Euphytica (2024) 220:40
https://doi.org/10.1007/s10681-024-03301-w

RESEARCH

Machine learning for classification of soybean populations for industrial technological variables based on agronomic traits

Larissa Pereira Ribeiro Teodoro · Maik Oliveira Silva · Regimar Garcia dos Santos · Júlia Ferreira de Alcântara · Paulo Carteri Coradi · Bárbara Biduski · Carlos Antonio da Silva Junior · Francisco Eduardo Torres · Paulo Eduardo Teodoro

Received: 4 May 2023 / Accepted: 21 January 2024 / Published online: 22 February 2024

Abstract A current challenge of genetic breeding programs is to increase grain yield and protein content and at least maintain oil content. However, evaluations of industrial traits are time and cost-consuming. Thus, achieving accurate models for classifying genotypes with better industrial technological performance based on easier and faster to measure traits, such as agronomic ones, is of paramount importance for soybean breeding programs. The objective was to classify groups of soybean genotypes to industrial technological variables based on agronomic traits measured in the field using machine learning (ML) techniques. Field experiments were carried out in two sites in a randomized block design with two replications and 206 F_2 soybean populations. Agronomic traits evaluated were: days to maturation (DM), first pod height (FPH), plant height (PH), number of branches (NB), main stem diameter (SD), mass of one hundred grains (MHG), and grain yield (GY). Industrial technological variables evaluated were oil yield, crude protein, crude fiber, and ash contents, determined by high-optical accuracy near-infrared spectroscopy (NIRS). The models tested were: support vector machine (SVM), artificial neural network (ANN), decision tree models J48 and REPTree, random forest (RF), and logistic regression (LR, used as control). A genotype clustering was performed using PCA and k-means algorithm, and then the clusters formed were used as output variables of the ML models, while the agronomic traits were used as input variables. ML techniques provided accurate models to classify soybean genotypes for more complex variables (industrial technological) based on agronomic traits. RF outperformed the other models and can be used to contribute to soybean breeding programs by classifying genotypes for industrial technological traits.

Keywords *Glycine max* (L.) Merril · Fiber · Oil · Protein · Random forest

L. P. R. Teodoro · M. O. Silva · J. F. de Alcântara · P. E. Teodoro (✉)
Federal University of Mato Grosso Do Sul (UFMS), Chapadão Do Sul, MS 79560-000, Brazil
e-mail: eduteodoro@hotmail.com

R. G. dos Santos
Department of Agronomy, State University of São Paulo (UNESP), Ilha Solteira, SP 15385-000, Brazil

P. C. Coradi
Department of Agricultural Engineering, Federal University of Santa Maria, Cachoeira Do Sul, RS 96503-205, Brazil

B. Biduski
Department of Food Science and Technology, University of Passo Fundo, Passo Fundo, RS 99052-900, Brazil

C. A. da Silva Junior
Department of Geography, State University of Mato Grosso (UNEMAT), Sinop, MT 78550-000, Brazil

F. E. Torres
State University of Mato Grosso Do Sul (UEMS), Aquidauana, MS 79200-000, Brazil

Introduction

Soybean [*Glycine max* (L.) Merril] is the most economically important oilseed in the world. Due to its high seed oil content and high protein content with balanced amino acid composition, soybean is an important alternative food in animal and human nutrition (Alaswad et al. 2021). The main use of soybean, both in Brazil and worldwide, is as raw material for the industry, producing meal and oil. Meal, which is rich in protein, is used mainly in the feed industry for poultry, pigs, and cattle (Goldsmith 2008). On the other hand, oil is used both in the food and biodiesel industries. About 78% of biodiesel produced in Brazil comes from soybean oil (Ramos et al. 2017), being the most used raw material in biodiesel production in the country given its availability of large-scale cultivation (André Cremonez et al. 2015).

Oil and protein contents in soybean are complex quantitative traits controlled by multiple genes and affected by environmental factors (Burton 1985). A current major challenge is increasing yield and protein content and at least maintaining oil content. However, there is a negative relationship between grain yield and protein and oil contents (Bandillo et al. 2015; Cober and Voldeng 2000; Kambhampati et al. 2020; Pipolo et al. 2015), which hinders the selection of genotypes combining good performance for yield and industrial technological variables such as oil and protein contents. Additionally, measuring such industrial variables is a time-consuming and destructive task, requiring grain sample collection and laboratory analysis with the development of new equipment and methodologies, such as NIRS, it has allowed fast, accurate and non-destructive evaluations. In this scenario, obtaining accurate models for classifying genotypes for industrial technological variables based on easier and faster to measure traits, such as agronomic ones, is of paramount importance for soybean breeding programs.

One approach that has been successfully employed in regression and classification problems on complex datasets is using machine learning (ML) techniques. ML is a subgroup of the artificial intelligence area in which algorithms can learn from data and then discover patterns in the dataset, deciding on new and similar information (Marques Ramos et al. 2020; Singh et al. 2016; van Dijk et al. 2021). In this sense, algorithms such as artificial neural network (ANN), support vector machine (SVM), decision tree models, and random forest (RF) can be used to build models that allow the classification of the data of interest. Several studies (Batista et al. 2022; Marques Ramos et al. 2020; Teodoro et al. 2021) have reported meaningful improvements in the accuracy of estimates when ML models are implemented compared to traditional methods. Schwalbert et al., (2020) used ML models applied to remote sensing data for soybean yield prediction, in which ANN outperformed other algorithms. Marques Ramos et al., (2020), when using ML techniques combined with different vegetation indices, achieved satisfactory results in predicting maize yield, with RF algorithm standing out. Fletcher and Reddy, (2016), archived accurate classification models using RF algorithm with leaf multispectral data to differentiate three soybean varieties from two pigweeds. Zhou et al., (2020) used SVM algorithm to classify soybean leaf wilting due to drought stress by UAV-based imagery. However, studies on soybean genotype classification for industrial technological variables are scarce. To the best of our knowledge, there are still no studies classifying soybean genotypes for oil and protein contents based on agronomic traits using ML techniques.

Our hypothesis is that it is possible to classify soybean genotypes with better performance for oil yield, and crude protein, crude fiber and ash contents, whose measurement is expensive in terms of time and financial resources, using information from variables that are easier to measure and can be collected in the field, such as the agronomic ones. The objective was to identify the best ML technique to classify groups of soybean F_2 populations clustered by their performance for industrial technological variables using agronomic traits as input variables in the models.

Material and methods

Conducting the experiments

Field experiments were carried out during the 2019/2020 crop season at two sites. The first experiment was carried out at the Universidade Federal de Mato Grosso do Sul, campus of Chapadão do Sul, State of Mato Grosso do Sul (MS), Brazil (located at 18°46 "South, 52°37 "W and average altitude of 810 m). The region's climate is classified as humid tropical, with mean annual rainfall of 1850 mm and mean annual temperature of 20.5 ± 7.5 °C. The soil of the experimental area was identified as Red Dystrophic Latossolo (Santos et al., 2018) and presents the following chemical properties: pH ($CaCl_2$) = 4.8; organic matter = 17.6 (g dm^{-3}); P = 5.0 (mg dm^{-3}); H + Al = 5.3; K = 69.0 (mg dm^{-3}); Ca = 1.6 ($cmol_c$ dm^{-3}); Mg = 0.5 ($cmol_c$ dm^{-3}); cation exchange capacity (CEC) = 7.6 ($cmol_c$ dm^{-3}); base saturation (V) = 30.0%. The second experiment was conducted at the State University of Mato Grosso do Sul, campus of Aquidauana, MS, Brazil (located at 20°27' South, 55°48 "W and average altitude of 170 m). The region's climate is classified as Tropical Savanna, with mean annual rainfall of 1200 mm and mean annual temperature of 24ºC. The soil of the area was classified as Red Dystrophic Argissolo (Santos et al., 2018) of sandy texture with the following chemical properties: pH ($CaCl_2$) = 6.1; organic matter = 19.74 (g dm^{-3}); P = 67.5 (mg dm^{-3}); K = 0.3 (mg dm^{-3}); Ca = 5.1 ($cmol_c$ dm^{-3}); cation exchange capacity (CEC) = 5.1 ($cmol_c$ dm^{-3}); base saturation (V) = 45.0%. The location of the study sites is shown in Fig. 1.

In both experiments, liming was performed three months before sowing in each season to raise the base saturation to 60%, as recommended by Sousa and Lobato (2017). The limestone used had a relative total neutralizing power (TNP) of 90% and a neutralizing power (NP) of 107%. The percentage of CaO and MgO is 31 and 21%, respectively. Sowing occurred in October 2019 using a conventional tillage system. A randomized block design with two repetitions and 206 F_2 soybean populations was used. The plots consisted of one row three meters long with spacing of 0.45 m between rows and 15 plants m^{-1}. The seeds were treated with fungicide (Pyraclotrobin + Methyl Thiophanate) and insecticide (Fipronil) at a dose of 200 mL of the commercial product for every 100 kg seeds to protect against the attack of pests and soil fungi. For biological nitrogen fixation (BNF), the seeds were inoculated with *Bradyrhizobium* spp. bacteria at a dose of 200 mL of concentrated liquid inoculant for each 100 kg seeds. The cultural treatments were performed according to the needs of the crop. Figure 1 shows the weather conditions during the experiment.

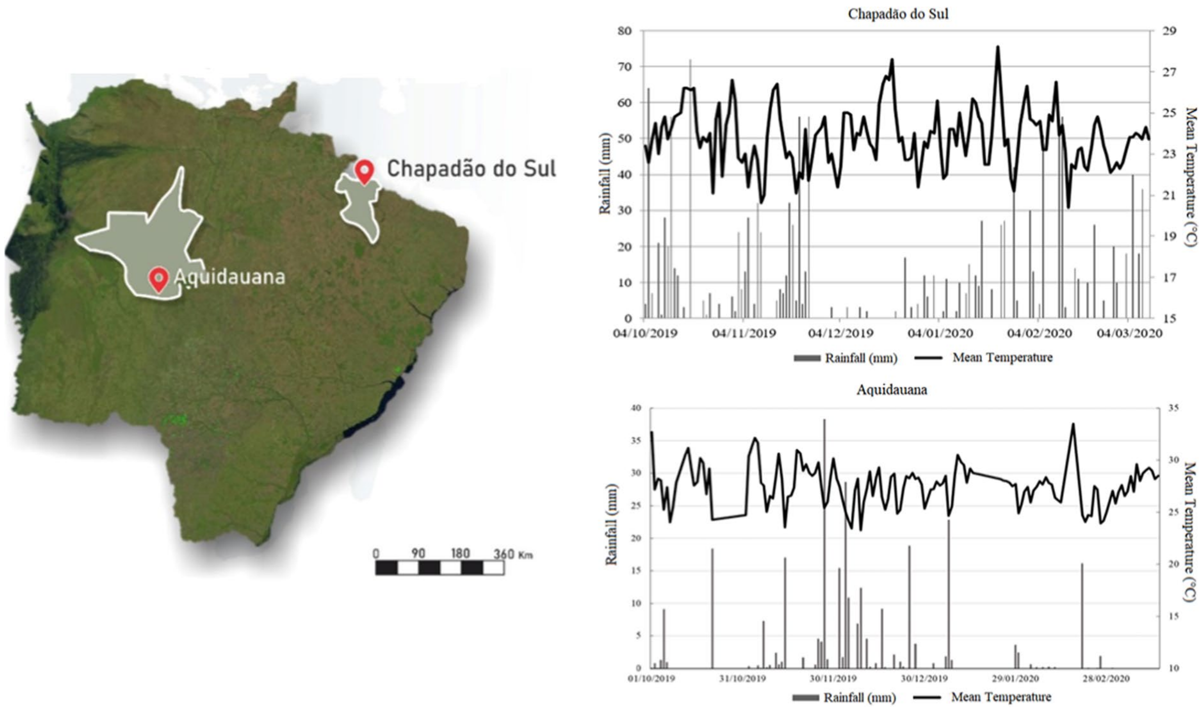


Fig. 1 Location of the study areas: municipalities of Chapadão do Sul and Aquidauana, Mato Grosso do Sul, Brazil

Agronomic traits

At maturity, the following agronomic traits were evaluated: days to maturity (DM), first pod insertion height (PIH, cm), plant height (PH, cm), number of branches (NB), main stem diameter (SD, cm), mass of one hundred grains (MHG, g), and grain yield (GY, kg ha^{-1}). DM corresponded to the days between emergence and maturation of more than 50% of plants in each experimental unit. The traits PIH, PH, SD, and NB and were evaluated in five plants per plot, with the three first being evaluated with the aid of a tape measure. To obtain the MHG, a sample was taken from the harvested grains and the humidity corrected to 13%. GY was evaluated by harvesting the central two meters of each plot and correcting for 13% moisture.

Industrial technological variables

To determine the crude protein (%), oil (%), fiber (%), and ash (%) contents in the F_2 soybean populations, near-infrared spectroscopy (NIRS) (Metrohm, DS2500 spectrometer, Herisau, Switzerland) with high optical precision was used. Samples were homogenized and placed in a sampling dish. The analysis was based on illuminating a sample with specific wavelength radiation in the near-infrared region and then measuring the difference between the amount of energy emitted by the spectroscope and reflected by the sample to the detector. The recording of spectral data was performed in the reflectance mode, within the spectral range of 400–2500 nm (Barnes et al. 1989). The result obtained was compared to a calibration set (Horwitz et al. 1970).

Machine learning models

The ML models tested were: artificial neural network (ANN), support vector machine (SVM), decision tree algorithms J48 and REPTree, and Random Forest (RF). Conventional logistic regression (LR) technique was used as a control model. The SVM performs prediction tasks by building hyperplanes in a multidimensional space to distinguish different classes (Rajvanshi and Chowdhary 2017). The ANN tested consists of the Multilayer Perceptron with a single hidden layer formed by a number of neurons equal to the number of attributes plus the number of classes, all divided by 2 (Egmont-Petersen et al. 2002). The J48 decision tree model is an adaptation of the C4.5 classifier that can be used in regression problems with an additional pruning step based on an error reduction strategy (Snousy et al. 2011). REPTree uses decision tree logic and creates multiple trees at different repetitions. It then selects the best tree using information gain and performs error reduction pruning as the splitting criteria (Kalmegh 2015). RF model can produce multiple prediction trees for the same dataset and use a voting scheme among all these learned trees to predict new values (Belgiu and Drăgu 2016). The six models tested were run on an AMD® PRO A10-8770E R7 CPU with 8 GB RAM, and all hyperparameters were set according to the default setting of the Weka software (Version 3.9.4, University of Waikato, Hamilton, New Zeland).

To generate the genotype groups from the populations, the data were submitted to principal component analysis (PCA). A biplot was constructed with the first two principal components due to the easy interpretation of these results. In this biplot, three clusters (C1, C2, C3) were defined based on the performance of the genotypes for the industrial technological variables for subsequent use of the k-means algorithm, which clusters treatments whose centroids are closest until there is no significant variation in the minimum distance of each observation to each of the centroids. These analyses were performed with the help of the "ggfortify" package of the free R application (R Development Core Team 2014). For the ML analyses, the supervised learning approach was adopted, in which the three clusters formed were used as output variables of the models, while the agronomic traits (DM, PIH, PH, NB, SD, MHG, and GY) were used as input variables of the models. Cluster classification was performed by the six ML models in a stratified cross-validation with k-fold = 10 and ten repetitions (100 runs for each model).

Statistical analyses

To evaluate the performance of classifier models, the following metrics were used: percentage of correct classifications (CC) and F-score. These metrics use a confusion matrix, which indicates the correct or incorrect classification of the classes in use, grouping the results into four classes: False Negative (FN), False Positive (FP), True Positive (TP) and True Negative (TN). The number of correct classifications obtained the percentage of correct classification for each algorithm by the machine learning algorithm in relation to the group that each genotype belonged to divided by the total number of classifications performed. F-score, also known as F-measure or F1 Score, is a precision measure of a test that considers both the precision and the recall of the test to calculate the score. F-measure can be interpreted as a weighted harmonic average of precision and recall, where an F1 score reaches its best value at 1 and the worst score at 0. Precision, also called positive predictive value, is the proportion of positive results that are truly positive. Recall, also called sensitivity, is the ability of a test to correctly identify the positive results to get the true positive rate (Cornelissen and Loureiro 2020). These performance metrics were obtained on the Weka software.

To evaluate the ML models' performance, the means of correct classifications (CC, %) and F-score for all models were grouped by the Scott-Knott test (Scott and Knott, 1974) at a 5% significance level. Boxplots were then generated to express the results graphically. These analyses were run on R software using the packages "ggplot2" and "ExpDes.en".

Results and discussion

Analysis of variance

The individual analyzes of variance for each location are contained in Table 1, while the joint analysis of variance is included in Table 2. For all traits, block effects were non-significant at both sites and in the joint analysis. The effects of genotypes (G) were significant in individual and joint analyzes for all traits evaluated. In the joint analysis, the effect of environments (E) and the GxE interaction were significant for all traits. It is important to highlight that the coefficient of variation values were less than 20% for all variables in all cases.

The presence of genetic variability between soybean genotypes and significant GxE interaction is important to verify the ability of machine learning algorithms to classify soybean genotypes

Table 1 P-value of individual analysis of variance for traits days to maturity (DM), first pod insertion height (PIH, cm), plant height (PH, cm), number of branches (NB), main stem diameter (SD, cm), mass of one hundred grains (MHG, g), and grain yield (GY, kg ha^{-1}), protein (%), oil (%), fiber (%), and ash (%) contents evaluated on 206 F$_2$ soybean genotypes in two locations

Trait	Environment	Block	Genotype (G)	Mean	CV (%)
DM	Aquidauana	0.56	0.00	110.33	7.71
	Chapadão do Sul	0.43	0.00	107.50	6.45
PIH	Aquidauana	0.31	0.01	8.53	12.34
	Chapadão do Sul	0.76	0.01	8.45	8.94
PH	Aquidauana	0.09	0.00	81.25	13.98
	Chapadão do Sul	0.15	0.00	78.19	11.36
NB	Aquidauana	0.91	0.02	3.67	14.13
	Chapadão do Sul	0.56	0.01	4.02	12.22
SD	Aquidauana	0.65	0.01	3.33	8.99
	Chapadão do Sul	0.47	0.02	3.67	9.07
MHG	Aquidauana	0.34	0.00	17.75	13.23
	Chapadão do Sul	0.22	0.00	18.90	14.98
GY	Aquidauana	0.09	0.00	2350.45	15.45
	Chapadão do Sul	0.17	0.00	2980.21	9.97
Protein	Aquidauana	0.34	0.00	36.41	8.45
	Chapadão do Sul	0.59	0.00	37.21	9.43
Oil	Aquidauana	0.67	0.01	19.71	11.25
	Chapadão do Sul	0.91	0.00	20.03	10.03
Fiber	Aquidauana	0.20	0.03	5.56	6.45
	Chapadão do Sul	0.31	0.04	5.82	6.02
Ash	Aquidauana	0.44	0.04	4.96	7.09
	Chapadão do Sul	0.51	0.03	4.83	6.34

CV, Coefficient of variation

Table 2 *P* value of jointed analysis of variance for traits days to maturity (DM), first pod insertion height (PIH, cm), plant height (PH, cm), number of branches (NB), main stem diameter (SD, cm), mass of one hundred grains (MHG, g), and grain yield (GY, kg ha^{-1}), protein (%), oil (%), fiber (%), and ash (%) contents evaluated on 206 F$_2$ soybean genotypes in two locations

Trait	Block/E	Genotype (G)	Environment (E)	GxE	Mean	CV (%)
DM	0.47	0.00	0.00	0.04	108.92	6.71
PIH	0.56	0.01	0.00	0.03	8.49	9.61
PH	0.11	0.01	0.00	0.01	79.72	10.04
NB	0.72	0.04	0.01	0.02	3.85	13.19
SD	0.58	0.03	0.01	0.02	3.50	9.03
MHG	0.29	0.02	0.00	0.01	18.33	14.78
GY	0.16	0.00	0.02	0.01	2665.33	15.41
Protein	0.46	0.01	0.03	0.02	36.81	9.16
Oil	0.72	0.01	0.04	0.04	19.87	10.71
Fiber	0.26	0.04	0.03	0.04	5.89	6.21
Ash	0.47	0.04	0.02	0.01	4.90	6.59

CV, Coefficient of variation

according to industrial traits. There are some studies that have carried out this procedure using spectral variables obtained with remote sensors (Santana et al. 2023). However, for this to be done there is an additional cost for the breeding program. The use of agronomic traits, which are routinely evaluated in breeding programs, as input into machine learning models is a low-cost alternative to access some technological traits of soybean genotypes (Fig. 2).

Principal component analysis—PCA

Based on the PCA results, three homogeneous clusters regarding the industrial variables were formed (Fig. 3). This analysis aims to measure the interrelationship between the treatments (genotypes contained in the clusters) and the variables. Most genotypes of cluster 1, which are associated with higher Protein and Ashes contents, are in the second quadrant. In the third quadrant are contained most of the genotypes in cluster 2, which is more associated with higher Protein and Fiber contents. Lastly, the gray points (cluster 3) are scattered throughout the other quadrants, showing no relationship with the variables. According to Hongyu et al. (2015), for applications in various areas of knowledge, the number of components used has been the one that accumulates 70% or more proportion of the total variance. Therefore, the total variance obtained by PC1 plus PC2 (74.4%) indicates that the graph can be interpreted accurately.

Correlation analysis

According to the correlations and scatterplot (Fig. 4), the relationship of the genotypes of the cluster 3 with the variables Oil and Protein showed high variability, while the genotypes of clusters 1 and 2 showed low variability. The genotypes in C2 showed high variability for Fiber and C1 for Ashes. The genotypes with higher variability are interesting for selection of individuals for genetic improvement. The existence of variability among the genotypes reveals the possibility of selecting individuals with higher means for industrial variables, as well as discarding genotypes with inferior means, contributing to decision-making in breeding programs.

There was significance among all clusters for the Oil x Protein correlation, but there was a positive correlation only for C3. For Fiber x Protein, there was only a significant negative correlation for C1. There was no significant correlation for Fiber x Oil considering the clusters formed. Only C1 showed a significant positive correlation for Ashes x Protein, while for Ashes x Fiber, there was a positive correlation only for cluster C2. It can be observed that the genotypes clustered into C1 and C2 presented a negative correlation for Oil and Protein, corroborating the difficulty in obtaining high means for both variables concomitantly.

Many studies have reported the existence of a negative association between protein content and oil yield. Thus, as protein content is increased, oil yield is reduced, and vice versa (Lee et al. 2019).

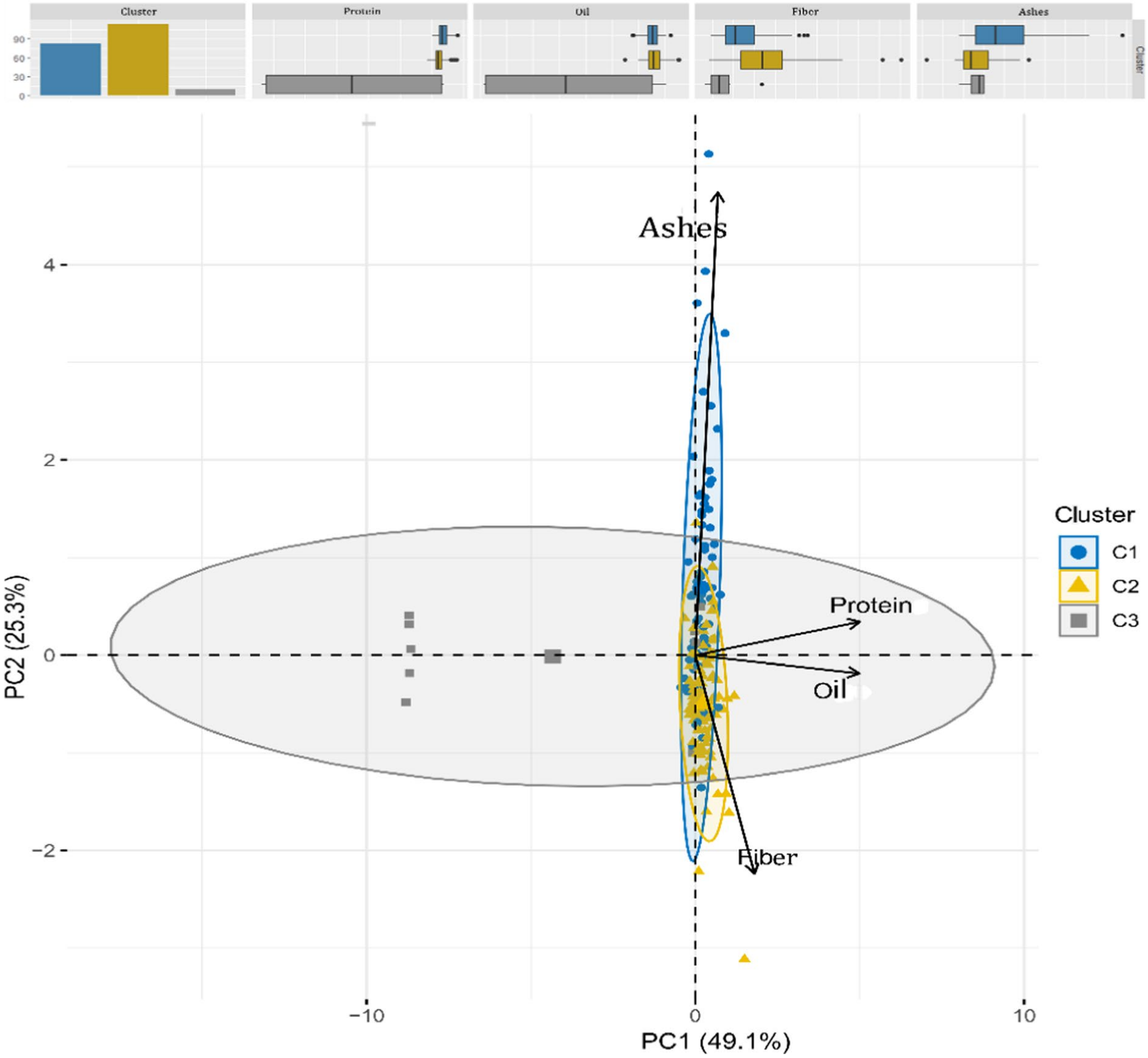


Fig. 2 Principal Component Analysis (PCA) for the genotypes clustered by the k-means algorithm. Blue sample points (circles) belong to cluster 1 (C1), yellow points (triangles) to cluster 2 (C2), and gray points (squares) to cluster 3 (C3)

However, the C3 genotypes showed a positive correlation for these variables, evidencing that although some authors have reported a negative relationship between oil yield and protein content (Bandillo et al. 2015; Kambhampati et al. 2020; Pipolo et al. 2015), there are genotypes in which this correlation is positive. This finding is of crucial importance for soybean breeding since it reveals that it is possible to select genotypes with high means for both variables.

Performance of ML models

Figure 4 shows the percentage of correct classifications (CC) based on the different ML models tested. ANN, REPTree, RF, and SVM showed the highest percentage of correct classifications, while J48 and RL had the lowest CC means. For F-score, J48 and RF showed better performance. Therefore, RF was the best performing model considering both accuracy metrics.

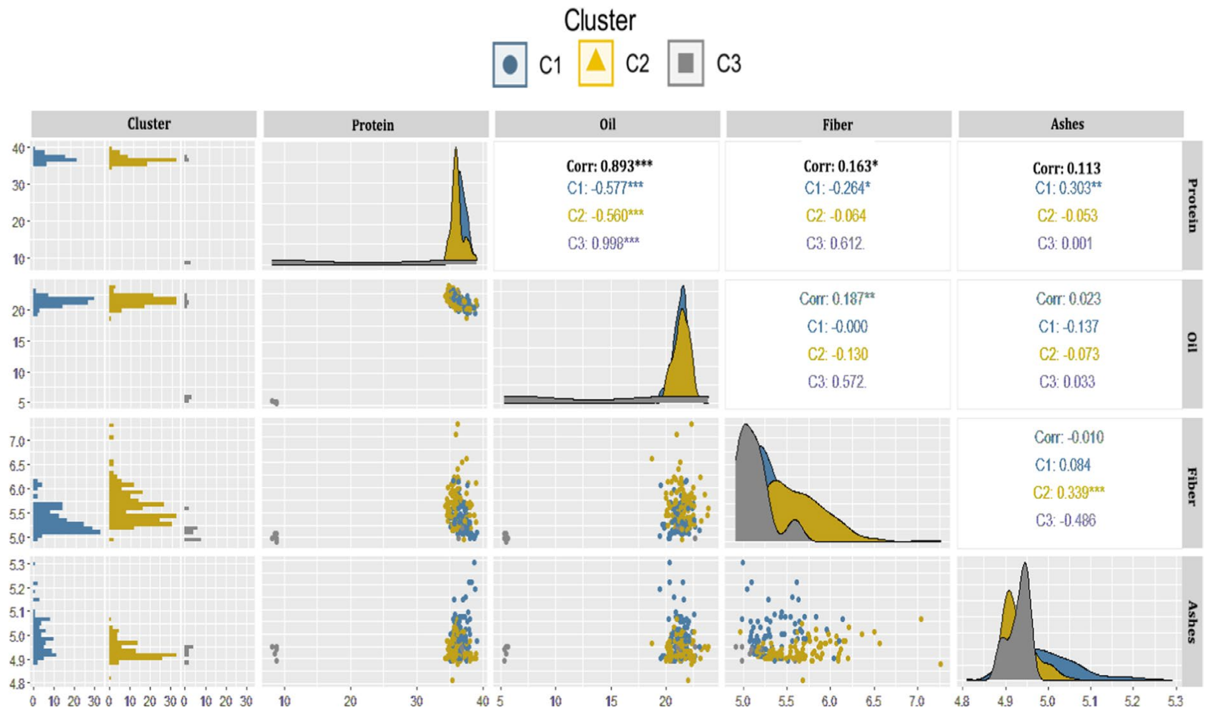
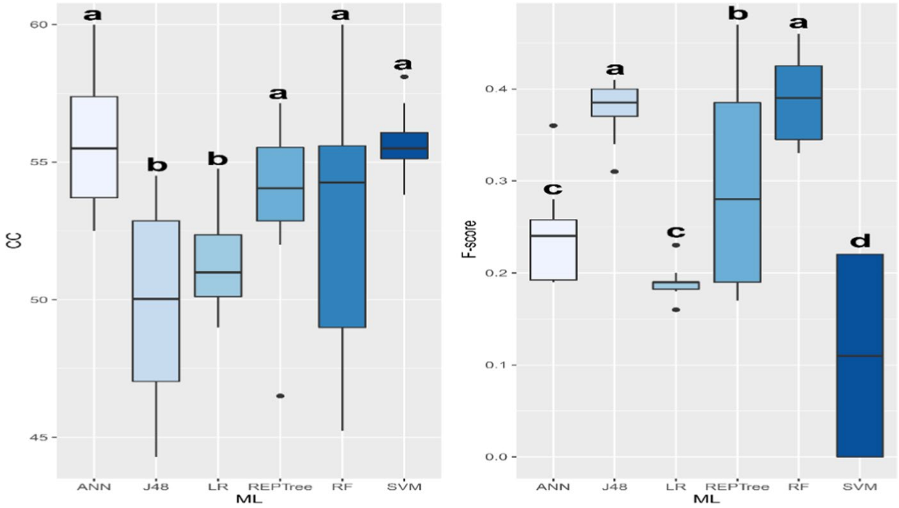


Fig. 3 Correlations and scatter plot between the clusters and the variables crude protein (Protein), oil yield (Oil), crude fiber (Fiber), and Ashes contents; ***, ** and *: significant at 0.1%, 1% and 5% probability, respectively, by the F test

Fig. 4 Boxplot for mean correct classifications (CC, %) and F-score considering logistic regression analysis (LR) and the ML models: neural network (ANN), decision tree algorithms J48 and REPTree, random forest (RF), and support vector machine (SVM). Groups of means with equal letters do not differ by the Scott-Knott test at 5% significance level



RF is commonly used in data modeling studies, and has achieved superior results compared to other techniques, especially in classification studies using spectral, multispectral, and hyperspectral data (Fletcher and Reddy 2016; Teodoro et al. 2021; Zhou et al. 2020). Although the accuracy values obtained with RF cannot be considered high magnitude (50% of CC and 0.40 of F-score), they are in line with those reported in the literature regarding the classification of soybean genotypes for nutritional traits. It is possible to predict the technological traits of soybeans using only the agronomic characters combined with this algorithm, without the need for additional analysis. This information is of great importance for soybean breeding programs.

However, in studies of genotype classification for traits of interest in plant breeding, this technique is still little explored. The RF algorithm is a method that proposes to group input variables using several decision trees that are built at the time of training step within the vector of characteristics of each tree, and then some of the tree attributes are randomly selected. Once this is done, the entropy presented by each attribute is calculated and the one with the highest entropy is chosen to separate the classes in that position of the tree. The output of the classifier will be the one in which the class was returned as the answer by most of the trees belonging to the forest (Breiman 2019). The main advantage of using RF is the elimination of overfitting, a very common problem when using decision trees (Belgiu and Drăgu 2016), which can justify its superior performance.

Our findings reveal that the clustering of genotypes with better performance for industrial variables obtained by k-means and PCA, and the use of these groups as output in supervised ML models using agronomic traits obtained still in the field as input data is a promising strategy for complex data analysis in soybean crop. The approach used here allows time, labor and funding savings, thus contributing to better decision-making in soybean breeding programs aimed at obtaining genotypes with higher oil yield, protein content and productive performance, which is one of the major current challenges in soybean improvement. Future studies should test other data modeling, such as the prediction of industrial technological variables based on agronomic ones aiming at identifying auxiliary variables in the selection of genotypes combining good performance for industrial and agronomic variables, which will require even more powerful ML models due to the complex relationship between these variables.

Conclusion

The use of machine learning techniques enables moderate accurate classification of soybean genotypes for industrial technological variables based on agronomic traits as input in models. Based on the percentage of correct classifications and F-score, Random Forest is the most efficient classification technique.

Acknowledgements The authors would like to thank the Universidade Federal de Mato Grosso do Sul (UFMS), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Grant numbers 303767/2020-0, and 304979/2022-8, and Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul (FUNDECT) TO numbers 88/2021, 07/2022, 318/2022 and 94/2023, and SIAFEM numbers 30478, 31333, 32242 and 33111. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil (CAPES) – Financial Code 001.

Author contributions L.P.R.T., B.B., F.E.T. and P.E.T. collected the data. L.P.R.T., M.O.S., P.E.T., and P.C.C. produced a draft of the manuscript. L.P.R.T., P.E.T., and M.O.S. performed all statistical analyses. C.A.S.J. and F.E.T. contributed with a critical review of the manuscript. All authors read and approved the final manuscript.

Funding The authors have not disclosed any funding.

Data availability The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Declarations

Conflicts of interest The authors declare no conflict of interest.

References

Alaswad AA, Song B, Oehrle NW, Wiebold WJ, Mawhinney TP, Krishnan HB (2021) Development of soybean experimental lines with enhanced protein and sulfur amino acid content. Plant Sci 308:110912. https://doi.org/10.1016/j.plantsci.2021.110912

André Cremonez P, Feroldi M, Cézar Nadaleti W, De Rossi E, Feiden A, De Camargo MP, Cremonez FE, Klajn FF (2015) Biodiesel production in Brazil: current scenario and perspectives. Renew Sustain Energy Rev 42:415–428. https://doi.org/10.1016/j.rser.2014.10.004

Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J, Lorenz A (2015) A population structure and genome-wide association analysis on the usda soybean germplasm collection. Plant Genome. https://doi.org/10.3835/plantgenome2015.04.0024

Barnes RJ, Dhanoa MS, Lister SJ (1989) Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl Spectrosc 43:772–777. https://doi.org/10.1366/0003702894202201

Batista TS, Teodoro LPR, Azevedo GB, de Azevedo GTDOS, Poersch NL, Borges MVV, Teodoro PE (2022) Artificial neural networks and non-linear regression for quantifying the wood volume in eucalyptus species. South For J For Sci. 84:1–7. https://doi.org/10.2989/20702620.2021.1976604

Belgiu M, Drăgu L (2016) Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 114:24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Breiman L (2019) Random forests. Random for. 1–122. https://doi.org/10.1201/9780429469275-8

Burton JW (1985) No titlworld soybean research conference III: Proceedingse, 1st editio. ed. Boca Raton. https://doi.org/10.1201/9780429267932

Cober ER, Voldeng HD (2000) Cs-40–1–39 (1) 1994–1997

Cornelissen W, Loureiro M (2020) Automatic onset detection using convolutional neural networks 199–200. https://doi.org/10.5753/sbcm.2019.10446

Egmont-Petersen M, De Ridder D, Handels H (2002) Image processing with neural networks–a review. Pattern Recognit 35:2279–2301. https://doi.org/10.1016/S0031-3203(01)00178-9

Fletcher RS, Reddy KN (2016) Random forest and leaf multispectral reflectance data to differentiate three soybean varieties from two pigweeds. Comput Electron Agric 128:199–206. https://doi.org/10.1016/j.compag.2016.09.004

Goldsmith PD (2008) Economics of soybean production, marketing, and utilization. Soybeans Chem Prod Process. https://doi.org/10.1016/B978-1-893997-64-6.50008-1

Hongyu K, Jorge G, Junior DO (2015) Análise de Componentes Principais : resumo teórico aplicação e interpretação principal component analysis : theory interpretations and applications. E&S Eng Sci 1:83–90. https://doi.org/10.18607/ES20165053

Horwitz W, Chichilo P, Reynolds H (1970) Official methods of analysis of the Association of Official Analytical Chemists, Washington, DC, USA: Association of Official Analytical Chemists

Kalmegh S (2015) Analysis of WEKA data mining algorithm REPTree, simple cart and randomtree for classification of indian news. Int J Innov Sci Eng Technol 2:438–446

Kambhampati S, Aznar-Moreno JA, Hostetler C, Caso T, Bailey SR, Hubbard AH, Durrett TP, Allen DK (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 10:1–15. https://doi.org/10.3390/metabo10010018

Lee S, Van K, Sung M, Nelson R, LaMantia J, McHale LK, Mian MAR (2019) Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I–IV. Theor Appl Genet 132:1639–1659. https://doi.org/10.1007/s00122-019-03304-5

Marques Ramos AP, Prado Osco L, Elis Garcia Furuya D, Nunes Gonçalves W, Cordeiro Santana D, Pereira Ribeiro Teodoro L, da Silva Antonio, Junior C, Fernando Capristo-Silva G, Li J, Henrique Rojo Baio F, Marcato Junior J, Eduardo Teodoro P, Pistori H (2020) A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices. Comput Electron Agric 178:105791. https://doi.org/10.1016/j.compag.2020.105791

Pipolo EA, Hungria M, Franchinio JC, Junior AAB, Debiasi H, Mandarino JMG, (2015) Comunicado técnico 86: teores de óleo e proteína em soja: fatores envolvidos e qualidade para a indústria. In Portuguese 1–15

R Development Core Team (2014) R: a language and environment for statistical computing

Rajvanshi N, Chowdhary KR (2017) Comparison of SVM and naïve bayes text classification algorithms using WEKA. Int J Eng Res. https://doi.org/10.17577/ijertv6is090084

Ramos LP, Kothe V, César-oliveira MAF, Nakagaki S, Krieger N, Wypych F, Cordeiro CS (2017) Artigo biodiesel : matérias-primas , tecnologias de produção e propriedades combustíveis biodiesel : matérias-primas , tecnologias de produção e propriedades combustíveis. https://doi.org/10.21577/1984-6835.20170020

Santana DC, Teodoro LPR, Baio FHR, dos Santos RG, Coradi PC, Biduski B, Shiratsuchi LS (2023) Classification of soybean genotypes for industrial traits using UAV multispectral imagery and machine learning. Remote Sens Appl Soc Environ 29:100919

Santos et al., (2018) Sistema brasileiro de classificação de solos, Embrapa Solos

Schwalbert RA, Amado T, Corassa G, Pott LP, Prasad PVV, Ciampitti IA (2020) Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agric for Meteorol 284:107886. https://doi.org/10.1016/j.agrformet.2019.107886

Singh A, Ganapathysubramanian B, Singh AK, Sarkar S (2016) Machine learning for high-throughput stress phenotyping in plants. Trends Plant Sci 21:110–124. https://doi.org/10.1016/j.tplants.2015.10.015

Snousy MBA, El-Deeb HM, Badran K, Khlil IAA (2011) Suite of decision tree-based classification algorithms on cancer gene expression data. Egypt Informatics J 12:73–82. https://doi.org/10.1016/j.eij.2011.04.003

Sousa, DMG, Lobato E (2017) Cerrado–Correção do solo e adubação

Teodoro PE, Teodoro LPR, Baio FHR, da Silva Junior CA, Dos Santos RG, Ramos APM, Pinheiro MMF, Osco LP, Gonçalves WN, Carneiro AM, Marcato Junior J, Pistori H, Shiratsuchi LS (2021) Predicting days to maturity, plant height, and grain yield in soybean: a machine and deep learning approach using multispectral data. Remote Sens. https://doi.org/10.3390/rs13224632

van Dijk ADJ, Kootstra G, Kruijer W, de Ridder D (2021) Machine learning in plant science and plant breeding. Iscience 24:101890. https://doi.org/10.1016/j.isci.2020.101890

Zhou J, Zhou J, Ye H, Ali ML, Nguyen HT, Chen P (2020) Classification of soybean leaf wilting due to drought stress using UAV-based imagery. Comput Electron Agric 175:105576. https://doi.org/10.1016/j.compag.2020.105576

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.