UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Previsão do preço do arroz no Brasil usando modelos de aprendizado de máquina e dados de oferta e demanda

Lucas Valle Mielke

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

ICMC USP SÃO CARLOS

Lucas Valle Mielke

Previsão do preço do arroz no Brasil usando modelos de aprendizado de máquina e dados de oferta e demanda

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria. *VERSÃO REVISADA*

Área de Concentração: Matemática, Estatística e Computação

Orientador: Prof. Dr. Paulino Ribeiro Villas-Boas

USP – São Carlos
Abril de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Valle Mielke, Lucas
V631p Previsão do preço do arroz no Brasil usando modelos de aprendizado de máquina e dados de oferta e demanda / Lucas Valle Mielke; orientador Paulino Ribeiro Villas-Boas. -- São Carlos, 2024.
73 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Previsão da Preço. 2. Aprendizagem de Máquina. 3. Commodity. 4. Arroz. I. Ribeiro Villas-Boas, Paulino, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Lucas Valle Mielke

Rice Price Prediction in Brazil using Machine Learning Models and Supply and Demand Data

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master – Professional Masters in Mathematics, Statistics and Computing Applied to Industry. *FINAL VERSION*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Prof. Dr. Paulino Ribeiro Villas-Boas

USP – São Carlos
April 2024

Este trabalho é dedicado a Alan Turing e a todos os que contribuem para a ciência, pesquisa e educação.

AGRADECIMENTOS

Sou imensamente grato ao Professor Doutor Paulino Ribeiro Villas-Boas por sua orientação, à USP e ao ICMC pela oportunidade no programa de mestrado. Agradeço ao meu parceiro Lucas, meus pais, amigos e colegas pelo apoio. A todos que contribuíram, meu mais sincero obrigado. Espero que este trabalho retribua de alguma forma o que recebi nesta jornada acadêmica.

"Todos os modelos estão errados, mas alguns são úteis."
(George E. P. Box)

"Me perguntas por que compro arroz e flores? Compro arroz para viver e flores para ter algo pelo que viver."
(Confúcio)

RESUMO

MIELKE, L. V. **Previsão do preço do arroz no Brasil usando modelos de aprendizado de máquina e dados de oferta e demanda**. 2024. 73 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

O arroz é um cereal essencial consumido por cerca de 2,5 bilhões de pessoas no mundo, e o Brasil se destaca entre os dez maiores produtores. A produção brasileira é reconhecida por sua produtividade, tecnologia e fiscalização, se concentrando no Rio Grande do Sul que contribui com cerca de 70% da produção total. Assim como qualquer *commodity* agrícola, o preço do arroz está sujeito às leis de mercado, sendo afetado por diversos fatores, como condições climáticas e preços dos insumos, além da demanda, refletida pelo poder de compra da população. Essa oscilação dos preços pode ser prejudicial tanto para os consumidores quanto produtores, especialmente considerando o tempo de 5 meses entre o plantio e a colheita. Diante dessas questões, o objetivo principal deste trabalho é desenvolver modelos de aprendizagem de máquina capazes de prever o preço dessa *commodity*, considerando um horizonte de 5 meses e utilizando variáveis representativas da oferta e da demanda. Embora existam pesquisas que buscam prever o preço do arroz e de outras *commodities* agrícolas utilizando diferentes modelos de aprendizagem de máquina, não foram encontrados estudos abordando especificamente a previsão com a mesma antecedência deste trabalho, nem utilizando variáveis representativas da oferta e da demanda. Portanto, este projeto preenche essa lacuna. Para a realização desta pesquisa, foram adotados diversos modelos de aprendizagem de máquina que foram aplicados com e sem a técnica de Eliminação Recursiva de Variáveis (RFE), utilizando subconjuntos de dados de treinamento e teste com diferentes períodos. Além disso, dois procedimentos de ajuste na base de dados foram realizados para prever com 5 meses de antecedência: um por meio de defasagem direta e outro utilizando variáveis independentes simuladas, como explicado no capítulo de Materiais e Métodos. Os resultados revelaram que foi possível desenvolver tais modelos, os quais apresentaram uma média de erro de aproximadamente 17%, notando-se erro mais elevado em períodos específicos, especialmente na segunda metade de 2020. O modelo de melhor desempenho na previsão com 5 meses de antecedência foi o Extreme Gradient Boosting com a técnica RFE no procedimento de defasagem direta, alcançando um MAPE de 10%.

Palavras-chave: Previsão da Preço. Aprendizagem de Máquina. *Commodity*. Arroz.

ABSTRACT

MIELKE, L. V. **Rice Price Prediction in Brazil using Machine Learning Models and Supply and Demand Data**. 2024. 73 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Rice is an essential cereal consumed by around 2.5 billion people worldwide, and Brazil stands out among the top ten producers. Brazilian production is recognized for its productivity, technology, and monitoring, mainly concentrated in Rio Grande do Sul, contributing to about 70% of the total production. Like any agricultural commodity, the price of rice is subject to market forces, influenced by various factors such as weather conditions, input prices, and demand, reflected by the population's purchasing power. Price fluctuations can be detrimental to both consumers and producers, especially considering the 5-month period between planting and harvesting. Given these concerns, the main objective of this work is to develop machine learning models capable of predicting the price of this commodity, considering a 5-month horizon and using variables representing supply and demand. While there is existing research aiming to predict the price of rice and other agricultural commodities using different machine learning models, no studies were found specifically addressing forecasting with the same lead time as this work, nor using variables representing supply and demand. Therefore, this project fills this gap. For this research, various machine learning models were adopted, applied both with and without the Recursive Feature Elimination (RFE) technique, using subsets of training and test data with different periods. Additionally, two data adjustment procedures were performed to forecast 5 months in advance: one through direct lagging and another using simulated independent variables, as explained in the Materials and Methods chapter. The results revealed that it was possible to develop such models, which had an average error of approximately 17%, with higher errors noted in specific periods, especially in the second half of 2020. The best-performing model in the 5-month-ahead prediction was the Extreme Gradient Boosting with RFE technique in the direct lagging procedure, achieving a MAPE of 10%.

Keywords: Price Forecast. Machine Learning. Commodity. Rice.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	Rede neural artificial
ANP	Agência Nacional do Petróleo
API	Interface de Programação de Aplicativos
ARIMA	Autorregressivo integrado de médias móveis
ARIMAX	Autorregressivos integrados de médias móveis com variável exógena
ARMA	modelos autorregressivos de média móvel
Ascar	Associação sulina de crédito e assistência rural
CBOT	Bolsa de valores de Chicago
Cepea	Centro de estudos avançados em economia aplicada
Conab	Companhia Nacional de Abastecimento
ELM	Aprendizagem de máquina extremo
Emater	Empresa de assistência técnica e extensão rural
FGV	Fundação Getulio Vargas
IA	Inteligência artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IGP-M	Índice Geral de Preços do Mercado
INMET	Instituto Nacional de Meteorologia
Irga	Instituto rio grandense do arroz
KNN	K-ésimo vizinho mais próximo
MAE	Erro Absoluto Médio
MAPE	Erro Percentual Absoluto Médio
MPL	Rede neural perceptron multicamada
MSE	Erro Quadrático Médio
PAM	Pesquisa Agrícola Municipal
PGPM	Política de Garantia de Preços Mínimos
Pnad	Pesquisa nacional por amostra de domicílios
RFE	Eliminação Recursiva de Variáveis
RMSE	Raiz Quadrática do Erro Quadrático Médio
SARIMA	Autorregressivos integrados de médias móveis com sazonalidade
SARIMAX	Modelo autorregressivo integrado de médias móveis com sazonalidade e fatores exógenos

SVM	Máquina de vetores de suporte
UCM	Modelo de decomposição em componentes não observáveis
USP	Universidade de São Paulo

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Contextualização

O arroz é um dos alimentos mais importantes do mundo, consumido por bilhões de pessoas e desempenhando um papel crucial na luta contra a fome devido à sua fácil adaptação a diferentes condições de solo e clima, além de suas qualidades nutricionais. No Brasil, o arroz também desempenha um papel fundamental ao compor, com o feijão, a base da alimentação dos brasileiros. Além disso, o país possui uma produção significativa de arroz, destacando-se como o único país não asiático entre os 10 maiores produtores do mundo, com o Rio Grande do Sul como seu principal estado produtor.

O arroz é também considerado uma *commodity* agrícola, sujeita a flutuações de preços influenciadas por diversos fatores, incluindo oferta e demanda, condições climáticas e políticas governamentais específicas. Essas oscilações de preço podem ter impactos significativos na qualidade de vida das pessoas, especialmente naquelas de baixa renda, que reduzem muitas vezes o consumo de alimentos, ficando sujeitas à subnutrição e à fome. Além disso, a renda dos agricultores também é fortemente afetada por essas flutuações, especialmente devido ao intervalo de meses entre o plantio e a venda, o que os deixa ainda mais expostos à volatilidade de preços e às suas incertezas.

Por conta da importância do tema exposto, cada vez mais encontramos estudos sobre o uso de modelos computacionais aplicados à predição de preço de *commodities* agrícolas. Esses modelos utilizam geralmente dados climáticos, históricos de preços e de custos importantes, entre outros, para estimar preços futuros do cultivo estudado. O objetivo dessa técnica geralmente é permitir que produtores, comercializadores e outros agentes importantes tomem decisões mais informadas sobre produção, venda e compra de produtos agrícolas (LUDOVICO, 2020). Alguns modelos computacionais mais avançados podem incorporar informações adicionais, como o comportamento do mercado financeiro e até mesmo eventos geopolíticos noticiados que podem

afetar os preços das *commodities* (BARBOSA, 2022). Apesar das previsões baseadas em modelos computacionais estarem sujeitas a erros e incertezas, elas podem ser bem úteis como ferramenta auxiliar para tomada de decisão combinadas com outras informações e análises.

Diante do exposto sobre a importância do arroz e da influência de fatores externos em seu preço, este trabalho visa responder às seguintes questões de pesquisa:

1. É viável coletar e analisar dados que representem os fatores que impactam a oferta e demanda de arroz, com o propósito de desenvolver modelos de aprendizagem de máquina para prever o preço do arroz?

2. É possível desenvolver modelos de aprendizagem de máquina capazes de prever o preço do arroz com uma antecedência de 5 meses, ou seja, considerando o intervalo entre o planejamento do plantio e a venda da produção do arroz?

1.2 Objetivos

O objetivo geral deste trabalho é obter pelo menos um modelo preditivo de aprendizagem de máquina capaz de prever razoavelmente o preço do arroz com 5 meses de antecedência. Para elaborar os modelos preditivos, serão utilizados os registros históricos de preço da *commodity* Arroz do estado do Rio Grande do Sul, sendo a maior região produtora, registrado pelo Centro de estudos avançados em economia aplicada (Cepea), assim como dados públicos representantes de fatores que impactam a oferta do grão, como temperatura, precipitação, área de plantio; bem como fatores representantes da demanda desta cultura, tais como renda média da população.

Como objetivos específicos, têm-se:

1. Consolidar uma base de dados contendo o preço histórico do arroz no Rio Grande do Sul, que será a variável dependente do estudo, e outras variáveis públicas representantes de fatores relacionados à oferta e demanda desta *commodity*, que serão as variáveis independentes analisadas.

2. Elaborar análise descritiva dos dados para compreensão precisa e objetiva dos mesmos, a fim de facilitar a análise dos resultados, bem como contribuir com trabalhos futuros.

3. Desenvolver modelos de predição para prever o preço do arroz com 5 meses de antecedência utilizando os seguintes modelos de aprendizagem de máquina: Regressão Linear, Regressão de Ridge, Árvore de Regressão, Extreme Gradiend Boosting, Modelo autorregressivo integrado de médias móveis com sazonalidade e fatores exógenos (SARIMAX) e Modelo de decomposição em componentes não observáveis (UCM).

4. Treinar e testar os modelos de predição em diversos períodos, registrando os resultados e erros de predição. Verificar quais são os modelos capazes de prever razoavelmente a variável dependente, destacando as melhores alternativas.

1.3 Organização

Este trabalho está organizado em cinco capítulos: 1-Introdução, 2-Fundamentação e Referencial Teórico, 3-Materiais e Métodos, 4-Resultados e Discussão e 5-Conclusão. No segundo capítulo, são apresentados a fundamentação teórica junto com estudos relacionados ao tema de previsão de *commodities*. O terceiro capítulo apresenta o fluxo das atividades para elaboração deste trabalho, detalhes da elaboração da base de dados e das ferramentas utilizadas. No quarto capítulo são apresentadas as análises exploratórias dos dados e análise comparativa da capacidade preditiva de cada um dos modelos estudados seguido de discussão dos resultados. O quinto e último capítulo conclui os resultados deste estudo e apresenta propostas de trabalhos futuros.

CAPÍTULO

2

FUNDAMENTAÇÃO E REFERENCIAL TEÓRICO

Este capítulo visa detalhar a importância do tema apresentado na introdução, analisar trabalhos científicos disponíveis na literatura que utilizaram modelos de aprendizagem de máquina para a predição de preços de *commodities* agrícolas, incluindo o arroz, e mostrar a fundamentação dos modelos de aprendizado de máquina utilizados. O capítulo está dividido em cinco seções, onde a primeira (2.1) se dedica a uma breve apresentação do arroz e de sua importância. A segunda seção (2.2) faz uma análise de publicações que possuem como foco a previsão do preço do arroz. A terceira seção (2.3) apresenta a análise de trabalhos que visam prever o preço de outras *commodities* através do uso da aprendizagem de máquina. As duas últimas seções, a quarta (2.4) e a quinta (2.5), apresentam a fundamentação dos modelos de aprendizado de máquina estudados e a avaliação de desempenho destes, respectivamente.

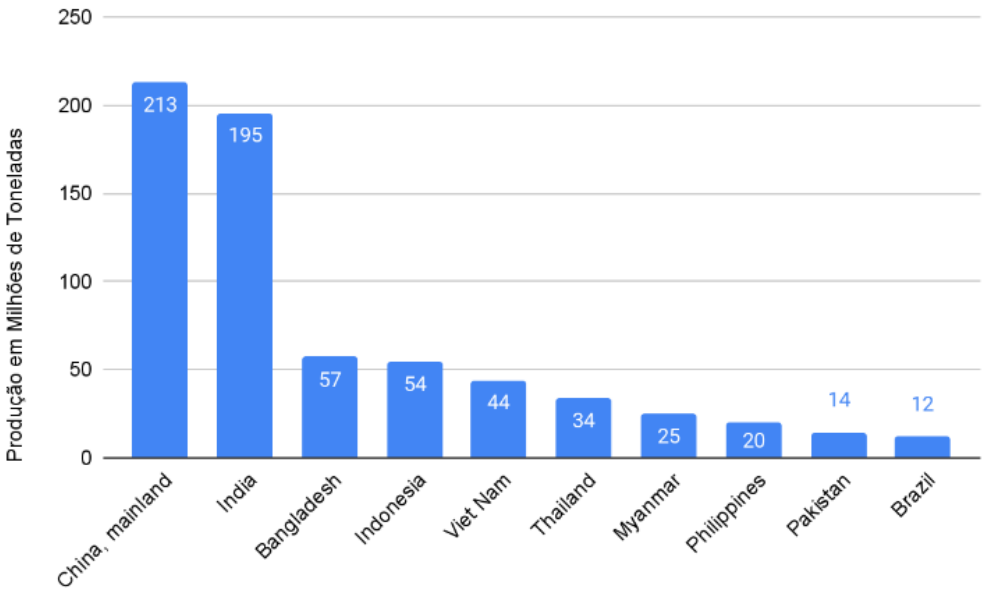
2.1 O Arroz e sua importância

O arroz (*Oryza sativa L.*) é um dos cereais mais relevantes do planeta, cultivado em todos os continentes e consumido por cerca de 2,5 bilhões de pessoas. Além disso, é um alimento que pode ajudar muito no combate à fome no mundo, devido às suas qualidades nutricionais e à facilidade de cultivo em diferentes condições de solo e clima (SILVA; WANDER; FERREIRA, 2021).

Em 2021, a produção mundial de cereais foi aproximadamente de 787 milhões de toneladas, cultivadas em mais de 165 milhões de hectares. Nesse período, o Brasil produziu cerca de 12 milhões de toneladas em quase 1,7 milhão de hectares, o que o coloca como o único país não asiático do ranking de 10 maiores produtores, apresentado na Figura 1 (FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS, 2023). O Brasil também se destaca pela sua produtividade superior à média mundial, pelas tecnologias que utiliza e pela

boa fiscalização governamental (FERREIRA; WANDER; SILVA, 2021).

Figura 1 – Ranking mundial de produção de arroz



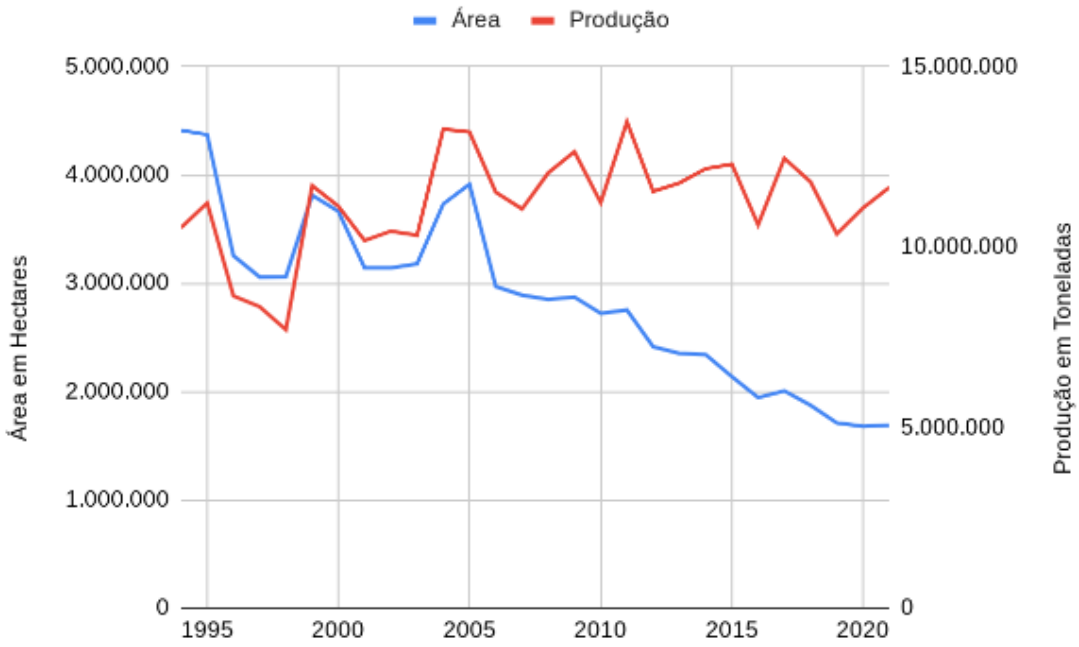
Fonte: Food and Agriculture Organization of the United Nations (2023).

Segundo o Instituto Brasileiro de Geografia e Estatística (2021), o maior estado produtor de arroz em casca é tradicionalmente o Rio Grande do Sul, que responde por 70% da produção nacional. Ferreira, Wander e Silva (2021) também aponta que a produção deste estado é amplamente feita em sistema irrigado com excelente produtividade. Os principais municípios produtores deste estado estão situados nas regiões sul e sudoeste, destacando-se, Uruguaiana, Santa Vitória do Palmar, Itaqui, Alegrete, São Borja, Dom Pedrito, Arroio Grande e São Gabriel, que, juntos, representam 46% da produção gaúcha (ATLAS SOCIOECONOMICO DO RIO GRANDE DO SUL, 2022).

Apesar da produção brasileira de arroz ter se mantido inalterada nas últimas duas décadas, a área cultivada tem se reduzido anualmente, conforme é possível visualizar no gráfico da Figura 2. Isto pode ser preocupante, dada à importância cultural deste alimento, consumido diariamente em todas as regiões, formando, com o feijão, a base da alimentação Brasileira (SILVA; WANDER; FERREIRA, 2021). O arroz é importante não apenas para a população, mas também para os agricultores produtores, especialmente os da Agricultura Familiar, que representam cerca de 10% da produção total (NETO; SILVA; ARAúJO, 2020).

Como o arroz é um produto agrícola básico, cultivado em grande escala e exportado por diversos países em todo o mundo, ele se enquadra na definição de *Commodity* Agrícola (BELLUZZO; FRISCHTAK; LAPLANE, 2014). Os preços das *commodities* são influenciados por diversos fatores, como a oferta e a demanda global, as condições climáticas, as políticas governamentais, entre outros (GUGLIELMETTI, 2016), sendo que quando a oferta é menor

Figura 2 – Evolução da área e produção de arroz no Brasil



Fonte: Food and Agriculture Organization of the United Nations (2023).

do que a demanda, seu preço tende a subir e vice-versa. No caso do arroz, sua oferta, ou seja, a produção total, é diretamente impactada pela área colhida e produtividade, que por sua vez é afetada por fatores climáticos tais como precipitação, insolação e temperatura; e por insumos, tais como fertilizantes, defensivos, combustíveis entre outros. Do outro lado, a demanda da *commodity* arroz é representada pelo consumo total, que possui relação com fatores dos consumidores como desocupação e renda média.

Outra característica marcante das *commodities* é a instabilidade dos preços, que podem sofrer oscilações inesperadas e significativas, causando problemas socioeconômicos importantes. No caso do arroz, que se trata de um alimento básico, um aumento de preço, além de agravar a inflação, pode afetar diretamente a qualidade de vida e a saúde das pessoas, especialmente aquelas de baixa renda, as quais são mais suscetíveis a uma alimentação insuficiente e pouco nutritiva. Por outro lado, uma redução inesperada dos preços pode afetar a renda no campo, principalmente se a redução for acompanhada do aumento dos custos produtivos.

Um exemplo marcante da instabilidade do preço da *commodity* arroz foi o aumento superior a 90% entre janeiro de 2020 e dezembro de 2022 (CENTRO DE ESTUDOS AVANçADOS EM ECONOMIA APLICADA, 2022). Essa oscilação ocorreu devido a fatores como a pandemia da COVID-19, que afetou as cadeias de produção e de distribuição de alimentos, e a variação cambial, que influenciou os custos de importação e exportação (OLIVEIRA; CECHIN, 2021). Aumentos semelhantes também ocorreram com outros alimentos neste período e também se observou claramente as consequências sociais, como o agravamento da fome que afetou milhões de pessoas no Brasil no período.

Na perspectiva do agricultor, há ainda um fator complicante que é o tempo existente entre o plantio e a venda de um produto agrícola, que deixa o produtor rural exposto ao risco de oscilação dos custos produtivos e consequentemente dificulta a previsão de rentabilidade da safra. No caso do cultivo do arroz na maior região produtiva, o estado do Rio Grande do Sul, que apresenta cultivo irrigado (NETO; SILVA; ARAúJO, 2020), o tempo entre plantio e colheita é de até 140 dias (BASF, 2022), ou aproximadamente 5 meses.

Visando dar mais garantias de rentabilidade aos produtores, o Brasil estabeleceu a Política de Garantia de Preços Mínimos (PGPM) que estabelece um preço mínimo para determinados produtos agrícolas, como arroz, milho, feijão e trigo, visando garantir uma renda mínima aos produtores rurais e incentivar a produção desses produtos considerados estratégicos. O preço mínimo é fixado pelo governo a cada safra e pode ser reajustado com base nos custos de produção, preços de mercado e outros fatores. Além disso, a PGPM oferece aos produtores rurais a possibilidade de vender seus produtos ao governo a preços mínimos garantidos, por meio de leilões eletrônicos realizados pela Companhia Nacional de Abastecimento (Conab) (COMPANHIA NACIONAL DE ABASTECIMENTO, 2017). Entretanto, apesar do mecanismo da PGPM trazer benefícios aos agricultores, a frequência de reajuste pode ser insuficiente dada a imprevisibilidade dos preços.

2.2 Previsão de preço do arroz utilizando modelos de aprendizagem de máquina

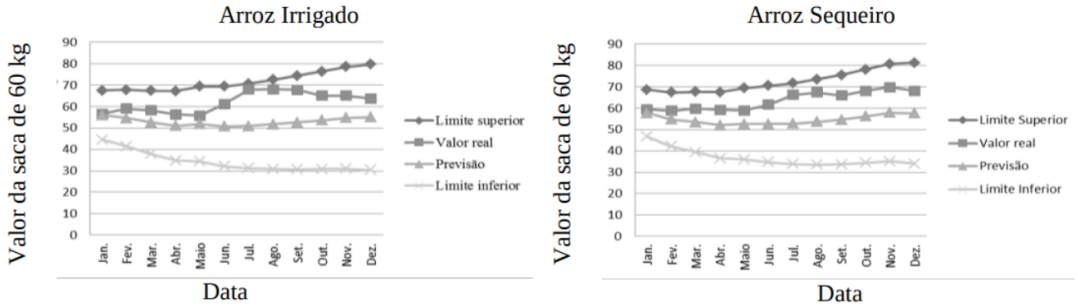
Pinheiro, Tavares e Oliveira (2017) utilizaram técnicas computacionais para prever o preço do arroz no estado do Paraná, com base em uma série temporal de preços históricos. O estudo analisou séries contínuas de preços nominais mensais da *commodity*, referentes à saca de sessenta quilos, no estado entre 1995 e 2015, perfazendo um total de 20 anos, para prever o preço no ano de 2016. O estado do Paraná foi escolhido pelos autores por conveniência, dado haver disponibilidade de todos os dados necessários para a análise, além de ser autossuficiente na produção de arroz.

Desde o princípio, Pinheiro, Tavares e Oliveira (2017) tinham a intenção de utilizar um modelo clássico para prever os preços do cereal. Por conseguinte, ao perceberem que a série temporal possuía uma sazonalidade constante, concluíram que o modelo ideal a ser empregado seria o sazonal simples com suavização exponencial. Esse modelo é uma média ponderada que atribui maior peso às observações mais recentes e utiliza a suavização para ajustar polinômios de baixa ordem, resultando na eliminação de tendências lineares. Para esse método, é calculada uma constante de suavização, sendo que um valor menor conduz a previsões mais estáveis, com maior influência dos últimos valores e, consequentemente, mais aleatórias. Por outro lado, uma constante maior leva a efeitos análogos. Na sequência, esses autores corrigiram o preço pelo Índice Geral de Preços do Mercado (IGP-M) da Fundação Getulio Vargas (FGV) e

aplicaram o modelo escolhido. Posteriormente, os autores identificaram a presença de tendências, variabilidade e *outliers* por meio da análise da autocorrelação, visando evitar possíveis viés nos resultados das previsões. Em seguida, realizaram os cálculos para prever os preços do arroz e compararam os valores previstos com os valores reais.

Pinheiro, Tavares e Oliveira (2017) afirmaram que o modelo criado no estudo apresentou uma capacidade preditiva satisfatória, devido ao coeficiente de correlação próximo a 0,94 nas estatísticas de ajuste do período analisado, além de obterem o MAPE inferior a 4%. Conforme os autores, o teste realizado no ano de 2016 foi bem-sucedido, já que todos os dados reais estavam nos limites do intervalo de confiança, apesar do ano ter sido marcado por instabilidade econômica e geadas. A Figura 3 apresenta a previsão do modelo para o ano de 2016.

Figura 3 – Valores reais dos preços pagos na saca de 60 kg ao produtor de arroz irrigado e sequeiro do estado do Paraná comparados com a previsão e o intervalo de confiança



Fonte: Pinheiro, Tavares e Oliveira (2017).

Conforme destacado por Pinheiro, Tavares e Oliveira (2017), embora o modelo desenvolvido apresente capacidade satisfatória de previsão, é crucial não atribuir-lhe exclusividade na tomada de decisões. É mais apropriado considerá-lo como uma ferramenta auxiliar para análise e planejamento das atividades no agronegócio brasileiro. Os autores ressaltam a limitação do modelo em abordar apenas a série temporal de preços, sublinhando a necessidade de incorporar outras variáveis influentes nos preços dos produtos. Eles recomendam pesquisas adicionais para identificar tais fatores e também explorar diferentes modelos de previsão. O presente trabalho acolhe essas sugestões, investigando os determinantes das oscilações de preço do arroz e avaliando modelos alternativos.

Rathod *et al.* (2022) empregaram técnicas de aprendizado de máquina para prever os preços do arroz na Índia, o segundo maior produtor mundial desse alimento (Figura 1), durante a pandemia da COVID-19 em 2020. Ao contrário do trabalho de Pinheiro, Tavares e Oliveira (2017), esse trabalho incluiu o contexto dos choques de preços devido ao *lockdown* como uma das variáveis a fim de desenvolver um modelo preditivo específico para cenários de crise semelhantes. Em contrapartida, os dados de preços foram coletados durante um período bem mais curto, de janeiro a junho de 2020, divididos em dois períodos: pré-intervenção (1º de janeiro a 24 de março) e pós-intervenção (25 de março a 30 de junho). Foram aplicados modelos gerais de séries

temporais, como o modelo Autorregressivo integrado de médias móveis (ARIMA), o modelo de Rede neural artificial (ANN) e o modelo de Aprendizagem de máquina extremo (ELM). Os modelos utilizaram como conjuntos de treinamento o período entre 1º de janeiro e 23 de junho e como validação as datas de 24 a 30 de junho de 2020. Os resultados indicam que todos os modelos testados apresentaram um MAPE inferior a 1,2% durante o teste, e os modelos de ELM se destacaram como a melhor opção para a modelagem e previsão de dados de preços em crises e descontrole de preços, como o ocorrido durante o *lockdown*. Isso se deve à capacidade dos modelos de ELM de capturar a não-linearidade dos dados de séries temporais.

Embora o período analisado pelo trabalho de Rathod *et al.* (2022) seja relativamente curto, com cerca de 6 meses de duração, o que poderia limitar a capacidade dos modelos em capturar padrões complexos e nuances da série temporal, a metodologia empregada pelos autores é notável por incorporar a ocorrência de um evento externo que afetou a série temporal em questão. Essa variável exógena foi incluída no modelo para explicar a mudança na série temporal, o que é um fator importante a ser considerado para aprimorar a análise e interpretação dos resultados obtidos.

Hasan *et al.* (2020) propuseram o uso de abordagens de aprendizado de máquina para prever os preços do arroz em Bangladesh, o terceiro maior produtor mundial do grão (Figura 1), país onde o arroz é o principal cultivo e a inflação alimentar é uma grande preocupação. Este trabalho se diferencia dos demais por adotar uma abordagem classificatória, iniciando pela classificação do preço em baixo, médio e alto e medindo os resultados por acurácia, uma vez que o problema é classificatório. Foram testados cinco diferentes algoritmos de aprendizado de máquina, a saber, Máquina de vetores de suporte (SVM), K-ésimo vizinho mais próximo (KNN), Naïve Bayes, Árvore de Decisão e Floresta Aleatória. Os dados foram coletados entre 2018 e 2019, totalizando dois anos de amostra e consistiram basicamente na variável dependente, o preço do arroz, e em variáveis independentes relacionadas principalmente ao período do ano (dia, mês, estação) e localização. Embora todos os cinco algoritmos tenham apresentado desempenho semelhante segundo os autores, a Floresta Aleatória se destacou e obteve o melhor resultado com acurácia de 98,17%. Na conclusão, os autores comentaram sobre a dificuldade da coleta de dados e que não puderam coletar dados que representassem todo o país, mas apenas uma região da capital, Dhaka.

Embora os três trabalhos mencionados nesta seção sejam centrados na previsão de preços do arroz, cada um apresenta suas particularidades. No entanto, nenhum deles incorpora variáveis independentes mais complexas, tais como dados climáticos, preços de insumos ou informações populacionais, para a previsão do preço do arroz, como está sendo proposto nesta dissertação. Além disso, nenhum estudo tenta efetuar projeções de valores com maior antecedência. Portanto, este estudo visa preencher essa lacuna e oferecer uma abordagem mais abrangente e precisa para a previsão de preços do arroz.

2.3 Uso de aprendizagem de máquina na predição de preço de outras *commodities*

Porto (2022) realizou uma revisão bibliométrica visando identificar as lacunas na pesquisa sobre modelos de previsão de preços de *commodities* agrícolas e mostrar as principais tendências nessa área. Seus resultados indicaram que as abordagens ARIMA e redes neurais são as mais utilizadas para prever preços de *commodities* agrícolas. Seus resultados também sugerem que os modelos híbridos de Inteligência artificial (IA) geralmente geram previsões com melhores níveis de acurácia em comparação aos métodos estatísticos tradicionais, incluindo ARIMA, modelos individuais e redes neurais, indicando uma tendência crescente na utilização de modelos híbridos. No entanto, apesar dos resultados geralmente superiores em acurácia dos modelos híbridos mencionados por Porto (2022), Shah, Vaidya e Shah (2022) realizaram uma revisão bibliográfica específica para esse tipo de modelo e apontaram limitações, como a elevada complexidade e o elevado custo computacional desses modelos.

Entre as pesquisas analisadas na revisão bibliométrica de Porto (2022), apenas um estudo abordou o preço do arroz, mas de forma conjunta com as *commodities* milho e soja. O estudo em questão é de Marchezan e Souza (2010), que utilizou técnicas de modelos ARIMA e Autorregressivos integrados de médias móveis com sazonalidade (SARIMA) para prever o preço do arroz no Rio Grande do Sul em 2007, usando dados de janeiro de 1995 a dezembro de 2006 da Associação sulina de crédito e assistência rural (Ascar) e da Empresa de assistência técnica e extensão rural (Emater) do Rio Grande do Sul. O método SARIMA (1,1,0) (1,0,1) foi o que apresentou melhor ajuste com erro quadrático médio para prever o preço do arroz.

Porto (2021) realizou um estudo que combinou uma pesquisa bibliométrica sobre modelagem de preços de *commodities* agrícolas com a previsão de preços da soja, utilizando os modelos identificados na pesquisa bibliométrica. Para isso, o autor modelou um conjunto de dados de preços mensais da soja entre 2011 e 2012 nos estados do Paraná, Rio Grande do Sul e Mato Grosso, bem como o preço do contrato futuro da Bolsa de valores de Chicago (CBOT) como variável exógena. Três modelos de previsão foram elaborados: Autorregressivos integrados de médias móveis com variável exógena (ARIMAX), Rede neural perceptron multicamada (MPL) e ELM. Os resultados indicaram que o ARIMAX foi o melhor modelo para prever os preços no Paraná, enquanto a MLP obteve o melhor desempenho de previsão para os preços no Rio Grande do Sul. Para o Mato Grosso, a MLP e a ELM tiveram a melhor acurácia entre as comparações das previsões de preços. Com base nesses resultados, conclui-se que não há um único modelo que seja o melhor para todos os casos, e que a escolha do modelo ideal deve considerar as condições e variáveis específicas do problema. Além disso, o trabalho também evidencia que embora os modelos de IA possam ser úteis em algumas situações, a abordagem estatística ARIMAX ainda é uma opção sólida para a previsão de preços da soja.

Diferentemente de Marchezan e Souza (2010) e Porto (2021), que conduziram pesquisas

de previsão de preços de *commodities* com poucas ou nenhuma variável exógena, o estudo de Carvalho *et al.* (2021) se destacou por incluir várias variáveis exógenas em sua análise de previsão de preços do contrato futuro de milho. O objetivo desse estudo foi avaliar a eficácia de vários modelos de aprendizado de máquina, como SVM, Floresta aleatória, Rede neural, KNN e regressão linear. Para isso, foram coletadas variáveis como a cotação do dólar, produção e produtividade do milho e área plantada para o período entre 2014 e 2018. Este estudo também se diferenciou dos demais desta seção por realizar o treinamento e teste em diversos períodos. Os autores avaliaram o desempenho dos modelos e os compararam utilizando o coeficiente de determinação (R^2). A Regressão Linear se destacou, obtendo R^2 acima de 0,9 em todas as partições de treinamento e validação, indicando que, em alguns casos, modelos de aprendizado de máquina simples podem ser mais úteis para auxiliar estratégias de previsão de preços.

Galyfianakis, Drimbetas e Sariannidis (2016) conduziram um estudo de previsão de preços de cinco séries temporais de *commodities* (petróleo bruto, óleo de aquecimento, gasolina, diesel e querosene de aviação) entre 2005 e 2015. Eles utilizaram um modelo de regressão dinâmica com mudança de regime de Markov, que permite modelar diferentes períodos com comportamentos distintos, visando identificar dois regimes de preços distintos, antes e depois da crise financeira norte-americana de 2008. Os autores destacaram que o estudo foi interessante por identificar várias "recessões" nos preços de mercado ao longo da série temporal estudada, e não apenas no período esperado. Durante a pesquisa, este modelo nos despertou bastante curiosidade devido à sua abordagem diferenciada do problema e por não ser tão comum quanto outros modelos. No entanto, infelizmente, não foi possível encontrar instruções suficientes para a implantação prática do modelo para a predição de valores em problemas que envolvam diversas variáveis independentes.

Esta seção discute trabalhos que se concentram na previsão de outras *commodities* de forma mais ampla do que na seção anterior (2.2), a qual se restringe especificamente à *commodity* arroz. Cada um dos estudos apresentados tem particularidades, especialmente no que diz respeito aos modelos escolhidos para previsão de preço. Embora o trabalho de Carvalho *et al.* (2021) se destaque por utilizar uma base de dados com algumas variáveis independentes, não há nenhum estudo que busque prever o preço de *commodities* agrícolas com um período de previsão mais longo, superior a 1 mês, ou que considere variáveis climáticas ou fatores impactantes na oferta do grão, como o custo de insumos, ou na demanda, como a renda média da população. Nesta dissertação, propõe-se uma abordagem que utiliza um conjunto mais amplo de variáveis e tem como objetivo gerar previsões de preços de arroz com um horizonte temporal maior, diferenciando-se assim dos trabalhos desta seção.

2.4 Modelos de aprendizagem de máquina

2.4.1 Regressão linear múltipla

A regressão linear múltipla é uma técnica estatística utilizada para modelar a relação linear entre uma variável dependente e várias variáveis independentes. Essa relação é obtida por meio de um plano que melhor se ajusta aos dados, minimizando a soma dos erros quadráticos entre as observações reais e as previstas pelo plano (CHEIN, 2019). A regressão linear múltipla é expressa pela equação:

$$Y = a_0 + a_1X_1 + a_2X_2 + ... + a_nX_n + E \tag{2.1}$$

em que:

- Y é a variável dependente a ser prevista;
- $X_1, X_2, ..., X_n$ são as variáveis independentes;
- a_0 é o intercepto da reta, ou seja, o valor de Y quando todas as variáveis independentes são iguais a zero;
- $a_1, a_2, ..., a_n$ são os coeficientes de inclinação da reta, que indicam a mudança em Y para cada unidade de mudança em cada uma das variáveis independentes;
- E é o termo de erro, que representa a variação não explicada pela relação linear entre Y e as variáveis independentes.

O objetivo principal da regressão linear múltipla é ajustar um modelo que encontre os valores de $a_0, a_1, a_2, ..., a_n$ que minimizam uma função de custo, no caso do pacote usado é a soma dos quadrados dos resíduos (ou erros) entre os valores previstos pelo modelo e os valores observados. Isso é feito para encontrar uma equação que melhor se ajuste aos dados e possa ser usada para prever o valor da variável dependente com base nas variáveis independentes (CHEIN, 2019). A função de soma dos quadrados dos resíduos é representada pela equação abaixo:

$$J(\theta) = \sum (y_n - \hat{y}_n)^2 \tag{2.2}$$

em que:

- $J(\theta)$ é a função de custo;
- θ são os parâmetros do modelo (coeficientes de inclinação e intercepto);
- y_n é o valor observado da variável dependente para a n-ésima observação;

- $\hat{y}_n$ é o valor previsto da variável dependente para a n-ésima observação, dado os valores das variáveis independentes e os parâmetros do modelo.

2.4.2 Regressão de ridge

A regressão de ridge, também conhecida como regressão de cume ou regressão de penalização L2, é uma técnica estatística utilizada para modelar a relação linear entre uma variável dependente e várias variáveis independentes. Ao contrário da regressão linear múltipla ordinária, a regressão Ridge adiciona uma penalidade L2 à soma dos erros quadráticos, controlada pelo parâmetro *alpha*. Quanto maior o valor de *alpha*, maior será a penalidade L2 e menor será a magnitude dos coeficientes da regressão, reduzindo a variância do modelo. A regressão Ridge é particularmente útil quando há multicolinearidade entre as variáveis independentes, ou seja, quando duas ou mais variáveis independentes estão altamente correlacionadas entre si.

A equação geral é semelhante à da Regressão Linear Múltipla, com a diferença na função de custo utilizada para encontrar os valores dos coeficientes da regressão, que inclui a penalização de L2 (WIERINGEN, 2015). A função de custo da regressão Ridge é dada por:

$$J(\theta) = \sum((y_n - \hat{y}_n)^2 + \alpha\|\theta\|^2) \tag{2.3}$$

em que:

- $J(\theta)$ é a função de custo;
- θ são os parâmetros do modelo (coeficientes de inclinação e intercepto);
- $\sum(y_n - \hat{y}_n)^2$ é a função de custo da regressão linear;
- α é o parâmetro de regularização que controla a magnitude da penalidade L2;
- $\|\theta\|^2$ é a norma L2 dos coeficientes θ da regressão. No presente trabalho, α é mantida como $\alpha = 1$.

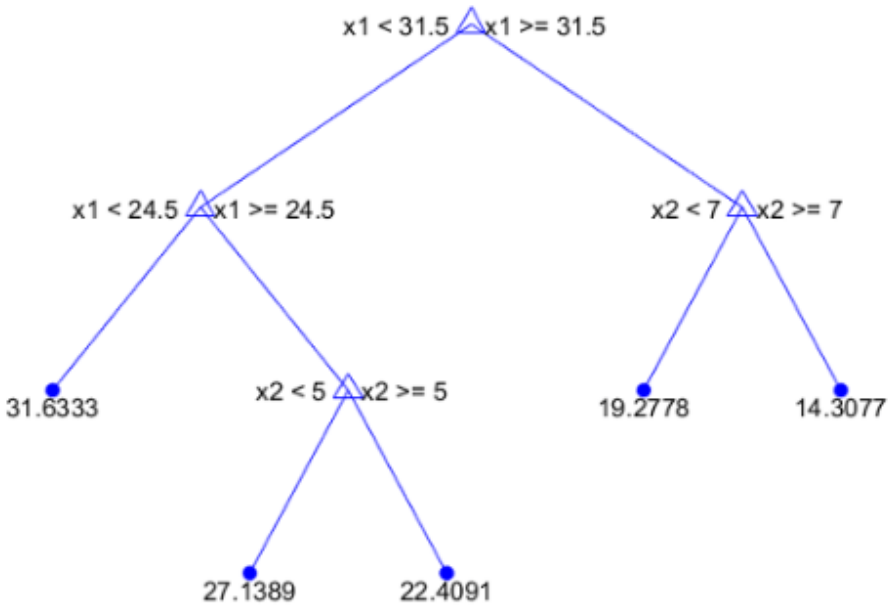
2.4.3 Árvore de regressão

Uma árvore de regressão é um modelo de aprendizado de máquina que, por meio de particionamento recursivo, divide iterativamente o conjunto de dados em subconjuntos menores com base nas características dos dados. Durante o treinamento, os dados são subdivididos em subgrupos menores de acordo com os valores de determinados preditores. Para cada subgrupo resultante, o modelo ajusta uma função constante aos dados contidos em cada nó terminal da árvore, representando uma estimativa numérica da variável de resposta (SOUZA *et al.*, 2021). Geralmente, essa função constante é calculada como uma estatística resumida, como a média dos valores de resposta para todas as observações no nó terminal.

A qualidade das divisões realizadas pela árvore é avaliada usando uma função de custo. Neste modelo, a função de custo utilizada é a soma dos quadrados dos resíduos (descrita detalhadamente na subseção 2.4.1), que mede a discrepância entre os valores previstos pela árvore e os valores reais das observações. Isso implica que a árvore é construída para minimizar essa discrepância, resultando em um modelo que se ajusta aos dados de treinamento e consegue fazer previsões precisas em novos dados.

O modelo é então configurado para selecionar a variável que minimiza a função de erro para cada subdivisão, dividindo os dados recursivamente em subconjuntos menores e criando um ramo da árvore para cada divisão. Esse processo continua até que um critério de parada seja atingido, como um número mínimo de amostras em cada nó ou um número máximo de níveis na árvore (SOUZA *et al.*, 2021). Essas divisões baseadas em regras geram uma representação hierárquica dos dados, como exemplificado na Figura 4.

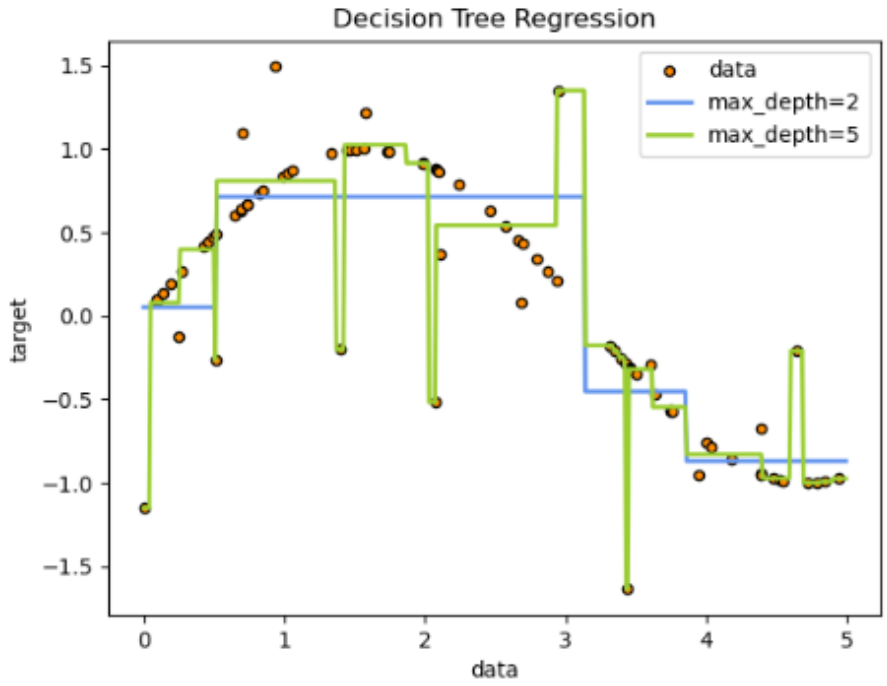
Figura 4 – Exemplo de árvore de decisão



Fonte: MATLAB (2023).

Durante a predição e avaliação, os dados são aplicados às regras de cada nó, gerando um valor de saída. No entanto, é importante destacar que esse modelo apresenta uma limitação: ele prevê resultados somente dentro dos limites dos dados utilizados no treinamento, conforme ilustrado na Figura 5, onde se observa que, embora os modelos (linhas) façam previsões nos limites dos nós de decisão, há pontos com dados reais que ficam distantes dessas previsões, resultando em um aspecto de "degrau" nas previsões. Portanto, é fundamental avaliar cuidadosamente a aplicabilidade do modelo para novos conjuntos de dados e considerar outras técnicas de modelagem quando se deseja prever além dos limites do conjunto de dados de treinamento.

Figura 5 – Previsões da regressão de árvore de decisão (linha) e resultados reais (pontos)



Fonte: scikit-learn.org (2023).

2.4.4 Extreme gradient boosting

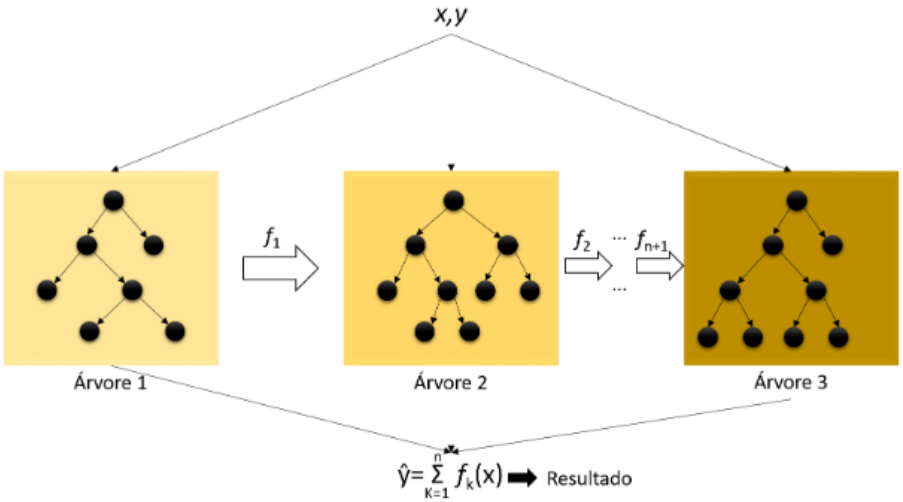
O Extreme gradient boosting, também conhecido como XGBoost ou aumento extremo de gradiente, é um modelo de aprendizado de máquina que constrói sequencialmente um conjunto de modelos de árvore de decisão, em que cada modelo é treinado para capturar os resíduos do modelo anterior, usando uma técnica chamada *boosting*, e a previsão final é a soma das previsões de todos os modelos (CHEN; GUESTRIN, 2016). A Figura 6 apresenta um esquema simplificado deste processo.

A função de custo usada no modelo é a soma dos erros quadrados, como detalhado anteriormente na subseção 2.4.1. No entanto, o XGBoost usa uma otimização de gradiente para ajustar os pesos dos modelos de árvore, a fim de minimizar a função de perda de regressão. A cada iteração, quando um novo modelo de árvore é adicionado ao conjunto, os pesos de cada árvore são atualizados para se concentrarem nas amostras classificadas incorretamente. Esse processo é repetido várias vezes até que o modelo final atinja um nível de precisão aceitável ou até que um limite seja atingido para o número máximo de árvores (CHEN; GUESTRIN, 2016).

2.4.5 Modelo autorregressivo integrado de médias móveis com sazonalidade e fatores exógenos

O modelo autorregressivo integrado de médias móveis com sazonalidade e fatores exógenos (SARIMAX) se trata de uma extensão dos modelos autorregressivos de média móvel

Figura 6 – Conjunto de árvores no XGBoost



Fonte: Junior (2022).

(ARMA). Cada letra representa uma parte do modelo, conforme descrito abaixo (KNAAK; PINTO, 2022):

- Sazonalidade (S): representa o componente sazonal ou o número de pontos de dados em uma temporada.
- Autorregressiva (AR): é o número de termos de atraso da série temporal incluídos no modelo.
- Integrado (I): é o número de vezes que a série temporal foi diferenciada para torná-la estacionária.
- Média Móvel (MA): é o número de erros de previsão de termos de atraso incluídos no modelo.
- Variáveis Exógenas (X): quando há a opção de adicionar variáveis exógenas, usadas para melhorar a precisão da previsão, fornecendo informações adicionais que não estão contidas na série temporal.

Além do significado de cada letra, o modelo é definido com números na sequência, no formato `SARIMAX(p, d, q)(P, D, Q, s)`, em que (KNAAK; PINTO, 2022):

- *p*: representa a ordem da parte autorregressiva (AR) do modelo, ou seja, o número de termos autorregressivos que serão utilizados para modelar a série temporal.
- *d*: representa a ordem da parte integrada (I) do modelo, ou seja, o número de vezes que a série temporal deve ser diferenciada para torná-la estacionária.

- q: representa a ordem da parte de média móvel (MA) do modelo, ou seja, o número de termos de média móvel que serão utilizados para modelar a série temporal.
- P: ordem dos termos autorregressivos sazonais.
- D: ordem das diferenças sazonais.
- Q: ordem dos termos de média móvel sazonais.
- s: periodicidade sazonal.

A equação geral do modelo SARIMAX incorpora tanto a estrutura autorregressiva e de média móvel da série temporal quanto a estrutura sazonal e variáveis exógenas na previsão (KNAAK; PINTO, 2022), sendo expressa por:

$$y(t) = c + \sum_{i=1}^{p} \phi_i y(t-i) + \sum_{j=1}^{q} \theta_j e(t-j) + \sum_{k=1}^{P} \phi_{ks} y(t-k) + \sum_{l=1}^{Q} \theta_{ls} e(t-ls) + X(t)\beta + e(t) \quad (2.4)$$

em que:

- $y(t)$ é o valor da série temporal na época t.
- c é uma constante que representa o intercepto do modelo.
- ϕ_i e θ_j são os coeficientes dos termos autorregressivos e de média móvel, respectivamente.
- ϕ_{ks} e θ_{ls} são os coeficientes dos termos autorregressivos sazonais e de média móvel sazonal, respectivamente.
- $X(t)$ é um vetor de variáveis exógenas na época t, se houver.
- β é o vetor de coeficientes das variáveis exógenas.
- $e(t)$ é o erro de previsão na época t.

2.4.6 Modelo de decomposição em componentes não observáveis

Os modelos de decomposição em componentes não observáveis (UCM) são modelos de espaço de estados que desagregam uma série temporal em suas componentes fundamentais, como tendência, sazonalidade, ciclo e ruído. O objetivo é modelar cada componente separadamente, considerando que as componentes observadas são geradas a partir de componentes não observadas, modeladas como processos estocásticos (SELUKAR, 2016). A equação geral do modelo UCM é dada por (SELUKAR, 2016):

$$Y(x) = T(x) + S(x) + C(x) + I(x) + Z(x) + E(x) \quad (2.5)$$

em que:

- $Y(x)$ é a observação na época x;
- $T(x)$ é a tendência no tempo x;
- $S(x)$ é a sazonalidade no tempo x;
- $C(x)$ é o ciclo no tempo x;
- $I(x)$ é a componente irregular ou estocástica no tempo x;
- $Z(x)$ é a matriz de variáveis exógenas no tempo x;
- $E(x)$ é o erro no período x.

2.5 Avaliação do desempenho dos modelos

Esta seção descreve a avaliação de desempenho dos modelos de aprendizagem de máquina testados neste trabalho, a fim de verificar o quão bem o modelo está se ajustando à variável dependente e permitir uma comparação entre eles. Uma das maneiras mais comuns de avaliar modelos de aprendizagem de máquina é usando métricas de erro, que fornecem uma medida quantitativa das diferenças entre as previsões do modelo e os valores reais. As métricas de erro mais comuns incluem (BOTCHKAREV, 2019):

- O Erro Absoluto Médio (MAE), sendo calculado encontrando a diferença absoluta entre cada previsão e o valor real correspondente e dividindo essa soma pelo número de observações.
- O Erro Quadrático Médio (MSE), sendo calculado somando os erros quadráticos, que nada mais são que a diferença entre cada previsão e o valor real elevada ao quadrado, e dividindo essa soma pelo número de observações.
- A Raiz Quadrática do Erro Quadrático Médio (RMSE), que corresponde exatamente à raiz quadrada do MSE, sendo mais útil para comparar modelos, por estar na mesma escala da variável alvo.
- O Erro Percentual Absoluto Médio (MAPE), que é semelhante ao MAE, mas divide as diferenças absolutas pelo valor real correspondente antes de calcular a média, resultando em um valor percentual e, por isso, uma medida mais útil quando os valores da variável alvo estão em escalas diferentes.

Este trabalho avalia os modelos utilizando todas as métricas mencionadas. No entanto, a comparação final exibida nos Resultados e Discussão utiliza o MAPE, que fornece uma medida mais intuitiva, especialmente quando a variável alvo é volátil e não é familiar para todos os leitores.

CAPÍTULO

3

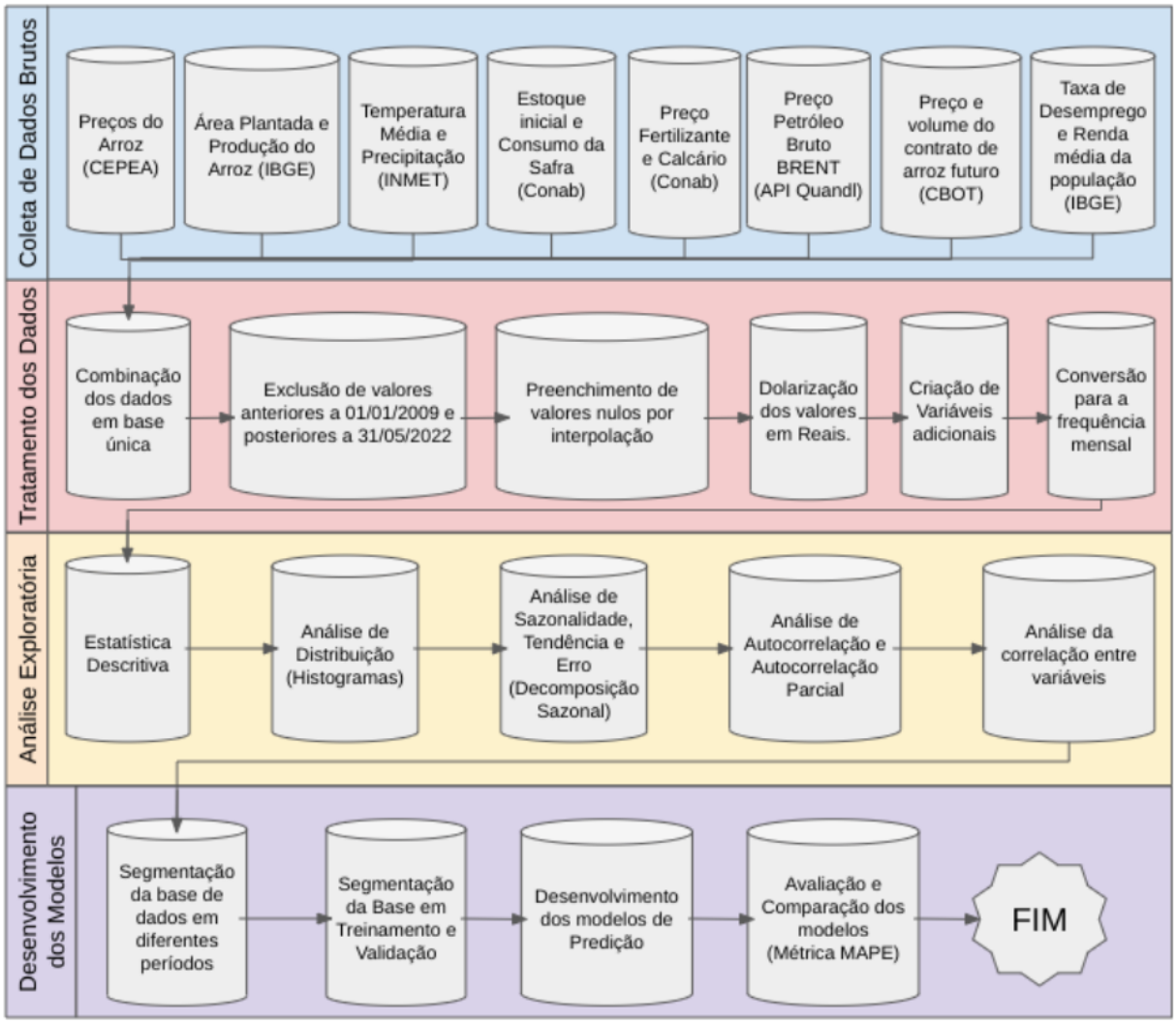
MATERIAIS E MÉTODOS

Este capítulo apresenta os materiais e métodos utilizados na elaboração deste trabalho. A seção 3.1 oferece um resumo do fluxo de atividades da pesquisa, enquanto as demais seções trazem detalhes sobre as etapas realizadas. A seção 3.2 descreve como o conjunto de dados é elaborado, incluindo a origem das informações, os tratamentos realizados e como ele é preparado para o treinamento e avaliação dos modelos. Já a seção 3.3 apresenta as ferramentas utilizadas em toda pesquisa, desde a obtenção e tratamento dos dados até as análises exploratórias e avaliação dos modelos de aprendizagem de máquina implantados. Durante a pesquisa, a coleta e o levantamento dos dados foram considerados os maiores desafios, devido à grande variabilidade de fontes e frequências, bem como ao elevado volume de dados a serem analisados.

3.1 Fluxo de atividades

A Figura 7 apresenta o fluxo das principais atividades adotadas no desenvolvimento da pesquisa, resumindo a metodologia utilizada, que engloba desde o tratamento dos dados até a aplicação dos modelos de predição e mensuração do desempenho destes. As atividades envolvidas no desenvolvimento da pesquisa são: (1) coleta dos dados brutos, oriundos de diversas fontes; (2) tratamento dos dados, incluindo limpeza, transformação, combinação e criação de variáveis adicionais; (3) análise exploratória dos dados, com estatística descritiva e visualizações gráficas para análise de distribuição, tendência, sazonalidade, correlação e autocorrelação, entre outros; (4) desenvolvimento do modelo, com opção de utilizar a Eliminação Recursiva de Variáveis (RFE), separação em amostras de treinamento, validação e testes em diferentes períodos, criação de modelos de previsão, avaliação do desempenho; e (5) comparação da acurácia dos modelos. É importante destacar que este fluxo apresentado é uma simplificação do processo de pesquisa, que é, na realidade, iterativo e interativo. São realizados retornos a etapas anteriores após a obtenção de dados da análise exploratória e dos resultados dos primeiros modelos, a fim de aprimorar o processo e obter melhores resultados.

Figura 7 – Diagrama de atividades descrevendo a metodologia empregada neste trabalho



Fonte: Elaborado pelo autor.

3.2 Construção do conjunto de dados

3.2.1 Descrição dos dados brutos

Ao escolher as variáveis para o modelo, consideramos tanto sua relevância para o problema em análise, bem como sua disponibilidade. Por isso, este trabalho opta por utilizar dados facilmente disponíveis na internet, que possuem um bom registro histórico, para facilitar a replicação em estudos futuros. A coleta dos dados brutos é realizada manualmente em algumas ocasiões, quando as bases estão em planilhas eletrônicas ou arquivos de texto, enquanto em outras é realizada de forma programática por meio de algoritmos em *Python* quando estão disponíveis em Interface de Programação de Aplicativos (API).

A variável dependente deste estudo é o preço do arroz, que é obtido por meio dos dados da pesquisa de preços à vista da saca de 50 kg do arroz com casca no Rio Grande do Sul. Essa pesquisa é conduzida diariamente pelo Cepea da Universidade de São Paulo (USP), que

registra os preços da saca tanto em reais como em dólares americanos (CENTRO DE ESTUDOS AVANçADOS EM ECONOMIA APLICADA, 2022).

Conforme descrito no capítulo 1, as variáveis independentes deste estudo visam representar os fatores que impactam tanto a oferta quanto a demanda de arroz. As primeiras variáveis representantes da oferta são a área colhida, medida em hectares, e a produção, medida em toneladas, no estado do Rio Grande do Sul. Ambas as variáveis são coletadas anualmente pelo Instituto Brasileiro de Geografia e Estatística (IBGE), por meio da Pesquisa Agrícola Municipal (PAM) (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATíSTICA, 2021). Para o ano de 2022, excepcionalmente, os dados são coletados por meio da projeção de safra realizada pelo Instituto rio grandense do arroz (Irga) (INSTITUTO RIO GRANDENSE DO ARROZ, 2022), uma vez que o IBGE ainda não havia divulgado os resultados da pesquisa no momento da coleta de dados.

Para também representar o grupo de variáveis relacionadas à oferta, são coletados a temperatura média em graus Celsius e a precipitação em milímetros, armazenados diariamente no banco de dados do Instituto Nacional de Meteorologia (INMET) e disponíveis via uma API (INSTITUTO NACIONAL DE METEOROLOGIA, 2022). Como há grande quantidade de estações meteorológicas espalhadas pelo estado, sem uma consolidação oficial, opta-se por calcular a média dos dados das estações dos municípios de Rio Grande, Bagé, Quaraí, Uruguaiana, São Borja, Camaquã, Santa Rosa, Campo Bom, Erechim, Rio Pardo, Santiago, Soledade, Santa Maria, Lagoa Vermelha e Tupanciretã. O critério de seleção das estações baseia-se na obtenção de dados de estações meteorológicas com bases mais antigas, consequentemente com mais dados históricos, além de buscar estações suficientes para representar toda a região produtora do estado.

Além das variáveis climáticas, são obtidas informações de estoque inicial e consumo anuais, ambas medidas em mil toneladas, a partir da base de dados do "Portal de Informações" da Conab (COMPANHIA NACIONAL DE ABASTECIMENTO, 2022b). Também são coletados dados do preço médio mensal da tonelada de insumos agrícolas, como calcário e fertilizantes minerais, para o estado do Rio Grande do Sul, disponíveis na base de "Insumos Agrícolas" também da Conab (COMPANHIA NACIONAL DE ABASTECIMENTO, 2022a). Essas informações, em conjunto com as variáveis de estoque e consumo, também compõem o grupo de variáveis representantes da oferta.

Para representar os movimentos do custo do diesel agrícola, também pertencente às variáveis relacionadas com oferta, são coletadas as informações diárias de preço do Petróleo Brent em dólares por barril, por meio de uma API da empresa QUANDL (QUANDL, 2022). A coleta inicialmente visava obter o custo do diesel nos postos de gasolina brasileiros, que era monitorado semanalmente pela Agência Nacional do Petróleo (ANP) no momento de planejamento da pesquisa. No entanto, a partir do segundo semestre de 2022, a pesquisa foi interrompida em certos períodos, gerando perda de informação, além de ter outros problemas como indisponibilidade de bases de dados históricas e piora da divulgação, que deixou de ser

feita em uma base de dados unificada estruturada. Esses problemas com os dados de diesel da ANP foram consequência de questões contratuais (AGêNCIA NACIONAL DO PETRóLEO, GáS NATURAL E BIOCOMBUSTíVEIS, 2022).

Para complementar o conjunto de variáveis relacionadas à oferta de arroz, são coletados o preço futuro e o volume negociado do arroz na CBOT. Essas informações são padronizadas para contratos de arroz com volume de 200.000 libras, equivalente a aproximadamente 90,7 toneladas. A coleta desses dados é realizada a partir de uma base diária armazenada no site de finanças da Yahoo (YAHOO, 2022).

Para completar a etapa de coleta de dados brutos, são coletadas informações sobre a taxa de desemprego e o rendimento médio da população para incluir variáveis que representassem a demanda de arroz. Essas métricas são medidas mensalmente pela Pesquisa nacional por amostra de domicílios (Pnad) realizada pelo IBGE (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATíSTICA, 2023). A Taxa de Desemprego é medida em porcentagem e representa a proporção de pessoas na força de trabalho que estão desempregadas, enquanto o Rendimento Médio da população é representado em Reais.

Abaixo, na Tabela 1, são apresentadas as informações resumidas dos dados brutos descritos neste capítulo e utilizados para a elaboração da base de dados desta pesquisa.

Tabela 1 – Dados brutos coletados

Tipo Variável	Variável	Detalhe	Frequência	Unidade	Fonte
Dependente	Preço do Arroz	Comercializado no RS	Diária	Reais e Dólares por saca de 50 kg	CEPEA USP
Independente, relacionado à Oferta	Área Colhida de Arroz	Total do RS	Anual	Hectares	PAM IBGE
Independente, relacionado à Oferta	Produção de Arroz	Total do RS	Anual	Toneladas	PAM IBGE
Independente, relacionado à Oferta	Temperatura Média	Média de algumas cidades do RS	Diária	Grau Celsius	INMET
Independente, relacionado à Oferta	Volume de Chuva	Média de algumas cidades do RS	Diária	Milímetro	INMET
Independente, relacionado à Oferta	Estoque Inicial no Início do Ano	Nacional	Anual	Mil Toneladas	CONAB
Independente, relacionado à Oferta	Consumo	Nacional	Anual	Mil Toneladas	CONAB
Independente, relacionado à Oferta	Preço médio Calcário	Média do RS	Mensal	Reais por Tonelada	CONAB
Independente, relacionado à Oferta	Preço médio Fertilizante	Média do RS	Mensal	Reais por Tonelada	CONAB
Independente, relacionado à Oferta	Combustível	Petróleo Brent	Diária	Dólare por Barril	QUANDL
Independente, relacionado à Oferta	Valor do Contrato Futuro de Arroz	Negociação na CBOT	Diária	Por contrato (de 200.000 libras)	YAHOO
Independente, relacionado à Oferta	Volume Negociado de Contratos	Negociação na CBOT	Diária	Por contrato (de 200.000 libras)	YAHOO
Independente, relacionado à Demanda	Taxa de Desemprego	Total Brasil	Mensal	Porcentagem	PNAD IBGE
Independente, relacionado à Demanda	Renda Média do Brasileiro	Média Brasileira	Mensal	Reais	PNAD IBGE

Fonte: Elaborado pelo autor.

É importante salientar que, para este trabalho, opta-se por utilizar valores dolarizados, tendo em vista que tanto o arroz quanto a maioria das variáveis independentes monetárias são *commodities* mundiais, negociadas em dólares e apresentando maior estabilidade nesta moeda. Ademais, a utilização de dados em dólares se mostrou vantajosa, uma vez que a inflação desta moeda se manteve relativamente baixa ao longo dos anos, registrando uma elevação somente a partir de 2020, com o impacto da pandemia de Covid-19. Dessa forma, opta-se também por não realizar ajustes de valores por inflação, a fim de minimizar as intervenções nos dados.

3.2.2 Tratamentos realizados

Após a coleta dos dados brutos usados no processo de aprendizado de máquina, é crucial realizar tratamentos para ajustar e melhorar o desempenho do modelo. Esses tratamentos incluem combinar todos os dados em uma única base de dados, transformações de formato e criação de novas variáveis, visando fornecer mais parâmetros para o treinamento e avaliação do modelo, garantindo melhor compreensão do problema e aumento das chances de eficácia do modelo de aprendizado de máquina.

Para reduzir os erros de manipulação manual e aumentar a eficiência nos tratamentos, especialmente após a atualização dos dados brutos, são utilizados algoritmos em *Python* para realizar os tratamentos descritos. A primeira etapa de tratamento consiste na combinação de todos os dados em uma única base, respeitando a frequência da base de dados da variável dependente "preço do arroz", que é diária. Em seguida, é feita a remoção dos dados anteriores a 01/01/2009 e posteriores a 31/05/2022, uma vez que o intervalo entre essas datas continha dados válidos em quase sua totalidade. Após essa etapa, os poucos valores faltantes são preenchidos utilizando o método de interpolação linear, utilizado para estimar valores ausentes com base em valores adjacentes conhecidos, o que permite preencher as lacunas de forma precisa e confiável.

Conforme mencionado na subseção anterior (3.2.1), optamos por trabalhar com dados dolarizados. Cabe destacar que a variável dependente está disponível em uma base de dados que apresenta valores tanto em Reais quanto em Dólares, portanto, não há necessidade de conversão. No entanto, os demais dados monetários em Reais são convertidos para Dólares, utilizando a taxa de câmbio calculada a partir da base de dados da variável dependente. Para isso, realiza-se a divisão do preço do arroz em Reais pelo preço do arroz em Dólares. Decidimos adotar essa taxa de câmbio para evitar a coleta de mais um dado bruto, uma vez que o Cepea, fornecedor da variável dependente, já realiza a conversão para Dólares corretamente, utilizando a taxa de câmbio comercial de venda cotado às 16h30 (CENTRO DE ESTUDOS AVANçADOS EM ECONOMIA APLICADA, 2022).

Após a dolarização das variáveis monetárias que estão em Reais, realiza-se a etapa de criação de variáveis adicionais com base nos dados existentes, conhecida como parte da "engenharia de atributos" (ou "feature engineering" em inglês) na literatura. Vale ressaltar que, como mencionado na seção 3.1, todo o processo é iterativo e as variáveis adicionais são criadas com base em análises exploratórias. A primeira etapa consiste na adição de médias móveis dos períodos de 4, 8 e 12 semanas para todas as variáveis. Toma-se essa decisão ao observar nas análises que a variável dependente "preco" apresenta correlação considerável com algumas variáveis independentes após defasagens. Ou seja, a variável dependente tem boa correlação com algumas variáveis independentes tanto no período atual (n), quanto em períodos anteriores ($n-1$, $n-2$, etc.). Mais detalhes sobre esse assunto são abordados na seção de Análise exploratória de dados (4.1).

Como destacado nos trabalhos relacionados da seção 2.3, o Modelo de Mudança de Regime de Markov desperta nosso interesse, no entanto, sua implantação para um problema com múltiplas variáveis independentes mostra-se inviável neste trabalho. Apesar disso, durante nossas análises, conseguimos usar este modelo para identificar dois regimes distintos na série temporal de preços do arroz, seguindo o exemplo de Galyfianakis, Drimbetas e Sariannidis (2016). Essa análise acaba gerando uma variável binária incorporada à base de dados dos modelos de aprendizagem de máquina propostos neste estudo. Mais detalhes sobre essa abordagem são apresentados na seção de Análise exploratória de dados (4.1).

Para atender aos objetivos do trabalho, a base de dados é convertida em frequência mensal. A base resultante contém 161 amostras e 61 atributos, apresentados na Tabela 2, representando um intervalo de 12 anos e meio de dados. Apesar do grande número de variáveis, cada modelo é treinado e avaliado com a base de dados completa e com uma versão reduzida, após a etapa de RFE, limitando o número de variáveis independentes a 10. Essa redução de variáveis é realizada visando simplificar os modelos e torná-los mais fáceis de reproduzir.

Com o objetivo de avaliar se a inclusão das variáveis independentes exógenas mencionadas na subseção anterior (3.2.1) contribui para aprimorar a capacidade preditiva dos modelos, são criadas bases de dados que contêm apenas os atributos provenientes da variável dependente Y. Isso permite testar os mesmos modelos descritos adiante exclusivamente com esta variável, buscando compreender sua eficácia preditiva sem o suporte das variáveis independentes adicionais.

3.2.3 Particionamento dos dados

A situação do estudo pode ser expressa na forma de uma equação em que temos uma variável Y de interesse, chamada variável dependente, explicada por vetores X, denominados variáveis explicativas ou independentes, da mesma forma que em um modelo linear clássico. A base de dados utilizada contém mais de 12 anos de amostras e é empregada para criar outras 12 bases, cada uma com um intervalo de 8 anos de janela deslizante, cujo período é definido de forma sistemática.

Cada uma das 12 bases, com 8 anos de dados cada, é utilizada para treinar e avaliar os diferentes modelos de aprendizagem de máquina propostos. Para isso, utilizam-se os primeiros 7 anos e 7 meses de dados para treinamento (correspondendo a 95% dos dados iniciais de cada base), enquanto os 5 meses finais são reservados para avaliação. Essa divisão dos dados é escolhida para atender aos objetivos do trabalho, que consistem em criar um modelo capaz de prever o preço do arroz com 5 meses de antecedência. A Tabela 3 apresenta os detalhes dessa divisão dos dados em cada base.

Utilizando esse método, é possível testar os modelos em uma ampla gama de intervalos, aproveitando mais de 12 anos totais de amostras disponíveis na base de dados. Cada modelo é

Tabela 2 – Variáveis finais utilizadas nos modelos de aprendizagem de máquina, sem RFE

Atributo	Descrição	Transformação Realizada
Preco	Preço do Arroz em dólares (variável dependente)	Os dados foram mantidos no formato bruto, porém com a frequência ajustada para mensal. Além disso, foram excluídos os dados anteriores a 01/01/2009 e posteriores a 31/05/2022. Também foi feito o preenchimento de valores nulos, caso existissem.
Temp_Med	Temperatura Média	
Chuva	Volume de Chuva	
Diesel	Preço do Petróleo Brent	
Producao	Produção total de Arroz no RS	
Area	Área colhida de Arroz no RS	
Estoque_Inicial	Estoque nacional de Arroz	
Consumo	Consumo Nacional de Arroz	
Arroz_Futuro	Valor do Contrato Futuro de Arroz	
Volume_Futuro	Volume Negociado de Contratos	
Adubos	Preço médio Fertilizante	
Calcario	Preço médio Calcário	
Desocupacao	Taxa de Desemprego	
Renda	Taxa de Desemprego	
FX	Taxa de Câmbio	A variável foi criada a partir da divisão do preço do arroz em reais pelo preço do arroz em dólares na base de dados do CEPEA.
Markov	Variável binária representando 2 regimes do modelo de Markov	A identificação dos regimes foi realizada por meio do Modelo de Mudança de Regime de Markov disponível na biblioteca Statsmodel do Python.
Preco_Media_Movel(1, 2 ou 3)	A variável representa a média móvel das semanas anteriores da variável correspondente. O número final indica a quantidade de semanas consideradas na média móvel: 1 para 4 semanas, 2 para 8 semanas e 3 para 12 semanas.	Cálculo da média móvel, utilizando as 4, 8 ou 12 semanas anteriores à semana atual, sem incluir a própria semana em questão
Temp_Med_Media_Movel(1, 2 ou 3)		
Chuva_Media_Movel(1, 2 ou 3)		
Diesel_Media_Movel(1, 2 ou 3)		
Fx_Media_Movel(1, 2 ou 3)		
Producao_Media_Movel(1, 2 ou 3)		
Area_Media_Movel(1, 2 ou 3)		
Estoque_Inicial_Media_Movel(1, 2 ou 3)		
Consumo_Media_Movel(1, 2 ou 3)		
Arroz_Futuro_Media_Movel(1, 2 ou 3)		
Volume_Futuro_Media_Movel(1, 2 ou 3)		
Adubos_Media_Movel(1, 2 ou 3)		
Calcário_Media_Movel(1, 2 ou 3)		
Desocupacao_Media_Movel(1, 2 ou 3)		
Renda_Media_Movel(1, 2 ou 3)		

Fonte: Elaborado pelo autor.

treinado e testado individualmente em cada base, visando avaliar sua capacidade de generalização para diferentes conjuntos de dados. No entanto, é importante destacar que, idealmente, teriam sido criados mais períodos, realizando validação cruzada, mas infelizmente a quantidade de bases criadas é limitada pela capacidade computacional disponível.

3.2.4 Ajustes para previsão 5 meses à frente

Para cumprir o objetivo de prever dados com cinco meses de antecedência, exploramos métodos distintos para utilizar as variáveis independentes do período t na previsão da variável

Tabela 3 – Intervalos de dados total, de treinamento e de avaliação, utilizados na aprendizagem de máquina

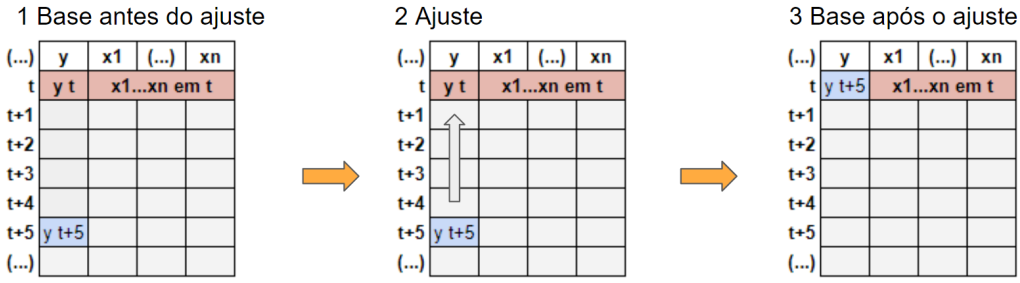
ID	Período Total	Treinamento	Avaliação
1	[2009-01-01 - 2016-12-31]	[2009-01-01 - 2016-07-31]	[2016-08-01 - 2016-12-31]
2	[2009-06-01 - 2017-05-31]	[2009-06-01 - 2016-12-31]	[2017-01-01 - 2017-05-31]
3	[2010-01-01 - 2017-12-31]	[2010-01-01 - 2017-07-31]	[2017-08-01 - 2017-12-31]
4	[2010-06-01 - 2018-05-31]	[2010-06-01 - 2017-12-31]	[2018-01-01 - 2018-05-31]
5	[2011-01-01 - 2018-12-31]	[2011-01-01 - 2018-07-31]	[2018-08-01 - 2018-12-31]
6	[2011-06-01 - 2019-05-31]	[2011-06-01 - 2018-12-31]	[2019-01-01 - 2019-05-31]
7	[2012-01-01 - 2019-12-31]	[2012-01-01 - 2019-07-31]	[2019-08-01 - 2019-12-31]
8	[2012-06-01 - 2020-05-31]	[2012-06-01 - 2019-12-31]	[2020-01-01 - 2020-05-31]
9	[2013-01-01 - 2020-12-31]	[2013-01-01 - 2020-07-31]	[2020-08-01 - 2020-12-31]
10	[2013-06-01 - 2021-05-31]	[2013-06-01 - 2020-12-31]	[2021-01-01 - 2021-05-31]
11	[2014-01-01 - 2021-12-31]	[2014-01-01 - 2021-07-31]	[2021-08-01 - 2021-12-31]
12	[2014-06-01 - 2022-05-31]	[2014-06-01 - 2021-12-31]	[2022-01-01 - 2022-05-31]

Fonte: Elaborado pelo autor.

dependente em $t+5$. Considerando a falta de referências na literatura para esse tipo específico de projeção, desenvolvemos dois procedimentos distintos de ajuste da base de dados, detalhados nos parágrafos a seguir: um procedimento de defasagem direta e outro utilizando variáveis independentes simuladas.

No procedimento de defasagem direta, aplicamos um deslocamento nos dados, associando as variáveis independentes x_{1t}, x_{2t},... x_{nt} de um período t à variável dependente $Y_{(t+5)}$, projetando-a cinco períodos adiante, conforme esquematizado na Figura 8. Essa abordagem permite treinar e avaliar a previsão direta do valor em $t+5$ com base nas variáveis independentes do período t.

Figura 8 – Transformação na base de dados no procedimento de defasagem direta

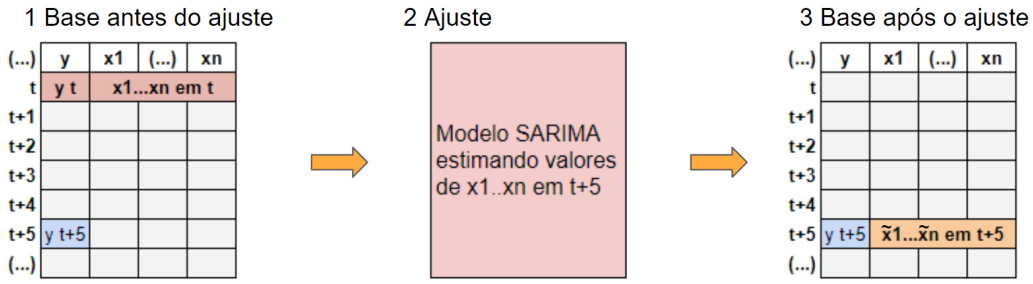


Fonte: Elaborado pelo autor.

Por outro lado, no procedimento com variáveis independentes simuladas, desenvolvemos modelos usando variáveis independentes simuladas $\tilde{x}$ do período $t+1$ até $t+5$ através de um modelo SARIMA. Em outras palavras, $Y(t+5)$ é determinado por $\tilde{x}_1(t+5)$, $\tilde{x}_2(t+5)$,... $\tilde{x}_n(t+5)$, onde as variáveis $\tilde{x}$ são simuladas a partir das variáveis x pelo modelo SARIMA mencionado. A Figura 9, a seguir, esquematiza o procedimento de variáveis independentes simuladas. A

adoção desse método visa testar a hipótese de que ele pode melhorar a precisão da previsão do modelo, permitindo a projeção de valores no mesmo período e capturando de forma mais eficiente padrões de tendência e sazonalidade das variáveis que não foram capturados pelo procedimento de defasagem direta. No entanto, é importante notar que essa segunda abordagem acarreta um aumento significativo no custo computacional.

Figura 9 – Transformação na base de dados no procedimento de variáveis independentes simuladas



Fonte: Elaborado pelo autor.

3.3 Ferramentas

Todas as etapas desta pesquisa são desenvolvidas utilizando programação *Python* na plataforma *Google Colab*.

A etapa de coleta de dados conta com procedimentos manuais, onde uma base de dados é baixada manualmente em formato de planilha eletrônica ou arquivo de texto, e também com descarregamento programático utilizando API. Para a aquisição dos dados, são utilizados pacotes de *Python* como *requests*, *json*, *quandl* e *pygbe*, enquanto, para tratamento dos mesmos, utilizam-se os pacotes *pandas* e *numpy*.

As análises exploratórias, que buscaram fazer estatísticas descritivas e visualizações gráficas, são realizadas com o auxílio dos pacotes *pandas*, *math*, *seaborn*, *matplotlib*, *fitter* e *statsmodel*.

A aplicação dos modelos de aprendizado de máquina, descritos ao longo da seção 2.4, é realizada utilizando pacotes estatísticos do *Python*. Mais precisamente, utilizamos o pacote *sklearn.linear_model.LinearRegression* para o modelo de regressão linear múltipla, o pacote *sklearn.linear_model.Ridge* para a regressão Ridge, o pacote *sklearn.tree.DecisionTreeRegressor* para a árvore de regressão, o pacote *XGBoost* para o extreme gradient boosting, o pacote *statsmodels.tsa.statespace.sarimax* para o SARIMAX, e o pacote *statsmodels.tsa.statespace.structural_-.UnobservedComponents* para os modelos UCM. Vale ressaltar que, especificamente para o modelo SARIMAX, é realizada uma etapa adicional de auto-arima utilizando o pacote *statsmodels* com o objetivo de determinar automaticamente o modelo SARIMAX mais adequado.

É importante mencionar que é aplicada a RFE do pacote *sklearn* do *Python*, visando reduzir o número de atributos dos modelos, a fim de simplificá-los.

As análises de desempenho são realizadas comparando os resultados de todos os modelos descritos na seção 2.4, assim como entre os procedimentos de ajuste de 5 meses detalhados na subseção 3.2.4. Adicionalmente, são feitas comparações entre os resultados das bases de dados completas e aquelas que contêm atributos derivados exclusivamente da variável dependente Y, conforme descrito no final da subseção 3.2.2.

CAPÍTULO

4

RESULTADOS E DISCUSSÃO

No capítulo 3, descrevemos como construímos e parametrizamos a base de dados para adequá-la aos modelos de aprendizagem de máquina, além de apresentar uma explanação sobre as ferramentas utilizadas em todas as etapas da pesquisa. Neste capítulo, abordamos os resultados obtidos, começando com a análise exploratória, apresentada na seção 4.1, seguida da análise de desempenho de cada modelo, destacando aqueles que produzem melhores resultados na seção 4.2, e encerrando com a discussão na seção 4.3.

4.1 Análise exploratória de dados

A Análise Exploratória de Dados é uma abordagem sistemática e descritiva para a análise de dados que visa compreender a natureza dos dados e suas propriedades. Seu principal objetivo é identificar padrões e tendências nos dados, incluindo correlações, agrupamentos de valores semelhantes e sazonalidades. Nessa análise, é comum começar com um sumário estatístico para compreender as medidas de posição e dispersão dos atributos. Vale relembrar que todos os atributos, juntamente com sua identificação e descrição, estão disponíveis na Tabela 2 apresentada na subseção 3.2.2.

Na Tabela 4, apresentamos o sumário estatístico das variáveis brutas coletadas, excluindo os atributos derivados destas, como as médias móveis e o indicador binário de Markov. É importante destacar que todos os 61 atributos, detalhados na Tabela 2, são analisados. A exclusão dos atributos derivados na tabela tem como finalidade permitir a visualização dos dados brutos de forma mais clara e concisa.

No sumário estatístico, destaca-se a diferença entre a média e a mediana da variável *Renda*, que representa a renda média mensal dos brasileiros em dólares por mês. Observa-se que a média deste atributo é 17% maior do que a mediana, sugerindo que, na maioria da série histórica, a renda está abaixo da média. Em contrapartida, para os demais atributos apresentados

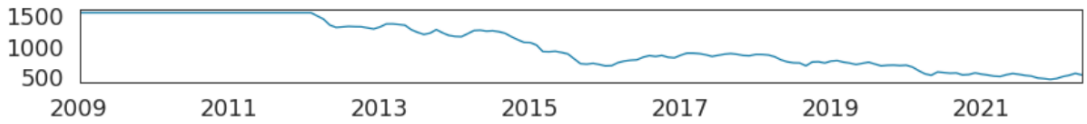
Tabela 4 – Sumário estatístico dos atributos, excluídos aqueles produzidos por engenharia de atributos

	Preco	Temp_Med	Chuva	Diesel	FX	Producao	Area	Estoque_Inicial	Consumo	Arroz_Futuro	Volume_Futuro	Adubos	Calcário	Desocupacao	Renda
Média	13,7	18,9	4,5	75,4	3,1	8.028.400	1.058.384	2.395	11.490	1.297	581	501	42,6	10,0	1.030
Desvio Padrão	2,3	4,1	2,1	26,0	1,2	584.481	67.820	421	600	196	306	139	9,2	2,6	382
Mínimo	9,6	10,0	0,5	18,9	1,6	6.875.077	949.611	1.737	10.545	948	116	313	23,0	6,3	457
Percentil 25%	11,5	15,5	3,0	54,6	2,0	7.692.223	966.937	2.122	10.800	1.150	331	396	36,9	8,0	716
Percentil 50%	14,0	19,4	4,2	71,6	3,1	8.099.357	1.067.581	2.374	11.821	1.280	582	493	41,5	9,0	882
Percentil 75%	15,4	22,6	5,9	102,7	3,9	8.401.787	1.109.976	2.682	12.062	1.459	774	587	50,7	12,3	1.370
Máximo	19,4	26,0	11,8	125,5	5,7	8.940.432	1.168.958	3.312	12.216	1.751	2.254	1.135	60,8	14,9	1.554

Fonte: Elaborado pelo autor.

na Tabela 4, a diferença relativa é significativamente menor, com um desvio inferior a 3% em termos absolutos para 11 dos 15 atributos. Ao se analisar a série histórica do atributo *Renda*, apresentada na Figura 10, é possível observar uma queda constante na renda média a partir da segunda metade do ano de 2011.

Figura 10 – Renda média do brasileiro em dólares por mês

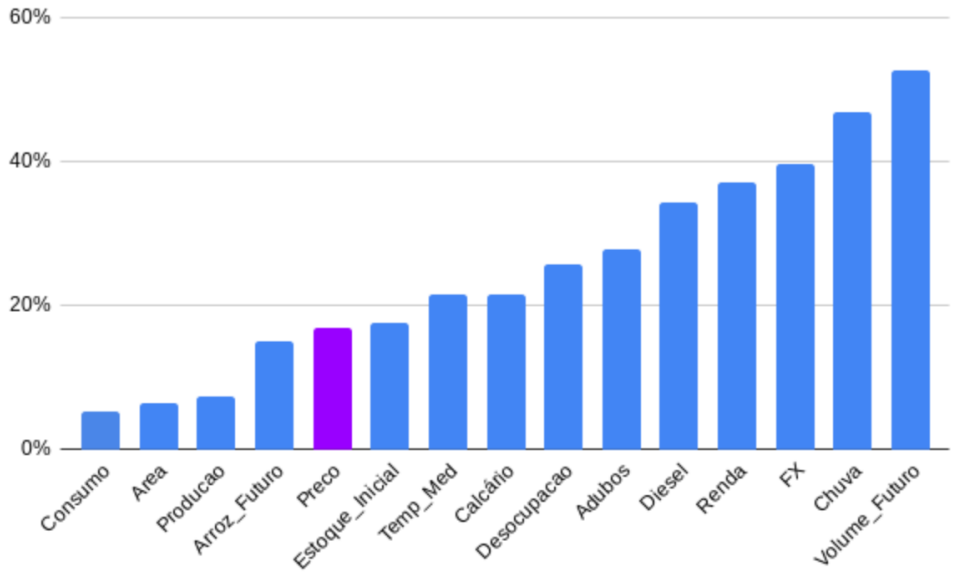


Fonte: Elaborado pelo autor.

O sumário estatístico fornece informações interessantes sobre os atributos analisados, incluindo o Coeficiente de Variação (CV), calculado pela divisão entre o desvio padrão e a média. Ao avaliar o CV para os atributos, notamos que o Preço do Arroz (*variável dependente*), apesar de ser conhecido por sua instabilidade e imprevisibilidade, possui um CV menor do que 10 dos atributos coletados. Isso indica que os valores da variável dependente estão relativamente mais próximos da média em comparação com a maioria dos atributos, sugerindo que a variável dependente é relativamente mais estável do que a maioria das variáveis independentes coletadas. Além disso, a proximidade do CV entre a variável dependente e o atributo *Arroz_futuro* é digna de nota, uma vez que este último representa o preço do contrato futuro do arroz, cuja correlação com o preço atual do arroz é esperada. Essas informações podem ser observadas na Figura 11.

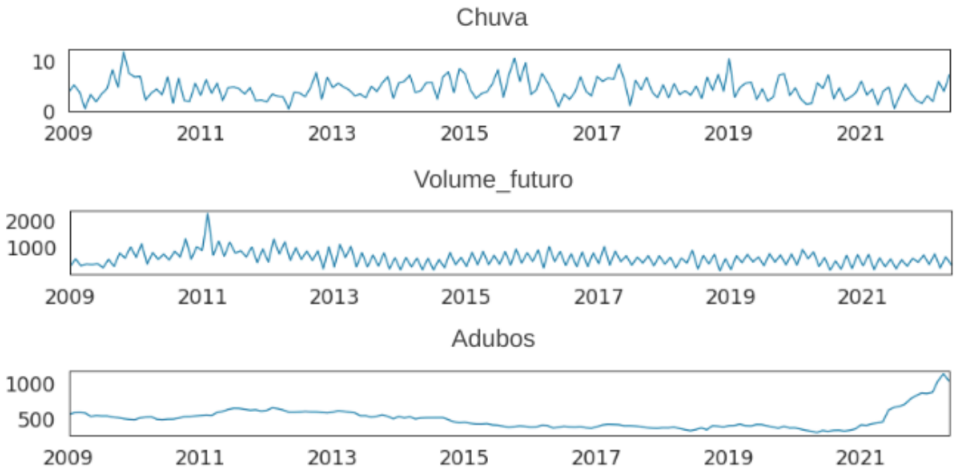
Ao analisar detalhadamente as distribuições dos atributos por meio do gráfico de *box-plot*, é possível observar *outliers* nos atributos *Chuva*, *Volume_Futuro* e *Adubos*. Ao analisar as séries temporais desses atributos, é possível observar que, enquanto nos dois primeiros os *outliers* são pontuais, sugerindo a ocorrência de eventos extremos ou excepcionais, no caso dos Adubos há uma clara tendência de crescimento no preço médio dos fertilizantes, mesmo em dólares, a partir de 2021 conforme evidenciado na Figura 12. Esse aumento pode ser atribuído à Guerra entre Rússia e Ucrânia e seu impacto nos preços dos fertilizantes no Brasil, como apontado por Senar (2022).

Figura 11 – Coeficiente de variação dos atributos, excluídos aqueles produzidos por engenharia de atributos, apresentados no sumário estatístico, em ordem crescente



Fonte: Elaborado pelo autor.

Figura 12 – Séries temporais dos atributos chuva, volume_futuro e adubos



Fonte: Elaborado pelo autor.

As análises das séries temporais, como apresentado na Figura 12, possibilitam uma análise visual das tendências de todos os atributos. Para complementar essa análise, são conduzidos dois testes de estacionariedade em todos os atributos. O primeiro teste, o teste de Dickey-Fuller Aumentado (ADF), avalia a hipótese nula (H0) de que a série temporal possui raízes unitárias, enquanto a hipótese alternativa (H1) sugere a estacionariedade dos dados. O segundo teste aplicado é o teste Kwiatkowski-Phillips-Schmidt-Shin (KPSS), que, diferentemente do ADF, é usado para testar a hipótese nula (H0) de estacionariedade nos dados. Portanto, um valor p menor que um determinado nível de significância no teste ADF indica evidências de que a série

temporal é estacionária, enquanto no teste KPSS isso indica a presença de não estacionariedade na série.

Conforme resumido na Tabela 5, ambos os testes, com um nível de significância de 5%, indicam que os atributos *Temp_med* e *Chuva* são estacionários, o que era esperado devido à natureza dessas variáveis como fatores climáticos. Surpreendentemente, os resultados também indicaram que a variável dependente, o preço do arroz dolarizado, é estacionária, uma descoberta inédita não encontrada na literatura consultada. É relevante destacar que, embora os atributos *Producao*, *Arroz_futuro* e *Adubos* sejam considerados estacionários apenas no teste KPSS, em contradição aos resultados do teste ADF, análises gráficas conduzidas durante a pesquisa a não estacionariedade desses atributos.

Tabela 5 – Resultados dos testes de normalidade e estacionariedade dos atributos, excluídos aqueles produzidos por engenharia de atributos

	ADF		KPSS	
Atributo	**valor p**	**Resultado**	**valor p**	**Resultado**
Preco	2,85E-02	Estacionária	7,49E-02	Estacionária
Temp_med	3,20E-02	Estacionária	1,00E-01	Estacionária
Chuva	1,23E-08	Estacionária	1,00E-01	Estacionária
Diese	4,27E-01	Não Estacionária	2,71E-02	Não Estacionária
FX	9,03E-01	Não Estacionária	1,00E-02	Não Estacionária
Producao	5,50E-02	Não Estacionária	1,00E-01	Estacionária
Area	9,02E-01	Não Estacionária	1,00E-02	Não Estacionária
Estoque_inicial	3,44E-01	Não Estacionária	1,00E-02	Não Estacionária
Consumo	7,72E-01	Não Estacionária	1,00E-02	Não Estacionária
Arroz_futuro	1,38E-01	Não Estacionária	7,52E-02	Estacionária
Volume_futuro	5,43E-01	Não Estacionária	1,00E-02	Não Estacionária
Adubos	5,44E-01	Não Estacionária	1,00E-01	Estacionária
Calcario	4,97E-01	Não Estacionária	1,00E-02	Não Estacionária
Desocupacao	4,24E-01	Não Estacionária	1,00E-02	Não Estacionária
Renda	8,06E-01	Não Estacionária	1,00E-02	Não Estacionária

Fonte: Elaborado pelo autor.

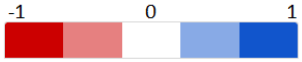
A fim de obter informações adicionais sobre os atributos e detectar possíveis problemas, como viés de dados ou atributos redundantes, a correlação entre os atributos é avaliada resultando na matriz de correlação apresentada na Figura 13. Observa-se que a variável dependente *Preco* não possui uma correlação forte com nenhum outro atributo, apresentando uma correlação moderada com os atributos *Diesel*, *Renda*, *Calcario* e *Arroz_futuro*. Surpreendentemente, a variável *Preco* não possui correlação com os atributos *Area* e *Temp_med*, contradizendo as expectativas iniciais. São identificados cinco casos de correlação positiva forte, acima de 0,7, entre os atributos: *Area* e *Consumo*, *Consumo* e *Renda*, *Desocupacao* e *FX*, *Calcario* e *Diesel*, e *Calcario* e *Renda*. A relação entre *Consumo* e *Renda* é esperada, uma vez que o consumo de alimentos tende a aumentar com a renda dos consumidores. Além disso, a correlação entre *Desocupacao* e *FX* indica que momentos de crise, com aumento do desemprego, costumam coincidir com a desvalorização do Real em relação ao Dólar Americano. A correlação entre *Area* e *Consumo* é curiosa, pois seria esperado que a área colhida impactasse o consumo devido

à sua relação direta com oferta e, consequentemente, inversa com o preço, no entanto, não há correlação entre a área colhida e o preço do arroz o que nos deixou sem hipótese explicativa. As demais correlações positivas fortes que observamos devem ocorrer por casualidade.

Também observamos 6 casos de correlação negativa forte, abaixo de -0,7, sendo que 4 desses casos ocorrem com a variável *FX*, que representa a cotação do Dólar Americano no Brasil, com os atributos *Area*, *Consumo*, *Calcario* e *Renda*. O fenômeno entre *FX* e os dois primeiros atributos podem estar relacionados com uma redução de oferta do Arroz quando o real se desvaloriza, possivelmente por conta da opção dos agricultores por *commodities* de exportação como a Soja, porém são necessários mais estudos para confirmar essa hipótese, enquanto a relação de *FX* com os dois últimos podem ser consequência simplesmente da conversão dos valores de Real para dólar, conforme explicado no capítulo de Materiais e Métodos. Os outros casos de correlação negativa forte ocorre entre *Calcario* e *Desocupacao*, provavelmente por casualidade, e entre *Desocupacao* e *Renda*, o que é bastante esperado, uma vez que a renda média de uma população é menor quando a desocupação desta aumenta. A Figura 13 apresenta a Matriz de Correlação dos Atributos.

Figura 13 – Matriz de correlação entre os atributos, excluídos aqueles produzidos por engenharia de atributos

	Preco	Temp_Med	Chuva	Diesel	Fx	Producao	Area	Estoque_Inicial	Consumo	Arroz_Futuro	Volume_Futuro	Adubos	Calcario	Desocupacao	Renda
Preco	1,00	0,00	-0,01	0,41	-0,28	-0,19	0,00	0,25	0,28	0,37	0,03	0,30	0,37	-0,29	0,38
Temp_Med	0,00	1,00	0,12	-0,04	0,05	0,01	-0,04	-0,04	-0,03	-0,01	0,12	0,01	-0,09	-0,01	-0,04
Chuva	-0,01	0,12	1,00	-0,07	-0,08	0,10	0,19	-0,10	0,15	-0,09	-0,04	-0,12	0,02	-0,08	0,00
Diesel	0,41	-0,04	-0,07	1,00	-0,54	0,14	0,26	0,60	0,31	0,68	0,21	0,64	0,75	-0,61	0,55
Fx	-0,28	0,05	-0,08	-0,54	1,00	0,00	-0,77	-0,51	-0,80	-0,23	-0,29	-0,14	-0,88	0,82	-0,95
Producao	-0,19	0,01	0,10	0,14	0,00	1,00	0,49	0,25	0,30	0,12	0,01	0,09	-0,02	0,00	-0,06
Area	0,00	-0,04	0,19	0,26	-0,77	0,49	1,00	0,39	0,82	0,02	0,21	-0,03	0,60	-0,63	0,67
Estoque_Inicial	0,25	-0,04	-0,10	0,60	-0,51	0,25	0,39	1,00	0,49	0,49	0,27	0,53	0,54	-0,46	0,67
Consumo	0,28	-0,03	0,15	0,31	-0,80	0,30	0,82	0,49	1,00	0,18	0,21	0,16	0,67	-0,68	0,78
Arroz_Futuro	0,37	-0,01	-0,09	0,68	-0,23	0,12	0,02	0,49	0,18	1,00	0,12	0,59	0,45	-0,43	0,36
Volume_Futuro	0,03	0,12	-0,04	0,21	-0,29	0,01	0,21	0,27	0,21	0,12	1,00	0,09	0,29	-0,20	0,29
Adubos	0,30	0,01	-0,12	0,64	-0,14	0,09	-0,03	0,53	0,16	0,59	0,09	1,00	0,46	-0,38	0,26
Calcario	0,37	-0,09	0,02	0,75	-0,88	-0,02	0,60	0,54	0,67	0,45	0,29	0,46	1,00	-0,81	0,84
Desocupacao	-0,29	-0,01	-0,08	-0,61	0,82	0,00	-0,63	-0,46	-0,68	-0,43	-0,20	-0,38	-0,81	1,00	-0,80
Renda	0,38	-0,04	0,00	0,55	-0,95	-0,06	0,67	0,67	0,78	0,36	0,29	0,26	0,84	-0,80	1,00

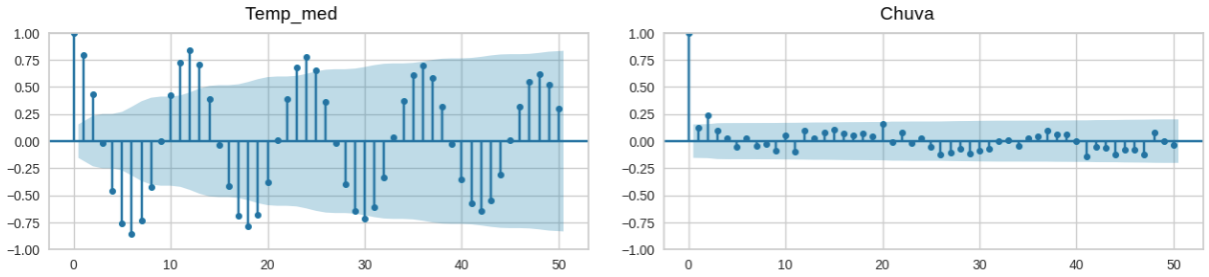


Fonte: Elaborado pelo autor.

Por fim, a autocorrelação e a autocorrelação parcial dos atributos são avaliadas para identificar padrões e dependências nas observações ao longo do tempo, considerando o processo de previsão da variável dependente com 5 meses de antecedência, conforme explicado na

subseção 3.2.4. Os resultados revelam que todos os atributos apresentam autocorrelação ao longo de pelo menos 5 meses, com exceção de *Temp_med*, que exibe autocorrelação sazonal, e *Chuva*, que praticamente não demonstra nenhum caso significativo de autocorrelação, conforme mostrado na Figura 14.

Figura 14 – Autocorrelação dos atributos “Temp_med” e “Chuva”

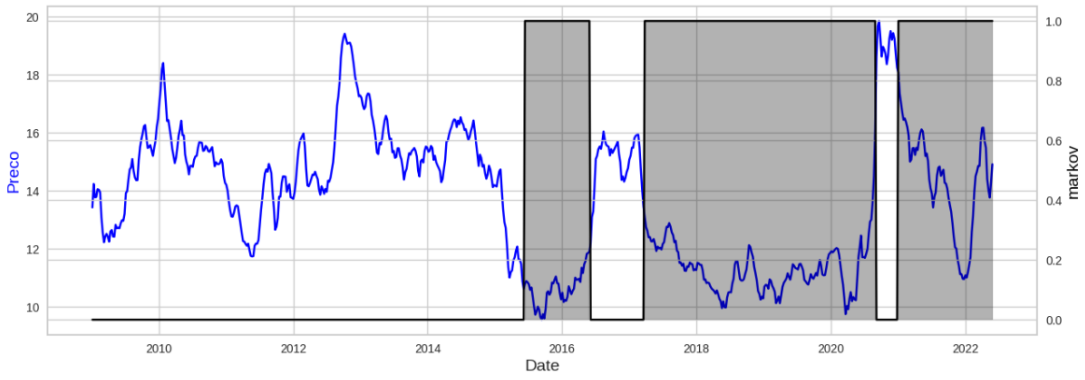


Fonte: Elaborado pelo autor.

Como mencionado nos Materiais e Métodos, desenvolvemos atributos através da engenharia de atributos. Com exceção do atributo *Markov*, que é discutido posteriormente, os demais atributos produzidos consistem em médias móveis dos atributos originais e reproduzem, em certa medida, as características dos atributos originais e não mostram nenhum destaque significativo que valesse a pena mencionar para as análises descritas até o momento.

O atributo *Markov*, mencionado anteriormente, é um atributo binário que identificou dois regimes para a variável dependente *preco*. O regime 0 representa aproximadamente 70% do período total, predominando especialmente no início da série temporal, entre 2009 e meados de 2015. Por sua vez, o regime 1 corresponde aos 30% restantes. A Figura 15 apresenta um gráfico com a série temporal do preço do arroz em linhas azuis, sobrepondo-se à variável *Markov* em preto, permitindo observar como o modelo de Markov identifica os dois regimes de preço.

Figura 15 – Série temporal da variável dependente “preco” e os regimes de Markov



Fonte: Elaborado pelo autor.

Embora não seja possível atribuir com 100% de certeza as causas e explicações para esses regimes, podemos inferir que o regime 0 está relacionado a uma relativa estabilidade com maiores preços da *commodity*, enquanto o regime 1 parece estar associado a momentos de queda no preço desta. É relevante ressaltar que, durante o primeiro período do regime 1, que compreendeu o período de meados de 2015 até meados de 2016, o Brasil enfrentou uma crise política e econômica tanto interna, com aumento do déficit público e da inflação, quanto externa, com a redução das importações da China, especialmente de *commodities* (BARRUCHO, 2015). Já o segundo e o terceiro períodos do regime 1 podem ser considerados praticamente um único período, que se estendeu da primeira metade de 2017 até o fim da série temporal, com uma pequena interrupção no final de 2020. Esse período foi marcado por grande instabilidade no Brasil e no mundo, envolvendo crises econômicas, crises diplomáticas entre EUA e China, bem como a crise sanitária do COVID-19 e suas consequências (TREVIZAN, 2017; CELASUN; MILESI-FERRETTI; OBSTFELD, 2018; PESSOA, 2021).

Na Tabela 6 abaixo, apresentamos os coeficientes selecionados pela RFE para cada modelo de aprendizagem de máquina. É evidente a predominância de variáveis de médias móveis geradas na engenharia de atributos, com quase todos os atributos apresentando essa característica. Além disso, notamos que as variáveis climáticas foram empregadas apenas na árvore de regressão, possivelmente devido à falta de correlação com outros fatores, conforme observado na matriz da Figura 13. Por fim, destacamos a presença da variável de mudança de regime de Markov em todos os modelos exceto "Ridge Regression"e "Tree Regression", indicando a importância desse indicador de regime na modelagem.

Tabela 6 – Coeficientes selecionados pela RFE

Modelo	Coeficientes
Extreme Gradient Boosting	preco_media_movel1', 'estoque_inicial_media_movel1', 'consumo_media_movel3', 'arroz_futuro_media_movel3', 'volume_futuro_media_movel3', 'Adubos_media_movel3', 'Calcário_media_movel3', 'Desocupacao_media_movel3', 'Renda_media_movel3', 'markov'
Linear Regressão	FX', 'preco_media_movel1', 'FX_media_movel1', 'Desocupacao_media_movel1', 'preco_media_movel2', 'FX_media_movel2', 'Desocupacao_media_movel2', 'preco_media_movel3', 'Desocupacao_media_movel3', 'markov'
Ridge Regression	FX', 'preco_media_movel1', 'FX_media_movel1', 'Desocupacao_media_movel1', 'preco_media_movel2', 'FX_media_movel2', 'Desocupacao_media_movel2', 'preco_media_movel3', 'FX_media_movel3', 'Desocupacao_media_movel3'
SARIMAX	FX', 'preco_media_movel1', 'FX_media_movel1', 'Desocupacao_media_movel1', 'preco_media_movel2', 'FX_media_movel2', 'Desocupacao_media_movel2', 'preco_media_movel3', 'Desocupacao_media_movel3', 'markov'
Tree Regression	CHUVA', 'preco_media_movel1', 'CHUVA_media_movel1', 'Calcário_media_movel1', 'Desocupacao_media_movel1', 'preco_media_movel2', 'TEMP_MED_media_movel2', 'preco_media_movel3', 'TEMP_MED_media_movel3', 'producao_media_movel3'
Unobserved Components Model	FX', 'preco_media_movel1', 'FX_media_movel1', 'Desocupacao_media_movel1', 'preco_media_movel2', 'FX_media_movel2', 'Desocupacao_media_movel2', 'preco_media_movel3', 'Desocupacao_media_movel3', 'markov'

Fonte: Elaborado pelo autor.

4.2 Resultados dos modelos de aprendizagem de máquina

No capítulo Materiais e Métodos, descrevemos o processo de preparação da base de dados, que envolve 12 particionamentos dos dados, detalhados na subseção 3.2.3, e a aplicação de 2 procedimentos distintos para prever dados com 5 meses de antecedência, mencionados na subseção 3.2.4. Detalhamos a implementação de 6 modelos de aprendizagem de máquina, com ou sem RFE, e explicamos a criação de bases de dados completas e simplificadas. Essas últimas possuem atributos gerados exclusivamente a partir da variável dependente, permitindo avaliar o impacto das variáveis independentes exógenas nos resultados, conforme explicado no final da subseção 3.2.2. Ao final desses processos, obtemos 432 resultados de teste, utilizando o MAPE como métrica de comparação, e organizamos esses resultados em 4 matrizes.

Essas 4 matrizes são organizadas para diferenciar os dois procedimentos distintos e também diferenciar os resultados das bases de dados completas das que continham apenas atributos derivados da variável dependente. Cada matriz representa os modelos de aprendizagem de máquina nas linhas e os 12 períodos de treino e teste nas colunas. É relevante destacar que os resultados dos modelos com e sem RFE são dispostos nessas matrizes, identificados pela adição do sufixo “RFE” ao nome do modelo de aprendizagem de máquina.

Antes de apresentar os resultados, é importante mencionar que é necessário estabelecer um limite razoável para o MAPE dos modelos, uma vez que não encontramos na literatura nenhum estudo sobre esse limite específico. Portanto, para este estudo, optamos por estipular o limite máximo do MAPE considerado razoável como 17%. Essa escolha é baseada na proximidade desse valor com o coeficiente de variação da variável dependente *preco* e, coincidentemente, também com o valor médio do MAPE de todos os modelos e períodos para dois os procedimentos realizados com a base completa, cuja informação pode ser observada nas matrizes apresentadas nas Figuras 16 e 18.

A Figura 16 mostra uma matriz no estilo de um “mapa de calor”, exibindo o MAPE de todos os modelos de aprendizagem de máquina em cada período do procedimento de defasagem direta, conforme detalhado em Materiais e Métodos, para a base de dados completa. Observa-se que seis modelos alcançam um MAPE abaixo do limite estabelecido de 17%, como representado na Figura. Entre esses modelos, aqueles que utilizam Extreme Gradient Boosting, com e sem a aplicação de RFE, obtêm os melhores resultados gerais, com um MAPE em torno de 10%. No entanto, é importante ressaltar que o modelo com RFE tem um desempenho consideravelmente inferior ao modelo sem a RFE no período de 2011-06-01 a 2019-05-31, sugerindo que a remoção de certos atributos pode ter influenciado negativamente a precisão do modelo durante esse período específico.

Além disso, ao examinar os resultados de cada modelo em diferentes períodos, observamos um aumento significativo no erro, especialmente nos estágios iniciais da pandemia. Essa tendência é notável ao analisar a série temporal da variável Y_{real} em comparação com a previsão

Figura 16 – MAPE dos modelos e períodos do procedimento de defasagem direta feita na base de dados completa, em ordem crescente por modelo

Model	2009-01-01 2016-12-31	2009-06-01 2017-05-31	2010-01-01 2017-12-31	2010-06-01 2018-05-31	2011-01-01 2018-12-31	2011-06-01 2019-05-31	2012-01-01 2019-12-31	2012-06-01 2020-05-31	2013-01-01 2020-12-31	2013-06-01 2021-05-31	2014-01-01 2021-12-31	2014-06-01 2022-05-31	Total
Extreme Gradient Boosting RFE	4,3	10,1	4,9	6,3	4,7	20,8	2,8	6,3	37,5	4,3	7,7	10,2	10,0
Extreme Gradient Boosting	5,1	12,3	6,9	7,8	7,3	2,3	10,7	9,5	31,7	2,7	10,5	16,1	10,2
Tree Regression RFE	6,5	6,5	9,9	6,6	8,7	3,7	16,0	6,8	33,0	20,4	8,6	25,3	12,7
Tree Regression	10,7	11,0	18,0	6,9	6,4	6,9	3,9	6,5	32,3	24,6	11,7	16,3	12,9
Ridge Regression RFE	12,0	14,6	20,2	15,9	5,8	5,4	1,6	6,7	36,6	42,1	9,8	15,7	15,6
Linear Regressão RFE	11,4	10,1	17,5	14,6	6,8	7,4	4,1	9,7	35,1	45,0	9,8	18,5	15,8
Unobserved Components Model RFE	16,0	29,0	23,9	9,3	7,4	6,0	5,4	10,1	26,8	31,6	16,9	19,5	16,8
SARIMAX RFE	16,0	29,0	23,9	9,3	7,4	6,0	5,4	10,1	26,8	31,6	16,9	19,5	16,8
Ridge Regression	2,6	8,9	7,7	22,9	31,8	10,0	36,7	16,4	47,4	24,7	12,4	11,9	19,4
Linear Regressão	15,7	7,8	21,3	32,9	21,9	12,2	28,2	21,0	25,9	29,4	6,6	11,0	19,5
SARIMAX	53,7	13,1	13,3	26,2	15,1	12,0	27,8	24,8	40,5	29,4	22,8	19,6	24,9
Unobserved Components Model	53,7	13,2	13,3	26,2	15,9	11,8	28,0	24,6	40,3	29,7	23,0	19,7	24,9
Total	17,3	13,8	15,1	15,4	11,6	8,7	14,2	12,7	34,5	26,3	13,0	17,0	16,6

Mínimo Máximo

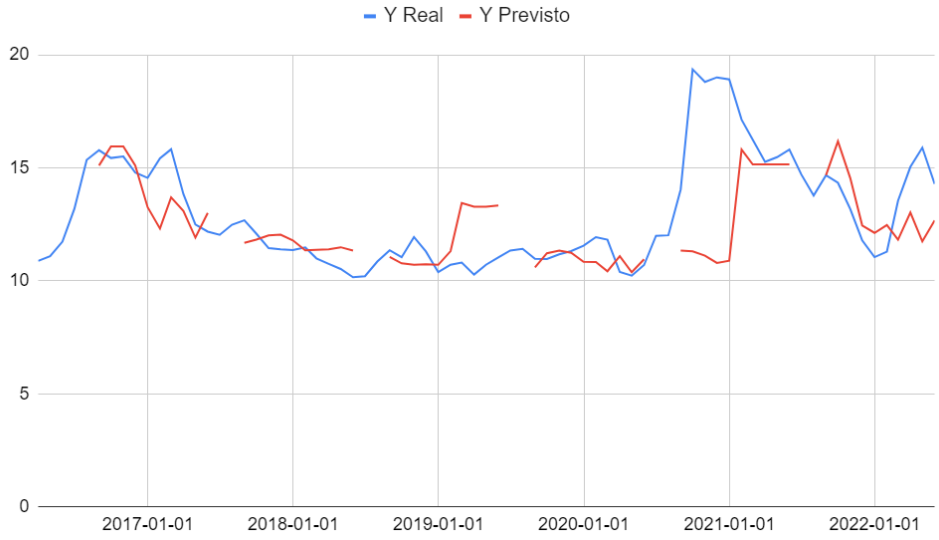
Fonte: Elaborado pelo autor.

Y_{previsto} dos modelos. Na Figura 17, apresentamos a série temporal real da variável dependente e a predição do modelo Extreme Gradient Boosting com RFE. Observamos falhas em capturar nuances, especialmente evidentes no período de 2011-06-01 a 2019-05-31 e no segundo semestre de 2020. As falhas desse último período também são observadas em todos os modelos, como ilustrado na Figura 16. Esses aumentos no erro estão possivelmente associados aos eventos extraordinários ocorridos durante a pandemia, os quais não são completamente refletidos nos atributos usados para treinar os modelos. Isso aponta para a ausência de elementos relevantes para uma previsão mais precisa nesses momentos específicos.

A Figura 18 exibe um "mapa de calor" que ilustra os resultados dos modelos desenvolvidos com e sem RFE, utilizando o procedimento de variáveis independentes simuladas, conforme detalhado em Materiais e Métodos, para a base de dados completa. A média total se aproxima bastante daquela obtida no procedimento de defasagem direta, e identificamos seis modelos com um MAPE abaixo de 17%. Esses modelos se distinguem dos seis melhores modelos do mapa anterior (Figura 16), com a inclusão do modelo "Ridge Regression" entre os melhores, enquanto o modelo "Linear Regression" é excluído do *Top 6*.

Os resultados do procedimento de variáveis independentes simuladas se assemelham aos do procedimento de defasagem direta, tanto pela média geral semelhante, quanto pelo desempenho geral significativamente inferior nos períodos de 2013-01-01 a 2020-12-31 e de 2013-06-01 a 2021-05-31. Entretanto, observa-se um desempenho insatisfatório nos dois últimos períodos, de 2014-01-01 a 2021-12-31 e 2014-06-01 a 2022-05-31, indicando uma metodologia menos eficaz para esses momentos específicos. Além disso, nota-se uma maior variação nos valores de MAPE entre diferentes períodos e modelos, sugerindo que a metodologia desse último procedimento pode não trazer benefícios para os períodos mais recentes e que os resultados

Figura 17 – Valores reais e previstos da variável dependente para o modelo extreme gradient boosting com RFE para o procedimento de defasagem direta na base de dados completa



Fonte: Elaborado pelo autor.

Figura 18 – MAPE dos modelos e períodos procedimento com variáveis independentes simuladas na base de dados completa, em ordem crescente por modelo

Model	2009-01-01 2016-12-31	2009-06-01 2017-05-31	2010-01-01 2017-12-31	2010-06-01 2018-05-31	2011-01-01 2018-12-31	2011-06-01 2019-05-31	2012-01-01 2019-12-31	2012-06-01 2020-05-31	2013-01-01 2020-12-31	2013-06-01 2021-05-31	2014-01-01 2021-12-31	2014-06-01 2022-05-31	Total
Ridge Regression RFE	7,0	20,0	3,1	10,0	6,3	7,0	5,6	3,1	35,9	11,5	8,8	9,4	10,7
Tree Regression RFE	11,5	13,4	4,0	2,4	9,4	3,2	4,6	6,5	34,0	19,7	5,1	24,5	11,5
Extreme Gradient Boosting RFE	8,6	12,2	11,0	2,4	11,8	5,1	4,3	5,8	30,2	16,8	5,2	26,0	11,6
Extreme Gradient Boosting	9,1	11,6	12,8	3,2	12,3	4,0	4,6	5,6	29,2	16,8	5,3	25,2	11,6
Tree Regression	10,7	10,8	9,3	4,4	6,2	8,1	2,7	6,4	32,2	19,9	5,5	24,7	11,7
Ridge Regression	19,2	7,6	2,6	8,6	15,3	23,4	7,5	2,0	34,2	10,7	4,4	26,5	13,5
Linear Regressão	2,7	11,0	4,5	23,4	3,7	13,9	11,2	24,8	52,2	18,8	54,3	20,0	20,0
Linear Regressão RFE	6,9	23,6	4,4	23,1	7,7	7,8	3,0	15,4	48,7	22,0	61,4	31,4	21,3
SARIMAX RFE	6,2	26,8	5,8	25,5	8,4	8,2	3,1	15,8	48,6	22,0	62,2	31,2	22,0
Unobserved Components Model RFE	6,2	26,8	5,8	25,5	8,4	8,2	3,1	15,8	48,6	22,0	62,2	31,2	22,0
Unobserved Components Model	6,3	19,2	12,8	35,0	6,3	8,2	10,8	24,3	53,4	19,3	54,0	17,0	22,2
SARIMAX	6,1	19,3	12,8	34,9	6,4	8,2	10,8	24,4	53,4	19,3	54,0	17,0	22,2
Total	8,4	16,9	7,4	16,5	8,5	8,8	5,9	12,5	41,7	18,2	31,9	23,7	16,7

Mínimo Máximo

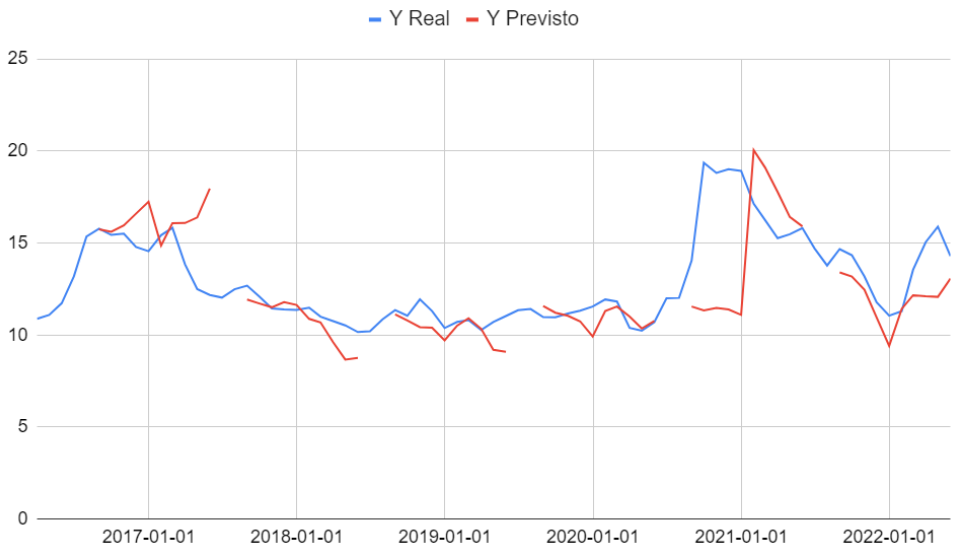
Fonte: Elaborado pelo autor.

podem variar consideravelmente entre os modelos.

É interessante ressaltar que, durante os períodos de baixo desempenho, o modelo Ridge Regression com RFE se destaca significativamente, sugerindo uma maior robustez desse modelo para a metodologia empregada. Na Figura 19, apresentamos a série temporal de Y_{real} e Y_{previsto} desse modelo, onde observamos, mais uma vez, alguns erros de tendência, especialmente no segundo semestre de 2020, semelhantes ao procedimento de defasagem direta.

As Figuras 20 e 21 apresentam os "mapas de calor" para os modelos nos procedimentos

Figura 19 – Valores reais e previstos da variável dependente para o modelo de regressão de ridge com RFE para o procedimento de variáveis independente simuladas na base de dados completa



Fonte: Elaborado pelo autor.

de defasagem direta e variáveis independentes simuladas, utilizando bases de dados contendo apenas atributos provenientes da variável dependente, ou seja, sem variáveis independentes exógenas, como descrito no final da subseção 3.2.2. Devido ao número reduzido de atributos, não aplicamos RFE nesses casos.

Figura 20 – MAPE dos modelos e períodos do procedimento de defasagem direta, da bases de dados com atributos derivados apenas da variável dependente, em ordem crescente por modelo

Model	2009-01-01 2016-12-31	2009-06-01 2017-05-31	2010-01-01 2017-12-31	2010-06-01 2018-05-31	2011-01-01 2018-12-31	2011-06-01 2019-05-31	2012-01-01 2019-12-31	2012-06-01 2020-05-31	2013-01-01 2020-12-31	2013-06-01 2021-05-31	2014-01-01 2021-12-31	2014-06-01 2022-05-31	Total
Ridge Regression	13,5	10,0	9,7	14,3	5,6	8,8	3,2	9,6	35,0	8,9	9,4	19,9	12,3
Linear Regressão	14,0	10,1	10,0	13,8	5,6	8,4	3,1	9,7	34,8	10,1	9,9	20,1	12,5
SARIMAX	13,9	9,9	12,0	12,2	6,0	7,4	2,1	8,1	36,7	15,0	16,6	18,1	13,2
Extreme Gradient Boosting	18,8	10,3	20,3	5,7	13,5	11,3	10,6	6,4	33,0	13,8	5,9	18,7	14,0
Unobserved Components Model	14,3	10,1	19,8	16,3	6,6	7,9	2,9	10,3	34,7	15,6	18,0	20,7	14,8
Tree Regression	23,1	11,8	18,3	5,8	17,1	11,4	8,5	20,6	33,5	14,5	9,3	19,8	16,1
Total	16,3	10,4	15,0	11,4	9,1	9,2	5,1	10,8	34,6	13,0	11,5	19,5	13,8

Mínimo Máximo

Fonte: Elaborado pelo autor.

Após analisar os resultados das Figuras 20 e 21, percebemos uma média geral ligeiramente abaixo de 14% em ambos os casos, o que se aproxima do desempenho dos seis melhores modelos das Figuras 16 e 18, embora o MAPE total destas seja significativamente inferior. Embora o Extreme Gradient Boosting RFE no procedimento de defasagem direta com a base completa (Figura 16) ainda se mantenha como o melhor modelo em todos os testes, o modelo Ridge Regression se destacou nas últimas duas análises realizadas com a base simplificada, atingindo um MAPE de 12,3% no procedimento de defasagem direta e apenas 10,4% no

Figura 21 – MAPE dos modelos e períodos procedimento com variáveis independentes simuladas, da bases de dados com atributos derivados apenas da variável dependente, em ordem crescente por modelo

Model	2009-01-01 2016-12-31	2009-06-01 2017-05-31	2010-01-01 2017-12-31	2010-06-01 2018-05-31	2011-01-01 2018-12-31	2011-06-01 2019-05-31	2012-01-01 2019-12-31	2012-06-01 2020-05-31	2013-01-01 2020-12-31	2013-06-01 2021-05-31	2014-01-01 2021-12-31	2014-06-01 2022-05-31	Total
Ridge Regression	5,4	20,8	5,5	6,1	7,4	7,2	4,7	2,9	33,5	10,3	12,9	8,5	10,4
Extreme Gradient Boosting	8,7	13,8	6,4	3,1	6,3	4,1	4,5	6,9	31,1	18,1	7,8	24,3	11,2
Tree Regression	8,2	15,4	6,9	5,1	8,2	6,9	4,6	5,8	27,7	18,6	4,5	25,5	11,4
Linear Regressão	5,0	31,5	6,3	5,9	4,6	6,0	5,0	8,0	42,3	14,9	28,2	18,9	14,7
Unobserved Components Model	5,7	32,7	6,6	6,3	4,8	5,6	4,9	8,5	42,6	15,8	29,5	20,8	15,3
SARIMAX	6,6	42,3	5,0	7,4	2,8	6,8	7,7	13,9	42,8	31,2	35,0	38,4	20,0
Total	6,6	26,1	6,1	5,6	5,7	6,1	5,2	7,7	36,7	18,1	19,6	22,7	13,9

Mínimo Máximo

Fonte: Elaborado pelo autor.

procedimento de variáveis independentes simuladas. Esses resultados são surpreendentes ao considerarmos o processo muito mais simples de criação da base, usando apenas atributos derivados da variável dependente “preco”.

Adicionalmente, notamos uma degradação geral no desempenho dos modelos apresentados nas Figuras 20 e 21 durante os mesmos períodos identificados nos procedimentos realizados com a base completa, demonstrado anteriormente nas matrizes das Figuras 16 e 18. Essa observação reforça a ideia de que o impacto repentino nos preços durante a pandemia de Covid-19 é desafiador para os modelos de previsão, sugerindo que as variáveis exógenas consideradas neste estudo não são capazes de explicar adequadamente essa mudança abrupta nos preços.

4.3 Discussão

Os resultados apresentados neste capítulo confirmam ser possível desenvolver modelos capazes de prever o preço do arroz com 5 meses de antecedência, atendendo aos objetivos da pesquisa. Adicionalmente, esse estudo também incorpora diversas variáveis independentes relacionadas à oferta e demanda, preenchendo, mesmo que parcialmente, uma lacuna existente nas pesquisas, uma vez que não encontramos estudos semelhantes conforme consideramos capítulo de Fundamentação e Referencial Teórico.

A Análise exploratória de dados revela informações preocupantes sobre os atributos utilizados. Por exemplo, observamos uma queda constante na renda média do brasileiro em dólares desde 2011. Também notamos que o coeficiente de variação da variável dependente “Preco” é semelhante ao do atributo “arroz_futuro”, que representa o preço futuro do arroz negociado em bolsa. Além disso, identificamos alguns *outliers* pontuais nas variáveis “Chuva” e “Volume_futuro”, bem como *outliers* resultantes de uma rápida tendência de crescimento do atributo “Adubos”, que representa o valor dos fertilizantes em dólares, possivelmente relacionado ao conflito entre Ucrânia e Rússia.

Os testes de correlação revelam que a variável dependente “Preco” não apresenta uma correlação significativa (superior a 0,7 ou inferior a -0,7) com nenhum dos atributos. Além disso, surpreendentemente, constata-se uma correlação praticamente nula entre a variável dependente e os atributos relacionados ao clima, contrariando as expectativas iniciais deste trabalho. No entanto, são identificados casos de correlação forte, tanto positiva quanto negativa, entre algumas variáveis. Destaca-se a correlação entre “Consumo” e “Renda”, que era esperada, pois a queda na renda leva à redução do consumo. Também observamos uma correlação forte e negativa entre “Desocupacao” e “FX”, indicando uma relação entre a desvalorização do real frente ao dólar americano e o aumento da desocupação, provavelmente devido aos dois serem indicadores afetados por crises. Além disso, observa-se uma correlação negativa forte entre “FX” e “Area”, possivelmente porque o aumento da taxa de câmbio impulsiona a maior produção de *commodities* de exportação, como soja e milho. No entanto, essa correlação deve ser explorada em estudos futuros, incluindo indicadores de preço e produção dessas *commodities* de exportação ou até mesmo substituindo a *commodity* na variável dependente “preco”.

Para a previsão com 5 meses de antecedência usando o procedimento de defasagem direta, observamos um MAPE geral em torno de 17%, com 6 modelos abaixo desse limite. Os modelos de destaque incluem o “Extreme Gradient Boosting” e a “Regression Tree”, com e sem RFE, além dos modelos “Ridge Regression RFE” e “Linear Regression RFE”. Uma observação interessante é notada na figura 16 e, especialmente, na 17, onde há uma redução significativa no desempenho dos modelos em alguns períodos, principalmente no segundo semestre de 2020. Isso ressalta a importância de realizar testes em estudos futuros que abranjam mais períodos e incluam atributos relevantes para os anos da pandemia.

O Procedimento de variáveis independentes simuladas apresenta um MAPE médio próximo ao procedimento anterior, também em torno de 17%. Este procedimento teve 6 modelos com MAPE abaixo do limite de 17%, com uma lista de melhores modelos semelhante ao procedimento anterior, exceto pela exclusão do modelo “Linear Regression” e inclusão do modelo “Ridge Regression”. Entretanto, observamos períodos com desempenho inferior, com maior frequência do que no procedimento anterior. Também é importante destacar que o custo computacional desse procedimento é consideravelmente maior.

Ao analisar o impacto das variáveis exógenas, nota-se, que os melhores modelos feitos com bases completas, cujos resultados estão representados nas figuras 16 e 18, possuem um MAPE inferior aos melhores modelos gerados nas bases simplificadas, que foram construídas considerando somente os atributos a partir da variável Y “preco”. Isso sugere uma vantagem para essas variáveis, embora seja bastante importante considerar o trabalho e custo adicionais associados ao manuseio de bases de dados mais extensas e complexas.

No entanto, é crucial destacar o caso específico do modelo “Ridge Regression” elaborado com o procedimento de variáveis independentes simuladas, contendo apenas atributos endógenos, ou seja, derivados da própria variável dependente Y. Este modelo apresenta um MAPE de 10,4%,

como evidenciado na figura 21, inferior aos demais modelos criados pelo mesmo procedimento na base de dados completa, destacando-se como um dos melhores modelos de todas as análises. Esse resultado desperta interesse para investigações e análises posteriores, além de levantar questionamentos sobre a viabilidade de incluir tantas variáveis independentes, considerando a complexidade na elaboração da base de dados completa, bem como a carga computacional envolvida.

Embora este estudo não seja diretamente comparável aos trabalhos de Pinheiro, Tavares e Oliveira (2017) e Rathod *et al.* (2022), descritos na seção 2.2, devido às diferenças metodológicas e regionais, é relevante destacar que ambos os estudos buscaram estimar o preço do arroz utilizando o MAPE como critério de avaliação, alcançando médias de erro inferiores às apresentadas neste estudo, com MAPE de 4% e 1,2%, respectivamente. Essa discrepância sugere a necessidade de uma investigação mais aprofundada em trabalhos futuros para compreender as razões por trás dessas diferenças.

CAPÍTULO

5

CONCLUSÃO

No início deste trabalho, na Introdução, levantamos duas questões:

1. É viável coletar e analisar dados que representem os fatores que impactam a oferta e demanda de arroz, com o propósito de desenvolver modelos de aprendizagem de máquina para prever o preço do arroz?

2. É possível desenvolver modelos de aprendizagem de máquina capazes de prever o preço do arroz com uma antecedência de 5 meses, ou seja, considerando o intervalo entre o planejamento do plantio e a venda da produção do arroz?

Nossas investigações confirmam ser possível coletar e analisar dados representando os fatores que impactam a oferta e a demanda do arroz, respondendo positivamente à primeira pergunta levantada. Além disso, estabelecemos com sucesso uma metodologia para prever o preço do arroz com 5 meses de antecedência por meio de modelos de aprendizado de máquina, respondendo também positivamente à segunda questão levantada. Assim, os resultados apresentados neste estudo confirmam a viabilidade de desenvolver modelos capazes de prever o preço do arroz com 5 meses de antecedência, atendendo aos objetivos geral e específicos propostos.

No entanto, é crucial ressaltar algumas questões relacionadas à localidade que merecem atenção em estudos subsequentes. Utilizamos dados de preço do estado do Rio Grande do Sul, mas os dados climáticos são municipais, sem representação oficial para toda a região produtora de arroz ou para o estado como um todo. Além disso, os dados de diesel estão relacionados ao preço da *commodity* negociada e não ao preço do insumo vendido aos produtores no estado. Esses aspectos podem impactar a precisão e a representatividade dos modelos, sendo cruciais para futuras pesquisas.

Outro ponto a considerar em estudos posteriores é a análise da lucratividade do produtor rural. Este estudo concentra-se apenas no preço da *commodity* do arroz negociada no estado.

Para uma compreensão mais holística, seria interessante avaliar a lucratividade, comparando o preço no momento da colheita com os custos estimados dos insumos agrícolas durante o plantio e tratamento da cultura.

Além disso, considerando o capítulo de Fundamentação e Referencial Teórico, este estudo previu o preço do arroz com 5 meses de antecedência, utilizando diversas variáveis independentes relacionadas à oferta e demanda. Isso preencheu, mesmo que parcialmente, uma lacuna existente nas pesquisas, uma vez que não encontramos estudos semelhantes. No entanto, é importante mencionar que, em trabalhos futuros, seria interessante realizar uma divisão temporal mais detalhada para avaliar os modelos com maior precisão, o que não foi possível neste estudo devido às limitações computacionais. Seria também relevante testar outros tipos de modelos de aprendizagem de máquina, como redes neurais e modelos "híbridos", que não foram explorados devido às mesmas limitações computacionais.

Na seção de Resultados dos modelos de aprendizagem de máquina, constatamos que os procedimentos de defasagem direta e de variáveis independentes simuladas geraram modelos de previsão de preço. No entanto, seria interessante explorar outras metodologias ou modelos que possam ter uma janela maior de previsão, mesmo utilizando várias variáveis independentes. Além disso, uma análise mais aprofundada dos modelos de mudança de regime de Markov seria relevante.

Utilizamos o MAPE para medir e comparar o desempenho dos modelos de aprendizagem de máquina, estabelecendo um limite de erro em 17% com base no coeficiente de variação. Em futuras pesquisas, buscar um limite de erro mais preciso com base em estudos ou melhores evidências seria uma abordagem interessante.

Por fim, conseguimos desenvolver modelos de previsão com 5 meses de antecedência utilizando dados públicos. Para estudos futuros, explorar mais variáveis e modelos, bem como realizar uma segmentação mais detalhada dos períodos, seria bastante proveitoso. Contudo, é crucial considerar que essa ampliação requererá recursos computacionais mais substanciais. Se os recursos avançados não estiverem disponíveis, talvez seja prudente reduzir o número de variáveis independentes ou dispensar o procedimento de variáveis independentes simuladas, devido à sua considerável demanda em termos de custo computacional e tempo, sem oferecer vantagens substanciais nesta análise específica.

REFERÊNCIAS

AGêNCIA NACIONAL DO PETRóLEO, GáS NATURAL E BIOCOMBUSTíVEIS. **Levantamento de Preços de Combustíveis (últimas semanas pesquisadas)**. 2022. Disponível em: <https://www.gov.br/anp/pt-br/assuntos/precos-e-defesa-da-concorrencia/precos/levantamento-de-precos-de-combustiveis-ultimas-semanas-pesquisadas#:~:text=Em%2013%2F09%2F2022%2C,os%20munic%C3%ADpios%20ser%C3%A3o%20acrescentados%20gradualmente>. Acesso em: 17 ago. 2022. Citado na página 46.

ATLAS SOCIOECONOMICO DO RIO GRANDE DO SUL. **O Rio Grande do Sul é o maior produtor de arroz em casca do Brasil**. 2022. Disponível em: <https://atlassocioeconomico.rs.gov.br/arroz>. Acesso em: 10 out. 2022. Citado na página 28.

BARBOSA, B. A. **Predição do movimento de ações da Petrobras a partir de notícias**. 72 p. Dissertação (Instituto de Ciências Matemáticas e de Computação) — Instituto de Ciências Matemáticas e de Computação, UNIFAL-MG, 2022. Disponível em: <https://www.teses.usp.br/teses/disponiveis/55/55137/tde-06092022-100818/en.php>. Acesso em: 10 out. 2022. Citado na página 24.

BARRUCHO, L. **'Tempestade perfeita'? Quatro fatos que fizeram de 2015 o ano das más notícias econômicas**. 2015. Disponível em: <https://www.bbc.com/portuguese/noticias/2015/12/151224_retrospectiva_economia_lgb>. Acesso em: 16 dez. 2022. Citado na página 59.

BASF. **Arroz: quanto tempo leva entre o plantio e a colheita?** 2022. Disponível em: <https://agriculture.basf.com/br/pt/conteudos/cultivos-e-sementes/arroz/plantio-e-colheita.html#:~:text=Elas%20podem%20ter%20tempo%20de,o%20arroz%20cultivado%20em%20sequeiro>. Acesso em: 11 out. 2022. Citado na página 30.

BELLUZZO, L. G. d. M.; FRISCHTAK, C. R.; LAPLANE, M. **Produção de Commodities e Desenvolvimento Econômico**. [s.n.], 2014. Disponível em: <https://www.eco.unicamp.br/neit/images/stories/arquivos/producao_de_commodities_e_desenvolvimento_economico.pdf>. Acesso em: 26 out. 2022. Citado na página 28.

BOTCHKAREV, A. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. **Interdisciplinary Journal of Information, Knowledge, and Management**, v. 14, p. 45–79, 2019. Disponível em: <https://doi.org/10.28945/4184>. Acesso em: 17 ago. 2022. Citado na página 41.

CARVALHO, R.; BISTON, J. V.; FAVAN, R.; DEOLINDO, D. Avaliação de algoritmos de machine learning na cotação do preço do contrato futuro de milho. **Revista Eletrônica e-F@TEC**, Fatec Shunji Nishimura, v. 11, n. 1, 2021. Disponível em: <https://pesquisafatec.com.br/ojs/index.php/efatec/article/view/249>. Acesso em: 15 out. 2022. Citado na página 34.

CELASUN, O.; MILESI-FERRETTI, G. M.; OBSTFELD, M. Cinco gráficos que explicam a economia mundial em 2018. **IMF Blog**, dez. 2018. Disponível em: <https://www.imf.org/pt/Blogs/Articles/2018/12/20/blog122018-5-charts-that-explain-the-global-economy-in-2018>. Acesso em: 16 dez. 2022. Citado na página 59.

CENTRO DE ESTUDOS AVANçADOS EM ECONOMIA APLICADA. 2022. Disponível em: <https://www.cepea.esalq.usp.br/br/arroz/>. Acesso em: 20 ago. 2022. Citado nas páginas 29, 45 e 47.

CHEIN, F. **Introdução aos modelos de Regressão Linear**: Um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas. [s.n.], 2019. Escola Nacional de Administração Pública - ENAP. Disponível em: <https://repositorio.enap.gov.br/bitstream/1/4788/1/Livro_Regress%C3%A3o%20Linear.pdf>. Acesso em: 26 Nov. 2022. Citado na página 35.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. **KDD**, 2016. Disponível em: <https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf>. Acesso em: 30 Nov. 2022. Citado na página 38.

COMPANHIA NACIONAL DE ABASTECIMENTO. **POLÍTICA DE GARANTIA DE PREÇOS MÍNIMOS**. 2017. Disponível em: <https://www.conab.gov.br/precos-minimos>. Acesso em: 15 out. 2022. Citado na página 30.

____. **Insumos Agropecuários**. 2022. Disponível em: <https://consultaweb.conab.gov.br/consultas/consultaInsumo.do?method=acaoListarConsulta>. Acesso em: 12 Nov. 2022. Citado na página 45.

____. **Portal de Informações**. 2022. Disponível em: <https://portaldeinformacoes.conab.gov.br>. Acesso em: 11 Nov. 2022. Citado na página 45.

FERREIRA, C. M.; WANDER, A. E.; SILVA, O. F. d. **Mercado, comercialização e consumo do Cultivo de Arroz**. 2021. Disponível em: <https://www.embrapa.br/agencia-de-informacao-tecnologica/cultivos/arroz/pre-producao/socioeconomia/importancia-economica-e-social>. Acesso em: 24 nov. 2022. Citado na página 28.

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. **Food and agriculture data**. 2023. Disponível em: <https://www.fao.org/faostat>. Acesso em: 10 out. 2022. Citado nas páginas 27, 28 e 29.

GALYFIANAKIS, G.; DRIMBETAS, E.; SARIANNIDIS, N. Modeling energy prices with a markov-switching dynamic regression model: 2005-2015. **Bulletin of Applied Economics**, Risk Market Journals, v. 3, n. 1, p. 11–28, 2016. Disponível em: <https://ideas.repec.org/a/rmk/rmkbae/v3y2016i1p11-28.html>. Acesso em: 17 out. 2022. Citado nas páginas 34 e 48.

GUGLIELMETTI, L. C. Commodity: formação de preço muito além da oferta e demanda. **Jusbrasil**, Set. 2016. Disponível em: <https://jus.com.br/artigos/52401/commodity-formacao-de-preco-muito-alem-da-oferta-e-demanda>. Acesso em: 26 out. 2022. Citado na página 28.

HASAN, M. M.; ZAHARA, M. T.; SYKOT, M. M.; NUR, A. U.; SAIFUZZAMAN, M.; HAFIZ, R. Ascertaining the fluctuation of rice price in bangladesh using machine learning approach. In: **2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)**. [s.n.], 2020. p. 1–5. Disponível em: <https://doi.org/10.1109/ICCCNT49239.2020.9225468>. Acesso em: 10 out. 2022. Citado na página 32.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATíSTICA. **Sistema IBGE de Recuperação Automática - SIDRA**. 2021. Disponível em: <https://sidra.ibge.gov.br/tabela/1612>. Acesso em: 17 ago. 2022. Citado nas páginas 28 e 45.

____. **Pesquisa Nacional por Amostra de Domicílios Contínua**. 2023. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/trabalho/9171-pesquisa-nacional-por-amostra-de-domicilios-continua-mensal.html?edicao=34731&t=series-historicas>. Acesso em: 25 abr. 2023. Citado na página 46.

INSTITUTO NACIONAL DE METEOROLOGIA. **Estações Automáticas**. 2022. Disponível em: <https://portal.inmet.gov.br/servicos/esta%C3%A7%C3%B5es-autom%C3%A1ticas>. Acesso em: 10 Nov. 2022. Citado na página 45.

INSTITUTO RIO GRANDENSE DO ARROZ. **Estimativas Safra 2021/2022**. 2022. Disponível em: <https://irga.rs.gov.br/irga-faz-projecao-do-custo-de-producao-da-safra-2021-2022>. Acesso em: 16 Nov. 2022. Citado na página 45.

JUNIOR, O. O guia do xgboost com python. **Data Science Machine Learning Python**, 2022. Disponível em: <https://dadosaocubo.com/o-guia-do-xgboost-com-python/>. Acesso em: 30 Nov. 2022. Citado na página 39.

KNAAK, J.; PINTO, W. d. P. Aplicação do modelo sarimax para modelar e prever a concentração de material particulado inalável, no espírito santo, brasil. **Ciência Matura**, 2022. Disponível em: <https://doi.org/10.5902/2179460X63466>. Acesso em: 30 Nov. 2022. Citado nas páginas 39 e 40.

LUDOVICO, S. N. **Previsão de indicadores diários de preços no mercado futuro de commodities agrícolas utilizando aprendizagem de máquina**. 155 f. Dissertação (Programa de Pós-Graduação em Estatística Aplicada e Biometria) — Universidade Federal de Alfenas, UNIFAL-MG, 2020. Disponível em: <https://bdtd.unifal-mg.edu.br:8443/handle/tede/1762>. Acesso em: 10 out. 2022. Citado na página 23.

MARCHEZAN, A.; SOUZA, A. M. Previsão do preço dos principais grãos produzidos no rio grande do sul. **Ciencia Rural**, v. 40, n. 11, p. 2368–2374, nov. 2010. Disponível em: <https://doi.org/10.1590/S0103-84782010001100019>. Acesso em: 15 out. 2022. Citado na página 33.

MATLAB. **Train Regression Trees Using Regression Learner App**. 2023. Disponível em: <https://www.mathworks.com/help/stats/train-regression-trees-using-regression-learner-app.html>. Acesso em: 17 ago. 2022. Citado na página 37.

NETO, C. R.; SILVA, F. d. A. C.; ARAúJO, L. V. d. **Qual é a participação da agricultura familiar na produção de alimentos no Brasil e em Rondônia?** 2020. Disponível em: <https://www.embrapa.br/busca-de-noticias/-/noticia/55609579/artigo---qual-e-a-participacao-da-agricultura-familiar/_-na-producao-de-alimentos-no-brasil-e-em-rondonia>. Acesso em: 23 nov. 2022. Citado nas páginas 28 e 30.

OLIVEIRA, W. R. S. d.; CECHIN, A. Efeitos da pandemia da covid-19 nos preÇos dos alimentos no brasil. **Revista Catarinense De Economia**, v. 5, n. 2, p. 141–155, 2021. Disponível em: <https://doi.org/10.54805/RCE.2527-1180.v5.n2.109>. Acesso em: 26 out. 2022. Citado na página 29.

PESSOA, S. **Pandemia e crise econômica: primeiro ano**. 2021. Disponível em: <https://blogdoibre.fgv.br/posts/pandemia-e-crise-economica-primeiro-ano>. Acesso em: 16 dez. 2022. Citado na página 59.

PINHEIRO, D. R. O.; TAVARES, M.; OLIVEIRA, K. G. de. Previsão de preços para a cultura do arroz irrigado e seco do estado do paraná utilizando séries temporais. In: **Anais do Congresso Contabilidade, Gestão e Agronegócio**. Uberlândia, MG: [s.n.], 2017. Disponível em: <https://eventos.ufu.br/sites/eventos.ufu.br/files/documentos/9614_-_previsao_de_precos_para_a_cultura_do_arroz_irrigado_e_sequeiro_do_estado_do_parana_utilizando_series_temporais.pdf>. Acesso em: 10 out. 2022. Citado nas páginas 30, 31 e 66.

PORTO, B. M. **Revisão bibliométrica sobre modelagem e previsão do preço da soja: uma comparação entre os modelos ARIMAX, redes neurais e máquina de aprendizado extremo**. Tese (Doutorado) — Fundação Universidade Federal de Mato Grosso do Sul, 2021. Disponível em: <https://repositorio.ufms.br/handle/123456789/3677>. Acesso em: 15 out. 2022. Citado na página 33.

PORTO, B. M. Previsão de preços das commodities agrícolas: uma revisão bibliométrica sobre modelos. **Revista De Gestão E Secretariado (Management and Administrative Professional Review)**, v. 13, n. 3, p. 881–912, 2022. Disponível em: <https://doi.org/10.7769/gesec.v13i3.1380>. Acesso em: 10 out. 2022. Citado na página 33.

QUANDL. **Financial Data API**. 2022. Disponível em: <https://demo.quandl.com/tools/api>. Acesso em: 15 jun. 2022. Citado na página 45.

RATHOD, S.; CHITIKELA, G.; BANDUMULA, N.; ONDRASEK, G.; RAVICHANDRAN, S.; SUNDARAM, R. M. Modeling and forecasting of rice prices in india during the COVID-19 lockdown using machine learning approaches. **Agronomy**, v. 12, n. 9, p. 2133, 2022. Disponível em: <https://doi.org/10.3390/agronomy12092133>. Acesso em: 10 out. 2022. Citado nas páginas 31, 32 e 66.

SCIKIT-LEARN.ORG. **Decision Tree Regression**. 2023. Disponível em: <https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html#sphx-glr-auto-examples-tree-plot-tree-regression-py>. Acesso em: 17 ago. 2022. Citado na página 38.

SELUKAR, R. **Time Series Modeling with Unobserved Components**. 2016. Disponível em: <https://forecasters.org/wp-content/uploads/gravity_forms/7-621289a708af3e7af65a7cd487aee6eb/2016/07/Selukar_Rajesh_ISF2016.pdf>. Acesso em: 30 Ago. 2022. Citado na página 40.

SENAR. Guerra rússia-ucrânia: o panorama do abastecimento de fertilizantes. **CNABrasil**, mar. 2022. Disponível em: <https://cnabrasil.org.br/noticias/guerra-russia-ucrania-o-panorama-do-abastecimento-de-fertilizantes>. Acesso em: 17 ago. 2022. Citado na página 54.

SHAH, J.; VAIDYA, D.; SHAH, M. A comprehensive review on multiple hybrid deep learning approaches for stock prediction. **Intelligent Systems with Applications**, v. 16, p. 200111, 2022. ISSN 2667-3053. Disponível em: <https://doi.org/10.1016/j.iswa.2022.200111>. Acesso em: 15 out. 2022. Citado na página 33.

SILVA, O. F. d.; WANDER, A. E.; FERREIRA, C. M. **Importância Econômica e Social do Cultivo de Arroz**. 2021. Disponível em: <https://www.embrapa.br/agencia-de-informacao-tecnologica/cultivos/arroz/pre-producao/socioeconomia/importancia-economica-e-social>. Acesso em: 24 nov. 2022. Citado nas páginas 27 e 28.

SOUZA, F. M.; PRADO, T. N. d.; WERNECK, G. L.; LUIZ, R. R.; MACIEL, E. L. N.; FAERSTEIN, E.; TRAJMAN, A. Classification and regression trees for predicting the risk of a negative test result for tuberculosis infection in brazilian healthcare workers: a cross-sectional study. **Revista brasileira de epidemiologia**, 2021. Disponível em: <https://doi.org/10.1590/1980-549720210035>. Acesso em: 28 Nov. 2022. Citado nas páginas 36 e 37.

TREVIZAN, K. **Brasil enfrenta pior crise já registrada poucos anos após um boom econômico**. 2017. Disponível em: <https://g1.globo.com/economia/noticia/brasil-enfrenta-pior-crise-ja-registrada-poucos-anos-apos-um-boom-economico.ghtml>. Acesso em: 16 dez. 2022. Citado na página 59.

WIERINGEN, W. N. van. Lecture notes on ridge regression. **arXiv preprint arXiv:1509.09169**, 2015. Disponível em: <https://arxiv.org/abs/1509.09169>. Acesso em: 26 Nov. 2022. Citado na página 36.

YAHOO. **Rough Rice Futures,Nov-2023 (ZR=F)**. 2022. Disponível em: <https://finance.yahoo.com/quote/ZR%3DF/history?period1=946857600&period2=1643673600&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true&guccounter=1>. Acesso em: 15 jun. 2022. Citado na página 46.


ICMC
USP
SÃO CARLOS