VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i1.6476

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

APPLYING TREE BASED MODEL FOR CROP RECOMMENDATION SYSTEM BASED ON SOIL PARAMETERS AND WEATHER CONDITIONS

Asrul Abdullah[1*]; Muhammad Iwan[2]; Sinta Rama Dani[1]

Department of Informatics[1], Department of Mechanical Engineering[2]
Universitas Muhammadiyah Pontianak, Pontianak, Indonesia[1,2]
https://www.unmuhpnk.ac.id[1,2]
asrul.abdullah@unmuhpnk.ac.id[*]; muhammad.iwan@unmuhpnk.ac.id; 212220091@unmuhpnk.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract—*The massive population in Indonesia needs to be supported by various sectors so that the population's needs are met. One of these sectors is agriculture. The problems are unpredictable climate change and weather and changes in land use from previously agricultural land to housing. In addition, plant quality is also influenced by soil quality and other abiotic factors, comprising rainfall, temperature, and humidity. Plant quality affects the increase in crop yields. A plant recommendation system based on plant parameters must help farmers determine the best plants according to agricultural land conditions. The recommended plants to be used include mango, cotton, rice, mungbeans, and apple. This work aims to create a plant recommendation system utilizing criteria related to plant requirements through a machine learning methodology. The stages in this study start with data collection, preprocessing, partitioning, modelling, performance evaluation, and a recommender system. This study's results indicate that the Random Forest method achieved the best accuracy at 0.9981, followed by XGBoost at 0.9909 and Decision Tree at 0.9873. The system provided recommendations for plant types based on user input.*

Keywords: *agricultural, decision tree, random forest, recommendation system, xgboost.*

Intisari—*Populasi penduduk di Indonesia yang sangat besar perlu ditunjang oleh berbagai sektor agar kebutuhan penduduk terpenuhi. Salah satu sektor tersebut adalah pertanian. Masalah yang ada perubahan iklim dan cuaca yang tidak bisa diprediksi dan perubahan fungsi lahan dari sebelumnya merupakan lahan pertanian berubah menjadi perumahan. Selain itu, kualitas tanaman juga dipengaruhi oleh kualitas tanah dan faktor abiotik lain seperti curah hujan, temperature dan kelembaban. Kualitas tanaman mempengaruhi peningkatan hasil panen. Sistem rekomendasi tanaman berdasarkan parameter tanaman perlu untuk dibuat agar membantu petani dalam menentukan tanaman terbaik sesuai dengan kondisi lahan pertanian. Adapun rekomendasi tanaman yang diterapkan antara lain mango, cotton, rice, mungbeans, apple. Tujuan dari penelitian ini adalah sistem rekomendasi tanaman berdasarkan parameter kebutuhan tanaman dengan menggunakan pendekatan machine learning. Metode dalam penelitian ini berawal dari pengumpulan data, data preprocessing, data partition, modelling, evaluasi performance dan recommender system. Hasil dari penelitian ini adalah akurasi tertinggi diraih oleh algoritma Random Forest dengan skor 0.9981, diikuti oleh XGBoost dengan skor 0.9909 dan Decision Tree dengan skor 0.9873. Sistem berhasil bekerja dengan baik memberikan rekomendasi jenis tanaman berdasarkan inputan dari pengguna.*

Kata Kunci: *pertanian, pohon keputusan, random forest, sistem rekomendasi, xgboost.*

INTRODUCTION

The population of Indonesia in 2020 was ± 270 million people. Compared to 2010, there was an increase of 32.56 million people (BPS, 2020). The increase in the population of Indonesia is also in line with the fulfilment of basic needs, especially in the agricultural sector. This sector can be a problem if

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6476

not managed seriously, as it is related to suitable crops for site-specific soils. Inaccurate and imprecise crop recommendations may result in significant material and capital losses [1]. Issues in this sector include limited farmland and rapid weather changes. Some crops, such as strawberries and apples, have been experimented with in tropical regions. In reality, these crops are not suited to tropical and dry climates. While the crops are able to adapt to the soil and climate with conditions such as fertilizer and optimal temperature control. In addition to these factors, plant quality is essential in increasing crop production. Climate change and soil quality interact to impact agricultural regions' capacity to produce enough food[2].

Abiotic factors greatly influence plant quality. Abiotic factors encompass physical and chemical factors. Physical elements include vibration, noise, and climate and weather conditions. Chemical aspects such as sulfur dioxide, fluorine, nitrogen dioxide, fertilizers, and pesticides [3]. Crop production highly depends on land area, pest attacks, plant diseases, fertilizers, and soil parameters. Soil parameters involve soil pH, electrical conductivity, nitrogen, phosphorus, potassium, humidity, and temperature. To solve this issue, a recommendation system is required to meet the objectives of crop effectiveness in accordance with significant crop metrics. The novelty of this research is to assist farmers in determining the perfect crops for a particular agricultural land.

This paper contributes to crop recommendation by applying various Machine Learning algorithms. The web application has been made more accessible and easier to use. The performance evaluation of this model uses a confusion matrix. The crop recommendation system employs diverse ML algorithms and AI to guide farmers in cultivating optimal crops according to their particular agricultural conditions.

AI is a domain of computer science that can address the issue of harvest efficiency, including agriculture, where it facilitates real-time monitoring of crop yields and soil conditions, diagnoses plant diseases, and forecasts optimal planting times [4]. The process of developing and implementing computer algorithms that can learn from past data or experience to produce models using statistical approaches is known as machine learning[5]. It classifies crop selection by applying machine learning methodologies, such as LR, SVM, KNN, DT, RF, Bagging, AB and Extra Trees (ET). Random Forest has the best accuracy with 99,31% among these models[6].

Additional research indicates that soil type and location can be used to forecast crop production based on the user's selection. The machine learning methods are SVM, ANN, RF, MLR, and KNN. Of the models, Random Forest exhibits the greatest accuracy at 90%. [7]. As detailed by [8], crop classification and forecasting methods comprise DT, SVM, KNN, RF, Naïve Bayes, and XGBoost, all based on soil nutrient levels. This study employs confusion metrics to determine the achievement of the model outcomes. Ensemble methods, including the XGBoost model, improve efficacy in crop type prediction.

This study gathers data on soil properties and proposed weather conditions, including rainfall, humidity, temperature, sunshine, and pH levels. In [8] employed machine learning methodologies, including SVM, SVM with Kernel, and DT, to classify three categories: rice, wheat, and sugarcane. This study effectively conducted graph modifications and regression analysis to identify statistical relationships between nutrition and atmospheric variables. The primary challenge farmers encounter in crop selection is climate change. Machine learning algorithms have demonstrated optimal efficiency in forecasting appropriate crops to enhance yields by selecting the correct criteria, which include humidity, temperature, rainfall, pH, and NPK. Selecting characteristics that align with the appropriate Machine Learning algorithm is essential.

Crop losses can be mitigated by selecting appropriate crops[9]. In [10], an innovative crop yield selection that utilizes ML methods to integrate meteorological conditions and soil factors was introduced. The meteorological evaluation utilizes LSTM RNN, whereas the crop selection action employs the RF Classifier. According to the final findings, the LSTM RNN model performs better than the artificial neural network. In [11], the authors create a machine learning algorithm that assists farmers in choosing crops wisely. Start by collecting information on weather, humidity, pH, temperature, nitrogen, phosphorus, and potassium values. The model encompassed Gaussian Naive Bayes (GNB), SVM, RF, and DT. The result is that GNB achieves an accuracy of 99%. In [12], a crop advice framework utilizing 250 sensors was deployed in various locations throughout Maharashtra, India. The collected data were evaluated utilizing multiple machine learning methods, achieving 95% accuracy with the random forest (RF) approach. In[13], it is possible to make accurate soybean production forecasts as early as the pod-setting phase. We used feature significance and Shapley additive explanations (SHAP) approaches to assess how input features affected the XGBoost model during the training and prediction phases, respectively.

This research employed three machine learning classifiers, specifically tree-based classifiers: decision trees, random forests, and XGBoost. Alternative classifiers, equally SVM, MLP, and KNN, exhibit suboptimal performance on extensive datasets, are significantly influenced by their parameter configurations, and yield unstable models that lack interpretability [14][15]. This research aims to develop a plant recommendation system utilizing fundamental plant attributes and diverse machine learning techniques.

MATERIALS AND METHODS

This research activity began with data collection, wherein soil parameters and weather conditions were obtained to establish a foundation for classification. Thereafter, data preparation was performed to ensure data integrity and reliability. The primary objective during preprocessing is to remove missing values and outliers utilizing the interquartile range (IQR) approach, while also assessing data consistency. After preparing the dataset, it was sent to the classification models phase. This phase entailed using multiple tree-based models, including DT, RF, and XGB. Multiple model were developed and evaluated to determine their accuracy in categorizing crops. Finally, during the model evaluation phase, the effectiveness of each classification model in categorizing crop production was evaluated using confusion measures.

Data Acquisition

This research uses a dataset retrieved from the Kaggle dataset with the link source (https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset). The dataset was selected because it provides pertinent agricultural data to classify crops through varied machine learning algorithms, such as methods-based tree.

Table 1. Sample dataset

N	P	K	temp	humi	pH	rainfall	label
90	42	43	20.87	82.00	6.50	202.93	Rice
85	58	41	21.77	80.31	7.03	226.65	Rice
60	55	44	23.00	82.32	7.84	263.96	Rice
74	35	40	26.49	80.15	6.98	242.86	Rice
78	42	42	20.13	81.60	7.62	262.71	rice

Source : (Research Result, 2025)

Preprocessing Data

Some crucial steps are taken during the data preprocessing stage to ensure the consistency and quality of the dataset that will be used. The Interquartile Range (IQR) approach finds outliers; values outside the criteria Q1−1.5×IQR or Q3+1.5×IQR are outliers. Outliers are addressed to prevent extreme values from affecting the analysis and the model. Furthermore, absent values are addressed utilizing the data. To determine which variables have missing values, use the isnull().sum() function. The mean or median was used to impute missing values to ensure data completeness and preserve an appropriate distribution where they were found. Additionally, duplication is verified using the data.

The duplicated() task guarantees no data is replicated, preventing redundancy and preserving model accuracy. Improving model performance and the precision of research findings requires efficiently handling duplicate data and missing variables. This crop dataset's class distribution is balanced, despite the common class imbalance problem in agricultural datasets. This means that equally oversampling or undersampling is not mandatory now.

Classification Models

After the data preprocessing, crop yield is analyzed and categorized in this study using a classification model. These methods are anticipated to enhance the precision of crop yield classification.

1. Decision Tree

This algorithm is an extensively employed supervised learning technique frequently utilized in machine learning. The Decision Tree is a widely recognized method and one of the most often employed models in classification tasks[16]. DT divides a dataset into more manageable subgroups using decision rules built from input attributes [17]. The decision tree approach is designed to enhance the homogeneity or purity of its segments by reducing impurity metrics involved in entropy or Gini impurity. These measurements evaluate a specific data set's degree of disorder or unpredictability. Based on the decrease via the split, impurity is attained. This method determines the best set of traits and values to partition the data at each node.

The process continues until a stopping requirement is met, such as reaching a leaf node's maximum depth or minimum sample size[18]. The decision tree consists of three categories of nodes: the root, decision, and leaf nodes. The utility of the decision tree resides in its relevance to both regression and classification tasks. The response of their base learners primarily dictates the sensitivity of ensemble models to the scaling technique. Consequently, when the base model is inherently insensitive to scaling, such as a decision tree, the ensemble also exhibits stability across different scaling transformations[19].

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6476

2. Random Forest

The algorithm employs ensemble learning, encompassing bagging, boosting, and stacking. The integration of data augmentation techniques with ensemble learning can markedly enhance classification efficacy on imbalanced datasets [20]. Random forest constructs several trees using a random subset of all features at a one-by-one split to decrease variation among correlated trees. The system uses the mean value to enhance forecast accuracy and mitigate overfitting [21].

3. XGBoost

Extreme Gradient Boosting (XGBoost) can be utilized to predict the annual rice production in Bangladesh. XGBoost outperforms ARIMA [22]. The XGBoost wherein each tree learns from its predecessor and influences the subsequent tree to enhance model completion. Each new tree is associated with the error rate of the preceding prediction tree, continuing until a specified number of trees is reached. At this point, the sample scores are necessary to forecast the score for a particular sample.

The ultimate predicted score is the aggregate of sample values across various trees. XGBoost provides benefits in managing imbalanced datasets and mitigates the likelihood of overfitting by employing sophisticated regularisation methods. XGBoost is a very effective and extensively employed ensemble method for various classification applications [23], including agricultural yield classification.

Tree-based models (DT, RF, and XGB) excel at managing skewed features and heterogeneous data types. They exhibit resilience to noise and are appropriate for visualizing feature interactions. Among tree models, ensemble methods such as RF and XGB excel by more effectively managing outliers[24].

Model Evaluation

Model evaluation is a critical stage in machine learning to assess the model's effectiveness in data classification. Metrics are obtained from the confusion matrix [25].

Table 2. Confusion matrix

	Predictive Positive	Predictive Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Source: (Research Result, 2025)

Accuracy is the fraction of correct predictions. Accuracy measures the classifier's condition prediction, as shown in Equation 1 :

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Precision is the fraction of correct predictions among all predictions. Precision can be expressed in the formula Equation 2 :

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Sensitivity/recall refers to the model's capacity to detect positive cases accurately. It is calculated as the ratio of actual positive forecasts to all positive predictions. The formula may be articulated in terms of recall in Equation 3 :

$$Sensitivity/Recall = \frac{TP}{TP+FN} \quad (3)$$

The F-Score is a statistic that equally evaluates both Precision and Sensitivity. It denotes the harmonic mean of precision and sensitivity as articulated in Equation 4 :

$$F1\ score = \frac{2*Precision*Sensitivity}{Precision+Sensitivity} \quad (4)$$

RESULTS AND DISCUSSION

Dataset

The dataset has 2200 instances with seven features: nitrogen, phosphorus, potassium, temperature, humidity, pH, and rainfall. One designation is referred to as label/crop type. The label comprises 22 categories: rice, maize, chickpeas, kidney beans, pigeonpeas, moth beans, mungbean, black gram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, and coffee.

Exploratory Data Analysis

This dataset is a multi-class classification problem. The 22 classes combined in 1 label have the same distribution. Each class has 100 data. The features contained in this dataset are numerical. The univariate analysis carried out was count, mean, standard deviation, min, quartile 1, median (50%), quartile 3, and maximum, as shown in Table 3.

Table 3. Univariate analysis

	N	P	K	temp	humi	ph	rainfall
count	2200	2200	2200	2200	2200	2200	2200
max	140	145	205	43.67	99.98	9.93	298.56
75%	84.25	68	49	28.56	89.94	6.92	124.26
mean	50.55	53.36	48.14	25.61	71.48	6.46	103.46
50%	37	51	32	25.59	80.47	6.42	94.86
25%	21	28	20	22.76	60.26	5.97	64.55
std	36.91	32.98	50.64	5.06	22.26	0.77	54.95
min	0	5	5	8.82	14.25	3.504	20.21

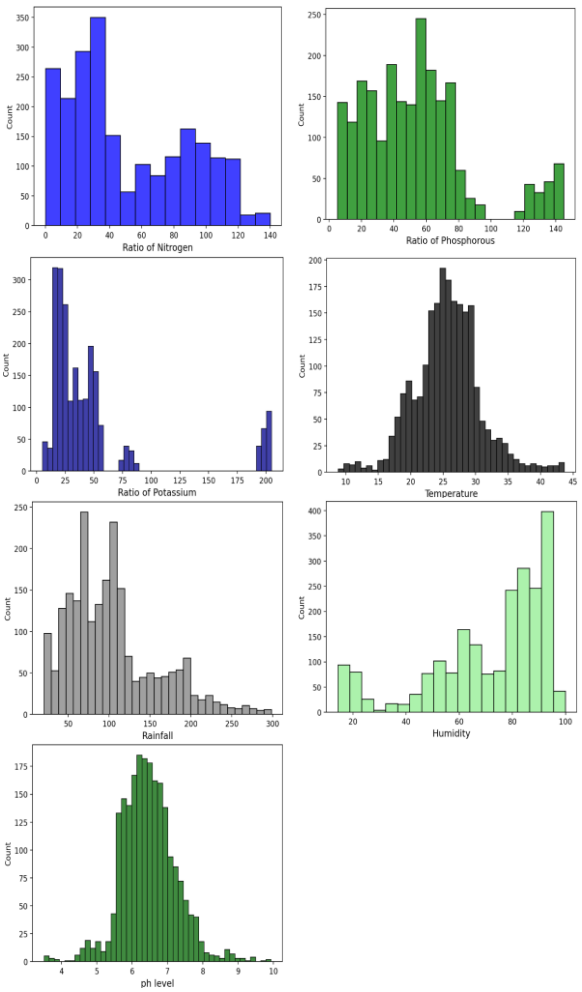
Source : (Research Result, 2025)

Table 3 presents that there is no missing or empty data. The distribution of data on each feature, such as N is mainly at an average value of 50,551 ±

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i1.6476

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

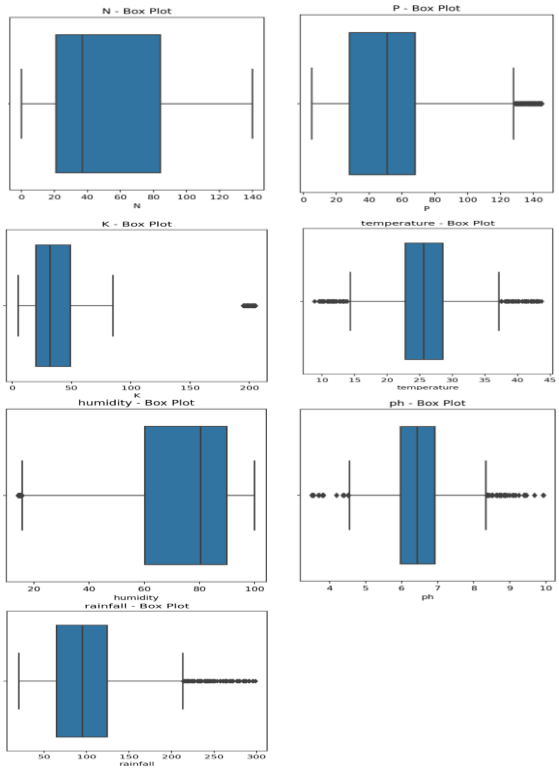
36,917 (standard deviation), P is at an average value of 53,362 ± 32,985 (standard deviation), K is at an average value of 48,149 ± 50,647 (standard deviation), temperature is at an average value of 25,616 ± 5,063 (standard deviation), humidity is at an average value of 71,481 ± 22,263 (standard deviation), pH is at a value of 6,469 ± 3,504 and rainfall is at an average value of 103,463 ± 54,958 (standard deviation). The distribution of data across all features is illustrated in Figure 1.



Source : (Research Result, 2025)
Figure 1. Data distribution of each feature

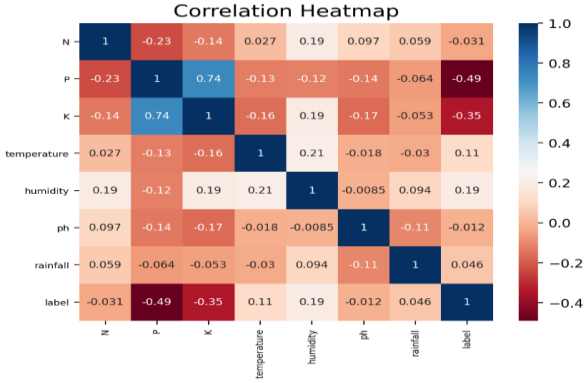
Data preparation

In this stage, the dataset will be examined for null values, NaN, and missing values. There are no missing values, NaNs or null values among the 2200 data. In addition, it is necessary to check outlier detection using Boxplot. Of the seven features, only the N feature does not have an outlier, while the P, K, temperature, humidity, pH, and rainfall features do. For more details, see Figure 2.



Source : (Research Result, 2025)
Figure 2. Outlier detection

Outliers in the P, K, temperature, humidity, pH, and rainfall features will be handled using the IQR (interquartile range) method. Data that is far from the IQR value is deleted. The result is that 16.09% of the data is deleted, so the total data becomes 1846. The amount of data for each class in the label/crop type feature is also balanced so that there is no need for oversampling/undersampling. The last stage is the correlation between features using a heatmap, as shown in Figure 3. Because the label of this dataset is categorical data, encoding needs to be done with LabelEncoder so that the class name changes to sequential numbers.



Source : (Research Result, 2025)
Figure 3. Correlation between features using with Pearson method that visualizes with a Heatmap

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6476

Figure 3 illustrates the absence of a significant association, whether positive or negative, between the attributes and labels. The temperature and humidity features exhibit a mild positive association, with scores of 0.11 and 0.19, respectively, whereas the P and K features demonstrate a weak negative correlation, with scores of -0.49 and -0.35. The temperature and humidity variables exert a marginal influence on the label, demonstrating a positive correlation with it. The P and K characteristics exert a minimal impact on the label, exhibiting an initial positive boost followed by a subsequent decline. Moreover, the employed algorithm, namely tree-based models consisting of decision trees, random forests, and XGBoost, does not necessitate feature scaling techniques such as normalization and standardization, as it is not sensitive to such scaling.

Upon finalizing preliminary preprocessing tasks, including managing missing values and identifying outliers, the subsequent stage involves partitioning the sanitized dataset into two subsets, training and testing, via the Scikit-Learn library's (train_test_split) function. Because we will use a tree-based modeling algorithm, feature scaling and encoding (all features are numerical) are no longer necessary, and the split dataset will be directed to the modeling stage.

Data partition

The dataset is partitioned into training data, which comprises 70%, and testing data, which comprises 30%. The cumulative training data amounts to 1292 and 594.

Modeling

This stage uses three tree-based methods with parameters, as shown in Table 4.

Table 4. Algorithms with parameters

Algorithms	Parameter
Decision Tree	*criterion : 'entropy', splitter : best, max_depth : None, min_samples_split : 2, min_samples_leaf : 1,*
Random Forest	*n_estimators=100, *, criterion='entropy', max_depth=None, min_samples_split=2, min_samples_leaf=1*
XGBoost	*booster = gbtree, device=cpu, verbosity = 1, validate_parameters= True*

Source : (Research Result, 2025)

From the results of applying models with three algorithms, the accuracy obtained is shown in Table 5.

Table 5. Models result in an accuracy score

Algorithms	Accuracy
Decision Tree	0.9873
Random Forest	0.9981
XGBoost	0.9909

Source : (Research Result, 2025)

Table 5 shows that the Random Forest algorithm has the highest accuracy, followed by XGBoost and Decision Tree.

Performance Evaluation

Performance evaluation using a confusion matrix with four outputs can be obtained: accuracy, precision, recall/sensitivity, and f1-score. In this performance evaluation, three algorithms with the highest accuracy are used in the random forest algorithm, followed by xgboost and decision tree. Complete details can be seen in Table 6.

Table 6. Confusion matrix results from three algorithms

Algorithms	Accuracy	Precision	Recall	F1-score
Decision Tree	0.987	0.987	0.987	0.987
Random Forest	0.998	0.998	0.998	0.998
XGBoost	0.990	0.991	0.990	0.990

Source : (Research Result, 2025)

Table 6 shows Random Forest is the algorithm with the highest accuracy, precision, recall, and f1-score, followed by XGBoost and Decision Tree. Furthermore, to see the ability of this algorithm to provide recommendations for data that has never been seen and as an effort to avoid overfitting, cross-validation using *StratifiedKFold with n_split=5* is needed. XGBoost 0.9845, Random Forest 0.9934, and Decision Tree 0.9853. Next, the recommendation stage is by the parameters for unknown data. The following recommendations will be made based on the 7 latest input data, as shown in Table 7.

Table 7. Crop recommendation using Random Forest

Parameter							Crop
N	P	K	temp	humi	pH	rainfall	
14	20	40	30	45	4.20	120	Mango
49	28	93	36	91	6.25	157	Cotton
21	42	116	15	15	8.13	91	Jute
48	50	148	27	58	5.26	198	Mungbeans
72	37	98	30	78	6.22	239	Apple

Source : (Research Result, 2025)

The plant recommendation system performed well using three tree-based machine learning algorithms: accuracy> 98%, precision > 98%, recall > 98%, and f1-score > 98%. This dataset is a bit difficult because it shows true positives, false positives, and false negatives.

Table 8 is the confusion matrix of class 0 against other classes.

Table 8. Confusion matrix at class 0

True Class	Predicted Class	
	0 →Yes	(1 – 19) → No
0 → Yes	TP → 30	FN → 0
(1 – 19) → No	FP → 0	TN → 524

Source : (Research Result, 2025)

Table 8 shows accuracy, precision, recall, and f1-score can be calculated.

Accuracy (class - 0) = $\frac{TP+TN}{N} = \frac{30+524}{30+0+0+524} = \frac{554}{554} = 1$

Precision (class – 0) = $\frac{TP}{TP+FP} = \frac{30}{30+0} = 1$

Recall (class – 0) = $\frac{TP}{TP+FN} = \frac{30}{30+0} = 1$

F1-score (class – 0) = $\frac{2*1*1}{1+1} = 1$

The next example is the manual calculation in class 18, as shown in Table 9.

Table 9. Confusion matrix at class 18

True Class	Predicted Class	
	18 →Yes	(0-17 & 19) → No
18 → Yes	TP → 18	FN → 0
(0-17 & 19) → No	FP → 0	TN → 536

Source : (Research Result, 2025)

From Table 9, accuracy, precision, recall, and f1-score can be calculated.

Accuracy (class - 18) = $\frac{TP+TN}{N} = \frac{18+536}{18+0+0+536} = 1$

Precision (class – 18) = $\frac{TP}{TP+FP} = \frac{18}{18+0} = 1$

Recall (class – 18) = $\frac{TP}{TP+FN} = \frac{18}{18+0} = 1$

F1-score (class – 18) = $\frac{2*1*1}{1+1} = 1$

This dataset presents a multi-class problem with one label comprising more than two classes. The generated confusion matrix must compare class 1 with the other classes. The recommendation system can be seen on the site https://crop-recommendation-machine-learning.streamlit.app/

CONCLUSION

The plant suggestion system can function efficiently using three separate algorithms. The performance test demonstrated that the Random Forest algorithm achieved the top metrics for accuracy, precision, recall, and F1-score followed by XGBoost and Decision Tree. The performance assessment criteria are as follows: The accuracy is 0.9981, the precision is 0.9982, the recall is 0.9981, and the F1-score is 0.9982. The accuracy of XGBoost is 0.9909, the precision is 0.9912, the recall is 0.9909, and the F1-score is 0.9909. The Decision Tree demonstrates an accuracy of 0.9873, precision of 0.9878, and recall of 0.987.

ACKNOWLEDGEMENT

The authors thank the Institute for Research and Community Service (LPPM) of Muhammadiyah University of Pontianak for funding our research.

REFERENCE

[1] J. Cock, D. Jiménez, H. Dorado, and T. Oberthür, 'Operations research and machine learning to manage risk and optimize production practices in agriculture: good and bad experience', *Curr Opin Environ Sustain*, vol. 62, p. 101278, Jun. 2023, doi: 10.1016/j.cosust.2023.101278.

[2] L. Qiao *et al.*, 'Soil quality both increases crop production and improves resilience to climate change', *Nat Clim Chang*, vol. 12, no. 6, pp. 574–580, Jun. 2022, doi: 10.1038/s41558-022-01376-8.

[3] S. P. Raja, B. Sawicka, Z. Stamenkovic, and G. Mariammal, 'Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers', *IEEE Access*, vol. 10, pp. 23625–23641, 2022, doi: 10.1109/ACCESS.2022.3154350.

[4] E. Elbasi *et al.*, 'Artificial Intelligence Technology in the Agricultural Sector: A Systematic Literature Review', *IEEE Access*, vol. 11, pp. 171–202, 2023, doi: 10.1109/ACCESS.2022.3232485.

[5] S. O. Abioye *et al.*, 'Artificial intelligence in the construction industry: A review of present status, opportunities and future challenges', *Journal of Building Engineering*, vol. 44, p. 103299, Dec. 2021, doi: 10.1016/j.jobe.2021.103299.

[6] F. S. Prity *et al.*, 'Enhancing Agricultural Productivity: A Machine Learning Approach to Crop Recommendations', *Human-Centric Intelligent Systems*, vol. 4, no. 4, pp. 497–510, Sep. 2024, doi: 10.1007/s44230-024-00081-3.

[7] S. M. PANDE, P. K. RAMESH, A. ANMOL, B. R. AISHWARYA, K. ROHILLA, and K. SHAURYA, 'Crop Recommender System Using Machine Learning Approach', in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, Apr. 2021, pp. 1066–1071. doi: 10.1109/ICCMC51019.2021.9418351.

[8] R. Dash, D. K. Dash, and G. C. Biswal, 'Classification of crop based on macronutrients and weather data using machine learning techniques', *Results in*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

Engineering, vol. 9, p. 100203, Mar. 2021, doi: 10.1016/j.rineng.2021.100203.

[9] F. Shahbazi, S. Shahbazi, M. Nadimi, and J. Paliwal, 'Losses in agricultural produce: A review of causes and solutions, with a specific focus on grain crops', *J Stored Prod Res*, vol. 111, p. 102547, May 2025, doi: 10.1016/j.jspr.2025.102547.

[10] S. Rani, A. K. Mishra, A. Kataria, S. Mallik, and H. Qin, 'Machine learning-based optimal crop selection system in smart agriculture', *Sci Rep*, vol. 13, no. 1, p. 15997, Sep. 2023, doi: 10.1038/s41598-023-42356-y.

[11] P. S. Kiran, G. Abhinaya, S. Sruti, and N. Padhy, 'A Machine Learning-Enabled System for Crop Recommendation', in *The 3rd International Electronic Conference on Processes*, Basel Switzerland: MDPI, Sep. 2024, p. 51. doi: 10.3390/engproc2024067051.

[12] S. D. Shingade and R. P. Mudhalwadkar, 'Sensor information-based crop recommendation system using machine learning for the fertile regions of Maharashtra', *Concurr Comput*, vol. 35, no. 23, Oct. 2023, doi: 10.1002/cpe.7774.

[13] Y. Li *et al.*, 'A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering', *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, p. 103269, Apr. 2023, doi: 10.1016/j.jag.2023.103269.

[14] P. P. Šimović, C. Y. T. Chen, and E. W. Sun, 'Classifying the Variety of Customers' Online Engagement for Churn Prediction with a Mixed-Penalty Logistic Regression', *Comput Econ*, vol. 61, no. 1, pp. 451–485, Jan. 2023, doi: 10.1007/s10614-022-10275-1.

[15] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, 'Machine Learning Applications for Precision Agriculture: A Comprehensive Review', *IEEE Access*, vol. 9, pp. 4843–4873, 2021, doi: 10.1109/ACCESS.2020.3048415.

[16] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, 'Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung', *Informatik : Jurnal Ilmu Komputer*, vol. 18, no. 3, p. 239, Dec. 2022, doi: 10.52958/iftk.v18i3.4694.

[17] B. Charbuty and A. Abdulazeez, 'Classification Based on Decision Tree Algorithm for Machine Learning', *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[18] A. Arifuddin, G. S. Buana, R. A. Vinarti, and A. Djunaidy, 'Performance Comparison of Decision Tree and Support Vector Machine Algorithms for Heart Failure Prediction', *Procedia Comput Sci*, vol. 234, pp. 628–636, 2024, doi: 10.1016/j.procs.2024.03.048.

[19] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, 'The choice of scaling technique matters for classification performance', *Appl Soft Comput*, vol. 133, p. 109924, Jan. 2023, doi: 10.1016/j.asoc.2022.109924.

[20] A. A. Khan, O. Chaudhari, and R. Chandra, 'A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation', *Expert Syst Appl*, vol. 244, p. 122778, Jun. 2024, doi: 10.1016/j.eswa.2023.122778.

[21] M. Yousefi D.B. *et al.*, 'Classification of oil palm female inflorescences anthesis stages using machine learning approaches', *Information Processing in Agriculture*, vol. 8, no. 4, pp. 537–549, Dec. 2021, doi: 10.1016/j.inpa.2020.11.007.

[22] M. Noorunnahar, A. H. Chowdhury, and F. A. Mila, 'A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh', *PLoS One*, vol. 18, no. 3, p. e0283452, Mar. 2023, doi: 10.1371/journal.pone.0283452.

[23] A. R. Al Musyaffa, Y. Pristyanto, and N. Mauliza, 'Comparison Of Ensemble Methods For Decision Tree Models In Classifying E. Coli Bacteria', *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 10, no. 3, pp. 514–522, Feb. 2025, doi: 10.33480/jitk.v10i3.5972.

[24] H. Ali, I. Niazi, D. White, M. Akhter, and S. Madanian, 'Comparison of Machine Learning Models for Predicting Interstitial Glucose Using Smart Watch and Food Log', *Electronics (Basel)*, vol. 13, no. 16, p. 3192, Aug. 2024, doi: 10.3390/electronics13163192.

[25] D. Müller, I. Soto-Rey, and F. Kramer, 'Towards a guideline for evaluation metrics in medical image segmentation', *BMC Res Notes*, vol. 15, no. 1, p. 210, Dec. 2022, doi: 10.1186/s13104-022-06096-y.