ISSN 1807-1929

Revista Brasileira de Engenharia Agrícola e Ambiental

Brazilian Journal of Agricultural and Environmental Engineering

v.28, n.11, e281001, 2024

Campina Grande, PB – http://www.agriambi.com.br – http://www.scielo.br/rbeaa

DOI: http://dx.doi.org/10.1590/1807-1929/agriambi.v28n11e281001

Original Article

Estimating energy efficiency of the aeration process of stored grains through machine learning[1]

Estimativa da eficiência energética do processo aeração de grãos armazenados através de machine learning

Weder N. Ferreira Junior[2], Osvaldo Resende[2], Daniela C. de Oliveira[2], Daniel E. C. de Oliveira[2]* & Elivânio dos S. Rosa[2]

[1] Research developed at Instituto Federal de Educação, Ciência e Tecnologia Goiano, Campus Rio Verde, Rio Verde, GO, Brazil

[2] Instituto Federal de Educação, Ciência e Tecnologia Goiano/Campus Rio Verde, Rio Verde, GO, Brazil

HIGHLIGHTS:

The model for estimating the energy efficiency of the aeration process proved to be efficient.
The proposed model for evaluating aeration efficiency has applicability of use in predictive analysis of the process.
From data mining and modeling with machine learning, it was possible to develop a Web tool.

ABSTRACT: Aeration is carried out by blowing external air into the silo, with the aim to keep the temperature in the mass of stored grains at safe levels. In the present study, the energy efficiency of aeration of stored sunflower grains was estimated, and a model was proposed and tested to estimate the energy efficiency of aeration, using different algorithms in supervised and unsupervised machine learning. The objective of the work was to develop a Web application based on data mining and modeling with machine learning. The database was composed of information on the average temperature at the height of the sensors, average temperature of the silo, external ambient temperature, occurrence of aeration, if there was cooling, heating and direct heating during aeration, and the energy efficiency of the aeration process. The model for estimating the energy efficiency of the aeration process proved to be efficient, identifying that the energy efficiency was 97.78% during the aeration of stored sunflower grains. Among the classifier algorithms tested, Support Vector Machine (SVM-Poly) showed the best metrics and indicators, hence being recommended for implementation in system development networks capable of predicting the aeration status of stored grains.

Key words: Weka, support vector machine, K-means

RESUMO: A aeração é realizada por meio da insuflação do ar externo para dentro do silo, tendo como objetivo manter a temperatura da massa de grãos armazenados em níveis seguros. No presente estudo foi estimada a eficiência energética da aeração de grãos de girassol armazenados, assim como proposto e testado um modelo de estimativa da eficiência energética da aeração, utilizando diferentes algoritmos no aprendizado de máquinas supervisionado e não supervisionado. O objetivo no trabalho foi desenvolver uma aplicação Web a partir da mineração e modelagem dos dados com o aprendizado de máquinas. O banco de dados foi composto pelas informações da temperatura média do nível dos sensores, temperatura média do silo, temperatura ambiente externa, ocorrência de aeração, se houve resfriamento, aquecimento e aquecimento direto durante aeração, e a eficiência energética do processo de aeração. O modelo de estimativa da eficiência energética do processo de aeração demonstrou-se eficiente, identificando que durante a aeração de grãos de girassol armazenados a eficiência energética foi de 97,78%. Dentre os algoritmos classificadores testados na Máquina de Vetores de Suporte (SVM-Poly) apresentou as melhores métricas e indicadores, sendo recomendado para implementação em redes de desenvolvimento de sistemas capaz de predizer o status da aeração de grãos armazenados.

Palavras-chave: Weka, máquina de vetores de suporte, K-means

• Ref. 281001 – Received 30 Nov, 2023
* Corresponding author - E-mail: oliveira.d.e.c@gmail.com
• Accepted 05 Jun, 2024 • Published 29 Jun, 2024
Editors: Toshik Iarley da Silva, Walter Esfrain Pereira & Leda Rita Dantonino Faroni

Introduction

Aeration aims to keep the temperature in the mass of stored grains at safe levels, as well as the intergranular relative humidity, besides contributing to the uniformity of grain mass temperature (Lopes & Steidle Neto, 2019). Aeration management is carried out by blowing external air into the silo; therefore, before and during the process, attention should be paid to grain mass temperature, through thermometry data, and external air temperature (Durks et al., 2019; Lopes & Steidle Neto, 2019).

To carry out aeration, the best external air conditions must be taken advantage of to make the process economical and preserve the quality of the product (Panigrahi et al., 2020). Inefficiency in the aeration process compromises grain quality, as the increase in temperature can contribute to increasing the respiratory rate of the product, thus directly affecting its quantity and quality (Mohapatra et al., 2017). In addition, an inefficient process results in unnecessary energy expenditure, increasing the costs of the storage.

Therefore, the energy efficiency of these processes need to be monitored to improve planning management, since there is no predictive model for the energy efficiency of aeration. Therefore, it is important to propose a model capable of estimating the efficiency of the process. Machine learning techniques can be employed in this management process, as they use the interference principle called induction, which allows drawing generic conclusions through a set of raw data, and this learning can be supervised or unsupervised (Setiawan et al., 2009; Lorena & Carvalho, 2013).

Thus, the objective in the present study was to estimate the energy efficiency of aeration of stored sunflower grains, as well as proposing and testing a model to estimate the energy efficiency of aeration, using different algorithms in supervised and unsupervised machine learning. Building upon these goals, the study also aimed to develop a Web application based on machine learning.

Material and Methods

All the information (temperature at the height of the sensors, average temperature of the silo, external ambient temperature) was classified into two databases, primary and secondary; the first was composed of raw data, while the second was obtained from the first through data processing with noise removal, correction of inconsistencies, and elimination of redundant and missing data. The experiment was carried out in a grain storage unit in the municipality of Morrinhos, Goiás, Brazil.

The experimental data used were obtained from the storage of sunflower grains, specifically from the beginning of storage, during the silo filling stage. In this period, the grains already stored were managed with the aeration strategy aimed at cooling, considering standards related to internal and external climatic conditions. Aeration fans were turned on when the external temperature was 4 °C lower than the average internal temperature, i.e., the temperature of the mass of sunflower grains, as well as under conditions of no rainfall, and outside peak energy hours, between 5:30 and 8:00 p.m.

The grains were stored in a vertical metal silo, 22 m wide, with 22 rings of 0.917 m, forming a body of 20.19 m in height and 26.44 m in total height. The silo body volume is 7,674.87 m^3, and the total volume is 8,466.81 m^3. Considering the specific mass of sunflower grains of 0.39 ton m^{-3}, the static capacity of the silo body is 2,993.20 tons of sunflower grains. The aeration system of the silos was composed of centrifugal fans with forward-curved blades driven by a three-phase motor with a power of 50 hp (36.77 kW) and which delivers an air flow (specific flow rate) of 0.05 m^3 min^{-1} ton^{-1} of grain to the aeration system.

The internal temperature of the silo was monitored by a digital thermometry system, whose sensors were distributed inside the silo for each 66 m^3 of grains. External climatological data were monitored by means of a weather station in the Storage Unit. The datasets were saved in intercalated periods of approximately 2 hours, or with each significant change in sensor readings.

The database was composed of data referring to 37 days of storage, and thermometry data only from half of the silo were used during this period. As this is the silo filling period, it was not possible to use the data from all the sensors present in the silo during the data analysis period, so the data up to the seventh height of sensors (S07) were used; the sensors were vertically spaced at 1.5 m (Figure 1). During the period, the thermometry cable 03 (P03) was deactivated for maintenance, so the data from sensors on this cable were discarded.

Figure 1 shows the front view of the silo up to the seventh height of sensors (S07), which were used to compose the database, in addition to illustrating the distribution scheme of the thermometry cables and sensors throughout the silo, in a total of nine cables. The primary database was organized according to the height of sensors per period, i.e., the data of the same height of sensors were arranged in sequence from the beginning to the end of the data period, followed by the other heights in the same format.

The height of sensors is composed of one sensor from each thermometry cable (P01, P02, P04, P05, P06, P07, P08 and

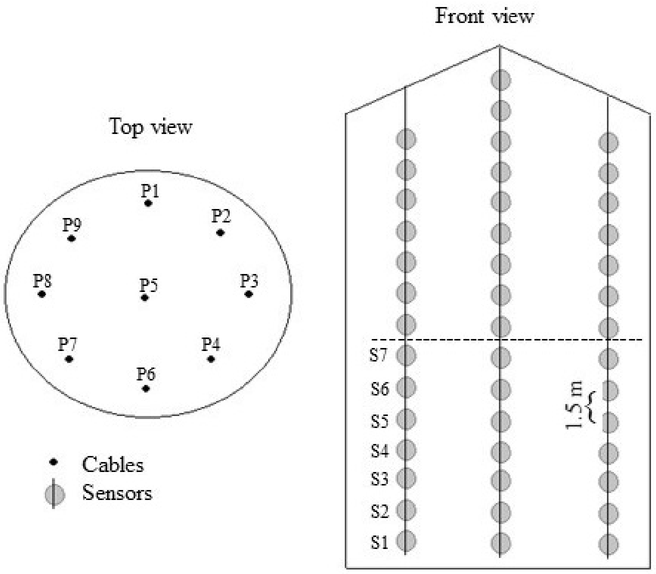


Figure 1. Organizational scheme of temperature data collected from a silo with digital thermometry system

P09), and the average temperature for each height of sensors was calculated for the different moments. The primary database was composed of the temperature of each sensor per cable at the same height, average temperature of the height of sensors at the time, and overall average of the silo temperature at the time, considering the average of only the seven heights of sensors, besides considering the average of the external ambient temperature at the time and the condition of whether or not there was aeration in that period. When there was aeration, the data was coded 1, whereas 0 indicated the absence of aeration in the period evaluated.

The secondary database was obtained from the primary database and included variables to indicate the efficiency of aeration; however, there are no standards in the literature for estimating efficiency, so some standards were tested in the present study in order to propose a model for predicting efficiency.

The first variable obtained was cooling during aeration. Cooling was considered when: there was aeration and the mean temperature of the height of sensors at the time was lower than or equal to the average temperature at the height of sensors in the period prior to the period analyzed. The heating during aeration variable was obtained by considering when there was aeration in the period, but no cooling, as well as when the average temperature of the height of sensors at the time was higher than 0.2 °C compared to the average temperature at the height of sensors in the previous period.

Heating during aeration data were filtered, considering direct heating during aeration, which can occur when the external air is blown into the silo at an inappropriate temperature. To evaluate whether the heating of the grains is directly related to aeration, it was considered to be the case when there was heating during aeration and when the average temperature at the height of sensors at the time was higher than or equal to the external ambient temperature + 2 °C.

The model for evaluating the energy efficiency of aeration was estimated considering the periods in which there was aeration, and the indication of efficiency was obtained through the global analysis of the database, and for the process to be considered efficient the average temperature of the height of sensors at the time had to be lower than 2 °C and above the previous average temperature of the same height of sensors. In addition, the average temperature of the silo at a given time must be lower than 1.5 °C and above the previous average temperature, and the data must also not indicate that there was direct heating during aeration.

For the variables cooling, heating during aeration and direct heating during aeration, the responses were coded; positive responses were coded with 1 and negative responses were coded with 0, whereas for efficiency the data were classified as efficient and non-efficient. The summary of the data analysis set is presented in Table 1.

Therefore, the complete database on which the tests were performed was obtained. The database was composed of eight columns and 3,808 rows, totaling 30,464 values. It should be highlighted that data in the rows were separated between the seven heights of sensors, thus increasing the variability of data throughout the silo.

Table 1. Summary of data processing to obtain responses during the aeration process

Process	Occurrence when
Cooling	There was aeration; and, $\text{TavgSensors}_{\text{actual}} - \text{TavgSensors}_{\text{previous}} \leq 0$ °C.
Heating during aeration	There was aeration; There was no cooling; $\text{TavgSensors}_{\text{actual}} - \text{TavgSensors}_{\text{previous}} > 0.2$ °C.
Direct heating	There is heating during aeration; $\text{TavgSensors}_{\text{actual}} \leq \text{Tambient}_{\text{actual}} + 2$ °C.
Aeration (energy) efficiency	There is aeration; $\text{TavgSensors}_{\text{actual}} \leq \text{TavgSensors}_{\text{previous}} + 2$ °C; $\text{TavgSilo}_{\text{actual}} \leq \text{TavgSilo}_{\text{previous}} + 1.5$ °C; There is direct heating during aeration.

TavgSensors - Average temperature of the height of sensors, °C; Tambient - External ambient temperature, °C; and TavgSilo - Average silo temperature, °C

For predictive modeling of the responses of efficiency of the aeration process, different supervised machine learning algorithms were fitted to the data obtained experimentally, later treated and organized in a file in the notepad tool. The classifier algorithms used were multilayer perceptron (MLP), support vector machine with polynomial kernel (SVM-Poly), support vector machine with radial kernel (SVM-Radial), and the decision tree algorithms J48 and random forest.

The predictive models were processed by the Weka 3.8.5 tool, using cross-validation with 10 folds. The efficiency of the classifiers was evaluated based on the analysis of performance indicators and metrics. The indicators analyzed were the correct classification of the instances (CCI) (Eq. 1), incorrect classification of the instances (ICI) (Eq. 2), and the errors in the confusion matrix for efficiency (EF) and non-efficiency (NEF) of the aeration process.

$$\text{CCI} = \frac{\text{Total of hits}}{\text{Total of data in the set}} \times 100 \tag{1}$$

$$\text{ICI} = \frac{\text{Total of errors}}{\text{Total of data in the set}} \times 100 \tag{2}$$

The metrics analyzed from the classifiers' results were the Kappa coefficient, relative absolute error (RAE) (Eq. 3), root mean squared error (RMSE) (Eq. 4), and root relative squared error (RRSE) (Eq. 5).

$$\text{RAE} = \frac{\sum_{i=1}^{n} \left| p_i - a_i \right|}{\sum_{i=1}^{n} \left| \bar{a} - a_i \right|} \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(p_i - a_i \right)^2}{n}} \tag{4}$$

$$\text{RRSE} = \frac{\sum_{i=1}^{n} \left(p_i - a_i \right)^2}{\sum_{i=1}^{n} \left(\bar{a} - a_i \right)^2} \tag{5}$$

where:

pi - value of the i-th observation;

ai - predicted value for i-th observation;

a - mean of predicted values; and,

n - number of observations.

The partitioning clustering algorithm was used for unsupervised learning, and it was performed in Weka 3.8.5. using the K-means algorithm, which aims to partition n records into k clusters, where k < n. The data are separated into clusters, where centroids represent the center of the cluster, so the data are grouped into clusters according to the shortest distance of the centroid of the different clusters, following a distance metric (Corcovia & Alves, 2019), which is the Euclidean distance for the K-means algorithm (Oliveira et al., 2022).

According to Corcovia & Alvez (2019), the application of the K-means algorithm requires determining which number of clusters will be generated by the algorithm; for these two clusters were standardized in the estimation of the separation of the data on the status of the energy efficiency of aeration into efficient and non-efficient. The results were evaluated based on the values of the incorrect classification of the instances, as well as on the assessment of the errors of the clusters as a function of the observed data of aeration, cooling, heating and direct heating.

This study involved the use of the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology for project management and development of a Web application, as illustrated in Figure 2, which was able to predict whether the aeration of the silo was efficient or non-efficient.

The dataset obtained was in *.xls* format, consisting of 3,808 rows and eight variables, totaling 34,272 data. The dataset was converted to comma-separated-values and then subjected to exploratory data analysis, which corresponds to summarizing, organizing, and interpreting the collected data. The following are the variables of the initial dataset and their descriptions:

TempSensors: temperature of the sensors inside the silo;

TempSILO: temperature inside the silo;

TempAMB: ambient temperature at the time of the experiment;

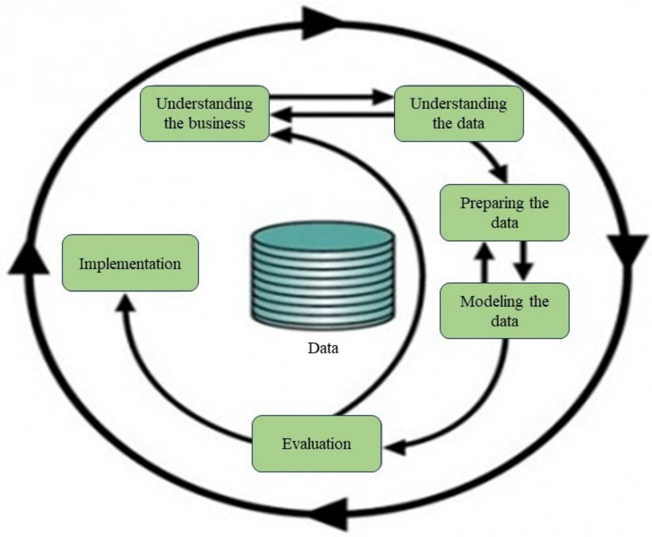


Figure 2. Simplified representation of the project

Aeration: determines whether or not there has been aeration;

Cooling Aeration: determines whether or not there has been cooling;

Heating Aeration: determines whether or not there has been heating;

Direct Heating Aeration: determines whether or not there has been direct heating at the time of aeration;

Aeration Efficiency: Determines whether or not aeration was efficient.

In the data pre-processing stage, outliers and missing data were identified and excluded. A column named efficiency, of the aeration system, which is the target variable of the machine learning model, was also created to determine whether the aeration was efficient or non-efficient.

After mining the data for the aeration system for sunflower grains, it was possible to conclude that the dataset was composed of 84.56% corresponding to 3220 aeration non-efficient data and 15.44% corresponding to 588 aeration efficient data, according to the data pre-processing step, for prediction of the efficiency of the aeration system for sunflower grains.

The statistical analysis of the dataset was performed to obtain the following information: mean (mean), standard deviation (std), minimum (min), quartiles (1^{st}, 2^{nd}, 3^{rd}) and maximum (max) for each column. Subsequently, Pearson's (r) correlation was performed.

For data modeling, the following technologies were used: GitHub tool to control project version, Phyton programming language to develop the machine learning model and the aeration system API. The following libraries for Python programming language were also used: pandas, pandas-profiling, seaborn, matplotlib, numPy, scikit-learn, PyCaret and Streamlit.

SQlite3 was chosen as the database, and Anaconda was chosen as the package manager. Jupyter Notebook tool was used as an environment for development, training, testing, and evaluation of the results of the machine learning model. The integrated development environment (IDE) Visual Studio Code was used.

For the training of the dataset, the following features were defined: sensor temperature, silo temperature, ambient temperature, aeration, cooling aeration, heating aeration, direct heating aeration and the target variable. The training set was defined as the data presented to the machine learning algorithm to create the model with 70% of the data. The test set was presented to the model after its creation, simulating real predictions that the model made, thus allowing the actual performance to be assessed, i.e., 30% of the data.

The classification module used was that of PyCaret, a supervised machine learning module that classifies elements into groups, capable of predicting categorical class labels (discrete and unordered), with graphs to assess the performance of the trained models. A total of 15 algorithms were compared with 10-fold cross-validation.

The metrics used were: accuracy indicated the performance of the model; area under the curve (AUC) provided the performance measure of the classification limits; recall measured the number of disapproved comments that the system approved; Precision indicated the classifications with positive class, i.e., how many are correct; F1 score indicated the calculated harmonic mean based on precision and recall; Kappa measured reliability between evaluators; and Matthews correlation coefficient (MCC) measured the quality of binary classifiers.

Results and Discussion

Table 2 shows that 15.44% (588) of the information that made up the database went through the aeration process, of which 60.54% (356) of the data indicated cooling. It was estimated that 6.46% (38) of the aeration process caused heating of the grains, and of these approximately 29% (11) of the heating data indicate direct heating due to the aeration process.

Considering the data of efficient aeration over the total aeration processes, the model tested in this study indicated an energy efficiency level of 97.79% (575) for the aerations performed during the analyzed period, which is a good efficiency level since inefficient aeration can represent unnecessary costs for the storage unit, in addition to compromising the quality of the grains.

It should be noted that the evaluation of aeration energy efficiency was carried out by silo height, so the aeration inefficiency cannot be generalized for the entire silo. Among the heights of sensors analyzed, only thermometry cable S05 did not have any energy inefficiency of aeration during the study period (Figure 3A).

For the S05 height, it is possible to notice that during the aeration moment shown there was a decrease in the temperature of the grains at the height of the sensor, as well as in the total grain mass of the silo. The temperature of the grains at the height of this sensor changed from 17.68 to 14.64 °C and the average temperature of the grain mass in the entire silo changed from 16.08 to 14.47 °C; under this condition, the aeration air at ambient temperature had an average temperature of 9.75 °C.

In Figure 3B it is possible to observe a moment in which the aeration showed energy inefficiency, according to the efficiency model tested, with pronounced reduction in the average temperature of the grains in the entire silo, from 17.20 to 16.97 °C. However, for grains located at the height of this sensor, there was an increase in temperature, from 13.81 to 16.76 °C, representing a 21.36% increment.

When analyzing a grain aeration simulation process without the control system, Sá et al. (2022) found that when the grain is exposed to temperatures higher than its temperature under continuous conditions, the heating process occurs, as the energy balance is positive. In other words, in the case of Figure 3B, the aeration was efficient to reduce grain mass temperature in the silo, but this did not occur at all monitored points.

Table 2. Result of database treatment

Process	Yes	No
Aeration	588	3220
Cooling	356	3452
Heating	38	3770
Direct heating	11	3797
Aeration efficiency	575	3233

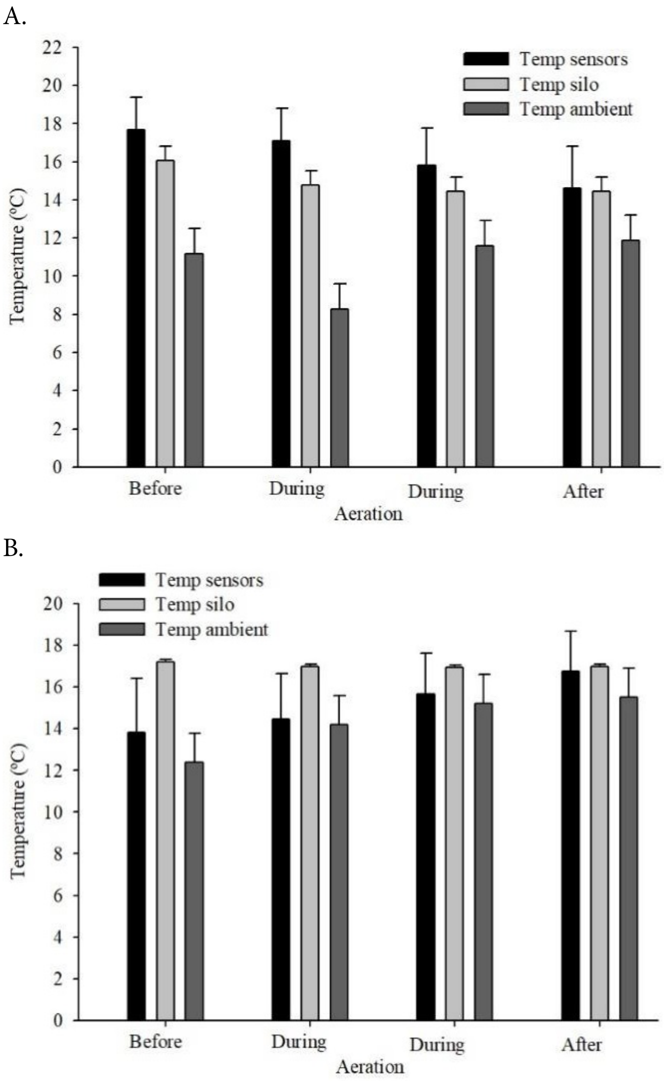


Figure 3. Behavior of grain mass and ambient temperatures for the S05 (A) and S04 (B) heights during an aeration process of stored sunflower grains

The inefficiency in this case occurred due to the increase in this temperature, as it represented an energy expenditure for the storage unit, since energy was spent for cooling the grain in the first instance, which involuntarily resulted in heating of the grain mass. However, two important points must be remembered. The first one is that aeration is carried out to cool the grain mass of the silo, that is, it can be performed to reduce heat concentration at specific points, such as heat pockets, so heating situations are normal from a practical point of view.

Another situation is related to the practical efficiency of aeration. Beyond the energy point of view, in the situation of heating of the S04 height (Figure 3B), although the temperature was high, it was within a safe range for grain storage, up to 25 °C, because above this value the storage time is reduced and the grain is subject to changes in moisture content, thousand grain mass, electrical conductivity, fatty acid profile etc. (Paraginki et al., 2015).

Therefore, in the cases raised during the experimentation of this aeration efficiency prediction model, only the evaluation of the process efficiency from the energy point of view was considered, suggesting that future studies can be conducted to test aeration efficiency models in order to address issues such as the safe temperature for grain storage.

Table 3 presents the performance indicators of the algorithms tested in supervised machine learning to classify the aeration status of stored sunflower grains. In relation to the MLP classifier, it had 99.9212% of correct answers in the classification of the status of the aeration process tested, and there was an incorrect classification: the classifier made a mistake 1 time classifying the process as efficient and 2 times classifying it as non-efficient. Of the total of 588 moments in which the aeration process was efficient, SVM-Poly obtained 100% accuracy, but this classifier made two mistakes when evaluating the status of non-efficiency of the aeration as efficient.

Among the classifiers tested, the SVM-Radial had the lowest results for CCI and the highest for ICI, indicating that this algorithm was the one that made the most errors during the classification of the aeration status. The error of this classifier occurred for the evaluation of the non-efficiency of aeration; in 12 moments of the analyzed period, the classifier predicted the process as efficient.

When comparing the decision tree algorithms, J48 and Random Forest, the J48 algorithm stood out with higher values for CCI and lower values for ICI and errors in the confusion matrix. Figure 4 shows the two errors from the algorithm's decision tree, and it is possible to notice that there were 588 evaluations of the aeration process as efficient.

The J48 classifier was efficient to identify situations of inefficiency for aeration when there was direct heating of the grains due to aeration, in 10 situations. In the classifier's decision tree, the two incorrect classifications occurred at times when there was no aeration.

Table 3. Indicators of the performance of the classifier algorithms, with correct classification of the instances (CCI), incorrect classification of the instances (ICI), and the errors in the confusion matrix for efficiency (EF) and non-efficiency (NEF) of the aeration process

Algorithm	CCI (%)	ICI (%)	Error in confusion matrix	
			EF	NEF
Multilayer perceptron (MLP)	99.9212	0.0788	2	1
SVM-Poly	99.9475	0.0525	2	0
SVM-Radial	99.6849	0.3151	12	0
J48	99.9475	0.0525	2	0
Random forest (RF)	99.9212	0.0788	2	1

SVM-Poly - Support vector machine with polynomial kernel; SVM - Radial: support vector machine with radial kernel; and J48 - Decision trees

Regarding the indicators (Table 3), the algorithms SVM-Poly, J48 and Random Forest showed the same values for CCI, ICI and errors in the confusion matrix. Figure 5 shows the errors predicted by the SVM-Poly algorithm as a function of the status of the aeration process and the incidence of direct heating during aeration.

It can be observed in Figure 5A that the erroneous classification occurred for the non-efficiency status, in which the classifier separated it as efficient. Figure 5B shows that the

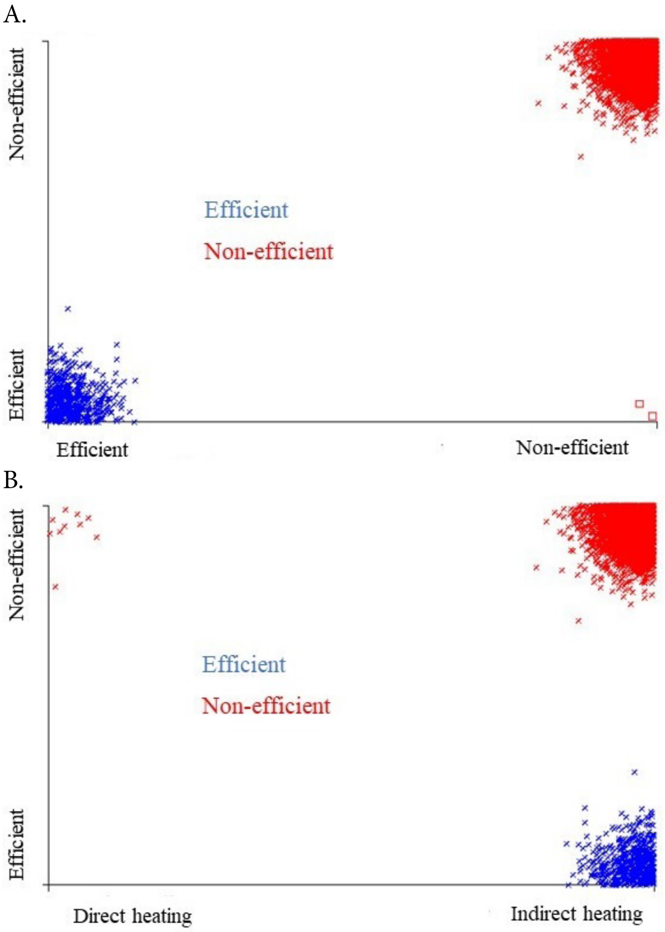
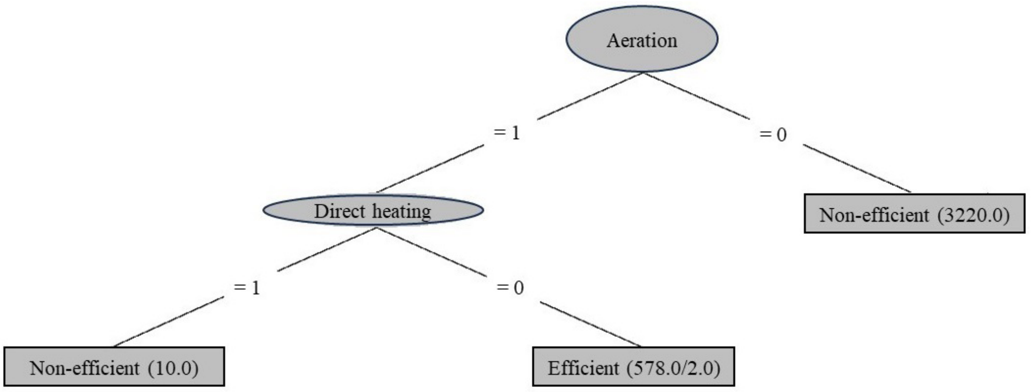


Figure 5. Incorrect classifications of the instances by the SVM-Poly algorithm as a function of aeration efficiency (A) and as a function of direct heating of the grain mass during aeration (B)



0 indicates no, and 1 indicates yes

Figure 4. J48 algorithm's decision tree for the aeration process of stored sunflower grains

classification as a function of direct heating was performed correctly, with the 10 moments of aeration being non-efficient due to direct heating during aeration, This result is similar to that observed in the classification of the J48 algorithm (Figure 4). Given these similar results generated by the different classifiers, the proposed model for evaluating aeration efficiency has applicability of use in predictive analyses of the process, since at least two algorithms were able to estimate the process correctly. Table 4 presents metrics to evaluate the performance of the algorithms tested to classify the status of the aeration process of the mass of stored sunflower grains.

The Kappa coefficient was higher for the SVM-Poly and J48 algorithms, both of which had the same magnitude (0.9980), except for the SVM-Radial classifier, which showed a lower value for the coefficient (0.9878). The other algorithms showed Kappa coefficients higher than 0.9969. According to Kotz & Johnson (1983), this coefficient is used to describe and test the reliability and precision of the classification. For Landis & Koch (1977), Kappa coefficients greater than 0.75 are characterized as excellent agreement, so all the algorithms tested in the present study showed classification in agreement.

The highest value of relative absolute error was shown by the SVM-Radial algorithm, which also had higher values for RMSE and RRSE. Regarding RAE, values close to zero indicate an ideal classification scenario (Charles, 2017), so the SVM-Poly classifier stood out, followed by J48. For the RMSE and RRSE parameters, the SVM-Poly classifier showed the lowest values, indicating a shorter distance between the classified data and the experimentally observed data (Kuhn & Jhonson, 2013).

According to the metrics analyzed, the SVM-Poly algorithm stands out in the classification of aeration status in supervised machine learning compared to the other classifiers tested in the present study, and this model is therefore recommended for implementation in neural networks to predict the aeration status of stored grains.

In the results of the clustering, the similarities of the data were found by means of the Euclidean distance and thus defined by the unsupervised machine learning, that is, the clusters were defined according to the variable of aeration efficiency: efficient and non-efficient from two clusters. Figure 6 represents the clusters grouped by the K-means algorithm, and it is possible to see the clusters performed as a function of the observed data of the aeration efficiency status estimated in the present study.

The efficient aeration statuses were grouped in cluster 0; however, 12 situations in which aeration was non-efficient were erroneously grouped in this cluster, as observed for cluster 1, in which 26 moments in which aeration was efficient were categorized in the non-efficient aeration cluster (Figure 6). Overall, there were 38 instances incorrectly clustered, approximately 1% of the data; cluster 0 was composed of 562 data (15%), and cluster 1 was composed of 3,246 data (85%).

Table 4. Metrics for evaluating the performance of the classifier algorithms, based on Kappa coefficient, relative absolute error (RAE), root mean squared error (RMSE) and root relative squared error (RRSE)

Algorithm	Kappa coefficient	RAE (%)	RMSE	RRSE (%)
Multilayer perceptron (MLP)	0.9969	0.6906	0.0276	7.6947
SVM-Poly	0.9980	0.2044	0.0229	6.3961
SVM-Radial	0.9878	1.2266	0.0561	15.6672
J48	0.9980	0.4076	0.0229	6.3953
Random forest (RF)	0.9969	0.7155	0.0266	7.4294

SVM-Poly - Support vector machine with polynomial kernel; SVM - Radial: support vector machine with radial kernel; and J48 - Decision trees

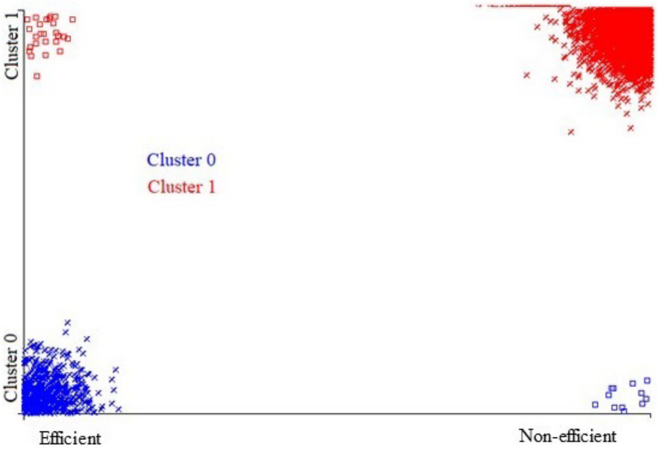


Figure 6. Result of clustering in unsupervised machine learning as a function of aeration status

Figure 7 details the grouping of the clusters, and it is possible to observe where the clustering estimation errors occurred for the aeration, cooling, heating and indirect heating situations.

Cluster 0 representing the efficiency of aeration was colored blue, and for this cluster there were situations in which when there was aeration, it was observed as non-efficient, but learning categorized it in cluster 0 (Figure 7A). As well as for non-efficiency, most of the data on aeration inefficiency are due to the fact that aeration did not occur at the evaluated moments, demonstrating that unsupervised machine learning was able to identify such a situation.

However, it can be seen that, in 26 moments when aeration was efficient, these data were grouped in cluster 1, which received a red color and was the cluster of non-efficiency of aeration. This may have occurred due to the wide variety of temperature data used in the tested database, as the algorithm may have identified some pattern between these values and the others in the same cluster, since K-means uses Euclidean distance as a metric to find the similarities of the clusters (Oliveira et al., 2022).

For the situations in which aeration promoted cooling (Figure 7B), the learning correctly clustered the data, except for the 26 data already listed previously in cluster 1, in which these were shown to be efficient in practice, but were judged as inefficient in learning. In this case the most plausible justification is that the K-means has not learned from the tests, so regardless of whether or not there was cooling, aeration can be efficient (Table 1).

Considering the behavior of clustering for the heating situations (Figure 7C), there is similarity with the results observed for cooling, since for conditions in which heating was not observed, the aeration can be efficient if it has occurred and non-efficient for the opposite case. Therefore, it can be seen that the 26 failures of unsupervised learning are due to these two rules (cooling and heating), which the system failed to learn. It

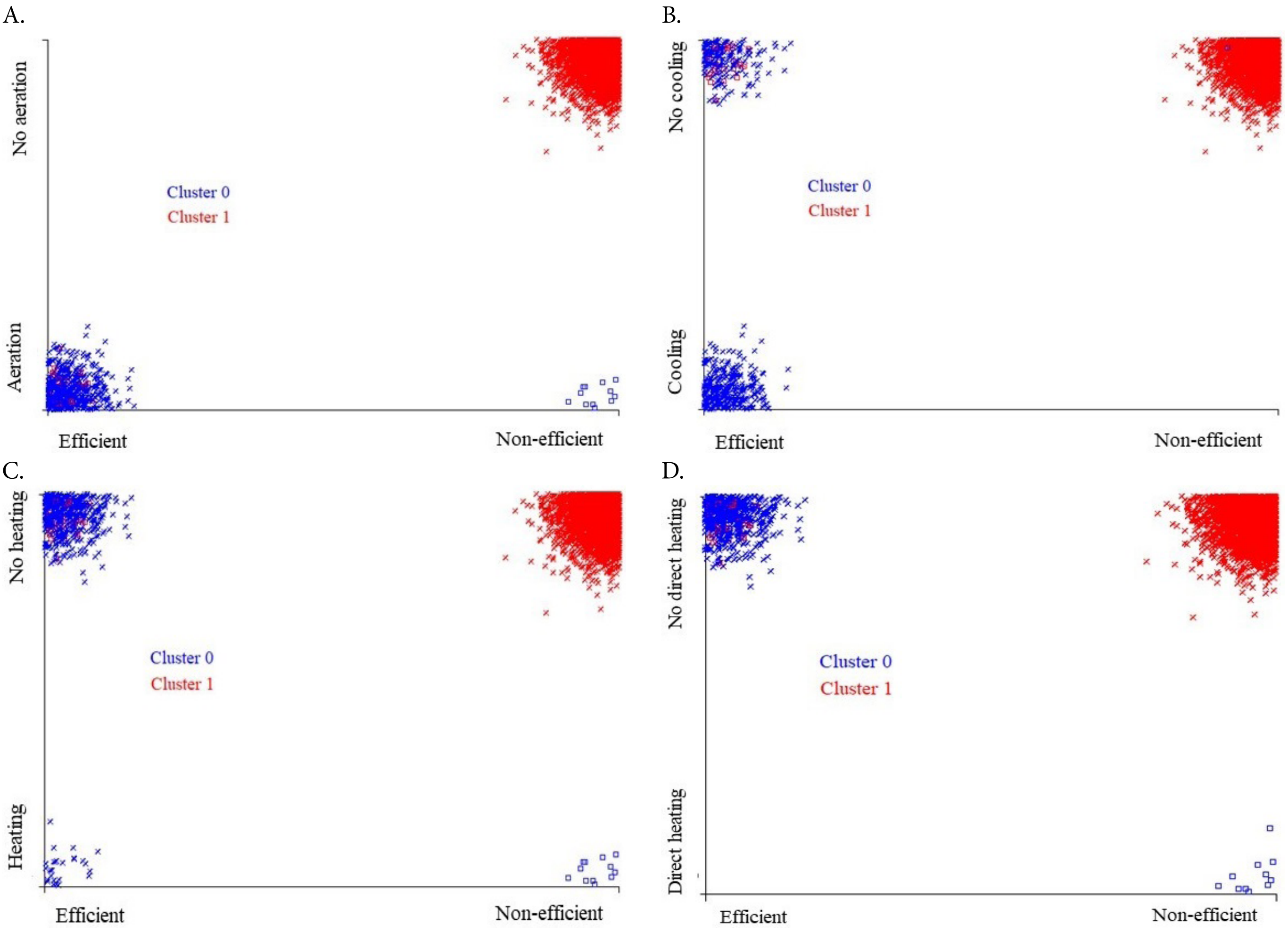


Figure 7. Representation of unsupervised machine learning clusters as a function of aeration efficiency with: aeration process (A); cooling of the grain mass (B); heating of the grain mass (C); direct heating to aeration (D)

was also observed that there were 12 situations in which there was heating and, although the aeration was non-efficient, the data were grouped in cluster 0, that is, the system could not connect the two situations of heating and non-efficiency.

Figure 7D shows the same behavior in the representation of the direct heating situation, in which the system was unable to learn from the tested data, since for the 10 moments in which there was direct heating due to aeration and, therefore, the aeration was non-efficient, the data were grouped in cluster 0.

With these results, it is possible to notice that despite having a high level of accuracy, 99%, the unsupervised machine learning showed some limitations in the estimation of decisive rules of the aeration status evaluation process, and supervised machine learning was able to estimate these rules through the J48 and SVM-Poly algorithms. The estimation errors by unsupervised learning may have occurred due to the high variety of data present in the database, more precisely due to the low concentration of indirect heating data, these being the only data in this situation that the system could not learn.

Therefore, according to what was observed and tested in the present study, the use of K-means in unsupervised machine learning is not recommended for the management of aeration efficiency evaluation processes. New tests with a more balanced database can be run to test the estimation of unsupervised learning with K-means. This algorithm is considered to be traditional unsupervised machine learning (Oliveira et al., 2022).

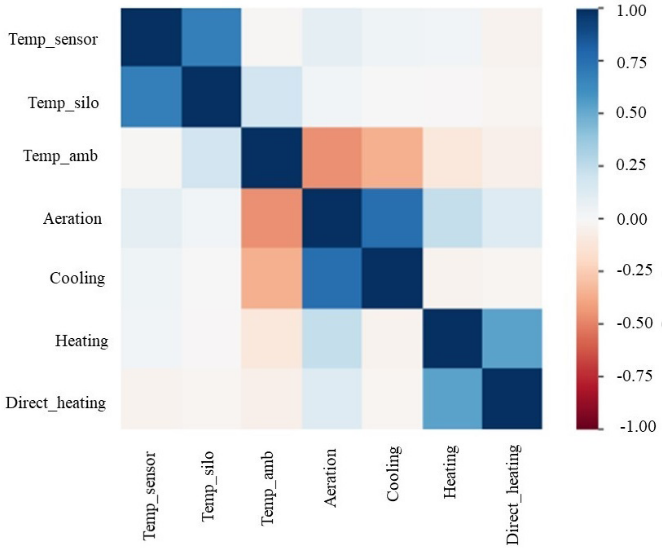
The metrics used to evaluate the supervised machine learning indicate that the Naive-Bayes classifier achieved the following means of the metrics: accuracy of 99.95, AUC of 99.98, recall of 99.94, precision of 100%, F1 score of 99.97, Kappa of 99.82 and MCC of 99.82%.

As the result of the data mining, the statistical analysis of the data was obtained through the Pearson matrix. In turn, Pearson's correlation matrix (r) signals through the intensity of the colors the correlation between the variables, that is, when the intensity of the color is dark or close to 1, the variables have a higher correlation, while the opposite indicates a lower correlation (Figure 8).

In this context, through Pearson's correlation matrix it was possible to conclude that the variables sensor temperature and silo temperature, aeration and cooling aeration, and finally heating aeration and direct heating aeration are highly correlated. On the other hand, the variables ambient temperature and silo temperature, as well as sensor temperature, have a low correlation, as well as cooling aeration and ambient temperature and silo temperature with low correlation.

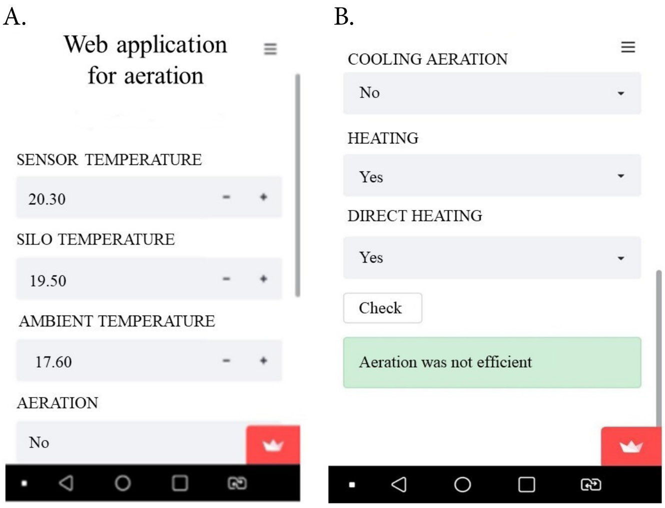
After the statistical analysis was completed, the supervised machine learning and the Web site were implemented using the Rest API with the Streamlit library to make the supervised machine learning model available run in real time (Figure 9).

The Web site was made available in real time to the producer and/or the storage unit, who were required to enter the data of sensor temperature, silo temperature, ambient temperature,



Temp_sensor: Temperature at the height of the sensors; Temp_silo: Average temperature of the silo; Temp_amb: External ambient temperature; Aeration: Occurrence of aeration; Resfrig: Occurrence of cooling aeration; Heating: Occurrence of Heating aeration; Direct_heating: Occurrence of direct heating aeration

Figure 8. Pearson's correlation matrix between the dataset variables


A.

Web application for aeration

SENSOR TEMPERATURE

20.30

SILO TEMPERATURE

19.50

AMBIENT TEMPERATURE

17.60

AERATION

No

B.

COOLING AERATION

No

HEATING

Yes

DIRECT HEATING

Yes

Check

Aeration was not efficient

Figure 9. Screenshot of the aeration system (A), and Web site made available in real time (B)

aeration, cooling aeration, heating aeration and direct heating aeration.

The Web site processes the prediction of the sunflower grain aeration system, so it was possible to check whether the aeration was efficient or non-efficient, based on specialized literature. It is worth mentioning that the data mining and modeling steps directly influenced the supervised machine learning and that it obtained a very relevant accuracy rate of 99.98%, making the result robust and accurate.

The Web application was registered as a computer program under registration number BR512022000174-8, with the title "SisAeração – Webapp para sistema de aeração utilizando aprendizado de máquinas".

Conclusion

1. The model for estimating the energy efficiency of the aeration process proved to be efficient, identifying that the energy efficiency was 97.78% during the aeration of stored sunflower grains.

2. The recommended models, J48 and SVM-Poly, for evaluating aeration efficiency have applicability of use in predictive analysis of the process.

3. Among the classifier algorithms tested in supervised machine learning, SVM-Poly showed the best metrics and indicators, being recommended for predicting the aeration status of stored grains.

4. Unsupervised machine learning with the K-means algorithm was not recommended for aeration efficiency evaluation with the tested database.

5. From data mining and modeling with machine learning, it was possible to develop the Web tool, registered under number BR512022000174-8, capable of predicting aeration efficiency.

Contribution of authors: Weder N. Ferreira Junior - Research design, data collection, analysis and interpretation, manuscript preparation and literature review. Osvaldo Resende - Research design, data analysis and interpretation, manuscript preparation, work supervision, administration and funding acquisition. Daniela C. de Oliveira - Research design, data analysis and interpretation, manuscript preparation and work supervision. Daniel E. C. de Oliveira - Data analysis and interpretation, manuscript preparation, literature review and work supervision. Elivânio dos S. Rosa - Collection, preparation of the manuscript, supervision of the work and acquisition of funding.

Supplementary documents: There are no supplementary sources.

Conflict of interest: The authors declare no conflict of interest.

Financing statement: There are no financing statements to declare.

Acknowledgement: To the Instituto Federal Goiano (Process Number: 23220.001881.2023-91), Procer, Caramuru, EMBRAPII - Empresa Brasileira de Pesquisa e Inovação Industrial, CAPES -Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, FAPEG - Fundação de Amparo à Pesquisa do Estado de Goiás, FINEP - Financiadora de Estudos e Projetos and CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico for the indispensable support to conduct this study.

Literature Cited

Charles, A. C.; Namen, A. A; Rodrigues, P. P. G. W. Comparison of data mining models applied to a surface meteorological station. Revista Brasileira de Recursos Hídricos, v.22, e58, 2017. http://dx.doi.org/10.1590/2318-0331.0217170029

Corcovia, L. O.; Alves, R. S. Aprendizagem de máquinas e mineração de dados. Revista Interface Tecnológica, v.16, p.90-101, 2019.

Durks, J. M.; Botelho, F. M.; Botelho, S. C. C.; Ruffato, S.; Hoscher, R. H. Perdas quantitativas e qualitativas em soja armazenada com aeração convencional e resfriamento artificial. Revista de Ciências Agroambientais, v.17, p.31-39, 2019. https://doi.org/10.5327/rcaa.v17i1.2861

Kotz, S.; Johnson, N. L. Encyclopedia of statistical sciences. New York: John Wiley & Sons. 1983. 352p.

Kuhn, M.; Johnson, K. Applied Predictive Modeling. New York: Springer, 2013. 593p.

Landis, J. R.; Koch, G. G. The measurement of observer agreement for categorical data. Biometrics, v.33, p.159-174, 1977. https://doi.org/10.2307/2529310

Lopes, D. de C.; Steidle Neto, A. J. Effects of climate change on the aeration of stored bean in Minas Gerais State, Brazil. Biosystems Engineering, v.188, p.155-164, 2019. https://doi.org/10.1016/j.biosystemseng.2019.10.010

Lorena, A. C.; Carvalho, A. C. P. L. F. Uma introdução às Support Vector Machines. Revista de Informática Teórica e Aplicada, v.14, p.1-25, 2013. https://doi.org/10.22456/2175-2745.5690

Mohapatra, D.; Kumar, S.; Kotwaliwale, N.; Singh, K. K. Critical factors responsible for fungi growth in stored food grains and non-chemical approaches for their control. Industrial Crops & Products, v.108, p.162-182, 2017. https://doi.org/10.1016/j.indcrop.2017.06.039

Oliveira, D. C.; Barbosa, U. C.; Bergland, A. C. R. O.; Resende, O.; Oliveira, D. E. C. G-Soja - Website with prediction on soybean classification using machine learning. Engenharia Agrícola, v.42, e20210140, 2022. http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v42nepe20210140/2022

Panigrahi, S. S.; Singh, C. B.; Fielke, J. CFD modelling of physical velocity and anisotropic resistance components in a peaked stored grain with aeration ducting systems. Computers and Electronics in Agriculture, v.179, e105820, 2020. https://doi.org/10.1016/j.compag.2020.105820

Paraginski, R. T.; Rockenbach, B. A.; Santos, R. F.; Elias, M. C.; Oliveira, M. Qualidade de grãos de milho armazenados em diferentes temperaturas. Revista Brasileira de Engenharia Agrícola e Ambiental, v.19, p.358-363, 2015. http://dx.doi.org/10.1590/1807-1929/agriambi.v19n4p358-363

Sá, T. D. V.; Amantéa, R. P.; Rocha, K. S. O.; Martins, J. H. Um modelo em dinâmica de sistemas para controle de sistemas de aeração de grãos em silos. Scientia Generalis, v.3, p.360-384, 2022. https://purl.org/27363/v3n1a32

Setiawan, N. A.; Venkatachalam, P. A.; Hani, A. F. M. Diagnosis of coronary artery disease using artificial intelligence based decision support system. Proceedings of the International Conference on Man-Machine Systems (ICoMMS), v.6, p.153-160, 2009. https://doi.org/10.48550/arXiv.2007.02854