Previsão de custos e preços de commodities agrícolas: Uma abordagem com Modelos Estatísticos e Aprendizado de Máquina

Fernanda A. Hamatsu[1], **Ivan C. A. de Oliveira**[1]

[1]Ciência da Computação
Faculdade de Computação e Informática
Universidade Presbiteriana Mackenzie
São Paulo – SP – Brasil

0395952@mackenzista.com.br, ivan.oliveira@mackenzie.br

Resumo. *Este trabalho tem como tema a previsão de custos e preços dos cultivos agrícolas soja, milho e trigo, visando obter dados de até 2 anos à frente através do uso de modelos estatísticos e aprendizado de máquina que possam auxiliar o produtor em seu posicionamento estratégico no mercado, calculando os preços de commodities. Será realizada a análise exploratória dos dados para encontrar traços de sazonalidade e variáveis que proporcionem melhorias significativas ao modelo. Serão explorados modelos simples estatísticos como Regressão Linear, como também modelos estatísticos avançados como o Auto Regressive Integrated Moving Average (ARIMA) e modelos que aplicam Machine Learning como o XGBoost.*

Palavras-chave: *Agricultura, Previsão de preços e cultivos , aprendizado de máquina.*

1. Introdução

A agricultura no Brasil representa aproximadamente 25% de seu Produto Interno Bruto (PIB), um mercado que cresce a cada ano e apresenta um grande potencial para soluções tecnológicas e inovadoras. Soluções utilizando dados podem auxiliar os agricultores em diversas áreas, sendo uma delas sua estratégia comercial para precificação de seus cultivos e posicionamento de mercado, visando preços mais justos e que o proporcionem grandes margens de lucro.

O mercado de agricultura realiza seus planejamentos em períodos de safra ou safrinha, a depender de seus cultivos, o que os torna menos flexíveis a ajustes de preços de forma semanal ou mensal. Dessa forma, realizar a previsão de custos e preços utilizando dados históricos se torna uma solução para o agricultor, apresentando a ele uma visão de futuro e o auxiliando na precificação de seus cultivos.

Para realizar a previsão de preços de cultivos será necessária uma base com dados históricos dos custos de commodities agrícolas, como também de fatores que os influenciam como câmbio e custo do plantio. Técnicas de Séries Temporais (JAMES DURBIN, 2012) com modelos estatísticos e de aprendizado de máquina serão utilizadas para obtenção das predições a depender de suas métricas de assertividade como Mean Absolute Percentage Error (MAPE) e Root Mean Squared Error (RMSE) (BOTCHKAREV, 2019).

2. Referencial Teórico

As soluções digitais inovadoras para o mercado agrícola, que representa 23,8% do PIB brasileiro (CENTRO DE ESTUDOS AVANÇADOS EM ECONOMIA APLICADA (CEPEA), 2024), tem sido exploradas pelas grandes empresas e industrias, visando extrair mais do potêncial economico dentre as diversas áreas que compoem este mercado (ANTÔNIO DA LUZ, 2023). Do ponto de vista do agricultor, levam em consideração fatores como: preparo do solo (fertilizantes e defensivos agrícolas), semeadura (custos de sementes), irrigação, controle de pragas, colheita, maquinário (combustíveis utilizados) e a mão de obra utilizada durante todo processo. (EMBRAPA, 2006)

Produtos agrícolas fazem parte das *soft commodities* ou commodities leves, as quais podem ser separadas entre Grãos (soja, café, milho e trigo) e Proteína animal (carne, leite e derivados), para este trabalho serão estudados grãos visto que o Brasil é o maior produtor de soja do mundo (EMBRAPA, 2023) com a produção de 154.566,3 milhões de toneladas e a produtividade de 3.508 kg/ha (CONAB) na safra 2022/23.

Índice de Preço das Exportações - Abril de 2024

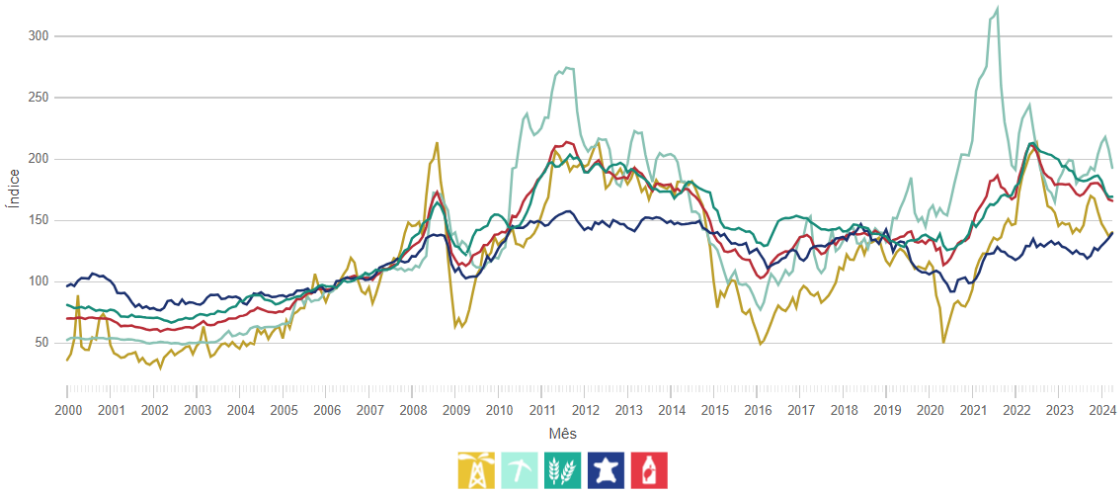

Figure 1. Fonte: Ministério do Desenvolvimento, Indústria, Comércio e Serviços

Consumo interno e exportação de soja brasileira em 2022

Discriminação	Exportação Mil Ton	Consumo Mil Ton	Total Mil Ton	Exportação Milhões US$
Soja em grão	78.730	50.932	129.662	46.664
Farelo de soja	20.353	18.661	39.014	10.336
Óleo de soja	2.597	7.342	9.939	3.927
Total	101.680	76.935	178.615	60.927

Figure 2. Fonte: Abiove; Ministério da Economia/Secex.

Em uma pesquisa realizada pela EMBRAPA, a quantidade e diversidade de dados disponíveis no agronegócio têm o potencial de causar profundas transformações nas pesquisas e inovações na agricultura, apontando em quais áreas a aplicação de Ciência de Dados resultariam em grandes benefícios a pesquisadores, agricultes e agentes públicos. (SOUZA et al., 2017).

Interaçào entre as respectivas áreas de Ciência de Dados na Agricultura

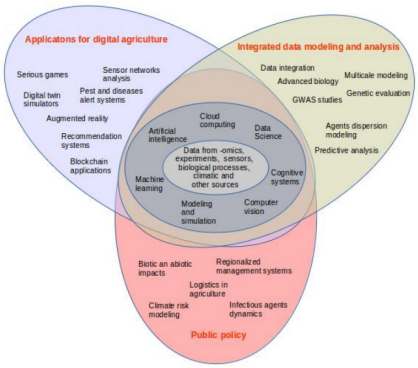

Figure 3. Fonte: SOUZA, K. X. S. de; TERNES, S.; OLIVEIRA, S. R. de M.; MOURA, M. F.; BARIONI, L. G.; HIGA, R. H.; FASIABEN, M. do C. R.

3. Materiais e Métodos

3.1. Solução

Foram utilizados como base dados históricos de cotações de cada cultivo, obtidos através do Portal Agrolink que disponibiliza uma base com valores históricos a partir de 2004 e que são atualizados mensalmente, podendo ser filtrados por Estado e tipo de produto (como sacas de diferentes volumes). Através destes dados foi possível realizar a previsão univariada da cotação de cada cultivo, como também poderão ser aplicados dados como câmbio monetário, obtido através da bibliteca do Banco Central (*python-bcb*), custos de combustível obtidos através da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis, e outras possíveis variáveis.

A aplicação experimental foi desenvolvida através da linguagem de programação Python, visto que ela fora a linguagem mais utilizada em 2023 (IEEE SPECTRUM) e a mais utilizada para ciência de dados (FLATIRON SCHOOL) devido a diversidade de bibliotecas para manuseio de dados e o desenvolvimento de modelos preditivos de forma otimizada. Os seguintes modelos foram utilizados durante a pesquisa:

- Auto Regessivo (*Auto Regressive*): modelo estatístico que utiliza variação da análise de Regressão Linear (CRUZ-RAMÍREZ et al., 2023) (CRUZ-RAMÍREZ et al., 2023) para prever as futuras sequências para uma variável em um determinado período de tempo. (MURTAZA DALAL ALEXANDER C. LI, 2019)
- ARMA (*Auto-Regressive Moving Average*) & ARIMA (*Auto-Regressive Integrated Moving Average*): modelos estatísticos avançados que levam em consideração a auto-regressão da média móvel da variável a ser prevista, sendo que no ARIMA é aplicada a Integrada (I) que permite a retirada da não-estacionariedade do modelo. (BIANCA REICHERT, 2020)

- Prophet (Meta): framework open-souce criado pelo time de Cientistas de Dados da Meta para previsão de dados de séries históricas que apresentam forte sazonalidade, aplicando sobre os modelos variáveis como feriados e datas históricas.(TRIEBE et al., 2021)
- XGBoost: biblioteca de *Gradient Boosting* que implementa modelos de *Machine Learning* para solucionar problemas de ciência de dados como previsões de séries temporáis. (CHEN; GUESTRIN, 2016)

Para realizar a avaliação de resultados dos modelos e a escolha dos melhores resultados foram aplicadas as métricas de válidação de performance *Mean Absolute Percentage Error* (MAPE) [Figura 4] ou Erro Percentual Médio Absoluto, e a *Root Mean Squared Error* (RMSE) [Figura 5] ou Raiz do Erro Quadrático Médio, utilizados em modelos de previsão (CHAI, 2014).

$$MAPE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i + \hat{y}_i|}{max(\epsilon, |y_i|)}$$

Figure 4. Equação do Erro Percentual Absoluto.

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Figure 5. Equação da raiz do erro quadrático médio.

3.2. Coleta de Dados

A partir do histórico de cotações dos cultivos Soja, Milho, Trigo e Algodão disponíveis através do Portal Agrolink, foi desenvolvido um algoritmo de web-scaping para realizar a coleta e criação das bases a serem utilizadas nos modelos. Utilizamos da biblioteca BeautifulSoup para desenvolver o algoritmo que obtem os dados dos cultivos a partir de Abril de 2004 até a última data atualizada pelo portal. Para variáveis independentes do modelo como os preços das commodities na Bolsa de Chicago, Gás Natural e Biocombustíveis foram obtidos os valores manualmente através de portais especializados em cada área e entregues em tabelas. A variável de câmbio foi obtida através da Biblioteca do Banco Central disponível na linguagem Python. Os dados de Precipitação e Temperatura dos Estados Mato Grosso e Paraná, maiores produtores dos cultivos estudados, foram obtidos através do Instituto Nacional de Metereologia (INME).

3.3. Preparação de Dados

Os dados obtidos para o preço histórico das cotações e os dados das variáveis independetes obtidos manualmente foram agrupados em dataframes unicos para cada cultivo, assim resultando em 4 tabelas, com as colunas:

1. Data: Data para cada linha (formato Internacional)

2. Estadual: Preço da commoditie no Estado líder na produção de cada cultivo
3. País: Preço da commoditie no País
4. Id: Identificação de cada cultivo
5. Último: Valor de encerramento do mês da commoditie na Bolsa de Valores de Chicago
6. Abertura: Valor de abertura do mês da commoditie na Bolsa de Valores de Chicago
7. USD: Valor do Dólar em Reais
8. Precipitação: Média mensal dos dados de precipitação dos Estados
9. Temperatura: Média mensal dos dados de temperatura dos Estados
10. lag_Estado: Valores com *lag* -1 ou valores do mês anterior.

A análise exploratória dos dados (EDA) foi realizada através da observação dos gráficos de linha para cada cultivo, comparando com as variáveis independentes e identificando similaridades, além do estudo de gráficos de Sazonalidade e Tendência [Figura 7], Função de Auto-Correlação [Figura 6] e *Heatmap* de Correlação [Figura 8].

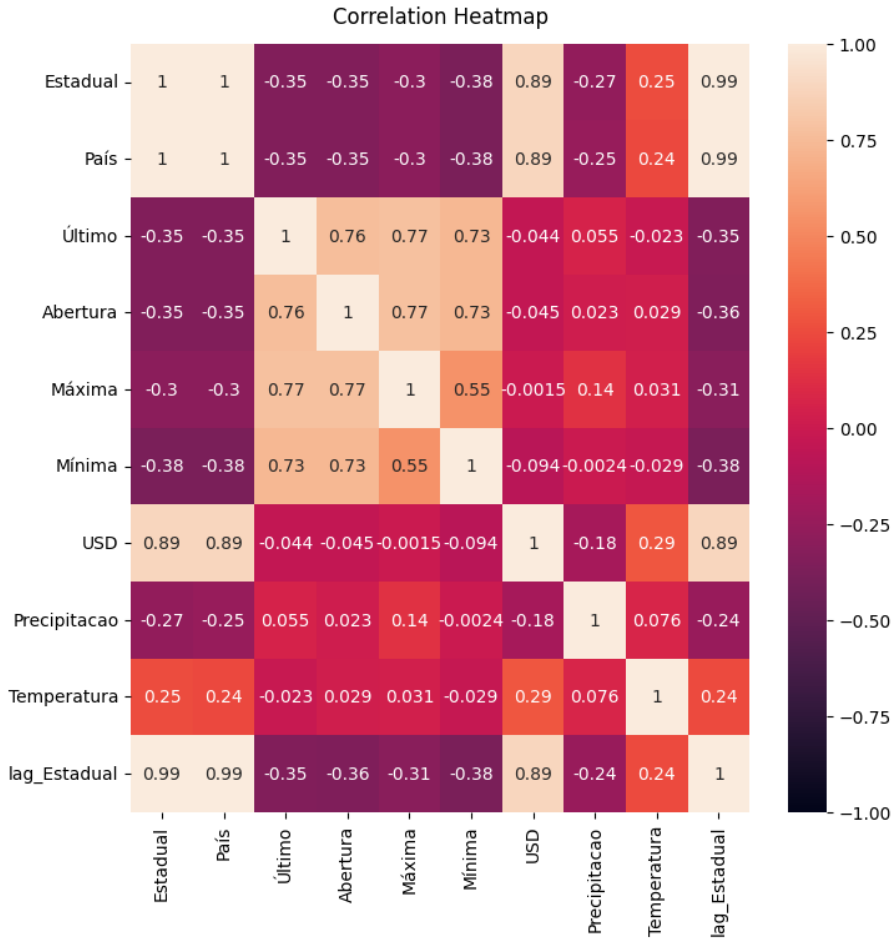

Figure 6. Heatmap para os dados da Soja

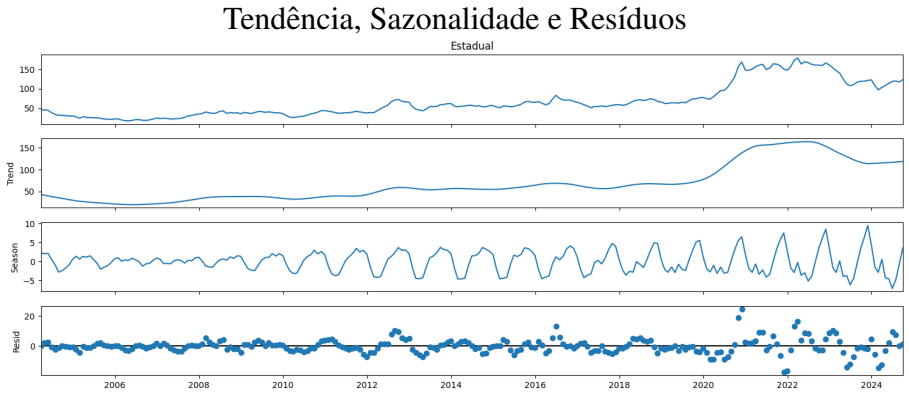

Figure 7. Análise dos dados da Soja

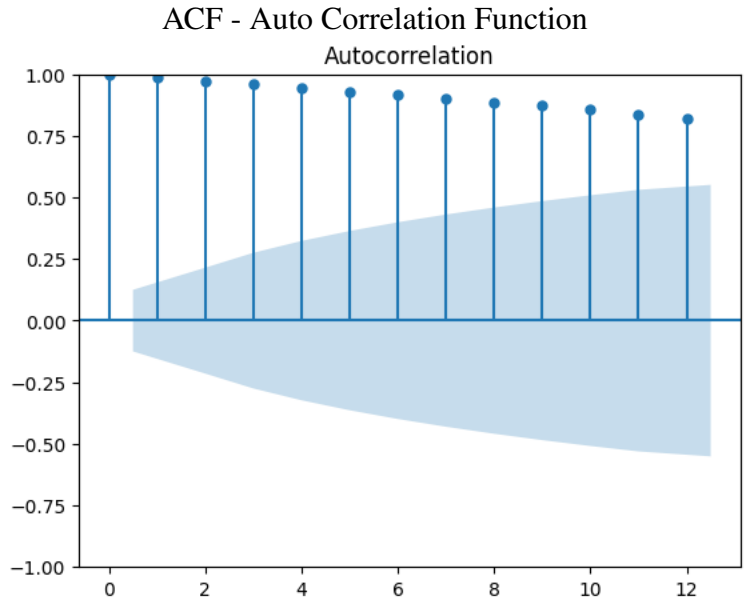

Figure 8. Análise dos dados da Soja

Para cada modelo foram realizadas alterações internas dos dataframes para que fossem lidos corretamente, como alterações nos formatos das datas ou nomes das colunas.

4. Resultados parciais

Esta seção apresenta os resultados obtidos da pesquisa, focando em modelos predutivos univariados e multivariados para previsão dos preços das Commodities Agrícolas em seus Estados de maior produtividade. Serão detalhados os resultados obtidos dos modelos Univariados, modelos Multivariados, e por fim serão apresentadas as métricas para cada modelo estudados para o estudo.

4.1. Algoritmos Univariados

A seguir, serão apresentados os resultados dos modelos univariados, subdivididos entre em: modelo Auto Regressivo, modelo Auto Regressivo Integrado de Médias Móveis, modelo XGBoost e modelo Prophet.

4.1.1. Auto Regressivo

O modelo univariado Auto Regressivo foi desenvolvido utilizando da biblioteca statsmodels utilizando da função AutoReg, os dados utilizados foram as colunas de Data e Estadual, sem filtros e alterações. Os resultados apresentaram baixa performance e foram utilizados como base para comparação com os modelos avançados posteriormente testados, conforme a Figura 9.

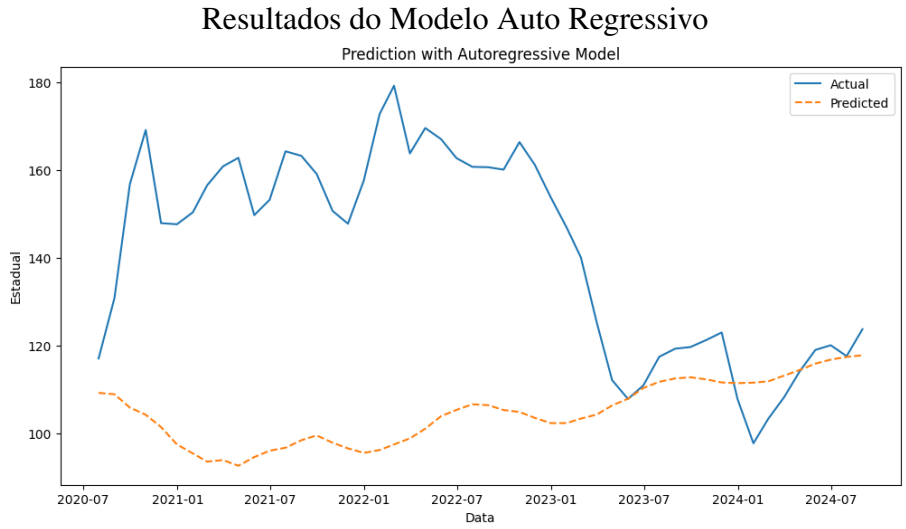

Figure 9. Previsão para Soja

4.1.2. ARIMA

O modelo univariado ARIMA ou *Auto-Regressive Integrated Moving Average* foi desenvolvido utilizando da biblioteca pmdarima utilizando da função autoarima, o qual encontra os melhores parâmetros p, d e q onde:

- p: ordem do modelo auto-regressivo
- d: grau de diferenciação
- q: ordem do modelo de média móvel

Inicialmente o modelo foi testado utilizando os dados Data e Estadual sem alterações, onde os dados analisados eram não estacionários o que estabelece os valores de grau de diferenciação como 0. Este modelo apresentou uma performance baixa com métricas altas, conforme a figura 10 e os resultados das métricas observados na tabela 1:

Para lidar com isso, foi realizada a transformação dos dados da variável Estado de não-estacionários para estacionáios, assim podendo ser utilizados diferentes valores para o grau de diferenciação do modelo. Este novo modelo apresentou uma performance baixa mas devido a sua proximidade com a curva de dados reais as suas métricas, na tabela 1 foram baixos, conforme a figura 11:

Resultados do Modelo Auto-Arima Univariado

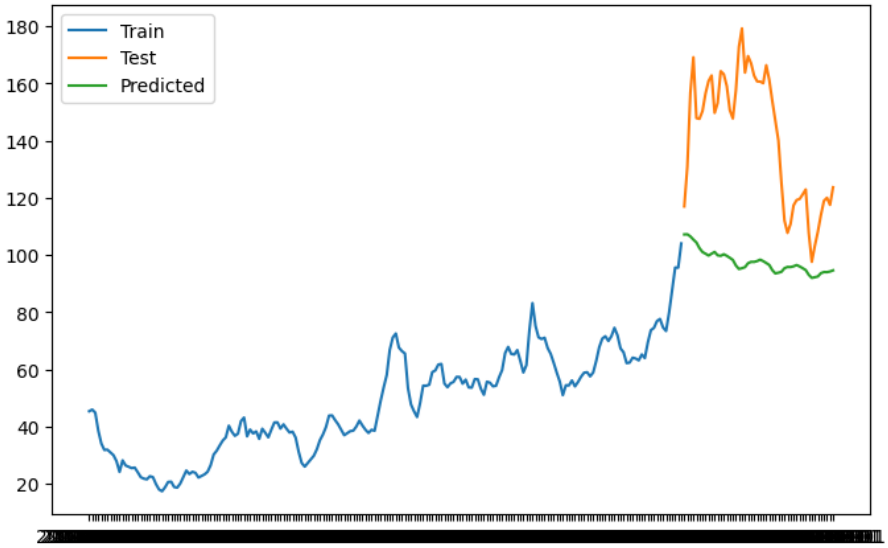

Figure 10. Previsão para Soja

Resultados do Modelo Auto-Arima Univariado com Dados Estacionários

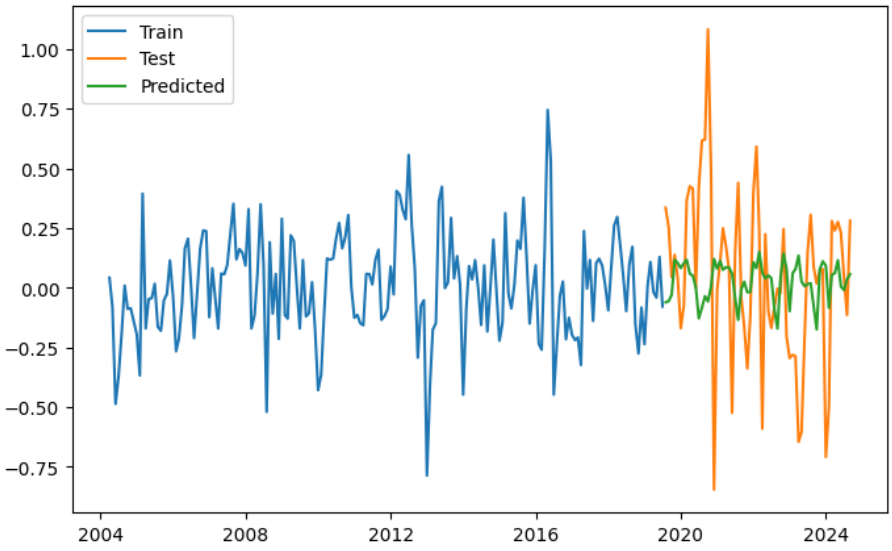

Figure 11. Previsão para Soja

4.1.3. XGBoost

O modelo *XGBoost* implementa algoritmos de Machine Learning e Gradient Boosting foi inicialmente testado para obter as previsões da variável objetivo Preço Estadual, obtendo a melhor performance dentre os modelos univariados sem a necessidade do processamento dos dados de não-estacionários para estacionários. Os resultados apresentavam forte sazonalidade com ciclos periódicos e repetitivos, assim mesmo com métricas boas o modelo não apresentavam resultados realistas, conforme a figura 12.

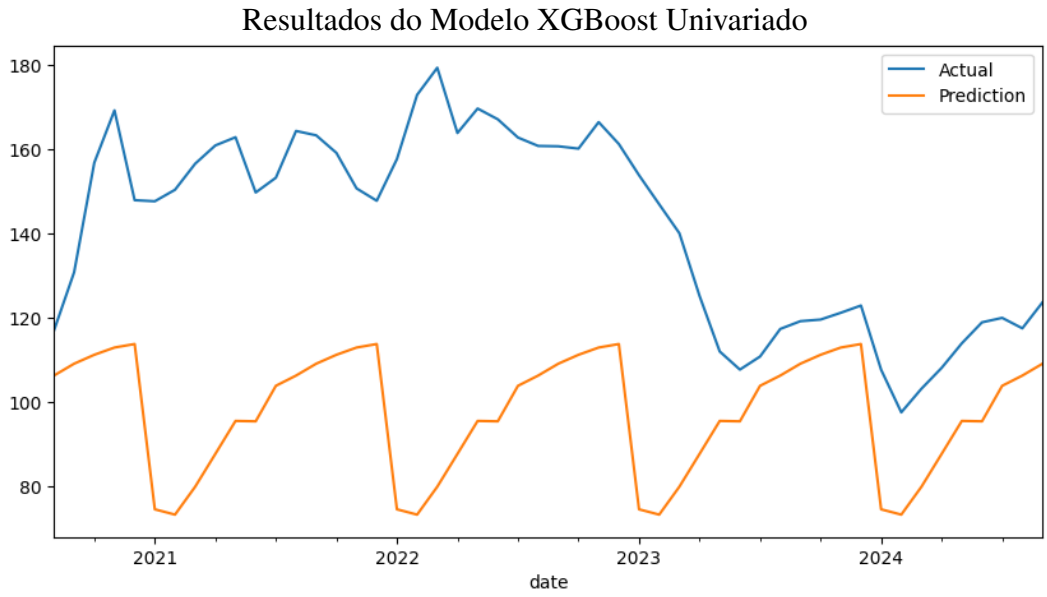

Figure 12. Previsão para Soja

4.1.4. Prophet - Meta

O modelo *Prophet* desenvolvido pela Meta para dados que apresentam forte sazonalidade e grandes quantidades de dados, para previsão univariada o modelo, observado na figura 13, apresentou uma baixa performance em comparação ao Auto-ARIMA e o XGBoost, mas apresentou o menor tempo de execução dentre os modelos testados.

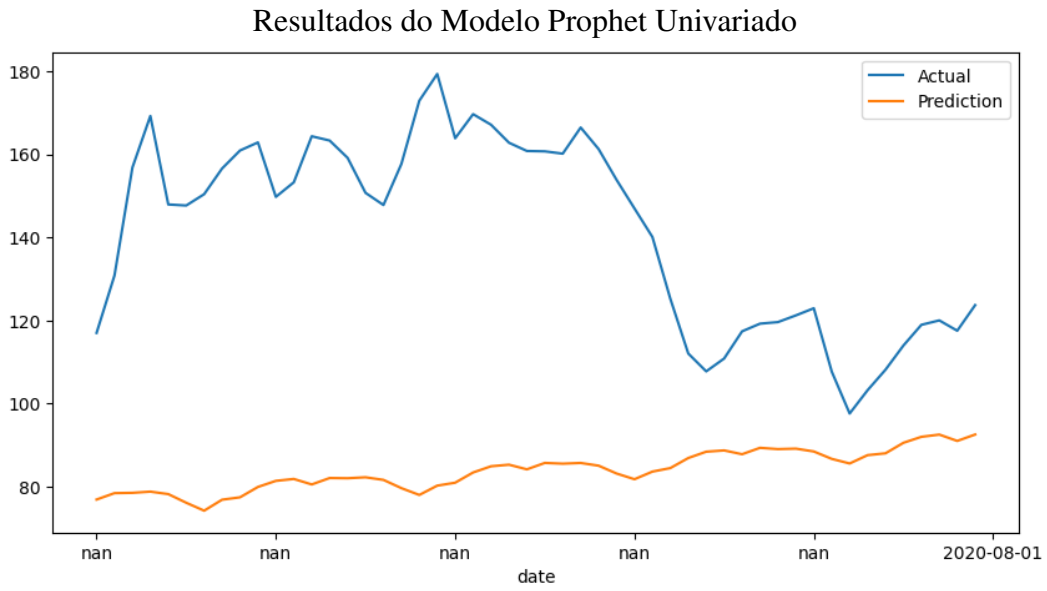

Figure 13. Previsão para Soja

4.2. Algoritmos Multivariados

A seguir, serão apresentados os resultados dos modelos multivariados, que utilizam variáveis além da variável objetivo Preço das Commodieties por Estado, como: preço médio das commodities no País, Câmbio (proporção entre Real e USD, preços das commodities na Bolsa de Chicago, médias de Precipiação e Temperatura Estaduais e o lag). Eles serão subdivididos entre em: modelo Auto Regressivo Integrado de Médias Móveis, modelo XGBoost e modelo Prophet.

4.2.1. ARIMA

Utilizando do modelo *Auto-ARIMA* novamente, variáveis independentes foram adicionadas afim de obter melhores resultados dos modelos. A partir da análise dos dados e a seleção manual das variáveis que se apresentavam correlacionadas à variável objetivo, foram utilizadas as variáveis: USD, País, Última, Máxima, Preciptação e lag_Estadual. Os resultados apresentaram melhora significativa em comparação aos resultados obtidos através dos modelos univariados, conforme a figura 14.

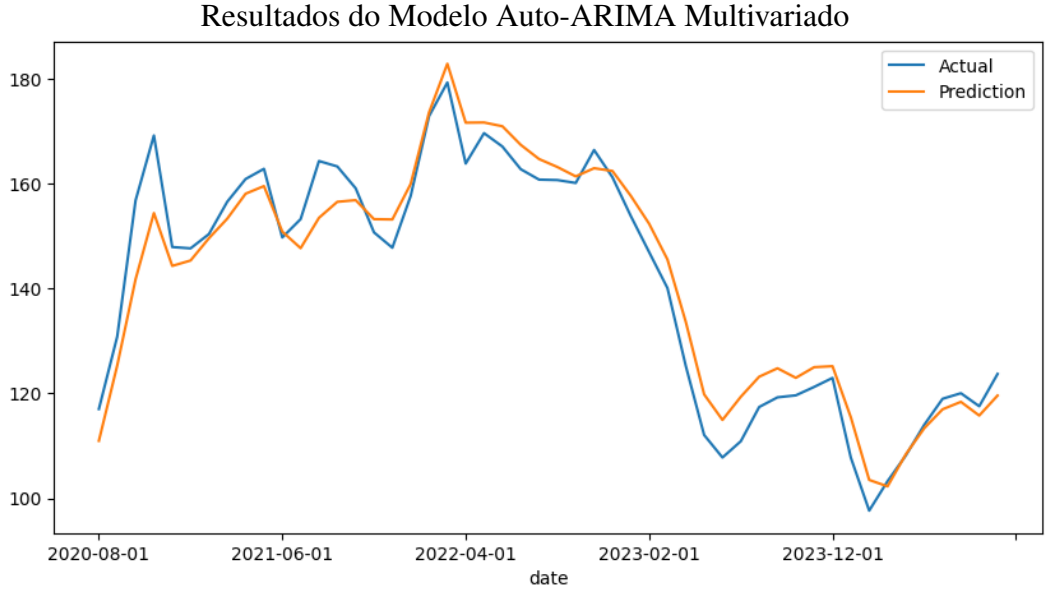

Figure 14. Previsão para Soja

4.2.2. XGBoost

Utilizando do modelo *XGBoost* novamente, as variáveis independentes USD (figura 15) e País (figura 16) foram adicionadas afim de obter melhores resultados dos modelos. Apesar de ambas variáveis indicarem melhorias ao modelo, quando o modelo é executado com as duas variáveis simultaneamente ele perde performance signifcativamente (figura 17).

Resultados do Modelo XGBoost com USD

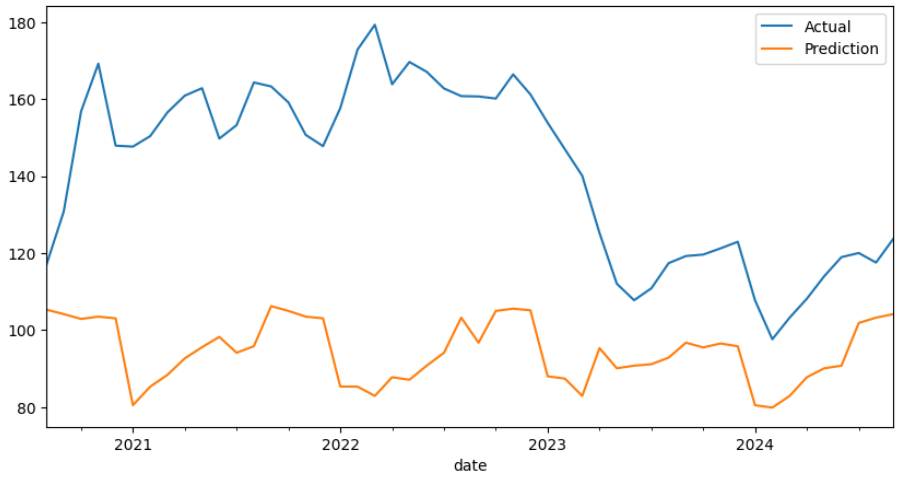

Figure 15. Previsão para Soja

Resultados do Modelo XGBoost com País

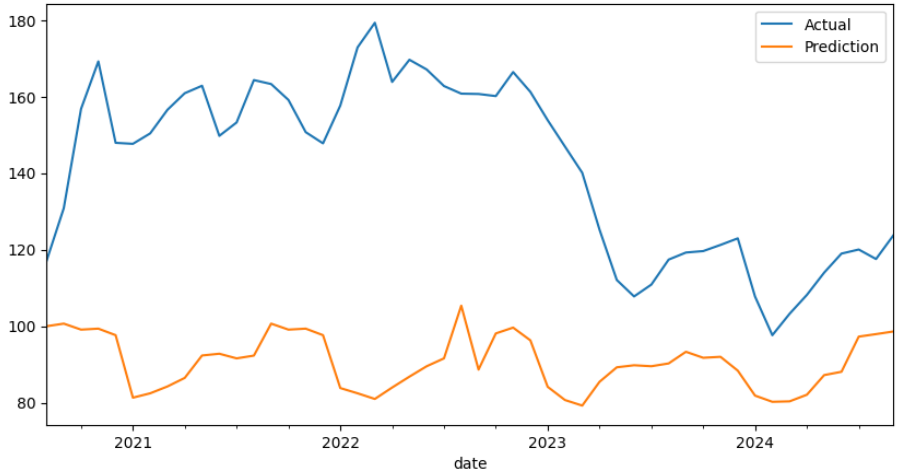

Figure 16. Previsão para Soja

Resultados do Modelo XGBoost com USD e País

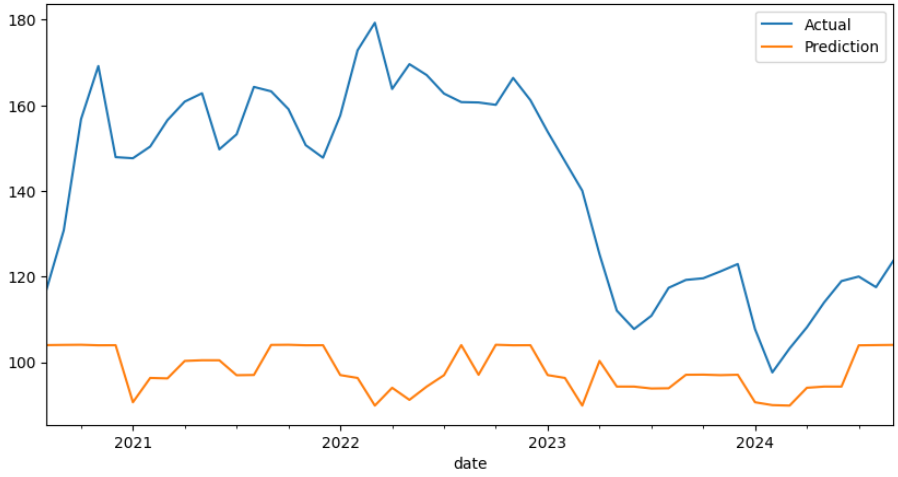

Figure 17. Previsão para Soja

4.2.3. Prophet - Meta

Utilizando do modelo *Prophet* para a previsão multivariada, utilizando das variáveis selecionadas USD, País, Máxima, Última e lag_Estadual. O modelo apresentou bons resultados e o menor tempo de execução dentre os modelos multivariados, observado na figura 18.

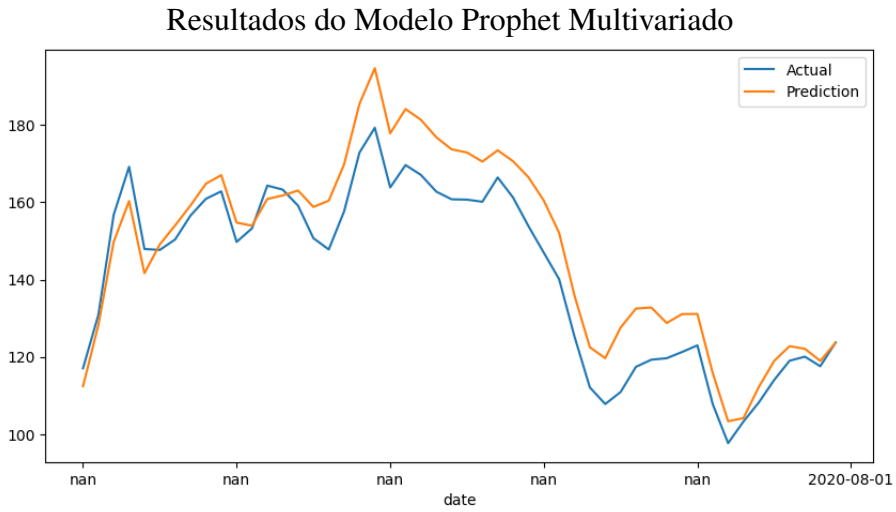

Figure 18. Previsão para Soja

4.2.4. Tabela de Resultados

Métricas Modelos para Soja			
Modelo	MAE	RMSE	R-Squared
Auto-Regressivo	45.612	49.576	-3.785
ARIMA Univariado Não-Estacionário	43.999	48.913	-3.658
ARIMA Univariado Estacionário	0.283	0.377	-0.129
XGBoost Univariado	44.101	51.334	-4.130
Prophet Univariado	57.664	63.165	-6.768
ARIMA Multivariado	4.424	5.470	0.941
XGBoost Multivariado	43.545	48.734	-3.624
Prophet Multivariado	7.674	8.762	0.850

Table 1. Tabela com resultados dos modelos testados para Soja

5. Conclusões finais

Os modelos multivariados apresentaram os melhores resultados em comparação aos univariados, tendo o câmbio como a variável de forte influência positiva nos resultados sendo utilizada em todos os modelos. Apesar das métricas baixas no modelo ARIMA Univariado Estacionário ele não se apresentou acertivo pois apesar dos valores baixos ele não era capaz de acertar nas tendências de crescimento ou queda e os resultados foram decorrentes da função Log aplicada. O modelo XGBoost apresentou métricas altas mas ele

apresentou acertos nas tendências, assim permitindo a aplicação de funções *escaler* para aproximar os valores da tendência real dos dados. O modelo Auto ARIMA Multivariado apresentou os melhores resultados ao ser acertivo nas tendências e estar próximo as curvas com valores das métricas baixos. Por fim, o modelo Prophet multivariado apresentou os melhores resultados levando em consideração métricas, acertividade e o tempo de execução, diferente da versão univariada do modelo que apresentou baixa performance.

Para os próximos passos é importante a análise de ocorrência de *overfitting* nos modelos que apresentaram métricas baixas, além de realizar a expansão das previsões para os 24 meses futuros em todos os cultivos e a partir delas realizar a apresentação dos dados para possíveis usuários das informações, como profissionais nas áreas de mercado de commodities e agricultores.

6. Cronograma

Cronograma do trabalho iniciando a partir do mês de abril de 2024. Foram desenvolvidas as seguintes atividades a partir de junho de 2024 (itens 6 a 13):

Cronograma para o Desenvolvimento do TCC

ATIVIDADE	**MÊS**											
	1	2	3	4	5	6	7	8	9	10	11	12
1. Levantamento Bibliográfico Complementar	X											
2. Elaboração do Projeto		X										
3. Apresentação do Pitch		X										
4. Estudo de área & mercado			X									
5. Estudo de acontecimentos & atualidades			X									
6. Estudar bases de dados			X									
7. Análisar possíveis aplicações				X								
8. Criar base estruturada para desenvolvimento					X							
9. Iniciar o desenvolvimento dos modelos						X	X					
10. Executar modelos								X	X			
11. Análise de resultados obtidos										X		
12. Documentação											X	
13. Apresentação do projeto												X

Referências

ANTÔNIO DA LUZ, Adelar Fochezatto. O transbordamento do PIB do Agronegócio do Brasil: uma análise da importância setorial via Matrizes de Insumo-Produto. **Revista De Economia E Sociologia Rural, 61(1), e253226.**, 2023.

BIANCA REICHERT, Adriano Mendonça Souza. Previsão e interação dos preços da celulose brasileira nos mercados interno e externo. **Ciênc. Florest. 30 (2)**, 2020.

BOTCHKAREV, Alexei. Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. **Interdisciplinary Journal of Information, Knowledge, and Management, 2019, 14, 45-79**, 2019.

CENTRO DE ESTUDOS AVANÇADOS EM ECONOMIA APLICADA (CEPEA), Confederação da Agricultura e Pecurária do Brasil (CNA). PIB DO AGRONEGÓCIO BRASILEIRO. **CEPEA - ESALQ-USP**, 2024.

CHAI, Tianfeng. Root mean square error (RMSE) or mean absolute error (MAE)? **Geoscientific Model Development**, 2014.

CHEN, Tianqi; GUESTRIN, Carlos. **XGBoost: A Scalable Tree Boosting System**. [S.l.]: ACM, 2016. P. 785–794.

CRUZ-RAMÍREZ, Angel Saul et al. Price trends of Agave Mezcalero in Mexico using multiple linear regression models. **AGROBUSINESS**, 2023.

EMBRAPA. **MANUAL de segurança e qualidade para a cultura da soja.** [S.l.]: Embrapa Informação Tecnológica, 2006.

JAMES DURBIN, Siem Jan Koopman. Time Series Analysis by State Space Methods. **Oxford University Press**, 2012.

MURTAZA DALAL ALEXANDER C. LI, Rohan Taori. **Autoregressive Models: What Are They Good For?** [S.l.], 2019.

SOUZA, Kleber Xavier Sampaio de et al. A prospective study on the application of Data Science in agriculture. **Embrapa Agricultura Digital**, 2017.

TRIEBE, Oskar et al. **NeuralProphet: Explainable Forecasting at Scale**. [S.l.], 2021.