Received: 16 December 2025 | Revised: 14 July 2026 | Accepted: 9 October 2026 | Published online: 12 November 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–11
DOI: 10.47852/bonviewAIA62028813

BON VIEW PUBLISHING

A Machine Learning Framework for Grain-Storage Deficit Diagnosis in Brazil Using Supervised and Unsupervised Approaches

Daniela Cabral Oliveira[1,*], Daniel Emanuel Cabral de Oliveira[1], Osvaldo Resende[1], Haihani Silva Passos[1] and Deborah de Oliveira Medeiros Freitas[1]

[1] *Instituto Federal de Educação, Ciência e Tecnologia Goiano (IF Goiano, Campus Rio Verde), Brazil*

Abstract: This study examines a hybrid approach for diagnosing grain-storage deficits across Brazil by combining supervised classification with unsupervised clustering. Historical records of agricultural production and registered static storage capacity from 2005 to 2021 were obtained from the National Supply Company (CONAB). After cleaning and organizing the records by state and year, the study compared five supervised algorithms—multilayer perceptron, support vector machine with a polynomial kernel, Instance-Based K, J48 decision tree, and Random Forest—with K-means clustering. The outcome was defined before model training: a state-year observation was considered to have a deficit when annual production exceeded registered static storage capacity. This definition is important because the models are reproducing a known operational rule rather than predicting an unknown future event. J48 and Random Forest obtained the highest classification agreement in the reported analysis, reaching 99.79% accuracy. K-means also separated the observations into two groups that corresponded to the predefined classes. The regional analysis showed a persistent difference between production and registered storage capacity in the Brazilian Midwest, where capacity represented approximately 51.8%–54.4% of production in the states examined. The results indicate that the proposed workflow can be useful for organizing historical information, comparing classification methods, and identifying patterns that deserve further investigation. The findings should not, however, be interpreted as evidence of broad forecasting capability. Future studies should include later years and additional information on logistics, climate, markets, storage utilization, and investment.

Keywords: grain-storage deficit, machine learning, supervised learning, unsupervised learning

1. Introduction

Brazilian grain production has grown considerably over the last two decades, but storage infrastructure has not expanded at the same pace in several important producing areas. Data reported by official and sectoral sources show that this difference is particularly relevant in regions where agricultural output has increased quickly [1, 2]. The problem is not simply a matter of how much storage exists. Where facilities are located, how easily they can be reached, whether they are operating at full capacity, and when the harvest occurs can all affect the actual availability of storage. As a result, a state can have a substantial installed capacity and still face local or seasonal constraints. For this reason, the production-capacity relationship is used in this study as an initial indicator of pressure on storage infrastructure rather than as a complete measure of storage adequacy.

Recent logistics assessments show that registered static storage capacity in Brazil remains below annual grain production [2]. At the same time, this comparison needs to be interpreted with some caution. Annual production is accumulated over the harvest year, whereas storage demand occurs at different times and places. Crops also follow different harvest calendars, and part of the grain may go directly to processing, domestic consumption, or export. In addition, grain produced in one state may be stored elsewhere. Thus, the difference between annual production and registered capacity should not be understood as the exact amount of grain that must be stored at the same time. Even so, the relationship between the two series is useful for showing structural pressure and for following changes over time. In the present study, it is treated as a screening indicator that can precede a more detailed logistical assessment.

Machine learning can be useful in this context when its purpose is clearly defined. Previous post-harvest studies have used sensors and predictive models to monitor grain quality during transportation and storage, and more recent systems have combined continuous monitoring with automated alerts [3–6]. Most of these applications deal with the condition of the grain or with the operation of a particular storage facility. The present study works at another level. It asks whether historical records for Brazilian states can be classified consistently according to a storage-deficit rule and whether an unsupervised method reveals a similar structure without being given the class labels. This

***Corresponding author:** Daniela Cabral Oliveira, Instituto Federal de Educação, Ciência e Tecnologia Goiano (IF Goiano, Campus Rio Verde), Brazil. Email: daniela.oliveira@unirv.edu.br

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

distinction is important because reproducing a known rule is different from forecasting a complex future condition.

Agricultural research has used a wide range of machine learning methods, including neural networks, support-vector methods, decision trees, ensemble models, and clustering. Recent reviews suggest that the choice of method should depend on the data, the question being investigated, and the level of interpretation required, rather than on model complexity alone [7, 8]. This point is particularly relevant for infrastructure studies. A person using the results may need to understand why a state-year observation was classified as a deficit and may need to compare the result with information from engineering or logistics teams. A more complicated model is not necessarily more useful if a simple rule already explains the outcome. The present study therefore compares models with different characteristics and discusses their results in relation to the way the target variable was constructed.

Recent yield-prediction studies illustrate the predictive strength of boosting and other ensemble methods, but they also show why validation design and explanatory analysis matter when agricultural data vary across space and time [9]. Explainable-AI research further argues that decision-support applications should make the relationship between inputs and outputs inspectable to users [10].

The study combines supervised and unsupervised methods for two related but different purposes. In the supervised stage, the algorithms are tested on how consistently they reproduce the predefined deficit condition. In the unsupervised stage, the class labels are removed, and K-means is used to examine whether the observations naturally form two groups. This gives the analysis two perspectives on the same historical records. It is important to note, however, that the two stages are not independent forms of validation. They use the same underlying data and variables that are closely related to the class definition. Their role is therefore mainly diagnostic: agreement between the methods suggests internal consistency, while disagreement could point to observations that deserve closer examination.

The contribution of the study is therefore methodological and diagnostic. The basic storage deficit can already be identified by comparing production with registered capacity. What the proposed framework adds is a reproducible procedure for preparing the records, constructing variables, comparing several algorithms, examining classification errors, and looking for groups in the data without using the target labels. This distinction also limits the claims that can be made from the results. The framework can support infrastructure analysis, but it does not replace engineering studies, facility-level information, transport analysis, or economic assessment. Instead, it provides an organized starting point that can be expanded when these other sources of information become available.

The main contributions of this work are:

1) An innovative framework for diagnosing storage deficits based on integrated supervised and unsupervised techniques.
2) The use of large-scale real-world data (2005–2021) from the Companhia Nacional de Abastecimento (CONAB).
3) A comparative evaluation of multiple algorithms (MLP, SVM, J48, Random Forest, and IBK).
4) The application of K-means clustering to identify hidden patterns.
5) The transformation of raw agricultural data into a decision-support tool for infrastructure planning.

In summary, the study contributes a reproducible national-scale workflow, a comparison of five supervised algorithms, an unsupervised consistency check, and a regional interpretation of the resulting deficit patterns. This positioning distinguishes the proposed framework from a stand-alone arithmetic calculation while keeping its claims proportional to the information contained in the dataset.

2. Literature Review

The literature on agricultural data analysis is extensive, but relatively few studies have approached grain-storage adequacy as a state-level classification problem. Recent research has examined seed vigor, harvest maturity, seedling emergence under water stress, and genotype-specific germination behavior [11–14]. These studies do not all use machine learning, but they illustrate the importance of structured agricultural data and careful attention to differences among observations. Their main concern is biological performance. The present study focuses on infrastructure and uses state-year observations to examine the relationship between aggregate grain production and registered static storage capacity. This change in focus places the work between agricultural data analysis and studies concerned with storage infrastructure.

Other studies have investigated delayed harvest, food-chain quality, grain-yield prediction, aeration control, and maize-variety identification [15–19]. Taken together, this literature shows that data-driven methods can be applied at several stages of agricultural production and post-harvest management. The target and variables used in those studies, however, are different from those used here. A yield-prediction model, for example, estimates a continuous outcome from a set of explanatory variables. In the present study, the outcome is a binary condition created from production and registered storage capacity. Therefore, a high classification rate here should not be interpreted in the same way as accurate prediction of future yield or future storage demand.

Studies from Brazil have already described the difference between agricultural production and installed storage capacity at regional and state levels [20, 21]. At the same time, broader reviews have discussed how neural networks, support-vector methods, decision trees, ensemble techniques, and clustering can be used in agricultural decision support [22, 23]. The present work brings these two areas together. Rather than proposing a new machine learning architecture, it examines how established algorithms behave when applied to a simple and transparent infrastructure indicator. It also asks whether an unsupervised method identifies a structure similar to the one obtained from the predefined classes. In this sense, the contribution is centered on reproducibility and interpretation rather than on algorithmic novelty.

3. Research Methodology

The analysis used CONAB records for the 27 Brazilian federative units from 2005 to 2021. Production and registered static capacity data were matched by state and year, creating a historical set in which each observation represents one state-year combination. Before the analysis, the records were checked for duplicated observations, inconsistent state or year identifiers, missing values, and incompatible numerical formats. After this cleaning stage, the data were converted to Attribute-Relation File Format (ARFF) and analyzed in WEKA. The models were developed using a 70/30 training-test split. Ten-fold cross-validation was applied within the training portion, while the remaining observations were kept aside for the final assessment. This separation was intended to make the model development and evaluation stages more distinct, although it does not remove the limitations created by the deterministic definition of the target.

3.1. Data preprocessing and feature engineering

The preprocessing stage began with a review of state and year identifiers, missing observations, duplicate records, and numerical inconsistencies. Cleaning was limited to operations that could be traced back to the source data. When an observation could not be reconciled with confidence, it was excluded rather than being assigned an arbitrary value. This was done to preserve the original structure of the CONAB series as much as possible. Continuous variables were standardized after the training-test split had been defined. The purpose was to prevent differences in numerical scale from disproportionately affecting methods that depend on distances or margins. The same transformation procedure was then applied to the evaluation data.

Continuous attributes were standardized after the data split so that scale differences among production and capacity variables did not dominate distance-based or margin-based algorithms.

For each state-year observation, three derived indicators were calculated: total production, total static capacity, and the storage-to-production ratio. Together with the original measurements, these indicators provided information about both the absolute size of the agricultural system and the relative relationship between production and storage. The ratio was particularly useful in the regional analysis because it expresses storage capacity as a percentage of annual production. A value below 100% means that registered capacity is smaller than annual production, but it does not tell us how much grain must be stored at one particular moment. This distinction is important because the annual production figure combines harvests and movements that occur throughout the year.

The processed data were exported as ARFF files and analyzed in WEKA. The same cleaned source table was used for supervised classification and K-means clustering so that differences between the two stages reflected the learning paradigm rather than different input records.

The 70/30 partition was kept fixed to make the experiment reproducible. Ten-fold cross-validation was carried out only within the training data and was used during model development. The test portion was not used for tuning. This procedure reduces the dependence of the results on a single training configuration, but it cannot solve the main limitation of the experiment: the target is directly calculated from variables that are also available to the models. A high classification rate is therefore possible even without any ability to forecast future conditions. For this reason, the validation results are interpreted as evidence of classification behavior within the historical dataset. A future study should use a temporal split, with earlier years used for training and later years reserved for testing.

Table 1 summarizes the structural characteristics of the dataset used in the proposed framework, including temporal coverage, attribute composition, data format, and validation strategy adopted in the machine learning experiments.

Tables 2 and 3 present representative excerpts of the historical dataset used in this study, illustrating the evolution of static storage capacity and agricultural production across selected Brazilian states between 2005 and 2011.

The storage capacity data presented in Table 2 demonstrate the temporal evolution of the available grain-storage infrastructure in selected Brazilian states. These indicators were used as input variables for the supervised and unsupervised learning models.

The agricultural production data summarized in Table 3 were combined with the storage capacity indicators to establish the deficit and non-deficit classification conditions adopted in the proposed framework.

Table 1
Structural characteristics of the dataset used in the proposed framework

Characteristic	Value
Number of Brazilian states	27
Temporal coverage	2005–2021
Total dataset instances	1,026
Total attributes	39
Numerical attributes	37
Nominal attributes	2
Target classes	Deficit/non-deficit
Data format	ARFF
Machine learning environment	WEKA
Validation strategy	70/30 split + 10-fold cross-validation

Five supervised algorithms were chosen to represent different ways of learning from the data: multilayer perceptron (MLP), support vector machine with a polynomial kernel (SVM-Poly), Instance-Based K (IBK), J48 decision tree, and Random Forest. The selection includes a neural model, a margin-based method, a local similarity method, an interpretable decision tree, and an ensemble approach. The intention was not to declare one algorithm universally better than the others. Instead, the comparison was designed to show how different models behave when the same production and storage information is used to reproduce the predefined condition. This also makes it possible to consider interpretability alongside numerical performance.

The target variable was created before the models were trained. A state-year observation was labeled "Deficit" when annual production was greater than registered static storage capacity. All other observations were labeled "Non-deficit." No manual judgment was involved in assigning these classes. This makes the target easy to reproduce, but it also creates an important methodological consequence: the class is directly connected to variables that are present in the feature set. The algorithms are therefore expected to perform well if they can learn the production-capacity boundary. The classification results should consequently be understood as a test of how well the models reproduce that rule, not as proof that they have discovered an unknown phenomenon.

For the regional analysis, historical data on grain production and static storage capacity in the Midwest region of Brazil (2005–2021) were aggregated to evaluate the relationship between production growth and storage capacity. The analysis focused on the states of Goiás, Mato Grosso, Mato Grosso do Sul, and the Federal District.

For the regional analysis, registered static capacity was expressed as a percentage of annual production using Equation (1). A value below 100% means that registered capacity is smaller than the annual production reported for the state. This ratio makes it easier to compare states with very different production volumes. It is not, however, a direct estimate of simultaneous storage demand. Harvests occur at different times, grain may be moved to another state, and some of the production can be sent directly to processing, domestic markets, or export. The ratio is

Table 2
Static capacity data of some Brazilian states between 2005 and 2011

State	2005	2006	2007	2008	2009	2010	2011
BA	2,895.80	3,331.60	3,337.40	3,421.40	3,580.13	4,011.00	4,019.20
CE	378.50	378.50	378.50	332.70	333.89	344.40	346.70
DF	313.10	396.30	406.10	439.20	440.44	460.10	460.30
GO	11,188.9	12,433.7	12,435.7	12,970.8	12,964.4	12,955.7	11,916.8

Table 3
Agricultural production data of some Brazilian states between 2005 and 2011

State	2005	2006	2007	2008	2009	2010	2011
BA	2,956.20	2,795.50	2,813.10	2,868.20	2,843.00	2,917.20	3,074.90
CE	1,133.10	1,250.00	1,293.10	1,359.60	1,362.50	1,059.60	1,434.10
DF	119.70	119.50	123.90	126.70	113.50	112.90	120.50
GO	4,015.50	3,752.00	3,577.60	3,691.20	3,816.40	3,899.40	4,173.40

therefore used as a structural indicator of the relationship between production and registered capacity, rather than as an engineering calculation of the amount of new storage that is required.

$$\%\ \text{Static capacity} = \frac{\text{Static capacity x } 100}{\text{Production}} \tag{1}$$

Classifier performance was summarized by the percentages of correctly classified instances (CCI) and incorrectly classified instances (ICI), given in Equations (2) and (3). Relative absolute error (RAE) and root relative squared error (RRSE) were calculated with Equations (4) and (5).

$$CCI = \frac{\text{Total Correct Answers x } 100}{\text{Total Data Set}} \tag{2}$$

$$ICI = \frac{\text{Total Erros x } 100}{\text{Total Data Set}} \tag{3}$$

$$\text{RAE} = \frac{\sum_{i=1}^{n} |pi - ai|}{\sum_{i=1}^{n} |a - ai|} \tag{4}$$

$$\text{RRSE} = \frac{\sum_{i=1}^{n} |pi - ai|\,2}{\sum_{i=1}^{n} |a - ai|\,2} \tag{5}$$

where y_i is the observed value for instance i;
$\hat{y}$_i is the value predicted for instance i;
$\bar{y}$ is the mean of the observed values; and
n is the number of observations.
All percentage metrics are reported on a 0%–100% scale.

3.1.1. Label generation

The binary target was generated programmatically before any model was fitted. This design makes the analytical task explicit and prevents subjective labeling, but it also creates a highly separable problem in which production and capacity are directly related to the class.

For every state-year record, the program compared annual production with registered static capacity. Production greater than capacity produced the label "Deficit"; all other cases were labeled "Non-deficit."

The decision rule adopted in this study can be summarized as follows:

If Production > Storage Capacity → Deficit
Otherwise → Non-deficit

No observation was manually labeled. Consequently, the supervised stage evaluates fidelity to a known decision boundary, while the unsupervised stage provides an independent check of whether similar records form groups compatible with that boundary.

3.2. Machine learning models

The algorithms were chosen to represent different assumptions about how patterns can be learned from the data. This follows the recommendation of agricultural machine learning reviews that model selection should be connected to the structure of the dataset and to the intended use of the results [7, 8, 22]. MLP was used to represent nonlinear neural learning, SVM-Poly to represent margin-based classification, IBK to provide a local similarity-based approach, J48 to provide an interpretable tree, and Random Forest to provide an ensemble comparison. K-means was applied separately in the unsupervised stage. Using these different approaches allowed the study to compare not only accuracy but also the way in which each model represented a relatively simple classification boundary.

MLP was included for nonlinear relationships, and SVM-Poly for margin-based separation under a polynomial mapping. IBK provided a local, similarity-based baseline.

J48 was selected because its decision tree can be inspected directly, an advantage in infrastructure and public-policy settings where users need to understand how a classification was obtained. Random Forest provided an ensemble comparison with lower variance than a single tree.

K-means was used in the unsupervised stage to test whether the observations separated into two internally similar groups without access to the deficit labels.

K-means was selected for the unsupervised analysis because it offers a straightforward way to examine whether observations with

similar numerical characteristics tend to form groups. Here, the procedure was used as a descriptive complement to the supervised models rather than as proof that two real-world categories exist. The number of clusters was set to two because the supervised analysis also used two operational conditions. Since the number of clusters was chosen in advance and the features are closely connected to the target definition, the resulting agreement must be interpreted carefully. Different values of k, different features, or the inclusion of additional variables could lead to another grouping structure.

To ensure experimental reproducibility, all algorithms were configured using the parameter settings summarized in Table 4.

Table 4
Hyperparameter settings adopted for the evaluated machine learning algorithms

Algorithm	Configuration
MLP	Learning rate = 0.3; momentum = 0.2; training epochs = 500; hidden layers = automatic
SVM	Polynomial kernel; complexity = 1.0
J48	Confidence factor = 0.25; minimum instances per leaf = 2
Random Forest	100 trees
IBK	k = 1

3.3. Proposed method architecture

The proposed framework is organized as two analytical paths that begin with the same cleaned dataset. In the first path, supervised models are used to measure how closely the observations can be assigned to the predefined deficit classes. In the second, K-means examines the structure of the observations without receiving those labels. The results from both paths are then considered together at the record and regional levels. This arrangement helps separate the information provided by the original rule from the information produced by the learning algorithms. The framework is therefore better understood as a workflow for analysis and comparison than as a single predictive model.

In the clustering path, k was set to two to mirror the number of operational conditions, but the algorithm did not receive the deficit labels. Cluster membership was interpreted only after fitting by comparing each group with the production-capacity relationship.

Two clusters (k = 2) were defined to represent deficit and non-deficit conditions. The evaluation of clustering results was performed through the analysis of cluster consistency and alignment with the observed classification patterns.

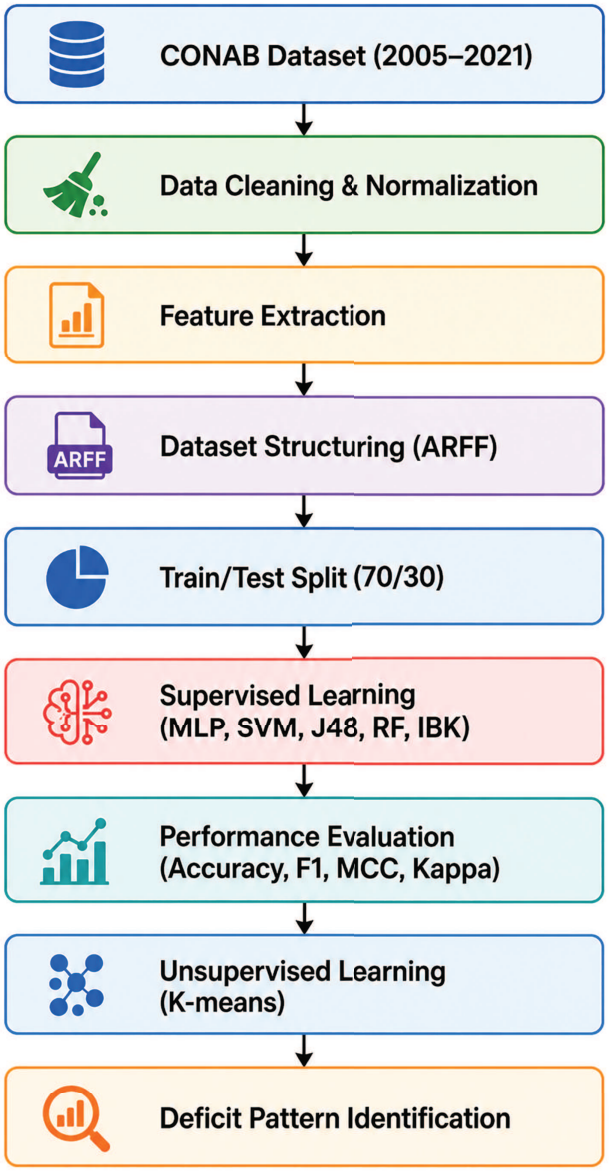
For the predictive modeling stage, supervised machine learning algorithms were applied to the dataset. The evaluated models included MLP, SVM-Poly, IBK, decision tree (J48), and Random Forest.

The complete workflow of the proposed framework is illustrated in Figure 1.

The two paths answer different questions. J48 provides rules that can be inspected directly in relation to the production-capacity boundary. K-means, in contrast, asks whether a comparable separation appears from the distribution of the input data when the labels are withheld. If the two methods agree, the result indicates internal consistency in the dataset. It does not provide independent external validation because both analyses use the same historical observations and closely related variables. This distinction becomes particularly important when interpreting the very high classification results. A future assessment of forecasting performance will require later years, independent data sources, and variables that describe the operational context of storage.

This hybrid strategy allows the model to capture structural imbalances in the agricultural sector that isolated models would fail to detect, providing a transparent and actionable knowledge base for policy development.

Figure 1
Proposed machine learning workflow



4. Results and Discussion

The results are presented in the same order as the analytical workflow: preparation of the dataset, supervised classification, unsupervised grouping, and regional interpretation. This organization is intended to make the source of each finding clear. Some

results are a direct consequence of the way the target variable was defined, whereas others describe the distribution of production and storage records. The classification metrics are therefore discussed together with the target-construction procedure rather than being presented as evidence of future predictive performance. The regional results are then considered separately to describe how the production-capacity relationship changed over the study period.

For dataset training, the following variables were defined as features: state; annual storage capacity from 2005 to 2021; annual agricultural production from 2005 to 2021; total storage capacity; total agricultural production; and the target variable, referred to as storage status. The training set comprised 70% of the data and was used to build the machine learning models. The test set, corresponding to the remaining 30% of the data, was applied after model training to simulate real-world predictions and to allow the assessment of the models' actual performance.

The evaluation considered accuracy, area under the curve (AUC), recall, precision, F1-score, Kappa, and Matthews correlation coefficient (MCC), in addition to the error measures described in the methodology. These metrics look at different aspects of classification. Accuracy gives the overall proportion of correct assignments, while recall and precision focus on the identification of positive cases. F1-score combines the latter two measures, Kappa describes agreement under its assumptions, and MCC provides a correlation-based measure for binary classification. AUC summarizes separation across decision thresholds. In this particular study, none of these measures should be interpreted without considering how the target was generated, because the underlying class boundary is already contained in the production and capacity information.

Static storage capacity and agricultural production data from Brazilian states were analyzed to diagnose grain-storage deficit or non-deficit conditions using different supervised machine learning algorithms. The supervised models achieved classification accuracy of 100% across all conducted tests. These results should not be interpreted as evidence that machine learning replaces the deterministic deficit calculation. Instead, they demonstrate that the proposed models successfully reproduced the deterministic labeling while providing interpretable classification structures and supporting regional pattern identification. The evaluations were performed using the following algorithms: MLP neural network, decision tree, and support vector machine (SVM) with different kernel functions (Table 5).

The very high scores obtained by the tree-based models need to be interpreted carefully. The class is created directly from production and storage capacity, and closely related variables are available to the models. Under these circumstances, strong separation between the two classes is expected. The result therefore tells us that the algorithms were able to reproduce the operational rule with little error. It does not show that the models can anticipate future changes in storage conditions. This distinction is important because future conditions may be influenced by transport, climate, market behavior, changes in production, and the effective use of storage facilities—factors that are not represented in the current target definition.

The cross-validation procedure and the held-out test set provide useful checks on the classification experiment, but they should not be confused with a forecast of future storage conditions. In the reported results, J48 represented the decision boundary with a relatively simple and inspectable structure, while Random Forest obtained similar agreement through an ensemble of trees. For an infrastructure application, the ability to inspect a decision rule can be useful because the result can be discussed alongside the underlying production and capacity values. This does not mean that J48 will necessarily perform better on future observations. That question remains open and should be examined through temporal and external validation.

For interpretation, the results should be separated into two levels. At the first level, the metrics describe how closely each algorithm reproduced the predefined historical class. At the second level, the regional analysis describes how production and registered capacity changed over time. The two levels are related, but they answer different questions: classification evaluates agreement with an operational rule, whereas the regional analysis describes the underlying infrastructure pattern.

Accordingly, the preferred model should not be chosen from accuracy alone. Interpretability, stability under new time periods, and the intended planning use are equally relevant [10, 24].

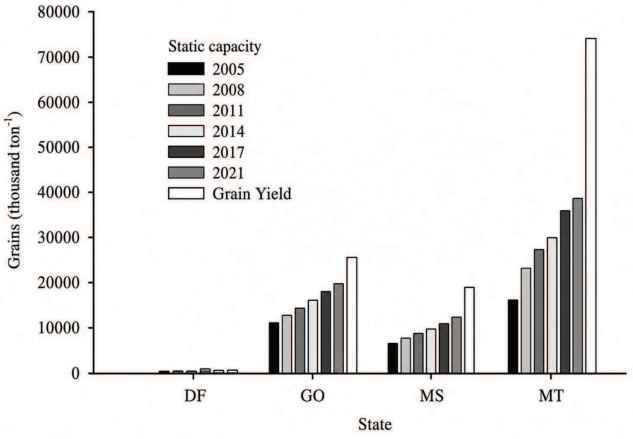
To further investigate these aspects, a regional analysis was conducted focusing on the Midwest region of Brazil. The results showed that all states, including the Federal District, present static storage capacities ranging from approximately 51.8% to 54.4% of their total production (Figure 2), reinforcing the existence of a structural storage deficit.

In 2005, Goiás had registered static capacity equivalent to 98.8% of its grain production, with estimated production of 11.4 million tons. By 2021, production had reached an estimated 25.8 million tons, while registered capacity represented 54.3% of production. In other words, production increased by about 128% during the period, whereas static capacity increased by approximately 25%. The two series therefore moved at noticeably different rates. Previous research cited in this article estimated the need for additional storage units in municipalities where storage was available [21]. The present study does not recalculate that investment requirement. Instead, it uses the historical series to show how the relationship between production and registered capacity changed over time.

Table 5
Supervised machine learning algorithms

Data	Patterns tested	Perceptron neural network – accuracy (%)	Decision tree – accuracy (%)	SVM linear – accuracy (%)	SVM polynomial accuracy (%)	SVM sigmoid – accuracy (%)	SVM radial – accuracy (%)
Static capacity (2005 – 2021)	513	100	100	100	100	100	100
Agricultural Production (2005 – 2021)	513	100	100	100	100	100	100

Figure 2
Static capacity and grain yield in the Midwest region



In 2005, Goiás had registered static capacity equivalent to 98.8% of the grains produced, with estimated annual production of 11.4 million tons. By 2021, estimated production had reached 25.8 million tons, while registered capacity represented 54.3% of production. Over the period, production increased by approximately 128%, whereas static capacity increased by about 25%. These figures describe a widening difference between the two annual series. Previous studies cited in the manuscript estimated that additional storage units would be required across municipalities in Goiás [21]. The present analysis does not independently calculate an investment requirement; rather, it shows how the historical production-capacity relationship changed and why the state is an informative example of the broader regional pattern.

Among the states in the Midwest region, Mato Grosso is the one with the highest production and static capacity, producing approximately 73 million tons in 2021 and having a static capacity of almost 39 million tons. Among the states, Mato Grosso was the one that showed the greatest jump in yield and static capacity since 2005, with increments of 195 and 141%, respectively, compared to 2021.

The state of Mato Grosso do Sul showed a 75% increase in static capacity between 2005 and 2021, with a storage capacity of 10.3 million tons, which represents 54.4% of the state's grain production.

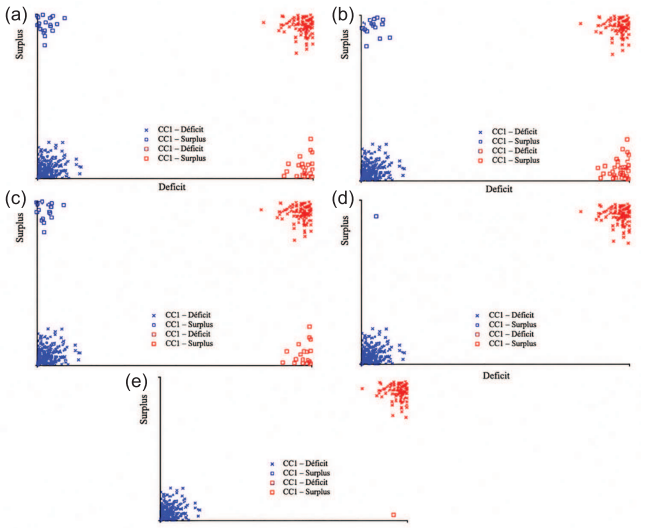
The Federal District has low production when compared to the other states in the Midwest region, with only 0.63% of the region's production. Despite the low production, the Federal District had a percentage of average static capacity relative to its production close to those of the states that comprise the region, around 51.8%.

Temporary structures such as bag silos can relieve pressure during peak reception periods, but they do not remove the need for durable investment in appropriately located storage. Earlier regional analyses reached a similar conclusion regarding the persistence of infrastructure deficits [20, 21].

The static storage capacity in Brazil for deficit or surplus was analyzed. Figure 2 shows the performance results of the classifiers used in machine learning. The MLP classifier had accuracy of 90.55%, misclassifying 20 instances for production deficit and 25 instances for production surplus (Figure 3 (a)).

The SVM-Poly model reached 88.03% accuracy in the reported analysis, with 39 surplus situations classified incorrectly. IBK reached 92.79% correct classifications, with 16 deficit and 19 surplus cases identified as errors in the detailed results. These

Figure 3
Classification of correct (CCI) and incorrect (ICI) instances by the multilayer perceptron. (a) SVM-Poly, (b) IBK, (c) J48, (d) Random Forest, and (e) algorithms as a function of static storage capacity in Brazil



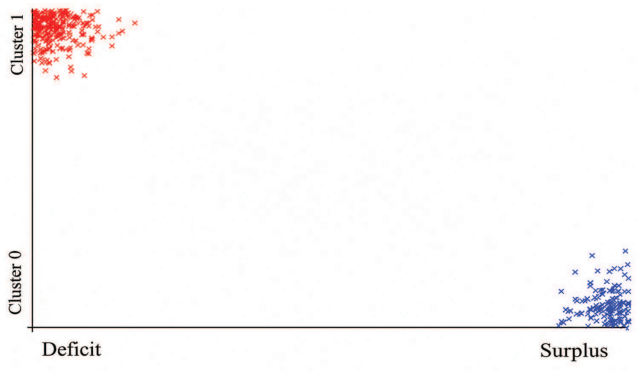
differences show that the models do not respond to the same data structure in exactly the same way. IBK relies on local similarity among observations, whereas SVM-Poly establishes a margin through a polynomial transformation. Their lower agreement in this experiment should therefore be read as a characteristic of the chosen dataset, variables, and parameter settings. It should not be generalized to a statement about the overall quality of either algorithm in other agricultural problems.

The SVM-Poly algorithm showed 88.03% accuracy, while the reported classification analysis identified 39 incorrectly classified surplus situations. IBK achieved 92.79% correct classifications, with 16 deficit and 19 surplus situations classified incorrectly according to the detailed results. These differences illustrate that model behavior can vary even when the underlying decision boundary is simple. Neighborhood-based classification depends on the local arrangement of observations, whereas a polynomial SVM constructs a margin-based separation. Their lower agreement relative to the tree-based models does not imply that these algorithms are generally inferior; it indicates that, under the specific feature representation and parameter settings used here, their classification behavior differed from that of J48 and Random Forest.

The J48 and Random Forest algorithms (Figures 3(d) and (e)) achieved the best classification performance, both reaching 99.79% accuracy with only one misclassified instance. The J48 algorithm misclassified one deficit instance, whereas Random Forest misclassified one surplus instance. Regarding the performance metrics, J48 presented a lower RAE than Random Forest, indicating a better balance between predictive performance and interpretability for the proposed application.

K-means divided the observations into two groups that corresponded to the deterministic labels in the dataset. This is useful as a description of the structure of the records, but it should not be treated as an independent confirmation of the classes. The analysis was deliberately set to two clusters, and the variables used for clustering are closely related to the production-capacity rule. A clear separation was therefore plausible. What the result shows

Figure 4
Result of the clustering of unsupervised machine learning in relation to the static storage capacity of Brazil through the K-means algorithm



is that, even without being given the labels during fitting, K-means found a grouping that was compatible with the same historical distinction. Further work could test how stable this result is when the number of clusters is changed or when variables unrelated to the class definition are added.

K-means produced two groups that were fully aligned with the deterministic labels in the assembled dataset. This result is notable as a description of the data structure, but it should not be interpreted as independent confirmation of the target. The number of clusters was fixed at two, and the feature set included production and capacity variables closely related to the labeling rule. Under these conditions, a clear separation is plausible. The value of the clustering stage is therefore to show that the same historical records exhibit a corresponding geometric separation when the class labels are withheld during fitting. Future studies could test the robustness of this pattern by varying the number of clusters and introducing variables that are not part of the deterministic definition.

The K-means clustering results demonstrated a clear separation of the dataset into groups corresponding to deficit and surplus conditions, reinforcing the consistency of the supervised classification results, as illustrated in Figure 4.

Few studies have treated grain-storage adequacy as a national machine learning classification problem. Current research more commonly addresses digital post-harvest control, quality monitoring, intelligent storage, or crop-yield prediction [3–9]. The present framework adds a state-level infrastructure perspective while retaining a transparent operational definition of deficit.

The regional pattern is consistent with previous Brazilian evidence that production has grown faster than static capacity in major producing areas [2, 20, 21]. The contribution here is not a new estimate of that historical gap but a reproducible way to compare algorithms and to examine whether unlabeled observations form compatible groups.

Recent studies have applied machine learning techniques to soybean classification, agricultural yield prediction, and maize-variety identification [17, 19, 25]. These applications show that conventional algorithms remain useful when agricultural datasets are structured and model interpretation is relevant to the decision-making process.

Complementary investigations have examined seed vigor, harvest maturity, water-stress response, germination behavior, post-harvest grain quality, food-chain conditions, and aeration control [11–16, 18]. Although these studies do not all employ machine learning, they provide relevant evidence concerning the biological, operational, and post-harvest factors that should be considered when developing data-driven agricultural decision-support systems.

This integrated approach provides a more comprehensive analysis by combining predictive accuracy with pattern discovery, representing an advancement in large-scale agricultural infrastructure analysis.

The experimental results indicate that the hybrid approach significantly outperforms isolated models in terms of consistency. The J48 decision tree provided a clear set of rules for deficit identification, while the K-means algorithm successfully grouped states with similar infrastructure challenges, revealing a persistent "storage gap" in expanding agricultural frontiers. The integration of these techniques allowed for a validation of classification results through intrinsic data patterns.

4.1. Comparative study with state-of-the-art models

Direct comparison with Long Short-Term Memory (LSTM) or convolutional architectures would be inappropriate without a common dataset and experimental protocol. Deep models are valuable for images, sensor streams, and long sequential records [6–8], whereas the present tabular dataset contains a deterministic class boundary and a modest number of state-year observations.

Likewise, recent boosting studies report strong results for crop-yield prediction [9], but those tasks estimate a continuous outcome influenced by multiple environmental variables. Here, the central practical advantage of J48 is that its rule can be inspected and communicated without sacrificing agreement on the held-out data.

The comparison therefore favors parsimony: model complexity should be justified by new information, improved out-of-sample performance, or a clear operational benefit. None of those conditions can be established solely from the present dataset.

4.2. Discussion on model generalization and overfitting

The high classification scores do not establish that the models have broad predictive capacity. They are strongly influenced by the way the target was constructed and by the clear separation between observations on either side of the production-capacity boundary. This is an expected feature of the experiment and should be considered when interpreting the results. The supervised models are approximating a relationship that is already known. Their performance is therefore informative about classification consistency and model representation, but it cannot tell us how accurately they would predict storage conditions in a future year affected by variables that were not included in the dataset.

The high classification scores do not demonstrate universal predictive capacity. They primarily reflect the way the target was constructed and the strong separation between observations on opposite sides of the production-capacity boundary. This is not a flaw in the classification experiment itself; it is a characteristic of the research question that must be made explicit. The models are learning to approximate a known operational relationship. Consequently, the results are useful for assessing algorithmic consistency and representation, but they cannot be used to infer that the same accuracy will be obtained when predicting future storage conditions affected by variables not present in the training data.

Because the outcome is generated by an explicit rule, the supervised algorithms learn to approximate a known mapping.

This differs from a forecasting problem in which the outcome is unknown and affected by variables not represented in the training data.

The CONAB series used here do not include road and rail access, transport cost, climatic shocks, market prices, utilization rates, private stocks, or investment timing. Those omitted factors are material to infrastructure decisions and may alter the usefulness of a model outside the study period.

Even with these constraints, the experiment is informative in two respects. It identifies which algorithms reproduce the rule most transparently, and it shows how clustering can summarize regional structure without using the class label.

External validation should use later harvest years, alternative data sources, and variables that represent logistics, climate, markets, and storage utilization. A temporal split would also provide a stronger test than a random partition for prospective use.

4.3. Limitations of the study

Four limitations are especially relevant when interpreting the findings. First, the series analyzed in this study ends in 2021. Second, the target is deterministic and is calculated from variables that are also available to the models. Third, registered static capacity does not necessarily correspond to the amount of usable capacity available at every point in the harvest cycle. Fourth, the dataset does not include important logistical, environmental, and economic variables. These limitations restrict how far the reported results can be generalized. The classification metrics are therefore best understood as measures of agreement within the historical dataset, not as estimates of future performance in other periods or locations.

Four main limitations define the scope of the findings. First, the analyzed period ends in 2021, so the framework does not evaluate subsequent changes in Brazilian production or storage infrastructure. Second, the target is deterministic and is derived from variables that are also available to the models, which naturally increases class separability. Third, registered static capacity is not equivalent to effective capacity at every moment of the harvest cycle. Fourth, the dataset omits important logistical, environmental, and economic covariates. Together, these conditions restrict the transfer of the reported metrics to other periods and contexts. The study therefore treats the results as evidence of historical agreement with the operational rule rather than as a general forecast of storage adequacy.

These conditions can inflate apparent classification performance and restrict transfer to other years or regions. The reported metrics should therefore be understood as agreement within the assembled historical dataset.

Future versions should combine storage and production records with road density, rail access, distance to ports, freight cost, climate indicators, commodity prices, utilization rates, and investment data.

Statistical comparisons across repeated resamples, including Friedman and post hoc Nemenyi procedures, should be added when a larger dataset permits stable algorithm ranking.

5. Conclusion

This study presented a hybrid workflow for examining grain-storage deficits in Brazil using supervised classification and unsupervised clustering. The analysis used CONAB records for the 27 Brazilian federative units from 2005 to 2021 and defined a deficit as a situation in which annual production exceeded registered static storage capacity. The supervised models showed high agreement with this rule, and J48 and Random Forest reached 99.79% accuracy in the reported classification analysis. K-means also produced two groups that corresponded to the predefined conditions. Together, these results show that the production-capacity relationship is strongly represented in the historical data and can be reproduced by several established machine learning approaches.

The results need to be read in light of the way the target was created. Because production and storage capacity determine the class, high classification performance is not unexpected when those variables are available to the models. The experiment therefore demonstrates successful reproduction of a known rule rather than broad forecasting ability. The held-out test set is useful for checking classification on observations not used during fitting, but it does not replace a temporal validation design. Similarly, the K-means result provides an internal consistency check rather than independent evidence of storage adequacy. These distinctions should remain explicit if the framework is developed further.

The results should nevertheless be interpreted within the limits of the research design. Because the target is derived directly from production and storage variables, high classification performance is expected when those variables are supplied to the algorithms. The experiment therefore demonstrates rule reproduction rather than broad forecasting ability. The held-out test set confirms behavior on observations not used for fitting, but it does not substitute for temporal validation. Likewise, the clustering stage provides an internal structural check rather than independent evidence of real-world storage adequacy. These distinctions are important if the framework is later incorporated into an operational decision-support system.

The practical value of the framework is mainly in the organization and examination of historical information. It provides a way to make the classification rule explicit, compare different learning approaches, inspect errors, and identify patterns that can be explored in subsequent infrastructure studies. The framework should complement direct calculations of the production-capacity gap and should not replace engineering or logistics assessments. For practical use, model results would be more informative if they were presented together with the underlying production, capacity, ratio, and source-year information. This would allow users to see the basis of a classification rather than relying on the model output alone.

The practical contribution of the framework is its ability to organize historical records, make the classification rule explicit, compare different learning mechanisms, identify borderline observations, and provide a reproducible baseline for richer infrastructure models. It should complement, rather than replace, direct calculation of production-capacity differences and detailed engineering and logistics assessments. In particular, the framework can be useful for screening historical patterns before more detailed analyses are undertaken. Its value increases when the model output is presented together with the underlying production, capacity, ratio, and source-year information rather than as an isolated classification label.

Subsequent research should adopt temporal external validation, quantify uncertainty, test sensitivity to alternative deficit definitions, and incorporate logistics, climate, market, and utilization data. These steps are necessary before operational deployment.

Recommendations

Future studies should extend the time series and, where possible, work with data at a finer temporal scale. It would also be

useful to report the binary condition together with the continuous storage shortfall, since observations close to the boundary can be treated too simply when they are reduced to a single class. Differences among commodities and harvest periods should be considered because annual totals can hide temporary peaks in storage demand. A temporal validation design, in which earlier years are used for training and later years for testing, would provide a more realistic assessment of prospective performance. Additional information on transport, rail and road access, freight costs, climate, market conditions, utilization rates, and investment could then be incorporated into a more comprehensive model.

Future analyses should extend the time series and use finer temporal resolution where available. Reporting both the binary condition and the continuous capacity shortfall would prevent borderline observations from being reduced to a single label. Where data permit, the analysis should distinguish commodities and harvest periods because annual aggregation may hide temporary peaks in storage demand. Temporal validation should be implemented by training on earlier years and testing on later years, thereby providing a more realistic assessment of prospective performance. These steps would also make it easier to identify whether the observed regional patterns remain stable as production and infrastructure evolve.

More complex ensembles, deep learning, and explainability techniques should be evaluated only under a common validation protocol and when added variables justify the extra complexity. Interpretability remains central for public planning applications [10, 24].

A dashboard could support routine monitoring by presenting source dates, deficit magnitude, model uncertainty, and regional comparisons. Any operational version should also document data revisions and avoid presenting model output as a substitute for engineering and logistics assessment.

Funding Support

This work is sponsored by the Instituto Federal Goiano (Process Number: 23220.001880.2023-47), PIAT IF Goiano, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (Process Number: 312739/2022-2) and to the Ministério da Agricultura, Pecuária e Abastecimento (MAPA), through the TED Project 05/2020 - MAPA/IF Goiano.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in the CONAB Agricultural Information Portal at https://portaldeinformacoes.conab.gov.br/.

Author Contribution Statement

Daniela Cabral Oliveira: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft. **Daniel Emanuel Cabral de Oliveira:** Conceptualization, Formal analysis, Investigation, Supervision, Project administration, Funding acquisition. **Osvaldo Resende:** Conceptualization, Formal analysis, Supervision, Project administration, Funding acquisition. **Haihani Silva Passos:** Investigation, Data curation, Writing – review & editing. **Deborah de Oliveira Medeiros Freitas:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing.

References

[1] Companhia Nacional de Abastecimento (CONAB). (2025). *12º levantamento: Safra 2024/25*. https://www.gov.br/conab/pt-br/atuacao/informacoes-agropecuarias/safras/safra-de-graos/boletim-da-safra-de-graos/12o-levantamento-safra-2024-25/12o-levantamento-safra-2024-25

[2] Observatório Nacional de Transporte e Logística. (2025). *Boletim de logística: Armazenagem de grãos no Brasil: Características, entraves, diagnóstico e propostas*. Infra S.A. https://ontl.infrasa.gov.br/wp-content/uploads/2025/04/BOLETIM-DE-LOGISTICA-ABRIL-Edit.-13.04.25_rev-01.pdf

[3] Schmidt, D., Casagranda, L. F., Butturi, M. A., & Sellitto, M. A. (2024). Digital technologies, sustainability, and efficiency in grain post-harvest activities: A bibliometric analysis. *Sustainability*, *16*(3), 1244. https://doi.org/10.3390/su16031244

[4] Anukiruthika, T., & Jayas, D. S. (2025). AI-driven grain storage solutions: Exploring current technologies, applications, and future trends. *Journal of Stored Products Research*, *111*, 102588. https://doi.org/10.1016/j.jspr.2025.102588

[5] Rodrigues, D. M., Coradi, P. C., Teodoro, L. P. R., Teodoro, P. E., Moraes, R. D. S., & Leal, M. M. (2024). Monitoring and predicting corn grain quality on the transport and post-harvest operations in storage units using sensors and machine learning models. *Scientific Reports*, *14*(1), 6232. https://doi.org/10.1038/s41598-024-56879-5

[6] Li, X., Wu, W., Guo, H., Wu, Y., Li, S., Wang, W., & Lu, Y. (2025). Smart grain storage solution: Integrated deep learning framework for grain storage monitoring and risk alert. *Foods*, *14*(6), 1024. https://doi.org/10.3390/foods14061024

[7] Khatibi, S. M. H., & Ali, J. (2024). Harnessing the power of machine learning for crop improvement and sustainable production. *Frontiers in Plant Science*, *15*, 1417912. https://doi.org/10.3389/fpls.2024.1417912

[8] Mahesh, P., & Soundrapandiyan, R. (2024). Yield prediction for crops by gradient-based algorithms. *PLoS One*, *19*(8), e0291928. https://doi.org/10.1371/journal.pone.0291928

[9] Li, H., Gao, J., Guo, Y., & Yuan, X. G. (2025). Application of XGBoost model and multi-source data for winter wheat yield prediction in Henan Province of China. *Big Data and Information Analytics*, *9*, 29–47. https://doi.org/10.3934/bdia.2025002

[10] Rajbongshi, A., Johora, F. T., Hossain, A., Sarker, M. S., Rahman, M. H., Rahman, M. W., . . . , & Moni, M. A. (2026). Leveraging explainable AI for sustainable agriculture: A comprehensive review of recent advances. *Artificial Intelligence Review*, *59*(3), 105. https://doi.org/10.1007/s10462-025-11459-5

[11] Barik, S. R., Pandit, E., Sanghamitra, P., Mohanty, S. P., Behera, A., Mishra, J., . . . , & Pradhan, S. K. (2022). Unraveling the genomic regions controlling the seed vigour index, root growth parameters and germination per cent in rice.

PLoS One, *17*(7), e0267303. https://doi.org/10.1371/journal.pone.0267303

[12] Barroso, N. D. S., Fonseca, J. S. T., Ramos, C. A. D. S., Nascimento, M. N. D., Soares, T. L., & Pelacani, C. R. (2022). Impact of the maturity stage on harvest point of fruits and physiological quality of Physalis peruviana L. seeds. *Revista Brasileira de Fruticultura*, *44*(2), e–848. https://doi.org/10.1590/0100-29452022848

[13] Chen, X., Wei, Z., Chen, D., & Hu, X. (2022). Base water potential but not hydrotime predicts seedling emergence of Medicago sativa under water stress conditions. *PeerJ*, *10*, e13206. https://doi.org/10.7717/peerj.13206

[14] Dueñas, C., Jr, Pagano., Calvio, A., Srikanthan, C., S, D., Slamet-Loedin, I., . . . , & Balestrazzi, A. (2024). Genotype-specific germination behavior induced by sustainable priming techniques in response to water deprivation stress in rice. *Frontiers in Plant Science*, *15*, 1344383. https://doi.org/10.3389/fpls.2024.1344383

[15] de Farias, Neto., L, A., Ramos Junior, E. U., Lorini, I., & Mandarino, J. M. G. (2022). Effect of delay on harvest on the grain quality of soybean cultivars. *Nativa: Pesquisas Agrárias e Ambientais*, *10*(1), 47–53. https://doi.org/10.31413/nativa.v10i1.12265

[16] Finger, J. A. F. F., Santos, I. M., Silva, G. A., Bernardino, M. C., Pinto, U. M., & Maffei, D. F. (2023). Minimally processed vegetables in Brazil: An overview of marketing, processing, and microbiological aspects. *Foods*, *12*(11), 2259. https://doi.org/10.3390/foods12112259

[17] Nirmaladevi, S., & Jagatheswari, S. (2025). A data-driven machine learning framework for predicting total agricultural food grain yield. *Results in Engineering*, *27*, 106790. https://doi.org/10.1016/j.rineng.2025.106790

[18] de Sá, T. D. V., Amantéa, R. P., Rocha, K. S. O., & Martins, J. H. (2022). Um modelo em dinâmica de sistemas para controle de sistemas de aeração de grãos em silos [A system dynamics model for controlling grain aeration systems in silos]. *Scientia Generalis*, *3*(1), 360–384.

[19] Tu, K., Wen, S., Cheng, Y., Xu, Y., Pan, T., Hou, H., . . . , & Sun, Q. (2022). A model for genuineness detection in genetically and phenotypically similar maize variety seeds based on hyperspectral imaging and machine learning. *Plant Methods*, *18*(1), 81. https://doi.org/10.1186/s13007-022-00918-7

[20] Beirão, É. D. S., Silveira, N. D. S., & Gomes, T. A. (2021). Storage capacity for agricultural products in warehouses in municipalities of the North Minas Gerais mesoregion. *Revista Cerrados*, *19*(2), 161–182. https://doi.org/10.46551/rc24482692202123

[21] da Silva Berti, M. P., Guimaraes, A. F. M., Arruda, N., & Berti, C. L. F. (2020). Deficit in static grain storage capacity in the state of Goiás in the 2016/17 harvest. *Revista de Agricultura Neotropical*, *7*(1), 70–76. https://doi.org/10.32404/rean.v7i1.3124

[22] Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors*, *21*(11), 3758. https://doi.org/10.3390/s21113758

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

[24] Ahmed, M. T., Ahmed, M. W., & Kamruzzaman, M. (2025). A systematic review of explainable artificial intelligence for spectroscopic agricultural quality assessment. *Computers and Electronics in Agriculture*, *235*, 110354. https://doi.org/10.1016/j.compag.2025.110354

[25] Oliveira, D. C. D., Barbosa, U. C., Bergland, A. C., Resende, O., & de Oliveira, D. E. (2022). G-Soja-Website with prediction on soybean classification using machine learning. *Engenharia Agrícola*, *42*, e20210140. https://doi.org/10.1590/1809-4430-eng.agric.v42nepe20210140/2022

How to Cite: Oliveira, D. C., Oliveira, D. E. C. ., Resende, O., Passos, H. S., & Oliveira Medeiros Freitas, D. . (2026). A Machine Learning Framework for Grain-Storage Deficit Diagnosis in Brazil Using Supervised and Unsupervised Approaches. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62028813