RELIMAT
Revista Eletrônica do Curso de Licenciatura em Matemática

Uni-FACEF
Centro Universitário Municipal de Franca

OTIMIZANDO A TOMADA DE DECISÕES NO VAREJO:

uma abordagem com árvore de decisão

Ana Laura Rodrigues Silva
Discente do Curso de Licenciatura em Matemática – Uni-FACEF
anarodriguesmat@gmail.com

Letícia Faleiros Chaves Rodrigues
Mestre em Matemática Universitária e Docente do Uni-FACEF
leticia@facef.br

RESUMO

Há algumas décadas, as preocupações das empresas do setor de varejo eram fundamentalmente ligadas à gestão de estoques, investimentos e resultados financeiros. No entanto, com o aumento exponencial de informações disponíveis através do comércio eletrônico e do uso de dispositivos móveis, as empresas estão reconhecendo a importância da análise de informações para o sucesso dos negócios. Consequentemente, o tratamento e administração de informações tornaram-se cada vez mais relevantes, e modelos estatísticos estão sendo cada vez mais empregados. Esses modelos oferecem dados para orientar as operações e a modelagem estatística, permitindo a avaliação de cenários e previsões que aprimoram o processo de tomada de escolhas. Devido à imensa quantidade de informações que são geradas a cada instante no setor de varejo, é imperativo desenvolver algoritmos capazes de analisar e processar essas informações de forma eficaz. O propósito principal deste estudo é examinar a viabilidade dos modelos de árvore de escolhas no setor de varejo. Isso implica destacar como a aplicação de modelos estatísticos, como a árvore de escolhas, pode aperfeiçoar a eficácia e incentivar decisões mais bem fundamentadas e orientadas por informações nas empresas deste ramo, considerando as particularidades e variáveis específicas de cada contexto. A árvore de escolhas identifica questões essenciais que necessitam de resposta para a tomada de uma escolha, representando essas questões como ramos na árvore. Cada ramo da árvore representa uma escolha ou alternativa possível, e as folhas representam as consequências ou resultados associados a essa escolha. Ao utilizar a árvore de escolhas, será viável auxiliar as empresas do varejo na tomada de decisões estratégicas, levando a resultados cada vez mais favoráveis e eficazes.

Palavras-chave: Árvore de decisão. Dados. E-commerce. Modelo estatístico.

RELIMAT
Revista Eletrônica do Curso de Licenciatura em Matemática
Uni-FACEF
Centro Universitário Municipal de Franca

ABSTRACT

A few decades ago, retail companies' concerns were essentially related to inventory management, investments, and financial returns. However, with the exponential increase in data available through e-commerce and the use of mobile devices, companies are realizing the importance of data analysis for business success. Consequently, data processing and management have become increasingly relevant, and statistical models are being more widely employed. These models provide inputs to guide operations and statistical modeling, enabling scenario analysis and predictions that enhance decision-making. Due to the vast volume of data generated every second in the retail sector, it is imperative to develop algorithms capable of efficiently analyzing and processing this data. The primary objective of this study is to investigate the feasibility of decision tree models in the retail sector. This involves highlighting how the application of statistical models, such as decision trees, can improve efficiency and promote more informed and data-driven decisions in companies within this industry, considering the specific characteristics and variables of each situation. The decision tree identifies key questions that need to be answered to make a choice, representing these questions as branches on the tree. Each branch of the tree represents a possible decision or option, and the leaves represent the consequences or outcomes associated with that decision. By using decision trees, it will be possible to assist retail companies in making strategic decisions, leading to increasingly positive and efficient results.

Palavras-chave: Decision tree. Data. E-commerce. Statistical model.

1 INTRODUÇÃO

A inovação é uma palavra-chave em muitos campos, especialmente na tecnologia e na ciência. A capacidade de criar e implementar novas ideias é fundamental para o desenvolvimento de soluções eficientes em diversas áreas. Nesse contexto, os modelos matemáticos são ferramentas importantes para a análise e a previsão de comportamentos em diferentes áreas, incluindo finanças, engenharia e ciências da saúde. Uma das aplicações desses modelos é a criação de árvores de decisão, que são formas de representar graficamente o processo de tomada de decisão que foram foco dos capítulos 2 e 3 com os respectivos temas: Modelo Matemático e Árvore de Decisão.

A tecnologia impulsiona a inovação, permitindo a criação de sofisticados softwares e ferramentas de análise de dados. Neste trabalho, foi utilizado o RStudio, um ambiente de desenvolvimento renomado para análise de dados, e o Power BI, uma poderosa ferramenta de visualização e Business Intelligence da Microsoft. O RStudio capacita análises estatísticas avançadas e criação de modelos preditivos, enquanto o Power BI transforma dados em relatórios interativos e painéis de controle, facilitando a comunicação das descobertas às partes interessadas.

Para avaliar a eficácia dos modelos desenvolvidos, é crucial que se utilizem métricas de avaliação apropriadas. No Capítulo 4, intitulado "Métricas de Avaliação", foram exploradas as métricas empregadas para medir o desempenho dos modelos. Essas métricas incluem a acurácia da matriz de confusão, precisão, recall, F1-score, curva Auroc, SMAPE entre outras. Essas ferramentas permitem a avaliação da qualidade dos resultados obtidos e o aprimoramento contínuo dos modelos.

No Capítulo 5, uma análise prática foi elaborada após a aplicação dos modelos a um contexto real, tendo sido analisada uma empresa de varejo. Nesse contexto, foram desenvolvidos modelos que exploraram os estudos realizados ao longo deste trabalho, utilizando dados reais. Foi explicitado como as métricas discutidas anteriormente, como a acurácia da matriz de confusão, precisão, recall, F1-score e outras, puderam ser aplicadas de maneira eficaz na resolução de desafios do mundo real. Essa abordagem possibilitou a avaliação não apenas da teoria por trás dos modelos, mas também de sua aplicação prática e de sua capacidade de impulsionar a tomada de decisões informadas no contexto de uma empresa de varejo.

Com o crescente aumento na produção de dados dentro desse setor, torna-se imprescindível contar com ferramentas e algoritmos capazes de lidar com esse grande volume de informações. Dessa forma, o presente estudo teve a intenção de contribuir para o aumento da eficiência e produtividade das empresas, impulsionando a inovação e embasando a tomada de decisões em dados sólidos.

2.MODELO MATEMÁTICO

O mundo está em constante evolução, o que vem acompanhado de diversas transformações que também afetam o comportamento do ser humano. Nesse cenário, é essencial que as pessoas desenvolvam a capacidade de se adaptar a esse crescimento e saibam lidar com as novas tecnologias. Essas adaptações trazem uma série de benefícios, como a melhora na qualidade de vida, acesso rápido ao conhecimento e facilitação da troca de informações.

A frente desse desenvolvimento, as evoluções de métodos tecnológicos provocam diferentes resultados em nossas vidas, os quais podem ser positivos, mas também negativos. De acordo com Oliveira et al. (2022), "um dos efeitos desfavoráveis que é muito comum na sociedade atual é o chamado 'technostress', que é desencadeado pelo excesso de conectividade e o bombardeio de informações em diferentes meios comunicativos".

Dentro das empresas, esse “boom tecnológico” ampliou significativamente o modo de produzir. Atualmente para se sobressair em um mercado competitivo, as empresas se encontram obrigadas a realizar mais em menos tempo, é preciso manter a redução de custos, buscando maior produção, tendo a preocupação com a segurança e qualidade, levando sempre em consideração uma margem de erro a respeito da decisão, além de elevados investimentos nesse setor. A automatização de tarefas é uma consequência da implantação da tecnologia e, também, uma ajuda à economia das empresas.

Dado aos muitos desafios enfrentados pelo setor de empreendimento, uma estratégia útil seria a implantação de um sistema de gestão financeira eficiente.

Tal sistema permitiria a identificação de gastos excessivos com matéria-prima e o monitoramento de desperdícios. Além disso, identificaria áreas em que é possível aumentar os lucros e identificar clientes potenciais.

A operabilidade desses projetos pode ser realizada por meio de modelos matemáticos, principalmente pelo elevado nível de desenvolvimento e poder computacional dos nossos dias.

Logo, quando decidimos por explorar modelo matemático, surgem algumas questões: o que preciso saber? O que me preocupa? Minhas ações teriam consequências? E se tivessem, quais seriam? Qual a importância de fazer uma escolha certa? Suas escolhas teriam resultados imediatos ou a longo prazo? Por qual motivo você está fazendo o que faz? Com base em cada ação, pense em como isso afetaria você e os demais.

A tomada de decisões consiste em fazer uma escolha importante com o objetivo de alcançar melhores resultados, porém, escolher algo não é tão simples assim. Segundo Pereira et al. (2021), "o processo de decisão é contínuo e deve ser trabalhado, analisando seus riscos e soluções". Assim, é importante desenvolver métodos com o objetivo de analisar os comportamentos de sistemas complexos em situações que geralmente são difíceis de serem observadas.

Um modelo matemático é uma representação matemática simplificada de um sistema ou fenômeno real. É uma ferramenta valiosa para entender, descrever e prever o comportamento de sistemas complexos, incluindo física, biologia, economia, engenharia e muitas outras áreas.

O tratamento de dados é um processo crítico na construção de modelos matemáticos precisos e confiáveis. É importante realizar uma análise cuidadosa e aplicar técnicas de limpeza, transformação e manipulação de dados para prepará-los para a modelagem. Além disso, é importante selecionar o modelo matemático adequado e ajustar os parâmetros do modelo para obter previsões precisas e confiáveis (Johnson, S., & Smith, R., 2022).

Alguns exemplos de técnicas comuns de tratamento de dados são:

- **Limpeza de dados:** envolve a identificação e correção de dados inconsistentes, incompletos ou incorretos. Por exemplo, se houver dados faltantes em um conjunto, é possível preenchê-los com a média ou a mediana dos dados existentes. Se houver duplicados, é possível remover as duplicatas.

- **Transformação de dados:** envolve a conversão de dados de um formato para outro, para facilitar a análise. Por exemplo, se houver uma coluna em formato de texto, é possível convertê-la em um formato numérico para facilitar a análise estatística.
- **Normalização de dados:** envolve a escala dos dados para uma faixa específica, como uma escala de 0 a 1. Isso é útil para comparar conjuntos que têm escalas diferentes. Por exemplo, se houver duas colunas, uma com valores entre 0 e 100 e outra com valores entre 0 e 1000, a normalização permitirá que sejam comparadas diretamente.
- **Padronização de dados:** envolve a transformação dos dados para que tenham uma média de zero e um desvio padrão de um. Isso é útil para reduzir o impacto de valores extremos na análise estatística.
- **Agregação de dados:** envolve a combinação de dados em grupos ou categorias com base em critérios específicos. Por exemplo, se houver dados de vendas por dia, é possível agregá-los por semana ou mês para obter uma ampla visão das tendências de vendas ao longo do tempo.
- **Filtragem de dados:** envolve a exclusão de dados que não são relevantes para a análise. Por exemplo, se houver dados de vendas em todo o país, é possível filtrar para incluir apenas as vendas em uma região específica.

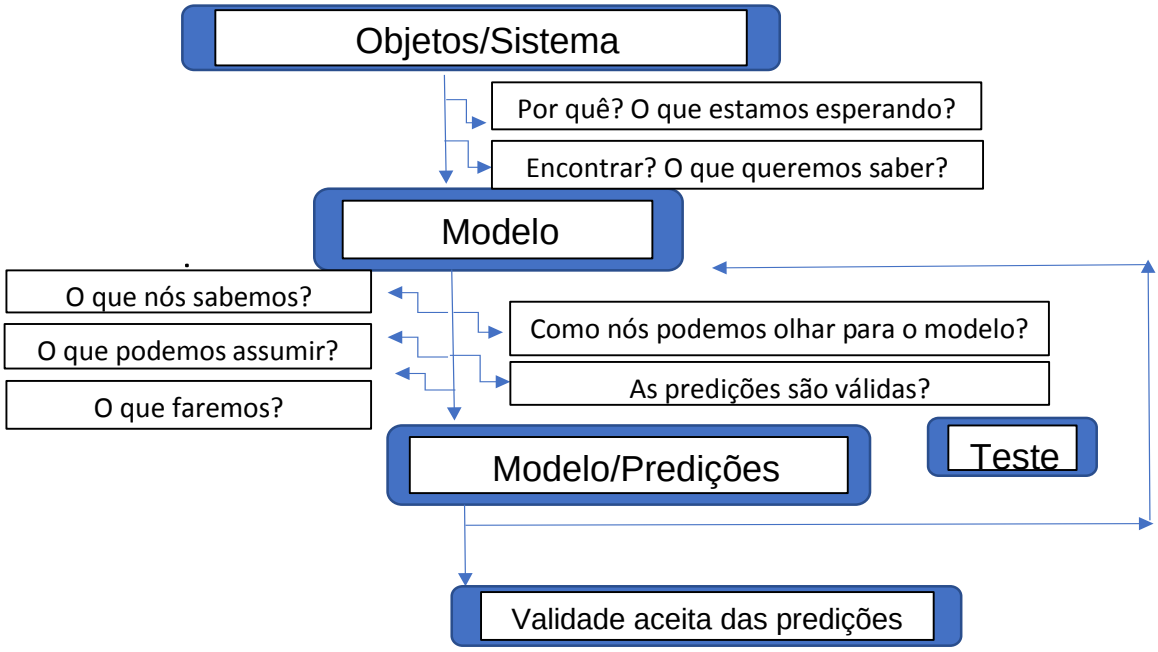
Essas são apenas algumas das técnicas comuns de tratamento de dados. A escolha das técnicas dependerá do conjunto de dados específico e dos objetivos da análise.

Depois de limpar e preparar, o próximo passo é selecionar o modelo matemático adequado para o problema em questão. Existem vários tipos de modelos matemáticos, como lineares, de regressão, de séries temporais, entre outros.

Uma vez selecionado o modelo matemático, é necessário ajustar os parâmetros aos dados. Isso pode envolver a aplicação de técnicas de otimização para encontrar os valores ideais dos parâmetros que melhor se ajustam aos dados. Além disso, é importante autenticar o modelo usando técnicas de validação cruzada ou outras técnicas para garantir que seja capaz de fazer previsões precisas em novos dados.

Desta forma, com o objetivo de tornar mais prático e ilustrativo o processo de desenvolvimento de um modelo matemático e facilitar a visualização das decisões tomadas, foi elaborado o esquema a seguir.

Figura 1- Esquema de desenvolvimento de modelo



Fonte: Autoria Própria.

Agora, a jornada terá continuidade, com foco nas características e propriedades dos modelos matemáticos. Características referem-se a elementos e atributos comuns encontrados em diferentes modelos, como análise de variáveis, parâmetros, restrições e relações entre variáveis. Propriedades são qualidades específicas que variam entre modelos, como simplicidade, objetividade, sensibilidade, estabilidade e universalidade, fornecendo informações detalhadas sobre o comportamento e a avaliação do modelo.

2.1 CARACTERÍSTICAS E ELEMENTOS DE UM MODELO MATEMÁTICO

Modelos matemáticos são utilizados para analisar a relação entre duas ou mais variáveis e podem ser reais ou abstratos. Segundo Santos e Oliveira (2021), "apesar de parecer um conceito teórico, na realidade existem muitos aspectos da vida cotidiana regidos por eles". Isso fica evidente ao considerar a aplicação daquele construído para pesquisa, por exemplo, tal como nos casos em

que se usam números para analisar os efeitos de uma dada doença que acomete uma parcela considerável da população, como ocorrido em 2020 com o COVID-19.

No que diz respeito à função da origem da informação utilizada, um modelo matemático pode classificar-se em heurístico, que é analisado com base em possíveis explicações sobre as causas dos fenômenos observados, e o empírico que busca usar informações de experimentação real.

De acordo com cada tipo de representação, o modelo é composto por uma categoria, sendo ela quantitativa, onde os resultados obtidos têm um valor específico com um determinado significado (pode ser exato ou relativo), e qualitativa que se refere a uma análise da qualidade ou tendência de um fenômeno sem calcular um valor exato.

Outros vieses que devem ser estudados incluem a aleatoriedade, que é dividida em dois tipos: a determinística, que envolve valores constantes já conhecidos, e a estocástica, que abrange valores ainda não conhecidos. Além disso, o objetivo do modelo pode variar, sendo voltado para uma simulação com a intenção de simular e descrever um fenômeno, ou para uma otimização, usado para encontrar resultados adequados à pesquisa, ou um controle, que visa a organização de um resultado.

Devemos sempre destacar que os modelos matemáticos são aproximações, suposições da realidade e não refletem resultados exatos. Conforme apontado por Johnson e Silva (2019), "é utilizar dados para entender padrões e buscar evidências melhores para embasar decisões". Nesse sentido, é melhor ter uma ideia de, por exemplo, 50% do comportamento de determinado fenômeno do que não fazer ideia nenhuma.

Dependendo do objetivo pretendido, os modelos matemáticos podem ser aplicados para prever o valor das variáveis no futuro, fazer hipóteses, avaliar os efeitos de uma determinada política ou atividade, entre outros objetivos, já citados anteriormente.

Seus modelos podem variar diante de sua complexidade, porém todos são constituídos diante de um mesmo conjunto de características básicas:

- **Variáveis**: buscam compreender ou analisar. Principalmente no que diz respeito à sua relação com outras variáveis. Assim, por exemplo, uma variável pode ser o salário dos trabalhadores e o que queremos analisar são os seus

principais determinantes (anos de estudo, escolaridade dos pais, naturalidade etc.).

- **Parâmetros:** esses são valores conhecidos ou controláveis do modelo.
- **Restrições**: são certos limites que indicam que os resultados da análise são razoáveis. Por exemplo, se uma das variáveis for o número de filhos de uma família, uma restrição natural é que esse valor não possa ser negativo.
- **Relações entre variáveis:** o modelo estabelece uma certa relação entre as variáveis com base em teorias econômicas, físicas, químicas etc.
- **Representações simplificadas**: uma das características essenciais de um modelo matemático é a representação das relações entre as variáveis estudadas por meio de elementos da matemática como: funções, equações, fórmulas etc.

2.2 PRINCIPAIS PROPRIEDADES DE UM MODELO MATEMÁTICO

Quando um modelo matemático é desenvolvido, é concebido pela intenção de alcançar a eficácia, com esse objetivo, tem propriedades que ajudam a garantir esse propósito. São elas:

- **Simplicidade:** é uma das características desejáveis de um modelo matemático, pois permite que seja mais fácil de entender, analisar e aplicar. Um modelo matemático simples pode ser mais útil do que um mais complexo, especialmente se o objetivo fornece uma descrição geral do fenômeno em estudo ou uma previsão de comportamento futuro. A simplicidade em um modelo matemático pode ser alcançada de diversas formas, como por exemplo, através da redução do número de variáveis, ou a redução de detalhes desnecessários.
- **Objetividade:** é uma das principais vantagens. Um modelo matemático é objetivo porque é baseado em princípios matemáticos bem definidos e em dados observáveis. Isso significa que as previsões e conclusões derivadas do modelo são baseadas em fatos, em vez de serem influenciadas por opiniões pessoais ou preconceitos. A objetividade pode ser alcançada de diversas formas, como por exemplo aplicação de princípios matemáticos bem definidos, ou definição clara das variáveis.

- **Sensibilidade:** é uma característica importante, pois permite que se avalie como as mudanças em diferentes variáveis e parâmetros afetam as previsões do modelo. Um modelo matemático sensível é capaz de mostrar como pequenas mudanças nos dados de entrada ou nos parâmetros afetam as previsões de saída. A sensibilidade pode ser alcançada de diversas formas, como por exemplo o uso de diferentes cenários ou o teste de validação.
- **Estabilidade**: é uma característica importante, pois garante que o modelo produza resultados consistentes e confiáveis ao longo do tempo. Um modelo matemático estável é aquele em que pequenas mudanças nas entradas ou parâmetros produzem apenas pequenas mudanças nas saídas. A estabilidade pode ser alcançada de diversas formas, como por exemplo o controle de erros.
- **Universalidade**: é uma característica desejável, pois permite que ele seja aplicado a uma ampla variedade de situações e problemas. Um modelo matemático universal é aquele que é capaz de descrever uma ampla gama de fenômenos ou processos, independentemente do domínio ou campo de aplicação. A universalidade pode ser alcançada de diversas formas, como por exemplo a flexibilidade, pois permite que ele seja adaptado ou ajustado para diferentes situações ou problemas.

2.3 TIPOS DE MODELOS MATEMÁTICOS

À medida que o mundo se torna mais orientado por dados, a narrativa por meio da análise está se tornando um componente vital. A análise de dados é e deve ser um aspecto crítico de todas as organizações, para ajudar a determinar o impacto sobre os negócios, incluindo a avaliação de opiniões de clientes, a realização de pesquisa de mercado e a identificação de tendências ou outros insights de dados. Embora o processo de análise se concentre nas tarefas de limpeza, modelagem e visualização, o conceito geral de análise de dados e a importância dele para os negócios não devem ser menosprezados. Para analisar dados, os componentes principais são divididos nas categorias a seguir: Modelos descritivos, Modelos prescritivos, Modelos preditivos.

2.3.1 MODELOS DESCRITIVOS

Os modelos descritivos são utilizados para representar sistemas reais ou propostos e experimentar diferentes cenários e políticas de ação (Johnson & Santos, 2022). Eles oferecem flexibilidade na representação de modelos complexos e facilidade de aplicação para prever o comportamento do sistema em um horizonte de planejamento escolhido. Ao realizar experimentos com diferentes políticas de ação, os resultados obtidos podem fornecer uma visão futura do sistema, auxiliando no processo de tomada de decisões.Diferentemente dos métodos prescritivos, os modelos descritivos não buscam indicar a melhor solução possível, mas sim comparar políticas de ação já escolhidas.

Um modelo descritivo pode ser utilizado em diferentes contextos para descrever e analisar fenômenos existentes. Por exemplo:

- **Análise de Séries Temporais:** na área de vendas, um modelo descritivo pode ser usado para analisar a demanda por um produto ou serviço ao longo do tempo, identificando tendências e sazonalidades. Assim, uma empresa pode analisar as vendas mensais de seus produtos para entender como elas variam ao longo do ano.
- **Análise de Dados de Pesquisa:** na área de pesquisa de mercado, um modelo descritivo pode ser usado para analisar os resultados de uma pesquisa de opinião e descrever as características demográficas e comportamentais dos respondentes. Por exemplo, uma pesquisa sobre hábitos de consumo pode descrever a idade, gênero, renda e localização geográfica dos participantes.
- **Análise de Dados de Tráfego:** um modelo descritivo pode ser usado para analisar os padrões de tráfego em uma rodovia ou em um cruzamento urbano, identificando os horários de pico e a velocidade média dos veículos. Essas informações podem seru sadas para melhorar a segurança e a eficiência do tráfego.

- **Análise de Dados de Saúde Pública:** pode ser usado para descrever a prevalência de uma doença em uma população, identificando fatores de risco e características demográficas associadas. Isso pode ajudar as autoridades de saúde pública a implementar medidas preventivas e de controle da doença.

2.3.2 MODELOS PRESCRITIVOS

Os modelos prescritivos visam encontrar soluções otimizadas para um processo através da representação dos objetivos e restrições envolvidas (Smith & Silva, 2022). Eles podem ser resolvidos de maneira exata ou aproximada, sendo que na resolução exata é encontrada a melhor solução possível para o problema. No entanto, a resolução exata pode ser computacionalmente cara para alguns tipos de problemas, enquanto a resolução aproximada busca uma solução de boa qualidade com menor custo computacional. É preciso analisar caso a caso e decidir qual é o método mais adequado para cada problema.

Um modelo prescritivo é um tipo de modelo que usa dados e análises para recomendar ações específicas a serem tomadas em uma determinada situação. Aqui estão alguns exemplos práticos de modelo prescritivo:

- **Modelos de Otimização de Estoque:** esses modelos usam dados históricos de vendas, informações de fornecedores e prazos de entrega para recomendar níveis ideais de estoque que minimizem o estoque excedente e as faltas de estoque.
- **Sistemas de Recomendação:** esses modelos usam dados de compras anteriores e outras informações do cliente para recomendar produtos específicos que possam ser do interesse do cliente. Por exemplo, uma loja online pode usar um modelo prescritivo para recomendar produtos com base no histórico de compras do cliente.
- **Modelos de Roteamento de Veículos:** esses modelos usam informações sobre a localização dos clientes, tempo de entrega, disponibilidade de motoristas e outros fatores para otimizar as rotas de entrega. Isso pode levar a reduções

significativas nos custos de transporte e melhorias na qualidade do serviço ao cliente.

- **Modelos de Previsão de Demanda:** esses modelos usam dados históricos de vendas, informações do mercado e outras variáveis para prever a demanda futura de produtos ou serviços. Isso pode ajudar empresas a planejar melhor a produção e os recursos necessários para atender à demanda esperada.

- **Modelos de Precificação Dinâmica:** esses modelos usam informações sobre a oferta e demanda em tempo real para ajustar os preços de produtos ou serviços para maximizar a receita. Por exemplo, uma empresa de transporte aéreo pode usar um modelo prescritivo para ajustar os preços dos bilhetes com base na demanda atual e na disponibilidade de assentos.

2.3.3 MODELOS PREDITIVOS

Um modelo preditivo utiliza dados históricos e outras informações relevantes para prever um resultado futuro, analisando as relações entre variáveis e identificando padrões e tendências (Johnson et al., 2020). Há muitos exemplos práticos de modelos preditivos em diversas áreas, incluindo:

- **Previsão de Demanda:** muitas empresas usam modelos preditivos para prever a demanda futura de seus produtos. Esses modelos podem usar dados históricos de vendas, fatores sazonais, tendências do mercado, preços e promoções, entre outros, para prever a demanda futura e ajudar as empresas a planejarem sua produção e estoque.
- **Previsão de Risco de Crédito:** as empresas financeiras usam modelos preditivos para avaliar o risco de crédito de seus clientes. Esses modelos usam dados financeiros, histórico de crédito e outros fatores para prever a probabilidade de inadimplência do cliente e, assim, ajudam a reduzir os riscos de empréstimos inadimplentes.
- **Previsão de Evolução de Doenças:** modelos preditivos podem ser usados para prever a evolução de doenças e epidemias. Eles usam dados epidemiológicos, histórico de doenças e fatores ambientais para prever o curso da doença e ajudar as autoridades de saúde a tomar medidas preventivas.

- **Previsão de Cancelamento de Clientes:** empresas de telecomunicações e serviços de assinatura usam modelos preditivos para prever quando os clientes estão em risco de cancelar seus serviços. Esses modelos usam dados comportamentais dos clientes, como frequência de uso e reclamações, para prever quando um cliente está prestes a cancelar e tomar medidas preventivas.
- **Previsão de Preços de Imóveis:** modelos preditivos podem ser usados para prever os preços futuros de imóveis. Eles usam dados históricos de preços, localização, tamanho, idade e outros fatores para prever os preços futuros e ajudar os compradores e vendedores a tomar decisões informadas.

Os modelos preditivos têm se mostrado instrumentos essenciais para tomar decisões mais informadas e estratégicas em diversos setores. Segundo Smith e Jones (2022), esses modelos têm a capacidade de prever tendências de mercado, riscos de crédito e demanda de produtos, entre outros aspectos. Além disso, os modelos preditivos são moldados de acordo com as necessidades específicas de quem os cria e alimenta com dados, permitindo que sejam adaptados para transações em tempo real. Isso possibilita, por exemplo, a identificação ágil de clientes com dificuldades no pagamento e a análise dos produtos que foram vendidos em maiores quantidades. Com o avanço da tecnologia disponível atualmente, é possível prever tendências de maneira muito mais ágil e eficiente, auxiliando na tomada de decisões estratégicas para o sucesso de negócios e instituições em um ambiente cada vez mais dinâmico.

Com base nos tipos e características dos modelos matemáticos apresentados, nos próximos capítulos será realizada a exploração do modelo de árvore de decisão, que é o foco do presente trabalho. Os modelos baseados em árvore de decisão são amplamente utilizados em áreas como marketing, finanças e ciência de dados. Além especialmente úteis quando há muitas variáveis disponíveis e é necessário identificar a importância relativa de cada uma para fazer uma previsão precisa. O objetivo deste trabalho é apresentar um modelo matemático capaz de analisar e otimizar os dados do setor de varejo, tendo como

base uma empresa, mostrando que é possível melhorar a eficiência e a tomada de decisão das empresas deste ramo. Para isso, será realizada uma análise dos dados e a escolha das variáveis mais relevantes para a construção do modelo. Uma vez construído o modelo, serão avaliadas sua eficácia e eficiência em prever os resultados desejados. Será avaliado se o modelo é adequado para o problema em questão e se é robusto o suficiente para lidar com dados futuros e desconhecidos. Por fim, serão discutidos os resultados obtidos e as limitações do modelo, bem como a possibilidade de explorar outras abordagens, como o uso de diferentes técnicas de modelagem, para futuros trabalhos.

3. ÁRVORE DE DECISÃO

Um desejo muito comum é o de prever o futuro. Mas será possível? Através de uma busca rápida via internet, percebemos que existem muitas ideias a respeito, alguns acreditam que sim é possível, outros já acham algo impossível. Diante dessas opiniões, surgiram métodos passíveis de desenvolvimento, que podem influenciar em algumas mudanças de ações futuras desenvolvendo suas percepções extrassensoriais e aplicando-as na sua vida.

> Colin Wilson conta a história de um homem morando na Índia que andava por um caminho sujo para ir ao rio todos os dias para nadar. Um dia, no caminho de volta para casa, ele percebeu algumas pegadas, e que em certo ponto do caminho, ele tinha "aleatoriamente" trocado de lado e andado do outro lado do caminho. Ele não conseguiu parar de pensar no porquê tinha feito isso. Por que nesse momento ele tinha trocado de lado no caminho? Quando parou para examinar o caminho, ele olhou por entre as árvores e percebeu rastros de um tigre enorme, bem do lado do caminho, bem na hora que tinha trocado de lado. Inconscientemente, talvez, o homem tinha consciência do perigo, e se afastou do tigre, provavelmente salvando a própria vida. (WILSON. Colin, 2004)

Afinal, a “decisão” do homem de trocar de lado, influenciou ou não no seu futuro? Não existem escolhas erradas! Tudo é apenas uma questão de sob qual ponto de vista se está olhando para a situação. A vida está cheia de opções e o rumo que ela terá é definido pelas decisões que tomamos. Realizamos escolhas desde o momento em que acordamos. Decisões das mais simples até aquelas capazes de mudar o nosso destino, haverá momentos em que iremos errar e em outros acertar. É possível que para alguns a dificuldade em decidir esteja envolvida com a falha na elaboração de valores pessoais e na identificação de suas referências, a indecisão pode atrapalhar sua rotina durante dias, antes de qualquer coisa, primeiramente é preciso se preparar para todo o trabalho de decidir. Prever ações não é mais uma exclusividade de ficções científicas.

A inteligência artificial envolve a criação de sistemas de computador que podem realizar tarefas inteligentes usando algoritmos e dados, como aprendizado de máquina e reconhecimento de padrões, com aplicações em diversas áreas. A aprendizagem de máquinas é uma subárea da inteligência artificial que se concentra no desenvolvimento de algoritmos e técnicas que permitem que um sistema computacional aprenda a partir de dados. O objetivo é criar sistemas que possam detectar padrões, fazer previsões e tomar decisões com base nesses padrões, sem que precisem ser programados explicitamente para fazer isso. Assim, a árvore de decisão é uma das muitas técnicas de aprendizagem de máquina.

Ao utilizar a árvore de decisão como método de aprendizagem de máquina, o objetivo é construir um sistema capaz de fazer previsões com base em dados históricos. O sistema é desenvolvido através do treinamento da árvore de decisão com um conjunto de dados de treinamento e, posteriormente, é testado quanto à sua precisão utilizando um conjunto de dados de teste.

Durante o processo de treinamento, a árvore de decisão é ajustada para otimizar sua precisão na previsão de resultados. Isso é feito através da seleção cuidadosa das variáveis e características a serem incluídas no modelo, bem como da definição das regras de decisão para cada nó da árvore.

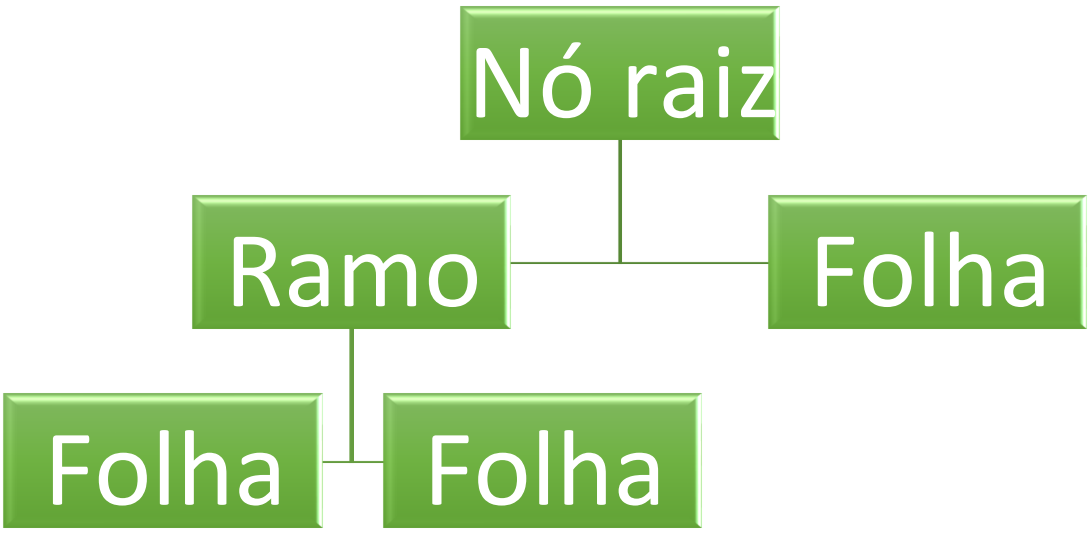
As árvores de decisão, são essenciais ao tomar decisões complexas, além de auxiliarem a compreensão das incertezas, isso não quer dizer que o propósito do diagrama seja conceber soluções, em geral, uma árvore de decisão é composta por perguntas e respostas, classificando um indivíduo ou entidade de acordo com o conjunto de respostas obtidas pelo conjunto de perguntas formuladas.

"A árvore de decisão é uma técnica de aprendizado de máquina que é utilizada para tomar decisões em problemas que envolvem várias alternativas, um mapa dos possíveis resultados de uma série de escolhas relacionadas." (Russell, S., & Norvig, P., 2016). Permite que um indivíduo ou organização compare possíveis ações com base em seus custos, probabilidades e benefícios. É construída com base em um conjunto de regras e dados de treinamento, e é usada para prever o resultado para novos dados. Por exemplo, uma árvore de decisão pode ser usada para prever se um cliente vai comprar um produto ou não, com base em suas características, como idade, renda e localização.

Além de ser amplamente utilizada em várias aplicações, como análise de dados, inteligência artificial, mineração de dados, marketing e finanças. É uma técnica de fácil compreensão e implementação, oferecendo boa precisão em muitos problemas. Ela funciona criando uma estrutura em forma de árvore, onde cada nó representa uma decisão ou uma condição, e as folhas representam as conclusões ou as ações a serem tomadas.

Basicamente, a árvore de decisão é iniciada por uma ideia, chamada de nó inicial (também conhecido como nó de raiz); após listar as opções, entra-se em uma segunda etapa, que são os nós internos, nos quais são utilizados conectores para listar essas opções e conectá-las com ramificações no nó raiz. Após essa etapa, deve-se colocar em prática, aplicar os testes nas opções, fase que também pode ser denominada como “ramos”. Por fim, chega-se na última etapa, nomeada de “folhas”, que é considerada a extremidade de cada fluxo, chegando-se em conclusões.

Figura 2- Árvore de Decisão



Fonte: Autoria Própria.

Assim, as árvores são subdivididas em dois grupos, um que classifica um registro (para problemas de classificação) e outro que estima um valor (para problemas de regressão).

3.1 ÁRVORE DE CLASSIFICAÇÃO

A árvore de classificação surge a partir de um processo de fazer uma série organizada de consultas, e as indagações feitas em cada etapa no processo são influenciadas pelas respostas fornecidas às consultas anteriores. O ponto de partida de uma árvore de classificação é chamado de nó raiz, que consiste em todo o conjunto de aprendizado, e está no topo da árvore. Um nó é um subconjunto do conjunto de atributos, e pode ser terminal ou não-terminal. Um nó não-terminal é um nó que se divide em nós filhos.

A árvore de classificação é uma técnica eficiente e de fácil interpretação, é amplamente utilizada em análise de dados e inteligência artificial, tornando-a uma ferramenta útil para muitos tipos de problemas de classificação, como a previsão de risco de crédito, a identificação de fraudes financeiras e a classificação de clientes em segmentos de mercado.

Ela se baseia em uma estrutura de árvore para representar relações hierárquicas entre as características dos dados e a classe a ser prevista.

Suponha que se tenha um conjunto de dados que contém informações sobre a decisão de uma pescaria. O objetivo é criar um modelo para prever a probabilidade de um dia viável a pesca. Pode-se utilizar uma árvore de classificação para construir esse modelo.

A árvore de classificação divide o conjunto de dados em diferentes subconjuntos com base em diferentes critérios, como a temperatura da água, se o dia está ensolarado, com nuvens ou chuvoso, se o vento está forte ou fraco, para criar ramos que levam a diferentes resultados. No final de cada ramo, temos um nó terminal que corresponde a uma decisão sobre a possibilidade de o pescador pescar ou não. Adiante, retornaremos a esse exemplo conforme mostrado na figura 8, onde poderemos observar como as informações nos nós anteriores influenciam nas decisões tomadas pelo pescador.

3.2 ÁRVORE DE REGRESSÃO

A análise de regressão é uma técnica estatística amplamente utilizada em diversas áreas, como a econometria, a psicologia e a medicina. Ela permite modelar a relação entre uma variável dependente e uma ou mais variáveis independentes, possibilitando a identificação de quais fatores são mais importantes e como eles interagem entre si.

Existem vários tipos de regressão, sendo os principais a linear simples e a linear múltipla (Montgomery et al., 2012). Na regressão linear simples, a relação entre a variável dependente e uma única variável independente é modelada por meio de uma equação linear. Já na linear múltipla, a relação entre a variável dependente e duas ou mais variáveis independentes é modelada por meio de uma equação linear múltipla (Montgomery et al., 2012).

Para ajustar um modelo de regressão, é necessário coletar dados e estimar os parâmetros da equação, geralmente por meio do método dos mínimos quadrados. Em seguida, é possível avaliar a qualidade do modelo por meio de diversas métricas, como o coeficiente de determinação (R^2) e o erro padrão da estimativa.

A regressão também pode ser utilizada para fins de previsão, ou seja, para estimar o valor da variável dependente com base nos valores das variáveis independentes. Para isso, basta substituir os valores das variáveis independentes na equação do modelo ajustado.

A equação geral para a regressão linear simples é:

$$y = a + bx$$

Onde y é a variável dependente, x é a variável independente, b é o coeficiente angular (também conhecido como declive ou inclinação), a é o coeficiente linear (também conhecido como interceptação ou constante) e a soma dos resíduos é igual a zero.

O valor de b pode ser calculado usando a seguinte fórmula:

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

Onde (é o número de pares de dados, $\sum x\,y$ é a soma dos produtos cruzados de x e y , $\sum x$ e $\sum y$ são as somas de x e y ,respectivamente, e $\sum x^2$ é a soma dos quadrados de x. O valor de a pode ser calculado usando a seguinte fórmula:

$$a = \frac{(\Sigma y - b\Sigma x)}{n}$$

Onde $\sum y$ é a soma de y e $\sum x$ é a soma de x.

Além da regressão linear, existem outros tipos de regressão, como a regressão logística e a regressão de Poisson, que são utilizadas para modelar relações entre variáveis de natureza diferente.

Por outro lado, na árvore de regressão, o objetivo é prever um valor numérico, e não uma classe. Para isso, a árvore utiliza os conceitos de média e desvio padrão, que possibilitam um resultado numérico. No contexto da regressão, os preditores nos nós principais são determinados através da análise do desvio padrão dos valores da resposta para cada característica preditora. Isso contribui para a identificação das características mais significativas na previsão da resposta e orienta a construção da árvore de decisão.

Assim como citado anteriormente, uma árvore de regressão é um modelo de aprendizado de máquinas que prevê valores numéricos de uma variável dependente com base em variáveis independentes. Diferentemente da árvore de regressão, uma árvore de classificação prevê classes ou categorias de uma variável dependente. Por exemplo, uma árvore de regressão pode prever o preço de um determinado produto com base em características como utilidade, enquanto uma árvore de classificação pode prever se um cliente fará uma compra com base em seu comportamento de compras anteriores.

Outra diferença que pode ser destacada é o tipo de saída gerada pelo modelo. A árvore de regressão produz uma previsão numérica, geralmente um valor médio, que representa a estimativa do valor da variável dependente. Já a árvore de classificação gera uma previsão categórica, indicando a qual classe está pertencendo.

Assim, o processo de construção de uma árvore de regressão é baseado na minimização do erro de predição, utilizando critérios como o erro quadrático médio (Breiman et al., 1984). Na árvore de classificação, o processo de construção é baseado na pureza das classes nos nós da árvore, utilizando critérios de impureza, como o índice Gini ou a Entropia (Breiman et al., 1984). Esses critérios ajudam a identificar a melhor divisão dos dados em grupos homogêneos, tornando a árvore um modelo eficaz para classificação de novos dados.

Em resumo, a árvore de regressão é usada para prever valores numéricos, enquanto a árvore de classificação é usada para prever classes ou categorias.

3.3 ALGORITMO DE CONSTRUÇÃO DA ÁRVORE

A árvore de decisão é um algoritmo de aprendizado de máquina que constrói uma estrutura de árvore a partir de um conjunto de dados para tomar decisões. Essas decisões são baseadas nas condições especificadas em cada nó da árvore e, finalmente, levam a uma folha que representa a classe ou o valor previsto para a entrada dada.

O algoritmo de árvore de decisão usa um critério de divisão para determinar qual atributo dividir em cada nó da árvore. Uma medida comum usada para medir a impureza do conjunto de dados é a entropia. A entropia mede a quantidade de incerteza presente em um conjunto de dados. Quanto maior a entropia, maior é a incerteza (Mitchell, 1997).

O cálculo da entropia é feito para cada atributo que pode ser dividido em um nó. O atributo com a menor entropia após a divisão é escolhido como o atributo de divisão. Isso significa que a divisão desse atributo reduzirá a incerteza e aumentará a pureza das classes resultantes (Mitchell, 1997).

A árvore de decisão pode ser complexa ou sobreajustar-se aos dados de treinamento. Para evitar esses problemas, é necessário realizar poda e utilizar técnicas de validação cruzada para encontrar um equilíbrio entre complexidade e generalização. Para evitar isso, a técnica de poda é frequentemente usada. A poda consiste na remoção de sub-árvores que são partes de uma árvore que não contribuem para o aprimoramento da precisão da árvore. Isso é alcançado ao eliminar as folhas cuja presença teria o potencial de reduzir a acurácia da árvore.

A técnica de validação cruzada é amplamente utilizada para avaliar a precisão da árvore de decisão tanto antes como depois da poda. Essa abordagem envolve dividir o conjunto de dados em várias partes, treinando o modelo em algumas dessas partes e testando-o em outras. Esse processo é repetido várias vezes, garantindo que todas as partes do conjunto de dados sejam usadas tanto para treinamento quanto para teste.

Ao utilizar a validação cruzada, é possível obter uma estimativa mais confiável do desempenho da árvore de decisão em dados não vistos, ajudando a identificar se o modelo está sofrendo de overfitting1 (sobreajuste) ou não. Essa técnica é especialmente útil ao determinar a importância da poda, pois permite comparar a precisão do modelo antes e depois da poda, ajudando a encontrar o ponto ideal de equilíbrio entre complexidade e generalização. Dessa forma, a validação cruzada é uma ferramenta valiosa no processo de otimização e avaliação de árvores de decisão.

A construção de uma árvore de decisão tem três objetivos, quais sejam: diminuir a entropia (a aleatoriedade da variável objetivo), ser consistente com o conjunto de dados e possuir o menor número de nós.

3.3.1 ENTROPIA

A entropia é uma medida de incerteza ou desordem em um conjunto de dados. O cálculo da entropia é usado para determinar a pureza de um conjunto de dados em relação a uma variável de classe.

Em teoria de probabilidade um sistema completo de eventos $A_1, A_2 \cdots, A_J$ é um conjunto de eventos tal que um e somente um deles deve ocorrer a cada tentativa (por exemplo, o aparecimento de 1, 2, 3, 4, 5 ou 6 pontos no lançamento de um dado). Se tivermos os eventos $A_1, A_2 \cdots, A_J$ de um sistema completo, juntamente com suas probabilidades estimadas $p_1, p_2 \cdots, p_J$ ($p_i > 0$, $\Sigma_{i=1}^{j} p_i = 1$), então temos um esquema finito

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_j \\ p_1 & p_2 & \cdots & p_j \end{pmatrix}$$

Todo esquema finito descreve um estado de incerteza, onde se deseja prever o resultado de um experimento com base nas probabilidades de cada evento. Claramente, o grau de incerteza é diferente para esquemas diferentes.

A medida de entropia busca, então, medir o grau de incerteza presente em cada esquema finito, de acordo com a função

$$H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \times log_2 p_i \qquad 1$$

onde os logaritmos são tomados numa base fixa qualquer e é atribuído

$$\mathrm{p_i} \times \log_2 p_i = 0 \text{ sempre que } p_i = 0.$$

A entropia é máxima quando as possibilidades são equiprováveis, ou seja, $p_i = p_j, \forall i \neq j$, a equação $H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \times log_2 p_i$ 1 toma a seguinte forma:

$$H = \log_2 n$$

A entropia é nula, se, e somente se, algum dos números p_i for igual a 1 e todos os demais iguais a 0.

Um caso particular é a entropia binária, que é definida tendo uma variável aleatória com dois valores possíveis, ou seja,

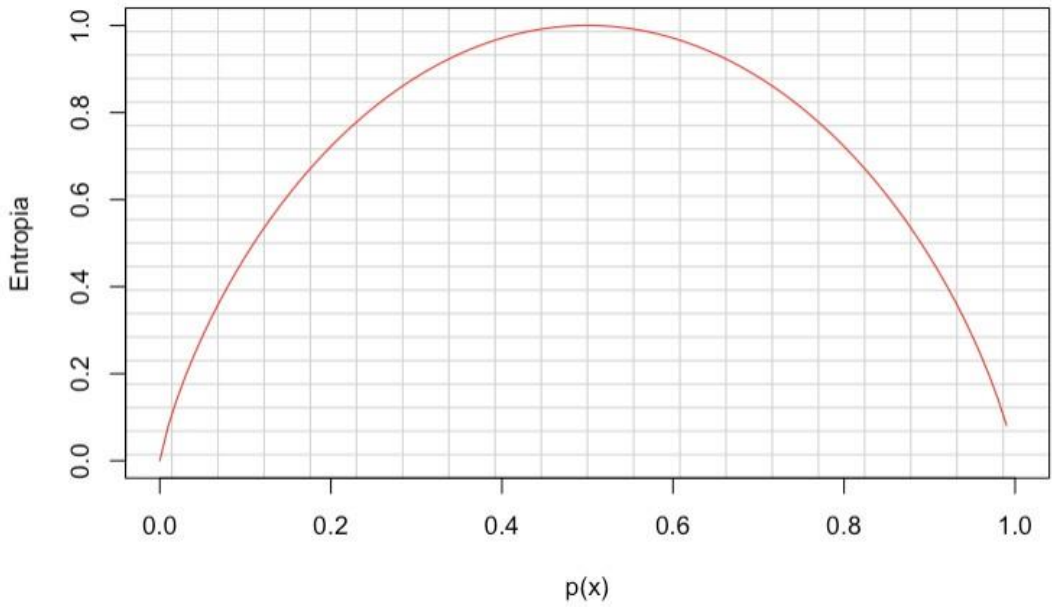
$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$

Quando a proporção de exemplos positivos chega a 0,5 temos o grau mais alto de entropia, pois em um conjunto binário esse é o ápice da mistura, depois ela vem diminuindo outra vez.

$$H(p_1, p_2) = -0{,}5 \log_2 0{,}5 - (1 - 0{,}5) \log_2 (1 - 0{,}5)$$
$$= -0{,}5 \log_2 0{,}5 - 0{,}5 \log_2 0{,}5$$
$$= -\log_2 0{,}5 = \log_2 \left(\frac{1}{2}\right)^{-1} = \log_2 2$$

Na Figura 3 é apresentado um gráfico da função entropia binária onde podemos verificar um comportamento côncavo da função $H(p)$.

Figura 3- Gráfico Função entropia binária.



Fonte: Autoria própria

Sendo assim, a entropia é uma medida comum de impureza em um conjunto de dados e é usada para selecionar o atributo mais informativo para dividir em um nó da árvore de decisão.

3.3.2 GANHO DE INFORMAÇÃO

A construção de uma árvore de decisão tem três objetivos, quais sejam: diminuir a entropia (a aleatoriedade da variável objetivo), ser consistente com o conjunto de dados e possuir o menor número de nós, ou seja, obter o maior ganho de informações possíveis.

O ganho de informação é uma medida importante para avaliar a relevância de um atributo na construção de uma árvore de decisão em algoritmos de aprendizado de máquina. Ele quantifica a redução da entropia esperada do conjunto de dados quando é feita a partição com base em um determinado atributo.

A entropia é uma medida de quão "desorganizado" ou "aleatório" é um conjunto de dados (Mitchell, 1997). Quanto maior a entropia, maior a falta de homogeneidade no conjunto de dados. Desta forma, para um atributo selecionado, o primeiro passo é ordenar as variáveis do nó em ordem crescente dos valores do atributo, e então agrupar as variáveis adjacentes de mesma classe. Nos passos subsequentes, agrupar os intervalos de maneira a diminuir a perda global dentro do nó, obtendo assim o maior ganho de informação (Mitchell, 1997). O ganho de informação, então, é dado pela diferença entre a entropia do conjunto de dados original e a entropia esperada após a partição com base no atributo A. Ele é calculado como:

Ganho (S, A) = Entropia(S) - E(A)

Quanto maior o ganho de informação, maior a redução da entropia esperada e, portanto, maior a relevância do atributo A na construção da árvore de decisão.

Ao usar um atributo A para particionar o conjunto de dados S em subconjuntos S_X, onde x é um valor possível do atributo A, podemos calcular a entropia esperada após a partição, denotada como E(A), somando as entropias dos subconjuntos ponderadas pelas proporções dos dados em cada subconjunto. A fórmula para calcular E(A) é

$$E(A) = \sum_{x \in P(A)} \frac{|S_x|}{|S|} \times Entropia\ (S_x)$$

onde P(A) é o conjunto de valores que o atributo A pode assumir.

Na construção de uma árvore de decisão, o objetivo é diminuir a entropia do conjunto de dados, ou seja, torná-lo mais homogêneo em relação à variável objetivo. Além disso, a árvore de decisão deve ser consistente com o conjunto de dados, ou seja, deve representar adequadamente os padrões presentes nos dados de treinamento. O ganho de informação é uma métrica útil na escolha dos atributos a serem utilizados na construção da árvore de decisão, levando em consideração esses objetivos. Por fim, busca-se construir uma árvore de decisão com o menor número de nós possível, para evitar a supercomplexidade e aumentar a interpretabilidade do modelo.

3.3.3 CRITÉRIO DE PODA

A podagem de árvores de decisão pode ocorrer em duas circunstâncias diferentes: pré-podagem ou poda descendente, que é usada para interromper o crescimento da árvore mais cedo, e pós-podagem ou poda ascendente, que ocorre após a árvore ter sido completamente construída. O processo pode ser realizado da seguinte forma:

- A árvore é percorrida em profundidade.
- Para cada nó de decisão, são calculados o erro no nó e a soma dos erros nos nós descendentes.
- Se o erro do nó for menor ou igual à soma dos erros dos nós descendentes, então o nó é transformado em folha. A diferença entre o erro do nó corrente e seus descendentes é medida por uma função que leva em consideração as proporções das classes, valorizando assim a pureza das partições.

Na pós-podagem, a árvore é inicialmente construída com o tamanho máximo e, em seguida, é podada usando métodos confiáveis de evolução. A subárvore com o melhor desempenho é escolhida como resultado.

No entanto, esse processo pode ser computacionalmente ineficiente, uma vez que uma árvore grande é gerada primeiro e, em seguida, reduzida para uma árvore mínima. Por outro lado, a pré-podagem é um método que interrompe o crescimento de uma árvore de decisão com base na confiabilidade das divisões feitas durante seu desenvolvimento. Embora seja um processo mais rápido, a prépodagem é menos eficiente do que a pós-podagem, pois existe um risco associado à escolha de uma sub-árvore que não é ótima em termos de desempenho e capacidade de generalização.

3.3.4 APLICAÇÃO PRÁTICA

Aqui está um exemplo com dados fictícios, utilizados para ilustrar a construção de uma árvore de decisão, juntamente com as medidas de entropia e ganho de informação.

A Tabela 1 representa um conjunto de treinamento com 14 dias observados e três atributos:

Tabela 1- Exemplo prático árvore de decisão

Dias	Temperatura água	Expectativa tempo	Vento	Pescar
1	Quente	Sol	Fraco	Sim
2	Quente	Sol	Forte	Sim
3	Quente	Nublado	Fraco	Sim
4	Moderada	Chuva	Forte	Não
5	Fria	Chuva	Fraco	Não
6	Fria	Chuva	Forte	Não
7	Fria	Nublado	Forte	Não
8	Moderada	Sol	Fraco	Sim
9	Fria	Sol	Fraco	Não
10	Moderada	Chuva	Fraco	Sim
11	Moderada	Sol	Forte	Sim
12	Moderada	Nublado	Forte	Sim
13	Quente	Nublado	Fraco	Sim
14	Moderada	Chuva	Forte	Não

Fonte: Autoria própria

Neste exemplo, deseja-se construir uma árvore de decisão para prever a condição do tempo com base nos atributos de Temperatura.

O próximo passo seria calcular a entropia do conjunto de treinamento, que é uma medida de impureza ou desordem dos dados. Quanto maior a entropia, mais impuro ou desordenado é o conjunto de dados. O objetivo é encontrar a divisão de atributos que minimize a entropia e, assim, maximize o ganho de informação.

O próximo passo é calcular a entropia de todo o conjunto. Verificando a tabela, é encontrado 8 exemplos positivos e 6 negativos, é muito comum usar a notação [8+,6-], logo a entropia é dada por:

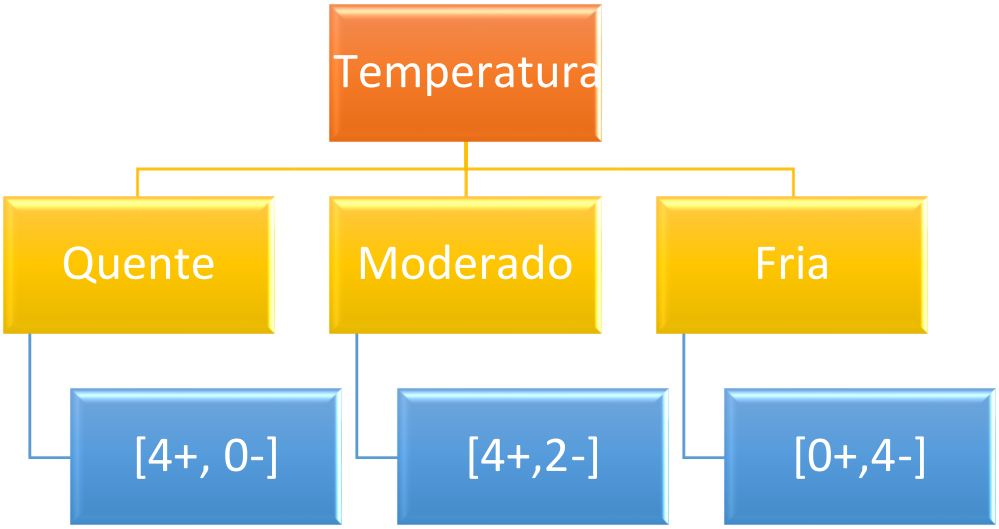
$$H = -\frac{8}{14}\log_2\frac{8}{14} - \frac{6}{14}\log_2\frac{6}{14} = 0{,}9852281$$

Esse índice será usado como parâmetro para outros índices e um deles é o ganho de informação.

Após determinar a entropia do conjunto de treinamento, é possível empregar uma fórmula específica para determinar o incremento de informação em relação a cada característica. Para decidir qual característica apresenta o maior incremento de informação entre as três variáveis, é essencial efetuar tanto a avaliação da entropia quanto do incremento de informação para cada característica.

- **Temperatura**

Figura 4- Análise do atributo Temperatura



Fonte: Autoria própria

$$E(Quente) = 0, \text{ pois } p_i = 0$$

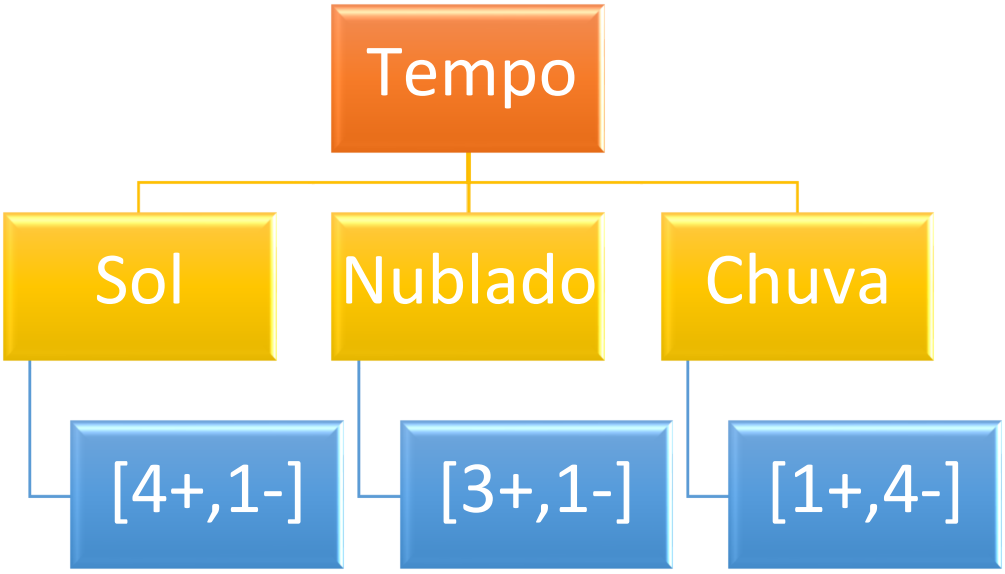
$$E(Moderada) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0{,}9182956$$

$$E(Fria) = 0, \text{ pois } p_i = 0$$

$$Ganho(S, temperatura) = 0{,}985 - \frac{4}{14} \times 0 - \frac{6}{14} \times 0{,}918 - \frac{4}{14} \times 0 = 0{,}5916728$$

- **Tempo**

Figura 5- Análise do atributo Tempo



Fonte: Autoria própria

$$E(Sol) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0{,}7219$$

$$E(Nublado) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0{,}8113$$

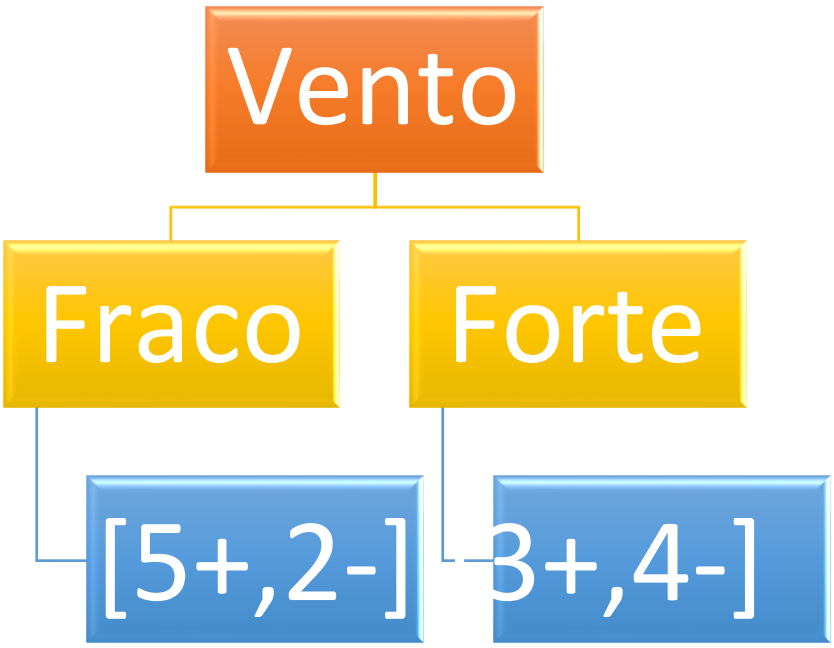
$$E(Chuva) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0{,}7219$$

$$Ganho(S, tempo) = 0{,}985 - \frac{5}{14} \times 0{,}7219 - \frac{4}{14} \times 0{,}8113 - \frac{5}{14} \times 0{,}7219$$

$$= 0{,}2377714$$

- **Vento**

Figura 6- Análise do atributo Vento



Fonte: Autoria própria

$$E(Fraco) = -\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7} = 0{,}8631$$

$$E(Forte) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0{,}9852$$

$$Ganho(S, vento) = 0{,}985 - \frac{7}{14} \times 0{,}8631 - \frac{7}{14} \times 0{,}9852 = 0{,}06105375$$

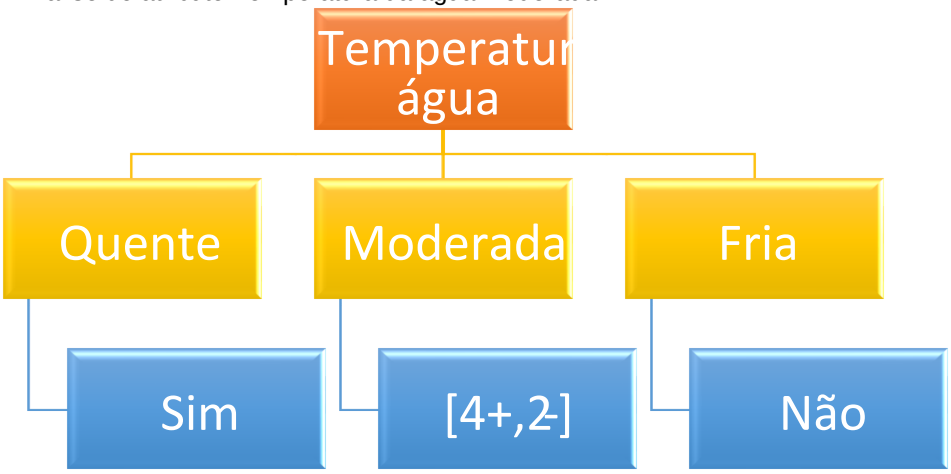
Após analisar as três variáveis - Temperatura, Tempo e Vento – conclui-se que a Temperatura apresentou o maior ganho de informação, tornandose o nó raiz da árvore de decisão.

$$Ganho(S, temperatura) = 0{,}5916728$$

$$Ganho(S, tempo) = 0{,}2377714$$

$$Ganho(S, vento) = 0{,}06105375$$

Figura 7- Análise do atributo Temperatura da água moderada



Fonte: Autoria própria

Com base em uma análise atualizada, é necessário considerar a temperatura da água como o ponto de partida e examinar as novas métricas, que serão divididas em três categorias: quente, moderada e fria. De acordo com a tabela examinada, se a água estiver quente, a pesca será bem-sucedida, enquanto se estiver fria, a pesca não será bem-sucedida. Portanto, é importante analisar a métrica moderada, levando em consideração o tempo e o vento.

$$E(Moderada) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0{,}9182958$$

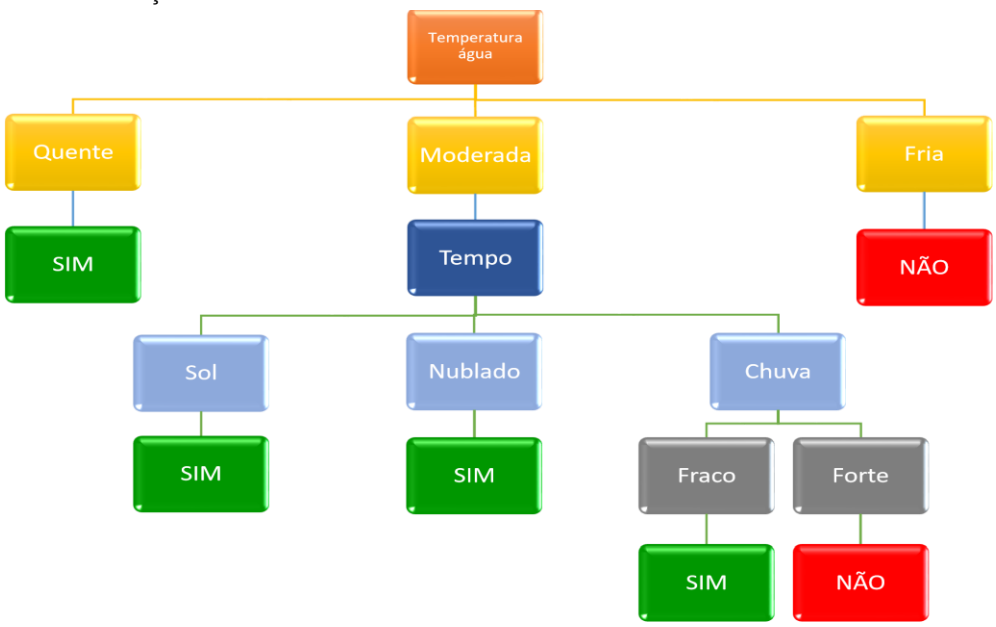
$$E(Moderada, tempo) = 0{,}9182958 - \frac{2}{6}\times 0 - \frac{1}{6}\times 0 - \frac{3}{6}\times 0{,}9182958$$

$$= 0{,}4591479$$

$$E(Moderada, vento) = 0{,}9182958 - \frac{3}{6}\times 0{,}9183 - \frac{3}{6}\times 0{,}9183 = 0$$

Depois de revisar as métricas, constata-se que o vento não exerce influência em nossa tomada de decisão. Portanto, considerando apenas as condições climáticas, teremos três novos cenários: sol, nublado e chuva. Se o dia estiver ensolarado ou nublado, iremos pescar. No entanto, se o dia estiver chuvoso, será analisada a métrica do vento. Se o vento for fraco, teremos um bom dia de pesca. Caso contrário, com ventos fortes, não haverá pescaria.

Figura 8- Construção da Árvore de decisão



Fonte: Autoria própria

O atributo com o maior ganho de informação é escolhido como o atributo de divisão na construção da árvore. Esse processo de cálculo de entropia e ganho de informação é repetido recursivamente para cada nó da árvore até que os critérios de parada sejam atingidos, resultando em uma árvore de decisão final.

Assim, o exemplo apresentado acima envolve a construção de uma árvore de decisão com base em medidas de entropia e ganho de informação para prever a condição do tempo com base em dados de temperatura e umidade observados em 14 dias.

Desta forma, após a explanação detalhada do algoritmo de árvore de decisão passo a passo, realizado manualmente, no próximo capítulo, será mostrada a existência de ferramentas que facilitam essa construção. Nessa seção, será abordado especificamente o RStudio, uma poderosa plataforma que oferece recursos e funcionalidades que simplificam a implementação e análise dos modelos de árvore de decisão. Será visto como o RStudio pode enriquecer nossa experiência, permitindo uma abordagem eficiente e produtiva na construção desses modelos.

3.4 SOFTWARES

Softwares são programas de computadores que foram desenvolvidos para ajudar na execução de tarefas específicas. Eles são criados para melhorar a eficiência e a produtividade em uma ampla gama de setores, incluindo negócios, educação, saúde e entretenimento.

A origem dos softwares remonta aos primeiros dias da computação, quando as primeiras linguagens de programação foram desenvolvidas. O primeiro software comercialmente disponível foi o compilador FORTRAN, lançado pela IBM em 1957 (NUNES, Lucas, 2021). Desde então, a indústria de software cresceu exponencialmente, com a introdução de novas tecnologias e ferramentas de desenvolvimento.

A importância dos softwares é amplamente reconhecida em todo o mundo. Eles são cruciais para a automação de tarefas repetitivas, melhorando a qualidade e a eficiência do trabalho. Eles também ajudam na tomada de decisões, fornecendo insights valiosos a partir dos dados e informações coletadas. Além disso, o desenvolvimento de softwares tem sido uma importante fonte de emprego e inovação em todo o mundo.

Em 2021, o mercado global de software foi avaliado em US$ 540 bilhões, com previsão de crescimento contínuo nos próximos anos (Statista, 2021). Isso reflete a importância e a influência dos softwares na economia global e em nossas vidas diárias.

Em resumo, softwares são uma parte fundamental de nossas vidas e da economia global, ajudando na automação, na tomada de decisões e na inovação. Como disse o co-fundador da Microsoft, Bill Gates: "O software é um grande espírito criador. As pessoas usam o software para criar coisas incríveis e eu acredito que o software pode capacitar cada pessoa e cada organização a fazer mais e alcançar mais do que jamais imaginaram ser possível". (2007)

Os softwares e as árvores de decisão estão intimamente relacionados. Afinal, as árvores de decisão são construídas usando algoritmos e implementadas em softwares para serem usadas na tomada de decisões.

Existem vários softwares disponíveis no mercado que permitem a construção de árvores de decisão, desde ferramentas mais simples até plataformas de análise de dados mais avançadas. Alguns exemplos de softwares populares que podem ser usados para construir árvores de decisão incluem o Microsoft Excel, o IBM SPSS, o R e o Python.

No contexto deste trabalho, optamos por programar os dados e realizar nossas análises utilizando o software RStudio, além disso, para visualizar os resultados usaremos o Power BI.

A escolha pelo R se deve, em grande parte, à sua versatilidade e capacidade de manipulação de dados, bem como à ampla gama de pacotes disponíveis para a construção de modelos estatísticos e visualizações de dados.

O R é uma linguagem de programação de código aberto e um ambiente de desenvolvimento integrado (IDE) amplamente utilizado para análise de dados, modelagem estatística e construção de visualizações de dados. Ele foi criado originalmente por Ross Ihaka e Robert Gentleman na Universidade de Auckland, na Nova Zelândia, em 1993, e desde então tem sido continuamente atualizado e mantido por uma grande comunidade de desenvolvedores em todo o mundo.

Com o R, é possível aplicar uma ampla variedade de técnicas de análise de dados, desde a análise exploratória inicial até a modelagem mais avançada, incluindo a construção de árvores de decisão. Além disso, o R também oferece recursos para visualização de dados, permitindo a criação de gráficos e figuras informativas e atraentes.

A seguir serão apresentadas as principais funções as serem utilizadas dentro do software RStudio para a construção do modelo, bem como as previsões.

A função “rpart” criará a árvore de decisão com uma expressão de entrada e a base de dados que utilizaremos para criar o modelo.

Com o modelo criado, é possível analisar como ficou à disposição da árvore e os respectivos graus de entropia, utilizando a função “prp”.

Com a função “rpart.plot” pode-se plotar a árvore e assim analisar a quantidade de observações por nó folha existem no modelo criado.

Utilizando a função “summary”, é possível entrar no detalhe do modelo previsto para analisar o grau de importância de cada variável e as proporções para cada classe.

A função “cptable” é usada para visualizar a tabela de critério de poda associada a uma árvore de decisão, e assim verificar algumas métricas que são úteis para avaliar a qualidade da árvore de decisão em relação à sua complexidade.

Por fim, a função “predict” no RStudio é usada para fazer previsões ou inferências com base em um modelo treinado. Ela permite que se aplique um modelo a novos dados e obtenha as previsões correspondentes.

Com todas essas ferramentas disponíveis no RStudio, é possível realizar uma exploração detalhada dos dados, treinar diferentes modelos de árvores de decisão e selecionar o mais adequado para o problema em questão.

Já a escolha do Power BI, se deve pela sua interface amigável, recursos poderosos de visualização e análise, além de permitir explorar os dados de maneira interativa e dinâmica. Através da criação de dashboards e relatórios personalizados, foi possível acompanhar métricas-chave, identificar tendências, padrões e oportunidades de negócio, apresentar resultados de forma clara e concisa para os tomadores de decisão.

Combinando a robustez da base de dados criada no RStudio com a capacidade de criação de relatórios e gráficos interativos do Power BI, a empresa terá em mãos uma poderosa ferramenta para extrair insights acionáveis e tomar decisões informadas. A visualização de dados de forma dinâmica e intuitiva permitirá explorar cenários e entender o impacto de diferentes variáveis no desempenho do negócio.

Além disso, a integração do Power BI com outras fontes de dados e ferramentas permitirá obter uma ampla visão do negócio e enriquecer ainda mais as análises. Com a possibilidade de compartilhar relatórios e dashboards de forma segura, a informação será disseminada de maneira ágil, proporcionando uma

cultura orientada a dados e favorecendo a tomada de decisão baseada em evidências.

O Power BI, desenvolvido pela Microsoft, é uma plataforma de análise de negócios que transforma dados brutos em informações úteis e visuais interativas. Essa poderosa ferramenta de business intelligence auxilia na tomada de decisões, oferecendo recursos avançados de análise, visualização de dados e compartilhamento de informações.

A importância do Power BI reside na sua capacidade de reunir dados de diversas fontes, como bancos de dados, arquivos, serviços na nuvem e aplicativos, e transformá-los em relatórios e dashboards interativos e de fácil compreensão. Ele permite que as empresas acompanhem métricas-chave, identifiquem tendências e padrões, descubram insights e compreendam melhor o desempenho de seus negócios.

Dessa forma, a combinação do RStudio com o Power BI representa uma abordagem completa e eficaz para a análise de dados e geração de insights, trazendo a empresa um passo mais perto de alcançar seus objetivos estratégicos e fortalecendo sua vantagem competitiva no mercado. Com uma base sólida e uma ferramenta poderosa, a empresa estará apta a enfrentar desafios complexos e a capitalizar oportunidades para o crescimento contínuo e o sucesso duradouro.

No próximo capítulo, exploraremos métricas de avaliação, como acurácia, precisão, recall, F1-Score, matriz de confusão, Curva ROC e AUC. Essas métricas servem para mensurar o desempenho do modelo e tomar decisões informadas para melhorar suas previsões.

4. MÉTRICAS DE AVALIAÇÃO

Atualmente, quando se fala em tecnologia, um termo que tem se tornado cada vez mais comum é o de acurácia. De acordo com o Dicionário Online de Português DICIO, define-se “acurácia” como “Exatidão e precisão numa medição ou no resultado apresentado por um instrumento de medição” (ACURÁCIA, 2023). Pensando na importância desse assunto e na forma como ele impacta o universo da tecnologia, buscamos estudar sua origem e conceitos.

Originado do “accuracy”, o termo passou a ganhar destaque através das áreas da Matemática e da Física, com o objetivo de analisar a proximidade de um resultado experimental com o real valor obtido, dessa forma, o termo visa instituir o grau de exatidão entre o valor desejado e o valor real.

A acurácia é uma métrica de avaliação de desempenho comum em problemas de aprendizado de máquina, especialmente em classificação. Ela representa a porcentagem de classificações corretas feitas pelo modelo em relação ao total de classificações. Em outras palavras, a acurácia mede quanto o modelo acerta na sua previsão (IDWALL, 2019).

Muitas vezes é confundida com “precisão”, porém, vale ressaltar que, embora possam possuir significados parecidos, são conceitos diferentes, onde a precisão é o grau de variação gerado por diferentes medições, assim, quanto mais preciso for um processo, menor será a variação entre os valores obtidos. Já a acurácia, é estabelecida pela “soma” entre exatidão e precisão. Veja na Figura 9.

Figura 9- Acurácia e Precisão

Fonte: IDWALL. O que é acurácia? [online]. [S.l.], 2019.

Mas afinal, qual a finalidade da acurácia? Ela deve ser aplicada considerando as necessidades distintas, pois, em alguns casos, o percentual de precisão precisa ser maior, portanto, para ser capaz de decidir se o indivíduo se encaixa nas regras de validação estabelecidas por sua empresa, contar com níveis altos de acurácia é fundamental.

As medidas de acurácia são usadas para avaliar a precisão de um modelo de machine learning em prever resultados corretos. A seguir, será apresentado algumas das medidas de acurácia.

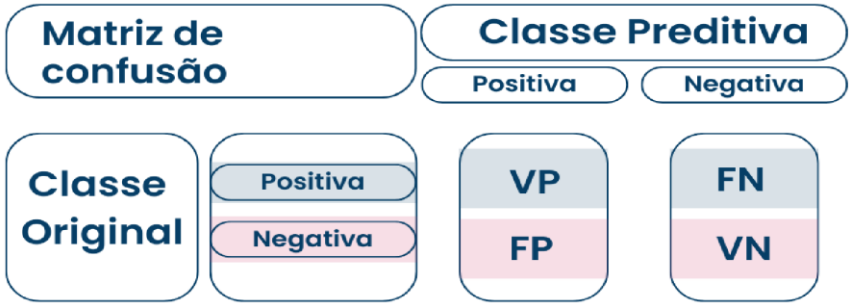
4.1 MATRIZ CONFUSÃO

De acordo com Diego Mariano (s.d.), a matriz de confusão é uma ferramenta fundamental na análise de resultados de modelos de classificação em Machine Learning. Ela fornece uma representação visual das previsões feitas pelo modelo em relação aos valores reais dos dados. A matriz de confusão organiza as previsões em quatro categorias:

- **VP- Verdadeiro positivo (True Positive — TP):** Representa os casos em que o modelo previu corretamente uma instância como pertencente à classe positiva, e de fato ela pertence a essa classe.
- **FN- Falso negativo (False Negative — FN):** Indica os casos em que o modelo classificou erroneamente uma instância como pertencente à classe negativa, mas na verdade ela pertence à classe positiva.
- **FP- Falso positivo (False Positive — FP):** Indica os casos em que o modelo classificou erroneamente uma instância como pertencente à classe positiva, mas na verdade ela pertence à classe negativa.
- **VN- Verdadeiro negativo (True Negative - TN):** Refere-se aos casos em que o modelo classificou corretamente uma instância como pertencente à classe negativa, e ela realmente pertence a essa classe.

Confira na Figura 10:

Figura 10- Matriz de confusão



Fonte: Autoria própria

Essa matriz permite uma compreensão detalhada do desempenho do modelo, avaliando sua capacidade de distinguir corretamente as classes e identificar possíveis erros ou tendências. A partir da matriz de confusão, diversas métricas de avaliação podem ser derivadas, como a acurácia, sensibilidade, precisão, especificidade e o F-Score, fornecendo uma visão abrangente do desempenho do modelo de classificação.

- Acurácia que é a proporção de previsões corretas em relação ao total de previsões feitas. É calculada como a soma das previsões corretas dividido pelo número total de previsões, por exemplo, se o modelo fizer 90 previsões corretas e 10 erradas, sua acurácia será de 90% (90/100).

$$\text{Acurácia} = \frac{VP + VN}{VP+FP+FN+VN}$$

- Precisão é uma métrica que avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores positivos:

$$Precision = \frac{VP}{VP + FP}$$

- Sensibilidade ou Recall, essa métrica avalia a capacidade do modelo de detectar com sucesso resultados classificados como positivos. Ela pode ser obtida pela equação:

$$Sensibilidade = \frac{VP}{FN + VP}$$

• VPN é uma métrica que avalia a quantidade de falsos positivos sobre a soma de todos os valores negativos:

$$VPN = \frac{FP}{FP + FN}$$

• Especificidade, avalia a capacidade do modelo de detectar resultados negativos. Podemos calculá-lo usando a equação:

$$Especificidade = \frac{FP}{FP + VN}$$

• F-score (F1), é uma média harmônica calculada com base na precisão e na sensibilidade. Ela pode ser obtida com base na equação:

$$F1 = 2 \times \frac{Precision \times Sensibilidade}{Precision + Sensibilidade}$$

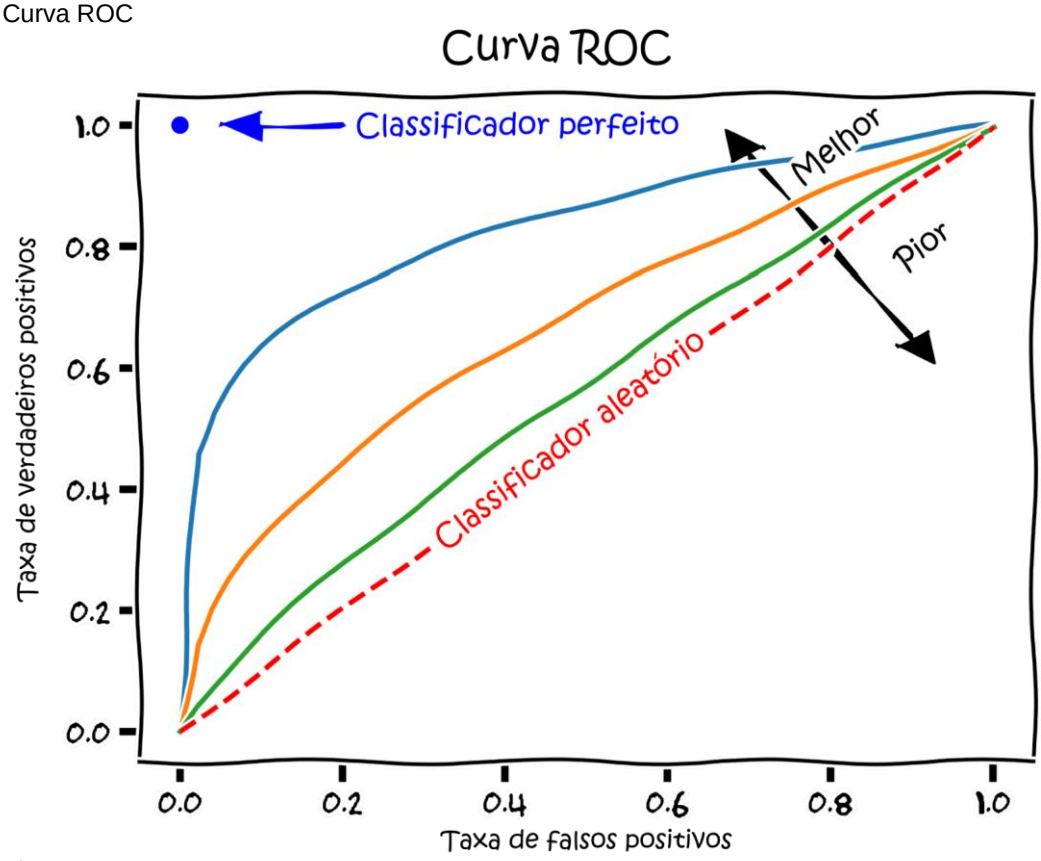
4.2 CURVA AUROC

De acordo com Polo (2020), AUC-ROC (Área sob a Curva ROC) é uma medida que fornece uma visão geral da capacidade discriminativa do modelo em classificar corretamente as instâncias, considerando tanto a sensibilidade quanto a especificidade. É uma medida comumente utilizada na avaliação de modelos de classificação binária. Ela representa graficamente o desempenho do modelo em relação à taxa de verdadeiros positivos (sensibilidade) em função da taxa de falsos positivos (1 - especificidade) para diferentes pontos de corte na classificação.

A curva AUROC é construída plotando-se os valores de sensibilidade no eixo y e os valores de 1 - especificidade no eixo x. Cada ponto na curva representa um limiar de decisão específico utilizado pelo modelo para classificar as instâncias. A área sob a curva (AUROC) varia entre 0 e 1, sendo que o valor de 1 indica um modelo perfeito, capaz de distinguir corretamente todas as instâncias das duas classes, enquanto o valor de 0.5 indica um modelo que realiza classificações aleatórias.

A interpretação da curva AUROC é a seguinte: quanto maior a área sob a curva, melhor é o desempenho do modelo em distinguir corretamente as instâncias das duas classes. Em outras palavras, uma maior área sob a curva indica que o modelo tem uma maior capacidade de classificar corretamente os casos positivos e negativos. A curva AUROC também possibilita a comparação do desempenho de diferentes modelos de classificação. Por exemplo, se temos dois modelos A e B, e a área sob a curva do modelo A é maior do que a do modelo B, podemos concluir que o modelo A possui uma melhor capacidade de classificação. Isso torna a curva AUROC uma ferramenta valiosa na seleção do melhor modelo para um determinado problema, ajudando a escolher aquele que oferece o melhor equilíbrio entre sensibilidade e especificidade.

Figura 11- Curva ROC



Fonte: Bioinfo, 2023.

4.3 OUTRAS MEDIDAS DE ACURÁCIA

A Validação Cruzada é uma técnica amplamente utilizada para avaliar a capacidade de generalização de um modelo. Ela envolve a divisão do conjunto de dados em conjuntos de treinamento e teste repetidas vezes, permitindo avaliar o desempenho do modelo em diferentes subconjuntos de dados. No contexto da análise de modelos de previsão e séries temporais, a métrica SMAPE (Symmetric Mean Absolute Percentage Error - Erro Percentual Médio Absoluto Simétrico) é frequentemente empregada para medir a precisão das previsões em relação aos valores reais.

Além do SMAPE, outras métricas importantes incluem o RAE (Relative Absolute Error - Erro Absoluto Relativo) e o MASE (Mean Absolute Scaled Error - Erro Médio Absoluto Escalado). O RAE calcula o erro absoluto médio do modelo em relação a um modelo de referência simples, como a média histórica, proporcionando uma comparação valiosa entre o modelo e um benchmark básico. Por sua vez, o MASE escala a diferença entre o erro absoluto médio do modelo e o erro do modelo de referência pelo desvio padrão da série temporal, levando em consideração a variabilidade dos dados.

Além disso, o índice de correlação de Pearson, muitas vezes referido como coeficiente de correlação, é usado para avaliar a relação linear entre as previsões do modelo e os valores reais. Um índice de correlação próximo de 1 indica uma forte correlação positiva, sugerindo que as previsões estão fortemente relacionadas aos valores reais, enquanto um índice próximo de -1 indica uma forte correlação negativa. A seguir apresentamos os parâmetros de correlação adotados.

Tabela 2- Parâmetros de correlação

Valor de correlação – COR (+ ou -)	Interpretação
$0 \leq COR < 0,2$	Bem fraca
$0,2 \leq COR < 0,4$	Fraca
$0,4 \leq COR < 0,7$	Moderada
$0,7 \leq COR < 0,9$	Forte
$0,9 \leq COR < 1,0$	Muito forte

Fonte: Autoria própria

Juntas, essas métricas (SMAPE, RAE, MASE e índice de correlação) são ferramentas valiosas para avaliar e comparar o desempenho de modelos de previsão em diferentes cenários e conjuntos de dados durante o processo de Validação Cruzada.

As equações para encontrar SMAPE, RAE, MASE, estão descritas abaixo:

$$SMAPE = \frac{2}{n}\sum_{t=1}^{n}\frac{|Previsto - real|}{Previsto + real}$$

$$RAE = \frac{\sum_{i=1}^{n}|Previsto - real|}{\sum_{i=1}^{n}|real - Média_{valores\ reais}|}$$

$$MASE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Previsto - real|}{\frac{1}{n-1}\sum_{i=2}^{n}|real_i - real_{i-1}|}$$

Cada medida de acurácia é apropriada para diferentes tipos de problemas e é importante entender as implicações de cada uma antes de escolher a medida mais apropriada para avaliar o desempenho de um modelo.

5. MODELO DE ÁRVORE DE DECISÃO APLICADO AO VAREJO

O avanço tecnológico e o aumento exponencial de informações têm desempenhado um papel fundamental no profundo remodelamento do setor varejista. Devido à crescente competitividade no mercado, as estratégias de tomada de decisão tradicionais, como análise de histórico de vendas e gerenciamento de estoque, já não são mais suficientes para impulsionar o crescimento das empresas. Nesse contexto, a utilização de dados provenientes de diversas fontes, incluindo transações de compras, interações online e feedback dos clientes, tornou-se imperativa para compreender a jornada do consumidor e criar experiências de compra mais personalizadas e contínuas.

A relevância dessas inovações para o progresso do setor é inegável. Um exemplo notável é o caso da Kroger, a maior rede de supermercados dos EUA, que em 2019 tomou uma medida audaciosa para se manter à frente. Reconhecendo a importância de se adaptar às transformações, a Kroger investiu na contratação de engenheiros especializados para desenvolver prateleiras equipadas com sensores de reconhecimento e capacidade de interação com os clientes. Essa iniciativa não apenas demonstra a integração entre tecnologia e varejo, mas também destaca como as empresas estão dispostas a abraçar a inovação para melhorar a experiência do cliente e impulsionar seus próprios resultados.

> "prateleiras digitais" que mostram anúncios (que podem até ser traillers de filmes) e alteram preços automaticamente, além de terem sensores que monitoram os produtos e ajudam o público a passar mais rápido pelos corredores -quem usa o aplicativo da rede é guiado para encontrar os itens de sua lista de compras e a prateleira digital mostra um ícone personalizado abaixo do produto escolhido quando você chega até ele. ("Desenvolvimento do Varejo com o Avanço da Tecnologia", 2019)

O exemplo da Kroger reflete uma tendência crescente no setor de varejo, em que as empresas estão se afastando das abordagens tradicionais e buscando ativamente soluções tecnológicas para se diferenciarem em um mercado cada vez mais competitivo. A capacidade de aproveitar dados e tecnologia não apenas transformou a maneira como as empresas gerenciam suas operações, mas também redefiniu a relação entre varejistas e consumidores, impulsionando uma nova era de experiências de compra personalizadas e aprimoradas. Diante dessa evolução, é evidente que o uso de modelos matemáticos pode ser uma ferramenta crucial para orientar as empresas na maximização de seus ganhos e minimização de perdas. Como mencionado anteriormente, existem vários modelos estatísticos que podem ser usados para análise de dados e tomada de decisões. A árvore de decisão é um desses modelos que pode ser útil para uma empresa em sua busca por otimizar seus resultados financeiros.

A criação de um modelo de árvore de decisão é uma ferramenta poderosa para empresas identificarem variáveis cruciais, como demanda de mercado, custos de produção e preços dos produtos, que afetam seus ganhos e perdas. Essas variáveis são organizadas em um modelo de árvore de decisão, onde cada uma delas influencia diretamente as decisões estratégicas no varejo. No contexto do varejo, as árvores de decisão desempenham um papel fundamental, auxiliando de diversas maneiras como:

- Segmentação de clientes: uma árvore de decisão pode ser usada para segmentar os clientes em diferentes grupos com base em variáveis relevantes, como histórico de compras, preferências de produtos, demografia, entre outros. Isso pode ajudar os varejistas a personalizar suas estratégias de marketing e oferecer produtos e serviços direcionados para cada segmento.
- Previsão de demanda: as árvores de decisão podem ser usadas para prever a demanda futura de produtos com base em variáveis como histórico de vendas, preço, promoções, eventos sazonais, entre outros. Essas previsões podem ser úteis para otimizar o estoque, planejar a produção e gerenciar a cadeia de suprimentos.

- Detecção de fraudes: árvores de decisão podem ser aplicadas para identificar padrões suspeitos ou fraudulentos em transações de varejo, como transações de cartão de crédito. Com base em variáveis como valor da transação, localização, horário, tipo de produto, entre outros, é possível construir uma árvore de decisão que ajude a detectar atividades fraudulentas e reduzir perdas.
- Recomendação de produtos: ao analisar as preferências e comportamentos de compra dos clientes, uma árvore de decisão pode ser usada para fazer recomendações personalizadas de produtos. Com base nas características e histórico de compras do cliente, a árvore pode direcionar a recomendação para determinadas categorias ou produtos específicos, aumentando as chances de venda adicional.
- Análise de precificação: as árvores de decisão também podem ser usadas para determinar estratégias de precificação. Com base em variáveis como custo, concorrência, elasticidade de preço e segmentação de clientes, uma árvore pode ajudar a identificar os melhores pontos de preço para maximizar a receita e o lucro.
- Previsão de lucro: as árvores de decisão podem ser usadas para prever os lucros futuros com base em variáveis como vendas, despesas, margens de lucro, investimentos e fatores econômicos. Essa previsão pode ajudar os varejistas a tomar decisões estratégicas, como ajustar preços, reduzir custos ou lançar novos produtos, para maximizar os lucros.
- Previsão de quantidade de produtos: uma árvore de decisão pode ser aplicada para prever a quantidade de produtos que serão vendidos em um determinado período com base em variáveis como histórico de vendas, promoções, sazonalidade, tendências de mercado e outros fatores relevantes. Essa previsão auxilia os varejistas na gestão de estoque, na definição de estratégias de produção e na otimização de recursos para atender à demanda esperada.

Em resumo, a criação de um modelo matemático como a árvore de decisão pode ser uma ferramenta poderosa para ajudar uma empresa a tomar decisões informadas e baseadas em dados. Isso pode levar a melhores resultados financeiros, otimização de recursos e aumento da competitividade no mercado.

5.1 BASE ANALISADA

O site Kaggle é uma plataforma amplamente reconhecida por abrigar conjuntos de dados diversificados e desafios de análise, serve como uma comunidade e um recurso valioso para cientistas, analistas e entusiastas da área,

fornecendo acesso a uma variedade de conjuntos de dados reais de diferentes domínios, incluindo comércio varejista, saúde, finanças e muito mais. É nesse ambiente que foi encontrado o conjunto de dados que contém informações essenciais sobre as vendas e lucros da "SuperLoja"[2].

É importante ressaltar que "SuperLoja" é um nome fictício dado a uma empresa do setor de atacado nos Estados Unidos. Embora a denominação seja fictícia, os dados reais contidos neste conjunto oferecem uma visão valiosa das operações e do desempenho financeiro de uma empresa semelhante no mercado. Após a determinação da base, procedeu-se ao seu tratamento, foi realizada a eliminação de inconsistências, valores ausentes e erros, tornando os dados mais confiáveis e prontos para análise.

Após a finalização do processo de tratamento e depuração dos dados, o conjunto de dados passou a conter 9.994 registros, englobando 21 variáveis, sendo elas:

- Row ID => ID Único para cada linha.
- Order ID => ID Único do Pedido para cada Cliente.
- Order Date => Data do Pedido do Produto.
- Ship Date => Data de Envio do Produto.
- Ship Mode => Modo de Envio especificado pelo Cliente.
- Customer ID => ID Único para identificar cada Cliente.
- Customer Name => Nome do Cliente.
- Segment => O segmento ao qual o Cliente pertence.
- Country => País de residência do Cliente.
- City => Cidade de residência do Cliente.
- State => Estado de residência do Cliente.
- Postal Code => Código Postal de cada Cliente.
- Region => Região à qual o Cliente pertence.
- Product ID => ID Único do Produto.
- Category => Categoria do Produto encomendado.
- Sub-Category => Subcategoria do Produto encomendado.

[2] https://www.kaggle.com/datasets/vivek468/superstore-dataset-final

- Product Name => Nome do Produto.
- Sales => Valor de Venda do Produto.
- Quantity => Quantidade do Produto.

- Discount => Desconto fornecido.
- Profit => Lucro/Prejuízo incorrido.

Pensando numa melhor avaliação do modelo, novas variáveis foram introduzidas, sendo elas:

- Diferenca_em_dias_envio => Que representa a diferença de dias entre a data de compra e data de envio do produto.
- Mes_ano => Representa o mês e ano de compra
- Receita_total => Foi obtido pelo produto do valor de venda (Sales) pela quantidade (Quantity)
- Lucro_total => Foi obtido pelo produto do Lucro (Profit) pela quantidade (Quantity)
- Custo_total => Diferença do Lucro_total pela Receita_total
- Custo_produto => Divisão do Custo_total pela quantidade (Quantity)
- Lucro_binário => 0 não ter lucro e 1 ter lucro.

Algumas outras variáveis preexistentes foram deliberadamente excluídas da modelagem devido a razões particulares, representadas na imagem abaixo. Este processo de seleção de variáveis foi conduzido meticulosamente para garantir que apenas os fatores mais relevantes e impactantes fossem considerados no escopo do estudo.

Figura 12- Variáveis eliminadas

```
#Eliminando variáveis semelhantes

```{r}
df$Customer_Name <- NULL    #Deixar Customer_ID
df$Product_ID <- NULL #Deixar Product_name
df$Postal_Code <- NULL #deixar o estado e cidade
df$Row_ID <- NULL #não tem necessidade do número de linhas
df$Order_ID <- NULL #não tem necessidade de deixar o número do pedido
```

Fonte: Autoria própria

Desta forma, após a definição da base e limpeza dos dados, a seguir foi realizada uma análise descritiva dos dados, explorando as principais características e tendências presentes na base.
```

5.2 ANÁLISE DESCRITIVA DAS VARIÁVEIS

No contexto da análise descritiva, este capítulo será subdividido em duas seções. No primeiro subcapítulo, serão exploradas as variáveis quantitativas e no segundo as qualitativas, com o intuito de identificar outliers e outros insights relevantes.

5.2.1 VARIÁVEIS QUANTITATIVAS

Foi utilizada a função summary do RStudio, que é uma ferramenta útil para obter informações resumidas sobre as variáveis em um dataframe. Ela fornece estatísticas descritivas básicas, como média, mediana, mínimo, máximo e quartis para variáveis numéricas.

- **Sales (valor de venda)**

Na Figura 13, é possível visualizar que o menor valor de venda foi de $0,44, de média $229,86 e o máximo $22638,48

Figura 13- Estatística básica do valor de venda

```
```{r}
summary(df$Sales)
```
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.444  17.280  54.490 229.858 209.940 22638.480 
```

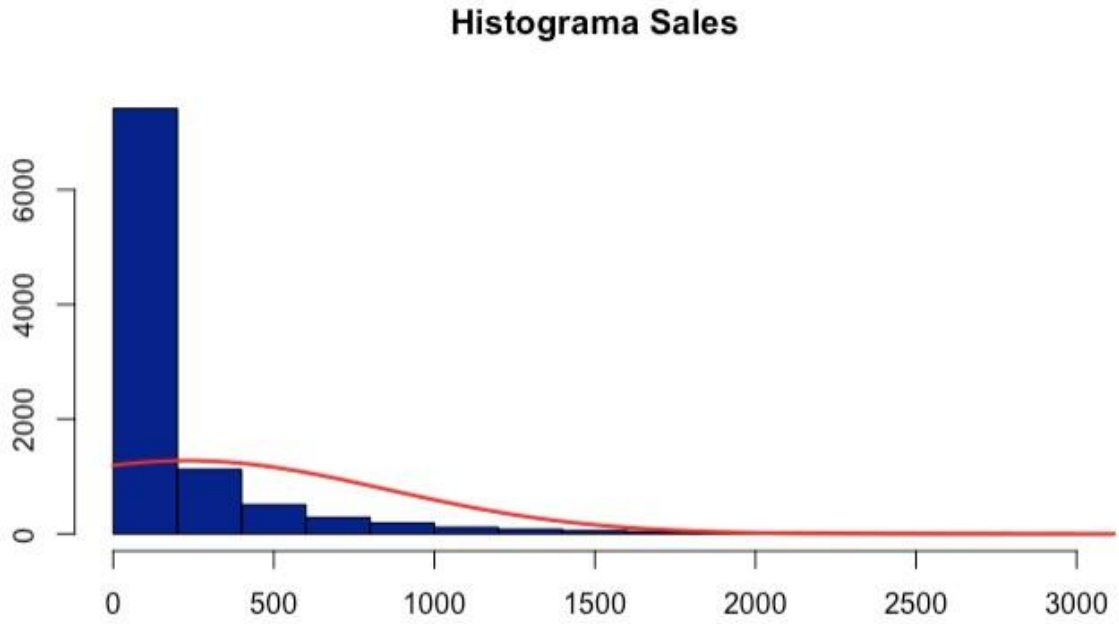
Fonte: Autoria própria

As estatísticas iniciais revelam que o menor valor de venda é de $0,44, indicando vendas muito baixas, enquanto a média é de $229,86, sugerindo vendas mais altas em média, mas com distorção devido a valores extremos, como $22.638,48.

Após análises adicionais, a mediana de $54 aponta que metade das vendas está abaixo desse valor, e o terceiro quartil de $209 indica que 75% das vendas o valor é relativamente baixo. No entanto, valores extremos, como o máximo de $22.638,48, influenciam significativamente a média, puxando-a para cima. Calculando o nono decil (90%), o valor obtido foi de $572,706, indicando que somente 10% dos valores estão acima deste montante.

A presença de valores extremos afeta a média das vendas, destacando a importância de considerá-los na interpretação das estatísticas. A mediana e o terceiro quartil indicam uma concentração de vendas em valores mais baixos, proporcionando uma visão mais completa da distribuição dos dados. Os resultados podem ser visualizados no histograma abaixo.

Figura 14- Histograma dos valores de vendas “Sales”



Fonte: Autoria própria

- **Quantity (Quantidade)**

Agora, analisando a variável quantidade por pedido (Quantity), obtém-se os seguintes resultados ilustrados abaixo:

Figura 15- Estatística básica da quantidade por pedido

```{r}
summary(df$Quantity)
```

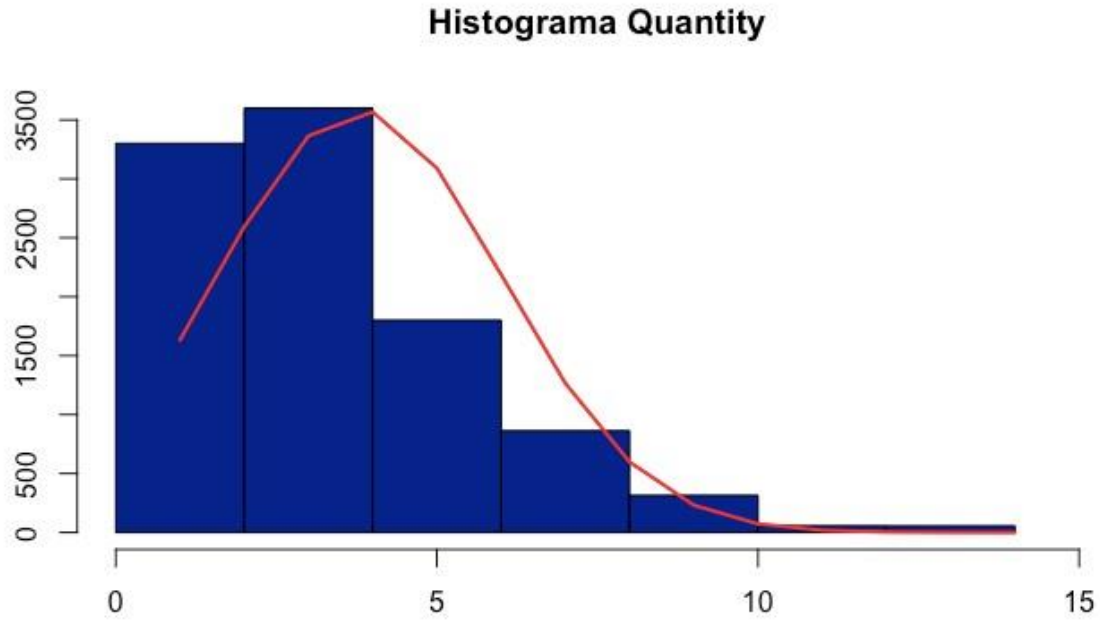
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.00    3.00    3.79    5.00   14.00
```

Fonte: Autoria própria

Ao analisar a quantidade de produtos por pedido, observa-se que a mediana é de 3, indicando que metade dos pedidos contém três ou menos itens. O primeiro quartil é de 2, sugerindo que 25% dos pedidos têm até 2 itens. A média é de 3,79, um pouco acima da mediana, devido à influência de valores extremos. O terceiro quartil é de 5, indicando que 75% dos pedidos têm até 5 itens. O valor máximo encontrado foi 14, representando um pedido com uma quantidade excepcionalmente alta de itens.

A maioria dos pedidos contém quantidades relativamente baixas, conforme indicado pela mediana e pelo terceiro quartil. No entanto, é importante estar ciente de valores extremos que podem afetar a média. Os resultados podem ser visualizados no gráfico abaixo:

Figura 16- Histograma da quantidade vendida por pedido “Quantity”



Fonte: Autoria própria

Dado que foram identificadas apenas três categorias, optou-se por realizar uma análise da quantidade de produtos por categoria, a fim de compreender a concentração das vendas em cada uma delas. Conforme o gráfico abaixo verificouse que a office-supplies possuí a maior quantidade de produtos, equivalente a 60,5% do total das vendas.

17- Gráfico Quantity x Category

Fonte: Autoria própria

- **Profit (Lucro por produto)**

Nesta etapa, optou-se por avaliar o lucro por produto, observando que o primeiro quartil foi de 1,729, a mediana de 8,666, o terceiro quartil de 29,36, com o valor máximo de 8399 e o mínimo de -6599,978. Essa análise revelou que a distribuição do lucro por produto é bastante variada, com alguns produtos gerando prejuízo (como evidenciado pelo valor mínimo negativo) e outros gerando lucros significativos (como indicado pelo valor máximo). A mediana e os quartis são úteis para entender a tendência central e a dispersão dos lucros, enquanto os valores mínimo e máximo destacam a presença de produtos com desempenho extremo, tanto positivo quanto negativo. Isso sugere a necessidade de avaliar e tomar medidas específicas em relação aos produtos que estão resultando em prejuízo.

18- Estatística básica do lucro por produto

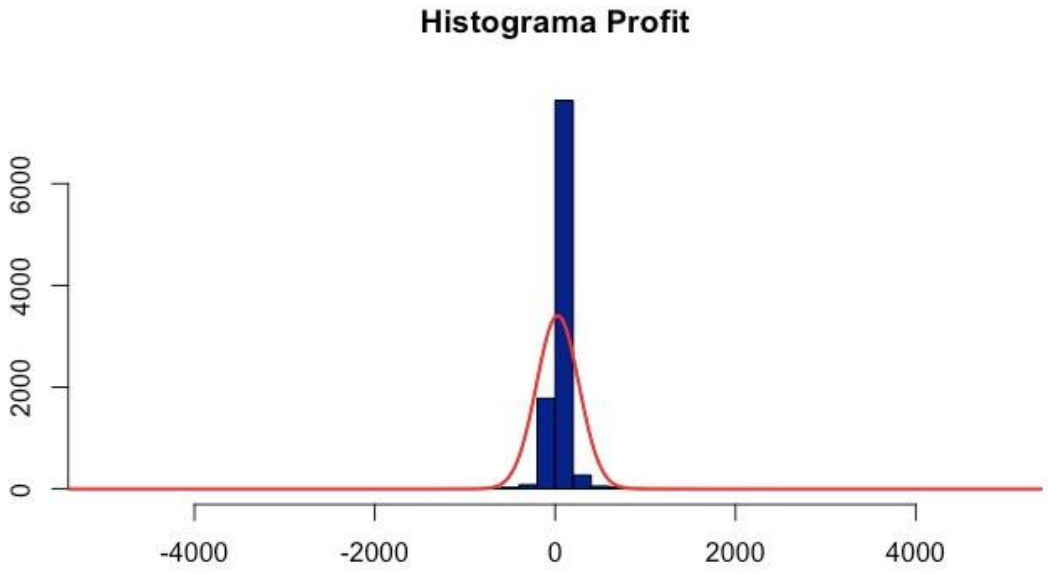
```{r}
summary(df$Profit)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
-6599.978     1.729     8.666    28.657    29.364  8399.976 
```

Fonte: Autoria própria

Considerando a análise anterior, ao examinar o lucro por exemplo, notase que os valores estão concentrados próximos a zero, com apenas alguns valores extremos. Essa observação é relevante, pois há produtos que, inclusive, resultaram em prejuízos, evidenciando a influência desses casos atípicos na análise, como visto na Figura *19*.

Figura 19-Histograma dos valores de lucro por produto “Profit”

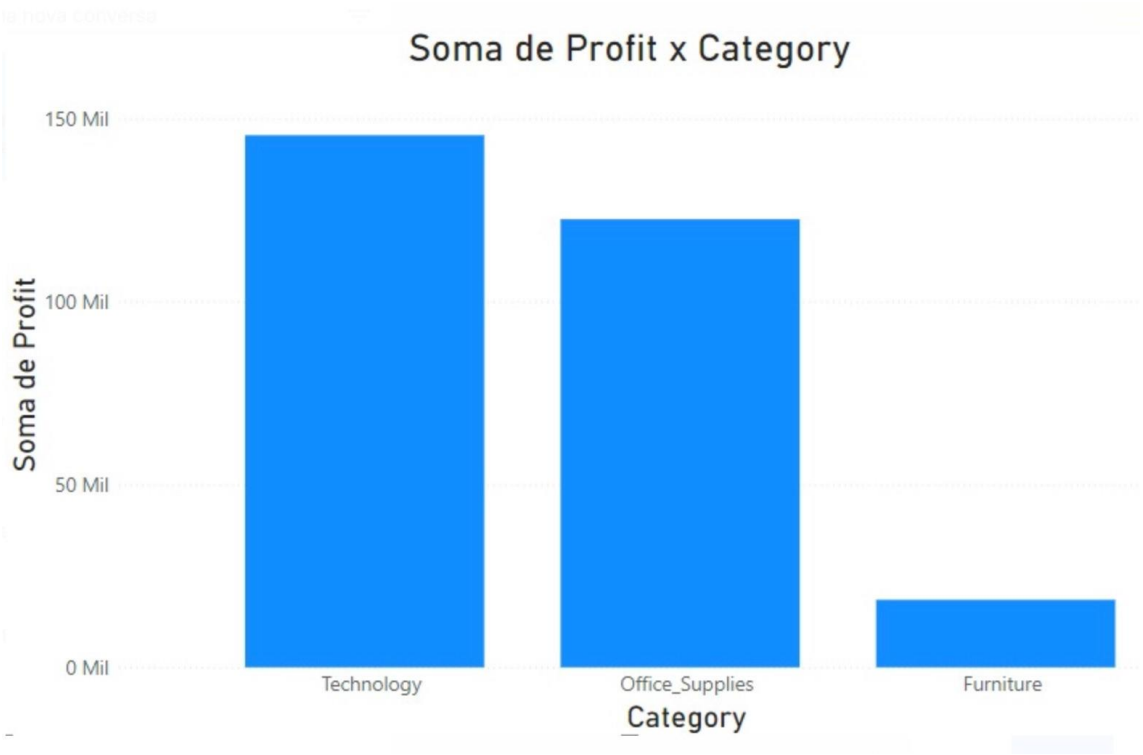


Fonte: Autoria própria

Na análise de desempenho por categoria, destaca-se que os produtos de tecnologia sobressaem ao gerar um lucro superior. Conforme ilustrado na imagem abaixo.

Revista Eletrônica do Curso de Licenciatura em Matemática
Centro Universitário Municipal de Franca

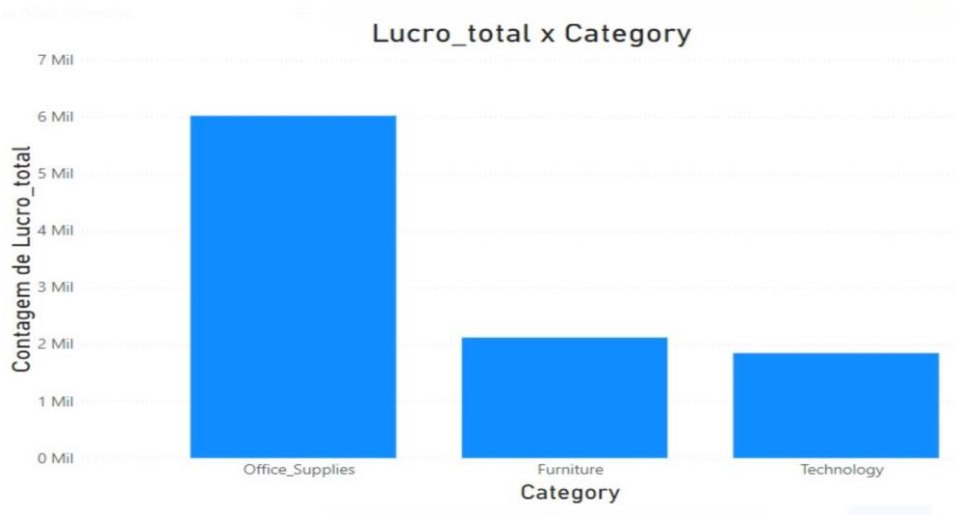
20- Gráfico Soma de Profit x Category



Fonte: Autoria própria

Ao analisar o desempenho das categorias de produtos, é evidente que os produtos de escritório se destacam devido à sua ampla demanda, o que, por sua vez, contribui para um lucro maior.

21- Gráfico Lucro total x Category

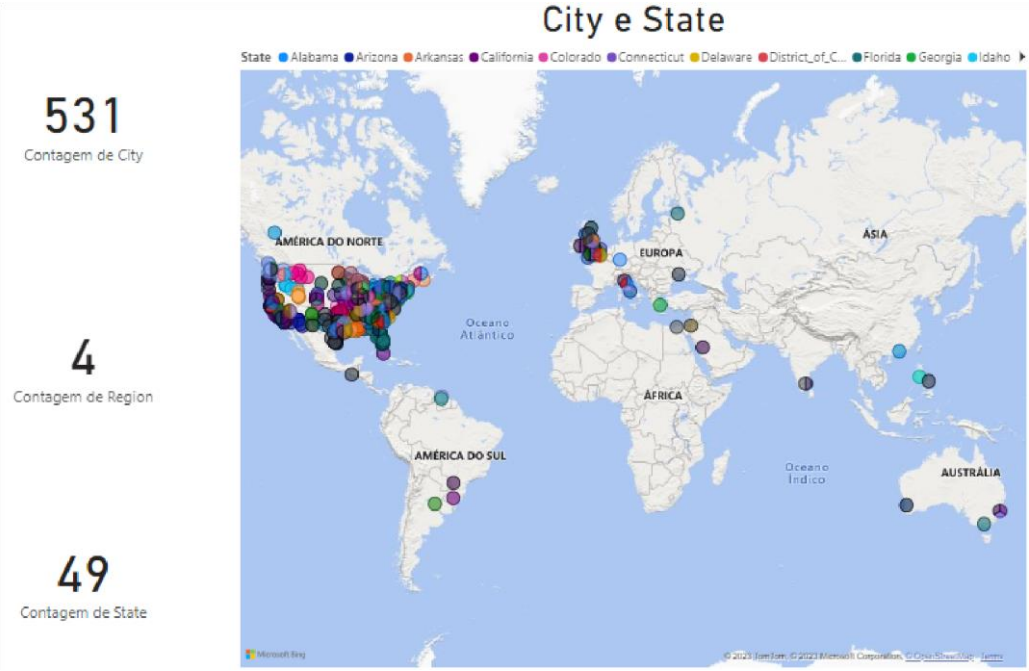


A análise de lucro por categoria mostra que os produtos de tecnologia geram mais lucro, enquanto os produtos de escritório têm mais vendas. Isso sugere que, embora os produtos tecnológicos sejam mais lucrativos individualmente, os produtos de escritório têm uma demanda maior.

5.2.2 VARIÁVEIS QUALITATIVAS

No contexto da análise realizada, foi observada a presença das variáveis categóricas: região, estado e cidade. As quais estão diretamente relacionadas com a localização dos clientes no momento das compras. A fim de compreender o perfil dos consumidores, conduziu-se uma análise descritiva que abrangeu simultaneamente todas essas variáveis, proporcionando insights valiosos sobre a distribuição geográfica dos clientes e suas preferências de compra em diferentes regiões, cidades e estados. Essa abordagem integrada permitiu uma exploração mais abrangente do comportamento dos clientes em relação à localização geográfica, como será visto na figura abaixo:

Figura 22- Análise descritiva das regiões de venda



Fonte: Autoria própria

Na análise realizada, foi constatado que a base de dados abrange quatro regiões distintas: Central, Sul, Leste e Oeste, englobando um total de 531 cidades e abrangendo 49 estados diferentes. Um destaque relevante é a região Central, que apresentou o maior fluxo de consumidores, com um alcance abrangendo 181 cidades e 13 estados, representados na Tabela *3*.

Tabela 3- Análise descritiva das regiões

Região	Cidades (City)	Estados (Stade)
Central	**181**	**13**
Leste	**108**	**14**
Sul	**125**	**11**
Oeste	**169**	**11**
Total	**531**	**49**

Fonte: Autoria própria

Além disso, vale ressaltar que a concentração de consumidores é mais significativa nos Estados Unidos, entretanto, observou-se a presença de consumidores em outros continentes, como América do Sul, Austrália, Europa e América Central. Essa diversidade geográfica da base de dados adiciona um aspecto interessante à análise, destacando a amplitude e a abrangência das operações da empresa em uma escala global.

No contexto das variáveis relacionadas à categoria, subcategoria e tipos de produtos, uma análise conjunta foi realizada, revelando um total de 1.850 tipos de produtos distribuídos em 17 subcategorias e 3 categorias diferentes. Notou-se que a concentração mais expressiva de produtos está na categoria de "Suprimentos de Escritório", cerca de 60,48% do total, e engloba nove subcategorias distintas. Isso destaca a extensa variedade e amplitude do portfólio de produtos oferecidos pela empresa nessa categoria específica, ressaltando sua importância e diversificação no mercado. Conforme pode ser visualizado na Tabela 4.

Tabela 4 - Análise descritiva categorias, subcategorias e tipos de produtos.

Categoria	Subcategoria	Tipos de Produtos	Quantidade
Furnitute (Mobília)	**4**	**380**	**8.028**
Office_suplies (Suprimentos de escritórios)	**9**	**1.058**	**22.906**
Technology (Tecnologia)	**4**	**412**	**6.939**
Total	**17**	**1.850**	**37.873**

Fonte: Autoria própria

Com isso, encerra-se este capítulo dedicado à análise dos dados. Ao longo desse processo, foram exploradas as variáveis geográficas, categorias de produtos e segmentos de clientes, proporcionando uma visão abrangente do cenário atual da empresa. Os insights obtidos têm o potencial de orientar decisões estratégicas futuras, aprimorando a eficiência das operações e impulsionando o crescimento no mercado. No próximo capítulo, serão abordados os resultados obtidos a partir das análises.

5.2.3 RESULTADOS

Neste capítulo, apresentam-se as conclusões e implicações dos resultados da análise conduzida. Os insights obtidos forneceram informações valiosas sobre o comportamento dos clientes e a amplitude do portfólio de produtos da empresa.

- Variáveis Categóricas: a análise das variáveis categóricas revelou uma distribuição geográfica significativa dos clientes em quatro regiões distintas, com destaque para a região Central. Esses achados indicam uma presença global de consumidores em diversos continentes, evidenciando a dimensão global das operações da empresa.
- Variáveis Relacionadas à Categoria de Produtos: no que diz respeito às variáveis relacionadas à categoria de produtos, notou-se uma diversidade no portfólio, com 1.850 tipos de produtos distribuídos em 17 subcategorias e 3 categorias diferentes. A categoria de "Suprimentos de Escritório" se destacou pela sua amplitude e diversificação.

Com base nas conclusões mencionadas, surgem algumas implicações importantes:

- Segmentação de Mercado: a compreensão da distribuição geográfica dos clientes permite uma segmentação de mercado mais precisa, facilitando a orientação de estratégias de marketing e vendas para regiões específicas.
- Ampliação do Portfólio: a diversidade de produtos, principalmente na categoria de "Suprimentos de Escritório," oferece oportunidades para expandir ainda mais o portfólio e atender às diversas demandas dos clientes.
- Globalização: a presença de consumidores em diferentes partes do mundo destaca a natureza global das operações da empresa, o que pode levar a considerações estratégicas adicionais para expandir a presença global ou adaptar produtos e serviços a mercados específicos.

Com base nas conclusões e implicações acima, as seguintes recomendações são apresentadas:

- Estratégias de Marketing Georreferenciadas: desenvolver estratégias de marketing personalizadas com foco nas diferentes regiões para maximizar o alcance e a eficácia das campanhas.
- Diversificação de Produtos: continuar diversificando o portfólio de produtos, explorando oportunidades em categorias com potencial de crescimento.
- Análise de Expansão Global: avaliar a viabilidade de expandir ainda mais a presença global, identificando mercados-chave com demanda potencial.

Os resultados da análise proporcionaram uma visão abrangente da base de clientes e do portfólio de produtos da empresa.

5.3 APLICAÇÕES DE MODELOS DE ÁRVORES DE CLASSIFICAÇÃO

Com o objetivo de organizar os estados (ou regiões) com base nos produtos vendidos, buscou-se obter insights que podem informar uma variedade de estratégias de negócios, incluindo logística, marketing e gerenciamento de estoque. É uma abordagem valiosa para personalizar as operações e estratégias de acordo com as características específicas de cada área geográfica.

Assim, ao buscar identificar padrões ou relações entre os estados (Numero_state) e as características dos produtos vendidos (Product_Name e Quantity), foi empreendida a construção de uma árvore de decisão com o propósito de categorizar os estados com base nessas características, conforme exemplificado no código abaixo, Figura *23*.

Figura 23- Código de Identificação de padrões

```
```{r}


modelo_1 = rpart( Numero_state ~ Product_Name+Quantity, data = df.train, control = rpart.control(cp
= .0001, maxdepth = 25),
  method = "class")

```
```

Fonte: Autoria própria

No entanto, devido à variável Product_Name conter muitas categorias (1.850 tipos de produtos), não foi viável desenvolver um modelo dessa maneira. Logo, optou-se por trabalhar com a variável subcategoria, que contém 17 tipos, cujo produtos estão agrupados por subcategoria. E assim, foi desenvolvido o modelo apresentado na Figura *24*.

Figura 24- Modelo desenvolvido por subcategorias

```{r}

modelo_2 = rpart( Numero_state ~ Numero_subcategoria+Segment+Quantity, data = df.train, control =
rpart.control(cp = .0001, maxdepth = 25),
  method = "class")

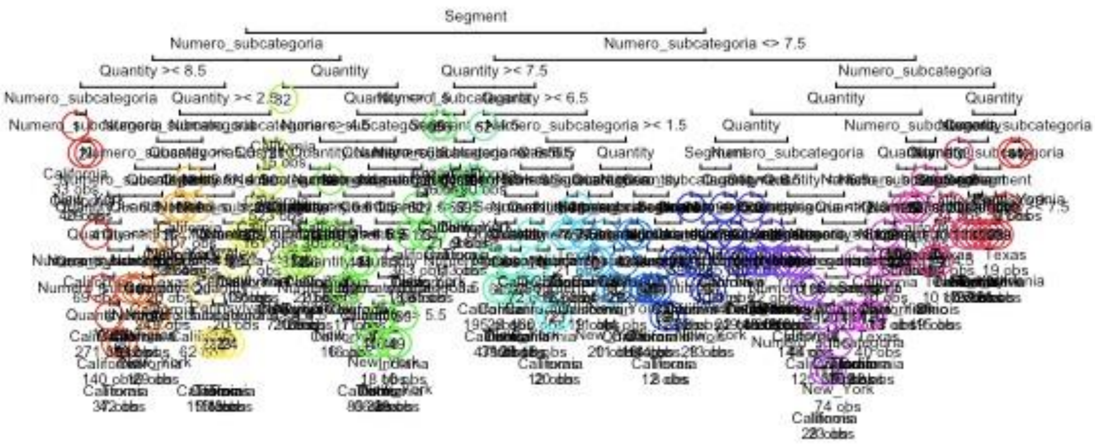
draw.tree(modelo_2, cex=0.5, print.levels = FALSE, size=2)

```

Fonte: Autoria própria

Dessa forma, obteve-se a árvore descrita a seguir, mas constatou-se que ela estava excessivamente complexa. Ao analisar os critérios de poda, percebeu-se que o erro relativo estava elevado, o que indica que o modelo não está conseguindo identificar padrões significativos nos dados.

Figura 25- Árvore do modelo por subcategorias



Fonte: Autoria própria

Como até o momento não foi possível desenvolver um modelo eficaz, capaz de agrupar os produtos com base na localização geográfica, foi adotada uma abordagem diferente. Houve a escolha de criar uma variável binária, 'Lucro_binaria' (1 para obteve lucro e 0 para não obteve lucro), que será a variável dependente ou alvo.

E assim, a intenção foi criar uma árvore de classificação utilizando várias variáveis independentes, incluindo 'Product_Name', 'State', 'Numero_subcategoria', 'Numero_regiao' e 'Quantity'. Com essa abordagem, buscou-se identificar padrões que permitem classificar se uma transação resultou em lucro ou não, com base nas informações disponíveis. Desta forma, temos o seguinte modelo, Figura *26*.

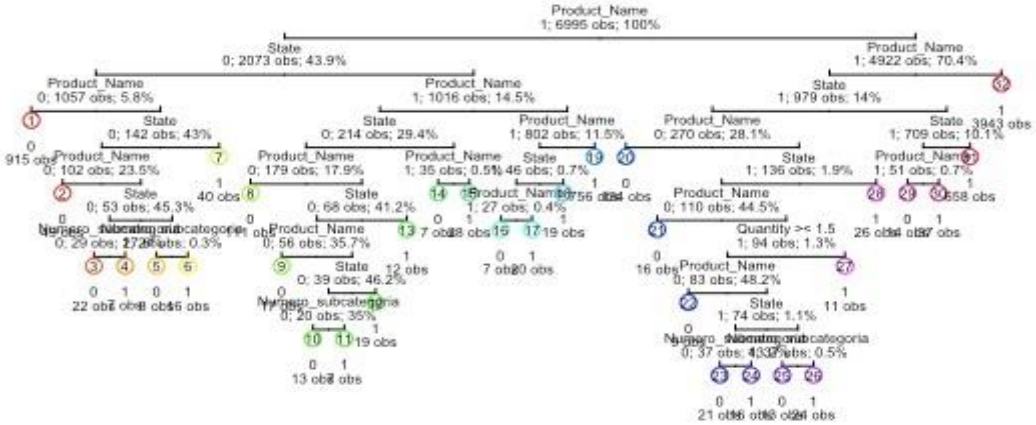
Figura 26- Árvore de classificação: Determinando o lucro

```
# Árvore de classificação modelo 3
```{r}
modelo_3 <- rpart(
  formula = Lucro_binaria ~ Quantity+ State+Numero_regiao+Numero_subcategoria+Product_Name,
  data = df.train,
  control = rpart.control(cp = 0.0001),
  method = "class"
)
draw.tree(modelo_3, cex = 0.45, nodeinfo = TRUE, print.levels = FALSE)
```
```

Fonte: Autoria própria

Este modelo gerou 32 folhas, considerando o critério de poda de 0,0001.

Figura 27- Árvore de Classificação: Modelo de lucro



Fonte: Autoria própria

Para aprimorar ainda mais a simplicidade da árvore de decisão, optamos por revisar o critério de poda, conforme a Figura *28*.

Figura 28- Análise do critério de poda

```
# Análise Critério de Poda
```{r}
modelo_3$cptable
```
```

	CP	nsplit	rel error	xerror	xstd
1	0.3404952658	0	1.00000000	1.0000000	0.02419448
2	0.0640932265	2	0.31900947	0.5637291	0.01910892
3	0.0429715950	3	0.25491624	0.5491624	0.01889071
4	0.0196649672	5	0.16897305	0.4756009	0.01772170
5	0.0123816460	6	0.14930808	0.4559359	0.01738835
6	0.0101966497	8	0.12454479	0.4304443	0.01694162
7	0.0087399854	9	0.11434814	0.4238893	0.01682394
8	0.0057052683	10	0.10560816	0.4173343	0.01670506
9	0.0050983248	16	0.07137655	0.4151493	0.01666516
10	0.0038844380	18	0.06117990	0.4107793	0.01658496
11	0.0014566642	22	0.04442826	0.4064093	0.01650420
12	0.0012138869	27	0.03423161	0.4049527	0.01647715
13	0.0007283321	30	0.03058995	0.4027677	0.01643647
14	0.0001000000	31	0.02986162	0.3998543	0.01638200

Fonte: Autoria própria

Após análise, optou-se por utilizar o valor de CP = 0.0057052683. Esse critério resultou em uma árvore de decisão simplificada com 11 folhas, conforme apresentado abaixo, Figura *30*.

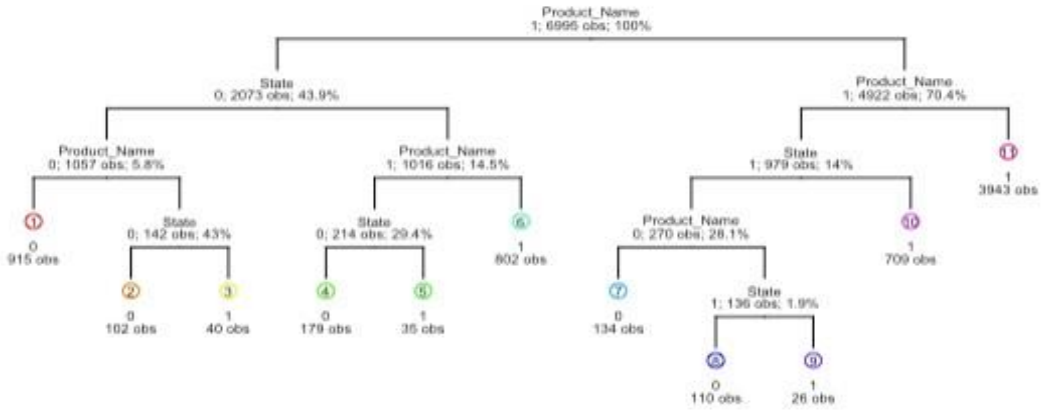
Figura 29- Realização da poda

```
# Realizar poda
```{r echo=TRUE}

modelo_3_v2 <- prune.rpart(modelo_3 ,cp = 0.0057052683)
draw.tree(modelo_3_v2, cex = 0.45, nodeinfo = TRUE, print.levels = FALSE)
```
```

Fonte: Autoria própria

Figura 30- Árvore Pós-Poda



Fonte: Autoria própria

Após a definição da árvore, o próximo passo é avaliar sua eficácia na classificação de transações como lucrativas ou não. Para isso, será aplicado medidas de acurácia que permitem verificar o desempenho do modelo em relação aos dados reais.

É crucial destacar a importância de avaliar dados que não fizeram parte do processo de modelagem. Ao aplicar as medidas de avaliação à base de teste, que foi mantida separada durante o desenvolvimento do modelo, garante uma avaliação imparcial e realista do desempenho da árvore de decisão. Isso permite obter uma visão confiável de como o modelo se comporta com novos dados, refletindo sua capacidade de generalização e aplicabilidade no mundo real.

Assim, temos a seguinte matriz confusão, Tabela 5:

Tabela 5- Matriz Confusão

MATRIZ CONFUSÃO		Valores previstos		
		0	1	Total
Valores Reais	0	389	150	539
	1	67	2254	2321
	Total	456	2404	2860

Fonte: Autoria própria

Logo, temos

$$Acurácia = \frac{389 + 2254}{2860} = 92\%$$

$$Precision = \frac{2254}{2404} = 94\%$$

$$Sensibilidade = \frac{2254}{2321} = 97\%$$

$$F1 = 2 \times \frac{94\% \times 97\%}{94\% + 97\%} = 95{,}4\%$$

Diante das medidas apuradas, podemos verificar que o modelo apresentou um desempenho considerável. Com uma acurácia de 92,41%, uma precisão de 93,76%, uma sensibilidade de 97,11%, e um valor F1 de 95,41%, podemos afirmar que o modelo demonstrou uma capacidade excepcional de fazer previsões precisas em relação à variável alvo, devido as medidas apresentarem valores próximos a 100%.

Além disso, foi calculado o valor do AUC-ROC (Área sob a Curva da Característica de Operação do Receptor) cujo valor obtido foi de 84%, este, indica que o modelo é capaz de distinguir eficazmente entre as classes positivas e negativas. Essas métricas refletem a eficácia do modelo na classificação de transações como lucrativas ou não lucrativas.

Com este modelo em mãos, a "SuperLoja" poderá direcionar seus esforços para compreender quais combinações de produtos, estados, subcategorias, regiões e quantidades têm maior probabilidade de gerar lucro. Essa capacidade permitirá à empresa tomar decisões estratégicas em relação às operações e campanhas de marketing.

5.4 APLICAÇÕES DE MODELOS DE ÁRVORES DE REGRESSÃO

Neste capítulo, será explorado o uso da árvore de regressão como uma ferramenta fundamental para prever tanto o custo do produto quanto a quantidade de produtos. Em seguida, o desempenho dessas previsões será avaliado utilizando diversas métricas, tais como índices de correlação, RAE (Erro Absoluto Relativo), MASE (Erro Médio de Escala Absoluta Média) e SMAPE (Erro Percentual Médio de Simetria Absoluta).

5.4.1 PREVISÃO DE CUSTOS

A previsão de custos é uma prática fundamental em qualquer empreendimento ou organização. Ela envolve a estimativa e análise dos gastos futuros que uma empresa pode incorrer ao realizar suas operações ou projetos. Através da previsão de custos, as empresas podem tomar decisões informadas sobre alocação de recursos, definição de preços de produtos ou serviços e

elaboração de orçamentos. Isso não apenas ajuda a controlar os gastos, mas também a otimizar a eficiência operacional, garantindo que os recursos sejam utilizados de forma eficaz e que metas financeiras sejam alcançadas.

Assim, ao avaliar os dados em análise, foi desenvolvida uma árvore de previsão de custos de produtos por meio do RStudio, como mostrado no seguinte script, Figura *31*.

Figura 31- Modelo de previsão de custo

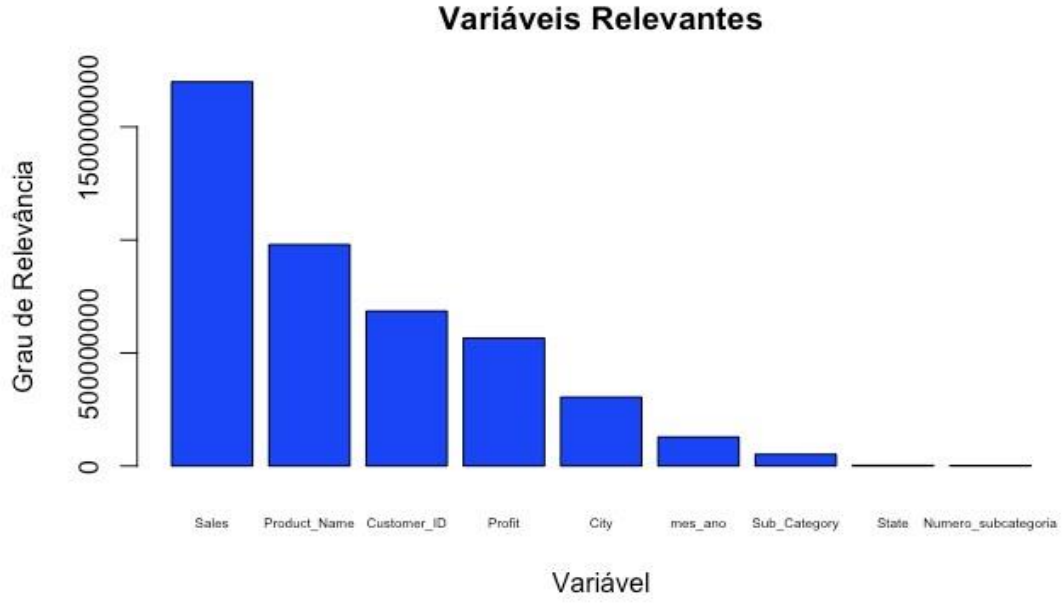
```
# Árvore previsão custo
```{r}
modelo_custo <- rpart(
  formula =  Custo_produto ~.,
  data = df.train,
  control = rpart.control(minsplit = 11, maxdepth = 8)
)
draw.tree(modelo_custo, cex = 0.45, nodeinfo = TRUE, print.levels = FALSE)

```
```

Fonte: Autoria própria

As variáveis consideradas na formação dessa árvore incluem as seguintes: Sales, Product_Name, Custumer_ID, Profit, City, mes_ano, Sub_category, State e Numero_subcategoria.

Figura 32- Variáveis Relevantes

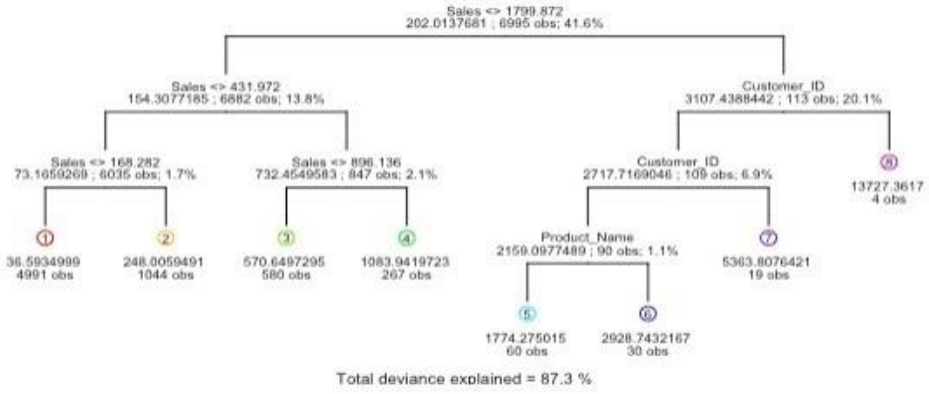


Fonte: Autoria própria

No entanto, apenas três se destacaram em termos de relevância, sendo elas: sales, customer ID, Product_Name.

A seguir está a representação da árvore de custos, a qual demonstrou uma notável capacidade de explicar os dados, atingindo 87,3% de capacidade de explicação dos dados, um resultado considerado altamente satisfatório.

Figura 33- Árvore de Custos



Fonte: Autoria própria

Essa abordagem resultou na criação de 8 ramos na árvore, permitindo-nos categorizar os produtos que compartilham características semelhantes, variando desde custos mais baixos até custos mais elevados. Portanto, em vez de lidar com a complexidade de 1850 produtos, agora podemos simplificar a análise trabalhando com apenas 8 grupos distintos de custos, o que torna a gestão financeira mais eficiente e estratégica.

Para avaliar o modelo de previsão do custo do produto, foram aplicadas as seguintes métricas de desempenho: RAE (Erro Absoluto Relativo), MASE (Erro Médio de Escala Absoluta Média), SMAPE (Erro Percentual Médio de Simetria Absoluta) e o índice de correlação. Tais medidas foram calculadas através do pacote 'metrics' no ambiente RStudio, conforme mostrado na Figura *34*.

Figura 34- Métricas de desempenho

```
# métricas de acurácia modelo custo

```{r }
 df.test$pred_custo <- predict(modelo_custo, newdata = df.test)

smape(df.test$Custo_produto, df.test$pred_custo)
cor(df.test$Custo_produto, df.test$pred_custo)
rae(df.test$Custo_produto , df.test$pred_custo)
smape(df.test$Custo_produto, df.test$pred_custo)

```

[1] 0.6969693
[1] 0.7961025
[1] 0.3420184
[1] 0.6969693
```

Fonte: Autoria própria

O modelo de árvore de regressão desenvolvido para prever os custos dos produtos mostrou um bom desempenho, conforme indicado pelas seguintes medidas de avaliação:

Tabela 6- Métricas aplicada a árvore de custo

COR	RAE	MASE	SMAPE
0.7961025	0.3420184	0.2720567	0.6969693

Fonte: Autoria própria

Essas métricas apontam para um modelo que é capaz de fazer previsões precisas e está bem alinhado com os dados reais. O RAE e o MASE são baixos, indicando que os erros de previsão são relativamente pequenos em relação à variabilidade dos dados. O SMAPE, embora um pouco mais elevado, ainda está em um nível razoável. Além disso, o índice de correlação de 0,796 sugere uma forte relação entre as previsões do modelo e os valores reais.

Com base nesses resultados, é válido afirmar que o modelo de árvore de regressão é eficaz na previsão de custos de produtos, o que pode ser um ativo valioso para tomar decisões relacionadas aos custos e estratégias de precificação.

Desta forma, a empresa pode trabalhar com apenas 8 grupos distintos de custos, direcionando seus recursos de forma mais eficiente, identificando áreas de oportunidade para redução de gastos e alocação estratégica

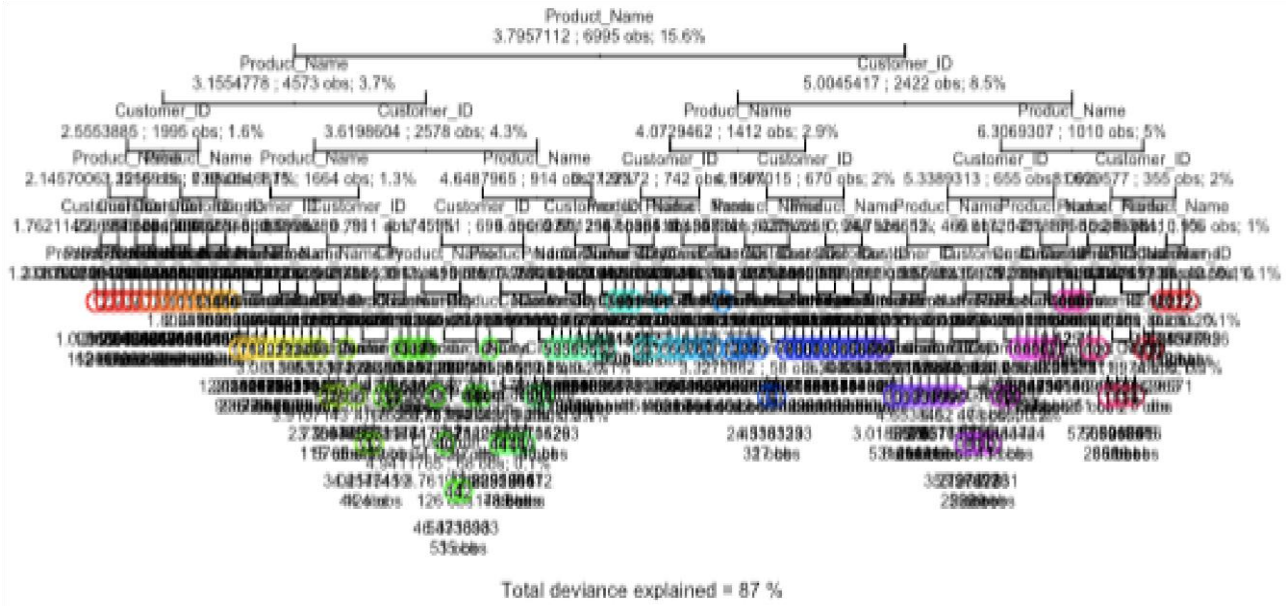
de investimentos. Além disso, a segmentação dos produtos permite uma abordagem personalizada para otimizar as estratégias de precificação e marketing, impulsionando o desempenho financeiro da empresa.

5.4.2 PREVISÃO DE DEMANDA

A previsão de demanda é crucial para a gestão de estoques e a eficiência da produção de uma empresa, pois permite que a organização se prepare para atender às futuras demandas dos clientes, evitando a falta de produtos em estoque ou o excesso de estoque obsoleto. Além disso, a previsão de demanda também desempenha um papel importante no planejamento estratégico, ajudando a empresa a tomar decisões informadas sobre investimentos, expansão e estratégias de preços.

Pensando nisso, o próximo passo é utilizar esses conhecimentos para analisar a demanda prevista para a base em análise. Nesse sentido, será criada uma árvore de previsão de demanda por meio do software RStudio, cujo resultado obtido está na Figura *35*.

Figura 35- Árvore de previsão de demanda

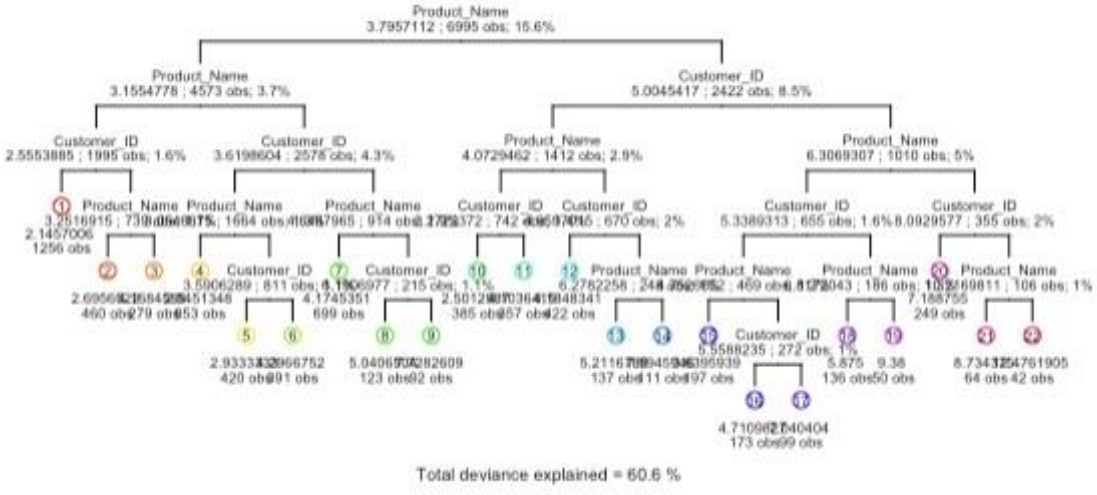


Fonte: Autoria própria

Conforme ilustrado, a árvore de decisão gerada exibiu um excesso de galhos, muitos dos quais não são relevantes para as análises em questão. Uma árvore de decisão com 112 galhos (nós) pode ser considerada muito complexa e suscetível a overfitting, a complexidade excessiva da árvore pode resultar em um modelo que se ajusta demasiadamente aos dados de treinamento, tornando-se menos capaz de generalizar para novos dados e, assim, prejudicando seu desempenho em previsões futuras.

Com o objetivo de simplificar a árvore, que originalmente apresentava 112 galhos, tornando-a mais interpretável e evitando o overfitting, optou-se por realizar a poda da árvore de decisão, utilizando o critério de poda de 0.009867173 sugerido pelo algoritmo do RStudio, resultando em uma árvore com 22 folhas, com capacidade de explicação dos dados de 60,6%, conforme a Figura *36*.

Figura 36- Árvore de previsão de demanda pós poda



Fonte: Autoria própria

Tabela 7- Métricas aplicada a árvore de demanda

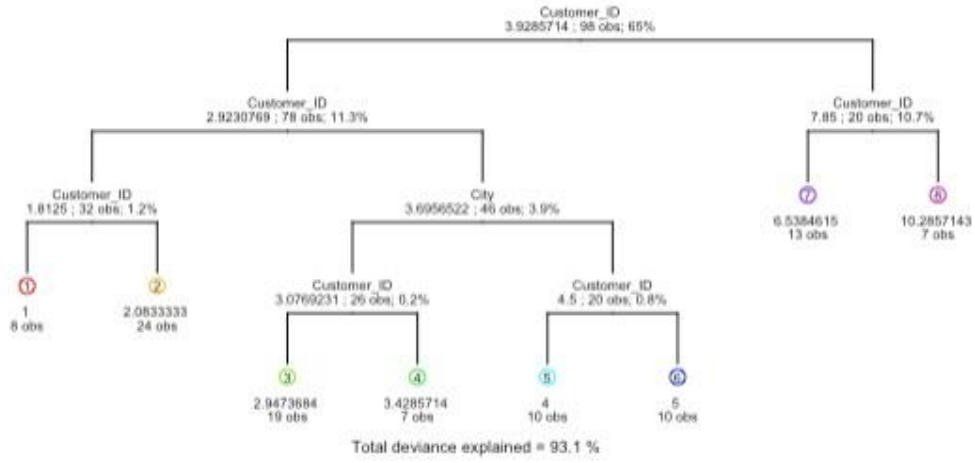
Modelo	COR	RAE	MASE	SMAPE
CP=0,001	0,029	1,265	0,943	0,565
CP=0,0098	-0,001	1,197	0,893	0,529

Fonte: Autoria própria

Os resultados iniciais indicam que o modelo teve um desempenho satisfatório, apesar de a correlação (COR) não apresentar uma relação muito forte. No entanto, é importante observar que lidamos com uma ampla diversidade de produtos em nosso conjunto de dados, o que pode afetar o desempenho do modelo em relação a diferentes produtos. Com isso, optou-se por criar uma árvore de previsão exclusiva para os três produtos mais vendidos: Staples, Staple_envelope e Easy_staple_paper. Essa estratégia visa melhorar a precisão das previsões, direcionando recursos para os produtos mais relevantes para a empresa.

A nova árvore de previsão, conforme a Figura *37*, levou em consideração informações relacionadas aos consumidores e às cidades onde ocorrem as vendas. Essa abordagem permitiu identificar padrões detalhados sobre onde esses produtos são mais vendidos e quem são os consumidores que os adquirem. A árvore resultante apresenta 8 folhas, cada uma representando um segmento distinto de mercado, e fornece insights para direcionar estratégias de marketing e operações em áreas específicas.

Figura 37- Árvore com os produtos: Staples, Staple_envelope e Easy_staple_paper



Fonte: Autoria própria

Com este direcionamento, após aplicar as métricas na base teste, obtivemos os resultados da tabela abaixo, que refletem uma melhoria significativa no desempenho do modelo:

Tabela 8- Métricas aplicada a árvore top 3 produtos

Modelo	COR	RAE	MASE	SMAPE
Top3_produto	0,505	0,856	0,563	0,345

Fonte: Autoria própria

Esses resultados destacam a eficácia da abordagem de concentração nos três produtos mais vendidos, pois as métricas de desempenho apresentaram melhorias substanciais em comparação com o modelo anterior. Isso fortalece a confiança na capacidade de prever a demanda desses produtos de forma precisa e alinhada com as necessidades do mercado.

Em resumo, a previsão da demanda ou do custo dos produtos é uma ferramenta poderosa e versátil para orientar as estratégias de negócios no varejo. É importante ressaltar que a capacidade de prever a demanda ou custo não se limita apenas aos produtos ou períodos específicos, mas é adaptável às necessidades e objetivos de cada cliente ou empresa. Através da aplicação de filtros, como período, estados e outros critérios relevantes, podemos ajustar e personalizar nossos modelos para fornecer previsões precisas e relevantes que impulsionam o sucesso de empresas.

6. CONSIDERAÇÕES FINAIS

O estudo do presente trabalho proporcionou uma jornada de aprendizado abrangente, indo desde a exploração dos modelos matemáticos até a análise detalhada das árvores de decisão, a familiarização com as ferramentas de software, a compreensão das medidas de acurácia e, finalmente, a aplicação prática desses conhecimentos em um banco de dados real. Ao longo desse percurso, destacou-se a capacidade das árvores de decisão como uma ferramenta fundamental, unindo teoria e prática, e exemplificou como a aquisição de competências analíticas e o domínio das tecnologias relevantes podem capacitar as empresas a tomar decisões estratégicas mais informadas e alcançar um sucesso duradouro no ambiente dinâmico do mercado.

Desta forma, percebeu-se a importância da análise de dados na tomada de decisões empresariais e na previsão da demanda no varejo. Nos últimos tempos, com fácil acesso aos dados e a crescente competição, modelos estatísticos, como a árvore de decisão, pode se destacar como uma ferramenta poderosa. Pois, a árvore de decisão possibilita identificar as variáveis mais importantes que impactam os resultados financeiros das empresas e representá-las visualmente de forma clara e compreensível.

E assim, as empresas podem tomar decisões embasadas para maximizar ganhos e minimizar perdas. Contudo, para que a análise de dados seja efetiva, é crucial realizar a limpeza adequada e transformação dos dados, definir objetivos claros e utilizar informações confiáveis de boa qualidade, pois os dados são a base do desenvolvimento do modelo.

A construção de modelos preditivos no varejo, tais como árvores de decisão, requer paciência, experimentação e ajustes contínuos. Explorar uma variedade de dados revela inúmeras opções e abordagens possíveis, sendo desafiador determinar quais são as relevantes para os objetivos da empresa, quais são capazes de explicar os dados, se são eficientes ao replicá-la em amostras aleatórias. Sendo importante, buscar iterar e aprimorar, para contribuir para modelos mais precisos, que se tornam valiosos para impulsionar o desempenho do setor varejista.

As árvores de decisão desempenham um papel crucial ao fornecer insights interpretáveis e acionáveis, mas é importante reconhecer que não há um único modelo que sirva para todas as situações. A adaptabilidade e a disposição em explorar diferentes variáveis, cenários e metodologias são essenciais para obter sucesso.

Ao longo desse processo, uma ferramenta extremamente útil no desenvolvimento do trabalho, foi o RStudio, uma plataforma de código aberto que facilita a análise de dados, modelagem estatística e visualização. Sua integração com diversas bibliotecas e interface amigável tornam a criação e o ajuste de modelos mais acessíveis.

No presente trabalho, foi apresentado um estudo de caso específico, explorando um banco de dados, desenvolvendo árvores de classificação para explorar os produtos por localização geográfica e árvore de regressão para prever a demanda e os custos, bem como avaliando a eficiência desses modelos, porém há ainda um vasto campo de possibilidades não exploradas. Poderiam ser realizadas análises adicionais, aplicando diferentes filtros e investigando outras abordagens, apresentando assim as diferentes oportunidades que a análise de dados e a modelagem estatística oferecem.

Desta forma, o uso de modelos estatísticos, como árvores de decisão, aliado à correta gestão de dados e à disposição para ajustes contínuos, pode ser essencial para o sucesso empresarial no mercado atual. Essa jornada pode ser desafiadora, mas também repleta de oportunidades para descobertas e melhorias, permitindo traduzir informações em estratégias sólidas no mercado de varejo em constante evolução.

REFERÊNCIAS

ARGE, Helton José. Modelos Matemáticos e Simulação Computacional em Engenharia Química. Rio de Janeiro: UFRJ, 2015. Disponível em: http://www2.peq.coppe.ufrj.br/Pessoal/Professores/Arge/COQ790/Cap_2.pdf. Acesso em: 09 mar. 2023.

BERTONE, Ana Maria Amarillo; BASSANEZI, Rodney Carlos; JAFELICE, Rosana Sueli da Motta. Modelagem Matemática. 2. ed. Uberlândia, MG: UFU, 2014. 352 p.

BIOINFO. Métricas de Avaliação em Machine Learning: Acurácia, Sensibilidade, Precisão, Especificidade e F-Score. Bioinfo, [S.l.], [20--]. Disponível em: https://bioinfo.com.br/metricas-de-avaliacao-em-machine-learning-acuraciasensibilidade-precisao-especificidade-e-f-score/. Acesso em: 30 mai. 2023.

BOYER, Carl B. História da matemática. Revisão: Uta C. Merzbach. Tradução: Elza F. Gomide. 2. ed. São Paulo: Edgard Blücher, 1996.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. Classification and Regression Trees. CRC Press, 1984.

Desenvolvimento do Varejo com o Avanço da Tecnologia. Disponível em: https://amis.org.br/plus/modulos/noticias/ler.php?cdnoticia=720. Acesso em: 08 de agosto de 2023.

DICIO. ACURÁCIA. In: DICIO, Dicionário Online de Português. Porto: 7Graus, 2020. Disponível em: https://www.dicio.com.br/acuracia/. Acesso em: 30 jan. 2023.

ECONOMY-PEDIA. Mathematical Model. Economy-Pedia, s.l. Disponível em: https://pt.economy-pedia.com/11030332-mathematical-model. Acesso em: 03 jan. 2023.

ENTROPIA, Ganho de informação e Decision trees. Disponível em: http://www.esalq.usp.br/lepse/imgs/conteudo_thumb/Entropia--Ganho-de-informa-o-e-Decision-trees.pdf. Acesso em: 11 de abril de 2023.

EQUIPE EDITORIAL DE CONCEITO.DE. Conceito de modelo matemático. Conceito.de. (28 de abril de 2012). Disponível em: https://conceito.de/modelomatematico. Acesso em: 27 mar. 2023.

Exame. 80% das compras em e-commerce são realizadas por mulheres. Bússola, [https://exame.com/bussola/80-das-compras-em-e-commerce-sao-realizadas-pormulheres/]. Acesso em: 15 de agosto de 2023.

FREITAS, Ana. Bioinformatics.ath.cx. Disponível em: http://web.tecnico.ulisboa.pt/ana.freitas/bioinformatics.ath.cx/bioinformatics.ath.cx/i ndexf23d.html?id=199. Acesso em: 11 de abril de 2023.

GATES, Bill. O software é um grande espírito criador. As pessoas usam o software para criar coisas incríveis e eu acredito que o software pode capacitar cada pessoa e cada organização a fazer mais e alcançar mais do que jamais imaginaram ser possível. In: FRASES FAMOSAS. Disponível em: https://www.frasesfamosas.com.br/frase/bill-gates-o-software-e-um-grande-espiritocriador/ Acesso em: 18 abr. 2023.

IBM. The FORTRAN Coding Sheet. Disponível em: https://www.ibm.com/ibm/history/ibm100/us/en/icons/fortrancodingsheet/. Acesso em: 04 abr. 2023.

IDWALL. O que é acurácia? [online]. [S.l.], 2019. Disponível em: https://blog.idwall.co/o-que-e-acuracia/. Acesso em: 26 abr. 2023.

iFood Blog Parceiros. Datas Comemorativas: Celebre conosco esses momentos especiais. Disponível em: <https://blog-parceiros.ifood.com.br/datascomemorativas/>\. Acesso em: 18 Agosto 2023.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. Journal of computational and graphical statistics, 5(3), 299-314.

Johnson, L., Adams Becker, S., Cummins, M., Estrada, V., Freeman, A., & Ludgate, H. (2020). NMC/CoSN Horizon Report: 2020 Higher Education Edition. Austin, TX: The New Media Consortium.

Johnson, M., & Santos, C. D. (2022). Modelos descritivos na análise de sistemas e tomada de decisões. Revista de Modelagem Matemática, 29(1), 23-35.

Johnson, M., & Silva, A. B. (2019). Modelos matemáticos como ferramenta para compreensão de padrões e tomada de decisões. Revista de Matemática Aplicada, 22(4), 67-80.

Johnson, S., & Smith, R. (2022). Técnicas de tratamento de dados para construção de modelos matemáticos. Revista de Matemática Aplicada, 25(1), 78-92.

LAURETTO, Marcelo S. Árvores de Decisão. São Paulo: Universidade de São Paulo
(USP), ano. Disponível em: https://edisciplinas.usp.br/pluginfile.php/4469825/mod_resource/content/1/ArvoresD ecisao_normalsize.pdf. Acesso em: 10 abr. 2023.

MARIANO, Diego. Métricas de avaliação em Machine Learning. Diego Mariano, [s.d.]. Disponível em: https://diegomariano.com/metricas-de-avaliacao-em-machinelearning/. Acesso em: 03 de agosto de 2023.

MARTINS, R. Modelos matemáticos. Londrina: UEL, 2006. Disponível em: http://www.uel.br/projetos/matessencial/superior/pdfs/modelos.pdf. Acesso em: 09 dez. 2022.
MICROSOFT. (2023). Power BI [Software]. Recuperado de https://powerbi.microsoft.com/

MITCHELL, T. M. Machine Learning. McGraw-Hill, 1997.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. Introduction to Linear Regression Analysis. 5th ed. John Wiley & Sons, 2012.

MORETTI, João Antonio. Aula de slides: MS428 - Análise de Séries Temporais. Campinas: Instituto de Matemática, Estatística e Computação Científica (IMECC), Universidade Estadual de Campinas (UNICAMP), 2º semestre de 2010. Disponível em: https://www.ime.unicamp.br/~moretti/ms428/2sem2010/aula_slides.pdf.
Acesso em: 10 abr. 2023.

MORETTI, Márcio. Introdução ao Cálculo Numérico. Slides da aula. Instituto de Matemática, Estatística e Computação Científica - Universidade Estadual de Campinas (UNICAMP). Campinas, SP, Brasil, 2010. Disponível em: https://www.ime.unicamp.br/~moretti/ms428/2sem2010/aula_slides.pdf.
Acesso em: 11 abr. 2023.

NUNES, Lucas. Fortran: um dinossauro das linguagens de programação volta depois de 10 anos. [online]. [S.l.], 04 de maio de 2021. Disponível em: https://www.ufsm.br/pet/sistemas-de-informacao/2021/05/04/fortran-umdinossauro-das-linguagens-de-programacao-volta-depois-de-10-anos.
Acesso em: 03 de agosto de 2023.

Oliveira, A. B., Santos, C. D., & Souza, E. F. (2022). Technostress: impactos do excesso de conectividade na sociedade atual. Revista de Estudos em Tecnologia e Comunicação, 15(2), 78-92.

QUINLAN, J. R. Entropia, ganho de informação e árvores de decisão. Piracicaba, SP: Escola Superior de Agricultura "Luiz de Queiroz", 1986. Disponível em: http://www.esalq.usp.br/lepse/imgs/conteudo_thumb/Entropia--Ganho-de-informa-o-e-Decision-trees.pdf. Acesso em: 12 dez. 2023.

RSTUDIO TEAM. (2023). RStudio: Integrated Development for R [Software]. Recuperado de https://www.rstudio.com/

Santos, A. B., & Oliveira, C. D. (2021). A aplicabilidade de modelos matemáticos na análise de fenômenos da vida cotidiana. Revista de Matemática Aplicada, 24(3), 4558.

Seu Futuro Vale Mais. Homens gastam 40% mais que mulheres em compras online. Seu Futuro Vale Mais, [https://www.seufuturovalemais.com.br/noticia/homensgastam-40-mais-que-mulheres-em-compras-online#:~:text=Segundo%20um%20estudo%20da%20Confedera%C3%A7%C3%A 3o,que%20mulheres%20em%20compras%20online.]. Acesso em: 15 de agosto de 2023.

SILVA, A. B. Evolução da gestão de pessoas. [apostila]. São Cristóvão: CESAD/UFS, 2012. Disponível em: https://cesad.ufs.br/ORBI/public/uploadCatalago/11534127032012Evolucao_Aula_ 10.pdf. Acesso em: 09 mar. 2023.

Smith, J., & Silva, A. B. (2022). Modelos prescritivos: abordagens exatas e aproximadas. Revista de Modelagem Matemática, 30(2), 45-59.

SOUSA, J. Por que analisar dados no corte de gastos da empresa? NetSource, [s.l.], 2021. Disponível em: https://www.netsource.com.br/por-que-analisar-dados-nocorte-de-gastos-da-empresa/. Acesso em: 09 fev. 2023.

STATISTA. Global software market revenue from 2016 to 2021 (in billion U.S. dollars). Disponível em: https://www.statista.com/statistics/442271/worldwidesoftware-market-revenue/. Acesso em: 18 abr. 2023.

Pagar.me. Produtos mais vendidos na internet. Blog Pagar.me, [s.d.]. Disponível em: https://pagar.me/blog/produtos-mais-vendidos-na-internet/. Acesso em: 18 ago. 2023.

Pereira, A. B., Silva, C. D., & Santos, E. F. (2021). Métodos para análise de sistemas complexos na tomada de decisões. Revista de Gestão e Tecnologia, 14(3), 87-102.

Polo TCF, Miot HA. Aplicações da curva ROC em estudos clínicos e experimentais. J Vasc Bras. 2020;19: e20200186. https://doi.org/10.1590/1677-5449.200186

PUCCINI, A. Técnicas de Mineração de Dados em um Ambiente de Aprendizado Colaborativo. Disponível em: https://www.maxwell.vrac.pucrio.br/7587/7587_4.PDF. Acesso em: 10 abr. 2023.

RUSSELL, S.; NORVIG, P. Inteligência Artificial. 3. ed. São Paulo: Editora Campus, 2016.

WikiHow. Prever o Futuro. Disponível em: https://pt.wikihow.com/Prever-o-Futuro. Acesso em: 26 fev. 2023.