Inteligência artificial e NIRS: Uma dupla poderosa para identificação de genótipos na pós-colheita

Ruan Bernardy[1]; Maria Antônia Fagundes de Leon[1]; Cesar Augusto Gaioso[1]; Lázaro da Costa Corrêa Cañizares[1]; Silvia Naiane Jappe[1]; Silvia Leticia Rivero Meza[1]; Maurício de Oliveira[1]

RESUMO

Com a crescente produção de soja no Brasil nos últimos anos, é fundamental o estudo aprofundado dos constituintes desse grão, aos quais podem variar de acordo com os genótipos e as condições de cultivo. Com isso, o objetivo desse estudo foi realizar a classificação de genótipos de soja, cultivados em diferentes ambientes e épocas de semeadura, utilizando como base a composição química obtida por infravermelho próximo (NIRs), época e ambiente de semeadura. Para isso, foi empregada a inteligência artificial e sua técnica de aprendizado de máquina. Foram utilizados 10 genótipos de soja, semeados em duas épocas de semeadura e em 7 cidades do Rio Grande do Sul. A composição química das amostras foi analisada através do equipamento FOSS NIRS DS2500, selecionando a banda entre 807 e 817nm. Os classificadores aplicados foram J48, Random Forest, CVR, IBk, MLP, utilizando o filtro Resample. Foi empregado o software Weka, versão 3.8.6, para mineração de dados. O algoritmo IBk conseguiu o melhor desempenho, alcançando 89% de classificação correta dos atributos. A partir da Matriz de Confusão, observou-se que todos os genótipos obtiveram resultados com pelo menos 60/70 valores preditos corretamente, destacando o bom desempenho dos classificadores. Nas métricas, o IBk obteve 0,89 de Precisão, Recall e F-Measure, e 0,94 de ROC Area. Foi possível classificar os genótipos, de acordo com a sua composição química relacionada aos dados obtidos na curva espectral, época e ambiente de semeadura, a partir da inteligência artificial e aprendizado de máquina.

Palavras-chave: *Glycine max* (L.); Aprendizado de Máquina; Agricultura 4.0

[1] Laboratório de Pós-Colheita, Industrialização e Qualidade de Grãos (LABGRÃOS), Faculdade de Agronomia Eliseu Maciel (FAEM), Universidade Federal de Pelotas (UFPEL).

INTRODUÇÃO

Nas últimas décadas, a agricultura no Brasil tornou-se destaque no cenário mundial, onde a soja é um dos principais produtos agrícolas cultivados no país, fixando-se como o maior produtor e exportador de soja do mundo (CONAB, 2022). Esse aumento da área semeada e da produção de soja deu-se devido ao aumento da demanda por esse grão, impulsionando os preços, além do alto investimento em tecnologias em todas as etapas da cadeia produtiva, melhorando as condições de cultivo e proporcionando produções recordes.

Nesse contexto, é fundamental o entendimento da composição química desses grãos, que podem variar entre os genótipos, para auxiliar a indústria no momento de compor os alimentos à base de soja. A composição do grão de soja é muito influenciada pelo genótipo e pelas condições de cultivo. Em média, o grão de soja é composto de 35 a 45% de proteína. A maior parte da proteína de soja (90%) é formada por globulinas, que podem ser solubilizadas em soluções salinas diluídas (LIU et al. 2008). A quantidade de globulinas influencia o rendimento de extração e a qualidade dos produtos alimentícios feitos com soja, como tofu e leite de soja (ZIEGLER et al., 2016). Outro componente relevante do grão de soja é o óleo, que representa cerca de 20%, sendo usado na preparação de alimentos em todo o mundo. O óleo de soja tem uma alta concentração de ácidos graxos poli-insaturados, que são mais propensos à oxidação e à digestão enzimática do que os ácidos graxos saturados (NIKOLIC et al., 2014). Nesse contexto, a aplicação de novas tecnologias pode interessar ao mercado da soja, pois permite resultados rápidos para a indústria (CAVALCANTE et al. 2023).

Nesse contexto, o uso de métodos não destrutivos, na avaliação da qualidade dos grãos, pode ser mais preciso com a inserção da tecnologia, possibilitando a reutilização de amostras em outros ensaios. As informações obtidas a partir de infravermelho próximo, aliado ao aprendizado de máquina, proporciona maior rapidez na interpretação dos dados, visto que há um grande volume de informações sobre os componentes de qualidade da semente (PINHEIRO et al. 2022).

Desta forma, o objetivo desse estudo foi avaliar a classificação de genótipos de soja, a partir da composição química obtida por infravermelho próximo, relacionada a época e ambiente de semeadura, a partir da inteligência artificial e sua técnica de aprendizado de máquina.

MATERIAL E MÉTODOS

A pesquisa foi realizada no Laboratório de Pós-Colheita, Industrialização e Qualidade de Grãos (LabGrãos) da Universidade Federal de Pelotas. Foram utilizados 10 genótipos de soja, em 1ª e 2ª época de semeadura, cultivados em 7 cidades do Rio Grande do Sul. Os Genótipos avaliados foram: BMX LANÇA IPRO; PONTA; VALENTE; 5909; 95R51; BMS 5601 RR; BMX DELTA IPRO; GARRA; DM 57I52RFS IPRO e BMX ZEUS IPRO. Os locais de semeadura foram: São Gabriel; Santo Augusto; Bagé; Tupanciretã; Vacaria; Passo Fundo e São Luiz Gonzaga.

A composição química das amostras foi realizada através do equipamento FOSS NIRS DS2500, realizando a seleção da banda entre 807 e 817nm. Foi analisado o teor de proteína, lipídios, fibras, cinzas e amido dos grãos.

Para a classificação dos genótipos, inicialmente foi necessário a realização de pré-processamento dos dados, de modo a preparar o conjunto, para que a ferramenta realizasse a correta leitura e análise. Essa etapa foi baseada no trabalho de Bernardy et al. (2023).

Os classificadores aplicados foram J48, Random Forest, CVR, IBk, MLP. Foi utilizada a validação cruzada para o aprendizado dos algoritmos, dividindo o conjunto de dados, treinamento e teste, em 10 subconjuntos (10 folds). Com isso, os dados foram subdivididos em dez partes, utilizando nove para treino e uma para teste, repetindo esse processo por dez vezes (folds), sempre alterando as partes utilizadas para treino e teste. Essa técnica reduz a probabilidade de que coincidências subavaliem ou sobreavaliem o desempenho, para uma determinada configuração (BERNARDY et al., 2023). A média dessas precisões correspondeu ao desempenho do algoritmo sobre o conjunto de dados fornecido. Para certificar a precisão dos algoritmos foram utilizadas as seguintes métricas de avaliação: Acurácia, Precisão, Recall, F-measure e Área ROC, de acordo com Lever et al. (2016).

Para a tarefa de mineração de dados, utilizando os métodos de aprendizagem de máquina, foi utilizado o software Weka, versão 3.8.6. Após o pré-processamento dos dados, totalizaram 700 linhas para análise dos algoritmos, com 70 linhas para cada genótipo.

Os dados de análises são desbalanceados por natureza. Com intuito de resolver esse problema e não tendenciar o algoritmo, ou melhorar seu desempenho, utilizou-se o filtro Resample, de instância não supervisionada, que mantém a distribuição das classes na subamostra, onde, alternativamente, pode ser configurado para enviesar a distribuição dos atributos para uma distribuição uniforme (GADOTTI et al., 2022a,b). A amostragem pode ser realizada com (padrão) ou sem reposição (WITTEN et al., 2011). A Figura 1 ilustra e resume a metodologia aplicada nesse trabalho.

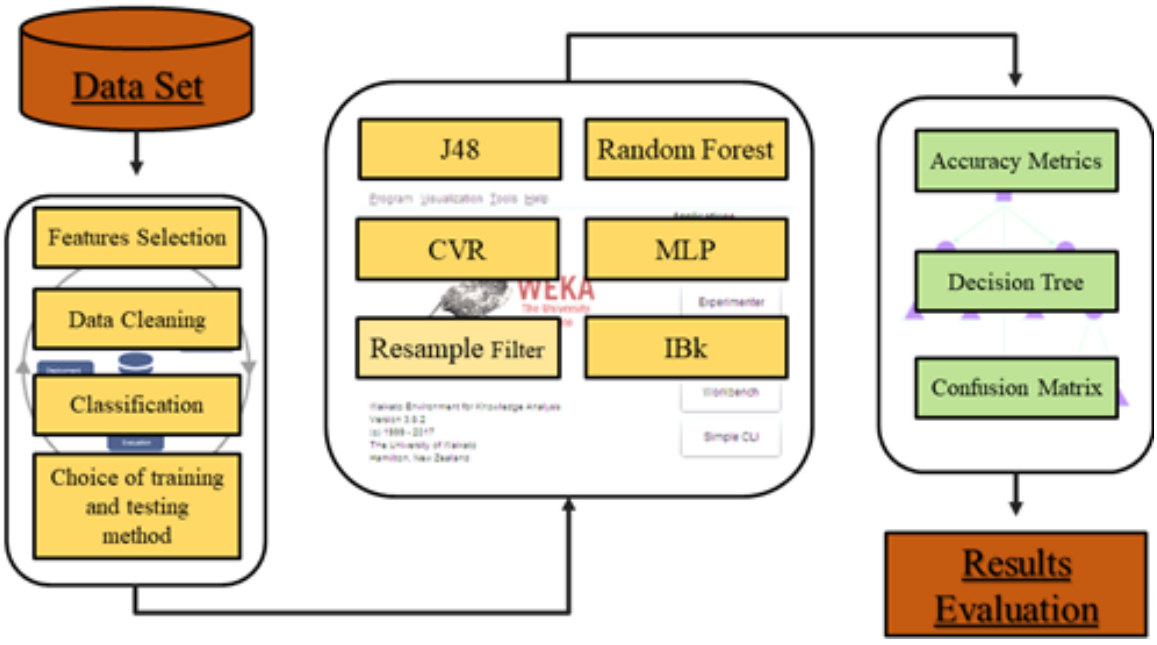

Figura 1. Metodologia aplicada na classificação de genótipos de soja.

RESULTADOS E DISCUSSÃO

A composição química dos grãos de soja determina sua utilização, ou seja, cultivares ricas em óleo são indicadas para a indústria de óleos vegetais ou produção de biocombustíveis. Por outro lado, quando os grãos são destinados à alimentação humana, altos teores de proteína e açúcar são desejáveis (JIANG et al., 2018).

A partir da Tabela 1, é possível observar que o algoritmo IBk conseguiu melhor desempenho para essa tarefa de classificação dos genótipos, alcançando 89,43% de classificação correta dos atributos.

Tabela 2. Acurácia dos algoritmos após a classificação dos genótipos de soja.

Algoritmo	Classificação Correta dos Atributos (%)
J48	80,28
Random Forest	87,86
CVR	80,29
IBk	89,43
MLP	86,29

Visto que, o algoritmo IBk possuindo maior desempenho e ajuste ao banco de dados proposto, foi extraído a Matriz de Confusão e as métricas de acurácia, para conhecer melhor seu desempenho frente as classes, de modo individual. A Matriz de Confusão se torna importante para avaliar melhor os erros e acertos dos classificadores empregados, conseguindo escolher uma técnica que possua classificação mais realista. Essa matriz é expressada em função das classes utilizadas, exibindo a distribuição dos dados de acordo com as classes reais e preditas pelo algoritmo, para comparar se o dado de uma determinada categoria foi classificado corretamente ou entendido como sendo de outra, pelo modelo computacional proposto (IBM, 2021). A matriz do algoritmo IBk pode ser analisada na Tabela 2.

Tabela 2. Matriz de confusão do algoritmo IBk para classificação de genótipos de soja

		Predição									
		Lança	**Ponta**	**Valente**	**5909**	**95R51**	**5601**	**Delta**	**Garra**	**57I52**	**Zeus**
Real	**Lança**	**63**	1	1	1	0	1	2	0	1	0
	Ponta	2	**61**	1	0	0	0	0	1	3	2
	Valente	0	0	**65**	1	3	0	0	1	0	0
	5909	2	0	1	**59**	3	0	0	4	0	1
	95R51	0	1	2	3	**64**	0	0	0	0	0
	5601	1	1	0	0	1	**66**	1	0	0	0
	Delta	2	0	0	0	0	0	**60**	2	2	4
	Garra	0	1	2	1	2	0	0	**64**	0	0
	57I52	0	2	0	0	0	0	2	2	**61**	3
	Zeus	0	1	0	2	1	1	0	1	1	**63**

A partir da Tabela 2, analisando a diagonal da matriz, onde encontram-se os valores classificados corretamente, observa-se que todos os genótipos obtiveram resultados com pelo menos 60/70 valores preditos corretamente, demonstrando que os classificadores conseguiram encontrar o padrão relacionado especificamente para cada cultivar.

A partir disso, foi analisado as métricas de acurácia. Em trabalho realizado por Gadotti et al. (2022a), os autores afirmam que a F-Measure é calculado através dos valores médios de recall e precisão. Já a ROC Area (*Receiver Operator Characteristic)* apresenta a relação entre a sensibilidade e a especificidade do classificador, ou seja, quanto maior o valor, mais ajustado é a curva. A Tabela 3 apresenta as médias de cada métrica, para melhor visualização e avaliação dos resultados.

Tabela 3. Acurácias dos diferentes algoritmos utilizados, sendo Recall (sensibilidade), Precisão, Curva ROC (*Receiver Operating Characteristic*) e F Measure

Classificador	**Média das Métricas de Acurácia**			
	Precisão	**Recall**	**F-Measure**	**ROC Area**
J48	0,807	0,803	0,803	0,928
Random Forest	0,880	0,879	0,879	0,987
CVR	0,804	0,803	0,803	0,962
IBk	0,895	0,894	0,894	0,942
MLP	0,863	0,863	0,862	0,957

As médias de cada métrica de avaliação são superiores a 0,80 em todos os classificadores (Tabela 3). Contudo, a ROC Area demonstra resultado superior no Random Forest. Esse utiliza uma série de árvores decisórias para classificar o lote em questão, montando uma “floresta”, para ao final efetuar uma espécie de eleição entre os resultados que mais aconteceram, mostrando a resposta que mais ocorreu quando se realizou a predição do atributo.

Além disso, pode-se observar valores próximos a 0,9 em ROC Area e entre 0,8 e 0,89 nas demais métricas, alcançando resultados satisfatórios, pois, em aprendizado de máquina, respostas próximas ou iguais a 1,00 significa super ajustamento do modelo com os dados, ou seja, o algoritmo pode estar “viciado” no conjunto de dados em questão, não conseguindo predizer outro diferente.

O algoritmo J48 (0,928 de ROC Area), em sua árvore de decisão, escolheu os valores da curva, em 810nm, como principal parâmetro para iniciar a predição dos dados. Posteriormente, utilizou a composição química (fibras, proteína e amido) para as tomadas de decisão seguintes. Isso demonstra ser possível utilizar os valores da curva espectral para predizer a composição química dos grãos. Além disso, a época de semeadura e ambiente, em um primeiro momento, não foram selecionados pelo algoritmo. No entanto, a composição e a qualidade do produto podem ser influenciadas por fatores como o genótipo, data de semeadura, fertilidade do solo, condições ambientais de cultivo e as etapas de pós-colheita, como secagem, armazenamento e industrialização (LEE e LEE 2009; ZIEGLER et al., 2018). Desta forma, é fundamental aprofundar essa temática em trabalhos futuros.

Em trabalho realizado por Soares et al. (2018), os autores destacam que a vantagem principal da mineração de dados é a capacidade de oferecer soluções para questões complexas em qualquer área de conhecimento, pois, se assemelha ao raciocínio lógico, ajudando no processo de classificação. Com isso, essa ferramenta é fundamental para auxiliar em tomadas de decisão que demandam tempo, devido à complexidade e grande quantidade de dados a serem processados.

Como conclusão, foi possível classificar os genótipos, de acordo com a sua composição química relacionada aos dados obtidos na curva espectral, época e ambiente de semeadura, a partir do emprego da inteligência artificial e sua técnica de aprendizado de máquina. Os valores da curva, na faixa entre 807 e 817nm, foram importantes para iniciar a tomada de decisão pelo algoritmo J48, seguido pela composição química, ambiente e época de plantio. O algoritmo IBk apresentou os melhores resultados, sendo mais indicado para estudos futuros em análise com genótipos de soja.

REFERÊNCIAS BIBLIOGRÁFICAS

BERNARDY, R.; GADOTTI, G. I.; MONTEIRO, R. C. M.; PINTO, K. V. A.; PINHEIRO, R. M. Fitting data mining settings for ranking seed lots. **Engenharia Agrícola**, v. 43, n. 2, 2023.

CONAB. COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento de safra brasileiro – grãos:** sexto levantamento, março de 2022, safra 2021/2022. Brasília:

Companhia Nacional de Abastecimento. 2022. Disponível em: <https://www.conab.gov.br/info-agro/safras>. Acesso em: 14 mar. 2022

GADOTTI, G. I.; ASCOLI, C. A.; BERNARDY, R.; MONTEIRO, R. C. M.; PINHEIRO, R. M. Machine learning for soybean seeds lots classification. **Revista Engenharia Agrícola**, v. 42, special issue, 2022a.

GADOTTI, G. I.; MORAES, N. A. B.; SILVA, J. G.; PINHEIRO, R. M.; MONTEIRO, R. C. M. Prediction of ranking of lots of corn seeds by artificial intelligence. **Revista Engenharia Agrícola**, v. 42, n. 4, 2022b.

IBM. **Visualização da Matriz de Confusão**. 2021. Disponível em: https://www.ibm.com/docs/en/db2/9.7?topic=visualizer-confusion-matrix-view. Acesso em: 23 ago. 2023.

JIANG, G.L.; CHEN, P.; ZHANG, J.; FLOREZ-PALACIOS, L.; ZENG, A.; WANG, X.; BOWEN, R.A. ; MILLER, A.; BERRY, H. Genetic analysis of sugar composition and its relationship with protein, oil, and fiber in soybean. **Crop Science**, v.58, p. 2413-2421, 2018.

LEE, S.; LEE, J. Effects of oven-drying, roasting, e explosive puffing process on isoflavone distributions in soybeans. **Food Chemistry**, v. 112, n. 2, p. 316-320, 2009.

LEVER, J.; KRZYWINSKI, M.; ALTMAN, N. Classification evaluation. **Nature Methods**, v. 13, n. 8, p. 603-604, 2016.

LIU, C.; WANG, X.; MA, H.; ZHANG, Z.; GAO, W.; XIAO, L. Functional properties of protein isotardias from soybeans stored under various conditions. **Food Chemistry**, v. 111, p. 29-37, 2008.

NIKOLIC, Z.; VASILJEVIC, I.; ZDJELAR, G.; DORDEVIC, V.; IGNJATOV, M.; JOVICIC, D.; MILOŠEVIC, D. Detection of genetically modified soybean in crude soybean oil. **Food Chemistry**, v. 145, p. 1072-1075, 2014.

PINHEIRO, R. M.; GADOTTI, G. I.; BERNARDY, R.; MONTEIRO, R. C. M.; MOREIRA, I. B. Processamento de imagens como ferramenta importante para inteligência artificial no setor de sementes. **Revista Agrária Acadêmica**, v. 5, p. 89-101, 2022.

SOARES, E. A. de M. G.; DAMASCENA, L. C. L. de; LIMA, L. M. M.; MORAES, R. M. de. Analysis of the Fuzzy Unordered Rule Induction Algorithm as a Method for Classification. In: CONGRESSO BRASILEIRO DE SISTEMAS FUZZY, 5., 2018, Fortaleza. **Recentes Avanços em Sistemas Fuzzy.** Fortaleza: Sbmac, 2018. p. 17-28.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining**: pratical machine learning tools and techniques. 3. ed. USA: Morgan Kaufmann Publishers, 2011. 665 p.

ZIEGLER, V.; FERREIRA, C. D.; HOFFMAN, J. F.; OLIVEIRA, M.; ELIAS, M. C. Effects of moisture e temperature during grain storage on the functional properties e isoflavone 519 profile of soy protein concentrate. **Food Chemistry**, v. 242, p. 37-44, 2018.

ZIEGLER, V.; FERREIRA, C.D.; VANIER, N.L.; DOS SANTOS, M.A.Z.; DE OLIVEIRA, M.; ELIAS, M.C. Physicochemical e technological properties of soybean as a function of storage conditions. **Brazilian Journal of Food Research**, v. 7, n.3, 2016.