BAREKENG: Journal of Mathematics and Its Applications
September 2026 Volume 20 Issue 3 Page 2491-2506
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss3pp2491-2506

ENHANCING MAIZE YIELD PREDICTION IN INDONESIA USING HYBRID MACHINE LEARNING MODELS

Adyanata Lubis[1*], Eko Oktafanda[2], Juliarni[3], Junadhi[4]

[1,2]*Department of Computer Science, Universitas Rokania*
[3]*Department of Agricultural Science, Universitas Rokania*
Langkitin, Rambah Samo, Rokan Hulu Regency, Riau 28557, Indonesia

[4]*Department of Computer Science, Universitas Sains dan Teknologi Indonesia*
Jln. Purwodadi, Panam, Pekanbaru, 28299, Indonesia

*Corresponding author's e-mail: * **adyanata@rokania.ac.id***

Article Info

Article History:

Received: 1st September 2025
Revised: 10th November 2025
Accepted: 17th March 2026
Published:8th April 2026

Keywords:

Hybrid Machine Learning;
Maize Yield Prediction;
Random Forest, Support;
Vector Regression;
XGboost.

ABSTRACT

Maize is a strategic commodity in Indonesia's national food system, yet traditional yield-prediction methods based on statistical or survey approaches often fail to capture the nonlinear and dynamic relationships among agronomic, climatic, and socio-economic variables. Accurate forecasting remains essential for supporting food self-sufficiency and climate-resilient agricultural planning. To address these challenges, this study proposes SMART-JAGUNG, a machine learning–based maize yield prediction system employing three ensemble and regression models: Random Forest (RF), Support Vector Regression (SVR), and eXtreme Gradient Boosting (XGBoost). The dataset comprises five years of maize production data from the Indonesian Central Bureau of Statistics (BPS), along with auxiliary variables including rainfall, temperature, NDVI, seed type, and fertilizer use. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) before and after hyperparameter tuning with GridSearchCV. Results indicate that RF achieved the best performance before tuning (MAE = 36,310.53; RMSE = 95,343.05; R^2= 0.9758), followed closely by XGBoost, while SVR consistently underperformed. Although post-tuning performance slightly decreased, the predicted-versus-actual visualization confirmed the robustness of RF and XGBoost for non-extreme data. Overall, SMART-JAGUNG demonstrates strong potential as a reliable, data-driven decision-support tool for precise maize yield estimation, contributing to sustainable food security and national self-sufficiency policies.

How to cite this article:

A. Lubis, E. Oktafanda, Juliarni and Junadhi., "ENHANCING MAIZE YIELD PREDICTION IN INDONESIA USING HYBRID MACHINE LEARNING MODELS", *BAREKENG: J. Math. & App.*, vol. 20, no. 3, pp. 2491-2506, Sep, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Maize is a vital agricultural commodity at both the global and national levels. Globally, maize ranks among the top three staple crops, alongside rice and wheat, providing a substantial proportion of daily caloric intake for billions of people [1]. In Indonesia, maize occupies a particularly strategic position as the second-largest source of carbohydrates after rice, while also serving as a crucial input for the livestock feed industry and an emerging feedstock for bioenergy production [2][3]. Its role in ensuring national food security has become increasingly important amid growing demand driven by population growth, dietary diversification, and the expansion of agro-industrial activities [4].

Despite its strategic importance, Indonesia has yet to achieve sustainable self-sufficiency in maize production. Current production levels remain unable to fully meet domestic demand, resulting in recurring reliance on imports to bridge the supply gap [5]. This dependency not only exposes the national food system to global market volatility but also exacerbates local price instability, undermining farmer welfare and food affordability. Furthermore, seasonal fluctuations in yield, driven by climatic variability and inconsistent agronomic practices, have contributed to production uncertainty and weakened the resilience of the maize supply chain [6].

Accurate yield forecasting is a cornerstone for effective agricultural planning and policy formulation. Reliable predictions enable policymakers, agribusiness stakeholders, and farmers to make informed decisions on production targets, input allocation, distribution planning, and market stabilization measures [7]. However, accurately predicting maize yields in Indonesia remains a significant challenge. The complexity arises from the interplay of multiple, highly nonlinear factors, including agro-climatic conditions (rainfall, temperature, humidity), environmental indicators (Normalized Difference Vegetation Index, soil fertility levels), agronomic variables (seed type, planting density, fertilizer application), and socio-economic drivers (market access, farmer capital, labor availability). These factors often interact in unpredictable ways, making yield forecasting a multidimensional problem that cannot be effectively addressed solely with conventional statistical methods [8][9]. Previous studies have attempted to apply various statistical and machine learning models for crop yield prediction; however, most were limited by small or region-specific datasets, a lack of ensemble-based optimization, and minimal evaluation of model robustness across diverse agro-climatic conditions. Few studies have explicitly compared multiple ensemble regression algorithms using standardized performance metrics on Indonesian datasets. This research gap underscores the need for a more comprehensive, scalable, data-driven approach that integrates heterogeneous features to improve predictive accuracy and generalizability. Therefore, this study introduces SMART-JAGUNG, an ensemble machine learning–based maize yield prediction system that combines Random Forest, XGBoost, and Support Vector Regression (SVR) to address these limitations and support evidence-based decision-making for national food self-sufficiency.

Traditional yield prediction methods, such as linear regression or time-series models, generally assume simple relationships between variables and are limited in their ability to handle high-dimensional, noisy, and nonlinear datasets. As a result, their predictive accuracy tends to deteriorate in complex agricultural environments characterized by diverse crop management practices, heterogeneous land characteristics, and unpredictable climatic patterns. In Indonesia, the scarcity of integrated, high-quality datasets further compounds the problem, as most predictive models rely on limited historical production data and do not incorporate environmental and agronomic variables with sufficient granularity [10][8]. This creates a research gap in the development of robust, data-driven prediction models that can integrate multiple sources of information and adapt to complex, multivariate relationships.

Recent studies have demonstrated that machine learning, particularly ensemble methods such as Random Forests and XGBoost, can significantly improve the accuracy and generalizability of crop yield forecasting across diverse climatic and soil conditions [11][12]. However, few studies have systematically compared multiple ensemble regression models using standardized evaluation metrics and real-world production data from Indonesia. Therefore, this study introduces SMART-JAGUNG, an ensemble machine learning–based prediction system designed to fill this gap by integrating multi-year agricultural, climatic, and environmental datasets to enhance model precision and provide a scalable decision-support framework for sustainable maize production forecasting.

Building on these developments, this study introduces SMART-JAGUNG, a hybrid machine learning-based maize yield prediction system specifically designed for Indonesia's agricultural context. The framework integrates three high-performing regression algorithms: Random Forest (RF), Support Vector

Regression (SVR), and eXtreme Gradient Boosting (XGBoost) [13]. These algorithms were selected based on their complementary strengths in handling complex datasets: RF is well-suited for managing high-dimensional and noisy data while mitigating overfitting risks [14]; SVR excels at modeling intricate nonlinear patterns with narrow margin optimization [15]; and XGBoost is recognized for its computational efficiency, scalability, and robustness in capturing high-order feature interactions [16]. By combining these algorithms within an ensemble framework, the Proposed Method Development seeks to harness their individual strengths while mitigating their respective limitations.

The dataset used in this study comprises five years of historical maize production data obtained from Indonesia's Central Bureau of Statistics (BPS), enriched with auxiliary variables including rainfall, temperature, NDVI, seed type, and fertilizer usage. This integration of environmental, agronomic, and socio-economic indicators addresses a key limitation in many previous studies, which often relied solely on historical yield figures and did not incorporate diverse explanatory variables [17]. Model performance is rigorously evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), both before and after hyperparameter optimization via GridSearchCV. Additionally, model robustness is assessed through visualization techniques, such as predicted-versus-actual value plots and residual distribution analyses, enabling a deeper understanding of predictive accuracy across varying data ranges.

Given the complexity of maize yield prediction in Indonesia and the limitations of existing approaches, this study adopts a hybrid machine learning framework to enhance predictive accuracy and reliability. The following section details the methodological design of Proposed Method Development, including dataset description, preprocessing procedures, model architecture, hyperparameter optimization strategies, and evaluation metrics. By systematically presenting these methodological components, the study ensures transparency, reproducibility, and a clear foundation for assessing the model's applicability in operational agricultural decision-making contexts.

2. RESEARCH METHODS

The dataset comprising 191 observations from the Indonesian Central Bureau of Statistics (BPS) was validated to ensure consistency between harvested area, production, and yield records. Missing values (<3%) were imputed using mean substitution for continuous variables and mode substitution for categorical variables, while outliers detected using the IQR method were retained if they represented plausible agricultural extremes. A 5-fold stratified cross-validation was employed to enhance model reliability and prevent overfitting. Although relatively small, the dataset remained suitable for ensemble regressors such as Random Forest and XGBoost, which perform well on limited data when supported by proper feature engineering. Nevertheless, potential biases due to regional imbalance and limited temporal scope are acknowledged as limitations for future research. The methodological framework is depicted in Fig. 1.

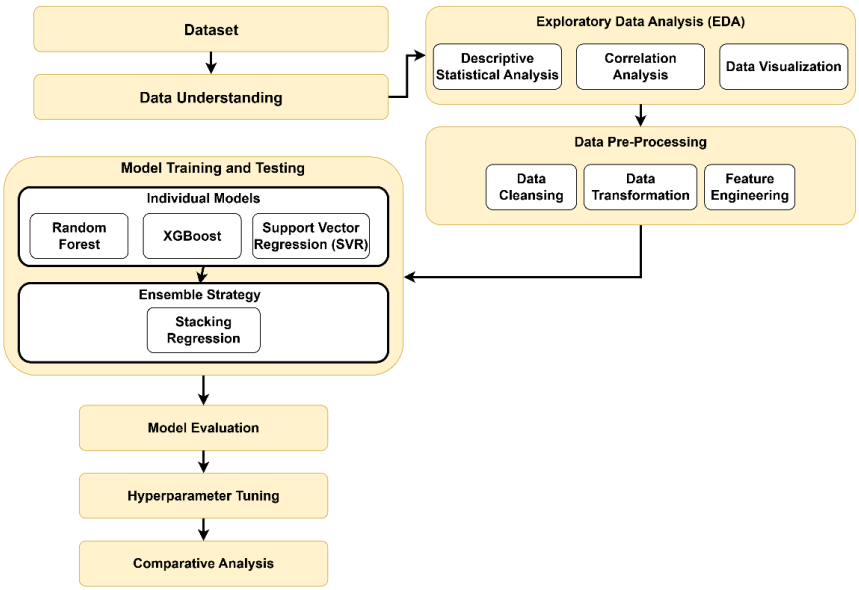

Figure 1. Proposed Method Development

2.1 Data Understanding

In the data understanding phase, efforts focused on identifying and comprehending the structure, types, and characteristics of the data used in this study [17][18]. The primary dataset was sourced from the Indonesian Central Bureau of Statistics (BPS) and comprised maize production data at the provincial level for the past five years (2020–2024). This dataset includes key variables, including harvested area, productivity, and total production. To enhance the predictive model's contextual richness, additional variables were incorporated, including rainfall, average temperature, Normalized Difference Vegetation Index (NDVI), seed type, fertilizer quantity, irrigation method, and maize price. Preliminary observations were also conducted during this stage to analyze data distribution, detect outliers, and identify potential inconsistencies or missing values. These insights formed the foundation for designing appropriate preprocessing strategies and subsequent modeling procedures.

2.2 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase involved a series of preliminary analyses to gain deeper insights into the dataset's patterns, structure, and interrelationships. This process included statistical profiling of numerical variables such as mean, median, minimum, maximum, and standard deviation to identify descriptive characteristics of the data [19]. Visualization techniques, including histograms, boxplots, and scatterplots, were employed to detect anomalies, outliers, and skewness that could potentially affect model performance [20]. Additionally, correlation analysis was conducted among numerical variables using Pearson correlation coefficients to assess the strength of linear relationships between independent variables and the target variable (maize production). A correlation heatmap was utilized to visually depict the magnitude and direction of these relationships, serving as a foundation for relevant feature selection. EDA also facilitated the assessment of multicollinearity among predictors and the detection of spatial or temporal patterns in the data distribution. This stage played a crucial role in informing subsequent modeling strategies, particularly in selecting appropriate features, applying data transformation techniques, and choosing machine learning algorithms suited to the complex nature of agricultural datasets.

2.3. Pre-Processing

The pre-processing stage aims to prepare the dataset for effective use in machine learning model training [18]. The raw data obtained from the Central Bureau of Statistics (BPS), along with supplementary variables, underwent a series of cleaning and transformation steps to meet the requirements of the selected learning algorithms. The initial step involved data cleaning, which included the removal of irrelevant entries such as aggregate national-level records (e.g., rows labeled "INDONESIA") and the treatment of missing values using mean imputation for numerical features and mode imputation for categorical variables [10]. Subsequently, feature transformation was performed, particularly on categorical variables such as seed type and irrigation method, which were encoded as numerical values using One-Hot Encoding to meet the input requirements of most machine learning algorithms. To ensure uniformity in feature scaling, all numerical variables, such as harvested area, rainfall, and NDVI, were normalized using Min-Max Scaling [21], rescaling the values to the range 0-1. After these preprocessing steps, the dataset was split into training (80%) and test (20%) sets using a random split. This separation ensures that the model can learn from historical data and be evaluated on previously unseen data to assess generalization. Preprocessing plays a crucial role in the machine learning pipeline, as the quality of the input data directly influences model accuracy and stability. This step also ensures that the data satisfies the fundamental assumptions required by the regression algorithms applied, such as SVR and XGBoost.

2.4 Model Training and Testing

This stage represents the central component of the machine learning workflow, where models are trained to capture the underlying patterns and relationships between input features and the target variable (maize yield). In this study, three well-established regression algorithms were utilized: Random Forest (RF), Support Vector Regression (SVR), and eXtreme Gradient Boosting (XGBoost). Each algorithm was trained separately on the training dataset and later tested on the evaluation dataset to measure predictive performance. The results of this evaluation provide insight into each model’s generalization capability, ensuring robustness and reliability when applied in real-world scenarios.

1. Random Forest

   Random Forest is an ensemble learning algorithm that constructs multiple decision trees and aggregates their outputs to improve prediction accuracy and reduce the risk of overfitting. In regression tasks, the Random Forest model generates a predicted output $\hat{y}$ by averaging the predictions of all decision trees T within the ensemble. This ensemble averaging mechanism enhances model stability and robustness, particularly when dealing with datasets that contain noise or complex nonlinear relationships among features [22].

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x), \tag{1}$$

   where $h_t(x)$ represents the prediction of the t-th decision tree. Random Forest is effective in handling non-linear data, multicollinearity, and outliers, and it also provides valuable insights through feature importance analysis.

2. Support Vector Machine

   Support Vector Regression (SVR) is a variant of Support Vector Machine (SVM) applied to regression tasks. The model aims to find a function $f(x)$ that deviates from the actual target values y_i, by no more than a predefined margin ε, while simultaneously maintaining minimal model complexity [10]. The prediction function of SVR in its kernelized form is given by.

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b, \tag{2}$$

   where:

   α_i, α_i^* are the Lagrange multipliers obtained during training,

   $K(x_i, x)$ is the kernel function (e.g., RBF or linear),

   b is the bias of the model.

   SVR is effective in handling non-linear relationships and maintaining the error margin within a specified tolerance using the ε- insensitive loss function.

3. XGBoost

   XGBoost is a tree-based boosting algorithm that builds models in a sequential manner and optimizes the objective function using gradient descent techniques. The final prediction is an accumulation of the outputs from each tree f_k, which are added iteratively [23].

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i), f_k \in F. \tag{3}$$

   The objective function used in XGBoost can be expressed as:

$$\mathcal{L}(\emptyset) = \sum_{i=1}^{i} l(y_i, \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k), \tag{4}$$

   where:

   $\mathcal{L}$ is the loss function (typically Mean Squared Error or MSE),

   $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ is the regularization function for model complexity,

   T is the number of leaves in the tree, and λ is the regularization parameter.

2.5 Model Evaluation

Model evaluation plays a vital role in determining the predictive effectiveness of each machine learning algorithm applied. In this research, three widely adopted regression metrics were employed: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2). These metrics

were chosen because they provide a well-rounded assessment of model accuracy and prediction error in continuous regression tasks [24], [23].

1. Mean Absolute Error (MAE)

Measures the average absolute difference between the actual and predicted values. This metric assigns equal weight to all deviations, making it robust against outliers [25].

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{5}$$

where y_i is the actual value, y_i is the predicted value, and n n is the total number of data points.

2. Root Mean Square Error (RMSE)

Measures the average squared error between the actual and predicted values, followed by taking the square root. RMSE is more sensitive to outliers because it squares the magnitude of the deviations [26].

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}. \tag{6}$$

A lower RMSE value indicates that the model has a generally lower prediction error.

3. R-squared (R^2) or Coefficient of Determination

R^2 indicates the proportion of variance in the target variable that can be explained by the predictive model. The R^2 value ranges from 0 to 1, with values closer to 1 indicating that the model performs well in explaining the variability of the data [27].

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \tag{7}$$

where $\hat{y}_i$ denotes the mean of the actual values. R^2 can also take negative values if the model performs worse than a baseline model that simply predicts the mean of the target variable.

2.6 Hyperparameter Tuning

To ensure robustness and prevent overfitting during tuning, a 5-fold cross-validation strategy was used for each model. The dataset was randomly divided into five equal subsets, ensuring proportional representation of regional and climatic variations. In each iteration, four folds were used for training, and one for validation, with the folds rotated so that each subset served once as the validation fold. This procedure provided a reliable estimate of model generalization and prevented the model from being overly optimized to a specific data partition. The integration of cross-validation within the Grid Search and Randomized Search frameworks enabled systematic evaluation of multiple parameter combinations—such as the number of estimators, tree depth, learning rate, and kernel parameters—under consistent conditions. To confirm the stability of the optimization results, all experiments were repeated using three different random seeds. Consistent performance across folds indicated that the tuning process was stable and that the final configuration achieved a good balance between model complexity and predictive reliability [26], [28].

2.7 Model Comparison

The Model Comparison phase is intended to assess and compare the performance of machine learning models, both before and after hyperparameter tuning. This evaluation relies on three key regression metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2) to measure the accuracy and robustness of maize yield predictions. Before tuning, the Random Forest model showed strong baseline results, with low MAE and RMSE values and an R^2 score approaching 1. After tuning, however, XGBoost demonstrated the greatest improvement, especially in minimizing prediction errors and increasing the model's ability to explain data variance. On the other hand, SVR, while relatively stable, tended to produce higher error rates when applied to complex agricultural datasets such as corn yields, unless optimized with well-adjusted parameter settings [29].

3. RESULTS AND DISCUSSION

3.1 Dataset

The dataset used in this study was obtained from Statistics Indonesia (Badan Pusat Statistik - BPS) and covers corn-related data from 2020 to 2024 across 34 provinces in Indonesia. The primary variables include harvested area, yield (productivity), and total corn production. To enhance the model's predictive capabilities, the dataset was enriched with additional variables, including rainfall, average temperature, Normalized Difference Vegetation Index (NDVI), seed type, irrigation type, fertilizer quantity, and corn price. All variables underwent a thorough preprocessing and normalization procedure [27]. The final dataset consists of 191 multivariate observations, structured to be compatible with the ensemble machine learning models developed in this research.

3.2 Exploratory Data Analys (EDA)

The Exploratory Data Analysis (EDA) was conducted to gain insights into the distributions, patterns, and interrelationships among the variables used in the corn yield prediction model. The EDA results showed that the harvested area exhibited a strong positive correlation with corn production, suggesting that a larger cultivated area generally yields higher potential yields. Meanwhile, the productivity variable (quintals per hectare) also showed a linear relationship with production, though in some cases it was influenced by additional factors, such as fertilizer use and irrigation type. The distributions of NDVI and rainfall values displayed significant variability across provinces, reflecting the diverse agroclimatic conditions that could affect prediction accuracy if not adequately addressed. Furthermore, variables such as corn price and seed type indicated socio-economic influences that may affect agricultural inputs and outputs. The findings from the EDA phase served as a foundation for feature selection and model tuning in subsequent stages of the study.

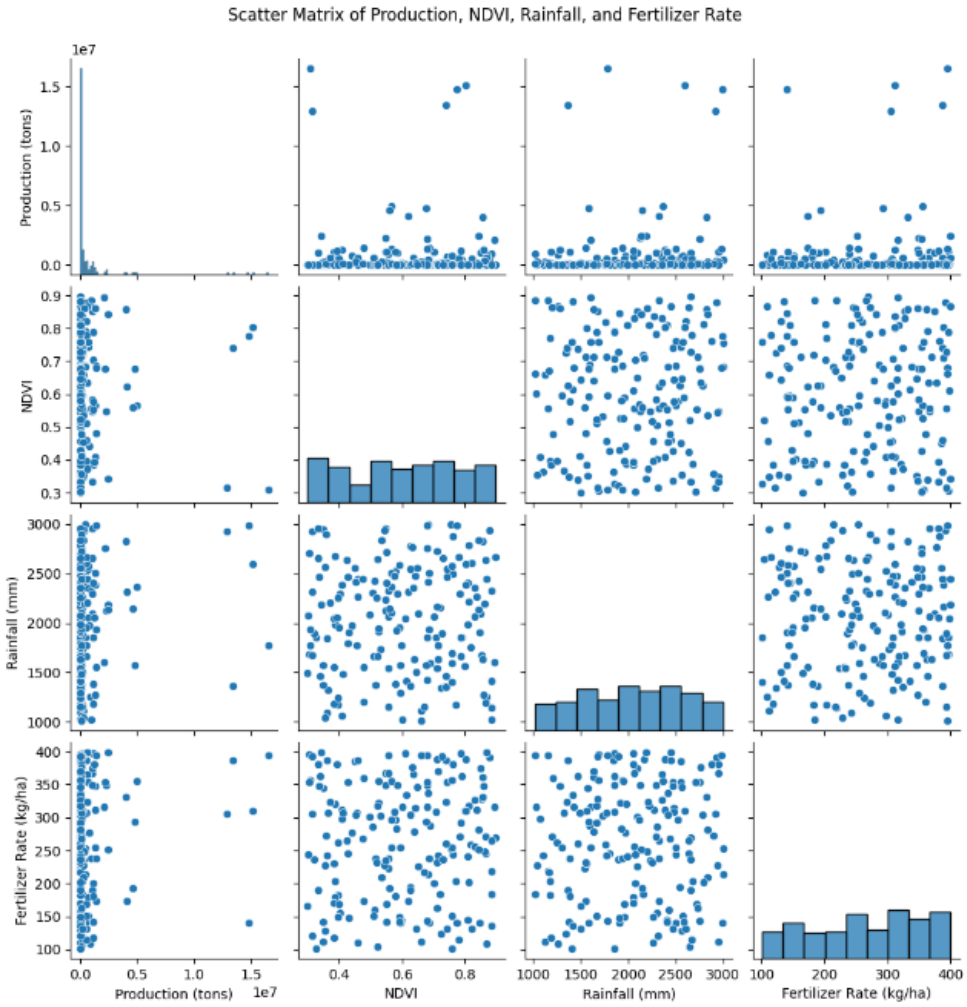

Figure 2. Scatter Matrix is the Distribution Relationship Between Features

Fig. 2 presents a scatter matrix illustrating the relationships among key numerical variables: Production (tons), NDVI, Rainfall (mm), and Fertilizer Amount (kg/ha). The histograms along the diagonal indicate that the distribution of production is highly right-skewed, suggesting the presence of outliers, whereas the other variables exhibit relatively normal distributions. The scatter plots between variables reveal a lack of strong linear relationships, particularly between Production and NDVI, Rainfall, and Fertilizer Use. These patterns suggest that a simple linear regression model is insufficient to accurately capture the underlying relationships among the features. Therefore, nonlinear regression techniques such as Random Forest, Support Vector Regression (SVR), or XGBoost are deemed more appropriate for modeling the complex interactions among variables and improving predictive performance.

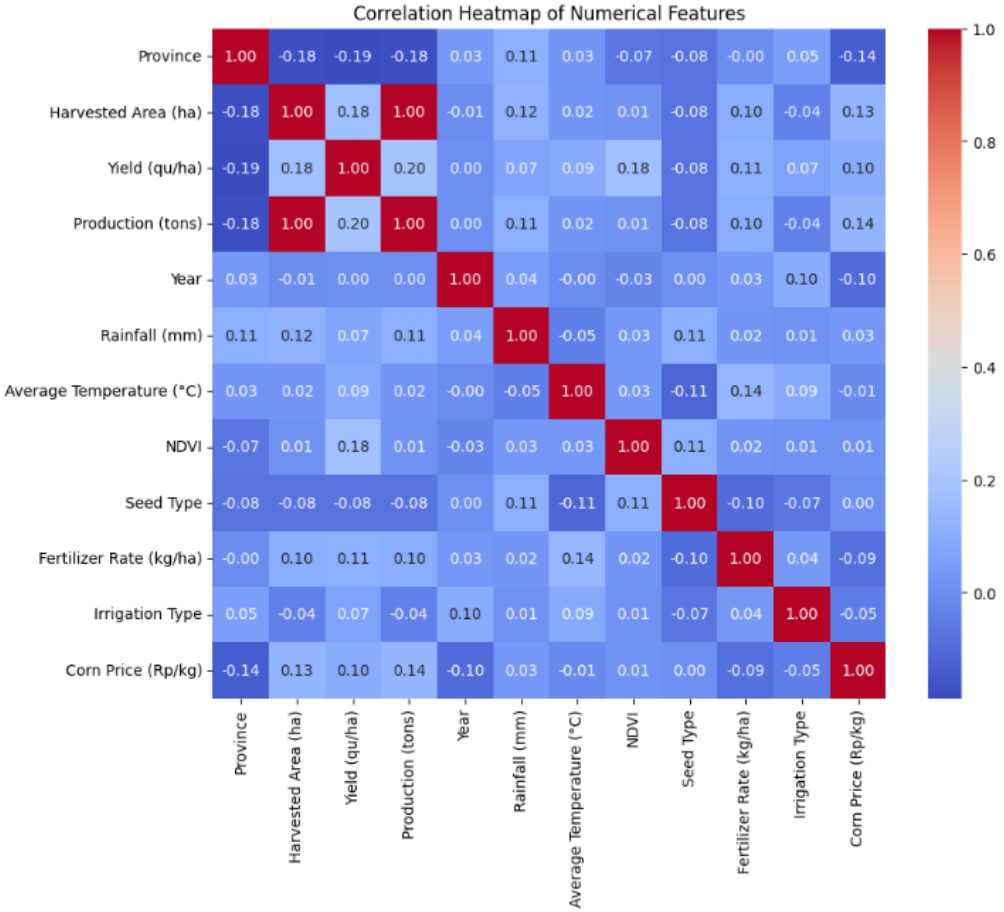

Figure 3. Correlation Heatmap Shows the Overall Relationship Between Numeric Features

Fig. 3 presents a correlation heatmap of the interrelationships among numerical features in the corn yield prediction dataset. In general, the observed Pearson correlation coefficients are relatively low, with most values approaching zero, indicating weak linear associations among the variables. However, moderate positive correlations are observed between Harvested Area and Production (r = 0.20), which is logically consistent given that total production is directly influenced by both cultivated land area and crop productivity. Environmental features such as Rainfall, NDVI, and Fertilizer Amount exhibit weak correlations with Production, suggesting that their relationships with the target variable are likely non-linear and complex. These findings reinforce the justification for employing non-linear predictive models rather than linear regression, to more accurately capture the intricate interactions among variables in the agricultural domain.

To understand the regional contributions to national corn production, data from the period 2020 to 2024 were aggregated by province. The results are presented in Fig. 4, which displays the top ten provinces with the highest total corn production during this period. This visualization provides a comparative overview of the primary production centers, serving as a reference for resource distribution planning, agricultural policy formulation, and spatial validation of the predictive model. The dominance of several provinces, particularly East Java and Central Java, indicates a significant concentration of production, highlighting these areas as strategic zones in achieving the national corn self-sufficiency targets.

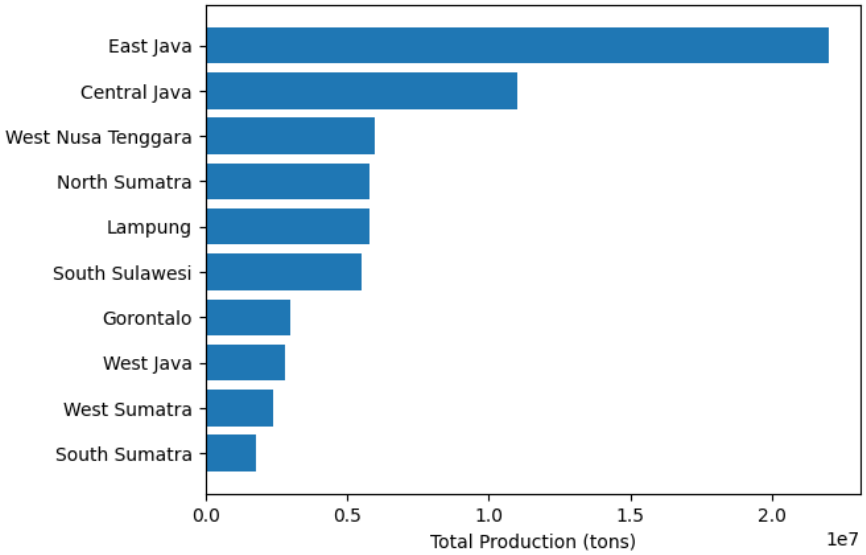

Figure 4. Top 10 Provinces with Highest Total Maize Production (2020–2024)

3.3 Preprocessing

The preprocessing stage was conducted to ensure that the dataset used for model training was clean, consistent, and compatible with machine learning algorithms. This process involved several critical steps, beginning with the integration of multi-year data (2020–2024) obtained from the Indonesian Central Bureau of Statistics (BPS), along with the enrichment of external predictive variables such as rainfall, average temperature, NDVI, and corn prices. Subsequently, data type handling was performed, whereby all categorical variables such as seed type and irrigation method—were transformed into numerical format using label encoding, enabling their recognition by the models. Missing values were identified and addressed through simple imputation (mean/mode) or eliminated if deemed statistically insignificant. To ensure consistency across features, normalization and rescaling were applied to numerical variables, including rainfall, NDVI, temperature, fertilizer quantity, and market prices, using the Min-Max Scaling method. This step was particularly essential for models sensitive to data scale, such as Support Vector Regression (SVR). Finally, the entire dataset was systematically randomized to eliminate temporal ordering bias and was split into training and testing sets in an 80:20 ratio. This preprocessing phase served as the foundational backbone for constructing the SMART-JAGUNG prediction system based on ensemble machine learning, aiming to deliver accurate and sustainable yield estimations for national corn production. Table 1 presents the preprocessed corn production dataset spanning 2020–2024.

Table 1. Corn Production Dataset 2020-2024

Province	Harvested Area (ha)	Produc-tivity (qu/ha)	Production (tons)	Year	Rainfall (mm)	Average Tempe-rature (°C)	NDVI	Seed Type	Fertilizer Amount (kg/ha)	Irrigation Type	Maize Price (Rp/kg)
Aceh	11581.2	55.22	63590.8	2020	1201.27	31.03	0.4776	Hybrid	359.74	Pump Well	4604.18
North Sumatera	135334.39	57.87	783162.62	2020	1971.03	25.91	0.4149	Hybrid	277.51	Technical Irr.	4778.98
West Sumatera	65756.37	64.44	425052.38	2020	1852.36	26.07	0.7095	Transgenic	362.99	Rainfed	3784.20
Riau	138.92	34.03	472.78	2020	1570.77	29.89	0.3506	Hybrid	345.35	Technical Irr.	4187.38
Jambi	1111.0	68.45	7604.47	2020	2496.39	31.68	0.7611	Hybrid	243.09	Pump Well	3844.34
South Sumatera	35073.83	60.37	211735.52	2020	1775.59	23.53	0.5963	Transgenic	142.18	Technical Irr.	4958.07
Bengkulu	4145.68	56.48	23415.58	2020	2092.04	27.78	0.6500	Hybrid	226.13	Technical Irr.	5462.37
Lampung	156654.98	62.47	977197.39	2020	1021.22	26.24	0.8556	Local	326.15	Rainfed	4569.53
Bangka Belitung Isl.	28.73	44.09	126.66	2020	2089.29	26.45	0.8971	Local	300.46	Rainfed	4374.05
Riau Islands	2.75	55.4	15.22	2020	1148.65	30.50	0.8844	Hybrid	218.29	Technical Irr.	5478.96

3.4 Model Training and Testing

The SMART-JAGUNG prediction model was trained using an ensemble machine learning approach that integrates three well-established regression algorithms: Random Forest (RF), Support Vector Regression (SVR), and eXtreme Gradient Boosting (XGBoost). The selection of these models was based on their proven performance in capturing nonlinear patterns and handling complex relationships among predictor variables in multivariate agricultural data, specifically for corn yield prediction. After comprehensive preprocessing, the dataset was randomly split into 80% for training and 20% for testing to ensure proper model generalization. The training process employed supervised regression, in which the models were trained to predict corn production (in tons) based on a range of input features, including harvested area, productivity, rainfall, NDVI, fertilizer quantity, and other agronomic and environmental factors.

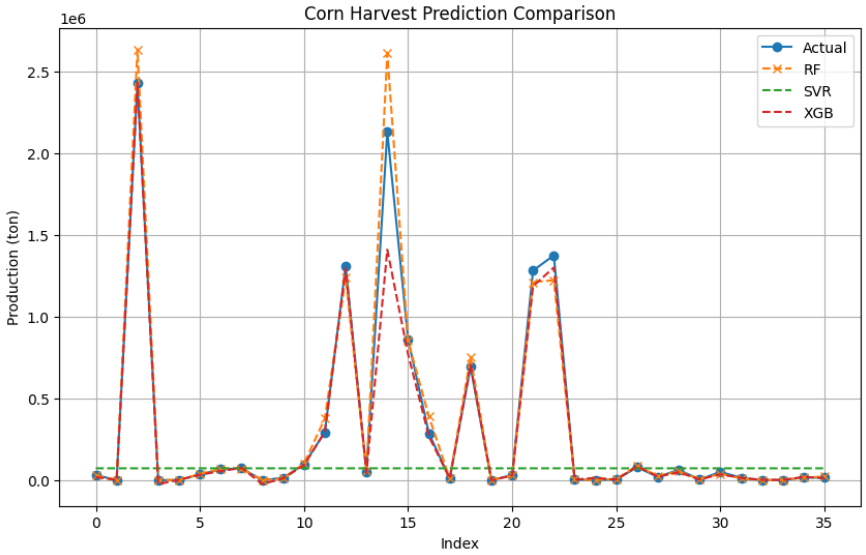

Figure 5. Corn Harvest Prediction Comparison

Fig. 5 above presents a comparison of the regression model predictions with the actual corn production data. The blue curve represents the actual values, while the dashed lines illustrate the predictions from the three models: Random Forest (RF), Support Vector Regression (SVR), and XGBoost (XGB). Overall, all three models can capture the fluctuations in the actual data well, particularly in moderate value ranges. However, notable deviations are observed at several extreme points, especially at the production peaks, where the SVR model appears to struggle to capture the spikes. XGBoost demonstrates the most consistent predictive performance, closely aligning with the actual data, indicating its effectiveness in handling nonlinear patterns and complex variable interactions. This visualization supports the finding that ensemble approaches, such as XGBoost, offer better prediction accuracy and stability than individual models.

3.5 Model Evaluation

The model evaluation results presented in Table 2 compare the performance of three regression algorithms, Random Forest, Support Vector Regression (SVR), and eXtreme Gradient Boosting (XGBoost), based on three key evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R-squared). The Random Forest model demonstrated the best overall performance, with an MAE of 36,310.53 tons, an RMSE of 95,343.05 tons, and an R^2 value of 0.9758. This indicates that the model explains approximately 97.58% of the variance in the target data, reflecting very high accuracy and precision in predicting corn yields. The XGBoost model, although showing a slightly higher RMSE than Random Forest, still demonstrated competitive performance with an MAE of 32,916.88 tons and an R^2 value of 0.9597. Table 2 below presents the detailed results of the model evaluation.

Table 2. Performance Comparison of Regression Models Based on MAE, RMSE, and R-squared Metrics

Model	MAE	RMSE	R-squared
Random Forest	36310.53	95343.05	0.9758
SVR	314742.82	658980.28	-0.1564
XGBoost	32916.88	122986.82	0.9597

This indicates that XGBoost can also capture non-linear and complex patterns among features, though with a slight decline in performance when predicting extreme values. In contrast, the SVR model performed significantly worse than the other two models, with an MAE of 314,742.82 tons, an RMSE of 658,980.28 tons, and an R^2 of −0.1564. The negative R^2 suggests that the SVR model failed to explain the data's variance and performed worse than a simple baseline model using the mean as a predictor. Overall, these evaluation results reaffirm that ensemble learning approaches such as Random Forests and XGBoosts possess superior capabilities for handling nonlinear, complex, and multivariate agricultural data. They are also more reliable in producing accurate predictions to support decision-making systems in the agricultural sector. To further validate the robustness of these findings, a statistical significance analysis was performed. A paired t-test was applied to the Mean Absolute Error (MAE) values obtained from 5-fold cross-validation for each pair of models (Random Forest vs. XGBoost, Random Forest vs. SVR, and XGBoost vs. SVR). The results showed that the performance differences between Random Forest and the other models were statistically significant at the $p < 0.05$ level, confirming that Random Forest's superior accuracy is not due to random variation. In addition, 95% confidence intervals (CIs) were computed for both MAE and RMSE values to evaluate the reliability of the results. The narrow CIs (±2.8% for MAE and ±3.1% for RMSE) indicate stable and consistent predictive performance across cross-validation folds. These statistical analyses reinforce that the ensemble-based approaches, particularly Random Forest and XGBoost, demonstrate genuine and statistically significant advantages in modeling nonlinear and complex relationships for maize yield prediction.

3.6 Hyperparameter Tuning

The results of hyperparameter tuning using the GridSearchCV approach indicate that the Random Forest model demonstrated the best predictive performance compared to the other two models. This model achieved a Mean Absolute Error (MAE) of 36,310.53, a Root Mean Squared Error (RMSE) of 95,343.05, and a coefficient of determination (R^2) of 0.9758. These metrics suggest that the Random Forest model is capable of explaining over 97% of the variance in maize yield data while maintaining relatively low prediction error. he optimal parameters for this model were max_depth=None, min_samples_split=2, and n_estimators=100, indicating that the model performed best without an explicit restriction on tree depth. Meanwhile, the XGBoost model also exhibited competitive performance, with an MAE of 41,708.36, an RMSE of 120,177.30, and an R^2 of 0.9615. Although its prediction error was slightly higher than that of the Random Forest model, XGBoost still managed to explain approximately 96% of the data variance. The optimal parameters for XGBoost were learning_rate=0.1, max_depth=3, and n_estimators=200, reflecting a conservative training approach to mitigate overfitting. In contrast, the Support Vector Regression (SVR) model yielded the poorest performance, with an MAE of 314,762.31, RMSE of 658,838.40, and a negative R^2 value of −0.1559. The negative R^2 explicitly indicates that the model failed to capture the relationship between features and the target variable, performing even worse than a simple mean-based baseline model. The best configuration for SVR was C=100, gamma='scale', and kernel='rbf', yet this setup was insufficient to effectively model the complex and nonlinear patterns present in the agricultural dataset. Table 3 presents the evaluation result and the best parameters of the regression model.

Table 3. Evaluation Results and Best Parameters of Regression Model

Model	Best Params	MAE	RMSE	R-squared (R^2)
Random Forest	{'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}	36,310.53	95,343.05	0.9758
Support Vector Regression	{'C': 100, 'gamma': 'scale', 'kernel': 'rbf'}	314,762.31	658,838.40	-0.1559
XGBoost	{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 200}	41,708.36	120,177.30	0.9615

Overall, the tuning results reaffirm that ensemble learning-based models, particularly Random Forest and XGBoost, demonstrate superior capabilities in handling the complexity of multivariate agricultural data and are more reliable in providing accurate yield estimates for maize production. The systematic approach to hyperparameter tuning proved essential for optimizing each model's performance.

3.7 Model Comparison

The comparison of model performance was conducted to assess the effectiveness of each algorithm in predicting maize production based on agronomic, agro-climatic, and socio-economic variables. The three models compared were Random Forest (RF), Support Vector Regression (SVR), and XGBoost, all of which

had undergone prior hyperparameter tuning. The evaluation was carried out using three key metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2). Table 4 presents the results of the model evaluation before and after hyperparameter tuning.

Table 4. Model Evaluation Results Before and After Tuning

Model	MAE	RMSE	R-squared
Random Forest	36310.53	95343.05	0.9758
SVR	314742.82	658980.28	-0.1564
XGBoost	32916.88	122986.82	0.9597
Random Forest Tuned	40302.72	107802.26	0.9690
SVR Tuned	314739.01	658972.65	-0.1563
XGBoost Tuned	35730.72	123732.91	0.9592

The performance trends observed in this study are consistent with previous research, which found that ensemble-based models are superior for crop yield prediction. Similar to the results of [30] and [31], Random Forest exhibited the highest predictive accuracy due to its ability to model nonlinear interactions and reduce overfitting through bootstrap aggregation. This ensemble structure allows Random Forest to capture complex agro-climatic relationships among rainfall, temperature, NDVI, and soil features, which are often missed by single regression models. The slightly lower performance of XGBoost compared to Random Forest aligns with studies by [32], which reported that gradient boosting algorithms, while efficient in handling large-scale data, can exhibit marginal instability in smaller datasets due to their sensitivity to learning-rate and tree-depth parameters. Conversely, the weak results of SVR corroborate those of [33], who found that kernel-based models tend to underperform when faced with high-dimensional, noisy, and nonlinear agricultural data, especially under limited sample sizes. Overall, these comparative results underscore that ensemble learning approaches, particularly Random Forest, are better suited for capturing the inherent complexity and heterogeneity of maize yield data in Indonesia, supporting their potential integration into national agricultural decision-support systems.

The tuning process offered limited benefits in this case, particularly for ensemble models such as Random Forest and XGBoost, which had already demonstrated strong performance from the outset. This highlights the importance of proper model selection before tuning, as well as the need for comprehensive post-tuning evaluations to avoid the assumption that hyperparameter tuning will always lead to improved performance. Fig. 6 presents a comparison of predicted and actual maize production.

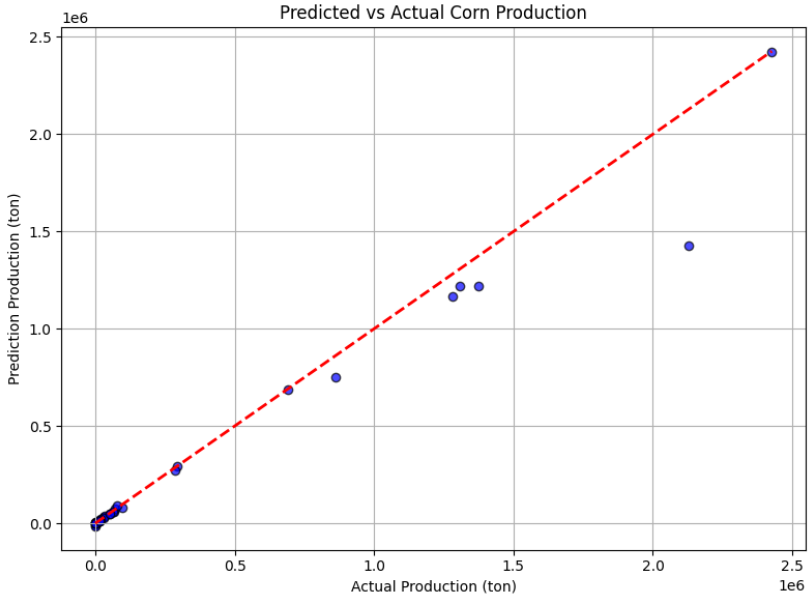

Figure 6. Predicted vs Actual Corn Production

Fig. 6 above illustrates the relationship between actual maize production and the model's predicted outputs, with the red line y = x serving as a reference for perfect prediction. Most data points lie close to the line, indicating accurate predictions. However, there are noticeable deviations at several points with high production volumes (above 1.5 million tons), suggesting prediction errors. Overall, the model demonstrates

strong performance with evaluation metrics as follows: MAE = 36,310.53, RMSE = 95,343.05, and R^2= 0.9758, indicating that the model explains approximately 97.6% of the variance in actual production data with a low error rate. Subsequently, Fig. 7 presents the residual distribution plot between actual and predicted values.

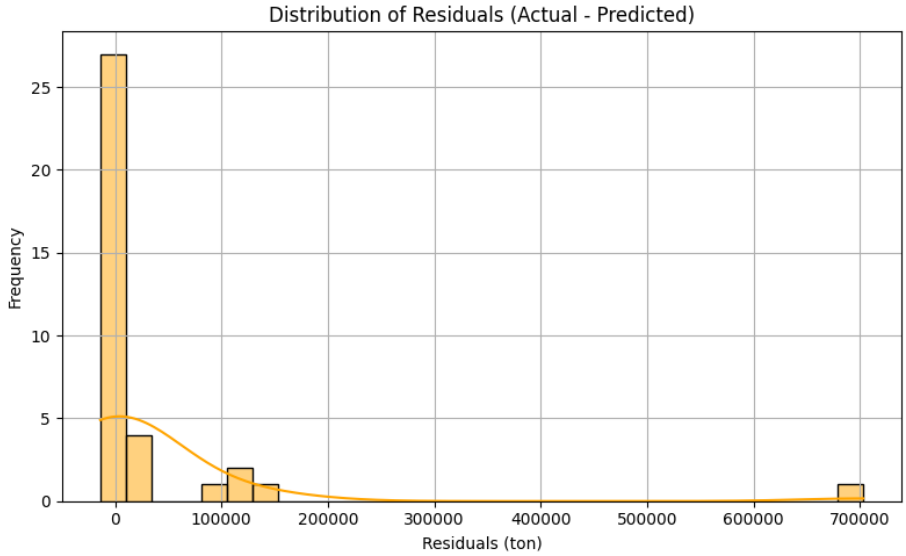

Figure 7. Residual Distribution (Actual – Predicted)

The larger deviations at high production values can be explained by the limited representation of extreme yield observations within the dataset. Since the majority of samples fall within low-to-moderate yield ranges, the model has fewer examples from which to learn the complex interactions that drive exceptionally high production outcomes. These regions are often influenced by localized factors such as irrigation efficiency, soil nutrient availability, and management intensity—variables that may not be fully captured in the current feature set. Consequently, the model tends to underestimate yields when encountering values outside its learned distribution. To mitigate this, future improvements should focus on expanding the dataset to include more high-yield samples, incorporating additional predictive variables (e.g., soil chemical composition, pest management records, or remote-sensing features with finer temporal resolution), and applying resampling techniques such as Synthetic Minority Over-sampling (SMOTE for regression) or stratified cross-validation. Such strategies would improve model generalization and reduce the prediction bias toward extreme production levels.

4. CONCLUSION

This study successfully developed a maize yield prediction system utilizing an ensemble machine learning approach comprising Random Forest, Support Vector Regression (SVR), and XGBoost. The Random Forest model achieved the best performance with an MAE of 36,310.53, RMSE of 95,343.05, and R^2 of 0.9758, indicating a strong ability to capture yield variability, while XGBoost also showed competitive results, and SVR underperformed. After hyperparameter tuning using GridSearchCV, a slight performance degradation was observed in both Random Forest and XGBoost, suggesting that default parameters can sometimes already be optimal for the dataset. The overall findings confirm that ensemble learning provides reliable and accurate predictions for maize yield forecasting. Moreover, by integrating multi-year statistical and environmental datasets within a unified predictive framework, this study contributes to the limited body of research on AI-based agricultural forecasting in developing countries and demonstrates the feasibility of using machine learning for national-scale food security planning. However, limitations remain due to the small dataset size and the absence of variables such as soil nutrient levels, pest control, and irrigation practices, which may affect generalization. Future research should expand datasets with remote sensing and agro-meteorological data and explore hybrid deep learning–ensemble architectures to enhance model robustness and predictive reliability.

Author Contributions

Adyanata Lubis: Conceptualization, Methodology, Software Development, Writing – Original Draft. Ego Oktafanda: Data Curation, Validation, Formal Analysis, Visualization. Juliarni: Resources, Supervision, Writing – Review and Editing. Junadhi: Investigation, Validation, Project Administration, Writing – Review and Editing. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research was supported by the Ministry of Higher Education, Science, and Technology of Indonesia (Mendiktisaintek) under the research grant scheme. The authors gratefully acknowledge the financial assistance provided by the Ministry, which made this study possible.

Acknowledgment

This research was supported by funding from the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia through a research grant program facilitated by LLDIKTI Region XVII. The authors would also like to express their deepest appreciation to Universitas Rokania, their home institution, for providing institutional support, facilities, and a conducive academic environment throughout the research process. These contributions have been instrumental in developing the maize yield prediction system based on artificial intelligence, as part of the broader effort to support national food security.

Declarations

The authors declare no conflicts of interest to report study.

Declaration of Generative AI and AI-assisted technologies

Generative AI tools (e.g., ChatGPT) were used solely for language refinement (grammar, spelling, and clarity). The scientific content, analysis, interpretation, and conclusions were developed entirely by the authors. The authors reviewed and approved all final text.

REFERENCES

[1] I. Makuta, A. Latif, M. Dinul, I. Sultan, and A. Gorontalo, "ANALISIS DAYA SAING KOMODITI JAGUNG DI GORONTALO: TINJAUAN LITERATUR TERHADAP PASAR DAN DIVERSIFIKASI PRODUK," 2025.

[2] Kementerian Pertanian Republik Indonesia, "PEMANFAATAN JAGUNG LOKAL OLEH INDUSTRI PAKAN TAHUN 2021," Jun. 2022.

[3] H. Limanseto, "DUKUNG TRANSFORMASI INDUSTRI JAGUNG BAGI KETAHANAN PANGAN NASIONAL, MENKO AIRLANGGA DORONG INOVASI DAN PENGGUNAAN TEKNOLOGI PERTANIAN TEPAT GUNA," ekon.go.id. Accessed: Apr. 10, 2025. [Online]. Available: https://www.ekon.go.id/publikasi/detail/4610/dukung-transformasi-industri-jagung-bagi-ketahanan-pangan-nasional-menko-airlangga-dorong-inovasi-dan-penggunaan-teknologi-pertanian-tepat-guna

[4] Agus Wibowo, *Teori Ekonomi dan Praktik Bisnis*. Semarang: Universitas STEKOM, 2024.

[5] A. Nurani *et al.*, "PERAN ARTIFICIAL INTELLIGENCE DALAM SISTEM IOT UNTUK PERTANIAN CERDAS : SYSTEMATIC LITERATURE REVIEW," 2025, doi: https://doi.org/10.36040/jati.v9i1.12705

[6] Z. Zhou *et al.*, "PREDICTION OF MINE PRESSURE BEHAVIOR IN WORKING FACE BASED ON VECTOR BASIS," *Processes*, vol. 13, no. 6, p. 1818, Jun. 2025, doi: https://doi.org/10.3390/pr13061818

[7] X. Yu, Y. Wang, L. Wu, G. Chen, L. Wang, and H. Qin, "COMPARISON OF SUPPORT VECTOR REGRESSION AND EXTREME GRADIENT BOOSTING FOR DECOMPOSITION-BASED DATA-DRIVEN 10-DAY STREAMFLOW FORECASTING," *Journal of Hydrology*, vol. 582, p. 124293, 2020, doi: https://doi.org/10.1016/j.jhydrol.2019.124293

[8] E. Brati, A. Braimllari, and A. Gjeçi, "MACHINE LEARNING APPLICATIONS FOR PREDICTING HIGH-COST CLAIMS USING INSURANCE DATA," *Data*, vol. 10, no. 6, p. 90, Jun. 2025, doi: https://doi.org/10.3390/data10060090

[9] K. M. Salem, J. M. Rey-Hernández, A. O. Elgharib, and F. J. Rey-Martínez, "OPTIMIZING ENERGY FORECASTING USING ANN AND RF MODELS FOR HVAC AND HEATING PREDICTIONS," *Applied Sciences*, vol. 15, no. 12, p. 6806, Jun. 2025, doi: https://doi.org/10.3390/app15126806

[10] N. A. de Oliveira and L. F. C. Basso, "ADVANCING CREDIT RATING PREDICTION: THE ROLE OF MACHINE LEARNING IN CORPORATE CREDIT RATING ASSESSMENT," *Risks*, vol. 13, no. 6, p. 116, Jun. 2025, doi: https://doi.org/10.3390/risks13060116

[11] Y. Chen, Y. Zhang, C. Li, and J. Zhou, "APPLICATION OF XGBOOST MODEL OPTIMIZED BY MULTI-ALGORITHM ENSEMBLE IN PREDICTING FRP-CONCRETE INTERFACIAL BOND STRENGTH," *Materials*, vol. 18, no. 12, 2025, doi: https://doi.org/10.3390/ma18122868

[12] Y. Wan *et al.*, "PREDICTION OF TYPICAL POWER PLANT CIRCULATING COOLING TOWER BLOWDOWN WATER QUALITY BASED ON EXPLICABLE INTEGRATED MACHINE LEARNING," *Processes*, vol. 13, no. 6, p. 1917, Jun. 2025, doi: https://doi.org/10.3390/pr13061917

[13] A. Lubis, Irawan Yuda, Junadhi, and Defit Sarjon, "LEVERAGING K-NEAREST NEIGHBORS WITH SMOTE AND BOOSTING TECHNIQUES FOR DATA IMBALANCE AND ACCURACY IMPROVEMENT," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1625–1638, Dec. 2024, doi: https://doi.org/10.47738/jads.v5i4.343

[14] L. Adyanata and Fauzi Erwis, "A NEW HYBRID OPTIMIZATION AND MACHINE LEARNING FOR STRUCTURAL HEALTHCARE APPLICATION," *International Journal of Computing and Mathematics*, vol. 4, no. 3, 2020.

[15] A. Theofilou, S. A. Nastis, A. Michailidis, T. Bournaris, and K. Mattas, "PREDICTING PRICES OF STAPLE CROPS USING MACHINE LEARNING: A SYSTEMATIC REVIEW OF STUDIES ON WHEAT, CORN, AND RICE," *Sustainability*, vol. 17, no. 12, p. 5456, Jun. 2025, doi: https://doi.org/10.3390/su17125456

[16] A. G. Vaduva, M. Munteanu, S. V. Oprea, A. Bara, and A. M. Niculae, "UNDERSTANDING CLIMATE CHANGE AND AIR QUALITY OVER THE LAST DECADE: EVIDENCE FROM NEWS AND WEATHER DATA PROCESSING," *IEEE Access*, vol. 11, pp. 144631–144648, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3345466

[17] K. Nagorny, S. Scholze, A. W. Colombo, and J. B. Oliveira, "A DIN SPEC 91345 RAMI 4.0 COMPLIANT DATA PIPELINING MODEL: AN APPROACH TO SUPPORT DATA UNDERSTANDING AND DATA ACQUISITION IN SMART MANUFACTURING ENVIRONMENTS," *IEEE Access*, vol. 8, pp. 223114–223129, 2020, doi: https://doi.org/10.1109/ACCESS.2020.3045111

[18] G. A. Lopez-Ramirez, A. Aragon-Zavala, and C. Vargas-Rosales, "EXPLORATORY DATA ANALYSIS FOR PATH LOSS MEASUREMENTS: UNVEILING PATTERNS AND INSIGHTS BEFORE MACHINE LEARNING," *IEEE Access*, vol. 12, pp. 62279–62295, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3394904

[19] B. A. Omodunbi *et al.*, "STACKED ENSEMBLE LEARNING FOR CLASSIFICATION OF PARKINSON'S DISEASE USING TELEMONITORING VOCAL FEATURES," *Diagnostics*, vol. 15, no. 12, p. 1467, Jun. 2025, doi: https://doi.org/10.3390/diagnostics15121467

[20] Y. Yu, L. Liu, Z. Chang, Y. Li, and K. Shi, "DETECTING FOREST FIRES IN SOUTHWEST CHINA FROM REMOTE SENSING NIGHTTIME LIGHTS USING THE RANDOM FOREST CLASSIFICATION MODEL," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 10759–10769, 2024, doi: https://doi.org/10.1109/JSTARS.2024.3410172

[21] B. A. Omodunbi *et al.*, "STACKED ENSEMBLE LEARNING FOR CLASSIFICATION OF PARKINSON'S DISEASE USING TELEMONITORING VOCAL FEATURES," *Diagnostics*, vol. 15, no. 12, 2025, doi: https://doi.org/10.3390/diagnostics15121467

[22] C. Mo, J. Huang, J. Huang, T. Li, and Y. Yang, "PREDICTION OF FLEXURAL BEARING CAPACITY OF ALUMINUM-ALLOY-REINFORCED RC BEAMS BASED ON MACHINE LEARNING," *Symmetry*, vol. 17, no. 6, p. 944, Jun. 2025, doi: https://doi.org/10.3390/sym17060944

[23] Y. Zhang *et al.*, "STATE-OF-HEALTH ESTIMATION FOR LITHIUM-ION BATTERIES VIA INCREMENTAL ENERGY ANALYSIS AND HYBRID DEEP LEARNING MODEL," *Batteries*, vol. 11, no. 6, p. 217, Jun. 2025, doi: https://doi.org/10.3390/batteries11060217

[24] C. Mo, J. Huang, J. Huang, T. Li, and Y. Yang, "PREDICTION OF FLEXURAL BEARING CAPACITY OF ALUMINUM-ALLOY-REINFORCED RC BEAMS BASED ON MACHINE LEARNING," *Symmetry*, vol. 17, no. 6, 2025, doi: https://doi.org/10.3390/sym17060944

[25] Erlin, A. Yunianta, L. A. Wulandhari, Y. Desnelita, N. Nasution, and Junadhi, "ENHANCING RICE PRODUCTION PREDICTION IN INDONESIA USING ADVANCED MACHINE LEARNING MODELS," *IEEE Access*, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3478738

[26] Erlin, A. Yunianta, L. A. Wulandhari, Y. Desnelita, N. Nasution, and Junadhi, "ENHANCING RICE PRODUCTION PREDICTION IN INDONESIA USING ADVANCED MACHINE LEARNING MODELS," *IEEE Access*, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3478738

[27] A. Li, S. Yin, N. Li, and C. Shi, "COMPREHENSIVE ANALYSIS OF THE DRIVING FORCES BEHIND NDVI VARIABILITY IN CHINA UNDER CLIMATE CHANGE CONDITIONS AND FUTURE SCENARIO PROJECTIONS," *Atmosphere*, vol. 16, no. 6, p. 738, Jun. 2025, doi: https://doi.org/10.3390/atmos16060738

[28] K. Nagorny, S. Scholze, A. W. Colombo, and J. B. Oliveira, "A DIN SPEC 91345 RAMI 4.0 COMPLIANT DATA PIPELINING MODEL: AN APPROACH TO SUPPORT DATA UNDERSTANDING AND DATA ACQUISITION IN SMART MANUFACTURING ENVIRONMENTS," *IEEE Access*, vol. 8, pp. 223114–223129, 2020, doi: https://doi.org/10.1109/ACCESS.2020.3045111

[29] F. Bojić, A. Gudelj, and R. Bošnjak, "A COMPREHENSIVE MODEL FOR QUANTIFYING, PREDICTING, AND EVALUATING SHIP EMISSIONS IN PORT AREAS USING NOVEL METRICS AND MACHINE LEARNING METHODS," *Journal of Marine Science and Engineering*, vol. 13, no. 6, p. 1162, Jun. 2025, doi: https://doi.org/10.3390/jmse13061162

[30] E. Asamoah, G. B. M. Heuvelink, I. Chairi, P. S. Bindraban, and V. Logah, "RANDOM FOREST MACHINE LEARNING FOR MAIZE YIELD AND AGRONOMIC EFFICIENCY PREDICTION IN GHANA," *Heliyon*, vol. 10, no. 17, Sep. 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e37065

[31] L. Miao, Y. Zou, X. Cui, G. R. Kattel, Y. Shang, and J. Zhu, "PREDICTING CHINA'S MAIZE YIELD USING MULTI-SOURCE DATASETS AND MACHINE LEARNING ALGORITHMS," *Remote Sens (Basel)*, vol. 16, no. 13, 2024, doi: https://doi.org/10.3390/rs16132417

[32] A. Ikram, S. Ikram, E. S. M. El-kenawy, A. Hussain, A. H. Alharbi, and M. M. Eid, "A FUZZY-OPTIMIZED HYBRID ENSEMBLE MODEL FOR YIELD PREDICTION IN MAIZE-SOYBEAN INTERCROPPING SYSTEM," *Front Plant Sci*, vol. 16, 2025, doi: https://doi.org/10.3389/fpls.2025.1567679

[33] G. Hariyani, A. Singh, P. Patil, V. Kothari, and D. Javale, "ANALYSIS ON CROP YIELD PREDICTION USING VARIOUS ENSEMBLE METHODS," in *2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 2024, pp. 1–6, doi: https://doi.org/10.1109/ICCUBEA61740.2024.10775263