eISSN 1981-1829

Ciência e Agrotecnologia
Science and Agrotechnology

Identifying soybean genotypes with artificial intelligence and near infrared reflectance

Identificação de genótipos de soja com o uso inteligência artificial e reflectância por infravermelho próximo

Ruan Bernardy[1*], Lázaro da Costa Corrêa Cañizares[1], Silvia Leticia Rivero Meza[1], Larissa Alves Rodrigues[1], Silvia Naiane Jappe[1], Maurício de Oliveira[1]

ABSTRACT

With the increasing soybean production in Brazil, and the demand for soybeans with high protein and oil content, it is essential to conduct an in-depth study of the constituents of this grain, which can vary according to genotypes and growing conditions. Therefore, the objective of this study was to classify soybean genotypes, cultivated in different environments and sowing seasons, according to their chemical composition and the spectrum generated by near-infrared spectroscopy (NIRS). For this purpose, artificial intelligence and its machine learning technique were employed. 10 soybean genotypes were used, sown in two sowing seasons and cultivated 7 cities in Rio Grande do Sul. The chemical composition of the samples was analyzed using the FOSS NIRS DS2500 equipment, selecting the band between 807 and 817 nm. The applied algorithms were J48, Random Forest, CVR, IBk, MLP, using the Resample filter. The Weka software, version 3.8.6, was employed for data mining. The IBk algorithm achieved the best performance, reaching 89% correct classification of attributes. From the Confusion Matrix, it was observed that all genotypes obtained results above 60/70 for correctly predicted values, highlighting the algorithms' good performance. In the metrics, IBk achieved 0.89 Precision, Recall, and F-Measure, and 0.94 ROC Area. Thus, it was possible to classify the genotypes according to their chemical composition related to the data obtained in the spectral curve, sowing season, and environment, using artificial intelligence and machine learning.

Index terms: *Glycine max* (L.); machine learning; agricultural technology; agriculture 4.0.

RESUMO

Com a crescente produção de soja no Brasil e a demanda por grãos de soja com alto teor de proteína e óleo, é fundamental o estudo aprofundado dos constituintes desse grão, os 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, de acordo com a composição química e o espectro gerado por infravermelho próximo (NIRs). 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 algoritmos 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 superiores a 60/70 para os valores preditos corretamente, destacando o bom desempenho dos algoritmos. 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.

Termos para indexação: *Glycine max* (L.); aprendizado de máquina; tecnologia agroindustrial; agricultura 4.0

Introduction

Brazil has consolidated itself as the world's largest producer and exporter of soybeans. This increase in the cultivated area and production of soybeans in Brazil has occurred due to the rising demand for these grains, which has driven up prices and investment in technologies across all stages of the soybean production chain (Companhia Nacional de Abastecimento -Conab, 2024). The increase in demand for soybeans is attributed to their high protein content (35-45%) and oil content (18-22%) (Liu et al., 2008; Ziegler et al., 2016).

The protein and oil content of soybeans are complex quantitative traits controlled by many genes and influenced by environmental and cultivation factors (Duan et al., 2023). A major challenge for breeders is to increase protein content without affecting oil content. However, protein content has a negative

Agricultural Sciences

Ciênc. Agrotec., 48:e005224, 2024
http://dx.doi.org/10.1590/1413-7054202448005224

Editor: Renato Paiva

1Universidade Federal de Pelotas, Faculdade de Agronomia Eliseu Maciel, Capão do Leão, RS, Brasil

*Corresponding author: ruanbernardy@yahoo.com.br

Received in March 14, 2024 and approved in June 6, 2024

correlation with oil content (Rodrigues et al., 2013; Kambhampati et al., 2020), making it difficult to select genotypes with high levels of both protein and oil. In this context, studying the chemical composition of genotypes cultivated in Brazil is essential to assist breeders in the process of developing new genotypes.

The selection, identification, and analysis of the chemical composition of different soybean genotypes in a database is a time-consuming and destructive task. The application of new technologies for segregating genotypes according to their composition becomes necessary for the industry (Santana et al., 2023).

The use of near-infrared spectroscopy (NIRS) technology in assessing grain composition, combined with machine learning, provides faster data interpretation, as there is a large volume of information to be processed (Pinheiro et al., 2022). Machine learning (ML) techniques are an approach that has been successfully applied in classifying complex datasets. This technique, belonging to the field of artificial intelligence, can discover patterns in a database, learning and improving results (Singh et al., 2016; Ramos et al., 2020; Van Dijk et al., 2021).

The algorithms used in machine learning (artificial neural networks (ANNs), decision tree models, and random forests) can be employed to create models that classify the data of interest (Teodoro et al., 2024), significantly improving accuracy and reducing time for data analysis compared to traditional methods (Ramos et al., 2020; Teodoro et al., 2021; Batista et al., 2022).

In a study conducted by Schwalbert et al. (2020), machine learning models applied to remote sensing data were used to predict soybean productivity, where the artificial neural network algorithm achieved a prediction with a mean absolute error (MAE) of 0.42 mg.ha^{-1} 70 days before harvest. Similarly, Ramos et al. (2020) obtained satisfactory results in predicting maize productivity by combining different cultivation parameters with machine learning techniques. However, the use of near infrared combined with machine learning to classify soybean genotypes is still relatively unexplored. Some studies have already tried to show important results using these two techniques (Santana et al., 2023), but more effort is still needed from researchers.

Therefore, the objective of this study was to evaluate and classify soybean genotypes, sown at different times and grown in distinct environments, according to their chemical composition and the spectrum obtained by near-infrared spectroscopy, using artificial intelligence and its machine learning techniques.

Material and Methods

Experimental Design

The research was conducted at the Grain Post-Harvest, Processing, and Quality Laboratory (LabGrãos) of the Federal University of Pelotas. Ten soybean genotypes were used, cultivated in two sowing seasons: a standard (October 25th) and a late sowing date (November 15th), across 7 cities in Rio Grande do Sul: São Gabriel, Santo Augusto, Bagé, Tupanciretã, Vacaria, Passo Fundo, and São Luiz Gonzaga. The evaluated genotypes were: BMX LANÇA IPRO; PONTA; VALENTE; 5909; 95R51; BMS 5601 RR; BMX DELTA IPRO; GARRA; DM 57I52RFS IPRO, and BMX ZEUS IPRO. The samples were provided by the Pró-Sementes Foundation for Research Support (Passo Fundo – RS – Brazil).

The study was conducted using a randomized block design with four replications (biological replicates). Eight rows were used per block, with a spacing of 40 cm between rows, and within the rows, the spacing between plants was 10 cm (equivalent to 10 plants per linear meter), totaling a sowing density of 250,000 plants per hectare. For analysis, only the four central rows of each treatment were collected. Soil management before sowing was identical in all treatments, as well as all agrochemical applications during cultivation, in order to standardize the treatments. Disease and pest control were rigorously conducted, with no differences observed between treatments due to management during cultivation. The harvested soybeans were fully homogenized to compose the biological replicates.

The chemical composition of the samples was analyzed using NIRS equipment, model DS2500 (FOSS Analytical, Denmark). The protein, lipid, fiber, ash and starch contents of the whole grains were analyzed. For each reading, 200 grams of grain were used, and the analyses were carried out in triplicate. In each replicate, the sample was mixed again to make the choice of soybeans random.

To avoid noise, data from the spectral curve in the 807~817nm range was used. The range of wavelengths used for the analysis avoided wavelengths susceptible to interference from water and chlorophyll (França-Silva et al., 2022).

Data preprocessing

For genotype classification, data preprocessing was initially performed to prepare the dataset for correct reading and analysis (Bernardy et al., 2023). The classifiers applied and tested were J48, Random Forest, Classification Via Regression (CVR), Instance Based k (IBk), and Multilayer Perceptron (MLP).

The J48 algorithm aims to build a decision tree based on the training data set, which is easily interpreted by anyone (Costa, Bernardini & Viterbo Filho, 2014). It is considered the most popular algorithm in the Weka software, being an open-source Java implementation of the C4.5 algorithm, dividing a complex problem into classes of sub-problems, applying this strategy repeatedly (Witten, Frank & Hall, 2011). In this way, it produces a decision tree, showing the path followed to classify the proposed data.

Along these lines, and also derived from the C4.5 algorithm, the Random Forest method was defined in 1995 by Tim Kam Ho in his work "Random Decision Forests". The author proposed this algorithm due to the limitations found in decision trees, such as the fact that very complex trees suffer from the phenomenon

of overfitting, when a statistical model fits a set of data perfectly but later proves to be totally ineffective in predicting new results (Schlenger, 2024). Random Forest creates hundreds of independent decision trees at random, where each tree will be used to choose the final result, making it more robust than J48. However, this is not a rule and both algorithms should always be analyzed.

On the other hand, the CVR algorithm works by estimating the probability of an instance belonging to a given class, using a linear approach. This approach is used to simplify the logistic regression model, which makes the algorithm more efficient. It can be applied to solve binary (two-class) or multi-class classification problems.

The IBk algorithm, on the other hand, classifies data by calculating the distance between each training instance and the new data, classifying it from the nearest instance, i.e. another training data next to it, in order to decide which class the new data belongs to. It is a non-parametric algorithm, making no assumptions about the distribution of the data. This makes the algorithm flexible and capable of working with a wide variety of data.

Finally, MLP uses a non-linear computational process and is highly efficient for classifying and regressing complex data (Chen & Wang, 2020; Hesami et al., 2020). It was created to solve classification problems using hidden neurons structured in layers, which in turn process the information obtained from the previous layers and send the knowledge generated to the layers in front, in order to arrive at an answer to the problem (Hecht-Nielsen, 1990). It works in much the same way as human neurons.

Cross-validation, using the k-fold technique, was employed for algorithm training, dividing, training, and testing the dataset into 10 subsets (10 folds). This technique reduces the likelihood of overfitting and underfitting of the model. The average of these accuracies corresponded to the algorithm's performance on the provided dataset. To ensure the accuracy of the algorithms, the following evaluation metrics were used: Accuracy, Precision, Recall, F-measure, and ROC Area, according to Lever Krzywinski and Altman (2016).

Data processing and mining

Data mining was performed using machine learning techniques. The software Weka, version 3.9.6, was utilized, running on an NVIDIA GeForce MX250 processor, integrated with an Intel® Core™ i5-10210U CPU running at 2.11 GHz, with 8GB of RAM. After data preprocessing, there were a total of 700 rows for algorithm analysis, with 70 rows for each genotype.

It aims to use machine learning techniques based on algorithms from various established approaches, and the greatest benefit of using this tool is the range of algorithms available (Silva, 2018). Weka uses the Java language, which has the greatest advantage of portability and can be implemented on different operating systems, as well as being free and open-source software (Silva, 2018).

Due to the imbalanced nature of the data, the Resample filter was initially used to avoid biasing the algorithm and improving its performance. Unsupervised analysis was used, maintaining the distribution of classes in the subsample (Gadotti et al., 2022 a,b). Sampling can be performed with or without replacement (Witten, Frank, & Hall, 2011). Figure 1 illustrates and summarizes the methodology applied in this work.

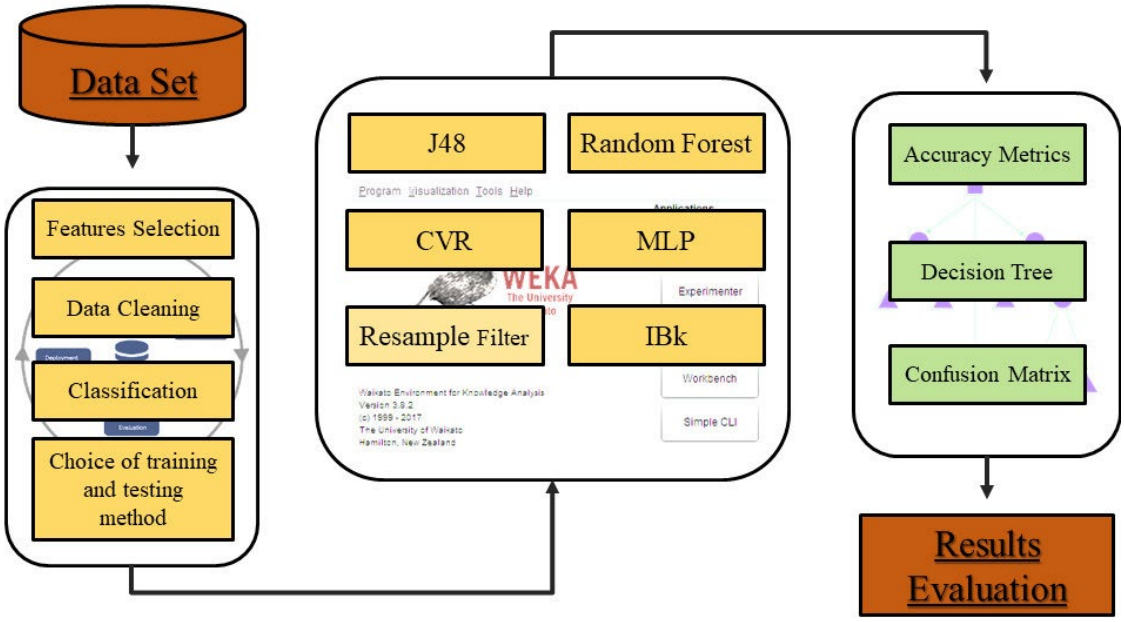

Figure 1: Methodology applied in soybean genotype classification.

Results and Discussion

The protein and oil content present in soybean grains are important parameters for the processing industry. Higher oil content is desirable for the vegetable oil industry or biofuel production. On the other hand, higher protein content is prioritized for human consumption (Jiang et al., 2018).

For the soluble protein analysis, considering the growing season, the highest value was found in the BMX GARRA IPRO genotype (77.86%) and the lowest in the NA 5909 RG genotype (48.34%), sown at a late date. The BMX ATIVA RR genotype behaved similarly to the aforementioned genotype in terms of quality parameters. No differences were observed when sown late and there was a greater impact on the characteristics associated with grain defects as storage time increased.

The sowing date affected the quality parameters: higher values of soluble protein and germination, and lower values of electrical conductivity, acidity and lipase activity were observed with delayed sowing. In general, the sowing date did not affect the quality parameters of soybeans under the conditions analyzed.

When analyzing the growing environment, the highest soluble protein results were found in the BMX ALVO RR genotype (68.82%) grown in Bagé, the 95r51 genotype (68.51%) grown in Cachoeira do Sul, the BMX ALVO RR genotype (67.23%) grown in Vacaria and the DM 57I52RSF IPRO genotypes (69, 96%) and BRS 5601 RR (69.95%) grown in São Luiz Gonzaga, for the other genotypes no significant differences were found, except for genotype 95r51 (52.93%), which obtained the lowest soluble protein result when grown in São Luiz Gonzaga.

In this regard, the use of ML models makes it possible to obtain more information about the crop from spectral variables, such as the identification of characteristics related to the cycle (Santana et al., 2022), prediction of nitrogen content and plant height (Osco et al., 2020) and association between characteristics related to productivity (Santana et al., 2021; Santana et al., 2022). This makes decision-making in genotype selection more agile and accurate, contributing to the industrial process of soy-based products (Gao, Guan, & Ma, 2022).

The performance of each algorithm used is presented in Table 1. The IBk algorithm showed the best performance in classifying the studied genotypes (89.43% correct classification of attributes), demonstrating a better fit to the proposed study data.

Table 1: Accuracy of the algorithms after the classification of soybean genotypes.

Algorithm	Correct Classification of Attributes (%)
J48	80.28
Random Forest	87.86
CVR	80.29
IBk	89.43
MLP	86.29

Classification is one of the functionalities of algorithms in machine learning. However, each algorithm has its own performance on a dataset, requiring testing of different models (Karakatič & Podgorelec, 2016). Oliveira et al. (2021), in a study conducted with eucalyptus species, achieved good accuracy in using ML through decision tree models (RNAs). Additionally, RNAs are also widely employed in various fields of knowledge, especially in agriculture for prediction, classification (Beucher, Møller, & Greve, 2019), and identification of diseases in soybean seeds (Singh et al., 2021).

Based on the results presented in Table 1, the Confusion Matrix of the IBk algorithm was constructed to detail its performance against the classes individually. The Confusion Matrix is important for evaluating the errors and successes of the employed classifiers, aiming to choose a technique that provides a realistic classification without causing overfitting or underfitting. This matrix is expressed in terms of the classes used, displaying the distribution of data according to the actual classes and those predicted by the algorithm, aiming to compare whether data from a particular category was classified correctly by the proposed computational model (IBM, 2021). The IBk algorithm's matrix can be analyzed in Table 2.

As described above, Table 2 shows the confusion matrix after classifying the genotypes using the IBk algorithm. This matrix shows all the data classified correctly and incorrectly, for example: the Lance genotype was identified and classified correctly in 63 of the 70 data points (Real and Prediction), but was confused with other genotypes throughout the matrix. The correct data will always be on the diagonal of the matrix, with the class name Real corresponding to Prediction. The matrix is used to calculate all the accuracy metrics.

Analyzing the diagonal of the matrix, where the correctly classified values are found, it is observed that all genotypes obtained results exceeding 60/70 correctly predicted values, demonstrating that the classifiers were able to identify the pattern specifically related to each cultivar.

In a study conducted by Santana et al. (2023), protein and oil concentrations showed a positive correlation. However, the characteristics present in soybean grains can vary depending on genetic factors and the environment in which they are cultivated, especially during the filling period (Pípolo & Mandarino, 2016).

In a study by Jiang et al. (2018), high and negative phenotypic and genotypic correlations were found between soybean fiber and oil, suggesting that changes in oil content during soybean breeding also lead to an increase in fiber content. The observed correlations between ash and oil (0.709) and between ash and fiber (-0.850) indicate the important nutritional value of the grain, which aids in the improvement of varieties and breeding programs (Azam et al., 2021). The grain's oil content is a quantitative trait influenced by genetic factors and the environmental conditions in which the genotype is introduced (Turquetti-Moraes et al., 2022).

Based on this, the accuracy metrics were analyzed (Table 3). In a study conducted by Gadotti et al. (2022a), the authors state that the F-Measure is calculated through the average values of recall and precision. The ROC Area (Receiver Operator Characteristic) presents the relationship between the classifier's sensitivity and specificity, meaning the higher the value, the more adjusted the curve is.

The accuracy of all algorithms showed values close to 0.9 in ROC Area and between 0.8 and 0.89 in the other metrics, achieving satisfactory results. In machine learning, responses close to or equal to 1.00 indicate overfitting of the model to the data (overfitting in the training data), corresponding to underfitting in the test data (Bernardy et al., 2023).

The ROC Area demonstrates a superior result in the Random Forest classifier. This classifier utilizes a series of decision trees to classify the dataset in question, assembling a "forest" to ultimately select the most occurring outcomes, indicating the response that happened most frequently when predicting the attribute. Considering the good accuracy results presented by the Random Forest algorithm (Table 3) in this study, the Confusion Matrix of this algorithm was conducted (Table 4).

The performance of this classifier was slightly inferior to IBk. However, it can be observed that the results on the diagonal presented more than 57 correctly predicted values, making it another model that can be applied in the segregation of soybean genotypes based on the analyzed characteristics.

In a study conducted by Alves et al. (2019), the authors demonstrated that decision trees or similar algorithms can extract quick and accurate information about crop health from data obtained from drones (Alves et al., 2019). They also efficiently detect diseases in rice plants with a precision of 97% (Rumy et al., 2021). When used for classification, it can achieve better accuracy and efficiency in data processing (Pandey & Prabhakar, 2016).

However, the composition and quality of soybean grains can be influenced by factors such as genotype, sowing date, soil fertility, environmental growing conditions, and post-harvest stages (drying, storage, and processing) (Cañizares et al., 2023; Ziegler et al., 2018). Therefore, it is essential to delve into this topic in future studies.

In a study conducted by Santana et al. (2023), the J48 algorithm was applied to classify soybean genotypes regarding grain yield, using the spectrum generated by NIRs as the database. This algorithm showed high efficiency and can be used to select genotypes for high grain yield in soybean breeding programs.

Table 2: Confusion matrix of the IBk algorithm for soybean genotype classification.

		Prediction									
		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

Table 3: Accuracies of the different algorithms used, including Recall (sensitivity), Precision, ROC Curve (Receiver Operating Characteristic), and F Measure.

Algorithms	Mean Accuracy Metrics			
	Precision	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

Table 4: Confusion Matrix of the Random Forest algorithm for soybean genotype classification.

		Prediction									
		Lança	Ponta	Valente	5909	95R51	5601	Delta	Garra	57I52	Zeus
Real	Lança	63	1	0	0	0	1	5	0	0	0
	Ponta	1	61	0	0	0	0	1	2	3	2
	Valente	0	0	67	0	2	0	0	1	0	0
	5909	1	0	0	62	3	1	0	2	0	1
	95R51	0	0	4	1	65	0	0	0	0	0
	5601	1	2	1	0	2	61	1	0	1	1
	Delta	3	0	0	0	0	3	57	0	5	2
	Garra	3	1	3	1	1	0	1	60	0	0
	57I52	2	0	0	0	0	0	3	2	57	6
	Zeus	1	0	0	0	0	3	1	2	1	62

The J48 algorithm (ROC Area of 0.928), in its decision tree, selected the values of the spectral curve (810nm) as the main parameter to initiate data prediction. Subsequently, it used the chemical composition (fiber, protein, and starch) for the following decision-making processes. This demonstrates that it is possible to use the values of the spectral curve to predict the chemical composition of the grains. Furthermore, the sowing season and cultivation environment were not selected by the algorithm. This indicates that data obtained from spectral reflectance response can be used for segregating soybean genotypes.

Therefore, machine learning algorithms can be used to overcome the problem of variable nonlinearity that commonly occurs between industrial and spectral variables. These algorithms process this data more rigorously, effectively overcoming the nonlinearities between variables. The integration of computational intelligence techniques with new technologies and field data allows for more reliable information for a variety of research objectives, such as outcomes (Schwalbert et al., 2020).

The main advantage of data mining is its ability to provide solutions to complex issues in any area of knowledge, making it a fundamental tool to assist in decision-making that requires time due to the complexity and large amount of data to be processed (Zhang et al., 2022).

Conclusions

It was possible to classify the genotypes through the chemical composition related to the data obtained in the spectral curve, planting season, and environment, using artificial intelligence and its machine learning technique. The curve values, in the range between 807 and 817nm, were important to initiate the decision-making process by the J48 algorithm, followed by chemical composition, environment, and planting season. The IBk algorithm yielded the best results, being more suitable for future studies in soybean genotype analysis.

Author Contribution

Conceptual idea: Oliveira, M.; Cañizares, L.C.C.; Bernardy, R.; Methodology design: Meza, S.L.R.; Data collection: Cañizares, L.C.C.; Rodrigues, L.A.; Bernardy, R.; Data analysis and interpretation: Cañizares, L.C.C.; Bernardy, R.; Meza, S.L.R.; Rodrigues, L.A.; and Writing and editing: Bernardy, R.; Cañizares, L.C.C.; Meza, S.L.R.; Rodrigues, L.A.; Jappe, S.N; Oliveira, M.

Acknowledgments

We want to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Fundação Pró Sementes de Apoio a Pesquisa (Pró-Sementes). This study was financed in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance code 001, Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) – Finance code 21/2551-0002255-8, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Finance codes 315822/2021-0.5.

References

Alves, T. M. et al. (2019). Optimizing band selection for spectral detection of *Aphis glycines* Matsumura in soybean. *Pest Management Science*, *75*(4):942-949.

Azam, M. et al. (2021). Profiling and associations of seed nutritional characteristics in Chinese and USA soybean cultivars. *Journal of Food Composition and Analysis*, *98*:103803.

Batista, T. S. et al. (2022). Artificial neural networks and non-linear regression for quantifying the wood volume in eucalyptus species. *Southern Forests: A Journal of Forest Science*, *84*(1):1-7.

Bernardy, R. et al. (2023). Fitting data mining settings for ranking seed lots. *Engenharia Agrícola*, *43*(2):e20220193.

Beucher, A., Møller, A. B., & Greve, M. H. (2019). Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark. *Geoderma*, *352*:351-359.

Cañizares, L. C. C. et al. (2023). Isoflavone profile identification and storage stability of different soybean genotypes sown at standard and late dates in a subtropical climate. *Biocatalysis And Agricultural Biotechnology*, *51*:e102739.

Chen, J. C., & Wang, Y. M. (2020). Comparing activation functions in modeling shoreline variation using multilayer perceptron neural network. *Water*, *12*(5):1281.

Companhia Nacional de Abastecimento - CONAB. (2024). *Acompanhamento de safra brasileiro - grãos: Sétimo levantamento, abril de 2024, safra 2023/2024*. Brasília: Companhia Nacional de Abastecimento, 120p.

Costa, J. J., Bernardini, F. C., & Viterbo Filho, J. (2014). A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. *AtoZ: novas práticas em informação e conhecimento*, *3*(2):139-157.

Duan, Z. et al. (2023). Genetic regulatory networks of soybean seed size, oil and protein contents. *Frontiers In Plant Science*, 14:1160418.

França-Silva, F. et al. (2022). Quantification of chlorophyll fluorescence in soybean seeds by multispectral images and their relationship with physiological potential. *Journal of Seed Science*, *44*:e202244023.

Gadotti, G. I. et al. (2022a). Machine learning for soybean seeds lots classification. *Engenharia Agrícola*, *42*:e20210101.

Gadotti, G. I. et al. (2022b). Prediction of ranking of lots of corn seeds by artificial intelligence. *Engenharia Agrícola*, *42*(4):e20210005.

Gao, S., Guan, H., & Ma, X. (2022). A recognition method of multispectral images of soybean canopies based on neural network. *Ecological Informatics*, *68*:101538.

Hecht-Nielsen, R. (1990). *Neurocomputing*. Massachusetts, United States: Addison-Wesley, 433p.

Hesami, M. et al. (2020). Development of support vector machine-based model and comparative analysis with artificial neural network for modeling the plant tissue culture procedures: effect of plant growth regulators on somatic embryogenesis of chrysanthemum, as a case study. *Plant Methods*, *16*:112.

International Business Machines Corporation - IBM. (2021). *Visualização da matriz de confusão*. IBM Documentation Help, Brasil: 1p. Available in: <https://www.ibm.com/docs/pt-br/db2/10.5?topic=visualizer-confusion-matrix-view>.

Jiang, G. L. et al. (2018). Genetic analysis of sugar composition and its relationship with protein, oil, and fiber in soybean. *Crop Science*, *58*:2413-2421.

Kambhampati, S. (2020). On the inverse correlation of protein and oil: Examining the effects of altered central carbon metabolism on seed composition using soybean fast neutron mutants. *Metabolites*, *10*(1)18.

Karakatič, S., & Podgorelec, V. (2016). Improved classification with allocation method and multiple classifiers. *Information Fusion*, *31*:26-42.

Lever, J., Krzywinski, M., & Altman, N. (2016). Classification evaluation. *Nature Methods*, *13*(8):603-604.

Liu, C. et al. (2008). Functional properties of protein isotardias from soybeans stored under various conditions. *Food Chemistry*, *111*:29-37.

Osco, L. P. et al. (2020). Leaf nitrogen concentration and plant height prediction for maize using uav-based multispectral imagery and machine learning techniques. *Remote Sens*, *12*:3237.

Oliveira, B. R. de et al. (2021). Eucalyptus growth recognition using machine learning methods and spectral variables. *Forest Ecology and Management, 497*:119496.

Pandey, P., & Prabhakar, R. (2016). An analysis of machine learning techniques (J48 & AdaBoost)-for classification. *India International Conference on Information Processing*, p.1-6.

Pinheiro, R. M. et al. (2022). Processamento de imagens como ferramenta importante para inteligência artificial no setor de sementes. *Revista Agrária Acadêmica*, *5*:89-101.

Pípolo, A. E., & Mandarino, J. M. G. (2016). Os teores de proteína da soja e a qualidade para a indústria. *Sociedade Brasileira de Ciência do Solo*, *42*(2):31-33.

Ramos, A. P. et al. (2020). A random forest ranking approach to predict yield in maize with UAV-based vegetation spectral indices. *Computers and Electronics in Agriculture*, *178*:105791.

Rodrigues, J. I. da S. et al. (2013). Associação de marcadores microssatélites com teores de óleo e proteína em soja. *Pesquisa Agropecuária Brasileira*, *48*(3):255-262.

Rumy, S. M. S. H. et al. (2021). An IoT based system with edge intelligence for rice leaf disease detection using machine learning. *IEEE International IOT, Electronics and Mechatronics Conference*, p.1-6.

Santana, D. C. et al. (2023). Classification of soybean genotypes for industrial traits using UAV multispectral imagery and machine learning. *Remote Sensing Applications: Society and Environment*, *29*:e100919.

Santana, D. C. et al. (2021). UAV-based multispectral sensor to measure variations in corn as a function of nitrogen topdressing. Remote Sensing Applications: *Society and Environment*, *23*:100534.

Santana, D. C. et al. (2022). High-throughput phenotyping allows the selection of soybean genotypes for earliness and high grain yield. *Plant Methods*, *18*:13.

Schlenger, J. (2024). Random forest. In D. Memmert. *Computer science in sport*. Springer, Berlin: Heidelberg, (pp.201-207).

Schwalbert, R. A. et al. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology*, *284*:107886.

Silva, R. (2018). *Introdução ao Weka: Software para mineração de dados*. Software para mineração de dados. Bagé, Brasil: Linkedin, 1p.

Singh, A. et al. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science*, *21*(2):110-124.

Singh, P. et al. (2021). Development of an intelligent laser biospeckle system for early detection and classification of soybean seeds infected with seed-borne fungal pathogen (*Colletotrichum truncatum*). *Biosystems Engineering*, *212*:442-457.

Teodoro, L. P. R. et al. (2024). Machine learning for classification of soybean populations for industrial technological variables based on agronomic traits. *Euphytica*, *220*:40.

Teodoro, P. E. et al. (2021). Predicting days to maturity, plant height, and grain yield in soybean: A machine and deep learning approach using multispectral data. *Remote Sens*, *13*(22):4632.

Turquetti-Moraes, D. K. et al. (2022). Integrating omics approaches to discover and prioritize candidate genes involved in oil biosynthesis in soybean. *Gene*, *808*:145976.

Van Dijk, A. D. J. et al. (2021). Machine learning in plant science and plant breeding. *iScience*, *24*(1):101890.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Burlington, United States: Morgan Kaufmann Publishers, 665p.

Zhang, Q. et al. (2022). E-commerce information system management based on data mining and neural network algorithms. *Computacional Intelligence and Neuroscience, 499801*:11.

Ziegler, V. et al. (2018). Effects of moisture e temperature during grain storage on the functional properties e isoflavone 519 profile of soy protein concentrate. *Food Chemistry*, *242*:37-44.

Ziegler, V. (2016). Physicochemical e technological properties of soybean as a function of storage conditions. *Brazilian Journal of Food Research*, *7*(3):117-132.