JOÃO VITOR ANDRADE ALVES DE SOUZA

APLICAÇÃO DE MÉTODOS BASEADOS EM LÓGICA FUZZY E DE APRENDIZADO DE MÁQUINA PARA A PREVISÃO DE PREÇOS DE COMMODITIES

LAVRAS – MG

2026

JOÃO VITOR ANDRADE ALVES DE SOUZA

APLICAÇÃO DE MÉTODOS BASEADOS EM LÓGICA FUZZY E DE APRENDIZADO DE MÁQUINA PARA A PREVISÃO DE PREÇOS DE COMMODITIES

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Prof. Dr. Geraldo Magela da Cruz Pereira
Orientador

LAVRAS – MG
2026

Ficha Catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Souza, João Vitor Andrade Alves de.
Aplicação de métodos baseados em lógica fuzzy e de aprendizado de máquina para a previsão de preços de commodities. / João Vitor Andrade Alves de Souza. - 2026.
71 p. : il.

Orientador: Geraldo Magela da Cruz Pereira

Dissertação (Mestrado Acadêmico) - Universidade Federal de Lavras, 2026.
Bibliografia.

1. Previsão de séries temporais. 2. Séries temporais fuzzy 3. Aprendizado de máquina. 4. commodities agrícolas.. 3. Aprendizado de máquina. 4. Commodities agrícolas. I. Pereira, Geraldo Magela da Cruz. II. Universidade Federal de Lavras. III. Título.

JOÃO VITOR ANDRADE ALVES DE SOUZA

APLICAÇÃO DE MÉTODOS BASEADOS EM LÓGICA FUZZY E DE APRENDIZADO DE MÁQUINA PARA A PREVISÃO DE PREÇOS DE COMMODITIES

APPLICATION OF FUZZY LOGIC-BASED AND MACHINE LEARNING METHODS FOR COMMODITY PRICE FORECASTING

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 06 de Fevereiro de 2026.

Prof. Dr. Geraldo Magela da Cruz Pereira	UFLA
Prof. Dr. Sérgio Domingos Simão	UNIFAL
Profa. Dra. Gabi Nunes Silva	UNIR
Prof. Dr. Anderson Castro Soares de Oliveira	UFMT

gov.br
Documento assinado digitalmente
GERALDO MAGELA DA CRUZ PEREIRA
Data: 08/09/2026 13:53:12-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Geraldo Magela da Cruz Pereira
Orientador

LAVRAS – MG
2026

Dedico este trabalho a Deus, por Sua presença constante, força e sabedoria ao longo desta caminhada, e a todos aqueles que me apoiaram, incentivaram e contribuíram de alguma forma para a realização desta jornada. .

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, por me conceder força, sabedoria e perseverança ao longo desta trajetória.

Ao meu orientador, Professor Geraldo, pela orientação, dedicação, confiança e valiosos ensinamentos, fundamentais para a realização deste trabalho.

À minha mãe, Sandra, pelo amor incondicional, apoio constante e por sempre acreditar em meu potencial. Aos meus irmãos, Giovana, José Gabriel e Maria Júlia, pelo carinho, incentivo e por estarem presentes em todos os momentos.

Ao meu namorado, Andrey, pela compreensão, companheirismo e apoio durante esta caminhada acadêmica.

À minha amiga Yasmin, que esteve ao meu lado desde o início desta jornada, compartilhando desafios, conquistas e momentos inesquecíveis.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Universidade Federal de Lavras (UFLA), ao Instituto de Ciências Exatas e Tecnológicas (ICET), ao Departamento de Estatística e ao núcleo de estudos nSTATs pelo apoio institucional, estrutura e oportunidades oferecidas ao longo da minha formação.

A todos que, de alguma forma, contribuíram para a realização deste trabalho, os meus mais sinceros agradecimentos.

“Não é o mais forte que sobrevive, nem o mais inteligente, mas o que melhor se adapta às mudanças.”

(Charles Darwin)

RESUMO

O milho e a soja são duas das principais commodities agrícolas do Brasil, desempenhando um papel estratégico para a economia brasileira. A previsão de seus preços é de suma importância para apoiar a tomada de decisões no setor agroindustrial. Este estudo tem por objetivo avaliar a utilização de métodos *Fuzzy Time Series* (FTS) para a previsão dos preços dessas commodities, considerando abordagens univariadas e multivariadas. Esses modelos foram aplicados e comparados a algoritmos de *Machine Learning* (ML), como *Random Forest*, *XGBoost*, *LightGBM*, entre outros. O conjunto de dados utilizado no primeiro artigo compreende dados diários de preços de milho no período de 2007 a 2025, no estado de Minas Gerais, tendo três recortes, com início em 2007, 2014 e 2021. Já os dados utilizados no segundo artigo são referentes ao preço global da soja, de outras commodities e de seus derivados, coletados mensalmente no período de 1960 a 2025. Após o particionamento dos dados em treinamento e teste, os modelos FTS e os algoritmos de aprendizado de máquina foram treinados e avaliados no conjunto de teste. Foram utilizadas métricas de avaliação de desempenho, como Erro Percentual Absoluto Médio (Mean Absolute Percentage Error – MAPE), o Erro Absoluto Escalonado Médio (Mean Absolute Scaled Error – MASE), a Acurácia Direcional Média (Mean Directional Accuracy – MDA), entre outras medidas amplamente empregadas na avaliação de modelos de previsão. No caso univariado, os resultados indicam que o modelo Probabilistic Weighted FTS (PWFTS) apresentou melhor desempenho, com menores erros de previsão e maior acurácia direcional quando treinado com dados iniciados em 2014, e com horizonte de previsão de um dia. Em contraste, os modelos high order FTS (HOFTS) e Weighted HOFTS (WHOFTS) apresentaram desempenho levemente inferior ao observado para o PWFTS. Ainda assim, HOFTS e WHOFTS mantiveram desempenho consistente, embora com variações entre as diferentes estratégias de treinamento utilizadas. Para o caso multivariado, os resultados indicaram que os modelos baseados em lógica fuzzy, em especial o Weighted Multivariate FTS (WMVFTS) com 20% de dados de teste e 60 partições, obtiveram o melhor desempenho geral, apresentando os menores erros de previsão e elevada acurácia direcional. Conclui-se que os métodos Fuzzy Time Series constituem uma boa alternativa para a previsão de preços de commodities agrícolas, mostrando-se adequados tanto em contextos univariados quanto multivariados e em diferentes escalas temporais.

Palavras-chave: previsão de séries temporais; séries temporais fuzzy; aprendizado de máquina; commodities agrícolas

ABSTRACT

Corn and soybeans are two of the main agricultural commodities in Brazil, playing a strategic role in the Brazilian economy. Forecasting their prices is of utmost importance to support decision-making in the agro-industrial sector. This study aims to evaluate the use of *Fuzzy Time Series* (FTS) methods for forecasting the prices of these commodities, considering univariate and multivariate approaches. These models were applied and compared to *Machine Learning* (ML) algorithms, such as *Random Forest*, *XGBoost*, *LightGBM*, among others. The dataset used in the first article comprises daily corn price data for the period from 2007 to 2025, in the state of Minas Gerais, with three subsets, beginning in 2007, 2014, and 2021. The data used in the second article refer to the global price of soybeans, other commodities, and their derivatives, collected monthly from 1960 to 2025. After partitioning the data into training and test sets, the FTS models and machine learning algorithms were trained and evaluated on the test set. Performance evaluation metrics were used, such as Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), Mean Directional Accuracy (MDA), among other measures widely used in the evaluation of forecasting models. In the univariate case, the results indicate that the Probabilistic Weighted FTS (PWFTS) model presented the best performance, with lower forecasting errors and higher directional accuracy when trained with data beginning in 2014, and with a one-day forecasting horizon. In contrast, the high order FTS (HOFTS) and Weighted HOFTS (WHOFTS) models presented slightly lower performance than that observed for PWFTS. Nevertheless, HOFTS and WHOFTS maintained consistent performance, although with variations among the different training strategies used. For the multivariate case, the results indicated that the models based on fuzzy logic, especially the Weighted Multivariate FTS (WMVFTS) with 20% of test data and 60 partitions, obtained the best overall performance, presenting the lowest forecasting errors and high directional accuracy. It is concluded that Fuzzy Time Series methods constitute a good alternative for forecasting agricultural commodity prices, proving to be suitable both in univariate and multivariate contexts and at different time scales.

Keywords: time series forecasting; fuzzy time series; machine learning; agricultural commodities

INDICADORES DE IMPACTO

Esta pesquisa, desenvolvida no Programa de Pós-Graduação em Estatística e Experimentação Agropecuária da Universidade Federal de Lavras (UFLA), contribui para o desenvolvimento de métodos de previsão dos preços do milho e da soja por meio de sistemas fuzzy e técnicas de aprendizado de máquina, com potencial de aplicação no agronegócio. Os resultados obtidos podem subsidiar produtores rurais, cooperativas, empresas do setor e formuladores de políticas públicas na tomada de decisões relacionadas à produção, à comercialização e à gestão de riscos.

Os principais impactos da pesquisa são de natureza tecnológica, econômica e educacional, ao promover o uso de técnicas de aprendizado de máquina aplicadas na previsão de séries temporais e na análise de mercados agrícolas. Embora o estudo tenha como contexto inicial o estado de Minas Gerais, seus resultados podem ser estendidos a outras regiões produtoras do Brasil. Os principais grupos impactados são produtores rurais, cooperativas agrícolas, pesquisadores, estudantes e demais agentes do setor agropecuário.

A pesquisa não envolveu ações extensionistas diretas nem contou com a participação de parceiros externos à UFLA durante sua execução. Entretanto, apresenta potencial para a transferência do conhecimento à sociedade por meio da aplicação prática dos modelos desenvolvidos e da divulgação dos resultados científicos em publicações e eventos.

Os impactos do trabalho enquadram-se, principalmente, nas áreas temáticas de Educação (Área 4), Tecnologia e Produção (Área 7) e Trabalho (Área 8) da Política Nacional de Extensão.

Além disso, os resultados estão alinhados aos Objetivos de Desenvolvimento Sustentável (ODS) da Organização das Nações Unidas (ONU), especialmente ao ODS 2 (Fome Zero e Agricultura Sustentável), ODS 8 (Trabalho Decente e Crescimento Econômico) e ODS 9 (Indústria, Inovação e Infraestrutura), ao contribuir para o desenvolvimento de soluções tecnológicas voltadas ao setor agropecuário e para o fortalecimento da tomada de decisões baseada em dados.

IMPACT INDICATORS

This research, developed within the Graduate Program in Statistics and Agricultural Experimentation at the Federal University of Lavras (UFLA), contributes to the development of forecasting methods for corn and soybean prices through fuzzy systems and machine learning techniques, with potential applications in agribusiness. The results obtained can support farmers, cooperatives, companies in the sector, and public policymakers in decision-making related to production, commercialization, and risk management.

The main impacts of the research are technological, economic, and educational in nature, by promoting the use of machine learning techniques applied to time series forecasting and agricultural market analysis. Although the study has the state of Minas Gerais as its initial context, its results can be extended to other producing regions of Brazil. The main groups impacted are farmers, agricultural cooperatives, researchers, students, and other stakeholders in the agricultural sector.

The research did not involve direct extension activities nor the participation of external partners outside UFLA during its execution. However, it presents potential for the transfer of knowledge to society through the practical application of the developed models and the dissemination of scientific results in publications and conferences.

The impacts of this work fall mainly within the thematic areas of Education (Area 4), Technology and Production (Area 7), and Work (Area 8) of the National Extension Policy.

Furthermore, the results are aligned with the United Nations (UN) Sustainable Development Goals (SDGs), especially SDG 2 (Zero Hunger and Sustainable Agriculture), SDG 8 (Decent Work and Economic Growth), and SDG 9 (Industry, Innovation and Infrastructure), by contributing to the development of technological solutions aimed at the agricultural sector and to strengthening data-driven decision-making.

SUMÁRIO

PRIMEIRA PARTE - UM PANORAMA GERAL

1 INTRODUÇÃO

O milho, originário da Mesoamérica, é uma das culturas agrícolas mais importantes do mundo. Ele apresenta uma ampla variedade de usos, incluindo alimento básico, ração animal e matéria-prima para a produção de biocombustíveis e compostos industriais, sendo essencial para a economia e para a sustentabilidade agrícola, conforme destacam García-Lara & Serna-Saldivar (2019). Segundo Sharma (2021), a soja é uma leguminosa de grande importância econômica e nutricional, destacando-se principalmente pelo elevado teor e qualidade de suas proteínas.

Nos últimos anos, o Brasil consolidou-se como um dos maiores produtores mundiais de soja e milho, tendo relevância nos mercados globais dessas commodities, conforme analisado por Avileis & Mallory (2022). Os preços domésticos desses produtos são influenciados por variáveis como taxa de câmbio e as cotações internacionais, sendo o mercado da soja mais sensível a essas oscilações (Tessmann; Carrasco-Gutierrez; Lima, 2022). Já em Minas Gerais, o milho figura entre as principais culturas agrícolas, estando presente na maioria dos municípios do estado e destacando-se como uma das atividades mais relevantes para a economia local, tanto pela geração de renda quanto pela contribuição ao abastecimento regional (Carmo, 2016).

A previsão de preços de *commodities* auxilia na tomada de decisões no setor agrícola. Métodos de aprendizado de máquina, como Máquinas de Vetores de Suporte (*Support Vector Machines* – SVM), Regressão, Memória de Longo e Curto Prazo (*Long Short-Term Memory* – LSTM) e Vetores Autorregressivos (VAR), entre outros, podem ser utilizados para identificar padrões em séries temporais, compreender a dinâmica dos mercados globais e prever tendências e riscos (Gupta *et al.*, 2024). Além disso, o uso de técnicas como lógica fuzzy e cadeias de *Markov* tem contribuído para uma melhor previsão de preços agrícolas, aumentando a acurácia das estimativas e beneficiando agentes do mercado (Chen, 2002). Outro desafio refere-se à previsão dos custos de produção dessas culturas, frequentemente afetados pela alta volatilidade dos mercados. Nesse contexto, simulações de Monte Carlo e análises de séries temporais têm se mostrado ferramentas importantes para a previsão dos custos e receitas líquidas, contribuindo para decisões mais acertadas em cenários complexos (Amorim *et al.*, 2024).

As *commodities* milho e soja têm grande relevância, podendo seus preços influenciar de forma expressiva a cadeia de preços de outros produtos. De acordo com Šarac *et al.* (2023), o milho é um dos produtos centrais na composição de rações para aves, exercendo impacto direto

e estatisticamente relevante sobre o preço do frango de corte, ao contrário da soja, cujo efeito é considerado pouco representativo. Já Khanna (2010) destacam que o aumento no preço do milho afeta positivamente sua produtividade e área cultivada nos Estados Unidos, enquanto o preço da soja não apresenta influência consistente sobre seu rendimento. Sendo assim, a alta interdependência e volatilidade desses preços mostram a importância de previsões acuradas, as quais são úteis para o planejamento de mercado e formulação de políticas, sendo técnicas avançadas capazes de capturar padrões complexos e melhorar a precisão das projeções (Manogna; Dharmaji; Sarang, 2025).

A lógica fuzzy é conhecida como uma ferramenta eficiente para representar e inferir conhecimento em contextos de incerteza, promovendo análises e tomadas de decisões flexíveis em diferentes áreas (Zadeh, 1992). Fundamentada em relações de similaridade, essa abordagem aprimora o tratamento de incertezas e oferece avanços na definição de funções de similaridade para análises mais detalhadas (Ruspini, 1991). Assim, a lógica fuzzy pode ser estruturada em abordagens univariadas, focadas em uma única série de dados, ou multivariadas, que incorporam múltiplas variáveis para capturar interações complexas (Lee; Wang; Chen, 2007). Além disso, sua integração com técnicas de aprendizado de máquina tem se mostrado promissora na gestão de cadeias de suprimentos, permitindo decisões mais precisas e eficazes em ambientes industriais (Rodríguez; Gonzalez-Cava; Pérez, 2020). O aprendizado de máquina tem se destacado como uma ferramenta para automação de modelos analíticos e aumento de eficiência em diversas áreas, incluindo mercados digitais, saúde e finanças (Janiesch; Zschech; Heinrich, 2021; Jordan; Mitchell, 2015). Ademais, sua aplicação na previsão de séries temporais, combinando redes neurais e algoritmos baseados em árvores, é promissora em economia e finanças, na análise de dados financeiros de alta frequência (Masini; Medeiros; Mendes, 2023).

Diante desse panorama, a previsão dos preços do milho e da soja é importante para a estabilidade econômica e para o estabelecimento de estratégias no setor agroindustrial. O desenvolvimento de modelos contribui para a redução de riscos financeiros e de investimentos. Neste sentido, métodos baseados em Lógica Fuzzy despontam como uma alternativa para a modelagem de relações complexas e não lineares presentes em séries temporais, permitindo capturar incertezas e imprecisões típicas de dados econômicos e agrícolas. Diversos estudos demonstram o potencial dessa abordagem na previsão de variáveis agrícolas, como café e cana-de-açúcar, como preços e produtividade, em função de sua flexibilidade diante da variabilidade dos dados. Assim, a aplicação da Lógica Fuzzy na previsão do preço do milho na região do

Triângulo Mineiro justifica-se pela necessidade de instrumentos preditivos, capazes de apoiar a tomada de decisão por parte de produtores, cooperativas e agentes do mercado regional. De forma complementar, no contexto do mercado global da soja, a utilização de modelos de Lógica Fuzzy multivariada mostra-se uma alternativa interessante, uma vez que as séries são tratadas de maneira conjunta em um contexto multivariado, incluindo variáveis como preços de outras commodities e de seus derivados. A avaliação via métodos de Lógica Fuzzy pode fornecer uma melhor compreensão da dinâmica internacional do preço da soja.

Considerando o exposto, este estudo teve como objetivo geral avaliar a aplicabilidade de métodos de Séries Temporais Fuzzy (*Fuzzy Time Series Methods* – FTSM) na previsão dos preços do milho e da soja, considerando abordagens univariadas e multivariadas, diferentes estratégias de treinamento e horizontes de previsão. Especificamente, objetivou-se: i) Avaliar o desempenho de diferentes estratégias de particionamento do universo de discurso e do número de partições na modelagem de Séries Temporais Fuzzy; ii) Investigar o efeito de diferentes períodos de treinamento e de horizontes de previsão sobre o poder preditivo dos modelos univariados aplicados à previsão dos preços de milho em Minas Gerais; iii) Avaliar a contribuição da inclusão de outras commodities e seus derivados na previsão dos preços internacionais da soja por meio de métodos de Séries Temporais Fuzzy; e iv) Comparar o desempenho preditivo dos métodos de Séries Temporais Fuzzy com algoritmos de aprendizado de máquina, incluindo XGBoost, LightGBM, *Random Forest*, *Support Vector Regression* e *Prophet*.

2 REFERENCIAL TEÓRICO

Neste capítulo, são apresentados os fundamentos teóricos que sustentam o desenvolvimento deste trabalho, abordando conceitos relacionados a séries temporais, modelos baseados em lógica fuzzy, algoritmos de aprendizado de máquina e métricas de avaliação utilizadas na análise preditiva.

2.1 Séries Temporais

As séries temporais consistem em sequências de dados coletados em intervalos regulares ao longo do tempo, sendo utilizadas para analisar o comportamento de fenômenos com base em sua evolução temporal. A compreensão da natureza do fenômeno é fundamental para descrever seu comportamento e, posteriormente, realizar estimativas e análises que permitam identificar fatores influentes e possíveis relações de causa e efeito entre diferentes variáveis (Latorre; Cardoso, 2001).

Segundo Fu (2011), essas séries representam uma forma estruturada e complexa de dados organizados cronologicamente, cuja análise tem estimulado avanços na mineração de dados. Já Liu *et al.* (2021) destacam que a previsão baseada em séries temporais busca identificar padrões nos dados organizados ao longo do tempo, enfrentando desafios como a complexity dos dados e a baixa capacidade de generalização dos modelos preditivos.

A análise de séries temporais é amplamente utilizada em diferentes áreas do conhecimento, como economia e meteorologia. Na economia, essas séries podem representar preços e ações observadas ao longo do tempo, permitindo avaliar tendências e comportamento dos mercados. Na meteorologia, são utilizadas para previsão de precipitação, monitoramento hidrológico e análise climática (Box *et al.*, 2015). Araujo, Sousa & Santos (2008) realizaram uma análise da exportação de melão considerando séries temporais do volume exportado e sua relação com preços internos e taxa de câmbio. Em outro estudo, Isensee, Pinheiro & Detzel (2021) analisaram vazões de rios e identificaram tendências relevantes para avaliação da segurança operacional de estruturas hidráulicas.

Os métodos estatísticos para previsão de séries temporais são importantes para reduzir incertezas na tomada de decisão. Esses métodos normalmente assumem determinadas propriedades estocásticas da série; entretanto, quando tais pressupostos não são atendidos, transforma-

ções como diferenciação podem ser necessárias para alcançar a estacionariedade (Sousa *et al.*, 2012).

Por outro lado, métodos baseados em lógica fuzzy e diversos algoritmos de aprendizado de máquina constituem abordagens não paramétricas, pois não exigem pressupostos rígidos sobre a distribuição dos dados. Métodos fuzzy têm apresentado resultados promissores na modelagem de sistemas complexos e incertos (Zhang *et al.*, 2022). Além disso, abordagens híbridas que combinam lógica fuzzy e aprendizado de máquina apresentam ganhos em desempenho preditivo e interpretabilidade (Sherly; Christo; Elizabeth, 2025).

2.1.1 Decomposição de Séries Temporais

Séries temporais são compostas por elementos responsáveis pela formação dos padrões observados ao longo do tempo. Esses componentes incluem tendência, sazonalidade, ciclos e erro aleatório. A decomposição permite separar esses componentes para compreender melhor seus padrões (Dudek, 2023).

A tendência representa a direção de evolução da série ao longo do tempo, caracterizada por um comportamento crescente, decrescente ou de estabilização. A identificação de tendências pode ser realizada por modelos de regressão polinomial, suavização (como médias móveis e LOESS) ou outras técnicas apropriadas à estrutura dos dados (Latorre; Cardoso, 2001). Segundo Li *et al.* (2015), tendências de curto e longo prazo podem coexistir e sua separação melhora tarefas de previsão e detecção de padrões.

A sazonalidade representa comportamentos que se repetem em intervalos regulares e estão associados a fatores periódicos, como meses, trimestres e dias da semana. Sua modelagem evita interpretações equivocadas de tendências estruturais (Hannan; Terrell; Tuckwell, 1970).

Os ciclos correspondem a oscilações de longo prazo ou aos desvios em tempo de rota de tendência, não associadas diretamente à sazonalidade e auxiliam na compreensão das relações entre diferentes séries econômicas (Cubadda, 1999).

O componente aleatório (ou ruído) representa flutuações aleatórias presentes nas observações, dificultando a identificação de padrões relevantes (Wunderlich; Sklar, 2023). Segundo Bagchi, Characiejus & Dette (2018), métodos específicos podem ser empregados para verificar se a série ou os resíduos do modelo apresentam comportamento compatível com ruído branco.

A análise dos resíduos constitui uma etapa importante na avaliação dos modelos, permitindo verificar hipóteses relacionadas à independência e estabilidade da variância (Velasco, 2022).

Modelos aditivos são amplamente utilizados por fornecerem estimativas consistentes da variabilidade temporal (Dozie; Ijomah, 2020). O modelo aditivo é definido por:

$$Y_t = T_t + S_t + E_t \tag{2.1}$$

em que: Y_t representa o valor observado no instante t; T_t representa a tendência; S_t corresponde à sazonalidade; e E_t representa o componente aleatório.

De forma alternativa, o modelo multiplicativo pode ser representado por:

$$Y_t = T_t \times S_t \times E_t \tag{2.2}$$

em que: Y_t, T_t, S_t e E_t representam os mesmos componentes descritos anteriormente no modelo aditivo (Dozie; Ijomah, 2020).

2.1.2 Modelo de Lógica Fuzzy para Previsões

A lógica fuzzy é uma extensão dos sistemas lógicos clássicos que oferece uma estrutura conceitual capaz de lidar com a representação do conhecimento em cenários marcados por incertezas e imprecisões, utilizando graus de pertinência e modos de raciocínio aproximado, em vez de exato (Zadeh, 1989b).

Os modelos de Fuzzy Time Series (FTS) constituem uma abordagem para a previsão de séries temporais na qual os valores observados são representados por conjuntos fuzzy, permitindo a modelagem das relações temporais entre os estados da série. Conforme apresentado por Lucas *et al.* (2021), o processo de modelagem pode ser organizado em duas etapas principais: treinamento e previsão. Na etapa de treinamento, inicialmente é definido o Universo de Discurso U a partir dos valores observados da série temporal. Para uma série temporal univariada $Y(t) \in \mathbb{R}$, o Universo de Discurso é definido pelos limites mínimo e máximo observados acrescidos de duas constantes positivas D_1 e D_2, de acordo com:

$$U = [\min(Y) - D_1, \max(Y) + D_2].$$

A partir do Universo de Discurso, são definidos os conjuntos fuzzy que representam os valores linguísticos da série. Cada conjunto fuzzy A_j está associado a uma função de pertinência μ_{A_j}, responsável por indicar o grau com que uma determinada observação pertence ao respectivo conjunto fuzzy. O Universo de Discurso pode ser particionado por diferentes procedimentos, sendo o particionamento em intervalos de mesmo tamanho uma das alternativas apresentadas por Lucas *et al.* (2021). Após a definição dos conjuntos fuzzy, realiza-se a fuzzificação dos valores observados. Nesse procedimento, a série temporal fuzzy é definida como $F(t)$, sendo uma observação no instante t dada por $F(t)$, cujos valores são conjuntos fuzzy sobre U. A representação fuzzificada de uma observação $y(t)$ no instante t corresponde ao conjunto fuzzy A_j para o qual o grau de pertinência $\mu_{A_j}(y(t))$ é máximo, expressa por:

$$F(t) = A_j \quad \text{tal que} \quad j = \arg\max_k \mu_{A_k}(y(t)),$$

em que: $y(t)$ representa a observação da série temporal no instante t e μ_{A_j} corresponde à função de pertinência associada ao conjunto fuzzy A_j (Lucas *et al.*, 2021; Song; Chissom, 1993b). Lucas *et al.* (2021) apresentam diferentes estratégias para a fuzzificação, entre elas o método baseado no maior grau de pertinência, a abordagem holística e o método α-cut. Após a fuzzificação, são extraídas as relações temporais existentes entre os estados fuzzy da série, dando origem às Relações Lógicas Fuzzy (*Fuzzy Logical Relationships* – FLR). Uma relação lógica fuzzy representa a transição entre dois estados fuzzy sucessivos e pode ser expressa por:

$$A_i \rightarrow A_j,$$

em que: A_i representa o estado fuzzy no instante $t-1$ e A_j representa o estado fuzzy no instante t. As relações lógicas fuzzy podem ser agrupadas de acordo com seus antecedentes. Chen (1996) propôs os Grupos de Relações Lógicas Fuzzy (*Fuzzy Logical Relationship Groups* – FLRG), nos quais relações que apresentam o mesmo antecedente são reunidas em um único grupo. Assim, relações como:

$$A_1 \rightarrow A_1, \quad A_1 \rightarrow A_2, \quad A_1 \rightarrow A_3$$

podem ser agrupadas na forma

$$A_1 \rightarrow A_1, A_2, A_3.$$

Outra abordagem apresentada na literatura é a representação das relações fuzzy por meio de matrizes de relações. Song & Chissom (1993b), Song & Chissom (1993a) utilizaram uma matriz de relações fuzzy $R(t,t-1)$ e o operador de composição Max-Min para representar a relação entre os estados fuzzy da série. Nessa formulação, a equação relacional fuzzy é expressa por

$$F(t) = F(t-1) \circ R(t,t-1),$$

em que: $F(t)$ representa o estado fuzzy da série no instante t, $R(t,t-1)$ representa a matriz de relações fuzzy e $\circ$ corresponde ao operador de composição Max-Min (Song; Chissom, 1993b; Song; Chissom, 1993a).

2.1.3 Lógica Fuzzy Multivariada

Segundo Wu *et al.* (2021), a previsão em sistemas de lógica fuzzy multivariada destaca-se por sua capacidade de trabalhar com múltiplas variáveis e relações complexas, utilizando diferentes entradas, como indicadores financeiros ou variáveis climáticas, para construir regras fuzzy que melhor representam o comportamento dos dados. A combinação de técnicas como agrupamento e regressão melhora a construção dessas relações e aumenta a precisão das previsões. Já Yazdanbakhsh & Dick (2017) mostram que modelos fuzzy mais avançados conseguem trabalhar com várias entradas e saídas simultaneamente, mantendo elevada precisão.

Perveen, Rizwan & Goel (2019) destacam que sistemas neuro-fuzzy, como o ANFIS, conseguem integrar múltiplos fatores, por exemplo, condições climáticas, e produzir previsões mais confiáveis do que métodos tradicionais. Assim, a lógica fuzzy multivariada, quando combinada com técnicas de aprendizado de máquina, apresenta-se como uma abordagem versátil para problemas de previsão.

A aplicação da lógica fuzzy em séries temporais multivariadas tem sido bastante explorada na literatura recente. Modelos baseados em previsões fuzzy têm incorporado técnicas como clusterização, otimização e aprendizado híbrido, demonstrando melhorias significativas em métricas de desempenho (Iqbal *et al.*, 2020). Exemplos de abordagens que integram métodos de ponderação e estratégias híbridas têm mostrado ganhos na modelagem de dados complexos e não lineares. Assim, estudos recentes têm combinado sistemas fuzzy com técnicas de aprendizado profundo, evidenciando que a inclusão de múltiplas variáveis e estruturas híbridas

permite capturar de forma mais eficiente relações complexas e dinâmicas presentes nos dados, aprimorando o desempenho preditivo (Wang; Shao; Jumahong, 2023).

Conforme Iqbal *et al.* (2020) e Wang, Shao & Jumahong (2023), os modelos de lógica fuzzy multivariada realizam previsões a partir da transformação simultânea de múltiplas variáveis de entrada em estados fuzzy e da aplicação de regras que capturam as interações entre essas variáveis. Considerando um vetor de entrada

$$X_t = (x_{1,t}, x_{2,t}, \ldots, x_{m,t})$$

torna-se possível modelar relações não lineares e dependências complexas entre os dados. A formulação geral do modelo pode ser expressa por

$$\tilde{Y}_t = F(x_{1,t-1}, x_{2,t-1}, \ldots, x_{m,t-1})$$

ou, de forma equivalente, considerando múltiplas regras fuzzy,

$$\tilde{Y}_t = \bigcup_{i=1}^{N} w_i \circ R_i$$

em que o grau de ativação de cada regra é dado por

$$w_i = \prod_{j=1}^{m} \mu_{A_{j,i}}(x_{j,t-1})$$

ou, alternativamente,

$$w_i = \min\left(\mu_{A_{1,i}}(x_{1,t-1}), \ldots, \mu_{A_{m,i}}(x_{m,t-1})\right)$$

em que: $\mu_{A_{j,i}}(x_{j,t-1})$ representa o grau de pertinência da j-ésima variável de entrada ao conjunto fuzzy $A_{j,i}$ na i-ésima regra, enquanto R_i representa a relação fuzzy multivariada definida por regras do tipo SE-ENTÃO. A operação $\circ$ corresponde à composição fuzzy, usualmente baseada em operadores do tipo produto ou mínimo, que combinam os graus de pertinência das múltiplas entradas para produzir a saída fuzzy $\tilde{Y}_t$ (Iqbal *et al.*, 2020; Wang; Shao; Jumahong, 2023).

Após a etapa de inferência, a saída do modelo pode ser obtida por meio de um sistema do tipo Takagi-Sugeno, em que cada regra possui um consequente funcional. Nesse caso, a

saída numérica é dada por

$$\hat{y}_t = \frac{\sum_{i=1}^{N} w_i f_i(x_t)}{\sum_{i=1}^{N} w_i}$$

em que:

$$f_i(x_t) = a_{i1}x_{1,t} + a_{i2}x_{2,t} + \cdots + a_{im}x_{m,t} + b_i$$

em que: a_{ij} e b_i são parâmetros associados à i-ésima regra, estimados a partir dos dados. Esse tipo de formulação permite interpretar o modelo como uma combinação ponderada de modelos locais (Wang; Shao; Jumahong, 2023).

A definição dos conjuntos fuzzy e da estrutura do modelo constitui um conjunto de hiperparâmetros que impactam diretamente a avaliação e o desempenho do modelo. Assim, parâmetros associados às funções de pertinência, como centro e dispersão no caso de funções gaussianas, também devem ser ajustados, podendo ser definidos por métodos de clusterização ou técnicas de otimização (Iqbal *et al.*, 2020).

2.2 Aprendizado de Máquina

Nas últimas décadas, tem-se observado um período de intensas transformações tecnológicas, frequentemente caracterizado como uma nova revolução industrial. Esse processo é impulsionado pelo avanço das tecnologias digitais e, principalmente, pela inteligência artificial (IA), que se destaca como uma das principais forças motrizes dessa transformação. As máquinas evoluíram rapidamente, não apenas assumindo tarefas manuais, mas também desempenhando atividades cognitivas que, por muito tempo, foram consideradas exclusivas da inteligência humana (Ludermir, 2021).

De acordo com Janiesch, Zschech & Heinrich (2021), o aprendizado de máquina refere-se à capacidade de sistemas computacionais aprenderem a partir de dados específicos de um problema, auxiliando na construção de modelos analíticos e na execução de tarefas. Complementarmente, Jordan & Mitchell (2015) destacam que o aprendizado de máquina possibilita o desenvolvimento de sistemas capazes de aprimorar seu desempenho à medida que novas experiências e informações são incorporadas. Essa área é fortemente influenciada por duas disciplinas fundamentais: a estatística e a ciência da computação. Seu crescimento tem sido impulsionado pelo aumento da disponibilidade de dados em ambientes digitais e pela redução dos custos computacionais. Como consequência, o aprendizado de máquina passou a ser aplicado em diversos

campos do conhecimento, tornando-se uma importante ferramenta para a tomada de decisões baseada em dados.

O aprendizado de máquina constitui uma abordagem empírica amplamente utilizada para resolver problemas de regressão e classificação, tanto em contextos supervisionados quanto não supervisionados, especialmente em sistemas caracterizados por relações não lineares. Essa área engloba um amplo conjunto de técnicas computacionais voltadas à extração de padrões e à geração de conhecimento a partir dos dados. Entre os algoritmos mais empregados destacam-se as redes neurais artificiais, as máquinas de vetor de suporte (*Support Vector Machines* – SVM), as árvores de decisão e as florestas aleatórias (*Random Forest*) (Lary *et al.*, 2016).

No aprendizado supervisionado, os modelos são treinados utilizando um conjunto de dados composto por variáveis de entrada e respectivas saídas conhecidas. Durante esse processo, o algoritmo aprende a relação existente entre os preditores e a variável resposta, tornando-se posteriormente capaz de realizar previsões para novos dados não observados. Esse tipo de abordagem pode ser aplicado tanto a problemas envolvendo variáveis categóricas quanto a problemas de natureza contínua (Sidey-Gibbons; Sidey-Gibbons, 2019).

Por outro lado, o aprendizado não supervisionado busca identificar estruturas, padrões e relações presentes nos dados sem a utilização de variáveis-resposta previamente conhecidas. Segundo Hinton & Sejnowski (2017), essa abordagem procura compreender como os sistemas representam padrões de entrada e capturam a estrutura subjacente dos dados observados. Dessa forma, os algoritmos são responsáveis por identificar automaticamente agrupamentos, associações ou representações latentes, sem a necessidade de informações rotuladas. Os métodos apresentados nas seções subsequentes são classificados como técnicas de aprendizado supervisionado.

2.2.1 *Floresta Aleatória*

Uma árvore de decisão é um diagrama dos possíveis resultados de uma série de escolhas associadas, utilizado para representar problemas e auxiliar nas tomadas de decisão. Ela é estruturada de modo semelhante a uma árvore, em que cada ramo representa uma escolha e cada nó terminal (folha) mostra o resultado dessa escolha. É uma metodologia útil quando há várias decisões a serem tomadas em sequência, e quando cada decisão pode levar a diferentes resultados (Dobra, 2009).

Este método se baseia no particionamento recursivo dos dados em regiões menores, facilitando a modelagem e a previsão. Ele é amplamente utilizado para tarefas de classificação e regressão. No entanto, pode ser sensível a pequenas variações nos dados, o que pode comprometer sua estabilidade (Talekar; Agrawal, 2020).

Visando superar essa limitação, foi desenvolvido o algoritmo RF, que se baseia no treinamento de várias árvores de decisão em paralelo. Neste algoritmo, cada árvore é criada a partir de valores de um vetor aleatório amostrado de forma independente e com a mesma distribuição probabilística para todas as árvores de decisão. Após o treinamento, os resultados obtidos para as árvores são combinados com o propósito de predição. Esse procedimento alcança maior precisão e robustez, melhorando a estabilidade das previsões. Tais características tornam a floresta aleatória uma das abordagens mais populares em Ciência de Dados, com aplicações em mineração de dados, aprendizado de máquinas e reconhecimento de padrões (Talekar; Agrawal, 2020). Ao utilizar uma seleção aleatória de características para dividir cada nó, a técnica consegue fornecer taxas de erro comparáveis às do *AdaBoost*. À medida que o número de árvores na floresta aumenta, o erro de generalização tende a diminuir até se estabilizar. A precisão do modelo depende tanto da força individual das árvores quanto da correlação entre elas (Breiman, 2001).

Segundo Schonlau & Zou (2020), o algoritmo de floresta aleatória consiste em construir múltiplas árvores de decisão geradas a partir de amostras *bootstrap* e, em cada nó, selecionar aleatoriamente um subconjunto de variáveis para otimizar o critério de divisão (entropia para classificação ou MSE para regressão), reduzindo o *overfitting* por meio do processo de *bagging*. O algoritmo a seguir descreve em detalhes o funcionamento da floresta aleatória.

Segundo Breiman (2001), a floresta aleatória para regressão é construída a partir de múltiplos conjuntos *bootstrap*, representados por:

$$D_b = \{(x_i, y_i)\}_{i=1}^{n}, \quad D_b \sim \text{amostragem com reposição.}$$

em que: x_i representa o vetor de atributos da i-ésima observação, y_i é o valor da variável resposta correspondente, n é o número total de observações no conjunto de dados original, e D_b indica o subconjunto criado para a b-ésima árvore, obtido via amostragem com reposição.

Cada árvore utiliza apenas um subconjunto aleatório de variáveis em cada divisão, definido como:

$$M \subseteq \{1, \ldots, p\}$$

em que: p representa o número total de variáveis explicativas e M é o conjunto de variáveis selecionadas aleatoriamente para avaliação da divisão em um nó específico.

A construção da floresta aleatória depende de um conjunto de hiperparâmetros, dentre os quais destacam-se: o número de árvores B (*ntree*), o número de variáveis consideradas em cada divisão m (*mtry*), a profundidade máxima das árvores (*max depth*), o número mínimo de observações em um nó terminal (*nodesize*) e o tamanho mínimo para divisão de nós (*min samples split*). Esses parâmetros influenciam diretamente o viés e a variância do modelo, bem como sua capacidade de generalização (Breiman, 2001; Hastie; Tibshirani; Friedman, 2009).

A melhor divisão é escolhida ao minimizar a soma dos erros quadráticos nos nós resultantes:

$$\theta^* = \arg\min_{\theta} \left[\sum_{x_i \in R_1(\theta)} (y_i - \bar{y}_{R_1})^2 + \sum_{x_i \in R_2(\theta)} (y_i - \bar{y}_{R_2})^2 \right].$$

em que: θ representa o parâmetro de divisão (como a variável e o ponto de corte), $R_1(\theta)$ e $R_2(\theta)$ correspondem aos dois nós resultantes da divisão, e $\bar{y}_{R_1}$ e $\bar{y}_{R_2}$ são as médias da variável resposta em cada nó.

A predição individual de cada árvore corresponde à média dos valores da variável resposta no nó terminal:

$$\hat{f}_b(x) = \frac{1}{|R_b(x)|} \sum_{i \in R_b(x)} y_i.$$

A estimativa final da floresta aleatória é a média das predições de todas as árvores:

$$\hat{f}_{RF}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x).$$

O modelo utiliza o erro *out-of-bag* (OOB) como mecanismo interno de validação:

$$\hat{f}_{OOB}(x_i) = \frac{1}{|B_i|} \sum_{b \in B_i} \hat{f}_b(x_i).$$

A importância das variáveis é estimada por:

$$\text{Imp}(X_j) = \frac{1}{B}\sum_{b=1}^{B}\left(MSE_{OOB,b}^{(j)} - MSE_{OOB,b}\right).$$

2.2.2 *Support Vector Regression*

Segundo Vapnik, Golowich & Smola (1996), a Regressão por Vetores de Suporte (*Support Vector Regression* – SVR) é uma técnica de aprendizado supervisionado derivada das SVM, projetada para problemas de regressão. Sua principal característica é a capacidade de lidar com dados não lineares por meio do uso de funções *kernel*, mantendo alta robustez frente a *outliers* e boa generalização, mesmo em espaços de alta dimensionalidade. Conforme Awad & Khanna (2015), esse método utiliza uma função de perda simétrica com margem de tolerância (ε), que ignora erros menores que ε e penaliza apenas desvios superiores. Para Huang *et al.* (2024), sua aplicação tem sido frequente em análises de séries temporais, sendo considerada uma alternativa eficaz à regressão linear em cenários complexos.

Estudos recentes têm aprofundado o uso da SVR em diferentes contextos. Dash *et al.* (2023) propuseram um modelo de SVR ajustado para previsão de ações, no qual a escolha da função *kernel* e seus hiperparâmetros é otimizada por busca em grade. Nesse caso, os vetores de suporte definem a função de predição ao minimizar apenas os erros mais relevantes, resultando em alta precisão e baixo custo computacional. Parbat & Chakraborty (2020) aplicaram a SVR com *kernel* RBF para prever casos de COVID-19 na Índia, obtendo boa capacidade de generalização na previsão de mortes, recuperações e casos acumulados. Já Du, Chen & Cong (2021) utilizaram algoritmo genético para otimização dos hiperparâmetros da SVR, incluindo o parâmetro de penalidade, a função *kernel* e a margem de tolerância, em um modelo para previsão de preços de imóveis. Os resultados indicaram maior precisão e velocidade de convergência, demonstrando a eficácia da abordagem em mercados complexos.

A base do método das Máquinas de Vetores de Suporte foi proposta por *Vladimir Vapnik* e colaboradores na década de 1990. Segundo Drucker *et al.* (1997), a função de regressão não linear no SVM pode ser expressa pela formulação dual:

$$f(u_j) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(u_i, u_j) + b$$

em que: o parâmetro C controla a penalização atribuída aos erros de ajuste, α_i e α_i^* representam os multiplicadores de *Lagrange*, u_i representa o i-ésimo vetor de entrada (amostra) do conjunto de dados, u_j corresponde ao vetor de entrada para o qual se deseja obter a predição e b é o termo de intercepto (*bias*) do modelo.

Sujeita às restrições:

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \quad 0 \leq \alpha_i, \alpha_i^* \leq C$$

O *kernel* é definido como:

$$K(u_i, u_j) = \langle \Phi(u_i), \Phi(u_j) \rangle$$

em que: $\Phi(\cdot)$ representa o mapeamento para um espaço de maior dimensionalidade.

As funções *kernel* mais utilizadas são:

$$K(u_i, u_j) = u_i \cdot u_j \qquad \text{(Linear)}$$

$$K(u_i, u_j) = (\gamma u_i \cdot u_j + c)^d \qquad \text{(Polinomial)}$$

$$K(u_i, u_j) = \tanh(\gamma u_i \cdot u_j + c) \qquad \text{(Sigmoide)}$$

$$K(u_i, u_j) = \exp\left[-\gamma \|u_i - u_j\|^2\right] \qquad \text{(Gaussiano/RBF)}$$

com: γ é o termo que define a influência de cada amostra no cálculo do *kernel* em modelos RBF e polinomiais, d determina o grau do *kernel* polinomial e c atua como termo de deslocamento.

A formulação primal do SVM é dada por:

$$R = C\sum_{i=1}^{n}(\xi_i + \xi_i^*) + \frac{1}{2}\|w\|^2,$$

em que: ξ_i e ξ_i^* representam as folgas associadas aos pontos que excedem a margem tolerada, e $\|w\|^2$ expressa a complexidade do modelo, sendo penalizada para evitar o sobreajuste.

2.2.3 *XGBoost*

O *XGBoost* (*Extreme Gradient Boosting*) é um algoritmo de aprendizado de máquina inspirado em conjuntos de árvores de decisão, conhecido por sua alta precisão, velocidade e escalabilidade, sendo bem eficaz na análise de grandes volumes de dados e variables complexas, conforme destacado por Chen & Guestrin (2016). Sua eficiência e desempenho o tornaram uma ferramenta popular entre cientistas de dados em diversas competições e aplicações reais, por sua habilidade de oferecer bons resultados em cenários com grandes bases de dados (Cai; Rodavia, 2023). Este algoritmo tem sido utilizado em contextos financeiros, como na modelagem preditiva do mercado de ações, onde se mostrou capaz de processar mais de 16 anos de dados históricos diários e integrar diversas fontes de informação para ajudar nas tomadas de decisões estratégicas em ambientes voláteis, como apresentado por Hussain *et al.* (2023).

O modelo *XGBoost* possui diversas aplicações. Uma delas é a previsão do tempo de viagem, identificando as melhores rotas e contribuindo para a redução do congestionamento de tráfego. No estudo apresentado por Chen & Fan (2021), foi utilizada essa modelagem para investigar dados históricos e gerar previsões baseadas em padrões temporais. Outra aplicação está na previsão de resultados esportivos. No estudo de Ouyang *et al.* (2024), foram feitas previsões de partidas de basquete da NBA, com base em dados coletados de 2021 a 2023. O autor utilizou modelos preditivos em tempo real, com destaque para o *XGBoost* na fase de aprendizado. O modelo demonstrou alta acurácia na previsão dos resultados, evidenciando sua eficiência.

De acordo com Amjad *et al.* (2022), a função objetivo do *XGBoost* é definida pela soma da perda entre valores previstos e observados, acrescida de um termo de regularização que controla a complexidade do modelo. Essa função geral pode ser expressa como:

$$O = \sum_{i=1}^{n} L(y_i, F(x_i)) + \sum_{k=1}^{t} R(f_k) + C$$

em que: $L(y_i, F(x_i))$ representa a perda associada a cada previsão, $R(f_k)$ corresponde ao custo da árvore adicionada em cada iteração e C é uma constante que pode ser omitida sem afetar o processo de otimização.

O termo de regularização pode ser formulado como:

$$R(f_k) = \alpha H + \frac{1}{2}\eta \sum_{j=1}^{H} \omega_j^2$$

em que: α controla a complexidade estrutural das folhas, H é o número de folhas da árvore, η corresponde ao coeficiente de penalização e ω_j representa o valor preditivo associado a cada nó terminal.

Considerando a aproximação de Taylor de segunda ordem, a função objetivo pode ser reescrita como:

$$O = \sum_{i=1}^{n} \left[g_i \omega_{q(x_i)} + \frac{1}{2} h_i \omega_{q(x_i)}^2 \right] + \alpha H + \frac{1}{2}\eta \sum_{j=1}^{H} \omega_j^2$$

em que: g_i e h_i são, respectivamente, a primeira e a segunda derivada da função de perda, e $q(x_i)$ define a folha correspondente a cada observação.

Agrupando os termos por folha, chega-se à seguinte formulação:

$$O = \sum_{j=1}^{T} \left[P_j \omega_j + \frac{1}{2}(Q_j + \eta)\omega_j^2 \right] + \alpha H$$

em que:

$$P_j = \sum_{i \in U_j} g_i, \qquad Q_j = \sum_{i \in U_j} h_i$$

e U_j representa o conjunto de amostras pertencentes à folha j.

2.2.4 *Prophet*

O modelo *Prophet* é utilizado para a previsão de séries temporais e foi desenvolvido pelo *Facebook*, sendo inspirado na decomposição aditiva da série em componentes de tendência, sazonalidade e efeitos irregulares, sendo indicado para dados que apresentam padrões sazonais complexos e flutuações não lineares. Segundo Arslan (2022), o *Prophet* preserva a sazonalidade da série original e pode haver junções com outros modelos para maximizar a precisão preditiva. Para Navratil & Kolkova (2019), o modelo é eficaz na identificação e interpretação de tendências sazonais nas áreas econômicas e empresariais.

O modelo auxilia na análise de séries temporais ao permitir a decomposição clara dos dados, mesmo quando a série é curta, esparsa ou irregular. De acordo com Atamimi, Witanti &

Abdillah (2025), essa classificação torna o *Prophet* útil em contextos empresariais, pois melhora a precisão das previsões por meio do ajuste de seus componentes internos. De forma complementar, Sherly, Christo & Elizabeth (2025) destacam que o *Prophet* é eficiente na modelagem de padrões não lineares e sazonais, contribuindo para previsões mais eficazes e interpretáveis, inclusive quando integrado a outras abordagens.

Estudos recentes mostram que o modelo *Prophet* é eficaz para a previsão de preços de *commodities*. Na agricultura, o modelo apresentou bons resultados na previsão dos preços do trigo ao incorporar fatores climáticos e macroeconômicos (Kasianova; Popov; Dehtiarov, 2025) e também se mostrou superior a métodos tradicionais na previsão dos preços do tomate, oferecendo estimativas mais precisas para auxiliar a tomada de decisão (Kostaridou *et al.*, 2024). No contexto do abastecimento de alimentos, o *Prophet* demonstrou robustez e alta adaptabilidade na previsão de preços de itens de consumo de massa, reforçando sua versatilidade em diferentes cenários de mercado (Anggraini; Nabila; Fajaryanti, 2026).

Do ponto de vista matemático, o modelo *Prophet* é formulado como um modelo aditivo decomponível, no qual a série temporal observada é representada como a soma de diferentes componentes estruturais. Essa formulação é expressa pela Equação 2.3:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t, \tag{2.3}$$

em que: $y(t)$ representa o valor observado da série temporal no instante t; $g(t)$ corresponde à função de tendência, podendo ser especificada como crescimento linear por partes ou crescimento logístico com pontos de mudança; $s(t)$ é o componente sazonal, modelado por uma expansão em série de *Fourier*, dada por:

$$s(t) = \sum_{k=1}^{K} \left[a_k \cos\left(\frac{2\pi kt}{P}\right) + b_k \sin\left(\frac{2\pi kt}{P}\right) \right],$$

com: P o período da sazonalidade e K o número de termos considerados, enquanto a_k e b_k são coeficientes a serem estimados; e o termo $h(t)$ incorpora os efeitos de feriados ou eventos especiais,

$$h(t) = \sum_{l=1}^{L} \kappa_l D_l(t)$$

sendo: $D_l(t)$ variáveis indicadoras associadas a cada evento e κ_l seus respectivos impactos. Por fim, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ representa o termo de erro aleatório, assumido como ruído branco. Segundo Taylor & Letham (2018), essa estrutura aditiva enquadra o *Prophet* na classe dos modelos aditivos generalizados, proporcionando interpretabilidade e flexibilidade na modelagem de múltiplas sazonalidades e mudanças estruturais na tendência.

2.2.5 *LightGBM Regression*

Segundo Yan *et al.* (2021), o *Light Gradient Boosting Machine* (*LightGBM*) é um algoritmo de aprendizado de máquinas inspirado no *gradient boosting*, desenvolvido para aumentar a eficiência computacional, a estabilidade e a precisão em tarefas preditivas. Utiliza a estratégia de crescimento de árvore folha a folha (*leaf-wise*), o que possibilita identificar a folha com maior ganho, realizando a divisão apenas para esta folha. Esta estratégia permite lidar com grandes volumes de dados de forma rápida e estável. Para os autores Chen & Seo (2023), sua arquitetura o torna eficaz em aplicações de predição, apresentando desempenho superior em relação a métodos tradicionais, como k-vizinhos mais próximos, árvores de decisão e RF.

O *LightGBM* tem sido utilizado em diferentes áreas da ciência, destacando-se em aplicações preditivas que demandam precisão e robustez. No mercado de criptomoedas, apresentou desempenho superior a outros modelos ao incorporar indicadores econômicos e múltiplos ativos, aumentando a acurácia das previsões (Sun; Liu; Sima, 2020). Zhu *et al.* (2023) demonstraram que, em plataformas de empréstimos *online*, o *LightGBM* superou técnicas tradicionais como regressão logística e árvores de decisão, com apoio de métodos explicativos que garantem maior transparência. No estudo realizado por Huang *et al.* (2024), na área da saúde, o *LightGBM* demonstrou eficácia na previsão da mortalidade hospitalar de pacientes com demência, mostrando seu potencial em contextos clínicos críticos.

O *LightGBM* é um modelo preditivo baseado em *gradient boosting*, no qual a predição final é dada pela soma das contribuições das árvores f_k. Assim, a predição do i-ésimo elemento é expressa por:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

em que: K é o número total de árvores no modelo.

O treinamento do *LightGBM* busca minimizar a função objetivo

$$L(\theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

em que: $l(\cdot)$ representa a função de perda utilizada e $\Omega(f_t)$ corresponde ao termo de regularização associado à árvore adicionada no passo t.

Esse termo de regularização é definido como

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

em que: T denota o número de folhas da árvore, w_j representa o peso da folha j, γ é a penalização pela complexidade, controlando o número de folhas, e λ é o fator de regularização L2 aplicado aos pesos das folhas.

Os principais hiperparâmetros do *LightGBM* podem ser representados pelo conjunto:

$$H = \{K, \text{max\_depth}, \text{num\_leaves}, \lambda, \gamma, \text{learning\_rate}, \text{feature\_fraction}, \text{bagging\_fraction}\}$$

em que: K corresponde ao número total de árvores (*n_estimators*); *max_depth* define a profundidade máxima permitida para cada árvore; *num_leaves* especifica o número máximo de folhas; λ representa a regularização L2 aplicada aos pesos das folhas; γ controla a penalização pela complexidade da árvore; *learning_rate* determina a taxa de aprendizado do *boosting*; *feature_fraction* indica a fração de variáveis selecionadas em cada iteração; e *bagging_fraction* corresponde à fração de amostras utilizada em cada etapa do processo de *bagging*.

2.3 Pré-processamento

Segundo Meisenbacher *et al.* (2022), a previsão de séries temporais é essencial em diversas áreas e segue cinco etapas principais: pré-processamento dos dados, engenharia de atributos, otimização de hiperparâmetros, escolha do método e obtenção das previsões. Com o avanço da automação, técnicas como AutoML e métodos estatísticos tornam esse processo mais eficiente e capaz de lidar com dados complexos.

O pré-processamento dos dados possui grande relevância, pois visa preparar os dados brutos para análise, aumentando a consistência e reduzindo ruídos e problemas como valores

faltantes. Segundo Han, Kamber & Pei (2012), procedimentos como limpeza, normalização e transformação dos dados são importantes para garantir que os algoritmos de aprendizagem de máquinas operem de forma eficiente e produzam resultados confiáveis.

A engenharia de atributos refere-se ao processo de construção, transformação e seleção de variáveis que melhor representam o fenômeno estudado. De acordo com Guyon & Elisseeff (2003), a escolha adequada dos atributos influencia diretamente o desempenho dos modelos, uma vez que atributos relevantes aumentam a capacidade preditiva e reduzem a complexidade computacional.

A identificação de defasagens relevantes em séries temporais é realizada por meio da análise das funções de autocorrelação (ACF) e autocorrelação parcial (PACF), juntamente com testes estatísticos como o de *Ljung-Box*, que permitem verificar a presença de dependência temporal e avaliar a hipótese de ruído branco. A partir desses diagnósticos, selecionam-se *lags* significativos que capturam estruturas de dependência de curto e médio prazo, contribuindo para uma modelagem mais adequada do comportamento da série (Box *et al.*, 2015; Hyndman; Athanasopoulos, 2018).

A adição de variáveis derivadas em previsões, como defasagens e médias móveis, é utilizada para capturar padrões dinâmicos e aumentar o desempenho preditivo dos modelos. As defasagens permitem representar a dependência temporal, enquanto as médias móveis contribuem para a suavização de flutuações de curto prazo e para a identificação de tendências subjacentes (Box *et al.*, 2015; Hyndman; Athanasopoulos, 2018; Cleveland *et al.*, 1990).

A otimização de hiperparâmetros auxilia na identificação dos valores ideais dos hiperparâmetros que controlam o processo de aprendizado dos modelos. Segundo Bergstra & Bengio (2012), técnicas sistemáticas de busca por hiperparâmetros são capazes de melhorar expressivamente o desempenho preditivo dos modelos, tornando-os mais adaptados às características específicas dos dados analisados.

A escolha do método de modelagem consiste na seleção do algoritmo mais adequado ao problema e à estrutura dos dados disponíveis. De acordo com Hastie, Tibshirani & Friedman (2009), diferentes métodos apresentam vantagens e limitações conforme o tipo de dado, a dimensionalidade e o objetivo do estudo, sendo necessária a comparação entre alternativas para garantir uma modelagem consistente.

A obtenção das previsões corresponde à etapa final da análise de modelagem, na qual o modelo treinado é utilizado para estimar valores futuros ou desconhecidos. Segundo Hyndman

& Athanasopoulos (2018), previsões confiáveis dependem não apenas da escolha adequada do modelo, mas também da correta avaliação de seu desempenho e da validação de sua capacidade de generalização.

2.3.1 Validação cruzada para séries temporais

Segundo Hyndman & Athanasopoulos (2018), a validação da performance preditiva de modelos para séries temporais pode ser realizada inicialmente por meio da separação dos dados em amostras de treino e teste, permitindo o cálculo de erros. Contudo, devido à arbitrariedade na escolha dessas amostras, a validação cruzada específica para séries temporais é apresentada como um método mais robusto para avaliar a capacidade preditiva dos modelos.

Alguns dos desafios enfrentados na validação cruzada em séries temporais incluem a dependência serial, a não estacionariedade, o *overfitting* e a dificuldade na escolha do ponto de corte para separação dos dados. Segundo Bergmeir, Hyndman & Koo (2018), a validação cruzada com janela expansiva é aplicada em problemas de previsão como uma estratégia adequada para avaliação de modelos preditivos, uma vez que respeita a ordem temporal dos dados e permite a incorporação progressiva de novas observações ao conjunto de treinamento. Esse procedimento possibilita uma análise mais realista do desempenho preditivo em diferentes horizontes, sendo empregado tanto em estudos com modelos de aprendizado de máquina quanto em abordagens híbridas e redes neurais, contribuindo para maior confiabilidade e estabilidade das previsões (Albrecht; Rausch; Derra, 2021; Takara *et al.*, 2024).

A otimização de hiperparâmetros visa aprimorar a acurácia e a eficiência de modelos aplicados à previsão de séries temporais em ambientes de *big data*. Para modelos *fuzzy*, técnicas específicas possibilitam a realização de otimizações distribuídas e escaláveis por meio do paradigma *MapReduce*, viabilizando o ajuste eficiente dos parâmetros mesmo em grandes volumes de dados (Silva *et al.*, 2020). De forma complementar, abordagens baseadas em algoritmos genéticos paralelos têm se mostrado eficazes na seleção automatizada de hiperparâmetros, contribuindo para a redução do sobreajuste e para o aumento da capacidade preditiva dos modelos (Wu *et al.*, 2023).

2.4 Métricas para Avaliação dos Previsores

As métricas de desempenho exercem funções distintas na avaliação de modelos preditivos. O *Mean Squared Error* (MSE) calcula a média dos erros ao quadrado, sendo sensível a desvios elevados (Chai; Draxler, 2014). O *Root Mean Squared Error* (RMSE) mantém a escala original da variável e penaliza erros maiores, facilitando a interpretação dos resultados. O *Mean Absolute Error* (MAE) mede a magnitude média dos erros absolutos, enquanto o *Mean Absolute Percentage Error* (MAPE) expressa o erro em termos percentuais, embora seja sensível a valores reais próximos de zero (Jierula *et al.*, 2021).

O *Mean Squared Error* (MSE) é definido por:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2.$$

O *Root Mean Squared Error* (RMSE) é dado por:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}.$$

O *Mean Absolute Error* (MAE) é expresso como:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|.$$

Por fim, o *Mean Absolute Percentage Error* (MAPE) é definido por:

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|.$$

Em todas as métricas, n representa o número de observações, y_t corresponde ao valor observado no instante t, e $\hat{y}_t$ representa o valor previsto pelo modelo.

REFERÊNCIAS

ALBRECHT, T.; RAUSCH, T. M.; DERRA, N. D. Call me maybe: Methods and practical implementation of artificial intelligence in call center arrivals' forecasting. **Journal of Business Research**, Elsevier, v. 123, p. 267–278, 2021.

AMJAD, M. *et al.* Prediction of pile bearing capacity using XGBoost algorithm: modeling and performance evaluation. **Applied Sciences**, MDPI, v. 12, n. 4, p. 2126, 2022.

AMORIM, F. R. d. *et al.* Forecasting cost risks of corn and soybean crops through Monte Carlo Simulation. **Applied Sciences**, MDPI, v. 14, n. 17, p. 8030, 2024.

ANGGRAINI, D.; NABILA, S. A.; FAJARYANTI, J. Stepping up food price prediction using prophet model. **International Journal of Informatics and Computation**, v. 8, n. 1, p. 533–542, 2026.

ARAUJO, A. A.; SOUSA, A. G.; SANTOS, R. B. N. d. EXPORTAÇÃO BRASILEIRA DE MELÃO: UM ESTUDO DE SÉRIES TEMPORAIS. **Revista de Economia e Agronegócio**, 2008.

ARSLAN, S. A hybrid forecasting model using LSTM and prophet for energy consumption with decomposition of time series data. **PeerJ Computer Science**, PeerJ Inc., v. 8, p. e1001, 2022.

ATAMIMI, F. M. H.; WITANTI, W.; ABDILLAH, G. Enhanching prophet time series forecasting on sparse data via hyperparameter optimizattion: A case study in retail. **Sinkron: Jurnal dan Penelitian Teknik Informatika**, v. 9, n. 2, p. 1000–1007, 2025.

AVILEIS, F. G.; MALLORY, M. L. The impact of Brazil on global grain dynamics: A study on cross-market volatility spillovers. **Agricultural Economics**, Wiley Online Library, v. 53, n. 2, p. 231–245, 2022.

AWAD, M.; KHANNA, R. **Efficient learning machines: theories, concepts, and applications for engineers and system designers**. [*S.l.*]: Apress, 2015.

BAGCHI, P.; CHARACIEJUS, V.; DETTE, H. A simple test for white noise in functional time series. **Journal of Time Series Analysis**, Wiley Online Library, v. 39, n. 1, p. 54–74, 2018.

BERGMEIR, C.; HYNDMAN, R. J.; KOO, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. **Computational Statistics & Data Analysis**, Elsevier, v. 120, p. 70–83, 2018.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **Journal of Machine Learning Research**, v. 13, n. 2, 2012.

BOX, G. E. P. *et al.* **Time series analysis: forecasting and control**. [*S.l.*]: John Wiley & Sons, 2015.

BREIMAN, L. Random forests. **Machine Learning**, Kluwer Academic Publishers, v. 45, p. 5–32, 2001.

CAI, K.; RODAVIA, M. R. XGBoost analysis based on consumer behavior. **Frontiers in Computing and Intelligent Systems**, v. 5, n. 2, p. 85–89, 2023.

CARMO, C. R. S. Atividade agrícola: uma análise sobre a sua contribuição para a economia do estado de minas gerais e seus possíveis determinantes agrícolas. **Revista em Agronegócio e Meio Ambiente**, v. 9, n. 2, p. 223–249, 2016.

CHAI, T.; DRAXLER, R. R. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. **Geoscientific Model Development**, Copernicus Publications, v. 7, n. 3, p. 1247–1250, 2014.

CHEN, C.; SEO, H. Prediction of rock mass class ahead of TBM excavation face by ML and DL algorithms with Bayesian TPE optimization and SHAP feature analysis. **Acta Geotechnica**, Springer, v. 18, n. 7, p. 3825–3848, 2023.

CHEN, S.-M. Forecasting enrollments based on fuzzy time series. **Fuzzy Sets and Systems**, v. 81, n. 3, p. 311–319, 1996.

CHEN, S.-M. Forecasting enrollments based on high-order fuzzy time series. **Cybernetics and Systems**, Taylor & Francis, v. 33, n. 1, p. 1–16, 2002.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [*S.l.: s.n.*], 2016. p. 785–794.

CHEN, Z.; FAN, W. A freeway travel time prediction method based on an XGBoost model. **Sustainability**, v. 13, p. 8577, 2021.

CLEVELAND, R. B. *et al.* A seasonal-trend decomposition procedure based on Loess. **Journal of Official Statistics**, v. 6, n. 1, p. 3–73, 1990.

CUBADDA, G. Common cycles in seasonal non-stationary time series. **Journal of Applied Econometrics**, Wiley Online Library, v. 14, n. 3, p. 273–291, 1999.

DASH, R. K. *et al.* Fine-tuned support vector regression model for stock predictions. **Neural Computing and Applications**, Springer, v. 35, n. 32, p. 23295–23309, 2023.

DOBRA, A. Decision trees. **Encyclopedia of Database Systems**, Springer US, p. 769–769, 2009.

DOZIE, K. C. N.; IJOMAH, M. A. A comparative study on additive and mixed models in descriptive time series. **American Journal of Mathematical and Computer Modelling**, v. 5, n. 1, p. 12–17, 2020.

DRUCKER, H. *et al.* Support vector regression machines. In: **Advances in Neural Information Processing Systems**. [*S.l.*]: MIT Press, 1997. p. 155–161.

DU, W.; CHEN, R.; CONG, Z. Application of support vector regression in prediction model using genetic algorithm optimized. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [*S.l.*], 2021. v. 1982, n. 1, p. 012048.

DUDEK, G. STD: a seasonal-trend-dispersion decomposition of time series. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 35, n. 10, p. 10339–10350, 2023.

FU, T.-c. A review on time series data mining. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 24, n. 1, p. 164–181, 2011.

GARCÍA-LARA, S.; SERNA-SALDIVAR, S. O. Corn history and culture. **Corn**, Elsevier, p. 1–18, 2019.

GUPTA, H. *et al.* Forecasting commodity prices using machine learning. **International Journal of Scientific Research in Science and Technology**, v. 11, n. 1, 2024.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research**, v. 3, p. 1157–1182, 2003.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: concepts and techniques**. [*S.l.*]: Morgan Kaufmann Publishers, 2012.

HANNAN, E. J.; TERRELL, R. D.; TUCKWELL, N. E. The seasonal adjustment of economic time series. **International Economic Review**, JSTOR, v. 11, n. 1, p. 24–52, 1970.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. [*S.l.*]: Springer Science & Business Media, 2009.

HINTON, G.; SEJNOWSKI, T. Unsupervised learning. In: SHEKHAR, S.; XIONG, H.; ZHOU, X. (Ed.). **Encyclopedia of GIS**. Cham: Springer International Publishing, 2017. p. 2375. ISBN 978-3-319-17885-1. Disponível em: https://doi.org/10.1007/978-3-319-17885-1_101437.

HUANG, C.-C. *et al.* Artificial intelligence prediction of In-Hospital mortality in patients with dementia: A multi-center study. **International Journal of Medical Informatics**, Elsevier, v. 191, p. 105590, 2024.

HUSSAIN, N. *et al.* Stock market performance analytics using XGBoost. In: SPRINGER. **Mexican International Conference on Artificial Intelligence**. [*S.l.*], 2023. p. 3–16.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. 2. ed. [*S.l.*]: OTexts, 2018.

IQBAL, M. *et al.* Multivariate fuzzy time series forecasting using clustering and optimization techniques. **IEEE Access**, v. 8, p. 123456–123470, 2020.

ISENSEE, L. J.; PINHEIRO, A.; DETZEL, D. H. M. Vazão de projeto de barragens sob condição de não-estacionariedade. **Revista Brasileira de Geografia Física**, v. 14, n. 5, p. 2975–2987, 2021.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. **Electronic Markets**, Springer, v. 31, n. 3, p. 685–695, 2021.

JIERULA, A. *et al.* Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. **Applied Sciences**, MDPI, v. 11, n. 5, p. 2314, 2021.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015.

KASIANOVA, N.; POPOV, Y.; DEHTIAROV, V. Data analytics as the basis for shaping marketing strategies of agricultural producers. **Black Sea Economic Studies**, Helvetica Publishing House, 2025.

KHANNA, M. An econometric analysis of US crop yield and cropland acreage: implications for the impact of climate change. **SSRN Electronic Journal**, 2010.

KOSTARIDOU, E. *et al.* Empirical comparison of Facebook prophet and traditional models for tomato price forecasting in Greece. **Research on World Agricultural Economy**, v. 5, n. 2, p. 594–607, 2024.

LARY, D. J. *et al.* Machine learning in geosciences and remote sensing. **Geoscience Frontiers**, Elsevier, v. 7, n. 1, p. 3–10, 2016.

LATORRE, M. d. R. D. d. O.; CARDOSO, M. R. A. Análise de séries temporais em epidemiologia: uma introdução sobre os aspectos metodológicos. **Revista Brasileira de Epidemiologia**, SciELO Brasil, v. 4, n. 3, p. 145–152, 2001.

LEE, L.-W.; WANG, L.-M.; CHEN, S.-M. A multivariate heuristic model for fuzzy time-series forecasting. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, v. 37, n. 4, p. 836–846, 2007.

LI, L. *et al.* Trend modeling for traffic time series analysis: An integrated study. **IEEE Transactions on Intelligent Transportation Systems**, IEEE, v. 16, n. 6, p. 3430–3439, 2015.

LIU, Z. *et al.* Forecast methods for time series data: A survey. **IEEE Access**, IEEE, v. 9, p. 91896–91912, 2021.

LUCAS, P. O. *et al.* A tutorial on fuzzy time series forecasting models: Recent advances and challenges. **Learning and Nonlinear Models**, v. 19, n. 2, p. 29–50, 2021.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. **Estudos Avançados**, SciELO Brasil, v. 35, n. 101, p. 85–94, 2021.

MANOGNA, R. L.; DHARMAJI, V.; SARANG, S. Enhancing agricultural commodity price forecasting with deep learning. **Scientific Reports**, Nature Publishing Group, v. 15, n. 1, p. 20903, 2025.

MASINI, R. P.; MEDEIROS, M. C.; MENDES, E. F. Machine learning advances for time series forecasting. **Journal of Economic Surveys**, Wiley Online Library, v. 37, n. 1, p. 76–111, 2023.

MEISENBACHER, S. *et al.* Review of automated time series forecasting pipelines. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 12, n. 6, p. e1475, 2022.

NAVRATIL, M.; KOLKOVA, A. Decomposition and forecasting time series in the business economy using prophet forecasting model. **Central European Business Review**, v. 8, n. 4, p. 26–41, 2019.

OUYANG, Y. *et al.* Integration of machine learning XGBoost and SHAP models for NBA game outcome prediction and quantitative analysis methodology. **PLoS ONE**, Public Library of Science, v. 19, n. 7, p. e0307478, 2024.

PARBAT, D.; CHAKRABORTY, M. A Python based support vector regression model for prediction of COVID-19 cases in India. **Chaos, Solitons & Fractals**, Elsevier, v. 138, p. 109942, 2020.

PERVEEN, S.; RIZWAN, M.; GOEL, N. Anfis-based prediction models: A review. **International Journal of Intelligent Systems and Applications**, 2019.

RODRÍGUEZ, G. G.; GONZALEZ-CAVA, J. M.; PÉREZ, J. A. M. An intelligent decision support system for production planning based on machine learning. **Journal of Intelligent Manufacturing**, Springer, v. 31, n. 5, p. 1257–1273, 2020.

RUSPINI, E. H. On the semantics of fuzzy logic. **International Journal of Approximate Reasoning**, Elsevier, v. 5, n. 1, p. 45–88, 1991.

ŠARAC, V. *et al.* Influence of corn and soybean prices on broiler prices. **Journal on Processing and Energy in Agriculture**, v. 27, n. 2, p. 73–75, 2023.

SCHONLAU, M.; ZOU, R. Y. The random forest algorithm for statistical learning. **The Stata Journal**, SAGE Publications, v. 20, n. 1, p. 3–29, 2020.

SHARMA, M. A brief description on soybean and its food products. **Asian Journal of Research in Business Economics and Management**, 2021.

SHERLY, A.; CHRISTO, M. S.; ELIZABETH, J. V. A hybrid approach to time series forecasting: Integrating ARIMA and prophet for improved accuracy. **Results in Engineering**, Elsevier, v. 27, p. 105703, 2025.

SIDEY-GIBBONS, J. A. M.; SIDEY-GIBBONS, C. J. Machine learning in medicine: a practical introduction. **BMC Medical Research Methodology**, Springer, v. 19, n. 1, p. 64, 2019.

SILVA, P. C. L. *et al.* Distributed evolutionary hyperparameter optimization for fuzzy time series. **IEEE Transactions on Network and Service Management**, IEEE, v. 17, n. 3, p. 1309–1321, 2020.

SONG, Q.; CHISSOM, B. S. Forecasting enrollments with fuzzy time series — part i. **Fuzzy Sets and Systems**, v. 54, n. 1, p. 1–10, 1993.

SONG, Q.; CHISSOM, B. S. Fuzzy time series and its models. **Fuzzy Sets and Systems**, v. 54, n. 3, p. 269–277, 1993.

SOUSA, A. P. d. *et al.* Análise comparativa de métodos de previsão de séries temporais através de modelos estatísticos e rede neural artificial. **Trabalho de Graduação**, Pontifícia Universidade Católica de Goiás, 2012.

SUN, X.; LIU, M.; SIMA, Z. A novel cryptocurrency price trend forecasting model based on LightGBM. **Finance Research Letters**, Elsevier, v. 32, p. 101084, 2020.

TAKARA, L. de A. *et al.* Optimizing multi-step wind power forecasting: Integrating advanced deep neural networks with stacking-based probabilistic learning. **Applied Energy**, Elsevier, v. 369, p. 123487, 2024.

TALEKAR, B.; AGRAWAL, S. A detailed review on decision tree and random forest. **Bioscience Biotechnology Research Communications**, v. 13, n. 14, p. 245–248, 2020.

TAYLOR, S. J.; LETHAM, B. Forecasting at scale. **The American Statistician**, v. 72, n. 1, p. 37–45, 2018.

TESSMANN, M. S.; CARRASCO-GUTIERREZ, C. E.; LIMA, A. V. Determinants of corn and soybean futures prices traded on the Brazilian stock exchange: An ARDL approach. **International Journal of Economics and Finance**, v. 15, n. 1, p. 65–75, 2022.

VAPNIK, V.; GOLOWICH, S.; SMOLA, A. Support vector method for function approximation, regression estimation and signal processing. **Advances in Neural Information Processing Systems**, v. 9, p. 281–287, 1996.

VELASCO, C. Estimation of time series models using residuals dependence measures. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 50, n. 5, p. 3039–3063, 2022.

WANG, Y.; SHAO, Z.; JUMAHONG, M. Hybrid deep learning and fuzzy systems for multivariate time series forecasting. **Expert Systems with Applications**, v. 228, 2023.

WU, D. *et al.* AutoML with parallel genetic algorithm for fast hyperparameters optimization in efficient IoT time series prediction. **IEEE Transactions on Industrial Informatics**, IEEE, v. 19, n. 9, p. 9555–9564, 2023.

WU, Y. *et al.* A survey of multivariate fuzzy time series forecasting methods. **Applied Soft Computing**, 2021.

WUNDERLICH, A.; SKLAR, J. Data-driven modeling of noise time series with convolutional generative adversarial networks. **Machine Learning: Science and Technology**, IOP Publishing, v. 4, n. 3, p. 035023, 2023.

YAN, J. *et al.* LightGBM: accelerated genomically designed crop breeding through ensemble learning. **Genome Biology**, Springer, v. 22, n. 1, p. 271, 2021.

YAZDANBAKHSH, O.; DICK, S. A systematic review of type-2 fuzzy logic systems for modeling uncertainty. **Information Sciences**, v. 416–417, p. 192–220, 2017.

ZADEH, L. A. Knowledge representation in fuzzy logic. In: **An introduction to fuzzy logic applications in intelligent systems**. [*S.l.*]: Springer, 1992. p. 1–25.

ZHANG, W. *et al.* Variational learning of deep fuzzy theoretic nonparametric model. **Neurocomputing**, Elsevier, v. 506, p. 128–145, 2022.

ZHU, X. *et al.* Explainable prediction of loan default based on machine learning models. **Data Science and Management**, Elsevier, v. 6, n. 3, p. 123–133, 2023.

SEGUNDA PARTE - ARTIGOS

ARTIGO 1 - Uso de métodos de séries temporais fuzzy para a previsão do preço do milho

Uso de métodos de séries temporais fuzzy para a previsão do preço do milho

Informações do artigo

Palavras-chave:
Lógica Fuzzy
Séries Temporais Fuzzy
Commodities Agrícolas

RESUMO

Este estudo tem como objetivo prever os preços diários do milho no estado de Minas Gerais utilizando métodos baseados em lógica fuzzy. Foram considerados diferentes recortes temporais da série de preços do milho semiduro (R$/sc), com início em 2007, 2014 e 2021. Os dados, obtidos no CEPEA, abrangem o período de 2007 a 2025, totalizando 4.727 observações. Após o pré-processamento, os dados foram organizados em conjuntos de treino e teste, sobre os quais foram ajustados e avaliados os modelos High Order Fuzzy Time Series (HOFTS), Weighted High Order Fuzzy Time Series (WHOFTS) e Partitioned Weighted Fuzzy Time Series (PWFTS) considerando diferentes ordens de defasagem e estratégias de particionamento do universo do discurso. Os resultados indicam que o modelo PWFTS apresentou o melhor desempenho para a série iniciada em 2014, no horizonte de previsão de 1 dia à frente, obtendo os menores valores para as métricas de erro e a maior acurácia direcional. De modo geral, os modelos HOFTS e WHOFTS também apresentaram desempenho satisfatório, embora esse desempenho tenha variado entre os recortes analisados. Além disso, os resultados sugerem que os modelos fuzzy apresentam redução de desempenho preditivo em horizontes de previsão mais longos, como sete e 30 dias à frente, e que o aumento do volume de dados de treinamento não proporciona ganhos expressivos de desempenho para os modelos FTS. Os achados demonstram que os métodos baseados em lógica fuzzy constituem uma abordagem competitiva para a previsão diária dos preços do milho, com potencial para apoiar a tomada de decisão no agronegócio.

1. Introdução

A cultura do milho tem relevância em escala global, contribuindo para a segurança alimentar, nutricional e energética, além de gerar impactos ambientais, culturais e econômicos. Estudos demonstram que a produção mundial dessa commodity apresenta expressiva participação no comércio internacional (Wang and Hu, 2021). A elevada demanda pelo milho está associada à sua versatilidade de uso na alimentação humana, na produção de ração animal e na geração de biocombustíveis (Kumar, Kumar, Kumar and Kumar, 2019). No contexto brasileiro, o milho destaca-se como um dos principais cereais cultivados, ocupando posição estratégica na agricultura nacional e apresentando significativa relevância socioeconômica (Silva, Sales Silva, Souza, Lima and Santos, 2020).

Além de sua importância econômica, o milho possui valor cultural e simbólico para diversos povos, estando associado a práticas religiosas e tradições históricas (Rivas, 2021). Atualmente, essa cultura agrícola também se destaca como uma das principais commodities globais, integrando cadeias produtivas complexas e demandando capacidade de adaptação frente a riscos externos (Wu and Zhu, 2025). Seu valor nutricional e energético, aliado à ampla diversidade de aplicações, reforça sua importância econômica e tecnológica (Bakhtiyorovna, O'Gli and Qizi, 2021). O crescimento da demanda por milho, tanto para alimentação animal quanto para aplicações industriais e tecnológicas, tem ampliado o interesse por estudos voltados à compreensão dos fatores que influenciam sua produção e comercialização (Kerchner et al., 2024).

Nos últimos anos, os preços do milho passaram a apresentar maior volatilidade em decorrência de eventos como a pandemia da COVID-19 e tensões geopolíticas internacionais. Nesse contexto, Wu, Wang and Wang (2024) propuseram um modelo que incorpora fatores relacionados à oferta, demanda, políticas públicas e riscos internacionais para compreender o comportamento dos preços. A previsão de preços de commodities tornou-se uma ferramenta importante para o planejamento e a tomada de decisões, sendo observados avanços com a utilização de algoritmos genéticos e modelos

híbridos (Drachal and Pawlowski, 2021). No Brasil, a expansão da segunda safra contribuiu para ampliar a participação do país no mercado internacional de milho (Mallory and Avileis, 2022). Adicionalmente, Justus et al. (2024) verificaram que os preços internacionais exerceram influência significativa sobre os preços domésticos entre 2005 e 2023.

A modelagem de preços agrícolas constitui um desafio devido às características de não linearidade e não estacionariedade frequentemente observadas nessas séries temporais. Em razão dessa complexidade, modelos estatísticos tradicionais podem apresentar limitações em sua capacidade preditiva, motivando o desenvolvimento de abordagens mais sofisticadas e técnicas híbridas (Choudhary et al., 2025). Estudos anteriores demonstram que a combinação de diferentes metodologias pode melhorar significativamente o desempenho preditivo em séries de preços agrícolas (Xiong, Li and Bao, 2018).

A qualidade das previsões de produtividade e preços do milho é fundamental para o planejamento agrícola e econômico. Nesse sentido, tecnologias baseadas em sensoriamento remoto e dados de satélite têm sido empregadas para aumentar a precisão das estimativas (Teste et al., 2024). Além disso, Wu et al. (2024) desenvolveram um modelo baseado em séries temporais multivariadas e aprendizado profundo capaz de incorporar fatores econômicos e geopolíticos, aumentando a precisão das previsões e contribuindo para a estabilidade do mercado.

A previsão dos preços do milho também tem se beneficiado dos avanços em técnicas de inteligência artificial. Redes neurais artificiais apresentam resultados satisfatórios em previsões de curto prazo ao explorar informações históricas e preços futuros (Xu and Zhang, 2021). De forma semelhante, modelos do tipo Long Short-Term Memory (LSTM) têm demonstrado ganhos de desempenho ao incorporar variáveis externas ao processo de previsão (Gu and Li, 2024). Outros métodos, como Random Forest, Support Vector Machine (SVM), AdaBoost e ARIMA, também têm sido aplicados com sucesso quando combinados a técnicas de seleção de atributos e interpretação baseadas em valores SHAP (Gaur et al., 2024). Apesar desses avanços, muitos modelos ainda apresentam limitações relacionadas à interpretabilidade, à dependência de grandes volumes de dados e à dificuldade em lidar adequadamente com incertezas (Makridakis, Spiliotis and Assimakopoulos, 2018).

Como alternativa, a lógica fuzzy oferece uma estrutura capaz de representar e processar informações imprecisas por meio de regras linguísticas. Segundo Tanscheit (2005), essa abordagem permite traduzir o raciocínio humano aproximado em sistemas computacionais, possibilitando a tomada de decisão em ambientes de incerteza. Os modelos de Séries Temporais Fuzzy (FTS) têm recebido destaque por sua capacidade de lidar com incertezas presentes em dados temporais, fornecendo previsões interpretáveis e com menor complexidade computacional quando comparados a diversos métodos estatísticos tradicionais (Lucas et al., 2021b). Além disso, modelos baseados em lógica fuzzy apresentam elevada flexibilidade para cenários caracterizados por volatilidade e incerteza (Frantti and Mähönen, 2001).

Fatih (2022) utilizaram modelos FTS para prever preços internacionais mensais do café entre 2000 e 2022, empregando etapas como definição do universo de discurso, construção dos conjuntos fuzzy, fuzzificação dos dados, estabelecimento das relações lógicas fuzzy e geração das previsões. Apesar do crescente interesse por essa abordagem, não foram identificados estudos que utilizem modelos FTS para previsão de preços diários do milho, especialmente considerando dados de alta frequência, e com diferentes configurações de treinamento e horizontes de previsão. A maioria das pesquisas concentra-se em frequências semanais ou mensais, evidenciando uma lacuna na literatura.

Diante desse contexto, o presente estudo tem como objetivo avaliar a aplicabilidade de métodos baseados em lógica fuzzy na previsão de preços diários do milho, expressos em reais por saca de 60 kg (R$/sc), no estado de Minas Gerais. Busca-se investigar a capacidade desses modelos em representar incertezas e capturar variações de curto prazo nos preços dessa commodity, contribuindo para o avanço das técnicas de previsão aplicadas ao mercado agrícola.

2. Metodologia

2.1. Dados e Pré-processamento

Os dados foram obtidos do Centro de Estudos Avançados em Economia Aplicada (CEPEA), referentes à modalidade de comercialização "Balcão (produtor)" para o produto milho semiduro na região do Triângulo Mineiro. A série contempla o período de 2 de outubro de 2007 a 6 de junho de 2025, totalizando 4.727 observações diárias de preços por saca de 60 kg (R$/sc), à vista (CDI), sem ICMS. Esse intervalo temporal permite observar flutuações nos preços influenciadas por fatores relacionados à oferta, demanda, mercado agropecuário e condições climáticas.

Para avaliação dos modelos, foram considerados três recortes distintos do conjunto de dados: o primeiro abrangendo toda a série histórica disponível (2007–2025), o segundo iniciando em 1 de janeiro de 2014 e o terceiro a partir de 1 de janeiro de 2021. Em todos os cenários, o conjunto de teste foi padronizado em 300 observações mais recentes, visando manter a comparabilidade entre os experimentos e reduzir o custo computacional. Foram analisadas defasagens (lags) de 1, 2, 3 e 4 períodos, partições de 30, 60, 90 e 120 intervalos e horizontes de previsão de 1, 7 e 30 dias.

A otimização dos hiperparâmetros foi realizada por meio da validação cruzada para séries temporais (Time Series Cross-Validation), utilizando o Root Mean Squared Error (RMSE) como função de custo para seleção da melhor configuração em cada cenário. Essa abordagem permite uma avaliação mais robusta da capacidade preditiva dos modelos, respeitando a estrutura temporal dos dados e reduzindo riscos de sobreajuste (Bergmeir, Costantini and Benítez, 2014). Como a base de dados não apresentou valores ausentes, não foi necessária a aplicação de técnicas de imputação.

Modelagem de Séries Temporais Fuzzy

Os modelos de Séries Temporais Fuzzy (FTS, *Fuzzy Time Series*) apresentam vantagens significativas em relação às técnicas estatísticas tradicionais de previsão por prescindirem de pressupostos rígidos sobre a distribuição dos dados, além de fornecerem resultados robustos na modelagem de incertezas conceituais e flutuações de dados históricos Lucas, Orang, Silva, Mendes and Guimarães (2021a). Conforme amplamente revisado na literatura desses frameworks, os métodos baseados em lógica fuzzy estruturam-se como ferramentas computacionais flexíveis capazes de lidar com dados imprecisos em problemas complexos de previsão preditiva sem a necessidade de parametrizações probabilísticas prévias (Wang, Liu, Chi and Li, 2024; Orang, Silva, Silva and Guimarães, 2020).

Do ponto de vista conceitual, a base teórica fundamenta-se na teoria dos conjuntos fuzzy clássica estabelecida por Zadeh (1965), na qual o universo de discurso é particionado em partições ou intervalos definidos por conjuntos fuzzy cujos elementos possuem um grau de pertinência mapeado continuamente no intervalo $[0, 1]$. No escopo dos modelos de previsão temporal, essa estrutura linguística foi formalizada pioneiramente por Song e Chissom Song and Chissom (1993), permitindo traduzir séries numéricas em séries de termos fuzzy estruturados sequencialmente.

No escopo metodológico, o processo de modelagem de uma Série Temporal Fuzzy (FTS) univariada toma como ponto de partida uma série temporal numérica convencional, representada

por $Y(t) \in \mathbb{R}$, para $t = 0, 1, \ldots, T$ (Lucas et al., 2021b). A partir dessa série de dados, o Universo de Discurso U é delimitado pelos limites operacionais mínimos e máximos identificados nas observações históricas disponíveis, podendo ser expandido por margens de segurança D_1 e D_2 (Lucas et al., 2021b):

$$U = [\min(Y) - D_1, \max(Y) + D_2].$$

Sobre o universo de discurso U, são definidos k conjuntos fuzzy A_j (para $j = 1, \ldots, k$), em que cada um possui sua respectiva função de pertinência μ_{A_j} (Lucas et al., 2021b). A série temporal fuzzy $F(t)$ representa a sequência de estados linguísticos associados a $Y(t)$, em que cada observação $y(t)$ é mapeada no conjunto fuzzy A_j de maior grau de pertinência (Song and Chissom, 1993; Lucas et al., 2021b):

$$F(t) = A_j \quad \text{tal que} \quad \mu_{A_j}(y(t)) = \max_m \mu_{A_m}(y(t)).$$

A base de conhecimento do modelo é composta por regras estruturadas na forma de Relações Lógicas Fuzzy (FLR) de primeira ordem, cuja representação padrão é expressa por (Song and Chissom, 1993; Lucas et al., 2021b):

$$A_i \rightarrow A_j,$$

em que: A_i atua como o antecedente (estado fuzzy no tempo $t-1$) e A_j corresponde ao consequente (estado fuzzy no tempo t) (Lucas et al., 2021b).

Para capturar dependências temporais estendidas, a abordagem expande-se para modelos de alta ordem (Lucas et al., 2021b). Conforme demonstrado por Lucas et al. (2021b) e Züge and Coelho (2024), uma Série Temporal Fuzzy de Alta Ordem (HOFTS) de ordem m ($m \geq 2$) considera múltiplos momentos antecedentes, sendo representada por:

$$A_{i_{t-m}}, \ldots, A_{i_{t-1}} \rightarrow A_{i_t}.$$

No modelo WHOFTS (Yu, 2005), atribui-se um peso w_j a cada relação lógica fuzzy contida em um grupo de relações lógicas (FLRG), baseado na frequência de ocorrência histórica da transição $A_{i_{t-m}}, \ldots, A_{i_{t-1}} \rightarrow A_j$. A previsão defuzzificada $\hat{Y}(t)$ é calculada pela média ponderada dos centroides c_j dos conjuntos fuzzy consequentes:

$$\hat{Y}(t) = \frac{\sum_{j=1}^{k} w_j \cdot c_j}{\sum_{j=1}^{k} w_j},$$

em que: w_j representa o peso/frequência acumulada da j-ésima regra no FLRG e c_j é o centroide (ponto médio) do conjunto fuzzy A_j.

O modelo PWFTS introduzido por Silva et al. (2020) aplica uma matriz de transição de probabilidades empíricas para modelar a incerteza do consequente dado o histórico do antecedente de ordem m. A probabilidade condicional de transição $P(A_j \mid A_{i_{t-m}}, \ldots, A_{i_{t-1}})$ é definida por (Silva et al., 2020):

$$P(A_j \mid A_{i_{t-m}}, \ldots, A_{i_{t-1}}) = \frac{N(A_{i_{t-m}}, \ldots, A_{i_{t-1}} \rightarrow A_j)}{\sum_{l=1}^{k} N(A_{i_{t-m}}, \ldots, A_{i_{t-1}} \rightarrow A_l)},$$

em que: $N(\cdot)$ denota a contagem exata de ocorrências da sequência de antecedentes para o consequente A_j no conjunto de treinamento.

A previsão quantitativa final do PWFTS combina a distribuição de probabilidade condicional sobre todos os k conjuntos fuzzy consequentes com seus respectivos centroides c_j (Silva et al., 2020):

$$\hat{Y}(t) = \sum_{j=1}^{k} P(A_j \mid A_{i_{t-m}}, \ldots, A_{i_{t-1}}) \cdot c_j.$$

2.2. Hiperparâmetros

De acordo com Bouktif, Fiaz, Ouni and Serhani (2020) a escolha inadequada de hiperparâmetros pode comprometer o desempenho de modelos de previsão, uma vez que influenciam de forma direta a capacidade dos padrões presentes em séries temporais. Os autores também enfatizam que a definição de hiperparâmetros constitui um problema de otimização complexo.

As Tabelas 1, 2 e 3 apresentam os hiperparâmetros ótimos dos modelos HOFTS, WHOFTS e PWFTS para diferentes horizontes de previsão. A seleção foi realizada com base no menor valor de RMSE-CV obtido durante o processo de validação cruzada para séries temporais, considerando os anos de 2007, 2014 e 2021, respectivamente.

Tabela 1. Melhores hiperparâmetros obtidos para os modelos fuzzy em diferentes horizontes de previsão do ano de 2007.

Horizonte	Modelo	Npart	Ordem	RMSE-CV
1 dia	HOFTS	60	4	1.431
	WHOFTS	60	4	1.440
	PWFTS	15	3	1.743
7 dias	HOFTS	60	2	2.422
	WHOFTS	60	4	2.444
	PWFTS	15	3	2.661
30 dias	HOFTS	60	4	4.968
	WHOFTS	60	4	4.973
	PWFTS	15	4	5.115

Tabela 2. Melhores hiperparâmetros obtidos para os modelos fuzzy em diferentes horizontes de previsão do ano de 2014.

Horizonte	Modelo	Npart	Ordem	RMSE-CV
1 dia	HOFTS	60	2	1.898
	WHOFTS	45	3	1.931
	PWFTS	15	2	1.954
7 dias	HOFTS	60	2	3.043
	WHOFTS	60	3	3.085
	PWFTS	15	3	3.187
30 dias	HOFTS	60	3	6.146
	WHOFTS	60	4	6.165
	PWFTS	15	4	6.289

Tabela 3. Melhores hiperparâmetros obtidos para os modelos fuzzy em diferentes horizontes de previsão do ano de 2021.

Horizonte	Modelo	Npart	Ordem	RMSE-CV
1 dia	HOFTS	60	2	2.835
	WHOFTS	60	2	2.839
	PWFTS	15	3	3.120
7 dias	HOFTS	30	3	3.848
	WHOFTS	30	3	3.919
	PWFTS	15	3	3.886
30 dias	HOFTS	30	3	6.977
	WHOFTS	30	3	7.043
	PWFTS	15	3	7.160

2.3. Métricas de Avaliação de Desempenho

O desempenho preditivo dos modelos foi avaliado por meio das métricas Mean Absolute Percentage Error (MAPE), Mean Directional Accuracy (MDA) e Root Mean Squared Error (RMSE), amplamente empregadas em estudos de previsão de séries temporais. A utilização conjunta dessas métricas permite avaliar tanto a magnitude dos erros quanto a capacidade dos modelos em identificar corretamente a direção das variações observadas Lehna, Scheller and Herwartz (2022).

O Mean Absolute Percentage Error (MAPE) mede o erro percentual médio entre os valores observados e previstos, sendo calculado por:

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|.$$

A Mean Directional Accuracy (MDA) avalia a capacidade do modelo em prever corretamente a direção das variações da série temporal, sendo particularmente útil em aplicações voltadas ao suporte à tomada de decisão Astore (2022). A métrica é definida por:

$$MDA = \frac{1}{n} \sum_{t=2}^{n} \mathbf{1} \left[\text{sign}(y_t - y_{t-1}) = \text{sign}(\hat{y}_t - y_{t-1}) \right],$$

em que: $\mathbf{1}(\cdot)$ é uma função indicadora que assume valor 1 quando a direção prevista coincide com a direção observada e 0 caso contrário.

O Root Mean Squared Error (RMSE) quantifica a magnitude média dos erros de previsão, atribuindo maior penalização aos desvios mais elevados devido ao termo quadrático. Essa métrica é amplamente utilizada para avaliar a proximidade entre os valores observados e previstos Hyndman and Koehler (2006). Seu cálculo é dado por:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}.$$

Em todas as métricas, n representa o número de observações, y_t corresponde ao valor observado no instante t, e $\hat{y}_t$ representa o valor previsto pelo modelo.

2.4. Recursos Computacionais

As análises foram realizadas em ambiente Python Software Foundation (2024), utilizando as bibliotecas *pandas*, *NumPy*, *pyFTS*, *Matplotlib* e *scikit-learn*. O ambiente computacional foi composto por um processador Intel(R) Xeon(R) E5-2630 v4 com frequência de 2,20 GHz, 40 *threads* de processamento e 78 GB de memória RAM, executando o sistema operacional Linux.

3. Resultados

A Tabela 4 apresenta os resultados das métricas de avaliação dos modelos HOFTS, WHOFTS e PWFTS para os horizontes de previsão de 1, 7 e 30 dias no ano de 2007. Foram consideradas as métricas MAPE, MDA e RMSE.

Tabela 4. Métricas de avaliação dos modelos fuzzy para diferentes horizontes de previsão no ano de 2007.

Horizonte	Modelo	RMSE	MAPE (%)	MDA
1 dia	HOFTS	3.706	4.223	0.433
	WHOFTS	3.706	4.198	0.460
	PWFTS	3.256	4.263	0.456
7 dias	HOFTS	4.933	6.433	0.401
	WHOFTS	4.564	5.570	0.438
	PWFTS	4.432	5.838	0.408
30 dias	HOFTS	8.162	9.238	0.387
	WHOFTS	8.252	9.305	0.413
	PWFTS	8.237	9.554	0.401

De forma geral, os modelos apresentam desempenho competitivo ao longo dos horizontes de 1, 7 e 30 dias, com variações consistentes entre as métricas avaliadas. No horizonte de 1 dia, o PWFTS obteve o menor RMSE (3.256), enquanto o WHOFTS destacou-se com o menor MAPE (4.198%) e o maior MDA (0.460).

Para o horizonte de 7 dias, o PWFTS manteve o menor RMSE (4.432), enquanto o WHOFTS registrou o menor MAPE (5.570%) e o maior MDA (0.438). Já no horizonte de 30 dias, o HOFTS

obteve o menor RMSE (8.162) e o menor MAPE (9.238%), indicando melhor ajuste absoluto e relativo, enquanto o WHOFTS alcançou o maior MDA (0.413). Dessa forma, o PWFTS destaca-se pela precisão absoluta de ajuste (RMSE) nos horizontes de curto e médio prazo (1 e 7 dias), ao passo que o WHOFTS sobressai no direcionamento de tendência (MDA) na maioria dos cenários.

A Tabela 5 apresenta os resultados das métricas de avaliação dos modelos HOFTS, WHOFTS e PWFTS para os horizontes de previsão de 1, 7 e 30 dias no ano de 2014. Foram consideradas as métricas MAPE, MDA e RMSE.

Tabela 5. Métricas de avaliação dos modelos fuzzy para diferentes horizontes de previsão no ano de 2014.

Horizonte	Modelo	RMSE	MAPE (%)	MDA
1 dia	HOFTS	4.099	5.608	0.356
	WHOFTS	3.715	4.461	0.446
	PWFTS	3.257	4.235	0.480
7 dias	HOFTS	4.929	6.641	0.397
	WHOFTS	4.491	5.495	0.486
	PWFTS	4.416	5.862	0.401
30 dias	HOFTS	8.186	9.471	0.383
	WHOFTS	8.287	9.568	0.424
	PWFTS	8.329	9.526	0.390

Para o ano de 2014, observa-se que os modelos tambem apresentam desempenho competitivo ao longo dos diferentes horizontes de previsão (1, 7 e 30 dias), com variações consistentes entre as métricas avaliadas. No horizonte de 1 dia, o modelo PWFTS apresentou o melhor desempenho entre todos os modelos e métricas, com o menor RMSE (3.257) e o menor MAPE (4.235%), além de apresentar o maior MDA (0.480), com superioridade simultânea em erro absoluto, erro percentual e acerto direcional.

No horizonte de 7 dias, o PWFTS manteve bom desempenho, destacando-se no RMSE (4.416), mas com MDA (0.401) inferior ao dos modelos HOFTS e WHOFTS, enquanto o WHOFTS obteve o menor MAPE (5.495%). Já no horizonte de 30 dias, o HOFTS apresentou o menor RMSE (8.186), indicando melhor ajuste absoluto ao longo prazo, enquanto os demais modelos apresentaram desempenho próximo entre si.

Assim, o PWFTS foi o modelo dominante no horizonte de 1 dia para o ano de 2014, apresentando superioridade clara nas três métricas avaliadas.

A Tabela 6 apresenta os resultados das métricas de avaliação dos modelos HOFTS, WHOFTS e PWFTS para os horizontes de previsão de 1, 7 e 30 dias no ano de 2021. Foram consideradas as métricas MAPE, MDA e RMSE.

Tabela 6. Métricas de avaliação dos modelos fuzzy para diferentes horizontes de previsão no ano de 2021.

Horizonte	Modelo	RMSE	MAPE (%)	MDA
1 dia	HOFTS	3.857	5.074	0.430
	WHOFTS	3.722	4.564	0.460
	PWFTS	3.450	4.579	0.396
7 dias	HOFTS	4.887	6.180	0.332
	WHOFTS	4.849	6.131	0.384
	PWFTS	4.522	5.903	0.445
30 dias	HOFTS	8.251	9.229	0.361
	WHOFTS	8.271	9.276	0.435
	PWFTS	8.294	9.448	0.320

Por fim, para o ano de 2021 os modelos apresentam desempenho competitivo ao longo dos horizontes de 1, 7 e 30 dias, com variações entre as métricas avaliadas. No horizonte de 1 dia, o WHOFTS apresentou melhor desempenho em MAPE (4.564%) e MDA (0.460), enquanto o PWFTS obteve o menor RMSE (3.450).

No horizonte de 7 dias, o PWFTS se destacou com os melhores resultados nas três métricas (RMSE = 4.522, MAPE = 5.903% e MDA = 0.445). Já no horizonte de 30 dias, o HOFTS apresentou os menores valores de RMSE (8.251) e MAPE (9.229%), enquanto o WHOFTS obteve o maior MDA (0.435).

Considerando todos os modelos e anos analisados, conclui-se que o modelo PWFTS apresenta o melhor desempenho global no horizonte de 1 dia, em que se destaca de forma mais consistente ao longo dos anos avaliados. Em 2014, o PWFTS apresentou seu melhor desempenho absoluto neste horizonte, com RMSE de 3.257, MAPE de 4.235% e MDA de 0.480, caracterizando o ponto mais forte e consistente do modelo nas três métricas avaliadas.

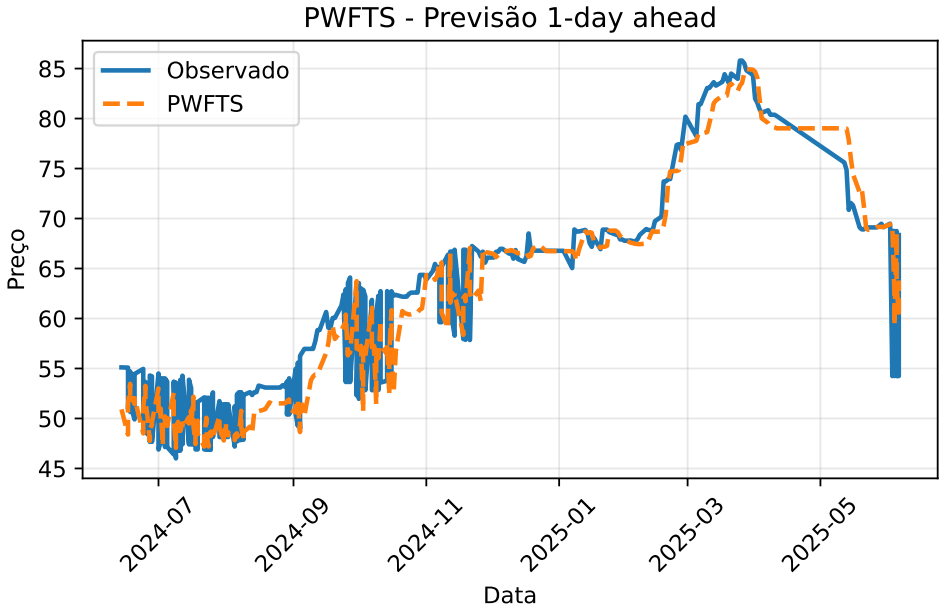

Figura 1. Valores observados versus valores preditos pelo modelo PWFTS, com horizonte de 1 dia e conjunto de treinamento com iníco em 2014.

A Figura 1 apresenta a série temporal do preço do milho, comparando os valores observados (linha azul) com os valores previstos pelo modelo PWFTS (linha laranja tracejada) no horizonte de previsão de 1 dia à frente.

Assim, observa-se que o modelo PWFTS consegue acompanhar o comportamento da série real, reproduzindo com boa precisão as principais tendências de alta e baixa ao longo do período analisado. O modelo captura bem o movimento de crescimento gradual dos preços, seguido por uma fase de elevação no início de 2025, além de refletir a posterior queda dos preços no final da série.

A série do milho apresenta um comportamento típico de mercado de commodities, com oscilações moderadas ao longo do tempo, alternando períodos de estabilidade (entre aproximadamente 50 e 70) com maior volatilidade. O modelo PWFTS acompanha essas variações de forma próxima, embora em alguns trechos apresente pequenas defasagens e suavizações em relação aos valores reais, o que é esperado em modelos de previsão fuzzy.

O gráfico indica que o preço do milho passou por um período inicial mais estável, seguido por uma tendência de alta gradual e posterior pico, antes de iniciar um movimento de correção e queda. O modelo PWFTS consegue representar bem essa dinâmica global.

4. Discussão

Os resultados obtidos evidenciam o potencial dos modelos baseados em lógica fuzzy para a previsão dos preços diários do milho em Minas Gerais, demonstrando sua capacidade de representar adequadamente séries temporais caracterizadas por volatilidade e incerteza. Embora a aplicação de Séries Temporais Fuzzy (FTS) à previsão de preços do milho ainda seja pouco explorada na literatura, estudos envolvendo outras commodities agrícolas têm reportado resultados promissores. Nesse contexto, modelos FTS têm sido apontados como alternativas eficientes aos métodos tradicionais de previsão, especialmente em cenários marcados por elevada variabilidade (Fatih, 2022). De forma semelhante, a incorporação de variáveis climáticas em modelos fuzzy tem contribuído para melhorar a qualidade das estimativas de produtividade e preço do trigo, evidenciando a flexibilidade dessa abordagem em aplicações agrícolas (Semeraro et al., 2019).

No caso específico do milho, diferentes metodologias de previsão têm sido propostas na literatura. Modelos lineares baseados em LASSO combinados com componentes extraídas de imagens de satélite apresentaram elevada precisão preditiva, superando abordagens convencionais utilizadas no mercado agrícola (Teste et al., 2024). De maneira complementar, redes neurais artificiais enriquecidas com informações provenientes dos mercados futuros podem aumentar a precisão das previsões de curto prazo em ambientes de alta volatilidade (Xu and Zhang, 2021). Esses resultados indicam que diferentes abordagens podem produzir previsões satisfatórias, enquanto os modelos fuzzy se destacam pela simplicidade estrutural e pela facilidade de interpretação dos resultados.

Outra característica relevante dos modelos fuzzy refere-se à sua capacidade de produzir previsões consistentes mesmo em conjuntos de dados relativamente reduzidos. As transformações não lineares inerentes à lógica fuzzy permitem extrair informações adicionais da série temporal, reduzindo simultaneamente a necessidade de recursos computacionais elevados (Li, Liu and Hu, 2011). Essa característica pode representar uma vantagem prática em aplicações relacionadas ao mercado de commodities agrícolas, especialmente em contextos nos quais a disponibilidade de infraestrutura computacional é limitada.

De forma geral, os resultados obtidos neste estudo corroboram o potencial da lógica fuzzy como ferramenta para previsão de preços agrícolas, demonstrando desempenho competitivo

na modelagem dos preços do milho. Além disso, a combinação entre capacidade preditiva, interpretabilidade e baixo custo computacional sugere que essa abordagem pode constituir uma alternativa viável para apoiar processos de tomada de decisão por produtores, agentes de mercado e formuladores de políticas relacionadas ao setor agrícola.

5. Conclusão

Conclui-se que o modelo PWFTS apresentou o melhor desempenho global entre os modelos avaliados, com destaque no horizonte de previsão de 1 dia, onde demonstrou maior estabilidade e consistência ao longo dos anos analisados. Em particular, no ano de 2014, o modelo alcançou seu melhor resultado, apresentando os menores valores de RMSE e MAPE e o maior MDA, mostrando superioridade nas métricas de erro e acerto direcional.

Já o comportamento da série do preço do milho, observa-se um padrão típico de commodities, com oscilações ao longo do tempo, alternando períodos de estabilidade e fases de maior volatilidade, além de tendência de crescimento seguida por momentos de reversão. O modelo PWFTS conseguiu reproduzir a dinâmica da série, acompanhando suas principais variações.

Apesar de pequenas defasagens e suavizações pontuais, esperadas em modelos fuzzy de previsão, o modelo captura adequadamente a tendência geral do preço do milho. Dessa forma, o PWFTS se mostra eficiente na modelagem e previsão de curto prazo da série analisada.

References

Astore, L.M., 2022. Data-driven spatio-temporal modeling with cellular automata and fuzzy time series methods. Master's thesis. Universidade Federal de Minas Gerais. Belo Horizonte.

Bakhtiyorovna, M., O'Gli, A., Qizi, M., 2021. Nutritional and economic importance of corn .

Bergmeir, C., Costantini, M., Benítez, J.M., 2014. On the usefulness of cross validation for directional forecast evaluation. Computational Statistics & Data Analysis 76, 132–143. doi:10.1016/j.csda.2014.02.001.

Bouktif, S., Fiaz, A., Ouni, A., Serhani, M.A., 2020. Multi-sequence lstm-rnn deep learning and metaheuristics for electric load forecasting. Energies 13, 391. URL: https://www.mdpi.com/1996-1073/13/2/391, doi:10.3390/en13020391.

Choudhary, K., et al., 2025. A genetic algorithm optimized hybrid model for agricultural price forecasting based on vmd and lstm network. Scientific Reports 15, 94173. doi:10.1038/s41598-025-94173-0.

Drachal, K., Pawlowski, M.E., 2021. A review of the applications of genetic algorithms to forecasting prices of commodities. Economies 9, 6. doi:10.3390/economies9010006.

Fatih, C., 2022. Forecasting models based on fuzzy logic: An application on international coffee prices. Econometrics. Ekonometria. Advances in Applied Data Analysis 26. doi:10.15611/eada.2022.4.01.

Frantti, T., Mähönen, P., 2001. Fuzzy logic based forecasting model. Engineering Applications of Artificial Intelligence 14, 189–201. doi:10.1016/S0952-1976(00)00076-2.

Gaur, S., et al., 2024. Precision corn price prediction with advanced ml techniques, in: International Conference on Trends in Quantum Computing and Emerging Business Technologies, pp. 1–5. doi:10.1109/TQCEBT59414.2024.10545288.

Gu, S., Li, M., 2024. Forecasting corn futures prices using the lstm model. Financial Economics Research doi:10.70267/st9q2c95.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688. doi:10.1016/j.ijforecast.2006.03.001.

Justus, M., et al., 2024. Did the entry of the corn ethanol industry in brazil affect the relationship between domestic and international corn prices? GCB Bioenergy 16. doi:10.1111/gcbb.13181.

Kerchner, A., et al., 2024. Corn culture: From the first records to harvesting and storage. Colloquium Agrariae doi:10.5747/ca.2024.v20.h528.

Kumar, S., Kumar, P., Kumar, M., Kumar, A., 2019. The global maize agri-food system and sustainable development goals. Global Food Security 23, 100327. doi:10.1016/j.gfs.2019.100327.

Lehna, M., Scheller, F., Herwartz, H., 2022. Forecasting day-ahead electricity prices: a comparison of time series and neural network models taking external regressors into account. Energy Economics 106, 105742. doi:10.1016/j.eneco.2021.105742.

Li, D.C., Liu, C.W., Hu, S.C., 2011. A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. Artificial Intelligence in Medicine 52, 45–52. doi:10.1016/j.artmed.2011.02.001.

Lucas, P.O., Orang, O., Silva, P.C.L., Mendes, E.M.A.M., Guimarães, F.G., 2021a. A tutorial on fuzzy time series forecasting models: Recent advances and challenges. Learning and Nonlinear Models 19, 29–50.

Lucas, P.O., et al., 2021b. A tutorial on fuzzy time series forecasting models: Recent advances and challenges. Learning and Nonlinear Models 19, 29–50.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. Statistical and machine learning forecasting methods: Concerns and ways forward. PLOS ONE 13, e0194889. doi:10.1371/journal.pone.0194889.

Mallory, M.L., Avileis, F.G., 2022. The impact of brazil on global grain dynamics: A study on cross market volatility spillovers. Agricultural Economics 53, 231–245. doi:10.1111/agec.12693.

Orang, O., Silva, R., Silva, P.C.d.L.e., Guimarães, F.G., 2020. Solar energy forecasting with fuzzy time series using high-order fuzzy cognitive maps, in: Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8. doi:10.1109/FUZZ42860.2020.9177708.

Python Software Foundation, 2024. Python Language Reference, version 3.12.3. Wilmington. URL: https://www.python.org/. acessado em: 7 maio 2026.

Rivas, R., 2021. El maíz, fuente de cultura mesoamericana. KOOT , 44–53doi:10.5377/KOOT.V0I11.10737.

Semeraro, T., et al., 2019. Modelling fuzzy combination of remote sensing vegetation index for durum wheat crop analysis. Computers and Electronics in Agriculture 156, 684–692. doi:10.1016/j.compag.2018.12.027.

Silva, L., Sales Silva, J., Souza, W., Lima, L., Santos, R., 2020. Desenvolvimento da cultura do milho (zea mays l.): revisão de literatura. Diversitas Journal 5, 1636–1657. doi:10.17648/diversitas-journal-v5i3-869.

Song, Q., Chissom, B.S., 1993. Fuzzy time series and its models. Fuzzy Sets and Systems 54, 269–277.

Tanscheit, R., 2005. Sistemas fuzzy.

Teste, F., et al., 2024. Early forecasting of corn yield and price variations using satellite vegetation products. Computers and Electronics in Agriculture 221, 108962. doi:10.1016/j.compag.2024.108962.

Wang, B., Liu, X., Chi, M., Li, Y., 2024. Bayesian network based probabilistic weighted high-order fuzzy time series forecasting. Expert Systems with Applications 237, 121430.

Wang, J., Hu, X., 2021. Research on maize production efficiency and influencing factors in typical farms: Based on data from 12 maize-producing countries from 2012 to 2019. PLOS ONE 16, e0254423. doi:10.1371/journal.pone.0254423.

Wu, B., Wang, Z., Wang, L., 2024. Interpretable corn future price forecasting with multivariate time series. Journal of Forecasting doi:10.1002/for.3099.

Wu, J., Zhu, J., 2025. Risk transmission and resilience of china's corn import trade network. Foods 14. doi:10.3390/foods14081401.

Xiong, T., Li, C., Bao, Y., 2018. Forecasting agricultural commodity prices using hybrid and combination approaches .

Xu, X., Zhang, Y., 2021. Corn cash price forecasting with neural networks. Computers and Electronics in Agriculture 184, 106120. doi:10.1016/j.compag.2021.106120.

Yu, H.K., 2005. Weighted fuzzy time series models for taiex forecasting. Physica A: Statistical Mechanics and its Applications 349, 609–624.

Zadeh, L.A., 1965. Fuzzy sets. Information and Control 8, 338–353.

Züge, S.W., Coelho, L.S., 2024. High-order fuzzy time series based on fuzzy c-means and bayesian networks for multivariate time series forecasting. Applied Soft Computing 152, 111234. doi:10.1016/j.asoc.2023.111234.

ARTIGO 2 - Previsão dos Preços da Soja: Avaliação Comparativa de Modelos de Aprendizado de Máquina e séries temporais fuzzy

Previsão dos Preços da Soja: Avaliação Comparativa de Modelos de Aprendizado de Máquina e séries temporais fuzzy

Informações do artigo

Palavras-chave:
Lógica Fuzzy
Preços da Soja
Aprendizado de Máquina

Resumo

A soja é uma das principais commodities agrícolas globais, com grande relevância econômica, nutricional e industrial. Sua produção e seus preços são influenciados por fatores climáticos, ambientais e geopolíticos, tornando a previsão um desafio complexo. Nesse contexto, métodos de aprendizado de máquina e de *Fuzzy Time Series* (FTS) destacam-se por capturar não linearidades e incertezas, oferecendo maior precisão na previsão dos preços da soja. Os dados utilizados foram obtidos do World Bank – Commodity Markets, no período de 1960 a 2025, com frequência mensal e 791 observações, tendo o preço da soja como variável resposta. Após a obtenção dos dados, foram realizadas etapas de pré-processamento, incluindo a definição de defasagens, o particionamento em conjuntos de treinamento e teste e a definição das partições fuzzy. Também foram avaliadas diferentes configurações de treino, *lags* e partições fuzzy. Em seguida, foram aplicados modelos de FTS multivariada e de aprendizado de máquina, incluindo Floresta aleatória, *Support Vector Regression* (SVR), *XGBoost, LightGBM e Prophet*. As análises foram realizadas em *Python* e R, utilizando bibliotecas consolidadas para modelagem e avaliação preditiva. Os resultados mostraram que os modelos de lógica fuzzy, o *Weighted Multivariate FTS (WMVFTS)* com teste de 20% e 60 partições, apresentaram o melhor desempenho global, com menores erros e alta precisão direcional. O modelo *Prophet* superou os métodos tradicionais de aprendizado de máquina, enquanto estes exibiram maior erro e instabilidade. Os modelos FTS demonstraram maior capacidade de capturar não linearidades e incertezas da série de preços da soja. Conclui-se que os modelos de lógica fuzzy superaram os métodos tradicionais de aprendizado de máquina, sobretudo com 80% de treinamento e particionamento 60, capturando adequadamente a dinâmica histórica e a tendência recente de estabilização e leve recuperação do preço da soja.

1. Introdução

A soja é uma leguminosa originária do Leste da Ásia, cultivada há cerca de 3.000 a 5.000 anos, principalmente na China, onde serviu como importante fonte alimentar (Page, 1924). Antes de sua domesticação, era conhecida como *Glycine soja*, a soja selvagem que deu origem às variedades atuais, resultado de processos de seleção e aprimoramento conduzidos no Sul e, posteriormente, na região Central da China (Zhou, 2022). Ao longo do tempo, transformou-se em uma das culturas mais importantes do mundo, integrando cadeias agrícolas globais (Norberg and Deutsch, 2023).

A soja apresenta grande valor nutricional, tecnológico e fisiológico, pois seus peptídeos bioativos exercem funções antioxidantes, anticancerígenas e reguladoras, além de melhorarem a absorção de aminoácidos (Kim, Yang and Kim, 2021). No mercado mundial, constitui uma das commodities mais produzidas e processadas, majoritariamente destinada ao farelo e ao óleo (Lee, 2016). Sua produção global depende de fatores ambientais e microbiológicos e possui grande importância para a segurança alimentar e industrial (Hamza, 2024). Em nível internacional, a soja também lidera a transmissão de choques entre commodities agrícolas e energéticas, principalmente em contextos geopolíticos críticos, como a guerra entre Rússia e Ucrânia, que eleva os preços pela dinâmica de oferta e demanda (Martignone, 2024).

A rede global de comércio de soja tem aumentado, com forte concentração das exportações nos Estados Unidos, Brasil e Argentina, enquanto a China permanece como principal importadora (Ma, 2023). Eventos climáticos extremos afetam a oferta mundial e elevam os preços, gerando forte impacto no mercado chinês (Hu, 2025). Aspectos geopolíticos e ambientais também remodelam os fluxos internacionais e reforçam a centralidade do Brasil e dos Estados Unidos no sistema comercial

(Oliveira, 2024). Diante desse cenário, a previsão dos preços da soja demanda modelos estatísticos mais robustos para garantir maior precisão, sendo os modelos de aprendizado de máquina uma alternativa adequada para essa finalidade (Srichaiyan, Tippayawong and Boonprasope, 2025).

Estudos recentes apontam que modelos de aprendizado de máquina têm se destacado na previsão do preço e da produtividade da soja por sua maior capacidade de capturar relações não lineares e interações complexas entre as variáveis, que modelos tradicionais têm maior dificuldade em representar devido ao não atendimento de pressupostos como linearidade, estacionariedade e resíduos do tipo ruído branco. A crescente utilização dessas técnicas tem ampliado as possibilidades de previsão em sistemas agrícolas. Nesse contexto, Srichaiyan et al. (2025) observaram que modelos adaptativos, como o XGBoost, apresentaram desempenho superior aos de aprendizado profundo. Pesquisas de produtividade mostraram que a floresta aleatória (RF) manteve desempenho robusto mesmo com amostras reduzidas (Santos, 2021), enquanto Santos (2024) reforçaram a eficácia da RF, das Regressão por Vetores de Suporte (SVR) e das redes neurais em diferentes sistemas produtivos. Em complemento, abordagens fuzzy multivariadas ampliaram a capacidade preditiva ao integrar múltiplos fatores e reduzir erros de previsão (Liu, 2023).

A lógica fuzzy multivariada constitui uma alternativa adequada para lidar com incertezas e múltiplas covariáveis na previsão temporal, integrando informações linguísticas e produzindo estimativas estáveis (Mehta, 2021). Em séries temporais, modelos fuzzy multivariados apresentaram maior acurácia quando comparados a métodos tradicionais e a modelos convencionais de aprendizado de máquina, devido à capacidade de processar múltiplas entradas e saídas simultaneamente (Yazdanbakhsh and Dick, 2017).

Este trabalho tem por objetivo avaliar a aplicação de modelos de aprendizado de máquina, incluindo XGBoost, RF, (SVR) e outros métodos correlatos, além de abordagens de lógica fuzzy multivariada, na previsão dos preços mensais da soja. Examina-se a capacidade desses métodos de incorporar múltiplas covariáveis, lidar com incertezas e capturar padrões não lineares do mercado. Após o desenvolvimento dos modelos, será realizada uma comparação sistemática entre eles para identificar aquele que apresenta o melhor desempenho preditivo no contexto analisado.

2. Metodologia

2.1. Base de dados e pré-processamento

O conjunto de dados utilizado neste estudo foi obtido no portal World Bank – Commodity Markets e compreende observações mensais entre janeiro de 1960 e novembro de 2025, totalizando 791 registros. A variável de interesse corresponde ao preço internacional da soja, expresso em dólares americanos (US$), enquanto as demais variáveis foram consideradas potenciais preditoras. Entre elas estão os preços de commodities agrícolas, energéticas e pecuárias, incluindo ureia, milho, trigo, cevada, óleo de soja, farelo de soja, óleo de palma, petróleo Brent, petróleo bruto, petróleo Dubai, gás natural, frango e carne bovina. Durante a etapa inicial de inspeção dos dados, valores ausentes foram identificados e tratados.

Como se trata de uma série temporal, as observações foram mantidas em sua sequência cronológica original. Posteriormente, foram geradas variáveis auxiliares a partir da própria série de preços da soja, incluindo valores defasados, médias móveis e componentes associados à sazonalidade. Essas variáveis foram incorporadas aos modelos com o objetivo de representar características temporais relevantes, tais como persistência, tendência e comportamento sazonal.

A avaliação dos modelos foi realizada por meio da separação da base em conjuntos de treinamento e teste, considerando as proporções 80/20 e 90/10. Para preservar a estrutura temporal

dos dados, as observações mais antigas foram destinadas ao treinamento, enquanto as observações mais recentes foram utilizadas na validação preditiva. Essa estratégia impede que informações futuras sejam utilizadas durante o ajuste dos modelos.

Cabe destacar que os algoritmos de Aprendizado de Máquinas empregados neste estudo não incorporam naturalmente a dependência temporal existente na série, exigindo a inclusão explícita de atributos temporais derivados dos dados observados. Por outro lado, os modelos de Séries Temporais Fuzzy Multivariadas realizam essa representação diretamente em sua estrutura de modelagem, permitindo considerar a dinâmica temporal durante o processo de previsão.

2.2. Modelos

A lógica fuzzy univariada é uma forma de modelagem em única variável de entrada representada por conjuntos fuzzy, permitindo tratar incertezas por meio de graus de pertinência em vez de valores crisp, sendo aplicada principalmente em séries temporais fuzzy para previsão Lucas, Orang, Silva, Mendes and Guimarães (2021); Züge and Coelho (2024); Wang, Liu, Chi and Li (2024). Já as séries temporais multivariadas são estudadas devido à escassez de métodos capazes de capturar as interdependências entre múltiplas variáveis ao longo do tempo. As abordagens baseadas em lógica fuzzy complexa têm apresentado elevada precisão em problemas de previsão multivariada (Yazdanbakhsh and Dick, 2017). Além disso, diferentes técnicas estatísticas e de aprendizado de máquinas podem apresentar desempenhos diferentes em cenários multivariados, mostrando a importância da investigação de métodos adequados para estruturas complexas de dados (Mariño, Arrieta-Prieto and Calderón, 2025). Revisões recentes também destacam avanços metodológicos relacionados à aplicação de algoritmos de aprendizado de máquinas em séries temporais de maior dimensionalidade (Masini, Medeiros and Mendes, 2020).

2.2.1. Fuzzy Time Series

Os modelos de Séries Temporais Fuzzy (FTS) tratam incertezas e imprecisões de séries temporais mapeando dados numéricos em conjuntos fuzzy e estruturando Relações Lógicas Fuzzy (FLR) para descrever a dependência temporal entre observações (Song and Chissom, 1993; Chen, 1996).

Nos modelos de alta ordem FTS (High Order FTS - HOFTS), a modelagem incorpora informações de múltiplas defasagens históricas, seguindo a estrutura geral:

$$F(t-1), F(t-2), \ldots, F(t-\Omega) \rightarrow F(t),$$

em que: Ω representa a ordem estabelecida para o modelo.

O modelo Weighted High Order Fuzzy Time Series (WHOFTS) avança ao ponderar essas relações lógicas conforme a frequência relativa de suas ocorrências, utilizando a equação:

$$\omega_i = \frac{\#A_i}{\#RHS},$$

em que: $\#A_i$ é o número de ocorrências da i-ésima relação lógica fuzzy específica e $\#RHS$ indica o total de relações que compartilham o mesmo antecedente histórico.

Os modelos de Séries Temporais Fuzzy Multivariadas (MVFTS) geram previsões por meio da integração simultânea de múltiplos fatores e variáveis quantificadas por partições fuzzy isoladas (Severiano, Silva, Cohen and Guimarães, 2021). O processo inicia-se com a conversão dos dados contínuos em conjuntos fuzzy mediante o particionamento do universo de discurso de cada variável explicativa. Na sequência, as dependências temporais e interações entre as variáveis são

representadas por Relações Lógicas Fuzzy Multivariadas (MFLR) de ordem p e dimensão d, estruturadas por:

$$(F_1(t-p), \ldots, F_d(t-p)), \ldots, (F_1(t-1), \ldots, F_d(t-1)) \rightarrow F_1(t), F_2(t), \ldots, F_d(t),$$

em que: o primeiro subíndice denota a i-ésima variável explicativa ($i = 1, \ldots, d$) e o termo subtrativo indica a defasagem temporal de até p períodos, enquanto $F_i(t)$ corresponde ao conjunto fuzzy estimado para a variável de interesse no instante t. Essa arquitetura possibilita modelar relações não lineares e dependências dinâmicas de forma interpretável.

Para cenários de maior complexidade, a extensão ponderada desse modelo, o Weighted Multivariate FTS (WMVFTS) avança ao calibrar pesos específicos para os antecedentes fuzzy, refletindo com maior precisão o nível de influência e relevância de cada variável preditora no processo de inferência.

Por fim, o modelo granular WMVFTS (GWMVFTS) aperfeiçoa essa dinâmica ao adotar Grânulos de Informação Fuzzy (FIG), nos quais cada vetor multivariado compõe um conjunto fuzzy associado (Züge and Coelho, 2024). Essa abordagem consolida padrões temporais estruturados que tornam o sistema capaz de gerar previsões pontuais, intervalares e probabilísticas, incorporando e modelando diretamente as incertezas intrínsecas à série temporal.

2.2.2. Floresta Aleatória

O Floresta Aletória é um método de aprendizado por ensemble na construção de múltiplas árvores de decisão a partir de amostras bootstrap do conjunto de treinamento, com seleção aleatória de subconjuntos de variáveis em cada divisão para garantir diversidade entre os modelos (Salman, Kalakech and Steiti, 2024; Abdulkareem and Abdulazeez, 2021; Liu, Wang and Zhang, 2012; Breiman, 2001). Essa estratégia reduz a variância do modelo e diminui problemas de sobreajuste, sendo utilizada principalmente em problemas de regressão.

Para problemas de regressão, a estimativa final pode ser obtida pela média das predições fornecidas pelas B árvores de decisão, conforme definido por:

$$\hat{y}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x),$$

em que: $T_b(x)$ representa a previsão produzida pela b-ésima árvore da floresta e B corresponde ao número total de árvores. Durante o treinamento, o desempenho do modelo pode ser monitorado pelo erro *out-of-bag* (OOB), calculado a partir das observações não utilizadas na construção de cada árvore (Breiman, 2001).

2.2.3. Support Vector Regression

A Support Vector Regression (SVR) apresenta-se como uma extensão das Máquinas de Vetores de Suporte para problemas de regressão, podendo assim modelar relações lineares e não lineares entre variáveis explicativas e uma variável resposta contínua por meio de funções kernel (Zhang and O'Donnell, 2020). O método apresenta uma alta capacidade de generalização e pode ser aplicado em problemas com estruturas não lineares e múltiplas covariáveis (Valente and Maldonado, 2020). Portanto, sua aplicação em séries temporais multivariadas permite incorporar no mesmo momento distintas fontes de informação ao processo de previsão (Wang and Zhang, 2023).

A função de regressão é dada por:

$$f(x) = \langle w, x \rangle + b = w^T x + b,$$

em que: w representa o vetor de pesos e b o viés (Awad and Khanna, 2015).

O ajustamento do modelo é realizado por meio da função de perda ϵ-insensível:

$$L_\epsilon(y, f(x)) = \begin{cases} 0, & \text{se } |y - f(x)| \leq \epsilon, \\ |y - f(x)| - \epsilon, & \text{caso contrário.} \end{cases}$$

Essa formulação estabelece uma região de tolerância em torno da função estimada, ignorando erros inferiores ao limite ϵ e penalizando apenas desvios que excedem esse valor, característica que contribui para a robustez do método frente à presença de ruídos nos dados (Awad and Khanna, 2015).

2.2.4. Extreme Gradient Boosting

O Extreme Gradient Boosting (XGBoost) é um algoritmo que utiliza a técnica de *gradient boosting*, muito aplicado em previsão devido à sua capacidade de modelar relações complexas e incorporar múltiplas covariáveis (Zhang, 2021). O método constrói sequencialmente árvores de decisão, em que cada nova árvore busca corrigir os erros cometidos pelas árvores previamente ajustadas (Wang, Cheng and Dong, 2022; Noorunnahar, Chowdhury and Mila, 2023; Zhang, 2025).

“‘latex De acordo com Chen and Guestrin (2016), considerando a adição da árvore f_t na iteração t, a função objetivo é definida por:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t).$$

em que l é uma função de perda diferenciável e $\Omega(f_t)$ é o termo de regularização associado à complexidade da árvore. Para uma árvore f, esse termo é definido por:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2.$$

em que T representa o número de folhas da árvore e w representa o vetor de pesos das folhas.

Para otimizar a função objetivo, Chen and Guestrin (2016) utiliza uma aproximação de Taylor de segunda ordem em torno da predição anterior:

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t),$$

em que g_i e h_i correspondem, respectivamente, às derivadas de primeira e segunda ordem da função de perda, definidas por:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$$

e

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right).$$

Removendo os termos constantes, que não dependem de f_t, obtém-se a função objetivo simplificada:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t).$$

Definindo $I_j = \{i \mid q(x_i) = j\}$ como o conjunto de instâncias pertencentes à folha j, a função objetivo pode ser reescrita, após a expansão do termo de regularização, como:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T.$$

2.2.5. Prophet

O Prophet é um modelo de regressão aditiva desenvolvido para a previsão de séries temporais, cuja especificação pode ser estendida pela inclusão de covariáveis externas (*regressors*), permitindo incorporar informações adicionais à modelagem da variável de interesse e, consequentemente, à sua previsão (Huang, 2022). O método apresenta estrutura flexível e interpretável, sendo adequado para séries que apresentam componentes de tendência e sazonalidade bem definidas (Oo and Phyu, 2020).

A estrutura básica do modelo é representada por:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t,$$

em que: $g(t)$ representa a componente de tendência de longo prazo, $s(t)$ descreve os padrões sazonais recorrentes, $h(t)$ corresponde aos efeitos associados a eventos específicos ou feriados e ϵ_t representa o termo de erro aleatório (Shakeel, Chong and Wang, 2023).

Essa decomposição aditiva permite que o modelo capture diferentes padrões temporais de forma independente, favorecendo a interpretação dos resultados e proporcionando robustez na modelagem de séries temporais com mudanças estruturais e sazonalidades complexas (Taylor and Letham, 2018).

2.2.6. LightGBM

O Light Gradient Boosting Machine (LightGBM) é um algoritmo de aprendizado de máquina baseado em *gradient boosting* que utiliza um conjunto de árvores de decisão para modelar relações complexas entre covariáveis e a variável alvo (Yan, Wang, Liu and Zhang, 2021). O método tem sido amplamente empregado em previsão devido à sua eficiência computacional, capacidade de lidar com conjuntos de dados de alta dimensionalidade e elevado desempenho preditivo em tarefas multivariadas (Kopyt, Piotrowski and Baczyński, 2024).

A estrutura do modelo fundamenta-se em uma expansão aditiva de funções, conforme formulado por Friedman (2001):

$$F\left(x; \{\beta_m, a_m\}_1^M\right) = \sum_{m=1}^{M} \beta_m h(x; a_m),$$

em que: $\{\beta_m, a_m\}_{m=1}^M$ representa o conjunto de parâmetros, β_m é o coeficiente associado à m-ésima função e $h(x; a_m)$ denota a função base parametrizada por a_m. O caso de particular interesse ocorre quando cada $h(x; a_m)$ é uma árvore de regressão, cujos parâmetros a_m correspondem às variáveis de divisão, aos pontos de corte e aos valores dos nós terminais.

Para dados finitos, Friedman (2001) considera uma abordagem denominada "greedy-stagewise". Para $m = 1, 2, \ldots, M$, os parâmetros são determinados iterativamente por:

$$(\beta_m, a_m) = \arg\min_{\beta, a} \sum_{i=1}^{N} L\left(y_i, F_{m-1}(x_i) + \beta h(x_i; a)\right),$$

seguida da atualização do modelo:

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m).$$

Como a distribuição conjunta de (y, x) é representada por uma amostra finita, o gradiente pode ser aproximado nos pontos observados por:

$$-g_m(x_i) = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}.$$

No algoritmo de Gradient Boosting, esses valores são utilizados como resposta para o ajuste da função base $h(x; a_m)$ a cada iteração (Friedman, 2001).

2.3. Métricas de Avaliação das Previsões

As métricas de avaliação são medidas quantitativas que mensuram o grau de concordância entre valores observados e previstos, permitindo avaliar a precisão, a confiabilidade e o desempenho de modelos de previsão. Diferentes métricas quantificam aspectos específicos da qualidade preditiva, auxiliando na comparação e validação de modelos (Owens, 2018).

Neste estudo, o desempenho dos modelos foi avaliado por meio das métricas Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), Mean Directional Accuracy (MDA), Normalized Root Mean Square Error (nRMSE), Mean Absolute Scaled Error (MASE) e Symmetric Mean Absolute Percentage Error (sMAPE).

As métricas de acurácia preditiva utilizadas nesta avaliação seguem as definições e padronizações propostas por Hyndman and Koehler (2006a). A Raiz do Erro Quadrático Médio (RMSE) é expressa por:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}$$

em que: y_t representa o valor observado no instante t, $\hat{y}_t$ denota a previsão pontual e N é o número de observações.

Para avaliar a acurácia dos modelos em séries temporais de diferentes escalas, utiliza-se a Raiz do Erro Quadrático Médio Normalizada (nRMSE), conforme formalizado por Taieb, Bontempi, Atiya and Sorjamaa (2012). O nRMSE escala o erro absoluto do modelo dividindo a raiz do erro quadrático médio pela amplitude (diferença entre os valores máximo e mínimo) dos valores observados na série temporal:

$$nRMSE = \frac{RMSE}{y_{\max} - y_{\min}}$$

em que: $y_{\max}$ corresponde ao maior valor observado e $y_{\min}$ ao menor valor observado na série. Valores menores de NRMSE indicam maior precisão preditiva do modelo independente da escala dos dados originais.

Para comparações entre diferentes séries temporais, Hyndman and Koehler (2006a) recomendam o Erro Absoluto Médio Escalonado (MASE), definido pela razão entre o erro absoluto do modelo e o erro absoluto médio de uma previsão ingênua do tipo random walk:

$$MASE = \frac{\frac{1}{N}\sum_{t=1}^{N} |y_t - \hat{y}_t|}{\frac{1}{n-1}\sum_{t=2}^{n} |y_t - y_{t-1}|}$$

em que: n representa o tamanho do conjunto de treinamento e N representa o número de observações no horizonte de previsão (Hyndman and Koehler, 2006b).

O Erro Percentual Absoluto Médio (MAPE) é expresso por:

$$\text{MAPE} = \frac{100\%}{N}\sum_{t=0}^{N-1}\frac{\left|y_i - \hat{y}_i\right|}{\left|y_i\right|},$$

com: y_t, $\hat{y}_t$ e N como definidos anteriormente Hyndman and Koehler (2006b).

O Erro Percentual Absoluto Médio Simétrico (Symmetric Mean Absolute Percentage Error – sMAPE) é empregado na avaliação de modelos de previsão por utilizar simultaneamente os valores observados e previstos em sua normalização. Essa característica reduz a influência da escala dos dados e permite comparações entre diferentes séries temporais. Valores menores de sMAPE indicam maior acurácia das previsões (Godahewa, Bergmeir, Webb, Montero-Manso and Hyndman, 2021):

$$sMAPE = \frac{100\%}{h}\sum_{t=1}^{h}\frac{|\hat{y}_t - y_t|}{\frac{|y_t|+|\hat{y}_t|}{2}},$$

em que: h representa o horizonte de previsão.

A Acurácia Direcional Média, também conhecida pela sigla MDA (*Mean Directional Accuracy*), é uma métrica utilizada para avaliar a capacidade de um modelo de prever corretamente a direção da variação de uma série temporal. Com base em Ezzat, Malek, AbdelKader et al. (2026), a MDA é definida como:

$$\text{MDA} = \frac{1}{N}\sum_{t=1}^{N} d_t \times 100$$

em que a variável indicadora de acerto direcional é dada por:

$$d_t = \begin{cases} 1, & \text{se } (y_t - y_{t-1})(\hat{y}_t - y_{t-1}) \geq 0, \\ 0, & \text{caso contrário.} \end{cases}$$

sendo y_t o valor observado no instante t, $\hat{y}_t$ o valor previsto pelo modelo e N o número total de observações.

2.4. Recursos computacionais

As análises estatísticas foram realizadas nas linguagens de programação Python (versão 3.12.3) e R (versão 4.5.1). A modelagem em lógica fuzzy foi desenvolvida em Python, utilizando as bibliotecas Pandas para manipulação e tratamento dos dados, NumPy para operações numéricas e vetorização matricial, pyFTS para modelagem e previsão de séries temporais com lógica fuzzy, matplotlib.pyplot para geração de gráficos e visualizações, e sklearn.metrics para o cálculo das métricas de avaliação dos modelos.

As análises de aprendizado de máquina foram implementadas em R, no ambiente RStudio. Foram empregados pacotes consolidados na literatura de ciência de dados e séries temporais, o tidyverse e tidyr na organização e transformação dos dados. A análise temporal e a engenharia de atributos foram conduzidas com os pacotes timetk e tsibble, enquanto a modelagem preditiva utilizou bonsai, modeltime, lightgbm, xgboost e Prophet, permitindo a aplicação de diferentes abordagens de modelagem aprendizado de máquinas.

A seleção das variáveis mais importantes foi realizada com o pacote Boruta, e a organização visual contou com o auxílio do patchwork. Todas as análises foram executadas em um ambiente computacional de alto desempenho, composto por processador Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz, com 40 threads, 78 GiB de memória RAM e sistema gráfico Matrox Electronics Systems Ltd. MGA G200e [Pilot] ServerEngines (SEP1), assegurando eficiência e reprodutibilidade dos experimentos.

3. Resultados

Os modelos de Lógica Fuzzy foram treinados e avaliados considerando duas proporções entre os conjuntos de treinamento e teste: 80% dos dados destinados ao treinamento e 20% ao teste, totalizando 159 observações no conjunto de teste, e 90% destinados ao treinamento e 10% ao teste, correspondendo a 80 observações no conjunto de teste.

A Figura 1 apresenta a divisão da série temporal considerando 80% dos dados para treinamento e 20% para teste. A separação entre os conjuntos ocorre nas proximidades do ano de 2013.

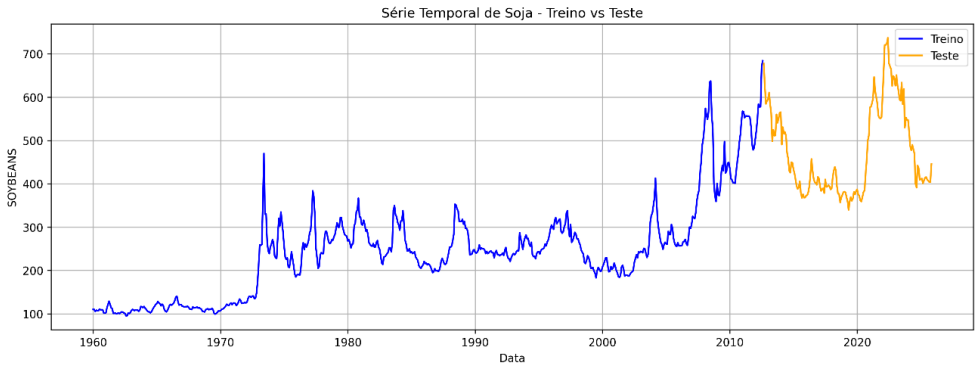

Figura 1. Divisão da série temporal em 80% para treinamento e 20% para teste.

A Figura 2 apresenta a divisão da série temporal considerando 90% dos dados para treinamento e 10% para teste.

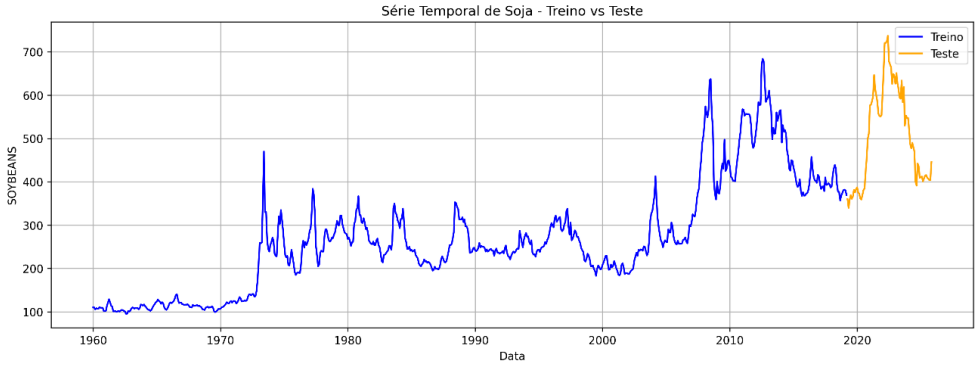

Figura 2. Divisão da série temporal em 90% para treinamento e 10% para teste.

Para os algoritmos de aprendizado de máquina, a modelagem foi realizada utilizando 20% dos dados para teste, configuração que apresentou melhor equilíbrio entre desempenho preditivo e estabilidade dos modelos durante as análises preliminares.

A Tabela 1 apresenta os valores obtidos para as métricas de avaliação do desempenho preditivo, para os modelos de aprendizado de máquina e para os modelos baseados em Lógica Fuzzy, considerando diferentes configurações do parâmetro de pertinência (PF). Os melhores resultados foram identificados com base nos menores valores das métricas MAPE, MASE, sMAPE, RMSE e nRMSE, bem como no maior valor obtido para a métrica MDA.

Tabela 1: Desempenho preditivo dos modelos avaliados no conjunto de teste

Modelo	MAPE	MASE	sMAPE	RMSE	MDA	nRMSE
KERNLAB	1.300	0.353	0.013	8.020	0.910	0.017
LIGHTGBM	4.920	1.420	0.049	35.100	0.568	0.074
PROPHET	0.066	0.017	0.001	3.360	1.000	0.007
RANGER	5.320	1.430	0.052	30.600	0.671	0.065
XGBOOST	4.460	1.360	0.045	35.500	0.606	0.075
MVFTS (PF 0.8 – F20)	1.116	0.259	0.545	9.348	0.918	0.020
Weighted MVFTS (PF 0.8 – F20)	1.082	0.252	0.529	9.047	0.930	0.019
MVFTS (PF 0.8 – F40)	0.058	0.013	0.029	0.924	0.981	0.002
Weighted MVFTS (PF 0.8 – F40)	0.058	0.013	0.029	0.924	0.981	0.002
MVFTS (PF 0.8 – F60)	0.008	0.002	0.004	0.303	0.994	0.001
Weighted MVFTS (PF 0.8 – F60)	0.008	0.002	0.004	0.303	0.994	0.001
MVFTS (PF 0.9 – F20)	1.273	0.264	0.614	11.671	0.911	0.023

Modelo	MAPE	MASE	sMAPE	RMSE	MDA	nRMSE
Weighted MVFTS (PF 0.9 – F20)	1.038	0.219	0.505	9.375	0.911	0.019
MVFTS (PF 0.9 – F40)	0.185	0.036	0.091	2.813	0.962	0.006
Weighted MVFTS (PF 0.9 – F40)	0.183	0.035	0.090	2.749	0.975	0.005
MVFTS (PF 0.9 – F60)	0.031	0.006	0.016	0.593	1.000	0.001
Weighted MVFTS (PF 0.9 – F60)	0.031	0.006	0.016	0.593	1.000	0.001

Nota: MAPE: Erro Percentual Absoluto Médio; MASE: Erro Escalonado Absoluto Médio; sMAPE: Erro Percentual Absoluto Médio Simétrico; RMSE: Raiz do Erro Quadrático Médio; MDA: Acurácia Direcional Média; nRMSE: Raiz do Erro Quadrático Médio Normalizada; KERNLAB: Modelos de Aprendizado de Máquina Baseados em Kernel (Máquinas de Vetores de Suporte); LIGHTGBM: Máquina de Impulsionamento por Gradiente Leve; PROPHET: Modelo de Previsão Prophet do Facebook; RANGER: Implementação Rápida de Florestas Aleatórias; XGBOOST: Impulsionamento por Gradiente Extremo; MVFTS: Séries Temporais Fuzzy Multivariadas; Weighted MVFTS: Séries Temporais Fuzzy Multivariadas Ponderadas; PF: Fator de Perturbação; F: Número de Conjuntos Fuzzy. Os valores destacadas em azul representam o melhor desempenho para cada métrica.

Observa-se que os algoritmos de aprendizado de máquina apresentam desempenho satisfatório (MAPE $<$ 10%), com erros superiores aos dos modelos baseados em lógica fuzzy. Os modelos RANGER, LIGHTGBM, e XGBOOST, nesta ordem, apresentaram os maiores valores para as métricas MAPE e sMAPE e os menores valores de MDA, sendo observados valores superiores a 1 para a métrica MASE. Estes algoritmos demonstraram comportamentos semelhantes, refletindo maior sensibilidade a ruídos e possíveis dificuldades em lidar com séries mais longas. O modelo KERNLAB apresentou resultados moderados, com MAPE de 1,3 e MASE de 0,353. Esses valores indicam boa capacidade de ajuste local, mas ainda com limitações na captura de variações mais complexas da série. De forma geral, os modelos de ML apresentaram valores de MDA entre 0,568 e 0,910, o que demonstra capacidade razoável de prever a direção da série, porém com erros absolutos e relativos ainda elevados. Ou seja, apesar de apresentarem bom desempenho em tarefas de regressão, esses modelos não capturaram integralmente o comportamento específico da série de dados analisada.

O Prophet apresentou desempenho expressivamente superior aos modelos de aprendizado de máquina, com MAPE de 0,066, RMSE de 3,360 e MASE de 0,017. Além disso, o MDA de 1,000 indica acerto perfeito na direção das variações da série, evidenciando elevado poder preditivo. Os resultados mostram que a modelagem explícita de tendência e sazonalidade contribuiu para a redução dos erros.

Os modelos baseados em lógica fuzzy apresentaram os melhores resultados globais, especialmente nas configurações com níveis de particionamento fuzzy (F40 e F60) e para a proporção de 80% dos dados no conjunto de treinamento. Para esta proporção, observa-se melhoria considerável dos resultados, especialmente para F60, em que o RMSE cai para 0,303, o MAPE para 0,008 e o MDA atinge 0,994, evidenciando elevada precisão tanto em magnitude quanto na direção da série.

O melhor desempenho absoluto é observado tanto no MVFTS quanto no Weighted MVFTS. Nessas ocasiões, os modelos atingem RMSE de 0,303, MAPE de 0,008, sMAPE de 0,004 e MDA de 0,994, evidenciando elevada capacidade preditiva e baixos níveis de erro. A versão ponderada (Weighted MVFTS) apresenta resultados levemente superiores em estabilidade.

Ao comparar os algoritmos de aprendizado de máquina e os métodos baseados em lógica fuzzy, observa-se uma ampla variação de desempenho. Os modelos de aprendizado de máquina tradicionais apresentaram os maiores erros e menor consistência, sendo indicados como referência, mas não como soluções ótimas para o problema em questão. O Prophet apresenta desempenho preditivo superior aos demais algoritmos de aprendizado de máquina, ao combinar modelagem

temporal com decomposição de tendência e sazonalidade. No entanto, seu desempenho ainda foi inferior aos melhores modelos de lógica fuzzy. A lógica fuzzy, por meio dos modelos nas configurações MVFTS e Weighted MVFTS com PF = 0,8 e F60, demonstrou superioridade em praticamente todas as métricas utilizadas, demonstrando melhor equilíbrio entre precisão, estabilidade e capacidade de previsão da direção da série.

A figura 3 apresenta a comparação entre os modelos de aprendizado de máquina e o modelo *Prophet* na previsão dos preços da soja. Observa-se que o modelo *PROPHET* apresentou as melhores previsões para a série observada praticamente em todo o período analisado, reproduzindo com maior precisão os movimentos de alta e baixa dos preços. O modelo SVR também demonstrou bom desempenho na captura das oscilações mais acentuadas observadas entre 2021 e 2023, mantendo-se próximo aos dados reais em grande parte do horizonte temporal. Já os algoritmos RF, *LIGHTGBM* e *XGBoost* apresentaram maior suavização da série e desvios mais evidentes em períodos de elevada volatilidade, resultando em menor capacidade de representar os extremos observados. De forma prática, os resultados corroboram com as métricas apresentadas na Tabela 1, mostrando a superioridade do modelo *Prophet* e o desempenho competitivo do modelo SVR na previsão dos preços da soja.

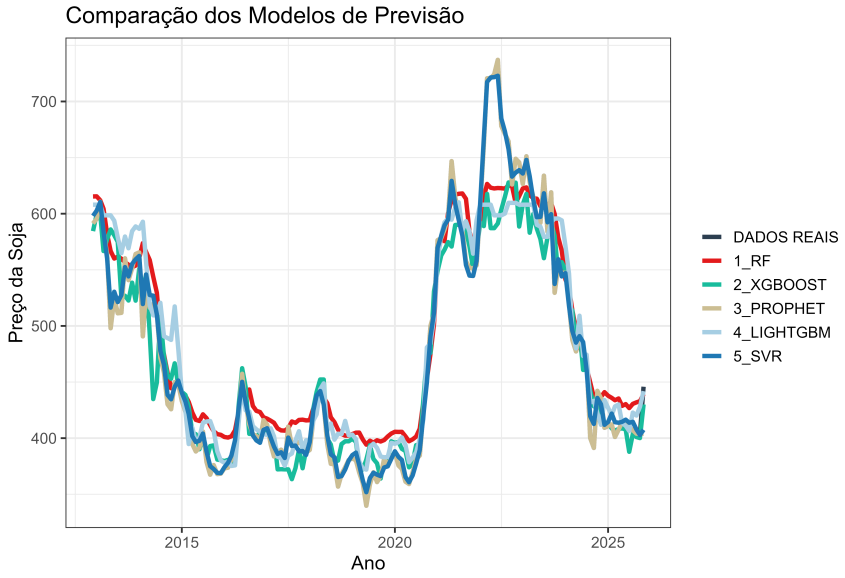

Figura 3. Valores observados e preditos pelos algoritmos de aprendizado de máquina e pelo Prophet no conjunto de teste, correspondente a 20% do conjunto de dados.

A Figura 4 apresenta o melhor modelo obtido na análise com 80% dos dados destinados ao treino e particionamento fuzzy igual a 60. Observa-se que a série inicia em um patamar constante e baixo até aproximadamente 1970, passando, em seguida, a apresentar uma tendência acompanhada por oscilações, com alternância entre picos de alta e de baixa. Em relação às previsões, nota-se um comportamento de queda semelhante ao observado no ano de 2023, seguido por uma estabilização em torno da faixa de 400. Por fim, identifica-se um leve início de elevação dos preços já no horizonte de previsão.

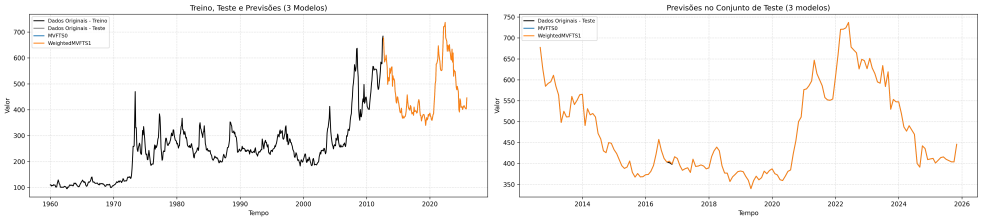

Figura 4. Previsões obtidas pelo modelo MVFTS no conjunto de teste, correspondente a 20% do conjunto de dados, com 60 partições.

4. Discussão

Para o modelo Prophet e os de aprendizado de máquina, as avaliações foram realizados utilizando 20% do conjunto de dados, enquanto os 80% restantes foram destinados ao treinamento. Essa escolha está de acordo com a literatura, que aponta a divisão 80%/20% como uma prática amplamente adotada em estudos de previsão de séries temporais, por permitir uma quantidade suficiente de dados para o treinamento do modelo e uma amostra adequada para a validação do desempenho fora da amostra Karim, Majumdar, Darabi and Chen (2017).

Ao comparar as abordagens de aprendizado de máquina tradicional e o modelo Prophet, observa-se que os modelos de aprendizado de máquina, apesar de eficientes em regressão, apresentam limitações na captura de padrões estruturais de longo prazo, como tendência e sazonalidade, quando aplicados de forma isolada Makridakis, Spiliotis and Assimakopoulos (2018). Nesse contexto, o Prophet representa um avanço ao integrar decomposição temporal e componentes de modelagem estatística, melhorando a capacidade preditiva, embora ainda apresente limitações diante de séries altamente não lineares (Taylor and Letham, 2018).

Em contraposição, os modelos de lógica fuzzy apresentaram desempenho superior na maioria das métricas, principalmente nas configurações MVFTS e Weighted MVFTS com PF = 0,8 e F60. Esse resultado está em consonância com a literatura de *Fuzzy Time Series*, que destaca a eficácia desses modelos no tratamento de incertezas, não linearidades e transições graduais presentes em séries temporais reais (Song and Chissom, 1993; Chen, 1996). Assim, os achados reforçam que abordagens fuzzy são adequadas para problemas de previsão temporal complexos, superando abordagens tradicionais.

5. Conclusão

Conclui-se que, para a base de dados analisada, as abordagens baseadas em lógica fuzzy apresentaram desempenho superior aos modelos tradicionais de aprendizado de máquina, consolidando-se como alternativas mais eficazes para problemas de previsão temporal complexos. Os resultados indicam que o desempenho dos modelos fuzzy está diretamente associado à adoção de partições elevadas para o conjunto de treinamento e proporções reduzidas para o conjunto de teste, uma vez que cenários com 30% de dados destinados ao teste não se mostraram adequados.

O melhor desempenho foi obtido com 80% dos dados destinados ao treinamento e particionamento fuzzy igual a 60. Observa-se que o preço da soja permaneceu em níveis baixos e estáveis até aproximadamente 1970, passando posteriormente a apresentar oscilações com ciclos de alta e baixa. No horizonte de previsão, o preço apresenta uma queda semelhante à observada em 2023,

seguida por estabilização em torno do patamar de 400 e um leve sinal de recuperação, indicando boa capacidade do modelo em capturar a dinâmica e a variação do preço da soja ao longo do tempo.

Referências

Abdulkareem, N., Abdulazeez, A., 2021. Machine learning classification based on random forest algorithm: a review. Zenodo 5, 128–142. doi:10.5281/zenodo.4471118.

Awad, M., Khanna, R., 2015. Support vector regression, in: Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Springer, pp. 67–80.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Chen, S.M., 1996. Forecasting enrollments based on fuzzy time series. Fuzzy Sets and Systems 81, 311–319. doi:10.1016/0165-0114(95)00220-0.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA. pp. 785–794. doi:10.1145/2939672.2939785.

Ezzat, M., Malek, Y.M., AbdelKader, T., et al., 2026. Spatio-temporal epidemic forecasting with graph-based transformer. International Journal of Health Geographics 25, 33. URL: https://doi.org/10.1186/s12942-026-00476-4, doi:10.1186/s12942-026-00476-4.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29, 1189–1232. doi:10.1214/aos/1013203451.

Godahewa, R., Bergmeir, C., Webb, G.I., Montero-Manso, P., Hyndman, R.J., 2021. Monash time series forecasting repository. Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks doi:10.48550/arXiv.2105.06643.

Hamza, M.e.a., 2024. Global impact of soybean production: a review. Asian Journal of Biochemistry, Genetics and Molecular Biology 16. doi:10.9734/ajbgmb/2024/v16i2357.

Hu, X.e.a., 2025. The impact of extreme weather events on global soybean markets and china's imports. Journal of Agricultural Economics doi:10.1111/1477-9552.12632.

Huang, Q., 2022. Forecasting stock prices using multi-macroeconomic regressors based on the facebook prophet model. BCP Business & Management 25. doi:10.54691/bcpbm.v25i.1762.

Hyndman, R.J., Koehler, A.B., 2006a. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688. doi:10.1016/j.ijforecast.2006.03.001.

Hyndman, R.J., Koehler, A.B., 2006b. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688.

Karim, F., Majumdar, S., Darabi, H., Chen, S., 2017. Lstm fully convolutional networks for time series classification. IEEE access 6, 1662–1669.

Kim, I., Yang, W., Kim, C., 2021. Beneficial effects of soybean-derived bioactive peptides. International Journal of Molecular Sciences 22. doi:10.3390/ijms22168570.

Kopyt, M., Piotrowski, P., Baczyński, D., 2024. Short-term energy generation forecasts at a wind farm—a multi-variant comparison of gradient-boosted decision tree models. Energies 17. doi:10.3390/en17236194.

Lee, T.e.a., 2016. Major factors affecting global soybean and products trade projections. Amber Waves , 1–1doi:10.22004/ag.econ.244273.

Liu, Y., Wang, Y., Zhang, J., 2012. New machine learning algorithm: Random forest, in: International Conference on Information Computing and Applications, Springer. pp. 246–252. doi:10.1007/978-3-642-34062-8_32.

Liu, Z., 2023. Forecasting stock prices based on multivariable fuzzy time series. AIMS Mathematics doi:10.3934/math.2023643.

Lucas, P.O., Orang, O., Silva, P.C.L., Mendes, E.M.A.M., Guimarães, F.G., 2021. A tutorial on fuzzy time series forecasting models: Recent advances and challenges. Learning and Nonlinear Models 19, 29–50.

Ma, J.e.a., 2023. The evolution of global soybean trade network pattern based on complex network. Applied Economics 56, 3133–3149. doi:10.1080/00036846.2023.2204218.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. The m4 competition: Results, findings, conclusion and way forward. International Journal of Forecasting 34, 802–808. doi:10.1016/j.ijforecast.2018.06.001.

Mariño, J., Arrieta-Prieto, M., Calderón, S., 2025. Comparison between statistical models and machine learning for forecasting multivariate time series. Communications in Statistics 11, 56–91. doi:10.1080/23737484.2025.2463905.

Martignone, G.e.a., 2024. The rise of soybean in international commodity markets: a quantile investigation. Heliyon 10. doi:10.1016/j.heliyon.2024.e34669.

Masini, R., Medeiros, M., Mendes, E., 2020. Machine learning advances for time series forecasting. arXiv doi:10.1111/joes.12429.

Mehta, R., 2021. Digital library service model for predictive analysis based on multivariable fuzzy logic. Digital Library Perspectives 37, 366–383. doi:10.1108/DLP-07-2020-0071.

Noorunnahar, M., Chowdhury, A., Mila, F., 2023. Xgboost model to forecast annual rice production. PLOS ONE 18. doi:10.1371/journal.pone.0283452.

Norberg, M., Deutsch, L., 2023. The soybean through world history. Routledge, London. doi:10.4324/9780367822866.

Oliveira, H.F.d.e.a., 2024. Comércio global de soja: uma análise de rede complexa. SEA – Aplicação Prática da Ciência doi:10.70147/s36203216.

Oo, Z., Phyu, S., 2020. Time series prediction based on facebook prophet. International Journal of Applied Mathematics, Electronics and Computers doi:10.18100/ijamec.816894.

Owens, M.J., 2018. Time-window approaches to space-weather forecast metrics: A solar wind case study. Space Weather 16, 1847–1861. doi:10.1029/2018SW002059.

Page, H., 1924. The soybean. Nature 113, 813–815. doi:10.1038/113813a0.

Salman, A., Kalakech, M., Steiti, M., 2024. Random forest based forecasting model for time series prediction. Applied Artificial Intelligence .

Santos, L.B.e.a., 2024. Soybean yield prediction using machine learning algorithms. Smart Agricultural Technology doi:10.1016/j.atech.2024.100442.

Santos, V.e.a., 2021. Machine learning algorithms for soybean yield prediction in the brazilian cerrado. Journal of the Science of Food and Agriculture doi:10.1002/jsfa.11713.

Severiano, C.A., Silva, P.C.d.L.e., Cohen, M.W., Guimarães, F.G., 2021. Evolving fuzzy time series for spatio-temporal forecasting in renewable energy systems. Renewable Energy 171, 764–783.

Shakeel, A., Chong, D., Wang, J., 2023. Load forecasting of district heating system based on improved fb-prophet model. Energy 278, 127637. doi:10.1016/j.energy.2023.127637.

Song, Q., Chissom, B.S., 1993. Fuzzy time series and its models. Fuzzy Sets and Systems 54, 269–277. doi:10.1016/0165-0114(93)90372-O.

Srichaiyan, P., Tippayawong, K., Boonprasope, A., 2025. Forecasting soybean futures prices with adaptive ai models. IEEE Access 13, 48239–48256. doi:10.1109/ACCESS.2025.3546786.

Taieb, S.B., Bontempi, G., Atiya, A.F., Sorjamaa, A., 2012. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn3 data set. International Journal of Forecasting 28, 706–730. doi:10.1016/j.ijforecast.2011.03.001.

Taylor, S.J., Letham, B., 2018. Forecasting at scale. The American Statistician 72, 37–45. doi:10.1080/00031305.2017.1380080.

Valente, J., Maldonado, S., 2020. Svr-ffs: Feature selection for time series forecasting. Expert Systems with Applications 160. doi:10.1016/j.eswa.2020.113729.

Wang, B., Liu, X., Chi, M., Li, Y., 2024. Bayesian network based probabilistic weighted high-order fuzzy time series forecasting. Expert Systems with Applications 237, 121430.

Wang, J., Cheng, Q., Dong, Y., 2022. Xgboost-based multivariate framework for stock forecasting. Kybernetes 52, 4158–4177. doi:10.1108/K-12-2021-1289.

Wang, Y., Zhang, H., 2023. Multivariate time series forecasting using support vector regression. Applied Soft Computing .

Yan, J., Wang, X., Liu, Y., Zhang, H., 2021. Lightgbm: Accelerated crop breeding through ensemble learning. Genome Biology 22. doi:10.1186/s13059-021-02492-y.

Yazdanbakhsh, O., Dick, S., 2017. Forecasting multivariate time series via complex fuzzy logic. IEEE Transactions on Systems, Man, and Cybernetics: Systems 47, 2160–2171. doi:10.1109/TSMC.2016.2630668.

Zhang, J., 2025. Enhancing predictive models using xgboost. ITM Web of Conferences doi:10.1051/itmconf/20257002014.

Zhang, L.e.a., 2021. Time series forecast of sales volume based on xgboost. Journal of Physics: Conference Series 1873. doi:10.1088/1742-6596/1873/1/012067.

Zhang, X., O'Donnell, J., 2020. Support vector regression for multivariate time series forecasting. Expert Systems with Applications .

Zhou, X.e.a., 2022. Morphological changes during soybean domestication. Frontiers in Plant Science 13. doi:10.3389/fpls.2022.913238.

Züge, C.V., Coelho, L.d.S., 2024. Granular weighted fuzzy approach applied to short-term load demand forecasting. Technologies 12, 182.

3 CONCLUSÃO GERAL

Conclui-se, de forma geral, que as abordagens baseadas em lógica fuzzy apresentaram desempenho satisfatório, tanto no caso univariado quanto no multivariado, consolidando-se como uma alternativa robusta para a previsão de séries temporais complexas. Os resultados também indicam a influência da configuração dos conjuntos de treinamento, teste e das partições fuzzy no desempenho dos modelos, destacando a importância da escolha desses hiperparâmetros para a previsão das séries temporais analisadas.

No primeiro estudo, referente ao preço do milho, observou-se que o melhor desempenho foi alcançado com 80% dos dados destinados ao treinamento e 60 partições fuzzy. O modelo demonstrou boa capacidade de representar a dinâmica da série ao longo do tempo, incluindo períodos de estabilidade e oscilações estruturais. As previsões também foram capazes de refletir movimentos de queda e posterior estabilização, indicando que a metodologia adotada conseguiu capturar adequadamente o comportamento histórico da série.

No segundo estudo, relacionado ao preço da soja, o modelo PWFTS apresentou o melhor desempenho global, sendo superior aos modelos de machine learning, no horizonte de previsão de um dia à frente, evidenciando maior estabilidade e precisão ao longo dos anos analisados. Em determinados períodos, como em 2014, o modelo obteve resultados mais expressivos, com menores erros e maior acurácia direcional. Embora algumas suavizações e pequenas defasagens tenham sido observadas, o modelo conseguiu representar de forma satisfatória a dinâmica da série, incluindo suas oscilações e mudanças de tendência, mostrando-se eficaz para previsões de curto prazo.