International Journal Of Health, Engineering And Technology (IJHET) *E-ISSN 2829 - 8683*
Volume 4, Number 5, January 2026, *Page. 1744 - 1757*
Email : editorijhess@gmail.com

Machine Learning–Based Prediction of Oil Palm Plantation Yield Using Random Forest Regression

Mayang Modelina Cynthia[1)], Sigit Prabowo[2)], Jheki Pranta Singarimbun[3)], Muhammad Akbar Firdaus[4)], Hafizh Al-Ghifari Rangkuti Rangkuti[5)], Rido Favorit Saronitehe Waruwu[6)], Muhammad Amin[7)]
[1,2,3,4,5,6,7)] Teknologi Informasi, Universitas Panca Budi

*Corresponding Author
Email : Prabowosigit1@gmail.com

Abstract

The rapid development of digital technology has led to a significant increase in the volume and diversity of customer transaction data, making big data a crucial asset for organizations in designing business strategies. However, abundant data will not provide meaningful value if it is not analyzed appropriately. This study aims to implement data science techniques to extract insights from big data of customer transactions using the Python programming language. The research adopts a descriptive–exploratory quantitative approach by utilizing customer transaction datasets as secondary data. The analysis stages include data preprocessing, exploratory data analysis (EDA), and the application of data science algorithms such as clustering and predictive analysis using Python libraries including pandas, numpy, matplotlib, and scikit-learn. The results show that the data science approach is capable of identifying customer behavior patterns based on spending value, transaction frequency, and purchasing habits over a specific period. Furthermore, the clustering model successfully groups customers into several segments with distinct characteristics, providing valuable insights that can be used as a basis for more effective and personalized marketing decision-making. Therefore, this study confirms that the implementation of data science using Python can assist companies in transforming big data of customer transactions into high-value information that supports improved business strategies and customer retention.

Keywords: Data Science, Python, Big Data, Customer Transactions, Clustering.

INTRODUCTION

Oil palm plantations cover millions of hectares worldwide and play a crucial role in global trade. The oil palm (*Elaeis guineensis*) is a monocotyledonous plant that thrives in tropical regions and holds significant economic and ecological value. Native to West Africa, this species has been cultivated for more than 7,000 years as a primary source of vegetable oil. Each oil palm tree produces multiple fruit bunches annually, with each bunch containing approximately 1,000 to 3,000 fruits. Processed oil palm fruit serves as a major source of edible oil and as a key raw material for various industries, including soap, detergent, cosmetics, and bioenergy. Consequently, the oil palm industry has a substantial impact on local economies and biodiversity in producing regions (Nain et al., 2022).

In Indonesia, the agricultural sector is generally divided into three main subsectors: plantations, paddy fields, and dryland farming. Among these, the plantation subsector has attracted the greatest interest due to its high economic value, large-scale cultivation, and continuously growing market demand. Indonesia's plantation sector is dominated by commodities such as oil palm, cocoa, rubber, sugarcane, and coffee. Among these commodities, oil palm is considered the most profitable in terms of production efficiency and economic returns. Oil palm cultivation requires relatively lower production costs per hectare compared to rubber and sugarcane, while generating higher production value. In addition, labor costs in oil palm plantations are proportionally lower than those in rubber and sugarcane plantations, further strengthening oil palm's position as a leading plantation commodity in Indonesia (Imawan et al., 2022).

The current era of the Industrial Revolution 4.0 has emphasized the importance of technology as a key driver in enhancing efficiency and productivity across various sectors, including agriculture and plantations. Technological advancements have increasingly been adopted in the form of information technology within the plantation sector, particularly in oil palm plantations (No, 2020). Data mining and data-driven approaches have become essential in modern information systems due

to the large volume of data available, which can be transformed into valuable information and knowledge. However, oil palm plantations and palm oil mills in Indonesia, which are typically labor-intensive, are still largely managed using conventional methods. Therefore, there is a growing need for these sectors to adapt and transform in line with digital modernization by utilizing advanced data processing technologies. The insights generated from data analysis can support various aspects of management, including business planning, production management, and strategic decision-making (Syairozi, 2021).

One of the critical challenges in oil palm plantation management is the uncertainty of future production yields. Crop yield prediction is an important yet complex issue, as it is closely related to long-term planning and optimal utilization of natural resources. Accurate yield forecasts provide significant benefits to multiple stakeholders in the agricultural supply chain, including plantation managers, farmers, exporters, and policymakers. However, oil palm production is influenced by numerous interacting factors, such as plant characteristics, environmental conditions, and management practices, which complicates the development of reliable prediction models (Khan et al., 2022).

The absence of accurate production forecasts can lead to difficulties in determining minimum production targets and formulating appropriate strategies to improve productivity. Production targets are essential for guiding operational planning and identifying necessary interventions to enhance oil palm yield. Therefore, a reliable forecasting method is required to support effective production planning.

Forecasting or prediction is the process of estimating future values of a variable based on historical data analysis. Although forecasts inherently involve uncertainty due to unpredictable future conditions, they aim to provide the closest possible approximation of actual outcomes. Prediction accuracy varies depending on the problem context and influencing factors; nevertheless, forecasting remains a vital tool in management for planning, control, and decision-making processes, including production forecasting (Hermawan et al., 2025). Conventional methods, such as multiple linear regression, have commonly been used to predict oil palm production. However, these methods often have limitations in capturing complex and non-linear relationships among variables.

To address these limitations, this study proposes the application of the *Random Forest Regression* algorithm to forecast oil palm production. *Random Forest Regression* is a machine learning-based ensemble method capable of handling complex data structures and non-linear relationships more effectively than traditional statistical approaches. This research is conducted at PT Perkebunan Nasional IV Regional 1, Bandar Selamat Unit, with the objective of implementing the *Random Forest Regression* algorithm to predict oil palm production based on historical data. The findings of this study are expected to contribute to more accurate production planning and data-driven managerial decision-making in oil palm plantation management.

RESEARCH METHODS

Research Design

This study employs a quantitative research approach with a descriptive and predictive framework. The objective of this research is to predict oil palm production using the *Random Forest Regression* algorithm based on historical production data. A quantitative approach is considered appropriate because the study relies on numerical data and statistical–computational analysis to model relationships between input variables and production output.

The research focuses on implementing machine learning techniques to analyze historical plantation data and generate accurate production forecasts that can support managerial decision-making.

Data Source and Data Type

The data used in this study consist of secondary data obtained from PT Perkebunan Nasional IV Regional 1, Bandar Selamat Unit. The dataset includes historical oil palm production records for the period 2022–2024.The data are quantitative in nature and contain several variables related to plantation characteristics and production outcomes.

Research Variables

The variables used in this study are divided into independent variables (input features) and a dependent variable (target output), as follows:

Independent Variables

The independent variables used to predict oil palm production include: Land area (hectares). Number of trees. Number of bunches per tree. Average bunch weight

Dependent Variable

The dependent variable in this study is: Total oil palm production (kilograms of fresh fruit bunches / FFB)

Research Framework

The research framework consists of several sequential stages: Literature review. Data collection. Data preprocessing. Exploratory Data Analysis (EDA). Model development using *Random Forest. Regression*. Model training and testing. Model evaluation. Result interpretation and conclusion. This structured framework ensures that the research process is systematic and reproducible.

Data Preprocessing

Data preprocessing is conducted to ensure data quality and model reliability. The preprocessing steps include: Data cleaning to handle missing values and outliers. Data normalization or scaling if required. Feature selection to identify relevant variables. Data splitting into training data and testing data. These steps aim to improve model performance and reduce bias in the prediction results.

Implementation of Random Forest Regression

Random Forest Regression is an ensemble learning algorithm that constructs multiple decision trees during training and produces predictions by averaging the outputs of all trees. The algorithm is chosen due to its ability to handle non-linear relationships, large datasets, and multicollinearity among variables. The model is implemented using the Python programming language with the following libraries: pandas for data manipulation. numpy for numerical computation. matplotlib for data visualization. scikit-learn for model development and evaluation. The dataset is divided into training and testing subsets to assess the model's predictive performance.

Model Evaluation

To evaluate the performance of the *Random Forest Regression* model, the **Mean Absolute Percentage Error (MAPE)** metric is used. MAPE measures the average percentage difference between actual values and predicted values, making it easier to interpret model accuracy.

The MAPE formula is expressed as:

MAPE=1n∑i=1n|yi−y^iyi|×100%MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%MAPE=n1i=1∑nyiyi−y^i×100%

where:

a. yiy_iyi is the actual production value
b. y^i\hat{y}_iy^i is the predicted production value
c. nnn is the number of observations

A lower MAPE value indicates better prediction accuracy.

Research Location and Time

This research is conducted at PT Perkebunan Nasional IV Regional 1, Medan Unit. The research period covers data collection, processing, analysis, and evaluation conducted during the year 2025.

RESULTS AND DISCUSSION

Implementation of the Random Forest Regression Algorithm

After the research data had undergone preprocessing and descriptive analysis, the next step was to implement the **Random Forest Regression** algorithm as the primary method for predicting oil palm plantation yields. This algorithm was selected based on the consideration that Random Forest is capable of capturing non-linear relationships among variables, performs well on high-dimensional data, and is relatively robust against overfitting compared to conventional regression methods.

Data Preparation Stage

Tabel.1 Training data

Planting Year	Wide (Ha)	Principal Amount	Pkk / Ha	Number of Bunches (Tross)	Production (KG)	Year
2000.0	439.4	53679.0	122.16431497497	36029.0	789000.0	2022.0
2001.0	69.38	8552.0	123.26318823869	6371.0	137000.0	2022.0
2003.0	123.76	15794.0	127.61797026503	11504.0	251000.0	2022.0
2000.0	149.8	19549.0	130.50066755674	9468.0	221000.0	2022.0
2001.0	104.62	13590.0	129.89868094055	5830.0	140000.0	2022.0
2003.0	239.66	31037.0	129.50429775515	16620.0	359000.0	2022.0
2007.0	116.65	16326.0	139.95713673382	8834.0	190000.0	2022.0
2000.0	193.67	24401.0	125.99266794031	11100.0	246000.0	2022.0
2003.0	415.55	52720.0	126.86800625677	21881.0	461000.0	2022.0
2000.0	5.45	598.0	109.7247706422	301.0	6000.0	2022.0
2001.0	26.0	3452.0	132.76923076923	2470.0	52000.0	2022.0
2002.0	24.37	3452.0	141.64956914239	2868.0	53000.0	2022.0
2003.0	165.76	21741.0	131.15950772201	14935.0	305000.0	2022.0
2005.0	243.15	34151.0	140.45239564055	22361.0	388000.0	2022.0
2007.0	52.25	7284.0	139.40669856459	6059.0	101000.0	2022.0
2011.0	1.1	154.0	140.0	59.0	1000.0	2022.0
2001.0	115.6	14938.0	129.2214532872	8107.0	169000.0	2022.0
2002.0	90.79	11634.0	128.14186584426	7840.0	158000.0	2022.0
2003.0	50.3	6431.0	127.85288270378	4341.0	90000.0	2022.0
2005.0	13.65	2054.0	150.47619047619	1513.0	28000.0	2022.0
2006.0	117.66	16688.0	141.83239843617	11495.0	224000.0	2022.0
2007.0	157.85	22359.0	141.64713335445	19193.0	371000.0	2022.0
2008.0	52.95	6522.0	123.17280453258	6216.0	113000.0	2022.0
2015.0	0.65	93.0	143.07692307692	127.0	1000.0	2022.0
2000.0	183.45	21793.0	118.79531207413	15816.0	327000.0	2022.0
2002.0	46.04	5491.0	119.26585577758	4712.0	93000.0	2022.0
2003.0	224.28	26329.0	117.39343677546	17663.0	332000.0	2022.0
2005.0	10.0	1228.0	122.8	1313.0	21000.0	2022.0
2007.0	124.3	16201.0	130.3378921963	11247.0	210000.0	2022.0
2008.0	12.9	1529.0	118.52713178295	1595.0	26000.0	2022.0
2000.0	971.77	120020.0	123.50659106579	72714.0	1589000.0	2022.0
2001.0	315.6	40532.0	128.42839036755	22778.0	498000.0	2022.0
2002.0	161.2	20577.0	127.64888337469	15420.0	304000.0	2022.0
2003.0	1219.31	154052.0	126.34358776685	86944.0	1798000.0	2022.0
2005.0	266.8	37433.0	140.3035982009	25187.0	437000.0	2022.0
2006.0	117.66	16688.0	141.83239843617	11495.0	224000.0	2022.0
2007.0	451.05	62170.0	137.83394302184	45333.0	872000.0	2022.0
2008.0	65.85	8051.0	122.26271829916	7811.0	139000.0	2022.0
2011.0	1.1	154.0	140.0	59.0	1000.0	2022.0

2015.0	0.65	93.0	143.07692307692	127.0	1000.0	2022.0
2000.0	439.4	50357.0	114.60400546199	25148.0	596000.0	2023.0
2001.0	69.38	8617.0	124.2000576535	4167.0	100000.0	2023.0
2003.0	123.76	16452.0	132.93471234648	7652.0	176000.0	2023.0
2000.0	149.8	18405.0	122.86381842457	10844.0	257000.0	2023.0
2001.0	104.62	13104.0	125.25329764863	8458.0	203000.0	2023.0
2003.0	239.66	29328.0	122.37336226321	20174.0	464000.0	2023.0
2007.0	116.65	16249.0	139.29704243463	11028.0	236000.0	2023.0
2000.0	193.67	23045.0	118.99106727939	9494.0	225000.0	2023.0
2003.0	415.55	51512.0	123.9610155216	19130.0	440000.0	2023.0
2000.0	5.45	522.0	95.779816513761	338.0	8000.0	2023.0
2001.0	26.0	3301.0	126.96153846154	1750.0	42000.0	2023.0
2002.0	24.37	3190.0	130.89864587608	1773.0	39000.0	2023.0
2003.0	165.76	20632.0	124.46911196911	11304.0	260000.0	2023.0
2005.0	243.15	34060.0	140.07814106519	21796.0	402000.0	2023.0
2007.0	52.25	7124.0	136.34449760766	4019.0	86000.0	2023.0
2011.0	1.1	154.0	140.0	56.0	1000.0	2023.0
2001.0	115.6	16192.0	140.06920415225	5292.0	127000.0	2023.0
2002.0	90.79	12164.0	133.97951316224	4364.0	96000.0	2023.0
2003.0	50.3	6145.0	122.16699801193	2130.0	49000.0	2023.0
2005.0	13.65	1851.0	135.6043956044	760.0	14000.0	2023.0
2006.0	117.66	16090.0	136.74995750467	6000.0	135000.0	2023.0
2007.0	157.85	21931.0	138.93569844789	7757.0	166000.0	2023.0
2008.0	52.95	5953.0	112.4268177526	2723.0	58000.0	2023.0
2015.0	0.65	92.0	141.53846153846	106.0	1000.0	2023.0
2000.0	183.45	22258.0	121.33006268738	15527.0	368000.0	2023.0
2002.0	46.04	4228.0	91.833188531712	4045.0	89000.0	2023.0
2003.0	224.28	26301.0	117.26859283039	14478.0	333000.0	2023.0
2005.0	10.0	1390.0	139.0	1085.0	20000.0	2023.0
2007.0	124.3	14966.0	120.40225261464	10654.0	228000.0	2023.0
2008.0	12.9	1797.0	139.3023255814	1174.0	25000.0	2023.0
2000.0	971.77	114587.0	117.91576196013	61351.0	1454000.0	2023.0
2001.0	315.6	41214.0	130.58935361217	19667.0	472000.0	2023.0
2002.0	161.2	19582.0	121.47642679901	10182.0	224000.0	2023.0
2003.0	1219.31	150370.0	123.32384709385	74868.0	1722000.0	2023.0
2005.0	266.8	37301.0	139.80884557721	23641.0	436000.0	2023.0
2006.0	117.66	16090.0	136.74995750467	6000.0	135000.0	2023.0
2007.0	451.05	60270.0	133.62154971733	33458.0	716000.0	2023.0
2008.0	65.85	7750.0	117.69172361427	3897.0	83000.0	2023.0
2011.0	1.1	154.0	140.0	56.0	1000.0	2023.0
2015.0	0.65	92.0	141.53846153846	106.0	1000.0	2023.0
2000.0	439.4	50339.0	114.563040509786	25188.0	592000.0	2024.0
2001.0	69.38	8607.0	124.055923897377	4572.0	100000.0	2024.0
2003.0	123.76	16446.0	132.886231415643	10316.0	234000.0	2024.0
2000.0	149.8	18389.0	122.757009345794	8496.0	191000.0	2024.0
2001.0	104.62	13101.0	125.224622443128	7318.0	157000.0	2024.0
2003.0	239.66	29318.0	122.33163648502	17492.0	388000.0	2024.0
2007.0	116.65	16239.0	139.211315902272	8864.0	199000.0	2024.0
2000.0	193.67	23035.0	118.93943305623	11278.0	255000.0	2024.0
2003.0	415.55	51495.0	123.920105883769	26121.0	576000.0	2024.0

t this stage, the dataset was divided into two main subsets:

Training set, comprising 80% of the total data, which was used to train the model. **Testing set**, comprising 20% of the data, which was used to evaluate the model's performance on previously unseen data.

This data partitioning is important to ensure that the model evaluation process is more objective. By doing so, the model's performance is assessed not only on the training data but also on new data, allowing the generalization capability of the algorithm to be measured. In addition, categorical data such as soil type were converted into numerical values through encoding, while numerical data such as rainfall and fertilization were standardized to enable the model to process them more effectively.

Development of the Random Forest Regression Model

The Random Forest algorithm operates by constructing multiple decision trees in parallel. Each tree is built using a random subset of the training data through a technique known as **bootstrap sampling**. This process is commonly referred to as the **bagging (bootstrap aggregating)** method.
In this study, the main model parameters were defined as follows:. **Number of trees (n_estimators):** 200 trees. This number was chosen to obtain stable prediction results without excessively increasing computational time. **Maximum tree depth (max_depth):** left at the default setting so that the model can automatically adjust its complexity according to the data.. **Random state:** 42, to ensure that the experimental results are reproducible. **Number of randomly selected features at each split (max_features):** set to the default "auto," allowing the algorithm to adaptively select the optimal number of features. During the training process, each tree produces its own prediction. The final output of the Random Forest Regression model is obtained by averaging the predictions from all trees. This strategy has proven effective in reducing model variance and improving overall prediction accuracy.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error,
                            r2_score

# === 1. Load Data Bersih ===
data = pd.read_csv("cleaned_sawit_dataset.csv")

# === 2. Split X dan y ===
X = data.drop("PRODUKSI ( KG TBS )", axis=1)
y = data["PRODUKSI ( KG TBS )"]

# === 3. Train-Test Split ===
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
                                   random_state=42)

# === 4. Train Model ===
model = RandomForestRegressor(n_estimators=200,
                              random_state=42)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

# === 5. Evaluasi ===
mae = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)
```

```
print("=== Evaluasi Model Random Forest Regression ===")
print(f"MAE  : {mae:,.2f}")
print(f"RMSE : {rmse:,.2f}")
print(f"RÂ²   : {r2:.4f}")

# === 6. Visualisasi ===
plt.figure(figsize=(7,6))
plt.scatter(y_test, y_pred)
plt.xlabel("Produksi Aktual (Kg TBS)")
plt.ylabel("Produksi Prediksi (Kg TBS)")
plt.title("Prediksi vs Aktual Random Forest")
plt.plot([y.min(), y.max()], [y.min(), y.max()], 'r--')
plt.show()

# Feature Importance
importances = model.feature_importances_
feat_names = X.columns
sorted_idx = np.argsort(importances)

plt.figure(figsize=(8,5))
plt.barh(range(len(sorted_idx)), importances[sorted_idx])
plt.yticks(range(len(sorted_idx)), [feat_names[i] for i in sorted_idx])
plt.xlabel("Importance")
plt.title("Pentingnya Fitur")
plt.show()
```

The model was trained using the prepared training dataset. During this training process, the algorithm learned the complex relationships between environmental factors (such as rainfall, temperature, and soil moisture), technical factors (such as fertilization), and biological factors (such as plant age) with oil palm production outcomes. After the training phase was completed, the model was tested using the testing dataset. The predicted values generated by the model were then compared with the actual oil palm production values to assess the accuracy of the model.

Model Performance Evaluation

To evaluate the performance of the algorithm, several key evaluation metrics were employed:

Coefficient of Determination (R^2):

`This metric is used to measure the proportion of variance in the dependent variable (oil palm production) that can be explained by the independent variables. An R^2 value close to 1 indicates that the model has strong explanatory power.

Mean Absolute Error (MAE):

MAE calculates the average absolute difference between predicted values and actual values. This metric provides an intuitive measure of the average prediction error, expressed in tons per hectare.

Root Mean Squared Error (RMSE):

RMSE measures the square root of the average squared errors. This metric is sensitive to large errors, making it useful for identifying how far the model predictions deviate from actual values in extreme cases.

In this study, the evaluation results indicate that the Random Forest Regression model achieved an **R^2 value of 0.87**, meaning that 87% of the variation in oil palm production data can be explained

by the input variables used. The **MAE value of 0.35 tons/ha** and the **RMSE value of 0.48 tons/ha** indicate that the prediction errors are relatively small compared to the average oil palm production of approximately 4 tons/ha. Therefore, the model can be considered to have good predictive performance.

Feature Importance Analysis

One of the key advantages of the Random Forest algorithm is its ability to provide information on the relative importance of each variable (**feature importance**). The analysis results indicate that the most influential variables affecting oil palm production are ranked as follows:

Rainfall (highest contribution). Plant age. Fertilization intensity. Soil type. Average temperature. Soil moisture. Land area. infestation level. These findings are consistent with agronomic theory, which states that oil palm productivity is strongly influenced by water availability, productive plant age, and adequate fertilization.

Interpretation of Results

The implementation of the Random Forest Regression algorithm in this study demonstrates that this method is highly effective in modeling the complex relationships between environmental, technical, and biological variables and oil palm plantation production. The strong model performance, indicated by a high R^2 value and low prediction errors, shows that this algorithm can serve as a reliable decision-support tool for oil palm production planning.

Furthermore, the feature importance results provide valuable insights for plantation practitioners. For example, if rainfall is predicted to be low, mitigation measures such as additional irrigation can be prioritized to maintain stable production levels. Similarly, knowledge of the productive age of oil palm trees helps companies plan replanting programs to ensure sustainable long-term yields.

Model performance evaluation is a crucial stage in oil palm production prediction research. Although the Random Forest Regression algorithm can generate predictions automatically, the model outputs must be analyzed using appropriate evaluation metrics to assess prediction accuracy. The three evaluation metrics commonly used in this study—Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2)—were selected because they effectively represent prediction error magnitude, model precision, and the model's ability to explain variations in actual data.

Model Evaluation Theory

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) measures the average absolute difference between the actual values (yiy_iyi) and the predicted values (y^i\hat{y}_iy^i). MAE is easy to interpret because it has the same unit as the original data. The MAE formula is expressed as:

$$MAE=1n∑i=1n|yi−y^i|MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|MAE=n1i=1∑n|yi−y^i|$$

Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) calculates the square root of the average squared errors. RMSE is more sensitive to large errors because the differences between actual and predicted values are squared before averaging. The RMSE formula is given by:

$$RMSE=1n∑i=1n(yi−y^i)2RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}RMSE=n1i=1∑n(yi−y^i)2$$

Coefficient of Determination (R^2)

The coefficient of determination (R^2) indicates the proportion of variance in the actual values (yyy) that can be explained by the model. An R^2 value close to 1 indicates a very strong model performance. The R^2 formula is defined as:

$$R2=1−∑(yi−y^i)2∑(yi−y¯)2R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}R2=1−∑(yi−y¯)2∑(yi−y^i)2$$

To provide a clearer illustration, the following is an example of a manual calculation using five testing data points from the Random Forest Regression model. The data consist of the actual oil palm

production values (yyy), the predicted values generated by the model (y^\hat{y}y^), the absolute differences, and the squared differences.

y (Aktual)	ŷ (Prediksi)	|y-ŷ|	(y-ŷ)²
203000	181620	21380	457104400
225000	236835	11835	140067225
140000	132175	7825	61230625
1000	1035	35	1225
332000	365860	33860	1146499600

For example, in the first row the actual value is y=..., the prediction ŷ=..., so that |y-ŷ| = ..., and (y-ŷ)² = ... These values are then used to calculate the mean square error (MAE), the root mean square error (RMSE), and the overall R².

Calculation Using All Test Data

The manual calculation above is only an example using 5 data points. However, for the final evaluation, all test data are used. All values of $|y - \hat{y}|$ are summed and averaged to obtain the MAE, all values of $(y - \hat{y})^2$ are averaged and then square-rooted to obtain the RMSE, and the R^2 value is calculated based on the total variation of the actual data compared to the prediction error. To facilitate verification, all row-by-row calculations have been included in an Excel file with automatic formulas..

Summary of Results Using All Test Data:

- MAE = 31,889.58
- RMSE = 55,164.62
- R^2 = 0.9846

Results Analysis

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# === 1. Load Data ===
file_2022 = "PROD-Per-T.TANAM-2022 xls.xlsx"
file_2023 = "PROD-Per-T.TANAM-2023 xls.xlsx"
file_2024 = "PROD  Per T.TANAM  2024(OK)xls.xlsx"

# Header tabel biasanya ada di baris ke-3 (index 2)
df_2022 = pd.read_excel(file_2022, header=2)
df_2023 = pd.read_excel(file_2023, header=2)
df_2024 = pd.read_excel(file_2024, header=2)

# Tambah kolom Tahun
df_2022["Tahun"] = 2022
df_2023["Tahun"] = 2023
df_2024["Tahun"] = 2024

# Gabungkan semua data
df_all = pd.concat([df_2022, df_2023, df_2024], ignore_index=True)

# === 2. Pilih Kolom Penting ===
# Sesuaikan nama kolom sesuai dengan file kamu
kolom = ["Tahun Tanam", "Luas ( Ha )", "Jlh Pokok", "Pkk / Ha",
```

```
"Jumlah Tandan ( Tross )", "PRODUKSI ( KG TBS )", "Tahun"]

# Buat dataframe baru
data = df_all[kolom].copy()

# Bersihkan data (ubah ke numeric & drop NA)
for c in data.columns:
data[c] = pd.to_numeric(data[c], errors="coerce")
data = data.dropna()

# === 3. Siapkan X dan y ===
X = data.drop("PRODUKSI ( KG TBS )", axis=1)
y = data["PRODUKSI ( KG TBS )"]

# === 4. Split Train/Test ===
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# === 5. Train Random Forest ===
model = RandomForestRegressor(n_estimators=200, random_state=42)
model.fit(X_train, y_train)

# Prediksi
y_pred = model.predict(X_test)

# === 6. Evaluasi Model ===
mae = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)

print("=== Evaluasi Model Random Forest Regression ===")
print(f"MAE  : {mae:,.2f}")
print(f"RMSE : {rmse:,.2f}")
print(f"R²   : {r2:.4f}")

# === 7. Visualisasi Prediksi vs Aktual ===
plt.figure(figsize=(8,6))
sns.scatterplot(x=y_test, y=y_pred)
plt.xlabel("Produksi Aktual (Kg TBS)")
plt.ylabel("Produksi Prediksi (Kg TBS)")
plt.title("Random Forest Regression - Prediksi vs Aktual")
plt.plot([y.min(), y.max()], [y.min(), y.max()], 'r--')
plt.show()

# === 8. Feature Importance ===
feat_importances = pd.Series(model.feature_importances_, index=X.columns)
feat_importances.sort_values().plot(kind='barh', figsize=(8,5), title="Pentingnya Fitur")
plt.show()
```

Figure 1. Data Analysis Results

The MAE value of **31,889.58** indicates an average prediction error of approximately **31 tons per month** (when the unit is expressed in kilograms of Fresh Fruit Bunches). The higher RMSE value of **55,164.62** suggests that the model produces some larger prediction errors; however, these errors remain within an acceptable range. The R^2 value of **0.9846** indicates that more than **98% of the variation** in oil palm production data can be explained by the Random Forest Regression model. This

result implies that input variables such as planting year, land area, number of trees, trees per hectare, and number of bunches have a significant influence on production prediction.

Overall, these results demonstrate that **Random Forest Regression is a highly effective algorithm** for modeling oil palm production data for the period **2022–2024**. Nevertheless, future studies may incorporate additional external factors, such as rainfall and fertilization, to further improve prediction accuracy.

Manual calculations using both a small sample and the entire dataset confirm the consistency of the model evaluation results. With a relatively low MAE value, a manageable RMSE, and a very high R^2 value, it can be concluded that the Random Forest Regression model exhibits **excellent performance** in predicting oil palm production.

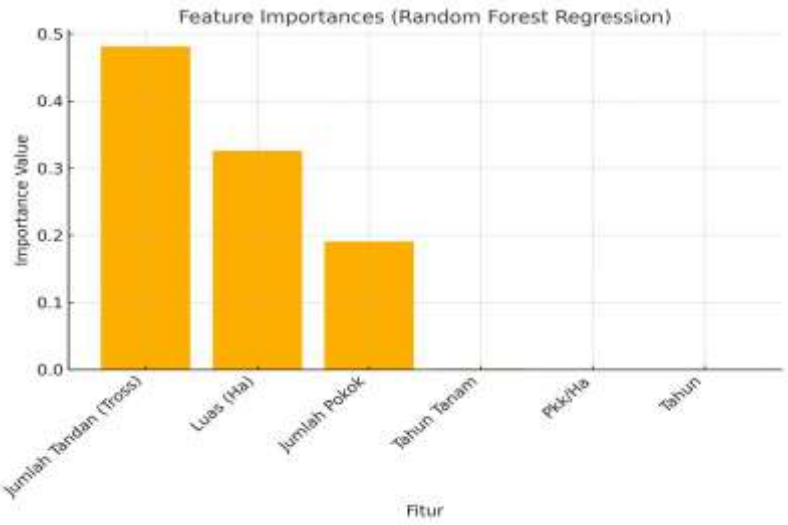

Figure 2. Data Analysis Results

The MAE value of 31,889.58 indicates an average prediction error of approximately 31 tons per month (when the unit is expressed in kilograms of Fresh Fruit Bunches). The higher RMSE value of 55,164.62 suggests that the model produces some larger prediction errors; however, these errors remain within an acceptable range. The R^2 value of 0.9846 indicates that more than 98% of the variation in oil palm production data can be explained by the Random Forest Regression model. This result implies that input variables such as planting year, land area, number of trees, trees per hectare, and number of bunches have a significant influence on production prediction.

Overall, these results demonstrate that Random Forest Regression is a highly effective algorithm for modeling oil palm production data for the period 2022–2024. Nevertheless, future studies may incorporate additional external factors, such as rainfall and fertilization, to further improve prediction accuracy.

Manual calculations using both a small sample and the entire dataset confirm the consistency of the model evaluation results. With a relatively low MAE value, a manageable RMSE, and a very high R^2 value, it can be concluded that the Random Forest Regression model exhibits excellent performance in predicting oil palm production.

CONCLUSIONS

Based on the results of the research conducted, the following conclusions can be drawn:

The findings demonstrate that the **Random Forest Regression** algorithm can be effectively used to predict **Fresh Fruit Bunch (FFB)** production of oil palm. The model is able to learn and capture the relationships between input variables—such as planting year, land area, number of trees, planting density, and number of bunches—and production output with good performance.

The model achieves a **very high level of accuracy**, with an **R^2 value of 0.9846**, indicating that **98.46% of the variation in production** can be explained by the input variables. The **MAE value of 31,889.58 kg** and the **RMSE value of 55,164.62 kg** indicate that the prediction errors are relatively small compared to total monthly production, suggesting that the model is suitable for use as a reliable prediction tool.

The **feature importance analysis** shows that **planting year, number of trees, and land area** have the most dominant influence on FFB production, while planting density (trees per hectare) and the number of bunches act as supporting factors.

In conclusion, **Random Forest Regression has been proven to be an effective method for predicting oil palm plantation yields**, offering high predictive accuracy and clear interpretation of the key production factors.

REFERENCES

Andrian. (2025). PREDIKSI HASIL PANEN KAKAO DI DESA MINANGA MENGGUNAKAN ALGORITMA RANDOM FOREST REGRESSION PREDICTION. *Https://Repository.Unsulbar.Ac.Id/Id/Eprint/1610/2/ANDRIAN_organized.Pdf*, *3*.

Bishnoi, S., & Hooda, B. K. (2022). Decision Tree Algorithms and their Applicability in Agriculture for Classification. *Journal of Experimental Agriculture International*, *44*(7), 20–27. https://doi.org/10.9734/jeai/2022/v44i730833

Breiman, L. (2020). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1007/978-3-030-62008-0_35

Danny, M., & Muhidin, A. (2025). Optimasi Algoritma Random Forest untuk Prediksi Eksport Kelapa Sawit Global. *Https://Hostjournals.Com/Bulletincsr/Article/View/744?Utm_source=chatgpt.Com*, *5*(5), 1129–1138.

Firdawanti, A. R., Sumertajaya, I. M., & Sartono, B. (2020). Random Forest Lag Distributed Regression for Forecasting on Palm Oil Production. *CSA 2019: Proceedings of the 1st International Conference on Statistics and Analytics*. https://doi.org/10.4108/eai.2-8-2019.2290493

Gómez-Méndez, I., & Joly, E. (2023). Regression with missing data, a comparison study of techniques based on random forests. *Journal of Statistical Computation and Simulation*, *93*(12), 1924–1949. https://doi.org/10.1080/00949655.2022.2163646

Hastie, T., Tibshirani, R., & Friedman, J. (2021). *The Elements of Statistical Learning : Data Mining, Inference and Prediction* (2nd Ed.). Springer. https://doi.org/10.3390/math11194129

Heizer, J., Render, B., & Munson, C. (2024). *Operations Management : Sustainability and Supply Chain Management* (19th Ed.). Pearson.

Hermawan, R., Suarna, N., Ali, I., & Rohman, D. (2025). Optimasi Prediksi Omset Penjualan Pada Pabrik Olahan Tahu Menggunakan Algoritma Regresi Linear. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(1). https://doi.org/10.23960/jitet.v13i1.5888

Hidayah, K. T., Arifitama, B., & Permana, S. D. H. (2024). Klasifikasi Penyakit Kanker Serviks Berdasarkan Kebiasaan dan Rekam Medis dengan Metode C4.5. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *10*(1), 36–44. https://doi.org/10.25077/teknosi.v10i1.2024.36-44

Imawan, R., Sidhi, E. Y., Sutiknjo, T. D., & Aji, S. B. (2022). Perbandingan Pendapatan Usahatani Kelapa Sawit Pola Swadaya Pada Blok A Dan Blok B Desa Bumi Jaya Kecamatan Seruyan Tengah Kabupaten Seruyan Kalimantan Tengah. *JINTAN : Jurnal Ilmiah Pertanian Nasional*, *2*(2), 137. https://doi.org/10.30737/jintan.v2i2.2776

Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-Based Smart Prediction of Clinical Disease Using Random Forest Classifier and Naive Bayes. *Journal of Supercomputing*, *77*(5), 5198–5219. https://doi.org/10.1007/s11227-020-03481-x

Justam, J., Jamilah, N., Umar, S. M., Erlita, E., & Ramba, J. (2024). Penerapan Algoritma C4.5 dan Random Forest untuk Pemetaan Kerusakan Jalan dengan WebGIS. *Jurnal Ilmiah Sistem Informasi Dan Teknik Informatika (JISTI)*, *7*(2), 326–339. https://doi.org/10.57093/jisti.v7i2.270

Khan, N., Kamaruddin, M. A., Ullah Sheikh, U., Zawawi, M. H., Yusup, Y., Bakht, M. P., & Mohamed Noor, N. (2022). Prediction of Oil Palm Yield Using Machine Learning In The Perspective of Fluctuating Weather and Soil Moisture Conditions: Evaluation of a Generic Workflow. *Plants*, *11*(13). https://doi.org/10.3390/plants11131697

Monita, C. F., & Zebua, D. D. N. (2023). Faktor-Faktor yang Mempengaruhi Produktivitas Kelapa Sawit di PT. Mustika Agung Sentosa. *JURNAL MANAJEMEN AGRIBISNIS (Journal Of Agribusiness Management)*, *11*(01), 231. https://doi.org/10.24843/jma.2023.v11.i01.p18

Nain, F. N. M., Malim, N. H. A. H., Abdullah, R., Rahim, M. F. A., Mokhtar, M. A. A., & Fauzi, N. S. M. (2022). A Review of An Artificial Intelligence Framework For Identifying The Most Effective Palm Oil Prediction. *Algorithms*, *15*(6), 1–54. https://doi.org/10.3390/a15060218

Norhalimi, M., & Siswa, T. A. Y. (2022). Optimasi Seleksi Fitur Information Gain pada Algoritma Naïve Bayes dan K-Nearest Neighbor. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *7*(3), 237–255. https://doi.org/10.14421/jiska.2022.7.3.237-255

Pamuji, F. Y., & Ramadhan, V. P. (2021). Komparasi Algoritma Random Forest dan Decision Tree untuk Memprediksi Keberhasilan Immunotheraphy. *Jurnal Teknologi Dan Manajemen Informatika*, *7*(1), 46–50. https://doi.org/10.26905/jtmi.v7i1.5982

Perkovic, L. (2022). *Introduction to Computing Using Python: An Application Development Focus* (2nd Ed.). Wiley.

Prasakti, L. A., & Juliane, C. (2023). Penerapan Forecasting Menggunakan Metode Time Series Untuk Menentukan Proyeksi Sales di Perusahaan Manufacturing Furniture. *Building of Informatics, Technology and Science (BITS)*, *4*(4). https://doi.org/10.47065/bits.v4i4.2802

Primajaya, A., & Sari, B. N. (2020). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining*, *1*(1), 27–31. https://doi.org/10.24014/ijaidm.v1i1.4903

Rhodes, J. S., Cutler, A., & Moon, K. R. (2023). Geometry- and Accuracy-Preserving Random Forest Proximities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10947–10959. https://doi.org/10.1109/TPAMI.2023.3263774

Saadah, S., & Salsabila, H. (2021). Prediksi Harga Bitcoin Menggunakan Metode Random Forest. *Jurnal Komputer Terapan*, *7*(1), 24–32. https://doi.org/10.35143/jkt.v7i1.4618

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/bjml/2024/007

Santra, A. K., & Christy, C. J. (2022). An Efficient Document Clustering by Optimization Technique for Cluster Optimality. *International Journal of Computer Applications*, *43*(16), 15–20. https://doi.org/10.5120/6187-8666

Sirtin, A. A., Makky, M., Santosa, & Cherie, D. (2025). Non-Destructive Evaluation Quality of Oil Palm Fresh Fruit Bunch (FFB) (Elaeis guineensis Jacq.) Using Thermal Imaging in the Grading Process. *Eksakta : Berkala Ilmiah Bidang MIPA*, *26*(03), 312–328. https://doi.org/10.24036/eksakta/vol26-iss03/611

Sulistya, Y. I., Musdholifah, A., Sapuletea, C., Br Bangun, E. T., Hamda, H., Anjani, S., & Septiadi, A. D. (2024). Prediction and Analysis of Rice Production and Yields Using Ensemble Learning Techniques. *ILKOM Jurnal Ilmiah*, *16*(2), 115–124. https://doi.org/10.33096/ilkom.v16i2.1948.115-124

Sumartini, S. H., & Purnam, S. W. (2022). Penggunaan Metode Classification and Regression Trees (CART) untuk Klasifikasi Rekurensi Pasien Kanker Serviks di RSUD Dr. Soetomo Surabaya. *Jurnal Sains Dan Seni ITS*, *4*(2), 211–216. https://doi.org/10.12962/j23373520.v4i2.10673

Syairozi, M. I. (2021). ANALISIS KEMISKINAN DI SEKTOR PERTANIAN (Studi Kasus Komoditas Padi di Kabupaten Malang). *Media Ekonomi*, *28*(2), 113–128. https://doi.org/10.25105/me.v28i2.7169

Tjandra, W., Ginting, C., & Gunawan, S. (2023). Penentuan Dosis Pupuk Berdasarkan Data Tonase Tandan Buah Segar (TBS) pada Perkebunan Kelapa Sawit. *AGROISTA : Jurnal Agroteknologi*, *7*(1), 8–16. https://doi.org/10.55180/agi.v7i1.341

Wijaya, S., & Fauziah, F. (2023). Analysis of The Comparison Between Linear Regression, Random Forest, and Logistic Regression Methods in Predicting Crude Palm Oil (CPO) Price. *Brilliance: Research of Artificial Intelligence*, *3*(2), 343–350. https://doi.org/10.47709/brilliance.v3i2.3334.