

Previsão de preços de Commodities: Um *case* real de Machine Learning aplicado ao mercado de farelo de soja na região de Campinas

Sérgio Domingos Simão
Pós-doutorando do Programa em
Estatística e Experimentação
Agropecuária, DES/ICET,
Universidade Federal de Lavras e
Professor na Escola Estadual João
Melo Gomide – Perdões/MG
MG, Brasil
sergiodomingos@ufla.br

João Paulo Assis Bonifácio
Graduando em Estatística
Universidade Federal de Lavras
MG, Brasil
joao.bonifacio2@estudante.ufla.br

Thelma Sáfadi
Departamento de Estatística –
DES/ICET
Universidade Federal de Lavras
MG, Brasil
safadi@ufla.br

Geraldo Magela da Cruz Pereira
Departamento de Estatística –
DES/ICET
Universidade Federal de Lavras
MG, Brasil
geraldo.pereira@ufla.br

Izabela Regina Cardoso de Oliveira
Departamento de Estatística –
DES/ICET
Universidade Federal de Lavras
MG, Brasil
izabela.oliveira@ufla.br

Resumo— Este estudo tem como objetivo aplicar e avaliar modelos de Machine Learning (ML) para compreender a dinâmica dos preços diários do farelo de soja na região de Campinas, São Paulo, com foco na previsão do preço por tonelada. A previsão de preços desse insumo é fundamental para a cadeia produtiva da pecuária, uma vez que o farelo de soja é amplamente utilizado na formulação de rações e seus preços impactam diretamente os custos de produção animal. Os dados utilizados foram obtidos por meio da plataforma Broadcast, cobrindo o período de janeiro de 2007 a dezembro de 2024, totalizando 4.398 observações. A análise preditiva foi conduzida com o uso do software R, empregando as bibliotecas *tidyverse* e *tidymodels*, que possibilitam um fluxo estruturado e reproduzível para manipulação de dados, engenharia de variáveis, treinamento e validação de modelos. Como variáveis explicativas, foram utilizados o tempo (como índice) e um parâmetro de defasagem baseado em componentes autorregressivos de primeira ordem (*AR-1*), visando capturar a dependência temporal presente na série. Foram ajustados e comparados diferentes algoritmos de ML, com destaque para os modelos de regressão por máquinas de vetores de suporte (SVM), redes neurais e árvores de decisão (*Random Forest*). O desempenho dos modelos foi avaliado com base em métricas de erro, como o Erro Médio Absoluto (MAE), a Raiz do Erro Quadrático Médio (RMSE), o Erro Percentual Absoluto Médio (MAPE) e o Coeficiente de Determinação (R^2). Entre os modelos testados, o SVM com kernel linear apresentou o melhor desempenho, destacando-se pela capacidade de generalização e pela baixa magnitude dos erros preditivos e baixo tempo computacional (25,4 segundos). Os resultados demonstram o potencial do uso de técnicas de ML para antecipar oscilações no preço do farelo de soja em nível regional, contribuindo para decisões mais assertivas por parte de produtores, indústrias e agentes de mercado.

Index terms—Aplicação, modelos, automatização, dashboards.

I. INTRODUÇÃO

A soja (*Glycine max*) é uma oleaginosa de origem asiática e a mais produzida e comercializada no mundo. Dentre seus

principais derivados estão o óleo e o farelo, utilizados tanto para alimentação humana e animal quanto para a produção de biocombustíveis e outros químicos. O Brasil, juntamente com os Estados Unidos e a Argentina, são os principais produtores do grão, detendo 80% do total mundial [1]. A soja, como matéria-prima, é amplamente utilizada na produção de óleo e farelo, sendo o farelo um subproduto crucial na alimentação animal e na indústria de rações [2].

A previsão dos preços das commodities, especialmente no curto prazo, tem sido um desafio significativo devido à alta volatilidade e à complexidade dos fatores que influenciam esses preços. Diversos estudos têm abordado a utilização de modelos de *Machine Learning* (ML) e técnicas avançadas para melhorar a precisão das previsões [3], [4], [5], [6]. O farelo de soja é uma das matérias-primas mais utilizadas na alimentação animal, e seu preço influencia diretamente os custos de produção de alimentos para pecuária, especialmente no Brasil, onde a produção de carne e leite tem uma forte correlação com o preço do farelo de soja. De acordo com Almeida [7], a volatilidade nos preços do farelo de soja pode afetar significativamente a rentabilidade de produtores de proteína animal, e, por conseguinte, impactar a economia local e nacional.

Diversos modelos de previsão têm sido empregados na tentativa de antecipar o comportamento dos preços de commodities [8], de forma geral, e do grão de soja [9]. Entre as abordagens mais comuns, destacam-se os métodos tradicionais de séries temporais, como ARIMA (*Autoregressive Integrated Moving Average*). Técnicas mais recentes baseadas em ML, como Redes Neurais Artificiais (RNAs), Máquinas de Vetores de Suporte (SVM) e Florestas Aleatórias (*Random Forest* - RF) [10], [11], redes neurais recorrentes LSTM (*Long Short-Term Memory*) [9] e modelos híbridos combinando algoritmos genéticos e gradiente descendente [3], têm mostrado bons resultados em cenários de alta volatilidade. Modelos como LSTNet [12] e *K-Nearest Neighbours* (KNN) [13] também têm sido explorados, com foco na precisão das previsões. Além disso, estudos mais

recentes, como o de Liu [6], têm introduzido novos modelos híbridos, como o CNN-ILSTM-SA (redes neurais convolucionais - CNN, ILSTM - *Improved Long Short-Term Memory* e SA - *Self-Attention Mechanism*), para melhorar a acurácia das previsões de preços de soja grão, destacando-se pela alta precisão e baixo erro. Além disso, modelos baseados na arquitetura *Transformer* vêm se destacando na modelagem de séries temporais multivariadas, com aplicações crescentes na agricultura e nos mercados futuros de grãos [14]. O modelo *Prophet*, desenvolvido pelo Facebook, tem se mostrado eficaz na previsão de safras e preços com forte tendência e sazonalidade, como ocorre com derivados da soja [15]. Adicionalmente, abordagens híbridas — como ARIMA-LSTM (*AutoRegressive Integrated Moving Average* combinada com *Long Short-Term Memory*), STL-*Prophet* (*Seasonal and Trend decomposition using Loess* combinada com o modelo *Prophet*) ou métodos de decomposição como EMD (*Empirical Mode Decomposition*) integrados a redes neurais — vêm sendo exploradas com o objetivo de aprimorar a acurácia preditiva em contextos agrícolas e de comercialização de insumos [16], [17], [5]. No entanto, esses estudos, não fazem a previsão para o farelo de soja, e nem consideram as especificidades regionais, como a de Campinas.

A previsão precisa do preço do farelo de soja é fundamental para diversos agentes do mercado, incluindo produtores, processadores, cooperativas, investidores e formuladores de políticas públicas. No caso dos produtores rurais, esse tipo de previsão auxilia no planejamento da produção e na definição do melhor momento para comercialização. Indústrias de ração animal podem utilizá-la para otimizar compras e gestão de estoques, minimizando riscos decorrentes da volatilidade. Além disso, informações preditivas confiáveis subsidiam políticas públicas voltadas à regulação do mercado de soja e seus derivados, como discutido por Barbosa [18].

Com o avanço das técnicas de aprendizado de máquina, a análise preditiva de preços de commodities tornou-se mais eficiente. Modelos como redes neurais artificiais (RNAs), LSTM e algoritmos híbridos, especialmente quando combinados com métodos de otimização, têm superado abordagens tradicionais como o ARIMA, oferecendo maior precisão em ambientes de alta volatilidade.

Apesar de avanços na previsão de preços em nível nacional e global, ainda são escassos os estudos que abordam a previsão do preço do farelo de soja em contextos regionais. Campinas, por sua relevância no agronegócio e presença de indústrias processadoras de soja, apresenta um mercado com dinâmicas próprias. Nesse sentido, torna-se necessário o desenvolvimento de modelos específicos que considerem tanto as variáveis locais quanto as tendências globais. Além disso, dado o alto grau de interdependência entre o preço da soja grão e do farelo, é essencial incorporar essa relação dinâmica nos modelos preditivos.

Neste contexto, o objetivo deste estudo é ajustar e comparar diferentes modelos de aprendizado de máquina para compreender a dinâmica e prever o preço diário do farelo de soja na região de Campinas/SP. Embora existam estudos aplicando modelos como LSTM e ARIMA à previsão do farelo de soja em nível estadual — como no Paraná, com séries mensais [19], não foram identificados trabalhos que explorem dados diários nem que avaliem um conjunto mais amplo de técnicas preditivas baseadas em *Machine Learning*.

Uma das principais contribuições deste trabalho consiste na aplicação e avaliação de modelos como XGBoost, LightGBM, *Random Forest*, SVM (linear, polinomial e *Radial*

Basis Function - RBF), KNN, MLP, MARS e *Cubist Rules*, utilizando dados diários de preços do farelo de soja. A comparação entre os modelos considera diferentes métricas de desempenho e cenários de teste, buscando identificar a abordagem mais eficaz para o contexto regional.

Este estudo integra um projeto mais amplo, realizado em parceria com uma empresa do setor agropecuário com forte atuação na cadeia produtiva do leite. O projeto envolve o desenvolvimento de modelos preditivos para preços de insumos estratégicos para a nutrição animal, como milho, farelo de soja, polpa cítrica e caroço de algodão. Os modelos são ajustados em linguagem R [20] e disponibilizados por meio de dashboards interativos, construídos com a ferramenta Quarto. Essa iniciativa visa oferecer uma solução prática e visual para o monitoramento das dinâmicas de mercado, contribuindo para a tomada de decisão empresarial. Além dos dados da região de Campinas, foram modelados preços de insumos em outras praças de negociação, totalizando aproximadamente 150 modelos atualizados diariamente com dados reais de mercado.

Este estudo se destaca por integrar técnicas modernas de aprendizado de máquina à previsão diária de preços do farelo de soja, com foco regional na praça de Campinas/SP — uma lacuna ainda pouco explorada na literatura. Ao empregar uma ampla variedade de modelos preditivos, aliando métodos tradicionais e algoritmos de ponta, o trabalho oferece uma análise comparativa robusta e replicável. Além disso, ao incorporar o modelo preditivo a dashboards interativos desenvolvidos em ambiente R/Quarto, o estudo avança não apenas no aspecto metodológico, mas também na aplicabilidade prática, fornecendo uma ferramenta concreta de apoio à decisão para agentes do agronegócio. A abordagem proposta pode ser expandida para outras regiões e produtos, contribuindo com soluções preditivas voltadas ao planejamento estratégico e à mitigação de riscos em cadeias agroindustriais.

Esse trabalho é organizado como segue. Na metodologia, apresentamos a fonte dos dados e as principais bibliotecas em R utilizadas. Também indicamos os modelos avaliados e as métricas adotadas na comparação. Na seção de resultados, abordamos a seleção de hiperparâmetros, apresentamos uma discussão detalhada sobre a comparação dos modelos a partir das medidas de desempenho nos dados de teste e de treinamento. A seção também inclui a previsão dos preços para o primeiro mês seguinte e é finalizada com uma breve apresentação do dashboard desenvolvido para essa aplicação.

II. METODOLOGIA

A. Base de Dados

Os dados deste estudo foram obtidos na plataforma Broadcast, cobrindo o período de 2007 a 2024, com 4398 observações diárias dos preços da soja (R\$ por tonelada) na região de Campinas, São Paulo. Esse intervalo de 17 anos permite analisar as variações de preço da soja, influenciadas por fatores como eventos climáticos, mudanças econômicas e oscilações na oferta e demanda global. Antes de serem utilizados nos modelos de previsão, os dados passaram por um processo de pré-processamento e validação. Este trabalho foi realizado em parceria com a empresa JPA Agro do Grupo JPA, localizada em Lavras, Minas Gerais. O Grupo JPA é uma *holding* que atua no segmento agro desde 2009, propondo soluções inovadoras na conexão, comercialização, produção e fabricação de produtos.

Um dos principais desafios metodológicos deste estudo foi a automatização do processo de entrada e estruturação dos dados, etapa fundamental para garantir a atualização contínua e a operacionalização das soluções preditivas no ambiente corporativo. A coleta dos dados ocorre por meio de uma função automatizada disponibilizada pela plataforma Broadcast, implementada como extensão no Microsoft Excel, permitindo a extração periódica e padronizada das informações necessárias. Posteriormente, os dados são organizados e armazenados com o uso da linguagem R [20] no pacote *duckdb* [21]. O *DuckDB* é um sistema de gerenciamento de banco de dados analítico embutido, projetado para executar consultas SQL de forma eficiente diretamente em arquivos locais, sem necessidade de servidor, ideal para análise de dados em R e Python. Essa abordagem possibilitou a estruturação dos dados em formato relacional, facilitando seu gerenciamento, rastreabilidade e integração com os modelos preditivos. No entanto, a automatização deste fluxo representou um desafio técnico relevante, exigindo a adoção de práticas para assegurar a consistência, escalabilidade e robustez do pipeline, aspectos indispensáveis para a sua incorporação nas rotinas operacionais da empresa.

A implementação foi realizada utilizando a linguagem R [20]. As bibliotecas adotadas foram a *tidyverse* [22] e a *tidymodels* [23]. *tidyverse* é um conjunto de pacotes que facilitam a manipulação, visualização e transformação de dados de forma coesa e intuitiva. *tidymodels* oferece ferramentas para construção, avaliação e ajuste de modelos preditivos, integrando todas as etapas do processo de modelagem. Ambos visam otimizar fluxos de trabalho de análise e técnicas de ML, proporcionando consistência e eficiência.

C. Modelos

Na literatura sobre previsão de preços de soja e farelo de soja, diversos modelos de ML são utilizados para capturar relações não lineares e padrões complexos nos dados, sendo particularmente úteis em mercados voláteis. Sendo assim, uma breve descrição dos modelos de ML analisados e das medidas de desempenho utilizados neste trabalho será apresentada.

Os 10 modelos de ML comparados incluem: o XGBoost, que é uma implementação eficiente de *Gradient Boosting*, utilizando árvores sequenciais para corrigir erros das anteriores; o LightGBM é uma versão otimizada do XGBoost, usando histogramas e crescimento "leaf-wise" para maior desempenho em grandes volumes de dados; o *Cubist Rules* combina regressão com árvores de decisão, ajustando funções lineares em cada nó; o MARS (*Multivariate Adaptive Regression Splines*) é uma técnica de regressão não linear que utiliza splines para modelar interações complexas entre variáveis; o *Multilayer Perceptron* (MLP), uma rede neural que ajusta pesos de camadas interconectadas para capturar padrões não lineares; o KNN que se baseia na média dos valores dos K vizinhos mais próximos para fazer previsões; o RF, um conjunto de árvores de decisão que reduz o *overfitting* e melhora a precisão preditiva; o SVM que busca um hiperplano ótimo para separar classes, com diferentes tipos de kernels como o Linear, Polinomial e RBF (*Radial Basis Function*) para lidar com dados lineares e não lineares.

D. Medidas de Avaliação dos Modelos

Com base na literatura e em pesquisas acadêmicas, este estudo adota quatro medidas para avaliar o desempenho dos

modelos preditivos. O Erro Médio Absoluto (MAE) que avalia a média das diferenças absolutas entre valores observados e previstos, refletindo a precisão geral. A Raiz do Erro Quadrático Médio (RMSE) atribui maior peso a erros elevados, sendo útil para penalizar grandes desvios. O Erro Percentual Absoluto Médio (MAPE) expressa os erros como uma porcentagem dos valores reais, facilitando comparações entre modelos. Por fim, o Coeficiente de Determinação (R^2) indica a proporção da variabilidade explicada pelo modelo, com valores próximos de 1 sugerindo melhor ajuste. Já para as medidas de MAE, RMSE e MAPE, quanto menores seus valores, melhor o desempenho do modelo.

E. Arquitetura

A previsão do preço do farelo de soja foi diária. As covariáveis utilizadas para ajuste foram o tempo e o parâmetro de componentes autorregressivos de primeira ordem (AR-1). O uso de outras covariáveis que podem afetar o preço desse insumo tais como a cotação do dólar americano, oferta global e o seu preço em mercados futuros, estão fora do escopo desse trabalho. O conjunto de dados históricos diários de 2007 a 2021 (julho) foi utilizado para treinar todos os modelos empregados neste trabalho de pesquisa. A estrutura proposta é ilustrada na Fig. 1.



Figura 1. Arquitetura proposta. Fonte: Do Autor (2025).

O conjunto de dados de 2021 (julho) a 2024 (dezembro) foi reservado como conjunto de teste. Vale ressaltar que o conjunto de treinamento também foi dividido na proporção de 20%–80%, onde 20% foram estritamente reservados para validação, enquanto o restante foi utilizado para o treinamento. No treinamento foi utilizada a validação cruzada de *10-fold* para ajuste dos hiperparâmetros e avaliação dos modelos. Assim, a cada iteração, um conjunto constituído por *k-1 fold* é utilizado para treinamento e o restante para validação. O preditor é treinado 10 vezes, mudando os conjuntos de treinamento e garantindo que cada subconjunto será utilizado para validação em algum *fold* da avaliação do modelo [24].

A. Seleção de Hiperparâmetros

A otimização de hiperparâmetros é o processo de escolher os melhores valores para parâmetros externos de um modelo, influenciando seu desempenho e generalização. Exemplos incluem o número de árvores em RF, a taxa de aprendizado em RNAs e a regularização em SVM.

No presente estudo, os hiperparâmetros foram otimizados individualmente para cada modelo, com o objetivo de maximizar a precisão preditiva e minimizar o erro. Para isso, foi adotada uma abordagem sistemática baseada em validação cruzada e algoritmos de otimização, como *Grid Search* e *Bayesian Optimization*.

A otimização foi implementada por meio da função *tune_grid*, do pacote *tidymodels* do R [22], [23], que realiza uma busca estruturada em grade, equilibrando a qualidade da solução com a eficiência computacional. Esse procedimento permitiu encontrar a melhor combinação de hiperparâmetros para cada modelo com base nos dados históricos de preços do farelo de soja, melhorando o desempenho preditivo e reduzindo o tempo de processamento.

A Tabela I retrata os valores dos hiperparâmetros otimizados através da função *grid_search* para cada um dos modelos analisados. Com base nos melhores hiperparâmetros encontrados, as medidas de desempenho dos modelos podem ser comparadas para os dados de treino e de teste.

TABELA I. VALORES DOS HIPERPARÂMETROS OTIMIZADOS PARA CADA MODELO

Modelos	Hiperparâmetros otimizados
XGBoost	nrounds = 500; max_depth = 5; eta = 0,0243; subsample = 0,9448; colsample_bytree = 1; gamma = 0,0222; min_child_weight = 5
LightGBM	num_leaves = 500; tree_depth = 3; learning_rate = 0,0188; mtry = 1; subsample = 0,000000735; min_n = 14
Cubist Rules	max_rules = 26; committees = 71; number_of_samples = 3518
MARS	prod_degree = 2; pruning = 0; nprune = 3
MLP	hidden_units = 5; activation = linear; learning_rate = 0,001; epochs = 783; decay = $5,092931 \times 10^{-10}$
KNN	k = 3; distance_metric = 1,1728; kernel = triangular
RF	n_estimators = 500; min_n = 3; mtry = 1
SVM Linear	C = 0.106; epsilon = 0,00353; type = 11
SVM Polinomial	C = 2.0602; epsilon = 0,0902; degree = 2; scale = 0,0129; offset = 1; number_of_support_vectors = 166
SVM RBF	C = 31,5032; epsilon = 0,00787; sigma = 0,000504; number_of_support_vectors = 2547

B. Medidas de Desempenho dos Modelos

Os modelos foram avaliados com base nas medidas de desempenho RMSE, MAE, MAPE e R^2 . Na Fig. 2 têm-se o resumo dessas medidas para os modelos analisados para os dados de treinamento e a Fig. 3, retrata os resultados dos modelos nos dados de teste. Como foi realizado o *K-fold*, os valores em parênteses refletem o desvio médio.

Para as métricas de erro MAE, RMSE e MAPE, valores mais baixos indicam melhor desempenho, enquanto para o R^2 , valores mais próximos de 1 ou 100% representam maior capacidade explicativa do modelo. A escala de cores do *heatmap* reforça essas avaliações: tons azulados indicam desempenho superior, enquanto tons avermelhados apontam desempenho inferior quando consideradas as métricas MAE, RMSE e MAPE. Para o R^2 , a coloração em tons avermelhados apontam desempenho superior e tons de cor azulados indicam desempenho inferior. Essa representação gráfica permite rápida identificação dos modelos mais adequados e menos eficazes para a tarefa em questão (Fig. 2 e Fig. 3)

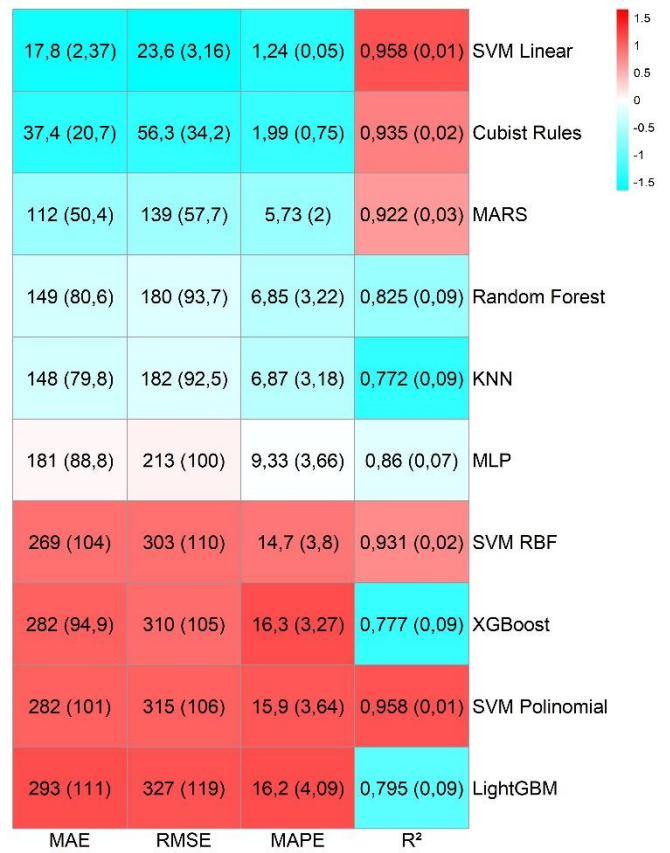


Figura 2. Medidas de desempenho dos dados de Treinamento. Fonte: Do Autor (2025).

O modelo XGBoost, com um MAE de 282,0 e um R^2 de 0,777, demonstra um bom equilíbrio entre complexidade e precisão. O valor de R^2 , embora inferior ao de alguns outros modelos, ainda é representativo de uma forte capacidade explicativa, visto que ele captura mais de 77% da variabilidade dos preços do farelo de soja. Essa medida de desempenho corroborou os achados de Zhang [25], que demonstram a eficácia dos modelos baseados em boosting, como o XGBoost, em tarefas de previsão de séries temporais complexas, incluindo commodities agrícolas como o farelo de soja.

No entanto, o RMSE de 310 e o MAE de 282 indicam que, embora o modelo seja robusto, há uma margem de erro considerável (Fig. 2). Por outro lado, o modelo LightGBM,

com um MAE de 293,0 e R^2 de 0,795, apresentou resultados ligeiramente superiores ao XGBoost, sugerindo um desempenho mais estável e preciso. Essa maior precisão pode ser explicada pela maior eficiência do modelo em lidar com grandes volumes de dados e com características de alta dimensionalidade. O valor de R^2 de 0,795 sugere que o modelo explica quase 80% da variabilidade nos preços do farelo de soja, o que é um bom indicativo de sua aplicabilidade para previsões no longo prazo. Este desempenho reforça estudos como o de Yuan [26], destacam o desempenho superior do LightGBM em dados de alta dimensionalidade e complexidade, o que também se aplica ao mercado de farelo de soja de Campinas, onde diversas interações entre fatores locais e globais afetam os preços.

Os modelos *Cubist Rules* e MARS com R^2 de 0,935 e 0,922, respectivamente, apresentam um desempenho excelente ao lidar com variáveis não lineares e complexas, características típicas do mercado de farelo de soja. O MARS, por exemplo, possui a capacidade de ajustar de forma flexível as relações não lineares entre variáveis explicativas. Os resultados indicam que o modelo não apenas tem um R^2 alto, mas também uma boa capacidade de generalização. O valor de MAE de 112 e RMSE de 139 do MARS refletem um desempenho sólido na previsão dos preços do farelo de soja, mostrando que o modelo consegue reduzir os erros em relação aos modelos de boosting, com a vantagem adicional de ser mais interpretável. Já os modelos de SVM Linear e SVM Polinomial, com um desempenho impressionante em termos de R^2 (0,958), sugerem que, embora seus valores de MAE sejam mais elevados, esses modelos são eficientes em capturar as tendências e as variáveis latentes que afetam o preço do farelo de soja no contexto de Campinas. A medida de desempenho de R^2 reflete uma boa capacidade de explicação dos dados, mas os modelos ainda são sensíveis ao ajuste de seus parâmetros e kernels, como discutido por Liu [27].

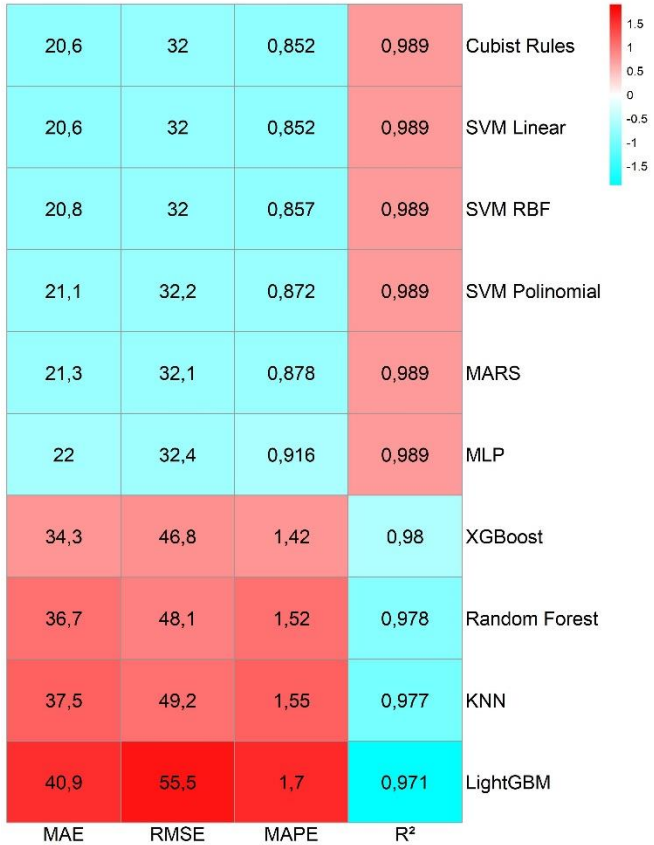


Figura 3. Medidas de desempenho dos dados de Teste. Fonte: Do Autor (2025).

A Fig. 4, apresenta um gráfico comparando os valores reais com as previsões dos modelos, destacando a separação entre os períodos de treino e teste.

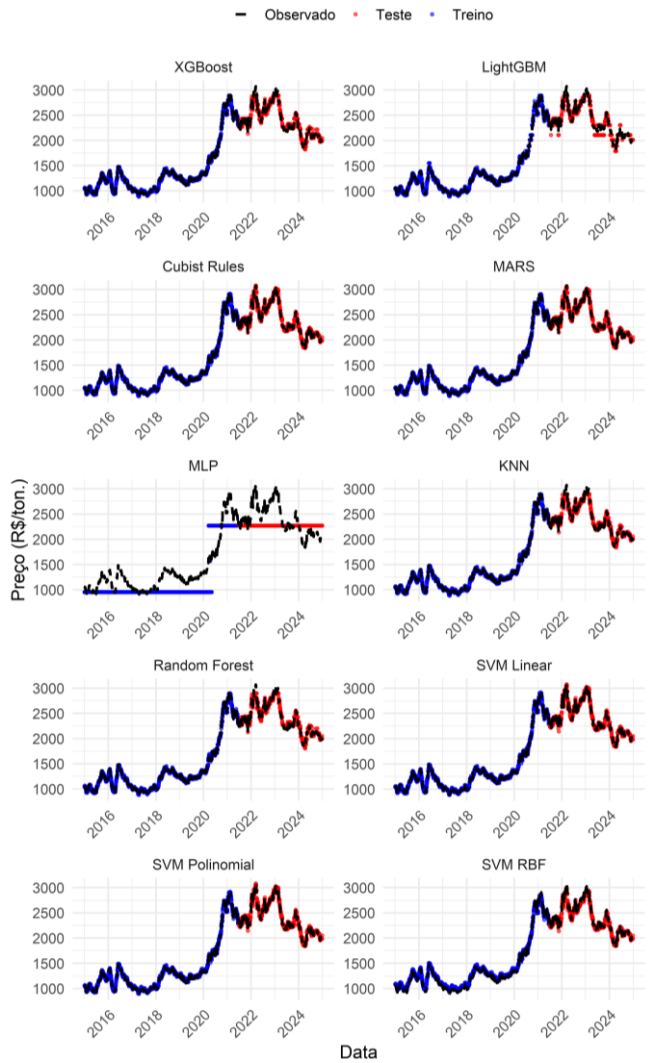


Figura 4. Valores reais com as previsões dos modelos, destacando a separação entre os períodos de treino e teste. Fonte: Do Autor (2025).

A análise dos resultados dos modelos de previsão do preço do farelo de soja por tonelada revela importantes insights sobre a capacidade de generalização e o ajuste de diversos algoritmos aplicados a esse tipo de previsão. Modelos como SVM Linear, SVM Polinomial e Cubist Rules se destacaram pelo bom desempenho (MAE, RMSE e MAPE) (Fig. 3 e Fig. 4). Os modelos SVM Linear e SVM Polinomial têm se mostrado eficazes em diversos contextos de previsão de séries temporais, especialmente em problemas de alta dimensionalidade. Isso está em linha com estudos de Gao [28], que demonstram a robustez do SVM na previsão de preços de commodities agrícolas, como a soja.

Vapnik [29], ao introduzir o SVM, destacou sua capacidade de maximizar margens e obter boa generalização, uma característica que é especialmente útil na previsão do preço do farelo de soja, dado que este mercado está sujeito a muitas variáveis externas e incertezas. O elevado R^2 observado nos dados de teste, próximo de 0,989 para SVM Linear e Polinomial indicam um bom ajuste e forte capacidade de generalização. Estudos, como Zhou [30], também destacam a alta precisão dos modelos SVM na previsão de preços de commodities, sugerindo sua eficácia na captura de tendências e flutuações.

O modelo *Cubist Rules* apresentou bom desempenho, com R^2 elevado e baixa variabilidade. Estudos, como de Kuhn [31], destacam sua interpretabilidade e eficácia em problemas agrícolas, especialmente quando há interações complexas entre variáveis, como no mercado de farelo de soja. A robustez do *Cubist Rules* e sua capacidade de lidar com dados ruidosos explicam sua eficiência. Silva [32] também destacam que modelos baseados em árvores de decisão, como o *Cubist*, oferecem maior precisão na previsão de preços agrícolas em comparação com algoritmos mais simples, como a regressão linear.

A análise indica que SVM Linear e *Cubist Rules* são ideais para prever os preços do farelo de soja em Campinas, devido à sua adaptabilidade e eficácia na modelagem de dados econômicos e agrícolas [28], [30]. A comparação dos resultados reforça que SVM Linear, SVM Polinomial e *Cubist Rules* são os mais eficazes, tiveram desempenho superior em precisão e generalização.

C. Tempo Computacional

Os modelos foram avaliados com base na análise de tempo computacional e custo-benefício. Na Fig. 5 têm-se o resumo dessas medidas para os modelos analisados para os dados de treinamento.

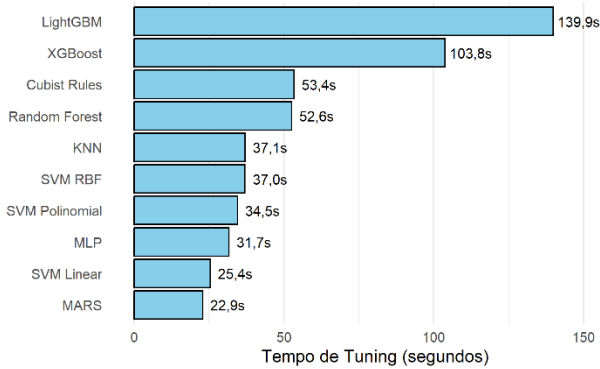


Figura 5. Tempo computacional (em segundos) para o *tuning* dos diferentes modelos de Machine Learning utilizados no estudo. Fonte: Do Autor (2025).

Entre os modelos avaliados, o SVM Linear apresentou o melhor desempenho global (Fig. 5). Com um tempo de *tuning* inferior a 30 segundos, esse modelo obteve os menores valores de erro (MAE = 17,8; RMSE = 23,6; MAPE = 1,24) e o maior R^2 (0,958; Fig. 2), evidenciando sua elevada capacidade de generalização associada a um custo computacional baixo. Tal resultado destaca o potencial deste modelo para aplicações em ambientes de produção que demandem atualizações frequentes e baixo tempo de latência.

O *Cubist Rules* também se mostrou eficiente, apresentando desempenho competitivo ($R^2 = 0,935$) com tempo computacional reduzido (53 segundos), o que reforça sua adequação para cenários que demandem interpretabilidade aliada a boa acurácia. O modelo MARS, embora apresente menor acurácia ($R^2 = 0,922$), destacou-se pelo menor tempo de *tuning* (22,9 segundos), o que pode justificar seu uso em sistemas com severas restrições computacionais.

Por outro lado, modelos tradicionalmente robustos como XGBoost e LightGBM apresentaram desempenho inferior ao esperado no contexto avaliado. Ambos demandaram tempos de *tuning* superiores a 100 segundos, mas obtiveram métricas de erro elevadas (RMSE > 300; MAPE > 15%), além de R^2 relativamente baixos (0,777 e 0,795, respectivamente). Este resultado sugere que, para o conjunto de dados em questão,

modelos com maior complexidade e custo computacional não necessariamente proporcionam melhor desempenho, sendo portanto menos indicados quando se considera a relação custo-benefício.

Os modelos baseados em vizinhos mais próximos (KNN) e florestas aleatórias (*Random Forest*) apresentaram desempenho intermediário, mas não competitivo, com R^2 abaixo de 0,83 e erros consideráveis (MAE e RMSE acima de 140 e 180, respectivamente). O desempenho do MLP também foi inferior ao dos melhores modelos, mesmo com tempo computacional semelhante, o que reforça a necessidade de cuidados no *tuning* de redes neurais simples. Destaca-se ainda o comportamento inconsistente do SVM Polinomial e SVM RBF, que apesar de apresentarem R^2 elevados (0,931–0,958), incorreram em erros absolutos muito altos (MAE > 260), sugerindo *overfitting* ou inadequação das parametrizações.

Esses resultados reforçam a importância de considerar o tempo de ajuste e a estabilidade dos modelos, além das métricas de erro usuais. Modelos mais simples, como o SVM Linear e o *Cubist Rules*, mostraram-se não apenas mais eficientes computacionalmente, mas também superiores em acurácia preditiva neste cenário, demonstrando que maior complexidade algorítmica não garante necessariamente melhor desempenho.

D. Previsão

A Fig. 6 e a Tabela II apresentam os preços reais do farelo de soja e os preços previstos pelo melhor modelo ML no período de 27 de dezembro de 2024 a 27 de janeiro de 2025 (intervalo de 30 dias).

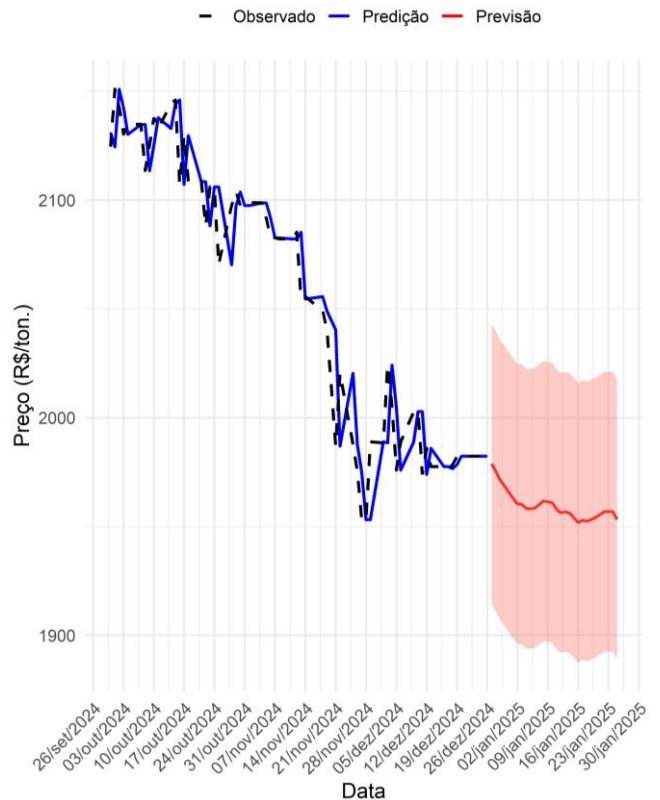


Figura 6. Preço diários real e previstos do farelo de soja para o melhor modelo de Machine Learning. Fonte: Do Autor (2025).

Observa-se uma tendência de queda nos preços do farelo de soja ao longo do período analisado (Fig. 6; Tabela II). O modelo foi capaz de captar corretamente essa dinâmica, com estimativas centrais próximas aos valores reais observados.

TABELA II. PREÇO REAL DO FARELO DE SOJA E PREÇOS DIÁRIOS PREVISTOS PARA O MELHOR MODELO DE MACHINE LEARNING DE 27 DE DEZEMBRO DE 2024 A 27 DE JANEIRO DE 2025

Data	Preço real (R\$/ton.)	Preço previsto (R\$/ton.)	Limite inferior	Limite superior
27/12/2024	2047	1978	1914	2043
30/12/2024	2047	1968	1904	2032
02/01/2025	2047	1960	1896	2024
03/01/2025	2034	1960	1895	2024
06/01/2025	2001	1958	1894	2022
07/01/2025	1993	1960	1895	2024
08/01/2025	1992	1961	1987	2025
09/01/2025	1992	1961	1987	2025
10/01/2025	1992	1961	1896	2025
13/01/2025	2015	1956	1892	2021
14/01/2025	1991	1956	1891	2020
15/01/2025	1981	1954	1889	2018
16/01/2025	1967	1951	1887	2016
17/01/2025	1967	1952	1888	2017
20/01/2025	1967	1954	1889	2018
21/01/2025	1974	1955	1891	2019
22/01/2025	1955	1956	1892	2021
23/01/2025	1955	1956	1892	2021
27/01/2025	1955	1956	1892	2021
27/01/2025	1949	1953	1888	2017

Além disso, os intervalos de previsão incluíram os preços efetivos, o que demonstra a confiabilidade das projeções para fins de planejamento de compras e vendas. Apesar de pequenas variações pontuais, o desempenho do modelo se manteve robusto e coerente com o comportamento real do mercado.

Embora este artigo tenha utilizado dados da praça de Campinas/SP como estudo de caso para fins de demonstração, o sistema desenvolvido é flexível e escalável, podendo ser adaptado para diferentes regiões do país com simples alterações na base de dados de entrada. Isso torna a proposta especialmente útil para instituições e empresas que atuam em diversas regiões, necessitando de monitoramento constante e previsões de curto prazo para otimizar decisões operacionais e comerciais.

Outro aspecto relevante deste trabalho foi a estruturação de um fluxo automatizado para o ajuste, comparação e seleção de modelos preditivos, aplicados não apenas ao mercado de soja, mas também a outros insumos utilizados na alimentação animal, como milho, caroço de algodão e polpa cítrica. Esse fluxo foi implementado na linguagem R, utilizando majoritariamente os pacotes do ecossistema *tidymodels* [22], como *parsnip* para a especificação de modelos, *recipe* para o pré-processamento de dados, *workflows* para a integração entre etapas, e *tune_grid* para a otimização automática de hiperparâmetros por meio de validação cruzada.

A automação do processo permite que, a partir da entrada de novos dados, sejam executadas de forma sequencial e sistemática as etapas de pré-processamento, ajuste dos modelos, avaliação de desempenho e seleção da melhor configuração com base em métricas padronizadas. Essa estrutura reduz significativamente o tempo necessário para a atualização periódica dos modelos, facilita sua manutenção contínua e possibilita a rápida replicação do processo para

múltiplas praças e produtos, ampliando substancialmente o potencial de uso prático da ferramenta em diferentes contextos e mercados.

Entretanto, a construção desse fluxo apresentou desafios importantes, especialmente relacionados à padronização do pré-processamento para bases de dados com características distintas, à definição de estratégias adequadas de validação para séries temporais e à necessidade de garantir a escalabilidade do pipeline frente ao aumento do volume de dados e à diversidade de insumos analisados. A integração com o banco de dados local, estruturado via *DuckDB* [21], também exigiu cuidados para assegurar a eficiência na extração e transformação dos dados utilizados na modelagem.

De modo geral, a automação promoveu ganhos expressivos em eficiência, reprodutibilidade e robustez analítica, além de estabelecer uma base sólida para a futura expansão do sistema, permitindo a incorporação de novos insumos, mercados e técnicas de modelagem.

E. Comunicação dos Resultados Através de Dashboards

A comunicação dos resultados da modelagem preditiva de preços do farelo de soja foi realizada por meio de dashboards interativos e responsivos. Estes foram desenvolvidos utilizando as ferramentas Quarto e Shiny da linguagem R. A finalidade foi apresentar e concentrar em apenas um lugar a análise da dinâmica dos preços, as previsões de mercado, métricas dos modelos e outros dados relevantes para a empresa parceira neste projeto

A arquitetura dos dashboards foi organizada em abas, cada uma dedicada a um tipo de análise. Conforme ilustrado na Fig. 7, que apresenta a página de previsão, é possível observar a organização com uma barra de navegação superior e um painel de controle contendo as métricas de desempenho dos modelos (RMSE, MAPE, MAE e R²), apresentadas em *ValueBoxes*. A área principal exibe o gráfico da série temporal, incluindo as previsões.

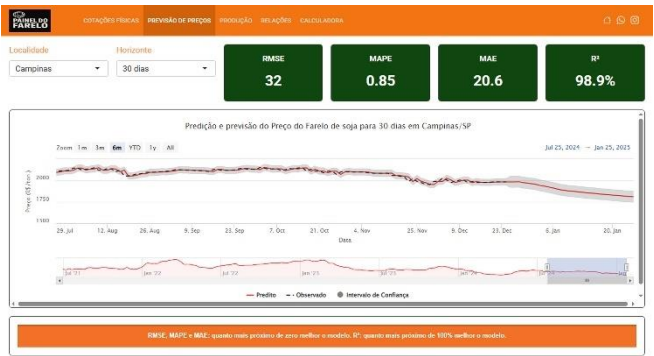


Figura 7. Página de previsão do Dashboard de Farelo de Soja. Fonte: Do Autor (2025).

A reatividade do sistema otimizou a experiência do usuário, permitindo que a seleção da praça (localidade) e a alteração do horizonte de previsão atualizassem dinamicamente as informações. Além disso, a fim de aprimorar a interatividade, a biblioteca *highcharter* [33] foi utilizada, possibilitando funcionalidades como zoom dinâmico e exibição de detalhes ao passar o mouse, o que permitiu a exploração detalhada dos dados.

O desenvolvimento dos dashboards foi um esforço colaborativo com a equipe da JPA Agro. O feedback contínuo

da empresa garantiu que a interface e as funcionalidades estivessem alinhadas com as necessidades dos tomadores de decisão.

IV. CONCLUSÃO

Este estudo evidenciou o potencial dos modelos de aprendizado de máquina (ML) na previsão dos preços diários do farelo de soja na região de Campinas/SP. Dentre os modelos avaliados, a Máquina de Vetores de Suporte com kernel linear (SVM Linear) apresentou o melhor desempenho preditivo, destacando-se como a alternativa mais precisa para esta aplicação específica.

Os resultados obtidos não apenas capturaram adequadamente a tendência de preços do insumo, como também demonstraram aplicabilidade prática, ao fornecer subsídios que podem contribuir para a gestão de riscos e o planejamento operacional de diversos agentes da cadeia produtiva do leite, como produtores, cooperativas e indústrias de nutrição animal.

Como direções futuras, propõe-se o aprimoramento dos modelos por meio da incorporação de variáveis exógenas como clima, câmbio, preços internacionais e custos logísticos, além da realização de análises comparativas que considerem também o tempo computacional e a escalabilidade dos algoritmos, visando aplicações em larga escala e em tempo real.

AGRADECIMENTOS

Agradecemos à FAPEMIG pelo apoio financeiro e à empresa JPA Agro pela parceria no projeto de pesquisa ao qual esse trabalho está vinculado.

REFERÊNCIAS

- [1] Lemos, M. L. F., Guimarães, D. D., Maia, G. B. S. and Amaral, G. F. (2018). Agregação de valor na cadeia da soja. BNDES Setorial, v. 46, p. 167–217.
- [2] Sindirações (2024). Boletim Informativo do Setor. Disponível em: <https://sindiracoes.org.br/boletim-informativo-do-setor/>. Acesso em: 02 fev. 2025.
- [3] Zhang, J., Li, Y. and Wang, T. (2018). "A Hybrid Model for Soybean Price Forecasting using Radial Basis Function Neural Networks and Genetic Algorithms." *Agricultural Economics Review*, 23(3), 202–217.
- [4] Reis Filho, J., Santos, V. and Lima, M. (2020). "Text Mining for Predictive Analysis of Agricultural Commodities Prices." *Journal of Economic Forecasting*, 12(2), 98–112.
- [5] Palazzi, R. B., Maçaira, P., Meira, E. and Klotzle, M. C. (2023). Forecasting commodity prices in Brazil through hybrid SSA-complex seasonality models. *Production*, 33.
- [6] Liu, H., Pan, Y., Liu, W. and Wang, J. (2024). A Prediction Model for Soybean Meal Futures Price Based on CNN-ILSTM-SA. *IEEE Access*, 12, 111663–111672.
- [7] Almeida, F. F., et al. (2021). "The Impact of Soybean Price Volatility on Brazilian Livestock Industry." *Agricultural Economics*, 52(1), 111–124.
- [8] Ribeiro, C. O. and Oliveira, S. M. (2011). A hybrid commodity price-forecasting model applied to the sugar-alcohol sector. *The Australian Journal of Agricultural and Resource Economics*, 55(2), 180–198.
- [9] Wang, Y. and Gao, S. (2018). Predicting the Future High and Low Prices of Soybeans using LSTM Networks. *International Journal of Financial Modeling*, 14(1), 45–58.
- [10] Silva, G. A., et al. (2020). "Prediction of Soybean Prices Using Machine Learning Algorithms." *Journal of Machine Learning in Agriculture*, 15(4), 47–60.
- [11] Lima, R. S., et al. (2022). Machine Learning Applications for Soybean Price Prediction in Brazil. *Computers in Agriculture*, 40(2), 79–89.
- [12] Ouyang, Y., Wei, C. and Wu, Y. (2019). "Long-and Short-Term Temporal Patterns with Neural Networks for Commodity Price Prediction." *Computational Economics and Finance*, 17(4), 300–315.
- [13] Prashantha, S., Kumar, R. and Patel, P. (2020). Machine Learning Techniques for Agricultural Commodity Price Forecasting. *Agricultural Systems Journal*, 9(1), 23–38.
- [14] Nayak, G.H.H., Alam, M.W., Naik, B.S. et al. (2025) Meta-transformer: leveraging metaheuristic algorithms for agricultural commodity price forecasting. *J Big Data* 12, 138.
- [15] Yoo, T.-W. and Oh, I.-S. (2020). Time Series Forecasting of Agricultural Products' Sales Volumes Based on Seasonal Long Short-Term Memory. *Applied Sciences*, 10(22), 8169.
- [16] Guo, Y.; Tang, D.; Tang, W.; Yang, S.; Tang, Q.; Feng, Y.; Zhang, F. (2022) Agricultural Price Prediction Based on Combined Forecasting Model under Spatial-Temporal Influencing Factors. *Sustainability*, 14, 10483.
- [17] Sun, C., Pei, M., Cao, B., Chang, S. and Si, H. (2023) A Study on Agricultural Commodity Price Prediction Model Based on Secondary Decomposition and Long Short-Term Memory Network, *Agriculture*, 14.
- [18] Barbosa, D. D., et al. (2020). "Public Policy and the Influence of Soybean Price Forecasting in Brazil." *Brazilian Journal of Agricultural Studies*, 45(3), 299–312.
- [19] Dos Santos, J. A. A. "Aplicação de modelos de séries temporais na previsão do preço do farelo de soja," *Revista Mundi Engenharia, Tecnologia e Gestão*, vol. 5, no. 40, pp. 117–130, 2020. [Online]. Available: <https://revistas.ifpr.edu.br/index.php/mundietg/article/view/1215/1078>
- [20] R Core Team. (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Disponível em: <https://www.R-project.org>.
- [21] H. Mühleisen, M. Raasveldt, K. Müller, et al., "duckdb-r: DBI package for the DuckDB Database Management System," ver. 1.2.1, CRAN, Mar. 14, 2025. [Online]. Available: <https://r.duckdb.org/>
- [22] Wickham, H., Haselden, C. and Rivera, M. (2019). tidyverse: Easily install and load the 'tidyverse'. Disponível em: <https://cran.r-project.org/web/packages/tidyverse/index.html>.
- [23] Kuhn, M., Hastie, T., Kolaczek, J. and Hurd, J. (2020). tidymodels: Easily install and load the 'tidymodels' collection. Disponível em: <https://cran.r-project.org/web/packages/tidymodels/index.html>. Acesso: 26 jan. 2025.
- [24] Alpaydin, E. (2020). Introduction to Machine Learning. Cambridge, EUA: MIT Press.
- [25] Zhang, Y., et al. (2020). A Comprehensive Review of XGBoost: A Machine Learning Algorithm for Predicting Agricultural Prices. *Journal of Machine Learning Research*.
- [26] Yuan, S., et al. (2021). Efficient Predictive Models for Commodity Price Forecasting using LightGBM. *IEEE Transactions on Neural Networks and Learning Systems*.
- [27] Liu, J., et al. (2021). Support Vector Machines and Kernel Methods in Agricultural Price Prediction. *Journal of Computational and Graphical Statistics*.
- [28] Gao, Q., Yang, C. and Zhang, Z. (2021). Machine learning for predicting agricultural commodity prices: A case study on soybean. *Computers and Electronics in Agriculture*, 178, 105789.
- [29] Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer-Verlag.
- [30] Zhou, X., Liu, F. and Wang, H. (2020). Predicting agricultural commodity prices using machine learning algorithms: A comparative study. *Agricultural Systems*, 178, 102736.
- [31] Kuhn, M. (2019). The Cubist: A powerful decision tree algorithm for regression. *Journal of Statistical Software*, 90(3), 150–168.
- [32] Silva, J. M., Souza, R. P. and Costa, A. B. (2022). Aplicação de modelos de aprendizado de máquina para previsão de preços de soja. *Journal of Agricultural Science*, 60(2), 130–145.
- [33] Kunst, J. (2025). highcharter: A Wrapper for the 'Highcharts' Library. R package version 0.12.2.0. Disponível em: <https://github.com/jbkunst/highcharter>