# Data mining for ranking sorghum seed lots

# Mineração de dados no ranqueamento de lotes de sementes de sorgo

Luciana D. Rocha[1], Gizele I. Gadotti[2], Ruan Bernardy[2]*, Romário de M. Pinheiro[1], Rita de C. M. Monteiro[1]

[1]Faculty of Agronomy Eliseu Maciel, Universidade Federal de Pelotas, Pelotas, RS, Brazil. [2]Engineering Center, Universidade Federal de Pelotas, Pelotas, RS, Brazil.

**ABSTRACT** - The ranking of seed lots is a fundamental process for all companies in the seed industry. This work aims to demonstrate data mining methods for ranking sorghum seed lots during the seed processing through analysis of quality control data. Germination and cold tests were performed to verify the physiological quality of the lots. Seed samples from each lot were evaluated in two moments: post-cleaning and finished product (ready for marketing). The results after pre-processing totaled 188 rows of data with six attributes, encompassing 150 lots accepted for marketing, 6 rejected, and 32 intermediate lots. The classifiers used were J48, Random Forest, Classification Via Regression, Naive Bayes, Multilayer Perceptron, and IBk. The Resample filter was used for adjustment of the data. The k-fold technique was used for training, with ten folds. The metrics of Accuracy, Precision, Recall, F-measure, and ROC Area were used to verify the accuracy of the algorithms. The results obtained were used to determine the best machine-learning algorithm. IBk and J48 presented the highest accuracy of data; the IBk technique presented the best results. The Resample filter was essential for solving the data imbalance problem. Sorghum seed lots can be classified with great accuracy and precision through artificial intelligence and machine learning technique.

**Keywords**: Quality. Post-harvest technology. Artificial intelligence.

**RESUMO** - A classificação de lotes de sementes é um processo fundamental para todas as empresas do setor sementeiro. O objetivo do trabalho é demonstrar os métodos de mineração de dados de ranqueamento de lotes de sementes de sorgo durante o processo de beneficiamento, através de análises de dados do controle de qualidade. Os testes realizados foram germinação e teste de frio, com o objetivo de verificar a qualidade fisiológica dos lotes. As amostras de sementes de cada lote foram avaliadas em dois momentos: pós-limpeza e produto acabado (pronto para comercialização). Os dados gerados, após o pré-processamento, totalizaram 188 linhas com seis atributos, contabilizando 150 lotes aceitos para comercialização, seis rejeitados e 32 denominados intermediários. Os classificadores utilizados foram J48, Random Forest, Classification Via Regression, Naive Bayes, Multilayer Perceptron e IBk. Utilizou-se o filtro Resample para ajustamento dos dados. A técnica empregada para treinamento foi a k-fold, com 10 folds. Para verificar a precisão dos algoritmos foram utilizadas as métricas de Acurácia, Precisão, Recall, F-measure e Área ROC. Com os resultados obtidos determinou-se o melhor algoritmo de aprendizagem de máquina. Verificou-se que o IBk e o J48 obtiveram maior acurácia nos dados, sendo que a técnica de IBk obteve o melhor resultado. O filtro Resample foi importante para resolver o problema do desequilíbrio dos dados. Concluímos ser possível classificar lotes de sementes de sorgo com grande acurácia e precisão através de inteligência artificial e sua técnica de aprendizado de máquina.

**Palavras-chave**: Qualidade. Tecnologia de pós-colheita. Inteligência artificial.

## INTRODUCTION

*Sorghum bicolor* [(L.) Moench)] is native to Central Africa and part of Asia and is considered the fifth most-grown cereal in the world (CAÑIZARES et al., 2020). It is currently an important alternative food for humans and animals. This species adapts well to regions with low water availability, presents seeds rich in proteins, vitamins, carbohydrates, and mineral salts, and is tolerant to droughts and high temperatures (CARVALHO; NAKAGAWA, 2012).

Sorghum crops have been expanding in Brazil, mainly as a second crop in succession to summer crops, focused on seed quality. The main types of sorghum grown in Brazil are grain, forage, saccharine, lignocellulosic biomass, and broomcorn (CAÑIZARES et al., 2020).

The importance of economic aspects of sorghum crops denotes the importance of ensuring high crop yields; thus, understanding the physiological quality of seeds is essential. Different biotic and abiotic factors can affect the seed physiological quality, including local climate conditions, which may affect several phases of the plant development and directly affect the seed physiological potential when they reach the maturity stage (MARCOS-FILHO, 2015). In addition, the choice of sowing season should be under more favorable climate conditions, according to the plant demands in the different developmental stages.

Seed physiological quality is commonly determined through laboratory tests, evaluating different aspects of seedling growth. For example, germination tests are used to determine the germination potential of seed lots under ideal conditions in laboratory, which generates information on germination aspects (abnormal and normal seedlings, hard and dead seeds) (BRASIL, 2009). However, vigor tests are used to evaluate or detect significant differences in the physiological quality of lots with similar germination. These differences have been safely distinguishing high from low vigor seed lots, separating or classifying lots at different levels of vigor, proportionally to their dynamics regarding seedling emergence in the field and storage potential (MARCOS-FILHO, 2021).

The use of information technology in the seed sector can improve results generated by seed quality tests, resulting in fast responses for decision-making in the planting, application of inputs, and harvest and post-harvest processes, leading to a more intelligent agriculture (PINHEIRO et al., 2021).

The ranking lots into high, medium, and low vigor are essential for all seed companies to speed up the delivery of standardized quality seeds to producers, as well as seed mapping regions to be grown, through results of physiological quality tests. In this sense, the demand for efficient and safe methods is increasing, and the information technology is a support tool for the seed sector, as the combination of adequate information regarding time and careful decision-making, which are essential for the success of this market (PINHEIRO et al., 2021).

In this context, the objective of this work was to demonstrate data mining methods for ranking sorghum seed lots during processing through physiological quality analysis.

## MATERIAL AND METHODS

The study was carried out using data collected from the analysis quality control of the laboratory of the Seeds Lab company, contracted by a seed company in Uberlândia, MG, Brazil. Three hundred seventy-seven sorghum seed lots were grown in the 2021 crop season. In addition, germination and cold tests were carried out to assess the physiological quality of the lots, since germination is the official test for marketing seeds required by the Brazilian Ministry of Agriculture, and the cold test is the most used vigor test for sorghum crops.

Seed samples of each lot were evaluated at the stages of post-cleaning and finished product (ready for marketing). The results after pre-processing totaled 188 rows of data with six attributes (Table 1), which encompassed 150 lots accepted for marketing, 6 rejected, and 32 intermediate lots (when they are not promptly considered as high nor low vigor).

**Table 1**. Description of data of the sorghum seed attributes analyzed for data mining.

| Attribute | Description | Value |
|---|---|---|
| Lots | Lots | {1-187} |
| Germination | Post-cleaning | {0-100} |
| | Finished product | |
| Cold test | Post-cleaning | {0-100} |
| | Finished product | |
| Lot classification | Decision | {Accepted, rejected, intermediate} |

An expert in the field classified the data, as shown in Table 1; they were termed supervised data. Algorithms work with a known classification focused on standards that direct the analyst to inform the referred class to a lot. According to Monard and Baranauskas (2003), this learning form training the algorithm using a dataset in which the class attribute is known. Moreover, it builds a classifier that can correctly determine the class of other non-labeled data, predicting the class to which each dataset belongs, based on the learned characteristics at the training stage.

The germination test was conducted using four 50-seed subsamples per lot, which were sown in paper rolls moistened with water at a quantity equivalent to 2.5-fold the weight of the dry paper. The rolls were maintained in a germinator set at 25 °C, and the evaluations followed the criteria of the Rules for Seed Analysis (BRASIL, 2009); the results were expressed as percentages (%) of normal seedlings.

The seed vigor was evaluated through the cold test, which simulates unfavorable conditions (excess water in the soil and low temperatures) that may occur in the sowing period in the field. The tested samples were placed in a chamber at 10 °C for seven days and then taken to another chamber at 25 °C where they remained for five days, following the evaluations of Rohr et al. (2023).

Regarding data processing and prediction of lots, firstly, a pre-analysis of the information generated during the

phases of physiological quality analysis was needed to prepare the dataset to enable the tool to perform the correct reading and learning analysis. The file was then converted into .csv format and opened in the Microsoft Windows Notepad application, which required adequate commas and semicolons. In addition, rows with missing values or misleading data were excluded during this processing.

The software Weka 3.8.5, developed by the University of Waikato, was used for the data mining task using machine learning methods.

The training and test of the dataset intended to cross-validation of these data were carried out by subdividing them into ten subsets (10 folds). One of the ten subsamples was retained for model validation and the others were used for training. As there were ten folds, this process was repeated ten times. This technique reduces the probability of coincidences underestimating or overestimating the performance of a configuration. All results reported in the present work were found using this technique.

The classifiers used were: J48; Random Forest, which works through decision-making trees; Classification via Regression; Naive Bayes; Multilayer Perceptron (neural networks); and IBk.

Classification filters are used to examine the characteristics of a dataset and frame them into classes to generalize and specialize the data that distinguish these classes for prediction of data or records not automatically classified (VASCONCELOS; CARVALHO, 2018). According to Beniwal and Arora (2012), some algorithms use this work concept with decision-making trees, neural or Bayesian networks, and closest neighbors.

The seed data analyses acquired are naturally imbalanced, mainly those from companies focusing on high-quality lots. The Resample filter was used to solve this problem and not bias the algorithm, as it is a non-supervised instance filter that keeps the class distribution in the subsample and, alternatively, can be set to bias the class distribution for a uniform distribution (GADOTTI et al., 2022a).

The following evaluation metrics were used to assess the precision of the algorithms: Accuracy, Precision, Recall, F-measure, and ROC Area, according to Lever, Krzywinski and Altman (2016).

True positive (TP), false positive (FP), true negative (TN), and false negative (FN) values were extracted through a confusion matrix to calculate the Recall and Precision metrics using Equations 1 and 2, as proposed by Medeiros et al. (2020). Finally, the results obtained were used to determine the best learning technique.

$$Recall = TP / (TP + FP) \qquad (1)$$

where:
TP = true positive
FP = false positive

$$Precision = (TP + TN) / (TP + TN + FP + FN) \qquad (2)$$

where:
TP = true positive
TN = true negative
FP = false positive
FP = false negative

After these steps, the dataset was ready to be processed by the main task of the process: mining. The algorithms were used several times and repeatedly sought standards and rules in the data. The information found was then interpreted and evaluated through graphics or reports, selecting the most helpful information (VASCONCELOS; CARVALHO, 2018). Figure 1 shows the phases of the methodology used.
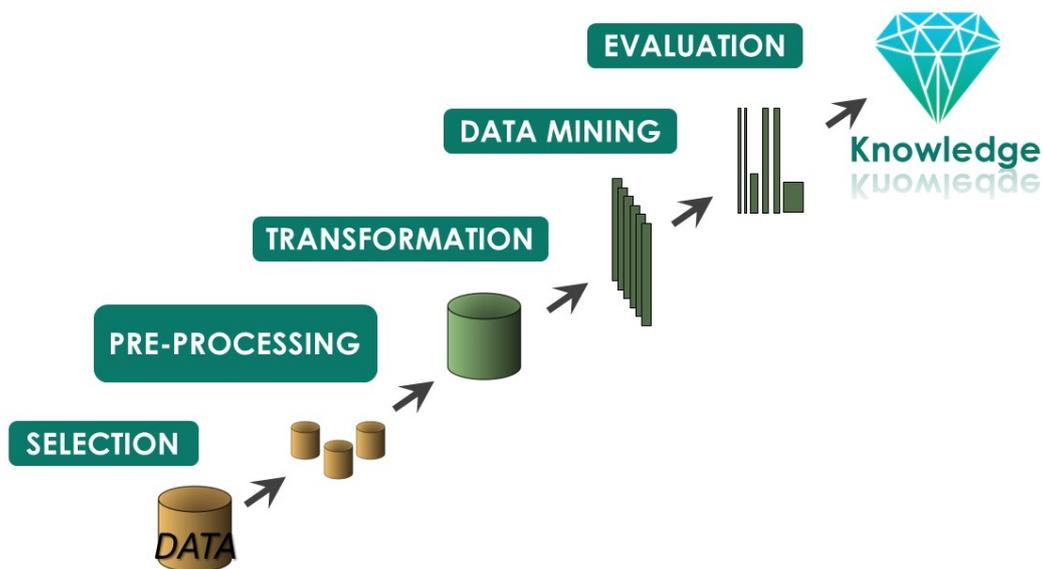


**Figure 1**. Sequencing of operations using data mining.

## RESULTS AND DISCUSSION

The results obtained were used to determine the best learning algorithm. The test data were also subjected to statistical procedures to confirm the improvement of the results from the application of the resample filter, through analysis of variance (ANOVA). The means were compared by the Tukey's test at 5% significance level when statistical differences were found. This test was not used for the choice of the best model, the choice was made through the metrics of accuracy and percentage of assertiveness.

## RESULTS AND DISCUSSION

The results presented and extracted by machine learning enabled the separation of seed lots according to the seed physiological quality; however, many models can be tested to ensure the best ranking technique. All models evaluated for separating sorghum seed lots presented accuracy above 80%, except the Classification via Regression (CVR), which presented accuracy of 79.8% (Table 2). Thus, not all models present good performance, which can be connected to the quantity of data. Furthermore, the number of samples used was insufficient to present more robust predictions, and not all machine learning algorithms classify the same data equally (JAGTAP et al., 2022).

Jin et al. (2022) evaluated rice seed viability and vigor predictions through machine learning, based on 212 images from 870 seeds, and found differences between models (logistic regression, neural networks convolutional, and support vector machine), but no evident advantage from one model to the other, and that the response efficiency of each model depends on the quantity of data used. In the present study, the CVR model showed a relatively small difference (0.2%) from the accuracy of 80%, which was the worse result among the models tested. The predictions established by the models were based on information in which the class values are known, based on a dataset obtained from existing systems (computational or human) that support the decision for the level of complexity of the crop modeling parameters on seed physiological quality when integrated into machine learning (GADOTTI et al., 2022b).

Seed quality control tests should be considered, since each agricultural crop season requires information on the lots produced. Seed quality standards are required and should be within minimum legal requirements; the companies can conduct internal control tests that generate this information (GADOTTI et al., 2022a).

However, physiological analyses are mandatory in the sowing period for transporting and marketing seed lots; thus, quality analyses are carried out in two phases, generating information for decision-making: post-harvest for storage and planting.

After evaluating the accuracy of algorithms for ranking the seed lots by physiological quality and using the Resample filter, the best models were: J48 (96.3% precision) and IBk (96.8% precision). Considering the machine learning model results, these models showed that the filter improves the performance of algorithms when they analyze unbalanced data, which is consistent with the findings of Gadotti et al. (2022a). Gadotti et al. (2022b) found higher accuracy for J48 and CVR; contrastingly, Gadotti et al. (2022a) found lower accuracy for Random Forest (79.56%) and CVR (81.72%) and reported that each species has specificities. Thus, there is not only one best classifier, but a dependency on the database used.

Table 2 shows that the Resample filter improved the accuracy of all algorithms. Resample is a non-supervised instance filter that keeps the class distribution in the subsample. Alternatively, it can be set to bias the class dispersal for a uniform distribution. The sampling can be carried out with (standard) or without replacing (WITTEN; FRANK; HALL, 2011).

This filter enables the production of a random subsample in a dataset by using sampling with or without replacing it. The seed physiological quality data are biased. For example, seed germination tends to be high (above 80%); thus, the data present a bias and can be used as standard data. The classification is always expected to have opposite or the most random possible data. In addition, conventional statistics enabled to differentiate the results of the algorithms evaluated, but it was determined that the choice of the best algorithms would be by criteria of the metric of precision and percentage of assertiveness of each model used, calculated by the WEKA software.

**Table 2**. Accuracy of algorithms after classification.

| Algorithm | Correct attribute classification (%) | |
| --- | --- | --- |
| | No resample | With resample |
| CVR | 79.8 cb | 86.2 c |
| J48 | 93.1 a | 96.3 a |
| Naïve Bayes | 92.5 a | 94.7 b |
| Random Forest | 80.8 b | 93.1 b |
| Multilayer Perceptron | 89.4 b | 93.08 b |
| IBk | 93.1 a | 96.8 a |

The statistical difference found by the Tukey's test was not used as a criterion for choosing the best method, but the metrics of accuracy and the percentage of assertiveness of each algorithm (Table 2).

High values were found for detecting physiological aspects of sorghum seeds; the IBk algorithm generated 98.7% recall, and 97.4% mean precision for the accepted class. The results were 100% recall and precision for the rejected class and 87.1% recall and 93.1% precision for the intermediate class (Table 3).

Gadotti et al. (2022a) evaluated soybean seed lots and found the best results for the Random Forest algorithm, with 92.6% recall and 90.9% precision for the accepted class, 92.6% recall and 73.59% precision for the rejected class, and 8.3% recall and 25% precision for the intermediate class, denoting that the precision of algorithms depends on the dataset used.

The F-Measure presents combined values of mean recall and precision, then the classes with higher values are consequently the accepted and rejected ones. The ROC (Receiver Operator Characteristic) curve or area denotes the correlation between the sensitivity and specificity of the classifier, the higher the value, the best the fit to the curve. The IBk and J48 classifiers presented the highest ROC values for the rejected (1.00) and accepted (0.94) classes, respectively (Table 3). In these classes, the ROC curve was more well-defined than in the other classes.

**Table 3**. Accuracy results in relation to the performance of the analyzed models, Recall (sensitivity), Precision, ROC (Receiver Operating Characteristic) and F Measure.

| Classifiers | Accuracy | | | | |
|---|---|---|---|---|---|
| | Recall | Precision | ROC Area | F-Measure | Class |
| IBk | 0.987 | 0.974 | 0.924 | 0.980 | Accept |
| | 1.000 | 1.000 | 1.000 | 1.000 | Reject |
| | 0.871 | 0.931 | 0.908 | 0.900 | Intermediary |
| J48 | 0.987 | 0.987 | 0.940 | 0.987 | Accept |
| | 0.857 | 0.667 | 0.916 | 0.750 | Reject |
| | 0.871 | 0.931 | 0.901 | 0.900 | Intermediary |
| CVR | 0.877 | 0.971 | 0.974 | 0.922 | Accept |
| | 0.769 | 1.000 | 1.000 | 0.870 | Reject |
| | 0.810 | 0.436 | 0.952 | 0.567 | Intermediary |
| MLP | 0.987 | 0.974 | 0.996 | 0.980 | Accept |
| | 0.000 | 0.000 | 1.000 | 0.000 | Reject |
| | 0.871 | 0.750 | 0.990 | 0.806 | Intermediary |
| Naïve Bayes | 0.948 | 0.986 | 0.992 | 0.967 | Accept |
| | 1.000 | 1.000 | 1.000 | 1.000 | Reject |
| | 0.905 | 0.704 | 0.985 | 0.792 | Intermediary |
| Random Forest | 1.000 | 0.922 | 1.000 | 0.960 | Accept |
| | 0.769 | 1.000 | 1.000 | 0.870 | Reject |
| | 0.524 | 1.000 | 0.997 | 0.688 | Intermediary |

The choice of decision tree by the algorithm J48 (Figure 2) is because this is a derivation, in Java, of the C4.5 algorithm, which is one of the most used and reliable statistical classifiers. It builds the decision tree using the entropy concept; the algorithm chooses the attribute that partitions most of the data through the gain of normalized information (GADOTTI et al., 2022b).
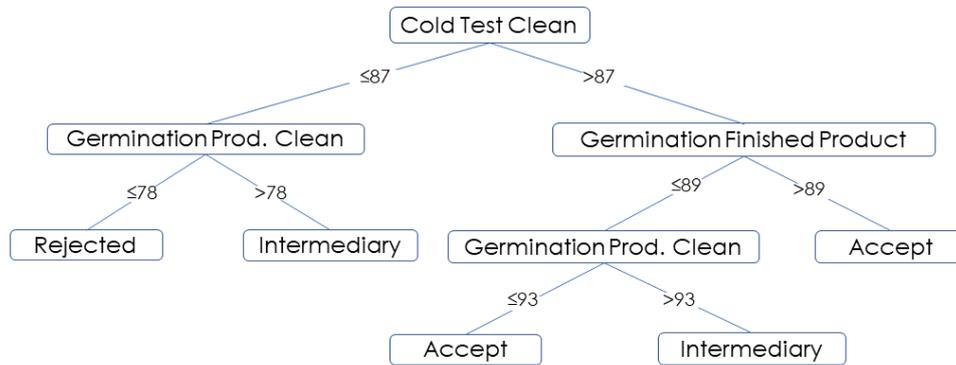


**Figure 2**. Decision tree for prediction of classification of sorghum seed lots by the J48 algorithm.

According to the data distribution analyzed, the decision tree generated indicates that the learning model defines the post-cleaning cold test attribute as the decision-making on the primary quality control test for sorghum seeds. This decision is one of the criteria also determined by the seed analyst. Thus, vigor test is essential for ranking seed lots and should be evaluated in further studies to determine the effect of each one on the segregation of seed quality. Furthermore, this is one of the most recommended tests for sorghum seeds due to seed dormancy.

However, a more significant number of attributes, combined with other vigor tests, would be interesting for seed science and technology studies. It is interesting to understand the importance of tests for more efficient classification of seed lots, as the cold test needs to be standardized for most crops (TILLMANN; TUNES; ALMEIDA, 2019; GADOTTI et al., 2022a).

The lots evaluated showed that the choice established by the decision tree presents higher values than the minimum germination recommended for marketing sorghum seeds, which was established as 80% (BRASIL, 2009). Seeds with germination below this established standard for marketing present lower possibility to express their physiological potential and originate normal vigorous seedlings that can survive under non-favorable field conditions (OLIVEIRA et al., 2015).

Thus, several tests are used to assess seed vigor. Silva et al. (2016) found that first germination count and accelerated aging test are more efficiently in detecting differences in vigor between sorghum genotypes. However, cold stress (0-15 °C) has been recognized as one of the significant abiotic constraints to sorghum production, especially in cold climate regions. Sorghum is a C4 plant that evolved under hot conditions in tropical Africa, where temperatures during the growing season are usually higher than 20 °C (PEREIRA FILHO; RODRIGUES, 2015). Therefore, cold stress in high-altitude regions affects the sorghum growth and development. Low temperature during the crop growth season affects almost all growth stages and decreases crop yield. Adverse effects caused by low temperatures are commonly visible during the initial stages of the plant development in most crops sensitive to low temperatures. The stress caused by low temperatures can reduce germination and emergence rates, limit the establishment of seedlings, atrophy buds and root development (RUTAYISIRE et al., 2021) and, consequently, generate low-quality seeds.

The confusion matrices showed that the predictions of the generated results and the two algorithms were similar; however, J48 distributes the error into the other classes, and IBk concentrates the error in a single class (Tables 4 and 5). This result explains the highest accuracy found for IBk (Table 1) and its sensitivity (Recall) (Table 2).

**Table 4**. Confusion matrix of the IBk algorithm.

| | | Prediction | | |
|---|---|---|---|---|
| | | Accepted | Rejected | Intermediary |
| | Accepted | 27 | 4 | 0 |
| Actual class | Rejected | 2 | 148 | 0 |
| | Intermediary | 0 | 0 | 7 |

**Table 5**. Confusion matrix of the J48 algorithm.

| | | Prediction | | |
|---|---|---|---|---|
| | | Accepted | Rejected | Intermediary |
| Actual class | Accepted | 27 | 2 | 2 |
| | Rejected | 1 | 148 | 1 |
| | Intermediary | 1 | 0 | 6 |

Gadotti et al. (2022a) evaluated soybean seed lots and found lower rates of false positives in rejected and intermediate classes for the CVR algorithm. The precision stood out in this case, classifying a rejected lot correctly in practically all of them, mistaken it in only two.

The intermediate class is more complex for decision-making, which is affected when different analysts analyze the dataset; this algorithm's complexity is also shown for predicting lots. In the study of Gadotti et al. (2022a), this class had mixed values in the confusion matrix, i.e., the algorithm did not understand the standard to inform whether the lot was intermediate, classifying some as rejected and others as accepted.

The ranking and association of lots are essential for companies to expedite seed deliveries to growers and for local mapping requirements to be met upon seed vigor and germination results.

In a study on corn seeds conducted by Gadotti et al. (2022b), data mining presented several forms and models for prediction of classification of lots using different machine learning algorithms. These models were created from lists whose class values are known and obtained from existing systems that perform the task for which the model is desired.

Conditions for high crop production include high-vigor seeds. According to the Normative Instruction No. 45 of September 17, 2013, of the Brazilian Ministry of Agriculture, Livestock and Supply, good quality seeds are those with at least 80% germination, with a minimum purity of 98% (BRASIL, 2013). Furthermore, it is recommended that the seed vigor has a percentage close to that found in the germination of the lot, increasing the productivity of the crop (SCHEEREN et al., 2010).

Thus, the whole detailing of physiological quality within the processing complements is needed for quality analyses, since when the lot is clean, the seed mass is ready, requiring only the treatment and bagging, with no changes in seed quality, as these operations do not include separations.

However, vigor tests are limited for sorghum crops; therefore, more physiological detailing is required, which needs further studies to improve methodologies and expedite processes. The tetrazolium test is faster than other tests; thus, it is more used at pre-harvest. The cold test is used after processing and carrying out the germination test based on laboratory routines and field stress described by Pereira Filho and Rodrigues (2015), and Rutayisire et al. (2021), thus resulting in no options for testing the volumes produced.

## CONCLUSION

Sorghum seed lots can be classified with great accuracy and precision through artificial intelligence and machine learning technique. The algorithms IBk and J48 stood out when applying the machine learning technique. The Resample filter was essential to solve the data unbalance problem.

## ACKNOWLEDGMENTS

## REFERENCES

BENIWAL, S.; ARORA, J. Classification and feature selection techniques in data mining, **International Journal of Engineering Research & Technology**, 1: 1-6, 2012.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. **Regras para análise de sementes**. 1. ed. Brasília, DF: MAPA, 2009. 398 p.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. **Instrução Normativa nº45 de 17 de setembro de 2013**. 2013. Disponível em: <https://www.gov.br/agricultura/pt-br/assuntos/insumos-agropecuarios/insumos-agricolas/sementes-e-mudas/publicacoes-sementes-e-mudas/ copy_of_INN45de17desetembrode2013.pdf>. Acesso em: 22 out. 2021.

CAÑIZARES, L. C. C. et al. Tecnologia e industrialização de grãos de canola, girassol, linhaça, algodão, amendoim, sorgo, milho pipoca, lentilha e ervilha. In: FERREIRA, C. D.; OLIVEIRA, M. de; ZIEGLER, V. (Eds.). **Tecnologia industrial de grãos e derivados**. 1 ed. Curitiba, PR: Editora CRV, 2020. v. 1, cap. 9, p. 225-276.

CARVALHO, N. M.; NAKAGAWA, J. **Sementes: ciência, tecnologia e produção**. 5. ed. Jaboticabal, SP: Funep, 2012. 590 p.

GADOTTI, G. I. et al. Aprendizado de máquina para classificação de lotes de sementes de soja. **Engenharia Agrícola**, 42: e20210101, 2022a.

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

JAGTAP, S. T. et al. Towards application of various machine learning techniques in agriculture. **Materials Today**, 51: 793-797, 2022.

JIN, B. et al. Determination of viability and vigor of naturally-aged rice seeds using hyperspectral imaging with machine learning. **Infrared Physics & Technology**, 122: e104097, 2022.

LEVER, J.; KRZYWINSKI, M.; ALTMAN, N. Classification evaluation. **Nature Methods**, 13: 603-604, 2016.

MARCOS-FILHO, J. Seed vigor testing: an overview of the past, present and future perspective. **Scientia Agricola**, 72: 363-374, 2015.

MARCOS-FILHO, J. Teste de envelhecimento acelerado. In: KRZYZANOWSKI, F. C. et al. (Eds.). **Vigor de sementes**: conceitos e testes. Londrina, PR: ABRATES, 2021. cap. 2, p. 1-24.

MEDEIROS, A. D. et al. Interactive machine learning for soybean seed and seedling quality classification. **Scientific Reports**, 10: 1-10, 2020.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: SOLANGE, O. R. (Ed.) **Sistemas Inteligentes: fundamentos e aplicações**. Barueri, SP: Manole Ltda, 2003. v. 1, cap. 4, p. 89-114.

OLIVEIRA, L. M. et al. Qualidade de sementes de feijão-caupi tratadas com produtos químicos e armazenadas em condições controladas e não controladas de temperatura e umidade. **Semina**: **Ciências Agrárias**, 36: 1263-1276, 2015.

PEREIRA FILHO, I. A.; RODRIGUES, J. A. S. **O produtor pergunta, a Embrapa responde**. Brasília, DF: Embrapa Cerrados, 2015. 332 p.

PINHEIRO, R. M. et al. Inteligência artificial na agricultura com aplicabilidade no setor sementeiro. **Diversitas Journal**, 6: 2984-2995, 2021.

ROHR, L. A. et al. Soybean seeds treated with zinc evaluated by X-ray micro-fluorescence spectroscopy. **Scientia Agricola**, 80: e20210131, 2023.

RUTAYISIRE, A. et al. Response of sorghum to cold stress at early developmental stage. **International Journal of Agronomy**, 2021: 1-10, 2021.

SCHEEREN, B. R. et al. Qualidade fisiológica e produtividade de sementes de soja. **Revista Brasileira de Sementes**, 32: 35-41, 2010.

SILVA, R. S. et al. Qualidade fisiológica de sementes de sorgo biomassa (*Sorghum bicolor* L. Moench). **Revista Espacios**, 37: 12-19, 2016.

TILLMANN, M. A. A.; TUNES, L. M.; ALMEIDA, A. S. Análise de Sementes. In: PESKE, S. T.; VILLELA, F. A.; MENEGHELLO, G. E. (Eds.). **Sementes: fundamentos científicos e tecnológicos**. Pelotas, RS: Becker, 2019. v. 4, cap. 3, p. 147–257.

VASCONCELOS, L. M. R.; CARVALHO, C. L. de. Aplicação de regras de associação para mineração de dados na web. **Technical Report**, 1: 1-20, 2018.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining**: pratical machine learning tools and techniques. 3. ed. Burlington: Morgan Kaufmann Publishers, 2011. 665 p.