**BIMA**
**BULLETIN OF INTELLIGENT MACHINES AND ALGORITHMS**
Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i3.17

# Comparative Analysis of Machine Learning Regression Models for Paddy Yield Prediction

**Chery Cardinawati Sitohang[1*], Fitri Kinkin[2], Budiman[3]**

Information System[1,2]
Informatics[3]
Universitas Informatika dan Bisnis Indonesia, Bandung, Indonesia[1,2,3]
https://unibi.ac.id[1,2,3]
cherycardinawatisitohang23@student.unibi.ac.id[1*], fitrikinkin23@student.unibi.ac.id[2], budiman@unibi.ac.id[3]

**Abstract**. Accurate paddy yield prediction is essential to support food security, agricultural planning, and data-driven decision-making. The increasing availability of agricultural data has encouraged the adoption of machine learning approaches to overcome the limitations of conventional yield estimation methods. This study presents a comparative analysis of five regression-based machine learning algorithms—Linear Regression, K-Nearest Neighbors Regressor, Decision Tree Regressor, Random Forest Regressor, and Support Vector Regression—for paddy yield prediction. The experiments were conducted using the Paddy dataset from the UCI Machine Learning Repository, which consists of 2,789 samples and 45 variables (44 input features and 1 target variable). The dataset was preprocessed through data cleaning, feature standardization, and an 80:20 train–test split. Model performance was evaluated using Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and the coefficient of determination ($R^2$). Experimental results show that Linear Regression achieved the best overall performance with an $R^2$ value of 0.9896 and an RMSE of 942.09, indicating strong predictive accuracy and stability. Despite its simplicity, Linear Regression outperformed more complex models, suggesting that the underlying relationships between input variables and paddy yield in the dataset are predominantly linear. These findings highlight the importance of systematic model evaluation and demonstrate that simpler regression models can remain effective and interpretable for practical paddy yield prediction and agricultural decision support systems.

**Keywords:** Paddy yield prediction, machine learning, regression algorithms, agricultural data, Linear Regression.

## 1. INTRODUCTION

The agricultural sector plays a strategic role in ensuring national food security, particularly for rice, which serves as the primary staple food for the majority of Indonesia's population[1]. Paddy production directly influences rice availability, farmers' economic stability, food price control, and government policy planning in the agricultural sector [7]. Consequently, accurate estimation and prediction of paddy yield are essential to support sustainable agricultural development and effective decision-making. However, paddy yield productivity remains challenged by various complex factors, including weather variability, heterogeneous soil conditions, suboptimal fertilizer usage, and limitations in conventional yield prediction methods.

Traditional approaches to paddy yield estimation generally rely on statistical analysis or empirical observations, which often struggle to capture the nonlinear and complex relationships among multiple influencing variables. The increasing availability of agricultural data—such as rainfall, temperature, humidity, soil characteristics, and cultivation practices—has encouraged the adoption of data-driven approaches for yield prediction. In this context, machine learning has emerged as a promising solution due to its ability to model complex patterns and relationships within large and heterogeneous datasets more effectively than conventional methods[2], [3].

Machine learning techniques have been widely applied in agricultural yield prediction, particularly through regression-based models that aim to estimate crop yield as a continuous variable. Algorithms such as Linear Regression, K-Nearest Neighbor (KNN), Decision Tree, Random Forest, and Support Vector Regression (SVR) have demonstrated potential in modeling agricultural data due to their flexibility in handling both linear and nonlinear relationships [4], [5], [6]. Nevertheless, the performance of each algorithm varies significantly depending on data characteristics, feature selection, and preprocessing techniques, indicating that no single model can be universally considered optimal for all agricultural datasets.

Several previous studies have investigated machine learning approaches for crop yield prediction, reporting varying results depending on dataset characteristics and modeling techniques. For instance, Joshua et al. [2] applied several machine learning models to predict paddy yield in a specific region of India, while Cao et al. [3] integrated multi-source environmental data with machine learning and deep learning

Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i3.17

methods for large-scale rice yield prediction. Although these studies demonstrate the effectiveness of machine learning for agricultural applications, many of them focus on specific datasets or regions, and often evaluate only a limited number of algorithms. Consequently, the comparative performance of different regression-based machine learning models under a unified experimental framework remains insufficiently explored.

Therefore, this study aims to address this research gap by conducting a systematic comparative analysis of five regression-based machine learning algorithms—Linear Regression, K-Nearest Neighbors Regressor, Decision Tree Regressor, Random Forest Regressor, and Support Vector Regression—for paddy yield prediction. All models are evaluated using the same dataset, preprocessing procedures, and evaluation metrics within a unified experimental framework to ensure a fair and consistent comparison. The findings of this study provide insights into the effectiveness and interpretability of different regression models and support the development of reliable machine learning-based decision-support systems for agricultural yield prediction.

This study contributes to the existing literature on agricultural yield prediction in several ways. First, it provides a systematic comparative evaluation of five widely used machine learning regression algorithms—Linear Regression, K-Nearest Neighbors Regressor, Decision Tree Regressor, Random Forest Regressor, and Support Vector Regression—for paddy yield prediction. Second, the study employs a unified experimental framework in which all models are trained and evaluated using the same dataset, preprocessing procedures, and evaluation metrics to ensure a fair and consistent comparison. Third, the study identifies Linear Regression as the most reliable and interpretable model for the given dataset, demonstrating that simpler regression approaches can remain highly effective when the underlying relationships among agricultural variables are predominantly linear. These findings provide useful insights for researchers and practitioners in selecting appropriate machine learning models for paddy yield prediction and agricultural decision-support systems.

## 2. RELATED WORK

Machine learning-based approaches for crop yield prediction have been extensively explored in recent years due to their effectiveness in handling complex agricultural datasets. Joshua *et al.* [2] investigated multiple machine learning models for paddy yield prediction in Tamil Nadu, India, and reported that ensemble-based models achieved superior prediction accuracy compared to traditional regression methods. Their findings emphasize the importance of selecting appropriate algorithms based on data characteristics. Cao *et al.* [3] integrated multi-source data, including meteorological and environmental variables, to predict rice yield across China using machine learning and deep learning approaches. The study demonstrated that advanced learning models significantly outperformed conventional statistical techniques, particularly in capturing spatial and temporal variability in large-scale agricultural data.

Weather-based paddy yield prediction has also gained significant attention. Sakthipriya and Chandrakumar [4] applied several regression-based machine learning algorithms using meteorological parameters and showed that nonlinear models achieved better performance than linear models. Similarly, Satpathi *et al.* [5] compared statistical and machine learning techniques for rice yield forecasting and concluded that machine learning models provided more robust predictions under varying climatic conditions.

Comprehensive reviews have further highlighted the growing adoption of machine learning in agricultural yield prediction. Jabed and Murad [8] presented an extensive review of machine learning and deep learning approaches for crop yield prediction, identifying regression-based models as among the most commonly used techniques due to their interpretability and predictive capability. Their study also emphasized the need for comparative evaluations across multiple algorithms to identify the most effective models for specific crops and datasets.

Several comparative studies have focused on regression models for crop yield estimation. Panigrahi *et al.* [9] and Jorvekar *et al.* [10] conducted comparative analyses of supervised regression algorithms and reported that ensemble methods, such as Random Forest, often outperform single-model approaches. However, they also noted that simpler models like Linear Regression remain competitive under certain conditions, particularly when data relationships are predominantly linear.

Although previous studies have demonstrated the effectiveness of machine learning for crop yield prediction, most research either focuses on specific algorithms or different crops and regions. Consequently, comparative studies specifically targeting paddy yield prediction using multiple regression algorithms within a unified experimental framework remain limited. This research addresses this gap by providing a systematic comparison of five regression-based machine learning algorithms using the same dataset and evaluation criteria.

Several studies have also explored alternative data sources and modeling approaches for rice and paddy yield prediction. For instance, remote sensing data and satellite imagery have been utilized to capture spatial variability in crop conditions, showing promising results in large-scale yield estimation [11], [12]. In addition, image-based approaches using rice panicle features have been investigated to enhance prediction accuracy at the plant level [13].

**TABLE 1.** Comparison of previous studies on crop yield prediction

| Author | Dataset | Method | Key Results | Limitations |
|---|---|---|---|---|
| [2] | Paddy yield data from Tamil Nadu, India | Multiple ML models | Ensemble models achieved higher prediction accuracy | Study limited to a specific regional dataset |
| [3] | Multi-source environmental and meteorological | Machine learning and deep | Deep learning models improved large-scale | High model complexity and limited interpretability |

Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i3.17

| Author | Dataset | Method | Key Results | Limitations |
|---|---|---|---|---|
| | data across China | learning models | yield prediction | |
| [4] | Weather-based paddy dataset | Regression-based ML algorithms | Nonlinear models outperformed linear models | Limited comparison of different algorithms |
| [5] | Agricultural yield dataset from India | Statistical vs ML models | ML models provided better prediction accuracy | Evaluation focused on specific climatic conditions |
| [9] | Crop yield dataset | Supervised regression algorithms | Ensemble methods performed well | Lack of standardized experimental settings |
| [10] | Agricultural yield dataset | Regression models | Random Forest showed strong performance | Limited exploration of model interpretability |

Although previous studies demonstrate the effectiveness of machine learning for crop yield prediction, several limitations remain evident, as shown in Table 1. Many studies rely on region-specific datasets, which may limit the generalizability of their findings to other agricultural environments. In addition, some research focuses on a limited set of algorithms or emphasizes complex models such as deep learning and ensemble methods without systematically evaluating simpler regression approaches under the same experimental conditions. Differences in datasets, preprocessing techniques, and evaluation metrics also make it difficult to fairly compare model performance across studies. Therefore, a unified experimental framework that evaluates multiple regression-based machine learning algorithms using the same dataset and evaluation criteria is necessary to provide a clearer understanding of their comparative performance in paddy yield prediction.

## 3. METHODS

This study adopts a systematic machine learning methodology consisting of four main stages: data understanding, data preprocessing, model development, and model evaluation. The overall research framework is designed to ensure methodological rigor and reproducibility in predicting paddy yield using regression-based machine learning algorithms.

### 3.1. Data Collection

The dataset used in this research was obtained from the UCI Machine Learning Repository, specifically the Paddy Dataset. This dataset contains agricultural data related to paddy production and environmental factors that influence crop yield.

The dataset consists of 2,789 records (samples) and 45 variables**,** which include 44 input features and 1 target variable**.** Each record represents a single agricultural observation describing cultivation conditions and environmental parameters that affect paddy productivity.

The features contained in the dataset include agronomic attributes such as cultivated area, seed rate, fertilizer application at different growth stages, pesticide usage, and environmental conditions such as rainfall, temperature, humidity, and wind speed. The target variable represents paddy yield measured in kilograms, which is used as the prediction output in this study. The dataset was downloaded from the UCI repository in CSV format and consists of 2,789 rows and 45 columns**.** An overview of the dataset variables used in this research is presented in

**TABLE 2.** Overview of Paddy Dataset Variables

| Category | Variables |
|---|---|
| Agronomic attributes | Agriblock, Hectares, Variety, Seed Rate |
| Fertilizer application | DAP_20Days, Urea_40Days, Potash_50Days, Micronutrients_70Days |
| Pest management | Pets_60Days |
| Weather conditions | Rainfall, Temperature (Min), Temperature (Max), Wind Speed, Wind Direction, Relative Humidity |
| Target variable | Paddy Yield (kg) |

Table 2 summarizes the main variables contained in the paddy dataset. These variables represent both agronomic practices and environmental conditions that potentially influence paddy yield production. The combination of these attributes provides a comprehensive dataset for developing machine learning models for yield prediction.

### 3.2. Data Understanding

The dataset used in this study was obtained from the paddy yield project dataset, which contains agricultural attributes related to paddy production. The dataset includes several numerical features representing environmental and agronomic conditions, such as climate-related variables and cultivation parameters, while the target variable represents paddy yield as a continuous value.

An initial data exploration process was conducted to understand the structure, distribution, and characteristics of the dataset. This stage aimed to identify potential issues such as missing values, data imbalance, and outliers that could affect model performance. Understanding the data characteristics is essential, as agricultural data are often heterogeneous and influenced by complex, nonlinear relationships between variables.

### 3.3. Data Preprocessing.

Data preprocessing was performed to improve data quality and ensure compatibility with machine learning algorithms. The preprocessing stage began with data cleaning to handle missing values and inconsistencies. Numerical missing values were treated using statistical imputation methods, while extreme outliers were examined and handled to prevent distortion of the learning process.

After data cleaning, feature scaling was applied using the StandardScaler method to normalize numerical variables. This process transforms each feature to have zero mean and unit variance, which is mathematically defined as Eq. (1).

Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i3.17

$$z = \frac{x-\mu}{\sigma} \tag{1}$$

where $x$ is the original feature value, $\mu$ is the mean of the feature, and $\sigma$is the standard deviation. Feature standardization is particularly important for distance-based and kernel-based algorithms to ensure that all variables contribute proportionally during model training.

The preprocessed dataset was then divided into training and testing subsets using a random split approach with an 80:20 ratio. The training data were used to build the prediction models, while the testing data were reserved for independent performance evaluation to assess the generalization capability of each algorithm.

### 3.4. Modeling

The modeling stage involved the implementation and comparison of five supervised machine learning regression algorithms: Linear Regression, K-Nearest Neighbors Regressor, Decision Tree Regressor, Random Forest Regressor, and Support Vector Regression. These algorithms were selected to represent diverse learning paradigms, including linear modeling, instance-based learning, tree-based learning, ensemble methods, and kernel-based approaches. This selection enables a comprehensive evaluation of model behavior on agricultural data, which often exhibit nonlinear and complex patterns.

Linear Regression models the relationship between independent variables and the target variable by fitting a linear equation, which can be expressed as Eq (2).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{2}$$

where $y$denotes the predicted paddy yield, $x_i$represents the input features, and $\beta_i$are the regression coefficients.

The K-Nearest Neighbors Regressor predicts the target value based on the average of the $k$nearest data points in the feature space. The similarity between data points is measured using the Euclidean distance as Eq (3).

$$d(x_i, x_j) = \sqrt{\textstyle\sum_{m=1}^{n} (x_{im} - x_{jm})^2 \mathrm{m}} \tag{3}$$

Decision Tree Regressor predicts output values by recursively partitioning the feature space into smaller regions based on decision rules that minimize prediction error. This approach enables the model to capture nonlinear relationships and feature interactions within the data.

Random Forest Regressor is an ensemble learning method that constructs multiple decision trees using random subsets of data and features. The final prediction is obtained by averaging the outputs of all trees, which can be formulated as Eq (4).

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} y_t \tag{4}$$

where $T$ is the total number of trees and $y_t$is the prediction from the $t$-th tree. This ensemble strategy improves prediction accuracy and reduces overfitting.

Support Vector Regression applies the principle of structural risk minimization by finding a function that deviates from the actual target values by no more than a predefined margin $\varepsilon$. The SVR optimization problem is defined Eq (5).

$$\min \frac{1}{2} \parallel w \parallel^2 \tag{5}$$

subject to:

$$\mid y_i - (w \cdot x_i + b) \mid \leq \varepsilon \tag{6}$$

This approach allows SVR to model complex nonlinear relationships using kernel functions.

All models were developed using the Python programming language and implemented with the scikit-learn library to ensure reproducibility and consistency.

### 3.5. Evaluation

The evaluation stage was conducted to assess the predictive performance of each regression model on unseen data. Several evaluation metrics were employed, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination ($R^2$). MAE measures the average magnitude of prediction errors, defined Eq (7).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid \tag{7}$$

MSE and RMSE emphasize larger prediction errors and are expressed Eq (8).

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$RMSE = \sqrt{MSE} \tag{8}$$

The $R^2$metric evaluates how well the predicted values explain the variance in the actual paddy yield data:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{9}$$

These metrics were selected to provide a comprehensive evaluation of model accuracy, error sensitivity, and goodness of fit. The comparison of results across models enables the identification of the most effective algorithm for paddy yield prediction based on the given dataset.

## 4. RESULTS AND DISCUSSIONS

### 4.1. Results

The results of the simulation were obtained through a comprehensive evaluation of five regression algorithms,

BIMA
BULLETIN OF INTELLIGENT MACHINES AND ALGORITHMS
Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i3.17

namely Linear Regression, K-Nearest Neighbors Regressor, Decision Tree Regressor, Random Forest Regressor, and Support Vector Regression, in predicting paddy yield. The evaluation focused on measuring predictive accuracy and error magnitude using the coefficient of determination ($R^2$) and Root Mean Squared Error (RMSE). These metrics were selected to ensure an objective assessment of model performance in capturing both linear and non-linear relationships within the agricultural dataset.

**TABLE 3.** Comparative Results of All Regression Models

| Model | R2 | RMSE |
|---|---|---|
| Linear Regression | 0.989633 | 942.091280 |
| KNN Regressor | 0.988509 | 991.850281 |
| Decision Tree | 0.988788 | 979.702955 |
| Random Forest | 0.989037 | 968.788493 |
| SVR | -0.012141 | 9308.561551 |

Table 3 presents the comparative results of all regression models. As shown in the table, Linear Regression achieved the highest $R^2$ value and the lowest RMSE, indicating superior performance in explaining yield variability and minimizing prediction error. This result demonstrates that the relationship between the selected input variables and paddy yield can be effectively modeled using a linear approach.

K-Nearest Neighbors Regressor showed moderate predictive capability but was highly sensitive to data distribution and neighborhood selection. Decision Tree Regressor was able to capture non-linear patterns; however, its tendency to overfit resulted in increased prediction errors. Random Forest Regressor improved stability by aggregating multiple decision trees, yet its performance remained slightly below that of Linear Regression. Support Vector Regression produced consistent results but required extensive parameter tuning to achieve comparable accuracy.

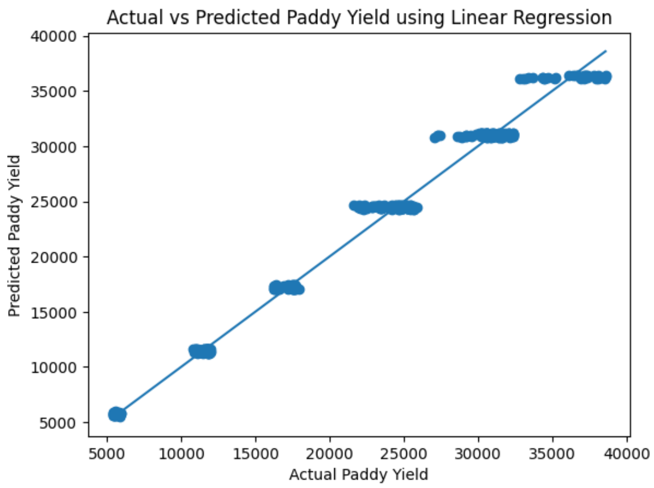


**Fig.1.** Actual vs predicted paddy yield using linear regression

To further analyze model behavior, Figure 1 illustrates the relationship between actual and predicted paddy yield using the Linear Regression model. The scatter points closely follow the diagonal reference line, indicating a strong agreement between predicted values and observed data. This visual evidence supports the numerical results reported in Table 2, confirming the robustness of Linear Regression in yield prediction.

These results are consistent with previous studies that reported strong performance of regression-based approaches for crop yield prediction. Setiya et al. demonstrated that multiple linear regression models achieved competitive accuracy in predicting rice yield using weather-related variables, emphasizing the effectiveness of linear relationships in agricultural data[14]. Similarly, Ansarifar et al. proposed an interaction regression model that highlighted the importance of capturing interactions among input variables to improve yield prediction accuracy[15] . These findings support the results obtained in this study, where Linear Regression exhibited robust performance and reliable generalization capability.

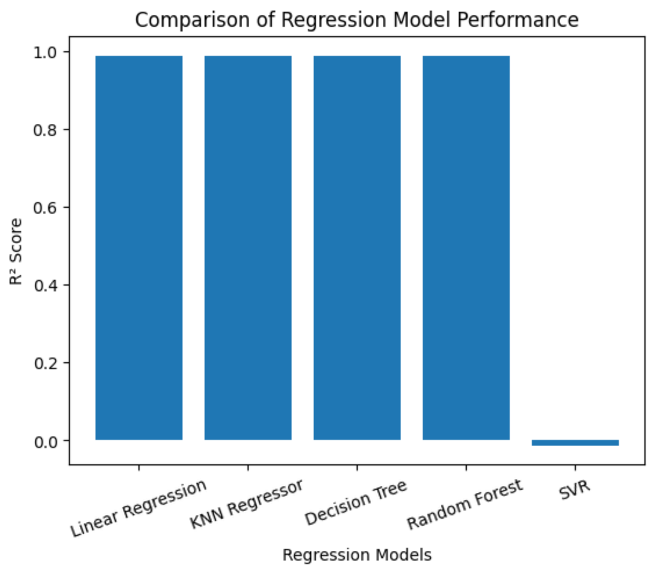


**Fig.2.** Model performance comparison

A comparative visualization of model performance based on $R^2$ values is presented in Figure 2. The figure clearly highlights the superiority of Linear Regression over other regression models, demonstrating its ability to capture the dominant trend of paddy yield data more effectively.

BIMA
BULLETIN OF INTELLIGENT MACHINES AND ALGORITHMS
Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i3.17

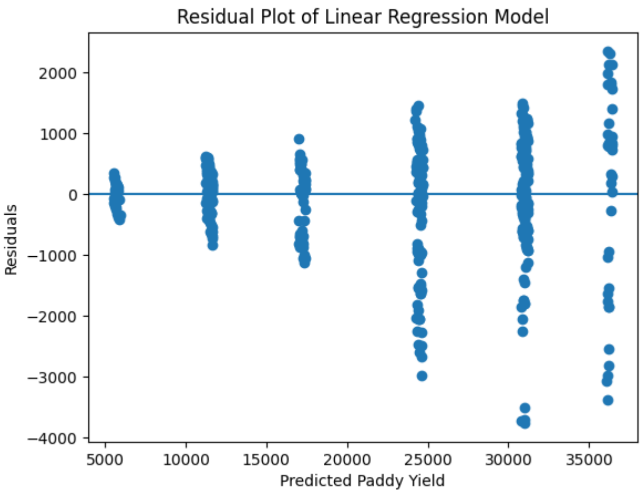


**Fig.3.** Residual plot

In addition, Figure 3 displays the residual distribution of the Linear Regression model. The residuals are symmetrically distributed around zero with no discernible pattern, indicating that the model does not suffer from systematic bias and maintains good generalization capability.

### 4.2. Discussions

The findings of this study indicate that Linear Regression provides the most reliable and interpretable solution for paddy yield prediction among the evaluated models. Despite the availability of more complex algorithms such as Random Forest and Support Vector Regression, the results suggest that increased model complexity does not necessarily lead to improved predictive performance in this context.

The strong performance of Linear Regression can be attributed to the underlying structure of the dataset, where paddy yield is predominantly influenced by linear relationships between agronomic factors. Moreover, the interpretability of Linear Regression makes it particularly suitable for agricultural decision-support systems, where transparency and ease of understanding are essential for practical implementation.

The findings of this study indicate that Linear Regression provides the most reliable and interpretable solution for paddy yield prediction among the evaluated models. Despite the availability of more complex algorithms such as Random Forest and Support Vector Regression, the results suggest that increased model complexity does not necessarily lead to improved predictive performance in this context, particularly when the underlying relationships among agronomic variables are predominantly linear [15].

Compared to tree-based and instance-based models, Linear Regression demonstrated greater stability and lower variance, as evidenced by its residual distribution in Figure 3. This characteristic is crucial for ensuring consistent predictions across different planting seasons and environmental conditions.

Overall, the simulation results confirm that Linear Regression not only achieves the highest predictive accuracy but also provides a balanced trade-off between performance, interpretability, and computational efficiency. Therefore, this model is well suited for supporting data-driven decision-making in paddy yield estimation and agricultural planning.

## 5. CONCLUSIONS

This study presented a comparative analysis of several machine learning regression algorithms for predicting paddy yield using a unified experimental framework. Five widely used regression models—Linear Regression, K-Nearest Neighbors Regressor, Decision Tree Regressor, Random Forest Regressor, and Support Vector Regression—were evaluated using the same dataset, preprocessing procedures, and evaluation metrics to ensure a consistent comparison of predictive performance. The experimental results indicate that Linear Regression achieved the best overall performance among the evaluated models, as reflected by its highest coefficient of determination ($R^2$) and relatively lower prediction error. These results suggest that the relationships between the input variables and paddy yield in the dataset can be effectively represented using a linear model. While more complex algorithms such as Random Forest and Support Vector Regression are capable of modeling nonlinear patterns, their performance in this study did not surpass that of the simpler regression approach.

The findings of this research highlight that simpler and more interpretable models can remain effective for agricultural yield prediction when the underlying relationships among variables exhsibit predominantly linear patterns. In addition to providing a comparative evaluation of multiple regression algorithms under a consistent experimental setup, this study offers practical insights for researchers and practitioners in selecting suitable machine learning models for paddy yield prediction and agricultural decision-support systems.

Future research may extend this work by incorporating additional datasets from different regions, exploring feature selection techniques, and evaluating advanced machine learning or hybrid models to further improve predictive performance and generalization capability in diverse agricultural environments.

## REFERENCES

[1] "A. Ashari, E. Ariningsih, S. Saptana, H. P. Saliem, and P. Laksono, 'An overview of reducing rice yield loss to improve national food security,' BIO Web of Conferences, vol. 119, p. 01011, 2024.".
[2] "V. Joshua, S. M. Priyadharson, and R. Kannadasan, 'Exploration of machine learning approaches for paddy yield prediction in eastern part of Tamilnadu,' Agronomy, vol. 11, no. 10, p. 2068, 2021.".
[3] "J. Cao et al., 'Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches,' Agricultural and Forest Meteorology, vol. 297, p. 108275, 2021.".

Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i3.17

[4] "D. Sakthipriya and T. Chandrakumar, 'Weather based paddy yield prediction using machine learning regression algorithms,' Journal of Agrometeorology, vol. 26, no. 3, pp. 344–348, 2024.".

[5] "A. Satpathi et al., 'Comparative analysis of statistical and machine learning techniques for rice yield forecasting for Chhattisgarh, India,' Sustainability, vol. 15, no. 3, p. 2786, 2023.".

[6] "I. Michael, 'Evaluation metrics and benchmarking for predictive accuracy,' 2025.".

[7] "A. Badshah, B. Y. Alkazemi, F. Din, K. Z. Zamli, and M. Haris, 'Crop classification and yield prediction using robust machine learning models for agricultural sustainability,' IEEE Access, vol. 12, pp. 1–15, 2024,".

[8] "M. A. Jabed and M. A. A. Murad, 'Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability,' Heliyon, vol. 10, no. 24, 2024.".

[9] "B. Panigrahi, K. C. R. Kathala, and M. Sujatha, 'A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models,' Procedia Computer Science, vol. 218, pp. 2684–2693, 2023.".

[10] "P. P. Jorvekar, S. K. Wagh, and J. R. Prasad, 'Predictive modeling of crop yields: A comparative analysis of regression techniques for agricultural yield prediction,' Agricultural Engineering International: CIGR Journal, vol. 26, no. 2, 2024.".

[11] "M. Kuradusenge et al., 'Crop yield prediction using machine learning models: Case of Irish potato and maize,' Agriculture, vol. 13, no. 1, p. 225, 2023.".

[12] "M. D. Islam et al., 'Rapid rice yield estimation using integrated remote sensing and meteorological data and machine learning,' Remote Sensing, vol. 15, no. 9, p. 2374, 2023.".

[13] "B. Kumar Pankaj et al., 'Paddy yield prediction based on 2D images of rice panicles using regression techniques,' The Visual Computer, vol. 40, no. 6, pp. 4457–4471, 2024.".

[14] "P. Setiya, A. Satpathi, and A. S. Nain, 'Predicting rice yield based on weather variables using multiple linear, neural networks, and penalized regression models,' Theoretical and Applied Climatology, vol. 154, no. 1–2, pp. 365–375, 2023.".

[15] "J. Ansarifar, L. Wang, and S. V. Archontoulis, 'An interaction regression model for crop yield prediction,' Scientific Reports, vol. 11, no. 1, p. 17754, 2021.".