UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS AGRÁRIAS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA RURAL

SÁVIO MEDEIRO VIANA

PREVISÃO DE PREÇOS DE SOJA NO BRASIL POR MEIO DE TÉCNICAS DE *MACHINE LEARNING*

FORTALEZA
2024

SÁVIO MEDEIRO VIANA

PREVISÃO DE PREÇOS DE SOJA NO BRASIL POR MEIO DE TÉCNICAS DE *MACHINE LEARNING*

Dissertação apresentada à Coordenação do Programa de Pós-Graduação em Economia Rural do Centro de Ciências Agrárias da Universidade Federal do Ceará, como requisito parcial para obtenção do título de Mestre. Linha de Pesquisa: Economia Aplicada ao Agronegócio.

Orientador: Prof. Vitor Hugo Miro Couto Silva.

FORTALEZA

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

V668p Viana, Sávio Medeiro.
Previsão de preços de soja no Brasil por meio de técnicas de machine learning / Sávio Medeiro Viana. – 2024.
58 f.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências Agrárias, Programa de Pós-Graduação em Economia Rural, Fortaleza, 2024.
Orientação: Prof. Dr. Vitor Hugo Miro Couto Silva.

1. Commodities agrícolas. 2. Previsão de preços. 3. Machine learning. 4. Soja. I. Título.
CDD 338.1

SÁVIO MEDEIRO VIANA

PREVISÃO DE PREÇOS DE SOJA NO BRASIL POR MEIO DE TÉCNICAS DE *MACHINE LEARNING*

Dissertação apresentada à Coordenação do Programa de Pós-Graduação em Economia Rural do Centro de Ciências Agrárias da Universidade Federal do Ceará, como requisito parcial para obtenção do título de Mestre. Linha de Pesquisa: Economia Aplicada ao Agronegócio.

Aprovada em: 29/10/2024.

BANCA EXAMINADORA

_____
Prof. Dr. Vitor Hugo Miro Couto Silva (Orientador)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Francisco José Silva Tabosa
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Leandro de Almeida Rocco
Universidade Federal do Ceará (UFC)

_____
Prof.a Dra. Isadora Gonçalves Costa Osterno
Universidade de Fortaleza (UNIFOR)

A Deus.
Aos meus pais, Regivaldo e Izanira.

AGRADECIMENTOS

A Deus, por ser minha fortaleza e refúgio, e ser o Pai que nunca abandona. Mesmo nas dificuldades e nas dúvidas, sempre está presente comigo, me guiando e me fortalecendo nessa grande jornada chamada vida, em que sempre estou em constante mudança.

A Ele sou reconhecido pela existência e a graça de me proporcionar a vida dos meus pais, Iza e Regivaldo, que, mesmo diante de todas as dificuldades, estabeleceram um lar com afeto e união, qual aprendi a importância da família e da perseverança. Juntamente com meus irmãos, Caio, Davi e Miguel, vocês todos são a minha base, porto seguro e meus maiores exemplos de amor. Sou grato aos amigos e colegas que estiveram junto a mim nesta etapa do Mestrado. Em particular, agradeço aos que adentraram comigo essa peleja: Leopodina, Ronaldo e Nilson. Os que vieram depois: Vitória Biana (amiguinha desde a graduação), Ivan, Moisés, Ramon, Ana Cristina, Mauro, Vitória Oliveira, Karla, Edglê, Tanara, Marcos Paulo, Leudiane e tantos outros no dia a dia acadêmico.

Sobra aqui, também, uma menção honrosa aos, não acadêmicos, mas que me concederam e concedem todo o apoio na amizade: Lucas Andrade (meu amiguinho), Nícolas, Darlyson, Lorena, Mylena, Felipe Mesquita, Ádamo, Bia, Fábio, Marto, Felipe Lima, Alex, Bruno, Natália, Kleison, Jauro. E outros com os quais, à extensão desse decurso, jamais perdi o contato, e que sempre torceram por mim: Caio, Samuel Levi, Niqueline, Thiago, Fernanda, Paulo Victor, Weider, Thales, Guilherme, Pedro Reis, Indira, Felipe Jucá, Armando, Luan e muitos outros. Minha eterna gratidão à Universidade Federal do Ceará, onde concluí a Graduação em Ciências Econômicas e remato, agora, o Mestrado em Economia Rural. Ao Departamento de Economia Agrícola. Aos professores doutores Kilmer, Edward, Nilton, Taboza, Cristiano e outros. A Carlene, Roberta, Margareth e a todas as pessoas do âmbito administrativo. Ao professor doutor Rafael, da Faculdade de Economia, Administração, Atuária e Contabildade – FEAAC da UFC, que me motivou a adentrar o PPGER.

Um agradecimento, também, mais do que especial, à CAPES, pelo incentivo a este trabalho e toda a sua importância no fomento da pós-graduação no Brasil.

Ao professor doutor Vitor Hugo Miro, pela excelente orientação. Sua dedicação, paciência e conhecimento foram indescartáveis em todo o processo. Agradeço pela confiança depositada em mim, as críticas construtivas, os incentivos constantes e por todos os ensinamentos que levarei para a vida. Inspirou-me a ser um pesquisador melhor e me proporcionou uma experiência acadêmica enriquecedora. Sou profundamente grato por sua orientação e amizade.

“É necessário sair da ilha para ver a ilha. ”

José Saramago – *O Conto da Ilha Desconhecida.*

RESUMO

As *commodities* agrícolas detêm função importante na economia global, sendo estes produtos responsáveis por cerca de 40% das exportações agrícolas do Brasil. A previsão de preços nesse segmento é um desafio importante, pois habilitada a influenciar positiva e significativamente os resultados da atividade, como os lucros dos produtores e a competitividade do setor, bem assim as estratégias de empreendedores e investidores. Este experimento acadêmico investiga e desenvolve modelos de previsão de preços de *commodities* no Brasil, com aplicação de métodos avançados de *machine learning*, em específico, os preços de soja no Brasil. Objetiva contribuir para o entendimento de como a aplicação de técnicas de aprendizado de máquina aprimora as previsões de preços de *commodities*, assumindo a hipótese de que tais modelos proveem performance preditiva em relação aos modelos mais tradicionais de séries temporais. A justificativa é fornecer às partes interessadas informações críticas para a tomada de decisões informadas e o fortalecimento da economia do País. Para certificar este objetivo, foram explorados modelos de séries temporais, como o SARIMA, e padrões de *machine learning*, como os baseados em árvores de decisão (Random Forest) e métodos de regularização (LASSO). Os resultados obtidos com a aplicação de variados modelos indicaram que, embora o SARIMA capte a sazonalidade dos preços, os de aprendizado de máquina, especialmente o Lasso, cheguem quase à precisão e sejam eficientes para prever o comportamento do preço da soja no Brasil.

Palavras-chave: c*ommodities* agrícolas; previsão de preços; *machine learning*; soja.

ABSTRACT

Agricultural commodities play an important role in the global economy, with these products accounting for approximately 40% of Brazil's agricultural exports. Price forecasting in this segment is a major challenge, as it has enabled a positive and significant influence on the results of the activity, such as producer profits and the competitiveness of the sector, as well as the strategies of entrepreneurs and investors. This academic experiment investigates and develops commodity price forecasting models in Brazil, with the application of advanced machine learning methods, specifically, soybean prices in Brazil. It aims to contribute to the understanding of how the application of improved machine learning techniques based on commodity prices, assuming the hypothesis that such models prove predictive performance in relation to more traditional time series models. The justification is to provide stakeholders with critical information for informed decision-making and strengthening the country's economy. To certify this objective, time series models, such as SARIMA, and machine learning standards, such as those based on decision trees (Random Forest) and regularization methods (LASSO), were explored. The results obtained with the application of different models indicated that, although SARIMA captures the seasonality of prices, machine learning, especially Lasso, will reach near accuracy and will be efficient in predicting the behavior of soybean prices in Brazil.

Key words: agricultural commodities; price forecast; machine learning; soy.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O mercado de *commodities* ocupa estratégica posição na economia global, dando ensejo a influências profundas em distintos setores e na dinâmica econômica de inúmeros países. Essa atividade é essencial para a estruturação dos fluxos de comércio internacional, influenciando, não apenas, a oferta e a demanda, mas também a volatilidade dos mercados financeiros e, por extensão, o crescimento econômico. No que se refere às *commodities* agrícolas, seu papel é ainda mais relevante, dado o vínculo direto com a produção e oferta de alimentos, afetando pontos críticos de segurança alimentar e nutricional. A volatilidade nesses mercados amplifica desafios, especialmente em economias emergentes e dependentes da exportação de produtos primários. Flutuações nos preços quase sempre mexem no equilíbrio comercial e na estabilidade socioeconômica, não apenas moldando as relações comerciais externas, mas, também, adaptando a distribuição de riqueza e a segurança econômica de várias regiões.

Característica relevante, e que justifica diversos estudos sobre o tema, está no fato de que esses mercados exprimem grandes instabilidades que, geralmente, são observadas na volatilidade de preços das mercadorias transacionadas. Além dos elementos comuns na formação de preços, como em qualquer outro comércio, os preços de *commodities* diversas são negativamente influenciados pela negociação de contratos em mercados financeiros. Com efeito, os ativos respondem não apenas aos condicionantes da dinâmica entre oferta e demanda, mas, também, são influenciados por uma combinação de variegados fatores que incluem componentes econômicos, políticos e sociais (Banco Mundial, 2023).

No Brasil, a produção e o comércio de *commodities* cumpriu um papel indispensável na história, ainda sendo por demais relevante para a economia do País, além de determinar as dinâmicas sociais e culturais. A instabilidade e a volatilidade nos preços destes produtos, *in hoc sensu,* afetam diretamente nas condições de vida dos agricultores, no acesso a alimentos e nos índices de inflação (Silva e Oliveira, 2022). Em ocorrendo assim, compreender os mecanismos que influenciam os valores de tais recursos faz-se relevante para desenvolver estratégias eficazes para gestão de riscos, formulação e análise de políticas públicas e estratégias de negócios. Logo, a projeção de preços das *commodities* agrícolas possui a função de guiar a tomada de decisão de agentes distintos.

Com efeito, a capacidade de prever com precisão os preços de *commodities* tornou-se um desafio de subida importância, e, nesse contexto, o avanço das técnicas de aprendizado de máquina revoluciona a capacidade de análise e previsão. Esses atributos concedem oportunidade à exploração de grandes volumes de indicativos, identificação de padrões ocultos

e adaptação a mudanças nas condições do mercado (Oliveira e Silva, 2022). Em expressas circunstâncias, o objetivo deste ensaio é aplicar modelos de previsão de preços de *commodities* no Brasil, comparando métodos já bem estabelecidos neste ambiente investigativo, com a metodologia de *machine learning.*

De sorte a exprimir melhor delimitação deste texto dissertativo, optou-se por aplicar essa proposta para realizar previsões do preço da soja, em vista da sua relevância para a economia brasileira, representando cerca de 40% das exportações agrícolas do País (MAPA, 2023). Por certo, a precificação da soja é um fator importante para o setor agropecuário, pois perturba os lucros dos produtores, a competitividade das indústrias e o bem-estar dos consumidores. Embora exista alguma literatura dedicada a este esforço de desenvolver e aplicar modelos de previsão, têm curso algumas lacunas importantes a respeito da contribuição dos modelos de aprendizado de máquina para a melhora dos modelos preditivos, o que justifica a abordagem adotada nesta Dissertação de Mestrado.

Outros estudos, como o de Souza (2021), avaliaram o desempenho de variadas técnicas de aprendizado de máquina na previsão de preços e demandas de produtos do varejo, ao passo que Carvalho (2021) estimou o desempenho de diversos algoritmos de *machine learning* na previsão do preço do contrato futuro de milho, indicando que o algoritmo ARIMA obteve os melhores resultados para um período de até 60 dias. Já Lima e Pereira (2021) aplicaram redes neurais artificiais para prever os preços de ações. Por isso, este exercício inquisitivo de ordem acadêmica inova, ao oferecer uma demanda orientada para um produto específico, utilizando-se de indicativos históricos de preços de soja, em que os resultados foram comparados com os outros obtidos por diversos métodos de previsão por meio de *machine learning*.

Foram explorados vários modelos de árvores de decisão (Random Forest), os de regularização (LASSO) e os de séries temporais (SARIMA), para determinar qual deles é mais adequado para prever os distintos preços de soja no contexto brasileiro. Portanto, a investigação agora relatada e sustida no concerto acadêmico da UFC se propõe contribuir significativamente para o entendimento e a aplicação de técnicas de aprendizado de máquina na previsão de preços de soja no Brasil, com o objetivo de fornecer às partes interessadas informações críticas para a tomada de decisões informadas e o fortalecimento da economia do País.

A Dissertação encontra-se organizada conforme está à continuidade. A seção 2, imediatamente após a Introdução (unidade capitular 1), descreve o levantamento bibliográfico para a completude do trabalho proposto. O segmento 3 introduz, a seu turno, a metodologia e os dados editados úteis. O módulo 4, por sua vez, contém os resultados e discussões. No plano

capitular 5, estão as considerações finais, acompanhadas da imediata relação das obras às quais se recorreu a fim de conceder substrato empírico e teórico à demanda efetivada.

2 REVISÃO DE LITERATURA

A literatura sobre previsões de preços oferece uma variedade de estratégias para que se compreendam o comportamento e a evolução dos valores de *commodities*, empregando uma diversidade de técnicas - e mesclando-as - com o propósito de alcançar prognósticos mais acurados. Com efeito, a dinâmica entre o preço futuro e o preço à vista, conhecida como base, é um pilar para a compreensão do mercado (Silveira, 2017). Essa diferença de preços, que também se manifesta entre contratos futuros com vencimentos distintos, é objeto de estudo de duas teorias proeminentes: a Teoria da Estocagem e a Teoria do *Normal Backwardation*.

Decerto, a Teoria do *Normal Backwardation,* proposta por John Maynard Keynes em 1930, postula o argumento de que, em condições normais, os preços futuros tendem a ser inferiores àqueles à vista. Essa diferença se justifica pela procura dos produtores por proteção contra a incerteza dos custos que estão por vir, oferecendo seus produtos a valores mais baixos em contratos posteriores. Em contrapartida, os especuladores, atraídos pela possibilidade de ganhos na posteridade, adquirem esses contratos, impulsionando a demanda e equilibrando o mercado (Keynes, 1930).

De outra vertente, a Teoria da Estocagem, proposta inicialmente por Nicholas Kaldor, em 1939, e posteriormente desenvolvida por parte de Holbrook Working, em 1948, atribui a diferença entre os preços futuros e à vista aos custos de armazenamento das *commodities*. Se os custos de armazenamento forem altos, os preços futuros tenderão a ser mais elevados do que aqueles à vista, refletindo essa despesa adicional (Kaldor, 1939; Working, 1948).

Fama e French (1987), em um estudo abrangente com 21 *commodities*, encontraram maior suporte empírico para a Teoria da Estocagem, indicando que a variação da taxa de juros é uma grande fração da variação da base, especialmente para produtos agrícolas.

Nessa direção, um conceito-chave na Teoria da Estocagem é o rendimento de conveniência, que se relaciona inversamente ao nível de estoque. A bibliografia explora a relação entre estoques e preços, evidenciando que os estoques atuam como "amortecedores" (Deaton e Laroque, 1992) e que as dinâmicas de preços, produção e estoques são inter-relacionadas (Pindyck, 2001). Estudos empíricos com diversas *commodities*, como metais industriais (Ng e Pirrong, 1994) e metais comuns (Omura e West, 2015), toparam resultados consistentes em relação à Teoria da Estocagem, com rendimentos de conveniência negativamente relacionados aos níveis de estoque.

No contexto brasileiro, pesquisas sobre a base em *commodities* agrícolas investigaram a relação entre preços à vista e futuros, a efetividade do *hedge* e o risco de base, especialmente

em mercados como soja e boi gordo. Resultados como os de Costa *et al*. (2006) e Gonçalves *et al*. (2007) e Tonin, Tonin e Tonin (2011) indicam que o mercado futuro é passível de ser líder de preços e que há uma relação de bi causalidade entre os preços futuro e à vista. Demais disso, o comportamento da base e seu risco em diversificadas localidades e períodos do ano foram analisados (Fontes, Castro Junior e Azevedo, 2005; Barros e Aguiar, 2005; Fernandes, Lima e Aguiar, 2005), demonstrando a importância da sazonalidade na formação da base.

Estudos no Brasil, também, testaram a Teoria do Normal Backwardation (Santos, 1998; Ende, 2002; Pereira, 2009) e a relação entre a base e o rendimento de conveniência (Corsini e Ribeiro, 2008; Pinto, 2015), com resultados variados e por vezes contraditórios. A pesquisa de Pinto (2015), no entanto, sobre o mercado de petróleo ianque encontrou resultados compatíveis com a tradicional Teoria da Estocagem, com o benefício de conveniência marginal sendo uma função decrescente do nível de estoques.

A literatura internacional mais recente sobre a base e o rendimento de conveniência abrange modelos teóricos e empíricos que exploram a relação entre preços à vista e futuros, estoques, volatilidade e microestrutura do mercado (Casassus e Collin-Dufresne, 2005; Gorton, Hayashi e Rouwenhorst, 2013; Brooks, Prokopczuk e Wu, 2013; Frankel, 2014; Joseph, Irwin e Garcia, 2015; Dockner, Eksi e Rammerstorfer, 2015; Milonas e Paratsiokas, 2017; Kim, Kim e Heo, 2017; Nakajima, 2017). Esses estudos evidenciam a complexidade da dinâmica da base e a importância de considerar múltiplos fatores em sua análise, incluindo a heterogeneidade dos agentes de mercado (Nakajima, 2015) e a microestrutura do mercado (Dockner, Eksi e Rammerstorfer, 2015).

Nessa conjuntura, a literatura sobre previsão de preços no mercado de *commodities* abrange um vasto conjunto de modelos e técnicas, desde abordagens econométricas tradicionais, como modelos de equilíbrio parcial e de ordem estrutural (Pindyck, 2001; Working, 1948) até métodos mais recentes baseados em aprendizado de máquina. As pesquisas intentam aprimorar a compreensão da dinâmica complexa dos preços das *commodities*, considerando fatores como oferta e demanda, custos de armazenamento, eventos climáticos e variáveis macroeconômicas (Pindyck, 2001; Fama e French, 1987; Working, 1948). A combinação de técnicas diversas e a utilização de modelos híbridos, como o Filtro de Kalman (Corsini e Ribeiro, 2008), exprimem-se como promissoras na demanda por previsões precisas e robustas.

Nessas circunstâncias, a aplicação de técnicas de *machine learning* tem ganhado destaque na literatura, impulsionada pela crescente disponibilidade de dados e pelo avanço da capacidade computacional. Algoritmos como redes neurais artificiais (Pinto, 2015), árvores de

decisão e modelos de *ensemble learning* são aplicados para capturar padrões complexos e não lineares nos dados, potencialmente superando as limitações dos modelos econométricos tradicionais. Adicionalmente, a flexibilidade e a capacidade de aprendizado contínuo dos modelos de *machine learning* propiciam a incorporação de novas informações em tempo real, o que se torna cada vez mais relevante em um mercado dinâmico e globalizado como o de *commodities*.

De fato, Mohanty; Thakurta; Kar (2023) oferecem um *framework* de *machine learning* para previsão de preços de *commodities* agrícolas composto por quatro blocos funcionais: previsão dos rendimentos das culturas, determinação da oferta, presciência da demanda e conjectura do preço. O trabalho se diferencia por abordar uma técnica eficiente baseada em aprendizagem de máquina para obter o erro reduzido na previsão de preços, usando a oferta e a demanda daquela cultura específica. Os indicativos foram obtidos de várias fontes, incluindo Ministério da Agricultura, Pecuária e Abastecimento (MAPA), Instituto Brasileiro de Geografia e Estatística (IBGE) e Instituto Nacional de Meteorologia (INMET). O conjunto inclui históricos de preços, rendimentos, condições climáticas e outros fatores para as *commodities* arroz, trigo e milho. Esse período vai de 2000 a 2022 e foi dividido em dois blocos: um conjunto de treinamento, no qual se utiliza para treinar o modelo de previsão, e uma conjunção de testes, empregando-se para avaliar o desempenho do modelo.

O modelo utilizado no artigo é o de regressão de árvore de decisão, com vistas a prever valores contínuos, como o preço de uma *commodity*. Com efeito, a técnica empregada para avaliar o desempenho do modelo é o erro médio quadrático (RMSE). O RMSE é uma medida de quanto o modelo está longe dos valores reais. Desse modo, os resultados indicaram que o modelo de regressão de árvore de decisão foi capaz de prever o preço das *commodities* agrícolas, com um erro médio de 5%. Os autores postularam o argumento de que a técnica de regressão de árvore de decisão é conveniente para a previsão, porque é capaz de capturar a complexidade das relações entre os fatores que influenciam os preços. Além deste modelo, o artigo também avalia dois outros: o modelo de regressão linear, que prevê o preço com base na oferta e na demanda, e o de séries temporais, antevendo o preço com base em dados históricos.

Os resultados, no entanto, mostraram que a regressão de árvore de decisão foi superior à dos outros dois modelos. Primeiro, pela capacidade de capturar a complexidade das relações entre os fatores que influenciam os valores. Os modelos de regressão linear e de séries temporais são métodos simples, assumindo que as relações entre os fatores são lineares ou que seguem um padrão previsível. As conexões que influenciam, entretanto, os preços de *commodities* agrícolas são, muitas vezes, complexas e não lineares. A árvore de decisão é habilitada a

capturar essa complexidade, resultando em melhores previsões. Segundo, a base de dados de alta qualidade, que incluía dados históricos de preços, rendimentos, condições climáticas e outros fatores, deu ensejo a que o modelo 'aprendesse' as relações entre os fatores de maneira mais achegada à precisão. Assim, de maneira mais específica, os resultados de Mohanty; Thakurta; Kar (2023) mostraram que o modelo de regressão de árvore de decisão teve um erro médio quadrático (RMSE) de 5%, enquanto o modelo de regressão linear experimentou um RMSE de 7% e o modelo de séries temporais registrou um RMSE de 9%. Assim tendo sido, o estudo concluiu que o *framework* proposto é uma ferramenta valiosa para a previsão de preços de *commodities* agrícolas, suscetível de ser utilizado por agricultores, *traders* e outros agentes do mercado agrícola para tomar decisões informadas sobre a produção, comercialização e investimento.

Com base um conjunto de abordagens de previsão híbridas, Palazzi *et al* (2023) denotam uma combinação entre Análise Espectral Singular (SSA) com distintos métodos de séries temporais univariados, alternando métodos de sazonalidade complexa a aprendizado de máquina e modelos autorregressivos para prever preços *spot* mensais de milho, soja e açúcar no Brasil. Os autores acompanham a arquitetura de dados de Mohanty; Thakurta; Kar (2023), empregando um conjunto de dados que vai de 2000 a 2022, divididos em dois conjuntos semelhantes, o de teste e o de treinamento. No artigo, todavia, o conjunto de treinamento inclui dados de preços *spot*, rendimentos, condições climáticas e outros fatores para os anos de 2000 a 2018, ao passo que o conjunto de testes inclui dados de preços *spot*, rendimentos, condições climáticas e outros fatores para os anos de 2019 a 2022. Ambos os dados foram coletados das fontes MAPA, IBGE e INMET.

O modelo de SSA, por sua vez, conforma uma técnica de decomposição de séries temporais utilizável para identificar e remover componentes de tendência, sazonalidade e ruído. Os métodos de sazonalidade complexa foram utilizados para capturar padrões de sazonalidade não lineares e dinâmicos. Por outra perspectiva, os métodos de aprendizado de máquina são utilizados para aprender as relações entre os fatores que influenciam os preços das *commodities* agrícolas, sendo que os modelos autorregressivos são aplicados para prever o preço futuro com base nos preços passados. Os modelos híbridos SSA a que o artigo recorreu são os seguintes: SSA-ARIMA, que combina a SSA com o modelo ARIMA (Autoregressive Integrated Moving Average); o SSA-GARCH, que cruza a SSA com o modelo GARCH (Generalized Autoregressive Conditional Heteroskedasticity); o SSA-ANN, na qual condiz a SSA com uma rede neural artificial (ANN); e o SSA-SVM, unindo a SSA com uma máquina de vetores de suporte (SVM).

Os resultados do trabalho indicaram que os modelos híbridos SSA superaram aqueles de referência em termos de precisão de previsão, uma vez que foram capazes de capturar a complexidade da sazonalidade dos preços das *commodities* agrícolas e se mostraram robustos a choques externos. Dessa maneira, na *commodity* milho, o modelo híbrido SSA-ANN exprimiu o melhor desempenho, com um erro médio quadrático (RMSE) de 5,2%, ao passo que o modelo de referência - o ARIMA - apontou um RMSE de 7,1%. Já na soja, o modelo híbrido SSA-SVM exprimiu a melhor performance, com um RMSE de 4,7%, enquanto o modelo de referência - GARCH – denotou um RMSE de 6,3%. Por último, no açúcar, a SSA-ARIMA apontou o melhor desempenho, com um RMSE de 3,8%, enquanto seu o modelo ARIMA exibiu um RMSE de 5,1%. Os autores concluem que os modelos híbridos SSA são uma ferramenta valiosa para a previsão de preços de *commodities* agrícolas no Brasil porque estes são capazes de capturar a complexidade da sazonalidade dos preços e são robustos a choques externos.

Levando em consideração fatores externos, que se mostram relevantes no mercado agrícola, Oktoviany, Knobloch e Korn (2021) propõem um modelo híbrido de duas etapas, baseado em métodos de aprendizado de máquina para agrupamento e classificação. Decerto, foram utilizados dados futuros de milho e fatores externos no período de janeiro de 2015 a dezembro de 2019, sendo o de treinamento compreendido de 01-01-2015 a 31-12-2018, e o de teste de 01-01-2019 a 31-12-2019. De tal modo, primeiramente, eles atribuíram estados de preços a valores históricos usando agrupamento K-*means*; em segundo lugar, as previsões dos estados de preços futuros são, pois, obtidas com suporte em previsões de curto prazo dos fatores externos por meio de K-vizinhos mais próximos ou de classificação florestal aleatória. Com efeito, os autores aplicaram o modelo a dados reais de futuro de milho, produzindo situações de preços por meio de uma simulação de Monte Carlo, comparando com Sørensen (Mark, 2002). Desse modo, obtiveram melhor aproximação dos preços futuros reais pelos preços futuros simulados no que diz respeito às medidas de erro MAE, RMSE e MAPE.

Os autores optaram por modelos híbridos em decorrência da relação não linear entre os fatores que determinam os preços e o processo de preços, em que os modelos lineares de séries temporais não são capazes de capturar totalmente as relações complexas. Com esse intento, essas abordagens híbridas que combinam modelos estatísticos clássicos com métodos de máquina desenvolvidas nesse experimento deram ensejo a uma caracterização de variados comportamentos de preços e desenvolvimento em distintos estados de valores, fragmentados em três componentes principais. Na primeira etapa, foi aplicado o algoritmo *K-means* a retornos históricos de *log* de preços futuros para a identificação de estados de retornos logarítmicos, em que os resultados do agrupamento servem, então, como dados de entrada para as tarefas

subsequentes de classificação e calibração. No segundo componente, foi vinculada uma seleção de fatores externos dos preços das mercadorias, usando um algoritmo de classificação que atribui cada conjunto de valores dos fatores externos a um estado dos preços das mercadorias. Nesta etapa, foram considerados os dois algoritmos de classificação K-vizinhos mais próximos (K-nn) e florestas aleatórias (RF) com dados externos sobre clima e dados de oferta e demanda no cultivo de milho como fatores externos.

Em síntese, o modelo proposto previu, com a requerida precisão o estado dos preços de *commodities* agrícolas, atingindo uma média de 75% de acerto em todos os três insumos estudados (milho, soja e açúcar). Em adição, também demonstrou alta sensibilidade e especificidade nas previsões. Sensibilidade, aqui, se refere à capacidade do modelo de identificar corretamente os verdadeiros positivos - isto é, prever um aumento no estado do preço quando realmente ocorre. Entrementes, especificidade concerne à competência do modelo de apontar de modo correto os verdadeiros negativos. *In aliis verbis*, significa prever uma diminuição no estado do preço quando realmente ocorre. Assim, foram alcançadas uma sensibilidade de 80% e uma especificidade de 70%, indicando sua capacidade de identificar com precisão aumentos e diminuições no estado do preço. Outrossim, o estudo comprovou a competência de previsão em variados horizontes de tempo, desde curto prazo (um mês) até extenso tempo (um ano). Isso o torna útil para uma variedade de aplicações, como planejamento de produção, comércio e investimento, além de indicar robustez ao ruído nos dados.

Com uma premissa semelhante, Yuan, San e Leong (2020) propuseram outra abordagem de aprendizado de máquina para melhorar a precisão da previsão de preços de *commodities* agrícolas na Malásia. Em ultrapasse à da abordagem envolvendo o uso de algoritmos de aprendizado de máquina, os autores inovam quando se louvam numa nova função de seleção de tempo de defasagem para identificar os períodos de descompassos ideais entre dados históricos e preços futuros. Eles utilizaram dados mensais de preços da borracha, do óleo de palma e do cacau, de janeiro de 2000 a dezembro de 2018. Em complementação, aproveitaram indicadores de vários fatores externos considerados como influentes nos preços das *commodities* agrícolas, entre eles as taxas de câmbio (USD/MYR), os valores globais do petróleo bruto (USD/barril) e os índices da Oscilação Sul (ENOS). Nessa contextura, o trabalho consiste em dois componentes principais: i) o modelo de aprendizado de máquina utilizado foi um *ensemble* de redes neurais artificiais e máquinas de vetor de suporte, sendo este capaz de capturar as relações complexas entre os preços das *commodities* e os vários fatores externos; ii) a função de seleção de tempo de defasagem foi utilizada para identificar os períodos de defasagem ideais entre dados históricos e preços futuros. De fato, a função proposta foi baseada

em uma técnica de otimização evolutiva, que encontrou uma solução ótima para o problema de seleção de tempo de defasagem.

Dessa maneira, o modelo proposto superou significativamente os métodos tradicionais de previsão, como ARIMA e GARCH. Ademais, a função de seleção de tempo de defasagem referido foi um componente importante do estudo, contribuindo expressivamente para sua precisão, uma vez que este capturou as relações entre os preços das *commodities* e os fatores externos. Em aditamento, o modelo foi avaliado usando uma variedade de métricas de desempenho, incluindo erro médio absoluto (MAE), erro médio quadrático de raiz (RMSE) e erro médio percentual absoluto (MAPE). Os resultados mostraram que o modelo atingiu elevado nível de desempenho em todas as métricas consideradas.

Em suma, a revisão de literatura aqui operada foi reveladora de um panorama abrangente das principais teorias relacionadas à precificação de *commodities*, com foco em modelos econométricos e métodos de aprendizado de máquina. Consoante lobrigado até aqui, abordaram-se temas como a dinâmica entre preço futuro e preço à vista, a teoria da estocagem, o rendimento de conveniência, a base em *commodities* agrícolas no Brasil e a aplicação de técnicas de *machine learning* na previsão de preços. Embora esta Dissertação compartilhe com a literatura revisada, o interesse na previsão de preços de *commodities* e na aplicação de *machine learning*, nesse contexto, ele se diferencia em diversos aspectos.

A abordagem aqui adotada se concentra na aplicação de modelos de aprendizado de máquina, especificamente Random Forest e LASSO, para prever os preços da soja no Brasil, considerando sua importância para a economia patrial. Os resultados expressos são originais e comparam o desempenho preditivo desses modelos com o padrão de ordem tradicional SARIMA, contribuindo para o entendimento e aplicação de técnicas de *machine learning* na previsão de preços da soja e fornecendo informações relevantes para a tomada de decisões no setor.

Cuida-se, na seção imediatamente próxima, da metodologia empregada na pesquisa sob relação, detalhando os dados utilizados, as etapas de pré-processamento e o estabelecimento dos modelos de previsão. Estão expressas as técnicas de seleção de variáveis, sendo mostrados os métodos de avaliação de desempenho e as ferramentas computacionais às quais se recorreu.

3 METODOLOGIA

A metodologia ora sugestionada tem como objetivo geral treinar um modelo para realizar previsões do preço da soja com a melhor performance preditiva, avaliando se modelos de aprendizado de máquina produzem previsões mais próximas da precisão do que aqueles tradicionais de séries temporais.

Aplica-se, aqui, um modelo *Seasonal Autoregressive Integrated Moving Average* (SARIMA), dadas as características da série temporal empregada e os objetivos da Dissertação. Este cumpriu o papel de *benchmark*, ou seja, um modelo de referência para avaliar se ferramentas de aprendizado de máquina contribuem para produzir previsões melhores. Neste texto, verificam-se distintas opções de modelo, mas a escolha é se concentrar na aplicação de dois modelos baseados em algoritmos de *machine learning:* o Random Forest e o LASSO.

Seguiu-se um fluxo de trabalho estruturado e comum na aplicação de modelos de *machine learning*, que envolve, *ab initio,* a coleta dos indicativos brutos, passando por etapas de preparação dos dados até a modelagem, validação e implantação do modelo para gerar previsões. De modo sintético, breve descrição da sequência empregada está contida à continuação.

- Com base na série temporal de preços da soja, novos atributos são criados com amparo nos dados originais, com vistas a prover variáveis preditoras ao modelo e aprimorar a sua capacidade preditiva. Isso inclui o cálculo de variáveis que insertam médias móveis e volatilidade, entre outros.
- Os dados foram divididos em conjuntos de treino (e validação) e teste. Essa divisão é necessária para avaliar a capacidade do modelo de generalizar para novos dados.
- Aplicação dos algoritmos de *machine learning,* usando o conjunto de dados de treino. Durante essa fase, o modelo ajusta seus parâmetros para minimizar determinado critério de erro. Como descrito acima, optou-se por efetivar a aplicação de um modelo SARIMA, tradicional em análises e previsões com suporte em séries temporais, e dois modelos com apoio em algoritmos de *machine learning*, o Random Forest e o LASSO.
- No caso dos modelos de *machine learning*, foram empregadas técnicas de validação cruzada e ajuste de *hiperparâmetros*. Estes são configurações não aprendidas diretamente pelos dados, mas exercem grande influência sobre o desempenho do modelo.
- O modelo treinado é avaliado usando o conjunto de testes. Para avaliar o desempenho, foram calculadas e comparadas distintas métricas de avaliação. Na situação sob exame,

para mensurar a performance preditiva do modelo, recorreu-se ao erro quadrático médio (e raiz quadrada) e ao erro absoluto médio.

No caso de um modelo de séries temporais, a temporalidade é cuidadosamente considerada na divisão entre conjuntos de treinamento e de teste, na engenharia de atributos e nos processos de modelagem e validação cruzada.

De modo a otimizar o desempenho dos modelos, foram estabelecidas séries derivadas, obtidas com algumas transformações sobre os dados originais. Tais transformações são benéficas, tanto para visualização, quanto para a modelagem de séries temporais. A aplicação de funções como a média móvel, por exemplo, é útil para identificar variações de alta magnitude na série temporal, uma etapa crítica nos modelos da família ARIMA (Ahmed *et al*, 2010).

Outro tipo de transformação suscetível de ser implementada é a normalização dos dados. Este tipo de transformação é muito comum em modelos de *machine learning* e tem como objetivo padronizar os dados dentro de um intervalo comum, desempenhando função de relevo na preparação dos dados antes da aplicação efetiva dos modelos, contribuindo para a consistência e aprimoramento das informações. Essa etapa é fundamental na aplicação do algoritmo LASSO, embora não seja determinante no desenho do Random Forest.

3.1 Origem e tratamento dos dados

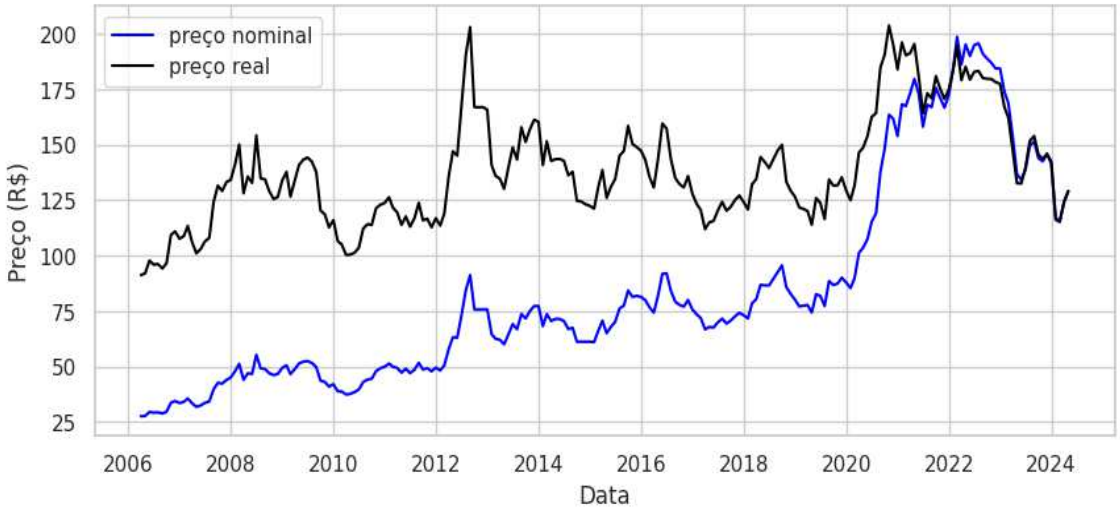
Os indicadores iniciais desta investigação compreendem uma série de preços da soja em um período de 18 anos, de 13/03/2006 a 30/04/2024. São dados proveniente do Centro de Estudos Avançados em Economia Aplicada (CEPEA) da Escola Superior de Agricultura Luiz de Queiroz – ESALQ, em Piracicaba, unidade da Universidade de São Paulo. O Indicador de Preço da Soja ESALQ/BM & FBOVESPA é uma média aritmética dos preços da soja comercializada no porto de Paranaguá, referente ao valor nominal da saca de 60 quilos.

Na primeira visualização dos dados, verificou-se uma tendência de alta nos preços. Considerando, todavia, que os valores da série são nominais, essa tendência reflete também efeitos da inflação ao extenso do período analisado. Uma vez que os dados originais são diários e são os índices de inflação calculados mensalmente, optou-se por efetivar uma agregação mensal, obtendo uma série mensal de preços médios e não ajustar os valores diários pela inflação. Com apoio na série de preços mensais, realizou-se a correção dos efeitos da inflação com base no Índice Geral de Preços (IGP-M), calculado mensalmente pelo Instituto Brasileiro de Economia da Fundação Getúlio Vargas (FGV IBRE)[1].

[1] Os preços foram assim corrigidos: $Preço_{real_t} = Preço_{nominal_t} \cdot \left(\frac{IGPM_{base}}{IGPM_t}\right)$.

O gráfico da Figura 1 contém os valores reais e nominais da série de preços da soja considerada na análise.

Figura 1: Preço nominal x Preço Real da soja (em R$/saca de 60 kg).



Fonte: Dados da pesquisa.

Do gráfico, evidencia-se a dinâmica do preço da soja ao curso de 18 anos, em que o preço nominal não reflete com precisão as oscilações do poder de compra real da soja. Observando o preço real, ajustado pela inflação, identificam-se os períodos de maior volatilidade, conforme delineado ao prosseguimento.

- 2008-2009: a crise financeira global impactou o preço real da soja, expondo a vulnerabilidade do setor agrícola a crises internacionais, dando azo a queda na demanda e retração do crédito.
- 2010-2014: o *boom* das *commodities*, impulsionado pelo crescimento da China, elevou o preço real da soja, com demanda crescente por alimentos e a valorização do real em relação ao dólar.
- 2014-2016: a desaceleração da economia chinesa e a crise política no Brasil contribuíram para a redução da demanda externa e a desvalorização do real, influindo negativamente na rentabilidade da cultura.
- 2020-2021: a pandemia de COVID-19 e a Guerra na Ucrânia causaram intensa volatilidade no preço real da soja, evidenciando a influência de eventos inesperados no mercado, com disfunções nas cadeias de suprimentos.

A análise dessas flutuações e dos eventos subjacentes orienta, assim, o desenvolvimento da modelagem, que intentou capturar as sazonalidades que afetam a formação de preços no mercado brasileiro.

3.2 Seasonal Autoregressive Integrated Moving Average (SARIMA)

O modelo *Seasonal Autoregressive Integrated Moving Average* (SARIMA) é uma extensão do *Autoregressive Integrated Moving Average* (ARIMA), projetado especificamente para lidar com dados de séries temporais denotativas de padrões sazonais. Na previsão de preços da soja, a modelagem de componentes sazonais é particularmente importante, em razão da natureza cíclica das colheitas e de outros fatores sazonais que influenciam os preços *pro rata temporis,* como variações climáticas e flutuações na demanda.

O modelo SARIMA é representado pela notação SARIMA $(p, d, q) \times (P, D, Q)_s$, onde:

- p: ordem da parte autorregressiva (AR);
- d: número de diferenciações necessárias para tornar a série estacionária;
- q: ordem da parte de média móvel (MA);
- P: ordem da parte autorregressiva sazonal;
- D: número de diferenciações sazonais;
- Q: ordem da parte de média móvel sazonal; e
- s: período da sazonalidade (12 para dados mensais).

Matematicamente, o modelo SARIMA é assim descrito:

$$\Phi_P(B^S)\phi_p(B)(1-B)^d(1-B^S)^D y_t = \Theta_Q(B^S)\theta_q(B)\varepsilon_t \quad [1]$$

Nesta expressão, B é o operador de defasagem, tal que $By_t = y_{t-1}$; $\phi_p(B)$ e $\Phi_P(B^S)$ são polinômios de ordem p e P, respectivamente, representando as componentes autorregressivas não sazonais e sazonais:

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \quad [2a]$$

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{s^2} - \cdots - \Phi_p B^{sp} \quad [2b]$$

O termo $(1-B)^d$ representa a diferenciação não sazonal de ordem d, enquanto o termo $(1-B^S)^D$ significa a diferenciação sazonal de ordem D. ε_t, por sua vez, é um termo de erro ruído branco, com média zero e variância constante.

A formulação SARIMA enseja modelar tanto as características não sazonais quanto as sazonais da série temporal. As componentes sazonais são incorporadas diretamente ao modelo, capturando-se padrões que se repetem em intervalos regulares. No caso dos preços da soja, a sazonalidade está ligada ao ciclo de produção, à variabilidade da oferta em função das colheitas

e à demanda que flutua no decorrer do ano, sendo influenciada, por exemplo, pelo consumo e pela exportação.

Ademais, fatores climáticos exercem influência significativa sobre a produção, tornando o comportamento dos preços ainda mais suscetível a variações periódicas. Assim, o método SARIMA é especialmente eficaz para capturar essas variações sazonais e, consequentemente, melhorar a precisão das previsões.

Na prática, ajustar um modelo SARIMA envolve determinar a ordem dos parâmetros sazonais e não sazonais (p, d, q) e $(P, D, Q)_s$. Isso é geralmente feito por meio de uma análise dos gráficos de autocorrelação (ACF) e autocorrelação parcial (PACF), bem como pela aplicação de testes de raiz unitária, como o Teste de Dickey-Fuller aumentado (ADF), para verificar a condição estacionária da série.

Uma vez ajustado, por via do modelo SARIMA, descreve-se a tendência de extenso prazo dos preços da soja, além de identificar e prever padrões cíclicos sazonais, proporcionando previsões precisas e robustas.

Como indicado por James *et al.* (2013), modelos que capturam características sazonais são particularmente úteis em contextos em que a periodicidade dos dados é evidente, conforme sucede com a soja.

3.3 Florestas Aleatórias (Random Forests)

As Florestas Aleatórias (Random Forests) são uma técnica de aprendizado de máquina amplamente utilizada para tarefas de regressão e classificação. Sua origem está no Método de *Bagging* (*bootstrap aggregating*), técnica que visa a melhorar a proximidade da precisão em modelos instáveis, como as árvores de decisão, ao reduzir a variância. No Método de *Bagging*, diversas amostras são geradas com arrimo na amostra de dados original, aplicando a técnica de *bootstrap*, enquanto uma árvore de decisão é treinada em cada uma dessas amostras. Em seguida, a previsão é gerada, no caso de variáveis contínuas, com a média das previsões dessas árvores.

O Random Forest estende a abordagem de bagging, ao introduzir um componente adicional de aleatoriedade pois, além de ensejar diversas amostras do conjunto de dados, ele também seleciona um subconjunto aleatório de variáveis em cada divisão da árvore. Isso reduz a correlação entre as árvores e melhora, ainda mais, a robustez do modelo, especialmente em problemas com alta dimensionalidade e variáveis correlacionadas (James *et al*., 2013). De tal jeito, o Random Forest oferece uma solução poderosa e precisa para lidar com problemas de alta dimensionalidade, variáveis correlacionadas e padrões não lineares.

As Florestas Aleatórias destacam-se pela sua robustez e capacidade de melhorar a proximidade de precisão das previsões, especialmente em contextos complexos como a previsão dos preços da soja, no âmbito da qual diversas variáveis, como fatores climáticos, macroeconômicos e de oferta e demanda, influenciam os preços.

A técnica de Random Forest foi proposta por Breiman (2001) como um conjunto de classificadores (ou regressores, no caso de previsão contínua) composto por múltiplas árvores de decisão, sendo aí cada árvore constituída com esteio numa amostra *bootstrap* do conjunto de dados; ou seja, cada árvore é ajustada com uma amostra aleatória, com reposição, do conjunto de dados original, dando oportunidade a que algumas observações sejam repetidas e outras excluídas em cada subconjunto. Em senso assim, logra o modelo capturar variados aspectos da variabilidade dos dados, contribuindo para maior diversidade das árvores.

Cada árvore de decisão em um modelo de Random Forest é uma estrutura hierárquica estabelecida com arrimo no conjunto de dados de treinamento, utilizando uma abordagem recursiva para dividir o espaço de preditores em regiões. O objetivo de cada divisão é encontrar subconjuntos de dados que sejam mais homogêneos em termos da variável de resposta. No caso específico da previsão de preços da soja, isso significa encontrar divisões dos preditores que reduzam ao máximo a variabilidade do preço dentro de cada nó (região).

A construção da árvore começa com o nó raiz, que contém todos os dados disponíveis. Em cada etapa, o algoritmo seleciona um conjunto aleatório de preditores (de tamanho m) e, entre esses, escolhe aquele que gera a melhor divisão do nó atual. A métrica usada para avaliar a qualidade da divisão é o erro quadrático médio (MSE), que aquilata a variabilidade do valor da variável de resposta (preço da soja) em cada subgrupo resultante. O objetivo é selecionar a variável e o ponto de corte que minimizem a soma dos erros quadráticos médios de cada um dos nós filhos gerados pela divisão:

$$MSE = \sum_{j=1}^{K} \sum_{i \in R_j}^{k} \left(y_i - \hat{y}_{R_j} \right)^2 \qquad [3]$$

Nesta expressão, R_j são as regiões resultantes da divisão, y_i são os valores observados de cada ponto de dados na região, e $\hat{y}_{R_j}$ é a média dos valores na região R_j.

A elaboração da árvore continua até que algum critério de parada seja atingido. Esse critério é suscetível de ser um tamanho mínimo do nó, uma profundidade máxima da árvore ou um valor-limite para o ganho na redução do erro.

No Random Forest, as árvores, geralmente, são deixadas crescer até a profundidade máxima possível (sem poda). O risco de *overfitting* que uma árvore tão profunda teria é

mitigado pelo fato de que várias árvores são combinadas, e cada uma delas é ajustada em uma amostra diferente e com distintos subconjuntos de preditores.

Ao final, cada árvore é formada por uma série de nós terminais (ou folhas), que correspondem às regiões terminais do espaço dos preditores. Para realizar a previsão de uma nova observação, o algoritmo atravessa a árvore, do nó raiz até um nó terminal, dependendo dos valores dos preditores dessa observação. A previsão do valor da variável resposta para um ponto que termina em um determinado nó terminal é simplesmente a média dos valores de resposta dos pontos de dados que caíram naquele nó durante o treinamento.

Matematicamente, se descreve o Random Forest de acordo com o que está expresso à continuação.

1. **Construção de Árvores Individuais**: cada árvore é gerada de uma amostra *bootstrap* do conjunto de dados de treinamento. Desde essa amostra, uma árvore de decisão completa é construída sem poda. Em cada nó da árvore, a divisão é feita selecionando a melhor variável dentre um subconjunto aleatório de m preditores.

2. **Predição Final**: para um problema de regressão, a predição do Random Forest para uma nova observação x_i é dada pela média das predições de todas as árvores:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x_i) \qquad [4]$$

Nesta expressão, B é o número total de árvores no modelo, e $T_b(x_i)$ é a predição da b-ésima árvore para a observação x_i.

A redução da correlação entre as árvores ocorre porque, em cada divisão, apenas um subconjunto de variáveis é considerado. Árvores de decisão constituem modelos de alta variância, ou seja, pequenas mudanças nos dados estão prontas a resultar em grandes modificações na estrutura da árvore. Ao reduzir a correlação entre as árvores por via da seleção aleatória de variáveis, a agregação das árvores tende a reduzir a variância do modelo sem aumentar significativamente o viés, o que resulta em um modelo mais robusto e confiável.

A introdução da retirada aleatória dos preditores em cada divisão de uma árvore – *in hoc sensu* - é uma inovação fundamental em relação ao *bagging*. Ao invés de utilizar todos os preditores para decidir a melhor divisão em cada nó da árvore, um número limitado de preditores m é pinçado aleatoriamente entre os p disponíveis. Esse número m, geralmente, é escolhido como $\sqrt{p}$ para problemas de classificação e aproximadamente $p/3$ para regressão.

Essa técnica reduz a correlação entre as árvores do modelo, o que é crucial para melhorar o desempenho do Random Forest em relação ao simples *bagging*.

Diferentemente de um modelo de árvore única, que tende a se ajustar demais aos dados de treinamento, o Random Forest tem uma alta resistência ao *overfitting*, especialmente quando o número de árvores (B) é grande. O erro de generalização do Random Forest se estabiliza à medida que o quantitativo de árvores aumenta, e mesmo adicionar muitas árvores não transporta ao *overfitting*. Esse comportamento é uma grande vantagem quando se lida com os conjuntos de dados complexos e ruidosos, como aqueles que envolvem preços de *commodities* agrícolas.

No contexto da previsão dos preços da soja, o Random Forest é particularmente poderoso em decorrência da complexidade e da alta dimensionalidade dos indicadores. As variáveis climáticas, econômicas, de demanda e oferta são passíveis de ser altamente correlacionadas e exibem relações não lineares que métodos lineares como a regressão linear comum não capturam bem. O Random Forest, por meio de sua estrutura de árvores e a capacidade de explorar interações complexas e não lineares, é hábil para capturar essas relações de maneira mais chegada à precisão, resultando em melhores previsões.

Em transposição, o uso de amostragem *bootstrap* e a seleção de preditores para cada divisão introduzem variação suficiente para reduzir o risco de ajustar ruídos específicos dos dados, tornando o modelo menos sensível a *outliers* ou a observações particularmente influentes.

3.3.1 Ajuste de hiperparâmetros

O ajuste de *hiperparâmetros* é uma etapa terminante no desenvolvimento de modelos de *machine learning*, pois influencia diretamente a performance do modelo. No caso dos de Random Forest, essa fase é particularmente importante, pois o desempenho do modelo depende de escolhas, como o número de árvores, a profundidade máxima das árvores e a porção de variáveis consideradas em cada divisão. Esses parâmetros controlam o equilíbrio entre variância e viés no modelo: enquanto um número insuficiente de árvores ou uma profundidade muito rasa é conducente a um modelo com alto viés, um ajuste inadequado de variáveis é propício a resultar em sobreajuste, comprometendo a capacidade de generalização do modelo para novos dados. Assim, o ato de otimizar esses *hiperparâmetros* garante que o Random Forest chegue a capturar de maneira eficiente os padrões do conjunto de dados, sem superestimar ruídos.

Na sequência, pontoam-se os principais hiperparâmetros a serem ajustados na estimação do modelo de Floresta Aleatória.

- Número de árvores da Floresta Aleatória (*B*, *n_estimators*): o *quantum* de árvores que compõem a floresta é um parâmetro relevante, pois, quanto maior o total de árvores, mais estável e preciso o modelo tende a ser. A principal restrição relativamente ao uso de um número muito grande de árvores decorre da capacidade de processamento. Perfilhou-se, neste passo, um valor típico de 100 árvores.
- Profundidade máxima de cada árvore (*max_depth*): esse parâmetro limita a profundidade de cada árvore na floresta. Árvores mais profundas capturam mais nuanças dos dados, mas conduzem ao ajuste excessivo (*overfitting*), caso em que o modelo se ajusta demais aos dados de treino e perde a capacidade de generalização.
- Quantidade mínima de amostras para dividir um nó interno (*min_samples_split*): define a quantidade mínima de amostras necessárias para que um nó interno seja dividido. Valores menores fazem, as mais das vezes, com que o modelo seja mais sensível aos dados, enquanto valores maiores resultam em árvores mais conservadoras e menos complexas.
- Quantitativo mínimo de amostras que uma folha (*min_samples_leaf*): indica o número mínimo de amostras que uma folha (nó terminal) deve ter. Ajustar este hiperparâmetro impede a criação de folhas com poucas amostras, promovendo generalização. Valores maiores tornam a árvore menos complexa.
- Número máximo de variáveis consideradas para dividir um nó (*max_features*): ajuda a diversificar as árvores, já que distintas combinações de variáveis são usadas em cada divisão, aumentando a robustez do modelo. Um valor mais baixo é propício a reduzir o risco de *overfitting*, mas valores muito altos prejudicam a precisão.

Amostragem com substituição (*bootstrap*): este parâmetro indica se o modelo deve usar amostragem com substituição (*bootstrapping*) ao estabelecer as árvores. Quando ativado (*bootstrap=True*), cada árvore é treinada em uma amostra diferente dos dados originais, promovendo variabilidade e robustez. Se desativado (*bootstrap=False*), todas as árvores são treinadas com o mesmo conjunto de dados, o que é passível de reduzir a robusteza do modelo.

A escolha adequada desses parâmetros é geralmente feita utilizando métodos de validação cruzada, que verificam qual configuração produz o menor erro preditivo no conjunto de validação. A validação cruzada é uma técnica usada para avaliar o desempenho de um modelo de maneira mais confiável, dividindo os dados em múltiplos conjuntos de treino e teste. É uma modalidade de garantir que o modelo não esteja apenas memorizando os dados de treino, o que resulta em *overfitting*. Em vez de treinar o modelo uma vez e avaliar em um conjunto de

teste, na validação cruzada, o modelo é testado em distintas partes dos dados, garantindo que o desempenho avaliado seja mais representativo e generalizável para dados fora da amostra.

Uma vez que se está lidando com uma série temporal, o método divide sequencialmente os dados, respeitando a ordem temporal, em que a dependência de tempo não há de ser ignorada. A validação cruzada temporal conduz o modelo a ser treinado em uma parte dos dados e validado em outra, sem violar a sequência temporal.

Em conjunto com a validação cruzada, se faz necessária a aplicação de um método para avaliar o melhor conjunto de *hiperparâmetros*. Os dois métodos mais usuais são a demanda em grade (*Grid Search*) e a procura aleatória (*Randomized Search*).

O procedimento de *Grid Search* define uma grade, combinando os valores possíveis para cada *hiperparâmetro* do modelo. Este é treinado em cada combinação possível e avaliado com base em uma métrica de desempenho (como erro quadrático médio, por exemplo). A combinação de hiperparâmetros que produz o melhor desempenho médio na validação cruzada é selecionada. O *Grid Search* procura a melhor combinação de hiperparâmetros para um modelo de modo exaustivo, testando todas as possíveis combinações dentro de um espaço predefinido de valores para cada hiperparâmetro. Essa tarefa é computacionalmente custosa se o espaço de procura for muito grande.

O *Randomized Search* é um método tenciona a melhor combinação de hiperparâmetros, selecionando, executando e avaliando, aleatoriamente, variadas combinações em uma grade de *hiperparâmetros* definida. Este método oferece um formato mais eficiente de explorar o espaço de *hiperparâmetros*, testando um subconjunto aleatório das combinações. Isso faz com que se encontrem boas soluções com menor custo computacional, embora não garanta que a melhor combinação possível seja testada.

3.3.2 Importância das variáveis

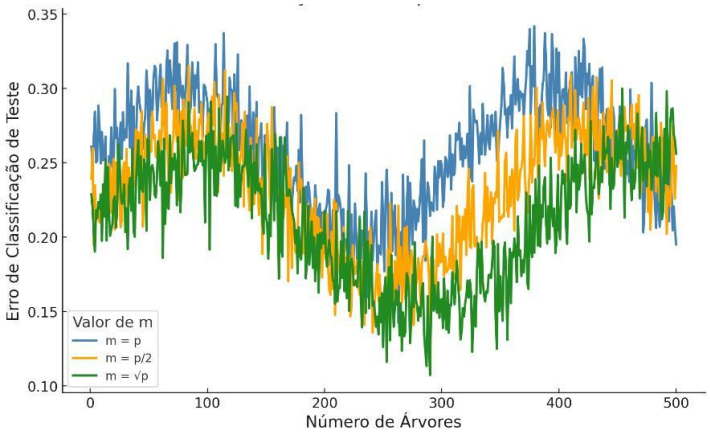
Outro aspecto de relevo do Random Forest é a possibilidade de calcular a importância das variáveis, que indica quais preditores têm maior impacto na predição do modelo. A importância de uma variável é geralmente medida pelo aumento do erro preditivo, quando essa variável é permutada (ou seja, suas observações são embaralhadas). Maior perda de acurácia implica que a variável era importante para as decisões das árvores, fornecendo *insights* sobre quais fatores são mais relevantes na previsão dos preços da soja.

Essa aleatoriedade adicional é indescartável para reduzir a correlação entre as árvores de decisão individuais, resultando em uma média das previsões das árvores, menos variável e mais confiável. Ademais, a resistência ao *overfitting* é outra característica importante das

Random Forests, especialmente quando muitas árvores B é são utilizadas. Com isso, o erro do modelo se estabiliza, à medida que B aumenta, proporcionando previsões mais robustas (James *et al*., 2013).

Nesse sentido, a previsão de preços de soja envolve a análise de uma série de variáveis econômicas, climáticas e de mercado, frequentemente correlacionadas e exibindo comportamento não linear. A capacidade de as Florestas lidarem com essas complexidades as torna particularmente adequadas para essa tarefa. Ao introduzir a aleatoriedade na seleção dos preditores em cada divisão, o modelo captura relações complexas entre as variáveis, proporcionando previsões precisas em comparação com métodos tradicionais.

Figura 2: Erro de classificação de teste, por números de árvores.



Fonte: Adaptado de James *et al*., 2013.

Ao aplicá-la a um conjunto de dados de previsão de preços de soja, é notória a eficácia da técnica na redução do erro preditivo. A Figura 2 (adaptada de James *et al*., 2013) ilustra como a escolha do parâmetro *m* afeta a taxa de erro do modelo. Observa-se que, ao utilizar $m = \sqrt{p}$, a Floresta Aleatória tende a produzir uma leve melhoria no erro de teste em comparação ao *bagging*, onde $m = p$. Este resultado indica que a redução da correlação entre as árvores realmente contribui para maior acurácia preditiva.

3.4 *Least Absolute Shrinkage and Selection Operator* (Lasso)

O método *Least Absolute Shrinkage and Selection Operator* (Lasso) é uma técnica de regressão que melhora a previsibilidade e a possibilidade de interpretar os modelos lineares, aplicando uma penalização à soma dos valores absolutos dos coeficientes das variáveis. Introduzido por Tibshirani em 1996, o Lasso tem como principal objetivo a seleção de variáveis e a regularização para prevenir o *overfitting*, especialmente em contextos em que o número de preditores é maior ou próximo da quantidade de observações.

De efeito, o Lasso resolve o problema de minimização:

$$\left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \quad [5]$$

A primeira parte da equação é o erro de ajuste (soma dos quadrados dos resíduos). A segunda é a penalização que o Lasso aplica, sendo controlada pelo parâmetro de regularização, $\lambda \geq 0$. Essa penalização é a soma dos valores absolutos dos coeficientes das variáveis explicativas (β_j). Quando $\lambda = 0$, o Lasso se reduz à regressão linear comum; à medida que λ aumenta, mais coeficientes β_j são reduzidos a zero, resultando em um modelo mais simples.

Assim, o valor de λ deve ser otimizado, pois valores muito grandes removem variáveis importantes, enquanto valores muito pequenos incluem muitas variáveis irrelevantes. Assim como os *hiperparâmetros* comentados no modelo Random Forest, o parâmetro de penalização λ deve ser ajustado para encontrar o valor ideal. Esse ajuste é crucial, porque λ controla a força da penalização imposta aos coeficientes, influenciando diretamente o desempenho do modelo e a seleção de variáveis.

Com efeito, uma das características mais notáveis do Lasso é a capacidade de selecionar variáveis. Diferentemente da regressão Ridge, que também aplica uma penalização, mas nunca zera os coeficientes, o Lasso, realmente, elimina variáveis, ao definir seus coeficientes exatamente a zero, identificando um subconjunto de preditores mais informativos. Desse modo, ele é particularmente útil em circunstâncias onde existe multicolinearidade entre as variáveis preditoras, pois este tende a selecionar uma variável de cada grupo de variáveis correlacionadas, em vez de todas. Uma limitação do método, entretanto, está no fato de que, em situações nas quais o número de variáveis preditoras é maior do que o quantitativo de observações, ele somente seleciona até n variáveis antes de zerar os coeficientes adicionais (James *et al*., 2013).

No contexto de previsão de preços de soja, o Lasso é utilizável para identificar os principais fatores que influenciam o preço, ao mesmo tempo em que descarta variáveis irrelevantes ou redundantes, resultando em um modelo mais interpretável e, possivelmente, com melhor desempenho preditivo.

Como o Lasso aplica uma penalização aos coeficientes, um aspecto importante é que ele se torna muito sensível às unidades das variáveis. Quando as variáveis têm escalas muito diferentes, os coeficientes associados a elas são influenciados, não apenas, pela importância real de cada variável, mas, também, por suas magnitudes relativas. Isso cria um viés de seleção

no Lasso, uma vez que o modelo é propício a preferir excluir (reduzir a zero) variáveis de alta magnitude, não porque elas sejam menos relevantes, mas simplesmente porque seus coeficientes são maiores em termos absolutos. Em assim ocorrendo, a padronização se faz necessária para evitar colocar todas as variáveis na mesma escala, dando ensejo a comparações justas.

3.5 Engenharia de atributos (preditores)

No contexto de um modelo de séries temporais univariado, a engenharia de atributos é uma etapa essencial para enriquecer o conjunto de dados com informações derivadas, de modo a melhorar a capacidade do modelo de capturar padrões, tendências e sazonalidades da série original. Para a previsão de preços, foi gerado um conjunto de *dummies* e instituídos diversos atributos derivados do preço real da soja, visando a capturar variadas dinâmicas e características da série temporal.

A literatura empírica demonstra a efetividade das variáveis *dummy* na modelagem de preços de *commodities* agrícolas, capturando choques de oferta e demanda, eventos climáticos extremos e mudanças nas políticas agrícolas (Adjemian *et al*., 2018; Isengildina-Massa *et al*., 2008). Assim, ao introduzir essas variáveis, analisa-se o comportamento dos preços ao extenso de cada mês. Isso é particularmente relevante para os preços agrícolas, uma vez que eles reagem de maneira muito sensível aos períodos de safra e entressafra.

Por sua vez, a criação de vários atributos com suporte na própria dinâmica da variável analisada, leva a se capturar distintas características da série temporal, de modo a facilitar o aprendizado do modelo. A inclusão das médias móveis e da diferença sazonal tem como objetivo fornecer ao modelo informações sobre a tendência geral dos preços e padrões recorrentes à medida que passam os anos.

Por sua vez, as variações percentuais ajudam a descrever a dinâmica da mudança nos preços, enquanto as medidas de volatilidade, associadas aos preços em diferentes janelas de tempo, são decisivas para entender períodos de maior risco, o que se faz útil para a tomada de decisão no contexto de mercado.

O indicador de *momentum*, em conjunto com as médias móveis, captura a dinâmica recente dos preços, indicando se há persistência em uma tendência de alta ou de baixa. Isso é útil para prever o comportamento futuro dos preços, assumindo que as tendências de curto prazo continuem por um período determinado.

A seguir se encontra uma lista de atributos produzidos desde a série de preço real da soja.

- Médias móveis de 3, 12 e 24 Meses (***mm_3***, ***mm_12***, **mm_24**): as médias móveis ajudam a suavizar ruídos e, por seu intermédio, se identificam tendências de curto, médio e comprido prazos dos preços.
- Variação percentual mensal e anual (***var_pct_mensal***, ***var_pct_anual***): a variação percentual mensal (***var_pct_mensal***) mede a mudança no preço do mês atual ao mês anterior. Esse atributo é útil para entender a dinâmica de curto prazo e identificar períodos de alta volatilidade ou mudanças bruscas no comportamento dos preços. Por sua vez, a variação percentual anual (***var_pct_anual***) fornece uma medida de como o preço se compara ao mesmo mês do ano anterior, sendo particularmente relevante para capturar padrões sazonais e avaliar como os preços estão se comportando em relação a possíveis ciclos anuais.
- Volatilidade de preços em janelas de 3 e 12 meses (***volatilidade_3m***, ***volatilidade_12m***): a volatilidade foi calculada como o desvio-padrão dos preços ao largo de janelas de 3 e 12 meses. A volatilidade de 3 meses (volatilidade_3m) ajuda a identificar períodos de instabilidade de curto prazo nos preços, enquanto a volatilidade de 12 meses (volatilidade_12m) mede a variação ao extenso de um ciclo anual, sendo útil para avaliar como a incerteza nos preços varia no percurso das estações.
- *Momentum* de 3 meses (***momentum_3m***): o indicador de *momentum* de 3 meses foi definido como a diferença entre o preço atual e o de meses atrás. Esse atributo é útil para identificar a direção e a magnitude das tendências recentes. Valores positivos indicam um pendor de alta, enquanto valores negativos apontam para uma propensão de queda. O *momentum* auxilia a prever se uma disposição está se acelerando ou perdendo força.
- Diferença Sazonal de Preço (***sazonal_diff***): foi calculada como a diferença entre os preços atual e do mesmo mês do ano anterior. Esse atributo é particularmente útil para capturar variações sazonais e diferenças de preço entre os anos, transportando o modelo a compreender padrões que se repetem ao tempo dos mesmos meses a cada ano.

3.6 Critérios de avaliação do modelo

No trabalho de previsão de preços de soja, a avaliação dos modelos preditivos é fundamental para garantir a precisão e a confiabilidade dos resultados. Eis dois critérios amplamente utilizados para essa avaliação: o Erro Absoluto Médio (MAE) e o Erro Quadrático Médio (MSE).

O MAE avalia a média das diferenças absolutas entre os valores previstos e os observados, oferecendo uma visão direta sobre a magnitude dos erros sem dar peso a erros maiores ou menores (Hyndman; Koehler, 2006). Matematicamente, o MAE é definido como:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|,$$

em que n é o número de previsões realizadas.

Relacionado ao MAE, o Erro Absoluto Percentual Médio (Mean Absolute Percentage Erro – MAPE) mensura o erro percentual absoluto médio entre os valores reais e os previstos. O cálculo do MAPE é assim realizado:

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

O MAPE é uma métrica intuitiva que fornece o erro em termos percentuais, facilitando a interpretação.

Já o MSE calcula a média dos quadrados das diferenças entre os valores previstos e os observados, penaliza mais severamente os grandes erros, tornando-os sensíveis a *outliers* e mais apropriados para modelos onde grandes desvios são críticos (Chai; Draxler, 2014). Matematicamente, o MSE é definido como:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Uma vez que o MSE está na escala dos erros quadrados, ele não tem uma interpretação direta. Com vistas a contornar isso, é bastante comum reportar a raiz quadrada do MSE, denominada Root Mean Squared Error (RMSE).

A escolha entre MAE e MSE deve ser orientada pela natureza dos dados e pela finalidade específica da previsão. Enquanto o MAE fornece uma interpretação mais intuitiva e é menos sensível a *outliers*, o MSE é preferido em contextos nos quais grandes erros são menos toleráveis, uma vez que sua estrutura quadrática aumenta a penalização de desvios maiores (Willmott; Matsuura, 2005).

Em modelos de previsão de preços de *commodities* como a soja, em que a volatilidade e a atuação de *outliers* são características frequentes, o uso combinado dessas métricas é

propício a empreender uma análise mais robusta da performance do modelo, dando oportunidade a uma avaliação mais equilibrada entre precisão e estabilidade dos resultados.

4 RESULTADOS E DISCUSSÕES

Nesta seção, relatam-se os principais resultados obtidos no decurso de desenvolvimento empírico da Dissertação, com foco na modelagem do comportamento do preço real da soja e na avaliação dos modelos de previsão, recorrendo a variadas técnicas.

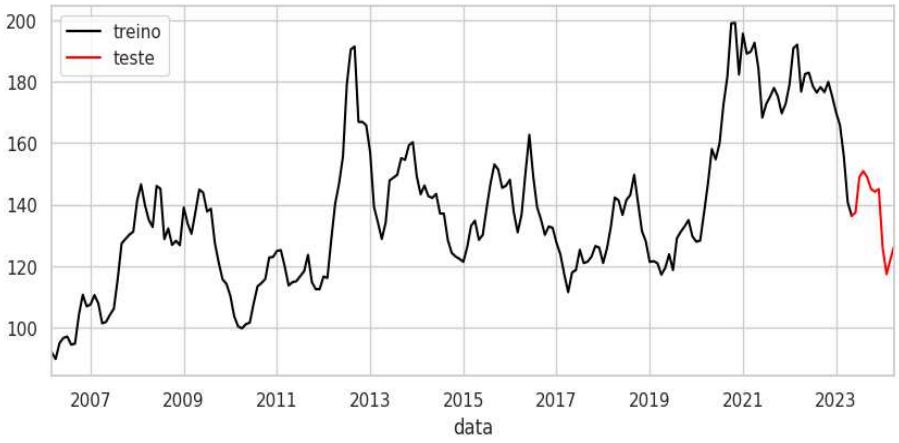
4.1 Divisão em dados de treino e teste

Uma das primeiras etapas do processo de aplicação de um modelo preditivo é a divisão dos dados em dois conjuntos: o de treino e o de teste.

O conjunto de treino é usado para ajustar o modelo. Nele, o modelo aprende os padrões, as relações entre as variáveis e ajusta seus parâmetros de acordo com os dados fornecidos. Basicamente, o modelo é "treinado" para encontrar uma função que descreva bem o comportamento dos indicadores. Após o treinamento, o modelo precisa ser avaliado em um conjunto de dados que ele não tenha visto durante o treino. Esse é o propósito do conjunto de teste. Ele contém novas observações conduzem a se verificar se o modelo logra generalizar os padrões aprendidos para novos indicativos. A avaliação do modelo no conjunto de teste leva ao cálculo mais realista de sua performance preditiva, pois testa sua habilidade de prever valores que ele não conhece.

No estudo sob glosa, foi definido um conjunto de testes com dados, no período de março de 2006 a abril de 2023, correspondendo a 207 observações. Por sua vez, o conjunto de teste foi definido com os últimos 12 meses da amostra, de maio de 2023 a abril de 2024.

Figura 3: Preço Real da soja (em R$/saca de 60 kg): dados de treino e teste.



Fonte: Dados da pesquisa.

4.2 Resultados - Modelo SARIMA

Na aplicação do modelo SARIMA, assim como no emprego de qualquer modelo de série temporal, a análise foi iniciada com testes de estacionariedade. Na análise do preço real da soja, foram aplicados dois testes sobre os dados de treino: o Teste de Dickey-Fuller Aumentado (ADF) e o Teste KPSS.

O Teste ADF toma como hipótese nula a ideação de que a série temporal possui raiz unitária - uma característica das séries temporais não estacionárias. De outra vertente, a hipótese alternativa (sob a qual a hipótese nula é rejeitada) é que a série é estacionária. Como a hipótese nula pressupõe uma raiz unitária, o *valor p* obtido deve ser inferior a um nível de significância especificado, muitas vezes fixado em 0,05, para rejeitar esta hipótese.

O Teste KPSS verifica se uma série temporal é estacionária em torno de uma tendência média ou linear. Neste teste, a hipótese nula é de que os dados são estacionários e procuram-se evidências de que a hipótese nula seja falsa. Consequentemente, com um *valor p* pequeno (menos de 0,05) rejeita a hipótese nula, sugerindo que a diferenciação é necessária.

Se a série não for estacionária, impõe-se diferenciá-la, removendo tendências ou sazonalidades, o que conduz à segunda etapa. Se a série não for estacionária, a técnica de diferenciação é aplicada. Para séries com padrões sazonais, é, decerto, necessário aplicar tanto uma diferenciação regular quanto uma sazonal para remover a tendência e a sazonalidade. No SARIMA, isso é expresso pelos parâmetros d e D, indicativos do número das diferenciações necessárias para alcançar a circunstância estacionária.

Para as séries em nível (escala original) e em primeira diferença, os resultados dos testes estão inscritos na Tabela 1.

Tabela 1 - Estatísticas dos Testes ADF e KPSS para os preços da soja

	Série em nível	**Série em 1ª diferença**
ADF Statistic	-3.1819 (p-valor: 0.0211)	-5.9456 (p-valor: 2.208e-07)
KPSS Statistic	0.9244 (p-valor: 0.01)	0.0948 (p-valor: >0.1)

Fonte: Dados da pesquisa.

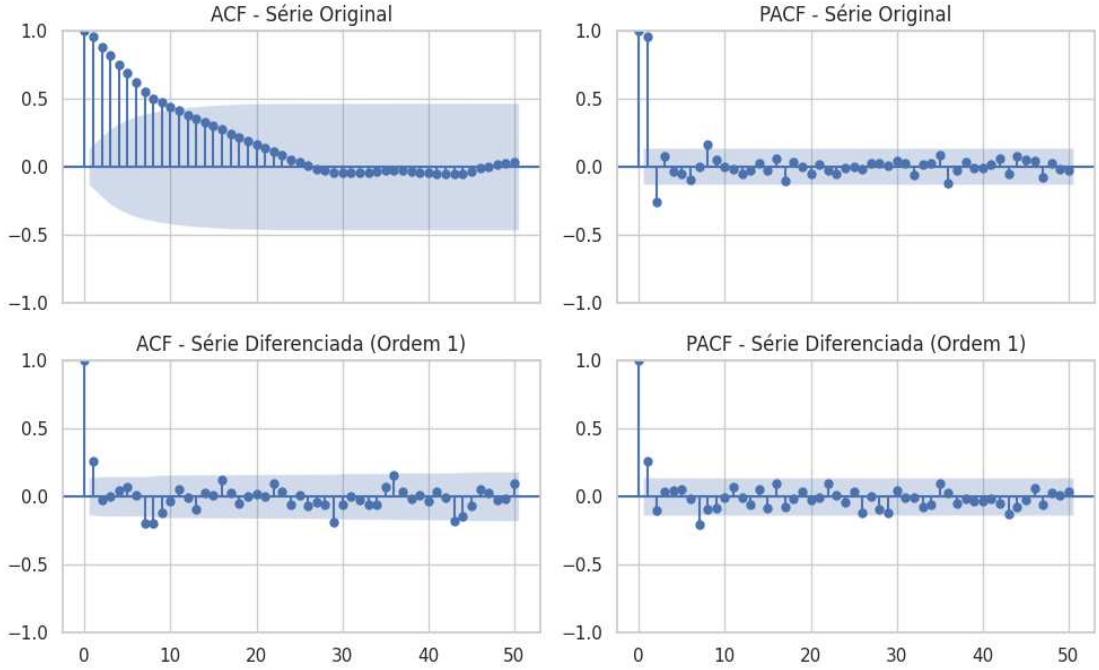
Para a série original, tem-se que o teste ADF rejeita a hipótese nula, realidade a sugerir que a série é estacionária, uma vez que o valor-p (0.0211) é menor do que o nível de

significância adotado (0.05). Por sua vez, o Teste KPSS indica a rejeição da hipótese nula de que a série seja estacionária, pois o valor-p (0.01) é menor do que o nível de significância.

Para a série em primeira diferença, o Teste ADF mostra um valor-p excessivamente baixo (muito menor do que 0.05), transportando à rejeição da hipótese nula e insinuando que a série diferenciada de primeira ordem é estacionária. E o Teste KPSS também aponta que a série é estacionária, uma vez que o valor-p é maior do que 0.1. Como ambos os testes concordam com a ideia de que a série é estacionária após a primeira diferenciação, é dado definir-se que a série é integrada de ordem 1.

De maneira complementar, foram plotados os gráficos das funções de autocorrelação (ACF) e autocorrelação parcial (PACF). Os gráficos mostram evidências de que a série original não é estacionária, denotando vigorosa autocorrelação com defasagens mais distantes. Isso é indicado com um decaimento gradual da ACF e pela primeira defasagem significativa na PACF. Após a diferenciação de primeira ordem, a série se torna estacionária. Isso é confirmado pelo padrão de ruído branco na ACF e pela ausência de defasagens significativas após a primeira na PACF.

Figura 4: Funções de autocorrelação e autocorrelação parcial da série de preço real da soja (em R$/saca de 60 kg).



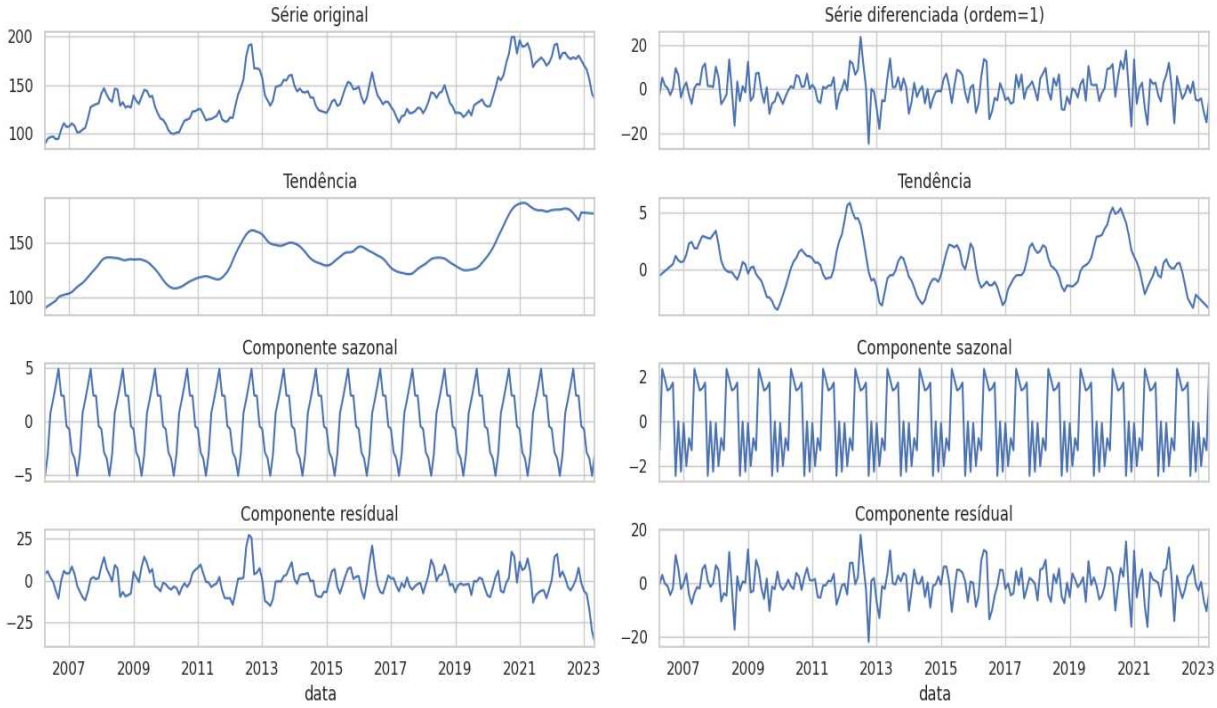
Fonte: Dados da pesquisa.

Foi realizada a decomposição da série em componentes de tendência, sazonalidade e termo residual. Essa decomposição foi aplicada com o uso da função *seasonal_decompose,* da Biblioteca *statsmodels*, em linguagem *Python*. Os resultados dessa decomposição encontram-se dispostos nos gráficos da Figura 5.

Uma nova etapa de testes foi realizada, considerando tanto tendências de longo prazo quanto variações sazonais que poderiam causar não-circunstância estacionária na série original. Para isso, realizou-se a diferenciação regular (ordem 1) e efetivou-se uma diferenciação sazonal de 12 períodos (diferenciação sazonal típica quando há sazonalidade anual em dados mensais). A estatística de Teste do ADF foi de -3.5579 (com valor-p de 0.0066). Já a estatística de Teste do KPSS foi estimada em 0.0464 (com valor-p de 0.1). Estes resultados indicam que, após aplicar a diferenciação de primeira ordem e a diferenciação sazonal, a série tornou-se estacionária.

Esses resultados apontam que é razoável realizar a diferenciação da série em termos regulares e de componentes sazonais para ser estacionária. Tal sugere a adequação de um modelo SARIMA com $d = 1$ e $D = 1$ (diferenciação sazonal de ordem 1).

Figura 5: Decomposição do preço real da soja (em R$/saca de 60 kg).



Fonte: Dados da pesquisa.

Considerando as análises da série do preço real da soja, foi realizada a estimação de um modelo SARIMA $(0,1,1) \times (1,1,0)_{12}$, ou seja, com uma diferenciação regular ($d = 1$) e um componente de média móvel (MA) de ordem 1 ($q = 1$), além de componente sazonal com diferenciação sazonal ($D = 1$), um componente autorregressivo sazonal de ordem 1 ($P = 1$) com sazonalidade de 12 meses.

O coeficiente MA (ma.L1) foi estimado em 0.2899 com um p-valor muito pequeno, indicando que é estatisticamente significativo. O coeficiente do componente AR sazonal (ar.S.L12) foi estimado em -0.4538, e se mostrou significativo. A variância dos resíduos (sigma2) foi estimada em 69.2754.

O Teste de Ljung-Box (L1), de Q=0.04 (com p-valor de 0.85), aponta que os resíduos não exprimem autocorrelação significativa no lag 1. O Teste de Normalidade dos Resíduos, de Jarque-Bera (JB), expressou valor de 1.11 (p-valor de 0.58), sugerindo que os resíduos são normalmente distribuídos. Entrementes, a medida de heterocedasticidade (H), H=1.43 (com p-valor de 0.15), sinaliza que não há evidências de heterocedasticidade nos resíduos.

Tabela 2. Resultados da estimação do modelo SARIMA $(0, 1, 1) \times (1, 1, 0)_{12}$

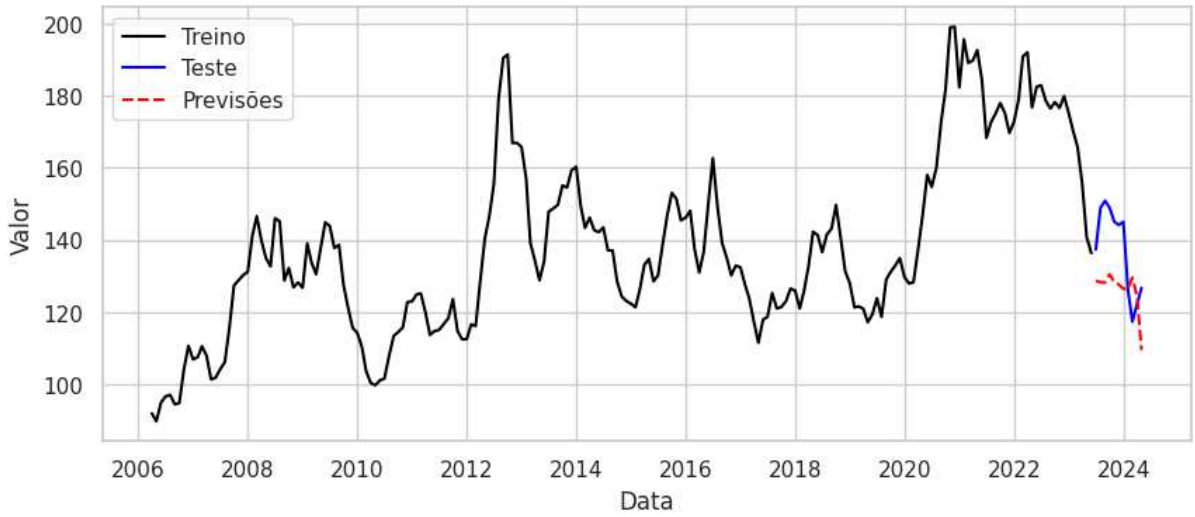
Coeficientes	
MA(1)	0.2899
	(erro padrão: 0.073)
	(estatística z: 3988)
AR(12)	-0.4538
	(erro padrão: 0.061)
	(estatística z: -7490)
σ^2	692.754
	(erro padrão: 7781)
	(estatística z: 8904)
Análise Residual	
Ljung-Box (Q)	0.04
	(p-valor: 0.85)
Jarque-Bera (JB)	1.11
	(p-valor: 0.58)
Heterocedasticidade (H)	1.43
	(p-valor: 0.15)
Critérios de Seleção	
AIC	1.381.592
BIC	1.391.396
HQIC	1.385.562

Fonte: Dados da pesquisa.

Os critérios de informação expressaram os seguintes valores: AIC = 1381.592, BIC= 1391.396 e HQIC= 1385.562. Dentre variadas especificações de modelos testadas, estes foram os valores mais baixos, o que contribuiu para definir a especificação do SARIMA estimado.

Com o modelo ajustado/treinado, este foi aplicado ao conjunto de dados de teste para realizar previsões do preço da soja. Considerando o período do conjunto de testes, a ideia é simular a aplicação do modelo para a previsão de 12 meses. O gráfico da Figura 6 aponta a série original com a separação em dados de treino e teste e a previsão gerada.

Figura 6: Série observada e previsão do preço real da soja (em R$/saca de 60 kg) com base no SARIMA $(0, 1, 1) \times (1, 1, 0)_{12}$.



Fonte: Dados da pesquisa.

Para avaliar o desempenho deste modelo no conjunto de dados de testes, foram calculadas as medidas do Erro Absoluto Médio (MAE) e o Erro Quadrático Médio (MSE), bem como a raiz quadrada deste último. O MAE exibiu valor de 14.03, enquanto o MSE foi calculado em 245.27, com raiz quadrada (RMSE) de 15.66.

Também foi calculado o erro percentual médio (*Mean Absolute Percentage Error* - MAPE) entre as previsões e os valores reais, ou seja, a média dos erros absolutos em termos percentuais. O MAPE foi calculado em 9.96%, indicando que, em média, o modelo comete um erro de cerca de 9.96% nas suas previsões.

4.3 Resultados – Random Forest

O modelo de Random Forest contou com a mesma divisão dos dados em conjuntos de treino e teste. Após a divisão, foi realizada a geração de atributos, como relatado na seção 3.3.

Durante o treinamento do modelo, foi implementado o procedimento de validação cruzada temporal, com o objetivo de avaliar o desempenho do modelo à extensão de variadas divisões (*splits*) temporais, calculando a medida de RMSE para os conjuntos de treino e teste em cada iteração. Em cada iteração foram realizadas cinco divisões, onde cada conjunto de teste é composto por uma parte sucessiva dos dados temporais, preservando a ordem cronológica.

Em conjunto com o processo de validação cruzada, foram testados variados conjuntos de hiperparâmetros por meio do procedimento de demanda aleatória na grade de possíveis valores (*Randomized Search*)[2]. Os resultados da validação cruzada são os postados sequencialmente.

- *Split* 1: *RMSE* Treino = 2.07, *RMSE* Teste = 17.65
- *Split* 2: *RMSE* Treino = 2.33, *RMSE* Teste = 9.05
- *Split* 3: *RMSE* Treino = 2.05, *RMSE* Teste = 4.37
- *Split* 4: *RMSE* Treino = 2.08, *RMSE* Teste = 9.69
- *Split* 5: *RMSE* Treino = 1.76, *RMSE* Teste = 19.94

Notória é uma variabilidade elevada nos valores, fato indicativo de que o modelo expressou dificuldades em prever corretamente dentro de alguns períodos específicos. A principal hipótese para este comportamento é que mudanças nos padrões dos dados *pro rata temporis* é propício a ensejar diferenças tão grandes. Os melhores hiperparâmetros foram definidos com base no desempenho médio da combinação de hiperparâmetros nos cinco *splits*.

O modelo foi ajustado com 100 árvores, algo significativo de que o algoritmo está fazendo previsões com base na média dos resultados de 100 árvores individuais. O parâmetro que ajusta o número de amostras em cada nó foi melhor ajustado de modo que cada nó tenha pelo menos cinco amostras antes de ser dividido em nós derivados (nós filhos). Conforme, também, foi ajustado, pelo menos duas amostras estão em cada folha, o que significa que nenhuma folha pode ser criada com apenas uma amostra. Outro parâmetro, que controla quantas variáveis explicativas são consideradas em cada divisão de um nó, foi definido como *'sqrt'*, de sorte que o modelo escolhe a raiz quadrada do número total de variáveis em cada divisão.

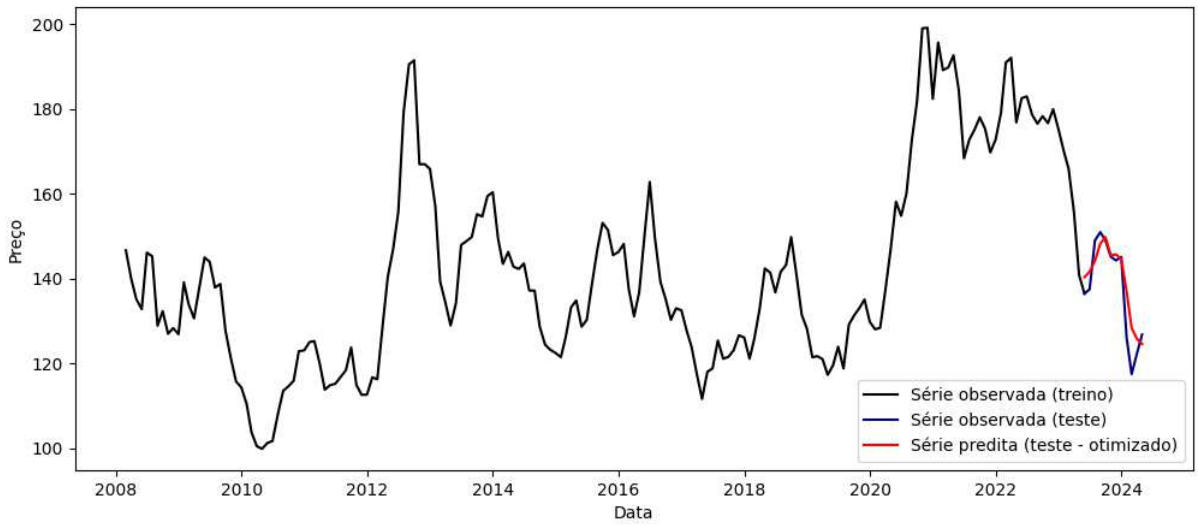
[2] Os valores de *hiperparâmetros* da grade criada são baseados nos seguintes: 'n_estimators': [50, 100, 150, 200, 250, 300], 'max_features': ['auto', 'sqrt'], 'max_depth': [5, 10, 20, 30, 40, 50], 'min_samples_split': [2, 5, 10, 15, 20], 'min_samples_leaf': [1, 2, 4, 6, 8, 10].

Embora não seja usual limitar o crescimento das árvores em um modelo de *bagging*, o algoritmo executado indicou que a melhor opção foi definir que cada árvore crescesse até 20 níveis de profundidade. Para estes parâmetros, o modelo obteve boa performance no conjunto de teste com MAE de 3,84, MAPE de 2,96%, MSE de 26,80 e RMSE de 5,18.

Um MAPE de 2.96% significa que, em média, as previsões do modelo têm um erro de cerca de 2.96% em relação aos valores reais. Os valores de MAE (3.84) e RMSE (5.18) são relativamente próximos, ideia sugestiva de que o modelo não está cometendo muitos erros extremos (*outliers*).

O gráfico da Figura 7 ilustra a série de dados, mas destaca o desempenho preditivo do modelo ao plotar a série de valores da previsão realizada com os dados da sequência de 12 meses correspondente ao conjunto de teste.

Figura 7: Série observada e previsão do preço real da soja (em R$/saca de 60 kg) com base no Random Forest.



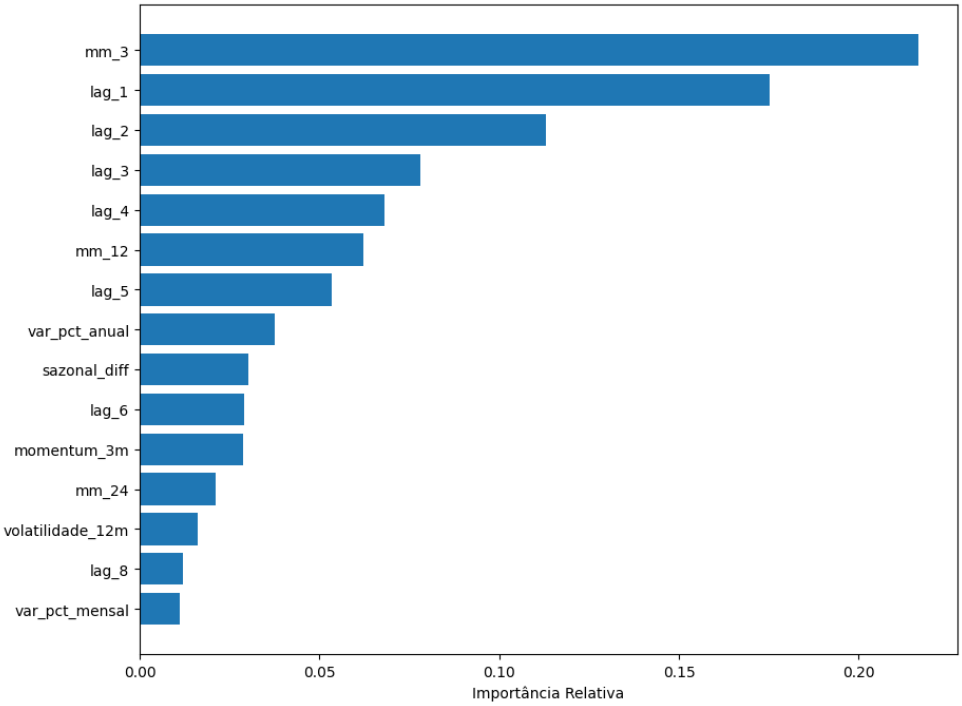
Fonte: Dados da pesquisa.

Após a previsão, investigamos a medida de importância das variáveis, para estabelecer quais dos atributos empregados apresentaram maior contribuição na redução do erro quadrático médio (MSE) do modelo. O gráfico de importância das variáveis, expresso na Figura 7 a seguir, mostra as 15 variáveis mais relevantes para o modelo Random Forest durante a modelagem do preço real da soja.

Neste gráfico, observa-se que as médias móveis e defasagens são cruciais para o modelo - o que é esperado em modelos de séries temporais. A importância da média móvel de três meses (mm_3) e das defasagens até a 5ª ordem (lag_5) sugere que o comportamento recente dos preços da soja tem intensiva influência nas previsões. A variável de *momentum*

(*momentum*_3m), que é a diferença entre o preço atual e o preço de três meses atrás, também se mostrou relevante. Ela capta a direção e magnitude da variação de preços em períodos curtos, o que será, decerto, útil para prever se o preço vai continuar subindo ou descendo.

Figura 8: Importância de variáveis (atributos) no Random Forest para a previsão do preço da soja.



Fonte: Dados da pesquisa.

O fato de que as variáveis como variação percentual anual (*var_pct_anual*) e diferença sazonal (*sazonal_diff*) estejam nesta lista sugere que o modelo também está captando padrões sazonais e mudanças estruturais ao largo de ciclos anuais, o que valida a inclusão destas variáveis. A existência de variáveis de volatilidade aponta que este aspecto, em consideração a incerteza e a estabilidade dos preços, também deve ser considerado.

4.4 Resultados – Lasso

Visando a aplicar uma abordagem alternativa no contexto dos modelos de *machine learning,* foi utilizado o Lasso. Vale evidenciar que, para garantir a consistência das penalizações impostas pelo Lasso, os preditores foram previamente padronizados.

Da mesma maneira que, no modelo Random Forest, foi aplicado um processo de validação cruzada sobre o conjunto de dados de treino, com o objetivo de otimizar o parâmetro de regularização λ. O valor ótimo encontrado foi $\lambda = 0{,}0225$, o que assegurou a melhor performance do modelo ao balancear ajuste e regularização.

A estimação do modelo Lasso produz coeficientes ajustados por meio de regularização. Esse processo reduz o valor de alguns coeficientes, aproximando-os de zero, e elimina completamente outros, definindo seus coeficientes como exatamente zero. Isso leva o Lasso a selecionar, automaticamente, as variáveis mais importantes, promovendo a simplicidade do modelo, ao manter apenas as variáveis mais relevantes para a predição. No caso deste experimento, os coeficientes das variáveis selecionadas pelo Lasso são expressos na Tabela 3, destacando os preditores mais importantes para a previsão do preço da soja.

Tabela 3. Coeficiente – Lasso.

Variável	Coeficiente
lag_3	12.0483
*momentum*_3m	7.7628
lag_1	5.6964
mm_3	5.4583
var_pct_mensal	2.4969
mm_12	0.2959
sazonal_diff	0.1217
mm_24	0.0739
var_pct_anual	0.0722
lag_4	0.0358
mes_9	0.0188
mes_10	0.0116
mes_2	0.0079
mes_12	0.0079
mes_6	-0.0015
mes_7	-0.0063
mes_3	-0.0098
volatilidade_3m	-0.0172
mes_5	-0.0202
lag_2	-0.5437

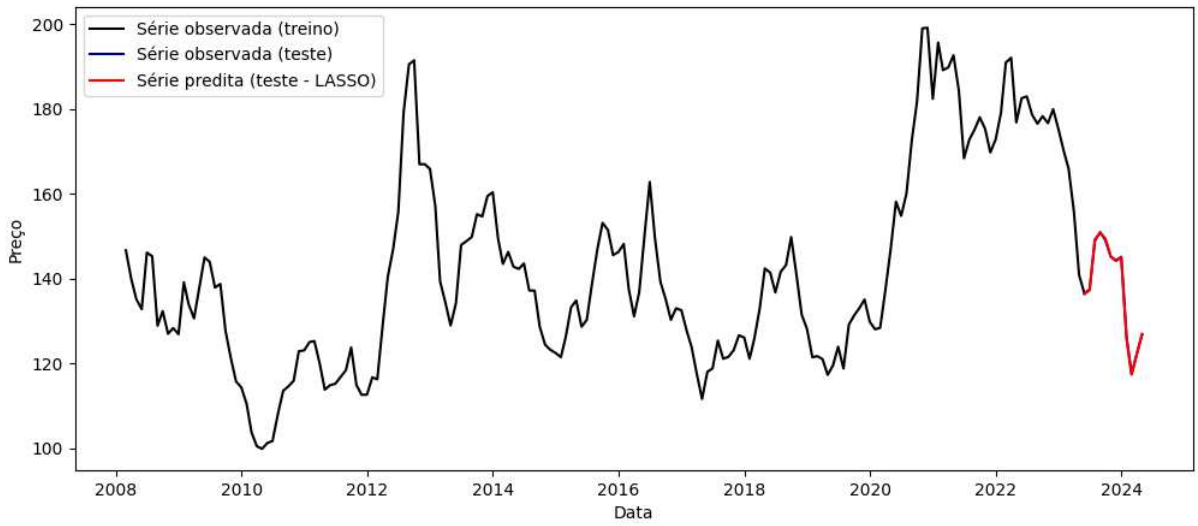
Fonte: Dados da pesquisa.

Conforme os preditores ora listados, é dado realizarem-se algumas inferências. Os valores defasados do preço, de um e três meses, também indigitaram vigorosa influência na previsão atual, sugerindo que os preços defasados são relevantes. O indicador de *momentum* (diferença entre o preço atual e o de três meses atrás) também é um coeficiente significativo, indicando que o modelo detectou um padrão importante relacionado ao movimento dos preços no período considerado. A média móvel de três meses e a variação percentual mensal também exibiram coeficientes relativamente altos, sugerindo que essas variáveis têm intensiva influência positiva na previsão do preço da soja. Além dessas variáveis, o modelo selecionou preditores que capturam efeitos de outras defasagens, diferenças sazonais e *dummies* para meses do ano, indicando que estes aspectos também são relevantes.

Após a fase de treino, o modelo otimizado foi aplicado aos dados de teste. As métricas obtidas foram as seguintes: MAE = 0.18, MAPE = 0,13%, MSE = 0.05, RMSE = 0.21. Estas medidas indicam que o modelo Lasso oferta um bom desempenho, com erros absolutos, quadráticos e percentuais muito baixos. Isso sugere que o modelo está fazendo previsões precisas, devendo ser havido como um bom preditor dos valores no conjunto de teste.

O gráfico da Figura 9 destaca o ajuste da previsão aos dados de teste. Percebe-se haver boa sobreposição entre a série de valores preditos e a série de valores observados, reforçando a ideia de que o modelo Lasso treinado demonstrou um bom desempenho preditivo.

Figura 9: Série observada e previsão do preço real da soja (em R$/saca de 60 kg) com base no Lasso.



Fonte: Dados da pesquisa.

4.5 Desempenho dos modelos de previsão

Neste estudo, foram testados três modelos de previsão - SARIMA, Random Forest e Lasso - aplicados a dados históricos dos preços da soja. Um dos principais objetivos foi avaliar o desempenho preditivo desses modelos, explorando a hipótese de que os métodos de *machine learning* uma performance preditiva superior em relação aos modelos tradicionais, como o SARIMA.

O desempenho de cada modelo foi avaliado utilizando um conjunto clássico de métricas de erro, incluindo Erro Médio Absoluto (MAE), Erro Percentual Absoluto Médio (MAPE), Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio (RMSE). Essas métricas fornecem uma visão abrangente da precisão dos modelos, medindo tanto a magnitude dos erros quanto os desvios em termos percentuais e quadráticos.

A Tabela 4 exprime os resultados das métricas de desempenho preditivo para cada modelo testado.

Tabela 4. Métricas de avaliação de desempenho/erro de previsão

Modelo	MAE	MAPE	MSE	RMSE
SARIMA	14,03	9,96%	245,27	15,66
Random Forest	3,84	2,96%	26,80	5,18
Lasso	0,18	0,13%	0,05	0,21

Fonte: Dados da pesquisa.

Os resultados indicam que o modelo SARIMA exprimiu uma performance inferior em relação aos métodos de *machine learning,* com erros mais elevados em todas as métricas. Entre os modelos testados, o Lasso se destacou, alcançando a melhor performance preditiva, com os menores valores de MAE, MAPE, MSE e RMSE, o que sugere sua maior adequação para a previsão dos preços da soja nesse contexto. O Random Forest também exibiu resultados superiores ao SARIMA, mas ficou aquém do desempenho obtido pelo Lasso em termos de precisão.

Desse modo, os resultados recolhidos nesta Dissertação encontram respaldo na literatura especializada, que destaca a superioridade de algoritmos de aprendizado de máquina em relação a abordagens tradicionais de séries temporais, especialmente em realidades complexas. Mohanty; Thakurta; Kar (2023) observaram que a regressão de árvore de decisão teve um desempenho superior a métodos como regressão linear e séries temporais, pois consegue captar as intricadas relações entre os fatores que afetam os preços. Palazzi *et al.* (2023), também, verificaram que a união da Análise Espectral Singular (SSA) com *machine learning* em

modelos híbridos ensejou previsões acuradas, melhores do que as dos modelos ARIMA e GARCH.

Os achados deste ensaio, em que o modelo Lasso superou o Random Forest e o SARIMA, vão ao encontro de trabalhos diversos mencionados na revisão de literatura. Yuan, San e Leong (2020) e Oktoviany; Knobloch; Korn (2021), por exemplo, obtiveram resultados promissores com modelos de *machine learning* para prever preços de *commodities* agrícolas, demonstrando a capacidade desses métodos de aprender relações complexas e se adaptar às flutuações do mercado. O Lasso, em particular, destaca-se por selecionar variáveis relevantes e estabelecer modelos parcimoniosos, o que justifica sua performance superior em relação ao Random Forest, mais propenso a *overfitting*.

Em conclusão, é fundamental ter em mente a ideia de que a performance dos modelos é passível de variar, a depender das particularidades dos dados, do horizonte de previsão e dos algoritmos de *machine learning* utilizados. Malgrado isso, os resultados aqui expressos e a literatura consultada apontam para a mesma direção: os métodos de *machine learning* demonstram resultados melhores para a projeção de preços de *commodities* agrícolas, com potencial para superar as limitações dos modelos tradicionais e fornecer informações precisas para auxiliar na tomada de decisões no setor.

5 CONSIDERAÇÕES FINAIS

A utilização de modelos de *machine learning* na previsão de preços de *commodities* agrícolas representa uma abordagem complementar aos métodos tradicionais de séries temporais, como ARIMA e SARIMA. Enquanto os modelos clássicos são eficazes em capturar padrões temporais e sazonalidades, os de *machine learning* têm a capacidade de assimilar uma ampla gama de variáveis e padrões não lineares componentes dos indicadores.

Tais ferramentas têm destaque pela sua flexibilidade na análise de preços de *commodities* agrícolas em decorrência da sua capacidade de aprender relações complexas e não lineares entre diversas variáveis. Esses modelos são propícios à adaptação automática a mudanças nas condições, identificando padrões que, com facilidade ou nem tanto, serão capturados por métodos tradicionais.

O desempenho superior dos algoritmos de aprendizado de máquina, evidenciado por métricas como MAE, MAPE e RMSE, reforça a relevância de metodologias mais flexíveis para lidar com múltiplos fatores e não linearidades que influenciam os preços. Em comparação, embora o SARIMA tenha se mostrado eficiente na captação de padrões sazonais, denotou desempenho inferior em um mercado tão volátil quanto o da soja, resultado coerente com estudos que evidenciam as limitações de modelos tradicionais perante dados complexos (Pindyck, 2001; James et *al*., 2013).

Em ultrapasse ao aprimoramento preditivo, os resultados confirmam a importância prática de previsões acuradas para o setor agropecuário, fornecendo subsídios mais robustos para a tomada de decisão. Informações precisas conduzem produtores, investidores e outros agentes econômicos a ajustarem suas estratégias para diminuir riscos e explorar oportunidades em um ambiente incerto, como sugerido por Oktoviany; Knobloch; Korn (2021) e Yuan; San; Leong (2020). Nessa contextura, a aplicação de técnicas de aprendizado de máquina demonstra-se a alternativa coerente para modernizar a gestão no agronegócio brasileiro e ampliar sua competitividade.

Havendo assim ocorrido, o experimento ora sob remate evidenciou o potencial da aprendizagem de máquina na previsão de preços de *commodities* agrícolas, com o modelo Lasso exprimindo performance superior ao Random Forest e ao SARIMA. O estudo contribuiu para a literatura, ao examinar a aplicação desses modelos na previsão de preços de soja no Brasil, utilizando dados do CEPEA e técnicas de engenharia de atributos. As limitações do ensaio, como a utilização de apenas uma série de preços e período relativamente curto dos dados, abrem caminhos para pesquisas em devir mais próximo ou mais distante, suscetíveis de insertar

variáveis explicativas adicionais, conjuntos de dados mais abrangentes e variegadas técnicas de *machine learning*. Outras abordagens, como redes neurais profundas e métodos de *ensemble learning*, são investigáveis, sendo também relevante expandir a análise para variadas *commodities*, reforçando a aplicabilidade e a robusteza dessas metodologias em distintos contextos econômicos.

REFERÊNCIAS

AHMED, N. *et al*. An empirical comparison of machine learning models for time series forecasting. **Econometric Reviews**, Örebro, v. 29, n. 5–6, p. 594–621, 2010.

ALPAYDIN, E. **Introduction to machine learning**. Cambridge: MIT Press, 2020.

BANCO MUNDIAL. **Commodity markets outlook**: The impact of the war in Ukraine on commodity markets. Washington, D.C.: World Bank, 2023. Disponível em: https://www.worldbank.org/commodity-markets. Acesso em: 17 out. 2024.

BARROS, Á. de M.; AGUIAR, D. R. D. Gestão do risco de preço de café arábica: uma análise por meio do comportamento da base. **Revista de Economia e Sociologia Rural,** Brasília, v. 43, n. 3, p. 443-464, 2005.

BARROS, G.; SPOLADOR, H.; BACCHI, M. Supply and demand shocks and the growth of the brazilian agriculture. **Revista Brasileira de Economia**, São Paulo, v. 63, n. jan-mar. 2009, p. 35-50, 2009.

BOX, G. *et al*. **Time series analysis**: forecasting and control. Hoboken: John Wiley & Sons, 2015.

BREIMAN, L. Random forests. **Machine Learning**, Berkeley, v. 45, n. 1, p. 5-32, 2001.

BROOKS, C.; PROKOPCZUK, M.; WU, Y. Commodity futures prices: more evidence on forecast power, risk premia and the theory of storage. **The Quarterly Review of Economics and Finance**, Illinois, v. 53, n. 1, p. 73-85, 2013.

CARVALHO, R. *et al*. Avaliação de algoritmos de machine learning na cotação do preço do contrato futuro de milho. **Revista Eletrônica e-Fatec**, Garça, v. 11, n. 1, 2021.

CASASSUS, J.; COLLIN-DUFRESNE, P. Stochastic convenience yield implied from commodity futures and interest rates. **The Journal of Finance**, Chicago, v. 60, n. 5, p. 2283-2331, 2005.

CORSINI, F. P.; RIBEIRO, C. de O. Dinâmica e previsão de preços de commodities agrícolascom o filtro de Kalman. *In:* ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO, 28, Rio de Janeiro, **Anais...** Rio de Janeiro: ABEPRO, 2008, p. 21-37.

COSTA, F. A.; ISHII, K. S.; MARTINES-FILHO, J. G.; MIQUELETO, G. J. Relação entre os preços dos mercados futuro e físico da soja: evidências para o mercado brasileiro. *In:* CONGRESSO DA SOCIEDADE BRASILEIRA DE ECONOMIA E SOCIOLOGIA RURAL – SOBER, 44, Fortaleza, **Anais...** Brasília, 2006. Disponível em: http://ageconsearch.umn.edu/bitstream/149189/2/1031.pdf. Acesso em: 17 out. 2024.

CUESTA, H.; KUMAR, S. **Practical data analysis**. 2nd. Birmingham: Packt Publishing, 2016. 306 p.

DEATON, A.; LAROQUE, G. On the behaviour of commodity prices. **The Review of Economic Studies**, Oxford, v. 59, n. 1, p. 1-23, 1992.

DOCKNER, E. J.; EKSI, Z.; RAMMERSTORFER, M. A Convenience Yield Approximation Model for Mean-Reverting Commodities. **Journal of Futures Markets**, Hoboken, v. 35, n. 7, p. 625- 654, 2015.

ENDE, M. V. **Comportamento dos preços dos contratos agropecuários negociados na BM&F**: a hipótese de "*normal backwardation*" no mercado futuro brasileiro. Dissertação de Mestrado em Administração – Escola de Administração, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

FAMA, E. F.; FRENCH, K. R. Commodity futures prices: Some evidence on forecast power, premiums, and the theory of storage. **Journal of Business**, Chicago, v. 60, n. 1, p. 55-73, 1987.

FOOD AND AGRICULTURE ORGANIZATION (FAO). **O estado da segurança alimentar e nutrição no mundo 2021**: Transformar os sistemas alimentares para que promovam dietas acessíveis e saudáveis para todos. Roma: FAO, 2021. Disponível em: https://www.fao.org/publications/sofi/2021/pt. Acesso em: 17 out. 2024.

FERNANDES, E. A.; LIMA, A. M. S.; DE AGUIAR, D. R. D. O comportamento do preço à vista e futuro de café no brasil: o hedge como opção de redução de risco. *In*: CONGRESSO DA SOCIEDADE BRASILEIRA DE ECONOMIA, ADMINISTRAÇÃO E SOCIOLOGIA RURAL – SOBER, 43, Ribeirão Preto, **Anais...** Ribeirão Preto, 2005, p. 36-42.

FRANKEL, J. A. Effects of speculation and interest rates in a “carry trade” model of commodity prices. **Journal of International Money and Finance**, Washington, v. 42, p. 88-112, 2014.

FONTES, R. E.; CASTRO JUNIOR, L. G. de; AZEVEDO, A. F. Estratégia de comercialização em mercados derivativos-descobrimento de base e risco de base da cafeicultura em diversas localidades de Minas Gerais e São Paulo. **Ciência e Agrotecnologia,** Lavras**,** v. 29, n. 2, p. 382-389, 2005.

GOMIDE, J.; VIGNOLI, L.; MACEDO, Y. Machine Learning Algorithms to Estimate Composite Mechanical Properties. *In*: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL, 20, Belo Horizonte, **Anais...** Belo Horizonte, 2023, p. 461-472.

GONÇALVES, D. F.; FRANCISCHINI, A. A.; ALVES, A. F.; PARRÉ, J. L. Análise de cointegração, causalidade e efetividade do hedge para os preços à vista e futuro do contrato de boi gordo para a região noroeste do Paraná. *In*: CONGRESSO DA SOCIEDADE BRASILEIRA DE ECONOMIA E SOCIOLOGIA RURAL – SOBER, 45, Londrina, **Anais...** Londrina, 2007, p. 27-32.

GORTON, G.; ROUWENHORST, K. Geert. Facts and fantasies about commodity futures. **Financial Analysts Journal**, New York, v. 62, n. 2, p. 47-68, 2006.

INSTITUTO DE PESQUISA ECONÔMICA APLICADA (IPEA). Mercados e preços agropecuários: análise do cenário recente e perspectivas. **Carta de Conjuntura,** Nota 5, maio 2023. Disponível em: https://www.ipea.gov.br/cartadeconjuntura/wp-content/uploads/2023/05/230503_nota_5_mercados_e_precos_agropecuarios.pdf. Acesso em: 13 out. 2024.

IZBICKI, R.; SANTOS, TM. **Machine Learning sob a ótica estatística**: uma abordagem preditivista para a estatística com exemplos em R, 2018. Disponível em: http://www.rizbicki.ufscar.br/sml.pdf. Acesso em: 06 dez. 2023. [versão em desenvolvimento]

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning**: With Applications in R. New York: Springer, p. 303-314, 2013.

JOSEPH, K.; IRWIN, S. H.; GARCIA, P. Commodity Storage under Backwardation: Does the Working Curve Still Work? **Applied Economic Perspectives and Policy**, Hoboken, v. 38, n. 1, p. 152-173, 2015.

KALDOR, N. Speculation and economic stability. **The Review of Economic Studies**, Oxford, v. 7, n.1, p. 1-27, 1939.

KEYNES, J. M. **Treatise on money**: Pure theory of money. v.1. London: Macmillan, 1930.

KHAN, U. *et al*. A robust regression-based stock exchange forecasting and determination of correlation between stock markets. **Sustainability**, Basel, v. 10, n. 10, p. 3702, 2018.

KIM, S.; KIM, J.; HEO, E. Convenience yield of accessible inventories and imports: A case study of the Chinese copper market. **Resources Policy**, Amsterdã, v. 52, p. 277-283, 2017.

MILONAS, N.; PARATSIOKAS, N. Convenience Yields in Electricity Prices: Evidence from the Natural Gas Market**. Journal of Futures Markets**, Hoboken, v. 37, n. 5, p. 522-538, 2017.

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO (MAPA). **Anuário da soja 2022/2023**. Brasília, DF: MAPA, 2023.

MOHANTY, M. K..; THAKURTA, P. K. G.; KAR, S. Agricultural commodity price prediction model: a machine learning framework. **Neural Computing and Applications**, Londres, v. 35, n. 20, p. 15109-15128, 2023.

NAKAJIMA, K. Commodity spot and futures prices under supply, demand, and financial trading. Tóquio: **Browser Download This Paper**, 2015.

NAKAJIMA, K. Commodity Spot, Forward Prices, and Convenience Yield Under Incomplete Market. **Working paper /SSRN,** mar. 2017. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2952662. Acesso em: 17 out. 2024.

NG, V. K.; PIRRONG, S. C. Fundamentals and volatility: Storage, spreads, and the dynamics of metals prices. **Journal of Business**, Chicago, p. 203-230, 1994.

OKTOVIANY, P.; KNOBLOCH, R.; KORN, R. A machine learning-based price state prediction model for agricultural commodities using external factors. **Decisions in Economics and Finance**, Milão, v. 44, n. 2, p. 1063-1085, 2021.

OLIVEIRA, J.; SILVA, L. Aprendizado de máquina para previsão de preços de commodities: Uma revisão sistemática da literatura. **Revista de Administração da Universidade Federal de Minas Gerais**, Belo Horizonte, 57, n.3, p. 330-352, 2022.

OMURA, A.; WEST, J. Convenience yield and the theory of storage: applying an option-based approach. **Australian Journal of Agricultural and Resource Economics**, Canberra, v. 59, n. 3, p. 355-374, 2015.

PALAZZI, R. *et al*. Forecasting commodity prices in Brazil through hybrid SSA-complex seasonality models. **Production**, São Carlos, v. 33, 2023.

PEREIRA, L. M. **Modelo de formação de preços de commodities agrícolas aplicado ao mercado de açúcar e álcool**. 2009. Tese (Doutorado em Administração) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2009.

PINDYCK, R. S. The dynamics of commodity spot and futures markets: a primer. **The energy journal**, Cleveland, v. 22, n. 3, p. 1-29, 2001.

PINTO, L. T. B. M. **Estoques de Óleo Cru e o Benefício de Conveniência (*Convenience Yield*): uma Análise do Mercado Norte-Americano**. 2015. Dissertação (Mestrado em Planejamento Energético) - Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

RASCHKA, S. **Python machine learning**. Birmingham: Packt Publishing Ltd, 2015.

RASMUSSEN, C. *et al*. **Gaussian processes for machine learning**. Cambridge, MA: MIT Press, 2006.

SANTOS, J. E. Normal backwardation é normal no mercado futuro brasileiro? **Relatório de Pesquisa nº 15/1998**. Escola de Administração de Empresas de São Paulo, Fundação Getúlio Vargas, Núcleo de Pesquisas e Publicações, 1998. Disponível em: http://gvpesquisa.fgv.br/sites/gvpesquisa.fgv.br/files/publicacoes/P00138_1.pdf. Acesso em: 10 out. 2024.

SILVA, A.; OLIVEIRA, M. Commodities agrícolas e desenvolvimento econômico no Brasil. **Revista de Economia Política**, São Paulo, v. 42, n. 4, p. 637-660, 2022.

SILVEIRA, G.; AMBROZINI, M. Análise da teoria da estocagem sobre a base dos contratos futuros de soja no Brasil. **Nucleus**, Ituverava, v. 15, n. 2, p. 401-421, 2018.

SOUZA, T. **Previsão de preços e demandas de produtos do varejo utilizando técnicas de aprendizado de máquina**. 2021. 106 p. Dissertação (Mestrado em Engenharia de Sistemas e Automação) – Universidade Federal de Lavras, Lavras, 2021.

TONIN, J. M.; TONIN, J. R.; TONIN, G. M. Operações de Hedge no Mercado da Soja: uma análise comparativa para o Estado do Paraná. **Revista Paranaense de Desenvolvimento-RPD**, Curitiba, n. 115, p. 7-30, 2011.

TORRES, S. **Previsão do preço de ações brasileiras utilizando redes neurais artificiais**. 2021. 51 f. Dissertação (Mestrado em Administração de Empresas) - Universidade Presbiteriana Mackenzie, São Paulo, 2021.

WITTEN, D.; JAMES, G. **An introduction to statistical learning**: with applications in R. New York: Springer; 2013.

WITTEN, I. H. *et al*. **Practical machine learning tools and techniques**. *In*: Data mining. Amsterdam, The Netherlands: Elsevier, p. 403-413, 2005.

WOOLDRIDGE, J. M. **Econometric Analysis of Cross Section and Panel Data**. Cambridge: The MIT Press, 2010.

WORKING, H. Theory of the inverse carrying charge in futures markets. **Journal of Farm Economics**, Oxford, v. 30, n. 1, p. 1-28, 1948.

YUAN, C. Z.; SAN, W. W.; LEONG, T. W. Determining optimal lag time selection function with novel machine learning strategies for better agricultural commodity prices forecasting in Malaysia. *In*: ASSOCIATION FOR COMPUTING MACHINERY (ACM). **Proceedings of the 2020 2nd International Conference on Information Technology and Computer Communications**. New York: ACM, p. 37-42, 2020.