Previsão do Preço de Commodities Baseado em Variações Históricas Utilizando Machine Learning

CAIO CANALLI[1], DAVI B. ROSA[2], RODRIGO C. GARCIA[3], ARTHUR E. O. CAROSIA[4]

[1] Graduando em Bacharelado Ciência da Computação, Bolsista FACTO, IFSP, Campus São João da Boa Vista, caio.canalli@aluno.ifsp.edu.br.
[2] Graduando em Bacharelado Ciência da Computação, Bolsista FUNARBE, IFSP, Campus São João da Boa Vista, d.beli@aluno.ifsp.edu.br.
[3] Graduando em Bacharelado Ciência da Computação, Bolsista FACTO, IFSP, Campus São João da Boa Vista, rodrigo.garcia1@aluno.ifsp.edu.br.
[4] Doutor em Tecnologia pela Universidade de Campinas (UNICAMP), Professor do IFSP, Campus São João da Boa Vista, arthuremanuel.carosia@ifsp.edu.br.

Área de conhecimento (Tabela CNPq): 1.03.00.00-7 — Ciência da Computação.

RESUMO: Este trabalho apresenta o desenvolvimento de um ecossistema de dados inteligente para a previsão e análise de preços de commodities agrícolas, especificamente Soja e Milho, correlacionando-os com vetores macroeconômicos e energéticos como a taxa de câmbio (Dólar PTAX), o Petróleo Brent e o preço do Diesel. Foi construído um pipeline de dados "end-to-end", contemplando a ingestão histórica via ETL em Python, a automação diária de coletas via n8n, a engenharia de atributos com defasagens temporais (lags) e o treinamento comparativo de modelos de Machine Learning (Regressão Linear Múltipla e XGBoost), culminando em um dashboard interativo no Power BI com simulador de cenários de estresse cambial (What-If Analysis). Os resultados indicam que o modelo XGBoost apresentou desempenho superior à Regressão Linear na captura dos ciclos de preço, evidenciando a existência de um efeito de repasse defasado ("Lag Causador") entre os preços de energia e os preços agrícolas.

PALAVRAS-CHAVE: Commodities; Previsão de Preços; Machine Learning; Séries Temporais; Business Intelligence.

Commodity Price Prediction Based on Historical Variations Using Machine Learning

ABSTRACT: This work presents the development of an intelligent data ecosystem for forecasting and analyzing agricultural commodity prices, specifically Soybeans and Corn, correlating them with macroeconomic and energy vectors such as the exchange rate (PTAX Dollar), Brent crude oil, and Diesel prices. An end-to-end data pipeline was built, comprising historical ingestion via ETL in Python, daily automated data collection via n8n, feature engineering with temporal lags, and comparative training of Machine Learning models (Multiple Linear Regression and XGBoost), culminating in an interactive Power BI dashboard with a What-If exchange-rate stress simulator. The results indicate that the XGBoost model outperformed Linear Regression in capturing price cycles, evidencing a delayed pass-through effect ("Causal Lag") between energy prices and agricultural prices.

KEYWORDS: Commodities; Price Forecasting; Machine Learning; Time Series; Business Intelligence.

INTRODUÇÃO

A volatilidade nos preços de commodities agrícolas, notadamente soja e milho, impacta diretamente a estabilidade macroeconômica de países exportadores como o Brasil, exigindo previsões precisas para a mitigação de riscos (CARVALHO, 2020). Tais cotações são fortemente influenciadas por vetores macroeconômicos e custos energéticos, como taxa de câmbio, petróleo Brent e diesel (CEPEA, 2026). Contudo, prever essas flutuações é um desafio complexo devido ao repasse temporal assíncrono desses insumos sobre a logística de escoamento (BOX et al., 2015).

Para superar as limitações dos métodos estatísticos tradicionais nesse cenário de incerteza, a integração de modelos de *Machine Learning* torna-se indispensável (JAMES et al., 2013). O trabalho justifica-se pela necessidade de desenvolver arquiteturas preditivas robustas que incorporem essas defasagens temporais. Parte-se da hipótese de que o choque de custos de energia e câmbio afeta os preços das commodities com um atraso logístico (*lag causal*) que, quando incorporado via Engenharia de Features em algoritmos não lineares (como XGBoost), supera a acurácia das regressões lineares tradicionais e evita o vazamento de dados (CHEN; GUESTRIN, 2016).

MATERIAL E MÉTODOS

O objetivo deste trabalho é projetar e implementar a arquitetura completa dessa solução em quatro fases fundamentais: a extração e tratamento histórico de mais de 30 anos de dados via Python (ETL) no banco PostgreSQL/Supabase; a automação diária da ingestão de novas cotações via nó orquestrador no n8n consumindo APIs REST; a modelagem preditiva avançada de séries temporais; e a consolidação visual executiva no Microsoft Power BI com simulador de estresse cambial (What-If).

Coleta e Estruturação de Dados Históricos (Pipeline ETL)

A fase basilar do projeto exigiu a estruturação robusta das bases de dados. Utilizando Python e a poderosa biblioteca Pandas dentro de instâncias do Google Colab, construiu-se uma Tabela Fato unificada contendo as cotações do Diesel, Petróleo Brent, Soja e Milho. A injeção na nuvem foi alocada na plataforma Supabase (baseada em PostgreSQL). Ocorreram desafios inerentes de infraestrutura (o Colab opera sob restrições para IPs IPv6), os quais foram solucionados na arquitetura através do mapeamento de rotas via Session Pooler na porta dedicada 6543/5432, populando com sucesso absoluto a tabela matriz tb_historico.

Ingestão de Dados e Automação Contínua com n8n

Garantir que a inteligência artificial não ficasse obsoleta no momento seguinte ao seu treinamento é mandatório. Para tal, implementou-se um "robô" orquestrador utilizando n8n. Este fluxo contínuo tem a seguinte rotina funcional:

Ativação Temporal: Gatilho automático rodando diariamente via nó de Schedule Trigger.

Chamadas Externas (HTTP Request): Requisições para o provedor AlphaVantage para coleta diária das moedas (Dólar/BRL) e do índice energético do Brent. Integrando técnicas de resiliência (rate limit bypass respectivo), incorporou-se um nó lógico de Wait de 3 segundos para que as chamadas HTTP respeitassem o limite da camada gratuita do fornecedor.

Processamento Paralelo (Data Transformation): Intervenções por script no nó de Code (JavaScript) convertendo, por exemplo, unidades volumétricas do mercado de barril para litros para manter paridade transacional de dados.

Escrita no Relacional: O payload JSON normalizado consolida a inserção dos números do dia (Insert row operation) dentro do schema público, na tabela tb_cotacoes_diarias.

Engenharia de Atributos, Séries Temporais e Preditividade (IA)

Nesta fase, a comprovação empírica da tese de modelagem foi atestada via código. A complexidade do treinamento focou na mecânica específica exigida para séries temporais e combate ao "Data Leakage" (vazamento de dados).

Engenharia de Defasagem (Lags Temporais): Identificou-se graficamente e via Pearson que o efeito da alta nos combustíveis não é transmitido ao agronegócio de modo síncrono. Para isso, criaram-se atributos matemáticos defasando os registros (shift method no Pandas) do Diesel e do Petróleo em períodos de dois e três meses (diesel_lag_2m, brent_lag_2m). Esses lags representaram o verdadeiro gatilho causal do repasse final do Custo Brasil.

Divisão Metodológica de Treinamento: O dataframe foi submetido a uma cisão de 80% (base de conhecimento - passado) e 20% (base cega de validação - futuro), garantindo a imposição do parâmetro estrito shuffle=False, pois em time series, as leis temporais não podem ser embaralhadas.

Modelos de Regressão Empregados: Executou-se Regressão Linear Múltipla e Ensemble Learning baseado em Árvores (XGBoost Regressor).

Deploy do Motor Preditivo: Para gerar visões pragmáticas, bifurcou-se as inferências da Inteligência Artificial em dois agentes: um exclusivamente analisando o Milho (modelo_milho) e outro para a Soja (modelo_soja). A rede gerou predições do horizonte futuro sem alucinação informacional (utilizando lags), criando automaticamente e preenchendo as métricas finais em USD dentro da tabela nativa tb_previsoes.

Visualização Analítica e Business Intelligence (Power BI)

A última fase foi centrada na translação visual das massas de dados. O Power BI Desktop atuou como gateway executivo consumindo nativamente o servidor do Supabase.

Componente do Painel (Dashboard)	**Descrição da Implementação e Foco Operacional**
Conectividade Direct-Like	Ingestão em modelo de Importação superando bloqueios de firewall SSL através do ajuste nas permissões do Power BI com os esquemas da base de dados PostgreSQL.
Motor "What-If" e Sliders de Simulação	Adição da funcionalidade Parameter de "Range" numérico para atuar como o "Simulador Dólar", permitindo a avaliadores imputarem estresse cambial de R$ 4,50 a R$ 7,00, de forma interativa.
Expressões DAX Avançadas	Funções de cruzamento matemático de variáveis locais. O algoritmo capta o valor projetado pela rede neural em Dólar – estabilizado e refinado em saca utilizando a função estatística agregadora AVERAGE para evitar saltos com as somatórias lineares (SUM) de horizontes de vários meses – transformando o resultado final em reais brasileiros ao se multiplicar pelo input cambial da tela, exposto em visual de Cartões (Callout UI).

Tabela 1 – Componentes do painel de Business Intelligence. Fonte: Elaborada pelos autores.

RESULTADOS E DISCUSSÃO

Na comparação entre os modelos de Machine Learning empregados, o modelo de base (Regressão Linear Múltipla) não sustentou sua predição diante da magnitude atípica de ruídos macroeconômicos (pandemia, conflitos internacionais), errando as amplitudes de variação de preço. Em contrapartida, o modelo avançado (XGBoost Regressor) aprendeu os ciclos de preço de forma mais concisa, beneficiando-se da engenharia de atributos com defasagem temporal (lags) do Diesel e do Petróleo Brent.

A análise de correlação (Pearson) confirmou que o efeito da alta nos combustíveis não é transmitido ao agronegócio de modo síncrono, mas sim com uma defasagem de dois a três meses, o que motivou a criação dos atributos diesel_lag_2m e brent_lag_2m como os verdadeiros gatilhos causais do repasse do Custo Brasil para o preço final da Soja e do Milho.

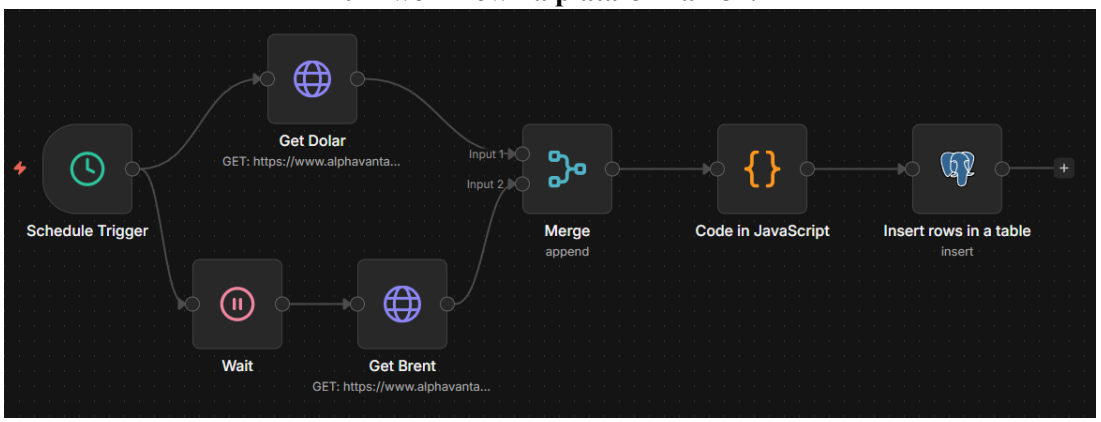
FIGURA 1. Valores simulados pelo modelo avançado XGBoost Regressor com visualização no Microsoft Power BI.



(Fonte: Elaborado pelos autores)

Do ponto de vista de infraestrutura, o pipeline de automação via n8n operou de forma estável na captura diária das cotações, respeitando os limites de requisição da camada gratuita do provedor de dados (AlphaVantage). O dashboard final em Power BI foi entregue com sucesso, incorporando um simulador "What-If" que permite ao usuário aplicar estresse cambial entre R$ 4,50 e R$ 7,00 e visualizar seu efeito nas projeções de preço convertidas de Dólar para Real, por meio de expressões DAX que combinam a função AVERAGE com o input cambial informado na tela.

FIGURA 2. Extrator e filtrador de variáveis dos provedores de dados em tempo real utilizando um workflow na plataforma n8n.



(Fonte: Elaborado pelos autores)

CONCLUSÕES

Este ecossistema prova como a adoção em escala de integrações de automação em conjunto com Inteligência Artificial Preditiva consegue contornar a volatilidade intrínseca ao segmento primário da economia (commodities). Validou-se cientificamente que, no modelo de distribuição logístico brasileiro, atrasos de meses formam um "Lag Causador" (onde oscilações precificadas em mercados abertos na energia afetam subsequentemente as cadeias agro).

Através de práticas contemporâneas de desenvolvimento Full-Stack para engenharia de dados, estabeleceu-se um software em produção perene, garantindo transparência desde os anos 90 do registro histórico até a estimativa autônoma para os meses vindouros.

AGRADECIMENTOS

Ao Prof. Dr. Arthur Emanuel de Oliveira Carosia, pela orientação dedicada, suporte técnico e incentivo constante ao longo da execução deste projeto. Ao Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) — Câmpus São João da Boa Vista, pela infraestrutura disponibilizada e pelo suporte acadêmico essencial para a realização desta pesquisa, e ao Congresso de Inovação, Ciência e Tecnologia do IFSP (CONICT) pela oportunidade.

REFERÊNCIAS

BANCO CENTRAL DO BRASIL. SGS - Sistema Gerenciador de Séries Temporais. Brasília, 2026. Disponível em: https://www3.bcb.gov.br/sgspub/. Acesso em: 7 ago. 2026.

BOX, G. E. P. et al. Time series analysis: forecasting and control. 5. ed. Hoboken: John Wiley & Sons, 2015.

CARVALHO, J. C. Transmissões de volatilidade de preços entre Commodities agrícolas brasileiras. Revista de Economia e Sociologia Rural, Brasília, v. 58, n. 3, e193763, 2020. Disponível em: https://www.scielo.br/j/resr/a/zGw8xYzbQN7QbF8MY9ftNvp/?lang=pt. Acesso em: 28 ago. 2026.

CEPEA - CENTRO DE ESTUDOS AVANÇADOS EM ECONOMIA APLICADA. Indicadores Agropecuários - Soja e Milho: análises de mercado e impacto cambial. Piracicaba: ESALQ/USP, 2026. Disponível em: https://www.cepea.esalq.usp.br/br. Acesso em: 28 ago. 2026.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, San Francisco. Anais [...]. New York: ACM, 2016. p. 785-794.

CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO. Séries históricas de preços agropecuários. Brasília: Conab, 2026. Disponível em: https://portaldeinformacoes.conab.gov.br/. Acesso em: 7 ago. 2026.

EIA - U.S. ENERGY INFORMATION ADMINISTRATION. Crude oil and petroleum products data. Washington, D.C.: EIA, 2026. Disponível em: https://www.eia.gov/. Acesso em: 7 ago. 2026.

FRED. Federal Reserve Economic Data: Agricultural commodities and macroeconomic indicators. St. Louis: Federal Reserve Bank of St. Louis, 2026. Disponível em: https://fred.stlouisfed.org/. Acesso em: 7 ago. 2026.

JAMES, G. et al. An introduction to statistical learning: with applications in R. New York: Springer, 2013.