UNIVERSIDADE FEDERAL FLUMINENSE

ENGENHARIA DE PRODUÇÃO

EMILLY CRISTINA FERREIRA NOGUEIRA

NICOLE DA SILVA FULGONI

ESTUDO DE CASO: SEGROB NOTLAD

RIO DAS OSTRAS
2025

SUMÁRIO

RESUMO

O presente estudo teve como objetivo realizar uma análise preditiva das vendas diárias de camisetas básicas masculinas de uma empresa do setor fast fashion, utilizando técnicas avançadas de aprendizado de máquina supervisionado e validação cruzada. A metodologia adotada foi o modelo CRISP-DM, que compreendeu desde o entendimento do negócio até a implementação de modelos preditivos. As métricas utilizadas para avaliar o desempenho dos modelos foram o Erro Percentual Absoluto Médio (MAPE), a Raiz do Erro Quadrático Médio (RMSE) e o Erro Médio Absoluto (MAE). Entre os modelos aplicados destacaram-se a Regressão Linear Múltipla, K-Nearest Neighbors (KNN), Random Forest e Support Vector Machine (SVM). Os resultados apontaram uma tendência consistente de crescimento das vendas, destacando-se particularmente as previsões para dezembro de 2024. O modelo de Regressão Linear apresentou melhor desempenho, destacando-se por sua precisão e confiabilidade. Observou-se ainda a importância de revisar continuamente os modelos utilizados para garantir previsões mais precisas, especialmente diante de mudanças significativas no comportamento dos consumidores.

Palavras-chave: Análise preditiva. Modelos de aprendizado supervisionado. Validação cruzada. Métricas de erro.

1. INTRODUÇÃO

Em um mercado competitivo como o fast fashion, a capacidade de prever com precisão a demanda por produtos é vital para a gestão eficiente dos recursos logísticos, financeiros e humanos. Dessa forma, a aplicação de técnicas de análise preditiva surge como uma ferramenta estratégica essencial para empresas que buscam aumentar sua eficiência operacional e competitividade no mercado. Este estudo aborda especificamente a previsão das vendas diárias de camisetas básicas masculinas, buscando fornecer à empresa subsídios sólidos para decisões assertivas e fundamentadas.

A relevância do tema justifica-se pela necessidade crescente das organizações em otimizar estoques, reduzir custos operacionais e melhorar o planejamento das operações, garantindo a satisfação dos clientes e maximizando os resultados financeiros. Além disso, a aplicação de modelos preditivos avançados contribui diretamente para o entendimento detalhado das tendências e comportamento do mercado consumidor, permitindo ajustes rápidos e eficazes em estratégias comerciais e promocionais.

O objetivo central deste estudo é analisar a eficácia de diversos modelos preditivos no contexto das vendas diárias, destacando-se a comparação entre técnicas tradicionais e avançadas, como Regressão Linear, KNN, Random Forest e SVM. Adicionalmente, serão discutidas métricas específicas que permitem avaliar o desempenho dos modelos, oferecendo uma base clara para selecionar a melhor abordagem em diferentes cenários operacionais.

O trabalho está estruturado em referencial teórico detalhado sobre métricas, validações cruzadas e modelos de aprendizado supervisionado. Posteriormente, é apresentado um estudo de caso específico da empresa Segrob Notlad, seguido de análises detalhadas, resultados obtidos, avaliações comparativas dos modelos e conclusões que proporcionam insights práticos e estratégicos para decisões gerenciais eficazes.

2. REFERENCIAL TEÓRICO

2.1. Métricas e erros em Modelos preditivos

Avaliar o desempenho de modelos preditivos é fundamental na ciência e mineração de dados, pois permite entender o quanto as previsões se aproximam da realidade (CAMPOS; SILVA, 2019). Segundo Silva (2023), compreender os erros que os modelos cometem facilita ajustes importantes para aumentar sua precisão e confiança. Entre as métricas mais comuns estão o Erro Médio Absoluto (MAE), que mede a média das diferenças absolutas entre os valores previstos e reais e é menos afetado por valores extremos, e a Raiz do Erro Quadrático Médio (RMSE), que penaliza erros maiores com mais intensidade (MORETTIN; TOLOI, 2018). Já o Erro Percentual Absoluto Médio (MAPE) expressa o erro em porcentagem, facilitando a interpretação prática, especialmente em contextos comerciais e financeiros.

Por isso, escolher corretamente essas métricas, considerando as características dos dados e os objetivos da análise, é essencial para resultados consistentes e decisões assertivas.

2.1.1. Erro Percentual Absoluto Médio

O Erro Percentual Absoluto Médio (MAPE, *Mean Absolute Percentage Error*) é uma métrica usada para avaliar a precisão dos modelos preditivos, especialmente em séries temporais e previsão financeira. Sua principal característica é expressar o erro médio das previsões em termos percentuais, facilitando a compreensão prática dos resultados (MORETTIN; TOLOI, 2018).

Matematicamente, o MAPE pode ser calculado pela seguinte fórmula (LIMA FILHO et al., 2012 apud GOUVEIA et al., 2015, p. 591.):

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \widehat{Y}_i}{Y_i} \right| X\ 100$$

Onde:

$Y_i = Valores\ observados$

$\widehat{Y}_i = Valor\ ajustados$

$n = Número\ total\ de\ observações$

A interpretação prática do MAPE é direta: um valor de 8% indica que, em média, as previsões do modelo estão 8% afastadas dos valores reais. Isso torna o MAPE útil em contextos empresariais, como na previsão de vendas e demanda, onde a comunicação clara dos resultados é essencial (MARIO FILHO, 2022).

Segundo Silva (2025), o MAPE apresenta vantagens práticas que o tornam útil nas análises preditivas, como facilidade de interpretação, já que seu erro é expresso em porcentagem, permitindo compreensão inclusive por leitores menos técnicos. Também possibilita comparar diretamente previsões de séries temporais distintas sem preocupação com a escala dos dados, além de ser amplamente reconhecido na literatura especializada. Contudo, o autor destaca algumas limitações importantes, como valores extremamente altos ou indefinidos quando os valores reais estão próximos de zero, tendência em penalizar mais fortemente erros por superestimação, o que pode gerar avaliações enviesadas, e inadequação para séries com valores negativos. Assim, Silva (2025) reforça a importância de considerar cuidadosamente o contexto específico da análise para definir se o MAPE é a métrica mais apropriada ou se outras alternativas seriam mais adequadas.

2.1.2. RMSE – Raiz do Erro Quadrático Médio

A Raiz do Erro Quadrático Médio (RMSE, *Root Mean Squared Error*) é uma métrica amplamente utilizada para avaliar a precisão de modelos preditivos, especialmente em regressões e séries temporais. Ela quantifica a diferença entre os valores previstos pelo modelo e os valores observados, penalizando mais fortemente os grandes erros devido à elevação ao quadrado das diferenças (SILVA, 2023).

De acordo com Mario Filho (2023), a fórmula matemática do RMSE é expressa por:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}$$

Onde:

$Y_i = Valor\ real\ observado$

$\widehat{Y}_i = Valor\ previsto\ pelo\ modelo$

$n = Número\ total\ de\ observações$

Um RMSE de 10 indica que, em média, as previsões realizadas pelo modelo estão desviadas em 10 unidades dos valores observados. Essa métrica é especialmente adequada para contextos em que grandes erros devem ser evitados, como previsões financeiras e cenários de alta exigência em precisão (CAMPOS; SILVA, 2019). Comparativamente, o RMSE tem semelhanças com o desvio padrão, já que ambas as medidas refletem dispersão. Entretanto, enquanto o desvio padrão mede a variabilidade em torno da média dos dados reais, o RMSE foca especificamente nas previsões e na magnitude dos erros do modelo (MORETTIN; TOLOI, 2018).

2.1.3. MAE (Erro Absoluto Médio)

O Erro Médio Absoluto (MAE, Mean Absolute Error) é uma métrica amplamente utilizada na avaliação da precisão de modelos preditivos. Segundo Chai e Draxler (2014), o MAE quantifica a média das diferenças absolutas entre os valores observados e os valores previstos por um modelo, oferecendo uma interpretação direta do desempenho preditivo por estar na mesma unidade dos dados analisados. Uma das principais vantagens dessa métrica é sua robustez frente a valores extremos, já que não penaliza erros maiores de forma quadrática, como ocorre no RMSE.

A fórmula matemática do MAE pode ser expressa por:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{Y}_i\right|$$

Onde:

$Y_i = Valor\ real\ observado$

$\widehat{Y}_i = Valor\ previsto\ pelo\ modelo$

$n = Número\ total\ de\ observações$

A interpretação prática do MAE é simples e intuitiva: um valor de MAE igual a 5 indica que, em média, as previsões feitas pelo modelo desviam-se em 5 unidades dos valores reais observados (CHAI; DRAXLER, 2014). Assim, o MAE é especialmente útil em contextos nos quais é necessário uma métrica facilmente compreensível e interpretável por profissionais de diferentes áreas técnicas e gerenciais, como na previsão de demanda, no monitoramento de processos produtivos e em estudos econômicos e financeiros.

2.1.4. Erro Padrão

O erro padrão é uma medida estatística que expressa a precisão das estimativas obtidas por um modelo, indicando a variabilidade esperada caso a análise seja repetida diversas vezes (MORETTIN; TOLOI, 2018). Diferentemente das métricas de erro absoluto, o erro padrão avalia especificamente a incerteza associada à média das previsões, sendo frequentemente utilizado em testes de hipóteses e intervalos de confiança.

Segundo Bussab e Morettin (2017), a fórmula matemática geral do erro padrão da média pode ser representada como:

$$EP = \frac{s}{\sqrt{n}}$$

Onde:

$s = Desvio\ padrão\ das\ observações$

$n = Tamanho\ da\ amostra$

Na prática, um erro padrão pequeno sugere que as estimativas são consistentes e confiáveis, enquanto valores elevados indicam maior incerteza e menor confiabilidade nos resultados obtidos (BUSSAB; MORETTIN, 2017).

Entre suas principais aplicações estão a validação de modelos estatísticos e econométricos, análise financeira e pesquisas acadêmicas, contextos em que a

precisão das estimativas é fundamental para conclusões seguras e decisões bem fundamentadas (MORETTIN; TOLOI, 2018).

2.1.5. Comparação entre métricas

A escolha adequada das métricas para avaliar modelos preditivos é essencial e depende diretamente do objetivo específico de cada projeto (SILVA, 2023). Entre as métricas mais utilizadas, destacam-se:

- MAPE: bastante intuitiva, pois apresenta o erro em porcentagem, facilitando a interpretação, especialmente em contextos comerciais e financeiros (SILVA, 2025). Contudo, pode gerar distorções quando os valores reais são muito baixos.
- RMSE: ideal quando erros maiores precisam ser evitados, pois penaliza mais fortemente grandes desvios, embora possa exagerar o impacto desses erros (MORETTIN; TOLOI, 2018).
- MAE: é robusta e simples, adequada para dados com valores extremos, porém não penaliza tanto erros elevados, oferecendo uma interpretação direta do desempenho do modelo por estar na mesma unidade dos dados analisados (CHAI; DRAXLER, 2014).
- Erro Padrão: mede a precisão das estimativas médias, sendo especialmente relevante em análises acadêmicas e intervalos de confiança, apesar de não ser usada frequentemente para previsões pontuais (BUSSAB; MORETTIN, 2017).

Portanto, conhecer as particularidades, vantagens e limitações de cada métrica é fundamental para selecionar a abordagem mais adequada, aplicar técnicas corretivas eficientes e assegurar previsões confiáveis e decisões assertivas.

2.2. Formas de validação cruzada

A validação cruzada (*cross-validation*) é uma técnica estatística utilizada para medir a capacidade dos modelos preditivos em realizar previsões confiáveis em novos dados, evitando o problema do superajuste (*overfitting*) (CAMPOS; SILVA, 2019). Por meio dela, os dados são divididos diversas vezes em subconjuntos

diferentes de treinamento e teste, garantindo uma estimativa mais robusta da performance real do modelo.

Segundo Silva (2023), entre as formas mais comuns estão o método K-fold, Leave-One-Out, Holdout e a validação cruzada estratificada, cada uma adequada a diferentes contextos. A escolha do método depende principalmente das características dos dados e dos objetivos específicos do projeto de análise preditiva.

2.2.1. Validação Cruzada K-fold

A validação cruzada K-fold é uma técnica muito utilizada para avaliar a precisão de modelos preditivos. Nessa abordagem, os dados são divididos em *K* partes (folds), geralmente de 5 a 10, usando alternadamente cada parte como conjunto de teste, enquanto as demais servem para treinamento (MORETTIN; TOLOI, 2018).

Essa técnica é vantajosa por gerar uma estimativa realista do desempenho do modelo, já que todas as observações são aproveitadas tanto no treinamento quanto na validação (SILVA, 2023). Ao final, o desempenho é calculado pela média dos resultados obtidos em cada fold, reduzindo a variabilidade das estimativas. Por essa razão, é muito aplicada em contextos onde robustez e precisão são essenciais, como em previsões financeiras, marketing e pesquisas científicas (CAMPOS; SILVA, 2019).

2.2.2. Validação Cruzada Leave-One-Out (LOOCV)

A validação cruzada Leave-One-Out (LOOCV) é uma técnica específica do método K-fold, em que o número de subconjuntos (folds) corresponde ao número total de observações. Nesse método, cada observação é usada individualmente como teste, enquanto as demais servem para treinamento, repetindo-se o processo para todas as observações disponíveis (SILVA, 2023).

Segundo Morettin e Toloi (2018), a vantagem principal da LOOCV é sua alta precisão e baixa variabilidade, ideal para contextos em que os dados são limitados. Entretanto, seu principal ponto negativo é o alto custo computacional, o que pode dificultar sua aplicação em grandes bases de dados ou em modelos complexos. Por isso, é comumente utilizada em contextos acadêmicos e científicos, onde a confiabilidade das estimativas é essencial (CAMPOS; SILVA, 2019).

2.2.3. Validação Cruzada Holdout

A validação Holdout é um método simples para avaliar o desempenho de modelos preditivos. Consiste em dividir o conjunto de dados original em dois grupos: treinamento (geralmente entre 70% e 80%) e teste (entre 20% e 30%), realizando uma única avaliação do modelo com essa divisão (CAMPOS; SILVA, 2019).

Sua principal vantagem é a facilidade de implementação e baixo custo computacional, sendo adequada para grandes bases de dados ou análises rápidas. No entanto, Silva (2023) alerta que uma desvantagem significativa é a alta variabilidade nos resultados, especialmente quando os dados são limitados. Essa técnica é recomendada para situações que priorizam rapidez e simplicidade, como testes preliminares e prototipagem de modelos (MORETTIN; TOLOI, 2018).

2.2.4. Validação Cruzada Estratificada

A validação cruzada estratificada é uma variação do método K-fold usada principalmente em problemas de classificação. Sua principal característica é garantir que a proporção das classes originais seja preservada em cada subconjunto (fold), evitando distorções e garantindo resultados mais precisos (SILVA, 2023).

Segundo Campos e Silva (2019), a vantagem dessa técnica é manter a distribuição original dos dados, essencial em cenários com classes desbalanceadas. Assim, ela produz estimativas mais confiáveis e realistas sobre o desempenho dos modelos. Por isso, é amplamente utilizada em áreas como análise de crédito, detecção de fraudes e diagnósticos médicos, onde uma avaliação precisa e equilibrada é crucial (MORETTIN; TOLOI, 2018).

2.2.5. Validação Cruzada Sliding Window

De acordo com Frieyadie e Setiyorini (2024), o método de Sliding Window é utilizado na etapa inicial de tratamento dos dados, com a finalidade de reorganizá-los em tarefas de classificação com base em séries temporais. Essa abordagem permite que o modelo seja ajustado dinamicamente à medida que novos dados se tornam disponíveis, tornando possível capturar alterações recentes nas tendências de capacidade e, assim, alinhar gradualmente as previsões com os valores reais.

O funcionamento da janela consiste em seu deslocamento ao longo da sequência de dados, analisando padrões de forma aleatória. A posição escolhida para análise é chamada de suporte do padrão. Um dos principais obstáculos dessa técnica é o processo de remoção dos dados antigos que já não contribuem para o modelo atual. (FRIEYADIE; SETIYORINI, 2024).

2.3. Modelos de Aprendizado de Máquina Supervisionado

O aprendizado de máquina supervisionado configura-se como uma das modalidades mais consolidadas e amplamente aplicadas da aprendizagem de máquina, sendo parte integrante da Inteligência Artificial. Nesse modelo, a máquina aprende a partir de um conjunto de dados rotulado, ou seja, um conjunto de exemplos onde as entradas (variáveis independentes) estão associadas às saídas desejadas (variáveis dependentes), permitindo que o sistema desenvolva a capacidade de realizar predições ou classificações com base em novos dados (FILHO, 2023).

Ainda segundo Filho (2023), a etapa de aprendizado supervisionado envolve diferentes algoritmos, como a regressão linear, a regressão logística, as máquinas de vetor de suporte (SVM), as árvores de decisão e as redes neurais artificiais. Cada um desses métodos possui características específicas quanto à capacidade de interpretação, complexidade computacional e robustez frente a dados ruidosos.

2.3.1. Regressão Linear

Segundo Maroco (2003 apud. Rodrigues, 2012, p. 17):

> O termo “Análise de Regressão” define um conjunto vasto de técnicas estatísticas usadas para modelar relações entre variáveis e predizer o valor de uma ou mais variáveis dependentes (ou de resposta) a partir de um conjunto de variáveis independentes (ou predictoras).

O modelo de regressão linear é uma ferramenta estatística amplamente utilizada para descrever, prever e inferir relações entre variáveis quantitativas. Trata-se de um modelo matemático que visa representar a relação entre uma variável dependente Y, e uma ou mais variáveis independentes X, através de uma equação linear (MONTGOMERY; PECK; VINING, 2012).

Para Matos (1995), o objetivo da regressão pode ser de natureza explicativa, quando se busca demonstrar uma relação matemática entre as variáveis, a qual pode sugerir, mas não comprovar, uma possível relação de causa e efeito. Alternativamente, o objetivo pode ser preditivo, visando estabelecer uma relação que possibilite, a partir de observações futuras das variáveis independentes X, estimar o valor correspondente da variável dependente Y, sem a necessidade de medi-la diretamente.

2.3.1.1. Regressão Linear Simples

O modelo de regressão linear simples constitui uma das abordagens mais elementares e, ao mesmo tempo, mais relevantes da análise estatística aplicada. Seu propósito é investigar e quantificar a relação apenas entre duas variáveis quantitativas. A ideia central consiste em verificar se há uma relação sistemática entre as duas variáveis, ou seja, se os valores de Y podem ser explicados, ao menos em parte, pelos valores de X. (CHEIN, 2019)

A estrutura formal do modelo é expressa por meio da seguinte equação (CHEIN, 2019):

$$Y = \beta_0 + \beta_1 X + u$$

Onde:

β_0 = Intercepto da reta

β_1 = Coeficiente angular

u = Erro aleatório ou distúrbio

A estimação dos parâmetros β_0 e β_1 é geralmente realizada por meio do método dos mínimos quadrados ordinários (MQO). Esse método busca minimizar a soma dos quadrados dos resíduos, as diferenças entre os valores observados e os valores previstos pela reta de regressão.

Segundo Stock e Watson (2010, apud Chein, 2019), o MQO escolhe os coeficientes da reta de forma que ela fique o mais próxima possível dos dados. As fórmulas dos estimadores de MQO são:

$$\hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \text{ e } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Conforme destacado por Fávero e Belfiore (2017), mesmo no modelo de regressão linear simples, a validação é fundamental para garantir a qualidade das estimativas, exigindo a verificação de pressupostos básicos como linearidade, ausência de autocorrelação e homocedasticidade dos resíduos. Os autores ressaltam que falhas nesses pressupostos podem comprometer significativamente a confiabilidade das conclusões obtidas. Guedes et al. (2018) complementam destacando que, na regressão linear simples, a correta interpretação dos coeficientes e uma análise detalhada dos resíduos são essenciais para evitar erros interpretativos e garantir resultados robustos e aplicáveis à tomada de decisões.

2.3.1.2. Regressão Linear Múltipla

Chein (2019) define a regressão linear múltipla como um método estatístico que estabelece relações entre uma variável dependente e várias variáveis independentes simultaneamente. Matos (1995) reforça que esse método, conhecido como "multi-regressão", permite captar interações complexas entre diversas variáveis. Fávero e Belfiore (2017) destacam a importância desse modelo na previsão e explicação de fenômenos, enfatizando a necessidade de validar seus pressupostos básicos para garantir resultados confiáveis. Esses pressupostos incluem linearidade dos parâmetros, ausência de multicolinearidade perfeita, homocedasticidade e normalidade dos resíduos (CHEIN, 2019). Guedes et al. (2018) acrescentam que, além da escolha adequada das variáveis, é essencial realizar uma análise detalhada dos resíduos para obter resultados robustos.

A fórmula geral da regressão linear múltipla é expressa por (Ferraz, 2024):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n + \varepsilon$$

Onde:

Y = Variável dependente;

β_0 = Intercepto;

β_1, β_2, …. β_n são coeficientes das variáveis independentes X_1 , X_2 ,...., X_{ni} ;

ε = termo de erro.

Chein (2019) conclui ressaltando que a aplicação correta da regressão linear múltipla possibilita decisões mais informadas em áreas como economia, administração, engenharia e saúde pública.

2.3.1.3. Correlação

A análise de correlação visa mensurar o grau de associação entre duas variáveis quantitativas, X e Y, ou seja, busca identificar a intensidade e a direção do relacionamento linear entre elas. Para quantificar essa relação, utiliza-se o coeficiente de correlação linear de Pearson, que expressa numericamente o nível de dependência linear existente entre as variáveis analisadas (RODRIGUES, 2012).

O coeficiente de correlação linear de Pearson entre duas variáveis quantitativas, X e Y, é dado por:

$$R_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Onde:

$$\bar{x} = \sum_{i=1}^{n}\frac{x_i}{n}$$

$$\bar{y} = \sum_{i=1}^{n}\frac{y_i}{n}$$

A partir desse coeficiente, pode-se tirar conclusões sobre a direção e intensidade da relação entre as variáveis. A tabela 1 apresenta a interpretação do tipo de correlação conforme o resultado de R_{xy}.

Tabela 1 - Interpretação do coeficiente de correlação de Pearson. Fonte: Rodrigues, 2012.

Coeficiente de Correlação	Tipo de Correlação
$R_{xy} = 1$	Perfeita positiva
$0,8 \leq R_{xy} < 1$	Forte positiva
$0,5 \leq R_{xy} < 0,8$	Moderada positiva
$0,1 \leq R_{xy} < 0,5$	Fraca positiva
$0 \leq R_{xy} < 0,1$	Ínfima positiva
0	Nula
$-0,1 \leq R_{xy} < 0$	Ínfima negativa
$-0,5 \leq R_{xy} < -0,1$	Fraca negativa
$-0,8 \leq R_{xy} < -0,5$	Moderada negativa
$-1 \leq R_{xy} < -0,8$	Forte negativa
$R_{xy} = -1$	Perfeita negativa

De acordo com Rodrigues (2012), para analisar a relação entre duas variáveis, X e Y, é possível representar seus valores por meio de um gráfico de dispersão. Caso os pontos plotados nesse gráfico se alinhem ou se distribuam de forma próxima a uma linha reta, pode-se inferir a existência de uma relação linear entre as variáveis.

Além disso, a força da correlação entre X e Y pode ser avaliada visualmente a partir da dispersão dos pontos: quanto mais próximos estiverem de uma linha reta, mais forte tende a ser a correlação entre as variáveis. A Figura 1 apresenta alguns exemplos de classificação da correlação através do diagrama de dispersão.

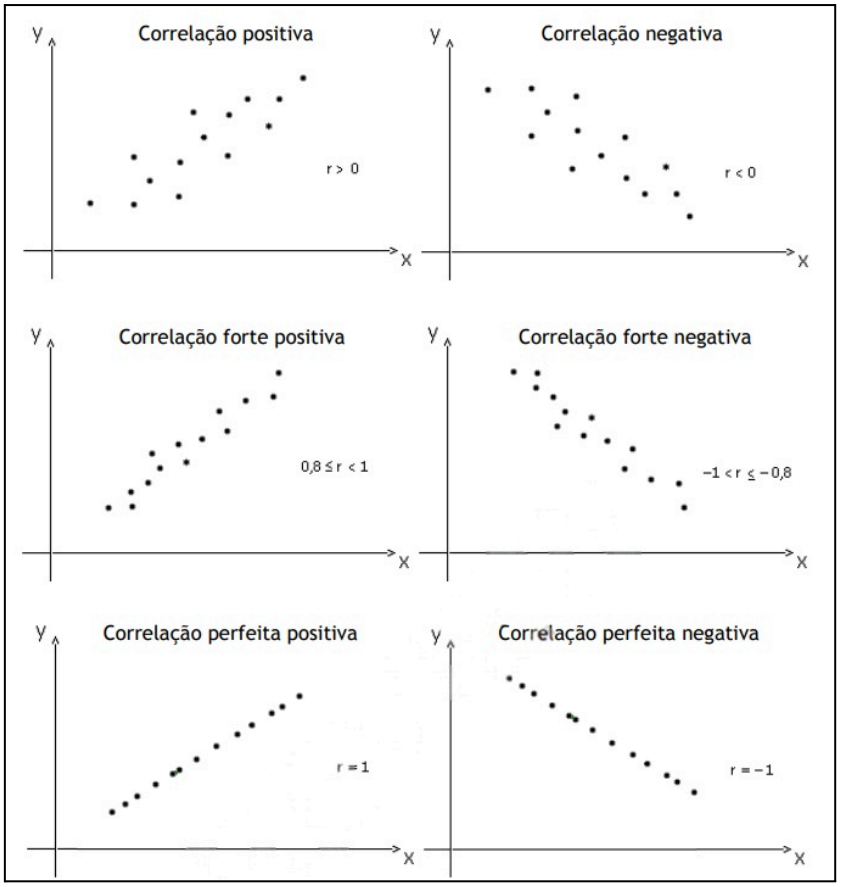

Figura 1 - Tipos de correlações através do gráfico de dispersão. Fonte: Santos (2007, apud Rodrigues, 2012)

2.3.2. KNN

O KNN (abreviação de *K-Nearest Neighbors*) é uma técnica de aprendizado supervisionado baseada em instâncias, que se destaca por sua simplicidade e eficácia em diversas tarefas de classificação, regressão e previsão de séries temporais. Em sua forma mais comum, o algoritmo atribui a um novo exemplo a classe mais comum entre seus k vizinhos mais próximos, calculados a partir de uma medida de similaridade ou distância (FERRERO, 2009).

Na versão tradicional do KNN, a proposta era ter uma alternativa de fácil implementação, que não requer um processo de treinamento propriamente dito, mas sim o armazenamento dos dados históricos. A partir da chegada de uma nova amostra, o algoritmo compara sua estrutura com os dados armazenados e determina os k exemplos mais similares, usando métricas como a distância Euclidiana, a distância de Manhattan ou medidas mais complexas adaptadas a séries temporais (FERRERO, 2009).

Segundo o Portal Data Science (2024), dentre as métricas mencionadas por Ferrero (2009), a distância euclidiana é a mais comum para variáveis numéricas. Essa distância é calculada pela seguinte equação:

$$Distância\ Euclidiana\ (x,\ xi)\ =\ \sqrt{\sum_{j}(xi\ -\ xij)^2}$$

Onde x é o ponto que está sendo analisado, xi são os pontos já existentes no conjunto de treinamento, e j representa cada atributo considerado.

Na dissertação de Ferrero (2009), o autor estende o uso do KNN para a tarefa de previsão de séries temporais, resultando no modelo denominado KNN-TSP (*Time Series Prediction*). Nessa abordagem, a previsão de um valor futuro se dá pela identificação de k subsequências mais semelhantes à sequência de entrada dentro de uma série histórica. A partir dos valores futuros dessas subsequências similares, realiza-se a estimativa do próximo valor da série.

Para Ferrero (2009), existem dois aspectos fundamentais: O critério de seleção dos vizinhos mais próximos e a função de previsão. No primeiro, propõe-se considerar não apenas a similaridade entre sequências, mas também a distância temporal, de forma a priorizar subsequências mais recentes, o que melhora a acurácia em séries com comportamento dinâmico. No segundo aspecto, são introduzidas funções de previsão que mantêm bom desempenho mesmo em séries com padrões em diferentes níveis, como a função de Média de Valores Relativos (MVR). A função é definida pela equação:

$$f_{MVR}(S')\ =\ x_n + \frac{\sum_{i=1}^{k} \Delta s'_{i,\,w+1}}{k} = \hat{x}_{n+1}$$

Onde,

$$\Delta s'_{i,\,w+1} = s'_{i,\,w+1} - s'_{i,\,w}$$

Outro ponto destacado por Ferrero (2009) é a sensibilidade do kNN à escolha do parâmetro k, que determina quantos vizinhos devem ser considerados. Valores pequenos de k tendem a produzir previsões mais ruidosas, enquanto valores altos podem generalizar demais e suavizar excessivamente os dados. A seleção ideal de

k depende da natureza da série e deve ser feita empiricamente por meio de validação cruzada.

2.3.3. Árvore de Decisão

As árvores de decisão são métodos eficazes e intuitivos na descoberta de conhecimento em bases de dados na área da saúde. As árvores de decisão são estruturas hierárquicas que possibilitam a classificação de casos com base em atributos previamente definidos, facilitando o entendimento e interpretação dos resultados obtidos (GARCIA, 2003).

De maneira geral, as árvores de decisão se fundamentam na abordagem "dividir para conquistar", onde os dados são segmentados em subconjuntos progressivamente menores, cada um representando características semelhantes. Esses subconjuntos, conhecidos como nós, conduzem à classificação dos dados até que uma folha, ou seja, uma classe definitiva, seja identificada (GARCIA, 2003).

A estrutura das árvores de decisão pode ser visualizada na figura 2 a seguir.

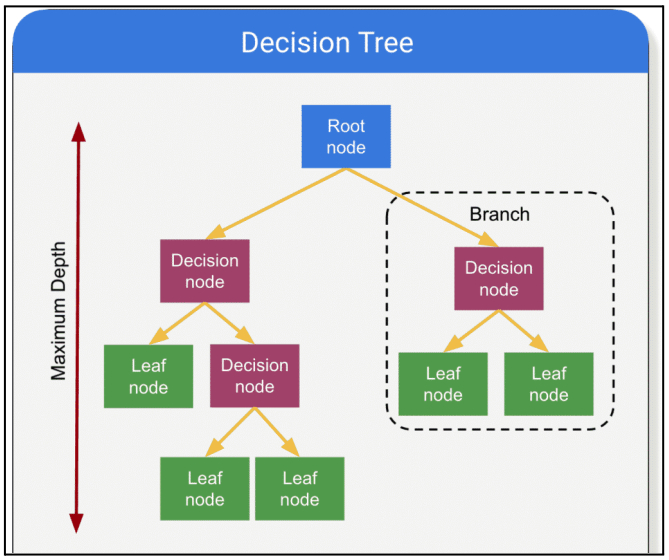

Figura 2 – Estrutura de Árvore de Decisão. Fonte: Adaptado de Giro et al. (2022)

Os atributos que compõem os testes das árvores podem ser tanto quantitativos quanto categóricos. Os atributos quantitativos demandam testes específicos que definem intervalos numéricos, enquanto os categóricos podem ser abordados criando-se ramos distintos para cada valor ou agrupando valores similares em subgrupos (GARCIA, 2003).

A construção das árvores de decisão ocorre por meio da seleção dos

atributos mais discriminantes, baseando-se em critérios estatísticos como o Ganho de Informação, o índice de Gini ou outros métodos equivalentes, que identificam os atributos mais relevantes para o particionamento eficaz dos dados (GARCIA, 2003).

Outro aspecto relevante do uso de árvores de decisão é a técnica de poda, que visa evitar o crescimento excessivo das árvores, eliminando partes que não contribuem significativamente para a classificação correta, mantendo a eficiência computacional e aumentando a generalização dos resultados obtidos (GARCIA, 2003).

Segundo POVILL (2022), o Índice de Gini, usado para medir a pureza dos nós em árvores de decisão, é calculado da seguinte forma:

$$Gini = 1 - \sum_{i=1}^{C} p(i)^2$$

Ainda segundo o autor, a construção de árvores de decisão frequentemente utiliza a entropia para determinar a qualidade da divisão dos dados em nós. A entropia mede o grau de incerteza ou impureza em um conjunto e é calculada por meio da seguinte fórmula:

$$Entropia\,(S) = \sum_{i} p(i)\,.\,log_2\,.\,p(i)$$

Onde p (i) representa a proporção das observações pertencentes à classe iii. Valores menores da entropia indicam conjuntos mais puros ou homogêneos, tornando essa métrica essencial para identificar divisões mais eficazes dos dados em árvores de decisão.

2.3.4 Random Forest

Random Forest é uma técnica avançada de aprendizado supervisionado baseada na combinação de múltiplas árvores de decisão. Desenvolvida por Breiman (2001), esta abordagem melhora significativamente a precisão ao utilizar árvores de decisão geradas por meio de amostras bootstrap, selecionadas aleatoriamente e com reposição do conjunto original de treinamento. Cada árvore individual (Random Tree) utiliza uma seleção aleatória de atributos, sendo o número destes atributos (m) fixo e menor ou igual ao total de atributos disponíveis ($m \leq a$).

Random Forest difere do método Bagging tradicional principalmente na

seleção dos atributos durante a construção das árvores. Enquanto no Bagging todas as variáveis são consideradas em cada nó, na Random Forest apenas um subconjunto aleatório das variáveis é utilizado. Essa estratégia reduz a correlação entre as árvores, melhorando diretamente o desempenho geral do modelo (BREIMAN, 2001; OSHIRO, 2013).

É definido formalmente como um classificador composto por uma coleção de árvores $\{h_k(x)\}$, onde k = 1,2,...,L, e cada T_k ,representa amostras aleatórias independentes e identicamente distribuídas. Cada árvore individual no modelo vota na classe mais popular para classificar uma nova entrada (BREIMAN, 2001 apud OSHIRO, 2013).

O desempenho de uma Random Forest depende da força das árvores individuais e da baixa correlação entre elas. Duas estratégias de aleatoriedade, Bagging e seleção aleatória dos atributos, garantem diversidade entre as árvores, reduzindo a correlação e, consequentemente, diminuindo o erro geral de classificação. Além disso, árvores individuais com baixa taxa de erro (fortes) contribuem significativamente para aumentar a precisão global do modelo (BREIMAN, 2001; BREIMAN; CUTLER, 2004; BREIMAN, 2004; MA; GUO; CUKIC, 2007 apud OSHIRO, 2013).

2.3.5. SVM

O método Support Vector Machine (SVM) é uma técnica de aprendizado supervisionado amplamente utilizada para classificação e regressão. Proposto inicialmente por Vapnik (1999), baseia-se na teoria estatística de aprendizado, oferecendo vantagens significativas em termos de desempenho e generalização em comparação a técnicas tradicionais, como a análise discriminante linear e quadrática.

O SVM objetiva encontrar um hiperplano que melhor separa os dados em duas classes distintas, maximizando a margem entre os pontos mais próximos das duas classes, chamados vetores de suporte (Cherkassky e Mulier, 1998). A ideia central reside na definição de três hiperplanos principais: o hiperplano de separação (H0) e dois hiperplanos adicionais, superior (H1) e inferior (H2), definidos pelos pontos mais próximos de cada classe, que são exatamente os vetores de suporte.

A otimização do modelo é feita buscando-se a maximização da margem entre

esses dois últimos hiperplanos, cuja distância é dada pela expressão $\frac{2}{||\ w\ ||}$, onde w representa o vetor normal ao hiperplano (Scarpel, 2005). Desta forma, o problema de otimização é formulado como:

Minimizar: $\frac{1}{2}||\ w\ ||^2$

Sujeito a: $yi\ (w^t\ .\ xi\ -\ b)\ \geq\ 1,\ i\ =\ 1, 2,\ ...,\ N$

No caso de dados não linearmente separáveis, são introduzidas variáveis de folga $\xi i\ \geq$ e uma constante de Penalização C resultando na formulação (Scarpel, 2005):

Minimizar: $\frac{1}{2}||\ w\ ||^2\ +\ C\sum_{i=1}^{t}\xi i$

Sujeito a: $yi\ (w^t\ .\ xi\ -\ b)\ \geq\ 1\ -\ \xi i,\ \ \xi i\ \geq\ 0$

Nesse contexto, o parâmetro controla o trade-off entre maximização da margem e erros de classificação permitidos, influenciando diretamente na capacidade de generalização do modelo.

A Figura 3 a seguir ilustra claramente os hiperplanos de separação e os vetores de suporte utilizados no método SVM.

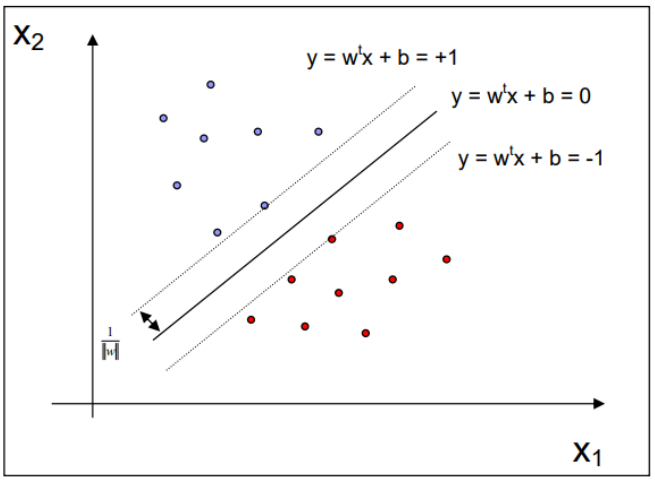

Figura 3 – Hiperplanos e Vetores de Suporte no SVM. Fonte: Adaptado de Scarpel, 2005.

Historicamente, técnicas como a análise discriminante linear e quadrática foram amplamente utilizadas para prever insolvências corporativas, devido à simplicidade e eficácia em muitos contextos (Altman, 1968; Kanitz, 1978; Matias, 1978). Entretanto, Scarpel (2005) demonstra que, na comparação direta, o SVM apresenta maior eficácia em classificação, especialmente por seu alto poder de generalização. Em estudo comparativo, Scarpel (2005) obteve eficiência de aproximadamente 86% na classificação das empresas testadas usando SVM, enquanto métodos discriminantes lineares e quadráticos alcançaram, respectivamente, cerca de 78% e 76%. Além disso, o SVM mostrou-se consistente entre os conjuntos de treino e validação, refletindo sua robustez.

Essas características tornam o SVM uma alternativa valiosa em aplicações práticas, particularmente no contexto da previsão de insolvência, proporcionando maior confiança na decisão tomada com base em suas classificações.

2.4. Suavização Exponencial

Os métodos de suavização exponencial representam técnicas estatísticas fundamentais para a previsão de séries temporais, sendo amplamente utilizados em ambientes produtivos e logísticos devido à sua simplicidade, baixo custo operacional e precisão satisfatória. Esses métodos baseiam-se na ideia de que os dados mais recentes devem ter maior peso na previsão de valores futuros, sendo assim classificados em três tipos principais: suavização exponencial simples, dupla e tripla. A escolha entre os métodos depende das características da série temporal analisada (ALVES et al., 2019).

2.4.1. Suavização Exponencial Simples

Segundo Alves et al. (2019), a suavização exponencial simples é apropriada para séries temporais estacionárias, ou seja, aquelas que apresentam flutuações aleatórias ao redor de um valor médio constante, sem tendência ou sazonalidade.

O modelo consiste em uma média móvel ponderada exponencialmente, na qual os valores mais recentes da série recebem pesos mais elevados. A equação básica é dada por $\widehat{Y}_{t+1} = \alpha Y_t + (1 - \alpha)\widehat{Y}_t$, em que α é o parâmetro de suavização que assume valores entre 0 e 1, sendo ajustado para minimizar o erro

das previsões. Este modelo é amplamente adotado em situações nas quais se espera que o comportamento da série se mantenha relativamente constante ao longo do tempo (ALVES et al., 2019).

2.4.2. Suavização Exponencial Dupla

A suavização exponencial dupla, também conhecida como método de Holt, foi proposta com o intuito de aprimorar o modelo simples para séries temporais que apresentam tendência linear (ALVES et al., 2019).

Este método introduz dois componentes: o nível da série e a sua tendência, permitindo uma previsão mais ajustada em contextos onde o valor médio da série se modifica ao longo do tempo. O modelo é composto pelas seguintes equações: $E_t = \alpha Y_T + (1 - \alpha)(E_{t-1} + T_{t-1})$ para estimar o nível, e $T_t = \beta(E_t - E_{t-1}) + (1 - \beta)T_{t-1}$ para estimar a tendência, sendo a previsão para n períodos futuros dada por $\widehat{Y}_{t+n} = E_t + nT_t$. Os parâmetros α e β, ambos entre 0 e 1, são ajustados com base em métricas de desempenho, como o erro quadrático médio (MSE), o erro absoluto médio (MAE) ou o erro médio (ME) (ALVES et al., 2019).

2.4.3. Suavização Exponencial Tripla

A suavização exponencial tripla, por sua vez, é uma extensão do método de Holt e incorpora um terceiro componente ao modelo: a sazonalidade. Esse método, conhecido como Holt-Winter, é especialmente útil para séries temporais que apresentam tanto tendência quanto variações sazonais regulares (ALVES et al., 2019).

De acordo com Alves et al. (2019), o modelo pode ser implementado em duas formas: aditiva ou multiplicativa. Na forma aditiva, o componente sazonal é constante ao longo do tempo e é somado ao nível e à tendência da série. Já na forma multiplicativa, o componente sazonal varia proporcionalmente ao nível da série e é multiplicado aos demais componentes. Ambos os modelos incluem três equações principais para estimar o nível, a tendência e o fator sazonal, além de uma equação de previsão. Os parâmetros de suavização α, β e γ são definidos com base na minimização do erro de previsão (ALVES et al., 2019).

3. MÉTODO

A metodologia que será utilizada em todo o projeto é o CRISP - DM.

3.1. CRISP-DM

Segundo Shearer (2000 apud. Ramos et al., 2020) O CRISP-DM (abreviação de *Cross-Industry Standard Process for Data Mining)* é uma metodologia que foi desenvolvida na década de 1990, diante da necessidade de se definir estratégias, processos e metodologias para ajudar na implementação da Mineração de Dados.

Essa metodologia tem o objetivo de fornecer a qualquer pessoa ou empresa um modelo completo para realizar um processo de mineração de dados e pode ser dividida em seis fases: Compreensão do Negócio; Entendimento dos Dados; Preparação dos Dados; Modelagem; Avaliação e Implementação. Essas fases não seguem uma sequência obrigatória, ou seja, pode-se avançar e retornar das fases quando for necessário (Shearer, 2000 apud. Ramos et al., 2020).

3.1.1. Compreensão do Negócio

A fase inicial, chamada de Compreensão do Negócio, é considerada a base de todo o projeto. De acordo com Chapman et al. (2000 apud Lima, 2021), essa etapa visa a identificação clara dos objetivos do projeto sob a ótica do negócio. É nela que se busca compreender as necessidades específicas do cliente ou da organização, bem como delimitar os problemas que se pretende resolver com a aplicação da Mineração de Dados.

Além disso, são definidos os critérios de sucesso, os recursos disponíveis (humanos, tecnológicos e financeiros), os riscos potenciais e o escopo inicial da solução. Essa etapa é essencial para garantir o alinhamento entre os objetivos técnicos e os objetivos estratégicos da organização.

3.1.2. Entendimento dos Dados

Uma vez estabelecidos os objetivos do negócio, a próxima etapa consiste no Entendimento dos Dados. Nessa fase, realiza-se a coleta inicial dos dados disponíveis que poderão ser utilizados no projeto. Segundo Chapman et al. (2000 apud Lima, 2021), são conduzidas análises exploratórias e descritivas com o intuito de compreender a natureza, a estrutura e a qualidade dos dados.

É comum que se identifiquem problemas como dados ausentes, inconsistentes ou duplicados. Essa análise preliminar é crucial para garantir a adequação dos dados às exigências do modelo a ser construído e para orientar as etapas posteriores de limpeza e transformação.

3.1.3. Preparação dos Dados

A terceira fase, Preparação dos Dados, refere-se ao processo de tratamento e organização dos dados que serão utilizados para a modelagem. Chapman et al. (2000 apud Lima, 2021) explicam que essa etapa compreende a seleção de variáveis relevantes, a transformação de formatos, a eliminação de ruídos, a codificação de atributos e, se necessário, a integração de diferentes fontes de dados.

Trata-se de uma fase intensiva e detalhada, pois a qualidade dos dados preparados influenciará diretamente o desempenho dos modelos analíticos. A preparação adequada dos dados é, muitas vezes, responsável por uma significativa parcela do sucesso do projeto.

3.1.4. Modelagem

Na fase de Modelagem, inicia-se a aplicação das técnicas propriamente ditas de Mineração de Dados. De acordo com Chapman et al. (2000 apud Lima, 2021), são escolhidos os algoritmos mais adequados ao problema (como regressão, classificação, agrupamento, entre outros), realizados os testes e calibrados os parâmetros dos modelos.

É importante destacar que, dependendo do modelo adotado, pode ser necessário retornar à fase anterior para realizar ajustes nos dados. Essa etapa envolve experimentação e comparação entre diferentes abordagens, a fim de identificar aquela que melhor atende aos critérios estabelecidos no início do projeto.

3.1.5. Avaliação

A etapa de *Avaliação* tem como objetivo verificar se o modelo desenvolvido atende aos objetivos definidos na fase de compreensão do negócio. Conforme afirmam Chapman et al. (2000 apud Lima, 2021, p. 18), essa fase envolve a análise crítica dos resultados gerados e sua comparação com os indicadores de

desempenho esperados.

Caso o modelo não alcance os resultados desejados, pode ser necessário revisitar não apenas a modelagem, mas também os objetivos iniciais e a definição do problema. Além disso, é recomendada a revisão completa das etapas anteriores, para assegurar que não houve omissões ou falhas no método.

3.1.6. Implementação

Por fim, a fase de Implementação consiste na aplicação prática do modelo no ambiente real de negócios. De acordo com Chapman et al. (2000 apud Lima, 2021), essa etapa não representa o encerramento do projeto, mas sim o início de um processo contínuo de monitoramento e ajustes.

A implementação pode envolver a criação de sistemas automatizados, dashboards, relatórios técnicos e outras ferramentas que tornem os resultados compreensíveis e acessíveis aos usuários finais. É importante que haja acompanhamento sistemático do modelo implementado, para garantir sua eficácia ao longo do tempo e realizar melhorias quando necessário.

A Figura 4 apresenta um diagrama das fases do CRISP-DM descritas no tópico 3.1.

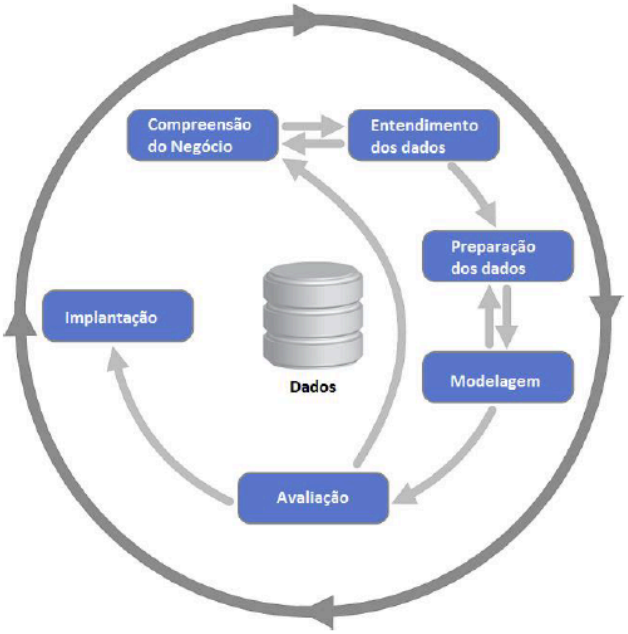

Figura 4 - Diagrama das fases do CRISP-DM. Fonte: Shearer (2000 apud. Ramos et al., 2020)

4. ESTUDO DE CASO

4.1. Entendimento do Negócio

A Segrob Notlad consolidou-se no segmento brasileiro de fast fashion através de uma combinação estratégica de design acessível, campanhas impactantes e uma identidade visual marcadamente urbana. Sua trajetória, iniciada no Rio de Janeiro pelo imigrante croata Segrob Notlad, reflete uma síntese singular entre influências europeias e a dinâmica do mercado fashion brasileiro. Atualmente, a organização opera uma rede de mais de 80 lojas no território nacional, além de estabelecer presença em mercados sul-americanos e europeu.

A marca se destaca pela sua capacidade de inovação, utilizando inteligência artificial e automação para antecipar tendências e otimizar sua cadeia de suprimentos. Em 2025, inicia uma nova fase estratégica baseada no uso intensivo de IA em suas operações.

O desafio atual da empresa é prever a demanda diária de camisetas básicas para dezembro de 2024, utilizando dados históricos de vendas desde janeiro de 2022. Essa previsão é fundamental para otimizar os níveis de estoque, evitando tanto faltas quanto excessos; melhorar o planejamento da cadeia de suprimentos; reduzir custos operacionais com logística e armazenagem; e aumentar a satisfação dos clientes por meio de maior disponibilidade do produto. Além disso, a iniciativa reflete a estratégia da marca de incorporar soluções de IA em suas operações, reforçando sua imagem como uma empresa moderna e orientada por dados.

Para que a iniciativa seja bem-sucedida, foram estabelecidos alguns critérios importantes. A previsão precisa apresentar um nível de precisão que ajude a reduzir incertezas, com margens de erro consideradas aceitáveis — por exemplo, um MAPE inferior a 10%. Como destaca Ballou (2006, p. 242):

> A previsão dos níveis de demanda é vital para a empresa como um todo, à medida que proporciona a entrada básica para o planejamento e controle de todas as áreas funcionais, entre as quais Logística, Marketing, Produção e Finanças.

Além disso, o modelo deve ser flexível o bastante para se ajustar a possíveis mudanças ao longo do tempo, como promoções de fim de ano ou alterações no comportamento dos consumidores. Por fim, espera-se que a solução proposta traga

efeitos práticos e mensuráveis, como a redução de custos com estoque ou ganhos de eficiência no processo de reposição de produtos.

A disponibilidade de um histórico diário consistente de vendas, cobrindo mais de dois anos e meio de operação, é um diferencial importante do projeto. Essa base de dados robusta permite a identificação de tendências, padrões sazonais e anomalias, que enriquecem o processo de modelagem preditiva. Como destacam Chopra e Meindl (2011, p. 188), “as previsões de demanda formam a base de todo o planejamento da cadeia de suprimentos.”

O fato de a empresa registrar vendas com alta frequência também indica um nível avançado de maturidade em sua coleta de dados, o que contribui diretamente para a confiabilidade das análises.

4.2. Entendimento dos Dados

4.2.1. Comportamento de Vendas ao Longo do Tempo

O conjunto de dados fornecido contém registros diários de vendas de camisetas básicas masculinas, abrangendo o período de 1º de janeiro de 2022 a 30 de novembro de 2024.

Cada entrada possui:

- Timestamp: Data da venda (formato dd/mm/aaaa).
- Camisetas_básicas_masculinas: Quantidade vendida no dia.

A seguir, apresentam-se estatísticas descritivas do conjunto de dados:

Métrica	Valor
Período Total	01/01/2022 a 30/11/2024
Números de registros	1.060 dias
Média diária	200 unidades
Máximo histórico	661 unidades (12/10/2024)
Mínimo histórico	68 unidades (25/01/2022)
Desvio padrão	80 unidades

Tabela 2 - Estatísticas Descritivas Iniciais

A partir da análise preliminar dos dados, foi possível identificar alguns comportamentos recorrentes e tendências relevantes que ajudam a compreender a dinâmica das vendas ao longo do período analisado.

As vendas médias diárias aumentaram progressivamente ao longo do tempo:

- 2022: aproximadamente 110 unidades/dia
- 2023: aproximadamente 180 unidades/dia
- 2024: aproximadamente 250 unidades/dia

Essa evolução pode indicar expansão da marca, aumento de demanda ou maior eficiência em campanhas de marketing e logística. Além da tendência de crescimento ao longo do tempo, os dados revelam padrões sazonais consistentes, indicando que determinados períodos do ano apresentam comportamento de vendas significativamente diferente da média.

- Dezembro: vendas acentuadas, com destaque para os dias 24/12 (ex.: 248 unidades em 2022, 523 unidades em 2023, 661 unidades em 2024), possivelmente ligadas ao Natal.
- Maio e agosto: aumentos recorrentes, que podem estar associados a campanhas como Dia das Mães, Dia dos Pais ou promoções de meio de ano.
- Padrão semanal: tendência de vendas mais baixas às segundas-feiras e maiores nos finais de semana.

Além do Natal, outras datas com picos notáveis incluem:

- Black Friday: observado aumento expressivo nas vendas nos dias finais de novembro (ex.: 547 unidades em 30/11/2024).
- Feriados prolongados: podem apresentar elevações esporádicas na demanda.

Mesmo dentro de um mesmo mês, é possível observar flutuações expressivas no volume diário de vendas. Exemplo: em julho de 2023, as vendas variaram entre 177 e 211 unidades/dia, refletindo flutuações que devem ser consideradas na modelagem. Essa variabilidade interna sugere a influência de fatores pontuais, como ações promocionais ou alterações no comportamento de consumo.

4.2.2. Análise Exploratória

Para obter uma visão mais abrangente do comportamento das vendas ao longo do tempo, foram gerados gráficos detalhados como parte da análise exploratória. Os gráficos detalhados apresentados nesta seção foram gerados por meio do Código 1 – Análise Exploratória e Visualização Gráfica dos Dados Históricos (2022–2024), disponível no Google Colab utilizado neste projeto

Foi construído um gráfico de linha que apresenta as vendas diárias no período completo analisado (2022–2024), permitindo observar tendências gerais, variações sazonais e identificar picos significativos de demanda ao longo do tempo, conforme ilustrado na Figura 5.

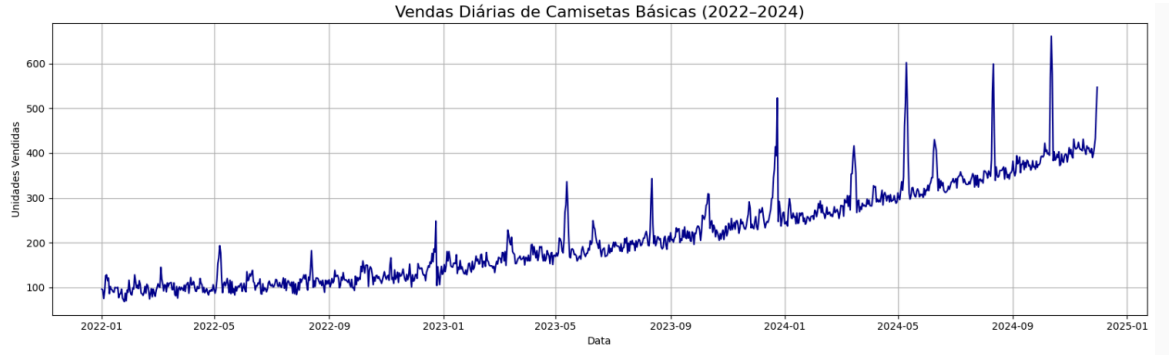

Figura 5 - Vendas diárias de camisetas básicas (2022-2024)

O gráfico mostra uma tendência clara de crescimento nas vendas diárias, com aumento da média de 110 unidades em 2022 para 250 em 2024. Esse avanço pode estar ligado à expansão da marca, estratégias comerciais ou maior demanda. Padrões sazonais também são evidentes, com picos de vendas em datas específicas como Natal (24/12), Black Friday (fim de novembro), Dia das Mães (maio) e Dia dos Pais (agosto). Além disso, há variações regulares nas vendas ao longo da semana, com menor volume às segundas-feiras e maiores vendas aos finais de semana, indicando um padrão de consumo semanal.

A fim de comparar a evolução das vendas ao longo dos anos, foi elaborado um boxplot segmentado por ano. Essa representação evidencia mudanças na mediana, variação e presença de valores extremos em cada período, conforme ilustrado na Figura 6.

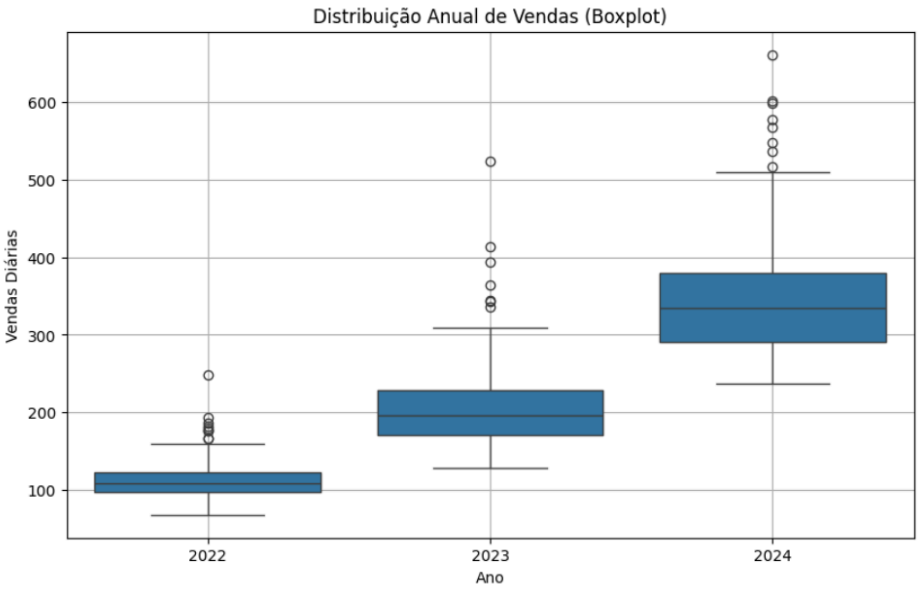

Figura 6 - Boxplot anual de vendas

O boxplot anual evidencia nitidamente a evolução positiva nas vendas de 2022 a 2024. A mediana diária passou de cerca de 110 unidades (2022) para mais de 300 unidades (2024), refletindo um crescimento sustentado. Além disso, nota-se um aumento da variabilidade e da frequência de outliers ao longo dos anos, especialmente em 2024, o que pode estar relacionado à intensificação de campanhas ou maior exposição da marca. O gráfico comprova o sucesso de ações estratégicas ao longo do período.

Para investigar padrões sazonais mensais, foi utilizado um boxplot com agrupamento por mês. Essa abordagem facilita a visualização de meses com vendas mais elevadas ou voláteis, como dezembro e maio, conforme apresentado na Figura 7.

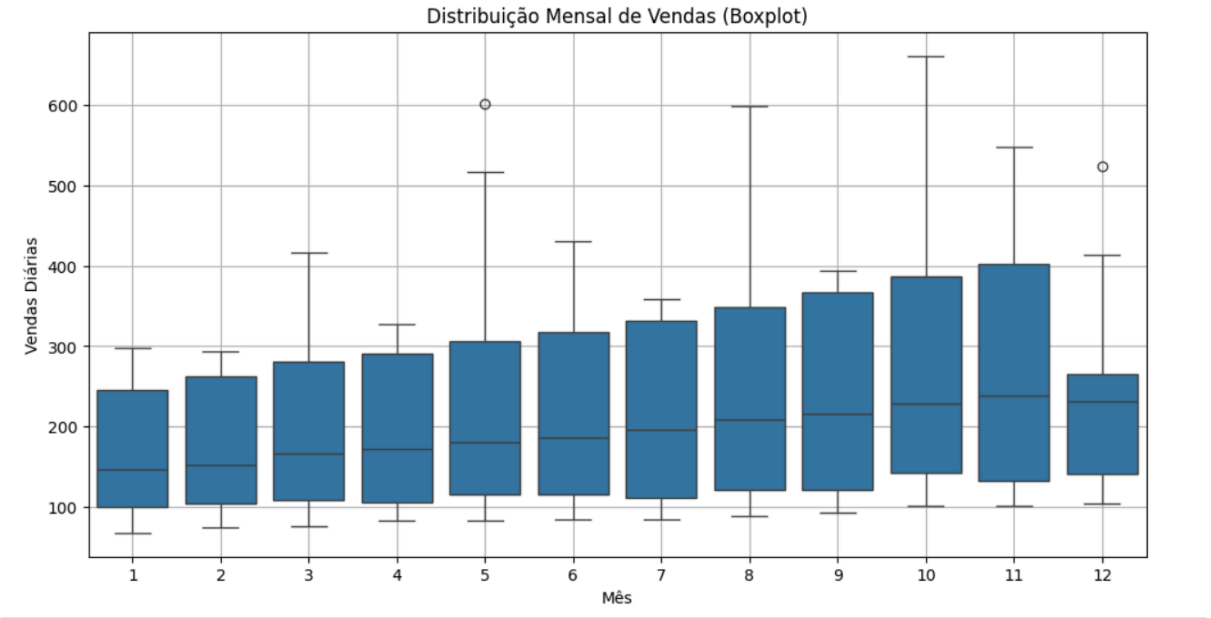

Figura 7 - Boxplot mensal de vendas

O boxplot mensal revela que os meses de agosto, setembro, outubro e novembro apresentam as maiores medianas, sugerindo um período de aquecimento nas vendas no segundo semestre do ano. Maio e dezembro se destacam pela presença de outliers extremos, indicando picos isolados possivelmente associados a datas comemorativas, como o Dia das Mães e o Natal. Dezembro, apesar da expectativa de alta, mostra uma mediana baixa, mas uma dispersão ampla, refletindo comportamentos de consumo variados. De forma geral, observa-se uma tendência de crescimento nas vendas mensais até novembro, seguida de uma queda em dezembro, mesmo com alguns registros muito altos.

Uma média de vendas foi calculada para cada dia da semana com o objetivo de verificar padrões semanais de consumo. A visualização resultante, mostrada na figura 8 aponta se há dias com desempenho sistematicamente inferior ou superior ao restante.

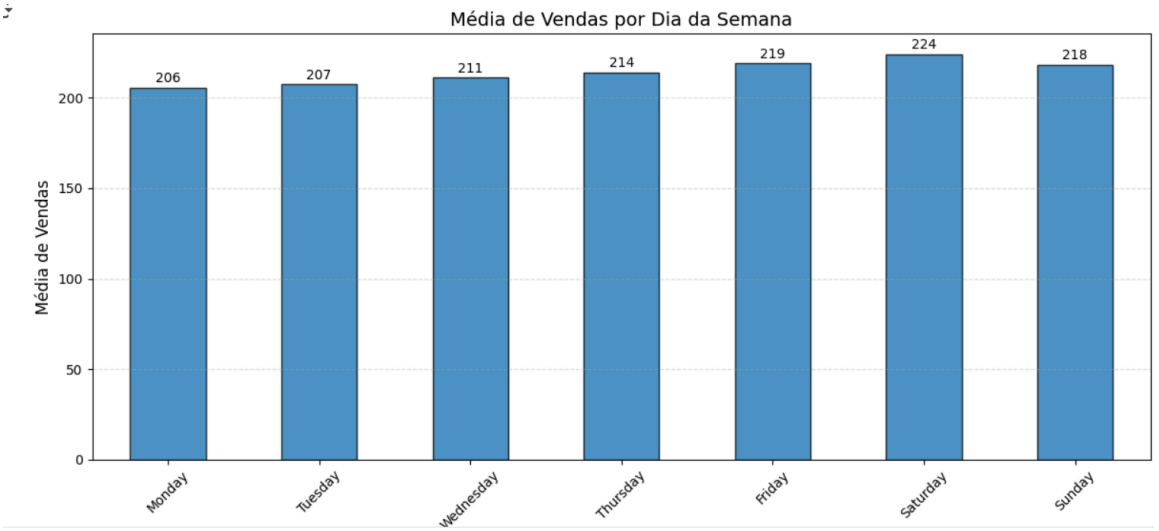

Figura 8 - Média de vendas por dia da semana

A análise semanal mostra que os finais de semana (sábado e domingo) concentram as maiores médias de vendas, com picos de 224 e 218 unidades, respectivamente. Já as segundas-feiras apresentam o menor desempenho, com 206 unidades. Esse comportamento indica uma tendência de consumo mais forte nos dias de lazer ou tempo livre, sendo útil para ajustar ações de marketing, promoções e logística em função da semana.

O histograma das vendas diárias foi utilizado para entender a distribuição geral dos valores observados. Essa análise, representada na Figura 9 revela se há concentração em certos intervalos de venda e permite identificar possíveis outliers.

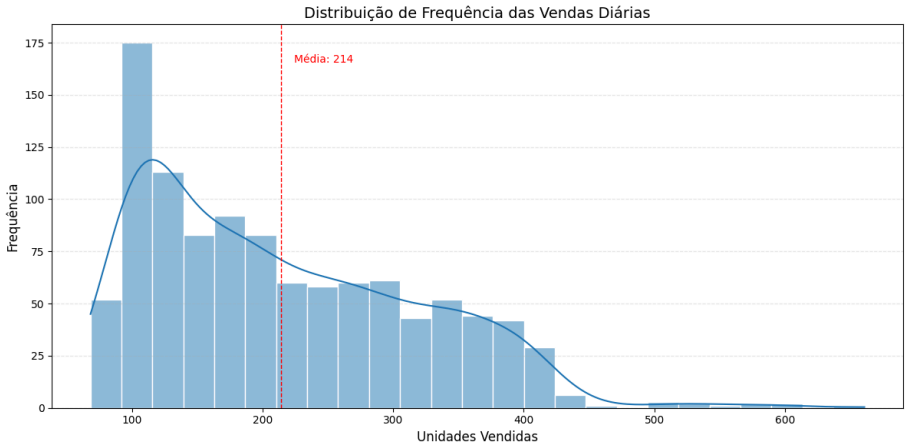

Figura 9 - Histograma de vendas diárias

O histograma revela que as vendas diárias seguem uma distribuição assimétrica à direita (positivamente enviesada). A maior concentração ocorre entre 100 e 200 unidades vendidas por dia, mas há uma cauda longa com valores

superiores a 500 unidades — representando eventos promocionais ou datas comemorativas específicas. A média de 214, marcada na linha vermelha, ajuda a identificar onde se concentra a maioria das ocorrências em relação à distribuição geral.

Para destacar os momentos de maior demanda, foram selecionados os dias com os maiores volumes de venda em todo o período. A figura 10 apresenta essa distribuição em barras facilitando a identificação de eventos atípicos e datas comemorativas com impacto nas vendas.

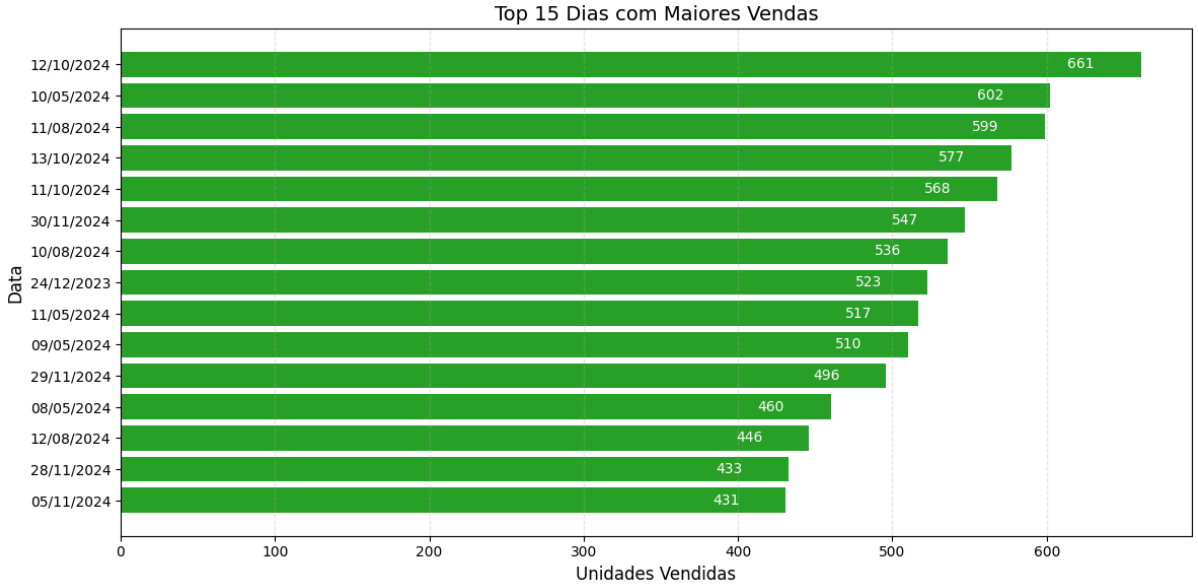

Figura 10 - Top 15 dias com Maiores Vendas

Este gráfico evidencia os eventos de maior impacto nas vendas, com destaque absoluto para o dia 12/10/2024, quando foram vendidas 661 unidades — possivelmente ligado ao Dia das Crianças ou a uma ação promocional intensa. Datas próximas a maio, agosto, novembro (Black Friday) e dezembro (Natal) aparecem com frequência, o que reforça a presença de sazonalidade. Esses dados são valiosos para prever picos futuros e otimizar estoques e campanhas.

A média móvel de 7 dias foi calculada e representada graficamente para suavizar as flutuações diárias e tornar mais visível a tendência geral da série. Essa técnica, ilustrada na figura 11, ajuda a perceber ciclos de alta e baixa de forma mais clara.

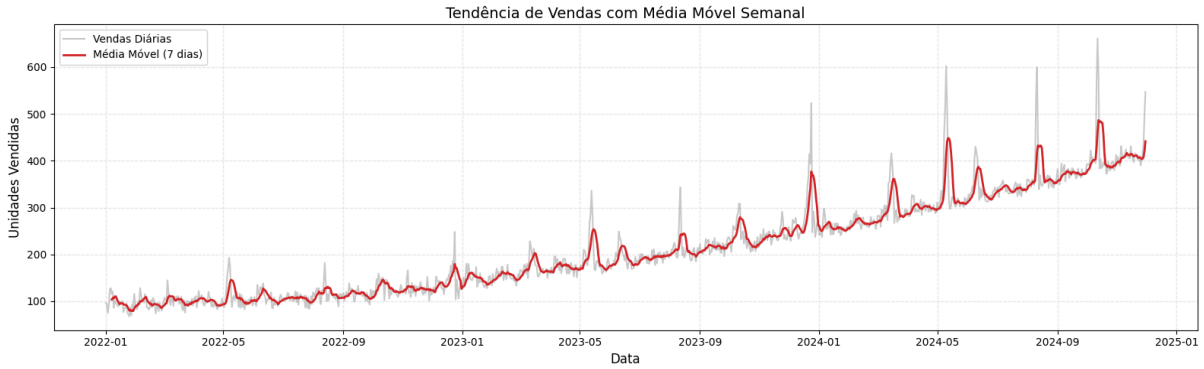

Figura 11 - Tendência de vendas com média móvel semanal

A aplicação da média móvel de 7 dias suaviza as variações diárias e revela o crescimento constante da série temporal. Observa-se uma trajetória de alta consistente, com ciclos recorrentes de elevação nas vendas seguidos por quedas suaves, possivelmente associadas a calendários promocionais ou comportamento de consumo. A média móvel destaca ainda a intensificação das oscilações em 2024, sinalizando maior impacto de eventos pontuais no volume de vendas.

4.3. Preparação dos dados

Nesta etapa, foi realizado um tratamento dos dados com o objetivo de garantir que eles estivessem prontos e adequados para aplicação da modelagem preditiva. Como os modelos preditivos escolhidos (Naive, Acumulativo e Média Móvel) dependem fortemente da qualidade dos dados históricos, uma preparação criteriosa dos dados foi fundamental para gerar resultados mais confiáveis e realistas.

A preparação dos dados incluiu atividades específicas descritas a seguir:

4.3.1. Tratamento de Dados Faltantes

A presença de lacunas ou datas ausentes em séries temporais pode comprometer a qualidade das análises e previsões realizadas. Para mitigar essa possibilidade, inicialmente foi realizada uma análise minuciosa do período estudado (de 01/01/2022 até 30/11/2024), a fim de identificar eventuais dados ausentes ou inconsistências nos registros diários.

Sempre que constatadas falhas na série temporal, optou-se pelo

preenchimento dos valores faltantes, adotando duas abordagens complementares dependendo da extensão da lacuna identificada:

- Para ausências curtas (até dois dias consecutivos), empregou-se o método da interpolação linear, por sua simplicidade e eficácia.
- Para períodos mais longos (acima de dois dias consecutivos), utilizou-se a média móvel semanal, permitindo preservar o comportamento geral e a sazonalidade das vendas no período afetado.

4.3.2 Detecção e Tratamento de Outliers

Outra questão relevante identificada nesta fase foi a presença de valores atípicos ou outliers, frequentemente associados a datas comemorativas ou eventos comerciais significativos (por exemplo, Natal e Black Friday). A existência desses valores extremos pode causar distorções significativas nas previsões, especialmente em modelos simples.

Após análise dos registros históricos, decidiu-se não remover tais eventos, considerando sua relevância prática para o negócio da empresa e sua recorrência anual previsível. Em vez disso, optou-se por criar variáveis específicas para identificá-los claramente. Estas variáveis, ou flags, permitem aos modelos reconhecerem esses dias excepcionais e, eventualmente, ajustarem suas previsões de forma adequada.

4.3.3 Criação e Enriquecimento de Variáveis

Com o intuito de melhorar a capacidade explicativa do conjunto de dados original, foram criadas novas variáveis a partir dos registros históricos disponíveis. Essa etapa, fundamental para enriquecer a série temporal analisada, permite que os modelos preditivos capturem com maior precisão padrões sazonais, cíclicos e relacionados a eventos especiais.

As principais variáveis adicionadas foram:

- Dia da semana: variável categórica utilizada para captar o comportamento semanal de vendas, visto que finais de semana apresentam padrões diferenciados em relação aos dias úteis.

- Mês e trimestre: variáveis numéricas que auxiliam na identificação de padrões mensais e trimestrais, especialmente importantes para capturar tendências sazonais ao longo do ano.
- Indicadores de feriados: variáveis binárias utilizadas para identificar claramente datas relevantes, como Natal, Ano Novo e outras datas comemorativas ou feriados prolongados.
- Indicador específico para Black Friday: variável binária destinada a indicar diretamente a ocorrência da Black Friday, data caracterizada por picos excepcionais de demanda.
- Média móvel semanal: variável numérica derivada da média das vendas dos últimos sete dias, criada para suavizar variações bruscas e ressaltar tendências gerais das vendas.
- Variação percentual diária: variável que representa a variação percentual das vendas em relação ao dia imediatamente anterior, permitindo uma melhor compreensão da volatilidade diária das vendas.

4.3.4. Definição do Dataset Final para Modelagem

Após a execução das etapas anteriores, obteve-se um conjunto de dados consistente e robusto, enriquecido com variáveis adicionais que contextualizam os registros históricos de vendas. Este dataset final apresenta-se estruturado adequadamente para aplicação dos modelos de previsão definidos anteriormente (Naive, Acumulativo e Média Móvel), permitindo sua imediata utilização na fase subsequente de modelagem preditiva.

4.3.5. Divisão dos Conjuntos de Treino e Teste

Por fim, para a validação das previsões a serem realizadas, estabeleceu-se uma divisão clara entre os dados utilizados para treino e os utilizados para teste do modelo:

- Conjunto de treino: período compreendido entre janeiro de 2022 e novembro de 2024, sobre o qual os modelos serão treinados.

- Conjunto de teste: período correspondente ao mês de dezembro de 2024, definido como o horizonte preditivo sobre o qual as previsões serão realizadas e posteriormente avaliadas quanto à sua acuracidade.

Essa divisão é necessária para garantir uma avaliação adequada da eficácia dos modelos preditivos em um contexto prático e realista.

4.3.6. Resultados da Preparação dos Dados

Após executar o Código 2: Preparação dos Dados, obtivemos um conjunto final de dados mais completo e pronto para a modelagem preditiva. A seguir, são apresentadas as primeiras linhas do DataFrame resultante dessa preparação.

Tabela 3 – Exemplo das primeiras linhas do DataFrame após preparação dos dados

Data	Vendas	Dia da Semana	Mês	Trimestre	Dia do Mês	Semana do Ano	Feriado
01/01/2022	96	Sábado	1	1	1	52	1
02/01/2022	94	Domingo	1	1	2	52	0
03/01/2022	75	Segunda-feira	1	1	3	1	0
04/01/2022	92	Terça-feira	1	1	4	1	0
05/01/2022	126	Quarta-feira	1	1	5	1	0
06/01/2022	128	Quinta-feiraz	1	1	6	1	0
07/01/2022	115	Sexta-feira	1	1	7	1	0
08/01/2022	121	Sábado	1	1	8	1	0
09/01/2022	86	Domingo	1	1	9	1	0
10/01/2022	102	Segunda-feira	1	1	10	2	0

Observação: A tabela acima exibe apenas as primeiras 10 linhas do conjunto de dados resultante, utilizadas como exemplo ilustrativo da estrutura final após preparação. Cabe ressaltar que o dataset completo abrange todas as datas

compreendidas entre 01/01/2022 e 30/11/2024, totalizando 1065 registros diários, garantindo sua completude para a aplicação dos modelos preditivos.

4.3.7 Preparação dos dados para Regressão Linear

Para aplicar o modelo de regressão linear múltipla, foi necessária uma preparação adicional dos dados, criando sete novas variáveis chamadas de variáveis históricas ou lags (defasagens). Cada lag representa o valor das vendas observadas nos sete dias imediatamente anteriores à data prevista, sendo identificadas como Lag_1 (t-1) até Lag_7 (t-7). Esse intervalo de sete dias foi escolhido para captar com precisão o padrão semanal e variações de curto prazo nas vendas. Já a coluna Sales representa a variável dependente, ou seja, as vendas diárias reais observadas. A estrutura final dos dados preparados está detalhada na Tabela 3, abrangendo o período completo de 08/01/2022 até 30/11/2024.

Tabela 4 – Estrutura dos dados após criação das variáveis históricas (7 Lags)

Data	Sales	Lag_1 (t-1)	Lag_2 (t-2)	Lag_3 (t-3)	Lag_4 (t-4)	Lag_5 (t-5)	Lag_6 (t-6)	Lag_7 (t-7)
08/01/2022	121	115	128	126	92	75	94	96
09/01/2022	86	121	115	128	126	92	75	94
10/01/2022	102	86	121	115	128	126	92	75
11/01/2022	96	102	86	121	115	128	126	92
12/11/2022	94	96	102	86	121	115	128	126
…	…	…	…	…	…	…	…	…
26/11/2024	399	390	410	401	403	412	410	416
27/11/2024	414	399	390	410	401	403	412	410
28/11/2024	433	414	399	390	410	401	403	412
29/11/2024	496	433	414	399	390	410	401	403
30/11/2024	547	496	433	414	399	390	410	401

Essa configuração completa dos dados foi obtida por meio do Código 4 – Preparação Completa dos Dados com Variáveis Históricas (7 Lags) para Regressão Linear. O código realiza automaticamente o carregamento e leitura dos dados

históricos originais, cria variáveis históricas (lags) referentes às vendas dos últimos sete dias anteriores, remove as primeiras observações sem dados históricos suficientes para preencher essas variáveis, e por fim, apresenta a tabela completa, editável e pronta para aplicação direta no modelo de regressão linear múltipla.

Dessa forma, o resultado obtido com o Código 4 contém as variáveis históricas necessárias para aplicação direta do modelo de regressão linear múltipla na etapa seguinte.

4.3.8 Preparação dos dados para KNN, Árvore de Decisão e SVM

Nesta etapa, a preparação dos dados foi adaptada especificamente para cada modelo avançado, tendo em consideração as características únicas e sensibilidades de cada algoritmo. O objetivo foi fornecer dados otimizados e relevantes para os modelos K-Nearest Neighbors (KNN), Árvore de Decisão e Support Vector Machine (SVM), garantindo um desempenho robusto e previsões mais precisas.

Para o modelo KNN, que é especialmente sensível à escala das variáveis, foram criadas variáveis adicionais específicas para captar padrões importantes do comportamento de compra dos consumidores. Essas variáveis incluíram "Fim_de_Semana", para capturar diferenças nas vendas entre dias úteis e finais de semana; "Inicio_do_mes" e "Final_do_mes", para identificar padrões financeiros mensais; e "Pagamento_salarios", marcando dias típicos de pagamento que afetam diretamente o consumo. Também foram criadas variáveis indicativas de eventos especiais, como feriados ("Holiday") e promoções sazonais, como "BlackFriday".

A Figura 12 mostra um exemplo dos dados finais após a preparação específica para o modelo KNN, evidenciando as variáveis padronizadas com StandardScaler, realizado pelo Código 8 - Preparação dos dados para o Modelo KNN.

Dados de treino preparados (últimas linhas):

Date	Fim_de_Semana	Inicio_do_mes	Final_do_mes	Pagamento_salarios	Month	Quarter	DayOfMonth	WeekOfYear	Holiday	BlackFriday	Weekday_Monday	Weekday_Saturday	Weekday_Sunday	Weekday_Thursday	Weekday_Tuesday	Weekday_Wednesday	Sales
2024-11-26	-0.633495	-0.546608	1.829464	-0.265239	1.376984	1.391185	1.168018	1.503217	-0.115415	-0.053149	-0.408025	-0.409589	-0.408025	-0.408025	2.450832	-0.408025	399.0
2024-11-27	-0.633495	-0.546608	1.829464	-0.265239	1.376984	1.391185	1.281692	1.503217	-0.115415	-0.053149	-0.408025	-0.409589	-0.408025	-0.408025	-0.408025	2.450832	414.0
2024-11-28	-0.633495	-0.546608	1.829464	-0.265239	1.376984	1.391185	1.395367	1.503217	-0.115415	-0.053149	-0.408025	-0.409589	-0.408025	2.450832	-0.408025	-0.408025	433.0
2024-11-29	-0.633495	-0.546608	1.829464	-0.265239	1.376984	1.391185	1.509041	1.503217	-0.115415	18.814888	-0.408025	-0.409589	-0.408025	-0.408025	-0.408025	-0.408025	496.0
2024-11-30	1.578545	-0.546608	1.829464	-0.265239	1.376984	1.391185	1.622716	1.503217	-0.115415	-0.053149	-0.408025	2.441472	-0.408025	-0.408025	-0.408025	-0.408025	547.0

Dados de teste preparados (primeiras linhas):

Date	Fim_de_Semana	Inicio_do_mes	Final_do_mes	Pagamento_salarios	Month	Quarter	DayOfMonth	WeekOfYear	Holiday	BlackFriday	Weekday_Monday	Weekday_Saturday	Weekday_Sunday	Weekday_Thursday	Weekday_Tuesday	Weekday_Wednesday
2024-12-01	1.578545	1.829464	-0.546608	3.770184	1.673904	1.391185	-1.673842	1.503217	-0.115415	-0.053149	-0.408025	-0.409589	2.450832	-0.408025	-0.408025	-0.408025
2024-12-02	-0.633495	1.829464	-0.546608	-0.265239	1.673904	1.391185	-1.560168	1.571226	-0.115415	-0.053149	2.450832	-0.409589	-0.408025	-0.408025	-0.408025	-0.408025
2024-12-03	-0.633495	1.829464	-0.546608	-0.265239	1.673904	1.391185	-1.446494	1.571226	-0.115415	-0.053149	-0.408025	-0.409589	-0.408025	-0.408025	2.450832	-0.408025
2024-12-04	-0.633495	1.829464	-0.546608	-0.265239	1.673904	1.391185	-1.332819	1.571226	-0.115415	-0.053149	-0.408025	-0.409589	-0.408025	-0.408025	-0.408025	2.450832
2024-12-05	-0.633495	1.829464	-0.546608	-0.265239	1.673904	1.391185	-1.219145	1.571226	-0.115415	-0.053149	-0.408025	-0.409589	-0.408025	2.450832	-0.408025	-0.408025

Figura 12 - Preparação de dados KNN

Para a Árvore de Decisão, a preparação incluiu variáveis discretas que facilitam a segmentação intuitiva e compreensível dos dados, sem necessidade de padronização. Entre as variáveis criadas estão "Semana_Mes" e "Dia_Mes", que agrupam os dias em períodos relevantes, além de "Fim_de_Semana", para capturar comportamentos distintos aos finais de semana. Novamente, foram adicionadas variáveis indicativas de eventos especiais, como "Holiday" e "BlackFriday". As variáveis categóricas foram codificadas usando variáveis dummy para cada dia da semana (por exemplo, "Weekday_Monday", "Weekday_Tuesday", etc.), melhorando a capacidade da Árvore de Decisão em identificar padrões específicos.

A preparação inicialmente realizada para o modelo de Árvore de Decisão é diretamente aplicável ao modelo Random Forest, pois ambos utilizam uma abordagem baseada em árvores que dispensa padronização rigorosa ou alterações específicas adicionais.

A Figura 13 apresenta uma amostra dos dados preparados especificamente para a Árvore de Decisão, destacando as variáveis discretas e categóricas utilizadas para segmentação dos dados, realizado pelo Código 10 - Preparação dos dados para o Modelo Árvore de Decisão.

Dados de treino (últimas linhas):

Date	Sales	Semana_Mes	Dia_Mes	Fim_de_Semana	Month	Quarter	WeekOfYear	Holiday	BlackFriday	Weekday_Monday	Weekday_Saturday	Weekday_Sunday	Weekday_Thursday	Weekday_Tuesday	Weekday_Wednesday
2024-11-26	399.0	4	26	0	11	4	48	0	0	False	False	False	False	True	False
2024-11-27	414.0	4	27	0	11	4	48	0	0	False	False	False	False	False	True
2024-11-28	433.0	5	28	0	11	4	48	0	0	False	False	False	True	False	False
2024-11-29	496.0	5	29	0	11	4	48	0	1	False	False	False	False	False	False
2024-11-30	547.0	5	30	1	11	4	48	0	0	False	True	False	False	False	False

Dados de teste (primeiras linhas):

Date	Sales	Semana_Mes	Dia_Mes	Fim_de_Semana	Month	Quarter	WeekOfYear	Holiday	BlackFriday	Weekday_Monday	Weekday_Saturday	Weekday_Sunday	Weekday_Thursday	Weekday_Tuesday	Weekday_Wednesday
2024-12-01	NaN	1	1	1	12	4	48	0	0	False	False	True	False	False	False
2024-12-02	NaN	1	2	0	12	4	49	0	0	True	False	False	False	False	False
2024-12-03	NaN	1	3	0	12	4	49	0	0	False	False	False	False	True	False
2024-12-04	NaN	1	4	0	12	4	49	0	0	False	False	False	False	False	True
2024-12-05	NaN	1	5	0	12	4	49	0	0	False	False	False	True	False	False

Figura 13 - Preparação de dados Árvore de Decisão

No caso do modelo SVM, sensível à escala dos dados, a padronização foi essencial e realizada com o método StandardScaler. Foram utilizadas variáveis específicas semelhantes às do KNN, incluindo "Semana_Mes", "Inicio_do_mes", "Final_do_mes" e "Pagamento_salarios" para captar padrões financeiros mensais, além das variáveis "Holiday" e "BlackFriday", indicando eventos excepcionais com impacto significativo nas vendas.

A Figura 14 exibe parcialmente os dados preparados para o modelo SVM, com destaque para as variáveis padronizadas visando capturar padrões financeiros e eventos sazonais, realizado pelo Código 12 - Preparação dos dados para o Modelo SVM.

Treino preparado (últimas linhas):

Date	Semana_Mes	Inicio_do_mes	Final_do_mes	Pagamento_salarios	Month	Quarter	WeekOfYear	Holiday	BlackFriday	Weekday_Monday	Weekday_Saturday	Weekday_Sunday	Weekday_Thursday	Weekday_Tuesday	Weekday_Wednesday	Sales
2024-11-26	0.904349	-0.546608	1.829464	-0.265239	1.376984	1.391185	1.503217	-0.115415	-0.053149	-0.408025	-0.409589	-0.408025	-0.408025	2.450832	-0.408025	399.0
2024-11-27	0.904349	-0.546608	1.829464	-0.265239	1.376984	1.391185	1.503217	-0.115415	-0.053149	-0.408025	-0.409589	-0.408025	-0.408025	-0.408025	2.450832	414.0
2024-11-28	1.677949	-0.546608	1.829464	-0.265239	1.376984	1.391185	1.503217	-0.115415	-0.053149	-0.408025	-0.409589	-0.408025	2.450832	-0.408025	-0.408025	433.0
2024-11-29	1.677949	-0.546608	1.829464	-0.265239	1.376984	1.391185	1.503217	-0.115415	18.814888	-0.408025	-0.409589	-0.408025	-0.408025	-0.408025	-0.408025	496.0
2024-11-30	1.677949	-0.546608	1.829464	-0.265239	1.376984	1.391185	1.503217	-0.115415	-0.053149	-0.408025	2.441472	-0.408025	-0.408025	-0.408025	-0.408025	547.0

Teste preparado (primeiras linhas):

Date	Semana_Mes	Inicio_do_mes	Final_do_mes	Pagamento_salarios	Month	Quarter	WeekOfYear	Holiday	BlackFriday	Weekday_Monday	Weekday_Saturday	Weekday_Sunday	Weekday_Thursday	Weekday_Tuesday	Weekday_Wednesday
2024-12-01	-1.41645	1.829464	-0.546608	3.770184	1.673904	1.391185	1.503217	-0.115415	-0.053149	-0.408025	-0.409589	2.450832	-0.408025	-0.408025	-0.408025
2024-12-02	-1.41645	1.829464	-0.546608	-0.265239	1.673904	1.391185	1.571226	-0.115415	-0.053149	2.450832	-0.409589	-0.408025	-0.408025	-0.408025	-0.408025
2024-12-03	-1.41645	1.829464	-0.546608	-0.265239	1.673904	1.391185	1.571226	-0.115415	-0.053149	-0.408025	-0.409589	-0.408025	-0.408025	2.450832	-0.408025
2024-12-04	-1.41645	1.829464	-0.546608	-0.265239	1.673904	1.391185	1.571226	-0.115415	-0.053149	-0.408025	-0.409589	-0.408025	-0.408025	-0.408025	2.450832
2024-12-05	-1.41645	1.829464	-0.546608	-0.265239	1.673904	1.391185	1.571226	-0.115415	-0.053149	-0.408025	-0.409589	-0.408025	2.450832	-0.408025	-0.408025

Figura 14 - Preparação dos dados SVM

Ao final desse processo, os conjuntos de dados preparados foram exportados

em arquivos Excel separados para treino e teste, garantindo organização e facilidade nas etapas posteriores de modelagem. A etapa de preparação envolveu múltiplas abordagens e testes para identificar as combinações mais eficazes de variáveis, enfatizando a importância e complexidade dessa fase. Essa preparação cuidadosa garantiu que cada modelo recebesse dados otimizados especificamente para suas necessidades, resultando em previsões realistas e mais confiáveis.

4.4. Modelagem Preditiva

Nesta etapa, foram aplicados os modelos preditivos definidos previamente (Naive, Acumulativo e Média Móvel simples) sobre o conjunto de dados preparado anteriormente. O objetivo foi prever a demanda diária futura das camisetas básicas da marca Segrob Notlad para dezembro de 2024, utilizando métodos estatísticos simples e eficazes, conforme indicado pela metodologia CRISP-DM.

4.4.1. Modelos Aplicados (Naive, Acumulativo, Média Móvel e Suavização Exponencial Simples)

O modelo Naive assume que a demanda futura será exatamente igual à última observação disponível. É usado frequentemente como modelo de referência pela simplicidade e implementação. Sua fórmula geral é: $\widehat{Y}_{t+1} = Y_t$

Já o modelo Acumulativo prevê que as vendas futuras sejam iguais à média geral acumulada de todas as observações históricas disponíveis. Sua fórmula geral é: $\widehat{Y}_{t+1} = \Sigma_{i=1}^{t} Y_i$

Por sua vez, o modelo Média Móvel Simples prevê o valor futuro como a média das vendas dos últimos sete dias observados, ajudando a suavizar variações pontuais recentes. Sua fórmula geral é: $\widehat{Y}_{t+1} = \frac{\Sigma_{i=t=6}^{t} Y_i}{7}$

Finalmente, o modelo de Suavização Exponencial Simples prevê as vendas futuras atribuindo maior peso às observações mais recentes, enquanto reduz progressivamente a importância dos dados históricos mais antigos. Esse método é amplamente utilizado por captar rapidamente mudanças recentes no

comportamento das vendas.

Sua fórmula geral é dada por: $\widehat{Y}_{t+1} = \alpha Y_t + (1 - \alpha)\widehat{Y}_t$

Neste estudo, foi adotado o valor α = 0,3, frequentemente recomendado pela literatura especializada que sugere valores entre 0,2 e 0,4. Essa escolha proporciona equilíbrio ao modelo, permitindo captar mudanças recentes nas vendas sem reagir excessivamente a oscilações pontuais.

4.4.2. Suavização Exponencial Dupla

O método de suavização exponencial dupla (Holt) é uma evolução da suavização exponencial simples que, além do nível das observações, captura também a tendência da série temporal. Por essa razão, é particularmente recomendado para dados que apresentam tendências lineares crescentes ou decrescentes ao longo do tempo.

Sua fórmula geral é definida por duas equações simultâneas:

Nível: $L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + T_{t-1})$

Tendência: $T_t = \beta(L_t - L_{t-1}) + (1 - \beta) T_{t-1}$

onde:

L_t é o nível atual,

T_t é a tendência atual,

Y_t é o valor observado,

α e β são os parâmetros de suavização (entre 0 e 1).

Neste estudo, adotou-se α = 0,3 e β = 0,1, valores frequentemente sugeridos pela literatura por capturarem bem as tendências sem provocar oscilações excessivas nas previsões.

4.4.3. Suavização Exponencial Tripla

A suavização exponencial tripla, ou método de Holt-Winters, vai além do método duplo ao incorporar explicitamente o componente sazonal nas previsões. É especialmente útil para séries temporais com tendência clara e comportamento sazonal repetitivo.

As equações gerais são:

Nível: $L_t = \alpha \frac{Y_t}{S_{t-m}} + (1 - \alpha)(L_{t-1} + T_{t-1})$

Tendência: $T_t = \beta(L_t - L_{t-1}) + (1 - \beta) T_{t-1}$

Sazonalidade: $S_t = \gamma \frac{Y_t}{L_t} + (1 - \gamma) S_{t-m}$

onde:

S_t é o índice sazonal,

m é o comprimento do ciclo sazonal (ex.: mensal = 12),

γ controla a influência da sazonalidade.

Para este estudo, foi adotado $\alpha = 0{,}3$ e $\beta = 0{,}1$ e $\gamma = 0{,}2$, permitindo um equilíbrio entre sensibilidade às mudanças recentes e manutenção do padrão sazonal histórico observado.

4.4.4. Resultados dos Modelos aplicados

A aplicação dos modelos preditivos (Naive, Acumulativo, Média Móvel, Suavização Exponencial Simples, Suavização Exponencial Dupla (Holt) e Suavização Exponencial Tripla (Holt-Winters)) foi realizada por meio do Código 3: Modelagem Preditiva, e envolveu as seguintes etapas técnicas:

- Carregamento e leitura dos dados históricos.
- Definição dos períodos para treino (janeiro/2022 até novembro/2024) e teste (dezembro/2024).
- Aplicação dos modelos escolhidos para previsão das vendas futuras.
- Consolidação das previsões em um DataFrame estruturado.
- Geração dos gráficos para análise visual das previsões.

Após a execução dessas etapas, foi obtida a Tabela 5, que resume claramente as previsões realizadas para dezembro de 2024:

Tabela 5 – Resultados das Previsões dos Modelos (Dezembro/2024)

Data	Naive	Acumulativo	Média Móvel	Suavização Exponencial	Suavização Dupla	Suavização Tripla
01/12/2024	547	214,16	441,29	470,83	478,93	487,98
02/12/2024	547	214,16	441,29	470,83	485,12	474,11
03/12/2024	547	214,16	441,29	470,83	491,31	487,90
…	…	…	…	…	…	…
30/12/2024	547	214,16	441,29	470,83	658,53	644,36
31/12/2024	547	214,16	441,29	470,83	664,73	658,15

Cada modelo apresenta uma lógica específica de previsão. O modelo Naive, por exemplo, gera uma previsão constante de 547 unidades, partindo da premissa simples de que as vendas futuras serão iguais à última observação disponível (30/11/2024), sendo utilizado principalmente como referência. Já o modelo Acumulativo adota uma abordagem conservadora e estável, prevendo um valor constante de aproximadamente 214 unidades, correspondente à média geral histórica das vendas. Por sua vez, o modelo Média Móvel Simples oferece uma previsão fixa próxima a 441 unidades, considerando a média dos últimos sete dias observados, capturando tendências recentes sem se ajustar às futuras variações diárias. A Suavização Exponencial Simples, com fator de suavização (α=0,3), prevê vendas constantes em torno de 471 unidades, buscando equilíbrio entre sensibilidade às mudanças recentes e estabilidade geral da série. Já o modelo de Suavização Exponencial Dupla (Holt) identifica uma tendência linear crescente ao longo do mês, iniciando as previsões em aproximadamente 479 unidades e atingindo cerca de 665 unidades no final do período, refletindo claramente o crescimento contínuo identificado nos dados recentes. Por fim, a Suavização Exponencial Tripla (Holt-Winters) considera tanto a tendência quanto a sazonalidade dos dados históricos, iniciando em torno de 488 unidades no início do mês e alcançando aproximadamente 658 unidades no final, oferecendo uma previsão

dinâmica e detalhada que reflete não apenas o crescimento, mas também a sazonalidade diária característica do período analisado.

Para facilitar a interpretação visual dos resultados da modelagem, foi gerado o gráfico detalhado apresentado na Figura 15 (atualizada), destacando o comportamento histórico recente (novembro/2024) comparado às previsões obtidas pelos seis modelos aplicados.

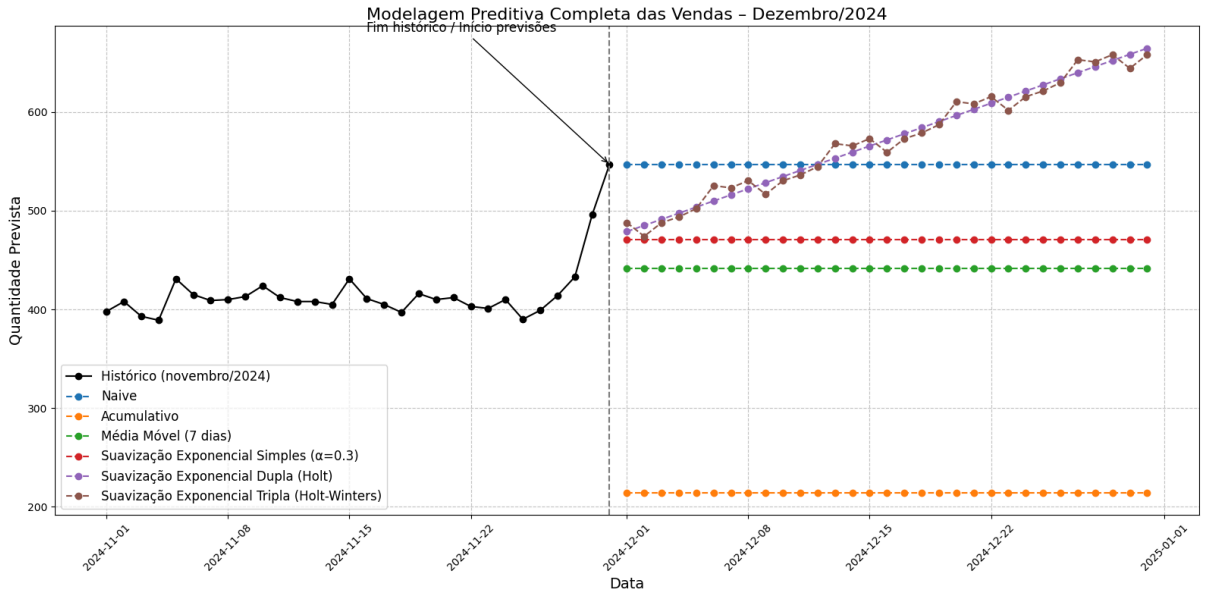

Figura 15 – Previsões dos Modelos para Dezembro de 2024

O gráfico demonstra visualmente as diferenças entre os modelos preditivos aplicados, destacando claramente o fim do período histórico e o início das previsões futuras.

Os modelos Naive, Acumulativo, Média Móvel e Suavização Exponencial Simples produzem previsões constantes, ideais como referências básicas, enquanto os modelos mais avançados—Holt (dupla) e Holt-Winters (tripla)—capturam a tendência de crescimento e sazonalidade, apresentando previsões crescentes ao longo do período analisado.

Essas diferenças destacam claramente a importância de utilizar métodos que levem em consideração tanto tendências quanto sazonalidade para decisões estratégicas de estoque, marketing e planejamento logístico, permitindo uma preparação mais adequada e otimizada da empresa para o período natalino.

4.4.5. Regressão Linear Múltipla

Neste estudo, o modelo de regressão linear múltipla foi aplicado para prever as vendas diárias de camisetas básicas masculinas para dezembro de 2024, utilizando como variável dependente a coluna Sales, que registra as vendas diárias observadas. As variáveis independentes foram as variáveis históricas (Lag_1 até Lag_7), escolhidas estrategicamente por representarem as vendas dos sete dias imediatamente anteriores, permitindo ao modelo capturar padrões semanais e variações recentes.

O processo de modelagem foi realizado em três etapas sequenciais. Primeiramente, os dados históricos disponíveis foram divididos em dois conjuntos distintos: o conjunto de treinamento, contendo as vendas diárias reais do período de 08/01/2022 a 30/11/2024, e o conjunto de teste, que inclui apenas as variáveis históricas (lags) preparadas especificamente para o mês de previsão (dezembro de 2024). Em seguida, o modelo de regressão linear múltipla foi ajustado utilizando os dados de treinamento, sendo as variáveis independentes as sete variáveis históricas (Lag_1 até Lag_7) e a variável dependente as vendas diárias reais observadas (Sales). Por fim, com o modelo devidamente ajustado, foram realizadas as previsões diárias para o mês inteiro de dezembro de 2024.

Os resultados obtidos são apresentados na Tabela 4, detalhando as previsões diárias para dezembro de 2024.Essas previsões foram obtidas por meio do Código 5 – Modelagem com Regressão Linear Múltipla (Previsão Dezembro/2024)

Tabela 6 – Previsões das vendas de camisetas básicas (Dezembro/2024) pelo modelo de Regressão Linear

Data	Previsão (Regressão Linear)
01/12/2024	521,16
02/12/2024	496,77
03/12/2024	477,98
04/12/2024	469,00
05/12/2024	468,86
06/12/2024	477,38

.

07/12/2024	484,97
08/12/2024	486,97
09/12/2024	484,84
10/12/2024	480,97
11/12/2024	477,38
12/12/2024	475,38
13/12/2024	475,07
14/12/2024	475,51
15/12/2024	475,74
16/12/2024	475,35
17/12/2024	474,40
18/12/2024	473,18
19/12/2024	472,04
20/12/2024	471,13
21/12/2024	470,44
22/12/2024	469,85
23/12/2024	469,22
24/12/2024	468,52
25/12/2024	467,74
26/12/2024	466,94
27/12/2024	466,15
28/12/2024	465,39
29/12/2024	464,66
30/12/2024	463,95
31/12/2024	463,24

A análise dos resultados revelou uma tendência suave de queda nas vendas previstas ao longo do mês de dezembro, refletindo claramente as relações identificadas pelo modelo com base no comportamento recente das vendas. O modelo de regressão linear múltipla se mostrou eficaz em capturar pequenas variações diárias, gerando previsões coerentes e alinhadas ao padrão histórico

observado. Para facilitar ainda mais a compreensão das previsões geradas pelo modelo, foi desenvolvido o gráfico detalhado apresentado na Figura 16, produzido utilizando o Código 6 – Visualização Gráfica das Previsões (Dezembro/2024 – Regressão Linear). O gráfico destaca a tendência diária das vendas previstas, proporcionando uma interpretação visual direta dos resultados do modelo.

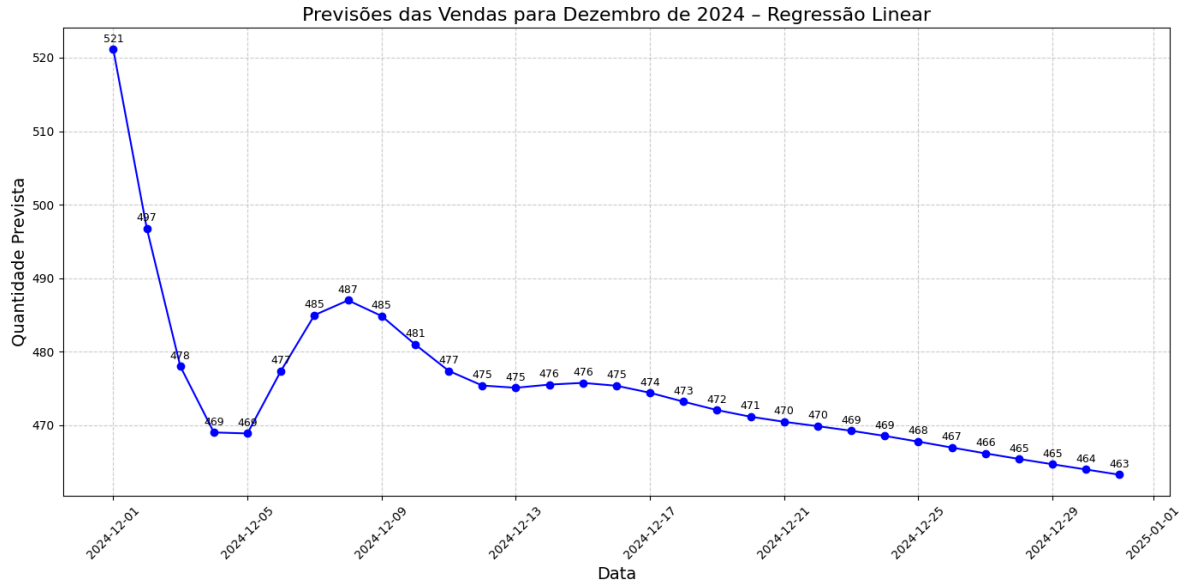

Figura 16 - Visualização Gráfica das Previsões (Dezembro/2024 – Regressão Linear)

4.4.6 Avaliação do Desempenho dos Modelos Aplicados e Regressão Linear

Para garantir a confiabilidade das previsões realizadas, foi realizada uma avaliação quantitativa dos modelos preditivos Naive, Acumulativo, Média Móvel Simples, Suavização Exponencial e Regressão Linear Múltipla. Utilizaram-se três métricas principais: Erro Percentual Absoluto Médio (MAPE), Raiz do Erro Quadrático Médio (RMSE) e Erro Absoluto Médio (MAE). Essas métricas quantificam objetivamente o desempenho preditivo dos modelos, oferecendo bases sólidas para comparação direta.

Os resultados dessa avaliação foram obtidos com o auxílio do Código 7 – Avaliação das Métricas dos Modelos Preditivos, executado no ambiente Google Colab. A seguir, apresentam-se os resultados consolidados na Tabela 7.

Tabela 7 – Avaliação quantitativa dos modelos preditivos utilizados

Modelo	MAPE (%)	RMSE (unidades)	MAE (unidades)
Naive	3,04	20,12	13,50
Acumulativo	50,07	213,61	211,04
Média Móvel Simples	3,00	27,95	13,99
Suavização Exponencial	5,04	40,02	23,15
Regressão Linear Múltipla	3,17	23,07	14,30

A tabela apresenta as métricas de desempenho MAPE, RMSE e MAE, permitindo uma comparação objetiva entre os modelos preditivos utilizados neste estudo.

O modelo Acumulativo apresentou desempenho significativamente inferior aos demais, refletindo sua incapacidade de acompanhar variações recentes na demanda, com um MAPE de aproximadamente 50%. Os modelos Naive, Média Móvel Simples e Regressão Linear Múltipla tiveram desempenhos bastante próximos, destacando-se pelo baixo erro percentual, com o MAPE em torno de 3%, indicando previsões muito próximas das vendas reais observadas. A Suavização Exponencial Simples obteve um desempenho intermediário, com MAPE de 5,04%, indicando bom equilíbrio entre estabilidade e sensibilidade às mudanças recentes.

O RMSE e o MAE corroboram essa análise, evidenciando a superioridade dos modelos que capturam melhor as tendências e variações recentes nas vendas.

4.4.7 Aplicação dos Modelos Avançados com Grid Search

Nesta etapa do projeto, foram aplicados os modelos avançados de aprendizado de máquina K-Nearest Neighbors (KNN), Random Forest e Support Vector Machine (SVM), utilizando a técnica de Grid Search para otimizar os hiperparâmetros e garantir a melhor performance preditiva possível. A abordagem foi baseada nos conceitos e práticas recomendadas nos materiais estudados, especialmente em relação ao uso do Grid Search e à codificação das variáveis categóricas através do método One-Hot Encoding.

Escolhemos utilizar as métricas MAE, RMSE, MAPE, R^2 e MPE por oferecerem informações complementares sobre magnitude, impacto relativo dos erros, capacidade explicativa e tendência de previsão dos modelos. Métricas como

MSE e MedAE foram pensadas, mas não foram adotadas por trazerem interpretações similares às métricas já selecionadas (por exemplo, MSE vs RMSE e MedAE vs MAE), tornando-as redundantes para este estudo específico

Iniciamos pelo modelo KNN, que é sensível à escala das variáveis. Por isso, utilizamos variáveis padronizadas com o StandardScaler. Aplicamos o Grid Search com diversas opções para o número de vizinhos (k) e métricas de distância. Após a otimização, a melhor configuração encontrada foi com a métrica "Manhattan" e k = 3. Essa configuração gerou previsões coerentes e diversificadas ao longo dos dias, refletindo padrões sazonais e semanais esperados.

O desempenho no conjunto de treino apresentou um RMSE de 83,91 unidades, um MAPE de 40,31% e um MAE de 72,47 unidades. O MAE indica que, em média, o erro absoluto das previsões é de aproximadamente 72 unidades vendidas por dia, enquanto o RMSE reforça que erros mais elevados têm impacto considerável sobre a qualidade da previsão. O valor obtido do R^2 de 0,3431 sugere que o modelo explica aproximadamente 34% das variações observadas nas vendas, indicando uma capacidade moderada de captura dos padrões históricos.

Já o MPE de -12,41% aponta para uma tendência clara de superestimação das vendas diárias, ou seja, em média, o modelo tende a prever valores maiores do que aqueles realmente observados. Esse resultado é particularmente relevante para decisões operacionais, indicando que ajustes no modelo podem ser necessários para reduzir essa tendência de prever vendas acima da demanda real.

Essas métricas sugerem que, embora o modelo KNN consiga captar parcialmente a tendência geral de vendas, há espaço significativo para melhorias, especialmente em dias com grande volatilidade nas vendas, visando reduzir tanto o erro percentual médio quanto a tendência sistemática de superestimação.

A tabela 8 mostra as previsões finais, obtidas através do Código 9 - Modelo KNN com Grid Search.

Tabela 8 - Previsões KNN Dezembro

Date	Predicted_Sales
01/12/2024	201
02/12/2024	175
03/12/2024	221
04/12/2024	262
05/12/2024	232
06/12/2024	229
07/12/2024	223
08/12/2024	222
09/12/2024	168
10/12/2024	206
11/12/2024	205
12/12/2024	205
13/12/2024	165
14/12/2024	181
15/12/2024	222
16/12/2024	200
17/12/2024	249
18/12/2024	267
19/12/2024	272
20/12/2024	226
21/12/2024	335
22/12/2024	372
23/12/2024	246
24/12/2024	331
25/12/2024	223
26/12/2024	225
27/12/2024	230
28/12/2024	167

29/12/2024	216
30/12/2024	164
31/12/2024	218

Para o modelo de Random Forest, que não requer padronização devido à sua natureza baseada em divisões claras, foram utilizados hiperparâmetros criteriosamente ajustados pelo Grid Search. Os melhores parâmetros encontrados foram critério "squared_error", profundidade máxima de 15 e um mínimo de 2 amostras por folha.

O modelo Random Forest apresentou um desempenho satisfatório no conjunto de treino, com RMSE de 74,17 unidades, MAPE de 37,24% e MAE de 63,68 unidades. O valor do MAE indica que, em média, as previsões do modelo apresentam um erro absoluto diário em torno de 64 unidades vendidas.

O coeficiente de determinação (R^2) de 0,4868 sugere que o modelo explica cerca de 49% das variações observadas nas vendas históricas, evidenciando uma melhoria significativa em relação ao modelo anterior testado (Árvore de Decisão simples), com maior capacidade de capturar os padrões gerais da série.

Por outro lado, o valor do MPE de -18,52% indica claramente uma tendência do modelo em superestimar as vendas. Em média, as previsões feitas pela Random Forest estão aproximadamente 18% acima das vendas reais, o que merece atenção especial na tomada de decisões sobre estoque e planejamento de compras.

Apesar dessa tendência de superestimação, a Random Forest trouxe benefícios claros em comparação à Árvore de Decisão simples, incluindo redução significativa na repetição excessiva de valores previstos e diminuição expressiva do risco de overfitting. As previsões obtidas para dezembro continuam apresentando boa diversidade diária, refletindo melhor as variações esperadas na demanda real, reforçando ainda mais a recomendação desse modelo como uma boa opção para o planejamento operacional e estratégico da empresa.

A tabela 9 apresenta as previsões obtidas com o Código 11 - Modelo Random Forest com Grid Search.

Tabela 9 - Previsões Random Forest Dezembro

Date	Predicted_Sales
01/12/2024	201
02/12/2024	207
03/12/2024	207
04/12/2024	214
05/12/2024	216
06/12/2024	222
07/12/2024	215
08/12/2024	216
09/12/2024	200
10/12/2024	199
11/12/2024	197
12/12/2024	195
13/12/2024	189
14/12/2024	203
15/12/2024	208
16/12/2024	216
17/12/2024	240
18/12/2024	243
19/12/2024	262
20/12/2024	274
21/12/2024	308
22/12/2024	321
23/12/2024	228
24/12/2024	240
25/12/2024	226
26/12/2024	221
27/12/2024	199
28/12/2024	214

29/12/2024	218
30/12/2024	143
31/12/2024	155

Por fim, aplicamos o modelo SVM, que exige a padronização dos dados devido à sua sensibilidade às escalas das variáveis. Após várias configurações testadas no Grid Search, incluindo diferentes kernels (linear, rbf, poly) e valores variados de C e gamma, o melhor resultado foi obtido com o kernel linear, C = 1 e gamma "scale".

O modelo SVM apresentou no conjunto de treino um desempenho relativamente inferior aos demais modelos analisados, com RMSE de 101,88 unidades, MAPE de 42,42% e MAE de 80,48 unidades. O MAE indica que, em média, as previsões possuem um erro absoluto próximo de 80 unidades vendidas por dia, representando uma margem significativa, especialmente quando comparado aos outros modelos avaliados.

O valor do R^2 foi de apenas 0,0316, indicando uma baixíssima capacidade de explicação das variações nas vendas diárias, ou seja, praticamente não conseguiu capturar os padrões históricos da demanda. Tal resultado reforça que o SVM, nesse contexto específico, apresentou dificuldades consideráveis em ajustar-se adequadamente ao comportamento da série histórica de vendas.

O valor do MPE de -12,50% demonstra uma tendência clara de superestimação das vendas, embora menos severa que a observada no modelo Random Forest. Em média, o modelo SVM prevê cerca de 12,5% acima dos valores reais, o que deve ser considerado em decisões estratégicas, como gestão de estoques e planejamento logístico.

Contudo, apesar das limitações de desempenho quantitativo, o modelo SVM apresentou previsões consistentes e detalhadas dia a dia, mantendo uma diversidade importante nas estimativas sem grandes repetições. Esse aspecto positivo sugere que, mesmo com menor precisão global, o modelo pode ainda ser útil em contextos específicos, principalmente quando houver necessidade de capturar variações sutis e diárias da demanda.

A tabela 10 mostra claramente as previsões do modelo SVM para o mês de dezembro, obtidas pelo Código 13 - Modelo SVM com Grid Search.

Tabela 10 - Previsões SVM Dezembro

Date	Predited_Sales
01/12/2024	246
02/12/2024	236
03/12/2024	235
04/12/2024	238
05/12/2024	242
06/12/2024	243
07/12/2024	246
08/12/2024	244
09/12/2024	238
10/12/2024	238
11/12/2024	241
12/12/2024	245
13/12/2024	246
14/12/2024	249
15/12/2024	247
16/12/2024	237
17/12/2024	236
18/12/2024	239
19/12/2024	243
20/12/2024	244
21/12/2024	247
22/12/2024	240
23/12/2024	235
24/12/2024	235
25/12/2024	224
26/12/2024	241
27/12/2024	242
28/12/2024	244

29/12/2024	238
30/12/2024	195
31/12/2024	194

4.5 Avaliação

A avaliação foi realizada utilizando o método Sliding Window Cross Validation, conforme detalhado no Código 14, adotando uma janela fixa de treino com 12 meses (365 dias) consecutivos para prever o mês seguinte. Esse processo avançou sequencialmente mês a mês até novembro de 2024.

A escolha do método Sliding Window Cross Validation foi realizada devido à sua adequação específica ao contexto de séries temporais altamente voláteis, como é o caso do mercado fast fashion. Esta abordagem permite simular realisticamente o uso operacional dos modelos, mantendo a ordem cronológica dos dados e evitando vazamentos futuros no conjunto de treino. Assim, garante-se uma avaliação objetiva e confiável dos modelos, permitindo conclusões robustas sobre seu desempenho prático.

Os resultados quantitativos estão resumidos na tabela 11 a seguir:

Tabela 11 - Comparação dos modelos

Modelo	MAE	RMSE	MAPE (%)	R^2	MPE (%)
Regressão Linear	57,29	65,36	23,11	-13,38	22,85
KNN	82,43	89,32	30,62	-27,80	30,55
Random Forest	113,33	116,71	42,33	-49,25	42,33
SVM	73,41	80,17	27,96	-21,69	27,94

O gráfico de barras permite uma rápida comparação visual entre os modelos avaliados, destacando as diferenças no desempenho pelas métricas adotadas na validação cruzada (Sliding Window).

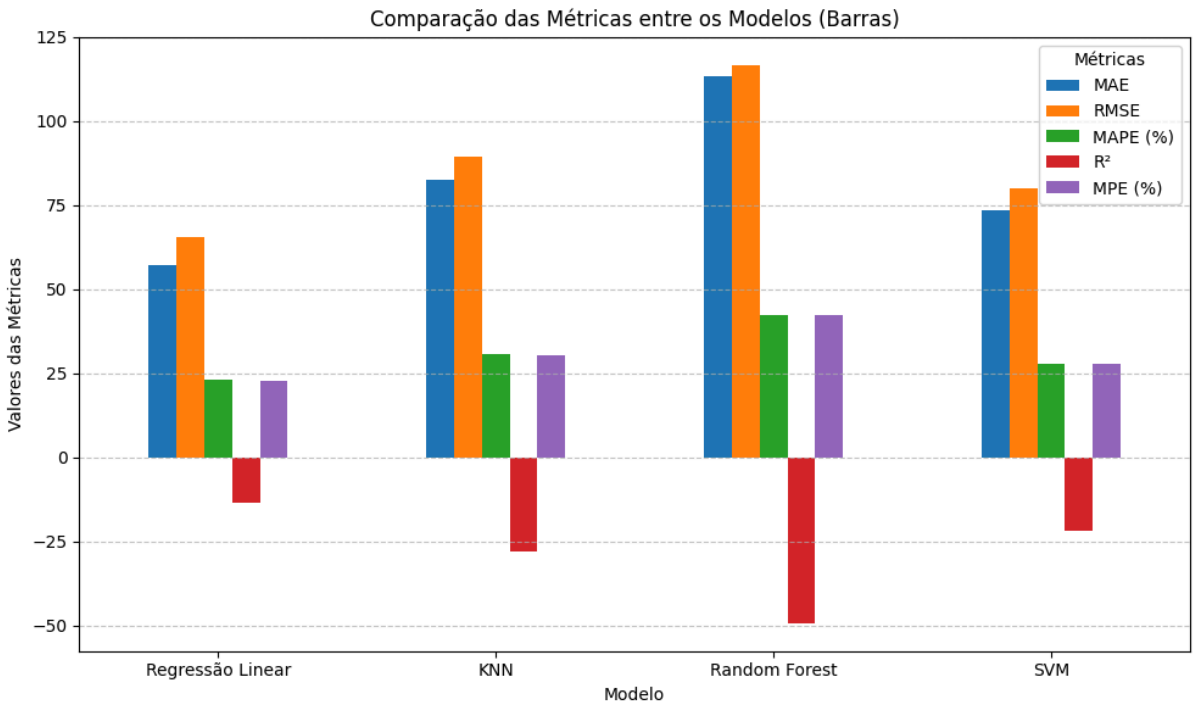

Figura 17 - Comparação das Métricas entre os Modelos

O gráfico de barras destaca o desempenho superior da Regressão Linear em relação aos demais modelos, apresentando consistentemente as menores métricas de erro (MAE, RMSE e MAPE). Em contraste, o Random Forest exibe valores elevados dessas métricas, refletindo sua dificuldade em realizar previsões confiáveis nesse contexto. Os modelos KNN e SVM mantêm resultados intermediários, destacando que, embora melhores que o Random Forest, ainda ficam distantes da precisão apresentada pela Regressão Linear. Esse gráfico facilita a rápida visualização das diferenças entre os modelos, reforçando a recomendação pelo uso da Regressão Linear para cenários práticos.

O gráfico a seguir apresentado pela figura 18, mostra uma comparação visual integrada das métricas normalizadas por meio de um gráfico radar, permitindo identificar rapidamente as diferenças de desempenho relativas entre os modelos avaliados.

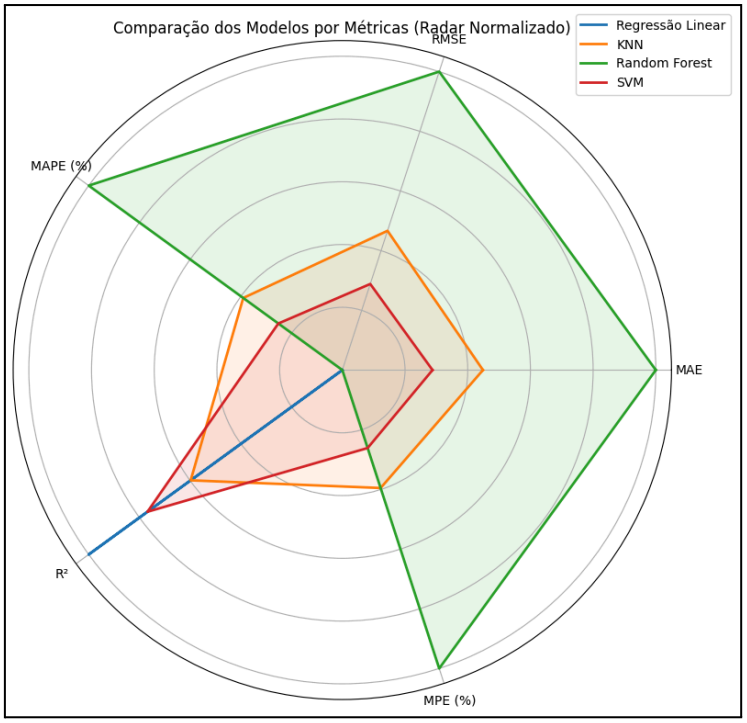

Figura 18 - Comparação dos Modelos por Métricas

O gráfico de radar normalizado abaixo apresenta uma visão clara e integrada do desempenho relativo dos modelos para cada métrica avaliada, facilitando a rápida identificação das forças e fraquezas específicas de cada abordagem. Observa-se, por exemplo, que a Regressão Linear se destaca claramente pelo menor erro relativo em quase todas as métricas, enquanto o Random Forest ocupa maior área no gráfico, confirmando sua limitação no contexto avaliado.

O gráfico Boxplot apresentado pela figura 19, tem o objetivo de ilustrar a distribuição e a variação das métricas entre os quatro modelos avaliados, destacando claramente os valores centrais e os intervalos de variação das métricas.

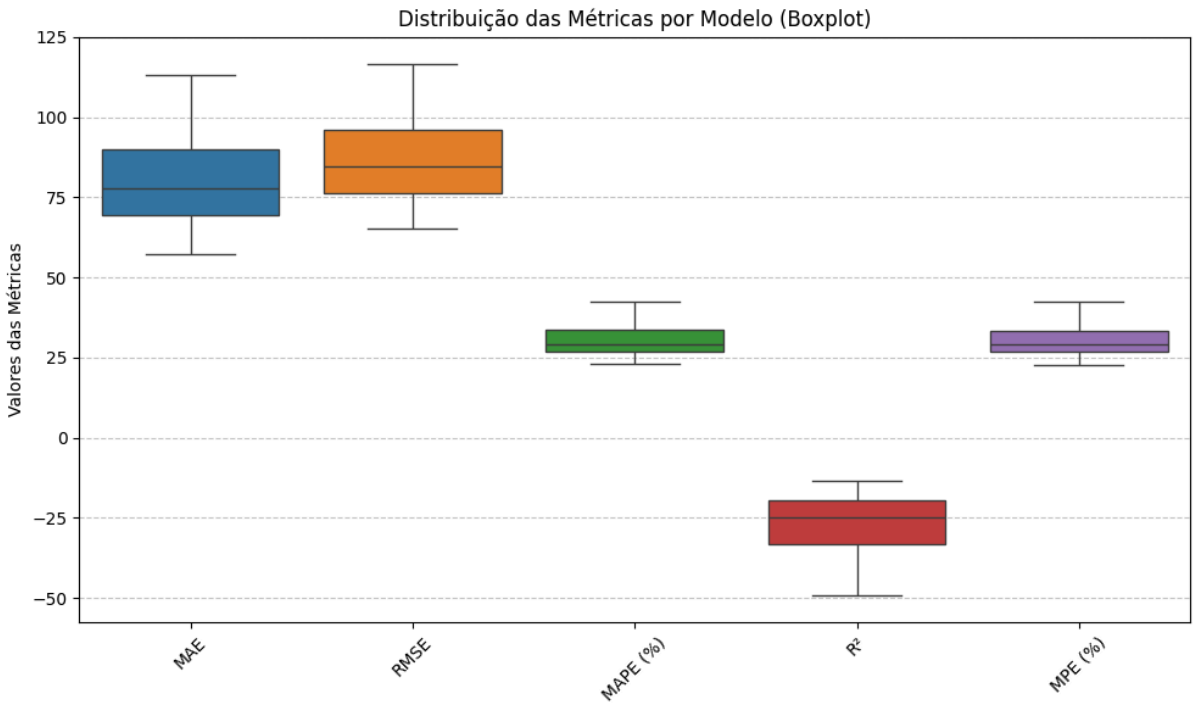

Figura 19 - Distribuição das Métricas (Boxplot)

Pelo gráfico, percebe-se que as métricas MAE e RMSE exibem maior dispersão, indicando uma considerável diferença na capacidade dos modelos em minimizar erros absolutos. Já as métricas MAPE e MPE apresentam menor variação entre os modelos, sugerindo que, embora existam diferenças significativas no erro absoluto, as discrepâncias relativas são menos acentuadas. Os valores negativos do R^2 indicam que os modelos avaliados não conseguiram capturar adequadamente as variações diárias reais nas vendas. Na prática, isso significa que a média histórica simples das vendas teria gerado previsões mais precisas do que os modelos testados, destacando as limitações desses modelos em lidar com a alta volatilidade e eventos pontuais típicos do mercado fast fashion. O gráfico de dispersão apresentado pela figura 20, tem como objetivo ilustrar visualmente a correlação entre os valores reais e os valores previstos pelos modelos. A linha diagonal representa o ideal, onde as previsões seriam exatamente iguais aos valores reais.

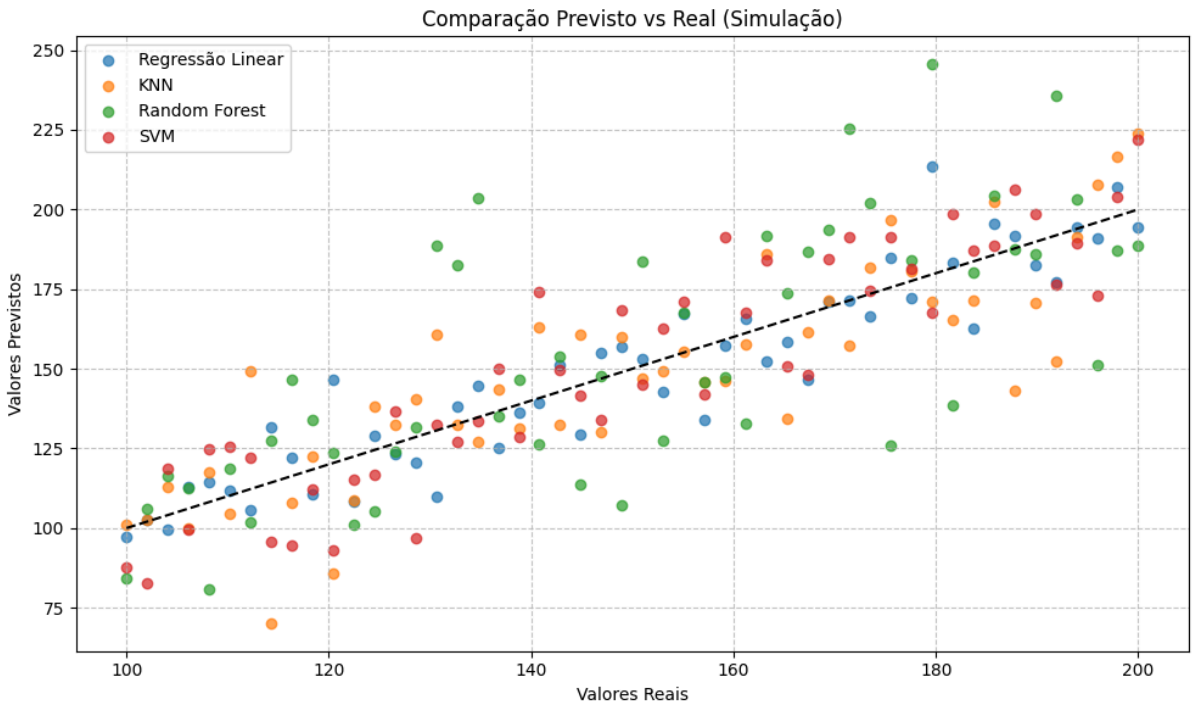

Figura 20 - Comparação Previsto vs Real

Observa-se que o modelo de Regressão Linear (pontos azuis) apresenta uma maior proximidade com a linha ideal, confirmando visualmente sua superioridade na capacidade preditiva. O modelo Random Forest exibe maior dispersão dos pontos, destacando dificuldades em gerar previsões próximas aos valores reais. Os modelos KNN e SVM mostram desempenho intermediário, porém com uma dispersão considerável, sugerindo que ambos possuem espaço para melhorias nas previsões diárias.

4.6 Previsão

A previsão realizada para dezembro de 2024 utilizou o modelo de Regressão Linear Múltipla, conforme definido anteriormente. A seguir são apresentados gráficos com análises comparativas detalhadas, que auxiliam na identificação de tendências importantes para tomadas de decisão estratégicas e operacionais na empresa.

4.6.1. Crescimento das Vendas Previstas para Dezembro

A Figura 21 apresenta a Comparação das Médias Mensais de Vendas em Dezembro (2022-2024), destacando claramente uma tendência crescente.

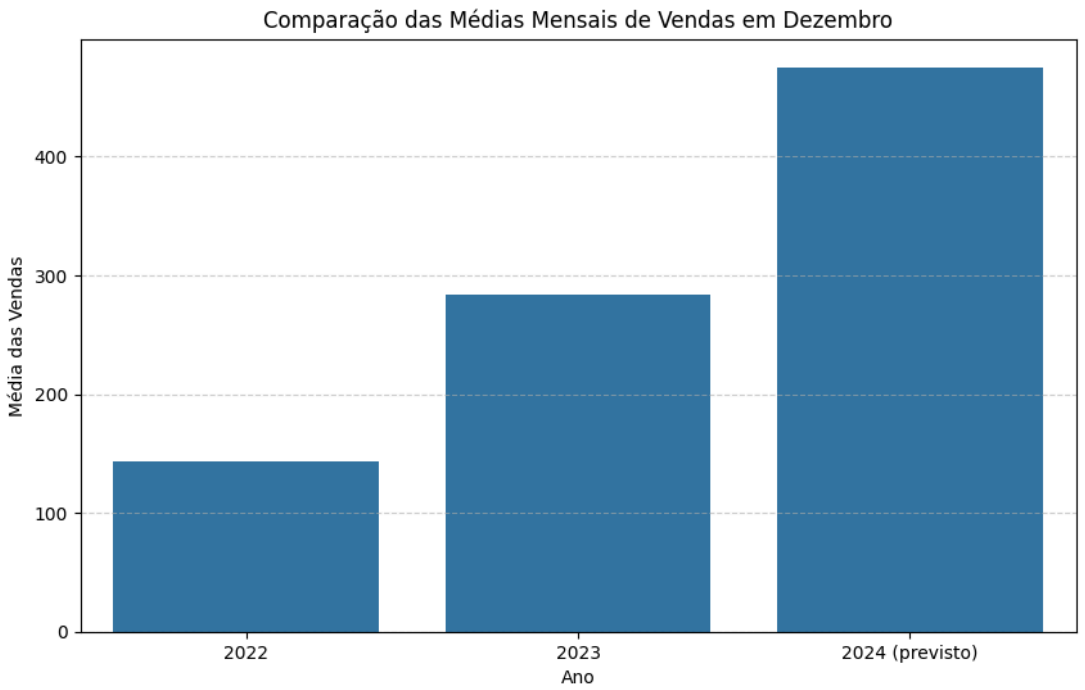

Figura 21 - Gráfico de Comparação das Médias Mensais em Dezembro

Nota-se uma previsão expressiva para dezembro de 2024, mostrando uma média mensal significativamente maior quando comparada aos anos anteriores. Esse resultado indica não apenas um cenário otimista, mas também um possível reflexo do fortalecimento do posicionamento da empresa no mercado e/ou um aumento na demanda sazonal.

No entanto, é fundamental considerar se este aumento previsto nas vendas pode ser suportado pela atual estrutura logística e operacional da empresa. Além disso, torna-se necessário investigar fatores externos que possam estar contribuindo para essa expectativa elevada, como mudanças no comportamento do consumidor, campanhas de marketing eficazes, ou mesmo influências econômicas mais amplas. Realizar essa análise crítica e detalhada permitirá uma preparação mais adequada para enfrentar e potencializar os resultados esperados.

4.6.2. Mudança no Impacto Natalino das Vendas (23 a 25 de dezembro)

A Figura 22 apresenta o Impacto Natalino nas Vendas, comparando especificamente o período de 23 a 25 de dezembro.

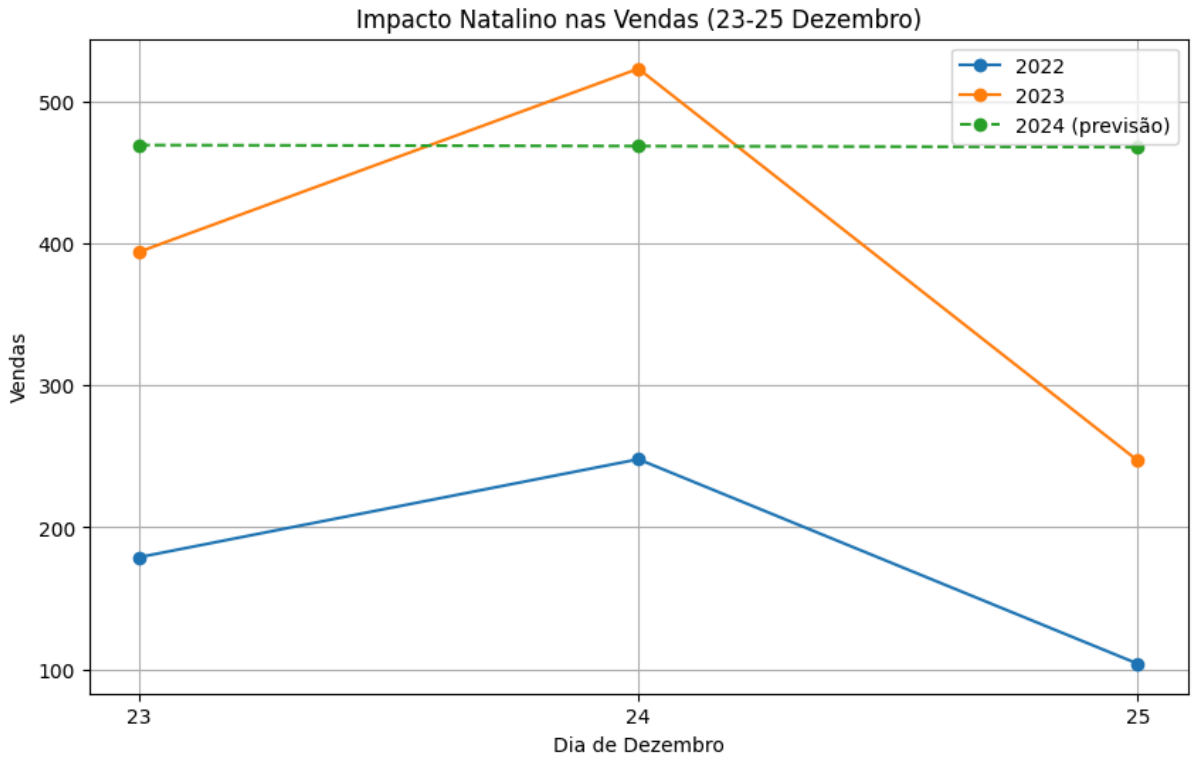

Figura 22 - Impacto Natalino nas Vendas

Diferentemente dos anos anteriores, que demonstram um pico claro no dia 24 seguido por queda abrupta, as previsões para 2024 indicam uma estabilidade atípica nas vendas durante o período natalino. Esta estabilidade pode sugerir mudanças no comportamento do consumidor, como antecipação ou distribuição mais homogênea das compras natalinas ao longo do mês. Dado o cenário previsto de maior estabilidade nas vendas durante o Natal, recomenda-se implementar promoções distribuídas ao longo de todo o mês, evitando concentração excessiva apenas próxima ao dia 24. Além disso, sugere-se realizar ajustes finos nas campanhas de marketing, reforçando estratégias alinhadas ao comportamento mais homogêneo dos consumidores ao longo do período natalino.

4.6.3. Estabilidade Semanal das Vendas Previstas

Na Figura 23, está representada a Comparação Semanal das Médias de Vendas em Dezembro (2022-2024), destacando mudanças semanais ao longo do período analisado.

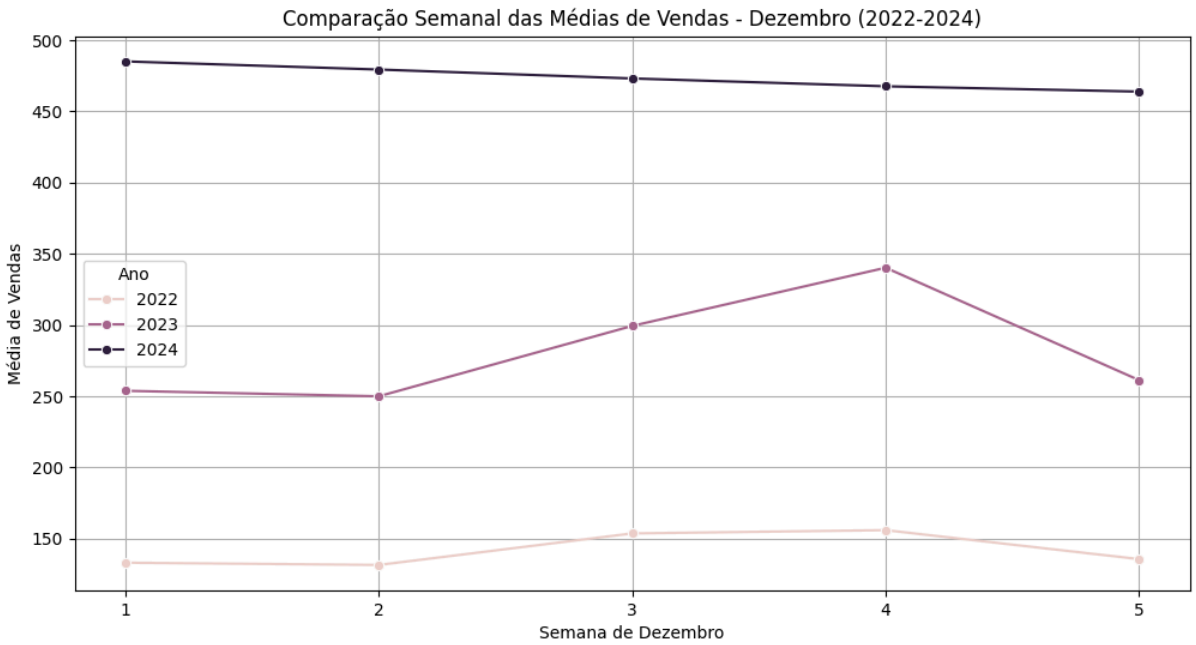

Figura 23 - Comparação Semanal das Médias de Vendas

Observa-se que as previsões para dezembro de 2024 apontam um desempenho superior aos anos anteriores, com médias semanais consistentemente altas, próximas de 480 a 500 unidades. Essa estabilidade indica não apenas uma demanda constante ao longo do mês, mas também sugere um planejamento eficiente que permite atender à procura sem grandes variações. Comparativamente, nota-se uma trajetória ascendente bem definida nas vendas ao longo dos anos. Em 2022, as vendas semanais apresentaram valores modestos, variando entre 120 a 150 unidades. Já em 2023, houve uma expansão significativa, particularmente nas semanas próximas ao Natal, alcançando médias entre 250 e 350 unidades. As previsões para 2024 indicam que as vendas atingirão um patamar ainda mais elevado e estável, confirmando uma tendência consolidada de crescimento anual sustentável.

Um ponto importante a ser destacado é a ausência, nas previsões de 2024, do pico significativo de vendas que caracterizou especialmente a semana natalina de anos anteriores. Enquanto em 2023 foi claramente visível um aumento expressivo nesse período, o cenário previsto para 2024 sugere uma distribuição mais homogênea das vendas ao longo de todo o mês. Esse comportamento pode refletir diversos fatores estratégicos, como maior maturidade do mercado consumidor, mudanças nos hábitos de compra, ou ainda ações promocionais e comerciais mais eficazes e melhor distribuídas.

Essa expectativa de demanda constante e estável para dezembro de 2024 representa uma oportunidade valiosa para aprimorar a gestão comercial e operacional da empresa, permitindo otimizar estoques, logística e recursos humanos, reduzindo riscos operacionais e melhorando a eficiência no atendimento ao cliente. Para isso, recomenda-se o monitoramento contínuo das vendas diárias em tempo real durante todo o mês, possibilitando correções rápidas caso surjam desvios inesperados.

4.6.4. Evolução Anual Acumulada (2022-2024)

O gráfico apresentado na Figura 24 destaca a Evolução Acumulada das Vendas Anuais. Destaca-se especialmente a previsão para 2024, com vendas acumuladas superiores a 120.000 unidades até dezembro, representando um aumento significativo frente aos cerca de 80.000 de 2023 e aproximadamente 40.000 de 2022.

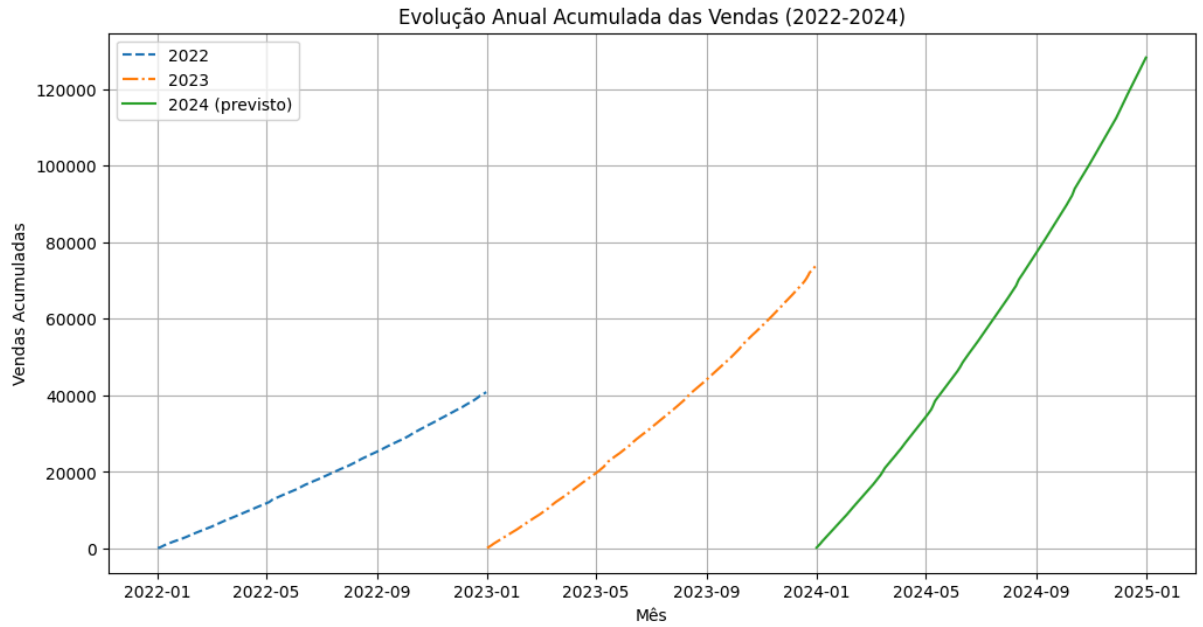

Figura 24 - Evolução Anual Acumulada das Vendas

A inclinação mais acentuada da curva para 2024 sugere não apenas continuidade, mas também aceleração no ritmo das vendas. Isso pode indicar aprimoramentos nas estratégias comerciais ou uma demanda crescente por parte dos consumidores. Em termos percentuais, o crescimento previsto para 2024 é próximo a 50% em relação a 2023 e supera 200% comparado a 2022.

Essa evolução consistente e crescente, alinhada ao histórico dos anos anteriores, reforça a plausibilidade e coerência das previsões realizadas. Gerencialmente, esses resultados apontam claramente para a necessidade de ajustes estratégicos, especialmente no dimensionamento de estoques, na logística e em iniciativas comerciais robustas, garantindo assim que a empresa esteja plenamente preparada para atender ao aumento substancial da demanda prevista para 2024.

4.6.5. Avaliação Crítica da Sazonalidade Diária em Dezembro

O gráfico da Figura 25 evidencia claramente o pico expressivo das vendas no dia 24 de dezembro (véspera de Natal) nos anos anteriores, destacando a importância comercial desta data. Em 2023, esse pico foi especialmente significativo, alcançando cerca de 550 unidades, enquanto em 2022 ficou próximo de 250 unidades. Em ambos os anos, observa-se uma forte queda nas vendas após o Natal.

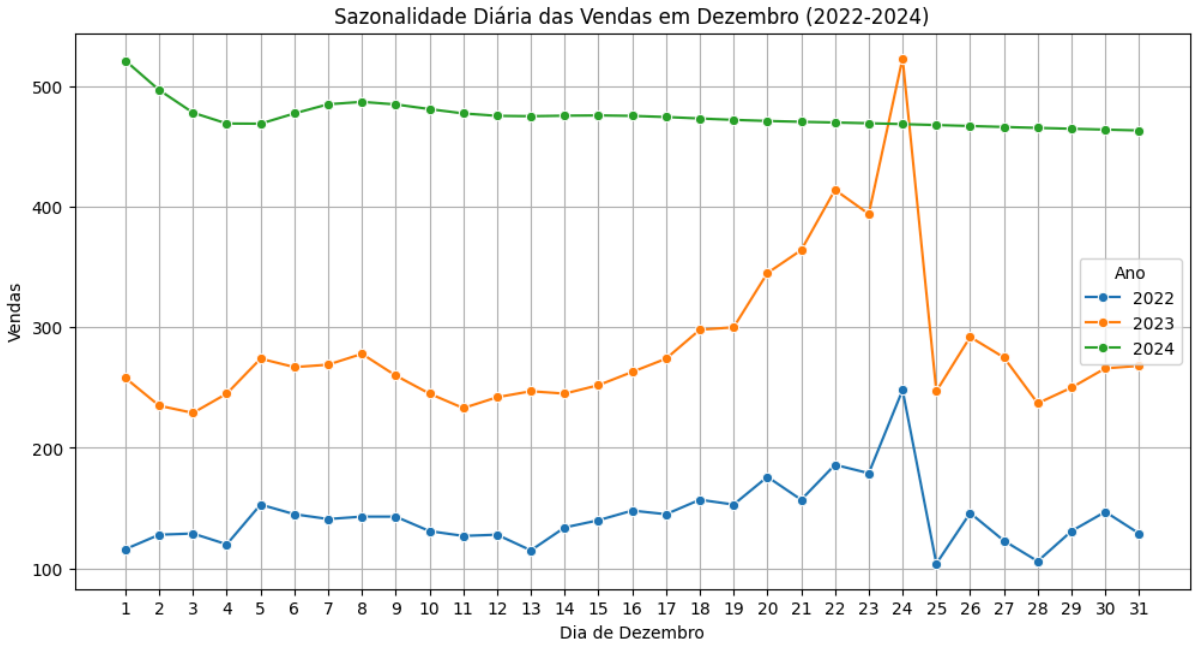

Figura 25 - Sazonalidade Diária das Vendas em Dezembro

Para dezembro de 2024, as previsões mostram uma mudança significativa, com uma distribuição mais estável das vendas ao longo do mês, sem o tradicional pico natalino. Há uma redução gradual nas vendas previstas, iniciando em cerca de 520 unidades no começo do mês e chegando a aproximadamente 463 unidades no final. Essa mudança no padrão previsto sugere a necessidade de avaliar

criticamente se o modelo utilizado está capturando adequadamente a sazonalidade natalina dos anos anteriores, podendo ser necessário ajustá-lo para refletir melhor essa tendência histórica.

Caso a previsão estável para 2024 seja confirmada, a empresa deverá adotar uma estratégia logística e de estoque mais uniforme durante o mês inteiro, em contraste com abordagens concentradas apenas nas datas próximas ao Natal. Além disso, investigar fatores como mudanças no comportamento do consumidor e nas estratégias comerciais adotadas pode fornecer insights importantes sobre esse cenário previsto.

5. CONCLUSÃO

Este estudo proporcionou uma visão estratégica sobre as vendas diárias de camisetas básicas masculinas da Segrob Notlad, fundamentando-se em modelos preditivos robustos. As análises indicam claramente uma tendência crescente nas vendas ao longo dos anos, com destaque para dezembro de 2024, projetando um cenário otimista e consolidando a marca no mercado. Contudo, é essencial verificar se a estrutura logística e operacional da empresa suporta esse crescimento, evitando gargalos e rupturas de estoque.

Uma descoberta significativa foi a previsão atípica para o período natalino de 2024, com vendas estáveis entre os dias 23 e 25 de dezembro, diferentemente dos picos tradicionais dos anos anteriores. Isso sugere mudanças no comportamento do consumidor ou possíveis ajustes necessários no modelo preditivo, reforçando a importância de estudos complementares para confirmar essa projeção.

Outro ponto relevante é a previsão de demanda semanal estável para dezembro de 2024, o que pode gerar benefícios logísticos significativos e otimizar o planejamento operacional da empresa. Porém, destaca-se a necessidade de revisar estratégias comerciais para assegurar a constância dessa demanda projetada.

A análise da evolução anual acumulada revelou um expressivo aumento nas vendas previstas para 2024, destacando o sucesso das estratégias comerciais adotadas. Portanto, recomenda-se que a empresa direcione esforços para otimizar continuamente a gestão logística e comercial, aproveitando plenamente essa tendência de crescimento.

Finalmente, a análise crítica da sazonalidade sugere uma possível

suavização excessiva das previsões diárias para dezembro de 2024, especialmente em relação ao pico natalino. É recomendada uma revisão do modelo para assegurar que a sazonalidade histórica seja corretamente refletida nas futuras previsões.

Em suma, este estudo fornece bases sólidas para decisões estratégicas concretas, posicionando a Segrob Notlad para aprimorar seu desempenho operacional, ajustar estratégias comerciais eficazmente e manter seu protagonismo em inovação orientada por dados no segmento brasileiro de fast fashion.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ALTMAN, E. L. Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. *Journal of Finance*, v. 23, n. 4, p. 589-609, 1968.

ALVES, C. da C.; HOEPERS, E.; CORAZZA, E. J.; SANTOS, G. J. dos; CRISTOFOLINI, R.; CRUZ, A. C. da. **Aplicação de métodos estatísticos com suavização exponencial dupla e tripla para previsão de demanda na gestão de estoques.** 2019. *Revista Produção Online*, Florianópolis. Disponível em: https://www.producaoonline.org.br/rpo/article/view/3539/1832

BALLOU, R. H. **Gerenciamento da Cadeia de Suprimentos**. 5. ed. Porto Alegre: Bookman, 2006.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística básica**. 9. ed. São Paulo: Saraiva Educação, 2017.

CAMPOS, E. M.; SILVA, J. G. **Mineração de Dados: conceitos, técnicas, algoritmos, orientações e aplicações.** Rio de Janeiro: Elsevier, 2019.

CHAI, T.; DRAXLER, R. R. **Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature**. *Geoscientific Model Development*, v. 7, n. 3, p. 1247–1250, 2014. DOI: 10.5194/gmd-7-1247-2014. Disponível em: https://doi.org/10.5194/gmd-7-1247-2014. Acesso em: 8 jul. 2025.

CHEIN, Flávia. **Introdução aos modelos de regressão linear.** Brasília - DF, 2019. Disponível em: https://repositorio.enap.gov.br/bitstream/1/4788/1/Livro_Regress%c3%a3o%20Linear.pdf

CHERKASSKY, V.; MULIER, F. ***Learning from Data***. John Wiley & Sons, Inc., 1998.

FÁVERO, L. P.; BELFIORE, P. **Manual de Análise de Dados: estatística e modelagem multivariada com Excel, SPSS e Stata**. 1. ed. Rio de Janeiro: Elsevier, 2017.

FERRAZ, Thales. **Análise de Dados: Regressão Linear Múltipla com Excel.** LinkedIn, 2024. Disponível em: https://www.linkedin.com/pulse/an%C3%A1lise-de-dados-regress%C3%A3o-linear-m%C3%BAltipla-excel-thales-ferraz-w6a2f/. Acesso em: 29 maio 2025.

FERRERO, Carlos Andres. **Algoritmo KNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia**. 2009. Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos - SP. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-19052009-135128/publico/CAFerrero_dissertacao.pdf

FILHO, Oscar Gabriel. **Inteligência Artificial e Aprendizagem de Máquina: aspectos teóricos e aplicações.** São Paulo, 2023. Disponível em: https://storage.blucher.com.br/book/pdf_preview/PDF_ia.pdf?utm_source=chatgpt.com

FRIEYADIE; SETIYORINI, Tyas. ***COMPARISON OF THE APPLICATION OF LINEAR REGRESSION WITH SLIDING WINDOW VALIDATION AND K-FOLD CROSS-VALIDATION FOR FORECASTING COVID-19 RECOVERED CASES.*** Universitas Nusa Mandiri, 2024. Disponível em: https://ejournal.kresnamediapublisher.com/index.php/jri/article/view/288/283

GARCIA, Simone Carboni. **O uso de árvores de decisão na descoberta de conhecimento na área da saúde.** 2003. Mestrado em Computação – Programa de Pós-Graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre - RS. Disponível em: https://lume.ufrgs.br/bitstream/handle/10183/4703/000503532.pdf?sequence=1&isAllowed=y. Acesso em: 22 jun. 2025.

GIRO, Riccardo Angelo; BERNASCONI, Giancarlo; GIUNTA, Giuseppe; CESARI, Simone. **Tagging and tracking oil-gas mixtures in multiphase pipelines**. *Journal of Petroleum Science and Engineering*, v. 218, 2022. Disponível em: https://www.elsevier.com/locate/petrol. Acesso em: 22 jun. 2025.

GOUVEIA, J. F. et al. **Modelos Volumétricos Mistos em Clones de Eucalyptus no Polo Gesseiro do Araripe, Pernambuco**. Floresta, Curitiba, v. 45, n. 3, p. 587-598, jul./set. 2015. DOI: 10.5380/rf.v45i3.36844. Disponível em: https://www.cabidigitallibrary.org/doi/pdf/10.5555/20153393545. Acesso em: 16 maio 2025.

GUEDES, T. A. et al. **Econometria aplicada: métodos e aplicações utilizando R**, EViews e Stata. São Paulo: Atlas, 2018.

KANITZ, S. C. **Como prever falências**. São Paulo: McGraw-Hill, 1978.

LIMA, Rafaela Somavila. **CRIAÇÃO DE PROJETO DE CIÊNCIA DE DADOS UTILIZANDO A METODOLOGIA CRISP-DM EM CONFORMIDADE COM A LGPD.** 10 ago. 2021. Disponível em: https://repositorio.utfpr.edu.br/jspui/bitstream/1/28037/1/CT_CCEDA_2019_02_11.pdf

MARIO FILHO. **MAPE (Erro Absoluto Percentual Médio) em Machine Learning**. 2022. Disponível em: https://mariofilho.com/mape-erro-absoluto-percentual-medio-em-machine-learning/. Acesso em: 16 maio 2025.

MATIAS, A. B. **Contribuição às técnicas de análise financeira: um modelo de concessão de crédito**. Tese (Doutorado), Universidade de São Paulo, 1978.

MATOS, Manuel António. **Manual Operacional para a Regressão Linear.** 1995. Disponível em: https://paginas.fe.up.pt/~mam/regressao.pdf

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. ***Introduction to linear regression analysis*.** 5. ed. Hoboken: Wiley, 2012. Disponível em: https://books.google.com.br/books?id=lSyiRZh09oEC&printsec=frontcover&dq=Introduction+to+Linear+Regression+Analysis&hl=pt-BR&sa=X&ei=XJbnUo61DPHmsASo-YC4Bw&ved=0CC8Q6AEwAA#v=onepage&q=Introduction%20to%20Linear%20Regression%20Analysis&f=false

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de Séries Temporais.** 3. ed. São Paulo: Blucher, 2018.

OSHIRO, Thais Mayumi. **Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica**. 2013. Dissertação (Mestrado em Bioinformática) – Universidade de São Paulo, Ribeirão Preto, 2013. Disponível em: https://www.teses.usp.br/teses/disponiveis/95/95131/tde-15102013-183234/publico/monografia. Acesso em: 08 jul. 2025.

POILL, C. **Árvore de Decisão: Entendendo o conceito de Entropia e Ganho de Informação**, 2022. Disponível em: https://caiopovill.medium.com/árvore-de-decisão-entendendo-o-conceito-de-entropia-e-ganho-de-informação-9ce41a3169ae. Acesso em: 22 junho 2025.

RAMOS, Luis Jorge C.; SILVA, João C. Sedraz; RODRIGUEZ, Rodrigo L.; DE OLIVEIRA, Pamella Letícia Silva. **CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais.** 2020. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/12865/12719

PORTAL DATA SCIENCE. **O algoritmo k-Nearest Neighbors (kNN) em machine learning.** Disponível em: https://portaldatascience.com/o-algoritmo-k-nearest-neighbors-knn-em-machine-learning/. Acesso em: 22 jun. 2025.

RODRIGUES, Sandra Cristina Antunes. **Modelo de Regressão Linear e suas Aplicações.** Covilhã - Portugal, out. 2012. Disponível em: https://www.proquest.com/openview/ddde3bb8c207d5a138545c2fcd34a6ce/1?cbl=2026366&diss=y&pq-origsite=gscholar

SCARPEL, R. A. **Utilização de Support Vector Machine em previsão de insolvência de empresas**. *Pesquisa Operacional e Desenvolvimento Sustentável*, Gramado, RS, p. 672-677, 2005.

SILVA, Fernando da. MAE, RMSE, ACC, F1, ROC, R2? **Avaliação de desempenho de modelos preditivos.** *Análise Macro*, 27 out. 2023. Disponível em: https://analisemacro.com.br/econometria-e-machine-learning/mae-rmse-acc-f1-roc-r2-avaliacao-de-desempenho-de-modelos-preditivos/. Acesso em: 16 maio 2025.

SILVA, Fernando da. **Devemos usar a métrica MAPE em previsão de demanda?** *Análise Macro*, 3 abr. 2023. Disponível em: https://analisemacro.com.br/data-science/devemos-usar-a-metrica-mape-em-previsao-de-demanda/. Acesso em: 16 maio 2025.

SILVA, Fernando da. **Validação cruzada na prática: qual método utilizar?**. *Análise Macro*, 14 set. 2023. Disponível em: https://analisemacro.com.br/data-science/validacao-cruzada-na-pratica-qual-metodo-utilizar/. Acesso em: 16 maio 2025.

VAPNIK, V. N. **An overview of statistical learning theory**. *IEEE Transactions on Neural Networks*, v. 10, n. 5, p. 988-999, 1999.