Agronomia

Avaliação de métodos de seleção de variáveis e de redução de dimensionalidade para a predição da Resistência à Brusone em Arroz na regressão logística

Douglas Geovanini de Paiva Mosca Leite - Graduando em Engenharia Ambiental e Sanitária

Nayara Maria Barbosa De Sousa - Mestranda em Estatística e Experimentação Agropecuária

João Vitor Andrade Alves de Souza - Mestrando em Estatística e Experimentação Agropecuária

João Paulo Assis Bonifácio - Mestrando em Estatística e Experimentação Agropecuária

Igor Pereira Gomes - Mestre em Engenharia Elétrica, UFMG

Geraldo Magela da Cruz Pereira - Professor do Departamento de Estatística, UFLA. - geraldo.pereira@ufla.br - Orientador(a)

Resumo

O arroz é um alimento fundamental, sendo base nutricional para cerca de 2,5 bilhões de pessoas no mundo, representando aproximadamente 20% da ingestão mundial de energia e 15% do aporte de proteínas. No Brasil, assume grande relevância alimentar juntamente com o feijão, constituindo um dos principais componentes da dieta da população. Dentre os diversos problemas fitossanitários relacionados à produção do arroz, a Brusone, (doença causada pelo fungo Pyricularia oryzae) se destaca, provocando perdas anuais estimadas em US$66 bilhões. Este estudo tem como objetivo avaliar o impacto da utilização dos métodos de redução de dimensionalidade Principal Component Analysis (PCA) e Independent Component Analysis (ICA), e do método de seleção de variáveis Boruta no desempenho do modelo de regressão logística para a predição da Brusone. Estes métodos foram comparados com a utilização da seleção em duas etapas, seleção univariada, seguida da aplicação do stepAIC, e com o procedimento avaliado em Ossifo et al. (2022). Estes procedimentos foram aplicados a um banco de dados composto por fenótipos de 413 plantas de arroz de 82 países, com um total de 15 fenótipos avaliados. No pré-processamento de dados, a variável discreta, resistência à brusone, foi convertida em binária, 0 para valores menores ou iguais à mediana, e 1, caso contrário. Em seguida, plantas com valores ausentes para mais de sete variáveis foram removidas, reduzindo o conjunto para 364 indivíduos. Como filtro, foram descartadas variáveis altamente correlacionadas (cutoff maior igual0.8), resultando na exclusão de Seed.volume e Seed.width. Por fim, para a imputação dos 93 valores ausentes (1,60%), utilizou-se o algoritmo K-Nearest Neighbors (KNN). Os métodos foram avaliados pelas métricas AUC e F1-Score. Os resultados mostram que os métodos de seleção de variáveis em duas etapas (AUC: 0.7281, F1-Score: 0.7177) e o Boruta (AUC: 0.7296, F1-Score: 0.7126) apresentaram o melhor desempenho global, superando as técnicas de redução de dimensionalidade PCA, ICA e procedimento avaliado em Ossifo et al. (2022). Dessa forma, os resultados evidenciam o potencial dessas técnicas para auxiliar a seleção de fenótipos relacionados à tolerância e resistência ao fungo, oferecendo uma ferramenta promissora para mitigar perdas.

Palavras-Chave: Redução de dimensionalidade, Seleção de variáveis, Regressão Logística.
Instituição de Fomento: UFLA

Link do pitch: https://youtu.be/xbL5WSjW12Q?si=59Gnu4Tcis0aV-lW