josif 2025
17ª Jornada Científica e Tecnológica
14º Simpósio de Pós-Graduação
do IFSULDEMINAS
ISSN: 2319-0124
REALIZAÇÃO
INSTITUTO FEDERAL Sul de Minas Gerais
APOIO
CNPq
FAPEMIG

MACHINE LEARNING NA PREVISÃO DE RENDIMENTO DA CULTURA DE CAFÉ NO SUL DE MINAS GERAIS: um estudo comparativo

Caio E. T. FERREIRA[1]; Paulo C. dos SANTOS[2]; Diego SAQUI[3]; Lucas E. de O. APARECIDO[4]

RESUMO

Este estudo investigou a aplicação de algoritmos de aprendizado de máquina na previsão do rendimento da cultura cafeeira no Sul de Minas Gerais, região de destaque na produção nacional. Foram utilizados dados históricos de produtividade e variáveis meteorológicas para treinar modelos supervisionados de regressão. A avaliação do desempenho dos modelos, com base nas métricas R^2, MAE e RMSE, indicou baixo poder preditivo em todos os casos, com o modelo de Deep Learning apresentando o melhor desempenho relativo (R^2=0,295, MAE=353,10 sc ha-1 e RMSE=513,36 sc ha-1). Os resultados sugerem que as variáveis climáticas utilizadas, embora relevantes, não são suficientes para explicar a complexa variabilidade da produção, reforçando a necessidade de incorporar atributos agronômicos, edáficos e fitossanitários em estudos futuros para aprimorar a acurácia das previsões.

Palavras-chave: Inteligência Artificial; Modelos Preditivos; Agricultura Digital.

1. INTRODUÇÃO

A cafeicultura brasileira ocupa posição estratégica no cenário econômico nacional, sendo responsável pela expressiva geração de divisas e empregos. Dados da CONAB (2024) mostram que o país produziu aproximadamente 59 milhões de sacas de café em 2024, com Minas Gerais respondendo por mais de 50% dessa produção, o que justifica a busca por métodos avançados de previsão de produtividade.

O advento do *Machine Learning* (ML) tem revolucionado a agricultura de precisão, oferecendo ferramentas poderosas para análise preditiva. Algoritmos como *Random Forest* e Redes Neurais Artificiais vêm demonstrando superioridade na modelagem de sistemas agrícolas complexos (KHAKI; WANG, 2019). Particularmente, as técnicas de *Deep Learning*, uma subárea das Redes Neurais que utiliza arquiteturas com múltiplas camadas ocultas para aprender representações de dados em diferentes níveis de abstração, têm se mostrado promissoras (BENGIO, 2009).

No contexto deste trabalho, o rendimento do café é medido em sacas de 60kg (sc ha-1), uma unidade padrão que quantifica a produtividade de uma área. Uma vez que o objetivo é prever essa quantidade, que é um valor numérico contínuo, o problema é caracterizado como uma tarefa de regressão (JAMES et al., 2013). Diante disso, este trabalho propõe comparar o desempenho de cinco algoritmos de ML: Regressão Linear, Árvore de Decisão, *Random Forest* e duas arquiteturas

[1]Bolsista PIBIC/CNPq, IFSULDEMINAS – *Campus* Muzambinho. E-mail: caio.tomaz@alunos.ifsuldeminas.edu.br.
[2]Orientador, IFSULDEMINAS – *Campus* Muzambinho. E-mail: paulo.santos@muz.ifsuldeminas.edu.br.
[3]Orientador, UFLA – *Campus* Paraíso. E-mail: diego.saqui@ufla.br.
[4]Orientador, IFSULDEMINAS – *Campus* Muzambinho. E-mail: lucas.aparecido@muz.ifsuldeminas.edu.br.

distintas de Redes Neurais Artificiais (uma rede MLP com duas camadas ocultas e uma arquitetura de Deep Learning com quatro camadas). O objetivo é determinar a eficácia desses modelos na previsão do rendimento de café no Sul de Minas Gerais, utilizando um conjunto de dados históricos de produtividade e variáveis meteorológicas.

2. MATERIAL E MÉTODOS

Trata-se de uma pesquisa quantitativa aplicada, conduzida no Laboratório de Tecnologias de Software e Computação Aplicada à Educação (LabSoft) do IFSULDEMINAS - Campus Muzambinho. As etapas metodológicas consistiram em:

i) Revisão bibliográfica, com levantamento de estudos relacionados ao uso de ML na agricultura;

ii) Coleta e Caracterização dos Dados: A base de dados utilizada neste estudo foi construída a partir de duas fontes principais: dados de produtividade (sacas de 60kg) cedidos pela Cooperativa Cooxupé e dados meteorológicos de acesso público. O conjunto de dados abrange um período de 10 anos e compreende 200 municípios distintos do Sul de Minas Gerais. A agregação dos dados em nível municipal foi adotada para garantir a anonimização dos produtores.

O conjunto final, após limpeza e remoção de valores ausentes, totalizou 2.000 amostras, onde cada amostra representa a observação de um município em um determinado ano. A variável-alvo, produtividade, não apresentou valores ausentes e foi utilizada conforme fornecida. O tratamento de valores discrepantes (*outliers*) foi realizado especificamente na variável preditora de precipitação, utilizando o método do Intervalo Interquartil (IQR), de modo a mitigar o impacto de eventos pluviométricos extremos e pontuais que poderiam distorcer a representatividade climática do município. O conjunto de atributos preditores é composto por 9 variáveis: ano, altitude e 7 variáveis meteorológicas (precipitação, pressão, radiação, temperatura do ar, temperatura de orvalho, umidade e vento);

iii) Pré-processamento: Os dados foram normalizados (escalonados) para os modelos de redes neurais, a fim de garantir a convergência do algoritmo;

iv) Modelagem e Validação: Para a avaliação dos modelos, os dados foram divididos em conjuntos de treino (80%) e teste (20%). Foi empregada uma abordagem de validação hold-out simples (random_state=42 para garantir a reprodutibilidade). Embora métodos como a validação cruzada temporal sejam robustos, optou-se pelo hold-out como uma estratégia inicial de avaliação, alinhada ao caráter exploratório deste estudo e às limitações de escopo. Para contextualizar os resultados e estabelecer uma base de comparação, um modelo de linha de base (baseline) que prevê a média histórica do conjunto de treino foi incluído na análise. Foram implementados os seguintes algoritmos: Regressão Linear (modelo padrão); Árvore de Decisão e Random Forest (configurações

padrão do scikit-learn com random_state=42); uma Rede Neural MLP com duas camadas ocultas (100, 50) e max_iter=500; e uma MLP Profunda com quatro camadas ocultas (256, 128, 64, 32) e max_iter=1000, ambas com critério de parada antecipada.

v) Avaliação de desempenho dos modelos sob as métricas R^2, *Mean Absolute Error* (MAE) e *Root Mean Square Error* (RMSE);

vi) Análise comparativa, visando identificar o modelo mais eficaz com base nos dados disponíveis.

3. RESULTADOS E DISCUSSÃO

O desempenho dos algoritmos, medido por meio das métricas R^2 (Coeficiente de Determinação), MAE (Erro Absoluto Médio) e RMSE (Raiz do Erro Quadrático Médio), é apresentado na Tabela 1.

Tabela 1 - Métricas de desempenho dos modelos

Modelo	R^2 Score	MAE (sc ha-1)	RMSE (sc ha-1)
Deep Learning (4 camadas)	0,2954	353,10	513,36
Random Forest	0,2537	360,46	528,33
Regressão Linear	0,0953	413,80	581,68
Redes Neurais (2 camadas)	0,0901	423,04	583,36
Baseline (Média)	-0,0031	437,20	612,51
Árvore de Decisão	-0,1836	448,51	523.43

Fonte: do autor (2025)

A análise da Tabela 1 revela um poder preditivo limitado em todos os modelos. O modelo de Deep Learning alcançou o melhor desempenho relativo, com um R^2 de 0,2954. Isso indica que o modelo explicou apenas 29,5% da variabilidade no rendimento do café com base nos dados climáticos, o que é insuficiente para aplicações práticas de previsão de safra. Os demais modelos tiveram desempenho inferior. Destaca-se o R^2 negativo (-0,1836) da Árvore de Decisão, que performou pior que o modelo de baseline (-0,0031), o qual simplesmente prevê a média de rendimento. Este resultado sugere um sobreajuste (overfitting) severo do modelo de árvore, que não conseguiu generalizar os padrões dos dados de treino.

O desempenho consistentemente baixo entre todos os algoritmos testados levanta uma forte hipótese: as variáveis climáticas, isoladamente, não são suficientes para modelar a complexidade da produtividade agrícola, especialmente quando os dados são agregados em uma escala municipal. A

heterogeneidade de fatores como manejo agronômico (e.g., fertilização, poda), variedades genéticas e microclimas dentro de um mesmo município introduz uma variabilidade que os modelos, baseados apenas em dados meteorológicos gerais, não conseguem capturar. Fatores cruciais como as propriedades físico-químicas do solo, o manejo agronômico (e.g., fertilização, irrigação, poda), a variedade genética da planta, a idade do cafezal e a incidência de pragas e doenças não foram incluídos neste estudo, mas são reconhecidos pela literatura como determinantes para o rendimento (KHAKI; WANG, 2019; FERREIRA et al., 2022). A ausência desses atributos limita a capacidade dos modelos de aprenderem a relação completa entre as condições de cultivo e a produção final, resultando no baixo poder preditivo observado.

4. CONCLUSÃO

Embora os algoritmos de Machine Learning sejam ferramentas poderosas, seu desempenho na previsão de safras de café é limitado quando se utilizam apenas dados meteorológicos. O melhor modelo, Deep Learning, alcançou um R^2 de apenas 0,295, reforçando que a produtividade do café é um fenômeno complexo.

AGRADECIMENTOS

Agradece-se ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio à pesquisa por meio do Programa PIBIC. Estende-se o reconhecimento ao IFSULDEMINAS, Reitoria e *Campus* Muzambinho, ao LabSoft e à Cooxupé pela disponibilização dos dados utilizados neste estudo.

REFERÊNCIAS

BENGIO, Y. Learning Deep Architectures for AI. Foundations and Trends in Machine Learning, v. 2, n. 1, p. 1-127, 2009.

CONAB. Acompanhamento da safra brasileira de café: safra 2024. Brasília: Companhia Nacional de Abastecimento, 2024. 60 p.

FERREIRA, W. P. M. et al. Mapping the mountainous climate in the Matas de Minas region. Research, Society and Development, v. 11, n. 3, e25411326540, 2022.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.

KHAKI, S.; WANG, L. Crop Yield Prediction Using Deep Neural Networks. Frontiers in Plant Science, v. 10, p. 621, 2019.