Engenharia Agrícola

ISSN: 1809-4430 (on-line)

www.engenhariaagricola.org.br

Scientific Paper
Doi: http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v42n4e20210005/2022

PREDICTION OF RANKING OF LOTS OF CORN SEEDS BY ARTIFICIAL INTELLIGENCE

Gizele I. Gadotti[1*], Nicacia A. B. Moraes[2], Joseano G. da Silva[2], Romário de M. Pinheiro[2], Rita de C. M. Monteiro[2]

[1*]Corresponding author. Centro de Engenharias, Universidade Federal de Pelotas/Pelotas-RS, Brasil.
E-mail: gizeleingrid@gmail.com | ORCID ID: https://orcid.org/0000-0001-9545-6577

KEYWORDS

quality control, classification, artificial intelligence, corn, data mining.

ABSTRACT

The seed sector faces several challenges when it comes to ensuring a quick and accurate decision making when working with large amounts of data on physiological quality of seed lots, which makes the process time-consuming and inefficient. Thus, artificial intelligence (AI) emerges as a new technological option in the seed sector to solve database problems in the post-harvest stages. This study aims to use machine learning to classify maize seed lots. Data were obtained from eight maize seed crops from a private company. These data were mined using the following classifiers: J48 (DecisionTree), RandomForest, CVR (ClassificationViaRegression), lBk (lazy.IBK), MLP (MultiLayerPercepton), and NäiveBayes. Cross-validation was used for data measurement, with the data set, including training and testing data, being divided into 10 subsets. The described steps were performed using the Weka software. It is concluded that results obtained allow the classification of maize seed lots with high accuracy and precision, and these algorithms can better classify the maize seed lot through vigor attributes, thus enabling more accurate decision making based on vigor tests on a reduced evaluation time.

INTRODUCTION

Seed producing companies use vigor tests associated with germination as a tool for internal quality control, estimating the potential performance in the field under favorable or adverse conditions. The decision to sell or discard the seed lots, which can contain a high number of seeds, is made based on these results (Grzybowski et al., 2015).

Seed technology, as a segment of the production process, has sought to improve tests that assess the physiological potential (germination and vigor) of seeds and show the potential performance of the lots under field conditions. The results obtained during the quality control of seed lots must comply with the minimum legal requirements, thus creating a database with lots of information.

The seed sector still faces several challenges when it comes to ensuring a quick and accurate decision making when working with large amounts of seed lots. In this sense, the demand for efficient and safe methods of food analysis, production, and marketing has been increasing. Information technology is one of the tools for this purpose (Patrício & Rieder, 2018). Artificial intelligence and machine learning can be applied for promoting sustainable development in the agricultural sector, facilitating the understanding of predictive models that support in different sectors of agriculture and contributing to the optimization of resources. Technological innovations are increasingly accessible in the field which opens the door to the integration of the seed quality analysis into technological processes.

The search for new techniques that contribute to the automation of seed classification and quality evaluation leads to the emergence of alternatives such as digital image processing and computer vision (Liu et al., 2020a). This has contributed to a more accurate evaluation of certain seed physiological quality markers. These markers include: seed size, evaluation of seed surface color space by image analysis, computer-assisted spectrometry, X-ray inspection combined with quantitative imaging, and improved detection of chlorophyll fluorescence signal by

[2] Faculdade de Agronomia Eliseu Maciel, Universidade Federal de Pelotas/Pelotas - RS, Brasil.

Area Editor: Ednilton Tavares de Andrade
Received in: 1-14-2021
Accepted in: 7-1-2022

laser technology, and the evaluation of seeds by thermographic images through infrared, among others (Dell'Aquila, 2009b; Arruda et al., 2016; Huang et al. 2015; Brunes et al. 2019; Monteiro et al., 2019; Xia et al., 2019; Aboukarima et al., 2020; Liu et al., 2020b; Medeiros, et al., 2020; Torres et al., 2020; Monteiro et al., 2021).

Today, machines and automation are required everywhere to process everyday tasks. Machine learning, a field of computer science and a central branch of AI that uses statistics to provide results, is one of these methods used to simplify everyday problems. It usually refers to the ability of a specific machine to learn from its previous results and algorithms, so that it can improve on its own not needing regular guidance to update its system. Moreover, machine learning defines steps to monitor the machine's performance, learning from its historical inputs. It focuses on the development of programs in computer systems that can access data and use it to learn by themselves (Pooja et al., 2018). In this context, this study aimed to use the machine learning technique for classifying the maize seed lots as to their physiological quality.

MATERIAL AND METHODS

Data were obtained from a private maize production company, considering eight crops (14/15, 15/15, 15/16, 16/16, 16/17, 17/17, 17/18, 18/ 18) and four cultivars (Table 1), totaling 5800 seed lots (rows). Legal issues (germination, purity, number of other seeds, percentage of infested seeds) and other attributes related to seed identification (material, sieve, harvest) were obtained (Brasil, 2009). The designation of the sieves followed the pattern of maize seeds, being R for round, F for flat, and other combinations between them (Strieder et al., 2014).

TABLE 1. Description of attributes analyzed by data mining.

Attribute	Description	Value
Material	Cultivate and Treatment	{A,B,D,E,A1,B1,D1,E1}
Sie	Sieves	{R2, R3, C3, R4, C2, C4, C1, R1, R2C, C2C, C3C, R3C, R4C, C4C, P0}
Harvest	Harvests	{14/15, 15/15, 15/16, 16/16, 16/17, 17/17, 17/18, 18/18}
Germination	Germination Test Result	{0-100}
Vigor	Vigot Test Result	{0-100}
%Infest	Percentage of Infested Seeds	{0-100}
NumOth	Number of Other Seeds	$\{0-\infty\}$
%Pure	Purity Test Percentage	{0-100}
Accept or Reject	Decision taken	{High vigor(Accept), Medium vigor(Accept/Reject), Low vigor(Reject), Hold}

Initially, there were discrepant data (outliers), incoherent data, and rows missing data. The training file included 80 Rejections (34%) and 237 Acceptances (66%) to obtain representative and balanced values (especially the rejected ones), considering the sample of 5000 lots. First, cultivar and harvest were analyzed separately. Subsequently, we analyzed each cultivar with all the harvests and then all the cultivars and crops combined.

The classifiers used were J48 (DecisionTree), which is easy to explain as it deals with non-linear data; RandomForest which is characterized by the creation of a set of decision trees; CVR (ClassificationViaRegression) which transforms problems into regression functions, combines the principles of decision tree algorithm with the principles of linear regression in several constructed subtrees (leaves), and delimitates an ordinary decision tree, separating criteria/parameters/attributes from their variations based on target/output values which were obtained by calculating deviation variance reduction. The subdivisions of this tree are placed in several possible subtrees according to the regression function (linear model), usually in the leaves (Arora & Dhir, 2017; Yu-Xun et al., 2014). lBk(lazy.IBK) is a distance weightier of K-nearest neighbors, selecting the appropriate value of K based on cross-validation; MLP (MultiLayerPercepton) is composed of an output layer and one or two intermediate layer; and NäiveBayes use independent data as it is a probabilistic classifier (Frank et al., 2016 and Patel & Kathiriya, 2017; Harrison, 2019).

Cross-validation was used for data validation, with the data set, including training and testing data, being divided into 10 subsets. The average accuracy corresponds to the algorithm's performance on the given dataset. This technique reduces the likelihood that duplicate values under or overestimate the performance for a given configuration. All results here reported use this technique. The described steps were performed using the Weka 3.8 software (Frank et al., 2016).

The accuracy and the confusion matrix from each model were considered when choosing the algorithms. After choosing the classifier model with the best accuracy in the training test, the model was used with all the data. The SimpleKMeans and FarthestFirst algorithms were used for clustering and the unsupervised evaluation.

RESULTS AND DISCUSSION

The classification technique was applied and the accuracy of the algorithms was determined (Table 2). The highest accuracy was obtained using the J48 technique and ClassificationViaRegression because they used a limited number of techniques. Their confusion matrices are identical as the data is sorted perfectly and without false positives or negatives.

The accuracy data are higher than that of Marko et al. (2017), who studied the selection of seed portfolios, since, as reported in several studies, the data have small variance and standard deviation (Vergara et al., 2019), which is confirmed by the tolerance tables of several rules such as ISTA, AOSA, and RAS (2009) with less than 10%. For Barbosa et al. (2020), the greater the variability the better, but when it comes to the seed sector this is not possible. Gadotti et al (2022) discusses this topic stating that in ANOVA tests, the number of letters resulting from the test confuses those who are classifying the lots.

TABLE 2. Accuracy of algorithms after classification.

Algorithms	Accuracy (%)
J48	100,0000
ClassificationViaRegression	100,0000
RandomForest	99,6845
lazy.IBk	90,8517
MultilayerPerceptron	98,1073
NaiveBayes	96,5300

The training data in Table 3 demonstrate that the number of accepted lots of maize seeds is high, with low rejection. Thus, it was expected that there would be a greater rigor when evaluating the vigor attribute, which occurred according to Figure 2. These values, even with more than 5000 evaluated lots, may have biased the result because they had few vigor attributes. Gadotti et al. (2022), when evaluating soybeans, obtained high results with less bias because they evaluated many vigor parameters.

TABLE 3. Confusion matrix of the J48 algorithm and Classification Via Regression.

Classified as	A	B
a = Accept	237	0
b = Reject	0	80

The J48 algorithm is a Java derivation of the C4.5 algorithm, one of the most used and reliable statistical classifiers. J48 builds the decision tree using the entropy concept. Through entropy, it chooses the attribute that most partitions the data through normalized information gain (Frank et al., 2016).

Figure 2 presents a generated decision tree, in which the attribute that most influences the classification of maize seed lots is the vigor attribute. However, a greater number of attributes related to vigor would be recommended for the Seed Science and Technology area, since, according to IN 45, the only attribute considered was germination (Brasil, 2013). Considering that vigor is not a standardized test and a single test was used in the present database, more vigor tests are necessary for a more efficient classification of lots, as already stated by Tillmann et al. (2019).

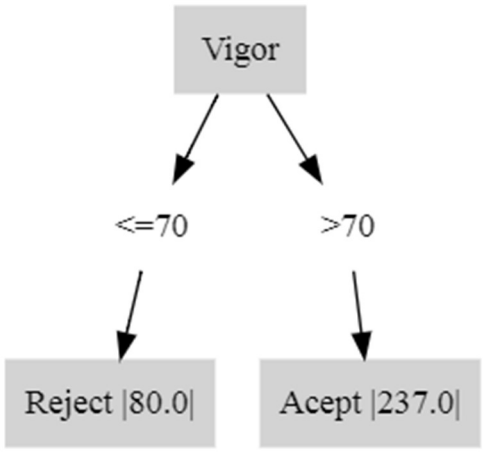

FIGURE 1. Decision tree for predicting maize seed lot classification.

This result is explained by the fact that the company has many years of control and information about the behavior of this species and all materials. The test used is a company secret and was not detailed. It is considered highly accurate by the company, thus enabling the company to have satisfactory and reliable results with just one test. However, as previously discussed, this is not congruent with the theory in the area. In this case, the company considers that using a single test is enough and provides good cost-benefit.

When the classification of lots was predicted without supervision, the FarthestFirst algorithm obtained greater prediction accuracy in the clusters (Table 4).

TABLE 4. Percentage of uncertainties of the algorithms after clustering.

		---------- Grouping ---------	
Algorithms	Uncertainty (%)	0	1
SimpleKMeans	32,8076	283 (89%)	34 (11%)
FarthestFirst	29,6530	281 (89%)	36 (11%)

During the training of the clusters, both algorithms behave in the same way, attributing only one lot to the wrong cluster, which can be considered as good accuracy (Table 5).

TABLE 5. Clusters generated from the classification of maize lots.

Classifier		
Simple K Means		
Cluster 0	Cluster 1	← assigned to cluster
0	283 (89%)	
1	34 (11%)	
FarthestFirst		
Cluster 0	Cluster 1	← assigned to cluster
0	281 (89%)	
1	36 (11%)	

There is a lot of uncertainty in the confusion matrices due to false positives (rejected in accepted) as shown in Table 6.

TABLE 6. Confusion Matrix in the cluster of SimpleKMeans and FarthestFirst algorithms.

Assigned to cluster	0 – Accept	1 – Reject	
Simple K Means	208	29	Accept
	75	5	Reject
FarthestFirst	212	25	Accept
	69	11	Reject

The most efficient algorithms according to the prediction were J48 and ClassificationViaRegression. The decision tree (J48) (Table 7) would be the most interesting option as it can be visualized, being more friendly for a data supervision later. Table 7. Accuracy of the J48 algorithm with all data. Cross-validation with 10 folds.

TABLE 7. Accuracy of the J48 algorithm with all data. Cross-validation with 10 folds. "?" decision-making of the analyst in loco *a posteriori*.

	Accuracy	Recall	ROC
Accept	0,988	0,999	0,881
Accept/Reject	0,619	0,356	0,967
Reject	0,667	0,154	0,696
Espera	?	0,000	0,056

=== Confusion Matrix ===

a	b	c	d	Classified as
5651	7	0	0	a = Accept
44	26	3	0	b = Reject/Accept
24	9	6	0	c = Reject
1	0	0	0	d = Esperar

FIGURE 2. Confusion Matrix of the J48 algorithm.

Gadotti et al. (2022), when evaluating soybean seeds, obtained better ROC results in the reject class. The area under the ROC curve shows the relationship between the sensitivity and specificity of the classifier; the higher the value, the more adjusted the curve. Therefore, the ROC curve here was better defined in the medium vigor class than in the high and low vigor classes.

TABLE 8. Accuracy of the J48 algorithm with all data. Cross-validation with 10 folds and 500 lots. "?" decision-making of the analyst in loco *a posteriori*.

	Accuracy	Recall	ROC
Accept	0,988	0,999	0,881
Accept/Reject	0,619	0,356	0,967
Reject	0,667	0,154	0,696
Espera	?	0,000	0,056

=== Confusion Matrix ===

a	b	c	d	Classified as
5651	7	0	0	a = Accept
44	26	3	0	b = Reject/Accept
24	9	6	0	c = Reject
1	0	0	0	d = Esperar

FIGURE 3. Confusion Matrix of the J48 algorithm with all data. Cross-validation with 10 folds and 500 lots.

TABLE 9. Accuracy of the J48 algorithm with all data. Cross-validation with 20 folds and 1000 lots. "?" decision-making of the analyst in loco *a posteriori*.

	Accuracy	Recall	ROC
Accept	0,988	**0,998**	**0,841**
Accept/Reject	**0,636**	**0,384**	**0,959**
Reject	0,667	0,154	**0,621**
Espera	?	0,000	0,031

=== Confusion Matrix ===

a	b	c	d	<-- classified as
5649	9	0	0	a = Accept
42	**28**	3	0	b= Reject/Accept
26	7	6	0	c= Reject
1	0	0	0	d= Esperar

FIGURE 4. Confusion Matrix of the J48 algorithm with all data. Cross-validation with 20 folds and 1000 lots.

TABLE 10. Accuracy of the J48 algorithm with all data. Cross-validation with 30 folds and 1000 lots. "?" decision-making of the analyst in loco *a posteriori*.

	Accuracy	Recall	ROC
Accept	0,988	**0,999**	**0,862**
Accept/Reject	**0,643**	0,370	**0,967**
Reject	0,545	0,154	**0,656**
Espera	?	0,000	0,031

=== Confusion Matrix ===

a	b	c	d	<-- classified as
5650	7	1	0	a = Accept
42	27	4	0	b= Reject/Accept
25	8	6	0	c= Reject
1	0	0	0	d= Esperar

FIGURE 5. Confusion Matrix of the J48 algorithm with all data. Cross-validation with 30 folds and 1000 lots.

TABLE 11. Accuracy of the J48 algorithm with all data. Cross-validation with 50 folds and 1000 lots. "?" decision-making of the analyst in loco *a posteriori*.

	Accuracy	Recall	ROC
Accept	0,988	**0,999**	**0,867**
Accept/Reject	0,605	0,356	0,965
Reject	**0,625**	0,128	0,650
Espera	?	0,000	0,018

=== Confusion Matrix ===

a	b	c	d	<-- classified as
5649	9	0	0	a = Accept
44	26	3	0	b = Reject/Accept
26	8	5	0	c = Reject
1	0	0	0	d = Esperar

FIGURE 6. Confusion Matrix of the J48 algorithm with all data. Cross-validation with 30 folds and 1000 lots.

It is possible to conclude from Tables 6 to 11 and Figures 2 to 6, which show the accuracy indices and confusion matrices obtained by the J48 algorithm, that increasing the number of folds was efficient up to 20 folds with 1000 lots. After that, it did not bring more efficiency to the data, with an increase in false positives.

When choosing seed lots, experts want germination to be close to 100% and vigor to be as close as possible to the percentage of germination.

In companies, it is extremely necessary that there are no errors in the classification of lots, because a low vigor lot can be released as a high vigor lot, which can bring losses and lack of credibility with the customer or a high vigor lot can be discarded for being a low vigor lot, bringing irreversible financial losses and compromising the company's image. In this sense, this technique can speed up the classification of seed lots with artificial intelligence techniques and reduce the human error associated with their classification.

CONCLUSIONS

Maize seed lots can be classified using artificial intelligence and machine learning techniques.

An even more consistent mining requires more attributes, such as vigor tests.

REFERENCES

Aboukarima A, El-Marazky M, Elsoury H, Zayed M, Minyawi M (2020) Artificial neural network-based method to identify five varieties of egyptian faba bean according to seed morphological features. Engenharia Agrícola 40(6):791-799. DOI: https://doi.org/10.1590/1809-4430-eng.agric.v40n6p791-799/2020

Arruda N, Cicero SM, Gomes-Junior FG (2016) Radiographic analysis to assess the seed structure of Crotalaria juncea L. Journal of Seed Science 38(2):161-168. DOI: http://dx.doi.org/10.1590/2317-1545v38n2155116.

Bansal R, Singh J, Kaur R (2019) Machine learning and its applications: A Review. Journal of Applied Science and Computations 6(6):1392-1398.

Barbosa A, Trevisan R, Hovakimyan N, Martin NF (2020) Modeling yield response to crop management using convolutional neural networks. Computers and Electronics in Agriculture 170:1-8. DOI: https://doi.org/10.1016/j.compag.2019.105197.

Brasil - Ministério da Agricultura, Pecuária e Abastecimento (2009) Regras para análise de sementes. Secretaria de Defesa Agropecuária. Brasília, DF. 399p.

Brasil - Ministério da Agricultura e do Abastecimento (2013) Instrução Normativa MAPA nº 45, 17 set. 2013. Padrões de identidade e qualidade para a produção e a comercialização de sementes. Diário Oficial da União, Brasília, Distrito Federal, 18 set. 2013, Seção 1, pag 13.

Brunes AP, Araújo ÁS, Dias LW, Antoniolli J, Gadotti GI, Villela FA (2019) Rice seeds vigor through image processing of seedlings. Ciência Rural 49(8):1-6. DOI: https://doi.org/10.1590/0103-8478cr20180107.

Dell' Aquila A (2009) Development of novel techniques in conditioning, testing and sorting seed physiological quality. Seed Science and Technology 37(3):608-624. DOI: https://doi.org/10.15258/sst.2009.37.3.10

Frank E, Hall MA, Witten IH (2016) The WEKA Workbench. Online appendix for "data mining: practical machine learning tools and techniques". Cambridge, Morgan Kaufmann.

Gadotti GI, Ascoli CA, Bernardy R, Monteiro RCM, Pinheiro RM (2022) Machine learning for soybean seeds lots classification. Engenharia Agrícola 42(spe):e20210101. DOI: https://doi.org/10.1590/1809-4430-Eng.Agric.v42nepe20210101/2022.

Grzybowski CRS, Vieira RD, Panobianco M (2015) Testes de estresse na avaliação do vigor de sementes de milho. Revista Ciência Agronômica 46(3):590-596. DOI: https://doi.org/10.5935/1806-6690.20150042.

Harrison M (2019) Machine Learning – Guia de referência rápida: trabalhando com dados estruturados em Python. São Paulo, Novatec Editora, p. 272.

Huang M, Wang QG, Zhu QB, Qin JW, Huang G (2015) Review of seed quality and safety tests using optical sensing technologies. Seed Science & Technology 43:337–366. DOI: https://doi.org/10.15258/sst.2015.43.3.16

Liu W, Liu C, Jin J, Li D, Fu Y, Yuan W (2020a) High-throughput phenotyping of morphological seed and fruit characteristics using x-ray computed tomography. Frontiers in Plant Science 11:601475 1-10.DOI: https://doi.org/10.3389/fpls.2020.60147

Liu L, Wang Z, Li J, Zhang X, Wang R (2020b) A Non-Invasive Analysis of Seed Vigor by Infrared Thermography. Plants 9(6):768. DOI: http://dx.doi.org/10.3390/plants9060768.

Marko O, Brdar S, Panić M, łałić I, Despotović D, Knežević M, Crnojević V (2017) Portfolio optimization for seed selection in diverse weather scenarios. Plos One 12(9):1-27. DOI: http://dx.doi.org/10.1371/journal.pone.0184198.

Medeiros ADD, Silva LJD, Ribeiro JPO, Ferreira KC, Rosas JTF, Santos AA, Silva CBD (2020) Machine Learning for Seed Quality Classification: An Advanced Approach Using Merger Data from FT-NIR Spectroscopy and X-ray Imaging. Sensors 20(15):1-12. DOI: https://doi.org/10.3390/s20154319

Monteiro RCM, Gadotti GI, Araújo AS (org.). (2019) Processamento de imagens para identificação de defeitos no arroz. In: ZUFFO, Alan Mario (org.). A produção do conhecimento nas ciências agrárias e ambientais. Ponta Grossa, Atena, p. 298-306. DOI: https://doi.org/10.22533/at.ed.87619260427

Monteiro RCM, Gadotti GI, Maldaner V, Curi ABJ, Bárbara Neto M (2021) Image processing to identify damage to soybean seeds. Ciência Rural 51(2):1-8. DOI: https://doi.org/10.1590/0103-8478cr20200107

Patel AA, Kathiriya DR (2017) Data mining trends in agriculture: a review. AGRES – An International E. Journal 6(4):637-645. Available: http://arkgroup.arkgroup.co.in/upload/Article/6442%20%20review.pdf Accessed Jan 13, 2021

Patrício DI, Rieder R (2018) Computer vision and artificial intelligence in precision agriculture for grain crops: a systematic review. Computers and Electronics in Agriculture 153:69–81. DOI: https://doi.org/10.1016/j.compag.2018.08.001

Pooja I, Sharma A, Sharma A (2018) Machine Learning: A review of techniques of machine learning. JASC: Journal of Applied Science and Computations 5(7):538-541.

Strieder G, Foguesatto RJ, Gadotti GI, Luz MLGS, Luz CAS, Gomes MC, Scherer VS (2014) Estudo técnico e de cenários econômicos para implantação de uma unidade de tratamento industrial de sementes de soja e trigo. Informativo Abrates 24(3):118-123.

Tilmann MAA, Tunes LVM, Almeida AS (2019) Análise de Sementes. In: Peske ST, Villela FA Meneghello, GE. Sementes: fundamentos científicos e tecnológicos. Orbia.

Torres MFO, Ferreira RA, Prata LCD, Silva-Mann R (2020) Seed Longevity of Enterolobium contortisiliquum (Vell.) Morong. Journal of Seed Science 42:1-14. DOI: https://doi.org/10.1590/2317-1545v42239618

Vergara RO, Gazolla-Neto A, Gadotti GI (2019) Space distribution of soybean seed storage potential. Revista Caatinga 32(2):399-410. DOI: https://doi.org/10.1590/1983-21252019v32n213rc.

Xia Y, Xu Y, Li J, Zhang C, Fan S (2019) Recent advances in emerging techniques for non-destructive detection of seed viability: A review. Artificial Intelligence in Agriculture 1:35-47. DOI: https://doi.org/10.1016/j.aiia.2019.05.001.