

Machine learning-based territorial classification of agricultural profiles in Colombia: A random forest approach

Jorge Saul Fandiño Pelayo¹, Ana Maria Porras Ravelo¹, Johan Leandro Tellez Garzón¹, Rocio Cazes Ortega¹, Renato Gonzalez Robles¹

¹ Department of Natural Sciences and Engineering, Technological Units of Santander (UTS), Student Street 9-82, Bucaramanga, 680005, Colombia.

*Corresponding author E-mail: jfandino@correo.uts.edu.co

ABSTRACT

Each year, Colombia generates a vast amount of agricultural data reflecting the evolution of production systems across different regions. Municipal Agricultural Assessments (EVA) annually compile records of production, planted areas, harvested areas, and yields for municipalities throughout the country, making them one of the main public sources of information for the agricultural sector. However, these records have primarily been used for compiling statistics and official reports, while their potential to explain differences in agricultural performance across regions using machine learning techniques has been largely unexplored. To leverage this information from a territorial perspective, this research developed a territorial intelligence framework that transforms EVA database records into interpretable agricultural profiles using derived indicators and machine learning models. To this end, five territorial indicators were constructed: real productivity, harvest efficiency, loss index, crop share, and historical variability. These indicators were derived from agricultural records collected between 2019 and 2024 and allowed for the representation of different dimensions of agricultural performance beyond traditional yield measures. The proposed territorial indicators made it possible to distinguish four agricultural profiles describing different patterns of agricultural performance across Colombian agricultural records. Decision Tree and Random Forest models were evaluated for their ability to recognize the proposed agricultural profiles using municipality-level data separation for model validation. Random Forest achieved a classification accuracy of $76.4\% \pm 0.9\%$, while Loss Index, Historical Variability, and Crop Participation showed the highest variable importance among the predictors. The resulting agricultural profiles describe different patterns of territorial agricultural performance across Colombian municipalities.

Keywords: Machine Learning, Territorial Classification, Agricultural Performance, Spatial Analysis, Random Forest, Agricultural Intelligence.

1. Introduction

The Municipal Agricultural Assessments (EVA) database provides annual information on crop production, cultivated area, harvested area, and other agricultural variables reported by municipalities throughout Colombia [1]. Because these records describe agricultural production under diverse environmental conditions, management practices, and levels of resource availability, they capture substantial differences in agricultural performance across Colombian municipalities. Understanding these territorial differences is important for agricultural planning and for designing public policies adapted to local conditions. Despite the broad territorial coverage of these records, they have been used mainly for descriptive analyses, while their potential for predictive and territorial studies has received much less attention [2].

The availability of large agricultural datasets has opened new opportunities to analyze agricultural systems using machine learning, allowing complex relationships within production data to be identified beyond the capabilities of conventional statistical methods. Recent studies have applied machine learning techniques to crop mapping,

crop classification, production forecasting, agricultural monitoring, and remote sensing, reporting encouraging results under different agricultural conditions [3]–[6]. Likewise, several studies have reported successful applications of machine learning for yield prediction in crops such as maize and rice, as well as comparative analyses of predictive algorithms designed to improve agricultural decision-making [7]–[9]. Other studies have combined artificial intelligence with sensor networks, predictive monitoring, irrigation systems, and automated control strategies to improve resource management and enhance the sustainability of agricultural production systems [10]–[13]. Among the available algorithms, Random Forest has consistently demonstrated good predictive performance when agricultural datasets contain nonlinear relationships and heterogeneous variables [14], [15].

Despite these advances, much of the existing literature focuses on estimating crop yields, predicting production, or monitoring crop conditions. Based on the review of these studies, only a limited number have explored how machine learning can be used to characterize agricultural performance from a territorial perspective by identifying agricultural records associated with similar productive conditions across municipalities. Even with these advances, distinguishing territories with stable agricultural performance from those affected by production losses, low harvest efficiency, or high temporal variability remains difficult [16], [17].

Agricultural records contain information that extends beyond production values alone. When these records are transformed into derived indicators, they can represent complementary dimensions of territorial agricultural performance, including productive efficiency, specialization, and production stability. Feature engineering, data preparation, and data quality are recurrent topics in agricultural machine learning studies [18], [19]. Agricultural information has also been combined with analytical techniques in studies related to precision agriculture and Internet of Things (IoT) systems for agricultural management [20].

This research proposes a territorial classification approach for Colombian agricultural records using information from the Municipal Agricultural Assessments (EVA). The proposed methodology is based on five territorial indicators derived from the original agricultural records: Real Productivity, Harvest Efficiency, Loss Index, Crop Participation, and Historical Variability. Together, these indicators describe complementary dimensions of agricultural performance and define four agricultural profiles: Efficient, Stable, Productive, and Vulnerable. Decision Tree and Random Forest models were subsequently evaluated to determine their ability to classify these profiles and to identify the variables that contribute most to agricultural differentiation. Municipality-level data separation was used for model validation to prevent records from the same municipality from being included in both the training and testing sets.

2. Methodology

Agricultural production records were obtained from the Municipal Agricultural Assessments (EVA), an open database maintained by the Colombian Ministry of Agriculture [1]. The records correspond to the period from 2019 to 2024 and cover municipalities throughout Colombia.

Before the analysis, numerical formats were corrected, duplicate observations were removed, missing values in key variables were addressed, records with invalid productivity values were excluded, and crop names were standardized. These procedures resulted in a final dataset containing 125,227 agricultural records from 1,020 municipalities representing 163 crops.

Five territorial indicators were calculated from the processed agricultural records: Real Productivity, Harvest Efficiency, Loss Index, Crop Participation, and Historical Variability. Together, these indicators describe complementary dimensions of agricultural performance and provided the basis for defining territorial agricultural profiles. The methodology comprised four sequential stages: dataset preprocessing and validation, calculation of the territorial indicators, definition of agricultural profiles according to productivity and harvest efficiency criteria, and training and evaluation of Decision Tree and Random Forest models for territorial agricultural profile classification.

2.1. Feature engineering

Five territorial indicators were calculated from the original EVA database to represent different dimensions of agricultural performance beyond conventional production variables [21].

Real Productivity represents the amount of agricultural production obtained per planted hectare:

$$\text{Real Productivity} = \frac{\text{Production}}{\text{Planted Area}} \quad (1)$$

Harvest Efficiency represents the proportion of planted area that was successfully harvested:

$$\text{Harvest Efficiency} = \frac{\text{Harvested Area (hectares)}}{\text{Planted Area (hectares)}} \quad (2)$$

Loss Index measures the proportion of planted area that was not harvested:

$$\text{Loss Index} = \frac{\text{Planted Area} - \text{Harvested Area}}{\text{Planted Area}} \quad (3)$$

Loss Index measures the proportion of planted area that was not harvested. Before applying the lower-bound truncation used in the preprocessing, this indicator is mathematically complementary to Harvest Efficiency, such that $\text{Loss Index} = 1 - \text{Harvest Efficiency}$.

Crop Participation represents the contribution of each crop to the total agricultural production within a municipality:

$$\text{Crop Participation} = \frac{\text{Crop Production}}{\text{Total Municipal Production}} \quad (4)$$

Historical Variability was calculated as the standard deviation of Real Productivity over time for each municipality–crop combination:

$$\text{Historical Variability} = \text{Std}(\text{Real Productivity}) \quad (5)$$

These indicators represent productivity, harvesting efficiency, production losses, crop specialization, and temporal variability.

2.2. Agricultural profile construction

Agricultural profiles were defined according to the median values of Real Productivity and Harvest Efficiency, which were used as classification thresholds. The median was selected as the threshold because it divides the observed records into two groups according to their relative position within the empirical distribution, while reducing the influence of extreme values that may occur in agricultural production data. Similar productivity-based classification approaches have been reported for agricultural risk assessment and crop productivity categorization using machine learning techniques [22].

Each agricultural record, representing an observation associated with a municipality, crop, and year, was assigned to one of four agricultural profiles. The Efficient profile corresponds to records with high productivity and high harvest efficiency. The Productive profile includes records with high productivity and low harvest efficiency. The Stable profile represents records with low productivity and high harvest efficiency, whereas the Vulnerable profile includes records with both low productivity and low harvest efficiency.

These four categories formed the target variable used to train the supervised classification models and represent different patterns of agricultural performance at the record level. The resulting profiles were subsequently used for territorial characterization by aggregating the assigned profiles at the municipality level. For this territorial representation, the modal agricultural profile was calculated for each municipality. This procedure allowed the spatial distribution of the agricultural profiles to be represented while preserving the original agricultural records as the unit of classification.

2.3. Model training and evaluation

The predictive variables included Loss Index, Crop Participation, Dominant Crop, and Historical Variability, while the response variable corresponded to the Agricultural Profile. Real Productivity and Harvest Efficiency were used to define the agricultural profiles and therefore were not included as predictive variables in the classification models. These variables are consistent with previous studies on agricultural segmentation and territorial characterization [23]. Decision Tree and Random Forest classifiers were implemented in Python using the Scikit-learn library [24].

To evaluate model generalization while avoiding the presence of records from the same municipality in both training and validation data, a five fold Stratified Group K Fold cross validation strategy was applied using municipality as the grouping variable. This approach preserved the distribution of the agricultural profiles as much as possible while ensuring that each municipality was assigned exclusively to either the training or validation partition within each fold. A fixed random seed (`random_state = 42`) was used to ensure reproducibility. Model performance was evaluated using Accuracy, Precision, Recall, F1 score, and Confusion Matrices to assess multiclass classification performance [25].

Random Forest variable importance was estimated for each predictor to identify the variables that contributed most to the classification of agricultural profiles. The performance of the Decision Tree and Random Forest models was compared across the five validation folds using the mean and standard deviation of the evaluation metrics. This procedure provided a municipality independent assessment of model performance and reduced the possibility that observations from the same municipality influenced both model training and validation.

3. Results

3.1. Data preparation and territorial coverage

The EVA database contained 141,073 agricultural records collected from Colombian municipalities between 2019 and 2024. After data cleaning, which removed duplicated observations, records with missing values in key variables, and records with invalid productivity values, the final dataset included 125,227 agricultural records. Crop names were also standardized during the preprocessing stage to ensure consistency across the database. The resulting dataset retained 88.8% of the original agricultural records for the subsequent analysis.

Table 1. Preprocessing steps applied to the EVA database

Data Processing Stage	Records	Percentage (%)
Original EVA records	141,073	100.0
Excluded records	15,846	11.2
Records retained for analysis	125,227	88.8

The final dataset included records from 1,020 municipalities and 163 agricultural crops across Colombia. These records provided the basis for calculating the five territorial indicators and defining the four agricultural profiles described in the subsequent analysis. The geographic coverage allowed agricultural records from municipalities across the country to be considered in the classification process. The resulting database was subsequently used for the construction of agricultural profiles and the evaluation of the machine learning models.

3.2. Descriptive analysis of productive and territorial indicators

Five indicators were constructed from the EVA database: Real Productivity, Harvest Efficiency, Loss Index, Crop Participation, and Historical Variability. Table 2 presents the descriptive statistics of the territorial indicators, including the mean, standard deviation, minimum, median, and maximum values. Real Productivity showed a mean of 9.20, a standard deviation of 11.30, and a median of 5.73, with values ranging from 0.00 to 126.00. These results provide a descriptive overview of the productive and territorial characteristics represented in the dataset.

Table 2. Descriptive statistics of the territorial indicators

Indicator	Mean	Std	Min	Median	Max
Real Productivity	9.20	11.30	0.00	5.73	126.00
Harvest Efficiency	0.91	0.21	0.00	1.00	2.00
Loss Index	0.10	0.18	0.00	0.00	1.00
Crop Participation	0.05	0.12	0.00	0.01	1.00
Historical Variability	2.88	3.90	0.00	1.60	45.15

Harvest Efficiency presented a mean of 0.91, a standard deviation of 0.21, and a median of 1.00, with values ranging from 0.00 to 2.00. Loss Index showed a mean of 0.10, a standard deviation of 0.18, and a median of 0.00, with a maximum value of 1.00. Crop Participation presented a mean of 0.05 and a median of 0.01, while Historical Variability showed a mean of 2.88 and a median of 1.60, with a maximum value of 45.15. The final dataset included 1,020 municipalities and 163 agricultural crops across Colombia.

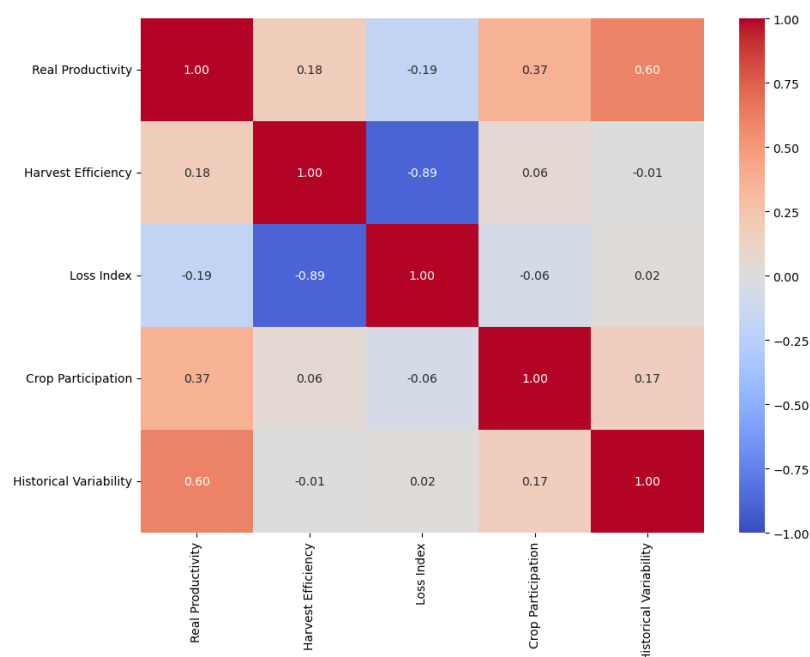


Figure 1. Correlation matrix of the territorial indicators

The correlation matrix showed different relationships among the territorial indicators. The strongest correlation was observed between Harvest Efficiency and Loss Index ($r = -0.89$), reflecting their complementary relationship in the measurement of harvested and non harvested planted area. Real Productivity showed a moderate positive correlation with Historical Variability ($r = 0.60$). The remaining indicators showed relatively weak correlations, with values close to zero in most cases.

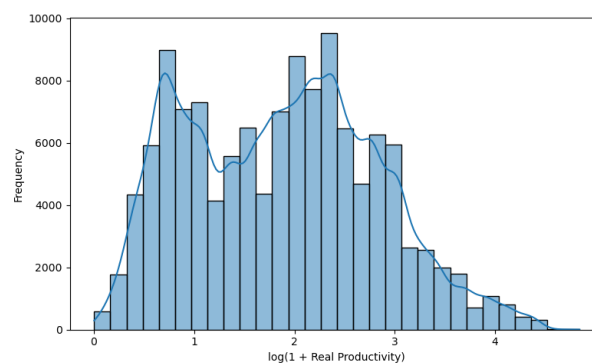


Figure 2. Distribution of Real Productivity on a logarithmic scale

The distribution of $\log(1 + \text{Real Productivity})$ was concentrated at low and intermediate productivity values, whereas fewer observations were found at higher productivity levels. Several local peaks were also observed in the distribution. The logarithmic transformation reduced the influence of extreme values while preserving the variability of the data.

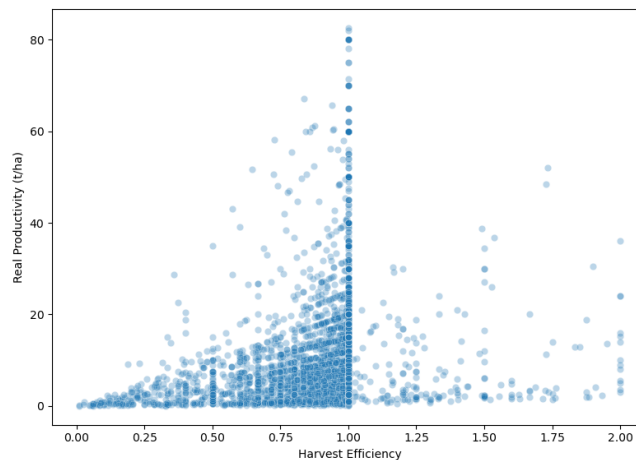


Figure 3. Scatter plot of Harvest Efficiency versus Real Productivity

Figure 3 shows the relationship between Harvest Efficiency and Real Productivity. For municipalities with Harvest Efficiency values close to 1.0, Real Productivity ranges from very low values to more than 80 t/ha. Similar levels of Harvest Efficiency are associated with different productivity values across the dataset.

3.3. Territorial and crop characterization

Table 3 presents the Colombian departments with the highest values of Real Productivity and Harvest Efficiency.

Table 3. Departments with the highest Real Productivity and Harvest Efficiency

Department	Real Productivity	Harvest Efficiency
Valle del Cauca	12.71	0.93
Risaralda	11.06	—
Quindío	10.95	—
Caldas	10.73	—
Cundinamarca	10.65	—
San Andrés and Providencia	—	0.94
Norte de Santander	—	0.94
Tolima	—	0.93
Magdalena	—	0.92

Valle del Cauca recorded the highest Real Productivity (12.71 t/ha), followed by Risaralda, Quindío, Caldas, and Cundinamarca, all with values above 10 t/ha. The highest Harvest Efficiency was recorded in San Andrés and Providencia and Norte de Santander (0.94), followed by Valle del Cauca and Tolima (0.93) and Magdalena (0.92). Real Productivity and Harvest Efficiency were not ranked identically across departments. Table 4 summarizes the crops with the highest Real Productivity and the lowest Historical Variability.

Table 4. Crops with the highest Real Productivity and the lowest Historical Variability

Crop	Real Productivity	Historical Variability
Sugar cane	40.92	—
Tomato	28.71	—
Carrot	21.97	—

Crop	Real Productivity	Historical Variability
Pineapple	21.88	—
Cabbage	20.92	—
Cocoa	—	0.13
Coffee	—	0.21
Bean	—	0.51
Pea	—	0.60
Oil palm	—	0.70

Sugar cane recorded the highest Real Productivity (40.92 t/ha), followed by tomato (28.71 t/ha), carrot (21.97 t/ha), pineapple (21.88 t/ha), and cabbage (20.92 t/ha). Cocoa recorded the lowest Historical Variability (0.13), followed by coffee (0.21), bean (0.51), pea (0.60), and oil palm (0.70). The crops with the highest Real Productivity did not coincide with those showing the lowest Historical Variability.

3.4. Agricultural profile construction and spatial distribution

Table 5 presents the distribution of agricultural records across the four agricultural profiles.

Table 5. Agricultural profile distribution

Agricultural Profile	Records
Efficient	37,600
Vulnerable	33,532
Stable	29,073
Productive	25,022

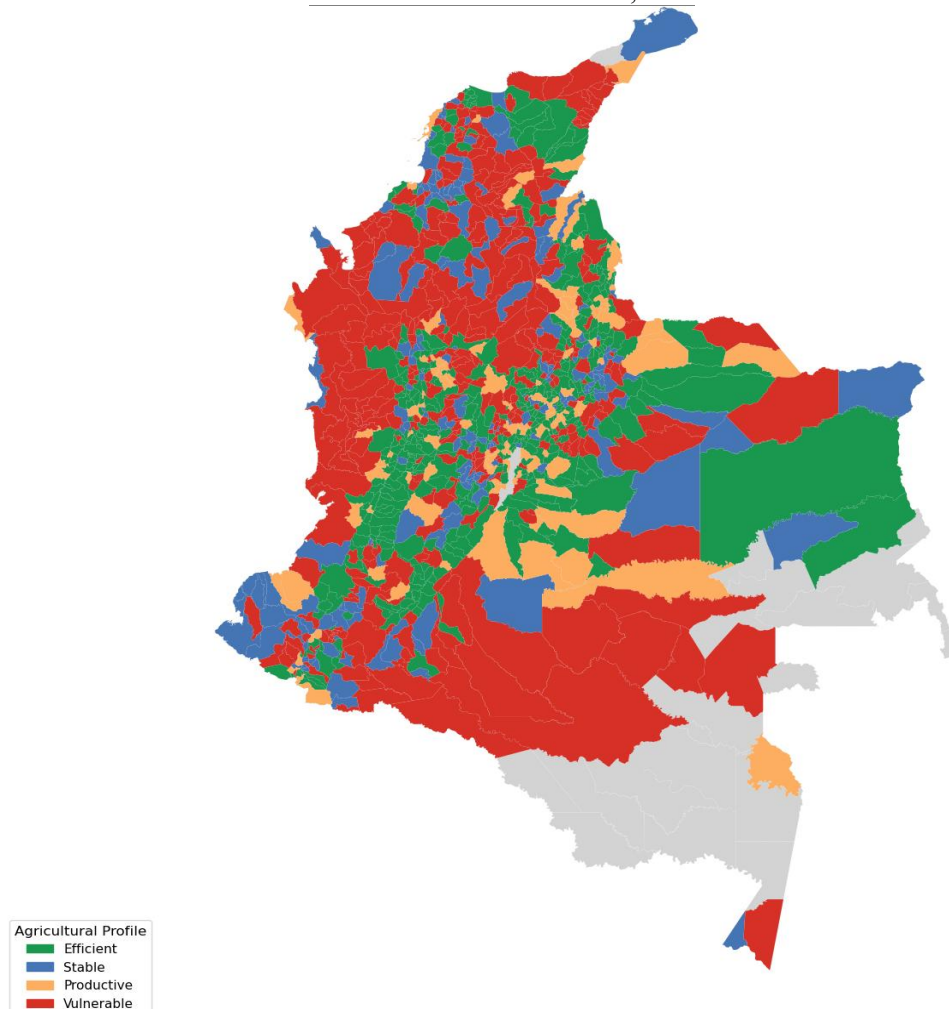


Figure 4. Agricultural profiles across Colombian municipalities

The Efficient profile contained the largest number of records (37,600), representing approximately 30% of the dataset. The Vulnerable profile included 33,532 records, followed by the Stable profile with 29,073 records and the Productive profile with 25,022 records. The four agricultural profiles showed a similar distribution across the dataset.

Figure 4 presents the spatial distribution of the four agricultural profiles across Colombian municipalities. The Vulnerable profile was widely distributed across the country, while Efficient profiles were also observed in different regions. Stable and Productive profiles appeared in smaller groups and were distributed among municipalities with different agricultural characteristics. Different agricultural profiles were also observed in neighboring municipalities, particularly in the Andean region, indicating spatial heterogeneity in agricultural performance.

3.5 Classification model performance

Table 6 compares the performance of the decision tree and random forest models using Accuracy, Precision, Recall, and F1-score.

Table 6. Classification performance of decision tree and random forest

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.777 ± 0.015	0.780 ± 0.014	0.777 ± 0.015	0.776 ± 0.015
Random Forest	0.764 ± 0.009	0.764 ± 0.010	0.764 ± 0.009	0.764 ± 0.009

The Decision Tree achieved an Accuracy of 0.777 ± 0.015 , Precision of 0.780 ± 0.014 , Recall of 0.777 ± 0.015 , and F1-score of 0.776 ± 0.015 , whereas the Random Forest obtained values of 0.764 ± 0.009 for Accuracy, Precision, Recall, and F1-score. These results show relatively consistent classification performance across the five validation folds for both models. Figure 5 presents the confusion matrix obtained with the Decision Tree model.

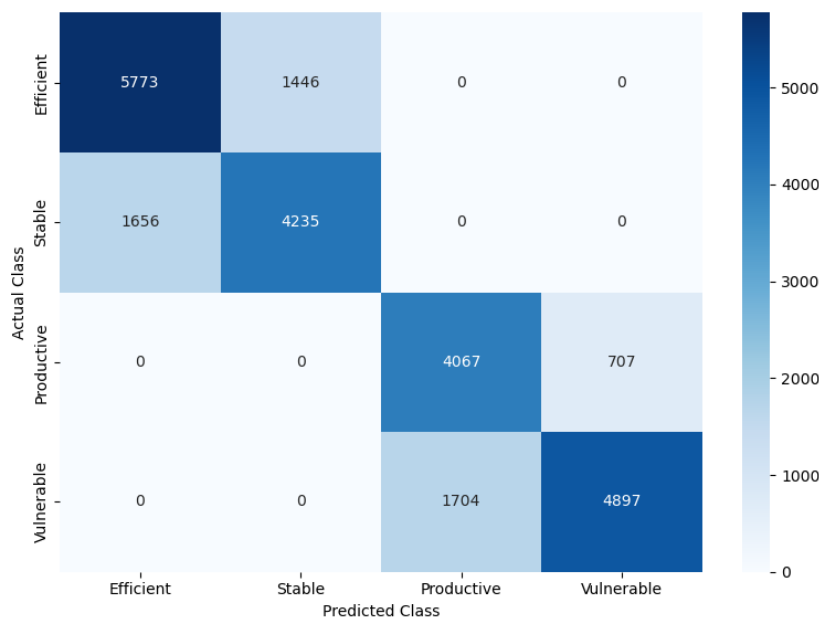


Figure 5. Confusion matrix for the decision tree model

The Decision Tree correctly classified most agricultural records. Most classification errors were observed between the Efficient and Stable profiles and between the Productive and Vulnerable profiles. Misclassification between profiles from different productivity groups was limited. Figure 6 presents the confusion matrix obtained with the Random Forest model.

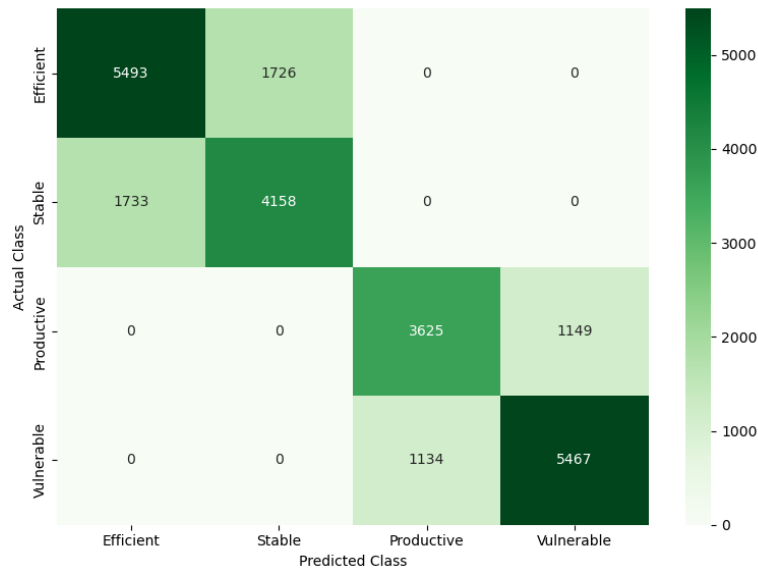


Figure 6. Confusion matrix for the random forest model

The Random Forest model showed a similar classification pattern to the Decision Tree model, with most agricultural records assigned to the corresponding profile. Most classification errors occurred between the Efficient and Stable profiles and between the Productive and Vulnerable profiles. Misclassification between profiles from different productivity groups was not observed. Figure 7 presents the Random Forest variable importance for the predictors used during model training.

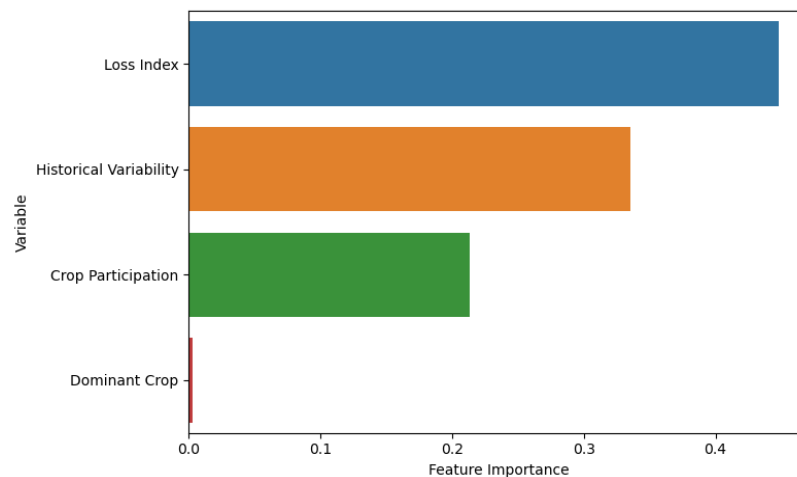


Figure 7. Variable importance estimated by the random forest model

Loss Index showed the highest variable importance (44.8%), followed by Historical Variability (33.6%) and Crop Participation (21.3%). Dominant Crop showed the lowest importance (0.3%). These results indicate that Loss Index, Historical Variability, and Crop Participation contributed most to the classification of agricultural profiles, whereas Dominant Crop had a limited contribution. Figure 7 presents the variable importance of the predictors used in the Random Forest model.

4. Discussion of the findings

Agricultural databases contain large amounts of information, although they continue to be used mainly for descriptive purposes. In this study, agricultural production records were transformed into territorial indicators that describe different dimensions of agricultural performance. This representation made it possible to characterize agricultural records associated with different municipalities, crops, and years using information

beyond cultivated area or production volume. Recent studies have also pointed out that digital agriculture is increasingly supported by machine learning, big data, and analytical tools capable of extracting useful information from large agricultural databases [25].

The combination of these indicators describes complementary aspects of agricultural performance beyond productivity alone. Real Productivity and Harvest Efficiency were used to define the agricultural profiles, while Loss Index, Crop Participation, and Historical Variability provided additional information for the classification models. Similar approaches have been reported in studies on agricultural performance evaluation, productive segmentation, and territorial assessment, where multiple indicators are used to characterize agricultural systems [21], [23]. The use of multiple indicators therefore provides a broader representation of agricultural conditions than the use of production or yield alone.

The spatial representation showed that neighboring municipalities may belong to different agricultural profiles. Municipal-level visualization provides information beyond broader regional summaries and allows differences between nearby territories to be observed. The map also showed that Efficient, Stable, Productive, and Vulnerable profiles were distributed across different parts of the country rather than being restricted to a single region. Similar spatial patterns have been described in studies that applied machine learning to territorial agricultural assessment and spatial decision support [23], [26].

The five-fold stratified group validation showed consistent classification performance for both models when municipalities were kept separated across the validation folds. The Decision Tree achieved an Accuracy of 0.777 ± 0.015 , while the Random Forest obtained an Accuracy of 0.764 ± 0.009 , with similar values for Precision, Recall, and F1-score. These results indicate that both models were able to recognize the proposed agricultural profiles under municipality-level separation, although their performance was slightly different across the validation folds. Similar applications of machine learning in agriculture have reported the usefulness of classification models for datasets characterized by nonlinear relationships and heterogeneous variables [14], [22].

Loss Index showed the highest variable importance (44.8%), followed by Historical Variability (33.6%) and Crop Participation (21.3%), whereas Dominant Crop showed the lowest importance (0.3%). The importance of Loss Index should be interpreted with caution because this indicator is mathematically related to Harvest Efficiency, which was used to construct the agricultural profiles. This relationship may contribute to the strong contribution of Loss Index to the classification and represents an important consideration when interpreting the model results. Historical Variability and Crop Participation also contributed substantially to the classification, suggesting that temporal production variation and the relative importance of crops within municipalities provide relevant information for differentiating agricultural profiles. The agricultural profiles defined in this study classify individual agricultural records representing municipality–crop–year observations according to their productive characteristics. The resulting classification identifies different combinations of productivity and harvest efficiency and provides a territorial representation based on the proposed indicators. For the spatial analysis, the profile assigned to individual records was aggregated by municipality using the most frequent profile, allowing the distribution of the profiles to be represented geographically. Similar applications have recently been discussed within the context of intelligent agricultural ecosystems and territorial agricultural analytics [25], [26].

Most previous studies have focused on crop yield prediction, crop mapping, or agricultural forecasting [2]–[9]. In contrast, this study classified agricultural records into profiles according to their agricultural performance using publicly available agricultural information. The incorporation of territorial indicators together with spatial representation provides an additional perspective for interpreting agricultural performance and complements traditional productivity-based evaluations. This approach also contributes to current research seeking to combine machine learning with territorial information for agricultural planning and decision support [23], [25], [26].

This study combines territorial indicators with machine learning to characterize agricultural profiles using information from the Municipal Agricultural Assessments. The proposed methodology provides a framework for analyzing agricultural records according to productivity, harvest efficiency, crop participation, and temporal variability. The approach can also be implemented using other agricultural databases that contain comparable territorial information [25]. These characteristics support the use of agricultural databases not only for statistical reporting but also for analytical approaches aimed at understanding differences in agricultural performance across municipalities.

5. Limitations of the study

This study was based exclusively on information from the Municipal Agricultural Assessments (EVA) database. The territorial indicators were calculated from the information reported by Colombian municipalities. Differences in data collection procedures, incomplete records, and reporting inconsistencies may influence the calculated indicators. Therefore, the results should be interpreted within the scope and quality of the available agricultural records.

The analysis focused on variables derived from agricultural production records. Climatic conditions, soil characteristics, irrigation availability, technological adoption, market accessibility, and socioeconomic conditions were not included in the analysis. These variables represent additional sources of information for describing territorial agricultural differences. Their incorporation could provide additional information for explaining variations in agricultural performance among municipalities and agricultural systems.

The agricultural profiles were derived from historical records collected between 2019 and 2024 and were assigned to individual agricultural records representing municipality–crop–year observations. Accordingly, the classifications represent the agricultural conditions observed during that period, while the spatial representation summarizes the resulting profiles at the municipal level. The four profiles were defined using the median values of Real Productivity and Harvest Efficiency, and the sensitivity of the resulting classification to alternative threshold definitions was not evaluated. In addition, Loss Index is mathematically related to Harvest Efficiency, which should be considered when interpreting its relatively high importance in the classification model.

Future studies may incorporate satellite information, climatic variables, soil properties, and socioeconomic indicators together with the territorial indicators used in this study. Additional machine learning algorithms and spatiotemporal models also represent alternatives for extending the territorial characterization of agricultural performance. Alternative thresholding methods could also be evaluated to determine whether different profile definitions produce similar classifications. These extensions could provide a broader representation of the factors associated with agricultural performance and support further analysis of territorial agricultural patterns.

6. Conclusions

Agricultural profiles were classified for Colombian agricultural records using information from the Municipal Agricultural Assessments (EVA) database and supervised machine learning models. The analysis was based on five territorial indicators derived from the original agricultural records: Real Productivity, Harvest Efficiency, Loss Index, Crop Participation, and Historical Variability. The profiles were assigned to individual agricultural records representing municipality–crop–year observations. The resulting classifications were subsequently represented spatially at the municipal level to describe territorial agricultural patterns.

Productivity alone did not describe all differences in agricultural performance across Colombian municipalities. The combination of territorial indicators identified additional differences among agricultural records beyond production values. The descriptive and spatial analyses documented variability in agricultural conditions across the country. The results also showed that neighboring municipalities were not always associated with the same dominant agricultural profile.

The five-fold stratified group validation showed consistent classification performance for both models when municipalities were separated across the validation folds. Decision Tree achieved an Accuracy of 0.777 ± 0.015 ,

Precision of 0.780 ± 0.014 , Recall of 0.777 ± 0.015 , and F1-score of 0.776 ± 0.015 , while Random Forest obtained an Accuracy of 0.764 ± 0.009 , Precision of 0.764 ± 0.010 , Recall of 0.764 ± 0.009 , and F1-score of 0.764 ± 0.009 . Loss Index showed the highest variable importance (44.8%), followed by Historical Variability (33.6%) and Crop Participation (21.3%). Dominant Crop showed a limited contribution to the classification, with an importance of 0.3%.

The spatial distribution of the agricultural profiles identified different territorial patterns across Colombian municipalities. Neighboring municipalities were not always classified within the same agricultural profile, indicating heterogeneity in agricultural performance across nearby territories. The spatial representation therefore complements the descriptive analysis by showing how the proposed profiles are distributed geographically. These patterns provide a territorial perspective on agricultural conditions that is not obtained from production values alone.

The proposed methodology combines territorial indicators, spatial representation, and machine learning to classify agricultural profiles using publicly available agricultural information. The methodology extends the use of the Municipal Agricultural Assessments (EVA) database beyond descriptive reporting by incorporating derived indicators and supervised classification. The municipality-level validation also provides a more stringent assessment of model performance by preventing records from the same municipality from being shared between training and validation folds. Overall, the framework provides a basis for characterizing heterogeneous agricultural conditions across Colombian territories using large-scale agricultural records.

Statements

The authors report no financial, professional, or personal conflicts of interest associated with this study.

Data availability

The Municipal Agricultural Assessments (EVA) database used in this study is publicly available through the Colombian Government Open Data portal. The processed datasets and the Python scripts used for data preprocessing, indicator calculation, machine learning, and spatial analysis are available from the corresponding author upon reasonable request.

Information on financing

This research did not receive any external funding.

References

- [1] Ministry of Agriculture and Rural Development, "Municipal Agricultural Assessments (EVA) 2019–2024," Open Data Colombia, 2024. [Online]. Available: <https://www.datos.gov.co/browse?q=Evaluaciones%20Agropecuarias%20Municipales%20EVA>
- [2] J. van Klompenburg, A. Kassahun, and C. Catal, "Crop Yield Prediction Using Machine Learning: A Systematic Literature Review," *Computers and Electronics in Agriculture*, vol. 177, p. 105709, 2020. doi: 10.1016/j.compag.2020.105709.
- [3] M. A. Machichi et al., "Crop Mapping Using Supervised Machine Learning and Deep Learning: A Systematic Literature Review," *International Journal of Remote Sensing*, vol. 44, pp. 2717–2753, 2023. doi: 10.1080/01431161.2023.2205984.
- [4] H. Hoppe, P. Dietrich, P. Marzahn, T. Weiß, C. Nitzsche, U. Freiherr von Lukas, T. Wengerek, and E. Borg, "Transferability of Machine Learning Models for Crop Classification in Remote Sensing Imagery Using a New Test Methodology: A Study on Phenological, Temporal, and Spatial Influences," *Remote Sensing*, vol. 16, no. 9, p. 1493, 2024. doi: 10.3390/rs16091493.
- [5] K. Singh et al., "Machine-Learning-Based Frameworks for Reliable and Sustainable Crop Forecasting," *Sustainability*, vol. 17, no. 10, p. 4711, 2025. doi: 10.3390/su17104711.

-
- [6] S. A. Mohamed et al., “A Hybrid Deep Learning and Rule-Based Model for Intelligent Weather Forecasting and Crop Recommendations Using Satellite Imagery,” *Scientific Reports*, vol. 15, p. 36102, 2025. doi: 10.1038/s41598-025-21506-4.
- [7] E. Asamoah, G. B. M. Heuvelink, P. S. Bindraban, and V. Logah, “Modeling Maize Yield and Agronomic Efficiency Using Machine Learning Models: A Comparative Analysis,” *Agronomy Journal*, 2025. doi: 10.1002/agj2.70206..
- [8] M. Croci, G. Impollonia, M. Meroni, and S. Amaducci, “Dynamic Predictions of Maize Yield Using Machine Learning on Data from Multiple Sources,” *Remote Sensing*, vol. 15, no. 1, p. 100, 2022. doi: 10.3390/rs15010100.
- [9] B. M. Lionel, R. Musabe, O. Gatera et al., “A Comparative Study of Machine Learning Models in Predicting Crop Yield,” *Discover Agriculture*, vol. 3, p. 151, 2025. doi: 10.1007/s44279-025-00335-z.
- [10] J. S. Fandiño Pelayo, L. S. Mendoza Castellanos, R. Cazes Ortega, and L. G. Hernández-Rojas, “AI-Driven Monitoring for Fish Welfare in Aquaponics: A Predictive Approach,” *Sensors*, vol. 25, no. 19, p. 6107, 2025. doi: 10.3390/s25196107.
- [11] J. Fandiño Pelayo, “Engineering Evaluation of Water Quality Correlations for Passive Aquaponic Systems in Tropical Environments,” *Periodicals of Engineering and Natural Sciences*, vol. 13, no. 4, pp. 785–792, 2025. doi: 10.21533/pen.v13.i4.636.
- [12] J. Fandiño Pelayo, L. S. Mendoza Castellanos, R. Cazes Ortega, L. G. Hernández Rojas, and O. Lengerke, “Intelligent Water Quality Regulation in Sustainable Aquaponics Using PID Control and Long-Term Performance Analysis,” *Sustainable Engineering and Innovation*, vol. 7, no. 2, pp. 521–536, 2025. doi: 10.37868/sei.v7i2.id606.
- [13] R. Cazes Ortega, A. Pardo Garcia, J. S. Fandiño Pelayo, and C. L. Sandoval Rodriguez, “Analysis of the Energy Efficiency of an Edge-AI Irrigation System for Cocoa Production in Rural Colombia,” *Heritage and Sustainable Development*, vol. 7, no. 2, pp. 1031–1040, 2025. doi: 10.37868/hsd.v7i2.1480.
- [14] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.
- [15] G. G. Otoro and K. Komai, “Integration of Machine Learning and Remote Sensing to Evaluate the Effects of Soil Salinity, Nitrate, and Moisture on Crop Yields and Economic Returns in the Semi-Arid Region of Ethiopia,” *Agriculture*, vol. 15, no. 22, p. 2378, 2025. doi: 10.3390/agriculture15222378.
- [16] T. Tian, J. Yang, N. Zheng, and J. Tang, “Characteristics of Spatiotemporal Trends and Influencing Factors of Eco-Efficiency in Cultivated Land Use in the Mountainous Regions of Southern China,” *Frontiers in Sustainable Food Systems*, vol. 9, 2025. doi: 10.3389/fsufs.2025.1760810.
- [17] X. Chen, “The Role of Modern Agricultural Technologies in Improving Agricultural Productivity and Land-Use Efficiency,” *Frontiers in Plant Science*, vol. 16, 2025. doi: 10.3389/fpls.2025.1675657.
- [18] J. R. Luque-Reyes, A. Zidi, A. Peña-Acevedo, and R. Gallardo-Cobos, “Assessing Agri-Food Digitalization: Insights from Bibliometric and Survey Analysis in Andalusia,” *World*, vol. 6, no. 2, p. 57, 2025. doi: 10.3390/world6020057.
- [19] A. Cravero, S. Pardo, S. Sepúlveda, and L. Muñoz, “Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review,” *Agronomy*, vol. 12, no. 3, p. 748, 2022. doi: 10.3390/agronomy12030748.
- [20] A. Khanna and S. Kaur, “The Evolution of the Internet of Things (IoT) and Its Significant Impact on Precision Agriculture,” *Computers and Electronics in Agriculture*, vol. 157, pp. 218–231, 2019. doi: 10.1016/j.compag.2018.12.039.[22] J. L. Hintze and R. D. Nelson, "Violin Plots: A Box Plot-Density
-

- Trace Synergism," *The American Statistician*, vol. 52, no. 2, pp. 181-184, 1998. <https://doi.org/10.1080/00031305.1998.10480559>
- [21] O. Coca, D. Creangă, Ș. Viziteu, I. S. Brumă, and G. Ștefan, "Analysis of the Determinants of Agriculture Performance at the European Union Level," *Agriculture*, vol. 13, no. 3, p. 616, 2023. doi: 10.3390/agriculture13030616.
- [22] J. P. M. Xavier, K. Schenatto, G. V. Miranda, C. L. Bazzi, R. Sobjak, and M. Rodrigues, "A Machine Learning Framework for Crop Productivity Classification and Risk Assessment," *AgriEngineering*, vol. 8, no. 6, p. 203, 2026. doi: 10.3390/agriengineering8060203.
- [23] N. D. Muñoz, J. Barón-Velandia, and S.-C. Vanegas-Ayala, "Descriptive Analysis and Clustering-Based Productive Scale Segmentation of Colombian Transitory Crop Production: A Departmental-Level Approach," *Agriculture*, vol. 16, no. 9, p. 980, 2026. doi: 10.3390/agriculture16090980.
- [24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [25] A. Cravero, S. Sepúlveda, F. Gutiérrez, and L. Muñoz, "From Precision Agriculture to Intelligent Agricultural Ecosystems: A Systematic Review of Machine Learning and Big Data Applications," *Agronomy*, vol. 16, no. 5, p. 516, 2026. doi: 10.3390/agronomy16050516.
- [26] G. Giannarakis, V. Sitokonstantinou, R. S. Lorilla, and C. Kontoes, "Towards Assessing Agricultural Land Suitability with Causal Machine Learning," in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, Jun. 2022, pp. 1441–1451. doi: 10.1109/CVPRW56347.2022.00150.