UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Faculdade de Ciências e Tecnologia
Câmpus de Presidente Prudente

Previsão do Preço do Etanol no Estado de São Paulo Utilizando Técnicas de Aprendizado de Máquina

Thomaz Augusto Ferreira Assis

Orientadora: Profa. Dra. Marilaine Colnago

Programa de Pós-Graduação em Matemática Aplicada e Computacional

Presidente Prudente, março de 2026

UNIVERSIDADE ESTADUAL PAULISTA

Faculdade de Ciências e Tecnologia de Presidente Prudente

Programa de Pós-Graduação em Matemática Aplicada e Computacional

Previsão do Preço do Etanol no Estado de São Paulo Utilizando Técnicas de Aprendizado de Máquina

Thomaz Augusto Ferreira Assis

Orientadora: Profa. Dra. Marilaine Colnago

Dissertação apresentada ao Programa de Pós-Graduação em Matemática Aplicada e Computacional da Faculdade de Ciências e Tecnologia da UNESP para obtenção do título de Mestre em Matemática Aplicada e Computacional.

Presidente Prudente, abril de 2025

A848p

Assis, Thomaz Augusto Ferreira
Previsão do preço do etanol no estado de São Paulo utilizando técnicas de aprendizado de máquina / Thomaz Augusto Ferreira Assis. -- Presidente Prudente, 2026
93 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientadora: Marilaine Colnago

1. Matemática. 2. Inteligência artificial. 3. Machine learning. 4. Deep learning (Machine learning). 5. Ciência da computação Matemática. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Presidente Prudente

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: "Previsão do Preço do Etanol no Estado de São Paulo Utilizando Técnicas de Aprendizado de Máquina".

AUTOR: THOMAZ AUGUSTO FERREIRA ASSIS
ORIENTADORA: MARILAINE COLNAGO

Aprovado como parte das exigências para obtenção do Título de Mestre em Matemática Aplicada e Computacional, pela Comissão Examinadora:

Documento assinado digitalmente
gov.br
MARILAINE COLNAGO
Data: 04/03/2026 17:16:29-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. MARILAINE COLNAGO (Participaçao Virtual)
Departamento de Ciências de Computação e Estatística / UNESP / Câmpus de São José do Rio Preto - Ibilce

Documento assinado digitalmente
gov.br
LARISSA FERREIRA MARQUES
Data: 05/03/2026 08:44:53-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. LARISSA FERREIRA MARQUES (Participaçao Virtual)
Departamento de Matemática / UNESP / Câmpus de Bauru - FC

Documento assinado digitalmente
gov.br
DIEGO RENAN BRUNO
Data: 05/03/2026 11:38:14-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. DIEGO RENAN BRUNO (Participaçao Virtual)
Departamento de Ciências de Computação e Estatística / UNESP / Câmpus de São José do Rio Preto - Ibilce

Presidente Prudente, 04 de março de 2026.

À minha mãe, Ana Maria Ferreira, por ter me mostrado que a matemática pode ser bela e apaixonante, por despertar em mim o amor pela ciência e por sempre me apoiar em cada decisão.

Ao meu tio e padrinho, Alexandre Ferreira, pelo exemplo e pela inspiração do que significa ser um verdadeiro estudioso, bem como pelo suporte emocional e financeiro ao longo da minha jornada acadêmica.

Ao meu sobrinho e afilhado, Fernando Ferreira, pelo seu amor incondicional e por cada abraço e sorriso que trouxeram aconchego aos meus dias de cansaço.

E, com todo o meu coração, ao meu parceiro de vida, Caio Cesar Costa. Obrigado pelos 12 anos de companheirismo, pelo incentivo constante, por me aturar nos dias difíceis e por enfrentar cada obstáculo ao meu lado. Esta conquista também é fruto do amor e da vida que construímos juntos.

Agradecimentos

Expresso minha profunda gratidão à minha orientadora, Profa. Dra. Marilaine Colnago, por ter aceitado o desafio de me orientar neste trabalho. Sua dedicação, parceria, o auxílio técnico fundamental e, sobretudo, sua paciência ao longo de toda esta jornada foram pilares essenciais para a conclusão deste estudo.

Agradeço ao Prof. Dr. Wallace Casaca, meu coorientador não oficial e professor. Suas contribuições e sua trajetória serviram como fonte genuína de inspiração para o desenvolvimento desta pesquisa.

À Profa. Dra. Analice Costacurta Brandi, manifesto meu sincero agradecimento, pela ajuda em momentos difíceis e pelo incentivo à ciência, proporcionando-me meu primeiro contato com a matemática computacional.

À Universidade Virtual do Estado de São Paulo (Univesp), pelo apoio financeiro concedido por meio da bolsa de estudos. Essa viabilização foi fundamental para que eu pudesse concluir este ciclo e me dedicar integralmente ao desenvolvimento desta dissertação.

"Do. Or do not. There is no try."
Mestre Yoda

Resumo

Esta dissertação investiga a previsão dos preços do etanol no estado de São Paulo, com o objetivo de reduzir incertezas de mercado por meio de técnicas de Aprendizagem de Máquina (AM). O estudo compara o desempenho dos modelos *Random Forest* (RF) e *Long Short-Term Memory* (LSTM) aplicados a séries temporais semanais no período de 2004 a 2024, abrangendo diferentes regimes econômicos. A metodologia avaliou a acurácia preditiva dos algoritmos e o papel de variáveis exógenas, como o índice ATR e o preço do açúcar. Os resultados evidenciam desempenho dependente do regime econômico: em cenários estáveis (pré-pandemia), o RF apresentou maior precisão (MAPE de 1,70%). Contudo, sob elevada volatilidade, o modelo mostrou perda de robustez, com erros superiores a 20%. Em contraste, o LSTM otimizado demonstrou maior capacidade de adaptação, destacando-se durante a pandemia (MAPE de 7,89%) e no período pós-pandêmico (4,88%), especialmente quando incorporado o preço do açúcar. Como contribuição prática, a pesquisa disponibiliza um *dashboard* interativo para visualização dos resultados obtidos. Conclui-se que, embora o RF seja adequado em ambientes estáveis, o LSTM oferece maior confiabilidade para a formulação de políticas públicas e o planejamento estratégico em mercados dinâmicos e sujeitos a instabilidades.

Palavras-Chave: *Preços do Etanol, Aprendizagem de Máquina, Random Forest, Long Short-Term Memory.*

Abstract

This dissertation investigates the forecasting of ethanol prices in the state of São Paulo, aiming to reduce market uncertainty through Machine Learning (ML) techniques. The study compares the performance of Random Forest (RF) and Long Short-Term Memory (LSTM) models applied to weekly time series from 2004 to 2024, encompassing different economic regimes. The methodology evaluated predictive accuracy and the role of exogenous variables, such as the ATR index and sugar prices. The results show that model performance depends on the economic regime: during stable periods (pre-pandemic), RF achieved the highest accuracy (MAPE of 1.70%). However, under high volatility, the model exhibited reduced robustness, with errors exceeding 20%. In contrast, the optimized LSTM demonstrated greater adaptability, outperforming RF during the pandemic (MAPE of 7.89%) and post-pandemic period (4.88%), particularly when incorporating sugar price as an explanatory variable. As a practical contribution, the research provides an interactive dashboard for visualizing the obtained results. It is concluded that, while RF performs well in stable environments, LSTM offers greater reliability for public policy formulation and strategic planning in dynamic and unstable markets.

Keywords: *Ethanol Price, Machine Learning, Random Forest, Long Short-Term Memory.*

Lista de Figuras

2.1 Evolução da produção brasileira de etanol, de 1979 a 2010 (em m^3) 15
2.2 Ciclo renovável da cana-de-açúcar. 16
2.3 Comparação de produtividade entre diferentes matérias-primas do etanol . 17

3.1 Representação gráfica da arquitetura do algoritmo *Random Forest*. 28
3.2 Exemplo de Rede Neural Artificial. 30

4.1 Preço de revenda do etanol no estado de São Paulo (2004-2023). 37
4.2 Tendência Spline a partir dos dados pré-pandêmicos. 38
4.3 Preço do açúcar (2004-2023). 39
4.4 Preço do petróleo (2004-2023). 39
4.5 ATR (2004-2023). 40
4.6 Preço da gasolina (2004-2023). 40
4.7 Cotação do dólar (2004-2023). 41
4.8 Açúcar NY (em US$) no período de 2004 a 2023. 42
4.9 Inflação (%) no período de 2004 a 2023. 42
4.10 ICMS (%) no período de 2004 a 2023. 43
4.11 Exportação de etanol (m^3) no período de 2004 a 2023. 43
4.12 Importação de etanol (m^3) no período de 2004 a 2023. 44
4.13 Importação de etanol (m^3) no período de 2004 a 2023. (recorte) 44
4.14 Correlação de Pearson entre as variáveis analisadas. 46
4.15 Importância das variáveis. 47
4.16 Preço do Etanol observado vs. preço previsto pelo modelo RF no conjunto de teste, utilizando a base CCA. 50
4.17 Preço do Etanol observado vs. preço previsto pelo modelo RF pré-pandemia, utilizando a base CCA. 51
4.18 Preço do Etanol observado vs. preço previsto pelo modelo RF durante a pandemia, utilizando a base CCA. 52
4.19 Preço do Etanol observado vs. preço previsto pelo modelo RF pós-pandemia, utilizando a base CCA. 53

4.20 Preço do Etanol observado vs. preço previsto pelo modelo RF no conjunto total, utilizando a base CR-CCA. 53
4.21 Preço do Etanol observado vs. preço previsto pelo modelo RF no conjunto de teste, utilizando a base CR-ATR. 55
4.22 Preço do Etanol observado vs. preço previsto pelo modelo RF pré-pandemia, utilizando a base CR-ATR. 55
4.23 Preço do Etanol observado vs. preço previsto pelo modelo RF durante a pandemia, utilizando a base CR-ATR. 56
4.24 Preço do Etanol observado vs. preço previsto pelo modelo RF pós-pandemia, utilizando a base CR-ATR. 57
4.25 Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto total, utilizando a base CR-ATR. 57
4.26 Preço do Etanol observado vs. preço previsto pelo modelo RF no conjunto de teste, utilizando a base CR-AC. 59
4.27 Preço do Etanol observado vs. preço previsto pelo modelo RF pré-pandemia, utilizando a base CR-AC. 59
4.28 Preço do Etanol observado vs. preço previsto pelo modelo RF durante a pandemia, utilizando a base CR-AC. 60
4.29 Preço do Etanol observado vs. preço previsto pelo modelo RF pós-pandemia, utilizando a base CR-AC. 61
4.30 Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto total, utilizando a base CR-AC. 61
4.31 Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto de teste, utilizando a base CCA. 63
4.32 Preço do Etanol observado vs. preço previsto pelo modelo LSTM pré-pandemia, utilizando a base CCA. 64
4.33 Preço do Etanol observado vs. preço previsto pelo modelo LSTM durante a pandemia, utilizando a base CCA. 64
4.34 Preço do Etanol observado vs. preço previsto pelo modelo LSTM pós-pandemia, utilizando a base CCA. 65
4.35 Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto total, utilizando a base CCA. 65
4.36 Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto de teste, utilizando a base CR-ATR. 67
4.37 Preço do Etanol observado vs. preço previsto pelo modelo LSTM pré-pandemia, utilizando a base CR-ATR. 67
4.38 Preço do Etanol observado vs. preço previsto pelo modelo LSTM durante a pandemia, utilizando a base CR-ATR. 68
4.39 Preço do Etanol observado vs. preço previsto pelo modelo LSTM pós-pandemia, utilizando a base CR-ATR. 69
4.40 Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto total, utilizando a base CR-ATR. 69

4.41 Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto de teste, utilizando a base CR-AC. 71

4.42 Preço do Etanol observado vs. preço previsto pelo modelo LSTM pré-pandemia, utilizando a base CR-AC. 71

4.43 Preço do Etanol observado vs. preço previsto pelo modelo LSTM durante a pandemia, utilizando a base CR-AC. 72

4.44 Preço do Etanol observado vs. preço previsto pelo modelo LSTM pós-pandemia, utilizando a base CR-AC. 73

4.45 Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto total, utilizando a base CR-AC. 73

4.46 Página 1: Comparação entre Séries Históricas e Projeções 77

4.47 Página 2: Visualização Comparativa de Variáveis 78

Lista de Tabelas

3.1 Conjunto principal de dados coletados, organizados semanalmente. 24

4.1 Fator de Inflação de Variância (VIF) das variáveis selecionadas. 48

4.2 Fator de Inflação de Variância (VIF) após rotina de exclusão. 48

4.3 Fator de Inflação de Variância (VIF) após rotina de exclusão, sem ATR. . . 49

4.4 Métricas de erro do modelo *Random Forest* nos diferentes períodos. 54

4.5 Métricas de erro do modelo *Random Forest* utilizando a base CR-ATR, nos diferentes períodos. 58

4.6 Métricas de erro do modelo *Random Forest* utilizando a base CR-AC, nos diferentes períodos. 62

4.7 Métricas de erro do modelo LSTM utilizando base de dados CCA nos diferentes períodos. 66

4.8 Métricas de erro do modelo LSTM utilizando a base CR-ATR, nos diferentes períodos. 70

4.9 Métricas de erro do modelo LSTM utilizando a base CR-AC, nos diferentes períodos. 74

4.10 Resultados de MAPE, MAE e MSE para os Modelos RF e LSTM nas diferentes bases de dados e períodos. 75

Sumário

Resumo **5**

Abstract **6**

Lista de Figuras **7**

Lista de Tabelas **9**

1 Introdução **13**

1.1 Objetivos 14

2 Contextualização e Justificativa do Estudo **15**

2.1 O papel do etanol na matriz energética brasileira 15

2.2 Impactos econômicos da COVID-19 18

2.3 Aprendizado de Máquina na Previsão de Preços no Mercado de Biocombustíveis 19

3 Materiais e Métodos **22**

3.1 Linguagem de programação 22

3.2 Base de Dados 22

3.3 Pré-processamento e Tratamento dos Dados 24

3.4 Divisão da Base de Dados 25

3.5 Análise e Seleção de Variáveis 25

3.5.1 Análise de Correlação 26

3.5.2 Seleção Estatística de Atributos 26

3.5.3 Importância das Variáveis (*Feature Importance*) 27

3.5.4 Análise de Multicolinearidade 27

3.5.5 Engenharia de Atributos 27

3.6 Modelos de Aprendizado de Máquina 27

3.6.1 Random Forest 28

3.6.2 Redes Neurais Artificiais 30

3.6.3 Aprendizado profundo (Deep Learning) . . . 31
3.6.4 Redes Neurais Recorrentes . . . 32
3.6.5 Long Short-Term Memory (LSTM) . . . 33
3.7 Métricas de Avaliação . . . 34
3.8 Desenvolvimento de Dashboard Interativo . . . 35

4 Resultados e Discussão **37**
4.1 Análise Exploratória . . . 37
4.2 Importância e Seleção das Variáveis . . . 45
4.3 Previsão do Modelo Random Forest . . . 49
4.3.1 Base CCA . . . 50
4.3.2 Base CR-ATR . . . 54
4.3.3 Base CR-AC . . . 58
4.4 Previsão do Modelo LSTM . . . 62
4.4.1 Base CCA . . . 63
4.4.2 Base CR-ATR . . . 66
4.4.3 Base CR-AC . . . 70
4.5 Comparação e discussão das métricas . . . 74
4.6 Visualização de Resultados por Dashboard . . . 76

5 Conclusões **79**

Referências **80**

Capítulo

1

Introdução

O etanol é um biocombustível estratégico para o Brasil, especialmente no estado de São Paulo, responsável por mais da metade da produção nacional (SEADE, 2023). Sua relevância ultrapassa a esfera energética, desempenhando papel central na promoção do desenvolvimento socioeconômico regional e nacional, além de contribuir significativamente para a mitigação das emissões de gases de efeito estufa (Goldemberg, 2008). Ao substituir parcial ou integralmente combustíveis fósseis, o etanol configura-se como uma alternativa renovável que fortalece a sustentabilidade ambiental e energética do país (Macedo et al., 2008). Ademais, sua cadeia produtiva gera empregos e renda em áreas urbanas e rurais, consolidando-se como importante vetor de desenvolvimento sustentável.

Em 2023, o etanol representou 48% da matriz energética brasileira, evidenciando sua forte presença no *mix* de combustíveis do país (EPE, 2023). O Brasil, como segundo maior produtor mundial, respondeu por cerca de 30% da oferta global, destacando-se por uma produtividade média significativamente superior à do etanol de milho produzido nos Estados Unidos (Association, 2023). Esse protagonismo decorre de uma combinação entre políticas públicas voltadas ao setor, avanços tecnológicos no cultivo e processamento da cana-de-açúcar e fatores estruturais, como condições climáticas favoráveis e disponibilidade de terras agricultáveis.

Apesar dessas vantagens competitivas, o setor sucroenergético brasileiro enfrenta desafios relevantes, entre os quais se sobressai a elevada volatilidade dos preços do etanol. Essa flutuação decorre de múltiplos fatores, incluindo condições climáticas, variações cambiais, políticas tributárias e dinâmicas do mercado internacional (Lucena et al., 2010). Entre 2010 e 2020, os preços do etanol no estado de São Paulo apresentaram um coeficiente de variação de 18,7%, superior ao de outras *commodities* relevantes, como soja (12,4%) e café (14,9%) (Ipea, 2022). Tal instabilidade gera incertezas para produtores, distribuidores e formuladores de políticas públicas, dificultando o planejamento tanto no curto quanto no longo prazo nos setores agrícola e energético.

Nesse contexto, torna-se essencial o desenvolvimento de ferramentas analíticas capazes de fornecer previsões mais precisas e confiáveis dos preços do etanol. Técnicas de Inteligência Artificial (IA) e Aprendizado de Máquina (AM) têm se mostrado promissoras, especialmente por sua capacidade de capturar relações não lineares e padrões temporais complexos em grandes volumes de dados (Hyndman and Athanasopoulos, 2018b). Modelos como o *Random Forest* (RF), baseados em métodos de aprendizado por conjunto, e as redes neurais recorrentes *Long Short-Term Memory* (LSTM), particularmente eficazes

em séries temporais, vêm sendo amplamente aplicados em contextos financeiros, agrícolas e energéticos, com resultados consistentes.

Dessa forma, este trabalho concentra-se na aplicação de técnicas de aprendizado de máquina para a previsão dos preços do etanol no Brasil, com ênfase nos modelos *Random Forest* (RF) e *Long Short-Term Memory* (LSTM). Ao explorar o potencial dessas abordagens, busca-se fornecer subsídios que contribuam para o planejamento estratégico de agentes públicos e privados, promovendo maior previsibilidade no setor energético e fortalecendo a sustentabilidade econômica e ambiental do país.

1.1 Objetivos

O objetivo geral deste estudo é avaliar o desempenho dos modelos *Random Forest* (RF) e *Long Short-Term Memory* (LSTM) na previsão dos preços do etanol no Brasil, visando gerar subsídios para o planejamento estratégico no setor energético.

Os objetivos específicos são:

- Analisar a evolução histórica dos preços do etanol e sua relação com variáveis econômicas e setoriais relevantes;
- Implementar modelos preditivos baseados em aprendizado de máquina, com foco nas arquiteturas RF e LSTM;
- Avaliar o desempenho preditivo dos modelos por meio de métricas estatísticas apropriadas;
- Desenvolver um *dashboard* interativo de acesso aberto para visualização dos resultados da pesquisa, permitindo a exploração dinâmica dos modelos, métricas e padrões identificados na série histórica analisada;
- Comparar os resultados obtidos e discutir suas implicações para a formulação de políticas públicas e para estratégias empresariais no setor sucroenergético.

A estrutura deste trabalho está organizada da seguinte forma: o Capítulo 2 discute o papel socioeconômico do etanol e os principais fatores que influenciam seus preços; o Capítulo 3 apresenta os fundamentos teóricos dos modelos RF e LSTM; o Capítulo 4 descreve os dados utilizados e a metodologia adotada; o Capítulo 5 apresenta os resultados e sua análise crítica; e o Capítulo 6 reúne as principais contribuições do estudo, suas limitações e propostas para pesquisas futuras.

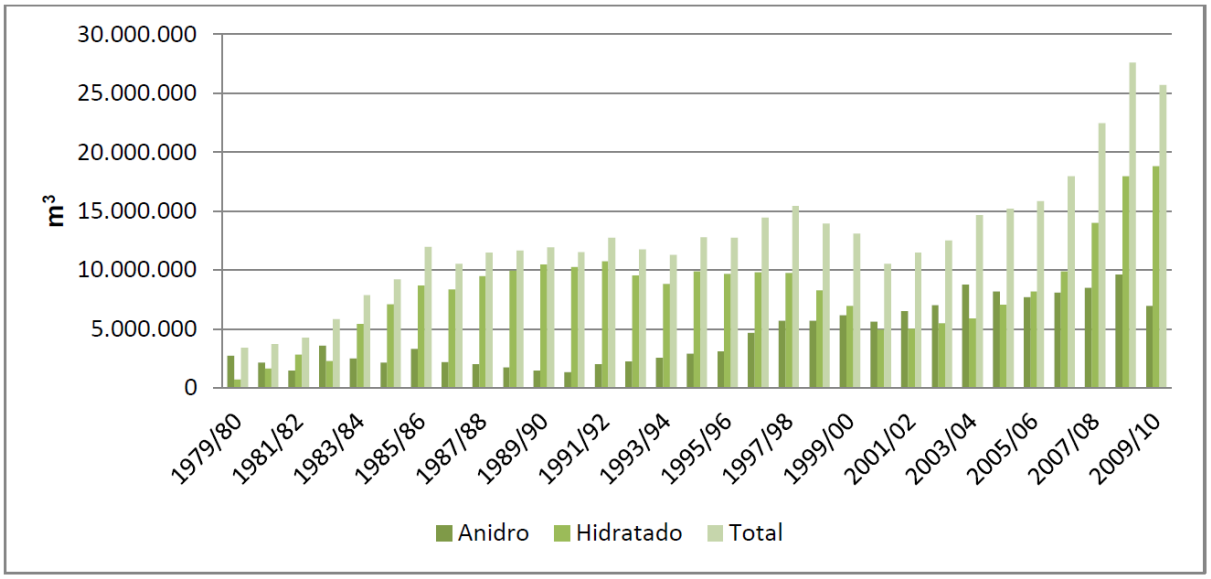
Capítulo 2

Contextualização e Justificativa do Estudo

2.1 O papel do etanol na matriz energética brasileira

O etanol, como alternativa à gasolina, desempenha papel relevante na diversificação da matriz energética e na redução da dependência de combustíveis fósseis. No Brasil, sua consolidação está diretamente associada ao Programa Nacional do Álcool (Proálcool), instituído na década de 1970 em resposta à crise mundial do petróleo (Dunham et al., 2011b). Esse programa impulsionou o desenvolvimento de tecnologias de produção e fomentou políticas públicas que viabilizaram sua adoção em larga escala, especialmente com a posterior disseminação dos veículos bicombustíveis. A evolução da produção nacional de etanol ao longo do tempo pode ser observada na Figura 2.1.

Figura 2.1: Evolução da produção brasileira de etanol, de 1979 a 2010 (em m^3)



Fonte: Santos (2012b).

Os benefícios do etanol extrapolam a redução da dependência do petróleo. Conforme destacam Tonin and Tonin (2014), a busca por combustíveis renováveis voltou ao centro das discussões entre os *stakeholders*, e o etanol emergiu como biocombustível de elevado potencial produtivo, favorecido pela disponibilidade de terras agricultáveis e por condições climáticas adequadas ao cultivo da cana-de-açúcar. Do ponto de vista ambiental, o etanol apresenta menores emissões de gases de efeito estufa em comparação à gasolina, contribuindo para uma matriz energética mais limpa. Além disso, sua produção é renovável, baseada no ciclo da cana-de-açúcar, que absorve CO_2 da atmosfera durante seu crescimento, como ilustrado na Figura 2.2 (Santos and Peixoto, 2008).

Figura 2.2: Ciclo renovável da cana-de-açúcar.



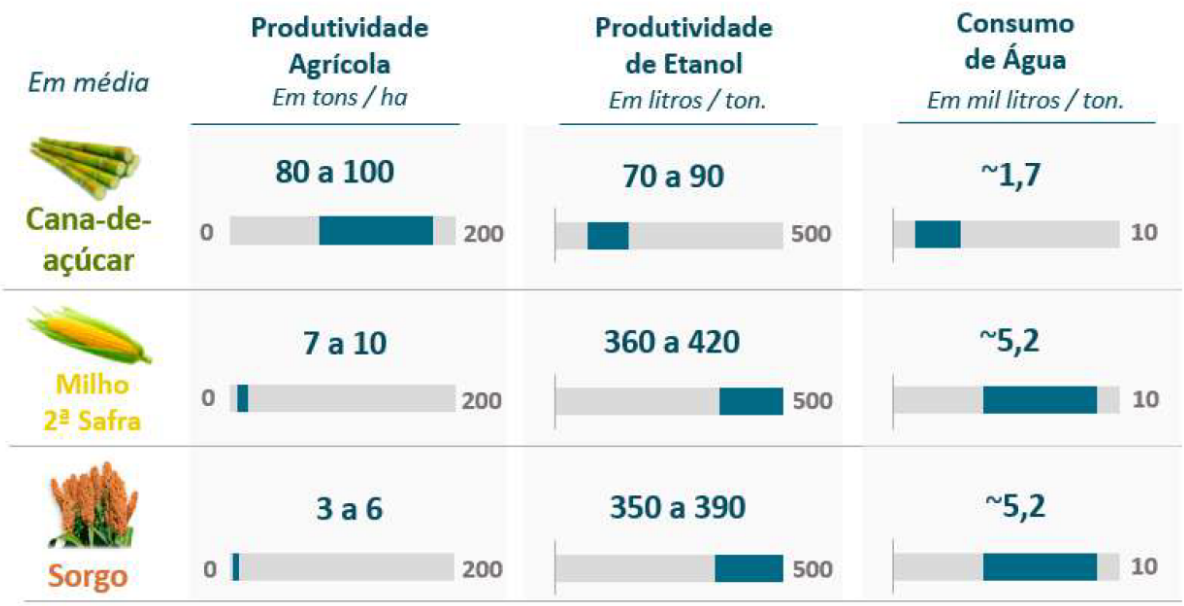
Fonte: Usina Açucareira de São Manoel (2025).

Um diferencial importante do modelo brasileiro é o aproveitamento de resíduos do processo produtivo, como o bagaço e a palha da cana, utilizados na geração de bioeletricidade. Essa prática aumenta a eficiência energética da cadeia produtiva e reforça seu caráter sustentável. Como observado por Santos (2012a), o crescimento do setor sucro-

alcooleiro ampliou significativamente o potencial de geração de bioeletricidade a partir dessa biomassa, fortalecendo sua participação na matriz energética nacional.

A predominância da cana-de-açúcar como matéria-prima para o etanol no Brasil, em contraste com o milho utilizado nos Estados Unidos, deve-se principalmente à maior produtividade por hectare e à menor demanda relativa por insumos agrícolas (Silva et al., 2020). O milho, além de requerer volumes expressivos de água e fertilizantes, compete diretamente com a cadeia alimentar. A cana-de-açúcar, por sua vez, apresenta maior rendimento energético e possibilita o aproveitamento integral da planta, consolidando-se como a matéria-prima mais eficiente para a produção de etanol no contexto brasileiro (Santos et al., 2012), conforme ilustrado na Figura 2.3.

Figura 2.3: Comparação de produtividade entre diferentes matérias-primas do etanol



Fonte: Neves et al. (2021).

A produção de etanol envolve diversas etapas, desde a colheita da cana-de-açúcar — manual ou mecanizada — até os processos de moagem e fermentação, nos quais os açúcares da planta são convertidos em álcool. No estado de São Paulo, essa cadeia produtiva é favorecida por condições climáticas adequadas, infraestrutura consolidada e políticas de incentivo que consolidaram o estado como líder nacional na produção e comercialização de etanol (Rodrigues Filho and Juliani, 2013). São Paulo responde por aproximadamente metade da produção brasileira e figura também como o principal consumidor, devido à ampla frota de veículos e à concentração de usinas e distribuidoras.

O mercado de etanol é dinâmico e sensível a variações sazonais e econômicas. Fatores como condições climáticas, políticas governamentais e os preços internacionais do petróleo influenciam diretamente sua produção e comercialização (Piacente, 2006). O estado de São Paulo destaca-se nesse contexto por sua estrutura de mercado robusta, infraestrutura de distribuição desenvolvida e acesso a portos estratégicos para exportação, configurando-se como ator central no cenário global de biocombustíveis. Conforme ressalta Santos (2012b), a proximidade com o principal centro consumidor do país e com os portos de exportação reforça a centralidade de São Paulo na cadeia sucroenergética.

Diante desses fatores, torna-se evidente a relevância de estudos que investiguem a produção e a comercialização do etanol em São Paulo, contribuindo para o aprimoramento de políticas públicas e para o desenvolvimento de estratégias empresariais mais eficientes no setor sucroenergético.

2.2 Impactos econômicos da COVID-19

A pandemia de COVID-19 desencadeou uma instabilidade econômica sem precedentes em escala global, resultante de medidas de *lockdown*, interrupções nas cadeias de suprimentos e queda abrupta da demanda em diversos setores produtivos (Ozili and Arun, 2023). Estudos iniciais destacaram o choque financeiro imediato enfrentado pelas empresas, especialmente pequenas e médias (PMEs), que lidaram com crises de liquidez e aumento da incerteza econômica (Apeksha and Nandi, 2021). Intervenções governamentais, como pacotes de estímulo fiscal e políticas de flexibilização monetária, foram fundamentais para mitigar os impactos imediatos, embora sua eficácia tenha variado entre os países, conforme as estratégias adotadas e as condições econômicas pré-existentes (Baldwin and Weder di Mauro, 2020).

Além dos efeitos macroeconômicos, a literatura evidencia que os impactos da pandemia foram distribuídos de forma desigual. Famílias de baixa renda e populações vulneráveis sofreram de maneira mais intensa com a perda de empregos e a redução da renda, ampliando desigualdades sociais já existentes (Apeksha and Nandi, 2021). Paralelamente, a transição acelerada para o trabalho remoto e o aumento da dependência de tecnologias digitais impulsionaram mudanças estruturais no mercado de trabalho, com possíveis efeitos persistentes sobre o emprego em determinados setores. No comércio internacional, as disrupções logísticas evidenciaram a fragilidade das cadeias globais de produção, estimulando revisões nas estratégias organizacionais e produtivas das empresas (Espitia et al., 2023).

No período pós-agudo da pandemia, pesquisas indicam que, apesar de sinais de recuperação impulsionados pela demanda reprimida e pelo apoio governamental, vulnerabilidades econômicas estruturais permaneceram. Gargalos nas cadeias de suprimentos continuaram a pressionar os preços, contribuindo para o avanço da inflação e dificultando um crescimento sustentado (Apeksha and Nandi, 2021). Ademais, o aumento da dívida pública e a possibilidade de novas ondas do vírus configuraram riscos adicionais à estabilidade econômica (Baldwin and Weder di Mauro, 2020). Em conjunto, esses estudos ressaltam a necessidade de políticas voltadas à construção de resiliência e ao crescimento econômico inclusivo no cenário pós-pandemia.

Os efeitos da COVID-19 também se manifestaram de forma marcante em setores específicos, como turismo, aviação e hotelaria, que enfrentaram interrupções prolongadas (Ozili and Arun, 2023). O colapso das viagens internacionais impactou severamente economias dependentes dessas atividades, reforçando a importância da diversificação produtiva e de estratégias adaptativas para o chamado "novo normal" (Espitia et al., 2023). Essas evidências reforçam que choques sistêmicos podem gerar repercussões duradouras, exigindo intervenções direcionadas e reformas estruturais.

No caso da cadeia produtiva do etanol no Brasil, a pandemia provocou impactos negativos em praticamente todos os estados produtores, com risco de colapso em determinadas regiões. Ainda assim, São Paulo e Mato Grosso apresentaram maior capacidade de adaptação. Em São Paulo, esse desempenho esteve associado, entre outros fatores, à receita

proveniente de créditos de carbono vinculados à Política Nacional de Biocombustíveis (RenovaBio), que estimula a produção de biocombustíveis e a comercialização de créditos de descarbonização (CBIOs), condicionados ao volume de etanol comercializado e ao Índice de Energia-Ambiental RenovaBio (Nocera Alves Junior et al., 2022).

Diante desse contexto, torna-se fundamental analisar o setor de etanol considerando os períodos pré, durante e pós-pandemia. Cada fase apresenta características distintas que influenciam diretamente a dinâmica de produção, demanda e formação de preços. A crise sanitária evidenciou vulnerabilidades do setor frente a choques inesperados, enquanto o período subsequente revela uma recuperação ainda incerta, moldada por novas políticas energéticas, mudanças nos padrões de consumo e transformações econômicas estruturais. A compreensão desses diferentes contextos é essencial para avaliar a resiliência do setor, aprimorar modelos de previsão e subsidiar decisões estratégicas e políticas públicas voltadas à estabilidade e sustentabilidade do mercado.

2.3 Aprendizado de Máquina na Previsão de Preços no Mercado de Biocombustíveis

O uso de técnicas de aprendizado de máquina (AM) tem se consolidado como uma abordagem relevante na previsão de preços e na modelagem de sistemas associados aos biocombustíveis. Em comparação com modelos estatísticos tradicionais, essas técnicas demonstram maior capacidade de lidar com relações não lineares, interações complexas entre variáveis e grandes volumes de dados, características comuns em mercados de commodities energéticas (Wang et al., 2023; Sharma et al., 2023; Paula et al., 2022; Sumayli, 2023; Manogna et al., 2025). Métodos clássicos, como ARIMA, embora amplamente utilizados, apresentam limitações em cenários de alta volatilidade e dinâmicas estruturais complexas (Wu et al., 2024).

No setor de biocombustíveis, a precisão na previsão de preços tem um impacto direto na tomada de decisões estratégicas ao longo da cadeia produtiva. A capacidade de prever com acurácia os preços futuros não só minimiza os riscos financeiros e operacionais, mas também auxilia no planejamento de investimentos, estocagem e estratégias de negociação (Arroyo-Marioli et al., 2023). Essa habilidade de prever com precisão é vital para otimizar a alocação de recursos e o planejamento da produção, além de contribuir para a redução de custos e melhoria da eficiência geral (Ludovico et al., 2022). Ao mesmo tempo, a modelagem por meio de ML permite substituir técnicas computacionalmente intensivas, como simulações de Monte Carlo, por soluções preditivas que são não apenas mais rápidas e eficientes, mas também mais adaptáveis a mudanças rápidas no mercado (Wu et al., 2024).

Na literatura recente, diversos algoritmos têm sido empregados na modelagem de mercados de biocombustíveis, incluindo redes neurais recorrentes, métodos baseados em vetores de suporte e modelos ensemble. As Redes Neurais Artificiais, como LSTM (*Long Short-Term Memory*) e GRU (*Gated Recurrent Units*), destacam-se por sua capacidade de capturar dependências temporais de longo prazo e padrões não lineares em séries temporais (Manogna et al., 2025; Tami and Owda, 2024). Essa habilidade é crucial em mercados como o de biocombustíveis, que são altamente sensíveis a fatores sazonais, políticas públicas e até mudanças climáticas.

Esses algoritmos são amplamente utilizados por sua flexibilidade em lidar com diferentes tipos de dados, como séries temporais, dados espectrais e até parâmetros de processos

produtivos que influenciam diretamente a produção e a previsão de preços de biocombustíveis (Sharma et al., 2023; Paula et al., 2022; Xing et al., 2021b). O modelo LSTM se destaca por ser capaz de preservar informações importantes ao longo do tempo, superando limitações de modelos tradicionais como ARIMA, especialmente em cenários de previsão de demanda e preços (Hochreiter and Schmidhuber, 1997; Márquez et al., 2021; Monner and Reggia, 2012; Tami and Owda, 2024). Além disso, o LSTM se adapta bem a choques externos no mercado, tornando-se uma ferramenta robusta para enfrentar as incertezas características desses setores.

Modelos baseados em árvores, como *Random Forest* e *Gradient Boosting*, também têm sido amplamente utilizados devido à sua robustez frente a ruído e à capacidade de modelar interações complexas entre variáveis (Sharma et al., 2023; Paula et al., 2022). Neste caso, o modelo *Random Forest* se destaca quando aplicado a dados complexos e de alta dimensionalidade, como a previsão de eficiência na produção de bioetanol ou a avaliação de características de biomassa lignocelulósica, apresentando resultados sólidos e confiáveis em uma variedade de contextos (Sharma et al., 2023; Ludovico et al., 2022).

No mercado de biocombustíveis, tanto os processos produtivos quanto as flutuações de preços de commodities são influenciados por múltiplos fatores interdependentes e dinâmicos. A interação entre aspectos econômicos, políticos, climáticos e sociais pode gerar padrões complexos que são difíceis de modelar com métodos tradicionais. Algoritmos como LSTM e Redes Neurais são especialmente eficazes nesse sentido, pois têm a capacidade de capturar essas interações complexas de maneira mais eficiente (Naumzik and Feuerriegel, 2020; Marcjasz et al., 2020; Theofilou et al., 2025). Nesse contexto, a seleção adequada de *features* e a inclusão de variáveis exógenas (como dados climáticos e macroeconômicos) mostram-se fundamentais para o desempenho dos modelos (Bhavani and Mounika, 2022).

Outro fator importante na escolha desses algoritmos é o seu alto desempenho preditivo, mesmo quando os conjuntos de dados são reduzidos ou apresentam ruído significativo. Isso é especialmente relevante em setores como o de biocombustíveis, onde dados completos e de alta qualidade podem ser difíceis de obter. Modelos como o *Random Forest*, por exemplo, são bem conhecidos pela sua robustez diante de conjuntos de dados limitados ou imprecisos, tornando-os ferramentas valiosas para previsão de preços e eficiência de produção, mesmo em cenários de dados não ideais (Naumzik and Feuerriegel, 2020).

Outras abordagens, como SVM/SVR, Gaussian Process Regression (GPR) e métodos híbridos — por exemplo, o ARIMA-LSTM, combina a linearidade do ARIMA com a não linearidade das redes neurais, ou modelos otimizados por algoritmos genéticos, como o o GA-ELM (*Genetic Algorithm optimized Extreme Learning Machine*) — têm sido investigadas na literatura com resultados promissores em contextos específicos (Ray et al., 2023; Sari et al., 2024). Estudos comparativos, como o de (Ludovico et al., 2022) e a revisão de (Theofilou et al., 2025), demonstram que o SVM e o SVR (*Support Vector Regression*) têm performance destacada na previsão de preços de commodities, superando outros modelos em termos de acurácia em cenários específicos. Entretanto, a escolha do modelo ideal depende fortemente das características da série analisada, da disponibilidade de dados e do objetivo do estudo (Bhavani and Mounika, 2022).

Diante dessa diversidade metodológica, este estudo opta por concentrar a análise nos modelos *Random Forest* (RF) e *Long Short-Term Memory* (LSTM), por representarem duas abordagens consolidadas e complementares: uma baseada em métodos ensemble e outra em redes neurais recorrentes profundas. Essa escolha permite avaliar o desempenho de modelos amplamente reconhecidos na literatura, mantendo o foco na comparação entre

técnicas robustas e aplicáveis a séries temporais econômicas, sem a complexidade adicional associada a arquiteturas híbridas ou fortemente parametrizadas.

Capítulo

3

Materiais e Métodos

3.1 Linguagem de programação

A linguagem de programação *Python* foi utilizada como base para a realização das etapas deste trabalho, abrangendo o pré-processamento, a análise de variáveis, o treinamento dos modelos, a geração de gráficos e a avaliação das métricas de desempenho. As principais bibliotecas utilizadas foram:

- ***Pandas***: responsável pela estruturação, limpeza e segmentação da base histórica de dados;
- ***Numpy***: utilizada para operações matriciais e o tratamento matemático das séries temporais;
- ***Matplotlib***: empregada na geração de gráficos técnicos e visualização das tendências e oscilações de mercado;
- ***Scikit-learn***: aplicada na implementação do modelo *Random Forest* e na otimização de hiperparâmetros;
- ***Keras***: fundamental para a construção, configuração e treinamento da arquitetura da rede neural *LSTM*.

Cabe destacar que o *Python* é uma linguagem de código aberto que conta com uma ampla comunidade de desenvolvedores e suporte contínuo. Tais fatores tornam a ferramenta altamente confiável e eficiente para a execução de projetos desse porte (Van Rossum and Drake, 2011).

3.2 Base de Dados

O presente estudo utiliza séries temporais semanais relacionadas ao mercado de etanol no estado de São Paulo, abrangendo o período de 09 de maio de 2004 a 26 de novembro de 2023, totalizando 1.021 observações. A variável alvo é o preço de revenda do etanol hidratado, enquanto as demais variáveis foram selecionadas para representar fatores energéticos, agrícolas, macroeconômicos e sazonais que influenciam sua formação de preços.

Os dados foram obtidos a partir de fontes públicas e oficiais, garantindo reprodutibilidade e confiabilidade. Em particular, foram coletados os conjuntos de dados dos seguintes repositórios públicos:

- *Agência Nacional do Petróleo, Gás Natural e Biocombustíveis* (ANP) (ANP, 2023);
- *Instituto Brasileiro de Geografia e Estatística* (IBGE) (IBGE, 2023);
- *Centro de Estudos Avançados em Economia Aplicada da USP* (CEPEA) (CEPEA, 2023);
- *Companhia Nacional de Abastecimento* (CONAB) (CONAB, 2023);
- *Plataforma de mercados financeiros Investing.com* (Investing.com, 2023);
- *Administração de Informação de Energia dos EUA* (EIA) (EIA, 2024), que é a agência estatística do Departamento de Energia dos EUA, e;
- *Ipeadata* (IPEA, 2024).

As variáveis consideradas refletem: (i) condições do mercado de energia (preço da gasolina e do petróleo), (ii) indicadores agrícolas e de matéria-prima (ATR da cana e preços do açúcar), (iii) fatores macroeconômicos (inflação e taxa de câmbio), (iv) dinâmica comercial (importações e exportações de etanol) e (v) efeitos sazonais. Para capturar a sazonalidade da produção, foi criada a variável binária ENSAFRA, que assume valor 0 durante o período de safra (abril–novembro) e 1 na entressafra (dezembro–março).

Uma variável binária, denominada ENSAFRA, foi criada, assumindo o valor 1 se o período for fora da safra (dezembro a março) ou 0 se for o período de safra (abril a novembro). Para melhor leitura dos dados adquiridos, a Tabela 3.1 resume as principais variáveis utilizadas para orientar a análise de dados e as previsões com os modelos de AM.

Tabela 3.1: Conjunto principal de dados coletados, organizados semanalmente.

Variável	**Descrição**	**Unidade**	**Repositório**
Preço do Açúcar Brasileiro	Preço por saco de 50 kg de açúcar cristal	BRL	CEPEA
Preço da Gasolina	Preço de revenda da gasolina por litro	BRL	ANP
Preço do Açúcar NY	Cents por libra	USD	Investing.com
Preço do Petróleo	Preço Spot do WTI FOB por barril	USD	EIA
Preço ATR da Cana	Açúcar total recuperável por kg	BRL	CONAB
Inflação	Índice de Preços ao Consumidor Amplo (IPCA)	% a.m.	IBGE
Dólar	Cotação comercial do dólar	BRL	Ipeadata
ENSAFRA	-	Binário	-
Importação	Volume de etanol importado	m^3	ANP
Exportação	Volume de etanol exportado	m^3	ANP
Preço do Etanol	Preço de revenda de etanol hidratado por litro	BRL	ANP

3.3 Pré-processamento e Tratamento dos Dados

O tratamento dos dados foi realizado para assegurar consistência, integridade e adequação à modelagem preditiva. Inicialmente, as séries foram sincronizadas temporalmente, garantindo periodicidade semanal uniforme.

Foram realizadas etapas de limpeza de dados, incluindo remoção de registros duplicados, padronização de unidades e verificação de consistência com as fontes originais. A limpeza de dados é uma etapa crítica para garantir a confiabilidade das análises subsequentes (Kim et al., 2003). Valores ausentes foram tratados por imputação baseada em média ou mediana, conforme a distribuição da variável.

As variáveis numéricas foram normalizadas para reduzir efeitos de escala e melhorar o desempenho dos modelos de aprendizado de máquina. A normalização é essencial para garantir que todas as variáveis contribuam de forma equitativa para o modelo, evitando viés devido a diferenças de escala (Han et al., 2011). Essa etapa é particularmente relevante para algoritmos sensíveis à magnitude das variáveis, como redes neurais.

Em seguida, foi conduzida uma Análise Exploratória de Dados (*Exploratory Data Analysis - EDA*), com o objetivo de identificar padrões, tendências e relações entre variáveis. Foram utilizadas estatísticas descritivas e visualizações, como histogramas e matrizes de correlação, auxiliando na compreensão da estrutura dos dados e na validação das va-

riáveis selecionadas. A EDA é uma etapa essencial para a compreensão dos dados e para a seleção de variáveis relevantes para a modelagem (Tukey, 1977).

3.4 Divisão da Base de Dados

A pandemia de COVID-19 representou uma ruptura estrutural relevante no comportamento dos preços de combustíveis, introduzindo níveis atípicos de volatilidade. Dessa forma, para avaliar a robustez dos modelos sob diferentes regimes econômicos, a série temporal foi segmentada em três períodos distintos baseados em observações semanais:

- **Pré-pandemia:** de 09/05/2004 a 31/12/2019, composto por 817 observações;
- **Pandemia:** de 01/01/2020 a 31/12/2022, contendo 157 observações;
- **Pós-pandemia:** de 01/01/2023 a 26/11/2023, compreendendo as 47 observações finais.

O período pré-pandemia foi o único utilizado para as fases de ajuste dos modelos, com uma divisão de 95% para treinamento (776 observações) e 5% para teste (41 observações). A escolha de destinar 95% dos dados para treino é fundamental devido à alta volatilidade do preço do etanol, que exige um volume robusto de informações históricas para capturar tendências de longo prazo e ciclos de preços antes da avaliação em dados não vistos (Vabalas et al., 2019). Os períodos de pandemia e pós-pandemia foram reservados exclusivamente para validação externa, funcionando como um "teste de estresse"para medir a capacidade de generalização dos modelos em cenários de mudança brusca de regime. Embora a divisão interna inicial tenha seguido a proporção de 95/5, a configuração final do estudo dispõe de 204 observações para validação externa. Esse volume representa, aproximadamente, 20% das 1.021 entradas totais, o que assegura um elevado rigor estatístico ao testar a eficácia do modelo em condições extremas de mercado (Hyndman and Athanasopoulos, 2018c).

A normalização das variáveis foi realizada utilizando apenas os parâmetros estimados a partir do conjunto de treinamento, evitando vazamento de informação (*data leakage*) e contribuindo para reduzir o risco de sobreajuste. Essa estratégia garante uma avaliação mais realista do desempenho preditivo fora da amostra e sob condições de mercado estruturalmente distintas.

3.5 Análise e Seleção de Variáveis

A seleção de variáveis foi conduzida com o objetivo de reduzir redundâncias, minimizar multicolinearidade e preservar os preditores mais informativos para os modelos de aprendizado de máquina. O processo seguiu três etapas complementares: (i) análise de correlação, (ii) avaliação estatística de relevância e (iii) verificação de redundância estrutural.

3.5.1 Análise de Correlação

Inicialmente, utilizou-se a correlação de Pearson para avaliar relações lineares entre variáveis explicativas e entre estas e a variável alvo (Benesty et al., 2009). O coeficiente de correlação de Pearson (ρ) entre duas variáveis X e Y é definido como:

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{3.1}$$

onde $\bar{X}$ e $\bar{Y}$ são as médias de X e Y, respectivamente. Valores próximos de ± 1, indicam forte relação linear.

Em geral, pares de variáveis explicativas com correlação superior a um determinado limite (como 0.8 ou 0.9) podem ser reduzidos, mantendo-se apenas um dos preditores no modelo (Guyon and Elisseeff, 2003). Essa estratégia contribui para reduzir a multicolinearidade e melhorar a interpretabilidade dos resultados. Sendo assim, variáveis altamente correlacionadas entre si ($|\rho| > 0{,}8$) foram avaliadas quanto à redundância, mantendo-se apenas aquela com maior relação com a variável alvo ou maior relevância econômica (Guyon and Elisseeff, 2003).

3.5.2 Seleção Estatística de Atributos

Para complementar a análise de correlação, aplicou-se o método *SelectKBest*, que é uma técnica de seleção de variáveis amplamente utilizada na redução da dimensionalidade de conjuntos de dados, especialmente em problemas de aprendizado de máquina supervisionado. Esse método seleciona as k melhores variáveis com base em testes estatísticos específicos, eliminando aquelas que possuem menor relevância para a previsão do alvo (Pedregosa et al., 2011).

A seleção de variáveis desempenha um papel crucial na construção de modelos preditivos eficientes, pois pode melhorar o desempenho do modelo ao reduzir a complexidade computacional e mitigar o risco de *overfitting* (Guyon and Elisseeff, 2003). No caso do *SelectKBest*, o critério de seleção pode ser baseado em diferentes métricas estatísticas, como o teste ANOVA F, que avalia a relação entre variáveis preditoras e a variável resposta contínua, ou o teste do qui-quadrado, utilizado para variáveis categóricas (Duda et al., 2001).

O método opera atribuindo um escore a cada variável preditora de acordo com a métrica escolhida e, em seguida, seleciona as k variáveis com os melhores escores. Dessa forma, garante-se que apenas as características mais relevantes para a tarefa sejam mantidas, melhorando a interpretabilidade do modelo e, em alguns casos, reduzindo o custo computacional (Jovic et al., 2015).

A escolha do valor de k depende do problema em questão e pode ser definida empiricamente por meio de validação cruzada ou outras estratégias de ajuste de hiperparâmetros. Além disso, a aplicação da técnica deve considerar o contexto do conjunto de dados, garantindo que a exclusão de variáveis não prejudique a capacidade preditiva do modelo.

3.5.3 Importância das Variáveis (*Feature Importance*)

Adicionalmente, foi analisada a importância das variáveis estimada por modelos baseados em árvores, como o *Random Forest* (Breiman, 2001a). Essa abordagem fornece uma medida do impacto relativo de cada atributo na redução do erro do modelo, incorporando interações não lineares que não são capturadas por métodos puramente estatísticos.

3.5.4 Análise de Multicolinearidade

Quanto à multicolinearidade, uma das formas mais utilizadas para quantificá-la é o Fator de Inflação da Variância (VIF, do inglês, *Variance Inflation Factor*). O valor mede o grau de colinearidade de uma variável em relação às demais do modelo e é definido como:

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{3.2}$$

onde R_i^2 representa o coeficiente de determinação da regressão auxiliar da variável i sobre as demais variáveis independentes.

Matematicamente, o R^2 quantifica a proporção da variância total da variável dependente que é explicada pelos regressores, sendo calculado pela expressão $R^2 = 1 - \frac{SQR}{SQT}$, em que SQR é a soma dos quadrados dos resíduos e SQT é a soma total dos quadrados (Wooldridge, 2015).

Valores de VIF superiores a 5, e definitivamente acima de 10, indicam uma alta multicolinearidade, o que pode inflacionar a variância dos coeficientes do modelo, comprometendo sua estabilidade e a interpretação das estimativas. O ideal é que o VIF permaneça abaixo de 5, embora valores até 10 sejam frequentemente considerados aceitáveis, dependendo do campo de estudo (Marquardt, 1970). Nesses casos, pode ser necessário remover ou transformar variáveis para melhorar a estabilidade do modelo.

3.5.5 Engenharia de Atributos

Por fim, foram criadas variáveis derivadas a partir das séries originais, incluindo médias móveis, com o objetivo de suavizar flutuações de curto prazo e evidenciar tendências temporais. Essa transformação auxilia os modelos a capturar padrões sazonais e reduzir o impacto de ruídos pontuais, característica particularmente relevante em séries econômicas voláteis (Gupta and Zheng, 2020). Essa abordagem também pode reduzir a necessidade de transformações avançadas, facilitando a interpretação dos resultados.

3.6 Modelos de Aprendizado de Máquina

Após as etapas de pré-processamento, seleção de variáveis e divisão temporal da base de dados, foram aplicados dois modelos de aprendizado de máquina amplamente utilizados em previsão de séries temporais: *Random Forest* (RF) e *Long Short-Term Memory* (LSTM). A escolha desses algoritmos buscou contemplar abordagens distintas e complementares para modelagem de relações não lineares e dependências temporais.

O Random Forest é um método do tipo *ensemble* baseado na combinação de múltiplas árvores de decisão (Breiman, 2001a). Cada árvore é construída a partir de subconjuntos aleatórios dos dados e das variáveis, e a predição final resulta da média das previsões individuais. Essa estratégia reduz a variância do modelo, melhora a generalização e confere robustez frente a ruídos e relações não lineares. Além disso, o RF permite estimar a importância relativa das variáveis, contribuindo para a interpretabilidade dos resultados.

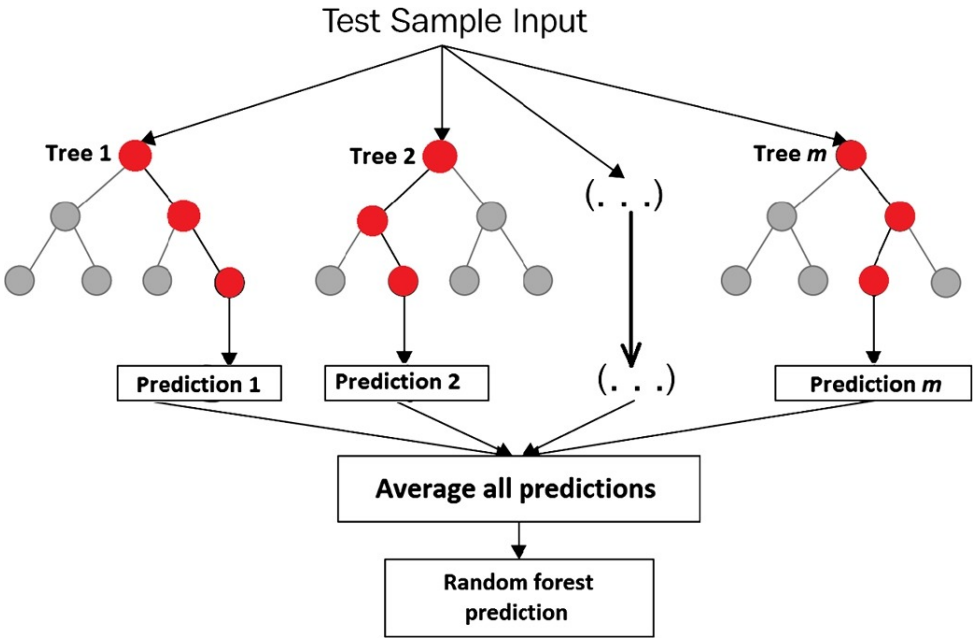
A rede neural LSTM é uma arquitetura de redes neurais recorrentes projetada para capturar dependências de longo prazo em dados sequenciais (Hochreiter and Schmidhuber, 1997). Diferentemente de redes neurais tradicionais, a LSTM incorpora mecanismos de memória interna, compostos por células de memória e portões de entrada, esquecimento e saída, que controlam o fluxo de informação ao longo do tempo. Essa estrutura permite modelar padrões temporais complexos e é particularmente adequada para séries econômicas sujeitas a sazonalidades, tendências e choques externos, como o mercado de combustíveis.

A utilização desses modelos possibilita comparar o desempenho de uma abordagem baseada em métodos ensemble e outra fundamentada em aprendizado profundo, avaliando suas capacidades de generalização em diferentes regimes econômicos.

3.6.1 Random Forest

O *Random Forest* (RF) é um algoritmo de aprendizado de máquina baseado em um conjunto de múltiplas árvores de decisão, que são construídas de forma independente a partir de diferentes amostras do conjunto de treinamento (*bootstrap sampling*) e subconjuntos aleatórios das variáveis preditoras (ver Figura 3.1). Essa estratégia de aleatoriedade introduz diversidade entre as árvores, reduzindo a correlação entre elas e mitigando o risco de *overfitting*, uma limitação comum em árvores de decisão individuais (Hastie et al., 2009).

Figura 3.1: Representação gráfica da arquitetura do algoritmo *Random Forest*.



Fonte: Afzal et al. (2020).

A construção de uma árvore de decisão dentro do RF ocorre por meio da divisão recursiva do espaço de atributos em subconjuntos cada vez menores, conforme um critério de impureza (Quinlan, 1986). Para problemas de classificação, esse critério pode ser a entropia ou o índice de Gini, enquanto para regressão, a métrica mais utilizada é o erro quadrático médio (MSE, do inglês *Mean Squared Error*). Formalmente, o índice de Gini para um nó t com K classes é definido como:

$$Gini(t) = 1 - \sum_{k=1}^{K} p_k^2,$$

onde p_k representa a proporção de amostras pertencentes à classe k. Quanto menor o valor de $Gini(t)$, mais homogêneo é o nó resultante da divisão. No caso da regressão, que é o caso do presente estudo, o critério é a minimização do MSE, dado por:

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

em que y_i são os valores reais da variável resposta e $\hat{y}$ representa a média das previsões.

Após a construção das árvores individuais, o RF realiza a agregação das previsões. Para problemas de classificação, a decisão final é baseada no princípio da maioria dos votos (*majority voting*), enquanto para regressão, a saída é a média das previsões individuais:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x),$$

onde T é o número total de árvores na floresta e $h_t(x)$ representa a previsão da t-ésima árvore.

Além de sua robustez contra *overfitting*, o RF apresenta diversas vantagens. Ele pode lidar com variáveis categóricas e contínuas sem necessidade de transformações complexas, é menos sensível a *outliers* do que modelos individuais de árvores e possui um mecanismo interno de validação por meio do erro fora da amostra (*out-of-bag error, OOB*), permitindo estimar a performance do modelo sem a necessidade de um conjunto de validação separado (Liaw and Wiener, 2002).

O desempenho do RF é altamente influenciado por hiperparâmetros essenciais. O número de árvores na floresta (*ntree*) controla a estabilidade das previsões; usualmente, valores acima de 500 garantem uma boa convergência. O número de variáveis consideradas em cada divisão (*mtry*) afeta a diversidade das árvores; recomenda-se um valor igual à raiz quadrada do número total de variáveis para classificação e um terço para regressão (Carvalho Junior, 2016). Por fim, o tamanho mínimo dos nós terminais (*nodesize*) regula a profundidade das árvores, influenciando o viés e a variância do modelo.

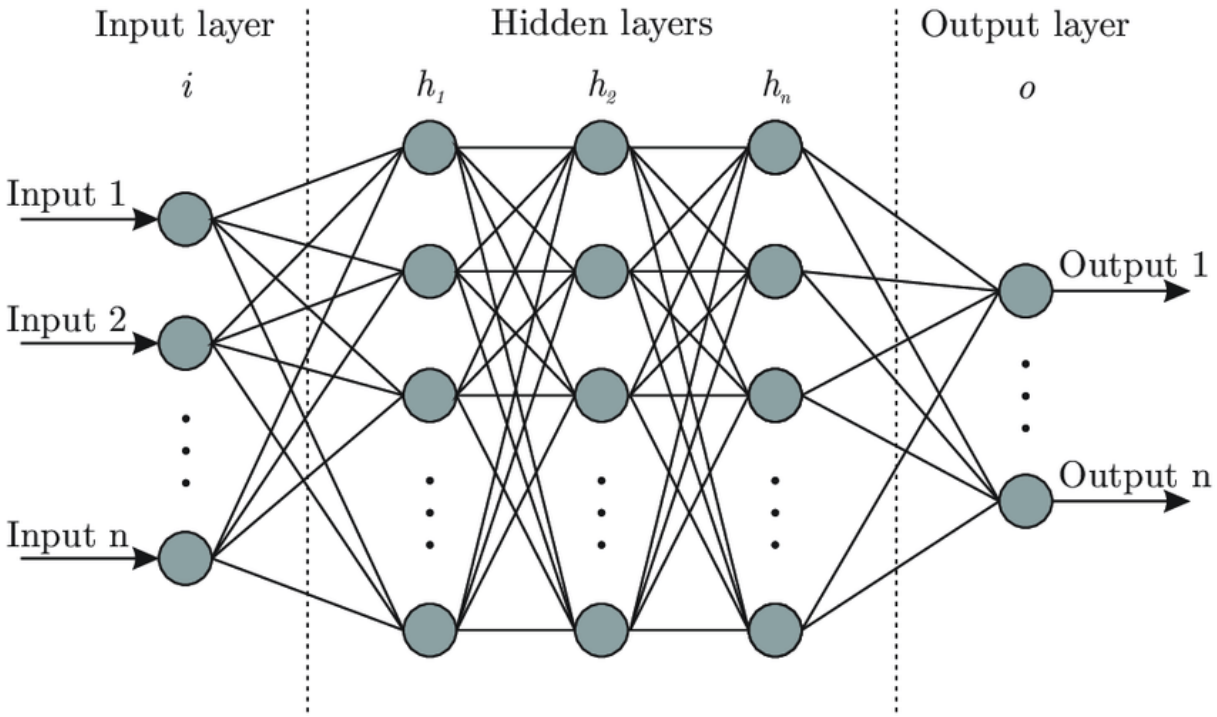
Em síntese, o RF se destaca como uma abordagem versátil e eficiente para modelagem preditiva, combinando a interpretabilidade das árvores de decisão individuais com a robustez de um conjunto de modelos, tornando-se uma das técnicas mais populares em aprendizado de máquina.

3.6.2 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são modelos computacionais inspirados na estrutura do cérebro humano, projetados para reconhecer padrões e aprender a partir de dados. Elas são compostas por unidades chamadas neurônios, organizados em camadas (Kovács, 2006). As três camadas principais em uma RNA são:

- **Camada de Entrada:** Recebe os dados de entrada e os passa para a próxima camada.
- **Camadas Ocultas:** Executam a transformação não linear dos dados recebidos, extraindo características relevantes.
- **Camada de Saída:** Produz a saída final, que pode ser uma classe para problemas de classificação ou um valor contínuo para problemas de regressão (como fim ilustrativo, ver Figura 3.2).

Figura 3.2: Exemplo de Rede Neural Artificial.



Fonte: Bre et al. (2017).

Cada neurônio de uma camada está conectado a todos os neurônios da camada anterior e da próxima, formando uma rede densamente conectada. As conexões entre neurônios possuem pesos que determinam a influência de uma entrada sobre a saída do neurônio. Durante o treinamento, o objetivo é ajustar esses pesos para minimizar o erro de previsão (Santos et al., 2005).

A operação básica de um neurônio pode ser descrita pela seguinte equação:

$$z_j = \sum_{i=1}^{n} w_{ij} x_i + b_j, \tag{3.3}$$

onde x_i é a entrada do neurônio, w_{ij} é o peso da conexão entre os neurônios i e j, b_j é um peso entre um valor constante e o neurônio k conhecido como *bias*, e z_j é o valor ponderado da soma das entradas (Fleck et al., 2016).

Após calcular z_j, aplica-se uma função de ativação para introduzir não linearidade, permitindo que a rede aprenda e represente padrões complexos:

$$a_j = \sigma(z_j). \tag{3.4}$$

Exemplos comuns de funções de ativação incluem:

- Sigmoide: $\sigma(z) = \dfrac{1}{1 + e^{-z}}$;
- Tangente hiperbólica (tanh): $\sigma(z) = \dfrac{e^z - e^{-z}}{e^z + e^{-z}}$;
- ReLU (sigla, do inglês, *Rectified Linear Unit*): $\sigma(z) = \max(0, z)$.

O treinamento das RNAs é realizado através de algoritmos de otimização, como o *backpropagation*, que ajusta os pesos utilizando o gradiente do erro em relação aos pesos, e métodos como o Gradiente Descendente para atualizar os pesos com base na taxa de aprendizado.

O termo *Deep Learning* (DL) refere-se a RNAs com múltiplas camadas ocultas, capazes de aprender representações hierárquicas dos dados (Min et al., 2017). As redes neurais profundas são treinadas ajustando os pesos das conexões entre neurônios para minimizar uma função de perda, que mede a diferença entre as previsões da rede e os valores reais (Reis, 2021). O processo de treinamento envolve otimização de milhões de parâmetros, o que requer grandes volumes de dados e poder computacional significativo. Avanços tecnológicos, como o uso de GPUs e *frameworks* como *TensorFlow* e *Keras*, facilitaram a implementação e treinamento de modelos de DL, possibilitando seu uso em larga escala (Grattarola and Alippi, 2021). Redes profundas têm demonstrado alto desempenho em problemas complexos devido à sua capacidade de modelar relações altamente não lineares.

3.6.3 Aprendizado profundo (Deep Learning)

O *Deep Learning*, ou aprendizado profundo, é uma subárea da aprendizagem de máquina que utiliza redes neurais profundas para modelar representações complexas dos dados. Para Min et al. (2017), essas redes são compostas por múltiplas camadas ocultas, o que lhes permite aprender características hierárquicas, começando por padrões simples nas primeiras camadas e avançando para representações mais abstratas em camadas posteriores. Diferentemente das redes neurais tradicionais, que geralmente possuem uma ou poucas camadas ocultas, as redes profundas podem capturar relações mais complexas.

As redes neurais profundas são treinadas ajustando os pesos das conexões entre neurônios para minimizar uma função de perda, que mede a diferença entre as previsões da rede e os valores reais (Reis, 2021). O processo de treinamento envolve otimização de milhões de parâmetros, o que requer grandes volumes de dados e poder computacional significativo. Avanços tecnológicos, como o uso de GPUs e *frameworks* como *TensorFlow* e *Keras*, facilitaram a implementação e treinamento de modelos de *deep learning*, possibilitando seu uso em larga escala (Grattarola and Alippi, 2021).

Entre os modelos mais conhecidos de redes profundas estão as Redes Neurais Convolucionais (CNNs) e as Redes Neurais Recorrentes (RNNs). As CNNs são amplamente utilizadas em tarefas de visão computacional, como reconhecimento de imagens e detecção de objetos. Elas se destacam por sua capacidade de extrair características locais por meio de camadas convolucionais, que aplicam filtros sobre as entradas, e camadas de *pooling*, que reduzem a dimensionalidade dos dados, preservando as informações mais relevantes. Por exemplo, em um sistema de reconhecimento de dígitos manuscritos, as primeiras camadas da CNN podem identificar bordas e texturas, enquanto as camadas mais profundas combinam essas características para reconhecer o número como um todo (O'Shea and Nash, 2015).

As RNNs, por outro lado, são projetadas para lidar com dados sequenciais, como séries temporais e texto (Medsker and Jain, 1999). Elas possuem conexões que permitem que informações de estados anteriores sejam mantidas e utilizadas para processar novos dados, tornando-as ideais para capturar dependências temporais. No entanto, as RNNs tradicionais enfrentam dificuldades em capturar dependências de longo prazo devido ao problema do gradiente em desaparecimento. Além disso, Sutskever et al. (2011) também ressalta que

> Embora os gradientes da RNN sejam fáceis de calcular, as RNNs são fundamentalmente difíceis de treinar, especialmente em problemas com dependências temporais de longo prazo, devido à sua natureza iterativa não linear. Uma pequena mudança em um processo iterativo pode agravar-se e resultar em efeitos muito grandes muitas iterações depois; isso é conhecido coloquialmente como "efeito borboleta". A implicação é que em uma RNN, a derivada da função de perda em um determinado momento pode ser exponencialmente grande em relação às ativações ocultas em um momento muito anterior. Assim, a função de perda é muito sensível a pequenas mudanças, tornando-se efetivamente descontínua.

Para contornar essa limitação, de acordo com Shiri et al. (2023), variantes como as *Gated Recurrent Units* (GRUs) e as *Long Short-Term Memory* (LSTM) foram desenvolvidas, oferecendo melhor desempenho em tarefas como tradução automática e previsão de séries temporais.

3.6.4 Redes Neurais Recorrentes

As Redes Neurais Recorrentes (RNNs) são uma classe de redes neurais artificiais projetadas para lidar com dados sequenciais, como séries temporais, processamento de linguagem natural e reconhecimento de fala (Rumelhart et al., 1986). A principal característica das RNNs é a presença de conexões recorrentes, que permitem que a informação persista ao longo do tempo, possibilitando a modelagem de dependências temporais (Hochreiter and Schmidhuber, 1997).

A estrutura de uma RNN é similar a uma rede neural artificial tradicional, mas cada neurônio na camada oculta tem uma conexão adicional para si mesmo, permitindo que a saída do neurônio em um determinado tempo seja influenciada pela sua saída anterior. Essa propriedade torna as RNNs mais adequadas para modelar sequências temporais do que redes neurais tradicionais, que tratam cada entrada de maneira independente. A equação que governa a dinâmica de uma RNN pode ser expressa como:

$$h_t = \sigma(W_h x_t + U_h h_{t-1} + b_h), \tag{3.5}$$

onde x_t é a entrada no tempo t, h_t é o estado oculto no tempo t, h_{t-1} é o estado oculto no tempo anterior, W_h e U_h são matrizes de pesos treináveis, b_h é o viés, e σ é a função de ativação, geralmente a função tangente hiperbólica (tanh) ou a função ReLU.

As RNNs têm sido amplamente aplicadas em tarefas como tradução automática, modelagem de linguagem, previsão de séries temporais e classificação de sequências (A. Graves and Hinton, 2013). Apesar de seu potencial, as RNNs tradicionais enfrentam desafios significativos, como o problema do gradiente desaparecendo ou explodindo, tornando difícil o aprendizado de dependências de longo prazo (Y. Bengio, 1994). Para contornar essas limitações, arquiteturas avançadas foram desenvolvidas, como as redes do tipo *Long Short-Term Memory* (LSTM) e *Gated Recurrent Units* (GRU), que incorporam mecanismos especializados para preservar informações relevantes ao longo do tempo.

3.6.5 Long Short-Term Memory (LSTM)

As redes LSTM foram propostas para superar as limitações das RNNs tradicionais (Hochreiter and Schmidhuber, 1997). A LSTM introduz uma estrutura mais complexa, composta por células de memória que regulam o fluxo de informação ao longo do tempo por meio de três portas principais:

- **Porta de Entrada (i_t)**: Controla quais informações da entrada atual devem ser armazenadas na célula de memória.
- **Porta de Esquecimento (f_t)**: Determina quais informações antigas devem ser descartadas da célula de memória.
- **Porta de Saída (o_t)**: Decide quais informações da célula de memória serão propagadas para a próxima etapa.

As equações que regem essas portas são definidas como:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{3.6}$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{3.7}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{3.8}$$

onde W_i, W_f, W_o são as matrizes de pesos, b_i, b_f, b_o são os vieses e σ representa a função sigmoide logística.

Além dessas portas, a LSTM possui um estado de célula c_t, atualizado da seguinte forma:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{3.9}$$

onde $\tilde{c}_t$ é uma atualização candidata para o estado da célula.

O treinamento das redes LSTM ocorre por meio do algoritmo de *Backpropagation Through Time* (BPTT) (Werbos, 1990). Esse método estende o algoritmo tradicional de

retropropagação (*backpropagation*) para redes recorrentes, computando os gradientes ao longo de toda a sequência temporal.

O processo de treinamento segue as seguintes etapas:

1. **Forward Pass**: A entrada sequencial é propagada através das camadas LSTM, atualizando os estados ocultos h_t e os estados da célula c_t.
2. **Cálculo da Função de Perda**: A saída do modelo é comparada com os valores reais, e a função de erro é calculada, geralmente utilizando *Mean Squared Error* (MSE) para regressão ou *Cross-Entropy Loss* para classificação (Lipton et al., 2015).
3. **Backward Pass (BPTT)**: O erro é propagado para trás ao longo do tempo, ajustando os pesos da rede com base no gradiente da função de perda.
4. **Atualização dos Pesos**: Os gradientes calculados são utilizados para atualizar os pesos da rede por meio de algoritmos de otimização, como *Adam* e *RMSprop* (Kingma and Ba, 2014).

Para evitar o alto custo computacional, pode-se aplicar uma versão truncada do BPTT, onde a retropropagação ocorre apenas por um número fixo de passos temporais. Além disso, arquiteturas como GRUs (Cho et al., 2014) e *Transformers* (Vaswani et al., 2017) foram propostas para lidar com as limitações de eficiência das LSTMs.

3.7 Métricas de Avaliação

As métricas de validação desempenham um papel crucial na avaliação do desempenho dos modelos de aprendizagem de máquina, permitindo quantificar a discrepância entre as previsões e os valores reais (Bishop, 2006; Willmott and Matsuura, 2005). Para quantificar o desempenho dos modelos e validar os resultados obtidos nas fases de teste e validação (períodos pandêmico e pós-pandêmico), foram empregadas as métricas de erro apresentadas e fundamentadas no capítulo de revisão bibliográfica: *Mean Absolute Percentage Error* (MAPE), *Mean Absolute Error* (MAE) e *Mean Squared Error* (MSE). A utilização conjunta destas métricas permite uma avaliação multidimensional, considerando tanto a magnitude absoluta dos erros quanto a sensibilidade a *outliers* e a precisão percentual das previsões.

O MAPE é amplamente empregado para medir o erro percentual médio das previsões em relação aos valores reais, sendo definido como:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100, \tag{3.10}$$

onde y_i representa o valor observado, $\hat{y}_i$ a previsão do modelo, e n o número total de observações. Essa métrica é útil para comparar a precisão entre diferentes modelos, especialmente quando se deseja interpretar o erro em termos relativos. No entanto, sua principal limitação é a sensibilidade a valores próximos de zero, o que pode distorcer os resultados (Armstrong and Collopy, 1985).

O MAE, por sua vez, expressa a média das diferenças absolutas entre as previsões e os valores reais:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\,. \tag{3.11}$$

Essa métrica é útil quando se deseja medir a magnitude dos erros sem penalizar discrepâncias extremas, tornando-se uma escolha apropriada em cenários onde os erros precisam ser minimizados de forma uniforme. Diferentemente do MSE, o MAE não enfatiza *outliers*, o que pode ser vantajoso dependendo da natureza dos dados (Chai and Draxler, 2014).

Por fim, o MSE é definido como:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{3.12}$$

O uso do quadrado da diferença faz com que erros maiores tenham um impacto desproporcional na métrica, o que pode ser benéfico quando se deseja penalizar previsões muito distantes dos valores reais. Entretanto, essa característica também pode torná-lo inadequado em situações onde *outliers* são comuns e podem distorcer a avaliação do modelo (Hyndman and Koehler, 2006).

A escolha da métrica ideal depende do contexto e do objetivo da análise. Modelos voltados para previsões financeiras, por exemplo, frequentemente utilizam o MAPE devido à sua interpretação percentual, enquanto modelos de séries temporais podem preferir o MSE quando grandes desvios precisam ser mais rigorosamente controlados. Assim, a compreensão das diferenças entre essas métricas é essencial para a seleção adequada da abordagem de validação de um modelo.

3.8 Desenvolvimento de Dashboard Interativo

Como etapa final da metodologia, foi desenvolvido um *dashboard* interativo com o objetivo de organizar e disponibilizar de forma visual os resultados obtidos pelos modelos preditivos. A ferramenta foi implementada na plataforma *Microsoft Power BI* (versão 2.150.1704.0), escolhida pela capacidade de integração de dados, recursos de visualização dinâmica e facilidade de compartilhamento.

O painel foi estruturado para permitir a exploração interativa das séries temporais analisadas e das previsões geradas pelo modelo LSTM. O *dashboard* é composto por duas páginas principais:

- **Página 1:** Comparação entre valores observados e previstos do preço do etanol ao longo do tempo, por meio de gráficos de área e tabela associada, além de filtros temporais que permitem selecionar intervalos específicos e os períodos pré-pandemia, pandemia e pós-pandemia.
- **Página 2:** Visualização das variáveis preditoras e da variável alvo, possibilitando análise conjunta da evolução das séries e comparação entre diferentes períodos selecionados pelo usuário.

O desenvolvimento dessa ferramenta complementa a modelagem ao fornecer uma interface de apoio à análise dos resultados, permitindo que os padrões identificados e o desempenho dos modelos sejam examinados de maneira visual e acessível.

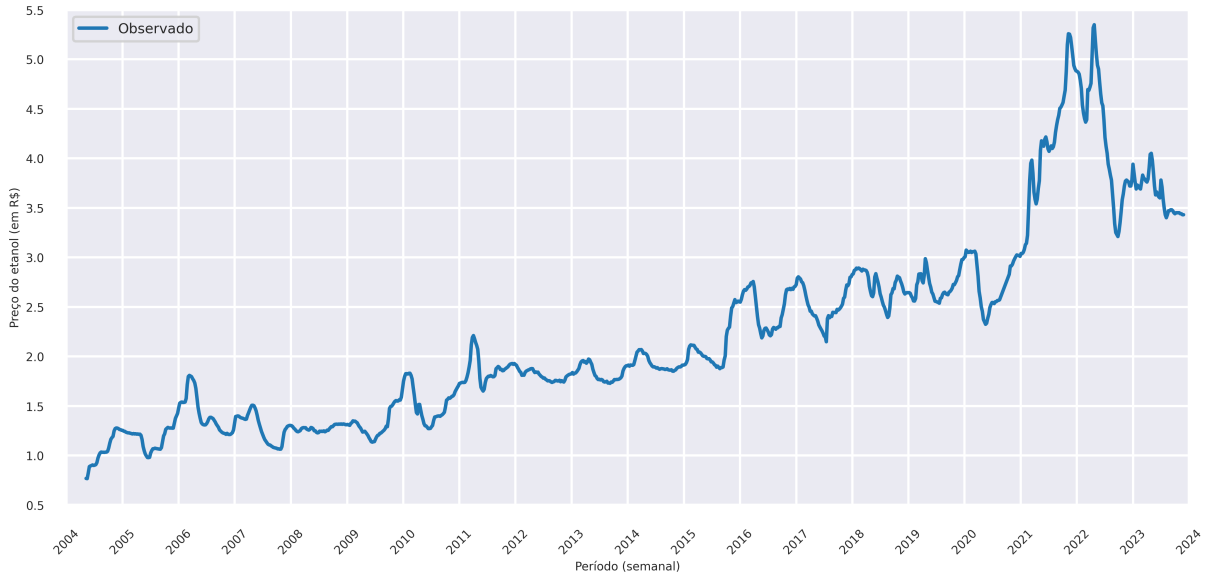
Capítulo 4

Resultados e Discussão

4.1 Análise Exploratória

A análise exploratória da série semanal de preços do etanol (2004–2023), que é a variável *target* do presente trabalho, evidencia um comportamento não estacionário, caracterizado por tendência de longo prazo, ciclos de volatilidade e quebras estruturais associadas a eventos macroeconômicos (Figura 4.1). Observa-se elevada variabilidade entre 2006 e 2011, seguida por um período de relativa estabilidade até 2014. A partir de 2015 inicia-se uma trajetória de crescimento, culminando no pico histórico entre 2021 e 2022, associado aos impactos da pandemia de COVID-19, desorganização das cadeias produtivas e choques energéticos globais.

Figura 4.1: Preço de revenda do etanol no estado de São Paulo (2004-2023).

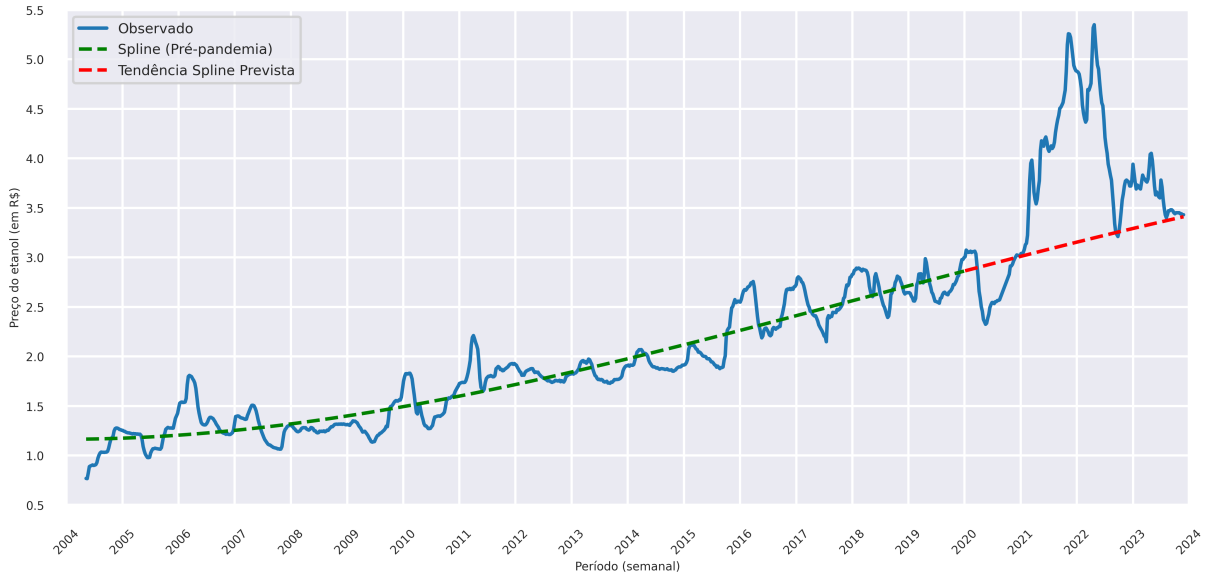


Fonte: Elaboração própria.

A modelagem spline aplicada ao período pré-pandêmico (Figura 4.2) evidencia uma tendência suavizada que descreve o comportamento estrutural do mercado em condições

consideradas normais. período pré-pandêmico oferece uma base mais estável para a modelagem, pois não sofreu as anomalias e intervenções atípicas observadas nos mercados durante e após a pandemia. Ao isolar e analisar essa fase, foi possível estabelecer uma marca de referência para comparar com o comportamento dos preços em cenários de maior instabilidade observados durante a pandemia. A análise permite, ainda, verificar se o período pós-pandêmico está retornando aos padrões anteriores ou se estabeleceu uma nova dinâmica no mercado de etanol, impactando as estratégias de previsão de preços.

Figura 4.2: Tendência Spline a partir dos dados pré-pandêmicos.

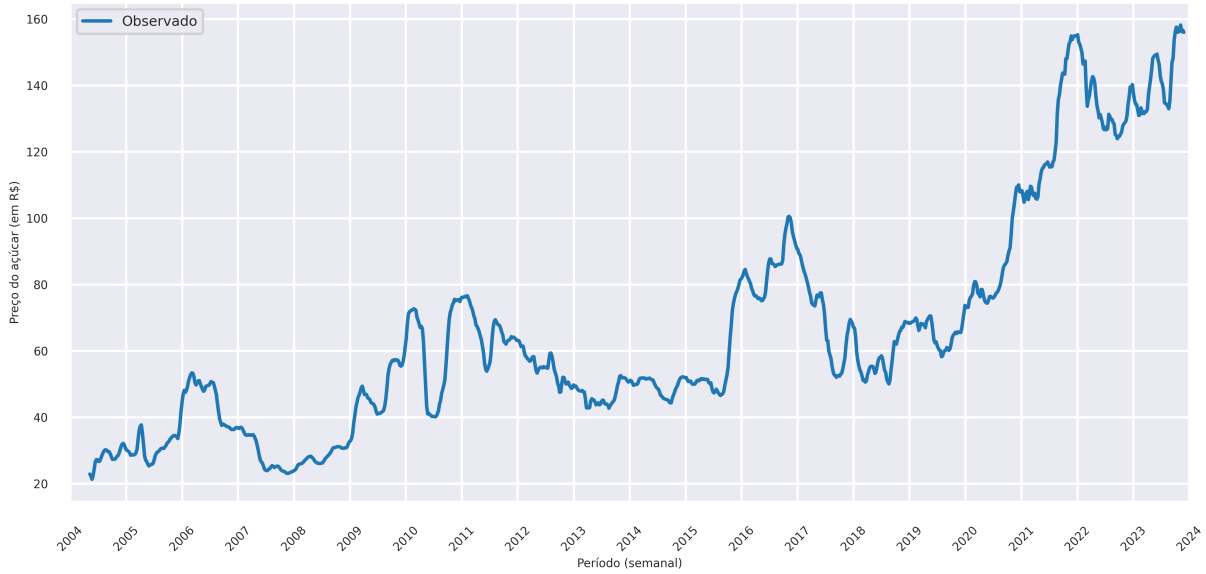


Fonte: Elaboração própria.

As variáveis explicativas (preditoras) exibem comportamentos coerentes com a dinâmica do setor sucroenergético. O preço do açúcar (Figura 4.3) apresenta trajetória semelhante à do etanol, com ciclos de alta coincidindo com períodos de valorização das commodities agrícolas, reforçando o caráter concorrencial entre a destinação da cana para açúcar ou etanol. Entre 2018 e 2020, houve relativa estabilidade, mas a partir de 2021 os preços atingiram valores recordes, reflexo da pandemia e de dinâmicas globais de mercado.

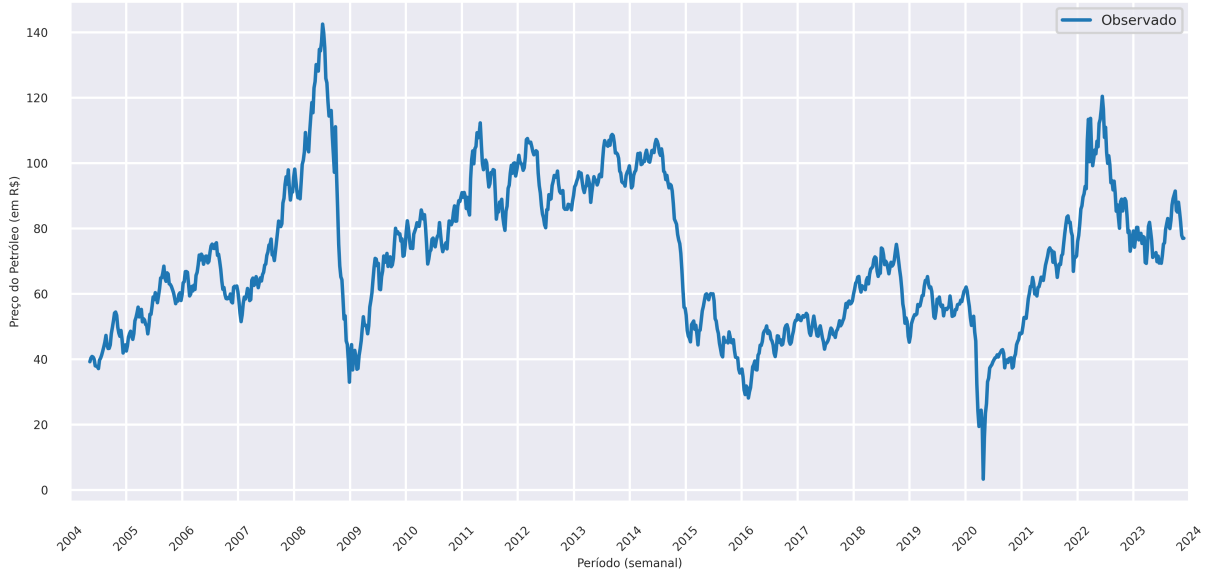
O petróleo (Figura 4.4) apresentou fortes oscilações ao longo do tempo, revelando forte influência de crises globais (2008, 2020) e tensões geopolíticas (2022), fatores que impactam indiretamente o etanol via paridade energética. Além disso, entre 2017 e 2019, os preços mantiveram-se relativamente estáveis, mas a pandemia de COVID-19 resultou em uma forte retração em 2020. A partir de 2021, os preços voltaram a subir, atingindo novo pico em 2022, refletindo fatores como tensões geopolíticas e pressões no mercado energético.

Figura 4.3: Preço do açúcar (2004-2023).



Fonte: Elaboração própria.

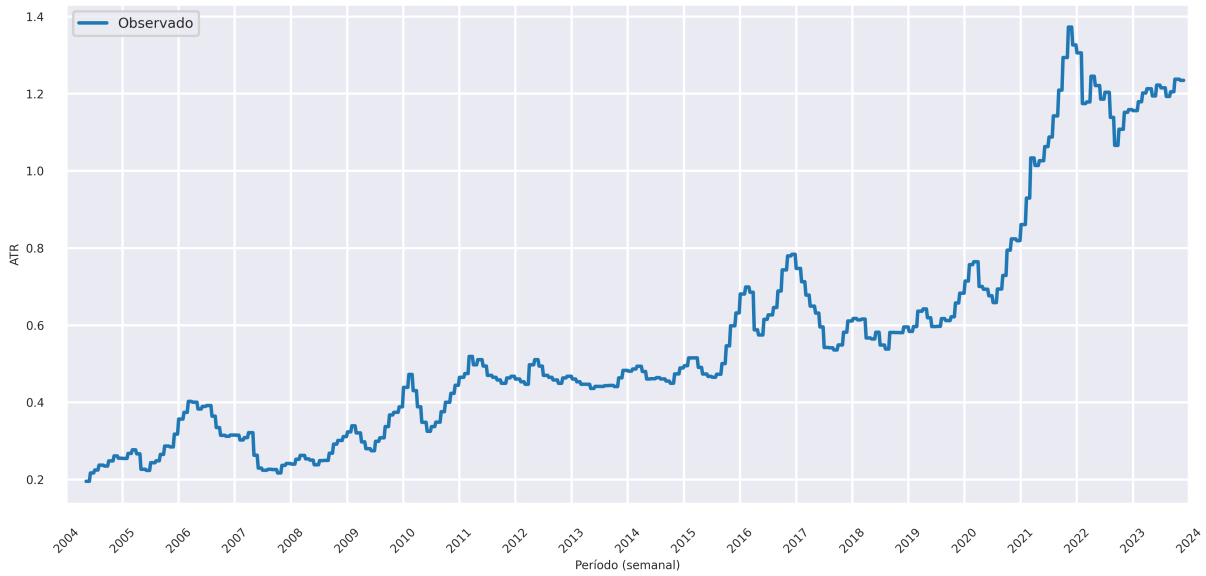
Figura 4.4: Preço do petróleo (2004-2023).



Fonte: Elaboração própria.

Já o ATR (Figura 4.5) mostra tendência de crescimento estrutural, evoluindo de R$ 0,20 em 2004 para valores próximos de R$ 1,40 em 2021, refletindo a valorização dos derivados da cana como açúcar e etanol, impulsionada por demanda crescente, valorização do dólar e incentivos ao setor. Além disso, o aumento dos custos de produção e a inflação contribuíram para esse crescimento.

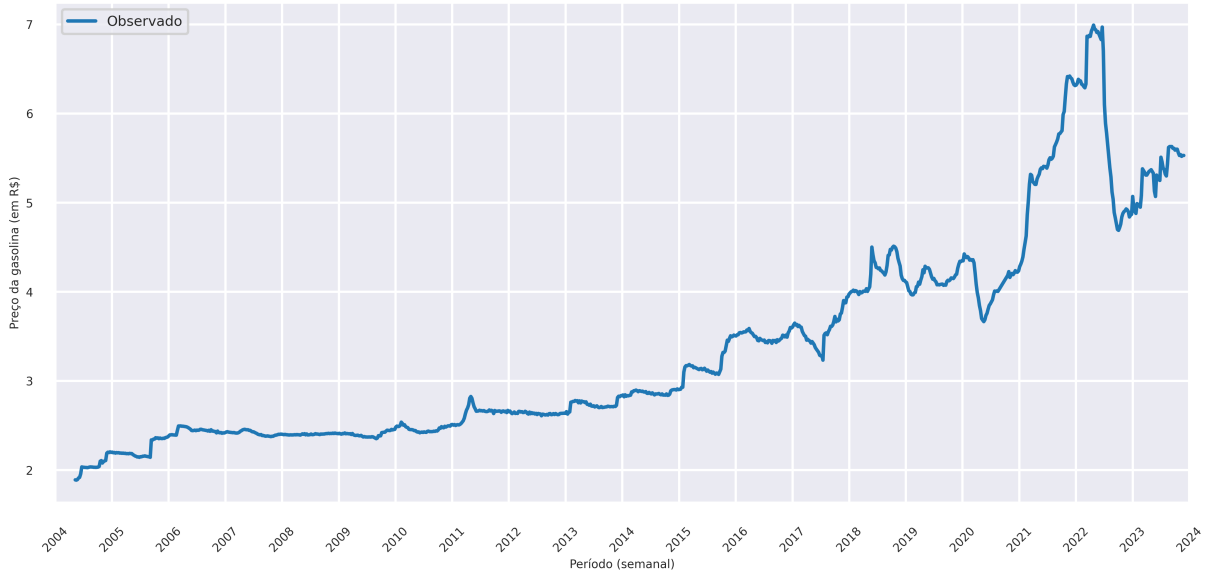
Figura 4.5: ATR (2004-2023).



Fonte: Elaboração própria.

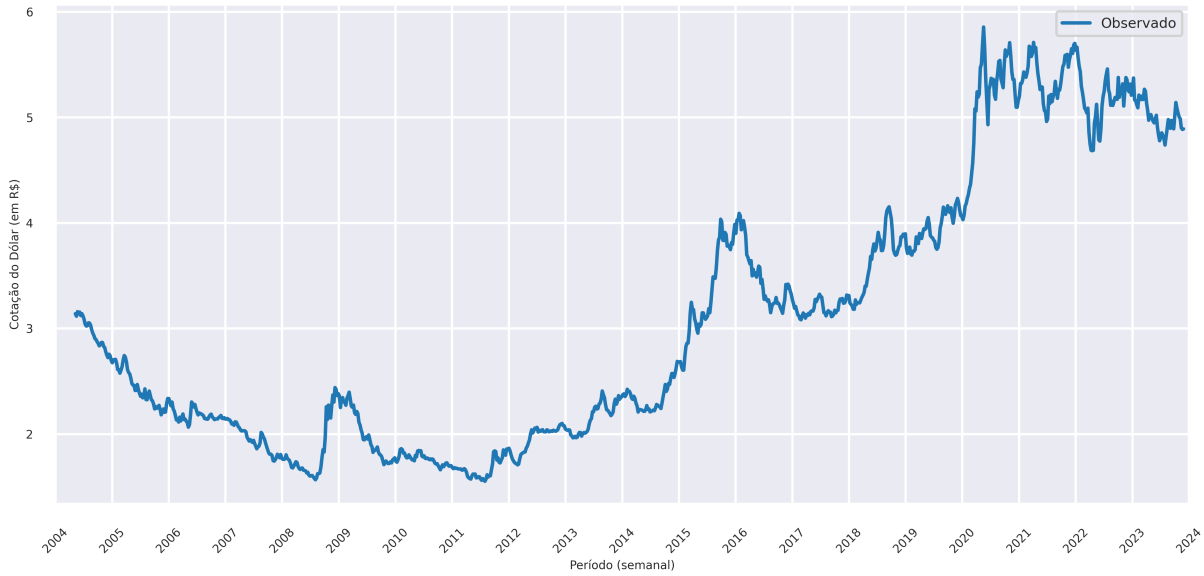
Variáveis de mercado interno também apresentam relevância. O preço da gasolina (Figura 4.6) apresenta trajetória semelhante à do etanol no período pós-2020, influenciada por fatores comuns ao mercado energético, como valorização cambial, choques de oferta e instabilidade global. Essa dinâmica reforça a interdependência entre os combustíveis e o efeito de substituição na demanda. Já a taxa de câmbio do dólar (Figura 4.7) demonstra picos de valorização em períodos de instabilidade econômica, influenciando exportações de açúcar e etanol e, consequentemente, os preços domésticos.

Figura 4.6: Preço da gasolina (2004-2023).



Fonte: Elaboração própria.

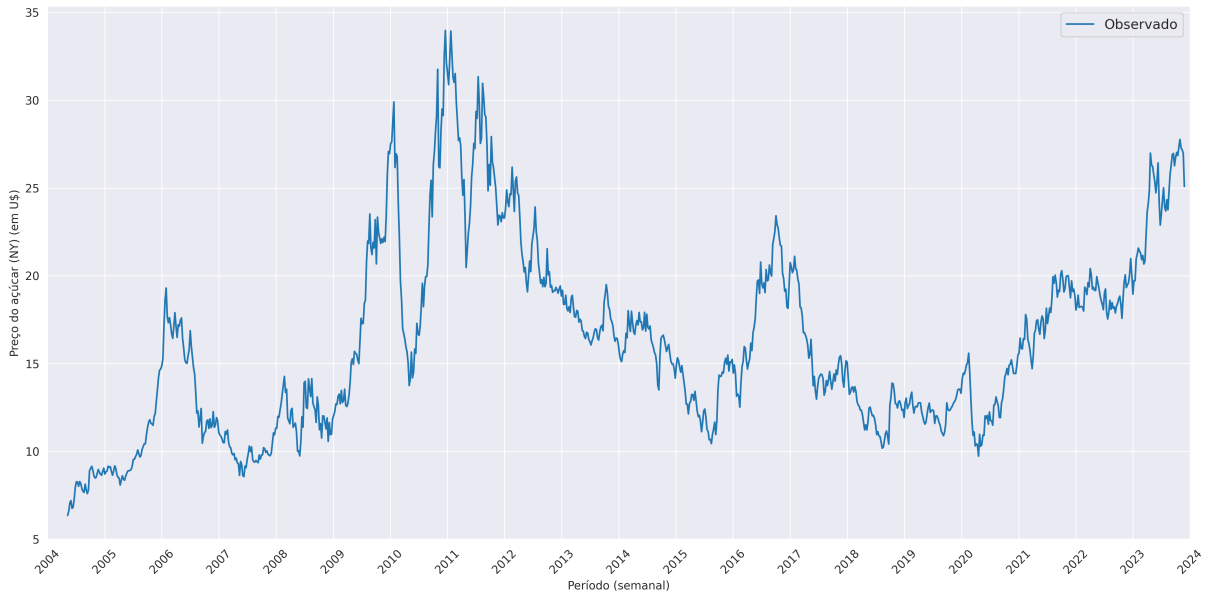
Figura 4.7: Cotação do dólar (2004-2023).



Fonte: Elaboração própria.

O açúcar negociado em NY (Figura 4.8) reforça a conexão do mercado brasileiro com o cenário internacional. A trajetória apresentou uma tendência geral de alta ao longo do período analisado, embora com momentos de maior instabilidade e picos de preço. A partir de 2011, observou-se um aumento expressivo nos preços, impulsionado por fatores como a crise econômica global, a valorização do dólar e a crescente demanda internacional pelo açúcar, atingindo um pico de US$ 35,00. Entre 2017 e 2020, a volatilidade foi mais pronunciada, com alternâncias entre altas e baixas, iniciando, a partir de 2021, um novo ciclo de alta, influenciado pela pandemia de COVID-19, pela crise energética global, pelo conflito entre Ucrânia e Rússia e pela valorização do dólar.

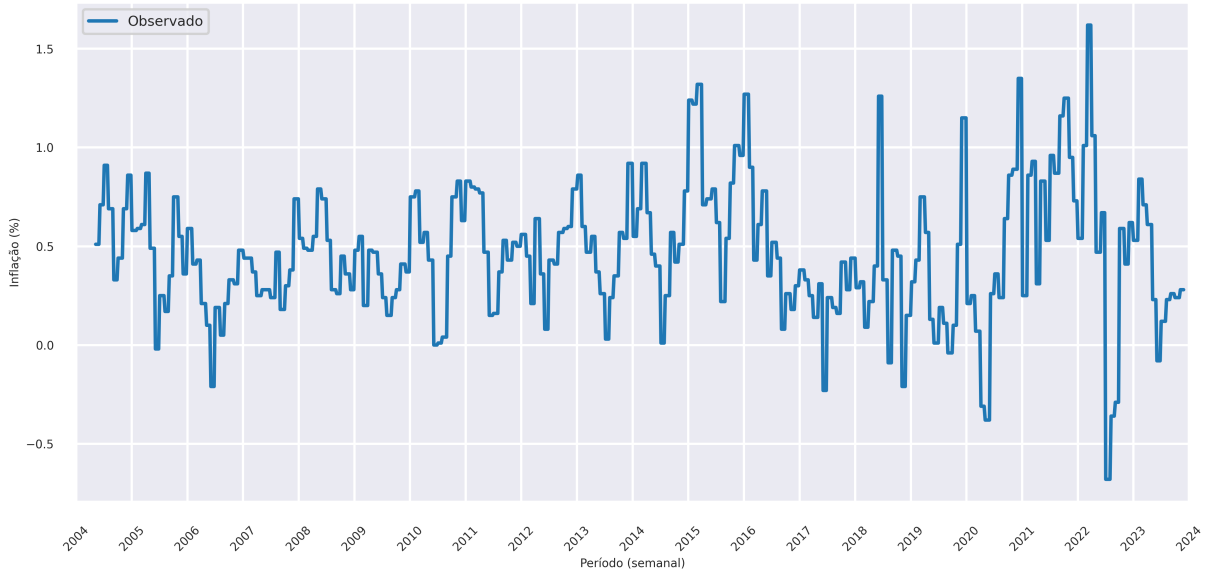
Figura 4.8: Açúcar NY (em US$) no período de 2004 a 2023.



Fonte: Elaboração própria.

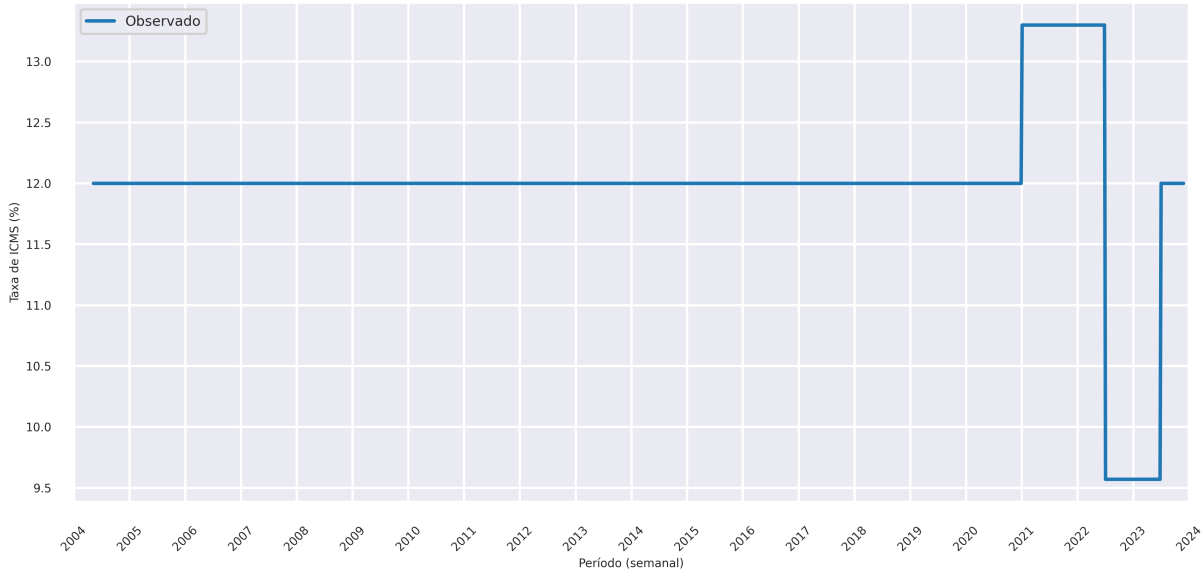
A inflação (Figura 4.9) e a variação do ICMS (Figura 4.10) representam componentes de custo e política econômica que também contribuem para explicar a formação de preços. A partir de 2011 houve um período de alta da inflação, seguido de uma desaceleração gradual entre 2017 e 2020. A partir de 2021, em decorrência da pandemia de COVID-19 e de outros fatores econômicos e políticos, a inflação voltou a subir. Já sobre o ICMS, destaca-se o aumento em 2021, seguido de uma baixa em meados de 2022.

Figura 4.9: Inflação (%) no período de 2004 a 2023.



Fonte: Elaboração própria.

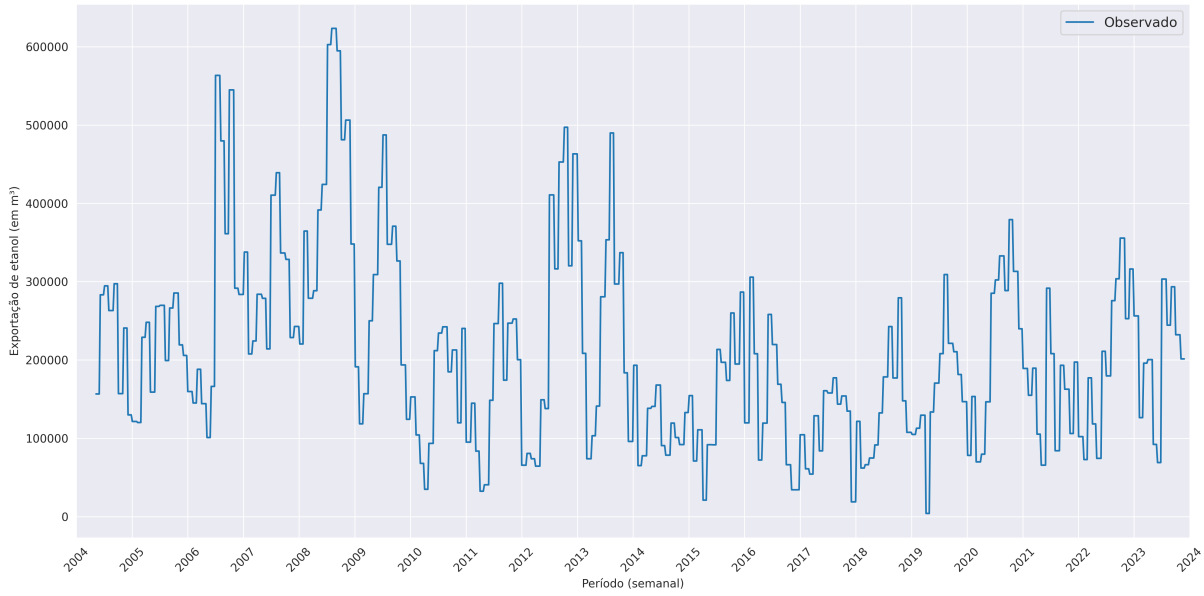
Figura 4.10: ICMS (%) no período de 2004 a 2023.



Fonte: Elaboração própria.

As exportações de etanol (Figura 4.11) evidenciam a importância do Brasil no mercado mundial, apresentando volumes que, apesar das variações de safra, mantêm patamares elevados ao longo do tempo. Os dados mostram que o envio do produto para o exterior é muito influenciado pelos preços internacionais e pela valorização do dólar. Em diversos momentos, o volume exportado ultrapassou a marca de 600.000 m^3.

Figura 4.11: Exportação de etanol (m^3) no período de 2004 a 2023.

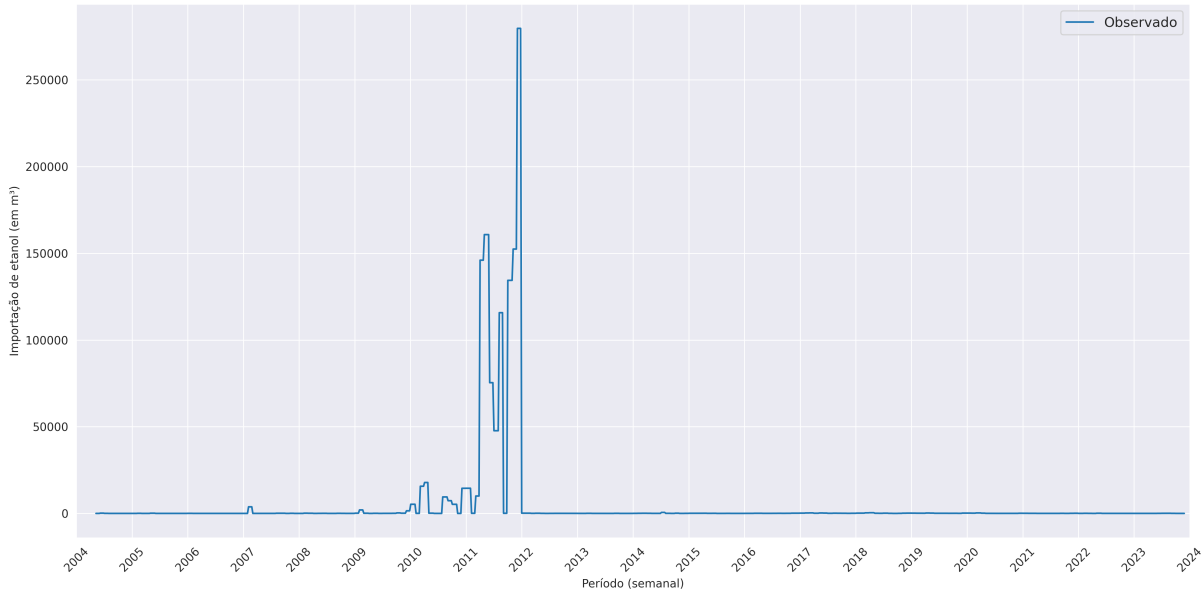


Fonte: Elaboração própria.

Em relação às importações de etanol (Figura 4.12), os volumes costumam ser baixos na maior parte do período, funcionando apenas como um ajuste para o abastecimento na-

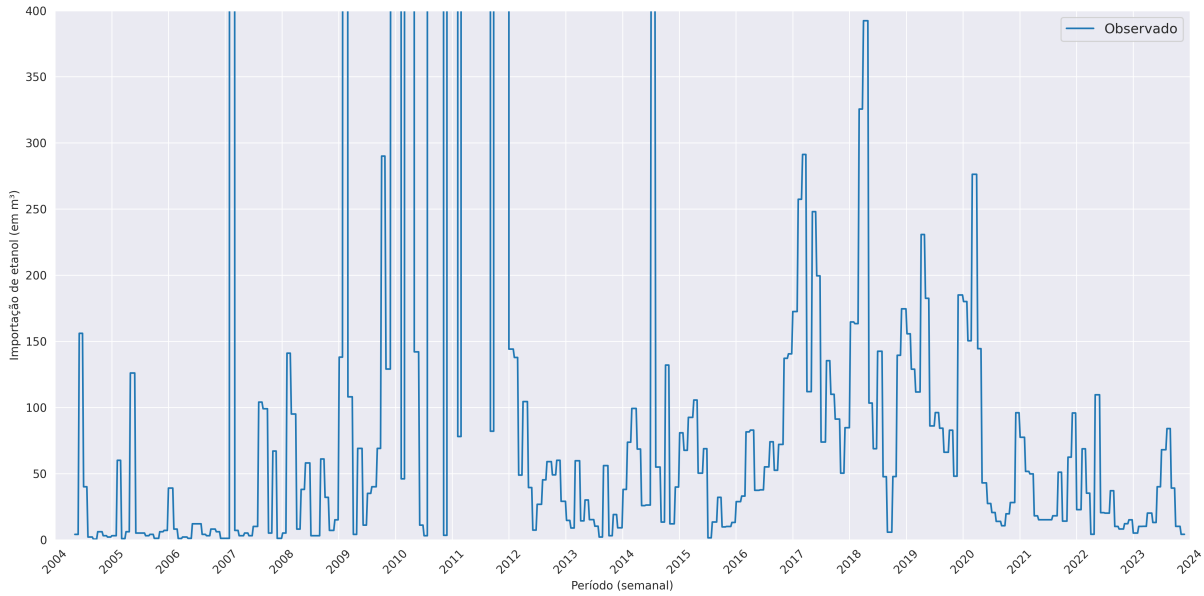
cional. Entretanto, observa-se um pico excepcional em 2011, causado pela quebra de safra ocorrida naquele ano. Esse problema na produção nacional gerou uma alta exorbitante nos números de importação, necessária para garantir o atendimento do mercado interno e evitar o desabastecimento. Nessas janelas de escassez, os volumes importados chegam a superar marcas de 160.000 m^3 e, em casos mais graves, ultrapassam 290.000 m^3. Para permitir uma observação mais clara do comportamento habitual do setor, a Figura 4.13 apresenta um recorte com o eixo vertical limitado entre 0 e 400 m^3, facilitando a análise dos dados fora do intervalo atípico de 2011.

Figura 4.12: Importação de etanol (m^3) no período de 2004 a 2023.



Fonte: Elaboração própria.

Figura 4.13: Importação de etanol (m^3) no período de 2004 a 2023. (recorte)



Fonte: Elaboração própria.

De forma geral, a análise exploratória confirma que o preço do etanol é resultado de um sistema multivariado complexo, envolvendo fatores energéticos, agrícolas, cambiais e fiscais. A presença de tendências, sazonalidade implícita e choques estruturais reforça a necessidade de modelos capazes de capturar não linearidades e dependências temporais de longo prazo, justificando a adoção das abordagens de aprendizado de máquina empregadas neste estudo.

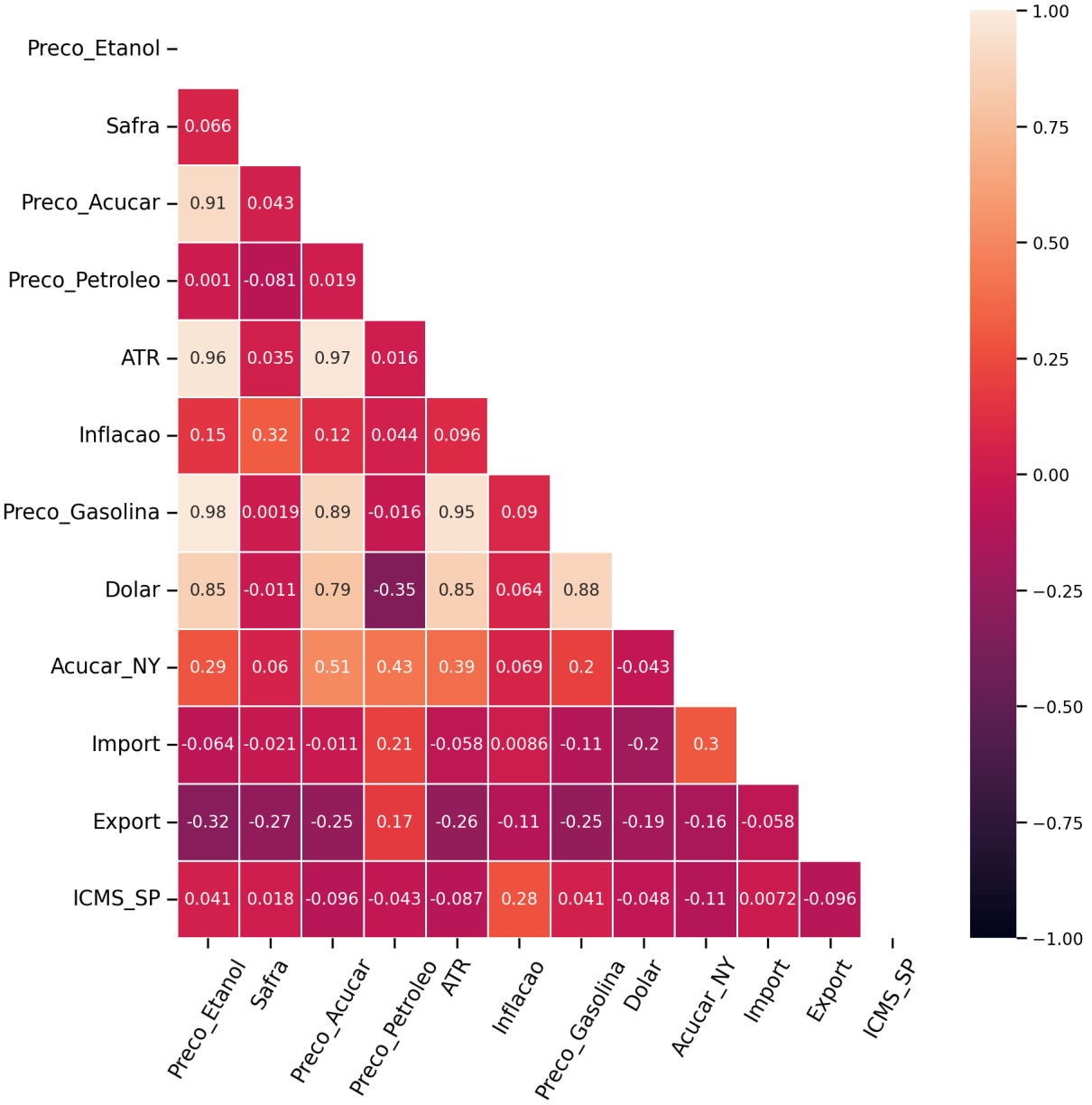
4.2 Importância e Seleção das Variáveis

A análise dos coeficientes de correlação de Pearson entre as variáveis, apresentada na Figura 4.14, fornece evidências relevantes sobre a estrutura de dependência do conjunto de dados, sendo fundamental para a construção de modelos preditivos mais robustos. Observou-se forte correlação positiva entre o preço do etanol e o preço da gasolina (0,98), o preço do ATR (0,96) e o preço do açúcar (0,91), indicando a forte influência conjunta dos mercados de combustíveis e da cadeia sucroenergética sobre a formação do preço do etanol.

Em contraste, verificou-se correlação negativa significativa com a cotação do dólar (-0,85), sugerindo que a valorização cambial está associada à redução do preço interno do etanol, possivelmente devido a ajustes nos fluxos de exportação e nos preços relativos das commodities. A variável exportação apresentou correlação negativa moderada (-0,32), o que pode indicar realocação de oferta entre mercado interno e externo.

As demais variáveis apresentaram correlações fracas com o preço do etanol, como safra (0,066), preço do petróleo (0,001) e inflação (0,15). Correlações negativas de baixa magnitude também foram observadas para açúcar NY (-0,29), importação (-0,064) e ICMS (-0,041), sugerindo influência marginal direta dessas variáveis na escala linear da série analisada.

Figura 4.14: Correlação de Pearson entre as variáveis analisadas.



Fonte: Elaboração própria.

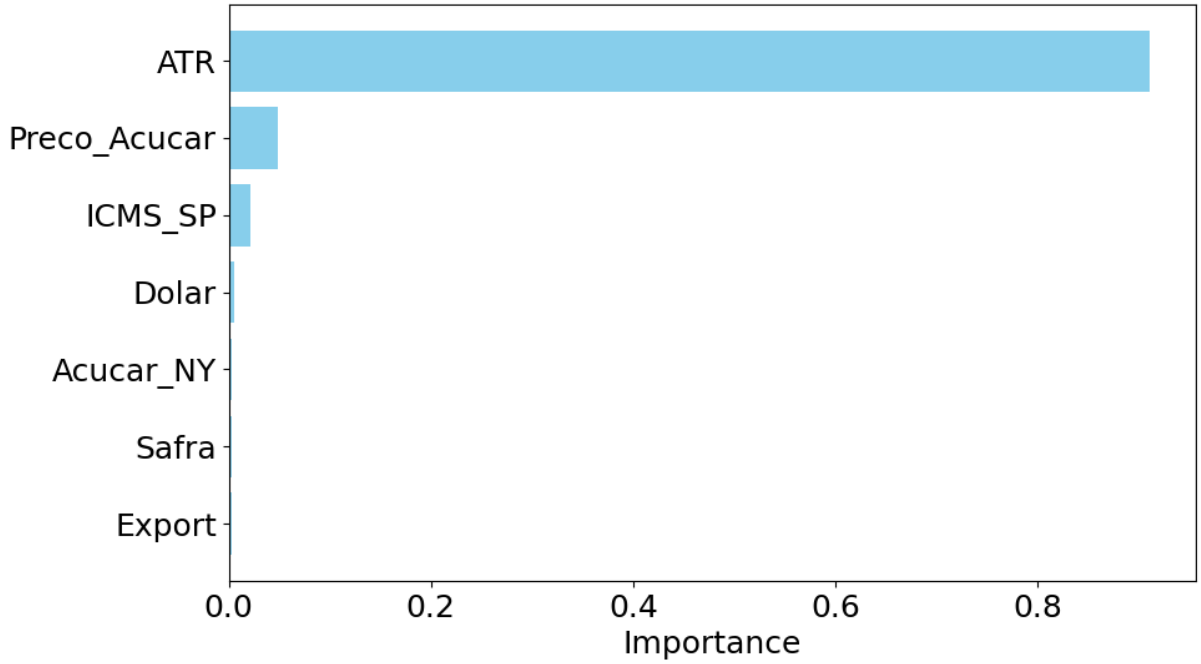
Com base nos resultados apresentados na Figura 4.14, optou-se inicialmente pela retirada da variável Preço da Gasolina. A correlação de 0,98 com o Preço do Etanol evidencia uma associação praticamente linear entre as duas séries, caracterizando forte redundância informacional. A permanência simultânea dessas variáveis no conjunto de dados tenderia a introduzir colinearidade excessiva, com potencial de inflar a variância dos estimadores em modelos lineares e favorecer *overfitting*) nos algoritmos de aprendizado de máquina.

Após a exclusão do Preço da Gasolina, o Preço do Petróleo também deixou de fazer sentido na base de dados. Sua principal relação ocorria justamente por meio da gasolina, e, isoladamente, o Preço do Petróleo não apresentou relação relevante com o Preço do Etanol ou com as demais variáveis explicativas. Dessa forma, sua manutenção no modelo não acrescentaria informação útil ao processo de predição, motivo pelo qual também foi retirado da análise.

A Figura 4.15 apresenta a importância relativa das variáveis estimada pelo modelo *Random Forest*. Diferentemente da correlação de Pearson, que mensura a intensidade da associação linear entre pares de variáveis, a métrica *Feature Importance* baseia-se na redução média da impureza proporcionada por cada preditor ao longo das árvores que compõem o conjunto. Assim, a medida reflete a contribuição efetiva de cada variável para a diminuição do erro de previsão no processo de treinamento.

Os resultados evidenciam que o ATR se destaca de forma expressiva como a variável de maior relevância para a predição do preço do etanol, apresentando importância substancialmente superior às demais. Na sequência, o Preço do Açúcar aparece como o segundo atributo mais influente. As demais variáveis analisadas apresentam importância relativa inferior, sugerindo contribuição complementar na estrutura preditiva capturada pelo algoritmo.

Figura 4.15: Importância das variáveis.



Fonte: Elaboração própria.

A Tabela 4.1 apresenta os valores do Fator de Inflação da Variância para as variáveis consideradas no modelo. Observa-se que o Preço do Açúcar (156.88) e o ATR (137.44) exibem valores substancialmente superiores aos das demais variáveis, caracterizando um quadro de multicolinearidade acentuada. Esse resultado é consistente com a forte interdependência entre variáveis associadas à cadeia sucroenergética, já evidenciada na análise de correlação e na hierarquização de importância pelo *Random Forest*. Adicionalmente, variáveis como Dólar (64.26), ICMS SP (52.35) e Açúcar NY (43.25) também apresentam níveis elevados de VIF, indicando presença de colinearidade relevante no conjunto de dados. Em contraste, Exportação (4.40), Inflação (3.49), Safra (1.75) e Importação (1.19) mantêm valores baixos, sugerindo menor redundância linear com os demais preditores.

Tabela 4.1: Fator de Inflação de Variância (VIF) das variáveis selecionadas.

Atributo	VIF
Preço Açúcar	156.88
ATR	137.44
Dólar	64.26
ICMS SP	52.35
Açúcar NY	43.25
Exportação	4.40
Inflação	3.49
Safra	1.75
Importação	1.19

Fonte: Elaboração própria.

Considerando os elevados níveis de multicolinearidade identificados anteriormente, foi implementada uma rotina iterativa de exclusão sucessiva das variáveis com maior VIF, seguida do recálculo da métrica, até a obtenção de um conjunto em que todos os preditores apresentassem VIF inferior ou igual a 5. O primeiro conjunto resultante desse procedimento é apresentado na Tabela 4.2. Nota-se que o Preço do Açúcar foi excluído por possuir o VIF mais alto na análise anterior. Entretanto, com a saída dessa variável, o ATR se manteve entre as cinco variáveis com VIF abaixo de 5 ao final da análise.

Tabela 4.2: Fator de Inflação de Variância (VIF) após rotina de exclusão.

Atributo	VIF
Inflação	3.02
ATR	2.81
Exportação	2.24
Safra	1.64
Importação	1.03

Fonte: Elaboração própria.

Considerando que o objetivo do procedimento é obter um conjunto de variáveis com níveis reduzidos de multicolinearidade, de modo a não comprometer a confiabilidade das estimativas, a utilização simultânea de ATR e Preço do Açúcar mostrou-se inadequada, dado o elevado grau de dependência linear entre esses atributos. Por outro lado, a exclusão do Preço do Açúcar não se mostrou metodologicamente razoável, uma vez que essa variável apresentou, de forma consistente, alta importância preditiva nas análises realizadas anterioremente.

Diante disso, optou-se pela remoção do ATR e pela realização de uma nova varredura do VIF. O conjunto resultante, apresentado na Tabela 4.3, atende ao critério estabelecido de VIF máximo inferior ou igual a 5. Observa-se que, nesse novo arranjo, com a ausência inicial do ATR, o Preço do Açúcar permaneceu entre as variáveis apresentadas. Dessa forma, foi possível preservar a estabilidade estatística do modelo sem abdicar de um dos atributos mais relevantes para a predição do preço do etanol.

Tabela 4.3: Fator de Inflação de Variância (VIF) após rotina de exclusão, sem ATR.

Atributo	VIF
Inflação	3.04
Preço Açúcar	2.89
Exportação	2.25
Safra	1.65
Importação	1.03

Fonte: Elaboração própria.

Após a criteriosa análise das variáveis presentes no conjunto de dados original, foram definidos três conjuntos de atributos (features) para avaliar o impacto da multicolinearidade e da redundância de informação na performance dos modelos, a partir dos resultados apresentados na Figura 4.15 e nas Tabelas 4.1, 4.2 e 4.3:

1. **Conjunto Completo de Atributos (CCA):** inclui todas as variáveis inicialmente coletadas.
2. **Conjunto Reduzido – Variante ATR (CR-ATR):** obtido após exclusão de variáveis com alta multicolinearidade ($|\rho| \geq 0{,}75$ e VIF > 5), mantendo o ATR como principal variável representativa da cadeia sucroenergética.
3. **Conjunto Reduzido – Variante Açúcar (CR-AC):** configuração alternativa ao CR-ATR, na qual o Preço do Açúcar substitui o ATR como variável representativa, dado o elevado grau de correlação entre ambas.

Assim, a seleção levou à construção das bases com as seguintes variáveis:

- **CCA:** Preço do Etanol, Safra, Preço do Açúcar, ATR, Inflação, Dólar, Açúcar NY, Importação, Exportação, ICMS SP.
- **CR-ATR:** Safra, Inflação, Importação, Exportação, ATR.
- **CR-AC:** Safra, Inflação, Importação, Exportação, Preço do Açúcar.

As previsões foram feitas em três períodos distintos: pré-pandemia (até 2019), pandemia (2020–2022) e pós-pandemia (a partir de 2023). A análise em blocos temporais permite avaliar o desempenho do modelo em contextos econômicos e sociais contrastantes, destacando tanto sua capacidade de capturar padrões históricos estáveis quanto suas limitações diante de choques exógenos, como os observados durante a pandemia de COVID-19.

4.3 Previsão do Modelo Random Forest

O modelo *Random Forest* foi implementado através da classe *RandomForestRegressor* da biblioteca *scikit-learn*. Para encontrar a configuração ideal de floresta, utilizou-se a ferramenta *RandomizedSearchCV*, que realizou uma busca aleatória em uma grade de hiperparâmetros definida com os seguintes critérios:

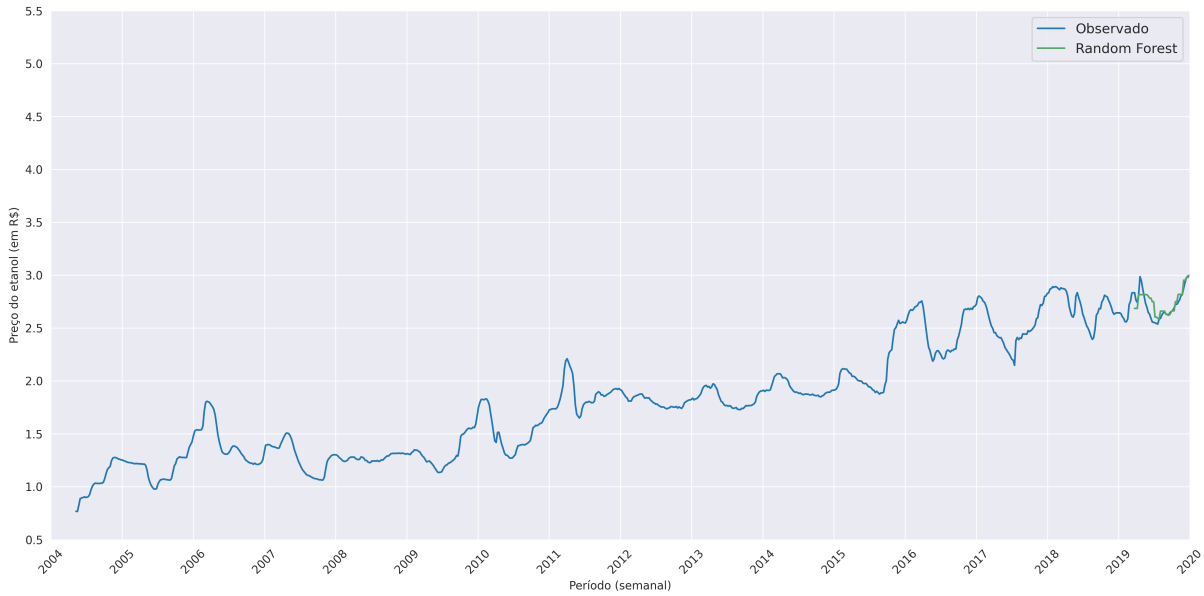
- *n_estimators*: Testou-se a quantidade de árvores na floresta entre 25, 50, 100, 200, 500 e 1000 unidades;
- *max_depth*: A profundidade máxima das árvores variou de níveis rasos (2) até profundos (30), incluindo a opção de profundidade ilimitada;
- *min_samples_split* e *min_samples_leaf*: Definiram-se os requisitos mínimos de amostras para divisão de nós (2 a 10) e para a formação de folhas (1 a 8);
- *max_features*: A seleção de recursos para cada divisão incluiu as opções automática, raiz quadrada e logaritmo na base 2;
- *bootstrap*: Avaliou-se a aplicação ou não da técnica de amostragem com reposição.

A busca foi configurada para 10 iterações aleatórias com validação cruzada de 5 dobras, utilizando o erro quadrático médio negativo como métrica de desempenho. O modelo final foi treinado com os melhores parâmetros identificados, mantendo o estado de aleatoriedade fixado ($random_state = 42$) para garantir a reprodutibilidade dos resultados sobre os dados.

4.3.1 Base CCA

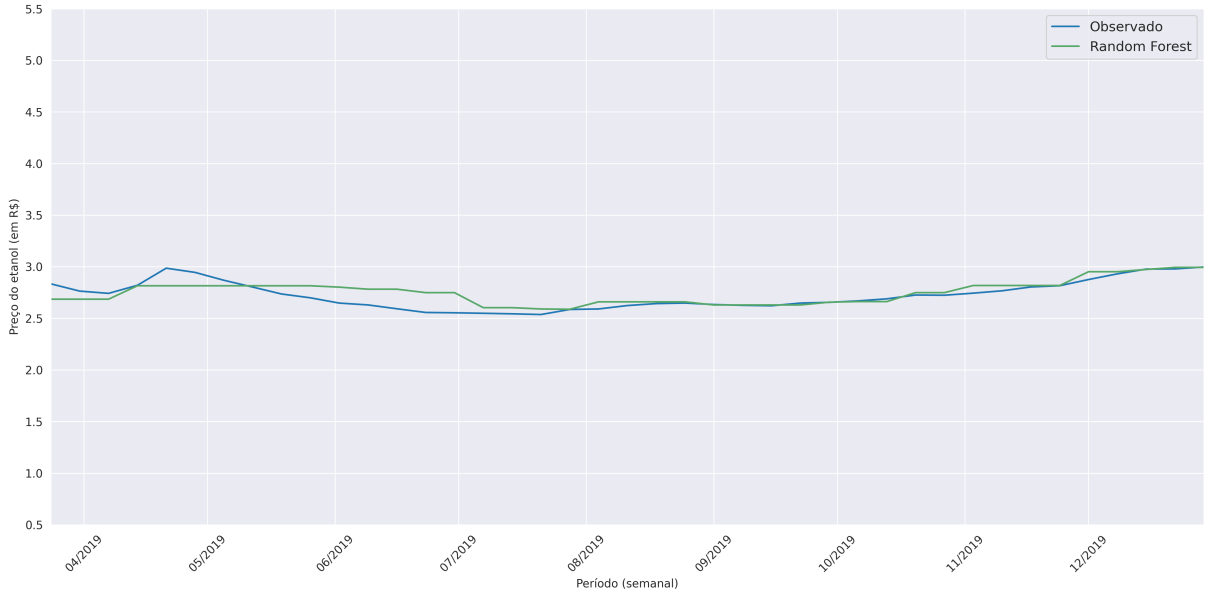
A Figura 4.16 ilustra o desempenho preditivo do modelo *Random Forest* (RF), treinado com a base CCA, em comparação com os valores observados no conjunto de teste referente ao período pós-2019. Conforme evidenciado também na Figura 4.17, nota-se uma correspondência satisfatória entre as previsões geradas pelo modelo e os preços efetivamente registrados, o que reforça sua capacidade de captar a dinâmica do mercado no intervalo analisado. Além disso, as estimativas apontam para uma tendência de elevação nos preços do etanol nos períodos subsequentes, indicando que o modelo consegue sugerir direções gerais mesmo em cenários prospectivos.

Figura 4.16: Preço do Etanol observado vs. preço previsto pelo modelo RF no conjunto de teste, utilizando a base CCA.



Fonte: Elaboração própria.

Figura 4.17: Preço do Etanol observado vs. preço previsto pelo modelo RF pré-pandemia, utilizando a base CCA.

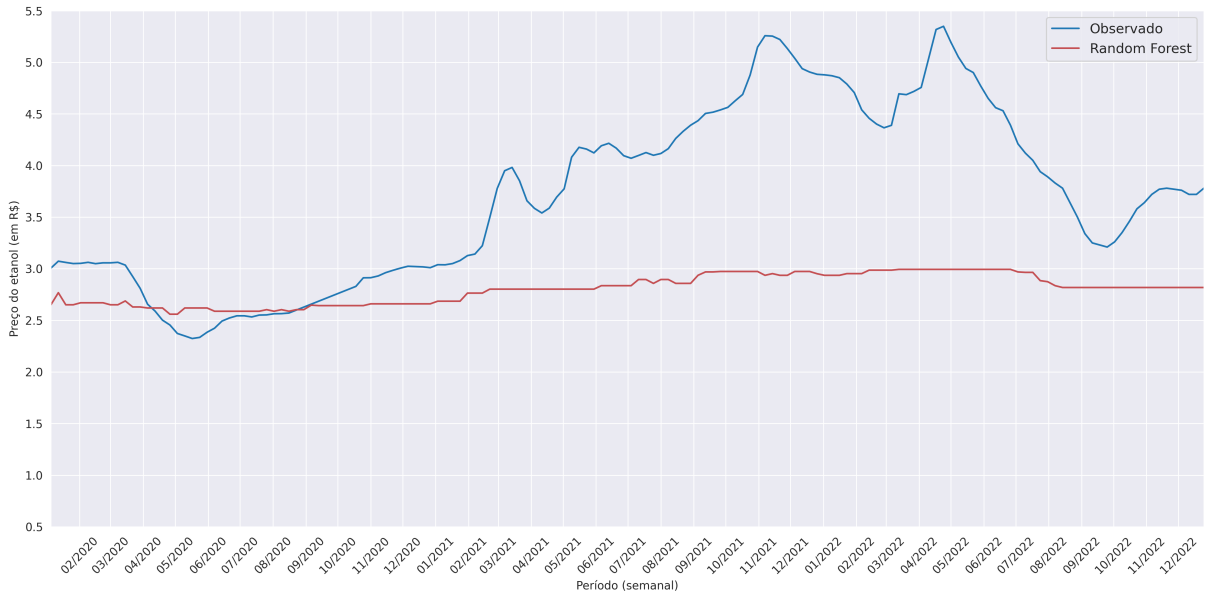


Fonte: Elaboração própria.

No intervalo pré-pandêmico, o modelo *Random Forest* (RF), calibrado com a base CCA, apresentou desempenho robusto, caracterizado por baixa discrepância entre os valores estimados e os dados observados. Tal desempenho é evidenciado pelas métricas de erro obtidas: MAPE de 2.18%, MAE de 0.059 e MSE de 0.007, indicando elevada precisão preditiva no período em questão.

Por sua vez, o período pandêmico (2020–2022) introduziu rupturas significativas no comportamento do mercado de combustíveis, aspecto claramente refletido nos resultados do modelo. Como ilustrado na Figura 4.18, o algoritmo não conseguiu antecipar a queda abrupta nos preços do etanol no início da crise sanitária, revelando limitações frente a eventos exógenos e de difícil previsão. Ao longo de grande parte desse intervalo, o modelo tendeu à subestimação dos preços, o que evidencia uma menor sensibilidade a variações bruscas no cenário econômico. Nesse contexto, os indicadores de erro foram sensivelmente mais elevados: MAPE de 22.26%, MAE de 0.934 e MSE de 1.341.

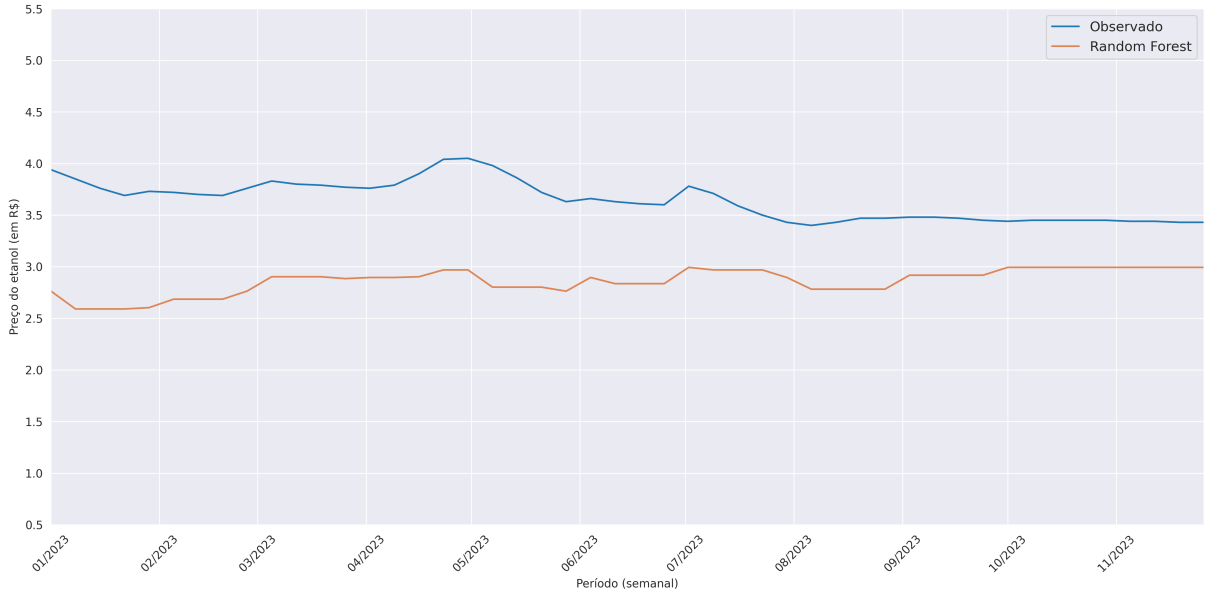
Figura 4.18: Preço do Etanol observado vs. preço previsto pelo modelo RF durante a pandemia, utilizando a base CCA.



Fonte: Elaboração própria.

Em 2023, observa-se uma relativa estabilidade nos preços do etanol, marcada por menor amplitude nas oscilações e ausência de variações abruptas. Apesar desse cenário menos volátil, o modelo *Random Forest* (RF), baseado na base CCA, manteve a tendência de subestimar os valores observados, resultando em discrepâncias sistemáticas ao longo do período analisado, conforme ilustrado na Figura 4.19. Os indicadores de erro para esse intervalo apontam para um MAPE de 21.25%, MAE de 0.834 e MSE de 0.756, sugerindo que, embora o comportamento dos preços tenha se tornado mais previsível, o modelo não acompanhou adequadamente essa mudança no padrão de variabilidade.

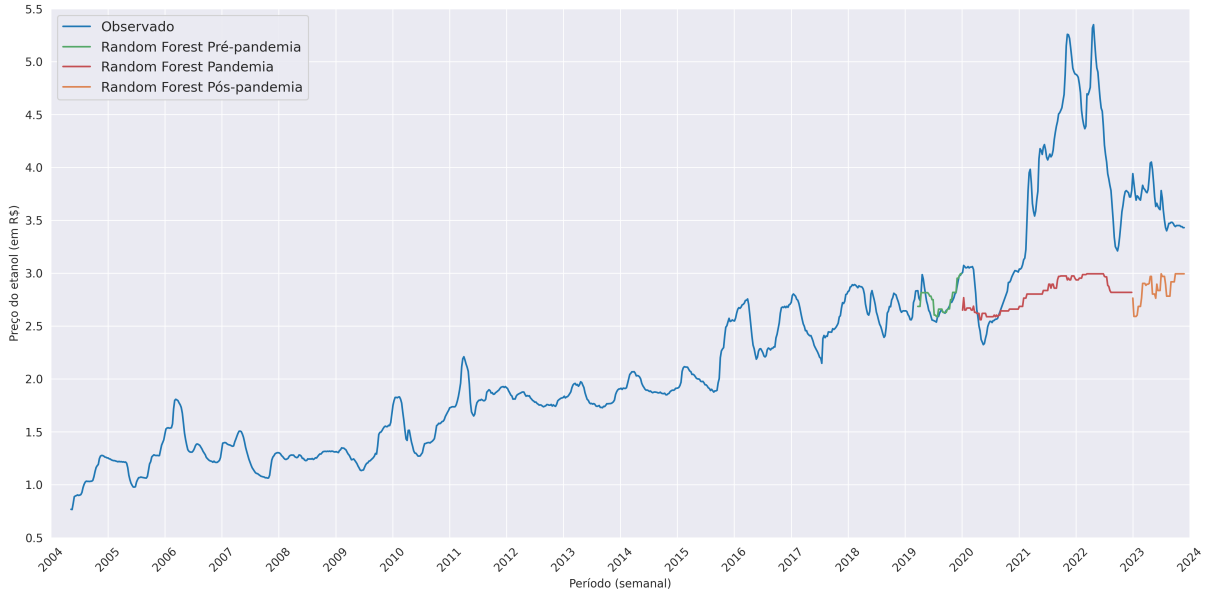
Figura 4.19: Preço do Etanol observado vs. preço previsto pelo modelo RF pós-pandemia, utilizando a base CCA.



Fonte: Elaboração própria.

Em resumo, a Figura 4.20 apresenta uma síntese dos resultados do modelo para os três períodos analisados nesta base de dados.

Figura 4.20: Preço do Etanol observado vs. preço previsto pelo modelo RF no conjunto total, utilizando a base CR-CCA.



Fonte: Elaboração própria.

A Tabela 4.4 sintetiza os indicadores de erro obtidos pelo modelo *Random Forest* ao longo dos três recortes temporais analisados: pré-pandemia, pandemia e pós-pandemia. No período anterior à crise sanitária, o desempenho do modelo foi bastante satisfatório,

refletido nos baixos níveis de erro: MAPE de 2.18%, MAE de 0.0589 e MSE de 0.0071, o que denota elevada aderência entre as previsões e os valores observados.

Durante o período pandêmico, os erros apresentaram crescimento expressivo, com MAPE de 22.26%, MAE de 0.9349 e MSE de 1.3415. Tal deterioração no desempenho era, em certa medida, previsível, uma vez que o modelo demonstrou limitações em capturar os choques abruptos e as disrupções incomuns no mercado de combustíveis decorrentes do contexto excepcional imposto pela COVID-19.

Já no intervalo pós-pandêmico, verifica-se uma atenuação nos erros em relação ao período anterior, com MAPE reduzido para 21.25%, MAE de 0.7833 e MSE de 0.6750. Embora se observe uma tendência de recuperação no desempenho preditivo, os valores permanecem substancialmente mais altos quando comparados ao período pré-crise, indicando persistência de desafios modelísticos frente às flutuações de um mercado ainda em processo de reconfiguração.

De modo geral, os resultados apontam para a sensibilidade dos modelos de previsão frente a eventos extremos, ressaltando a importância de estratégias complementares — como o uso de variáveis contextuais ou técnicas de modelagem adaptativa — para melhorar a acurácia em cenários de alta volatilidade.

Tabela 4.4: Métricas de erro do modelo *Random Forest* nos diferentes períodos.

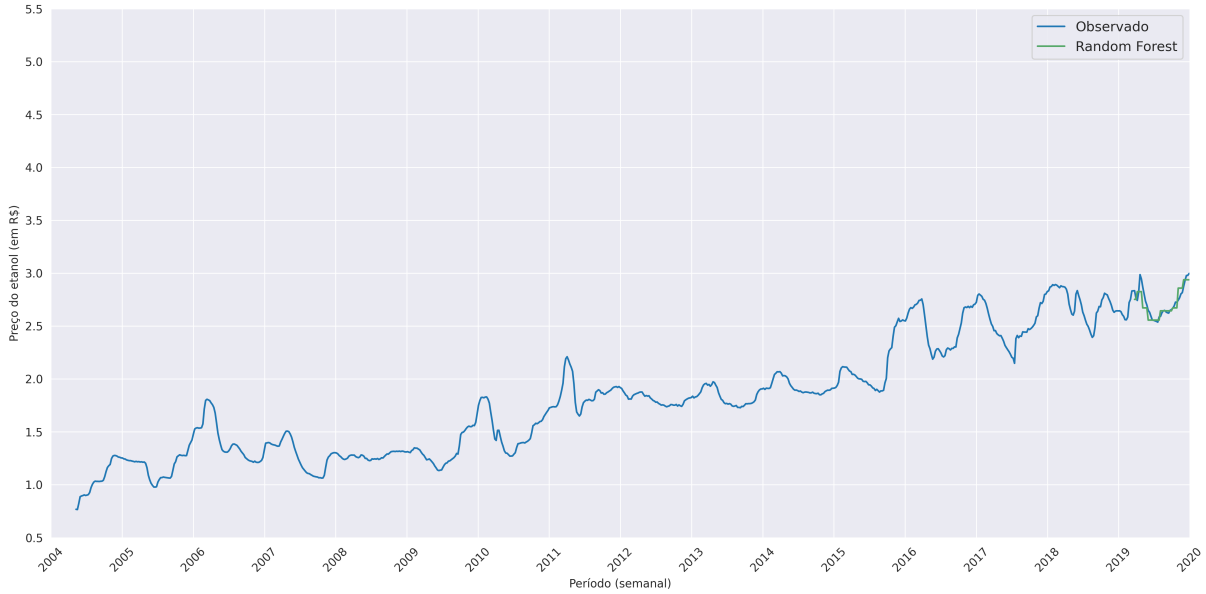
Período	**MAPE (%)**	**MAE**	**MSE**
Pré-pandemia	2.18	0.0589	0.0071
Pandemia	22.26	0.9349	1.3415
Pós-pandemia	21.25	0.7833	0.6750

Fonte: Elaboração própria.

4.3.2 Base CR-ATR

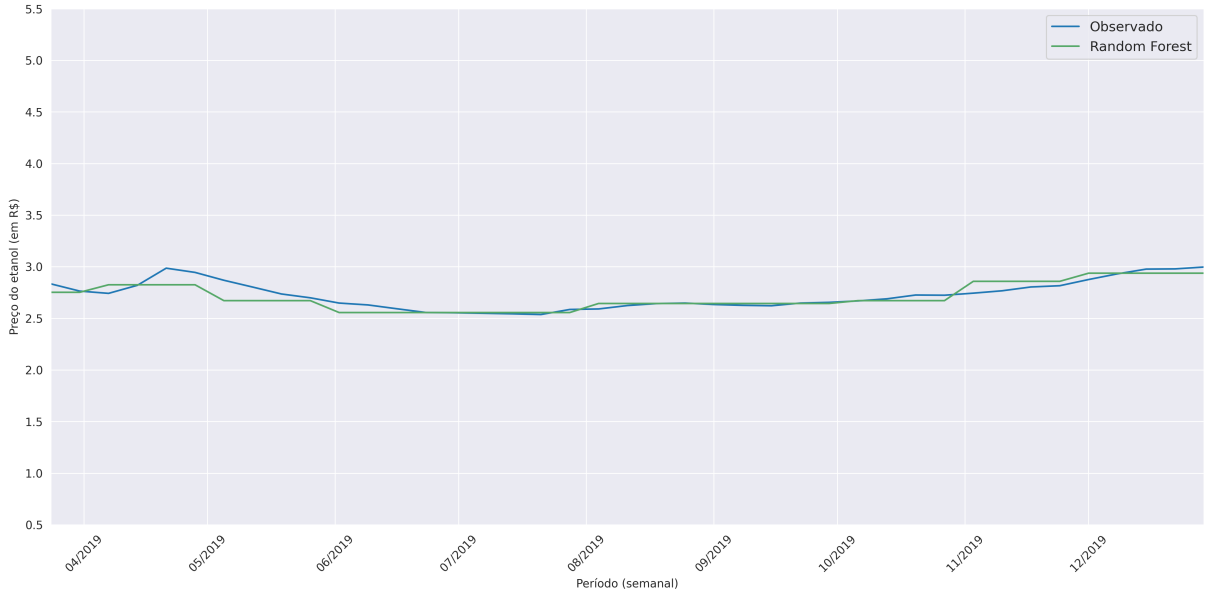
A Figura 4.21 ilustra o desempenho do modelo *Random Forest* alimentado com a base que inclui a variável ATR, no conjunto de teste referente ao período posterior a 2019. Verifica-se que as estimativas do modelo apresentam boa aderência aos valores observados, com trajetória compatível à dos dados empíricos. Além disso, a projeção aponta uma tendência ascendente nos preços do etanol para os períodos subsequentes, sinalizando a capacidade do modelo de capturar dinâmicas de crescimento no comportamento da série.

Figura 4.21: Preço do Etanol observado vs. preço previsto pelo modelo RF no conjunto de teste, utilizando a base CR-ATR.



Fonte: Elaboração própria.

Figura 4.22: Preço do Etanol observado vs. preço previsto pelo modelo RF pré-pandemia, utilizando a base CR-ATR.

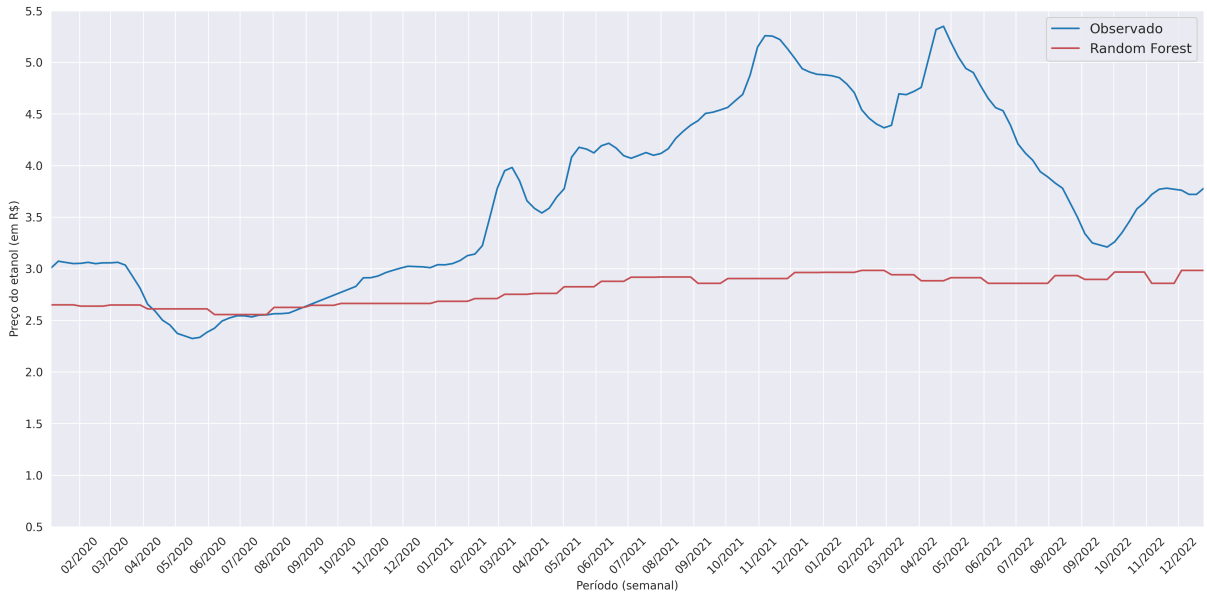


Fonte: Elaboração própria.

No recorte pré-pandêmico, o modelo *Random Forest* alimentado com a base que incorpora a variável ATR apresentou desempenho expressivo, com baixos níveis de erro: MAPE de 1.70%, MAE de 0.047 e MSE de 0.004. Esses resultados indicam elevado grau de acurácia nas previsões, bem como boa aderência entre os valores estimados e os dados observados, o que reforça a robustez do modelo em contextos de estabilidade relativa.

Durante o período pandêmico (2020–2022), contudo, verifica-se uma deterioração significativa no desempenho preditivo. Conforme ilustrado na Figura 4.23, o modelo foi incapaz de antecipar a brusca retração nos preços do etanol no início da crise sanitária, evidenciando sua limitação frente a eventos inesperados e de alta volatilidade. A subestimação dos preços ao longo de grande parte desse intervalo revela baixa responsividade do modelo às rápidas transformações de cenário econômico. Os indicadores de erro nesse período — MAPE de 22.24%, MAE de 0.938 e MSE de 1.379 — confirmam esse quadro de menor precisão.

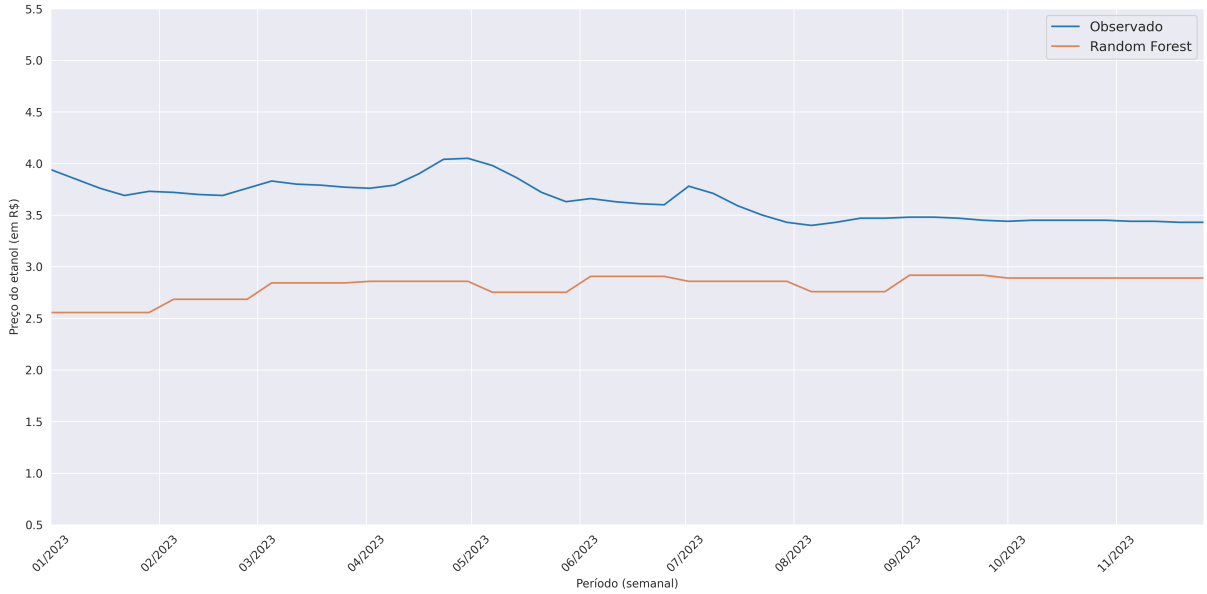
Figura 4.23: Preço do Etanol observado vs. preço previsto pelo modelo RF durante a pandemia, utilizando a base CR-ATR.



Fonte: Elaboração própria.

No ano de 2023, observa-se um cenário de maior estabilidade nos preços do etanol, caracterizado por menor frequência de oscilações acentuadas. Apesar desse contexto menos volátil, o modelo *Random Forest*, mesmo com a inclusão da variável ATR, manteve a tendência de subestimar os preços reais, conforme ilustrado na Figura 4.24. As divergências sistemáticas entre os valores projetados e os efetivamente observados indicam persistência de limitações na capacidade preditiva do modelo. As métricas de erro para o período — MAPE de 22.66%, MAE de 0.835 e MSE de 0.756 — corroboram esse diagnóstico, sinalizando que, embora o ambiente tenha se tornado mais previsível, a acurácia do modelo não apresentou melhora significativa em relação ao período pandêmico.

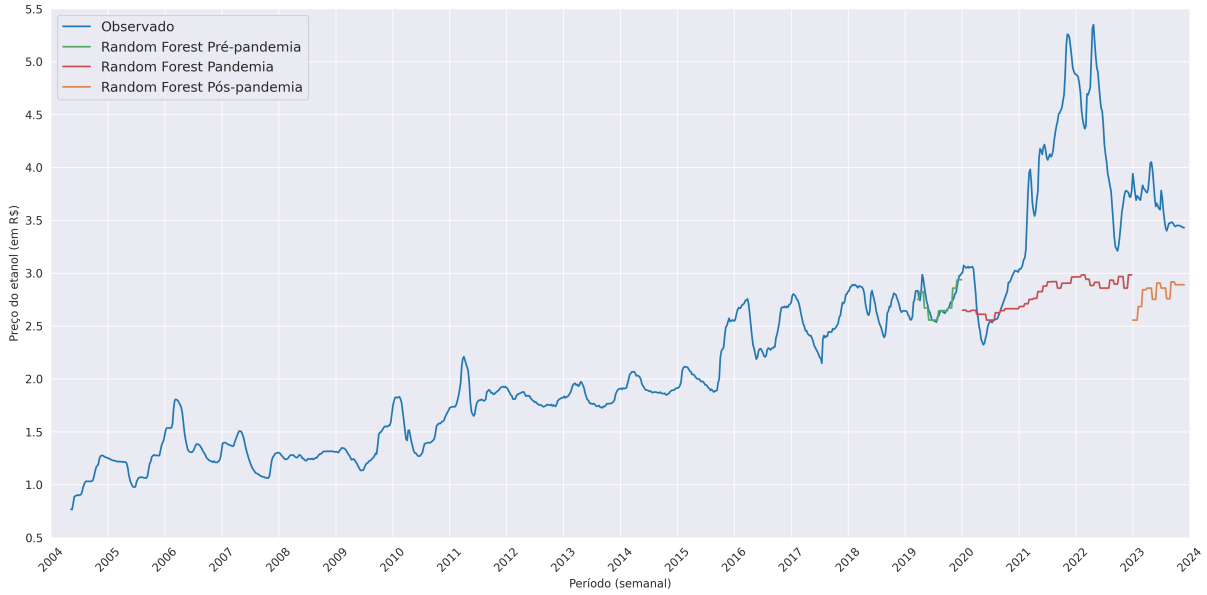
Figura 4.24: Preço do Etanol observado vs. preço previsto pelo modelo RF pós-pandemia, utilizando a base CR-ATR.



Fonte: Elaboração própria.

A Figura 4.25 sintetiza os resultados obtidos pelo modelo ao longo dos três períodos considerados nesta base.

Figura 4.25: Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto total, utilizando a base CR-ATR.



Fonte: Elaboração própria.

A Tabela 4.5 sintetiza os indicadores de desempenho do modelo *Random Forest* construído com base na seleção de variáveis que inclui o ATR, ao longo dos períodos pré-pandemia, pandemia e pós-pandemia. Os resultados evidenciam que, no intervalo ante-

rior à crise sanitária, o modelo alcançou elevada precisão preditiva, com MAPE de 1.70%, MAE de 0.0473 e MSE de 0.0044, demonstrando forte aderência entre os valores estimados e os observados.

Com o advento da pandemia, no entanto, nota-se uma deterioração expressiva do desempenho do modelo, cujos erros aumentaram substancialmente (MAPE de 22.24%, MAE de 0.9378 e MSE de 1.3793). Tal comportamento reforça a limitação da abordagem em capturar eventos exógenos abruptos e não-lineares, típicos de contextos de crise.

No cenário pós-pandêmico, embora os níveis de erro tenham se mantido próximos aos registrados durante a pandemia (MAPE de 22.66%, MAE de 0.8348 e MSE de 0.7581), observa-se certa estabilidade nas métricas. Isso sugere que, mesmo com sinais de retomada econômica, o modelo ainda enfrenta dificuldades para ajustar-se a um ambiente caracterizado por incertezas persistentes e padrões de mercado menos previsíveis.

Tais resultados reiteram o impacto dos choques externos na performance de modelos de previsão e apontam para a importância de abordagens complementares ou híbridas, especialmente em contextos marcados por instabilidades econômicas significativas.

Tabela 4.5: Métricas de erro do modelo *Random Forest* utilizando a base CR-ATR, nos diferentes períodos.

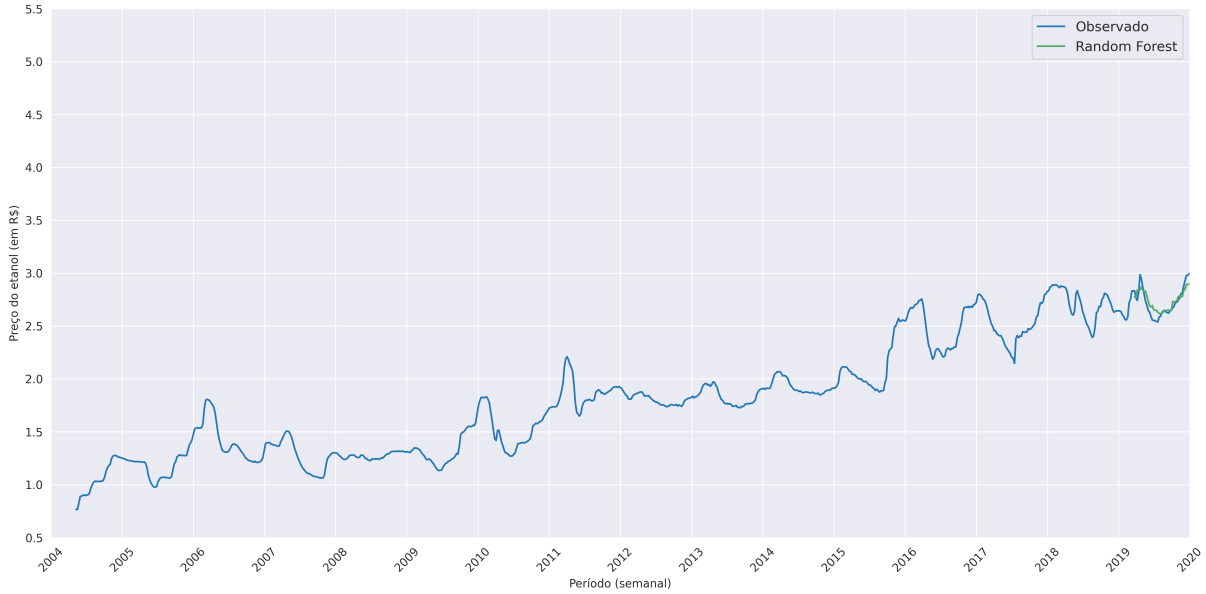
Período	**MAPE (%)**	**MAE**	**MSE**
Pré-pandemia	1.70	0.0473	0.0044
Pandemia	22.24	0.9378	1.3793
Pós-pandemia	22.66	0.8348	0.7581

Fonte: Elaboração própria.

4.3.3 Base CR-AC

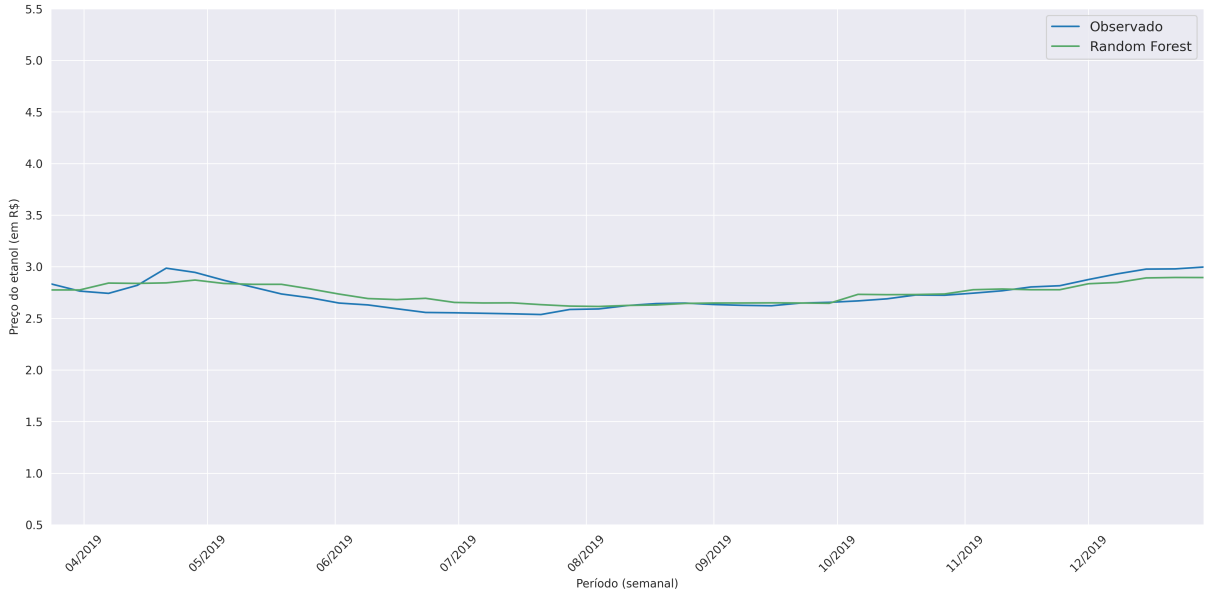
O gráfico exibido na Figura 4.26 ilustra as previsões do modelo *Random Forest*, utilizando a base que inclui a variável preço do açúcar, em comparação com os preços observados no período subsequente a 2019, dentro do conjunto de teste. A partir da análise da Figura 4.27, é possível observar que os valores previstos se alinham de forma bastante precisa aos dados efetivamente registrados. Além disso, a projeção sugere uma tendência de elevação nos preços para os períodos seguintes, indicando a capacidade do modelo em capturar tendências de alta no mercado de açúcar.

Figura 4.26: Preço do Etanol observado vs. preço previsto pelo modelo RF no conjunto de teste, utilizando a base CR-AC.



Fonte: Elaboração própria.

Figura 4.27: Preço do Etanol observado vs. preço previsto pelo modelo RF pré-pandemia, utilizando a base CR-AC.



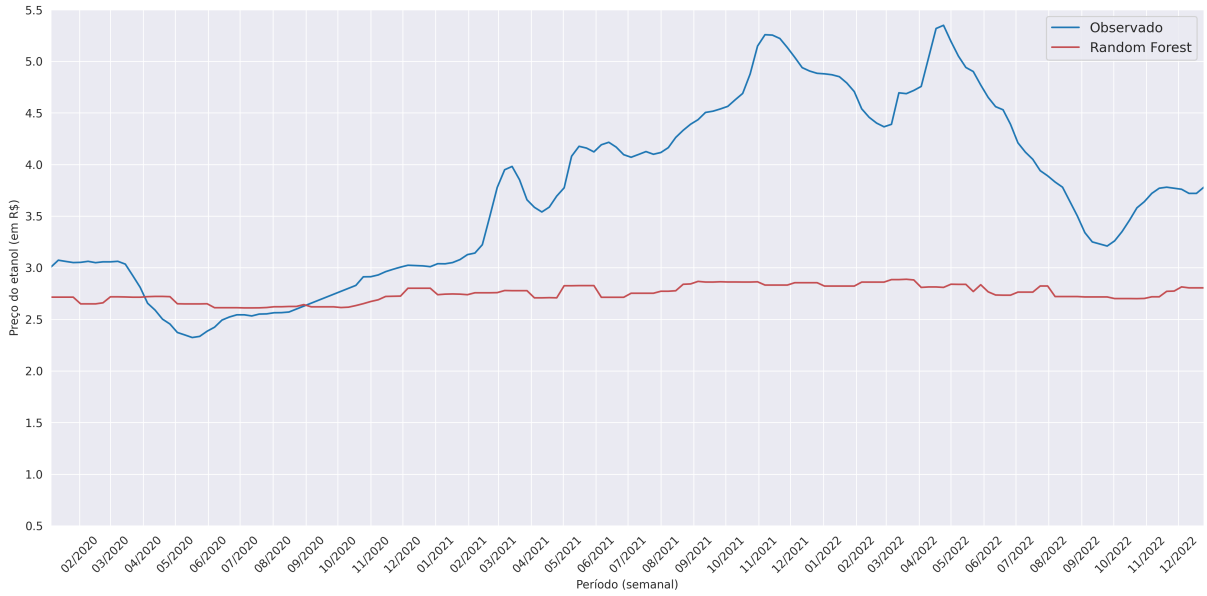
Fonte: Elaboração própria.

O modelo *Random Forest*, utilizando a base CR-AC, apresentou um bom desempenho durante o período pré-pandêmico, com um baixo grau de discrepância entre os valores previstos e observados. Esse comportamento indica um ajuste adequado para o recorte analisado. Os indicadores de erro do modelo foram favoráveis, com um MAPE de 1.97%,

MAE de 0.054 e MSE de 0.004, sugerindo alta precisão na previsão dos preços durante esse período.

Entretanto, o cenário da pandemia (2020–2022) causou um impacto substancial no mercado de combustíveis, o qual foi refletido nos dados analisados. Como ilustrado na Figura 4.28, no início da crise sanitária, em 2020, o modelo *Random Forest* com base CCA não conseguiu prever a queda abrupta nos preços do etanol. Esse resultado revela uma limitação do modelo em capturar eventos imprevistos e de grande magnitude. Consequentemente, o modelo apresentou uma subestimação dos preços ao longo de grande parte do período pandêmico, evidenciando sua menor sensibilidade às mudanças drásticas no contexto econômico. Durante esse intervalo, o modelo obteve MAPE de 23.73%, MAE de 0.997 e MSE de 1.537, refletindo a dificuldade em se ajustar às flutuações inesperadas do mercado.

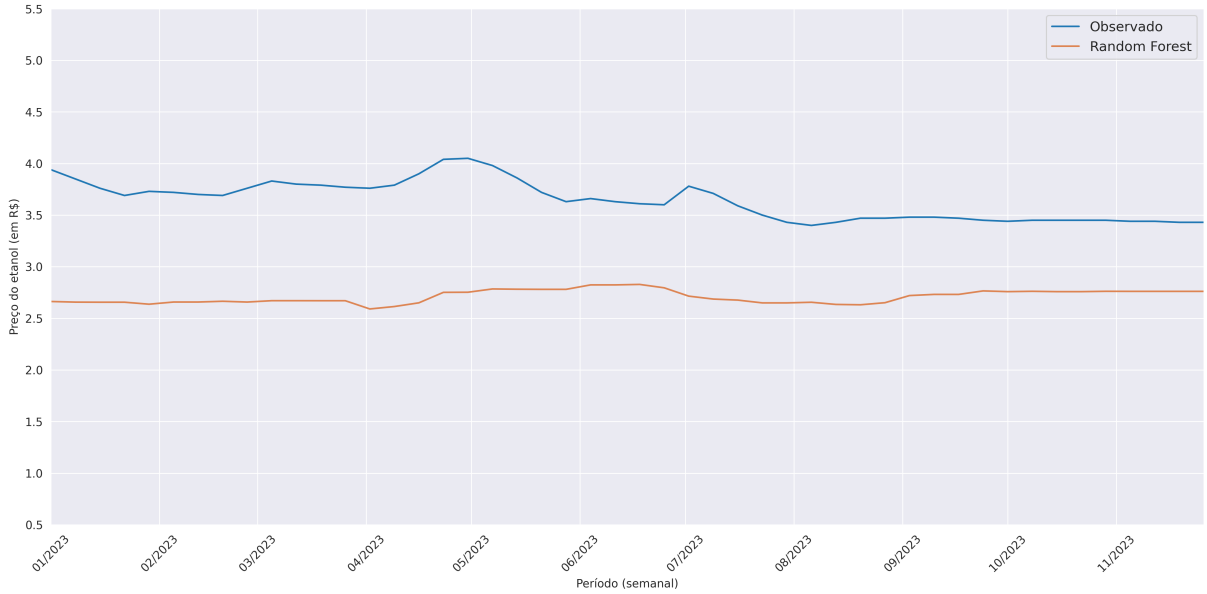
Figura 4.28: Preço do Etanol observado vs. preço previsto pelo modelo RF durante a pandemia, utilizando a base CR-AC.



Fonte: Elaboração própria.

No ano de 2023, observou-se uma maior estabilidade nos preços, com uma diminuição nas oscilações e variações abruptas. No entanto, o modelo *Random Forest*, utilizando a base CR-AC, continuou a subestimar os preços durante esse período, apresentando discrepâncias consistentes entre os valores previstos e os dados observados (conforme mostrado na Figura 4.29). Durante esse intervalo, os indicadores de erro foram de MAPE de 25.42%, MAE de 0.933 e MSE de 0.912, refletindo a persistente limitação do modelo em captar com precisão as flutuações do mercado, mesmo em um cenário de maior estabilidade.

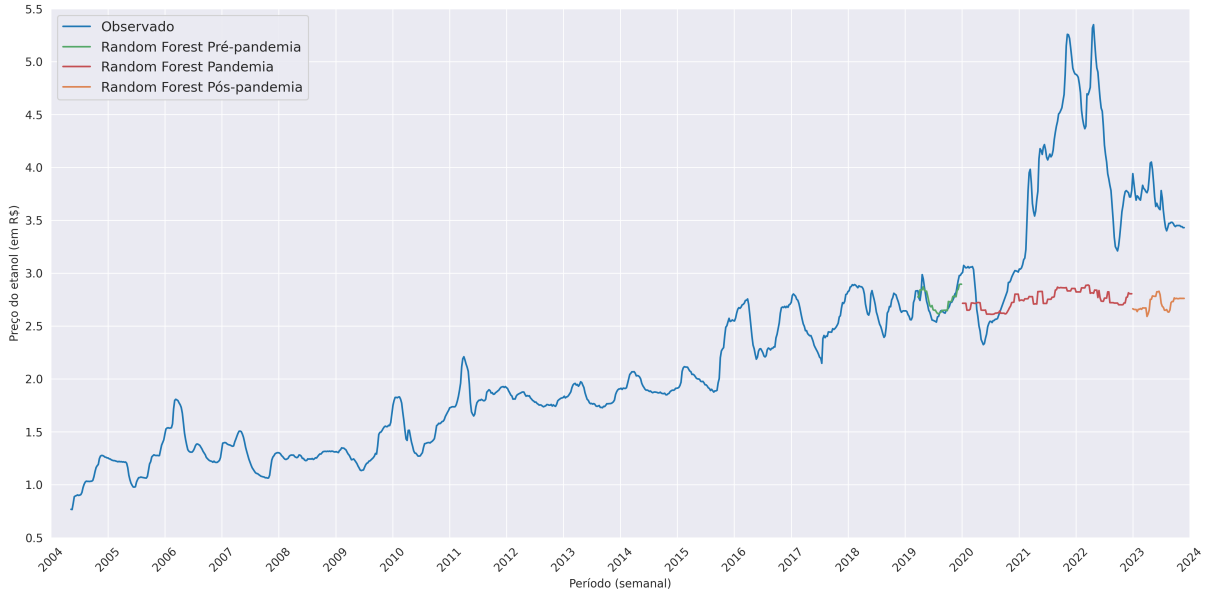
Figura 4.29: Preço do Etanol observado vs. preço previsto pelo modelo RF pós-pandemia, utilizando a base CR-AC.



Fonte: Elaboração própria.

Os resultados do modelo, referentes aos três períodos avaliados nesta base, estão resumidamente apresentados na Figura 4.30.

Figura 4.30: Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto total, utilizando a base CR-AC.



Fonte: Elaboração própria.

A Tabela 4.6 apresenta as métricas de erro do modelo *Random Forest*, utilizando a base com seleção de variáveis, incluindo a variável Preço do Açúcar, para os períodos pré-pandemia, pandemia e pós-pandemia. No período pré-pandêmico, o modelo demonstrou

um desempenho satisfatório, apresentando um MAPE de 1.97%, MAE de 0.0536 e MSE de 0.0044, o que indica uma baixa discrepância entre os valores previstos e observados.

Durante a pandemia, os erros aumentaram de forma significativa, com MAPE de 23.73%, MAE de 0.9974 e MSE de 1.537, refletindo a dificuldade do modelo em acompanhar as flutuações abruptas no mercado, exacerbadas pelo impacto da crise sanitária global.

No período pós-pandemia, observou-se que o modelo manteve-se relativamente estável, apresentando MAPE de 25.42%, MAE de 0.9334 e MSE de 0.9122. Apesar da recuperação econômica parcial, esses valores ainda sugerem desafios na previsão precisa de preços em cenários de alta volatilidade e incerteza econômica.

Esses resultados destacam a sensibilidade dos modelos de previsão a eventos extremos e reforçam a necessidade de desenvolvimento de estratégias complementares que possam lidar com contextos caracterizados por grande incerteza e flutuações imprevisíveis.

Tabela 4.6: Métricas de erro do modelo *Random Forest* utilizando a base CR-AC, nos diferentes períodos.

Período	**MAPE (%)**	**MAE**	**MSE**
Pré-pandemia	1.97	0.0536	0.0044
Pandemia	23.73	0.9974	1.5370
Pós-pandemia	25.42	0.9334	0.9122

Fonte: Elaboração própria.

4.4 Previsão do Modelo LSTM

Para a rede neural recorrente LSTM, utilizou-se a biblioteca *TensorFlow / Keras*. Devido à dependência temporal da série de preços, os dados foram reestruturados em janelas de histórico com um tamanho de 15 semanas. A arquitetura foi otimizada em duas etapas. Inicialmente, estabeleceu-se um modelo base sequencial composto por duas camadas LSTM (6 e 12 unidades), seguidas por uma camada densa de 6 neurônios e uma saída linear. O treinamento utilizou o otimizador *Adam* e a função de perda de erro quadrático médio, incorporando a técnica de interrupção precoce (*EarlyStopping*) para monitorar a perda e evitar o sobreajuste. Posteriormente, aplicou-se o *Keras Tuner* (*RandomSearch*) para realizar uma busca refinada na arquitetura. O espaço de busca permitiu:

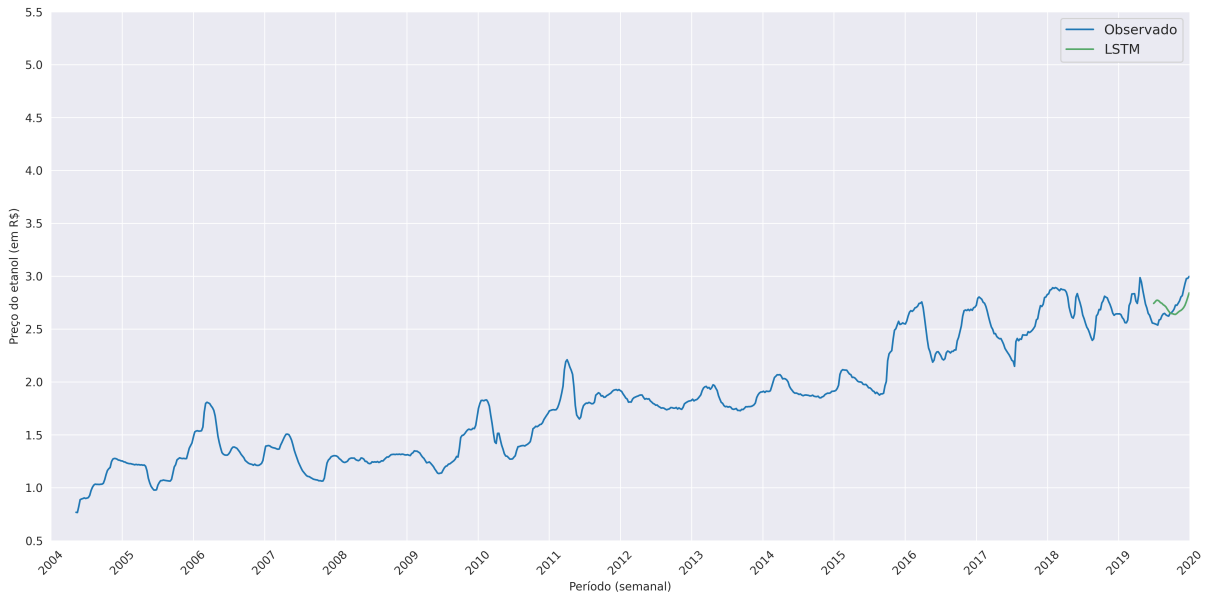
- Número de camadas: Variação dinâmica entre 1 e 10 camadas ocultas;
- Unidades LSTM: Ajuste entre 32 e 128 neurônios por camada, com incrementos de 32;
- Taxa de *Dropout*: Inclusão de camadas de descarte com taxas entre 0 e 0,5 para aumentar a capacidade de generalização;
- Execução: O processo foi configurado para 10 tentativas, com 3 execuções por tentativa para garantir a estabilidade estatística dos resultados.

O ajuste final do melhor modelo foi realizado com um tamanho de lote de 128 ao longo de 20 épocas, validando o desempenho frente aos dados de teste da série.

4.4.1 Base CCA

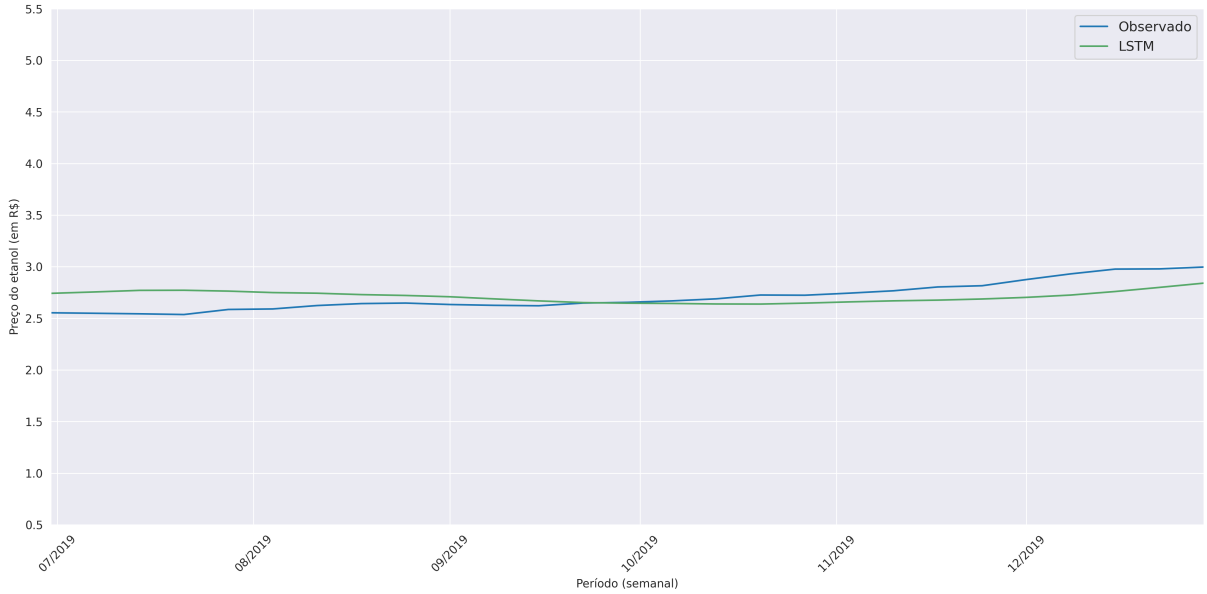
Conforme ilustrado na Figura 4.31, o modelo LSTM demonstrou bom desempenho na previsão dos preços do etanol durante o período de teste pré-pandemia (observável no final do gráfico, próximo a 2020), acompanhando de forma consistente os valores observados e capturando adequadamente as tendências e variações sazonais. A previsão sugere uma leve tendência de alta no início de 2020, o que se alinha com a inclinação da linha de previsão. As métricas de avaliação corroboram a qualidade do ajuste, com MAPE de 4.48%, MAE de 0.121 e MSE de 0.019.A Figura 4.32 apresenta um recorte específico do período de teste, permitindo uma análise mais detalhada da aproximação entre os valores previstos e observados.

Figura 4.31: Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto de teste, utilizando a base CCA.



Fonte: Elaboração própria.

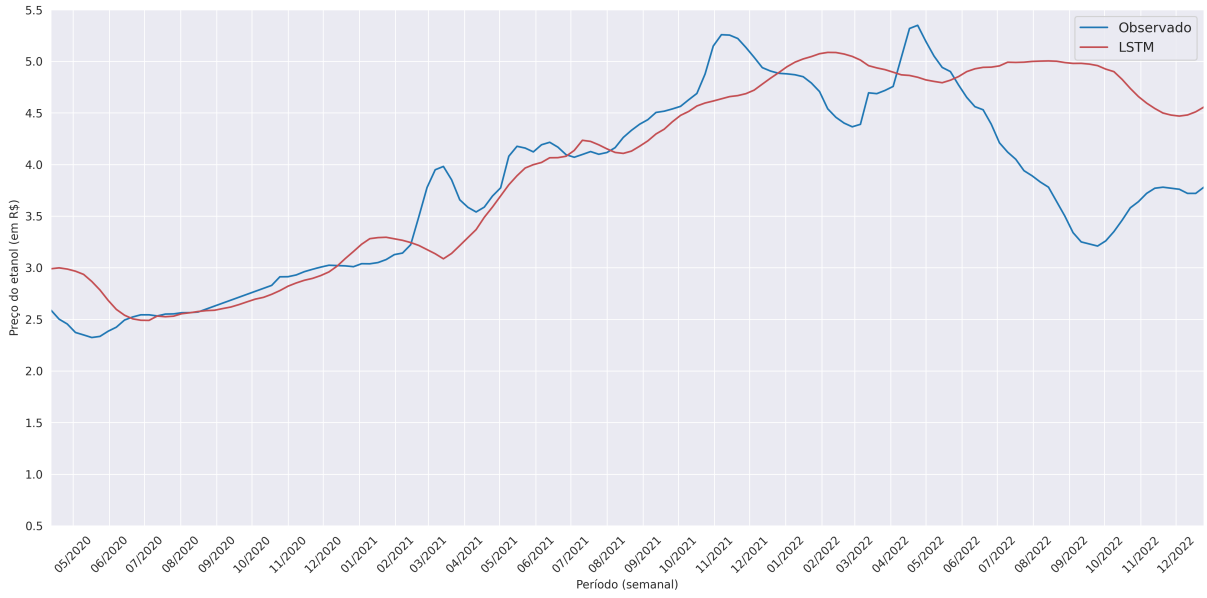
Figura 4.32: Preço do Etanol observado vs. preço previsto pelo modelo LSTM pré-pandemia, utilizando a base CCA.



Fonte: Elaboração própria.

Para o período pandêmico ilustrado no gráfico (Figura 4.33), o modelo LSTM acompanhou a tendência geral de alta e as flutuações no preço do etanol, embora com um certo atraso e suavização das variações mais bruscas. Observam-se discrepâncias entre os valores observados e os valores projetados ao longo de todo o período. As métricas de desempenho para esse recorte temporal indicaram um MAPE de 10.91%, MAE de 0.399 e MSE de 0.342.

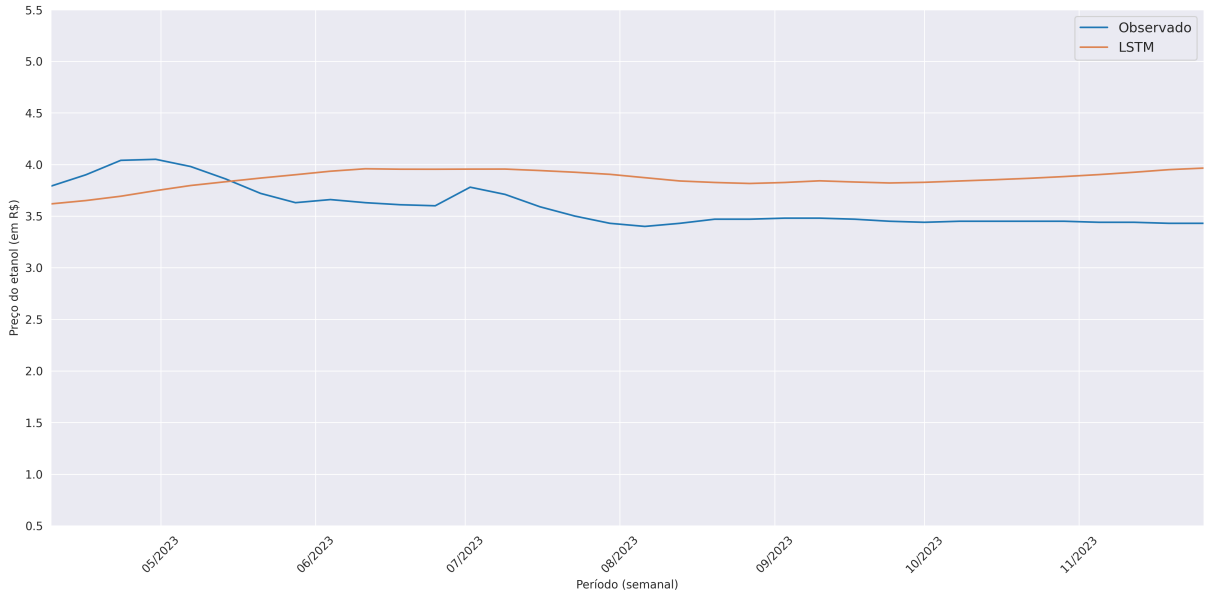
Figura 4.33: Preço do Etanol observado vs. preço previsto pelo modelo LSTM durante a pandemia, utilizando a base CCA.



Fonte: Elaboração própria.

No ano de 2023, os preços apresentaram maior estabilidade e menor flutuação, porém o modelo superestimou os valores, conforme ilustrado na Figura 4.34. Durante esse período, o modelo obteve um MAPE de 4.88%, MAE de 0.179 e MSE de 0.056.

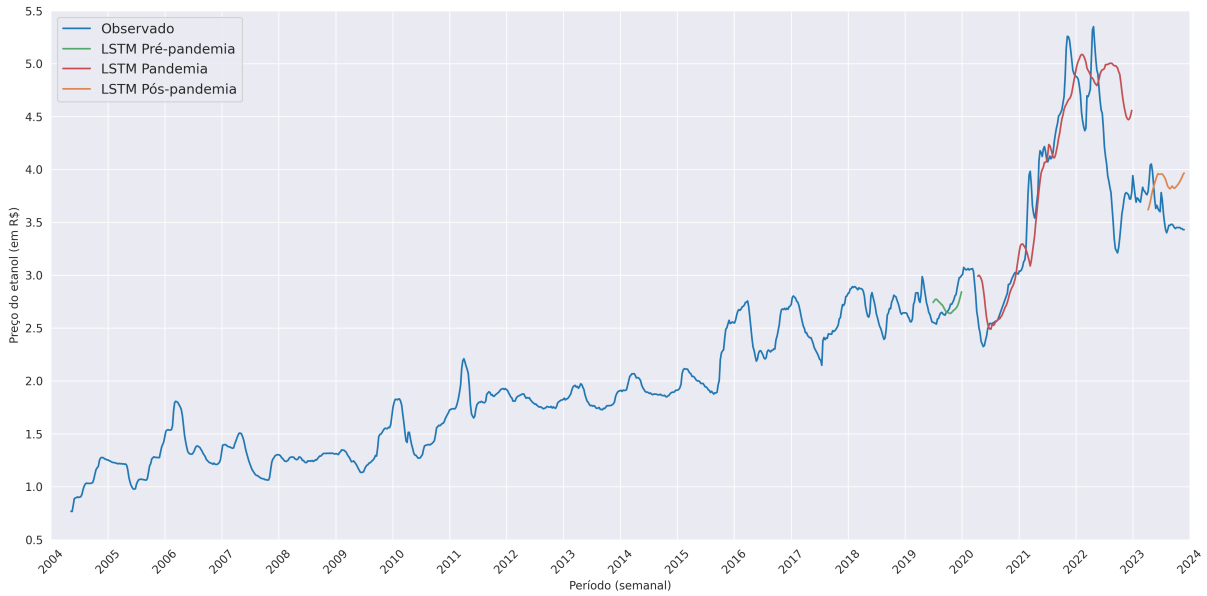
Figura 4.34: Preço do Etanol observado vs. preço previsto pelo modelo LSTM pós-pandemia, utilizando a base CCA.



Fonte: Elaboração própria.

Em suma, os resultados dos três períodos para o modelo nessa base encontram-se sintetizados na Figura 4.35.

Figura 4.35: Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto total, utilizando a base CCA.



Fonte: Elaboração própria.

A Tabela 4.7 apresenta as métricas de erro do modelo LSTM para os períodos pré-pandemia, pandemia e pós-pandemia. Observa-se que, no período pré-pandêmico, o modelo obteve um desempenho satisfatório, com MAPE de 4.48%, MAE de 0.1214 e MSE de 0.0194, indicando boa aderência entre os valores observados e previstos. Durante a pandemia, os erros aumentaram, com MAPE de 10.91%, MAE de 0.3991 e MSE de 0.3422, refletindo a dificuldade do modelo em capturar os choques abruptos no mercado de combustíveis durante esse período atípico. No período pós-pandêmico, apesar da melhoria em relação à pandemia, os erros permaneceram superiores aos observados no período pré-crise, com MAPE de 9.74%, MAE de 0.3449 e MSE de 0.1313. Isso indica que, embora o modelo tenha conseguido se ajustar melhor às condições de mercado no contexto pós-crise, ele ainda apresenta limitações para prever com precisão os preços em cenários marcados por instabilidade.

Tabela 4.7: Métricas de erro do modelo LSTM utilizando base de dados CCA nos diferentes períodos.

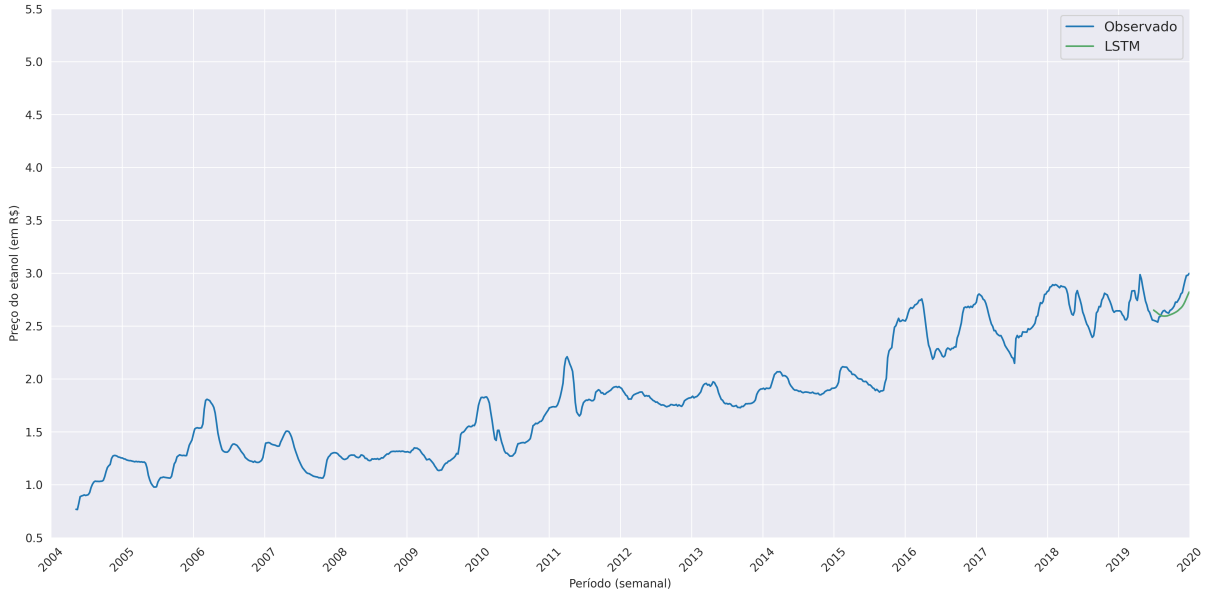
Período	**MAPE (%)**	**MAE**	**MSE**
Pré-pandemia	4.48	0.1214	0.0194
Pandemia	10.91	0.3991	0.3422
Pós-pandemia	9.74	0.3449	0.1313

Fonte: Elaboração própria.

4.4.2 Base CR-ATR

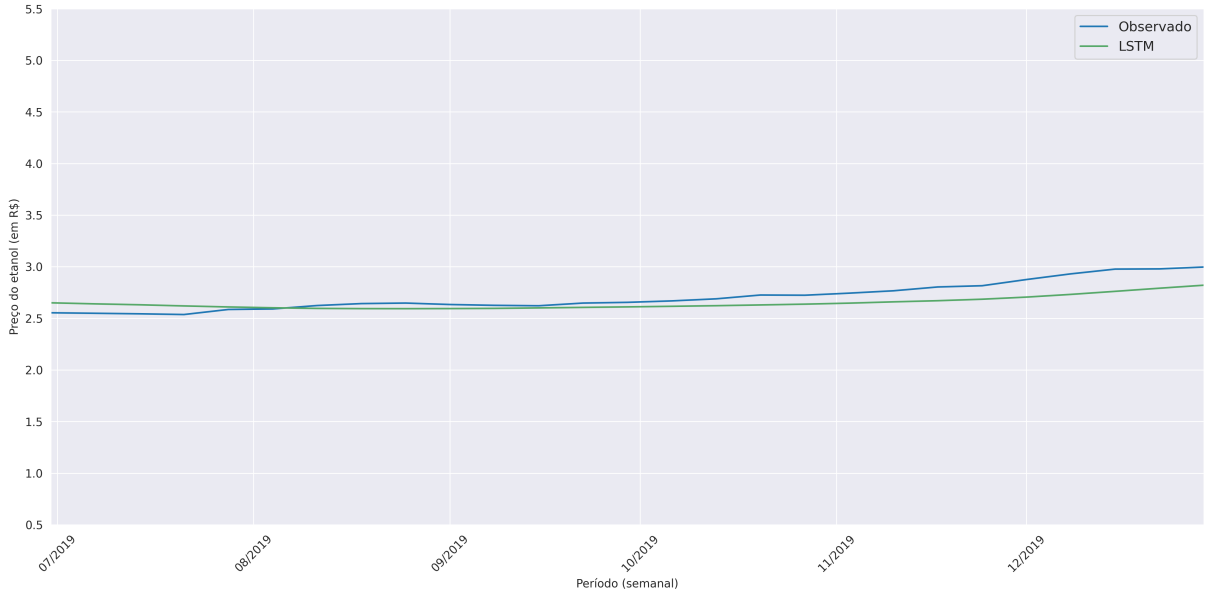
Conforme ilustrado na Figura 4.36, o modelo LSTM apresentou bom desempenho na previsão dos preços do etanol no período de teste pré-pandemia, acompanhando de forma consistente os valores observados e capturando adequadamente as tendências e variações sazonais. A previsão indicou uma leve tendência de alta no início de 2020. As métricas de avaliação reforçam a qualidade do ajuste, com MAPE de 3.22%, MAE de 0.089 e MSE de 0.011. A Figura 4.37 apresenta um recorte específico do período de teste, permitindo uma visualização mais detalhada da aproximação entre os valores previstos e observados.

Figura 4.36: Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto de teste, utilizando a base CR-ATR.



Fonte: Elaboração própria.

Figura 4.37: Preço do Etanol observado vs. preço previsto pelo modelo LSTM pré-pandemia, utilizando a base CR-ATR.

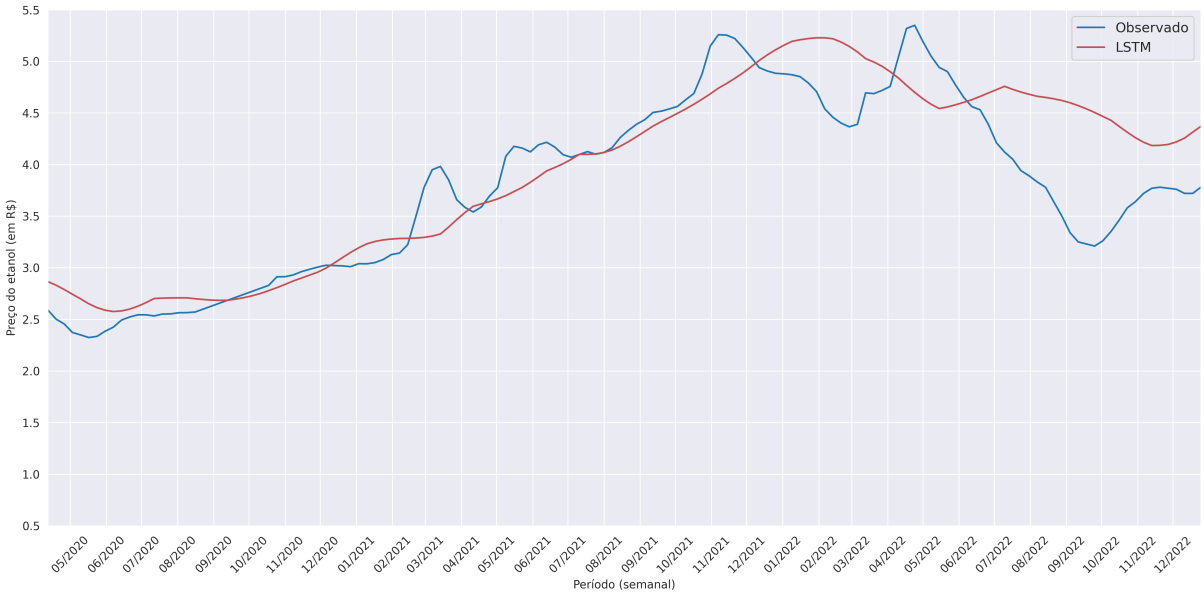


Fonte: Elaboração própria.

Para o período pandêmico ilustrado na Figura 4.38 (a partir de maio de 2020), o modelo LSTM acompanhou a tendência de alta do preço do etanol até meados de 2021, embora com algumas diferenças em relação aos valores observados. A partir do início de 2022, os preços observados apresentaram uma queda significativa, enquanto o modelo inicialmente continuou a projetar uma tendência de alta antes de também iniciar um

declínio, mantendo-se acima dos valores observados no final do período. As métricas de desempenho para um determinado recorte temporal indicaram um MAPE de 8.81%, MAE de 0.329 e MSE de 0.205.

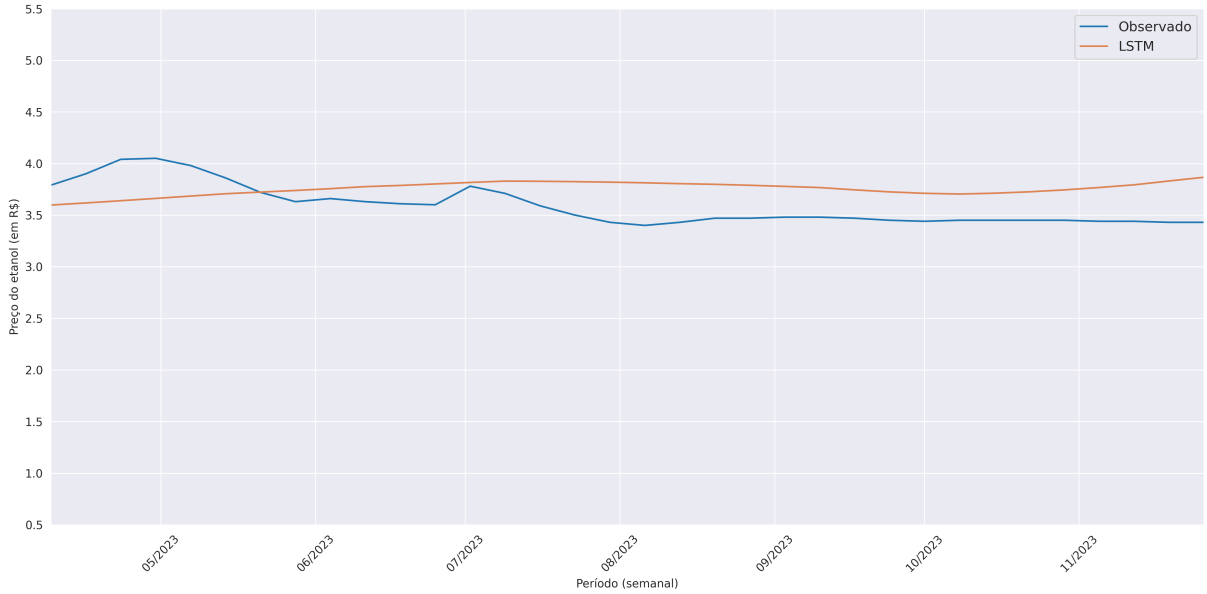
Figura 4.38: Preço do Etanol observado vs. preço previsto pelo modelo LSTM durante a pandemia, utilizando a base CR-ATR.



Fonte: Elaboração própria.

Em 2023, os preços apresentaram maior estabilidade e menores flutuações, mas o modelo LSTM continuou a subestimar os valores observados, com variações projetadas de menor amplitude, como mostrado na Figura 4.39. Para este período, o valor de MAPE foi de 7.43%, MAE de 0.265 e MSE de 0.082.

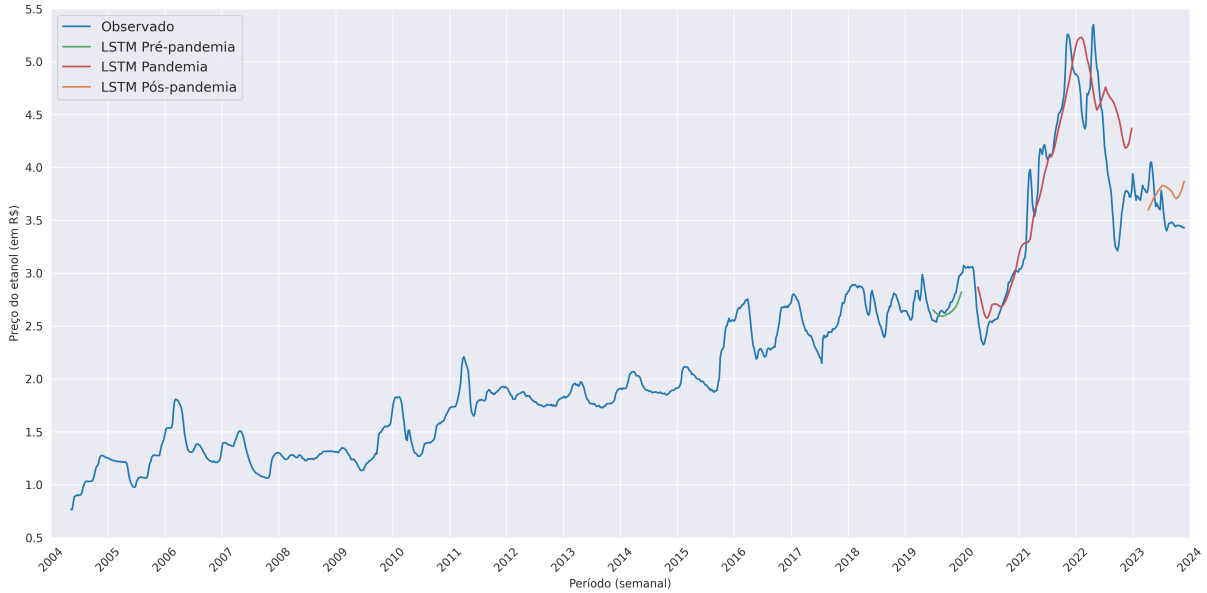
Figura 4.39: Preço do Etanol observado vs. preço previsto pelo modelo LSTM pós-pandemia, utilizando a base CR-ATR.



Fonte: Elaboração própria.

De forma sintética, a Figura 4.40 ilustra os resultados do modelo para os três períodos em questão nesta base.

Figura 4.40: Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto total, utilizando a base CR-ATR.



Fonte: Elaboração própria.

A Tabela 4.8 apresenta as métricas de erro do modelo LSTM para os períodos pré-pandemia, pandemia e pós-pandemia. Observa-se que, no período pré-pandêmico, o modelo apresentou um desempenho satisfatório, com MAPE de 3.22%, MAE de 0.0895 e

MSE de 0.0113, indicando boa aderência entre os valores observados e previstos. Durante a pandemia, houve um aumento nos erros, com MAPE de 8,81%, MAE de 0,3286 e MSE de 0,2054, refletindo a dificuldade do modelo em capturar choques abruptos no mercado de combustíveis nesse período atípico. No período pós-pandêmico, embora os indicadores tenham melhorado em relação à pandemia, os erros ainda foram superiores ao período pré-crise, com MAPE de 7.43%, MAE de 0.2645 e MSE de 0.0816. Isso indica que, apesar de o modelo ter conseguido se ajustar melhor às condições de mercado no contexto pós-crise, ele ainda enfrenta limitações para prever com precisão os preços em cenários marcados por instabilidade.

Tabela 4.8: Métricas de erro do modelo LSTM utilizando a base CR-ATR, nos diferentes períodos.

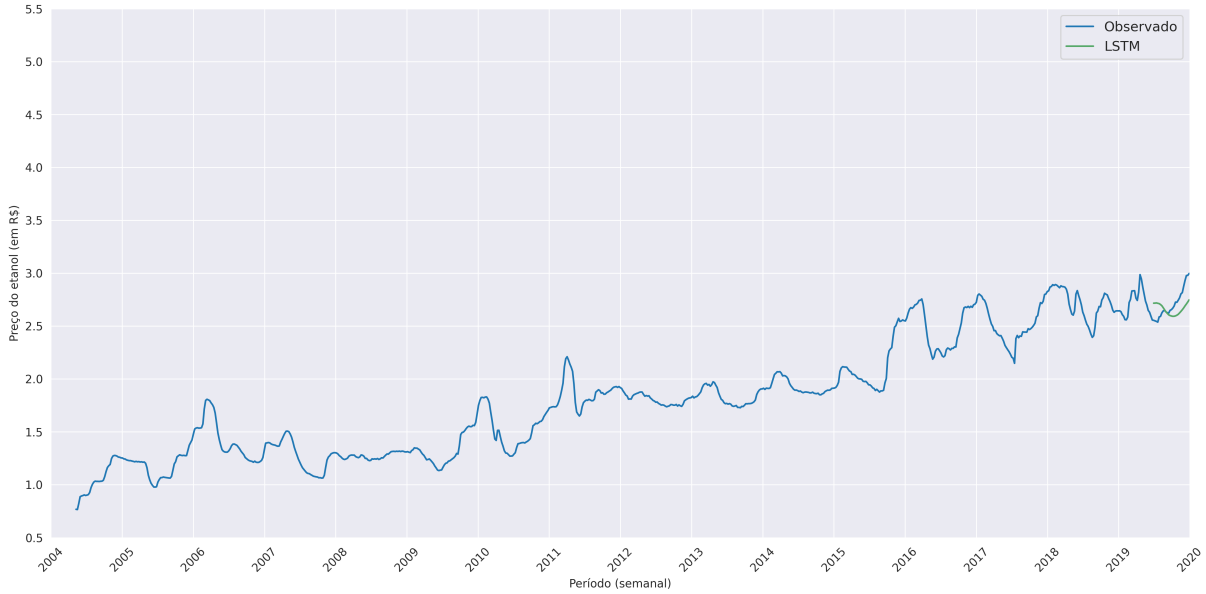
Período	**MAPE (%)**	**MAE**	**MSE**
Pré-pandemia	3.22	0.0895	0.0113
Pandemia	8.81	0.3286	0.2054
Pós-pandemia	7.43	0.2645	0.0816

Fonte: Elaboração própria.

4.4.3 Base CR-AC

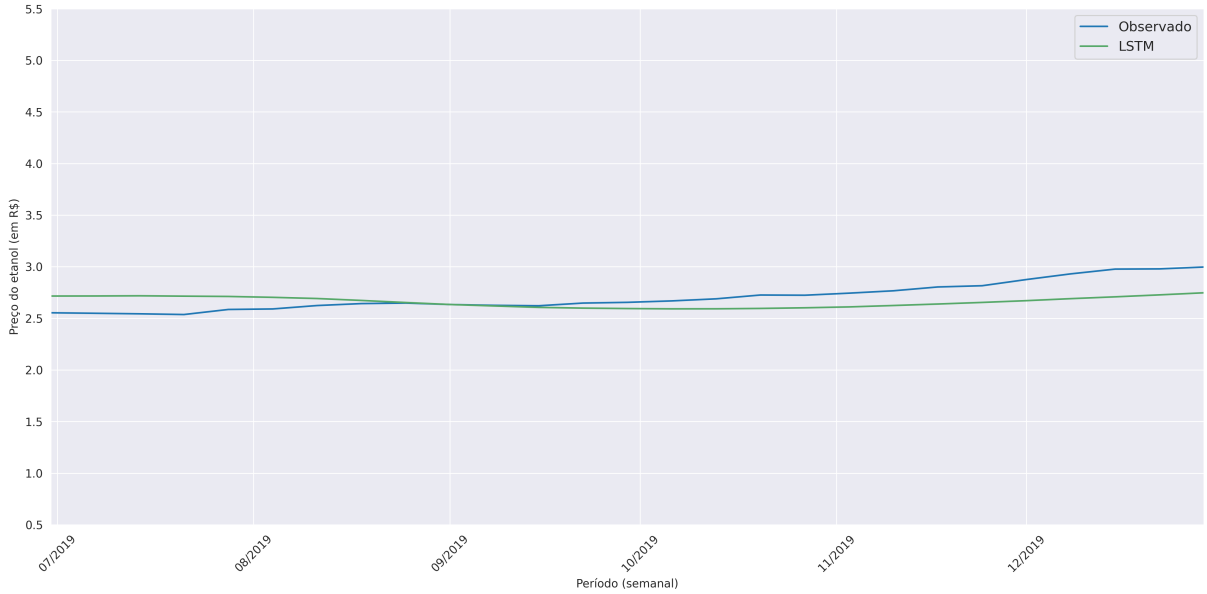
Conforme ilustrado na Figura 4.41, o modelo LSTM apresentou bom desempenho na previsão dos preços do etanol no período de teste pré-pandemia, acompanhando de forma consistente os valores observados e capturando adequadamente as tendências e variações sazonais. A previsão indica uma leve tendência de alta no início de 2020. As métricas de avaliação reforçam a qualidade do ajuste: MAPE de 4.57%, MAE de 0.126 e MSE de 0.022. A Figura 4.42 apresenta um recorte específico do período de teste, permitindo uma visualização mais detalhada da aproximação entre os valores previstos e observados.

Figura 4.41: Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto de teste, utilizando a base CR-AC.



Fonte: Elaboração própria.

Figura 4.42: Preço do Etanol observado vs. preço previsto pelo modelo LSTM pré-pandemia, utilizando a base CR-AC.

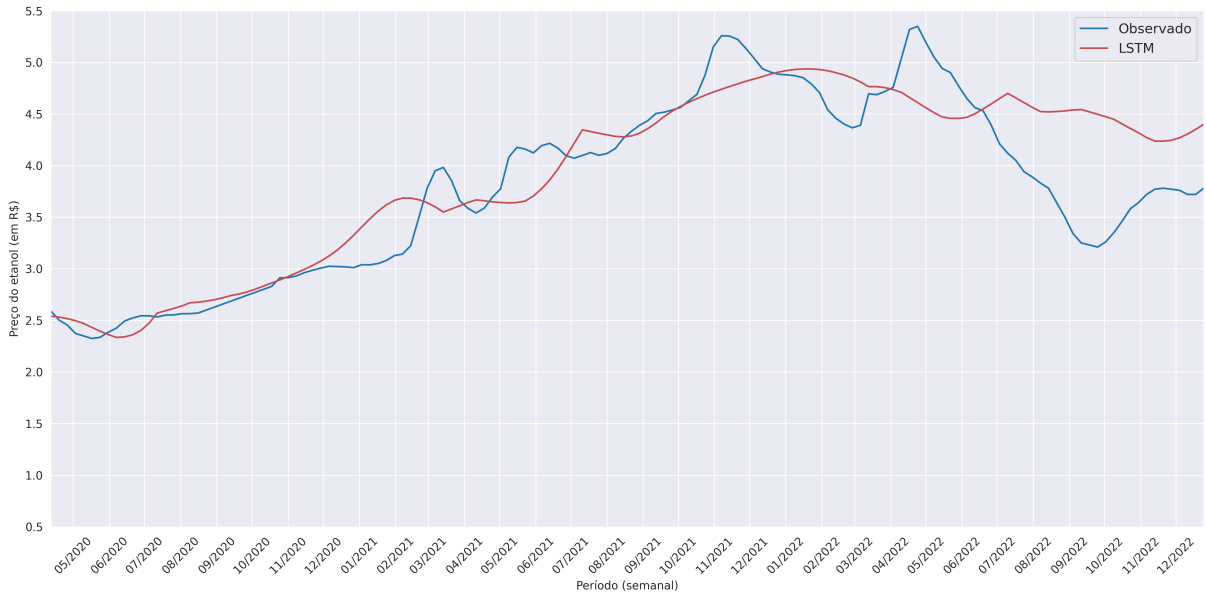


Fonte: Elaboração própria.

Para o período pandêmico ilustrado na Figura 4.43 (a partir de maio de 2020), durante a fase de recuperação (aproximadamente de meados de 2020 até meados de 2021), o modelo LSTM acompanhou relativamente bem a tendência de alta do preço do etanol. No entanto, a partir de abril de 2022, enquanto o modelo inicialmente continuou a projetar uma tendência de alta, os preços observados apresentaram uma queda significativa. As

métricas de desempenho para um determinado recorte temporal indicaram um MAPE de 8.17%, MAE de 0.305 e MSE de 0.187.

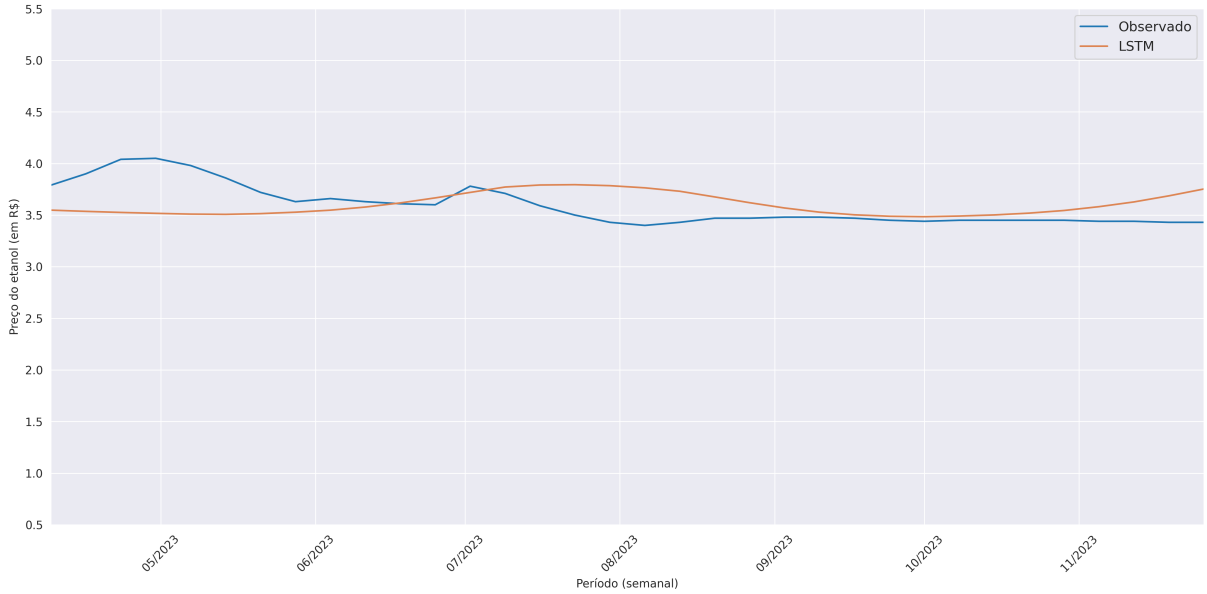
Figura 4.43: Preço do Etanol observado vs. preço previsto pelo modelo LSTM durante a pandemia, utilizando a base CR-AC.



Fonte: Elaboração própria.

Em 2023, com a estabilização dos preços e a redução das flutuações, o modelo ainda apresentou certa tendência a subestimar os valores, projetando variações de menor magnitude em comparação aos dados observados, como mostrado na Figura 4.44. Nesse ano, as métricas de desempenho indicaram um MAPE de 5.18%, MAE de 0.189 e MSE de 0.058.

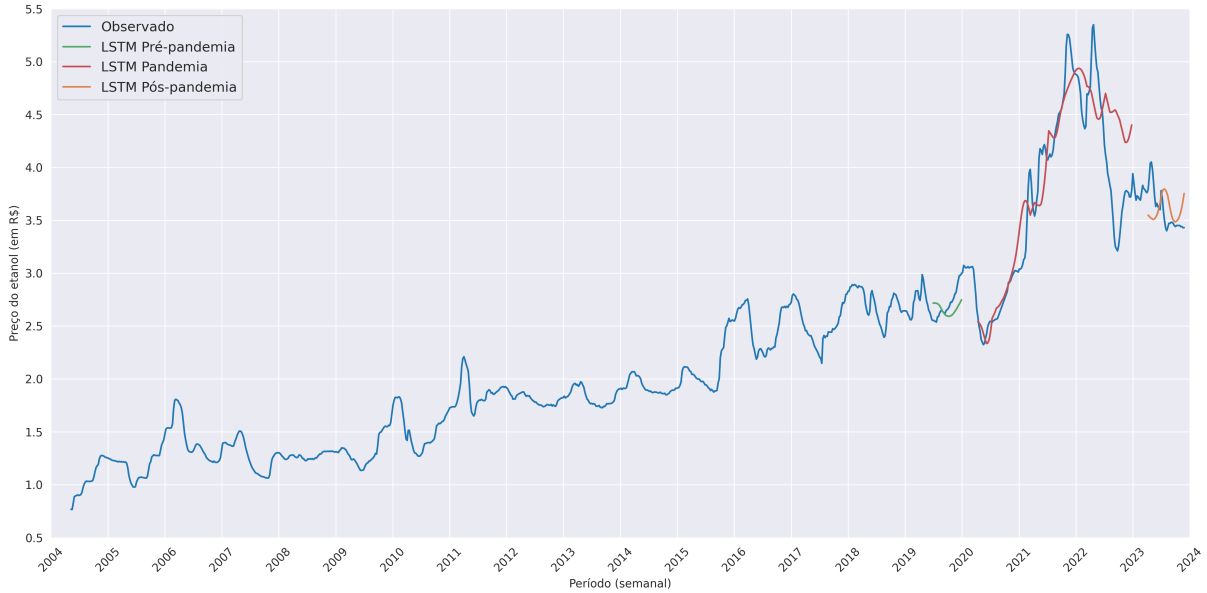
Figura 4.44: Preço do Etanol observado vs. preço previsto pelo modelo LSTM pós-pandemia, utilizando a base CR-AC.



Fonte: Elaboração própria.

Em síntese, conforme a Figura 4.45, encontram-se os resultados do modelo para os três períodos analisados sobre esta base de dados.

Figura 4.45: Preço do Etanol observado vs. preço previsto pelo modelo LSTM no conjunto total, utilizando a base CR-AC.



Fonte: Elaboração própria.

A Tabela 4.9 apresenta as métricas de erro do modelo LSTM nos períodos pré-pandemia, pandemia e pós-pandemia. Durante o período pré-pandêmico, o modelo obteve um desempenho bastante satisfatório, com MAPE de 4.95%, MAE de 0.1362 e MSE de

0.0251. Esses resultados indicam uma boa correspondência entre os preços previstos e os observados, com um erro relativamente baixo.

Contudo, com a chegada da pandemia, o modelo teve dificuldades em acompanhar os choques abruptos no mercado de combustíveis, refletindo-se em um aumento nos erros. No período pandêmico, as métricas foram de MAPE de 7.89%, MAE de 0.2947 e MSE de 0.1681, o que evidenciou a incapacidade do modelo de capturar adequadamente as flutuações imprevisíveis daquele período.

Após a pandemia, o modelo se ajustou de maneira mais eficaz às condições do mercado, mas os resultados ainda ficaram aquém do desempenho observado no período pré-crise. O MAPE foi de 4.88%, o MAE de 0.1789 e o MSE de 0.0564, refletindo uma recuperação, mas com erros ainda superiores ao que se esperava antes da crise. Isso sugere que, embora o modelo tenha se adaptado melhor ao cenário pós-pandemia, ele continua enfrentando desafios para prever com alta precisão os preços em um contexto instável.

Tabela 4.9: Métricas de erro do modelo LSTM utilizando a base CR-AC, nos diferentes períodos.

Período	**MAPE (%)**	**MAE**	**MSE**
Pré-pandemia	4.57	0.126	0.022
Pandemia	8.17	0.305	0.187
Pós-pandemia	5.18	0.189	0.058

Fonte: Elaboração própria.

4.5 Comparação e discussão das métricas

A Tabela 4.10 apresenta os resultados da avaliação de dois modelos de previsão, *Random Forest* (RF) e Long Short-Term Memory (LSTM), aplicados a diferentes seleções de variáveis (Base CCA, Base CR-ATR e Base CR-AC) em três períodos distintos: pré-pandemia, pandemia e pós-pandemia. O desempenho dos modelos é quantificado por meio de três métricas de erro: Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) e Mean Squared Error (MSE). Uma análise detalhada dos resultados em cada período e para cada modelo se faz necessária para identificar as abordagens mais eficazes.

Tabela 4.10: Resultados de MAPE, MAE e MSE para os Modelos RF e LSTM nas diferentes bases de dados e períodos.

Modelo	Seleção de Variáveis	Período	MAPE (%)	MAE	MSE
RF	Base CCA	Pré-pandemia	2.18	0.0589	0.00710
	Base CR-ATR	Pré-pandemia	**1.70**	**0.0473**	**0.00436**
	Base CR-AC	Pré-pandemia	1.97	0.0536	0.00444
	Base CCA	Pandemia	22.26	**0.9349**	**1.3415**
	Base CR-ATR	Pandemia	**22.24**	0.9378	1.3793
	Base CR-AC	Pandemia	23.73	0.9974	1.5370
	Base CCA	Pós-pandemia	**21.25**	**0.7833**	**0.6750**
	Base CR-ATR	Pós-pandemia	22.66	0.8348	0.7581
	Base CR-AC	Pós-pandemia	25.42	0.9334	0.9122
LSTM	Base CCA	Pré-pandemia	4.48	0.1214	0.0194
	Base CR-ATR	Pré-pandemia	**3.22**	**0.0895**	**0.0113**
	Base CR-AC	Pré-pandemia	4.57	0.1258	0.0219
	Base CCA	Pandemia	10.91	0.3991	0.3422
	Base CR-ATR	Pandemia	8.81	0.3286	0.2054
	Base CR-AC	Pandemia	**8.17**	**0.3052**	**0.1872**
	Base CCA	Pós-pandemia	9.74	0.3449	0.1313
	Base CR-ATR	Pós-pandemia	7.43	0.2645	0.0816
	Base CR-AC	Pós-pandemia	**5.18**	**0.1893**	**0.0578**

Fonte: Elaboração própria.

No período pré-pandêmico, caracterizado por maior estabilidade estrutural da série, ambos os modelos apresentaram desempenho satisfatório, com baixos valores de erro em todas as configurações de atributos. Entretanto, o *Random Forest* (RF) obteve resultados consistentemente superiores aos do LSTM. Destaca-se o desempenho do RF com o Conjunto Reduzido – Variante ATR (CR-ATR), que apresentou os menores erros do estudo nesse período (MAPE = 1.70%, MAE = 0.0473, MSE = 0.00436). Esse resultado sugere que, em contextos de menor volatilidade e com padrões históricos relativamente estáveis, modelos baseados em árvores e particionamento do espaço de atributos conseguem capturar de forma eficiente relações não lineares de curto e médio prazo, sem a necessidade de modelar dependências temporais complexas. O LSTM, embora também tenha apresentado bom desempenho, exibiu erros sistematicamente superiores, indicando que sua maior capacidade de modelagem temporal não foi plenamente explorada em um regime de baixa instabilidade.

Durante o período pandêmico, observa-se uma mudança estrutural clara no comportamento dos modelos. A elevação acentuada dos erros em ambos os casos reflete o caráter exógeno e abrupto dos choques de mercado, que alteraram drasticamente os padrões históricos de preços. Nesse contexto, o LSTM demonstrou maior resiliência que o RF. Utilizando o Conjunto Reduzido – Variante Açúcar (CR-AC), o LSTM atingiu MAPE de 7.89%, enquanto o RF ultrapassou 23% na mesma configuração. A queda acentuada de desempenho do RF indica que modelos baseados em partições fixas do espaço de atributos apresentam dificuldade em extrapolar para regimes não observados no treinamento. Já o LSTM, por incorporar mecanismos de memória e modelar dependências temporais de longo prazo, mostrou maior capacidade de adaptação a mudanças abruptas na dinâmica da série.

No período pós-pandêmico, verifica-se recuperação parcial da capacidade preditiva de ambos os modelos, embora os níveis de erro não retornem ao patamar pré-2020, sugerindo que o mercado não voltou integralmente à estrutura anterior, mas sim a um novo regime. O LSTM manteve desempenho superior ao RF, novamente com destaque para o CR-AC (MAPE = 4.88%, MAE = 0.1789, MSE = 0.0564). Esse comportamento reforça a hipótese de que o modelo recorrente consegue se ajustar melhor a transições de regime e a novas dinâmicas de mercado. O RF, por outro lado, manteve erros elevados (MAPE $>$ 21%), evidenciando limitação em cenários onde a estrutura estatística da série se altera de forma persistente.

A análise conjunta dos três períodos revela um padrão consistente: o RF apresenta vantagem em cenários de estabilidade estrutural, enquanto o LSTM se mostra mais adequado em ambientes de não estacionariedade e choques externos. A robustez do LSTM associada ao conjunto de variáveis relacionadas ao mercado de açúcar indica que essa cadeia produtiva atua como um importante transmissor de informação econômica para o preço do etanol, especialmente em contextos de instabilidade. O *Random Forest*, apesar de sua eficácia em condições estáveis, parece ser mais sensível a choques de mercado.

Em síntese, os resultados evidenciam uma diferença fundamental entre os modelos: o RF opera de forma mais eficiente quando a série mantém regularidade histórica, ao passo que o LSTM apresenta maior capacidade de generalização diante de mudanças estruturais. Enquanto o *Random Forest* se destacou no período pré-pandêmico, a capacidade do LSTM de mitigar os efeitos da pandemia e manter um desempenho relativamente melhor no período pós-pandêmico o torna uma escolha mais promissora para a modelagem desta série temporal em um ambiente de mercado dinâmico.

4.6 Visualização de Resultados por Dashboard

Como desdobramento aplicado dos resultados obtidos, as previsões geradas pelo modelo LSTM, utilizando a base CR-AC, foram integradas a um *dashboard* interativo desenvolvido na plataforma *Microsoft Power BI Desktop*. O fluxo de dados consistiu na importação da base histórica original em formato *.csv* e dos resultados preditivos, previamente processados e salvos em *Python* utilizando a biblioteca *Pandas*, também em formato *.csv*. Dentro do ambiente *Power BI*, os dados passaram por etapas de tratamento para permitir a utilização simultânea de filtros em múltiplos gráficos. Foram criadas colunas e medidas adicionais para calcular e exibir a imprecisão nominal (Preço Observado - Previsão) e a imprecisão percentual, permitindo uma auditoria precisa do erro semanalmente.

O painel foi estruturado em duas páginas que priorizam a interatividade por meio de segmentadores de dados. Na primeira página (Figura 4.46), o usuário pode filtrar as séries por períodos específicos (Pré-Pandemia, Pandemia e Pós-Pandemia) ou selecionar intervalos de datas personalizados. As ferramentas visuais incluem gráficos de área para a sobreposição das séries e tabelas detalhadas. Um diferencial técnico é o uso de cartões de indicadores (KPI) dinâmicos, que se ajustam automaticamente ao intervalo selecionado para exibir métricas como a Média do Preço Observado, Média da Previsão LSTM e a Média de Imprecisão.

Figura 4.46: Página 1: Comparação entre Séries Históricas e Projeções

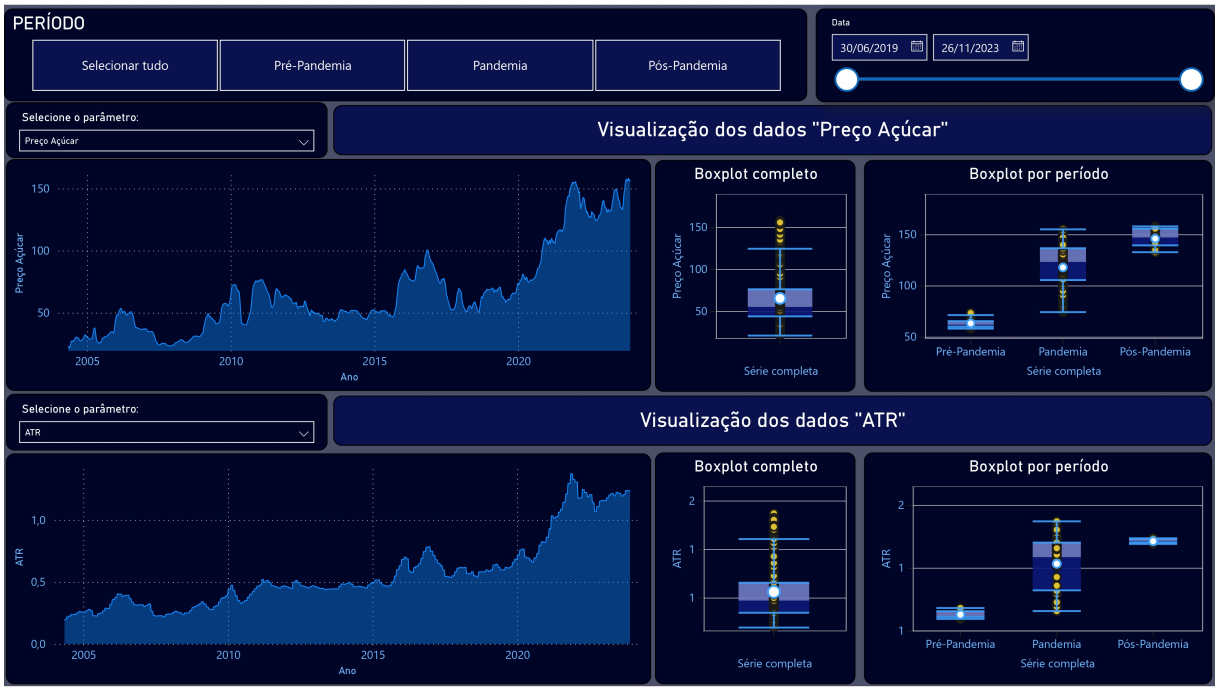


Data (dd/mm/aaaa)	Preço Observado	Previsão (LSTM)	Imprecisão (obs-prev)	Imprecisão (%)
30/06/2019	2,55	2,72	-0,16	-6,35%
07/07/2019	2,55	2,72	-0,17	-6,58%
14/07/2019	2,54	2,72	-0,17	-6,86%
21/07/2019	2,54	2,72	-0,18	-7,02%
28/07/2019	2,59	2,71	-0,13	-4,88%
04/08/2019	2,59	2,70	-0,11	-4,34%
11/08/2019	2,62	2,69	-0,07	-2,60%
18/08/2019	2,64	2,67	-0,03	-1,15%
25/08/2019	2,65	2,65	-0,01	-0,20%
01/09/2019	2,63	2,63	0,00	-0,00%
08/09/2019	2,63	2,62	0,01	0,23%
15/09/2019	2,62	2,61	0,02	0,60%
22/09/2019	2,65	2,60	0,05	1,84%
29/09/2019	2,66	2,59	0,06	2,26%
06/10/2019	2,67	2,59	0,08	2,88%
13/10/2019	2,69	2,59	0,10	3,59%
20/10/2019	2,73	2,60	0,13	4,77%
27/10/2019	2,72	2,60	0,12	4,47%
03/11/2019	2,74	2,61	0,13	4,85%
10/11/2019	2,77	2,62	0,14	5,15%
17/11/2019	2,80	2,64	0,17	5,89%
24/11/2019	2,82	2,65	0,16	5,75%
01/12/2019	2,88	2,67	0,20	7,11%
08/12/2019	2,93	2,69	0,24	8,20%
15/12/2019	2,98	2,71	0,27	9,02%
22/12/2019	2,98	2,73	0,25	8,45%

Fonte: Elaborado pelo autor (2026).

A segunda página (Figura 4.47) é voltada à análise conjunta das variáveis preditivas e da variável *target*. As séries são apresentadas de forma comparativa através de gráficos de área, permitindo observar co-movimentos, defasagens visuais e possíveis relações estruturais entre os fatores explicativos e o preço do etanol. Para a análise de distribuição e variabilidade, foi incorporada a extensão *Box and Whisker*, que gera *boxplots* detalhados por período. Essa funcionalidade permite comparar visualmente como a dispersão e as medianas das variáveis preditoras mudaram sob diferentes regimes econômicos. Essa organização favorece uma leitura exploratória orientada por hipóteses, auxiliando o usuário a compreender como choques em variáveis como açúcar, ATR, câmbio ou safra se refletem na dinâmica do etanol.

Figura 4.47: Página 2: Visualização Comparativa de Variáveis



Fonte: Elaborado pelo autor (2026).

Assim, o *dashboard* cumpre dupla função: (i) instrumento de validação visual dos resultados do modelo e (ii) ferramenta de apoio à tomada de decisão. Ao permitir a exploração interativa das séries temporais, o painel amplia o alcance dos resultados da pesquisa, tornando-os acessíveis a usuários não especialistas e promovendo a aplicação prática das evidências obtidas. Dessa forma, o produto tecnológico desenvolvido consolida a dimensão aplicada do estudo, alinhando modelagem preditiva e usabilidade informacional.

O *dashboard* encontra-se disponível para acesso público no seguinte endereço eletrônico: `https://app.powerbi.com/links/BlWIjKLdf9?ctid=fe8787bc-c914-4665-8547-268e15cb0d9a&pbi_source=linkShare`, possibilitando a replicação das análises apresentadas e a exploração autônoma das informações por parte do leitor.

Capítulo

5

Conclusões

A presente dissertação cumpriu seu objetivo geral ao avaliar e comparar o desempenho dos modelos *Random Forest* (RF) e *Long Short-Term Memory* (LSTM) na previsão dos preços do etanol. A pesquisa demonstrou que, embora distintas em suas abordagens, ambas as técnicas de Aprendizagem de Máquina oferecem ferramentas relevantes para mitigar as incertezas do setor energético brasileiro, com aplicabilidades que variam conforme o cenário econômico.

No que tange à análise da evolução histórica e relação com variáveis, a investigação confirmou que a inclusão de dados setoriais foi determinante para a acurácia dos modelos. O índice ATR destacou-se como preditor relevante em períodos de estabilidade, enquanto o preço do açúcar mostrou-se fundamental para a resiliência dos modelos em momentos de crise, evidenciando a interdependência estrutural entre as *commodities* energéticas e agrícolas, especialmente durante a pandemia e no período subsequente.

Quanto à implementação e avaliação dos modelos, os resultados revelaram uma distinção clara entre precisão em condições estáveis e robustez em cenários de choque. No período pré-pandêmico, caracterizado pela estabilidade, o modelo *Random Forest* apresentou desempenho superior, alcançando um notável *Mean Absolute Percentage Error* (MAPE) de 1,70% na base com ATR, superando o MAPE de 3,22% do LSTM na mesma configuração. Contudo, a arquitetura do RF mostrou-se frágil diante de rupturas estruturais: durante e após a pandemia, seus erros saltaram para patamares superiores a 20%, evidenciando dificuldades de adaptação a novos regimes de dados.

Em contrapartida, o modelo LSTM destacou-se pela capacidade de generalização temporal e gestão de riscos. Embora não tenha sido o mais preciso em períodos de estabilidade, demonstrou maior robustez sob volatilidade elevada. Utilizando a base com preço do açúcar, o LSTM limitou o erro a 7,89% durante o período pandêmico e apresentou recuperação no pós-pandemia (MAPE de 4,88%), evidenciando melhor adaptação às mudanças estruturais do mercado.

A incorporação dos resultados em um *dashboard* interativo amplia a contribuição prática da pesquisa, conferindo ao trabalho um caráter analítico e exploratório, ao possibilitar a navegação fluida entre previsões geradas pelo modelo LSTM e os valores observados, bem como a seleção e comparação de diferentes combinações de variáveis preditivas e da variável *target*. A disposição dos elementos visuais e dos controles interativos favorece a realização de comparações em recortes temporais distintos, a detecção de padrões sazonais

e de rupturas associadas a eventos externos, e a investigação de relações de correlação e de causalidade potencial entre os preditores e o comportamento da série.

A inclusão de tabelas numéricas complementares e de filtros pré-definidos permite verificar consistência entre representações gráficas e valores brutos, aumentando a transparência e a reprodutibilidade das análises. Como resultado, os achados metodológicos deixam de estar restritos a relatórios estáticos e passam a ser acessíveis por meio de uma interface que facilita a interpretação por públicos técnicos e não técnicos. Essa integração entre rigor analítico e usabilidade prática amplia a aplicabilidade dos resultados da pesquisa, oferecendo subsídios concretos para a formulação de políticas públicas, decisões de mercado e investigações futuras.

Por fim, a comparação e discussão dos resultados fornecem subsídios claros para o setor. De forma geral, conclui-se que o *Random Forest* é adequado para análises em contextos de estabilidade econômica, enquanto o LSTM se mostra mais apropriado para cenários de incerteza e transições de regime. Assim, a principal contribuição do estudo reside na evidência de que modelos baseados em redes neurais recorrentes oferecem maior segurança para planejamento estratégico em mercados voláteis, reforçando o papel do etanol como componente estratégico da matriz energética nacional.

Como perspectivas futuras, destacam-se duas direções principais. A primeira envolve a evolução do *dashboard* para um sistema de monitoramento em tempo real, integrando fluxos contínuos de dados de mercado, o que permitiria previsões dinâmicas e suporte imediato à tomada de decisão. A segunda refere-se ao avanço metodológico por meio da adoção de modelos híbridos, como ARIMA-LSTM ou abordagens *ensemble* combinando modelos lineares e não lineares, capazes de capturar simultaneamente componentes estruturais e padrões complexos. Essas extensões podem ampliar a robustez preditiva e a capacidade de adaptação a choques exógenos, consolidando uma linha de pesquisa voltada à inteligência analítica aplicada ao setor sucroenergético.

Referências

A. Mohamed A. Graves and G. Hinton. Speech recognition with deep recurrent neural networks. *ICASSP*, pages 6645–6649, 2013.

Asif Afzal, Abdul Aabid, Ambareen Khan, Sher Afghan Khan, Upendra Rajak, Tikendra Nath Verma, and Rahul Kumar. Response surface analysis, clustering, and random forest regression of pressure in suddenly expanded high-speed aerodynamic flows. *Aerospace Science and Technology*, 107:106318, 2020. ISSN 1270-9638. doi: https://doi.org/10.1016/j.ast.2020.106318. URL `https://www.sciencedirect.com/science/article/pii/S1270963820310002`.

Poonam Allawadi, Rajesh Kumar, Prashant Kumar, Anand Mohan Singh, Anand Kumar, Sandeep Singh, and Brijesh Singh. Machine learning-based models for the prediction of biofuel properties: A critical review. *Fuel*, 321:123658, 2022. doi: 10.1016/j.fuel.2022.123658.

ANP. Brazilian national agency for petroleum, natural gas and biofuels. série histórica do levantamento de preços, 2023. Avaiable online: `https://www.gov.br/anp` (accessed on 25 November 2023).

Alexandre Hild Aono. *Inovações em análises ômicas de cana-de-açúcar: desvendando caracteres de importância econômica*. PhD thesis, [sn], 2024.

Apeksha and Nilanjan Nandi. Covid-19 and economic instability: The indian experience. *Journal of Public Affairs*, 21(4):e2649, 2021.

J. Scott Armstrong and Fred Collopy. Long-range forecasting: from crystal ball to computer. *Wiley*, 1985.

Francisco Arroyo-Marioli, Jeetendra Khadan, Franziska Ohnsorge, and Takefumi Yamazaki. Forecasting industrial commodity prices: Literature review and a model suite. Policy Research Working Paper 10611, World Bank, November 2023.

Renewable Fuels Association. Annual ethanol production report 2023, 2023. URL `https://ethanolrfa.org`. Disponível em: `https://ethanolrfa.org`.

Richard Baldwin and Beatrice Weder di Mauro. Economics in the time of covid-19. *Available at SSRN 3563249*, 2020.

Eduardo de Castro Barrero and Leandro Santos Leão. Inteligência artificial aplicada na gestão de manutenção. 2022.

Valéria Delgado Bastos. Etanol, alcoolquímica e biorrefinarias. 2007.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise Reduction in Speech Processing*, pages 1–4, 2009.

Eduardo Bezerra. Introdução à aprendizagem profunda. *Artigo–31º Simpósio Brasileiro de Banco de Dados–SBBD2016–Salvador*, 2016.

Mudrakola Bhavani and Podila Mounika. A novel model selection framework for forecasting agricultural commodity prices using time series features and forecast horizons. *International Journal of Scientific Research in Science and Technology (IJSRST)*, 9: 134–144, 2022. doi: 10.32628/IJSRST229535.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

William Blyth, Rob Bradley, Derek Bunn, Cathryn Clarke, and Tom Wilson. Decarbonising the electricity sector: sensitivity of renewable and carbon capture and storage technologies to carbon prices. *Energy Policy*, 35(10):4754–4767, 2007. doi: 10.1016/j.enpol.2007.03.025.

Brad Boehmke and Brandon M. Greenwell. *Hands-On Machine Learning with R*. CRC Press, 2020. ISBN 9780367144427.

Benedict du Boulay. Inteligência artificial na educação e ética. *RE@ D–Revista de Educação a Distância e eLearning*, page e202301, 2023.

Facundo Bre, Juan Gimenez, and Víctor Fachinotti. Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings*, 158, 11 2017. doi: 10.1016/j.enbuild.2017.11.045.

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

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001b.

P. Brignoli, D. Ferraz, and R. R. Righi. Machine learning to predict grains futures prices. *Agricultural Economics*, 55(3):481–495, 2024. doi: 10.1111/agec.12828.

W. et al. Carvalho Junior. Regressão linear múltipla e modelo random forest para estimar densidade do solo. *Pesquisa Agropecuária Brasileira*, 51(9):1428–1437, 2016.

CEPEA. Center for advanced studies on applied economics, 2023. Avaiable online: `https://www.cepea.esalq.usp.br/en/cepea-1.aspx` (accessed on 25 November 2023).

Tianfeng Chai and Roland R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Seulki Chung. Modelling and forecasting energy market volatility using garch and machine learning approach. *arXiv preprint arXiv:2405.19849*, 2024.

CONAB. National supply company, 2023. Avaiable online: https://www.conab.gov.br (accessed on 25 November 2023).

German Creamer, Hamed Ghoddusi, and Nima Rafizadeh. Machine learning in energy economics and finance: A review. *SSRN Electronic Journal*, 01 2018. doi: 10.2139/ssrn.3270251.

D. R. Cutler, T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

Bruna Gomes de Lima and Guilherme Canuto da Silva. Protótipo virtual de um sistema automático de manufatura automotiva baseado em redes neurais artificiais. *Journal of Production and Automation (JPAUT) ISSN 2595-9573*, 3(2):21–36, 2020.

Dusan Drabik, Harry De Gorter, David R Just, and Govinda R Timilsina. The economics of brazil's ethanol-sugar markets, mandates, and tax exemptions. *American Journal of Agricultural Economics*, 97(5):1433–1450, 2015.

Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2001.

D. Dunham, J. Lewandrowski, J. Baker, J. Ballerene, G. M. Souza, and G. Wiltsee. Brazil's biofuel industry: opportunities and challenges for sugarcane ethanol. *USDA Economic Research Report*, (125):1–61, 2011a.

Fabrício Brollo Dunham, José Vitor Bomtempo, and Denise Lima Fleck. A estruturação do sistema de produção e inovação sucroalcooleiro como base para o proálcool. *Revista Brasileira de Inovação*, 10(1):35–72, 2011b.

EIA. U.s. energy information administration, 2024. Avaiable online: https://www.eia.gov (accessed on 25 January 2024).

EPE. Relatório anual de energia 2023, 2023. URL https://www.epe.gov.br. Disponível em: https://www.epe.gov.br.

Tatiana Escovedo and Adriano Koshiyama. *Introdução a Data Science: Algoritmos de Machine Learning e métodos de análise*. Casa do Código, 2020.

Alvaro Espitia, Aaditya Mattoo, and Roberta Rocha. Trade, production and organization in the covid-19 pandemic. *World Economy*, 46(1):7–33, 2023.

Donald E. Farrar and Robert R. Glauber. Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics*, 49(1):92–107, 1967.

Leandro Fleck, Maria Hermínia Ferreira Tavares, Eduardo Eyng, Andrieli Cristina Helmann, and Minéia Aparecida de Moares Andrade. Redes neurais artificiais: Princípios básicos. *Revista Eletrônica Científica Inovação e Tecnologia*, 1(13):47–57, 2016.

José Goldemberg. The brazilian biofuels industry. *Biotechnology for Biofuels*, 1:6, 2008. doi: 10.1186/1754-6834-1-6.

Dennis dos Santos Gomes. Inteligência artificial: conceitos e aplicações. *Revista Olhar Científico*, 1(2):234–246, 2010.

Daniele Grattarola and Cesare Alippi. Graph neural networks in tensorflow and keras with spektral [application notes]. *IEEE Computational Intelligence Magazine*, 16(1): 99–106, 2021.

Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, 2017.

Alice Gupta and Bob Zheng. Feature engineering for machine learning: Principles and techniques for data scientists. *O'Reilly Media*, 2020.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2nd edition, 2018a. URL `https://otexts.com/fpp2/`.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018b. URL `https://otexts.com/fpp2/`.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018c.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

IBGE. Brazilian institute of geography and statistics. sugarcane production, 2023. Avaiable online: `https://www.ibge.gov.br` (accessed on 25 November 2023).

Investing.com, 2023. Avaiable online: `https://br.investing.com` (accessed on 25 November 2023).

Ipea. Estudos do ipea (2022) identificaram coeficiente de variação de 18,7% nos preços paulistas entre 2010-2020, superando a volatilidade de commodities como soja (12,4%) e café (14,9%). 2022. Please provide full citation details.

IPEA. Institute for applied economic research, 2024. Avaiable online: `http://www.ipeadata.gov.br` (accessed on 25 January 2024).

Kumara Jati. Sugar price analysis in indonesia. *International Journal of Social Science and Humanity*, 3(4):369, 2013.

Harikrushna Jena, Bijay Kumar Panda, Chitrasen Pradhan, Jitendriya Nayak, Prasanta Kumar Mishra, and Haribandhu Das. Deep learning-based hybrid models for prediction of biofuel properties. *Fuel Processing Technology*, 223:106860, 2021. doi: 10.1016/j.fuproc.2021.106860.

Aleksandar Jovic, Kresimir Brkic, and Nikola Bogunovic. A review of feature selection methods with applications. *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1200–1205, 2015.

B. I. A. Kaipper, H. X. Corseuil, and V. A. Todeschini. Influência do etanol na solubilidade de hidrocarbonetos totais de petróleo em aquíferos contaminados por óleo diesel. *Anais do XII Congresso Brasileiro de Águas Subterrâneas*, 2011. Publicado em agosto de 2011. Artigo sobre contaminação por HTPs, CETESB, análise RBCA e técnicas como CG-DIC e Headspace.

Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, and Doheon Lee. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 26(4): 3–11, 2003.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.

Zsolt László Kovács. *Redes neurais artificiais*. Editora Livraria da Fisica, 2006.

Samantha Vieira Lanzelotti. Análise comparativa de ferramentas do aprendizado de máquina para predição da produção de gasolina e etanol. 2022.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

TPF et al. Lima. Death risk and the importance of clinical features in elderly patients with covid-19 using random forest algorithm. *Revista Brasileira de Saúde Materno Infantil*, 21(2):433–442, 2021.

Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv*, 2015.

André F.P. Lucena, Alexandre S. Szklo, and Roberto Schaeffer. Least-cost adaptation strategies for global climate change impacts on the brazilian electric power system. *Energy Policy*, 38:332–342, 2010. doi: 10.1016/j.enpol.2009.09.051.

S. N. Ludovico, R. M. Salgado, L. A. Beijo, E. C. Miguel, and M. L. Rezende. Previsão de preços de commodities via modelos de aprendizagem de máquina. *Sigmae*, 2022. Análise de previsão de preços de açúcar, boi, café, etanol, milho e soja com uso de machine learning.

N. Tavakoli M. Siami-Namini and A. Siami Namin. The performance of lstm and bilstm in forecasting time series. *IEEE Big Data*, pages 3285–3292, 2019.

Isaias C. Macedo, Joaquim E.A. Seabra, and João E.A.R. Silva. Greenhouse gases emissions in the production and use of ethanol from sugarcane in brazil: The 2005/2006 averages and a prediction for 2020. *Biomass and Bioenergy*, 32:582–595, 2008. doi: 10.1016/j.biombioe.2007.12.006.

Isaías de Carvalho Macedo and LA Horta. Avaliação da expanção da produção do etanol no brasil. *Centro de Gestão e Estudos Estratégicos. July*, 2004.

R. L. Manogna, Vijay Dharmaji, and S. Sarang. Enhancing agricultural commodity price forecasting with deep learning. *Applied Soft Computing*, 2025. Data de aceitação: 30 de maio de 2025.

Grzegorz Marcjasz, Jesus Lago, and Rafał Weron. Neural networks in day-ahead electricity price forecasting: Single vs. multiple outputs. *arXiv preprint arXiv:2008.08006*, 2020.

Donald W Marquardt. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12(3):591–612, 1970.

Larry Medsker and Lakhmi C Jain. *Recurrent neural networks: design and applications*. CRC press, 1999.

Ednaldo Michellon, Ana Aracelly Lima Santos, and Juliano Ricardo Alves Rodrigues. Breve descrição do proálcool e perspectivas futuras para o etanol produzido no brasil. 2008.

Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017.

Derek Monner and James A. Reggia. A hybrid lstm and encoder–decoder neural network for modeling sequences. *Neural Networks*, 33:76–89, 2012. doi: 10.1016/j.neunet.2012.05.014.

Adán Márquez, Carlos Vázquez, David De la Fuente, and Pablo Priore. A comparative study of lstm and gru models for learning ecg signals. *Applied Sciences*, 11(3):1236, 2021. doi: 10.3390/app11031236.

Christof Naumzik and Stefan Feuerriegel. Forecasting electricity prices with machine learning: Predictor sensitivity. *International Journal of Energy Sector Management*, 2020.

Marcos Fava Neves, Patrícia MILAN, Flávio R VALÉRIO, Vitor N MARQUES, Fábio G DELSIN, Vinícius CAMBAÚVA, Leticia F MARTINEZ, Marcelo MR MOREIRA, Sofia ARANTES, and Gabriel O TEIXEIRA. Etanol de milho: cenário atual e perspectivas para a cadeia no brasil. *Ribeirão Preto*, 2021.

Paulo Nocera Alves Junior, Isotilia Costa Melo, Rodrigo de Moraes Santos, Fernando Vinícius da Rocha, and José Vicente Caixeta-Filho. How did covid-19 affect green-fuel supply chain? - a performance analysis of brazilian ethanol sector. *Research in Transportation Economics*, 93:101137, 2022. ISSN 0739-8859. doi: https://doi.org/10.1016/j.retrec.2021.101137. URL `https://www.sciencedirect.com/science/article/pii/S0739885921001098`. Disruption of global and regional supply-chains in the aftermath of Covid-19 pandemic. Analyses and forecasts.

F. A. de Oliveira, C. N. Nobre, and L. E. Zárate. Applying artificial neural networks to prediction of stock price and improvement of the directional prediction index–case study of petr4, petrobras, brazil. *Expert systems with applications*, 40(18):7596–7606, 2013.

David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999a. doi: 10.1613/jair.614.

David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999b. doi: 10.1613/jair.614.

K O'Shea and R Nash. An introduction to convolutional neural networks. *arXiv*, 2015.

Peterson K Ozili and Thankom Arun. Spillover of covid-19: impact on the global economy. In *Managing inflation and supply chain disruptions in the global economy*, pages 41–61. IGI Global, 2023.

J. S. Paula, L. L. Teixeira, S. B. Rodrigues, T. Hickmann, J. M. Correa, and L. S. Ribeiro. Aplicação de técnicas de aprendizado de máquina e estatística na previsão da demanda de biocombustíveis. *Revista GeSec*, 13(4):2559–2572, 2022. doi: 10.7769/gesec.v13i4.1488.

M. Paula, W. Casaca, M. Colnago, José R da Silva, Kleber Oliveira, Mauricio A Dias, and R. Negri. Predicting energy generation in large wind farms: A data-driven study with open data and machine learning. *Inventions*, 8(5):126, 2023.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

Erik Augusto Piacente. Perspectivas do brasil no mercado internacional de etanol. *Campinas, SP: Unicamp*, 189, 2006.

Camila Proni. Desenvolvimento de modelos neurais e florestas aleatórias para estudo de tintas da indústria gráfica. 2022.

J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Soumik Ray, Achal Lama, Pradeep Mishra, Tufleuddin Biswas, Soumitra Sankar Das, and Bishal Gurung. An arima-lstm model for predicting volatile agricultural price series with random forest technique. *Applied Soft Computing*, 149:110939, 2023. doi: 10.1016/j.asoc.2023.110939.

Carlos Henrique Reis. Otimização de hiperparâmetros em redes neurais profundas. *Minas Gerais*, 2021.

Reuters. Ai leading to faster, cheaper oil production, executives say, 2025. `https://www.reuters.com/business/energy/ceraweek-ai-leading-faster-cheaper-oil-production-executives-say-2025-03-13/`.

Saulo Rodrigues Filho and Antonio José Juliani. Sustentabilidade da produção de etanol de cana-de-açúcar no estado de são paulo. *estudos avançados*, 27:195–212, 2013.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

Alcione Miranda dos Santos, José Manoel de Seixas, Basílio de Bragança Pereira, and Roberto de Andrade Medronho. Usando redes neurais artificiais e regressão logística na predição da hepatite a. *Revista Brasileira de Epidemiologia*, 8:117–126, 2005.

Antonio Carlos Santos and Roberto Aguiar Peixoto. Efeito da adição de etanol à gasolina na emissão de gases do efeito estufa. *Science & Engineering Journal*, 17:33–44, 2008.

Fernando A Santos, José H de Queiróz, Jorge L Colodette, Sergio A Fernandes, Valéria M Guimarães, and Sebastião T Rezende. Potencial da palha de cana-de-açúcar para produção de etanol. *Química nova*, 35:1004–1010, 2012.

Fernando Alves dos Santos. *Análise da aplicação da biomassa da cana como fonte de energia elétrica: usina de açúcar, etanol e bioeletricidade.* PhD thesis, Universidade de São Paulo, 2012a.

Jaqueline Zani dos Santos. Poder de mercado no varejo de etanol no estado de são paulo. 2012b.

M. F. A. Santos, J. T. Faria, J. M. M. Lima, and A. M. R. Júnior. Aplicações de algoritmos de inteligência artificial na simulação de processos e na predição de propriedades de biocombustíveis – uma revisão sistemática. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 44(12):100242, 2022. doi: 10.1016/j.joei.2022.100242.

Murat Sari, Serbay Duran, Huseyin Kutlu, Bulent Guloglu, and Zehra Atik. Various optimized machine learning techniques to predict agricultural commodity prices. *Neural Computing and Applications*, 36:11439–11459, 2024. doi: 10.1007/s00521-024-09679-x.

Fundação SEADE. São paulo lidera produção de etanol no país. Online, 2021. Disponível em: https://www.seade.gov.br/sao-paulo-lidera-producao-de-etanol-no-pais. Acessado em: 22/03/2024.

Fundação SEADE. Dados estatísticos do estado de são paulo, 2023. URL `https://www.seade.gov.br`. Disponível em: `https://www.seade.gov.br`.

Shabir A. Shah, Ajay Kumar, Fatima Noor, Pankaj Sharma, M. Asif Mehmood, Hermann Mitter, Bilal A. Bhat, and Mohd Mushtaq. Artificial intelligence and machine learning for ensuring food, nutritional security and smart bioresource management under changing climate: Mitigation and adaptation strategies. *Biotechnology Reports*, 31:e00642, 2021. doi: 10.1016/j.biotech.2021.07.005.

Vishal Sharma, Mei-Ling Tsai, Chiu-Wen Chen, Pei-Pei Sun, Parushi Nargotra, and Cheng-Di Dong. Advances in machine learning technology for sustainable biofuel production systems in lignocellulosic biorefineries. *Science of The Total Environment*, 886: 163972, 2023. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2023.163972. URL `https://www.sciencedirect.com/science/article/pii/S0048969723025937`.

Farhad Mortezapour Shiri, Thinagaran Perumal, Norwati Mustapha, and Raihani Mohamed. A comprehensive overview and comparative analysis on deep learning models: Cnn, rnn, lstm, gru. *Journal on Artificial Intelligence*, 2023.

Haroldo José Torres da Silva, Peterson Felipe Arias Santos, Enilson Carlos Nogueira Junior, and Carlos Eduardo de Freitas Vian. Aspectos técnicos e econômicos da produção de etanol de milho no brasil. *Revista de Política Agrícola*, 29(4):142–159, 2020.

João G.C. Siqueira, João E.G.W. Lima, and Hildeberto F.M. Souza Filho. Price volatility of ethanol in brazil: Effects of the international sugar market and gasoline prices. *Energy Economics*, 45:520–532, 2014. doi: 10.1016/j.eneco.2014.08.017.

Celso Correia de Souza, Joao Bosco Arbués Carneiro Junior, Márcia Ferreira Cristaldo, Raul Assef Castelão, and Daniel Massen Frainer. Aplicação de inteligência artificial na previsão de produções de açúcar e etanol anidro, no período de 2020 até 2025, nas macrorregiões centro-sul e norte-nordeste do brasil. *Revista Brasileira de Geografia Física*, 15(5):2200–2211, 2022.

J. G. M. Souza and F. M. Pompermayer. Variações no preço do etanol em comparação ao preço da gasolina: Uma análise da resposta do consumidor. Texto para Discussão 39, Instituto de Pesquisa Econômica Aplicada (Ipea), 2015.

Abdullah Sumayli. Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil: Gaussian process regression (gpr), multilayer perceptron (mlp), and k-nearest neighbor (knn) regression models. *Arabian Journal of Chemistry*, 2023. doi: 10.1016/j.arabjc.2023.104833. URL `https://doi.org/10.1016/j.arabjc.2023.104833`.

Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1017–1024, 2011.

Mohammad Tami and Amani Yousef Owda. Efficient commodity price forecasting using long short-term memory model. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 14:994–1004, 2024. doi: 10.11591/ijai.v13.i1.pp994-1004.

Narayanaswamy Tattar, Suresh Hanumanthappa, and Sharmila Kumar. *Practical Data Science with R*. Packt Publishing, 2018. ISBN 9781788291463.

RLM et al. Tavares. Prediction of soil co_2 flux in sugarcane management systems using random forest algorithm. *Scientia Agricola*, 75(3):242–250, 2018.

Asterios Theofilou, Stefanos A. Nastis, Anastasios Michailidis, Thomas Bournaris, and Konstadinos Mattas. Predicting prices of staple crops using machine learning: A systematic review of studies on wheat, corn, and rice. *Sustainability*, 17:5456, 2025. doi: 10.3390/su17125456.

João Ricardo Tonin and Julyerme Matheus Tonin. Do proálcool ao "próetanol": novos desafios na produção do etanol brasileiro. *Informe Gepec*, 18(1):61–76, 2014.

John W. Tukey. Exploratory data analysis. *Addison-Wesley*, 1977.

USDA. U.s. department of agriculture, foreign agricultural service – brazil: Sugar annual, 2024. Avaiable online: `https://fas.usda.gov/data/brazil-sugar-annual-10` (accessed on 02 May 2024).

Usina Açucareira de São Manoel. Usina açucareira de são manoel, 2025. URL `https://www.saomanoel.com.br/`. Acessado em: 22 de abril de 2025.

Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J Casson. Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11):e0224365, 2019.

Guido Van Rossum and Fred L. Drake. *Python Programming for the Absolute Beginner*. CreateSpace Independent Publishing Platform, 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Maria de Fátima Vidal. Produção e mercado de etanol. 2020.

Chen Wang, Xuemeng Zhang, Guohua Zhao, and Yinguang Chen. Mechanisms, methods and applications of machine learning in bio-alcohol production and utilization: A review. *Chemosphere*, 342:140191, 2023. ISSN 0045-6535. doi: https://doi.org/10.1016/j.chemosphere.2023.140191. URL https://www.sciencedirect.com/science/article/pii/S004565352302461X.

Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30(1):79–82, 2005.

Jeffrey M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, 6th edition, 2015.

Chao Wu, Yan Wang, and Ling Tao. Machine learning-enabled techno-economic uncertainty analysis of sustainable aviation fuel production pathways. *Chemical Engineering Journal Advances*, 2024. doi: 10.1016/j.ceja.2024.100650. URL https://doi.org/10.1016/j.ceja.2024.100650.

Y. Xing, X. Zhang, and B. Xu. A review on machine learning application in biodiesel. *International Journal of Chemical Engineering*, 2021:2154258, 2021a. doi: 10.1155/2021/2154258.

Yuanzhi Xing, Zile Zheng, Yike Sun, and Masoome Agha Alikhani. A review on machine learning application in biodiesel production studies. *International Journal of Chemical Engineering*, 2021(1):2154258, 2021b. doi: https://doi.org/10.1155/2021/2154258. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/2154258.

P. Frasconi Y. Bengio, P. Simard. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.

Yuan Zhuang, Xiao Sun, You Li, Jianzhu Huai, Luchi Hua, Xiansheng Yang, Xiaoxiang Cao, Peng Zhang, Yue Cao, Longning Qi, Jun Yang, Nashwa El-Bendary, Naser El-Sheimy, John Thompson, and Ruizhi Chen. Multi-sensor integrated navigation/positioning systems using data fusion: From analytics-based to learning-based approaches. *Information Fusion*, 95:62–90, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.01.025. URL https://www.sciencedirect.com/science/article/pii/S1566253523000398.

Diego Zilio. Inteligência artificial e pensamento: redefinindo os parâmetros da questão primordial de turing. *Ciências & Cognição*, 14(1):208–218, 2009.