APLICAÇÃO DE APRENDIZADO DE MÁQUINA NA DIFERENCIAÇÃO DE GENÓTIPOS DE ARROZ PARA DETECÇÃO DE ADULTERAÇÕES

RUAN BERNARDY[1]; JANETE V. DA ROSA MONTEIRO[2]; SILVIA NAIANE JAPPE[3]; BRENDA DANNENBERG KASTER[4]; BETINA BUENO PERES[5]; MAURÍCIO DE OLIVEIRA[6]

[1]Universidade Federal de Pelotas – ruanbernardy@yahoo.com.br
[2]Universidade Federal de Pelotas – janete.monteiro.ppgcta@hotmail.com
[3]Universidade Federal de Pelotas – jappesilvia@gmail.com
[4]Universidade Federal de Pelotas – brendadannenbergkaster@gmail.com
[5]Universidade Federal de Pelotas – betinabuenop@gmail.com
[6]Universidade Federal de Pelotas – mauricio@labgraos.com.br

1. INTRODUÇÃO

A Inteligência Artificial (IA) tem se consolidado como uma das áreas mais promissoras da ciência da computação, buscando compreender e desenvolver sistemas capazes de reproduzir o raciocínio humano. Essas tecnologias possibilitam o armazenamento e a análise de grandes volumes de informações, permitindo que as máquinas evoluam continuamente por meio da interação com os dados (JING et al., 2025). Dentro desse campo, o Aprendizado de Máquina (*Machine Learning* – ML) se destaca por oferecer a capacidade de aprendizado autônomo, no qual os algoritmos extraem padrões e tomam decisões a partir das informações fornecidas, sem a necessidade de programação explícita (ZARBAKHSH *et al*., 2025).

Nos últimos anos, as aplicações de IA têm se expandido significativamente na agricultura e na indústria de alimentos, sendo utilizadas em equipamentos, no manejo de lavouras e em processos de industrialização. Apesar desses avanços, um dos maiores desafios do setor alimentício continua sendo o combate às fraudes, que geralmente envolvem a mistura de produtos de menor valor com aqueles de maior qualidade e preço. Um exemplo frequente é a adulteração em grãos é a adição de sabor artificial ao arroz na China, misturando arroz comum com arroz tailandês de jasmim, que é semelhante, comprometendo sua autenticidade e causando prejuízos econômicos (JU *et al*., 2021).

Dessa forma, distinguir os genótipos de arroz por meio de suas propriedades físico-químicas torna-se uma estratégia fundamental para identificar e prevenir adulterações (SHARMA *et al*., 2024). Nesse contexto, algoritmos de aprendizado de máquina, como o Random Forest, apresentam-se como ferramentas eficazes para a análise e classificação de dados complexos, oferecendo respostas rápidas e precisas.

Assim, a aplicação de métodos de IA e ML possibilita não apenas a correta classificação dos genótipos de arroz, mas também a seleção das variáveis mais determinantes nesse processo, aumentando a confiabilidade das análises. Para esse fim, destaca-se o uso da linguagem de programação *Python*, amplamente adotada na área de ciência de dados devido à sua simplicidade de uso e ao vasto conjunto de bibliotecas dedicadas ao aprendizado de máquina. Desta forma, o objetivo deste trabalho foi classificar genótipos de arroz através da técnica de machine learning, avaliando diferentes algoritmos desenvolvidos em *python*.

2. METODOLOGIA

O trabalho foi desenvolvido no Laboratório de Pós-Colheita, Industrialização e Qualidade de Grãos da Universidade Federal de Pelotas (LabGrãos/UFPel). Foram avaliados 40 genótipos de arroz provenientes da empresa Camil Alimentos S/A, do Instituto Rio Grandense do Arroz (IRGA) e da Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina (EPAGRI). Foram analisados composição centesimal pelo equipamento NIRS modelo FOSS DS2500L (teor de proteína, óleo, amido, cinzas e fibras), percentual de amilose, rendimento volumétrico e gravimétrico, e parâmetros de textura (dureza, coesividade, gumosidade e mastigabilidade), através de protocolos padrões para análises do LabGrãos.

Após a caracterização dos genótipos, os dados foram pré-processados para padronização dos nomes de colunas e aplicou-se o filtro Resample, para garantir a padronização dos valores. Foram realizadas cinco repetições por genótipo, com foco na uniformidade entre as classes, para ter representatividade de cada variedade.

Utilizou-se os algoritmos Random Forest, KNN, J48, MLP e Naive Bayes para classificar os genótipos com base nas variáveis preditoras analisadas. A técnica utilizada para o treinamento e teste dos algoritmos foi a validação cruzada estratificada com 4 folds (*StratifiedKFold*). O desempenho dos modelos foi realizado por meio das métricas: precisão, recall, F1-score, matriz de confusão e área sob a curva ROC (AUC), além da média da importância das variáveis. Todas as etapas da metodologia foram realizadas dentro do ambiente de programação Google Colaboratory (Colab), usando a linguagem Python.

3. RESULTADOS E DISCUSSÃO

Entre os modelos testados neste trabalho, o Random Forest apresentou o melhor resultado, com assertividade média de 0,939, além de elevada precisão (0,913) e recall (0,933). Em seguida, o Naive Bayes (0,807) e o J48 (0,647) mostraram desempenho satisfatório, mantendo equilíbrio entre precisão e recall. Já os algoritmos KNN (0,322) e MLP (0,206) tiveram resultados inferiores, indicando menor capacidade de classificação. As métricas variam de 0 a 1, onde valores próximos de 1 indicam alto desempenho do modelo e próximos de 0 representam baixa capacidade de classificação. A Tabela 1 apresenta os valores obtidos em cada métrica para comparação do desempenho entre os algoritmos.

Tabela 1 – Métricas de acurácia dos algoritmos utilizados para classificação

Algoritmos	**Métricas**				
	Precisão	**Recall**	**F1-Score**	**AUC ROC**	**Assertividade**
Random Forest	0,913	0,933	0,913	0,998	0,939
KNN	0,154	0,231	0,170	0,734	0,322
J48	0,561	0,638	0,574	0,814	0,647
MLP	0,052	0,074	0,055	0,644	0,206
Naive Bayes	0,740	0,776	0,733	0,981	0,807

Fonte: Elaborado pelos autores (2025).

A Figura 1 apresenta um recorte de uma das árvores de decisão gerada pelo algoritmo Random Forest, a qual ilustra os critérios utilizados para a classificação dos genótipos de arroz. Nesta visualização, é possível identificar as variáveis que exercem maior influência no processo de decisão, destacando-se óleo, proteína, gumosidade e cinzas, que surgem nas divisões superiores da árvore. A profundidade da árvore (total de 21 níveis) evidencia a complexidade da classificação, considerando o número de genótipos avaliados, enquanto os ramos refletem o encadeamento lógico aplicado pelo modelo para separar corretamente as amostras em seus respectivos grupos.

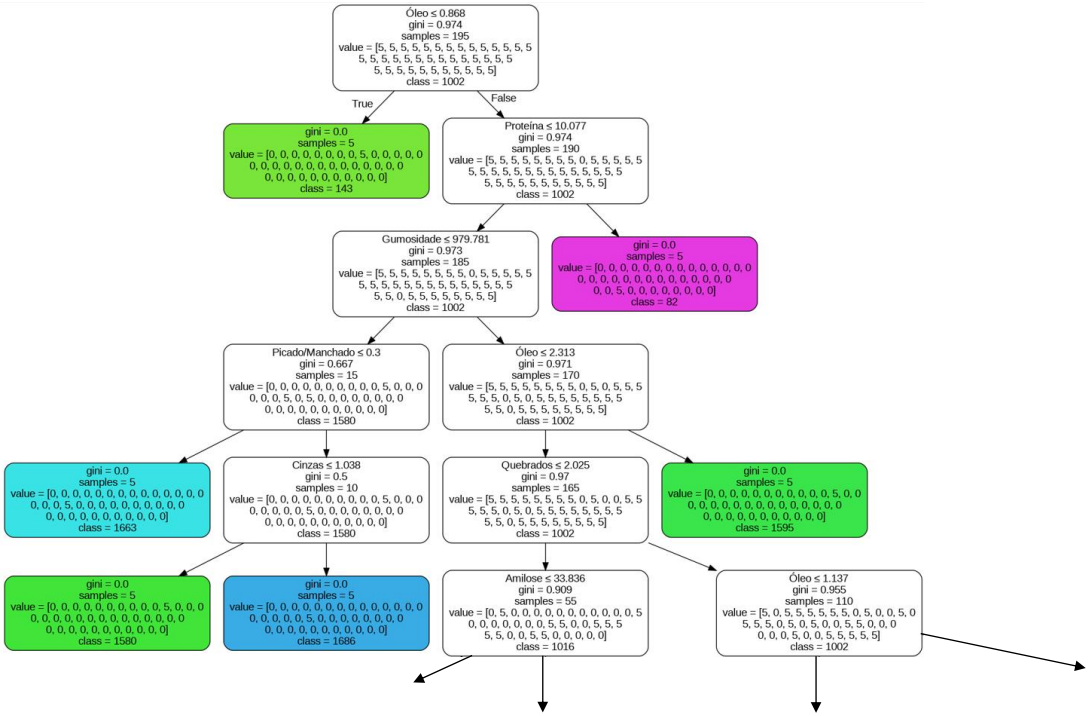

Figura 1 – Recorte de uma árvore de decisão do algoritmo Random Forest.
Fonte: Script do Google Colaboratory (2025).

A árvore de decisão apresentada evidencia que a variável inicial de separação foi o teor de óleo, com ponto de corte em 0,868. Essa condição direciona as amostras para distintos caminhos, permitindo a classificação de forma precisa. Em determinados nós, como nos que apresentam índice de Gini igual a 0, há total pureza na classificação, o que garante confiabilidade no modelo. Além disso, variáveis como proteína, gumosidade, picado/manchado e amilose se mostraram relevantes em níveis subsequentes de divisão, demonstrando a capacidade do algoritmo em explorar características físico-químicas e morfológicas para discriminar entre os diferentes genótipos.

A presença de múltiplas variáveis explicativas na árvore destaca a representatividade de atributos importantes para a diferenciação entre genótipos de arroz. O fato de características químicas, como óleo, proteínas e amilose, serem constantemente utilizadas pelo algoritmo confirma a robustez do banco de dados, que reúne informações consistentes e diversificadas. Essa riqueza de variáveis garante maior profundidade nas análises, ampliando o poder de classificação do modelo e assegurando que diferentes aspectos do grão sejam considerados no processo de decisão (GUO *et al*., 2023).

Diante disso, a aplicação de algoritmos de ML demonstra grande relevância no contexto da detecção de fraudes alimentares, especialmente em situações de rotulagem inadequada de arroz. A utilização de classificadores baseados em atributos físico-químicos e espectrais possibilita identificar adulterações e inconsistências com elevada precisão, fortalecendo os sistemas de rastreabilidade e assegurando maior confiabilidade ao consumidor (GUO *et al*., 2023). Dessa forma, o emprego de ML não apenas contribui para o controle de qualidade, mas também se consolida como uma ferramenta estratégica na promoção da segurança alimentar (BIAN *et al*., 2022).

4. CONCLUSÕES

O algoritmo Random Forest (0,939 e o Naive Bayes obtiveram os melhores desempenhos ao classificar 40 genótipos de arroz com diversas variáveis combinadas, com assertividade média de 0,939 e 0,807 respectivamente. Além disso, foi possível verificar as variáveis com maior influência na classificação, com destaque para o percentual de óleo, proteína bruta e cinzas, além do parâmetro de textura gumosidade, que surgem nas divisões superiores das árvores de decisão gerada pelo Random Forest. Isso desempenha um papel essencial no progresso tecnológico dos setores agrícola e industrial, reforçando a confiabilidade da cadeia produtiva para o consumidor.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BIAN, C.; SHI, H.; W., S.; ZHANG, K.; WEI, M.; ZHAO, Y.; SUN, Y.; ZHUANG, H.; ZHANG, X.; CHEN, S. Prediction of Field-Scale Wheat Yield Using Machine Learning Method and Multi-Spectral UAV Data. **Remote Sensing**, v.14, n.6, p.1474, 2022.

GUO, Y.; XIAO, Y.; HAO, F.; ZHANG, X.; CHEN, J.; BEURS, K. de; HE, Y.; FU, Y. H. Comparison of different machine learning algorithms for predicting maize grain yield using UAV-based hyperspectral images. **International Journal of Applied Earth Observation and Geoinformation**, v.124, e103528, 2023.

JING, G.; HE, Y.; WANG, M.; LIU, W.; RAN, K.; YU, J.; LI, W.; LIU, W. Rapid detection of fragrant rice adulteration by CD-IMS with machine learning. **Microchemical Journal**, v.209, e112740, 2025.

JU, X.; LIAN, F.; GE, H.; JIANG, Y.; ZHANG, Y.; XU, D. Identification of Rice Varieties and Adulteration Using Gas Chromatography-Ion Mobility Spectrometry. **Ieee Access**, v.9, p.18222-18234, 2021.

SHARMA, R.; NATH, P.C.; LODH, B.K.; MUKHERJEE, J.; MAHATA, N.; GOPIKRISHNA, K.; TIWARI, O.N.; BHUNIA, B. Rapid and sensitive approaches for detecting food fraud: a review on prospects and challenges. **Food Chemistry**, v. 454, e139817, 2024.

ZARBAKHSH, S.; FAKHRZAD, F.; RAJKOVIC, D.; NIEDBAŁA, G.; PIEKUTOWSKA, M. Approaches and challenges in machine learning for monitoring agricultural products and predicting plant physiological responses to biotic and abiotic stresses. **Current Plant Biology**, v.43, e100535, 2025.