UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS HUMANAS, SOCIAIS E AGRÁRIAS
DEPARTAMENTO DE CIÊNCIAS SOCIAIS APLICADAS
GRADUAÇÃO EM ADMINISTRAÇÃO

ANÁLISE COMPARATIVA ENTRE MODELOS ARIMA E MODELOS DE MACHINE LEARNING PARA PREVISÃO DE SÉRIES TEMPORAIS DE COMMODITIES

LUCAS SOUZA SANTOS

Bananeiras - PB
Maio, 2024

LUCAS SOUZA SANTOS

ANÁLISE COMPARATIVA ENTRE MODELOS ARIMA E MODELOS DE MACHINE LEARNING PARA PREVISÃO DE SÉRIES TEMPORAIS DE COMMODITIES

Trabalho de Curso apresentado como parte dos requisitos necessários à obtenção do tulo de Bacharel em Administração, pelo Centro de Ciências Humanas Sociais e Agrárias, da Universidade Federal da Paraíba / UFPB.

Orientador(a): Prof. O Dr. Gustavo Correa Xavier

Bananeiras - PB
Maio, 2024

Catalogação na publicação
Seção de Catalogação e Classificação

S237aa Santos, Lucas Souza.
Análise comparativa entre modelos Arima e modelos de Machine Learning para revisão de séries temporais de commodities / Lucas Souza Santos. - Bananeiras, 2024.
25 f. : il.

Orientação: Gustavo Correia Xavier.
TCC (Graduação) - UFPB/CCHSA.

1. Commodities. 2. Séries Temporais. 3. ARIMA. 4. SVR. I. Correia Xavier, Gustavo. II. Título.

UFPB/CCHSA-CHÃ CDU 658 (043)

Elaborado por BRUNA ISABELLE MEDEIROS DE MORAIS - CRB-15/813

Folha de aprovação

Trabalho apresentado à banca examinadora como requisito parcial para a Conclusão de Curso do Bacharelado em Administração

Aluno: Lucas Souza Santos

Trabalho: Análise Comparativa Entre Modelos ARIMA e Modelos de Machine Learning Para Previsão de Séries Temporais de Commodities

Data de aprovação: 10 de maio de 2024

Banca Examinadora

_____

Prof. O Dr. Gustavo Correa Xavier
Orientador(a)

_____

Prof. O Dr. Claudio Germano Dos Santos Oliveira
Membro(a) 1

AGRADECIMENTOS

Após alguns anos de dedicação, chegou este momento tão significante para mim, a conclusão deste ciclo da graduação. Agradeço primeiramente a Deus, “Porque dele, e por meio dele, e para ele são todas as coisas. A ele seja a glória para sempre. Amém!” - Rm 11:36.

Agradeço imensamente aos meus pais, por todo esforço, companheirismo e ensinamentos que me ajudaram a chegar até aqui. Agradeço igualmente aos meus irmãos por sempre me apoiarem e torcerem por mim, sempre me inspirando. E aos amigos feitos nesse caminho, colegas e professores, agradeço por ter compartilhado essa difícil jornada com vocês, com seus conselhos em sala de aula e boas conversas e risadas nos corredores e R.U. Agradeço ao meu orientador pelas oportunidades em projetos e seu auxílio neste trabalho. A todos vocês dedico essa pesquisa de TCC.

"Essencialmente, todos os modelos estão errados, mas alguns são úteis."

(George Edward Pelham Box).

RESUMO

Este estudo aborda a análise comparativa entre modelos Arima e modelos de machine learning para previsão de séries temporais de commodities aplicados à soja, commodity brasileira. A abordagem metodológica deste estudo pode ser categorizada como quantitativa e descritiva. Esta pesquisa se concentra na descrição e análise dos dados relacionados à série temporal de preços da soja. Além disso, a abordagem é quantitativa, uma vez que faz uso de métodos estatísticos e de Machine Learning (ML) para análise preditiva dos dados. Quanto aos modelos econométricos (ARIMA), utilizou-se a metodologia de Box, Jenkins e Reinsel (2008) e para os modelos de ML (Floresta Aleatória e SVR), utilizou-se a metodologia de Parmezan (2016). Os dados foram obtidos no website da CEPEA/ESALQ, usando a série de preços em dólar diferenciados em 1 dia. Tanto os modelos econométricos como os modelos de ML se destacaram com precisão preditiva, obtendo resultado satisfatório, mostrando que simples modelos lineares como o ARIMA ainda são úteis e também o potencial do ML para lidarmos com ST financeiras, tendo a habilidade de antecipar as mudanças com o tempo. Ao averiguar com a revisão inicial da literatura, a escassez de trabalhos aplicando esses métodos ao mercado financeiro brasileiro e a escassez na última década de trabalhos no Brasil sobre commodities, nota-se que há várias direções promissoras para pesquisas futuras. Por fim, foi possível alcançar todos os objetivos deste trabalho e ainda contribuir para o avanço do conhecimento no campo da previsão de séries temporais aplicado a finanças e commodities.

Palavras-Chave: Commodities; Séries Temporais; ARIMA; SVR

ABSTRACT

This study addresses the comparative analysis between Arima models and machine learning models for forecasting commodity time series applied to soybeans, a Brazilian commodity. The methodological approach of this study can be categorized as quantitative and descriptive. This research focuses on the description and analysis of data related to the time series of soybean prices. Furthermore, the approach is quantitative, as it makes use of statistical and Machine Learning (ML) methods for predictive data analysis. As for the econometric models (ARIMA), the methodology of Box, Jenkins and Reinsel (2008) was used and for the ML models (Random Forest and SVR), the methodology of Parmezan (2016) was used. The data were obtained from the CEPEA/ESALQ website, using the series of prices in dollars differentiated in 1 day. Both econometric models and ML models stood out with predictive accuracy, obtaining satisfactory results, showing that simple linear models such as ARIMA are still useful and also the potential of ML to deal with financial ST, having the ability to anticipate changes with the time. When investigating the scarcity of work applying these methods to the Brazilian financial market and the scarcity of work in Brazil on commodities in the last decade in Brazil, it is noted that there are several promising directions for future research. Finally, it was possible to achieve all the objectives of this work and also contribute to the advancement of knowledge in the field of time series forecasting applied to finance and commodities.

Keywords: Commodities; Times Series; ARIMA; SVR

LISTA DE SIGLAS

ADF……..Dickey Fuller aumentado
ALC……...América Latina e Caribe
AR……..….Autor-egressivo
ARIMA……Autorregressivo Integrado de Médias Móveis
B3……Brasil, Bolsa, Balcão
CBOT…….Chicago Board of Trade
CEPEA……Centro de Estudos Avançados em Economia Aplicada
ESALQ…….Escola Superior de Agricultura Luiz de Queiroz
ETF……..Exchange-Traded Fund
ML……...Machine Learning
MFC……...Mercado Futuro de Commodity
MSE…….Erro Médio Quadrático
ST…….…….Série Temporal
SVM……..Máquinas de Suporte Vetorial
SVR…….Máquinas de Suporte Vetorial Regressiva
USP……Universidade de São Paulo

SUMÁRIO

1 INTRODUÇÃO

O mercado financeiro é enormemente vasto em produtos de investimento e instrumentos através dos quais se pode investir, é comum deparar-nos com instrumentos para investir como Ações, ETFs (Exchange-Traded Fund), Derivativos, dentre outros. Um dos produtos financeiros que é de grande importância para as economias e no qual pode-se investir são as commodities (plural de commodity). Uma commodity pode ser grãos, ouro, carne bovina, petróleo, gás natural etc. Contanto que possa ser definida como um ativo físico que possui características padronizadas, de ampla negociação em diversas localidades e que possa ser transportado e armazenado por um longo período sem perda da qualidade, também sendo conhecidos por insumos ou matérias-primas (MOLERO; MELLO, 2021).

Molero e Mello (2021) afirmam ainda que há tipos de mercados diferentes em que as commodities podem ser negociadas, a primeira é no mercado à vista, que envolve entrega física e imediata da mercadoria, a segunda é ser negociada como referência nos mercados derivativos. Segundo Giambiagi (2017) os derivativos são instrumentos financeiros cujo resultado depende da realização de uma variável aleatória, normalmente representado por um preço de um ativo em uma data futura. Eles existem para permitir a transferência de risco entre agentes da economia. Os tipos de derivativos mais comuns são: termo, opção, swap e futuro, sendo este último um dos mais comuns.

Ao longo das últimas décadas vários autores se dedicaram ao estudo e avaliação de produtos financeiros, tal qual as commodities. Sørensen (2002) por exemplo, modelou séries temporais de commodities agrícolas para estimar preços de futuros de cereais como milho, soja e trigo. Richter e Sørensen (2002) estimaram um modelo de volatilidade estocástica de dados de Futuros e Opções da soja. Lima et al. (2010) obteve resultados satisfatórios ao realizar estudos baseados na realização de previsões dentro das subséries decompostas de preços da soja através de Wavelets (onduletas) em conjunto com modelos econométricos e comparando estes com previsões de modelos de Redes Neurais.

Entender e estimar o comportamento dos preços de commodities mostra-se um atraente e relevante campo de estudo e de geração de análises para produtores e investidores. Estes preços variam no tempo, ao passo que suas negociações acontecem no mercado, a essa variação do preço no tempo podemos chamar de *Times Series* ou Séries Temporais (ST). As ST sejam, a temperatura de uma cidade ao longo de um ano, quantidade de vendas em um semestre, ou o preço do índice Ibovespa ao longo de sua existência, são objetos valiosos para extração de insights que ajudem seus stakeholders (partes interessadas) na sua tomada de decisão.

Vários modelos de econometria e de Machine Learning (ML) têm sido aplicados nas análises de séries temporais financeiras, por serem ferramentas interdisciplinares. Uma dessas técnicas mais recentes são, os chamados modelos de Máquinas de Suporte Vetorial (SVM) ou Support Vector Machine no inglês, que é uma técnica de ML aplicada em tarefas de aprendizado supervisionado (KIM, 2003; LAHMIRI, 2013; TAY; CAO, 2001).

Como também as já conhecidas técnicas de econometria aplicadas por autores como Lima et al. (2010) que utilizou modelos ARIMA-GARCH, na modelagem de tendência, sazonalidade e volatilidade. Pai e Lin (2005) que utilizaram um modelo híbrido ARIMA e SVM na previsão de preços de ações.

Assim o objetivo deste estudo é compreender descritivamente séries temporais da commodity brasileira, a soja, para entender como essas características se relacionam e assim definir os modelos de previsão, aplicar as técnicas de decomposição a ST e modelos de previsão através do SVR (Máquinas de Suporte Vetorial Regressiva) um sub-modelo do SVM e comparar esses resultados com a clássica econometria. Para assim compreender e avaliar a validade do uso destas técnicas na previsão de variáveis temporais financeiras de commodities.

Este trabalho encontra-se dividido além desta parte (1) introdutória, em (2) referencial teórico, em (3) método da pesquisa, em (4) análise e discussão dos resultados e em (5) considerações finais.

1.1 OBJETIVOS

1.1.1 Objetivo geral

Explorar a viabilidade da aplicação da metodologia de decomposição e previsão de séries temporais de preços de commodities, por meio de modelos de regressão de Máquinas de Vetor de Suporte (SVM) e compará-lo a modelos econométricos tradicionais.

1.1.2 Objetivos específicos

a) Descrever o comportamento da ST.
b) Realizar uma análise da decomposição das ST da soja, identificando os componentes de tendência, sazonalidade e ruído.
c) Implementar o modelo Máquinas de Suporte Vetorial Regressiva (SVR) para previsão da ST e compará-lo a outros modelos, através de métricas de desempenho, a fim de verificar a viabilidade prática dessa abordagem.

2 REFERENCIAL TEÓRICO

2.1 IMPACTO DAS COMMODITIES NO MERCADO BRASILEIRO

As commodities tiveram forte influência no crescimento econômico brasileiro, durante o famoso boom das commodities que perdurou dos anos 2000-2014. E recentemente, porém, o crescimento desacelerou, em parte como consequência do declínio do preço das commodities (BLANCHARD, 2017). Além do Brasil, toda a América Latina e Caribe (ALC) se beneficiou da formação de capital e crescimento das exportações nesse período. Em seu estudo Kristjanpoller, Olson e Salazar (2016) mostram que as commodities no geral tiveram efeito positivo no crescimento da ALC durante o período de boom, mais especificamente os combustíveis e os produtos industriais.

Mostrando a força da indústria extrativa e agrícola brasileira, as commodities representam hoje cerca de 70% das exportações brasileiras e cresceram 16,8% em volume na comparação entre julho de 2022 e 2023 (FGV IBRE, 2023, p.23).

Compreende-se então a relevância que as commodities têm para a economia nacional e mundial, se tornando assim nas últimas décadas base de vários produtos no mercado financeiro em todo o mundo. Na análise entre esses mercados com o mercado acionário, Brooks e Prokopczuk (2013) concluem que as commodities podem ser um diversificador útil da volatilidade e dos retornos das ações, podendo ser um recurso útil para a constituição de portfólio.

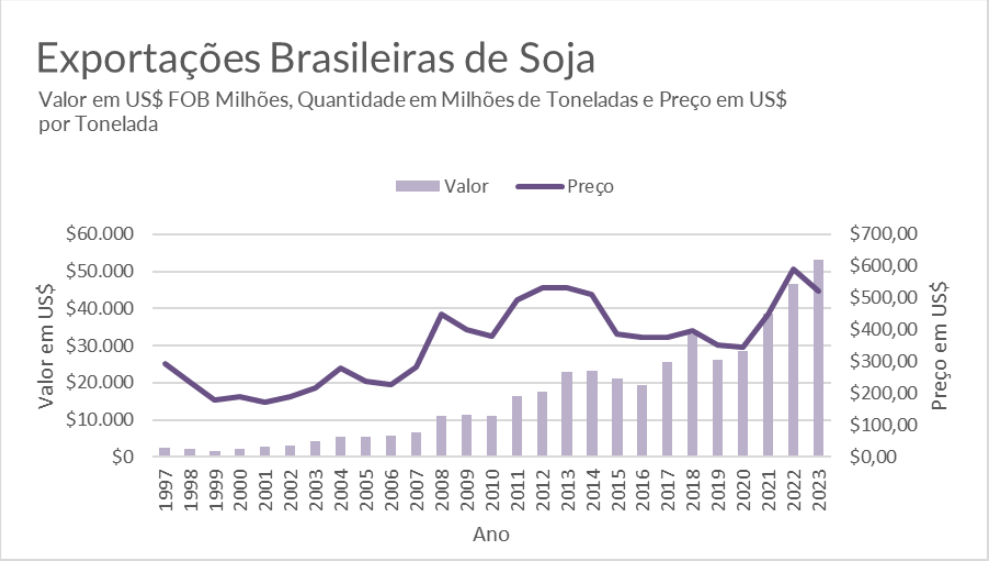
> Uma das áreas de crescimento mais rápido nas finanças empíricas, e uma das menos rigorosamente analisadas, especialmente do ponto de vista da econometria financeira, é a análise econométrica de derivados financeiros. Juntamente com as ações e as obrigações, os derivados constituem a terceira categoria principal de instrumentos financeiros e são normalmente negociados em bolsa ou no mercado de balcão (CHANG; MCALEER, 2015, p.1)

Haase, Zimmermann e Zimmermann (2016) revisaram estudos empíricos sobre o impacto da especulação no Mercado Futuro de Commodity (MFC), e concluem que não há consenso entre os economistas sobre o impacto da especulação no MFC, porém a especulação de preços nesse mercado para commodities agrícolas é vista como uma preocupação, pois pode levar a aumentos nos preços dos alimentos, afetando negativamente as economias menos desenvolvidas.

Assim, é importante para investidores, produtores, consumidores e demais tomadores de decisão ter uma boa compreensão do comportamento dos preços das commodities e de suas interdependências, bem como da sua relação com o mercado acionista (BROOKS; PROKOPCZUK, 2013, p.1).

Como visto anteriormente, existem várias commodities, dentre as que mais se destacam no Brasil estão as agrícolas, principalmente a soja que segundo os dados consolidados de 2023 da Secretaria do Comércio Exterior do Brasil (Comex) é a mais exportada do país estando em 1º no ranking.

Gráfico 01: Exportações brasileiras de soja (1997-2023).



Fonte: Elaboração própria a partir dos dados de (BRASIL, 2023).

O "Gráfico: 01" acima apresenta a evolução do valor da soja brasileira exportada e o preço da tonelada de 1997 a 2023, no consolidado de 2023 ela atingiu seu recorde de US$ 53.244,60 milhões em soja exportada, a um preço de US$ 522,70 a tonelada.

2.2 PREVISÃO DE VARIÁVEIS EM SÉRIES TEMPORAIS

Segundo Petropoulos et al. (2022) apesar da vasta literatura econométrica e de modelos de previsão, há poucos trabalhos até agora sobre a aplicação de tais modelos para commodities agrícolas e suas séries temporais, além de vários desses trabalhos serem recentes, o que indica que existem muitos caminhos abertos para investigação futura sobre estes tópicos e, em particular, para a previsão aplicada.

> Os rápidos avanços na computação permitiram a análise de conjuntos de dados maiores e mais complexos e estimularam o interesse em análise e ciência de dados. Como resultado, a caixa de ferramentas de métodos dos previsores cresceu em tamanho e sofisticação. A ciência da computação abriu o caminho com métodos como redes neurais e outros tipos de Machine Learning, que estão recebendo muita atenção de analistas e tomadores de decisão (PETROPOULOS et al, 2022, p.7).

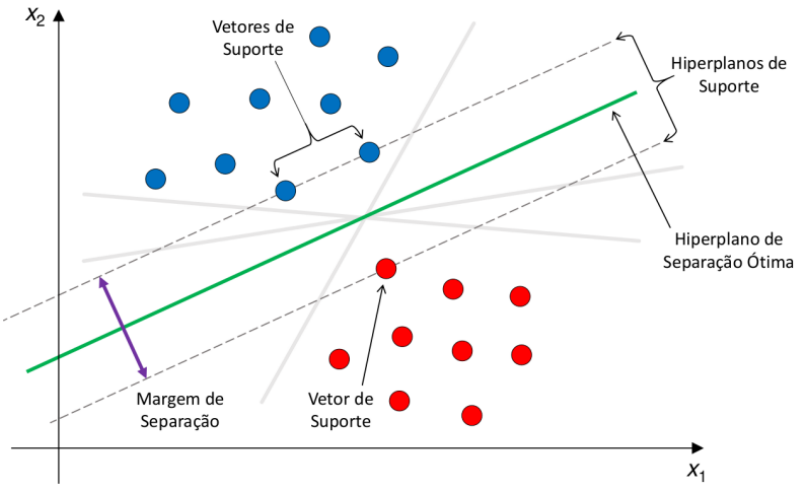
Porém, como Haase, Zimmermann e Zimmermann (2016) ressaltam, a previsão de preços de commodities agrícolas é um desafio por vezes complexo devido sua natureza volátil e à influência de múltiplos fatores, dependendo da commodity analisada. Sendo assim é necessária uma variedade de informações e modelos para obter previsões mais precisas.

Séries temporais financeiras podem gerar vários insights e possibilidades de análise, e no uso financeiro da previsão de preços, Lahmiri (2013) faz previsões da tendência futura do índice de preços do S&P500 por meio de séries temporais de baixa frequência extraídas de uma análise Wavelet.

Existem vários métodos de previsão na econometria e no Machine Learning que têm sido amplamente encontrados na literatura econômica e financeira. Por exemplo, Gujarati, Yamagama e Girvilitto (2019) apresentam alguns dos principais, os quais são (1) modelos de regressão, (2) o processo autorregressivo integrado de médias móveis (ARIMA), popularizado pelos estatísticos Box e Jenkins, também conhecido como metodologia Box-Jenkins (BJ) e (3) vetores autorregressivos (VAR).

Entretanto, Petropoulos et al. (2022) apresentam outros modelos, dentre eles o modelo de Máquinas de Suporte Vetorial Regressiva (SVR) para problemas de previsão, como bastante úteis para previsão de consumo de energia e previsão de preços de petróleo, por forneceram soluções poderosas para reconhecer padrões não lineares e irregulares. O modelo SVR é uma versão do SVM, que foi proposto por Drucker et al (1996).

Figura 01: Hiperplano de separação ótima e seus hiperplanos de suporte. Os eixos ordenados x1 e x2 representam as dimensões das amostras no espaço 2D.



Fonte: (CAMPOS, 2020).

Conforme a “Figura: 01” apresentada acima, é colocado por Campos (2020, p.23) que o conceito de generalização das SVM pode ser exemplificado com uma classificação binária. A partir de duas classes e um conjunto de dados, as SVM determinam o hiperplano que os separa, de maneira a colocar a maior quantidade possível de pontos da mesma classe do mesmo lado.

Tendo em vista a relevância da soja não só para o Brasil, ela já vem sendo abordada em alguns estudos financeiros sobre commodities, mercados de Futuros e em estudos de avaliação de modelos para série temporais. O “Quadro: 01” apresenta um resumo integrado com alguns dos principais estudos revisados sobre análise de séries temporais de commodities, apresentando autores, modelos utilizados e variáveis chave.

Quadro 01: Quadro integrador de revisão da literatura.

Referências	**Modelos**	**Variáveis Utilizadas**
Tay e Cao (2001)	Modelos: SVM; Redes Neurais BP	Preços de Contratos Futuros de Commodities da CBOT
Sørensen (2002)	Modelo: com parâmetros por filtro de Kalman	Preços de Contratos Futuros de Commodities da CBOT
Richter e Sørensen (2002)	Modelo: de volatilidade estocástica	Preços de Contratos Futuros de Commodities da CBOT
Kim (2003)	Modelos: SVM; Redes Neurais BP e CBR	Índices de preços de ações
Pai e Lin (2005)	Modelo: híbrido SVM-ARIMA	Preços de ações
Lima et al. (2010)	Modelos: ARIMA-GARCH; Rede Neural Recorrente	Preços à vista da soja (dados da ESALQ)
Lahmiri (2013)	Modelo: SVM	Índice S&P 500
Zhang et al. (2020)	Modelos: SVM; Random Forest; ELM…	Preços à vista de commodities
Campos (2021)	Modelos: AR; ARIMA; SARIMA; Random Forest; SVM; LSTM	Índice Ibovespa e preços da Vale
Nunes et al. (2023)	Modelos: ARIMA; SVM	Séries do setor elétrico

Fonte: Elaboração própria a partir de revisão da literatura.

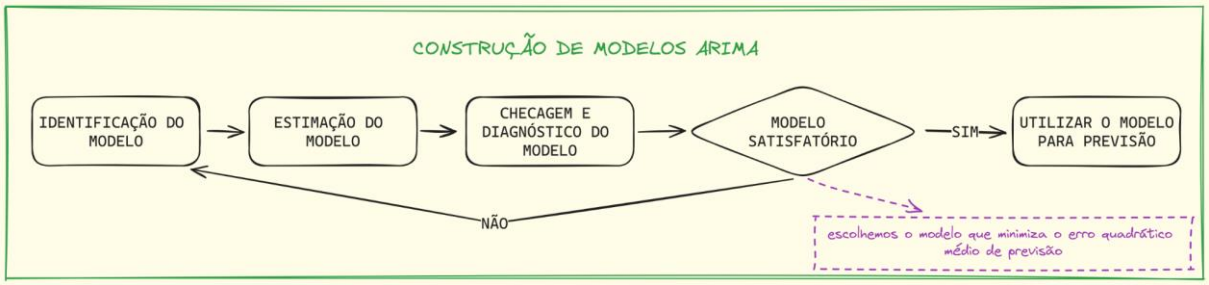
3 MÉTODO DA PESQUISA

A metodologia adotada neste estudo se baseia na aplicação da decomposição de séries temporais e no modelo de regressão do SVM (Support Vector Machine) de aprendizado não supervisionado, mas daqui para frente chamaremos este modelo apenas de SVR, já que usaremos a regressão e não a classificação, com o objetivo de analisar e prever a ST não linear de preços da commodity soja, além de comparar com modelos ARIMA, considerando sacas de 60kg conforme dados fornecidos pela CEPEA/ESALQ/USP. Atualmente, o Cepea e a B3 mantêm parceria para a elaboração e divulgação dos indicadores do Boi Gordo, Bezerro, Milho, Etanol, Açúcar, Soja e Algodão.

A abordagem metodológica deste estudo pode ser categorizada como quantitativa e descritiva. Quanto aos objetivos, esta pesquisa se concentra na descrição e análise dos dados relacionados à série temporal de preços da soja, o que a caracteriza como descritiva. Além disso, a abordagem é quantitativa, uma vez que faz uso de métodos estatísticos e de ML para análise dos dados.

Primeiramente, para os modelos econométricos, utilizou-se a metodologia de Box, Jenkins e Reinsel (2008), o ARIMA (Autorregressivo Integrado de Médias Móveis) para uma ST consiste em três estágios principais conforme apresentado na “Figura: 02”, caso após esses três estágios o modelo seja satisfatório, o mesmo passa a ser utilizado para previsão.

Figura 02: Representação da construção de modelos ARIMA.



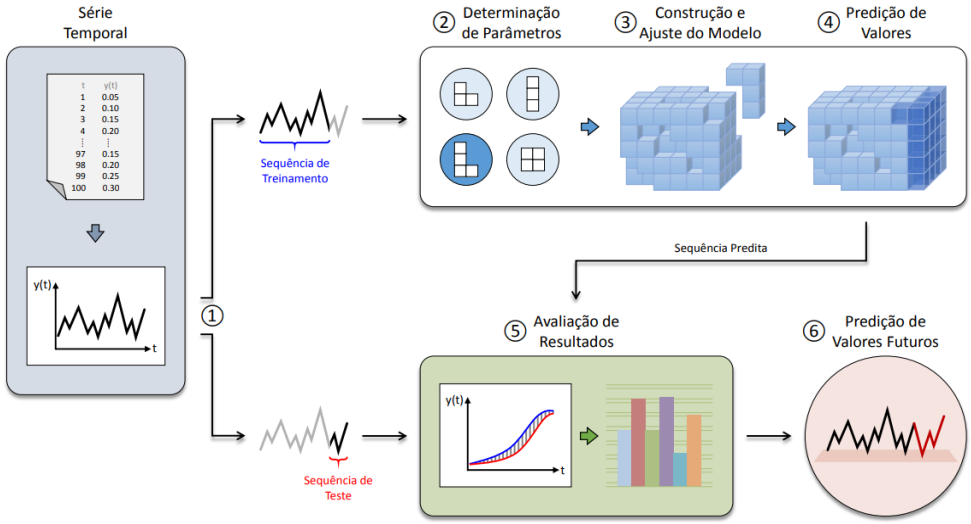
Fonte: Elaboração própria, a partir de Box, Jenkins e Reinsel (2008).

A ordem de elementos dos modelos ARIMA, segundo Campos (2021) são:

- ARIMA(p, 0, 0) = AR(p);
- ARIMA(0, 0, q) = MA(q);
- ARIMA(p, 0, q) = ARMA(p, q);
- No caso do SARIMA, que explora a autocorrelação sazonal, será SARIMA(p,d,q).

Segundamente, na aplicação de Machine Learning a sequência metodológica utilizada partiu da demonstração de Parmezan (2016), também empregada por Campos (2021) conforme ilustrado na "Figura: 03":

Figura 03: Processo de predição de valores em ST com ML.



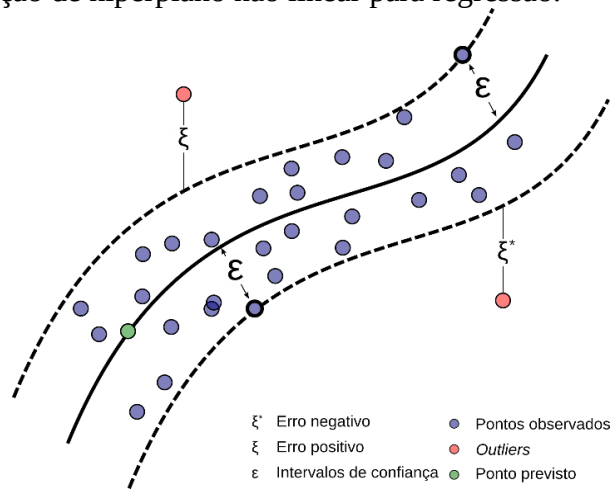
Fonte: (PARMEZAN, 2016).

Principais etapas do processo de previsão com os modelos estatísticos e de ML conforme apresenta Campos (2021):

- Primeira etapa: a ST é dividida em dados de treinamento e teste, neste trabalho os primeiros 80% dos dados da ST são dados de treinamento e a parte final, correspondente a 20%, são dados de testes.
- Segunda etapa: seguindo para a estruturação do modelo. Aqui são encontrados os melhores parâmetros empregando alguma técnica de busca. Usualmente, essa técnica é implementada por um algoritmo que recebe como entrada a sequência.
- Terceira etapa: a partir dos parâmetros identificados na etapa anterior, o modelo de interesse é construído e ajustado aos dados da sequência de treinamento.

- Quarta etapa: é a previsão de valores, ou seja, nesta etapa que são executados os modelos estatísticos e de Machine Learning.
- Quinta etapa: é a etapa de avaliação do modelo. Nesta etapa é realizada a avaliação dos erros de previsão, ou seja, mensurar o quanto os valores da sequência predita se distanciam dos valores da sequência de teste e visualizar através de gráficos adequados.

Figura 04: Representação de hiperplano não linear para regressão.



Fonte: (ALMEIDA; CARVALHO; MENINO, 2017).

No SVR conforme a "Figura: 04" os valores são previstos através dos hiperplanos, que utilizam derivadas parciais para os cálculos dos intervalos de confiança. As margens (ϵ) representam os intervalos de confiança e os vetores de suporte que as delimitam, representam os limites para os erros positivos (ξ) e negativos (ξ^*) (ALMEIDA; CARVALHO; MENINO, 2017).

Para a avaliação do desempenho preditivo, ou seja, a precisão dos modelos, foi utilizado o Erro Médio Quadrático (MSE) amplamente utilizado pela literatura revisada. A seguir estão as fórmulas dos indicadores.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (Yi - \widehat{Y\imath})^2$$

onde: N é o número de períodos de teste; Yi é o valor real e $\widehat{Y\imath}$ é o valor previsto.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Com base nos objetivos estabelecidos anteriormente, esta seção traz a modelagem dos previsores Estatísticos Econométricos e de Machine Learning das séries temporais obtidas.

4.1 OBTENÇÃO, MANIPULAÇÃO E PRÉ-PROCESSAMENTO DOS DADOS

A obtenção dos dados da soja foi realizada a partir do site do CEPEA/ESALQ/USP além de outros dados sobre a commodity obtidos nos sites do IBGE (Instituto Brasileiro de Geografia e Estatística) e do SIDRA. No site do CEPEA foi obtida a tabela 'SOJA ESALQ/BM&FBOVESPA - PARANAGUÁ' contendo informações como data, preço em R$ (reais) e preço em U$ (Dólar), a principal aplicação desse indicador é a liquidação financeira do contrato futuro de soja na BM&FBOVESPA – por código: SFI, além disso representam o mercado à vista. A série de preços obtida contém todo o período desde quando a CEPEA

começou a medi-lo em 13/03/2006, até 28/12/2023 quando os dados foram coletados, somando quase dezessete anos, possuindo uma granularidade diária, ou seja, dados de fechamento de todos os dias em que houve negociação.

Após a obtenção dos dados, a tabela foi carregada dentro do *VScode* (que é um interpretador de código fonte) usando a linguagem de programação *Python*, onde foi feita a configuração da coluna 'Data' como um elemento de data e em granulação diária, para que o *Python* entenda os dados como uma série temporal, esta etapa é muito importante para evitar erros futuros de execução. Pelo fato dos preços estarem indexados ao tempo (diário), faz-se necessário os valores ausentes nos dias em que não houve negociação, para isso foi utilizado o método '*interpolate(method='time')*' do Python que preenche os valores ausentes de forma proporcional ao intervalo de tempo entre o índices.

4.2 ANÁLISE EXPLORATÓRIA DA SÉRIE

A variável alvo utilizada foi o preço em dólar, por possuir melhores características de uma ST para análise preditiva. Vejamos algumas estatísticas da variável na "Tabela: 01":

Tabela 01: Resumo das estatísticas relacionadas a variável utilizada.

Medidas	**Variável: preço (U$)**
Média	26,04
Mediana	25,16
Moda	20,98
Desvio Padrão	6,15
Mínimo	12,40
Máximo	45,32

Fonte: Elaboração própria a partir dos dados do (CEPEA/ESALQ/USP, 2023).

Em resumo as estatísticas da variável são: média de 26,04; mediana de 25,16; moda de 20,98; desvio padrão de 6,15; mínimo de 12,40 e máximo de 45,32. Essas estatísticas também mostram que a série em Dólar é melhor que a série em Real, pelo fato da segunda, respectivamente, sofrer mais com variações e distúrbios macroeconômicos.

Ao processar a série notou-se que a diferenciação da série dos preços em Dólar trouxe melhores resultados que a própria série dos preços. A diferenciação é um método para deixar séries não ou pouco estacionárias em estações, em que cada valor na nova série é calculado como a diferença entre o valor original na série temporal e seu valor anterior.

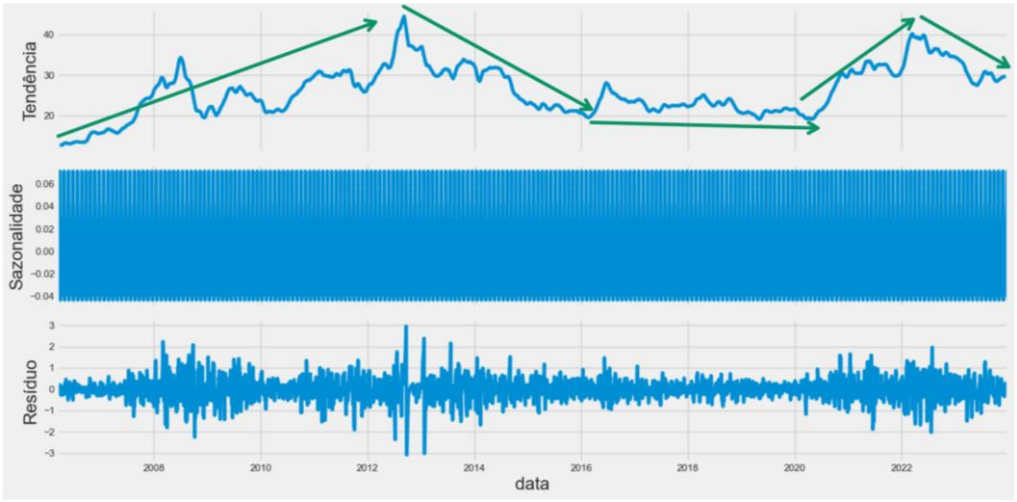
Gráfico 02: Série original de preços da soja brasileira em dólar (2006-2023).

Fonte: Elaboração própria a partir dos dados do (CEPEA/ESALQ/USP, 2023).

A partir do "Gráfico: 02" acima, é possível observar algumas características como que: a uma tendência de alta longa entre 2006 e 2013, uma tendência de baixa entre 2013 e 2016 e uma lateralização entre 2016 e 2020; seguida de uma tendência de alta durante o período da pandemia de Covid-19 entre 2020 e início de 2022 devido a alta generalizada das commodities, seguida de mais uma tendência de baixa curta pós pandemia, isso fica mais evidente na decomposição da série.

Assim como a distribuição, em ST é imprescindível a análise da decomposição da série em tendência, sazonalidade e ruído para que possamos distinguir as características da ST, o "Gráfico: 03" demonstra esses três componentes. Além da tendência já mencionada podemos observar que não há uma sazonalidade pouco distinguível, a qual aparenta ser de um período igual ou inferior a 30 dias e ainda um ruído elevado e frequente nos períodos de pico de alta ou baixa das tendências.

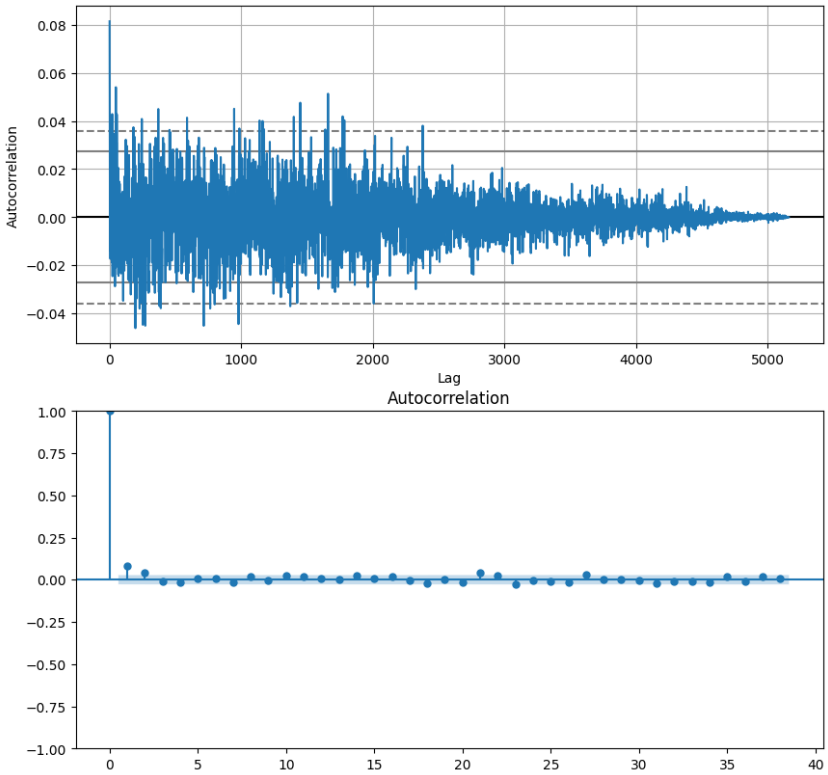
Gráfico 03: Decomposição da série de preços da soja brasileira em dólar.



Fonte: Elaboração própria a partir dos dados do (CEPEA/ESALQ/USP, 2023).

Devemos analisar também nesta parte de exploração da ST, a autocorrelação dos dados, correlação significa a relação de um valor com outro (não implicando causa e efeito necessariamente). Nesse caso foi utilizado a função de autocorrelação ACF, comparando um valor presente com um valor do passado da mesma série, com intervalo de confiança de 95%.

Gráfico 04: Avaliação da autocorrelação da ST (frequência diária).



Fonte: Elaboração própria a partir dos dados do (CEPEA/ESALQ/USP, 2023).

O Gráfico 04, apresenta duas configurações gráficas, ambas representam a autocorrelação da nossa ST diferenciada, o eixo x contém o lag ou os dias e o eixo y as autocorrelações. Nota-se que: as autocorrelações decaem ao longo do tempo e que as linhas diárias (linhas verticais) se contêm majoritariamente dentro da banda de confiança, na parte superior do gráfico representada pelas linhas horizontais em cinza.

4.3 VERIFICANDO A ESTACIONARIEDADE DA ST

Segundo Box, Jenkins e Reinsel (2008), os modelos estacionários assumem que o processo permanece em equilíbrio estatístico, ou seja, média e variância da série temporal não mudam ao longo do tempo. Por outro lado, muitas séries temporais na indústria, negócios e economia não se encaixam nesse modelo estacionário. Elas são chamadas de não estacionárias, como a ST em real ou pouco estacionária como a ST em Dólar.

Tabela 02: Resultados do teste de Dickey Fuller aumentado (ADF).

Teste de Dickey-Fuller	**Resultados**
Test Statistic	-53.01
p-value	0.0
Lags Used	1.0
Number of Observations Used	6448.0
Critical Value (1%)	-3.43
Critical Value (5%)	-2.86
Critical Value (10%)	-2.57

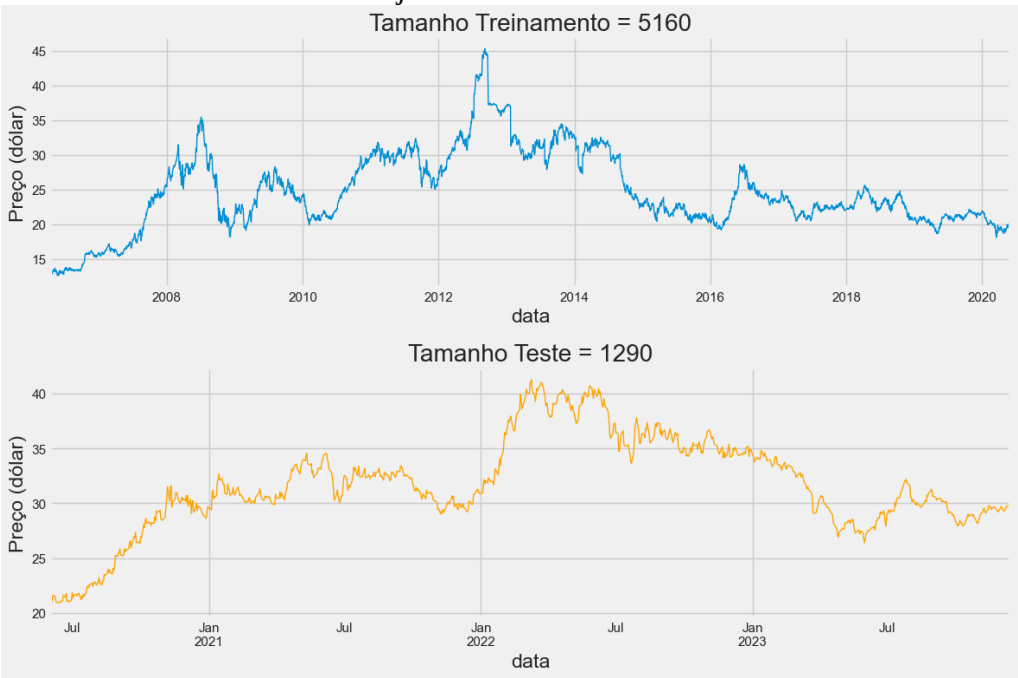
Fonte: Elaboração própria a partir dos dados do (CEPEA/ESALQ/USP, 2023).

Porém como podemos ver na “Tabela: 02”, com a aplicação do teste de Dickey Fuller aumentado (ADF), da biblioteca *Stats Models* do *python*, a ST dos preços da soja em dólar depois de ser diferenciada se mostrou estacionária com ‘p-value > 0.05’ e ‘test statistic > critical value (1%)’. Assim como foi notado na autocorrelação.

4.4 DIVISÃO E TREINAMENTO DOS DADOS

No desenvolvimento de modelos de previsão de ST uma das etapas que estão sempre presentes é a divisão dos dados, e a metodologia de divisão utilizada aqui foi, conjuntos de dados de treino e teste. Optou-se pela proporção de divisão convencional, normalmente utilizada nestes casos, que foi de 80% dos dados para treinamento e 20% para testes. Os “Gráficos: 05” a seguir, mostram como ficaram os conjuntos divididos.

Gráfico 05: Tamanho da série em conjuntos de treinamento e de teste.



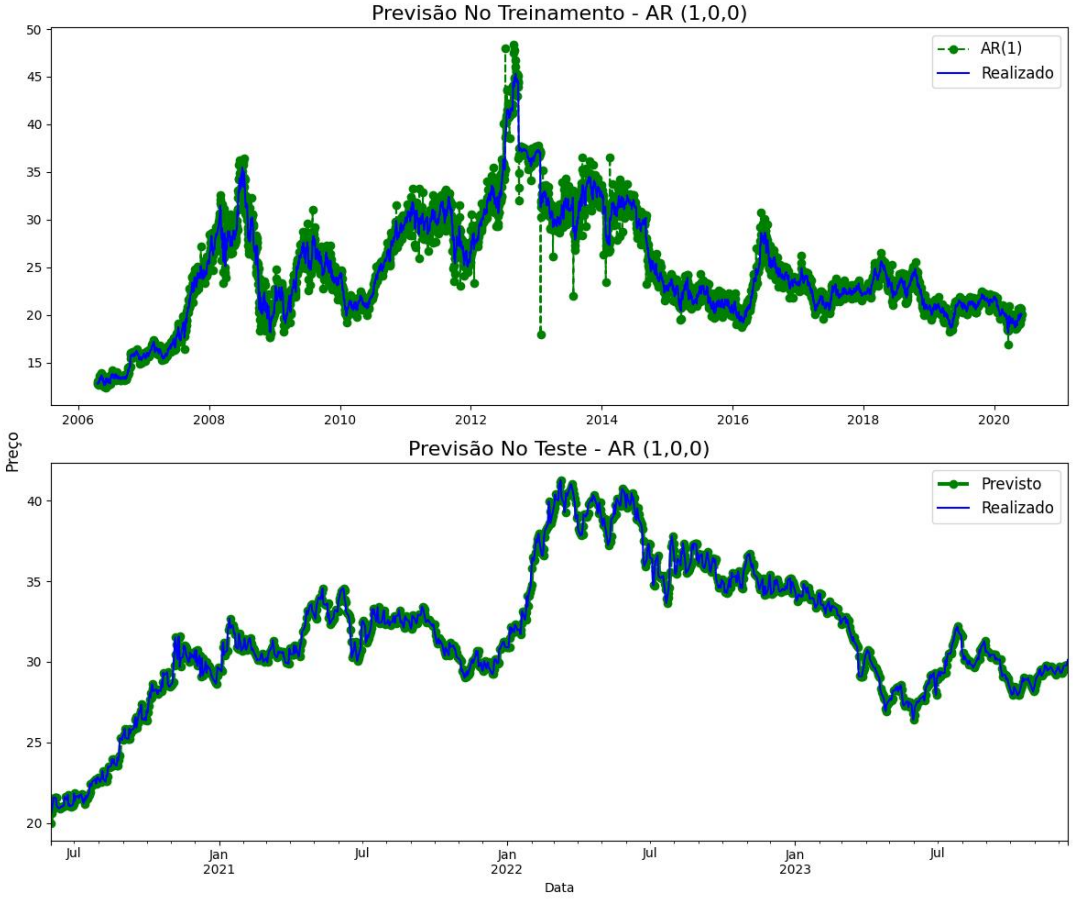
Fonte: Elaboração própria a partir dos dados do (CEPEA/ESALQ/USP, 2023).

Após a divisão dos dados, o conjunto de treinamento apresentou um tamanho de 5.160 negociações (dias) e o conjunto de teste apresentou 1.290 negociações (dias).

4.5 IMPLEMENTAÇÃO DOS MODELOS ECONOMÉTRICOS

Para implementação do modelo auto-regressivo (AR), o número de defasagens (ou lags) analisado pelo gráfico da autocorrelação é de $p=1$. O “Gráfico: 06” demonstra a comparação entre os dados realizado e previsto para o modelo AR, no qual foi obtido um MSE = 0,1130, ou seja, um erro quadrático médio baixo, tendo bom desempenho. Apesar de não conseguir acertar os pontos de máximo e mínimo relevantes.

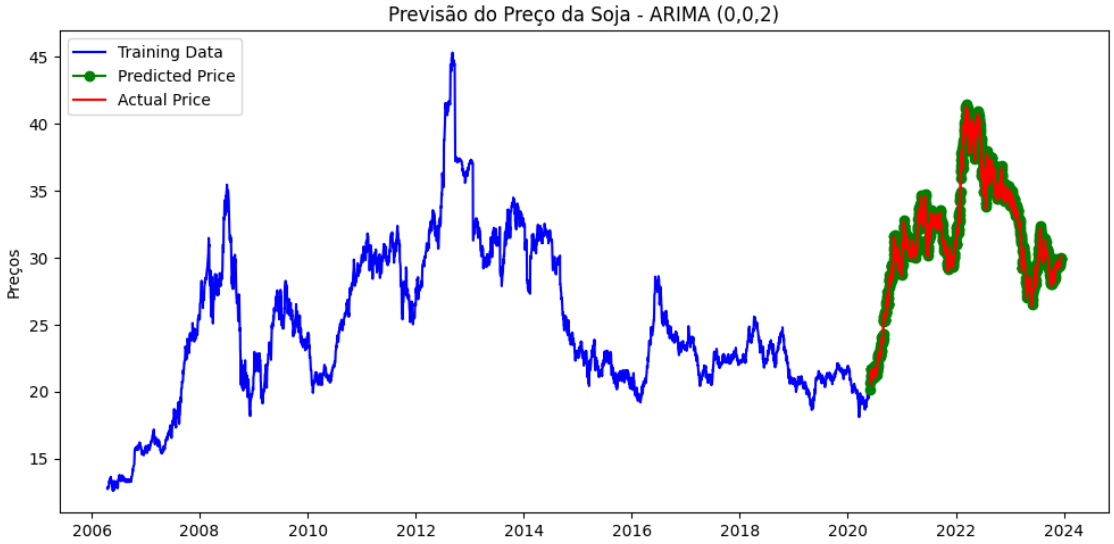
Gráfico 06: Comparação entre dados de testes e dados previstos pelo modelo AR.



Fonte: Elaboração própria a partir dos dados do (CEPEA/ESALQ/USP, 2023).

Já o modelo ARIMA recebeu como parâmetros (p, d, q) os valores de (0, 0, 2). Foi obtido um MSE = 0.1129, ou seja, um erro quadrático médio baixo, tendo bom desempenho, assim como o AR, veja no "Gráfico: 07":

Gráfico 07: Comparação entre dados de testes e dados previstos pelo modelo ARIMA.



Fonte: Elaboração própria a partir dos dados do (CEPEA/ESALQ/USP, 2023).

Assim, podemos concluir que os modelos econométricos clássicos da família ARIMA, são modelos lineares simples, que apesar de simples ainda se mostram eficientes. E ao contrário do resultado obtido por Campos (2021) que não obteve bom desempenho com os modelos ARIMA na previsão de preços de fechamento de ações e do Ibovespa, estes modelos se mostraram eficientes em sua aplicação a séries de commodities em dólar, devidamente transformadas em estacionárias.

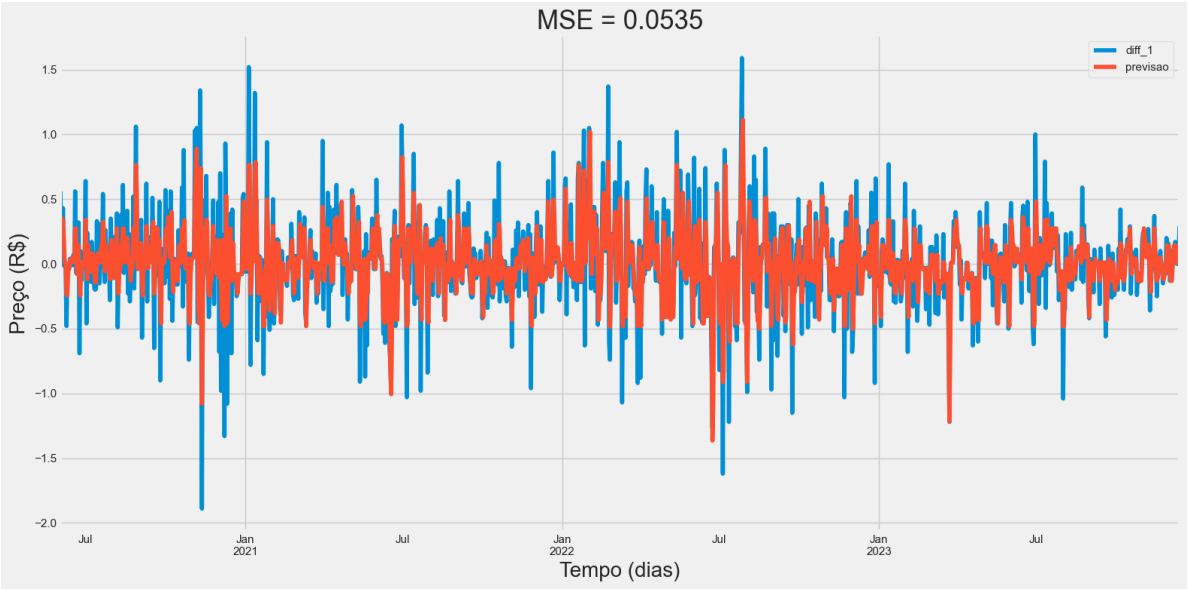
4.6 IMPLEMENTAÇÃO DOS MODELOS DE MACHINE LEARNING

Como a preparação dos dados para os modelos econométricos e de ML diferem um pouco e queremos compará-los diretamente, foram geradas novas características que serviram como entrada nos modelos de Floresta Aleatória e SVR. Seguindo a demonstração de Campos (2021) as novas características são as seguintes:

- Resíduos: dados obtidos através da decomposição da ST.
- Tendência: dados obtidos através da decomposição da ST.
- Sazonalidade: dados obtidos através da decomposição da ST.
- Diferença 1, 2, 3, 4 e 5: dados obtidos calculando a diferença do preço do dia atual menos o preço de um, dois, três, quatro e cinco dias atrás.

O modelo de Floresta Aleatória recebeu como parâmetros, 1000 estimadores e árvores com profundidade máxima de 5 níveis. O MSE foi de 0,0535, métricas baixas, ou seja, melhor previsto. O “Gráfico: 08” compara os dados de teste e os dados previstos do modelo, lembrando que neste gráfico a série está no nível de diferenciação.

Gráfico 08: Comparação entre dados de testes e dados previstos pelo modelo de Floresta Aleatória.

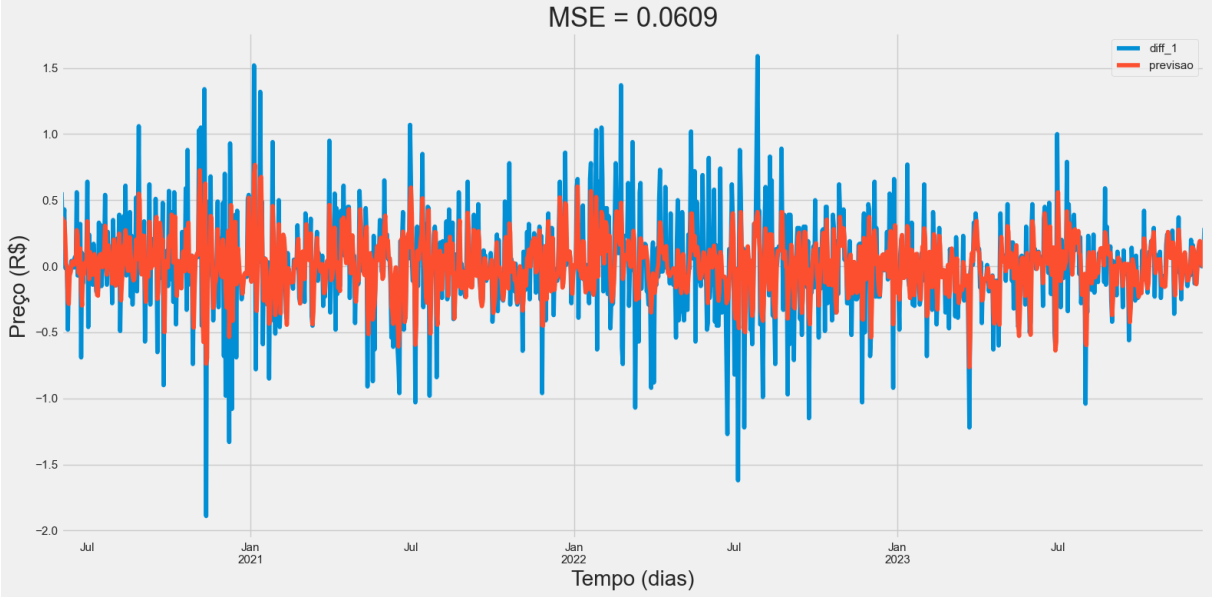


Fonte: Elaboração própria a partir dos dados do (CEPEA/ESALQ/USP, 2023).

As florestas aleatórias são uma escolha sólida para prever séries temporais financeiras, especialmente no contexto de previsão de preços, devido à sua interpretabilidade e à facilidade de compreensão dos resultados. O modelo de floresta aleatória aplicado neste trabalho demonstrou habilidade em identificar a tendência geral do movimento, mas falhou em capturar picos de alta e baixa, em que há mais ruído de informação.

O modelo de SVR foi configurado com o kernel radial e gamma de 0,1 para garantir uma boa generalização do modelo. O MSE foi de 0,0609. O “Gráfico: 09” compara os dados de teste e os dados previstos.

Gráfico 09: Comparação entre dados de testes e dados previstos pelo modelo SVR.



Fonte: Elaboração própria a partir dos dados do (CEPEA/ESALQ/USP, 2023).

O SVR se mostrou semelhante ao Floresta Aleatória, apesar de demonstrar uma precisão na previsão um pouco menor. No entanto, o desafio reside na amplitude desses movimentos na série temporal (ST). A máquina de vetores de suporte (SVR) conseguiu identificar corretamente os períodos de alta e baixa na série original dos preços nos nossos testes, mas não tanto com a sua diferenciação em um dia.

Tabela 03: Resultado comparativo da precisão dos modelos.

Modelo	**MSE**
AR	0,1130
ARIMA	0,1129
Random Forest	0,0535
SVR	0,0609

Fonte: Elaboração própria a partir dos dados do (CEPEA/ESALQ/USP, 2023).

A “Tabela: 03” acima, integra os resultados de MSE obtidos com cada modelo, onde o modelo de melhor desempenho foi o Random Forest corroborando com as conclusões de outros estudos que mostraram os modelos de ML como mais eficientes, outra coisa a ser mencionada é que estes modelos também são mais rápidos.

5 CONSIDERAÇÕES FINAIS

Ao longo deste estudo, exploramos a complexidade da modelagem e previsão de séries temporais, em nosso contexto, dos preços da soja em dólar, obtidos na CEPEA/ESALQ, empregando tanto métodos econométricos quanto técnicas de Machine Learning.

Iniciamos com uma análise exploratória da série temporal, revelando padrões significativos, como tendências predominantes de alta ao longo dos anos. A decomposição da série em tendência, sazonalidade e ruído nos proporcionou insights valiosos sobre a estrutura subjacente dos dados, permitindo uma compreensão mais profunda da escolha dos modelos preditivos a serem utilizados, além das flutuações do mercado. Nesse aspecto a série em dólar se mostrou melhor que a série em Real para análise, pelo fato da segunda, respectivamente, sofrer mais com variações e distúrbios macroeconômicos, e pré-processamos a série em diferenciação de um dia para torná-la estacionária.

A aplicação de modelos econométricos tradicionais, como AR e ARIMA, apesar de ter limitações na captura de padrões complexos e variações abruptas, se mostrou eficiente em nossa abordagem. Já os modelos de Machine Learning, como Floresta Aleatória e SVR, emergiram como destaques em nossa busca por previsões com menor erro. A integração de características derivadas da decomposição da série temporal permitiu que esses modelos capturassem nuances sutis nos dados e produzissem previsões mais confiáveis. Mostrando o potencial do ML para lidarmos com ST financeiras.

Averiguou-se com a revisão inicial da literatura, a escassez de trabalhos aplicando esses métodos ao mercado financeiro brasileiro e a escassez na última década de trabalhos no Brasil sobre commodities. Há várias direções promissoras para pesquisas futuras. Recomenda-se explorar ainda mais o uso de modelos de Machine Learning avançados, como redes neurais profundas e SVR. Além disso, investigações sobre a inclusão de variáveis externas, como fatores climáticos e políticas governamentais, podem enriquecer nossa compreensão da dinâmica dos preços dos ativos financeiros e melhorar a precisão das previsões. Além de revisões da literatura que tragam o Estado da Arte deste campo de estudo no Brasil atualmente.

Este estudo contribui para o avanço do conhecimento no campo da previsão de séries temporais aplicado a finanças, fornecendo insights valiosos sobre as metodologias e as melhores práticas para lidar com dados complexos e voláteis.

Em resumo, este trabalho representa um recorte daquilo que a aplicação da inteligência artificial pode gerar no estudo de séries temporais financeiras, e ainda como as commodities podem conter insights pouco explorados neste campo. Esperamos que este estudo sirva como uma peça importante com contribuições adicionais nesta área dinâmica.

REFERÊNCIAS

ALMEIDA, A; CARVALHO, F; MENINO, F. **Introdução ao Machine Learning:** de alunos para alunos. [S. L.]: Dataat, 2017.

BLANCHARD, Olivier. **Macroeconomia**. 7. ed. São Paulo, SP: Pearson, 2017. E-book. Disponível em: https://plataforma.bvirtual.com.br. Acesso em: 19 ago. 2023.

BOX, George E. P.; JENKINS, Gwilym M.; REINSEL, Gregory C. **Time Series Analysis**: forecasting and control. 4. ed. Hoboken, Nj: John Wiley & Sons, Inc, 2008.

BRASIL. Mdic. Secretaria de Comércio Exterior. **Comex Vis**. Disponível em: http://comexstat.mdic.gov.br/pt/comex-vis. Acesso em: 01 set. 2023.

BROOKS, C; PROKOPCZUK, M. The dynamics of commodity prices. **Quantitative Finance**, [S.L.], v. 13, n. 4, p. 527-542, abr. 2013. Informa UK Limited. http://dx.doi.org/10.1080/14697688.2013.769689.

CAMPOS, B. A. R de. M. **Análise Comparativa de Técnicas para a Previsão de Séries Temporais no Contexto de Mercado Financeiros**. 2020. 76 f. TCC (Graduação) - Curso de Sistemas de Informação, Departamento de Informática e Estatística, Universidade Federal de Santa Catarina, Florianópolis - Sc, 2021.

CEPEA/ESALQ/USP – Centro de Estudos Avançados em Economia Aplicada / Escola Superior de Agricultura Luiz de Queiroz / Universidade de São Paulo. (2023). https://www.cepea.esalq.usp.br/br.

CHANG, C; MCALEER, M. Econometric analysis of financial derivatives: an overview. **Journal Of Econometrics**, [S.L.], v. 187, n. 2, p. 403-407, ago. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.jeconom.2015.02.026.

DRUCKER, H. et al. **Support Vector Regression Machines**. Em: NIPS. 3 dez. 1996. Disponível em: <https://www.semanticscholar.org/paper/Support-Vector-Regression-Machines-Drucker-Burges/e52fb14e4beccc5e88a33c1fe5c7d6e780831ae1>. Acesso em: 4 set. 2023.

FGV IBRE. **Boletim Macro**. 2023. Disponível em: https://portalibre.fgv.br/sites/default/files/2023-08/2023%2008%20Boletim%20Macro_1.pdf. Acesso em: 27 ago. 2023.

GIAMBIAGI, F. **Derivativos e Risco de Mercado**. Rio de Janeiro - RJ: Grupo GEN, 2017. E-book. ISBN 9788595154742.

GUJARATI, D; YAMAGAMI, C; VIRGILITTO, S. B. **Econometria**. São Paulo - SP: Editora Saraiva, 2019. E-book. ISBN 9788553131952.

HAASE, M; ZIMMERMANN, Y. S; ZIMMERMANN, H. The impact of speculation on

commodity futures markets – A review of the findings of 100 empirical studies. **Journal Of Commodity Markets**, [S.L.], v. 3, n. 1, p. 1-15, set. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.jcomm.2016.07.006.

KIM, K. Financial time series forecasting using support vector machines. **Neurocomputing**, [S.L.] v. 55, n. 1–2, p. 307–319, set. 2003. Elsevier BV. http://dx.doi.org/10.1016/s0925-2312(03)00372-2.

KRISTJANPOLLER, W; OLSON, J. E; SALAZAR, R. I. Does the commodities boom support the export led growth hypothesis? Evidence from Latin American countries. **Latin American Economic Review**, [S.L.], v. 25, n. 1, 20 out. 2016. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s40503-016-0036-z.

LAHMIRI, S. Forecasting Direction of the S&P 500 Movement Using Wavelet Transform and Support Vector Machines. **International Journal of Strategic Decision Sciences**, [S.L.], v. 4, n. 1, p. 79–89, 1 jan. 2013. IGI Global. http://dx.doi.org/10.4018/jsds.2013010105.

LIMA, F. G. et al. Previsão de preços de commodities com modelos ARIMA-GARCH e redes neurais com ondaletas: velhas tecnologias - novos resultados. **Revista de Administração**, [S.L.], v. 45, n. 2, p. 188-202, abr. 2010. Elsevier BV. http://dx.doi.org/10.1016/s0080-2107(16)30537-4.

NUNES, L. R. M. et al. Uso do ARIMA e SVM para previsão de séries temporais do sistema elétrico brasileiro. **Research, Society And Development**, [S.L.], v. 12, n. 3, p. 1-16, 25 fev. 2023. Research, Society and Development. http://dx.doi.org/10.33448/rsd-v12i3.40438.

MOLERO, L; MELLO, E. M. **Derivativos – Negociação e precificação 2ª edição**. São Paulo - SP: Saint Paul Publishing (Brazil), 2021. E-book. ISBN 9786586407150.

PAI, P; LIN, C. A hybrid ARIMA and support vector machines model in stock price forecasting. **Omega**, [S.L.], v. 33, n. 6, p. 497-505, dez. 2005. Elsevier BV. http://dx.doi.org/10.1016/j.omega.2004.07.024.

PARMEZAN, Antonio Rafael Sabino. **Predição de séries temporais por similaridade**. 2016. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2016. doi:10.11606/D.55.2016.tde-21112016-150659. Acesso em: 2024-03-23.

PETROPOULOS, F. et al. Forecasting: theory and practice. **International Journal of Forecasting**, [S.L.], v. 38, n. 3, p. 705-871, jul. 2022. Elsevier BV. http://dx.doi.org/10.1016/j.ijforecast.2021.11.001.

RICHTER, M. C.; SØRENSEN, C. Stochastic Volatility and Seasonality in Commodity Futures and Options: The Case of Soybeans. **SSRN Electronic Journal**, [S.L.], 2002. Elsevier BV. http://dx.doi.org/10.2139/ssrn.301994.

SØRENSEN, C. Modeling seasonality in agricultural commodity futures. **Journal Of Futures**

Markets, [S.L.], v. 22, n. 5, p. 393-426, 7 mar. 2002. Wiley. http://dx.doi.org/10.1002/fut.10017.

TAY, F. E. H; CAO, L. Application of support vector machines in financial time series forecasting. **Omega**, [S.L.], v. 29, n. 4, p. 309-317, ago. 2001. Elsevier BV. http://dx.doi.org/10.1016/s0305-0483(01)00026-3.

ZHANG, D. et al. Forecasting Agricultural Commodity Prices Using Model Selection Framework with Time Series Features and Forecast Horizons. **Ieee Access**, [S.L.], v. 8, p. 28197-28209, 2020. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/access.2020.2971591.