APRENDIZADO DE MÁQUINA SUPERVISIONADO *VERSUS* NÃO SUPERVISIONADO NO MELHORAMENTO DE PLANTAS

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

INTRODUÇÃO

O melhoramento de plantas é fundamental para o aumento da produção de alimentos, sendo responsável por cerca de 50% do incremento da produção por área (EVENSON; GOLLIN, 2003). A partir dessa importância, os programas modernos geram inúmeros dados, o que desafia a obtenção de informações com análises estatísticas clássicas. Nesse cenário, o Aprendizado de Máquina surge como uma alternativa visando auxiliar o melhorista. Assim, o presente trabalho tem como objetivo comparar abordagens de aprendizado de máquina não supervisionado e supervisionado, com foco no potencial de cada uma para melhorar o processo de desenvolvimento de cultivares superiores.

DESENVOLVIMENTO

O experimento foi conduzido no Instituto de Melhoramento e Genética Molecular (IMEGEM), localizado no Centro de Ciências Agroveterinárias. Utilizou-se 79 populações de feijão (*Phaseolus vulgaris L.*) com diferentes níveis de heterozigose. Foram mensurados os caracteres, estatura (cm), altura de inserção do primeiro legume (cm), diâmetro do caule (mm), número de legumes, número de grãos e massa de grãos por planta (g). Ao todo, foram mensuradas 5.136 plantas, totalizando 30.816 observações. Este conjunto de dados serviu como base para aplicação dos algoritmos (PEDREGOSA *et al.*, 2011). O aprendizado de máquina não supervisionado foi empregado com o objetivo de agrupar os genótipos semelhantes com base no conjunto de dados. O número de grupos foi determinado com base no critério de Mojena, conforme Souza Neto (2022). No aprendizado supervisionado, a finalidade foi realizar a previsão do mérito dos genótipos, utilizando três classes aplicadas ao conjunto de dados: *i)* plantas ruins (média -1 desvio); *ii)* plantas médias (± 1 desvio); *iii)* plantas boas (+ 1 desvio). No treino do modelo utilizou-se todas as características avaliadas, como variáveis preditoras, exceto a massa de grãos por planta, que foi usada como a variável resposta, a ser prevista. Foram utilizados 80% dos dados para o treino e 20% no teste do modelo (PAIXÃO *et al.*, 2022). Todas as análises computacionais foram realizadas utilizando a linguagem de programação Python (3.13.6) e as respectivas bibliotecas: Matplotlib, Numpy, Pandas, Scikitlearn, Seaborn.

RESULTADOS

Na abordagem não supervisionada foi obtido como resultado um dendrograma com correlação cofenética de 72%, que reflete a concordância da matriz de distâncias entre os genótipos que está representada no dendrograma (Figura 1A). Este dendrograma elucida as relações de dissimilaridade entre todos os genótipos, evidenciando a formação de agrupamentos distintos entre os genitores e suas respectivas progênies. Essa informação sugere que, para o conjunto de dados avaliados, há divergência genética. Com base nisso, o melhorista pode direcionar futuros cruzamentos de forma estratégica, visando incrementar a variabilidade genética nas populações segregantes ao hibridar artificialmente genitores de grupos distintos com características

complementares. Além disso, populações segregantes promissoras podem ser identificadas mediante a média do grupo formado. Por outro lado, a abordagem supervisionada foi empregada para a previsão do mérito dos genótipos quanto ao caráter massa de grãos por planta. A avaliação do desempenho desta previsão releva que na classe boa, que possui os indivíduos com maior massa de grãos, o algoritmo identificou corretamente 90% dessas plantas corretamente e 10% ele destinou de forma errônea a classe média (Figura 1B). Além disso, o aprendizado não classificou nenhuma planta boa como ruim, o que acarretaria no descarte de uma planta promissora. Com isso, a principal vantagem dessa abordagem, está compreendida na previsão de indivíduos com maior massa de grãos por planta. Para este estudo foram utilizadas características vegetativas e reprodutivas, porém seria possível utilizar apenas características vegetativas para saber se uma planta será promissora?

CONSIDERAÇÕES FINAIS

Os resultados obtidos com o uso dos aprendizados de máquina é promissor no melhoramento de plantas. A abordagem não supervisionada revelou a divergência genética entre os genótipos de feijão, podendo auxiliar na definição de cruzamentos estratégicos e identificação de populações segregantes promissoras; enquanto a abordagem supervisionada apresentou boa eficiência na previsão fenotípica, permitindo predizer genótipos superiores. Assim, a integração dessas técnicas pode otimizar o processo de melhoramento vegetal e acelerar o desenvolvimento de cultivares de feijão promissoras.

Palavras-chave: *Phaseolus vulgaris*, Ciência de dados, Aprendizado de máquina.

ILUSTRAÇÕES

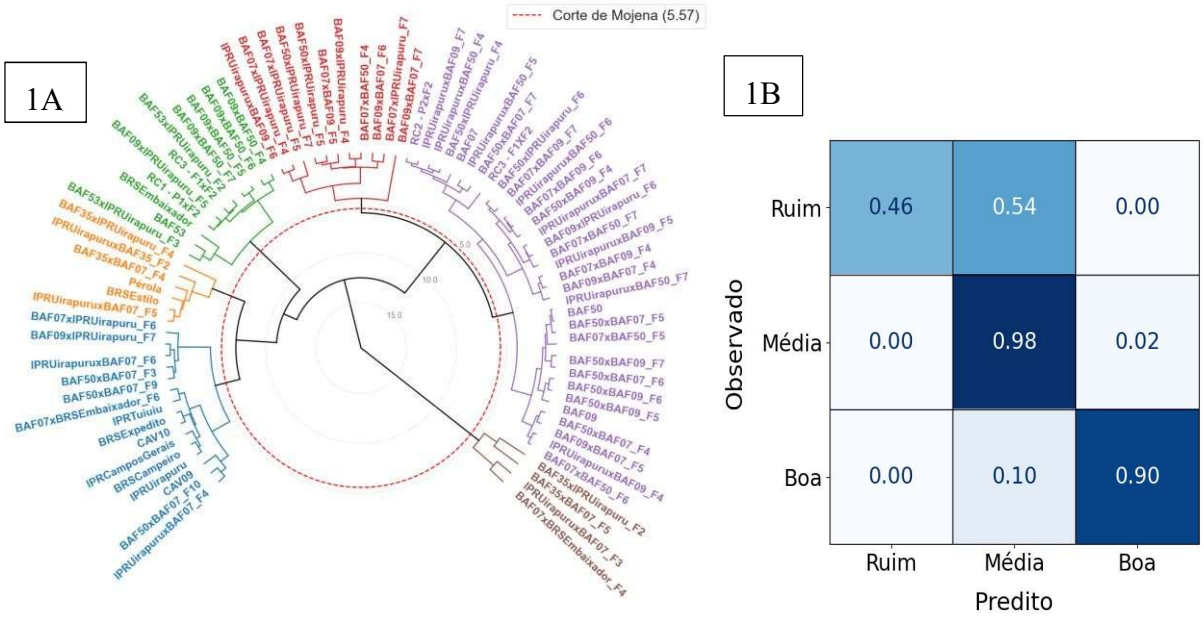

Figura 1. *Dendrograma em forma de leque obtido pelo aprendizado de máquina não supervisionado (1A) e matriz de confusão normalizada para avaliação do modelo referente ao aprendizado de máquina supervisionado (1B).*

REFERÊNCIAS BIBLIOGRÁFICAS

EVENSON, R. E.; GOLLIN, D. Assessing the impact of the green revolution, 1960 to 2000. **Science**, v. 300, n. 5620, p. 758-762, 2 maio 2003.

PAIXÃO, Gabriela Miana de Mattos et al. Machine learning na medicina: revisão e aplicabilidade. **Arquivos Brasileiros de Cardiologia**, v. 118, n. 1, p. 95-102, 2022.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011.

SOUZA NETO, Thaynara Aparecida de. Estudo de divergência genética em acessos de Capsicum annuum L., utilizando métodos de agrupamento. 2022.

DADOS CADASTRAIS

BOLSISTA: Lucas Daniel Chaves
MODALIDADE DE BOLSA: PIBIC/CNPq
VIGÊNCIA: 04/2025 a 08/2025 – Total: 5 meses
ORIENTADOR(A): Jefferson Luís Meirelles Coimbra
CENTRO DE ENSINO: CAV
DEPARTAMENTO: Agronomia
ÁREAS DE CONHECIMENTO: Ciências Agrárias / agronomia
TÍTULO DO PROJETO DE PESQUISA: Consequência da Epistasia na expressão do sistema radicular de feijão dos grupos gênicos andino e mesoamericano
Nº PROTOCOLO DO PROJETO DE PESQUISA: NPP3750-2021