Pantanal Editora
Trends in Agricultural and Environmental Sciences
10.46420/TAES. e240004

A qualitative Decision Tree model for Common Beans and Cowpea classification

Bruno Rodrigues de **Oliveira**[1,*], Jorge González **Aguilera**[1,2], Fabio **Steiner**[2], Diógenes Martins **Bardiviesso**[2], Luis **Morales-Aranibar**[3], and Leandris **Argentel-Martínez**[4]

[1] Pantanal Editora, Nova Xavantina-MT;
[2] Universidade Estadual de Mato Grosso do Sul/UEMS, Departamento de Agronomia, Cassilândia, MS, Brasil;
[3] Universidad Nacional Intercultural de Quillabamba (UNIQ), Cusco, Perú;
[4] Tecnológico Nacional de México, Instituto Tecnológico del Valle del Yaqui, Bácum, Sonora, México;
* Correspondence: bruno@pantanaleditora.com.br

Received: 2024-03-28
Accepted: 2024-04-25
Published: 2024-04-26

Main Editors
Alan Mario Zuffo

For citation: Oliveira, B. R.; Aguilera, J. G.; Steiner, F; Bardiviesso, D. M.; Morales-Aranibar, L; Argentel-Martínez, L (2024). A qualitative Decision Tree model for Common Beans and Cowpea classification. Trends in Agricultural and Environmental Sciences, (e240004), DOI: 10.46420/TAES.e240004

Abstract: Common beans and cowpea are two grains that form part of the preferred diet in several countries, mainly due to their nutritional value. Knowledge of their diversity is important for plant breeding and determines the conservation and use strategy. Previous analyzes show that there is variability for a set of qualitative and quantitative descriptors for this species. The objective of these work was to use data from qualitative descriptors to generate a decision tree model that makes it possible to classify common bean and cowpea genotypes. 17 bean genotypes were used, 12 of which were common beans and 5 were cowpeas. Eight qualitative descriptors were used to characterize the bean genotypes. Machine learning techniques were used to generate decision tree models for classifying bean genotypes. Using the accuracy, precision and F1-score metrics in the cross-validation approach, we select the best decision tree model. This model was adapted into a flowchart for use in various purposes, aiming to classify beans based on qualitative descriptors.

Keywords: *Phaseolus vulgaris L.*, *Vigna unguiculata L. Walp.*, selection, Machine learning.

1. Introduction

Common beans (*Phaseolus vulgaris* L.) and cowpeas (*Vigna unguiculata* L. Walp.) hold significant importance in the human diet across various countries worldwide, primarily due to their high nutritional value and protein content (Abebe & Alemayehu, 2022; Catarino et al., 2021; Singh, 2015). Both types of beans are rich in essential amino acids, dietary fiber, B vitamins and minerals such as iron, zinc and magnesium. Furthermore, they contain phytochemical compounds with antioxidant and anti-inflammatory properties. These foods play a significant role in promoting health and preventing disease when integrated into a balanced and varied diet (Didinger et al., 2022; Enyiukwu et al., 2020).

Understanding the diversity within any crop starts with the conservation and selection of genotypes. This initial step is crucial for identifying genetic resources, essential for maintaining germplasm banks and ensuring food security for future generations (Elsayed et al., 2023; Özkan et al., 2022; Sampaio et al., 2023; Wu et al., 2021). Studies on beans have elucidated the genetic dissimilarities within the crop (Cabral et al., 2011; Catarino et al., 2021; Coelho et al., 2007; Guimarães et al., 2023; Tavares et al., 2018). These investigations reveal variability in morphological characteristics and growth development. Genetic improvement and breeding programs are of paramount importance for crops like cowpea and common bean. In the case of cowpea, due to its adaptability to water stress and its nitrogen-fixing capacity (de Sousa Leite et al., 2023), as well as its wide geographical distribution, its genetic diversity, both in cultivated varieties and wild relatives, represents a valuable resource for breeding programs (Maia, 2023).

However, despite advances in breeding programs, the genetic base of cowpea remains narrow, necessitating the exploration of alien germplasm to broaden this base (Boukar et al., 2020; Catarino et al., 2021). Additionally, research on the development of transgenic cowpea varieties resistant to pests, such as the pod borer (*Maruca vitrata*), highlights a promising area for improving pest resistance and increasing productivity (Catarino et al., 2021). In the case of common bean, disease resistance is crucial to ensure production stability. Breeding strategies, including phenotypic selection and marker-assisted selection, have been employed to develop varieties resistant to a variety of pathogens, such as viruses, fungi, and bacteria, with marker-assisted selection offering advantages in terms of efficiency and accuracy in identifying resistance genes (Catarino et al., 2021; Watare, 2023). Considering the challenges faced by agriculture, such as climate change and increased pressure from diseases and pests, genetic improvement programs play a vital role in ensuring food security and agricultural sustainability.

In a recent study (Aguilera et al., 2023), the authors researched the genetic diversity of bean genotypes using qualitative and quantitative descriptors for characterization. The research, carried out with 17 bean cultivars, revealed significant differences between the genotypes in terms of weight, size and qualitative characteristics of the seeds. Principal component analysis (PCA) identified five divergent groups, highlighting the genetic variability present in bean germplasm. The general results emphasized the importance of germplasm characterization for the selection and conservation of genetic resources, fundamental for future food security. The diversity observed in the genotypes suggests a significant potential for genetic improvement programs, allowing the identification of superior parents and the achievement of genetic gains. The combination of qualitative and quantitative descriptors proved to be an effective strategy in the discrimination and selection of bean genotypes. This strategy has also been tested in works involving *Solanum lycopersicon* cultivation (Aguilera et al., 2019), *Capsicum* spp. (Sampaio et al., 2023), *Manihot esculenta* (Vilela Barros et al., 2020).

Machine learning has proven crucial in agricultural applications due to its ability to process large volumes of data and extract meaningful insights to improve agricultural productivity and efficiency (Liakos et al., 2018). Within this context, decision tree models play a fundamental role in knowledge discovery (Sivagama Sundhari, 2011), offering an interpretable and effective approach to agricultural data analysis. These models are capable of identifying complex patterns in data, allowing farmers to make informed decisions about cultivation, pest management, resource optimization and crop forecasting (de Oliveira et al., 2023; Marin et al., 2021). By integrating machine learning and decision tree models into agricultural practices, it is possible to boost sustainability and food security (Tariq et al., 2023), contributing to an agricultural sector that is more resilient and adaptable to environmental and climate change (Yeganeh-Bakhtiary et al., 2022).

Given the importance of beans in food and the need for constant genetic improvement for genotypes more adapted to severe conditions, mainly due to climate change, and also the relevance that machine learning has shown for the discovery of knowledge in agriculture, in this work we propose the use of decision tree algorithms to obtain a classification model for Common beans and cowpeas species.

2. Material and Methods

2.1 Experimental design

The experiment was conducted at the State University of Mato Grosso do Sul (UEMS) in Cassilândia, MS, Brazil. A total of 17 bean genotypes, 12 common bean genotypes, and 5 cowpea genotypes were purchased from the local seed market in the municipality of Cassilândia, MS, Brazil. The seeds of cowpea genotypes are part of the UEMS/Cassilândia seed bank. The detailed description of the 17 bean genotypes used in this study is shown in the work described

by Aguilera et al. (2023). Data were evaluated in a completely randomized design, with three repetitions of 25 seeds each.

2.2 Qualitative descriptors

To evaluate the genetic divergence among the 17 genotypes, seeds with a moisture content ranging from 12% to 14% were utilized (Ministry of Agriculture, 2009). The assessment involved three repetitions of 25 seeds each to ascertain seed width (SW, in mm), seed length (SL, in mm), and seed thickness (ST, in mm), used to calculate some of the qualitative descriptors.

Embrapa's recommendation (Silva, 2005) was employed to evaluate genetic divergence using qualitative descriptors for characterizing common bean cultivars/varieties (*Phaseolus vulgaris* L.). The assessment involved examining qualitative characteristics using a sample of 10 seeds.

- Seed color: The assessment involved evaluating the uniformity of seed color, with scores of 1 to "Uniform" or 2 to "Non-Uniform";
- Primary and Secondary color given in % by evaluating the percentage of color occurrence in the seed;
- Seed shape: the calculation is founded on the J coefficient (mm) = SL/SW, resulting in the following shapes: Spherical (1.16 to 1.42) and Elliptical (1.43 to 1.65), Oblong/Short Reniform (1.66 to 1.85), Oblong/Medium Reniform (1.86 to 2.00) and Oblong/Long Reniform (> 2.00) (Romero, 1961);
- Degree of seed flattening: calculation from the coefficient H (mm) = ST/SW, where: Flattened (< 0.69), Semi filled (0.70 to 0.79), and Filled (>0.8) (Romero, 1961);
- Seed brightness: involved considering the color shade of the seeds: Opaque, Intermediate, and Bright;
- Seed halo: by observing the presence of the seed halo and assigning the values: "Absent" or "Present";
- Color of the seed halo: evaluation involved considering the color of the seed halo and assigning "Same" color of the seed or "Different" color of the seed.

2.2 Machine learning and Decision Tree

Machine Learning (ML) is a fundamental area of computer science that allows systems to learn patterns and make predictions from data without being explicitly programmed. ML has applications in all areas of knowledge, including agronomy. An essential technique within ML is Pattern Classification, which involves assigning labels to data based on known characteristics (de Oliveira et al., 2021). To this end, in the training stage, examples are provided to the algorithm so that it learns the patterns and in the testing stage, the learned models are tested using performance metrics (Haykin, 2009).

Cross-validation is a crucial technique for evaluating the performance of an ML model. It involves dividing the dataset into training and testing subsets repeatedly in order to check the stability and generalization of the model (Kubat, 2021). Cross-validation helps mitigate bias from arbitrary selection of datasets and provides a more reliable estimate of model performance. To evaluate the effectiveness of an ML model, several performance metrics are used. Accuracy measures the proportion of correct predictions in relation to the total number of predictions.

Precision represents the proportion of correctly classified positive instances among all instances classified as positive. Recall, in turn, indicates the proportion of correctly classified positive instances in relation to all positive instances in the data set. The ROC (Receiver Operating Characteristic) Curve is a graphical representation of the performance of a classification model as the discrimination threshold varies. It illustrates the true positive rate as a function of the false positive rate.

The Confusion Matrix is a table that shows the performance of a classification model in terms of true positives, false positives, true negatives and false negatives, providing a detailed view of the model's performance in different classes (Nalini Durga & Usha Rani, 2020).

Decision Tree is one of the most popular and widely used machine learning algorithms for classification (Fletcher & Islam, 2020). This algorithm builds a hierarchical tree structure made up of nodes and edges. Each internal node of the tree represents a decision based on a specific characteristic, while the edges represent the possible outcomes of that decision. The leaves of the tree represent the output classes or prediction values. When making a prediction, the data travels through the decision tree, following the paths determined by the decisions at each node, until it reaches a leaf. Therefore, an interesting feature of the decision tree is its interpretability.

Because the tree structure can be visualized as a decision flowchart, it is easy to understand and explain how the model makes predictions. This makes the decision tree a popular choice in scenarios where model interpretability is important. Furthermore, this algorithm makes it possible to calculate the importance of variables in the classification task (Louppe et al., 2013), which is useful for analyzing agricultural data, when many variables are obtained in experiments and the aim is to know how they impact treatments (de Oliveira et al., 2023). This importance is obtained from different criteria, such as the reduction of impurity in the nodes, or the gain of information when making a division.

In the context of the scikit-learn package (Pedregosa et al., 2011), the main hyperparameters of a decision tree include: Split Criteria: Defines the function to measure the quality of a split, such as "gini" for the Gini index or "entropy" for information entropy; Maximum Tree Depth: Controls the maximum tree depth to avoid overfitting; Minimum number of samples per sheet: defines the minimum number of samples required on a sheet to perform a split; Minimum number of samples needed to split an internal node: Specifies the minimum number of samples needed in a node to consider splitting; Maximum number of features: limits the number of features to be considered in each division.

3. Results

Table 1 contains the experimental results obtained for each of the selected qualitative descriptors, according to the values they assume in accordance with the description made in Section 2.2 Qualitative Descriptors. In addition, the names of the genotypes and species are also included, which are used as classes in the machine learning scheme to learn the decision tree models.

The proposed approach aims to obtain a decision tree model for classifying beans into two species. To this end, cross-validation with 5 folds was used to overcome overfitting. Furthermore, we computed the weight of the classes as they were unbalanced and this information was used in the decision tree model learning algorithm in order to adjust the model appropriately. Figure 1 (a) shows the ROC curves for each evaluation of the models learned in each of the 5 folds. The Area Under Curve (AUC) values for each fold stand out, in addition to the average values. The boxplots in Figure 1 (b) show the distribution of performance metrics values in cross-validation.

Table 1. Experimental results of qualitative descriptors.

Genotype	Specie	Seed color	Primary color	Secondary color	Seed shape	Degree of seed flattening	Seed brightness	Seed halo	Color of Seed Halo
Paquito	Bean	Non-Uniform	95.00	5.00	Elliptical	Flattened	Intermediate	Present	Same
Rajado 1	Bean	Non-Uniform	90.00	10.00	Oblong/Long Reniform	Flattened	Intermediate	Present	Different
Caupi Sempre Verde	Cowpea	Uniform	100.00	0.00	Elliptical	Filled	Opaque	Present	Different
Caupi Nova Era	Cowpea	Uniform	100.00	0.00	Spherical	Flattened	Opaque	Present	Different
Caupi BRS Guariba	Cowpea	Uniform	100.00	0.00	Spherical	Semi filled	Opaque	Present	Different
Caupi BRS Itaim	Cowpea	Uniform	100.00	0.00	Oblong/Medium Reniform	Filled	Opaque	Present	Different
Caupi BRS Tamucumaqui	Cowpea	Uniform	100.00	0.00	Spherical	Flattened	Opaque	Present	Different
Vô Cid	Cowpea	Non-Uniform	90.00	10.00	Elliptical	Semi filled	Opaque	Present	Same
Rajado 2	Cowpea	Non-Uniform	90.00	10.00	Oblong/Medium Reniform	Filled	Bright	Present	Different
Vermelho Dark	Bean	Uniform	100.00	0.00	Oblong/Short Reniform	Filled	Intermediate	Present	Different
Vermelho	Bean	Uniform	100.00	0.00	Oblong/Long Reniform	Flattened	Bright	Present	Different
Bolhinha	Bean	Uniform	100.00	0.00	Oblong/Short Reniform	Filled	Bright	Present	Different
Bem Te-vi	Bean	Non-Uniform	95.00	5.00	Elliptical	Flattened	Opaque	Present	Different
Branco	Bean	Uniform	100.00	0.00	Oblong/Long Reniform	Filled	Opaque	Present	Same
Branco Dorama	Bean	Uniform	100.00	0.00	Oblong/Medium Reniform	Semi filled	Opaque	Absent	Same
Feijão Preto	Bean	Uniform	100.00	0.00	Elliptical	Semi filled	Intermediate	Present	Different
TAA-Marhe	Bean	Non-Uniform	95.00	5.00	Elliptical	Semi filled	Opaque	Present	Same

Bean: *Phaseolus vulgaris* L.; Cowpea: *Vigna unguiculata* (L.) Walp.

Figure 2 illustrates the confusion matrices for the training and testing stages, both calculated as the average of the confusion matrices for each fold in cross-validation. These performance results, together with those shown by the ROC curves in Figure 1 (a), are used to select the best decision tree model. This will be used to obtain a flowchart for classifying beans in relation to their species.

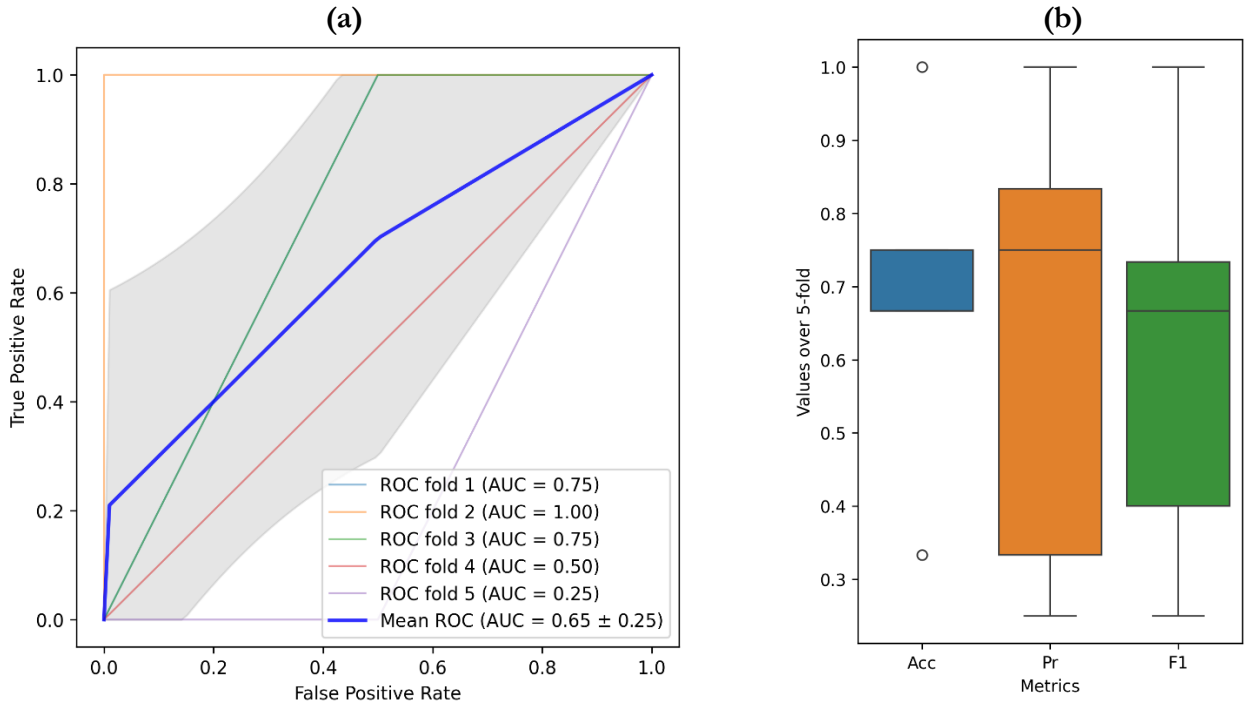

Figure 1. (a) ROC curve for each cross-validation fold, in addition to the average curve. (b) Boxplot of performance metrics values in cross-validation. Accuracy (Acc), precision (Pr), and F1-score (F1).

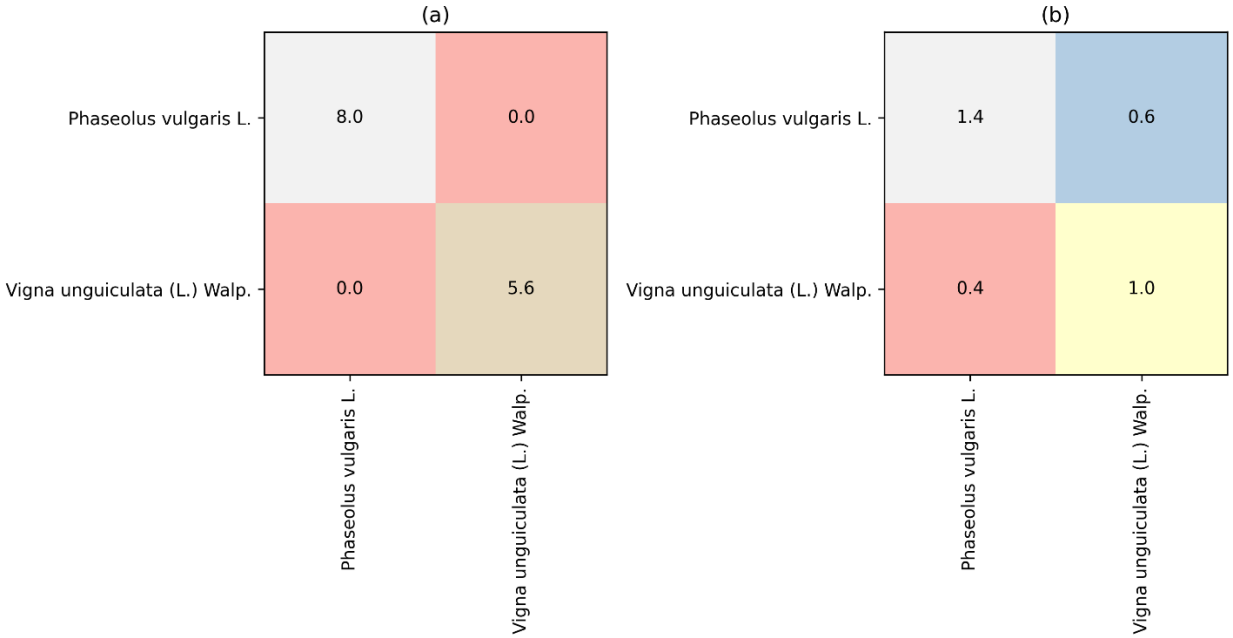

Figure 2. Confusion matrices calculated as the average of the confusion matrices of each fold in cross-validation. (a) train and (b) test stage.

Figure 3 highlights the importance of each of the variables (qualitative descriptors) for building the most accurate decision tree model. Which was selected according to the results shown in Figures 1 and 2. Therefore, among these variables, those with greater importance must compose the chosen decision tree model at some level.

Finally, in Figure 4 we have a flowchart construction of the best decision tree model obtained. Although, at each node (yes or no decision point) of the model, the decision about which class a certain sample belongs to can be made, we prefer to leave decisions about classes to the last leaves of the tree (model). This ensures greater accuracy in classification, when going from the branches to the leaves.

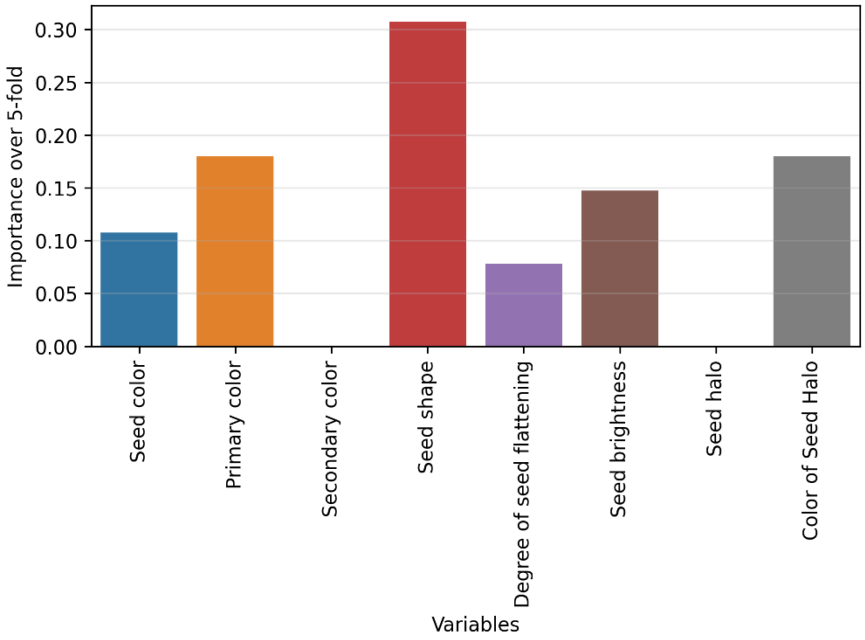

Figure 3. Bar chart with the importance of the variables for the best performing decision tree model.

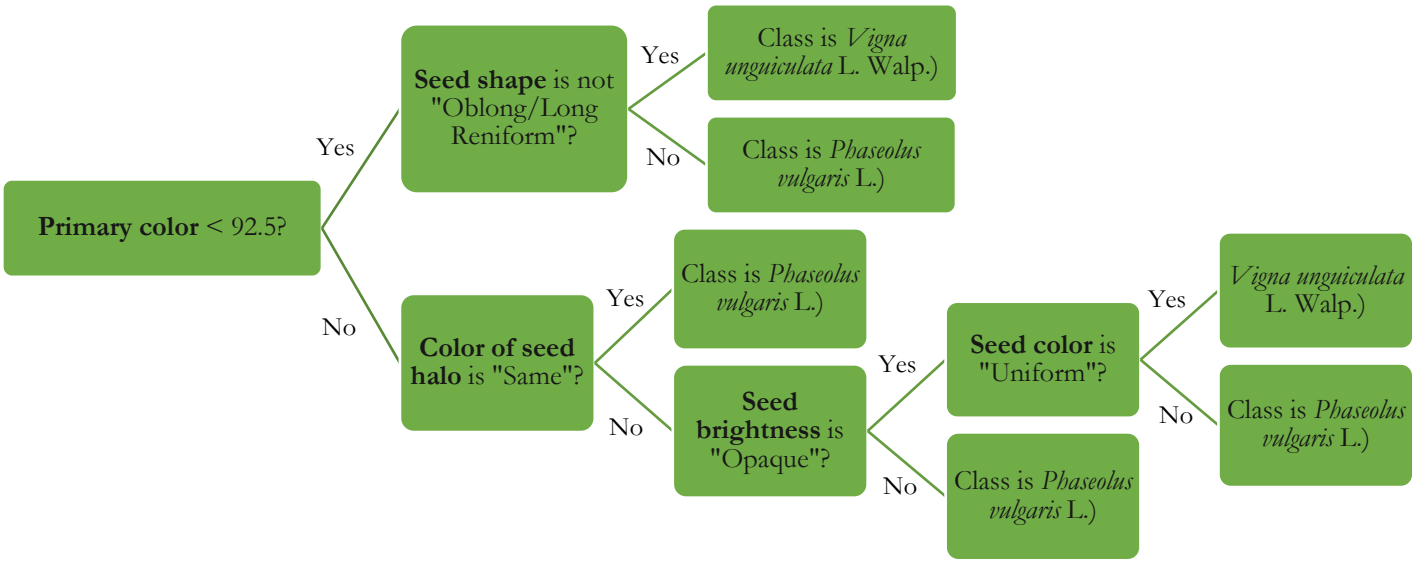

Figure 4. Decision tree model that resulted in the highest performance among the models generated in cross-validation.

4. Discussion

The presented experimental results in Table 1 outline the qualitative descriptors obtained for each selected genotype, alongside their respective species. These descriptors, as detailed in Section 2.2, include seed color, primary and secondary color, seed shape, degree of seed flattening, seed brightness, presence of seed halo, and the color of the seed halo. These descriptors serve as crucial inputs for the machine learning scheme employed to construct decision tree models.

Table 1 exhibits a diverse range of qualitative characteristics observed across the bean and cowpea genotypes. Notably, variations in seed color and shape are evident, with some genotypes displaying uniformity while others exhibit non-uniformity. For instance, genotypes like "Caupi Sempre Verde" and "Caupi Nova Era" demonstrate uniform seed color and shape, whereas genotypes like "Paquito" and "Rajado 1" display non-uniform characteristics. Additionally, the degree of seed flattening varies among the genotypes, with some having flattened seeds while others retain their original shape. Furthermore, the presence and characteristics of the seed halo contribute to the diversity observed among the genotypes. The presence of the seed halo, along with its color, differs across the genotypes, suggesting potential variations in seed composition or protective features (Table 1).

These experimental findings provide valuable insights into the genetic diversity present within the bean and cowpea populations studied. Such diversity is essential for breeding programs aimed at enhancing crop resilience, productivity, and nutritional value. Moreover, the utilization of machine learning techniques, utilizing these qualitative descriptors, facilitates the development of decision tree models for genotype classification and selection, thereby aiding in the advancement of agricultural research and crop improvement strategies. The integration of these findings into scientific discourse contributes to a deeper understanding of plant genetics

and informs future efforts in crop breeding and genetic resource conservation (Parmley et al., 2019; Yoosefzadeh Najafabadi et al., 2023).

The results of the five-fold cross-validation for the decision tree model to classify beans are presented in Figure 1 (a). The Area Under the Curve (AUC) is a metric used to assess the performance of binary classification models. It represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Hajian-Tilaki, 2013). In this case, a positive instance is a bean belonging to a specific class (specie), and a negative instance is a bean belonging to another class. The AUC values for each fold varied between 0.25 and 1.00, with an average AUC of 0.65 ± 0.25. A perfect classifier would have an AUC of 1.0, while a random classifier would have an AUC of 0.5 (Gorunescu, 2011). Therefore, the results indicate that the decision tree model performed moderately well in classifying the bean species. However, the high standard deviation (0.25) suggests that the performance may vary across different datasets. This is partly due to the small number of samples available. Nonetheless, fold 2 achieved a perfect classification (AUC = 1.00), which indicates that the model was able to correctly classify all bean samples in this fold. On the other hand, fold 5 achieved a very low AUC (0.25), which suggests that the model performed poorly in classifying the bean samples. This variability in performance across folds highlights the importance of cross-validation to assess the generalizability of the model (Kohavi, 1995).

The performance metrics of the decision tree model for classifying beans is summarized in the boxplots of Figure 1 (b). The boxplots depict the distribution of three performance metrics across the five folds of the cross-validation process: accuracy (Acc), precision (Pr), and F1-score (F1). The accuracy metric represents the proportion of bean samples that the model classified correctly. The boxplot for accuracy shows a median value of 0.8, indicating that the model achieved an accuracy of 80% on average across the folds. The interquartile range (IQR) spans from 0.7 to 0.9, suggesting that the accuracy remained relatively stable across most of the folds. However, there were outliers, with a minimum accuracy of 0.6 and a maximum accuracy of 1.0. The presence of outliers suggests that the model's accuracy may vary on different datasets.

Precision refers to the proportion of positive predictions that were truly positive. In this case, a positive prediction is a bean sample that the model classified as belonging to a specific class (specie). The precision boxplot shows a median value of 0.75, indicating that on average, 75% of the time the model predicted a bean to belong to a specific class, it was correct. The IQR for precision ranges from 0.65 to 0.9, which is similar to the IQR for accuracy. This suggests that the model's precision was also relatively stable across most of the folds. However, similar to accuracy, there were outliers for precision, with a minimum of 0.4 and a maximum of 1.0.

The F1-score is a harmonic mean between precision and recall. It considers both the model's ability to correctly identify positive samples (precision) and its ability to avoid incorrectly classifying negative samples (recall). A perfect F1-score of 1.0 indicates that the model is performing well in both aspects. The F1-score boxplot shows a median value of 0.8, which is consistent with the median accuracies and precision values. The IQR for F1-score ranges from 0.65 to 0.9, again similar to the IQRs for the other metrics. This suggests that the model's F1-score also remained relatively stable across most folds. There were outliers for F1-score, with a minimum of 0.5 and a maximum of 1.0.

Overall, the performance metrics in the boxplots suggest that the decision tree model achieved moderate performance in classifying bean samples. The model's accuracy, precision, and F1-score were all around 0.8 on average, with some variation across folds. However, the main objective of this research is to obtain a decision tree model with high accuracy, to convert it into a flowchart to assist in the designation of bean species. Therefore, this variation has little effect on the desired result, as long as we have at least one model with high accuracy in the test set.

The confusion matrices in Figure 2 provide a more detailed picture of the decision tree model's performance on both the training and testing stages. These confusion matrices are calculated as the average of the confusion matrices for each fold in the cross-validation process. A confusion matrix is a table that summarizes the number of correct and incorrect predictions made by a classification model. The rows represent the actual classes of the bean samples, and the columns represent the classes predicted by the model. Ideally, a good model would have most of its values concentrated on the diagonal, where the actual class and the predicted class match (Heydarian et al., 2022).

The confusion matrix for the training stage — Figure 2 (a) — shows the model's performance on the data it was trained on. It results is an ideal scenario, since the training confusion matrix have a high number of data points on the diagonal (zero on the secondary diagonal), indicating that the model learned the patterns in the training data perfectly and can classify those samples correctly. The confusion matrix for the testing stage — Figure 2 (b) — shows the model's performance on unseen data. This is more important than the training stage's performance, as it generalizes how well the model will perform on new data. On average, 0.4 samples from the *Vigna unguiculata* L. Walp. class were mistakenly classified as being from the *Phaseolus vulgaris* L. class. While on average 0.6 samples of the *Phaseolus vulgaris* L. species were mistakenly classified as being of the *Vigna unguiculata* L. Walp. species. In general, there were more classifications in the correct classes (species) than confusions in the classified ones, even for unknown data in the testing stage.

Using the best performing model based on the results illustrated in Figures 1 and 2, the importance of each attribute (variable) is calculated as shown in Figure 3. When building a decision tree, as part of the machine learning process, the importance of attributes is calculated to determine which characteristics have the greatest influence on the model's decision making. Gini importance is calculated by measuring how much each attribute improves the purity of the tree nodes (Tangirala, 2020). For each split in a node, the algorithm calculates the Gini gain, which is the difference between the Gini index of the parent node and the weighted sum of the Gini indexes of the child nodes. The Gini importance of an attribute is then determined by summing the Gini gains for all divisions in which the attribute is used (Sivagama Sundhari, 2011). In Figure 3 we note that the Secondary color and Seed halo variables are not important for classifying the samples into the species used here. Comparing these results with the values collected in the experiment (Table 1) it is not difficult to understand why. As we can see, the Secondary color variable has several null values for both species. While for the variable Seed halo all values are equal to "Present" except for the genotype "Branco Dorama". Therefore, the variance of these attributes is very low, and therefore, they do not contribute to the construction of the decision tree model.

Finally, in Figure 4 we have the decision tree model displayed as a flowchart. This model is the one that presented the greatest accuracy in cross-validation. The flowchart is an adaptation of the model generated by the scikit-learn package (Pedregosa et al., 2011), as it only works with numerical data for attributes. Then it was necessary to convert the cutoff values, returned as real numbers, to the qualitative values detailed in Section 2.2. We can notice that the variables with the greatest importance were used in it, according to Figure 4. Although Seed shape is the variable with the highest importance value, it appears on the second level of the flowchart. However, it is decisive in determining the class (species), as there is no level after it. We also noticed that the least important variable, namely Degree of seed flattening, did not appear in the selected model. Using this flowchart model, we were able to determine the class (specie) of any of the samples presented in Table 1 with 100% accuracy.

It is necessary to emphasize that although the model obtained can be generalized to other data, including other cultivars, the same accuracy cannot be guaranteed. This is an expected characteristic of this machine learning model, as the dataset used for training is not extensive.

And this generalization deficiency is reflected in the results in Figures 1 and 2, where we noticed that for certain folds the performance metrics were low.

We conclude that the model obtained is viable to be executed for bean classification and can be used for various purposes. From didactic purposes in teaching environments, through practical applications such as use on farms to determine whether a sample is presenting the appropriate characteristics, to breeding programs where the model can be used to determine whether a certain sample of a species is presenting qualitative characteristics common to the other.

5. References

Abebe, B. K., & Alemayehu, M. T. (2022). A review of the nutritional use of cowpea (Vigna unguiculata L. Walp) for human and animal diets. *Journal of Agriculture and Food Research*, *10*, 100383. https://doi.org/10.1016/J.JAFR.2022.100383

Aguilera, J. G., Barbosa, E., Ribeiro, E., Vidal Do Nascimento, A. C., Silva, M. V., Dos, R., Carvalho, S., Santos Cocco, A., Fernanda, A., Barreto, S., Martins, G. S., Pereira Barcelos, R., Augusto, J., Rodrigues, S., Steiner, F., & Martins Bardiviesso, D. (2023). *Qualitative and quantitative descriptors for quantifying the genetic diversity of bean seeds*. https://doi.org/10.46420/TAES.e230001

Aguilera, J. G., Marim, B. G., Setotaw, T. A., Zuffo, A. M., Nick, C., & da Silva, D. J. H. (2019). The combination of data as a strategy to determine the diversity of tomato subsambples. *Amazonian Journal of Plant Research*, *3*(1), 276–289. https://doi.org/10.26545/ajpr.2019.b00035x

Boukar, O., Abberton, M., Oyatomi, O., Togola, A., Tripathi, L., & Fatokun, C. (2020). Introgression Breeding in Cowpea [Vigna unguiculata (L.) Walp.]. *Frontiers in Plant Science*, *11*, 567425. https://doi.org/10.3389/FPLS.2020.567425/BIBTEX

Cabral, P. D. S., Soares, T. C. B., Lima, A. B. de P., Alves, D. de S., & Nunes, J. A. (2011). Diversidade genética de acessos de feijão comum por caracteres agronômicos. *Revista Ciência Agronômica*, *42*(4), 898–905. https://doi.org/10.1590/S1806-66902011000400011

Catarino, S., Brilhante, M., Essoh, A. P., Charrua, A. B., Rangel, J., Roxo, G., Varela, E., Moldão, M., Ribeiro-Barros, A., Bandeira, S., Moura, M., Talhinhas, P., & Romeiras, M. M. (2021). Exploring physicochemical and cytogenomic diversity of African cowpea and common bean. *Scientific Reports 2021 11:1*, *11*(1), 1–14. https://doi.org/10.1038/s41598-021-91929-2

Coelho, C. M. M., Coimbra, J. L. M., Souza, C. A. de, Bogo, A., & Guidolin, A. F. (2007). Diversidade genética em acessos de feijão (Phaseolus vulgaris L.). *Ciência Rural*, *37*(5), 1241–1247. https://doi.org/10.1590/S0103-84782007000500004

de Oliveira, B. R., da Silva, A. A. P., Teodoro, L. P. R., de Azevedo, G. B., de Oliveira Sousa Azevedo, G. T., Baio, F. H. R., Sobrinho, R. L., da Silva Junior, C. A., & Teodoro, P. E. (2021). Eucalyptus growth recognition using machine learning methods and spectral variables. *Forest Ecology and Management*, *497*, 119496. https://doi.org/10.1016/J.FORECO.2021.119496

de Oliveira, B. R., Zuffo, A. M., Steiner, F., Aguilera, J. G., & Gonzales, H. H. S. (2023). Classification of soybean genotypes during the seedling stage in controlled drought and salt stress environments using the decision tree algorithm. *Journal of Agronomy and Crop Science*, *209*(5), 724–733. https://doi.org/10.1111/jac.12654

de Sousa Leite, W., de Souza Miranda, R., de Moura Rocha, M., Dutra, A. F., Santos, A. S., da Silva, A. C., de Brito, F. M., de Sousa, R. S., de Araújo, A. S. F., do Nascimento, C. W. A., Zuffo, A. M., & de Alcântara Neto, F. (2023). Silicon alleviates drought damage by increasing antioxidant and photosynthetic performance in cowpea. *Journal of Agronomy and Crop Science*, *209*(6), 772–787. https://doi.org/10.1111/jac.12659

Didinger, C., Foster, M. T., Bunning, M., & Thompson, H. J. (2022). Nutrition and Human Health Benefits of Dry Beans and Other Pulses. *Dry Beans and Pulses*, 481–504. https://doi.org/10.1002/9781119776802.CH19

Elsayed, A. Y. A. M., Hassan, B. A. A., Hassanin, A. A., Zyada, H. G., Ismail, H. E. M., & Aguilera, J. G. (2023). Selection parameters for improvement of yield and quality in tomatillo. *Ciência e Agrotecnologia*, *47*. https://doi.org/10.1590/1413-7054202347013722

Enyiukwu, D. N., Chukwu, L. A., & Bassey, I. N. (2020). Nutrient and anti-nutrient compositions of cowpea (Vigna unguiculata) and mung bean (Vigna radiata) seeds grown in humid Southeast Nigeria: A comparison. *International Journal of Tropical Drylands*, *4*(2). https://doi.org/10.13057/tropdrylands/t040202

Fletcher, S., & Islam, Md. Z. (2020). Decision Tree Classification with Differential Privacy. *ACM Computing Surveys*, *52*(4), 1–33. https://doi.org/10.1145/3337064

Gorunescu, F. (2011). *Classification Performance Evaluation* (pp. 319–330). https://doi.org/10.1007/978-3-642-19721-5_6

Guimarães, J. B., Nunes, C., Pereira, G., Gomes, A., Nhantumbo, N., Cabrita, P., Matos, J., Simões, F., & Veloso, M. M. (2023). Genetic Diversity and Population Structure of Cowpea (Vigna unguiculata (L.) Walp.) Landraces from Portugal and Mozambique. *Plants*, *12*(4), 846. https://doi.org/10.3390/PLANTS12040846/S1

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627–635.

Haykin, S. (2009). *Neural Networks and Learning Machines* (3 rd). Pearson Prentice Hall.

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-Label Confusion Matrix. *IEEE Access*, *10*, 19083–19095. https://doi.org/10.1109/ACCESS.2022.3151048

Kohavi, R. (1995). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. 1137–1143.

Kubat, M. (2021). An Introduction to Machine Learning. *An Introduction to Machine Learning*, 1–458. https://doi.org/10.1007/978-3-030-81935-4/COVER

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*. https://doi.org/10.3390/s18082674

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 431–439). Curran Associates, Inc. http://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-trees.pdf

Maia, F. R. (2023). Cowpe bean production - (Vigna Unguiculata (l.) walp). A drought-resistant plant is very common in regions of the brazilian semi-arid. *Journal of Interdisciplinary Debates*, *4*(04), 242–263. https://doi.org/10.51249/jid.v4i04.1731

Marin, D. B., Santana, L. S., Barbosa, B. D. S., Barata, R. A. P., Osco, L. P., Ramos, A. P. M., & others. (2021). Detecting coffee leaf rust with UAV-based vegetation indices and decision tree machine learning models. *Computers and Electronics in Agriculture*, *190*. https://doi.org/10.1016/j.compag.2021.106476

Ministry of Agriculture. (2009). *Ministry of Agriculture, Livestock and Supply. Rules for seed analysis. Ministry of Agriculture, Livestock and Supply.* . Secretariat for Agricultural Defense. Brasília: Mapa/A.

Nalini Durga, S., & Usha Rani, K. (2020). *A Perspective Overview on Machine Learning Algorithms* (pp. 353–364). https://doi.org/10.1007/978-3-030-46939-9_30

Özkan, G., Haliloğlu, K., Türkoğlu, A., Özturk, H. I., Elkoca, E., & Poczai, P. (2022). Determining Genetic Diversity and Population Structure of Common Bean (Phaseolus vulgaris L.) Landraces from Türkiye Using SSR Markers. *Genes*, *13*(8), 1410. https://doi.org/10.3390/GENES13081410/S1

Parmley, K. A., Higgins, R. H., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2019). Machine Learning Approach for Prescriptive Plant Breeding. *Scientific Reports*, *9*(1), 17132. https://doi.org/10.1038/s41598-019-53451-4

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Romero, J. (1961). *Varieties of Judias grown in Spain*. Ministry of Agriculture.

Sampaio, A. P. L., Aguilera, J. G., Mendes, A. M. da S., Argentel-Martínez, L., Zuffo, A. M., & Teodoro, P. E. (2023). The role of the genetic diversity of Capsicum spp. in the conservation of the species: Qualitative and quantitative characterization. *Ciência e Agrotecnologia*, *47*. https://doi.org/10.1590/1413-7054202347009122

Silva, H. T. (2005). *Minimum descriptors to characterize cultivars/varieties of common bean (Phaseolus vulgaris L.)*. Documentos 184. Embrapa Arroz e Feijão.

Singh, B. B. (2015). Cowpea: The Food Legume of the 21st Century. *Cowpea: The Food Legume of the 21st Century*, 1–166. https://doi.org/10.2135/2014.cowpea

Sivagama Sundhari, S. (2011). A knowledge discovery using decision tree by Gini coefficient. *2011 International Conference on Business, Engineering and Industrial Applications*, 232–235. https://doi.org/10.1109/ICBEIA.2011.5994250

Tangirala, S. (2020). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm*. *International Journal of Advanced Computer Science and Applications*, *11*(2). https://doi.org/10.14569/IJACSA.2020.0110277

Tariq, A., Yan, J., Gagnon, A. S., Riaz Khan, M., & Mumtaz, F. (2023). Mapping of cropland, cropping patterns and crop types by combining optical remote sensing images with decision tree classifier and random forest. *Geo-Spatial Information Science*, *26*(3), 302–320. https://doi.org/10.1080/10095020.2022.2100287

Tavares, T. C. de O., Sousa, S. A. de, Lopes, M. B. S., Veloso, D. A., & Fidelis, R. R. (2018). Divergência genética entre cultivares de feijão comum cultivados no estado do tocantins. *REVISTA DE AGRICULTURA NEOTROPICAL*, *5*(3), 76–82. https://doi.org/10.32404/rean.v5i3.1892

Vilela Barros, P. P., González Aguilera, J., Rodrigues Molina Rezende, J., Cordeiro Taveira, A., Costa Martins, W., Silva Abreu, M., Mario Zuffo, A., & Argentel Martínez, L. (2020). Diversidade Genética Entre Acessos de Mandioca Por Meio de Caracteres Agronômicos. *Ensaios e Ciência C Biológicas Agrárias e Da Saúde*, *24*(1), 29–35. https://doi.org/10.17921/1415-6938.2020v24n1p29-35

Watare, G. W. (2023). *Marker assisted Selection for resistance to bean common Mosaic Necrosis virus In French Bean cultivars in Kenya* [Http://repository.embuni.ac.ke/handle/embuni/4256]. University of Embu.

Wu, X., Wang, B., Wu, S., Li, S., Zhang, Y., Wang, Y., Li, Y., Wang, J., Wu, X., Lu, Z., & Li, G. (2021). Development of a core set of single nucleotide polymorphism markers for genetic diversity analysis and cultivar fingerprinting in cowpea. *Legume Science*, *3*(3), e93. https://doi.org/10.1002/LEG3.93

Yeganeh-Bakhtiary, A., EyvazOghli, H., Shabakhty, N., Kamranzad, B., & Abolfathi, S. (2022). Machine Learning as a Downscaling Approach for Prediction of Wind Characteristics under Future Climate Change Scenarios. *Frontiers in Data-Driven Methods for Understanding, Prediction, and Control of Complex Systems 2022*, *2022*(Special Issue).

Yoosefzadeh Najafabadi, M., Hesami, M., & Eskandari, M. (2023). Machine Learning-Assisted Approaches in Modernized Plant Breeding Programs. *Genes*, *14*(4), 777. https://doi.org/10.3390/genes14040777

6. Additional Information

6.1 Acknowledgments

We thank the State University of Mato Grosso do Sul Cassilândia Unit for their support.

6.2 Funding

There were no funds sponsoring this research.

6.3 Conflicts of Interest

We declare that there is no conflict of interest.

6.4 Data availability

The computational scripts and data can be accessed via the link: https://github.com/brunobro/a-qualitative-decision-tree-model-for-common-beans-and-cowpea-classification