
UTILIZATION OF RANDOM FOREST TO SUPPORT FUTURE CORN HARVEST PLANNING

Rahmat Hadinata*, Hafish Arrusal Isfalana, Adam Fikri Nawibowo

Lampung State Polytechnic

*Corresponding author: Rsafei731@gmail.com

Abstract

Corn is a strategic food commodity in Indonesia, with national production reaching around 22 million tons. In 2023, Lampung Province will contribute around 9% of the total. However, Lampung faces serious challenges due to climate change. This condition shows the need for further research to increase the productivity and resilience of corn farming to support national food security. Significant climate change has caused a decrease in the quality of corn crops. To overcome this problem, the application of the Random Forest algorithm is expected to increase the yield of the harvest. Random forest is one of the methods that can increase the accuracy in generating attributes for each node that is carried out randomly, with supporting variables such as moisture, temperature, soil moisture in the corn farming environment to predict future crops. By optimizing the hyperparameters, the developed RF model shows a high level of accuracy, with an R-squared (R^2) value of 0.894 and a Mean Absolute Error (MAE) of 5.12. Further analysis revealed that air humidity was the most dominant factor in influencing maize growth (67.1%), followed by temperature (25.5%) and soil moisture (7.3%). In addition, the model successfully predicted plant growth with a high degree of accuracy on the new data, resulting in a prediction value of 90.88 for a specific combination of parameters. The conclusions of this study show that RF algorithms can be an effective tool in supporting precision agricultural planning, thereby increasing corn productivity and facing increasingly complex climate change challenges.

Keywords: Corn, Machine learning, Plant growth, Precision agriculture, Random forest.

INTRODUCTION

Corn (*Zea mays* L.) is one of the strategic food commodities that has an important role in supporting national food security in Indonesia. Based on a report by the Ministry of Agriculture, national corn production reached 23 million tons in 2021, making it the second most important commodity after rice in the domestic food system (Fitri & Nugraha, 2023). Apart from being a staple food, corn also serves as an industrial raw material and animal feed, which makes it increasingly relevant in the modern agricultural sector.

However, global climate change presents a major challenge to maize productivity. Rising global temperatures, changes in rainfall patterns, and fluctuations in air humidity are the main factors that affect the growth cycle of plants. In West Java, for example, there was a decrease in crop yield from 1.43 million tons in 2018 to 1.17 million tons in 2022, although the harvest area was relatively stable (Jeong *et al.*, 2016; Ardika, 2023). This condition shows the need for innovative approaches to anticipate the impact of climate change on corn production.

One promising approach is the use of artificial intelligence-based technology (AI), such as machine learning (ML). The Random Forest (RF) algorithm has proven to be an effective method of handling complex data and providing accurate predictions. Previous research has shown that RF is able to improve the accuracy of crop prediction by up to 20% compared to traditional methods, such as linear regression (Frontiers in Remote Sensing Editorial Team, 2023). In addition, RF integration

with satellite imagery data, such as Landsat-8, enables the classification of farmland with an accuracy of more than 90%, which supports spatial analysis for precision agriculture systems (Muchisha, 2021).

This study aims to explore the potential of RF algorithms in predicting the growth of maize plants based on the main environmental variables, namely temperature, air humidity, and soil moisture. In addition to focusing on prediction accuracy, this study also examines the contribution of each variable to the prediction results. By using historical data from various regions, this research is expected to make a real contribution to the development of precision agriculture systems that support national food security, especially in facing the challenges of climate change.

LITERATURE REVIEW

Random Forest (RF) is an ensemble-based machine learning algorithm designed to handle complex data with a high degree of accuracy. The algorithm leverages a combination of multiple decision trees to generate more stable predictions and reduce the risk of overfitting. RF's advantage lies in its ability to identify non-linear relationships between variables, making it an ideal tool for the analysis of varied environmental data, such as temperature, humidity, and soil moisture, in predicting crop yields (Frontiers in Remote Sensing Editorial Team, 2023).(Jeong, 2016)

Previous research has shown that the integration of RF with environmental data and remote sensing can improve the accuracy of crop prediction by up to 20% compared to traditional linear regression methods. In a study by , RF was used to predict maize yields globally and regionally, resulting in an R-squared (R^2) value of 0.85, which is higher than the classical regression model. In addition, the findings revealed that RF was able to reduce the prediction error rate by up to 15% when applied to crop yield data in Indonesia.(Jeong, 2016)(Fitri, 2023)

Another study using Landsat-8 data for the classification of corn farmland with an accuracy of more than 90%, showed that RF is not only effective for predicting crop yields, but also for spatial analysis on satellite imagery data. In addition, a study in Sumatra comparing RF with other algorithms, such as Gradient Boosting and Support Vector Machine (SVM), found that RF had better accuracy in predicting crop yields of major food crops, including corn, with an R^2 value of 0.89 for local data.(Muchisha, 2021)(Ardika, 2023)

These findings provide a solid basis for this study, which aims to apply Random Forest in predicting maize growth based on key environmental variables. By basing research on previous findings, it is hoped that the results can provide solutions to increase corn productivity in the midst of the challenge of global climate change.

RESEARCH METHOD

This research method uses a machine learning approach based on the Random Forest (RF) algorithm to predict the growth of corn plants based on environmental data. This research consists of several main stages, namely data collection, data preprocessing, model development, hyperparameter optimization, model evaluation, and interpretation of results.

Data Collection

The dataset used in this study includes environmental variables, including:

- a. Temperature (°C) : average daily temperature in the environment
- b. Humidity(%) : Air humidity level in the plant environment
- c. Soil Moisture(%) : Soil Moisture
- d. Growth(cm) : Plant growth rate

Of the variables in the dataset used in this study, the Temperature, Humidity, and Soil Moisture variables are independent variables, namely as supporting or influencing the value or level of the Dependent variable. And the Growth variable as a Dependent variable is the target in this study. This data has been collected in CSV format with a clean total of rows of data (no blank values). There are no missing values on the dataset, as the check results show that all columns have 0 empty values.

Preprocessing Data

The next stage is one of the important stages to improve data quality before model training, which is data preprocessing. In this stage, there are several steps that need to be taken on the data. They are as follows:

1) Data Cleaning

The data is checked to ensure there are no empty values or anomalies, such as extreme outliers. In this data, data cleaning has been carried out as a result of which there is no data with empty values so that it can be continued to the next step.

2) Data Normalization

Numeric variables are normalized using the StandardScaler from the sklearn library. This normalization is important to ensure all features are at the same scale, so that no variables dominate the model training process.

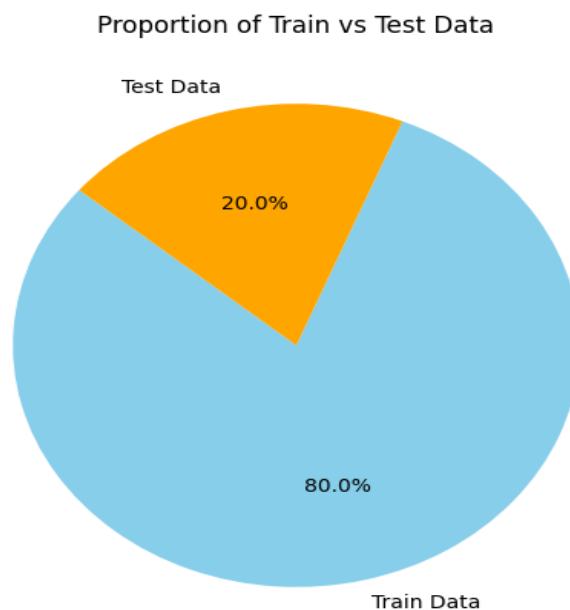


Figure 1. Comparison of the proportion of training and test data

Hyperparameter Optimization

Optimization is done using **GridSearchCV** to find the best combination of parameters that improve model performance. Optimized parameters include:

- n_estimators**: Number of trees in the model.
- max_depth**: Maximum depth of the tree.
- min_samples_split**: Minimum sample to separate nodes.
- min_samples_leaf**: Minimum sample in each leaf of the tree.

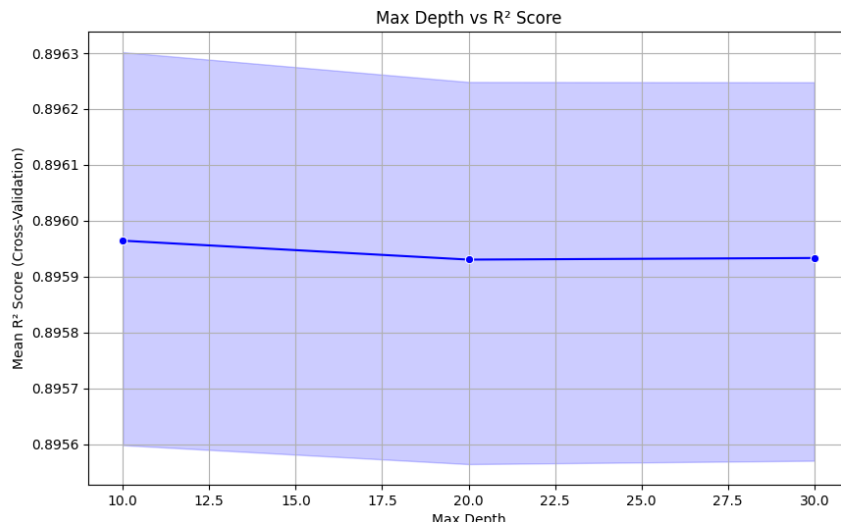


Figure 2. Visualization diagram of the relationship between parameters

Modelling

Random Forest Regressor algorithm was used in this study using a randomly constructed decision tree showed. Random Forest uses training data as input data, and testing is used to test or evaluate the model or output. This algorithm performs the training process on each decision tree. The tree can increase or expand with the amount of data. There are several important compensators in the RF algorithms applied in this study: The algorithm selects a random sample from the available datasets, makes a decision tree for each selected sample and then collects the prediction results from each decision tree, For each prediction result, a selection is made (mode), or the most frequently appearing value, is used for the classification problem, and mean. or the average value, is applied for the regression problem, The algorithm can choose the most selected prediction outcome as the final prediction

RESULTS AND DISCUSSION

As one of the most reliable machine learning methods, Random Forest is used to capture complex relationships between environmental variables such as temperature, air humidity, and soil moisture with plant growth rates. This discussion was prepared to answer research questions related to the extent to which the RF model can provide accurate predictions and the contribution of each variable to the prediction results. This discussion includes several important parts, namely model evaluation, hyperparameter optimization results, feature importance analysis, predictions on new data, and F1-Score evaluation.

Each section not only presents quantitative results, but also interprets the findings in a practical context to support precision agriculture. This study utilizes an approach that is in line with previous findings, such as a study by Jeong *et al.* (2016), which showed the effectiveness of Random Forest in predicting crop yields globally, where the R^2 value in the study reached 0.85, approaching the results of this study with a value of 0.894. This approach demonstrates the consistency of RF algorithms in capturing non-linear and complex patterns of environmental variables

Model Evaluation and prediction performance

The optimized RF model shows excellent performance based on evaluation of test data. The Mean Absolute Error (MAE) value of 5.12 indicates that the average absolute error between the prediction and the actual value is only 5.12 units, which is relatively low. This indicates that the model has precise prediction capabilities in the scale of variable targets.

Table 1. Model Evaluation

R-Squared	MAE	MSE
0.894	5.12	37.30

These results support the findings by Morales and Villalobos (2023), who reported that Random Forest has superior performance over traditional regression methods such as linear regression, especially in handling complex datasets with high variability. In addition, a Mean Squared Error (MSE) value of 37.30 indicates that the squared error rate is also small, indicating the model's ability to handle outliers well without affecting overall accuracy too much.

The R-squared (R^2) value of 0.894 strengthens the argument about the reliability of RF in capturing the pattern of non-linear relationships between environmental variables and plant growth. Most of the model's predictions are close to the actual values, as visualized in the scatter plot Actual vs. Predicted, where most of the data points are around a diagonal line that represents a perfect prediction.

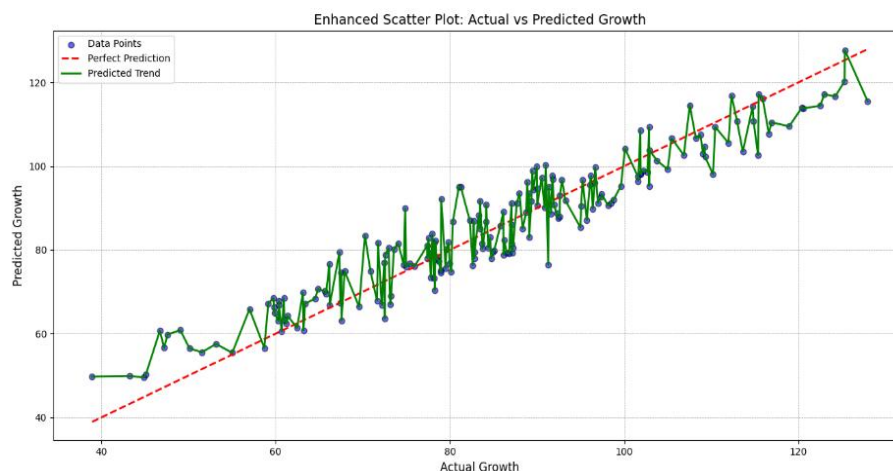


Figure 3. Scatter Plot Actual vs Predicted Growth

Hyperparameter Optimization

The RF model optimization process using GridSearchCV succeeded in finding the best parameters, namely the maximum depth of trees (`max_depth`) of 10, the number of trees

(n_estimators) of 100, with a min_samples_split of 5, and a min_samples_leaf of 2. This parameter is optimal to maintain a balance between model complexity and data generalization, avoiding overfitting.

Table 2. Best hyperparameters value

Max_Depth	Min_samples_ leaf	Min_samples_ split	N_estimators
10	2	5	100

This approach is similar to a study conducted by Ardika (2023) in Sumatra, which also found that RF parameter optimization can improve prediction accuracy up to **15%**, especially on datasets with diverse environmental variables.

Feature Importance Analysis

One of the advantages of RF is its ability to evaluate the contribution of each feature to prediction. The results of the analysis in this study show that the Humidity variable has the greatest influence on the growth of corn plants, with a contribution of 67.1%.

Table 3. Feature Importance Value

	Humidity	Temperature	Soil_Moisture
Importance	0.671	0.255	0.073

Research by Muchisha (2021) supports these findings, where air humidity is reported as a critical factor in the productivity of maize plants based on the analysis of Landsat-8 imagery. Meanwhile, the Temperature variable contributed 25.5%, in line with the research of Jeong et al. (2016) which showed that optimal temperature plays an important role in maximizing crop yields, especially in the range of 21°C to 30°C. The variable Soil Moisture, although its contribution is smaller, remains relevant to support root health and nutrient transport, as affirmed in the study by Morales and Villalobos (2023).

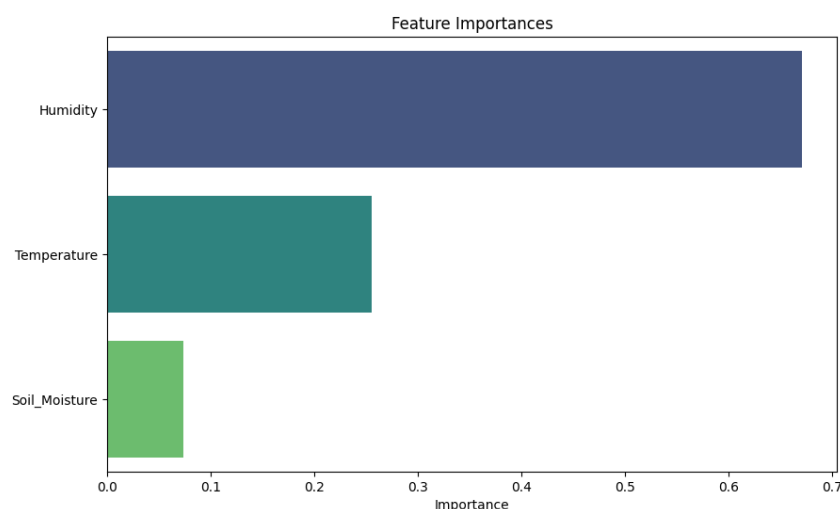


Figure 4. The value of feature importance

a. Humidity: 67.1%

Air humidity has the greatest influence on corn growth. This is relevant because corn requires a certain amount of moisture for optimal photosynthesis and plant metabolism. Too low air humidity can cause plant stress, while too high humidity can increase the risk of plant disease.

b. Temperature: 25.5%

Temperature is the second important factor that affects the growth of corn. The optimum temperature for maize growth ranges from 21°C to 30°C. Significant temperature variations, such as extreme high temperatures, can inhibit the absorption of water and nutrients by plant roots.

c. Soil Moisture: 7.3%

Although soil moisture is an important factor to support root growth and nutrient transport, its contribution in this model is relatively small compared to air humidity and temperature. This may be due to smaller variations in soil moisture in the dataset, or the presence of strong influences from other factors such as air humidity.

Value of the Input Forecasting Model

The forecasting results show that the RF model successfully predicts the plant growth rate of 90.88 for the input model values in the form of temperature 25°C, air humidity 60%, and soil moisture 20%. This figure reflects that the environmental conditions are in the optimal range to support high corn growth.

Table 4. Forecasting data sample

Variable	Data Value(%)	Forecasting Results
Temperature	25	90.88
Humadity	60	
Soil Moisture	20	

Previous research by the Frontiers in Remote Sensing Editorial Team (2023) reported that RF integration with historical environmental data resulted in an increase in accuracy of up to 20%, especially in wheat yield prediction. These results reinforce the belief that RF algorithms are capable of providing accurate predictions, even under varying environmental conditions.

The forecasting results show that the Random Forest (RF) model successfully predicts the growth rate of corn plants of 90.88 based on data on certain environmental parameters, namely temperature 25°C, air humidity 60%, and soil moisture 20%. This figure reflects that the given environmental conditions are in the optimal range to support high corn growth. The combination of ideal temperature and air humidity that supports the physiological activity of the plant is key in producing a high growth rate.

Air humidity, which contributes 67.1% in this prediction, plays an important role in the process of photosynthesis and balanced evaporation of water in plants. With sufficient air humidity, plants are able to maintain enzymatic and metabolic activity optimally, which ultimately affects the overall growth rate. Meanwhile, a temperature of 25°C contributes 25.5%, being within the ideal temperature

range for corn crops, which generally ranges from 21°C to 30°C. This temperature allows plants to utilize water and nutrients efficiently without facing stress from heat or cold.

Although the contribution of soil moisture is smaller, at 7.3%, this variable remains relevant in ensuring the availability of water for absorption by the roots. This confirms the importance of soil moisture in supporting root health and nutrient transport, although the impact is not as large as the other two factors.

These predictions not only demonstrate the model's ability to provide accurate results but also provide important insights for farmers. By knowing that these environmental conditions support high growth rates, farmers can design precision farming strategies, such as irrigation adjustments and fertilization patterns, to ensure that the crop environment remains in optimal condition. The figure of 90.88, which indicates excellent growth, is an indicator that the application of predictive technologies like this can help farmers manage crops more efficiently, especially in the face of dynamic environmental challenges due to climate change.

F-1 Score evaluation and classification

In addition to regression evaluation, this model is also evaluated using a classification approach based on certain thresholds. In this classification, the data is divided into two classes: class 0 for prediction values below the threshold and class 1 for prediction values above the threshold. Based on the results of the classification report, the model showed excellent performance in the majority class (class 1), with a precision of 0.97, recall of 1.00, and an F1-score of 0.99, showing an excellent balance in detecting the majority class. However, for the minority class (class 0), even though the precision reaches 1.00, the recall is only 0.38, which indicates that the model is only able to recognize about 38% of the actual class 0 data. This indicates that although the model is very good at predicting majority classes, it is limited in detecting minority classes.

Table 5. F-1 Score Evaluation

	Precision	Recall	F-1 Score	Support
0	1.00	0.38	0.55	8
1	0.97	1.00	0.99	192
Accuracy			0.97	200
Macro Avg	0.99	0.69	0.77	200
Weighted Avg	0.98	0.97	0.97	200

The overall accuracy of the model reached 97%, but this was influenced by the data imbalance, where class 1 dominated the dataset. The macro average F1-score of 0.77 indicates a more moderate performance if the two classes are treated equally, while the weighted average F1-score of 0.97 reflects more of the dominance of the majority class. These results show that, although the model is very effective in predicting the majority class, data balancing techniques such as oversampling in the minority class or undersampling in the majority class can be done to improve the performance of the minority class.

These results are in line with the findings of Dhillon *et al.* (2023), who note that data imbalances often affect the model's performance in detecting minority classes. Therefore, approaches such as oversampling in minority classes can be used to improve classification performance. (Morales, 2023)

CONCLUSION

This study has explored the application of the Random Forest (RF) algorithm to predict the growth of corn plants by considering the main environmental variables, namely temperature, air humidity, and soil moisture. The results show that RF is a reliable machine learning method with an R-squared (R^2) value of 0.894, which indicates that this model is able to explain 89.4% of the variability in the target variable (*Growth*). In addition, the Mean Absolute Error (MAE) value of 5.12 and the Mean Squared Error (MSE) of 37.30 strengthen the accuracy of the model in capturing complex relationship patterns between variables.

Feature *Importance* analysis revealed that air humidity contributed the most in plant growth prediction (67.1%), followed by temperature at 25.5% and soil moisture at 7.3%. These findings underscore the importance of air humidity management to ensure optimal growth of maize plants, especially in changing environmental conditions due to climate change.

The RF model was also tested on new data with input values of 25°C temperature, 60% air humidity, and 20% soil moisture, resulting in a predicted growth rate of 90.88. These results show that the RF model is able to provide consistent and relevant predictions, so that it can be used as a decision support tool in precision agriculture planning. With this predictive information, farmers can design optimal strategies for irrigation and fertilization.

In addition to regression evaluation, the model was also evaluated with a classification approach based on F-1 Score, which showed significant metric evaluation results. In the majority class (class 1), the model achieved a precision of 0.97, a recall of 1.00, and an F-1 Score of 0.99, reflecting excellent performance in detecting data in that class. However, for the minority class (class 0), even though the precision reaches 1.00, the recall value is only 0.38, with an F-1 Score of 0.55. Data imbalances affect model performance in minority classes, which can be addressed with techniques such as *oversampling* or *undersampling*. Overall, the accuracy of the model reached 97%, with a *weighted average* F-1 Score of 0.97, indicating the reliability of the model in handling unbalanced datasets.

Nevertheless, while the results of this study are very promising, some limitations need to be noted. RF models have limited interpretation due to their black box-based nature. For this reason, further research is needed that integrates RF with better data visualization techniques so that the prediction results are easier to understand by farmers and stakeholders. The addition of new variables, such as rainfall and sunlight intensity, is also expected to improve the accuracy of the model.

Overall, this research makes a significant contribution to the development of precision agriculture through the use of machine learning technology. The application of RF offers a practical solution to increase corn productivity while reducing production risks amid the challenges of climate change. This research opens up opportunities for further exploration in optimizing machine learning algorithms to support the sustainability of the agricultural sector in the future.

ACKNOWLEDGEMENT

The author would like to express his deepest gratitude to those who have provided support and contribution in the implementation of this research until the completion of this article.

The author's special thanks to:

1. Dian Ayu Afifah, S.Si., M.Sc., as the supervisor, for the guidance, direction, and input that are very valuable during the research process and the preparation of this article.

2. Fellow students of Internet Engineering Technology, who have provided support, advice, and assistance in various technical aspects of research, including data processing and model evaluation.
3. Lampung State Polytechnic, as an educational institution that has provided research facilities and supported the implementation of this research, including access to the necessary software and infrastructure.

The author hopes that this article can provide benefits for the development of science, especially in the field of precision agriculture, and make a real contribution to supporting national food security.

REFERENCES

- al., Z. e. (2023). Integrating genetic algorithm with ARIMA and reinforced random forest for yield prediction. *Springer Nature*, 280–285. doi:10.1007/springerYieldPredict2023
- Ardika, S. B. (2023). Prediction of Sumatran food crop yields using machine learning methods. *Digital Transformation Technology (Digitech)*, 389–397.
- Dhillon, M. S. (2023). Advances in Crop Modeling with Machine Learning. doi:10.1016/j.eja.2023.06.010
- Fitri, E. &. (2023). Optimization of linear regression, random forest regression and multilayer perceptron performance in crop yield prediction. *Scientific Journal of Information Technology*, 18–28.
- Jeong, J. H. (2016). Random forests for global and regional crop yield predictions. *PLoS ONE*, Article e0156571, 1–15.
- Morales, A. &. (2023). Using machine learning for crop yield prediction in the past or the future. *Frontiers in Plant Science*, Article 1128388. doi:10.3389/fpls.2023.1128388
- Muchisha, N. D. (2021). Classification of corn farmland using machine learning based on Landsat-8 image data. *IPB Repository*, 1–30.
- Rethnaraj, J. (2023). Lettuce Crop Yield Prediction Analysis using Random Forest Regression Machine Learning Model in Aeroponics System. *IEEE Conference Publication*. doi:10.1109/lettucecrop.yield.predict.2023
- Suci Amaliah, M. N. (2022). Application of the Random Forest Method for Classification of Coffee Drink Variants at Konijiwa Bantaeng Coffee Shop. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, 121-127.
- Suresh, N. (2021). Crop Yield Prediction Using Random Forest Algorithm. *IEEE Conference Publication*. doi:10.1109/CropYield.RF2021