UNIVERSIDADE FEDERAL DO PARANÁ

CARLOS AUGUSTO MATTOS DO BOMFIM

COMPARAÇÃO DE MÉTODOS DE IA PARA PREVISÃO DE VENDAS EM SÉRIE TEMPORAL COM SAZIONALIDADE: APLICAÇÃO DE *ARIMA*, *Holt-Winters Exponential Smoothing* E *XGBoost*

CURITIBA, PR
2023

CARLOS AUGUSTO MATTOS DO BOMFIM

COMPARAÇÃO DE MÉTODOS DE IA PARA PREVISÃO DE VENDAS EM SÉRIE TEMPORAL COM SAZIONALIDADE: APLICAÇÃO DE *ARIMA*, *Holt-Winters Exponential Smoothing* E *XGBoost*

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Especialista em Inteligência Artificial Aplicada no Programa de Pós-Graduação em Inteligência Artificial Aplicada, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná.

Orientador: Prof. Dr. Jaime Wojciechowski.

CURITIBA, PR
2023

MINISTÉRIO DA EDUCAÇÃO
SETOR DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
CURSO DE PÓS-GRADUAÇÃO INTELIGÊNCIA ARTIFICIAL APLICADA - 40001016348E1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação INTELIGÊNCIA ARTIFICIAL APLICADA da Universidade Federal do Paraná foram convocados para realizar a arguição da Monografia de Especialização de **CARLOS AUGUSTO MATTOS DO BOMFIM** intitulada: **Comparação de Métodos de IA para Previsão de Vendas em Série Temporal com Sazionalidade: Aplicação de Arima, Holt-Winters Exponential Smoothing e Xgboost**, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua Aprovação no rito de defesa.

A outorga do título de especialista está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 09 de Novembro de 2023.

JAIME WOJCIECHOWSKI
Presidente da Banca Examinadora

RAZER ANTHOM NIZER ROJAS MONTAÑO
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Comparação de métodos de IA para previsão de vendas em série temporal com sazionalidade: Aplicação de *ARIMA*, *Holt-Winters Exponential Smoothing* e *XGBoost*

Carlos Augusto Mattos do Bomfim
Setor de Educação Profissional e Tecnológica
Universidade Federal do Paraná
Curitiba, Brasil
cmattosdobomfim@gmail.com

Prof. Dr. Jaime Wojciechowski
Setor de Educação Profissional e Tecnológica
Universidade Federal do Paraná
Curitiba, Brasil
jaimewo@ufpr.br

***Resumo*—A previsão de vendas é uma parte crucial para um negócio de sucesso e este estudo tem como objetivo a comparação de dois métodos de inteligência artificial para identificar qual seria a mais adequada para o caso específico da base de dados que foi cedida para este estudo. Os métodos analisados foram – *ARIMA (Auto Regressive Integrated Moving Average)*, *Holt-Winters Exponential Smoothing* e *XGBoost(Extreme Gradient Boosting)*, utilizando as métricas E*rro Quadrático Médio (RMSE)* e *Mean Absolute Percentage Error (MAPE)* como formas de avaliação dos métodos. O estudo foi feito dentro do ambiente google colaboratory utilizando a linguagem Python e mostrou que é possivel utilizar esses três métodos para este caso, sendo a principal limitação a própria base de dados, que dificulta a utilização dos métodos e a interpretação dos resultados.**

***Palavras-chave*—Predição de vendas, Holt-Winters Exponential Smoothing, XGBoost, Extreme Gradient Boosting, ARIMA**

***Abstract*—The sales forecast is a crucial part of a successful business, and this study aims to compare two artificial intelligence methods to identify which one would be most suitable for the specific case of the dataset provided for this study. The analyzed methods were – *ARIMA (Auto Regressive Integrated Moving Average)*, *Holt-Winters Exponential Smoothing* and *XGBoost(Extreme Gradient Boosting)*, using the metrics *Root Mean Squared Error (RMSE)* and *Mean Absolute Percentage Error (MAPE)* as evaluation criteria. The study was conducted within the Google Colaboratory environment using the Python language and demonstrated that it is possible to use these two methods for this case. The main limitation observed was the dataset itself, which hinders the application of the methods and the interpretation of results.**

***Index Terms*—Sales forecast, Holt-Winters Exponential Smoothing, XGBoost, Extreme Gradient Boosting, ARIMA**

I. DESENVOLVIMENTO

A previsão de vendas desempenha um papel fundamental na gestão de qualquer empresa, principalmente em uma empresa de bens de consumo. Antecipar as vendas futuras com precisão é essencial para um melhor planejamento de operações, incluindo a aquisição de estoque, gerenciamento da cadeia de suprimentos e estratégias de marketing. A previsão precisa das vendas é a chave para evitar custos desnecessários de estoque e garantir que os produtos estejam prontamente disponíveis para atender às necessidades dos clientes. Além disso, uma previsão de vendas confiável é essencial para otimizar a alocação de recursos e embasar decisões estratégicas.

Nesse contexto, a inteligência artificial (IA) desempenha um papel crucial na melhoria da precisão e eficácia das previsões de vendas. Sua capacidade de processar grandes volumes de dados, identificar padrões complexos e adaptar-se a mudanças nas condições do mercado a torna uma ferramenta inestimável.

Este estudo concentra-se na comparação de três abordagens de previsão: *ARIMA (AutoRegressive Integrated Moving Average)*, *Holt-Winters Exponential Smoothing* e *XGBoost(Extreme Gradient Boosting)*. A comparação será realizada usando as métricas *RMSE (Root Mean Square Error)* e *MAPE (Mean Absolute Percentage Error)* implementados em Python dentro do ambiente Google Colab, conforme detalhado adiante.

A. Tecnologias

O script foi desenvolvido dentro do ambiente Google Colaboratory, utilizando as seguintes aplicações e bibliotecas:

- Linguagem Python, versão 3.10.12;
- RStudio, versão 2022.02.3;
- Pacote *google.colab*, versão 1.0.0;
- Pacote *numpy*, versão 1.23.5;
- Pacote *pandas*, versão 1.5.3;
- Pacote *matplotlib*, versão 3.7.1;
- Pacote *seaborn*, versão 0.12.2;
- Pacote *sklearn*, versão 1.2.2;
- Pacote *statsmodels*, versão 0.14.0;

B. Descrição dos dados

A base de dados utilizada foi gentilmente fornecida por uma empresa privada do setor de venda de produtos de consumo. A base de dados contém o histórico de vendas de uma das lojas da empresa e foi disponibilizada com algumas restrições com o intuito de proteger as informações sensíveis da empresa.

Os dados fornecidos não identificam os produtos específicos vendidos, mas se concentram exclusivamente na quantidade total de produtos vendidos no final do dia. Essa abordagem de anonimização dos dados preserva a confidencialidade dos produtos, porém faz com que no final ocorra uma previsão geral de vendas, não sendo possivel uma análise muito específica das previsões.

A base de dados é composta por um total de 937 registros, abrangendo o período de janeiro de 2014 a setembro de 2016. Dentro da base, as colunas são contituidas pelos atributos, que são eles: Data, com a concatenação de dia, mes e ano das vendas; Venda, com o volume total de produtos vendidos no final do dia; Estoque, com a quantidade de estoque total disponível no final do dia; Preço, que é o valor médio dos produtos vendidos no dia. A Tabela I descreve os atributos disponíveis em cada registro, e a tabela II fornecendo uma visão geral das informações contidas nos dados.

Esses atributos são fundamentais para a realização das previsões de vendas e serão utilizados nas diferentes abordagens de previsão comparadas neste estudo. A seguir, serão detalhadas as técnicas de previsão utilizadas e os resultados obtidos por meio das métricas RMSE e MAPE.

Tabela I
ATRIBUTOS DA BASE DE DADOS

Atributo	Descrição	tipo de dados
$Data$	Dia da venda	object
$Venda$	Total de produtos vendidos	int64
$Estoque$	Total de estoque	int64
$Preco$	Preço médio dos produtos	float64

Fonte: O autor (2023)

Tabela II
PRIMEIRAS LINHAS DO BANCO DE DADOS

Data	Venda	Estoque	Preco
2014-01-01	1	4972	1,29
2014-01-02	70	4902	1,29
2014-01-03	59	4843	1,29
2014-01-04	93	4750	1,29
2014-01-05	96	4654	1,29

Fonte: O autor (2023)

C. Métodos

A Figura 1 ilustra a divisão do processo de estudo deste trabalho em seis etapas, utilizando dados do histórico de vendas. Cada uma dessas etapas desempenha um papel crucial no desenvolvimento e na avaliação dos modelos de previsão baseados em inteligência artificial. O objetivo final é determinar qual dos modelos é o mais preciso. As etapas são as seguintes:

1) **Tratamento da Base de Dados:** Nesta etapa, os dados brutos do histórico de vendas são processados e preparados para análise. Isso inclui a limpeza de dados ausentes ou inconsistentes e a formatação adequada para análises subsequentes.
2) **Análise Estatística da Base:** Uma análise estatística é realizada para explorar a distribuição dos dados, identificar tendências e padrões dos registros de vendas.
3) **Pré-processamento dos dados:** Nesta etapa, os dados são divididos em conjuntos de treinamento e teste, permitindo que os modelos de previsão sejam treinados e avaliados de maneira adequada.
4) **Treinamento:** Os modelos de previsão são treinados usando os dados de treinamento, com o objetivo de aprender os padrões e relações subjacentes nos dados históricos.
5) **Predição:** Os modelos treinados são usados para realizar previsões das vendas futuras com base nos dados de testes, e é feito a comparação entre seus resultados utilizando recursos gráficos e as métricas RMSE (*Root Mean Square Error*) e MAPE (*Mean Absolute Percentage Error*)

O estudo visa identificar a técnica de previsão de vendas que teve a melhor eficaz dentro do contexto de uma loja de produtos de consumo, com base nos resultados obtidos nas etapas acima descritas.

1) tratamento da base de dados: Para garantir a precisão e eficácia das previsões de vendas, é de extrema importância assegurar a qualidade dos dados utilizados. Mesmo após a realização de um pré-processamento prévio pela própria empresa dona da base, é necessário uma revisão detalhada para verificar a integridade dos dados e prepará-los para a análise. Essa revisão inclui a identificação de valores nulos e a detecção de possíveis valores discrepantes (*outliers*). Após a análise foi contatado que o banco de dados utilizado nesse estudo está livre de valores nulos, e outliers.

No entanto, para alcançar previsões precisas e significativas, foi necessária a alteração do formato da variável de data que originalmente era um campo string, mas que para ser melhor trabalhada pelos algoritmos foi transformada em um campo de data. resultando em um banco de dados como mostram as tabelas III.

Tabela III
ATRIBUTOS DA BASE DE DADOS

Atributo	Descrição	tipo de dados
$Data$	Dia da venda	datetime64[ns]
$Venda$	Total de vendas no dia	int64
$Estoque$	Estoque no final do dia	int64
$Preco$	Preço médio dos produtos	float64

Fonte: O autor (2023)

2) Análise estatística da base: Antes de mais nada, é importante ter uma noção de como é a série que será trabalhada, e para isso a imagem 2 nos mostra o gráfico da série temporal de vendas.

O objetivo deste estudo é comparar três técnicas de previsão: ARIMA (*AutoRegressive Integrated Moving Average*), *Holt-*

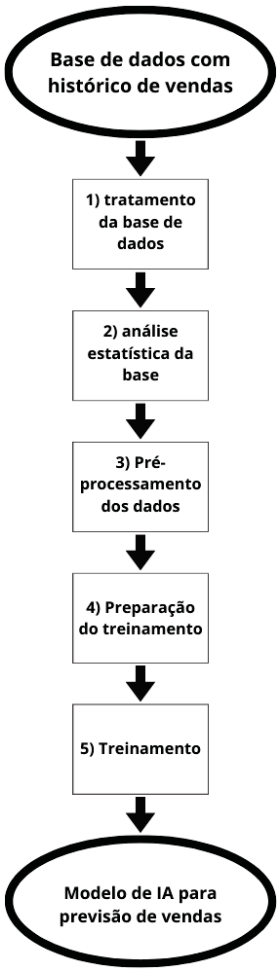

Figura 1. Etapas do método aplicado no estudo.
Fonte: O autor (2023)

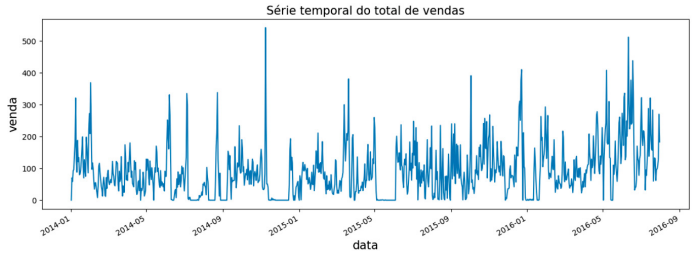

Figura 2. Série temporal de vendas.
Fonte: O autor (2023)

Winters Exponential Smoothing e *xgboost*. Para garantir a eficácia desses modelos dentro do contexto de séries temporais é muito relevante entender se a série que vai ser trabalhada é estacionariária ou não.

Para confirmar a estacionariedade da série temporal foi utilizado o teste ADF (*Augmented Dickey-Fuller*), que é uma extensão do teste Dickey-Fuller, utilizado para testar a hipótese de que uma série temporal possui raiz unitária, o que indica a presença de não estacionariedade. A hipótese alternativa é que a série tem a característica1 de não estacionária.

O teste ADF nos dá 4 métricas, a Estatística ADF e Valores Críticos de 1%, 5% e 10%.

Nesse teste temos a *Hipótese Nula* (H0): A presença de raiz unitária, indicando que a série temporal é não estacionária. *Hipótese Alternativa* (H1): A ausência de raiz unitária, indicando que a série temporal é estacionária.

E para julgar essas hipóteses podemos fazer de dois métodos:

1- Se a estatística de teste for menor que os valores críticos , pode-se rejeitar a hipótese nula e concluir que a série temporal é estacionária.

2- Se o p-value for menor ou igual a um determinado nível de significância, normalmente, igual a 0.01 (1%), 0.05 (5%) ou até mesmo 0.1 (10%) o H0 é rejeitado e é possivel concluir que esta série temporal é estacionária.

Seguindo essa análise na tabela IV resultante do teste ADF, é possivel de concluir que esta série temporal é estacionária.

Tabela IV
RESULTADO TESTE ADF

Estatística ADF	**Valor crítico**		
	1%	**5%**	**10%**
-5,326902	-3,437	-2,865	-2,568

Fonte: O autor (2023)

Além do teste ADF se faz necessário uma análize das componentes de tendência e sazonalidade, pois além de entender melhor a série temporal, auxilia na escolha ente os três modelos de suavização que foram levantados para esse experimento, o algoritmo de Suavização Simples, que utiliza apenas a componente de nível, que é o valor médio da série ao longo do tempo para a suavisação; O algoritmo de Suavização Holt, que além da componente de nível utiliza a componente de tendência; e o algoritmo Holt-Winter, que além do nível e tendência utiliza a componente de sazonalidade, que representa a mudança de longo prazo no nível de uma série temporal [1] e é uma característica de uma série temporal na qual os dados sofrem mudanças regulares e previsíveis [2].

Analisando o gráfico da componente de sazonalidade 3, gerado utilizando as biblitecas matplotlib, é perceptível que existe sim uma sazonalidade, por conta do padrão de repetição dentro da componente de sazonalidade.

Observando o gráfico da componente de tendência 4 é possivel dizer que existe uma tendência a partir de janeiro de 2016, até o fim da série, para que as vendas aumentem. Porém analisando junto com o teste do teste ADF, é difícil chegar a conclusão que exista uma tendência dentro desta série,

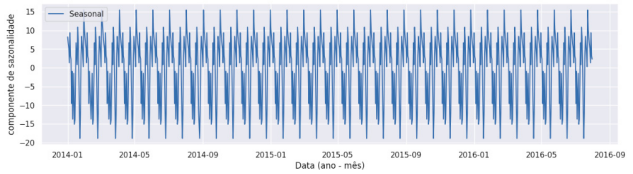

Figura 3. gráfico da componente de sazonalidade das vendas.
Fonte: O autor (2023)

pois isso implicaria que a mesma não é estacionária, mas não podemos excluir essa possibilidade.

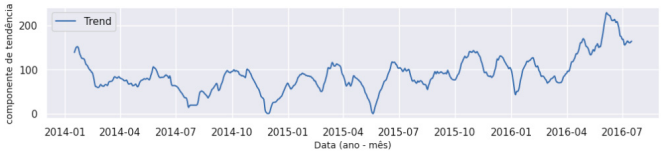

Figura 4. gráfico da componente de tendência das vendas.
Fonte: O autor (2023)

Dentro do contexto do modelo ARIMA(p,d,q) existem os três parâmetros: de Diferenciação (d), ordem do Termo de médias móveis (q) e ordem do Termo auto-regressivo (p), onde os valores próximos do ideal de p e q p podem ser identificados por meio das análises dos gráficos de ACF (*Função de Autocorrelação)* e PACF (*Função de Autocorrelação Parcial)*. Assim os gráficos de autocorrelação 5 e autocorrelação parcial 6 foram feitos.

Nos gráficos de autocorrelação 5 e autocorrelação parcial 6, o eixo x mostra quantos momentos anteriores estão sendo comparados, enquanto no eixo y indica o nível de relação [3]. Dentro dos gráficos estão os limites críticos, ou seja, 5% da auto correlação, ou, segundo a fórmula 2, onde N é o número de pontos do dataframe e d é o número de vezes que diferenciamos os dados.

$$\frac{1.96}{\sqrt{N-d}} \tag{1}$$

Os limites Críticos são úties pois a partir deles que identificaremos os parâmetros a serem usados no ARIMA.

Também é possível observar nos gráficos gráficos do ACF e PACF que existem picos que acontecem aproximadamente a cada 7 marcações de tempo, indicando que pode haver sazonalidade na série temporal.

Além disso, foram realizadas análises de correlação entre as variáveis, que indicou a ausência de uma tendência significativa, mas evidenciaram a presença de sazonalidade [4].

Além disso, não foi observada correlação entre as variáveis, como indica a matriz de correlação na tabela V, resultante da função de correlação da biblioteca Pandas [5]. Isso é relevante, pois se existisse uma correlação, seria um indício de que técnicas de regressão teriam uma boa taxa de acurácia, como explicado por Ferber em [6].

Portanto, com base na análise realizada, é apropriado empregar o modelo ARIMA e de Holt-Winters para as previsões de vendas.

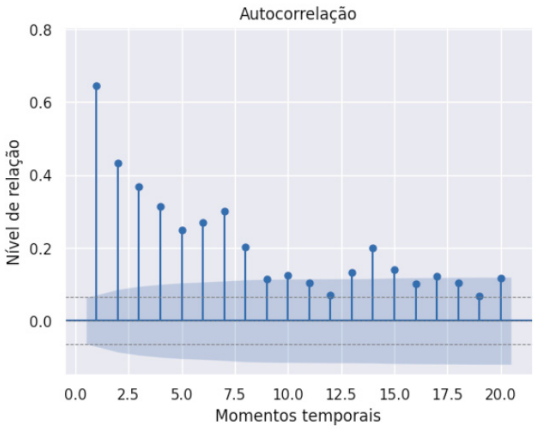

Figura 5. gráfico de autocorrelação das vendas.
Fonte: O autor (2023)

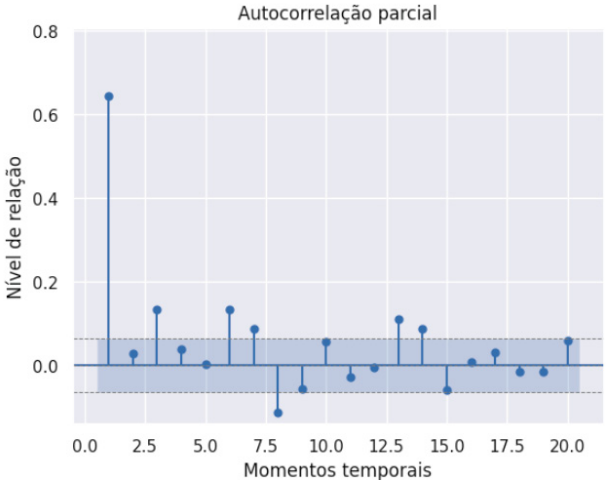

Figura 6. gráfico de autocorrelação parcial das vendas.
Fonte: O autor (2023)

Tabela V
CORRELAÇÃO ENTRE VARIÁVEIS

Variável	Venda	Estoque	Preço
Venda	1,000000	0,153659	0,094779
Estoque	0,153659	1,000000	-0,032604
Preço	0,094779	-0,032604	1,000000

Fonte: O autor (2023)

3) Pré-processamento dos dados: Uma vez que a verificação de valores nulos e informações irrelevantes já foi realizada anteriormente para determinar a adequação da base de dados ao estudo, a etapa subsequente abordou a sazonalidade e a possibilidade de tendência na série temporal. A presença da sazonalidade pos sí só não indica que a série não é estacionária, mas a tendência sim mudaria esse aspecto, o que pode ser um

problema para os modelos. Portanto, como existe a incerteza sobre a natureza desta série, o banco de dados foi duplicado, e nesta cópia foi utilizado a técnica de diferenciação nos dados de venda, que consiste em calcular a diferença entre valores consecutivos da série [7] [8], conforme mostrado na equação 2.

$$\Delta y_t = y_t - y_{t-1} \tag{2}$$

Onde Δy_t representa a diferença entre os valores atual y_t e o valor anterior y_{t-1} da série temporal.

É importante destacar que, embora essa etapa não seja um fator crítico para o método XGBoost, conforme discutido por [9].

Após a diferenciação, a base ficou como mostra a imagem 7. E Assim os modelos vão ser treinados e testados uma vez com cada banco, para ver se a diferenciação resulta em alguma melhora nas previsões dos modelos.

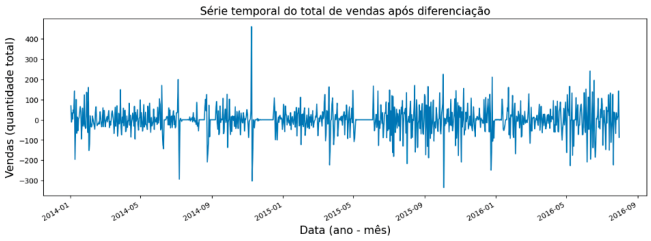

Figura 7. Série temporal após diferenciação.
Fonte: O autor (2023)

4) Preparação do treinamento: Após o pré-processamento dos dados, foram construídos os modelos de previsão ARIMA (AutoRegressive Integrated Moving Average) e o Holt-Winters Exponential Smoothing , utilizando a linguagem de programação Python com a TensorFlow para implementação dos algoritmos de aprendizado de máquina.

Para o conjunto de dados foi criado uma nova base de dados, contendo apenas as informações de data e do total de vendas, pois apenas elas que serão preditas neste trabalho, por conta do tratamento na base feita pela da empresa dona dos dados, que somou a venda de todos os produtos em apenas uma coluna, tirando a possibilidade de fazer uma análise mais profunda por produto. sendo que após o novo tratamento da base os atributos ficaram igual a tabela VI, e a tabela VII mostra os primeiros gegistros do banco de treinamento.

Tabela VI
ATRIBUTOS DA BASE DE DADOS DE TREINAMENTO E TESTE

Atributo	Descrição	tipo de dados
$Data$	Dia da venda	datetime64[ns]
$Venda$	Total de vendas no dia	int64

Fonte: O autor (2023)

O conjunto de dados foi dividido em dois conjuntos distintos, a base de treino ficou com os primeiros 842 dos 937 totais da série temporal, cerca de 90% da quantidade de registros

Tabela VII
PRIMEIRAS LINHAS DO BANCO DE DADOS

Data	Venda
2014-01-01	1
2014-01-02	70
2014-01-03	59
2014-01-04	93
2014-01-05	96

Fonte: O autor (2023)

da base, enquanto a base de teste ficou com os ultimos 95 registros da série, aproximadamente 10% do total de registros da base.

A divisão dos dados ficou conforme a a figura 8, onde em cor preta estão as 842 observações que serão usadas para treinar os algoritmos e em vermelho as 95 observações de treino.

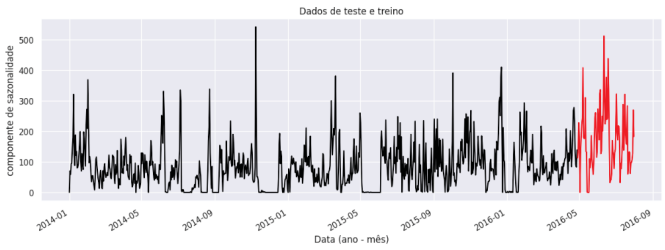

Figura 8. Divisão de treino e teste.
Fonte: O autor (2023)

No entanto, no caso do modelo XGBoost, foi necessária uma divisão adicional em quatro conjuntos (x treino, y treino, x teste e y teste).

Para avaliara qualidade do desempenho dos modelos de previsão, foram utilizados duas métricas, o RMSE e o MAPE.

Na equação 3, $RMSE$ representa o Erro Quadrático Médio, n é o número de observações, y_i são os valores reais observados e $\hat{y}_i$ são os valores previstos pelos modelos. O $RMSE$ é uma métrica que nos permite quantificar o erro das previsões em relação aos valores reais, facilitando a avaliação do desempenho dos modelos.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

Já na Equação 4, $MAPE$ representa o Mean Absolute Percentage Error, onde novamente n é o número de observações, y_i são os valores reais, e $\hat{y}_i$ são os valores previstos pelos modelos. O $MAPE$ fornece uma medida de erro em termos percentuais, sendo útil para avaliar o erro médio relativo das previsões.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{4}$$

5) Treinamento: Neste estudo, foram empregadas técnicas de IA ARIMA e o Holt-Winters Exponential Smoothing.

O ARIMA é um modelo estatístico que combina *componentes autorregressivos (AR)*, *diferenciação integrada (I)* e *médias móveis (MA)* para a modelagem e previsão de séries temporais, permitindo uma especialização em cada modelo, preenchendo a ordem dos *componentes autorregressivos (p)*, de *diferenciação (d)* e de *média móvel (q)*. Os valores dos hiperparâmetros que foram testados e tiveram os melhores desempenhos estão resumidos na Tabela VIII.

Tabela VIII
HIPERPARÂMETROS DOS MODELOS ARIMA

Arima Model	p	d	q
1	2	1	12
2	12	0	2
3	2	1	8
4	6	0	0
5	0	1	2

Fonte: O autor (2023)

É possivel fazer a análise no gráfico de autocorrelação **??** e no de autocorrelação parcial **??**, para encontrar os hiperparâmetros que possam ter o melhor desempenho, seguindo duas regras:

Para o hiperparâmetro p é possivel localizar no gráfico 6, de autocorrelação parcial de vendas onde, o PACF cruza o limite superior do intervalo pela primeira vez, neste caso seria perto de 2.

Para o hiperparâmetro q é possivel localizar no gráfico de autocorrelação de vendas **??** onde o ACF cruza o limite superior do intervalo pela primeira vez, neste caso seria perto de 12.

Este método de utilizar os gráficos de autocorrelação não garante ter o melhor desempenho, portanto é importante testar outros hiperparâmetros.

O algoritmo de suavização Holt-Winters utiliza três componentes, o **Nível** (*Nt*) que é o valor médio da série ao longo do tempo, a **tendência** (*Tt*) que são movimento de crescimento ou de decrescimento da série temporal e que pode ser estar em toda ou apenas em uma parte da série; a **sazonalidade** (*Ft*), que são comportamentos que ocorrem na série temporal de forma repetitiva e com periodicidade.

O algoritmo de Holt-Winters tem duas variações, que dependem da característica da série temporal que vai ser utilizada. O modelo Holt-Winter Aditivo é utilizado quando a variação sazonal é aproximadamente constante ao longo do tempo, quando o impacto da sazonalidade não depende do nível da série temporal. E o modelo multiplicativo é apropriado quando a variação sazonal é proporcional ao nível da série temporal, ou seja, que o impacto da sazonalidade é multiplicativo, ou seja, aumenta ou diminui com o nível da série.

A ideia central desse algoritmo é: O nível da série é estimado usando uma média ponderada entre as observações passadas. Os pesos decrescem exponencialmente e são maiores para observações mais recentes. Ou seja, dado a constante α,que varia entre 0 e 1.

As fórmulas para os componentes do modelo Holt-Winters Aditivo são:

1. Nível (N_t):

$$N_t = \alpha \cdot (Y_t - F_{t-L}) + (1 - \alpha) \cdot (N_{t-1} + T_{t-1}) \quad (5)$$

2. Tendência (T_t):

$$T_t = \beta \cdot (N_t - N_{t-1}) + (1 - \beta) \cdot T_{t-1} \quad (6)$$

3. Sazonalidade (F_t):

$$F_t = \gamma \cdot (Y_t - N_t) + (1 - \gamma) \cdot F_{t-L} \quad (7)$$

As fórmulas para os componentes do modelo Holt-Winters Multiplicativo são:

1. Nível (N_t):

$$N_t = \alpha \left(\frac{Y_t}{F_{t-L}} \right) + (1 - \alpha) \cdot (N_{t-1} + T_{t-1}) \quad (8)$$

2. Tendência (T_t):

$$T_t = \beta \cdot (N_t - N_{t-1}) + (1 - \beta) \cdot T_{t-1} \quad (9)$$

3. Sazonalidade (F_t):

$$F_t = \gamma \left(\frac{Y_t}{N_t} \right) + (1 - \gamma) \cdot F_{t-L} \quad (10)$$

Onde Y_t é a observação no período t, L é a periodicidade sazonal, e α, β, e γ são os parâmetros de suavização.

Para utilizar o algoritmo Holt-Winters com a biblioteca statsmodels é necessário definir três parâmetros, sendo eles o *seasonal periods* (número de períodos para completar um ciclo sazonal), *trend* (tipo de tendência) e *seasonal* (tipo de sazonalidade).

Quando o hiperparâmetro seasonal está com o valor add, significa que essa é uma análise de Holt-hinters aditiva, e quando está com o valor mul significa que é multiplicativa.

O hiperparâmetro trend apenas indica quando a série tem uma tendência, e foi testado a hipotese de ter a tendência positiva ou nenhuma tendência.

O seasonal periods é o tempo para se completar um ciclo sazonal, foi testado a hipótese desse ciclo ser semanal, mensal e trimestral.

No contexto deste trabalho, os parâmetros utilizados no modelo de Holt-Winters estão apresentados na tabela IX.

O XGBoost é um método que utiliza o *Gradient Boosting"*, treinando sequencialmente árvores de decisão, onde cada árvore tenta corrigir os erros das árvores anteriores, melhorando o desempenho do modelo.

Para a implementação do XGBoost, a biblioteca "xgboost"em Python oferece a capacidade de ajustar hiperparâmetros para refinar o desempenho do modelo. Neste estudo, os hiperparâmetros que demonstraram o melhor desempenho estão listados na Tabela 4. Esses hiperparâmetros incluem o *n.estimators*, que controla o número total de árvores criadas no modelo; o *early.stopping.rounds*, que interrompe o treinamento quando

Tabela IX
PARÂMETROS HOLT-WINTERS

Algoritmo	seasonal periods	trend	seasonal
Holt-Winters	7	add	add
Holt-Winters	7	None	add
Holt-Winters	7	add	mul
Holt-Winters	7	None	mul
Holt-Winters	30	add	add
Holt-Winters	30	None	add
Holt-Winters	30	add	mul
Holt-Winters	30	None	mul
Holt-Winters	90	add	add
Holt-Winters	90	None	add
Holt-Winters	90	add	mul
Holt-Winters	90	None	mul

Fonte: O autor (2023)

não há mais melhorias de acordo com as métricas de avaliação (RMSE e MAPE), evitando o overfitting.

Esses hiperparâmetros desempenham um papel fundamental no ajuste fino do modelo XGBoost, garantindo resultados mais precisos e evitando problemas de sobreajuste. testados estão an tabela X:

Tabela X
PARAMETROS XGBOOST

Algoritmo	n.estimators	stopping.rounds
10000	500	127,31
10000	50	127,31
1000	50	127,31
100	50	127,31
10	50	127,57

Fonte: O autor (2023)

II. RESULTADOS E DISCUSSÕES

A. *Resultado dos modelos*

A seguir estão os resultados obtidos com as três técnicas: *ARIMA*, *Holt-Winters Exponential Smoothing* e *xgboost*, utilizadas para a previsão de vendas em uma série temporal.

O método Holt-Winters Exponential Smoothing foi dividido em três rodadas de testes, variando entre série aditiva e multiplicativa e no período para completar um ciclo sazonal.

Para analisar os resultados obtidos pelo modelo de Holt-winter foram utilizadas duas formas, a análise gráfica, com o objetivo de ter uma visialização mais clara dos dados, e a análise por meio das métricas RMSE quanto MAPE.

A análise gráfica consiste em plotar em um gráfico só a série temporal da base de treino e os quatro modelos testados, isso foi feito uma vez para cada possibilidade de ciclo sazonal, e gerou os gráficos 9, 10 e 11.

Os resultados gerados pelo algoritmo de Holt-winters de acordo com as métricas RMSE e MAPE estão na tabela XI.

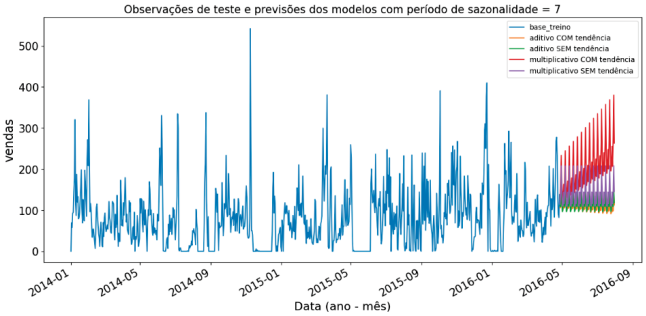

Figura 9. gráfico teste Holt 7 ciclos
Fonte: O autor (2023)

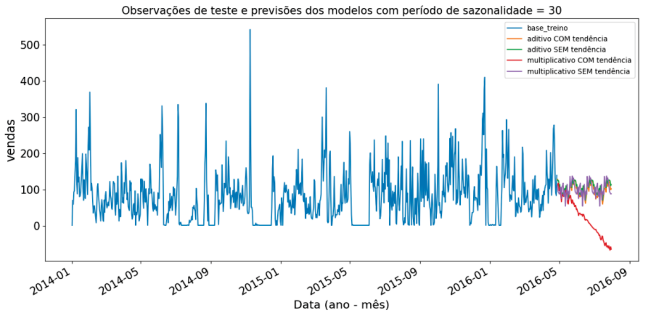

Figura 10. gráfico teste Holt 30 ciclos.
Fonte: O autor (2023)

Os valores mostram que o modelo Holt-Winters não teve uma precisão muito grande, talvez demonstrando que seja necessário fazer a diferenciação nos dados antes de treinar e testar o modelo.

O método ARIMA foi executado alternando os hiperparametros para tentar encontrar o que funcionasse melhor com a série, mas a primeira excussão foi com os hiperparametros retirados dos gráficos 5 e 6, que são de auto correlação e autocorrelação parcial,respectivamente. Isso é um indicio de que o hiperparâmetro p perto de 2 e hiperparâmetro q perto de 12 poderia ter a melhor performance.

Parecido com a análise de resultados do modelo holt-winters, os resultados foram divididos em duas etapas, a gráfica e a

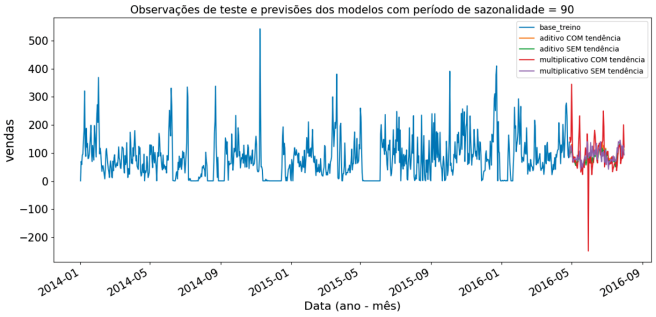

Figura 11. gráfico teste Holt 90 ciclos.
Fonte: O autor (2023)

Tabela XI
RESULTADOS DOS EXPERIMENTOS POR CONFIGURAÇÃO

Algoritmo	Ciclo Sazonais	Resultados	
		RMSE	MAPE
Aditivo com Tendência	7	109,21	64,61%
	30	129,26	49,74%
	90	101,73	52,27%
Aditivo sem Tendência	7	107,09	63,87%
	30	125,13	53,37%
	90	134,57	46,18%
multiplicador sem Tendência	7	101,93	95,2%
	30	159,72	51,34%
	90	112,53	85,26
multiplicador com Tendência	7	95,34	76,29%
	30	127,85	52,49%
	90	128,38	44,86

Fonte: O autor (2023)

das métricas RMSE quanto MAPE.

O gráfico ficou dividido entre os gráficos 12 e 13, para facilitar a visualização.

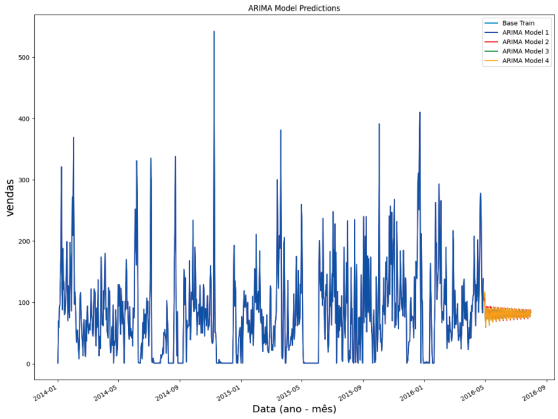

Figura 12. Primeiro gráfico de predição ARIMA.
Fonte: O autor (2023)

Os resultados levando em conta as métricas estão na Tabela XII, que utiliza a métrica RMSE e MAPE para demonstrar o desempenho de cada método.

Os resultados gerados pelo algoritmo ARIMA, assim como o Holt-winters não teve uma precisão muito grande, mostrando, novamente, que talvez seja necessário fazer a diferenciação nos dados antes de treinar e testar o modelo.

Usando os mesmos Parâmetros de antes da diferenciação, existe um problema com o modelo de Holt Winters multiplicativo, pois existem casos em que a diferença entre as vendas de cada dia acaba sendo negativa, e o modelo não multiplicativo não funciona nesses casos.

Mesmo asim foi obtido os resultados dos gráfico gráfico 14, 15 e 16 e da tabela XIII.

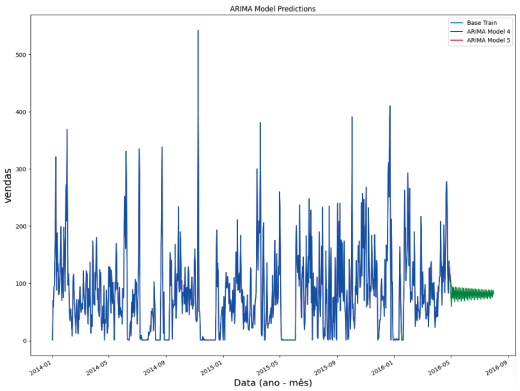

Figura 13. Segundo gráfico de predição ARIMA.
Fonte: O autor (2023)

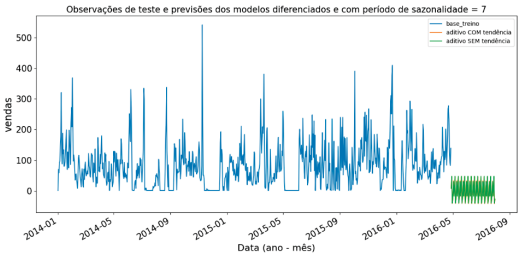

Figura 14. Gráfico teste Holt após diferenciação 7 períodos.
Fonte: O autor (2023)

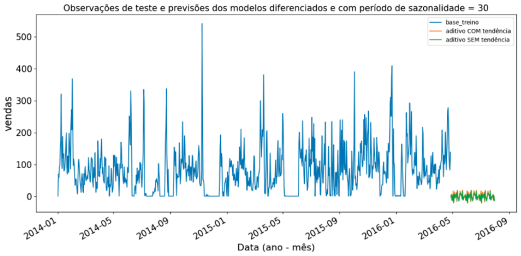

Figura 15. Gráfico teste Holt após diferenciação 30 períodos.
Fonte: O autor (2023)

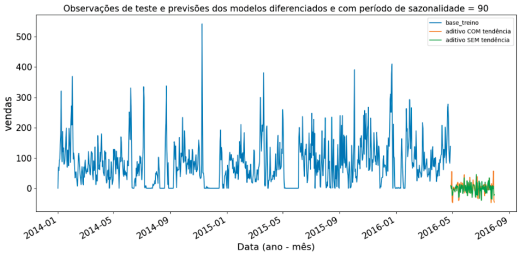

Figura 16. Gráfico teste Holt após diferenciação 90 períodos.
Fonte: O autor (2023)

Tabela XII
RESULTADOS DOS MODELOS ARIMA

Arima Model	p	d	q	RMSE	MAPE
1	2	1	12	133,11	45,32%
2	12	0	2	133,24	44,84%
3	2	1	8	132,64	46,45%
4	6	0	0	133,96	49,16%
5	0	1	2	133,11	45,32%

Fonte: O autor (2023)

Tabela XIII
RESULTADOS DOS EXPERIMENTOS POR CONFIGURAÇÃO APÓS DIFERENCIAÇÃO

Algoritmo	Ciclo Sazonais	RMSE	MAPE
Aditivo com tendência	7	205,97	76,00%
Aditivo sem tendência	7	205,93	79,82%
Aditivo com tendência	30	198,43	72,10%
Aditivo sem tendência	30	198,84	69,84%
Aditivo com tendência	90	198,18	64,80%
Aditivo sem tendência	90	199,32	69,49%

Fonte: O autor (2023)

O arima não teve resultados muito diferentes do Holt winters, onde a diferenciação não melhorou os resultados finais, segundo os gráficos de predição 17 e 18 , e a tabela de resultados XV.

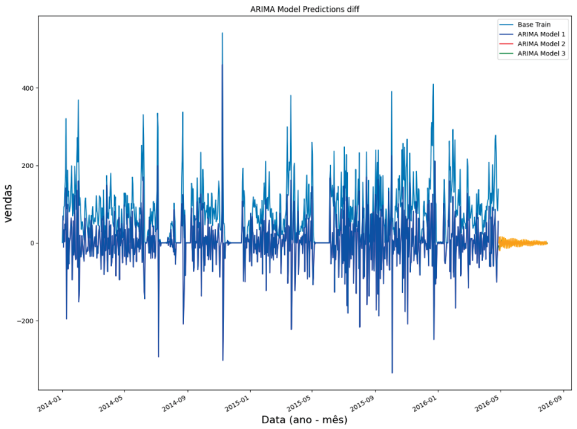

Figura 17. Primeiro gráfico de predição ARIMA após diferenciação.
Fonte: O autor (2023)

O XGboost teve os resultados conforme a tabela abaixo:

B. Discussões

Ambos os modelos mostraram ser capazes de fazer a predição das vendas dentro do caso específico deste estudo, sendo que é necessário analisar cada métrica para entender qual teve o melhor desempenho. Na métrica RMSE o Holt-Winters teve o melhor desempenho, com o algoritmo Aditivo sem Tendência, com o RMSE de 94,20 e MAPE 125,13%, enquanto o ARIMA teve o melhor desempenho no MAPE, com 44,84% e RMSE de

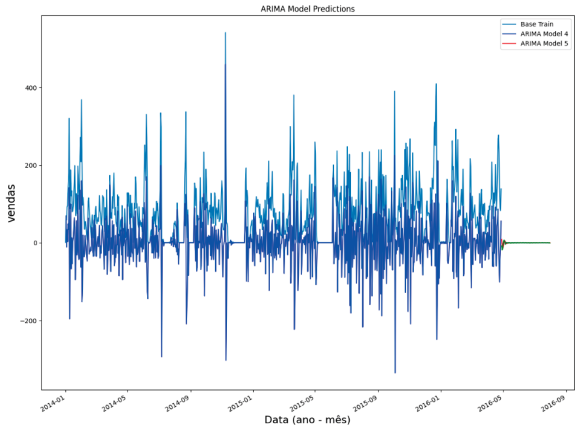

Figura 18. Segundo gráfico de predição ARIMA após diferenciação.
Fonte: O autor (2023)

Tabela XIV
RESULTADOS DOS MODELOS ARIMA DIFERENCIADO

Arima Model	p	d	q	RMSE	MAPE
1	2	1	12	199,60	100,06%
2	12	0	2	199,75	102,02%
3	2	1	8	199,53	144,03%
4	6	0	0	199,35	104,8%
5	0	1	2	199,60	103,00%

Fonte: Fonte: O autor (2023)

133,11, com os hiperparametros p = 2;d = 1;q = 12, enquanto o XGboost se encontra no meio termo, com o RMSE de 127,31 e mape de 48,10%, sendo interessante perceber que apartir de 100 arvores não teve melhora de precisão.

porém existem algumas considerações que são importantes de serem revisadas.

Ambos os modelos foram construidos para o tipo de caso desse estudo, foram feitos para a análise de séries temporais, mas mesmo assim não tiveram um desempenho muito aproveitoso. Isso pode se dever de vários fatores, sendo eles a construção errada dos modelos de teste, análise equivocada dos padrões estatísticos da base e, muito provavelmente, por conta da série temporal ter sido alterada no momento em que foi feito a soma de todos os produtos vendidos em apenas uma variável. Isso faz com que as características de venda de cada produto se misturem e se percam.

Mesmo assim ainda existem algumas ações que ainda podem ser feitas para tentar análisar este banco de dados, sendo elas uma limpeza mais profunda do banco de dados, análises estatísticas diferentes que possam dar mais informações sobre a natureza da série, como a utilização do x13-Arima Seats para a extração da sazonalidade. E é claro, a possibilidade de outros algoritmos para a análise da série.

Porém também é relevante levantar a questão de que mesmo se tivessemos um método de predição acurado, não seria fáciil chegar a alguma conclusão útil com essas informações, pois mesmo tendo os dados de vendas, não sabemos o preço

Tabela XV
RESULTADOS DOS MODELOS XGBOOST

n.estimators	stopping.rounds	RMSE	MAPE
10000	500	127,31	48,10
10000	50	127,31	48,10
1000	50	127,31	48,10
100	50	127,31	48,10
10	50	127,57	48,87

Fonte: Fonte: O autor (2023)

específico dos produtos vendidos, então não tem decidir a priorização de estoque de um produto ou até a época certa de fazer uma promoção, pois no momento em que um produto tem mais venda, o outro pode estar pior, e isso não é demonstrado dentro da série.

III. CONCLUSÕES

Este estudo teve como objetivo estudar a aplicação de diferentes técnicas na previsão de vendas em uma série temporal. A intenção era implementar e comparar essas técnicas para determinar se alguma deria mais adequada para este cenário específico.

O modelo Holt-Winters utiliza a média móvel ponderada exponencialmente (EWMA) para suavizar séries temporais com tendência e/ou sazonalidade. No entanto, mesmo sendo concebido para situações semelhantes às deste estudo como oescrito por [10] e [11], ele não apresentou um desempenho que pode ser considerado praticavel na realidade, sendo a menor taxa o RMSE de 75,410 e MAPE de 100,056%.

E o ARIMA O obteve um MAPE mais aceitável de 44,84%, porém é necessário olhar o todo, e com o RMSE de 133,111 seria necessário uma análise de viabilidade mais profunda.

O XGboost não teve destaque dentro das métricas, com o RMSE de 127,31 e mape de 48,10%, mas é interessante como o algoritmo não teve melhora de precisão apartir de 100 arvores.

É importante notar que todas as técnicas testadas são viáveis de implementar. Além disso, este estudo foi realizado com uma base de baixa complexidade. Em futuras pesquisas, seria interessante aplicar as técnicas a conjuntos de dados mais complexos, que incluam a separação de produtos ou outras variáveis mestras relevantes.

O tema da previsão em séries temporais é muito relevante e possui uma extensa literatura, como o livro de Hyndman et al. [12], que é uma ótima introdução ao mundo da previsão.

REFERÊNCIAS

[1] V. Cerqueira, "Understanding time series trend." Disponível em: <https://towardsdatascience.com/understanding-time-series-trend>. Acessado em: 15 de jun. de 2023.

[2] W. Kenton, "Seasonality: What it means in business and economics, examples." Disponível em: <https://www.investopedia.com/terms/s/seasonality.>. Acessado em: 15 de jun. de 2023.

[3] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," *Proceedings of the 9th Python in Science Conference*, vol. 57, 2010.

[4] L. Liu, *Identification of seasonal arima models using a filtering method*, vol. 57. 1989.

[5] W. McKinney, *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*. Califórnia, EUA: O'Reilly Media, 3 ed., 2022.

[6] R. Ferber, "Sales forecasting by correlation techniques," *Journal of Marketing*, vol. 18, 1954.

[7] R. Pindyck and D. Rubinfeld, "Econometria: modelos e previsões," 2004.

[8] A. S. Lopes, "Autocorrelação e heteroscedasticidade em regressões com séries temporais," 2021.

[9] D. Borkin, A. Némethová, G.Michaľčonok, and K. Maiorov, "Impact of data normalization on classification model accuracy.," *Sciendo*, vol. 27, pp. 79–84, 2019.

[10] A. Veríssimo, C. Alves, E. Henning, C. Amaral, and A. Cruz, "Métodos estatísticos de suavização exponencial, holt-winters para previsão de demanda em uma empresa do setor metal mecânico," *Revista Gestão Industrial*, vol. 8, 2012.

[11] C. Chatfield and M. Yar, "Holt-winters forecasting: Some practical issues," *ournal of the Royal Statistical Society Series D: The Statistician*, vol. 37, 1988.

[12] R. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. Melbourne, Australia: OTexts, 3 ed., 2021.