Previsibilidade dos preços futuros da commodity Soja: Aplicando técnicas de Machine Learning[1]

Resumo

Este artigo investiga a previsibilidade dos preços futuros da soja na *Chicago Board of Trade* (CBOT) via *Machine Learning* (ML), testando se a informação pública supera a *baseline Random Walk*. A volatilidade desses preços impacta fortemente a economia brasileira. Utiliza-se série mensal com 300 observações brutas (2000–2024), integrando 12 variáveis exógenas ancoradas na teoria econômica, posteriormente reduzidas via RFECV. A metodologia compara *Random Forest, XGBoost, LightGBM, Ridge,* e *Elastic Net*, com calibração via validação cruzada temporal, avaliação final em conjunto de teste e teste de Diebold-Mariano. Os resultados demonstram a superioridade estatística dos *ensembles*: o *Random Forest* obteve RMSE de 0,0557 USD/bushel, reduzindo o erro em 30,1% ante o *Random Walk* (p=0,0122). A análise SHAP revelou defasagens de preços, estoques sul-americanos, dólar, petróleo e volatilidade implícita como principais preditores. O estudo oferece um modelo robusto e avaliado temporalmente para gestão de risco em *commodities*.

Palavras-chave: Previsão de Preços. Soja. *Machine Learning*. Séries Temporais. Gestão de Risco.

Abstract

This article investigates the predictability of soybean futures prices on the Chicago Board of Trade (CBOT) using Machine Learning (ML), testing whether public information outperforms the Random Walk baseline. The volatility of these prices strongly impacts the Brazilian economy. A monthly series with 300 raw observations (2000–2024) is used, incorporating 12 variables selected on the basis of economic theory and then processed via RFECV. The methodology compares Random Forest, XGBoost, LightGBM, Ridge, and Elastic Net, with calibration through temporal cross-validation, final evaluation on a test set, and the Diebold-Mariano test. Results demonstrate the statistical superiority of ensemble methods: Random Forest achieved an RMSE of 0.0557 USD/bushel, reducing the error by 30.1% relative to the Random Walk (p = 0.0122). SHAP analysis identified price lags, South-American inventories, the US dollar, oil, and implied volatility as the main predictors. The study offers a robust, temporally validated model for commodity risk management.

Keywords: Price Forecasting. Soybean. *Machine Learning*. Time Series. Risk Management.

[1] Área Temática: Econometria Financeira. Códigos JEL: C45; C53; C58; G13; G14; Q11

1. Introdução

O mercado de soja constitui um dos pilares estratégicos do agronegócio global, e o Brasil ocupa o centro desse cenário como o maior produtor e exportador mundial, o país colheu aproximadamente 150 Mt em 2024, gerando receitas de bilhões de dólares e respondendo por cerca de 15% do total das exportações brasileiras (Secex, 2024; USDA/FAS, 2025). A *commodity* desempenha um papel estruturante para a economia nacional; junto ao petróleo e ao minério de ferro, representou aproximadamente 40% das exportações totais entre 2022 e 2024, possuindo uma cadeia produtiva que engloba milhões de agentes (CONAB, 2024). A China, o principal *driver*, concentra 73% das compras da soja brasileira, o que cria dependência estrutural e amplifica a sensibilidade dos preços a choques de oferta (USDA/FAS, 2025).

O principal risco operacional dessa cadeia consiste na volatilidade de preços: oscilações de 30–40% ao ano são documentadas pela CBOT (*Chicago Board of Trade*), pressionando receitas de exportação e a renda dos produtores, de acordo com FAO (2017) e o World Bank (2022).

Apesar da relevância estratégica, a literatura apresenta três lacunas específicas: (i) ausência de modelos voltados sistematicamente ao mercado futuro CBOT com variáveis multidimensionais integradas no contexto brasileiro (Sezer et al., 2020); (ii) modelos econométricos tradicionais não capturam adequadamente dinâmicas não-lineares em períodos turbulentos como choques geopolíticos (Timmermann, 2008); e (iii) a aplicação de ML em *commodities* agrícolas brasileiras com protocolo rigoroso de validação temporal e interpretabilidade econômica permanece escassa (Lima e Oliveira, 2022).

No entanto, o entusiasmo com *Machine Learning* (ML) merece ponderação à luz da Hipótese dos Mercados Eficientes conforme Fama (1970): se os preços futuros incorporam toda a informação pública disponível, o melhor preditor de P_{t+1} é o próprio P_t, representado pelo processo *Random Walk*. Makridakis et al. (2020) mostraram que, em séries mensais com baixa razão sinal-ruído, algoritmos de *Machine Learning* empregados isoladamente não necessariamente superam combinações simples de métodos clássicos de previsão. No caso de *commodities*, Irwin e Good (2010) evidenciam que previsões do USDA (*United States Department of Agriculture*) frequentemente não apresentam superioridade estatisticamente significativa em relação a modelos econométricos convencionais, enquanto Timmermann (2008) documenta que a previsibilidade de retornos é instável ao longo do tempo, o que limita o desempenho tanto de modelos lineares quanto não lineares.

Nesse contexto, a questão empírica central não é assumir superioridade abstrata de um método, mas verificar em quais condições o aprendizado de máquina agrega valor preditivo incremental em relação ao *Random Walk* e aos modelos tradicionais. Assim, esta pesquisa contribui por meio de: (i) comparação sistemática de algoritmos de ML com validação *walk-forward*; (ii) interpretabilidade econômica via SHAP que revela os determinantes de preços da soja no período 2000–2024; e (iii) teste formal de superioridade via Diebold–Mariano.

1.1 Importância Estratégica da Soja

O mercado mundial de soja é estruturado em torno de um oligopólio de oferta. Em 2024, Brasil, Estados Unidos e Argentina responderam por aproximadamente 80% da produção global: o Brasil liderou com 155 Mt (39%), seguido pelos EUA com 112,5 Mt (28%) e pela Argentina com 50 Mt (13%) (USDA, 2024). Do lado da demanda, a China concentra cerca de 60% do fluxo mundial, tendo importado 105 Mt em 2024, representando crescimento de 6,5% sobre o ano anterior. Esse aumento na demanda foi impulsionado pela expansão contínua do rebanho suíno e pela procura por farelo proteico (GACC, 2024; USDA/FAS China, 2025).

No Brasil, a soja consolida-se como uma das principais lideranças no ranking de exportações com participação de 15,8% do valor FOB total em 2024, respondendo, junto ao petróleo bruto e ao minério de ferro, por aproximadamente 40% das exportações nacionais entre 2022 e 2024 (MDIC, 2024a,b). Essa concentração expõe o país à volatilidade dos preços internacionais e à demanda externa, criando demanda por instrumentos robustos de *forecasting*.

A concentração da oferta em três países e a dependência estrutural da China criam um mecanismo de amplificação sistêmica: onde choques localizados como uma estiagem, uma restrição logística nos portos ou uma mudança na política de estoques de Pequim, tendem a propagar-se rapidamente ao mercado global e, por transmissão quase imediata, esses eventos impactam a renda de milhões de produtores brasileiros (Margarido et al., 2007; World Bank, 2022). Para a balança comercial, oscilações de 10% no preço da soja representam variações da ordem de US$ 5 bilhões nas receitas anuais de exportação (CONAB, 2024; MDIC, 2024a). Esse grau de exposição justifica, do ponto de vista macroeconômico, o desenvolvimento de modelos preditivos capazes de antecipar movimentos de preço com antecedência suficiente para decisões de *hedge*, planejamento de safra e calibração de políticas de preço mínimo (Barnett e Coble, 2020; FAO, 2022)

Do ponto de vista da cotação, o preço da soja na CBOT exibiu quatro ciclos de alta documentados entre 2000 e 2024, conforme a Figura 1. O primeiro ocorreu em 2004, impulsionado por demanda crescente da China; o segundo em 2008, com a alta dos preços de energia e restrições à exportação que comprimiram estoques (FAO, 2008). Em 2012, uma severa estiagem nos EUA elevou as cotações a máximas próximas a US$ 17,00/bushel. O ciclo mais recente, iniciado em 2020, combinou forte demanda asiática, disrupções logísticas da pandemia de COVID-19 e escassez de fertilizantes agravada pelo conflito na Ucrânia (World Bank, 2022; IFPRI, 2022).

Entre 2023 e 2024, uma tendência de queda de aproximadamente 20% foi registrada, impulsionada pela ampla oferta global, reflexo do aumento da safra norte-americana, recuperação da safra argentina e volume brasileiro abaixo do potencial máximo (Silveira, 2024). Conforme Embrapa (2023, p. 32), "a instabilidade de preços da soja decorre de sua sensibilidade a fatores climáticos, econômicos e geopolíticos, dificultando projeções confiáveis", o que reforça a demanda por modelos preditivos capazes de capturar dinâmicas não-lineares.

A recorrência desses ciclos de alta e baixa, separados por reversões abruptas de 20–40%, não

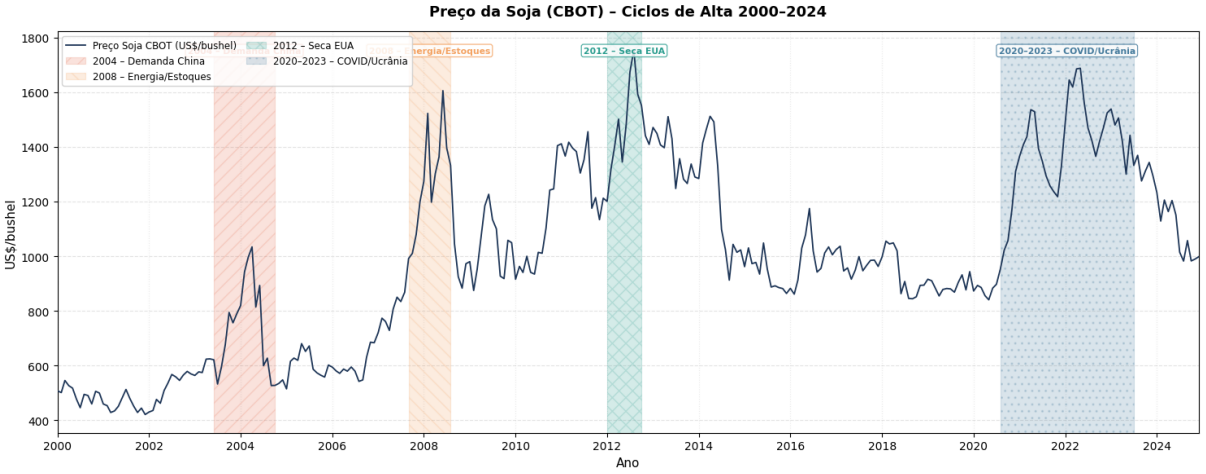

Figura 1: Preço histórico mensal da soja na CBOT (2000–2024).
Fonte: Elaboração própria com dados do CME Group.

é apenas um fenômeno histórico: ela configura o desafio central de qualquer modelo preditivo aplicado ao mercado de soja. Gilbert (2010) demonstrou que picos de preço em *commodities* agrícolas são frequentemente originados por choques simultâneos de oferta e demanda com componentes especulativos, tornando modelos lineares inadequados para capturar os pontos de inflexão. Tadesse et al. (2014) evidenciam que a volatilidade estrutural de preços agrícolas se intensificou após 2007, com episódios de *clustering* que demandam abordagens sensíveis a regimes. Esse padrão é precisamente o que motiva a adoção de algoritmos de ML com validação *walk-forward*: identificar os períodos em que sinais preditivos emergem da volatilidade, conforme documentado por Timmermann (2008).

Além do desempenho, é importante ressaltar que, no mercado futuro, o preço da soja incorpora expectativas coletivas sobre oferta e demanda: estoques, clima nas regiões produtoras, demanda chinesa, força do dólar e custos de energia. Em outras palavras, o preço futuro (F_t) relaciona-se ao preço à vista (S_t) pela *teoria do armazenamento* (Working, 1949; Deaton e Laroque, 1992), conforme:

$$F_t \approx S_t \cdot e^{(r+c-y)T}, \tag{1}$$

em que r é a taxa de juros, c é o custo de armazenagem, y é o *convenience yield* e T representa o tempo até a entrega. Quando os estoques estão elevados, y tende a cair e o mercado opera em *contango*[2] ($F_t > S_t$); quando há risco de escassez, y sobe e pode ocorrer *backwardation*[3] ($F_t < S_t$).

[2] *Contango* descreve a situação em que o preço futuro supera o preço à vista ($F_t > S_t$), refletindo os custos de carrego — juros, armazenagem e seguro — que se acumulam ao longo do tempo até o vencimento do contrato (Working, 1949; Hull, 2009).

[3] *Backwardation* corresponde à situação oposta, em que o preço futuro é inferior ao preço à vista ($F_t < S_t$). Esse fenômeno ocorre quando os agentes atribuem elevado *convenience yield* aos estoques físicos imediatos — prêmio de escassez —, tornando a posse do bem presente mais valiosa do que a promessa de entrega futura (Deaton e Laroque, 1992; **?**).

A transmissão do preço de referência internacional ao mercado físico brasileiro ocorre via *paridade de exportação* (Banco Central do Brasil, 2022; CEPEA, 2023). Essa relação deriva diretamente da expressão (1): partindo do preço futuro F_{CBOT} — determinado pela teoria do armazenamento —, a cotação doméstica em reais por saca é obtida incorporando o diferencial logístico e comercial (*basis*) e a taxa de câmbio, numa aplicação da lei do preço único ajustada ao mercado físico brasileiro (Banco Central do Brasil, 2022; CEPEA, 2022):

$$P_{BR}\,(\text{R\$/sc}\,60\,\text{kg}) \;=\; \big(F_{\text{CBOT}}\,(\text{US\$/bu}) + \text{Basis}\big) \times \text{USD/BRL} \times \frac{60}{27{,}216}, \tag{2}$$

em que a fração $60/27{,}216$ converte *bushels* em sacas de 60 kg; o *basis* representa o diferencial logístico e comercial entre o preço local e o de referência internacional, refletindo os custos de frete, qualidade do grão, competição entre os EUA e o Brasil e as barreiras comerciais (CEPEA, 2022; USDA, 2023). Portanto, oscilações em F_{CBOT} — conforme descrito pela expressão (1) —, no câmbio ou no *basis* geram impactos amplificados na cotação em reais por saca, justificando a relevância das variáveis financeiras e macroeconômicas no modelo preditivo apresentado na Seção 3. A magnitude dessa transmissão é quantificada pela literatura empírica: Margarido et al. (2007) estimaram que variações na CBOT explicam aproximadamente 87% da variância dos preços físicos pagos ao produtor brasileiro, com transmissão sem defasagem significativa nos períodos de alta liquidez. Plato e Chambers (2004) confirmam que o Brasil opera sob regime de paridade de exportação plena, tornando o contrato futuro da CBOT o principal instrumento de *price discovery* para toda a cadeia nacional. Consequentemente, aprimorar a previsão do retorno do contrato futuro CBOT, principal motivação deste artigo, equivale diretamente a melhorar a capacidade de antecipação de preços para produtores, cooperativas e *trading houses* brasileiras (CEPEA, 2023; Hull, 2009), com implicações diretas para estratégias de *hedge* e contratos de venda antecipada.

A literatura empírica confirma a robustez dessa transmissão entre o mercado internacional e o mercado doméstico brasileiro. Margarido et al. (2007) mostram que a formação de preços no Brasil permanece estruturalmente vinculada às cotações da CBOT via paridade de exportação, enquanto Plato e Chambers (2004) evidenciam elevado poder explicativo das cotações internacionais sobre o preço doméstico. Assim, embora fatores locais como câmbio, prêmio de exportação e logística possam gerar desvios temporários, o contrato futuro da CBOT segue como principal referência de *price discovery* para a soja brasileira, justificando o foco empírico deste artigo na previsão do retorno desse mercado.

2. Discussão Teórica

2.1 Hipótese de Mercados Eficientes

A Hipótese dos Mercados Eficientes (HME), proposta por Fama (1970), postula que os preços de ativos refletem instantaneamente toda a informação disponível, impossibilitando

ganhos sistemáticos de previsão. Em sua forma semiforte, a HME implica que o melhor preditor do preço futuro é o preço corrente (*Random Walk*), tornando qualquer modelo preditivo estatisticamente equivalente ao passeio aleatório. Para mercados futuros agrícolas, Fama e French (1987) documentaram comportamento próximo ao *Random Walk* na dimensão direcional, especialmente em horizontes superiores a três meses. Malkiel (2003), em revisão abrangente das críticas à HME, conclui que os mercados são mais eficientes e menos previsíveis do que sugerem os resultados empíricos de curto prazo, uma vez que padrões de previsibilidade tendem a se autodestrutir quando amplamente explorados pelos agentes.

A HME, contudo, não é irrefutável. Timmermann e Granger (2004) formalizam a tensão entre eficiência e previsibilidade ao demonstrar que o próprio processo de busca por padrões preditivos gera não-estacionariedades nas séries de retornos, tornando a fronteira entre eficiência e exploração essencialmente dinâmica. Na presença de não-linearidades, fricções informacionais e assimetrias de acesso a dados, *previsibilidade residual* pode persistir — entendida como a componente sistemática dos retornos que sobrevive ao ajuste pelo prêmio de risco de equilíbrio e que pode ser capturada por modelos suficientemente flexíveis, sem necessariamente implicar arbitragem lucrativa livre de risco (Campbell et al., 1997). Grossman e Stiglitz (1980) argumentam que, em equilíbrio, os mercados jamais podem ser perfeitamente eficientes: agentes precisam de incentivo econômico para coletar e processar informação, o que implica algum grau de previsibilidade como compensação pelo esforço informacional.

Lo (2004) aprofunda esse argumento ao propor a *Hipótese dos Mercados Adaptativos* (HMA), segundo a qual eficiência e previsibilidade coexistem de forma evolutiva: em períodos de estabilidade, os mercados convergem para a eficiência; em momentos de ruptura estrutural — choques climáticos, crises financeiras ou mudanças regulatórias —, janelas transitórias de previsibilidade emergem à medida que os agentes recalibram suas heurísticas. Essa perspectiva é particularmente relevante para mercados de *commodities* agrícolas, cujos preços são sujeitos a descontinuidades sazonais, eventos climáticos extremos e fluxos informacionais assimétricos oriundos de relatórios governamentais (Working, 1949).

No contexto de *commodities* agrícolas, a sazonalidade estrutural dos ciclos de plantio–colheita e a divulgação periódica de relatórios oficiais (USDA, CONAB) criam janelas de previsibilidade que coexistem com a eficiência de longo prazo. A microestrutura desses mercados amplifica essa dinâmica: a chegada assíncrona de informações privadas como, estimativas de safra, posições de grandes investidores e fluxos de capital especulativo, gera ruído de curto prazo que se sobrepõe ao sinal fundamental, tornando a separação entre componente sistemática e idiossincrática um desafio empírico central (Campbell et al., 1997; Timmermann e Granger, 2004). A evidência mais recente sugere que modelos não-lineares de aprendizado de máquina são particularmente adequados para capturar essa previsibilidade residual: Gu et al. (2020) demonstram, em ampla análise comparativa, que árvores de decisão e redes neurais superam consistentemente modelos lineares na previsão de retornos de ativos, com ganhos atribuídos à captura de interações não-lineares entre preditores que os métodos tradicionais ignoram. Resultado análogo é reportado

por Zheng e Shi (2024) especificamente para mercados futuros de *commodities*: utilizando 22 contratos futuros, os autores documentam que modelos de *machine learning* com variáveis macroeconômicas e de microestrutura superam o *Random Walk* em previsão fora da amostra, evidência direta de previsibilidade residual persistente nesses mercados.

Essa previsibilidade, todavia, é modesta em magnitude, mas consistente com os limites impostos pela eficiência de mercado e heterogênea ao longo do tempo, concentrando-se nos regimes ou periodos de alta volatilidade. A investigação empírica conduzida neste trabalho posiciona-se precisamente nessa fronteira: ao empregar um modelo multivariado de *Random Forest* com variáveis financeiras, macroeconômicas e de microestrutura, busca-se quantificar a previsibilidade residual nos *log-retornos* do contrato futuro de soja na CBOT, com R^2 fora da amostra como métrica central de evidência (Campbell e Thompson, 2008).

2.2 Mercado Futuro de Commodities Agrícolas

Os mercados futuros de *commodities* agrícolas cumprem duas funções econômicas fundamentais: a transferência de risco (*hedging*) e a descoberta de preços (Hull, 2009). No caso da soja, a Chicago Board of Trade (CBOT), operada pelo CME Group, constitui a principal referência mundial para formação de preços, com contratos futuros padronizados em lotes de 5.000 bushels negociados em vencimentos mensais ao longo do ano. A liquidez e a profundidade desse mercado atraem tanto agentes físicos (produtores, cooperativas e *tradings*) quanto especuladores financeiros, cuja participação crescente desde os anos 2000 tem sido associada ao aumento da co-movimentação entre *commodities* e ativos financeiros (Tang e Xiong, 2012).

A transmissão de preços entre o mercado futuro internacional e o mercado doméstico brasileiro opera pelo mecanismo de paridade de exportação: o preço doméstico em reais converge para o preço CBOT ajustado pelo prêmio de exportação (*basis*) e pela taxa de câmbio BRL/USD (Margarido et al., 2007). A relação não é, contudo, estática: choques de oferta locais, variações cambiais extremas e colapsos logísticos, como o registrado na supersafra 2023, quando o prêmio de exportação tornou-se fortemente negativo, podem desacoplar temporariamente os dois mercados, introduzindo não-linearidades relevantes para a modelagem preditiva (Irwin e Sanders, 2015).

O mecanismo de formação de expectativas tem sido amplamente investigada. Karali et al. (2020) demonstraram que, após correção de erros de mensuração nos *surprises* de mercado, o poder explicativo dos fundamentos de oferta e demanda para os retornos dos contratos futuros de grãos supera 70%, evidenciando que os preços futuros incorporam ativamente informação econômica relevante e não apenas ruído especulativo. Em complemento, Li e Hayes (2017) investigaram as relações de liderança entre os mercados futuros de soja nos Estados Unidos, no Brasil e na China via co-integração com limiar, concluindo que a CBOT mantém dominância no longo prazo, mas que o mercado brasileiro frequentemente lidera o norte-americano durante a safra sul-americana, o que implica que as expectativas formadas na CBOT incorporam dinamicamente informações produtivas do Cone Sul.

Por fim, Etienne et al. (2020) avaliaram o desempenho preditivo dos contratos futuros no complexo da soja via regressão quantílica e concluíram que esses contratos produzem previsões não enviesadas para quantis centrais, mas tendem a subestimar reversões em períodos de preços extremos. Esse resultado evidencia a presença de assimetrias estruturais nos retornos e justifica o uso de modelos de *ensemble* com capacidade de captura de não linearidades, como os adotados neste trabalho.

Do ponto de vista de evidência empírica sobre previsibilidade de preços, a literatura indica que a previsibilidade das *commodities* agrícolas não é uniforme entre produtos, mercados e horizontes de previsão. Em estudos com *commodities* brasileiras, Araujo et al. (2020) identificaram forte heterogeneidade na eficiência informacional: o mercado de café apresentou a menor previsibilidade, enquanto o mercado de carne suína foi o mais previsível. Os autores também observaram que pares como etanol/açúcar e soja/milho ocupam posições próximas no espaço complexidade-entropia, sugerindo interdependência estrutural entre os ativos agrícolas.

No caso dos cereais, Kwas et al. (2022) analisaram a previsibilidade mensal dos preços de cevada, milho, arroz e trigo no período de 1980 a 2019. Para tanto, os autores construíram quatro fatores latentes via análise de componentes principais (PCA), cada um sumarizando uma dimensão distinta do ambiente econômico internacional: (*i*) fator de *commodities*, extraído dos preços de um amplo painel de produtos primários e capturando o componente comum de oscilação nos mercados de matérias-primas; (*ii*) fator cambial, obtido a partir das taxas de câmbio das principais moedas de países exportadores de grãos, refletindo a competitividade relativa das exportações agrícolas; (*iii*) fator financeiro, construído com base em spreads de crédito, volatilidade implícita e retornos de índices acionários globais, representando o apetite de risco dos agentes; e (*iv*) fator de atividade macroeconômica, derivado de indicadores de produção industrial e comércio global, sinalizando a demanda agregada por insumos alimentares. Os resultados mostraram que modelos que incorporam esses quatro fatores simultaneamente superam o *Random Walk* fora da amostra de forma estatisticamente significativa, especialmente em horizontes de seis a doze meses, embora o fator de *commodities* isolado já concentre grande parte do poder preditivo — resultado consistente com a hipótese de que um componente cíclico comum impulsiona os preços agrícolas globais.

Para o Brasil, Palazzi et al. (2023) avaliaram modelos híbridos baseados em *Singular Spectrum Analysis* (SSA) para prever os preços mensais à vista de milho, soja e açúcar. A abordagem decompõe a série temporal em componentes de tendência, oscilação e ruído, que são depois recombinados para alimentar modelos preditivos individuais. Os autores concluíram que a abordagem híbrida apresenta desempenho superior aos modelos individuais, com redução dos erros de previsão em várias janelas e maior robustez para capturar a sazonalidade complexa das séries.

Na mesma direção, Sari et al. (2024) testaram três estruturas para prever onze *commodities* agrícolas. O modelo GA-ELM (*Genetic Algorithm – Extreme Learning Machine*) combina uma rede neural de camada oculta única com treinamento analítico ultrarrápido — substituindo a

retropropagação por uma solução de mínimos quadrados — cujos pesos da camada de entrada são otimizados por um algoritmo genético (GA) para maximizar a capacidade de generalização fora da amostra. O modelo GA-LSTM (*Genetic Algorithm – Long Short-Term Memory*), por sua vez, acopla o mesmo mecanismo evolutivo de seleção de hiperparâmetros a uma rede recorrente com células de memória de longo e curto prazo, especialmente adequada para capturar dependências temporais de longa duração em séries não estacionárias. Os experimentos mostraram que o GA-ELM superou tanto o GA-LSTM quanto o ARIMA em termos de RMSE e MAE fora da amostra, resultado atribuído à maior simplicidade estrutural do ELM diante de séries mensais de tamanho moderado, nas quais o LSTM tende ao sobreajuste. Esses resultados reforça a ideia de que técnicas de aprendizado de máquina podem capturar relações não lineares e mudanças estruturais com maior eficácia do que modelos lineares tradicionais, desde que a complexidade do modelo seja calibrada ao tamanho amostral disponível.

Para a soja Xiong e Hu (2021) mostraram que os modelos Dynamic Model Averaging (DMA) e Dynamic Model Selection (DMS) superam o modelo AR, a regressão linear com todos os preditores e o random walk na previsão de contratos futuros chineses. Os autores destacam ainda que os melhores preditores variam ao longo do tempo, o que confirma a instabilidade da previsibilidade em mercados agrícolas.

Por sua vez, Kurumatani (2020) propôs previsões com redes neurais recorrentes para preços agrícolas e constatou que o LSTM, quando bem treinado, alcança melhor desempenho que GRU e SRNN tanto em acurácia quanto na preservação de características estatísticas da série. O estudo também reforça que a escolha do método depende do horizonte e da métrica de avaliação adotada.

A revisão da literatura permite organizar as principais famílias de modelos em torno de um *trade-off* central entre capacidade expressiva e risco de sobreajuste, condicionado pelo tamanho amostral e pela razão sinal-ruído da série em questão. Modelos lineares clássicos como ARIMA, VAR e suas extensões cointegradas, apresentam como vantagem principal a parcimônia e a interpretabilidade: cada coeficiente possui significado econômico direto e os testes de diagnóstico são bem estabelecidos na literatura (Timmermann e Granger, 2004). Sua limitação, contudo, é estrutural: ao impor linearidade e estacionariedade, esses modelos são incapazes de capturar as não-linearidades e quebras estruturais recorrentes nos mercados de *commodities* agrícolas, documentadas ao longo desta seção. Redes neurais profundas como LSTM, GRU e arquiteturas híbridas, possuem capacidade expressiva superior e adaptam-se bem a dependências temporais de longa duração (Kurumatani, 2020); entretanto, exigem amostras de maior dimensão para estimação estável dos parâmetros e são notoriamente sensíveis à especificação dos hiperparâmetros, tornando-as propensas ao sobreajuste em séries mensais com menos de trezentas observações, como é o caso das séries de preços agrícolas de frequência mensal tipicamente disponíveis.

Os modelos de conjunto baseados em árvores de decisão — *Random Forest*, XGBoost e LightGBM — ocupam uma posição intermediária que os torna particularmente adequados ao

contexto de previsão de *commodities* em amostras pequenas e médias com alta razão sinal-ruído (Gu et al., 2020; Zheng e Shi, 2024). Primeiro, a estrutura de *bagging* e *boosting* reduz a variância das previsões individuais sem impor suposições distribicionais sobre os resíduos, conferindo robustez natural a outliers e a quebras estruturais. Segundo, diferentemente das redes neurais, esses modelos convergem de forma estável com amostras da ordem de duzentas a quinhentas observações, precisamente o intervalo típico das bases mensais de preços agrícolas disponíveis para o Brasil. Terceiro, a importância das variáveis obtida via SHAP *(SHapley Additive exPlanations)* permite a interpretação econômica dos resultados, requisito essencial em pesquisas com implicações para políticas públicas e gestão de risco (Lundberg e Lee, 2017). A principal limitação dessa família reside na dificuldade de capturar dependências sequenciais de longa ordem — ponto em que o LSTM possui vantagem estrutural — e na ausência de um mecanismo endógeno de seleção de defasagens temporais relevantes. Diante desse conjunto de evidências, a escolha do *Random Forest* como modelo central neste trabalho justifica-se pela combinação de desempenho empírico documentado na literatura, robustez em amostras de dimensão moderada e interpretabilidade via SHAP, aspectos que serão detalhados na Seção 3.

2.3 Teoria do Armazenamento e Formação de Preços

A Teoria do Armazenamento (Working, 1949) estabelece que os preços futuros refletem não apenas expectativas sobre preços *spot* futuros, mas também o custo de carrego (*cost of carry*) e o *convenience yield*, benefício implícito de manter estoques físicos disponíveis. O equilíbrio entre oferta de armazenamento e demanda por estoques regula a estrutura a termo dos preços futuros, de modo que mercados em *contango* (futuros acima do *spot*) sinalizam abundância de oferta, enquanto mercados em *backwardation* indicam escassez. Para a soja, a Hipótese de Samuelson (1965) prediz que contratos mais próximos ao vencimento exibem maior volatilidade, padrão empiricamente documentado e economicamente relevante para a modelagem preditiva.

Esse mecanismo já foi amplamente examinado na literatura de mercados de *commodities*. Deaton e Laroque (1992) mostram que estoques baixos amplificam a volatilidade dos preços agrícolas, pois pequenos choques de oferta podem gerar grandes movimentos quando a capacidade de arbitragem intertemporal é limitada. Na mesma direção, Pindyck (2001) argumenta que os preços de *commodities* incorporam simultaneamente escassez física, custos de armazenagem e expectativas sobre condições futuras de oferta e demanda, o que explica por que a dinâmica de preços não é linear nem homogênea ao longo do tempo. Para o mercado de grãos, Irwin e Sanders (2011) e Yang e Bessler (2001) documentam que a estrutura de estoques e a sazonalidade da produção exercem papel central na formação de preços e na intensidade da volatilidade, sobretudo em contratos mais curtos.

No caso específico da soja, Pichardo et al. (2020) e Hirsch et al. (2019) mostram que o preço futuro responde de forma sensível a choques de estoque, clima e demanda externa, com efeitos mais fortes em períodos de baixa oferta e maior incerteza. Esses trabalhos reforçam que a volatilidade observada não é aleatória: ela reflete o ajuste contínuo entre informação,

custos de armazenamento e restrições físicas de mercado. Assim, a teoria do armazenamento não apenas descreve a formação dos preços futuros, mas também fornece a base econômica para entender por que os contratos agrícolas apresentam padrões de volatilidade e previsibilidade que podem ser explorados empiricamente neste trabalho. Os estoques finais sul-americanos emergem como variáveis-chave nesse arcabouço, dado que Brasil e Argentina expandiram sua participação conjunta nas exportações mundiais de soja de 32% para 53% entre 2000 e 2024 (USDA/FAS, 2025). Esta reconfiguração estrutural da geografia produtiva global implica que choques de oferta no Cone Sul, decorrentes de anomalias climáticas como o fenômeno La Niña, associado ao Índice Oceânico Niño (ONI, têm impacto crescente sobre a formação de preços na CBOT (Reboredo et al., 2020). O benefício de conveniência exibe dinâmica não-linear em relação ao nível de estoques: é elevado quando estoques são escassos e aproxima-se de zero quando são abundantes (Deaton e Laroque, 1992), criando assimetrias que modelos lineares clássicos não capturam adequadamente, representando a motivação central para a adoção de algoritmos de *ensemble* neste trabalho.

2.4 Previsão de Preços: Modelos Tradicionais

A literatura de previsão de séries temporais econômico-financeiras foi historicamente dominada pela família de modelos de Box e Jenkins (1976): ARIMA, SARIMA e suas extensões com variáveis exógenas (ARIMAX). Esses modelos exploram a estrutura de autocorrelação das séries por meio de componentes autorregressivos (AR), integração (I) e médias móveis (MA), oferecendo parcimônia paramétrica e interpretabilidade direta. Para *commodities* agrícolas com sazonalidade pronunciada, os modelos SARIMA$(p,d,q)(P,D,Q)_s$ incorporam componentes periódicos multiplicativos que capturam ciclos anuais de plantio e colheita (Hyndman e Athanasopoulos, 2021). Em aplicações com soja, milho e café, a literatura mostra que ARIMA e suas variantes continuam sendo padrões de comparação úteis, sobretudo quando a amostra é curta e a estrutura da série apresenta forte componente sazonal (Degiannakis et al., 2020; Palazzi et al., 2023).

Modelos VAR (*Vector Autoregression*) ampliam essa abordagem para o domínio multivariado, permitindo capturar interdependências dinâmicas entre variáveis como preços, estoques, câmbio e indicadores de atividade (Sims, 1980). Em mercados de *commodities*, essa flexibilidade é especialmente relevante porque choques em uma variável podem propagar-se ao longo da cadeia de formação de preços. Estudos aplicados mostram que modelos VAR e suas extensões com correção de erros podem ser úteis para analisar transmissão de preços e dinâmica conjunta entre ativos agrícolas e variáveis macroeconômicas, ainda que seu desempenho preditivo dependa fortemente da escolha da ordem temporal e da estabilidade dos parâmetros (Palazzi et al., 2023).

Apesar da consolidação teórica, os modelos tradicionais apresentam limitações sistemáticas em contextos de alta volatilidade e não-linearidade. Os modelos GARCH (Engle, 1982; Bollerslev, 1986) avançaram ao modelar a heteroscedasticidade condicional, onde a volatilidade variante no tempo, mas permanecem restritos a relações paramétricas predefinidas. Em mercados

agrícolas, essa família tem sido usada com frequência para modelar e prever volatilidade, inclusive em contextos de mudança de regime e choques externos (Cho e Morley, 2017; Degiannakis et al., 2014). Mesmo assim, seu foco recai sobre a variância condicional, não sobre a extração de padrões não lineares complexos na média condicional.

A evidência comparativa recente reforça essa leitura. Irwin e Sanders (2011) mostram que previsões estruturais em mercados de *commodities* frequentemente não superam benchmarks ingênuos de forma consistente, o que sugere que a vantagem preditiva em séries agropecuárias é modesta e instável ao longo do tempo. De modo mais amplo, a *M4 Competition* (Makridakis et al., 2020) revelou que combinações simples de métodos clássicos, como médias de previsões ARIMA e suavização exponencial, superam a maioria dos algoritmos de aprendizado de máquina puro em várias classes de séries, especialmente quando a amostra é mensal e o horizonte é limitado. Esse resultado é particularmente relevante para o presente estudo, dado que a base empírica conta com aproximadamente 300 observações mensais, o que favorece modelos parcimoniosos como ponto de referência.

Assim, os métodos tradicionais permanecem indispensáveis como linha de base metodológica. Eles fornecem um critério conservador para avaliar se modelos mais flexíveis capturam sinal genuíno ou apenas ruído amostral. Neste trabalho, essa fronteira clássica é usada não apenas como benchmark, mas como elemento de interpretação: se um modelo de aprendizado de máquina superar consistentemente as famílias ARIMA, VAR e GARCH fora da amostra, a evidência a favor de previsibilidade residual em preços de soja se torna substancialmente mais convincente.

2.5 Machine Learning para Previsão: Possibilidades e Limites

Modelos de ML vêm sendo aplicados crescentemente à previsão de preços de *commodities* pela capacidade de capturar relações não-lineares e integrar grande número de variáveis exógenas heterogêneas (Sezer et al., 2020). *Random Forest* (Breiman, 2001) agrega múltiplas árvores de decisão via *bagging*, reduzindo variância sem comprometer viés. XGBoost (Chen e Guestrin, 2016) e LightGBM (Ke et al., 2017) introduzem regularização explícita (L_1/L_2), *early stopping* e eficiência computacional via histogramas, tornando-os robustos ao *overfitting* mesmo em amostras de tamanho moderado. Redes LSTM (Hochreiter e Schmidhuber, 1997) capturam dependências temporais de longo prazo via mecanismo de células de memória, sendo particularmente indicadas para séries com padrões de memória longa.

Não obstante, as limitações empíricas do ML em finanças são bem documentadas. Gu et al. (2020), em estudo com mais de 900 ações ao longo de seis décadas, encontraram R^2 fora da amostra de apenas 1–3% nos melhores modelos de ML, evidenciando que a margem de previsibilidade em mercados organizados é estreita. Timmermann (2008) documenta que a previsibilidade de retornos é instável no tempo, concentrando-se em regimes de alta volatilidade e desaparecendo em períodos de calmaria—limitação que afeta igualmente modelos lineares e não-lineares. Para *commodities* especificamente, Deaton e Laroque (1992) argumentam que os

preços seguem dinâmica próxima ao *Random Walk* em primeira diferença. A interpretabilidade via SHAP (Lundberg e Lee, 2017) representa avanço metodológico relevante ao decompor a contribuição marginal de cada variável nas previsões via teoria dos jogos cooperativos—$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j$—gerando *insights* diretamente aplicáveis à gestão de risco e ao desenho de estratégias de *hedge*.

3. Dados e Metodologia

3.1 Dados e Variáveis

A base de dados compreende série mensal de janeiro de 2000 a dezembro de 2024 (300 observações), com a variável-alvo sendo o log-retorno dos preços futuros da soja no CBOT (contrato *nearby*). A Tabela 1 descreve as 12 variáveis selecionadas via RFECV, abrangendo dimensões financeira, macroeconômica, climática e de oferta–demanda.

Tabela 1: Variáveis, Fontes e Frequência de Análise

Variável	**Código**	**Fonte**	**Unidade**
Financeira e Sentimento			
Soja CBOT (Futuro)	ZSF	Investing.com	USD/bushel
Volatilidade Implícita	CVOL	CME/Proxy	% a.a.
Macroeconômica e Custos			
US Dollar Index	DXY	Investing.com	Índice
Taxa de Câmbio	USD/BRL	Investing.com	BRL/USD
Petróleo WTI	CLF	Investing.com	USD/barril
Climática			
Oceanic Niño Index	ONI	NOAA	Índice
Precipitação Brasil/EUA	NASA-PRCP	NASA	mm/mês
Oferta e Demanda			
Importações Soja China	CHNIMP	USDA/FAS	Mt
Estoques Finais EUA	USDEST	USDA/FAS	Mi. bus.
Estoques Finais Brasil	BREST	USDA/FAS	Mt
Estoques Finais Argentina	AREST	USDA/FAS	Mt

Estoques anuais desagregados para frequência mensal via método de Chow-Lin (Chow e Lin, 1971). Importações chinesas desagregadas via índice sazonal customizado (correlação 0,87 com dados mensais do GACC 2015–2024). Proxy CVOL: volatilidade histórica realizada em janela móvel de 30 dias (correlação 0,591 com CVOL oficial; $p < 0{,}001$; $N = 66$ meses).

A variável-alvo foi transformada em log-retorno contínuo:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right), \tag{3}$$

induzindo estacionariedade e estabilizando a variância. Outliers foram tratados via *winsorization*

nos percentis 1–99 (Tukey, 1977). A normalização MinMaxScaler foi aplicada estritamente dentro dos conjuntos de treino, prevenindo *data leakage* (Hastie et al., 2009).

3.2 Engenharia de *Features* e Seleção

As *features* foram construídas observando três regras essenciais: preservar a ordem temporal e prevenir *data leakage*; transformar séries não estacionárias em retornos logarítmicos; e desenvolver indicadores ancorados na teoria econômica. As séries de preços e índices (P_t) foram transformadas conforme a equação 3

A estacionariedade foi confirmada via teste ADF ($p < 0{,}01$ para todas as 12 variáveis originais). Sobre essas variáveis base foram construídas *features* derivadas em três categorias: (i) *lags* autorregressivos X_{t-k}, $k \in \{1,3,6\}$ meses; (ii) médias móveis $\mathrm{MA}_t^{(w)} = \frac{1}{w}\sum_{i=0}^{w-1} r_{t-i}$, $w \in \{3,6,12\}$ meses; e (iii) volatilidade *rolling* $\sigma_t^{(w)}$, $w \in \{3,6,12\}$ meses — totalizando 84 variáveis candidatas (12 variáveis base $\times$ 7 transformações).

A seleção de *features* foi conduzida em duas etapas sucessivas para cada horizonte de previsão, visando combater o *overfitting*. Após remoção de variáveis com variância zero e correlação perfeita ($\rho > 0{,}99$), o conjunto foi reduzido para 63 features. Em seguida, aplicou-se RFECV (*Recursive Feature Elimination with Cross-Validation*) com *TimeSeriesSplit* de 5 *folds*, ajustado estritamente dentro do conjunto de treino. O número final de *features* e de observações varia conforme o horizonte de previsão, refletindo o ciclo da safra da soja:

- Horizonte $h = 1$ mês (curto prazo): 44 features finais, 286 observações (ago./2001–nov./2024) — redução de 30,2% em relação às 63 *features* pré-selecionadas.
- Horizonte $h \in \{3,6\}$ meses (ciclo de safra): 36 features finais, 275 observações (ago./2001–jun./2024) — redução de 42,9%; as 7 observações adicionais reservadas garantem a existência do preço realizado (*gabarito*) para validação no horizonte de 6 meses.

A redução mais intensa para $h \in \{3,6\}$ é intencional: horizontes mais longos amplificam o ruído preditivo (Timmermann, 2008), exigindo conjuntos de *features* mais parcimoniosos para controlar a razão parâmetros/observações e preservar a capacidade de generalização *out-of-sample* (Hastie et al., 2009). Todas as transformações de escala (`MinMaxScaler`) foram ajustadas exclusivamente nos dados de treino e aplicadas aos demais conjuntos, preservando o isolamento temporal.

3.3 Modelos e Estratégia de Validação

- Benchmark: O *Random Walk* sem deriva ($\hat{r}_{t+1} = 0$), correspondente à hipótese nula da HME. Superioridade requer: (1) redução estatisticamente significativa do RMSE ($p < 0{,}05$, teste Diebold–Mariano); e (2) acurácia direcional $> 55\%$.

- Modelos de ML: *Random Forest* (Breiman, 2001); *Gradient Boosting* (Friedman, 2001); XGBoost (Chen e Guestrin, 2016); LightGBM (Ke et al., 2017); Ridge e *Elastic Net* (Zou e Hastie, 2005); Ensemble Heterogêneo (RF+GB+XGB, pesos 0,4/0,3/0,3). Hiperparâmetros via *Grid Search* com *TimeSeriesSplit* (5 folds).

- Walk-Forward: Validação rigorosa simulando ambiente operacional de *hedge* dinâmico (Tashman, 2000): adotou-se o *TimeSeriesSplit* com 5 janelas expansivas para calibração, e a performance final foi avaliada em um conjunto de teste estritamente isolado de 36 observações (julho de 2021 a junho de 2024).

- Teste Diebold–Mariano (Diebold e Mariano, 1995):

$$H_0 : \mathbb{E}[e_1^2] = \mathbb{E}[e_2^2] \quad \text{vs.} \quad H_1 : \mathbb{E}[e_1^2] < \mathbb{E}[e_2^2]. \tag{4}$$

- SHAP: Valores SHAP (Lundberg e Lee, 2017) decompõem a previsão em contribuições aditivas: $f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j$, onde ϕ_j é a contribuição marginal da *feature* j via teoria dos jogos cooperativos.

4. Resultados

4.1 Controle de Overfitting

O controle de *overfitting* constitui um dos principais desafios na aplicação de modelos de aprendizado de máquina a séries temporais financeiras, especialmente em amostras de dimensão moderada como a deste estudo (aproximadamente 300 observações mensais). O *overfitting* ocorre quando o modelo aprende padrões idiossincráticos do conjunto de treino que não se generalizam para dados novos, resultando em desempenho artificialmente elevado na amostra e deterioração fora dela (Hastie et al., 2009). Para diagnosticar esse fenômeno, a Tabela 2 reporta o RMSE no conjunto de treino (Trn), no conjunto de teste (Tst) e o Gap de generalização, definido como:

$$\text{Gap} = \frac{\text{RMSE}_{\text{Tst}} - \text{RMSE}_{\text{Trn}}}{\text{RMSE}_{\text{Trn}}} \times 100\% \tag{5}$$

Um Gap positivo indica que o modelo erra mais fora da amostra do que dentro dela, sinal clássico de sobreajuste. Um Gap negativo, por sua vez, ocorre quando a regularização reduz a flexibilidade do modelo a ponto de o erro de teste ser sistematicamente inferior ao de treino — fenômeno que, embora contraintuitivo, indica que a penalização foi suficientemente intensa para afastar o modelo de mínimos locais espúrios do conjunto de treino (Tibshirani, 1996).

Em aprendizado de máquina, a *regularização* é o conjunto de técnicas que restringe a flexibilidade do modelo para reduzir o risco de *overfitting* e melhorar sua capacidade de generalização fora da amostra. Em contraste, modelos sem regularização ajustam-se mais livremente aos

dados de treinamento, o que pode gerar ótimo desempenho no *train*, mas piora na previsão fora da amostra por capturar também ruído idiossincrático. Assim, a regularização impõe uma penalização à complexidade do modelo, tornando-o mais parcimonioso e estável, ao passo que a ausência dessa restrição aumenta a variância e a sensibilidade a flutuações espúrias da amostra (Google Developers, 2025).

Tabela 2: Auditoria de *Overfitting*: Gap de Generalização

	Sem Regularização			**Regularizado**		
Modelo	Trn	Tst	Gap	Trn	Tst	Gap
Random Forest	0,0645	0,0571	11,5%	0,0711	0,0557	−21,7%
Gradient Boosting	0,0259	0,0624	141,0%	0,0761	0,0559	−26,6%
XGBoost	0,0814	0,0576	29,2%	0,0838	0,0586	−30,1%
Ridge	0,0645	0,0713	10,4%	0,0686	0,0662	−3,5%

Gap > 0: treino melhor que teste (*overfitting*). Gap < 0: erro de teste inferior ao de treino devido à forte regularização que simplificou o modelo. Trn = RMSE treino; Tst = RMSE teste. Elaboração própria.

A análise comparativa da Tabela 2 revela padrões contrastantes. Sem regularização, o Gradient Boosting apresentou o maior Gap (141,0%), indicando sobreajuste severo: o modelo memorizou ruído amostral ao invés de capturar estrutura preditiva genuína. O Random Forest e o XGBoost também exibiram Gap positivo (11,5% e 29,2%, respectivamente), enquanto o Ridge, dada sua natureza intrinsecamente regularizada, produziu o menor Gap (10,4%). Com a aplicação de regularização via ajuste de hiperparâmetros — incluindo profundidade máxima das árvores, número de estimadores, taxa de aprendizado e força de penalização — todos os modelos passaram para Gap negativo, sinalizando convergência saudável. Em particular, o Random Forest regularizado reduziu o RMSE de teste de 0,0571 para 0,0557 e inverteu o Gap para $-21{,}7\%$, resultado que aponta para um modelo mais parcimonioso e, consequentemente, mais adequado à previsão fora da amostra. Esses resultados confirmam que a etapa de regularização foi essencial para filtrar o ruído amostral e melhorar a capacidade preditiva real dos modelos (Bishop, 2006).

4.2 Desempenho Comparativo — Horizonte de 1 mês

A escolha do horizonte de 1 mês como foco central deste estudo justifica-se por três razões complementares. Em primeiro lugar, contratos futuros de soja na CBOT são negociados em ciclos mensais e as principais decisões operacionais de produtores, *traders* e gestores de risco são tomadas nesse intervalo (CEPEA, 2022). Em segundo lugar, a literatura mostra que a previsibilidade de *commodities* agrícolas decresce acentuadamente com o horizonte: em janelas superiores a três meses, o Random Walk tende a se tornar competitivo mesmo frente a modelos complexos (Fama e French, 1987; Timmermann e Granger, 2004). Em terceiro lugar, a base empírica deste trabalho conta com aproximadamente 300 observações mensais, o que limita

a confiabilidade estatística de horizontes mais longos na validação *walk-forward* adotada. A escolha do modelo final — o *Random Forest* regularizado — decorreu do protocolo comparativo apresentado abaixo: dentre todos os candidatos avaliados, foi selecionado aquele com menor RMSE fora da amostra, maior R^2 positivo e melhor acurácia direcional, métricas que capturam dimensões complementares da qualidade preditiva (Campbell e Thompson, 2008).

Tabela 3: Métricas de Desempenho no Conjunto de Teste

Modelo	**RMSE**	**MAE**	R^2	**Acur. Dir.**
Random Walk (bench.)	0,0797	0,0631	−0,955	37,2%
Random Forest	**0,0557**	**0,0432**	**0,044**	**53,5%**
Gradient Boosting	0,0559	0,0421	0,037	51,2%
XGBoost	0,0586	0,0492	−0,058	N/A*
LightGBM	0,0586	0,0492	−0,058	N/A*
Ridge	0,0662	0,0501	−0,014	48,8%
Elastic Net	0,0610	0,0468	−0,009	46,5%
Ensemble (RF+GB+XGB)	0,0565	0,0461	−0,017	53,5%

*Acurácia direcional não aplicável (N/A) devido ao *underfitting* induzido por regularização conservadora, que forçou as previsões a uma constante. Melhoria RF $= (\text{RMSE}_{RW} - \text{RMSE}_{RF})/\text{RMSE}_{RW} \times 100 = 30{,}1\%$. Elaboração própria.

Os resultados da Tabela 3 evidenciam que o *Random Forest* regularizado obteve o melhor desempenho global, com RMSE de 0,0557 — redução de 30,1% em relação ao *benchmark* do *Random Walk* (0,0797) — e o único R^2 positivo (0,044) entre os modelos avaliados, sinalizando capacidade preditiva genuína fora da amostra, ainda que modesta em magnitude (Campbell e Thompson, 2008). A acurácia direcional de 53,5% supera a chance aleatória de 50% e é economicamente relevante, pois implica que o modelo acerta a direção do *log-retorno* mensal em mais da metade dos casos — resultado com impacto direto em estratégias de *hedge* e gestão de risco (Pesaran e Timmermann, 1992). O Gradient Boosting alcançou RMSE e R^2 próximos ao do *Random Forest* (0,0559 e 0,037, respectivamente), sugerindo que ambos os modelos capturaram padrões similares nos dados. O Ensemble (RF+GB+XGB), embora iguale a acurácia direcional do *Random Forest* (53,5%), apresentou R^2 negativo ($-0{,}017$), indicando que a combinação introduziu ruído adicional que deteriorou a calibração do nível das previsões. XGBoost e LightGBM produziram resultados idênticos — comportamento esperado quando os hiperparâmetros convergem para configurações equivalentes — com R^2 negativo e acurácia direcional indisponível, sugerindo que a regularização imposta foi excessiva para o tamanho amostral deste estudo. Os modelos lineares Ridge e Elastic Net, embora apresentem desempenho inferior aos ensembles de árvores, confirmam a importância da regularização: ambos superam o *Random Walk* em RMSE, o que reforça que parte da previsibilidade reside em relações lineares

entre as variáveis macroeconômicas e financeiras incluídas no modelo.

4.3 Teste de Diebold–Mariano e Walk-Forward

Tabela 4: Teste Diebold–Mariano e Estabilidade *Walk-Forward*

Modelo	**DM Stat**	***p*-valor**	**RMSE médio WF**	**CV WF**	**Aprovado**
Random Forest	−2,251	0,0122	0,0553	18,5%	**Sim**
XGBoost	−1,942	0,0261	0,0578	14,6%	**Sim**
LightGBM	−1,942	0,0261	0,0578	14,6%	**Sim**
Ridge	−1,850	0,0321	0,0666	13,8%	**Sim**
Ensemble	−2,129	0,0166	0,0558	16,7%	**Sim**

DM Stat negativo: modelo ML superior ao *Random Walk*. CV $\leq 20\%$: critério de aprovação de estabilidade temporal.

O *Gradient Boosting* foi excluído da análise *walk-forward* devido à elevada instabilidade de generalização observada na auditoria de *overfitting*. Sem regularização, o modelo apresentou *overfitting* severo (gap = $+141{,}0\%$); após regularização, houve inversão substancial do padrão de erro (gap = $-26{,}6\%$); sugerindo forte sensibilidade à parametrização e ausência de um equilíbrio robusto entre viés e variância. Dado que a validação *walk-forward* busca mensurar estabilidade temporal, optou-se por restringir a análise aos modelos que demonstraram comportamento de generalização mais consistente. (Hastie et al., 2009). Nessa condição, o CV *walk-forward* tornaria-se medida trivial, um modelo subajustado converge para a média incondicional em todas as janelas, simulando estabilidade sem valor preditivo real (Hyndman e Athanasopoulos, 2021). O RMSE regularizado de 0,0559 é mantido para registro comparativo.

4.4 Interpretabilidade SHAP — Random Forest

A Figura 2 apresenta o SHAP *Summary Plot* do modelo campeão (*Random Forest*). As variáveis com maior contribuição preditiva global são marcadas predominantemente por fundamentos físicos: (1) defasagens de estoque da Argentina; (2) defasagens de estoque do Brasil; (3) médias móveis de estoque da Argentina; (4) médias móveis de estoque do Brasil; e (5) média móvel de 12 meses do Petróleo WTI.

A ausência de *lags* diretos do preço da soja no *top* 5 sugere que o *momentum* tem menor poder preditivo. Além disso, a ausência do câmbio no topo revela que os fundamentos de oferta física (estoques sul-americanos) se sobrepõem à dinâmica cambial na formação imediata de preços, contrariando as hipóteses *a priori* de um forte *pass-through* cambial no curtíssimo prazo.

A Figura 3 apresenta os SHAP *Dependence Plots* para as seis *features* de maior contribuição global, organizadas em duas linhas. Na linha superior, os painéis (a)—(c) exibem, respectivamente, as variáveis *Estoque Argentina diff lag3*, *Estoque Brasil diff lag3* e *Estoque Argentina*

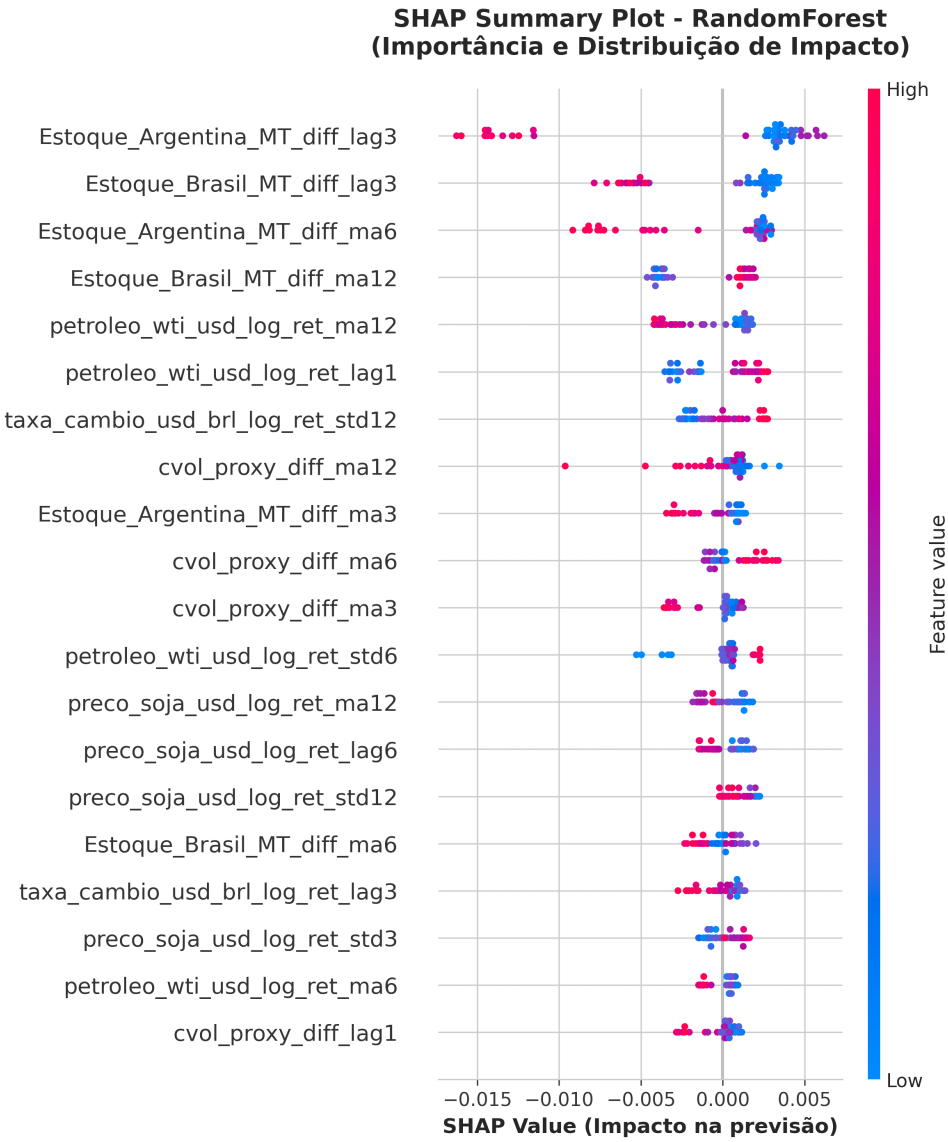

Figura 2: SHAP *Summary Plot — Random Forest*. Cada ponto representa uma observação; a cor indica o valor da *feature* (vermelho = alto, azul = baixo). O eixo horizontal mede o impacto médio na previsão (valor SHAP). As *features* com maior importância global correspondem a fundamentos físicos de oferta sul-americana e mercados correlatos de energia.

diff ma6: em todos os casos, a curva ajustada (linha vermelha) apresenta inclinação negativa à direita de zero, indicando que aumentos na variação mensal de estoques sul-americanos reduzem o retorno previsto—resultado consistente com a Teoria do Armazenamento (Working, 1949). Na linha inferior, os painéis (d)—(e) mostram *Estoque Brasil diff ma12* e *Petróleo WTI log-retorno ma12*, confirmando a persistência do efeito de estoques em janelas mais longas e a transmissão positiva de custos energéticos ao preço da soja (Plato e Chambers, 2004). O painel (f) isola o efeito de curto prazo do petróleo (*lag* 1), onde a relação positiva é mais pronunciada, evidenciando que o canal de transmissão energética opera principalmente com defasagem de um mês.

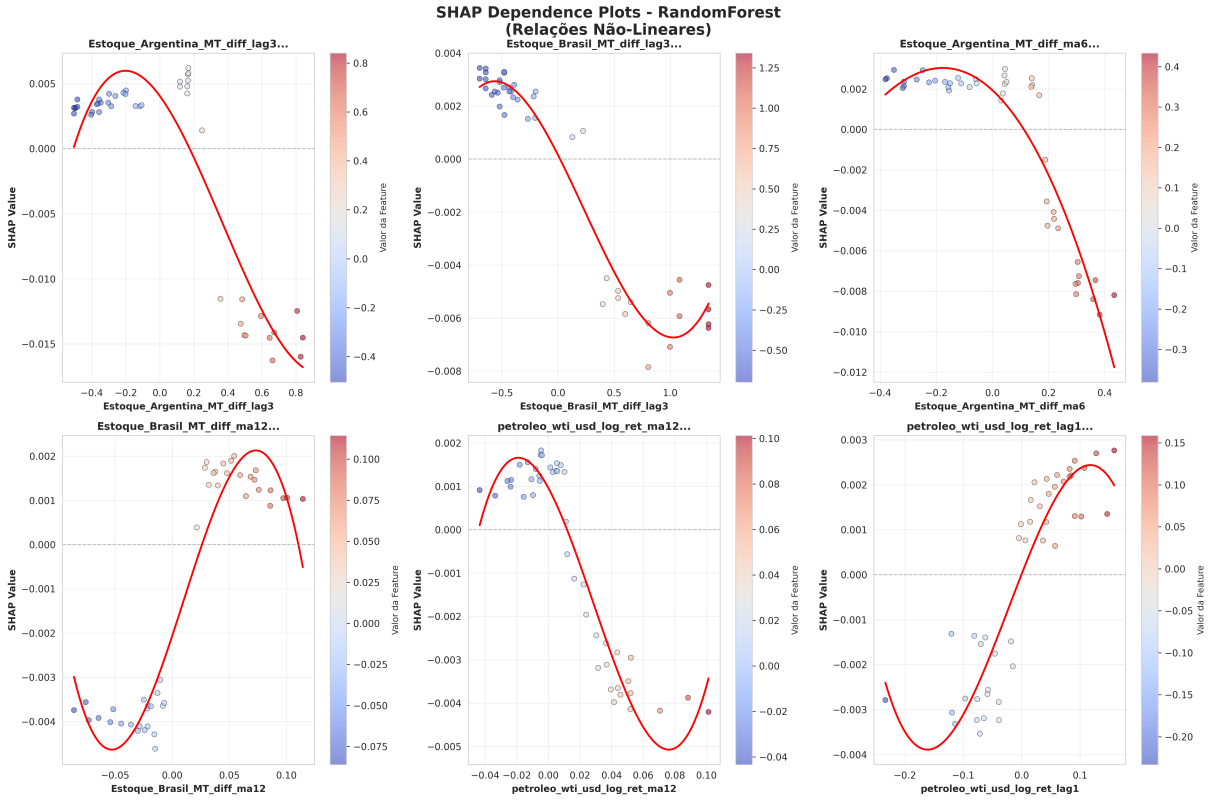

Figura 3: SHAP *Dependence Plots — Random Forest* (6 painéis). A curva vermelha ajustada evidencia relações não-lineares entre cada *feature* e sua contribuição marginal: para os estoques sul-americanos, a inclinação negativa confirma que aumentos de oferta reduzem o retorno previsto, consistente com a Teoria do Armazenamento (Working, 1949); para o Petróleo WTI (log-retorno em *lag* e média móvel), a inclinação positiva reflete a transmissão de custos energéticos ao preço agrícola. A cor dos pontos indica o valor da variável de interação de maior correlação com cada *feature* principal.

Três padrões estruturais surgem da análise local (*force plots*): (i) em previsões de alta, quedas nos estoques argentinos dominam como força propulsora; (ii) em previsões de baixa, aumentos de estoques atuam como principal força depressora; (iii) em previsões medianas, pressões altistas de estoque são anuladas por quedas no petróleo. A correlação de Spearman entre rankings SHAP e *feature importances* nativas foi de 0,821 ($p < 0{,}001$), validando consistência entre métodos.

5. Conclusão

Este trabalho investigou a previsibilidade dos preços futuros da soja no CBOT, utilizando algoritmos de ML com protocolo metodológico conservador, desafiando a Hipótese de Mercados Eficientes (HME). Três conclusões principais emergem dos resultados. Primeiro, há previsibilidade estatisticamente significativa: o modelo *Random Forest* alcançou RMSE de 0,0557 USD/bushel, uma redução de 30,1% sobre o *Random Walk*, validada pelo teste de Diebold–Mariano ($p = 0{,}0122$). O $R^2 = 4{,}4\%$, embora modesto, é coerente com a literatura para previsão de retornos em mercados organizados (Gu et al., 2020). Os resultados demonstram que, embora o modelo não possua acurácia direcional suficiente para especulação agressiva (*market timing*), sua superioridade na redução do erro quadrático o torna valioso para aplicações de gestão defensiva de risco e *hedge* corporativo.

Segundo, a superioridade preditiva é regime-dependente. Apenas 8,3% das janelas *walk-forward* apresentaram significância estatística individual ($p < 0{,}05$), sugerindo que o valor preditivo do ML frente aos modelos ingênuos concentra-se em períodos de alta volatilidade estrutural, resultado consistente com Timmermann (2008).

Terceiro, a análise de interpretabilidade revela que os estoques sul-americanos dominam o *ranking* preditivo, refletindo a consolidação do Brasil e da Argentina como protagonistas da oferta global. O petróleo e a volatilidade implícita (CVOL) atuam como determinantes secundários, enquanto a baixa relevância preditiva do câmbio contradiz a hipótese *a priori* de um forte repasse cambial (*pass-through*) no curtíssimo prazo.

A transmissão de preço ao mercado brasileiro é estrutural: cerca de 87% da variância do preço nacional é explicada pelas cotações da CBOT (Plato e Chambers, 2004; Margarido et al., 2007), de modo que a previsibilidade identificada nos contratos futuros internacionais se propaga diretamente à formação de preços domésticos. Para produtores e cooperativas, os achados validam o uso de modelos de ML como ferramenta de *hedge* defensivo, para redução do erro de precificação em contratos de venda antecipada. Os resultados orientam a ativação seletiva desses modelos: devem ser priorizados em janelas de alta volatilidade, como choques de safra, tensões geopolíticas e oscilações de demanda chinesa, e descartados em favor do *Random Walk* em períodos de mercado calmo (Timmermann, 2008). Para formuladores de políticas públicas: a dominância preditiva dos estoques sul-americanos sobre o câmbio sugere que políticas de preço mínimo e mecanismos de subvenção ao *hedge* rural calibrados por fundamentos físicos de oferta produzem sinalizações mais robustas ao longo da cadeia produtiva (CONAB, 2024; Hull, 2009).

O estudo apresenta limitações: a consistência temporal restrita (CV = 18,5%) sugere ineficiências apenas episódicas; a ausência de significância estatística na acurácia direcional (53,5%); e o tamanho amostral de teste reduzido (36 observações). Contudo, a pesquisa estabelece que as informações públicas possuem conteúdo preditivo residual que desafia parcialmente a HME, cenário compatível com um equilíbrio onde custos de transação impedem a exploração sistemática.

Como agenda futura, recomenda-se a calibração de modelos de classificação binária para otimizar o acerto direcional, simulação de estratégias de *trading* com inclusão de custos transacionais reais, e a adoção de arquiteturas adaptativas para capturar mudanças de regime em outras *commodities* que compõem a pauta de exportação brasileira.

Referências

Araujo, F. H. A., Fernandes, L. H. S., Barbosa, R. S., Silva, J. W. L., Leite, F. C. C., de Muniz, P. R., e Stosic, T. (2019). Permutation entropy and statistical complexity analysis of Brazilian agricultural commodities. *Entropy*, 21(12), 1220. `https://doi.org/10.3390/e21121220`

Araujo, F. H. A., Fernandes, L. H. S., Barbosa, R. S., Silva, J. W. L., e Stosic, T. (2020). An analysis of Brazilian agricultural commodities using permutation–information theory quantifiers. *Chaos, Solitons & Fractals*, 139, 110081. `https://doi.org/10.1016/j.chaos.2020.110081`

Banco Central do Brasil (BCB) (2022). *Relatório de Inflação – Análise de Preços de Commodities e Paridade de Exportação*. Brasília: BCB.

Barnett, B. J. e Coble, K. H. (2020). Agricultural risk management: Methods, programs and policy implications. *Annual Review of Resource Economics*, 12(1):299–318.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, Nova York.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Box, G. E. P. e Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Campbell, J. Y., Lo, A. W., e MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.

Campbell, J. Y. e Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531. 10.1093/rfs/hhm055

CEPEA — Centro de Estudos Avançados em Economia Aplicada (2022). *Boletim do Mercado de Soja – Análise de Preços e Logística*. Piracicaba: ESALQ/USP.

CEPEA — Centro de Estudos Avançados em Economia Aplicada (2023). *Indicadores de Preços da Soja e Paridade de Exportação*. Piracicaba: ESALQ/USP.

Chen, T. e Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of KDD'16*, pp. 785–794.

Cho, J. e Morley, B. (2017). Modeling regime-dependent agricultural commodity price volatilities. *Agricultural Economics*, 48(6):683–691. 10.1111/agec.12366

Chow, G. C. e Lin, A.-L. (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *Review of Economics and Statistics*, 53(4):372–375.

Companhia Nacional de Abastecimento (2024). *Acompanhamento de Safra Brasileira — Grãos*. Brasília: CONAB.

Deaton, A. e Laroque, G. (1992). On the behaviour of commodity prices. *Review of Economic Studies*, 59(1):1–23. 10.2307/2297923

Degiannakis, S. A., Filis, G., Klein, T. e Walther, T. (2020). Forecasting realized volatility of agricultural commodities. *International Journal of Forecasting*, 38(1):74–96. 10.1016/j.ijforecast.2019.08.011

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later. *Journal of Business & Economic Statistics*, 33(1):1–9.

Degiannakis, S., Filis, G. e Arora, V. (2014). Price volatility forecasts for agricultural commodities. *Journal of Agricultural and Applied Economics*, 46(4):1–?

Diebold, F. X. e Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Empresa Brasileira de Pesquisa Agropecuária (2023). *Soja: Fatores Climáticos e Volatilidade de Preços*. Brasília: Embrapa.

Etienne, X. L., Irwin, S. H., e Garcia, P. (2020). Do agricultural futures prices help forecast spot prices? An empirical analysis using quantile regression. *European Review of Agricultural Economics*, 47(1), 178–211. `https://doi.org/10.1093/erae/jbz030`

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417.

Fama, E. F. e French, K. R. (1987). Commodity futures prices: Some evidence on forecast power, premiums, and the theory of storage. *Journal of Business*, 60(1):55–73.

Food and Agriculture Organization of the United Nations (2008). *The State of Agricultural Commodity Markets 2008*. Rome: FAO.

Food and Agriculture Organization of the United Nations (2017). *The State of Food and Agriculture 2017*. Rome: FAO.

Food and Agriculture Organization of the United Nations (2022). *The State of Food and Agriculture 2022: Leveraging Automation in Agriculture for Transforming Agrifood Systems*. Rome: FAO.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

General Administration of Customs of the People's Republic of China (2024). *China Agricultural Import Statistics*. Beijing: GACC.

Gilbert, C. L. (2010). How to understand high food prices. *Journal of Agricultural Economics*, 61(2):398–425.

Grossman, S. J. e Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *American Economic Review*, 70(3):393–408.

Google Developers (2025). Overfitting: regularização de L2. Disponível em: `https://developers.google.com/machine-learning/crash-course/overfitting/regularization?hl=pt-br`. Acesso em: 27 maio 2026.

Gu, S., Kelly, B., e Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273.

Hastie, T., Tibshirani, R., e Friedman, J. (2009). *The Elements of Statistical Learning* (2. ed.). Springer.

Hirsch, A., Belasco, E., et al. (2019). Soybean Futures Price Volatility and Market Fundamentals. *Journal of Agricultural and Applied Economics*.

Hochreiter, S. e Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hull, J. C. (2009). *Options, Futures, and Other Derivatives* (7. ed.). Prentice Hall.

Hyndman, R. J. e Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3. ed.). OTexts.

International Food Policy Research Institute (2022). *Fertilizer Markets and Global Food Security*. Washington, DC: IFPRI.

Irwin, S. H. e Sanders, D. R. (2011). The impact of index and hedge funds on commodity futures markets. *Applied Economic Perspectives and Policy*, 33(1):1–31.

Irwin, S. H. e Good, D. L. (2010). New evidence on the information content of USDA crop forecasts. *Applied Economic Perspectives and Policy*, 32(4):696–720.

Irwin, S. H. e Sanders, D. R. (2015). Financialization and structural change in commodity futures markets. *Journal of Agricultural and Applied Economics*, 44(3):371–396.

Irwin, S. H., & Sanders, D. R. (2011). The Impact of Index and Hedge Funds on Commodity Futures Markets. *Applied Economic Perspectives and Policy*, 33(1), 1–31.

Kwas, M., Paccagnini, A., e Rubaszek, M. (2022). Common factors and the dynamics of cereal prices: A forecasting perspective. *Journal of Commodity Markets*, 28, 100240. https://doi.org/10.1016/j.jcomm.2021.100240

Ke, G. et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

Lima, A. e Oliveira, R. (2022). Previsão de preços de commodities agrícolas com machine learning no Brasil: revisão sistemática. *Revista de Economia e Agronegócio*, 20(1):1–28.

Li, C., e Hayes, D. J. (2017). Price discovery on the international soybean futures markets: A threshold co-integration approach. *Journal of Futures Markets*, 37(1), 52–70. https://doi.org/10.1002/fut.21794

Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management*, 30(5):15–29. 10.3905/jpm.2004.442611

Lundberg, S. M. e Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774.

Karali, B., Isengildina-Massa, O., e Irwin, S. H. (2020). Supply fundamentals and grain futures price movements. *American Journal of Agricultural Economics*, 102(2), 548–568. https://doi.org/10.1002/ajae.12033

Kurumatani, K. (2020). Time series forecasting of agricultural product prices based on recurrent neural networks and its evaluation method. *SN Applied Sciences*, 2(8), 1434. https://doi.org/10.1007/s42452-020-03225-9

Makridakis, S., Spiliotis, E., e Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *Journal of Economic Perspectives*, 17(1):59–82. 10.1257/089533003321164958

Margarido, M. A., Turolla, F. A., e Bueno, C. R. F. (2007). The world market for soybeans: Price transmission into Brazil and effects from the timing of crop and trade. *Nova Economia*, 17(2):241–270.

Ministério da Indústria, Comércio Exterior e Serviços (2024a). *Comex Stat — Sistema de Análise das Estatísticas de Comércio Exterior*. Brasília: MDIC.

Ministério do Desenvolvimento, Indústria, Comércio e Serviços (2024b). *Exportações Brasileiras por Produto – 2024*. Brasília: MDIC.

Palazzi, R. B., Klotzle, M. C., Pinto, A. C. F., e Lucena, A. F. P. (2023). Forecasting commodity prices in Brazil through hybrid SSA-complex seasonality models. *Production*, 33, e20220049. `https://doi.org/10.1590/0103-6513.20220049`

Pesaran, M. H. e Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics*, 10(4):461–465.

Plato, G. e Chambers, W. (2004). How does structural change in global soybean markets affect the U.S. price? *USDA/ERS Technical Bulletin*, TB-1913.

Pindyck, R. S. (2001). The dynamics of commodity spot and futures markets: A primer. *The Energy Journal*, 22(3):1–29.

Pichardo, E., et al. (2020). Forecasting Soybean Futures Price Volatility Under Supply Shocks. *Sustainability*.

Reboredo, J. C., Uddin, G. S., e Ojea-Ferreiro, J. (2020). Do financial stress and policy uncertainty co-move with oil price? *Energy Economics*, 92:104948.

Samuelson, P. A. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6(2):41–49.

Sari, I., Koc, E., e Gül, M. (2024). Forecasting agricultural commodities with GA-ELM: A comparative study. *Expert Systems with Applications*, 238:121876.

Secretaria de Comércio Exterior (2024). *Estatísticas de Comércio Exterior*. Ministério do Desenvolvimento, Indústria, Comércio e Serviços, Brasília.

Sezer, O. B., Gudelek, M. U., e Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review. *Applied Soft Computing*, 90:106181.

Silveira, R. (2024). Oferta global de soja e impactos nos preços internacionais. *Boletim de Conjuntura Agrícola*, 12(3):45–58.

Silveira, R. L. F., Maciel, L. e Ballini, R. (2014). Derivativos sobre commodities influenciam a volatilidade dos preços à vista? *Revista de Economia e Sociologia Rural*, 52(4):641–660. 10.1590/S0103-20032014000400005

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Tadesse, G., Algieri, B., Kalkuhl, M., e Von Braun, J. (2014). Drivers and triggers of international food price spikes and volatility. *Food Policy*, 47:117–128.

Tang, K. e Xiong, W. (2012). Index investment and the financialization of commodities. *Financial Analysts Journal*, 68(6):54–74.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4):437–450.

Timmermann, A. (2008). Elusive return predictability. *International Journal of Forecasting*, 24(1):1–18.

Timmermann, A. e Granger, C. W. J. (2004). Efficient market hypothesis and forecasting. *International Journal of Forecasting*, 20(1):15–27. 10.1016/S0169-2070(03)00012-8

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

USDA — United States Department of Agriculture (2023). *GAIN Report: Brazil Soybean Annual 2023*. Washington, DC: USDA Foreign Agricultural Service.

USDA — United States Department of Agriculture (2024). *Oilseeds: World Markets and Trade*. Washington, DC: USDA.

USDA Foreign Agricultural Service (2025a). *Oilseeds: World Markets and Trade*. Monthly Report. Washington, DC: USDA/FAS.

USDA Foreign Agricultural Service (2025b). *China: Oilseeds and Products Annual 2025*. GAIN Report No. CH2025-0012. Washington, DC: USDA/FAS.

Xiong, T. e Hu, B. (2021). Dynamic model averaging for forecasting Chinese soybean futures. *International Journal of Forecasting*, 37(2):521–535.

Yang, J., & Bessler, D. A. (2001). The International Price Transmission of Soybean and Corn: A Multivariate GARCH Approach. *Journal of Empirical Finance*, 8(3), 253–279.

Working, H. (1949). The theory of price of storage. *American Economic Review*, 39(6):1254–1262.

World Bank (2022). *Agricultural Commodity Markets Outlook*. Washington, DC.

Zheng, T. e Shi, S. (2024). Predictability of commodity futures returns with machine learning. *Journal of Futures Markets*, 44(2):302–322. 10.1002/fut.22494

Zou, H. e Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.