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

AGRUPAMENTO DE GENÓTIPOS DE ARROZ COM BASE EM CARACTERÍSTICAS FÍSICO-QUÍMICAS, DE COCÇÃO E TEXTURAIS

Ruan Bernardy[1]; Lázaro da Costa Corrêa Cañizares[2]; Miquele Sodré Novak[3]; Larissa Alves Rodrigues[4]; Maurício de Oliveira[5]

Palavras-chave: Aprendizado não-supervisionado, Machine learning, Fraude alimentar, *Oryza sativa*; Inovação

Introdução

O arroz (*Oryza sativa* L.) é considerado o alimento básico de mais da metade da população mundial, devido à sua facilidade de preparação, e pelas propriedades nutricionais que trazem grandes benefícios a saúde (Gunaratne et al., 2013). Esse cereal é a terceira cultura mais cultivada no mundo, ocupando as lavouras em mais de 100 países (Dalbhagat et al., 2019). Existem diversas variedades de arroz produzidas ao redor do globo, devido a necessário de adaptar-se as condições ambientais de cada local. Além disso, o genótipo de arroz desempenha um papel importante na seleção do consumidor, pois influencia os parâmetros de qualidade desse grão, tais como adesividade, dureza, secura, viscosidade, aroma e brancura (Lapcharoensuk; Sirisomboon, 2014). Portanto, para garantir a autenticidade e integridade da variedade de arroz, evitando fraudes por mistura, tornou-se indispensável determinar o genótipo com precisão e rapidez (Teye; Amuah, 2022).

Nesse sentido, surgem novos estudos para detectar adulterações em amostras de arroz, destacando a importância de encontrar soluções para segregação genética, coibindo fraudes comerciais (Ganopoulos et al., 2011). Além da segurança e autenticidade, a segregação de genótipos também influencia diretamente na qualidade tecnológica e industrial do arroz (Śliwińska-Bartel et al., 2021). Assim, selecionar genótipos com base em seu desempenho sob diferentes condições de processamento permite padronizar a qualidade final e atender às exigências de consumidores e do mercado internacional.

O avanço das ferramentas de Inteligência Artificial e Aprendizado de Máquinas, tornando-se cada vez mais comum no campo de alimentos, ajuda a enfrentar desafios práticos e fornecer suporte à decisão (Zeng et al., 2025). Técnicas como a Análise de Componentes Principais (PCA) e os algoritmos de agrupamento (*clustering*), como o K-Means, são alternativas promissoras para identificar padrões e grupos entre os genótipos com base em múltiplas variáveis. Desta forma, o objetivo deste trabalho foi identificar padrões de similaridade entre genótipos de arroz por meio de análises multivariadas, visando o agrupamento daqueles com características comuns entre si.

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 20 genótipos de arroz oriundos da Estação Experimental do Arroz do IRGA, localizada em Cachoeirinha-RS.

[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]Agrônomo, Doutor em Ciência e Tecnologia de Alimentos, Universidade Federal de Pelotas, Campus Universitário s/n, 96160-000, Capão do Leão – RS, lazarocoosta@hotmail.com
[3]Agrônoma, Universidade Federal de Pelotas, Campus Universitário s/n, 96160-000, Capão do Leão – RS. miquele_novak@hotmail.com
[4]Agrônoma, Mestre em Ciência e Tecnologia de Alimentos, Universidade Federal de Pelotas, Campus Universitário s/n, 96160-000, Capão do Leão – RS, larissaalvesrodrigues23@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

O banco de dados construído contém variáveis como rendimento de inteiros e quebrados, teor de proteína, óleo, amido, fibras, tempo de cocção, percentual de amilose e parâmetros de textura (dureza, coesividade, gumosidade, mastigabilidade e resiliência). Inicialmente, foi realizado um pré-processamento dos dados para padronização das colunas e remoção de valores discrepantes. Para garantir equilíbrio nas análises, foram mantidas até 4 repetições por genótipo, removendo amostras muito distantes da média interna de cada grupo.

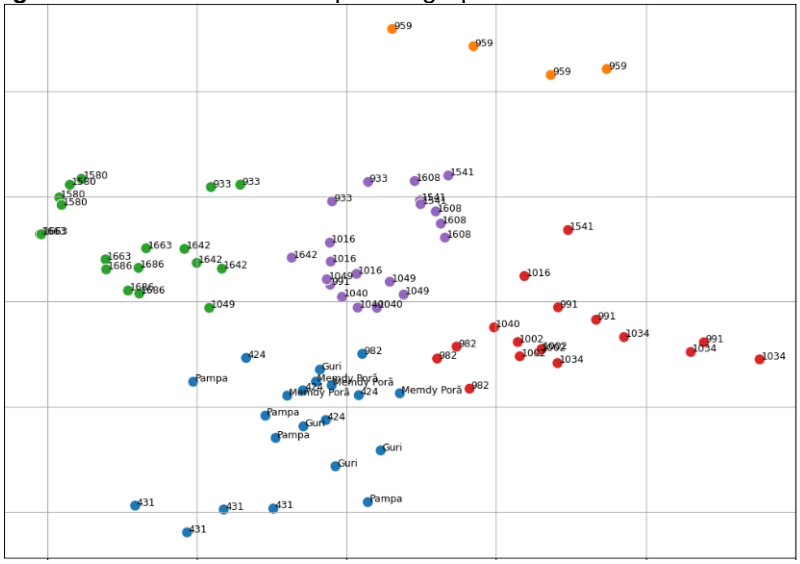
Aplicou-se Análise de Componentes Principais (PCA) para redução de dimensionalidade, retendo os 3 primeiros componentes principais que juntos explicam mais de 98% da variância total. Em seguida, realizou-se clusterização utilizando o algoritmo *K-Means*, com o número de *clusters* definido a partir da análise da variação interna entre os grupos, de modo a identificar o ponto ideal de separação, realizado de forma automática pelo próprio algoritmo. Os agrupamentos foram visualizados em gráfico de dispersão colorido e os genótipos associados a cada grupo foram listados. Todas as etapas dessa metodologia foram realizadas dentro do ambiente de programação Google Colaboratory (Colab), usando a linguagem *Python*, pois 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.

Resultados e Discussão

Conforme ressaltado por Raschka *et al*. (2020), a linguagem *Python* consolidou-se como a linguagem preferida para aplicações em ciência de dados e inteligência artificial, impulsionando a produtividade e facilitando a implementação de modelos complexos. No contexto da indústria alimentícia, Ding *et al*. (2023) demonstraram que técnicas de aprendizado de máquina, implementadas em *Python*, são eficazes na análise de grandes volumes de dados para prever tendências alimentares, aprimorar processos de produção e otimizar cadeias de suprimentos. Além disso, Arora *et al*. (2024) desenvolveram modelos de aprendizado de máquina e processamento de linguagem natural em *Python* para prever o grau de processamento de alimentos, utilizando perfis nutricionais e níveis de processamento NOVA, evidenciando a aplicabilidade da linguagem em tarefas complexas de classificação e análise de dados nutricionais.

Assim, a Figura 1 mostra os clusters criados de forma natural pelo algoritmo *K-Means*. A visualização dos clusters em gráfico 2D (com base nos dois primeiros componentes principais) revelou a formação de grupos coesos, indicando similaridade entre os perfis físico-químicos, de cocção e textura.

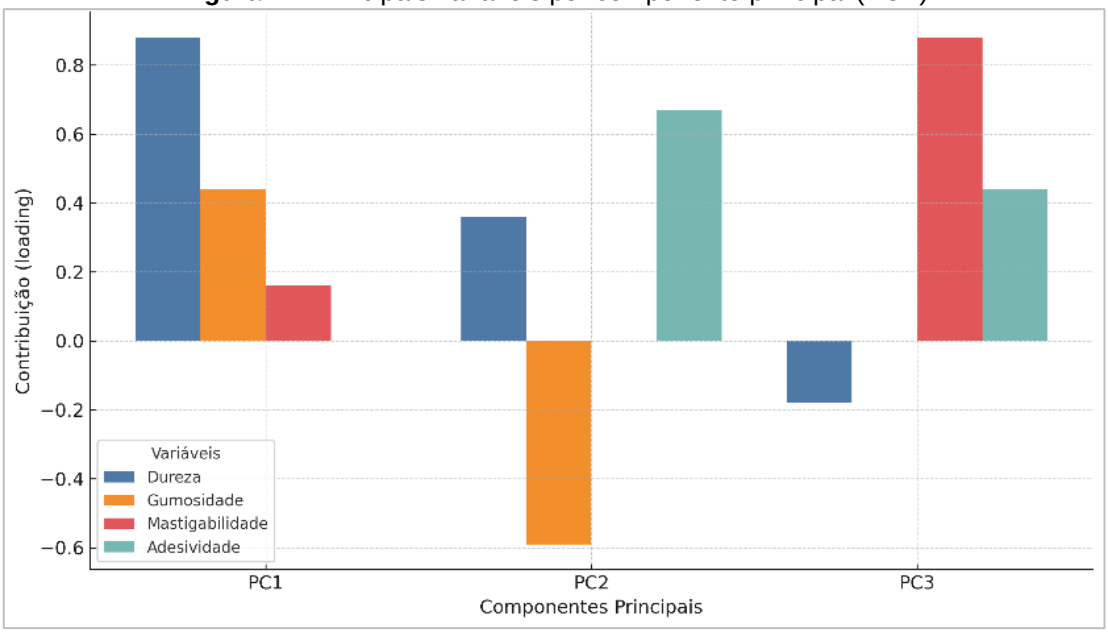
Figura 1 – Clusters criados após o agrupamento com K-Means e PCA



Fonte: Elaborado pelos autores, 2025.

A Figura 2 apresenta a contribuição de quatro variáveis – Dureza, Gumosidade, Mastigabilidade e Adesividade – nos três primeiros componentes (PC1, PC2 e PC3) resultantes de uma Análise de Componentes Principais (PCA). A PC1 é fortemente influenciada por Dureza e Gumosidade, indicando que essas variáveis têm maior peso na explicação da variabilidade inicial dos dados.

Figura 2 – Principais variáveis por componente principal (PCA)



Fonte: Elaborado pelos autores, 2025.

A PC2 mostra influência positiva de Adesividade e negativa de Gumosidade, sugerindo um contraste entre essas características. Já a PC3 é dominada por Mastigabilidade e Adesividade, o que indica que essas variáveis contribuem significativamente para a terceira dimensão de variação.

Desta forma, observa-se que a PC1 explica 97% da variância total dos dados, o que indica que quase toda a informação relevante do conjunto original está concentrada nessa dimensão. As variáveis com maior peso são Dureza (0,88), Gumosidade (0,44) e Mastigabilidade (0,16), demonstrando que essas características são as mais importantes para a variação entre as amostras. A PC2 possui 1,5% da variância explicada, enquanto a PC3 representa menos de 1% da variância. Assim, os resultados indicam que a análise pode ser simplificada praticamente à PC1, já que os demais componentes possuem contribuição estatisticamente irrelevante, sendo úteis apenas em contextos muito específicos ou exploratórios.

Isso demonstra a importância da análise de textura para o agrupamento desses genótipos, sendo coerente com o comportamento laboratorial observado. Por exemplo, genótipos com altos valores de dureza e baixo tempo de cocção tenderam a agrupar-se, sugerindo que essas variáveis desempenharam papel chave na diferenciação.

Conclusões

A aplicação da análise de agrupamento, em conjunto com PCA, permitiu identificar grupos de genótipos de arroz com perfis semelhantes sem a necessidade de rótulos prévios. A metodologia mostrou-se eficiente para revelar padrões naturais nos dados, auxiliando a seleção e caracterização de materiais. Estes resultados contribuem significativamente para a tomada de decisão na indústria, que por sua vez conseguirá com mais assertividade agrupar materiais genéticos que compartilham características, tornando o processo mais eficiente e estratégico. Isso reduz perdas decorrentes de decisões equivocadas, como a seleção de genótipos inadequados para determinados processos industriais.

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

ARORA, Nikhil; BHAGAT, Shreya; DHAMA, Ritu; BAGLER, Ganesh. Machine learning and natural language processing models to predict the extent of food processing. **arXiv preprint**, 2024. DOI: https://doi.org/10.48550/arXiv.2412.17217.

DALBHAGAT, Chandrakant Genu; MAHATO, Dipendra Kumar; MISHRA, Hari Niwas. Effect of extrusion processing on physicochemical, functional and nutritional characteristics of rice and rice-based products: a review. **Trends In Food Science & Technology**, v. 85, p. 226-240, mar. 2019. DOI: http://dx.doi.org/10.1016/j.tifs.2019.01.001.

DING, Haohan; TIAN, Jiawei; YU, Wei; WILSON, David I.; YOUNG, Brent R.; CUI, Xiaohui; XIN, Xing; WANG, Zhenyu; LI, Wei. The Application of Artificial Intelligence and Big Data in the Food Industry. **Foods**, v. 12, n. 24, p. 4511, 18 dez. 2023. DOI: http://dx.doi.org/10.3390/foods12244511.

GANOPOULOS, Ioannis; ARGIRIOU, Anagnostis; TSAFTARIS, Athanasios. Adulterations in Basmati rice detected quantitatively by combined use of microsatellite and fragrance typing with High Resolution Melting (HRM) analysis. **Food Chemistry**, v. 129, n. 2, p. 652-659, nov. 2011. Elsevier BV. http://dx.doi.org/10.1016/j.foodchem.2011.04.109.

GUNARATNE, A.; WU, K.; LI, D.; BENTOTA, A.; CORKE, H.; CAI, Y. Antioxidant activity and nutritional quality of traditional red-grained rice varieties containing proanthocyanidins. **Food Chemistry**. v. 138, p. 1153-1161, 2013

LAPCHAROENSUK, Ravipat; SIRISOMBOON, Panmanas. Eating quality of cooked rice determination using Fourier transform near infrared spectroscopy. **Journal Of Innovative Optical Health Sciences**, v. 07, n. 06, p. 1450003, 21 out. 2014. DOI: http://dx.doi.org/10.1142/s1793545814500035.

RASCHKA, Sebastian; PATTERSON, Josh; NOLET, Corey. Machine Learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. **arXiv preprint**, 2020. DOI: https://doi.org/10.48550/arXiv.2002.04803.

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**, v. 116, p. 36-46, out. 2021. DOI: http://dx.doi.org/10.1016/j.tifs.2021.06.042.

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**, v. 2, n. 2, p. 100218, dez. 2022. DOI: http://dx.doi.org/10.1016/j.afres.2022.100218.

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**, v. 209, p. 116306, maio 2025. DOI: http://dx.doi.org/10.1016/j.foodres.2025.116306.