

Machine Learning-Based Prediction of Palm Oil Production in Merangin Regency

Latifatus Sa'diyah^{1)*}, Brestina Gultom²⁾

^{1,2}Department of Information System, Faculty of Science and Technology, University of Adiwangsa Jambi

* Corresponding author : latifabangko2@gmail.com

Abstract

Palm oil production in Merangin Regency experiences fluctuations influenced by various factors, making a prediction model necessary to support planning and decision-making. This study aims to develop a prediction model for palm oil production using Multiple Linear Regression and Random Forest methods. The data used are secondary data obtained from the Plantation Office of Merangin Regency for the period 2016–2025, which include productive plantation area, plant age, number of farmers, sub-districts, and previous production. The Multiple Linear Regression method is used to analyze the relationships between variables, while Random Forest is used for prediction. Model evaluation is carried out using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) metrics with an 80% training and 20% testing data split. The results show that Random Forest vastly outperforms Multiple Linear Regression, with an MAE value of 0.60, MSE of 324.07, and RMSE of 18.00. This significant performance disparity occurs because Random Forest, as an ensemble method, is highly capable of capturing complex and non-linear interactions among the agricultural variables, whereas Multiple Linear Regression is strictly limited to rigid linear patterns. The previous production variable is the most dominant factor influencing production outcomes. The developed model has also been successfully implemented in a simple application-based prediction system. Thus, the Random Forest method is proven to be effective in predicting palm oil production and can be used as a decision-support tool in the plantation sector.

Keywords- Palm Oil, Production Prediction, Multiple Linear Regression, Random Forest, Machine Learning

1. INTRODUCTION

The plantation sector is one of the most important subsectors in Indonesian agriculture because it has high economic value and contributes significantly to regional economies. Plantations involve the systematic cultivation of perennial crops that are managed to generate economic value. Commodities such as palm oil, cocoa, rubber, sugarcane, and coffee are among the main contributors to this sector[1]. Plantation crops are perennial plants that are cultivated systematically and managed for plantation business purposes. Plantations are found in almost every province in Indonesia[2].

Each province in Indonesia has advantages in crop production that are adapted to its specific soil and land conditions, one of which is Jambi Province. Jambi Province consists of several regencies, including Merangin Regency, which is one of the regions that highly depends on the plantation sector.[3]. Based on data from the Plantation Office of Merangin Regency for the period 2016 to 2025, the plantation commodity with the highest production contribution is palm oil. Several key factors that must be considered in influencing palm oil production include the use of superior seedlings, fertilization, soil and climate suitability, as well as pest and disease control[4]. Merangin Regency has a tropical climate with warm temperatures and sufficient rainfall that is relatively evenly distributed throughout the year, making it highly suitable for palm oil growth. This combination of weather conditions and soil characteristics results in higher palm oil productivity compared to other plantation commodities[5]. Oil palm also has varying planting areas and production levels across sub-districts in Merangin Regency.

In plantation operations, one of the main challenges faced is the uncertainty in the amount of oil palm production in each period. Fluctuating production levels can be caused by various factors such as weather conditions, planting year, land area, number of trees, and other environmental factors[6].

Although production data are already available, their utilization is still limited to historical reporting without the application of predictive modeling. Prediction is the process of estimating future events based on historical data and emerging patterns; therefore, an analytical approach is needed to produce more accurate estimates[7][8]. In the context of the agriculture and plantation sector, the use of data mining and machine learning techniques has been widely applied to analyze production patterns, classify regions based on specific characteristics, and support data-driven decision-making[9].

However, most previous studies have focused on clustering production areas or on prediction studies at the company or specific commodity level, and have not specifically addressed the prediction of plantation production at the regency level by utilizing data on productive crops and annual production as the main variables[10][11]. In addition, the application of machine learning-based prediction models to plantation data in Merangin Regency is still limited.

Machine learning is a part of artificial intelligence, defined as a method that learns from experience and can adapt to new inputs to perform tasks automatically and make accurate and appropriate decisions[12][13]. One of the methods used is Multiple Linear Regression, a statistical method that models the relationship between one dependent variable and several independent variables in a linear manner[14]. In addition, the Random Forest method is also used, which is an ensemble learning technique based on a collection of decision trees to improve prediction accuracy and reduce overfitting[15].

The novelty of this study lies in addressing a critical research gap in regional agricultural forecasting, where previous machine learning applications have predominantly focused on macro-level provincial clustering or localized company-specific predictions. This study bridges that gap by developing a localized prediction framework that integrates multi-dimensional regional predictors: sub-district territorial dynamics, actual productive plantation areas, aggregate plant age, localized farmer counts, and historical sequential production lags. Crucially, rather than merely presenting a baseline model comparison, this research focuses on how palm oil production prediction can be assessed using Random Forest (RF) and Multiple Linear Regression (MLR) architectures within a structured spatial-temporal evaluation framework. Furthermore, the unique scientific and practical contribution of this research is the active deployment of the developed framework into an operational, web-based decision support system for regional planning. Therefore, this study aims to establish and evaluate this localized prediction framework, thereby offering a scalable system to mitigate uncertainty and support data-driven decision-making in the plantation sector. [16].

2. METHODS AND DATA

This study employs a quantitative approach using machine learning techniques to predict palm oil production in Merangin Regency. The quantitative approach was selected because the research utilizes numerical data that can be analyzed using statistical and computational methods. Rather than focusing solely on comparing prediction algorithms, this study proposes a localized palm oil production prediction framework that integrates regional agricultural variables into a structured prediction workflow.

The proposed framework begins with data collection from the Merangin Regency Plantation Office, followed by data preprocessing, including data cleaning, categorical encoding, and feature preparation. The processed dataset is then divided into training and testing sets using an 80:20 ratio to evaluate the generalization capability of the prediction models. The training data are used to develop Multiple Linear Regression and Random Forest models, while the testing data are utilized to assess their predictive performance using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Finally, the best-performing model is deployed in a web-based prediction system developed using Streamlit to support regional plantation planning and decision-making.

2.1 Data Source

The dataset used in this study was obtained from the Plantation Office of Merangin Regency. The data cover the period from 2016 to 2025 and contain palm oil production records from several sub-districts

in Mrangin Regency. The variables used in this study include sub-district, productive plantation area, plant age, number of farmers, previous production, and palm oil production. Palm oil production was used as the target variable, while the other variables were used as predictor variables.

2.2 Research Variables

The variables used in this study are presented in Table 1.

Table 1. Research Variables

No.	Variable	Type	Description
1	Sub-district	Independent variable	The administrative area where palm oil production is recorded
2	Productive plantation area	Independent variable	The area of productive palm oil plantation in hectares
3	Plant age	Independent variable	The age of palm oil plants in years
4	Number of farmers	Independent variable	The number of farmers involved in palm oil cultivation
5	Previous production	Independent variable	Palm oil production from the previous period
6	Palm oil production	Dependent variable	The target variable to be predicted

2.3 Palm Oil Prediction Production

This study proposes a palm oil production prediction framework to support regional plantation management in Merangin Regency. The framework integrates data collection, preprocessing, feature engineering, prediction modeling, performance evaluation, and deployment into a unified workflow. Two prediction algorithms, namely Multiple Linear Regression and Random Forest, are incorporated within the framework to assess palm oil production forecasting performance.

Following the preprocessing stage, the dataset is divided into training and testing subsets using an 80:20 ratio. The training dataset is utilized to develop the Multiple Linear Regression and Random Forest models, whereas the testing dataset is used to evaluate the predictive performance of the developed models. This train–test strategy ensures that the proposed framework is capable of assessing model generalization on unseen data before deployment into the web-based decision support system. The overall framework is illustrated in Figure 1.

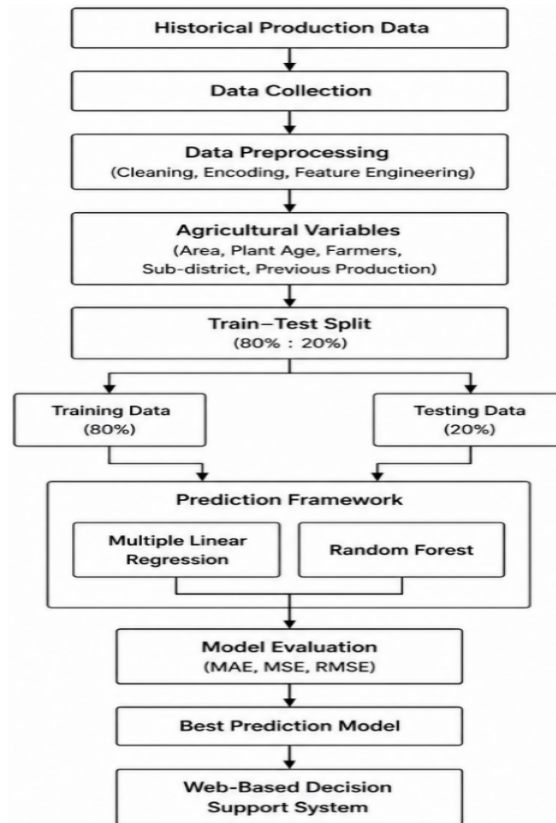


Figure 1. Proposed Palm Oil Production Prediction Framework

2.4 Data Processing

Data preprocessing was carried out to ensure that the dataset was suitable for machine learning analysis. This stage included data cleaning, missing value checking, variable transformation, and feature preparation. Missing value checking was performed to identify whether there were incomplete data in each variable. The results showed that no missing values were found in the dataset, so the data could be used for the next modelling stage without imputation.

The sub-district variable was categorical; therefore, it was converted into numerical form using encoding techniques. This transformation was necessary because machine learning algorithms require numerical input to process the data. In addition, the previous production variable was added as an additional predictor to capture historical production patterns. This variable is important because palm oil production in the previous period may influence production in the following period. After the preprocessing stage was completed, the dataset was ready to be divided into training and testing data for model development and evaluation.

2.5 Multiple Linear Regression

Multiple Linear Regression was used as a baseline model to analyze the linear relationship between the independent variables and palm oil production. This method is suitable for identifying how each independent variable contributes to changes in the dependent variable. In this study, palm oil production was used as the dependent variable, while productive plantation area, plant age, number of farmers, encoded sub-district, and previous production were used as independent variables. The multiple linear regression model used can be formulated as follows[17]:

$$Y = a + b_1 X_1 + b_2 X_2 + b_n X_n \quad (1)$$

where Y represents palm oil production, a represents the intercept, b1,b2,...,bn represent the regression coefficients, and X1,X2,...,Xn represent the independent variables.

In this model, the regression coefficients indicate the contribution of each independent variable to palm oil production. A positive coefficient indicates that an increase in the independent variable tends to increase the predicted production value, while a negative coefficient indicates the opposite relationship.

2.6 Random Forest

Random Forest represents the primary prediction model incorporated into the proposed framework. Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to produce a final prediction. This method can reduce overfitting and improve prediction accuracy by averaging the results of several decision trees. The Random Forest model can be formulated as follows[18]:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x) \quad (2)$$

In this study, the Random Forest model was built using 100 decision trees. The random state parameter was used to ensure consistent and reproducible results. Random Forest was selected because palm oil production may be influenced by complex and non-linear relationships among variables. Compared with Multiple Linear Regression, Random Forest is more flexible in capturing interactions among input variables.

2.7 Evaluation Model

The evaluation stage determines the most suitable prediction model within the proposed framework. The dataset was divided into training and testing data using an 80:20 ratio. The training data were used to build the model, while the testing data were used to evaluate the model's prediction performance. Three evaluation metrics were used in this study: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Mean Absolute Error measures the average absolute difference between actual and predicted values. Mean Squared Error measures the average squared difference between actual and predicted values. Root Mean Squared Error represents the square root of MSE and provides an error value in the same unit as the target variable. The formulas of MAE, MSE, and RMSE are presented as follows[13]:

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i| \quad (3)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} \quad (5)$$

where Y_i represents the actual production value, $\hat{Y}_i$ represents the predicted production value, and n represents the number of testing data. The model with the lowest MAE, MSE, and RMSE values was considered the best-performing model. Lower error values indicate that the predicted production values are closer to the actual production values.

Therefore, this study proposes a localized palm oil production prediction framework for Merangin Regency that integrates regional agricultural variables and employs Multiple Linear Regression and Random Forest to assess prediction performance. The proposed framework is intended to support data-driven decision-making in regional plantation management.

The proposed framework demonstrates that palm oil production prediction can be systematically performed by integrating regional agricultural variables, preprocessing procedures, prediction models, and deployment into a single workflow. This framework can be adapted for other plantation commodities or different regions by replacing the corresponding datasets while maintaining the same prediction process.

3. RESULTS AND DISCUSSION

The proposed palm oil production prediction framework was implemented using historical plantation data collected from Merangin Regency. The framework integrates data preprocessing, feature engineering, prediction modeling, model evaluation, and deployment into a unified workflow. Within this framework, Multiple Linear Regression and Random Forest were utilized to assess prediction performance, while the best-performing model was subsequently implemented in the web-based prediction system.

This section presents the results of the study obtained through a series of systematically designed stages, beginning with raw data preprocessing, followed by an analysis of the relationships between variables using Multiple Linear Regression. The dataset was then divided into training and testing sets, after which a prediction model was developed using the Random Forest algorithm. Finally, the model's performance was evaluated, and the trained model was implemented to predict palm oil production in subsequent periods.

3.1 Data Description

This stage represents the data preparation component of the proposed prediction framework. Before model development, the dataset was examined to understand its initial structure. The dataset consisted of date, sub-district, productive plantation area, plant age, number of farmers, and palm oil production. A sample of the initial dataset before preprocessing is presented in Table 2.

Table 2. Initial Dataset Before Preprocessing (2016–2025)

No.	Date	Sub-district	Productive	Plant Age	Farmers	Production
1	2016-01-08	Bangko	4186	8	2093	330.58
2	2018-03-05	Tiang Pumpung	132	10	66	9.62
3	2020-12-02	Tabir Timur	2239	12	1119	185.04
4	2021-02-26	Pamenang Barat	478	13	239	74.85
5	2022-09-17	Tabir Lintas	666	14	333	50.52
6	2023-10-08	Lembah Masurai	164	15	82	27.04
7	2024-07-29	Bangko Barat	1497	16	748	117.60
8	2025-01-06	Muara Siau	154	17	77	19.33

No.	Date	Sub-district	Productive	Plant Age	Farmers	Production
9	2025-05-19	Tabir Selatan	5690	17	2845	436.81
10	2025-10-13	Pamenang Selatan	5111	17	2556	388.80

The preprocessing stage was carried out to prepare the dataset for machine learning analysis. This stage included checking missing values, formatting numerical variables, encoding sub-district data, and adding previous production as a predictor variable. The dataset after preprocessing is presented in Table 3.

Table 3. Dataset After *Processing* (2016–2025)

No	Date	Sub-district	Area (ha)	Age	Farmers	Production	Sub-district Code	Prev. Production
1	2016-01-08	Bangko	4186	8	2093	330.577	0	330.577
2	2018-03-05	Tiang Pumpung	132	10	66	9.615	15	9.615
3	2020-12-02	Tabir Timur	2239	12	1119	185.042	14	185.042
4	2021-02-26	Pamenang Barat	478	13	239	74.850	8	74.850
5	2022-09-17	Tabir Lintas	666	14	333	50.519	12	50.519
6	2023-10-08	Lembah Masurai	164	15	82	27.038	5	27.038
7	2024-07-29	Bangko Barat	1497	16	748	117.600	1	117.600
8	2025-01-06	Muara Siau	154	17	77	19.330	6	19.330
9	2025-05-19	Tabir Selatan	5690	17	2845	436.810	13	436.810
10	2025-10-13	Pamenang Selatan	5111	17	2556	388.800	9	388.800

The missing value checking results show that all variables had zero missing values. This indicates that the dataset was complete and could be used for model development without imputation. The missing value checking results are presented in Table 4.

Table 4. Results of Missing Value Checking in the Dataset

Variable	Number of Missing Values	Percentage (%)
Date	0	0.0
Sub-district	0	0.0
Productive Area	0	0.0
Plant Age	0	0.0
Number of Farmers	0	0.0
Production	0	0.0
Encoded Sub-district	0	0.0
Previous Week Production	0	0.0

3.2 Training And Testing Split Data

The dataset was divided into training and testing data using an 80:20 ratio. The training data were used to build the model, while the testing data were used to evaluate the model's ability to predict unseen data. The data split is presented in Figure 2.[20].

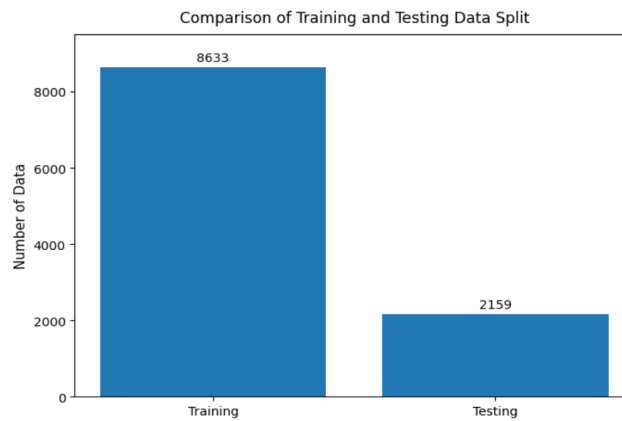


Figure 2. 80:20 Data Split Distribution

The use of training and testing data is important to evaluate whether the model can generalize well to new data. By separating the dataset into two parts, the model is trained on one portion of the data and tested on another portion that was not used during training.[21].

3.3 Multiple Linear Regression Analysis

Within the proposed framework, Multiple Linear Regression serves as the baseline prediction model. Multiple Linear Regression was used to determine the relationship between independent variables and palm oil production. Prior to evaluating the regression coefficients, a classical assumption test for multicollinearity was conducted using the Variance Inflation Factor (VIF). Multicollinearity is evaluated to ensure that the independent variables are not highly correlated with one another, which could distort the coefficient estimations. The VIF thresholds and baseline multicollinearity check results are summarized in Table 5.

Table 5. Multicollinearity Test Results (VIF)

No.	Variable	VIF Value	Status
1	Encoded Sub-district	1.05	No Multicollinearity
2	Productive plantation area	4.21	No Multicollinearity
3	Plant age	1.87	No Multicollinearity
4	Number of farmers	4.15	No Multicollinearity
5	Previous production	2.34	No Multicollinearity

As presented in Table 5, all independent variables yielded a VIF value well below the standard threshold of 10.0, indicating the complete absence of multicollinearity issues within the dataset. Consequently, this confirms that the negative coefficient observed in the productive plantation area is not a mathematical artifact of multicollinearity, but rather a reflection of non-linear dynamics in localized production. Based on the regression analysis, each independent variable had a different contribution to the target variable. The coefficient results are presented in Table 6.

Table 6. Results Multiple Linear Regression Coefecients

Variable	Coefficient
Intercept	-0.6340
Productive plantation area	-0.1573
Plant age	0.0570
Number of farmers	0.3176
Encoded sub-district	0.0238
Previous production	0.9808

Based on Table 6, previous production has the highest coefficient value of 0.9808. This indicates that production in the previous period has the strongest relationship with current palm oil production. The plant age and number of farmers variables show positive coefficients, indicating that increases in these variables tend to increase production. Meanwhile, productive plantation area shows a negative coefficient in the regression model. This result may occur because the relationship between plantation area and production is not purely linear and may be influenced by other factors such as plant condition, productivity level, maintenance, and regional characteristics.

Overall, Multiple Linear Regression can explain the linear relationship between variables. However, this method has limitations in capturing complex and non-linear patterns in the dataset. Therefore, Random Forest was used to improve prediction performance.

3.4 Random Forest Modelling

The Random Forest model is built using a parameter of 100 decision trees ($n_{estimators}$). The selection of this number of trees aims to improve the stability and accuracy of the model, as a larger number of trees generally produces more robust predictions and reduces sensitivity to data variation[22].

In addition, a random state parameter is used to ensure that the modeling results remain consistent when repeated. This parameter functions to control the random sampling process in the construction of decision trees [23].

3.5 Model Evaluation

The evaluation of the Random Forest model in this study is conducted using three metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These three metrics are used to measure the level of prediction error from different perspectives.[24].

```
===== MODEL EVALUATION RESULTS 80:20 =====
MAE : 0.5998654113373473
MSE : 324.0706755529518
RMSE : 18.00196310275498
```

	sub-district	productive_area	plant_age	number_of_farmers	Actual	Prediction
4597	Tabir	544	11	272	45.153846	45.153846
9215	Pamenang Selatan	4035	15	2017	381.096154	381.096154
7412	Bangko Barat	1018	14	509	113.384615	113.384615
9259	Pamenang	12471	15	6235	960.115385	960.115385
8057	Batang Masumai	490	14	245	37.442308	37.442308

Figure 3. Results Evaluation Random Forest Modelling

Based on the model testing results using an 80% training and 20% testing data split, the evaluation values obtained are MAE of 0.60, MSE of 324.07, and RMSE of 18.00. The relatively low MAE value indicates that the average prediction error of the model is small.

The MSE and RMSE values also show that the model's error level is still within an acceptable range, with an RMSE of 18.00 indicating that the average deviation between predicted and actual values is relatively small compared to the scale of production data.

These results are supported by the comparison between actual and predicted values, which shows that the model is able to follow the data pattern well. Therefore, the Random Forest model demonstrates good performance and can be used to predict palm oil production.

3.6 Proposed Prediction Framework Results

Based on the performance evaluation, the proposed palm oil production prediction framework successfully identified Random Forest as the best-performing prediction model. The framework integrates data preprocessing, train–test splitting, prediction modeling, performance evaluation, and system deployment into a unified workflow. Multiple Linear Regression served as the baseline model for comparison, whereas Random Forest achieved the lowest prediction errors and was therefore selected as the final prediction model for deployment.. The selected model is subsequently deployed into the web-based decision support system to support palm oil production prediction in Merangin Regency.

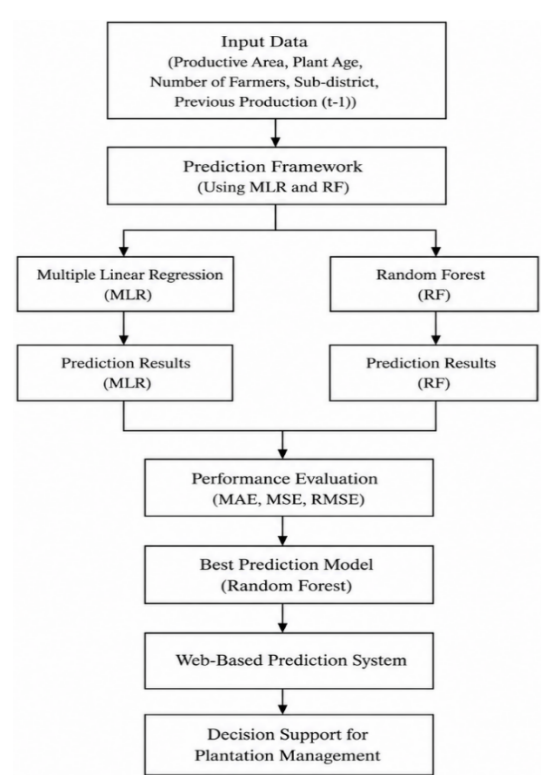


Figure 4. Result of the Proposed Palm Oil Production Prediction Framework

Figure 4 illustrates the outcome of the proposed palm oil production prediction framework after the evaluation stage. The framework evaluates the predictive performance of Multiple Linear Regression and Random Forest using MAE, MSE, and RMSE metrics. Based on the evaluation results, Random Forest was selected as the best prediction model and subsequently deployed in the web-based prediction system to support decision-making for palm oil production planning in Merangin Regency.

3.7 Comparison of Multiple Linear Regression and Random Forest Methods

The comparison is carried out based on model evaluation values using three metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The comparison results of the two methods can be seen in the following figure.

	Method	MAE	MSE	RMSE
0	Multiple Linear Regression	31.02	4443.93	66.66
1	Random Forest	0.60	324.07	18.00

Figure 5. Comparison of Multiple Linear Regression and Random Forest Methods

Based on the evaluation results, the Random Forest method shows lower error values compared to Multiple Linear Regression across all metrics. This indicates that Random Forest produces predictions that are closer to the actual values.

The better performance of Random Forest is due to its ability to capture complex and non-linear relationships between variables, unlike Multiple Linear Regression, which is limited to linear relationships. In addition, the ensemble learning approach in Random Forest enables the model to generate more stable and accurate predictions.

Thus, the Random Forest method is proven to be superior in predicting palm oil production compared to Multiple Linear Regression.

3.8 Implementation of The Prediction System

The prediction system was developed to implement the best-performing model into a simple application interface. Based on the evaluation results, Random Forest was selected as the prediction model because it produced the lowest error values compared with Multiple Linear Regression. The system was designed to help users estimate palm oil production by entering several input variables, including sub-district, productive plantation area, plant age, number of farmers, and date. The technologies used in the prediction system are presented in table 6.

Table 6. Technologies Used in the Prediction System

Component	Technology Used	Description
Programming language	Python	Used for data processing, model development, and implementation
Machine learning library	Scikit-learn	Used to build and apply the Random Forest model
Data processing library	Pandas and NumPy	Used for preprocessing and numerical computation
Web framework	Streamlit	Used to develop the application interface
Visualization	Map visualization	Used to display the selected sub-district location
Model storage	Pickle or Joblib	Used to save and load the trained model

The system architecture consists of three main layers: input layer, processing layer, and output layer. The input layer receives data entered by the user through the application interface. The processing layer validates the input, transforms categorical data into numerical form, loads the trained Random Forest

model, and generates the prediction result. The output layer displays the predicted production value and the selected sub-district location map.

The workflow of the prediction system starts when the user opens the application. The user then enters the required input variables. After the input data are submitted, the system checks whether the data are valid. If the data are incomplete or invalid, the system displays an input error and asks the user to revise the input. If the data are valid, the system preprocesses the input data and applies the trained Random Forest model to generate the predicted palm oil production value. The prediction result and sub-district location map are then displayed to the user. The prediction system flowchart is shown in Figure 6.

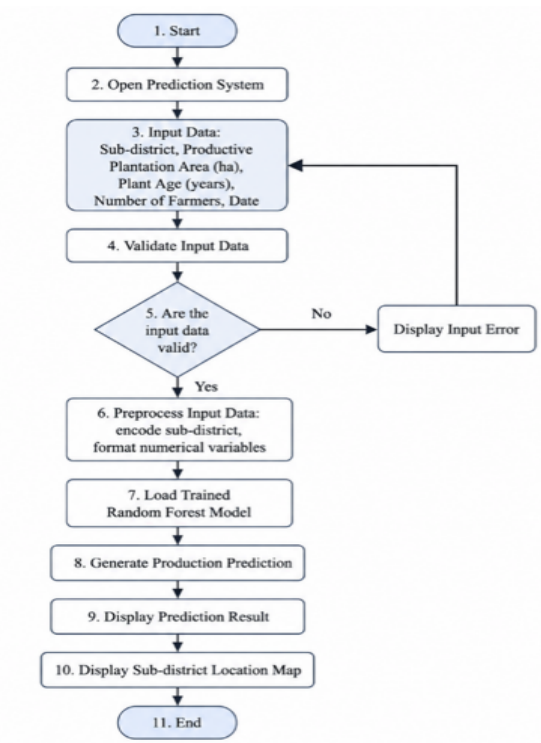
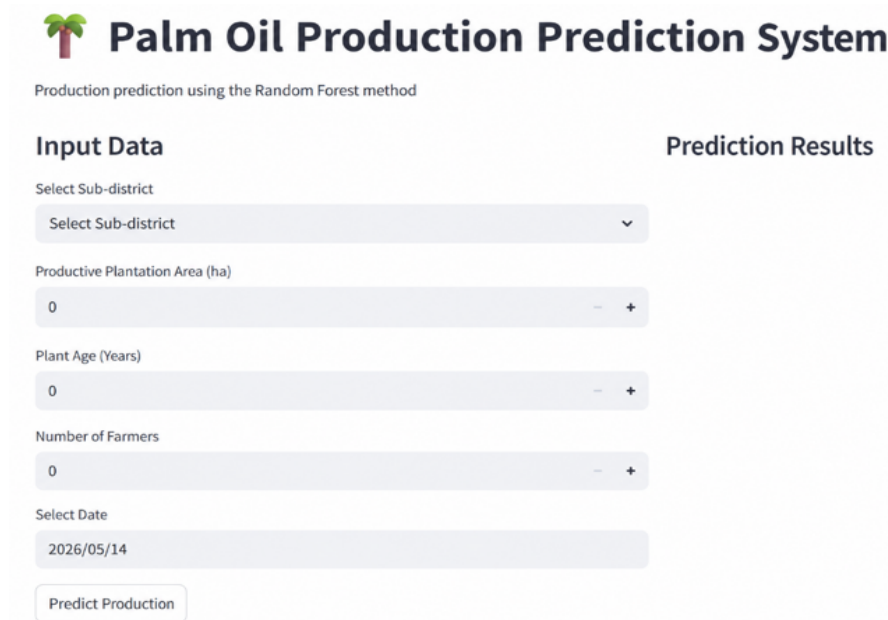


Figure 6. Prediction System Flowchart

The user interface provides input fields for sub-district, productive plantation area, plant age, number of farmers, and date. The sub-district variable is presented as a dropdown menu to reduce input errors. The numerical variables are entered directly by the user. After all data have been entered, the user can click the prediction button to obtain the predicted production value. The prediction system interface is shown in Figure 7.



Palm Oil Production Prediction System
Production prediction using the Random Forest method

Input Data

Select Sub-district

Select Sub-district

Productive Plantation Area (ha)

0

Plant Age (Years)

0

Number of Farmers

0

Select Date

2026/05/14

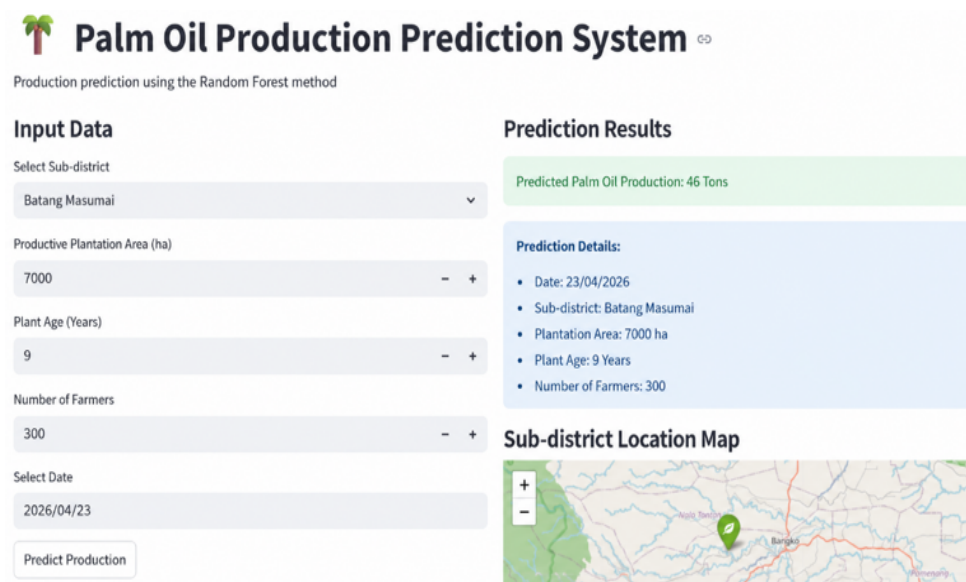
Predict Production

Prediction Results

Figure 7. Prediction System Interface

Based on Figure 7, the system provides input features that are used as parameters in the prediction process. The input variables include sub-district, productive plantation area, plant age, and number of farmers. The sub-district selection is presented in the form of a dropdown menu to make it easier for users and to minimize input errors.

After all data have been entered, the system processes the prediction and displays the result in the form of an estimated palm oil production value. The prediction results are displayed instantly, allowing users to easily determine production values based on the given parameters.



Palm Oil Production Prediction System ©
Production prediction using the Random Forest method

Input Data

Select Sub-district

Batang Masumai

Productive Plantation Area (ha)

7000

Plant Age (Years)

9

Number of Farmers

300

Select Date

2026/04/23

Predict Production

Prediction Results

Predicted Palm Oil Production: 46 Tons

Prediction Details:

- Date: 23/04/2026
- Sub-district: Batang Masumai
- Plantation Area: 7000 ha
- Plant Age: 9 Years
- Number of Farmers: 300

Sub-district Location Map

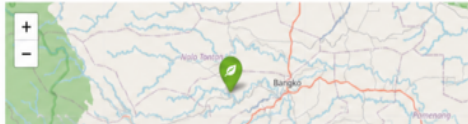


Figure 8. Output Prediction System Interface

The predicted value is the result of processing by the Random Forest model that has been trained using historical data. The model processes all input variables and generates a prediction value that represents the estimated production under those conditions.

In addition to displaying the prediction value in numerical form, the system is also equipped with a visualization in the form of a sub-district location map. This map functions to show the geographical position of the area selected by the user, thereby providing spatial context to the prediction results.

With this informative output display, users not only obtain prediction results directly but can also understand the relationship between location and production outcomes. This makes the system not only a prediction tool but also a supporting medium for analyzing palm oil production data.

4. CONCLUSION

This study successfully proposed a localized palm oil production prediction framework integrating regional agricultural variables, machine learning prediction, model evaluation, and web-based deployment. Based on the results of the study, it can be concluded that the variables of productive plantation area, plant age, number of farmers, sub-district, and previous production significantly influence palm oil production, with previous production and the number of farmers being the most dominant factors. The modeling evaluation proves that the Random Forest method yields superior performance compared to Multiple Linear Regression. Specifically, Random Forest achieved a Mean Absolute Error (MAE) of 0.60, a Mean Squared Error (MSE) of 324.07, and a Root Mean Squared Error (RMSE) of 18.00. In contrast, the Multiple Linear Regression model produced a substantially higher MAE of 31.02, MSE of 4443.93, and RMSE of 66.66.

This significant performance disparity indicates that Random Forest is highly effective in capturing complex, non-linear relationships and multi-variable interactions within agricultural datasets. Furthermore, the practical integration of the Random Forest model into a Streamlit web interface demonstrates that this method can be reliably deployed as an efficient, automated decision-support tool for real-time production forecasting.

Based on these findings, several concrete policy recommendations are proposed for the Merangin Regency Plantation Office to optimize regional palm oil supply chains and management, 1. Strengthening Regional Logistics and Storage Infrastructure: Since previous production trends strongly dictate consecutive weekly yields, the Plantation Office should establish buffer storage facilities (Crude Palm Oil collection hubs) in high-producing sub-districts to stabilize local prices and manage temporary regional oversupplies efficiently. 2. Targeted Smallholder Farmer Support Programs: Given that the number of farmers is a crucial driver of production volume, policies should prioritize equitable fertilizer distribution, continuous agricultural training, and the formation of structured farmer cooperatives to standardize cultivation practices and boost productivity per capita. 3. Data-Driven Agricultural Planning Expansion: The Plantation Office should fully adopt the developed machine learning web system to transition from historical retrospective reporting to proactive, predictive resource scheduling, allowing better logistical coordination for seasonal harvest peaks.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support provided by their academic supervisor. The authors also thank the Plantation Office of Merangin Regency for providing the data used in this study, and all parties who contributed to this research.

REFERENCES

- [1] S. Joya, A. Br, and R. K. R, "Prediksi Harga Tandan Buah Segar dengan Algoritma K-Nearest Neighbor," vol. 5, no. September, pp. 92–101, 2023, doi: 10.30865/json.v5i1.6818.
- [2] M. Sitorus, H. Vande, N. Sinaga, R. Pramono, and D. Kusumo, "Clusterisasi Perkebunan Kelapa Sawit Berdasarkan Luas Area Produksi Dengan Algoritma K-Means," vol. 3, no. 1, 2022.
- [3] F. Hestuningtias and A. H. Primandari, "Pengelompokan Kecamatan Di Kabupaten Merangin Berdasarkan Produksi Tanaman Perkebunan Tahun 2021 Menggunakan Agglomerative Hierarchical Clustering," vol. 1, no. 1, pp. 137–147, 2023.
- [4] R. M. R. Siregar, "Penerapan Metode Moora Dalam Pemilihan Bibit Kelapa Sawit Terbaik," vol.

- 16, no. 2, 2024.
- [5] K. I. Reza, “Faktor – Faktor Yang Berhubungan Dengan Keputusan Petani Dalam Melakukan Alih Komoditi Karet Menjadi Komoditi Sawit Di Kecamatan Renah Pembarap Kabupaten Merangin,” 2024.
 - [6] M. Irpan, U. Rio, Karpen, and Hamdani, “Model Prediksi Jumlah Produksi Kelapa Sawit Menggunakan Regresi Linear Berganda Di PT . Surya Argolika Reksa,” vol. 16, no. 02, pp. 122–130, 2025.
 - [7] H. Tantyoko, D. K. Sari, and A. R. Wijaya, “Prediksi Potensial Gempa Bumi Indonesia Menggunakan Metode Random Forest Dan Feature Selection,” vol. 6, pp. 83–89, 2023.
 - [8] E. Lette, M. Zunaidi, and W. R. Maya, “Prediksi Penjualan Crude Palm Oil (CPO) Menggunakan Metode Regresi Linear Berganda,” vol. 1, 2022.
 - [9] A. A. Simangunsong, I. Gunawan, and Z. M. Nasution, “Pengelompokkan Hasil Produksi Tanaman Perkebunan Berdasarkan Provinsi Menggunakan Metode K-Means Clustering Production of Plantation Crops by Province Using the K-Means Method,” vol. 1, no. 4, 2023, doi: 10.55123/jomlai.v1i4.1661.
 - [10] T. Risananto, E. Poerwandono, D. Sawit, K. J. Timur, and L. Regression, “Optimasi Proses Klasifikasi Topik Berita Berbahasa Indonesia,” vol. 9, no. 6, pp. 9644–9649, 2025.
 - [11] Y. P. Anggriani and A. Arif, “Implementasi Algoritma K-Means Clustering Dalam Menentukan Blok Tanaman Sawit Produktif Pada PT Arta Prigel,” vol. 8, no. 1, pp. 22–32, 2024.
 - [12] S. Prasetyo and T. Dewayanto, “Penerapan Machine Learning , Deep Learning , dan Data Mining Dalam Deteksi Kecurangan Laporan Keuangan - A Systematic Literature Review,” vol. 13, pp. 1–12, 2024.
 - [13] M. Rianto and R. Yunis, “Analisis Runtun Waktu Untuk Memprediksi Jumlah Mahasiswa Baru Dengan Model Random Forest,” vol. 23, no. 1, 2021.
 - [14] S. Adiguno, Y. Syahra, and M. Yetri, “Prediksi Peningkatan Omset Penjualan Menggunakan Metode Regresi Linier Berganda,” vol. 1, pp. 275–281, 2022.
 - [15] Y. Supriyanto, M. Ilhamsyah, and U. Enri, “Prediksi Harga Minyak Kelapa Sawit Menggunakan Linear Regression Dan Random Forest,” vol. 8, no. 7, 2022, doi: 10.5281/zenodo.6559603.
 - [16] B. M. Pratama *et al.*, “Kontribusi Subsektor Perkebunan Dalam Mendukung Pertumbuhan Ekonomi Masyarakat di Sumatera Utara,” vol. 4, no. 6, pp. 1637–1645, 2023.
 - [17] M. D. Irrawati and M. Mukaramah, “Implementasi Metode Regresi Linear Berganda untuk Mengatasi Pelanggaran Asumsi Klasik (Implementation of Multiple Linear Regression Methods to Overcome Violations of Classical Assumptions),” vol. 3, no. 2, pp. 83–94, 2024.
 - [18] F. Diba, M. S. Lydia, and P. Sihombing, “Analisis Random Forest Menggunakan Principal Component Analysis Pada Data Berdimensi Tinggi,” vol. 12, no. 1, pp. 2152–2160, 2023.
 - [19] B. G. Sudarsono and M. I. Leo, “Analisis Data Mining Data Netflix Menggunakan Aplikasi Rapid Miner Analisis Data Mining Netflix Data Using The Rapid Miner,” vol. 4, no. 1, pp. 13–21, 2021.
 - [20] A. N. Maharadja, I. Maulana, and B. A. Dermawan, “Penerapan Metode Regresi Linear Berganda untuk Prediksi Kerugian Negara Berdasarkan Kasus Tindak Pidana Korupsi,” vol. 5, no. 1, pp. 95–102, 2021.
 - [21] C. A. Rahayu, R. Hartono, and A. Sudiarjo, “Prediksi Penderita Diabetes Menggunakan Random Forest,” vol. 11, no. 3, 2023.
 - [22] N. H. Alfajr and S. Defiyanti, “Metode Random Forest dan Penerapan Principal Component Analysis (PCA),” vol. 12, no. 3, 2024.
 - [23] F. Azzahra, N. Suarna, and Y. A. Wijaya, “Penerapan Algoritma Random Forest dan Cross Validation Untuk Prediksi Data Stunting,” vol. 08, no. 01, pp. 1–6, 2024.
 - [24] R. D. Aprilyanto, R. Gustian, M. H. Hernawan, Sumanto, and A. S. Budiman, “Analisis Prediksi Kualitas Udara Dengan Metode Random Forest Berdasarkan Data Cuaca,” pp. 824–828, 2025.