UNIVERSIDADE FEDERAL FLUMINENSE – UFF
CAMPUS RIO DAS OSTRAS
GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

ANDRÉ AUGUSTO DA SILVA FERNANDES
LENILDO MACHADO RIBEIRO LIMA JÚNIOR

RELATÓRIO FINAL DE ANÁLISE PREDITIVA

Rio das Ostras
2025

ANDRÉ AUGUSTO DA SILVA FERNANDES

LENILDO MACHADO RIBEIRO LIMA JÚNIOR

RELATÓRIO FINAL DE ANÁLISE PREDITIVA

Relatório de Análise Preditiva, apresentado à Universidade Federal Fluminense – Campus Rio das Ostras, como requisito obrigatório para a obtenção de conceito na disciplina de Análise Preditiva Obrigatória do Curso de Graduação em Engenharia de Produção, sob a orientação do professor Dalton Borges.

Rio das Ostras

2025

Sumário

1. CONTEXTUALIZAÇÃO

A previsão de demanda é uma metodologia utilizada para estimar valores futuros de uma variável de interesse, como as vendas de um produto. Essa ferramenta é essencial para o gerenciamento eficiente das operações e para apoiar a tomada de decisões estratégicas em diversos setores, inclusive no varejo de moda (Luo; Chang; Xu, 2022).

Modelos de previsão inadequados podem levar a uma gestão ineficaz dos estoques, resultando em perdas financeiras e impactos negativos em toda a cadeia de suprimentos. Por isso, grandes empresas investem constantemente em técnicas de previsão cada vez mais precisas, buscando identificar padrões de comportamento nas séries históricas. Um modelo eficiente garante não apenas uma melhor organização dos estoques, mas também oferece vantagem competitiva (Giri; Chen, 2022).

No cenário atual, marcado pela globalização e pela aceleração do consumo digital, o comportamento do mercado se torna ainda mais volátil. Produtos podem ter variações bruscas de demanda ao longo dos anos, reforçando a necessidade do uso de dados históricos aliados a métodos de previsão para orientar as estratégias comerciais (Sousa; Loureiro; Miguéis, 2025).

Particularmente no setor de moda, a sazonalidade, a volatilidade e a sensibilidade a tendências são ainda mais intensas (Giri; Chen, 2022). Dentro desse contexto, o presente estudo foca na rede *Segrob Notlad*, uma empresa de fast fashion reconhecida pelo design acessível de seus produtos. Com mais de 80 lojas no Brasil e três unidades na Europa, a empresa busca aprimorar seu processo de reposição de camisetas básicas, um item-chave em seu portfólio.

O objetivo deste estudo é estimar o volume diário de vendas de camisetas básicas ao longo do mês de dezembro de 2024, a fim de apoiar o planejamento estratégico da empresa. A previsão de demanda permitirá um abastecimento mais eficiente das lojas, reduzindo excessos e rupturas de estoque, além de otimizar os custos logísticos e operacionais. Por fim, ao alcançar esses resultados, a empresa poderá obter uma vantagem competitiva em relação aos seus concorrentes, fortalecendo sua posição no mercado.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 PREVISÃO DE DEMANDA

A previsão de demanda é o processo de estimar valores futuros de uma variável de interesse, desempenhando papel fundamental na tomada de decisões em diversos contextos produtivos. No setor varejista, por exemplo, a previsão de demanda é essencial para o planejamento eficiente dos recursos e a definição de estratégias comerciais (Chen et al., 2020).

É importante destacar que a demanda é influenciada por uma variedade de fatores, como feriados, preferências culturais e regionais, eventos sazonais, controvérsias envolvendo a marca e ações de marketing. Todos esses elementos devem ser considerados para que a previsão seja mais precisa e alinhada à realidade do mercado (Seyedan; Mafakheri, 2020).

A previsão de demanda não deve ser vista apenas como uma extrapolação de dados históricos, mas também como uma ferramenta inteligente, capaz de se adaptar a mudanças no ambiente de negócios. Modelos preditivos modernos incorporam mecanismos de aprendizado e ajuste contínuo, permitindo que a previsão evolua em resposta a alterações na cadeia de suprimentos (Seyedan; Mafakheri, 2020).

Por fim, a previsão de demanda pode ser entendida como o processo de determinar valores futuros com base em dados históricos e estatísticos, utilizando uma metodologia clara, estruturada e replicável (Fonseca; Pedrosa; Cardoso, 2024). Ao permitir uma melhor alocação de recursos, redução de custos operacionais e aumento da capacidade de resposta às flutuações do mercado, a previsão de demanda torna-se importante para o crescimento das empresas.

2.2 ANÁLISE DE ERROS

A análise de erros representa uma etapa na avaliação de modelos preditivos, pois permite quantificar a discrepância entre os valores estimados e os valores observados (Chen et al., 2020). De forma geral, os erros de previsão correspondem às diferenças entre os dados reais e aqueles gerados pelo modelo, sendo essenciais para a compreensão do desempenho da modelagem.

Há ampla gama de métricas estatísticas para a mensuração desses erros, cada uma com características, vantagens e limitações específicas (Adeleke et al., 2021). Neste trabalho, serão

empregadas três principais métricas: o MAPE (Erro Percentual Médio Absoluto), que indica a precisão em termos percentuais; o RMSE (Raiz do Erro Médio Quadrático), que enfatiza grandes desvios ao penalizá-los com maior severidade; e o MAD (Erro Médio Absoluto), que oferece uma medida clara do erro médio, independente da direção. A adoção de múltiplos indicadores contribui para uma avaliação confiável do desempenho do modelo.

$$MAD = \frac{\sum_{t=1}^{n} |A_t - F_t|}{n} \qquad RMSE = \sqrt{\frac{\sum_{t=1}^{n} (A_t - F_t)^2}{n}} \qquad MAPE = \frac{\sum_{t=1}^{n} \left|\frac{A_t - F_t}{A_t}\right|}{n} * 100$$

Nesse contexto, At representa o valor real, Ft o valor previsto, e n o número total de observações (Geetha et al., 2022). As métricas utilizadas fornecem diferentes perspectivas sobre os erros de previsão, ou seja, enquanto algumas quantificam o desvio médio absoluto, outras destacam a variação percentual.

2.3 AVALIAÇÃO DE MODELO

A avaliação do modelo é uma etapa importante para a avaliação adequada do desempenho de um modelo preditivo, permitindo verificar sua precisão por meio de testes. De modo geral, utilizam-se métricas que quantificam a diferença entre os valores esperados e os valores estimados pelo modelo (Chen et al., 2020).

É importante destacar que esse processo é realizado utilizando-se um conjunto de dados distinto daquele empregado no treinamento do modelo, justamente para evitar respostas enviesadas decorrentes da familiaridade com os chamados dados de treinamento. Em síntese, a principal finalidade é avaliar a capacidade do modelo de generalizar seu desempenho para dados inéditos, denominados dados de validação (Maleki et al., 2020).

2.3.1 Validação Cruzada

A validação cruzada é uma técnica de reamostragem empregada na avaliação do desempenho de modelos preditivos, com o objetivo de fornecer estimativas mais imparciais do erro de generalização (Chen et al., 2020). Diferentemente da validação por retenção (o conjunto de dados é dividido, uma única vez, em subconjuntos exclusivos para treinamento e teste) a validação cruzada reduz a dependência da partição específica escolhida, mitigando o

viés e a variância da estimativa de erro. Essa característica torna a técnica particularmente vantajosa em contextos com conjuntos de dados de tamanho limitado, nos quais cada observação possui maior relevância na modelagem (Maleki et al., 2020).

Diversas variações de validação cruzada podem ser adotadas, como a validação cruzada k-fold, leave-one-out, entre outras. Neste trabalho, serão destacadas as principais abordagens utilizadas nas etapas de avaliação e seleção de modelos.

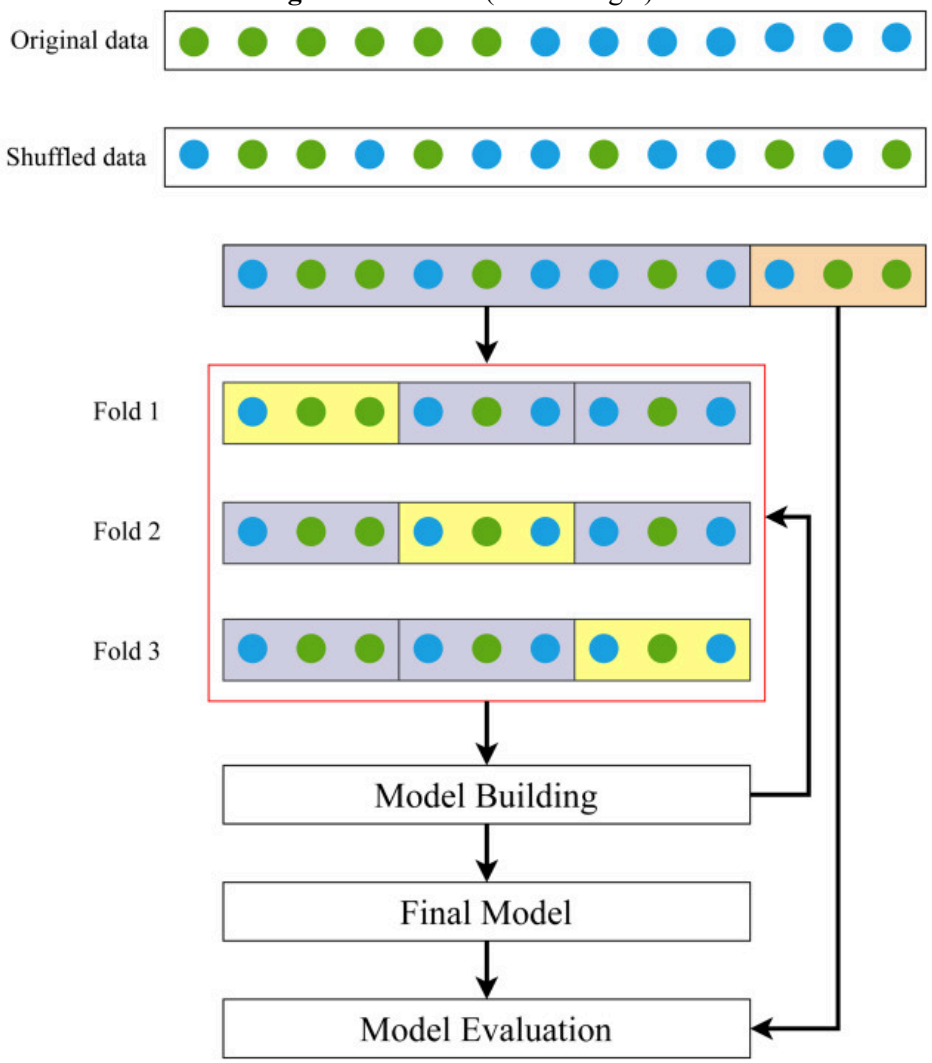
2.3.1.1 Validação Cruzada K-Fold (KFCV)

A validação cruzada k-fold consiste em dividir aleatoriamente os dados disponíveis em k subconjuntos mutuamente exclusivos (ou "folds") de tamanho semelhante. Em seguida, o modelo é treinado k vezes, cada vez utilizando $k - 1$ subconjuntos como conjunto de treinamento e o subconjunto restante como conjunto de validação. A média das métricas de desempenho obtidas ao longo das k iterações é então usada como uma estimativa do erro de validação (Allgaier; Pryss, 2024).

A escolha do valor de k influencia diretamente a performance e a confiabilidade do processo. Valores comuns são $k = 5$ e $k = 10$, pois oferecem um bom equilíbrio entre viés e variância na estimativa do erro, além de manter a viabilidade computacional (Cerqueira; Torgo; Mozetič, 2020).

É importante destacar que, em muitas aplicações, uma parte dos dados é separada previamente como conjunto de teste, sendo utilizada somente após a etapa de validação cruzada para fornecer uma estimativa não enviesada do erro de generalização (Allgaier; Pryss, 2024).

Figura 1 - K-Fold (Metodologia)



Fonte: Maleki et al. (2020)

Outro aspecto relevante na aplicação da validação cruzada é o desequilíbrio de classes, que ocorre quando há uma disparidade significativa entre a quantidade de amostras das classes, sendo a classe majoritária aquela com maior número de instâncias e a classe minoritária aquela com menor. Nesses casos, a validação cruzada pode resultar em medidas de desempenho instáveis. A validação cruzada k-fold estratificada realiza a divisão dos dados em cada uma das *k* dobras de forma a preservar a proporção original das classes no conjunto de dados (Maleki et al., 2020).

2.3.1.2 Validação Cruzada Leave-One-Out (LOOCV)

É uma forma de validação *k*-fold, na qual $k = n$, sendo n o número total de amostras no conjunto de dados. Em cada iteração, uma única amostra é separada para teste, enquanto o restante é utilizado para o treinamento do modelo. Esse procedimento é repetido n vezes, o que garante que todas as amostras sejam utilizadas uma vez como conjunto de teste (Allgaier; Pryss, 2024). Apesar de sua precisão, a LOOCV é computacionalmente intensiva, pois requer o treinamento de n modelos distintos. Essa limitação torna seu uso inviável em contextos com grandes volumes de dados (Maleki et al., 2020).

Além da LOOCV, outras variações da validação cruzada podem ser utilizadas em contextos específicos. A validação cruzada *leave-p-out* (LPOCV) é uma generalização da LOOCV, em que p amostras são deixadas de fora em cada iteração, ao invés de apenas uma. Embora forneça estimativas boas de desempenho, seu custo computacional cresce exponencialmente com o aumento de p (Maleki et al., 2020).

Em situações onde há dependência entre amostras, a validação cruzada *leave-one-group-out* (LOGOCV) é mais adequada. Nessa abordagem, um grupo inteiro de amostras correlacionadas é excluído a cada iteração, evitando viés na estimativa de desempenho (Maleki et al., 2020).

2.4 MODELAGEM (MODELOS PARAMÉTRICOS)

Modelos paramétricos são amplamente utilizados na previsão de vendas devido à sua simplicidade e facilidade de interpretação. Esses modelos assumem, desde o início, uma forma funcional específica para a relação entre as variáveis e trabalham com um número fixo de parâmetros (Lin et al., 2022).

Neste estudo, serão utilizados diversos modelos paramétricos clássicos, com destaque para abordagens baseadas em séries temporais, como os modelos Naive, Cumulativo, Média Móvel, Suavização Exponencial Simples, Suavização Exponencial Dupla e Suavização Exponencial Tripla, além do modelo de Regressão Dinâmica, que permite incorporar variáveis explicativas ao processo de previsão. Os benefícios dos modelos paramétricos incluem baixa necessidade de dados de treinamento, maior facilidade na interpretação dos resultados e simplicidade na análise de sensibilidade dos parâmetros (Lin et al., 2022).

Apesar das vantagens em termos de eficiência e interpretabilidade, a rigidez estrutural

desses modelos pode limitar sua capacidade de adaptação a padrões de consumo altamente voláteis, o que justifica, em alguns casos, o uso complementar de modelos não paramétricos (Yu et al., 2021).

2.4.1 Naive

O modelo *Naive* é uma abordagem simples, porém bastante útil em contextos nos quais a complexidade dos dados não justifica o uso de modelos mais sofisticados. A previsão para o próximo período é assumida como sendo igual ao valor observado no período anterior (Hyndman; Athanasopoulos, 2021).

Apesar de sua simplicidade, o modelo é frequentemente utilizado como *benchmark* para avaliar o desempenho de métodos mais complexos, justamente por sua facilidade de implementação. Além disso, em determinadas aplicações nas quais os dados são altamente voláteis ou há pouco histórico disponível, o modelo Naive pode apresentar desempenho comparável ao de métodos mais avançados (Gunter, 2021).

$$\hat{y}_{t+1} = y_t$$

Nesse contexto, $\hat{y}_{t+1}$ representa a previsão para o período t+1, enquanto y_t corresponde ao valor real observado no período anterior, t. O modelo pressupõe que o melhor palpite para o futuro é simplesmente repetir o último valor conhecido (Gunter, 2021).

2.4.2 Cumulativo

Os modelos de previsão cumulativos são amplamente utilizados em contextos nos quais a variável de interesse se acumula ao longo do tempo. A principal característica desses modelos é considerar não apenas os valores pontuais de determinado período, mas também o somatório dos eventos registrados até então, o que possibilita capturar a tendência geral do fenômeno observado (Hyndman; Athanasopoulos, 2021).

Apesar das vantagens em termos de suavização dos dados e identificação de tendências, os modelos cumulativos apresentam também algumas limitações. Uma das principais é a perda de sensibilidade a mudanças abruptas mais recentes, o que pode comprometer a capacidade de adaptação do modelo em ambientes dinâmicos (Kriston, 2020).

$$\hat{y}_{t+1} = \sum_{i=1}^{t} y_i$$

Nessa formulação, $\hat{y}_{t+1}$ representa o valor previsto acumulado até o instante t+1, e y_i são os valores observados em cada período i, com i=1,2,...,t. A utilização desse modelo é apropriada em contextos onde o comportamento acumulativo da série é relevante para a tomada de decisão (Kriston, 2020).

2.4.3 Média Móvel

A média móvel é um dos métodos mais simples e amplamente utilizados para previsão de séries temporais, especialmente quando se busca suavizar flutuações aleatórias e identificar tendências subjacentes nos dados (Hyndman; Athanasopoulos, 2021). O modelo consiste em calcular a média aritmética de um número fixo de observações anteriores, sendo atualizado conforme novos dados são inseridos, o que o torna uma técnica adaptável a diferentes contextos (Bhardwaj et al., 2021).

Embora seja um método de fácil implementação e interpretação, a média móvel apresenta limitações, como o atraso na resposta a mudanças abruptas e a incapacidade de capturar sazonalidades complexas sem ajustes adicionais (Hyndman; Athanasopoulos, 2021).

$$\hat{y}_{t+1} = \frac{1}{k} \sum_{i=t-k+1}^{t} y_i$$

Nesse caso, $\hat{y}_{t+1}$ é a previsão para o próximo período, baseada na média dos k últimos valores observados, representados por y_i. O parâmetro k determina a janela de tempo considerada e deve ser escolhido com base na característica da série, de modo a equilibrar suavização e sensibilidade às variações.

2.4.4 Suavização Exponencial Simples

A suavização exponencial simples é um dos métodos mais básicos e amplamente utilizados para a previsão de séries temporais univariadas, especialmente quando os dados não apresentam tendência ou sazonalidade (Kim; Kim, 2021). O modelo pressupõe que o valor

futuro da série é uma média ponderada entre o valor observado mais recente e a previsão anterior, sendo que os pesos decrescem exponencialmente à medida que os dados se afastam no tempo (Ensafi et al., 2022).

Tem um parâmetro α de suavização que varia entre 0 e 1. Valores de α mais próximos de 1 atribuem maior peso às observações mais recentes, enquanto valores menores conferem maior inércia ao modelo. Esse modelo se destaca por sua capacidade de responder rapidamente a mudanças nos dados, embora sua principal limitação seja a ineficácia diante de séries com comportamento de tendência (Kim; Kim, 2021).

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha)\hat{y}_t$$

Nessa equação, $\hat{y}_{t+1}$ é a previsão para o próximo período, enquanto y_t é o valor observado no tempo t, e $\hat{y}_t$ representa a previsão anterior. O parâmetro α controla o grau de suavização, com valores entre 0 e 1 (Kim; Kim, 2021).

2.4.5 Suavização Exponencial Dupla

Para lidar com séries temporais que apresentam tendência, foi desenvolvida a suavização exponencial dupla, também conhecida como modelo de Holt. Esse modelo estende a suavização simples ao introduzir um componente adicional que captura a tendência da série ao longo do tempo. A abordagem consiste em duas equações: uma para o nível da série e outra para a tendência, ambas ajustadas iterativamente a partir dos dados históricos. (Ensafi et al., 2022).

Os parâmetros de suavização α e β controlam a sensibilidade da estimativa ao comportamento recente da série e sua tendência, respectivamente. Ainda que apresente um desempenho superior ao da suavização simples em séries com tendência, o modelo de Holt não é adequado para dados sazonais, o que levou à formulação de modelos ainda mais completos, como o de Holt-Winters (Munim et al., 2023).

$$l_t = \alpha y_t + (1 - \alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1}$$

$$\hat{y}_{t+h} = l_t + hb_t$$

Nessa abordagem, l_t representa o nível estimado da série no tempo t, enquanto b_t é a tendência. O valor $\hat{y}_{t+h}$ indica a previsão para h períodos à frente. Os parâmetros α e β controlam a suavização do nível e da tendência, respectivamente (Munim et al., 2023).

2.4.6 Suavização Exponencial Tripla

A suavização exponencial tripla, ou modelo de Holt-Winters, é uma generalização da abordagem de Holt, incluindo um componente sazonal para lidar com séries que exibem variações sistemáticas em intervalos regulares. Essa técnica é particularmente apropriada para aplicações em que a demanda ou o comportamento da série varia conforme o período de tempo (Ensafi et al., 2022).

O modelo de Holt-Winters é composto por três equações de suavização: uma para o nível, uma para a tendência e outra para a sazonalidade, cada uma com seu respectivo parâmetro de suavização (α, β e γ). Embora seja mais complexo, o modelo de Holt-Winters oferece previsões bastante precisas em contextos onde os padrões sazonais são predominantes, como em vendas no varejo que é o caso deste trabalho (Trull; García-Díaz; Troncoso, 2020).

$$l_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1}$$

$$s_t = \gamma(y_t - l_t) + (1 - \gamma)s_{t-m}$$

$$\hat{y}_{t+h} = l_t + hb_t + s_{t+h-m(k+1)}$$

Nessa formulação, l_t representa o nível da série, b_t a tendência e s_t o componente sazonal. O parâmetro m corresponde ao número de períodos no ciclo sazonal, enquanto h é o horizonte de previsão. Os parâmetros α, β e γ controlam a suavização dos respectivos componentes. A previsão $\hat{y}_{t+h}$ é composta pela soma desses três elementos, permitindo capturar padrões complexos na série (Trull; García-Díaz; Troncoso, 2020).

2.4.7 Regressão Linear

A regressão linear é uma técnica estatística fundamental que visa descrever a relação entre duas variáveis quantitativas, sendo uma considerada dependente e a outra independente (Mohd Jais et al., 2024). A partir da suposição de que essa relação pode ser representada por uma equação linear, o modelo busca ajustar uma reta que melhor explica como as variações na variável independente influenciam os valores da variável dependente (Liaw et al., 2020).

Apesar de sua simplicidade, a regressão linear é amplamente utilizada em diversas áreas por fornecer estimativas e exigir relativamente poucos dados para aplicação (Toft et al., 2023). Contudo, a validade do modelo depende do atendimento a pressupostos como linearidade, normalidade dos erros, ausência de autocorrelação e independência das observações (Mohd Jais et al., 2024).

$$\hat{y} = \beta_0 + \beta_1 x$$

A variável x representa a variável independente, também chamada de preditora ou explicativa. O termo β_0 é o intercepto da reta, indicando o valor previsto de $\hat{y}$ quando x é igual a zero. Já o coeficiente β_1 representa a inclinação da reta de regressão, ou seja, quantifica a variação esperada em $\hat{y}$ para cada unidade de aumento em x (Mohd Jais et al., 2024).

Entre esses modelos, destacam-se aqueles voltados à análise de dados com estrutura temporal, como a regressão dinâmica. Esse tipo de abordagem amplia a regressão tradicional ao incorporar dependências entre observações ao longo do tempo, sendo particularmente útil em estudos com séries temporais (Chum et al., 2021).

2.4.7.1 Regressão Dinâmica

A regressão dinâmica é uma abordagem estatística que amplia os conceitos tradicionais da regressão linear ao incorporar a dimensão temporal dos dados. Essa técnica é especialmente relevante no contexto de séries temporais, pois permite modelar não apenas relações estáticas entre variáveis, mas também suas interdependências ao longo do tempo (Chum et al., 2021).

Em essência, trata-se de uma evolução dos modelos clássicos de regressão, projetada

para capturar a autocorrelação entre observações ao longo do tempo. Isso significa que o valor atual de uma variável pode depender de seus próprios valores passados (Alduailij et al., 2021). Além disso, o modelo pode incluir variáveis explicativas externas (variáveis exógenas), cujos efeitos muitas vezes se manifestam com defasagem. A inclusão dessas variáveis torna o modelo mais robusto na representação de fenômenos em que as influências não são imediatas (Muzalyova et al., 2021).

Em resumo, a regressão dinâmica é especialmente útil em aplicações voltadas à previsão de séries temporais que sofrem influência de múltiplos fatores, como indicadores econômicos, condições climáticas, políticas públicas, entre outros (Pérez-López et al., 2020). Assim, configura-se como uma ferramenta poderosa para modelar e prever fenômenos temporais em cenários complexos.

$$\hat{y}_t = \beta_0 + \sum_{i=1}^{p} \beta_i x_{i,t}$$

Nessa estrutura, $\hat{y}_t$ é o valor previsto no tempo t, β_0 é o intercepto, e β_i são os coeficientes associados a cada variável explicativa $x_{i,t}$. O número total de variáveis exógenas é representado por p (Muzalyova et al., 2021).

2.5 MODELAGEM (MODELOS NÃO PARAMÉTRICOS)

Os modelos não paramétricos não assumem uma forma funcional fixa para a relação entre as variáveis. Em vez disso, esses modelos buscam extrair padrões diretamente dos dados, oferecendo maior flexibilidade para capturar comportamentos complexos, sendo especialmente úteis em contextos com alta variabilidade ou mudanças abruptas no comportamento de compra (Lin et al., 2022).

Neste trabalho, serão utilizados três modelos não paramétricos: o algoritmo *k-Nearest Neighbors (k-NN)*, as Árvores de Decisão com *Random Forest* e o *Support Vector Machine* (SVM). Esses métodos, ao contrário dos modelos paramétricos, não impõem suposições rígidas sobre a estrutura dos dados e conseguem lidar melhor com relações não lineares. No entanto, geralmente exigem maior volume de dados e maior poder computacional (Yu et al., 2021).

A principal vantagem dos modelos não paramétricos é sua flexibilidade, o que os torna

especialmente úteis quando há incerteza sobre a forma da função que relaciona as variáveis de entrada e saída. No entanto, essa flexibilidade pode vir acompanhada de maior complexidade interpretativa e maior risco de sobreajuste, especialmente quando o conjunto de dados apresenta muito ruído (Yu et al., 2021).

2.5.1 Hiperparâmetros

Os hiperparâmetros desempenham um papel fundamental no desempenho de modelos de aprendizado de máquina, sendo definidos como parâmetros cujo valor é estabelecido antes do processo de treinamento, ao contrário dos parâmetros internos do modelo que são ajustados durante o treinamento. A escolha adequada desses hiperparâmetros pode impactar significativamente a capacidade do modelo de generalizar para dados não vistos (Yang; Shami, 2020).

Algoritmos como o *k-Nearest Neighbors* (k-NN), o *Support Vector Machines* (SVM) e o *Random Forest* dependem diretamente de hiperparâmetros que influenciam significativamente seu desempenho. No k-NN, o número de vizinhos é o principal fator a ser ajustado, afetando a sensibilidade do modelo a ruídos. No SVM, os hiperparâmetros mais relevantes são o parâmetro de custo C e o tipo de kernel, que definem a flexibilidade e a capacidade de separação. No *Random Forest*, destaca-se a quantidade de árvores, que impacta a estabilidade e a capacidade de generalização (Markovics; Mayer, 2022).

A determinação dos hiperparâmetros mais adequados é comumente realizada por meio de técnicas de validação cruzada combinadas a métodos de busca como *grid search*, *random search* ou algoritmos de otimização bayesiana. A validação cruzada, em particular, permite estimar a capacidade de generalização do modelo ao dividir os dados em múltiplos subconjuntos e avaliar o desempenho médio do modelo em diferentes partições. Essa abordagem reduz o risco de *overfitting* sobre o conjunto de validação (Elgeldawi et al., 2021).

Por fim, é importante destacar que a negligência na escolha dos hiperparâmetros pode comprometer significativamente a performance de modelos promissores. Um modelo sofisticado, mas mal configurado, pode ter desempenho inferior a modelos mais simples, porém bem ajustados. Assim, a etapa de definição e ajuste de hiperparâmetros não deve ser tratada como um procedimento secundário, mas como parte integrante no processo de modelagem preditiva (Nguyen et al., 2021).

2.5.2 k-Nearest Neighbors (k-NN)

O algoritmo *k-Nearest Neighbors (KNN)* é uma técnica de aprendizado supervisionado amplamente utilizada para tarefas de classificação e regressão, sendo reconhecido por sua simplicidade e efetividade (Uddin et al., 2022). A previsão de uma nova amostra é realizada com base na classe ou valor das k instâncias mais próximas no conjunto de treinamento, conforme uma medida de distância predefinida (Khaledian; Miller, 2020).

Entre os principais hiperparâmetros do KNN, destaca-se o valor de k, que representa o número de vizinhos a serem considerados. Quando k é muito pequeno, o modelo pode se tornar sensível ao ruído nos dados, resultando em sobreajuste. Por outro lado, valores muito altos tendem a suavizar excessivamente as fronteiras de decisão (Uddin et al., 2022).

Outro aspecto importante é a métrica de distância utilizada para identificar os vizinhos mais próximos. A distância Euclidiana é a mais comum no KNN, sendo apropriada para dados numéricos e com escala padronizada, mas outras métricas também podem ser utilizadas, como a distância de Manhattan e a de Mahalanobis, dependendo da natureza dos dados (Nabipour et al., 2020).

Por ser um algoritmo baseado em instâncias, o KNN não realiza uma etapa de treinamento propriamente dita, armazenando todos os dados e realizando o cálculo das distâncias apenas no momento da predição. Isso torna o modelo computacionalmente custoso em grandes conjuntos de dados. Ainda assim, sua facilidade de implementação e desempenho competitivo em diversos contextos justificam sua aplicação (Yağcı, 2022).

2.5.3 Árvores de Decisão

A árvore de decisão é um dos métodos mais amplamente utilizados na área de aprendizado de máquina supervisionado, sendo aplicada tanto em problemas de classificação quanto de regressão (Nasser; Rashad; Hussein, 2020). Estruturalmente, uma árvore de decisão é composta por nós internos que representam testes lógicos sobre atributos, ramos que correspondem aos resultados desses testes, e nós folhas que representam a predição final. Ao percorrer a árvore desde a raiz até uma folha, um conjunto de regras é aplicado sucessivamente, conduzindo à decisão final (Walker et al., 2020).

O critério mais comum na classificação para definir a qualidade de uma divisão é a medida de impureza, como a entropia ou o índice de Gini, que quantificam o grau de

homogeneidade dos subconjuntos gerados em cada nó. Já em regressão, são utilizadas métricas como o erro quadrático médio para avaliar a qualidade das divisões (Van Steenbergen; Mes, 2020). Um dos grandes desafios na construção de árvores de decisão consiste em evitar o *overfitting*, que ocorre quando a árvore se torna demasiadamente complexa e modela o ruído dos dados de treinamento (Walker et al., 2020).

Além de sua aplicabilidade direta, as árvores de decisão servem como base para métodos mais sofisticados, como o *Random Forest* e o *Decision Tree*, que combinam múltiplas árvores com o intuito de melhorar a acurácia preditiva e reduzir a variância do modelo (Brillinger et al., 2021). Enquanto as árvores individuais podem ser sensíveis a pequenas perturbações nos dados, esses métodos aumentam a robustez e mitigam problemas de instabilidade (Waqas Khan et al., 2020).

2.5.3.1 Random Forest

O *Random Forest* é um algoritmo de aprendizado supervisionado amplamente empregado em tarefas de classificação e regressão (Maimaitijiang et al., 2020). Sua estrutura baseia-se na construção de múltiplas árvores de decisão, combinadas por meio de uma abordagem denominada *bagging*, em que cada árvore é treinada sobre uma amostra aleatória com reposição do conjunto original (Schonlau; Zou, 2020). A predição final do modelo é obtida por votação majoritária, no caso da classificação, ou pela média das predições individuais, no caso da regressão, reduzindo significativamente o risco de sobreajuste (Zhang et al., 2021).

Entre os principais hiperparâmetros do *Random Forest,* destaca-se o número de árvores na floresta (*n_estimators*), que influencia diretamente o desempenho do modelo. Em geral, um maior número de árvores leva a uma redução da variância, mas com custo computacional crescente e ganhos marginais após certo ponto. Outro hiperparâmetro essencial é o número máximo de atributos considerados em cada divisão de nó (*max_features*), cuja definição afeta a diversidade entre as árvores e, por consequência, a eficácia da agregação. Valores mais baixos promovem maior variabilidade entre as árvores e tendem a melhorar a generalização do modelo (Schonlau; Zou, 2020).

A profundidade máxima permitida para as árvores (*max_depth*) também é um parâmetro relevante, pois regula a complexidade dos modelos individuais. Árvores muito profundas podem memorizar os dados de treinamento, acarretando em sobreajuste, enquanto

profundidades mais rasas favorecem a generalização, embora possam deixar de capturar padrões mais complexos. Associado a esse controle, os parâmetros *min_samples_split* e *min_samples_leaf* determinam, respectivamente, o número mínimo de amostras exigido para dividir um nó interno e o número mínimo de amostras que um nó terminal deve conter (Zhang et al., 2021).

Além disso, o critério de impureza utilizado para as divisões (Gini ou a entropia) no caso da classificação determina a maneira como os nós são particionados. Embora ambos apresentem desempenhos semelhantes em muitos contextos, pequenas variações podem ocorrer a depender da distribuição dos dados (Schonlau; Zou, 2020).

O *Random Forest* é especialmente vantajoso em cenários com grande número de variáveis explicativas, relações não lineares e dados ruidosos, apresentando excelente desempenho preditivo sem exigir extensivo pré-processamento ou normalização dos dados (Reddy et al., 2020).

2.5.4 Support Vector Machine (SVM)

O *Support Vector Machine (SVM)* é um algoritmo de aprendizado supervisionado amplamente utilizado para tarefas de classificação e regressão, sendo reconhecido por sua base teórica sólida e pela capacidade de lidar com conjuntos de dados de alta dimensionalidade (Uddin et al., 2019). Seu funcionamento consiste na construção de um hiperplano ótimo que maximiza a margem entre as classes no espaço de características, utilizando apenas os vetores de suporte, ou seja, amostras localizadas próximas à fronteira de decisão (Sidey-Gibbons; Sidey-Gibbons, 2019).

Entre os principais hiperparâmetros do SVM, destaca-se o parâmetro de custo C, que regula o equilíbrio entre a maximização da margem e a penalização por erros de classificação. Valores mais altos de C levam o modelo a tentar classificar corretamente todas as instâncias do treinamento, o que pode resultar em sobreajuste. Por outro lado, valores mais baixos permitem erros na classificação (Makridakis; Spiliotis; Assimakopoulos, 2018).

Em problemas de regressão, é introduzido o hiperparâmetro ε (epsilon), que define uma zona de tolerância em torno da função de predição, dentro da qual os erros não são penalizados. Isso torna o modelo mais robusto a pequenas flutuações nos dados, reduzindo a sensibilidade a ruídos (Uddin et al., 2019).

Outro hiperparâmetro essencial é o γ, particularmente relevante quando se utilizam

funções kernel não lineares, como o radial. O valor de γ determina a influência de cada ponto de dados na construção do modelo: valores baixos conferem maior generalização, já valores altos restringem essa influência a regiões muito próximas (Brisimi et al., 2018).

A escolha da função kernel é um dos aspectos mais críticos do SVM, pois define como os dados serão transformados em um novo espaço de características. O kernel linear é indicado para problemas em que as classes são aproximadamente separáveis por uma reta ou plano. Já o kernel polinomial permite capturar interações não lineares entre as variáveis por meio do ajuste do grau do polinômio. O kernel radial, por sua vez, é amplamente utilizado por sua capacidade de modelar relações complexas e não lineares com flexibilidade (Sidey-Gibbons; Sidey-Gibbons, 2019).

3. METODOLOGIA CRISP-DM

O CRISP-DM (Cross-Industry Standard Process for Data Mining) é um processo utilizado com frequência na área de ciência de dados. Essa metodologia tem como intuito garantir que os dados sejam tratados de forma confiável e os modelos construídos sejam avaliados de forma adequada (Schröer; Kruse; Gómez, 2021).

Desse modo, o CRISP-DM é um modelo independente para executar projetos de mineração de dados, tendo sido desenvolvido no final da década de 1990, mas que permanece relevante dentro do cenário tecnológico, apesar das diversas mudanças ao longo dos anos (Schröer; Kruse; Gómez, 2021).

O processo é constituído de seis fases, que são: Entendimento do Negócio (Business Understanding), Entendimento dos Dados (Data Understanding), Preparação dos Dados (Data Preparation), Modelagem (Modeling), Avaliação (Evaluation) e Implantação (Deployment) (Ramos et al., 2020).

1. Entendimento do Negócio (Business Understanding)

O entendimento do negócio possui como foco entender os objetivos do projeto com uma perspectiva empresarial, convertendo esse conhecimento em uma definição de um problema de mineração de dados. Sendo assim, um planejamento pode ser desenvolvido a fim de atingir esses objetivos (Shearer, 2000).

2. Entendimento dos Dados (Data Understanding)

O entendimento dos dados inicia com a coleta dos dados, para que assim o analista consiga verificar a qualidade das informações. Diante desse quadro, nessa fase observa-se possíveis problemas encontrados (como por exemplo, dados faltantes ou espaços em branco), além de percepções e hipóteses sobre o assunto abordado (Shearer, 2000).

3. Preparação dos Dados (Data Preparation)

A preparação dos dados envolve todas as atividades necessárias para construir o conjunto de dados final que será utilizado na modelagem. Nessa fase, está inclusa a seleção, limpeza, construção, integração e a formatação dos dados, mudando de forma a ficar com a maior eficiência possível para a próxima etapa do CRISP-DM (Shearer, 2000).

4. Modelagem (Modeling)

Nessa fase, algumas técnicas de modelagem são selecionadas e aplicadas visando possuir uma boa modelagem para o problema. Diante disso, são construídos e

avaliados modelos com base nas técnicas escolhidas (Shearer, 2000).

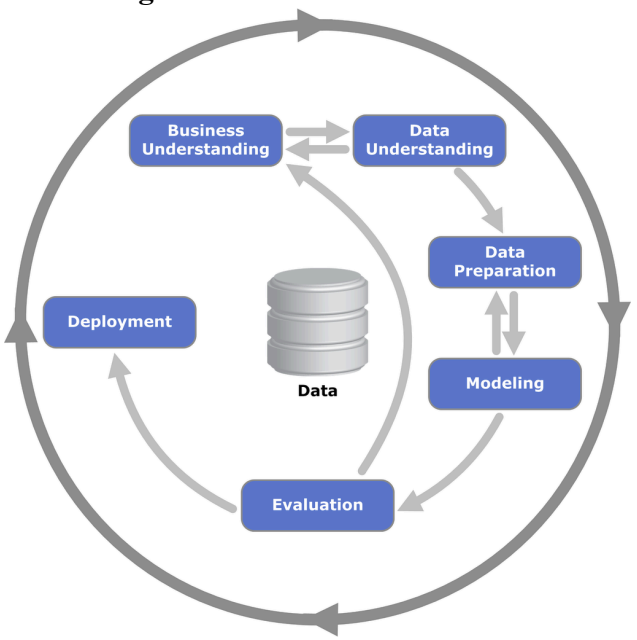
5. **Avaliação (Evaluation)**

Antes de implementar o modelo, é necessário avaliar de forma mais detalhada e revisar sua construção com o intuito de garantir que ele atinja os objetivos do negócio. Ademais, é importante determinar se alguma questão comercial não foi considerada anteriormente. Por último, o líder do projeto deve nessa fase, decidir como utilizar os resultados da mineração de dados, além de decidir os próximos passos do projeto (Shearer, 2000).

6. **Implantação (Deployment)**

Nessa última fase, é necessário desenvolver e documentar um plano para implementar o modelo construído nas fases anteriores do CRISP-DM. Sendo assim, esse plano deve conter um resumo do que foi elaborado em todas as etapas do projeto, além de realizar uma revisão do que foi bem implementado e do que poderia ter sido diferente para melhorar no futuro. Esse documento é importante para monitorar os resultados do projeto e manter um processo de melhoria contínua (Shearer, 2000).

Figura 2 - Processo CRISP-DM



Fonte: Shearer, 2000

4. ESTUDO DE CASO

4.1 ENTENDIMENTO DO NEGÓCIO

4.1.1 Mercado de Fast Fashion e Objetivo

Segundo a metodologia, tem-se como primeiro marco para uma análise de demanda assertiva o entendimento do negócio, fase na qual as atividades projetuais estarão focadas na determinação dos reais objetivos do projeto, ponderando a todo tempo as premissas do objeto de estudo, assim como suas restrições e características que o diferenciam dos demais concorrentes da indústria em que se insere. Nesse contexto, o objetivo central do desafio proposto é a captação de talentos na área de análise de dados, a fim de que estes possam auxiliar a empresa *Segrob Notlad* no desenvolvimento de um modelo preditivo capaz de estimar a demanda diária por camisetas básicas ao longo do mês de dezembro de 2024.

Entre as premissas estabelecidas no desafio está a flexibilidade do escopo a ser trabalhado, o que pode representar obstáculos adicionais à equipe envolvida com as adições de dados e mudanças oriundas do conglomerado estratégico organizacional. Tal complexidade decorre, também, da própria natureza do mercado de *fast fashion*, que equivale a uma cadeia de suprimentos internacional altamente dinâmica, composta por uma ampla rede de fornecedores, distribuidores e clientes conectados por fluxos de material, informação e capital (Oliveira, 2017). Diante disso, concretizar o objetivo da organização representa uma vantagem estratégica significativa, possibilitando uma gestão de estoques mais eficiente, alinhamento da produção à demanda real e redução de custos operacionais. Além disso, o sucesso na previsão fortalece a tomada de decisão baseada em dados, promovendo maior competitividade e inovação frente ao mercado (Giri; Chen, 2022).

O setor de *fast fashion*, por sua vez, é caracterizado pela produção acelerada e em grande escala de peças que acompanham as últimas tendências, com o intuito de disponibilizar novos produtos nas lojas em ciclos curtos e a preços acessíveis. Esse modelo demanda elevada agilidade nos processos de criação, produção e distribuição, sendo altamente sensível a fatores externos como sazonalidade, comportamento do consumidor, tendência e eventos pontuais (Luo; Chang; Xu, 2022).

4.1.2 Caracterização da Empresa

A *Segrob Notlad* é uma empresa brasileira do setor de moda varejista, inserida no

segmento de *fast fashion*. Fundada no Rio de Janeiro, a empresa consolidou-se nacionalmente com mais de 80 lojas distribuídas pelo território brasileiro, além de expandir sua presença para pontos estratégicos da América do Sul e três lojas conceito na Europa. Sua base de operações permanece concentrada no estado do Rio de Janeiro.

O público-alvo da organização é composto majoritariamente por jovens urbanos e digitalmente conectados, que valorizam produtos com apelo estético contemporâneo, preços acessíveis e campanhas diferenciadas. A marca diferencia-se pelo posicionamento cosmopolita e pela comunicação ousada, frequentemente ancorada em campanhas que promovem diversidade e crítica social, o que fortalece o engajamento com seu nicho de mercado.

A empresa opera em um ambiente de elevada volatilidade, tanto por fatores internos (necessidade constante de atualização de coleções) quanto externos, incluindo sazonalidade, comportamento de consumo imprevisível e flutuações macroeconômicas. Para lidar com essas variáveis, a *Segrob Notlad* tem investido sistematicamente em inteligência de mercado, automação da cadeia de suprimentos e, mais recentemente, em soluções baseadas em inteligência artificial.

A estrutura organizacional é pautada por um modelo de gestão orientado por dados, no qual decisões operacionais e estratégicas são cada vez mais baseadas em análises preditivas e ferramentas analíticas. Essa orientação tecnológica visa aprimorar a eficiência logística, reduzir perdas por excesso ou falta de estoque, e aumentar a acurácia das previsões de demanda.

No contexto atual, a empresa enfrenta o desafio de reestruturar sua estratégia de abastecimento, buscando soluções que otimizem a previsão de demanda. Tal desafio se insere em um esforço mais amplo de inovação empresarial, que reflete não apenas a competitividade do setor, mas também a crescente complexidade das decisões gerenciais em mercados orientados por dados.

4.2 ENTENDIMENTO DOS DADOS

Conforme mencionado anteriormente, os dados utilizados neste projeto são provenientes do *dataset* de vendas de camisetas pretas disponibilizado pela empresa *Segrob Notlad*. A primeira etapa consistiu em uma análise exploratória, cujo objetivo foi compreender melhor o conjunto de dados e embasar a elaboração de uma previsão de

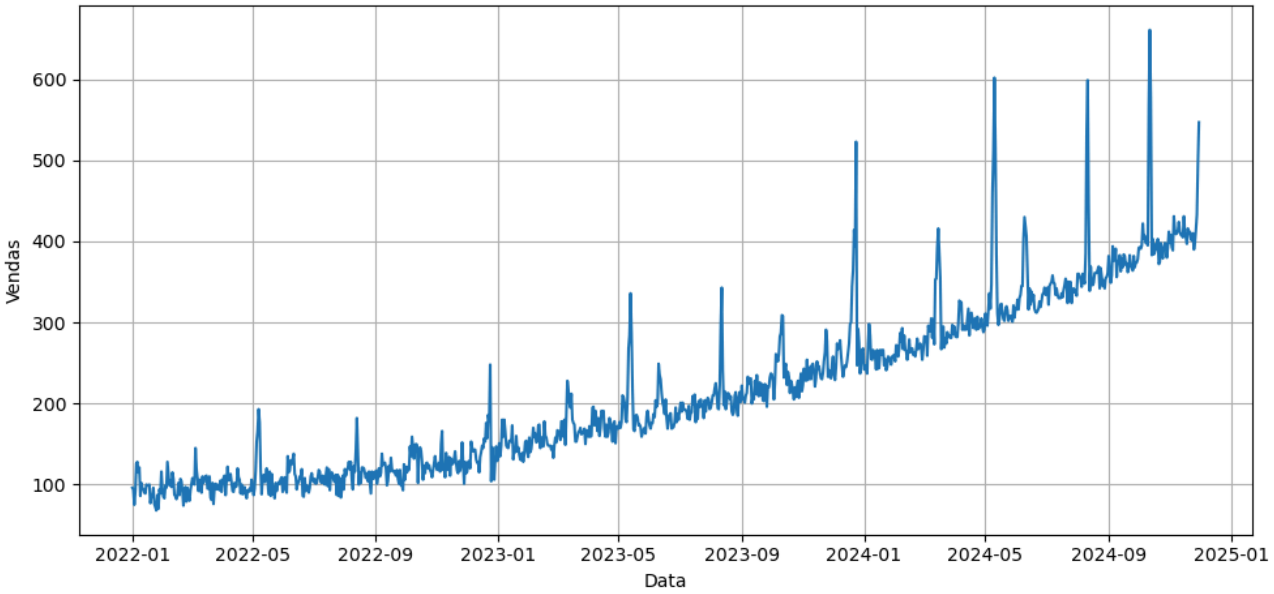
demanda mais precisa. Nessa fase inicial, todos os dados foram considerados relevantes, e verificou-se a inexistência de valores ausentes ou zerados nas colunas correspondentes ao período de 2022 a 2024. Para facilitar a visualização e a identificação de possíveis tendências, foram elaborados gráficos que proporcionam uma visão mais clara dos padrões existentes.

Apesar dessa base consistente, uma das principais limitações do projeto reside na restrição do conjunto de dados disponível, que contempla, inicialmente, apenas as variáveis de data e quantidade vendida de camisetas básicas. A ausência de atributos complementares, como indicadores promocionais, dados meteorológicos, feriados e localização das vendas, compromete a capacidade do modelo de capturar com precisão os padrões sazonais e as variações externas que influenciam o comportamento de compra. Essa limitação representa um desafio à qualidade das previsões, exigindo abordagens cautelosas e, idealmente, a futura incorporação de variáveis adicionais que possam enriquecer a análise e aumentar a robustez do modelo.

O gráfico 1 apresenta a série temporal das vendas de camisetas básicas masculinas, evidenciando uma tendência clara e consistente de crescimento ao longo do período analisado. Observa-se uma elevação gradual no volume de vendas, acompanhada por picos sazonais recorrentes, que se intensificam nos anos mais recentes. Esse padrão sugere não apenas uma ampliação da base de clientes, mas também possíveis melhorias na atuação comercial da empresa, como campanhas promocionais mais eficazes ou maior presença no mercado.

Além disso, a presença simultânea de tendência e sazonalidade indica que o comportamento das vendas segue uma estrutura relativamente estável, o que é favorável à aplicação de modelos preditivos baseados em séries temporais. A variação nos picos também pode sinalizar oportunidades para análises mais específicas, como identificar períodos com maior retorno sobre ações de marketing (dia das mães, dia dos pais, natal e etc).

Gráfico 1 - Dados Série Temporal

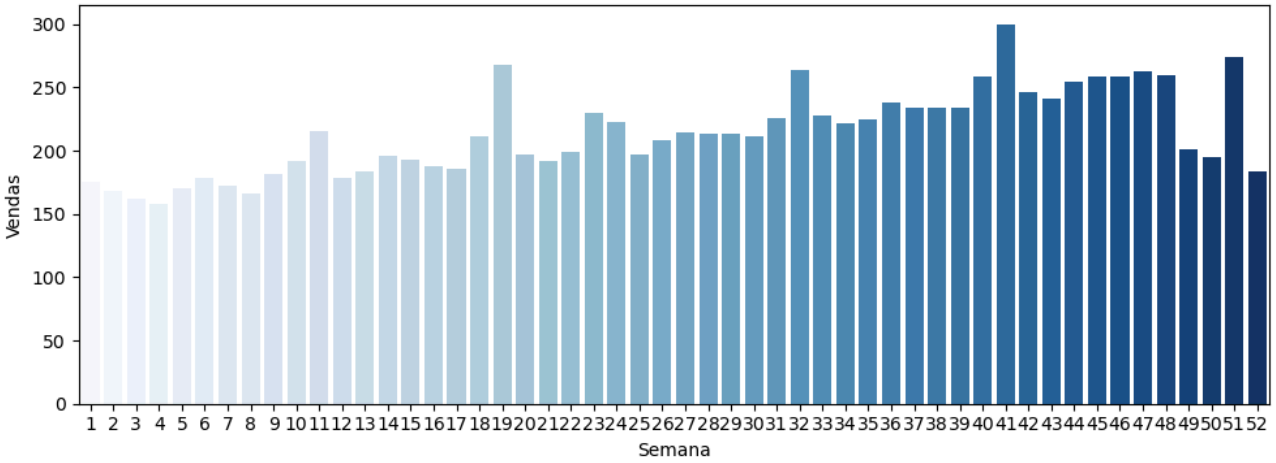


Fonte: Autoria Própria

O gráfico 2 apresenta a média de vendas por semana do ano, evidenciando um padrão sazonal bem definido ao longo das 52 semanas. Nota-se uma elevação gradual no volume médio de vendas a partir da semana 22, com picos significativos nas semanas 32 e 51, possivelmente associados a períodos promocionais estratégicos, como datas comemorativas e ações de fim de ano, respectivamente. Essa distribuição sugere que a demanda é influenciada por fatores sazonais recorrentes, o que pode ser explorado para planejamento de estoque e campanhas de marketing.

A tendência de crescimento ao longo do ano também é perceptível no aumento do patamar médio das barras, o que reforça os indícios de expansão observados na série temporal do gráfico 1. No entanto, é importante destacar que a redução nas vendas nas últimas semanas não representa, necessariamente, uma queda real de demanda, esse comportamento é explicado pela ausência de dados de dezembro de 2024, o que compromete a completude do período analisado.

Gráfico 2 - Vendas por Semana



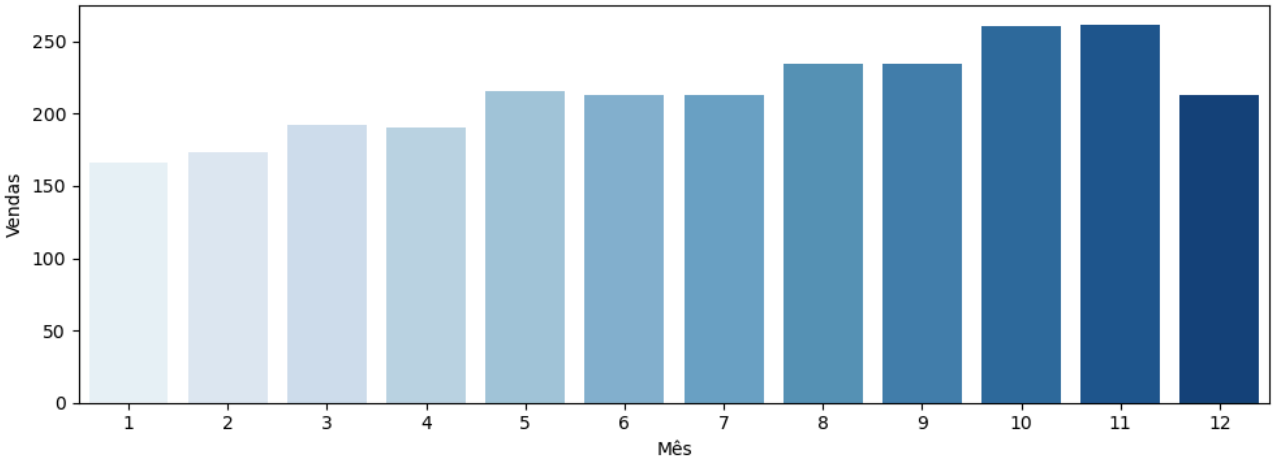
Fonte: Autoria Própria

O gráfico 3 exibe a média mensal de vendas ao longo do ano, evidenciando uma tendência de crescimento contínuo a partir de maio (mês 5), com picos em outubro e novembro (meses 10 e 11), quando a média mensal ultrapassa 250 unidades. Esse comportamento pode estar relacionado a ações promocionais ou campanhas.

Os quatro primeiros meses do ano apresentam médias inferiores a 200 unidades, sugerindo um período inicial de menor número de vendas, o que pode refletir tanto sazonalidade do mercado quanto o impacto de fatores externos como clima ou calendário promocional reduzido.

Destaca-se também a queda observada em dezembro (mês 12), que contrasta com os meses anteriores de alta performance. No entanto, essa diminuição é decorrente da incompletude dos dados de dezembro de 2024, o que compromete a representatividade da média para esse mês e não deve ser interpretado como um declínio real na demanda.

Gráfico 3 - Média Mensal de Vendas

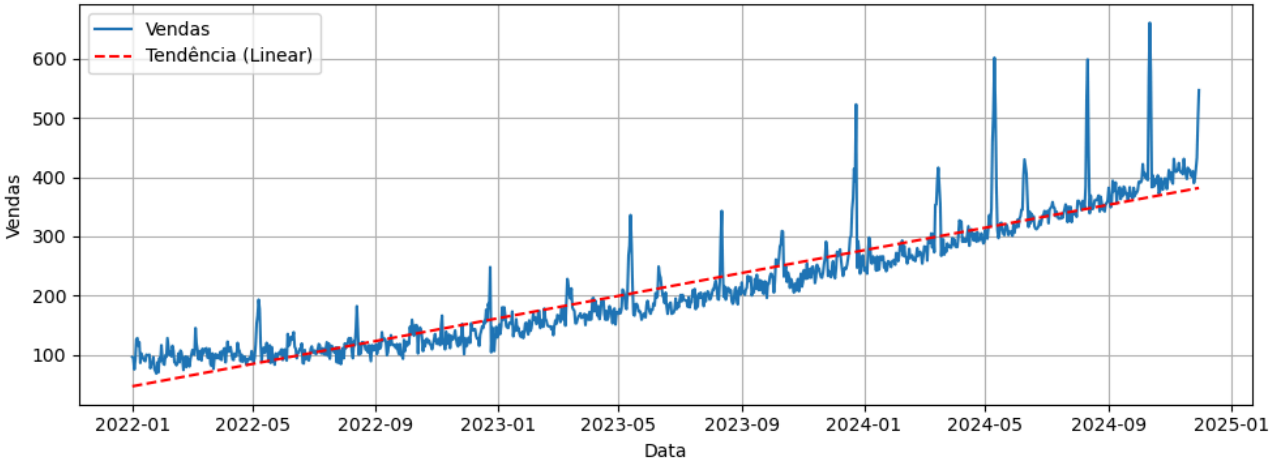


Fonte: Autoria Própria

O gráfico 4 reforça a presença de uma tendência de crescimento nas vendas de camisetas básicas masculinas entre 2022 e 2024. A linha azul demonstra oscilações regulares, com picos de alta intensidade, que coincidem com padrões sazonais previamente observados. No entanto, o destaque do gráfico é a linha de tendência linear (vermelha tracejada), que evidencia um aumento constante relevante no volume de vendas ao longo do tempo.

Esse crescimento linear indica que a base de clientes e a eficácia das estratégias de vendas vêm se expandindo. A consistência da tendência ascendente sugere estabilidade do modelo de negócio e potencial de escalabilidade, o que é particularmente relevante para decisões estratégicas em áreas como capacidade operacional, planejamento de estoque, contratação de pessoal e investimentos em marketing.

Gráfico 4 - Vendas por Semana (Tendência)



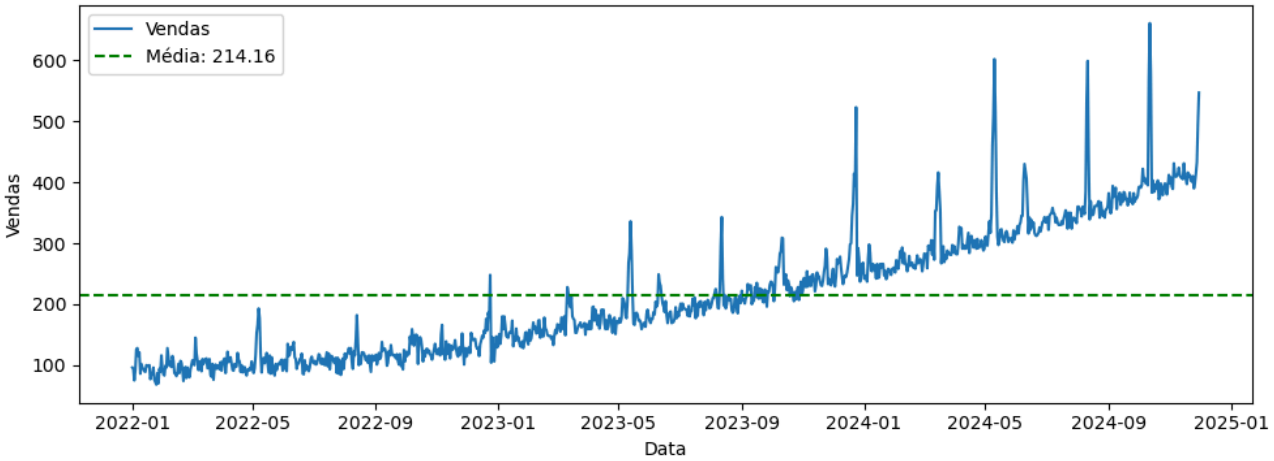
Fonte: Autoria Própria

Os gráficos 5 e 6, juntamente com a figura 3, oferecem uma visão estatística complementar que aprofunda a compreensão sobre o comportamento das vendas ao longo do tempo. O gráfico 5 apresenta a linha da média global de vendas, fixada em 214,16 unidades, destacando visualmente como grande parte das observações diárias flutuam em torno desse valor. A análise é enriquecida pela figura 3, que mostra um desvio padrão de 103,57, revelando uma alta dispersão dos dados e reforçando a existência de variações significativas no volume de vendas ao longo do período.

O gráfico 6 evidencia os limites extremos das vendas, com um valor mínimo de 68 unidades e um valor máximo de 661 unidades, indicando uma amplitude considerável nas quantidades comercializadas. Esses valores extremos ajudam a compreender a magnitude dos picos e vales nas vendas, e podem estar associados a fatores sazonais, promocionais ou operacionais. A distância entre os limites também sugere que, apesar da média fornecer uma

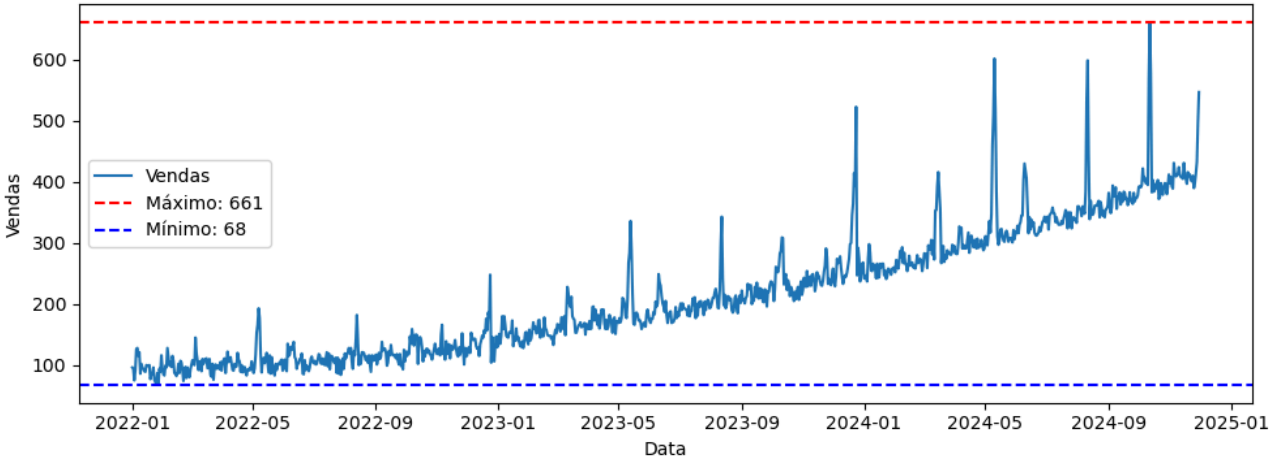
estimativa central, há grande volatilidade no comportamento de consumo.

Gráfico 5 - Média de Vendas



Fonte: Autoria Própria

Gráfico 6 - Máximo e Mínimo de Vendas



Fonte: Autoria Própria

Figura 3 - Informações Descritivas

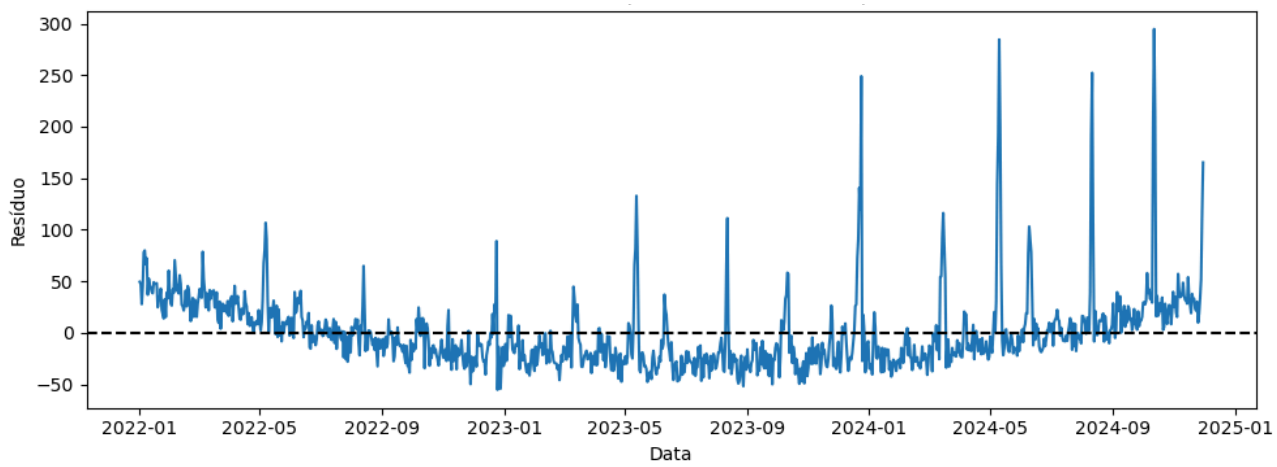
Média das Vendas	Desvio Padrão	Valor Máximo	Valor Mínimo
214.16	103.57	661	68

Fonte: Autoria Própria

O gráfico 7 apresenta os resíduos do modelo, evidenciando a variação das vendas em relação à tendência linear esperada. Essa visualização permite identificar desvios e verificar se os erros se comportam como ruído branco. Observa-se que, ao longo do período analisado, os resíduos flutuam em torno de zero, conforme indicado pela linha preta tracejada, o que confirma que, em média, os erros não apresentam viés. No entanto, há picos esporádicos, tanto positivos quanto negativos, que representam eventos atípicos. Esse comportamento é relevante

para a avaliação da qualidade do modelo, pois a predominância de resíduos próximos de zero sugere um bom ajuste, enquanto os picos indicam possíveis oportunidades de refinamento, seja por meio da inclusão de variáveis explicativas adicionais ou da adoção de uma abordagem alternativa de modelagem.

Gráfico 7 - Ruído Branco

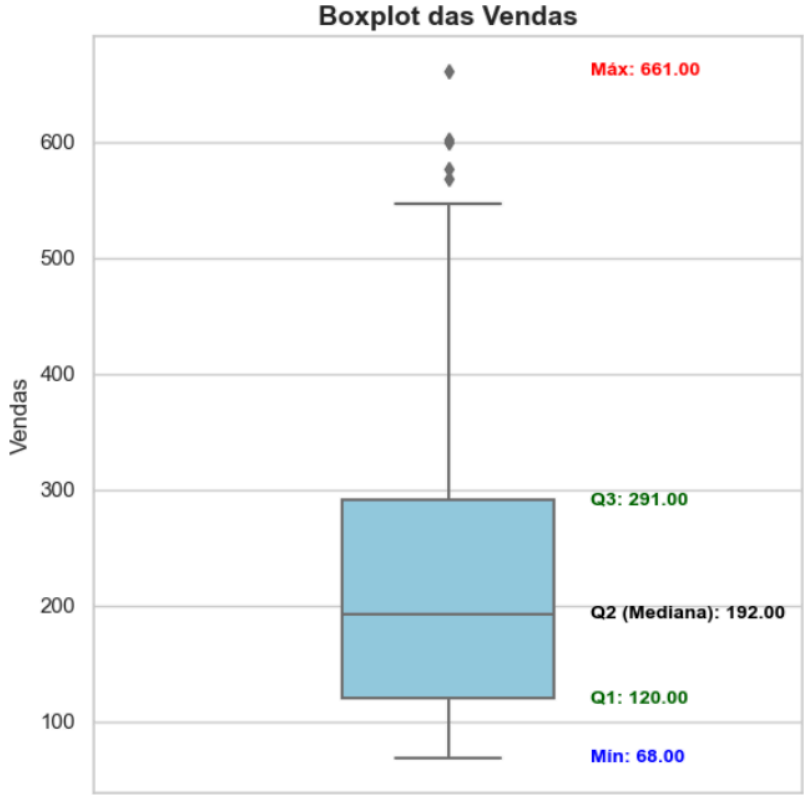


Fonte: Autoria Própria

O gráfico 8 apresenta um boxplot das vendas, que permite visualizar a distribuição dos dados, a dispersão e a presença de valores atípicos (*outliers*) ao longo do período analisado. O gráfico fornece informações importantes sobre a distribuição estatística das vendas, com destaque para os quartis: o primeiro quartil (Q1) foi de 120,00 unidades, a mediana (Q2) foi de 192,00 unidades, e o terceiro quartil (Q3) atingiu 291,00 unidades.

Os valores extremos observados foram 68,00 unidades (mínimo) e 661,00 unidades (máximo), indicando uma ampla amplitude de variação. Além disso, é possível notar a presença de *outliers* acima do limite superior, representando picos de vendas consideravelmente acima do padrão esperado. Esses picos fora do padrão esperado estão associados a datas comemorativas, como Dia das Mães, Dia dos Pais e Dia das Crianças.

Gráfico 8 - Boxplot

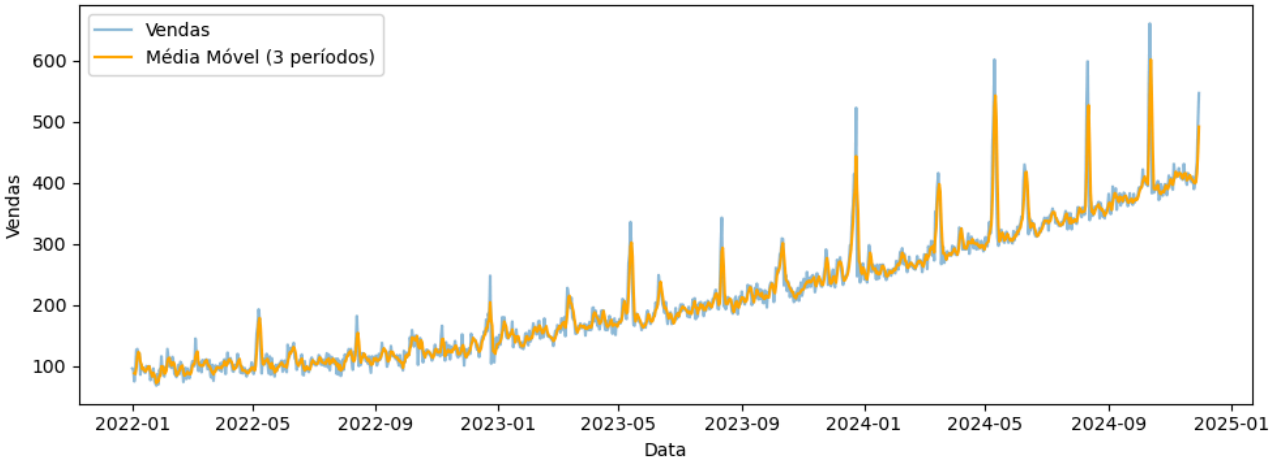


Fonte: Autoria Própria

O gráfico 9 apresenta a média móvel das vendas ao longo do período analisado, com o objetivo de suavizar as oscilações diárias e realçar tendências de comportamento dos dados. A linha laranja representa a média móvel de 3 períodos, que acompanha de maneira eficiente a série temporal original (linha azul), reduzindo a influência de variações pontuais e proporcionando uma visão mais clara do movimento geral das vendas.

Observa-se, com o auxílio da média móvel, uma tendência de crescimento gradual ao longo do tempo, bem como a repetição de picos sazonais, já identificados anteriormente em datas comemorativas específicas. Essa abordagem é especialmente útil em análises preditivas, uma vez que permite identificar padrões recorrentes e orientar a tomada de decisões estratégicas.

Gráfico 9 - Média Móvel



Fonte: Autoria Própria

4.3 PREPARAÇÃO DOS DADOS

4.3.1 Identificação dos Outliers

A etapa de preparação dos dados tem como objetivo transformar os dados brutos em uma base estruturada para a aplicação de modelos preditivos. A partir da análise exploratória, especialmente com o auxílio do boxplot, foram identificados pontos fora da curva (outliers) nos seguintes dias: 10/05/2024 (602 vendas), 11/08/2024 (599 vendas), 11/10/2024 (568 vendas), 12/10/2024 (661 vendas) e 13/10/2024 (577 vendas).

Com base em uma investigação contextual, observou-se que esses picos estão associados a datas comemorativas relevantes. O elevado número de vendas em 10/05/2024 está possivelmente relacionado ao aumento da demanda em função do Dia das Mães, comemorado poucos dias depois. Da mesma forma, o pico em 11/08/2024 coincide com o período do Dia dos Pais. Já os aumentos registrados em 11/10, 12/10 e 13/10 referem-se ao Dia das Crianças (12/10), abrangendo o dia anterior, o próprio feriado e o dia subsequente.

Adicionalmente, foi identificado um pico de vendas em 30/11/2024, com 547 unidades vendidas, logo após a Black Friday. Embora este valor não tenha sido classificado como outlier pelo boxplot, ele representa um comportamento atípico e relevante para o modelo.

Optou-se por não excluir os outliers identificados na análise exploratória, uma vez que estes representam comportamentos sazonais recorrentes e relevantes para a previsão de demanda. As datas com volumes atípicos de vendas coincidem com eventos comemorativos e promocionais, como o Dia das Mães, Dia dos Pais, Dia das Crianças e o período pós-Black

Friday, os quais influenciam significativamente o comportamento do consumidor. A remoção desses registros poderia resultar na perda de informações valiosas sobre padrões temporais que se repetem anualmente. Em vez disso, serão criadas variáveis indicadoras para representar esses eventos, permitindo que o modelo reconheça e aprenda tais variações.

4.3.2 Criação de Variáveis

Com base na etapa anterior, foram criadas variáveis representativas para datas comemorativas de reconhecida relevância comercial: Dia das Mães, Dia dos Pais, Dia das Crianças e Black Friday. Essas datas são tradicionalmente associadas a campanhas promocionais intensas e a um aumento na disposição dos consumidores para realizar compras, o que justifica sua inclusão como potenciais variáveis explicativas. Por exemplo, o Dia das Mães e o Dia dos Pais estão entre as principais datas do calendário varejista nacional, impulsionando significativamente a venda de presentes.

Além dessas, também foram incluídas variáveis para o Natal e o Dia dos Namorados. Embora essas datas não tenham se destacado inicialmente nos resultados preliminares, sua possível influência sobre o comportamento do consumidor motivou a inclusão para fins exploratórios.

As variáveis comemorativas foram organizadas em formato booleano, indicando se determinada data estava presente em cada observação. Complementarmente, foram adicionadas lags para dias da semana, mês e ano com o intuito de capturar o comportamento da demanda em períodos anteriores, partindo da premissa de que a demanda passada influencia a atual.

A análise da relevância dessas variáveis será realizada por meio de testes estatísticos, com base nos coeficientes estimados em uma regressão. Esses coeficientes indicam o impacto das variáveis comemorativas sobre a variável dependente (vendas), permitindo avaliar se há diferenças estatisticamente significativas durante esses períodos.

Os resultados indicam que, com exceção do Natal, todas as variáveis analisadas para os anos de 2023 e 2024 apresentaram coeficientes positivos, sugerindo que essas datas têm, de fato, um efeito relevante sobre as vendas.

Figura 4 - Influência das Variáveis

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9012	3.113	1.896	0.058	-0.210	12.013
lag_1	0.6461	0.039	16.572	0.000	0.570	0.723
lag_2	-0.0999	0.046	-2.167	0.031	-0.190	-0.009
lag_3	-0.0663	0.046	-1.453	0.147	-0.156	0.023
lag_4	0.0240	0.046	0.526	0.599	-0.066	0.114
lag_5	0.0166	0.046	0.364	0.716	-0.073	0.106
lag_6	0.0575	0.045	1.266	0.206	-0.032	0.147
lag_7	0.0268	0.017	1.553	0.121	-0.007	0.061
lag_7	0.0268	0.017	1.553	0.121	-0.007	0.061
lag_30	0.0921	0.024	3.901	0.000	0.046	0.138
lag_365	0.4446	0.050	8.920	0.000	0.347	0.542
natal_2022	-1.382e-14	4.2e-15	-3.292	0.001	-2.21e-14	-5.58e-15
dia_namorados_2022	-1.927e-14	4.43e-15	-4.356	0.000	-2.8e-14	-1.06e-14
dia_criancas_2022	9.153e-15	4.25e-15	2.153	0.032	8.07e-16	1.75e-14
natal_2023	34.5088	9.277	3.720	0.000	16.294	52.723
dia_namorados_2023	2.0250	8.895	0.228	0.820	-15.440	19.490
dia_criancas_2023	19.2400	8.847	2.175	0.030	1.869	36.612
natal_2024	-2.821e-15	1.86e-15	-1.514	0.130	-6.48e-15	8.37e-16
dia_namorados_2024	6.5253	9.276	0.703	0.482	-11.687	24.738
dia_criancas_2024	28.9570	9.314	3.109	0.002	10.669	47.245
dia_maes_2022	-6.352e-16	4.68e-16	-1.358	0.175	-1.55e-15	2.83e-16
dia_maes_2023	20.7875	8.346	2.491	0.013	4.400	37.175
dia_maes_2024	38.0440	8.858	4.295	0.000	20.651	55.436
dia_pais_2022	0	0	nan	nan	0	0
dia_pais_2023	12.0835	8.295	1.457	0.146	-4.203	28.370
dia_pais_2024	27.1616	8.523	3.187	0.002	10.427	43.896
black_friday_2022	0	0	nan	nan	0	0
black_friday_2023	13.7688	8.241	1.671	0.095	-2.412	29.950
black_friday_2024	6.2764	8.478	0.740	0.459	-10.369	22.922

Fonte: Autoria Própria

4.4 MODELAGEM

4.4.1 Modelos Baseline

Na fase de modelagem, foram primeiramente aplicados modelos preditivos baseline, considerados métodos simples e eficazes para estabelecer uma linha de base de desempenho. O objetivo principal desses modelos foi compreender o comportamento da série temporal e identificar estratégias de previsão mais adequadas ao contexto do projeto e aos dados disponíveis.

Modelos baseline, no cenário de previsão de demanda, são abordagens que utilizam regras simples e de fácil implementação, geralmente sem exigir grande volume de parametrização ou treinamento. Eles funcionam como referência comparativa para métodos mais sofisticados: se um modelo avançado não superar os modelos baseline, é sinal de que ele não está agregando valor ao processo preditivo. Esses modelos são essenciais nas primeiras etapas de modelagem por proporcionarem previsões rápidas e de fácil interpretação (Hyndman;

Athanasopoulos, 2021).

Os métodos aplicados inicialmente foram o Naive, a Média Cumulativa, a Média Móvel Simples com janela de 30 dias, a Suavização Exponencial Simples (SES), a Suavização Exponencial Dupla (DES, também conhecida como método de Holt) e a Suavização Exponencial Tripla (TES ou método de Holt-Winters). Cada um desses métodos foi empregado com o objetivo de estimar a curva de demanda real, tomando como base os dados históricos disponíveis e para isso, em todos os métodos, a série temporal foi dividida em duas partes: dados de treino e dados de teste. Essa divisão tem como finalidade simular um cenário real de previsão, em que o modelo é treinado com dados históricos conhecidos (treino) e posteriormente avaliado com dados mais recentes (teste), que também são conhecidos, mas são ocultados do modelo durante o treinamento.

Essa abordagem permite comparar a curva prevista pelo modelo com a curva real, avaliando o quão próxima está a previsão da realidade e, sendo assim, a comparação tem como finalidade validar a eficiência de cada modelo e direcionar a escolha da abordagem preditiva final, aquela que oferecer melhor equilíbrio entre simplicidade e desempenho.

Para as observações históricas, os dados de treino abrangem o período de janeiro de 2022 até o final. A justificativa para esta escolha de 30 dias como intervalo de teste está alinhada ao objetivo final do modelo preditivo desenvolvido neste projeto: compreender o comportamento da demanda de camisetas básicas ao longo de 2024, com foco específico no mês de dezembro. Assim, prever os dados serve como uma etapa intermediária para validar o desempenho dos métodos antes da aplicação definitiva no mês de interesse. A seguir, são apresentados os métodos aplicados e seus respectivos gráficos, contemplando tanto a série temporal completa quanto um recorte do período final da base de dados. Esse recorte destaca a transição entre os dados de treino e os dados de teste, permitindo uma visualização mais detalhada da performance preditiva dos modelos no intervalo mais recente da série.

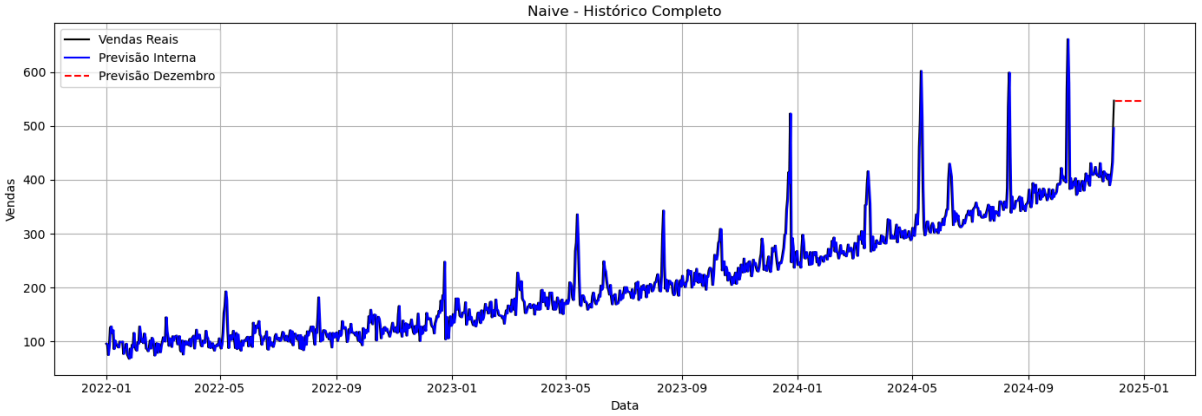
4.4.1.1 Naive

O modelo Naive assume que o próximo valor de previsão será igual ao último valor observado. É o modelo mais simples entre os métodos preditivos e serve como referência mínima de desempenho. Sua principal vantagem é a rapidez e facilidade de implementação, sendo especialmente útil quando a série apresenta pouca variação ou quando se deseja

estabelecer um comparativo inicial.

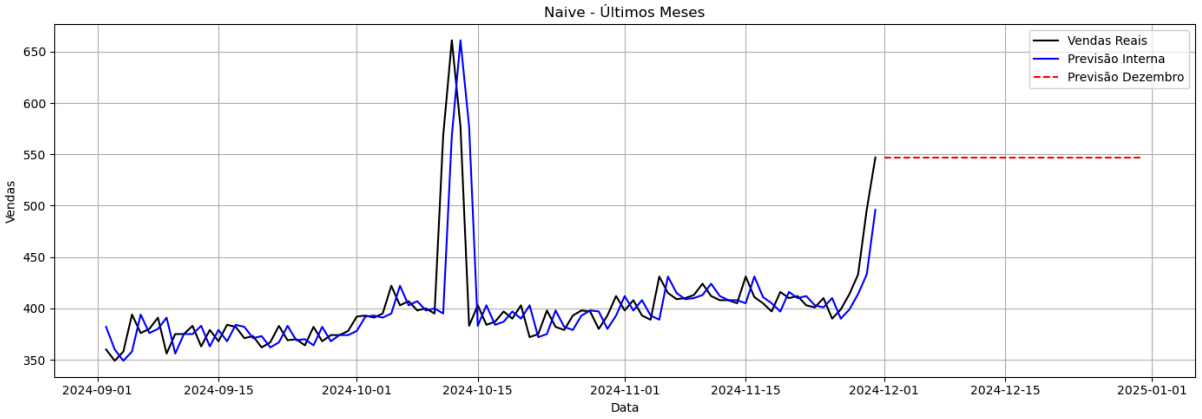
Conforme pode ser observado no gráfico 10, que apresenta toda a série temporal, a curva de previsão do modelo Naive segue exatamente o último ponto observado, estendendo-o ao longo do horizonte de previsão. Já no gráfico 11, é possível visualizar com mais clareza os últimos meses da série, evidenciando a sobreposição entre o último valor real e os valores previstos, o que reforça o funcionamento simples do modelo.

Gráfico 10 - Método Naive



Fonte: Autoria Própria

Gráfico 11 - Método Naive (Últimos Meses)



Fonte: Autoria Própria

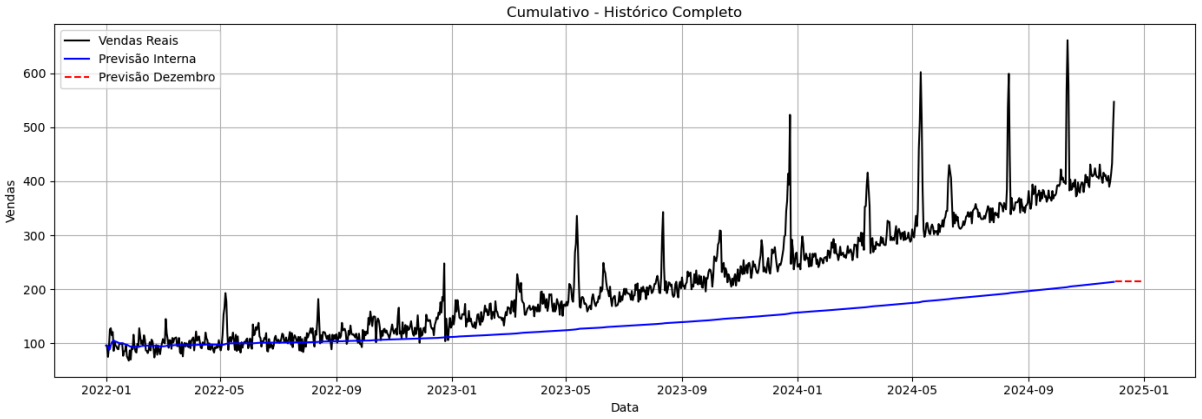
4.4.1.2 Média Cumulativa

Nesse método, a previsão é baseada na média aritmética de todos os valores observados até o momento da previsão. Ele suaviza ruídos ao longo do tempo, porém, não reage bem a mudanças repentinas na tendência ou na sazonalidade, sendo mais indicado para séries relativamente estáveis.

Conforme o gráfico 12, observa-se que a curva gerada pelo modelo acompanha o

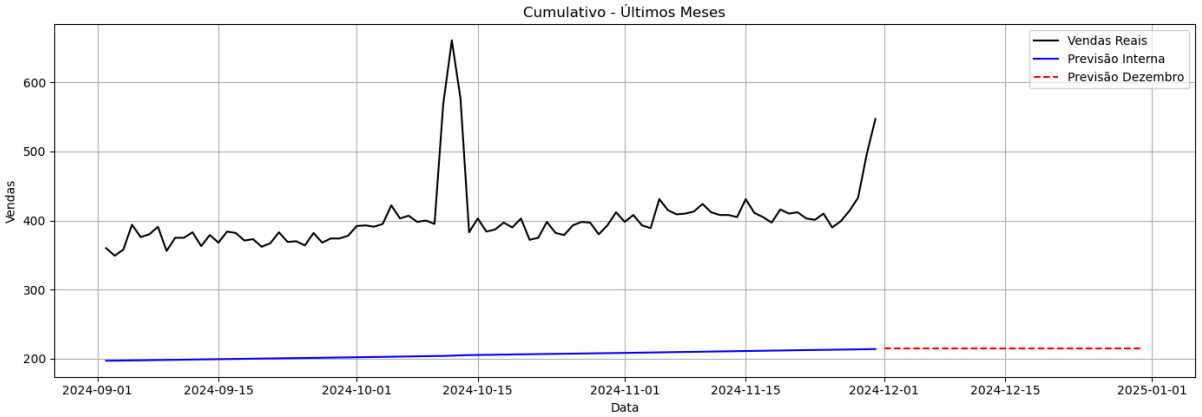
comportamento geral da série temporal. Já o gráfico 13 apresenta um recorte dos últimos meses da série, permitindo visualizar com maior clareza o desempenho da previsão em curto prazo.

Gráfico 12 - Média Cumulativa



Fonte: Autoria Própria

Gráfico 13 - Média Cumulativa (Últimos meses)



Fonte: Autoria Própria

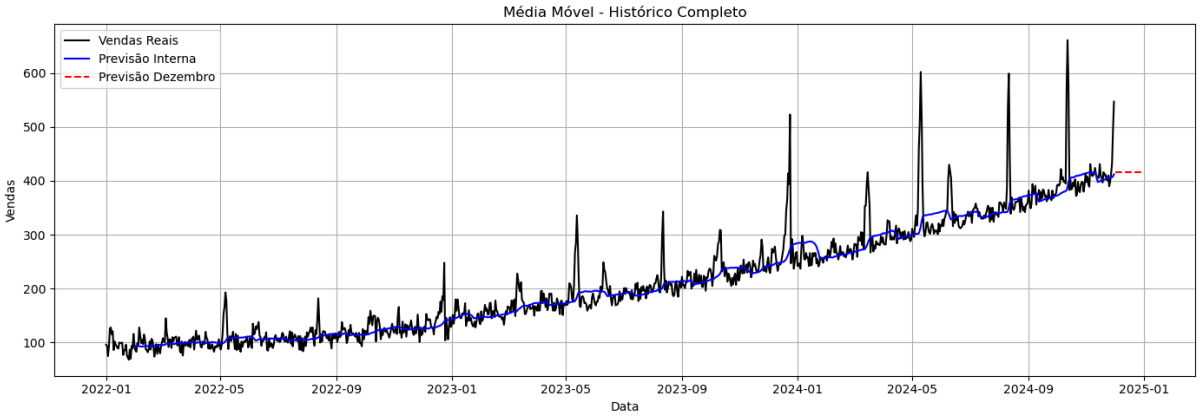
4.4.1.3 Média Móvel Simples (30 dias)

A média móvel simples calcula a média dos últimos n períodos — neste projeto, 30 dias — para gerar a previsão do próximo ponto da série. O uso de uma janela de 30 dias visa captar uma média mensal do comportamento da série, suavizando flutuações de curto prazo. No entanto, esse método não é capaz de capturar tendências de longo prazo ou padrões sazonais mais complexos.

Conforme apresentado no gráfico 14, que exibe toda a série temporal, é possível observar como a média móvel suaviza as oscilações nos dados originais, oferecendo uma visão mais estável do comportamento geral da variável ao longo do tempo. Já o gráfico 15

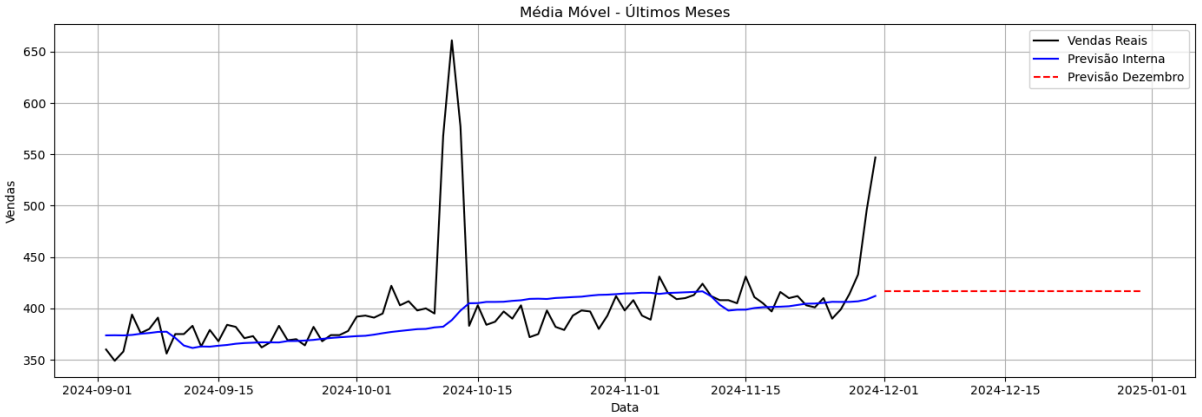
foca nos meses mais recentes, evidenciando a aplicação da média móvel para prever os valores futuros com base na média dos últimos 30 dias.

Gráfico 14 - Média Móvel



Fonte: Autoria Própria

Gráfico 15 - Média Móvel (Últimos Meses)



Fonte: Autoria Própria

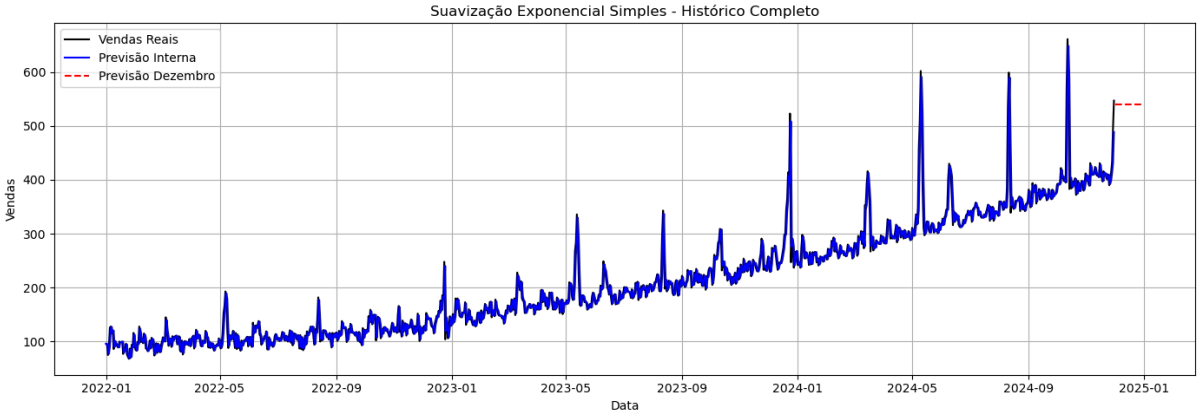
4.4.1.4 Suavização Exponencial Simples (SES)

A suavização exponencial simples atribui pesos decrescentes aos valores passados, conferindo maior importância às observações mais recentes por meio de um parâmetro de suavização α (alfa). Esse método é indicado para séries temporais que não apresentam tendência ou sazonalidade, sendo que seu desempenho está diretamente relacionado à estimativa adequada desse parâmetro.

Conforme ilustrado no gráfico 16, a aplicação da suavização exponencial simples permite acompanhar o comportamento da série completa ao longo do tempo, evidenciando a capacidade do modelo de capturar as variações mais recentes de forma responsiva. Já o gráfico 17 destaca os últimos meses da série, permitindo observar com maior clareza a projeção realizada pelo

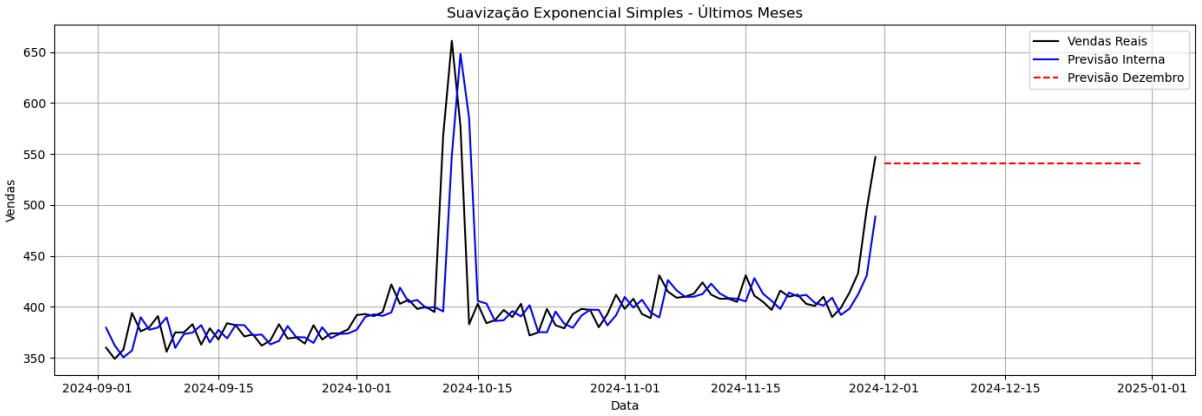
modelo e a aderência entre os valores previstos e os observados nesse intervalo mais recente.

Gráfico 16 - SES



Fonte: Autoria Própria

Gráfico 17 - SES (Últimos Meses)



Fonte: Autoria Própria

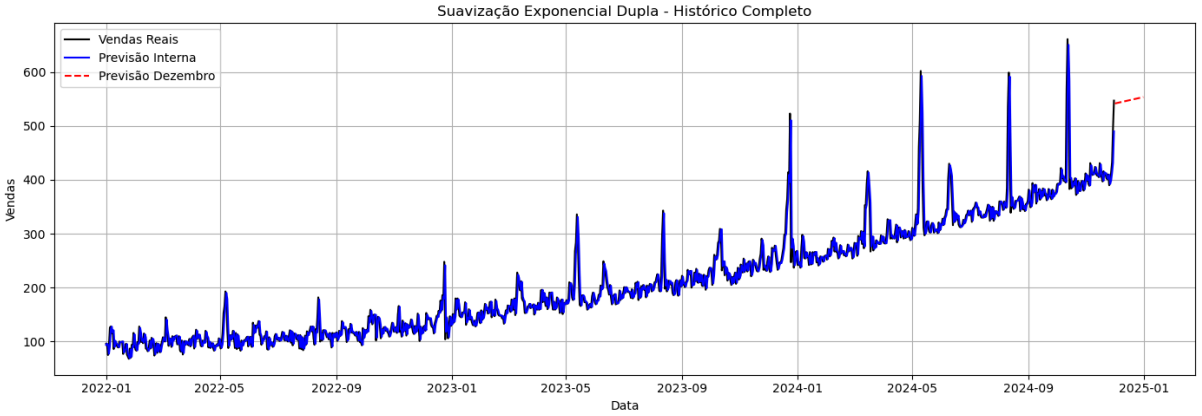
4.4.1.5 Suavização Exponencial Dupla (DES - Holt)

O método de Holt, também conhecido como suavização exponencial dupla, representa uma evolução em relação ao modelo SES (Suavização Exponencial Simples), ao incorporar um componente de tendência além do nível da série. Para isso, utiliza dois parâmetros: α (alfa), responsável por capturar o nível, e β (beta), que modela a tendência da série ao longo do tempo. Essa abordagem torna o modelo particularmente indicado para séries temporais que apresentam comportamentos de crescimento ou declínio consistentes.

Conforme pode ser observado no gráfico 18, a aplicação do método de Holt permite representar adequadamente toda a série temporal, destacando-se a adaptação da curva ajustada à tendência crescente dos dados. Já o gráfico 19, que foca nos últimos meses da série,

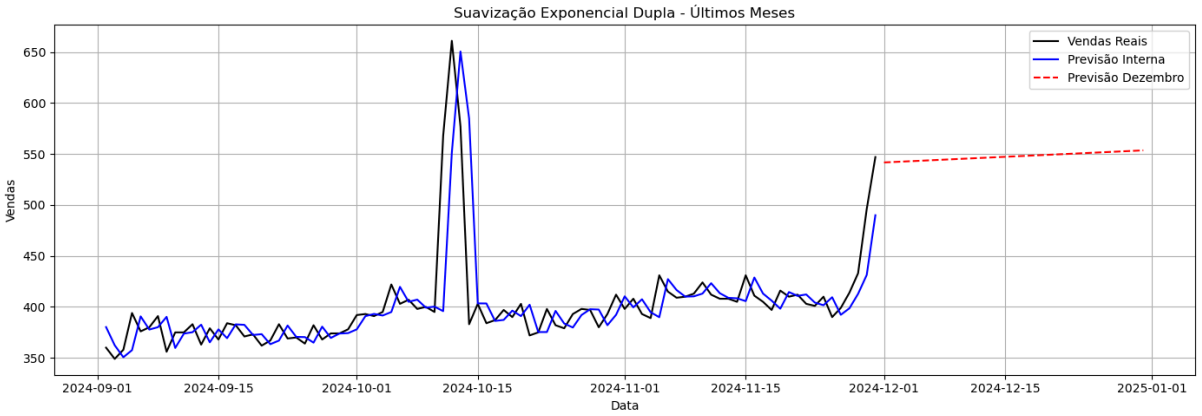
evidencia de forma mais clara a capacidade preditiva do modelo.

Gráfico 18 - DES



Fonte: Autoria Própria

Gráfico 19 - DES (Últimos Meses)



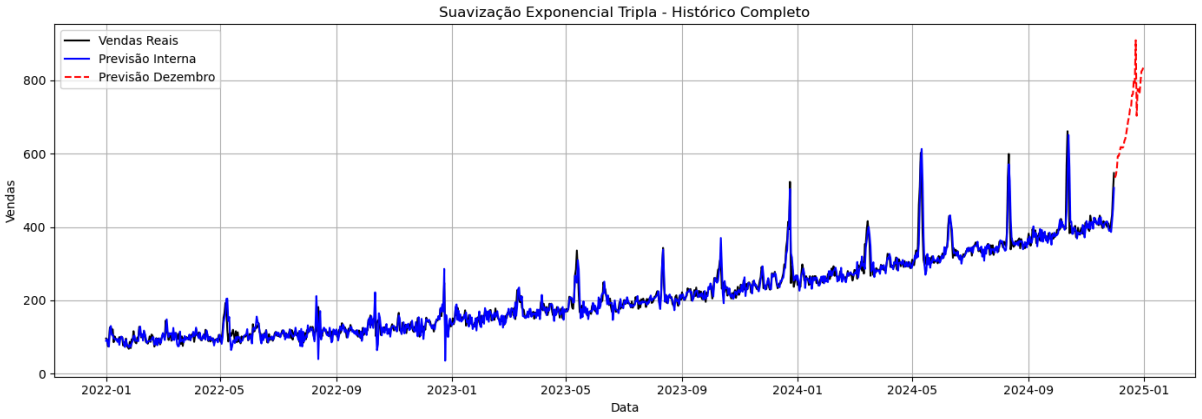
Fonte: Autoria Própria

4.4.1.6 Suavização Exponencial Tripla (TES - Holt-Winters)

Essa abordagem adiciona um terceiro componente ao modelo de Holt: a sazonalidade. Com isso, é possível prever séries que apresentam um comportamento sazonal recorrente. O método exige a definição de um período sazonal e envolve três parâmetros: nível (α), tendência (β) e sazonalidade (γ). Portanto, trata-se de uma das abordagens mais completas entre os modelos clássicos de séries temporais.

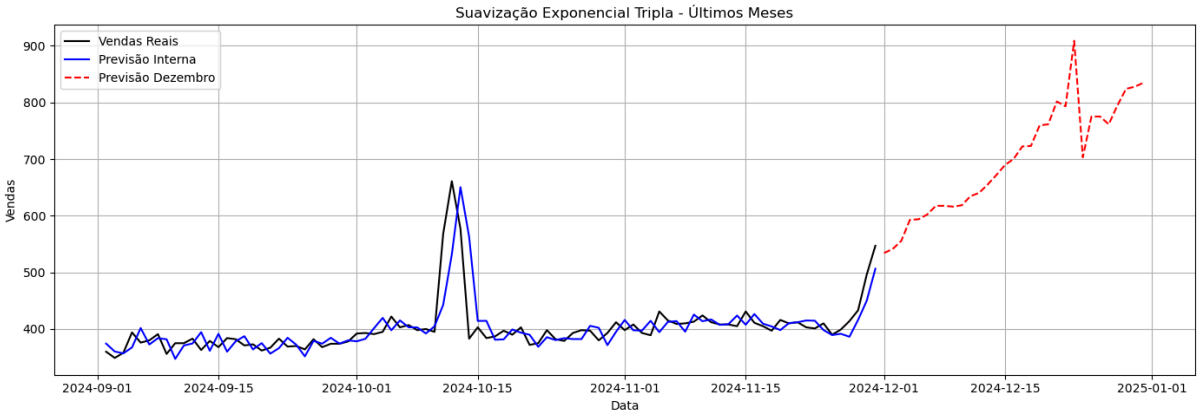
Conforme o gráfico 20, observa-se que o modelo é capaz de capturar o comportamento geral da série temporal ao longo de todo o período analisado, reproduzindo adequadamente os padrões sazonais. Já o gráfico 21, que foca nos últimos meses da série, evidencia a capacidade preditiva do modelo, mostrando que a curva ajustada segue de forma coerente as oscilações recentes e mantém a sazonalidade nas projeções.

Gráfico 20 - TES



Fonte: Autoria Própria

Gráfico 21 - TES (Últimos Meses)



Fonte: Autoria Própria

4.4.1.7 Coeficiente de Suavização

Com o objetivo de ajustar modelos de suavização exponencial às características da série temporal analisada, foi desenvolvido um código para estimar automaticamente os coeficientes dos modelos de Suavização Exponencial Simples (SES), Suavização Exponencial Dupla de Holt e Suavização Exponencial Tripla de Holt-Winters.

Os parâmetros obtidos foram os seguintes: para o modelo SES, o coeficiente de suavização (alfa) foi de 0,8872; para o modelo de Holt, os coeficientes alfa e beta foram de 0,9003 e 0,0001, respectivamente; e para o modelo de Holt-Winters, os coeficientes alfa, beta e gama foram de 0,8900, 0,0019 e 0,0242, respectivamente.

Observa-se que os valores de alfa são elevados em todos os modelos, indicando que as observações mais recentes da série têm maior influência nas previsões, o que é desejável em contextos onde os dados apresentam mudanças rápidas. Por outro lado, os valores muito

baixos de beta sugerem que a tendência da série é praticamente constante, ou sofre variações muito suaves ao longo do tempo. No caso do modelo Holt-Winters, o valor também reduzido de gama indica a presença de uma componente sazonal estável, com pouca variabilidade entre os ciclos.

Nesse contexto, o modelo Holt-Winters, apesar dos coeficientes reduzidos para tendência e sazonalidade, pode oferecer previsões mais robustas por incorporar múltiplas componentes estruturais, mesmo que com intensidade limitada.

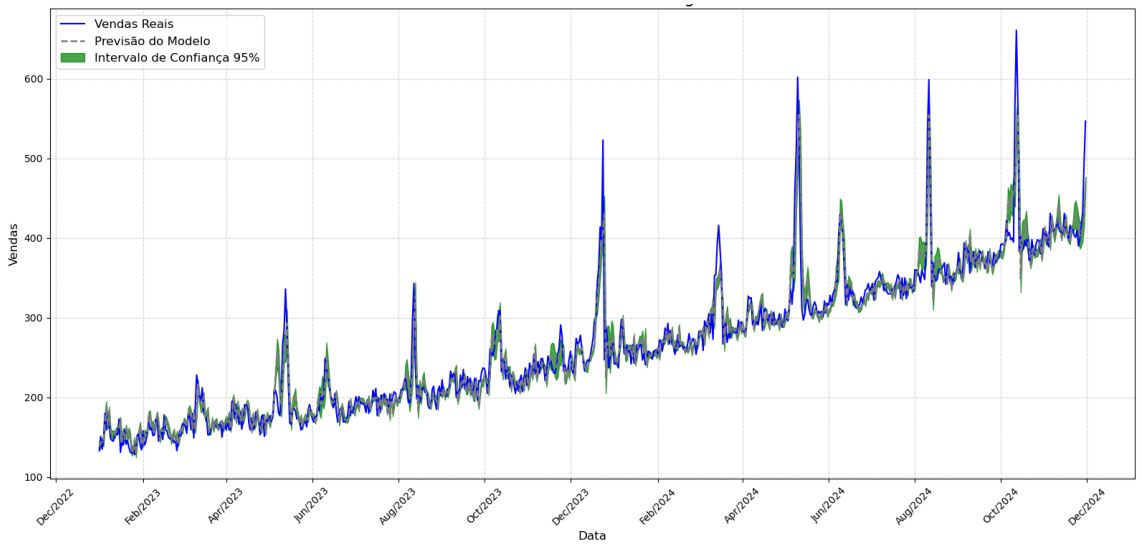
4.4.2 Regressão Dinâmica

Até o momento, haviam sido empregados modelos baseline, como médias móveis simples e métodos de suavização exponencial (simples, dupla e tripla), que, embora úteis como ponto de partida, apresentam limitações na capacidade de capturar padrões mais complexos nos dados, como efeitos sazonais específicos, comportamento promocional e influência de variáveis externas. Para superar essas limitações, optou-se pela adoção de um modelo de regressão dinâmica, uma abordagem estatística que permite incorporar variáveis explicativas ao processo de previsão, aumentando significativamente a capacidade interpretativa e preditiva do modelo.

Diferente dos modelos baseline, que se baseiam exclusivamente na estrutura temporal das séries, a regressão dinâmica possibilita a inclusão de variáveis como datas comemorativas, efeitos defasados de vendas e outros indicadores relevantes, tornando o modelo mais aderente à realidade comercial e comportamental observada no varejo. Além disso, esse tipo de modelo facilita a análise de significância estatística dos coeficientes, permitindo identificar quais fatores exercem maior influência sobre a variável dependente.

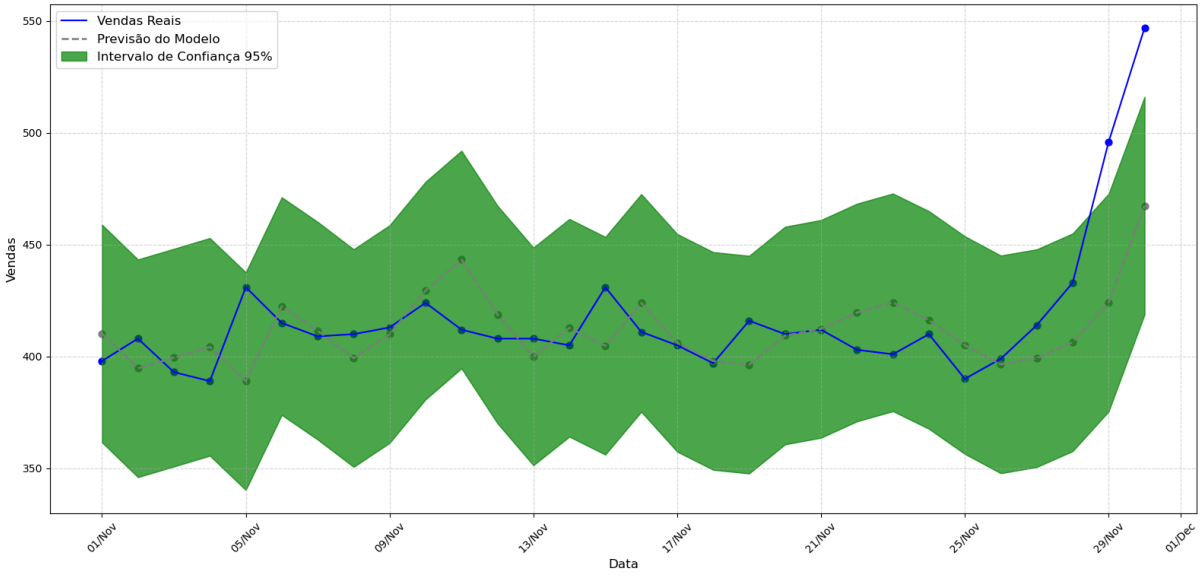
Com base nesse novo modelo, foram geradas previsões para o mês de dezembro, conforme apresentado a seguir. O gráfico 22 apresenta toda a série temporal juntamente com os valores ajustados e previstos, permitindo uma visão ampla do desempenho do modelo ao longo do tempo. Em complemento, o gráfico 23 destaca os últimos meses da série, facilitando a visualização da tendência recente e da projeção específica para dezembro.

Gráfico 22 - Regressão Dinâmica



Fonte: Autoria Própria

Gráfico 23 - Regressão Dinâmica (Novembro)



Fonte: Autoria Própria

4.4.3 Modelos Não Paramétricos

Após a aplicação dos modelos baseline e da regressão dinâmica, foram empregados modelos não paramétricos com o objetivo de ampliar a capacidade de previsão e capturar padrões mais complexos presentes nos dados. Esses modelos não assumem uma forma funcional específica entre as variáveis e, por isso, oferecem maior flexibilidade para lidar com relações não lineares e estruturas ocultas, comuns em séries temporais de demanda.

Foram utilizados, nessa etapa, os algoritmos *k-Nearest Neighbors (KNN)*, *Support*

Vector Regression (SVR) e *Random Forest*. A escolha por esses modelos se deu tanto por sua ampla aplicação em tarefas de previsão quanto pela capacidade de adaptação a diferentes formas de variabilidade nos dados históricos. Cada modelo foi treinado utilizando a mesma divisão entre dados de treino e teste adotada nas abordagens anteriores, o que assegura uma comparação direta de desempenho entre as diferentes estratégias preditivas avaliadas ao longo do estudo.

Além de seu potencial preditivo, a utilização dos modelos não paramétricos permitiu também avaliar a sensibilidade do problema à variação de hiperparâmetros e à presença de ruídos nos dados. Os resultados obtidos com esses modelos são apresentados a seguir, com foco na previsão da demanda para o mês de dezembro e na comparação com os métodos previamente testados. Essa abordagem busca identificar se técnicas mais robustas e baseadas em aprendizado estatístico oferecem ganhos reais em precisão, justificando sua aplicação no contexto analisado.

4.4.3.1 k-Nearest Neighbors (k-NN)

Diferente dos modelos clássicos de séries temporais, a aplicação do *k-Nearest Neighbors (KNN)* requer uma definição prévia de dois elementos principais: o número de vizinhos (K) e a métrica de distância a ser utilizada. Devido à sensibilidade do modelo a esses parâmetros, inicialmente foi conduzida uma varredura, testando diferentes combinações entre valores de K (de 1 a 10) e três tipos de distâncias: Euclidiana, Manhattan e Mahalanobis. O resultado dessa avaliação está sintetizado na figura 5, que apresenta os desempenhos dos modelos com base em métricas como MAPE, RMSE, MAD e MAE.

Figura 5 - Desempenho do KNN

Distância	K	MAPE (%)	RMSE	MAD	MAE
Euclidean	1	6.83	44.50	30.13	30.13
Euclidean	2	5.75	39.51	25.57	25.57
Euclidean	3	5.36	38.53	24.01	24.01
Euclidean	4	5.65	38.96	25.20	25.20
Euclidean	5	5.57	39.16	24.89	24.89
Euclidean	6	5.68	39.28	25.34	25.34
Euclidean	7	5.86	39.28	26.09	26.09
Euclidean	8	6.03	39.72	26.76	26.76
Euclidean	9	6.20	40.59	27.51	27.51
Euclidean	10	6.40	41.26	28.35	28.35
Mahalanobis	1	9.69	61.33	42.03	42.03
Mahalanobis	2	10.44	58.68	44.89	44.89
Mahalanobis	3	11.39	60.82	48.81	48.81
Mahalanobis	4	12.36	64.51	52.79	52.79
Mahalanobis	5	13.30	68.37	56.60	56.60
Mahalanobis	6	13.66	69.49	58.10	58.10
Mahalanobis	7	14.00	70.33	59.56	59.56
Mahalanobis	8	14.09	70.40	59.94	59.94
Mahalanobis	9	14.94	74.42	63.51	63.51
Mahalanobis	10	15.48	76.11	65.72	65.72
Manhattan	1	6.18	39.04	27.17	27.17
Manhattan	2	5.21	37.24	23.29	23.29
Manhattan	3	5.38	37.60	24.00	24.00
Manhattan	4	5.55	38.53	24.77	24.77
Manhattan	5	5.57	38.91	24.89	24.89
Manhattan	6	5.42	38.03	24.25	24.25
Manhattan	7	5.43	38.17	24.27	24.27
Manhattan	8	5.37	35.84	23.89	23.89
Manhattan	9	5.70	37.14	25.29	25.29
Manhattan	10	5.90	38.26	26.15	26.15

Fonte: Autoria Própria

Após essa etapa inicial, foi realizada uma filtragem dos melhores desempenhos para cada tipo de distância, conforme mostrado na figura 6. Essa abordagem permitiu uma análise mais focada, destacando as combinações que apresentaram os menores valores de erro absoluto percentual médio (MAPE), raiz do erro quadrático médio (RMSE), desvio absoluto médio (MAD) e erro absoluto médio (MAE). Embora os desempenhos com as distâncias Euclidiana e Manhattan tenham sido bastante próximos em termos de MAPE — 5,36% e 5,37%, respectivamente — o modelo com distância Manhattan e K = 8 se destacou por apresentar, de forma simultânea, os menores valores de RMSE (35,84), MAD (23,89) e MAE (23,89), o que indica uma performance superior na estimativa da demanda.

Figura 6 - Resumo da Análise

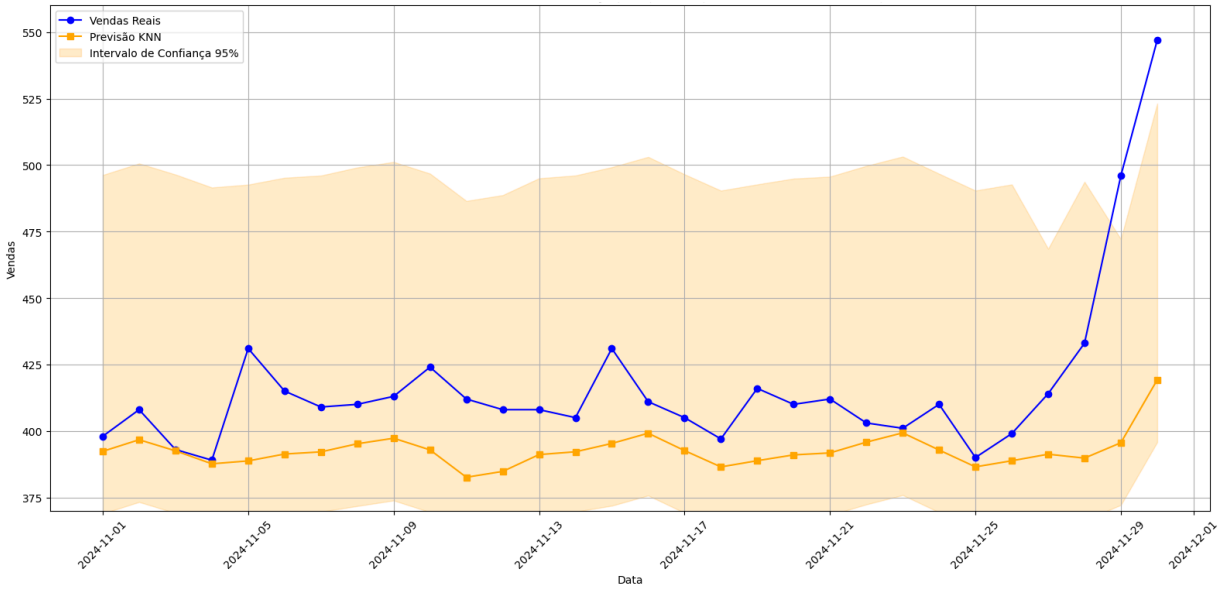
Distância	K	MAPE (%)	RMSE	MAD	MAE
Manhattan	8	5.37	35.84	23.89	23.89
Euclidean	3	5.36	38.53	24.01	24.01
Mahalanobis	2	10.44	58.68	44.89	44.89

Fonte: Autoria Própria

Essa escolha é justificada, sobretudo, pela capacidade do modelo Manhattan com K = 8 de capturar padrões locais da série temporal com maior suavidade, sem comprometer a generalização. A menor sensibilidade da métrica Manhattan a outliers, em comparação à Euclidiana, também pode ter contribuído para a melhoria do desempenho, considerando as flutuações da série de demanda analisada. Além disso, o valor de K = 8 proporcionou uma média local suficientemente boa para suavizar variações pontuais sem perder a representatividade dos dados históricos.

O desempenho do modelo selecionado é ilustrado no gráfico 24, que apresenta a curva ajustada ao longo da série temporal. Observa-se que o modelo foi capaz de acompanhar a tendência geral e adaptar-se razoavelmente às oscilações da série, especialmente no período mais recente.

Gráfico 24 - KNN (Manhattan K=8)



Fonte: Autoria Própria

4.4.3.2 Support Vector Machine (SVM)

Na sequência da aplicação do modelo KNN, foi implementado o modelo de *Support Vector Regression (SVR)*, visando explorar a capacidade desse algoritmo em capturar padrões complexos e comportamentos não lineares na série de demanda. Após a realização de uma busca em grade, foram definidos os melhores hiperparâmetros para o modelo: C = 10, epsilon = 0,5, kernel linear e gamma ajustado automaticamente via escala (Figura 7). A escolha por um kernel linear, ao final do processo de otimização, indica que o comportamento da série pôde ser adequadamente modelado por uma função linear regularizada, mesmo em presença de variações sazonais e picos pontuais.

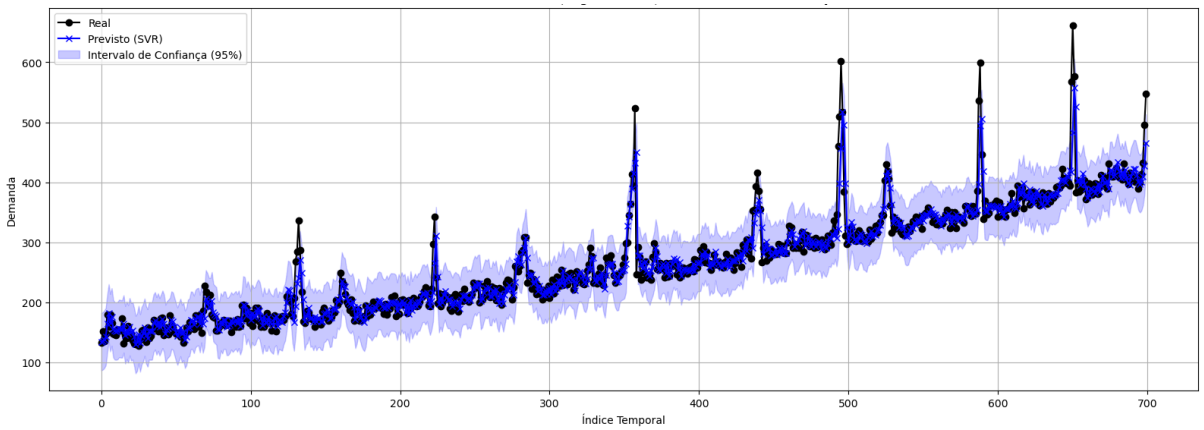
Figura 7 - Resumo da Análise

C	Epsilon	Gamma	Kernel
10	0.500000	scale	linear

Fonte: Autoria Própria

Conforme ilustrado no gráfico 25, o modelo SVR demonstrou boa capacidade de ajustar-se ao comportamento geral da série temporal ao longo de todo o período analisado. A linha de previsão segue de forma consistente a tendência crescente dos dados reais, conseguindo também capturar parcialmente os picos sazonais que ocorrem em determinados momentos da série.

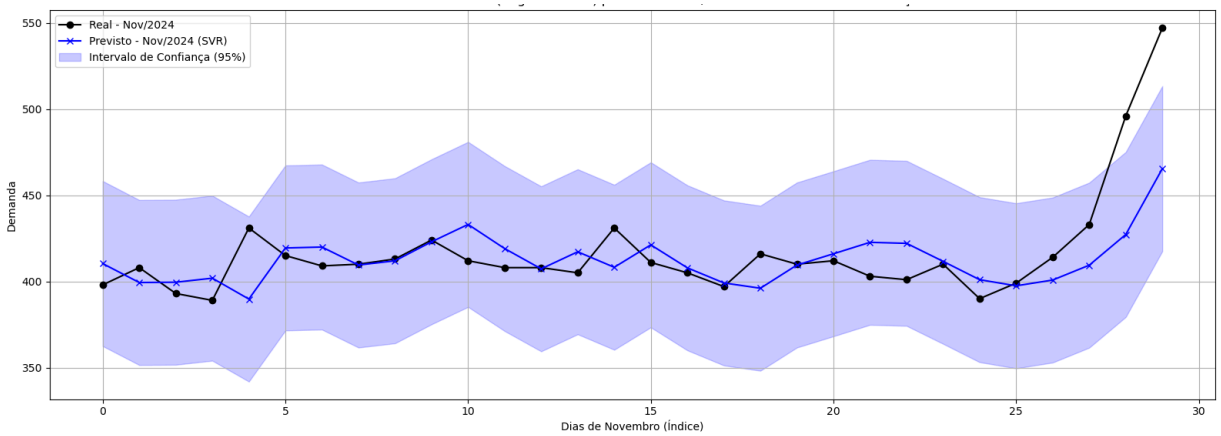
Gráfico 25 - SVM



Fonte: Autoria Própria

Já o gráfico 26, que detalha o desempenho do modelo para o mês de análise evidencia a capacidade preditiva do SVR em replicar o padrão da curva real com razoável aderência. Embora o modelo apresente uma tendência de suavização em relação aos valores observados, ele acompanha de maneira coerente as variações locais da demanda ao longo do mês. A proximidade entre os valores reais e previstos, especialmente nos dias centrais de novembro, confirma a adequação da configuração do SVR ao problema em questão.

Gráfico 26 - SVM (Últimos Meses)



Fonte: Autoria Própria

4.4.3.3 Random Forest

A etapa seguinte consistiu na aplicação do modelo *Random Forest*, com o objetivo de explorar o potencial de modelos baseados em aprendizado por conjunto na previsão da demanda. Após a realização da busca pelos hiperparâmetros ideais, definiu-se como melhor configuração: profundidade máxima das árvores igual a 10, número de estimadores (n_estimators) igual a 100, número mínimo de amostras por folha (*min_samples_leaf*) igual a

2, divisão mínima de amostras para formar um nó interno (*min_samples_split*) igual a 5, e o parâmetro *max_features* definido como a raiz quadrada do número total de variáveis (figura 8).

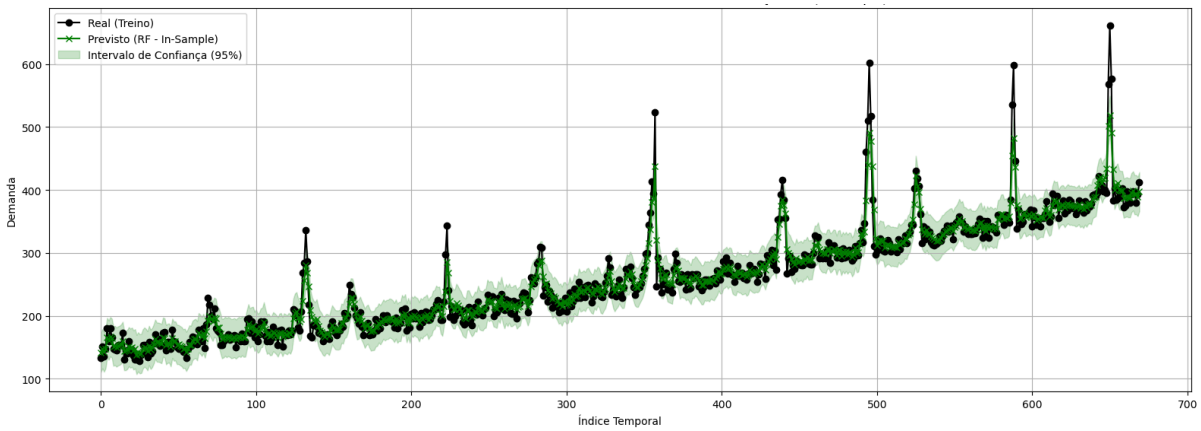
Figura 8 - Resumo da Análise

Max_depth	Max_features	Min_samples_leaf	Min_samples_split	N_estimators
10	sqrt	2	5	100

Fonte: Autoria Própria

Conforme apresentado no gráfico 27, o modelo demonstrou elevada capacidade de ajuste à série temporal ao longo de todo o histórico disponível. A curva de previsão acompanha com precisão os movimentos da série real, incluindo os picos sazonais que se repetem ciclicamente, o que evidencia a capacidade do *Random Forest* de capturar variações abruptas sem comprometer a estabilidade geral do modelo.

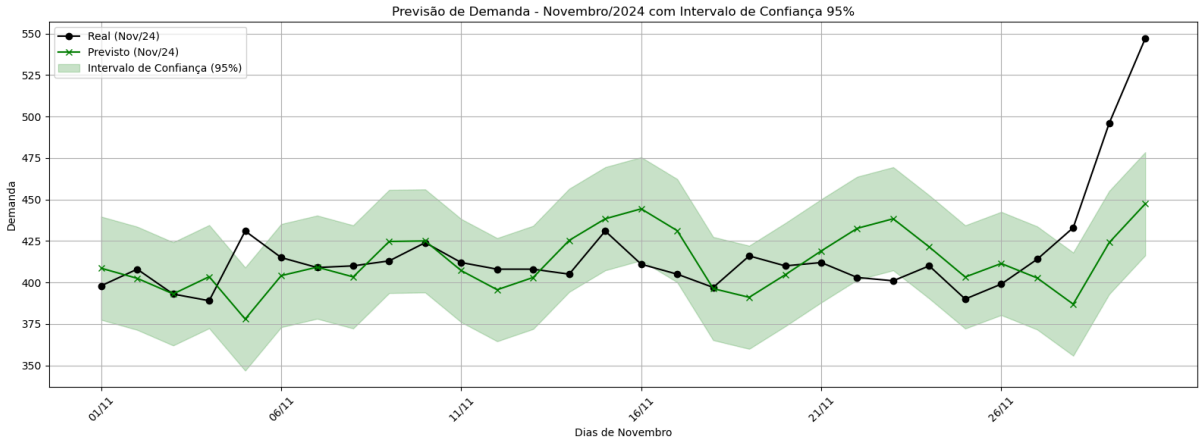
Gráfico 27 - Random Forest



Fonte: Autoria Própria

O Gráfico 28, focado especificamente no desempenho preditivo para o mês de análise, evidencia a performance do modelo na etapa de teste. Observa-se que o *Random Forest* foi capaz de seguir de maneira consistente a trajetória da série real, reproduzindo suas oscilações com razoável fidelidade.

Gráfico 28 - Random Forest (Últimos Mêses)



Fonte: Autoria Própria

4.4.4 Avaliação dos Modelos

A avaliação dos modelos preditivos foi conduzida por meio de métricas quantitativas amplamente reconhecidas na literatura: RMSE (Root Mean Squared Error), MAD (Mean Absolute Deviation), MAPE (Mean Absolute Percentage Error) e Erro Padrão. Essas métricas foram aplicadas de forma consistente em todos os modelos ao longo do estudo, com o objetivo de garantir uma comparação justa entre os diferentes métodos e identificar aquele que melhor representa o comportamento da demanda.

Inicialmente, observa-se que os modelos *baseline* apresentaram desempenho modesto, com destaque negativo para o modelo Cumulativo, que registrou o maior MAPE (29,51%), MAD (77,88) e RMSE (103,64), revelando uma limitação significativa em capturar variações ao longo da série. Os demais modelos clássicos, como Naive, Média Móvel, SES, DES e TES, apresentaram desempenhos mais consistentes, com erros relativamente baixos e próximos entre si. O modelo TES, por exemplo, destacou-se positivamente dentro desse grupo, alcançando o menor RMSE (20,32), MAD (12,77) e MAPE (6,91%), o que reforça sua capacidade de capturar padrões sazonais ao longo do tempo.

No grupo dos modelos paramétricos, a regressão aplicada diretamente ao mês de novembro apresentou MAPE de 3,74%, superando a regressão projetada exclusivamente para dezembro, que obteve 5,52%. Isso sugere que a regressão focada em um intervalo mais próximo do período de teste teve melhor aderência aos dados recentes, embora a regressão geral também tenha se mantido competitiva em relação aos modelos clássicos.

Entre os modelos não paramétricos, o *Random Forest* demonstrou o melhor

desempenho global. Com MAPE de 3,30%, RMSE de 15,86 e MAD de apenas 8,81, esse modelo apresentou os menores valores de erro em todas as métricas avaliadas. Seu erro padrão (13,19) também foi o mais baixo entre todos os modelos, indicando alta estabilidade nas previsões. O SVR obteve resultados consistentes, com MAPE de 5,09% e bom equilíbrio entre precisão e suavização.

De forma geral, os resultados apontam que os modelos baseados em aprendizado de máquina, especialmente o *Random Forest*, ofereceram ganhos expressivos em acurácia preditiva em comparação aos métodos tradicionais. Essa constatação reforça a importância da adoção de abordagens mais flexíveis em contextos de previsão de demanda.

Figura 9 - Análise das Métricas

	MAPE (%)	RMSE	MAD	Erro Padrão
Naive	7.87	24.30	14.63	24.29
Cumulativo	29.51	103.64	77.88	70.32
Média Móvel	8.06	32.24	17.42	31.85
SES	7.61	24.18	14.26	24.17
DES	7.64	24.17	14.28	24.17
TES	6.91	20.32	12.77	20.32
Regressão para Dezembro	5.52	22.73	14.60	17.43
Regressão para Novembro	3.74	24.88	16.50	17.43
KNN	5.37	35.84	23.89	27.18
SVR	5.09	24.43	13.59	20.30
Random Forest	3.30	15.86	8.81	13.19

Fonte: Autoria Própria

5. CONCLUSÃO

Este trabalho buscou aplicar conceitos e técnicas de análise preditiva com o intuito de estimar o volume diário de vendas de camisetas básicas da empresa *Segrob Notlad* ao longo do mês de dezembro de 2024. Inserido em um setor marcado pela alta volatilidade e rápida renovação de tendências, como o fast fashion, o desafio de previsão de demanda se mostrou particularmente relevante, exigindo uma abordagem estruturada e adaptável às especificidades do negócio. Para isso, foi utilizada a metodologia CRISP-DM, que guiou todas as etapas do projeto, desde o entendimento do problema até a avaliação dos modelos preditivos.

Na fase inicial, foi realizado o entendimento do negócio, que permitiu contextualizar a importância da previsão de demanda como ferramenta estratégica para otimização dos estoques, redução de perdas logísticas e melhor aproveitamento de datas promocionais. A empresa analisada, com forte atuação no mercado brasileiro e presença internacional emergente, apresenta um modelo de negócios que exige decisões ágeis e baseadas em dados confiáveis. Isso reforça o papel da previsão como instrumento central para a competitividade em mercados dinâmicos e altamente influenciados por sazonalidade e comportamento do consumidor.

A seguir, foi conduzida uma análise exploratória dos dados históricos disponibilizados. A partir dela, foram identificadas tendências de crescimento, padrões sazonais recorrentes e picos significativos de vendas em datas comemorativas como o Dia das Mães, Dia dos Pais, Dia das Crianças e o período pós-Black Friday. Esses insights embasaram a etapa de preparação dos dados, que incluiu a criação de variáveis representativas para esses eventos, bem como a incorporação de variáveis temporais como dias da semana, meses e anos, visando enriquecer o poder explicativo dos modelos.

Na modelagem, foram testadas diferentes abordagens, abrangendo tanto modelos paramétricos (como Naive, Média Móvel, Suavizações Exponenciais e Regressão Dinâmica) quanto não paramétricos (como *k-Nearest Neighbors*, *Random Forest* e *Support Vector Machine*). Os modelos foram avaliados com base em métricas como MAPE, RMSE e MAD, possibilitando uma comparação consistente do desempenho preditivo de cada um. Modelos mais sofisticados, como o Holt-Winters e modelos não paramétricos, apresentaram desempenho superior, sobretudo na captura de sazonalidades e variações complexas, embora exigissem maior esforço computacional e cuidado na escolha dos hiperparâmetros.

Ao longo do trabalho, ficou evidente que a escolha do modelo mais adequado depende não apenas da acurácia obtida, mas também da capacidade de interpretação, escalabilidade e

aderência ao contexto operacional da empresa. Modelos mais simples, embora menos precisos, podem ser úteis em situações com limitação de dados ou quando se busca rápida implementação. Por outro lado, modelos mais complexos oferecem maior desempenho preditivo, desde que bem ajustados e devidamente validados.

Dessa forma, este trabalho não apenas cumpriu seu objetivo de desenvolver modelos preditivos aplicáveis ao contexto do varejo de moda, mas também contribuiu para a formação analítica dos autores, promovendo o domínio de ferramentas estatísticas, técnicas de *machine learning* e metodologias estruturadas de ciência de dados. Espera-se que os conhecimentos adquiridos e os resultados obtidos possam ser utilizados como base para aplicações reais na empresa estudada, bem como para futuros estudos e projetos no campo da análise preditiva.

Bibliografia

ADELEKE, Oluwatobi *et al.* Application of artificial neural networks for predicting the physical composition of municipal solid waste: An assessment of the impact of seasonal variation. **Waste Management & Research: The Journal for a Sustainable Circular Economy**, v. 39, n. 8, p. 1058–1068, ago. 2021.

ALDUAILIJ, Mona A. *et al.* Forecasting peak energy demand for smart buildings. **The Journal of Supercomputing**, v. 77, n. 6, p. 6356–6380, jun. 2021.

ALLGAIER, Johannes; PRYSS, Rüdiger. Cross-Validation Visualized: A Narrative Guide to Advanced Methods. **Machine Learning and Knowledge Extraction**, v. 6, n. 2, p. 1378–1388, 20 jun. 2024.

BHARDWAJ, N. *et al.* NEURAL NETWORK AUTOREGRESSION AND CLASSICAL TIME SERIES APPROACHES FOR RICE YIELD FORECASTING. **The Journal of Animal and Plant Sciences**, v. 31, n. 4, p. 1126–1131, 30 jan. 2021.

BRILLINGER, Markus *et al.* Energy prediction for CNC machining with machine learning. **CIRP Journal of Manufacturing Science and Technology**, v. 35, p. 715–723, nov. 2021.

BRISIMI, Theodora S. *et al.* Federated learning of predictive models from federated Electronic Health Records. **International Journal of Medical Informatics**, v. 112, p. 59–67, abr. 2018.

CERQUEIRA, Vitor; TORGO, Luis; MOZETIČ, Igor. Evaluating time series forecasting models: an empirical study on performance estimation methods. **Machine Learning**, v. 109, n. 11, p. 1997–2028, nov. 2020.

CHEN, Xinqiang *et al.* Augmented Ship Tracking Under Occlusion Conditions From Maritime Surveillance Videos. **IEEE Access**, v. 8, p. 42884–42897, 2020.

CHEN, Yitian *et al.* Probabilistic forecasting with temporal convolutional neural network. **Neurocomputing**, v. 399, p. 491–501, jul. 2020.

CHUM, Antony *et al.* Changes in Public Response Associated With Various COVID-19 Restrictions in Ontario, Canada: Observational Infoveillance Study Using Social Media Time Series Data. **Journal of Medical Internet Research**, v. 23, n. 8, p. e28716, 25 ago. 2021.

ELGELDAWI, Enas *et al.* Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis. **Informatics**, v. 8, n. 4, p. 79, 17 nov. 2021.

ENSAFI, Yasaman *et al.* Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. **International Journal of Information Management Data Insights**, v. 2, n. 1, p. 100058, abr. 2022.

FONSECA, Larissa Rodrigues Da; PEDROSA, Maria Eduarda Correia; CARDOSO, Renata De Lima Andrade. ANÁLISE DO SISTEMA PRODUTIVO EM UMA HAMBURGUERIA ARTESANAL: UM ESTUDO DE CASO EM NATAL/RN. *In*: ENEGEP 2024 - ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO. **Anais**... PORTO ALEGRE/RS - BRASIL: 7 nov. 2024.

GEETHA, A. *et al.* Prediction of hourly solar radiation in Tamil Nadu using ANN model with different learning algorithms. **Energy Reports**, v. 8, p. 664–671, abr. 2022.

GIRI, Chandadevi; CHEN, Yan. Deep Learning for Demand Forecasting in the Fashion and Apparel Retail Industry. **Forecasting**, v. 4, n. 2, p. 565–581, 20 jun. 2022.

GUNTER, Ulrich. Improving Hotel Room Demand Forecasts for Vienna across Hotel Classes and Forecast Horizons: Single Models and Combination Techniques Based on Encompassing Tests. **Forecasting**, v. 3, n. 4, p. 884–919, 27 nov. 2021.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. **Forecasting: principles and practice**. Third print edition ed. Melbourne, Australia: Otexts, Online Open-Access Textbooks, 2021.

KHALEDIAN, Yones; MILLER, Bradley A. Selecting appropriate machine learning methods for digital soil mapping. **Applied Mathematical Modelling**, v. 81, p. 401–418, maio 2020.

KIM, Yunsun; KIM, Sahm. Forecasting Charging Demand of Electric Vehicles Using Time-Series Models. **Energies**, v. 14, n. 5, p. 1487, 9 mar. 2021.

KRISTON, Levente. **Predictive Accuracy of a Hierarchical Logistic Model of Cumulative SARS-CoV-2 Case Growth**. , 16 jun. 2020.

LIAW, Jiun-Jian *et al.* PM2.5 Concentration Estimation Based on Image Processing Schemes and Simple Linear Regression. **Sensors**, v. 20, n. 8, p. 2423, 24 abr. 2020.

LIN, Huasheng *et al.* An investigation of machine learning techniques to estimate minimum horizontal stress magnitude from borehole breakout. **International Journal of Mining Science and Technology**, v. 32, n. 5, p. 1021–1029, set. 2022.

LUO, Tian; CHANG, Daofang; XU, Zhenyu. Research on Apparel Retail Sales Forecasting Based on xDeepFM-LSTM Combined Forecasting Model. **Information**, v. 13, n. 10, p. 497, 15 out. 2022.

MAIMAITIJIANG, Maitiniyazi *et al.* Soybean yield prediction from UAV using multimodal data fusion and deep learning. **Remote Sensing of Environment**, v. 237, p. 111599, fev. 2020.

MAKRIDAKIS, Spyros; SPILIOTIS, Evangelos; ASSIMAKOPOULOS, Vassilios. Statistical and Machine Learning forecasting methods: Concerns and ways forward. **PLOS ONE**, v. 13, n. 3, p. e0194889, 27 mar. 2018.

MALEKI, Farhad *et al.* Machine Learning Algorithm Validation. **Neuroimaging Clinics of North America**, v. 30, n. 4, p. 433–445, nov. 2020.

MARKOVICS, Dávid; MAYER, Martin János. Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. **Renewable and Sustainable Energy Reviews**, v. 161, p. 112364, jun. 2022.

MOHD JAIS, Nurshahida Azreen *et al.* Improved accuracy in IoT-Based water quality monitoring for aquaculture tanks using low-cost sensors: Asian seabass fish farming. **Heliyon**, v. 10, n. 8, p. e29022, abr. 2024.

MUNIM, Ziaul Haque *et al.* Forecasting container throughput of major Asian ports using the Prophet and hybrid time series models. **The Asian Journal of Shipping and Logistics**, v. 39, n. 2, p. 67–77, jun. 2023.

MUZALYOVA, Anna *et al.* Forecasting Betula and Poaceae airborne pollen concentrations on a 3-hourly resolution in Augsburg, Germany: toward automatically generated, real-time predictions. **Aerobiologia**, v. 37, n. 3, p. 425–446, set. 2021.

NABIPOUR, Mojtaba *et al.* Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis. **IEEE Access**, v. 8, p. 150199–150212, 2020.

NASSER, Ahmed Abdel; RASHAD, Magdi Z.; HUSSEIN, Sherif E. A Two-Layer Water Demand Prediction System in Urban Areas Based on Micro-Services and LSTM Neural Networks. **IEEE Access**, v. 8, p. 147647–147661, 2020.

NGUYEN, Hoang *et al.* Efficient machine learning models for prediction of concrete strengths. **Construction and Building Materials**, v. 266, p. 120950, jan. 2021.

OLIVEIRA, H. B. de J. *O mercado das empresas fast fashion: um estudo de caso da cadeia de suprimentos da H&M e Zara*. **SIMPÓSIO DE ENGENHARIA DE PRODUÇÃO DE SERGIPE**, v; 9, 2017, São Cristóvão.

PÉREZ-LÓPEZ, Artemio *et al.* Postharvest respiration of fruits and environmental factors interaction: An approach by dynamic regression models. **Scientia Agropecuaria**, v. 11, n. 1, p. 23–29, 31 mar. 2020.

RAMOS, Jorge Luis Cavalcanti *et al.* CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. *In*: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. **Anais do XXXI Simpósio Brasileiro de Informática na Educação (SBIE 2020)**. Brasil: Sociedade Brasileira de Computação, 24 nov. 2020.

REDDY, G. Thippa *et al.* Analysis of Dimensionality Reduction Techniques on Big Data. **IEEE Access**, v. 8, p. 54776–54788, 2020.

SCHONLAU, Matthias; ZOU, Rosie Yuyan. The random forest algorithm for statistical learning. **The Stata Journal: Promoting communications on statistics and Stata**, v. 20, n. 1, p. 3–29, mar. 2020.

SCHRÖER, Christoph; KRUSE, Felix; GÓMEZ, Jorge Marx. A Systematic Literature Review on Applying CRISP-DM Process Model. **Procedia Computer Science**, v. 181, p. 526–534, 2021.

SEYEDAN, Mahya; MAFAKHERI, Fereshteh. Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. **Journal of Big Data**, v. 7, n. 1, p. 53, dez. 2020.

SIDEY-GIBBONS, Jenni A. M.; SIDEY-GIBBONS, Chris J. Machine learning in medicine: a practical introduction. **BMC Medical Research Methodology**, v. 19, n. 1, p. 64, dez. 2019.

SOUSA, M. S.; LOUREIRO, A. L. D.; MIGUÉIS, V. L. Predicting demand for new products in fashion retailing using censored data. Expert Systems with Applications, v. 259, p. 125313, jan. 2025.

TOFT, Håvard B. *et al.* Can big data and random forests improve avalanche runout estimation compared to simple linear regression? **Cold Regions Science and Technology**, v. 211, p. 103844, jul. 2023.

TRULL, Óscar; GARCÍA-DÍAZ, J. Carlos; TRONCOSO, Alicia. Stability of Multiple Seasonal Holt-Winters Models Applied to Hourly Electricity Demand in Spain. **Applied Sciences**, v. 10, n. 7, p. 2630, 10 abr. 2020.

UDDIN, Shahadat *et al.* Comparing different supervised machine learning algorithms for disease prediction. **BMC Medical Informatics and Decision Making**, v. 19, n. 1, p. 281,

dez. 2019.

UDDIN, Shahadat *et al.* Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. **Scientific Reports**, v. 12, n. 1, p. 6256, 15 abr. 2022.

VAN STEENBERGEN, R. M.; MES, M. R. K. Forecasting demand profiles of new products. **Decision Support Systems**, v. 139, p. 113401, dez. 2020.

WALKER, Shalika *et al.* Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings. **Energy and Buildings**, v. 209, p. 109705, fev. 2020.

YAĞCI, Mustafa. Educational data mining: prediction of students' academic performance using machine learning algorithms. **Smart Learning Environments**, v. 9, n. 1, p. 11, dez. 2022.

YANG, Li; SHAMI, Abdallah. On hyperparameter optimization of machine learning algorithms: Theory and practice. **Neurocomputing**, v. 415, p. 295–316, nov. 2020.

YU, Thomas *et al.* Model-informed machine learning for multi-component T 2 relaxometry. **Medical Image Analysis**, v. 69, p. 101940, abr. 2021.

ZHANG, Wengang *et al.* Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. **Geoscience Frontiers**, v. 12, n. 1, p. 469–477, jan. 2021.