



Article

Enhancing Agricultural Decision-Making: Banana Yield Forecasting in Colombia Using Tuned Ensemble Machine Learning Models

María I. Arrieta-Escobar ¹, Carlos D. Paternina-Arboleda ², Jorge I. Vélez ¹ and Guisselle A. García-Llinás ^{1,*}

¹ Department of Industrial Engineering, Universidad del Norte, Barranquilla 081007, Colombia; imarrieta@uninorte.edu.co (M.I.A.-E.); jvelezv@uninorte.edu.co (J.I.V.)

² Department of Management Information Systems, Fowler College of Business, San Diego State University, San Diego, CA 92182, USA; cpaternina@sdsu.edu

* Correspondence: gagarcia@uninorte.edu.co

Abstract

Accurate short-term forecasts of banana productivity can improve harvest scheduling, packing-capacity allocation, and export logistics, yet commercial forecasts often deviate substantially from realized yields. **Objective:** We evaluated whether tuned ensemble machine learning could reduce that error for seven-week-ahead forecasting of weekly productivity (boxes/Ha). **Methods:** Using weekly records from nine commercial farms in northern Colombia (2007–2018), we benchmarked six tuned ensembles—Random Forest, Extra Trees, Gradient Boosting, LightGBM, XGBoost, and CatBoost—against a company forecast and three statistical baselines (Seasonal Naive, Historical Mean, and a per-farm ARIMA). Hyperparameters were tuned by time-series cross-validation on 2007–2017, and models assessed on a 2018 holdout (405 observations). **Results:** CatBoost obtained the lowest errors (MAE = 3.95, $R^2 = 0.61$), a 57.0% MAE and 81.2% MSE reduction versus the company baseline (MAE = 9.18, $R^2 = -1.06$). Bootstrap 95% CIs and the Diebold–Mariano test showed CatBoost, XGBoost, LightGBM, Gradient Boosting, Random Forest, and the Historical Mean to be statistically equivalent (CatBoost’s lead was not significant), all outperforming the ARIMA (MAE = 5.64), Seasonal Naive (MAE = 7.31), and company baselines; an ablation confirmed no data leakage. **Conclusions:** Tuned ensembles can substantially improve short-term harvest planning in commercial banana production, with model choice guided by operational and computational constraints.

Keywords: banana productivity forecasting; ensemble learning; CatBoost; XGBoost; harvest planning; time-series prediction; precision agriculture; export agriculture



Academic Editor: Chrysanthos Maraveas

Received: 25 May 2026

Revised: 16 June 2026

Accepted: 1 July 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

Bananas (*Musa* spp.) are the most traded fresh fruit worldwide, and the sector continues to grow even as it faces weather variability, the spread of Fusarium Tropical Race 4 (TR4), and rising logistics costs that tighten quality tolerances and delivery windows [1–3]. Colombia is one of the leading exporters, with production concentrated in the Urabá and Magdalena regions and oriented toward the North American and European markets [2]. In this context, the operational challenge is not only how much will be produced but also when harvest-ready fruit will be available and how that timing aligns with packing capacity, transport bookings, and destination ripening programs [4–6].

For bananas, harvest timing is closely linked to fruit physiology and postharvest behavior. A central operational concept is the fruit's green life—the time span during which bananas remain green and suitable for shipping before ripening is triggered—which is affected by maturity at harvest, handling, and temperature management [4]. Postharvest losses during transportation and marketing are estimated at 25–40% worldwide, and can exceed 50% in certain rural markets, primarily due to physiological changes, flesh softening, and reduced resistance to microbial attack [7]. Because export programs depend on predictable ripening outcomes at destination, cutting/harvest schedules act as an upstream lever for managing green life, shelf-life, and rejection risk downstream [4,5]. Variability in harvest timing can therefore propagate into logistical inefficiencies (idle capacity or bottlenecks), inconsistent fruit condition, and avoidable losses, particularly when peaks or disruptions coincide with limited cold-chain and transport capacity [1,4–6].

From a physiological standpoint, the seven-week forecast horizon adopted in this study corresponds to a critical window in the banana production cycle. In commercial *Cavendish* plantations under tropical conditions, fruit development from bunch emergence to harvest spans approximately 11–13 weeks, during which dry matter accumulates exponentially in the pulp [8]. Harvest timing is commonly determined using thermal sums (accumulated degree-days since bunch emergence), with commercial maturity reached at approximately 900 degree-days [9]. Importantly, the rate of fruit growth is strongly modulated by environmental conditions: soil water deficit can reduce fruit fresh-weight accumulation by up to 39%, and temperature governs the overall pace of development [8]. A seven-week-ahead forecast therefore targets the window in which the agronomic trajectory of the standing crop is largely determined by existing bunch-age distributions—captured in the dataset by the Pct 8 through Pct 13 weeks variables—yet still provides sufficient lead time for operational adjustments to packing, labor, and logistics scheduling. This physiological grounding complements the logistical rationale and explains why bunch-age features carry meaningful predictive information for harvest volume.

These decisions sit within a broader planning context where agrifood supply chains must coordinate biological production processes with industrial constraints—labor, packing lines, storage, trucking, and shipping slots—under uncertainty. Planning research consistently shows that perishable systems benefit from integrating forecasts with scheduling rules because errors in timing, not only volume, can degrade quality and increase cost [6]. In the banana case, a short-term forecast that anticipates fluctuations in cutting dynamics can help align labor and maintenance plans, allocate packing capacity, and schedule shipments to match target arrival conditions. The growing adoption of data-driven decision tools in the banana sector [1,3] underscores the industry's appetite for quantitative forecasting methods that go beyond traditional expert judgment.

From a modeling perspective, time-series forecasting provides a pragmatic route for supporting these decisions when historical operational records exist and explanatory drivers are incomplete or difficult to measure consistently. Classical approaches such as ARIMA remain widely used and interpretable baselines because they are transparent, data-efficient, and often competitive at short horizons when autocorrelation structures are stable [10,11]. At the same time, methodological work has shown that performance gains can be overstated when evaluation protocols allow data leakage or do not reflect how forecasts are generated under rolling, real-time use; therefore, careful temporal validation is essential in operational settings [12]. A related concern is comparability: model families should be benchmarked against appropriate baselines and evaluated under consistent protocols, ideally with formal tests of predictive accuracy differences such as the Diebold–Mariano test [13]. Evidence from the M3 and M4 forecasting competitions has shown that, on monthly time series, traditional statistical methods (Theta, ARIMA, exponential

smoothing combinations) often outperform more complex machine learning approaches, and that computational cost does not necessarily translate into improved accuracy [14–16].

Nonetheless, over the past decade, ensemble machine learning methods have emerged as the dominant approach for structured, tabular prediction tasks where cross-sectional features are abundant. Random Forests, introduced by Breiman [17], reduce variance by averaging decorrelated decision trees. Gradient boosting methods sequentially correct residual errors, with several high-performance implementations now available: XGBoost [18] introduced regularized boosting with column subsampling; LightGBM [19] improved training efficiency through histogram-based learning and leaf-wise tree growth; and CatBoost [20] natively handles categorical features through ordered target statistics and oblivious decision trees. A recent study comparing these three gradient boosting algorithms for crop yield prediction found that all three outperformed other machine learning methods, with CatBoost achieving the highest accuracy in some contexts [21]. More recent evidence reinforces this pattern: Ennaji et al. [22] reported XGBoost as the most accurate among eight algorithms for maize-hybrid yield prediction. These methods have consistently dominated benchmarks on structured data, although their advantage over statistical baselines depends on the availability of informative cross-sectional features [14–16].

In agriculture, machine learning has been increasingly applied to crop yield prediction and decision-support tasks [23,24]. A systematic review by van Klompenburg et al. [25] identified artificial neural networks as the most commonly used algorithm family, with temperature, rainfall, and soil type as the most frequent input features; tree-based ensembles were also well represented but not the majority. More recent comprehensive reviews confirm the growing adoption of gradient boosting methods for yield forecasting, though they emphasize that reported gains depend strongly on data quality, competitive baselines, and validation designs that respect temporal structure [26–28]. Beyond single-model comparisons, recent work has explored ensemble combination strategies for yield prediction—for example, inverse-RMSE-weighted hybrids that adaptively blend base learners [29]—reflecting sustained interest in ensemble learning for agricultural monitoring. A persistent limitation is the lack of robust temporal validation: Srivastava et al. [30] found that convolutional neural networks, deep neural networks, and XGBoost all performed well when evaluated with spatial cross-validation on winter wheat in Germany, but performance degraded substantially when tested on temporally independent years, highlighting the risk of overstating generalization. This finding underscores the importance of strict temporal holdout designs like the one adopted in the present study.

Despite extensive work on banana postharvest behavior and green-life management [4,5,7], and a growing body of forecasting research in other crops [25–27,30], studies specifically targeting banana yield or cutting-time forecasting with machine learning remain scarce. Patrick et al. [31] used SARIMAX, state-space, and LSTM models to forecast banana crop yield in Tanzania under climate change scenarios, employing a weighted-average ensemble of these methods. Yáñez-Cajo et al. [32] applied Random Forest with UAV imagery and spectral indices for banana yield prediction in Ecuador. Ndayambaje and Ngango [33] compared Random Forest, SVR, and linear regression for banana yield prediction in Rwanda using meteorological and agricultural data, and Zhou et al. [34] estimated banana-orchard yield from RGB–depth imagery using a DeepSORT-based bunch-counting pipeline. However, none of these studies benchmarked the full suite of modern gradient boosting methods (XGBoost, LightGBM, CatBoost) against each other and against an operational company baseline under a strict temporal validation protocol. Furthermore, no prior study has applied formal statistical comparison tests—such as the Diebold–Mariano test [13]—to determine whether observed differences between competing models are statistically significant or merely artifacts of a single test set.

This study addresses these gaps by modeling banana productivity seven weeks ahead using historical records from nine commercial farms in northern Colombia (2007–2017). We evaluate two families of tree-based ensemble methods—bagging (Random Forest, Extra Trees) and boosting (Gradient Boosting, LightGBM, XGBoost, CatBoost)—against the company’s existing forecast baseline under a consistent temporal validation protocol using the 2018 holdout period. The choice of tree-based ensembles is motivated by the nature of the available data: the feature set comprises 15 production-related numerical variables and a categorical farm identifier, a regime in which gradient boosting methods have shown strong empirical performance [18–21] and where the limited sample size (~614 observations per farm) makes deep learning approaches less suitable [27]. Hyperparameters are tuned through time-series cross-validation with GridSearchCV, and accuracy is assessed using six complementary forecast-error measures. To strengthen inferential rigor, we report bootstrap 95% confidence intervals for all metrics and apply the Diebold–Mariano test to formally compare predictive accuracy between models [13]. The analysis also includes per-farm evaluation to reveal heterogeneity in model performance across production units, and feature importance analysis to identify the variables most relevant for forecast accuracy. The contribution of this work is methodological and applied rather than algorithmic: to our knowledge it is the first benchmark of the full modern gradient-boosting suite for banana yield under a strict temporal hold-out with formal Diebold–Mariano testing and bootstrap intervals; it quantifies the cross-farm heterogeneity that aggregate results conceal; and it translates forecast error into concrete export-logistics decisions, linking model accuracy to operational value. In addition to the company forecast, we benchmark the ensembles against three statistical baselines—a Seasonal Naive forecast, a Historical Mean, and a per-farm ARIMA model—so that the contribution of the machine learning methods can be assessed relative to both naive and classical time-series references. We acknowledge that the present study does not include neural network architectures (e.g., LSTM, CNN) or exponential-smoothing variants as additional comparators; the integration of such approaches, as well as the incorporation of external climatic and agronomic variables, is discussed as future work in Section 5.

2. Materials and Methods

2.1. Data Source, Processing, and Software

The data used for the computational experiments correspond to banana production between 2007 and 2018 from nine banana farms located in the northern zone of Colombia. The farms are situated in the banana-producing belt of the Magdalena department, on the Caribbean lowlands between the Sierra Nevada de Santa Marta and the Ciénaga Grande, an area characterized by a warm tropical climate and export-oriented *Cavendish* production (Figure 1). The data were provided under a strict confidentiality agreement and are available upon request to the authors; to preserve confidentiality, individual farm locations are not disclosed and only the study region is shown. To support reproducibility, a desensitized representative sample of these records is provided as Supplementary Materials (Dataset S1).

Forecasts were made seven weeks before the harvest date. This decision was driven by several factors that present significant challenges in the banana supply chain, including the perishable nature of the fruit, technical conservation requirements, transportation and packing conditions, and the long distance of travel from farms to ports. The target variable was constructed by shifting the observed productivity per hectare (Prod_real_Ha) seven weeks forward within each farm using a grouped shift operation, ensuring no data leakage across farm boundaries. The company’s forecast was similarly shifted to serve as the baseline comparator. The predictor variables contained no missing values over the study period, so no imputation was necessary. The only rows discarded were the last seven

weeks of each farm, for which the seven-week-ahead target is undefined. After that step, the dataset was split into a training period (years 2007–2017, 5166 observations) and a validation period (year 2018, 405 observations across the nine farms). Table 1 describes the variables available in the dataset, and Table 2 presents the derived variables used in the modeling process. All data processing, model training, and evaluation were carried out in Python (Python Software Foundation, Wilmington, DE, USA) using the pandas, scikit-learn, XGBoost, LightGBM, CatBoost, statsmodels, and pmdarima libraries.

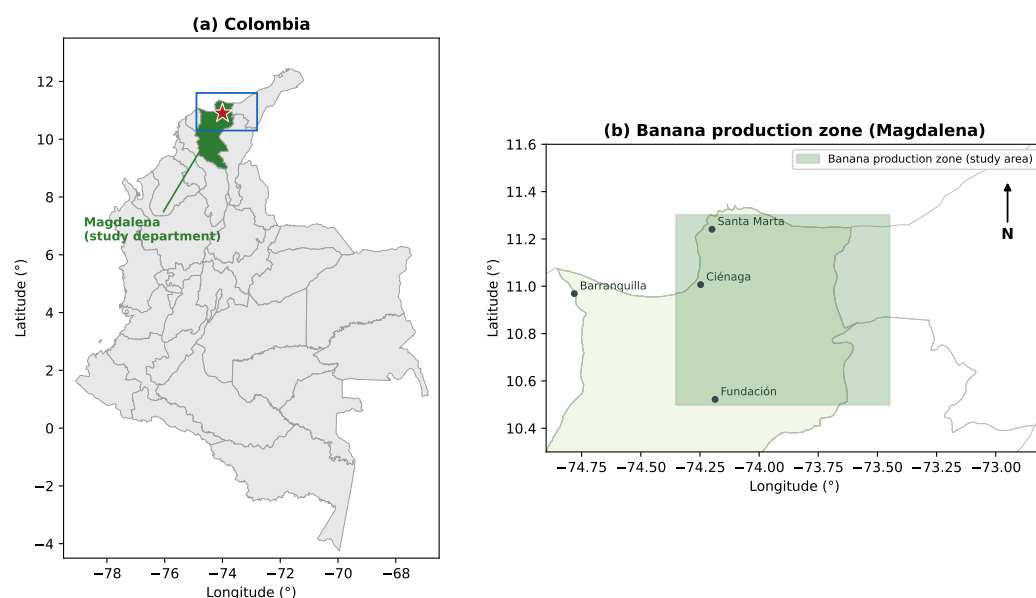


Figure 1. Study area. (a) Location of the Magdalena department within Colombia (the red star marks the studied banana-production zone and the blue rectangle delimits the area enlarged in panel (b)); (b) banana production zone in the Caribbean lowlands of Magdalena (shaded), with reference cities. Exact farm locations are withheld under a confidentiality agreement.

At the time a seven-week-ahead forecast is operationally needed, the current week's actual production per hectare (Prod_real_Ha) is already known, as it reflects harvest counts completed and recorded during that week. This variable is therefore a legitimate contemporaneous feature, not a future-looking variable. To verify this, we conducted an ablation study removing Prod_real_Ha from the feature set; the resulting XGBoost MAE changed by less than 0.01 boxes/Ha (from 4.030 to 4.025, a difference of -0.005), confirming that the model did not depend on that variable for predictive accuracy.

Python 3 was used for all computational experiments. The code made use of pandas for data manipulation, numpy for numerical computations, scikit-learn for preprocessing (StandardScaler), cross-validation (TimeSeriesSplit, GridSearchCV), and performance metrics, plus lightgbm, xgboost, and catboost for their respective gradient boosting implementations. All code and results are available upon request.

Table 1. Description of original variables in the dataset.

Variable	Description	Unit
Farm	Farm identifier (1–9)	Categorical
Week	Calendar week number	Integer
Hectares	Productive area	Ha
Actual Packing	Boxes packed in the current week	Boxes
Pct 8–13 weeks	Percentage of bunches aged 8–13 weeks	%
Company Forecast	Company's internal yield forecast	Boxes/Ha

Table 1. *Cont.*

Variable	Description	Unit
Pct Short	Percentage of short-grade fruit	%
Pct Premium	Percentage of premium-grade fruit	%
Boxes Short	Volume of short-grade boxes	Boxes
Boxes Premium	Volume of premium-grade boxes	Boxes
Prod_real_Ha	Actual production per hectare	Boxes/Ha

Table 2. Derived variables used in the modeling process.

Variable	Description
Target (y_{t+7})	Prod_real_Ha shifted 7 weeks forward within each farm
Baseline forecast	Company forecast shifted 7 weeks forward within each farm

2.2. Forecasting Models

Six ensemble machine learning models were trained and evaluated against the company's existing forecast baseline. All models used the same set of 15 numerical features plus a farm identifier: Week, Hectares, Actual Packing, six bunch-age percentages (Pct 8 through Pct 13 weeks), Company Forecast, Short Percentage, Premium Percentage, Short Boxes, Premium Boxes, and Actual Production per Hectare. Numerical features were standardized using a StandardScaler fit on the training set and applied to both training and validation sets.

The models evaluated were: (1) Random Forest Regressor, an ensemble of decorrelated decision trees that averages predictions to reduce variance [17]; (2) Extra Trees Regressor, which uses random split thresholds for additional diversity; (3) Gradient Boosting Regressor, which sequentially builds shallow trees that correct residual errors [35]; (4) LightGBM, an efficient gradient boosting framework using histogram-based learning and leaf-wise tree growth [19]; (5) XGBoost, a regularized gradient boosting method with column subsampling [18]; and (6) CatBoost, which natively handles categorical features through ordered target statistics and uses oblivious decision trees [20].

Three additional statistical baselines were included for comparison: (1) a Seasonal Naive forecast, which repeats the production observed at the same farm 52 weeks earlier (y_{t-52}); (2) a Historical Mean forecast, computed as the average target value for each farm-week combination across the training period (2007–2017); and (3) a per-farm ARIMA model. For each farm, a non-seasonal ARIMA order was selected automatically by minimizing the AIC over the pre-2018 series (auto_arima, maximum order 5) [11]. Forecasts were then generated under a genuine seven-week-ahead, rolling-origin scheme: at each 2018 origin the model was refit on all observations available up to that week (including already-realized 2018 weeks), and the seventh-step-ahead forecast was taken as the prediction. This ensured that ARIMA addressed exactly the same forecasting task as the ensemble models. These baselines provided reference points between the company forecast and the tuned machine learning models, allowing a more nuanced assessment of where the ensemble methods' predictive gains originated.

2.3. Hyperparameter Tuning

Hyperparameters for each model were tuned using GridSearchCV with a 3-fold TimeSeriesSplit to respect chronological ordering. The scoring criterion was the negative mean absolute error. All models, including CatBoost, used RMSE as the training loss function to ensure a consistent optimization objective across methods. Table 3 presents the hyperparameter grids explored for each model, and Table 4 reports the best hyperparameters selected via cross-validation along with the corresponding CV MAE. For CatBoost, the learning

rate was fixed at 0.03, a value adopted from the library's documentation and preliminary runs, while tree depth and the number of iterations were tuned. The search budget was deliberately concentrated on depth and iterations, which most directly govern model capacity, keeping the number of CatBoost configurations comparable in computational cost to the grids of the other models. Depth and iterations were therefore prioritized over an additional sweep of the learning rate.

Table 3. Hyperparameter grids explored for each model.

Model	Hyperparameter Grid
Random Forest	n_estimators: {100, 200} max_depth: {10, 20, None} max_features: {sqrt, log2, None}
Extra Trees	n_estimators: {100, 200} max_depth: {10, 20, None} max_features: {sqrt, log2, None}
Gradient Boosting	n_estimators: {100, 200} max_depth: {3, 5} learning_rate: {0.05, 0.1, 0.2}
LightGBM	n_estimators: {100, 200} max_depth: {5, 10} learning_rate: {0.05, 0.1, 0.2}
XGBoost	n_estimators: {100, 200} max_depth: {3, 4} learning_rate: {0.05, 0.1, 0.2}
CatBoost	iterations: {300, 500} depth: {4, 6} learning_rate: {0.03}

Table 4. Best hyperparameters selected via time-series cross-validation and corresponding CV MAE.

Model	Best Hyperparameters	CV MAE
Random Forest	max_depth=20, max_features=sqrt, n_estimators=200	6.66
Extra Trees	max_depth=None, max_features=sqrt, n_estimators=200	6.87
Gradient Boosting	learning_rate=0.1, max_depth=3, n_estimators=100	6.52
LightGBM	learning_rate=0.05, max_depth=10, n_estimators=200	6.73
XGBoost	learning_rate=0.05, max_depth=4, n_estimators=200	6.54
CatBoost	iterations=500, depth=6, learning_rate=0.03	6.48

2.4. Performance Measures

To assess predictive capability, we selected complementary metrics evaluating distinct aspects of forecast accuracy. The Explained Variance Score (EVS) and Coefficient of Determination (R^2) quantify the proportion of variability captured by the models [11,35]. Mean absolute error (MAE) and root-mean-square error (RMSE) evaluate error magnitude: MAE provides an intuitive measure of the average error in boxes per hectare, while RMSE penalizes larger deviations more severely [11]. Mean squared logarithmic error (MSLE) assesses relative errors across farms with varying production scales, ensuring evaluation is not biased towards farms with higher yields.

3. Results

The main results are presented in Table 5. The company baseline exhibited the worst performance across all metrics (MAE = 9.18, RMSE = 11.62, $R^2 = -1.06$), with negative R^2 and EVS values indicating predictions worse than simply using the mean of the observed values. The Seasonal Naive baseline also showed poor performance (MAE = 7.31, $R^2 = -0.32$). The per-farm ARIMA baseline improved on both (MAE = 5.64, $R^2 = 0.22$) but remained well above the Historical Mean and all ensemble models; its 95% bootstrap CI for MAE ([5.21, 6.07]) lies entirely above that of every ensemble (whose upper CI bounds reach at most 4.79, for Extra Trees), indicating a statistically meaningful gap without the need for a formal pairwise test. All six tuned ensemble models substantially outperformed the company baseline, the Seasonal Naive forecast, and the ARIMA baseline.

Table 5. Performance metrics for the forecasting models on the 2018 external validation set (405 observations). Best values in bold.

Model	MAE	RMSE	MSE	R^2	EVS	MSLE
Company Baseline	9.18	11.62	135.06	−1.06	−1.05	0.05
Seasonal Naive	7.31	9.32	86.83	−0.32	−0.32	0.03
ARIMA (per farm)	5.64	7.15	51.18	0.22	0.27	0.02
Historical Mean	4.28	5.35	28.58	0.56	0.66	0.01
Random Forest	4.27	5.37	28.84	0.56	0.58	0.01
Extra Trees	4.43	5.67	32.12	0.51	0.53	0.01
Gradient Boosting	4.12	5.17	26.75	0.59	0.63	0.01
LightGBM	4.03	5.09	25.89	0.61	0.62	0.01
XGBoost	4.03	5.07	25.71	0.61	0.64	0.01
CatBoost	3.95	5.04	25.39	0.61	0.65	0.01

MAE: mean absolute error; RMSE: root-mean-square error; MSE: mean squared error; R^2 : coefficient of determination; EVS: explained variance score; MSLE: mean squared logarithmic error. Units: MAE and RMSE in boxes/Ha; MSE in (boxes/Ha)²; R^2 , EVS, and MSLE are dimensionless. Best model: CatBoost (lowest MAE).

CatBoost achieved the lowest errors (MAE = 3.95, RMSE = 5.04, $R^2 = 0.61$, EVS = 0.65, MSLE = 0.01), followed closely by XGBoost (MAE = 4.03, $R^2 = 0.61$) and LightGBM (MAE = 4.03, $R^2 = 0.61$); as the statistical tests below show, these differences among the leading models are not significant. All ensemble models reduced the MAE by more than 51% compared to the company baseline, with CatBoost achieving a 57.0% reduction. The MSE reductions were similarly pronounced: CatBoost reduced MSE by 81.2% (from 135.06 to 25.39), confirming that the ensemble models not only produced smaller average errors but also substantially reduced large forecast deviations that are particularly costly for logistics planning. The Seasonal Naive baseline (MAE = 7.31) performed substantially worse than all ML models, while the Historical Mean baseline (MAE = 4.28) was competitive with the simpler ensemble methods, underscoring the importance of benchmarking against appropriate statistical references.

To assess statistical robustness, we computed 95% bootstrap confidence intervals (CIs) based on $B = 10,000$ resamples for all metrics (Table 6; Figure 2) and applied the Diebold–Mariano (DM) test [13] using squared-error loss differentials to formally compare each model’s predictive accuracy against XGBoost as the reference model (Table 7).

The bootstrap confidence intervals (Table 6) confirm that the performance differences between the ensemble models and the company baseline are substantial and not attributable to sampling variability. The DM test results (Table 7) show that XGBoost is statistically significantly better than the company baseline (DM = −5.73, $p < 0.0001$), the Seasonal Naive forecast (DM = −4.90, $p < 0.0001$), and Extra Trees (DM = −2.77, $p = 0.006$) but not statistically distinguishable from CatBoost (DM = 0.45, $p = 0.652$), LightGBM ($p = 0.906$),

Gradient Boosting ($p = 0.272$), Random Forest ($p = 0.053$), or the Historical Mean baseline ($p = 0.229$) at the 5% significance level. This result is consistent with the overlapping bootstrap CIs between CatBoost (MAE: [3.65, 4.26]) and XGBoost (MAE: [3.74, 4.34]), suggesting these models offer statistically equivalent predictive performance.

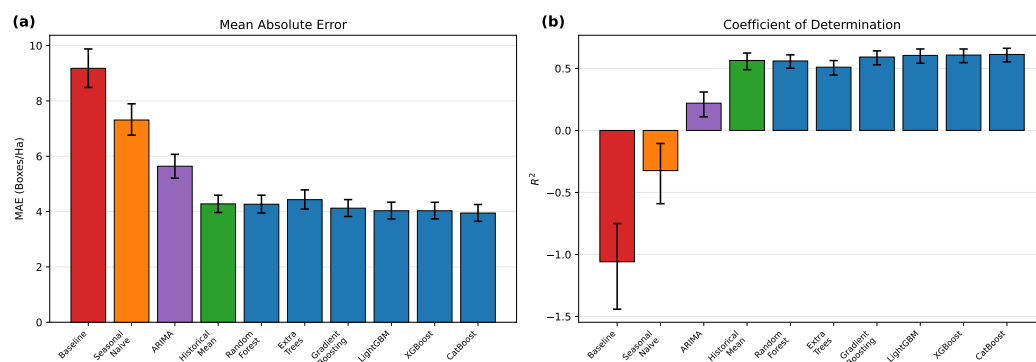


Figure 2. Model performance with 95% bootstrap CIs (error bars) for the 2018 validation set based on $B = 10,000$ resamples. Panel (a) shows MAE and panel (b) shows R^2 .

Table 6. Performance metrics with 95% bootstrap-based CI with $B = 10,000$ resamples.

Model	MAE [95% CI]	RMSE [95% CI]	R^2 [95% CI]
Company Baseline	9.18 [8.49, 9.88]	11.62 [10.74, 12.53]	−1.06 [−1.44, −0.75]
Seasonal Naive	7.31 [6.76, 7.90]	9.32 [8.64, 10.01]	−0.32 [−0.59, −0.11]
ARIMA (per farm)	5.64 [5.21, 6.07]	7.15 [6.67, 7.62]	0.22 [0.11, 0.31]
Historical Mean	4.28 [3.97, 4.59]	5.35 [4.99, 5.72]	0.56 [0.49, 0.62]
Random Forest	4.27 [3.95, 4.59]	5.37 [5.01, 5.73]	0.56 [0.50, 0.61]
Extra Trees	4.43 [4.09, 4.79]	5.67 [5.27, 6.08]	0.51 [0.45, 0.56]
Gradient Boosting	4.12 [3.82, 4.43]	5.17 [4.81, 5.55]	0.59 [0.53, 0.64]
LightGBM	4.03 [3.73, 4.34]	5.09 [4.72, 5.47]	0.61 [0.54, 0.66]
XGBoost	4.03 [3.74, 4.34]	5.07 [4.70, 5.47]	0.61 [0.55, 0.66]
CatBoost	3.95 [3.65, 4.26]	5.04 [4.66, 5.43]	0.61 [0.55, 0.66]

Table 7. Diebold–Mariano (DM) test for equal predictive accuracy (reference model: XGBoost, $h = 7$).

Model	DM Statistic	p -Value	Significance
Company Baseline	−5.73	<0.0001	Yes ***
Seasonal Naive	−4.90	<0.0001	Yes ***
Historical Mean	−1.20	0.229	No
Random Forest	−1.93	0.053	No
Extra Trees	−2.77	0.006	Yes **
Gradient Boosting	−1.10	0.272	No
LightGBM	−0.12	0.906	No
CatBoost	0.45	0.652	No

Negative DM statistic indicates XGBoost has smaller squared-error loss differentials. p -values are reported to three decimal places because rounding to two would obscure their proximity to the 0.05 threshold (e.g., Random Forest, $p = 0.053$). *** $p < 0.001$, ** $p < 0.01$.

Among the gradient boosting variants, CatBoost achieved the lowest MAE on the external test set (3.95), closely followed by XGBoost (4.03) and LightGBM (4.03), though none of the pairwise differences were statistically significant. CatBoost also achieved the best cross-validation MAE during training (6.48; Table 4), suggesting consistent generalization from training to test. Notably, even the Historical Mean baseline—a simple average of historical farm-week production—was not statistically distinguishable from XGBoost ($p = 0.229$), highlighting that much of the predictive signal in this dataset arises from stable

farm-level and seasonal patterns. The implications of this equivalence between a classical statistical baseline and the tuned ensembles are examined in Section 4.

To verify that the inclusion of contemporaneous production (Prod_real_Ha) did not introduce data leakage, we conducted an ablation study in which that feature was removed from the XGBoost model. The resulting MAE changed from 4.030 to 4.025 (a difference of -0.005 boxes/Ha), confirming that the model did not rely on that variable for its predictive accuracy and that no information from the forecast horizon leaked into the training features.

Table 8 presents the per-farm performance of XGBoost compared to the company baseline, revealing substantial heterogeneity across farms. The per-farm R^2 values for all models are visualized in Figure 3. Since CatBoost and XGBoost are statistically equivalent ($DM = 0.45$, $p = 0.652$), XGBoost is used here as a representative of the best-performing model group.

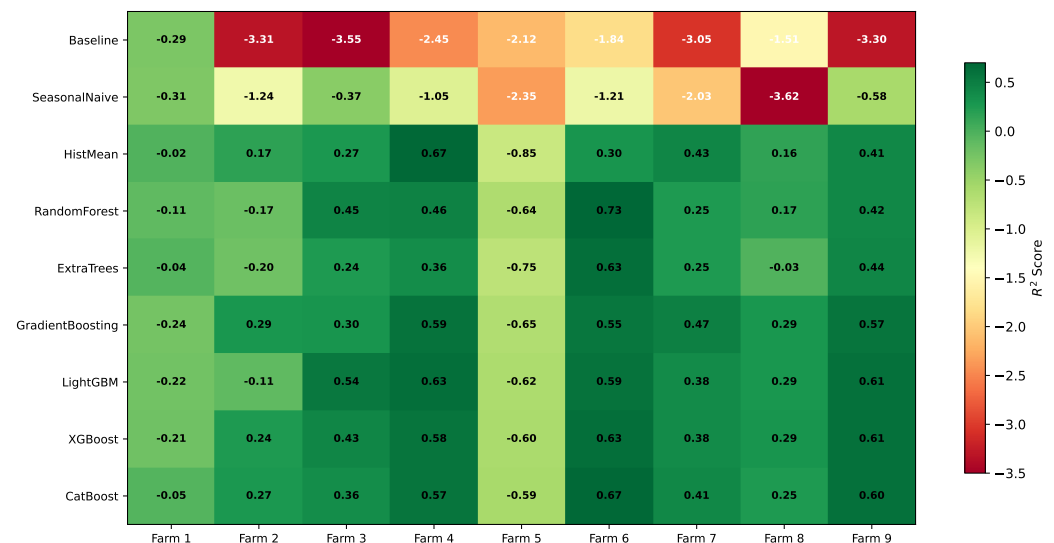


Figure 3. R^2 score heatmap by farm and model on the 2018 validation period. Darker green shading indicates higher explained variance.

Table 8. Performance metrics of the XGBoost and Company Baseline models, by farm, for the 2018 validation set.

Farm	XGBoost			Company Baseline		
	MAE	RMSE	R^2	MAE	RMSE	R^2
1	4.91	6.15	−0.22	4.62	6.34	−0.29
2	3.39	4.50	0.25	9.09	10.75	−3.31
3	4.87	5.57	0.43	12.62	15.80	−3.55
4	4.85	5.82	0.58	13.52	16.64	−2.46
5	5.63	6.62	−0.60	7.78	9.25	−2.12
6	3.05	3.66	0.63	8.32	10.10	−1.84
7	3.35	4.35	0.38	9.43	11.12	−3.05
8	3.04	4.26	0.29	6.49	8.05	−1.51
9	3.18	3.76	0.61	10.74	12.54	−3.30

MAE and RMSE are expressed in boxes/Ha; R^2 is dimensionless.

XGBoost reduced the MAE in eight of the nine farms (all except Farm 1). The observed and predicted weekly series by farm are shown in Figure 4, and the per-farm MAE with 95% bootstrap CIs in Figure 5. The most dramatic improvements occurred in farms with high baseline errors: Farm 3 (MAE reduced from 12.62 to 4.87, a 61.4% improvement), Farm 4 (MAE from 13.52 to 4.85, 64.1%), and Farm 9 (MAE from 10.74 to 3.18, 70.4%). Farm 5 was the most challenging for all models, with XGBoost achieving a negative R^2

(−0.60), indicating that high production variability exceeded the model’s capacity with available features. Farm 1 was the only case where XGBoost showed a higher MAE than the baseline (4.91 vs. 4.62), although both models produced negative R^2 values, suggesting that production variability in that farm was difficult to capture with the current feature set.

The observed heterogeneity in model performance across the nine farms suggests that local environmental factors and management practices significantly influence the predictability of banana yields. The negative R^2 values in Farms 1 and 5 indicate that the current feature set—primarily based on bunch-age distributions and historical production volumes—is insufficient to capture the high variance in these specific production units. Research on global banana yields has shown that the climate–productivity relationship varies substantially across regions, and that relying on a single model without accounting for local conditions can generate considerable prediction error [36]. In the northern Colombian banana zone, Black Sigatoka (*Mycosphaerella fijiensis*) is the principal phytosanitary constraint, and its severity has been documented to vary across farms within the Magdalena region depending on local soil and climatic conditions [37]. These unobserved sources of variability—microclimate, disease pressure, and soil heterogeneity—are not captured in the current feature set, which may explain the poor model performance in specific farms. Incorporating farm-level climatic records and disease-incidence indicators could improve forecast accuracy in these challenging production units, as demonstrated by recent studies showing that meteorological variables substantially enhance banana yield prediction [33]. A further, emerging phytosanitary threat is *Fusarium oxysporum* f. sp. *cubense* Tropical Race 4 (TR4), detected in Colombia’s La Guajira region in 2019, after our dataset ends. Any operational deployment of these models would therefore need to account for potential TR4-driven yield disruptions; adaptive models able to detect disease-related structural breaks in production patterns, possibly coupled with disruption-aware logistics simulation [38], are a promising direction for handling such events.

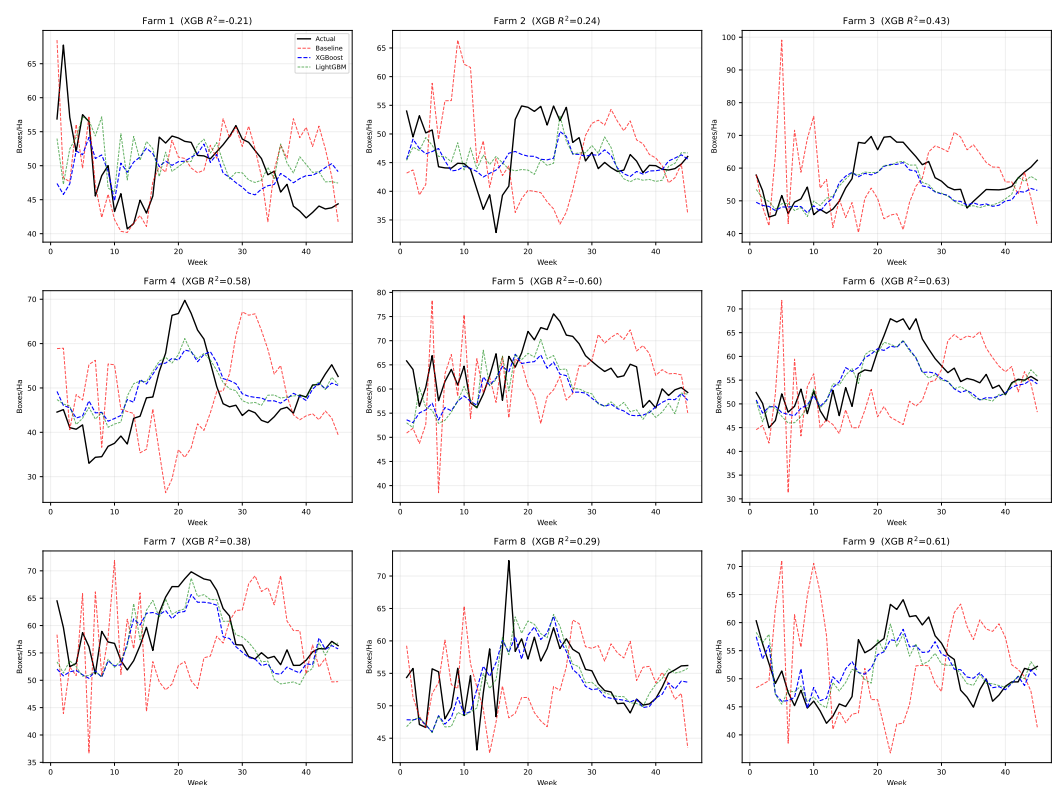


Figure 4. Yield predictions by farm (boxes/Ha) on the 2018 validation period. Solid lines show observed values; dashed lines show XGBoost predictions.

Feature importance analysis, shown in Figure 6, revealed that the farm identifier was the most influential variable (17.6% gain importance in XGBoost), reflecting substantial between-farm heterogeneity in production levels that the model leverages through farm-specific learned patterns. The importance of Week (11.4%) reflects the well-known seasonality in banana production cycles. Among the production-related features, Boxes Premium (8.5%), Prod_real_Ha (8.3%), and Actual Packing (8.3%) ranked third through fifth. It is important to note that features such as Company Forecast and Prod_real_Ha correspond to values observed at the time the forecast is issued (i.e., the current week), not at the forecast horizon. The Company Forecast reflects the company's own internal projection available at that point, and Prod_real_Ha represents the actual production already recorded during the current week. These are therefore legitimate contemporaneous inputs, not future-looking variables. As confirmed by the ablation study, removing Prod_real_Ha changes the XGBoost MAE by only -0.005 boxes/Ha, indicating that the model does not depend on this variable. These importance values should be interpreted with caution: they reflect the variables that are most useful for prediction within this particular dataset and seven-week horizon and may not generalize to other contexts or time scales.

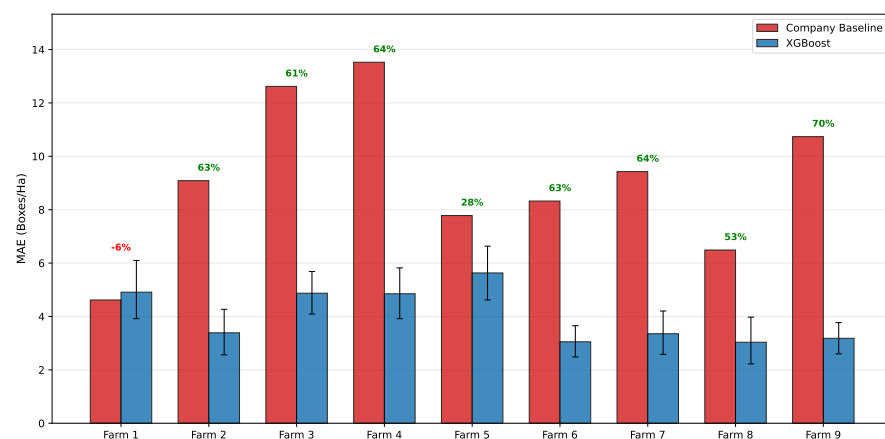


Figure 5. MAE for the XGBoost and Company Baseline by farm. Vertical segments correspond to 95% bootstrap-based CIs.

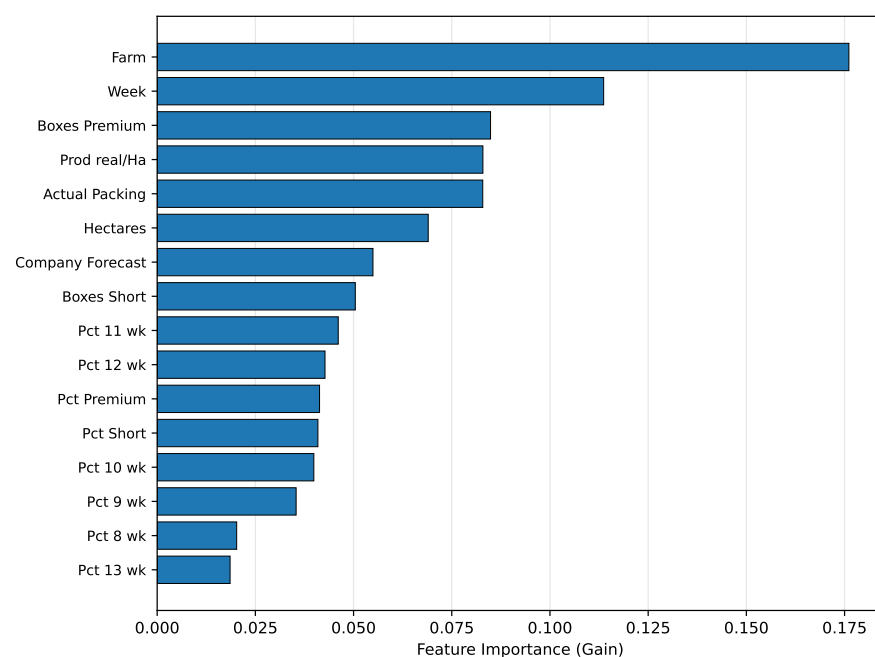


Figure 6. XGBoost feature gain-based importance ranking.

4. Discussion

The performance evaluation of six tuned ensemble machine learning models for banana productivity forecasting in Colombia's northern Caribbean region demonstrates that data-driven methods can substantially improve upon traditional expert-based forecasts. All ensemble models outperformed the company's existing forecast, with CatBoost achieving the best overall performance ($MAE = 3.95$, $R^2 = 0.61$), though statistically equivalent to XGBoost, LightGBM, Gradient Boosting, Random Forest, and the Historical Mean baseline.

The company baseline's negative R^2 and EVS values (-1.06 and -1.05 , respectively) indicate that the current forecasting approach introduces more error than a simple mean prediction would. We did not have access to the exact methodology used by the company. Still, the magnitude of the errors, together with the systematic pattern of overestimation on some farms and underestimation on others, suggests that the current approach relies on expert judgment or simple heuristics that miss the temporal patterns in productivity data. The ensemble models, by contrast, achieved positive R^2 values ranging from 0.51 (Extra Trees) to 0.61 (CatBoost), demonstrating meaningful predictive power. The bootstrap confidence intervals (Table 6) confirm that these improvements are robust and not attributable to sampling variability in the test set.

Among the models evaluated, the gradient boosting family (CatBoost, XGBoost, LightGBM) generally outperformed the bagging-based methods (Random Forest, Extra Trees), although the Diebold–Mariano test showed that the difference between XGBoost and Random Forest was not statistically significant ($p = 0.053$). This nuanced finding contrasts with the common assumption that boosting always dominates bagging on structured data and suggests that for this particular dataset with moderate sample size and limited feature complexity, the simpler Random Forest model can achieve competitive results. The DM test confirmed that CatBoost and XGBoost were statistically equivalent ($DM = 0.45$, $p = 0.652$) and that LightGBM was also indistinguishable from XGBoost ($p = 0.906$), meaning all three gradient boosting frameworks are viable choices for operational deployment. This finding is practically important because LightGBM offers advantages in training speed for larger datasets, and CatBoost provides native categorical feature handling.

The statistical equivalence between the Historical Mean and the tuned ensemble models ($p = 0.229$) warrants careful interpretation. This result reveals that banana productivity in the studied region follows a strong seasonal baseline, a finding that is agronomically expected, given that fruit development in *Cavendish* plantations is governed by temperature-driven phenological cycles that repeat with relative regularity across years [8,9]. However, the operational superiority of machine learning lies in its responsiveness to non-linear disruptions. Unlike a static historical average, gradient boosting models can dynamically adjust to shifts in bunch-age distributions (Pct 8–13 weeks) caused by localized weather events or management changes. While a historical mean assumes that production will follow its typical farm-week pattern, ensemble models integrate dynamic features—such as the current distribution of bunch ages and real-time packing volumes—to adjust forecasts when physiological development or operational conditions deviate from the norm. For agricultural exporters, the ability to anticipate a deviation from the historical norm is often more valuable than achieving a low average error, as these deviations are what trigger logistical bottlenecks and green-life expiration risks at the port [4,5]. During anomalous weeks (e.g., weather-driven production peaks or drops), a static historical average cannot adapt, whereas a tuned ensemble model can leverage contemporaneous signals to produce more accurate forecasts precisely when accuracy matters most for logistics planning. Future work incorporating real-time climatic variables, such as cumulative precipitation and temperature anomalies, could further enhance the ensemble models' ability to outperform historical baselines during extreme events [36].

The feature importance analysis (Figure 6) further reinforces the agronomic interpretability of the ensemble models. The prominence of the farm identifier (17.6%) captures between-farm heterogeneity in soil fertility, microclimate, and management practices that jointly determine baseline productivity levels. The importance of Week (11.4%) reflects the well-documented seasonality of *Cavendish* banana production, in which thermal accumulation drives fruit filling rates and harvest periodicity [8,9]. Critically, the bunch-age distribution variables (Pct 8–13 weeks) collectively contribute a substantial share of the predictive signal. These features encode the standing crop's developmental stage: bunches aged 8–10 weeks are in the rapid cell-expansion and starch-accumulation phase, while those at 11–13 weeks approach commercial maturity and are most sensitive to temperature-driven variations in filling rate [8]. The model's reliance on these physiological proxies means that it effectively tracks the biological pipeline of fruit approaching harvest readiness, rather than relying solely on statistical autocorrelation. This connection between feature importance and fruit development physiology—mediated by local climatic conditions such as temperature and soil water availability that modulate bunch filling rates by up to 39% [8]—bridges the gap between data-driven prediction and agronomic reality, providing a forecasting tool that respects the biological constraints of the crop while optimizing the industrial scale of the export supply chain.

The per-farm analysis revealed substantial heterogeneity in model performance, with R^2 values for XGBoost ranging from -0.60 (Farm 5) to 0.63 (Farm 6). Farms with higher baseline forecast errors tended to show the strongest improvements. Notably, Farm 1 was the only case where XGBoost exhibited a slightly higher MAE than the company baseline (4.91 vs. 4.62), though both models produced negative R^2 values (-0.22 and -0.29 , respectively). Both the company forecast and the ML model failed to explain the variance in that farm, which suggests that Farm 1's 2018 dynamics were driven by idiosyncratic factors not captured by the current feature set. These could include atypical management decisions, localized soil-drainage problems, or microclimate variability, none of which are reflected in the aggregate bunch-age distributions and historical production volumes the model relies on. This interpretation is consistent with research showing that the climate–productivity relationship in bananas varies substantially even across farms within the same region [36,37]. Farm 5 was the most challenging for all models, including the company baseline; its negative R^2 (-0.60) suggests that high production variability in this unit may be driven by factors not captured in the current feature set—such as localized weather events, irrigation variability, or pest outbreaks—and that farm-specific models or hierarchical approaches could yield additional gains.

The practical significance of these results is substantial. A reduction in MAE from 9.18 to 3.95 boxes per hectare at a seven-week horizon translates directly into improved capacity allocation for packing lines, more accurate transport bookings, and better alignment between harvest volumes and destination ripening programs. To illustrate the operational impact, consider that a standard 40-foot refrigerated container holds on the order of 960 boxes of export-grade bananas (approximately 20 pallets of 48 boxes each, at about 18 kg per box; the exact figure varies with box weight and packing configuration) [39]. For a farm of 200 Ha, the baseline forecast error of 9.2 boxes/Ha corresponds to a cumulative weekly deviation of approximately 1840 boxes—close to two full containers either over-booked (generating costly last-minute logistics) or under-booked (leaving idle cold-chain capacity). Reducing this error to 3.9 boxes/Ha narrows the deviation to approximately 780 boxes, allowing container bookings and packing schedules to be planned with substantially greater precision. The 57% reduction in forecast error also has implications for sustainable farm management. More precise allocation of packing materials, labor, and transport reduces the carbon footprint associated with under-utilized shipments or emergency logistics. It also

minimizes fruit rejection at destination ports due to green-life expiration caused by harvest-timing misalignment [4,5,7]. Recent work on AI-driven simulation for agri-fresh food supply chains has shown that intelligent network design and digital decision-making tools can significantly mitigate disruption impacts and improve logistics responsiveness [38], further supporting the operational value of integrating ML-based forecasts into supply chain planning. For a company managing nine farms, these improvements compound across production units, reducing idle packing capacity, limiting overtime during unexpected harvest peaks, and improving fruit condition at destination by enabling more precise cutting schedules aligned with green-life requirements.

To verify that these findings do not depend on the single 2018 test year, we performed an expanding-window rolling-origin validation. For each test year from 2016 to 2018, the models were retrained on all preceding years (from 2007), using the hyperparameters selected for the main analysis (Table 4), and evaluated on the held-out year. Table 9 reports the resulting MAE. The pattern is stable across the three years: every tuned ensemble reduces MAE by between 46% and 57% relative to the company baseline, CatBoost is consistently among the best models, and the Historical Mean remains a competitive statistical reference while the Seasonal Naive lags behind. The relative advantage of the ensembles is therefore not an artifact of the 2018 window, and it strengthens as the training history lengthens.

Table 9. Rolling-origin (expanding-window) validation: MAE (boxes/Ha) by model and test year. For each year, models were trained on all prior years (from 2007) using the hyperparameters of Table 4.

Model	2016	2017	2018
Company Baseline	10.92	11.64	9.18
Seasonal Naive	6.57	7.83	5.23
Historical Mean	6.18	5.88	4.28
Random Forest	5.88	5.90	4.27
Extra Trees	5.91	6.17	4.42
Gradient Boosting	5.68	5.72	4.12
LightGBM	5.81	5.70	4.13
XGBoost	5.70	5.74	4.05
CatBoost	5.55	5.55	3.95

Test years comprise 468 (2016 and 2017) and 405 (2018) observations. The 2018 column corresponds to the main-analysis setting (Table 5); minor differences for the tuned models arise because the rolling protocol refits each model with fixed, rather than re-tuned, hyperparameters. The Seasonal Naive entry differs because the rolling protocol indexes the same farm and calendar week of the previous year, whereas the main analysis uses a strict 52-week lag; the two definitions diverge when years span 53 weeks. Both implementations remain well above the ensemble errors. Best values per column in bold.

From a deployment standpoint, these models are light and practical to operate. The tuned ensembles train in under a minute on a standard workstation and produce a forecast in milliseconds, so an operator can refresh the seven-week-ahead outlook every week from records the company already keeps (bunch-age counts, packing volumes, and grade percentages), without new instrumentation. The forecast feeds the weekly cycle the company already runs: provisional packing-line and labor schedules, container bookings, and cold-chain allocation, with the model retrained at the end of each season as new weeks accumulate. Two cautions follow from the feature-importance analysis. Because the farm identifier and Week dominate the model, the learned patterns are specific to these nine farms and their seasonal cycle; applying the model to a new farm, or to a year with a structural change in management or climate, would require retraining and, most likely, the external climatic, soil, and disease-pressure variables discussed above. The model also

interpolates within historical operating conditions rather than extrapolating to unseen regimes, which is consistent with the poor performance on the idiosyncratic Farms 1 and 5. These constraints define the operating envelope within which the forecasts can be trusted.

Several limitations should be acknowledged. First, although we included a rolling-origin validation across 2016–2018 (Table 9), the external evaluation is still bounded by data ending in 2018; additional years and longer production histories would strengthen the robustness evidence further. Second, the feature set was limited to production-related variables available in the company's operational records. Systematic reviews consistently identify temperature, rainfall, and soil properties as important predictors of crop yield [25,27]; incorporating such variables could improve prediction accuracy, particularly for high-variability farms like Farm 5. Third, although the study included a classical autoregressive comparator (per-farm ARIMA) alongside the naive and historical baselines, it did not evaluate exponential-smoothing variants or deep learning architectures (LSTM, CNN). The ARIMA results are themselves informative: while the forecasting competition literature reports that statistical methods can be competitive at short horizons [16], the per-farm ARIMA here improved on the company and Seasonal Naive forecasts yet was clearly outperformed by both the Historical Mean and the ensembles. This indicates that, in this dataset, predictive skill arises less from univariate temporal autocorrelation than from stable cross-sectional (farm) and seasonal structure together with the contemporaneous operational signals exploited by the ensembles. LSTM models have been applied to banana yield forecasting in other contexts [31]; their evaluation, along with exponential smoothing, multi-year validation protocols, and farm-specific or hierarchical modeling, remains a direction for future work to further strengthen the evidence base.

5. Conclusions

This study evaluated six tuned ensemble machine learning models, three statistical baselines (Seasonal Naive, Historical Mean, and a per-farm ARIMA), and a company forecast baseline for seven-week-ahead forecasting of banana productivity per hectare using operational data from nine commercial farms in northern Colombia (2007–2018). All ensemble models substantially outperformed the company's existing forecast baseline, with CatBoost achieving the lowest errors on the 2018 external validation set (MAE = 3.95, RMSE = 5.04, $R^2 = 0.61$, EVS = 0.65)—though not significantly better than the other leading models—representing a 57.0% reduction in mean absolute error and an 81.2% reduction in mean squared error compared to the company baseline.

The Diebold–Mariano test revealed that CatBoost, XGBoost, LightGBM, Gradient Boosting, Random Forest, and the Historical Mean baseline were statistically equivalent in predictive accuracy ($p > 0.05$), while all significantly outperformed Extra Trees, the Seasonal Naive forecast, and the company baseline. An ablation study confirmed no data leakage from contemporaneous production features. This finding provides flexibility for operational deployment, as simpler models like Random Forest or faster implementations like LightGBM can be chosen based on computational constraints without sacrificing statistical performance.

Per-farm analysis revealed that the largest gains occurred in farms with high baseline forecast errors, while farms with highly volatile production remained challenging and may benefit from enriched feature sets incorporating climatic and agronomic variables.

From an operational perspective, these results demonstrate that tuned ensemble machine learning models can serve as practical decision-support tools for short-term harvest planning in banana export operations. The seven-week forecast horizon aligns with real operational decision timelines for packing capacity allocation, transport scheduling, and labor planning. Future work should explore integration of external weather and

agronomic variables, multi-year rolling validation protocols, and farm-specific model tuning to further enhance forecast accuracy and operational utility.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/agriengineering8070289/s1>, Dataset S1: banana_sample_desensitized.csv—desensitized representative weekly sample (2015–2018, nine farms) supporting reproducibility; README_sample.md—description of the sample and the desensitization procedure.

Author Contributions: Conceptualization, M.I.A.-E. and C.D.P.-A.; methodology, M.I.A.-E., C.D.P.-A., J.I.V. and G.A.G.-L.; software, M.I.A.-E.; formal analysis, M.I.A.-E.; data curation, M.I.A.-E. and C.D.P.-A.; writing—original draft, M.I.A.-E., C.D.P.-A., J.I.V. and G.A.G.-L.; visualization, M.I.A.-E., C.D.P.-A., J.I.V. and G.A.G.-L.; supervision, C.D.P.-A., J.I.V. and G.A.G.-L.; project administration, C.D.P.-A. and G.A.G.-L.; funding acquisition, C.D.P.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Becas de Excelencia Doctoral del Bicentenario program, through the Fondo de Ciencia, Tecnología e Innovación del Sistema General de Regalías (SGR), Colombia (2019). The APC was funded by Universidad del Norte.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The full dataset was provided under a strict confidentiality agreement that protects both farm identities and commercial production figures and therefore cannot be made publicly available. To support reproducibility, a desensitized representative sample (weekly records for 2015–2018, with farm identifiers randomly relabeled and all magnitude variables rescaled by an undisclosed factor and perturbed with random noise) is provided as Supplementary Materials; this sample preserves the data structure for demonstrating the pipeline but does not reproduce the exact reported results. The full data are available from the authors upon reasonable request, subject to the confidentiality agreement.

Acknowledgments: The authors thank the banana company for providing the production data used in this study.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

MAE	Mean Absolute Error
RMSE	Root-Mean-Square Error
MSE	Mean Squared Error
EVS	Explained Variance Score
MSLE	Mean Squared Logarithmic Error
DM	Diebold–Mariano
CV	Cross-Validation
CI	Confidence Interval
TR4	Tropical Race 4
UAV	Unmanned Aerial Vehicle

References

1. FAO. *Banana Market Review 2024*; Food and Agriculture Organization of the United Nations: Rome, Italy, 2025. Available online: <https://openknowledge.fao.org> (accessed on 30 June 2026).
2. IISD. *Global Market Report: Banana Prices and Sustainability*; International Institute for Sustainable Development: Winnipeg, MB, Canada, 2023.

3. Bebbier, D.P. The long road to a sustainable banana trade. *Plants People Planet* **2023**, *5*, 662–671.
4. Brat, P.; Bugaud, C.; Guillermet, C.; Salmon, F. Review of banana green life throughout the food chain. *Sci. Hortic.* **2020**, *262*, 109054. [[CrossRef](#)]
5. Jedermann, R.; Praeger, U.; Geyer, M.; Lang, W. Remote quality monitoring in the banana chain. *Philos. Trans. R. Soc. A* **2014**, *372*, 20130303. [[CrossRef](#)]
6. Ahumada, O.; Villalobos, J.R. Application of planning models in the agri-food supply chain: A review. *Eur. J. Oper. Res.* **2009**, *196*, 1–20. [[CrossRef](#)]
7. Al-Dairi, M.; Pathare, P.B.; Al-Yahyai, R.; Jayasuriya, H.; Al-Attabi, Z. Postharvest quality, technologies, and strategies to reduce losses along the supply chain of banana: A review. *Trends Food Sci. Technol.* **2023**, *134*, 177–191. [[CrossRef](#)]
8. Turner, D.W.; Fortescue, J.A.; Thomas, D.S. Environmental physiology of the bananas (*Musa* spp.). *Braz. J. Plant Physiol.* **2007**, *19*, 463–484. [[CrossRef](#)]
9. Ganry, J.; Chillet, M. Methodology to forecast the harvest date of banana bunches. *Fruits* **2008**, *63*, 371–373. [[CrossRef](#)]
10. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*, 5th ed.; Wiley: Hoboken, NJ, USA, 2015.
11. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 3rd ed.; OTexts: Melbourne, Australia, 2021. Available online: <https://otexts.com/fpp3/> (accessed on 30 June 2026).
12. Bergmeir, C.; Benítez, J.M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **2012**, *191*, 192–213. [[CrossRef](#)]
13. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [[CrossRef](#)]
14. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 Competition: Results, findings, conclusion and way forward. *Int. J. Forecast.* **2018**, *34*, 802–808. [[CrossRef](#)]
15. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M5 competition: Background, organization, and implementation. *Int. J. Forecast.* **2022**, *38*, 1325–1336. [[CrossRef](#)]
16. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE* **2018**, *13*, e0194889. [[CrossRef](#)] [[PubMed](#)]
17. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
18. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
19. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
20. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6639–6649.
21. Mahesh, P.; Soundrapandiyan, R. Yield prediction for crops by gradient-based algorithms. *PLoS ONE* **2024**, *19*, e0291928. [[CrossRef](#)] [[PubMed](#)]
22. Ennaji, O.; Baha, S.; Vergutz, L.; El Allali, A. Gradient boosting for yield prediction of elite maize hybrid ZhengDan 958. *PLoS ONE* **2024**, *19*, e0315493. [[CrossRef](#)] [[PubMed](#)]
23. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine Learning in Agriculture: A Review. *Sensors* **2018**, *18*, 2674. [[CrossRef](#)] [[PubMed](#)]
24. Hernández Hernández, G.C.; Gómez Gómez, J.; Jiménez-Cabas, J. Predictive Models Based on Artificial Intelligence to Estimate Crop Yield: A Literature Review. *Agriculture* **2025**, *15*, 2438. [[CrossRef](#)]
25. van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [[CrossRef](#)]
26. Javed, M.A.; Murad, M.A.A. Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches. *Heliyon* **2024**, *10*, e40836. [[CrossRef](#)] [[PubMed](#)]
27. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [[CrossRef](#)]
28. Oikonomidis, A.; Catal, C.; Kassahun, A. Deep learning for crop yield prediction: A systematic literature review. *N. Z. J. Crop Hortic. Sci.* **2022**, *51*, 1–26. [[CrossRef](#)]
29. Ravikumar, A.; Periyasamy, V.; Devi, K.; Krishnamoorthy, R.; Mariavalavan, O. A hybrid SERWI ensemble model for crop yield prediction using an inverse RMSE weighting strategy. *Sci. Rep.* **2025**, *15*, 45085. [[CrossRef](#)] [[PubMed](#)]
30. Srivastava, A.K.; Safaei, N.; Khaki, S.; Lopez, G.; Zeng, W.; Ewert, F.; Gaiser, T.; Rahimi, J. Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Sci. Rep.* **2022**, *12*, 3215. [[CrossRef](#)] [[PubMed](#)]
31. Patrick, S.; Mirau, S.; Mbalawata, I.; Leo, J. Time series and ensemble models to forecast banana crop yield in Tanzania, considering the effects of climate change. *Resour. Environ. Sustain.* **2023**, *14*, 100138. [[CrossRef](#)]

32. Yáñez-Cajo, D.; Vasconez-Montufar, G.; Villamar-Torres, R.O.; Godoy-Montiel, L.; Jazayeri, S.M.; Pérez-Porras, F.; Mesas-Carrascosa, F. Banana Yield Prediction Using Random Forest, Integrating Phenology Data, Soil Properties, Spectral Technology, and UAV Imagery in the Ecuadorian Littoral Region. *Sustainability* **2025**, *17*, 10098. [CrossRef]
33. Ndayambaje, A.; Ngango, J. Predicting banana crop yield in Eastern Rwanda using machine learning models. *Cogent Food Agric.* **2025**, *11*, 2578320. [CrossRef]
34. Zhou, L.; Yang, Z.; Fu, L.; Duan, J. Yield Estimation in Banana Orchards Based on DeepSORT and RGB-Depth Images. *Agronomy* **2025**, *15*, 1119. [CrossRef]
35. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.
36. Varma, V.; Bebbber, D.P. Climate change impacts on banana yields around the world. *Nat. Clim. Change* **2019**, *9*, 752–757. [CrossRef]
37. Aguirre, S.E.; Piraneque, N.V.; Rodríguez Barrios, J. Relationship between the nutritional status of banana plants and black sigatoka severity in the Magdalena region of Colombia. *Agron. Colomb.* **2015**, *33*, 348–355. [CrossRef]
38. Tascon, D.; Mejia, G.; Paternina, C. AI-Driven Computational Simulation for Enhancing Resilience in Agri-Fresh Food Supply Chains. In Proceedings of the 10th North American Conference on Industrial Engineering and Operations Management, Orlando, FL, USA, 17–19 June 2025. [CrossRef]
39. Notteboom, T.; Pallis, A.; Rodrigue, J.-P. The Transport and Handling of Bananas. In *Port Economics, Management and Policy*, 2nd ed.; Routledge: New York, NY, USA, 2026. Available online: <https://porteconomicsmanagement.org/pemp/contents/part5/bulk-breakbulk-terminal-design-equipment/handling-bananas/> (accessed on 30 June 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.