Previsão de Produtividade Agrícola Utilizando Machine Learning Aplicado a Dados Climáticos e de Solo

1st Bruno da Silva Godoy
Engenharia de Computação, FHO
FHO – Fundação Hermínio Ometto
Araras, Brasil
brunogodoy@alunos.fho.edu.br

2nd Caetano Orloski Moriconi
Engenharia de Computação, FHO
FHO – Fundação Hermínio Ometto
Araras, Brasil
caetanoorloski@alunos.fho.edu.br

***Abstract*—Agricultura é um dos setores mais impactados pelas variaçes climáticas, tornando a previso de produtividade um desafio essencial para o planejamento estratégico e a sustentabilidade do agronegócio. Este trabalho tem como objetivo desenvolver um modelo de aprendizado de máquina capaz de prever a produtividade agrícola a partir de variáveis climáticas e de solo. Tem com o intuito identificar o modelo de melhor desempenho. Pretende-se avaliar métricas como erro médio e coeficiente de determinação, além de discutir a relevância da aplicação de técnicas de inteligência artificial no apoio à agricultura de precisão.**

***Index Terms*—Machine Learning, Agricultura de Precisão, Previsão de Safras, Inteligência Artificial, Ciência de Dados**

I. Introdução

A agricultura desempenha um papel estratégico na economia global, sendo responsável pelo abastecimento alimentar e pela geração de riqueza em diversos países. Entretanto, a produtividade agrícola sofre grande influência de fatores climáticos, condições do solo e práticas de manejo, o que torna sua previsão uma tarefa complexa e de elevada importância.

Com o avanço das técnicas de *Machine Learning* e o crescente acesso a grandes volumes de dados, tornou-se possível explorar algoritmos de inteligência artificial para apoiar a tomada de decisão no setor agrícola. Modelos de aprendizado supervisionado permitem identificar padrões históricos entre variáveis climáticas e produtivas, possibilitando previsões mais robustas e confiáveis.

Este trabalho propõe o desenvolvimento de um modelo de aprendizado de máquina voltado para a previsão da produtividade agrícola a partir de dados climáticos e de solo. O projeto será estruturado em etapas que incluem a coleta e preparação de dados, implementação, avaliação de métricas de desempenho e comparação dos resultados obtidos. O objetivo é analisar a aplicabilidade das técnicas de IA no contexto da agricultura de precisão, contribuindo para a redução de riscos e o uso mais eficiente de recursos.

II. Trabalhos Relacionados.

A aplicação de Machine Learning para prever a produtividade de culturas agrícolas é um campo de pesquisa consolidado, com diversas abordagens documentadas na literatura científica. Os trabalhos podem ser categorizados com base nos algoritmos empregados e nas fontes de dados utilizadas.

Modelos de ensemble, como o Random Forest, são frequentemente destacados por sua precisão e robustez. Em um estudo de grande impacto, Jeong et al. (2016) utilizaram dados de sensoriamento remoto, clima e solo para prever a produtividade do milho nos Estados Unidos. Eles concluíram que o Random Forest superou outras técnicas, como Support Vector Machines (SVM), e que variáveis climáticas durante o período reprodutivo da cultura foram determinantes para o sucesso da previsão.

A integração de múltiplas fontes de dados é um fator crítico para a acurácia dos modelos. Paudel et al. (2021) investigaram o impacto de variáveis de clima, solo, topografia e imagens de satélite na previsão da safra de milho. O trabalho reforça que a combinação de dados de solo com variáveis climáticas melhora significativamente o desempenho preditivo, permitindo uma análise mais granular e útil para a agricultura de precisão.

No contexto brasileiro, o estudo de Johann, Oliveira e Siqueira (2016) aplicou Redes Neurais Artificiais para estimar a produtividade do milho safrinha no estado do Paraná, utilizando principalmente dados de satélite. Este estudo é relevante por validar a aplicação de técnicas de IA em larga escala no Brasil, demonstrando o potencial do uso de diferentes fontes de dados para monitorar a agricultura nacional.

Na fronteira da pesquisa, abordagens de Deep Learning têm sido exploradas. Um framework que combina Redes Neurais Convolucionais (CNNs) com Redes Neurais Recorrentes (LSTMs) foi proposto para interpretar dados espaciais e séries temporais de clima, obtendo um desempenho superior ao de modelos de Machine Learning tradicionais e mostrando o potencial de técnicas mais complexas (SUN et al., 2019).

Em suma, a literatura confirma que modelos de Machine Learning são ferramentas poderosas para a previsão de safras (JEONG et al., 2016; PAUDEL et al., 2021). Fica evidente a superioridade de algoritmos de ensemble (JEONG et al., 2016) e a necessidade de integrar dados de solo e clima (PAUDEL et al., 2021). O presente trabalho se posiciona ao realizar

uma comparação sistemática entre diferentes algoritmos, aplicados a um conjunto de dados específico para as condições edafoclimáticas do Brasil, contribuindo com a validação dessas técnicas em um contexto nacional, similarmente ao proposto por Johann, Oliveira e Siqueira (2016), porém com foco na integração de dados de solo.

III. Metodologia

O presente trabalho teve como objetivo integrar diferentes bases de dados públicas para a previsão da produtividade agrícola no estado de São Paulo, considerando variáveis climáticas e de produção.

Inicialmente, foram coletados dados a partir de duas fontes governamentais: o Instituto Nacional de Meteorologia (INMET) e o Instituto Brasileiro de Geografia e Estatística (IBGE), por meio do Levantamento Sistemático da Produção Agrícola (LSPA). O INMET forneceu dados climáticos históricos contendo informações mensais de precipitação total (mm), temperatura máxima, mínima e média (°C) e umidade relativa do ar (%), abrangendo o período de 2018 a 2024. Esses arquivos estavam organizados por estação meteorológica e apresentavam variações de formato e estrutura, o que exigiu um processo de padronização e tratamento para garantir a consistência e integridade das informações.

Paralelamente, os dados de produtividade agrícola foram obtidos do LSPA, referentes ao estado de São Paulo, contendo as colunas de Ano, Mês e Produção (t). Esses arquivos foram tratados manualmente, uma vez que as informações estavam distribuídas de forma fragmentada nas planilhas originais. Após o tratamento individual de cada base, os conjuntos de dados foram integrados, relacionando as condições climáticas com os valores mensais de produtividade agrí cola, formando uma base consolidada que compreende o período de 2018 a 2024.

O tratamento e a integração dos dados foram realizados utilizando a linguagem Python, com o auxílio das bibliotecas Pandas, NumPy, Matplotlib e Scikit-learn. Os dados foram limpos, normalizados e preparados para o processo de modelagem. Para a etapa de previsão da produtividade agrícola, aplicou-se uma regressão não linear, metodologia adequada para capturar relações complexas e não proporcionais entre as variáveis climáticas e a produtividade. Essa abordagem permitiu representar de forma mais realista as interações entre os fatores ambientais e a produção agrícola, resultando em previsões mais consistentes e alinhadas ao comportamento real observado nos dados históricos.

IV. Resultados

A análise dos resultados obtidos neste estudo oferece uma visão transparente sobre os desafios de se modelar a produtividade agrícola utilizando exclusivamente variáveis climáticas em um período histórico curto. Para validar o desempenho do modelo Random Forest, submetemos o algoritmo a um conjunto de dados de teste desconhecido durante o treinamento, correspondente a 20(%) da base original.

A. Análise das Métricas de Erro

Ao examinarmos os indicadores quantitativos apresentados na Tabela I, notamos que o modelo encontrou dificuldades significativas em generalizar os padrões de produção. O Coeficiente de Determinação (R^2) resultou em -0.7168. Em termos estatísticos, um valor negativo nesta métrica sinaliza que o modelo não conseguiu superar a eficácia de uma simples linha de média histórica. Isso não invalida o experimento, pelo contrário, lança luz sobre a complexidade do problema: a tentativa de explicar a produção agrícola baseando-se apenas em temperatura, precipitação e umidade mostrou-se insuficiente.

B. TABELA IRESUMO DAS MÉTRICAS DE DESEMPENHO

TABLE I
Resumo das Métricas de Desempenho

Métrica	Valor Obtido
R^2	-0.7168
MAE (Toneladas)	400.663,63
RMSE (Toneladas)	872.884,74

Os valores de erro absoluto (MAE) e quadrático (RMSE), na casa das centenas de milhares de toneladas, refletem a grande escala da produção total na região estudada. A discrepância entre o MAE e o RMSE sugere ainda que houve erros pontuais de grande magnitude, indicando que o modelo foi penalizado por tentar prever safras atípicas sem ter acesso a dados explicativos fundamentais, como o manejo do solo, o controle de pragas ou a tecnologia aplicada naquela safra específica.

A figura abaixo ilustra perfeitamente este cenário para o ano de 2020, onde o modelo não conseguiu capturar o pico atípico de produção no meio do ano.

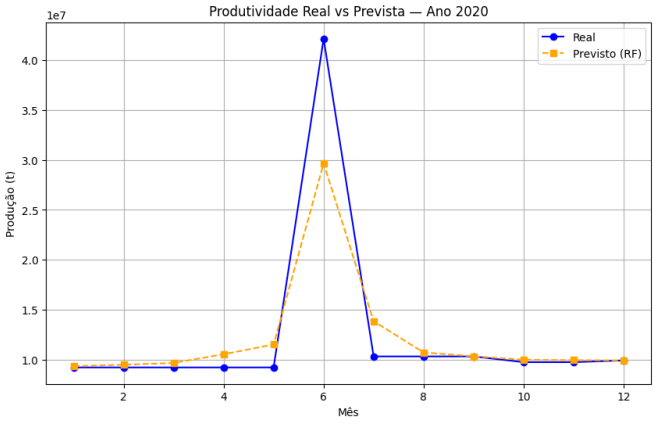

Fig. 1. Comparação entre Produtividade Real e Prevista (Ano 2020).

É importante ressaltar que, para os anos já presentes no conjunto de dados histórico (como 2020), o sistema opera em modo de validação. Neste cenário, não é necessário projetar as variáveis climáticas via ARIMA, uma vez que os dados reais de temperatura e precipitação já são conhecidos. Portanto, o

gráfico acima apresenta um confronto direto: a linha azul representa a produção real colhida, enquanto a linha laranja exibe a estimativa gerada pelo *Random Forest* ao ser alimentado com as condições climáticas daquele ano. Essa comparação visual serve como a principal prova de conceito da capacidade do modelo em aprender (ou não) os padrões do passado.

C. O Comportamento nas Previsões Futuras (2025-2026)

Um dos pontos mais relevantes desta discussão surge ao analisarmos a aplicação da abordagem híbrida (ARIMA + Random Forest) para os anos futuros de 2025 e 2026. Ao gerar essas projeções, observou-se um comportamento conservador por parte do modelo: os valores previstos mantiveram-se muito próximos à média dos anos anteriores, apresentando pouquíssima oscilação.

A figura a seguir demonstra essa estabilidade projetada para os próximos anos.

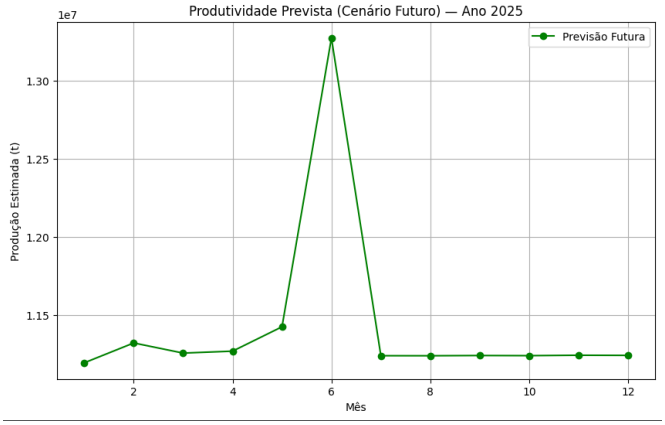

Fig. 2. Comparação entre Produtividade Real (Ano 2025).

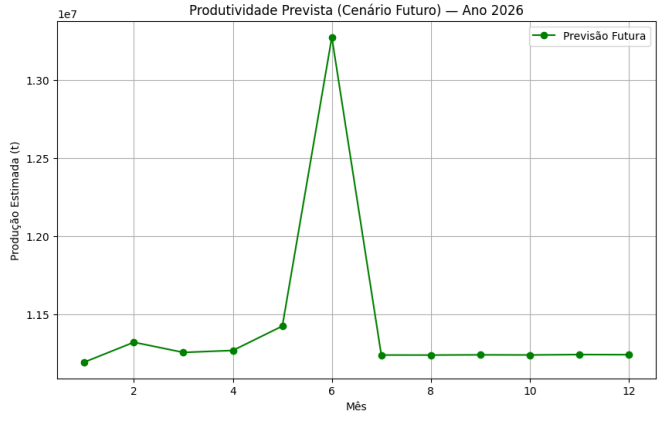

Fig. 3. Comparação entre Produtividade Real (Ano 2026).

Essa estabilidade nas previsões futuras não é um acaso, mas sim um reflexo direto das características do dataset utilizado. Como a série histórica coletada (2018 a 2024) apresentou uma variação de produtividade relativamente baixa, o algoritmo aprendeu que o padrão ”normal” é a estabilidade. O Random Forest, por sua natureza matemática, possui limitações em extrapolar tendências; ou seja, ele tem dificuldade em prever valores muito acima ou muito abaixo do que já viu no treinamento.

Consequentemente, para os anos futuros, o modelo replicou o padrão de estabilidade que observou no passado recente, demonstrando que, para capturar grandes quebras de safra ou recordes produtivos no futuro, seria necessário treinar a IA com um histórico mais longo e volátil. Portanto, os resultados evidenciam que a acurácia da Inteligência Artificial na agricultura é totalmente dependente da riqueza dos dados. O clima é apenas uma peça de um quebra-cabeça muito maior.

V. Conclusão

O presente trabalho alcançou seu objetivo principal ao desenvolver e implementar um pipeline completo de Inteligência Artificial para a previsão de produtividade agrícola, integrando técnicas de previsão de séries temporais (ARIMA) com algoritmos de aprendizado de máquina (Random Forest). A proposta de criar uma abordagem híbrida mostrou-se viável tecnicamente, permitindo não apenas a análise de dados passados, mas também a projeção de cenários climáticos e produtivos para safras futuras.

A. Recapitulação dos Achados

Os experimentos realizados evidenciaram que a modelagem de sistemas agrícolas é uma tarefa de elevada complexidade. Os resultados quantitativos (com R^2 negativo) demonstraram que as variáveis climáticas, embora essenciais, não são suficientes para explicar isoladamente a totalidade da variação produtiva. Confirmou-se a hipótese de que a produtividade no campo é um fenômeno multifatorial, fortemente influenciado também por manejo, características do solo e tecnologia — dados estes que não estavam disponíveis nas fontes públicas utilizadas neste estudo.

B. Limitações e Desafios

Uma constatação crucial, e também a principal limitação do projeto, refere-se à janela temporal restrita do conjunto de dados (2018 a 2024). Como discutido nos resultados, a baixa variabilidade produtiva neste curto período fez com que o modelo aprendesse um padrão de ”estabilidade excessiva”. Isso resultou em previsões para 2025 e 2026 que apenas replicaram a média histórica, mostrando pouca sensibilidade a possíveis eventos extremos. O modelo ficou limitado pela falta de exemplos históricos de alta volatilidade durante seu treinamento.

C. Trabalhos Futuros

Diante desse cenário, conclui-se que o passo mais urgente para a evolução deste projeto é a expansão robusta do dataset. A captação de dados de um período histórico mais longo — idealmente abrangendo as últimas duas ou três décadas — é indispensável para corrigir o viés de estabilidade identificado. Além disso, recomenda-se fortemente a inclusão de variáveis físico-químicas do solo e índices de vegetação via satélite

(NDVI), bem como o teste de arquiteturas de Deep Learning (como redes LSTMs), que podem oferecer uma capacidade superior de processar sequências temporais complexas.

Em suma, este trabalho serviu como um protótipo valioso, delimitando as fronteiras algorítmicas e os requisitos de dados necessários para a aplicação eficaz da Inteligência Artificial na agricultura de precisão.

REFERÊNCIAS

[1] J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, and E. J. Butler, "Random forests for global and regional crop yield predictions," PloS one, vol. 11, no. 6, p. e0156571, 2016.

[2] D. Paudel, H. Boogaard, A. de Wit, M. van der Velde, and T. A. D. Clingant, "Machine learning for large-scale crop yield forecasting," Agricultural and Forest Meteorology, vol. 298, p. 108287, 2021.

[3] J. A. Johann, A. R. de Oliveira, and M. L. G. de Siqueira, "Estimativa da produtividade do milho safrinha por meio de redes neurais artificiais e dados orbitais," Engenharia Agrícola, vol. 36, no. 6, pp. 1105-1115, 2016.

[4] J. Sun, Y. Wang, Y. Shu, and Z. Wang, "A deep learning framework for corn yield prediction using remote sensing and crop phenology data," ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-2/W13, pp. 1183-1188, 2019.

[5] INSTITUTO NACIONAL DE METEOROLOGIA (INMET). Banco de Dados Meteorológicos para Ensino e Pesquisa – BDMEP. Brasília, [s.d.]. Disponível em: https://bdmep.inmet.gov.br/ . Acesso em: 08 nov. 2025.

[6] INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). Levantamento Sistemático da Produção Agrícola (LSPA). Rio de Janeiro, 2025. Disponível em: https://sidra.ibge.gov.br/pesquisa/lspa . Acesso em: 08 nov. 2025.

[7] DA SILVA, Alexandre Marco; ALVARES, Clayton Alcarde. Levantamento de informações e estruturação de um banco dados sobre a erodibilidade de classes de solos no estado de São Paulo. Geosciences= Geociências, v. 24, n. 1, p. 33-41, 2005..

[8] https://github.com/brunogodoy23/IA-II.git