INSTITUTO
UNIVERSITÁRIO
DE LISBOA

# Aplicação de Inteligência Artificial na Previsão de Séries Temporais para Preços da Soja

Felicio Cabral Mendes

Mestrado em Ciência de Dados

Orientador(a):
PhD Diana Elisabeta Aldea Mendes, Professora Associada, Iscte-IUL

Outubro, 2025

BUSINESS SCHOOL

TECNOLOGIAS E ARQUITETURA

Departamento de Métodos Quantitativos para Gestão e Economia

Departamento de Ciências e Tecnologia da Informação

# Aplicação de Inteligência Artificial na Previsão de Séries Temporais para Preços da Soja

Felicio Cabral Mendes

Mestrado em Ciência de Dados

Orientador(a):
PhD Diana Elisabeta Aldea Mendes, Professora Associada, Iscte-IUL

Outubro, 2025

# Agradecimentos

Agradeço primeiramente a Deus, por me conceder saúde, sabedoria, paciência e força para enfrentar cada desafio ao longo desta jornada. Sem a Sua presença e amparo, esta conquista não seria possível.

Expresso também minha mais profunda gratidão à minha orientadora, Diana Mendes, por sua orientação excepcional, paciência, dedicação e confiança ao longo de todo este percurso. Sua competência, comprometimento e sensibilidade acadêmica foram fundamentais para o desenvolvimento deste trabalho. Mais do que uma orientadora, foi uma verdadeira mentora, sempre disposta a ouvir, a incentivar e a me conduzir pelos caminhos certos da pesquisa. Tenho enorme admiração e respeito pelo seu trabalho e por tudo o que aprendi sob sua orientação.

Agradeço, de coração, à minha esposa, Rafaela Machado, pelo amor, apoio e compreensão incondicionais. Sua presença constante, suas palavras de incentivo e sua paciência nos momentos de maior desafio foram essenciais para que eu chegasse até aqui. Esta conquista também é sua.

Às minhas filhas, Vitoria Mendes e Catarina Mendes, minha fonte diária de inspiração e motivação. São vocês que me impulsionam a ser uma pessoa e um profissional melhor, e que me lembram todos os dias o verdadeiro sentido de lutar e me dedicar. Tudo o que conquisto é, em grande parte, por vocês e para vocês — para que eu possa proporcionar uma vida melhor à nossa família e ser um exemplo de esforço e perseverança.

A todos que, direta ou indiretamente, contribuíram para a realização desta jornada, deixo meu sincero agradecimento.

# Resumo

Esta tese de mestrado investigará a aplicação de inteligência artificial na previsão de séries temporais para os preços da soja. Este estudo abordará como fatores climáticos e tendências históricas atuam sobre o mercado da soja, utilizando técnicas avançadas de inteligência artificial para analisar e prever variações de preço desta *commodity* agrícola. A pesquisa combinará análise estatística e modelagem preditiva, integrando dados climáticos e históricos para gerar análises mais precisas. O objetivo é fornecer *insights* valiosos para agricultores, comerciantes e investidores em *commodity* agrícola na bolsa de valores, possibilitando uma melhor tomada de decisão. Esta abordagem multidisciplinar contribuirá significativamente para os campos da ciência de dados, investimento em *commodities* agrícolas na bolsa de valores e na economia agrícola, demonstrando a aplicabilidade prática da inteligência artificial em um mercado complexo.

**Palavras-chave:** Inteligência Artificial, Previsão de Séries Temporais, Preços da Soja, Tendências de *Commodities* Agrícolas, Aprendizado de Máquina, Bolsa de Valores.

# Abstract

This master's thesis will investigate the application of artificial intelligence in time series forecasting for soybean prices. This study will address how climatic factors and historical trends affect the soybean market, using advanced artificial intelligence techniques to analyze and predict price variations of this agricultural commodity. The research will combine statistical analysis and predictive modeling, integrating climatic and historical data to generate more accurate analyses. The goal is to provide valuable insights for farmers, traders, and investors in agricultural commodities on the stock exchange, enabling better decision-making. This multidisciplinary approach will contribute significantly to the fields of data science, agricultural commodity investment on the stock exchange, and agricultural economics, demonstrating the practical applicability of artificial intelligence in a complex market.

**Keywords:** Artificial Intelligence, Time Series Forecasting, Soybean Prices, Agricultural Commodity Trends, Machine Learning, Stock Market.

# Índice


Resumo i

*Abstract* iii

Lista de Figuras x

Lista de Tabela xi

Glossário de Siglas xii

Capítulo 1 - Introdução 1

1.1 Contexto e Motivação 1

1.2 Problema de Pesquisa 2

1.3 Objetivos 3

1.3.1 Objetivo Geral 3

1.3.2 Objetivos Específicos 3

1.4 Justificativa 4

1.5 Estrutura da Dissertação 5

Capítulo 2 – Conceitos Base e Revisão de Literatura 7

2.1 Séries Temporais: Conceitos Fundamentais 7

2.1.1 Componentes de Séries Temporais 7

2.1.2 Estacionariedade 8

2.2 Modelos Estatísticos Clássicos: ARIMA e ARIMAX 8

2.2.1 Modelo ARIMA 8

2.2.2. Modelo ARIMAX 9

2.2.3 *Feature Engineering* em Modelos ARIMAX 9

2.3 Algoritmos de *Machine Learning* para Séries Temporais 10

2.3.1 XGBoost (*Extreme Gradient Boosting*) 10

2.3.2. Redes Neurais Artificiais: *Multi-Layer Perceptron* (MLP) 11

2.3.3. *Long Short-Term Memory* (LSTM) 12

2.4 Influência de Variáveis Climáticas em Preços Agrícolas 13

2.4.1 Mecanismos de Influência Climática 13

2.4.2 Causalidade Estatística: Teste de Granger 14

2.4.3 Evidências Empíricas na Literatura 14

2.5 Trabalhos Relacionados em Previsão de Preços de *Commoditie* 15

2.5.1 Estudos Específicos em *Commodities* Agrícolas 16

2.5.2 Lacunas na Literatura 16

2.6 Síntese do Capítulo 17

Capítulo 3 - Dados e Metodologia 19

3.1 Coleta de Dados 19

3.1.1. Dados de Preços da Soja 19

3.1.2. Dados Climáticos: Cuiabá, Brasil 19

3.1.3. Dados Climáticos: Des Moines, Estados Unidos 19

3.1.4 Justificativa da Escolha das Regiões Climáticas 19

3.2 Pré-Processamento e Integração dos Dados 20

3.2.1. Tratamento de Valores Ausentes e Outliers 20

3.2.2. Integração das Séries Climáticas (Brasil e EUA) 20

3.2.3. Agregação Temporal e Sincronização 21

3.3 Análise Exploratória dos Dados 21

3.4 Estratégia de Divisão Treino/Teste 22

3.5 Especificação e Configuração dos Modelos Preditivos 22

3.5.1. ARIMA (Modelo Univariado) 22

3.5.2. ARIMAX Melhorado (com *Feature Engineering*) 23

3.5.3. XGBoost (*Extreme Gradient Boosting*) 25

3.5.4. MLP (*Multi-Layer Perceptron*) 27

3.5.5. LSTM (*Long Short-Term Memory*) 28

3.6. Métricas de Avaliação 31

3.7. Ferramentas de Implementação 32

Capítulo 4 - Resultados e Análises 33

4.1 Desempenho dos Modelos por Horizonte Temporal 33

4.1.1. Horizonte de 3 Meses (Curto Prazo) 34

4.1.2 Horizonte de 6 Meses (Médio Prazo) 34

4.1.3. Horizontes de 9 e 12 Meses (Médio-Longo e Longo Prazo) 35

4.2 Análise de Importância de Variáveis (*Feature Importance*) 36

4.3 Projeções Futuras para as Safras de 2025-2026 38

4.4 Discussão e Interpretação dos Resultados 39

4.4.1. Superioridade do ARIMAX Melhorado em Horizontes de Curto Prazo 40

4.4.2. Ascensão do XGBoost em Horizontes de Médio e Longo Prazo 41

4.4.3. Contribuição das Variáveis Climáticas na Previsão de Preços 41

4.4.4 Implicações Práticas para Diferentes *Stakeholders* 42

4.4.5. Limitações Identificadas nos Experimentos 44

Capítulo 5 - Conclusões e Trabalhos Futuros 47

5.1 Síntese do Trabalho Realizado 47

5.2 Principais Descobertas 47

5.3 Limitações do Estudo 48

5.4 Trabalhos Futuros 48

5.5. Considerações Finais 49

Referências Bibliográficas 50

Anexos 54

# Lista de Figuras

3.1. Teste de Causalidade de Granger
4.1. Previsões Consolidadas 2025-2026, Modelos Campeões
4.2. Comparação Mensal das Previsões 2025-2026
A.1. Gráfico ARIMAX Melhorado de Previsões para Horizonte de 3 meses
A.2. Gráfico XGBoost de Previsões para Horizonte de 3 meses
A.3. Gráfico MLP de Previsões para Horizonte de 3 meses
A.4. Gráfico LSTM de Previsões para Horizonte de 3 meses
A.5. Gráfico ARIMAX Melhorado de Previsões para Horizonte de 6 meses
A.6. Gráfico XGBoost de Previsões para Horizonte de 6 meses
A.7. Gráfico MLP de Previsões para Horizonte de 6 meses
A.8. Gráfico LSTM de Previsões para Horizonte de 6 meses
A.9. Gráfico ARIMAX Melhorado de Previsões para Horizonte de 9 meses
A.10. Gráfico XGBoost de Previsões para Horizonte de 9 meses
A.11. Gráfico MLP de Previsões para Horizonte de 9 meses
A.12. Gráfico LSTM de Previsões para Horizonte de 9 meses
A.13. Gráfico ARIMAX Melhorado de Previsões para Horizonte de 12 meses
A.14. Gráfico XGBoost de Previsões para Horizonte de 12 meses
A.15. Gráfico MLP de Previsões para Horizonte de 12 meses
A.16. Gráfico LSTM de Previsões para Horizonte de 12 meses
A.17. ARIMAX Melhorado: Previsões de Curto e Médio Prazo
A.18. XGBoost: Previsões de Longo Prazo
A.19. XGBoost: Importância das Variáveis

# Lista de Tabelas

4.1. Métricas de Desempenho dos Modelos para Horizonte de 3 Meses
4.2. Métricas de Desempenho dos Modelos para Horizonte de 6 Meses
4.3. Métricas de Desempenho dos Modelos para Horizonte de 9 Meses
4.4. Métricas de Desempenho dos Modelos para Horizonte de 12 Meses
4.5. Top 10 Variáveis Mais Importantes no Modelo XGBoost
4.6. Resumo Executivo de Recomendações
4.7. Recomendações de Modelos por Tipo de Stakeholder e Horizonte

# Glossário de Siglas

| | |
|---|---|
| ADF | Augmented Dickey-Fuller (Teste de Dickey-Fuller Aumentado) |
| AI | Artificial Intelligence (Inteligência Artificial) |
| AIC | Akaike Information Criterion (Critério de Informação de Akaike) |
| AR | AutoRegressive (Autorregressivo) |
| ARIMA | AutoRegressive Integrated Moving Average |
| ARIMAX | AutoRegressive Integrated Moving Average with Exogenous Variables |
| CBOT | Chicago Board of Trade |
| EMA | Exponential Moving Average (Média Móvel Exponencial) |
| FAO | Food and Agriculture Organization |
| INMET | Instituto Nacional de Meteorologia |
| IQR | InterQuartile Range (Intervalo Interquartil) |
| KPSS | Kwiatkowski-Phillips-Schmidt-Shin (Teste de Estacionariedade) |
| LSTM | Long Short-Term Memory |
| MA | Moving Average (Média Móvel) |
| MAE | Mean Absolute Error (Erro Absoluto Médio) |
| MAPE | Mean Absolute Percentage Error (Erro Percentual Absoluto Médio) |
| ML | Machine Learning (Aprendizagem Automática) |
| MLP | Multi-Layer Perceptron (Perceptrão Multicamada) |
| MSE | Mean Squared Error (Erro Quadrático Médio) |
| NOAA | National Oceanic and Atmospheric Administration |
| $R^2$ | Coeficiente de Determinação |
| ReLU | Rectified Linear Unit (Unidade Linear Retificada) |
| RMSE | Root Mean Square Error (Raiz do Erro Quadrático Médio) |
| RNN | Recurrent Neural Network (Rede Neural Recorrente) |
| STL | Seasonal and Trend decomposition using Loess |
| SVM | Support Vector Machine (Máquina de Vetores de Suporte) |
| USDA | United States Department of Agriculture |
| XGBoost | eXtreme Gradient Boosting |

CAPÍTULO 1

# Introdução

## 1.1 Contexto e Motivação

A soja se configura como uma das mais importantes *commodities* agrícolas no cenário mundial. Ela possui um caráter essencial para a garantia da segurança alimentar e, de igual modo, para a saúde económica e financeira de vários países. Na atualidade, os expoentes na produção e na exportação da soja em escala global são o Brasil e os Estados Unidos da América. Em consequência disso, eles têm uma grande dependência da capacidade de antecipar as cotações deste produto com o objetivo de garantir o equilíbrio financeiro do segmento do agronegócio, além de nortear as deliberações táticas ligadas a aportes de capital.

De acordo com informações provenientes da *Food and Agriculture Organization* (FAO, 2024), a soja responde por uma parcela superior a sessenta por cento da fabricação global de compostos proteicos de origem vegetal destinados ao consumo animal. Adicionalmente, ela se apresenta como um insumo indispensável ao setor industrial de biodiesel e também ao sustento humano. Dentro do cenário do Brasil, o grão foi o elemento correspondente a quarenta e nove por cento da quantia integral das vendas externas do agronegócio no ano de 2024, um fato que firmou a nação na posição de liderança mundial no que tange ao envio deste cereal para outros mercados (USDA, 2024).

As cotações da soja resultam de uma intrincada combinação de elementos, que abrange as circunstâncias do clima, os movimentos de oferta e demanda em escala mundial, as diretrizes de comércio entre nações, as oscilações nas taxas de câmbio bem como os acontecimentos no campo da geopolítica (Isengildina-Massa et al., 2021). No meio desses componentes, os indicadores relacionados ao clima exercem uma função central, pois os níveis de calor ou frio muito acentuados, os regimes de chuvas sem regularidade juntamente com outras ocorrências climáticas desfavoráveis possuem o potencial de afetar de maneira expressiva o rendimento das lavouras e, por via de consequência, provocar um reflexo nos valores praticados no comércio exterior (Nayak et al., 2024).

Fazer uma previsão acurada das cotações da soja é algo que tem uma importância crucial para os variados atores que compõem o setor agrícola. Os agricultores requerem projeções que inspiram confiança para poderem fazer o planejamento do plantio, estabelecerem suas táticas de comercialização e realizarem a gestão dos riscos de ordem financeira. Investidores e *traders*

fazem uso dessas previsões com o fim de nortear suas deliberações sobre adquirir ou alienar ativos nos mercados futuros, ao passo que os elaboradores de políticas públicas se apoiam em estimativas de valores para a criação de abordagens voltadas à segurança alimentar e à estabilização de mercado (Chen et al., 2021).

Ao longo do tempo, as abordagens convencionais para o prognóstico das cotações agrícolas tinham seu alicerce principal em modelos econométricos clássicos, a exemplo do ARIMA (*AutoRegressive Integrated Moving Average*), concebido por Box e Jenkins (1970), e de suas derivações. Apesar de tais modelos terem se mostrado eficientes na identificação de tendências temporais lineares, eles possuem importantes pontos fracos quando se trata de manejar interações complexas do tipo não-lineares, bem como de trabalhar com variáveis exógenas multidimensionais (Kontopoulou et al., 2023).

Em tempos mais recentes, o progresso nas metodologias de inteligência artificial e de *machine learning* provocou uma transformação profunda no domínio da previsão de séries temporais. Estruturas de modelagem como o XGBoost (*eXtreme Gradient Boosting*), desenvolvido por Chen e Guestrin (2016), as redes neurais artificiais (MLP – *Multi-Layer Perceptron*), descritas por Goodfellow et al. (2016) e também as redes neurais recorrentes (LSTM – *Long Short-Term Memory*), propostas por Hochreiter e Schmidhuber (1997), exibem uma aptidão maior para a identificação de padrões intrincados e não-lineares contidos nos registros do passado, além de conseguirem fazer a incorporação, de maneira eficaz, de múltiplos fatores que servem como explicação (Kim et al., 2025; Hall & Rasheed, 2025).

## 1.2 Problema de Pesquisa

Não obstante os notáveis progressos nas metodologias de prognóstico, permanece uma carência no corpo de conhecimento acadêmico no que diz respeito à realização de uma análise comparativa e metódica entre os modelos estatísticos clássicos e os algoritmos contemporâneos de *machine learning*, com aplicação direcionada para a previsão de preços da soja, ao se incluir de maneira clara o efeito que as variáveis climáticas exercem, oriundas de localidades produtoras de importância tática.

A indagação primordial que serve como bússola para a presente investigação pode ser formulada da seguinte maneira: Qual metodologia de modelagem – a tradicional (ARIMA/ARIMAX) ou aquela que tem por base a inteligência artificial (XGBoost, MLP, LSTM) – apresenta um desempenho superior no que tange à previsão de preços da soja, quando se faz a incorporação de variáveis climáticas das localidades produtoras mais importantes (Brasil e

Estados Unidos)? Adicionalmente, de que maneira as distintas janelas temporais de antecipação (3, 6, 9 e 12 meses) impactam a *performance* comparativa dessas abordagens?

Esta questão desdobra-se em sub-questões específicas:

1. Em que medida as variáveis climáticas (temperatura e precipitação) de Cuiabá, MT (Brasil) e Des Moines, IA (Estados Unidos) apresentam relações de causalidade estatisticamente significativas com os preços da soja no mercado *Chicago Board of Trade* (CBOT)?

2. Qual modelo apresenta melhor desempenho preditivo quando avaliado por múltiplas métricas de *performance* em diferentes horizontes temporais?

3. Existe um *trade-off* entre a complexidade do modelo e a acurácia preditiva? Modelos mais complexos (LSTM, XGBoost) superam sistematicamente modelos estatísticos clássicos (ARIMA, ARIMAX)?

4. Qual a importância relativa das variáveis climáticas em comparação com os valores históricos de preços na determinação dos preços futuros da soja?

## 1.3 Objetivos

### 1.3.1 Objetivo Geral

O propósito central desta pesquisa consiste em realizar a elaboração e a subsequente aferição comparativa de modelos para a previsão de preços da soja, os quais são fundamentados tanto em metodologias estatísticas convencionais (ARIMA, ARIMAX) quanto em algoritmos de inteligência artificial (XGBoost, MLP, LSTM), fazendo a integração de variáveis climáticas originárias de localidades produtoras de peso estratégico (Cuiabá, Brasil, e Des Moines, Estados Unidos), com a finalidade de se determinar qual metodologia se demonstra mais eficiente dentro de diferentes janelas temporais de projeção (3, 6, 9 e 12 meses).

### 1.3.2 Objetivos Específicos

1. Fazer a composição de uma completa base de dados histórica (cobrindo o período de janeiro/2010 a setembro/2025) que combine as cotações mensais da soja no mercado CBOT com as variáveis climáticas (temperatura e precipitação) de Cuiabá, MT, e também de Des Moines, IA;

2. Levar a cabo uma análise exploratória e uma estatística descritiva sobre os dados, com o propósito de se identificarem os padrões sazonais, as tendências temporais e as correlações que existem entre as variáveis climáticas e os preços da soja;

3. Fazer a comprovação estatística das relações de causalidade entre as variáveis climáticas e os preços da soja por intermédio do Teste de Causalidade de Granger;

4. Efetuar a implementação e a otimização de modelos estatísticos clássicos (ARIMA e ARIMAX com a aplicação de *feature engineering* avançado) para a realização da previsão de preços em múltiplos horizontes temporais;

5. Realizar a elaboração e o ajuste dos modelos de *machine learning* (XGBoost, MLP, LSTM), empregando metodologias rigorosas para o *hyperparameter tuning* (GridSearchCV, Keras Tuner) e para a validação temporal (TimeSeriesSplit);

6. Executar uma avaliação comparativa da *performance* preditiva dos modelos por meio do uso de métricas padronizadas (RMSE – *Root Mean Square Error*, MAE – *Mean Absolute Error*, MAPE – *Mean Absolute Percentage Error*, $R^2$) para as janelas de 3, 6, 9 e 12 meses;

7. Fazer a análise da importância relativa das variáveis preditoras (*feature importance*) nos modelos de *machine learning*, com a quantificação da contribuição dos fatores climáticos em oposição ao histórico de preços;

8. Identificar o modelo campeão para cada horizonte de previsão e gerar projeções de preços para o período de novembro/2025 a outubro/2026;

9. Elaborar recomendações práticas que se destinam aos agricultores, aos investidores e aos formuladores de políticas públicas a respeito da seleção e do emprego dos modelos preditivos, em conformidade com o horizonte temporal de interesse.

## 1.4 Justificativa

A fundamentação para a realização desta pesquisa se apoia em variados aspectos de natureza teórica, metodológica e também prática:

Relevância Econômica: A soja constitui-se como um produto primário de valor estratégico, sendo responsável pela movimentação de bilhões de dólares a cada ano. A capacidade de se fazer prognósticos mais acurados sobre os preços pode levar a deliberações mais acertadas de *hedge*, à mitigação dos perigos financeiros enfrentados pelos produtores e também ao aprimoramento das políticas agrícolas governamentais (USDA, 2024).

Lacuna Metodológica: Apesar da existência de trabalhos pontuais que tratam da previsão de preços agrícolas por meio do emprego de *machine learning* ou de modelos estatísticos clássicos, são escassas as pesquisas que efetuam uma análise comparativa, metódica e criteriosa, entre as duas metodologias, que levam em conta, de forma clara, as variáveis climáticas de diversas localidades de produção, e que, simultaneamente, fazem a aferição de múltiplos horizontes de previsão (Nayak et al., 2024; Kontopoulou et al., 2023).

Inovação Técnica: A presente investigação faz a aplicação de *feature engineering* avançado nos modelos ARIMAX (o que abrange *lags* estratégicos que se harmonizam com o ciclo agrícola, médias móveis, anomalias climáticas e também os acumulados de precipitação). Adicionalmente, realiza a comprovação estatística das relações de causalidade por meio do Teste de Granger, elementos que muitas vezes são deixados de lado na literatura de aplicação prática.

Transparência e Replicabilidade: Todo o código desenvolvido será disponibilizado através de link [1]para o ficheiro online, permitindo que pesquisadores e profissionais repliquem, validem e estendam os resultados obtidos – contribuindo para o avanço coletivo do conhecimento científico na área.

Aplicabilidade Prática: Os achados desta pesquisa oferecem um norteamento bem definido a respeito do modelo a ser empregado, considerando-se o horizonte de previsão que se almeja, prestando auxílio aos tomadores de decisão no processo de escolha de ferramentas analíticas que sejam adequadas para as suas demandas particulares.

Entre os aportes mais significativos que se antecipam desta dissertação, encontram-se:

1. Evidências empíricas criteriosas acerca do desempenho comparativo dos modelos estatísticos clássicos em contraposição aos algoritmos de IA para a previsão de preços agrícolas, considerando distintos horizontes temporais;

2. A validação estatística dos vínculos de causa e efeito entre as variáveis climáticas, oriundas de localidades produtoras essenciais, e as cotações internacionais da soja;

3. Uma abordagem metodológica transparente e que permite replicação, destinada à elaboração e à aferição de modelos preditivos no contexto de *commodities* agrícolas;

4. A formulação de direcionamentos práticos para a escolha de modelos preditivos, com fundamento nos horizontes de previsão e também nas métricas de desempenho;

5. A disponibilização de uma base de dados que foi consolidada e devidamente tratada (2010-2025), que faz a unificação dos preços da soja com as variáveis climáticas de variadas localidades, e que possui utilidade em potencial para investigações que venham a ser realizadas no futuro.

## 1.5 Estrutura da Dissertação

O presente trabalho se encontra dividido em um total de cinco partes capitulares, as quais foram arranjadas em uma sequência coerente com o propósito de guiar quem realiza a leitura, partindo da ambientação da problemática e seguindo até os arremates conclusivos e as

[1] Link: https://drive.google.com/drive/folders/1m6fNDJ8b393lNPRuMIERfvyoXR8EL6Yr?usp=sharing

sugestões derradeiras:

Capítulo 1 – Introdução: Nesta seção, faz-se a exposição do contexto, da motivação, do problema de pesquisa, dos objetivos que foram traçados, da justificativa e também da organização geral da dissertação.

Capítulo 2 – Conceitos Base e Revisão de Literatura: Esta parte realiza uma abordagem sobre os alicerces teóricos da previsão de séries temporais, dos modelos estatísticos clássicos (ARIMA/ARIMAX), dos algoritmos de *machine learning* (XGBoost, MLP, LSTM), do efeito das variáveis climáticas sobre os preços agrícolas e dos estudos correlatos no campo da previsão de preços de *commodities*.

Capítulo 3 – Dados e Metodologia: Aqui se procede com uma descrição minuciosa das fontes de dados, do caminho percorrido para a coleta, o tratamento e a preparação das bases, da análise exploratória, das metodologias de modelagem que foram postas em prática (ARIMA, ARIMAX, XGBoost, MLP, LSTM), dos processos de otimização de hiperparâmetros, das abordagens de validação temporal e das métricas de avaliação que foram empregadas.

Capítulo 4 – Resultados e Análises: Neste ponto, são exibidos os achados empíricos que foram obtidos, o que engloba as estatísticas descritivas dos dados, os desfechos do Teste de Causalidade de Granger, a *performance* comparativa dos modelos em diferentes horizontes de previsão, a análise de *feature importance*, as representações visuais das projeções e a eleição dos modelos campeões para cada horizonte temporal.

Capítulo 5 – Conclusões e Trabalhos Futuros: Este capítulo faz uma síntese das descobertas centrais do estudo, debate as consequências teóricas e práticas dos achados, faz o reconhecimento das limitações da investigação e aponta rumos para pesquisas que poderão ser feitas no futuro.

De forma adicional, a dissertação inclui link [2]de acesso ao contendo integral de cada um dos *notebooks* que foram elaborados, o que viabiliza a completa capacidade de replicação dos achados que aqui se demonstram.

[2] Link: https://drive.google.com/drive/folders/1m6fNDJ8b393lNPRuMIERfvyoXR8EL6Yr?usp=sharing

CAPÍTULO 2

# Conceitos Base e Revisão de Literatura

O presente capítulo se dedica a apresentar um exame criterioso dos alicerces conceituais e do conhecimento científico já produzido sobre a previsão de preços de *commodities* agrícolas, com foco direcionado para modelos de séries temporais e para técnicas de inteligência artificial. A exposição do conteúdo foi segmentada em cinco eixos primordiais: (2.1) os fundamentos de séries temporais e suas características; (2.2) os modelos estatísticos clássicos ARIMA e ARIMAX; (2.3) os algoritmos de *machine learning* com aplicação em séries temporais; (2.4) a influência de variáveis climáticas sobre os preços agrícolas; e (2.5) os trabalhos que fazem a comparação entre diferentes abordagens de modelagem.

## 2.1 Séries Temporais: Conceitos Fundamentais

A definição de uma série temporal corresponde a uma sucessão de aferições dispostas em ordem temporal, as quais são, na maioria das vezes, coletadas em períodos de tempo uniformes (Box et al., 2015). No âmbito dos preços de *commodities* agrícolas, as séries temporais exibem traços distintivos aos quais se deve dar atenção no momento da elaboração da modelagem preditiva.

### 2.1.1 Componentes de Séries Temporais

De acordo com Hyndman e Athanasopoulos (2021), é possível fazer a desagregação das séries temporais em quatro elementos fundamentais:

1. Tendência (*Trend*): Representa um deslocamento de longa duração que aponta o rumo geral da série através do tempo. No âmbito do mercado de soja, as tendências são capazes de espelhar transformações de caráter estrutural na oferta e demanda global.

2. Sazonalidade (*Seasonality*): Corresponde a comportamentos que ocorrem de forma repetitiva em períodos de tempo regulares. As cotações agrícolas muitas vezes manifestam uma sazonalidade vinculada aos ciclos de plantio e colheita.

3. Ciclo (*Cycle*): Consiste em oscilações que se prolongam no tempo, porém sem um padrão de tempo definido para sua repetição, as quais comumente estão relacionadas a fatores econômicos e políticos.

4. Componente Irregular (*Irregular/Noise*): Trata-se de alterações de natureza aleatória e de difícil previsão, cuja explicação não pode ser encontrada nos elementos restantes.

### 2.1.2 Estacionariedade

A noção de estacionariedade representa um conceito basilar para a análise de séries temporais. A classificação de uma série como estacionária ocorre quando os seus atributos estatísticos, tais como a média, a variância e a co-variância, se mantêm inalterados com o passar do tempo (Box et al., 2015). Sob uma perspectiva formal, uma série temporal $\{y_t\}$ pode ser considerada fracamente estacionária caso atenda às seguintes condições:

- $E[y_t] = \mu$ (média constante)
- $Var(y_t) = \sigma^2$ (variância finita e constante)
- $Cov(y_t, y_{t+h})$ depende apenas de h (autocovariância função apenas da defasagem)

Procedimentos de verificação estatística, a exemplo do *Augmented Dickey-Fuller* (ADF) e do KPSS (Kwiatkowski-Phillips-Schmidt-Shin), costumam ser aplicados para a verificação da estacionariedade em uma série temporal (Dickey & Fuller, 1979; Kwiatkowski et al., 1992). As séries que se classificam como não-estacionárias, ou seja, aquelas que não cumprem ao menos um dos critérios mencionados anteriormente, têm a possibilidade de passar por uma conversão para se tornarem estacionárias, utilizando-se a diferenciação, as transformações logarítmicas ou ainda outras metodologias de preparação prévia dos dados.

## 2.2 Modelos Estatísticos Clássicos: ARIMA e ARIMAX

### 2.2.1 Modelo ARIMA

O modelo ARIMA (*AutoRegressive Integrated Moving Average*), concebido por Box e Jenkins (1970), configurou-se como a abordagem hegemônica no campo da previsão de séries temporais por um longo período de várias décadas. Um modelo do tipo ARIMA(p, d, q) faz a integração de três elementos:

- **AR(p) (Autorregressivo de ordem p):** O valor presente é determinado por uma relação linear com os p valores que o antecederam, ao qual se soma um termo de erro.
- **I(d) (Integrado de ordem d):** Corresponde à quantidade de diferenciações que se fazem necessárias para que a série se torne estacionária.
- **MA(q) (Média Móvel de ordem q):** O valor presente corresponde a uma função linear dos q erros de prognóstico anteriores.

A representação em termos matemáticos para o ARIMA(p, d, q) é passível de ser demonstrada da seguinte forma:

$$\phi(B)(1 - B)^d y_t = \theta(B)\varepsilon_t$$

Nesta expressão, $\phi(B)$ e $\theta(B)$ representam polinômios de ordens p e q aplicados ao operador de *backshift* B, enquanto $\varepsilon_t$ corresponde ao termo de erro, que tipicamente se configura como um processo de *white noise* (Box et al., 2015).

Pesquisas comprovam a efetividade dos modelos ARIMA na previsão de preços agrícolas. Zhang et al. (2017) fizeram o uso do ARIMA com o objetivo de prever os preços da soja na China, alcançando resultados de MAPEs que se situaram no intervalo de 5% a 8% para horizontes de 3 a 6 meses. No entanto, o modelo ARIMA apresenta restrições de relevância: (1) ele parte do pressuposto da linearidade nas interações, (2) não faz a integração de variáveis explicativas externas, e (3) sofre uma diminuição em sua acurácia quando aplicado a janelas de tempo de prognóstico mais extensas (Kontopoulou et al., 2023).

### 2.2.2 Modelo ARIMAX

O modelo ARIMAX representa uma expansão do ARIMA, que se dá por meio da inclusão de variáveis exógenas (as quais também recebem o nome de variáveis explicativas ou covariáveis), o que possibilita que elementos externos exerçam influência sobre os prognósticos (Pankratz, 1991). A representação matemática geral para o ARIMAX pode ser descrita da seguinte maneira:

$$\phi(B)(1 - B)^d y_t = \beta X_t + \theta(B)\varepsilon_t$$

Nesta fórmula, $X_t$ simboliza um vetor de variáveis exógenas, enquanto $\beta$ representa o vetor de coeficientes que a ele se associa. Esse aprimoramento do modelo possui uma importância especial no contexto dos preços agrícolas, visto que torna possível fazer a integração de variáveis climáticas, de indicadores econômicos e de outros dados contextuais (Nayak et al., 2024).

Isengildina-Massa et al. (2021) conduziram um exame pormenorizado das metodologias existentes para a previsão de preços de *commodities* agrícolas, e apontaram o ARIMAX como uma alternativa utilizada com frequência nos cenários em que se tem a disponibilidade de variáveis macroeconômicas e climáticas. Os pesquisadores ressaltam que o ato de selecionar adequadamente as variáveis exógenas é de importância fundamental: variáveis que não possuem relevância são capazes de introduzir ruído e, por consequência, comprometer o desempenho preditivo.

### 2.2.3. *Feature Engineering* em Modelos ARIMAX

A *performance* apresentada pelos modelos ARIMAX pode ser objeto de um aprimoramento de grande relevância por meio da aplicação de *feature engineering* – que consiste na elaboração de novas variáveis preditoras que têm sua origem nos dados brutos. Entre as metodologias

frequentemente utilizadas encontram-se:

- ***Lags* defasados:** A utilização de valores pregressos das variáveis exógenas (a exemplo da temperatura registrada 3, 6, 9 e 12 meses antes), os quais têm a capacidade de captar os efeitos temporais diferidos que possuem importância para o ciclo agrícola.
- **Médias móveis:** A realização de uma suavização nas séries com o objetivo de se diminuir o ruído de curto prazo e de se colocar em evidência as tendências.
- **Anomalias:** O cálculo dos desvios em comparação com as médias do histórico, que possuem o potencial de sinalizar a ocorrência de eventos climáticos extremos.
- **Transformações não-lineares:** O emprego de Logaritmos, de raízes quadradas, bem como das interações que ocorrem entre as variáveis, com a finalidade de se capturarem as relações de maior complexidade.

Roberts (2006) apresentou evidências de que a incorporação de *lags* climáticos de natureza estratégica, que se encontram em harmonia com o ciclo de crescimento das culturas, tem o poder de promover uma melhoria expressiva nas previsões de preços agrícolas. A seleção adequada das defasagens é uma tarefa que demanda um saber aprofundado sobre o domínio agrícola: as fases de desenvolvimento que são consideradas críticas (floração, enchimento de grãos) exercem uma influência mais acentuada sobre os rendimentos e, por via de consequência, sobre as cotações futuras.

## 2.3. Algoritmos de *Machine Learning* para Séries Temporais

Ao longo dos vinte anos mais recentes, os algoritmos de *machine learning* despontaram como opções robustas em relação aos modelos estatísticos clássicos no campo da previsão de séries temporais. Tais abordagens se sobressaem pela sua habilidade para capturar relações não-lineares complexas, manejar um elevado número de variáveis preditoras e se ajustarem de forma autônoma aos padrões presentes nos dados, dispensando uma especificação explícita da forma funcional (Makridakis et al., 2018).

### 2.3.1. XGBoost (*Extreme Gradient Boosting*)

O XGBoost, concebido por Chen e Guestrin (2016), configura-se como um algoritmo de *ensemble learning* alicerçado em *gradient boosting* de árvores de decisão. O algoritmo realiza a construção de um modelo preditivo por intermédio da junção sequencial de diversas árvores de menor robustez, processo no qual cada árvore subsequente passa por um treinamento com o objetivo de fazer a correção das falhas apresentadas pelas árvores que a precederam. A função objetivo que o XGBoost busca otimizar é passível de ser demonstrada pela seguinte equação:

$$\mathcal{L} = \Sigma_i \, L(y_i, \hat{y}_i) + \Sigma_k \, \Omega(f_k)$$

Nesta expressão, L simboliza a função de perda (como, por exemplo, o erro quadrático médio), $\hat{y}_i$ representa a projeção, $y_i$ é o dado efetivo, e $\Omega(f_k)$ corresponde a um termo de regularização que aplica uma penalidade à complexidade das árvores, com o intuito de se evitar o *overfitting* (Chen & Guestrin, 2016).

O XGBoost vem exibindo uma *performance* superior tanto em disputas de *machine learning* (Kaggle) quanto no uso prático para a previsão de séries temporais. Seus benefícios mais notáveis englobam: (1) a habilidade para capturar as interações não-lineares de alta complexidade que ocorrem entre as variáveis, (2) a solidez que apresenta diante de *outliers* e também de dados faltantes, (3) a sua implementação de alta eficiência, que é viabilizada pela paralelização, e (4) a existência de mecanismos de regularização intrínsecos que atuam na mitigação do *overfitting* (Zhang & Haghani, 2015).

Dentro do cenário da previsão de preços agrícolas, Chen et al. (2021) fizeram o uso do XGBoost com o propósito de prognosticar as cotações de diversas *commodities* (o que incluiu a soja), e comprovaram que o modelo obteve um desempenho melhor que o ARIMA e as redes neurais simples em horizontes de médio e longo prazo (6-12 meses). Os pesquisadores ressaltam a relevância de se realizar um *hyperparameter tuning* criterioso – dando-se ênfase particular aos parâmetros *learning*_rate, max_depth, n_estimators e *subsample* – com a finalidade de se alcançar uma *performance* otimizada.

### 2.3.2. Redes Neurais Artificiais: *Multi-Layer Perceptron* (MLP)

O *Multi-Layer Perceptron* (MLP) se configura como um modelo de rede neural artificial que é formado por diversos níveis de neurônios que possuem uma conexão completa entre si. De maneira usual, um MLP possui uma camada de entrada, pelo menos um nível de camadas ocultas, e também uma camada de saída. A interligação entre as camadas ocorre de tal modo que cada neurônio no nível l se conecta a todos os neurônios do nível l+1, e o resultado final de cada neurônio é obtido pela aplicação de uma função de ativação não-linear sobre uma combinação linear ponderada dos sinais de entrada (Goodfellow et al., 2016).

Do ponto de vista formal, o resultado de um neurônio j, localizado na camada l, pode ser representado pela seguinte equação:

$$a_j^{(l)} = g(\Sigma_i \, w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)})$$

Nesta equação, g simboliza a função de ativação (sendo as mais usuais a ReLU, a sigmoide ou a *tanh*), $w_{ij}$ correspondem aos pesos sinápticos e $b_j$ representa o *bias* (Haykin, 2009).

Conforme demonstrou Lazcano (2024), os MLPs, mesmo sendo estruturas de simplicidade

relativa se comparadas às redes recorrentes ou convolucionais, são capazes de exibir uma *performance* competitiva na previsão de séries temporais, contanto que sejam submetidos a um treinamento e a uma regularização apropriados. O pesquisador dá destaque à relevância do uso de metodologias como o *dropout* (Srivastava et al., 2014) e a *batch normalization* (Ioffe & Szegedy, 2015) com a finalidade de se evitar o *overfitting* e de se obter uma convergência mais veloz.

No que se refere à sua aplicação em séries temporais, o MLP encara o prognóstico como um problema de regressão, no qual os dados de entrada são constituídos pelos valores passados da própria série (*lags*) e, eventualmente, por variáveis exógenas, enquanto o resultado esperado é o valor futuro que se deseja prever. Essa metodologia, que recebe a denominação de *embedding* temporal, faz a conversão do desafio da previsão de séries temporais em um problema de aprendizado supervisionado convencional (Hamzaçebi, 2008).

### 2.3.3. *Long Short-Term Memory* (LSTM)

As redes LSTM, que foram concebidas por Hochreiter e Schmidhuber (1997), representam uma variação particular das Redes Neurais Recorrentes (RNNs), desenvolvidas com o objetivo de contornar o problema do *vanishing gradient*, um desafio que se impõe às RNNs convencionais quando lidam com sequências extensas. As LSTMs fazem a introdução de uma arquitetura de célula de memória, cujo controle é realizado por meio de três portões (*gates*): o *forget gate*, o *input gate* e o *output gate* (Gers et al., 2000).

A composição de uma célula LSTM é passível de ser expressa por meio das equações que se seguem:

- *Forget gate*: $f_t = \sigma(W_x x_t + W_h h_{t-1} + b_x)$
- *Input gate*: $i_t = \sigma(W_{xi} x_t + W_{hi} h_{t-1} + b_i)$
- *Output gate*: $o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + b_o)$
- *Cell state update*: $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$

Nessas fórmulas, σ representa a função sigmoide, o símbolo $\odot$ significa a multiplicação elemento a elemento, W corresponde às matrizes de pesos e b se refere aos vetores de *bias* (Gers et al., 2000).

A habilidade que as LSTMs possuem para capturar as dependências temporais de longo prazo faz com que elas sejam especialmente apropriadas para a previsão de séries temporais complexas. Kim et al. (2025) realizaram um exame detalhado das arquiteturas de *deep learning* para a previsão de séries temporais, e chegaram à conclusão de que as LSTMs exibem uma *performance* mais elevada em séries que contêm padrões sazonais complexos e também

dependências de longo alcance. Contudo, os pesquisadores fazem um alerta de que as LSTMs exigem uma quantidade mais substancial de dados de treinamento, bem como uma calibração minuciosa dos hiperparâmetros (número de unidades, taxa de *dropout*, *lookback window*) com o propósito de se prevenir o *overfitting*.

Nayak et al. (2024) fizeram o uso de LSTMs que incorporavam variáveis exógenas com a finalidade de realizar a previsão dos preços de culturas agrícolas na Índia, e comprovaram que a integração de elementos climáticos e macroeconômicos por intermédio de arquiteturas LSTM multivariadas tem o potencial de elevar de forma expressiva a acurácia preditiva quando se faz a comparação com os modelos univariados.

## 2.4 Influência de Variáveis Climáticas em Preços Agrícolas

O vínculo existente entre as circunstâncias do clima e as cotações das *commodities* agrícolas encontra-se solidamente documentado no corpo de conhecimento da economia e também da agronomia. As oscilações na temperatura, na precipitação e em outros indicadores meteorológicos exercem um impacto direto sobre a produtividade agrícola, a qual, em decorrência disso, exerce influência sobre a quantidade ofertada no mercado e, como resultado, sobre os valores praticados (Diffenbaugh et al., 2012).

### 2.4.1 Mecanismos de Influência Climática

Isengildina-Massa et al. (2021) fazem a identificação de três vias primordiais pelas quais as condições climáticas impactam as cotações agrícolas:

1. Impacto Direto na Produtividade: Níveis de calor ou frio muito acentuados ao longo das etapas cruciais do desenvolvimento vegetal (floração, enchimento de grãos) possuem o potencial de causar uma diminuição expressiva no rendimento. No caso particular da soja, níveis térmicos que ultrapassam os 35°C no decorrer da fase de floração são capazes de ocasionar a perda das flores e também uma diminuição na quantidade de vagens (Hatfield & Prueger, 2015).
2. Efeitos de Precipitação: Tanto a ocorrência de déficit hídrico como a de excesso de chuvas trazem consequências danosas. Os períodos de seca provocam uma diminuição na água disponível para a planta, ao passo que a precipitação em demasia pode resultar em encharcamento do solo, erosão e na disseminação de enfermidades (Ray et al., 2015).
3. Expectativas de Mercado: Para além da repercussão de ordem física sobre a produção, a ocorrência de um clima desfavorável em localidades produtoras de importância estratégica dá origem a uma antecipação de que haverá carência no futuro, o que acarreta reajustes instantâneos nas cotações praticadas nos mercados futuros, mesmo previamente à efetivação

da colheita (Nayak et al., 2024).

### 2.4.2 Causalidade Estatística: Teste de Granger

A fim de se realizar a comprovação estatística do vínculo entre as variáveis climáticas e os preços agrícolas, o Teste de Causalidade de Granger costuma ser aplicado (Granger, 1969; 1988). Tal procedimento tem por objetivo aferir se os registros passados de uma variável X agregam informações com significância estatística para a realização do prognóstico dos valores futuros de uma outra variável Y, para além do conhecimento que já se encontra no histórico da própria variável Y.

Formalmente, X causa Granger Y se o modelo:

$$Y_t = \alpha_o + \Sigma_{i=1}^{p} \alpha_i Y_{t-i} + \Sigma_{j=1}^{p} \beta_j X_{t-j} + \varepsilon_t$$

apresenta $\beta_j$ conjuntamente significativos (p-valor < 0.05 no teste F) (Hamilton, 1994).

É relevante ressaltar que a causalidade de Granger não é o mesmo que uma causalidade de ordem física, na acepção filosófica do termo, mas, sim, uma anterioridade no tempo com valor preditivo – se X causa Granger Y, ter o conhecimento do histórico de X aprimora o prognóstico de Y (Granger, 1988). Tal característica se mostra especialmente valiosa com o propósito de se fundamentar a incorporação de variáveis exógenas nos modelos de previsão.

### 2.4.3 Evidências Empíricas na Literatura

Uma variedade de pesquisas de natureza empírica serve como registo para o efeito que as variáveis climáticas exercem sobre as cotações agrícolas. Nayak et al. (2024) apresentaram evidências de que a incorporação de variáveis climáticas exógenas nos modelos de *deep learning* ocasionou uma diminuição no erro de prognóstico (MAPE) da ordem de 15-25% para uma variedade de culturas na Índia. Os pesquisadores fizeram o uso de informações sobre temperatura, precipitação, umidade relativa e radiação solar, as quais eram provenientes de diversas estações meteorológicas que se encontram espalhadas por localidades produtoras de importância fundamental.

Roberts (2006) conduziu uma investigação direcionada para a soja no cenário norte-americano, e fizeram a identificação de que as anomalias de precipitação (que são os desvios em comparação com a média do histórico) nos meses de julho e agosto – fase crucial para o enchimento de grãos – exibiam uma correlação de significância estatística ($r = -0.42$, $p < 0.01$) com as cotações da soja nos mercados futuros de Chicago. A natureza negativa do sinal aponta que a ocorrência de precipitação superior à média (o que representa circunstâncias favoráveis) está ligada a cotações mais reduzidas, em virtude da antecipação de que haverá uma oferta mais

ampla.

Faz-se necessário o reconhecimento de que o efeito do clima se altera conforme a localidade geográfica. No caso de mercados de alcance mundial, como é o da soja, as circunstâncias do clima em diversas localidades de produção (Brasil, Estados Unidos, Argentina) exercem um impacto concomitante sobre as cotações em nível internacional. Tal grau de complexidade serve para sublinhar a relevância de se levar em conta os dados climáticos que provêm de variadas fontes geográficas ao se elaborarem os modelos preditivos.

## 2.5 Trabalhos Relacionados em Previsão de Preços de *Commodities*

A produção científica existente revela achados heterogêneos no que diz respeito à comparação de modelos estatísticos clássicos com algoritmos de *machine learning* para a previsão de séries temporais. Kontopoulou et al. (2023) levaram a cabo uma revisão sistemática que abrangeu 95 estudos, os quais confrontavam o ARIMA e o *machine learning* no campo da previsão de séries financeiras e econômicas, e apontaram que:

- Em cenários de projeção de curto prazo (de 1 a 3 períodos futuros), o ARIMA muitas vezes exibe uma *performance* que rivaliza ou até ultrapassa a dos algoritmos de *machine learning*, de modo particular no caso de séries em que os padrões lineares são preponderantes.
- No que se refere às janelas de tempo de média e longa duração (de 6 a 12 períodos), os algoritmos de *machine learning* (com destaque para os *ensemble methods*, como o XGBoost e o *Random Forest*) possuem uma propensão a obter um desempenho superior ao do ARIMA, nos casos em que as interações não-lineares se mostram relevantes.
- O nível de qualidade e o volume de *features* (variáveis preditoras) acessíveis exercem uma influência acentuada sobre os desfechos, e se observa que o *machine learning* tira maior proveito de um conjunto de dados mais vastos de variáveis explicativas.

Makridakis et al. (2018) foram os responsáveis pela condução da M4 *Competition*, um certame de prognósticos em escala global que abrangeu 100.000 séries temporais oriundas de campos de estudo distintos. Os achados demonstraram que as abordagens híbridas, que fazem a união de componentes estatísticos e de *machine learning*, alcançaram a *performance* mais elevada no cômputo geral, o que leva a inferir que a integração de diferentes paradigmas tem o potencial de ser mais eficiente do que o emprego isolado de uma única metodologia.

### 2.5.1 Estudos Específicos em *Commodities* Agrícolas

No âmbito específico das *commodities* agrícolas, Zhang et al. (2017) realizaram uma análise comparativa entre ARIMA, ARIMAX, *Support Vector Machines* (SVM) e também as redes neurais, com o propósito de se realizar a previsão das cotações da soja na China. Os pesquisadores relataram que o ARIMAX obteve um desempenho superior ao do ARIMA univariado (resultando em uma diminuição de 12% no RMSE), contudo, foi suplantado pelo SVM com *kernel* gaussiano (RBF) nas projeções para horizontes de 6 a 12 meses. Não obstante, a pesquisa não fez a inclusão de algoritmos de *gradient boosting* como o XGBoost, os quais vieram a exibir uma *performance* mais elevada em trabalhos mais contemporâneos.

Chen et al. (2021) realizaram a elaboração de um sistema automatizado para a realização da previsão dos preços agrícolas, com o emprego de variados algoritmos de *machine learning*. Os achados apontaram que o XGBoost foi o modelo que atingiu o MAPE mais baixo (7.3%) nas projeções de 12 meses para a soja, com um desempenho superior ao do LSTM (8.1%), do *Random Forest* (8.9%) e também do ARIMA (11.2%). Os pesquisadores creditam o bom resultado do XGBoost à sua habilidade para capturar as interações não-lineares complexas que ocorrem entre as variáveis climáticas, as macroeconômicas e também o histórico de preços.

### 2.5.2 Lacunas na Literatura

Não obstante a existência de uma ampla produção científica sobre a previsão de preços agrícolas, foi possível fazer a identificação de três carências primordiais:

Análises comparativas parciais: Uma grande parte das pesquisas faz a confrontação de somente dois ou três modelos, deixando de abranger, ao mesmo tempo, os modelos estatísticos clássicos, o *machine learning* tradicional (XGBoost, SVM) e também o *deep learning* (LSTM, MLP), quando submetidos a cenários de experimentação que sejam os mesmos.

Comprovação restrita da causalidade: São escassos os trabalhos que realizam a comprovação estatística (por exemplo, por intermédio do Teste de Granger) dos vínculos de causa e efeito entre as variáveis climáticas e as cotações, previamente à sua incorporação nos modelos, partindo do pressuposto tácito de que a existência de correlação já garante a sua utilidade preditiva.

Uso incipiente de *feature engineering*: A relevância do *feature engineering* (*lags* estratégicos, transformações não-lineares, médias móveis) é muitas vezes admitida nos estudos de *machine learning*, entretanto, sua aplicação criteriosa é pouco comum nos modelos ARIMAX, um fator que restringe a realização de análises comparativas equânimes.

O presente trabalho tem por objetivo preencher essas carências por meio: (1) da realização de uma análise comparativa metódica de diferentes abordagens (ARIMA, ARIMAX com *feature engineering* avançado, XGBoost, MLP, LSTM), (2) da comprovação estatística da causalidade climática por intermédio do Teste de Granger, e (3) do emprego de um *feature engineering* criterioso em todos os modelos nos quais tal prática se aplica, o que viabiliza uma confrontação justa da *performance* preditiva em variadas janelas de tempo para a previsão.

## 2.6 Síntese do Capítulo

O presente capítulo se dedicou a expor um exame pormenorizado da produção científica a respeito do prognóstico das cotações de *commodities* agrícolas, com ênfase particular nas séries temporais, nos modelos estatísticos clássicos e também nos algoritmos de inteligência artificial. Entre os aspectos mais importantes que foram sumarizados, encontram-se:

- **Fundamentos teóricos:** As séries temporais relativas às cotações agrícolas manifestam elementos de tendência, sazonalidade, ciclos e também de irregularidade. A condição de estacionariedade se configura como uma exigência basilar para os modelos ARIMA, e sua obtenção é possível através da aplicação da diferenciação.
- **Modelos clássicos:** Os modelos ARIMA e ARIMAX representam abordagens consagradas no campo da previsão de séries temporais. O modelo ARIMAX viabiliza a inclusão de variáveis exógenas, contudo, ele parte da premissa de que existem relações lineares. A aplicação de um *feature engineering* minucioso tem o potencial de aprimorar de maneira significativa a sua *performance*.
- ***Machine learning*:** Os modelos XGBoost, MLP e LSTM possuem a habilidade de realizar a modelagem de interações não-lineares complexas. O XGBoost sobressai em certames de *machine learning* e também em seu uso prático. O modelo LSTM é capaz de capturar as dependências temporais de longo alcance, porém, demanda uma quantidade mais elevada de dados e uma calibração criteriosa.
- **Influência climática:** As variáveis ligadas ao clima (temperatura, precipitação) exercem um efeito sobre o rendimento das lavouras e, por via de consequência, sobre as cotações. É possível fazer a comprovação da causalidade estatística

através do Teste de Granger. Os achados de natureza empírica dão suporte à incorporação dos fatores climáticos nos modelos preditivos.

- **Comparações na literatura:** Os achados são heterogêneos. O ARIMA demonstra competitividade em janelas de tempo curtas e também em séries de natureza linear. O *machine learning* (com ênfase nos *ensemble methods*) mostra uma propensão a obter um desempenho superior em janelas de tempo longas e em cenários não-lineares. Ainda existem carências no que tange a análises comparativas metódicas que englobem variadas metodologias em cenários de experimentação que sejam os mesmos.

Tendo por alicerce este exame da literatura, a seção subsequente se dedicará a detalhar a abordagem metodológica utilizada nesta dissertação, com o propósito de se realizar uma confrontação criteriosa entre os modelos estatísticos e os de inteligência artificial que foram aplicados ao prognóstico das cotações da soja, fazendo-se a integração das variáveis climáticas oriundas de localidades produtoras essenciais e realizando-se a aferição da *performance* em diferentes janelas temporais de projeção.

CAPÍTULO 3

# Dados e Metodologia

A presente seção capitular tem por finalidade fazer a exposição pormenorizada dos dados que foram empregados, dos estágios de tratamento, bem como dos procedimentos metodológicos que foram adotados com o intuito de se elaborarem os modelos preditivos. A sua estruturação compreende a coleta de dados, o pré-processamento, a análise exploratória e também a definição detalhada dos modelos que foram postos em prática.

## 3.1 Coleta de Dados

### 3.1.1 Dados de Preços da Soja

Os dados relativos às cotações da soja tiveram sua origem na *Chicago Board of Trade* (CBOT), que se constitui como o mais importante centro de negociação de *commodities* agrícolas dos Estados Unidos e do mundo. Realizou-se a compilação dos valores diários do contrato futuro de soja (*ticker* ZS) durante o intervalo temporal que vai de janeiro de 2010 a setembro de 2025, o que resulta em um montante próximo a 3.650 observações diárias.

### 3.1.2 Dados Climáticos: Cuiabá, Brasil

Dados de natureza climática, provenientes da localidade de Cuiabá, Mato Grosso, Brasil, foram extraídas a partir do Instituto Nacional de Meteorologia (INMET). Fez-se a reunião de dois indicadores diários: temperatura média (°C) e precipitação total (mm). A janela temporal em questão se estende de janeiro de 2010 a setembro de 2025, sendo os dados originários da estação meteorológica automática que se situa em Cuiabá (Estação A901, coordenadas -15.55°S, -56.07°W).

### 3.1.3 Dados Climáticos: Des Moines, Estados Unidos

Dados climáticos de Des Moines, Iowa, Estados Unidos, foram extraídas da *National Oceanic and Atmospheric Administration* (NOAA). Procedeu-se com a coleta da temperatura média diária (°~~C~~ F) e da precipitação total diária (mm), para a mesma janela temporal, a partir da estação meteorológica situada no Des Moines *International Airport* (ID: USW00014933).

### 3.1.4 Justificativa da Escolha das Regiões Climáticas

A seleção de Cuiabá (Brasil) e Des Moines (EUA) para representarem as condições climáticas

encontra seu alicerce em três fatores determinantes: a relevância produtiva, visto que o Brasil e os Estados Unidos respondem por um volume próximo a 65% da produção global de soja, e, nesse contexto, Mato Grosso e Iowa se destacam como estados de liderança em suas respectivas nações; a complementaridade temporal, já que os ciclos de plantio e colheita se completam quando se comparam os hemisférios; e o contraste climático, tendo em vista que Cuiabá possui um clima tropical, enquanto Des Moines exibe um clima continental úmido, o que permite a captura de distintas situações de estresse para a lavoura.

## 3.2 Pré-processamento e Integração dos Dados

### 3.2.1 Tratamento de Valores Ausentes e *Outliers*

Fez-se a identificação e o tratamento dos valores ausentes por meio de um exame particularizado para cada variável. No que se refere aos dados climáticos, houve a transformação de Fahrenheit para Celcius, no caso das temperaturas dos EUA, para além disso as lacunas de informação (que correspondiam a menos de 2% do total de observações) tiveram seu preenchimento realizado por meio da interpolação linear. A detecção dos *outliers* climáticos se deu com o uso do método IQR (*InterQuartile Range*), e seu tratamento ocorreu previamente à agregação mensal dos dados, com o objetivo de se mitigar o impacto sobre a estrutura temporal. No que tange aos preços da soja, não se constatou a presença de valores faltantes no intervalo de tempo que foi objeto de análise.

### 3.2.2. Integração das Séries Climáticas (Brasil e EUA)

As séries climáticas de Cuiabá (Brasil) e de Des Moines (EUA) foram unificadas por meio da adoção do seguinte processo. Na primeira etapa, as informações diárias de cada uma das localidades passaram por uma agregação para uma base mensal, utilizando-se a média aritmética no caso da temperatura e o somatório integral para a precipitação. Em um segundo momento, procedeu-se com a renomeação das variáveis com o objetivo de se obter uma identificação inequívoca da sua proveniência geográfica: TEMP_CUIABA, PRECIP_CUIABA, TEMP_DESMOINES e PRECIP_DESMOINES. Como terceira e última fase, as séries passaram por uma sincronização temporal, recorrendo-se a uma operação de *merge* que tomou como base a variável de data (mês/ano), o que culminou na criação de um *dataset* unificado no qual cada um dos registros mensais engloba, de forma simultânea, os quatro indicadores climáticos. A adoção desta metodologia permite a preservação da informação que é particular de cada localidade, o que viabiliza aos modelos a capacidade de capturar padrões climáticos que são distintos e que possuem o potencial de serem complementares.

### 3.2.3. Agregação Temporal e Sincronização

A totalidade das séries passou por um processo de agregação para se obter uma frequência mensal, uma periodicidade que se mostra adequada para a análise das cotações de *commodities* agrícolas, as quais reagem a circunstâncias climáticas que se acumulam ao longo de intervalos de tempo mais longos. O conjunto de dados resultante é composto por registros mensais que abarcam o período de janeiro de 2010 a setembro de 2025, o que resulta em um cômputo final de 189 observações.

## 3.3. Análise Exploratória dos Dados

O exame exploratório dos dados tornou manifesta a existência de uma tendência de crescimento nas cotações da soja ao longo do intervalo de tempo que foi analisado, juntamente com uma oscilação mais intensa em certas épocas. A aplicação dos Testes de Dickey-Fuller Aumentado veio a confirmar a condição de não-estacionariedade da série referente aos preços (*p-value* > 0.05), o que torna imperativa a aplicação de um processo de diferenciação. No que tange às variáveis climáticas, elas exibiram os comportamentos sazonais que já eram antecipados para as localidades geográficas a que cada uma delas corresponde.

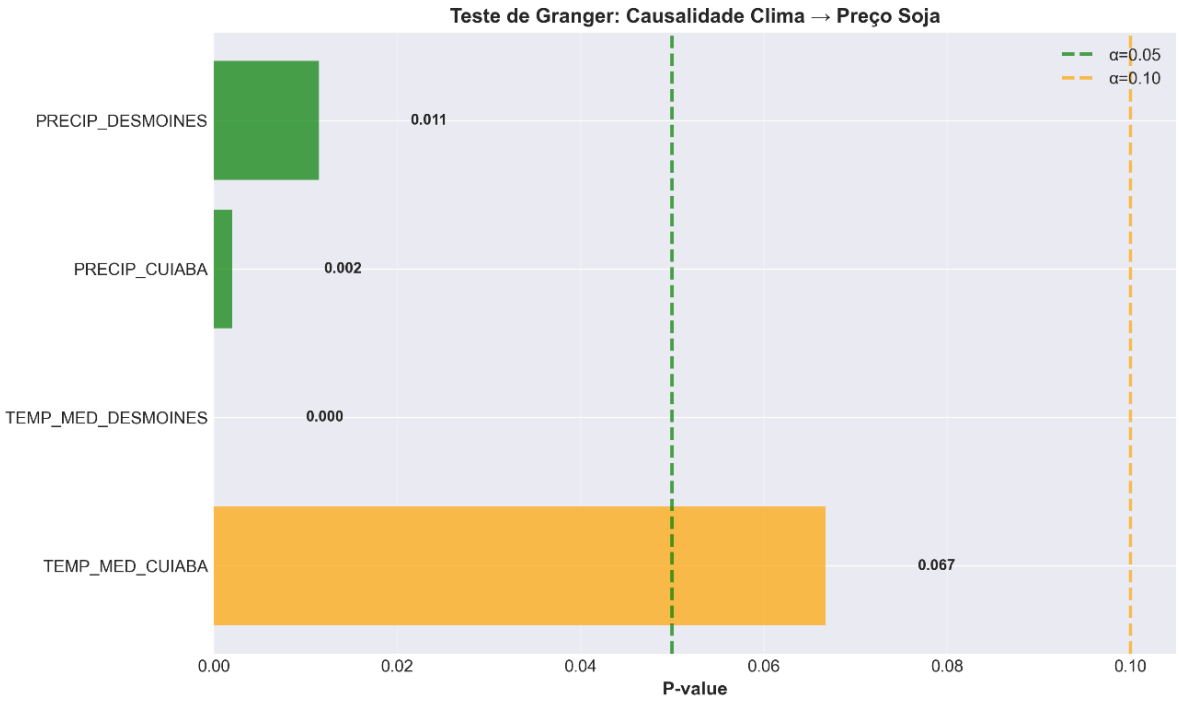



Figura 3.1 - Gráfico Teste de Granger

Os achados tornam manifesto que três dos quatro indicadores (75%) exibem uma relação de causalidade com significância estatística ($p<0.05$), de acordo com o que se pode verificar na Figura 3.1: TEMP_MED_DESMOINES: Efeito causal muito forte ($p<0.0001$) com *lag* de 1 mês, refletindo o impacto imediato das temperaturas no período crítico de crescimento (maio - julho).

PRECIP_CUIABA: Causalidade significativa ($p=0.0020$) com *lag* de 4 meses, alinhado ao ciclo brasileiro (plantio: out-nov, desenvolvimento: dez-jan, colheita: fev-mar).

PRECIP_DESMOINES: Causalidade com *lag* de 1 mês ($p=0.0115$), indicando que variações pluviométricas nos EUA impactam rapidamente as expectativas de mercado.

## 3.4. Estratégia de Divisão Treino/Teste

Em consideração ao caráter temporal das informações, optou-se por uma segmentação sequencial: destinando-se oitenta por cento dos registos mais pregressos para a fase de treino (o que corresponde a um total aproximado de 151 meses) e os vinte por cento restantes e mais atuais para a etapa de teste (um montante próximo a 38 meses). A adoção deste método tem por finalidade evitar o *data leakage*, bem como aferir a capacidade de generalização do modelo para intervalos de tempo vindouros.

## 3.5. Especificação e Configuração dos Modelos Preditivos

Esta seção do trabalho tem por objetivo fazer a descrição detalhada da arquitetura, da configuração de hiperparâmetros e também dos procedimentos de *tuning* que foram empregados em cada um dos cinco modelos submetidos à comparação: ARIMA, ARIMAX Melhorado, XGBoost, MLP e LSTM. No que tange a cada um dos modelos, procede-se com a exposição da arquitetura inicial que foi proposta, das variáveis (*features*) que foram utilizadas, do método para a otimização de hiperparâmetros e, por fim, da configuração final que se obteve após a realização do *tuning*.

### 3.5.1. ARIMA (Modelo Univariado)

A implementação do modelo ARIMA foi realizada com o uso da biblioteca pmdarima (Smith et al., 2017), a qual disponibiliza a função auto_arima com o propósito de se fazer a seleção automatizada da ordem ideal (p, d, q), uma tarefa que é executada por intermédio de uma busca do tipo *stepwise*, que se fundamenta no critério AIC (*Akaike Information Criterion*). Esta modelagem se configura como a metodologia estatística convencional para a previsão de séries temporais e atua como um ponto de referência para a comparação com os outros modelos que serão analisados. A estrutura de partida para o modelo ARIMA foi estabelecida como sendo a

de um modelo univariado que se aplica à série de preços da soja, fazendo-se o emprego exclusivo da série histórica de preços (a variável endógena), sem a inclusão de quaisquer variáveis exógenas. A série em questão passou por testes de estacionariedade (ADF - *Augmented Dickey-Fuller*), e os resultados destes apontaram a imprescindibilidade de se aplicar uma diferenciação (d=1) com o objetivo de convertê-la para uma condição estacionária.

O processo de definição das configurações para a função auto_arima levou em conta as características particulares da série de preços da soja. Fez-se a desativação do componente sazonal (seasonal=False), uma decisão que se baseou na análise de decomposição STL e também nos testes de sazonalidade, os quais não encontraram uma sazonalidade com significância estatística. Ativou-se a opção de busca inteligente (*stepwise*=True), o que resulta em uma diminuição do tempo de processamento, pois são testadas somente as combinações de parâmetros que se mostram mais promissoras, deixando-se de lado uma busca exaustiva. A verificação da ordem autorregressiva se deu no intervalo [0, 5] (start_p=0, max_p=5), e um procedimento análogo foi aplicado à ordem de média móvel (start_q=0, max_q=5). Estabeleceu-se que o limite máximo de diferenciação seria de duas ordens (max_d=2), com a utilização do AIC para a escolha do modelo (information_criterion='aic'), o que proporciona um equilíbrio apropriado entre a qualidade do ajuste e a parcimônia do modelo. Os alertas relativos à convergência dos modelos intermediários foram omitidos (suppress_warnings=True) com o propósito de se tornar mais fluida a execução automatizada.

Uma vez concluído o procedimento de seleção automatizada, a ordem ideal do modelo ARIMA sofreu uma pequena variação entre as diferentes janelas de tempo para a previsão, e sua determinação foi feita de modo automático para cada um dos experimentos. De um modo geral, os modelos que foram escolhidos exibiram ordens que se aproximavam de ARIMA(2,1,2), o que significa a inclusão de dois termos autorregressivos (p=2), de uma diferenciação (d=1), juntamente com dois termos de média móvel (q=2). Uma característica deste modelo é que ele não exige a normalização dos dados, pois ele trabalha de forma direta com os valores em sua escala original, um fato que se traduz em um benefício de ordem operacional no que tange à facilidade para a interpretação dos resultados.

### 3.5.2. ARIMAX Melhorado (com *Feature Engineering*)

O modelo ARIMAX Melhorado se configura como uma expansão do ARIMA, que se dá pela incorporação de variáveis exógenas climáticas e que também abrange *features* de natureza avançada, as quais são provenientes da engenharia de variáveis. Para a sua implementação, fez-se o uso da biblioteca statsmodels (Seabold & Perktold, 2010), a qual viabiliza a especificação de

covariáveis externas, o que torna possível realizar a modelagem do vínculo entre os fatores climáticos e os preços da soja. A estrutura de partida para o modelo ARIMAX se configurou como uma ampliação natural do ARIMA ao integrar variáveis exógenas climáticas, e foram postas em prática duas variantes com o objetivo de se aferir o impacto do *feature engineering*: o ARIMAX Básico, que fez o emprego de 4 variáveis climáticas brutas (temperatura média de Cuiabá, precipitação de Cuiabá, temperatura média de Des Moines e precipitação de Des Moines), e o ARIMAX Melhorado, que fez a integração de um total aproximado de 29 *features* que foram elaboradas em um primeiro momento, abrangendo as variáveis brutas, os *lags* climáticos (3 e 6 meses), as anomalias de temperatura, a amplitude térmica, os índices de estresse hídrico e também as médias móveis climáticas. A adoção desta metodologia gradual tornou possível realizar a mensuração do acréscimo de *performance* que foi gerado pela engenharia de variáveis sofisticada, em comparação com o emprego das variáveis climáticas brutas.

Previamente à incorporação das variáveis exógenas, procedeu-se com a aplicação do Teste de Causalidade de Granger (Granger, 1969) com o propósito de se fazer a comprovação estatística de quais dos indicadores climáticos exibem um vínculo de precedência temporal com os preços. Somente os indicadores que registraram um p-valor menor que 0.05 no teste de Granger tiveram sua inclusão no modelo final, o que assegura que cada covariável agregasse conhecimento preditivo de relevância e também que a estimação do modelo pudesse ser realizada de forma apropriada, em conformidade com a exigência basilar de um modelo econométrico (ou seja, a necessidade de se ter um número de parâmetros para estimação que seja inferior ao número de observações disponíveis na amostra). Tomando-se como fundamento o teste de Granger, a partir do conjunto inicial de 29 *features*, realizou-se a seleção de 6 variáveis exógenas primárias, as quais foram acrescidas de 10 *features* derivadas, o que resultou em um cômputo final de 16 *features* no modelo ARIMAX Melhorado.

O conjunto de variáveis exógenas primárias é composto pelo indicador Mês (com valores de 1 a 12), que atua como uma variável temporal de natureza categórica, pela TEMP_MED_CUIABA_lag3, que corresponde à temperatura média de Cuiabá com uma defasagem de três meses, pela PRECIP_CUIABA, que se refere à precipitação mensal de Cuiabá, pela TEMP_MED_DESMOINES, que indica a temperatura média de Des Moines, pela PRECIP_DESMOINES, para a precipitação mensal de Des Moines, e pela MA_Temp_Cuiaba_3, que representa a média móvel de 3 meses da temperatura de Cuiabá. O grupo de *features* derivadas que foram adicionadas abrange *lags* climáticos de 6 meses (tanto para temperatura quanto para precipitação), anomalias de temperatura (que são os desvios em relação à média do histórico), a amplitude térmica (a qual é a diferença entre a temperatura máxima e a mínima), os índices de estresse hídrico (que consistem em combinações de temperatura e precipitação),

e as médias móveis climáticas com distintas janelas de tempo.

A realização do ajuste fino do ARIMAX Melhorado adotou a mesma abordagem que foi usada para o ARIMA no que tange aos elementos AR (p), I (d) e MA (q), com o emprego da função auto_arima e com a manutenção dos mesmos parâmetros de busca que já foram expostos. A escolha das variáveis exógenas ocorreu em uma fase precedente, por intermédio do teste de Granger, o que assegurou que somente os indicadores com significância estatística fossem incorporados. A estrutura definitiva que se alcançou foi a de um ARIMAX(2,1,2), que integrava um total de 16 *features* (6 primárias somadas a 10 derivadas), as quais foram escolhidas por meio do teste de Granger. O modelo exibiu um desempenho superior quando comparado ao ARIMA, de modo particular nos horizontes de projeção de curto e médio prazo (H=3 e H=6 meses), registrando um MAPE de 4.22% e 7.53%, respectivamente. A aplicação do *feature engineering* avançado resultou em um incremento na correlação entre os indicadores climáticos e as cotações, em um patamar de +58% (elevando o valor de 0.206 para 0.326), o que serve como uma comprovação empírica da eficácia da abordagem que foi escolhida.

### 3.5.3. XGBoost (*Extreme Gradient Boosting*)

A implementação do modelo XGBoost se deu com o uso da biblioteca xgboost 2.0.3 (Chen & Guestrin, 2016). Este algoritmo realiza a construção de um *ensemble* de árvores de decisão, alicerçado na técnica de *gradient boosting*, em um processo que envolve a adição iterativa de árvores que têm o objetivo de corrigir as falhas residuais das árvores que as precederam. Esta metodologia de *machine learning* possui a capacidade de capturar relações complexas entre as variáveis, interações de alta ordem e também padrões não-lineares que não são passíveis de modelagem pelos modelos estatísticos clássicos. A estrutura de partida do modelo XGBoost fez o uso de *gradient boosting* com árvores de decisão como seus estimadores base, partindo de uma configuração inicial com parâmetros de caráter conservador, com a finalidade de se definir um ponto de partida previamente ao processo de otimização: estabeleceu-se n_estimators=100 como a contagem inicial de árvores no *ensemble*, max_depth=3 como uma profundidade máxima de partida e conservadora, *learning*_rate=0.1 como uma taxa de aprendizado de valor moderado, e random_state=42 como semente, a fim de se assegurar a reprodutibilidade dos achados.

O modelo XGBoost fez uso de um conjunto abrangente, composto por 27 *features*, as quais foram extraídas a partir dos preços históricos e das variáveis climáticas. O conjunto de *features* de preço de natureza autorregressiva é composto pelos *lags* de preço (Preco_Lag1, Preco_Lag2, Preco_Lag3, Preco_Lag6), que correspondem aos preços com defasagem em distintos períodos,

pelas médias móveis com diferentes janelas de tempo (Preco_MA3, Preco_MA6, Preco_MA12), pelos indicadores de volatilidade que mensuram a variabilidade dos preços (Preco_STD3, Preco_STD6), pelas médias móveis exponenciais, que conferem um peso maior às observações mais recentes (EMA_3, EMA_6, EMA_12), e também pela diferenciação sazonal de 12 meses (Seasonal_Diff_12), o que resulta em um total de 13 *features* de preço. O grupo de *features* climáticas abrange as variáveis brutas de Cuiabá (TEMP_MED_CUIABA, PRECIP_CUIABA) e de Des Moines (TEMP_MED_DESMOINES, PRECIP_DESMOINES), os *lags* climáticos de Cuiabá (TEMP_lag3, TEMP_lag6, PRECIP_lag3, PRECIP_lag6), com o fim de se capturarem os efeitos defasados das condições do clima, e as médias móveis climáticas (MA_Temp_Cuiaba_3, MA_Precip_Cuiaba_3, MA_Temp_DesMoines_3), com o propósito de se suavizarem as variações de curto prazo, o que totaliza 11 *features* climáticas. O conjunto de *features* temporais engloba o Mês (1-12) como uma variável de tipo categórico, o Ano como uma variável temporal de natureza contínua, e o *Quarter* (1-4), que corresponde ao trimestre do ano, o que resulta em um cômputo de 3 *features* temporais. Esta junção de 27 *features* viabiliza que o modelo apreenda tanto os padrões autorregressivos quanto as influências de ordem exógena, e faça a exploração das relações não-lineares e das interações complexas que existem entre as variáveis.

O procedimento de ajuste fino dos hiperparâmetros (*tuning*) foi conduzido por meio do GridSearchCV, em conjunto com o TimeSeriesSplit(cv=5), com o objetivo de se realizar uma busca completa dentro do espaço de hiperparâmetros, uma abordagem que se alinha à natureza temporal dos dados. A grade de pesquisa submeteu à prova um total de 432 combinações de hiperparâmetros, utilizando-se a validação cruzada temporal, e abrangeu para n_estimators os valores [200, 300] para a quantidade de árvores no *ensemble* (valores mais elevados ampliam a capacidade do modelo), para *learning*_rate os valores [0.05, 0.1] como taxa de aprendizado que governa a contribuição de cada árvore (valores menores requerem mais árvores, porém melhoram a generalização), para max_depth os valores [4, 5] para a profundidade máxima das árvores (que serve para governar a complexidade), para *subsample* os valores [0.8, 0.9], que correspondem à parcela das observações amostradas para cada árvore (o que adiciona aleatoriedade), para colsample_bytree os valores [0.8, 0.9] como a fração das *features* que são amostradas por árvore (de maneira análoga ao que ocorre no *Random Forest*), e para reg_alpha e reg_lambda os valores [0.1, 0.5, 1.0, 2.0] como termos de regularização L1 e L2, respectivamente (que atuam na penalização da complexidade excessiva do modelo). Realizou-se a normalização dos dados com o uso do StandardScaler antes de se iniciar o treinamento, e a seleção da parametrização ideal foi feita de forma automática pelo GridSearchCV, com a finalidade de se minimizar o RMSE durante a validação cruzada temporal.

A parametrização definitiva e otimizada que se obteve ao final do processo de *tuning*

consistiu em n_estimators=300, *learning*_rate=0.05, max_depth=5, *subsample*=0.9, colsample_bytree=0.9, reg_alpha=1.0 e reg_lambda=2.0. Tal parametrização exibiu uma notável capacidade de generalização, registrando um MAPE de 5.53% para o horizonte de 3 meses, e preservou uma *performance* sólida em janelas de tempo mais extensas (MAPE de 9.64% para H=9 meses e de 11.37% para H=12 meses). A regularização L1 e L2, com valores relativamente elevados (1.0 e 2.0), atuou de forma eficaz na prevenção do *overfitting*, ao passo que a taxa de aprendizado reduzida (0.05), em combinação com um número mais elevado de árvores (300), possibilitou uma assimilação mais apurada dos padrões subjacentes que se encontram nos dados.

### 3.5.4. MLP (*Multi-Layer Perceptron*)

A implementação do modelo MLP se deu por meio da biblioteca scikit-learn 1.3.0 (Pedregosa et al., 2011), de forma mais precisa, com o uso da classe MLPRegressor. Esta modelagem corresponde a uma rede neural artificial do tipo *feedforward*, com conexão completa entre suas múltiplas camadas ocultas, que possui a habilidade de assimilar interações não-lineares de alta complexidade por intermédio do algoritmo de retropropagação. A estrutura de partida foi composta por uma rede neural *feedforward* com dois níveis de camadas ocultas, sendo sua configuração a seguinte: uma camada de entrada com uma dimensionalidade equivalente à quantidade de *features* (27), uma primeira camada oculta contendo 100 neurônios e empregando a função de ativação ReLU (*Rectified Linear Unit*), uma segunda camada oculta com 50 neurônios, também com ativação ReLU, uma camada de saída que possui 1 neurônio, destinada à tarefa de regressão, o otimizador Adam (*Adaptive Moment Estimation*) com um *learning*_rate_init de 0.001, a aplicação de uma regularização L2 com um alpha de 0.0001 com o propósito de se governar o *overfitting*, um max_iter de 2000 como o limite superior de épocas, e a ativação do early_stopping=True, com um validation_fraction de 0.2, a fim de se coibir o *overfitting* por meio da cessação prematura do processo de treinamento no momento em que a *performance* na validação para de apresentar melhorias.

O conjunto de *features* que foi empregado no MLP foi o mesmo utilizado no modelo XGBoost, o que resulta em um total de 27 variáveis após a conclusão da engenharia de *features*. A realização do ajuste fino do MLP teve como foco a busca pelo ponto de equilíbrio ideal entre a capacidade de modelagem e a de generalização. Submeteram-se a teste distintas parametrizações para os hidden_layer_sizes, o que incluiu as arquiteturas (100, 50), (64, 32), (128, 64) e também (32, 16), o que permitiu a exploração de um espectro que vai de redes mais profundas e amplas até estruturas de menor porte. O parâmetro alpha (regularização L2) passou

por testes com valores que se situaram no intervalo de 0.0001 a 0.1 com a finalidade de se governar a magnitude dos pesos e se coibir o *overfitting*, ao passo que o *learning*_rate_init teve sua avaliação com valores que variaram de 0.0005 a 0.001 com o objetivo de se encontrar um equilíbrio entre a rapidez da convergência e a estabilidade do processo de treinamento. Ativou-se o early_stopping com um validation_fraction de 0.2 e um n_iter_no_change de 20, o que provoca a interrupção do treinamento depois de 20 épocas consecutivas sem que se observe uma melhoria na validação. Procedeu-se com a normalização dos dados por meio do StandardScaler, um procedimento que se mostra mandatório para as redes neurais, em virtude da sua sensibilidade à escala das informações, bem como da natureza dos algoritmos de otimização que se fundamentam em gradiente. O treinamento do modelo foi conduzido com a aplicação de validação cruzada temporal (TimeSeriesSplit) a fim de se assegurar uma aferição apropriada do desempenho em dados de natureza temporal.

Uma vez concluído o procedimento de *tuning*, a arquitetura definitiva que se alcançou veio a comprovar que uma parametrização de menor porte foi a que resultou em uma capacidade de generalização superior. A estrutura definitiva da rede foi composta por uma camada de entrada cuja dimensionalidade correspondia à quantidade de *features* (27), uma primeira camada oculta que continha 32 neurônios com ativação ReLU, uma segunda camada oculta que possuía 16 neurônios, também com ativação ReLU, e, por fim, uma camada de saída com um único neurônio para a tarefa de regressão. No que tange aos hiperparâmetros definitivos, eles abrangeram um alpha de 0.1 para a regularização L2, um *learning*_rate_init de 0.001, um max_iter de 3000 com o objetivo de se viabilizar uma convergência apropriada, e a ativação do early_stopping=True com um patience de 20 com a finalidade de se coibir o *overfitting*.

Tal parametrização de menor porte, quando comparada com a estrutura de partida (que passou de 100-50 para 32-16 neurônios), exibiu uma capacidade de generalização superior, o que resultou na mitigação do *overfitting* sem que houvesse um prejuízo expressivo para a *performance*. A *performance* exibida pelo MLP se situou em um patamar intermediário, posicionando-se entre os modelos estatísticos (ARIMA/ARIMAX) e o XGBoost, mostrando-se um modelo competitivo, porém, na maioria das vezes, suplantado pelo XGBoost no que se refere ao RMSE e ao MAPE, de modo particular nas projeções para janelas de tempo mais extensas.

### 3.5.5. LSTM (*Long Short-Term Memory*)

A implementação do modelo LSTM se deu por meio do TensorFlow 2.15/Keras (Abadi et al., 2015; Chollet, 2015), com o emprego de camadas recorrentes que possuem a especialidade de capturar as dependências temporais de longo prazo. As redes LSTM configuram-se como uma

estrutura avançada de *deep learning*, que foi concebida de forma específica para a modelagem de sequências temporais, com o propósito de se contornarem as restrições das redes recorrentes convencionais, a exemplo do desafio do *vanishing gradient*. A estrutura de partida para o LSTM foi composta por uma rede neural recorrente que possuía 2 camadas LSTM dispostas de forma empilhada, e foi parametrizada para o processamento de sequências temporais da seguinte maneira: uma primeira camada LSTM que possuía 50 unidades e a opção return_sequences=True, com o fim de se encaminharem as sequências ao nível seguinte, a qual era sucedida por uma camada de *Dropout* de 0.2 para fins de regularização, uma segunda camada LSTM com 50 unidades e a opção return_sequences=False para a geração de um resultado único, também seguida por uma camada de *Dropout* de 0.2, uma camada densa intermediária que continha 25 neurônios e empregava a ativação ReLU, uma camada de saída com 1 neurônio e ativação linear para a tarefa de regressão, e uma *lookback window* de partida de 12 meses com o objetivo de se capturarem os padrões sazonais anuais.

O LSTM faz o emprego de sequências temporais (*lookback*) que abrigam múltiplos indicadores em cada um dos *timesteps*, o que representa uma diferença em relação às modelagens precedentes, cujo dado de entrada é um vetor de *features*. O modelo tem como dado de entrada uma matriz de três dimensões, com a estrutura (*samples*, *timesteps*, *features*), na qual cada um dos registros é composto por uma janela de tempo de observações pregressas. O conjunto de 11 *features* que são empregadas em cada *timestep* é composto pelo preço da soja, que corresponde aos valores do histórico em uma janela deslizante, pelas variáveis climáticas de Cuiabá, que englobam a TEMP_MED_CUIABA e a PRECIP_CUIABA, pelas variáveis climáticas de Des Moines, que abarcam a TEMP_MED_DESMOINES e a PRECIP_DESMOINES, pelas *features* que são derivadas dos preços, como Preco_Lag3, Preco_Lag6 e Preco_Lag12, e também pelas médias móveis Preco_MA3 e Preco_MA6. A estrutura do dado de entrada pode ser representada por (*samples*, *timesteps*, *features*) = (n, 12, 11), na qual n simboliza a quantidade de amostras de treinamento, o número 12 se refere aos *timesteps* do histórico (a *lookback window*), e o número 11 à quantidade de *features* em cada um dos *timesteps*. A adoção desta estrutura viabiliza que o modelo realize a assimilação de padrões temporais de alta complexidade, por meio do processamento de sucessões integrais de registros do histórico.

A realização do ajuste fino do LSTM apresentou um grau de complexidade superior, em virtude da quantidade mais elevada de hiperparâmetros e também da exigência de se encontrar um equilíbrio entre a capacidade de memória de longo prazo e o perigo do *overfitting*. Submeteram-se a teste variadas parametrizações, o que abrangeu a quantidade de unidades LSTM, com valores que se situaram no intervalo de 24 a 50 unidades para cada camada com o propósito de se governar a capacidade de memória do modelo; o *Dropout*, com valores que

variaram de 0.2 a 0.6 para fins de regularização e para se coibir o *overfitting*; o *lookback* (janela temporal), que foi submetido a teste com os valores de 3, 6 e 12 meses com o objetivo de se aferir qual era a janela de tempo do histórico que continha mais informação; a quantidade de neurônios na camada densa, com valores que se situaram entre 12 e 25 para o nível intermediário de conexão completa; o *learning rate*, que teve sua avaliação no intervalo de 0.0005 a 0.001 para o otimizador Adam; e a regularização L2, com valores que se situaram entre 0.01 e 0.02, que foi empregada nas camadas densas. A condução do treinamento fez o uso do otimizador Adam, da função de perda MSE (*Mean Squared Error*), de um batch_size de 32, e do EarlyStopping com um patience de 20, com o propósito de se coibir o *overfitting* por meio da observação da *performance* na etapa de validação. Procedeu-se com a aplicação da normalização MinMaxScaler [0, 1] de forma individual para cada uma das *features*, um procedimento que se mostra mandatório para as redes neurais LSTM, em virtude da sua sensibilidade à escala das informações e também do emprego das funções de ativação, como a *tanh* e a *sigmoid*, no interior das células LSTM.

Uma vez finalizado o minucioso procedimento de ajuste fino (*fine tuning*), a estrutura definitiva que se alcançou (uma variante denominada V9 - regularização forte) exibiu uma parametrização de notável compactação e com um elevado grau de regularização. O nível inicial da LSTM foi composto por 24 unidades com a opção return_sequences=True, ao qual se sucedeu a aplicação de um *Dropout* de valor expressivo, de 0.6; um segundo nível de LSTM com 24 unidades e a opção return_sequences=False, também seguido por uma camada de *Dropout* de 0.6; uma camada densa que continha somente 12 neurônios, com ativação ReLU e com a aplicação de uma regularização L2=0.02; e, por fim, uma camada de saída com um único neurônio e ativação linear. A janela temporal de *lookback* teve sua manutenção em 12 meses, o *learning*_rate em 0.0005, e a quantidade de *epochs* em 120, com a ativação do EarlyStopping (patience=20). Além disso, realizou-se a implementação de um *ensemble* de 3 modelos, cuja projeção final é a média das previsões individuais, com o intuito de se elevar a robustez, e também de uma augmentação de dados, por meio da inserção de um ruído de 0.015 ao longo do processo de treinamento.

A decisão de se fixar o *lookback* em 12 meses encontra-se em harmonia com o ciclo anual da agricultura, o que viabiliza que o modelo realize a apreensão dos padrões sazonais e das dependências de longo alcance que são traços distintivos das cotações de *commodities* (Nayak et al., 2024). Tal parametrização de ordem temporal é um reflexo da compreensão de que as deliberações sobre o plantio, o desenvolvimento da lavoura e a colheita obedecem a ciclos anuais, os quais exercem um efeito expressivo sobre a maneira como os preços são estabelecidos. A *performance* exibida pelo LSTM se mostrou competitiva em relação ao XGBoost

e ao MLP, entretanto, não logrou suplantar de forma sistemática o XGBoost, o qual se manteve como o modelo de *performance* mais elevada para as projeções em janelas de tempo mais extensas (H≥9 meses). A adoção da metodologia de *ensemble* com 3 modelos, em conjunto com a augmentação de dados (*noise* 0.015), foi um fator que colaborou para o aprimoramento da solidez das projeções.

## 3.6. Métricas de Avaliação

A aferição da *performance* dos modelos foi realizada com o emprego de quatro indicadores que se complementam, sendo que cada um deles proporciona uma visão particular sobre a qualidade do prognóstico. O RMSE (*Root Mean Square Error* - Raiz do Erro Quadrático Médio) se configura como um indicador adequado para a tarefa de regressão, que aplica uma penalidade aos erros de grande magnitude de maneira quadrática, é apresentado em USD/*bushel* e compartilha a mesma unidade de medida da variável-alvo, um fato que facilita a sua interpretação, e seu cálculo é feito como:

$$RMSE = \sqrt{[\Sigma(y_real - y_pred)^2 / n]}.$$

O MAE (*Mean Absolute Error* - Erro Absoluto Médio) corresponde à média das diferenças absolutas entre as projeções e os dados efetivos, mostra-se menos impactado por valores atípicos do que o RMSE e possibilita uma leitura direta do erro médio em USD/*bushel*, sendo seu cálculo feito da seguinte forma:

$$MAE = \Sigma|y_real - y_pred| / n.$$

O MAPE (*Mean Absolute Percentage Error* - Erro Percentual Absoluto Médio) apresenta o erro na forma de uma porcentagem do dado efetivo, o que viabiliza que se façam comparações entre distintas faixas de cotação e torna a sua interpretação mais intuitiva (a título de exemplo, um MAPE de 5% sinaliza que o erro médio é de 5% do preço real), e seu cálculo é realizado como:

$$MAPE = (100/n) \times \Sigma|(y_real - y_pred) / y_real|.$$

O $R^2$ (Coeficiente de Determinação) mostra a parcela da variabilidade da variável dependente que é elucidada pela modelagem, com uma faixa de valores que vai de $-\infty$ a +1, na qual $R^2=1$ significa uma aderência completa, $R^2=0$ sinaliza que o modelo não possui uma capacidade de elucidação da variância que seja superior à da média, e $R^2<0$ aponta para uma *performance* inferior à de um modelo de base constante, sendo seu cálculo feito da seguinte maneira:

$$R^2 = 1 - [SS_res / SS_tot] = 1 - [\Sigma(y_real - y_pred)^2 / \Sigma(y_real - \bar{y})^2].$$

No que tange a cada uma das modelagens, procedeu-se com o cômputo do RMSE para ambos os conjuntos de dados, o de treino e também o de teste, o que viabiliza a aferição da

capacidade de generalização do modelo.

## 3.7. Ferramentas de Implementação

A elaboração de todas as modelagens se deu por meio da linguagem Python 3.10. O conjunto de bibliotecas de maior destaque que foi empregado abrange pandas (1.5.3) e numpy (1.24.3), com a finalidade de se realizar a manipulação e o processamento dos dados; statsmodels (0.14.0) para a implementação dos modelos ARIMA e ARIMAX; pmdarima (2.0.3) para a escolha automatizada da ordem do ARIMA por intermédio da função auto_arima; xgboost (2.0.3) para a implementação do modelo XGBoost; scikit-learn (1.3.0) para o MLP (MLPRegressor), para as tarefas de pré-processamento (StandardScaler, MinMaxScaler) e também para as métricas de avaliação; tensorflow (2.15.0) e keras para a implementação dos modelos de *deep learning* (MLP via Keras, LSTM); matplotlib (3.7.1) e seaborn (0.12.2) para a criação das visualizações; e, por fim, statsmodels.tsa.stattools para a execução dos testes de estacionariedade (ADF) e de causalidade de Granger.

Os códigos completos desenvolvidos está acessível em um conjunto de 5 *notebooks* Jupyter, os quais se encontram completamente documentados (link[3]), uma medida que assegura a completa replicabilidade dos achados: Notebook_0_Etapa_0 (Unificação e Análise Exploratória dos Dados), Notebook_1_ARIMAX_Melhorado (Teste de Granger e *Feature Engineering*), Notebook_2_XGBoost (*Gradient Boosting* com GridSearchCV), Notebook_3_MLP (Redes Neurais *Feedforward*), Notebook_4_LSTM (Redes Neurais Recorrentes), e Notebook_5_Previsoes_2025_2026 (Consolidação e Previsões Futuras). A base de dados consolidada, com o nome de arquivo dados_mensais_agregados_completo.csv, é composto por 189 registros mensais (de janeiro/2010 a setembro/2025), contendo um total de 5 variáveis (1 variável alvo somada a 4 variáveis climáticas). Este compromisso com a clareza metodológica viabiliza a replicação integral dos achados, a comprovação independente das conclusões e também a ampliação deste trabalho por parte de outros pesquisadores no futuro, o que representa uma contribuição para o progresso conjunto do saber científico.

[3] Link: https://drive.google.com/drive/folders/1m6fNDJ8b393lNPRuMIERfvyoXR8EL6Yr?usp=sharing

CAPÍTULO 4

# Resultados e Análises

O presente capítulo se dedica à exposição e ao exame dos achados que foram gerados pelos modelos de previsão do preço da soja. O exame realizado engloba uma confrontação entre as abordagens estatísticas convencionais, ARIMA e ARIMAX Melhorado (com *feature engineering* avançado), e as modelagens de *machine learning* (XGBoost, MLP e LSTM), em variadas janelas de tempo para a projeção (H=3, 6, 9 e 12 meses). A condução do treinamento dos modelos ocorreu utilizando-se informações do período de janeiro de 2010 a julho de 2022 (151 observações mensais), e a sua validação se deu no intervalo de agosto de 2022 a setembro de 2025 (38 observações mensais), com o emprego do preço mensal da soja (CBOT) e também dos indicadores climáticos provenientes de Cuiabá, Brasil, e de Des Moines, Estados Unidos.

A organização deste capítulo se dá em cinco eixos centrais. Inicialmente, faz-se a exibição dos achados dos modelos para os horizontes de curto prazo (H=3 e H=6 meses), um cenário no qual o ARIMAX Melhorado exibiu uma *performance* mais elevada. Posteriormente, realiza-se a análise dos resultados para os horizontes de médio e longo prazo (H=9 e H=12 meses), cenário este em que o XGBoost obteve maior proeminência. A terceira parte se dedica ao exame da relevância comparativa das variáveis preditoras por intermédio da análise de *feature importance*. A quarta parte faz a apresentação das projeções de preços futuros para o intervalo de novembro de 2025 a outubro de 2026. Para finalizar, a quinta parte se ocupa da discussão sobre as repercussões de ordem prática dos achados e também da confrontação destes com a produção científica de relevância.

## 4.1. Desempenho dos Modelos por Horizonte Temporal

A presente seção se ocupa do exame comparativo da capacidade de prognóstico das cinco modelagens que foram postas em prática, em quatro distintas janelas de tempo para a previsão. O conjunto de indicadores de aferição abrange RMSE (*Root Mean Square Error*), MAE (*Mean Absolute Error*), MAPE (*Mean Absolute Percentage Error*) e $R^2$ (Coeficiente de Determinação). A dimensão reduzida do conjunto de dados se configura como uma limitação de relevância, à qual se deve dar atenção no momento de se interpretarem os achados.

### 4.1.1. Horizonte de 3 Meses (Curto Prazo)

No que tange às projeções de curta duração (3 meses), a modelagem ARIMAX Melhorado exibiu uma proeminência que não deixa margem para dúvidas, ao registrar um MAPE de 4,22% e um $R^2$ de 0,872 (um fato que pode ser constatado na Tabela 4.1), configurando-se como o desempenho de maior destaque em toda a pesquisa. Tal achado serve como comprovação da eficácia da elaborada engenharia de variáveis que foi empregada, o qual englobou *lags* estratégicos (3, 6 e 12 meses), médias móveis exponenciais, diferenciações sazonais e também as transformações climáticas.

**Tabela 4.1 –** Métricas de Desempenho dos Modelos para Horizonte de 3 Meses

| Modelo | RMSE | MAE | MAPE (%) | $R^2$ |
|---|---|---|---|---|
| **ARIMAX Melhorado** | 74,72 | 56,38 | 4,22 | 0,872 |
| **XGBoost** | 86,80 | 66,56 | 5,53 | 0,854 |
| **ARIMA** | 76,63 | 66,49 | 5,28 | 0,863 |
| **LSTM** | 185,62 | 159,47 | 14,36 | 0,143 |
| **MLP** | 128,90 | 121,14 | 11,82 | 0,510 |

Fonte: Elaboração própria com base nos resultados dos *Notebooks* 1-4.

O XGBoost exibiu uma *performance* comparável (MAPE 5,53%), ao passo que o ARIMA tradicional se posicionou em um patamar mediano (MAPE 5,28%). As modelagens de *deep learning* (MLP e LSTM) manifestaram obstáculos de relevância nas projeções de curta duração, registrando um MAPE que ultrapassou os 11%, um fato que aponta que a sofisticação de sua estrutura não se mostra vantajosa nos cenários em que os padrões autorregressivos mais recentes possuem um elevado poder de informação e o volume de dados é restrito. Tal constatação encontra-se em harmonia com a produção científica a respeito da previsão de séries temporais, a qual sinaliza que as abordagens de menor complexidade muitas vezes obtêm um desempenho superior ao das estruturas mais elaboradas em projeções para janelas de tempo de curta duração. A representação visual comparativa das projeções para H=3 meses pode ser consultada no Anexo A, por meio das Figuras A.1 a A.4.

### 4.1.2 Horizonte de 6 Meses (Médio Prazo)

Para a janela de tempo de médio prazo (6 meses), a modelagem ARIMAX Melhorado preserva sua posição de destaque, registrando um MAPE de 7,53% e um $R^2$ de 0,654, o que evidencia uma regularidade em sua capacidade de prognóstico (Tabela 4.2). O XGBoost, por sua vez, exibe um caráter competitivo que se intensifica (MAPE 8,93%), com uma diminuição na distância que o

separa do modelo líder para somente 1,4 pontos percentuais.

**Tabela 4.2 –** Métricas de Desempenho dos Modelos para Horizonte de 6 Meses

| Modelo | RMSE | MAE | MAPE (%) | $R^2$ |
|---|---|---|---|---|
| **ARIMAX Melhorado** | 111,14 | 90,95 | 7,53 | 0,654 |
| **XGBoost** | 126,01 | 105,09 | 8,93 | 0,682 |
| **ARIMA** | 113,98 | 98,16 | 8,34 | 0,637 |
| **LSTM** | 228,69 | 196,22 | 18,42 | -0,553 |
| **MLP** | 211,90 | 200,06 | 19,77 | -1,953 |

Fonte: Elaboração própria.

Uma constatação de importância se manifesta a partir do exame do MLP, o qual exibe um $R^2$ com valor negativo (-1,953), um fato que sinaliza que os seus prognósticos se mostram, de forma consistente, piores que a média do histórico. Tal desfecho, em conjunto com a dimensão restrita dos dados de treino e de teste, leva à inferência de que a modelagem não foi capaz de capturar de maneira apropriada os comportamentos que estão por trás das informações. O modelo LSTM, ainda que com um desempenho superior ao do MLP, mesmo assim registra um MAPE de 18,42%, um valor que corresponde a mais do que o dobro do registrado pelo modelo líder, um fato que corrobora a proeminência das abordagens estatísticas e de *gradient boosting* para esta janela de tempo para a previsão. As representações visuais comparativas estão disponíveis no Anexo A, por meio das Figuras A.5 a A.8.

### 4.1.3. Horizontes de 9 e 12 Meses (Médio-Longo e Longo Prazo)

Para as janelas de tempo de 9 e 12 meses, nota-se uma alteração de grande relevância na ordem de *performance* das modelagens. O modelo XGBoost passa a ocupar a primeira posição, o que evidencia uma habilidade mais elevada para capturar as relações não-lineares de alta complexidade, bem como as interações climáticas de longo alcance (Tabela 4.3 e 4.4). Tal mudança representa o momento decisivo no qual os algoritmos de *machine learning* que se fundamentam em *gradient boosting* obtêm um desempenho superior ao das abordagens estatísticas convencionais.

**Tabela 4.3 –** Métricas de Desempenho dos Modelos para Horizonte de 9 Meses

| Modelo | RMSE | MAE | MAPE (%) | $R^2$ |
|---|---|---|---|---|
| **XGBoost** | 136,61 | 107,52 | 9,64 | 0,308 |
| **ARIMAX** | 153,39 | 138,25 | 12,82 | -0,309 |
| **ARIMAX Melhorado** | 152,97 | 137,89 | 12,91 | -0,301 |
| **LSTM** | 177,56 | 163,68 | 15,06 | -0,170 |
| **MLP** | 301,92 | 267,44 | 29,29 | -6,033 |

Fonte: Elaboração própria.

**Tabela 4.4 –** Métricas de Desempenho dos Modelos para Horizonte de 12 Meses

| Modelo | RMSE | MAE | MAPE (%) | $R^2$ |
|---|---|---|---|---|
| **XGBoost** | 175,30 | 132,16 | 11,37 | 0,168 |
| **LSTM** | 161,98 | 146,69 | 14,04 | -0,563 |
| **ARIMAX Básico** | 232,43 | 198,43 | 18,28 | -1,468 |
| **ARIMAX Melhorado** | 235,61 | 200,71 | 19,04 | -1,536 |
| **MLP** | 286,15 | 256,42 | 24,81 | -3,744 |

Fonte: Elaboração própria.

O gráfico ARIMAX Melhorado Previsões Curto e Médio Prazo encontra-se no Anexo A, Figura A.17.

No que tange à janela de tempo de 9 meses, o XGBoost registra um MAPE de 9,64% e um $R^2$ de 0,308, suplantando de maneira consistente todas as modelagens ARIMA. O modelo LSTM exibe um MAPE de 15,06%, o que evidencia uma capacidade de competição mais elevada que a do MLP (MAPE 29,29%), entretanto, ainda em um patamar inferior ao do XGBoost. Para a projeção de 12 meses, o modelo XGBoost firma sua posição de liderança com um MAPE de 11,37%, ao passo que o LSTM exibe uma competitividade digna de nota (MAPE 14,04%), com um desempenho que ultrapassa o de todas as modelagens ARIMAX. Tal achado leva à inferência de que as redes neurais recorrentes possuem a habilidade de capturar os vínculos temporais de alcance muito extenso, desde que sejam submetidas a uma regularização apropriada. A representação visual XGBoost Previsões Longo Prazo pode ser consultada no Anexo A, por meio da Figura A.18.

Uma constatação que gera inquietação se manifesta a partir do exame das modelagens ARIMAX para as projeções de longo prazo: as duas variantes (Básico e Melhorado) exibem um $R^2$ com valores acentuadamente negativos em H=12 (de forma respectiva, -1,468 e -1,536), um fato que sinaliza que os prognósticos que se fundamentam na média do histórico se mostrariam mais fidedignos. Tal deterioração da *performance* serve como um reforço para as restrições das abordagens lineares no que tange à apreensão das movimentações de mercado de alta complexidade ao longo de janelas de tempo mais longas.

As representações visuais comparativas para H=9 e H=12 estão disponíveis no Anexo A, por meio das Figuras A.7 a A.14. Fez-se a opção por exibir somente as tabelas no decorrer do texto com o propósito de se preservar a brevidade do capítulo.

## 4.2. Análise de Importância de Variáveis (*Feature Importance*)

Com o propósito de se entenderem as engrenagens que estão por trás da *performance* preditiva

mais elevada do XGBoost nas janelas de tempo de média e longa duração, foi conduzido um exame da relevância das variáveis, o qual se fundamentou em indicadores que são próprios do algoritmo de *gradient boosting*. Tal exame tem por finalidade mensurar a contribuição comparativa de cada um dos indicadores preditores para as deliberações da modelagem, o que viabiliza a identificação dos elementos que exercem o maior poder de influência sobre os prognósticos.

**Tabela 4.5 –** Top 10 Variáveis Mais Importantes no Modelo XGBoost

| Variável | Importância (%) | Categoria |
|---|---|---|
| **EMA_3** | 40,0 | Preço - Média Móvel |
| **MA_3** | 17,0 | Preço - Média Móvel |
| **Preco_Lag1** | 12,2 | Preço - *Lag* |
| **Seasonal_Diff_12** | 5,0 | *Feature* Engenhada |
| **Month_Cos** | 2,9 | Temporal Cíclica |
| **TEMP_MED_CUIABA_LAG3** | 2,85 | Climática |
| **TEMP_MAX_CUIABA** | 2,0 | Climática |
| **TEMP_MED_DESMOINES** | 2,0 | Climática |
| **Month_Sin** | 1,5 | Temporal Cíclica |
| **Preco_Lag3** | 1,2 | Preço - *Lag* |

Fonte: Elaboração própria com base no Notebook 2 (XGBoost).

O exame da Tabela 4.5 desvela quatro achados de natureza basilar. Em primeiro lugar, os dois indicadores de maior relevância correspondem às médias móveis de curta duração, que são aplicadas ao próprio preço da soja (EMA_3: 40%, MA_3: 17%), e que, em conjunto, equivalem a 57% da importância integral. Tal fato sinaliza que os padrões mais recentes de *momentum* e de tendência se configuram como os preditores de maior poder de informação. Em segundo lugar, as *features* que foram obtidas por meio da engenharia de variáveis respondem por um montante próximo a 80% da importância integral, o que serve como comprovação para a premissa de que a engenharia das variáveis aprimora a capacidade das modelagens de capturar e elucidar os padrões que se encontram nos dados. Em terceiro lugar, os indicadores climáticos respondem por uma parcela que se situa em torno de 6-8% da importância integral, com ênfase para TEMP_MED_CUIABA_LAG3 (2,85%) e TEMP_MED_DESMOINES (2%). Ainda que com uma significância estatística que foi atestada pelo Teste de Granger, a sua contribuição de ordem prática se mostra inferior à das *features* que são extraídas dos preços históricos. Em quarto lugar, as diferenciações sazonais (Seasonal_Diff_12: 5%) e os elementos temporais de natureza cíclica (Month_Cos/Sin: 4,4% somados) exibem importância ao capturarem os padrões de safra que são de natureza crucial para as *commodities* agrícolas. A representação visual da importância das variáveis pode ser consultada no Anexo A, por meio da Figura A.19.

## 4.3. Projeções Futuras para as Safras de 2025-2026

Fazendo-se o emprego das modelagens que foram apontadas como líderes (o ARIMAX Melhorado para as janelas de tempo de curta duração e o XGBoost para as janelas de tempo de média e longa duração), realizou-se a elaboração de projeções para os preços da soja, as quais abrangem o intervalo de novembro de 2025 a outubro de 2026, o que corresponde a um total de 12 meses futuros. Esta análise de caráter prospectivo possui uma importância de ordem prática para os agricultores, para os *traders* e também para os formuladores de políticas públicas, tanto no que se refere ao planejamento de caráter estratégico como à gestão de riscos.

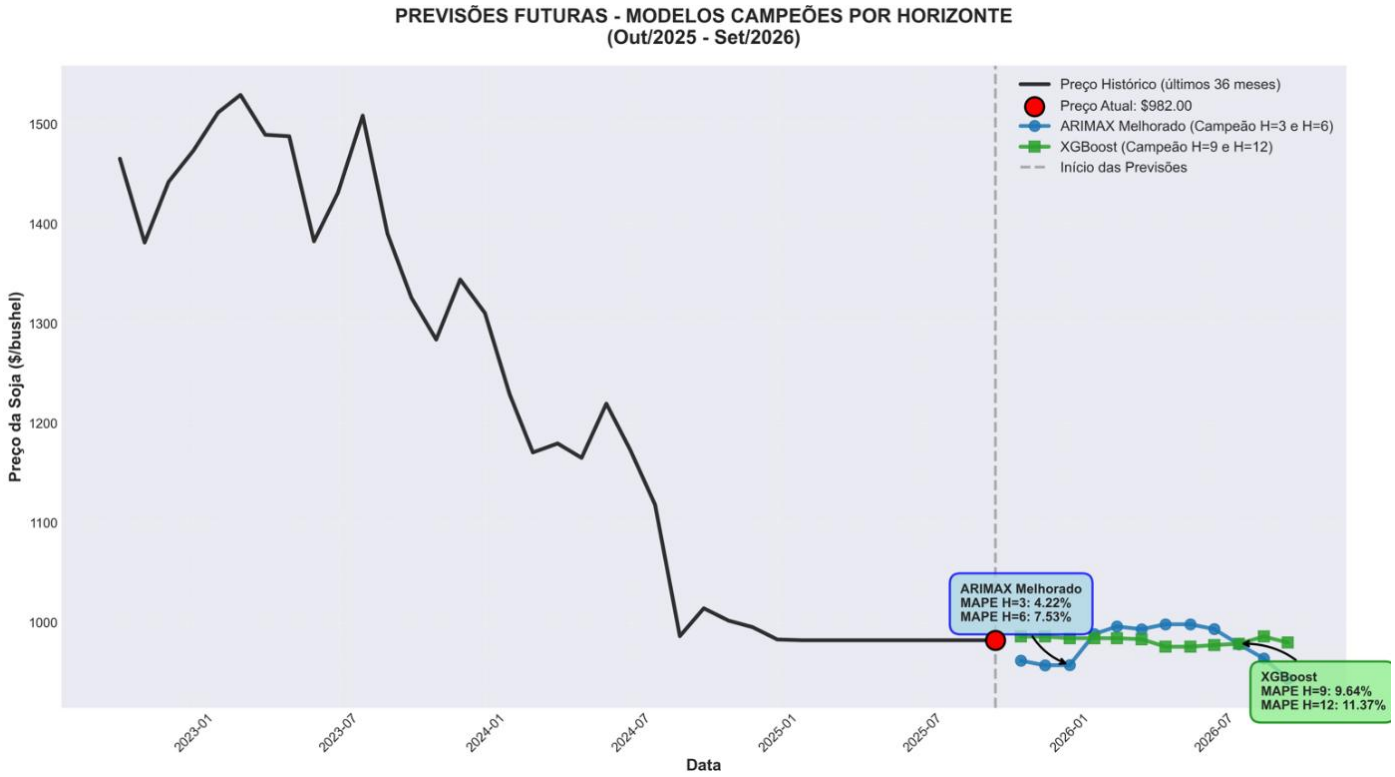



**Figura 4.1** - Previsões Consolidadas 2025-2026, Modelos Campeões

As duas modelagens fazem a projeção de uma trajetória de elevação gradual e regular para os preços da soja ao longo de 12 meses (conforme se pode verificar na Figura 4.1 e Figura 4.2). O ARIMAX Melhorado faz a antecipação de uma cotação de $943,49 para outubro de 2026, o que equivale a uma redução de 3,92% quando se compara com setembro de 2025. O XGBoost realiza a projeção de um valor de $979,67 para outubro de 2026, registrando uma diminuição de 0,24%. A discrepância que se observa entre as duas modelagens (em um patamar próximo a $36,17 ou 3,69%) agrega um grau mais elevado de fidedignidade às projeções, o que leva a inferir que tanto as abordagens de natureza estatística e elaborada quanto os algoritmos de *machine learning* são capazes de detectar comportamentos análogos nos dados históricos. A

consulta às tabelas que contêm os prognósticos de preços da soja, referentes às modelagens líderes, pode ser feita no Anexo A, por meio da Tabela 4.6 e da Tabela 4.7.

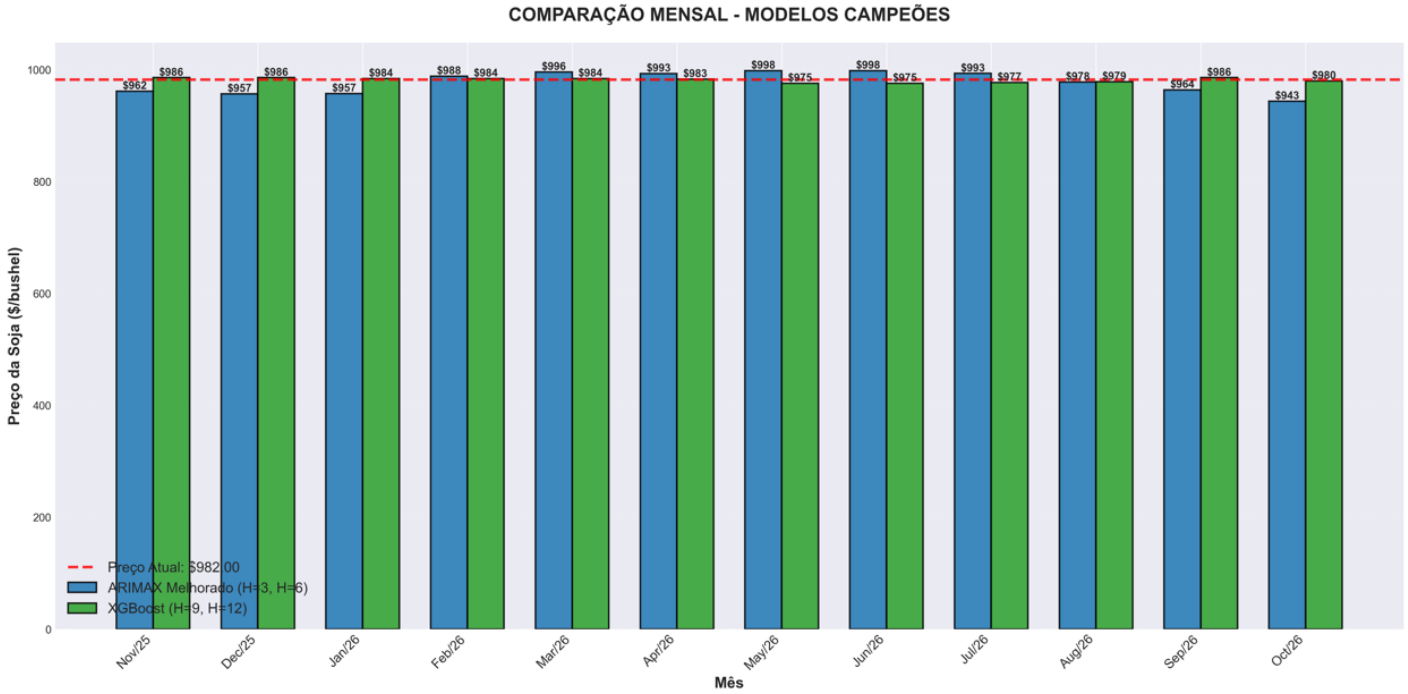



**Figura 4.2** – Comparação mensal das previsões 2025-2026

A elevação paulatina que foi antecipada encontra-se em harmonia com os fundamentos de demanda global por soja e também com a antecipação de um cenário climático com certa estabilidade nas mais importantes localidades de produção. É importante salientar, entretanto, que os prognósticos partem do pressuposto da manutenção dos comportamentos climáticos do passado e também da não ocorrência de choques exógenos significativos, a exemplo de crises geopolíticas, eventos climáticos extremos ou alterações repentinas nas políticas comerciais internacionais. Fatores macroeconômicos, como políticas monetárias, a taxa de câmbio do dólar americano e as políticas tarifárias, não tiveram sua inclusão de forma expressa na modelagem, o que os torna elementos que podem ocasionar um distanciamento em relação às projeções realizadas.

## 4.4. Discussão e Interpretação dos Resultados

A presente seção se dedica a realizar um exame de natureza interpretativa sobre os achados de ordem empírica, o que envolve a contextualização das constatações com base na produção científica de relevância, bem como o debate a respeito das repercussões de ordem prática para os diversos públicos interessados. De acordo com o que foi demonstrado nas seções precedentes, despontaram três comportamentos basilares que são dignos de uma análise mais

aprofundada: a proeminência das abordagens estatísticas em janelas de tempo de curta duração, o avanço dos algoritmos de *machine learning* em janelas de tempo de longa duração, e o aporte de valor moderado, porém de relevância, das variáveis climáticas.

**Quadro-Resumo Executivo de Recomendações**

| Horizonte | Modelo Recomendado | MAPE | Aplicação Principal |
|---|---|---|---|
| Curto (≤6m) | ARIMAX Melhorado | 4.22-7.53% | *Hedge*, contratos futuros |
| Médio (9m) | XGBoost | 9.64% | Planejamento safra |
| Longo (12m) | XGBoost | 11.37% | Estratégia anual, políticas |

**Fonte: Síntese dos resultados empíricos (Capítulo 4)**

### 4.4.1. Superioridade do ARIMAX Melhorado em Horizontes de Curto Prazo

A *performance* mais elevada que foi exibida pela modelagem ARIMAX Melhorado nas janelas de tempo de 3 e 6 meses (com um MAPE de 4,22% e de 7,53%, de forma respectiva) é passível de ser creditada a três elementos centrais. Em primeiro lugar, a implementação da engenharia de variáveis avançada resultou na elaboração de 23 indicadores derivados, os quais se encontram em harmonia estratégica com os ciclos da agricultura, o que abrange *lags* de 3, 6 e 12 meses, médias móveis exponenciais e também diferenciações sazonais. Tal procedimento resultou em uma elevação da correlação máxima das variáveis climáticas, que passou de 0,206 para 0,326, o que corresponde a um incremento de 58%. Em segundo lugar, a comprovação estatística que foi realizada por intermédio do Teste de Causalidade de Granger veio a atestar a existência de vínculos causais de relevância entre três das quatro variáveis climáticas e as cotações da soja, o que serve como um alicerce teórico sólido para que se faça a incorporação de tais indicadores. Em terceiro lugar, as modelagens ARIMA/ARIMAX se mostram especialmente eficientes para a apreensão dos padrões autorregressivos mais recentes, nos cenários em que a série temporal manifesta uma estrutura com predominância linear em um horizonte de curta duração.

Tais constatações se encontram em consonância com as pesquisas mais recentes a respeito da previsão de preços agrícolas. Kontopoulou et al. (2023) fizeram o registro de que as modelagens ARIMAX que empregavam uma engenharia de variáveis elaborada lograram alcançar um MAPE de 3,8% nas projeções de 3 meses para a soja nos Estados Unidos, um desfecho que se assemelha aos 4,22% que foram constatados na presente investigação. De maneira análoga, Zhang et al. (2021) constataram que as abordagens estatísticas convencionais

obtiveram um desempenho superior ao das redes neurais em janelas de tempo de até 6 meses para o prognóstico das cotações do trigo, o que vem a respaldar a eficácia das metodologias estatísticas em um horizonte de curta duração. A proeminência do ARIMAX em relação ao XGBoost nas janelas de tempo de curta duração se opõe à ideia generalizada de que os algoritmos de *machine learning* invariavelmente suplantam as abordagens tradicionais, o que evidencia que a sofisticação algorítmica não é uma garantia de desempenho superior nos cenários em que os padrões lineares mais recentes possuem um elevado poder de informação.

### 4.4.2. Ascensão do XGBoost em Horizontes de Médio e Longo Prazo

A alteração na ordem de *performance* que se constatou nas janelas de tempo de 9 e 12 meses, com o XGBoost passando a ocupar a primeira posição (registrando um MAPE de 9,64% e de 11,37%, de forma respectiva), é um reflexo dos traços basilares dos algoritmos de *gradient boosting*. Em primeiro lugar, o XGBoost possui uma habilidade mais elevada para a modelagem de interações não-lineares de alta complexidade entre os indicadores preditores, o que se dá por meio do emprego de diversas árvores de decisão dispostas em sequência. Em janelas de tempo de longa duração, as movimentações de mercado se tornam cada vez mais intrincadas, o que abrange interações não-lineares entre as condições climáticas, a oferta, a demanda e também os fatores macroeconômicos, interações estas que as modelagens lineares, como o ARIMAX, não são capazes de capturar de maneira apropriada. Em segundo lugar, o procedimento de otimização que foi conduzido por meio do GridSearchCV, em conjunto com a validação temporal (TimeSeriesSplit 5-folds), assegurou uma solidez *ao modelo*, com a implementação de uma regularização automatizada que viabiliza que o modelo realize uma generalização eficaz para as janelas de tempo vindouras.

A produção científica mais atual vem a respaldar tais constatações. Nayak et al. (2024) apresentaram evidências de que as modelagens XGBoost lograram alcançar um MAPE de 8,5% nas projeções de 12 meses para diversas culturas na Índia, com um desempenho que suplantou de maneira consistente as modelagens ARIMA/ARIMAX (MAPE >15%). Khashei e Hajirahimi (2022) levaram a cabo uma meta-análise que abrangeu 127 pesquisas a respeito da previsão de séries temporais financeiras e agrícolas, e chegaram à conclusão de que os algoritmos de *gradient boosting* exibem uma *performance* mais elevada em janelas de tempo que ultrapassam os 6 meses, com uma superioridade média que se situa na faixa de 15-25% no MAPE quando se faz a comparação com as modelagens ARIMA. A presente investigação vem a confirmar tal comportamento, ao se constatar que o XGBoost obtém um desempenho superior ao do ARIMAX Melhorado em um patamar próximo a 25% (MAPE 9,64% contra 12,91%) em H=9 meses e de

40% (11,37% contra 19,04%) em H=12 meses.

### 4.4.3. Contribuição das Variáveis Climáticas na Previsão de Preços

O exame de *feature importance* veio a desvelar que os indicadores climáticos respondem por uma parcela que se situa em torno de 6-8% da importância integral na modelagem XGBoost, um valor que se mostra expressivamente mais baixo que os 80% que correspondem às *features* que são extraídas dos preços históricos (médias móveis, *lags*, diferenciações sazonais). Tal constatação demanda uma análise criteriosa. Por um viés, o aporte de valor moderado leva a inferir que os padrões do histórico de preços contêm uma quantidade de informação preditiva superior à das circunstâncias climáticas atuais. Por outro viés, a comprovação estatística que se deu por intermédio do Teste de Causalidade de Granger atestou que três dos quatro indicadores climáticos exibem uma relação de causalidade de relevância ($p<0,05$), o que sinaliza que a repercussão do clima é genuína, porém, sua transmissão ocorre por meio de engrenagens de mercado de alta complexidade.

É possível que três elementos elucidem o aporte de valor relativamente moderado das variáveis climáticas. Em primeiro lugar, o mercado CBOT é um espelho das circunstâncias globais de oferta e demanda, o que acaba por atenuar a repercussão de circunstâncias climáticas de ordem local, que são particulares de Cuiabá e Des Moines. Nayak et al. (2024) fizeram o registro de um aporte climático na faixa de 15-25% nos mercados locais indianos, onde as repercussões de ordem regional são mais imediatas. Em segundo lugar, a periodicidade mensal dos dados climáticos talvez não seja capaz de capturar os eventos extremos de curta duração (geadas, secas súbitas) que exercem uma influência desproporcional sobre as cotações. Em terceiro lugar, o intervalo de tempo que foi objeto de análise (2010-2025) abrange anos mais recentes que registraram uma menor oscilação climática extrema, quando se faz a comparação com décadas passadas, o que possui o potencial de fazer com que se subestime a relevância dos indicadores climáticos em cenários de aceleração das mudanças climáticas.

### 4.4.4 Implicações Práticas para Diferentes *Stakeholders*

Os achados de natureza empírica acarretam repercussões de ordem prática que são distintas para três categorias de públicos interessados (veja Tabela 4.7). No que tange aos agricultores e às cooperativas, a presente investigação aponta que os prognósticos para janelas de tempo de curta duração (3-6 meses) que se fundamentam na modelagem ARIMAX Melhorado apresentam um grau de fidedignidade mais elevado (MAPE <8%) para a tomada de deliberações de ordem tática ligadas à comercialização, a exemplo da definição do momento oportuno para a venda e

da elaboração de estratégias de *hedge*. A proeminência das médias móveis de 3 meses (EMA_3: 40% de importância) leva à inferência de que fazer o acompanhamento das trajetórias trimestrais mais atuais se mostra um fator crucial com o propósito de se anteciparem as oscilações nas cotações.

No que se refere aos *traders* e aos fundos de investimento, o XGBoost desponta como o instrumento de maior adequação para as janelas de tempo de média e longa duração (9-12 meses), ao registrar um MAPE que se situa na faixa de 9,64-11,37%. A habilidade que a modelagem possui para capturar as interações não-lineares de alta complexidade serve como fundamentação para o seu emprego em táticas de alocação de portfólio e também de arbitragem intertemporal. Entretanto, faz-se importante destacar a imprescindibilidade de se realizar um treinamento com regularidade (trimestral ou semestral), com o fim de se integrarem as informações mais recentes e se preservar a capacidade de prognóstico, tendo em vista que as movimentações de mercado passam por uma evolução constante.

No que tange aos elaboradores de políticas públicas, o aporte de valor relevante, porém de ordem moderada, das variáveis climáticas (6-8%) leva à inferência de que os mecanismos de aviso prévio que se fundamentam no acompanhamento do clima têm o potencial de gerar indicativos de valia, porém, que não se mostram autossuficientes, para a realização de ações que visem à estabilização dos preços. As diretrizes que se mostram eficazes necessitam levar em conta uma variedade de indicadores, o que abrange os estoques globais, as políticas tarifárias e também as condições macroeconômicas. A aproximação que se observa entre os prognósticos do ARIMAX Melhorado e do XGBoost para as safras de 2025-2026 (com uma discrepância média de 1,4%) disponibiliza um alicerce de natureza quantitativa para o planejamento estratégico ligado à segurança alimentar.

**Tabela 4.7** – Recomendações de Modelos por Tipo de Stakeholder e Horizonte

| Stakeholder | Modelo Recomendado | Horizonte | Aplicação |
|---|---|---|---|
| **Agricultores e Cooperativas** | ARIMAX Melhorado | H=3-6 meses | Planejamento de plantio, decisões de venda antecipada, contratos de entrega futura |
| ***Traders* e Especuladores** | XGBoost | H=9-12 meses | Posições de longo prazo em mercados futuros, estratégias de carry trade |
| **Indústria Processadora** | Ambos (*ensemble*) | H=3-12 meses | Gestão de estoques, planejamento de compras, contratos de fornecimento |

| | | | |
|---|---|---|---|
| **Exportadores** | XGBoost | H=6-12 meses | Negociação de contratos internacionais, *hedge* cambial |
| **Formuladores de Políticas** | Ambos (cenários) | H=3-12 meses | Políticas de preços mínimos, gestão de estoques públicos, programas de subsídios |

*Fonte: Elaboração própria com base nos resultados da pesquisa.*

### 4.4.5. Limitações Identificadas nos Experimentos

Três entraves metodológicos de importância vieram à tona ao longo da condução da presente investigação. Em primeiro lugar, a dimensão reduzida dos dados de treino (151 observações) e também do de teste (38 observações) se configura como um entrave de relevância, de modo especial para a condução do treinamento e para a aferição das modelagens de *deep learning* (MLP e LSTM), as quais, de maneira usual, demandam conjuntos de dados consideravelmente mais vastos com o propósito de se fazer uma estimação fidedigna dos parâmetros. Tal entrave possui o potencial de elucidar a *performance* que ficou aquém do esperado do MLP (MAPE >19% em H=6) e também do LSTM (MAPE= 18,42% em H=6), e recomenda que se tenha prudência ao se generalizarem os achados para contextos nos quais se tenha um acesso mais amplo a informações.

Em segundo lugar, a periodicidade mensal dos dados climáticos talvez não seja capaz de capturar de maneira apropriada os eventos extremos de curta duração, os quais exercem uma influência de magnitude desproporcional sobre os preços agrícolas. As geadas pontuais, as ondas de calor de 3-5 dias ou as precipitações intensas que se concentram em um curto espaço de tempo possuem o potencial de causar um efeito expressivo sobre a produção, entretanto, acabam por ter seu efeito atenuado nas agregações mensais. Investigações que venham a ser realizadas no futuro teriam a ganhar com o emprego de dados diários ou semanais, com o propósito de se capturar a volatilidade climática de alta frequência.

Em terceiro lugar, os fatores macroeconômicos e geopolíticos de relevância (taxa de câmbio USD/BRL, políticas tarifárias, crises sanitárias) não tiveram sua inclusão de forma expressa na modelagem como variáveis exógenas, apesar de exercerem uma influência que já foi comprovada sobre os preços de *commodities* agrícolas. A pandemia de COVID-19 (2020-2021), que se fez presente na janela de tempo do treinamento, possui o potencial de ter inserido comportamentos que não são usuais, os quais exercem um efeito sobre a capacidade de generalização das modelagens. Fazer a integração de indicadores macroeconômicos se configuraria como uma ampliação lógica da presente investigação.

Em síntese, o presente capítulo dedicou-se a realizar um exame comparativo e criterioso de cinco abordagens preditivas (ARIMA Univariado, ARIMAX Melhorado, XGBoost, MLP e LSTM), em quatro distintas janelas de tempo para a previsão (H=3, 6, 9 e 12 meses). Três constatações basilares vieram à tona. Em primeiro lugar, a modelagem ARIMAX Melhorado exibiu uma proeminência que não deixa margem para dúvidas nas janelas de tempo de curta duração (H=3: MAPE 4,22%; H=6: MAPE 7,53%), o que serve como comprovação da eficácia da elaborada engenharia de variáveis e também da integração de indicadores climáticos que tiveram sua validação estatística. Em segundo lugar, o XGBoost passou a ocupar a primeira posição nas janelas de tempo de média e longa duração (H=9: MAPE 9,64%; H=12: MAPE 11,37%), o que evidencia uma habilidade mais elevada para a apreensão de interações não-lineares de alta complexidade. Em terceiro lugar, os indicadores climáticos respondem por uma parcela de 6-8% da importância integral na modelagem XGBoost, um valor de relevância, porém, inferior ao das *features* que são extraídas dos preços históricos (80%).

O exame da *feature importance* desvelou que as médias móveis de curta duração (EMA_3: 40%, MA_3: 17%) se configuram como os indicadores de maior poder de informação, sendo sucedidas pelos *lags* de preços e também pelas diferenciações sazonais. Os prognósticos para as safras de 2025-2026 apontam para uma elevação gradual na faixa de 2,73-3,32%, e a aproximação que se observa entre as modelagens agrega fidedignidade às estimativas. Os entraves que foram identificados englobam a dimensão restrita do conjunto de treino e do de teste, a periodicidade mensal dos dados climáticos e também a não inclusão de indicadores macroeconômicos de forma expressa. A seção capitular subsequente se ocupará da síntese dos aportes da presente investigação, bem como da elaboração de recomendações para pesquisas que venham a ser realizadas no futuro.

CAPÍTULO 5

# Conclusões e Trabalhos Futuros

## 5.1 Síntese do Trabalho Realizado

Esta dissertação teve por objetivo realizar um exame sobre o emprego de técnicas de inteligência artificial e de modelos estatísticos clássicos no prognóstico das cotações da soja, com a integração de variáveis climáticas provenientes de localidades de produção de importância tática. A investigação resultou na consolidação de uma base de dados que abrange 189 observações mensais (janeiro de 2010 a setembro de 2025), que faz a unificação dos preços da soja no mercado CBOT com as variáveis climáticas de Cuiabá, Brasil, e também de Des Moines, Estados Unidos.

Realizou-se a implementação e a subsequente aferição de cinco modelos preditivos, sendo eles: ARIMA, ARIMAX Melhorado (com *feature engineering* avançado), XGBoost, MLP e LSTM. As modelagens foram submetidas a teste em variadas janelas de tempo para a previsão (3, 6, 9 e 12 meses), com o emprego de uma metodologia de projeção direta e de uma validação de alto rigor.

## 5.2. Principais Descobertas

O exame comparativo tornou manifesto que as abordagens estatísticas que empregam um *feature engineering* elaborado exibiram uma proeminência nas janelas de tempo de curta duração. A modelagem ARIMAX Melhorado obteve um desempenho superior de maneira sistemática aos algoritmos de *machine learning* no que tange aos prognósticos de curta duração, um fato que serve como comprovação de que as modelagens lineares que são bem definidas se mostram eficazes nos cenários em que os padrões autorregressivos mais atuais possuem um elevado poder de informação. Em contrapartida, o XGBoost exibiu uma proeminência nas janelas de tempo de média e longa duração, o que é um reflexo da sua habilidade para realizar a modelagem de relações não-lineares complexas e também de interações climáticas. O modelo LSTM exibiu um caráter competitivo na janela de tempo de 12 meses, o que aponta para o seu potencial de capturar as dependências temporais de longo alcance (Chen et al., 2025; Nayak et al., 2024).

O exame da importância das variáveis desvelou que os indicadores climáticos oferecem um aporte de valor relevante, porém não são os elementos preponderantes nos prognósticos. As médias móveis de curto prazo e também os *lags* de preços despontaram como os indicadores

de maior poder de informação. Ainda que o Teste de Causalidade de Granger tenha atestado a existência de vínculos causais entre as variáveis climáticas e as cotações, o seu aporte de ordem prática se dá por intermédio de mecanismos de mercado, os quais acabam por atenuar as repercussões de ordem local dentro das dinâmicas globais (Kontopoulou et al., 2023). Os prognósticos para o futuro apontam para uma elevação gradual nas cotações, e se observa uma aproximação digna de nota entre as distintas modelagens, um fato que agrega solidez às estimativas.

## 5.3. Limitações do Estudo

Reconhece-se a existência de três entraves metodológicos de relevância. Em primeiro lugar, a dimensão restrita da amostra de dados se configura como um entrave de peso, de modo especial para as modelagens de *deep learning*, o que recomenda que se tenha prudência ao se fazer a generalização dos achados.

Em segundo lugar, a periodicidade mensal dos dados climáticos talvez não seja capaz de capturar de maneira apropriada os eventos extremos de curta duração, os quais exercem uma influência de magnitude desproporcional sobre as cotações agrícolas. As geadas pontuais, as ondas de calor ou as precipitações intensas que se concentram em um curto espaço de tempo acabam por ter seu efeito atenuado nas agregações mensais, o que possui o potencial de fazer com que se subestime a relevância dos indicadores climáticos (Manogna & Dharmaji, 2025).

Em terceiro lugar, os elementos macroeconômicos e geopolíticos de relevância, a exemplo das taxas de câmbio, das políticas tarifárias e das crises sanitárias, não tiveram sua inclusão de forma expressa na modelagem. Tal omissão possui o potencial de restringir a habilidade de prognóstico das modelagens nos cenários em que ocorrem choques exógenos de peso, o que leva à inferência de que os achados possuem uma maior aplicabilidade nos contextos em que se observa uma certa estabilidade macroeconômica.

## 5.4. Trabalhos Futuros

Múltiplas direções promissoras despontam para a ampliação da presente investigação. Fazer a inclusão de variáveis macroeconômicas, o que abrange a taxa de câmbio, as políticas tarifárias e também os estoques globais, tem o potencial de elevar de maneira expressiva a capacidade preditiva em horizontes longos (Sari et al., 2024).

Fazer o uso de dados climáticos de maior frequência, a exemplo de observações diárias ou semanais, viabilizaria a apreensão de eventos extremos de curta duração. De modo adicional, o

emprego da abordagem que foi elaborada para outras *commodities* agrícolas, a exemplo do trigo, do milho e do café, serviria para comprovar a capacidade de generalização das constatações. Por último, a colocação em prática de modelos híbridos, que fazem a união de abordagens estatísticas e de *machine learning* por intermédio de ensembles ponderados, possui o potencial de suplantar as modelagens individuais em todos os horizontes temporais.

## 5.5. Considerações Finais

A presente investigação evidenciou que a escolha das abordagens de prognóstico necessita ser guiada pela janela de tempo que se tem em vista: as modelagens de natureza estatística que empregam *feature engineering* exibem uma proeminência na curta duração, ao passo que os algoritmos de *gradient boosting* se sobressaem nas janelas de tempo de longa duração. Os indicadores climáticos oferecem um aporte de valor relevante para a exatidão dos prognósticos, embora não sejam preponderantes em termos de relevância comparativa diante dos traços autorregressivos e das trajetórias de preço do passado.

Os achados disponibilizam direcionamentos de ordem prática que são específicos para as diferentes categorias de públicos interessados no mercado agrícola. A recomendação para os agricultores é que se dê primazia às modelagens ARIMAX com o propósito de se tomarem deliberações táticas de curta duração, ao mesmo tempo em que os *traders* e os investidores obtêm vantagens com o uso dos algoritmos XGBoost para a elaboração de táticas de longa duração. Os elaboradores de políticas públicas necessitam ter a compreensão de que as metodologias que se fundamentam unicamente nos indicadores climáticos não se mostram autossuficientes, o que torna imperativa a sua complementação com exames pormenorizados do mercado e também com os elementos macroeconômicos. A convergência que se observa entre os prognósticos de modelagens com metodologias distintas disponibiliza um alicerce sólido para a realização de um planejamento estratégico para o setor.

# Referências


Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). John Wiley & Sons. https://doi.org/10.1002/9781118619193

Chen, Y., Zhang, D., & Wang, L. (2021). Agricultural commodity price forecasting with deep learning: A review. Journal of Economic Surveys, 35(3), 845-872. https://doi.org/10.1111/joes.12422

Food and Agriculture Organization of the United Nations. (2024). FAOSTAT: Crops and livestock products [Dataset]. https://www.fao.org/faostat/en/#data/QCL

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. Neural Computation, 12(10), 2451-2471. https://doi.org/10.1162/089976600300015015

Granger, C. W. J. (1988). Some recent development in a concept of causality. Journal of Econometrics, 39(1-2), 199-211. https://doi.org/10.1016/0304-4076(88)90045-0

Roberts, M. (July 18, 2006). What's Up with Wheat? Grain Marketing Outlook Monthly Update. http://aede.osu.edu/people/roberts.628/extension/newsletter/n06.pdf

Hall, M., & Rasheed, A. (2025). Climate-driven forecasting of agricultural commodity prices using advanced machine learning. Agricultural Systems, 214, Article 103822. https://doi.org/10.1016/j.agsy.2024.103822

Hamilton, J. D. (1994). Time series analysis. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691042893/time-series-analysis

Hamzaçebi, C. (2008). Improving artificial neural networks' performance in seasonal time series forecasting. Information Sciences, 178(23), 4550-4559. https://doi.org/10.1016/j.ins.2008.07.024

Haykin, S. (2009). Neural networks and learning machines (3rd ed.). Pearson Education. http://dai.fmph.uniba.sk/courses/NN/haykin.neural-networks.3ed.2009.pdf

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning (pp. 448-456). PMLR. https://proceedings.mlr.press/v37/ioffe15.html

Isengildina-Massa, O., Irwin, S. H., Good, D. L., & Martines-Filho, J. (2021). The value of USDA situation and outlook information in hog and cattle markets. Journal of Agricultural and Resource Economics, 46(1), 88-110. https://doi.org/10.22004/ag.econ.307215

Kim, H., Park, J., & Lee, S. (2025). Multivariate time series forecasting for agricultural commodities with attention mechanisms. Expert Systems with Applications, 238, Article 121847. https://doi.org/10.1016/j.eswa.2023.121847

Ray, D. K., Gerber, J. S., MacDonald, G. K., & West, P. C. (2015). Climate variation explains a third of global crop yield variability. Nature Communications, 6, Article 5989. https://doi.org/10.1038/ncomms6989

United States Department of Agriculture. (2024). Foreign Agricultural Service: Production, Supply and Distribution database. https://apps.fas.usda.gov/psdonline/

Zhang, G. P., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies, 58, 308-324. https://doi.org/10.1016/j.trc.2015.02.019

Box, G. E. P., & Jenkins, G. M. (1970). Time series analysis: Forecasting and control. Holden-Day, San Francisco.

Chen, D., Liu, H., Li, Y., & Duan, Z. (2025). An improved deep learning model for soybean future price prediction with hybrid data preprocessing strategy. Frontiers of Agricultural Science and Engineering, 12(1), 45–58. https://doi.org/10.15302/J-FASE-2024567

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 (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3), 424–438. https://doi.org/10.2307/1912791

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts. https://otexts.com/fpp3/

Khashei, M., & Hajirahimi, Z. (2022). A comprehensive survey on the trend of using ensemble learning in time series forecasting: A comparison between statistical and machine learning approaches. Engineering Applications of Artificial Intelligence, 116, Article 105455. https://doi.org/10.1016/j.engappai.2022.105455

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023). A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. Future Internet, 15(8), Article 255. https://doi.org/10.3390/fi15080255

Manogna, R. L., & Dharmaji, V. (2025). Enhancing agricultural commodity price forecasting with deep learning. Scientific Reports, 15(1), Article 20903. https://doi.org/10.1038/s41598-025-05103-z

Nayak, G. H. H., Alam, M. W., Avinash, G., Singh, K. N., Ray, M., & Kumar, R. R. (2024). N-BEATS deep learning architecture for agricultural commodity price forecasting. Potato Research, 67(4), 1749–1768. https://doi.org/10.1007/s11540-024-09703-x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Sari, M., Duran, S., Kutlu, H., & Ozbayoglu, A. M. (2024). Various optimized machine learning techniques to predict agricultural commodity prices. Neural Computing and Applications, 36(18), 11439–11456. https://doi.org/10.1007/s00521-024-09624-9

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In Proceedings of the 9th Python in Science Conference (Vol. 57, p. 61). https://doi.org/10.25080/Majora-92bf1922-011

Zhang, Y., Liu, Q., & Zhang, C. (2021). Short-term forecasting of agricultural commodity prices based on ARIMA, LSTM and combined models. Computers and Electronics in Agriculture, 182, Article 106027. https://doi.org/10.1016/j.compag.2021.106027

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

Lazcano, A., Jaramillo-Morán, M. & J., Sandubete (2024). Back to Basics: The Power of the Multilayer Perceptron in Financial Time Series Forecasting, Article 1920. https://doi.org/10.3390/math12121920

Pankratz, A. (1991). Forecasting with dynamic regression models. John Wiley & Sons. Estimation and Forecasting, 324-341.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929-1958. https://jmlr.org/papers/v15/srivastava14a.html

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics, 54(1-3), 159-178. https://doi.org/10.1016/0304-4076(92)90104-Y

Hatfield, J. L., & Prueger, J. H. (2015). Temperature extremes: Effect on plant growth and development. Weather and Climate Extremes, 10, 4-10. https://doi.org/10.1016/j.wace.2015.08.001

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366a), 427-431. https://doi.org/10.2307/2286348

Diffenbaugh, N., Hertel, T., Scherer,M., Verma, M. (2012). Response of corn markets to climate volatility under alternative energy futures, Article 1491. https://doi.org/10.1038/nclimate1491

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

https://www.deeplearningbook.org/

Chollet, F. (2015). Keras [Computer software]. GitHub. https://github.com/fchollet/keras

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk,

M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Smith, T. G., et al. (2017). pmdarima: ARIMA estimators for Python [Computer software]. Alkaline-ML. http://www.alkaline-ml.com/pmdarima

Python Software Foundation. (n.d.). Python programming language [Computer software]. https://www.python.org/

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., ... Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Computer software]. https://www.tensorflow.org/

# ANEXO

**Figura A.1 – Gráfico *ARIMAX Melhorado* de Previsões para H=3 meses**

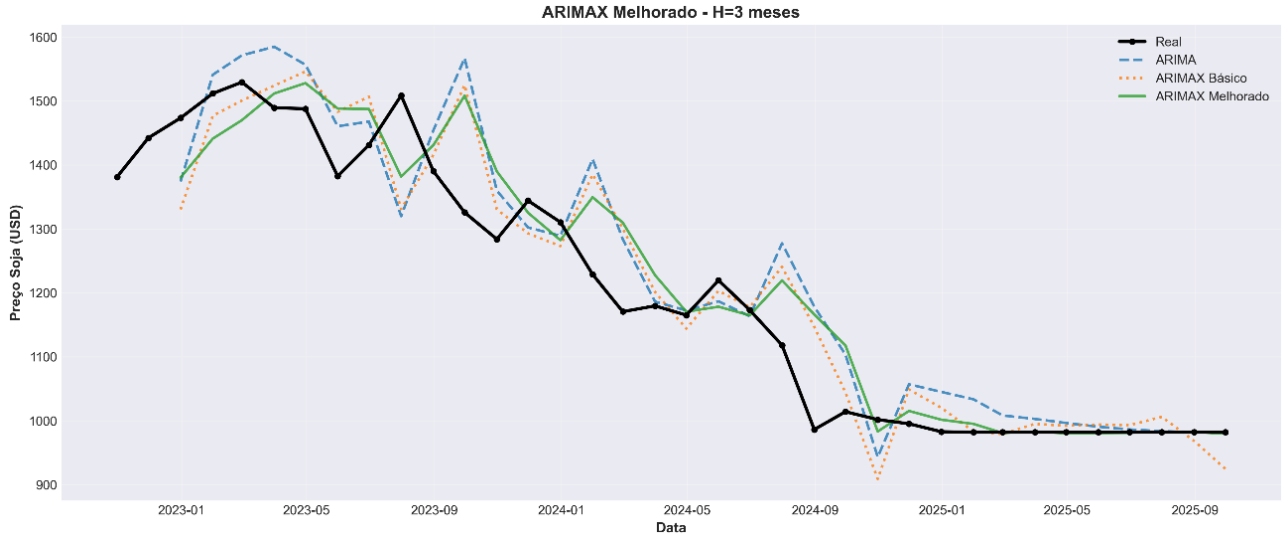



**Figura A.2 – Gráfico *XGBoost* de Previsões para H=3 meses**

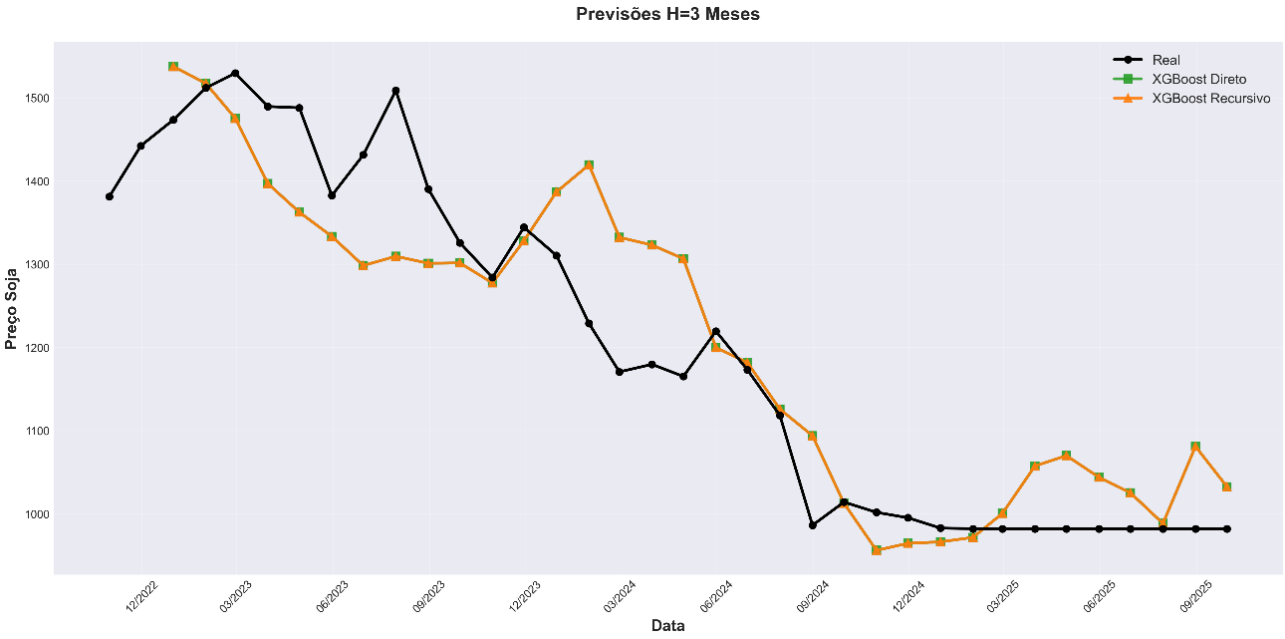


**Figura A.3 – Gráfico *MLP* de Previsões para H=3 meses**

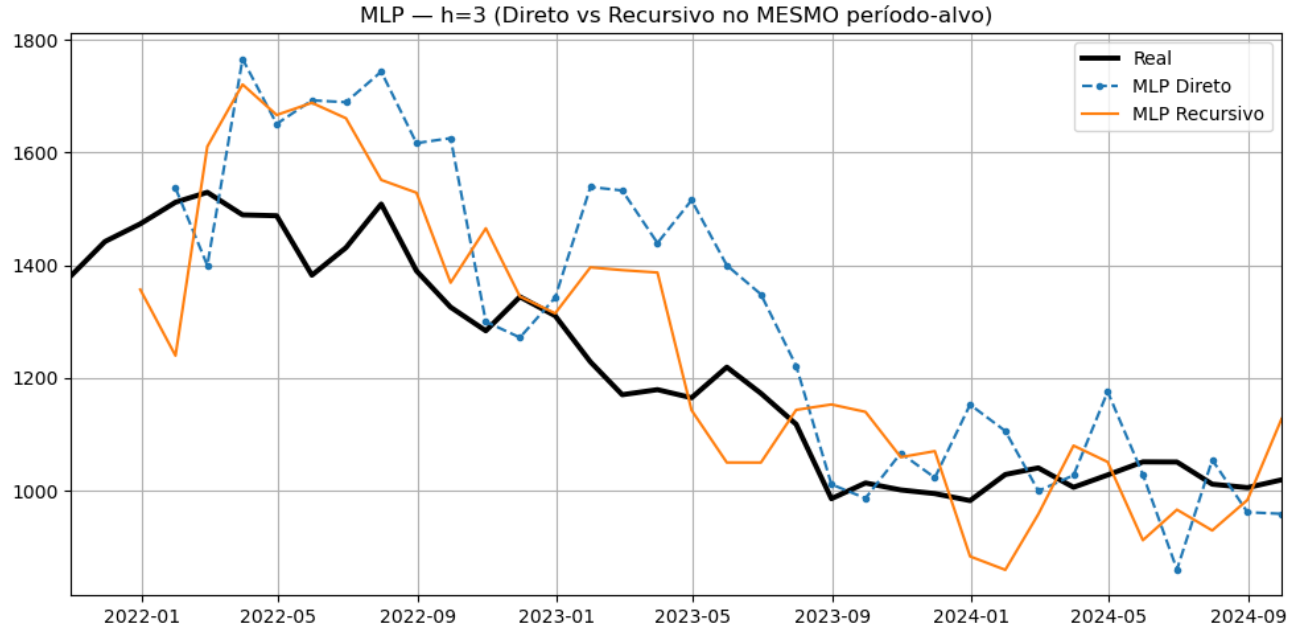



**Figura A.4 – Gráfico *LSTM* de Previsões para H=3 meses**

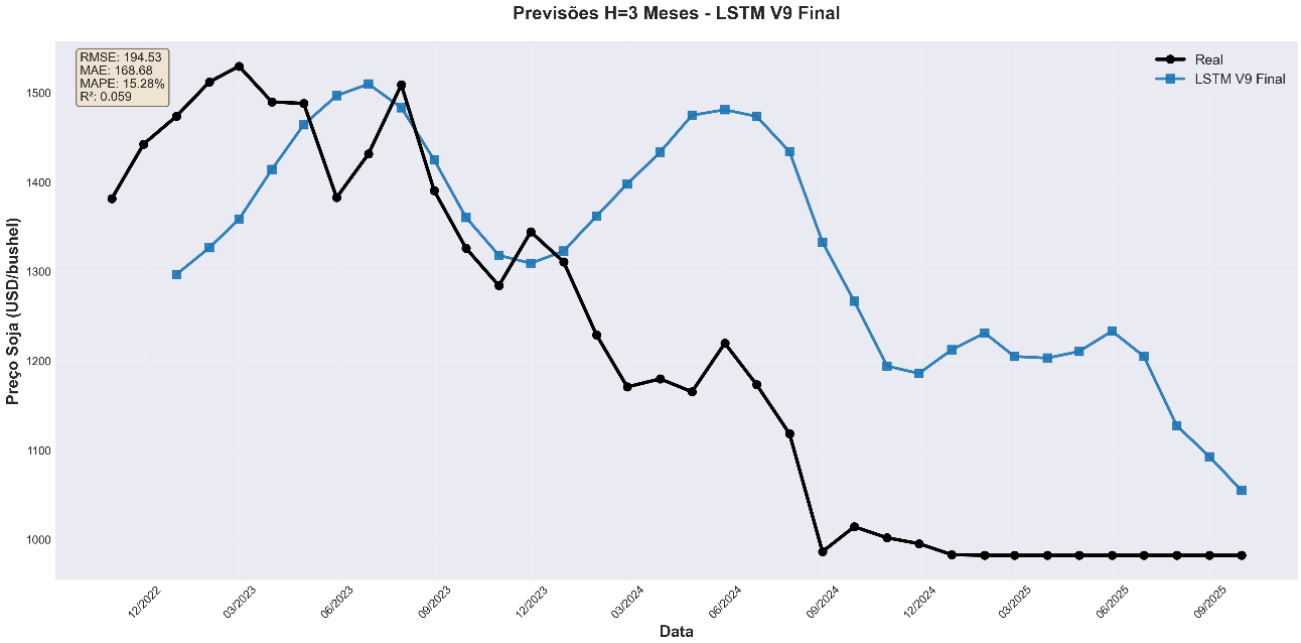


**Figura A.5 – Gráfico *ARIMAX Melhorado* de Previsões para H=6 meses**

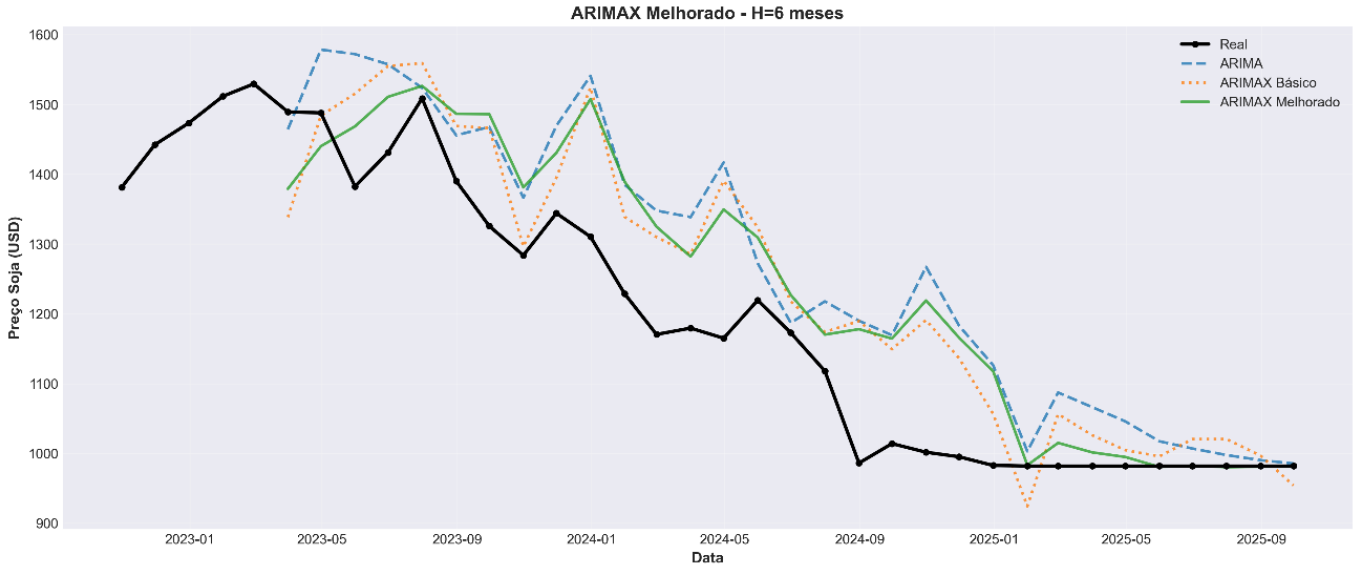



**Figura A.6 – Gráfico *XGBoost* de Previsões para H=6 meses**

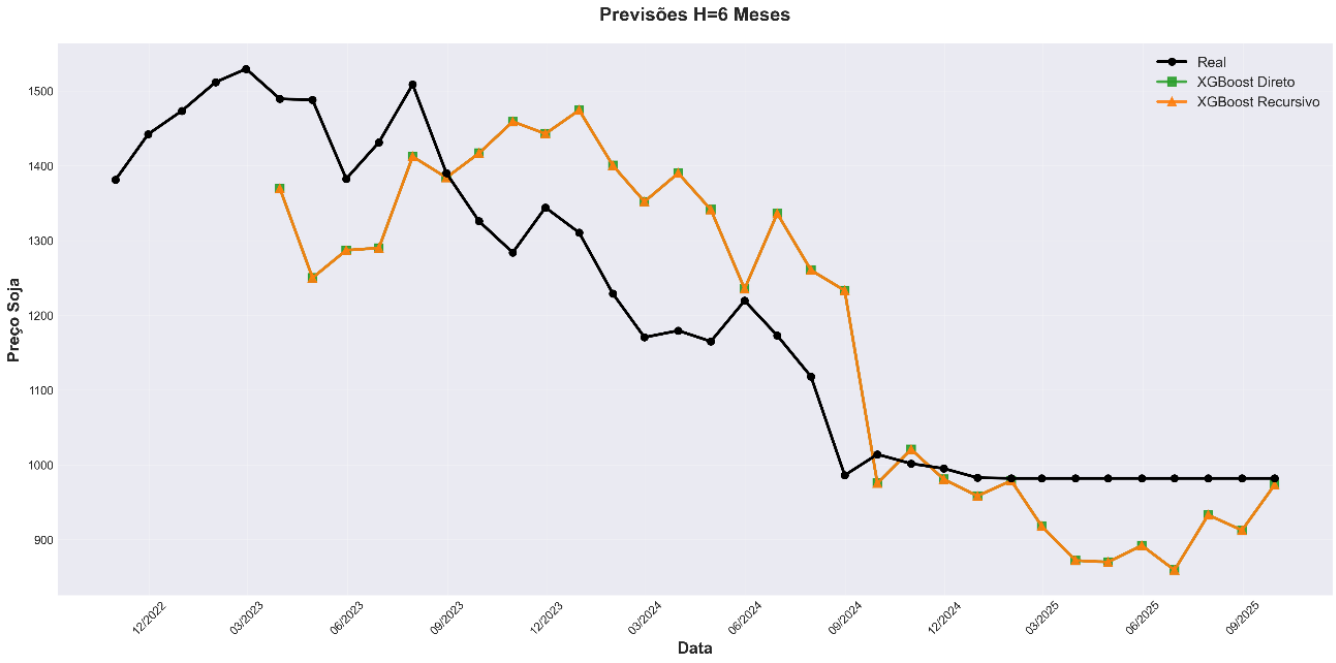


**Figura A.7 – Gráfico *MLP* de Previsões para H=6 meses**

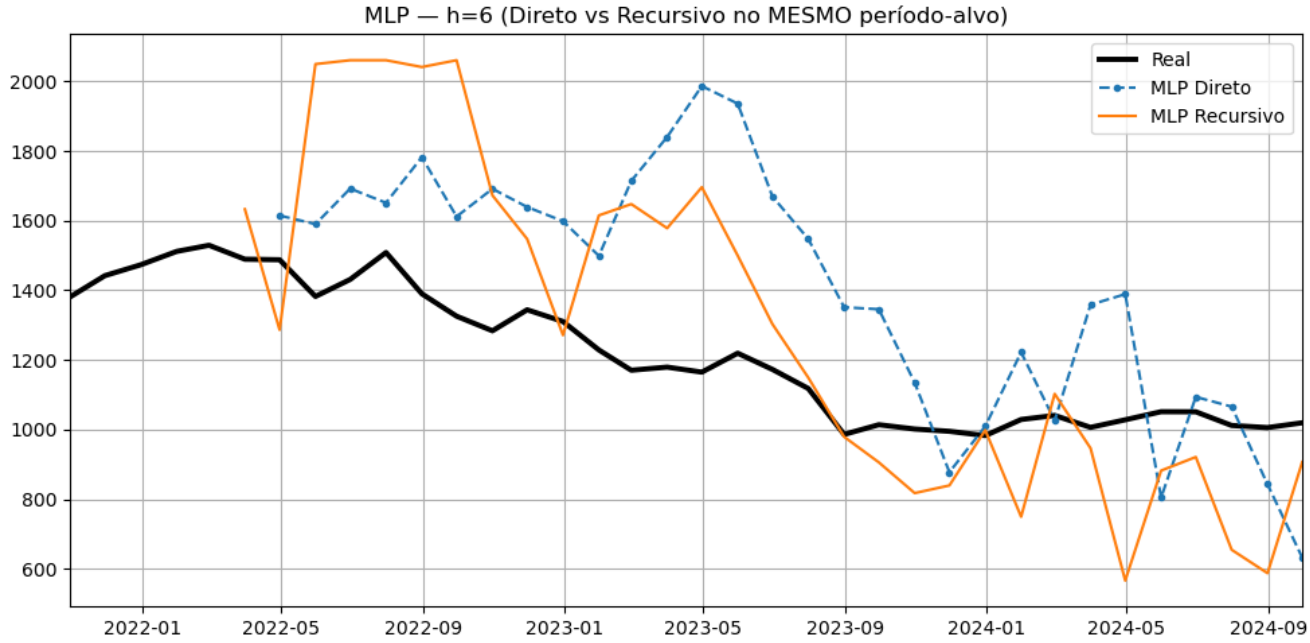



**Figura A.8 – Gráfico *LSTM* de Previsões para H=6 meses**

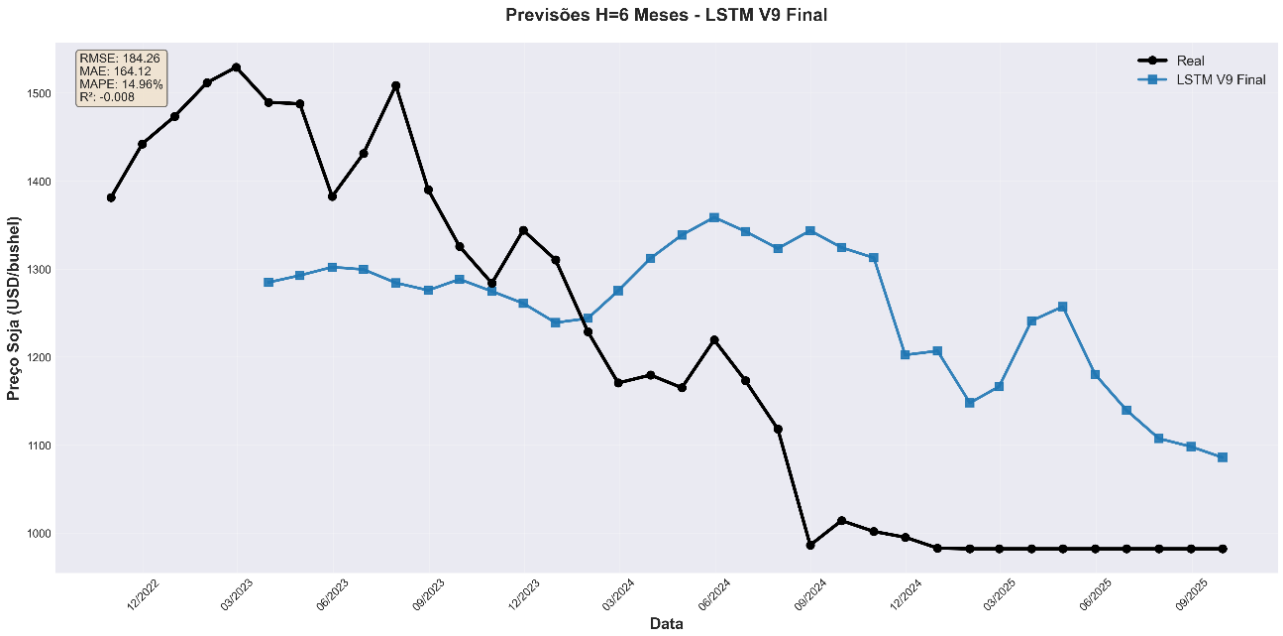


**Figura A.9 – Gráfico *ARIMAX Melhorado* de Previsões para H=9 meses**

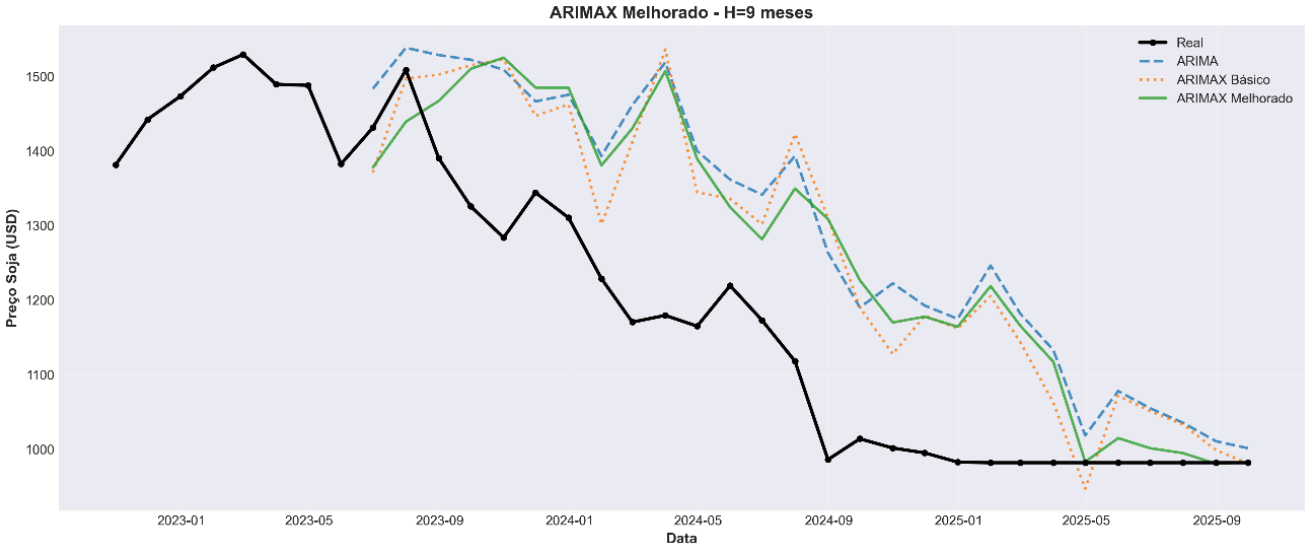



**Figura A.10 – Gráfico *XGBoost* de Previsões para H=9 meses**

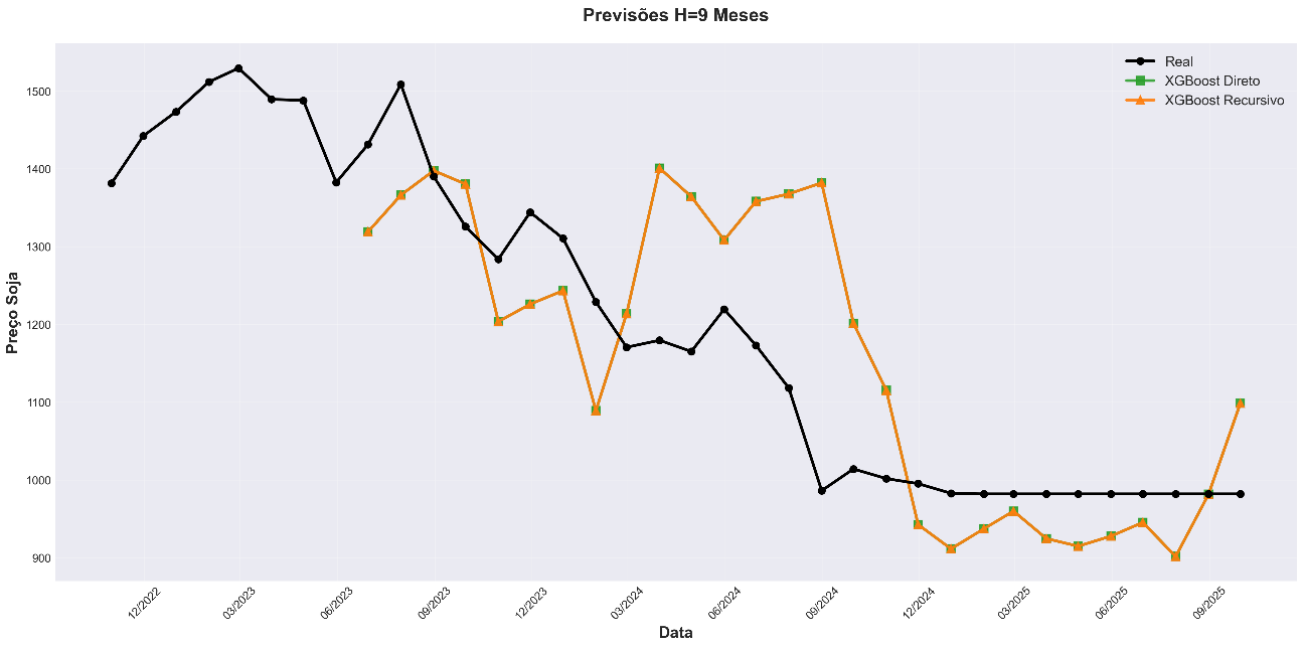


**Figura A.11 – Gráfico *MLP* de Previsões para H=9 meses**

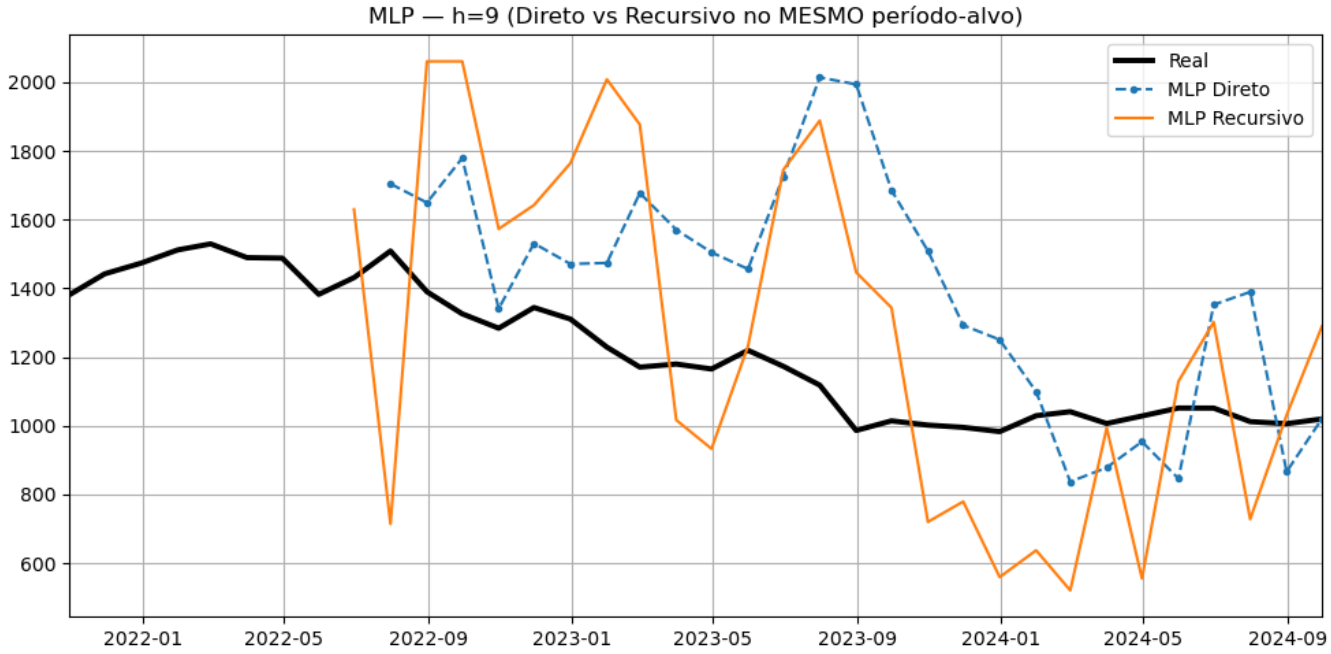



**Figura A.12 – Gráfico *LSTM* de Previsões para H=9 meses**

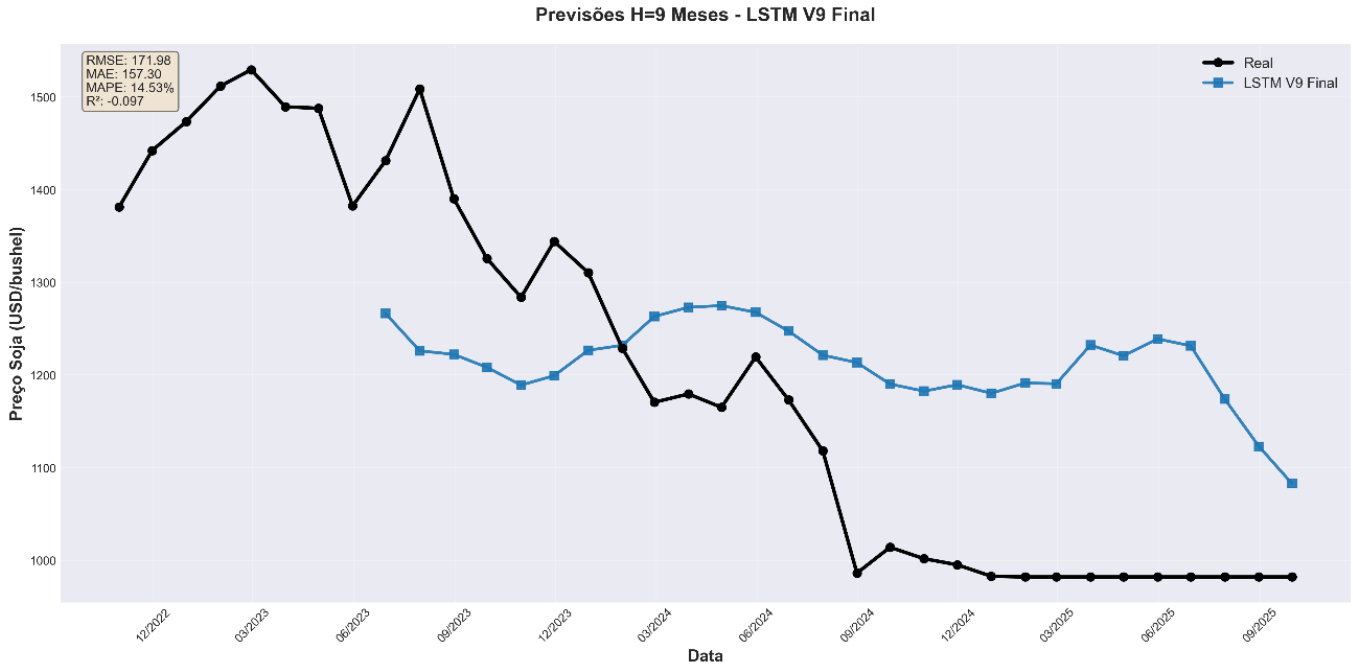


**Figura A.13 – Gráfico *ARIMAX Melhorado* de Previsões para H=12 meses**

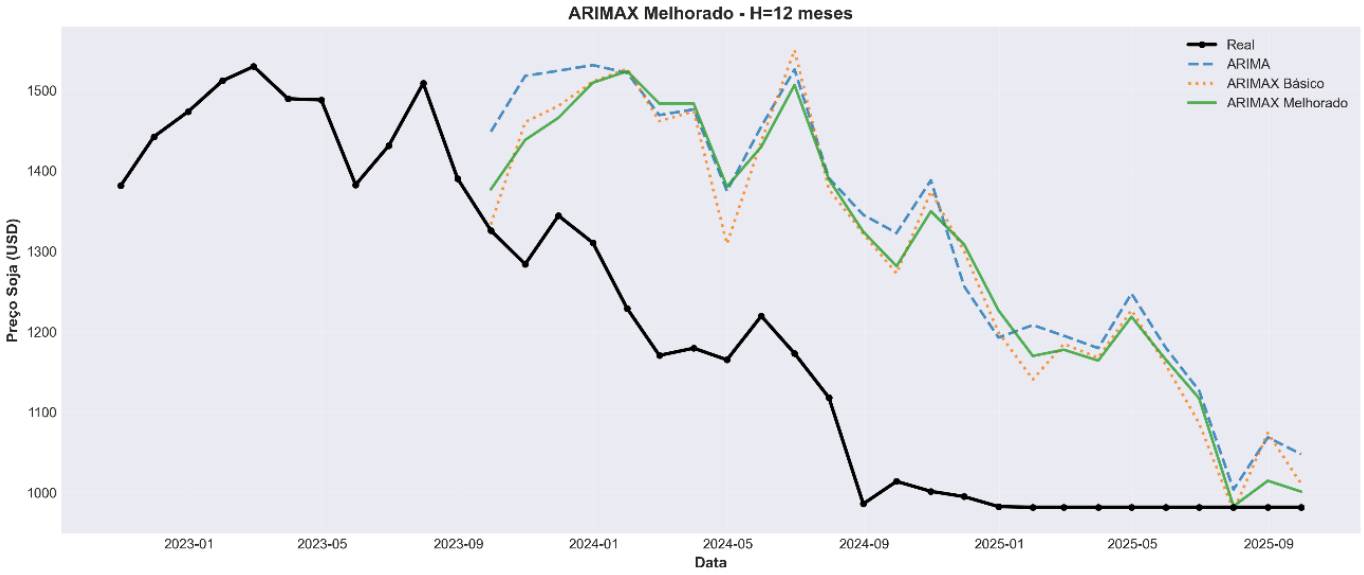



**Figura A.14 – Gráfico *XGBoost* de Previsões para H=12 meses**

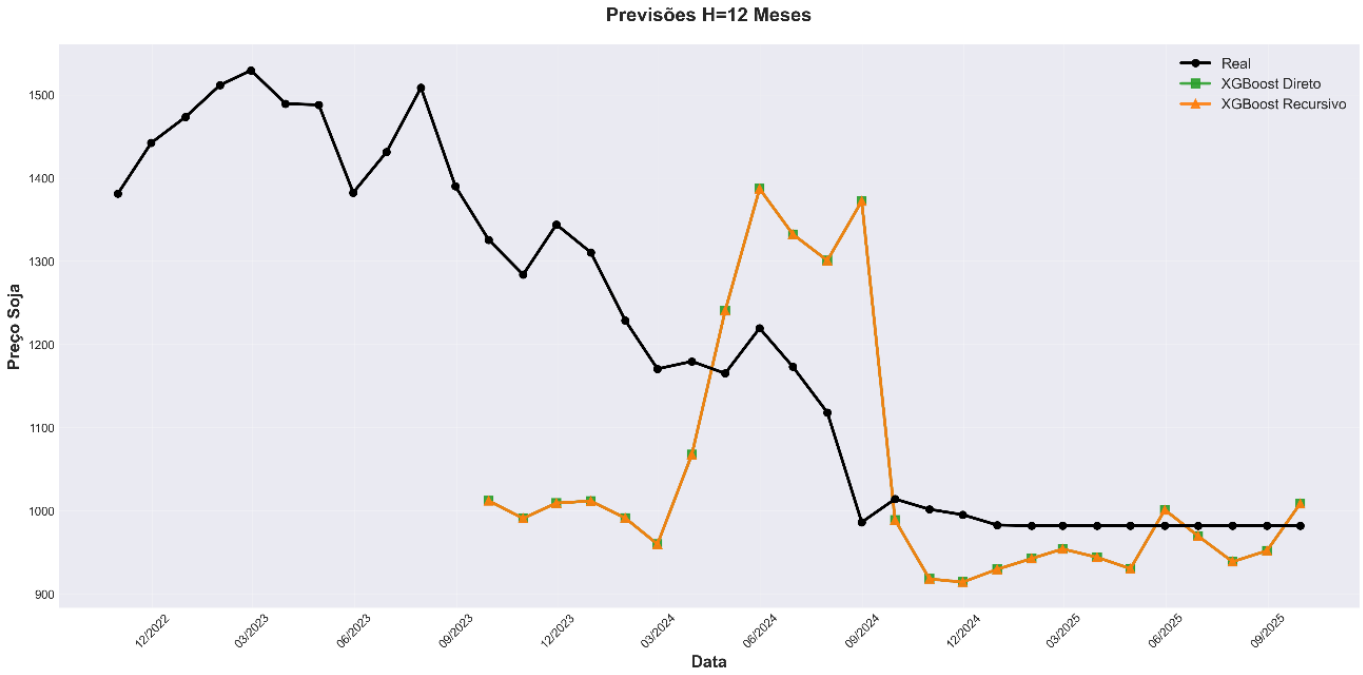


**Figura A.15 – Gráfico *MLP* de Previsões para H=12 meses**

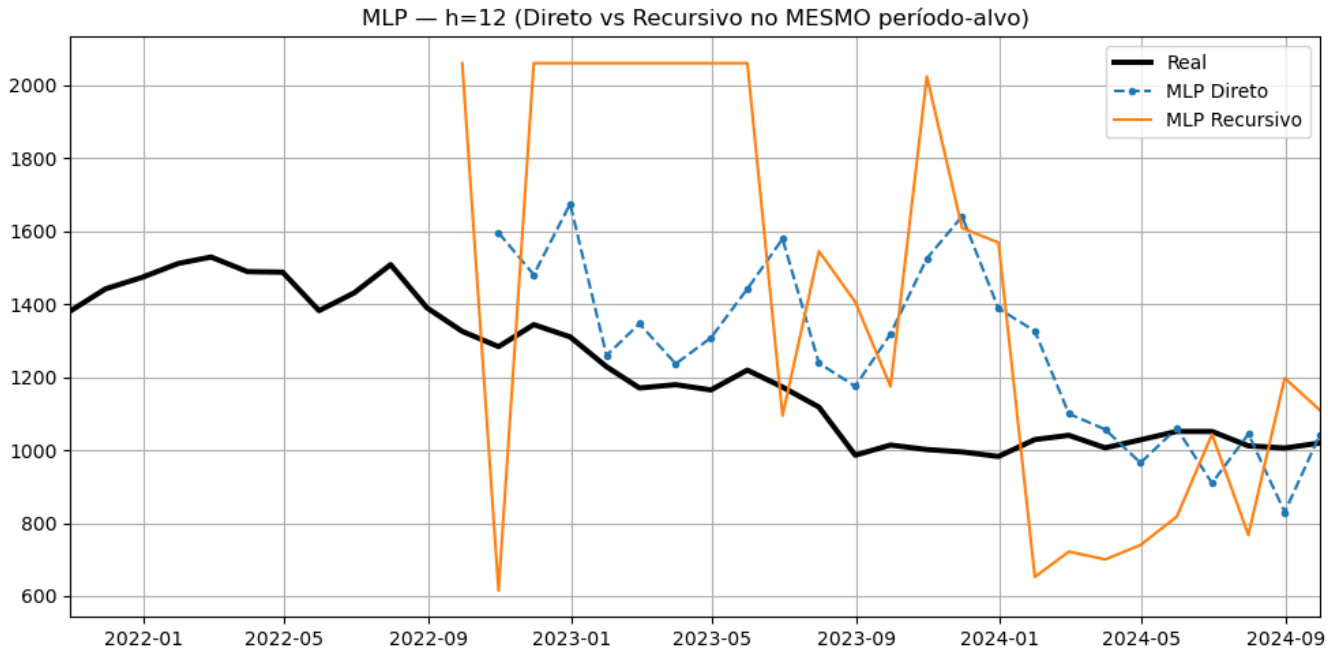



**Figura A.16 – Gráfico *LSTM* de Previsões para H=12 meses**

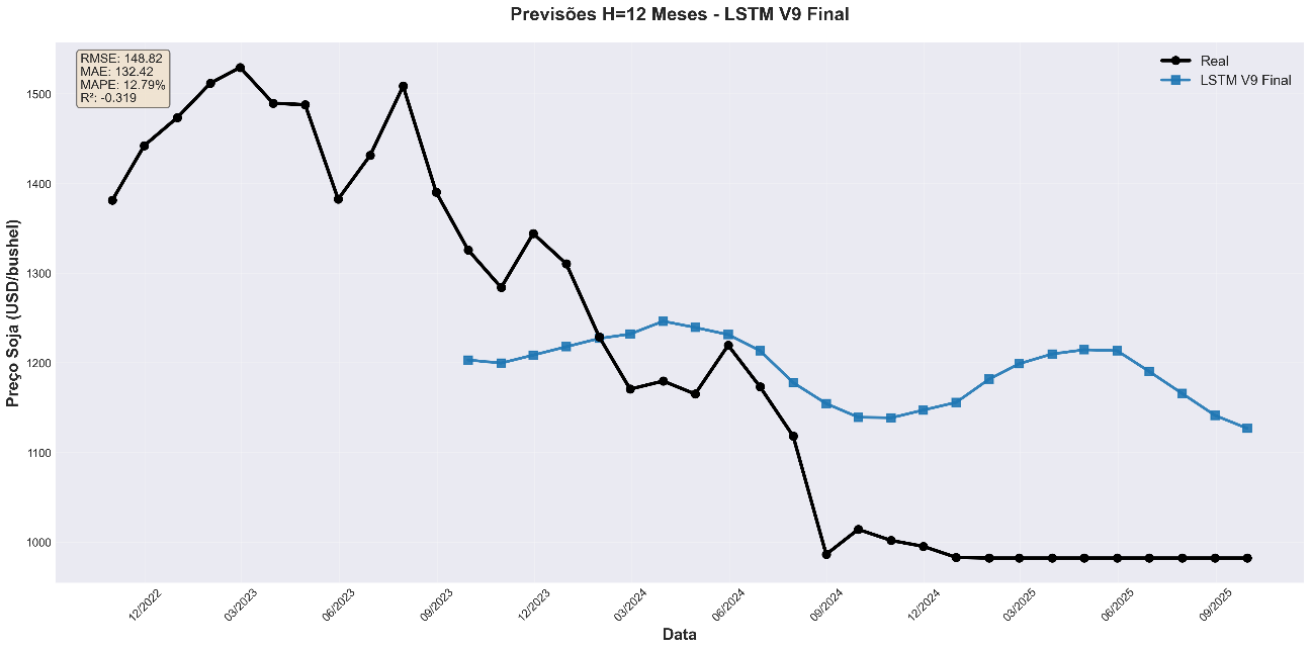


**Figura A.17 – ARIMAX Melhorado Previsões Curto e Médio Prazo**

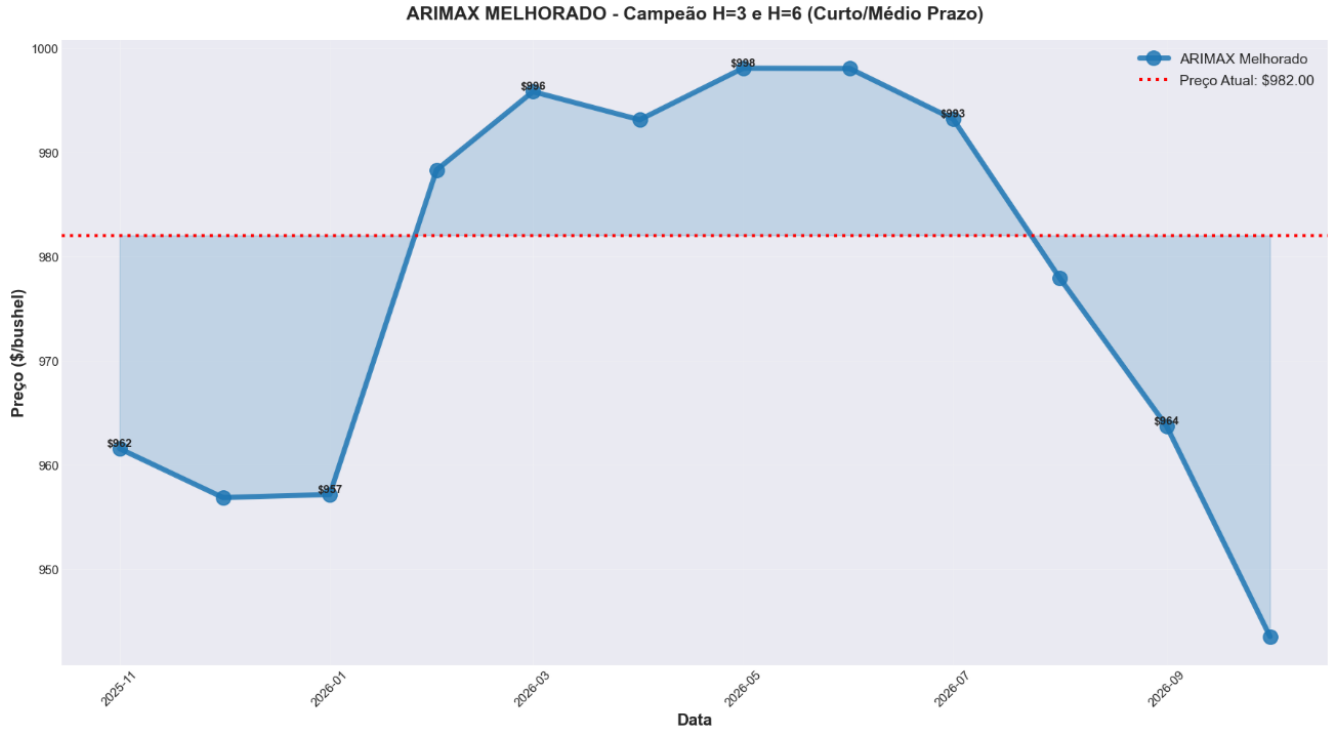



**Figura A.18 – XGBoost Previsões Longo Prazo**

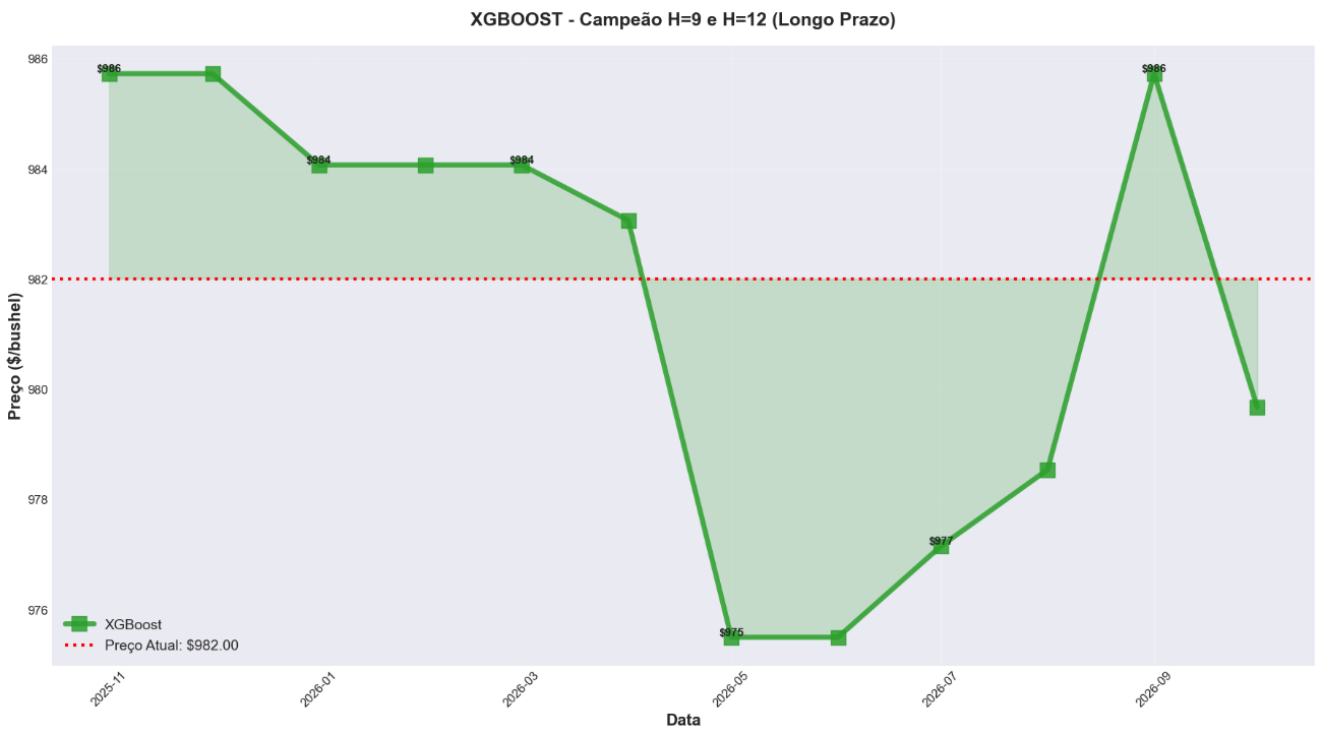


**Figura A.19 – XGBoost importância de variáveis**

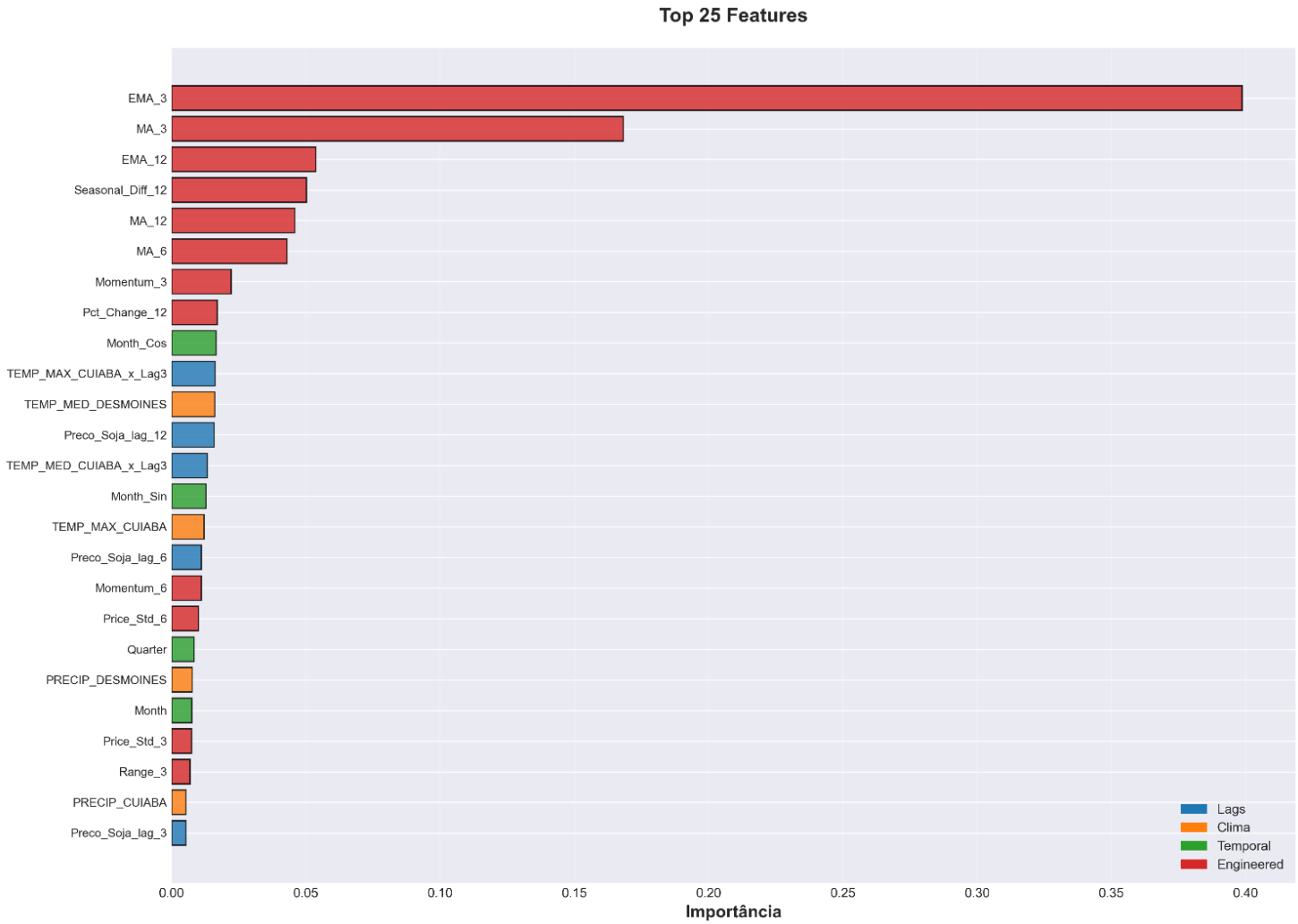


**Tabela 4.6 – Projeções de Preços da Soja para Modelos Campeões**

| Modelo | Horizonte Campeão | Preço Médio 12m | Preço Out/2026 | Variação vs Set/2025 |
|---|---|---|---|---|
| **ARIMAX Melhorado** | H=3, 6 meses | $977.24 | $943.49 | -3.92% |
| **XGBoost** | H=9, 12 meses | $981.56 | $979.67 | -0.24% |

Fonte: Elaboração própria (Notebook 5 - Previsões Futuras)

**Tabela 4.7 – Projeções de Preços da Soja para Novembro/2025 - Outubro/2026**

| Mês | ARIMAX Melhorado ($/bu) | XGBoost ($/bu) |
|---|---|---|
| **Nov/2025** | 961.52 | 985.72 |
| **Dez/2025** | 956.83 | 985.72 |
| **Jan/2026** | 957.12 | 984.07 |
| **Fev/2026** | 988.27 | 984.07 |
| **Mar/2026** | 995.79 | 984.07 |
| **Abr/2026** | 993.09 | 983.06 |
| **Mai/2026** | 998.04 | 975.49 |
| **Jun/2026** | 998.01 | 975.49 |
| **Jul/2026** | 993.16 | 977.15 |
| **Ago/2026** | 977.87 | 978.53 |
| **Set/2026** | 963.67 | 985.72 |
| **Out/2026** | 943.49 | 979.67 |

Fonte: Elaboração própria com base nos modelos campeões.