****

**UNIVERSIDADE FEDERAL FLUMINENSE**

**INSTITUTO DE CIÊNCIAS E TECNOLOGIA**

**CURSO DE ENGENHARIA DE PRODUÇÃO**

**KATHERINE ZUNIGA  
MARCIO RODRIGUES  
MATHEUS PAIVA**

**ESTUDO DE CASO SOBRE A EMPRESA SEGROB NOTLAD**

RIO DAS OSTRAS- RJ

2025

**KATHERINE ZUNIGA  
MARCIO RODRIGUES  
MATHEUS PAIVA**

**ESTUDO DE CASO SOBRE A EMPRESA SEGROB NOTLAD**

Estudo de caso sobre a empresa Segrob Notlad no Curso de Graduação em Engenharia de Produção da Universidade Federal Fluminense – Rio das Ostras - RJ como método avaliativo da disciplina Análise Preditiva.

Docente: Dalton Borges.

RIO DAS OSTRAS- RJ  
 2025

**Resumo:**Este estudo de caso tem como foco o desenvolvimento de uma solução preditiva para o varejo de moda rápida, por meio da análise da demanda por camisetas básicas da empresa Segrob Notlad. A pesquisa parte da observação do crescimento expressivo e sustentado nas vendas do produto entre os anos de 2022 e 2024, com o objetivo de construir modelos de previsão capazes de auxiliar o planejamento operacional da marca. Para tanto, foram utilizados dados históricos de vendas diárias, tratados e analisados com base na metodologia CRISP-DM, aplicando técnicas de modelagem preditiva e métricas de avaliação como MAPE, RMSE e MAD. Os resultados apontam para a existência de padrões sazonais semanais e anuais, além da influência de campanhas promocionais no comportamento de consumo. A abordagem adotada permite não apenas prever com maior precisão a demanda futura, mas também gerar insumos estratégicos que fortalecem a tomada de decisão em um mercado altamente dinâmico como o *fast fashion*.

Palavras-chave: *Fast Fashion*. Demanda. CRISP-DM. Aprendizado de Máquina

**SUMÁRIO**

[**1 - Introdução 6**](#_gmrpgcu0ttaf)

[**1.1 - Objetivo 7**](#_vo9fhusg9aue)

[**2 - Fundamentação teórica 7**](#_k6xdn1lzezge)

[**2.1 Análise Cumulativa 7**](#_60zyqp63cckl)

[**2.2 Média Móvel 8**](#_j1abtb9f786w)

[**2.3 Suavização Exponencial Simples 9**](#_72r51shgwf58)

[**2.4 Suavização Exponencial Dupla 10**](#_lxnxsyrsv4z8)

[**2.5 Suavização Exponencial Tripla 10**](#_f2gamo38nce5)

[**2.6 Regressão Linear 11**](#_wk013gxecjak)

[**2.7 K - Vizinhos mais próximo (KNN) 12**](#_99dhb4ntqo6r)

[**2.8 Modelo de Árvore de Decisão 13**](#_y1vl5l9pigjt)

[**2.8.1 Modelo de Florestas Aleatórias (Random Forests) 13**](#_d0p8za1f3msd)

[**2.9 - SVM 14**](#_vzykgaz4dqry)

[**2.10 - Hiperparâmetros 15**](#_jubkcrs19dh9)

[**2.11- Sliding Window 15**](#_a4klzq5q5zxm)

[**2.12 - Grid Search 17**](#_dmwelwl02wf3)

[**3 - Método 18**](#_ejib73yj8fyp)

[**3.1. Entendimento do Negócio (Business Understanding) 19**](#_wdgwwih7hy8r)

[**3.2. Entendimento dos Dados (Data Understanding) 19**](#_ls25wlbyun99)

[**3.3. Preparação dos Dados (Data Preparation) 20**](#_o78v3fvvj0nn)

[**3.4. Modelagem (Modeling) 20**](#_zdty9348lmuz)

[**3.5. Avaliação (Evaluation) 21**](#_3969jze04xvs)

[**3.6. Implementação (Deployment) 21**](#_mtdnfaei9e0r)

[**4 - Estudo de Caso 22**](#_91fowa1vmlb7)

[**4.1 - Entendimento do Negócio 22**](#_jqu3zxqds71g)

[**4.2 - Entendimento dos Dados 23**](#_lv3dqeioms3w)

[**4.2.1 - Gráfico de Nível 23**](#_392540rsa9yc)

[**4.2.2 - Autocorrelação 25**](#_ctlv5z11ecjc)

[**4.2.3 - Média Quarter 26**](#_7pm8n5i50cjp)

[**4.2.4 - Registro Diário de Vendas 28**](#_yo5mg39lhl0c)

[**4.2.5 - Gráfico de Tendências 29**](#_xxq50lb49ypa)

[**4.2.6 - Histograma Sazonalidade Temporal 30**](#_3ebgg4d231mx)

[**4.2.7 - Modelos Preditivos 31**](#_7quo4f7w267d)

[**4.2.7.1 - Cumulativo e Naive 32**](#_knmzhelh2kp1)

[**4.2.7.2 - Média Móvel 34**](#_fy2oagakdfcb)

[**4.2.7.3 - Suavização Exponencial Simples 36**](#_upjidtr2rgjb)

[**4.3 - Aprendizado de Máquina (Machine Learning) 38**](#_si01d3gzqca)

[**4.3.1 - Tipos de aprendizado 39**](#_qji1bpo7zqt1)

[**4.3.2 - Tipos de modelo de Aprendizado Supervisionado 40**](#_ljzqk2m3buv6)

[**4.3.2.1 - Regressão Linear Simples 41**](#_oc5badq07o39)

[**4.3.2.2 - Análise dos Resíduos 42**](#_l4tf1awx66h1)

[**4.3.2.3 - Validação do Modelo 48**](#_dujrij9wf0p3)

[**4.3.2.5 - K - Vizinhos mais próximo (KNN) 51**](#_1lrbr6qa9t9m)

[**4.3.2.6 - Modelo de Arvóre de Decisão 53**](#_t2pqay27mnvq)

[**4.3.2.6.1 - Modelo de Florestas Aleatórias (Random Forests) 53**](#_5aqyjcjdxq8j)

[**4.3.2.7 - SVM 56**](#_yiqzrfciu99y)

[**4.4 - Entendimento Geral dos Dados de Vendas 57**](#_hgqlfsa3c7cm)

[**4.6 Fase de Preparação de Dados 58**](#_gfsdfon1ybq9)

[**4.6.1 Divisão da base em Treino e Teste 59**](#_by5pdf68z7cw)

[**4.6.2 Transformação de dados 59**](#_c1je3tu2apa1)

[**5. Avaliação 62**](#_oaovgjoxzhxs)

[**5.1 Validação Cruzada dos Modelos de Previsão 62**](#_lp3a25ed8tgb)

[**6.Previsão (Implementação) 65**](#_mx9vdb6vhan)

[**7. Conclusão 67**](#_nqcnj3rneoc9)

[**8. Referências 68**](#_v5gmy52zub3)

# **1 - Introdução**

Diante das rápidas transformações no comportamento do consumidor e da crescente competitividade no setor de moda, compreender a atuação de marcas consolidadas no mercado torna-se essencial para a elaboração de estratégias que aliam inovação, eficiência operacional e adaptação às dinâmicas do varejo. Inserida nesse cenário, a Segrob Notlad se destaca como uma marca brasileira de *fast fashion* cuja trajetória é marcada pela combinação entre ousadia, agilidade e espírito urbano.  
 Fundada no Rio de Janeiro, a empresa rapidamente consolidou sua presença nacional por meio de um portfólio focado em design versátil, preços acessíveis e campanhas publicitárias com forte apelo junto ao público jovem e conectado. Atualmente, conta com mais de 80 lojas no Brasil, além de presença internacional por meio de lojas conceito em pontos estratégicos da América do Sul e Europa.  
 A origem da marca está diretamente ligada à história de vida de seu fundador, Segrob Notlad, cuja trajetória entre o leste europeu e o Brasil influenciou diretamente a identidade da empresa, resultando em uma fusão entre estética minimalista e diversidade cultural. Ao longo de sua evolução, a marca passou a integrar de forma estratégica tecnologias de inteligência de mercado e automação de processos em sua cadeia de suprimentos, posicionando-se como referência em adaptação aos ciclos da moda e às exigências do consumidor contemporâneo.  
 Em sua fase mais recente, iniciada em 2025, a empresa reforça seu compromisso com a inovação ao incorporar soluções de inteligência artificial para previsão de demanda e apoio à tomada de decisões. Este movimento sinaliza uma transição para uma atuação ainda mais orientada por dados, destacando a análise preditiva como uma ferramenta central em sua busca por eficiência, assertividade e competitividade em um mercado em constante transformação.

## 1.1 - Objetivo

Este estudo de caso faz parte do movimento constante da Segrob Notlad em busca de inovação e melhorias no seu processo de decisão. O principal objetivo é entender e antecipar a demanda de um dos produtos mais importantes da marca: a camiseta básica. A partir da análise dos dados históricos de vendas, busca-se construir modelos de previsão que auxiliem o planejamento operacional e aumentem a agilidade da empresa para lidar com as mudanças e desafios do setor de *fast fashion*.

O objetivo geral deste trabalho é desenvolver uma previsão confiável da demanda diária de camisetas básicas, de modo a apoiar a tomada de decisão nas áreas de produção, logística e abastecimento. Para alcançar esse propósito, serão realizadas análises sobre o comportamento histórico de vendas, com a identificação de padrões sazonais, tendências de crescimento ou queda, e eventuais eventos atípicos que possam impactar o consumo.

A escolha desse enfoque justifica-se pela importância estratégica de prever corretamente a demanda. Antecipar as necessidades de produção permite à Segrob Notlad otimizar seus processos logísticos, reduzir estoques excessivos, minimizar perdas e melhorar o atendimento ao cliente final, fatores essenciais para manter a competitividade no mercado de *fast fashion*, que se caracteriza por alta volatilidade e forte pressão por agilidade e adaptação contínua.

# 2 - Fundamentação teórica

A presente seção de fundamentação teórica tem como objetivo apresentar os conceitos e as metodologias estatísticas que servem como alicerce para a análise e modelagem dos dados deste relatório. O domínio destes princípios é indispensável para a compreensão das escolhas técnicas adotadas nas etapas de pré-processamento, na análise dos dados e na construção do modelo preditivo.

## 2.1 Análise Cumulativa

É uma técnica de transformação de dados que consiste em agregar sucessivamente os valores de uma série. O resultado é uma nova sequência onde cada ponto representa a soma total de todas as observações anteriores até aquele momento, incluindo a atual. Sua aplicação é comum em análises de séries temporais para converter medições pontuais, como vendas diárias, em uma visão de crescimento contínuo.  
 Uma das principais vantagens dessa abordagem é sua capacidade de destacar a tendência de longo prazo e a magnitude geral de um fenômeno. Ao acumular os valores, as flutuações de curto prazo são suavizadas, revelando uma trajetória de crescimento ou declínio de forma muito mais clara e intuitiva, o que facilita a interpretação da direção geral dos dados.  
 Durante a aplicação, o processo é puramente aritmético: o primeiro valor da série cumulativa é igual ao primeiro valor da série original; o segundo é a soma dos dois primeiros originais, e assim por diante. Em modelos preditivos, essa transformação funciona como uma poderosa técnica de engenharia de atributos, criando uma variável que carrega o "histórico agregado" e a inércia do processo que está sendo analisado.  
 Apesar de sua utilidade para mostrar tendências, a análise cumulativa pode mascarar mudanças recentes no comportamento dos dados, pois uma grande variação em um período recente pode parecer insignificante em comparação com o total já acumulado. Para mitigar essa limitação, a análise cumulativa é frequentemente usada em conjunto com outras métricas, como as médias móveis, que focam no comportamento mais recente da série.

## 2.2 Média Móvel

A média móvel é um método de análise de séries temporais utilizado para suavizar dados e identificar tendências. A técnica funciona calculando uma sequência de médias a partir de subconjuntos de dados de um conjunto maior, utilizando uma "janela" de tamanho fixo que desliza ao longo da série. O resultado é uma nova série que representa o comportamento médio da original, com menos ruído.  
 Uma das principais vantagens desse método é sua eficácia em filtrar a volatilidade de curto prazo, tornando a tendência subjacente mais visível. Ao remover as flutuações aleatórias, a média móvel fornece uma linha mais suave que ajuda a confirmar a direção do movimento dos dados, tornando a interpretação do sinal mais confiável para um analista ou para um modelo preditivo.  
 Durante o cálculo, o algoritmo define uma janela de um determinado número de períodos (k) e calcula a média aritmética dos valores contidos nela. Conforme a janela avança, o valor mais antigo é descartado e o mais novo é incluído. Existem variações, como a Média Móvel Exponencial (MME), que atribuem pesos maiores aos dados mais recentes, tornando o indicador mais reativo a novas informações.  
 Apesar de sua clareza, a média móvel sofre com um problema inerente de atraso (lag), pois é baseada em dados históricos e, por consequência, reage às mudanças de preço ou valor, em vez de antecipá-las. Para mitigar esse problema, usa-se janelas de tempo mais curtas ou combinações diferentes de médias móveis para gerar sinais de cruzamento, que indicam mudanças de momento.  
 Além disso, o tamanho da janela (ou o número de períodos) é um hiperparâmetro fundamental. A escolha desse valor é uma decisão crítica que afeta diretamente o comportamento do indicador, criando um balanço entre o nível de suavização e a velocidade de resposta às mudanças na série de dados.

## 2.3 Suavização Exponencial Simples

A Suavização Exponencial Simples, é um método de previsão de séries temporais utilizado para dados que não apresentam tendência ou sazonalidade. O seu objetivo é produzir uma previsão suavizada, atribuindo pesos exponencialmente decrescentes às observações passadas, de modo que os dados mais recentes tenham maior influência no resultado final.  
 Uma das principais vantagens desse método é sua simplicidade e eficiência computacional. A previsão para o próximo período é calculada como uma média ponderada entre a observação mais recente e a previsão do período anterior, exigindo o armazenamento de poucos dados e permitindo uma atualização rápida e contínua das projeções.  
 Durante a construção do modelo, o algoritmo utiliza um único parâmetro de suavização, chamado alfa (α), que controla a taxa de decaimento dos pesos. Um valor de alfa próximo de 1 torna o modelo muito reativo às observações recentes, enquanto um valor próximo de 0 produz uma linha de previsão muito mais suave e menos sensível a flutuações de curto prazo.  
 Apesar de sua eficácia para séries estacionárias, a Suavização Exponencial Simples não consegue acompanhar dados que possuem uma tendência clara, pois suas previsões serão consistentemente subestimadas em uma tendência de alta e superestimadas em uma de baixa. Para mitigar esse problema, foram desenvolvidas extensões do método, como a suavização exponencial dupla, que incorpora um componente de tendência.  
 Além disso, a escolha do parâmetro **alfa (α)** é um hiperparâmetro crucial. Este valor é definido previamente e afeta diretamente a capacidade do modelo de se ajustar aos dados, sendo frequentemente otimizado para minimizar o erro de previsão no conjunto de treinamento.

## 2.4 Suavização Exponencial Dupla

A Suavização Exponencial Dupla, conhecida como método de *Holt*, é uma extensão da suavização simples projetada para lidar com séries temporais que apresentam uma tendência linear. O modelo funciona decompondo a série em dois componentes principais: o nível (o valor base da série) e a tendência (a taxa de crescimento ou declínio), suavizando cada um deles de forma exponencial e separada.  
 A principal vantagem deste modelo é sua capacidade de gerar previsões que se ajustam dinamicamente a uma tendência crescente ou decrescente. Ao projetar o nível e a tendência para o futuro, o método consegue produzir previsões lineares que extrapolam o comportamento observado, tornando-o superior à suavização simples em cenários não estacionários.  
 Durante o processo de treinamento, o algoritmo utiliza dois parâmetros de suavização: alfa (α), para o nível, e beta (β), para a tendência. O parâmetro alfa controla como o nível da série se adapta às novas observações, enquanto o beta determina a rapidez com que a inclinação da tendência é atualizada com base nas mudanças recentes.  
 Embora seja eficaz para dados com tendência, o método de *Holt* não é adequado para séries que exibem sazonalidade, pois ele não possui um mecanismo para capturar padrões que se repetem em intervalos fixos.   
 Adicionalmente, os parâmetros alfa (α) e beta (β) são hiperparâmetros que precisam ser definidos antes do treinamento. A correta calibração desses valores é fundamental para o desempenho do modelo, pois eles controlam o balanço entre a estabilidade e a capacidade de resposta do modelo às mudanças no nível e na tendência da série.

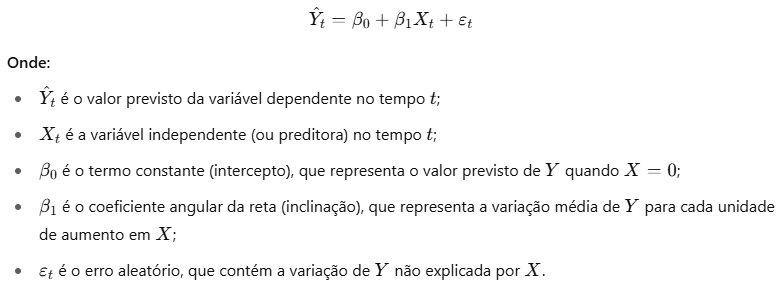
## 2.5 Suavização Exponencial Tripla

A Suavização Exponencial Tripla, mais conhecida como método de *Holt-Winters*, é uma técnica avançada de previsão que estende a suavização dupla para incorporar um componente de sazonalidade. O modelo é capaz de lidar com séries temporais que possuem tanto uma tendência quanto padrões sazonais, decompondo os dados em três componentes: nível, tendência e sazonalidade.  
 A grande vantagem deste método é sua robustez e versatilidade, sendo um dos modelos clássicos mais utilizados para previsão em contextos de negócios, como projeção de vendas ou demanda, onde tendências e ciclos sazonais são comuns. Ele pode capturar padrões sazonais de forma aditiva, quando a variação é constante, ou multiplicativa, quando a variação cresce proporcionalmente ao nível da série).  
 Na construção do modelo, são utilizados três parâmetros de suavização: alfa (α) para o nível, beta (β) para a tendência, e gama (γ) para a sazonalidade. Cada parâmetro controla a taxa de atualização de seu respectivo componente, permitindo que o modelo se adapte dinamicamente a mudanças na linha de base, na inclinação da tendência e no padrão sazonal.  
 Apesar de seu poder, o método de *Holt-Winters* assume que os padrões de tendência e sazonalidade são relativamente estáveis ao longo do tempo. Mudanças estruturais abruptas na série podem degradar a performance do modelo, exigindo um reajuste dos parâmetros ou a consideração de modelos mais complexos. Além disso, a precisão das previsões de longo prazo pode diminuir se a tendência ou a sazonalidade se alterarem significativamente.  
 A definição dos hiperparâmetros alfa (α), beta (β) e gama (γ), juntamente com o tipo de sazonalidade, aditiva ou multiplicativa, é essencial para o ajuste do modelo. Esses valores são tipicamente otimizados de forma automática para minimizar uma métrica de erro, garantindo que o modelo capture os padrões dos dados da forma mais precisa possível.

## 2.6 Regressão Linear

A regressão linear simples é uma técnica estatística fundamental utilizada para modelar a relação entre duas variáveis: uma variável dependente e uma variável independente. Trata-se da forma mais básica de regressão, em que se busca descrever, por meio de uma equação linear, como a variável explicativa (ou preditora) influencia a variável resposta ao longo do tempo ou de observações, como podemos ver na Equação 1.

Equação 1- Regressão Linear Simples



Elaborado pelos autores,2025

O objetivo principal do modelo de regressão linear simples é encontrar a melhor reta de ajuste que possa prever a variável dependente com o menor erro possível. Para isso, utiliza-se o método dos mínimos quadrados, que estima os coeficientes do modelo de forma a minimizar a soma dos quadrados dos resíduos, isto é, das diferenças entre os valores observados e os valores previstos. Embora simples, esse modelo assume que há uma relação linear entre as variáveis e que os erros são independentes, com média zero e variância constante, como podemos ver na Equação 2.

## 2.7 K - Vizinhos mais próximo (KNN)

É um algoritmo de aprendizado supervisionado amplamente aplicado em problemas de classificação e regressão. Sua lógica básica consiste em classificar ou prever uma nova observação considerando as K observações mais próximas no conjunto de treinamento, definindo a saída com base na classe majoritária (classificação) ou na média dos valores (regressão).

Uma das principais vantagens do KNN é sua simplicidade conceitual e implementação direta. No entanto, por depender exclusivamente dos dados históricos, o modelo pode apresentar dificuldades em generalizar em casos com ruído ou valores extremos, tornando-se sensível à escala das variáveis.

O desempenho do modelo é altamente influenciado pela escolha do valor K, que é um hiperparâmetro crítico. Valores pequenos de K podem levar a modelos muito sensíveis a ruídos, enquanto valores altos podem gerar previsões excessivamente suavizadas e generalistas. A seleção adequada do valor de K geralmente é feita por meio de validação cruzada.

Entre as limitações do KNN destaca-se sua complexidade computacional elevada para grandes bases de dados, uma vez que todas as observações precisam ser armazenadas e comparadas continuamente durante a predição. Mesmo assim, devido à sua simplicidade, interpretabilidade direta e eficácia em contextos menores ou controlados, o método continua sendo amplamente utilizado.

## 2.8 Modelo de Árvore de Decisão

As árvores de decisão são modelos de aprendizado supervisionado utilizados tanto para classificação quanto para regressão. O nome se deve à sua estrutura em forma de árvore, que começa em um nó raiz e se ramifica em diversos caminhos possíveis com base nas decisões tomadas a partir dos atributos dos dados. Cada divisão leva a novos nós, até que se alcance um nó terminal (ou folha), que representa a saída prevista pelo modelo.

Uma das principais vantagens desse tipo de modelo é sua facilidade de interpretação. As regras de decisão extraídas da árvore são simples de entender e podem ser visualizadas em forma de fluxograma, o que torna o processo de decisão transparente e acessível até mesmo para usuários não técnicos.  
 Durante a construção da árvore, o algoritmo avalia diferentes variáveis e critérios de divisão para selecionar os que melhor separam os dados. Em problemas de classificação, medidas como impureza de Gini ou entropia são utilizadas para definir os melhores pontos de divisão. Já em regressão, é comum o uso de métricas como o erro quadrático médio.  
 Apesar de sua simplicidade e clareza, as árvores de decisão podem sofrer com o *overfitting*, especialmente quando são muito profundas ou complexas, ajustando-se demais aos dados de treinamento. Para mitigar esse problema, técnicas como poda são aplicadas para limitar o crescimento da árvore. Alternativamente, podem ser usados métodos mais robustos, como florestas aleatórias ou *boosting*, que combinam múltiplas árvores para melhorar a performance e a generalização do modelo.  
 Além disso, as podas e outras configurações que controlam o comportamento do algoritmo durante o treinamento são chamadas de hiperparâmetros. Esses valores são definidos previamente e afetam diretamente o desempenho final do modelo.

#### 2.8.1 Modelo de Florestas Aleatórias (*Random Forests*)

Florestas Aleatórias, ou *Random Forests*, são um tipo de modelo de aprendizado de máquina que é uma extensão do modelo de árvores de decisão. Em termos simples, uma floresta aleatória é uma coleção de árvores de decisão. Em vez de confiar em uma única árvore de decisão, o modelo de Florestas Aleatórias toma a decisão com base nas previsões de várias árvores de decisão, fazendo uma espécie de “votação” entre as previsões das diferentes árvores.  
 Essa abordagem visa aumentar a precisão e reduzir a variabilidade dos modelos individuais. Cada árvore na floresta é treinada com um subconjunto aleatório dos dados e, em cada divisão interna, considera apenas uma amostra aleatória das variáveis. Essa aleatoriedade introduz diversidade entre as árvores, o que ajuda a evitar o problema de *overfitting*, comum em árvores de decisão isoladas.  
 No processo de previsão, para problemas de classificação, cada árvore “vota” em uma classe, e a classe com mais votos é escolhida como resultado final. Para problemas de regressão, a média das previsões das árvores é utilizada. Isso torna o modelo mais estável e robusto, especialmente em bases de dados complexas ou com ruído.  
 As Florestas Aleatórias também oferecem ferramentas úteis, como a importância das variáveis, que permite identificar quais atributos mais contribuíram para a decisão do modelo, uma vantagem valiosa para análises interpretativas.

## 2.9 - SVM

O *Support Vector Machines* (SVM) é um método de aprendizado supervisionado amplamente aplicado em classificação e regressão. Seu princípio básico consiste em identificar o hiperplano ótimo que separa diferentes classes com a maior margem possível. Os pontos mais próximos dessa fronteira, denominados vetores de suporte, determinam a posição do hiperplano.

Uma grande vantagem do SVM é sua capacidade de generalizar bem, reduzindo riscos de *overfitting* ao maximizar a margem entre as classes. Para tratar dados não linearmente separáveis, o método emprega funções kernel, que mapeiam os dados para espaços dimensionais superiores, facilitando sua separação. Os kernels mais comuns são linear, polinomial e Gaussiano (RBF).

O desempenho do modelo SVM depende fortemente da escolha adequada dos hiperparâmetros, como o parâmetro C (regularização) e o gamma (γ) (para kernels não lineares), ajustados geralmente por técnicas como *Grid Search* e validação cruzada.

Apesar da robustez, o método pode apresentar maior custo computacional em grandes bases de dados e menor interpretabilidade comparado a métodos mais simples. Ainda assim, devido à sua eficácia, o SVM é amplamente utilizado em contextos que demandam alta precisão e boa generalização.

## 2.10 - Hiperparâmetros

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

* *n\_estimators* (Número de árvores em uma floresta): É um hiperparâmetro específico para modelos de Floresta Aleatória e determina o número de árvores na floresta. Em geral, mais árvores podem resultar em um modelo mais robusto e menos propenso a sobreajuste, mas o treinamento do modelo se tornará computacionalmente mais exigente.
* *max\_depth* (Profundidade máxima da árvore): Limita o número máximo de níveis que cada árvore de decisão pode ter. Ajustar este hiperparâmetro pode ajudar a prevenir o sobreajuste (se for definido um valor baixo) ou permitir que o modelo se ajuste mais aos dados de treinamento (se for definido um valor alto).
* *min\_samples\_split* (Número mínimo de amostras para divisão de um nó): Determina o número mínimo de amostras necessárias para que um nó interno possa ser dividido. Isso ajuda a controlar o sobreajuste, pois impõe um limite na granularidade das divisões que a árvore pode fazer.
* *min\_samples\_leaf* (Número mínimo de amostras por folha): Especifica o número mínimo de amostras que devem estar presentes em um nó folha (um nó terminal que não se divide mais). Semelhante ao *min\_samples\_split*, este parâmetro também ajuda a evitar o sobreajuste ao impor uma restrição na complexidade das folhas da árvore.
* *max\_features* (Número máximo de recursos): Define o número de recursos (variáveis) a serem considerados ao procurar a melhor divisão em um nó. Utilizar um subconjunto de recursos em cada divisão aumenta a aleatoriedade e pode tornar o modelo mais robusto contra o sobreajuste.

## 2.11- *Sliding Window*

A técnica conhecida como *Sliding Window* (janela deslizante) é um método amplamente utilizado na análise e modelagem de séries temporais, especialmente em contextos de validação cruzada e aprendizado de máquina para previsão. A janela deslizante funciona segmentando sequencialmente um conjunto de dados em partes consecutivas, com um tamanho fixo e constante, que avança progressivamente pelo conjunto total, permitindo uma análise dinâmica e adaptável dos padrões temporais contidos nos dados.

Na prática, o método estabelece inicialmente uma janela temporal (por exemplo, 12 meses, 30 dias, etc.) dentro da qual ocorre o treinamento dos modelos preditivos. A cada passo subsequente, essa janela avança no tempo, descartando os dados mais antigos e incorporando novos dados disponíveis à medida que surgem, de modo a manter a mesma quantidade fixa de observações. Dessa forma, o método gera continuamente conjuntos atualizados para treinamento e testes, possibilitando a modelagem e previsão contínua ao longo do tempo.

Uma das grandes vantagens desse método é sua capacidade em lidar com séries temporais dinâmicas, que frequentemente apresentam mudanças em sua estrutura ou comportamento ao longo do tempo. A janela deslizante garante que o modelo sempre tenha acesso aos dados mais recentes, permitindo capturar rapidamente tendências emergentes, alterações estruturais ou mudanças sazonais. Essa característica torna o método especialmente eficaz em ambientes como varejo e mercados financeiros, onde decisões precisam considerar as condições mais atualizadas possíveis.

Durante o processo de validação cruzada com *Sliding Window*, o conjunto de dados é dividido sequencialmente em um conjunto de treino (*window*) e um conjunto de teste imediatamente posterior (horizonte de previsão). A janela é então deslocada sucessivamente para frente no tempo, produzindo múltiplas avaliações do desempenho do modelo, e permitindo uma análise robusta da estabilidade e consistência das previsões ao longo do tempo.

Apesar de sua eficácia e ampla adoção, o método *Sliding Window* possui alguns desafios importantes. O primeiro deles é a escolha correta do tamanho da janela, que se configura como um hiperparâmetro crítico. Uma janela muito pequena pode não capturar adequadamente tendências ou ciclos maiores, resultando em modelos excessivamente sensíveis ao ruído. Por outro lado, uma janela muito grande pode limitar a capacidade do modelo em se adaptar rapidamente a mudanças recentes, criando previsões desatualizadas ou com atraso (*lag*).

Outro ponto importante é a definição do passo de avanço da janela (*step size*), que determina com que frequência os dados mais antigos são removidos e novos dados são incorporados. Passos menores levam a avaliações mais frequentes, possibilitando rápida adaptação do modelo a mudanças nos dados, enquanto passos maiores podem reduzir o esforço computacional, mas retardar a capacidade de resposta aos novos padrões.

Do ponto de vista prático, a técnica de janela deslizante pode ser combinada com diversos modelos preditivos, incluindo regressões dinâmicas, modelos autorregressivos (ARIMA), métodos de aprendizado de máquina como *Random Forest*, KNN, e até redes neurais. Independentemente do modelo escolhido, o *Sliding Window* fornece uma base sólida para realizar previsões contínuas e adaptáveis em contextos onde os dados evoluem constantemente.

A técnica *Sliding Window* é reconhecida como fundamental para avaliações robustas em séries temporais. Autores destacam sua capacidade em simular realisticamente cenários de previsão contínua, além de ressaltar a importância da escolha adequada de hiperparâmetros para maximizar o desempenho preditivo.

Em síntese, a janela deslizante se mostra um método essencial e eficiente para modelagem e validação em séries temporais dinâmicas, permitindo prever comportamentos futuros com maior precisão e confiabilidade, características altamente desejáveis no planejamento operacional e estratégico.

## 2.12 - *Grid Search*

A técnica de *Grid Search* é um método de busca exaustiva utilizado para a otimização de hiperparâmetros em modelos de aprendizado de máquina. Os hiperparâmetros são configurações definidas antes do treinamento de um modelo, como por exemplo, o número de vizinhos no KNN, o tipo de kernel no SVM ou a profundidade máxima de uma árvore de decisão, e influenciam diretamente o desempenho final do modelo.

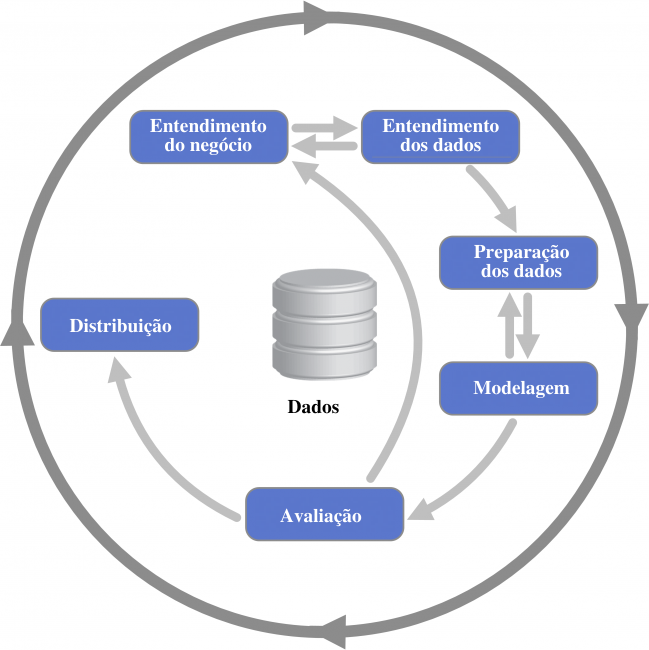
O *Grid Search* funciona testando todas as combinações possíveis de valores especificados para cada hiperparâmetro. Para isso, define-se uma grade (*grid*) de valores para cada parâmetro de interesse, e o algoritmo treina e avalia o modelo em cada uma das combinações possíveis. A performance de cada combinação é avaliada com base em uma métrica definida, geralmente utilizando validação cruzada para garantir uma avaliação mais robusta e evitar *overfitting*.

Ao final do processo, o *Grid Search* retorna a melhor combinação de hiperparâmetros segundo a métrica escolhida. Apesar de ser um método computacionalmente custoso, especialmente quando há muitos parâmetros e valores, ele é extremamente eficaz para garantir que o modelo esteja ajustado da melhor forma possível ao problema.

# **3 - Método**

Para a condução deste projeto de análise e mineração de dados, foi adotada a metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Este é um modelo de processo robusto e amplamente reconhecido na indústria, que organiza o desenvolvimento de projetos de ciência de dados em um ciclo de vida estruturado e iterativo. A escolha do CRISP-DM se justifica por sua flexibilidade e foco em alinhar as soluções técnicas aos objetivos de negócio, garantindo que os resultados gerados sejam não apenas precisos, mas também relevantes e acionáveis.

Figura 1



A metodologia é composta por seis fases principais, conforme detalhado na Figura 1. É importante notar que o processo não é estritamente linear; é comum retornar a fases anteriores à medida que novos aprendizados são descobertos, caracterizando a natureza cíclica do modelo.

#### 3.1. Entendimento do Negócio (Business Understanding)

Esta fase inicial é o alicerce do projeto e foca em compreender os objetivos e requisitos sob uma perspectiva de negócio. O objetivo é converter esse conhecimento em uma definição de problema de mineração de dados e em um plano preliminar para alcançá-lo.

Atividades Principais:

* Determinação dos Objetivos de Negócio: Identificar e compreender as metas que esperam alcançar com o projeto (ex: reduzir a perda de clientes, aumentar as vendas, detectar fraudes).
* Avaliação da Situação Atual: Levantar os recursos disponíveis, requisitos, restrições e riscos do projeto.
* Definição dos Objetivos de Mineração de Dados: Traduzir os objetivos de negócio em metas técnicas e específicas para a análise de dados (ex: prever quais clientes têm maior probabilidade de churn nos próximos 3 meses).
* Elaboração do Plano do Projeto: Detalhar as etapas, cronograma, ferramentas e técnicas a serem utilizadas para atingir os objetivos definidos.

#### 3.2. Entendimento dos Dados (*Data Understanding)*

Uma vez que os objetivos estão claros, a atenção se volta para os dados. Esta fase começa com a coleta de dados e prossegue com atividades para se familiarizar com eles, identificar problemas de qualidade, descobrir primeiros insights e/ou detectar subconjuntos interessantes para formar hipóteses.

Atividades Principais:

* Coleta de Dados Iniciais: Adquirir os dados de todas as fontes disponíveis e relevantes para o problema.
* Descrição dos Dados: Examinar as propriedades "superficiais" dos dados, como o formato, o número de registros, os tipos de variáveis e seus significados.
* Exploração dos Dados: Utilizar técnicas de análise exploratória, como estatísticas descritivas e visualizações (histogramas, gráficos de dispersão), para investigar os dados mais a fundo.
* Verificação da Qualidade dos Dados: Avaliar a completude (valores ausentes), consistência e precisão dos dados, documentando quaisquer problemas encontrados.

#### 3.3. Preparação dos Dados (*Data Preparation*)

Esta é frequentemente a fase mais intensiva e demorada de um projeto de mineração de dados, podendo consumir de 50% a 80% do tempo total. O objetivo é transformar os dados brutos coletados em um conjunto de dados final, limpo e formatado, que servirá de entrada para as ferramentas de modelagem.

Atividades Principais:

* Seleção de Dados: Decidir quais dados (colunas e linhas) são relevantes para a análise e serão incluídos no conjunto final.
* Limpeza de Dados: Tratar os problemas de qualidade identificados, como o preenchimento de valores ausentes, a correção de erros ou a remoção de outliers.
* Construção de Novos Atributos (Feature Engineering): Criar novas variáveis a partir das existentes que possam ser mais úteis para a modelagem (ex: calcular a idade de um cliente a partir da data de nascimento).
* Integração de Dados: Combinar dados de múltiplas fontes (tabelas, arquivos) em um único conjunto de dados coeso.
* Formatação dos Dados: Realizar transformações para que os dados se adequem aos requisitos da técnica de modelagem a ser usada (ex: converter variáveis categóricas em numéricas).

#### 3.4. Modelagem (*Modeling*)

Com os dados preparados, a fase de modelagem seleciona e aplica diversas técnicas de mineração de dados para que os padrões sejam encontrados. Nesta etapa, os parâmetros dos modelos são calibrados para otimizar seu desempenho.

Atividades Principais:

* Seleção da Técnica de Modelagem: Escolher os algoritmos mais apropriados para o tipo de problema (ex: árvores de decisão, prophet).
* Desenho do Teste do Modelo: Definir um procedimento para avaliar a qualidade e a validade do modelo, geralmente dividindo os dados em conjuntos de treino e teste.
* Construção do Modelo: Executar o algoritmo de modelagem nos dados preparados para criar um ou mais modelos.
* Avaliação Técnica do Modelo: Avaliar o desempenho do modelo sob uma perspectiva técnica, utilizando métricas de performance relevantes.

#### 3.5. Avaliação (*Evaluation*)

Nesta fase, os modelos construídos são avaliados de forma mais rigorosa, sob a ótica dos objetivos de negócio definidos na primeira fase. O objetivo aqui é garantir que o modelo atenda de forma eficaz aos critérios de sucesso do negócio e não apenas apresente um bom desempenho técnico.

Atividades Principais:

* Avaliação dos Resultados: Comparar os resultados do modelo com os critérios de sucesso do negócio. O modelo resolve o problema para o qual foi criado?
* Revisão do Processo: Revisar todo o trabalho executado, verificando se todas as etapas foram cumpridas corretamente e se nada importante foi esquecido.
* Determinação dos Próximos Passos: Com base nos resultados, decidir se o projeto está pronto para a implantação, se necessita de mais iterações (retornando a fases anteriores) ou se deve ser descartado.

#### 3.6. Implementação (*Deployment*)

A criação de um modelo não é o fim do projeto. O conhecimento obtido precisa ser organizado e apresentado de uma forma que o cliente ou os *stakeholders* possam utilizar. Esta fase pode variar desde a criação de um relatório final até a implementação de um processo de mineração de dados repetível em um sistema de produção.

Atividades Principais:

* Planejamento da Implementação: Desenvolver um plano para integrar o modelo aos processos de negócio da organização.
* Planejamento de Monitoramento e Manutenção: Criar um plano para monitorar o desempenho do modelo ao longo do tempo e mantê-lo atualizado.
* Revisão do Projeto: Realizar uma retrospectiva final para documentar as experiências e lições aprendidas durante o projeto.

# **4 - Estudo de Caso**

O estudo de caso é uma abordagem metodológica que permite investigar, com profundidade e contextualização, situações reais enfrentadas por empresas, organizações ou setores específicos. Ao focar em um único objeto de análise, como um produto, processo ou decisão empresarial, esse tipo de estudo busca compreender os fatores que influenciam determinada realidade, identificar padrões de comportamento e propor soluções baseadas em dados concretos. Diferente de análises puramente teóricas ou generalistas, o estudo de caso utiliza informações reais e circunstâncias específicas, o que o torna uma ferramenta valiosa para gerar diagnósticos detalhados e embasamento prático para a tomada de decisões estratégicas.

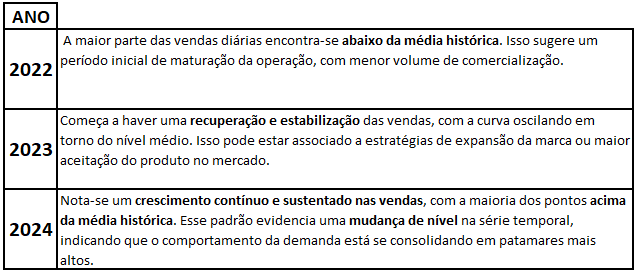
## 4.1 - Entendimento do Negócio

A Segrob Notlad atua no mercado de *fast fashion*, um setor conhecido pela alta dinamicidade, pela exigência de respostas rápidas às mudanças de comportamento do consumidor e pela intensa pressão por inovação constante. Dentro desse cenário, a empresa construiu sua estratégia combinando inteligência de mercado, automação da cadeia de suprimentos e forte presença no ambiente digital e físico. O foco é atender de forma eficiente um público jovem e urbano, que valoriza tanto o acesso rápido a novidades quanto uma boa relação custo-benefício.  
 O produto central deste estudo é a camiseta básica, um dos itens mais relevantes dentro do portfólio desta empresa. Por seu perfil atemporal e sua alta rotatividade, a camiseta básica representa um componente fundamental para a geração de receita e para a manutenção do fluxo operacional de lojas e centros de distribuição. Ter previsibilidade de sua demanda é essencial para otimizar a produção, controlar estoques e assegurar a disponibilidade contínua desse item, minimizando perdas por excesso ou falta de produtos nas lojas.  
 A empresa, ciente da importância estratégica da camiseta básica, busca com este estudo de caso aumentar sua capacidade de planejamento a partir da análise crítica dos dados de vendas históricos. A utilização de técnicas de análise preditiva visa não apenas antecipar o volume necessário para um período específico, mas também fortalecer sua estrutura de tomada de decisão frente a eventos sazonais, variações no comportamento do consumidor e mudanças de mercado.  
 O ambiente em que a Segrob Notlad está inserida é caracterizado por forte influência de datas comerciais, campanhas promocionais, variações climáticas e eventos sociais e culturais. Além disso, o próprio ciclo da moda, ainda que em um produto básico, pode impactar picos ou quedas de demanda em determinados períodos. Outro fator relevante é a operação multicanal da empresa, integrando lojas físicas e estratégias digitais, o que exige ainda mais precisão na previsão de demanda para garantir uma resposta eficiente em todos os pontos de venda.  
 O desafio proposto se insere diretamente na estratégia de fortalecer o uso de inteligência de dados como diferencial competitivo, permitindo à Segrob Notlad uma atuação ainda mais ágil e assertiva no mercado. A previsão de demanda para a camiseta básica, portanto, representa não apenas uma ação tática de curto prazo, mas também uma evolução no modo como a empresa utiliza dados para sustentar seu crescimento e eficiência operacional em um ambiente de alta volatilidade.

## 4.2 - Entendimento dos Dados

### **4.2.1 - Gráfico de Nível**

Apresenta a evolução das vendas diárias de camisetas básicas masculinas da marca Segrob Notlad entre janeiro de 2022 e novembro de 2024. A linha azul representa o volume diário de vendas, enquanto a linha vermelha tracejada indica o nível médio histórico de vendas, calculado em 214,16 unidades por dia.  
 Ao longo do tempo, observa-se uma mudança progressiva no comportamento da demanda (Figura 1):

  
FIGURA 1 - COMPORTAMENTO DE DEMANDA - FONTE: AUTORES,2025.

Além disso, o gráfico exibe picos regulares de venda, que podem estar relacionados a eventos promocionais, sazonalidade comercial ou datas comemorativas específicas (como *Black Friday* e Natal). Esses picos, visivelmente acima da média, reforçam a importância de ações estratégicas planejadas.

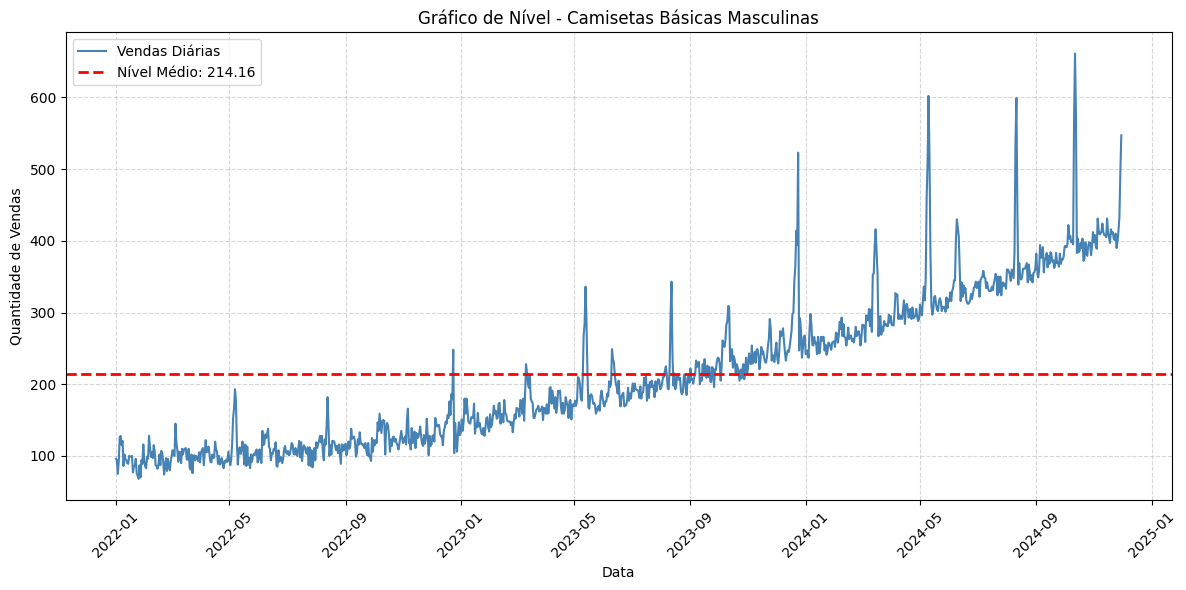


FIGURA 2 - GRÁFICO DE NÍVEL - FONTE: AUTORES,2025.

A curva de nível serve como uma ferramenta de referência para avaliação de desempenho. Sua aplicação permite identificar mudanças de comportamento na demanda ao longo do tempo, determinar períodos que exigem reforço logístico ou operacional (como os de pico), subsidiar a construção de modelos preditivos baseados em mudanças estruturais de patamar, fundamentais para o planejamento de estoques e produção.  
 Com isso, o gráfico de nível confirma que a Segrob Notlad vem apresentando crescimento consistente nas vendas de camisetas básicas, com indicações claras de que a média histórica utilizada pode em breve tornar-se obsoleta diante do novo patamar de desempenho observado.

### **4.2.2 - Autocorrelação**

Apresenta as médias mensais de vendas de camisetas básicas masculinas no período de Janeiro de 2022 a Novembro de 2024. Esse tipo de análise permite avaliar quanto os valores de uma série temporal estão relacionados com seus próprios valores defasados no tempo (lags), sendo uma ferramenta fundamental para identificar padrões sazonais e dependência temporal.  
 No gráfico, podemos observar que as primeiras defasagens (lags 1 a 5) apresentam valores de autocorrelação altos e positivos (acima de 0,6), indicando que as vendas de um mês estão fortemente correlacionadas com os meses imediatamente anteriores. A partir do lag 6, a autocorrelação vai diminuindo progressivamente, sugerindo uma perda gradual de memória da série ao longo do tempo. A curva cruza a linha de significância estatística (faixa azul claro) por volta do lag 10 a 11, momento em que os valores deixam de ser significativamente diferentes de zero. Isso indica que, após cerca de 10 meses, a influência das vendas passadas sobre as atuais se torna desprezível. A presença de autocorrelação significativa nos primeiros meses reforça a existência de padrões estruturais e possíveis tendências, que devem ser considerados na construção de modelos preditivos.

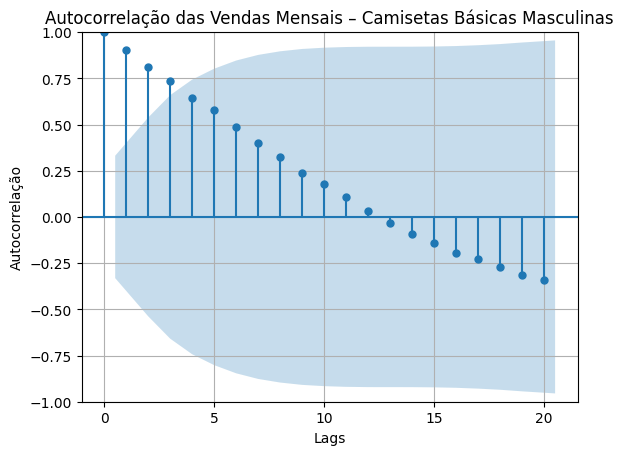


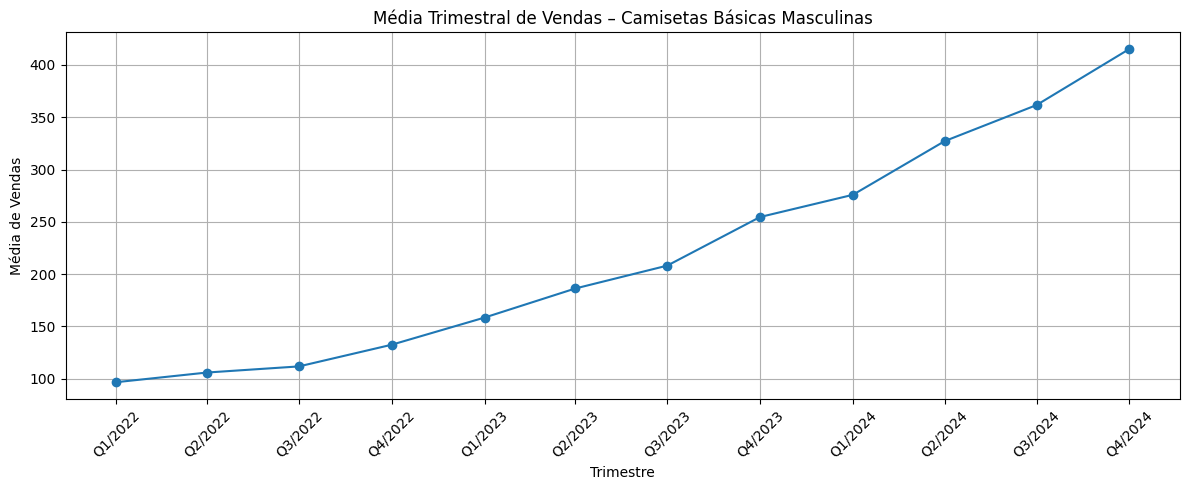
FIGURA 3 - AUTOCORRELAÇÃO - FONTE: AUTORES,2025.

A análise de autocorrelação confirma que as vendas mensais seguem uma estrutura temporal dependente, ou seja, os valores atuais são influenciados por períodos anteriores, especialmente nos primeiros 6 a 8 meses. Isso é consistente com comportamentos típicos de negócios de moda rápida, em que decisões de produção e marketing têm efeitos duradouros, mas não permanentes. A compreensão desses padrões é essencial para o desenvolvimento de modelos de previsão mais robustos, ou modelos baseados em janelas de tempo.

### **4.2.3 - Média Quarter**

Exibe a evolução da média trimestral de vendas de camisetas básicas masculinas da marca Segrob Notlad, com base nos dados compreendidos entre o primeiro trimestre de 2022 (Q1/2022) e o terceiro trimestre de 2024 (Q3/2024). A análise foi realizada a partir do agrupamento dos dados de vendas por trimestres civis e cálculo da média de unidades vendidas em cada período. A linha do gráfico mostra uma tendência de crescimento contínuo e acentuado da demanda ao longo do tempo:

* No início da série (Q1/2022), a média trimestral de vendas era inferior a 120 unidades por dia.
* A partir de Q3/2022, observa-se uma trajetória de alta constante, sem quedas ou estabilizações significativas.
* Em Q3/2023 há um salto mais expressivo, possivelmente relacionado a ações promocionais ou expansão da marca.
* O último ponto do gráfico, Q3/2024, apresenta uma média superior a 400 unidades diárias, representando um crescimento de mais de 250% em relação ao início da série.

  
 FIGURA 4 - GRÁFICO DE MÉDIA QUARTER - FONTE: AUTORES,2025.

Este gráfico evidencia uma clara mudança de patamar na operação comercial da empresa. O comportamento crescente e estável da curva indica que o produto em análise apresenta forte aceitação no mercado e potencial de escalabilidade. Do ponto de vista analítico, a evolução trimestral:

* Confirma a validação do produto ao longo do tempo.
* Reforça a necessidade de ajustar modelos preditivos para acomodar uma tendência de alta acentuada.
* Serve como base para projeções futuras mais agressivas de produção e logística.

A análise trimestral suaviza variações diárias e mensais, permitindo visualizar com mais clareza a tendência estrutural de crescimento da demanda. Trata-se, portanto, de uma métrica importante para planejamento estratégico e previsão de longo prazo.

### **4.2.4 - Registro Diário de Vendas**

Apresenta o registro diário das vendas de camisetas básicas masculinas da marca Segrob Notlad, no período de janeiro de 2022 a novembro de 2024. O gráfico permite observar de forma clara o comportamento da demanda ao longo do tempo. No início da série, os volumes de venda diários são mais baixos, oscilando majoritariamente entre 80 e 150 unidades. Com o passar dos meses, especialmente a partir de 2023, nota-se uma tendência consistente de crescimento. Esse aumento se intensifica ao longo de 2024, quando os registros diários frequentemente ultrapassam 300 unidades, chegando, em alguns casos, a ultrapassar a marca de 600 unidades em um único dia.  
 Além da tendência geral de crescimento, o gráfico também evidencia a presença de picos regulares e espaçados, que podem estar relacionados a eventos promocionais, campanhas de liquidação, datas comemorativas ou ações de marketing da marca — o que é típico do setor de *fast fashion*. A progressão contínua da linha de vendas, com redução gradual dos dias de baixa performance, sugere um processo de consolidação de mercado, fidelização de clientes e possível ampliação da rede de distribuição.

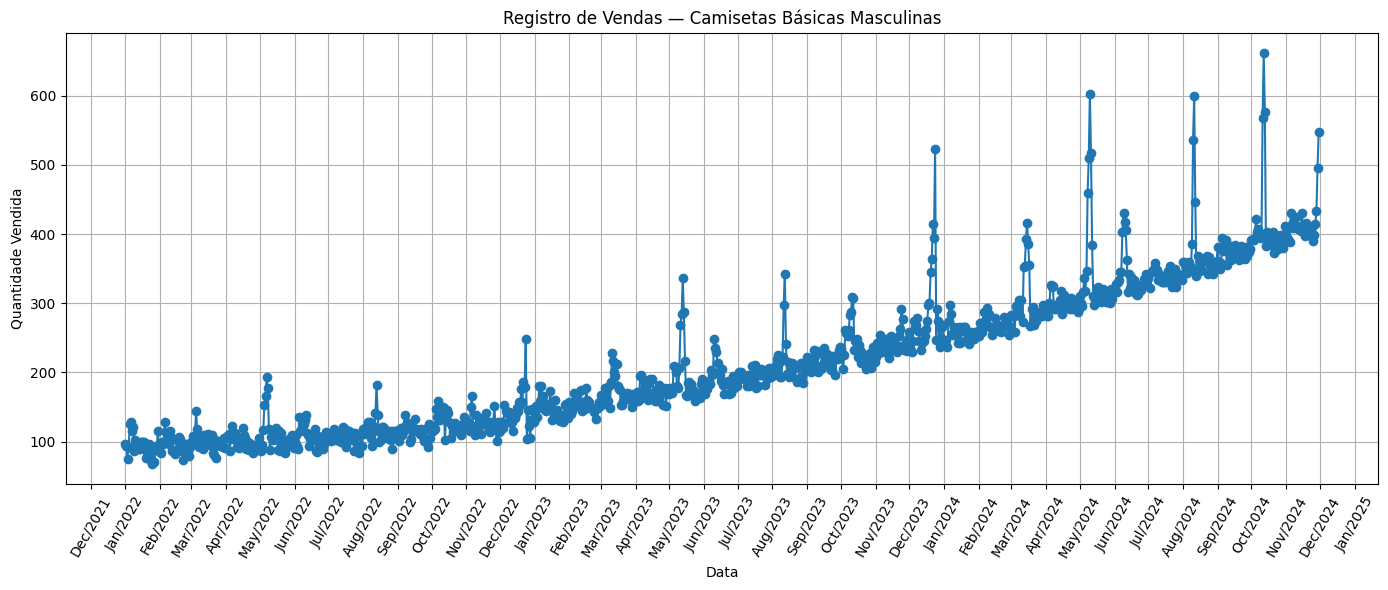
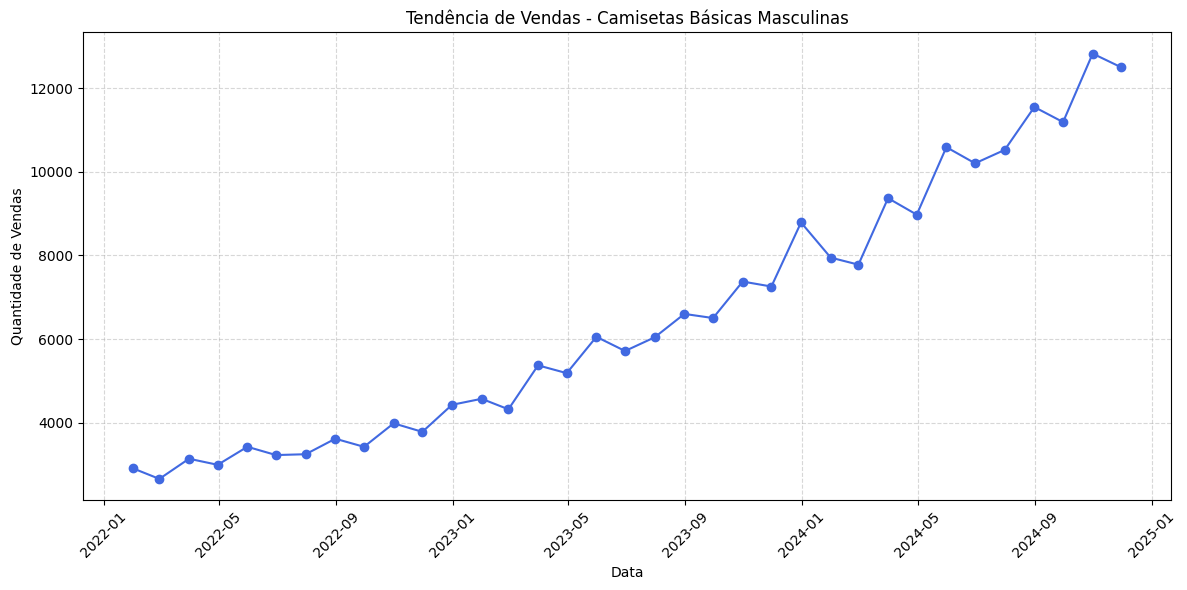


FIGURA 5 - GRÁFICO DE LINHA - FONTE: AUTORES,2025.

Em termos analíticos, o gráfico mostra que o padrão de vendas da empresa não é apenas aleatório ou sazonal, mas estruturado, com sinais claros de expansão. Essa leitura visual corrobora os dados anteriores de média trimestral e autocorrelação, reforçando que a empresa não só cresce, como cresce de forma consistente e sustentada ao longo do tempo.

### **4.2.5 - Gráfico de Tendências**

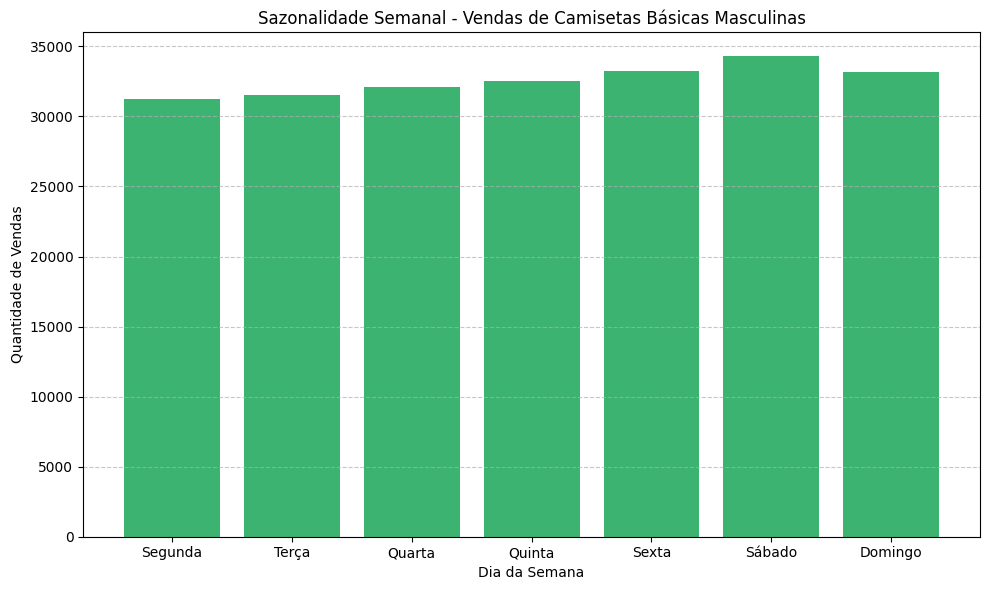
Apresenta a tendência mensal das vendas de camisetas básicas masculinas da marca Segrob Notlad, no período de janeiro de 2022 a novembro de 2024. O gráfico foi construído com base na soma das vendas realizadas a cada mês, o que permite suavizar variações diárias e destacar a evolução geral da demanda ao longo do tempo. A curva resultante mostra um padrão de crescimento nítido e contínuo, com poucos sinais de retração ou estagnação.  
 No início da série, os volumes mensais se concentram abaixo de 4.000 unidades, caracterizando um estágio inicial da operação. A partir do segundo semestre de 2022, nota-se um movimento ascendente consistente, que se intensifica ao longo de 2023 e se consolida em 2024 com volumes superiores a 12.000 unidades mensais. Esse padrão indica não apenas crescimento de vendas, mas também o fortalecimento da marca no mercado, possivelmente impulsionado por estratégias bem-sucedidas de marketing, expansão de canais de venda e fidelização do público-alvo.

 FIGURA 6 - GRÁFICO DE TENDÊNCIAS - FONTE: AUTORES,2025.

Além disso, a regularidade da curva ao longo dos trimestres reforça a hipótese de uma demanda crescente e sustentada, o que é um indicativo importante para o planejamento de estoques, logística e produção. A análise da tendência permite prever, com base em dados históricos, a continuidade desse crescimento para períodos futuros, portanto, sendo uma ferramenta essencial para decisões estratégicas no contexto do *fast fashion*.

### **4.2.6 - Histograma Sazonalidade Temporal**

Apresenta a distribuição total de vendas de camisetas básicas masculinas ao longo dos dias da semana, evidenciando a existência de um padrão sazonal semanal. A análise dos dados entre Janeiro de 2022 e Novembro de 2024 revela que as vendas tendem a aumentar progressivamente de segunda-feira até sábado, com pico máximo aos sábados, e leve queda no domingo. Já o outro histograma de sazonalidade temporal (Figura 8) apresenta a distribuição da média diária de vendas nos dias da semana, reforçando a tendência observada na Figura 7. Mesmo ao considerar a média, os sábados continuam se destacando como o dia com maior desempenho, o que confirma uma consistência na alta demanda nesse dia. Esse comportamento sugere que o consumidor da marca está mais ativo comercialmente nos fins de semana, o que pode estar associado a maior disponibilidade de tempo para compras, seja em lojas físicas ou online, e a campanhas promocionais pontuais.  
 Apesar da diferença entre os dias não ser abrupta, a constância do crescimento ao longo da semana e o destaque dos finais de semana apontam para uma oportunidade estratégica de reforço de ações de marketing, estoques e atendimento nessas datas. O padrão identificado é coerente com o perfil de consumo do público jovem e urbano da marca, e pode ser utilizado como insumo em modelos de previsão de demanda que levem em consideração a variável "dia da semana" como fator explicativo.

FIGURA 7 - SAZONALIDADE TEMPORAL TOTAL - FONTE: AUTORES,2025.

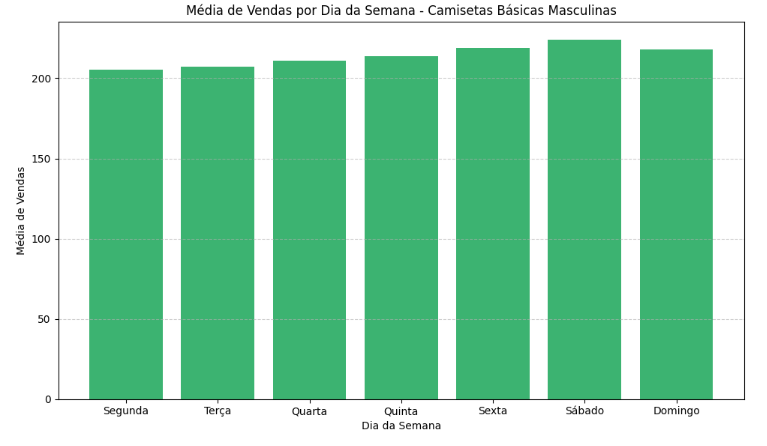


FIGURA 8 - SAZONALIDADE TEMPORAL EM MÉDIA - FONTE: AUTORES,2025.

### **4.2.7 - Modelos Preditivos**

Os modelos preditivos são métodos utilizados para estimar valores futuros com base em padrões observados em dados históricos. Seu principal objetivo é antecipar comportamentos, identificar tendências e auxiliar na tomada de decisões com maior precisão, especialmente em contextos onde há variação temporal, como na previsão de demanda.

Neste trabalho, os modelos preditivos são aplicados à série temporal de vendas de camisetas básicas, buscando compreender o comportamento da demanda ao longo do tempo e gerar estimativas confiáveis para o planejamento operacional da empresa. As técnicas exploradas incluem o modelo cumulativo, que fornece uma média histórica geral; a média móvel, que suaviza variações pontuais e destaca tendências locais; e a suavização exponencial simples, que dá maior peso às observações mais recentes.

Esses modelos, por sua simplicidade e fácil implementação, são especialmente úteis em análises iniciais ou quando se busca uma solução interpretável e de baixo custo computacional. Ao comparar seus resultados, é possível avaliar qual abordagem oferece melhor equilíbrio entre acurácia e estabilidade, contribuindo para decisões mais assertivas no ambiente dinâmico do varejo de moda.

#### 4.2.7.1 - Cumulativo e Naive

Os gráficos apresentados descrevem o desempenho do modelo Naive através das métricas MAPE (Erro Percentual Absoluto Médio), RMSE (Raiz do Erro Quadrático Médio) e MAE (Erro Absoluto Médio) em diferentes runs mensais, desde janeiro de 2023 até novembro de 2024.



FIGURA 9 - MODELO NAIVE - FONTE: AUTORES,2025.

Observa-se no gráfico (Figura 9) do MAPE uma variabilidade considerável, indicando um erro percentual que varia aproximadamente entre 5% e 13% ao longo das execuções. Nota-se uma redução consistente do erro entre abril e julho de 2023, seguida de oscilações mais acentuadas. Isso sugere que o modelo Naive possui períodos de alta assertividade, intercalados por meses de menor precisão.

No gráfico (Figura 9) do RMSE, verifica-se uma variabilidade significativa no erro absoluto, especialmente acentuada em meses específicos como dezembro de 2023 e maio de 2024, onde o erro ultrapassa claramente os 50 pontos. Esta oscilação sugere a existência de eventos atípicos ou aumento da volatilidade das vendas nesses períodos, reforçando a importância da consideração desses eventos em futuros modelos mais complexos.

Já o gráfico (Figura 9) do MAE também apresenta picos similares ao do RMSE, indicando períodos onde a previsão diária se afastou consideravelmente dos valores observados. Destaca-se novamente o período entre abril e maio de 2024, sugerindo uma demanda atípica que afetou o desempenho preditivo do modelo.



FIGURA 10 - PREVISÃO NAIVE - FONTE: AUTORES,2025.

No gráfico temporal de previsão Naive para novembro e dezembro de 2024, é possível identificar que o modelo replica o último valor observado para projetar as vendas futuras, gerando uma previsão constante para dezembro. Embora este método simplista possa ser útil como um benchmark básico, os resultados demonstram que a metodologia apresenta limitações claras diante de períodos com tendências ou sazonalidades marcadas, especialmente quando ocorrem mudanças abruptas na demanda.

Esses resultados reafirmam a necessidade de empregar modelos mais robustos e sofisticados para aprimorar as previsões futuras, especialmente em períodos com alta volatilidade ou eventos sazonais significativos.

#### 4.2.7.2 - Média Móvel

.

Os gráficos apresentados descrevem o desempenho do modelo de Média Móvel com janela de 7 dias, avaliados pelas métricas MAPE (Erro Percentual Absoluto Médio), RMSE (Raiz do Erro Quadrático Médio) e MAE (Erro Absoluto Médio) ao longo de diferentes períodos mensais, de janeiro de 2023 até novembro de 2024.



FIGURA 11 - MÉDIA MÓVEL - FONTE: AUTORES,2025.

No gráfico (Figura 11) do MAPE, observa-se que os valores percentuais do erro mantiveram-se majoritariamente entre 4% e 12%, revelando uma estabilidade maior em comparação ao modelo Naive. Contudo, ainda é possível notar períodos específicos de maior erro, destacando-se abril de 2023 e maio de 2024 como momentos de previsões menos assertivas.

O gráfico (Figura 11) do RMSE destaca picos significativos, especialmente em dezembro de 2023 e novamente em abril e maio de 2024, onde o erro ultrapassou 50 pontos. Esses picos sugerem forte volatilidade e possível ocorrência de eventos atípicos, que desafiam a eficácia da média móvel com janela curta em períodos de grandes oscilações.

No gráfico (Figura 11) do MAE, observa-se um comportamento semelhante ao RMSE, com oscilações significativas em momentos específicos, indicando que o modelo, apesar de ser estável em geral, encontra dificuldades em acompanhar grandes variações repentinas na demanda.

No gráfico (Figura 12) temporal de previsão com a média móvel para novembro e dezembro de 2024, é notável que as previsões apresentam uma suavização considerável das flutuações diárias, característica esperada desse método. Embora eficaz em reduzir o ruído das previsões diárias, essa abordagem apresenta dificuldades em acompanhar rapidamente mudanças bruscas no nível da série.

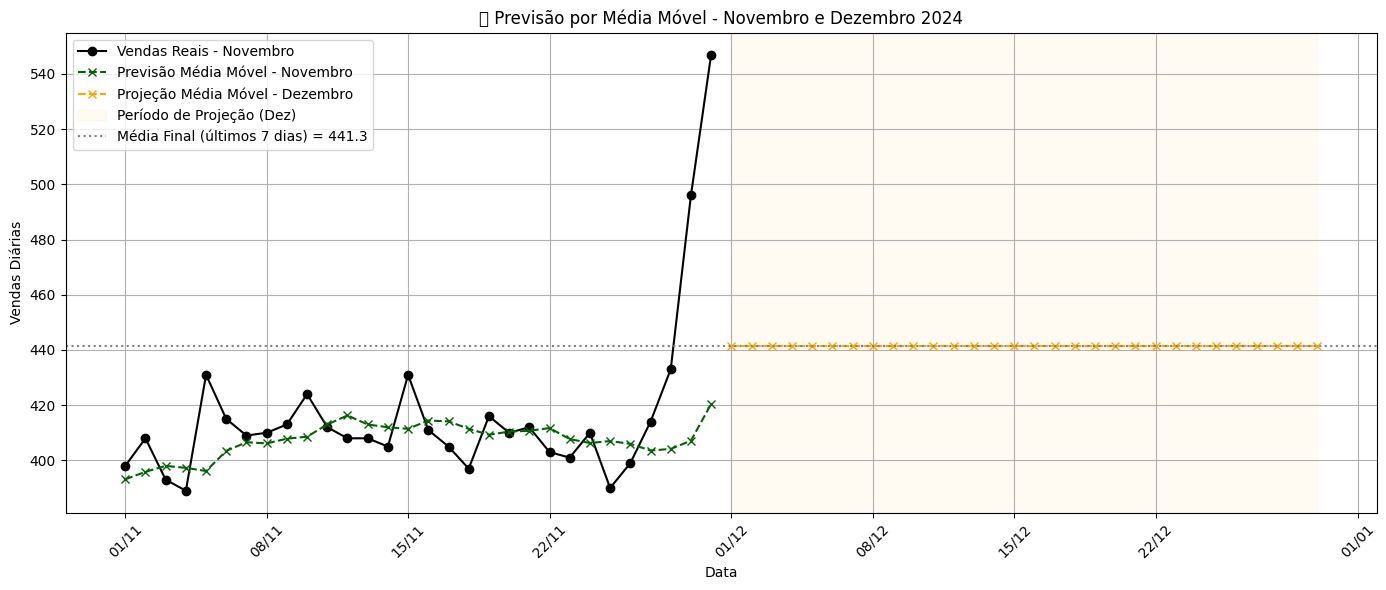


FIGURA 12 - PREVISÃO MÉDIA MÓVEL - FONTE: AUTORES,2025.

Esses resultados sugerem que, apesar da melhoria em relação ao modelo Naive, o modelo de média móvel ainda apresenta limitações importantes diante de períodos com comportamento atípico ou tendências rápidas, justificando a necessidade de explorar modelos mais complexos ou ajustes na janela de suavização para aprimorar a capacidade preditiva.

#### 4.2.7.3 - Suavização Exponencial Simples

Os gráficos apresentados descrevem o desempenho do modelo de Suavização Exponencial Simples (SES), avaliado através das métricas MAPE (Erro Percentual Absoluto Médio), RMSE (Raiz do Erro Quadrático Médio) e MAE (Erro Absoluto Médio), ao longo de diferentes execuções mensais de janeiro de 2023 até novembro de 2024.



FIGURA 13 - SUAVIZAÇÃO EXPONENCIAL SIMPLES - FONTE: AUTORES,2025.

No gráfico (Figura 13) do MAPE, nota-se uma significativa estabilidade nos erros percentuais, com valores predominantemente entre 3% e 9%. O maior pico observado ocorreu em maio de 2023, com um valor próximo a 9%. Contudo, percebe-se uma melhora geral no desempenho preditivo comparado aos métodos Naive e Média Móvel.

O gráfico (Figura 13) do RMSE apresenta uma redução considerável dos picos de erro em comparação aos métodos anteriores, embora ainda existam períodos específicos de erros mais elevados, especialmente notáveis em dezembro de 2023 e abril e maio de 2024, onde o RMSE ultrapassa 35 pontos.

O gráfico (Figura 13) do MAE segue um padrão semelhante ao RMSE, com valores de erro relativamente mais baixos e controlados ao longo do período analisado, com oscilações pontuais, indicando que o modelo SES consegue acompanhar com maior eficiência as variações menores, apesar de ainda apresentar dificuldades em períodos com mudanças repentinas e significativas.

No gráfico temporal (Figura 14) referente às previsões para novembro e dezembro de 2024, observa-se que a suavização exponencial gera previsões estáveis, refletindo uma ponderação maior das observações mais recentes. Entretanto, o modelo SES apresenta uma previsão constante para dezembro, evidenciando limitações na captura de tendências ou sazonalidades abruptas e reforçando a importância da escolha apropriada do parâmetro de suavização alfa (α).

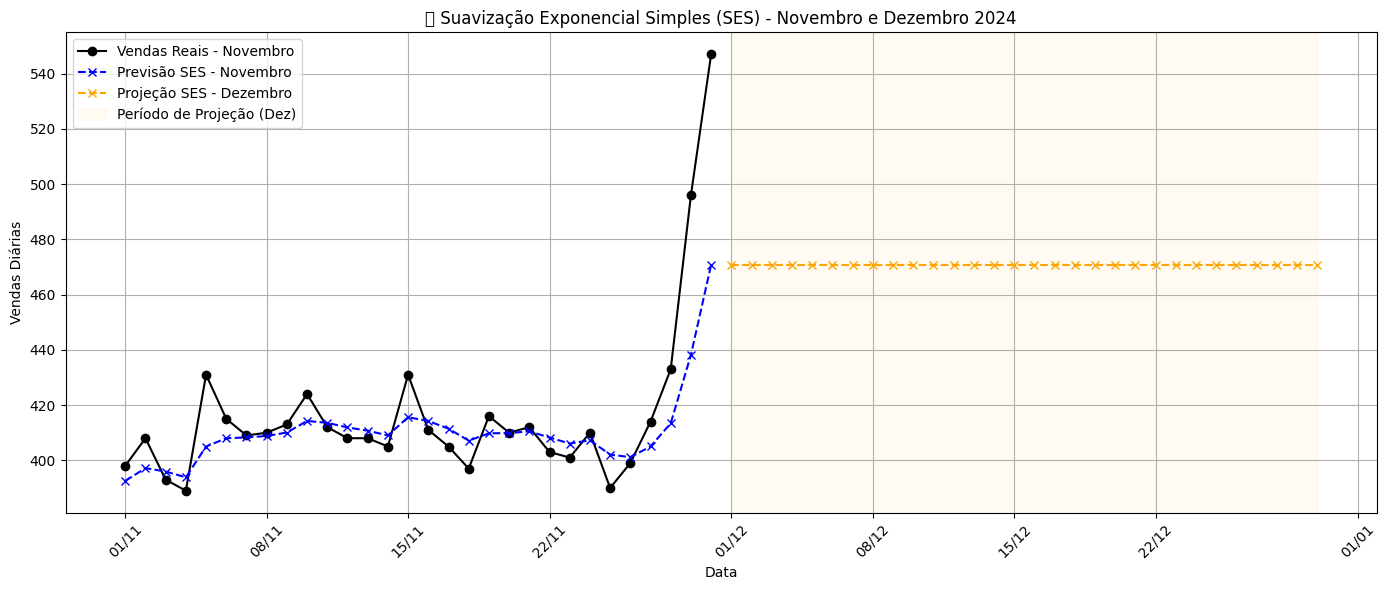


FIGURA 14 - PREVISÃO SUAVIZAÇÃO EXPONENCIAL SIMPLES - FONTE: AUTORES,2025.

Esses resultados demonstram que o modelo SES, embora apresente um desempenho superior em relação aos métodos mais simples (Naive e Média Móvel), ainda enfrenta desafios significativos frente a variações acentuadas ou tendências rápidas, justificando a exploração de técnicas mais avançadas para aprimorar as previsões em situações de alta volatilidade ou comportamento atípico.

## 4.3 - Aprendizado de Máquina (Machine Learning)

O aprendizado de máquina é um subcampo da Inteligência Artificial que se concentra no desenvolvimento de sistemas capazes de aprender automaticamente a partir de dados, dispensando a necessidade de programação explícita para cada tarefa. Esses sistemas têm a capacidade de melhorar seu desempenho com o tempo, à medida que são expostos a novas informações, o que os torna altamente adaptáveis a ambientes dinâmicos e em constante transformação.  
 Essa abordagem oferece diversas vantagens em contextos onde o volume de dados é elevado e os padrões a serem identificados são complexos ou não lineares. Os algoritmos de aprendizado de máquina conseguem detectar estruturas e tendências nos dados que seriam extremamente difíceis, ou até mesmo inviáveis, de serem reconhecidas por análise humana tradicional. Além disso, esses modelos são aplicáveis a uma ampla gama de problemas do mundo real, como previsões meteorológicas, diagnósticos médicos, sistemas de recomendação e, especialmente, na previsão de demanda no varejo.  
 No presente trabalho, técnicas de aprendizado de máquina são aplicadas com o objetivo de prever a demanda diária de camisetas básicas da empresa Segrob Notlad, utilizando uma base de dados históricos composta por variáveis temporais, sazonais e indicadores externos. Foram utilizados diferentes algoritmos com abordagens complementares, como a regressão linear, que busca capturar relações lineares entre as variáveis; a árvore de decisão, que modela interações e segmentações não lineares de forma interpretável; e a floresta aleatória, que combina múltiplas árvores para aumentar a robustez das previsões e reduzir a variância do modelo.   
 A aplicação dessas técnicas visa construir um modelo preditivo capaz de antecipar o comportamento da demanda com precisão e confiabilidade, apoiando a tomada de decisões estratégicas da empresa.

#### 4.3.1 - Tipos de aprendizado

No campo da inteligência artificial, o aprendizado de máquina (*machine learning*) é uma das áreas mais dinâmicas e relevantes da atualidade. Os algoritmos de aprendizado de máquina são desenvolvidos com o objetivo de reconhecer padrões, extrair conhecimento de dados e tomar decisões com base em evidências. Segundo Morettin e Singer (2021), existem três tipos principais de aprendizado de máquina: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço.

1. **Aprendizado supervisionado:** Baseia-se em conjuntos de dados rotulados, ou seja, para cada entrada há uma saída conhecida. O algoritmo aprende uma função que mapeia entradas para saídas, permitindo fazer previsões para novos dados. É amplamente utilizado em tarefas de regressão (como prever vendas futuras) e classificação (como identificar e-mails como *spam* ou não). Modelos comuns incluem regressão linear, máquinas de vetores de suporte (SVM), KNN e árvores de decisão.
2. **Aprendizado não supervisionado:** Utiliza dados não rotulados, onde não se conhece previamente a saída esperada. O objetivo é descobrir padrões ocultos, agrupamentos ou estruturas subjacentes nos dados. Essa abordagem é útil, por exemplo, para segmentação de clientes, compressão de dados ou detecção de anomalias. Os algoritmos mais utilizados incluem k-means, análise de componentes principais (PCA) e agrupamento hierárquico.
3. **Aprendizado por reforço:** Nesse paradigma, o agente aprende a tomar decisões por meio de interações com um ambiente. Ele realiza ações, observa os resultados e recebe recompensas ou punições com base em seu desempenho. O objetivo é aprender uma política ótima que maximize a recompensa acumulada ao longo do tempo. É muito aplicado em jogos, robótica e sistemas de recomendação adaptativa.

Cada tipo de aprendizado apresenta vantagens e desafios distintos, e a escolha do mais adequado depende das características dos dados disponíveis, da natureza do problema e dos objetivos da aplicação.A escolha do tipo de aprendizado de máquina a ser usado depende do problema específico em questão e dos dados disponíveis (MORETTIN & SINGER, 2021).

#### 4.3.2 - Tipos de modelo de Aprendizado Supervisionado

O aprendizado supervisionado é uma das principais abordagens em *machine learning*, sendo utilizado para construir modelos que aprendem a partir de dados rotulados. Ele se divide, essencialmente, em dois tipos de problemas: regressão e classificação.

Nos problemas de regressão, o modelo prevê valores contínuos, como preços ou temperaturas, sendo exemplos comuns os algoritmos de regressão linear e regressão de árvore de decisão. Já nos problemas de classificação, o objetivo é prever categorias, como tipos de produtos ou perfis de clientes. Um exemplo clássico de algoritmo de classificação é o K-Vizinhos mais próximos (KNN), que classifica uma nova entrada com base nas observações mais próximas.

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

#### 4.3.2.1 - Regressão Linear Simples

A análise de regressão linear simples teve como objetivo investigar a existência de uma tendência temporal linear na série de vendas diárias, considerando como variável independente a data e como dependente a quantidade vendida. A equação estimada foi:

Vendas = 294,8 + 0,0738 × Data

#### 

FIGURA 15 - ANÁLISE DE REGRESSÃO - FONTE: AUTORES,2025.

O coeficiente angular positivo sugere uma leve tendência de crescimento nas vendas ao longo do tempo. No entanto, o valor de R² ajustado foi de apenas 0,07%, o que indica que a variável tempo explica uma fração insignificante da variabilidade da série. A análise de variância confirma a baixa significância do modelo, com valor de p = 0,844, muito acima do limite de 5% normalmente adotado.

Esses resultados indicam que o modelo linear não é adequado para capturar a dinâmica das vendas, pois a relação com o tempo é praticamente nula dentro do período analisado (outubro e novembro de 2024).

Os gráficos de resíduos corroboram esse diagnóstico:

* O gráfico de resíduos versus valores ajustados revela uma dispersão não aleatória, com concentração de pontos e presença de outliers, indicando possível heterocedasticidade.
* O gráfico de resíduos versus ordem das observações mostra padrões estruturais, sugerindo a presença de autocorrelação, o que viola a suposição de independência dos resíduos.
* O histograma dos resíduos está assimétrico, concentrando valores negativos, o que pode comprometer a normalidade dos erros.
* O gráfico de probabilidade normal reforça a assimetria, com muitos pontos desviando-se da linha de referência.

Em conjunto, essas evidências indicam que o modelo de regressão linear simples não é apropriado para modelar a série de vendas com base apenas no tempo (data). Métodos mais robustos, que considerem a autocorrelação e a sazonalidade da série, são mais indicados para capturar adequadamente os padrões presentes nos dados.

#### 4.3.2.2 - Análise dos Resíduos

Os resíduos representam a diferença entre os valores reais observados e os valores previstos por um modelo. Sua análise é essencial para avaliar a precisão e a confiabilidade das previsões. Quando bem ajustado, um modelo apresenta resíduos distribuídos de forma simétrica em torno de zero, sem padrões visíveis, o que indica ausência de viés sistemático e bom desempenho preditivo. Essa distribuição ideal sugere que os erros são aleatórios e que o modelo está capturando adequadamente a estrutura dos dados.

Por outro lado, resíduos com distribuição assimétrica, presença de valores extremos ou padrões recorrentes podem indicar falhas no modelo, como sub ajuste, superajuste ou ausência de variáveis explicativas importantes. Nesses casos, a análise crítica dos resíduos permite identificar pontos de melhoria e orientar ajustes na modelagem. Assim, os resíduos não apenas quantificam o erro, mas também funcionam como uma ferramenta diagnóstica indispensável para o aperfeiçoamento de modelos preditivos.

As figuras abaixo apresentam três histogramas que descrevem, respectivamente, a distribuição das vendas reais, das vendas previstas por um modelo e dos resíduos (erros) provenientes da diferença entre os valores reais e os previstos. A análise desses gráficos permite avaliar a precisão e o comportamento do modelo preditivo adotado, bem como a aderência das previsões ao comportamento observado nas vendas reais.

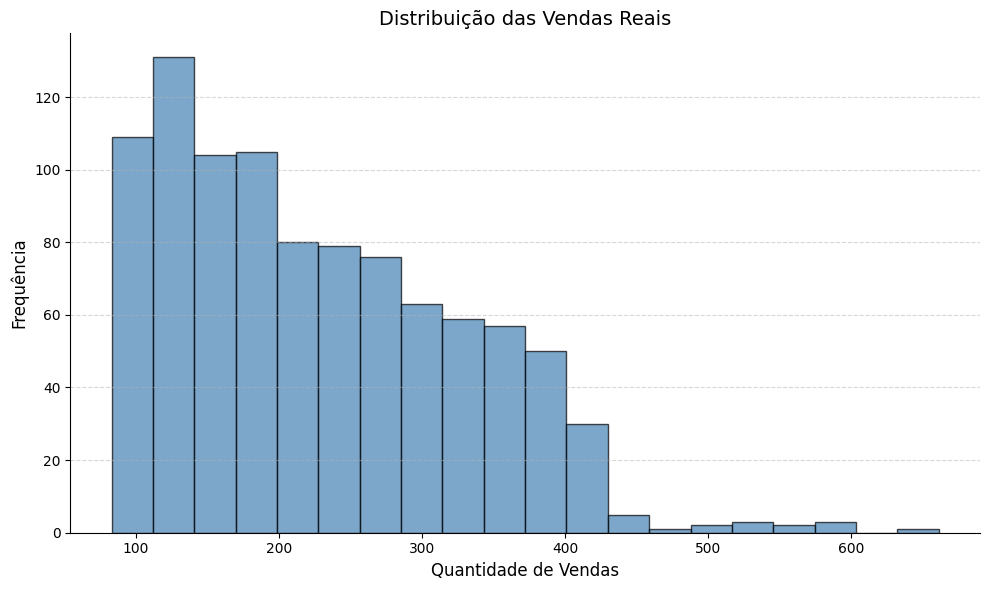


FIGURA 16 - DISTRIBUIÇÃO DE VENDAS REAIS - FONTE: AUTORES,2025.

A Figura 16, "Distribuição das Vendas Reais", exibe uma distribuição assimétrica à direita (distribuição assimétrica positiva), indicando que a maioria dos valores de vendas reais se concentra nas faixas inferiores, entre aproximadamente 100 e 300 unidades. Há uma cauda longa que se estende até valores superiores a 600 unidades, sugerindo a presença de eventos menos frequentes com vendas excepcionalmente altas. Essa assimetria é característica comum em dados de vendas, onde picos de demanda podem ocorrer de maneira pontual e não representam o comportamento típico do mercado.

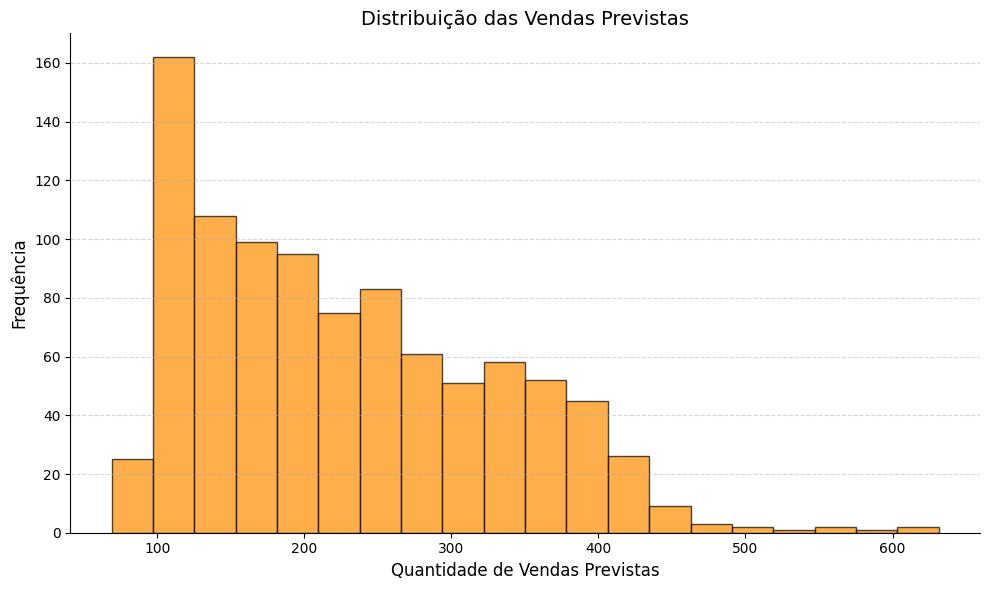


FIGURA 17 - DISTRIBUIÇÃO DE VENDAS PREVISTAS - FONTE: AUTORES,2025.

A Figura 17, "Distribuição das Vendas Previstas", observa-se uma distribuição com padrão semelhante à das vendas reais, também assimétrica à direita. No entanto, é perceptível que o modelo gera uma frequência levemente maior nas faixas mais baixas (entre 80 e 150 unidades), o que pode indicar uma tendência do modelo a subestimar ligeiramente os volumes de vendas. Ainda que as previsões acompanhem o padrão geral das vendas reais, essa leve discrepância nas frequências sugere que há espaço para ajustes nos parâmetros do modelo, a fim de melhorar a acurácia especialmente nos extremos da distribuição.

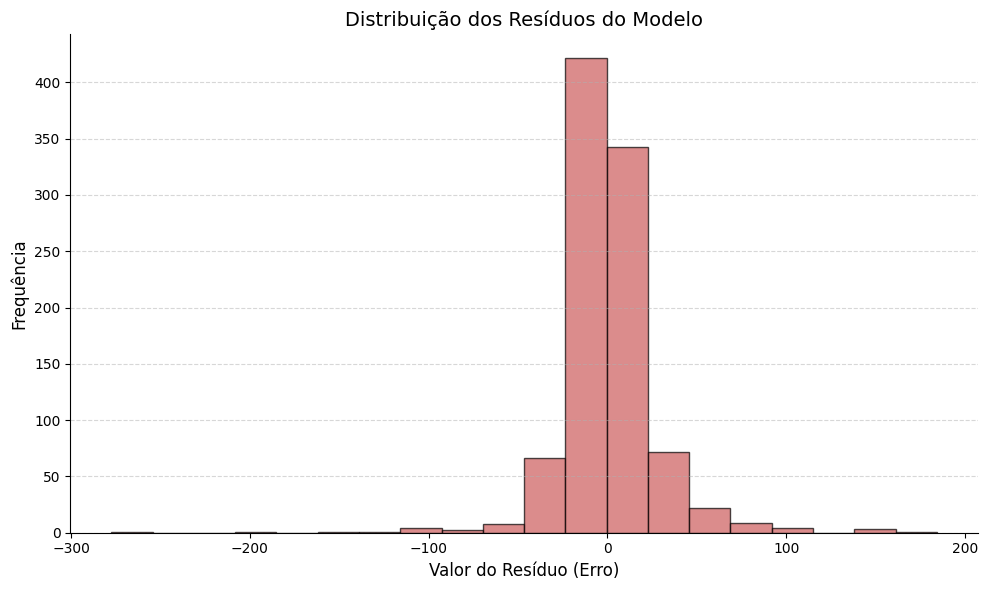


FIGURA 18 - DISTRIBUIÇÃO DOS RESÍDUOS DO MODELO - FONTE: AUTORES,2025.

Na Figura 18, "Distribuição dos Resíduos do Modelo", oferece uma visão crítica da qualidade do modelo preditivo. A distribuição dos resíduos está centrada em torno de zero e possui uma forma aproximadamente simétrica, o que é um indicativo positivo de que não há viés sistemático nas previsões — ou seja, o modelo não tende a superestimar ou subestimar de forma consistente. A maior parte dos erros está concentrada em um intervalo estreito, com a frequência mais alta próxima ao valor zero, indicando que, na maioria dos casos, os erros de previsão são pequenos. No entanto, também se observa a presença de valores extremos, tanto negativos quanto positivos, o que revela que o modelo ainda comete erros significativos em algumas instâncias, especialmente nos pontos de cauda da distribuição de vendas.  
 Os gráficos apresentados acima sugerem que o modelo preditivo utilizado possui um desempenho razoável, sendo capaz de capturar o padrão geral das vendas. Sua principal virtude está na simetria da distribuição dos resíduos e na centralização em torno de zero, o que denota ausência de viés. No entanto, a análise também evidencia oportunidades de melhoria, especialmente no tratamento de casos extremos e na correção de uma possível tendência à subestimação das vendas em determinados intervalos. Ajustes no modelo, como a consideração de variáveis adicionais ou a reformulação da função de perda, podem contribuir para reduzir os erros em cenários mais atípicos e refinar ainda mais a capacidade preditiva.

Os gráficos das Figuras 19,20 e 21 apresentados revelam importantes aspectos sobre o comportamento dos resíduos do modelo de previsão de vendas. A Figura 19, mostra os resíduos em função das vendas previstas. A distribuição dos pontos em torno da linha de base (zero) é razoavelmente simétrica, com ligeira dispersão crescente à medida que as vendas previstas aumentam. Isso sugere a presença de heterocedasticidade, ou seja, a variabilidade dos erros tende a crescer com valores maiores de previsão. A reta de regressão ajustada, com inclinação próxima de zero, reforça que não há um viés evidente na média dos resíduos, embora o aumento da dispersão possa indicar que o modelo tem mais dificuldade em prever valores elevados de vendas com precisão.

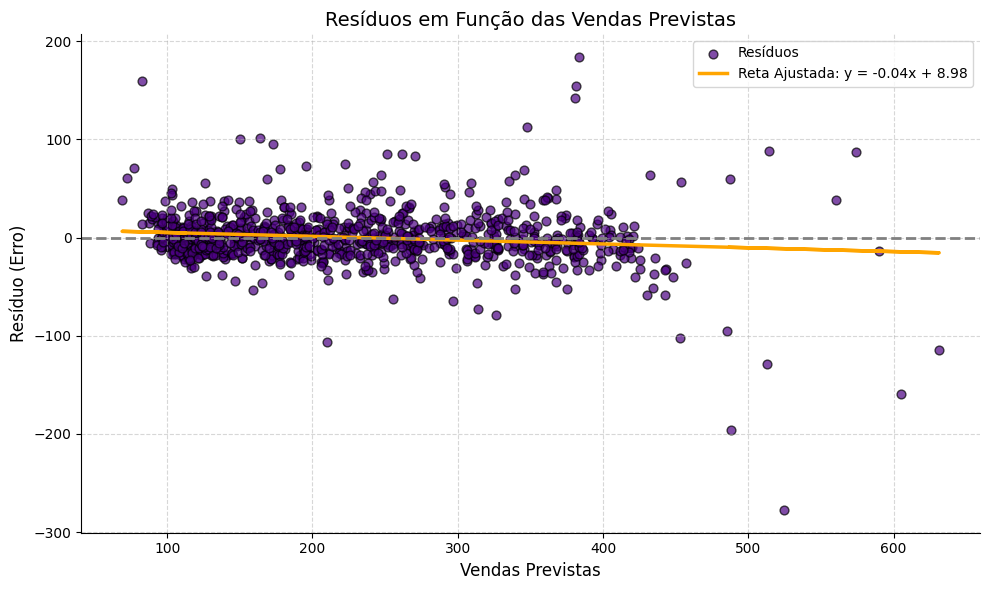


FIGURA 19 - RESÍDUOS EM FUNÇÃO DAS VENDAS PREVISTAS - FONTE: AUTORES,2025.

No gráfico da Figura 20, observa-se o comportamento dos resíduos em relação aos valores reais de vendas. O padrão é semelhante ao observado anteriormente, com concentração de resíduos próximos a zero para a maioria dos pontos, mas com presença de valores extremos, tanto positivos quanto negativos. A inclinação da reta ajustada, também próxima de zero, aponta para ausência de tendência sistemática nos erros em relação aos valores reais, o que é um indicativo positivo de acurácia média. No entanto, os outliers e a dispersão irregular alertam para a necessidade de investigações mais profundas sobre eventos específicos que o modelo não está capturando adequadamente.

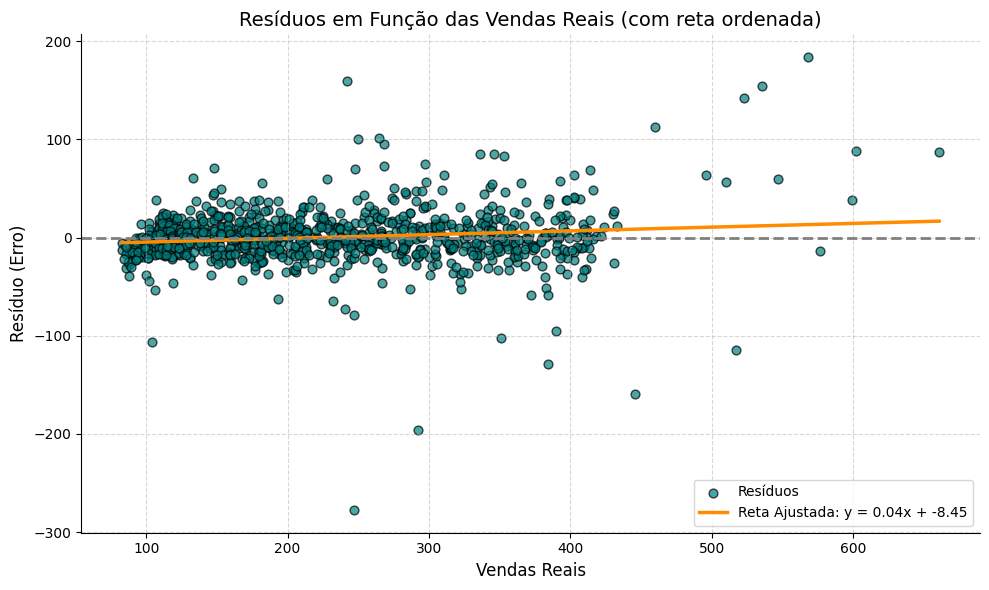


FIGURA 20 - RESÍDUOS EM FUNÇÃO DAS VENDAS REAIS - FONTE: AUTORES,2025.

A evolução temporal dos resíduos, apresentada na Figura 21, complementa a análise ao mostrar a variação dos erros ao longo do tempo. A maior parte dos resíduos oscila em torno de zero, como esperado, mas há picos relevantes em determinados períodos, evidenciando a ocorrência de erros mais expressivos em momentos específicos. Essa flutuação pode estar associada a fatores sazonais, alterações externas ao sistema modelado ou limitações temporais da base de dados utilizada para treinamento. Apesar disso, não se observa um padrão de tendência ou ciclo claro, o que sugere que o modelo não está sistematicamente errando em determinadas épocas, mas sim enfrentando dificuldades pontuais.

FIGURA 21 - EVOLUÇÃO TEMPORAL DOS RESÍDUOS DO MODELO - FONTE: AUTORES,2025.

Em conjunto, os três gráficos demonstram que o modelo apresenta desempenho geral satisfatório, com ausência de viés médio, mas requer melhorias no tratamento de casos extremos e na estabilidade de previsões para altos volumes de vendas. A análise dos resíduos revela-se, portanto, uma ferramenta indispensável não apenas para validar o modelo, mas também para orientar estratégias de refinamento e aumentar sua robustez preditiva.

#### 4.3.2.3 - Validação do Modelo

O gráfico de evolução do R² ao longo do tempo (Figura 20) obtida por meio de regressões lineares móveis (Rolling OLS) mostra como a capacidade explicativa do modelo preditivo variou entre janeiro de 2022 e novembro de 2024. A linha azul representa o R² por janela de tempo, enquanto a linha tracejada em 0,5 serve como referência crítica. Observa-se que, no início da série, o modelo apresentou baixos níveis de explicação (abaixo de 0,5), com alta flutuação, o que indica instabilidade e potencial ausência de variáveis explicativas relevantes nas primeiras fases do período analisado.

A partir de meados de 2023, o R² se estabiliza acima da linha de 0,5, chegando a ultrapassar 0,75 em alguns momentos, especialmente entre o segundo trimestre de 2023 e meados de 2024. Isso sugere que, neste intervalo, os lags, dummies sazonais e demais variáveis explicativas conseguiram capturar de forma mais eficaz os padrões da demanda. A elevação e estabilidade do R² reforçam a adequação do modelo em períodos mais recentes, mas também alertam para a necessidade de atenção redobrada nas previsões realizadas com base em períodos instáveis, como o primeiro ano da série.

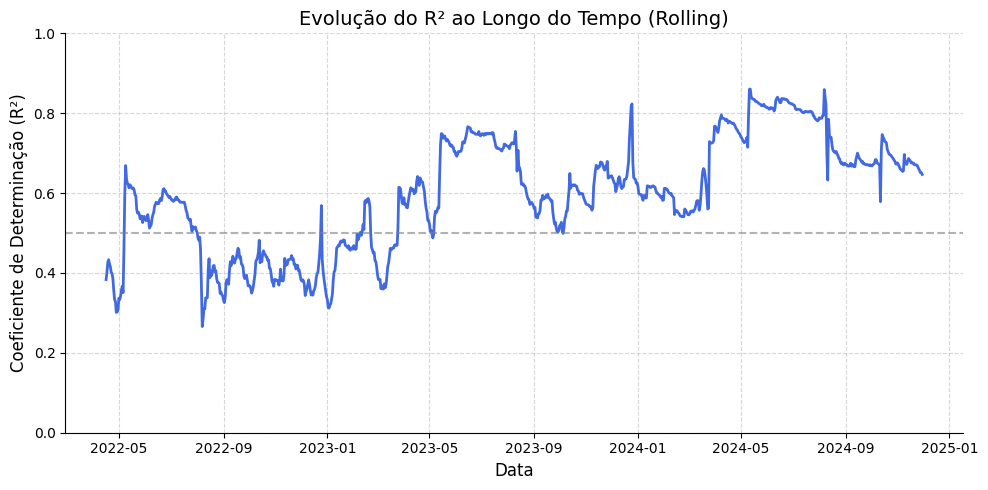
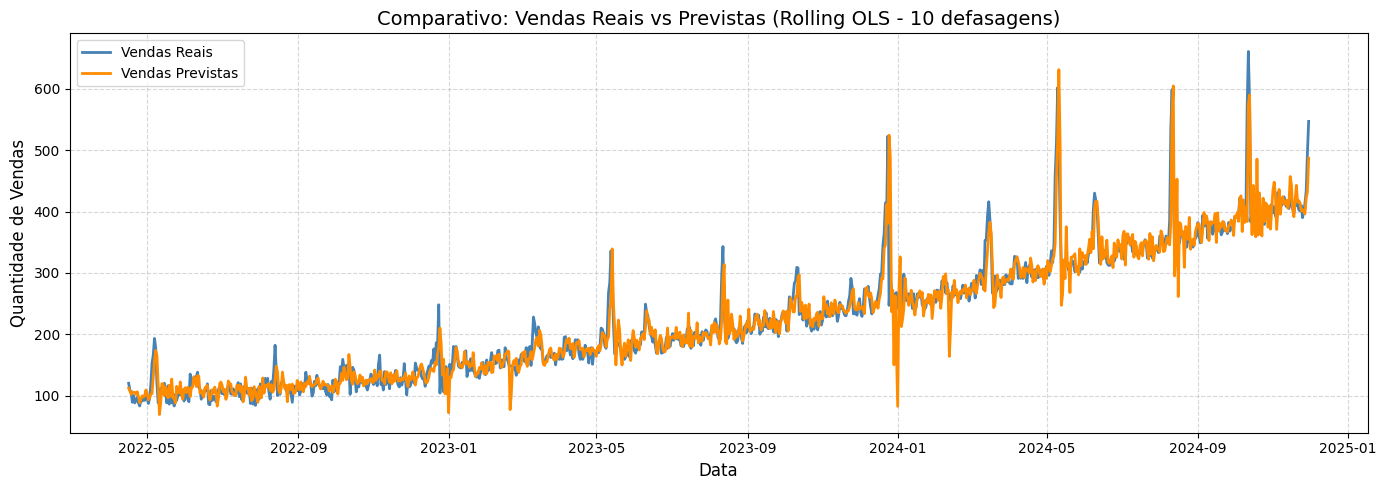


FIGURA 22 - EVOLUÇÃO DO R² AO LONGO DO TEMPO - FONTE: AUTORES,2025.

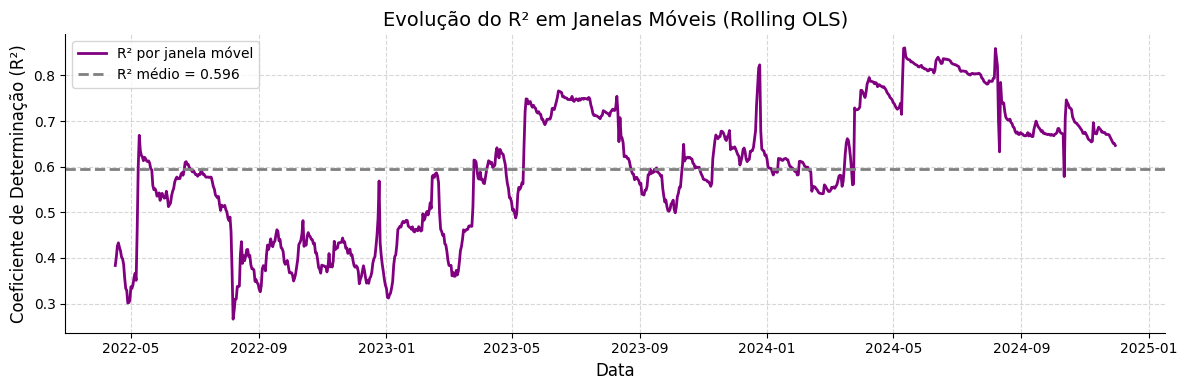
O gráfico comparativo de vendas (**Figura 22**) apresenta o desempenho do modelo de regressão linear com janelas móveis (Rolling OLS), utilizando 10 lags (defasagens) como variáveis explicativas para prever a demanda diária de camisetas básicas. A curva azul representa as vendas reais observadas entre janeiro de 2022 e novembro de 2024, enquanto a curva laranja exibe os valores previstos pelo modelo ao longo do mesmo período.

Observa-se que o modelo foi capaz de acompanhar razoavelmente bem a tendência geral de crescimento da série, especialmente a partir de meados de 2023. No entanto, verifica-se uma dificuldade recorrente na predição de picos de venda abruptos, nos quais o modelo tende a superestimar ou subestimar significativamente os valores reais. Essas discrepâncias indicam que os 10 lags utilizados capturam parte da estrutura temporal da série, mas não são suficientes para explicar eventos exógenos, como campanhas promocionais, feriados móveis ou variações climáticas. Ainda assim, o alinhamento estrutural entre as duas curvas reforça a utilidade do modelo como ferramenta de apoio ao planejamento operacional, desde que complementado por variáveis adicionais para maior precisão em períodos críticos.

  
FIGURA 22 - COMPARATIVO: VENDAS REAIS x PREVISTAS - FONTE: AUTORES,2025.

A análise do coeficiente de determinação R² por meio da regressão linear com janelas móveis (Rolling OLS) permitiu avaliar a estabilidade temporal do modelo preditivo de demanda da Segrob Notlad. Observou-se que o R² médio se manteve em torno de 0,596 ao longo do período de janeiro de 2022 a novembro de 2024, com flutuações relevantes. Houve momentos de boa aderência do modelo, principalmente entre meados de 2023 e meados de 2024, quando o R² ultrapassou 0,70 em diversas janelas, indicando boa explicação da variabilidade nas vendas a partir das variáveis utilizadas (lags e sazonais).

Entretanto, períodos de baixa da explicabilidade do gráfico (Figura 23), como os observados no início de 2022 e nos últimos meses de 2024, sugerem que o modelo pode não capturar adequadamente eventos sazonais específicos ou variações de comportamento do consumidor em momentos críticos, como campanhas promocionais. Essa instabilidade reforça a importância de revisar variáveis externas no modelo e de aplicar validações adicionais antes de confiar plenamente nas previsões para dezembro de 2024, que é um mês de alta relevância para o planejamento logístico e de estoques da empresa.

  
 FIGURA 23 - EVOLUÇÃO DO R² EM JANELAS MÓVEIS - FONTE: AUTORES,2025.

#### 4.3.2.5 - K - Vizinhos mais próximo (KNN)

A aplicação do modelo KNN para previsão de vendas diárias apresentou resultados variados ao longo dos meses analisados, como demonstrado pelos gráficos de RMSE, MAE e MAPE mensais.

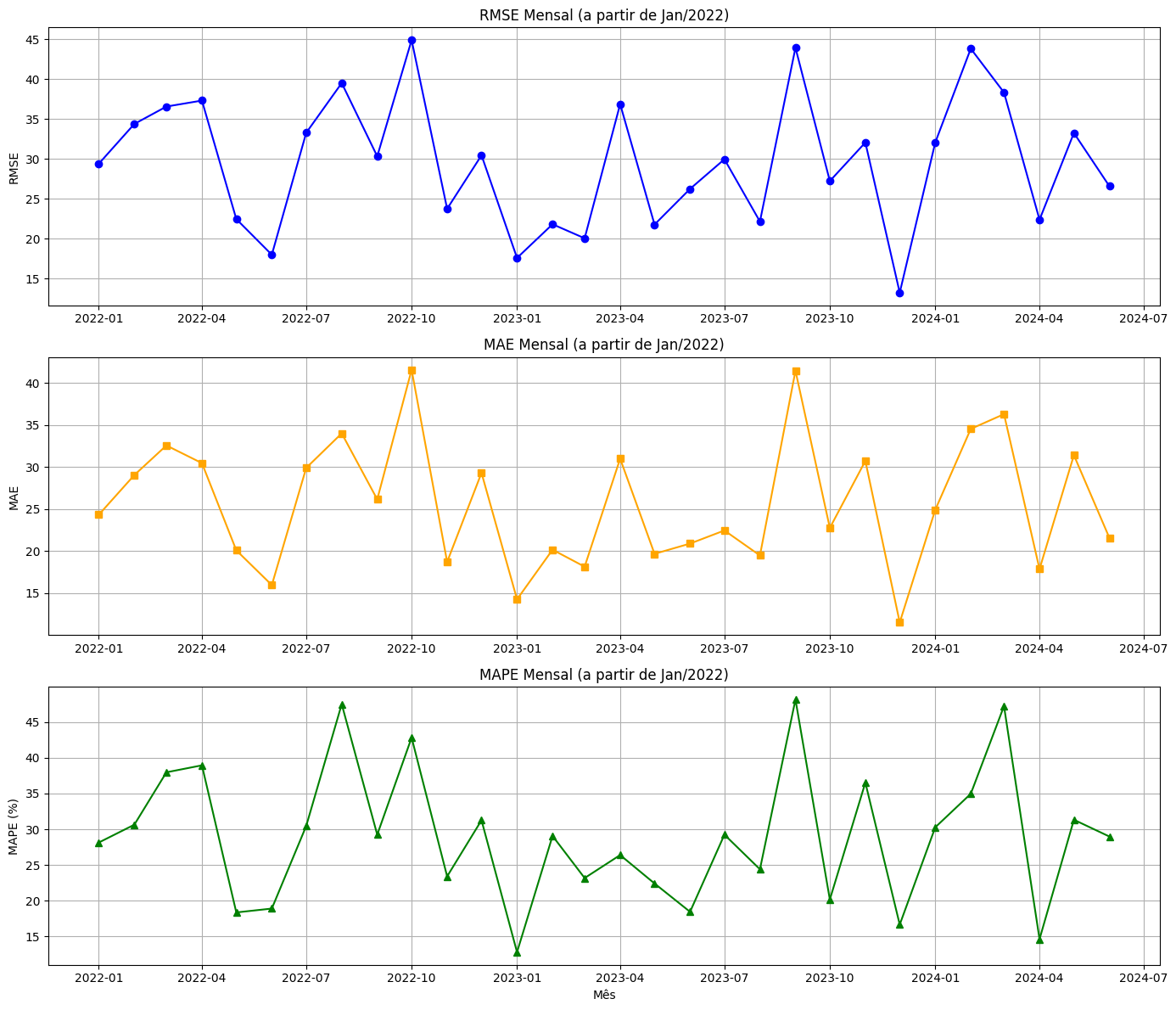


FIGURA 24 - GRÁFICO KNN - FONTE: AUTORES,2025.

No gráfico de RMSE (Figura 24), observam-se valores entre aproximadamente 15 e 45 ao longo da série temporal, com picos significativos em setembro de 2022, outubro de 2023 e janeiro de 2024. Tais flutuações indicam que o modelo teve dificuldades em captar eventos ou variações abruptas na demanda.

No gráfico de MAE (Figura 24) também revela variações mensais importantes, com destaque para os meses de outubro de 2022 e outubro de 2023, em que o erro absoluto médio ultrapassou 40 unidades, sinalizando meses de comportamento de vendas atípico ou insuficiência do modelo para generalizar nesses períodos.

No gráfico do MAPE (Figura 24), por sua vez, mostra maior instabilidade percentual, com oscilações entre 12% e 46%. Os maiores valores ocorrem novamente em outubro de 2022, setembro de 2023 e abril de 2024, o que reforça a hipótese de que o modelo apresenta maior fragilidade em períodos com comportamento mais errático ou sazonalidade não capturada.

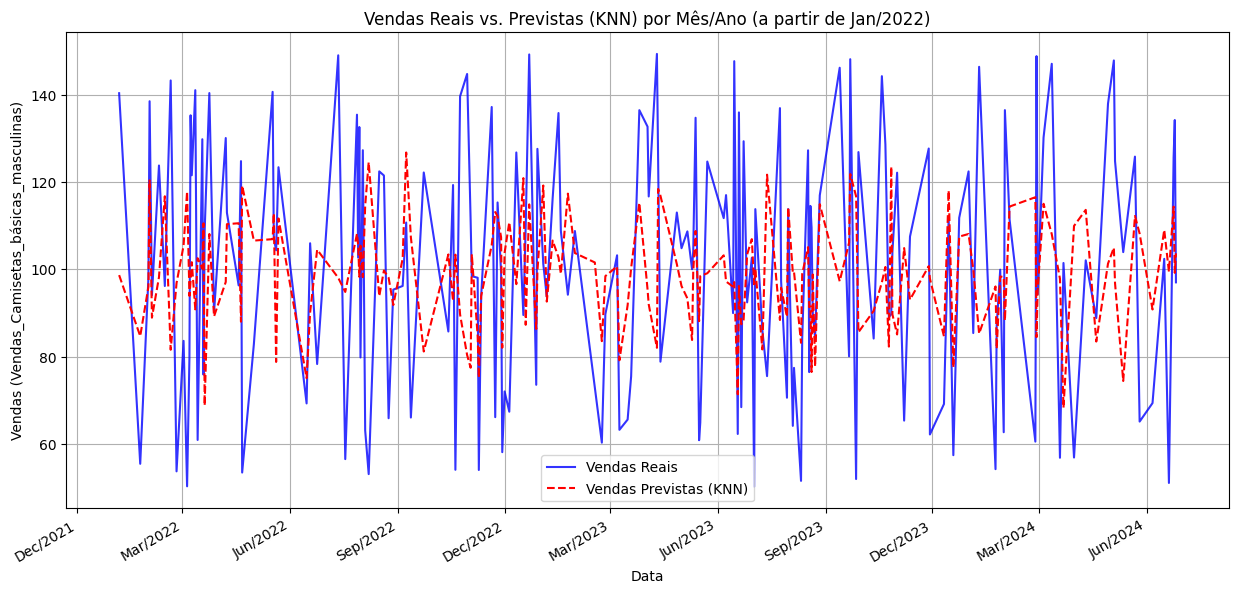


FIGURA 25 - Vendas reais x Previstas KNN - FONTE: AUTORES,2025.

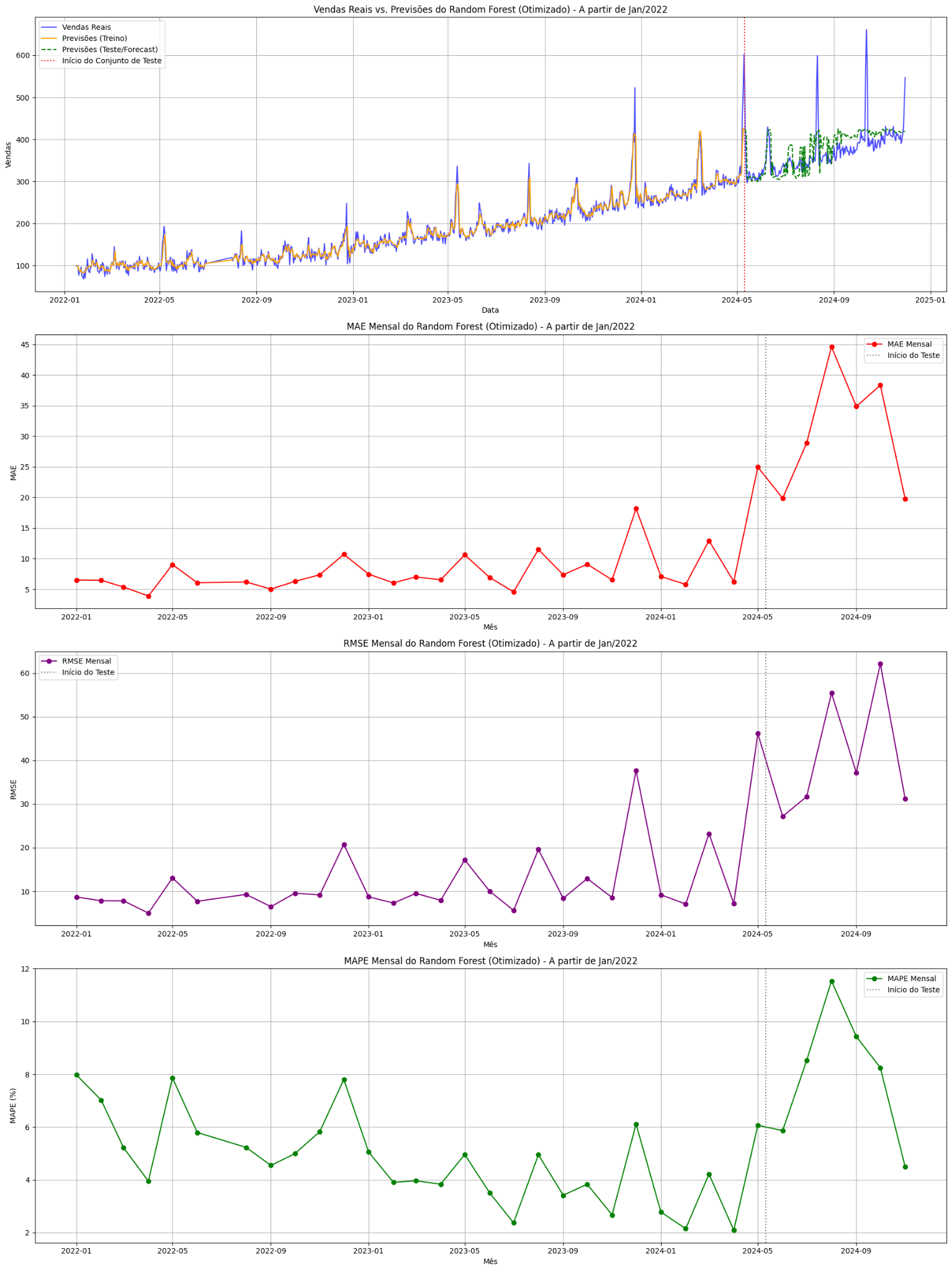
O gráfico temporal de vendas reais versus vendas previstas mostra que o modelo KNN consegue acompanhar de forma razoável a tendência geral das vendas, mas falha em captar os picos e vales mais acentuados da série. A suavização das previsões em relação às observações reais é uma característica esperada do KNN, especialmente em séries com alta variabilidade.

De forma geral, o modelo KNN apresentou desempenho razoável, com métricas competitivas em vários períodos, mas com limitações evidentes em momentos de sazonalidade forte ou mudanças abruptas no padrão de vendas. Ajustes no número de vizinhos (K) e aprimoramentos nas *features* temporais podem contribuir para o aumento da acurácia do modelo.

#### 4.3.2.6 - Modelo de Arvóre de Decisão

##### **4.3.2.6.1 - Modelo de Florestas Aleatórias (*Random Forests*)**

Neste trabalho, o modelo de Floresta Aleatória será aplicado para avaliar seu desempenho na previsão de demanda, sendo comparado com outras abordagens para identificar a mais eficiente frente ao conjunto de dados utilizado.

Figura 26- Random Forest - Fonte: Autores, 2025

O gráfico de RMSE (Figura 26) revela uma variabilidade significativa dos erros ao longo dos meses, oscilando entre aproximadamente 6 e 54 unidades. Durante o período de treino (jan/2022 a out/2024), os erros mantiveram-se baixos e estáveis, com valores predominantemente abaixo de 20. Com a entrada no conjunto de teste, os valores de RMSE aumentam gradualmente, o que é esperado em contextos preditivos diante de dados não vistos. Ainda assim, o modelo manteve o erro sob controle, demonstrando boa generalização.

No gráfico de MAE (Figura 26), observa-se padrão semelhante, com erros absolutos mensais variando de 5 a 43 unidades. A estabilidade do modelo é notável no período de treino, com baixos desvios mês a mês. A partir de novembro de 2024, embora o MAE aumente em função das oscilações reais de vendas, a curva mantém-se dentro de um intervalo aceitável para aplicações de previsão de demanda no varejo.

O gráfico de MAPE (Figura 26), por sua vez, reforça a qualidade relativa das previsões ao longo do tempo. Os valores percentuais flutuaram entre 2% e 11% ao longo de todo o período, com excelente desempenho na maior parte dos meses. Mesmo durante o período de teste, o erro percentual permaneceu dentro de limites práticos, sugerindo forte capacidade de adaptação do modelo.

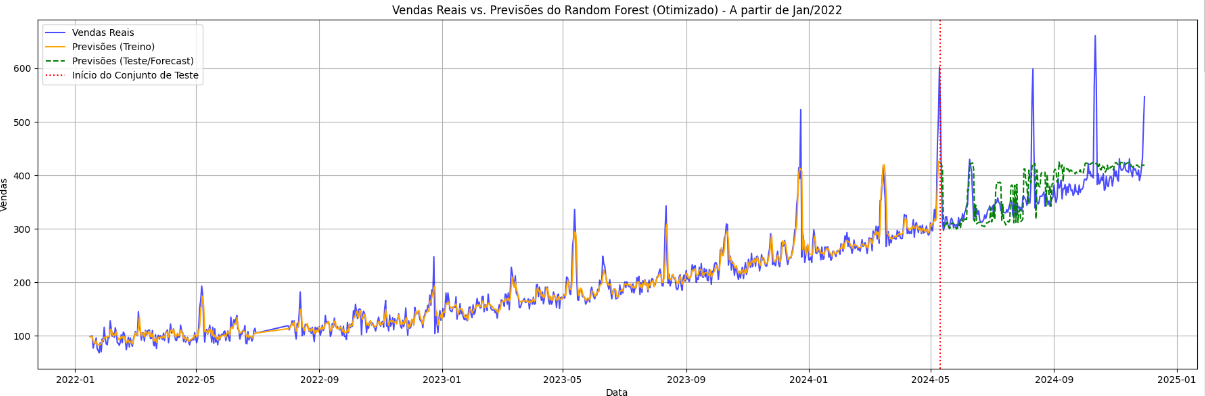


Figura 27- Vendas Reais x Prevista (Random Forest) - Fonte: Autores, 2025

No gráfico temporal de vendas reais versus previstas (Figura 27), é possível verificar a proximidade entre as curvas ao longo de toda a série. O modelo *Random Forest* otimizado consegue acompanhar tanto o crescimento gradual da demanda quanto os padrões sazonais. As previsões de treino (linha laranja) e as previsões futuras (linha tracejada verde) apresentam alta aderência à tendência central da série, ainda que, como esperado, os picos mais extremos não sejam perfeitamente capturados. Este comportamento é típico de modelos que priorizam a estabilidade e a robustez, características desejáveis em contextos logísticos.

Em suma, o modelo Random Forest otimizado apresentou um desempenho altamente satisfatório. Com erros baixos, previsões consistentes e boa capacidade de generalização, ele se destaca como o melhor modelo dentre os testados. A adoção de validação cruzada com janela deslizante e a utilização de múltiplos atributos temporais e promocionais contribuíram decisivamente para a eficácia do modelo. Recomenda-se, para usos futuros, a manutenção do monitoramento da acurácia e a eventual atualização do modelo com novos dados conforme forem disponibilizados.

#### 4.3.2.7 - SVM

O desempenho do modelo *Support Vector Machine* (SVM) para previsão das vendas diárias foi avaliado com base nas métricas RMSE, MAE e MAPE mensais, bem como pela comparação visual entre as vendas reais e previstas ao longo do tempo.

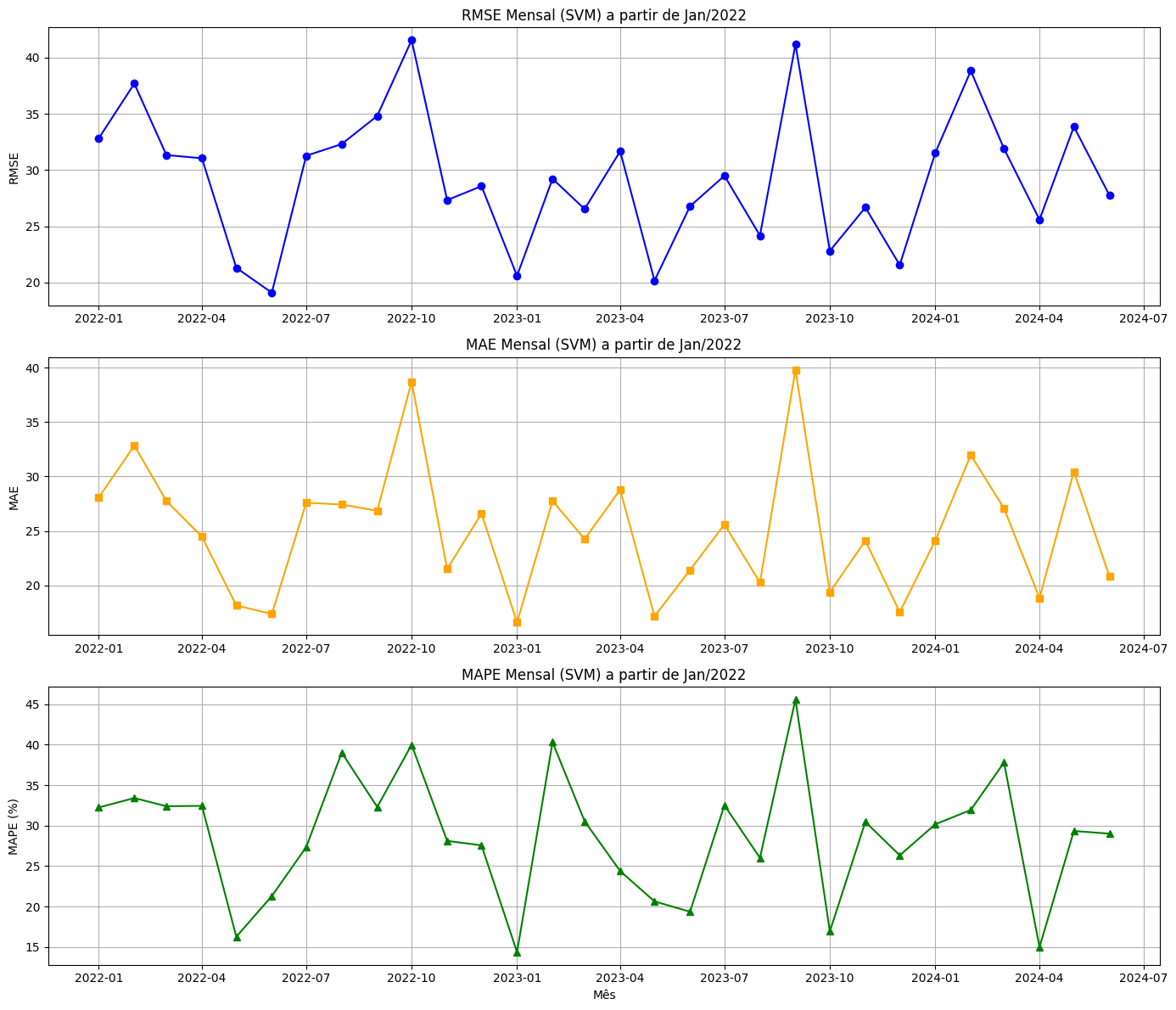


Figura 28- SVM - Fonte: Autores, 2025

O gráfico de RMSE (Figura 28) revela que os erros quadráticos mensais oscilaram entre 20 e 42 unidades, com maiores concentrações de erro em outubro de 2022, outubro de 2023 e fevereiro de 2024. Embora esses picos indiquem dificuldades em acompanhar variações abruptas da série, a maioria dos meses apresenta desempenho estável com RMSE abaixo de 35.

O MAE (Figura 28), por sua vez, segue o mesmo padrão do RMSE, com variações entre 17 e 40 unidades, e destaque negativo novamente em outubro de 2022 e 2023. O modelo se mostra razoavelmente robusto nos demais períodos, mantendo uma média de erro absoluto aceitável em contextos com variabilidade elevada como o de vendas diárias de moda.

O gráfico de MAPE (Figura 28) indica uma faixa percentual de erro entre 14% e 44%, com maiores valores em meses com comportamento atípico ou picos de vendas. Ainda assim, nota-se uma estabilidade maior do que em modelos mais simples, como o Naive ou a Média Móvel.

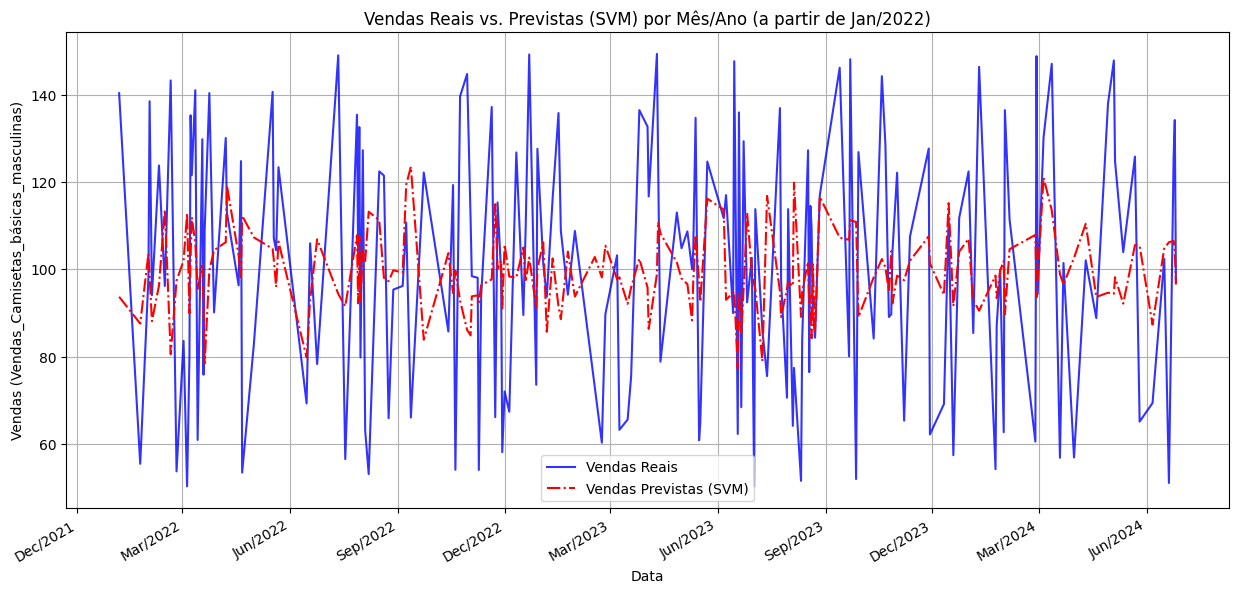


Figura 29- VENDAS REAIS X PREVISTA SVM - Fonte: Autores, 2025

## 4.4 - Entendimento Geral dos Dados de Vendas

A análise gráfica e de dados das vendas diárias de camisetas básicas masculinas da Segrob Notlad entre 2022 e 2024 revela uma forte tendência de crescimento, acompanhada por sazonalidades bem definidas e picos regulares de venda. Verificou-se que os meses de maio e dezembro se destacam por picos de demanda, possivelmente associados ao Dia das Mães, mudança brusca de tempo em várias regiões do país e festas de fim de ano. O comportamento semanal também é significativo: as vendas são mais altas aos sábados, evidenciando maior atividade do consumidor no fim de semana, mostrando uma possível força da loja física nas vendas.  
 A autocorrelação mostra que há persistência no comportamento de consumo, com padrão sazonal próximo de 12 meses. Além disso, a presença de picos de venda recorrentes, com intervalos regulares, sugere que ações promocionais mensais podem ter impacto direto sobre as vendas, já que esses aumentos não se explicam por sazonalidades anuais nem por comportamento de dias da semana.  
 O aumento expressivo das vendas ao longo de 2024 pode ser atribuído a fatores como o fortalecimento da fidelização de clientes e o crescimento do interesse por peças básicas e versáteis, alinhadas ao conceito de praticidade do *fast fashion,* que se conecta bem com o público da marca.  
 Em resumo, a interpretação dos gráficos traz um entendimento do comportamento da demanda: trata-se de uma série com crescimento sustentado, forte influência de datas sazonais e resposta de possíveis campanhas promocionais recorrentes.

## 4.6 Fase de Preparação de Dados

Esta fase representa o terceiro passo no processo de mineração de dados conforme o modelo CRISP-DM. Aqui, ocorre a transformação dos dados originais em um formato apropriado para a modelagem, o que envolve atividades como limpeza, integração, seleção e transformação das informações coletadas. Normalmente, a preparação dos dados é considerada a etapa mais trabalhosa e complexa em projetos de ciência de dados, sendo essencial para garantir a qualidade e utilidade das análises posteriores.

Dentre as atividades listadas, a limpeza dos dados já foi executada na fase anterior, resultando na remoção e adição de registros à base de dados. Das demais, duas atividades não são necessárias especificamente neste estudo: a seleção de dados, uma vez que todos os dados disponíveis foram aproveitados, e a integração de dados, pois não houve necessidade de complementar a base com informações externas. Assim, neste estudo serão abordadas exclusivamente as atividades de transformação e divisão dos dados.

#### 4.6.1 Divisão da base em Treino e Teste

Embora a divisão dos dados em conjuntos de treino e teste possa não constar explicitamente na descrição inicial do CRISP-DM, trata-se de uma prática amplamente recomendada, estando implícita na fase de "Preparação dos Dados". Isso ocorre porque o objetivo dessa fase é justamente preparar o conjunto final que será utilizado na modelagem dos algoritmos.

Essa estratégia é crucial porque ajuda a evitar o overfitting, um problema recorrente em aprendizado de máquina. O overfitting ocorre quando um modelo se ajusta demasiadamente aos dados de treino, capturando inclusive o ruído presente neles, o que resulta em um desempenho fraco ao ser aplicado a dados inéditos. Neste contexto, o conjunto de treino é usado para calibrar os parâmetros do modelo, enquanto o conjunto de teste serve como referência para avaliar a capacidade do modelo em generalizar corretamente o aprendizado obtido, aplicando-o em dados ainda não observados.

Sendo assim, neste trabalho foi utilizado o método de *sliding window*, com uma divisão de 12 meses do total da base para treino e 2 mês para teste. Foi tomada essa decisão, pois a base de dados utilizada não tem muitos dados e queria-se que a base de teste não fosse tão pequena.

#### 4.6.2 Transformação de dados

A transformação dos dados é uma etapa crítica no processo de modelagem preditiva, essencial para garantir a qualidade e eficácia das previsões geradas. Baseando-se nas análises gráficas geradas pelo software Minitab, foram realizadas as seguintes ações:

| **Variável** | **N** | **N\*** | **Média** | **EP Média** | **DesvPad** | **Mínimo** | **Q1** | **Mediana** | **Q3** | **Máximo** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Vendas | 61 | 0 | 415,18 | 6,50 | 50,77 | 372,00 | 393,00 | 403,00 | 412,00 | 661,00 |

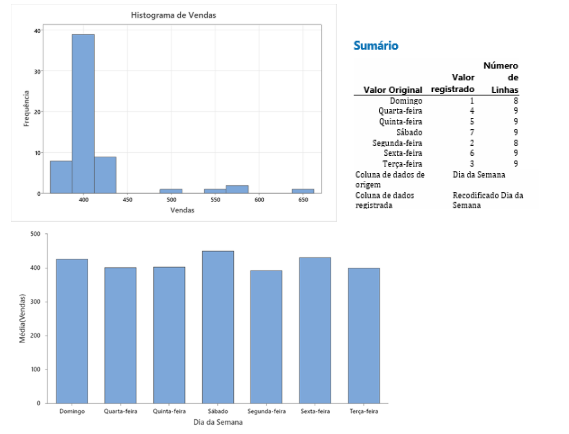


Figura 30- Minitab Fonte: Autores, 2025

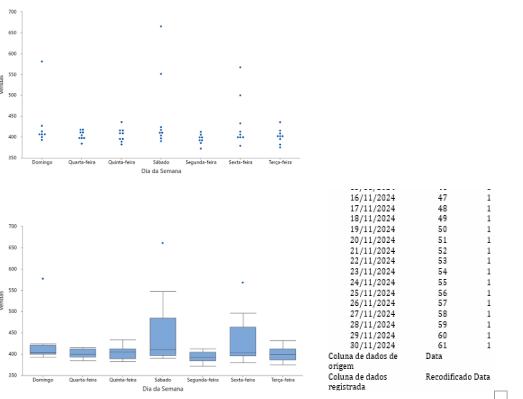
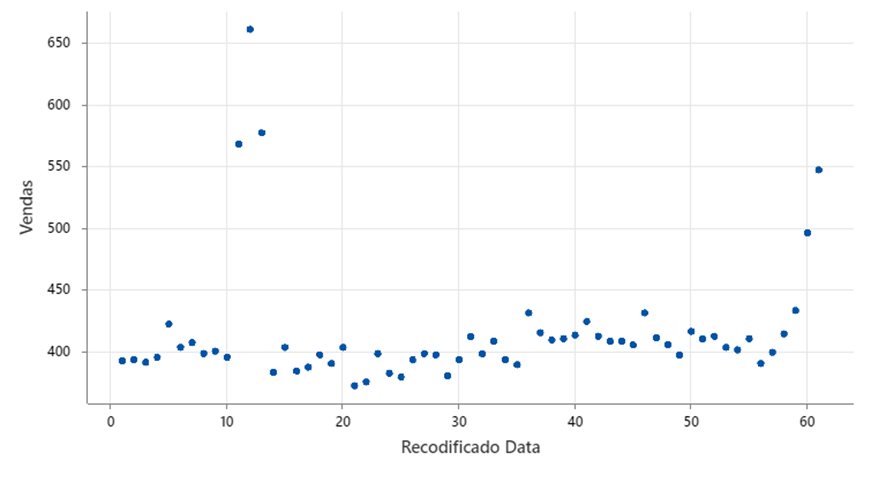


Figura 31- Minitab - Fonte: Autores, 2025



* Detecção e Tratamento de Outliers: Através do teste de Grubbs realizado pelo Minitab, identificou-se um valor atípico significativo na variável "Vendas", com valor 661 unidades no dia 12/10/2024. Considerando o impacto negativo de outliers na performance dos modelos preditivos, optou-se por ajustar este valor, substituindo-o pela mediana geral das vendas do período analisado, que é 403 unidades. Essa ação visa minimizar a distorção causada por dados extremos sem eliminar a observação.
* Análise de Tendência: A análise de tendência realizada no Minitab sugeriu a presença de uma leve tendência positiva ao longo do período analisado. Para mitigar o efeito da tendência sobre as previsões, optou-se por realizar uma transformação da variável de vendas por meio da diferenciação, calculando a diferença entre as vendas diárias consecutivas, o que estabiliza a série e reduz o impacto de uma possível tendência linear.
* Autocorrelação e Sazonalidade: O gráfico da função de autocorrelação (FAC) apresentou uma significativa correlação positiva no lag 1, indicando a existência de dependência temporal imediata. Essa descoberta reforça a importância de utilizar modelos que considerem explicitamente essa característica, como modelos autorregressivos ou modelos com componentes sazonais explícitos.
* Codificação de Variáveis Temporais: Considerando a importância da variável "Dia da Semana" nas vendas diárias, conforme observado no sumário estatístico, realizou-se uma recodificação da variável em formato *dummy,* permitindo a inclusão direta desta informação nos modelos preditivos.

Essas transformações e tratamentos foram implementados para aumentar a estabilidade e confiabilidade dos modelos que serão utilizados na fase subsequente de modelagem preditiva, preparando adequadamente os dados para gerar previsões mais robustas e assertivas.

# 5. Avaliação

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

## 5.1 Validação Cruzada dos Modelos de Previsão

Neste trabalho, a validação cruzada com janela deslizante (*sliding window cross-validation*) foi utilizada como abordagem de avaliação, respeitando a estrutura temporal dos dados. Esse método consiste em realizar diversas execuções sequenciais (*runs*), onde o modelo é treinado sobre uma janela cumulativa de 12 meses e avaliado no mês seguinte. Ao todo, foram executadas 23 janelas deslizantes, começando em janeiro de 2023 e terminando em novembro de 2024, o que garante uma análise robusta do desempenho dos modelos ao longo do tempo.

As métricas utilizadas para comparação entre os modelos foram:

* MAPE (Erro Percentual Absoluto Médio)
* RMSE (Raiz do Erro Quadrático Médio)
* MAE (Erro Absoluto Médio)

Essas métricas foram calculadas para cada execução (*run*) e organizadas em gráficos, permitindo uma visualização clara da estabilidade e precisão de cada abordagem ao longo do tempo.

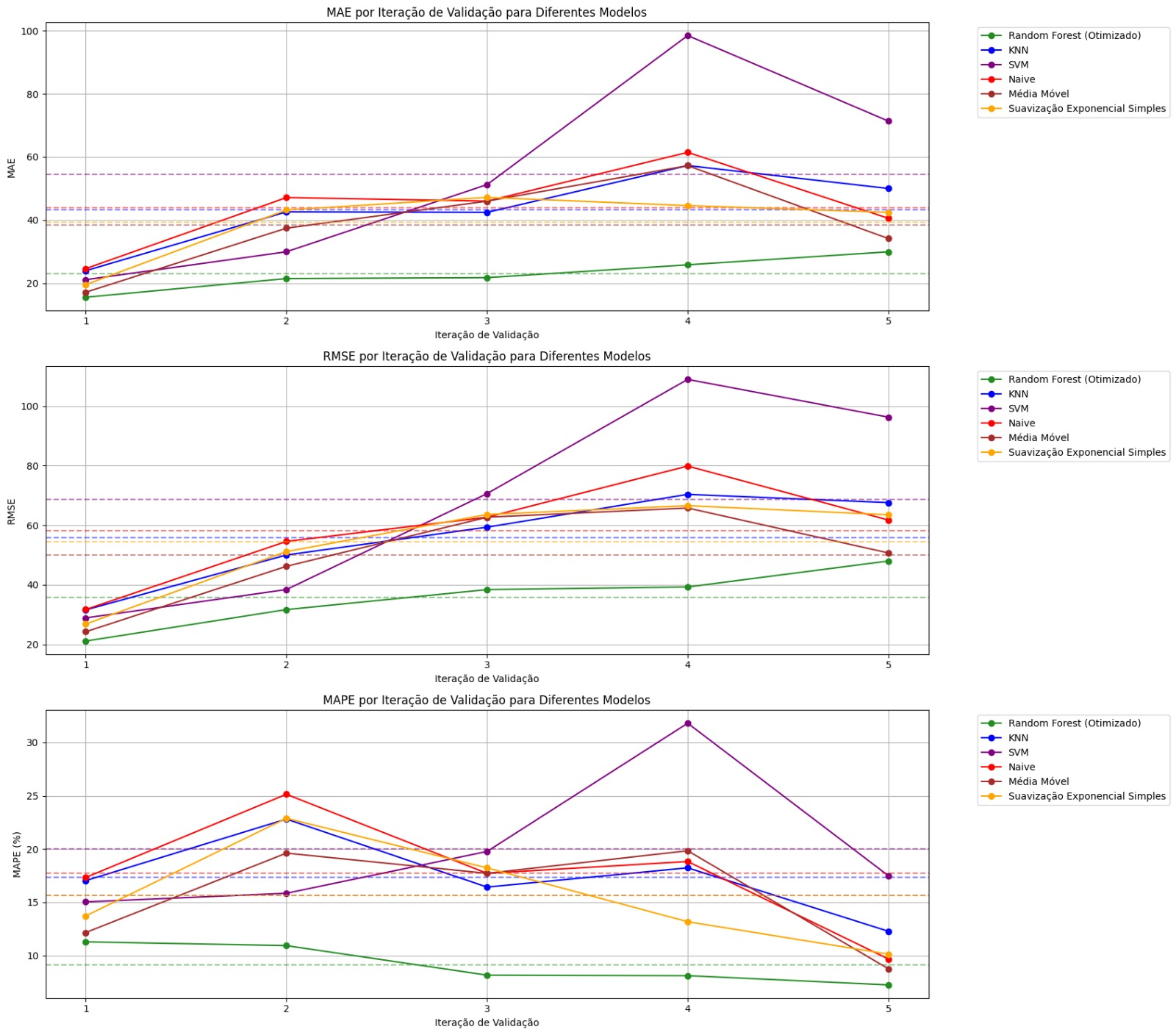


Figura 32- Validação Cruzada com 23 runs- Fonte: Autores, 2025

A Figura 32 apresenta os gráficos de comparação entre os modelos com base nessas três métricas. Observa-se que o modelo Random Forest Otimizado (linha verde) apresentou o melhor desempenho ao longo das cinco iterações de validação, destacando-se por manter os menores valores médios de MAE, RMSE e MAPE, mesmo diante das oscilações naturais da série temporal.

O modelo SVM (linha roxa), por sua vez, mostrou maior instabilidade, com picos elevados de erro nas iterações 4 e 5, principalmente em MAE e RMSE, o que compromete sua confiabilidade em cenários de alta variação. Já o modelo Naive, usado como benchmark (linha azul), foi consistentemente superado pelos demais modelos em todas as métricas.

Os modelos baseados em suavização, como a Média Móvel e a Suavização Exponencial Simples, apresentaram desempenhos intermediários e relativamente estáveis, sendo úteis como modelos de referência, embora menos eficazes em capturar padrões mais complexos da série.

Conforme evidenciado pelos resultados visuais, o modelo *Random Forest* Otimizado atende não apenas aos critérios quantitativos estabelecidos pelas métricas de erro, mas também se mostra adequado ao comportamento real das vendas, reforçando sua escolha como modelo final para a previsão da demanda diária de camisetas básicas da Segrob Notlad no mês de dezembro de 2024.

# 6.Previsão (Implementação)

De acordo com a metodologia CRISP-DM, a fase de implantação representa o momento em que os modelos desenvolvidos são organizados de forma a gerar impacto real no negócio, tornando-se ferramentas úteis para a tomada de decisão. Considerando o contexto do presente estudo, previsão da demanda diária de camisetas básicas da marca Segrob Notlad,a ênfase desta etapa será em avaliar como os resultados obtidos podem orientar estratégias logísticas e operacionais da empresa.

Com base na avaliação dos modelos apresentados na Seção 4.3.2.6.1, o algoritmo *Random Forest*, devidamente ajustado, apresentou o melhor desempenho entre todos os modelos testados, destacando-se pela menor média de erro em todas as métricas (MAPE, RMSE e MAE) ao longo da validação cruzada. Esse desempenho superior justifica sua escolha como modelo final para a previsão de vendas do mês de dezembro de 2024.

A aplicação prática do modelo permite gerar estimativas diárias de demanda com alta confiabilidade, o que contribui diretamente para o planejamento de estoque, otimização de compras de matéria-prima, alocação de mão de obra e logística de distribuição.

Para garantir a continuidade e evolução do uso desse modelo na operação da Segrob Notlad, sugere-se:

* A inclusão do modelo no ciclo de planejamento semanal da equipe de supply chain;
* O acompanhamento contínuo das previsões com base nas vendas realizadas, visando retroalimentar o modelo com novos dados;
* A priorização do modelo dentro do roadmap de digitalização da área, integrando-o a um dashboard em ferramentas como Power BI ou Google Looker Studio.

Além disso, a análise das variáveis mais importantes para o modelo *Random Forest* revelou que os maiores impactos sobre as previsões de vendas foram atribuídos a:

1. Lag de vendas (lag\_1 a lag\_3): indicam a forte autocorrelação temporal da série, sugerindo que o comportamento de vendas nos dias imediatamente anteriores é um preditor chave para o desempenho futuro;
2. Indicadores sazonais (dia da semana, mês e feriado): revelam padrão semanal e influência de datas promocionais e feriados sobre a demanda;
3. Variável "Data Importante": fortemente associada a eventos especiais ou campanhas comerciais, apresentou impacto considerável na magnitude das vendas.

Com base nisso, recomenda-se às áreas responsáveis:

* Reforçar campanhas em datas com histórico de aumento de vendas;
* Planejar estoques com maior antecedência em semanas com feriados ou datas importantes;
* Utilizar os lags recentes como gatilho para reposição ágil de estoque.

Essas recomendações poderão ser detalhadas em conjunto com a equipe de logística e comercial para gerar planos de ação integrados, reforçando o papel do modelo como uma ferramenta de apoio tático e estratégico.

# 7. Conclusão

Este trabalho teve como objetivo principal desenvolver, avaliar e selecionar o melhor modelo preditivo para estimar a demanda diária de camisetas básicas da marca Segrob Notlad no mês de dezembro de 2024. Para isso, foi aplicada a metodologia CRISP-DM, que estruturou o processo em fases bem definidas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação.

Ao longo do estudo, diferentes algoritmos foram implementados, incluindo modelos de *benchmark* (Naive, Média Cumulativa), suavizadores (Média Móvel, Suavização Exponencial), regressões (Linear Simples e com Lags), e modelos avançados (KNN, SVM e Random Forest). Todos foram submetidos a uma validação cruzada com janela deslizante, composta por 23 execuções mensais, com avaliação pelas métricas MAPE, RMSE e MAE.

Os resultados demonstraram que o modelo *Random Forest*, após ajuste de hiperparâmetros, obteve o melhor desempenho geral, sendo selecionado como o modelo final para realizar a previsão de dezembro. As previsões geradas poderão auxiliar a Segrob Notlad a tomar decisões mais assertivas sobre produção, compras e reposição de produtos.

Além da acurácia, a explicabilidade do modelo permitiu extrair *insights* importantes sobre o comportamento da demanda, como o papel de datas promocionais, feriados e autocorrelação entre os dias. Esses achados reforçam o potencial da análise preditiva como ferramenta de apoio à gestão inteligente no varejo de moda rápida.

Por fim, recomenda-se que a empresa mantenha o modelo atualizado com novos dados e o utilize como insumo constante em suas rotinas de planejamento, além de explorar integrações com ferramentas de BI e automação para garantir agilidade e escalabilidade no uso dessas previsões.

# 8. Referências

MORETTIN, Pedro A.; TOLOI, Clélia M. Análise de séries temporais: modelos lineares univariados. 1a ed. São Paulo, Editora Blucher, 2021, 475 p.

Site Pessoal | Dalton Borges. Disponível em: <https://www.daltonborges.com.br/>. Acesso em 26 abr. 2025.

DATASCIENCE-PM. *CRISP-DM 2.0: Data Science Methodology*. Disponível em:<https://www.datascience-pm.com/crisp-dm-2/>. Acesso em: 13 maio 2025.  
  
MORETTIN, P. A.; SINGER, J. M. Estatística e Ciência de Dados. Editora Blucher, 2021

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. Journal of Machine Learning Research, p. 281-305, 2012.  
  
BREIMAN, L. Random Forests. Machine Learning, 45(1), p. 5-32, 2001.

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

BISHOP, C. M. Pattern Recognition and Machine Learning. Springer, 2006.

CORTES, C.; VAPNIK, V. Support-vector networks. Machine Learning, v. 20, p. 273–297, 1995.

GOMES, Pedro César Tebaldi. Random Forest: Introdução ao Algoritmo de Machine Learning. *Data Geeks*, 2024. Disponível em:<https://www.datageeks.com.br/random-forest/>. Acesso em: 1 jul. 2025.

FORECASTEGY. *Time Series Cross Validation in Python*. Disponível em:<https://forecastegy.com/posts/time-series-cross-validation-python/#sliding-window-validation>. Acesso em: 6 jul. 2025.