Engenharia Agrícola

ISSN: 1809-4430 (on-line)

www.engenhariaagricola.org.br

Special Issue: Artificial Intelligence

Scientific Paper

Doi: http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v42nepe20210101/2022

MACHINE LEARNING FOR SOYBEAN SEEDS LOTS CLASSIFICATION

Gizele I. Gadotti[1], Carla A. Ascoli[1], Ruan Bernardy[1*],
Rita de C. M. Monteiro[1], Romário de M. Pinheiro[1]

[1*]Corresponding author. Universidade Federal de Pelotas/ Pelotas - RS, Brasil.
E-mail: ruanbernardy@yahoo.com.br | ORCID: https://orcid.org/0000-0001-9285-1993

KEYWORDS

artificial intelligence, agriculture, quality control.

ABSTRACT

The seed germination and vigor evaluation are essential for the sowing sector to measure the performance of different seed lots and improve the efficiency of storage and sowing processes. However, the analysis of various tests to determine seed quality generates a large amount of information, making it almost impossible for humans to perform a quick and effective quality control analysis. Therefore, the objective of this study was to evaluate the differences in the physiological quality of soybean seeds in different cultivars using machine learning techniques to rank the lots based on their quality. Three cultivars were used, and the analysis was germination, accelerated aging, tetrazolium treatment, seedling emergence, and 1000 seed weight from 65 lots were measured. The lots were evaluated in two phases, one immediately after harvest and the other after six months of storage. Random forest, multi-layer perceptron, J48, and classification via regression classifiers were used, aided by the feature resampler technique. Random forest and classification via regression obtained the highest accuracy, and the random forest technique obtained the best results. Therefore, it is possible to classify soybean seed lots with great accuracy and precision using artificial intelligence and machine learning techniques.

INTRODUCTION

Analyzing seed quality tests generates such a massive amount of information that it becomes almost impossible for humans to rapidly and effectively analyze such information quickly in a quality control laboratory (Pinheiro et al., 2021). Therefore, erroneous results may result in economic losses for seed companies.

Seed quality standards are required, with minimum legal requirements, and companies must perform internal control tests that generate important information. It can lead to copious data generated during an agricultural harvest depending on the company's size.

Based on this demand, seed technology research has focused on identifying aspects associated with the ranking of lots based on the physiological potential of the seeds. One tool that has attracted the attention of researchers is the use of machine learning and artificial intelligence to rank lots.

Data mining techniques consist of methods and classifications that generate more accurate information, where patterns are automatically extracted from the dataset (Reddy, 2021; Cardoso & Machado, 2008). Thus, data mining has emerged as an important tool for predicting the physiological quality of seeds.

Data generated during quality control tests of lots must be evaluated by adapting responses using machine learning techniques to reduce the time and resources spent on repetitive laboratory tests. Therefore, there is a need to streamline analyzing the large amount of data generated during the characterization of seed quality.

The objective of the present study was to evaluate the differences in the physiological quality of soybean seeds in different cultivars using machine learning techniques to rank the lots based on their quality that was evaluated immediately following harvest and after six months of storage.

[1] Universidade Federal de Pelotas/ Pelotas - RS, Brasil.

Area Editor: Adão Felipe dos Santos
Received in: 6-28-2021
Accepted in: 9-27-2021

MATERIAL AND METHODS

The study was conducted at the Internal Laboratory of Seed Analyzes of a company located in Sinop, Mato Grosso do Sul state, Brazil. Seeds cultivars were provided by the company using their genetic material for cultivation in the state of Mato Grosso and were produced during the 2018/19 harvest. The cultivars were classified as shown in Table 1.

TABLE 1. Cultivars, number of lots, sieves (mm), and the total of lots.

Cultivars	Number of lots	Sieves (mm)	Total of lots
A	36	6,0	45
	7	6,5	
	2	7,0	
B	22	6,0	36
	11	6,5	
	3	7,0	
C	7	6,5	12
	1	7,0	
	4	7,5	

The seeds were evaluated in two stages. The first stage was immediately following harvest and the second was after six months of storage. The storage conditions were those practiced by the company: refrigerated environment maintained at 13 °C and 60% relative humidity. Germination, accelerated aging, tetrazolium, seedling emergence, and 1000 seed weight tests were measured.

To determine the 1000 seed weight, eight replicates were used with 100 seeds that were weighed on an analytical balance and the seed weight was calculated according to the Rules for Seed Analysis (RAS) (Brasil, 2009). The moisture content was determined using the incubator method at 105 ± 3 °C for 24 h, using two subsamples of 5 g of seeds from each lot.

The germination test was conducted using four subsamples of 50 seeds per treatment sown in a Germitest paper roll moistened with water at a ratio of 2.5 × the mass of the paper. The rolls were maintained in a germinator set at 25 °C, and the evaluations followed the criteria established by the RAS (Brasil, 2009). The following germination aspects were considered: normal vigorous seedlings, normal weak seedlings, abnormal seedlings, dead seeds, and hard seeds. Normal vigorous seedlings and normal weak seedlings were the company protocols.

The accelerated aging test was performed with four subsamples of 50 seeds placed on an aluminum screen distributed in a single layer in plastic boxes containing 40 mL of distilled water. They were then placed in an incubator at a constant temperature of 41 °C for 48 h (Marcos Filho, 1999). After the aging period, the seeds were subjected to the germination test, according to the RAS (Brasil, 2009), with a single evaluation performed at five days and the percentage of normal seedlings calculated.

For the tetrazolium test, the procedure described by França Neto et al. (1988) was followed using two subsamples of 50 seeds that were preconditioned in moistened paper rolls and maintained under these conditions for 16 h at 25 °C. These samples were subsequently placed in plastic cups, submerged in tetrazolium solution (0.075%), and kept at a temperature of 35 °C for 180 min in the dark. After reaching the perfect color, the seeds were washed and sectioned longitudinally through the center of the embryo axis and classified according to vigor, viability, moisture, mechanical, and stink damage.

During the analysis of seedling emergence in the sand, four subsamples of 100 seeds per lot were sown at a 5 cm depth with a spacing of 40 cm between rows. In addition, those seedlings that emerged 14 days after sowing were counted (Brasil, 2009).

The data generated using these vigor tests were utilized for the machine learning technique, with 93 rows containing 42 attributes considered for the supervised machine learning training database (Table 2), with 54 lots accepted for commercialization, 27 rejected, and 12 termed intermediate.

TABLE 2. Description of attributes analyzed using data mining.

Attribute	Description	Value
Cultivar	Cultivar	{A, B, C}
Lots	Lots	{1-26}
Sieves	Sieves	{6,6.5,7}
Thousand of seed weight	TSW	{0-∞}
Moisture content	MC	{0-∞}
Tetrazolium initial	Vigor	{0-100}
	Viability	
	Moisture damage	
	Mechanical damage	
	Bug damage	
Accelerated aging initial	Vigor	{0-100}
	Normal Vigorous	
	Normal Weak	
	Abnormal	
	Dead	
	Hard	
Germination initial	Vigor	{0-100}
	Normal Vigorous	
	Normal weak	
	Abnormal	
	Dead	
	Hard	
Emergence in sand initial	Sand	{0-100}
Tetrazolium final	Vigor	{0-100}
	Viability	
	Moisture Damage	
	Mechanical Damage	
	Bug Damage	
Accelerated aging final	Vigor	{0-100}
	Normal Vigorous	
	Normal Weak	
	Abnormal	

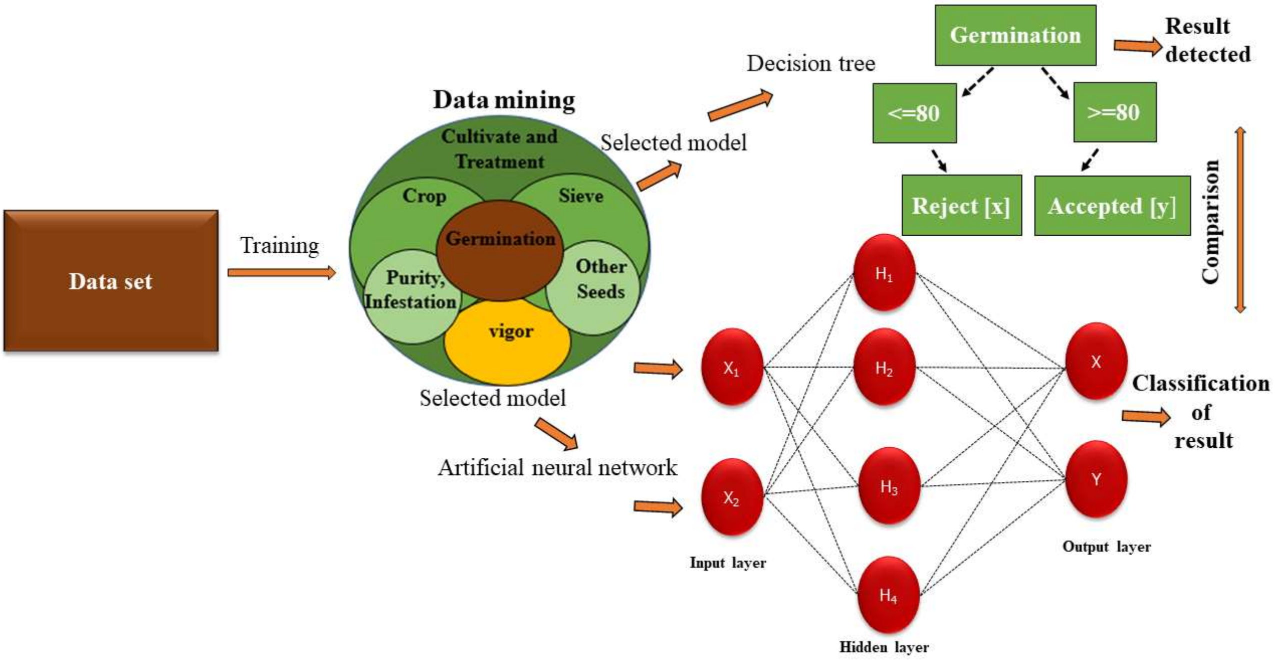
	Dead	
	Hard	
Germination final	Normal Vigorous	{0-100}
	Normal Weak	
	Abnormal	
	Dead	
	Hard	
	Normal Vigorous	
Emergence in sand final	Sand	{0-100}
Classification of lot	Decision	{accept, rejected, intermediary}

For the processing and prediction of lots, the data had to be first preprocessed so that the tool could perform the correct reading and analysis. For this step, the data were obtained in .xls format, and all attributes were single row; each value in columns below its respective attribute. Subsequently, the file was converted to .csv format and the dataset was executed using Microsoft Windows Notepad software, replacing the “commas”, when the assigned value was a decimal (number with commas), with “points” and “semicolons”, which divide the columns of attributes into “commas”. Rows with missing values or data considered erroneous were excluded from this process.

Four classifiers were used for data mining: J48, random forest, classification via regression (CVR), and multi-layer perceptron (MLP). The initial procedure was cross-validation, where the dataset, training, and test were divided into 10 subsets. This technique reduced the likelihood that coincidences underestimated or overestimated the performance of a given configuration. During the cross-validation without data duplication, the “Resample” filter was used to randomly produce a subsample of the dataset evaluated using sampling and maintain the distribution of classes toward a uniform distribution (Witten et al., 2011). The software informed of the ideal number of repetitions for training so the classifier could demonstrate its maximum performance to classify the dataset.

Weka software, version 3.8.5, developed by the University of Waikato, was used for the data mining (Eibe et al., 2020). When choosing which algorithms would be the most accurate, the following evaluation metrics were used: accuracy, precision, recall, F-measure, and area under a receiver operating characteristic (ROC) curve according to the methodology described by Lever et al. (2016). The values of true positives, false positives, true negatives, and false negatives extracted from the confusion matrix were used to calculate the recall and precision metrics using eqs (1) and (2) proposed by Medeiros et al. (2020). Finally, the best learning technique was determined based on the results obtained.

The process adopted can be better understood by the methodology described in Figure 1, demonstrating the steps adopted on the data generated in the quality control laboratory, where a set of data was formed that moved through the information treatment and proceeded to the training. Then, the data mining was established and, after performing the tests with the best algorithms, the values for decision-making were calculated.



Source: Pinheiro et al., 2021

FIGURE 1. Machine learning technique segmentation and selection process generating a decision tree and neural network in seed lots.

The data were also subjected to statistical procedures for comparing means using analysis of variance (ANOVA) and when a significant difference was found, the means were compared with Tukey's test at 5% significance.

RESULTS AND DISCUSSION

The evaluation of seed germination and vigor is essential for the sowing sector to measure the performance of different seed lots and improve the efficiency in storage and sowing processes, ensuring the crop's success. The selection of a high-yield seed lot results in germination close to 100% and vigor with a value near germination (Moraes, 2020). The applicability of traditional statistical methods in agricultural experimentation is mainly performed using a comparison of means (ANOVA), followed by a complementary test (e.g., Tukey's test) when significant results are obtained. It used various analyses or attributes for the seed sector's quality control of seed. For example, using ANOVA and Tukey's test makes selecting lots with several letters overlap or not differ from each other, as shown in Table 3; therefore, it is challenging to decide the best quality lots.

Thus, traditional statistical analyses make it difficult to decide on the classification of the vigor levels of seeds from several lots because it is common to have a high demand for evaluating seed lots in the seed industry. As shown in Table 3, the proposed statistical analysis does not allow the determination of lots with different vigor levels; therefore, the specialist needs to empirically establish the criteria for allocating lots, which may be for the disposal or commercialization of seeds. The pressure increases for the analyst. One lot has 30,000 kg, of approximately 750 bags worth US $70.00 each, and this value might directly influence the analyst's decision.

TABLE 3. Comparison of means using physiological performance tests of different cultivar seeds using four sieve sizes tested immediately following harvest and after six months of storage.

IEVES	TSW	MC	TZ VIGOR	TZ VIAB.	AA	AA NS	AA NW	AA AN	AA D	AA H	G	G NS	G NW	G AN	G D	G H	E
Initial Evaluation																	
6,00	138,45 a	12,76 a	91,17 a	95,27 a	81,36 a	69,72 a	11,63 ab	5,06 a	13,58 a	0,00	90,31 a	80,5 a	8,55 a	4,20 a	5,75 a	0,00	89,53 a
6,50	176,36 b	12,80 a	94,00 a	96,52 a	84,24 a	72,60 a	11,64 ab	5,08 a	11,32 a	0,00	91,72 a	82,64 a	9,08 a	3,52 a	4,68 a	0,00	91,40 a
7,00	197,07 c	12,71 a	94,16 a	97,00 a	88,66 a	71,16 a	17,50 b	7,33 a	5,66 a	0,00	95,66 a	86,33 a	9,33 a	1,16 a	3,16 a	0,00	94,16 a
7,50	212,02 d	12,72 a	93,75 a	96,25 a	89,75 a	81,00 a	8,75 a	7,25 a	3,00 a	0,00	96,25 a	89,5 a	6,75 a	3,00 a	0,75 a	0,00	94,75 a
Means	155,59	12,76	92,23	95,76	82,96	71,07	11,89	5,31	12,01	0,00	91,29	81,83	8,66	3,77	5,08	0,00	90,55
CV	4,66	4,26	5,86	3,44	13,29	19,47	44,63	52,1	87,17	0,00	7,73	16,95	67,9	99	114,15	0,00	6,37
dms	9,23	0,69	6,88	4,2	14,05	17,64	6,76	3,52	13,34	0,00	9	17,68	7,5	4,76	7,4	0,00	7,35
Final Evaluation (6 months of storage)																	
6,00	-	-	77,46 a	88,32 a	67,58 a	51,75 a	15,82 a	5,01 a	27,39 a	0,00	85,29 a	75,13 a	10,15 a	5,46 a	9,13 a	0,00	84,77 a
6,50	-	-	78,60 a	90,44 a	73,76 a	57,96 a	15,80 a	4,40 a	21,84 a	0,00	86,84 a	76,68 a	10,16 a	4,68 a	8,40 a	0,00	85,76 a
7,00	-	-	80,66 a	92,33 a	76,33 a	60,33 a	16,00 a	3,16 a	20,50 a	0,00	92,00 a	76,16 a	15,83 a	4,00 a	4,00 a	0,00	90,33 a
7,50	-	-	88,00 a	92,25 a	78,25 a	64,50 a	13,75 a	3,25 a	18,50 a	0,00	93,25 a	86,25 a	7,00 a	4,50 a	2,75 a	0,00	91,00 a
Means	-	-	78,43	89,32	70,26	54,52	15,74	4,65	25,07	0,00	86,48	76,09	10,38	5,11	8,33	0,00	85,66
CV	-	-	12,47	6,96	22,02	32,2	34,87	68,02	57,33	0,00	10,79	18,24	64,07	58,68	93,88	0,00	9,63
dms	-	-	12,47	7,92	19,72	22,38	6,99	4,03	18,32	0,00	11,89	17,69	8,48	3,82	9,97	0,00	10,51

Means followed by the same letter (uppercase in the column and lowercase in the row) do not differ by Tukey's test at 5% probability.
*TSW (thousand seeds weight), MC (moisture), TZ VIGOR (tetrazolium - vigor test), TZ VIAB. (tetrazolium test - viability), AA (accelerated aging), AA NS (accelerated aging - vigorous normal seedlings), AA NW (accelerated aging - weak normal seedlings), AA AN (accelerated aging - abnormal seedlings), AA D (aging accelerated - dead seeds), AA H (accelerated aging - hard seeds), G (germination pattern), G NS (germination pattern: vigorous normal seedlings), G NW (germination pattern: weak normal seedlings), G AN (germination pattern - abnormal seedlings), G D (germination pattern - dead seeds), G H (germination pattern - hard seeds), and E (emergence in sand).

The dataset evaluated in the present study using the machine learning models detected vigor levels based on germination performance. The suggested models achieved high mean accuracy values, above 60%, suggesting a highly significant predictive power. The training set for each model comprised 81.7% of data correctness for random forest and CVR, 79.6% for J48, and 74.2% for MLP, the latter being that with the lowest performance methods. In a study of predicted germination, Genze et al. (2020) found significant predictive power values > 90%; thus, obtaining a more accurate germination index using machine learning.

The evaluation components established regarding the stratification results of the soybean seed lots showed high values for detecting the physiological aspects of the seeds, where the random forest algorithm-generated 92.6% recall and 90.9% recall average accuracy relative to the actual test dataset for the accepted data class. This algorithm showed 92.6% recall and 73.59% accuracy for the reject class. The intermediate class exhibited an 8.3% recall and 25% accuracy (Table 4). Medeiros et al. (2020) studied an approach based on interactive and traditional machine learning methods to classify soybean seeds and seedlings based on their morphological characteristics and physiological potential. They obtained values with 93% precision, highlighting its good performance in classifying seeds based on their morphology (size, color, and damage). In studies of synthetic datasets for seed phenotyping, Toda et al. (2020) analyzed neural networks and obtained 96% recall and 95% average accuracy for the test dataset of the actual data.

The F-measure obtained mean values of recall and precision, facilitating the interpretation of only one metric instead of two or more (91%). Consequently, the classes with the highest values were accepted and rejected. The area under a ROC curve shows the relationship between the sensitivity and specificity of the classifier; the higher the value, the more adjusted the curve. However, when observing the data in Table 4 for classifiers J48 and MLP, the area under a ROC curve was higher for the reject class (0.94 and 0.89). Therefore, the ROC curve was better defined in the reject class than in the accepted class.

TABLE 4. Accuracy of the different algorithms used: recall (sensitivity), precision, ROC curve (receiver operating characteristic), and F-measure.

Classifiers	**Accuracy**				
	Recall	**Precision**	**ROC Area**	**F-Measure**	**Class**
Random Forest	0,926	0,909	0,974	0,917	Accepted
	0,926	0,735	0,978	0,820	Rejected
	0,083	0,250	0,904	0,125	Intermediary
Multi-Layer Perception (MLP)	0,907	0,875	0,934	0,891	Accepted
	0,741	0,741	0,941	0,741	Rejected
	0,000	0,000	0,597	0,000	Intermediary
J48	0,926	0,877	0,862	0,901	Accepted
	0,815	0,786	0,892	0,800	Rejected
	0,167	0,250	0,527	0,200	Intermediary
Classification Via Regression (CVR)	0,944	0,850	0,960	0,895	Accepted
	0,852	0,793	0,961	0,821	Rejected
	0,167	0,500	0,742	0,250	Intermediary

Hussain & Ajaz (2015) conducted a study on seed classification using Weka software and found 93.8% recall, 93.8% F-measure, and 98.9% ROC area. The 10-fold CVR classifier had 95.2% recall, 95.2% F-measure, and 99.6% ROC area using 10-fold MLP as a classifier, highlighting the good performance of the results found in Table 4.

The analysis of the decision trees generated by the CVR showed the practicality of understanding decision-making for the separation attributes of the lots. The best performance in the tests showed high vigor values through laboratory analyses in the quality control and segregation of sieves on seed sizes. Figure 2 shows the decision-making following a sequence defined by the best numerically expressed attribute obtained in the vigor tests.

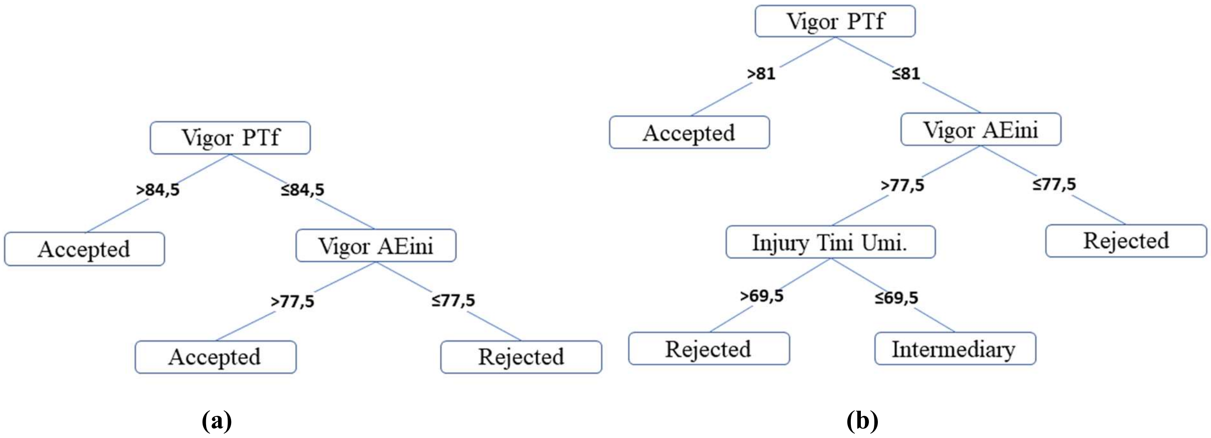


FIGURE 2. First decision tree from the CVR (A). Second decision tree from the CVR (B).

Germination is a standard test and is required by the Ministry of Agriculture, Livestock, and Supply (MAPA) as a mandatory item for seed commercialization in Brazil. The MAPA Normative Instruction 45 of 2013 established the commercialization pattern of soybean seeds, with at least 80% germination. Thus, the germination test is used to compare the quality of different lots and characterize the physiological quality, establish parameters for commercialization, and determine the sowing rate to determine the maximum germination potential (Marcos Filho, 2015).

From the decision tree of the first stage (Figure 2-A), the seed lots met the criteria pre-established by the supervisory body, with 84.5% accepted. Therefore, the initial quality of soybean seeds is fundamentally important for storage (Vergara et al., 2019b). From the vigor analysis, the physiological quality variability of seeds was observed, even in the production field; therefore, there was unevenness in the vigor of the same lots (Gazolla-Neto et al., 2015). However, the results generated by the decision tree met the criteria of the present study, and it was necessary to establish other standards for decision-making to analyze the parameters of each species or how the shape of the data should be treated.

The proposed methodology is adequate because the precision to discriminate seeds in their different classes of accepted lots was high, ranging from 0.85 to 0.90 accuracy between the machine learning models. However, human effort is still essential because manual tests must meet the standards (Boelt et al., 2018).

Another highlighted attribute was the tetrazolium test, the first criterion used in decision tree 2 (Figure 2-B). Tetrazolium analysis allows for the quick determination of seed viability, even for the most dormant seeds, compared to the germination test. It is crucial in the global seed market at present, where the industry requires reliable information on the viability of seed lots quickly to make decisions on seed commercialization and sowing (Soares et al., 2016).

We can still visualize other data in the tetrazolium treatment, such as moisture damage. Vergara et al. (2019a) stated that moisture-related damage significantly affected the physiological quality of soybean seeds suffering from a delayed harvest. In a study on soybean and precision agriculture, Vergara et al. (2019b) showed that fields with bug and moisture damage had low physiological quality and reduced protein levels.

In a study by Moraes (2020), the decision tree was based on vigor, although only one vigor test was used, suggesting a more significant number of vigor tests for future studies, as performed in the present study. Our study considered several vigor tests and the same tests were performed over two periods (initial and six months of storage). According to Tillmann et al. (2019), vigor tests are of fundamental importance for a more efficient classification of lots. There was also an increase of efficiency in the present study with greater accuracy in the “rejected” and “intermediate” classes because more attributes were obtained based on vigor.

The data from seed analyses are unbalanced, especially for companies with high lot quality (Table 3). A resample filter was used to solve the problem and not bias the algorithm. Unbalanced learning is a classification problem in which the number of observations of one class far exceeds that of another class. The subsampling technique is the best technique among conventional approaches to managing this problem (Sarada & Devi, 2019). Here, the feature resampler technique was used to resample the data. The selection of resources is vital because it decreases the dimensionality of the data and helps the classifier function faster, thus improving its accuracy (Sarada & Devi, 2019).

Oliveira et al. (2021), using fermented cocoa beans, also had unbalanced data and stated that the classes with more data had higher accuracy and precision values than the other classes. These classes would more often be classified incorrectly by classifiers sensitive to unbalanced data. If these classes are classified incorrectly, they influence the values of the performance metrics of their respective classes owing to their small number.

Detailing the accuracy resulting from the classification obtained by the algorithms, we found that the CVR had a lower rate of false positives in the three classes and greater accuracy in the rejected and intermediate classes. These latter classes are the most complex for making decisions. Thus, the higher the accuracy, the better, and we obtained 79% accuracy in the rejected class, which was promising.

CVR is a classification method that can transform problems into regression functions (Yu-Xun et al., 2014). This method combines the principles of the decision tree algorithm and linear regression in several constructed subtrees (leaves) and involves two main steps. First, an ordinary decision tree is delimited, maximizing the separation of criteria/parameters/attributes and their variations based on the target/output values. This was achieved by calculating the deviation reduction. Then, subdivisions of this tree are placed into several possible subtrees and, according to the regression function (linear model), usually in the leaves (Arora & Dhir, 2017). The data in the present study were quantitative; therefore, regression combined with the decision tree was a more assertive solution.

According to Pinheiro et al. (2021), the datasets used in machine learning training are usually enormous; therefore, manual analysis generates time-consuming responses. Furthermore, when information on cultivar, treatments, purity, germination, and other quality attributes is generated, the work becomes slow and inefficient decision-making. Therefore, testing classifier models is essential to match the algorithm's performance for the dataset provided (Pinheiro et al., 2021).

According to Jha et al. (2019), the only purpose of machine learning is to feed a system with data from previous experiences and statistical values to perform its assigned task and thus solve a specific problem. Therefore, machine learning is a mathematical approach to building intelligent systems.

CONCLUSIONS

It was possible to classify numerous soybean seeds with great accuracy and precision using artificial intelligence and machine learning techniques. The best algorithms were random forest and CVR when applying the machine learning technique. In addition, the feature resampler technique was necessary for solving the data imbalance problem.

REFERENCES

Arora T, Dhir R (2017) Correlation-Based Feature Selection and Classification Via Regression of Segmented Chromosomes Using Geometric Features Medical & biological engineering & computing 55(5): 733-745.

Boelt B, Shrestha S, Salimi Z, Jørgensen J, Nicolaisen M, Carstensen J (2018) Multispectral imaging - A new tool in seed quality assessment? Seed Science Research 28 (13): 222-228. DOI: https://doi.org/10.1017/S0960258518000235.

Brasil (2009) Regras para análise de sementes. Ministério da Agrícultura, Pecuária e Abastecimento, sistema BINAGRI-SISLEGIS, instrução normativa – IN 6/2009 de 18 de fevereiro de 2009.

Cardoso O, Machado R (2008) Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. Revista de Administração Pública 42 (3): 495-528.

Eibe F, Mark AH, Ian HW (2020) The WEKA Workbench. Online appendix for data mining: practical machine learning tools and techniques. Burlington, Morgan Kaufmann.

França Neto JB, Pereira LAG, Costa NP, Krzyzanowski FC, Henning AA (1988) Metodologia do teste de tetrazólio em sementes de soja. Londrina, Embrapa Soja. Empresa Brasileira de Pesquisa Agropecuária. 60p.

Gazolla-Neto A, Fernandes MC, Gomes AD, Gadotti GI, Villela FA (2015) Distribuição espacial da qualidade fisiológica de sementes de soja em campo de produção. Revista Caatinga 28 (12): 119-127 DOI: https://doi.org/10.1590/1983-21252015v28n314rc.

Genze N, Bharti R, Grieb M, Schultheiss, SJ, Grimm D (2020). Accurate machine learning-based germination detection, prediction and quality assessment of three grain crops. Plant Methods 16 (12). DOI: https://doi.org/10.1186/s13007-020-00699-x.

Hussain L, Ajaz R. (2015) Seed Classification using Machine Learning Techniques. Journal of Multidisciplinary Engineering Science and Technology 2(4): 1098-1102.

Jha, K, Doshi, A, Patel, P, Shah, M. (2019) A comprehensive review on automation in agriculture using artificial intelligence. Artificial Intelligence in Agriculture 2: 1-12. Doi: https://doi.org/10.1016/j.aiia.2019.05.004.

Lever J, Krzywinski M, Altman N (2016) Classification evaluation. Nat Methods 13 (8): 603-604. DOI: https://doi.org/10.1038/nmeth.3945.

Marcos Filho J (1999). Teste de envelhecimento acelerado. In: Krzyzanowski FC, Vieira RD, FrançaNeto JB. Vigor de sementes: conceitos e testes. Associação brasileira de tecnologia de sementes, Comitê de vigor de sementes, p1-24.

Marcos Filho J (2015) Seed vigor testing: an overview of the past, present and future perspective. Scientia Agricola 72 (12): 363-374. DOI: https://doi.org/10.1590/0103-9 016-2015-0007.

Medeiros AD, Capobiango NP, Silva JM da (2020) Interactive machine learning for soybean seed and seedling quality classification. Scientific Reports 10 (8). DOI: https://doi.org/10.1038/s41598-020-68273-y.

Moraes NAB (2020) Predição de ranqueamento de lotes de sementes de milho por inteligência artificial. Dissertação, Universidade Federal de Pelotas, Faculdade de Agronomia Eliseu Maciel.

Oliveira MM, Cerqueira BV, Barbon S, Barbin DF (2021) Classification of fermented cocoa beans (cut test) using computer vision. Journal of Food Composition and Analysis 97 (14). DOI: https://doi.org/10.1016/j.jfca.2020.103771.

Pinheiro RM, Gadotti GI, Monteiro RCM, Bernardy R (2021) Inteligência artificial na agricultura com aplicabilidade no setor sementeiro. Diversitas Journal 6 (3): 2984-2995. DOI: https://doi.org/10.48017/Diversitas_Journal-v6i3-1857

Reddy PVS (2021) Data mining and fuzzy data mining using map reduce algorithms. Data Mining: Methods, Applications and Systems (3): 1-25. DOI: https://doi.org/10.5772/intechopen.92232.

Sarada C, Devi MS (2019) Imbalanced big data classification using feature selection under-sampling. CVR Journal of Science and Technology 17 (14): 78-82. DOI: https://doi.org/10.32377/cvrjst1714.

Soares VN, Elias SG, Gadotti GI, Garay AE, Villela FA (2016) Can the tetrazolium test be used as an alternative to the germination test in determining seed viability of grass species? Crop Science 56 (13): p.707-716. DOI: https://doi.org/10.2135/cropsci2015.06.0399

Tillmann MAA, Tunes LM, Almeida AS (2019). Análise de sementes. In: Peske ST, Villela FA, Meneghello GE. Sementes: fundamentos científicos e tecnológicos. Pelotas, Editora Universitária, p147-258.

Toda Y, Okuda F, Ito J (2020) Training instance segmentation neural network with synthetic datasets for crop seed phenotyping. Communications Biology 3 (9): 1-12 DOI: https://doi.org/10.1038/s42003-020-0905-5.

Vergara RO, Silva RNO, Nadal AP, Gadotti GI (2019a) Harvest delay, storage and physiological quality of soybean seeds. Journal Seed Science 41 (13): 506-513. DOI: https://doi.org/10.1590/2317-1545v41n4222413.

Vergara RO, Gazolla-Neto A, Gadotti GI (2019b) Space distribution of soybean seed storage potential. Revista Caatinga 32 (13): 399-410. DOI: https://doi.org/10.1590/1983-21252019v32n213rc.

Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers (3).

Yu-Xun R, Hsuan-Tien L, Ming-Feng T (2014). Improving Ranking Performance With Cost-Sensitive Ordinal Classification Via Regression. Information retrieval 17 (17): 1-20. DOI: https://doi.org/10.1007/s10791-013-9219-2