MODELAGEM PREDITIVA COM DADOS INCOMPLETOS: INTELIGÊNCIA ARTIFICIAL NO MELHORAMENTO DE PLANTAS

Samuel da Silva Agostinho Branco, Luan Tiago dos Santos Carbonari, Carlos Zacarias Joaquim Júnior, Paulo Henrique Cerutti, Henrique de Sá Albino, Mauro Bitencourt de Souza, Lucas Daniel Chaves, João Antonio Dalmagro, Aires da Costa, Geovanna Alves Burdzaki, Jefferson Luís Meirelles Coimbra

INTRODUÇÃO

Em experimentos agrícolas, dados ausentes são comuns devido a falhas na coleta, medições ou perdas aleatórias, podendo comprometer análises e modelos (1). O método MICE (Imputação Multivariada por Equações Encadeadas) se destaca por gerar versões completas dos dados, preservando relações multivariadas e estimando a incerteza das imputações (2). Considerando esse cenário, o presente trabalho tem como objetivo examinar o reflexo de diferentes níveis de imputação de dados no desempenho de aprendizado de máquina aplicado a caracteres agronômicos em genótipos de feijão, visando compreender os impactos da imputação sobre a qualidade preditiva dos modelos em diferentes situações de perda de informação.

DESENVOLVIMENTO

O experimento foi realizado a campo com genótipos de feijão comum (*Phaseolus vulgaris* L.) organizados em três blocos adjacentes. Foram mensurados os caracteres estatura da planta (ET), diâmetro do caule (DC) e número de legumes por planta (NL), sendo a massa de grãos por planta (MGP) utilizada como variável resposta. Para simular cenários de ausência de dados, inseriram-se valores faltantes de forma aleatória nas variáveis preditoras, em proporções de 20%, 40%, 60% e 80%. Em cada cenário, as linhas contendo valores ausentes foram removidas, originando conjuntos incompletos. A imputação foi realizada pelo método MICE, que faz estimativas sucessivas a partir das próprias relações entre as variáveis, mantendo a coerência do conjunto de dados. A variável resposta (MGP) foi discretizada em quatro categorias com base nos quartis: valores <4,99 (“ruim”), entre 5,00-8,50 (“razoável”), entre 8,51-14,00 (“bom”) e >14,00 (“excelente”). Os dados completos e imputados foram utilizados na classificação por redes neurais artificiais, com o desempenho dos modelos sendo mensurado pela acurácia.

RESULTADOS

Observou-se, que o conjunto com 20% de dados ausentes (sem imputação) apresentou a maior acurácia (0,78). Isso pode estar relacionado à exclusão de linhas com valores faltantes, que reduziu a base de dados e possivelmente eliminou ruídos, porém deixando-a enviesada, refletindo melhor desempenho aparente sem garantir maior generalização. Nos cenários com 40% a 80% de dados ausentes, a acurácia manteve-se relativamente estável (0,70 a 0,73), indicando certa robustez, embora a redução das observações possa ter comprometido a representatividade do conjunto. Quanto à imputação pelo MICE, houve queda da acurácia em todos os níveis, especialmente em 60% (0,55) e 80% (0,50), sugerindo que imputações excessivas introduzem vieses e reduzem a variabilidade natural. Assim, de acordo com a

Literatura que estima um nível aceitável de ausência de dados de 20% (3), os resultados obtidos neste estudo indicaram que, para esse conjunto específico, a imputação pelo MICE mostrou-se razoável até mesmo em 40%, sem comprometer drasticamente o desempenho do modelo.

CONSIDERAÇÕES FINAIS

Está pesquisa evidenciou que, com base nos resultados obtidos, a imputação via MICE pode ser aplicada a um nível aceitável de até 40% de dados ausentes, acima do qual a perda de desempenho e a introdução de vieses tornam-se mais evidentes. Todavia a ausência de dados impacta diretamente o desempenho de modelos de aprendizado, a exclusão de linhas com valores faltantes pode melhorar momentaneamente a acurácia, mas implica perda de variabilidade genética. Nesse sentido, a imputação pelo método MICE se apresenta como alternativa para preservar informações.

Palavras-chave: *Phaseolus vulgaris* L.; Imputação de Dados; Ciência de Dados.

ILUSTRAÇÕES

Acurácia em Diferentes Cenários

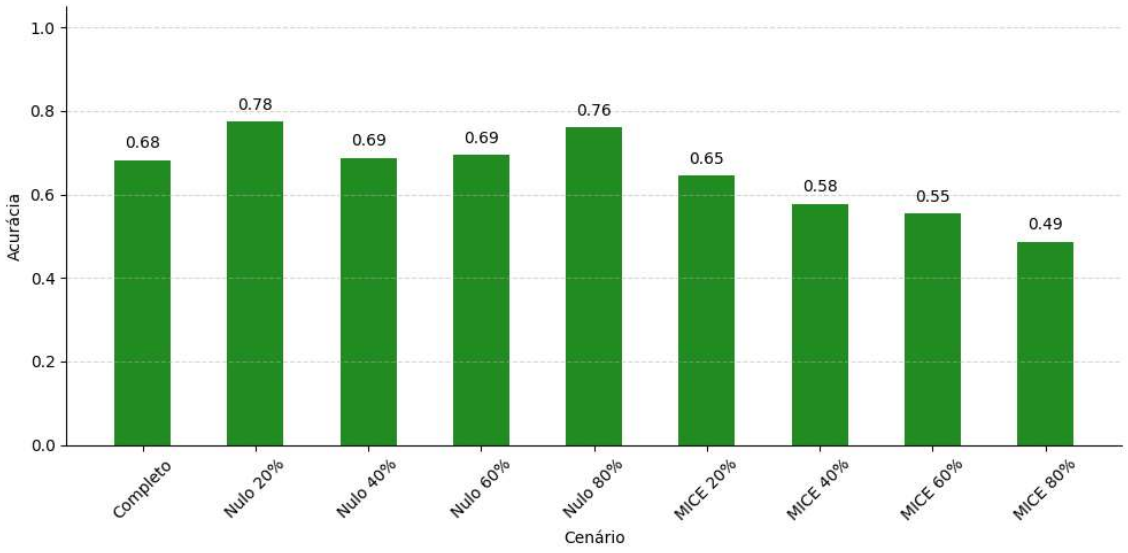

Gráfico 1. *Desempenho do modelo em diferentes cenários de dados completos, ausentes e imputados pelo método MICE.*

REFERÊNCIAS BIBLIOGRÁFICAS

1. SCHAFER, Joseph L.; OLSEN, Maren K. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, v. 33, n. 4, p. 545-571, 1998.
2. AZUR, Melissa J. et al. Multiple imputation by chained equations: what is it and how does it work?. *International Journal of Methods in Psychiatric Research*, v. 20, n. 1, p. 40-49, 2011.
3. PEDERSEN, A. B. et al. Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, v. 9, p. 157-166, 2017. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC4442050/
4. SCHMIDHUBER, J. Aprendizado profundo em redes neurais: uma visão geral. *Redes Neurais*, v. 61, p. 85-117, 2015.
5. ALI, N. F. et al. Comparison of five imputation methods in handling missing data in a continuous frequency table. *ResearchGate*, 2022. Disponível em: https://www.researchgate.net/publication/360174018_Comparison_of_five_imputation_methods_in_handling_missing_data_in_a_continuous_frequency_table

DADOS CADASTRAIS

BOLSISTA: Samuel da Silva Agostinho Branco
MODALIDADE DE BOLSA: PROBIC/UDESC
VIGÊNCIA: 09/2024 a 08/2025 – Total: 12 meses
ORIENTADOR(A): Jefferson Luís Meirelles Coimbra
CENTRO DE ENSINO: CAV
DEPARTAMENTO: Agronomia
ÁREAS DE CONHECIMENTO: Ciências Agrárias/ Agronomia/ Melhoramento Vegetal
TÍTULO DO PROJETO DE PESQUISA: O Impacto do Ambiente de Cultivo no Potencial Genético do Feijão: Análise e Melhoramento com Redes Neurais
Nº PROTOCOLO DO PROJETO DE PESQUISA: NPP3749-2021