Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 3, Juni 2025, Page. 1431-1442
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4681

Multivariate Forecasting of Paddy Production: A Comparative Study of Machine Learning Models

Feri Yasin[1], Muhammad Raafi'u Firmansyah[*2], Dasril Aldo[3], Muhammad Afrizal Amrustian[4]

[1,2,3]Department of Informatics, Telkom University, Indonesia
[4]Department of Information and Computer Science, King Fahd University of Petroleumand Minerals, Saudi Arabia

Email: [1]raafiu@telkomuniversity.ac.id

Received : May 3, 2025; Revised : Jun 3, 2025; Accepted : Jun 14, 2025; Published : Jun 23, 2025

Abstract

Accurate rice production forecasting plays an important role in supporting national food security planning. This study aims to evaluate the performance of four machine learning algorithms, namely Random Forest, XGBoost, Support Vector Regression (SVR), and Linear Regression, in predicting three target variables simultaneously: harvest area, productivity, and production. The dataset used includes annual data per province in Indonesia from 2018 to 2024 obtained from the Central Statistics Agency (BPS). Evaluation was conducted using five metrics: MAE, RMSE, MAPE, R^2, and training time. The results of the experiment showed that the Random Forest Regressor performed best in the 80:20 scenario, with an MAE of 76,259.52, an RMSE of 154,036.91, a MAPE of 0.61%, and an R^2 of 0.997. XGBoost showed a competitive performance with an MAE of 79,381.44 and faster training times. In contrast, the SVR showed the worst performance with the MAPE reaching 198.56% and the R^2 of 0.209. Linear Regression as baseline recorded an MAE of 1,194,355.28 and an R^2 of 0.503, indicating that the linear model is not effective enough for this data. The 80:20 scenario is considered the best configuration because it is able to balance the accuracy and generalization of the model. These findings show that the use of ensemble algorithms, especially Random Forest and XGBoost, has the potential to be applied practically by agricultural agencies or local governments in designing data-driven policies for more proactive and predictive rice production management. Furthermore, this study contributes to the advancement of applied informatics by demonstrating how machine learning models can be effectively used in multivariate forecasting for complex, real-world problems, thereby supporting the development of intelligent decision-support systems in the agricultural domain.

Keywords : *Forecasting, Machine Learning, Multivariate Regression, Paddy Production, Random Forest.*

1. INTRODUCTION

Rice is a strategic commodity related to food security and economic stability in almost all countries, including Indonesia [1], [2], [3]. As an agrarian country, Indonesia prioritizes rice as part of its food policy [4], [5]. Climate change, the rate of land degradation, the adoption of agricultural technology, and government policies are attributes in the variability of rice production [6]. Thus, the development of an accurate rice production forecasting model is essential for food security planning.

The adoption of machines and digitalization, especially machine learning technology [7], [8], has automatically transformed in the last decade in the agricultural sector. Various machine learning algorithms have been applied to improve the accuracy of predictions of agricultural yields [9], [10]. With the use of machine learning, conventional statistical methods are no longer used due to their limitations in handling complex and non-linear data patterns [11], [12].

Random Forest is one of the algorithms that has shown promising performance in previous studies [13]. Random Forest XGBoost, which is a development of the boosting method, is known for its efficiency as well as accurate prediction results on large and complex data sets [14], [15]. SVR (Support

Vol. 6, No. 3, Juni 2025, Page. 1431-1442
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4681

Vector Regression) has proven to be excellent for anticipating outliers and irrelevant features, especially in the process of predicting and modeling agricultural yields influenced by non-linear factors such as weather [16], [17]. Linear Regression, although known as a widely used model because of its proficiency and ease of interpretation and implementation, is known as a static model that is unable to account for nonlinear relationships [18]. Most of the previous studies used a univariate approach so that the calculations were limited to only one predictable variable such as total production, without considering the relationship between other variables such as harvest area and productivity. Some do cross analysis, compare two to three models, and do not conduct a comprehensive study of several models for multivariate regression.

In accordance with this background, this study aims to simultaneously analyze the performance of four machine learning models: Random Forest Regressor, XGBoost Regressor, Support Vector Regression (SVR), and linear regression, in multivariate forecasting on rice. The models were evaluated based on results by predicting three main key variables: rice, harvest area, productivity, and rice production time using province-level annual historical data. The use of province-level annual data is particularly relevant because agricultural conditions, productivity, and government interventions can vary significantly across regions in Indonesia. By analyzing this granularity, the model can better capture spatial heterogeneity and support localized forecasting, which is essential for regional policy planning and precision agriculture initiatives. The evaluation was carried out using several metrics, including MAE, RMSE, R^2, MAPE, and training time. The main contribution of this study is to provide a comprehensive comparative analysis of the performance of four machine learning algorithms in the context of multivariate predictions of the agricultural sector, especially rice commodities. The results of this study are expected to be a reference in the selection of the most suitable model to be applied to data-based agricultural prediction systems, as well as enrich the reference for the use of artificial intelligence in the field of precision agriculture in Indonesia.

Furthermore, the ability to forecast multiple agricultural indicators simultaneously allows policymakers to allocate resources more efficiently, plan targeted interventions for productivity improvement, and anticipate potential regional disparities in rice supply. This is particularly beneficial for optimizing subsidy distribution, managing food stock reserves, and strengthening national food resilience strategies at both local and national levels.

Previous studies have predominantly focused on univariate prediction models that only estimate a single output, such as rice production, without considering the interdependencies among other crucial variables like harvested area and productivity. Additionally, comparative evaluations often involve a limited set of models or overlook multivariate regression frameworks. This study addresses these gaps by introducing a multivariate forecasting approach that evaluates four distinct machine learning algorithms simultaneously on three interrelated targets. Such comprehensive analysis, especially in the context of Indonesian provincial data, has not been extensively explored, making this study a novel contribution to both agricultural informatics and applied machine learning research.

2. METHOD

The research methodology used in this study is illustrated in Figure 1. The main stages include data collection, data preprocessing, modeling with machine learning algorithms, and model performance evaluation. Each step is designed to support the multivariate forecasting process of rice production accurately and systematically.

Figure 1 shows the overall flow of the research methodology. Each stage is designed to support the modeling and evaluation process in multivariate forecasting of rice production. A detailed explanation of each stage is presented as follows.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 3, Juni 2025, Page. 1431-1442
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4681

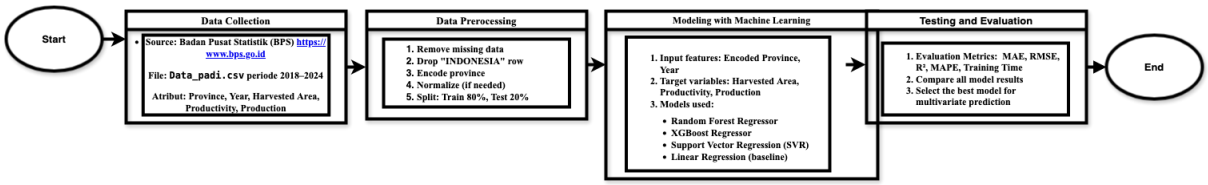

Figure 1. Research Methodology Flow Diagram

2.1. Data Collection

The data used in this study was obtained from the official website of the Central Statistics Agency (BPS) (https://www.bps.go.id). The dataset contains information on rice production per province in Indonesia for the period 2018 to 2024, in .csv format. Table 1 presents the attributes used in this study, which consist of input variables and target variables in the multivariate rice production forecasting process.

Table 1. Research Dataset

	Year	Harvested Area (ha)	Productivity (quintal/ha)	Production (ton)
Aceh	2018	329515.78	56.49	1861567.1
Bali	2018	110978.37	60.11	667069.06
Bengkulu	2018	65891.16	43.83	288810.52
….	….	….	….	….
Jawa Timur	2019	1702426.36	56.28	9580933.88
….	….	….	….	….
Sumatera Selatan	2019	539316.52	48.27	2603396.24
Aceh	2020	317869.41	55.28	1757313.07
….	….	….	….	….
Sumatera Utara	2020	388591.22	52.51	2040500.19
Aceh	2021	297058.38	55.03	1634639.6
….	….	….	….	….
Sumatera Utara	2021	385405	52	2004142.51
Sulawesi Utara	2024	59121.96	46.2	273134.94
Sumatera Barat	2024	295278.98	45.94	1356467.93
Sumatera Selatan	2024	521092.21	55.83	2909411.67
Sumatera Utara	2024	419463.48	52.56	2204875.51

Table 1 displays historical data on rice production in Indonesia by province for the period 2018 to 2024. The data includes three main indicators, namely harvest area (ha), productivity (ku/ha), and total production (tons) presented annually. This information is the basis for training and testing machine learning models to perform multivariate forecasting.

2.2. Data Preprocessing

In the preprocessing phase, the dataset was cleaned and standardized to ensure consistency before being used for machine learning model training [19]. The following steps were performed to prepare the dataset for machine learning model training:

2.2.1. Handing Missing Values

Vol. 6, No. 3, Juni 2025, Page. 1431-1442
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4681

The dataset was checked to ensure no missing values (null or NaN) existed in the target columns [20], Harvested Area, Productivity, and Production. Rows containing missing values in any of these three columns were removed to prevent errors during regression model training.

2.2.2. Removing National Aggregate Rows

Rows containing aggregated national-level data for Indonesia were identified by the value "Indonesia" in the province column. These entries were deleted as they do not represent individual provincial entities and could introduce bias if included alongside province-level data.

2.2.3. Encoding the Province Column

Since machine learning algorithms cannot directly process categorical string data [21], [22], the Province column was converted into numerical form using Label Encoding. Each province was assigned a unique integer label based on its order of appearance in the dataset.

2.2.4. Data Normalization

To ensure numerical stability in algorithms such as Support Vector Regression (SVR), numerical features were normalized (e.g., using Min-Max Scaling) to a [0, 1] range. This step is particularly crucial when features exhibit significantly different scales [23], [24].

2.2.5. Train-Test Split

The dataset was partitioned into two primary subsets: training data and test data, used for model training and performance evaluation, respectively [25]. The dataset was partitioned into two primary subsets: training data and test data, used for model training and performance evaluation, respectively:

a. 70% training and 30% test
b. 80% training and 20% test
c. 90% training and 10% test

2.3. Modeling with Machine Learning

The modeling stage is the core process in this study, where data that has gone through the preprocessing stage is used to train several machine learning models [26] in order to make multivariate predictions. The approach used focuses on multivariate regression to predict three targets at once, namely Harvested Area, Productivity, and Production.

2.3.1. Features and Target Variables

a. Input features (X)**:**
    1. Province (encoded): Numerical representation of province names using label encoding.
    2. Year: Observation year in numerical format.
b. Target outputs (Y):
    1. Harvested Area (ha)
    2. Productivity (quintal/ha)
    3. Production (ton)

All models were trained using the same feature set and were designed to predict all three target variables simultaneously.

2.3.2. Implemented Models

Four machine learning algorithms were selected and implemented in this study, chosen for their respective strengths in handling tabular data, modeling non-linear relationships, and training efficiency.

a. Random Forest Regressor

Vol. 6, No. 3, Juni 2025, Page. 1431-1442
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4681

The Random Forest is an ensemble method based on decision trees that employs bagging techniques [27]. The Random Forest algorithm constructs multiple decision trees through randomized processes and generates predictions by aggregating (averaging) the outputs of all individual trees. This ensemble approach demonstrates three key advantages: (1) inherent capability to model complex non-linear relationships, (2) enhanced resistance to overfitting through its majority-voting mechanism, and (3) robustness when processing noisy datasets due to its distributed error-correction property [28].

$$\acute{y} = \frac{1}{B}\sum_{i=1}^{B} h_i(x) \quad (1)$$

b. XGBoost Regressor

XGBoost (eXtreme Gradient Boosting) is an advanced gradient boosting framework renowned for its computational efficiency and predictive accuracy [29]. XGBoost utilizes a stage-wise training approach, where each new decision tree minimizes the loss function gradient relative to the ensemble's current predictions. The model's regularization framework, comprising both shrinkage (learning rate) and penalty terms, provides inherent protection against overfitting [30].\

$$\acute{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (2)$$

c. Support Vector Regression (SVR)

SVR is an extension of Support Vector Machines (SVM) for regression cases [31]. SVR finds the optimal prediction function within a specified error margin (ε-insensitive loss) and supports kernel use (such as RBF) to handle non-linear relationships. SVR is effective for small-to-medium-sized datasets and when feature relationships are complex [32].

$$f(x) = (w, x) + b \quad (3)$$

d. Linear Regression (Baseline)

Based on an assumed linear relationship between inputs and outputs, this model employs ordinary least squares estimation. While simple, it serves as a critical baseline for quantifying performance enhancements achieved by complex models [33], [34].

2.3.3. Testing and Evaluation

After the training process is completed, each machine learning model is evaluated using the test data that has been set aside from the dataset. This step aims to assess how accurately the model is able to predict the actual value of the three multivariate targets, which are Harvested Area, Productivity, and Production. The evaluation is conducted numerically by calculating the difference between the predicted results and the actual data.

The model was evaluated using the following five key metrics:

a. Mean Absolute Error (MAE)

MAE is calculated based on the mean of the absolute differences between the prediction and actual values. The lower the MAE, the better the model accuracy is in providing estimates regardless of the direction of deviation [35].

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (4)$$

b. Root Mean Squared Error (RMSE)

RMSE imposes a greater penalty on highly inaccurate forecasts since it computes the square root of the average of the errors. A lower value indicates better performance of the model [36].

Vol. 6, No. 3, Juni 2025, Page. 1431-1442
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4681

$$\text{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

c. Coefficient of Determination (R^2 Score)

R^2 refers to how well the model is able to explain the variability of a target variable. Values of R^2 that approach 1 indicate that the model captures the patterns in the dataset accurately [37].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (6)$$

d. Mean Absolute Percentage Error (MAPE)

MAPE expresses error as a percentage of the actual value, thus simplifying interpretation within the context of the domain. This metric is sensitive to very small target values [38].

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \acute{y}_t}{y_t}\right| * 100\% \quad (7)$$

e. Training Time

In addition to evaluating the accuracy of predictions, the time taken to train each model was also recorded in seconds. This information aids in assessing the computing efficiency of each method, especially in the context of real-world applications which require fast processing [39].

3. RESULT

This section presents the results of experiments conducted with four machine learning models used in this study, namely random forest regressor, xgboost regressor, support vector regression (SVR), and linear regression used as a reference. Evaluations for these models were carried out related to harvest area, productivity, and rice production as three target variables, applying several evaluation metrics, namely MAE, RMSE, R^2, MAPE, and training time. The full results of the model tests are shown in Table 2, which contains the average performance values of each model in relation to the three target variables from the test data.

Table 2. Experimental Results

Split	Model	MAE	RMSE	MAPE (%)	R^2 Score	Training Time (s)
70:30	Linear Regression	143,403.54	189,157.25	1537.79	0.9940	0.04
70:30	Random Forest	56,744.12	118,681.01	8.09	0.9976	0.82
70:30	XGBoost	92,637.47	247,457.25	56.01	0.9898	1.67
70:30	SVR	251,467.60	368,164.55	1606.02	0.9774	0.04
80:20	Linear Regression	146,907.79	191,461.17	1249.69	0.9954	0.02
80:20	Random Forest	76,259.52	154,036.91	8.81	0.9970	0.28
80:20	XGBoost	80,908.71	151,319.18	28.00	0.9971	0.08
80:20	SVR	295,422.28	431,515.13	2832.99	0.9767	0.01
90:10	Linear Regression	170,498.79	199,469.71	1185.63	0.9964	0.03
90:10	Random Forest	66,271.59	134,982.05	5.38	0.9984	0.47
90:10	XGBoost	63,228.60	108,900.52	18.04	0.9989	0.86
90:10	SVR	320,622.81	418,415.02	4175.81	0.9843	0.06

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 3, Juni 2025, Page. 1431-1442
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4681

The analysis of the results in Table 2 shows that the Random Forest Regressor consistently performs best in all data-sharing scenarios. For example, in the 80:20 scenario, Random Forest recorded the lowest MAE and RMSE of 76,259.52 and 154,036.91, with R^2 0.997, indicating high predictive multivariate accuracy. The XGBoost Regressor also performed very well with MAE and RMSE approaching the Random Forest results, as well as a relatively fast training time advantage at a 90:10 split (0.38 seconds). However, the MAPE value on XGBoost fluctuates. On the other hand, SVR produces poor prediction performance with extreme MAPE inaccuracy, even though SVR has higher MAE and RMSE errors. Linear Regression as the baseline shows the lowest performance in most scenarios. In general, the data split ratio of 90:10 remains dependent on the goal of producing the highest R^2 value across all models, however, it has the potential for greater overfitting. The 80:20 scenario is more optimal, considered to be the most balanced between precision and generalization.

Based on a general evaluation, the 80:20 data sharing scenario results in the best balance between the accuracy and generalization capabilities of the model. However, for a more in-depth analysis, this section presents the test results and performance of all models in the 70:30 data sharing scenario. The purpose of this analysis was to measure the consistency of model performance in a larger proportion of test data, as well as to measure the sensitivity of the model to training data size variations.

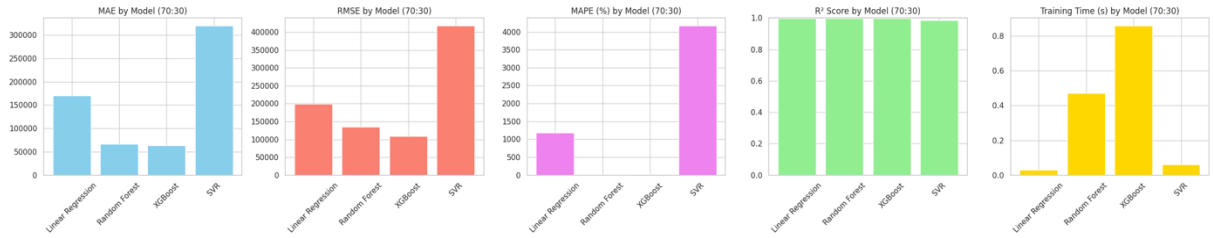

Figure 2. Comparison of Model Performance Under 70:30 Data Split Scenario

After evaluating the model's performance in the 70:30 train-test split scenario (Figure 2), we now present the results of the model evaluation for the 80:20 data separation scenario. The purpose of this analysis was to observe how the consistency and improved performance of the model increased as the proportion of training data increased. The results of the comparison of the various models examined are shown in Figure 3.

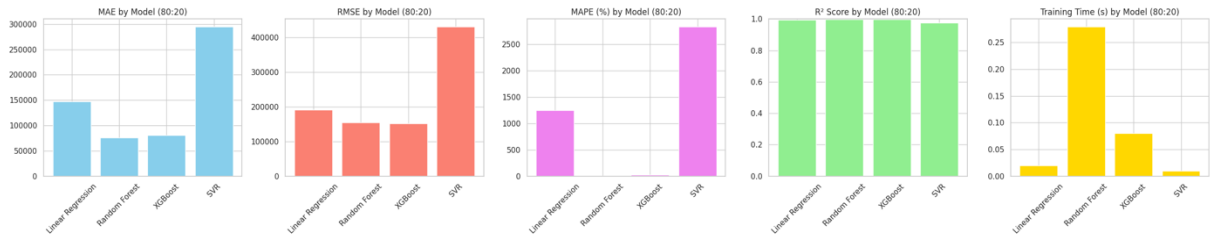

Figure 3. Model Performance Comparison Under 80:20 Data Split Scenario

As a continuation of the previous analysis, Figure 4 shows the results of the model evaluation for the 90:10 data separation scenario. With a larger proportion of training data, the model is expected to learn in an optimal way, albeit with a lower risk of generalization. The following visual illustrates the performance of each model based on the same evaluation metrics.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 3, Juni 2025, Page. 1431-1442
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4681

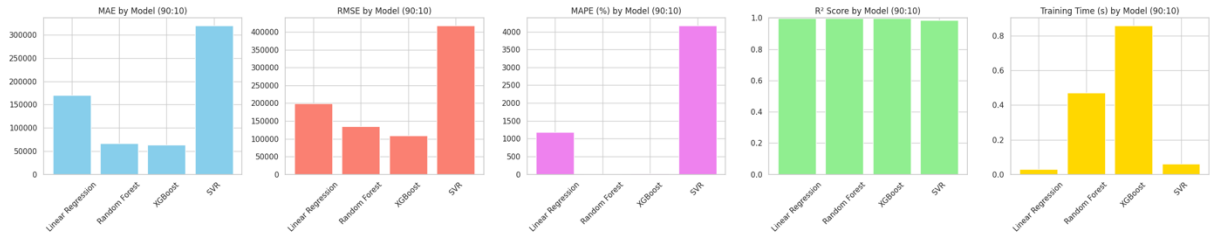

Figure 4. Model Performance Comparison Under 90:10 Data Split Scenario

The results of the evaluation showed that the Random Forest Regressor outperformed the other models in all three data-sharing scenarios. In a 70:30 scenario, the model achieves an MAE of 88,723.69 and an RMSE of 171,040.24 with an R^2 of 0.996. There was an improvement in the performance in the 80:20 scenario, where the MAE further increased to 76,259.52 and the RMSE to 154,036.91, while the R^2 increased to 0.997. The best results were obtained in a 90:10 scenario with an MAE of 61,076.90 and an RMSE of 123,674.90 and an R^2 of 0.998. MAPE values in all three scenarios were also low ranging from 0.61% to 0.67%, indicating consistently high predictive accuracy.

XGBoost Regressor is competitive against Random Forest. There is only a small difference in the score. In an 80:20 scenario, for example, XGBoost obtains an MAE of 79,381.44 and an RMSE of 157,595.33, R^2 is 0.997 with a MAPE of 0.71%. The advantage of XGBoost lies in the efficient training time, which is always faster than Random Forest, especially in 90:10 splits.

As a baseline, simpler models such as Linear Regression show much worse performance. MAE in all scenarios exceeds one million and the RMSE is very high, with R^2 values ranging from 0.503 to 0.573. The MAPE in this model is also very high with a maximum value of 19.24%, which means that the linear enclave cannot capture its patterns in more complex non-linear data.

Support Vector Regression (SVR) gives the worst results. This model obtained a very extreme MAPE value even reaching 198.56% and a very low R^2 of 0.209 in the optimal scenario. This suggests that SVRs in the context of this multivariate rice production dataset, SDRs are more likely to experience difficulties due to the selected kernel, or overly specific parameter tuning.

Overall, Random Forest was the most accurate and stable model, while XGBoost delivered results with competitive performance but more efficient in computing. The 80:20 split of data shows the best results in terms of accuracy and generalization, making it highly recommended as the optimal configuration for similar cases.

4. DISCUSSIONS

The results of this study show that the application of machine learning algorithms for multivariate forecasting of rice production performs very well, especially with the random forest and xgboost models. Both models obtained low mae and rmse values and very high r^2 scores above 0.99 in 80:20 and 90:10 data separation scenarios. These findings reinforce the results of previous research by Sengar et al. [40] and Sharma [41], who also reported the effectiveness of random forest and xgboost in improving the predictive accuracy of agricultural yields compared to traditional linear methods.

The performance of support vector regression (SVR) in this study is far below expectations. This is in contrast to the study conducted by S et al. [42] which showed success for svr in weather-based prediction scenarios. The difference in findings is likely due to the characteristics of the rice production dataset, which is more complex, nonlinear, and has more data complexity. In addition, the sensitivity of svr to scale and kernel parameters also contribute to the discrepancy. Svr is also limited by non-linearity reported by Jovanovic et al. [18] in the non-interdependence between features while linear regression, as a benchmark, is limited by capturing non-linearity between features.

Vol. 6, No. 3, Juni 2025, Page. 1431-1442
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4681

The advantage of random forest is its stability and ability to handle data that has multiple dimensions and is interrelated, while for xgboost the efficiency of model training is its advantage. These findings are in line with the analysis of Fatima et al. [43] which advocates the use of boosting and bagging to address the problem of multivariable prediction in agriculture.

On the other hand, this analysis still has some limitations. First, the available datasets contain only internal attributes (harvest area, productivity, and production), without considering external variables such as rainfall, fertilizer prices, or the use of superior varieties. Second, the multivariate approach applied is limited to regression per dependent variable; there is no exploration of multitask learning techniques or multivariable deep learning. For further research, we suggest extending features with spatial data, climate data, and agricultural policy interventions to improve the predictive power of the model. In addition, the application of deep learning such as lstm or multivariate attention-based models for temporal and intervariable dependencies simultaneously can be the next direction of research.

5. CONCLUSION

This study aims to evaluate the performance of four machine learning algorithms, namely Random Forest, XGBoost, Support Vector Regression (SVR), and Linear Regression in conducting multivariate forecasting of rice production data in Indonesia. The evaluation results showed that the Random Forest Regressor was the best performing model overall, followed by XGBoost which also showed accurate and computationally efficient results. The SVR and Linear Regression models show much lower performance, especially on the MAPE and R^2 metrics. This indicates that these models are not suitable to handle the complexity of rice production data.

The 80:20 data sharing scenario was found to be the most balanced configuration between accuracy and generalization. This research also confirms the importance of selecting the right algorithm in the task of multivariate forecasting in the agricultural sector. As a development, further research is suggested to include external variables such as weather, price, or policy interventions, as well as test a deep learning-based model approach to address dependencies between variables simultaneously and temporally.

ACKNOWLEDGEMENT

The author would like to thank the Institute for Research and Community Service (LPPM) Telkom University Purwokerto for the support and facilities provided in the implementation of this research.

REFERENCES

[1] S. Herliana, S. Ratnaningtyas, S. Nur Arifin, and N. Lawiyah, 'Analysis of Indonesia's Food Security Strategy: Rice Price Volatility', *Glob. Conf. Bus. Soc. Sci. Proceeding*, vol. 14, no. 2, pp. 1–1, Dec. 2022, doi: 10.35609/gcbssproceeding.2022.2(35).

[2] A. Suryana, M. D. Hartono, A. T. Suryana, M. R. Suryana, J. P. Sinaga, and A. R. Irawan, 'Stability of rice availability and prices in Indonesia during the COVID-19 pandemic and Russia-Ukraine war', *BIO Web Conf.*, vol. 119, p. 02013, 2024, doi: 10.1051/bioconf/202411902013.

[3] N. A. Yusrin, 'The Analysis Of Rice Massive Importing In Indonesia Based On Macroeconomics, Microeconomics, International Economics And Politic Economics', *Ultima Manag. J. Ilmu Manaj.*, pp. 308–329, Dec. 2023, doi: 10.31937/manajemen.v15i2.3411.

[4] S. Kholik, I. Nurlinda, Z. Muttaqin, and M. Priyanta, 'Reformulation of Policies to Prevent Land Conversion of Rice Fields In Achieving Indonesia's National Food Security', *F1000Research*, vol. 13, p. 945, Aug. 2024, doi: 10.12688/f1000research.151364.1.

[5] H. V. Christopher, A. A. Purnama, and S. M. M. Harahap, 'Application of K-Means Clustering and OR-Tools to Optimize Rice Distribution: A Case Study of Perum Bulog Indonesia', *Appl. Inf. Syst. Manag. AISM*, vol. 7, no. 2, Sep. 2024, doi: 10.15408/aism.v7i2.40618.

Vol. 6, No. 3, Juni 2025, Page. 1431-1442
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4681

[6] A. Massagony, T. Tam Ho, and K. Shimada, 'Climate change impact and adaptation policy effectiveness on rice production in Indonesia', *Int. J. Environ. Stud.*, vol. 80, no. 5, pp. 1373–1390, Sep. 2023, doi: 10.1080/00207233.2022.2099110.

[7] 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*, vol. 12, pp. 151161–151177, 2024, doi: 10.1109/ACCESS.2024.3478738.

[8] A. Kurniawan, T. R. Soeprobowati, and B. Warsito, 'Innovative Agricultural Solutions: Utilizing Machine Learning to Combat Rice Leaf Disease in Indonesia', in *2024 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, Indonesia: IEEE, Sep. 2024, pp. 247–252. doi: 10.1109/iSemantic63362.2024.10762341.

[9] P. Mahesh and R. Soundrapandiyan, 'Yield prediction for crops by gradient-based algorithms', *PLOS ONE*, vol. 19, no. 8, p. e0291928, Aug. 2024, doi: 10.1371/journal.pone.0291928.

[10] M. U. Maheswari and R. Ramani, 'A Comparative Study of Agricultural Crop Yield Prediction Using Machine Learning Techniques', in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India: IEEE, Mar. 2023, pp. 1428–1433. doi: 10.1109/ICACCS57279.2023.10112854.

[11] K. Geetha, B. V. Vidhya, and A. Kiran, 'An Extensive Study on Precision Farming Based on Crop Yield Using Integrated Approaches to Learning', in *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, Chennai, India: IEEE, Nov. 2023, pp. 1–8. doi: 10.1109/RMKMATE59243.2023.10369738.

[12] Z. Chen, H. S. Goh, K. L. Sin, K. Lim, N. K. H. Chung, and X. Y. Liew, 'Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques', *Adv. Sci. Technol. Eng. Syst. J.*, vol. 6, no. 4, pp. 376–384, Aug. 2021, doi: 10.25046/aj060442.

[13] O. A. Montesinos López, A. Montesinos López, and J. Crossa, 'Random Forest for Genomic Prediction', in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Cham: Springer International Publishing, 2022, pp. 633–681. doi: 10.1007/978-3-030-89010-0_15.

[14] S. N. K, V. K. E B, S. J, Y. H, and J. Jk, 'A Machine Learning approach for crop and fertilizer using Ensemble Model with XGBoost and Random Forest Algorithms', in *International Conference on Recent Trends in Computing & Communication Technologies (ICRCCT'2K24)*, International Journal of Advanced Trends in Engineering and Management, Nov. 2024. doi: 10.59544/CYJR6469/ICRCCT24P49.

[15] Z. Wang, 'Fruit and Vegetable Image Recognition Based on Multiple Tree Models: Applications of Random Forest, XGBoost and Decision Tree', *Sci. Technol. Eng. Chem. Environ. Prot.*, vol. 1, no. 9, Oct. 2024, doi: 10.61173/fvzhe382.

[16] F. García-Vázquez *et al.*, 'Prediction of Internal Temperature in Greenhouses Using the Supervised Learning Techniques: Linear and Support Vector Regressions', *Appl. Sci.*, vol. 13, no. 14, p. 8531, Jul. 2023, doi: 10.3390/app13148531.

[17] Akanksha Sharma and Dr. Charu Saraf, 'Comparing MLR and SVR in Evaluating the Impacts of Climate Change on Sugarcane Production in Saharanpur District', *Int. Res. J. Adv. Eng. Manag. IRJAEM*, vol. 3, no. 01, pp. 35–39, Jan. 2025, doi: 10.47392/IRJAEM.2025.0008.

[18] A. Jovanović, A. Krstić, S. Vujnović, and Ž. Durović, 'On Multivariate Linear Regression Applications', in *2024 11th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN)*, Nis, Serbia: IEEE, Jun. 2024, pp. 1–5. doi: 10.1109/IcETRAN62308.2024.10645121.

[19] S. Vinothkumar, S. Varadhaganapathy, R. Shanthakumari, E. Dhivya, K. B. Jayaharitha, and J. Livithasri, 'Crop Prediction Based on Factors of the Agricultural Environment Using Machine Learning', in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India: IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10725351.

[20] M. Alabadla *et al.*, 'Systematic Review of Using Machine Learning in Imputing Missing Values', *IEEE Access*, vol. 10, pp. 44483–44502, 2022, doi: 10.1109/ACCESS.2022.3160841.

Vol. 6, No. 3, Juni 2025, Page. 1431-1442
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4681

[21] A. Tran, C. Zuniga-Navarrete, L. J. Segura, A. Dourado, X. Wang, and C. R. Bego, 'Categorical Variable Coding for Machine Learning in Engineering Education', in *2024 IEEE Frontiers in Education Conference (FIE)*, Washington, DC, USA: IEEE, Oct. 2024, pp. 1–5. doi: 10.1109/FIE61694.2024.10893080.

[22] E. Valdez-Valenzuela, A. Kuri-Morales, and H. Gomez-Adorno, 'Measuring the Effect of Categorical Encoders in Machine Learning Tasks Using Synthetic Data', in *Advances in Computational Intelligence*, vol. 13067, I. Batyrshin, A. Gelbukh, and G. Sidorov, Eds., in Lecture Notes in Computer Science, vol. 13067. , Cham: Springer International Publishing, 2021, pp. 92–107. doi: 10.1007/978-3-030-89817-5_7.

[23] S. Kim, Y. Noh, Y.-J. Kang, S. Park, J.-W. Lee, and S.-W. Chin, 'Hybrid data-scaling method for fault classification of compressors', *Measurement*, vol. 201, p. 111619, Sep. 2022, doi: 10.1016/j.measurement.2022.111619.

[24] D. U. Ozsahin, M. Taiwo Mustapha, A. S. Mubarak, Z. Said Ameen, and B. Uzun, 'Impact of feature scaling on machine learning models for the diagnosis of diabetes', in *2022 International Conference on Artificial Intelligence in Everything (AIE)*, Lefkosa, Cyprus: IEEE, Aug. 2022, pp. 87–94. doi: 10.1109/AIE57029.2022.00024.

[25] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, 'Trade-off between training and testing ratio in machine learning for medical image processing', *PeerJ Comput. Sci.*, vol. 10, p. e2245, Sep. 2024, doi: 10.7717/peerj-cs.2245.

[26] Y. Li, C. Jia, H. Chen, H. Su, J. Chen, and D. Wang, 'Machine Learning Assessment of Damage Grade for Post-Earthquake Buildings: A Three-Stage Approach Directly Handling Categorical Features', *Sustainability*, vol. 15, no. 18, p. 13847, Sep. 2023, doi: 10.3390/su151813847.

[27] M. Ibrahim, 'Evolution of Random Forest from Decision Tree and Bagging: A Bias-Variance Perspective', *Dhaka Univ. J. Appl. Sci. Eng.*, vol. 7, no. 1, pp. 66–71, Feb. 2023, doi: 10.3329/dujase.v7i1.62888.

[28] J. Bi, E. Li, and Y. Luo, 'Petroleum Price Prediction Based on the Linear Regression and Random Forest', *Appl. Comput. Eng.*, vol. 8, no. 1, pp. 292–296, Aug. 2023, doi: 10.54254/2755-2721/8/20230170.

[29] D. F. Santos, 'Parkinson's Disease Detection using XGBoost and Machine Learning', Oct. 25, 2023. doi: 10.1101/2023.10.23.23297369.

[30] R. Sibindi, R. W. Mwangi, and A. G. Waititu, 'A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices', *Eng. Rep.*, vol. 5, no. 4, p. e12599, Apr. 2023, doi: 10.1002/eng2.12599.

[31] H. Muradi, A. Saefuddin, I. M. Sumertajaya, A. M. Soleh, and D. D. Domiri, 'Support Vector Regression (SVR) Method For Paddy Growth Phase Modeling Using Sentinel-1 Image Data', *MEDIA Stat.*, vol. 16, no. 1, pp. 25–36, Jun. 2023, doi: 10.14710/medstat.16.1.25-36.

[32] T. Uemoto and K. Naito, 'Support vector regression with penalized likelihood', *Comput. Stat. Data Anal.*, vol. 174, p. 107522, Oct. 2022, doi: 10.1016/j.csda.2022.107522.

[33] A. Wooditch, N. J. Johnson, R. Solymosi, J. Medina Ariza, and S. Langton, 'Ordinary Least Squares Regression', in *A Beginner's Guide to Statistics for Criminology and Criminal Justice Using R*, Cham: Springer International Publishing, 2021, pp. 245–268. doi: 10.1007/978-3-030-50625-4_15.

[34] S. Mao, 'Statistical derivation of linear regression', in *International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2021)*, K. Chen, N. Lin, R. Meštrović, T. A. Oliveira, F. Cen, and H.-M. Yin, Eds., Nanjing, China: SPIE, Apr. 2022, p. 141. doi: 10.1117/12.2628017.

[35] S. M. Robeson and C. J. Willmott, 'Decomposition of the mean absolute error (MAE) into systematic and unsystematic components', *PLOS ONE*, vol. 18, no. 2, p. e0279774, Feb. 2023, doi: 10.1371/journal.pone.0279774.

[36] M. Shanmugavalli and K. M. J. Ignatia, 'Comparative Study among MAPE, RMSE and R Square over the Treatment Techniques Undergone for PCOS Influenced Women', *Recent Pat. Eng.*, vol. 19, no. 1, p. e041223224190, Jan. 2025, doi: 10.2174/0118722121269786231120122435.

Vol. 6, No. 3, Juni 2025, Page. 1431-1442
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4681

[37] D. Chicco, M. J. Warrens, and G. Jurman, 'The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation', *PeerJ Comput. Sci.*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[38] P. G. G. Lagrazon and J. B. Tan, 'Predicting Crop Yield in Quezon Province, Philippines Using Gaussian Process Regression: A Data-Driven Approach for Agriculture Sustainability', in *2023 International Conference on Modeling & E-Information Research, Artificial Learning and Digital Applications (ICMERALDA)*, Karawang, Indonesia: IEEE, Nov. 2023, pp. 7–12. doi: 10.1109/ICMERALDA60125.2023.10458211.

[39] Y. Duan, N. Wang, and J. Wu, 'Minimizing Training Time of Distributed Machine Learning by Reducing Data Communication', *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 2, pp. 1802–1814, Apr. 2021, doi: 10.1109/TNSE.2021.3073897.

[40] H. S. Sengar and S. Rai, 'A Comparative Analysis of Different Machine Learning Approaches for Crop Yield Prediction', in *2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS)*, Kalaburagi, India: IEEE, Nov. 2024, pp. 1–5. doi: 10.1109/ICIICS63763.2024.10859455.

[41] N. Sharma and M. Dutta, 'Yield Prediction and Recommendation of Crops in the Northeastern Region Using Machine Learning Regression Models', *Üzüncü Il Üniversitesi Tarım Bilim. Derg.*, vol. 33, no. 4, pp. 700–708, Dec. 2023, doi: 10.29133/yyutbd.1321518.

[42] A. S, M. K. Debnath, and K. R, 'Statistical and machine learning models for location-specific crop yield prediction using weather indices', *Int. J. Biometeorol.*, vol. 68, no. 12, pp. 2453–2475, Dec. 2024, doi: 10.1007/s00484-024-02763-w.

[43] S. Fatima, A. Hussain, S. B. Amir, S. H. Ahmed, and S. M. H. Aslam, 'XGBoost and Random Forest Algorithms: An in Depth Analysis', *Pak. J. Sci. Res.*, vol. 3, no. 1, pp. 26–31, Oct. 2023, doi: 10.57041/pjosr.v3i1.946.