XIII CONGRESSO BRASILEIRO DE ARROZ IRRIGADO
12 A 15 DE AGOSTO 2025 | PELOTAS-RS

CLASSIFICAÇÃO DE GENÓTIPOS DE ARROZ COM ALGORITMOS DE MACHINE LEARNING

Ruan Bernardy[1]; Greice Neitzel[2]; Janete V. da Rosa Monteiro[3]; Silvia Leticia Rivero Meza[4]; Maurício de Oliveira[5]

Palavras-chave: Aprendizado supervisionado; Inteligência artificial; Fraude alimentar; *Oryza sativa*; Tecnologia agroindustrial

Introdução

A tecnologia de Inteligência Artificial (IA) é uma área cada vez mais popular da ciência da computação. Ela tenta analisar o mecanismo da inteligência humana e criar máquinas que possam imitar essa mentalidade, conseguindo um vasto armazenamento de conhecimento, aprendendo e progredindo continuamente através do trabalho com estes dados (Zeng et al., 2025). O Machine Learning (ML) é uma técnica dentro da IA que permite às máquinas aprenderem de forma autônoma a partir de dados fornecidos, o que significa que dá à máquina a capacidade de adquirir conhecimento e fazer julgamentos, encontrando padrões de forma autônoma (Clercq; Mahdi, 2025).

Essas tecnologias ganharam muito espaço na agricultura e na produção de alimentos, com o emprego em máquinas, sistemas para manejo da lavoura e também nos processos industriais. Contudo, um dos principais desafios do setor alimentício é combater a fraude alimentar, especialmente quando produtos de baixo valor são misturados a outros de maior valor agregado. No setor de industrialização de grãos, esse tipo de fraude ocorre especialmente com a adição de pequenas quantidades de arroz comum em lotes de arroz aromático, como é o caso da variedade de arroz Basmati (Ganopoulos et al., 2011), prejudicando os consumidores que adquirem esse tipo de arroz com custo mais elevado devido sua característica sensorial especifica.

Nesse sentido, a identificação e diferenciação de genótipos de arroz com base em suas características físico-químicas são essenciais para detectar e evitar esse tipo de adulteração (Hao et al., 2019). Em um estudo realizado por Panchbhai e Lanjewar (2025), os pesquisadores demonstraram a eficiência do espectrômetro de infravermelho próximo (NIR) na diferenciação de genótipos de arroz. No entanto, o processo industrial necessita de um sistema inteligente que permita uma rápida tomada de decisão. Nesse contexto, algoritmos de aprendizado de máquina, especialmente os baseados em árvores de decisão, como o Random Forest, surgem como soluções alternativas robustas para a classificação de dados multivariados.

[1]Eng. Agrícola, Mestre em Ciências Ambientais, Universidade Federal de Pelotas, Campus Universitário s/n, 96160-000, Capão do Leão – RS, ruanbernardy@yahoo.com.br
[2]Eng. Agrícola, Mestre em Ciência e Tecnologia de Alimentos, Universidade Federal de Pelotas, Campus Universitário s/n, 96160-000, Capão do Leão – RS, greice4321@gmail.com
[3]Tecnóloga em Gastronomia, Universidade Federal de Pelotas, Campus Universitário s/n, 96160-000, Capão do Leão – RS, janete.monteiro.ppgcta@hotmail.com
[4]Eng. de Alimentos, Mestre e Doutora em Ciência dos Alimentos, Universidade Federal de Pelotas, Campus Universitário s/n, 96160-000, Capão do Leão – RS, silvialrmezaufpel@gmail.com
[5]Agrônomo, Mestre e Doutor em Ciência e Tecnologia de Alimentos, Universidade Federal de Pelotas, Campus Universitário s/n, 96160-000, Capão do Leão – RS, mauricio@labgraos.com.br

Assim, o uso de técnicas de inteligência artificial permite não apenas classificar com precisão os genótipos de arroz, mas também identificar as variáveis mais relevantes para essa distinção, contribuindo para a eficiência no processo. Para alcançar esses resultados, pode-se utilizar a linguagem de programação Python, amplamente reconhecida por sua sintaxe intuitiva e pela extensa disponibilidade de bibliotecas voltadas ao aprendizado de máquina. Trata-se de uma ferramenta consolidada nas áreas de ciência de dados e inteligência artificial, o que a torna especialmente adequada para esse tipo de aplicação. Desta forma, o objetivo deste trabalho foi classificar genótipos de arroz através da técnica de *machine learning* aplicada em *python*.

Material e Métodos

O trabalho foi desenvolvido no Laboratório de Pós-Colheita, Industrialização e Qualidade de Grãos da Universidade Federal de Pelotas (LabGrãos/UFPel). Foram avaliados 5 genótipos de arroz disponibilizados pela empresa Camil Alimentos S/A: Guri INTA CL, IRGA 431 CL, IRGA 424 RI, BRS Pampa CL e Mendy Porã INTA CL. Foram analisados rendimento de inteiros e quebrados, composição centesimal pelo equipamento NIRS modelo FOSS DS2500L (teor de proteína, óleo, amido, cinzas e fibras), tempo de cocção, percentual de amilose e parâmetros de textura (dureza, coesividade, gumosidade, mastigabilidade e resiliência), através do protocolo padrão de análises do LabGrãos.

Após a caracterização dos genótipos, os dados foram pré-processados para padronização e remoção de valores discrepantes. Para garantir equilíbrio nas análises, foram realizadas 7 repetições por genótipo, com a remoção de 3 amostras que o algoritmo entendesse como distante da média interna de cada grupo, retirando discrepâncias.

Utilizou-se os algoritmos Random Forest, KNN, J48, MLP e Naive Bayes para classificar os genótipos com base nas variáveis preditoras analisadas. Os dados foram subdivididos em 70% para treinamento e 30% para teste, utilizando a técnica k-fold, com 4 folds, para calcular a acurácias da classificação. Foi empregada a codificação dos rótulos utilizando o LabelEncoder e validação por meio das métricas: acurácia, precisão, recall, F1-score, matriz de confusão e área sob a curva ROC (AUC), além do filtro Resample, que evita distorções por desequilíbrios nas amostras. Todas as etapas da metodologia foram realizadas dentro do ambiente de programação Google Colaboratory (Colab), usando a linguagem *python*.

Resultados e Discussão

Para avaliar o desempenho geral dos modelos, o enfoque foi dado para o F1-Score macro médio, que considera tanto a precisão quanto o recall de forma equilibrada entre todas as classes. O algoritmo Random Forest apresentou o melhor desempenho, com F1-score (0.900), seguido por Naive Bayes (0.717) e J48 (0.700). A Tabela 1 apresenta as outras métricas de acurácia analisadas dos algoritmos, para uma visualização completa do desempenho de cada modelo.

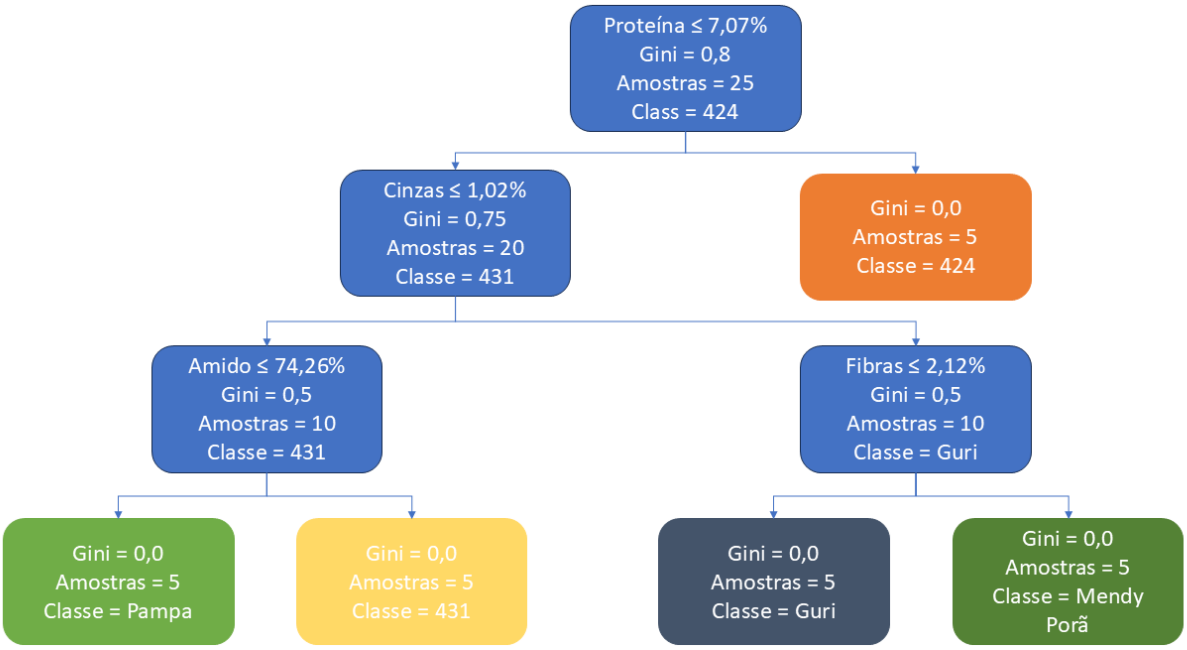
Tabela 1 – Métricas de acurácia dos algoritmos utilizados para classificação

Algoritmos	Métricas			
	Precisão	Recall	F1-Score	ROC Area
Random Forest	0,900	0,925	0,900	0,994
KNN	0,158	0,225	0,167	0,548
J48	0,700	0,725	0,700	0,825
MLP	0,071	0,200	0,101	0,445
Naive Bayes	0,692	0,775	0,717	0,906

Fonte: Elaborado pelos autores (2025).

A área sob a curva ROC média foi de 0,994 para Random Forest, 0,906 para Naive Bayes e 0,825 para J48, indicando boa separabilidade dos genótipos com esses três algoritmos. A árvore de decisão do J48 apresentada na Figura 1 demonstra a forte influência dos parâmetros de composição centesimal do NIRS na classificação de genótipos.

Figura 1 – Árvore de decisão do algoritmo J48



Fonte: Elaborado pelos autores (2025).

O primeiro nó utiliza o teor de proteína (≤ 7,07%), separando as amostras em dois grandes grupos. O grupo com teor inferior ou igual a 7,07% passa por divisões subsequentes com base nos teores de cinzas, amido e fibras. Esse ramo permite classificar corretamente os genótipos como “Pampa” e “431”, com pureza total (índice de Gini = 0) nos nós finais. Já o grupo com teor de proteína superior a 7,07% se divide diretamente em duas classes distintas: uma formada apenas por amostras do genótipo “424” e outra que, a partir do teor de fibras (≤ 2,12%), diferencia entre os genótipos “Guri” e “Mendy Porã”. A árvore apresenta uma boa capacidade discriminativa com nós terminais puros (Gini = 0), indicando que as variáveis utilizadas foram eficazes para separar os genótipos avaliados.

Assim, a aplicação da IA nessa classificação de genótipos demonstrou um valor significativo, posicionando-se como uma solução eficaz na redução de fraude alimentar. Isso corrobora com Zhang et al. (2024), que reforça a aplicação da IA em vários setores na produção de alimentos, incluindo também a indústria pesada, área da saúde e governança social. Contudo, é importante ressaltar que a aplicação da tecnologia de IA na indústria de alimentos tem um potencial significativo e pode gerar impactos profundos no futuro do setor (Lugo-Morin et al., 2024).

Conclusões

O uso do algoritmo Random Forest mostrou-se eficiente para a classificação de genótipos agrícolas com base em características físico-químicas. Além da alta acurácia obtida, a abordagem permitiu identificar variáveis-chave que podem ser utilizadas como marcadores indiretos em programas de melhoramento. Essa estratégia baseada em aprendizado supervisionado representa uma ferramenta promissora para análises rápidas e objetivas na seleção genotípica de arroz, podendo ser aplicada em diferentes culturas e condições experimentais.

Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES – Projeto 4732/UFPel), Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Unidade Embrapii InovaAgro pelo fornecimento de bolsas de pesquisas aos autores.

Referências

CLERCQ, Djavan de; MAHDI, Adam. Modern computational approaches for rice yield prediction: a systematic review of statistical and machine learning-based methods. **Computers And Electronics In Agriculture**, [S.L.], v. 231, p. 109852, abr. 2025. Elsevier BV. http://dx.doi.org/10.1016/j.compag.2024.109852.

HAO, Yong; GENG, Pei; WU, Wenhui; WEN, Qinhua; RAO, Min. Identification of Rice Varieties and Transgenic Characteristics Based on Near-Infrared Diffuse Reflectance Spectroscopy and Chemometrics. **Molecules**, [S.L.], v. 24, n. 24, p. 4568, 13 dez. 2019. MDPI AG. http://dx.doi.org/10.3390/molecules24244568.

LUGO-MORIN, Diosey Ramon. Artificial Intelligence on Food Vulnerability: future implications within a framework of opportunities and challenges. **Societies**, [S.L.], v. 14, n. 7, p. 106, 29 jun. 2024. MDPI AG. http://dx.doi.org/10.3390/soc14070106.

PANCHBHAI, Kamini G.; LANJEWAR, Madhusudan G.. Detection of amylose content in rice samples with spectral augmentation and advanced machine learning. **Journal Of Food Composition And Analysis**, [S.L.], v. 142, p. 107455, jun. 2025. Elsevier BV. http://dx.doi.org/10.1016/j.jfca.2025.107455.

ZENG, Fangye; ZHANG, Min; LAW, Chung Lim; LIN, Jiacong. Harnessing artificial intelligence for advancements in Rice / wheat functional food Research and Development. **Food Research International**, [S.L.], v. 209, p. 116306, maio 2025. Elsevier BV. http://dx.doi.org/10.1016/j.foodres.2025.116306.