GUILHERME STORTO ALVES

UTILIZAÇÃO DO CRISP-DM PARA CRIAÇÃO DE UM MODELO DE PREVISÃO DE NPS DA CENTRAL DE ATENDIMENTO DE UM GRANDE BANCO BRASILEIRO.

SÃO PAULO

2023

GUILHERME STORTO ALVES

UTILIZAÇÃO DO CRISP-DM PARA CRIAÇÃO DE UM MODELO DE PREVISÃO DE NPS DA CENTRAL DE ATENDIMENTO DE UM GRANDE BANCO BRASILEIRO.

Trabalho de Formatura apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do diploma de Engenheiro de Produção

Orientador: Prof. Dr. Guilherme Ary Plonski

SÃO PAULO

2023

GUILHERME STORTO ALVES

UTILIZAÇÃO DO CRISP-DM PARA CRIAÇÃO DE UM MODELO DE PREVISÃO DE NPS DA CENTRAL DE ATENDIMENTO DE UM GRANDE BANCO BRASILEIRO.

Trabalho de Formatura apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do diploma de Engenheiro de Produção

Orientador: Prof. Dr. Guilherme Ary Plonski

SÃO PAULO

2023

Catalogação-na-publicação

Alves, Guilherme Storto
UTILIZAÇÃO DO CRISP-DM PARA CRIAÇÃO DE UM MODELO DE PREVISÃO DE NPS DA CENTRAL DE ATENDIMENTO DE UM GRANDE BANCO BRASILEIRO. / G. S. Alves -- São Paulo, 2023.
132 p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Produção.

1.Aprendizado de Máquina 2.CRISP-DM 3.Experiência do Cliente 4.NPS 5.Banco de Varejo I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Produção II.t.

À minha família e amigos,

que sempre me apoiaram incondicionalmente.

AGRADECIMENTOS

À minha mãe, pelo amor incondicional e educação.

Ao meu irmão por sempre ter sido minha referência.

Aos amigos que fiz ao longo da minha jornada pelos momentos memoráveis.

Ao Centro Acadêmico da Engenharia de Produção pela oportunidade de fazer coisas que nem imaginava fazer e ser um reduto de fuga dos problemas.

À Cris e Osni, por serem tão importantes na graduação de tantas pessoas não só da Engenharia de Produção. Sem eles, certamente os dias seriam muito mais difíceis.

À Escola Politécnica e à Universidade de São Paulo pelos meios que me fizeram evoluir em todas as áreas da minha vida.

Ao professor Guilherme Ary Plonski (meu xará) pelos ensinamentos ao longo da graduação e na orientação deste trabalho.

À todos os professores e professoras que passaram pela minha vida e expandiram minha consciência.

Aos meus colegas de trabalho que me ajudaram durante toda a trajetória de desenvolvimento deste trabalho.

E a todos que leram estas palavras e meu trabalho, obrigado pela atenção, espero ter ajudado de alguma forma.

“*A mente grata é aquela que*
atrai para si as melhores coisas”

- Platão

RESUMO

A indústria bancária de varejo passou por transformações significativas nos últimos anos. Com a progressão da digitalização, essa indústria tem presenciado o surgimento de *fintechs* que fornecem alternativas de produtos e serviços focados na ideia de Experiência do Cliente. Neste cenário, o presente trabalho tem como objetivo desenvolver uma modelagem de aprendizado de máquina para previsão do NPS (*Net Promoter Score*) da Central de Atendimento de um banco de varejo tradicional e, assim, gerar insumos para melhoria da satisfação dos clientes nos atendimentos. Desta forma, foi feita uma análise dos dados da área em questão, considerando os efeitos das variáveis individualmente e coletivamente e, depois, foi criado um modelo de aprendizado supervisionado de regressão para previsão do NPS e compreensão das variáveis mais relevantes. Foi utilizada a ferramenta CRISP-DM como *framework* para a construção do trabalho, auxiliando no passo a passo da construção de um projeto de mineração de dados. Ao final, os insumos adquiridos ao longo de toda a análise foram utilizados para recomendar alguns planos de ação que a área poderia adotar para melhorar a nota média de NPS.

Palavras-chave: Aprendizado de Máquina. CRISP-DM. Experiência do Cliente. NPS. Banco de Varejo.

ABSTRACT

The retail banking industry has undergone significant transformations in recent years. With the progression of digitalization, this industry has witnessed the emergence of fintechs providing alternative product and service offerings focused on the concept of Customer Experience. In this scenario, this term paper aims to develop a machine learning model to forecast the NPS (Net Promoter Score) of a traditional retail bank's Contact Center, thereby generating inputs to enhance customer satisfaction in the service. Accordingly, an analysis of the data from the area in question was conducted, considering the effects of the variables both individually and collectively, and subsequently, a supervised regression learning model was created to forecast the NPS and understand the most important features. The CRISP-DM tool was used as a framework for the construction of the work, aiding in the step-by-step creation of a data mining project. Ultimately, the inputs acquired throughout the analysis were used to recommend some action plans that the area could adopt to improve the average NPS score.

Keywords: Machine Learning. CRISP-DM. Customer Experience. NPS. Retail Banking.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

EC: Experiência do Cliente

CX: *Customer Experience*

NPS: *Net Promoter Score*

TI: Tecnologia da Informação

EDA: *Exploratory data analysis*

AED: Análise Exploratória de Dados

FEBRABAN: Federação Brasileira de Bancos

CRM: *Customer Relationship Management*

LGPD: Lei Geral de Proteção de Dados

ECDF: *Empirical Cumulative Distribution Function*

MSE: *Mean Squared Error*

MAE: *Mean Absolute Error*

SUMÁRIO

1 INTRODUÇÃO

O crescimento da tecnologia ao longo dos anos também trouxe mudanças no mercado de bancos de varejo. Um elemento fundamental nessa mudança é o advento das *FinTechs*, isto é, empresas de "*Financial Technology*" (em português, tecnologia financeira) que oferecem soluções financeiras baseadas em plataformas de tecnologia de informação (FARIA, 2018). Elas trouxeram para um mercado tradicionalmente dominado por poucos grandes bancos, no Brasil, agilidade de transformação e competitividade nunca antes vista, de uma forma em que esses grandes bancos dominantes se sentiram incomodados e passaram a tomar medidas de transformação tecnológicas e estratégicas para combater esses novos entrantes.

Acompanhado dessa inovação tecnológica, o padrão de comportamento dos consumidores também mudou. Segundo pesquisa do PEW Research Center (2019), 60% dos adultos brasileiros possuem um smartphone e em países desenvolvidos como Coreia do Sul e Estados Unidos, essa penetração chega a 95% e 81%, respectivamente. A Pesquisa TIC Domicílios (2019) mostrou que 3 a cada 4 brasileiros têm acesso à internet. Assim, com essa adoção do digital cada vez mais difundida, os brasileiros também estão preferindo realizar transações financeiras por canais digitais devido principalmente a fatores como facilidade de uso e confiança nas plataformas desenvolvidas (FEBRABAN, 2019).

O que pode ser visto nas novas entrantes, as *fintechs*, é que elas cativam os consumidores ao diminuir a burocracia, os processos demorados e o valor cobrado pelos serviços. Essa fórmula é viável graças à tecnologia e à penetração dela nas sociedades mundo afora e o conceito que está por trás dessa metodologia é o *Customer Experience* (CX), ou Experiência do Cliente em português.

Dentro deste cenário, o trabalho em questão tem como objeto de estudo uma Central de Atendimento (*Contact Center*) de um banco brasileiro tradicional e tem dois objetivos principais:

i. criar um modelo para prever o NPS (*Net Promoter Score*), nota que mede a satisfação dos cliente em relação aos atendimentos;
ii. identificar os fatores mais relevantes que influenciam esta nota.

Para isso, o trabalho apresenta uma modelagem estatística baseada em *machine learning* utilizando o *framework* CRISP-DM como orientação.

1.1 A empresa e definição do problema

Por motivos de confidencialidade, a empresa tratada no trabalho será denominada como "Banco A". Ele é um dos líderes do mercado de banco de varejo no Brasil com mais de 90 mil colaboradores, mais de 2.500 agências, mais de 55 milhões de clientes (dezembro de 2022). Empresa de capital aberto na Bolsa de Valores, presente em mais de 15 países, teve um lucro líquido de mais de R$ 7,6 bi no quarto trimestre de 2022 e mais de 480 mil acionistas. Na sua agenda estratégica, um de seus objetivos é tornar-se *benchmark* em satisfação do cliente, sendo ela uma métrica-chave para toda a organização. O banco quer ser comparado a empresas de excelência em CX como Starbucks, Amazon, Apple, Tesla e Netflix.

Neste cenário, a Centralidade no Cliente tem se tornado um assunto crucial para o Banco A, que tem tomado diversas atitudes de transformação digital, gestão de pessoas e renovação de cultura para melhorar a experiência dos clientes. Parte dessa mudança foi exatamente motivada pela concorrência cada vez mais intensa das *fintechs* que, em sua maioria, já nascem com uma cultura de *Customer Experience* bem definida e que só a aperfeiçoam com o passar do tempo. Inclusive, esse é um diferencial estratégico, o que pode ser observado pelo posicionamento de marketing delas, que está atrelado à qualidade de seu atendimento e solução de problemas, sendo eles, muitas vezes, até previstos antes de acontecerem.

O Banco A tem enfrentado o desafio de criar projetos que alterem essa visão de "banco tradicional", muito vinculada a processos lentos, muito burocráticos e com taxas elevadas. Um dos passos para essa transformação foi a criação de uma área especializada em modelo de atendimento e satisfação de clientes na qual o autor faz parte.

Um dos principais desafios historicamente enfrentados por essa área é a definição e compreensão de quais variáveis de fato são expressivas na insatisfação do cliente quando em contato com a central de atendimento (*contact center*) que é objeto de estudo. Esse passo é crucial para que os esforços de melhoria sejam concentrados em ações que vão realmente impactar a experiência final do consumidor.

1.2 Motivação

A motivação para a criação do tema deste trabalho passou por algumas etapas que se complementam:

- Por se tratar de um tema relacionado ao trabalho do autor, é uma oportunidade para ele se aprofundar mais na metodologia que norteia todo o trabalho dele como

analista, o *Net Promoter System* e também na intersecção dele com os conceitos trabalhados durante a graduação em Engenharia de Produção;

- O autor participou ativamente da implementação de alguns projetos relacionados ao NPS, com mais de 3 anos trabalhando na instituição. Portanto faz sentido consolidar o que foi criado dentro do trabalho;
- O tema de Experiência do Cliente não é muito abordado na graduação de Engenharia de Produção, portanto a pesquisa sobre esse tema e todo o instrumental por trás dele é de interesse do autor;
- O autor tem como objetivo desenvolver a Ciência de Dados como nova habilidade de trabalho.
- O tema é muito atual e relevante para o mercado, não só de banco de varejo, mas também para todas as empresas que pretendem ter um atendimento de excelência, voltado para as necessidades do cliente.
- O impacto positivo que o projeto pode trazer para os clientes e, consequentemente, para a empresa.

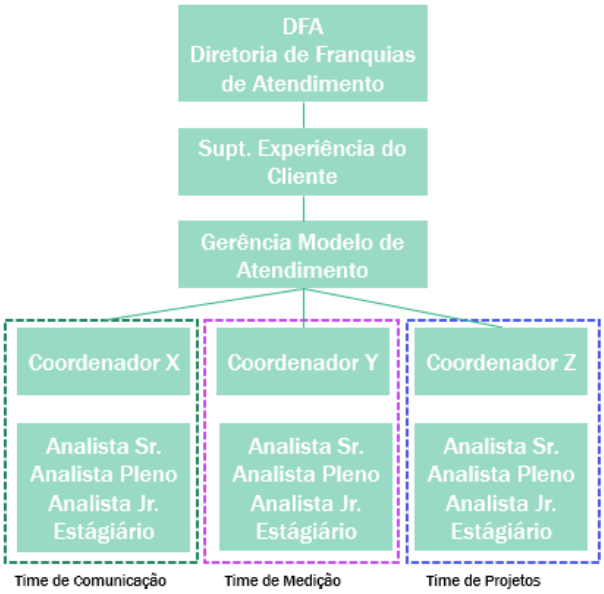
Em suma, este trabalho é a consolidação da prática com o embasamento teórico, dentro de um tema relevante para os tempos atuais, além de ser uma oportunidade de aprender novas áreas de conteúdo.

1.3 Atuação do autor no Banco A

O autor trabalhou por mais de 3 anos na instituição e durante esse período esteve inserido na estrutura da Diretoria de Franquias de Atendimento, na Superintendência de Experiência do Cliente, dentro da Gerência de Modelo de Atendimento. A gerência tem como principal balizador o sistema NPS, implementado em agosto de 2018 no banco e cabe a ela promover projetos que melhorem a satisfação dos clientes do banco e, consequentemente, aumentem o *Net Promoter Score*. Portanto, houve primeiro um entendimento da metodologia pelo autor e depois ele foi alocado na equipe de projetos.

O organograma mostrado na Figura 1 exemplifica a configuração de trabalho durante o período de vínculo do autor com a empresa.

Figura 1 – Organograma da estrutura de trabalho do autor



Fonte: Elaborado pelo autor.

As equipes, ou *squads*, trabalham seguindo metodologia ágil, realizando *sprints*, isto é, períodos programados de entregas.

1.4 Objetivos

O Trabalho de Formatura tem como objetivo analisar a agenda de centralidade no cliente dentro do Banco A e trazer uma solução de modelagem estatística para previsão do NPS do *contact center* do banco e, assim, auxiliar na melhoria da experiência do cliente nele. De uma forma mais específica, os principais objetivos são:

- Trazer a posição do Banco A dentro dos *frameworks* de estratégia de NPS;
- Investigar os dados do *contact center* criando um modelo matemático de previsão do NPS dessa área do banco;
- Propor um enfoque para melhoria dos resultados de satisfação a partir dos insumos do modelo, descrevendo quais características do *contact center* o banco pode focar para obter melhores resultados.

Sendo assim, será aplicada a teoria de Ciência de Dados, Aprendizado de Máquina (*Machine Learning*) e Estatística voltada para a Experiência do Cliente dentro do cenário real do Banco A e do ambiente que ele está inserido.

1.5 Estrutura do Trabalho

O Trabalho de Formatura foi dividido em 6 capítulos. Primeiro tem-se a Introdução, já apresentada. Ela conteve o cenário e as motivações iniciais para o trabalho e seus principais objetivos.

Depois, no segundo capítulo, vão ser discutidos os temas acadêmicos mais relevantes para o trabalho, sobretudo nos campos da Estatística, Ciência de Dados, *Machine Learning* e *Customer Experience*.

No terceiro capítulo é apresentada a metodologia que será utilizada ao longo do desenvolvimento do trabalho junto dos objetivos, avaliação do contexto e plano do projeto.

Já no capítulo quatro, tem-se uma apresentação mais prática do trabalho, sendo apresentado todo o passo a passo da construção de uma modelagem de mineração de dados para o *contact center* do Banco A. As análises são feitas e são apresentadas algumas propostas e encaminhamentos para resolver o problema apresentado.

Depois, o quinto capítulo demonstra as conclusões do trabalho, alguns resultados, limitações, lições aprendidas e os próximos passos.

Por fim, no sexto e último capítulo tem-se as referências bibliográficas utilizadas em todo o desenvolvimento do estudo, fundamentais para o embasamento teórico.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo irá apresentar os principais temas teóricos, conceitos e metodologias que embasam este trabalho de formatura, capacitando, assim, uma abordagem crítica e resolutiva do problema apresentado na introdução bem como o atingimento dos objetivos propostos. Alguns temas foram debatidos durante a graduação em Engenharia de Produção, mas também há temas complementares na abordagem do estudo. Em listagem, os conceitos estudados são:

- Experiência do cliente
- *Net Promoter System*
- *Net Promoter Score*
- CRISP-DM
- Análise exploratória de dados
- Ciência de dados
- Modelagem e métricas de regressão
- *Machine Learning*

2.1 Experiência do Cliente

2.1.1 Histórico e Definição

A origem da Experiência do Cliente (EC) ou em inglês *Customer Experience* (CX) remonta à década de 1990, quando Pine e Gilmore (1999) publicaram o livro "*The Experience Economy: Work Is Theater & Every Business a Stage*". Nesta obra, os autores argumentam que a economia está evoluindo para uma economia de experiência, na qual as empresas devem se concentrar em criar experiências únicas e diferenciadas para seus clientes, em vez de simplesmente vender produtos e serviços. Pine e Gilmore (1999) foram pioneiros ao abordar a experiência do cliente como um novo paradigma para a criação de valor.

Os autores afirmam que a economia de experiência se baseia em quatro estágios:

1. ***Commodities*:** os produtos são indiferenciáveis e vendidos com base no preço;
2. **Bens físicos:** os produtos são diferenciados com base em recursos tangíveis, como design e qualidade;
3. **Serviços:** as empresas agregam valor aos bens físicos por meio de serviços associados, como suporte técnico e garantia;

4. **Experiências:** as empresas criam valor por meio de experiências únicas e memoráveis para os clientes.

A partir desse conceito, surgiu a ideia de que a EC é um produto em si, que pode ser comercializado e precificado. Neste sentido, Pine e Gilmore (1999) argumentam que a EC pode ser considerada como uma forma de marketing, que busca criar valor para o cliente através de experiências únicas e diferenciadas.

Desde então, o conceito de CX vem sendo aprimorado e expandido por diversos autores. Meyer e Schwager (2007) propõem um modelo de três elementos para a EC:

1. **Funcional:** refere-se à qualidade do produto ou serviço em si;
2. **Acessível:** refere-se à facilidade de acesso e uso do produto ou serviço;
3. **Emocional:** refere-se à emoção que a experiência gera no cliente.

A Experiência do Cliente tem se mostrado uma abordagem eficaz para a fidelização de clientes e o aumento da satisfação do cliente. Para proporcionar uma boa experiência ao cliente, é necessário que a empresa conheça profundamente o seu público-alvo. Verhoef et al. (2009) destacam a importância da gestão estratégica da EC, que envolve a identificação dos determinantes da experiência, a análise da dinâmica da experiência e a implementação de estratégias de gestão da EC. Essas estratégias devem ser orientadas para a personalização e a individualização do atendimento, de forma a atender às expectativas dos clientes.

Complementar a isso, Rosenbaum e Massiah (2011) enfatizam que a EC deve ser vista como um processo contínuo, pois as necessidades e expectativas dos clientes mudam ao longo do tempo.

Em resumo, a origem da EC está ligada à evolução da economia de bens e serviços para a economia de experiência. A EC é um tema complexo e multidisciplinar que envolve aspectos funcionais, emocionais e estratégicos, é um conceito que se refere a todas as interações que um cliente tem com uma marca, desde a descoberta de um produto ou serviço até o pós-venda. É uma abordagem centrada no cliente que busca oferecer uma experiência positiva e memorável. A partir desse conceito, ela foi aprimorada e expandida por diversos autores, se tornando uma abordagem importante para a fidelização de clientes e aumento da satisfação do cliente.

2.1.2 Fatores que Influenciam a Experiência do Cliente

Os estudos de Kumar et al. (2013) e Grewal et al. (2017) identificaram vários fatores que influenciam a experiência do cliente, como a qualidade do produto, o atendimento ao cliente, o ambiente de compra e o preço. Além disso, a personalização é um elemento-chave na

experiência do cliente, conforme apontado por Peppers e Rogers (1997), que destacam a importância de adaptar a comunicação e a oferta aos interesses e necessidades individuais dos clientes.

1. **Qualidade do produto e serviço:** Zeithaml et al. (1996) afirmam que a qualidade percebida tem um efeito direto sobre a satisfação do cliente e a intenção de recompra. Por isso, é essencial que as empresas busquem sempre oferecer produtos e serviços de alta qualidade e trabalhem na melhoria contínua deles.
2. **Atendimento ao cliente:** Berry et al. (2002) destacam a importância de um atendimento eficiente, personalizado e empático para construir relacionamentos sólidos com os clientes e aumentar a sua satisfação. Treinar e capacitar os colaboradores para lidar com os clientes é fundamental para garantir um atendimento de qualidade.
3. **Ambiente de compra:** seja físico ou virtual, também afeta a experiência do cliente (KOTLER, 1973). Um ambiente agradável, confortável e que facilite a navegação dos clientes pode contribuir para uma experiência positiva e aumentar a probabilidade de recompra. Gentile et al. (2007) ressaltam a importância de criar ambientes sensoriais que estimulem a percepção dos clientes, aumentando a sua satisfação e engajamento.
4. **Preço:** empresas que adotam estratégias de precificação adequadas e justas podem melhorar a experiência do cliente e aumentar a sua lealdade (DODDS et al., 1991).
5. **Personalização:** a personalização das ofertas e da comunicação com os clientes é um elemento-chave na experiência do cliente (PEPPERS & ROGERS, 1997). O uso de tecnologias e análise de dados para identificar as necessidades e preferências dos clientes permite que as empresas ofereçam soluções e experiências personalizadas, aumentando a satisfação e a lealdade (VERHOEF et al., 2010).

2.1.3 Importância

A experiência do cliente tem se tornado cada vez mais importante para as empresas, pois a qualidade da experiência que um cliente tem ao interagir com uma empresa pode ter um impacto significativo em seu comportamento e nas decisões de compra futuras. Pode-se destacar quatro fatores relevantes: diferencial competitivo, lealdade e retenção de clientes, impacto positivo no desempenho financeiro e desenvolvimento de inovação.

Pine e Gilmore (1999) dizem que a criação de experiências memoráveis é essencial para as empresas prosperarem e se destacarem em um mercado cada vez mais competitivo. Em 2020, 81% das empresas esperavam competir principalmente com base na experiência do cliente (GARTNER, 2019).

Reichheld & Markey (2011) argumentam que a experiência do cliente é um fator crítico para a fidelidade e recomendação de clientes, e que as empresas devem priorizar a melhoria contínua da experiência do cliente para se destacarem em seus mercados. Clientes satisfeitos e que vivenciam experiências positivas têm maior probabilidade de continuar comprando e recomendar a marca a outras pessoas, gerando um efeito multiplicador e aumentando o valor do cliente ao longo do tempo (MORGAN & REGO, 2006).

Outro ponto importante relacionado à EC e fundamental para as empresas é a sua relação direta com receita. Empresas com uma experiência do cliente superior têm maior fidelidade de clientes, menor rotatividade de clientes e maiores receitas do que empresas com uma experiência do cliente inferior (FORRESTER, 2021). Empresas líderes em experiência do cliente também tendem a ter um crescimento de receita até duas vezes maior do que empresas não líderes e têm maior engajamento, satisfação e retenção dos clientes (MCKINSEY, 2023).

A experiência do cliente também está relacionada à inovação. Entender as necessidades e desejos dos clientes pode levar ao desenvolvimento de novos produtos, serviços e modelos de negócio (PRAHALAD & RAMASWAMY, 2004).

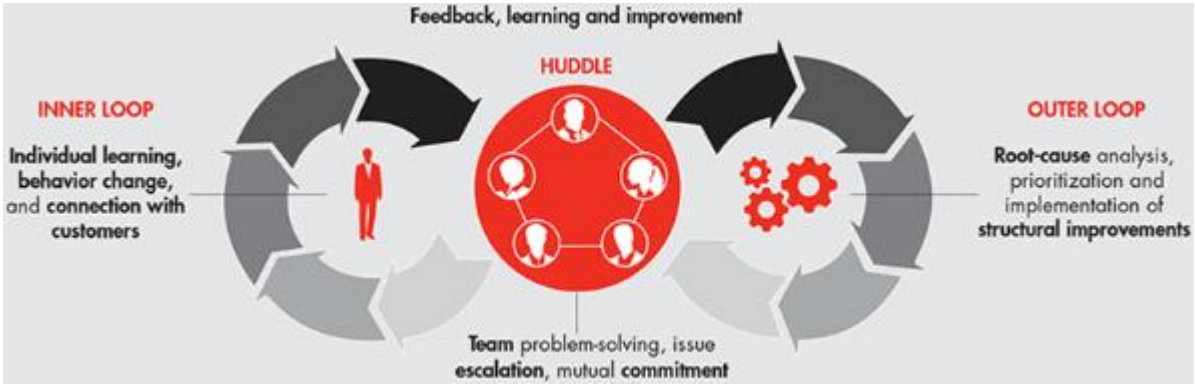
2.2 *Net Promoter System*

2.2.1 Origens e definição

O *Net Promoter System* é uma marca criada e registrada pela empresa Bain & Company, Satmetrix Systems e Fred Reichheld. É uma metodologia de gestão de experiência do cliente desenvolvida por Fred Reichheld (2003) que se concentra em medir e melhorar a lealdade do cliente. O sistema é pautado em uma estrutura de raciocínio (Figura 2) composta por três pilares:

1. *Inner loop* (ciclo interno): coletar *feedback* dos clientes;
2. *Huddle* (círculo de pessoas): reuniões de equipe dentro da empresa para agir sobre o *feedback*;
3. *Outer loop* (ciclo externo): fechar o *loop* com o cliente oferecendo soluções para os problemas identificados nos *feedbacks*.

Figura 2 – *Framework Net Promoter System*



Fonte: Markey & Reichheld (2011).

O *Inner Loop* é uma das etapas do *Net Promoter System* e se refere ao processo de coleta e análise de *feedback* imediato do cliente para identificar e resolver problemas rapidamente. É quando a empresa tem a oportunidade de demonstrar ao cliente que sua opinião é valorizada e que a empresa está comprometida em fornecer a melhor experiência possível. (REICHHELD, 2003).

A coleta de *feedback* pode ser feita por meio de diversos canais como questionários online, *e-mails* e *chats*. É importante que a empresa tenha uma abordagem proativa na coleta de *feedback*, buscando identificar problemas em tempo hábil e resolver rapidamente (REICHHELD, 2006).

Já a *Huddle* do *Net Promoter System* é uma reunião curta e interativa em equipe, que ocorre regularmente em todos os níveis da organização e serve a uma variedade de propósitos no sistema de gestão de clientes. Segundo Reichheld (2006), as *Huddles* são importantes para criar um ritmo regular de atividades e fazer os funcionários focarem no cliente, construindo o compromisso da equipe com os clientes. Além disso, oferecem um local para dar e receber ajuda dos colegas, e proporcionam a oportunidade de elevar questões mais amplas para outras partes da organização, onde podem ser resolvidas.

A *Huddle* ajuda as equipes a desenvolver um sentimento de propriedade pela experiência do cliente e sua própria experiência como funcionários. É um facilitador primário da cultura e mudança de comportamento em uma organização, e é um grande passo no caminho para criar uma força de trabalho autodirigida e autocorretiva (REICHHELD, 2006).

Por fim, o *Outer Loop* do *Net Promoter System* é o processo de análise e ação que as empresas utilizam para identificar as principais causas raiz dos problemas apontados pelos clientes e para implementar mudanças que resolvam esses problemas.

Ele começa com a análise de dados coletados no processo de *feedback* do cliente, incluindo os comentários dos clientes e as pontuações de classificação. As empresas usam esses dados para identificar tendências e padrões de comportamento do cliente, bem como para identificar as áreas problemáticas em sua experiência do cliente (REICHHELD, 2006).

Em seguida, as empresas usam técnicas de melhoria contínua para identificar as causas raiz dos problemas identificados. Isso pode incluir a realização de pesquisas de mercado, a análise de dados de processo interno e a colaboração com os departamentos relevantes para entender melhor as áreas problemáticas. As empresas também podem utilizar a análise de dados para prever problemas futuros antes que eles ocorram (REICHHELD, 2006).

Finalmente, as empresas usam as informações coletadas para implementar mudanças significativas em seus processos de negócios e melhorar a experiência do cliente. Isso pode incluir a melhoria dos processos internos, a implementação de novas tecnologias e/ou a reorganização da estrutura organizacional (REICHHELD, 2006).

2.2.2 *Net Promoter Score* (NPS)

2.2.2.1 Origens e objetivo

Dentro do *Net Promoter System*, mais precisamente na parte de coleta de *feedbacks* (*inner loop*), é utilizado o indicador do *Net Promoter Score* (NPS). Ele é uma métrica de lealdade do cliente desenvolvida por Fred Reichheld (2003) para avaliar a satisfação do cliente e prever o crescimento dos negócios.

O NPS foi introduzido por Fred Reichheld em seu artigo "*The One Number You Need to Grow*" na Harvard Business Review (2003). Reichheld (2003) argumentou que as empresas precisavam de uma única métrica simples para medir a lealdade do cliente e identificar áreas de melhoria. O NPS foi desenvolvido como uma solução para essa necessidade, baseado em uma única pergunta: "Numa escala de 0 a 10, qual é a probabilidade de você nos recomendar (ou recomendar este produto/serviço/marca) a um amigo ou colega?" e em complemento a ela, ainda instruiu as empresas a realizarem uma pergunta aberta complementar para entender a motivação da nota dada na primeira pergunta: "Qual é o motivo mais importante para a nota que você deu?" (REICHHELD & MARKEY, 2011). Isso leva os próprios clientes a usarem suas palavras, sem vieses, e ainda auxilia na descoberta da causa raiz dos problemas.

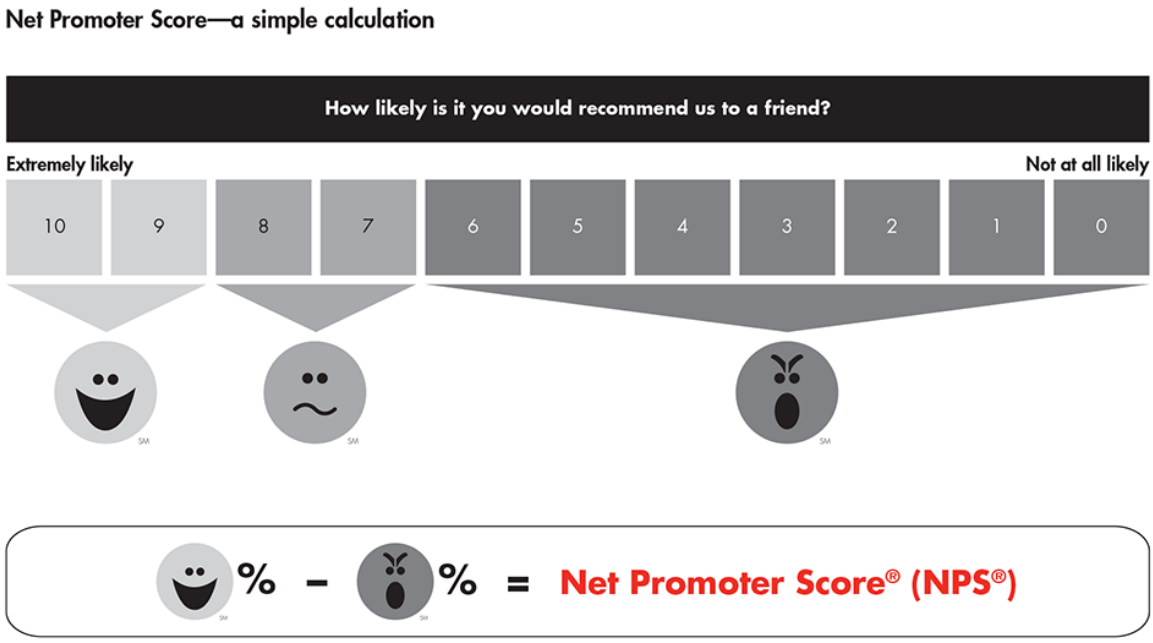
Para chegar à "pergunta definitiva", Reichheld (2003) e sua equipe fizeram 14 estudos de caso em diferentes setores de empresas para encontrar qual pergunta tinha a maior correlação

estatística com relação às compras repetidas ou as recomendações feitas pelos clientes e em 11 dos 14 foi essa que estava no *top* 1 e 2.

2.2.2.2 Cálculo e categorização do NPS

O NPS é calculado subtraindo a porcentagem de detratores (clientes que dão uma pontuação de 0 a 6) da porcentagem de promotores (clientes que dão uma pontuação de 9 ou 10). Os clientes que dão uma pontuação de 7 ou 8 são considerados neutros e não afetam diretamente o cálculo (REICHHELD & MARKEY, 2011). O resultado é um número entre -100 e 100, que pode ser usado como um indicador da lealdade geral do cliente.

Figura 3 – Cálculo e categorização do NPS



Fonte: Markey & Reichheld (2011).

Nos estudos de caso de Reichheld (2003) o grupo de promotores representavam as taxas de recompra mais altas e mais de 80% das recomendações, os neutros eram pessoas passivamente satisfeitas, mas que iam embora ou escolhiam outro produto/serviço caso eles tivessem uma oferta melhor, já os detratores representavam mais de 80% da propaganda negativa feita em relação ao produto/serviço.

2.2.2.3 Princípios do NPS

As empresas alteram o nome da metodologia, modificam a pergunta em alguns aspectos, mas há 3 elementos fundamentais que não podem ser dispensados:

1. “As empresas precisam caracterizar sistematicamente promotores e detratores.” (REICHHELD & MARKEY, 2011, P. 9). Isso deve fazer sentido para os funcionários da linha de frente e esse número deve ser compilado e divulgado para toda a organização;
2. “As empresas devem pautar-se pelo aprendizado de ciclo fechado e criar processos de melhoria, incorporando-os em suas operações diárias” (REICHHELD & MARKEY, 2011, P. 9).
3. “A missão fundamental de CEOs e outros líderes deve ser criar mais promotores e menos detratores” (REICHHELD & MARKEY, 2011, P. 9).

Além disso há 8 princípios que Reichheld (2011) argumenta serem necessários:

1. **“Faça a pergunta definitiva e nada muito além disso”:** ao fazer apenas a pergunta definitiva e nada muito além disso, as empresas evitam sobrecarregar os clientes com pesquisas longas e cansativas, aumentando a probabilidade de obter respostas mais honestas e úteis. Além disso, a simplicidade da pergunta facilita a comparação dos resultados entre diferentes segmentos de clientes, departamentos ou concorrentes (REICHHELD & MARKEY, 2011). Embora a pergunta definitiva seja o elemento central do NPS, Reichheld e Markey (2011) reconhecem que, em alguns casos, pode ser útil fazer perguntas adicionais para obter *insights* mais detalhados sobre a experiência do cliente. No entanto, é importante manter essas perguntas complementares focadas e limitadas em número. Exemplos de perguntas adicionais podem incluir solicitação de *feedback* aberto sobre o motivo da nota atribuída ou perguntas específicas sobre elementos-chave da experiência do cliente, como atendimento ao cliente, qualidade do produto ou tempo de entrega.
2. **“Escolha uma escala que funcione e fique com ela”:** este princípio destaca a importância de utilizar uma escala consistente e eficaz para medir a lealdade do cliente e fornecer resultados comparáveis ao longo do tempo e entre diferentes segmentos de clientes e empresas (REICHHELD & MARKEY, 2011).

   A escala NPS de 0 a 10 oferece várias vantagens em relação a outras escalas possíveis. Primeiro, ela é simples e fácil de entender para os clientes, o que facilita a coleta de *feedback* honesto e útil (REICHHELD & MARKEY, 2011). Além disso,

a escala de 0 a 10 permite uma ampla gama de respostas, o que possibilita a identificação de diferenças significativas na lealdade do cliente entre promotores, neutros e detratores.

Reichheld e Markey (2011) enfatizam a importância de escolher uma escala que funcione e manter-se fiel a ela. Ao utilizar consistentemente a mesma escala em pesquisas de NPS, as empresas podem comparar os resultados ao longo do tempo e entre diferentes segmentos de clientes, departamentos e concorrentes. A consistência da escala também facilita a comunicação dos resultados do NPS dentro da organização e a definição de metas e expectativas.

3. **"Evite confusão entre índice interno (de baixo para cima) e índice externo (de cima para baixo ou *benchmark*)":** este princípio refere-se à necessidade de distinguir claramente entre a análise interna do NPS e a comparação com *benchmarks* externos para evitar mal-entendidos e tomar decisões informadas sobre a melhoria da lealdade do cliente e do desempenho empresarial (REICHHELD & MARKEY, 2011).
    a. **Índice Interno** (de baixo para cima): é calculado com base nas respostas dos clientes à pergunta definitiva: "Em uma escala de 0 a 10, qual a probabilidade de você recomendar nossa empresa/marca/produto a um amigo ou colega?" (REICHHELD & MARKEY, 2011). Este índice é usado para medir e rastrear a lealdade do cliente dentro da organização e identificar áreas de melhoria na experiência do cliente. O índice interno pode ser analisado em diferentes níveis, como departamentos, unidades de negócios ou segmentos de clientes, para compreender as variações na lealdade do cliente e identificar oportunidades de melhoria (REICHHELD & MARKEY, 2011).
    b. **Índice Externo** (de cima para baixo ou *benchmark*): refere-se à comparação do NPS da empresa com o de outras empresas, geralmente concorrentes ou líderes do setor. Isso permite que a empresa avalie seu desempenho em relação ao mercado e identifique áreas em que pode aprender com as melhores práticas (REICHHELD & MARKEY, 2011). A comparação com *benchmarks* externos pode ser útil para estabelecer metas realistas de melhoria do NPS e avaliar o desempenho da empresa em relação às expectativas do mercado (REICHHELD & MARKEY, 2011).

É importante evitar confusões entre o índice interno e o índice externo do NPS. Misturar ou confundir essas duas abordagens pode levar a conclusões

incorretas e dificultar a tomada de decisões informadas sobre como melhorar a lealdade do cliente (REICHHELD & MARKEY, 2011). Para evitar essa confusão, as empresas devem comunicar claramente a diferença entre esses índices aos funcionários e garantir que as análises e relatórios do NPS sejam claramente rotulados como internos ou externos (REICHHELD & MARKEY, 2011).

4. **"Busque altas taxas de resposta dos clientes certos":** este princípio destaca a importância de obter *feedback* representativo e abrangente dos clientes para garantir que as empresas possam tomar decisões informadas e efetivas com base nas informações coletadas (REICHHELD & MARKEY, 2011).

As taxas de resposta são um fator importante a ser considerado ao coletar *feedback* dos clientes por meio do NPS. Uma alta taxa de resposta ajuda a garantir que a empresa receba *feedback* de uma amostra representativa de clientes, o que permite uma análise mais precisa e confiável dos resultados REICHHELD & MARKEY, 2011). Além disso, altas taxas de resposta fornecem uma base sólida para a tomada de decisões e a identificação de áreas de melhoria na EC.

O princípio ainda ressalta a importância de coletar *feedback* dos clientes mais relevantes para a empresa. Isto pode incluir clientes de diferentes segmentos, regiões geográficas, canais de venda ou histórico de compras (REICHHELD & MARKEY, 2011). Ao garantir que o *feedback* seja coletado dos clientes certos, as empresas podem obter uma visão abrangente das necessidades e preferências de seu público-alvo, o que permite uma melhor tomada de decisões e ações para melhorar a EC.

Reichheld e Markey (2011) oferecem várias sugestões para aumentar as taxas de resposta e obter *feedback* dos clientes certos. Algumas dessas estratégias incluem:

- Facilitar a participação dos clientes: Tornar o processo de fornecer *feedback* o mais fácil e conveniente possível, como oferecer um questionário curto e simples, e permitir que os clientes respondam por meio de diferentes canais, como e-mail, SMS ou plataformas online.
- Comunicar a importância do *feedback*: Informar aos clientes que sua opinião é valorizada e será usada para melhorar a experiência do cliente, o que pode incentivar a participação.
- Fornecer incentivos: Oferecer incentivos, como descontos, brindes ou sorteios, pode motivar os clientes a fornecer *feedback*.

- Segmentar e personalizar a abordagem: Adaptar a comunicação e a abordagem com base nas características e preferências dos diferentes segmentos de clientes, o que pode aumentar a probabilidade de obter *feedback* dos clientes certos.

5. **"Faça relatórios de dados de relacionamento e de dados financeiros com a mesma frequência":** esse princípio ressalta a importância de acompanhar e reportar tanto as métricas de relacionamento com o cliente quanto as métricas financeiras, para garantir uma visão holística do desempenho da empresa e facilitar a tomada de decisões informadas (REICHHELD & MARKEY, 2011). Ao reportar dados de relacionamento e financeiros com a mesma frequência, as empresas podem identificar tendências e correlações entre as métricas de desempenho do cliente e as métricas financeiras (REICHHELD & MARKEY, 2011). Isso pode ajudar a empresa a entender como as iniciativas de experiência do cliente estão impactando o desempenho financeiro e a identificar oportunidades para alinhar melhor as estratégias de negócios com as necessidades e expectativas dos clientes.
6. **"Aprenda mais rápido e aumente sua capacidade de atribuir responsabilidades com dados mais detalhados":** esse princípio enfatiza a importância de coletar e analisar dados detalhados para acelerar o aprendizado organizacional e melhorar a responsabilidade no processo de melhoria da experiência do cliente (REICHHELD & MARKEY, 2011).

   No que tange "aprender mais rápido", o uso de dados detalhados no contexto do NPS permite que as empresas identifiquem tendências e padrões mais rapidamente e tomem decisões informadas sobre como melhorar a experiência do cliente (REICHHELD & MARKEY, 2011). Ao coletar informações granulares sobre a satisfação do cliente, as empresas podem identificar áreas específicas de melhoria e priorizar as iniciativas que terão o maior impacto na lealdade do cliente e no desempenho financeiro.

   Dados mais detalhados também ajudam as empresas a "atribuir responsabilidades" de forma mais eficaz. Ao analisar as informações coletadas, é possível identificar quais equipes, departamentos ou unidades de negócios estão contribuindo mais significativamente para a EC, tanto positiva quanto negativamente (REICHHELD & MARKEY, 2011). Essa atribuição de responsabilidade ajuda a garantir que os recursos sejam alocados de forma eficiente e que as equipes sejam responsáveis por melhorar suas respectivas áreas de atuação.

Reichheld e Markey (2011) fornecem várias dicas para coletar dados detalhados no contexto do NPS. Algumas dessas estratégias incluem:

- Segmentar clientes: segmentar os clientes com base em critérios como comportamento de compra, demografia ou preferências ajuda a identificar padrões específicos e a entender melhor as necessidades e expectativas de diferentes grupos de clientes.
- Analisar comentários abertos: além da pergunta definitiva do NPS, é útil incluir uma seção de comentários abertos para que os clientes possam fornecer *feedback* mais detalhado e específico sobre sua experiência.
- Rastrear métricas adicionais: coletar e analisar métricas adicionais relacionadas à experiência do cliente, como tempo de resposta, taxas de resolução de problemas e satisfação com o atendimento, pode fornecer *insights* valiosos sobre áreas de melhoria.
- Integrar dados de várias fontes: combinar dados do NPS com outras fontes de informações, como análise de comportamento de compra, dados de suporte ao cliente e interações nas redes sociais, pode fornecer uma visão mais abrangente da experiência do cliente.

7. **“Faça auditorias para assegurar precisão e isenção”:** esse princípio está relacionado à necessidade de se antecipar fontes de possível origem de viés e reduzi-las através de métodos avançados de medição (REICHHELD & MARKEY, 2011).

Segundo Reichheld e Markey (2011), o NPS está sujeito a quatro tipos de viés:

1. **Medo de retaliação**: um cliente não dar uma resposta fiel a sua percepção simplesmente pelo medo de seu fornecedor utilizá-la como motivação para tratá-lo de forma pior, por isso, é importante ressaltar a confidencialidade das respostas;
2. **Fraude:** fornecedores podem oferecer benefícios aos clientes para estes darem notas mais altas na pontuação, por isso, é importante educar os clientes e empregados sobre a ética da pesquisa;
3. **Amostragem parcial:** não pesquisar pessoas possivelmente detratoras para aumentar a nota ou até mesmo pesquisar poucas pessoas e não ter massa suficiente para se ter uma percepção fiel da nota dado que detratores tendem a responder menos a pesquisa;

4. **Notas infladas:** nem sempre os clientes se sentem confortáveis em dar uma nota fiel à pesquisa dependendo da situação e que se encontram, por exemplo, dar *feedback* diretamente à pessoa que cuidou de seu atendimento/produto. Desta forma, uma alternativa para se diminuir este viés é deixar que um terceiro realize a pesquisa e que reforce mais uma vez o anonimato do pesquisado.

Reichheld e Markey (2011) ainda propõem algumas outras alternativas para se reduzir os viéses da pesquisa NPS:

- **Padronização da metodologia:** utilizar uma abordagem padronizada para a coleta de dados do NPS, garantindo que todos os entrevistados sejam tratados de maneira consistente e justa.
- **Treinamento de equipe:** treinar a equipe responsável pela coleta de dados e análise para garantir que entendam os objetivos da pesquisa NPS e sigam as melhores práticas.
- **Anonimato dos respondentes:** garantir que as respostas sejam anônimas para que os entrevistados se sintam mais confortáveis em fornecer *feedback* honesto e imparcial.
- **Seleção aleatória de participantes:** utilizar uma seleção aleatória de clientes para as pesquisas NPS, garantindo que a amostra seja representativa da base de clientes como um todo.
- **Períodos de coleta de dados consistentes:** conduzir a pesquisa NPS em intervalos regulares e consistentes, evitando períodos de tempo que possam introduzir viés sazonal ou eventos específicos que possam afetar os resultados.
- **Perguntas claras e objetivas:** formular perguntas de pesquisa de maneira clara e objetiva, evitando linguagem ambígua ou tendenciosa que possa influenciar as respostas dos entrevistados.
- **Uso de escalas consistentes:** utilizar uma escala consistente para medir a probabilidade de recomendação, como a escala de 0 a 10 proposta por Reichheld e Markey (2011), para facilitar a comparação e análise dos resultados.

- **Análise de dados imparcial:** realizar análises de dados objetivas e independentes, evitando a introdução de viéses pessoais ou organizacionais na interpretação dos resultados.
- **Auditorias internas e externas:** conduzir auditorias regulares, tanto internas quanto externas, para garantir a precisão e isenção dos dados coletados e dos processos de análise associados.
- **Melhoria contínua do processo:** monitorar continuamente o processo de pesquisa NPS e faça ajustes conforme necessário para garantir a precisão e isenção dos resultados ao longo do tempo.

8. **"Valide a relação entre resultados e comportamentos":** esse princípio enfatiza a necessidade de compreender a relação entre resultados de clientes individuais e os comportamentos e a experiência geral do cliente, bem como os resultados financeiros e operacionais da empresa (REICHHELD & MARKEY, 2011). Esse monitoramento irá garantir que o sistema de *feedback* está livre de viéses, fraudes e manipulação.

2.2.2.4 Benefícios e críticas ao NPS

Reichheld (2011) critica os métodos tradicionais de mensuração da satisfação dos clientes pelos seguintes principais motivos:

> - A maioria das pesquisas é longa e complexa demais, desperdiçando o tempo dos entrevistados.
> - As pesquisas servem apenas como relatório, sem influenciar no aprendizado prático do pessoal da linha de frente.
> - As pesquisas geralmente são anônimas, o que elimina a possibilidade de se fechar o ciclo com clientes individuais.
> - As pesquisas são estruturadas na linguagem do pesquisador, não na do cliente.
> - As taxas de resposta normalmente são baixas, de modo que não é possível confiar nos resultados.
> - Via de regra, quem responde às pesquisas é a pessoa errada – principalmente em contextos business-to-business, em que os altos executivos responsáveis pelas decisões de compra raramente têm tempo para isso.
> - Os resultados são facilmente manipuláveis (pense na última vez em que foi a uma concessionária e o vendedor implorou por uma avaliação positiva de sua parte). (REICHHELD & MARKEY, 2011, P. 72).

Com isso, ele defende o método NPS ser melhor que estes métodos tradicionais, porém, o seu próprio modelo também é criticado por outros autores pelos motivos:

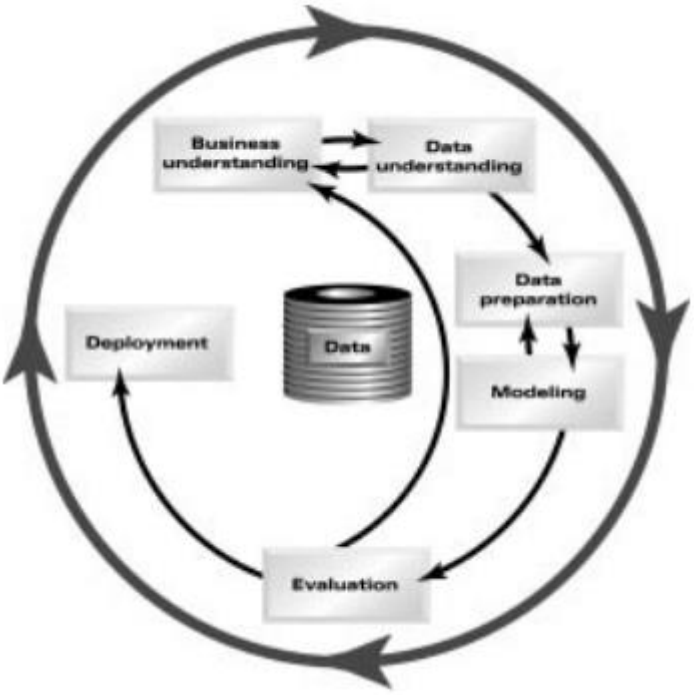
- **Simplificação excessiva:** a métrica é baseada em uma única pergunta, enquanto a satisfação e a lealdade do cliente podem ser influenciadas por uma ampla variedade de fatores (KEININGHAM et al., 2007).
- **Validade e confiabilidade:** falta de evidências empíricas sólidas que sustentem sua validade e confiabilidade como uma métrica de lealdade do cliente e crescimento dos negócios (KEININGHAM et al., 2007). Algumas pesquisas sugerem que o NPS pode não ser superior a outras métricas de satisfação do cliente na previsão do desempenho empresarial (KEININGHAM et al., 2007).
- **Sensibilidade cultural e contextual:** O uso de uma escala de 0 a 10 pode não ser apropriado em todos os contextos culturais, e as diferenças culturais podem afetar a maneira como os clientes respondem à pergunta do NPS (KLAUS & MAKLAN, 2013). Além disso, a aplicabilidade do NPS em diferentes setores e contextos de negócios também foi questionada (KLAUS & MAKLAN, 2013).
- **Foco excessivo na métrica:** pode levar as empresas a negligenciarem outros aspectos importantes da experiência do cliente (DIXON et al., 2010). A ênfase na métrica pode desviar a atenção dos negócios dos problemas subjacentes e das oportunidades de melhoria.

2.3 CRISP-DM

2.3.1 História e descrição

O CRISP-DM (*Cross-Industry Standard Process for Data Mining*) é uma metodologia amplamente adotada para orientar projetos de mineração de dados em diversos setores da indústria. Desenvolvido em 1996 por um consórcio de empresas, incluindo SPSS, Teradata e NCR Corporation, o CRISP-DM tem sido utilizado como um padrão para a realização de projetos de mineração de dados (CHAPMAN et al., 2000).

Figura 4 – Estrutura CRISP-DM



Fonte: Chapman et al. (2000).

Conforme a Figura 4, ele é composto por seis fases principais: Entendimento do Negócio (*Business Understanding*), Entendimento dos Dados (*Data Understanding*), Preparação dos Dados (*Data Preparation*), Modelagem (*Modeling*), Avaliação (*Evaluation*) e Implantação (*Deployment*). Essas fases não são estritamente sequenciais e podem ser iterativas, permitindo que os profissionais de mineração de dados retornem a fases anteriores conforme necessário. Cada fase inclui tarefas específicas e objetivos a serem alcançados, ajudando os profissionais de mineração de dados a estruturar seus projetos de forma eficiente (SHEARER, 2000).

2.3.1.1 Entendimento do Negócio (*Business Understanding*)

A etapa de Entendimento do Negócio é fundamental para o sucesso de um projeto de mineração de dados, pois ajuda a estabelecer o contexto e os objetivos do projeto, além de identificar as partes interessadas e os recursos necessários (CHAPMAN et al., 2000). Essa fase é composta por quatro tarefas principais:

1. **Determinação dos objetivos do projeto:** a equipe deve identificar os objetivos do projeto de mineração de dados, considerando as metas e necessidades do negócio.

Isso inclui a definição de questões específicas que o projeto pretende responder e a identificação de métricas de sucesso (SHEARER, 2000).

2. **Avaliação da situação:** a equipe deve analisar o contexto do projeto, identificando os recursos disponíveis, como dados, pessoal e tecnologias, bem como as restrições e desafios potenciais (CHAPMAN et al., 2000). Essa avaliação permite que a equipe compreenda melhor o ambiente em que o projeto será realizado e identifique possíveis obstáculos.
3. **Determinação dos objetivos de mineração de dados:** com base nos objetivos do projeto e na avaliação da situação, a equipe deve estabelecer objetivos específicos de mineração de dados que estejam alinhados às metas do negócio. Esses objetivos devem ser claros, mensuráveis e realistas, para garantir que a mineração de dados produza resultados úteis e acionáveis (SHEARER, 2000).
4. **Desenvolvimento do plano do projeto:** por fim, a equipe deve criar um plano detalhado do projeto, incluindo a estrutura do projeto, cronograma, alocação de recursos e plano de comunicação (CHAPMAN et al., 2000). O plano do projeto é essencial para garantir que todas as partes interessadas estejam cientes das expectativas e responsabilidades, facilitando a execução eficiente do projeto.

2.3.1.2 Entendimento dos Dados (*Data Understanding*)

A etapa de Entendimento dos Dados visa a coletar, descrever e explorar os dados disponíveis, além de verificar sua qualidade para apoiar o desenvolvimento de modelos de mineração de dados eficazes (CHAPMAN et al., 2000). Essa fase é composta por quatro tarefas principais:

1. **Coleta de dados inicial:** a equipe deve identificar e reunir os conjuntos de dados relevantes para o projeto, levando em consideração os objetivos de mineração de dados e as restrições do negócio (SHEARER, 2000). Isso pode envolver a coleta de dados de várias fontes, como bancos de dados internos, arquivos externos e APIs.
2. **Descrição dos dados:** a equipe deve examinar os dados coletados e criar uma descrição detalhada de seus atributos, incluindo o tipo de dados (numérico, categórico, etc.), o número de registros, a existêncica de valores ausentes e a distribuição dos valores (CHAPMAN et al., 2000). Essa descrição permite que a equipe entenda melhor as características dos dados e identifique possíveis problemas de qualidade.

3. **Verificação da qualidade dos dados:** a equipe deve avaliar a qualidade dos dados, identificando problemas como valores ausentes, erros de entrada e inconsistências (CHAPMAN et al., 2000). A verificação da qualidade dos dados é crucial para garantir que os modelos de mineração de dados sejam construídos com base em dados confiáveis e precisos.
4. **Exploração dos dados:** por fim, a equipe deve realizar uma análise exploratória dos dados, utilizando técnicas como estatísticas descritivas, visualizações e análise de correlações (SHEARER, 2000). Essa exploração ajuda a identificar padrões, tendências e possíveis anomalias nos dados, fornecendo *insights* úteis para o desenvolvimento dos modelos de mineração de dados.

2.3.1.3 Preparação dos Dados (*Data Preparation*)

A etapa de Preparação dos Dados envolve o tratamento dos dados identificados na fase anterior para que possam ser usados de forma eficaz nos modelos de mineração de dados (CHAPMAN et al., 2000). Esta fase inclui várias tarefas, como limpeza, transformação e seleção de dados:

1. **Limpeza de dados:** a equipe deve abordar os problemas de qualidade dos dados identificados na etapa de Entendimento dos Dados, corrigindo erros, tratando valores ausentes e eliminando inconsistências (SHEARER, 2000). A limpeza de dados é essencial para garantir que os modelos de mineração de dados sejam construídos com base em informações precisas e confiáveis.
2. **Transformação de dados:** a equipe deve transformar os dados conforme necessário para facilitar a aplicação de algoritmos de mineração de dados. Isso pode incluir a normalização de variáveis numéricas, a codificação de variáveis categóricas e a criação de novos atributos a partir dos dados existentes (CHAPMAN et al., 2000). A transformação de dados ajuda a garantir que os dados sejam compatíveis com os requisitos dos algoritmos de mineração de dados utilizados.
3. **Seleção de dados:** a equipe deve selecionar os atributos e registros mais relevantes para os objetivos de mineração de dados, eliminando informações desnecessárias ou redundantes (SHEARER, 2000). A seleção de dados é importante para reduzir a complexidade e o tempo de processamento dos modelos de mineração de dados, bem como para melhorar a qualidade das descobertas geradas.

4. **Integração de dados:** em alguns casos, a equipe pode precisar integrar dados de várias fontes para criar um conjunto de dados completo e abrangente (CHAPMAN et al., 2000). Isso pode envolver a junção de tabelas, a agregação de dados e a reconciliação de diferenças nos esquemas de dados.

2.3.1.4 Modelagem (*Modeling*)

A etapa de Modelagem envolve a aplicação de algoritmos e técnicas de mineração de dados para criar modelos que possam ser utilizados para atingir os objetivos de mineração de dados estabelecidos (CHAPMAN et al., 2000). Esta fase inclui várias tarefas, como a seleção de técnicas, a construção de modelos e a avaliação dos resultados:

1. **Seleção de técnicas de modelagem:** a equipe deve escolher as técnicas de mineração de dados mais adequadas para os objetivos do projeto, levando em consideração as características dos dados e os requisitos do negócio (SHEARER, 2000). Essa seleção pode incluir algoritmos de classificação, regressão, agrupamento ou associação, entre outros.
2. **Construção de modelos:** a equipe deve aplicar as técnicas selecionadas aos dados preparados, criando modelos de mineração de dados que possam ser utilizados para extrair conhecimento e padrões relevantes (CHAPMAN et al., 2000). Isso pode envolver a definição de parâmetros, a divisão dos dados em conjuntos de treinamento e teste, e a otimização dos modelos para melhorar seu desempenho.
3. **Avaliação dos modelos:** a equipe deve avaliar o desempenho dos modelos construídos, utilizando métricas e métodos apropriados para os objetivos de mineração de dados (SHEARER, 2000). Essa avaliação pode incluir a comparação dos modelos com *benchmarks*, a validação cruzada e a análise de métricas como acurácia, precisão, recall e F1-score, entre outras.
4. **Ajuste e otimização dos modelos:** com base nos resultados da avaliação, a equipe pode ajustar e otimizar os modelos para melhorar seu desempenho e garantir que atendam aos objetivos de mineração de dados (CHAPMAN et al., 2000). Isso pode envolver a alteração de parâmetros, a seleção de novas características ou a aplicação de técnicas de combinação de modelos, como *ensemble* ou *stacking*.

2.3.1.5 Avaliação (*Evaluation*)

A etapa de Avaliação tem como objetivo verificar se os modelos desenvolvidos na fase de Modelagem atendem aos objetivos de mineração de dados e às necessidades do negócio (CHAPMAN et al., 2000). Essa fase inclui várias tarefas, como a avaliação do desempenho dos modelos, a revisão dos resultados e a consideração do contexto do projeto:

1. **Avaliação do desempenho dos modelos:** a equipe deve analisar o desempenho dos modelos de mineração de dados gerados, utilizando métricas e métodos apropriados para os objetivos do projeto (SHEARER, 2000). Essa avaliação pode incluir a análise de métricas como acurácia, precisão, *recall* e *F1-score*, entre outras, bem como a validação cruzada e a comparação dos modelos com *benchmarks*.
2. **Revisão dos resultados:** a equipe deve revisar os resultados obtidos pelos modelos de mineração de dados, avaliando sua relevância e utilidade para os objetivos do projeto e as necessidades do negócio (CHAPMAN et al., 2000). Essa revisão pode envolver a análise dos padrões e conhecimentos extraídos dos dados, a comparação com o conhecimento prévio e a verificação da consistência dos resultados.
3. **Consideração do contexto do projeto:** a equipe deve levar em consideração o contexto do projeto, incluindo as restrições do negócio, os recursos disponíveis e as expectativas das partes interessadas (SHEARER, 2000). Isso pode envolver a avaliação dos custos e benefícios dos modelos, a análise de possíveis riscos e a identificação de possíveis melhorias.
4. **Decisão sobre a implementação:** com base na avaliação do desempenho, na revisão dos resultados e na consideração do contexto do projeto, a equipe deve decidir se os modelos de mineração de dados são adequados para implementação ou se são necessárias modificações adicionais (CHAPMAN et al., 2000).

2.3.1.6 Implantação (*Deployment*)

A etapa de Implantação envolve a aplicação dos modelos de mineração de dados e os conhecimentos obtidos no ambiente operacional do negócio, a fim de gerar valor e atingir os objetivos estabelecidos (CHAPMAN et al., 2000). Essa fase inclui várias tarefas, como o planejamento da implantação, a monitorização e a manutenção dos modelos:

1. **Planejamento da implantação:** a equipe deve desenvolver um plano detalhado para a implementação dos modelos e dos conhecimentos adquiridos, incluindo a

definição de responsabilidades, prazos e recursos necessários (SHEARER, 2000). O plano deve abordar questões como a integração dos modelos nos sistemas de TI existentes, a adoção de novos processos e a capacitação dos usuários.

2. **Implementação dos modelos e conhecimentos:** a equipe deve colocar em prática o plano de implantação, integrando os modelos de mineração de dados nos sistemas operacionais e aplicando os conhecimentos adquiridos para melhorar a tomada de decisões e os processos do negócio (CHAPMAN et al., 2000). Isso pode envolver a construção de aplicações de suporte à decisão, a criação de relatórios e *dashboards* e a comunicação dos resultados aos usuários finais.
3. **Monitorização e manutenção dos modelos:** após a implantação, a equipe deve monitorar o desempenho e a relevância dos modelos de mineração de dados no ambiente operacional, identificando possíveis problemas e oportunidades de melhoria (SHEARER, 2000). A manutenção dos modelos pode envolver a atualização dos dados de treinamento, a recalibração dos parâmetros e a incorporação de novos conhecimentos e *feedback* dos usuários.
4. **Avaliação dos resultados e impacto no negócio:** por fim, a equipe deve avaliar o impacto dos modelos e dos conhecimentos adquiridos no negócio, analisando métricas e indicadores relevantes para os objetivos estabelecidos (CHAPMAN et al., 2000). Essa avaliação pode incluir a análise de retorno sobre o investimento (ROI), a satisfação dos usuários e a contribuição dos modelos para a melhoria dos processos e resultados do negócio.

2.3.2 Benefícios e críticas

Um dos principais benefícios do CRISP-DM é sua estrutura bem definida e consistente, que facilita o gerenciamento e a execução de projetos de mineração de dados (CHAPMAN et al., 2000). Ao oferecer um processo padronizado, a metodologia promove a comunicação clara e eficiente entre as partes interessadas e permite o compartilhamento de melhores práticas.

Ele também é flexível e pode ser adaptado às necessidades específicas de diferentes contextos, setores e tipos de dados (CHAPMAN et al., 2000). Essa adaptabilidade permite que as equipes personalizem a metodologia de acordo com suas necessidades e requisitos, garantindo que o processo seja relevante e eficaz para uma ampla gama de projetos.

O CRISP-DM ainda fornece orientações detalhadas e abrangentes para cada uma das seis fases do processo (CHAPMAN et al., 2000). Essas orientações ajudam as equipes a planejar

e executar projetos de mineração de dados de maneira eficiente e eficaz, aumentando a probabilidade de sucesso e reduzindo o risco de erros e omissões.

Além disso, o modelo enfatiza a importância do entendimento do negócio como um componente fundamental do processo de mineração de dados (SHEARER, 2000). Essa ênfase ajuda a garantir que os projetos sejam conduzidos de acordo com os objetivos e necessidades do negócio, chegando em resultados mais relevantes e aplicáveis.

Por outro lado há certas críticas ao *framework* como a falta de atualizações significativas desde sua criação (KELLEHER et al., 2015). Essa limitação pode tornar a metodologia menos relevante para abordar desafios emergentes na área de mineração de dados, especialmente no contexto das rápidas mudanças tecnológicas e inovações na ciência de dados.

Além disso, o CRISP-DM é frequentemente criticado por não abordar adequadamente aspectos não técnicos, como questões éticas, legais e de privacidade, que são cada vez mais relevantes no campo da mineração de dados (KELLEHER et al., 2015). Essa lacuna pode levar a projetos que não consideram suficientemente as implicações sociais e legais de suas análises e resultados.

Por fim, outra crítica ao modelo é a falta de orientação sobre como lidar com a escalabilidade e automação dos processos de mineração de dados (MARISCAL et al., 2010). Em projetos de grande escala e complexidade, a abordagem sequencial e centrada no humano do CRISP-DM pode ser insuficiente, exigindo metodologias complementares para abordar esses desafios.

2.4 Análise Exploratória de Dados (AED)

A Análise Exploratória de Dados (AED) ou *Exploratory Data Analysis* (EDA), conforme descrita por Morettin e Singer (2021) é uma abordagem metodológica para o entendimento inicial e a descrição de um conjunto de dados. O objetivo é explorar e entender os dados antes de fazer inferências ou modelos preditivos.

A AED visa proporcionar uma noção inicial dos dados disponíveis, de suas características e de possíveis relações entre as variáveis. A partir dessa análise, é possível desenvolver uma estratégia de modelagem estatística ou de aprendizado de máquina adequada, bem como identificar possíveis *outliers* ou inconsistências nos dados.

Morettin e Singer (2021) destacam a importância de métodos gráficos na AED. Diagramas de caixa (*boxplots*), histogramas e gráficos de dispersão são alguns dos muitos tipos de visualizações que podem ser utilizados para resumir e representar dados. Esses métodos

gráficos podem ajudar a identificar tendências, padrões e possíveis exceções nos dados, que podem não ser facilmente visíveis a partir de medidas resumidas como a média e o desvio padrão.

A AED não se limita apenas à visualização. Morettin e Singer (2021) também discutem a aplicação de técnicas estatísticas como medidas de tendência central e variabilidade, assim como o uso de métodos mais sofisticados, como análise de componentes principais (PCA) e agrupamento (*clustering*) para explorar a estrutura dos dados.

O livro enfatiza que a AED é um passo crítico em qualquer análise de dados, seja na ciência de dados, na estatística ou em qualquer outra disciplina que lide com a interpretação de dados. Ela fornece uma compreensão inicial e uma orientação para os passos subsequentes da análise.

2.4.1 Preparação dos dados

2.4.1.1 Tipos de dados

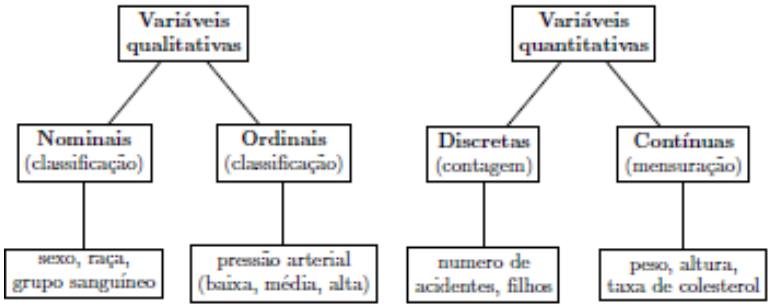
De acordo com Morettin e Singer (2021) os dados podem ser classificados em diferentes tipos, cada um com suas características e métodos de análise específicos:

- **Dados Quantitativos e Qualitativos (ou categóricos):**

Dados quantitativos são dados que representam quantidades, ou seja, podem ser medidos numericamente. Eles podem ser divididos em duas categorias: discretos e contínuos. Dados discretos são aqueles que podem assumir apenas um número finito ou uma sequência infinita enumerável de valores (por exemplo, o número de filhos de uma família). Dados contínuos, por outro lado, podem assumir qualquer valor em um intervalo específico (por exemplo, a altura de uma pessoa) (MORETTIN & SINGER, 2021).

Dados qualitativos (ou categóricos) são dados que representam características ou categorias. Eles também podem ser divididos em duas categorias: nominais e ordinais. Dados nominais são aqueles que não possuem uma ordem ou ranking inerente (por exemplo, cor dos olhos). Já os dados ordinais são aqueles que possuem uma ordem ou graduação (por exemplo, níveis de satisfação) (MORETTIN & SINGER, 2021).

Figura 5 – Classificação de variáveis



Fonte: Morettin & Singer (2021).

- **Dados Univariados, Bivariados e Multivariados:**

Dados univariados são conjuntos de dados que envolvem uma única variável. Por exemplo, a idade de um grupo de pessoas.

Dados bivariados são conjuntos de dados que envolvem duas variáveis. Por exemplo, a idade e a renda de um grupo de pessoas.

Dados multivariados são conjuntos de dados que envolvem três ou mais variáveis. Por exemplo, a idade, a renda e o nível de escolaridade de um grupo de pessoas (MORETTIN & SINGER, 2021).

2.4.1.2 Tratamento dos dados

Morettin e Singer (2021) destacam que essa etapa é fundamental para garantir que os dados sejam confiáveis e úteis para a análise. Envolve várias subetapas, incluindo a limpeza de dados, o tratamento de dados faltantes, a transformação de variáveis e a detecção de *outliers*.

- **Limpeza de dados:** esta etapa inclui a verificação da consistência dos dados, a correção de erros e a remoção de duplicatas. Isso pode envolver a verificação de erros tipográficos, inconsistências nas unidades de medida ou dados codificados de forma errada (MORETTIN & SINGER, 2021).
- **Tratamento de dados faltantes:** como discutido anteriormente, os dados faltantes podem ser tratados de várias maneiras, como imputação média, imputação por regressão ou imputação múltipla, entre outras. A escolha do método apropriado depende da natureza dos dados e do tipo de análise a ser realizada (MORETTIN & SINGER, 2021).
- **Transformação de variáveis:** em alguns casos, as variáveis podem precisar ser transformadas para atender aos pressupostos de um determinado método estatístico ou para melhorar a interpretabilidade. Isso pode incluir a transformação logarítmica,

normalização, padronização, binarização, ou criação de variáveis *dummy* (MORETTIN & SINGER, 2021).

- **Detecção de *outliers*:** *outliers* são valores extremos que podem distorcer a análise se não forem adequadamente tratados. Métodos como gráficos de caixa, gráficos de dispersão ou métodos estatísticos mais avançados, como o método de Tukey ou o método *Z-score*, podem ser usados para detectar *outliers* (MORETTIN & SINGER, 2021).

Finalmente, Morettin e Singer (2021) salientam a importância de entender a fonte e o contexto dos dados durante todo o processo de tratamento de dados. Isso inclui compreender o método de coleta de dados, as possíveis fontes de erro e a relação entre as variáveis.

2.4.1.3 Tratamento de dados ausentes

Existem várias técnicas para lidar com dados faltantes, as quais podem ser escolhidas dependendo da natureza e da quantidade de dados ausentes:

- **Exclusão de dados:** esta é a abordagem mais simples, na qual as linhas ou colunas com dados faltantes são excluídas. No entanto, esta abordagem pode levar à perda de informações importantes, especialmente se a ausência de dados não é completamente aleatória (LITTLE & RUBIN, 2002).
- **Imputação média/mediana/moda:** para cada coluna, os valores faltantes são substituídos pela média (para dados contínuos), mediana (para dados contínuos não-normalmente distribuídos) ou moda (para dados categóricos). Esta é uma técnica comum, mas pode resultar em uma subestimação da variabilidade (SCHAFER & GRAHAM, 2002). Embora seja uma técnica rápida e fácil de implementar, Schafer & Graham (2002) advertem que a imputação média/mediana/moda pode resultar em uma subestimação da variabilidade, uma vez que está substituindo os valores ausentes por uma constante, e não está levando em conta a variabilidade natural dos dados. Isso pode levar a uma subestimação do erro e um excesso de confiança nas estimativas.
- **Imputação por regressão:** aqui, um modelo de regressão é usado para prever os valores ausentes com base em outras variáveis. Esta abordagem pode preservar melhor a estrutura dos dados, mas também pode introduzir erro devido à incerteza na previsão (ALLISON, 2002).

- **Métodos de imputação múltipla:** a imputação múltipla cria múltiplos conjuntos de dados preenchidos e as análises são realizadas em cada um desses conjuntos. Os resultados são então combinados para obter estimativas e erros padrão que incorporam a incerteza da imputação (RUBIN, 1987).
- **Modelos que lidam diretamente com dados faltantes:** alguns modelos de *machine learning*, como o XGBoost, podem lidar diretamente com dados faltantes. Eles têm uma estrutura interna que permite dividir os nós de uma maneira que lidam com valores ausentes (CHEN & GUESTRIN, 2016).

É importante salientar que o tratamento adequado de dados faltantes depende do contexto e da natureza da ausência de dados, seja ela completamente aleatória ou não.

Aprofundando um pouco mais no caso de exclusão de dados, Little e Rubin (2002) discutem três tipos principais de mecanismos de ausência de dados: *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) e *Missing Not at Random* (MNAR).

A exclusão completa de observações (também chamada de exclusão casuística) com dados ausentes pode ser aplicada quando a ocorrência dos dados ausentes *é Missing Completely at Random* (MCAR). Neste cenário, a ausência de dados é completamente aleatória e não está relacionada nem com os valores observados nem com os não observados.

No entanto, é importante notar que a exclusão casuística de dados faltantes é frequentemente desaconselhada, mesmo quando os dados são MCAR, porque essa abordagem pode levar a uma redução considerável na quantidade de dados disponíveis para análise. A eliminação de dados só é adequada se a proporção de dados faltantes for muito pequena (LITTLE & RUBIN, 2002).

Para os casos em que a ausência de dados é Missing at Random (MAR) ou Missing Not at Random (MNAR), a exclusão casuística de dados não é apropriada, pois pode introduzir vieses nas análises. Para esses casos, outras técnicas de tratamento de dados ausentes, como a imputação, são preferíveis.

2.4.1.4 Estatística descritiva

Conforme detalhado por Morettin e Singer (2021) a Estatística Descritiva é um conjunto de técnicas que auxiliam na descrição e resumo de conjuntos de dados, permitindo uma compreensão geral das características dos dados. As principais ferramentas de estatística descritiva incluem:

- **Medidas de Posição:** Morettin e Singer (2021) descrevem que essas medidas são ferramentas estatísticas que fornecem informações sobre a posição de valores particulares dentro de um conjunto de dados. Há as que dão uma ideia do valor central em torno do qual os dados estão distribuídos como média, mediana e moda e as que são usados para entender a distribuição dos dados, e são especialmente úteis quando se quer comparar diferentes conjuntos de dados ou identificar *outliers* como por exemplos os quantis (percentis, decis, quartis). Elas incluem:
    - **Quantis (Percentis, Decis, Quartis):** são pontos tomados em intervalos regulares da função de distribuição cumulativa de uma variável aleatória. Os quartis dividem a distribuição em quatro partes iguais, os decis em dez e os percentis em cem. Por exemplo, o percentil 30 (P30) é o número abaixo do qual se encontram 30% das observações. O segundo quartil (Q2) é o valor que divide os dados ao meio, também conhecido como a mediana.
    - **Média (ou média aritmética):** é calculada somando todos os valores no conjunto de dados e dividindo pelo número total de valores. A média pode ser usada para conjuntos de dados em que os valores são distribuídos simetricamente e não há muitos *outliers*. Uma desvantagem da média é que ela é muito sensível a *outliers* ou valores extremos no conjunto de dados.
    - **Mediana**: é o valor do meio em um conjunto de dados ordenado. Se o número total de valores é ímpar, a mediana é o valor no meio; se o número total de valores é par, a mediana é a média dos dois valores do meio. A mediana tem a vantagem de ser menos sensível a *outliers* ou valores extremos do que a média.
    - **Moda:** é o valor que aparece com mais frequência em um conjunto de dados. Pode haver nenhum, um, ou vários modos em um conjunto de dados. A moda é a única medida de centralidade que pode ser usada com dados nominais (categóricos).

Essas medidas fornecem uma maneira de resumir um conjunto de dados com um único valor que representa um "centro" ou "localização típica" do conjunto de dados. No entanto, essas medidas não capturam a variabilidade ou dispersão dos dados, que é onde as medidas de dispersão entram (MORETTIN & SINGER, 2021).

- **Medidas de Dispersão:** conforme definidas por Morettin e Singer (2021), descrevem o grau de variabilidade em um conjunto de dados. Elas fornecem *insights* sobre o quão espalhados estão os dados em relação à média. As medidas de

dispersão mais comumente usadas são o desvio padrão, a variância, a amplitude e o intervalo interquartil.

- **Desvio Padrão:** é uma medida que expressa a quantidade de variação ou dispersão de um conjunto de valores. Um desvio padrão baixo indica que os valores estão próximos da média (ou esperado), enquanto um alto desvio padrão indica que os valores estão espalhados em uma ampla gama.
- **Variância:** é a média dos quadrados dos desvios da média em um conjunto de dados. É uma medida de dispersão que mostra o quão distantes cada valor neste conjunto está do valor central (médio). Assim como o desvio padrão, a variância maior sugere maior dispersão dos dados.
- **Amplitude:** é a diferença entre o valor máximo e mínimo em um conjunto de dados. É a medida de dispersão mais simples e dá uma noção rápida do grau de dispersão dos dados. Contudo, é altamente sensível a *outliers*.
- **Distância Interquartil:** é a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1) e descreve a dispersão da metade central dos dados. O IQR é uma medida robusta de dispersão, pois não é afetada por *outliers* ou valores extremos.

- **Tabelas de Frequência:** essas são tabelas que listam as categorias de dados e a frequência (o número de vezes) que cada categoria ocorre. As tabelas de frequência podem ser úteis para entender a distribuição dos dados.

- **Gráficos:** várias formas de gráficos e visualizações são usadas na estatística descritiva para representar dados. Isso inclui gráficos de barras, histogramas, gráficos de setor, *boxplots* e diagramas de dispersão. Cada tipo de gráfico tem suas próprias aplicações e pode ser usado para visualizar diferentes aspectos dos dados.

- **Correlação:** Morettin e Singer (2021) descrevem a correlação como uma medida estatística que expressa o grau de relação linear entre duas variáveis. A força da correlação é medida pelo coeficiente de correlação, geralmente denotado por 'r'. O coeficiente de correlação varia de -1 a +1. Um valor de correlação próximo de 1 indica uma forte correlação positiva, isto é, aumentam ou diminuem juntas, enquanto um valor próximo de -1 indica uma forte correlação negativa, ou seja, quando uma variável aumenta e a outra diminui. Um valor próximo a zero representa

uma correlação nula, não apresentando relação discernível entre duas variáveis. Vale ressaltar que correlação não implica causalidade, ou seja, mesmo que duas variáveis estejam fortemente correlacionadas, isso não significa que uma causa a outra.

Através dessas ferramentas, a estatística descritiva pode fornecer *insights* importantes e uma melhor compreensão de um conjunto de dados antes de avançar para análises mais complexas.

2.4.2 Análise de dados

2.4.2.1 Análise univariada

De acordo com Morettin e Singer (2021), a análise de dados conhecida como análise univariada, é o processo de coleta, sumarização e interpretação de dados de uma única variável. Nesta análise, a primeira etapa envolve a exploração dos dados brutos. Para dados quantitativos, isso pode ser feito usando um histograma ou ECDF (*Empirical Cumulative Distribution Function*), gráficos que mostram a distribuição dos dados. Para dados categóricos, um gráfico de barras pode ser mais apropriado (MORETTIN & SINGER, 2021).

Um gráfico ECDF (*Empirical Cumulative Distribution Function*) mostra a proporção de dados que são menores ou iguais a cada valor distinto do conjunto de dados. Em outras palavras, para qualquer ponto “x” no gráfico, a ECDF em “x” é a proporção de pontos de dados na amostra que são menores ou iguais a “x” (CASELLA & BERGER, 2002).

Um histograma é um gráfico que representa a distribuição de frequências de um conjunto de dados. Ele é construído dividindo o intervalo de dados em um conjunto de intervalos e então contando quantas observações caem em cada intervalo. O eixo x em um histograma representa as categorias em que os dados foram divididos, enquanto o eixo y representa a frequência ou a densidade de frequência das observações dentro dessas categorias. O formato geral do histograma pode fornecer uma visão sobre a natureza da distribuição dos dados, incluindo sua simetria, picos e a presença de valores discrepantes (*outliers*) (MORETTIN & SINGER, 2021).

A segunda etapa envolve calcular algumas medidas estatísticas descritivas como medidas de centralidade (média, mediana, moda), medidas de dispersão (desvio padrão, variância, amplitude e o intervalo interquartil.

Finalmente, a interpretação dos resultados envolve a compreensão do que essas estatísticas representam em termos do conjunto de dados. Por exemplo, uma grande variância

indica que os dados estão muito espalhados em torno da média, enquanto uma pequena variância indica que os dados estão próximos da média (MORETTIN & SINGER, 2021).

Em resumo, a análise univariada fornece uma visão detalhada de cada variável isoladamente, ajudando a entender as principais tendências, padrões e *outliers* em cada uma delas.

2.4.2.2 Análise bivariada

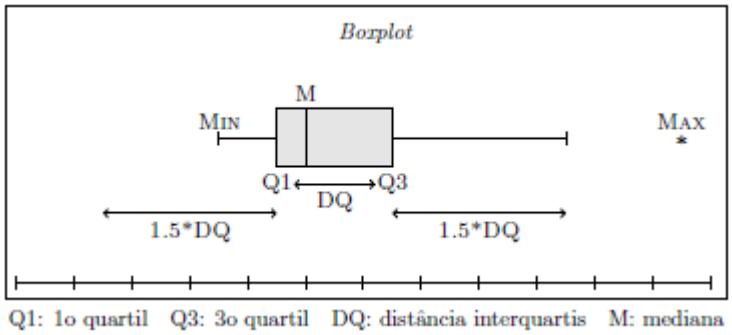
A análise bivariada investiga as relações entre duas variáveis e pode fornecer *insights* sobre como as variáveis estão associadas ou inter-relacionadas.

Na análise de dados bivariados envolvendo uma variável qualitativa e outra quantitativa o objetivo é entender como uma variável quantitativa se comporta em diferentes grupos definidos pela variável qualitativa.

De acordo com Morettin e Singer (2021) uma técnica comum utilizada para tratar este caso são os Gráficos de Caixa (*Boxplots*). Este é um gráfico que visualiza a mediana, quartis e possíveis *outliers* de uma variável quantitativa para diferentes categorias de uma variável qualitativa. Um *boxplot* (Figura 6) possui os seguintes componentes:

- **A caixa:** representa o intervalo interquartil (IIQ), que abrange do primeiro quartil (Q1, que é o valor no qual 25% dos dados estão abaixo) ao terceiro quartil (Q3, o valor abaixo do qual 75% dos dados se encontram). O tamanho da caixa dá uma indicação da dispersão dos dados.
- **A linha mediana:** indica a mediana dos dados (Q2), o valor no meio dos dados quando ordenados.
- **Os bigodes:** são as linhas que se estendem da caixa, indicam a variabilidade dos dados fora do intervalo interquartil. Normalmente, eles vão até o valor mais distante dentro de 1.5 vezes o IIQ a partir da caixa. Os pontos que estão além dos bigodes são considerados *outliers* e podem ser representados individualmente.
- ***Outliers*:** são observações que são notavelmente diferentes do restante dos dados. Eles são geralmente representados por pontos ou asteriscos

Figura 6 – Detalhes para a construção de *boxplots*



Fonte: Morettin & Singer (2021).

Já na análise de dados bivariados envolvendo duas variáveis quantitativas o objetivo é entender como duas variáveis numéricas se relacionam e influenciam uma à outra.

Segundo Morettin e Singer (2021) uma técnica comum utilizada para tratar este caso são os Gráficos de Dispersão (*scatterplot*). Cada ponto no gráfico representa um registro nos dados, com as coordenadas do ponto representando os valores das duas variáveis. O eixo horizontal normalmente representa a variável independente (ou explicativa), enquanto a variável dependente (ou resposta) é representada no eixo vertical. Os gráficos de dispersão permitem identificar várias características das relações entre as duas variáveis, incluindo:

- **Direção:** a relação entre as duas variáveis pode ser positiva (à medida que uma variável aumenta, a outra também aumenta) ou negativa (à medida que uma variável aumenta, a outra diminui).
- **Forma:** o gráfico pode revelar se a relação é linear (os pontos formam uma linha reta) ou não-linear (os pontos formam uma curva).
- **Força:** a força da relação é dada pela proximidade dos pontos à linha ou curva que melhor se ajusta a eles. Se os pontos estiverem bem próximos a essa linha ou curva, a relação é forte. Se estiverem dispersos, a relação é fraca.
- ***Outliers*:** observações que não se encaixam na tendência geral do restante dos dados podem ser facilmente identificadas em um gráfico de dispersão.

Por fim, na análise de dados bivariados envolvendo duas variáveis qualitativas o objetivo é entender como duas variáveis categóricas se relacionam. Conforme Morettin e Singer (2021) destacam, este tipo de análise envolve principalmente a construção e interpretação de tabelas de contingência.

Uma tabela de contingência é uma tabela que resume a relação entre duas variáveis categóricas. Cada entrada na tabela corresponde ao número de observações que correspondem a uma determinada combinação das categorias das duas variáveis. A partir de uma tabela de

contingência, é possível calcular proporções e porcentagens para entender a relação entre as variáveis. Por exemplo, pode-se calcular a porcentagem de observações em uma categoria da variável A que estão em uma categoria específica da variável B.

Outras métricas que podem ser calculadas a partir de uma tabela de contingência incluem o teste qui-quadrado para a independência, que testa a hipótese nula de que as duas variáveis são independentes (ou seja, não há relação entre elas).

A análise de dados bivariados com duas variáveis categóricas também pode incluir a construção de gráficos de barras ou gráficos de mosaico para visualizar a relação entre as variáveis.

2.4.2.3 Análise multivariada

A análise de dados multivariada, conforme descrita por Morettin e Singer (2021) em "Estatística e Ciência de Dados", envolve a observação e análise de mais de duas variáveis simultaneamente. Essa análise permite entender como as variáveis interagem e afetam umas às outras, ao invés de apenas olhar para elas isoladamente.

Quando há mais de duas variáveis, a representação gráfica das relações entre elas se torna mais complexa (MORETTIN & SINGER, 2021). Entretanto, há alguns gráficos que podem auxiliar, por exemplo, o Gráfico de Dispersão Simbólico.

O Gráfico de Dispersão Simbólico é uma técnica de visualização multivariada que é usada para visualizar a relação entre três ou mais variáveis. Morettin e Singer (2021) explicam que este tipo de gráfico funciona ao plotar duas variáveis no eixo x e y, de forma semelhante a um gráfico de dispersão regular, mas em vez de usar um ponto para representar cada observação, usa símbolos cujas propriedades (como tamanho, cor, ou forma) representam uma ou mais variáveis adicionais.

Por exemplo, pode-se traçar duas variáveis no eixo x e y, e usar o tamanho do símbolo para representar uma terceira variável e a cor do símbolo para representar uma quarta variável. Desta forma, o gráfico de dispersão simbólico pode fornecer uma maneira eficaz de visualizar a relação entre três ou quatro variáveis ao mesmo tempo.

No entanto, como Morettin e Singer (2021) apontam, os gráficos de dispersão simbólicos podem se tornar difíceis de interpretar se houver muitos pontos ou se as variáveis adicionais representadas pelos símbolos têm muitos valores únicos. Portanto, é importante usar essa técnica de forma criteriosa e sempre acompanhar o gráfico com uma explicação clara do que cada símbolo representa.

2.5 Aprendizado de Máquina (*Machine Learning*)

O aprendizado de máquina é um subcampo da Inteligência Artificial que se concentra na criação de sistemas capazes de aprender a partir de dados, sem que seja necessário programá-los explicitamente para cada tarefa. Isso permite que os sistemas melhorem automaticamente seu desempenho com a experiência, adaptando-se a novos dados e a mudanças no ambiente.

Morettin e Singer (2021) definem o aprendizado de máquina como a ciência (e a arte) de programar computadores para que eles possam aprender a partir dos dados. Isso envolve alimentar os computadores com dados e deixá-los ajustar e melhorar os modelos por conta própria.

Essa abordagem tem várias vantagens. Em primeiro lugar, os sistemas de aprendizado de máquina podem lidar com uma grande quantidade de dados e descobrir padrões complexos que seriam muito difíceis, se não impossíveis, para um humano identificar. Em segundo lugar, eles podem se adaptar a mudanças e aprender a partir de novos dados, tornando-os adequados para muitos problemas do mundo real.

Os algoritmos de aprendizado de máquina são usados em uma variedade de aplicações, desde sistemas de recomendação em sites de comércio eletrônico e serviços de *streaming*, até diagnósticos médicos, previsões meteorológicas e negociação algorítmica no mercado de ações.

Aprendizado de máquina é um campo vasto e em rápido crescimento, com uma variedade de técnicas e algoritmos disponíveis, cada um com suas próprias forças e fraquezas. A escolha do algoritmo adequado depende da natureza dos dados disponíveis e do problema específico que se está tentando resolver.

2.5.1 Tipos de aprendizado

Segundo Morettin e Singer (2021), existem três tipos principais de aprendizado de máquina: supervisionado, não supervisionado e por reforço.

1. **Aprendizado supervisionado:** o algoritmo é treinado em um conjunto de dados pré-rotulado, isto é, para cada entrada no conjunto de dados, a saída desejada é conhecida. O objetivo é que o algoritmo aprenda uma função que, dada uma nova entrada, possa prever a saída correta. Exemplos de algoritmos de aprendizado supervisionado incluem regressão linear e logística, máquinas de vetores de suporte, árvores e florestas e redes neurais.

2. **Aprendizado não supervisionado:** diferente do supervisionado, lida com conjuntos de dados onde as saídas não são conhecidas. O algoritmo é deixado para identificar padrões e estruturas nos dados por conta própria. Isso é frequentemente usado para agrupar dados semelhantes juntos ou para reduzir a dimensionalidade dos dados. Algoritmos de aprendizado não supervisionado incluem *clustering k-means*, hierárquico e análise de componentes principais (PCA).
3. **Aprendizado por reforço**: o algoritmo aprende através de tentativa e erro, realizando ações e recebendo recompensas ou punições. O objetivo é aprender uma política de ação, que é uma função que mapeia o estado do ambiente para as ações que o agente deve tomar.

Cada um desses tipos de aprendizado de máquina tem suas próprias aplicações e desafios. A escolha do tipo de aprendizado de máquina a ser usado depende do problema específico em questão e dos dados disponíveis (MORETTIN & SINGER, 2021).

2.5.2 Tipos de modelo de Aprendizado Supervisionado

De acordo com Morettin e Singer (2021), existem dois tipos principais de problemas que o aprendizado supervisionado pode resolver: regressão e classificação.

Nos modelos de regressão, a variável de saída é um número real, como um preço ou uma temperatura. O objetivo do modelo é prever um valor contínuo. Exemplos de algoritmos de regressão incluem regressão linear, regressão polinomial e regressão de árvore de decisão.

Por outro lado, nos modelos de classificação, a variável de saída é uma categoria, como “spam” ou “não spam” para *e-mails*, ou “cão”, “gato” e “coelho” para imagens de animais. O objetivo do modelo é prever a classe de uma entrada.

Vale ressaltar que tanto os modelos de regressão quanto de classificação exigem um conjunto de dados rotulado para treinamento. Esses modelos aprendem com os erros, ou seja, a diferença entre a saída prevista e a saída real. O modelo ajusta seus parâmetros para minimizar a soma dos erros ao longo do conjunto de dados de treinamento. Assim, os modelos de aprendizado supervisionado são capazes de prever a saída para novas entradas após serem treinados (MORETTIN & SINGER, 2021).

2.5.2.1 Modelo de Regressão Linear Múltipla

A regressão linear múltipla, conforme explorado por Morettin e Singer (2021) é uma extensão da regressão linear simples e é empregada quando temos mais de um preditor ou variável independente para o modelo. Matematicamente, este modelo pode ser expresso pela Equação 1:

Equação 1 – Modelo de Regressão Linear Múltipla

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p * X_{ip} + \varepsilon_{i,\ i = 1,\ ...,\ n.},\ \text{onde:}$$

- Y é a variável dependente (a variável a ser prevista ou estimada);
- X_1, X_2, ..., Xn são as variáveis independentes (os preditores);
- β_0 é o termo constante (também conhecido como intercepto);
- β_1, β_2, ..., βn são os coeficientes dos preditores (representam a mudança na variável de resposta Y para uma unidade de mudança em X, mantendo todos os outros preditores constantes);
- ε é o erro aleatório (contém a variação de Y que não pode ser explicada pelas variáveis independentes).

Fonte: Elaborado pelo autor.

O objetivo principal do modelo de regressão linear múltipla é encontrar a melhor linha de ajuste que possa prever a variável dependente com o menor erro. Para isso, normalmente, se utiliza o método dos mínimos quadrados para estimar os coeficientes que minimizam a soma dos quadrados dos resíduos (a diferença entre os valores observados e os previstos).

Entretanto, é importante ter em mente que a regressão linear múltipla pressupõe que as variáveis independentes não estejam altamente correlacionadas entre si, uma condição conhecida como multicolinearidade.

2.5.2.2 Modelo de Regressão “Lasso”

Lasso (*Least Absolute Shrinkage and Selection Operator*) é um método de regressão que realiza tanto a seleção de variáveis quanto a regularização para melhorar a precisão e a interpretabilidade do modelo estatístico produzido (TIBSHIRANI, 1996). Especificamente, o Lasso utiliza um parâmetro de penalidade que determina a quantidade de regularização a ser aplicada e é aplicado para minimizar a soma dos quadrados dos resíduos, semelhante à regressão linear padrão, mas com uma restrição adicional.

Essa restrição consiste na soma dos valores absolutos dos coeficientes do modelo, multiplicados pelo parâmetro de penalidade, sendo menor ou igual a um valor fixo. Isso tem o efeito de forçar alguns dos coeficientes estimados a serem exatamente zero quando o parâmetro de penalidade é suficientemente grande (JAMES et. al, 2013). Isso significa que o Lasso não só ajuda a evitar o sobreajuste, mas também realiza a seleção de variáveis.

Em termos simples, o Lasso diminui a complexidade do modelo ao penalizar os coeficientes de regressão com o valor absoluto de seus próprios coeficientes, ou seja, realiza um encolhimento dos coeficientes de regressão. Isso faz do Lasso uma abordagem particularmente útil para situações com conjuntos de dados de alta dimensão, nas quais a seleção de variáveis se torna crucial.

2.5.2.3 Modelo de Árvore de Decisão

As Árvores de Decisão são uma abordagem de aprendizado supervisionado usada tanto para classificação quanto para problemas de regressão. Conforme descrito por Morettin e Singer (2021), elas são chamadas de "árvores de decisão" porque começam com um único nó (a raiz), que se ramifica em possíveis resultados. Cada um desses resultados leva a nós adicionais, que por sua vez podem se ramificar em outros resultados. Isso continua até que um nó final (um "nó folha") seja alcançado, que fornece a previsão do resultado.

A principal vantagem das árvores de decisão é sua interpretabilidade. As regras de decisão inferidas a partir de uma árvore de decisão treinada são facilmente compreendidas, e o processo de tomada de decisões pode ser facilmente visualizado, o que não é sempre o caso com outros tipos de modelos de aprendizado de máquina.

No processo de construção de uma árvore de decisão, o algoritmo faz uma série de decisões, procurando pelo melhor recurso para dividir os dados com base em algum critério de avaliação. Para os problemas de classificação, a impureza de Gini ou a entropia são comumente usadas como critérios. Para problemas de regressão, o erro quadrado médio é frequentemente usado.

Embora as árvores de decisão sejam intuitivas e a sua interpretação seja relativamente simples, elas são frequentemente propensas ao sobreajuste (*overfitting*) quando há muitos recursos e a árvore cresce muito profunda. A flexibilidade do procedimento é tamanha que a árvore final pode ter um número de nós terminais igual ao número de observações, resultando em uma árvore onde cada instância do conjunto de treinamento é classificada de maneira

perfeita (MORETTIN & SINGER, 2021). Para combater isso, métodos como a poda da árvore são usados, ou então são utilizados métodos de conjuntos como florestas aleatórias ou *boosting*.

As “podas” são um tipo de hiperparâmetro. No contexto de aprendizado de máquina, hiperparâmetros são os parâmetros que são definidos antes do início do processo de treinamento de um modelo e que governam o próprio processo de treinamento. Em outras palavras, enquanto os parâmetros do modelo são aprendidos durante o treinamento - como os pesos em uma rede neural ou os coeficientes em uma regressão linear - os hiperparâmetros são definidos de antemão (GOODFELLOW, BENGIO & COURVILLE, 2016).

2.5.2.4 Modelo de Florestas Aleatórias (*Random Forest*)

Florestas Aleatórias, ou *Random Forests*, são um tipo de modelo de aprendizado de máquina que é uma extensão do modelo de árvores de decisão. Em termos simples, uma floresta aleatória é uma coleção de árvores de decisão. Em vez de confiar em uma única árvore de decisão, o modelo de Florestas Aleatórias toma a decisão com base nas previsões de várias árvores de decisão, fazendo uma espécie de “votação” entre as previsões das diferentes árvores.

De acordo com Morettin e Singer (2021), a ideia por trás das florestas aleatórias é combinar muitas árvores de decisão de alta variância e baixo viés para criar um modelo final que não apenas tenha baixo viés, mas também baixa variância. Cada árvore é construída usando uma amostra de *bootstrap* do conjunto de dados, e a divisão em cada nó é determinada por um subconjunto aleatório de preditores. Isso resulta em árvores altamente decorrelacionadas, o que é útil porque as árvores de decisão, embora sejam poderosas, têm a tendência de sobreajustar os dados de treinamento.

Portanto, uma Floresta Aleatória pode ser considerada como uma forma de realizar uma aprendizagem conjunta, a qual combina múltiplos modelos de aprendizado de máquina para criar um modelo mais potente e preciso. Como cada árvore é construída de maneira um pouco diferente, a floresta como um todo tem uma visão mais variada e mais completa dos dados de treinamento.

As Florestas Aleatórias são, portanto, um poderoso algoritmo de aprendizado de máquina que utiliza técnicas de *bagging* e aleatoriedade de atributos para formar um conjunto de árvores de decisão com desempenho superior. Com boa capacidade de generalização, as florestas aleatórias podem ser usadas tanto para tarefas de classificação quanto para regressão (MORETTIN & SINGER, 2021).

2.5.3 Hiperparâmetros

Hiperparâmetros podem ter um impacto significativo no desempenho de um modelo. Eles podem controlar aspectos como a complexidade do modelo, a velocidade de aprendizado, a regularização, entre outros. No caso de modelos de árvore de decisão e florestas aleatórias, alguns dos hiperparâmetros mais comuns são:

- **Profundidade máxima da árvore (*max_depth*):** limita o número máximo de níveis que a árvore de decisão pode ter. Ajustar este hiperparâmetro pode ajudar a prevenir o sobreajuste se for definido um valor baixo, ou permitir que o modelo se ajuste mais aos dados se for definido um valor alto (HASTIE, TIBSHIRANI & FRIEDMAN, 2009).
- **Número mínimo de amostras para divisão de um nó (*min_samples_split*):** determina o número mínimo de amostras necessárias para que um nó interno possa ser dividido. Isso também pode ajudar a controlar o sobreajuste, pois impõe um limite na granularidade das divisões que a árvore pode fazer (HASTIE, TIBSHIRANI & FRIEDMAN, 2009).
- **Número mínimo de amostras por folha (*min_samples_leaf*):** semelhante ao min_samples_split, mas especifica o número mínimo de amostras que devem estar presentes em um nó folha. Isso também pode ajudar a evitar o sobreajuste ao impor uma restrição na granularidade das folhas da árvore (HASTIE, TIBSHIRANI & FRIEDMAN, 2009).
- **Número de árvores em uma floresta (*n_estimators*):** é específico para modelos de Floresta Aleatória e determina o número de árvores na floresta. Em geral, mais árvores podem resultar em um modelo mais robusto e menos propenso a sobreajuste, mas o treinamento do modelo se tornará computacionalmente mais exigente (BREIMAN, 2001).
- ***Max_features*:** o número de recursos a serem considerados ao procurar a melhor divisão. A pesquisa com menos recursos pode levar a maior aleatoriedade, o que pode ajudar a tornar o modelo mais robusto contra o sobreajuste (BREIMAN, 2001).

No caso de modelo de regressão Lasso, alguns dos hiperparâmetros mais comuns são:

- **alpha**: este é o parâmetro de regularização. Quanto maior o valor de alpha, mais forte é a regularização, o que leva a modelos mais simples e, portanto, mais propensos a subajuste. Um valor menor para alpha produz modelos mais

complexos, que podem sofrer de sobreajuste (HASTIE, TIBSHIRANI & FRIEDMAN, 2009).

- ***max_iter***: determina o número máximo de iterações para que o algoritmo tente convergir para uma solução ótima. Se o algoritmo não atingir a convergência após "*max_iter*" iterações, ele será interrompido (SCIKIT-LEARN DOCUMENTATION, 2023).
- ***tol***: estabelece a tolerância para os critérios de parada. Se a melhoria do modelo em uma iteração é menor que o valor definido por 'tol', o algoritmo para de executar (SCIKIT-LEARN DOCUMENTATION, 2023).
- ***selection***: determina como os coeficientes são atualizados durante o processo de treinamento. Ele pode assumir dois valores: 'cyclic', para o qual os coeficientes são percorridos em um ciclo predefinido, ou 'random', para o qual um coeficiente é atualizado aleatoriamente a cada vez (SCIKIT-LEARN DOCUMENTATION, 2023).

Vale a pena notar que a escolha e o ajuste dos hiperparâmetros pode ser um processo desafiador e demorado, pois a performance de um modelo pode ser sensível a eles e o espaço de busca pode ser vasto. Além disso, não existe um conjunto “universal” de hiperparâmetros que funcionará melhor para todos os problemas e conjuntos de dados, o que torna este processo mais de uma arte do que uma ciência (GOODFELLOW, BENGIO & COURVILLE, 2016).

2.5.4 *Grid Search* e *Cross Validation*

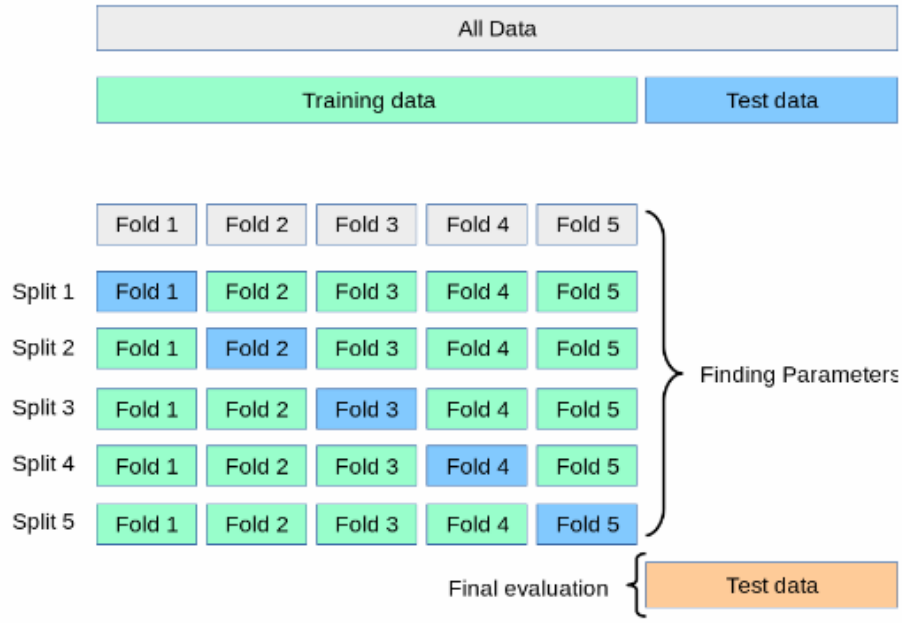
Ajustar hiperparâmetros é uma parte importante do processo de construção de um modelo de aprendizado de máquina. Uma maneira comum de fazer isso é por meio de técnicas como *Grid Search* ou *Random Search*, que envolvem a exploração sistemática de diferentes combinações de hiperparâmetros para encontrar a que produz o melhor desempenho no conjunto de validação ou por meio de validação cruzada (BERGSTRA & BENGIO, 2012).

No *Grid Search*, os pesquisadores definem um conjunto de possíveis valores para cada hiperparâmetro e o *Grid Search* examina todas as possíveis combinações desses valores. Para cada combinação, o modelo é treinado e avaliado usando uma métrica de desempenho específica (por exemplo, a acurácia para classificação ou o erro quadrado médio para regressão). Assim, a combinação de hiperparâmetros que produz o melhor desempenho será a melhor.

Embora o *Grid Search* seja uma abordagem de força bruta que pode ser computacionalmente intensiva, especialmente para modelos com muitos hiperparâmetros ou quando o conjunto de possíveis valores é grande, ele tem a vantagem de ser simples de entender e fácil de implementar. Além disso, ele garante que a melhor combinação de hiperparâmetros seja encontrada dentro do conjunto definido de possíveis valores (BERGSTRA & BENGIO, 2012). Há outras técnicas de ajuste de hiperparâmetros, como o *Random Search*, que procura valores aleatórios para os hiperparâmetros dentro de um intervalo definido, e que pode ser mais eficiente do que o *Grid Search* quando o número de hiperparâmetros é grande (BERGSTRA & BENGIO, 2012).

Comumente, é utilizada um método em conjunto ao *Grid Search* chamado Validação Cruzada (*Cross Validation*). Um tipo comum de validação cruzada é a validação cruzada k-fold (Figura 7), na qual os dados são divididos em "k" subconjuntos ou "*folds*". O modelo é então treinado "k" vezes, cada vez usando "k-1" subconjuntos como dados de treinamento e o subconjunto restante como dados de validação. A performance do modelo é então a média das performances em cada uma das "k" iterações (HASTIE, TIBSHIRANI & FRIEDMAN, 2009).

Figura 7 – Exemplo *5-Fold Cross Validation*



Fonte: Biblioteca Scikit Learn (2023).

2.5.5 *Feature importance*

A "importância dos recursos", ou "*feature importance*", é uma técnica que atribui uma pontuação para cada recurso (ou característica) de um modelo de aprendizado de máquina com

base em quão útil esse recurso é para prever a variável alvo. Nos modelos de Árvore de Decisão e Florestas Aleatórias, a importância do recurso é calculada com base em quão útil é um recurso para dividir os dados e, portanto, reduzir a impureza ou incerteza da variável alvo (BREIMAN, 2001).

No caso das Árvores de Decisão, a importância de um recurso é determinada pelo quanto esse recurso é capaz de diminuir a impureza total quando usado para dividir os dados.

No caso da Floresta Aleatória, a importância de um recurso é calculada de forma semelhante, mas é média entre todas as árvores na floresta. Isto é, para cada recurso, calcula-se a diminuição média na impureza do Gini (ou no erro quadrático médio, no caso de um problema de regressão) em todas as árvores onde esse recurso é usado para uma divisão. Em seguida, essas diminuições médias são normalizadas para que a soma das importâncias de todos os recursos seja 1 (BREIMAN, 2001).

Como mencionado por Breiman (2001), a importância dos recursos como calculada por Florestas Aleatórias tem a vantagem de ser uma medida "intrínseca", ou seja, não requer a construção de um modelo separado ou a alteração dos dados para ser calculada. No entanto, também é importante notar que a importância dos recursos não fornece uma medida da relação entre um recurso e a variável alvo, e pode ser influenciada por fatores como a correlação entre recursos e a escala dos recursos.

2.5.6 Validação de modelos

A validação permite verificar o desempenho do modelo em dados não vistos anteriormente, ajudando a garantir que o modelo é capaz de generalizar bem para novos dados, ao invés de apenas memorizar o conjunto de dados de treinamento - um fenômeno conhecido como sobreajuste ou *overfitting* (GOODFELLOW, BENGIO & COURVILLE, 2016).

A validação também fornece uma forma objetiva de comparar diferentes modelos ou diferentes configurações de hiperparâmetros, permitindo que os cientistas de dados selecionem o modelo ou configuração que é mais provável de ter o melhor desempenho em novos dados (HASTIE, TIBSHIRANI & FRIEDMAN, 2009).

Existem várias técnicas para a validação de modelos de aprendizado de máquina. Uma técnica comum é a divisão de treino/teste, na qual o conjunto de dados é dividido em um conjunto de treinamento, que é usado para treinar o modelo, e um conjunto de teste, que é mantido separado e usado para avaliar o desempenho do modelo. No entanto, essa técnica pode ser sensível à forma como os dados são divididos.

2.5.6.1 Divisão de base em treino e teste

A divisão de bases de dados em conjuntos de treinamento e teste é um dos procedimentos essenciais em projetos de *machine learning*. O conjunto de treinamento é utilizado para "ensinar" o modelo de *machine learning*. Isso significa que o modelo se esforça para identificar e aprender as relações entre os atributos (ou variáveis independentes) e o alvo (ou variável dependente) com base nos dados de treinamento. Por outro lado, o conjunto de teste não é utilizado durante a fase de treinamento e é empregado para avaliar o desempenho do modelo.

Hastie, Tibshirani e Friedman (2009) observam que a divisão de dados é feita para fornecer uma avaliação realista e imparcial do desempenho futuro do modelo. O conjunto de treinamento é usado para ajustar os parâmetros do modelo, enquanto o conjunto de teste serve como uma plataforma para avaliar a capacidade do modelo de generalizar os padrões aprendidos para novos dados não vistos.

Essa prática é crucial para evitar o sobreajuste (*overfitting*), um problema comum no aprendizado de máquina. Segundo Goodfellow, Bengio, & Courville (2016), o overfitting ocorre quando um modelo se ajusta muito bem aos dados de treinamento, a ponto de aprender até mesmo o ruído desses dados, resultando em um desempenho ruim ao lidar com novos dados. Por exemplo, ele pode capturar o ruído nos dados de treinamento, interpretando-o como uma tendência legítima. Nesse caso, embora o desempenho do modelo nos dados de treinamento seja excelente, seu desempenho nos dados de teste (dados não vistos em geral) é ruim.

Ao reservar uma parte dos dados para teste, tem-se uma maneira de avaliar a habilidade do modelo de generalizar para novos dados. Em outras palavras, a divisão ajuda a avaliar se o modelo aprendeu realmente os padrões nos dados ou se simplesmente memorizou o conjunto de treinamento. Essa é uma forma de validação conhecida como validação *hold-out*.

A proporção entre os dados de treinamento e teste pode variar, mas uma divisão comum é de 80% dos dados para treinamento e 20% para teste (KOHAVI, 1995). No entanto, essa proporção pode ser ajustada dependendo do tamanho e especificidades do conjunto de dados.

Em suma, a divisão dos dados em conjuntos de treinamento e teste é vital para desenvolver e validar modelos de *machine learning* robustos e generalizáveis.

2.5.7 Métricas de avaliação de desempenho para problemas de regressão

Na modelagem de regressão e no aprendizado de máquina, as métricas de avaliação de desempenho são cruciais para entender a qualidade do ajuste de um modelo aos dados. Para

problemas de regressão, erros e resíduos são conceitos chave usados nas métricas de avaliação. Os erros referem-se à diferença entre os valores previstos pelo modelo e os valores verdadeiros, enquanto os resíduos são as diferenças entre os valores observados e os valores previstos no conjunto de dados de treinamento.

2.5.7.1 Erro Quadrático Médio (MSE) e Erro Absoluto Médio (MAE)

Existem várias métricas de avaliação de desempenho para modelos de aprendizado de máquina, e a métrica correta a ser usada depende do tipo de problema e do objetivo específico do modelo. Aqui estão algumas das métricas de avaliação mais comuns para problemas de regressão:

- Erro Quadrático Médio (*Mean Squared Error* - MSE): é a média da soma dos erros quadrados entre as previsões e os valores reais (HASTIE, TIBSHIRANI & FRIEDMAN, 2009).
- Raiz do Erro Quadrático Médio (*Root Mean Squared Error* - RMSE): é a raiz quadrada do MSE.
- Erro Absoluto Médio (Mean Absolute Error - MAE): é a média da soma dos erros absolutos entre as previsões e os valores reais (WILLMOTT & MATSUURA, 2005).

A fórmula para o MSE pode ser visualizada na Equação 2:

Equação 2: Fórmula do Erro Quadrático Médio (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}[(y_i - \widehat{f}(\mathbf{x}_i)]^2$$

, onde:

- n é é o número de observações;
- y_i é o valor real;
- $f(x_i)$ é o valor previsto da resposta para a i-ésima observação.

Fonte: Morettin & Singer (2021).

Cada diferença (y_i - $f(x_i)$) é chamada de "resíduo" e representa o erro de previsão para a i-ésima observação. Ao elevar ao quadrado esses resíduos e calcular a média, obtém-se o MSE. A elevação ao quadrado tem o efeito de penalizar mais erros maiores mais do que menores. Isso significa que o MSE é sensível a *outliers* e tende a dar mais peso aos erros grandes (HASTIE, TIBSHIRANI & FRIEDMAN, 2009).

Uma das vantagens do MSE é que ele é fácil de calcular e compreender, no entanto, uma desvantagem é que ele pode ser difícil de interpretar diretamente, pois está em unidades quadradas. É comum usar a raiz quadrada do MSE, ou RMSE (*Root Mean Squared Error*), que tem as mesmas unidades que a variável de destino (HASTIE, TIBSHIRANI & FRIEDMAN, 2009).

Ao contrário do Erro Quadrático Médio (MSE), o MAE não eleva ao quadrado os erros e, portanto, não penaliza tanto os erros grandes. Isso significa que o MAE é menos sensível a *outliers* em comparação com o MSE. Essa característica faz do MAE uma escolha mais apropriada quando há *outliers* significativos nos dados (WILLMOTT & MATSUURA, 2005).

A fórmula para o MAE pode ser visualizada na Equação 3:

Equação 3: Fórmula do Erro Absoluto Médio (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

, onde:

- n é é o número de observações;
- y_i é o valor real;
- $\hat{y}_i$ é o valor previsto da resposta para a i-ésima observação.

Fonte: Willmot & Matsuura (2005).

2.5.8 Transformação de variáveis numéricas

Muitos modelos estatísticos partem da premissa de que os valores de uma ou mais variáveis possuem uma distribuição normal (MORETTIN & SINGER, 2021). Neste sentido, é muito comum no aprendizado de máquina normalizar os dados numéricos da base de dados.

A transformação de Yeo-Johnson é uma técnica de transformação de dados que é uma extensão da transformação de Box-Cox. Ambos os métodos são usados para normalizar os dados numéricos, tornando-os mais adequados para o uso em certos algoritmos de aprendizado de máquina. O objetivo dessas transformações é tornar a distribuição dos dados mais próxima de uma distribuição normal.

A transformação de Box-Cox só pode ser aplicada a dados positivos, enquanto a transformação de Yeo-Johnson é mais flexível e pode ser aplicada a dados com valores zero ou negativos (YEO & JOHNSON, 2000).

Outra normalização também utilizada é a Quantílica. Esta abordagem se mostra bastante útil ao lidar com distribuições assimétricas, distorcidas ou com a presença de *outliers*, pois mapeia os dados para sua distribuição quantílica (HUANG, LI, & LIU, 2008).

A transformação quantílica classifica os dados e, em seguida, substitui os valores originais pela sua posição percentual, ou quantil. Por exemplo, o valor mínimo no conjunto de dados é transformado para o valor 0, o valor máximo é transformado para 1, e todos os outros valores são mapeados para o seu quantil correspondente.

2.5.9 Transformação de variáveis categóricas

O processo de codificação *dummy*, também conhecido como codificação *one-hot*, é uma maneira comum de lidar com variáveis categóricas em aprendizado de máquina e estatística (HASTIE, TIBSHIRANI & FRIEDMAN, 2009).

Variáveis categóricas são aquelas que contêm um número limitado e geralmente fixo de categorias possíveis. No entanto, muitos algoritmos de aprendizado de máquina requerem variáveis numéricas como entrada, o que torna necessário algum tipo de transformação.

Na codificação *dummy*, cada categoria possível de uma variável categórica é convertida em uma nova variável binária (0 ou 1). Por exemplo, se há uma variável categórica "cor" com as categorias “vermelho”, “azul” e “verde”, a codificação *dummy* criará três novas variáveis binárias: “cor_vermelho”, “cor_azul” e “cor_verde” (MÜLLER & GUIDO, 2016).

A principal vantagem da codificação *dummy* é que ela mantém todas as informações das categorias originais sem impor uma ordem arbitrária (como aconteceria se as cores fossem codificadas como 1, 2 e 3). Isso pode ser importante em muitas situações, pois a imposição de uma ordem artificial pode levar a resultados distorcidos.

3 METODOLOGIA

Este capítulo tem como objetivo explicar o desenvolvimento do projeto, qual o embasamento e as metodologias seguidas, além de apresentar o contexto no qual ele está inserido. A metodologia é a espinha dorsal de qualquer estudo acadêmico, pois define como a pesquisa será conduzida e fornece a justificativa para as técnicas e métodos utilizados.

O capítulo está organizado da seguinte forma: a seção 3.1 detalha o método geral escolhido no trabalho e a justificativa para essa escolha, e a seção 3.2 descreve o início de aplicação desse método e todo o contexto no qual o trabalho foi criado.

Ao fornecer uma descrição clara e completa da metodologia, espera-se não apenas garantir a transparência e a replicabilidade da pesquisa, mas também contribuir para o corpo de conhecimento existente no campo, oferecendo *insights* sobre a aplicação de determinados métodos e técnicas em um contexto específico.

3.1 Método geral de projeto – CRISP-DM

Para apoiar o Trabalho de Formatura, foi utilizada a metodologia CRISP-DM. Ela é amplamente utilizada na academia e na indústria, inclusive no Banco A. Ela é comumente escolhida pelas empresas para esse tipo de trabalho por alguns motivos.

Em primeiro lugar, a flexibilidade do CRISP-DM é uma de suas principais vantagens. A metodologia pode ser aplicada a uma variedade de problemas de ciência de dados, em diferentes setores e escalas. Isso significa que pode ser adaptada a pequenos projetos com recursos limitados, bem como a grandes projetos complexos (SHEARER, 2000).

Em segundo lugar, o CRISP-DM enfatiza uma abordagem iterativa e incremental para projetos de ciência de dados. Essa abordagem permite que as equipes de ciência de dados refinem continuamente seus modelos e adaptem seus esforços à medida que novos dados e *insights* se tornam disponíveis. Isso é particularmente valioso em ambientes dinâmicos, onde os requisitos do projeto podem mudar ao longo do tempo (SHEARER, 2000). Neste sentido, essa metodologia é valiosa para o projeto em questão, pois um dos seus objetivos é ser sempre atualizado com novos dados de entrada.

Por fim, o CRISP-DM fornece orientações detalhadas para cada etapa do processo de ciência de dados, desde a compreensão dos objetivos e requisitos do negócio até a implantação e manutenção do modelo final. Isso fornece uma estrutura clara para as equipes seguirem, aumentando a eficiência e reduzindo o risco de erros (SHEARER, 2000).

3.2 Fase "Entendimento do negócio (*Business Understanding*)"

Como já explicado na revisão bibliográfica, o CRISP-DM é um modelo de processo adotado para projetos de ciência de dados. Ele consiste em seis fases principais: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação (SHEARER, 2000). Cada uma delas possui suas particularidades que foram abordadas durante o desenvolvimento do projeto e estão aqui explicitadas. No item 3.2 deste Trabalho de Formatura, apenas será explicitada a primeira fase do modelo: Entendimento do Negócio (*Business Understanding*). Como as outras etapas estão relacionadas diretamente aos dados, serão abordadas no capítulo 4 – Análise de Dados.

O Entendimento do Negócio é a fase inicial na qual se define claramente o problema a ser resolvido. Ela envolve a compreensão dos objetivos e requisitos do projeto a partir de uma perspectiva de negócios, e a conversão deste conhecimento em uma definição de problema de ciência de dados e um plano preliminar (SHEARER, 2000). Ela é composta por quatro tarefas principais: determinar os objetivos do projeto, avaliar a situação do ambiente em que o projeto será realizado, determinar os objetivos da mineração de dados e desenvolver o plano do projeto.

3.2.1 Determinação dos objetivos do projeto

Para esse tópico, foram pensados, em conjunto com os *stakeholders* do projeto, objetivos que se encaixam nas diretrizes SMART criadas por Doran (1981). Cada letra da sigla SMART representa uma característica que os objetivos devem possuir para aumentar a sua eficácia:

- **Específicos (*Specific*):** deve ser claro e bem definido. Isso significa que todos os envolvidos devem ter uma compreensão precisa do que o objetivo implica. Isso pode incluir a identificação do que precisa ser feito, quem precisa fazer isso, e onde precisa ser feito (DORAN, 1981);
- **Mensuráveis (*Measurable*)**: deve ter critérios claros de sucesso que possam ser medidos. Isso permite que o progresso seja rastreado e que se possa determinar quando o objetivo foi atingido (DORAN, 1981);
- **Alcançáveis (*Achievable*):** deve ser realista, considerando as habilidades, recursos e restrições existentes. Isso não significa que o objetivo não possa ser desafiador, mas deve ser possível de ser alcançado com esforço e dedicação (DORAN, 1981);

- **Relevantes (*Relevant*):** deve estar alinhado com as metas mais amplas da organização ou indivíduo. Isso ajuda a garantir que o esforço para alcançar o objetivo contribua para o progresso geral em direção a metas mais significativas (DORAN, 1981);
- **Temporais (*Time-bound*):** deve ter um prazo claro. Isso cria um senso de urgência que pode motivar a ação e ajuda evitar postergações (DORAN, 1981);

Como tratado na Introdução do Trabalho de Formatura, o projeto está inserido em um time dentro do Banco A que tem o objetivo de melhorar a experiência do cliente em diversas áreas. Nesta equipe, o autor teve como objeto de estudo a Central de Atendimento (*contact center*) do banco, área que lida direto com o cliente e suas necessidades.

A demanda inicial era observar formas de melhorar a satisfação dos clientes em relação ao atendimento. Como dentro do banco a satisfação do cliente é metrificada pelo Net Promoter Score, a questão a ser resolvido era "como podemos aumentar o NPS do *contact center*?", mas para isso, primeiro é necessário entender "o que afeta o NPS", sendo este o objetivo final do projeto.

O Banco A entende que para continuar sendo um líder de mercado, ele deve estrategicamente melhorar seu nível de serviço. Neste sentido, para reforçar esse pilar dentro da empresa, todos os funcionários do banco, do estagiário ao presidente, possuem uma parcela de suas metas anuais atreladas ao NPS.

Sendo assim, foi proposto um projeto de Ciência de Dados com prazo de 3 meses no qual seriam analisadas as bases disponíveis com as informações dos atendimentos realizados ao longo de alguns meses dentro do banco e esses dados sereiam tratados em alguns modelos estatísticos para identificar quais fatores mais impactam na nota final dada pelo cliente e, assim, propor planos de ação focados nesses fatores.

3.2.2 Avaliação do contexto

Esta tarefa envolve uma avaliação da situação atual, incluindo recursos disponíveis, restrições, suposições e condições que podem impactar o projeto. Isso pode envolver a análise da infraestrutura de TI existente, a qualidade dos dados disponíveis, os recursos humanos e a capacidade da organização de implementar as soluções propostas.

Para o trabalho em questão, havia um gargalo para contratar uma pessoa especialista em ciência de dados que pudesse se manter na estrutura organizacional da área após a finalização do projeto. Sendo assim, a decisão tomada foi de fornecer à pessoa já contextualizada pelo dia

a dia de trabalho da área de negócio (o Autor desse projeto) um contato da área de TI que pudesse tirar dúvidas e ser um ponto de apoio.

Em relação aos dados, eles vinham de bases sistêmicas e direto das equipes responsáveis por serem guardiãs dos dados, ou seja, não haviam dúvidas sobre a qualidade deles.

Em termos práticos, foi escolhida a utilização do Google Colab para criação dos códigos. O Google Colab é uma plataforma de notebook Jupyter hospedada na nuvem. Essa ferramenta fornece um ambiente acessível de codificação em Python, apresentando todas as bibliotecas necessárias para o projeto e poder de processamento suficiente, além de ser gratuito. Assim, esses foram os motivos que acabaram culminando no uso desta ferramenta somados ao fato da equipe já ter expertise no uso dela.

3.2.3 Determinação dos objetivos de mineração de dados

Com base nos objetivos do negócio e na avaliação da situação atual, os objetivos de mineração de dados devem ser definidos. Isso envolve a tradução dos objetivos de negócios em um problema de ciência de dados que pode ser abordado por técnicas de mineração de dados.

Neste contexto, visualizando a base de dados apresentada, foi identificado que tem-se a variável de saída desejada, o NPS dos atendimentos. Deste modo, conforme explicado no item 2.5.1, enquadra-se em um problema de Aprendizado Supervisionado. Além disso, pelo NPS ser um número real, o indicado é o uso de modelos de regressão.

Pensando neste tipo de problema apresentado, o objetivo da mineração de dados é testar alguns modelos estatísticos, identificar qual é o melhor deles e utilizá-lo como fonte de insumos para criação de um plano de ação para melhoria do NPS. Isso envolve:

- Identificar se o modelo é um modelo confiável;
- Identificar quais são as variáveis que mais influenciam a métrica de satisfação no modelo;
- Deixar o modelo o mais genérico possível para ser replicado quando a equipe tiver dados mais recentes/atualizados.

3.2.4 Desenvolvimento do plano do projeto

O plano de projeto deve detalhar a abordagem proposta, os passos a serem tomados, o cronograma e os recursos necessários. Isso proporciona um roteiro claro para a execução do projeto e serve como um ponto de referência para avaliar o progresso.

Como o trabalho dependia apenas do autor, foi elaborado um cronograma simples dentro do prazo estabelecido pelo seu gestor (6 meses):

Tabela 1 – Cronograma do projeto de *Machine Learning*

Semana	Tarefas
Semanas 1 a 12	• Estudar plataforma Google Colab • Estudar linguagem Python direcionada à Ciência de Dados (incluindo bibliotecas auxiliares)
Semanas 13 a 15	• Realizar análise Exploratória de Dados
Semanas 16 a 18	• Realizar preparação dos dados (Tratamento das variáveis e Divisão de Base Treino/Teste)
Semanas 19 a 21	• Realizar modelagem e avaliação dos resultados, revisando os códigos anteriores
Semanas 22 a 24	• Consolidar planos de ação/próximos passos

Fonte: Elaborado pelo autor.

Durante o planejamento ainda foi definido que seriam criados três diferentes *notebooks* (Anexo 1, 2 e 3), isto é, três páginas diferentes de códigos dentro do Google Colab seguindo as etapas do cronograma:

1. Análise Exploratória de Dados;
2. Preparação dos Dados;
3. Modelagem.

4 ANÁLISE DE DADOS

Este capítulo abordará a análise de dados realizada durante este projeto de graduação. Essa etapa é fundamental para o desenvolvimento do trabalho, pois é nesse momento que serão extraídas as informações mais relevantes dos dados coletados, contribuindo diretamente para a geração de *insights* e a tomada de decisões baseada em evidências.

O *framework* CRISP-DM (*Cross Industry Standard Process for Data Mining*) continuará sendo utilizado como guia para estruturar a análise, seguindo suas etapas de preparação dos dados, modelagem, avaliação e implementação (CHAPMAN et al., 2000).

A análise será apoiada por técnicas estatísticas e de aprendizado de máquina, cuja escolha será justificada de acordo com a natureza dos dados e os objetivos do estudo. Ao longo desta etapa, será feito o uso do ambiente de programação Python no Google Colab, devido à

sua acessibilidade, facilidade de uso e ampla gama de bibliotecas disponíveis para a ciência de dados. Para a manipulação de dados, foram utilizadas as bibliotecas Pandas e Numpy, para a visualização dos dados (gráficos, tabelas e árvores) Matplotlib.pyplot, Seaborn, Plotly.express e Graphviz. Finalmente, para processamento dos dados e modelagem foi utilizada a biblioteca Sklearn.

Finalmente, os resultados obtidos na análise serão apresentados e discutidos em detalhe. Serão destacadas as principais descobertas, bem como suas implicações para o problema em estudo. Além disso, será feita uma avaliação crítica dos métodos utilizados, abordando as limitações da análise e sugerindo possíveis melhorias para trabalhos futuros.

Espera-se que, ao final deste capítulo, o leitor tenha uma compreensão clara das técnicas de análise de dados aplicadas, dos resultados obtidos e de seu significado no contexto do estudo.

4.1 Fase "Entendimento dos Dados (*Data Understanding*)"

Esta fase é a segunda etapa no processo de mineração de dados do modelo CRISP-DM. Este estágio envolve uma análise aprofundada dos dados disponíveis para fornecer informações valiosas sobre a qualidade, estrutura e propriedades dos dados coletados (CHAPMAN et al., 2000).

Ao analisar os dados durante a etapa de "Entendimento dos Dados", é fundamental documentar as descobertas e quaisquer decisões ou suposições feitas com base nessas descobertas. Essa documentação será útil nas etapas subsequentes do projeto e pode ser uma referência valiosa para projetos futuros. As atividades típicas durante esta fase incluem: coleta de dados, descrição dos dados, análise exploratória dos dados e verificação da qualidade dos dados.

4.1.1 Coleta de Dados

Conforme discutido por Davenport e Harris (2007), sem uma coleta adequada e precisa de dados, o valor dos *insights* obtidos a partir da análise deles pode ser limitado. Os autores destacam que a coleta correta de dados envolve garantir que os dados sejam relevantes, completos, precisos e atuais. Por exemplo, dados irrelevantes podem levar a *insights* inúteis, enquanto dados incompletos ou imprecisos podem resultar em conclusões errôneas. Da mesma forma, se os dados não forem atualizados regularmente, eles podem não refletir a situação atual,

o que torna os *insights* derivados menos úteis para a tomada de decisões (DAVENPORT & HARRIS, 2007).

Além disso, Davenport e Harris (2007) salientam que a coleta correta de dados também envolve respeitar as leis e regulamentos de privacidade. As organizações devem garantir que têm permissão para coletar e usar os dados, e que estão tomando as medidas necessárias para proteger esses dados contra o uso indevido.

Neste sentido, foi recebido pelo autor diretamente da área de CRM (*Customer Relationship Management*) do *contact center* uma base sistêmica com dados sumarizados sem qualquer informação de clientes que possivelmente poderiam infringir a LGPD (Lei Geral de Proteção de Dados). Para tratar esses dados o autor foi responsável por toda a codificação com apoio de seu superior para orientação técnica e de gerenciamento das atividades.

4.1.2 Descrição dos dados

A base contém as seguintes características:

- Dados de 02/05/2022 até 31/10/2022, com linhas sumarizadas pela data dos atendimentos, senioridade dos atendentes, tipo de canal de atendimento e tipo de produto;
- Possui as seguintes colunas com suas descrições:
    - **Data_final:** data dos atendimentos
    - **Cargo**: nível de senioridade dos atendentes de contact center (júnior, pleno e sênior)
    - **Canal:** veículo de atendimento (telefone, e-mail ou chat)
    - **Produto:** assunto tratado nos atendimentos identificado por produto (cartão de crédito, conta corrente, seguros e demais produtos)
    - **Horas_atendimento_canal:**

      Tempo Médio de Atendimento (TMA) x Quantidade de atendimentos / Quantidade de Colaboradores (em horas)
    - **Qtd_atendimentos**: número de atendimentos realizados por todos os atendentes por canal em um dia
    - **Qtd_colaboradores_area**: número de atendentes por senioridade por canal no dia
    - **Tempo_medio_resposta_min:** tempo que o cliente espera para receber uma resposta do atendente (em minutos)

- **Tempo_medio_atendimento_min (TMA):** tempo médio que o atendente leva para realizar 1 atendimento (em segundos)
- **Nps_medio:** nota de NPS média dos atendentes no dia por canal por senioridade

Tabela 2 – Exemplo da base de dados utilizada no trabalho

Data_final	Cargo	Canal	Produto	Horas_atendimento_canal	Qtd_atendimentos	Qtd_colaboradores_area	Tempo_medio_resposta_min	TMA (segundos)	NPS_médio
2022-05-02	Júnior	Chat	Seguros	1,75	274	8	15,01	184,1	57
2022-05-02	Júnior	E-mail	Seguros	2,95	511	8	1317,5	166	39
2022-05-02	Júnior	Telefone	Seguros	1,15	280	8	1,98	118	35
2022-05-02	Pleno	Chat	Seguros	3	315	5	15,27	171,3	58
2022-05-02	Pleno	E-mail	Seguros	3,24	343	5	1299,65	170	33
2022-05-02	Pleno	Telefone	Seguros	0,79	126	5	2,26	113	35
2022-05-02	Sênior	Chat	Seguros	1,2	23	1	14,48	189,5	55
2022-05-02	Sênior	E-mail	Seguros	2,46	51	1	1256,54	172,8	59
2022-05-02	Sênior	Telefone	Seguros	1,42	46	1	1,41	112,4	67
2022-05-03	Júnior	Chat	Seguros	1,72	243	8	14,6	204,5	55
2022-05-03	Júnior	E-mail	Seguros	3,22	532	8	1265	174,4	31
2022-05-03	Júnior	Telefone	Seguros	1,16	296	8	1,08	112,7	35
2022-05-03	Pleno	Chat	Seguros	4,09	338	5	14,61	217,5	31
2022-05-03	Pleno	E-mail	Seguros	3,22	379	5	1252,43	153,1	36
2022-05-03	Pleno	Telefone	Seguros	1,66	265	5	2,77	112,8	45
2022-05-03	Sênior	Chat	Seguros	1,12	19	1	14,36	215,8	73
2022-05-03	Sênior	E-mail	Seguros	2,06	49	1	1285,22	152,5	66
2022-05-03	Sênior	Telefone	Seguros	1,65	54	1	2,66	110,2	67

Fonte: Elaborado pelo autor.

Figura 8 – Tipos de Dados da base (*Data Types*)

```
[ ]  df_analise.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6624 entries, 0 to 6623
Data columns (total 10 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   cargo                        6624 non-null   object
 1   canal                        6624 non-null   object
 2   horas_atendimento_canal      6624 non-null   float64
 3   qtd_atendimentos             6624 non-null   int64
 4   qtd_colaboradores_area       6624 non-null   int64
 5   nps_medio                    6624 non-null   int64
 6   tempo_medio_resposta_min     6624 non-null   float64
 7   data_final                   6624 non-null   datetime64[ns]
 8   produto                      6624 non-null   object
 9   tempo_medio_atendimento_min  6624 non-null   float64
dtypes: datetime64[ns](1), float64(3), int64(3), object(3)
memory usage: 517.6+ KB
```

Fonte: Elaborado pelo autor no Google Colab.

Como pode-se observar na Figura 8, a base de dados possui 6.624 observações (ou linhas) e 10 colunas e contém quatro tipos de dados: data, objeto, número racional e inteiro. As variáveis categóricas são “cargo”, “canal” e “produto” e elas possuem os seguintes valores:

- **Cargo:** “Júnior”, “Pleno” e “Sênior”
- **Canal:** “Chat”, “E-mail” e “Telefone”
- **Produto:** “Cartão de Crédito”, “Conta Corrente”, “Demais Produtos” e “Seguros”

4.1.3 Verificação da qualidade dos dados

Para averiguar a completude e a qualidade da base foram feitos alguns testes. O primeiro deles foi a verificação de valores ausentes e foi constatado que não havia nenhum valor ausente. O segundo foi a verificação de valores duplicados e também não havia nenhum valor duplicado. O terceiro foi verificar se os valores dos dados numéricos faziam sentido e descobriu-se que haviam 191 casos (2,9% da base) em que havia registro de NPS, mas não havia registro de atendimento. Quando observado no detalhe pode-se analisar que 184 dos 191 casos eram registros de “Sênior” no canal “Telefone” e assunto “Conta Corrente”, ou seja, todos os registros deste caso específico estavam com as variáveis “qtd_atendimentos” e “horas_atendimento_canal” iguais a zero.

Como esses 184 casos representam uma parcela significativa, sobretudo, uma parcela inteira de um público específico, foi escolhido um método simples de imputar dados para substituir a ausência deles, a imputação de mediana. Ela foi escolhida, pois ambas as variáveis

possuem dados contínuos não-normalmente distribuídos (SCHAFER & GRAHAM, 2002). A mediana de “qtd_atendimentos” para canal “Telefone” e assunto “Conta Corrente” imputado foi de “434,5” e a mediana de “horas_atendimento_canal” para os mesmos casos foi de “0,50292310475”.

Já para os 7 casos restantes, foi feita uma exclusão deles, pois atendem os dois critérios de Little e Rubin (2002), o primeiro de que são casos *Missing Completely at Random* (MCAR), isto é, a ausência de dados é aleatória e não está relacionada nem com os valores observados nem com os não observados, e o segundo de que representam uma proporção baixa de casos em relação ao todo.

4.1.4 Análise Exploratória dos Dados (AED)

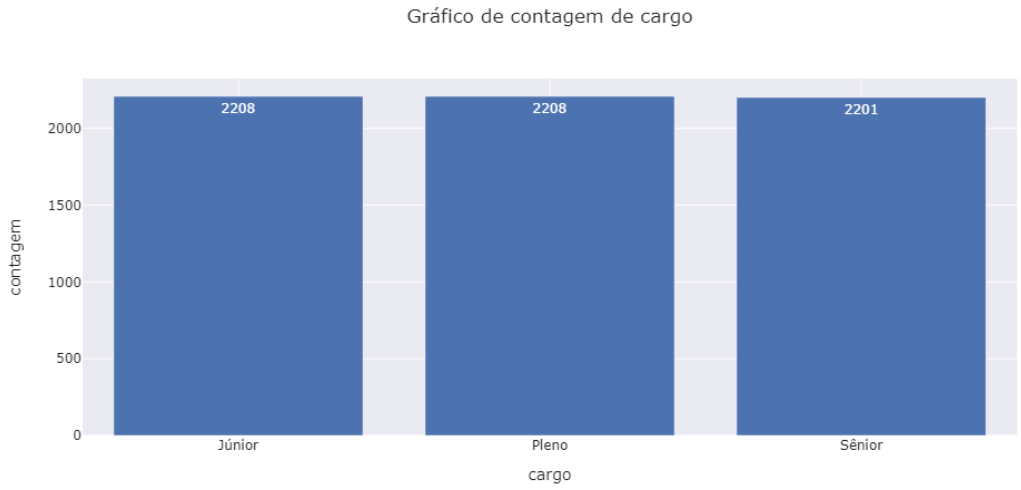
Segundo Morettin e Singer (2021) essa etapa da análise de dados envolve uma investigação minuciosa das características intrínsecas dos dados coletados. Por meio de gráficos, tabelas, medidas resumo e outros métodos, os pesquisadores podem obter uma visão geral do comportamento dos dados, identificar padrões, detectar anomalias e formular hipóteses. Este processo permite aos pesquisadores esclarecer quais métodos estatísticos são mais apropriados para um posterior e mais profundo processamento e análise dos dados. É, portanto, um pilar fundamental na ciência de dados e estatística, preparando o terreno para a inferência estatística e a modelagem preditiva.

Este item será dividido em três tipo de análises: univariada, bivariada e multivariada, considerando o tipo de dado, se ele é categórico (representam características) ou numérico (representam quantidades).

4.1.4.1 Análise Univariada

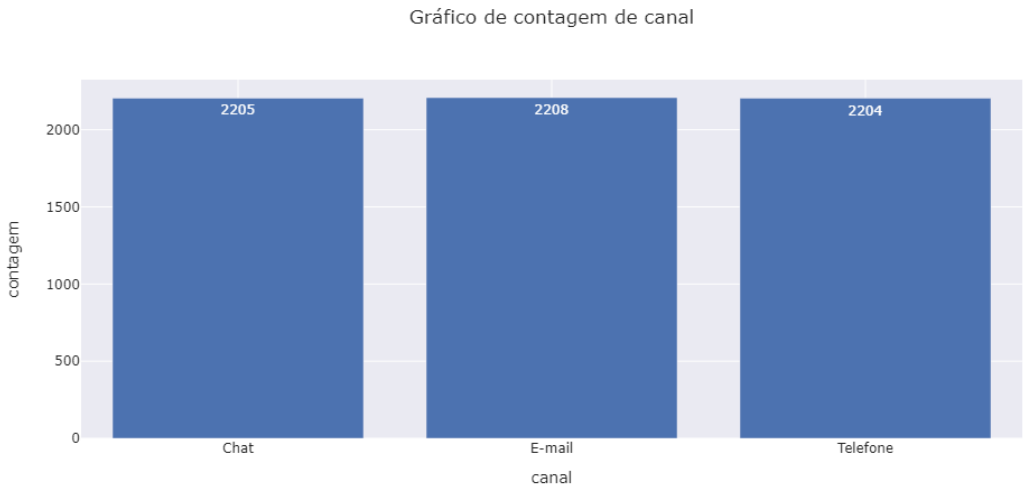
Para a análise univariada de variáveis categóricas foi realizado um estudo de contagem dos valores como pode ser visualizado nas Figuras 9, 10, 11 e 12. É de se observar que não há nenhuma anomalia nos gráficos, os números de contagem dos valores só não são exatamente iguais, pois existiram os sete casos de exclusão descritos na etapa anterior. Pode-se perceber que estes casos são todos do Cargo “Sênior”, três do Canal “Chat” e quatro do Canal “Telefone”, seis do Produto “Cartão de Crédito” e um do Produto “Conta Corrente” e todos são de datas distintas.

Figura 9 – Gráfico de contagem da variável “Cargo”



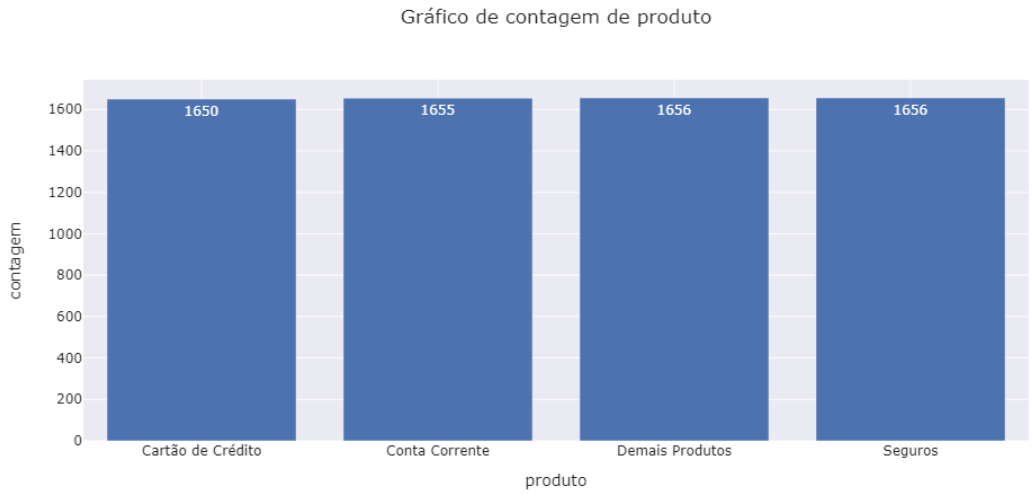
Fonte: Elaborado pelo autor no Google Colab.

Figura 10 – Gráfico de contagem da variável “Canal”



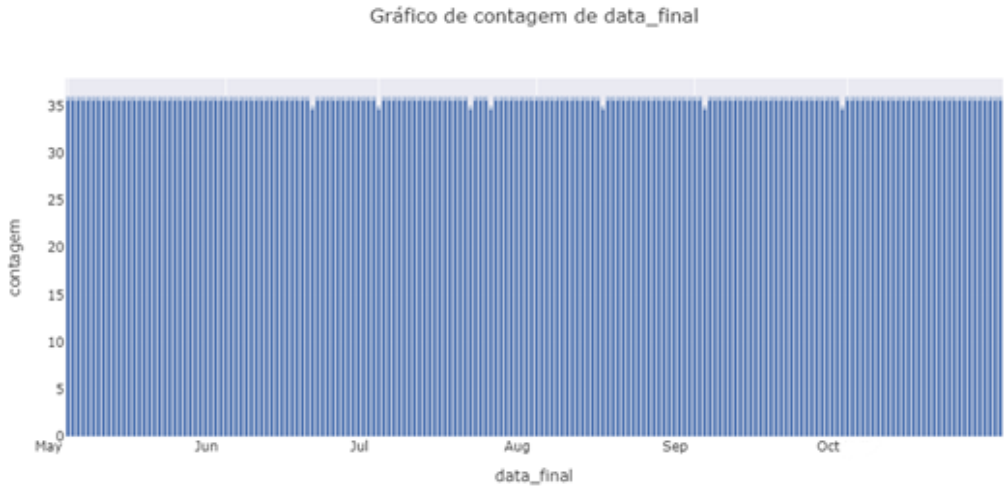
Fonte: Elaborado pelo autor no Google Colab.

Figura 11 – Gráfico de contagem da variável “Produto”



Fonte: Elaborado pelo autor no Google Colab.

Figura 12 – Gráfico de contagem da variável “data_final”



Fonte: Elaborado pelo autor no Google Colab.

Vale ressaltar que faz sentido os dados categóricos serem igualmente proporcionais, pois a base foi sumarizada em relação a eles.

Para as variáveis numéricas foram realizadas três tipos de análises: uma de estatística descritiva e duas gráficas. Observando a Tabela 3 pode-se constatar que também não há anomalias nos números apresentados. A amplitude (máximos e mínimos) fazem sentido dentro do contexto de cada variável. Algumas variáveis possuem desvios padrões elevados, mas isso também não indica nenhuma anomalia, só mostra que os dados estão mais dispersos em relação a centralidade.

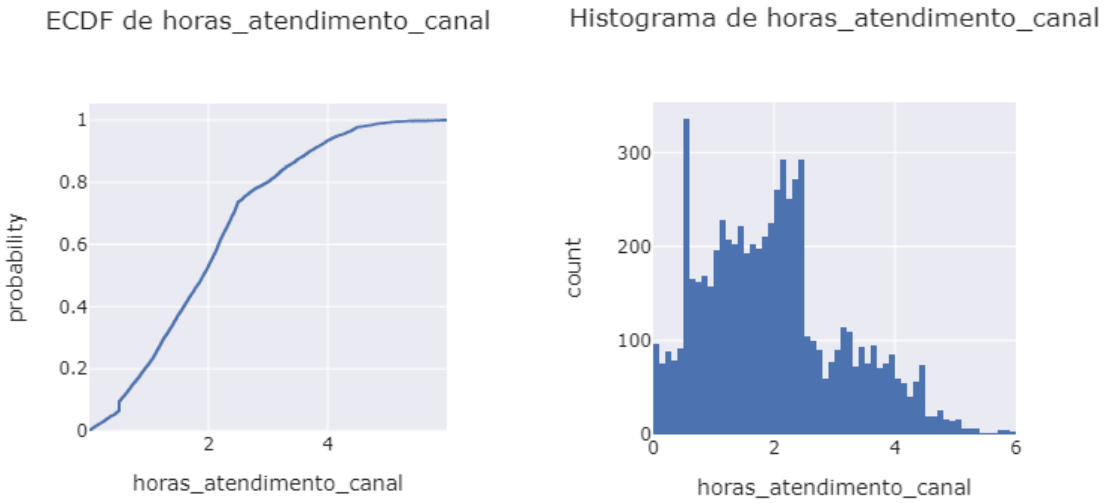
Tabela 3 – Dados de Estatística Descritiva das variáveis numéricas da base

	horas_atendimento_canal	qtd_atendimentos	qtd_colaboradores_area	nps_medio	tempo_medio_resposta_min	tempo_medio_atendimento_min
count	6617.000000	6617.000000	6617.000000	6617.000000	6617.000000	6617.000000
mean	1.994637	1022.086595	19.760163	70.426477	570.264330	2.574639
std	1.157508	1151.344236	14.238895	19.350700	992.172625	0.596608
min	0.002580	1.000000	1.000000	1.000000	1.000000	1.833333
25%	1.114315	291.000000	7.000000	57.000000	3.859678	1.916601
50%	1.912742	652.000000	18.000000	75.000000	9.123315	2.673317
75%	2.590000	1359.000000	29.000000	86.000000	997.575907	3.025864
max	5.997623	12553.000000	60.000000	100.000000	6864.573562	3.666131

Fonte: Elaborado pelo autor no Google Colab.

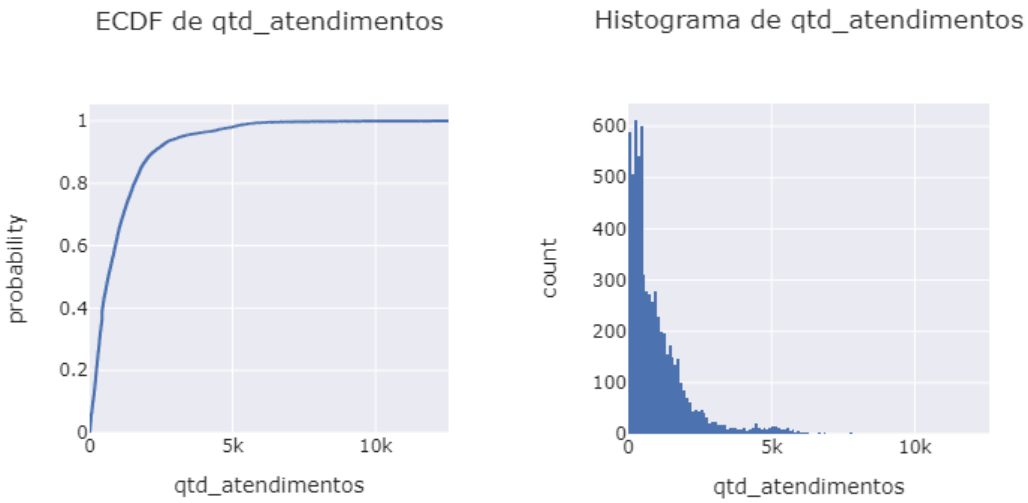
Na parte de análises gráficas foram feitos para cada uma das variáveis dois tipos de gráficos: ECDF e histogramas. Os gráficos ECDF representam a proporção da quantidade de dados acumulados (eixo y) em relação ao valor dele (eixo x) (CASELLA & BERGER, 2002), já os histogramas mostram a frequência de valores dentro de um conjunto de intervalos (MORETTIN & SINGER, 2021).

Figura 13 – Gráfico ECDF e Histograma de horas_atendimento_canal



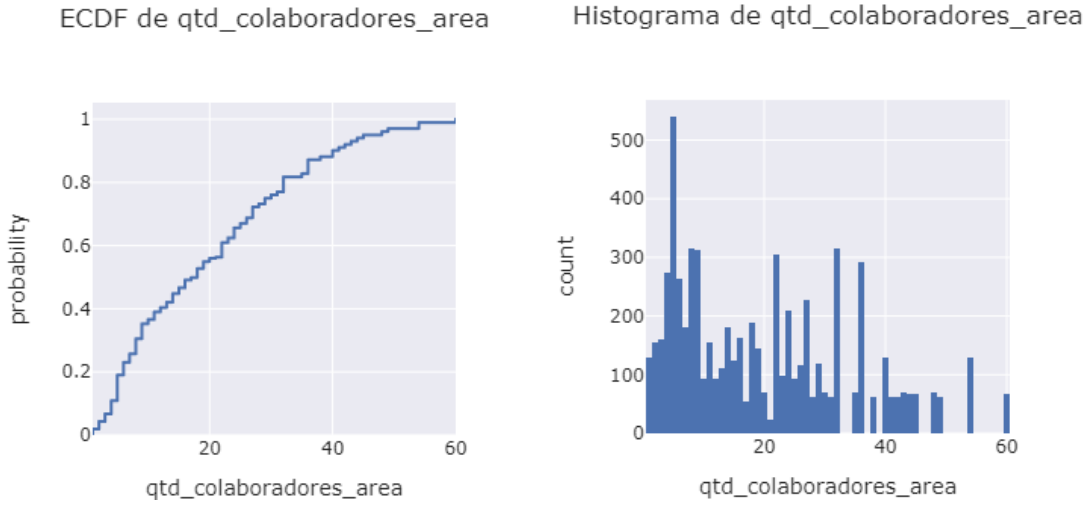
Fonte: Elaborado pelo autor no Google Colab.

Figura 14 – Gráfico ECDF e Histograma de qtd_atendimentos



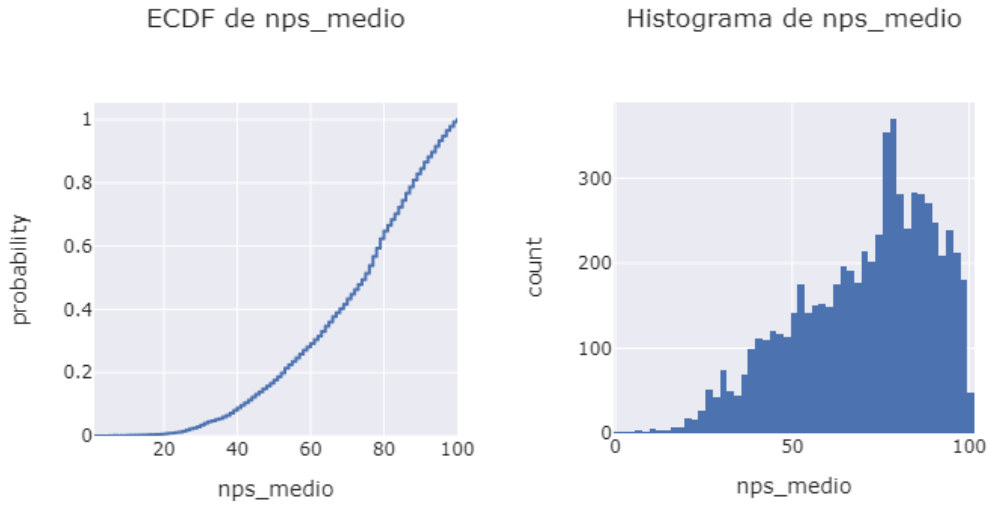
Fonte: Elaborado pelo autor no Google Colab.

Figura 15 – Gráfico ECDF e Histograma de qtd_colaboradores_area



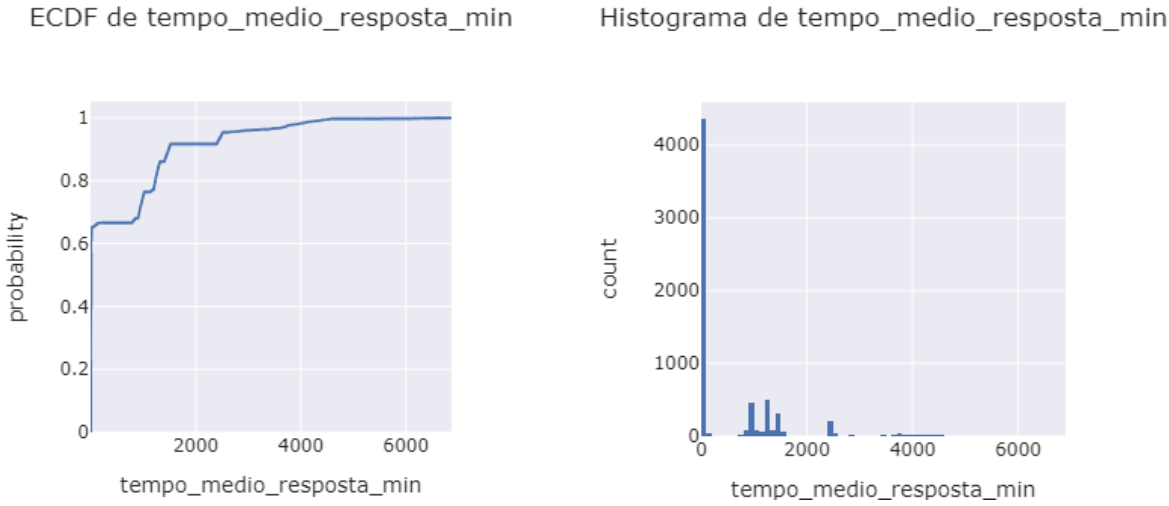
Fonte: Elaborado pelo autor no Google Colab.

Figura 16 – Gráfico ECDF e Histograma de nps_medio



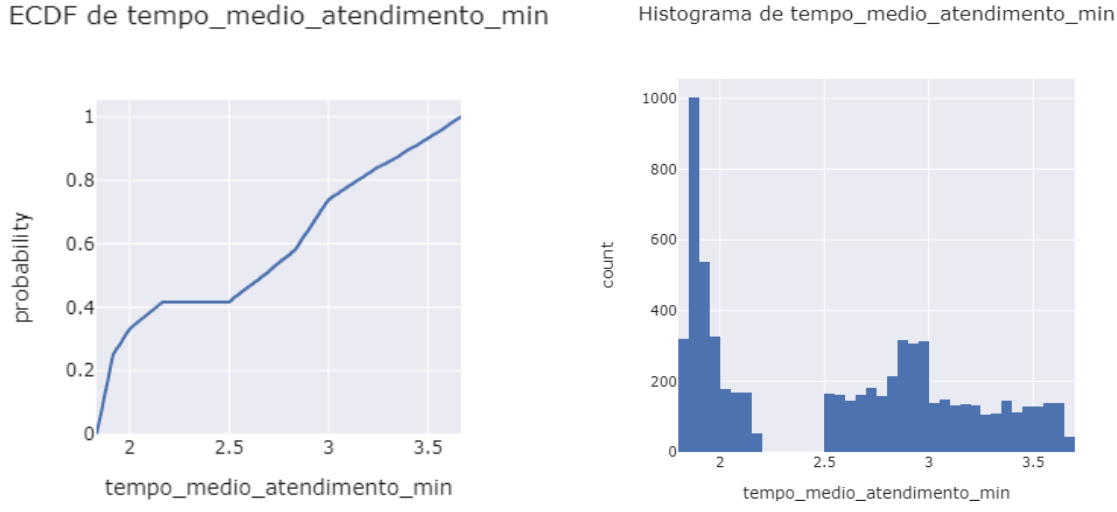
Fonte: Elaborado pelo autor no Google Colab.

Figura 17 – Gráfico ECDF e Histograma de tempo_medio_resposta_min



Fonte: Elaborado pelo autor no Google Colab.

Figura 18 – Gráfico ECDF e Histograma de tempo_medio_atendimento_min



Fonte: Elaborado pelo autor no Google Colab.

As Figuras 13 e 15 mostram que os valores de “horas_atendimento_canal” e “qtd_colaboradores_area” se concentram na primeira metade dos valores, representando 80% dos casos até o segundo quartil. Já as Figuras 14 e 17 mostram que os valores de “qtd_atendimentos” e “tempo_medio_resposta_min” se concentram no primeiro quartil dos valores. Por fim, a Figura 16 demonstra que a variável “nps_medio” está mais concentrada na segunda metade dos valores e a Figura 18 mostra que há uma distribuição mais homogênea ao longo dos valores com um pico logo no início.

Também é interessante comentar que para variáveis inteiras como “qtd_colaboradores_area” e “nps_medio” o gráfico ECDF tem um aspecto de escada, diferente das variáveis contínuas.

Esses gráficos são relevantes não só para ter-se um primeiro entendimento das variáveis, mas também para identificar qual é a distribuição dos valores de cada variável e, posteriormente, na etapa de Tratamento dos Dados, ser possível escolher qual é a melhor forma de tratá-los seguindo as devidas transformações necessárias.

4.1.4.2 Análise Bivariada

Segundo Morettin e Singer (2021), após o entendimento individual das variáveis, é necessário entender a relação entre elas. Sendo assim, primeiro será realizado um estudo bivariado (relação entre duas variáveis) e depois um multivariado (relação entre mais de 2 variáveis).

Como o trabalho busca entender quais fatores influenciam o NPS do *contact center* do Banco A, foram criadas algumas hipóteses (Tabela 4) de quais relações entre o “nps_medio” e as outras variáveis podem ser relevantes para serem posteriormente testadas em análises de estatística descritiva e gráficas. Dentro das hipóteses ainda surgiram alguns fatores derivados das variáveis originais da base de dados que também poderiam estar relacionadas ao NPS, por exemplo, o dia da semana.

Tabela 4 – Quadro de hipóteses da análise bivariada

Análise	Hipótese
Análise 1: Senioridade vs. NPS	Quanto mais sênior um atendente é, maior será o NPS.
Análise 2: Produto vs. NPS	Talvez determinados produtos apresentem maior NPS simplesmente por satisfazerem mais os clientes (Exemplo: crédito costuma ter maior NPS no banco, pois as pessoas ficam satisfeitas em terem crédito)
Análise 3: Canal vs. NPS	Canais de mais fácil acesso podem ter NPS maior.
Análise 4: Horas de atendimento no canal por pessoa vs. NPS	Canais que dependem de mais tempo do atendente podem impactar negativamente no NPS, pois isso os saturariam.
Análise 5: Quantidade de Atendimentos vs. NPS & Quantidade de Atendimentos por pessoa vs. NPS	Um dia muito cheio de atendimentos pode impactar negativamente o NPS.
Análise 6: Tempo médio de atendimento (TMA) vs. NPS	Quanto maior a demora no atendimento do cliente, menor seria o NPS.
Análise 7: Tempo médio de resposta vs. NPS	Quanto maior demora para iniciar um atendimento, menor seria o NPS.
Análise 8: Dia da semana vs. NPS	Algum dia pode influenciar positivamente ou negativamente o NPS.

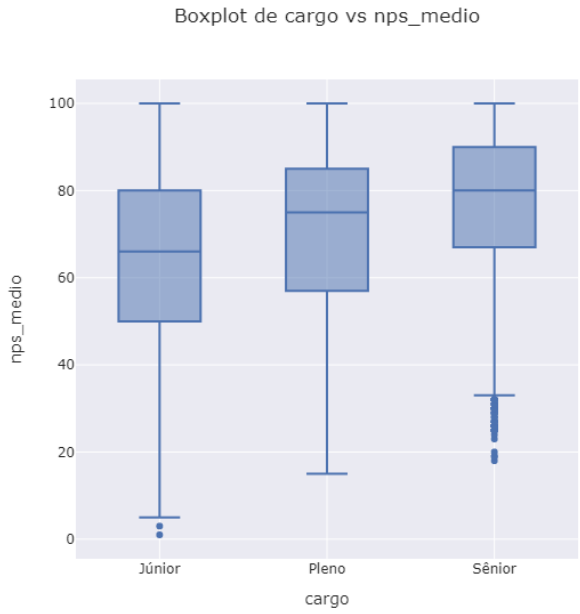
Fonte: Elaborado pelo autor no Google Colab.

4.1.4.2.1 Análise 1: Senioridade vs. NPS

Na primeira análise foi testada a relação da variável categórica “cargo” com a variável numérica “nps_medio”. De acordo com Morettin e Singer (2021), para essa combinação de tipos de variáveis pode-se utilizar o gráfico *Boxplot* na interpretação de suas relações.

Sendo assim, foi criado o gráfico da Figura 19, no qual pode ser observada uma tendência de alta no NPS médio à medida que a senioridade dos atendentes aumenta, seguindo a hipótese inicial. Lembrando que, isso não confirma a relevância dessa variável em relação ao NPS, mas indica que ela pode ser uma variável relevante. A confirmação virá quando os modelos estatísticos forem executados.

Figura 19 – Gráfico *Boxplot* de cargo vs. nps_medio



Fonte: Elaborado pelo autor no Google Colab.

Tabela 5 – Estatística descritiva da relação cargo vs. nps_medio

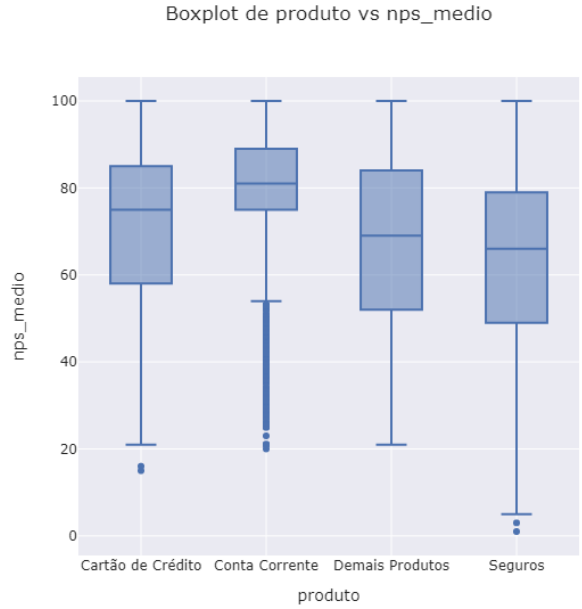
	nps_medio							
	count	mean	std	min	25%	50%	75%	max
cargo								
Júnior	2208.0	64.047554	20.045033	1.0	50.0	66.0	80.0	100.0
Pleno	2208.0	70.881341	18.294461	15.0	57.0	75.0	85.0	100.0
Sênior	2201.0	76.369378	17.631154	18.0	67.0	80.0	90.0	100.0

Fonte: Elaborado pelo autor no Google Colab.

4.1.4.2.2 Análise 2: Produto vs. NPS

Na segunda análise foi testada a relação da variável “produto” com a variável “nps_medio”. Como observado na Figura 20, essa relação parece ser constante, a não ser pelo produto “Conta Corrente”, ele tende a puxar a média do NPS para cima. Portanto, não podemos dispensar a possibilidade dessa variável também ser importante.

Figura 20 – Gráfico *Boxplot* de produto vs. nps_medio



Fonte: Elaborado pelo autor no Google Colab.

Tabela 6 – Estatística descritiva da relação produto vs. nps_medio

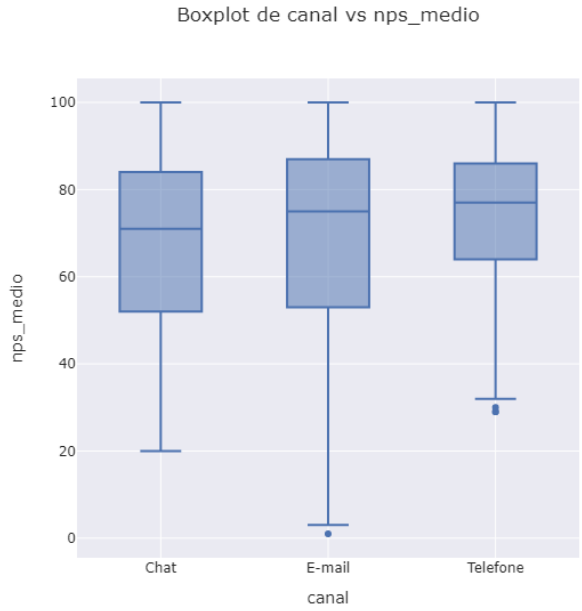
	nps_medio							
	count	mean	std	min	25%	50%	75%	max
produto								
Cartão de Crédito	1650.0	71.681818	17.585966	15.0	58.0	75.0	85.0	100.0
Conta Corrente	1655.0	78.115408	15.865116	20.0	75.0	81.0	89.0	100.0
Demais Produtos	1656.0	67.350242	20.221476	21.0	52.0	69.0	84.0	100.0
Seguros	1656.0	64.567633	20.586013	1.0	49.0	66.0	79.0	100.0

Fonte: Elaborado pelo autor no Google Colab.

4.1.4.2.3 Análise 3: Canal vs. NPS

Na terceira análise foi testada a relação da variável “canal” com a variável “nps_medio”. A Figura 21 mostra que não há uma tendência muito bem definida para esta relação, logo, essa pode ser uma variável não tão relevante para o NPS médio.

Figura 21 – Gráfico *Boxplot* de canal vs. nps_medio



Fonte: Elaborado pelo autor no Google Colab.

Tabela 7 – Estatística descritiva da relação canal vs. nps_medio

	nps_medio							
	count	mean	std	min	25%	50%	75%	max
canal								
Chat	2205.0	67.740136	20.192226	20.0	52.0	71.0	84.0	100.0
E-mail	2208.0	69.048460	21.501033	1.0	53.0	75.0	87.0	100.0
Telefone	2204.0	74.494555	15.093399	29.0	64.0	77.0	86.0	100.0

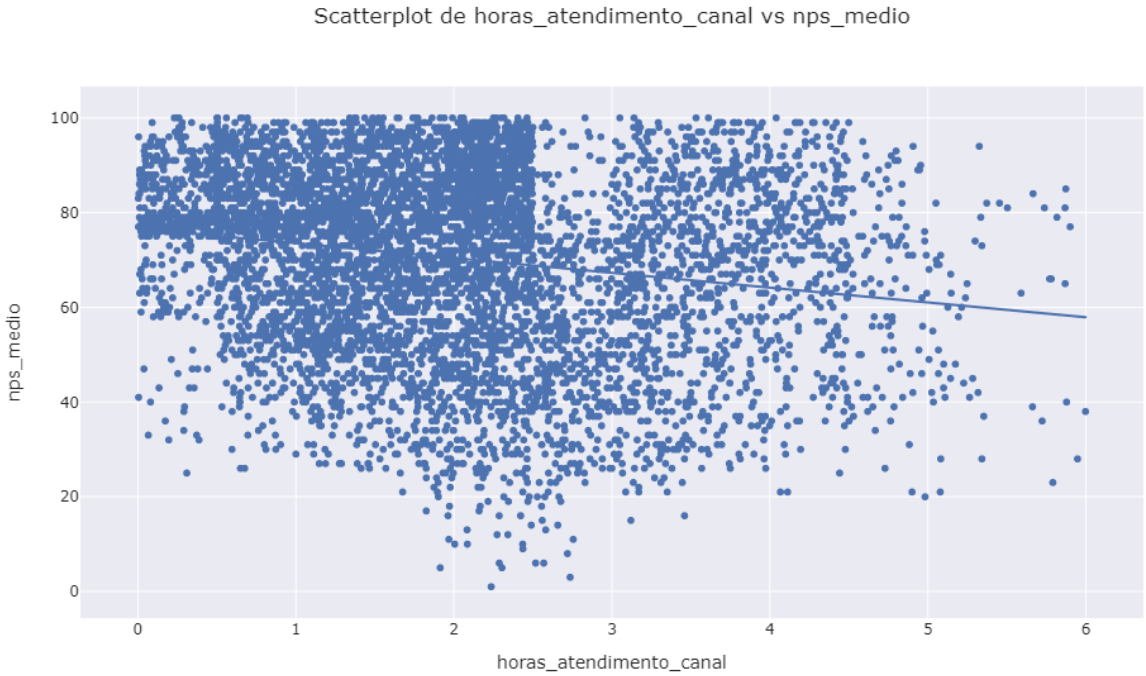
Fonte: Elaborado pelo autor no Google Colab.

4.1.4.2.4 Análise 4: Horas de atendimento no canal por pessoa vs. NPS

Na quarta análise foi testada a relação da variável numérica “horas_atendimento_canal” com a variável “nps_medio”, também númerica. Segundo Morettin e Singer (2021), para essa combinação de tipos de variáveis pode-se utilizar o Gráfico de Dispersão na interpretação de suas relações.

Analisando o gráfico da Figura 22, pode-se observar que ele possui uma linha de tendência bem vertical, isto é, não há uma tendência muito bem definida para esta relação, os dados são bem distribuídos ao longo dos valores. Portanto, essa variável não apresenta uma correlação muito forte com o NPS médio, logo, pode não ser tão relevante para o estudo.

Figura 22 – Gráfico de Disperção de horas_atendimento_canal vs. nps_medio

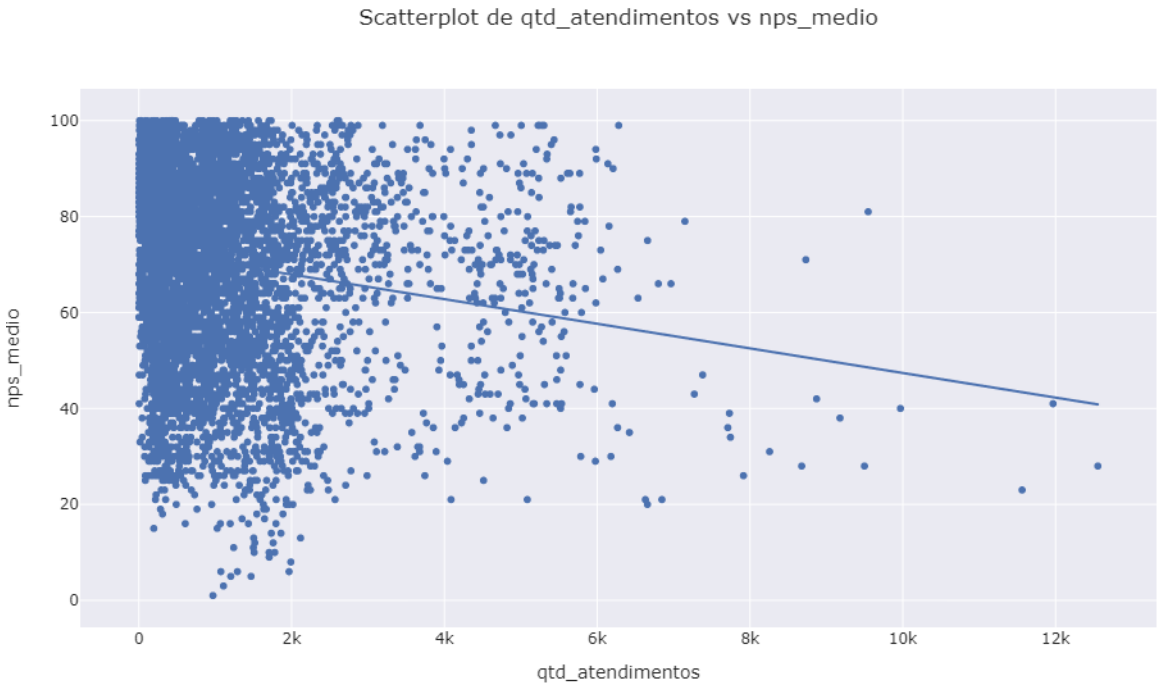


Fonte: Elaborado pelo autor no Google Colab.

4.1.4.2.5 Análise 5: Qtd. de Atendimentos (e Qtd. de Atendimentos/pessoa) vs. NPS

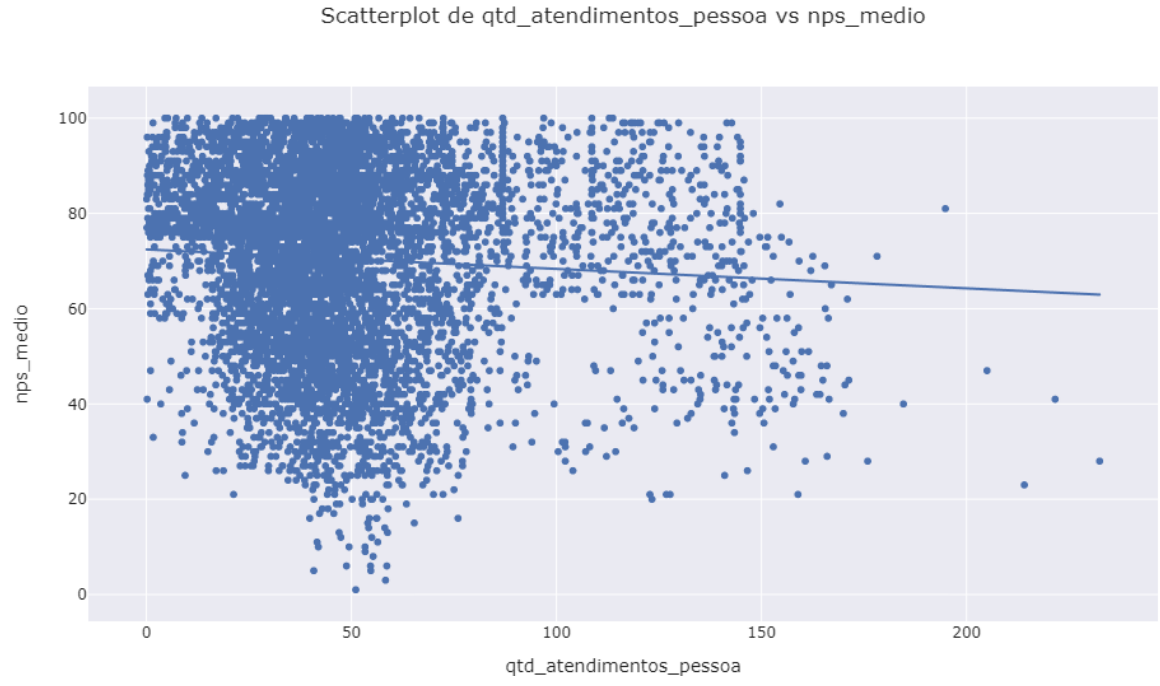
Na quinta análise foi testada não só a relação da variável “qtd_atendimentos” com a variável “nps_medio”, mas também da nova variável criada “qtd_atendimentos_pessoa” com a variável “nps_medio”. Esse foi um dos casos de criação de variável derivada da base original, pois queria-se entender se quanto maior o número médio de atendimentos por pessoa em um dia, menor seria o NPS médio. Em outras palavras, a variável “qtd_atendimentos” é um número relativo a todos os atendimentos feitos por todos os atendentes daquele turno específico e, portanto, não traduz a quantidade média de atendimentos que cada um dos atendentes de um turno acaba fazendo.

Figura 23 – Gráfico de Disperção de qtd_atendimentos vs. nps_medio



Fonte: Elaborado pelo autor no Google Colab.

Figura 24 – Gráfico de Disperção de qtd_atendimentos_pessoa vs. nps_medio



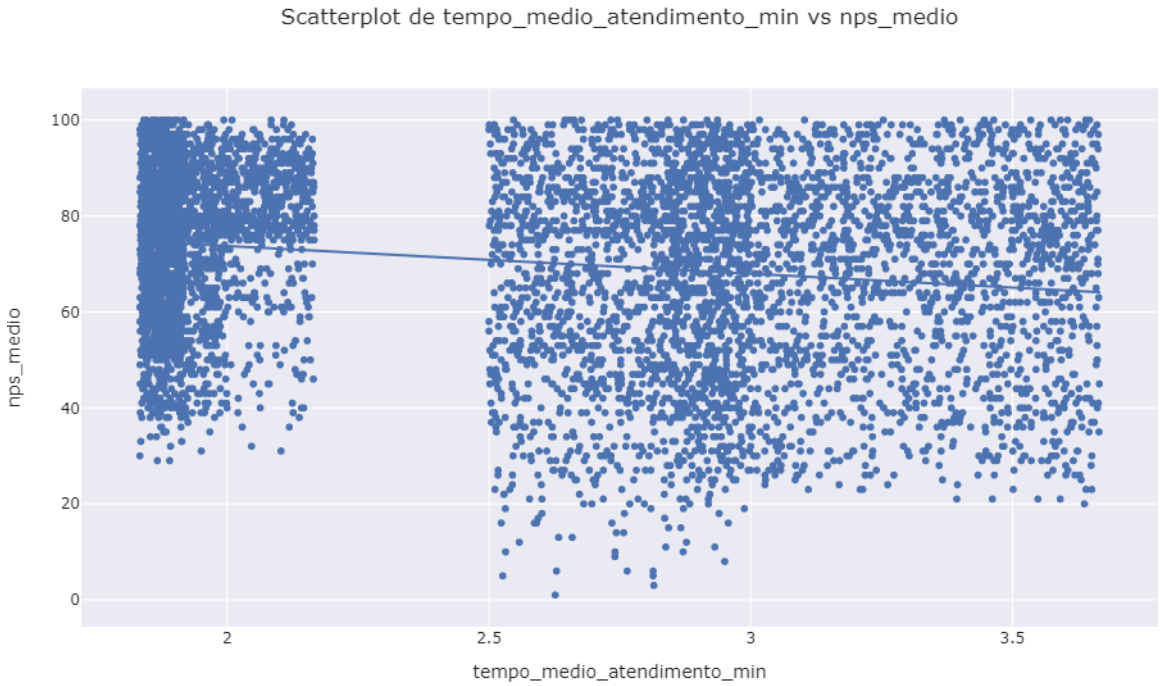
Fonte: Elaborado pelo autor no Google Colab.

Quando observado o gráfico da variável original (Figura 23), há uma leve tendência de queda no NPS quando a quantidade de atendimentos aumenta. Entretanto, quando observado o gráfico da variável na perspectiva por pessoa (Figura 24), essa tendência diminui e a linha fica bem horizontalizada.

4.1.4.2.6 Análise 6: Tempo médio de atendimento (TMA) vs. NPS

Na sexta análise foi testada a relação da variável “tempo_medio_atendimento” com a variável “nps_medio”. O gráfico da Figura 25 demonstra que não há uma tendência muito forte de relação entre as variáveis e também mostra um fato curioso, de que não há dados de TMA entre 2,3 e 2,5 minutos.

Figura 25 – Gráfico de Disperção de tempo_medio_atendimento_min vs. nps_medio

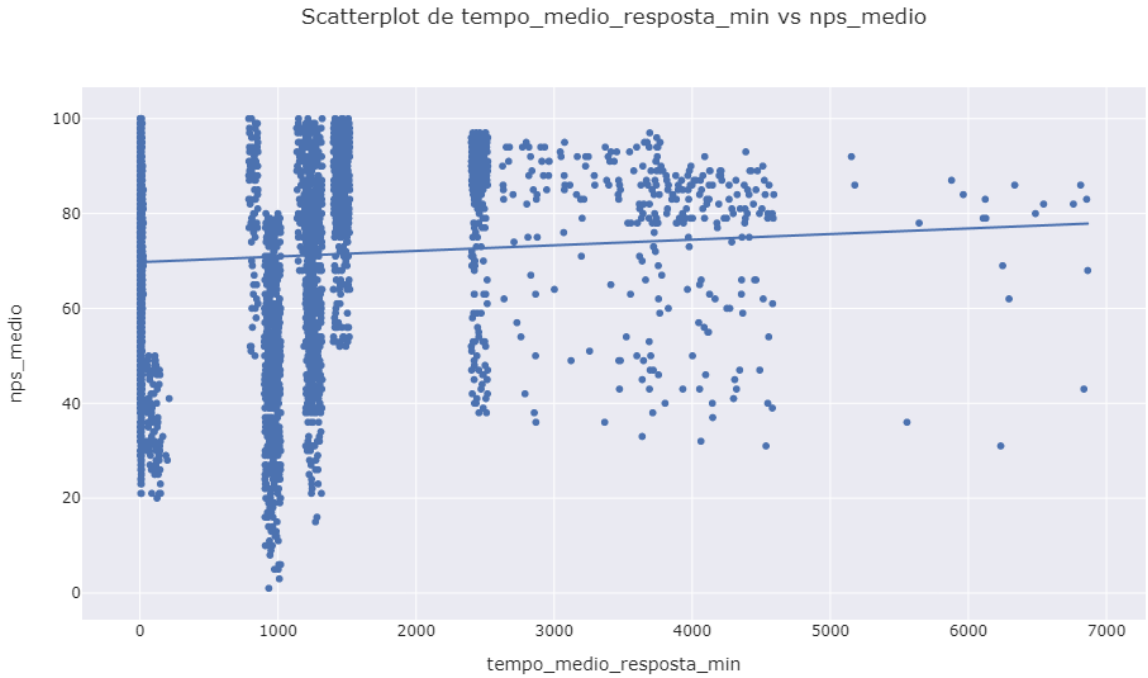


Fonte: Elaborado pelo autor no Google Colab.

4.1.4.2.7 Análise 7: Tempo médio de resposta vs. NPS

Na sétima análise foi testada a relação da variável “tempo_medio_resposta_min” com a variável “nps_medio”. O gráfico da Figura 26 demonstra que não há uma tendência muito forte de relação entre estas variáveis. Também pode-se observar que a distribuição dos dados nos valores de tempo é concentrada em alguns tempos específicos.

Figura 26 – Gráfico de Disperção de tempo_medio_resposta_min vs. nps_medio



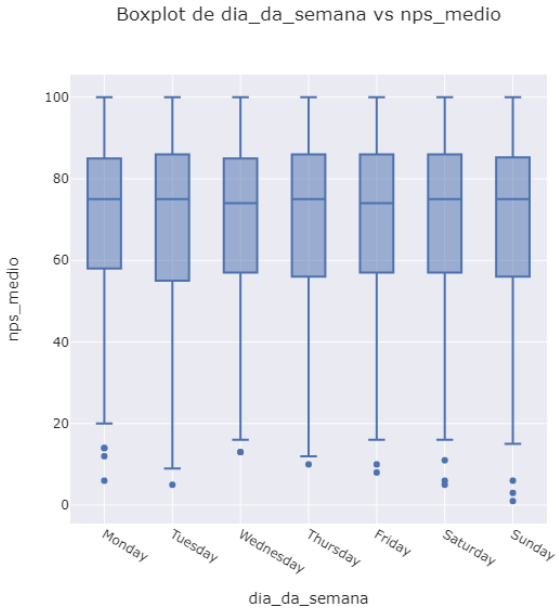
Fonte: Elaborado pelo autor no Google Colab.

4.1.4.2.8 Análise 8: Dia da semana vs. NPS

Na oitava análise foi testada a nova variável “dia_da_semana” com a variável “nps_medio”. Este novo fator foi criado a partir dos dados da data de atendimento e tem o objetivo de entender se algum dia da semana influencia os valores do NPS médio.

Com o apoio da Figura 27, pode-se perceber que também não há uma tendência muito forte de relação entre estas variáveis. Os valores de NPS médio são bem constantes ao longo de toda a semana, descartando a hipótese inicial.

Figura 27 – Gráfico de Disperção de dia_da_semana vs. nps_medio



Fonte: Elaborado pelo autor no Google Colab.

Tabela 8 – Estatística descritiva da relação dia_da_semana vs. nps_medio

	nps_medio							
	count	mean	std	min	25%	50%	75%	max
dia_da_semana								
Friday	936.0	70.549145	19.213068	8.0	57.0	74.0	86.0	100.0
Monday	971.0	70.813594	18.652030	6.0	58.0	75.0	85.0	100.0
Saturday	934.0	70.437901	19.307516	5.0	57.0	75.0	86.0	100.0
Sunday	933.0	70.095391	19.451009	1.0	56.0	75.0	85.0	100.0
Thursday	936.0	70.604701	19.871452	10.0	56.0	75.0	86.0	100.0
Tuesday	972.0	70.013374	19.689475	5.0	55.0	75.0	86.0	100.0
Wednesday	935.0	70.471658	19.309022	13.0	57.0	74.0	85.0	100.0

Fonte: Elaborado pelo autor no Google Colab.

4.1.4.2.9 Conclusões análise bivariada

Para complementar os gráficos de dispersão elaborados para a relação de duas variáveis numéricas, foi feita uma análise de Correlação de Pearson (Tabela 9) entre as variáveis numéricas. Algumas variáveis apresentam um maior grau de correlação como “horas_atendimento_canal” vs. “qtd_atendimentos” e “qtd_colaboradores_area” vs. “qtd_atendimentos”, mas ainda assim não é um grau muito elevado. Ademais, nenhuma variável numérica tem alguma correlação expressiva em relação a nossa variável de estudo, o “nps_medio”. Em todos os casos tem-se uma correlação menor que 0,2 (em módulo), confirmando as inferências feitas durante as análises gráficas.

Tabela 9 – Matriz de Correlação de Pearson entre as variáveis numéricas da base de dados

	horas_atendimento_canal	qtd_atendimentos	qtd_colaboradores_area	nps_medio	tempo_medio_resposta_min	tempo_medio_atendimento_min	qtd_atendimentos_pessoa
horas_atendimento_canal	1.000000	0.643283	0.168494	-0.187028	-0.150544	0.392451	0.787384
qtd_atendimentos	0.643283	1.000000	0.649325	-0.152629	-0.071624	-0.001460	0.727034
qtd_colaboradores_area	0.168494	0.649325	1.000000	-0.195789	0.036720	-0.011926	0.114208
nps_medio	-0.187028	-0.152629	-0.195789	1.000000	0.060691	-0.178267	-0.063303
tempo_medio_resposta_min	-0.150544	-0.071624	0.036720	0.060691	1.000000	-0.170837	-0.082942
tempo_medio_atendimento_min	0.392451	-0.001460	-0.011926	-0.178267	-0.170837	1.000000	-0.060382
qtd_atendimentos_pessoa	0.787384	0.727034	0.114208	-0.063303	-0.082942	-0.060382	1.000000

Fonte: Elaborado pelo autor no Google Colab.

Sendo assim, após todas essas análises, pode-se concluir que o foco maior de relevância em relação ao NPS médio está sob as variáveis categóricas, mas isso não quer dizer que as variáveis numéricas não podem ser relevantes posteriormente no modelo.

4.1.4.3 Análise Multivariada

Como Morettin e Singer (2021) salientam, a análise multivariada é uma extensão da análise bivariada que nos permite investigar e compreender as relações entre três ou mais variáveis. Este é um passo crucial na análise de dados, pois permite considerar a interação e o impacto mútuo entre várias variáveis no modelo.

Assim como na análise bivariada, foram criadas algumas hipóteses para guiar o projeto (Tabela 10) que depois foram testadas por meio de Gráficos de Dispersão Simbólicos. Este tipo de análise gráfica é recomendado por Morettin e Singer (2021) nos casos multivariados.

Tabela 10 – Quadro de hipóteses da análise multivariada

Análise	**Hipótese**
Análise 1: Senioridade vs. TMA vs. NPS	Pessoas mais seniores atendem mais rápido e melhor, gerando uma nota de NPS maior.
Análise 2: Canal vs. Tempo Médio de Resposta vs. NPS	O canal deve influenciar o tempo médio de resposta, o que deve influenciar o NPS. De forma negativa para canais mais demorados e positiva para canais mais rápidos.
Análise 3: Senioridade vs. Quantidade de Atendimento vs. NPS e Senioridade vs. Quantidade de Atendimento/Pessoa vs. NPS	A senioridade deve influenciar a quantidade de atendimentos realizada e isto pode aumentar ou diminuir o NPS.

Fonte: Elaborado pelo autor no Google Colab.

4.1.4.3.1 Análise 1: Senioridade vs. TMA vs. NPS

Na primeira análise foi testada a relação das variáveis “cargo”, “tempo_medio_atendimento_min” e “nps_medio”. De acordo com a Figura 28, as tendências são horizontais, não indicando uma relação muito forte entre estas variáveis.

Figura 28 – Gráfico de Disperção tempo_medio_atendimento_min vs. nps_medio vs. cargo

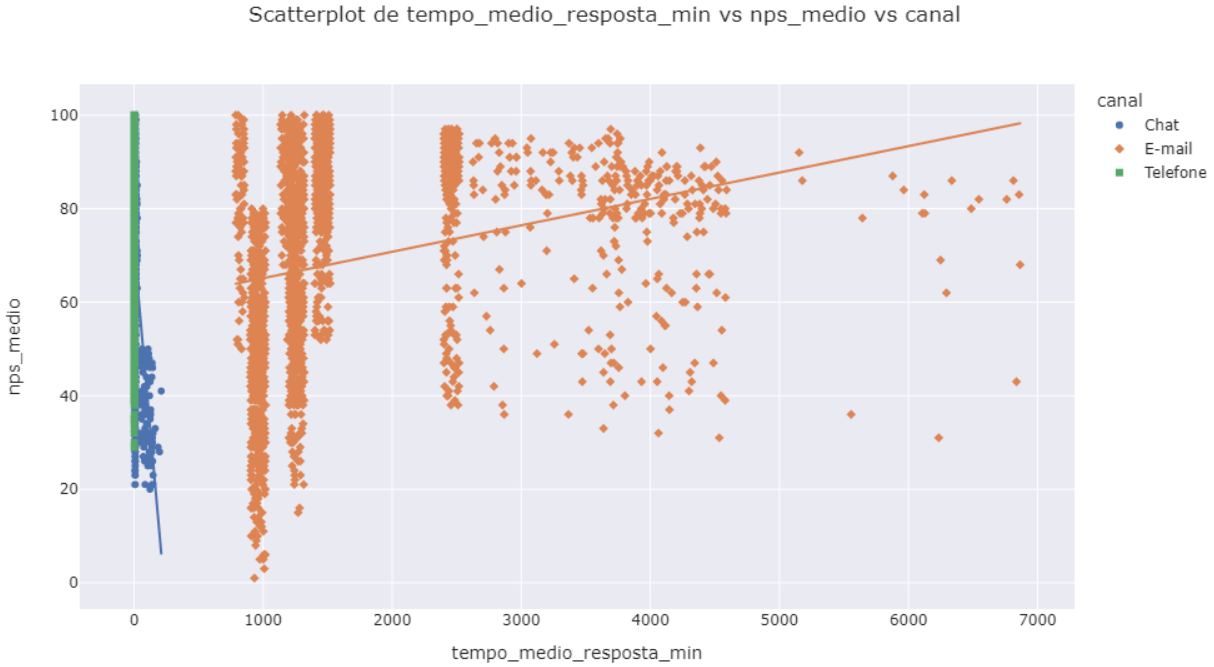
Fonte: Elaborado pelo autor no Google Colab.

4.1.4.3.2 Análise 2: Canal vs. Tempo Médio de Resposta vs. NPS

Na segunda análise foi testada a relação das variáveis “canal”, “tempo_medio_resposta_min” e “nps_medio”. Como pode-se observar na Figura 29, há uma tendência positiva do NPS médio para o tempo médio de resposta no canal “E-mail” e uma tendência negativa do NPS no tempo médio de resposta no canal “Chat”.

Diferente da análise bivariada, quando observada essa relação tripla, o “tempo_medio_resposta_min” acaba se tornando uma variável candidata a ser importante na relação com o NPS médio, confirmando que o que foi visto anteriormente não é definitivo.

Figura 29 – Gráfico de Disperção tempo_medio_resposta_min vs. nps_medio vs. canal



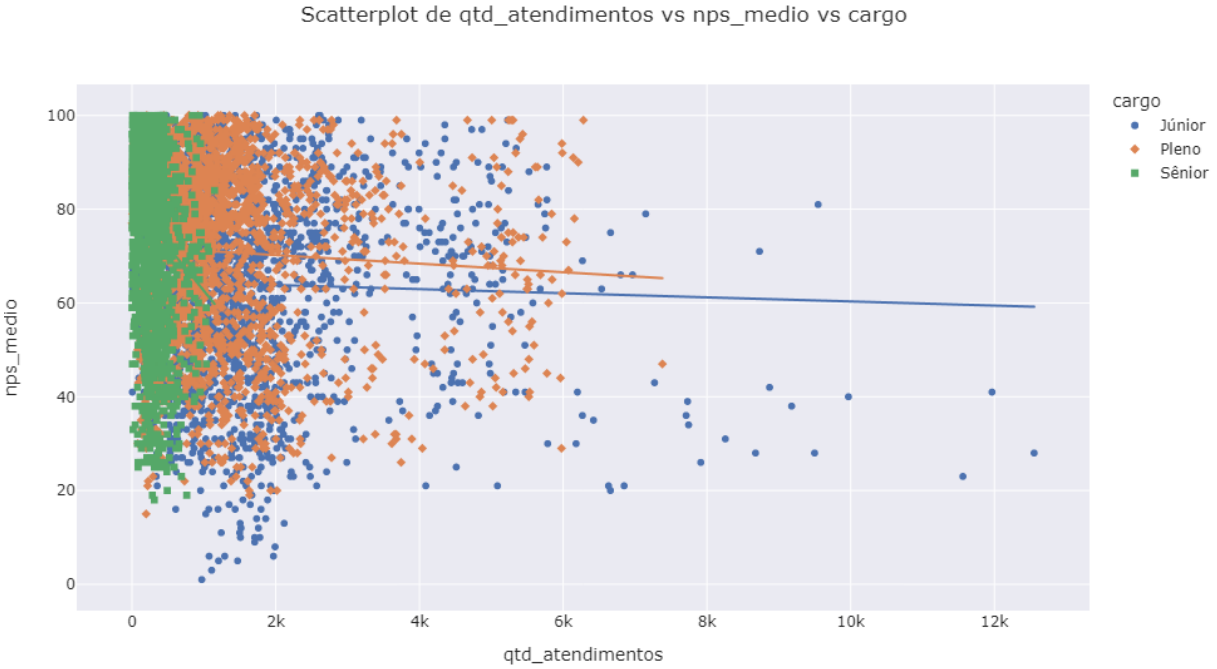
Fonte: Elaborado pelo autor no Google Colab.

4.1.4.3.3 Análise 3: Senioridade vs. Qtd. de Atendimento (e Qtd. de Atendimento/Pessoa) vs. NPS

Na terceira análise, também foi utilizada a variável criada na análise bivariada “qtd_atendimentos_pessoa”. Neste caso, foi testada tanto a relação das variáveis “qtd_atendimentos”, “cargo” e “nps_medio”, quanto a relação das variáveis “qtd_atendimentos_pessoa”, “cargo” e “nps_medio”.

Como observado nas Figuras 30 e 31, as tendências não indicam uma relação muito forte entre estas variáveis.

Figura 30 – Gráfico de Disperção qtd_atendimentos vs. nps_medio vs. cargo



Fonte: Elaborado pelo autor no Google Colab.

Figura 31 – Gráfico de Disperção qtd_atendimentos_pessoa vs. nps_medio vs. cargo

Fonte: Elaborado pelo autor no Google Colab.

4.2 Fase “Preparação dos Dados (*Data Preparation*)”

Esta fase é a terceira etapa no processo de mineração de dados do modelo CRISP-DM. Nesta fase, os dados brutos são transformados em um formato adequado para modelagem, envolvendo tarefas como limpeza, integração, seleção e transformação dos dados. Chapman et

al. (2000) sugerem que a preparação de dados é frequentemente o estágio mais demorado e desafiador de um projeto de ciência de dados, mas é essencial para garantir que os dados sejam confiáveis e úteis para a análise subsequente.

Desta lista de tarefas uma já foi realizada na etapa anterior, a limpeza de dados que culminou na exclusão e incrementação de dados na base. Das que restaram, duas são dispensáveis especificamente nesse trabalho. A primeira é a seleção de dados, pois foram utilizados todos os dados disponíveis; a segunda é a integração de dados, já que não foi necessário adicionar nenhuma outra informação na base de dados. Portanto, serão tratadas neste item do trabalho apenas as tarefas de divisão e transformação de dados.

4.2.1 Divisão da base em Treino e Teste

Embora a separação dos dados em conjuntos de treinamento e teste possa não ser explicitamente mencionada na descrição original do CRISP-DM, ela é uma prática recomendada e está implícita na etapa de "Preparação de Dados", pois o propósito desta etapa é preparar os dados finais que serão alimentados nos algoritmos de modelagem.

A importância dessa estratégia reside na sua capacidade de prevenir o *overfitting*, um problema comum em *machine learning*. Goodfellow, Bengio e Courville (2016) explicam que o *overfitting* acontece quando um modelo se ajusta excessivamente aos dados de treinamento, a ponto de aprender o ruído desses dados, e, consequentemente, performar mal quando exposto a dados inéditos. O conjunto de treinamento é utilizado para otimizar os parâmetros do modelo, e o conjunto de teste fornece um meio de verificar a eficiência do modelo ao generalizar os conhecimentos adquiridos para dados novos.

Sendo assim, neste trabalho foi utilizada uma divisão de 75% do total da base para treinamento e 25% para teste, próxima aos valores comuns de 80% - 20% (KOHAVI, 1995). Foi tomada essa decisão, pois a base de dados utilizada não tem muitos dados e queria-se que a base de teste não fosse tão pequena.

A base com as variáveis de entrada foram nomeadas de "X_train" e X_test" e as bases de saída "Y_train" e "Y_test". As de treino ficaram com 4.962 linhas e as de teste 1.655.

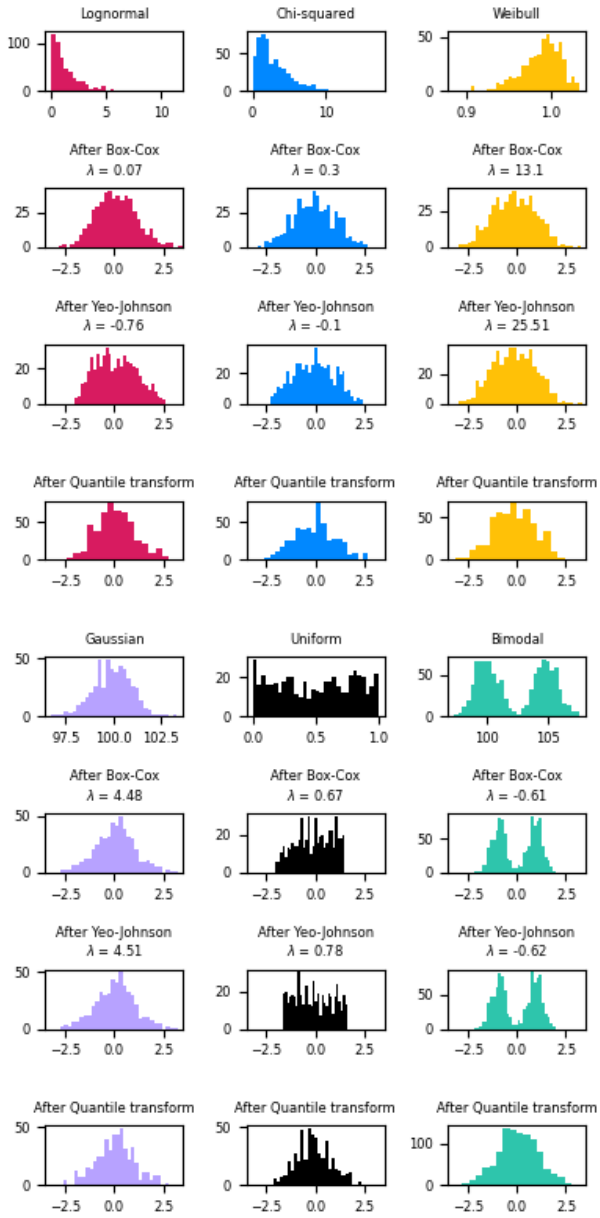
4.2.2 Transformação de dados

Como descrito por Morettin e Singer (2021) o objetivo da transformação de variáveis é modificar os dados de maneira a aprimorar a interpretação das informações contidas neles e/ou

adequá-los às premissas de métodos estatísticos a serem aplicados posteriormente. As transformações podem ser usadas tanto para lidar com variáveis numéricas quanto categóricas e elas devem ser realizadas tanto nas bases de treino (“X_train”), quanto de teste (“X_test”).

Muitos modelos estatísticos partem da premissa de que os valores de uma ou mais variáveis possuem uma distribuição normal (MORETTIN & SINGER, 2021) e este é o caso dos modelos de regressão linear utilizados neste projeto. Sendo assim, para as variáveis numéricas, foi necessário realizar um processo de normalização dos dados, utilizando a biblioteca Scikit Learn como apoio (Figura 32).

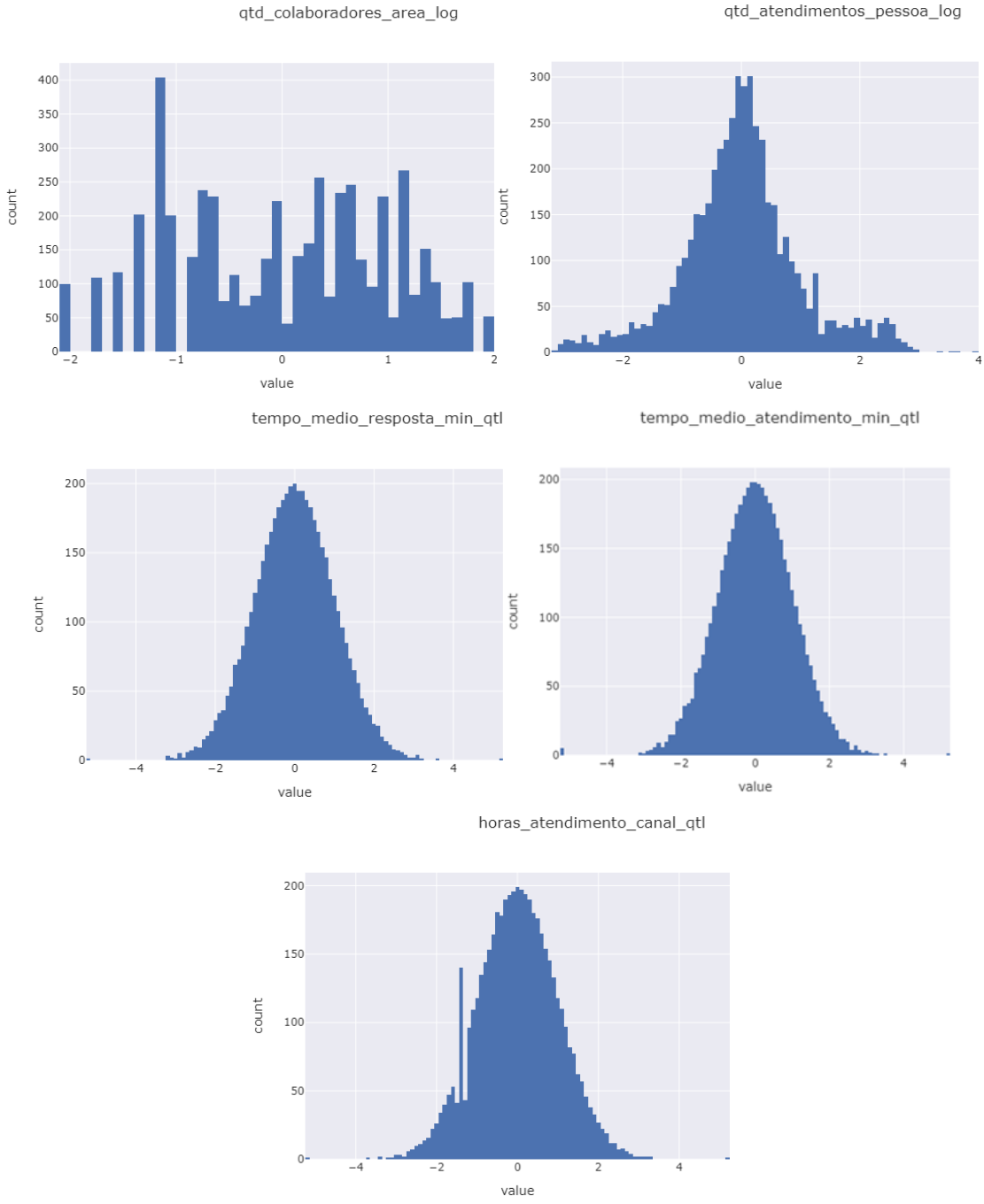
Figura 32 – Comparação gráfica de transformações para apoio na decisão de escolha



Fonte: Biblioteca Scikit Learn (2023).

Com o auxílio da Figura 32, foi comparada a distribuição dos valores observados na Análise Exploratória dos Dados e, assim, feita a escolha de quais transformações seriam utilizadas. As variáveis “qtd_colaboradores_area” e “qtd_atendimentos_pessoa” possuem um aspecto próximo ao lognormal, portanto, foi escolhida a transformação “Yeo-Johnson”. Já as variáveis “tempo_medio_resposta_min”, “tempo_medio_atendimento_min” e “horas_atendimento_canal”, possuem um aspecto próximo ao uniforme, portanto, foi escolhida a transformação “Quantílica”.

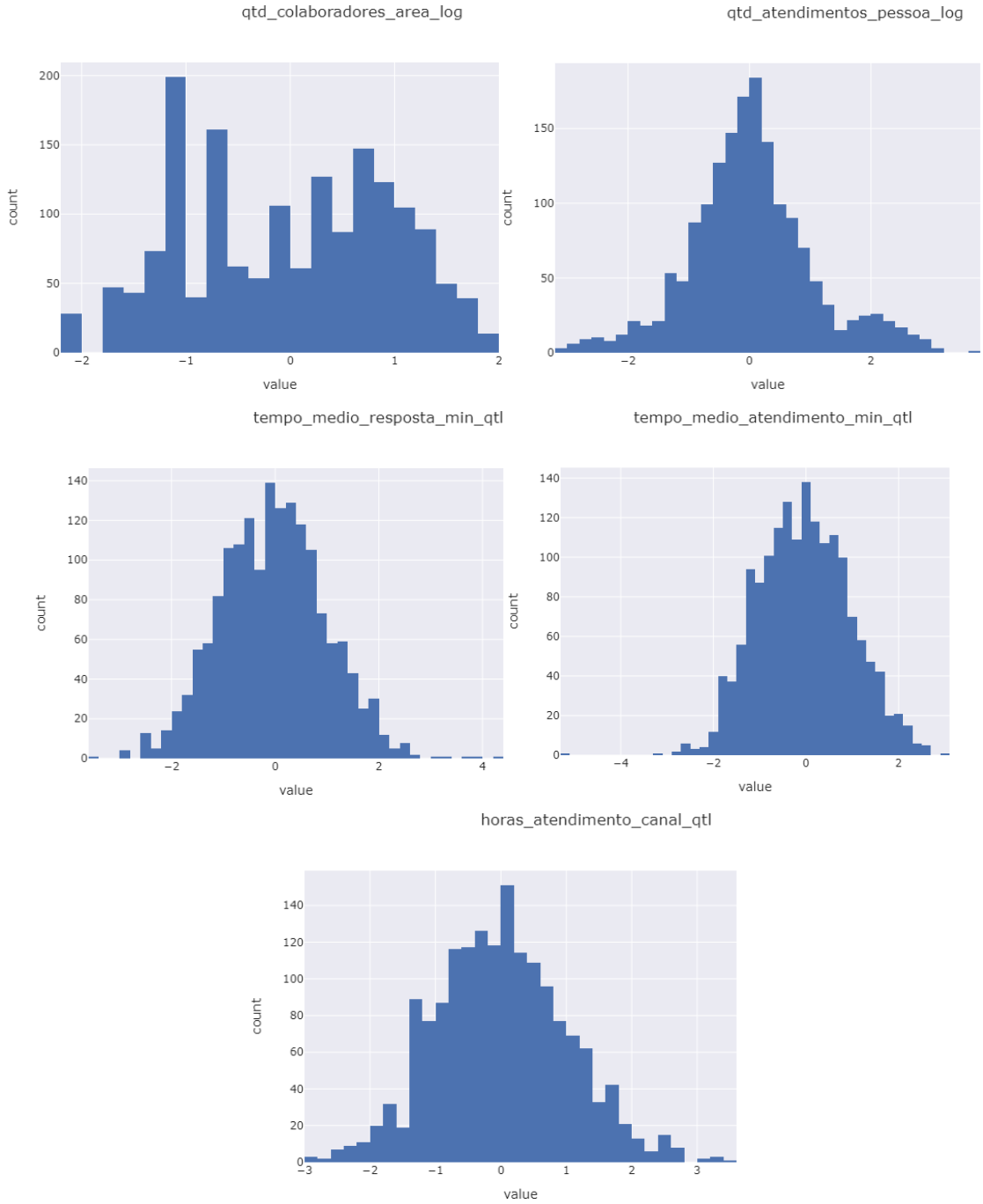
Figura 33 – Histogramas das variáveis numéricas na base de treino após transformações



Fonte: Elaborado pelo autor no Google Colab.

Conferindo a Figura 33, pode-se observar que quase todas as variáveis foram bem normalizadas a não ser pela “qtd_colaboradores_area”. Esta variável tem uma distribuição com um padrão não muito bem definido, por isso, decidiu-se continuar com o que foi feito.

Figura 34 – Histogramas das variáveis numéricas na base de teste após transformações



Fonte: Elaborado pelo autor no Google Colab.

O comportamento da distribuição das variáveis numéricas transformadas na base de teste (Figura 34) foi bem próximo ao da base de treino. Além disso, também pode-se perceber que foi criada uma nova nomenclatura para as variáveis transformadas. Para as transformações Lognormal por Yeo-Johson tem-se as variáveis “qtd_colaboradores_area_log” e “qtd_atendimentos_pessoa_log”, para as transformações quantílicas tem-se “tempo_medio_resposta_min_qtl”, “tempo_medio_atendimento_min_qtl” e “horas_atendimento_canal_qtl”.

Já para a transformação das variáveis categóricas foi utilizada a função “*get_dummies*” da biblioteca Pandas. Essa função converte variáveis categóricas em variáveis *dummy* ou indicadoras, um processo também conhecido como codificação *dummy*. Essencialmente, este método cria n-1 novas variáveis binárias para cada categoria possível da variável original. Em cada uma dessas novas variáveis, um valor de “1” indica a presença da categoria e “0” a ausência, como exemplificado na Tabela 11. As variáveis criadas foram: “cargo_Pleno”, “cargo_Sênior”, “canal_E-mail”, “canal_Telefone”, “produto_Conta Corrente”, “produto_Demais produtos” e “produto_Seguros”.

Essa transformação é importante, pois todos os algoritmos de modelagem utilizados no projeto esperam que os dados de entrada sejam numéricos. Além disso, ela permite que o modelo capture qualquer estrutura potencial e padrões que possam estar associados às diferentes categorias.

Tabela 11 – Exemplo da base de variáveis categóricas após transformação *dummy*

	cargo_Pleno	cargo_Sênior	canal_E-mail	canal_Telefone	produto_Conta Corrente	produto_Demais Produtos	produto_Seguros
3353	1	0	0	0	0	1	0
6174	0	1	1	0	0	0	1
4544	0	1	0	0	0	1	0
3771	0	1	1	0	0	1	0
2592	0	1	1	0	1	0	0
...	...	...	...	...	...	...	...

Fonte: Elaborado pelo autor no Google Colab.

4.3 Fase “Modelagem (*Modeling*)”

A fase de Modelagem se concentra na aplicação de algoritmos para a construção de modelos que possam desvendar padrões ou predizer resultados desconhecidos com base nos dados coletados e preparados nas etapas anteriores (CHAPMAN et al., 2000). Nessa fase, é comum a implementação de uma ampla variedade de modelos estatísticos ou de aprendizado de máquina e a comparação de seu desempenho. O objetivo é identificar o modelo que oferece

o melhor desempenho preditivo ou descritivo para o problema em questão. É importante lembrar que esta etapa é iterativa, os modelos podem precisar ser ajustados várias vezes antes de apresentarem resultados satisfatórios (WITH & HIPP, 2000).

Para selecionar quais técnicas de modelagem utilizar é necessário entender qual problema de modelagem está sendo discutido. Neste caso, o problema que está sendo trabalhado é um problema de regressão, pois a variável resposta a ser prevista é o NPS, uma variável numérica. Sendo assim, foram escolhidos quatro tipos de modelos a serem feitos, os dois primeiros mais simples e os dois últimos mais complexos:

1. Regressão Linear Múltipla (*Linear Regression*);
2. Regressão Linear de Lasso (*Lasso's Linear Regression*);
3. Árvores de Decisão de Regressão (*Decision Tree Regression*);
4. Florestas Aleatórias de Regressão (*Random Forest Regression*).

Foi tomada essa decisão, pois os dois primeiros modelos possuem uma base de cálculo diferente da dos dois últimos. Essa distinção é benéfica para o projeto, pois aumenta as possibilidades de interpretação e *insights*.

Vale ressaltar que a baixa correlação entre variáveis identificada na Análise Exploratória de Dados é benéfica para os modelos de regressão utilizados no trabalho, pois previne a multicolinearidade, um fenômeno em que duas ou mais variáveis independentes em um modelo de regressão estão altamente correlacionadas. Quando a multicolinearidade é alta, torna-se difícil determinar o efeito individual de cada variável na variável dependente, pois as alterações em uma variável independente estão associadas a alterações em outra. Além disso, a multicolinearidade pode levar a coeficientes de regressão instáveis e inflados, o que torna a interpretação do modelo desafiadora e pode resultar em um modelo sobreajustado com desempenho ruim em dados não vistos. Portanto, uma baixa correlação entre as variáveis geralmente leva a modelos de regressão mais estáveis, precisos e interpretáveis.

4.3.1 Modelo de Regressão Linear Múltipla

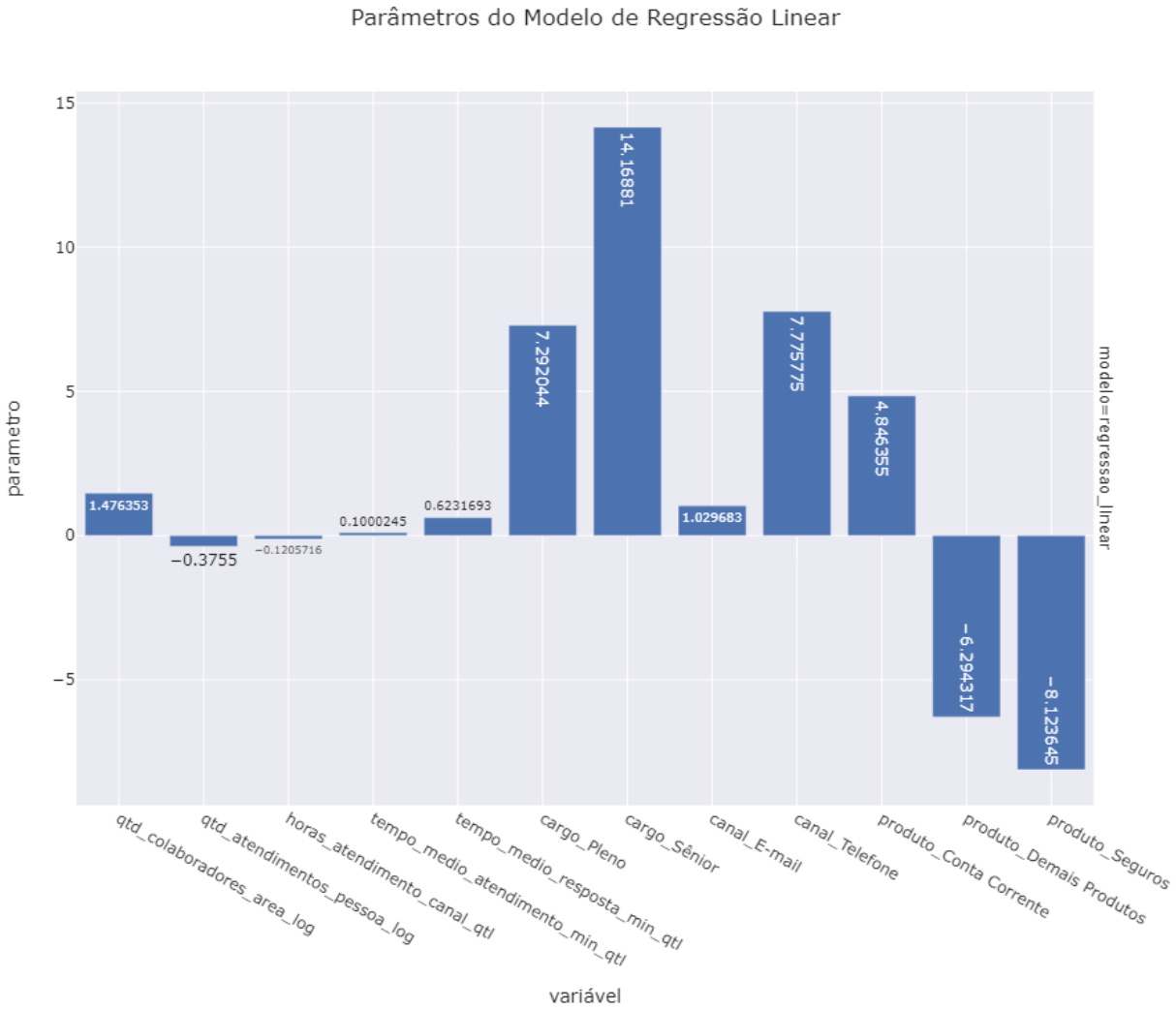
Na regressão linear, os parâmetros (Tabela 12 e Figura 35) são os coeficientes que determinam a linha (ou hiperplano, no caso de múltiplas variáveis) que melhor se ajusta aos dados. Cada um desses coeficientes indica o efeito dessa variável específica sobre a variável dependente, mantendo todas as outras variáveis constantes.

Tabela 12 – Parâmetros do Modelo de Regressão Linear Múltipla

	variável	modelo	parametro
0	qtd_colaboradores_area_log	regressao_linear	1.476353
1	qtd_atendimentos_pessoa_log	regressao_linear	-0.375500
2	horas_atendimento_canal_qtl	regressao_linear	-0.120572
3	tempo_medio_atendimento_min_qtl	regressao_linear	0.100024
4	tempo_medio_resposta_min_qtl	regressao_linear	0.623169
5	cargo_Pleno	regressao_linear	7.292044
6	cargo_Sênior	regressao_linear	14.168808
7	canal_E-mail	regressao_linear	1.029683
8	canal_Telefone	regressao_linear	7.775775
9	produto_Conta Corrente	regressao_linear	4.846355
10	produto_Demais Produtos	regressao_linear	-6.294317
11	produto_Seguros	regressao_linear	-8.123645

Fonte: Elaborado pelo autor no Google Colab.

Figura 35 – Gráfico dos parâmetros do Modelo de Regressão Linear Múltipla



Fonte: Elaborado pelo autor no Google Colab.

Considerando que 62.856007357050885 é o valor que intersecciona a equação da regressão no eixo y e que a variável da primeira linha (item 0 da Tabela 12) “qtd_colaboradores_area_log” representa X1, a segunda “qtd_atendimentos_pessoa_log” representa X2, a terceira “horas_atendimento_canal_qtl” representa X3 e, assim, sucessivamente, a Equação 4 é a que melhor representa a regressão, arredondando em 2 casas decimais os coeficientes:

Equação 4: Equação da Regressão Linear Múltipla

$$Y = 62{,}86 + 1{,}48 * X1 - 0{,}36 * X2 - 0{,}12 * X3 + 0{,}10 * X4 + 0{,}62 * X5 + 7{,}29 * X6 + 14{,}17 * X7 + 1{,}03 * X8 + 7{,}78 * X9 + 4{,}85 * X10 - 6{,}29 * X11 - 8{,}12 * X12$$

Fonte: Elaborado pelo autor.

Sendo assim, pode-se inferir que as variáveis que têm maior efeito sobre a variável respostas (o NPS) no modelo de Regressão Linear Múltipla são: “cargo” em geral, “canal_Telefone” e “produto” em geral. Isto é, são as variáveis que possuem maior impacto sobre o NPS do Banco A e, consequentemente, o banco deveria focar para obter melhores resultados de satisfação dos clientes segundo este modelo.

4.3.2 Modelo de Regressão Linear de Lasso

Assim como o modelo de Regressão Linear Múltipla, o modelo Lasso também é um modelo de regressão linear e, portanto, possui a mesma interpretação dos parâmetros. A única diferença é a necessidade de configurar hiperparâmetros.

Hiperparâmetros são configurações que podem ser ajustadas antes do treinamento de um modelo de aprendizado de máquina e sua seleção pode ter um impacto significativo no desempenho do modelo. Distinto dos parâmetros do modelo, que são aprendidos durante o treinamento, os hiperparâmetros devem ser configurados e ajustados pelo modelador.

No caso do trabalho, o hiperparâmetro a ser configurado para o modelo de regressão linear de Lasso é:

- Número máxima de iterações (*max_iter*);

Para otimizar a seleção do hiperparâmetro, foi utilizada uma técnica chamada “*Grid Search*” com “*5 - Fold Cross Validation*”. Nesta técnica, os modeladores definem um conjunto de possíveis valores para cada hiperparâmetro (Figura 36) e o *Grid Search* examina todas as possíveis combinações desses valores. Para cada combinação, o modelo é treinado e avaliado

usando uma métrica de desempenho específica (erro quadrado médio para regressão). Assim, a combinação de hiperparâmetros que produz o melhor desempenho será a escolhida (Figura 37).

Figura 36 – Valores imputados para busca da melhor combinação no *Grid Search*

```
modelo = Lasso(random_state = 123)
param_grid = {"max_iter": [1000,3000,5000]}
```

Fonte: Elaborado pelo autor no Google Colab.

Figura 37 – Escolha da melhor combinação pelo *Grid Search*

```
lasso.best_params_
```

```
{'max_iter': 1000}
```

Fonte: Elaborado pelo autor no Google Colab.

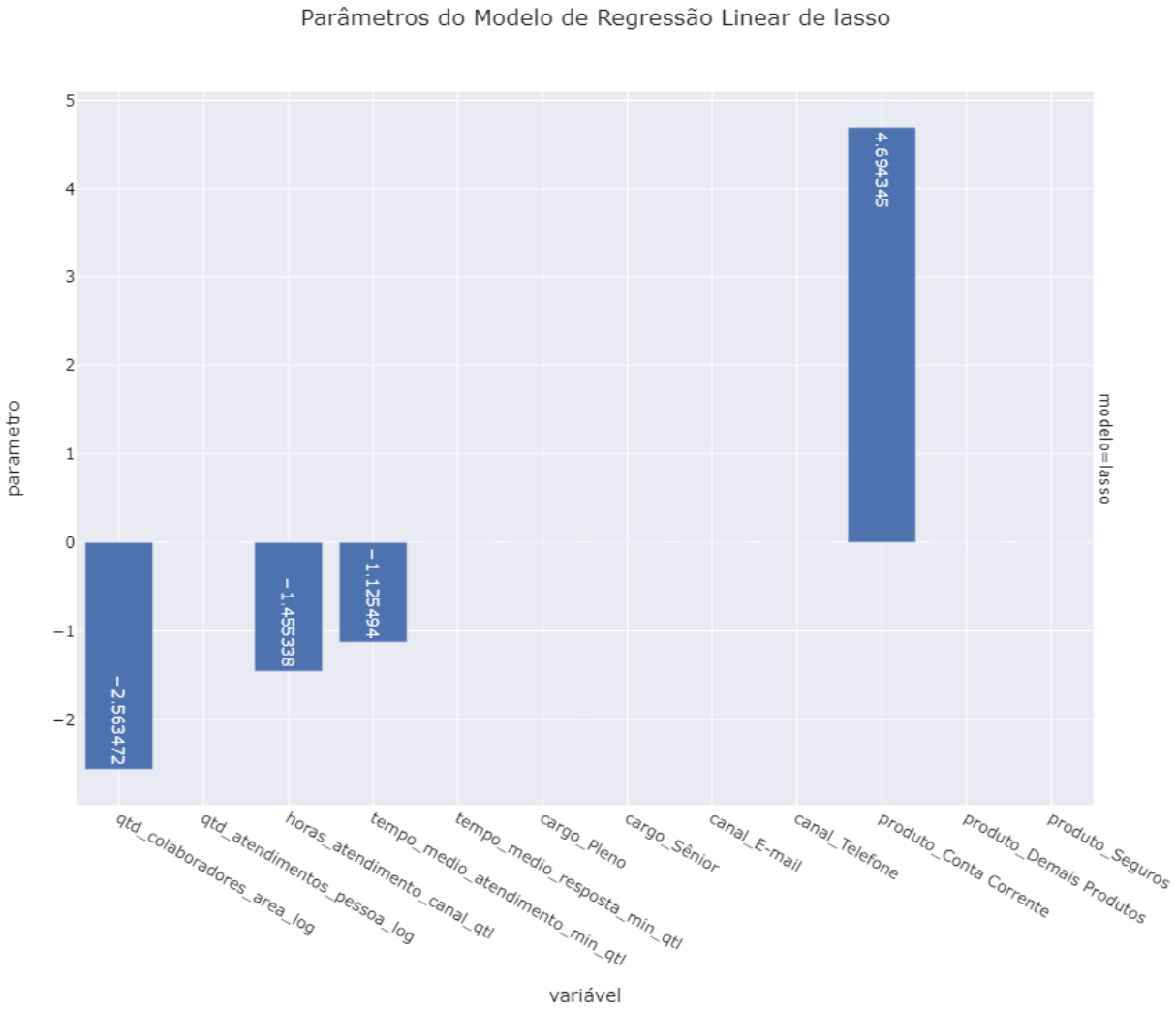
Dado os hiperparâmetros selecionados o modelo foi testado:

Tabela 13 – Parâmetros do Modelo de Regressão Linear de Lasso

	variável	modelo	parametro
0	qtd_colaboradores_area_log	lasso	-2.563472
1	qtd_atendimentos_pessoa_log	lasso	-0.000000
2	horas_atendimento_canal_qtl	lasso	-1.455338
3	tempo_medio_atendimento_min_qtl	lasso	-1.125494
4	tempo_medio_resposta_min_qtl	lasso	-0.000000
5	cargo_Pleno	lasso	0.000000
6	cargo_Sênior	lasso	0.000000
7	canal_E-mail	lasso	-0.000000
8	canal_Telefone	lasso	0.000000
9	produto_Conta Corrente	lasso	4.694345
10	produto_Demais Produtos	lasso	-0.000000
11	produto_Seguros	lasso	-0.000000

Fonte: Elaborado pelo autor no Google Colab.

Figura 38 – Gráfico dos parâmetros do Modelo de Regressão Linear de Lasso



Fonte: Elaborado pelo autor no Google Colab.

Como pode ser observado na Tabela 13 e Figura 38, as variáveis que têm maior efeito sobre o NPS no modelo de Regressão Linear de Lasso são: “produto_Conta Corrente”, “qtd_colaboradores_area_log”, “horas_atendimento_canal_qtl” e “tempo_medio_atendimento_min_qtl”. Lembrando que este método zera o coeficiente das variáveis que não têm efeito significativo.

Além disso, considerando que 69.27082167917024 é o valor que intersecciona a equação da regressão no eixo y e que a variável da primeira linha (item 0 da Tabela 12) “qtd_colaboradores_area_log” representa X1, a segunda “qtd_atendimentos_pessoa_log” representa X2, a terceira “horas_atendimento_canal_qtl” representa X3 e, assim, sucessivamente, a Equação 5 é a que melhor representa a regressão, arredondando em 2 casas decimais os coeficientes:

Equação 5: Equação da Regressão Linear de Lasso

$$Y = 69{,}27 - 2{,}56 * X1 - 1{,}46 * X3 - 1{,}13 * X4 + 4{,}69 * X10$$

Fonte: Elaborado pelo autor.

4.3.3 Modelo de Árvores de Decisão

Assim como no modelo Lasso, os modelos de árvore também necessitam do *input* de hiperparâmetros. No caso do trabalho, os hiperparâmetros a serem configurados para o modelo de Árvore de Decisão incluem:

- Profundidade máxima da árvore (*max_depth*);
- Número mínimo de amostras para divisão de um nó (*min_samples_split*);
- Número mínimo de amostras por folha (*min_samples_leaf*).

Também foi utilizada a técnica "*Grid Search*" com "*5 - Fold Cross Validation*". O conjunto de possíveis valores para cada hiperparâmetro foi definido (Figura 39) e o *Grid Search* examinou a melhor combinação (Figura 40).

Figura 39 – Valores imputados para busca da melhor combinação no *Grid Search*

```
modelo = DecisionTreeRegressor(random_state = 123)
param_grid = {"max_depth": [4, 5, 6, 7],
              "min_samples_split": [0.04, 0.06],
              "min_samples_leaf": [0.01]}
```

Fonte: Elaborado pelo autor no Google Colab.

Figura 40 – Escolha da melhor combinação pelo *Grid Search*

```
arv_dec.best_params_
```

```
{'max_depth': 7, 'min_samples_leaf': 0.01, 'min_samples_split': 0.04}
```

Fonte: Elaborado pelo autor no Google Colab.

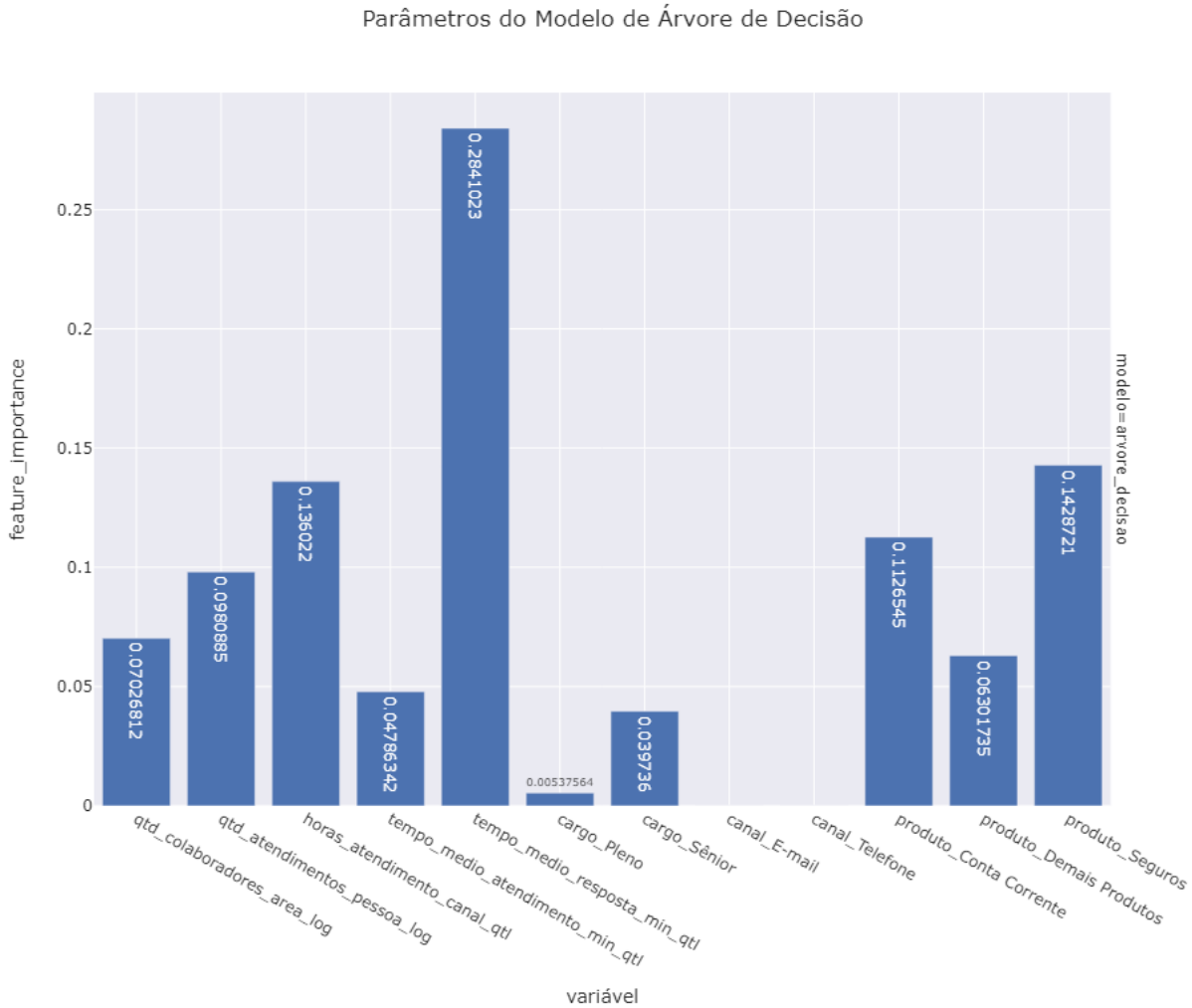
Dado os hiperparâmetros selecionados o modelo foi testado:

Tabela 14 – *Feature Importance* do Modelo de Árvore de Decisão

	variável	modelo	feature_importance
0	qtd_colaboradores_area_log	arvore_decisao	0.070268
1	qtd_atendimentos_pessoa_log	arvore_decisao	0.098089
2	horas_atendimento_canal_qtl	arvore_decisao	0.136022
3	tempo_medio_atendimento_min_qtl	arvore_decisao	0.047863
4	tempo_medio_resposta_min_qtl	arvore_decisao	0.284102
5	cargo_Pleno	arvore_decisao	0.005376
6	cargo_Sênior	arvore_decisao	0.039736
7	canal_E-mail	arvore_decisao	0.000000
8	canal_Telefone	arvore_decisao	0.000000
9	produto_Conta Corrente	arvore_decisao	0.112655
10	produto_Demais Produtos	arvore_decisao	0.063017
11	produto_Seguros	arvore_decisao	0.142872

Fonte: Elaborado pelo autor no Google Colab.

Figura 41 – Gráfico de *feature Importance* do Modelo de Árvore de Decisão



Fonte: Elaborado pelo autor no Google Colab.

Segundo a Tabela 14 e a Figura 41, as variáveis que têm maior importância sobre o NPS no modelo de Árvore de Decisão são: em primeiro lugar “tempo_medio_resposta_min_qtl”, seguido por “produto_Seguros” e em terceiro “horas_atendimento_canal_qtl”. Pode-se observar a imagem da Árvore de Decisão dividida em três partes no Anexo 4.

4.3.4 Modelo de Florestas Aleatórias

Os modelos de florestas aleatórias também precisam da configuração de hiperparâmetros. No caso do trabalho, os hiperparâmetros a serem configurados para o modelo de Florestas Aleatórias incluem:

- Profundidade máxima da árvore (*max_depth*);
- Número mínimo de amostras para divisão de um nó (*min_samples_split*);
- Número mínimo de amostras por folha (*min_samples_leaf*);
- Número de árvores em uma floresta (*n_estimators*).

Utilizando novamente a técnica de “*Grid Search*” com “*5 - Fold Cross Validation*” tem-se o conjunto de possíveis valores para cada hiperparâmetro (Figura 42) e a combinação de hiperparâmetros que produz o melhor desempenho (Figura 43).

Figura 42 – Valores imputados para busca da melhor combinação no *Grid Search*

```
modelo = RandomForestRegressor(random_state = 123)
param_grid = {"n_estimators": [100, 200],
              "max_depth": [4, 5, 6, 7],
              "min_samples_split": [0.04, 0.06],
              "min_samples_leaf": [0.01]}
```

Fonte: Elaborado pelo autor no Google Colab.

Figura 43 – Escolha da melhor combinação pelo *Grid Search*

```
rand_for.best_params_
```

```
{'max_depth': 7,
 'min_samples_leaf': 0.01,
 'min_samples_split': 0.04,
 'n_estimators': 200}
```

Fonte: Elaborado pelo autor no Google Colab.

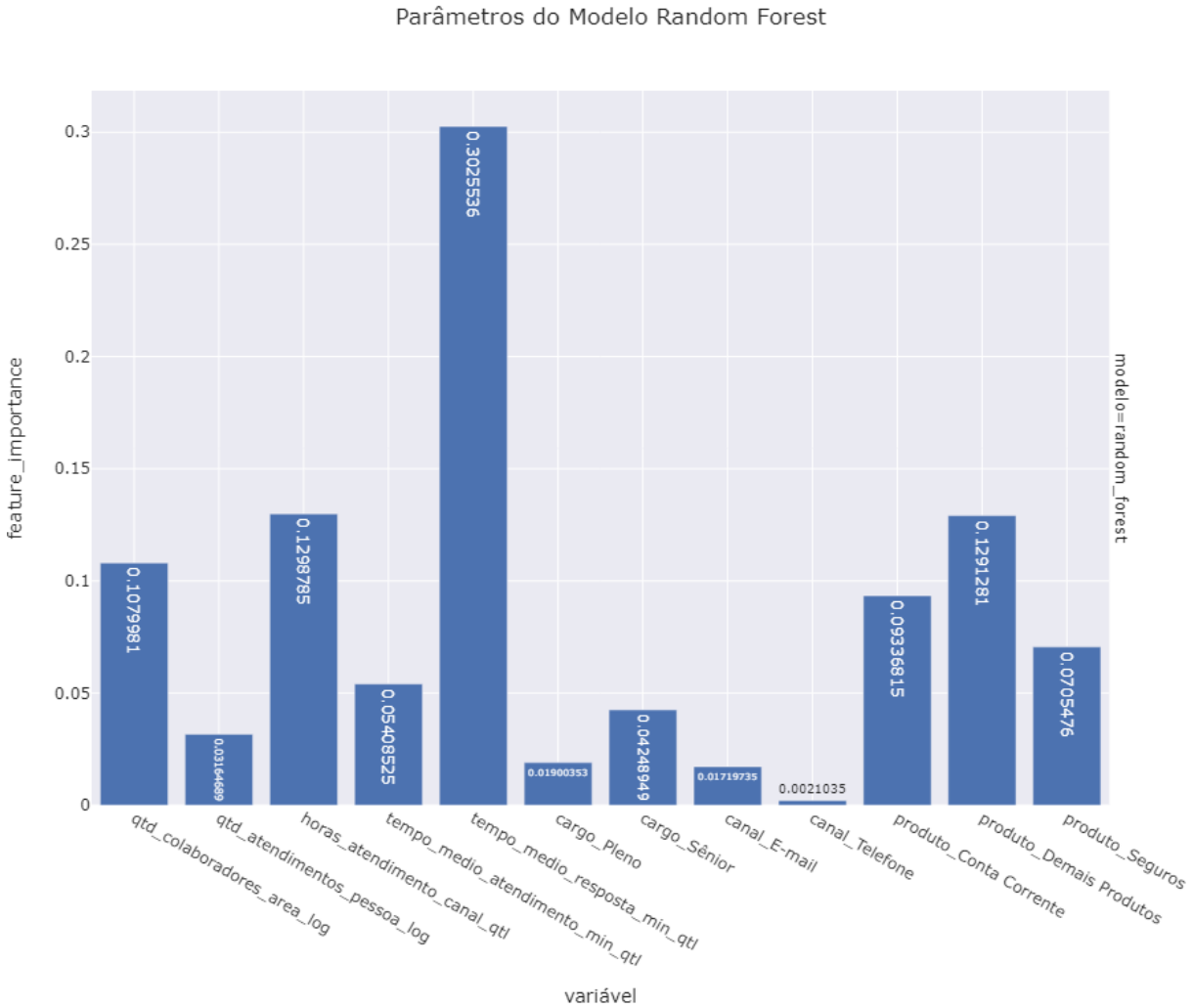
Dado os hiperparâmetros selecionados o modelo foi testado:

Tabela 15 – *Feature Importance* do Modelo *Random Forest*

	variável	modelo	feature_importance
0	qtd_colaboradores_area_log	random_forest	0.107998
1	qtd_atendimentos_pessoa_log	random_forest	0.031647
2	horas_atendimento_canal_qtl	random_forest	0.129878
3	tempo_medio_atendimento_min_qtl	random_forest	0.054085
4	tempo_medio_resposta_min_qtl	random_forest	0.302554
5	cargo_Pleno	random_forest	0.019004
6	cargo_Sênior	random_forest	0.042489
7	canal_E-mail	random_forest	0.017197
8	canal_Telefone	random_forest	0.002104
9	produto_Conta Corrente	random_forest	0.093368
10	produto_Demais Produtos	random_forest	0.129128
11	produto_Seguros	random_forest	0.070548

Fonte: Elaborado pelo autor no Google Colab.

Figura 44 – Gráfico de *feature Importance* do Modelo *Random Forest*



Fonte: Elaborado pelo autor no Google Colab.

Analisando a Tabela 15 e a Figura 44, as variáveis que têm maior importância sobre o NPS no modelo de Floresta Aleatória são: em primeiro lugar "tempo_medio_resposta_min_qtl", seguido por "horas_atendimento_canal_qtl" e em terceiro "produto_Demais Produtos".

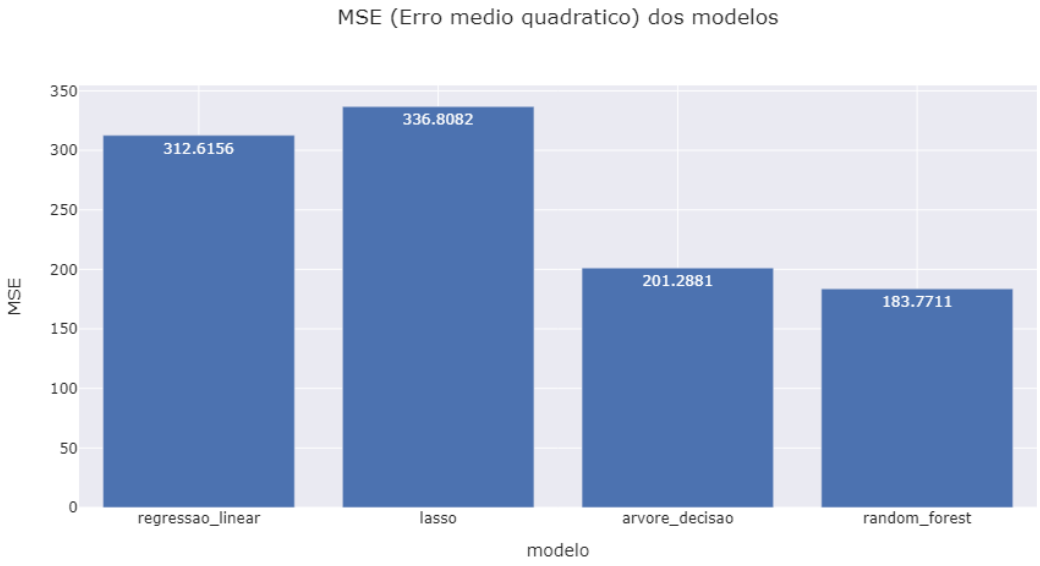
4.4 Fase "Avaliação (*Evaluation*)"

A etapa de Avaliação, no *framework* CRISP-DM, é quando os modelos de *data mining* desenvolvidos na fase de Modelagem são cuidadosamente analisados para determinar se eles satisfazem os requisitos do projeto (CHAPMAN et al., 2000). Este é um momento crucial em qualquer projeto de mineração de dados, pois é quando se estabelece a confiança de que os esforços realizados nas etapas anteriores irão entregar os resultados esperados.

Durante a Avaliação, vários aspectos dos modelos são examinados. Isso inclui a avaliação do desempenho dos modelos, revisão dos resultados, considerações do contexto do projeto e decisão sobre implantação.

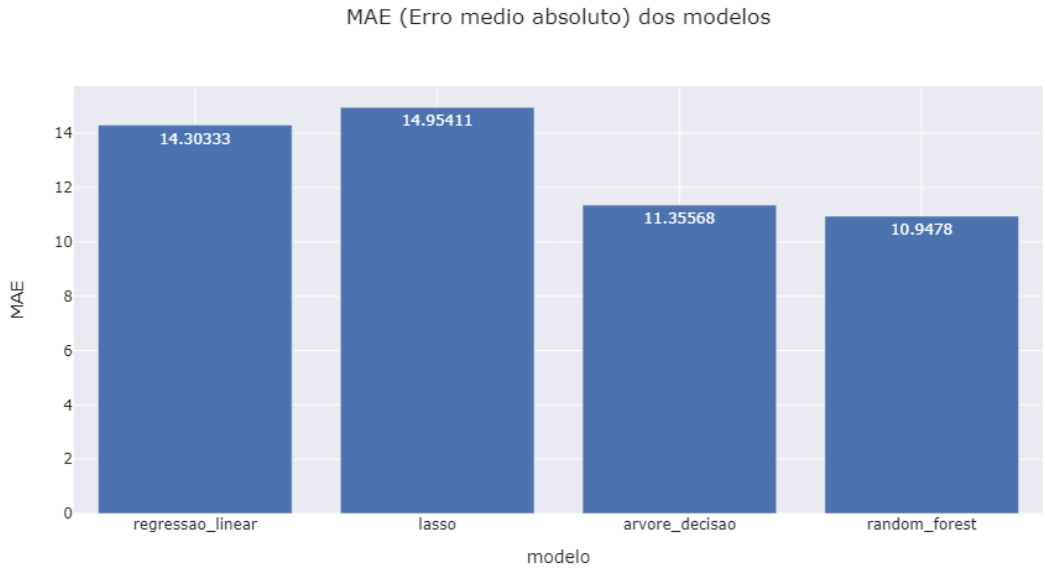
Para a análise de desempenho dos modelos foram utilizadas dois tipos de medidas de erro: o Erro Médio Quadrático (Figura 45) e o Erro Médio Absoluto (Figura 46).

Figura 45 – Gráfico de comparação do MSE dos modelos



Fonte: Elaborado pelo autor no Google Colab.

Figura 46 – Gráfico de comparação do MAE dos modelos



Fonte: Elaborado pelo autor no Google Colab.

Como esperado, o modelo *Random Forest* é o que apresentou melhor desempenho dentre os modelos elaborados, pois é o que teve menor erro nos dois tipos comparados. Ademais, este erro é aceitável para o contexto do projeto, visto que o NPS dentro do Banco A possui tolerância de margem de erro de mais ou menos 10 pontos, erro comparável ao MAE do modelo de Floresta Aleatória.

4.5 Fase "Implantação (*Deployment*)"

A etapa de Implantação do CRISP-DM é o ponto no qual os modelos de *data mining* são organizados e apresentados de maneira que possam ser usados pelo cliente do negócio ou equipe de tecnologia da informação (CHAPMAN et al., 2000). Como o projeto em questão priorizado na área em que o autor trabalha no Banco A não faz parte de uma área de ciência de dados, mas sim uma área de negócio que foi apoiada por uma análise mais especialista, o foco nessa etapa do CRISP-DM será em analisar os impactos do modelo no negócio (tarefa 4 mencionada no item 2.3.1.6 da revisão bibliográfica).

Para cumprir com as 3 primeiras tarefas dessa etapa do CRISP-DM (planejamento da implantação; implementação dos modelos e conhecimentos; monitorização e manutenção dos modelos), foi feita uma proposta de sugestão para a coordenação e gerência da área para que esse projeto fosse incluído no *backlog* da Tecnologia da Informação e pudesse ser mantido por pessoas especialistas em ciências de dados.

Para apoiar a importância e impacto desse projeto, os resultados mostrados no item 4.4 desse trabalho foram apresentados e documentados na ferramenta Google Colabs, facilitando o acesso, compreensão e manipulação do modelo e seus *outputs*.

Trazendo o foco para a última tarefa, um dos objetivos desse trabalho era indicar próximos passos e fazer recomendações para a área de Central de Atendimento melhorar o NPS dos seus serviços. Segundo os *insights* do modelo de Floresta Aleatória, o de melhor desempenho, as variáveis de maior importância, em ordem decrescente, foram: "tempo_medio_resposta_min_qtl" (tempo de espera para receber uma resposta), seguido por "horas_atendimento_canal_qtl" (horas de atendimento médio por pessoa em cada canal) e "produto_Demais Produtos" (assunto referente ao atendimento focado em "demais produtos").

Sendo assim, foram feitas sugestões iniciais de hipóteses do que pode melhorar o NPS com base nessas três variáveis. Essas sugestões serão levadas adiante para que se desdobrem em projetos multidisciplinares.

Considerando o tempo médio de resposta, concluiu-se, segundo as análises, que permitir que o cliente permaneça em uma longa espera por uma resposta tende a aumentar o NPS, o que pode ser contra-intuitivo. A hipótese levantada é que esta espera maior pode estar resultando em melhores resoluções de problema dos clientes, ou até levar a um sentimento de maior acolhimento. Para complementar os resultados quantitativos obtidos nessa pesquisa, o autor sugere que os atendimentos com maior tempo e maior NPS sejam analisados de forma qualitativa para que se comprove ou refute essa hipótese. Com base no resultado, sugere-se revisar o plano de incentivos dados aos funcionários da central, para que se alinhe com as ações que trarão maior satisfação. Por exemplo, o foco pode ser em priorizar o sentimento do cliente e resolver seu problema, e não o volume ou tempo rápido da ligação. Para esse plano de ação, seria necessário envolver a gerência da central de atendimentos e planejar treinamentos e reforços culturais.

No que tange a métrica relativa às horas dedicadas ao atendimento em cada canal, observou-se que quanto mais tempo o atendente passa em um canal, menor tende a ser a nota de NPS. A hipótese é de que uma quantidade excessiva de tempo requerido em cada um pode estar dificultando o desempenho e a qualidade do serviço oferecido pelos profissionais da central de atendimento, isto é, pode ser que eles estejam ficando saturados em atuar em um canal por longos períodos. Uma abordagem possível para amenizar esse problema seria adotar uma rotina com intervalos mais frequentes proporcionados aos atendentes. Esta estratégia permitiria que esses colaboradores pudessem recarregar suas energias e manter sua produtividade elevada, beneficiando tanto a equipe quanto os clientes atendidos.

Outra estretégia interessante seria entender melhor a rotina de trabalho dos atendentes utilizando o método "*go to Gemba*", isto é, os gerentes e os tomadores de decisão irem visitar o local de trabalho para observar diretamente o processo e entender completamente quaisquer problemas que estejam ocorrendo, ao invés de confiar unicamente em relatórios ou métricas (SUZAKI, 1987).

Finalmente, com relação aos demais produtos, é imprescindível realizar uma análise mais aprofundada para identificar quais deles estão influenciando positivamente ou negativamente o NPS. Essa investigação pode ser complexa e essa variável possui uma importância terciária quando comparada às duas anteriores, desta forma, a recomendação do autor é concentrar esforços inicialmente nas duas primeiras possibilidades. Assim, pode-se alimentar o modelo com novos dados e, em seguida, com base nos novos *insights* adquiridos, reorientar o foco para otimizar outros aspectos, incluindo a performance dos produtos.

5 CONCLUSÃO

5.1 Discussão Final

Este trabalho buscou aplicar o *framework* CRISP-DM para desenvolver e comparar quatro modelos de aprendizado de máquina - Regressão Linear Múltipla, Regressão Linear Lasso, Árvore de Decisão e Floresta Aleatória - com o objetivo de prever o *Net Promoter Score* (NPS) de uma central de atendimento de um tradicional banco brasileiro e identificar as variáveis que mais o influenciam.

Durante a fase de compreensão do negócio, foi identificado o NPS como uma métrica chave para o desempenho da central de atendimento ao cliente, com o potencial de oferecer *insights* valiosos para o aprimoramento do serviço.

Na fase de compreensão dos dados, foi realizada uma análise exploratória detalhada dos dados disponíveis, o que permitiu identificar características importantes que possivelmente influenciam o NPS.

Ao chegar na fase de modelagem, foram experimentados quatro modelos diferentes. A Regressão Linear Múltipla, dada sua simplicidade e interpretabilidade, forneceu uma primeira aproximação para entender a relação entre as variáveis e o NPS. A Regressão Linear Lasso, com sua capacidade de realizar a seleção de características, exibiu de forma mais enfática quais variáveis tinham maior influência sobre o NPS. A Árvore de Decisão e a Floresta Aleatória, sendo modelos não lineares, foram capazes de capturar relações mais complexas, com a Floresta Aleatória apresentando o melhor desempenho entre os quatro modelos.

Depois, na fase de avaliação, os modelos foram avaliados e validados, garantindo que eles tinham uma performance satisfatória.

Finalmente, na fase de implementação, foram apresentadas algumas recomendações para melhoria do NPS e, consequentemente, da satisfação dos clientes em relação aos atendimentos realizados pela central de atendimentos.

No geral, este trabalho demonstrou a eficácia dos modelos de aprendizado de máquina na previsão do NPS e enfatizou a importância de um processo estruturado, como o CRISP-DM, no desenvolvimento e avaliação desses modelos. Além disso, destacou a necessidade de equilibrar a complexidade dos modelos e a interpretabilidade, considerando tanto o desempenho quanto a compreensibilidade dos resultados para os tomadores de decisão.

Ao explorar esses quatro modelos, este trabalho sublinhou a importância do aprendizado de máquina na análise de dados do setor bancário, mais especificamente, na previsão do NPS.

Os *insights* obtidos a partir desta análise podem auxiliar o Banco A a identificar áreas para melhorias em seu serviço de atendimento ao cliente.

No entanto, é importante lembrar que, apesar do desempenho promissor dos modelos, eles são uma ferramenta para auxiliar na tomada de decisões. Decisões estratégicas ainda precisam levar em consideração uma gama de fatores, incluindo, mas não se limitando a, interpretações dos dados.

Por fim, espera-se que este trabalho inspire futuras pesquisas na aplicação de métodos de aprendizado de máquina no setor bancário, e em particular, no contexto de experiência do cliente.

5.2 Objetivos

No item 1.4 foram enunciados os principais objetivos do trabalho. O primeiro deles foi "Trazer a posição do Banco A dentro dos *frameworks* de estratégia de NPS". Esse objetivo foi cumprido ao longo do desenvolvimento do trabalho, pois foram apresentadas as vantagens estratégicas e a relevância da metodologia NPS, sobretudo na revisão bibliográfica. Ademais, o NPS foi fundamental para tornar tangível e palpável matematicamente a satisfação do cliente, uma percepção subjetiva.

O segundo objetivo "Investigar os dados do *contact center* criando um modelo matemático de previsão do NPS dessa área do banco" também foi alcançado durante a execução do trabalho. Isso se deu pela criação de um modelo robusto permitindo uma previsão confiável do NPS na Central de Atendimento. Este modelo viabilizou uma investigação mais detalhada das variáveis que têm impacto direto na satisfação do cliente.

Por último, o terceiro objetivo era "Propor um enfoque para melhoria dos resultados de satisfação a partir dos insumos do modelo, descrevendo quais características do *contact center* o banco pode focar para obter melhores resultados.". Esse objetivo foi concretizado, pois com os *insights* extraídos do modelo preditivo, foi fornecida uma orientação interessante para futuros esforços de aprimoramento, assim, a área do *contact center* pode utilizar seus esforços naquilo que realmente irá impactar a nota de satisfação dos atendimentos.

5.3 Aprendizados

O autor vivenciou a criação e execução de um projeto prático, essencial para o funcionamento da empresa em que estava integrado. Este contato direto com a solução de

problemas rotineiros dentro do ambiente corporativo proporcionou uma visão clara do papel e da relevância do engenheiro na dinâmica empresarial.

O projeto possibilitou a aplicação prática dos temas abordados durante o curso de Engenharia de Produção na Escola Politécnica, complementando e reforçando o conhecimento teórico acumulado ao longo do curso. Simultaneamente, houve a oportunidade de adquirir e desenvolver conhecimentos novos, especificamente relacionados à *Machine Learning* e Experiência do Cliente.

Em nível profissional, o autor foi capaz de obter um entendimento mais profundo sobre o negócio da empresa, o que certamente será útil para sua progressão de carreira dentro da organização.

Em suma, esta experiência se mostrou enriquecedora não somente pela aplicação de conhecimentos adquiridos durante o curso de Engenharia de Produção, mas também pelo aprendizado acerca das tendências do setor bancário e como elas estão sendo incorporadas em nossa realidade. Este projeto representou, assim, um passo significativo no caminho da evolução pessoal e profissional do autor.

5.4 Limitações

Ao longo do desenvolvimento do projeto, houve algumas limitações que merecem ser destacadas. Em primeiro lugar, embora foram construídos modelos e identificadas variáveis de grande importância, não foi realizada uma etapa para retirar as variáveis pouco relevantes e retreinar o modelo. Esse é um aspecto que poderia ter sido aprimorado e gerado melhores resultados para os modelos, porém não foi realizado, pois não foi priorizado dentro dos objetivos e cronograma.

Outra limitação diz respeito à seleção dos modelos. Existem alternativas mais robustas de modelos que poderiam potencialmente gerar melhores resultados, como o XGBoost e o Light GBM baseados em árvores e o SVM (Support Vector Machine) como opção mais robusto de modelo linear. Optou-se, contudo, por modelos de mais fácil interpretação, visando facilitar o entendimento e a utilização dos resultados por parte dos *stakeholders*.

Por fim, outro aspecto limitante foi a sensibilidade dos dados tratados. Por envolverem questões delicadas, não foi possível obter um volume de dados tão grande quanto se desejava.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ALLISON, P. D. **Missing Data**. Sage Publications, 2002.

BERGSTRA, J.; BENGIO, Y. **Random search for hyper-parameter optimization.** Journal of Machine Learning Research, p. 281-305, 2012.

BERRY, L. L.; SEIDERS, K.; GREWAL, D. **Understanding service convenience**. Journal of Marketing, 66(3), p. 1-17, 2002.

BOUGH, V.; EHRLICH, O.; FANDERL, H.; SCHIFF, R. **Experience-led growth: A new way to create value**. McKinsey & Company, 2023. Disponível em < https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/experience-led-growth-a-new-way-to-create-value > Acesso em: 20 abr. 2023.

BREIMAN, L. **Random Forests**. Machine Learning, 45(1), p. 5-32, 2001.

CASELLA, G.; BERGER, R. L. **Statistical Inference (2nd Edition)**. Duxbury Advanced Series, 2002.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0: Step-by-step data mining guide**. SPSS Inc, 2000.

CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System.** In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

DAVENPORT, T. H.; HARRIS, J. G. **Competing on analytics: The new science of winning**. Harvard Business Press, 2007.

DIXON, M.; FREEMAN, K.; TOMAN, N. **Stop trying to delight your customers**. Harvard Business Review, 88(7-8), p. 116-122, 2010.

DODDS, W. B.; MONROE, K. B.; GREWAL, D. **Effects of price, brand, and store information on buyers' product evaluations**. Journal of Marketing Research, 28(3), p. 307-319, 1991.

DORAN, G. T. **There's a S.M.A.R.T. way to write management's goals and objectives**. Management Review, 70(11), p. 35-36, 1981.

FARIA, E. **Fintechs de crédito e intermediários financeiros: uma análise comparativa de eficiência.** Curso de Pós Graduação em Administração, Universidade de São Paulo (USP). Dissertação de Mestrado, São Paulo, 2018. Disponível em < https://www.teses.usp.br/teses/disponiveis/12/12142/tde-07012019-112337/publico/CorrigidoEmerson.pdf >. Acesso em: 20 abr. 2023.

FEBRABAN. **Relatório Anual.** 2019. Disponível em < https://cmsarquivos.febraban.org.br/Arquivos/documentos/PDF/Relat%C3%B3rio%20anual%202019_pt.pdf > Acesso em: 20 abr. 2023.

FORRESTER CONSULTING. **The Business Impact of Customer Experience: How Experience-Driven Businesses Survive and Thrive in Uncertain Business Environments.** Forrester Research, junho, 2021. Disponível em < https://business.adobe.com/content/dam/dx/us/en/resources/reports/the-business-impact-of-investing-in-experience-forrester-thought-leadership-paper-2021/the-business-impact-of-investing-in-experience-forrester-thought-leadership-paper-2021.pdf > Acesso em: 20 abr. 2023.

GARTNER. **Gartner Customer Experience Management Survey.** 2019. Disponível em < https://www.gartner.com/en/marketing/research/customer-experience-survey > Acesso em: 20 abr. 2023.

GENTILE, C.; SPILLER, N.; NOCI, G. **How to sustain the customer experience: An overview of experience components that co-create value with the customer**. European Management Journal, 25(5), p. 395-410, 2007.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016.

GREWAL, D.; ROGGEVEEN, A. L.; NORDFÄLT, J. **The future of retailing**. Journal of Retailing, 93(1), p. 1-6, 2017.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer Series in Statistics, 2009.

HUANG, M.; LI, Y.; LIU, R. Y. **Transformation-Kernel Density Estimation of Multi-quantile Levels**. Journal of the American Statistical Association, 103(484), p. 1585–1594, 2008.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning (Vol. 112)**. Springer, 2013.

KEININGHAM, T. L.; COOIL, B.; ANDREASSEN, T. W.; AKSOY, L. **A longitudinal examination of net promoter and firm revenue growth**. Journal of Marketing, 71(3), p. 39-51, 2007.

KELLEHER, J. D.; MAC NAMEE, B.; D'ARCY, A. **Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies**. MIT Press, 2015.

KLAUS, P.; MAKLAN, S. **Towards a better measure of customer experience**. International Journal of Market Research, 55(2), 227-246, 2013.

KOHAVI, R. **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. Proceedings of the 14th International Joint Conference on Artificial Intelligence (Vol. 2, p. 1137-1143). Morgan Kaufmann, 1995.

KOTLER, P. **Atmospherics as a marketing tool**. Journal of Retailing, 49(4), p. 48-64, 1973.

KUMAR, V.; ANISH, A.; SONG, H. **The dynamics of marketing and sales capabilities in improving firm performance**. Journal of Marketing Research, 50(5), p. 661-677, 2013.

LITTLE, R. J.; & RUBIN, D. B. **Statistical Analysis with Missing Data**. John Wiley & Sons, 2002.

MARISCAL, G.; MARBÁN, Ó.; FERNÁNDEZ, C. **A survey of data mining and knowledge discovery process models and methodologies**. The Knowledge Engineering Review, 25(2), p. 137-166, 2010.

MARKEY, R.; REICHHELD, F.; DULLWEBER, A. **Closing the customer feedback loop**. Harvard Business Review, 87(12), p. 43-47, 2009. Disponível Em < https://hbr.org/2009/12/closing-the-customer-feedback-loop > Acesso em: 20 abr. 2023.

MARKEY, R., REICHHELD, F. **Introducing: The Net Promoter System®**. 2011. Disponível em < https://www.bain.com/pt-br/insights/introducing-the-net-promoter-system-loyalty-insights/ >. Acesso em: 20 abr. 2023.

MEYER, C.; SCHWAGER, A. **Understanding customer experience**. Harvard business review, 85(2), p. 116-126, 2007.

MORETTIN, P. A.; SINGER, J. M. **Estatística e Ciência de Dados**. Editora Blucher, 2021.

MORGAN, R. M.; REGO, L. L. **The value of different customer satisfaction and loyalty metrics in predicting business performance**. Marketing Science, 25(5), 426-439, 2006.

MÜLLER, A. C.; GUIDO, S. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. O'Reilly Media, Inc., 2016.

PEPPERS, D.; ROGERS, M. **Enterprise one to one: Tools for competing in the interactive age**. New York: Currency Doubleday, 1997.

PINE, B. J.; GILMORE, J. H. **The experience economy: Work is theatre & every business a stage**. Harvard Business Press, 1999.

PRAHALAD, C. K.; RAMASWAMY, V. **Co-creation experiences: The next practice in value creation**. Journal of Interactive Marketing, 18(3), p. 5-14, 2004.

REICHHELD, F. F. **The one number you need to grow**. Harvard Business Review, 81(12), p. 46-54, 2003. Disponível em < https://hbr.org/2003/12/the-one-number-you-need-to-grow > Acesso em: 20 abr. 2023.

REICHHELD, F. F. **The Ultimate Question: Driving Good Profits and True Growth**. Harvard Business Press, 2006.

REICHHELD, F.; MARKEY, R. G. **The Ultimate Question 2.0: How Net Promoter Companies Thrive in a Customer-Driven World**. Harvard Business Review Press, 2011.

ROSENBAUM, M. S.; MASSIAH C. **An expanded servicescape perspective**. Journal of Service Management, 22(4), p. 473, 2011.

RUBIN, D. B. **Multiple Imputation for Nonresponse in Surveys**. John Wiley & Sons, 1987.

SCHAFER, J. L.; GRAHAM, J. W. **Missing data: our view of the state of the art**. Psychological methods, 7(2), p. 147-177, 2002.

SCIKIT-LEARN DOCUMENTATION. **3.1. Cross-validation: evaluating estimator performance**. 2023. Disponível em < https://scikit-learn.org/stable/modules/cross_validation.html > Acesso em: 15 jun. 2023.

SCIKIT-LEARN DOCUMENTATION. **Map data to a normal distribution**. 2023. Disponível em < https://scikit-learn.org/stable/auto_examples/preprocessing/plot_map_data_to_normal.html > Acesso em: 15 jun. 2023.

SCIKIT-LEARN DOCUMENTATION. **Sklearn.linear_model.Lasso**. 2023. Disponível em < https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html > Acesso em 15 jun. 2023.

SHEARER, C. **The CRISP-DM model: The new blueprint for data mining**. Journal of Data Warehousing, 5(4), p. 13-22, 2000.

SILVER, L. **Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally**. PEW Research Center, 2019. Disponível em < https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/ > Acesso em: 20 abr. 2023.

SUZAKI, K. **The new manufacturing challenge: Techniques for continuous improvement**. New York: Free Press, 1987.

TIBSHIRANI, R. **Regression Shrinkage and Selection via the Lasso**. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), p. 267–288, 1996.

VALENTE, J. **Brasil tem 134 milhões de usuários de internet, aponta pesquisa.** Pesquisa TIC Brasil, Agencia Brasil, 2019: Disponível em < https://agenciabrasil.ebc.com.br/geral/noticia/2020-05/brasil-tem-134-milhoes-de-usuarios-de-internet-aponta-pesquisa > Acesso em: 20 abr. 2023.

VERHOEF, P. C.; LEMON, K. N.; PARASURAMAN, A.; ROGGEVEEN, A.; TSIROS, M.; SCHLESINGER, L. A. **Customer experience creation: Determinants, dynamics and management strategies**. Journal of retailing, 85(1), p. 31-41, 2009.

VERHOEF, P. C., REINARTZ, W. J., & KRAFFT, M. **Customer engagement as a new perspective in customer management**. Journal of Service Research, 13(3), p. 247-252, 2010.

WIRTH, R.; HIPP, J. CRISP-DM: **Towards a standard process model for data mining**. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). Springer-Verlag, 2000.

YEO, I.K.; JOHNSON, R. **A new family of power transformations to improve normality or symmetry**. Biometrika 87(4), p. 954-959, 2000.

ZEITHAML, V. A.; BERRY, L. L.; PARASURAMAN, A. **The behavioral consequences of service quality**. Journal of Marketing, 60(2), p. 31-46, 1996.

7 ANEXOS

ANEXO 1 – *Notebook* 1: Código em Python da Análise Exploratória de Dados

Import bibliotecas

```
#mount do colab
from google.colab import
drive
drive.mount('/content/gdri
ve')

#manipulacao de
dadosimport pandas
as pd import numpy
as np
from datetime import datetime

#visualizacao de dados
import matplotlib.pyplot as
pltimport seaborn as sns
import plotly.express as px
```

Import e ajuste da base

Import e ajuste das colunas da base de analise

```
#import da base de analise
df_analise =
                        pd.read_excel("./gdrive/MyDrive/projeto_cx
                        /case_cx_itau.xlsx",sheet_name = "BASE",
                        usecols = ["Data Final", "Cargo", "Produto", "Canal",
                                  "Qtd de atendimentos", "Qtd Colab", "TMA
                                  (segundos)","dedicação media de atend
                                  (horas/dia)",
                                  "NPS Médio", "Tempo medio de resposta (min)"])
#colunas da base
df_analise.columns = ["cargo", "canal", "horas_atendimento_canal","qtd_atendimentos",
                    "qtd_colaboradores_area", "tempo_medio_"nps_medio", "tempo_medio_resposta_min",
                    "data_final", "produto"]

#conversao do TMA de segundos para minutos
df_analise["tempo_medio_atendimento_min"] = df_analise["tempo_medio_atendimento_seg"]/60

#conversao da data de objeto para datetime
df_analise['data_final'] = pd.to_datetime(df_analise['data_final'], format='%d/%m/%Y')

#remoção da coluna de TMA de minutos
df_analise.drop(columns = "tempo_medio_atendimento_seg", inplace = True)

#visualização da
basedf_analise
```

Descrição das colunas da base:

- **cargo:** nível de senioridade dos atendentes de contact center (júnior,pleno e sênior)
- **canal:** veículo de atendimento (telefone,e-mail ou chat)
- **horas_atendimento_canal:** TMA x qtd_atendimentos / qtd_colaboradores

  OBS: Para se ter a dedicação total de cada atendente por dia basta somar os canais
- **qtd_atendimentos:** número de atendimentos realizados por todos os atendentes por canal em um dia
- **qtd_colaboradores_area:** número de atendentes por senioridade por canal no dia
- **nps_medio:** nota de NPS média dos atendentes no dia por canal por senioridade
- **tempo_medio_resposta_min:** tempo que o cliente espera para receber uma resposta do atendente
- **data_final**: data dos atendimentos
- **produto**: assunto tratado nos atendimentos clusterizado por produto (cartão de crédito, conta corrente, seguros e demais produtos)
- **tempo_medio_atendimento_min**: tempo médio que o atendente leva para realizar 1 atendimento (em segundos)

Analise de consistencia dos dados

Valores ausentes

Contagem de valores nulos por coluna

```
df_analise.isnull().sum()
```

Valores duplicados

Quantidade de valores duplicados na base

```
df_analise.duplicated().sum()
```

Valores unicos das variaveis categoricas

Valores unicos por coluna

```
for coluna in ["cargo", "canal", "produto"]:
    print(f"{coluna}:", df_analise[coluna].unique(), end
    = "\n")
```

Tipo do dado

```
df_analise.info()
```

Verificação de inconsistências

Verificação de atendimento nulo

```
df_analise.loc[df_analise.qtd_atendimentos == 0]
```

Existem 191 valores em que o atendimento é nulo, mas existe NPS. A maioria aparenta ser do cargo senior, canal de telefone e produto conta corrente. Vamos avaliar se existem valores preenchidos para esse filtro

```
df_analise.loc[(df_analise.cargo == "Sênior") &
              (df_analise.canal == "Telefone") &
              (df_analise.produto == "Conta
              Corrente")]
```

Como não existem valores preenchidos, pode ser que haja um erro de input na base. Para não perder a informação, vamos preencher com amediana do canal telefone e do produto conta corrente, com base nos cargos que não são senior

```
horas_atendimento_canal_input_senior, qtd_atendimentos_input_senior = df_analise.loc[(df_analise.canal ==
"Telefone") &
                                                                                     (df_analise.produto == "Conta
                                                                                     Corrente") (df_analise.cargo !=
                                                                                     "Sênior"), ["horas_a
print("horas_atendimento_canal_input_senior: ",
horas_atendimento_canal_input_senior)print("qtd_atendimentos_input_senior:
", qtd_atendimentos_input_senior)

df_analise.loc[(df_analise.cargo == "Sênior") &
              (df_analise.canal == "Telefone") &
              (df_analise.produto == "Conta Corrente"), ["horas_atendimento_canal", "qtd_atendimentos"]] =
              horas_atendimento
```

Vamos verificar quais valores de atendimento permanecem 0, re iremos removê-los da base

```
df_analise.loc[df_analise.qtd_atendimentos == 0]

df_analise =
df_analise.loc[df_analise.qtd_atendimentos > 0]
len(df_analise)
```

EDA

Analise Univariada

Variaveis Categoricas

Análise Gráfica

Plot da contagem de todas as variaveis categoricas

```
for coluna in ["cargo", "canal", "produto", "data final"]: df grafico = df analise.groupby(coluna, as index =
```

```
False)\
                        .qtd_atendimentos.count()\
                        .rename(columns = {"qtd_atendimentos": "contagem"})
    px.bar(df_grafico, x = coluna, y = "contagem", template = "seaborn",width=1000,
    height=500,
          text_auto = True, title = f"Gráfico de contagem de {coluna}").show()
```

Variaveis Numericas

```
#pd.to_datetime(df_analise.data_final)
```

Estatística Descritiva

```
df_analise.select_dtypes(include = ["number"]).describe()
```

Análise Gráfica

ECDF

```
for     coluna     in     ["horas_atendimento_canal",     "qtd_atendimentos",
             "qtd_colaboradores_area",                          "nps_medio",
             "tempo_medio_resposta_min",     "tempo_medio_atendimento_min"]:
             px.ecdf(df_analise,
                        x = coluna,
                        title = f"ECDF de {coluna}",
                        template = "seaborn",width=400, height=400).show()
```

Histograma

```
for     coluna     in     ["horas_atendimento_canal",     "qtd_atendimentos",
             "qtd_colaboradores_area",                          "nps_medio",
             "tempo_medio_resposta_min",     "tempo_medio_atendimento_min"]:
             px.histogram(df_analise,
                        x = coluna,
                        title = f"Histograma de {coluna}",
                        template = "seaborn",width=500, height=500).show()
```

Analise Bivariada

```
def grafico_bivariado(df, x, y = "nps_medio", grafico = "categorico", height = 600,
  width = 600):# função que plota o grafico bivariado de duas variaveis x e y
  if grafico ==
    "categorico":
    px.box(df,
        y =
        y,x =
        x,
        title = f"Boxplot de {x} vs
           {y}", height = height, width
           = width,
        template = "seaborn"
        ).show()
  if grafico ==
    "numerico":
    px.scatter(df,
              x =
              x,y =
              y,
              title = f"Scatterplot de {x} vs
              {y}",height = height, width =
              width, template = "seaborn",
              trendline = "ols").show()
  return

def estatistica_descritiva_bivariado(df, x, y = "nps_medio"):
  # função que fornece a estatistica descritiva de acordo com uma

  categoriadisplay(df_analise.groupby(x, as_index =

  True)[[y]].describe())

  return
```

Correlação entre as variáveis

```
df analise.corr(method = "pearson")
```

Analise 1: Senioridade vs NPS

- O que gostariamos de testar e porquê

```
grafico_bivariado(df_analise, x = "cargo", y = "nps_medio", grafico = "categorico", height = 600, width = 600)

estatistica_descritiva_bivariado(df_analise, x = "cargo", y = "nps_medio")
```

Analise 2: Especialidade vs. NPS

```
grafico_bivariado(df_analise, x = "produto", y = "nps_medio", grafico = "categorico", height = 600, width =
600)

estatistica_descritiva_bivariado(df_analise, x = "produto", y = "nps_medio")
```

Analise 3: Canal vs.NPS

```
grafico_bivariado(df_analise, x = "canal", y = "nps_medio", grafico = "categorico", height = 600, width = 600)

estatistica_descritiva_bivariado(df_analise, x = "canal", y = "nps_medio")
```

Analise 4: Horas de atendimento vs. NPS

```
grafico_bivariado(df_analise, x = "horas_atendimento_canal", y = "nps_medio", grafico = "numerico", height =
600, width = 100
```

Analise 5: Qtd de atendimento e qtd de atendimento/pessoa vs. NPS:

```
grafico_bivariado(df_analise, x = "qtd_atendimentos", y = "nps_medio", grafico = "numerico", height = 600,
width = 1000)

df_analise['qtd_atendimentos_pessoa'] = df_analise['qtd_atendimentos']/df_analise['qtd_colaboradores_area']
grafico_bivariado(df_analise, x = "qtd_atendimentos_pessoa", y = "nps_medio", grafico = "numerico", height =
600, width = 100
```

Analise 6: TMA vs. NPS

```
grafico_bivariado(df_analise, x = "tempo_medio_atendimento_min", y = "nps_medio", grafico = "numerico", height
= 600, width =
```

Analise 7: Tempo médio de resposta vs. NPS

```
grafico_bivariado(df_analise, x = "tempo_medio_resposta_min", y = "nps_medio", grafico = "numerico", height =
600, width = 10
```

Analise 8: Dia da semana vs. NPS

```
# Extrai dia da semana como string
df_analise['dia_da_semana'] = df_analise['data_final'].apply(lambda x:

datetime.strftime(x, '%A'))df_analise.dia_da_semana

grafico_bivariado(df_analise, x = "dia_da_semana", y = "nps_medio", grafico = "categorico", height = 600,
width = 600)

estatistica_descritiva_bivariado(df_analise, x = "dia_da_semana", y = "nps_medio")

df_analise.corr(method = "pearson")
```

Analise Multivariada

Analise 1: Senioridade vs. TMA vs. NPS

```
px.scatter(df_analise,
         x =
         "tempo_medio_atendimento_min"
         ,y = "nps_medio",
         title = "Scatterplot de tempo_medio_atendimento_min vs nps_medio
         vs cargo",template = "seaborn",
          height = 600, width =
         1000, color="cargo",
         symbol="cargo",trendline =
         "ols").show()
```

Analise 2: Canal vs. Tempo Médio de Resposta vs. NPS

```
px.scatter(df_analise,
         x =
         "tempo_medio_resposta_min",
         y = "nps_medio",
title = "Scatterplot de tempo medio resposta min vs nps medio
```

```
vs canal",template = "seaborn",
 height = 600, width =
1000, color="canal",
symbol="canal",trendline =
"ols").show()
```

Analise 3: Senioridade vs. Qtd de Atendimento/pessoa vs. NPS

```
px.scatter(df_analise,
         x =
         "qtd_atendimentos",y
         = "nps_medio",
         title = "Scatterplot de qtd_atendimentos vs nps_medio vs
         cargo",template = "seaborn",
         color="cargo",
          symbol="cargo",height =
          600, width = 1000,
         trendline = "ols").show()
```

```
px.scatter(df_analise,
         x =
         "qtd_atendimentos_pessoa",
         y = "nps_medio",
         title = "Scatterplot de qtd_atendimentos_pessoa vs nps_medio
         vs cargo",template = "seaborn",
         color="cargo",
          symbol="cargo",height =
          600, width = 1000,
         trendline = "ols").show()
```

Salvamento da base

```
salvar_base = False
if salvar_base == False:
  raise Exception("Sem
salvamento")else:
  df_analise.to_csv('./gdrive/MyDrive/projeto cx/df_analise_tratada.csv', index=False)
```

ANEXO 2 – *Notebook* 2: Código em Python da Preparação de Dados

Import bibliotecas

```
#mount do colab
from google.colab import
drive
drive.mount('/content/gdri
ve')

#manipulacao de
dadosimport pandas
as pd import numpy
as np
from datetime import datetime

#visualizacao de dados
import matplotlib.pyplot as
pltimport seaborn as sns
import plotly.express as px

#sklearn - preprocessing
from sklearn.preprocessing import
PowerTransformerfrom sklearn.preprocessing
import StandardScaler
from sklearn.preprocessing import QuantileTransformer

#sklearn - model_selection
from sklearn.model_selection import train_test_split
```

Import da base pre tratada

Import e ajuste das colunas da base de analise

```
#import da base de analise
df_analise = pd.read_csv("./gdrive/MyDrive/projeto_cx/df_analise_tratada.csv")

#visualização da
basedf_analise
```

Divisao da Base em Treino e Teste

- Utilizaremos uma separação padrão de base de 75% para treino e 25%
- para testeSepararemos também a base de features (X) da base de resposta (y)
- A semente 123 garante a reprodutibilidade dos resultados

```
#Separacao das bases em feature e
resposta            X             =
df_analise.drop("nps_medio", axis =
1)y = df_analise[["nps_medio"]]

#divisao da base de treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size = 0.25,
                                                    random_state = 123)

X_train

X_test

y_train

y_testMostrar saída oculta
```

Tratamento das Variaveis Numericas

Base de Treino

Padronização das colunas com escala log

```
#colunas com escala log
colunas_numericas_com_log = ["qtd_colaboradores_area","qtd_atendimentos_pessoa"]
colunas_numericas_com_log_novo =
["qtd_colaboradores_area_log","qtd_atendimentos_pessoa_log"]
```

```
#separação das colunas numericas com log
df_num_train_com_log =
X_train[colunas_numericas_com_log]

#instancia o transformador
power_transformer = PowerTransformer().fit(df_num_train_com_log)

#transformação dos valores originais
df_num_train_com_log_padronizado = power_transformer.transform(df_num_train_com_log)

#visualização
df_num_train_com_log_padronizado = pd.DataFrame(df_num_train_com_log_padronizado, columns =
colunas_numericas_com_log_novo, idf_num_train_com_log_padronizado
```

Transformação das colunas com escala quantilica

```
#colunas com escala log
colunas_numericas_qtl = ["horas_atendimento_canal", "tempo_medio_atendimento_min", "tempo_medio_resposta_min"]
colunas_numericas_qtl_novo = ["horas_atendimento_canal_qtl", "tempo_medio_atendimento_min_qtl",
"tempo_medio_resposta_min_qtl

#separação das colunas numericas
quantilicas df_num_train_qtl =
X_train[colunas_numericas_qtl]

#instancia o transformador
qtl_transformer = QuantileTransformer(output_distribution = "normal", random_state =
123).fit(df_num_train_qtl)

#transformação dos valores originais
df_num_train_qtl_padronizado = qtl_transformer.transform(df_num_train_qtl)

#visualização
df_num_train_qtl_padronizado = pd.DataFrame(df_num_train_qtl_padronizado, columns =
colunas_numericas_qtl_novo, index = X_tradf_num_train_qtl_padronizado
```

Base de Teste

Padronizacao das colunas com base nos objetos de treino

```
#power
df_num_test_com_log_padronizado = pd.DataFrame(power_transformer.transform(X_test[colunas_numericas_com_log]),
                                        columns =
                                        colunas_numericas_com_log_novo,
                                        index = X_test.index)
#quantilico
df_num_test_qtl_padronizado = pd.DataFrame(qtl_transformer.transform(X_test[colunas_numericas_qtl]),
                                        columns =
                                        colunas_numericas_qtl_novo,index
                                        = X_test.index)
```

Tratamento das Variaveis Categoricas

```
colunas_categoricas = ["cargo", "canal", "produto"]
```

Base de Treino

```
df_cat_train = pd.get_dummies(data = X_train[colunas_categoricas],
                              drop_first = True)
df_cat_train
```

Base de Teste

```
df_cat_test = pd.get_dummies(data = X_test[colunas_categoricas],
                             drop_first = True)
df_cat_test
```

Unificação das bases

```
X_train_treated = pd.concat([df_num_train_com_log_padronizado, df_num_train_qtl_padronizado,
df_cat_train], axis = 1)X_train_treated
```

```
X_test_treated = pd.concat([df_num_test_com_log_padronizado, df_num_test_qtl_padronizado,
df_cat_test], axis = 1)X_test_treated
```

Validação das Distribuições

Treino

```
for col in colunas_numericas_com_log_novo + colunas_numericas_qtl_novo:
  px.histogram(X_train_treated[col], title = col, template = "seaborn",width=800,
  height=500).show()
```

Teste

```
for col in colunas_numericas_com_log_novo + colunas_numericas_qtl_novo:
  px.histogram(X_test_treated[col], title = col, template = "seaborn",width=800,
  height=500).show()
```

Salvamento das bases

```
salvar_base = True
if salvar_base == False:
  raise Exception("Sem
salvamento")else:
  for (df, df_name) in zip([X_train_treated, X_test_treated, y_train, y_test],
                     ["X_train_treated", "X_test_treated", "y_train",
    "y_test"]):df.to_csv(f'./gdrive/MyDrive/projeto_cx/{df_name}.csv',
    index=False)
```

ANEXO 3 – *Notebook* 3: Código em Python da Modelagem

```
#mount do colab
from google.colab import
drive
drive.mount('/content/gdri
ve')

#manipulacao de
dadosimport pandas
as pd import numpy
as np
from datetime import datetime

#visualizacao de dados
import matplotlib.pyplot as
pltimport seaborn as sns
import plotly.express as px

#transforma em normal padrão
from sklearn.preprocessing import StandardScaler

#sklearn - modelos
from sklearn.model_selection import
GridSearchCV from sklearn.linear_model
import LinearRegressionfrom
sklearn.linear_model import Lasso
from sklearn.tree import
DecisionTreeRegressor from sklearn.ensemble
import RandomForestRegressorfrom sklearn
import tree

#sklearn - analise de erros
from sklearn.metrics import
mean_absolute_errorfrom sklearn.metrics
import mean_squared_error

#import biblioteca
!pip install
graphvizimport
graphviz
```

Import da base pre tratada

Import e ajuste das colunas da base de analise

```
#import da base de analise
X_train_treated =
pd.read_csv("./gdrive/MyDrive/projeto_cx/X_train_treated.csv")
X_test_treated =
pd.read_csv("./gdrive/MyDrive/projeto_cx/X_test_treated.csv") y_train
= pd.read_csv("./gdrive/MyDrive/projeto_cx/y_train.csv")
y_test = pd.read_csv("./gdrive/MyDrive/projeto_cx/y_test.csv")
```

Modelagem

```
def criar_modelo_e_previsao(modelo, param_grid,
                        X_train_treated,
                        X_test_treated,y_train,
                        y_test,
                        scoring = "neg_mean_squared_error"):
  # cria o modelo, para pegar os parametros posteriormente, e obtem o df de valores previstos de teste

  #cria o gridsearch
  grid_model = GridSearchCV(estimator = modelo,
                        param_grid =
                        param_grid,scoring =
                        scoring,
                        cv = 5)

#fit do grid com a base de treino grid model.fit(X train treated,
```

```
  np.ravel(y_train))

  #previsao do modelo na base de teste
  previsao = grid_model.predict(X_test_treated)

  #erro médio absoluto
  mae = mean_absolute_error(y_test, previsao)

  #erro medio quadratico
  mse = mean_squared_error(y_test, previsao)
  return grid_model, previsao, mae, mse
```

Modelo 1 - Regressão Linear

Definição dos hiperparametros

```
modelo =
LinearRegression()
param_grid = {}


reg_linear, previsao_reg_linear, mae_reg_linear, mse_reg_linear =
                        criar_modelo_e_previsao(modelo, param_grid,X_train_treated,
                        X_test_treated,
                        y_train, y_test)
```

Melhores hiperparametros (regressao linear nao tem)

```
reg_linear.best_params_
```

Modelo 2 - Lasso

Definição dos hiperparametros

```
modelo = Lasso(random_state = 123)
param_grid = {"max_iter":
[1000,3000,5000]}

lasso, previsao_lasso, mae_lasso, mse_lasso = criar_modelo_e_previsao(modelo,
                        param_grid,X_train_treated, X_test_treated,
                        y_train, y_test)
```

Melhores hiperparametros

```
lasso.best_params_
```

Modelo 3 - Árvore de Decisão

Definição dos hiperparametros

```
modelo = DecisionTreeRegressor(random_state
= 123)param_grid = {"max_depth": [4, 5, 6,
7],
         "min_samples_split": [0.04, 0.06],
         "min_samples_leaf": [0.01]}

arv_dec, previsao_arv_dec, mae_arv_dec, mse_arv_dec = criar_modelo_e_previsao(modelo,
                        param_grid,X_train_treated, X_test_treated,
                        y_train, y_test)
```

Melhores hiperparametros

```
arv_dec.best_params_
```

Modelo 4 - Random Forest

Definição dos hiperparametros

```
modelo = RandomForestRegressor(random_state
= 123)param_grid = {"n_estimators": [100,
200],

             "max_depth": [4, 5, 6, 7],
             "min_samples_split": [0.04, 0.06],
             "min_samples_leaf": [0.01]}

rand_for, previsao_rand_for, mae_rand_for, mse_rand_for =
                        criar_modelo_e_previsao(modelo, param_grid,X_train_treated,
                        X_test_treated,y_train, y_test)
```

Melhores hiperparametros

```
rand_for.best_params_
```

Resultados da previsão

Erro Medio Quadratico

```
df_resultados_mse = pd.DataFrame({"modelo": ["regressao_linear", "lasso", "arvore_decisao",
                          "random_forest"],"MSE": [mse_reg_linear, mse_lasso, mse_arv_dec,
                          mse_rand_for]})
df_resultados_mse
```

```
px.bar(df_resultados_mse, x = "modelo", y = "MSE", template = "seaborn",
       text_auto = True, title = f"MSE (Erro medio quadratico) dos modelos").show()
```

Erro Medio Absoluto

```
df_resultados_mae = pd.DataFrame({"modelo": ["regressao_linear", "lasso", "arvore_decisao",
                          "random_forest"],"MAE": [mae_reg_linear, mae_lasso, mae_arv_dec,
                          mae_rand_for]})
df_resultados_mae
```

```
px.bar(df_resultados_mae, x = "modelo", y = "MAE", template =
       "seaborn", text_auto = True, title = f"MAE (Erro medio absoluto)
       dos modelos").show()
```

Analise dos Parametros e Feature Importance

Modelos Lineares

```
df_parametros_modelos_lineares = pd.DataFrame(data = {"variável": X_train_treated.columns,"regressao_linear":
reg_linear.best_estimator_.coef_,"lasso": lasso.best_estimator_.coef_})
```

```
df_parametros_modelos_lineares = df_parametros_modelos_lineares.melt(id_vars = "variável",
                                                        value_vars =
                                                        ["regressao_linear","lasso"],
                                                        var_name = "modelo",
                                                        value_name = "parametro")
```

```
df_parametros_modelos_lineares
```

```
px.bar(df_parametros_modelos_lineares, x = "variável", y = "parametro", facet_row = "modelo",
       template = "seaborn",text_auto = True, height = 800, title = f"Parâmetros dos Modelos de
       Regressão Linear").show()
```

Modelos de Arvore

```
df_parametros_modelos_arvore = pd.DataFrame(data = {"variável": X_train_treated.columns,
                                           "arvore_decisao":
                                           arv_dec.best_estimator_.feature_importances_,
                                           "random_forest":
                                           rand_for.best_estimator_.feature_importances_})
```

```
df_parametros_modelos_arvore = df_parametros_modelos_arvore.melt(id_vars = "variável",
                                                        value_vars =
                                                        ["arvore_decisao","random_forest"],
                                                        var_name = "modelo",
                                                        value_name = "feature_importance")
```

```
df_parametros_modelos_arvore
```

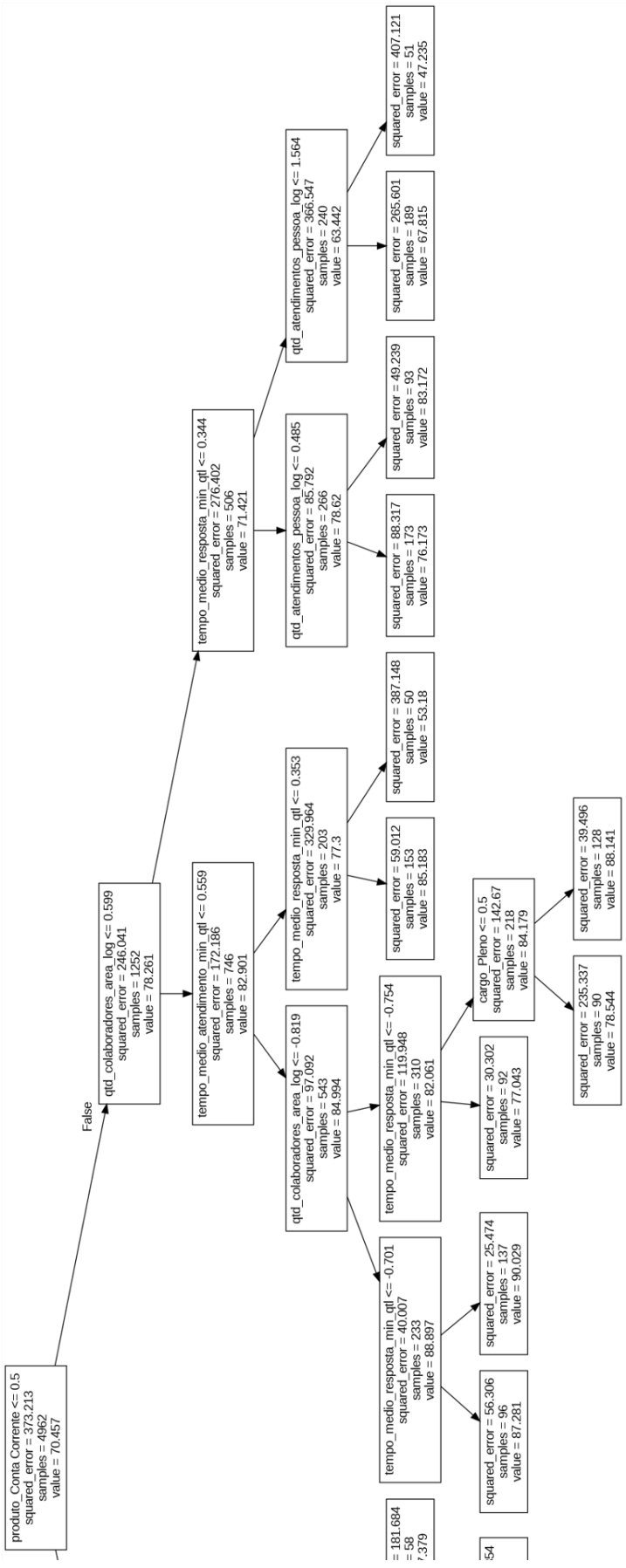
```
px.bar(df_parametros_modelos_arvore, x = "variável", y = "feature_importance", facet_row = "modelo",
       template = "seaborn",text_auto = True, height = 800, title = f"Parâmetros dos Modelos de
       Árvore").show()
```

Visualização da Árvore

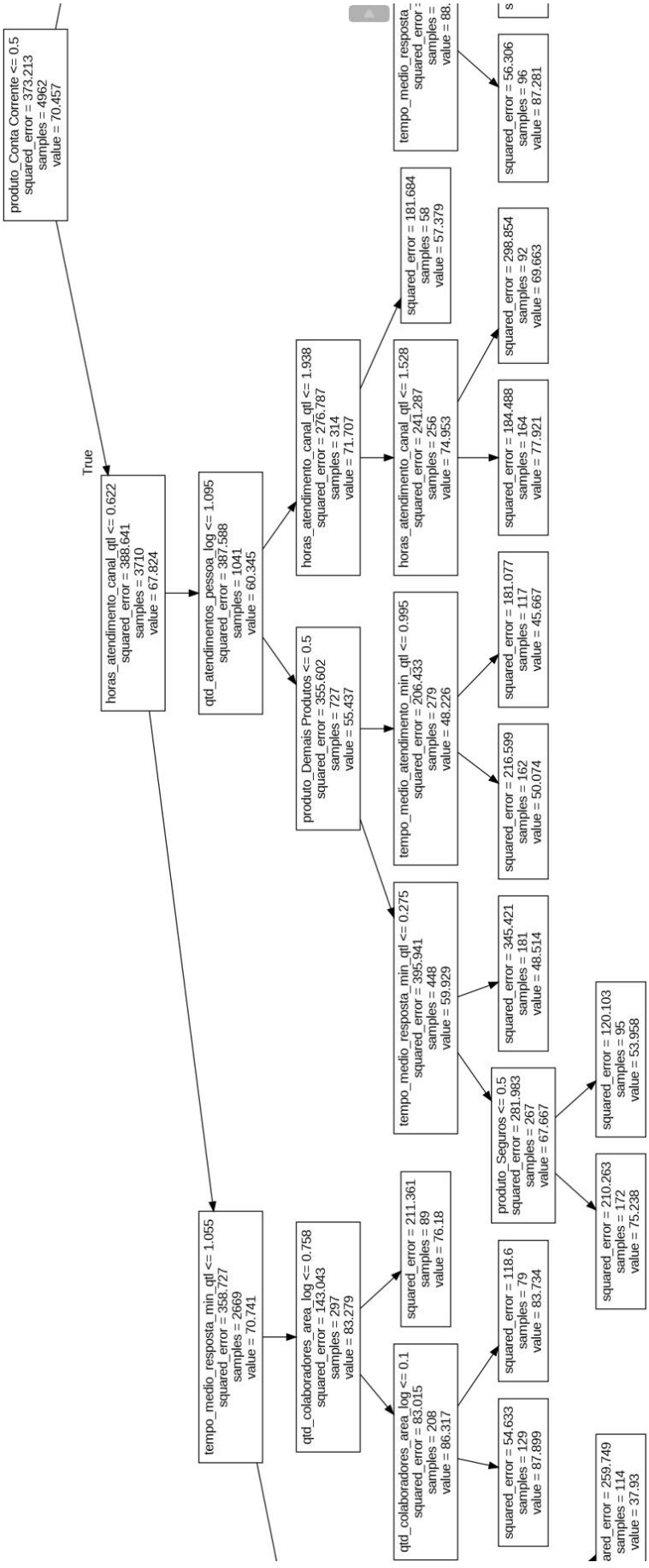
```
# Gerar o código da árvore em formato DOT
dot data = tree.export graphviz(arv dec.best estimator ,
```

```
                            feature_names=X_train_treated.col
                            umns)

 # Criar o gráfico da árvore usando o
 Graphvizgraph =
 graphviz.Source(dot_data)

 # Exibir o gráfico da árvore
 em PDF# graph.view()

 # Exibir o gráfico da árvore
 em PNGgraph.format = 'png'
graph.render(filename='arvore_regressao', directory='.', cleanup=True)
```

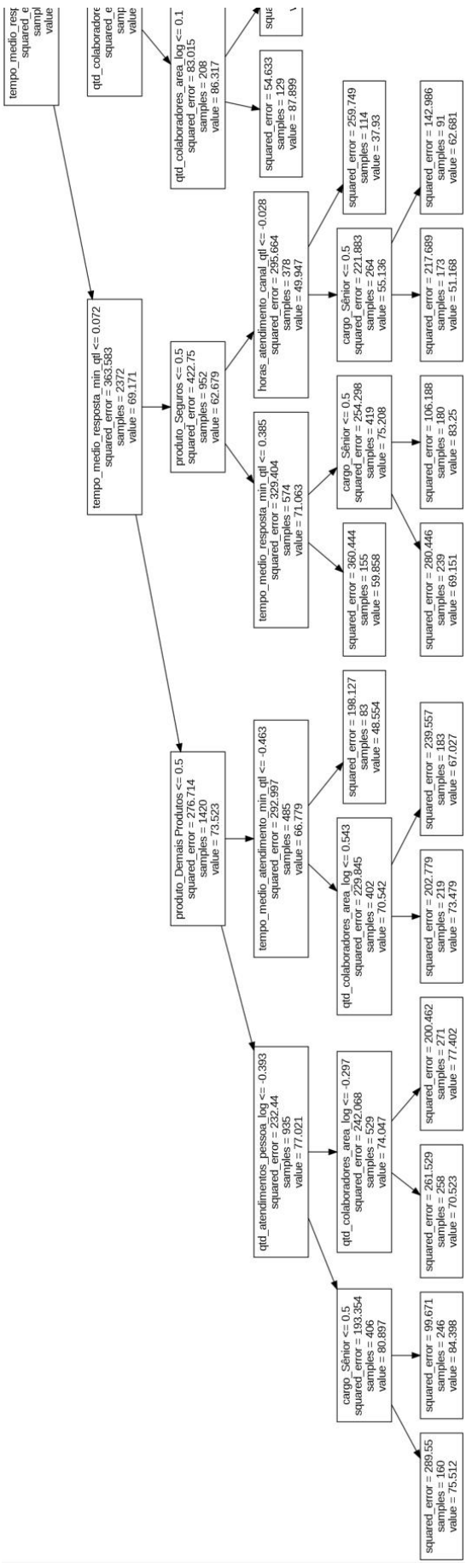
ANEXO 4 – Árvore de Decisão (Parte 1)



(Parte 2)


produto_Conta Corrente <= 0.5
squared_error = 373.213
samples = 4962
value = 70.457
True
horas_atendimento_canal_qtl <= 0.622
squared_error = 388.641
samples = 3710
value = 67.824
qtd_atendimentos_pessoa_log <= 1.095
squared_error = 387.588
samples = 1041
value = 60.345
horas_atendimento_canal_qtl <= 1.938
squared_error = 276.787
samples = 314
value = 71.707
squared_error = 181.684
samples = 58
value = 57.379
horas_atendimento_canal_qtl <= 1.528
squared_error = 241.287
samples = 256
value = 74.953
squared_error = 298.854
samples = 92
value = 69.663
squared_error = 184.488
samples = 164
value = 77.921
produto_Demais Produtos <= 0.5
squared_error = 355.602
samples = 727
value = 55.437
tempo_medio_atendimento_min_qtl <= 0.995
squared_error = 206.433
samples = 279
value = 48.226
squared_error = 181.077
samples = 117
value = 45.667
squared_error = 216.599
samples = 162
value = 50.074
tempo_medio_resposta_min_qtl <= 0.275
squared_error = 395.941
samples = 448
value = 59.929
squared_error = 345.421
samples = 181
value = 48.514
produto_Seguros <= 0.5
squared_error = 281.983
samples = 267
value = 67.667
squared_error = 120.103
samples = 95
value = 53.958
squared_error = 210.263
samples = 172
value = 75.238
tempo_medio_resposta_min_qtl <= 1.055
squared_error = 358.727
samples = 2669
value = 70.741
qtd_colaboradores_area_log <= 0.758
squared_error = 143.043
samples = 297
value = 83.279
squared_error = 211.361
samples = 89
value = 76.18
qtd_colaboradores_area_log <= 0.1
squared_error = 83.015
samples = 208
value = 86.317
squared_error = 118.6
samples = 79
value = 83.734
squared_error = 54.633
samples = 129
value = 87.899
ared_error = 259.749
samples = 114
value = 37.93
tempo_medio_resposta_
squared_error =
samples =
value = 88.
squared_error = 56.306
samples = 96
value = 87.281
s

(Parte 3)


tempo_medio_resposta_min_qtl <= 0.072
squared_error = 363.583
samples = 2372
value = 69.171
produto_Seguros <= 0.5
squared_error = 422.75
samples = 952
value = 62.679
qtd_colaboradores_area_log <= 0.1
squared_error = 83.015
samples = 208
value = 86.317
squared_error = 54.633
samples = 129
value = 87.899
horas_atendimento_canal_qtl <= -0.028
squared_error = 295.664
samples = 378
value = 49.947
tempo_medio_resposta_min_qtl <= 0.385
squared_error = 329.404
samples = 574
value = 71.063
squared_error = 360.444
samples = 155
value = 59.858
cargo_Sênior <= 0.5
squared_error = 254.298
samples = 419
value = 75.208
squared_error = 280.446
samples = 239
value = 69.151
cargo_Sênior <= 0.5
squared_error = 221.883
samples = 264
value = 55.136
squared_error = 259.749
samples = 114
value = 37.93
squared_error = 217.689
samples = 173
value = 51.168
squared_error = 106.188
samples = 180
value = 83.25
squared_error = 142.986
samples = 91
value = 62.681
produto_Demais Produtos <= 0.5
squared_error = 276.714
samples = 1420
value = 73.523
tempo_medio_atendimento_min_qtl <= -0.463
squared_error = 292.997
samples = 485
value = 66.779
squared_error = 198.127
samples = 83
value = 48.554
qtd_colaboradores_area_log <= 0.543
squared_error = 229.845
samples = 402
value = 70.542
squared_error = 239.557
samples = 183
value = 67.027
squared_error = 202.779
samples = 219
value = 73.479
qtd_atendimentos_pessoa_log <= -0.393
squared_error = 232.44
samples = 935
value = 77.021
qtd_colaboradores_area_log <= -0.297
squared_error = 242.068
samples = 529
value = 74.047
squared_error = 200.462
samples = 271
value = 77.402
squared_error = 261.529
samples = 258
value = 70.523
cargo_Sênior <= 0.5
squared_error = 193.354
samples = 406
value = 80.897
squared_error = 99.671
samples = 246
value = 84.398
squared_error = 289.55
samples = 160
value = 75.512