REVISTA OWL (*OWL Journal*)
www.revistaowl.com.br – ISSN: 2965-2634

INTELLIGENT MODEL FOR SOIL FERTILITY CLASSIFICATION BASED ON MACHINE LEARNING AND INTEGRATED INTO THE TELLURIUM SYSTEM

MODELO INTELIGENTE PARA CLASSIFICAÇÃO DA FERTILIDADE DO SOLO BASEADO EM APRENDIZADO DE MÁQUINA E INTEGRADO AO SISTEMA TELLURIUM

DOI: 10.5281/zenodo.22016973

Laenor Henrique Tenório Dantas Brandão[1]
Hector Alexis Miranda[2]
Wilker Alves Morais[3]
Rodes Angelo Batista da Silva[4]
Adriano Ferraz da Costa[5]
Heyde Francielle do Carmo França[6]
Renato Domiciano Silva Rosado[7]
Suelen Cristina Mendonça Maia[8]
Lucas Peres Angelini[9]
Marconi Batista Teixeira[10]

1 Programa de Pós-Graduação em Ciências Agrárias - Agronomia, Instituto Federal Goiano – Campus Rio Verde, Rio Verde, GO, Brasil, E-mail: laenor.henrique@gmail.com.
2 Programa de Pós-Graduação em Ciências Agrárias - Agronomia, Instituto Federal Goiano – Campus Rio Verde, Rio Verde, GO, Brasil, E-mail: laenor.henrique@gmail.com.
3 Doutor em Ciências Agrárias - Agronomia, Instituto Federal Goiano – Campus Rio Verde, Rio Verde, GO, Brasil.
4 Doutora em Engenharia Agrícola, Instituto Federal Goiano – Campus Rio Verde, Rio Verde, GO, Brasil.
5 Doutor em Ciência da Computação, Instituto Federal Goiano, Rio Verde, GO, Brasil.
6 Doutorado em Ciência da Computação, Universidade Federal de Goiás, UFG, Goiânia, GO, Brasil.
7 Doutor em Fitotecnia, Universidade Federal de Viçosa, Viçosa, MG, Brasil.
8 Doutorado em Agronomia, Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, Botucatu, SP, Brasil.
9 Doutorado em Física, Universidade Federal do Mato Grosso, Cuiabá, MT, Brasil.
10 Doutorado em Irrigação e Drenagem, Universidade de São Paulo, Piracicaba, SP, Brasil.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

ABSTRACT

The growing demand for technologies capable of supporting decision-making in agriculture has driven the development of intelligent systems focused on soil fertility management. In this context, this study aimed to deve lop and evaluate a machine learning model to complement the functionalities of the Tellurium platform, assisting in soil fertility classification and the validation of agronomic management recommendations. The platform already generates recommendations for fertilization, liming, and soil management based on established agronomic criteria, with the machine learning model incorporated as an additional decision-support layer. To achieve this, a pipeline was structured comprising the following stages: data preprocessing, data quality verification, feature selection, variable standardization, training and comparison of classification algorithms, stratified cross-validation, hyperparameter optimization via GridSearchCV, and performance evaluation using metrics such as accuracy, precision, recall, F1-score, and the confusion matrix. Among the evaluated algorithms, Random Forest demonstrated the best performance, achieving an accuracy of 93.56%, precision of 93.82%, recall of 93.56%, and an F1-score of 93.59%, thereby showing a high capacity for generalization to unseen data. The model was exported, serialized, and validated in a development environment, ready for future integration into the Tellurium platform. However, its final implementation remains in an adaptation phase due to version incompatibilities between libraries used in the interface development. The results highlight the potential of the proposed hybrid approach, which integrates agronomic knowledge and machine learning to enhance the reliability of recommendations and strengthen the platform as a decision-support tool for soil fertility management.

Keywords: Artificial intelligence; Decision support; Digital agriculture; Predictive algorithm.

RESUMO

A crescente demanda por tecnologias capazes de apoiar a tomada de decisões na agricultura impulsionou o desenvolvimento de sistemas inteligentes voltados para o manejo da fertilidade do solo. Nesse contexto, este estudo teve como objetivo desenvolver e avaliar um modelo de aprendizado de máquina para complementar as funcionalidades da plataforma Tellurium, auxiliando na classificação da fertilidade do solo e na validação de recomendações de manejo agronômico. A plataforma já gera recomendações de adubação, calagem e manejo do solo com base em critérios agronômicos estabelecidos, sendo o modelo de aprendizado de máquina incorporado como uma camada adicional de suporte à decisão. Para isso, estruturou-se um fluxo de trabalho (*pipeline*) composto pelas seguintes etapas: pré-processamento de dados, verificação da qualidade dos dados, seleção de atributos, padronização de variáveis, treinamento e comparação de algoritmos de classificação, validação cruzada estratificada, otimização de hiperparâmetros via *GridSearchCV* e avaliação de desempenho utilizando métricas como acurácia, precisão, *recall*, *F1-score* e matriz de confusão. Entre os algoritmos avaliados, o *Random Forest* apresentou o melhor desempenho, alcançando acurácia de 93,56%, precisão de 93,82%, *recall* de 93,56% e *F1-score* de 93,59%, demonstrando, assim, alta capacidade de generalização para dados inéditos. O modelo foi exportado, serializado e validado em um ambiente de desenvolvimento, estando pronto para futura integração à plataforma Tellurium. No entanto, sua implementação final permanece em fase de adaptação devido a incompatibilidades de versão entre as bibliotecas utilizadas no desenvolvimento da interface. Os resultados destacam o potencial da abordagem híbrida proposta, que integra conhecimento agronômico e aprendizado de

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

máquina para aumentar a confiabilidade das recomendações e fortalecer a plataforma como uma ferramenta de suporte à decisão para o manejo da fertilidade do solo.

Palavras-chave: Algoritmo preditivo; Agricultura digital; Inteligência artificial; Suporte à decisão.

1. INTRODUCTION

Given the need to reconcile increased agricultural productivity with the conservation of natural resources, proper soil management plays a strategic role in promoting sustainability and resilience in production systems (ABREU *et al*., 2020). Soil health assessment depends on the integrated analysis of its physical and chemical attributes, with laboratory analyses and fertility mapping being essential tools for diagnosing nutrient availability and supporting the rational application of amendments and fertilizers (SARMA *et al*., 2024). In this context, the adoption of adequate fertilization and management practices contributes to more efficient use of inputs, favors the maintenance of soil quality, and promotes sustainable productivity gains without compromising the productive capacity of this resource over time (SITOE *et al*., 2025).

However, soil analyses generate costs that become prohibitive for small- and medium-sized producers. Furthermore, proper management is only possible with laboratory analyses (ADOMAKO *et al*., 2022; WEN *et al*., 2022) and their correct interpretation, since the absence of good recommendations can cause damage to the system, such as over-liming or improper use of nitrogen-based fertilizers. In this scenario, the lack of tools to monitor and guide plant mineral nutrition is longstanding. Studies by Soares *et al*. (2009) have already demonstrated this need, where approximately 70% of producers did not know how to interpret soil analyses. In view of this, more recent studies have shown that the problem of data interpretation persists over the years, where smallholder farmers still face difficulties in data interpretation, in addition to limited access to quality technical assistance, which demonstrates that, despite technological advances, there is still a need for tools that serve this small segment of farmers (DESCONSI & de SÁ, 2024; FONSECA *et al*., 2024).

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

The existing limitations due to the lack of data integration, machinery with outdated technologies, and the scarcity of financial resources imply low adoption of technological tools, in addition to limiting possible support from adequate technical assistance (CASTRO *et al*., 2024; MMBANDO, 2025). In this scenario, the need for technological tools capable of keeping up with the dynamics of agricultural systems becomes evident. Among these technologies, the use of Artificial Intelligence (AI) has been used as a decision-support tool for agricultural management. Thus, the use of AI has been increasing both in the development of everyday technologies to increase productivity and has been increasingly consolidating alongside digital agriculture technologies (CASINI *et al*., 2026). However, there is a barrier regarding the use of these tools, with inexperience and misinterpretation of data being the greatest among them.

Therefore, concomitantly, the use of Machine Learning (ML), which is a subfield of artificial intelligence, has been disseminated in the global market as a promising tool in several areas (WOLFERT *et al*., 2023). Furthermore, the training of algorithms to assist practical functions, such as fertilization calculations and fertility prediction, has been developing and shows itself to be a promising tool in the agricultural market.

Unlike systems based exclusively on machine learning models, the Tellurium platform was designed to provide agronomic recommendations based on consolidated technical criteria. In this context, the machine learning model was incorporated as a complementary decision-support layer, acting in soil fertility classification and validation of recommendations generated by the platform. Given the above, this study aimed to develop and evaluate a Machine Learning model integrated into the Tellurium platform as a complementary mechanism to the agronomic recommendations already implemented, assisting in soil fertility classification and validation of management recommendations.

2. MATERIALS AND METHODS

2.1 Data source and types

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

The dataset used in this study consists of chemical attributes relevant to the assessment and classification of soil fertility, such as Nitrogen (N), Phosphorus (P), Potassium (K), Sulfur (S), Zinc (Zn), Copper (Cu), Manganese (Mn), Boron (B), as well as variables such as hydrogen potential (pH), Organic Carbon (OC), and Electrical Conductivity (EC). Initially, an Exploratory Data Analysis (EDA) step was conducted to characterize the dataset, identify variable types, detect missing values, inconsistencies, and possible anomalies, as well as to perform preprocessing and refinement necessary to ensure data quality. The dataset consists of 880 instances (samples) and 13 attributes (features), which were used for training and evaluating the machine learning models. The database was obtained from the public Kaggle repository (KAGGLE, 2023), available in Comma-Separated Values (CSV) format.

2.2 Data preprocessing and treatment

Data treatment was performed following the processing workflow provided by Faceli *et al*. (2021). For the development of this step, the Python® language (PYTHON SOFTWARE FOUNDATION, 2024) and the Google Colab® platform (GOOGLE COLABORATORY, 2024) were used. Initially, the dataset was imported using the Pandas library. Subsequently, an exploratory data analysis was performed to verify the structure of the tabular data, check for missing or inconsistent data, and assess data distribution. After that, the categorical fertility variables were encoded using LabelEncoder, while numerical attributes were standardized using StandardScaler.

After preprocessing, the data were split into Training (75%) and Test (25%) sets and evaluated using Stratified Cross-Validation, a technique that aims to ensure homogeneous balancing of the data and approximate sample sizes (SZEGHALMY & FAZEKAS, 2023). In addition, four algorithms were trained: Random Forest (RF), Gradient Boosting (GB), Logistic Regression (RL), and Support Vector Machine (SVM). The use of stratified cross-validation for model performance evaluation was also employed by Mahesh &

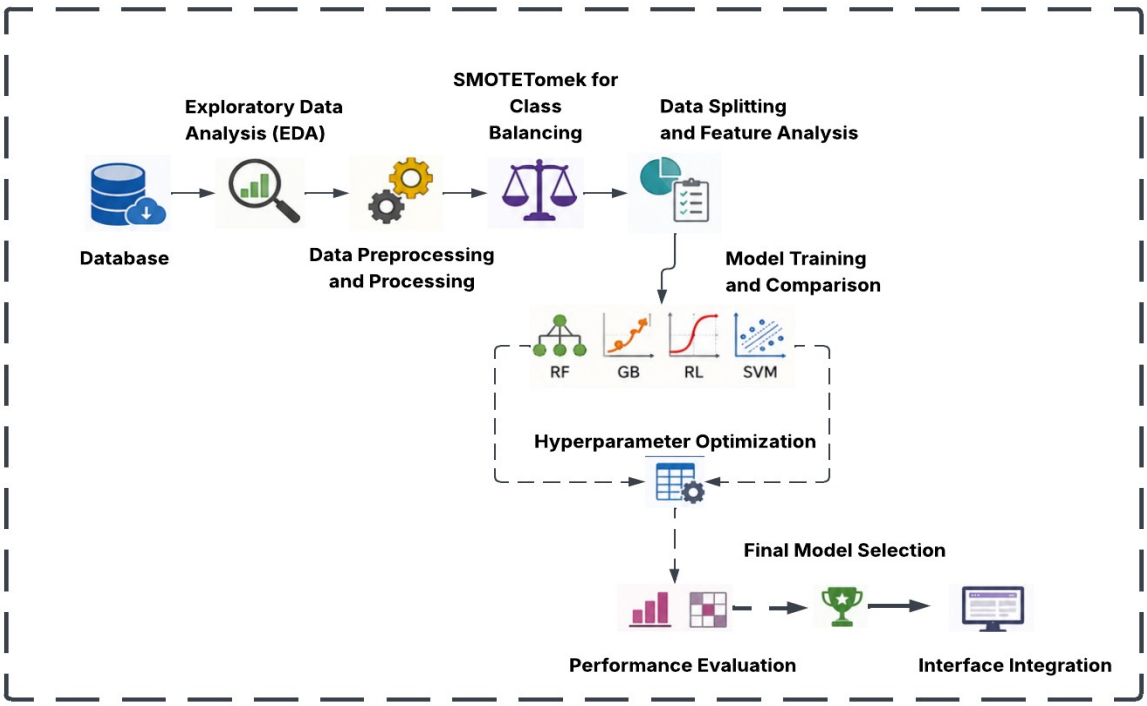
A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Soundrapandiyan (2026), who combined feature selection, hyperparameter optimization, and classification metrics to predict soil fertility.

Thus, the variables in this study were converted to numerical format via Label Encoding, which is used to improve data interpretability by the algorithm. However, in this study no outlier treatment was performed in order to preserve the real variability of Brazilian soils and thereby ensure the robustness of the model in different field scenarios.

Regarding class balancing, deliberately, due to the imbalanced nature of the original dataset, the SMOTE (Synthetic Minority Over-sampling Technique) technique (CHAWLA *et al*., 2002; LEMAÎTRE *et al*., 2017) was applied. This technique creates synthetic samples interpolated between neighbors from the minority class to balance the data; similar methods have also been used by Islam *et al*. (2022). Figure 1 illustrates, in a schematic form, the workflow of procedures performed throughout this study.

Figure 1 - Flowchart of data preprocessing, comparison, and model training.



A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Figure 1 - Flowchart of data preprocessing, comparison, and model training. Source: Prepared by the authors based on the literature and study workflow (2026). Steps include: dataset upload using Pandas; Exploratory Data Analysis (EDA); data preprocessing with LabelEncoder, StandardScaler, and SMOTE; data splitting and feature analysis; model training and comparison (Random Forest, Gradient Boosting, Logistic Regression, and Support Vector Machine); hyperparameter optimization using GridSearchCV and Stratified K-Fold Cross-Validation; performance evaluation and final model selection; model export and interface integration.

2.3 Feature selection and correlation analysis

To optimize performance and reduce possible dimensionality, the Pearson Correlation Matrix was applied. Its purpose is to measure how much one variable influences another, identifying highly correlated attributes, detecting possible multicollinearity issues, among other functions. The calculation performed by the matrix has a correlation coefficient that ranges from -1 to +1 and indicates the direction and intensity of the relationship between two variables (PEARSON, 1895).

2.4 Model architecture and training

Using the train_test_split function from the Scikit-learn library. The split was performed in a stratified manner (stratify), preserving the class proportion in both subsets. Additionally, random_state = 42 was adopted to ensure experiment reproducibility. The algorithms were implemented using the RandomForestClassifier, GradientBoostingClassifier, LogisticRegression, and SVC (Support Vector Classification) classes, available in the Scikit-learn library (PEDREGOSA *et al*., 2011). Four algorithms were compared in this study: Random Forest (BREIMAN, 2001); Gradient Boosting (FRIEDMAN, 2001); Logistic Regression (HOSMER; LEMESHOW; STURDIVANT, 2013); and Support Vector Machine (SVM) (CORTES; VAPNIK, 1995).

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Table 1 - Specific Hyperparameters

Algorithms	Hyperparameters	Assessed Values
Random Forest	*n_estimators;*	100, 200, 300
	max_depth	10, 15, 20, *None*
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 2, 4
Gradient Boosting	*n_estimators*	100, 200
	learning_rate	0,01; 0,05; 0,1
	max_depth	3, 5, 7
	min_samples_split	2, 5
Logistic Regression	C	0,01; 0,1; 1; 10
	penalty	12
Support Vector Machine (SVM)	C	0,1; 1,0; 10,0
	gamma	*scale, auto*
	kernel	*rbf*

Source: Research data (2026).

2.5 Model deployment and integration into the application

After identifying Random Forest as the best-performing model, the optimized classifier was persisted using the Joblib library (JOBLIB DEVELOPMENT TEAM, 2025), ensuring its storage for later use on the Tellurium platform. In addition to the trained model itself, the data standardization object (StandardScaler), the list of variables selected in the preprocessing step, and the metadata essential for reproducibility of the inference pipeline used during training were preserved. The recording of these elements aimed to ensure that

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

new samples submitted to the application received exactly the same treatment applied during the model development phase.

Thus, all steps of variable selection, attribute standardization, and classification remained consistent between the training and platform operation environments. Subsequently, a computational routine was developed to integrate the model into the application. This routine automates the loading of the trained model, preprocessing of user-input data, inference execution, and return of the recommendation generated by the algorithm, enabling its real-time application on the platform.

The machine learning model does not replace the pre-existing recommendation mechanism on the Tellurium platform. Its implementation will aim to complement the decision-making process, providing a second evaluation based on patterns learned from the data used in training. The recommendations remain grounded in consolidated agronomic criteria, while the model will act as an independent validation mechanism for the classifications generated by the platform. Thus, it is expected to increase system robustness and minimize the occurrence of inconsistencies, without replacing the technical knowledge that underlies agronomic recommendations.

2.6 Inference pipeline validation

After exporting and serializing the trained model, a functional test was performed in a development environment to validate the integrity of the inference pipeline (PEDREGOSA *et al*., 2011) prior to its future integration into the Tellurium platform. In this step, all serialized components were successfully restored, including the Random Forest model, the variable standardization object (StandardScaler), the list of selected attributes, and the metadata associated with training. The test allowed verification of the correct loading of files, consistency of the processing flow, and execution of predictions in a controlled environment, demonstrating the technical feasibility of integrating the model into the platform.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Subsequently, a sample from the test set was submitted to the same processing flow adopted during the training phase, encompassing variable selection, attribute standardization, prediction generation, and calculation of class membership probabilities. This procedure enabled verification of the consistency of the inferential process, confirming that the model's behavior remained unchanged after its integration into the application.

2.7 Use of Artificial Intelligence

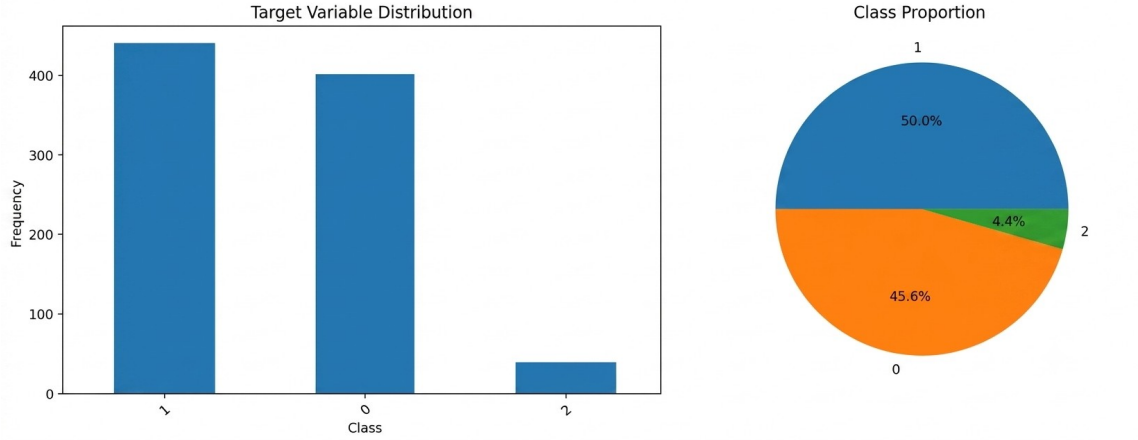
During the preparation of this work, the authors used conventional search platforms such as Google Scholar and ScienceDirect for article retrieval. Additionally, the authors used ChatGPT (version GPT-5.2) for specific article search and reference review, and DeepSeek (version DeepSeek-V4) for specific article search, formatting, Portuguese/English translation, interpretation of results, and identification of Python libraries. These AI tools were used as complementary resources and did not replace the authors' own intellectual work, analysis, or decision-making. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

3. RESULTS AND DISCUSSION

During data treatment, after performing exploratory data analysis, due to class imbalance, where the "medium" class represented about 5% of the data, as shown in Figure 2, the SMOTETomek technique (TOMEK, 1976; BATISTA, PRATI, & MONARD, 2004; LEMAÎTRE *et al*., 2017) was applied to generate synthetic samples and remove noise. Authors such as Abdullah (2025) also used data balancing methods in view of datasets with imbalanced classes, highlighting the importance of homogeneous data.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

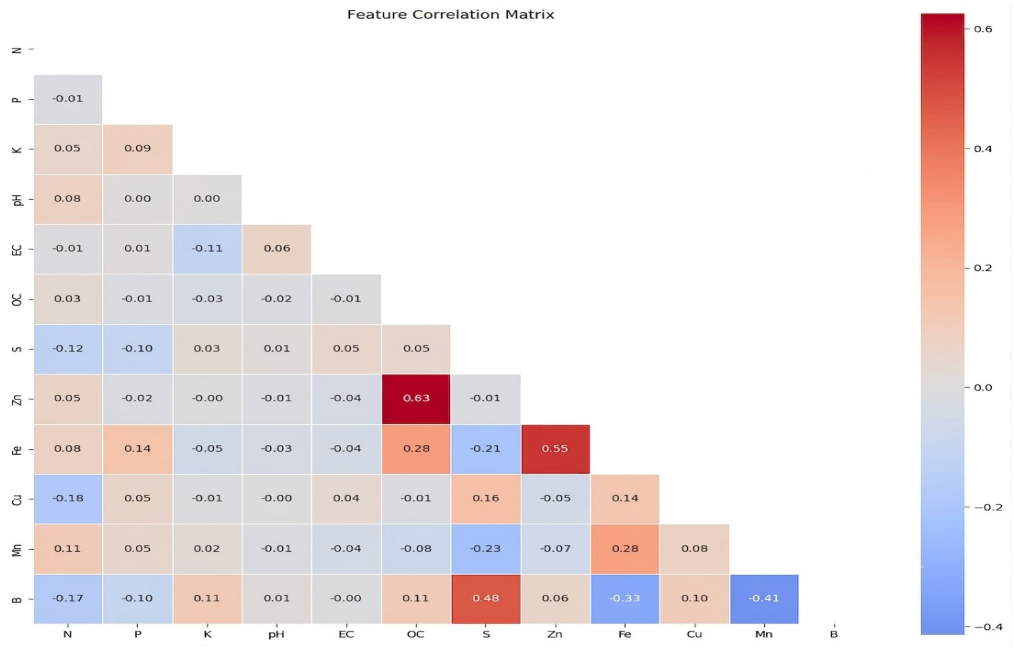
Figure 2 - Class distribution before normalization: 0 - Low; 1 - High; 2 - Medium.



Source: Research data (2026).

Furthermore, during processing, the Pearson Correlation Matrix was applied to identify the degree of linear association between the chemical attributes, verifying the existence of highly correlated variables and possible redundancies in the database, as shown in Figure 3.

Figure 3 - Correlation matrix of the features.



Source: Research data (2026).

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

The correlation matrix shows a predominance of low correlations among the soil chemical attributes, with most coefficients presenting values lower than 0.30 (Figure 3). Furthermore, only a few associations present moderate to strong magnitude, highlighting the positive correlation between Organic Carbon (OC) and Zinc (Zn) (r = 0.63), followed by Zn and Iron (Fe) (r = 0.55), Sulfur (S) and Boron (B) (r = 0.48), in addition to the negative correlation between Manganese (Mn) and Boron (B) (r = -0.41). Overall, the results indicate low multicollinearity among the variables, suggesting that the attributes provide complementary information for algorithm training. However, although some elements have correlation above 0.60, it is not high enough to justify removing one of the variables. Studies such as Carbajal-Llosa *et al.* (2025) highlight the importance of correlation matrix analysis for ML; the authors emphasize in their study that low correlation among variables favors information complementarity and reduces the risk of potential redundancy effects between them.

Table 2 presents the score classification of the analyzed variables. In view of this, it is observed that Nitrogen was the variable with the greatest statistical relevance for classification, presenting an approximate score of 507, a value higher than the others. This difference suggests that Nitrogen contents have a high capacity to distinguish themselves from the other classified attributes in the present database, thus being the variable with the greatest discriminative value among all analyzed elements. Despite its relevance, the fact that N presented much higher values demonstrates that it was significantly relevant for classification. Although Nitrogen has high values and its ability to classify in isolation, this does not reduce the significance of the other analyzed nutrients.

Table 2 - Classification of features.

Feature	***Score***
Nitrogen (N)	507.643522
Phosphorus (P)	21.274703

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Copper (Cu)	8.686957
Potential of Hydrogen (pH)	3.898211
Sulfur (S)	2.105332
Iron (Fe)	1.676912
Potassium (K)	1.641299
Manganese (Mn)	1.291842
Boron (B)	1.223354
Zinc (Zn)	0.924088

Source: Prepared using data from the study (2026). This method was implemented using the Pandas library (MCKINNEY, 2010). Additionally, feature selection was performed using the scikit-learn library (PEDREGOSA *et al*., 2011).

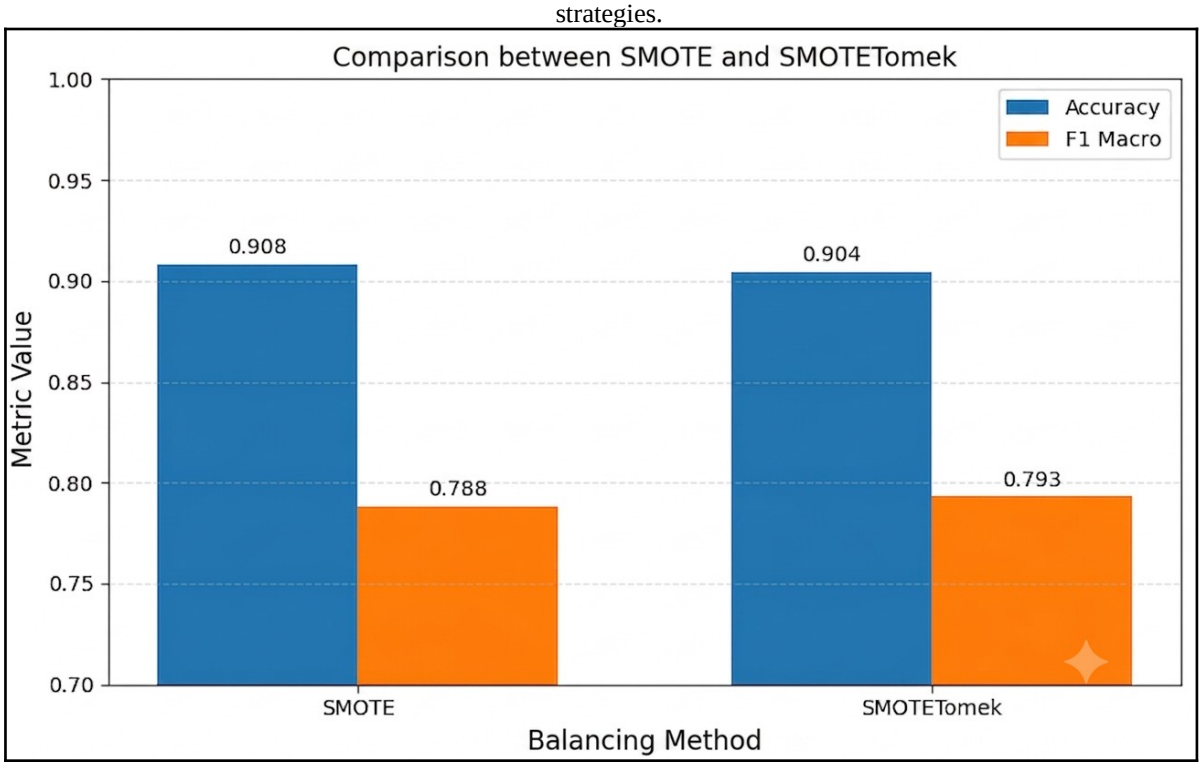
In second level of importance appears Phosphorus (P), with a score of approximately 21, followed by Copper (Cu), with 8.69, and hydrogen potential (pH), with 3.90. Although with high importance, these attributes present lower discriminative power than that observed for Nitrogen, indicating that they contribute to soil fertility classification, but with relatively less individual influence for class definition. The other variables such as Sulfur (S), Iron (Fe), Potassium (K), Manganese (Mn), Boron (B), and Zinc (Zn) present lower scores, but still with sufficient values to be classified among the most relevant by the method. Thus, it is observed that, although they exert less influence, they provide important complementary information for the model. Furthermore, variables such as Organic Carbon (OC) and Electrical Conductivity (EC), although not appearing among the scoreable variables, are found at the end of the list, generally meeting the established criteria for selection. In this work, none of the features were discarded, only classified by the f_classif function from the Scikit-learn library.

After processing, synthetic data were created to complement the database. Authors such as Nasaruddin *et al*. (2025) employed similar techniques for tabular data preprocessing, creating complementary synthetic data and using them for training algorithms such as

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Random Forest. The method used to identify pairs of very close samples in border regions between classes applied the SMOTETomek technique (combination of SMOTE with Tomek Links), as shown in Figure 4, to remove possible synthetic samples that could generate noise or ambiguity in class interpretation. This procedure removed samples considered "noise," representing about 2.7% of the samples generated during balancing. These results suggest that the removal of data in border regions contributed to a more consistent representation of the classes, without significantly compromising the predictive capacity of the trained model. After refining the dataset following the steps described above, the Feature Selection and Correlation Analysis was carried out.

Figure 4 - Comparison of Random Forest model performance using SMOTE and SMOTETomek balancing strategies.



Source: Research data (2026).

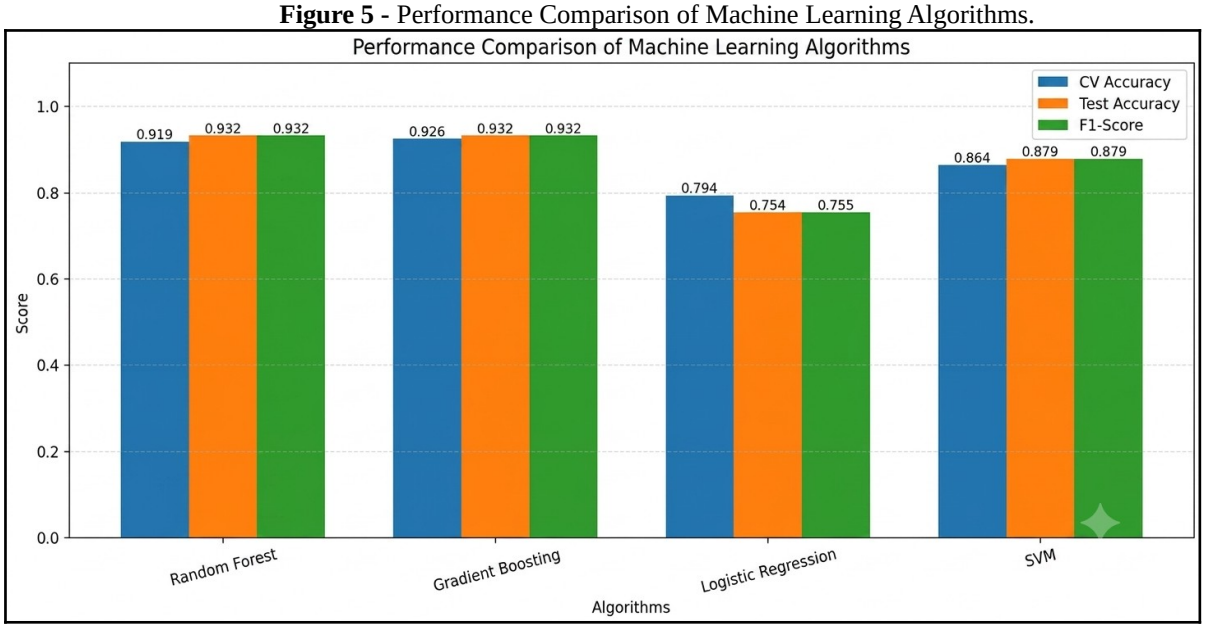
After class balancing and verification, it is observed that SMOTETomek promoted a small increase in the Macro F1-Score (from 0.788 to 0.793), indicating better balance in

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

classification among classes, although with a slight reduction in overall accuracy (from 0.908 to 0.905). This was performed in order to maintain homogeneity among classes, since synthetic data were integrated during preprocessing.

After processing, four algorithms were trained and subsequently hyperparameter optimization was performed using Grid Search. Furthermore, model comparison was also performed using Accuracy, Precision, Recall, F1-Score (Sokolova & Lapalme, 2009), and ROC Curve (PEDREGOSA *et al.*, 2011), selecting the best-performing algorithm for integration into the system. In this study, the best algorithm was Random Forest, standing out with a small percentage over the others, as shown in Figure 5. The comparison among different machine learning algorithms for soil fertility classification follows an approach similar to that used by Benedet *et al*. (2021), Abdullah (2025), and Mahesh & Soundrapandiyan (2026), who evaluated supervised models using performance metrics such as Accuracy, Precision, Recall, and F1-score. Furthermore, the authors also achieved satisfactory levels for the Random Forest model, similar to the present study.

Figure 5 - Performance Comparison of Machine Learning Algorithms.



Source: Research data (2026).

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

For this, the StratifiedKFold(n_splits=5) code was used, and each algorithm was trained five times to ensure good selection results, thus reducing the influence of data splitting and providing a more accurate estimate of model performance. According to the results obtained, the best performance was achieved by Random Forest, with values very close to Gradient Boosting, both with accuracy above 90%. The proximity between the accuracy obtained in cross-validation and on the test dataset indicates excellent generalization capacity and absence of relevant evidence of overfitting. In contrast, the slightly inferior performance of Logistic Regression suggests a slight nonlinearity between classes and attributes.

Furthermore, the optimization performed by GridSearchCV identified as the best configuration for Random Forest a total of 200 trees, maximum depth of 15, min_samples_split = 2, and min_samples_leaf = 1, resulting in an accuracy of 93.56%. After selecting the model with the highest performance during the training and optimization process, its evaluation was performed using the test set, previously separated and not used in training. Performance was also measured using accuracy, precision, recall, and F1-score metrics, in order to represent the overall classification performance of the classes.

Additionally, the classification report was generated, containing individual metrics for each class, and the confusion matrix was constructed, presented in Table 3, both with absolute values and in normalized form. The confusion matrix (SOKOLOVA & LAPALME, 2009) enabled analysis of the distribution of hits and classification errors, identifying which classes have higher or lower confusion rates among themselves. As the problem addressed in this study consists of a multiclass classification, the ROC curve (receiver operating characteristic curve) and the AUC-ROC metric (area under the receiver operating characteristic curve) (FAWCETT, 2006) were not used in the final evaluation, being calculated only in binary classification situations.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Table 3 - Accuracy, Precision, Recall, F-score and Report

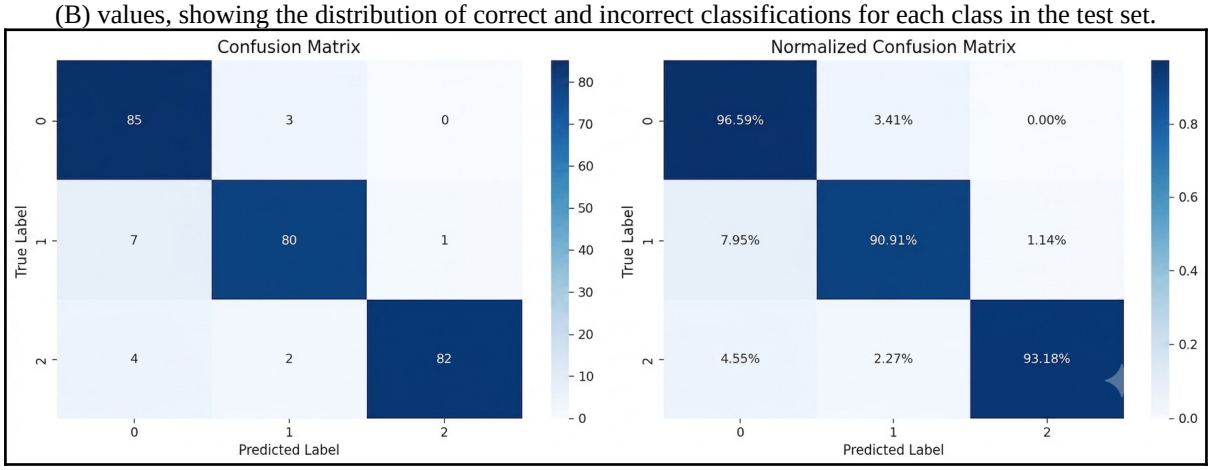
Class	*Precision*	*Recall*	*F1-score*	*Support*
0	0.89	0.97	0.92	88
1	0.94	0.91	0.92	88
2	0.99	0.93	0.96	88
Accuracy			0.94	264
Macro avg	0.94	0.94	0.94	264
weighted avg	0.94	0.94	0.94	264

Source: Research data (2026).

The optimization of the Random Forest model algorithm presented satisfactory performance in soil fertility classification, achieving accuracy of 93.56%, weighted precision of 93.82%, recall of 93.56%, and F1-score of 93.59%, evidencing its generalization capacity on data not used during training. Individual class analysis revealed consistent behavior among classes. Class 0 presented precision of 0.89, recall of 0.97, and F1-score of 0.92, indicating high capacity for identifying this category, although with a small occurrence of false positives. Class 1 presented a precision of 0.94, recall of 0.91, and F1-score of 0.92, while Class 2 obtained the best performance among all categories, with precision of 0.99, recall of 0.93, and F1-score of 0.96. In Figure 6, the confusion matrices are presented in absolute and normalized values, allowing visualization of the model's correct classifications and errors.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Figure 6 - Confusion matrix of the optimized Random Forest model, presented in absolute (A) and normalized (B) values, showing the distribution of correct and incorrect classifications for each class in the test set.



Source: Research data (2026).

The confusion matrix shows that most samples were correctly classified, with the highest values concentrated on the main diagonal. Of the 88 samples belonging to Class 0, 85 were correctly classified, resulting in a hit rate of 96.59%, while only three samples were misclassified, belonging to Class 1. For Class 1, it was observed that 80 of the 88 samples were correctly classified, with a hit rate of 90.91%, with seven classified as Class 0 and one as Class 2. Class 2 presented 82 correct classifications, with 93.18% accuracy, with only four samples being classified as Class 0 and two as Class 1. Overall, it appears that errors occurred mainly between Classes 0 and 1, indicating greater similarity between these categories. In contrast, Class 2 showed heterogeneity in relation to the other two classes, which is evidenced by the precision of 0.99 and the reduced number of classification errors.

The optimized Random Forest model was exported and integrated into the developed application, allowing new samples to be processed automatically. During application execution, user-input data undergo the same preprocessing steps used in training, including variable selection and attribute standardization, and are subsequently submitted to the trained model for soil fertility classification generation.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

3.1 Model implementation into the interface

The optimized Random Forest model was exported and prepared for integration into the Tellurium platform, preserving all stages of the processing pipeline employed during training. Functional tests performed in a development environment confirmed the correct loading of the trained model, the variable standardization object, and the selected attributes, enabling predictions and calculation of classification probabilities. Authors such as Chandra *et al*. (2023) also obtained similar results for the application of a predictive algorithm in an interactive interface using the Random Forest trained model.

Furthermore, the function responsible for integrating the model into the application was developed, automating the loading of pipeline components, preprocessing of user-input data, and inference execution. However, due to an update of one of the libraries used in the interface development, the final integration of the model into the application is currently in an adaptation phase. At present, this incompatibility prevents the result generated by the algorithm from being correctly returned and displayed to the user through the graphical interface, although the model remains functional and validated in the development environment.

Thus, the results obtained demonstrate the feasibility of incorporating the machine learning model into the Tellurium platform, with only the adaptation of the interface integration layer remaining to enable its full operation in the production environment.

Although the final integration still depends on interface adaptation, the results obtained demonstrate the potential of incorporating the machine learning model into the Tellurium platform. Its implementation was not intended to replace the pre-existing agronomic recommendation mechanism, but to act as a complementary layer of intelligence and validation of the recommendations generated by the application. While the platform continues to base its recommendations on established technical and agronomic criteria, the Random Forest model, trained and validated with high performance (accuracy of 93.56% and F1-score of 93.59%), will provide a second evaluation based on patterns learned from the

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

data. This hybrid approach allows comparing the recommendations produced by the system with those generated by the machine learning model, increasing the reliability of the decision-making process and reducing the possibility of inconsistencies. Thus, the integration between agronomic knowledge and artificial intelligence will expand the platform's potential as a decision-support tool, adding an additional verification mechanism that strengthens the robustness, security, and reliability of the recommendations provided to users.

3.2 Limitations and future work

As a limitation of the study, it is highlighted that the complete integration of the model into the interface is still in an adaptation phase due to the update of Python® libraries used in the application development, which caused incompatibilities between the trained model and the interface. This limitation is restricted to the application integration layer and does not compromise the model's performance, which was validated in a development environment. As future work, it is intended to complete this integration, expand the database used in training, and carry out continuous evaluation of new algorithms and optimization strategies, aiming to increase the robustness, generalization capacity, and reliability of the interface. It is noted that this research remains under development, enabling the incorporation of new functionalities and improvements to the Tellurium interface.

4. CONCLUSION

The present study aimed to develop and evaluate a machine learning model integrated into the Tellurium interface to assist small, medium, and large farmers in soil fertility classification and in the recommendation of fertilization management, liming, and soil conservation practices. For this purpose, a pipeline was structured involving data preprocessing, feature selection, algorithm training and optimization, culminating in the integration of the best-performing model into the interface.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

The incorporation of the machine learning model aims to expand the functionalities of the Tellurium platform without replacing its original recommendation mechanism. Thus, the model will act as an additional intelligence layer, complementing the agronomic recommendations already implemented and increasing the platform's potential as a decision-support tool. Among the evaluated algorithms, Random Forest presented the best performance, achieving accuracy of 93.56%, precision of 93.82%, recall of 93.56%, and F1-score of 93.59%, demonstrating high predictive capacity. The confusion matrix evidenced a high rate of correct classifications among the classes evaluated in this study, indicating good generalization capacity of the model for data not used during training.

In conclusion, the results demonstrate that the incorporation of machine learning techniques can expand the potential of intelligent systems aimed at agricultural management. The integration of the model into the interface shows how this technology, when properly developed and applied, can become an accessible tool for farmers of different profiles, contributing to decision-making, management planning, and more efficient use of resources in production systems.

5. ACKNOWLEDGMENTS

The authors acknowledge the Ministry of Science, Technology and Innovation (MCTI), the Brazilian Innovation Agency (FINEP), the Goiás State Research Foundation (FAPEG), the National Council for Scientific and Technological Development (CNPq), the Coordination for the Improvement of Higher Education Personnel (CAPES), the Goiás State Secretariat for Development and Innovation (SECTI), the Center of Excellence in Exponential Agriculture (CEAGRE), the Center of Excellence in Sustainable Irrigation (CEIS), and the Federal Institute of Goiás (IF Goiano) for their support of the research conducted.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

REFERENCES

ABDULLAH, W. A comparative study of machine learning models for soil fertility prediction based on soil properties. **Optimization in Agriculture**, v. 2, p. 1-9, 2025. Disponível em: https://doi.org/10.61356/j.oia.2025.2477. Acesso em: 12 abr. 2026.

ABREU, A. M.; SÁTIRO, G.; LITRE, G.; SANTOS, L.; OLIVEIRA, J. E.; SOARES, D.; ÁVILA, K. A interface entre saúde, mudanças climáticas e uso do solo no Brasil: uma análise da evolução da produção científica internacional entre 1990 e 2019. **Saúde e Sociedade**, v. 29, n. 2, 2020. Disponível em: https://doi.org/10.1590/S0104-12902020180866. Acesso em: 25 abr. 2026.

ADOMAKO, M. O.; ROILOA, S.; YU, F. H. Potential roles of soil microorganisms in regulating the effect of soil nutrient heterogeneity on plant performance. **Microorganisms**, v. 10, n. 12, p. 2399, 2022. Disponível em: https://doi.org/10.3390/microorganisms10122399. Acesso em: 08 maio 2026.

BABU, K. D.; GOUSIA, S. U.; DEEP, N.; CP, M. M.; GANGWAR, S.; BEHERA, S. K.; HIMSHIKHA; DUTTA, S. Sensor based soil monitoring: a new era in precision agronomic decision making. **International Journal of Research in Agronomy**, v. 8, n. 6, p. 743-754, 2025. Disponível em: https://doi.org/10.33545/2618060X.2025.v8.i6i.3113. Acesso em: 19 maio 2026.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD Explorations Newsletter**, v. 6, n. 1, p. 20-29, 2004.

BENEDET, L.; ACUÑA-GUZMAN, S. F.; FARIA, W. M.; SILVA, S. H. G.; CURI, N.; GUILHERME, L. R. G. Rapid soil fertility prediction using X-ray fluorescence data and machine learning algorithms. **CATENA**, v. 197, p. 105003, 2021. Disponível em: https://doi.org/10.1016/j.catena.2020.105003. Acesso em: 04 jun. 2026.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001. Disponível em: https://doi.org/10.1023/A:1010933404324. Acesso em: 16 jun. 2026.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

CARBAJAL-LLOSA, C.; BARJA, A.; PIZARRO, S. Ensemble machine learning for digital mapping of soil pH and electrical conductivity in the Andean agroecosystem of Peru. **Frontiers in Soil Science**, v. 5, p. 1673628, 2025. Disponível em: https://doi.org/10.3389/fsoil.2025.1673628. Acesso em: 27 jun. 2026.

CASTRO, C. N. Agricultura familiar e maquinário agrícola no Brasil: barreiras ao uso, pesquisa e desenvolvimento e políticas públicas. **Planejamento e Políticas Públicas**, n. 70, 2025. Disponível em: https://doi.org/10.38116/ppp70art10. Acesso em: 11 abr. 2026.

CASINI, S.; DUCANGE, P.; MARCELLONI, F.; POLLINI, L. Artificial intelligence in agri-robotics: a systematic review of trends and emerging directions leveraging bibliometric tools. **Robotics**, v. 15, n. 1, p. 24, 2026. Disponível em: https://doi.org/10.3390/robotics15010024. Acesso em: 30 abr. 2026.

CHANDRA, H.; PAWAR, P. M.; ELAKKIYA, R.; TAMIZHARASAN, P. S.; MUTHALAGU, R.; PANTHAKKAN, A. Explainable AI for soil fertility prediction. **IEEE Access**, v. 11, p. 97866-97878, 2023. Disponível em: https://doi.org/10.1109/ACCESS.2023.3311827. Acesso em: 14 maio 2026.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321-357, 2002. Disponível em: https://doi.org/10.1613/jair.953. Acesso em: 28 maio 2026.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273-297, 1995.

DESCONSI, C.; DE SÁ, M. O uso e a apropriação de tecnologias digitais na gestão de unidades agropecuárias: uma análise a partir da percepção de agricultores de Santa Catarina - Brasil. **Desenvolvimento em Questão**, v. 22, n. 60, 2024. Disponível em: https://doi.org/10.21527/2237-6453.2024.60.15011. Acesso em: 10 jun. 2026.

FACELI, K. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. 2. ed. Rio de Janeiro: LTC, 2021.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861-874, 2006. Disponível em: https://doi.org/10.1016/j.patrec.2005.10.010. Acesso em: 23 jun. 2026.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189-1232, 2001. Disponível em: https://doi.org/10.1214/aos/1013203451. Acesso em: 06 abr. 2026.

GITHUB. GitHub Docs. Disponível em: https://docs.github.com/. Acesso em: 21 abr. 2026.
GOOGLE. Google Colaboratory. 2024. Disponível em: https://colab.research.google.com/. Acesso em: 17 maio 2026.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. 3. ed. Hoboken: Wiley, 2013.

ISLAM, A.; HOSSAIN, M. S.; ANDERSEN, P.; RAHMAN, M. A. KNNOR: an oversampling technique for imbalanced datasets. **Applied Soft Computing**, v. 115, p. 108288, 2022. Disponível em: https://doi.org/10.1016/j.asoc.2021.108288. Acesso em: 29 maio 2026.

JOBLIB DEVELOPMENT TEAM. Joblib: Running Python functions as pipeline jobs. Disponível em: https://joblib.readthedocs.io/. Acesso em: 12 jun. 2026.

KAGGLE. Soil Fertility Dataset. 2023. Disponível em: https://www.kaggle.com/datasets/rahuljaiswalonkaggle/soil-fertility-dataset. Acesso em: 24 jun. 2026.

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. **Journal of Machine Learning Research**, v. 18, n. 17, p. 1-5, 2017. Disponível em: http://jmlr.org/papers/v18/16-365.html. Acesso em: 15 abr. 2026.

MAHESH, P.; SOUNDRAPANDIYAN, R. Enzyme-aware soil fertility prediction using dual optimization with improved SCSO. **Scientific Reports**, v. 16, p. 56124, 2026. Disponível em: https://doi.org/10.1038/s41598-026-56124-1. Acesso em: 18 jun. 2026.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

MCKINNEY, W. Data structures for statistical computing in Python. In: **PROCEEDINGS OF THE 9TH PYTHON IN SCIENCE CONFERENCE (SCIPY 2010)**. SciPy, 2010. p. 56-61. Disponível em: https://doi.org/10.25080/Majora-92bf1922-00a. Acesso em: 07 abr. 2026.

MMBANDO, G. S. Harnessing artificial intelligence and remote sensing in climate-smart agriculture: the current strategies needed for enhancing global food security. **Cogent Food & Agriculture**, v. 11, n. 1, 2025. Disponível em: https://doi.org/10.1080/23311932.2025.2454354. Acesso em: 18 abr. 2026.

NASARUDDIN, N. et al. A SMOTE PCA HDBSCAN approach for enhancing water quality classification in imbalanced datasets. **Scientific Reports**, v. 15, p. 13059, 2025. Disponível em: https://doi.org/10.1038/s41598-025-97248-0. Acesso em: 30 abr. 2026.

PEARSON, K. Notes on regression and inheritance in the case of two parents. **Proceedings of the Royal Society of London**, v. 58, p. 240-242, 1895. Disponível em: https://doi.org/10.1098/rspl.1895.0041. Acesso em: 12 maio 2026.

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011. Disponível em: https://jmlr.org/papers/v12/pedregosa11a.html. Acesso em: 21 maio 2026.

PYTHON SOFTWARE FOUNDATION. Python Language Site: Documentation. 2024. Disponível em: https://www.python.org/doc/. Acesso em: 03 jun. 2026.

SARMA, H. H.; BEZBARUAH, A.; TALUKDAR, N.; KURMI, K. K.; BORKOTOKY, B. Comprehensive assessment of various soil fertility evaluation techniques for estimating nutrient status of soil: a review. **Plant Archives**, v. 24, n. 2, p. 1131-1140, 2024. Disponível em: https://doi.org/10.51470/PLANTARCHIVES.2024.v24.no.2.135. Acesso em: 27 jun. 2026.

SITOE, E. S.; SCHENATO, R. B.; SANTOS, D. R. Potassium content in soil and plants in a long-term potassium fertilization experiment. **Ciência Rural**, v. 55, n. 7, e20240133, 2025. Disponível em: https://doi.org/10.1590/0103-8478cr20240133. Acesso em: 10 abr. 2026.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, v. 45, n. 4, p. 427-437, 2009. Disponível em: https://doi.org/10.1016/j.ipm.2009.03.002. Acesso em: 22 maio 2026.

SOARES, M. R. et al. Levantamento do consumo de fertilizantes e utilização da análise de solo por pequenos e médios produtores agrícolas da região de Araras-SP. **Revista Ciência em Extensão**, v. 5, n. 1, p. 56-75, 2009. Disponível em: https://ojs.unesp.br/index.php/revista_proex/article/view/95. Acesso em: 08 jun. 2026.

SZEGHALMY, S.; FAZEKAS, A. A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. **Sensors**, v. 23, n. 4, p. 2333, 2023. Disponível em: https://doi.org/10.3390/s23042333. Acesso em: 19 jun. 2026.

TOMEK, I. Two modifications of CNN. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 6, n. 11, p. 769-772, 1976.

VIANA F., A.; DE OLIVEIRA PAZ, A. M. As dificuldades de agricultores em transações comerciais bancárias e a oficina de letramento como estratégia de superação. **Revista Eletrônica Competências Digitais para Agricultura Familiar (RECoDAF)**, v. 10, n. 1, e1001189, 2024. Disponível em: https://owl.tupa.unesp.br/recodaf/index.php/recodaf/article/view/189. Acesso em: 15 jun. 2026.

WEN, Z.; CHEN, Y.; LIU, Z.; MENG, J. Biochar and arbuscular mycorrhizal fungi stimulate rice root growth strategy and soil nutrient availability. **European Journal of Soil Biology**, v. 113, p. 103448, 2022. Disponível em: https://doi.org/10.1016/j.ejsobi.2022.103448. Acesso em: 25 abr. 2026.

WOLFERT, S.; GE, L.; VERDOUW, C.; BOGAARDT, M.-J. Big data in smart farming – A review. Agricultural Systems, v. 153, p. 69-80, 2017. Disponível em: https://doi.org/10.1016/j.agsy.2017.01.023. Acesso em: 31 maio 2026.

Recebido em: 29/06/2026
Aprovado em: 17/07/2026
Publicado em: 13/08/2026

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)