CLASSIFICAÇÃO DE GENÓTIPOS DE ARROZ COM APRENDIZADO DE MÁQUINA PARA REDUÇÃO DE FRAUDES ALIMENTARES

GÜNTHER BLANK DA SILVA[1]; RUAN BERNARDY[2]; LARISSA ALVES RODRIGUES[3]; JULIANA PINO DE PAULA[4]; GUILHERME MACIEL MUNHOZ[5] MAURÍCIO DE OLIVEIRA[6]

[1]Universidade Federal de Pelotas – gunther.kath.blank@gmail.com
[2]Universidade Federal de Pelotas – ruanbernardy@yahoo.com.br
[3]Universidade Federal de Pelotas – larissaalvesrodrigues23@gmail.com
[4]Universidade Federal de Pelotas – jupino22@gmail.com
[5]Universidade Federal de Pelotas – guilhermemacielmunhoz@gmail.com
[6]Universidade Federal de Pelotas – mauricio@labgraos.com.br

1. INTRODUÇÃO

A Inteligência Artificial (IA) é uma tecnologia que revolucionou as atividades humanas. Diferentemente de quase tudo que já se viu, ela consegue armazenar, aprender, reproduzir e progredir no aprendizado com base em dados fornecidos (ZENG *et al*., 2025). Uma das técnicas que fazem parte da IA é o Machine Learning (ML), que traduzindo ao literal significa “aprendizado de máquina”. Esta ferramenta permite que as máquinas aprendam sozinhas, adquirindo conhecimento e encontrando padrões a partir dos dados brutos (CLERCQ; MAHDI, 2025).

Na agricultura e tecnologia de alimentos, a IA ganha cada vez mais espaço, visto que pode ser aplicada da pré a pós-colheita, especialmente nos processos industriais, tornando a cadeia produtiva cada vez mais eficiente e dinâmica (CUNHA *et al*., 2025). Neste contexto, diversos estudos alertam para fraudes alimentares no setor alimentício, especialmente na área de grãos (TEYE; AMUAH, 2022).

Produtos como mel, azeite e peixes são rotineiramente alvos de fraudes. Além disso, na área de grãos, o arroz (segundo alimento mais consumido no mundo) também é alvo de processos fraudulentos. No comércio global de arroz, é comum a rotulagem fraudulenta de variedades (ZHANG; XUE, 2016), especialmente do Basmati indiano e paquistanês (arroz de qualidade premium com base em seu aroma característico), que seguidamente é rotulado em variedades mais baratas (ŚLIWIńSKA-BARTEL, 2021).

Por isso, o objetivo do trabalho é utilizar as tecnologias de IA para classificação de genótipos de arroz visando o controle de qualidade e a fraude alimentar.

2. METODOLOGIA

O presente trabalho foi desenvolvido no Laboratório de Pós-Colheita, Industrialização e Qualidade de Grãos da Universidade Federal de Pelotas (LabGrãos), localizado no Campus Capão do Leão-RS.

Foram analisados 40 genótipos, oriundos da Estação Experimental do Arroz (EEA-IRGA) localizada em Cachoeirinha-RS. Foi realizada a caracterização dos genótipos, com quatro repetições, através da composição centesimal em espectroscopia por infravermelho próximo (NIRS), utilizando o equipamento DS2500 FOSS. Após, os dados obtidos foram pré-processados com padronização

do nome das colunas e filtragem, visando manter um padrão concreto na distribuição das classes e melhor desempenho dos algoritmos.

Para a realização da tarefa de classificação, foram utilizados diferentes algoritmos: Random Forest, KNN, J48, MLP e Naive Bayes, utilizando como base as variáveis do NIRS : proteína, óleo, fibras, cinzas e amido. Os algoritmos foram treinados e testados utilizando validação cruzada estratificada (*Stratifiedd K-Fold*), com K=4 folds. Foi empregada a codificação dos nomes dos genótipos com LabelEncoder e parâmetros de avaliação: precisão, recall, F1-score, AUC ROC e média de assertividade. Toda a metodologia foi implementada integralmente no ambiente *Google Colaboratory* (Colab), utilizando a linguagem computacional *Python*.

3. RESULTADOS E DISCUSSÃO

Para a avaliação de desempenho dos algoritmos, utilizou-se como base principal os dados obtidos na AUC ROC, desde que estivesse em equilíbrio com os demais parâmetros. O modelo Random Forest se mostrou mais promissor, em comparação aos demais, com AUC ROC (0,988). Naive Bayes e KNN apresentaram bons resultados AUC ROC (0,987) e (0,951), respectivamente. A Tabela 1 expõe todos os valores obtidos em cada métrica relacionada aos algoritmos para fim de comparação do desempenho.

Tabela 1 - Métricas de acurácia dos algoritmos utilizados para classificação

Algoritmos	**Métricas**				
	Precisão	**Recall**	**F1-score**	**AUC ROC**	**Assertividade**
Random Forest	0,783	0,821	0,778	0,988	0,843
KNN	0,607	0,692	0,621	0,951	0,741
J48	0,571	0,644	0,574	0,818	0,718
MLP	0,175	0,343	0,216	0,947	0,652
Naive Bayes	0,632	0,708	0,638	0,987	0,420

Fonte: Elaborado pelos autores (2025).

A Figura 1 mostra o início da árvore de decisão gerada pelo J48, que nesse trabalho é o único que possibilita essa visualização, além de ter alcançado AUC ROC de 0,818. Essa árvore apresenta o caminho percorrido pelo algoritmo para realizar a classificação correta das classes. Nesta é possível observar as variáveis que possuem maior influência no momento da classificação, sendo as primordiais: cinzas, óleo e fibras. É coerente esta afirmação visto que nos modelos de árvore as variáveis que aparecem nas divisões superiores possui maior influência sobre a classificação (ATA *et al*., 2023).

As setas na base da árvore apresentada demonstram a continuidade do processo de classificação, devido a profundidade de mais de quinze (15) camadas, gerada pelo número de classes (40 genótipos).

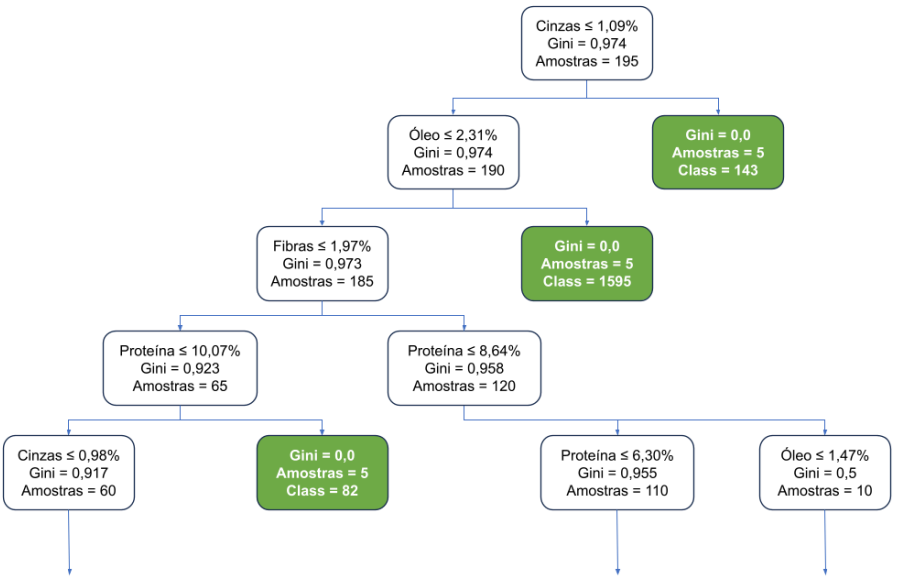

Figura 1 - Árvore de decisão do algoritmo J48
Fonte: Elaborado pelos autores (2025).

A primeira variável utilizada pelo algoritmo foi cinzas, onde valores que satisfazem a condição de menor que 1,09%, seguem para o próximo critério de avaliação, neste caso o teor de óleo. Contudo, se a condição não fosse verdadeira, a amostra já era classificada no genótipo 1595, onde o gini é de 0,0. O gini indica a certeza do modelo, onde valores próximos de zero indicam maior pureza da classificação, portanto o zero absoluto indica uma classificação confiável segundo o treinamento do algoritmo.

A utilização conjunta do NIRS e ML é relatada na literatura por Singh *et al.* (2024), onde os autores demonstraram bom desempenho empregando essas duas técnicas para a predição de conteúdo proteico em feijão e arroz.

Assim, a partir dos resultados desta pesquisa, infere-se que a classificação feita por algoritmos de ML em 40 genótipos de arroz é eficiente. Portanto, colabora para a diminuição das fraudes alimentares que ocorrem na indústria. É importante o aprofundamento nas pesquisas, para desenvolver novas tecnologias que visam o maior controle de qualidade no setor orizícola, elevando a confiabilidade dos consumidores e fortalecendo a segurança alimentar mundial.

4. CONCLUSÕES

O algoritmo Random Forest foi o mais eficiente na classificação de genótipos de arroz (com AUC ROC de 0,988 e Assertividade média de 0,843), baseando-se somente nas características químicas da composição centesimal em NIRS. Além disso, o percentual de Cinzas, Óleo e Fibras desempenharam maior influência durante o processo de classificação. Isso é fundamental para o avanço tecnológico na área agrícola e industrial, tornando a cadeia produtiva mais confiável para o consumidor.

5. REFERÊNCIAS BIBLIOGRÁFICAS

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. DOI: http://dx.doi.org/10.1016/j.foodres.2025.116306.

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. DOI: http://dx.doi.org/10.1016/j.compag.2024.109852.

CUNHA, Nicolau Brito da; FERNANDES, Fabiano Cavalcanti; GIL-LEY, Abel; FRANCO, Octavio L.; TIMAKONDU, Naagma; COSTA, Fabricio F.. Bridging BioSciences and technology: the impact of ai & genai in life sciences and agribusiness. **Gene**, [S.L.], v. 964, p. 149623, set. 2025. DOI: http://dx.doi.org/10.1016/j.gene.2025.149623.

ŚLIWIńSKA-BARTEL, Magdalena; BURNS, D. Thorburn; ELLIOTT, Christopher. Rice fraud a global problem: a review of analytical tools to detect species, country of origin and adulterations. **Trends In Food Science & Technology**, [S.L.], v. 116, p. 36-46, out. 2021. DOI: http://dx.doi.org/10.1016/j.tifs.2021.06.042.

ZHANG, Wenjing; XUE, Jianhong. Economically motivated food fraud and adulteration in China: an analysis based on 1553 media reports. **Food Control**, [S.L.], v. 67, p. 192-198, set. 2016. DOI: http://dx.doi.org/10.1016/j.foodcont.2016.03.004.

SINGH, Naseeb; KAUR, Simardeep; PHILANIM, W.s.; KUMAR, Amit; SHARMA, Paras; ANANTHAN, R.; BHARDWAJ, Rakesh. Integrating NIR spectroscopy with machine learning and heuristic algorithm-assisted wavelength selection algorithms for protein content prediction in rice bean (Vigna umbellata L.). **Food And Humanity**, [S.L.], v. 3, p. 100399, dez. 2024. DOI: http://dx.doi.org/10.1016/j.foohum.2024.100399.

TEYE, Ernest; AMUAH, Charles L.y.. Rice varietal integrity and adulteration fraud detection by chemometrical analysis of pocket-sized NIR spectra data. **Applied Food Research**, [S.L.], v. 2, n. 2, p. 100218, dez. 2022. DOI: http://dx.doi.org/10.1016/j.afres.2022.100218.