**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 utilizadas as métricas RMSE (Raiz do Erro Médio Quadrático), MAD (Desvio Médio Absoluto) e MAPE (Erro Percentual Médio Absoluto). 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.

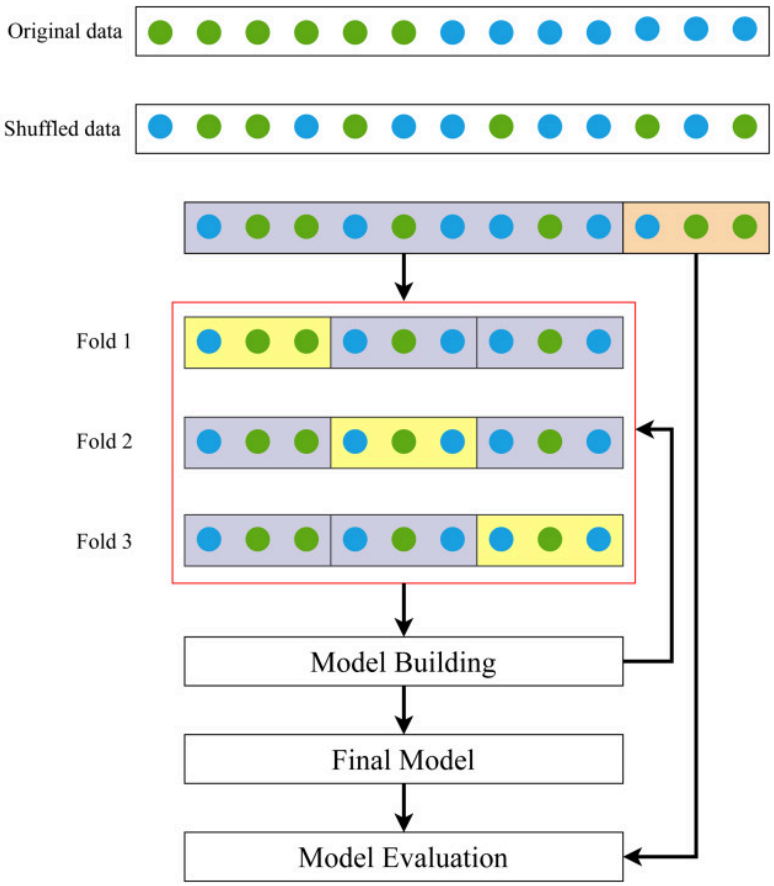
#### 2.3.1.1 Validação Cruzada K-Fold (KFCV)

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

### 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).

### 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).

### 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).

### 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 $\alpha$ de suavização que varia entre 0 e 1. Valores de $\alpha$ 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).

### 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 $\alpha$ e $\beta$ 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).

### 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 ($\alpha$, $\beta$ e $\gamma$). 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 desse trabalho (Trull; García-Díaz;

Troncoso, 2020).

## 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).

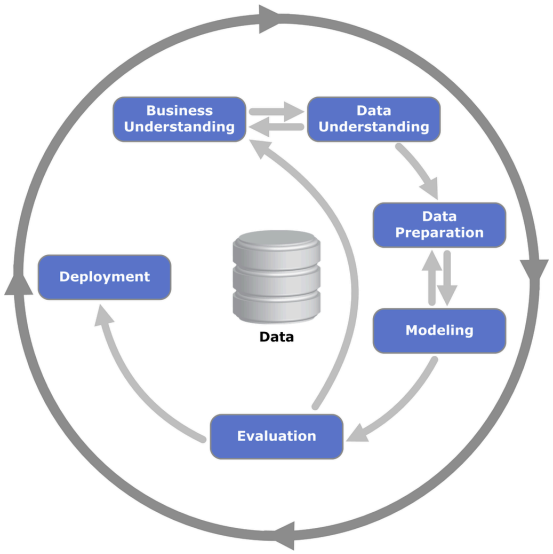
5. **Avaliação (Evaluation)**

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 1 -** Processo CRISP-DM




**Fonte:** Shearer, 2000

# 4. ESTUDO DE CASO

## 4.1 ENTENDIMENTO DO NEGÓCIO

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). Diante desse cenário, a construção de um modelo preditivo robusto torna-se essencial, exigindo o uso de métricas de performance reconhecidas na literatura. Serão empregadas, neste projeto, 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.

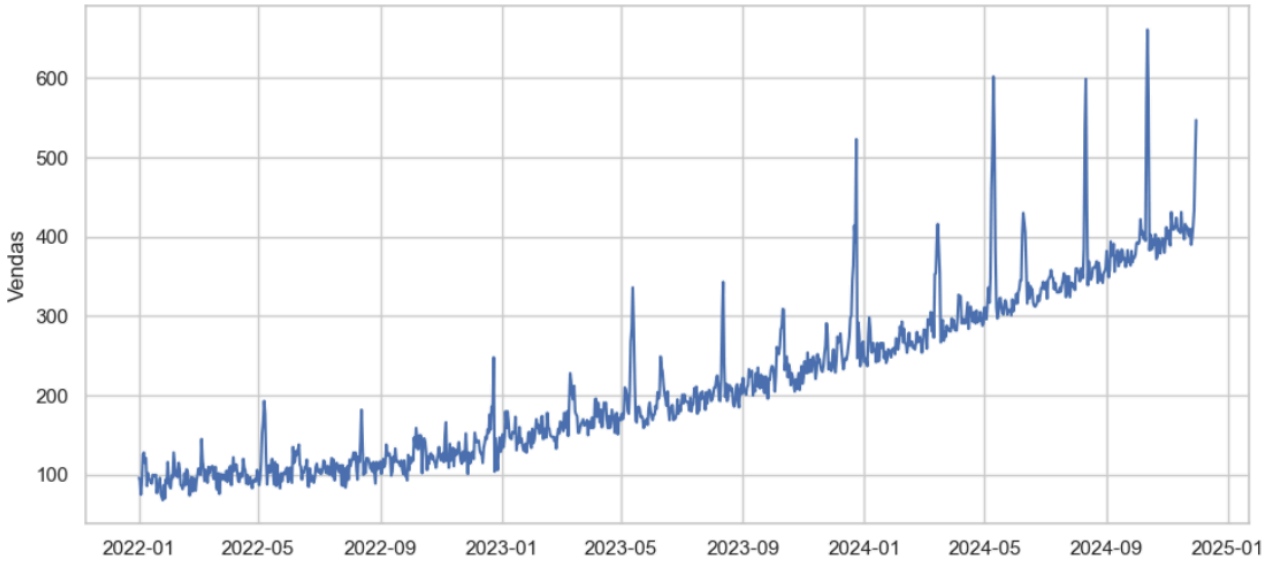
Contudo, uma das restrições mais relevantes do projeto diz respeito à limitação do conjunto de dados disponível, que, inicialmente, contempla apenas as variáveis de data e quantidade vendida de camisetas básicas. A ausência de atributos complementares, como indicadores promocionais, dados meteorológicos, 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 carência de informações pode representar um desafio à qualidade das previsões, exigindo, portanto, abordagens cuidadosas e, se possível, a incorporação futura de variáveis adicionais que enriquecem a análise.

## 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*. Inicialmente, foi realizada uma análise exploratória com o objetivo de compreender melhor os dados e embasar uma previsão de demanda mais assertiva. Todos os dados foram considerados relevantes nessa etapa inicial, e constatou-se que não há valores ausentes ou zerados nas colunas referentes ao período de 2022 a 2024. Para facilitar a visualização e identificação de possíveis tendências, foram elaborados gráficos que oferecem uma visão mais clara dos padrões presentes nos dados.

No gráfico 1, a série temporal de vendas de camisetas básicas masculinas revela uma tendência clara de crescimento na demanda ao longo do tempo. Os dados mostram um comportamento consistente, com média de 214 vendas e picos recorrentes que sugerem a influência de sazonalidades ou ações promocionais. Esses picos se intensificam nos anos mais recentes, indicando uma possível ampliação da base de clientes ou maior eficiência nas estratégias comerciais. A presença de tendência e padrões sazonais colabora para ela ser aplicável para modelagem preditiva.

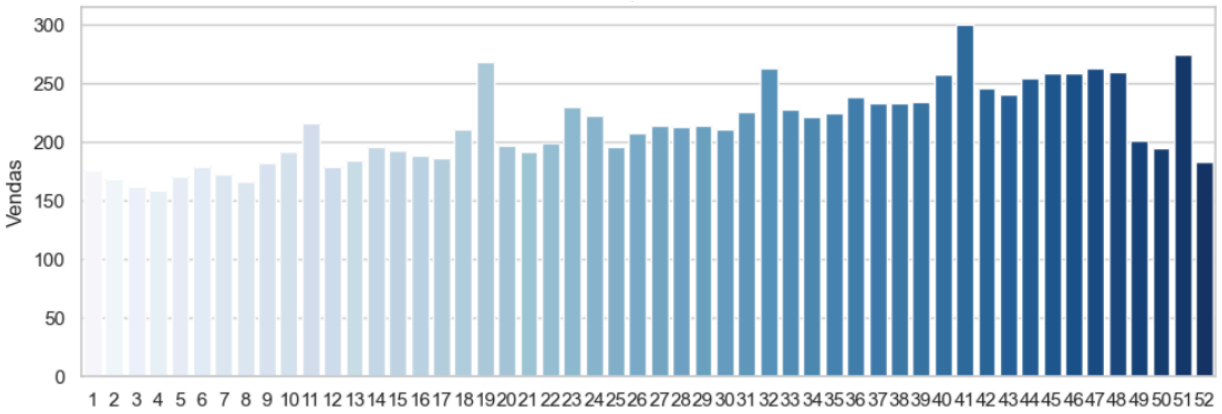
**Gráfico 1 -** Dados Série Temporal




**Fonte:** Autoria Própria

No gráfico 2, a análise da média de vendas por semana do ano revela uma distribuição sazonal bem definida ao longo das 52 semanas. Observa-se um aumento gradual nas vendas a partir da semana 22, com picos mais acentuados nas semanas 32 e 51. A elevação consistente no volume médio semanal ao longo do ano reforça a tendência de crescimento da demanda. Algo importante é se observar que o ano de 2024 não teve dados de dezembro o que justifica os números mais baixos nas últimas semanas do gráfico.

**Gráfico 2 -** Vendas por Semana

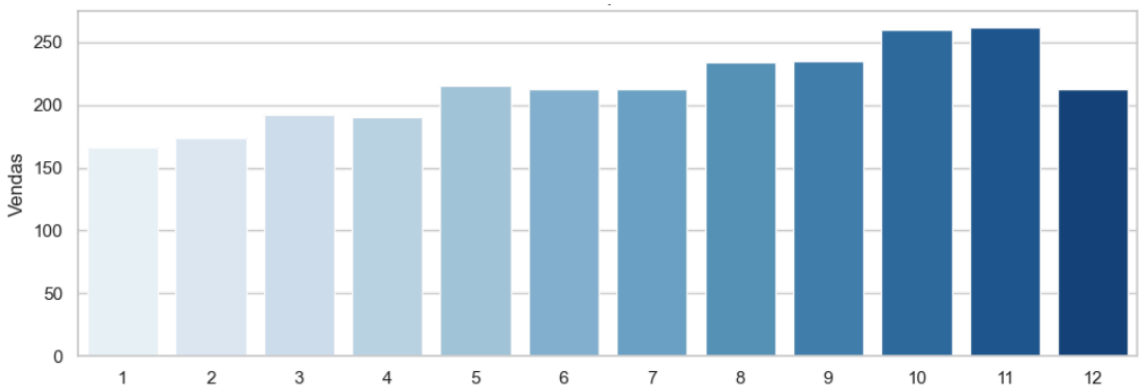



**Fonte:** Autoria Própria

No gráfico 3, é revelado uma tendência geral de crescimento ao longo do ano, com destaque para o aumento consistente a partir do mês 5 e atingindo picos nos meses 10 e 11, onde as vendas ultrapassam 250 unidades. Nos primeiros quatro meses, os valores

permanecem abaixo de 200 unidades, sugerindo um início de ano mais modesto. No entanto, em dezembro (mês 12), há uma queda perceptível nas vendas em comparação aos picos anteriores, o que está relacionado ao fato do último ano não ter dados desse mês o que gerou uma queda desse número.

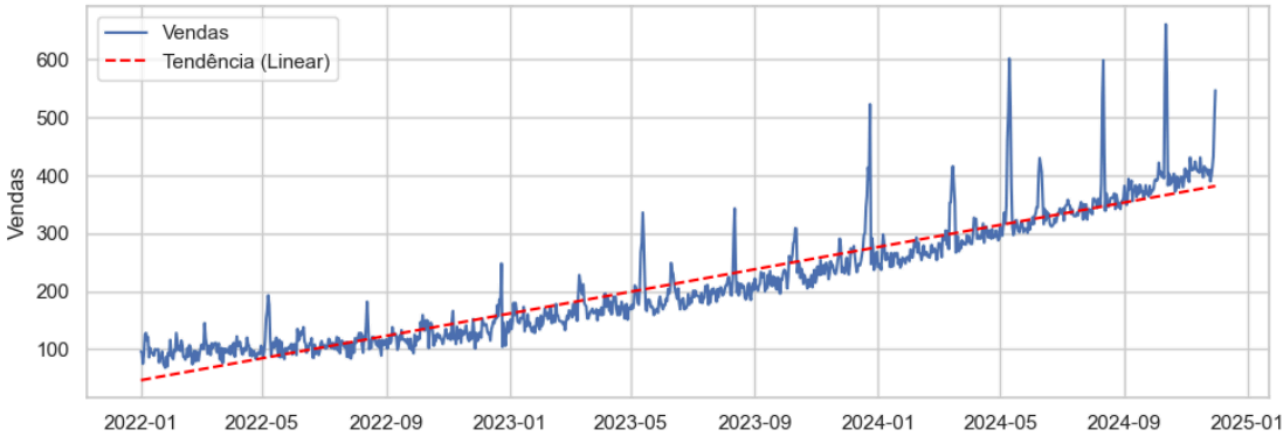
**Gráfico 3 -** Vendas por Mês




**Fonte:** Autoria Própria

No gráfico 4, a análise da tendência das vendas ao longo do período de 2022 a 2024 demonstra um crescimento consistente, evidenciado tanto pela linha azul, que representa as vendas diárias, quanto pela linha de tendência linear (vermelha tracejada), que confirma o aumento gradual. A tendência ascendente indica que o negócio está em expansão, o que reforça a importância de planejar capacidade operacional, estoque e recursos.

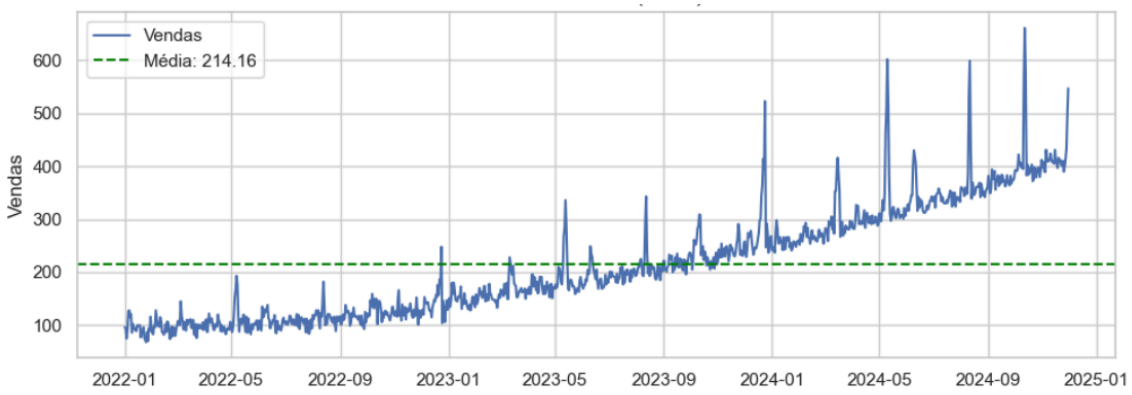
**Gráfico 4 -** Vendas por Semana (Tendência)




**Fonte:** Autoria Própria

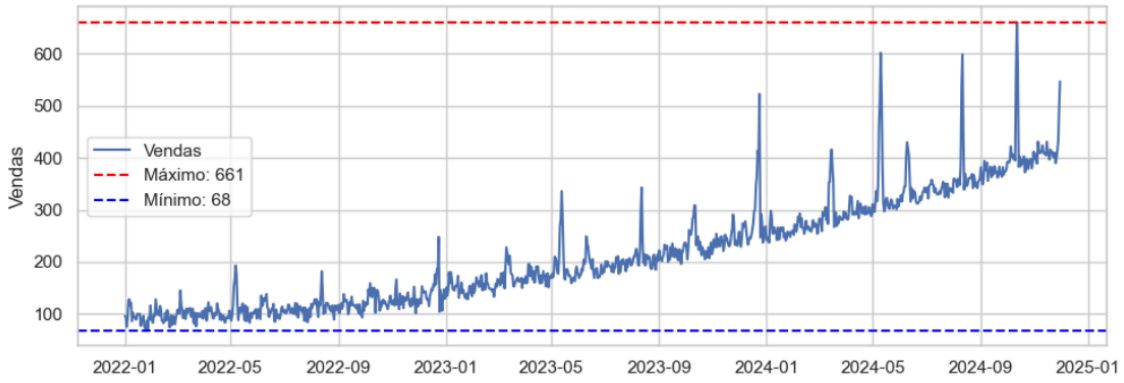
Posteriormente foram gerados gráficos e uma tabela para obter informações descritivas dos dados. A média das vendas ficou em 214,16 unidades, enquanto o desvio padrão foi de 103,57, indicando uma variação considerável em torno da média e reforçando a presença de oscilações no volume de vendas. O valor máximo registrado foi de 661 unidades, enquanto o valor mínimo observado foi de 68 unidades.

**Gráfico 5 -** Média de Vendas




**Fonte:** Autoria Própria

**Gráfico 6 -** Máximo e Mínimo de Vendas




**Fonte:** Autoria Própria

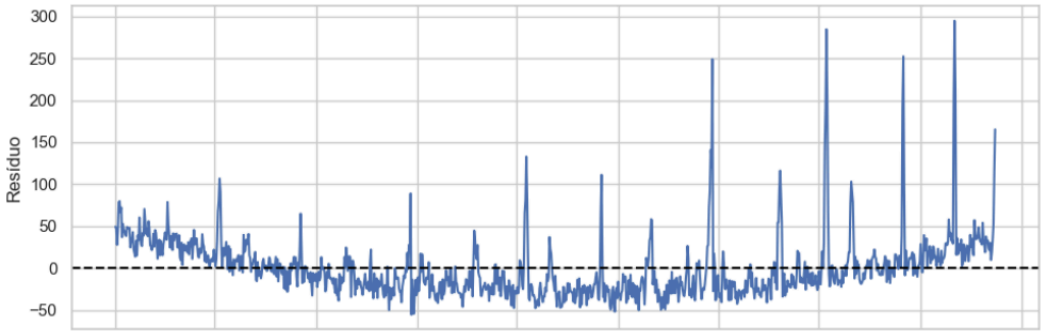
**Imagem 2 -** Informações Descritivas

```
Média das vendas: 214.16
Desvio padrão das vendas: 103.57
Valor máximo: 661.00
Valor mínimo: 68.00
```

**Fonte:** Autoria Própria

O gráfico de resíduos apresentado mostra a variação das vendas em relação à tendência linear esperada, permitindo 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, como indicado pela linha preta tracejada, confirmando que, em média, os erros não apresentam viés. Contudo, destacam-se picos esporádicos positivos e negativos, que representam eventos atípicos. Esse comportamento é importante para avaliar a qualidade do modelo, já que a predominância de resíduos próximos de zero sugere um bom ajuste, mas os picos apontam oportunidades de refinamento, seja incorporando variáveis explicativas adicionais ou uma outra alternativa.

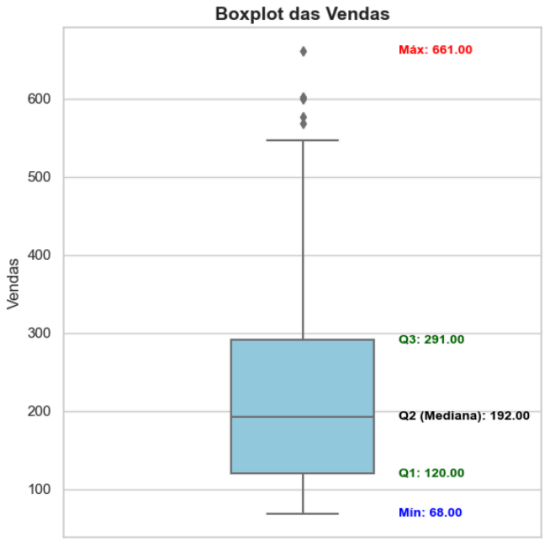
**Gráfico 7 -** Ruído Branco




**Fonte:** Autoria Própria

O boxplot das vendas abaixo permite visualizar a distribuição dos dados e identificar a presença de outliers, além de fornecer informações sobre os quartis e a dispersão das vendas ao longo do período analisado. O primeiro quartil (Q1) foi de 120,00, enquanto a mediana (Q2) foi de 192,00 e o terceiro quartil (Q3) foi 291. O valor mínimo registrado foi de 68,00 e o máximo, de 661,00, evidenciando uma ampla variação nos dados. Observa-se também a presença de outliers acima do limite superior, representando picos de vendas fora do padrão esperado.

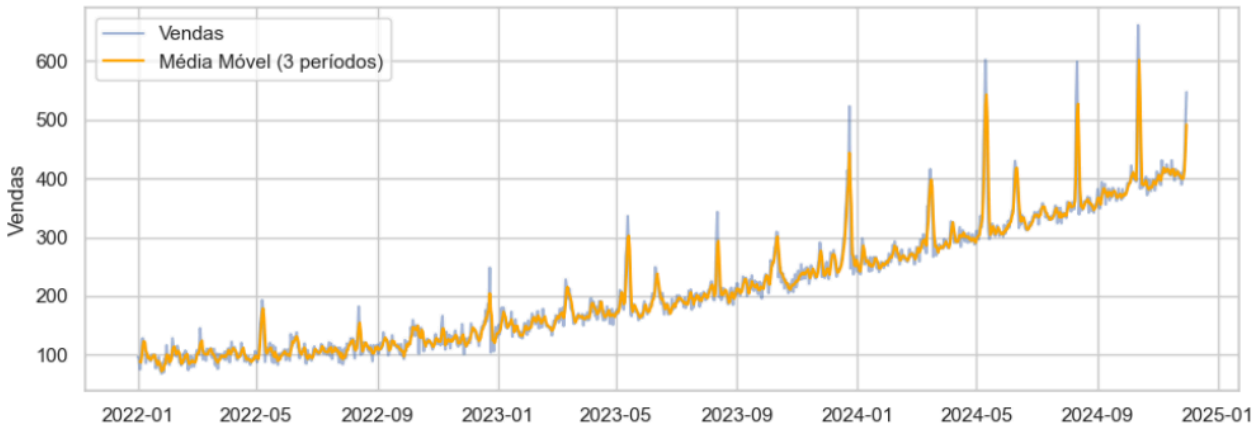
**Gráfico 8** - Boxplot




**Fonte:** Autoria Própria

Por fim, a imagem abaixo apresenta a média móvel das vendas ao longo do tempo, suavizando as variações diárias e permitindo identificar tendências gerais. A linha laranja, correspondente à média móvel de 3 períodos, acompanha de perto os dados reais, destacando os padrões de elevação gradual e suavizando os ruídos. Esse tipo de análise é útil para prever comportamentos futuros.

**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

Considerando a etapa anterior, serão geradas variáveis representativas para as semanas em que ocorrem datas comemorativas de reconhecida relevância comercial: Dia dos Pais, Dia das Mães, Dia das Crianças e Black Friday.

Essas datas são tradicionalmente associadas a campanhas promocionais intensas e a uma maior propensão do consumidor à compra, o que justifica sua inclusão como potenciais variáveis explicativas. Por exemplo, o Dia das Mães e o Dia dos Pais figuram entre as principais datas do varejo nacional, estimulando a compra de presentes.

Adicionalmente, serão incluídas variáveis correspondentes à semana do Natal e à semana do Dia dos Namorados, com o objetivo de investigar, por meio da análise estatística, a relevância dessas datas para o modelo. Apesar de não estarem inicialmente entre as datas selecionadas com base nos resultados preliminares, sua influência potencial sobre o comportamento do consumidor justifica a inclusão para fins exploratórios.

A análise da importância dessas variáveis será conduzida por meio do teste estatístico ANOVA (Análise de Variância), utilizando os coeficientes estimados na tabela de regressão. Esse teste permitirá verificar a significância estatística de cada variável correspondente às semanas comemorativas, avaliando se há diferenças substanciais no comportamento da variável dependente durante esses períodos.

Essas variáveis serão adicionadas posteriormente para aprimorar o modelo de previsão. Inicialmente, serão elaborados os modelos de suavização simples, dupla e tripla, além do naive, cumulativo e de média móvel.

## 4.4 MODELAGEM

Na fase de modelagem, foram primordialmente 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.

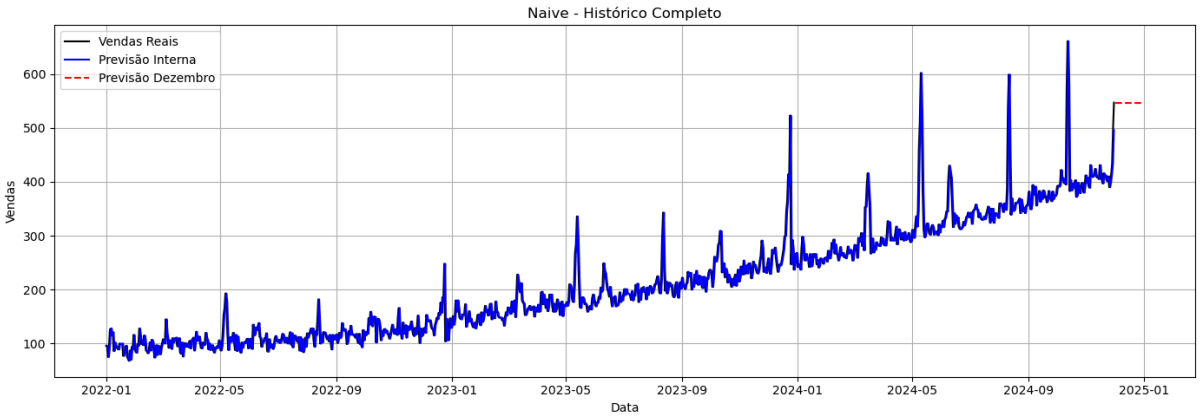
Para as observações históricas, os dados de treino abrangem o período de janeiro de 2022 até outubro de 2024. Já os dados de teste compreendem o mês de novembro de 2024. A justificativa para esta escolha de 30 dias como intervalo de teste está alinhada ao objetivo final do modelo preditivo desenvolvido 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 de novembro 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 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

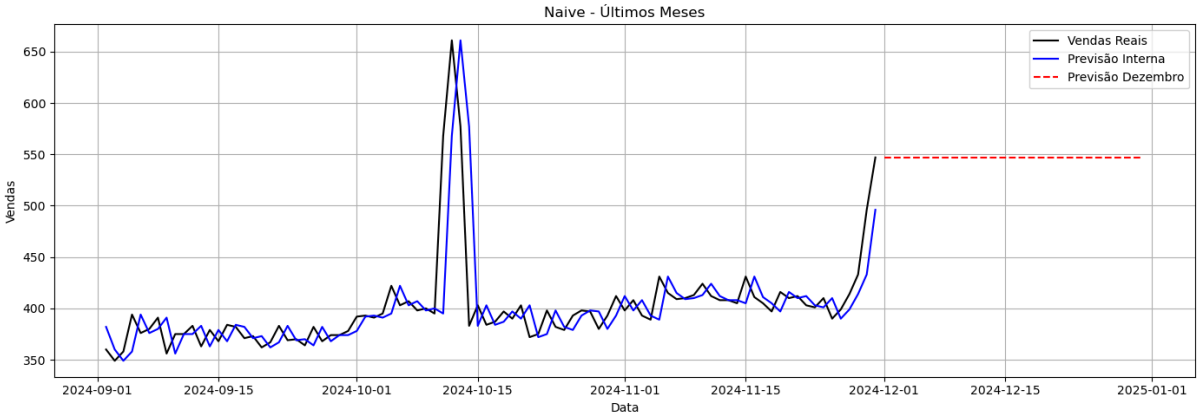
mínima de desempenho. Sua principal vantagem é a rapidez e facilidade de implementação, sendo especialmente útil quando a série apresenta pouca variação ou quando se deseja estabelecer um comparativo inicial.

**Gráfico 10 - Método Naive**




**Fonte:** Autoria Própria

**Gráfico 11 - Método Naive - Últimos meses**

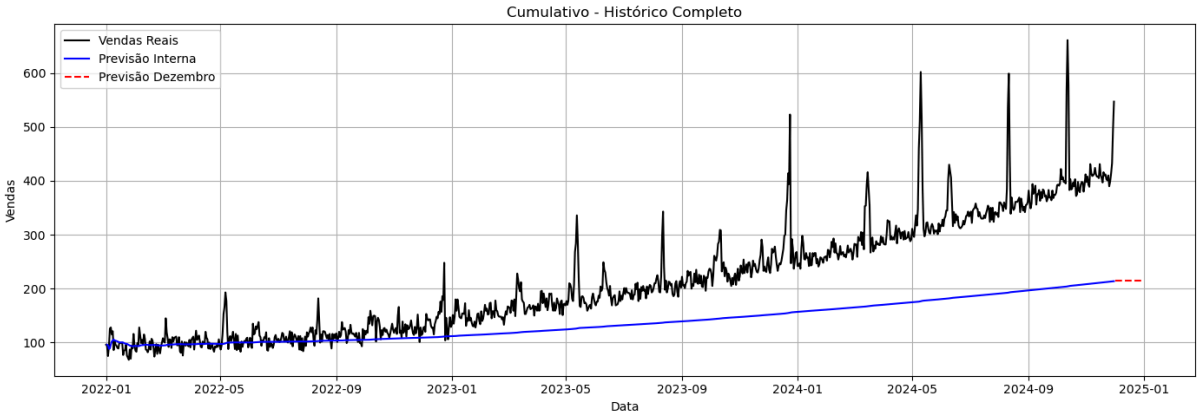



**Fonte:** Autoria Própria

### 4.4.2 Média Cumulativa

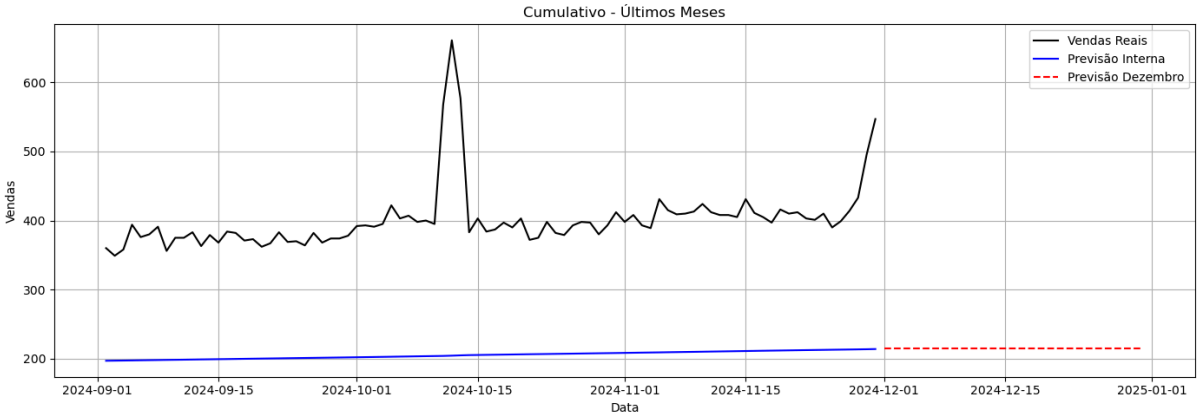
Nesse método, a previsão é baseada na média aritmética de todos os valores observados até o momento da previsão. Ele suaviza ruídos ao longo do tempo, porém, não reage bem a mudanças repentinas na tendência ou na sazonalidade. É útil em séries relativamente estáveis.

**Gráfico 12 - Média Cumulativa**




**Fonte:** Autoria Própria

**Gráfico 13 - Média Cumulativa - Últimos meses**

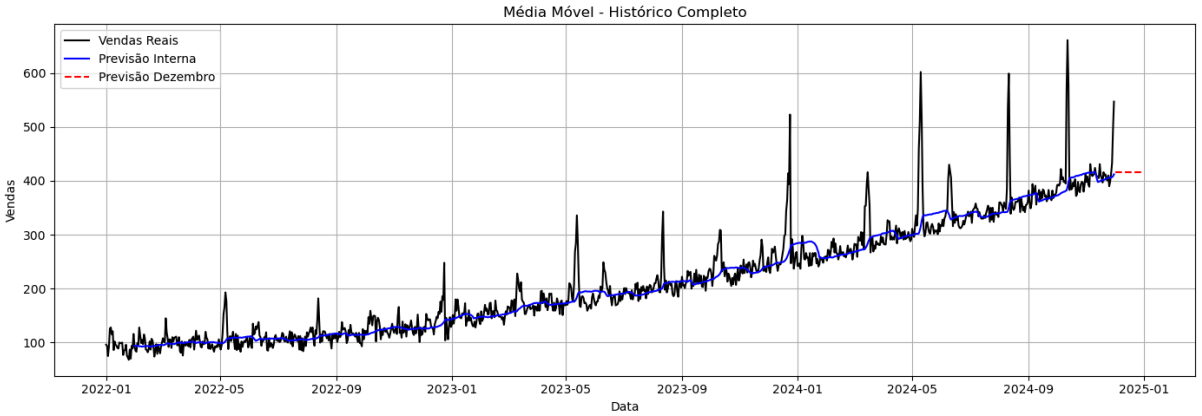



**Fonte:** Autoria Própria

### 4.4.3 Média Móvel Simples (30 dias)

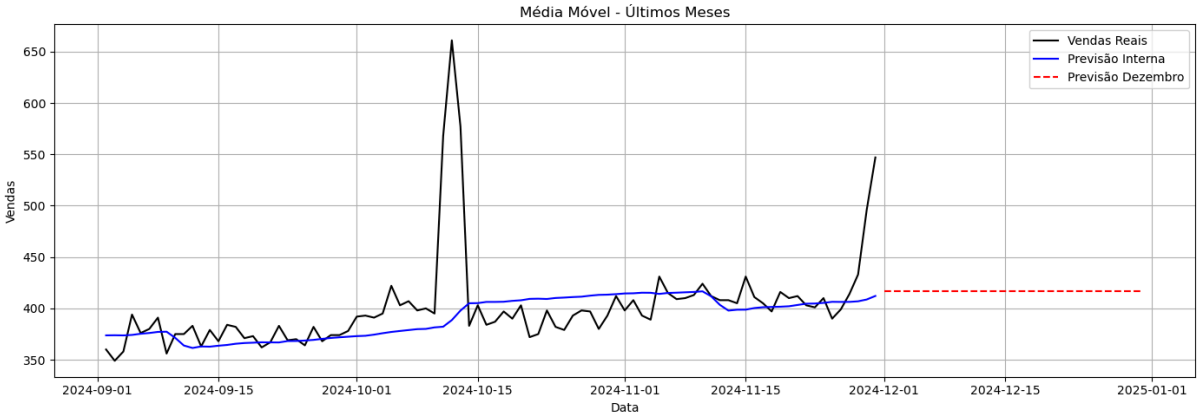
A média móvel simples calcula a média dos últimos *n períodos* – no caso deste projeto, 30 dias – para gerar a previsão do próximo ponto da série. O uso de uma janela de 30 dias está relacionado ao objetivo de captar uma média mensal de comportamento da série. Esse método é eficaz para suavizar flutuações de curto prazo, mas não captura tendências ou padrões sazonais.

**Gráfico 14 - Média Móvel**




**Fonte:** Autoria Própria

**Gráfico 15 - Média Móvel- Últimos meses**

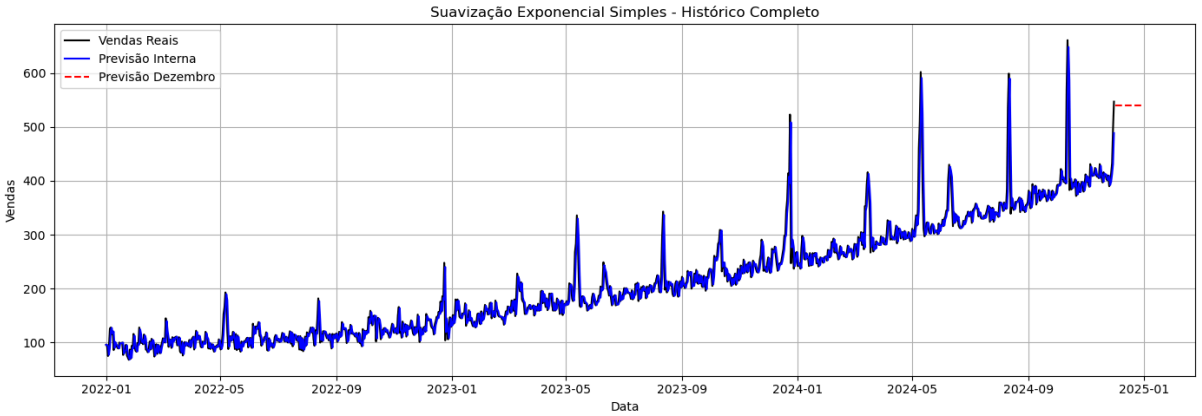



**Fonte:** Autoria Própria

### 4.4.4 Suavização Exponencial Simples (SES)

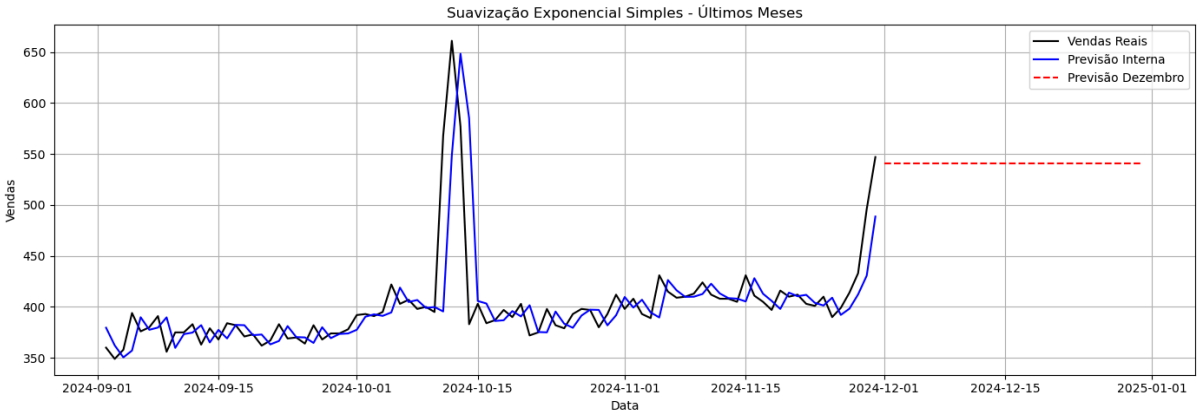
A suavização exponencial simples atribui pesos decrescentes aos valores passados, com maior peso para as observações mais recentes, por meio de um parâmetro de suavização α (alfa). É indicada para séries sem tendência ou sazonalidade, e seu desempenho depende diretamente do valor estimado para o parâmetro.

**Gráfico 16 - SES**




**Fonte:** Autoria Própria

**Gráfico 17 - SES - Últimos meses**

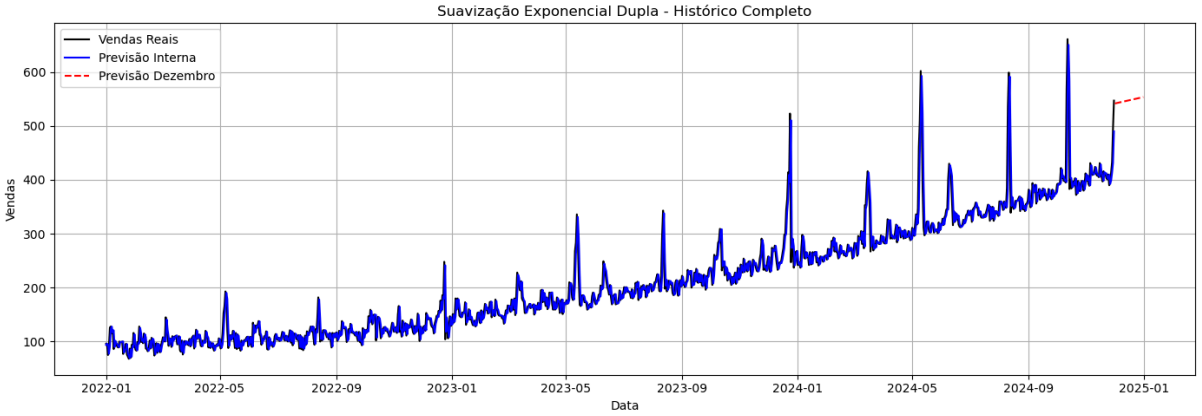



**Fonte:** Autoria Própria

### 4.4.5 Suavização Exponencial Dupla (DES - Holt)

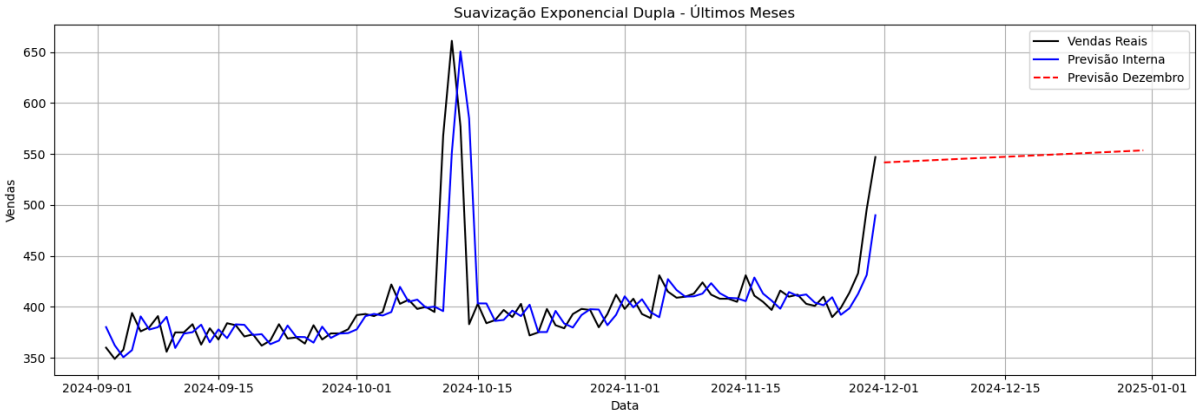
O método de Holt, ou suavização exponencial dupla, é evolução do SES visto que incorpora um componente de tendência, utilizando dois parâmetros: α (alfa), para o nível, e β (beta), para a tendência. É recomendado para séries que apresentam crescimento ou declínio ao longo do tempo.

**Gráfico 18 - DES**




**Fonte:** Autoria Própria

**Gráfico 19 - DES - Últimos meses**

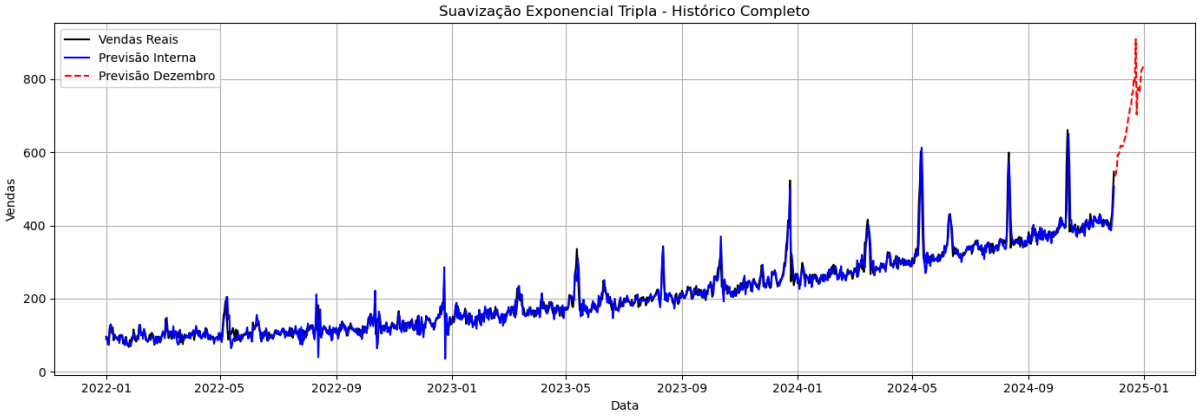



**Fonte:** Autoria Própria

### 4.4.6 Suavização Exponencial Tripla (TES - Holt-Winters)

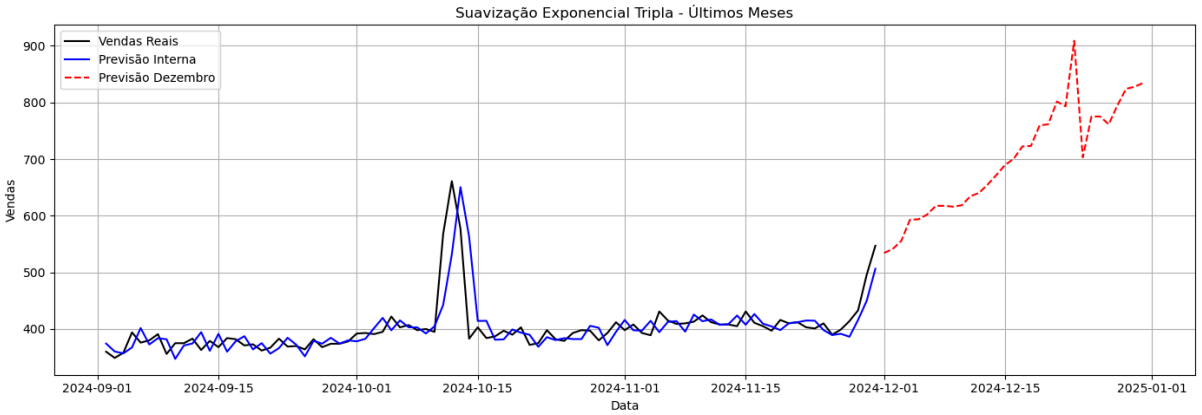
Essa abordagem adiciona um terceiro componente ao modelo de Holt ,a sazonalidade, e com isso é possível prever séries que apresentam comportamento sazonal recorrente. O método exige definição de um período sazonal e envolve três parâmetros: nível (α), tendência (β) e sazonalidade (γ). Portanto, é uma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

**Gráfico 20 - TES**




**Fonte:** Autoria Própria

**Gráfico 21 - TES - Últimos meses**




**Fonte:** Autoria Própria

### 4.4.7 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 absoluto.

O RMSE corresponde à raiz quadrada da média dos quadrados dos erros e tem como característica penalizar mais severamente os desvios maiores. O MAD, por sua vez, representa a média dos valores absolutos dos erros de previsão, oferecendo uma medida direta e menos sensível a valores extremos. O MAPE expressa o erro médio em termos percentuais, o que facilita a comparação entre séries de diferentes magnitudes, embora apresente limitações

quando os valores reais se aproximam de zero. Por fim, o erro absoluto simples corresponde à diferença absoluta entre o valor previsto e o valor observado para cada ponto da série, sendo útil para uma avaliação pontual das previsões.

**Imagem 3 -** Erros dos Modelos

| | MAPE | RMSE | MAD | Erro Padrão |
|---|---|---|---|---|
| **Naive** | 7.865569 | 24.297092 | 14.632519 | 24.293395 |
| **Cumulativo** | 29.514403 | 103.642450 | 77.876624 | 70.317621 |
| **Média Móvel** | 8.062236 | 32.237989 | 17.420354 | 31.849343 |
| **SES** | 7.610436 | 24.178836 | 14.262301 | 24.174250 |
| **DES** | 7.641050 | 24.171326 | 14.277851 | 24.170968 |
| **TES** | 6.906871 | 20.317505 | 12.765021 | 20.316507 |

**Fonte:** Autoria Própria

Com base nos resultados obtidos pelas métricas de avaliação, observa-se que o modelo TES (Triple Exponential Smoothing) apresentou o melhor desempenho geral entre os métodos analisados. Esse modelo obteve os menores valores para todas as métricas consideradas: MAPE (6,91%), RMSE (20,32), MAD (12,77) e Erro Padrão (20,32), indicando previsões mais precisas. Em contraste, o modelo Cumulativo apresentou o pior desempenho, com valores significativamente mais altos em todas as métricas, evidenciando sua limitação para fins preditivos neste contexto. Os modelos Naive, SES e DES apresentaram desempenhos bastante similares, com métricas próximas entre si e resultados melhores que os da média móvel, mas inferiores ao TES.

Bibliografia

ADELEKE, O. 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.

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.

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

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

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, L. R. D.; PEDROSA, M. E. C.; CARDOSO, R. D. L. A. **ANÁLISE DO SISTEMA PRODUTIVO EM UMA HAMBURGUERIA ARTESANAL: UM ESTUDO DE CASO EM NATAL/RN**. . Em: ENEGEP 2024 - ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO. 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, C.; CHEN, Y. 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.

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.

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

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

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.

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.

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

SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. A Systematic Literature Review on Applying CRISP-DM Process Model. **Procedia Computer Science**, v. 181, p. 526–534, 2021.

SEYEDAN, M.; MAFAKHERI, F. 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.

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.

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.