

INTEGRAÇÃO DA ANÁLISE UNIVARIADA E APRENDIZADO DE MÁQUINA NA AVALIAÇÃO DE GENÓTIPOS DE SOJA EM ENSAIOS DE VALOR DE CULTIVO E USO

INTEGRATION OF UNIVARIATE ANALYSIS AND MACHINE LEARNING FOR SOYBEAN GENOTYPE SELECTION IN VALUE FOR CULTIVATION AND USE TRIALS

Penélope Marques Chagas¹; Márcio Mandrá Júnior²; Eryadison Flávio Bonifacio de Araujo³; Laura Pinto Rodrigues⁴; Matheus Siqueira de Oliveira⁵; Jardel da Silva Souza⁶; Sandra Helena Unêda-Trevisoli⁷.

¹Mestranda em Agronomia (Genética e Melhoramento de plantas) pela Unesp/FCAV, ²Doutorando em Agronomia (Genética e Melhoramento de plantas) pela Unesp/FCAV, ³Doutorando em Agronomia (Genética e Melhoramento de plantas) pela Unesp/FCAV, ⁴Mestranda em Agronomia (Genética e Melhoramento de plantas) pela Unesp/FCAV, ⁵Doutorando em Agronomia (Genética e Melhoramento de plantas) pela Unesp/FCAV, ⁶Pós-doutorando em Agronomia (Genética e Melhoramento de plantas), Unesp/FCAV; ⁷Profa. Dra. em Agronomia (Genética e Melhoramento de plantas). Departamento de Produção Vegetal, Unesp/FCAV.

Resumo

A recomendação de cultivares de soja depende da seleção de genótipos com elevado desempenho produtivo, estabilidade e adaptação às diferentes condições edafoclimáticas e de cultivo. Mediante o exposto, o presente estudo objetivou avaliar a eficiência da integração de análises univariadas convencionais com algoritmos de aprendizado de máquina, na seleção de genótipos de soja em Ensaios de Valor de Cultivo e Uso. Para tanto, foram avaliados 30 genótipos de soja em Ensaios de Valor de Cultivo e Uso, durante a safra 2024/2025 em Jaboticabal e Piracicaba, SP. Os experimentos foram conduzidos no delineamento de blocos casualizados, com três repetições. De acordo com os resultados obtidos, observou-se que, a análise conjunta entre ambientes, evidenciou efeito significativo de genótipos e da interação genótipo × ambiente, indicando respostas diferenciadas dos genótipos entre os locais avaliados. Como abordagem complementar, os modelos Random Forest e Gradient Boosting apresentaram desempenho preditivo semelhante, com coeficientes de determinação próximos de 0,66, indicando capacidade moderada de explicar a variabilidade observada na produtividade de grãos dos genótipos avaliados. Os resultados indicam que a integração entre as análises clássicas e ferramentas estatísticas baseadas no aprendizado de máquina, pode ampliar a interpretação dos resultados obtidos em Ensaios de Valor de Cultivo e Uso, auxiliando no processo seletivo e na análise do desempenho genotípico em programas de melhoramento de soja.

Palavras-chave

Glycine max; Gradient Boosting; Interação genótipo × ambiente; Random Forest.

Abstract

Soybean cultivar recommendation relies on the identification of genotypes that combine high grain yield, stability, and broad adaptation across diverse edaphoclimatic and crop management conditions. In this context, the present study aimed to evaluate the effectiveness of integrating conventional univariate statistical analyses with machine learning algorithms for soybean genotype

selection in Value for Cultivation and Use (VCU) trials. A total of 30 soybean genotypes were evaluated during the 2024/2025 growing season in VCU trials conducted at Jaboticabal and Piracicaba, São Paulo State, Brazil. The experiments were established using a randomized complete block design with three replications. Combined analysis across environments revealed significant effects of genotype and genotype $\times$ environment interaction, indicating differential genotype performance across the evaluated locations. As a complementary analytical approach, the Random Forest and Gradient Boosting algorithms exhibited comparable predictive performance, with coefficients of determination (R^2) of approximately 0.66, indicating a moderate capacity to explain the observed variation in grain yield among the evaluated genotypes. Overall, the results demonstrate that integrating conventional statistical methods with machine learning approaches can improve the interpretation of data generated in VCU trials, thereby supporting genotype selection and enhancing the assessment of genotypic performance in soybean breeding programs.

Keywords

Glycine max; Gradient Boosting; Genotype $\times$ environment interaction; Random Forest.

1. Introdução

A soja destaca-se entre as principais culturas agrícolas do Brasil e do mundo, em razão de seu elevado valor econômico, produtivo e nutricional (SOUZA; SMIDERLE, 2024). Nesse cenário, o melhoramento genético é essencial para a obtenção de cultivares superiores, com maior produtividade, ampla adaptação edafoclimática e resistência a pragas e doenças (SILVA, 2024; VIEIRA, 2020). Dessa forma, a recomendação de cultivares superiores baseia-se na seleção de genótipos com elevado potencial produtivo, estabilidade e adaptação às diferentes condições edafoclimáticas existentes e diferentes condições de cultivos agrícolas (OBUA et al., 2026).

A introdução de novas cultivares superiores no mercado exige a validação dos genótipos em Ensaios de Valor de Cultivo e Uso (VCU) (MAPA, 2024). Esses ensaios constituem uma etapa fundamental dos programas de melhoramento, pois permitem avaliar o potencial produtivo dos genótipos e estimar a magnitude da interação genótipo $\times$ ambiente ($G \times A$), responsável por alterações no desempenho relativo entre ambientes de avaliação (OBUA et al., 2026). Tradicionalmente, esses dados são analisados por meio da análise de variância (ANOVA), que permite verificar a significância dos efeitos de genótipo, ambiente e $G \times A$ (ABEBE et al., 2024).

Embora a ANOVA seja uma técnica amplamente empregada, o avanço das técnicas de aprendizado de máquina tem ampliado a capacidade de exploração de dados experimentais, permitindo identificar padrões complexos e gerar informações complementares às análises estatísticas convencionais (CROSSA et al., 2025). Assim, o emprego complementar de métodos estatísticos convencionais e algoritmos de aprendizado de máquina representa uma abordagem promissora para aprimorar a interpretação do desempenho genotípico em ensaios multiambientes

(YAN et al., 2025). Mediante o exposto, o presente estudo objetivou avaliar a eficiência da integração de análises univariadas convencionais com algoritmos de aprendizado de máquina, na seleção de genótipos de soja em Ensaios de Valor de Cultivo e Uso.

2. Material e Métodos

Os experimentos foram conduzidos durante a safra agrícola 2024/2025 em dois ambientes: na Fazenda de Ensino, Pesquisa e Extensão da Faculdade de Ciências Agrárias e Veterinárias da UNESP, em Jaboticabal (SP), pertencente à região edafoclimática (REC) 302; e na Fazenda Experimental da Escola Superior de Agricultura Luiz de Queiroz, em Piracicaba (SP), pertencente à REC 203. Ambos os experimentos foram conduzidos em sistema de plantio direto, respeitando o zoneamento agrícola da cultura para cada região.

O delineamento experimental foi o de blocos casualizados, com três repetições. Cada parcela foi composta por quatro linhas de cinco metros de comprimento, espaçadas em 0,50 m, sendo consideradas úteis as duas linhas centrais. Foram avaliados 30 genótipos de soja, sendo 26 linhagens avançadas provenientes do Programa de Melhoramento Genético da UNESP/FCAV e quatro cultivares comerciais, utilizadas como testemunhas.

As análises estatísticas foram realizadas por meio de script desenvolvido em Python 3.13.9. Inicialmente, foram realizadas análises de variância individuais para cada ambiente. Posteriormente, foi conduzida a análise de variância conjunta, considerando os efeitos de ambiente, blocos dentro de ambiente, genótipos e interação genótipo $\times$ ambiente. Como abordagem complementar, foram ajustados e comparados modelos de Random Forest e Gradient Boosting para predição do caráter produtividade de grãos, utilizando ambiente, bloco e genótipo como variáveis preditoras. Os modelos foram avaliados por validação hold-out (75% para treinamento e 25% para teste), validação cruzada k-fold ($k = 5$), Repeated K-Fold e otimização de hiperparâmetros por GridSearchCV. O desempenho dos modelos foi comparado por meio do coeficiente de determinação (R^2), raiz do erro quadrático médio (RMSE), erro absoluto médio (MAE) e erro percentual absoluto médio (MAPE).

3. Resultados e Discussão

As análises de variância individuais evidenciaram diferenças significativas entre os genótipos em ambos os ambientes avaliados ($p < 0,05$), demonstrando a existência de variabilidade genética para o caráter produtividade de grãos e indicando potencial para discriminação de genótipos superiores. No ambiente de Jaboticabal, também foi observado efeito significativo de blocos,

evidenciando a eficiência do delineamento experimental em controlar parte da variabilidade ambiental da área experimental. Em Piracicaba, o efeito de blocos não foi significativo.

A análise conjunta, apresentada na Tabela 1, revelou efeito altamente significativo dos genótipos ($p < 0,001$) e da interação genótipo $\times$ ambiente ($p < 0,001$), enquanto o efeito de ambiente não foi significativo. Esse resultado indica que, embora a produtividade média dos dois ambientes tenha sido semelhante, os genótipos responderam de maneira distinta em cada local, caracterizando interação $G \times A$ e reforçando a importância da avaliação multiambiental para a seleção de genótipos adaptados. A presença dessa interação demonstra que a classificação relativa dos genótipos varia entre os ambientes avaliados, aspecto amplamente reconhecido como um dos principais desafios dos programas de melhoramento de soja.

TABELA 1. Resumo da análise de variância conjunta para produtividade de grãos de 30 genótipos de soja avaliados em dois ambientes.

Fonte de variação	GL	F	p-valor
Ambiente	1	<0,01	ns
Genótipo	29	3,61	<0,001
Genótipo $\times$ Ambiente	29	5,58	<0,001
Resíduo	92		

ns = não significativo ($p \geq 0,05$).

Os algoritmos Random Forest e Gradient Boosting apresentaram desempenho semelhante, com coeficientes de determinação próximos de 0,66, indicando capacidade moderada de explicar a variabilidade observada para o caráter produtividade de grãos. No entanto, ambos os modelos apresentaram desempenho consistente nas avaliações realizadas, evidenciando potencial para aplicação como ferramentas complementares na análise de dados provenientes de ensaios de Valor de Cultivo e Uso. Embora os modelos de aprendizado de máquina não tenham superado a inferência estatística proporcionada pela análise de variância, sua utilização apresentou-se promissora para complementar a interpretação dos resultados dos experimentos e apoiar a análise do desempenho dos genótipos. Além disso, os resultados obtidos pelos algoritmos foram coerentes com as conclusões da análise de variância, reforçando o potencial dos modelos de aprendizado de máquina como abordagem complementar à análise estatística convencional.

4. Conclusões

A integração entre análises estatísticas convencionais e algoritmos de aprendizado de máquina mostrou-se promissora para ampliar a interpretação dos resultados experimentais e apoiar a análise do desempenho genotípico em ensaios de Valor de Cultivo e Uso, contribuindo para a seleção de genótipos superiores.

Apoio financeiro

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES).

Referências

- ABEBE, A. T. et al. **Genotype × environment interaction and yield stability of soybean (*Glycine max* L.) genotypes in multi-environment trials (METs) in Nigeria**. *Heliyon*, v. 10, n. 19, e38097, 2024.
- BRASIL. Ministério da Agricultura e Pecuária. **Valor de Cultivo e Uso**. Brasília, DF, 2024. Disponível em: <https://www.gov.br/agricultura/pt-br/assuntos/insumos-agropecuarios/insumos-agricolas/sementes-e-mudas/registro-nacional-de-cultivares/valor-de-cultivo-e-uso-2013-vcu>. Acesso em: 5 jul. 2026.
- CROSSA, J. et al. **Machine learning algorithms translate big data into predictive breeding accuracy**. *Trends in Plant Science*, v. 30, n. 2, p. 167-184, 2025.
- OBUA, T. et al. **Soybean rust resistance and yield performance of elite soybean genotypes across diverse environments in Uganda**. *Plant-Environment Interactions*, v. 7, n. 1, e70125, 2026.
- SILVA, A. B.; OLIVEIRA, C. D.; SOUSA, E. F. **Melhoramento genético de plantas para adaptabilidade às mudanças climáticas**. *Revista de Gestão e Secretariado*, São José dos Pinhais, v. 15, n. 9, p. 3925-3948, 2024.
- SOUZA, A. G.; SMIDERLE, O. J. **Cultura da soja: radiação solar, acúmulo de matéria seca e suas interações para alta produtividade**. In: **Ciências Agrárias: tecnologia, sustentabilidade e inovação**. v. 1. São Paulo: Editora Científica Digital, 2024. p. 84-107. Disponível em: <https://doi.org/10.37885/240416357>. Acesso em: 5 jul. 2026.
- VIEIRA, B. L. **Desempenho agrônomo de genótipos de soja e cruzamentos potenciais para ciclo e produtividade**. 2020. 42 f. Trabalho de Conclusão de Curso (Graduação em Agronomia), Universidade Federal de Uberlândia, Instituto de Ciências Agrárias, Uberlândia, 2020.
- YAN, C. et al. **ResDeepGS: a deep learning-based method for crop phenotype prediction**. *Methods*, v. 244, p. 65-74, 2025.