Mathematical Modelling of Engineering Problems
Vol. 11, No. 11, November, 2024, pp. 3015-3024
Journal homepage: http://iieta.org/journals/mmep

Prediction of Horticultural Production Using Machine Learning Regression Models: A Case Study from Indramayu Regency, Indonesia

Mayanda Mega Santoni[1*], Didit Widiyanto[1], Desta Sandya Prasvita[1], Jayanta[1], Wan Suryani Wan Awang[2]

[1] Department of Informatics, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jakarta, South Jakarta 12450, Indonesia
[2] Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Besut 22200, Terengganu, Malaysia

Corresponding Author Email: megasantoni@upnvj.ac.id

https://doi.org/10.18280/mmep.111114

Received: 23 February 2024
Revised: 21 August 2024
Accepted: 3 September 2024
Available online: 29 November 2024

Keywords:
horticultural production, linear regression, random forest regression, gradient boosting regression, decision tree regression, mango, Indramayu

ABSTRACT

This study employs machine learning techniques to forecast horticultural production in Indramayu Regency, Indonesia, utilizing data from the Indramayu Regency Statistics Agency from 2009 to 2017. The variables under observation encompass mango fruit production volume, harvest area, rainfall, and the number of rainy days. Mango fruit production volume is the target variable, while the remaining data serves as features. Regression models comprise Linear Regression, Random Forest Regression, Gradient Boosting Regression, and Decision Tree Regression. The research unveils three key findings. Firstly, it underscores the significance of data preprocessing to eliminate noise or outliers, thereby enhancing the performance of regression models, as evidenced by amplified R^2 and reduced RMSE values alongside diminished MAPE. Elevated RMSE values highlight the presence of noise or outliers in unprocessed data. Secondly, it emphasizes the necessity of a representative test data proportion for precise prediction outcomes, as indicated by escalating MAPE and RMSE with increased test data proportions. Lastly, it shows the strong correlation between harvest area and mango production volume, culminating in commendable evaluation metrics. Among the regression models, Random Forest Regression emerges as the most robust, boasting the highest R^2 value and lowest RMSE, affirming its efficacy in this study.

1. INTRODUCTION

Horticultural production is vital in addressing food security and improving community welfare [1]. Horticulture is a crucial sector in Indonesia, encompassing a diverse array of plants, including vegetables, fruits, flowers, and ornamental species prized for their economic value. The country's expansive land and favorable climatic conditions provide an ideal environment for robust plant growth [2], particularly evident in regions like Indramayu. Here, the convergence of ample land and a supportive climate underscores the strategic importance of horticultural production. Moreover, the economic potential inherent in various horticultural crops further improves the welfare of Indramayu's people. Through effective horticultural practices, nutritional needs are addressed, and economic benefits are gained, leading to improved livelihoods and welfare for the people [3].

Indramayu Regency boasts vast resource potential, one of the most significant being the horticultural sector [4]. As a cornerstone of the region's agricultural landscape, horticulture is focal in bolstering food security and fostering economic prosperity for local farmers. Accurate forecasting of production levels equips policymakers, farmers, and stakeholders with valuable insights for resource allocation, market strategizing, and investment decisions [5]. For instance, anticipating a surge in horticultural output during a specific season empowers farmers to plan their planting schedules meticulously and bolster investments in irrigation systems and fertilizers [6]. This proactive approach enhances crop yields and profits and amplifies agricultural output, thereby invigorating the local economy. Conversely, in the event of anticipated low production levels, policymakers can pivot resources towards importing essential crops or devising alternative income streams for affected farmers. Such responsive measures mitigate economic setbacks triggered by seasonal fluctuations. Moreover, adept planning enables farmers to reduce the risk of crop failure and avoid significant investment losses [7]. Thus, the significance of agricultural production forecasting lies in its capacity to efficiently manage resources, ensuring the sustained profitability and resilience of both the farm sector and the local economy.

Artificial intelligence, particularly machine learning, presents a compelling avenue for forecasting agricultural and horticultural production [8]. Machine learning, a cornerstone of artificial intelligence, empowers computers to obtain insights from data and past experiences, enabling predictive analytics and decision-making without explicit programming. When predicting agricultural or horticultural yields, machine learning algorithms leverage historical data encompassing weather patterns, soil composition, and other factors influencing crop growth. Machine learning systems craft predictive models by discerning underlying patterns within

this data corpus, furnishing farmers and horticulturists with invaluable tools for informed planning and decision-making across crop production cycles.

The main objective of this research is to predict horticultural production, especially mango fruit in Indramayu Regency, Indonesia, by applying various machine learning models. This prediction model will utilize historical data published by the Indonesian government publicly through BPS (Statistics Indonesia) Indramayu data. Thus, we aim to answer the research question of what machine learning model is most suitable for predicting horticultural production, as well as how accurate each model is in predicting horticultural production. This research is expected to contribute to the optimization of horticultural production in Indramayu Regency, Indonesia.

2. RELATED WORKS

There are several previous studies that have applied machine learning in solving the problem of predicting agricultural production [9-11]. Rathod et al. [12] compared the ARIMA algorithm against Neural Network Autoregressive (NAR) and non-linear Support Vector Regression (NLSVR) in predicting mango fruit production in Karnataka, India. Here, the research revealed the machine learning approach outstripped the traditional ARIMA model, reaffirming the potential of machine learning techniques in agricultural yield prediction.

Jhajharia and Mathur [13] predicted crop yields in the state of Rajasthan, India. Prediction is done using a machine learning approach. The data used are yield data, area data, production data and rainfall data. The data range used was from 1997-2018. Machine learning regression models used are Decision Tree, Random Forest, Gradient Boosting Regression. In this study, it was found that the gradient boosting regression model provided the most superior performance compared to other machine learning regression models.

Jorvekar et al. [14] compared the evaluation of machine learning regression model performance metrics for crop yield prediction in agriculture. The data used in this study were taken from several data sources from 101 different countries. The data used are yield dataset, rainfall dataset, and pesticides dataset with data time span from 1961-2016. There are 8 regression models compared, namely Linear Regression, K-Nearest Neighbor, Support Vector Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, Lasso Regression, and Elasticnet Regression. In this study, the experimental results show that Random Forest Regression has the highest R^2 value of 0.973.

Rai et al. [15] also conducted a comparative analysis on various machine learning regression models to predict agricultural yields. The purpose of this research is to determine the most accurate regression model to predict crop yields. The data used in this study came from the Indian agriculture government portal. The data used includes covered area, yield, seasons, and years from 1997-2013. The regression models compared in this study are Decision Tree Regression, Linear Regression, Lasso Regression and Random Forest Regression. The results showed that the Random Forest Regression model outperformed the performance of other regression models.

Tiwari et al. [16] proposed the use of historical rainfall data to predict agricultural yields using a machine learning regression model approach. The data used includes several years of data (which does not mention the details of the year range), crop yields of two types of crops namely wheat and potatoes, as well as rainfall data. Machine learning regression models used are Random Forest Regression, Support Vector Regression, K-Nearest Neighbor, and Gradient Boosting Regression. Experiments in this research show that Random Forest Regression produces the highest accuracy. In addition, this study also mentioned that feature selection and hyperparameter tunning can improve the accuracy of machine learning regression models.

Ashwitha and Spoorthi [17] also reported that regression models in machine learning can improve the accuracy of crop yield predictions for agricultural progress. This study compared four regression models namely Random Forest Regression, Gradient Boosting Regression, Decision Tree Regression, and Linear Regression. This research also provides similar findings to other studies, that the Random Forest Regression model produces the most accurate prediction performance.

Meanwhile, Putra and Walmi [18] delved into agricultural prediction, employing the artificial neural network (ANN) algorithm to forecast rice production in West Sumatra, Indonesia. Their study encompassed tests across 19 regions within West Sumatra, yielding an impressive accuracy rate of 88.14% and a relatively minimal error rate of 11.86%. Similarly, other studies within Indonesia have explored machine learning applications in horticultural data analysis. Kaunang et al. [19] utilized a decision tree algorithm to predict food crop outcomes, Masdian et al. [20] employed the random forest algorithm to scrutinize rice productivity in Batang Regency, Indonesia, while Fareza et al. [21] leveraged the extreme learning machine method to forecast the yield of biopharmaceutical plants.

In the above studies, it can be concluded that the machine learning approach can be used to solve the problem of predicting agricultural production, including horticulture. The majority of machine learning models used are regression models such as Random Forest Regression, Decision Tree, Gradient Boosting Regression, Linear Regression and other regression models. Therefore, in this research we will focus on horticultural data, especially mango fruit in Indramayu Regency, as one of the regions that has potential resources in the horticultural sector.

These regression models were chosen based on their unique capabilities. Random Forest Regression was chosen because of its ability to handle data with many variables and its resistance to overfitting although it has limitations on large computations [22, 23]. In addition, the Random Forest Regression model is also widely mentioned in previous studies to produce more accurate prediction performance [14-17]. Gradient Boosting Regression was chosen because it has the advantage of gradually increasing prediction accuracy by correcting errors from the previous model but has a longer computation time [13, 24, 25]. Decision Tree was chosen due to its simplicity and ability to interpret results, which is important for intuitive understanding in an agricultural context, but is prone to overfitting [26, 27]. Meanwhile, Linear Regression was chosen as the base model to compare performance with other more complex models.

Therefore, in this research we investigate several machine learning regression models, namely Random Forest Regression, Decision Tree, Gradient Boosting Regression, Linear Regression, to predict the productivity of horticulture, especially mangoes in Indramayu Regency, Indonesia.

3. METHODOLOGY

3.1 Data collection

The data collection process in this study was carried out through secondary data collection. The secondary data used as data for this study were obtained from the data publication "Indramayu Regency in Figures" which is released annually by the Indramayu Regency's Indonesia Statistics (BPS) [28]. The data collected consists of the name of the sub-district, year, total mango production, mango harvest area, rainfall and number of rainy days.

The time span of the data is from 2009 to 2017, covering 31 sub-districts in Indramayu Regency, Indonesia. All data that has been collected is tabulated so that 279 rows of data are obtained. Data on the amount of mango production is the data used as the target variable to predict the amount of mango production. Meanwhile, other data will be used as feature variables to predict the amount of mango production. The pieces of data used can be seen in Figure 1.

The limited amount of data available each year has caused the data in this study to only cover the years 2009 - 2017. This is because since 2018, BPS no longer publishes data related to mango harvest areas.

To reduce the impact of this limitation, a thorough data quality check was carried out at the data preprocessing stage, so that the data used is representative and accurate in the context of this research.

The stages carried out during data preprocessing are identifying and deleting empty or missing value data and checking the distribution of each data using QQ-plot. QQ-plot is used to determine whether there are outliers in the data that will affect the performance of the model later in making predictions.

	District	Year	Mango Production	Harvested Area	Rainfall	Rainy Days
0	HAURGEULIS	2009	25144.580	74352	1124	77
1	GANTAR	2009	18680.000	36990	1046	72
2	KROYA	2009	12052.000	76200	1581	68
3	GABUSWETAN	2009	2756.200	32439	1455	72
4	CIKEDUNG	2009	5151.600	41213	873	59
...	...	...	...	...	...	...
274	KANDANGHAUR	2017	985.842	4124	1344	84
275	BONGAS	2017	2156.961	15670	1582	87
276	ANJATAN	2017	7241.936	30520	1171	107
277	SUKRA	2017	695.476	10000	1244	72
278	PATROL	2017	76.000	1000	6792	85

279 rows × 6 columns

Figure 1. Excerpts of data on mango production, harvested area, rainfall, and rainy days in Indramayu Regency, Indonesia

3.2 Data preprocessing

Once the data collection phase concludes, the next step involves data preprocessing, an important process to refine the dataset by eliminating noise to facilitate more precise predictions. Initially, the focus is on identifying and removing empty or missing values. Subsequently, a thorough examination of the distribution of each dataset ensues, typically conducted through techniques like QQ-plot analysis. This stage serves to uncover potential data outliers that could compromise the model's predictive accuracy. The results of this preliminary data distribution assessment are shown in Figure 2.

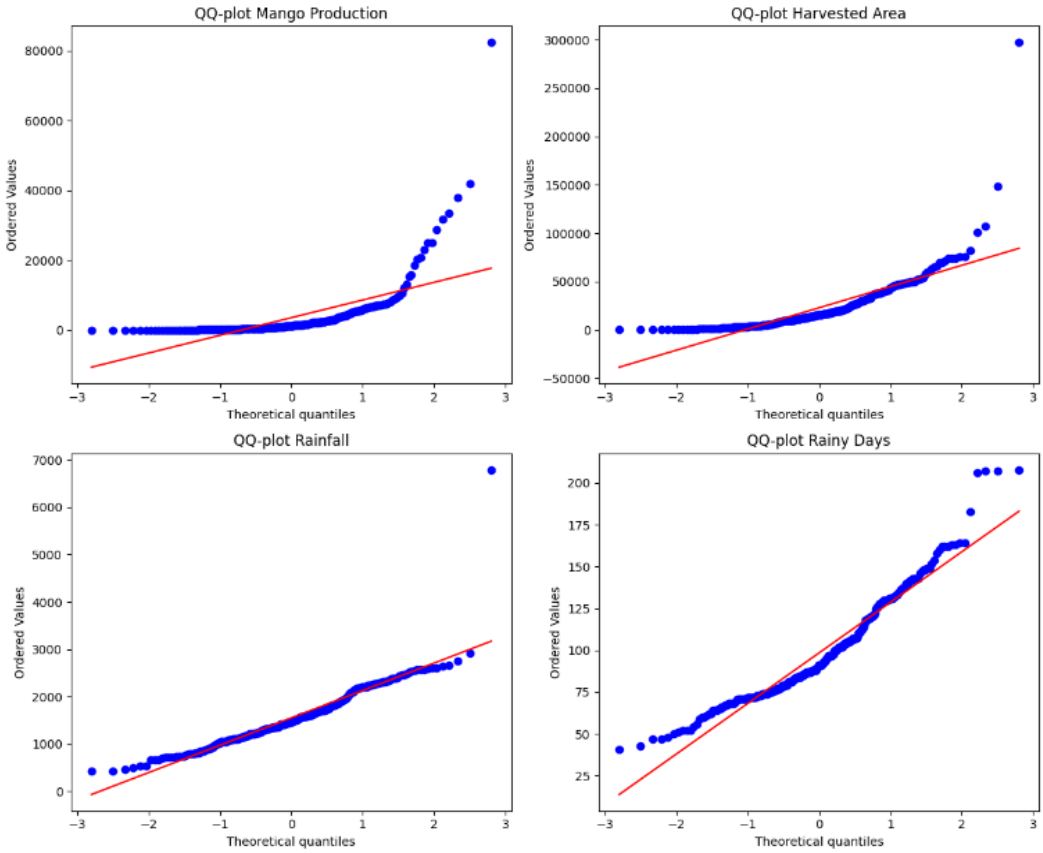

Figure 2. Initial data distribution

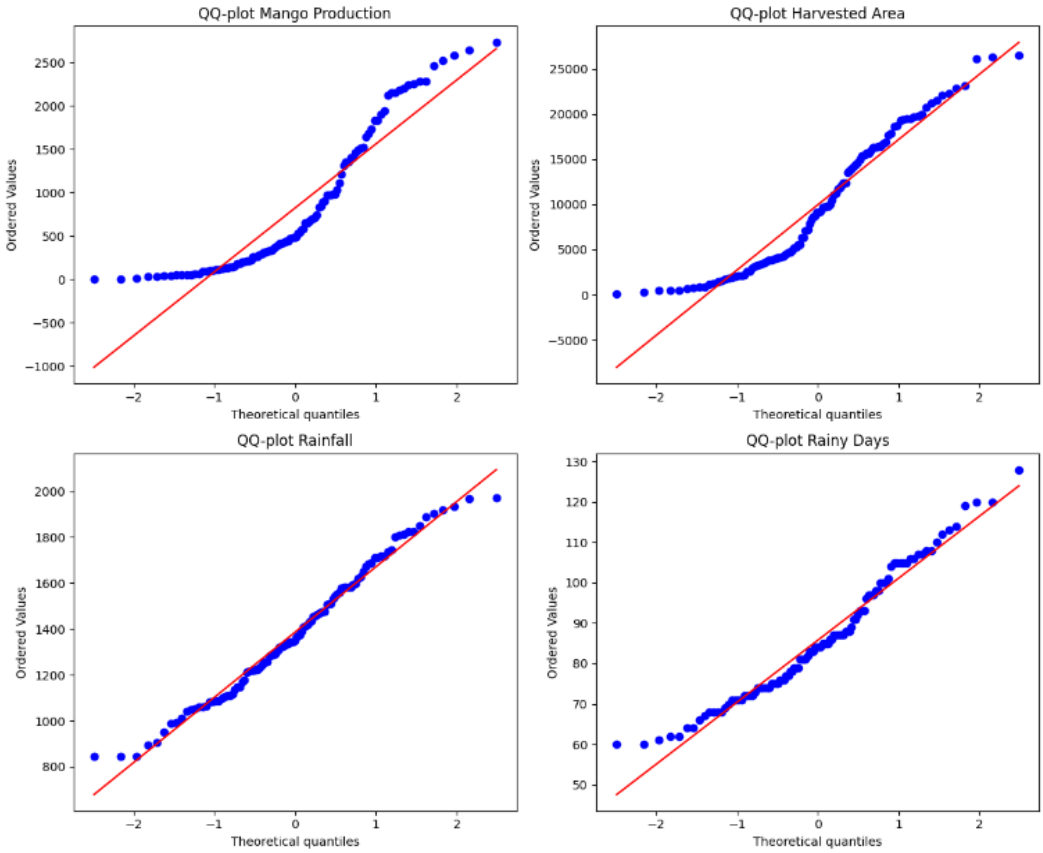

Figure 3. Data distribution after outlier removal

In Figure 2, it can be seen that the initial data distribution still contains outlier data. Therefore, the next step is to remove outliers using the Interquartile Range method. The results of the data distribution after removing outliers can be seen in Figure 3.

The next step is to transform the district's data using one-hot encoding, which will convert each district into a numeric representation. In addition, it also deletes year data that will not be used in forming a prediction model for the amount of mango production in Indramayu Regency, Indonesia. The last stage is to normalize the data using MinMaxScaler.

3.3 Model training dan testing

After the data undergoes preprocessing, the next step involves data partitioning. This entails dividing the dataset into two subsets: training and test data. The training data is utilized to construct regression models during the training phase, while the test data is reserved for assessing these models.

The training data undergoes training using various regression techniques, including linear regression, random forest regression, gradient boosting regression, and decision tree regression. Conversely, the test data, which remains unseen throughout the training phase, serves as a litmus test for evaluating the predictive performance of the trained regression models on fresh data. Notably, this training and testing process uses the Python programming language within the Google Colab environment.

3.3.1 Linear regression

The linear regression method is employed to predict a quantitative response, denoted as Y, based on a solitary predictor variable, X. This approach hinges on the assumption that X and Y exhibit a linear relationship [29], a mathematical expression formulated in Eq. (1).

$$Y \approx \beta_0 + \beta_1 X \tag{1}$$

Eq. (1) signifies a regression of Y on X, where X embodies factors about harvest productivity, such as mango harvest area, rainfall, number of rainy days, and districts in Indramayu. Meanwhile, Y represents the quantity of mango production. Utilizing a linear regression model, we can effectively regress the quantity of mango production against these harvest productivity factors. Within Eq. (1), β_0 and β_1 denote two unknown constants, signifying the intercept and slope within the linear model. β_0 and β_1 called coefficients or model parameters, are important in determining the relationship between X and Y.

3.3.2 Random forest regression

Random forest is an ensemble classifier that collects different decision trees [30]. Capable in both regression and classification tasks, this method combines multiple tree predictors. Each decision tree within the ensemble operates on random vectors as parameters, selecting features from samples and subsets of the dataset for training in a theoretical manner [31]. The strength of individual trees and their interdependence collectively determine the generalization error of the forest. Each node is split through a randomized feature selection process, yielding error rates that rival Adboost, particularly in noisy environments [32]. Known for its adaptability and user-friendly nature, random forest emerges as a favored machine learning algorithm among researchers. Its innate effectiveness often yields exceptional results even without hyper-parameter tuning, rendering it a

staple choice due to its simplicity.

3.3.3 Gradient boosting regression

Gradient Boosting Regression is commonly employed to uncover nonlinear relationships within tabular datasets [33]. When a machine learning model exhibits poor predictability, gradient boosting enhances model quality through interpretability. This iterative process optimizes the model's predictive capacity in each learning iteration, enabling it to handle missing values and outliers for improved generalization. The primary objective of gradient boosting is to enhance the model's predictive performance and optimize the loss function by bolstering weak learners, which measure the disparity between predicted and actual target values. The algorithm initiates by training a decision tree, assessing the weight of each tree, and classifying them based on their complexity. Gradient Boost blends multiple weak models with each iterative step to form stronger ones, thereby minimizing bias errors [34].

3.3.4 Decision tree regression

One notable decision tree algorithm is CART (Classification and Regression Tree), which is renowned for its versatility in handling regression and classification tasks [35]. CART operates as a recursive partitioning method, systematically dividing a subset of the dataset into two child nodes by utilizing all predictor variables. Beginning with the entire dataset, a decision tree is constructed iteratively. The selection of the best predictor is guided by the Gini index, serving as a measure of inequality (or impurity) within the sample. This index aids in establishing decision nodes (D-nodes) and partitioning the dataset into smaller subsets. The tree is then constructed through a recursive process until one of several conditions is met. All tuples possess the same attribute value; no remaining attributes or further instances exist. Eq. (2) delineates the calculation of the Gini index, wherein p_i denotes the probability of a tuple in D belonging to class-i.

$$Gini\ (D) = 1 - \sum_{i=1}^{m} {p_i}^2 \tag{2}$$

3.4 Model evaluation

Model evaluation is conducted to determine the performance of the model in predicting the amount of mango production. In the study, three evaluation values were used, namely MAPE, RMSE and R^2. By using a combination of these evaluation values, we can get a comprehensive picture of the performance of the regression model, rather than using only one evaluation value. So that we can ensure which regression model provides accurate and reliable results. Especially in the context of mango production prediction as an agricultural production. It is important to ensure that the model used can provide accurate and reliable predictions in determining the right agricultural policy.

3.4.1 Mean Absolute Percentage Error (MAPE)

MAPE serves as a valuable metric for assessing the error rate of predictions [36]. It quantifies the error percentage between predicted and actual values, averaging them across all data points. This calculation, as outlined in Eq. (3), involves summing the absolute differences between actual and predicted values, dividing by the true value, and then averaging across all data points, where n represents the total number of data, A_i signifies the true value of the i-ith data point, and F_i denotes the predicted value for the i-th data point.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| * 100\% \tag{3}$$

This method proves highly effective in evaluating the accuracy of predictive models, offering insight into their performance. MAPE's notable advantage lies in its straightforward interpretation, which yields results in percentage. Moreover, MAPE facilitates comparisons between various prediction models, enabling assessment of their accuracy. By quantifying the percentage error between predicted and actual values, MAPE clearly indicates a model's predictive prowess. Thus, MAPE provides easily understandable information and facilitates precise comparisons across diverse predictive models.

3.4.2 Root Mean Squared Error (RMSE)

RMSE is an essential metric in assessing the difference between observed and predicted values by statistical models or algorithms [37]. Computed as the square root of the average of squared differences between observed and predicted values, RMSE is depicted in Eq. (4), where n represents the total number of data points, A_i signifies the actual value of the i-th data point, and F_i denotes the predicted value for the i-th data point.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - F_i)^2} \tag{4}$$

In predictive modelling, RMSE functions as a crucial gauge of prediction quality, with lower RMSE values indicating more accurate predictions and reduced error rates. It also facilitates comparative analysis across various prediction models, allowing for the identification of a more precise model. Furthermore, RMSE is valuable in evaluating discrepancies between predicted and observed values, particularly in continuous data estimation.

3.4.3 Determinant Coefficient (R^2)

R^2 is a metric for calculating the explanatory power of independent variables in a regression model [38]. Calculated value of R^2, as depicted in Eq. (5), involves variables; total data points as n, A_i denoting the actual value of the i-th data, F_i representing the predicted value of the i-th and $\bar{A}$ signifying the average value of the real data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(A_i - F_i)^2}{\sum_{i=1}^{n}(\bar{A} - A_i)^2} \tag{5}$$

R^2 values fall within the range of 0 to 1, with higher values indicating a stronger ability of the regression model to elucidate variations within the data. This metric can be interpreted as the percentage of variation in the dependent variable that can be accounted for by the independent variables within the regression model. In the context of prediction or estimation, R^2 serves as a valuable indicator of the accuracy of model predictions. Higher R^2 values signify closer alignment between model predictions and actual data, while lower values

indicate greater unexplained variation within the dependent variable.

4. RESULT AND DISCUSSION

The optimum value of parameters (hyperparameter tuning) is obtained using the Randomized Search method to get the optimal linear regression model. The predetermined hyperparameter value will be randomly selected by this method. Furthermore, the model of each combination will be evaluated using an evaluation metric. Based on the results, this method will choose the hyperparameter combination with the best evaluation value. Table 1 shows the optimum parameter value for each regression model used in this study.

Table 1. Optimal parameter values for each regression model

Regression Model	Optimum Parameter
Linear Regression (**LReg**)	positive= True, n_jobs=1, fit_intercept=False, copy_X=True
Random Forest Regression (**RFReg**)	n_estimators=300, max_depth=3, random_state=0, min_samples_split= 10, min_samples_leaf=8, bootstrap=True
Gradient Boosting Regression (**GBReg**)	n_estimators=1000, max_depth=10, random_state=0, subsample=0.1, min_samples_split=2, min_samples_leaf= 4, learning_rate =0.01
Decision Tree Regression (**DTReg**)	splitter='random' min_samples_split=10, min_samples_leaf=4, max_features=None, max_depth=7, random_state=0

Table 2. Comparison of preprocessed vs. non-preprocessed data for a 90% training and 10% testing data split

Experiment	Regression Model	R^2	MAPE	RMSE
Without preprocessing	LReg	-0.152	867.507	0.093
	RFReg	-0.007	810.965	0.087
	GBReg	0.003	942.524	0.087
	DTReg	-0.258	337.034	0.097
With preprocessing	LReg	0.151	56.955	0.225
	RFReg	0.467	58.966	0.178
	GBReg	0.509	45.598	0.171
	DTReg	0.347	56.139	0.197

Table 3. Comparison of preprocessed and non-preprocessed data for an 80% training and 20% testing data split

Experiment	Regression Model	R^2	MAPE	RMSE
Without preprocessing	LReg	-0.107	1.25×10^{14}	0.0706
	RFReg	-0.085	5.53×10^{13}	0.0699
	GBReg	-0.003	2.32×10^{13}	0.0672
	DTReg	-0.157	4.83×10^{12}	0.0722
With preprocessing	LReg	0.290	97.310	0.255
	RFReg	0.613	67.462	0.188
	GBReg	0.577	63.848	0.197
	DTReg	0.615	57.389	0.188

Table 4. Comparison of preprocessed and non-preprocessed data for a 70% training and 30% testing data split

Experiment	Regression Model	R^2	MAPE	RMSE
Without preprocessing	LReg	-0.291	8.78×10^{13}	0.079
	RFReg	-0.198	2.95×10^{13}	0.076
	GBReg	-0.054	3.00×10^{13}	0.072
	DTReg	-0.269	1.25×10^{14}	0.079
With preprocessing	LReg	0.126	207.633	0.292
	RFReg	0.485	163.617	0.225
	GBReg	-0.080	222.110	0.325
	DTReg	0.310	139.902	0.260

Tables 2-4 highlight the efficacy of incorporating a preprocessing stage in improving the performance of the regression model, evidenced by the uptick in R^2 values and reduction in MAPE. Notably, in the gradient boosting regression model, the most significant enhancement is observed, with the R^2 value soaring from 0.0029 to 0.5087, while the MAPE plummets from 942.5236 to 45.5978, with a 90% to 10% split between training and test data.

Conversely, the RMSE values obtained from experiments without preprocessing stages are lower than those with preprocessing. This discrepancy suggests that data lacking preprocessing may contain noise or outliers, causing RMSE to weigh on outliers and yield lower values disproportionately.

These findings underscore the important role of the preprocessing stage in machine learning regression models. Specifically, the preprocessing step, focused on outlier removal in this experiment, is instrumental in ensuring the accuracy of regression model predictions. Outliers, by their deviation from the norm, can distort the evaluation of predictive model quality, making their removal imperative for robust analysis.

Tables 2-4 also show that the greater the proportion of test

data, the greater the MAPE and RMSE values. This indicates that the larger the sample of test data used, the higher the prediction error rate of the model [39, 40]. Therefore, using a representative sample of test data to obtain accurate prediction results is important.

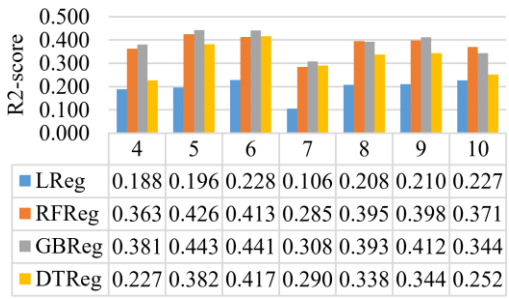

Figure 4. Comparison of R^2 values for various k values on k-fold cross validation

In the next experiment, data division is executed using the k-fold cross-validation technique to assess the model's overall performance across the entire dataset. This step determines the proportion of data allocated for training and testing in the ensuing experiment. Figure 4 presents a comparative analysis of R2 values for various k values in k-fold cross-validation. Notably, for k = 5, the gradient-boosting regression model yields the highest R2 value of 0.4426. With k = 5 indicating a test data distribution of 1/5 or 20% of the entire dataset, the subsequent experiment will adopt a training data to test data ratio of 80%:20%.

Table 5. Correlation value of features with target class

	Harvested Area	Precipitation	Rainy Days
Correlation	0.7	0.1	-0.083

In the subsequent investigation, the efficacy of different feature sets in predicting mango production quantities is scrutinized. As previously discussed, three features—mango harvest area, rain precipitation, and number of rainy days—were utilized in predicting mango production amounts. Table 5 elucidates the correlation values of each feature with the target variable, mango production quantity. Notably, the harvest area feature exhibits a robust correlation with mango production quantity, while precipitation and rainy-day features demonstrate weaker correlations.

Further analysis of this finding that the strong correlation between harvest area and total mango production confirms that the importance of harvest area management in increasing mango production, because the larger the harvest area, the greater the mango production produced [41]. This reinforces the fact that the harvest area will contribute significantly to the machine learning regression model. Meanwhile, the weak correlation between rainfall and rainy days variables with mango production illustrates that these variables have an indirect influence on mango production. For example, during the dry season, these variables play a role in mango production. The machine learning regression model can identify important features, so this rainfall and rainy day variables still need to be taken into account, even though their influence is not as great as the harvest area. Therefore, the next experiment will compare the performance of the regression model with various combinations of features.

Table 6. Comparison of R^2 values for each regression model on various feature combinations

Experiment	LReg	RFReg	GBReg	DTReg	Mean
3 features	0.290	0.613	0.577	0.615	0.524
2 features (a+b)	0.305	**0.643**	0.590	0.520	0.514
2 features (a+c)	0.278	0.591	0.586	0.520	0.494
1 feature (a)	0.338	0.624	0.600	0.590	**0.538**
Mean	0.303	**0.618**	0.588	0.561	

a = harvested area, b = rainfall, c = rainy days

Table 6 comprehensively compares R^2 values across different regression models and feature combinations. Notably, the random forest regression model achieves the highest R^2 value of 0.6430 when utilizing two features—harvest area and rainfall. However, upon closer examination of average values across experiments, employing only one feature, specifically harvest area, yields more robust R^2 results than other configurations. Furthermore, the table underscores the consistent performance of the random forest regression model across all experiments, suggesting that it remains stable regardless of the number of features utilized.

Table 7. Comparison of RMSE values for each regression model on various feature combinations

Experiment	LReg	RFReg	GBReg	DTReg	Mean
3 features	0.255	0.188	0.197	0.188	0.207
2 features (a+b)	0.252	**0.181**	0.194	0.209	0.209
2 features (a+c)	0.257	0.193	0.195	0.209	0.214
1 feature (a)	0.246	0.185	0.191	0.194	**0.204**
Mean	0.252	**0.187**	0.194	0.200	

a = harvested area, b = rainfall, c = rainy days

Meanwhile, if analyzed from the RMSE values shown in Table 7, the lowest RMSE value is 0.181 in the random forest regression model using two features, namely harvest area and rainfall. However, when considered for all regression models, experiments using one feature, harvest area, have the smallest average RMSE value of 0.204. This is in line with the correlation value of the harvest area feature, which is higher than the other features, indicating a strong relationship between the harvest area and the amount of mango production. Meanwhile, random forest regression provides a minor error value compared to all regression models, indicating that this model is the most accurate in predicting the number of mangoes. A low RMSE value indicates a closer fit between the predicted and actual values.

5. CONCLUSIONS

Effective machine learning techniques are employed in a regression context to forecast horticultural production in Indramayu Regency, Indonesia. Utilizing secondary data from the Central Bureau of Statistics of Indramayu Regency spanning 2009 to 2017 and encompassing 31 sub-districts, the dataset comprises key variables including mango production quantity, mango harvest area, rainfall, and number of rainy days. Notably, the quantity of mango production is the prediction's target variable. The study attempts to predict mango production quantities by employing various machine learning regression models such as Linear Regression, Random Forest Regression, Gradient Boosting Regression, and Decision Tree Regression.

The comparison of regression model performance hinges on three key evaluation metrics: MAPE, RMSE, and R^2. MAPE gauges the accuracy of regression model predictions, with smaller values indicating higher precision. Conversely, RMSE reflects the model's error rate, where smaller values signify superior accuracy, implying a closer alignment between predicted and actual values. Meanwhile, R^2 measures the regression model's ability to elucidate variation in predicted data, with higher values indicating a stronger relationship between independent and dependent variables, ranging from 0 to 1, where 1 denotes a perfect explanation of data variation.

This study comprises three experiments aimed at enhancing regression model performance. The initial experiment contrasts model performance with and without data preprocessing, highlighting the importance of cleaning data from noise or outliers. Preprocessing notably improves model performance, as evidenced by increased R^2 values and decreased MAPE. However, it also leads to higher RMSE values, indicative of noise or outliers in unprocessed data.

The second experiment determines the optimal proportion of training and test data. Results indicate that a larger proportion of test data correlates with higher MAPE and RMSE values, implying increased prediction error rates. Thus, employing a representative sample of test data for accurate predictions is crucial, with the 80%:20% training-to-test data ratio yielding superior regression model performance.

The third experiment compares different feature combinations for predicting mango production quantity. Harvest area emerges as the feature with the strongest correlation to mango production quantity, yielding higher R^2 values and lower RMSE values. Additionally, the random forest regression model exhibits superior robustness to other models, boasting the highest R^2 value and lowest RMSE value.

The implications of the findings in this study indicate that machine learning models, such as Random Forest Regression and Gradient Boosting Regression, can produce more accurate predictions of horticultural production in Indramayu Regency compared to machine learning models of Decision Tree Regression and Linear Regression. By utilizing historical data, these models allow farmers and policy makers to make more informed decisions regarding resource use and agricultural management. In other regions with similar characteristics, these models can be applied to improve agricultural efficiency through more accurate yield forecasting, optimization of resource use, and better risk management, thus supporting food security and increased agricultural productivity.

The results of this study have several limitations, such as the limited time period of data collection (2009-2017). In addition, there are also limitations related to the predictor variables available from BPS data publications. Because the accuracy of the regression model is also influenced by the quality of the predictor variables or features used in the formation of the regression model.

Based on the research limitations that have been presented, there are three potentials for further research development. The first potential is the addition of predictor variables or features such as soil quality, water availability, weather, air temperature, and other environmental factors that can improve the accuracy of the regression model. The second potential is to expand the time period of data collection and use other data sources besides BPS data that can also enrich the analysis and research results. The third potential is to conduct a comparative study on other regression models and also use an ensemble learning approach to improve the performance of the regression model. Thus, further research development in this case can bring significant benefits in improving the accuracy and validity of the regression model used.

ACKNOWLEDGMENT

This research was funded by the Universitas Pembangunan Nasional Veteran Jakarta under the RIKIN 2022 (Riset Kerjasama Internasional - International Cooperation Research) (Grant No.: 292/UN.61.0/HK.07/LIT.RIKIN/2022).

REFERENCES

[1] Euriga, E., Boehme, M.H., Amanah, S. (2021). Changing farmers' perception towards sustainable horticulture: A case study of extension education in farming community in Yogyakarta, Indonesia. AGRARIS: Journal of Agribusiness and Rural Development Research, 7(2): 225-240. https://doi.org/10.18196/agraris.v7i2.11510

[2] Sahara, Daryanto, A., Nugrahapsari, R.A., Perkasa, H., Reardon, T., Stringer, R. (2019). Improving market integration for high value fruit and vegetable production systems in Indonesia. In Australian Centre for International Agricultural Research (ACIAR), Canberra, Australia.

[3] Orsini, F., Kahane, R., Nono-Womdim, R., Gianquinto, G. (2013). Urban agriculture in the developing world: A review. Agronomy for Sustainable Development, 33: 695-720. https://doi.org/10.1007/s13593-013-0143-z

[4] Rasmikayati, E., Elfadina, E.A., Kusumo, R.A.B., Saefudin, B.R., Supriyadi, S. (2020). Policy analysis of mango's agribusiness development (A case in Cikedung district, Indramayu Regency). Jurnal Manajemen & Agribisnis, 17(1): 52-52. https://doi.org/10.17358/jma.17.1.52

[5] Sharma, R., Kamble, S.S., Gunasekaran, A., Kumar, V., Kumar, A. (2020). A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. Computers & Operations Research, 119: 104926. https://doi.org/10.1016/j.cor.2020.104926

[6] Bantacut, T. (2014). Indonesian staple food adaptations for sustainability in continuously changing climates. Journal of Environment and Earth Science, 4(21): 202-216.

[7] Benyam, A.A., Soma, T., Fraser, E. (2021). Digital agricultural technologies for food loss and waste prevention and reduction: Global trends, adoption opportunities and barriers. Journal of Cleaner Production, 323: 129099. https://doi.org/10.1016/j.jclepro.2021.129099

[8] Shaikh, T.A., Rasool, T., Lone, F.R. (2022). Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. Computers and Electronics in Agriculture, 198: 107119. https://doi.org/10.1016/j.compag.2022.107119

[9] Kumar, P., Agarwal, S., Gupta, S.K., Kaur, G. (2023). Agricultural crop yield prediction using regression models with prominent feature. In 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India, pp. 384-389.

https://doi.org/10.1109/ICIMIA60377.2023.10426396
[10] Garanayak, M., Sahu, G., Mohanty, S.N., Jagadev, A.K. (2021). Agricultural recommendation system for crops using different machine learning regression methods. International Journal of Agricultural and Environmental Information Systems, 12(1): 1-20. https://doi.org/10.4018/IJAEIS.20210101.oa1
[11] Raja, K. (2023). Comparison of machine learning algorithms for the prediction of rice crop yield in Karnataka. In 2023 Fourth International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India, pp. 1-8. https://doi.org/10.1109/ICSTCEE60504.2023.10585142
[12] Rathod, S., Vijayakumar, S., Bandumula, N., Chitikela, G. (2022). Prediction of mango production using machine intelligence techniques: A case study from Karnataka, India. Acta Scientific Agriculture, 6(9): 16-22. https://doi.org/10.31080/ASAG.2022.06.1174
[13] Jhajharia, K., Mathur, P. (2024). Machine learning based crop yield prediction model in Rajasthan Region of India. Iraqi Journal of Science, 65(1): 390-400. https://doi.org/10.24996/ijs.2024.65.1.32
[14] Jorvekar, P.P., Wagh, S.K., Prasad, J.R. (2024). Predictive modeling of crop yields: A comparative analysis of regression techniques for agricultural yield prediction. Agricultural Engineering International: CIGR Journal, 26(2): 125-140.
[15] Rai, S., Nandre, J., Kanawade, B.R. (2022). A comparative analysis of crop yield prediction using regression. In 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, pp. 1-4. https://doi.org/10.1109/CONIT55038.2022.9847783
[16] Tiwari, P., Raj, R., Das, H., Gourisaria, M.K. (2023). A Comparative analysis of regression models for crop yield prediction based on rainfall data: Experimental study and future perspective. In 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), Bengaluru, India, pp. 1-6. https://doi.org/10.1109/NMITCON58196.2023.10275902
[17] Ashwitha, K., Spoorthi, B. (2024). Advancing agricultural sustainability: Enhancing crop yield prediction through regression modeling. In 2024 Second International Conference on Data Science and Information System (ICDSIS), Hassan, India, pp. 1-6. https://doi.org/10.1109/ICDSIS61070.2024.10594257
[18] Putra, H., Walmi, N.U. (2020). Penerapan prediksi produksi padi menggunakan artificial neural network algoritma backpropagation. Jurnal Nasional Teknologi dan Sistem Informasi, 6(2): 100-107. https://doi.org/10.25077TEKNOSI.v6i2.2020.100-107
[19] Kaunang, F.J., Rotikan, R., Tulung, G.S. (2018). Pemodelan sistem prediksi tanaman pangan menggunakan algoritma decision tree. Cogito Smart Journal, 4(1): 213-218. https://doi.org/10.31154/cogito.v4i1.115.213-218
[20] Masdian, A.R., Bashit, N., Hadi, F. (2023). Analisis produktivitas padi menggunakan algoritma machine learning random forest di kabupaten batang tahun 2018-2022. Elipsoida: Jurnal Geodesi dan Geomatika, 6(1): 43-51. https://doi.org/10.14710/elipsoida.2023.19023
[21] Fareza, Z.A.N.A., Cholissodin, I., Muflikhah, L. (2022). Prediksi hasil panen tanaman biofarmaka di Indonesia dengan menggunakan metode extreme learning machine. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 6(11): 5331-5338.
[22] Danieli, M.G., Tonacci, A., Paladini, A., Longhi, E., Moroncini, G., Allegra, A., Sansone, F., Gangemi, S. (2022). A machine learning analysis to predict the response to intravenous and subcutaneous immunoglobulin in inflammatory myopathies. A proposal for a future multi-omics approach in autoimmune diseases. Autoimmunity Reviews, 21(6): 103105. https://doi.org/10.1016/j.autrev.2022.103105
[23] Rezaei, N., Jabbari, P. (2022). Chapter 14 - Practice examples. In Immunoinformatics of Cancers, N. Rezaei and P. Jabbari, pp. 223-261. https://doi.org/10.1016/B978-0-12-822400-7.00002-6
[24] Mesut, B., Başkor, A., Aksu, N.B. (2023). Role of artificial intelligence in quality profiling and optimization of drug products. In Handbook of Artificial Intelligence in Drug Delivery, pp. 35-54. https://doi.org/10.1016/B978-0-323-89925-3.00003-4
[25] Miller, A., Panneerselvam, J., Liu, L. (2022). A review of regression and classification techniques for analysis of common and rare variants and gene-environmental factors. Neurocomputing, 489: 466-485. https://doi.org/10.1016/j.neucom.2021.08.150
[26] Shobha, G., Rangaswamy, S. (2018). Chapter 8 - Machine learning. In Handbook of Statistics, pp. 197-228. https://doi.org/10.1016/bs.host.2018.07.004
[27] Apsemidis, A., Psarakis, S., Moguerza, J.M. (2020). A review of machine learning kernel methods in statistical process monitoring. Computers & Industrial Engineering, 142: 106376. https://doi.org/10.1016/j.cie.2020.106376
[28] Statistik, B.P. (2017). Kabupaten Indramayu Dalam Angka 2017. BPS Kabupaten Indramayu. Indramayu.
[29] James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J. (2023). Linear regression. In An Introduction to Statistical Learning, pp. 69-134. https://doi.org/10.1007/978-3-031-38747-0_3
[30] Provost, F., Hibert, C., Malet, J.P. (2017). Automatic classification of endogenous landslide seismicity using the Random Forest supervised classifier. Geophysical Research Letters, 44(1): 113-120. https://doi.org/10.1002/2016GL070709
[31] Bradter, U., Kunin, W. E., Altringham, J.D., Thom, T.J., Benton, T.G. (2013). Identifying appropriate spatial scales of predictors in species distribution models with the random forest algorithm. Methods in Ecology and Evolution, 4(2): 167-174. https://doi.org/10.1111/j.2041-210x.2012.00253.x
[32] Shakoor, M.T., Rahman, K., Rayta, S.N., Chakrabarty, A. (2017). Agricultural production output prediction using supervised machine learning techniques. In 2017 1st International Conference on Next Generation Computing Applications (NextComp), Mauritius, pp. 182-187. https://doi.org/10.1109/NEXTCOMP.2017.8016196
[33] Zhang, Z., Zhao, Y., Canes, A., Steinberg, D., Lyashevska, O. (2019). Predictive analytics with gradient boosting in clinical medicine. Annals of Translational Medicine, 7(7): 152. https://doi.org/10.21037/atm.2019.03.29
[34] Natekin, A., Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7: 21. https://doi.org/10.3389/fnbot.2013.00021

[35] Breiman, L., Friedman, J., Olshen, R., Stone, C.J. (2017). Classification and Regression Trees. New York: CRC Press. https://doi.org/10.1201/9781315139470
[36] Kim, S., Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. International Journal of Forecasting, 32(3): 669-679. https://doi.org/10.1016/j.ijforecast.2015.12.003
[37] Singla, P., Duhan, M., Saroha, S. (2022). Different normalization techniques as data preprocessing for one step ahead forecasting of solar global horizontal irradiance. In Artificial Intelligence for Renewable Energy Systems, pp. 209-230. https://doi.org/10.1016/B978-0-323-90396-7.00004-3
[38] Chicco, D., Warrens, M.J., Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Peerj Computer Science, 7: e623. https://doi.org/10.7717/peerj-cs.623
[39] Ihzaniah, L.S., Setiawan, A., Wijaya, R.W.N. (2023). Perbandingan kinerja metode regresi k-nearest neighbor dan metode regresi linear berganda pada data Boston housing. Jambura Journal of Probability and Statistics, 4(1): 17-29. https://doi.org/10.34312/jjps.v4i1.18948
[40] Nurani, A.T., Setiawan, A., Susanto, B. (2023). Perbandingan kinerja regresi decision tree dan regresi linear berganda untuk prediksi bmi pada dataset asthma. Jurnal Sains dan Edukasi Sains, 6(1): 34-43. https://doi.org/10.24246/juses.v6i1p34-43
[41] Rezaei, E.E., Ghazaryan, G., Moradi, R., Dubovyk, O., Siebert, S. (2021). Crop harvested area, not yield, drives variability in crop production in Iran. Environmental Research Letters, 16(6): 064058. https://doi.org/10.1088/1748-9326/abfe29