Article

Not peer-reviewed version

Explainable Machine Learning for Crop Yield Classification Using Foliar Nutrient Analysis and Management Data in Colombia

Yeison Eduardo Conejo Sandoval * and Andres Polo

Posted Date: 26 March 2026

doi: 10.20944/preprints202603.2121.v1

Keywords: crop yield classification; machine learning; foliar analysis; nutrient management; precision agriculture; Random Forest; agricultural data; Colombia

 doi:10.20944/preprints202603.2121.v1

Article

Explainable Machine Learning for Crop Yield Classification Using Foliar Nutrient Analysis and Management Data in Colombia

Yeison Eduardo Conejo Sandoval * and Andrés Polo

Department of Industrial Engineering, Fundación Universitaria Agraria de Colombia, Bogotá 110110, Colombia

* Correspondence: conejos.yeison@uniagraria.edu.co

Abstract

Accurate crop yield prediction is essential for improving agricultural productivity, resource management, and food security, particularly in heterogeneous environments such as Colombia. Recent advances in machine learning have enhanced predictive capabilities; however, most existing approaches rely predominantly on climatic or image-based data, limiting their direct applicability to agronomic decision-making. This study proposes a machine learning framework for crop yield classification based on foliar nutrient analysis, fertilization practices, and geographic variables using open-access agricultural data. The approach formulates yield prediction as a multi-class classification problem, enabling the identification of performance levels that are more interpretable and actionable in practical contexts. Four machine learning models—Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting—were evaluated. The results show that ensemble-based methods outperform alternative approaches, with Random Forest achieving the highest accuracy (96.27%) and macro F1-score (0.9261), followed by Gradient Boosting (95.34%). Feature importance analysis reveals that geographic location, crop type, and foliar nutrients such as sulphur, nitrogen, magnesium, calcium, potassium, and zinc are the most influential predictors. These findings demonstrate that nutrient-based variables provide a direct and meaningful representation of crop performance, offering advantages over models based solely on environmental proxies. By integrating foliar analysis with management practices, the proposed framework enhances interpretability and supports agronomic decision-making, contributing to the advancement of precision agriculture in data-scarce and heterogeneous contexts.

Keywords: crop yield classification; machine learning; foliar analysis; nutrient management; precision agriculture; Random Forest; agricultural data; Colombia

1. Introduction

Accurate crop yield prediction has become a central concern in modern agricultural systems, particularly under increasing pressure to ensure sustainability, optimize resource use, and strengthen food security. As global demand for agricultural products continues to grow in parallel with climate variability, the capacity to anticipate production levels is no longer optional but essential for both strategic planning and operational decision-making [1,2]. Within this evolving landscape, quantitative approaches combining statistical models, crop simulation, and artificial intelligence have progressively expanded the analytical capabilities available to agricultural stakeholders [3].

This challenge becomes even more critical in regions characterized by high environmental variability, such as South America, where agricultural productivity is strongly influenced by climate fluctuations, including droughts, irregular rainfall patterns, and extreme weather events [4]. In such contexts, yield prediction transcends its analytical role and becomes a key mechanism for reducing supply chain vulnerability, improving input allocation, and supporting national food security

 doi:10.20944/preprints202603.2121.v1

strategies [5]. In Colombia, where the agricultural sector contributes significantly to both employment and economic development, improving productivity and resilience remains a persistent priority [4,6]). Nevertheless, much of the existing work has focused on explaining yield variability retrospectively, rather than developing predictive tools capable of informing proactive decision-making [7].

The increasing adoption of machine learning (ML) has opened new possibilities for addressing the complexity of agricultural systems [8]. Unlike traditional statistical approaches, ML models can capture nonlinear relationships and interactions among heterogeneous variables, including climate, soil conditions, and management practices [9]. This capability has driven their application across multiple domains, ranging from yield prediction to disease detection and land suitability analysis [10]. Empirical evidence consistently shows that algorithms such as Random Forest and Gradient Boosting outperform linear models when dealing with high-dimensional and nonlinear agricultural data [11].

However, closer examination of recent advances reveals that improvements in predictive accuracy do not necessarily translate into practical applicability. A considerable portion of the literature remains centered on climate-driven variables or image-based datasets, which, although valuable, only partially represent the underlying agronomic processes governing crop performance [12]. Deep learning models developed for crop disease detection frequently achieve high accuracy under controlled experimental conditions but exhibit substantial performance degradation when exposed to real field environments, largely due to dataset imbalance, noise, and limited representativeness of rare conditions [13]. A similar pattern emerges in yield prediction models, where performance tends to decline when applied beyond the range of the training data, highlighting persistent issues of generalization and transferability.

At the same time, interpretability remains an unresolved concern. Although techniques such as feature importance and SHAP (SHapley Additive exPlanations) values have improved transparency, their use often remains confined to variable ranking, without establishing a clear connection to agronomic decision-making [14]. As a result, many ML models provide predictions that are difficult to translate into actionable strategies for farmers or policymakers. This limitation becomes more evident when considering that key agronomic drivers—such as nutrient availability, fertilization practices, and plant physiological status—are frequently underrepresented in predictive frameworks [12].

Within this context, foliar analysis offers a direct and agronomically meaningful perspective on crop performance. By capturing the nutritional status of plants, foliar data reflect physiological conditions that are closely linked to yield formation [7]. When integrated with management practices and environmental conditions, this type of information enables a more comprehensive representation of the production system, facilitating the development of models that are not only predictive but also interpretable and operational [15]. Despite its relevance, the incorporation of foliar and nutrient-based variables into machine learning models remains limited in the current literature.

Building on these observations, this study proposes a machine learning framework for crop yield classification that integrates foliar analysis and management variables within the Colombian agricultural context. The approach leverages open-access data from the Corporación Colombiana de Investigación Agropecuaria to develop a predictive model that categorizes crop performance into discrete levels. This formulation allows the results to be more easily interpreted and applied in decision-making processes, particularly in environments characterized by high heterogeneity and data limitations.

Rather than focusing exclusively on improving predictive accuracy, the proposed framework emphasizes the integration of agronomically relevant variables and the generation of interpretable outputs that can support nutrient management and fertilization strategies. In doing so, the study contributes to bridging the gap between data-driven modeling and practical agricultural decision-making, offering a context-aware approach aligned with the principles of precision agriculture and sustainable production systems.

 doi:10.20944/preprints202603.2121.v1

2. Literature Review

Crop yield prediction has evolved into a central research domain within precision agriculture due to its direct implications for resource optimization, sustainability, and food security. Recent studies highlight that the increasing complexity of agricultural systems, combined with climate variability, has driven the adoption of quantitative and data-driven approaches that integrate statistical modeling, crop simulation, and artificial intelligence [3]. In this context, machine learning has emerged as a powerful tool capable of capturing nonlinear relationships and interactions among multiple agronomic and environmental variables, surpassing the limitations of traditional linear approaches [2,7].

A significant portion of the literature has focused on modeling crop yield as a function of climatic and environmental variables [16,17]. These studies commonly incorporate precipitation, temperature, evapotranspiration, soil moisture, and vegetation indices derived from remote sensing. Evidence suggests that such variables are strong predictors of yield variability, particularly under drought conditions and extreme climatic events [5,18]. Drought-related indicators have been widely used to anticipate yield anomalies at regional scales, demonstrating that machine learning models can provide useful forecasts several months before harvest [19]. This line of research has contributed to improving large-scale agricultural planning and understanding climate-yield relationships.

Despite these advances, the strong dependence on climate-driven predictors introduces important limitations. Model performance often varies across regions, scales, and datasets, revealing challenges related to generalization and transferability. Studies have shown that models trained under specific environmental conditions tend to lose accuracy when applied outside their training domain, especially when extreme events are underrepresented in the data [20]. This issue is further exacerbated by the inherent heterogeneity of agricultural systems, where local soil conditions, management practices, and crop varieties significantly influence yield outcomes [21].

Another dominant research stream has explored the use of image-based and deep learning approaches for crop monitoring and disease detection [2]. These models, particularly convolutional neural networks, have demonstrated high accuracy in controlled datasets and have enabled early detection of plant diseases and stress conditions [14]. However, recent reviews emphasize that such performance is often achieved under laboratory conditions, with limited dataset diversity and significant class imbalance [7]. As a result, models frequently experience performance degradation when deployed in real field environments, where variability in lighting, background, and plant conditions introduces additional uncertainty [5]. This gap between experimental accuracy and real-world applicability remains a critical challenge in agricultural machine learning.

In parallel, interpretability has gained attention as a necessary component of machine learning models in agriculture [22]. Techniques such as feature importance analysis and SHAP values have been increasingly adopted to identify the contribution of different variables to model predictions. These approaches represent an important step toward transparency and trust in data-driven systems. However, existing studies often limit interpretability to variable ranking without establishing a direct connection to agronomic decision-making [23]. Consequently, the practical usefulness of these models remains constrained, as farmers and decision-makers require insights that can inform specific actions, such as fertilization strategies or crop management adjustments.

Within this context, agronomic variables directly related to plant physiology, particularly nutrient availability and management practices, have received comparatively limited attention in predictive modeling. This gap is notable given that crop yield is strongly influenced by nutrient balance, fertilization history, and soil-plant interaction [7]. Foliar analysis provides a direct assessment of the nutritional status of plants and enables the identification of deficiencies or excesses that affect crop performance. Previous studies have highlighted the relevance of integrating nutrient data with agronomic management to improve diagnostic accuracy and support site-specific fertilization strategies [15,20]. However, the incorporation of foliar variables into machine learning frameworks for yield prediction remains scarce and fragmented.

 doi:10.20944/preprints202603.2121.v1

This limitation is particularly evident in the Colombian context. Although the importance of yield prediction for agricultural planning and food security has been widely recognized, most regional studies have focused on descriptive analyses of climate impacts or general discussions on precision agriculture [1]. As a result, there is a lack of predictive models that integrate agronomically meaningful variables, such as nutrient status and fertilization practices, within a data-driven framework tailored to local conditions. This gap restricts the ability of producers to implement precision agriculture strategies and limits the operational value of existing analytical tools.

The present study contributes to addressing this gap by developing a machine learning framework for crop yield classification based on foliar analysis and management variables using open data from Colombia. This approach introduces a shift in perspective by focusing on nutrient-driven and management-oriented predictors that are directly linked to crop performance. In addition, the formulation of yield prediction as a classification problem enhances interpretability and facilitates its use in practical decision-making contexts, where categorical assessments of performance are often more actionable than continuous estimates.

By integrating foliar nutrient information, fertilization history, and geographic variability, the proposed framework advances beyond the predominant climate-centered and image-based approaches in the literature. Its contribution lies in bridging the disconnect between predictive modeling and agronomic decision-making, offering a methodology that is both analytically robust and operationally relevant for heterogeneous agricultural systems such as those found in Colombia.

3. Methodology

3.1. Research Design and Data Source

This study adopts a quantitative and data-driven research design aimed at developing an interpretable machine learning framework for crop yield classification. The approach is grounded in the integration of agronomic, nutritional, and management variables, with the objective of capturing the multidimensional nature of crop performance in heterogeneous agricultural systems. The dataset used in this study was obtained from the Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), which provides open-access agricultural data across multiple regions of Colombia. The dataset includes information on crop performance, foliar nutrient composition, fertilization practices, and geographic attributes. This combination of variables enables the representation of both physiological and management-related drivers of yield variability, which are often treated separately in existing studies.

Prior to modeling, the dataset was examined to ensure consistency, completeness, and suitability for classification tasks. Records with missing or inconsistent values were filtered or corrected where appropriate, and categorical variables were standardized to ensure compatibility across observations. Continuous variables, particularly nutrient concentrations, were preserved in their original scale to maintain their agronomic interpretability.

The dataset was constructed from open-access records provided by AGROSAVIA. Initial filtering criteria were applied to ensure statistical robustness, excluding crops with fewer than 20 observations. This process resulted in a final dataset comprising 3,219 observations across 27 crop types. A total of 17 explanatory variables were retained, including temporal, geographic, agronomic, and nutritional attributes. These variables capture key dimensions of crop performance, ranging from management practices to plant physiological status. The response variable corresponds to crop yield performance, expressed as a categorical variable.

Table 1. *Explanatory Variables Used in the Study.*

Variable	Data Type	Description
Year	Numerical	Year of data collection
Department	Categorical	Geographic location of the sample (department level)

 doi:10.20944/preprints202603.2121.v1

Crop	Categorical	Crop type analyzed
Fertilization	Categorical	Type of fertilization applied
Nitrogen (%)	Numerical	Nitrogen content determined using the Kjeldahl method
Phosphorus (%)	Numerical	Phosphorus concentration in plant tissue
Potassium (%)	Numerical	Potassium concentration in plant tissue
Calcium (%)	Numerical	Calcium concentration in plant tissue
Magnesium (%)	Numerical	Magnesium concentration in plant tissue
Sodium (%)	Numerical	Sodium concentration determined through laboratory analysis
Sulfur (%)	Numerical	Sulfur content determined after wet digestion
Iron (mg/kg)	Numerical	Iron concentration measured in plant tissue (ICP-OES)
Copper (mg/kg)	Numerical	Copper concentration measured in plant tissue (ICP-OES)
Manganese (mg/kg)	Numerical	Manganese concentration measured in plant tissue (ICP-OES)
Zinc (mg/kg)	Numerical	Zinc concentration measured in plant tissue (ICP-OES)
Boron (mg/kg)	Numerical	Boron concentration measured in plant tissue (ICP-OES)
Yield Performance	Categorical	Target variable representing crop yield classification (high, medium, low)

Categorical variables, including crop type, geographic location, and fertilization practices, were transformed using one-hot encoding to enable their integration into machine learning models. This transformation allows the representation of qualitative information in a numerical format while preserving category-specific effects. Nutrient-related variables were maintained in their original measurement scales, expressed either as percentage values or concentrations (mg/kg), depending on the nutrient type. These variables reflect the foliar composition of the plant and provide a direct representation of its nutritional status. The integration of encoded categorical variables with continuous nutrient data enables the model to capture both management-driven and physiological drivers of crop performance.

3.2. Feature Construction and Variable Selection

The modeling framework is built upon three main groups of predictors: foliar nutrient variables, agronomic management variables, and geographic descriptors. Foliar variables represent the nutritional status of the plant and include macro- and micronutrients such as nitrogen (N), phosphorus (P), potassium (K), calcium (Ca), magnesium (Mg), sulfur (S), sodium (Na), boron (B), and zinc (Zn). These variables provide a direct physiological signal associated with plant development and productivity, reflecting nutrient availability and uptake efficiency [7]. Management variables capture the history and characteristics of fertilization practices, including type, frequency, and intensity of nutrient application. These variables are essential for linking observed nutritional status with prior agronomic decisions, enabling the model to account for human-driven interventions in the production process [15,20]. Geographic variables incorporate spatial information that reflects environmental heterogeneity, such as location or region. These descriptors serve as proxies for unobserved factors, including soil properties, microclimatic conditions, and local management traditions, which may influence crop performance [3].

The selection of these variables is consistent with the objective of developing a model that is not only predictive but also interpretable and agronomically meaningful. Instead of relying exclusively on environmental proxies, the framework emphasizes variables that can be directly linked to management decisions and physiological processes.

 doi:10.20944/preprints202603.2121.v1

3.3. Yield Classification Framework

Crop yield was formulated as a classification problem to enhance interpretability and practical applicability. Instead of predicting continuous yield values, the model categorizes crop performance into three discrete classes: high, medium, and low yield. This transformation addresses two methodological considerations. On one hand, classification reduces sensitivity to noise and variability inherent in agricultural data, which often affects regression-based approaches. On the other hand, categorical outputs are more aligned with decision-making processes in agricultural practice, where producers frequently operate based on performance thresholds rather than precise numerical estimates.

The class labels were constructed based on the distribution of yield values in the dataset. Thresholds were defined to ensure a balanced representation of classes while preserving meaningful distinctions between performance levels. This approach allows the model to capture relative differences in productivity while maintaining statistical robustness.

3.4. Machine Learning Models

The predictive framework employs supervised machine learning algorithms, with a primary focus on Random Forest (RF) due to its robustness in handling nonlinear relationships, heterogeneous data, and interactions among variables [2]. Random Forest operates as an ensemble of decision trees constructed through bootstrap aggregation (bagging), where each tree is trained on a random subset of data and features. This structure reduces variance, mitigates overfitting, and improves generalization performance, making it particularly suitable for agricultural datasets characterized by noise and complexity.

The choice of RF is further supported by its interpretability. Feature importance measures derived from the model allow the identification of key predictors driving classification outcomes, facilitating the connection between model results and agronomic reasoning. To ensure consistency and reliability, the model was trained and evaluated using a randomized data-splitting strategy. The dataset was divided into training and testing subsets, typically following a 70–30 proportion. This separation allows the assessment of model performance on unseen data, providing a more realistic estimate of predictive capability.

3.5. Model Evaluation Metrics

Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. Given the potential imbalance among yield classes, the macro-average F1-score was used as a complementary metric to ensure that performance was not biased toward dominant classes. Accuracy provides a general measure of correct classifications, while precision and recall offer insights into the model's ability to correctly identify each class. The F1-score balances these two aspects, making it particularly useful in multi-class classification problems.

In addition to these metrics, variability in model performance was assessed through repeated training and testing iterations. This procedure allows the identification of potential instability in the model and ensures that the reported results are not dependent on a single random split of the data.

Beyond predictive performance, the methodological design emphasizes interpretability as a key component of the framework. Feature importance analysis was conducted to identify the variables that most strongly influence yield classification.

This analysis enables the translation of model outputs into agronomic insights. For instance, the identification of specific nutrients as dominant predictors can inform fertilization strategies, while the relevance of management variables can highlight the impact of prior interventions on crop performance. By linking predictive results with actionable variables, the proposed framework supports decision-making processes in agricultural systems. This integration of prediction and interpretation addresses a critical limitation in existing machine learning applications, where high accuracy is often achieved without providing guidance for practical implementation.

 doi:10.20944/preprints202603.2121.v1

4. Results

4.1. Predictive Performance of Machine Learning Models

The comparative evaluation of the four classification algorithms on the test dataset (n = 644) reveals clear differences in predictive performance, highlighting the superiority of ensemble-based methods over linear approaches. Logistic Regression achieved an overall accuracy of 91.92% with a macro-averaged F1-score of 0.8527. Its class-wise performance shows moderate consistency, with F1-scores of 0.81 for low yield, 0.80 for medium yield, and 0.95 for high yield. These results indicate a stronger ability to correctly identify high-performance crops, while lower performance is observed in distinguishing medium and low yield categories.

Decision Tree improves upon Logistic Regression, reaching an accuracy of 95.49% and a macro F1-score of 0.9096. Its classification report shows balanced performance across categories, with F1-scores of 0.85 (low), 0.90 (medium), and 0.97 (high). This improvement suggests that nonlinear decision boundaries enhance the model's ability to capture interactions among agronomic variables. Random Forest presents the best overall performance, achieving an accuracy of 96.27% and a macro F1-score of 0.9261. Its classification results indicate high precision and recall across all classes, with F1-scores of 0.88 (low), 0.92 (medium), and 0.98 (high). This performance reflects the robustness of ensemble learning in capturing complex patterns in the data. Gradient Boosting also shows strong and stable performance, with an accuracy of 95.34% and a macro F1-score of 0.9206. Its class-wise F1-scores are 0.90 (low), 0.89 (medium), and 0.97 (high), positioning it as the second-best model in terms of predictive capability. The overall performance of the evaluated models is summarized in Table 2.

Table 2. *Performance of Machine Learning Models for Yield Classification.*

Model	Accuracy	F1 (Macro)
Random Forest	0.9627	0.9261
Gradient Boosting	0.9534	0.9206
Decision Tree	0.9549	0.9096
Logistic Regression	0.9192	0.8527

These results confirm that ensemble-based models provide superior predictive performance, particularly in capturing nonlinear relationships among agronomic variables. In contrast, Logistic Regression exhibits lower performance, suggesting limitations in modelling complex interactions within the dataset.

The classification reports (Tables 3–6) provide a detailed evaluation of model behavior across yield categories (low, medium, high). Across all models, classification performance is consistently higher for the high-yield category, which shows the highest precision, recall, and F1-score values. For instance, Random Forest achieves an F1-score of 0.98 for this class, while Gradient Boosting and Decision Tree reach 0.97. This indicates that high-performance crops present clearer patterns in the feature space.

Table 3. *Classification Report for Logistic Regression.*

Class	Precision	Recall	F1-score
Low	0.88	0.75	0.81
Medium	0.88	0.73	0.8
High	0.93	0.98	0.95
Accuracy	—	—	0.9192

Table 4. *Classification Report for Decision Tree.*

Class	Precision	Recall	F1-score
Low	0.88	0.82	0.85

 doi:10.20944/preprints202603.2121.v1

Medium	0.91	0.9	0.9
High	0.97	0.98	0.97
Accuracy	—	—	0.9549

Table 5. *Classification Report for Random Forest.*

Class	Precision	Recall	F1-score
Low	1	0.79	0.88
Medium	0.94	0.9	0.92
High	0.97	0.99	0.98
Accuracy	—	—	0.9627

Table 6. *Classification Report for Gradient Boosting.*

Class	Precision	Recall	F1-score
Low	1	0.82	0.9
Medium	0.96	0.83	0.89
High	0.95	0.99	0.97
Accuracy	—	—	0.9534

In contrast, the medium-yield category exhibits slightly lower and more variable performance across models. This suggests that intermediate productivity levels are more difficult to classify, likely due to overlapping characteristics and compensatory effects among variables. The low-yield category also shows strong performance, although with slightly lower recall values in some models, indicating occasional misclassification with adjacent classes. The confusion matrix presented in Figure 1 supports these findings, showing that misclassification is more frequent in the medium-yield class, while high and low categories are more distinctly separated.

The analysis of variable importance for Decision Tree, Random Forest, and Gradient Boosting shows consistent patterns across the three models, confirming the robustness of the identified predictors. As illustrated in Figure 1, geographic location—particularly the department of Cundinamarca—emerges as the most influential variable, with an importance value of 0.062 in Random Forest. This dominance is also observed in Decision Tree (0.203) and Gradient Boosting (0.222), suggesting strong regional effects associated with agroecological conditions.

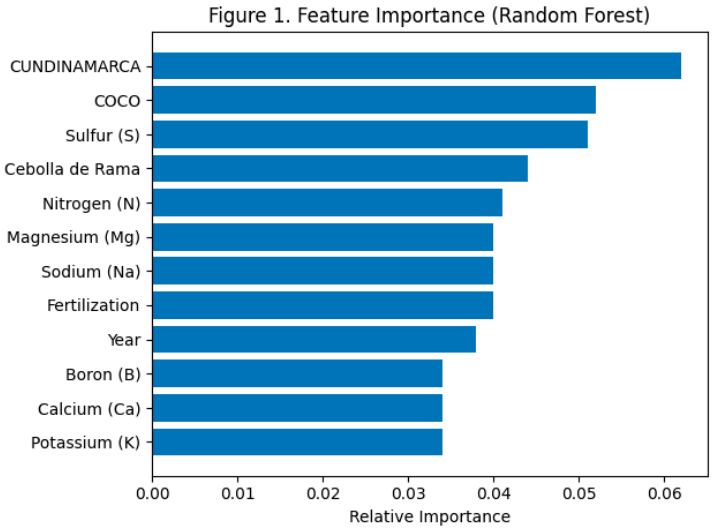

Figure 1. *Feature Importance (Random Forest).*

Crop type also plays a significant role, with COCO and CACAO consistently ranked among the top predictors across models. This indicates the presence of crop-specific performance patterns,

 doi:10.20944/preprints202603.2121.v1

particularly in perennial systems. Nutrient variables contribute substantially to model predictions. Among the most relevant are sulfur (S) (0.051), nitrogen (N) (0.041), magnesium (Mg) (0.040), sodium (Na) (0.040), calcium (Ca) (0.034), potassium (K) (0.034), and micronutrients such as boron (B), zinc (Zn), manganese (Mn), iron (Fe), and copper (Cu), all with values around 0.032–0.034. Fertilization practices also appear as key predictors, particularly combinations of nutrients such as N, P, K, Ca, S, B, Zn, Mg, Cu, Fe, Mn, and Mo (0.040), along with temporal variables such as year of record (0.038). These findings highlight the relevance of both management decisions and temporal variability in determining crop performance. Overall, approximately 67% of the most influential variables correspond to foliar nutrients, 20% to geographic factors, and 13% to management practices. This distribution confirms the dominant role of plant nutritional status in explaining yield variability.

4.2. Interaction Between Nutritional and Management Variables

The combined interpretation of model performance and feature importance indicates that crop yield is not determined by isolated variables but by the interaction between nutritional status, geographic conditions, and agronomic management.

Fertilization practices influence nutrient availability, which is reflected in foliar composition and ultimately affects yield classification. This interaction creates a system where management decisions are directly linked to physiological responses in plants. The ability of ensemble models to capture these interactions explains their superior predictive performance. These models effectively represent the combined influence of environmental conditions, nutrient dynamics, and management practices.

Figure 2 presents a comparative analysis of the top predictors identified by Decision Tree, Random Forest, and Gradient Boosting. A strong structural consistency is observed across the three models. The department of Cundinamarca ranks as the most important predictor in all cases, followed by COCO as the second most relevant variable. CACAO also appears consistently among the top predictors.

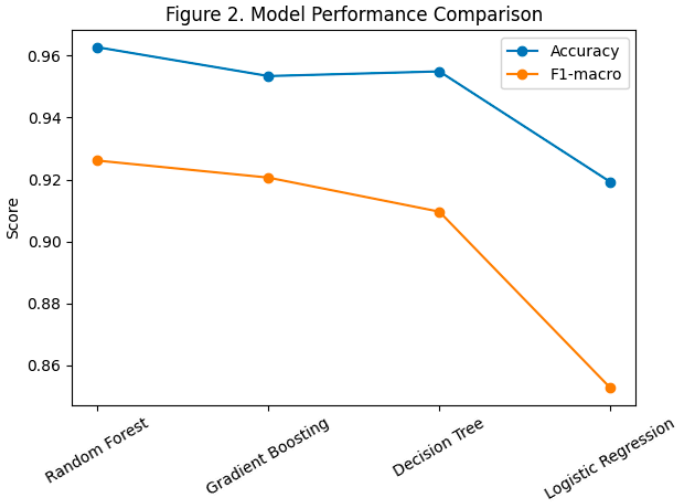

Figure 2. *Model Performance Comparison.*

While differences exist in the remaining variables, these reflect complementary perspectives rather than contradictions. Decision Tree and Gradient Boosting emphasize geographic and temporal variables, capturing broader agroecological patterns. Random Forest, on the other hand, assigns greater importance to foliar nutrient variables, indicating higher sensitivity to biochemical interactions. This convergence in the top predictors, combined with variation in secondary variables, supports the robustness of the findings and suggests that crop yield is driven by a combination of regional conditions, crop type, and nutritional status.

 doi:10.20944/preprints202603.2121.v1

5. Discussion

The results obtained in this study provide consistent evidence that ensemble-based machine learning models outperform linear and single-tree approaches in the classification of crop yield. This finding aligns with previous research indicating that agricultural systems exhibit strong nonlinear behavior, where interactions among environmental, agronomic, and physiological variables play a central role [2]. The superior performance of Random Forest and Gradient Boosting confirms that capturing such interactions is essential for accurate modelling of crop performance.

The performance differences observed across models can be directly linked to their structural capacity to represent complexity. Logistic Regression, constrained by its linear formulation, shows limitations in distinguishing between yield categories when variables interact in a non-additive manner. Decision Tree improves this behavior by introducing nonlinear splits, yet its single-structure nature restricts its generalization capacity. Ensemble methods overcome these limitations by aggregating multiple decision structures, allowing them to capture both local and global patterns within the data.

A closer examination of class-level results reveals that high-yield observations are consistently identified with greater precision across all models. This suggests that optimal crop performance is associated with more distinct and stable patterns in the feature space. In contrast, medium-yield conditions exhibit greater classification ambiguity, which can be interpreted as the result of compensatory interactions among variables. These findings are consistent with the understanding that agricultural systems rarely operate under isolated conditions, and that intermediate productivity levels often emerge from the balance between limiting and favorable factors.

The analysis of feature importance introduces a critical insight into the drivers of crop performance. Geographic location emerges as a dominant predictor across all tree-based models, indicating that regional agroecological conditions strongly influence yield outcomes. This observation is consistent with previous studies emphasizing the role of spatial heterogeneity in agricultural productivity [7,14]. However, the present results extend this understanding by showing that geographic effects do not act in isolation but interact with crop type and management practices. The prominence of crop-specific variables, particularly COCO and CACAO, highlights the importance of species-dependent production dynamics. Perennial crops tend to exhibit distinct nutritional and physiological responses, which are reflected in the model's predictive structure. This finding suggests that yield prediction models should account for crop-specific behavior rather than relying solely on generalized predictors.

One of the most relevant contributions of this study lies in the role of foliar nutrient variables. Elements such as sulphur, nitrogen, magnesium, sodium, calcium, potassium, and zinc consistently appear among the most influential predictors. This result contrasts with a significant portion of the existing literature, where predictive models are primarily based on climatic or remote sensing variables [13,19]. While environmental variables provide indirect signals of crop performance, foliar analysis offers a direct representation of plant physiological status, enabling a closer approximation to yield formation processes [7]. This shift toward nutrient-centred modelling addresses a key limitation identified in recent studies. Although machine learning applications in agriculture have advanced considerably, many models remain disconnected from agronomic decision-making. Interpretability is often limited to identifying important variables without translating those findings into actionable strategies. The present results demonstrate that nutrient and management variables can serve as both predictive and decision-relevant factors, bridging the gap between data-driven modelling and practical agricultural applications.

The interaction between fertilization practices and nutrient composition further reinforces this perspective. Fertilization variables influence nutrient availability, which is subsequently reflected in foliar measurements and ultimately determines yield classification. This relationship highlights the importance of integrating management history into predictive models, as it provides context for interpreting plant nutritional status. The ability of ensemble models to capture these interactions

 doi:10.20944/preprints202603.2121.v1

explains their superior performance and supports the view of agricultural systems as integrated and adaptive systems.

Another important aspect of the findings relates to model robustness and applicability. The use of open agricultural data introduces variability that more closely resembles real-world conditions compared to controlled experimental datasets. This characteristic enhances the external relevance of the results and partially addresses the generalization challenges reported in previous studies (Saini et al., 2026). At the same time, the inclusion of geographic and management variables contributes to stabilizing model performance across heterogeneous conditions. The consistency observed in the top predictors across different models strengthens the reliability of the results. The convergence of variables such as geographic location, crop type, and key nutrients suggests that these factors form the core explanatory structure of crop performance. Differences among models are mainly observed in secondary variables, reflecting complementary perspectives rather than contradictory findings. This structural agreement supports the robustness of the modelling approach and increases confidence in its applicability.

These findings highlight that crop yield cannot be adequately explained through isolated variables. Instead, it emerges from the interaction between nutritional status, geographic context, and agronomic management. This perspective aligns with the principles of precision agriculture, where decision-making is based on the integration of multiple data sources and localized conditions.

6. Conclusions and Future Research

This study developed and evaluated a machine learning framework for crop yield classification based on foliar analysis, agronomic management, and geographic variables using open agricultural data from Colombia. The results demonstrate that ensemble-based models, particularly Random Forest and Gradient Boosting, provide superior predictive performance compared to linear and single-tree approaches, confirming their ability to capture the nonlinear and interactive nature of agricultural systems. Beyond predictive accuracy, the findings highlight a structural pattern in the determinants of crop performance. Geographic location, crop type, and foliar nutrient composition consistently emerge as the most influential factors across models. This convergence indicates that crop yield is shaped by the interaction between regional agroecological conditions, species-specific characteristics, and plant nutritional status. The prominence of nutrients such as sulfur, nitrogen, magnesium, calcium, potassium, and zinc reinforces the relevance of foliar analysis as a direct indicator of physiological conditions linked to productivity.

The study contributes to the literature by shifting the focus from climate-driven or image-based prediction toward a nutrient-centered and management-oriented framework. This perspective addresses a key limitation in existing approaches, where predictive models often rely on indirect environmental proxies and provide limited support for agronomic decision-making. By integrating foliar data with fertilization practices, the proposed framework offers a more interpretable and actionable representation of crop performance.

The formulation of yield prediction as a classification problem further enhances the practical applicability of the model. Categorizing performance into discrete levels facilitates its use in real-world contexts, where decision-making is typically based on thresholds rather than continuous estimates. This characteristic strengthens the connection between data-driven modeling and precision agriculture practices.

Despite these contributions, several limitations should be acknowledged. The model was developed using a specific dataset with defined spatial and temporal coverage, which may affect its generalization to other regions or cropping systems. While the inclusion of geographic and management variables improves robustness, further validation in different agroecological contexts is necessary to confirm the transferability of the approach.

Future research can extend this work in several directions. One relevant avenue involves incorporating climatic and remote sensing variables to complement nutrient-based predictors, enabling a more comprehensive representation of crop systems. Another promising direction is the

 doi:10.20944/preprints202603.2121.v1

integration of temporal dynamics, allowing the analysis of yield evolution over time and improving early prediction capabilities. The development of hybrid modelling approaches that combine machine learning with process-based or agronomic models may also enhance interpretability and robustness. In addition, expanding the dataset to include socioeconomic variables and farm-level management decisions could provide a more complete understanding of productivity drivers.

Further work is also needed to strengthen the interpretability of the models by linking feature importance results with specific agronomic recommendations. This includes translating nutrient-related findings into fertilization strategies and decision-support tools that can be directly used by farmers and agricultural planners. Finally, the framework could be extended toward integrated agricultural systems by connecting yield prediction with supply chain and resource management models. Such integration would enable the evaluation of productivity not only at the field level but also within broader agro-industrial and logistical contexts.

References

1. Sierra-Forero BL, Baron-Velandia J, Vanegas-Ayala SC. Assessment of the relevance of features associated with corn crop yield prediction in Colombia, a country in the Neotropical zone. International Journal of Information Technology (Singapore) 2024;16:2129–38. https://doi.org/10.1007/s41870-024-01762-9.
2. Khaki S, Wang L. Crop yield prediction using deep neural networks. Front Plant Sci 2019;10. https://doi.org/10.3389/fpls.2019.00621.
3. Basso B, Liu L. Seasonal crop yield forecast: Methods, applications, and accuracies. Advances in Agronomy, vol. 154, Academic Press Inc.; 2019, p. 201–55. https://doi.org/10.1016/bs.agron.2018.11.002.
4. Hualpa Zúñiga AM, Rangel Díaz JE. Traceability in the agricultural sector: A review for the period 2017-2022. Agronomía Mesoamericana 2022;34:1–19. https://doi.org/10.15517/am.v34i2.51828.
5. Xiang K, Wang B, Liu DL, Chen C, Ji F, Yao S, et al. Future climate change increases the risk of wheat yield loss due to agricultural drought in southeastern Australia. European Journal of Agronomy 2026;173. https://doi.org/10.1016/j.eja.2025.127909.
6. Muñoz Ordoñez CC, Cobos Lozada CaA, Muñoz Ordoñez JF. Predicción del rendimiento de cultivos de café. Ingeniería y Competitividad 2023;23. https://doi.org/10.25100/iyc.v25i313171.
7. Silva BC, Prado RM, Baio FHR, Campos CNS, Teodoro LPR, Teodoro PE, et al. New approach for predicting nitrogen and pigments in maize from hyperspectral data and machine learning models. Res Sq 2022. https://doi.org/10.21203/rs.3.rs-2350350/v1.
8. Meng H, Qian L, Qiu R. Can machine-learning methods better characterize the relationships between crop yields and water disaster intensity at different growth stages? Field Crops Res 2026;336. https://doi.org/10.1016/j.fcr.2025.110231.
9. Lang H, Cui Y, Liu W, Tang S, Gao W, Zhai C, et al. Machine learning reveals molecular substructure drivers of organic contaminant translocation in crops. Science of the Total Environment 2026;1018. https://doi.org/10.1016/j.scitotenv.2026.181514.
10. Lang H, Cui Y, Liu W, Tang S, Gao W, Zhai C, et al. Machine learning reveals molecular substructure drivers of organic contaminant translocation in crops. Science of the Total Environment 2026;1018. https://doi.org/10.1016/j.scitotenv.2026.181514.
11. Kaur A, Randhawa GS, Farooque AA, Ali M, Singh H, Barrett R, et al. Potato crop disease prediction in Prince Edward Island using machine learning: A decision support approach for farmers. Ecol Inform 2026;94:103667. https://doi.org/10.1016/j.ecoinf.2026.103667.
12. Serra E, Debolini M, Fraga H, Trabucco A, Mereu V, Spano D. Machine learning and species distribution models for crops: A review. Ecol Inform 2026;93. https://doi.org/10.1016/j.ecoinf.2025.103563.
13. Saini Y, Somani V, Khatri N. Critical Analysis of Advanced Machine Learning and Deep Learning Models for Crop Disease Detection: Challenges, Innovations, and Future Directions. Franklin Open 2026:100584. https://doi.org/10.1016/j.fraope.2026.100584.
14. Valleggi L, Scutari M, Stefanini FM. Learning Bayesian Networks with Heterogeneous Agronomic Data Sets via Mixed-Effect Models and Hierarchical Clustering. Eng Appl Artif Intell 2024:1–8.

 doi:10.20944/preprints202603.2121.v1

15. de Oliveira Faria R, Filho ACM, Santana LS, Martins MB, Sobrinho RL, Zoz T, et al. Models for predicting coffee yield from chemical characteristics of soil and leaves using machine learning. J Sci Food Agric 2024;104:5197–206. https://doi.org/10.1002/jsfa.13362.
16. Saruk BS, Rayalu GM. Integrating Machine Learning for Enhanced Agricultural Productivity: A Focus on Bananas and Arecanut in the Context of India's Economic Growth. Journal of Statistical Theory and Applications 2024;23:408–26. https://doi.org/10.1007/s44199-024-00090-y.
17. Sai A, Ellafi A, Younes S Ben, Borgi MA. Morphological responses of four crops to heavy metal-rich phosphate processing wastewater: Experimental toxicology and machine learning modeling. Journal of Environmental Sciences 2026. https://doi.org/10.1016/j.jes.2026.03.052.
18. Torres-Herrera PA, Arce-Inga M, Tarrillo E, Ocupa E, Atalaya-Marin N, Cabrera-Hoyos H, et al. Effect of cover crops on soil quality, yield, and prediction using machine learning in papaya (Carica papaya L.). Smart Agricultural Technology 2026;14:101953. https://doi.org/10.1016/j.atech.2026.101953.
19. Sohoulande CDD, Khedun CP. A Framework for Early-Season Crop Yield Prediction Using Machine Learning and Drought Conditions in the Southeastern United States. Farming System 2026:100224. https://doi.org/10.1016/j.farsys.2026.100224.
20. Lisboa BB, Abichequer AD, de São José JFB, Moura-Bueno JM, Brunetto G, Vargas LK. Compositional Nutrient Diagnosis Methodology and Its Effectiveness to Identify Nutrient Levels in Yerba Mate (Ilex paraguariensis). Agriculture (Switzerland) 2024;14. https://doi.org/10.3390/agriculture14060896.
21. Sahu H, Garg PK, Vijay S, Dasgupta A. Predictive drivers and transferability of multi-scale machine learning based crop yield prediction under drought across European and Asian climates. Agric Water Manag 2026;327. https://doi.org/10.1016/j.agwat.2026.110254.
22. Serra E, Debolini M, Fraga H, Trabucco A, Mereu V, Spano D. Machine learning and species distribution models for crops: A review. Ecol Inform 2026;93. https://doi.org/10.1016/j.ecoinf.2025.103563.
23. Xia Z, Wang B, Guo L, Guo X. Estimating the crop coefficient of meadow ecosystems and its driving factors using machine learning on the Qinghai–Tibet Plateau. J Hydrol Reg Stud 2026;64:103204. https://doi.org/10.1016/j.ejrh.2026.103204.