UNIVERSIDADE FEDERAL FLUMINENSE

ENGENHARIA DE PRODUÇÃO

EMILLY CRISTINA FERREIRA NOGUEIRA

NICOLE DA SILVA FULGONI

ESTUDO DE CASO: SEGROB NOTLAD

RIO DAS OSTRAS
2025

SUMÁRIO

1. INTRODUÇÃO.....6
2. REFERENCIAL TEÓRICO.....6
2.1. Métricas e erros em Modelos preditivos.....6
2.1.1. Erro Percentual Absoluto Médio.....6
2.1.2. RMSE – Raiz do Erro Quadrático Médio.....7
2.1.3. MAD (Desvio Absoluto Médio).....8
2.1.4. Erro Padrão.....9
2.1.5. Comparação entre métricas.....10
2.2. Formas de validação cruzada.....10
2.2.1. Validação Cruzada K-fold.....11
2.2.2. Validação Cruzada Leave-One-Out (LOOCV).....11
2.2.3. Validação Cruzada Holdout.....11
2.2.4 Validação Cruzada Estratificada.....12
3. MÉTODO.....12
3.1. CRISP-DM.....12
3.1.1. Compreensão do Negócio.....13
3.1.2. Entendimento dos Dados.....13
3.1.3. Preparação dos Dados.....13
3.1.4. Modelagem.....13
3.1.5. Avaliação.....13
3.1.6. Implementação.....14
4. ESTUDO DE CASO.....14
4.1. Entendimento do Negócio.....14
4.2. Entendimento dos Dados.....16
4.2.1. Comportamento de Vendas ao Longo do Tempo.....16
4.2.2. Análise Exploratória.....17
4.3. Preparação dos dados.....23
4.3.1. Tratamento de Dados Faltantes.....23
4.3.2 Detecção e Tratamento de Outliers.....24
4.3.3 Criação e Enriquecimento de Variáveis.....24
4.3.4. Definição do Dataset Final para Modelagem.....25
4.3.5. Divisão dos Conjuntos de Treino e Teste.....25
4.3.6. Resultados da Preparação dos Dados.....26
4.4. Modelagem.....27
4.4.1. Modelo Naive.....27
4.4.2. Modelo Acumulativo.....27
4.4.3. Modelo Média Móvel Simples.....27
4.4.4. Resultado da Modelagem.....27
REFERÊNCIAS BIBLIOGRÁFICAS.....30

RESUMO

1. INTRODUÇÃO

2. REFERENCIAL TEÓRICO

2.1. Métricas e erros em Modelos preditivos

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

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

2.1.1. Erro Percentual Absoluto Médio

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

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

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

Onde:

$Y_i = Valores\ observados$

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

$n = Número\ total\ de\ observações$

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

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

2.1.2. RMSE – Raiz do Erro Quadrático Médio

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

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

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

Onde:

$Y_i = Valor\ real\ observado$

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

$n = Número\ total\ de\ observações$

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

2.1.3. MAD (Desvio Absoluto Médio)

O Desvio Absoluto Médio (MAD, *Mean Absolute Deviation*) é uma métrica frequentemente utilizada para avaliar a precisão das previsões em modelos preditivos. Diferente do RMSE, o MAD calcula o erro absoluto médio entre os valores reais e previstos, sem a penalização exagerada dos erros extremos, oferecendo uma visão equilibrada da precisão média do modelo (MORETTIN; TOLOI, 2018).

A fórmula matemática do MAD pode ser expressa por (SILVA, 2023):

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

Onde:

$Y_i = Valor\ real\ observado$

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

$n = Número\ total\ de\ observações$

Na prática, um MAD igual a 15 indica que, em média, as previsões feitas pelo modelo se desviam dos valores reais em 15 unidades. Por sua simplicidade, o MAD é uma métrica intuitiva, facilmente compreendida em contextos técnicos e gerenciais, sendo especialmente útil em situações onde os dados têm presença de valores extremos ou atípicos (outliers) (CAMPOS; SILVA, 2019).

Entre suas aplicações mais comuns, destacam-se a previsão de demanda, controle de qualidade e processos logísticos, áreas nas quais a robustez e a clareza na interpretação do erro são fundamentais para decisões precisas e eficientes (SILVA, 2023).

2.1.4. Erro Padrão

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

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

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

Onde:

$s = \ Desvio\ padrão\ das\ observações$

$n = \ Tamanho\ da\ amostra$

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

Entre suas principais aplicações estão a validação de modelos estatísticos e econométricos, análise financeira e pesquisas acadêmicas, contextos em que a precisão das estimativas é fundamental para conclusões seguras e decisões bem fundamentadas (MORETTIN; TOLOI, 2018).

2.1.5. Comparação entre métricas

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

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

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

2.2. Formas de validação cruzada

A validação cruzada (*cross-validation*) é uma técnica estatística utilizada para medir a capacidade dos modelos preditivos em realizar previsões confiáveis em novos dados, evitando o problema do superajuste (*overfitting*) (CAMPOS; SILVA, 2019). Por meio dela, os dados são divididos diversas vezes em subconjuntos diferentes de treinamento e teste, garantindo uma estimativa mais robusta da performance real do modelo.

Segundo Silva (2023), entre as formas mais comuns estão o método K-fold, Leave-One-Out, Holdout e a validação cruzada estratificada, cada uma adequada a

diferentes contextos. A escolha do método depende principalmente das características dos dados e dos objetivos específicos do projeto de análise preditiva.

2.2.1. Validação Cruzada K-fold

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

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

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

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

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

2.2.3. Validação Cruzada Holdout

A validação Holdout é um método simples para avaliar o desempenho de modelos preditivos. Consiste em dividir o conjunto de dados original em dois grupos:

treinamento (geralmente entre 70% e 80%) e teste (entre 20% e 30%), realizando uma única avaliação do modelo com essa divisão (CAMPOS; SILVA, 2019).

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

2.2.4 Validação Cruzada Estratificada

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

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

3. MÉTODO

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

3.1. CRISP-DM

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

Essa metodologia tem o objetivo de fornecer a qualquer pessoa ou empresa um modelo completo para realizar um processo de mineração de dados e pode ser dividida em seis fases: Compreensão do Negócio; Entendimento dos Dados; Preparação dos Dados; Modelagem; Avaliação e Implementação. Essas fases não

seguem uma sequência obrigatória, ou seja, pode-se avançar e retornar das fases quando for necessário (Shearer, 2000 apud. Ramos et al., 2020).

3.1.1. Compreensão do Negócio

Segundo Chapman et al. (2000 apud. LIMA, 2021), a fase de compreensão do negócio é considerada a etapa mais importante do projeto e tem o propósito definir os recursos, requisitos, critérios e objetivos do projeto, bem como o plano inicial para atingi-los e identificar as necessidades do cliente.

3.1.2. Entendimento dos Dados

No entendimento dos dados ocorre a coleta inicial dos dados que serão usados, a análise exploratória, validação e descrição desses dados (CHAPMAN et al., 2000 apud. LIMA, 2021).

3.1.3. Preparação dos Dados

A fase de preparação dos dados consiste na seleção do que realmente será usado como conjunto de dados e tratamento deles, como limpeza e transformação, caso haja necessidade (CHAPMAN et al., 2000 apud. LIMA, 2021).

3.1.4. Modelagem

De acordo com Chapman et al. (2000 apud. LIMA, 2021), na fase da modelagem são escolhidos, aplicados e testados os modelos a serem usados no projeto. Nessa fase, dependendo da modelagem escolhida, pode se fazer necessário o retorno na fase de preparação dos dados.

3.1.5. Avaliação

Como cita Chapman et al. (2000 apud. LIMA, 2021, p. 18):

> A avaliação consiste em identificar se o modelo escolhido está apto a cumprir os objetivos que foram definidos na primeira fase, caso não esteja, é necessário voltar à primeira etapa e rever o escopo e/ou objetivos.

Também é recomendado a revisão das fases anteriores, para se certificar que está tudo conforme o planejado.

3.1.6. Implementação

A fase da implementação tem como objetivo executar o modelo escolhido. De acordo com Chapman et al. (2000 apud. LIMA, 2021), a implementação do modelo não é o fim do projeto, pois deve se fazer um monitoramento e adaptação dos dados e resultados. Para melhor entendimento do cliente, pode ser necessário a criação de um relatório mais compreensível.

A Figura 2 apresenta um diagrama das fases do CRISP-DM descritas no tópico 3.1.

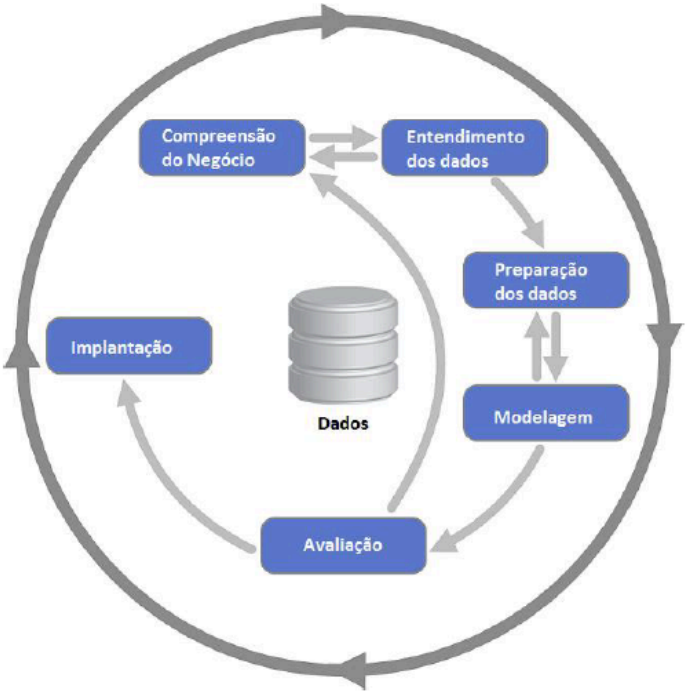

Figura 2 - Diagrama das fases do CRISP-DM. Fonte: Shearer (2000 apud. Ramos et al., 2020)

4. ESTUDO DE CASO

4.1. Entendimento do Negócio

A Segrob Notlad consolidou-se no segmento brasileiro de fast fashion através de uma combinação estratégica de design acessível, campanhas impactantes e uma identidade visual marcadamente urbana. Sua trajetória, iniciada no Rio de

Janeiro pelo imigrante croata Segrob Notlad, reflete uma síntese singular entre influências europeias e a dinâmica do mercado fashion brasileiro. Atualmente, a organização opera uma rede de mais de 80 lojas no território nacional, além de estabelecer presença em mercados sul-americanos e europeu.

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

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

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

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

Além disso, o modelo deve ser flexível o bastante para se ajustar a possíveis mudanças ao longo do tempo, como promoções de fim de ano ou alterações no comportamento dos consumidores. Por fim, espera-se que a solução proposta traga efeitos práticos e mensuráveis, como a redução de custos com estoque ou ganhos de eficiência no processo de reposição de produtos.

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

planejamento da cadeia de suprimentos.”

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

4.2. Entendimento dos Dados

4.2.1. Comportamento de Vendas ao Longo do Tempo

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

Cada entrada possui:

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

A seguir, apresentam-se algumas estatísticas descritivas iniciais do conjunto de dados, que permitem uma visão geral do comportamento das vendas ao longo do período analisado:

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

Tabela 1 - Estatísticas Descritivas Iniciais

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

As vendas médias diárias aumentaram progressivamente ao longo do tempo:

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

- 2024: aproximadamente 250 unidades/dia

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

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

Além do Natal, outras datas com picos notáveis incluem:

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

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

4.2.2. Análise Exploratória

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

Foi construído um gráfico de linha que apresenta as vendas diárias no período completo analisado (2022–2024), permitindo observar tendências gerais, variações sazonais e identificar picos significativos de demanda ao longo do tempo, conforme ilustrado na Figura 3.

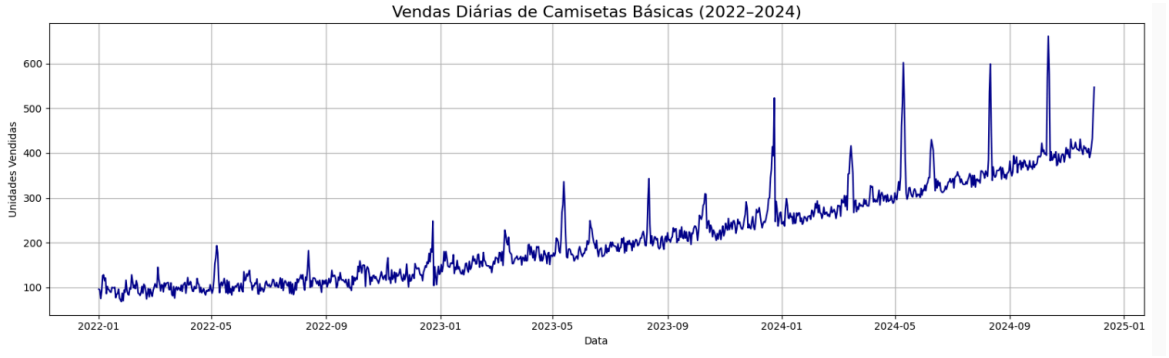

Figura 3 - Vendas diárias de camisetas básicas (2022-2024)

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

A fim de comparar a evolução das vendas ao longo dos anos, foi elaborado um boxplot segmentado por ano. Essa representação evidencia mudanças na mediana, variação e presença de valores extremos em cada período, conforme ilustrado na Figura 4.

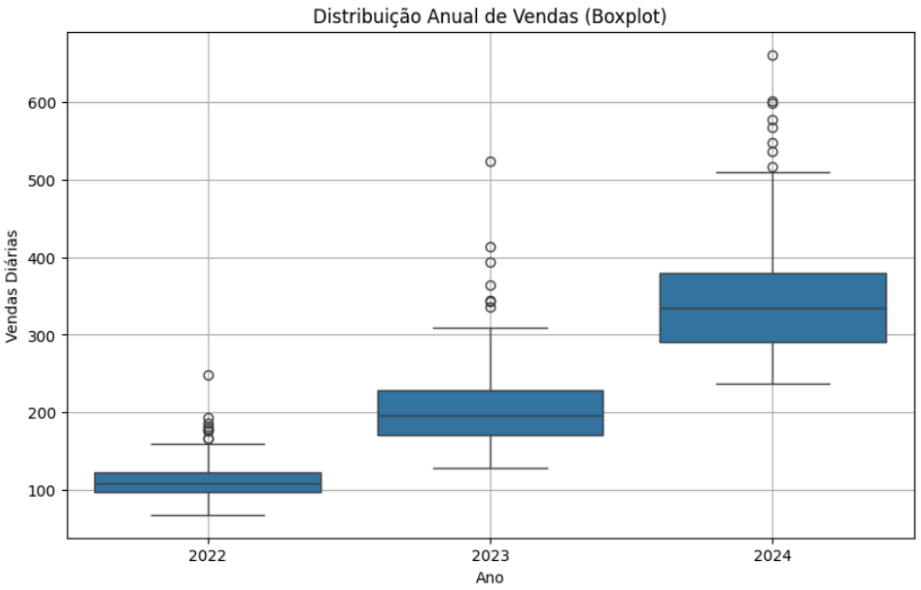

Figura 4 - Boxplot anual de vendas

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

Para investigar padrões sazonais mensais, foi utilizado um boxplot com agrupamento por mês. Essa abordagem facilita a visualização de meses com vendas mais elevadas ou voláteis, como dezembro e maio, conforme apresentado na Figura 5.

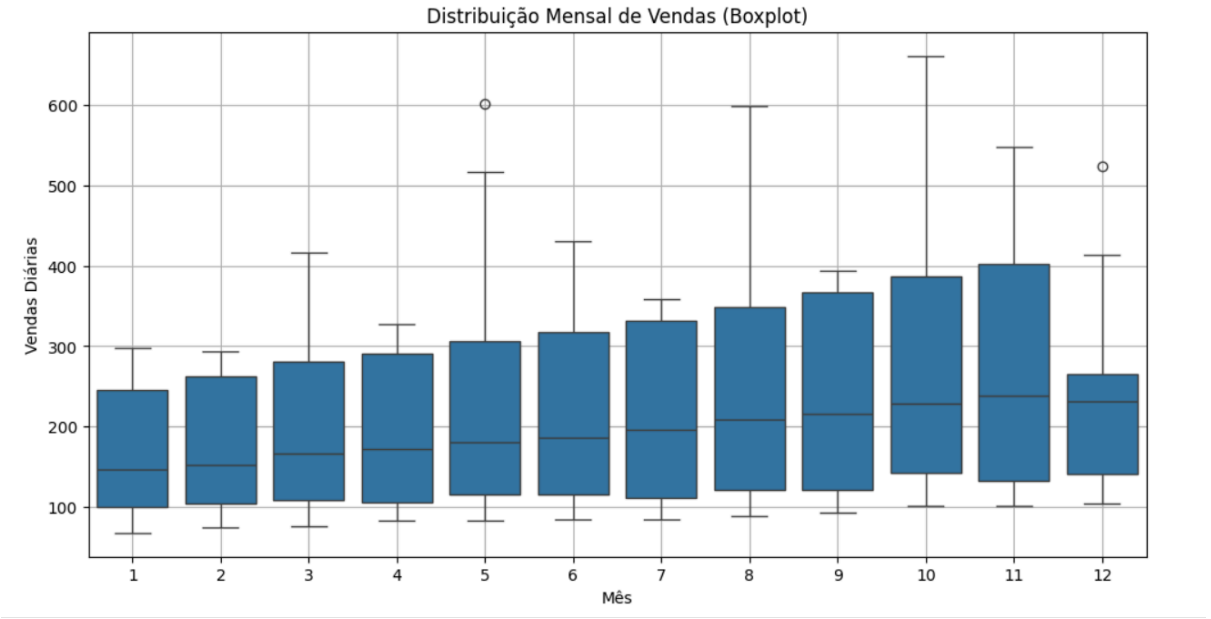

Figura 5 - Boxplot mensal de vendas

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

Uma média de vendas foi calculada para cada dia da semana com o objetivo de verificar padrões semanais de consumo. A visualização resultante, mostrada na figura 6 aponta se há dias com desempenho sistematicamente inferior ou superior ao restante.

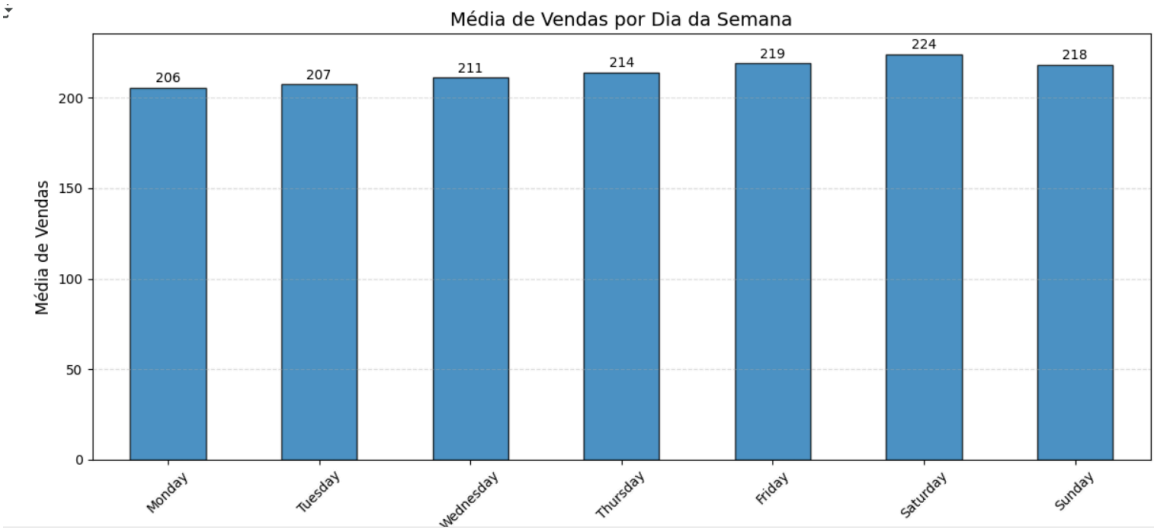

Figura 6 - Média de vendas por dia da semana

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

O histograma das vendas diárias foi utilizado para entender a distribuição geral dos valores observados. Essa análise, representada na Figura 7 revela se há concentração em certos intervalos de venda e permite identificar possíveis outliers.

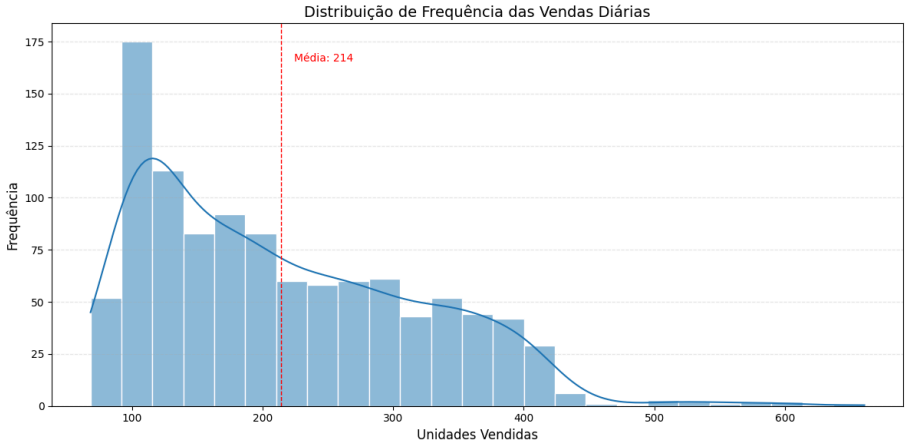

Figura 7 - Histograma de vendas diárias

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

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

Para destacar os momentos de maior demanda, foram selecionados os dias com os maiores volumes de venda em todo o período. A figura 8 apresenta essa distribuição em barras facilitando a identificação de eventos atípicos e datas comemorativas com impacto nas vendas.

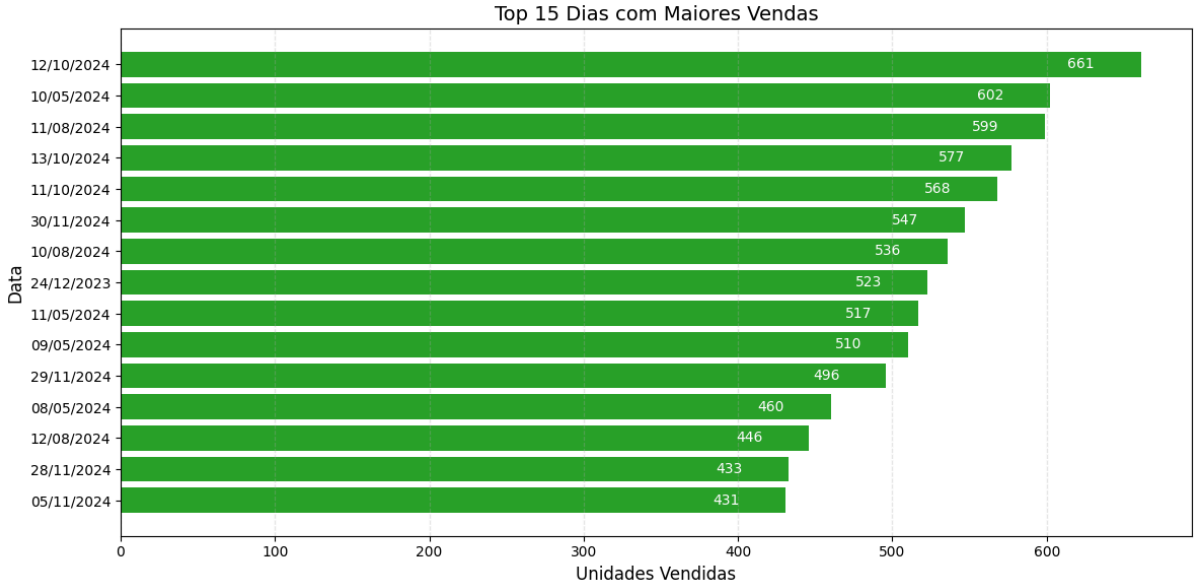

Figura 8 - Top 15 dias com Maiores Vendas

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

A média móvel de 7 dias foi calculada e representada graficamente para suavizar as flutuações diárias e tornar mais visível a tendência geral da série. Essa técnica, ilustrada na figura 9, ajuda a perceber ciclos de alta e baixa de forma mais clara.

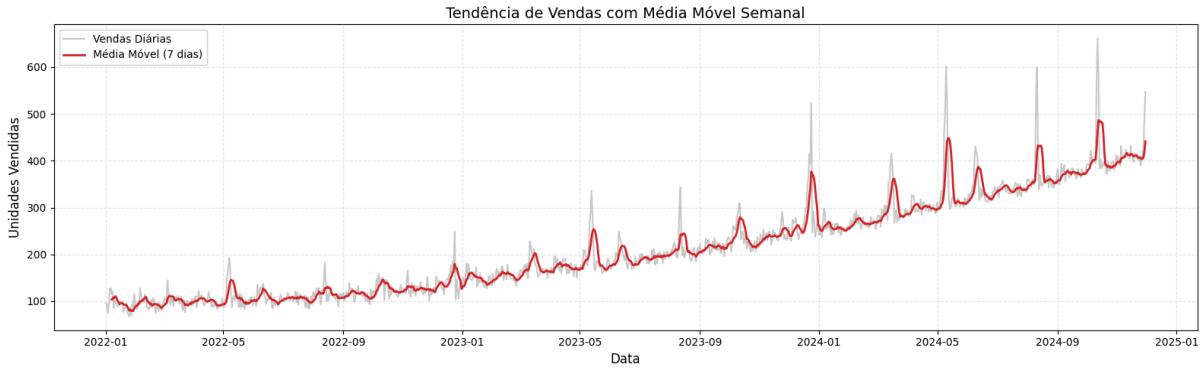

Figura 9 - Tendência de vendas com média móvel semanal

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

4.3. Preparação dos dados

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

A preparação dos dados incluiu atividades específicas descritas a seguir:

4.3.1. Tratamento de Dados Faltantes

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

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

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

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

4.3.2 Detecção e Tratamento de Outliers

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

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

4.3.3 Criação e Enriquecimento de Variáveis

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

As principais variáveis adicionadas foram:

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

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

4.3.4. Definição do Dataset Final para Modelagem

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

4.3.5. Divisão dos Conjuntos de Treino e Teste

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

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

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

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

4.3.6. Resultados da Preparação dos Dados

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

Tabela 2 – Exemplo das primeiras linhas do DataFrame após preparação dos dados

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

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

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

4.4. Modelagem

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

4.4.1. Modelo Naive

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

4.4.2. Modelo Acumulativo

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

4.4.3. Modelo Média Móvel Simples

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

4.4.4. Resultado da Modelagem

A aplicação dos modelos preditivos (Naive, Acumulativo e Média Móvel) foi realizada por meio do Código 3: Modelagem Preditiva, e envolveu as seguintes etapas técnicas:

- Carregamento e leitura dos dados históricos.

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

Após a execução dessas etapas, foi obtida a Tabela 3, que resume claramente as previsões realizadas para dezembro de 2024:

Tabela 3 – Resultados das Previsões dos Modelos (Dezembro/2024)

Data	Naive	Acumulativo	Média Móvel
01/12/2024	547	214,16	441,29
02/12/2024	547	214,16	441,29
03/12/2024	547	214,16	441,29
…	…	…	…
30/12/2024	547	214,16	441,29
31/12/2024	547	214,16	441,29

O modelo Naive, por exemplo, é bastante simples e assume que as vendas futuras serão iguais à última observação feita, que neste caso foi de 547 unidades. Já o modelo Acumulativo também produz previsões constantes, pois utiliza a média geral histórica das vendas, que corresponde a aproximadamente 214,16 unidades. Da mesma forma, o modelo de Média Móvel Simples emprega a média das vendas observadas na última semana, resultando num valor fixo de 441,29 unidades para todas as previsões futuras.\

Para facilitar a interpretação visual dos resultados da modelagem, foi gerado um gráfico, figura 10, detalhado que mostra o comportamento histórico recente (novembro/2024) comparado às previsões obtidas pelos três modelos aplicados.

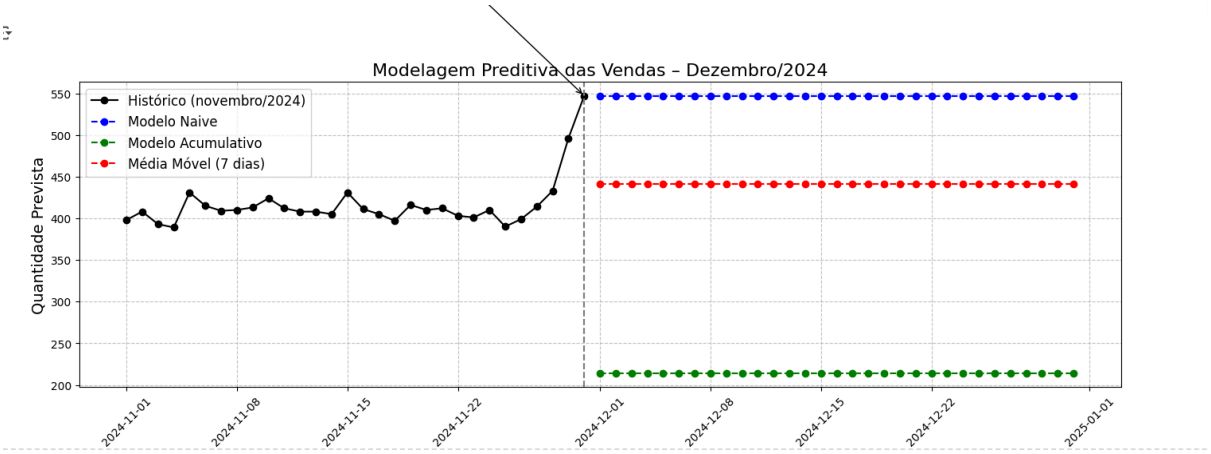

Figura 10 – Previsões dos Modelos para Dezembro de 2024

O gráfico demonstra como cada modelo prevê o comportamento das vendas para dezembro. A linha divisória destacada no gráfico ilustra o fim do período histórico e o início das previsões futuras, proporcionando uma interpretação mais clara do comportamento esperado das vendas segundo cada modelo.

Para o Modelo Naive, a previsão constante de 547 unidades se deve ao fato deste ter sido o último valor histórico observado (em 30/11/2024). Este modelo não considera tendências ou variações futuras, sendo apenas uma referência simples para comparação. Para o Modelo Acumulativo, a previsão constante de aproximadamente 214 unidades é resultado da média histórica geral (considerando todos os quase 3 anos de dados disponíveis). Reflete uma visão mais conservadora e estável, sem dar grande peso às flutuações recentes. Para a Média Móvel, a previsão constante de aproximadamente 441 unidades representa uma média dos últimos sete dias observados em novembro de 2024, oferecendo uma abordagem intermediária que capta tendências recentes, mas não é sensível a flutuações diárias extremas.

Essas diferenças mostram claramente como cada modelo interpreta os dados históricos, fornecendo distintas visões sobre o futuro.