ANÁLISE DA TÉCNICA DE IMPORTÂNCIA DE VARIÁVEIS POR PERMUTAÇÃO EM MODELOS DE CLASSIFICAÇÃO SUPERVISIONADA BASEADOS EM ÁRVORES

JANE PIANTONI[1], KAMILA CRISTINA DE CREDO ASSIS[2]

[1] Doutoranda em Tecnologia - Gestão, Processamento e Armazenamento da Informação, Analista de Capacitação, Flextronics Instituto de Tecnologia, Sorocaba, jane.piantoni@fit-tecnologia.org.br

[2] Doutoranda em Engenharia de Sistemas Agrícolas, Pesquisadora, Flextronics Instituto de Tecnologia, Sorocaba, kamila.assis@fit-tecnologia.org.br

Área de conhecimento (Tabela CNPq): 1.03.03.04-9 Sistemas de Informação

RESUMO: Este estudo examina a aplicação da técnica de Importância de Variáveis por Permutação (PFI) em modelos de classificação supervisionada, com foco em algoritmos baseados em árvores de decisão, como o Random Forest. A PFI tem sido empregada devido à sua simplicidade e natureza agnóstica ao modelo, o que permite sua aplicação em diferentes arquiteturas preditivas. Este estudo aborda as limitações dessa técnica, como a geração de instâncias fora da distribuição original e a sensibilidade a variáveis correlacionadas. O conjunto de dados utilizado abrange características do solo e condições climáticas, e a importância das variáveis foi calculada utilizando a implementação da PFI pela biblioteca scikit-learn. Os resultados demonstram a eficácia da técnica na identificação das variáveis mais relevantes. Comparações com outras técnicas, como SHAP e LIME, são apresentadas, destacando suas vantagens e limitações no contexto da classificação supervisionada.

PALAVRAS-CHAVE: Aprendizado de Máquina; Importância de Variáveis por Permutação; Floresta Aleatória; Modelos Baseados em Árvores; Classificação Supervisionada.

ANALYSIS OF PERMUTATION FEATURE IMPORTANCE IN TREE-BASED SUPERVISED CLASSIFICATION MODELS

ABSTRACT: This study examines the application of the Permutation Feature Importance (PFI) technique in supervised classification models, with a focus on tree-based algorithms, such as Random Forest. PFI has been employed due to its simplicity and model-agnostic nature, which allows its application across different predictive architectures. This study addresses the limitations of this technique, such as the generation of out-of-distribution instances and sensitivity to correlated variables. The dataset used includes soil characteristics and climatic conditions, and the importance of the variables was calculated using the PFI implementation from the scikit-learn library. The results demonstrate the technique's effectiveness in identifying the most relevant variables. Comparisons with other techniques, such as SHAP and LIME, are presented, highlighting their advantages and limitations in the context of supervised classification.

KEYWORDS: Machine Learning; Permutation Feature Importance; Random Forest; Tree-Based Models; Supervised Classification.

INTRODUÇÃO

A produção agrícola resulta da complexa interação entre variáveis ambientais, atributos do solo e a dinâmica de nutrientes no sistema solo-planta-atmosfera. Essas variáveis estão fortemente interligadas e são influenciadas pelas práticas de manejo agrícola, como a escolha de variedades, idade das plantas, corte, maturação, ataques de pragas e estresses climáticos. Assim, novos estudos sobre o potencial produtivo de culturas em solos brasileiros são essenciais. O avanço na ciência de dados oferece grande potencial para explorar e otimizar essas interações, contribuindo para práticas agrícolas mais eficientes. (Wolfert et al., 2017)

Modelos de Machine Learning baseados em árvores de decisão, como o Random Forest, têm demonstrado desempenho robusto em tarefas de classificação e regressão, principalmente ao lidar com grandes volumes de dados e variáveis complexas (Breiman, 2001). Porém, além da precisão preditiva, torna-se importante, em vários contextos, compreender quais variáveis mais contribuem para as predições do modelo. A técnica de Permutation Feature Importance (PFI) tem sido empregada para medir a importância das variáveis em modelos de aprendizado supervisionado.

A PFI oferece uma abordagem agnóstica ao modelo, avaliando a relevância de cada variável por meio da permutação de seus valores e observando o impacto na acurácia do modelo. Essa técnica pode ser aplicada a diversos modelos preditivos, como árvores de decisão e redes neurais, sem a necessidade de conhecer a estrutura interna do modelo (Breiman, 2001). Sua simplicidade e interpretabilidade tornam a PFI uma opção vantajosa para uso em diferentes áreas de domínio. O presente trabalho analisa a eficácia da PFI em um modelo de Random Forest, utilizando dados de características do solo e condições climáticas, além de comparar a técnica com métodos como SHAP (Lundberg et al., 2017) e LIME (Ribeiro et al., 2016).

MATERIAL E MÉTODOS

Os dados utilizados neste estudo foram obtidos do "Crop Recommendation Dataset" disponibilizado por Nalluri (2024) em Kaggle, e incluem observações relacionadas à composição do solo e variáveis ambientais. O dataset é composto por 2200 instâncias e 7 variáveis (features): Nitrogen (nitrogênio no solo), Phosphorus (fósforo no solo), Potassium (potássio no solo), Temperature (temperatura ambiente), Humidity (umidade relativa), pH_Value (pH do solo) e Rainfall (precipitação em milímetros). A variável alvo ("Crop") é categórica, representando a cultura recomendada para cada conjunto de condições.

Para a modelagem, foi utilizado o algoritmo Random Forest implementado na biblioteca scikit-learn. A configuração incluiu 100 árvores de decisão, com profundidade máxima indefinida, permitindo que o modelo capturasse as complexidades presentes nos dados. A técnica de Permutation Feature Importance (PFI) foi aplicada após o ajuste do modelo, avaliando a importância de cada variável ao comparar a acurácia do modelo original com a acurácia obtida após a permutação aleatória dos valores de uma variável específica. A PFI pode ser formalmente expressa da seguinte forma:

Permutation Feature Importance:

$$\Delta_{FI}(X_j) = \frac{1}{R}\sum_{r=1}^{R}(\mathcal{L}(f(X_{\text{perm},j}), y) - \mathcal{L}(f(X), y)) \quad (1)$$

em que,

X_j representa a j-ésima variável permutada;

$\mathcal{L}(f(X), y)$ é a função de perda do modelo original;

$\mathcal{L}(f(X_{\text{perm},j}), y)$ é a perda do modelo após a permutação da variável;

R é o número de permutações realizadas.

A PFI permite medir a relevância de uma variável observando a variação no erro do modelo após a permutação de seus valores. Variáveis que exercem maior impacto tendem a causar um aumento mais expressivo no erro preditivo, enquanto variáveis menos influentes resultam em variações menores na performance do modelo (Molnar, 2020).

RESULTADOS E DISCUSSÃO

Os resultados obtidos a partir da aplicação da técnica de Permutation Feature Importance (PFI) no modelo de Random Forest indicaram que as variáveis climáticas, especialmente Humidity e Rainfall, exerceram a maior influência sobre a acurácia do modelo. Conforme apresentado na Tabela 1, a permutação dessas variáveis resultou em reduções médias de 0,31 e 0,20, respectivamente.

Além disso, as variáveis relacionadas ao solo, como Nitrogen, Phosphorus, e Potassium, também mostraram impacto considerável, com reduções médias na acurácia entre 0,12 e 0,15. Por outro lado, pH_Value apresentou uma influência mínima, com uma redução de apenas 0,01, sugerindo que essa variável não desempenha um papel significativo no contexto do modelo em questão.

A produtividade de culturas em sistema de sequeiro é altamente dependente das interações entre suas fases fenológicas e as variações interanuais do tempo e clima. Um estudo recente de Nguru e Mwongera (2023) destaca que valores subótimos de umidade relativa podem limitar o rendimento das culturas e afetar a qualidade dos produtos agrícolas.

A precipitação, por sua vez, sempre figura entre os fatores ambientais mais críticos para a produção agrícola, pois influencia diretamente a disponibilidade de água para as plantas. Estudos recentes mostram que a sua variabilidade, como chuvas intensas ou secas prolongadas, prejudicam a produtividade agrícola e aumentam a perda de nutrientes essenciais no solo, como nitrogênio e fósforo, afetando a eficiência dos fertilizantes e a produtividade das culturas (Zambrano-Medina et al., 2024).

TABELA 1. Importância das Variáveis (Permutation Feature Importance) calculada para o modelo de Random Forest.

Variável	Importância Média	Desvio-Padrão
Nitrogen	0.15	0.03
Phosphorus	0.12	0.02
Potassium	0.14	0.03
Temperature	0.02	0.01
Humidity	**0.31**	**0.05**
pH_Value	0.01	0.01
Rainfall	**0.20**	**0.04**

Através da Figura 1 pode-se notar graficamente a redução média na acurácia de cada variável, juntamente com o desvio padrão associado às permutações. A maior influência da variável Humidity é clara, enquanto Temperature e pH_Value tiveram impacto mínimo no modelo.

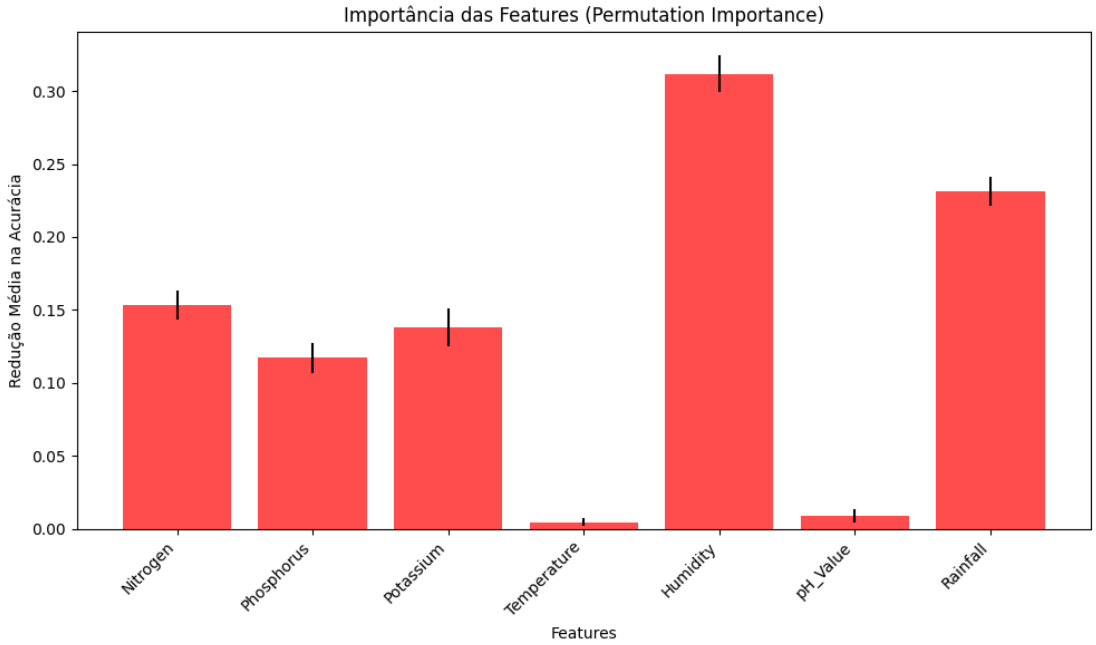

FIGURA 1. Gráfico de Importância das Variáveis (Permutation Feature Importance) calculado para o modelo de Random Forest.

Embora a técnica de PFI tenha sido utilizada, suas limitações devem ser reconhecidas. Primeiramente, como apontado por Strobl et al. (2008), a permutação das variáveis pode gerar instâncias fora da distribuição original dos dados, o que compromete a acurácia em datasets com forte correlação entre variáveis. Além disso, técnicas mais avançadas, como SHAP, oferecem uma decomposição mais detalhada da importância das variáveis, principalmente em modelos com alta dimensionalidade (Lundberg et al., 2017).

Por outro lado, o PFI se destaca por sua simplicidade e por ser uma técnica agnóstica ao modelo, o que permite sua aplicação em uma ampla variedade de arquiteturas preditivas. Essa característica faz com que o PFI seja especialmente útil em modelos como Random Forest, que combinam múltiplas árvores de decisão, onde a importância das variáveis pode variar significativamente entre as árvores individuais (Breiman, 2001).

Outra limitação do PFI é sua sensibilidade à dimensionalidade dos dados. Em datasets com muitas variáveis irrelevantes ou redundantes, o PFI pode superestimar ou subestimar a importância de algumas variáveis, uma vez que a permutação de uma variável com pouca importância relativa pode gerar flutuações menores na acurácia. Isso se agrava em modelos complexos, como os de alta dimensionalidade, onde a interação entre as variáveis desempenha um papel crucial no desempenho do modelo.

Em comparação, técnicas como SHAP e LIME têm a vantagem de fornecer explicações locais e globais mais precisas. O SHAP se baseia na teoria dos valores de Shapley, oferecendo uma decomposição da importância das variáveis que considera as interações entre elas. Essa abordagem é mais robusta para lidar com variáveis correlacionadas, algo que o PFI, em sua versão simples, não aborda adequadamente (Lundberg et al., 2017). No entanto, o SHAP é consideravelmente mais caro em termos computacionais, especialmente em modelos complexos com muitas variáveis.

O LIME, por sua vez, constrói modelos lineares simples ao redor de uma instância específica para explicar suas predições. Isso oferece uma boa explicabilidade local, mas não fornece uma visão abrangente sobre a importância das variáveis ao longo de todo o modelo, como faz o PFI.

Portanto, a escolha entre o PFI e outras técnicas de importância de variáveis deve ser guiada pelos requisitos do problema. Para modelos onde a simplicidade e a interpretabilidade global são prioridades, como na recomendação de culturas ou predição de características climáticas, o PFI demonstra ser uma escolha robusta e eficiente. No entanto, em cenários onde as interações entre variáveis ou correlações complexas são críticas para a performance do modelo, técnicas mais sofisticadas como o SHAP podem oferecer uma visão mais acurada da importância das variáveis.

CONCLUSÕES

A técnica de Permutation Feature Importance (PFI) aplicada ao modelo de Random Forest demonstrou ser uma ferramenta eficaz para a identificação das variáveis mais influentes em um problema de classificação supervisionada. As varíaveis com maior importância para o modelo foram a umidade do ar e a precipiação.

O uso do Random Forest permitiu capturar interações complexas entre as variáveis, e o PFI forneceu uma forma simples e intuitiva de medir o impacto de cada variável na acurácia do modelo.

A simplicidade e a generalidade do PFI o tornam uma técnica útil em muitas aplicações práticas de aprendizado supervisionado, especialmente quando é necessária uma visão global da importância das variáveis. Contudo, para problemas com alta dimensionalidade ou forte correlação entre variáveis, uma análise complementar com outras técnicas de interpretabilidade pode fornecer uma compreensão mais completa do comportamento do modelo.

CONTRIBUIÇÕES DOS AUTORES

Ambos os autores elaboraram a metodologia e realizaram conjuntamente análise de dados, experimentos e avaliação de resultados, bem como redigiram e revisaram este artigo e aprovaram a versão submetida.

AGRADECIMENTOS

Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei nº 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado Residência em TIC 03 - Aditivo, DOU 01245.013770/2020-64.

REFERÊNCIAS

BREIMAN, L. Random Forests. Machine Learning, Dordrecht, v. 45, p. 5–32, 2001. Disponível em: https://doi.org/10.1023/A:1010933404324. Acesso em: setembro 2024.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I.; VON LUXBURG, U.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Eds.). Advances in Neural Information Processing Systems, v. 30. Curran Associates, Inc., 2017.

MOLNAR, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2ª ed. 2024. Disponível em: https://christophm.github.io/interpretable-ml-book/. Acesso em: setembro 2024.

NALLURI, V. Crop Recommendation Dataset. Kaggle. Disponível em: https://www.kaggle.com/datasets/varshitanalluri/crop-recommendation-dataset. Acesso em: setembro 2024.

NGURU, W.; MWONGERA, C. Predicting the future climate-related prevalence and distribution of crop pests and diseases affecting major food crops in Zambia. PLOS Clim, v.2, n.1, p.e0000064, 2023. Disponível em: https://doi.org/10.1371/journal.pclm.0000064. Acesso em: setembro 2024.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). New York: Association for Computing Machinery, 2016. p. 1135–1144. Disponível em: https://doi.org/10.1145/2939672.2939778. Acesso em: setembro 2024.

STROBL, C.; BOULESTEIX, A. L.; KNEIB, T.; AUGUSTIN, T.; ZEILEIS, A. Conditional variable importance for random forests. BMC Bioinformatics, v. 9, p. 307, 2008. Disponível em: https://doi.org/10.1186/1471-2105-9-307. Acesso em: setembro 2024.

WOLFERT S.; GE L.; VERDOUW C.; BOGAARDT M. J. Big data in smart farming, a review. Agricultural Systems, v.153, p. 69-80, 2017. Disponível em: https://doi.org/10.1016/j.agsy.2017.01.02 . Acesso em: setembro 2024.

ZAMBRANO-MEDINA, Y. G.; ACEVES, E. A.; AGUILAR, L. Y; P.; MONJARDIN-ARMENTA, S. A.; PLATA-ROCHA, W.; FRANCO-OCHOA, C.; CHÁVEZ-MARTINEZ, O. The Impact of Climate Change on Crop Productivity and Adaptation and Mitigation Strategies in Agriculture. In: KANGA, S.; SINGH, S. K.; SHEVKANI, K.; PATHAK, V.; SAJAN, B. (eds) Transforming Agricultural Management for a Sustainable Future. 2024. World Sustainability Series. Springer, Cham. Disponível em: https://doi.org/10.1007/978-3-031-63430-7_1. Acesso em: setembro 2024.