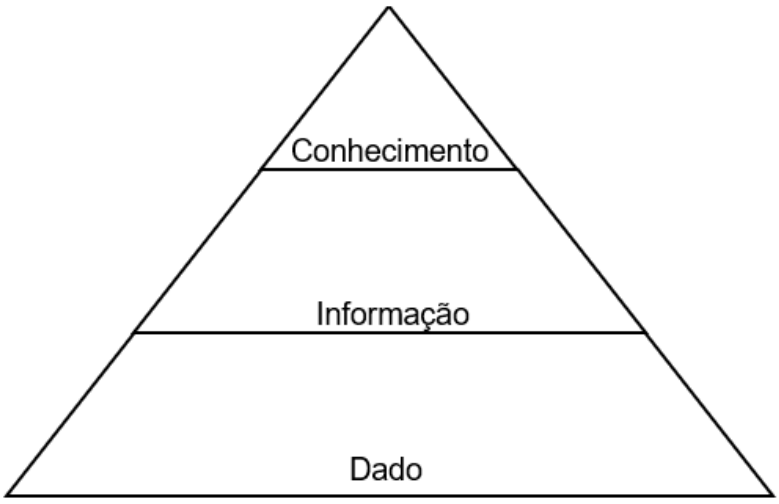

Figura 1: Hierarquia entre dado, informação e conhecimento.
Fonte: adaptado de Goldschmidt e Passos (2005).

Dados são um conjunto de fatos distintos e objetivos, relativos a eventos. Num contexto organizacional dados são utilitariamente descritos como registros

estruturados de transações (DAVENPORT; PRUSAK, 1998). São entendidos como fatos em sua forma primária, onde seu significado depende da sua relação com outros fatos e inseridos em um contexto.

Quando esses registros ou fatos são organizados ou combinados de forma significativa, eles se transformam numa informação (BEAL, 2004). Segundo McGee e Prusak (1994), informação consiste em dados coletados, organizados, orientados, aos quais são atribuídos significados e contexto. A informação é a base para a construção do conhecimento.

No topo da pirâmide encontra-se o conhecimento. Da mesma forma que a informação é originada a partir dos dados, o conhecimento também é produzido a partir de incrementos da informação. Davenport e Prusak (1998) conceituam o conhecimento como uma mistura fluida de experiência condensada, valores, informação contextual e *insight* experimentado, a qual proporciona uma estrutura para a avaliação e incorporação de novas experiências e informações. Dessa forma, o conhecimento é uma mistura de vários elementos que o transforma em algo complexo.

Em geral, o conhecimento não pode ser abstraído das bases de dados por recursos tradicionais da Tecnologia da Informação (GOLDSCHMIDT; PASSOS, 2005). Esse conhecimento se tornou um recurso essencial para as organizações. Não se trata mais de reduzir custos e atender com qualidade, mas interpretar os objetivos, expectativas e desejos dos clientes para conseguir se destacar no mercado global. O processo KDD é composto por várias etapas e provê um método automático para cobrir padrões em Banco de Dados (BD).

Segundo Goldschmidt e Passos (2005) o termo KDD foi formalizado em 1989 em referência ao amplo conceito de procurar conhecimento a partir de base de dados. O KDD evoluiu e continua a evoluir da interseção de campos de pesquisa como aprendizado de máquinas, reconhecimento de padrões, bancos de dados, estatísticas, inteligência artificial, aquisição de conhecimento, visualização de dados e computação de alto desempenho (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

2.1.1 Processo *Knowledge Discovery in Databases*

Uma das definições mais difundidas é a dos pesquisadores Fayyad, Piatetsky-Shapiro e Smyth (1996) como sendo um processo de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados.

Segundo Maimon e Rokach (2010) deve ser notado que em cada etapa o processo é “iterativo”, o que significa voltar para ajustar etapas anteriores pode ser necessário. O termo “interativo” indica a necessidade de atuação do homem como responsável pelo controle do processo (GOLDSCHMIDT; PASSOS, 2005). Segundo o autor a expressão “não trivial” alerta para a complexidade normalmente presente na execução de processos de KDD.

O processo de descoberta do conhecimento em base de dados é dividido em cinco etapas principais: seleção; pré-processamento e limpeza de dados; transformação dos dados; Data Mining; Interpretação e avaliação dos resultados. Na Figura 2 é ilustrado o processo.

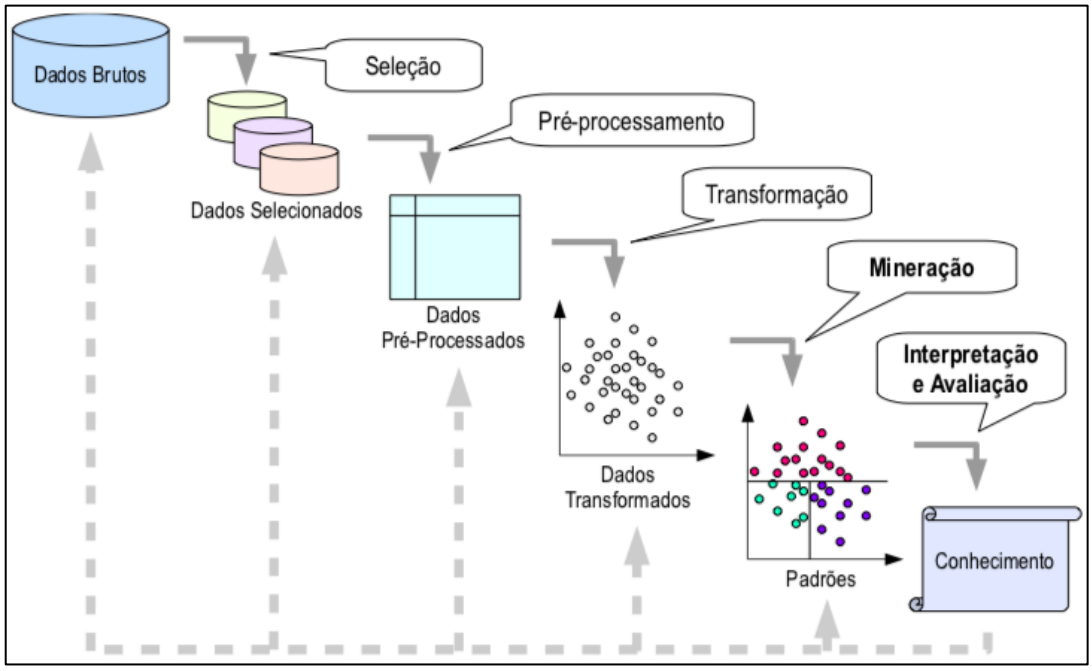

Figura 2: Processo de Descoberta de Conhecimento em Banco de Dados.
Fonte: Santos (2009).

O processo começa com a compreensão do domínio da aplicação, do conhecimento prévio necessário, e a determinação dos objetivos do usuário final do

processo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Esse passo, antecede o passo (seleção) ilustrado na figura 2.

Com os objetivos definidos, os dados que serão utilizados na aplicação do processo devem ser determinados. Goldschmidt e Passos (2005) explicam que em essência, é a identificação de quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo de KDD. Essa etapa é muito importante, pois com base nesses dados que mineração de dados aprende e descobre os padrões desejados. Se algum atributo importante estiver faltando então o estudo inteiro pode falhar (MAIMON; ROKACH, 2010).

Na segunda etapa, pré-processamento e limpeza, a confiabilidade dos dados é aprimorada. Informações ausentes, errôneas, inconsistentes ou discrepantes (*outliers,* que interferem inadequadamente o processo), são eliminadas do processo através de técnicas como manipulação de valores ausentes (*missing value*) e remoção de ruídos ou *outliers*.

Após as etapas de seleção e pré-processamento dos dados é necessário formatá-los da maneira mais adequada para obter o aprendizado buscado. Essa etapa consiste na transformação dos atributos com o objetivo de facilitar a etapa de mineração de dados. Métodos aqui incluem redução de dimensão (como seleção e extração de recurso e amostragem de registro) e transformação de atributo (como a discretização de atributos numéricos e transformação funcional), (MAIMON; ROKACH, 2010). Segundos os autores, essa etapa é crucial para o andamento do KDD, mas que é específica para cada projeto.

A quarta etapa é referente à mineração de dados, nessa fase, serão estabelecidos os algoritmos e os métodos utilizados para buscar padrões nos dados. A escolha deve buscar atingir os objetivos definidos na primeira etapa do processo. De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), essa etapa envolve a escolha dos modelos e parâmetros que correspondem com o método de mineração de dados particular com os critérios gerais do processo de KDD. A etapa será descrita mais detalhadamente devido a sua importância para o processo.

Por último, a interpretação dos resultados obtidos, ou seja, a análise dos padrões encontrados pelas tarefas de mineração aplicadas, a avaliação do atendimento aos objetivos propostos, e a aplicação do conhecimento obtido (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). A última etapa abrange o tratamento, interpretação e a avaliação dos resultados, ou seja, a análise do

conhecimento obtido através da mineração de dados. Aqui, o objetivo principal é utilizar as relações descobertas para tomar decisões importantes e avaliar se as causas do problema foram sanadas ou o objetivo da empresa alcançado (CARVALHO, 2002).

2.2 ETAPA DE MINERAÇÃO DE DADOS DO PROCESSO KDD

A etapa de mineração de dados (do inglês, *Data Mining* – DM) está no centro de todo o processo KDD, essa é a etapa onde é definido o algoritmo e as técnicas a serem utilizadas. Essa escolha é definida com base nos objetivos traçados no início do processo e nas etapas anteriores. Para Goldschmidt e Passos (2005, p.18) é nessa etapa que é realizada a busca efetiva por conhecimentos úteis no contexto da aplicação de KDD. É a principal etapa do processo de KDD. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), DM é uma das etapas do KDD, onde tem-se uma descoberta de padrões únicos a partir da aplicação de algoritmos específicos e análise de dados.

Tan, Steinbach e Kumar (2009) ressaltam que, mineração de dados é o processo automático de descoberta de informação valiosa em grandes bases de dados para identificar padrões úteis e que eram desconhecidos anteriormente, possibilitando até a previsão de resultado a partir do aprendizado com os dados. Para Witten, Frank e Hall (2011), DM é sobre resolver problemas a partir da análise de um banco de dados já existentes, ou ainda descobrir padrões nos dados de maneira automática ou semiautomática. Segundo Carvalho (2002), é o uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertos a olho nu pelo ser humano.

De acordo com Tan, Steinbach e Kumar (2009) as tarefas da mineração de dados está geralmente dividia em duas categorias:

a) Preditiva: consiste no uso de algumas variáveis ou campos da BD para prever valores futuros ou desconhecidos de variáveis de interesses. Frequentemente é referida como mineração de dados supervisionada (MAIMON; ROKACH, 2010);

b) Descritiva: visa buscar padrões (correlações, agrupamentos, trajetórias e anomalias) que resumem os relacionamentos subjacentes. Para Fayyad, Piatetsky-Shapiro e Smyth essa categoria foca em descobrir padrões nos dados, que possam ser interpretados por humanos. Mineração de dados descritiva inclui os aspectos não-supervisionados e visualização (MAIMON; ROKACH, 2010).

Para alcançar os objetivos desejados as duas categorias utilizam-se de tarefas como: classificação, sumarização, agrupamento, regressão, modelagem de dependência, e detecção de mudanças ou desvios.

2.2.1 Classificação

A classificação é uma das técnicas mais utilizadas do DM simplesmente porque é uma das tarefas cognitivas humanas mais realizadas no auxílio à compreensão do ambiente em que vivemos (CARVALHO, 2002). Essa tarefa consiste na descoberta de uma função preditiva que mapeia um conjunto de dados em um conjunto de variáveis predefinidas denominados classes. De acordo com Goldschmidt e Passos (2005, p.13) uma vez descoberta, tal função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram. Algumas das ferramentas utilizadas para essa tarefa são: redes neurais, *Back-Propagation,* classificadores Bayesianos, estatísticas e algoritmos genéticos

É possível utilizar classificação para desenvolver uma ideia do tipo de cliente, item ou objeto, descrevendo vários atributos para identificar uma determinada classe. Por exemplo, é possível classificar carros facilmente em diferentes tipos (sedan, 4x4, conversível) identificando atributos diferentes (número de lugares, formato do carro, rodas motrizes) (BROWN, 2012). Outro exemplo, é a de categorizar os dados para empréstimos bancários. Separa-se em duas classes, pessoas aprovadas para empréstimo e não aprovadas, com base nas receitas e os débitos dos clientes. Nesse caso o banco pode utilizar essa função de classificação para determinar empréstimos futuros e diminuir os riscos tomados como ilustrado na Figura 3.

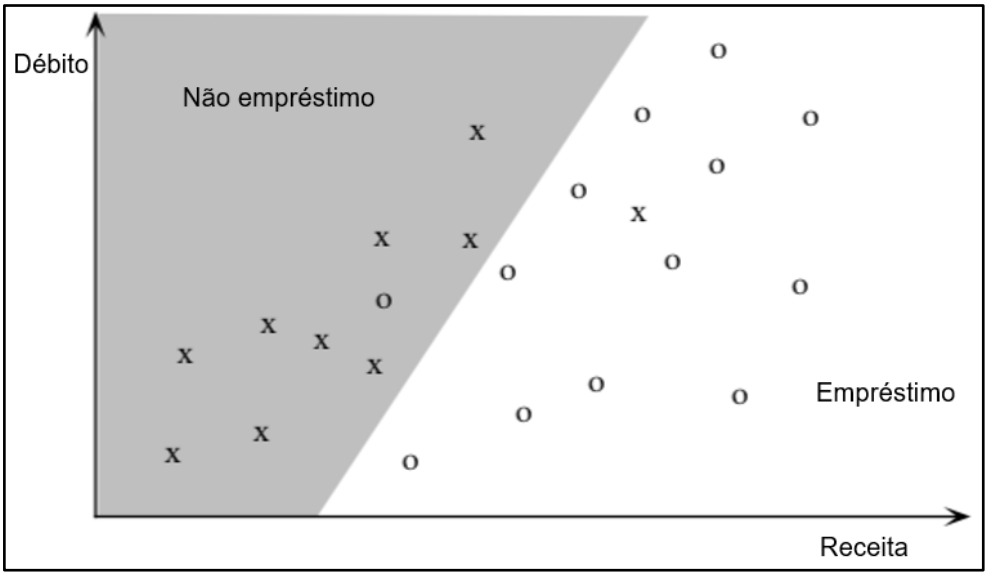

Figura 3: Classificação linear para banco de dados de empréstimos;
Fonte: adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

É importante notar que não é possível separar as classes perfeitamente utilizando a classificação linear, algumas variáveis não são abrangidas pelo limite de decisão devido as particularidades das dimensões e restrições do problema.

2.2.2 Sumarização

De acordo com Weiss e Indurkhya (1998), essa tarefa, muito comum em KDD, consiste em procurar identificar e indicar características comuns entre conjuntos de dados. É muito comum aplicar a tarefa de sumarização a cada um dos agrupamentos obtidos pela tarefa de agrupamento (GOLDSCHMIDT; PASSOS, 2005).

Um exemplo é a análise de perfil de clientes que assinam uma revista. A tarefa busca então características similares a maioria dos clientes. Assim, a título de exemplo poderia fornecer um resultado como: são assinantes da revista Y, mulheres na faixa etária 20 a 30 anos, trabalham na área de produção e possuem nível superior. Esse resultado auxilia o direcionamento do marketing da revista para a oferta de novos assinantes.

2.2.3 Agrupamento (Clustering)

Segundo Tan, Steinbach e Kumar (2009), essa tarefa procura encontrar grupos de observações intimamente relacionadas, para que as observações que pertencem ao mesmo *cluster* sejam mais semelhantes entre si do que as observações que pertencem a outros grupamentos. O objetivo dessa tarefa é maximizar similaridade intracluster e minimizar similaridades intercluster (GOLDSCHMIDT; PASSOS, 2005).

Diferentemente da classificação, o agrupamento não possui classes definidas, e cabe ao algoritmo descobrir os grupos. De acordo com Carvalho (2002), a dificuldade reside no fato de que pode não haver tais classes, ou seja, os dados distribuem-se equitativamente por todo o espaço possível, não caracterizando nenhuma categoria.

Alguns algoritmos utilizados nessa tarefa são: K-Means, K-Modes, K-Prototypes, K-Medoids, Kohohen, dentre outros.

A Figura 4 ilustra o exemplo de uma amostra de dados de vendas que compara a idade do cliente com o poder de compra. Pode ser identificado dois clusters, um em torno de R$ 2.000,00/grupo de idade de 20 a 30 anos e outro de R$ 7.000,00 a R$ 8.000/grupo de idade de 50 a 65 anos (BROWN, 2012).

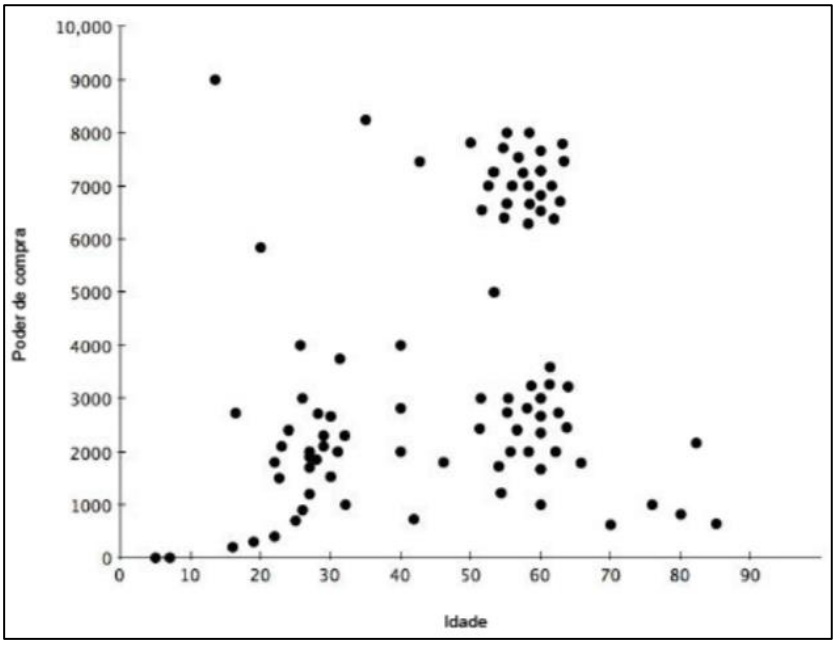

Figura 4: *Clustering* com base no poder de compra e idade.
Fonte: Brown (2012).

Vale ressaltar, que os grupos podem se sobrepor permitindo que alguns dados pertençam a mais de um *cluster*, esses casos indicam que os conceitos definidos para os grupos estão pouco definidos apresentando grande similaridade intercluster.

2.2.4 Regressão (Estimativa)

Para Santos (2009), é a descoberta de uma função preditiva de forma similar à feita em classificação, mas com o objetivo de calcular um valor numérico real ao invés de obter uma classe discreta. Segundo Goldschmidt e Passos (2005), compreende a busca por uma função que mapeie os registros de um banco de dados em valores reais, sendo restrita apenas a atributos numéricos.

As aplicações de regressão são diversas, estimar a probabilidade de um paciente sobreviver dado os resultados dos diagnósticos, prever a demanda de um novo produto como uma função dos gastos com marketing, definição do limite do cartão de crédito para cada cliente em um banco, dentre outros.

2.2.5 Modelagem de Dependências (Regras de Associação)

Provavelmente, a associação (ou relação) é a técnica de mineração de dados mais conhecida, mais familiar e mais direta (BROWN, 2012). Essa tarefa abrange a busca por itens que frequentemente ocorram de forma simultânea em transações da BD (GOLDSCHMIDT; PASSOS, 2005).

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), a associação é uma técnica de modelagem de dependência. Com ela é possível identificar um modelo que apresenta dependências (regras) significativas entre valores de um atributo de um conjunto de dados.

Um exemplo muito característico dessa tarefa é o carrinho de supermercado, do qual é extraído diversas informações sobre os produtos que os consumidores

compram em conjunto. Isso possibilita a realização de vendas dirigidas nas quais os itens oferecidos já em conjunto com preço ligeiramente menor (CARVALHO, 2002).

Algoritmos como o Apriori, GSH, DHP, entre outros, são exemplos de ferramentas utilizadas na implementação dessa tarefa de descoberta de associações.

2.2.6 Detecção de Mudanças ou Desvios

Segundo Santos (2009), são técnicas que permitem a descoberta e identificação de dados que não se comportam de acordo com um modelo aceitável dos dados (ou, por exemplo, mudanças em séries temporais ou em dados indexados por tempo). Esses registros observados são denominados *outliers.*

O objetivo do algoritmo de detecção de anomalias é descobrir as verdadeiras anomalias e evitar rotular falsamente objetos normais como anômalos (TAN; STEINBACH; KUMAR, 2009). Como exemplo de aplicação, é a tarefa de detectar desvios em compras no cartão de crédito dos clientes. O algoritmo deve buscar compras cujas características sejam diferentes do perfil normal de compras do dono do cartão.

2.3 E-COMMERCE

A Internet ganhou força e revolucionou o mundo somente a partir do advento da *World Wide Web* (WWW) ou somente *Web.* Esse marco, proporcionou a criação de novas formas de fazer negócios que avançaram gradativamente e evoluem constantemente com a expansão dos usuários mundiais. O comércio eletrônico ou *E-commerce* surgiu como consequência dos avanços tecnológicos e da popularização da internet, inicialmente tinha o propósito de manter a comunicação de diferentes bases militares durante a Guerra Fria, sendo atualmente um assunto em destaque no que tange as relações comerciais e perspectivas de faturamento.

Segundo Albertin (2010) comércio eletrônico (CE) é a realização de toda a cadeia de valor dos processos de negócio num ambiente eletrônico, por meio da

aplicação intensa das tecnologias de comunicação e de informação, atendendo aos objetivos de negócio.

Conforme Alves (2000), o comércio eletrônico é definido como qualquer forma de transação de negócios em que as partes interagem eletronicamente, ou seja, sem contatos físicos ou diretos. Dessa forma, o CE não precisa necessariamente utilizar a *Web* para realizar transações, as máquinas automáticas de refrigerantes podem ser consideradas comércio eletrônico, ou o pagamento através do cartão de crédito ou ainda uma transação via *fax*.

Os novos modelos de negócios permitem classificar o comércio eletrônico por meio da análise das modalidades de relacionamento entre os agentes que participam da rede, conforme demonstrado no Quadro 1.

Modelo	**Classificação do Segmento**
B2C – Business to Consumer	Transações de comercio eletrônico, de organização para consumidores.
B2B – Business to Business	Transações comerciais entre empresas ou entidades.
B2G – Business to Govern	Transações comerciais entre empresas privadas e governamentais.
B2I – Business to Institutions	Atividades comerciais entre empresas e instituições (educacionais, associações etc.).
B2E – Business to Employee	Comércio eletrônico em que empresas vendem seus serviços ou produtos aos seus funcionários.
E-Procurement	Comércio eletrônico utilizado pelas empresas para compra de suprimentos.
C2C – Consumer to Consumer	Comércio eletrônico entre consumidores de forma direta tais como: Mercado Livre, OLX e outros.

Quadro 1: Modelos e classificações de E-commerce.
Fonte: Adaptado de Turchi (2012).

Por volta de 1995 o comércio eletrônico começou a deslanchar nos EUA devido ao surgimento de muitas empresas virtuais. No Brasil, esse processo começou cinco anos depois, com várias lojas iniciando esse novo modelo de vendas *on-line* e até então as vendas não pararam de crescer (ALMEIDA; BRENDLE; SPINDOLA, 2014).

A tendência de crescimento do *e-commerce* é mundial, de acordo com *E-bit* (2019) o comércio eletrônico apresentou 24% de crescimento no mundo todo em 2018, atingindo a marca de 2,9 trilhões de dólares em vendas. As transações *on-line* apresentam crescimento maior que o varejo tradicional em praticamente todos os

países que já operam o comércio eletrônico. O volume do *e-commerce* representa 12% de todas as vendas mundiais e a América Latina se mostra como grande oportunidade, pois apenas 2,7% do total consumido é feito digitalmente (E-BIT, 2009).

Ainda conforme o *E-bit* (2019) o Brasil é o mercado mais desenvolvido em termos de *e-commerce* da América Latina, com importância de vendas de 4,3% e crescimento de dois dígitos. O faturamento alcançou cerca de R$ 53,2 bilhões em 2018 e a previsão para 2019 é de crescimento de 15% desse valor representando um faturamento de R$ 61,2 bilhões, como ilustrado no Gráfico 1.

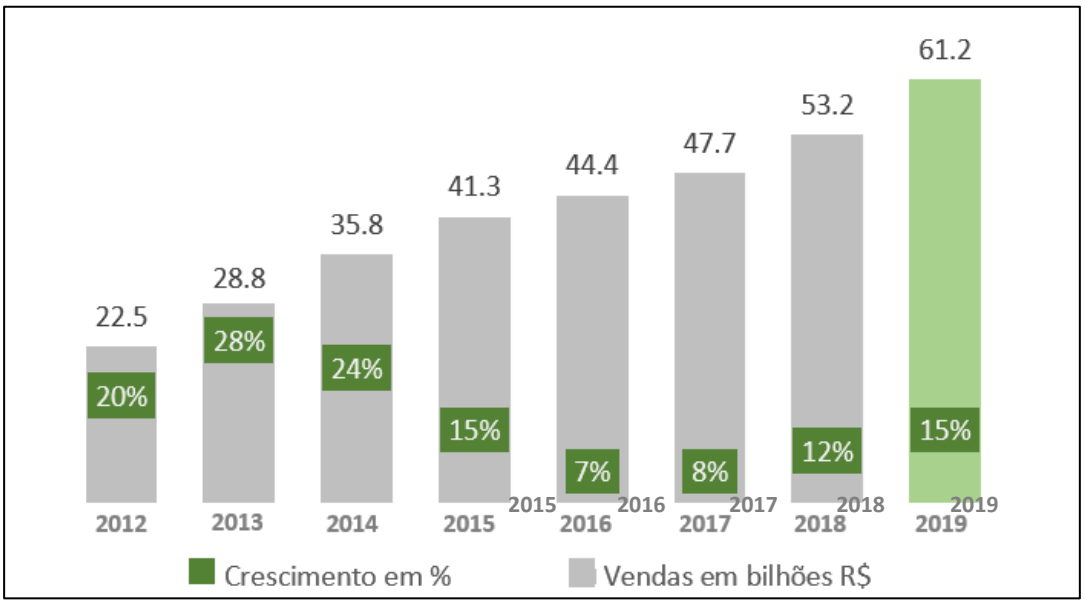

Gráfico 1: crescimento do e-commerce no Brasil.
Fonte: adaptado do relatório *Webshoppers* - *E-bit* (2019).

Sendo assim, à medida que a população possui acesso à internet, fatores como a facilidade de crédito, melhoria da segurança nas transações eletrônicas juntamente com a entrada de novos *players,* tiveram como consequência o aumento consecutivo no faturamento do *e-commerce.*

O CE deixa de ser um novo meio de comercializar seus produtos e se transforma em um mecanismo necessário para atingir as necessidades dos seus clientes. Diniz et al. (2011) afirmam que no decorrer dos anos o comércio virtual passou a ser uma ferramenta importantíssima para que as empresas possam conhecer as necessidades dos clientes e aumentar as vendas deixando de ser somente um diferencial. Neste contexto, a empresa precisa saber lidar com esta ferramenta, sendo importante estar sempre atenta aos fatores de sucesso para a venda e as mudanças do mercado.

2.3.1 E-consumidores

De acordo com Esteves (2011) consumidores eletrônicos são clientes que utilizam a internet para efetuarem suas compras virtuais. Além dessa denominação, o cliente virtual pode ser chamado de *e-consumer*, ou simplesmente consumidor, entre outras.

Em relação aos consumidores digitais Morais (2011), define os e-consumidores como pessoas que buscam na *web* algo além da compra, trata-se de consumidores incomuns. Esses consumidores desejam estreitar o relacionamento e interagir com a marca, tendo a oportunidade de pesquisar e comprar preços se baseando em todo o contexto oferecido por uma loja. O comportamento deles já não é mais o mesmo, eles não são limitados pela loja física. Uma compra pode se iniciar com uma experiência na loja física, ser pesquisada no *smartphone* e concluída no *desktop*. (E-BIT, 2019).

Dessa forma os canais se retroalimentam e trazem uma jornada de consumo muito mais complexa, o uso da internet e suas ferramentas para compartilhamento de informações sobre os produtos, é um grande influenciador no processo de decisão de compra, o consumidor se tornou contestador e mais consciente de sua decisão de compra. Segundo o *E-bit* a jornada de compras está se tornando mais complexa, em 2018 o brasileiro visitou, em média, 8 canais diferentes para fazer as suas compras. O consumidor pesquisa mais e tende a comprar da marca que oferece as informações que ele procura (SALOMÃO, 2018).

Em razão da versatilidade e facilidade apresentada pela *web,* o número de pessoas que utilizam métodos eletrônicos para efetuar uma aquisição está crescendo a cada dia. Segundo pesquisa do *E-bit* (2019)*,* o número de consumidores ativos no Brasil aumentou de 31,27 milhões em 2013 para 58,51 milhões em 2018.

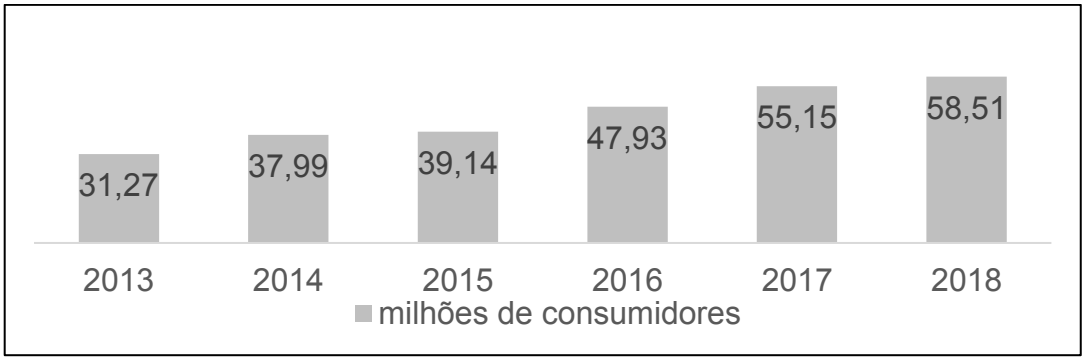

Gráfico 2: crescimento dos e-consumidores no Brasil.
Fonte: adaptado do relatório *Webshoppers* - *E-bit* (2019).

Neste contexto de crescimento, Macedo et al. (2010) afirmam que em um mercado competitivo, acelerado e de consumidores cada vez mais exigentes, pode fazer toda diferença traçar o perfil do consumidor e assim adotar melhores estratégias de negociação, consequentemente aumentando a satisfação do cliente. Assim, as organizações devem buscar entender quais são os fatores mais relevantes para seu público-alvo para efetuar um atendimento estratégico.

Dessa maneira é necessário que as empresas entendam este novo ambiente e busquem soluções rápidas que consigam satisfazer os anseios de seus atuais e potenciais clientes (ESTEVES, 2011).

2.4 SATISFAÇÃO E INSATISFAÇÃO

A satisfação dos consumidores é reconhecida como ponto central dos estudos de *marketing*. A justificativa é que através da análise dos comportamentos, decorridos da satisfação ou insatisfação, podem ser levantadas implicações gerenciais úteis, que resultam em aprimoramento no relacionamento com os clientes e práticas eficazes de *marketing.*

Para Kotler e Keller (2012) a satisfação é a comparação feita por uma pessoa sobre o desempenho percebido de um produto em relação as suas expectativas. Quando o desempenho não atinge as expectativas o cliente fica satisfeito. Já quando o desempenho supera as expectativas, o cliente fica encantado.

Zeithaml, Bitner e Gremler (2014) elucidam que a definição de ‘satisfação’ e ‘qualidade em serviço’ geralmente são confundidos, porém são fundamentalmente

diferentes quanto aos seus agentes causais e aos desfechos resultantes. Embora tenha alguns pontos em comum, a satisfação é entendida como um conceito mais amplo, ao passo que a qualidade do serviço se volta especificamente para as dimensões do serviço. Assim, a qualidade percebida é um componente da satisfação do cliente.

Os autores ainda acrescentam que a satisfação, por ser mais inclusiva, é influenciada pelas percepções da qualidade do serviço, pela qualidade do produto, e pelo preço, além de fatores situacionais e pessoais, ilustrado na figura 5. Sendo assim as organizações devem trabalhar para entender o que os clientes esperam, levando em consideração todos esses fatores, para conseguir atingir as expectativas críticas, e por fim, conseguir fidelizá-los.

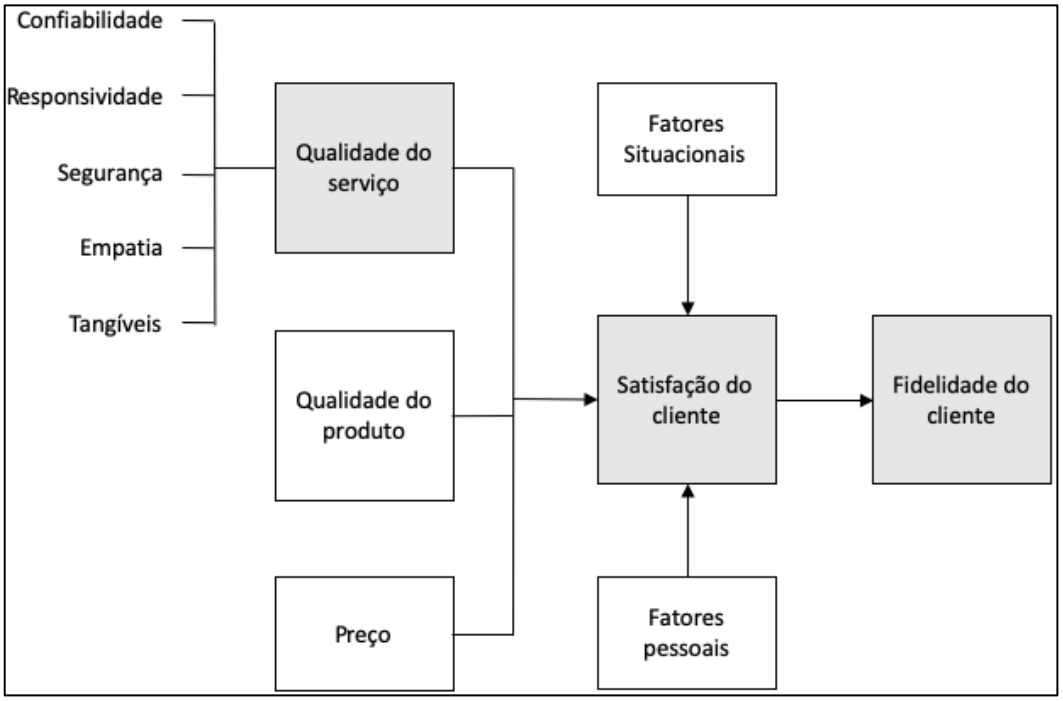

Figura 5: Satisfação e as percepções do cliente.
Fonte: adaptado de Zeithaml, Bitner e Gremler (2014).

Conforme ilustrado na Figura 5 a satisfação é influenciada por atributos específicos do produto ou serviço, pela compreensão da qualidade e pelo preço. Além desses, fatores oriundos da esfera pessoal, como humor ou atitude mental do cliente e fatores situacionais, como percepções e opiniões de colegas e funcionários. O Quadro 2 descreve esses fatores.

Fatores	Descrição
As características do produto e serviços	Satisfação influenciada pela avaliação das características e atributos do produto ou serviço
As emoções do consumidor	As emoções positivas ou negativas dos consumidores afetam a percepção de sua satisfação
As causas do sucesso ou do fracasso	Satisfação é influenciada pela avaliação dos motivos da discrepância entre a expectativa e a realidade
As percepções de igualdade ou justiça	Noções de justiça e igualdade influenciam a satisfação dos clientes com relação a produtos e serviços
Outros consumidores, familiares e colegas	Satisfação é influenciada pelas reações e sentimento de pessoas externas aos consumidores.

Quadro 2: Fatores Determinantes a Satisfação.
Fonte: adaptado de Zeithaml, Bitner e Gremler (2014).

Segundo Corrêa e Gianesi (2019) um cliente satisfeito pode ser um importante aliado na conquista de novos clientes, assim como um cliente insatisfeito pode arruinar um negócio. Nesse sentido, Oliver (1997) afirma que a relação entre satisfação e lucratividade se manifesta em uma sequência de etapas: qualidade (neste caso, uma performance excepcional) levaria à satisfação do cliente, que por sua vez seria o caminho à lealdade, culminando com uma maior lucratividade.

Essas relações se mostram ainda mais coerentes quando Anderson, Fornell e Lehmann ressaltam para a necessidade de visão de longo prazo, pois os retornos econômicos resultantes da melhoria da satisfação dos clientes não são percebidos imediatamente. A satisfação impactará então nos comportamentos futuros. Oliver (1997) reforça essa posição entendendo que o efeito da satisfação nos lucros ocorre por meio da retenção (ou lealdade). De acordo com o autor, clientes satisfeitos:

a) aumentam o volume de compra por esses produtos;
b) toleram elasticidades de preços mais facilmente;
c) estão mais isolados ou menos atentos a ofertas da concorrência;
d) são mais atentos a ações de comunicação dessas empresas;
e) diminuem os custos de transações futuras.

Assim a satisfação dos consumidores ganha um papel essencial nas organizações, seja pelas relações com a lucratividade e *market share* ou por sua relação com a lealdade (LOVELOCK; WRIGHT, 2001).

No outro lado, as atribuições de insatisfação resultam quando o consumidor está decepcionado com as compras, decorrente do não atendimento as suas expectativas. A fidelidade do cliente pode cair abruptamente quando ele atinge um

dado nível de insatisfação ou sempre que ele está insatisfeito com os atributos essenciais do serviço (ZEITHAML; BITNER; GREMLER, 2014).

2.5 QUALIDADE DO SERVIÇO

O conceito de qualidade como visto atualmente teve sua origem na década de 50, período posterior a segunda guerra mundial em que o Japão necessitava reestabelecer a economia que entrava em colapso devido as sanções da nova ordem mundial. Conceito este, muito difundido nas indústrias, mas que se diferencia em alguns pontos quando levado para a prestação de serviços.

De acordo com Kotler, Hayes e Bloom (2002), serviço são ações, desempenho ou ato que é essencialmente intangível e não acarreta necessariamente a propriedade do que quer que seja. Sua criação pode ou não estar vinculada a um produto material.

Anderson, Fornell e Lehmann (1994) definem qualidade percebida como uma avaliação global dos serviços realizados pelo fornecedor. Zeithaml, Bitner e Gremler (2014), reforçam essa interpretação definindo como o julgamento do consumidor sobre a excelência ou superioridade de um produto.

Entretanto, é importante explicitar que a qualidade percebida pode tomar diferentes conceitos levando em consideração o serviço prestado. No caso de serviços puros (cuidados de saúde, educação) a qualidade é sempre o componente dominante nas avaliações. Quando os serviços são oferecidos junto com um produto físico, a qualidade do serviço pode ser igualmente crucial na definição de satisfação do cliente (ZEITHAML; BITNER; GREMLER, 2014).

Dessa forma, as organizações necessitam entender os critérios que atraem os olhos do seu consumidor de maneira a compreender as suas expectativas e assim poder oferecer um serviço de qualidade e torná-lo o seu diferencial frente aos competidores. Esses critérios de avaliação devem refletir os fatores que determinam a satisfação do cliente, ou em outras palavras, qualidade do projeto e da prestação de serviço (CORRÊA; GIANESI, 2019).

Zeithaml, Bitner e Gremler (2014) sugerem, a partir de pesquisas, que os clientes não percebem a qualidade de modo unidimensional; ao contrário, eles julgam

a qualidade com base em fatores diversos, relevantes ao contexto. Levando isso consideração e, como o objetivo é conseguir atingir de forma eficaz o consumidor, não se deve priorizar os atributos sem antes realizar algum tipo de pesquisa com os clientes buscando minimizar os riscos no momento da escolha dos critérios.

Em relação ao dimensionamento da qualidade do serviço o trabalho de Zeithaml, Parasuraman e Berry (1990), é um dos mais reconhecidos na área. Os autores criaram um instrumento que busca mensurar a percepção dos consumidores a respeito da qualidade do serviço, utilizando cinco dimensões aplicáveis a diversos contextos.

a) Confiabilidade: a capacidade de executar o serviço prometido de forma confiável e precisa;
b) Responsividade: é a disposição de ajudar os clientes e fornecer o serviço imediatamente;
c) Segurança: remete ao conhecimento e a cortesia dos funcionários, e sua capacidade de transmitir credibilidade e certeza;
d) Empatia: é a atenção individualizada dispensada aos clientes;
e) Tangíveis: constituem a aparência das instalações físicas, do equipamento, dos funcionários e dos materiais de comunicação.

Essas dimensões refletem como os clientes organizam as informações sobre a qualidade do serviço em suas mentes. Zeithaml, Bitner e Gremler (2014) complementam que além do tipo de serviço prestado, a diferença cultural também afeta a importância dada às cinco dimensões.

Sendo assim, uma empresa que busca atingir seus consumidores deve objetivar fundamentalmente as suas satisfações, assim deve manter uma equipe capaz de compreender a real expectativa com o serviço e fornecer meios para que esta possa desenvolver um trabalho de qualidade.

2.6 QUALIDADE DO SERVIÇO DO E-COMMERCE

O amplo desenvolvimento do comércio eletrônico no decorrer dos últimos anos fez com que as empresas buscassem um diferencial significativo para manter e atrair seus consumidores. Yoo e Donthu (2001) indagam em seu artigo, sobre o

crescente aumento da pressão na obtenção do entendimento do quesito qualidade *on-line*, já que com a experiência do consumidor, as expectativas sobre os negócios *on-line* estão aumentando.

A análise da qualidade do serviço prestado é uma característica primordial para o desenvolvimento do comércio eletrônico. Em termos gerenciais, é essencial as empresas entenderem como os consumidores avaliam a qualidade de uma compra *on-line*, quais elementos são mais relevantes para o site e quais são os impactos da qualidade sobre as intenções de compras futuras.

Devido ao potencial do alcance da internet as lojas virtuais desempenham atualmente um papel estratégico para diversos negócios tradicionais e *on-line*. Produtos são apresentados a possíveis clientes, e estes podem ou não ser motivados a realizar a compra. Diante disso, Torres (2013) propõe que os projetos de CE devem ser estabelecidos com base no comportamento dos seus e-consumidores como resultado do acesso à informação devido à internet. Eles esperam estar no foco das atenções, independente do canal de conexão com a organização.

A explícita necessidade de entender sobre as expectativas do consumidor virtual, fez com que no início dos anos 2000 eclodisse diversos estudos sobre o tema. Em um dos estudos mais difundidos na área, Parasuraman, Zeithaml e Malhotra (2005) definem o constructo qualidade do serviço eletrônico como sendo “a extensão pelo qual o *site* facilita de modo eficiente e efetivo, a compra, aquisição e a entrega do bem”.

Wolfinbarger e Gilly (2003) reforçam afirmando que a qualidade é definida além da *interface* do *site,* em que a experiência de compra do consumidor consiste em todos os fatores desde a busca pela informação, a precificação do produto, a decisão, realização da transação, envio do bem, retorno e o serviço ao consumidor.

Após extensa pesquisa bibliográfica, aplicação de questionário, e análise com fator exploratório Parasuraman, Zeithaml e Malhotra (2005) identificaram quatro dimensões críticas à avaliação dos *websites*:

a) Eficiência: facilidade e velocidade de acessar o site;
b) Preenchimento/cumprimento: representa a extensão na qual o site promete entregar o pedido de modo correto e eficaz, juntamente com a sua disponibilidade;
c) Disponibilidade do sistema: compreende a funcionalidade correta e técnica do site;

d) Privacidade: representa o grau no qual o site demonstra a segurança e sigilo nas informações particulares do consumidor.

O estudo também revelou três dimensões que o cliente utiliza para julgar a recuperação do serviço quando algum problema ou dúvida aparece, sendo:

a) Responsividade: é o gerenciamento efetivo dos problemas e dos retornos através do *site*;
b) Compensação: é o grau pelo qual o *site* compensa o cliente pelo problema;
c) Contato: é a disponibilidade de assistência por meio do telefone ou atendimento *on-line.*

A pesquisa dos autores, foi elaborada exclusivamente para mensurar a qualidade em varejo eletrônico e não leva em consideração *sites* adversos de entretenimento, o que torna o estudo mais preciso quanto ao objetivo deste trabalho.

3 PROCEDIMENTOS METODOLÓGICOS

Neste capítulo é descrito a classificação da pesquisa e são apresentadas as etapas realizadas para execução da aplicação do estudo em busca de atingir os objetivos definidos neste trabalho.

3.1 CLASSIFICAÇÃO DA PESQUISA

De acordo com Kuark, Manhães e Medeiros (2010) é muito importante compreender os tipos de pesquisas para delimitação dos instrumentos e procedimentos que o pesquisador utilizará no planejamento de sua investigação. As pesquisas podem ser classificadas de diferentes formas e variam conforme a sua natureza, por sua abordagem ao assunto, propósito dos objetivos e aos procedimentos metodológicos.

Desta forma, a presente pesquisa pode ser classificada de natureza aplicada, de abordagem quantitativa e qualitativa, com objetivo da pesquisa descritivo e exploratório e os procedimentos utilizados para alcance dos dados foram o levantamento bibliográfico e pesquisa documental. Na figura 6 é observado a classificação da pesquisa.

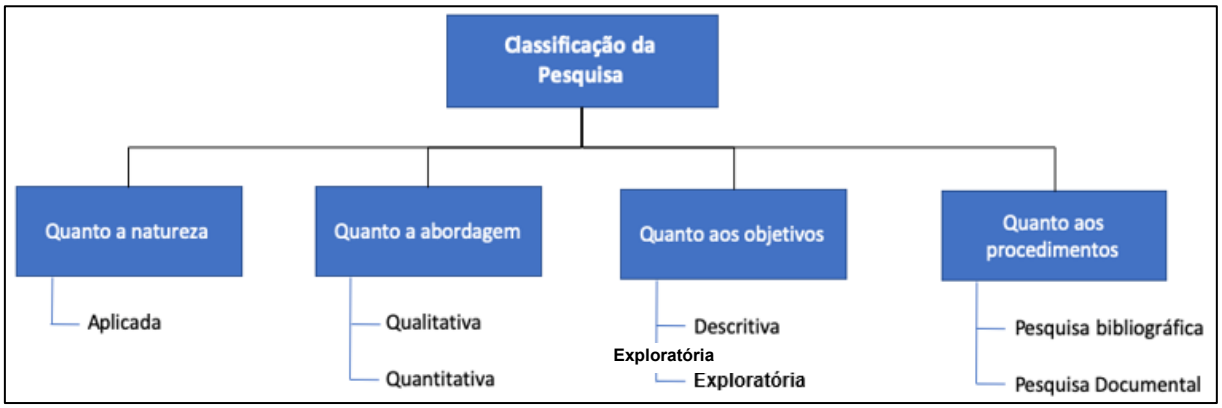

Figura 6: Classificação da pesquisa.
Fonte: Autoria própria (2020).

Segundo Prodanov e Freitas (2013), pesquisa aplicada visa gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos

envolvendo verdades e interesses locais. Para Gil (2010) a pesquisa aplicada abrange estudos desenvolvidos com o propósito de solucionar problemas observados no âmbito da sociedade em que o pesquisador se encontra. O trabalho busca utilizar o conhecimento da investigação em aplicações reais. Ou seja, a partir da teoria levantada no trabalho serão enumerados os fatores determinantes para a satisfação do cliente no *e-commerce* e através do processo KDD, os fatores serão analisados para gerar conhecimento sobre o grau de influência de cada. O resultado gerado, terá futuras aplicações práticas em organizações que praticam o comércio eletrônico.

Do ponto de vista da abordagem Oliveira (2011) afirma que a pesquisa qualitativa usa o ambiente natural como fonte direta para coleta de dados e o pesquisador é o instrumento-chave. Kauark, Manhães e Medeiros (2010) consideram que há uma relação dinâmica entre o mundo real e o sujeito, ou seja, o mundo objetivo e a subjetividade não podem ser traduzidos em números. Sendo a interpretação dos resultados e a atribuição de conhecimento são relacionados com a pesquisa qualitativa, na qual os pesquisadores analisam dado, com foco na abordagem, o processo e o significado do estudo.

A abordagem quantitativa considera o que pode ser quantificável, o que significa traduzir em números opiniões e informações para classificá-las e analisá-las. Requer o uso de recursos e de técnicas estatísticas (KAUARK; MANHÃES; MEDEIROS, 2010). Segundo Fonseca (2002), a pesquisa quantitativa é centrada na objetividade, e como as amostras são grandes, os resultados constituem um retrato de toda a população. O autor ainda afirma que a utilização conjunta da pesquisa quantitativa e qualitativa permite recolher mais informações do que se poderia conseguir isoladamente.

Considera-se a abordagem qualitativa e quantitativa, pois buscou utilizar dados para entender e analisar o comportamento dos e-consumidores do varejo Olist diante dos fatores de satisfação, também foram utilizados métodos estatísticos e algoritmos para encontrar as relações existentes no BD através da quantificação dos fatores tornando objetivo a interpretação dos resultados.

Quanto aos objetivos, essa pesquisa pode ser classificada de duas maneiras, objetivo descritivo e exploratório. A pesquisa descritiva segundo Gil (2010) tem o objetivo de descrever características de determinada população e podem ser também elaboradas com a finalidade de descobrir relações entre as variáveis. Nas pesquisas descritivas os fatos são observados, registrados, analisados, classificados e

interpretados, sem que o pesquisador interfira sobre eles, (PRODANOV; FREITAS, 2013). Tanto a identificação da associação entre os fatores e a satisfação dos consumidores e a análise, classificação e interpretação dos fatos observados no BD têm caráter descritivo.

Em relação à pesquisa exploratória, Gil (2010) explica que tem como principal finalidade desenvolver, esclarecer e modificar conceitos e ideias, tendo em vista a formulação de problemas mais precisos e hipóteses pesquisáveis para estudos posteriores. Devido à complexidade de obter dados complementares ao presente estudo, e pela proposta de desenvolvimento de hipóteses para criar análises gerais, o trabalho assume a condição exploratória, no que tange o objetivo.

Em relação ao ponto de vista dos procedimentos utilizados, a pesquisa documental é caracterizada pela fonte de coleta de dados estar restrita a documentos, escritos ou não. Podendo ser feitas quando ocorreu o fato ou depois, (MARCONI; LAKATOS, 2010). Gil (2010), aponta que normalmente o documento mais comum é constituído por um texto escrito em papel, mas estão se tornando cada vez mais usuais os documentos eletrônicos, disponíveis sob os mais diversos formatos. De acordo com o autor, vale-se de materiais que não receberam ainda um tratamento analítico, ou que ainda podem ser reelaborados de acordo com os objetivos da pesquisa.

Já a pesquisa bibliográfica é elaborada com base em material já publicado constituídos principalmente de livros e artigos científicos, (GIL, 2010). Sua finalidade é colocar o pesquisador em contato direto com tudo que foi escrito, dito ou filmado sobre determinado assunto, (MARCONI; LAKATOS, 2010). Segundo Gil (2010), a sua principal vantagem está no fato de possibilitar ao pesquisador uma gama de fenômenos muito mais ampla da aquela que poderia pesquisar diretamente.

Nesta pesquisa, foram utilizados arquivos eletrônicos de organizações particulares, constituídos principalmente pelo BD utilizado no presente trabalho, além do uso de dados estatísticos para caracterizar o aumento da aplicação da Internet e do *e-commerce*. A execução desse trabalho foi constituída pela revisão bibliográfica de diversos autores sobre os assuntos principais do estudo, sendo assim classificado como pesquisa bibliográfica.

3.2 ETAPAS PARA DESCOBRIMENTO DO CONHECIMENTO

Neste subcapítulo é realizado o detalhamento das etapas aplicadas no trabalho que garantiram o atingimento do objetivo. Os primeiros passos executados antecedem o processo do KDD. Neste momento inicial, buscou-se ter a compreensão da extensão do trabalho, a obtenção do conhecimento técnico necessário para sua realização e a determinação do objetivo a ser alcançado. Esses passos iniciais são fundamentais para nortear o descobrimento do conhecimento na base de dados.

A linguagem de programação python™ apresenta alto nível de flexibilidade, com aplicações diversas no mercado de trabalho e uma grande comunidade de suporte. A linguagem cresceu com a proposta de facilitar a análise de dados com bibliotecas cujas aplicações são focadas em gerar insights em diversos setores da economia (PYTHON, 2019). E para esse fim, o Google criou a plataforma Kaggle – um *site* que hospeda diversos dados públicos disponíveis permanentemente para fomentar o conhecimento em técnicas de ciência de dados utilizando a linguagem de programação python™.

Com a definição da ferramenta e o estudo prévio da linguagem, pode-se dar continuidade no processo do KDD. O código fonte do presente estudo é encontrado no Apêndice A. As atividades realizadas para solucionar o objetivo principal do trabalho “Quais fatores influenciam na satisfação do consumidor digital?” podem ser visualizadas na Figura 7 em forma de fluxograma.

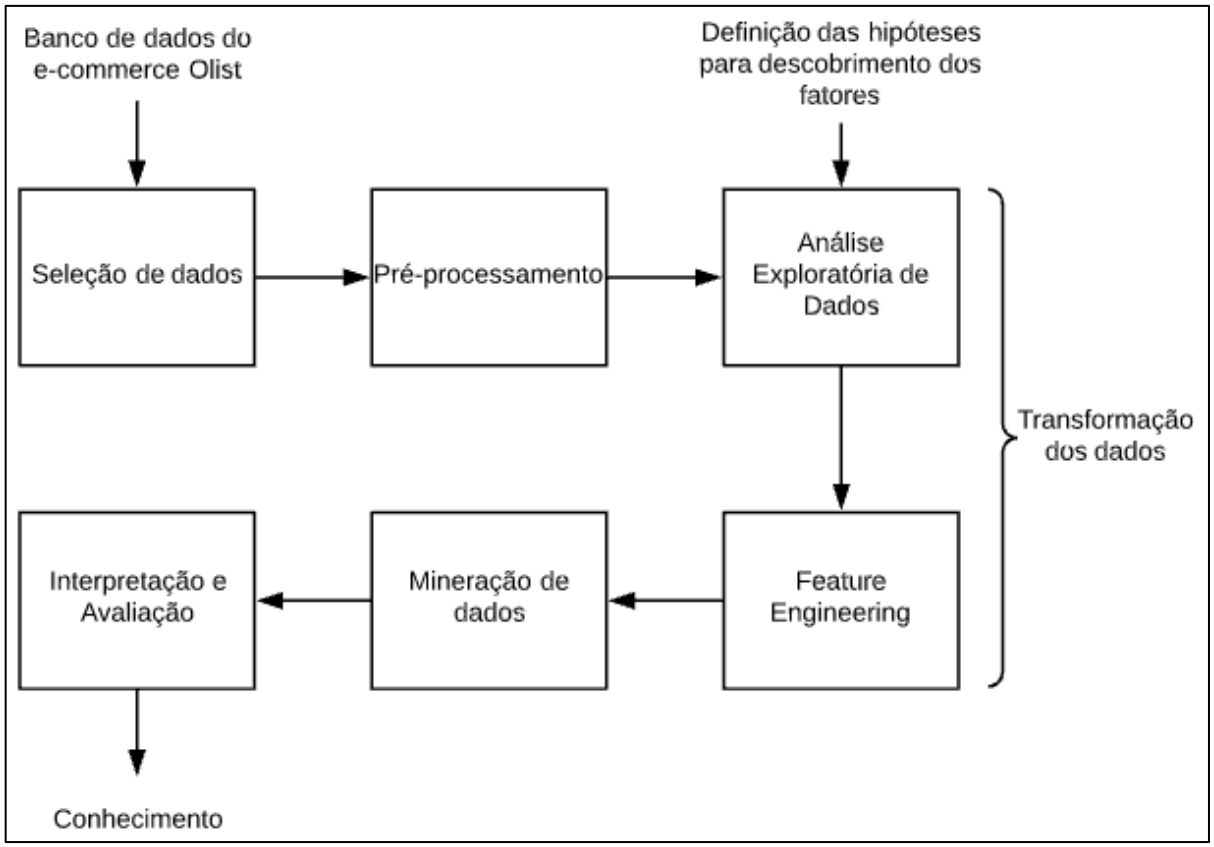

Figura 7: Fluxograma das etapas de execução da pesquisa.
Fonte: Autoria própria (2020).

A primeira etapa, Obtenção dos Dados, compreendeu a pesquisa e obtenção do banco de dados. A segunda etapa, Seleção de Dados, tem o foco em delimitar os atributos que serão utilizados limitando o escopo aos dados convergentes com o objetivo do trabalho. O Pré-processamento teve por objetivo consolidar os dados extraídos a partir de tratamento e formatação.

A quarta etapa, Análise Exploratória de Dados, consistiu na análise de conjunto dos dados selecionados visando resumir suas características principais. A quinta etapa, *Feature Engineering,* foi utilizada para extrair recursos dos dados brutos e melhorar o desempenho dos algoritmos utilizados na etapa seguinte.

Na sexta etapa, Mineração de Dados, utilizou-se modelos estatísticos para realizar o treinamento dos dados para encontrar padrões. E a sétima etapa, Interpretação e Avaliação, teve o objetivo de analisar os resultados e elaborar uma interface que permitiu visualizar o que foi encontrado na etapa seis. A seguir será apresentado com mais detalhes cada etapa do processo.

3.2.1 Obtenção de Dados

O início do KDD consiste na aquisição do banco de dados que será utilizado durante o processo. Os dados foram disponibilizados no *site* Kaggle, uma subsidiária da Google LLC. Essa plataforma é uma comunidade *online* para cientistas de dados que tem o propósito de fornecer ferramentas e recursos para o desenvolvimento do conhecimento sobre dados aos seus usuários.

Esse banco de dados pertence empresa de tecnologia brasileira chamada Olist. A escolha foi baseada na disponibilidade, facilidade de manipulação e o volume de informações fornecidas que se alinhavam com o objetivo de encontrar os fatores relevantes aos consumidores virtuais.

A empresa é a principal loja de departamentos inseridos nos marketplaces (*sites* especializados em venda de produtos) brasileiros e atua como um canal de venda entre pequenos lojistas e os principais *e-commerces* do Brasil. Os lojistas vendem seus produtos através desse *site* e a entrega aos consumidores também é realizada pelo sistema de logística da Olist.

Na Figura 8 ilustra-se o anúncio feito por um lojista em um *marketplace*, neste anúncio, é possível reparar que o nome que aparece como vendedor é a empresa Olist, e não o lojista anunciante.

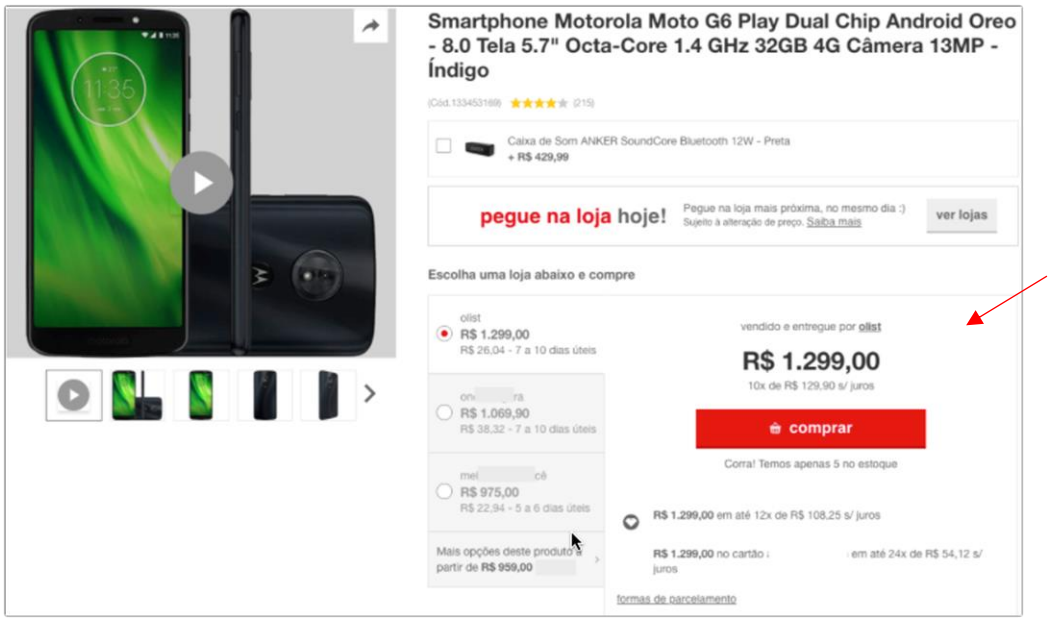

Figura 8: Anúncio feito pelo Olist em um *marketplace*.
Fonte: Adaptado de Kaggle (2020).

O banco de dados possui informações de 100 mil compras realizadas no período de 2016 a 2018. Os seus atributos permitem analisar a compra por diferentes dimensões: situação do pedido, preço, pagamento e desempenho do frete, atributos do produto e as avaliações feitas pelos consumidores. Assim, os fatores levantados por Parasuraman, Zeithaml e Malhotra, convergem com os disponíveis no BD.

Para melhor organização e compreensão os dados estão divididos em múltiplos BD, em formato *Comma Separated Values* (CSV), e seguem a estrutura do esquema ilustrado na Figura 9**.** Os atributos de relacionamento entre os BD estão destacados no centro das setas que representam essa conexão.

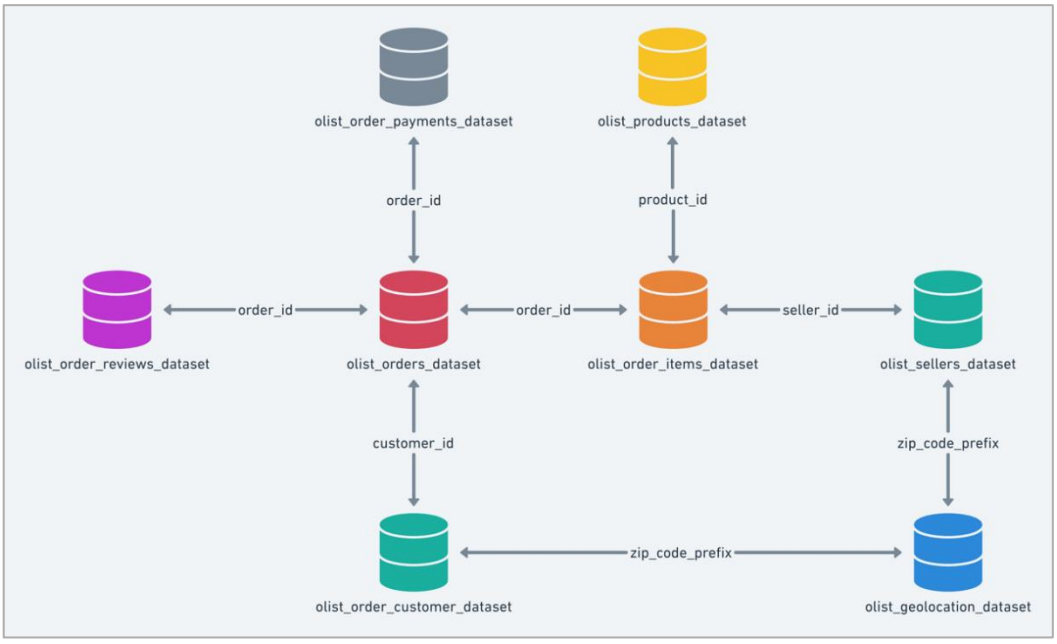

Figura 9: Esquema da estrutura do banco de dados.
Fonte: Kaggle, (2020).

Após a efetivação de uma compra de produto, o lojista recebe uma notificação para finalizar a ordem de compra. E assim que o produto é recebido pelo cliente ou a data estimada de entrega é superada, o consumidor recebe uma pesquisa de satisfação por *e-mail* onde ele pode dar uma nota para a experiência de compra. Essa nota é armazenada no banco de dados *'olist_order_reviews_dataset'* na variável dependente '*reviews_score'*. Elas variam de 1 a 5, e o objetivo do trabalho é prever a influência que outros atributos podem exercer na experiência de compra do cliente.

3.2.2 Seleção de Dados

De posse de todos os dados, o próximo passo é definir quais atributos são relevantes para manter no conjunto de dados estudado e quais não vão contribuir para o andamento do trabalho. Isso é importante para limitar o escopo de atuação nos passos subsequentes e otimizar o processamento de dados.

Uma vez que não há informação suficiente sobre as dimensões: eficiência, disponibilidades do sistema e privacidade, faz-se necessário concentrar os esforços na análise da dimensão de preenchimento/cumprimento. Dessa maneira, foram selecionadas as tabelas que focassem em avaliar se o pedido foi entregue de maneira correta e eficaz com base no anúncio disponível no *website.* Na Figura 10 é ilustrado o esquemático dos BD selecionados.

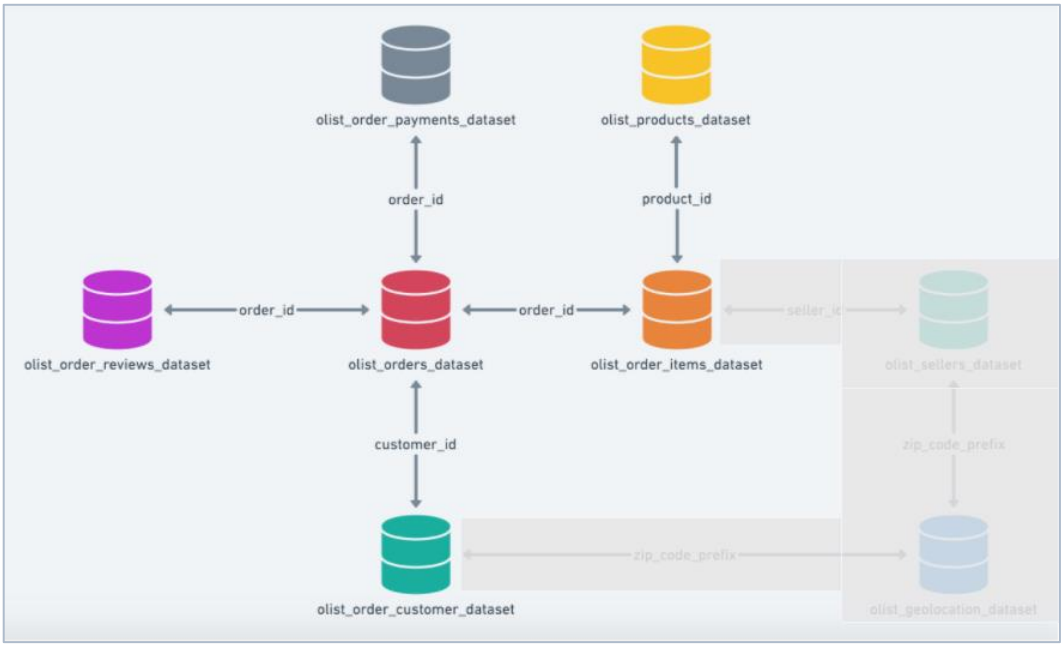

Figura 10: Estrutura do banco de dados com tabelas selecionadas
Fonte: Adaptado Kaggle (2020)

Como resultado, é retornada uma única tabela com todas as informações selecionadas que possibilitará o estudo da correlação entre os dados.

3.2.3 Pré-processamento

Na etapa de pré-processamento é realizada uma “limpeza” nos dados para se tornarem interpretáveis pelos processos computacionais que vão ser aplicados. É verificado a legibilidade dos dados pelo ambiente que será utilizado na etapa de análise. Para esta pesquisa, foi utilizado um notebook interpretador da linguagem Python com bibliotecas de análise de dados como pandas, *scikit-learn, seaborn, numpy*. Por haver dados preenchidos por usuários ou pela integração de BD, é possível encontrar erros não previstos que implicam em barreiras na hora de processá-los.

Primeiro, investiga-se os tipos de dados e se eles são interpretáveis. Nessa etapa é possível compreender como foram registrados os acontecimentos e quais são as regras que regem o negócio.

Em seguida, é apurado a existência de dados faltantes que serão adicionados utilizando regras estatísticas que suponham seu comportamento conforme a distribuição. A partir dessa etapa, foi possível inferir que, como trata-se de uma empresa de grande porte, os dados já vieram limpos, portanto, não houve necessidade de imputar dados faltantes.

Por fim, foi estimada a granularidade dos dados e se esta satisfaz o objetivo da análise. A Granularidade significa a entidade mínima que representa uma linha da tabela. Nesse caso, a tabela foi reestruturada para que cada atributo se refere a uma ordem de pedido, ou “*order_id*”. Assim, cada linha representa uma experiência de venda única com uma nota referente, evitando duplicidade de dados para uma mesma ordem de compra.

3.2.4 Transformação de Dados

A etapa de transformação de dados é dividida em duas fases que combinadas pretendem transformar dados brutos em informações complementares que possam alimentar o sistema. Tipicamente, é a etapa onde é dedicada a maior energia dos cientistas pois, ao interpretar minunciosamente os gráficos na Análise Exploratória de

Dados (EDA), é possível criar diferentes atributos a partir de suas descobertas durante a *Feature Engineering* e dar mais robustez ao processo. É comum experimentar uma melhora significativa nas métricas de avaliação após transformar os atributos de maneira assertiva.

3.2.4.1 Análise Exploratória de dados (EDA)

Trata-se de uma investigação estatística detalhista onde os dados tomam uma forma diferente e passam a ter significado real para quem conduz a análise. Um erro de interpretação nesse cenário pode afastar consideravelmente o projeto do seu objetivo, para tanto o processo é repetido várias vezes para que cada cenário possa ser considerado e as decisões tomadas a partir dessa etapa sejam mais assertivas.

Para melhor entendimento do processo, a EDA foi segmentada em três partes, conhecimento dos dados, exploração da variável *target* e desenvolvimento de hipóteses. Cada um dos processos é descrito a seguir.

a) Conhecimento dos dados: A primeira etapa é necessária para entendimento inicial dos dados. É observado os dados temporais associados ao banco de dados, onde busca-se obter padrões de tendência sazonal.
b) Exploração da variável target: Os recursos utilizados para minerar os dados referentes a esse banco de dados devem ser estudados minuciosamente para que a solução aplicada seja confiável para aquele conjunto de dados e ofereça a melhor explicação para a questão levantada pelo estudo. Portanto, investiga-se a variável dependente afim de entender como é a sua distribuição e se existe necessidade de limitar o escopo com amostras da população.
c) Desenvolvimento de hipóteses: A partir da análise inicial do banco de dados e o estudo minucioso da variável dependente, são propostas suposições sobre a causa dos fenômenos observados na relação das variáveis independentes com a variável *target*.

3.2.4.2 Feature engineering

Antes de qualquer aprofundamento no processo de KDD, é importante separar os dados em teste e treino. Essa é uma prática de aproveitamento de dados que se certifica que o tratamento recebido pelo conjunto tratado possa ser repetido em dados novos e não deve haver a possibilidade de *leakage* – quando o modelo teve acesso aos resultados antes que ele pudesse aprender e invalidando todo o processo de conhecimento.

Durante a *Feature Engineering* são criadas variáveis que explicam processos que não estão literalmente especificados nos dados mas que podem ser calculados a partir desses – de forma que aumentem a correlação entre os fatos e a variável dependente ou, remova variáveis com alta colinearidade, afim de diminuir a complexidade do conjunto de dados.

3.2.5 Mineração de Dados

A etapa de mineração de dados compreende aplicar o conjunto de dados a modelos estatísticos e de *Machine Learning* para que a variável dependente possa ser mapeada computacionalmente a partir das variáveis que compõem o *dataset.* Dessa maneira, a grande questão que iniciou esse estudo “O que faz o cliente ficar satisfeito com uma compra de um *e-commerce*?” pode ser respondida pelos fatos que constituem essas transações comerciais.

Para o KDD, essa é a etapa em que um conjunto de dados é revertido em padrões de comportamento e a informação é transformada em conhecimento. A sua execução segue um fluxo direto e simples, porém é na sua interpretação que se extrai o valor do KDD.

4 RESULTADOS E DISCUSSÃO

Nesta seção serão expostos os resultados obtidos após completadas todas as etapas do KDD descritas anteriormente e suas respectivas análises de forma a resolver a questão que envolve esse estudo, como os dados podem ser uteis para melhorar a satisfação do cliente, dentro de um *marketplace* online.

4.1 ANÁLISE EXPLORATÓRIA DOS DADOS

4.1.1 Conhecimento dos Dados

Inicialmente buscou-se entender o comportamento de todo o conjunto de dados ao longo do tempo para entender as características principais das vendas.

Nos Gráficos 3 e 4 são ilustrados os valores de pagamento final acumulados anualmente e mensalmente, respectivamente em milhões de reais. O valor de pagamento foi calculado considerando a quantidade de itens na compra pelo preço e acrescido ao frete.

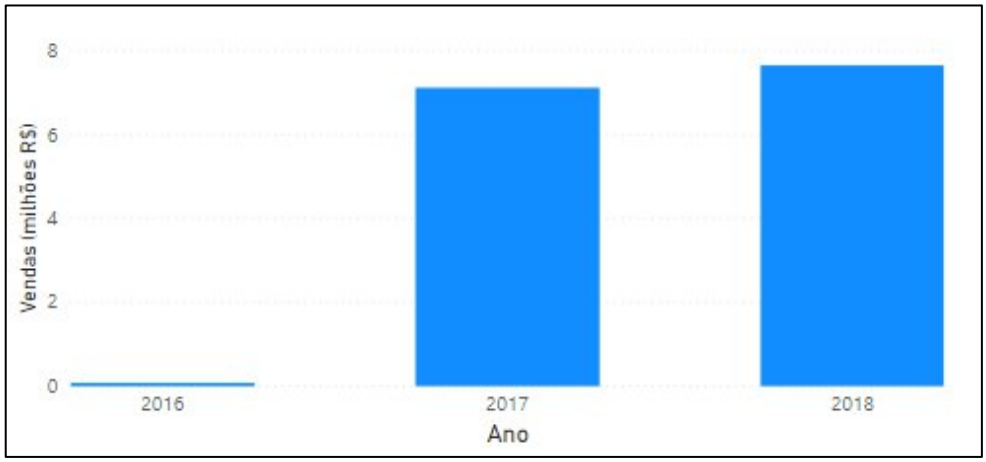

Gráfico 3: Somatória de vendas em milhões de reais através dos anos.
Fonte: Autoria própria (2020).

É possível observar que os dados são majoritariamente a partir de 2017, com um crescimento aproximado de 2 milhões de reais para o ano de 2018.

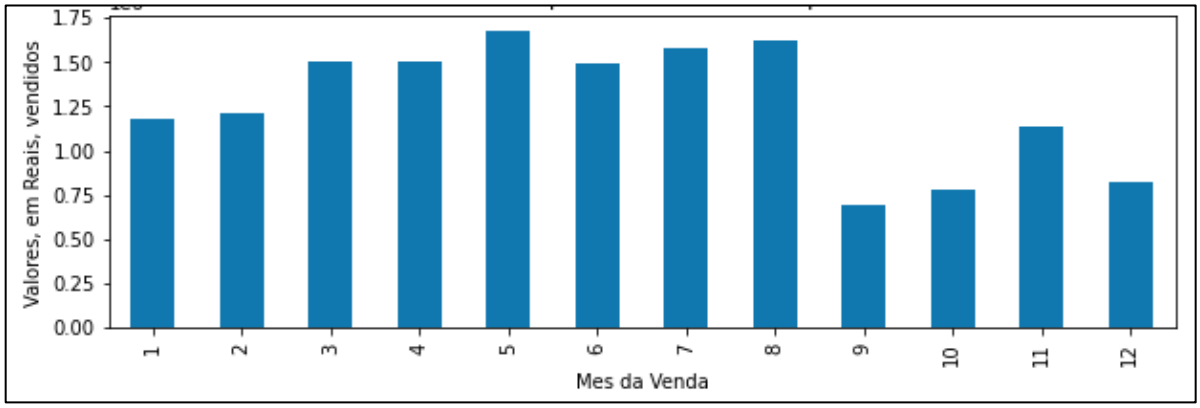

Gráfico 4: Somatória de vendas mensais em milhões de reais na plataforma.
Fonte: Autoria própria (2020).

Contrariando o senso comum, na qual as maiores vendas do varejo ocorrem no período de final do ano, os meses com maior faturamento foram maio, julho e agosto. No gráfico 4 é ilustrado o consolidado mensal das vendas nos três anos analisados, onde o pico de vendas ocorreu em maio.

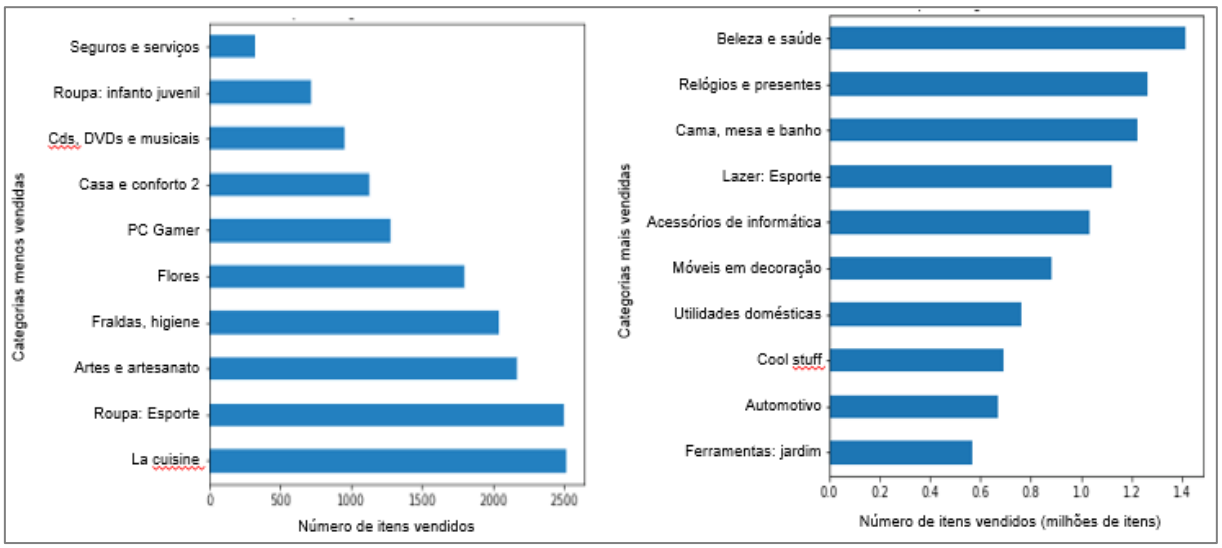

Gráfico 5: As dez categorias mais e menos vendidas no Olist.
Fonte: Autoria própria (2020).

A característica principal de um *marketplace* é a conexão de consumidores à vendedores terceiros através de uma plataforma. A intenção é disponibilizar uma grande variedade de produtos para criar uma rede de conexões e assim aumentar

número de transações. No Gráfico 5 é ilustrado o acontecimento desse fator, o *e-commerce* oferece 32.951 de produtos diferentes em 73 categorias diferentes.

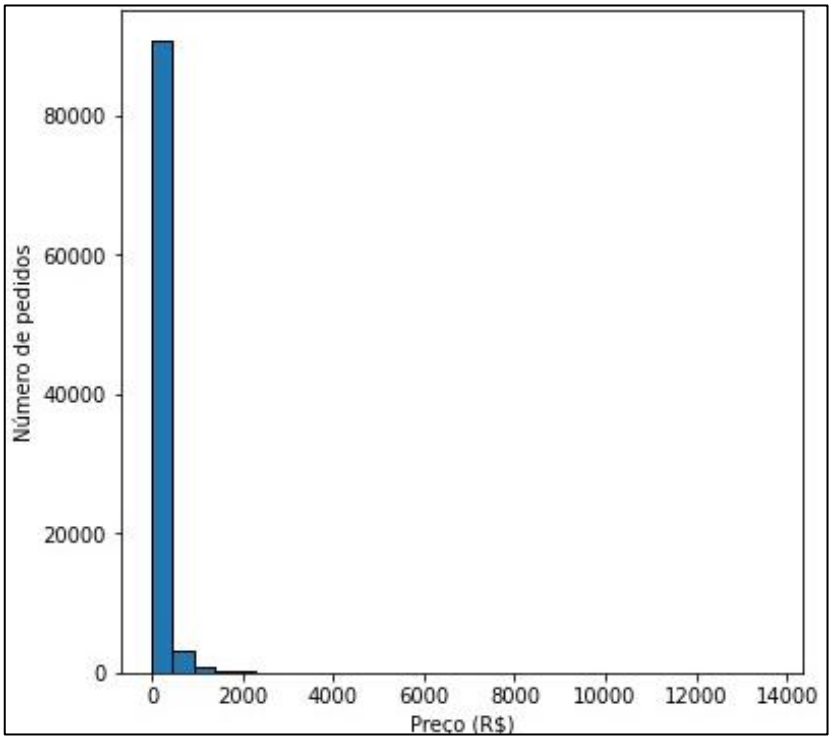

Gráfico 6: Histograma do preço pelo número de vendas dos produtos.
Fonte: Autoria própria (2020).

O histograma da variável preço unitário está apresentada no gráfico 6. O resultado expõe que o Olist é um *e-commerce* popular cujas vendas têm maior incidência em valores mais baixos, com a mediana da distribuição enviesada para a direita. As vendas com preço unitário inferior a mil reais compreendem quase a totalidade das vendas realizadas na plataforma reportadas através de um *dataset* de aproximadamente 95 mil *datapoints*. O que indica que a clientela é diversa e que o preço pode ser um fator que influência no tipo de serviço.

4.1.2 4.1.2 Exploração da variável dependente “Nota de avaliação”

Todo o processo de descobrimento da satisfação do cliente está sendo pautado na métrica de feedback recebido após a realização da entrega do produto vendido. A Nota de avaliação é um indicador que demonstra se o serviço prestado em

toda a jornada do consumidor atingiu as suas expectativas. Ela tem a proposta de ‘escutar’ o consumidor, é uma métrica estratégica para as empresas. Portanto, é importante entender como essa variável se comporta dependendo das condições impostas pelas demais variáveis, por isso ela é chamada variável dependente. No Gráfico 7 é ilustrado a distribuição da variável sem a influência de nenhum outro fator.

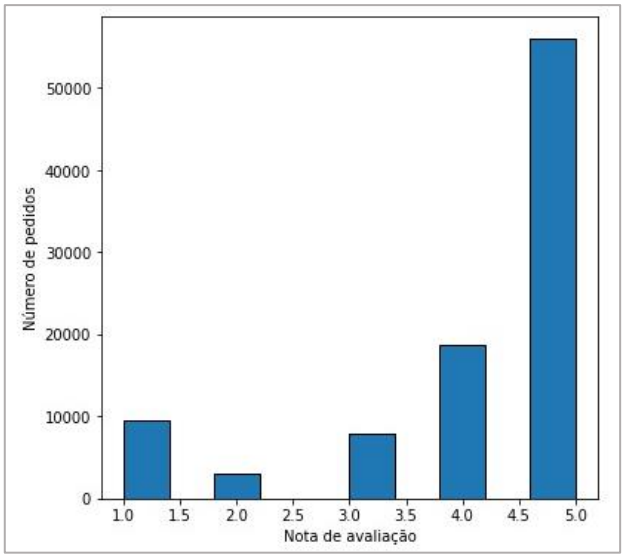

Gráfico 7: Histograma dos valores da variável Nota de avaliação.
Fonte: Autoria própria (2020).

Espera-se que modelos matemáticos tenham distribuição normal e assim exista uma diversidade de dados suficiente para não enviesar a forma na qual o modelo entende os resultados. É compreensível que essa distribuição não seja normal porque dentro desse espectro duas notas são mais esperadas: a nota máxima e a nota mínima, sendo na verdade uma reflexão da experiência da jornada de cada usuário, não representando uma nota de performance inerente ao serviço.

4.1.3 4.1.3 Desenvolvimento de Hipóteses

Com o estudo teórico dos fatores que influenciam na satisfação do e-consumidor e feita a análise inicial do BD. Foram propostas hipóteses a respeito da relação esperada das variáveis independentes com a nota de avaliação. Foram

levantadas seis hipóteses diferentes e complementares buscando abranger os principais fatores de satisfação do consumidor.

a) Número de caracteres na descrição do produto: cliente que possui mais informações sobre o produto terá uma relação positiva com a nota de avaliação.

b) Número de fotos do produto: mais fotos do produto terá uma relação positiva com a nota de avaliação.

c) Quantidade de itens comprados: clientes que compram mais de um produto terá uma relação positiva com a nota de avaliação.

d) Preço unitário do produto: preço mais alto influencia positivamente na nota de avaliação devido a expectativa de adquirir um produto de maior qualidade.

e) Valor pago no frete: entrega com valor mais elevado terá uma relação negativa com a nota de avaliação.

f) Prazo estimado e precisão na data de entrega: período de entrega menor e item que chega no prazo ou antes do que foi descrito deve ter uma relação positiva com a pontuação da revisão.

O estudo detalhado de cada hipótese está descrito a seguir no Gráfico 8.

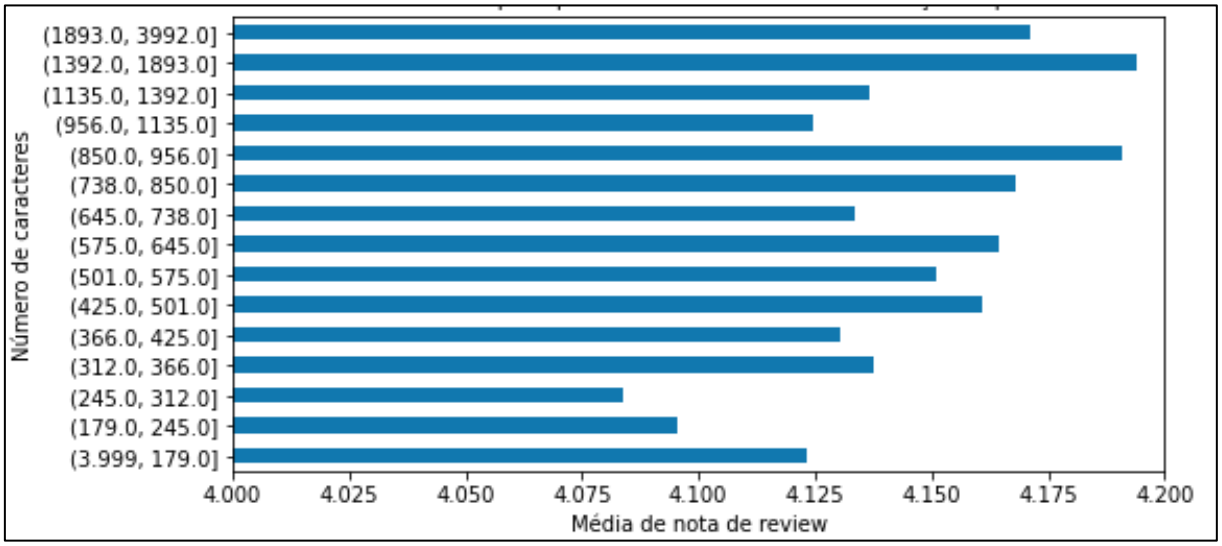

Gráfico 8: Notas de avaliação pelo número de caracteres na descrição do produto.
Fonte: Autoria própria (2020).

No gráfico da influência da descrição na nota de avaliação dada pelos consumidores (Gráfico 8), conforme a descrição do produto aumenta é percebido que as notas crescem de maneira não linear. Assim, é possível constatar que a qualidade do texto descritivo influencia na nota do consumidor. Outro quesito a ser destacado, é a variabilidade das notas, as descrições com textos longos sofrem menos variações e

tem limites inferior e superior mais altos, caracterizando menores riscos para o *e-commerce.* Supõe-se, por esse gráfico, que clientes com mais informações sobre o produto tendem a dar maiores notas de avaliação, por isso, uma descrição do produto no *site* do *e-commerce* mais longa pode afetar positivamente a relação com o cliente.

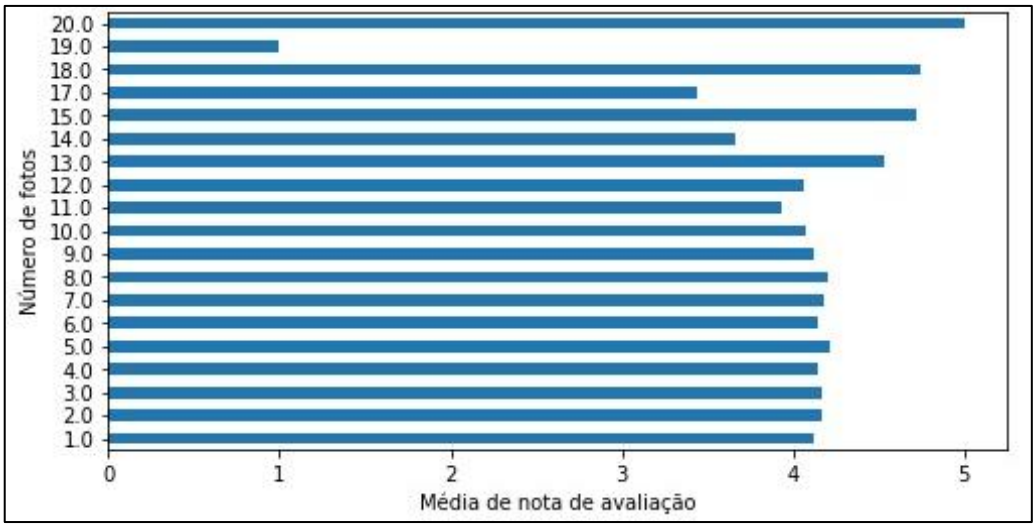

Gráfico 9: Notas de avaliação pelo número de fotos na descrição do produto.
Fonte: Autoria própria (2020).

Outro fator que caracteriza conhecimento sobre o produto são as imagens disponibilizadas no *e-commerce*. Era esperado que o número de fotos teria uma influência positiva para as notas de review, entretanto não é possível concluir através do Gráfico 9 que o aumento de imagens causou aumento na nota recebida pelo cliente. Foram investigados resultados muito anômalos da média esperada, como a intercorrência na faixa de 19 fotos. Para esse caso, a amostra de dados é estatisticamente insignificante – o que causa pouca segurança com o resultado pois induz a variável a assumir médias com valores extremos.

No Gráfico 10 é levantada a hipótese de a quantidade de itens comprados influenciar positivamente na média final da nota de review. A expectativa é haver um tratamento diferenciado para clientes com pedido alto e isso ser refletivo na nota final do pedido.

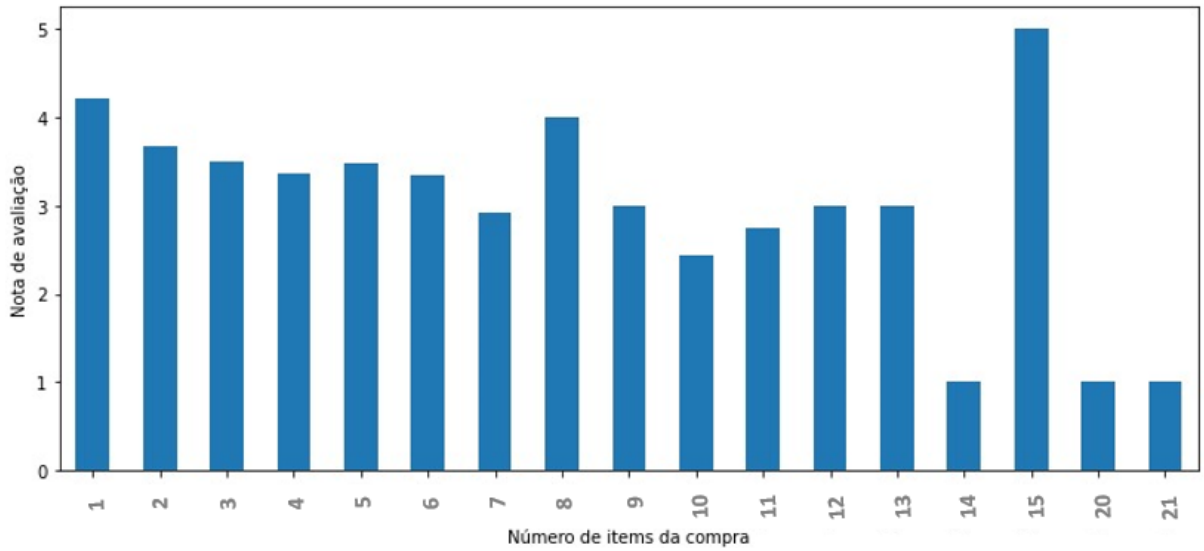

Gráfico 10: Média da nota de avaliação pela quantidade de itens comprados.
Fonte: Autoria própria (2020).

Porém, é notável uma tendência negativa, onde, com o aumento do número de produtos comprados a cresce a incidência de uma nota de avaliação mais baixa. Devido a pequena quantidade de ocorrência de casos com quatorze, ou mais, itens comprados, a análise não foi conclusiva, pois enviesaria o estudo por baixa amostragem.

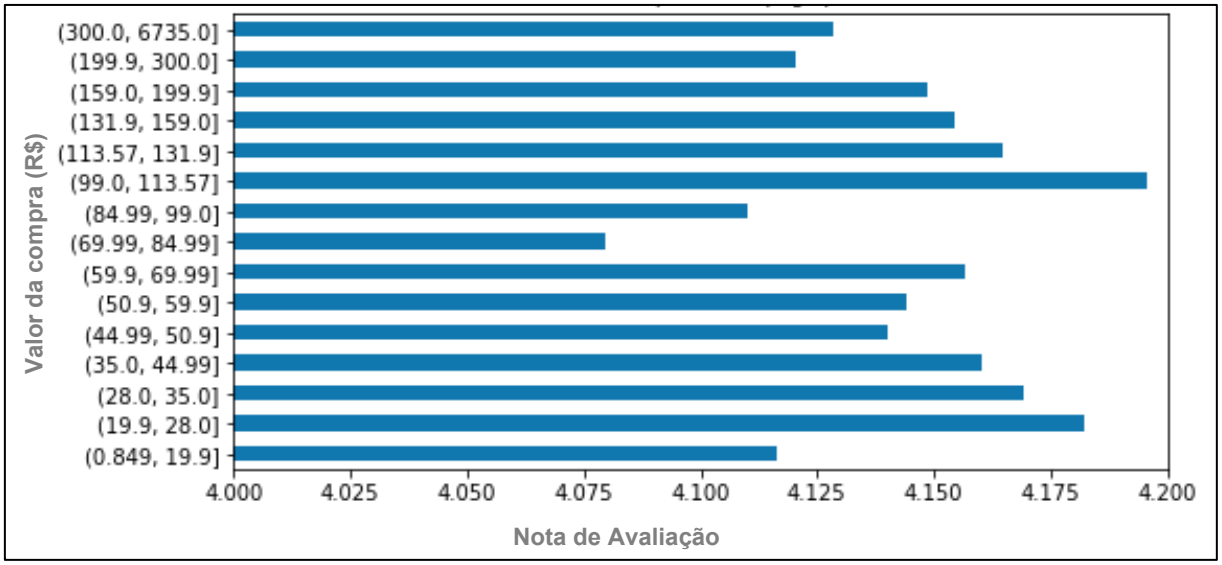

Gráfico 11: Média da nota de avaliação pela média do preço unitário.
Fonte: Autoria própria (2020).

Com o objetivo de entender como a nota de review reage aos elementos embutidos ao valor final da compra, este foi fragmentado em três elementos. Influência do preço unitário do produto, custo pago pelo frete e valor pago pela compra total

(preço unitário mais o frete). A influência desses três atributos na nota de avaliação está ilustrada nos Gráficos 11, 12 e 13. É percebido que o preço unitário não demonstra tendências em relação às avaliações dos consumidores.

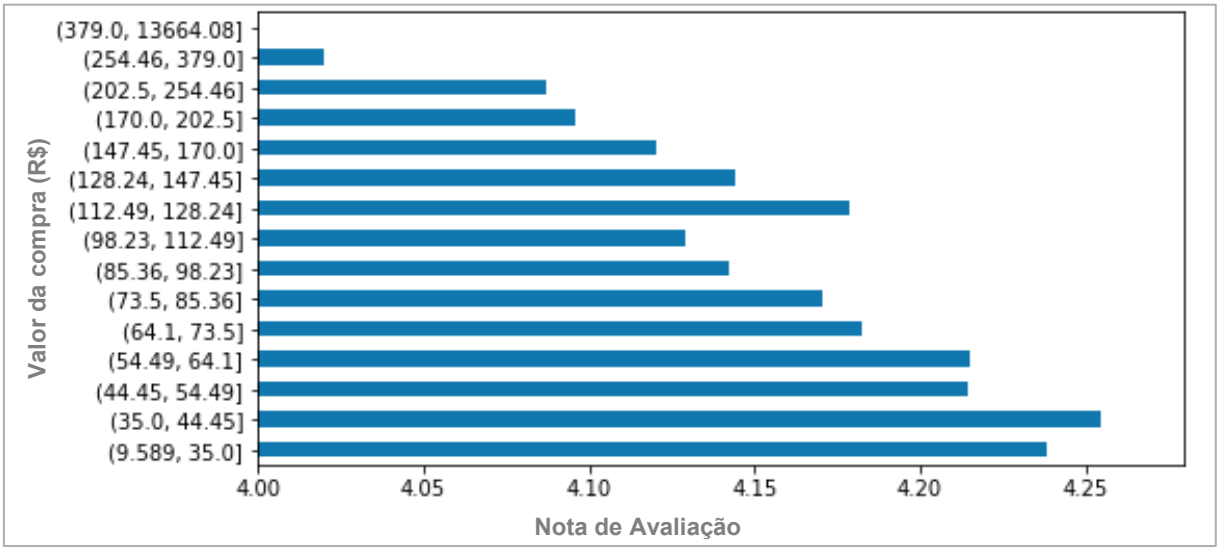

Gráfico 12: Média da nota de avaliação pelo valor médio pago pelo item mais frete.
Fonte: Autoria própria (2020).

No Gráfico 12 é possível perceber que a relação entre o valor pago pelo item mais o frete, é inversamente proporcional a nota de avaliação dada pelos consumidores, dessa forma, quanto menor for o preço total maior será a média da nota da avaliação.

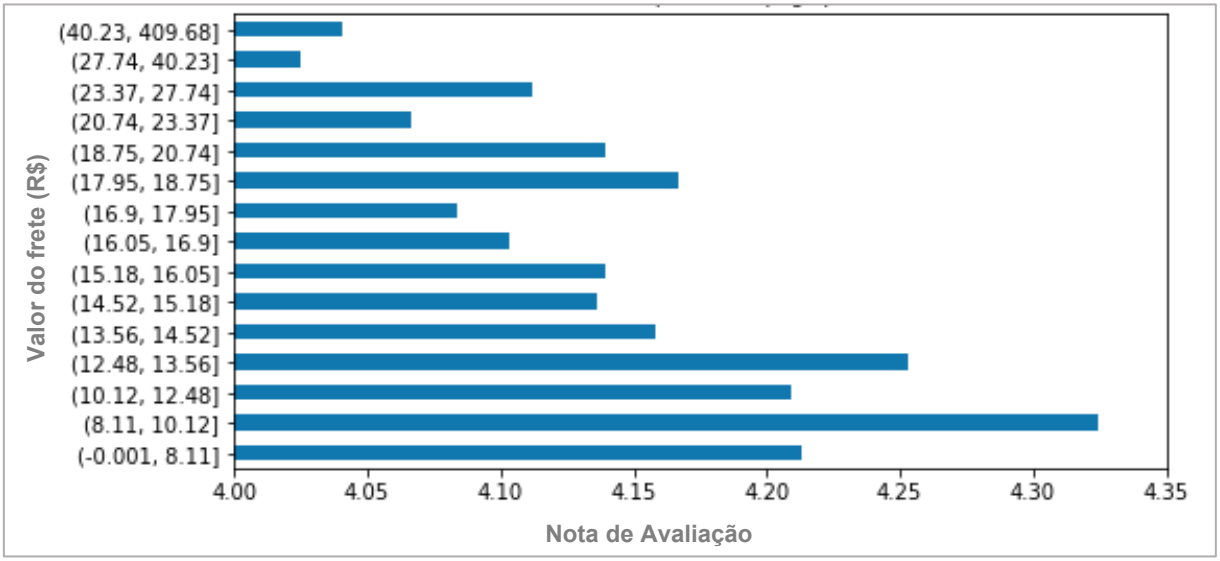

Gráfico 13: Média da nota de avaliação pelo valor médio pago pelo frete.
Fonte: Autoria própria (2020).

A partir do confronto dos Gráficos 11 e 12 foi levantada a hipótese de que o valor pago pelo frete do produto influenciaria na nota de forma tendenciosa. No Gráfico 13 é ilustrado a relação do frete e a média da nota de avaliação dada pelos consumidores. Foi percebido uma tendência similar ao Gráfico 12 apontando que o valor pago no frete efetua uma força maior do que o preço unitário do item, indicando que fatores relacionados à entrega do produto impactam mais na satisfação do cliente.

Com o objetivo de explorar mais o impacto do frete na nota de avaliação foi traçado um paralelo entre as médias das notas e a média do valor do frete por estado. Sabendo que o frete é a uma variável dependente da localização, o objetivo foi explorar o comportamento dos fatores frente a variação da localidade. O gráfico 14 está ordenado pelo frete médio por estado (em vinho), em ordem decrescente.

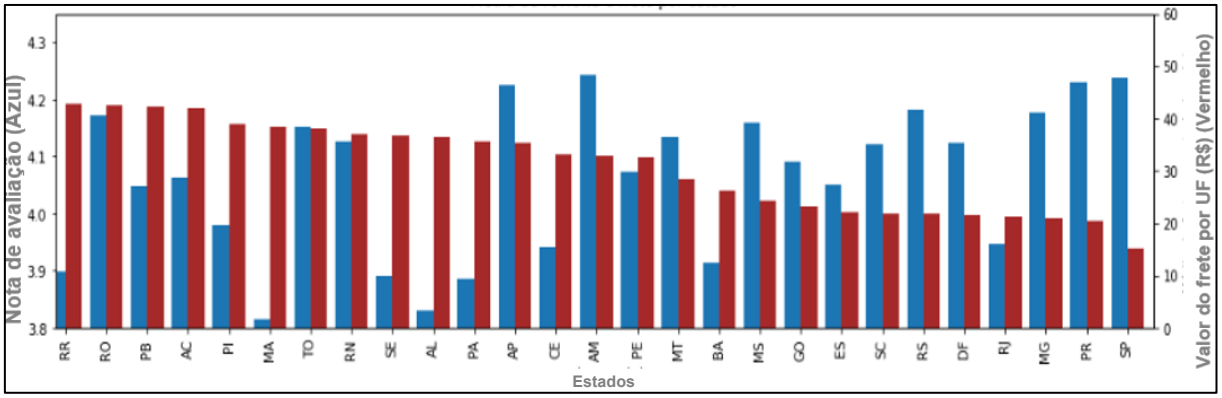

Gráfico 14: Média das notas de avaliação e valor do frete por estado
Fonte: Autoria própria (2020).

Observa-se que os estados localizados no Sul/Sudeste têm menores fretes médios e, nesses lugares, ocorreram altas médias de review. Nota-se através da análise do gráfico, que as menores médias das notas de avaliação (em azul) podem ser percebidas nas regiões Norte/Nordeste, onde, geograficamente ficam mais distantes dos centros econômicos, e apresentam valor médio de frete mais alto.

A correlação entre as variáveis não é estritamente linear; de forma que um estado com valor médio de frete entre as mais baixas, como o Rio de Janeiro, possui média baixa na nota de avaliação. Assim, a variável frete necessita uma análise mais profunda para buscar maior confiabilidade na sua relação com a variável *target,* pois esta não pode ser validada utilizando somente o gráfico.

A tabela 1, foi traçada para oferecer mais clareza na comparação das variáveis frete e valor pago com o target. A tabela está ordenada pela incidência de

pedidos no estado do *dataset*, a fim de se garantir um peso maior para incidentes com maior ocorrência.

Tabela 1: Média dos Valores dos Atributos Estudados

Estado	Quantidade	Nota média de avaliação	Valor médio pago (R$)	Valor médio de frete (R$)
SP	39940	4,2377	142,66	15,30
RJ	12167	3,948	166,83	21,19
MG	11184	4,1778	160,75	20,87
RS	5263	4,1816	161,19	21,86
PR	4854	4,2289	159,11	20,48
SC	3489	4,1218	168,25	21,86
BA	3214	3,914	181,63	26,31
DF	2050	4,1229	167,35	21,44
ES	1978	4,0516	159,66	22,16
GO	1910	4,0911	170,05	23,14
PE	1576	4,0723	193,07	32,61
CE	1265	3,9415	207,93	33,02
PA	929	3,887	218,21	35,59
MT	874	4,1344	206,17	28,44
MA	707	3,814	206,35	38,33
MS	692	4,1604	193,13	24,36
PB	509	4,0491	264,76	42,24
PI	469	3,9787	222,67	39,01
RN	468	4,1271	213,6	36,88
AL	393	3,8295	237,93	36,34
SE	333	3,8919	210,41	36,65
TO	271	4,1513	219,49	38,2
RO	238	4,1723	235,97	42,5
AM	144	4,2431	187,21	32,9
AC	78	4,0641	249,47	41,91
AP	67	4,2239	240,92	35,22
RR	39	3,8974	224,04	42,73

Fonte: Autoria própria (2020).

Ao observar o topo da tabela, onde se concentram a maior parte das vendas, é percebido que o frete assume valor médio menor. A partir dessa tabela foi possível calcular o R de Person afim de verificar o fator na qual elas se encontram correlacionadas por estado.

A correlação linear de Person mede o quanto da variável y pode ser explicada por x. É transmitida através de um valor no alcance de -1 a 1, podendo ser positivamente ou inversamente proporcional. No KDD, esta é apenas uma das métricas que se observa no conjunto de dados para buscar proximidade entre as variáveis independentes e a variável dependente.

Tabela 2: Valor do indicador R de Pearson da correlação com a variável dependente

R de Pearson	Valor médio pago	Valor frete médio de frete
Nota média de avaliação	-0,324	-0,397

Fonte: Autoria própria (2020).

As expectativas geradas pelo gráfico foram confirmadas ao verificar que o valor pago tem R^2 = -0,32 e o frete tem R^2 = -0,39, valores considerados altos para correlação.

As demais hipóteses que não puderam ser verificadas pelos atributos existentes foram investigadas na seção *Feature Engineering* de forma que o desenvolvimento dessa análise fosse modularizado e seus resultados pudessem ser incluídos no banco de dados. O processo do KDD é descrito como continuado e iterativo, sendo assim os escopos de cada etapa podem se sobrepor à medida que surjam novas descobertas. Esse é um dos comportamentos que garante o sucesso do processo.

4.2 4.2 *FEATURE ENGINEERING*

Essa etapa foi desenvolvida através de uma combinação entre experimentação e leitura subjetiva do problema acreditando conseguir aumentar o poder explicativo do conjunto de dados bruto. Abaixo estão listadas as variáveis criadas durante o processo.

Delivery accuracy (Exatidão de entrega) – valor, em dias, calculado a partir da diferença entre a data prevista para a entrega e a data que foi entregue. Esse atributo supõe que a relação com o cliente é beneficiada se a entrega obedecer ao prazo estimado durante a compra, independentemente da quantidade de dias prometida.

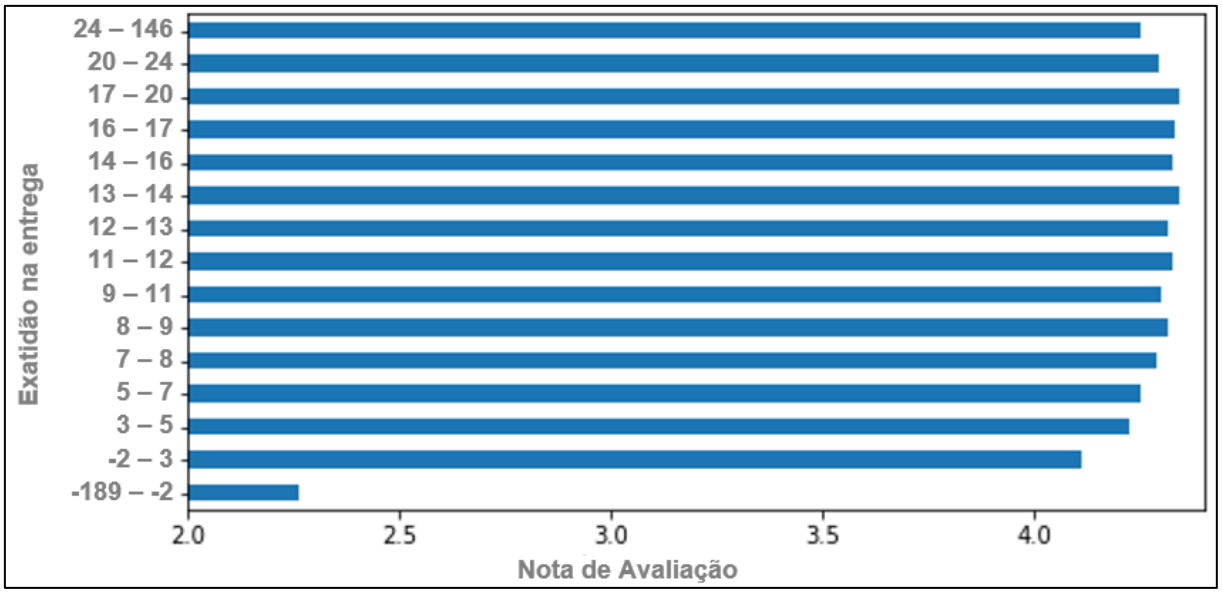

Gráfico 15: Média das notas de avaliação pela exatidão na entrega.
Fonte: Autoria própria (2020).

Os valores negativos indicam as vendas nas quais a data de entrega foi superior a data prevista. Através do gráfico é possível perceber o quanto isso tem um impacto negativo para a nota final pois representa baixa qualidade de serviço frente a expectativa do e-consumidor.

Days to deliver (dias para entrega) – valor, em dias, da previsão de dias para o produto chegar. Diferente do *Delivery accuracy*, esse atributo supõe uma relação inversa entre a nota final de *review* e a quantidade de dias líquidos de espera suposto pelo vendedor.

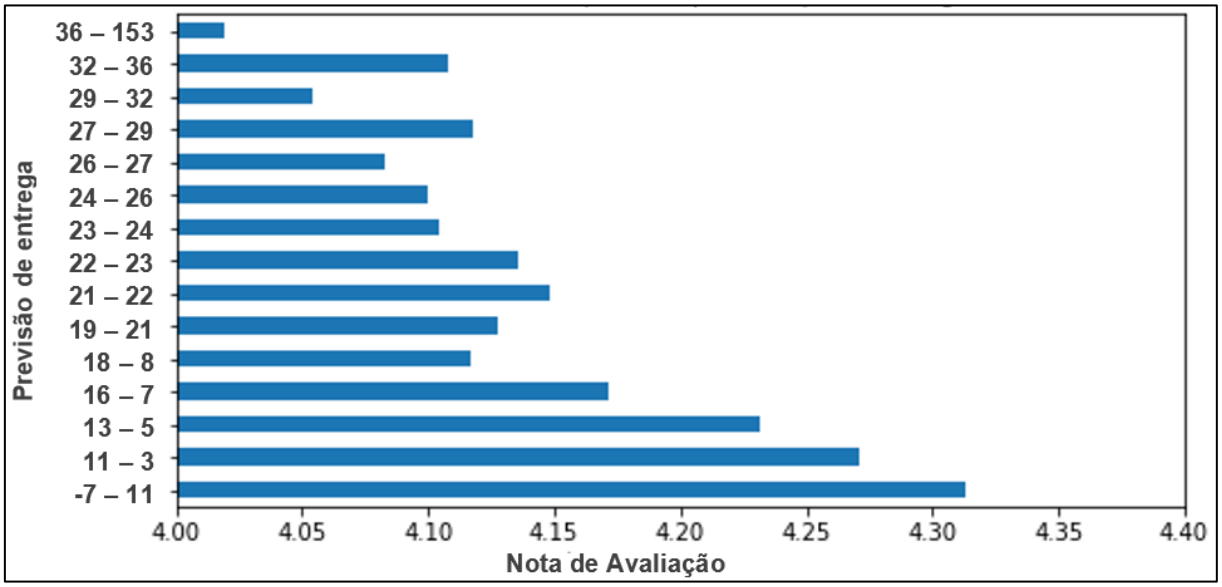

Gráfico 16: Média das notas de avaliação pelos dias previstos para a entrega.
Fonte: Autoria própria (2020).

A hipótese levantada é que quanto menor os dias previstos para a entrega, maior será a média da nota final para o grupo. No gráfico 16 é ilustrado os valores com o corte da média das notas acima de 4. Claramente o gráfico tem uma tendência que suporta essa hipótese.

Delivered before time (Entrega antes do tempo) – variável booleana, do tipo verdadeiro/falso, que complementa a informação de *Delivery accuracy,* ao informar positivo – se foi entregue antes do tempo, ou se negativo – houve atraso na entrega.

No Gráfico 17, os dados foram divididos em dois grupos, aqueles pedidos que foram entregues antes do tempo e os que foram entregues depois do tempo. A entrega se mostrou um fator que diferenciador da opinião do cliente durante sua experiência *online*. Existe uma diferença de média de quase dois pontos entre os grupos que corroboram com a ideia de cumprir o prazo está entre o topo das prioridades do cliente ao avaliar uma compra.

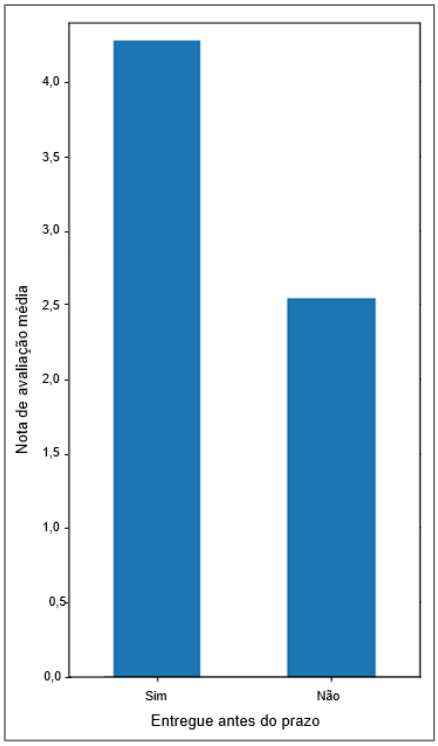

Gráfico 17: Média das notas de avaliação para se entregue no prazo ou não.
Fonte: Autoria própria (2020).

Risk Value (Valor de risco) – variável categórica que separa os valores da variável de preço em 4 percentis de forma a classificar os produtos de menor risco de investimento ao maior risco.

A ideia dessa variável é captar diferenças inerente no serviço geral que possam ser refletidas na diferença de preço e em maior qualidade que gerem maior satisfação ao cliente. Não foi possível distinguir no gráfico 18 o impacto de cada uma das decisões de compra. Essa pode não ser uma variável de muito impacto e, a partir da construção do atributo, poderemos verificar durante os resultados da mineração de dados se isso se confirmou.

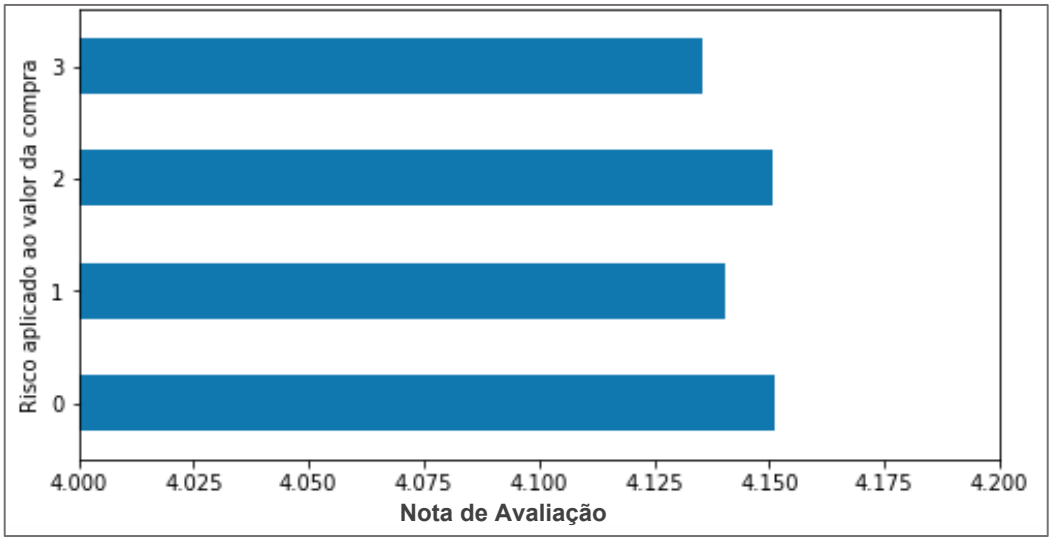

Gráfico 18: Média das notas de avaliação pelo risco envolvido na compra.
Fonte: Autoria própria (2020).

Variáveis estatísticas (mínimo, máximo, mediana, média, desvio padrão) – Outra prática comum nessa etapa do KDD é criar variáveis que mostrem estatísticas daquele grupo. Nesse caso, agrupou-se o *dataset* pelo estado do comprador, *status* da compra e categoria do produto e tirou as métricas estatísticas das notas para cada uma dessas agregações.

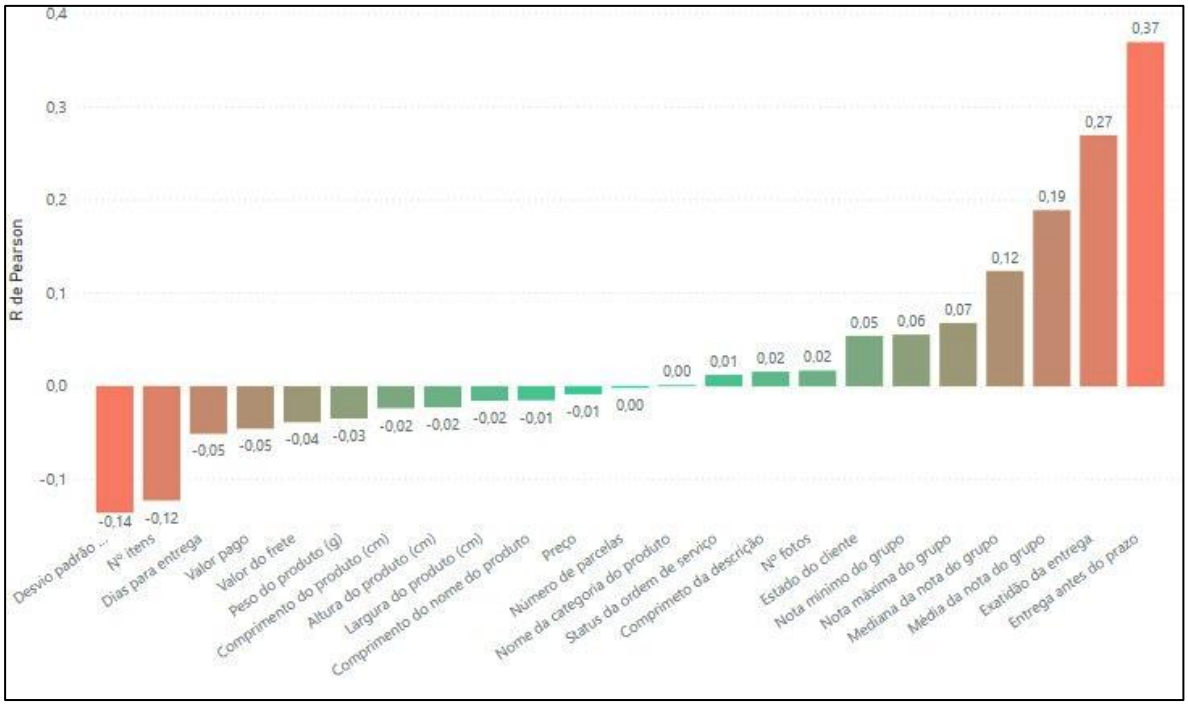

Gráfico 19: Correlação linear entre os atributos e a nota de avaliação.
Fonte: Autoria própria (2020).

O Gráfico 19 foi produzido com o conjunto de dados após a realização da etapa de transformação de dados e ilustra os índices de correlação linear de cada um dos atributos em função da variável dependente nota de *review* através do indicador R de *Pearson*.

Na Tabela 3 são mostrados os atributos com maior correlação linear positiva com a nota de avaliação e seu respectivo R de Pearson e as variáveis que apresentam maiores correlações negativas. O resultado aponta o quanto foi produtiva a etapa de transformação de dados uma vez que as maiores correlações, em módulo, foram variáveis criadas durante a etapa de *featuring engineering*. Além disso, indica que o banco de dados tem mais chance de sucesso na etapa de mineração de dados.

Tabela 3: Maiores Correlações Entre Atributos e a Variável "*review score*"

Atributo	R de Pearson
“delivered_before_time”	0,37
“delivery_accuracy”	0,27
“group_mean”	0,19
“group_max”	0,13
“group_min”	0,06
“group_std”	-0,13
“days_to_deliver”	-0,13

Fonte: Autoria própria (2020).

Para lidar com dados categóricos em um BD existem várias práticas que podem contribuir com a assertividade do modelo, porém para esse conjunto de dados foi utilizado a transformação numérica das variáveis categóricas chamada *Label Encoding.* Ela não oferece muita explicação porque trata-se apenas de criar variáveis numéricas correspondentes as variáveis categóricas, na mesma dimensão. Porém, essa foi uma decisão tomada baseada em dois princípios: (i) as variáveis categóricas eram muitas e nesse caso adicionar-se-ia uma complexidade muito maior aos dados com pouco ganho, (ii) mesmo assim, foi feito o proposto uma vez que um dos modelos matemáticos (regressão linear) não é tolerante a variáveis alfabéticas no seu algoritmo.

Por fim, algumas colunas que fizeram parte das etapas anteriores são retiradas do *dataframe* final por compatibilidade com o modelo matemático como:

- As colunas no formato de data (order_purchase_timestamp, order_approved_at,order_estimated_delivery_date,order_delivered_customer _date, shipping_limit_date, order_delivered_carrier_date) – que já estão representadas pelas colunas criadas na *Feature Engineering*;
- As colunas de identificação (*customer_id, customer_unique_id, order_id, order_item_id, product_id, seller_id*) por apresentaram grande cardinalidade e nenhum poder explicativo;
- As colunas de cidade – uma vez que existe a localização menos granular de estado e lidar com mais de 7 mil cidades aumentaria a complexidade do modelo e não a sua performance.

4.3 4.3 MINERAÇÃO DE DADOS

Nesse estudo foram desenvolvidos três algoritmos matemáticos para serem comparados: Regressão Linear, *Random Forest* e *Gradient Boosting*. O primeiro passo da mineração é submeter o conjunto de dados separadamente em cada um dos algoritmos e medir qual resultado apresenta o menor Erro Médio Quadrático (MSE em inglês). O MSE pode ser calculado a partir da seguinte equação:

$$E = \frac{1}{N}\sum_{i=1}^{n}(\widehat{\theta_i} - \theta_i)^2$$

em que:

- $\hat{\theta}$ representa todos os palpites de *target* (*review_score*) que cada um dos algoritmos supões ao fim do treinamento;
- θ representa todos os valores de *target* esperados;
- N representa a quantidade de *datapoints*, ou linhas, que existem no conjunto de dados;
- *E* representa o erro somado de todos as linhas calculado a partir da diferença entre o target e o que realmente foi atingido, elevado ao quadrado para capturar erros negativos e positivos, somados linha e linha e divididos pela quantidade de linhas.

O valor de MSE captura a performance geral de cada um dos algoritmos e funciona para decidir qual deles foi mais preciso e deve ser persistido para basear as nossas previsões.

Após a seleção do algoritmo, acontecem as etapas inerentes ao processo de mineração de dados.

a) *Fit* – onde o sistema é alimentado com os dados separados para treino (tanto as variáveis independentes como a variável dependente) e todo o processamento é desenvolvido de forma a calcular o *target,* através de algoritmos de busca, possuindo um conjunto de dados e uma coluna de *target* correspondente. Nessa etapa ainda não é devolvido nenhum resultado, porém, os parâmetros utilizados pelo algoritmo para calcular os resultados são gravados nesse método e podem ser usados na etapa posterior para calcular o *target* para qualquer outro conjunto de dados que lhe for entregue;

b) *Predict* – é a etapa onde o algoritmo, munido dos parâmetros encontrados na etapa anterior, consegue calcular os valores de *target* sem mais ter um parâmetro para se basear. Dessa forma apresenta-se um conjunto de dados diferente, (normalmente o conjunto de teste, apenas com as variáveis independentes) e o

algoritmo é capaz de devolver valores de *target* que nunca teve acesso.

Existem diferentes resultados que podem ser alcançados a partir da mineração de dados, face os objetivos do projeto. O produto do presente estudo é saber quais foram os atributos que mais contribuíram para o modelo alcançar o valor da nota de *review* correspondente.

Na etapa de transformação de dados foi apresentada a correlação de Pearson como um dos fatores que pode medir a contribuição de cada uma das variáveis para o mapeamento do valor esperado. Porém, na realidade, cada um dos algoritmos desenvolve dentro dele um tipo de "Importância do Atributo" (*feature importance)* que considera outros fatores além da correlação linear em razão da interação entre as variáveis durante a aplicação dos cálculos não-lineares.

Essa métrica, implementada pela biblioteca *Sci-kit Learn,* baseia-se na *Gini Importance* que é capaz de traduzir quantitativamente a contribuição individual das variáveis independentes para o algoritmo de decisão. O resultado para cada um deles é calculado pela somatória da quantidade de *splits* que incluíam esse atributo proporcionalmente ao número de *splits* realizado em todas as amostras.

Após aplicadas as cinco etapas do KDD, o modelo na qual atingiu o menor erro foi o *Random Forest.* Ele teve um resultado superior ao *Gradient Boosting* e a Regressão Linear, conforme ilustrado na Tabela 4.

Tabela 4: MSE dos três modelos após o treino comparativo

Algoritmo	MSE
Random Forest	1,3016
Gradient Boosting	1,3091
Regressão Linear	1,3648

Fonte: Autoria própria (2020).

Os parâmetros definidos antes do início do treinamento do modelo que puderam determinar como seria feita a mineração estão expressos da tabela 5.

Tabela 5: Parâmetros Utilizados no Modelo Random Forest

Parâmetros	Valores
Máxima profundidade	15
Máxima *features* em cada separação	8
Mínima quantidade de *datapoints* necessários para dividir um nó	80
Número de árvores numa floresta	60
Número de processadores	4

Fonte: Autoria própria (2020).

Esses valores não são definidos por nenhuma literatura e são específicos de cada caso. Eles foram encontrados empiricamente, partindo dos valores *default* e calibrando para o menor cenário de erro.

Finalmente, o Gráfico 20 retorna todos os atributos utilizados durante o treino e teste do modelo por ordem de importância, considerando o *Gini Importance*. Diferentemente do método de correlação linear, essa métrica não apresenta valores negativos e mede o impacto apenas no espectro positivo, não importando se a variável era positivamente ou negativamente impactante.

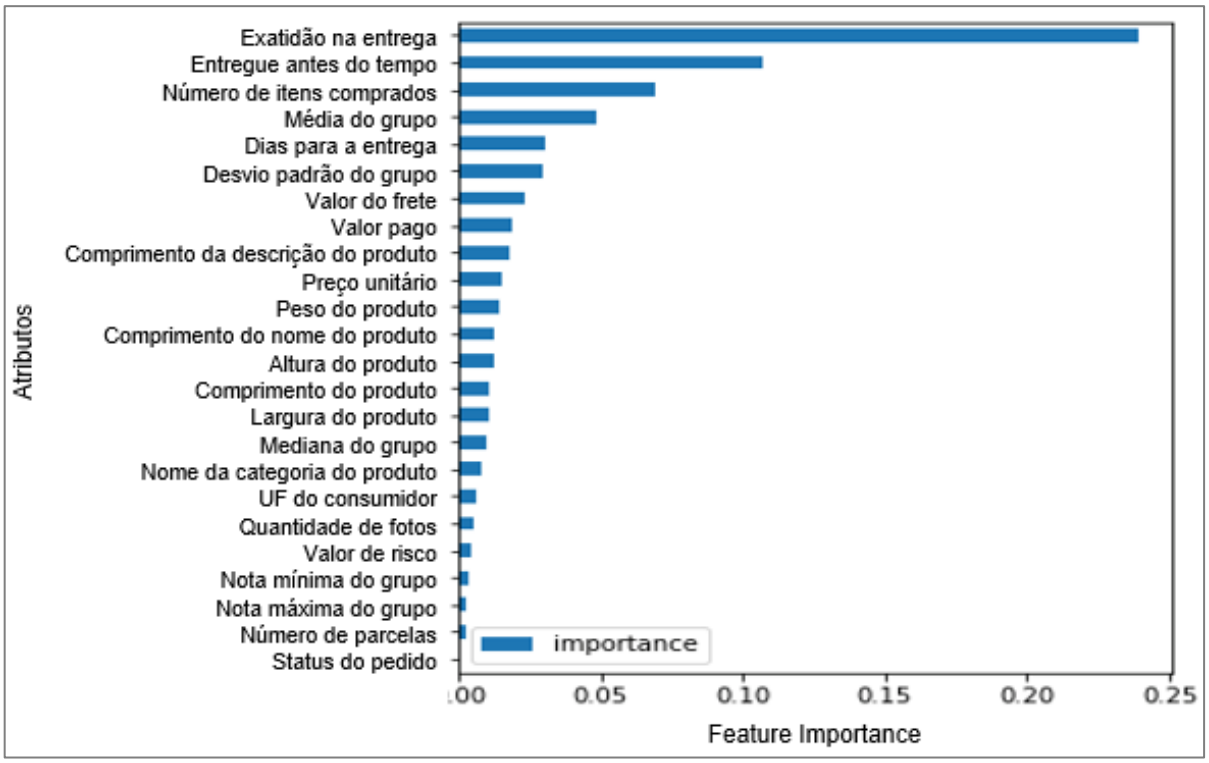

Gráfico 20: Feature Importance do Random Forest em ordem de importância dos atributos.
Fonte: Autoria própria (2020).

Algumas das previsões que foram propostas durante a transformação de dados foram confirmadas; como as variáveis Exatidão da Entrega (*Delivery Accuracy*)

e Entregue Antes do Tempo (*Delivered Before Time*) ocupando as primeiras posições do gráfico. Conclui-se a partir desse resultado o valor que a logística possui na aferição da qualidade do serviço de uma compra pela internet.

Um ponto que vale a pena ser destacado é a entrega ser efetivada dentro do prazo previsto tem mais valor para o cliente do que o produto chegar antes do prazo. Isso revela uma característica importante da análise de dados. A estima do comprador na confiança daquele serviço é alimentada por ações que abonem o caráter do vendedor acima das ações nas quais o comprador ganharia alguma vantagem.

Para a estratégia de negócio, esse resultado pode ser utilizado para calibrar as negociações que exponham a empresa a posições competitivas sem ter que sacrificar os processos previstos. A variável na quinta posição, “Dias para Entrega”, é análoga as variáveis logísticas nas duas primeiras colocações, podendo ser extraídas as mesmas conclusões.

O próximo destaque do gráfico é a variável “Número de Itens” e o seu peso no momento de quantificar a nota para o serviço. Durante a exploração, já havia sido previamente estabelecida uma relação negativa entre a quantidade de itens numa compra e a nota atribuída àquela transação. Ou seja, quanto mais itens havia numa compra, pior tinha sido a experiência do cliente, medida através da variável “*review score*”.

Dessa forma, esse resultado expõe um ponto na qual a plataforma *Olist* é capaz de se desenvolver. É possível que esta possa atrair uma clientela que compre em maior quantidade se houver investimento em políticas que beneficiem o comprador a atacado revelando um mercado a ser explorado e sustentando a necessidade de analisar os resultados da companhia.

Duas variáveis com pouca explicação prática, “*group_mean*” e “*group_std*” estão entre as seis primeiras colocações em importância. Essas foram acrescidas durante a transformação de dados como uma abstração matemática de forma a aumentar a quantidade de explicação existente numa linha angariando informações daquele grupo (categoria de produto, estado e situação da venda). Por se tratar de uma prática muito comum e recomendada do processo de análise de dados através de *Machine Learning*, esse resultado é visto como um sucesso do projeto, mesmo sem agregar muito na visão comercial de um negócio.

As próximas variáveis no *ranking* possuem um peso consideravelmente inferior na decisão da nota por parte do cliente diante dos outros atributos analisados

anteriormente. Pela natureza da métrica *Gini Importance*, é possível ter um peso para cada uma das variáveis compostas no banco de dados e o gráfico 20 expõe a dimensão da ordem de grandeza de cada fator.

As variáveis relativas ao preço (valor do frete, valor do pagamento final e preço unitário) ocupam respectivamente a 7, 8 e 10 posição, seguidas pelas variáveis de descrição do produto no *marketplace* (tamanho da descrição do produto (9), tamanho do nome do produto (12)) e, após isso, as variáveis intrínsecas do produto (tamanho do produto em gramas (11), altura do produto em centímetros (13), comprimento do produto em centímetros (14), largura do produto em centímetros (15)).

Este resultado indica uma realidade particular desse banco de dados. Para o modelo utilizado, não foram utilizadas nenhuma característica subjetiva – todas os atributos são quantitativos. Não foram explorados os comentários dos clientes ou dados não estruturados que pudessem contribuir com outra explicação.

Conclui-se que, de acordo com essa configuração específica, os atributos referentes à entrega influenciam mais no resultado do que aqueles referentes a descrição do anúncio e da própria natureza do produto. Ou seja, para o e-consumidor, a experiência da compra está mais relacionada com um frete favorável do que um anúncio de venda explicativo. Resguarda-se a observação que todos esses atributos não contribuem com um impacto tão relevante se olhados separadamente.

Por fim, vale ressaltar que todos os atributos subsequentes tem duas características importantes: (i) todas elas cabem no espectro de correlação linear entre -0,1 e 0,1 ou seja, baixíssimo poder explicativo (ii) são variáveis que possuem baixa variabilidade na distribuição: situação do pedido (duas possibilidades), quantidade de parcelas no pagamento (seis possibilidades), valor de risco (quatro possibilidades), nota mínima (maioria na faixa de 1), nota máxima (maioria entre 4 e 5). Como a variabilidade não destoa durante a distribuição, não há muito como aprender as nuances dos resultados. Nenhuma conclusão pode ser tirada acerca dessas hipóteses.

5 CONCLUSÃO

Esse estudo foi desenvolvido de forma a contribuir para a discussão que permeia uma ampla gama de campos dentro do *Marketing*, Operação, Logística e Gestão com um olhar sistêmico sobre a satisfação do cliente e desenvolvendo um método previsto na literatura que pode ser replicado afim de obter os mesmos *insights* para outras situações.

Durante a EDA e a *Feature Engineering* foram levantadas hipóteses de atributos como: número de caracteres na descrição do produto, número de fotos do produto, quantidade de itens comprados, preço unitário do produto, valor pago no frete, prazo estimado, precisão na data de entrega e entrega realizada antes do prazo, que poderiam exercer uma forte influência na relação entre o consumidor e os vendedores que utilizam o *marketplace* Olist.

Nos resultados apresentados, verificou-se que os fatores que englobam a operação logística; como exatidão na entrega, entrega antes do tempo, número de itens, e fatores geográficos e econômicos, como valor do frete e preço de pagamento, estão entre as maiores prioridades do cliente ao avaliar a sua experiência de compra.

Para futuros trabalhos cita-se a possibilidade de usar esse banco de dados para diferentes estudos como análise de comentários das ordens de serviços através de Processamento de Linguagem Natural (NLP em inglês) para identificar novos fatores relacionados a satisfação do cliente e identificar com mais precisão os resultados obtidos no presente estudo. Além deste, é possível expandir a pesquisa através da captação de banco de dados relacionados à *Design* da interface do *site* e assim aprimorar a experiência do cliente na usabilidade do *e-commerce.*

REFERÊNCIAS

ABERNETHY, Michael. **Mineração de dados com WEKA**, Parte 1: Introdução e regressão. 2010. Disponível em: <https://www.ibm.com/developerworks/br/opensource/library/os-weka1/>. Acesso em: 26 de abril de 2019.

ALBERTIN, Alberto Luiz. **Comércio Eletrônico: Modelo, aspectos e contribuições de sua aplicação**. 6. ed. São Paulo: Atlas, 2010.

ALMEIDA, Raimunda Eunice da Silva; BRENDLE, Vivian; SPINOLA, Noelio Dantaslé. **E- commerce: Evolução, processo de compra, e o desafio da entrega**. RDE: Revista de Desenvolvimento Econômico, Salvador, v. 16, n. 29, p.138-149, 2014.

ALVES, Luiz. **Vencendo na economia digital**. São Paulo: Makron Books, 2002.

ANDERSON, Eugene W.; FORNELL, Claes; LEHMANN, Donald R. Customer Satisfaction, Market Share, and Profitability: Findings from Sweden. **Journal Of Marketing**, [s.l.], v. 58, n. 3, p.53-66, jul. 1994. SAGE Publications. http://dx.doi.org/10.1177/002224299405800304.

BEAL, Adriana. **Gestão Estratégica da Informação:** Como Transformar a Informação e a Tecnologia da Informação em Fatores de Crescimento e de Alto Desempenho nas Organizações. 1. ed. São Paulo: Atlas, 2004.

BROWN, Martin. **Técnica de mineração de dados**. 2012. Disponível em: <https://www.ibm.com/developerworks/br/library/tecnicas-mineracao-de-dados/index.html> Acesso em: 03 de maio de 2019.

CARVALHO, Luís A. **Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração**. 2 ed. São Paulo: Ciência Moderna, 2002.

CORREA, Henrique Luiz; GIANESI, Irineu Gustavo Nogueira. **Administração estratégica de serviços**: operações para a satisfação do cliente. 2. ed. São Paulo: Atlas, 2019.

DAVENPORT, Thomas. H.; PRUSAK, Laurence. **Conhecimento empresarial**: como as organizações gerenciam o seu capital intelectual. 4. ed. Tradução de Lenke Peres. Rio de Janeiro: Campus, 1998. 237 p.

E-BIT (Brasil). **Webshoppers**. 39. ed. [s.i]: Nielsen, 2019. 41 p. Disponível em: <https://www.ebit.com.br/webshoppers>. Acesso em: 28 maio 2019.

EMC. ***The Digital Universe*** **in 2020**: *Big Data, Bigger Digital Shadows, Biggest Growth in theFarEast*. EMC. [Online] 2012. Acesso em: Abril 2019. Disponível em: http://www.emc.com/leadership/digital-universe/index.htm.

ESTEVES, Yohans de Oliveira. **Marketing, Internet e o Comportamento do E-Consumidor**. In: CONGRESSO NACIONAL DE EXCELÊNCIA EM GESTÃO, 7., 2011, Rio de Janeiro. Anais... . Rio de Janeiro: CNEG, 2011. p. 1 - 17. Disponível em:

<http://www.excelenciaemgestao.org/Portals/2/documents/cneg7/anais/T11_0393_185.pd>. Acesso em: 28 maio 2019.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. ***Advances in Knowlwdge Discovery and Data Mining***. Menlo Park, EUA: AAAI Press, 1996. 611 p

_____. PIATETSKY-SHAPIRO, Gregory; SMITH, Padhraic. *From Data Mining to Knowledge Discovery in Databases.* ***Ai Magazine***, Palo Alto, v. 17, n. 3, p.3-21, nov. 1996.

FONSECA, João José Saraiva da. **Metodologia da pesquisa científica**. Fortaleza: Eduece, 2002.

GERHARDT, Tatiana Engel. SILVEIRA, Denise Tolfo (Org.). Universidade Aberta do Brasil (Coord.). **Método de pesquisa**. Porto Alegre: Editora da UFRGS, 2009.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisas**. 5. ed. São Paulo: Atlas, 2010.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel Lopes. **Data mining: um guia prático**: conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. ***Data Mining: Concepts and Techniques***. 3. ed. Waltham: Elsevier, 2012.

HERNANDEZ, José. Mauro. **Entendendo melhor o processo de decisão de compra na internet: uma análise sobre o papel da confiança em diferentes**. In: Encontro Nacional da ANPAD, 2002, Salvador. Anais... Salvador, 2002.

KAUARK, Fabiana da Silva; MANHÃES, Fernanda Castro; MEDEIROS, Carlos Henrique. **Metodologia da pesquisa**: Um guia prático. Itabuna: Via Litterarum, 2010.

KOTLER, Philip, HAYES, Thomas, BLOOM, N. Paul. **Marketing de Serviços Profissionais** – 2a Ed. Barueri: Manole, 2002.

_____; KELLER, Kevin Lane. **Administração de Marketing**. Tradução de Sônia Midori Yamatto. 14. ed. São Paulo: Pearson, 2012.

LOVELOCK, Christopher; WRIGHT, Lauren. **Serviços: Marketing e Gestão**. São Paulo: Saraiva, 2001.

MAIMON, Oded; ROKACH, Lior (Ed.). ***Data Mining and Knowledge Discovery Handbook***. 2. ed. New York: Springer, 2010.

MANSANO, Adriana Toledo Rodrigues; GORNI, Patrícia Monteiro. SATISFAÇÃO DO CONSUMIDOR COM O COMÉRCIO ELETRÔNICO: ESTUDO DE CASO DE UMA FABRICANTE DE TAPETES. **Revista de Extensão e Iniciação Científica SOCIESC - REIS**, Santa Catarina, v. 1, n. 1, p.12-22, jun. 2014.

MANUAL de gestão de serviços de informação. Curitiba: TECPAR/ Brasília: IBICT, 1997. 257 p.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos da metodologia da científica**. 7. ed. São Paulo: Atlas, 2010.

McGEE, James; PRUSAK, Laurence. **Gerenciamento estratégico da informação**. 21. ed. Rio de Janeiro: Campus, 1994.

MORAIS, Felipe. **Quem são os e-consumidores?**. 2011. Disponível em: <http://imasters.com.br/artigo/20096/dotnet/quem-sao-os-e-consumidores/>. Acesso em: 28 maio 2019.

OLIVER, Richard. ***Satisfaction: A behavioral perspective on the consumer***. New York: Irwin/mcgraw-hill, 1997.

OLIVEIRA, Maxwell Ferreira de. **Metodologia científica: um manual para a realização de pesquisa em Administração**. Catalão: UFG, 2011.

PARASURAMAN, A.; ZEITHAML, Valarie A.; MALHOTRA, Arvind. *E-S-QUAL: A Multiple-Item Scale for Assessing Electronic Service Quality*. ***Journal Of Service Research***, [s.l.], v. 7, n. 3, p.213-233, fev. 2005. SAGE Publications. http://dx.doi.org/10.1177/1094670504271156.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do trabalho científico**: métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: Feevale, 2013.

SALOMÃO, Flávio. **A evolução e as tendências para os conteúdos no e-commerce**. 2018. Disponível em: <https://www.ecommercebrasil.com.br/artigos/evolucao-conteudos-no-e-commerce/>. Acesso em: 30 maio 2019.

SANTOS, Rafael. **Conceitos de Mineração de Dados na Web**. 2009. Disponível em: <http://www.lac.inpe.br/~rafael.santos/Docs/WebMedia/2009/webmedia2009.pdf>. Acesso em: 28 de abril de 2019.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao datamining**: mineração de dados. 1. ed. Ciência Moderna, 2009.

TORRES, Norberto. **Principais fatores de sucesso para o varejo online**. 2013. Disponível em: <https://www.ecommercebrasil.com.br/artigos/principais-fatores-de-sucesso-para-o- varejo-online/>. Acesso em: 04 junho 2019

TURCHI, Sandra R. **Estratégias de Marketing Digital e E-commerce**. São Paulo: Atlas S.a., 2012.

WEISS, S.; INDURKHYA, N. ***Predictive data mining****: a practical guide.* San Francisco: Morgan Kaufmann Publishers, 1998.

WITTEN, I. H.; FRANK, E. Data Mining: ***Practical Machine Learning Tools and Techniques***. 3 ed. Burlington, MA: Elsevier, 2011.

WOLFINBARGER, Mary; GILLY, Mary C. *ETailQ: dimensionalizing, measuring and predicting etail quality*. **Journal Of Retailing**, [s.l.], v. 79, n. 3, p.183-198, jan. 2003. Elsevier BV. http://dx.doi.org/10.1016/s0022-4359(03)00034-4.

Yoo, Boonghee, & Donthu, Naveen. (2001). *Developing a scale to measure the perceived quality of an Internet shopping site (SITEQUAL).* ***Quarterly Journal of Electronic Commerce***, 2(1), 31–46.

Zeithaml, Valarie., Parasuraman, Ananthanarayanan, Berry, Leonard., ***Delivering Service Quality****: Balancing Customer Perceptions and Expectations*. New York: Free Press, 1990.

ZEITHAML, Valarie; BITNER, Mary Jo; GREMLER, Dwayne. **Marketing de serviços**: A empresa com foco no cliente. 6. ed. São Paulo: AMGH, 2014.

APÊNDICES

APÊNDICE A – CÓDIGO FONTE

```
import pandas as pd
import numpy as np
import datetime
import seaborn as sns
from IPython import display
import matplotlib.pyplot as plt
from IPython.core.pylabtools import figsize

from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import train_test_split
from sklearn.utils import shuffle
from sklearn.metrics import mean_squared_error
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import LabelEncoder
from sklearn.pipeline import make_pipeline
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler

pd.set_option('display.max_columns', 40)
pd.options.mode.chained_assignment = None

data_file = 'data_1.csv'

city_cols = ['customer_zip_code_prefix','customer_city']

cat_cols = ['customer_state',
'order_status',
'product_category_name']

date_cols = ['order_purchase_timestamp',
 'order_approved_at',
 'order_estimated_delivery_date',
 'order_delivered_customer_date',
 'shipping_limit_date',
 'order_delivered_carrier_date']

target = 'review_score'

id_cols = ['customer_id', 'customer_unique_id', 'order_id', 'order_item_id',
 'product_id', 'seller_id']

num_cols = ['price',
 'freight_value',
 'no_items',
 'product_name_lenght',
 'product_description_lenght',
```

```
 'product_photos_qty',
 'product_weight_g',
 'product_length_cm',
 'product_height_cm',
 'product_width_cm',
 'payment_sequential',
 'payment_value',
 'days_to_deliver',
 'delivery_accuracy',
 'delivered_before_time']

class Data:
def __init__(self, data_file, target, cat_cols, date_cols, num_cols, id_cols, city_cols):
self.cat_cols = cat_cols
self.date_cols = date_cols
self.num_cols = num_cols
self.city_cols = city_cols
self.feature_cols = self.cat_cols + self.num_cols
self.target = target
self.label_encoders = {}
self.train_df, self.test_df = self._preprocessing(data_file)
#display(self.train_df)
#display(self.test_df)

def label_encode_df(self, df, cols):
'''creates one label encoder for each column in the data object instance'''
for col in cols:
if col in self.label_encoders:
#if label encoder already exits for col, use it
self._label_encode(df, col, self.label_encoders[col])
else:
self._label_encode(df, col)

def _label_encode(self, df, col, le=None):
'''label encodes data'''
if le:
df[col] = le.transform(df[col])
else:
le = LabelEncoder()
le.fit(df[col])
df[col] = le.transform(df[col])
self.label_encoders[col] = le

def inverse_encode_df(self, df, cols):
'''does inverse label encoding'''
for col in cols:
if col in self.label_encoders:
self._inverse_label_encode(df, col)
else:
```

```
raise ValueError("label_encoders must be define for each col before calling
inverse_encode_df")

def _inverse_label_encode(self, df, col):
'''inverse label encodes data'''
le = self.label_encoders[col]
df[col] = le.inverse_transform(df[col])

def _load_data(self, data_file):
return pd.read_csv(data_file, parse_dates = date_cols)

def _preprocessing(self, data_file, label_encode = True):
data = self._load_data(data_file)
train, test = train_test_split(data, test_size = 0.2, shuffle=True)
if label_encode:
self.label_encode_df(train, self.cat_cols)
self.label_encode_df(test, self.cat_cols)
return train, test


class FeatureGenerator(Data):
def __init__(self, data_file, target, cat_cols, date_cols, num_cols, id_cols, city_cols):
'''initializes class and creates groupby object for data'''
Data.__init__(self, data_file, target, cat_cols, date_cols, num_cols, id_cols, city_cols)
#self.cat_cols = data.cat_cols
self.groups = self.train_df.groupby(self.cat_cols)

def create_new_cols(self):
'''adds group statistics and time related columns to data stored in data object'''
#creates time related columns
self.train_df = self._data_association(self.train_df)
self.test_df = self._data_association(self.test_df)

#get group stats
group_stats_df = self._get_group_stats()
group_stats_df.reset_index(inplace=True)

#merge derived columns to original df
self.train_df = self._merge_new_cols(self.train_df, group_stats_df, self.cat_cols,
fillna=True)
self.test_df = self._merge_new_cols(self.test_df, group_stats_df, self.cat_cols,
fillna=True)

#drop unecessary columns
self.train_df = self._kicking_cols(self.train_df, city_cols, date_cols, id_cols, cleaning =
True)
self.test_df = self._kicking_cols(self.test_df, city_cols, date_cols, id_cols, cleaning =
True)

#update column lists
```

```
group_stats_cols = ['group_mean', 'group_max', 'group_min', 'group_std',
'group_median']
time_related_cols = ['delivery_accuracy', 'days_to_deliver', 'delivered_before_time',
'risk_value']
self._extend_col_lists(cat_cols=group_stats_cols, num_cols = time_related_cols)
# self._check_corr_status(df = self.train_df, target = self.target)

def _check_corr_status(self, df, target):
cor = df.corr()[target].sort_values()
cor = pd.DataFrame(cor).reset_index().reset_index()
cor = cor.dropna()

xs = list(cor.level_0)
ys = list(cor.review_score)
figsize(13,6)
sns.barplot(x='index', y='review_score', data=cor, orient='h')
plt.xlabel('Atributos',size=10)
plt.ylabel("Pearson's R",size=10)
plt.title('Correlação linear dos atributos pelo target', size=10)
plt.xticks(rotation=45)
for x,y in zip(xs,ys):

label = "{:.2f}".format(y)

plt.annotate(label, # this is the text
 (x,y), # this is the point to label
 textcoords="offset points", # how to position the text
 xytext=(0,10), # distance from text to points (x,y)
 ha='center')


def _data_association(self, df):
'''creates columns related to time'''
df['delivery_accuracy'] = (df.order_estimated_delivery_date -
df.order_delivered_customer_date)/(np.timedelta64(1, 'D'))
# df.delivery_accuracy = df.delivery_accuracy.apply(np.floor)

df['days_to_deliver'] = (df.order_estimated_delivery_date
- df.order_approved_at)/(np.timedelta64(1, 'D'))
# df.days_to_deliver = df.days_to_deliver.apply(np.floor)

df['delivered_before_time'] = np.where(df.delivery_accuracy <0, 0, 1)
df['risk_value'] = pd.qcut(df.price,15,np.arange(0,15))
return df

def _get_group_stats(self):
'''calculates group statistics'''
target_col = self.target
group_stats_df = pd.DataFrame({'group_mean': self.groups[target].mean()})
group_stats_df['group_max'] = self.groups[target_col].max()
```

```
group_stats_df['group_min'] = self.groups[target_col].min()
group_stats_df['group_std'] = self.groups[target_col].std()
group_stats_df['group_median'] = self.groups[target_col].median()
return group_stats_df

def _merge_new_cols(self, df, new_cols_df, keys, fillna=False):
'''merges engineered features with original df'''
df = pd.merge(df, new_cols_df, on=keys, how='left')
if fillna:
df.fillna(0, inplace=True)
#display(df)
return df

def _extend_col_lists(self, cat_cols=[], num_cols=[]):
'''addes engineered feature cols to data col lists'''
self.num_cols.extend(num_cols)
self.cat_cols.extend(cat_cols)
self.feature_cols.extend(num_cols + cat_cols)

def _kicking_cols(self, df, city_cols, date_cols, id_cols, cleaning = True):
cols_to_kick = city_cols + date_cols + id_cols
if cleaning:
cols_to_stay = [col for col in df.columns if col not in cols_to_kick]
df = df[cols_to_stay]
return df

%time
engineer_features = True
if engineer_features:
data = FeatureGenerator(data_file, target, cat_cols, date_cols, num_cols, id_cols,
city_cols)
data.create_new_cols()

class ModelContainer:
def __init__(self, models=[]):#, default_num_iters=10, verbose_lvl=0):
'''initializes model list and dicts'''
self.models = models
self.best_model = None
self.predictions = None
self.mean_mse = {}
#self.default_num_iters = default_num_iters
#self.verbose_lvl = verbose_lvl

def add_model(self, model):
self.models.append(model)

def cross_validate(self, data, k=3, num_procs=1):
'''cross validate models using given data'''
feature_df = data.train_df[data.feature_cols]
target_df = data.train_df[data.target]
```

```
for model in self.models:
neg_mse = cross_val_score(model, feature_df, target_df, cv=k, n_jobs=num_procs,
scoring='neg_mean_squared_error')
self.mean_mse[model] = -1.0*np.mean(neg_mse)

def select_best_model(self):
'''select model with lowest mse'''
self.best_model = min(self.mean_mse, key=self.mean_mse.get)

def best_model_fit(self, features, targets):
'''fits best model'''
self.best_model.fit(features, targets)

def best_model_predict(self, features):
'''scores features using best model'''
self.predictions = self.best_model.predict(features)

def save_results(self):
pass

@staticmethod
def get_feature_importance(model, cols):
'''retrieves and sorts feature importances'''
if hasattr(model, 'feature_importances_'):
importances = model.feature_importances_
feature_importances = pd.DataFrame({'feature':cols, 'importance':importances})
feature_importances.sort_values(by='importance', ascending=False, inplace=True)
#set index to 'feature'
feature_importances.set_index('feature', inplace=True, drop=True)
return feature_importances
else:
#some models don't have feature_importances_
return "Feature importances do not exist for given model"

def print_summary(self):
'''prints summary of models, best model, and feature importance'''
print('\nModel Summaries:\n')
for model in models.mean_mse:
print('\n', model, '- MSE:', models.mean_mse[model])
print('\nBest Model:\n', models.best_model)
print('\nMSE of Best Model\n', models.mean_mse[models.best_model])
print('\nFeature Importances\n', models.get_feature_importance(models.best_model,
data.feature_cols))

feature_importances = self.get_feature_importance(models.best_model,
data.feature_cols)
feature_importances.plot.barh()
plt.show()

return feature_importances
```

```
#define number of processors to use for parallel runs
num_procs = 4

#set verbose level for models
verbose_lvl = 0

#create model container
models = ModelContainer()

#create models -- hyperparameter tuning already done by hand for each model
models.add_model(LinearRegression())
models.add_model(RandomForestRegressor(n_estimators=60, n_jobs=num_procs,
max_depth=15, min_samples_split=80, \
 max_features=8, verbose=verbose_lvl))
models.add_model(GradientBoostingRegressor(n_estimators=40, max_depth=7,
loss='ls', verbose=verbose_lvl))

models.cross_validate(data, k=2, num_procs=num_procs)
models.select_best_model()
models.best_model_fit(data.train_df[data.feature_cols], data.train_df[data.target])
models.best_model_predict(data.test_df[data.feature_cols])

fi = models.print_summary()
```