Prediction of Rubber Crop Production in Malaysia Using Machine Learning

M. N. Shah Zainudin[1]*, P. H. Jing[2], N. A. Sulaiman[2], M. R. Kamarudin[2], N. Z. Nizam[3], Sufry Muhammad[4]

[1]Faculty of Artificial Intelligence and Cyber Security (FAIX), Universiti Teknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

[2]Faculty of Electronics and Computer Technology and Engineering (FTKEK), Universiti Teknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

[3]Faculty of Technology Management and Technopreneurship (FPTT), Universiti Teknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

[4]Faculty of Science and Information Technology (FSIT), Universiti Putra Malaysia (UPM), 43400 UPM Serdang, Selangor, Malaysia

***Corresponding Author**

DOI: https://dx.doi.org/10.47772/IJRISS.2025.909000305

Received: 27 August 2025; Accepted: 04 September 2025; Published: 09 October 2025

ABSTRACT

Agriculture is a cornerstone of economic growth in many countries, especially in developing nations where it provides a significant source of employment and raw materials for industries like sugar, palm oil, and rubber. In Malaysia, agriculture is particularly vital, with rubber products being a key export. To mitigate the economic impact of potential crop shortages, a reliable prediction system for rubber production is crucial. Traditional, human-based methods for predicting crop yields are often inefficient, time-consuming, and prone to errors. For example, the crop-cut method can lead to inaccurate measurements and overestimation of production. To address these limitations, a new study introduces a machine learning-based prediction system. This system, which was developed to forecast rubber crop production in the Malaysian states of Melaka, Perak, Pahang, and Johor, utilizes several machine learning algorithms, including Random Forest, Decision Tree, Linear Regression, and Neural Network. The performance of these models was evaluated using Mean Square Error (MSE) and Mean Absolute Error (MAE). The findings of the study demonstrate that Linear Regression was the most effective algorithm, consistently providing the most accurate predictions with the lowest MAE values. This new system offers a more efficient and precise alternative to traditional methods, enabling better financial planning and decision-making for the agricultural sector.

Keywords: agriculture, machine learning, prediction, rubber crop

INTRODUCTION

Agriculture plays a major role in the economic growth of many countries because it provides job opportunities to the vast majority of the population from developing countries. The other point of view, it will reduce the rate of unemployment and raise the national income level as well as people’s living standards. Some countries earn foreign exchange by exporting and importing their raw materials based on their natural resources. Malaysia's economy grew by 5.1% in 2024, a notable increase from the 3.5% growth in 2023. This brought the total economic value to RM1.65 trillion, up from RM1.57 trillion. The Services sector remained the primary driver of the economy, contributing 59.4% and expanding by 5.3%. The Manufacturing and Agriculture sectors also saw significant rebounds, growing by 4.2% and 3.1% respectively, a substantial improvement from their modest growth of 0.7% and 0.2% the previous year. Additionally, the Mining & Quarrying sector saw a slight

increase from 0.5% to 0.9% growth, while the Construction sector continued its strong performance, posting an impressive 17.5% growth rate (Department of Statistics, 2024). Malaysia is one of the top 10 rubber exporters and importers due to Malaysia is the 3rd largest rubber producer in the world during 2017 (Khin et al., 2019). The purpose of this study is to develop a system using machine learning which can predict rubber crop yield. In the current meantime, human-based prediction methods include various techniques which are crop cuts, farmer's surveys, and expert assessments. However, these techniques are considered labour and time-consuming. Inappropriate measurement in the crop cuts method will cause overestimation. In some cases, insufficient recall data by farmer's survey will affect the accuracy of prediction results. A large range of crops is challenging for an expert to predict the crop yield and it is highly relying on the level of expertise. Some of the work has been reported the accuracy of prediction achieved using machine learning is 71.88% and recorded higher than the accuracy of human-based prediction which is 65.5% (Charoen-Ung and Mittrapiyanuruk, 2019). Hence, machine learning is applied as a prediction mechanism while the crops data analysed as the input data to predict the outcome (Fulkerson et al., 1995), and the accuracy performance will be compared. The machine learning algorithms chosen are Random Forest, Decision Tree, Linear Regression and Neural Network. Four parameters are chosen as an attribute for the prediction of rubber crop yield are temperature, rainfall, humidity, and planted area. Then, the performance of each algorithm is evaluated and compared based on Mean Squared Error (MSE) and Mean Absolute Error (MAE).

In the article Smart Farming System: Crop Yield Prediction Using Regression Techniques (Shah et al., 2018), the authors developed a prediction system using Support Vector Machine, Random Forest Algorithm, and Multivariate Polynomial Regression. The parameters chosen are temperature, precipitation, and rainfall data. The performance of prediction models is evaluated using Root Mean Squared Error (RMSE), MAE, Median Absolute Error, and R-squared values. According to the article Predicting Early Crop Production by Analysing Prior Environment Factors (Osman et al., 2017), the author has carried out a system to forecast crop production. Temperature, rainfall, humidity, sunshine, cloud coverage, and wind speed are selected as parameters. Linear Regression and Neural Network are applied and evaluated using RMSE. A decision tree algorithm is one of the methods in data mining to extract useful information from large datasets (Song and Lu, 2015). The decision tree model is formed from the relevant input variables. After the relevant variables are identified, some variables are prioritized according to the reduction of model accuracy. The importance of a variable is depending on the impact of a variable. Random forests are constructed from multiple tree predictors and each tree contains predictor variables on its lead nodes and other dependent variables on internal nodes (Breiman, 2001). The decision tree from the random forests generates the decision and later it will be considered as a vote for that predictor (Jaiswal and Samikannu, 2017). Then, the random forests will consider the decision by measuring the majority votes from all of the trees created. Linear regression is one of the predictive models to figure out the relationship between independent and dependent variables. Linear regression's equation is given as below:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \qquad (1)$$

The regression equation is expressed as $y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon$ where y denotes crop yield, x_i represents the input attributes, β_i are the model coefficients, and ε is the error term. The independent variable is represented by y and the values of y are continuous or categorical. The dependent variable is represented by x and it is a continuous value. Linear regression separates the independent variables and dependent variables to determine the relationship between two different variables as similar to correlation (Kavitha et al., 2016). An artificial neural network can be divided into three parts which are input layers, hidden layers, and output layers. The input neurons are also known as the input layer, it is used to feed input patterns into the rest of the network (Silva et al., 2017). The following layer is intermediate layers of units, which are hidden layers. These layers are made up of neurons that are in charge of extracting patterns related to the processor system. The hidden layers are then followed by a final output layer. This layer is made up of neurons as well. It is in charge of generating and displaying the final network outputs.

METHODOLOGY

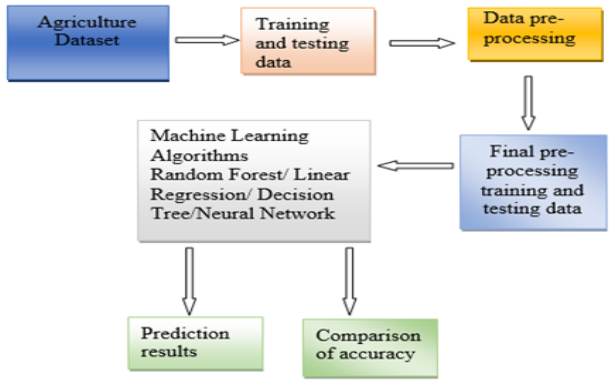

Figure 1: Block Diagram of the System

In this study, an agriculture dataset is collected from the Department of Statistic Malaysia and Open Government Data Malaysia. Then, the dataset is split into training and testing subsets shown in Figure 1. The data are cleaned and normalised in the data pre-processing step. The final pre-processed training data are fit to each machine learning algorithms (Random Forest, Linear Regression, Decision Tree and Neural Network). Testing subset later is used to evaluate the performance of prediction using the trained model. Lastly, the prediction results are analyse and average error of each model is compared. PyCharm IDE using Python is used in this study. All datasets from years 2000 to 2011 are collected from four different states which involving Melaka, Perak, Pahang and Johor. All datasets are sorted according to temperature, rainfall, humidity, planted area of rubber tree and production of rubber in CSV file. 80% of data are split for training data while 20% of data are used for testing. In this section, the split process is done by dividing the data from year 2000 to 2008 for training, while from years 2009 to 2011 are used for the testing. Data standardisation and normalisation are used as the method of feature scaling. The training and testing data are standardised to ensure the variables of dataset lie within a specific range. Hyperparameter optimisation is a method to choose the suitable parameter for each machine learning algorithm. The optimum performance of algorithm is related to the value of parameter chosen. Fine-tuning is a process to figure out which hyperparameters are optimal for each algorithm by enabling the machine learning algorithms to provide the most decisive output as it has variable numerical parameters that are adjusted through iterative optimisation (Naqa amd Murphy, 2015). The trained model is produced once the training data is fit into model. Lastly, the performance of prediction models is evaluated by regression metrics which are MSE and MAE.

RESULTS AND DISCUSSION

The initial prediction results of prediction of rubber production from year 2009 to 2011 are obtained from Random Forest, Decision Tree, Linear Regression and Neural Network for all four states Melaka, Perak, Pahang and Johor. The improvement of prediction results can be done with optimisation by tuning the hyperparameters of each prediction model. The comparison of results of before and after hyperparameter tuning are tabulated in Tables 1 and 2.

Table 1: Comparison between the MSE values

Type of model	Year	MSE value before tuning	MSE value after tuning	Difference before and after tuning
Random Forest	2009	0.4262	0.3946	0.0316
	2010	0.5100	0.4720	0.0380
	2011	0.5800	0.5370	0.0430
Decision	2009	0.2334	0.0794	0.1540

Tree	2010	0.2950	0.1000	0.1950
	2011	0.2650	0.0900	0.1750
Linear Regression	2009	0.0179	0.0089	0.0009
	2010	0.0250	0.0124	0.0126
	2011	0.0150	0.0075	0.0075
Neural Network	2009	0.0666	0.0446	0.0220
	2010	0.0800	0.0536	0.0264
	2011	0.0550	0.0369	0.0181

After hyperparameter optimisation, it is observed that the decreasing of MSE value is obtained in each model. In average, MSE in Decision Tree decreases from 0.2334 to 0.0794 and it has been reduced by 0.1540. MSE values of Random Forest, Linear Regression and Neural Network are recorded the lowest by 0.0316, 0.0009 and 0.0220 respectively in average.

Table 2: Comparison between the MAE values

Type of model	**Year**	**MAE value before tuning**	**MAE value after tuning**	**Difference before and after tuning**
Random Forest	2009	0.4262	0.3946	0.0316
	2010	0.6010	0.5630	0.0380
	2011	0.6710	0.6280	0.0430
Decision Tree	2009	0.2334	0.0794	0.1540
	2010	0.4625	0.2675	0.1950
	2011	0.4325	0.2575	0.1750
Linear Regression	2009	0.0179	0.0089	0.0009
	2010	0.1086	0.0960	0.0126
	2011	0.0986	0.0911	0.0075
Neural Network	2009	0.0666	0.0446	0.0220
	2010	0.2260	0.1996	0.0264
	2011	0.2010	0.1829	0.0181

Referring to Table 2, it is clearly shown that the value of MAE in each model decreases after tuning the parameter. Decision Tree MAE values decrease from 0.4009 to 0.2383, it has been reduced by 0.1626 for all three years. The MAE value for Random Forest, Decision Tree and Neural Network has the smallest by 0.0394, 0.0135 and 0.0555 respectively.

Results Random Forest Algorithm

The prediction results of rubber production and performance are generated and computed according to the state of each dataset. It is clearly can be observed that the differences between actual and predicted results are relatively high as tabulated in Table 3. The lowest MSE and MAE values in Pahang are 0.0618 and 0.2259 respectively. In contrast, the highest value of MSE and MAE values are computed in Johor, which are 0.4004 and 0.6180.

Table 3: Prediction results from Random Forest algorithm

State	Year	Actual Result (tonnes)	Predicted Result (tonnes)	Average MSE (3 years)	Average MAE (3 years)
Melaka	2009	2003	1919.5	0.3946	0.4782
	2010	1387	2020.25		
	2011	1479	3272.6		
Perak	2009	4373	5227.67	0.1222	0.3454
	2010	4184	5435.63		
	2011	4176	5227.67		
Pahang	2009	7837	8539.83	0.0618	0.2259
	2010	7905	8770.8		
	2011	8081	8302		
Johor	2009	7428	8755	0.4004	0.6180
	2010	6951	8755		
	2011	6450	8755		

Results Decision Tree Algorithm

This section focuses on prediction results of rubber production using decision tree. The results are computed according to the state of each dataset. Based on Table 4, it can be seen that the lowest value of MSE and MAE are computed in Melaka, which are 0.0794 and 0.2383 respectively. However, the MSE and MAE values obtained from Johor are the highest, which are 0.4155 and 0.6180.

Table 4: Prediction results from Decision Tree algorithm

State	Year	Actual Result (tonnes)	Predicted Result (tonnes)	Average MSE (3 years)	Average MAE (3 years)
Melaka	2009	2009	2003	0.0794	0.2383
	2010	2010	1387		
	2011	2011	1479		
Perak	2009	2009	4373	0.1123	0.3339
	2010	2010	4184		
	2011	2011	4176		
Pahang	2009	2009	7837	0.1241	0.3127
	2010	2010	7905		
	2011	2011	8081		
Johor	2009	2009	7428	0.4155	0.6180
	2010	2010	6951		
	2011	2011	6450		

Results Linear Regression Algorithm

This section illustrates the prediction results of rubber production are generated according to the state of each dataset using linear regression. According to Table 5, the results for Melaka show that the lowest value of MSE and MAE are 0.0089 and 0.0880. In contrast, the highest value of MSE and MAE are obtained in Johor, which are 0.1672 and 0.3446 respectively.

Table 5: Prediction results from Linear Regression algorithm

State	Year	Actual Result (tonnes)	Predicted Result (tonnes)	Average MSE (3 years)	Average MAE (3 years)
Melaka	2009	2003	2072.94	0.0089	0.0880
	2010	1387	1182.33		
	2011	1479	1666.08		
Perak	2009	4373	4617.9	0.0058	0.0758
	2010	4184	3962.44		
	2011	4176	3949.05		
Pahang	2009	7837	8025.26	0.0455	0.1761
	2010	7905	8815.35		
	2011	8081	8377.67		
Johor	2009	7428	7629.19	0.1672	0.3446
	2010	6951	8000.16		
	2011	6450	8230.82		

Results Neural Network Algorithm

The prediction results of rubber production are computed according to the state of each dataset using neural network is described. As can be seen from Table 6, the lowest value of MSE and MAE values are obtained in Perak, which are 0.0104 and 0.0693. On the other hand, Pahang has recorded the highest value of MSE is 0.1007 while MAE is 0.3077.

Table 6: Prediction results from neural network algorithm

State	Year	Actual Result (tonnes)	Predicted Result (tonnes)	Average MSE (3 years)	Average MAE (3 years)
Melaka	2009	2003	1856.81	0.0446	0.1571
	2010	1387	2007.7		
	2011	1479	1421.27		
Perak	2009	4373	4907.03	0.0104	0.0693
	2010	4184	4242.71		
	2011	4176	4135.5		
Pahang	2009	7837	8682.82	0.1007	0.3077
	2010	7905	8949.69		
	2011	8081	8628.25		

Johor	2009	7428	8578.97	0.0556	0.1688
	2010	6951	6952.56		
	2011	6450	6781.94		

Comparison between Machine Learning Models

This section focuses on the comparison of average MSE and MAE values obtained from each prediction model. The MAE achieved the lowest error and it is considered better than MSE. Thus, its value is been considered for choosing the most suitable machine learning algorithm. This is due to the MSE calculates the average error depends on total squared of errors. The total error is concentrated within a smaller number of increasingly large individual errors, the total square error will increase (Willmott and Matsuura, 2005). Since the MSE is computed by sum of squared error divided by n, the variance associated with the frequency distribution of error magnitudes will grow along with total square error. In comparison with learning model, Linear Regression has the least value of MAE and it could be concluded as the most accurate prediction result. Johor consistently exhibited higher error margins compared to other states, which may be attributed to larger plantation heterogeneity, diverse soil conditions, or variability in farming practices. This indicates that state-specific calibration or localized parameter inclusion may enhance prediction accuracy. Melaka recorded the lowest error margins across models, likely due to its smaller plantation scale and relatively homogeneous conditions, which reduce variability in production. Perak also showed lower errors, possibly because its plantation data is more consistent and rainfall patterns more stable. This contrast highlights how state-level differences in plantation size, soil heterogeneity, and climate variability strongly influence predictive accuracy.

Table 7: Comparison between models

Type of model	State	MSE (3 years)	Average MSE	MAE (3 years)	Average MAE
Random Forest	Melaka	0.3946	0.2448	0.4782	0.4169
	Perak	0.1222		0.3454	
	Pahang	0.0618		0.2259	
	Johor	0.4004		0.6180	
Decision Tree	Melaka	0.0794	0.1828	0.2383	0.3757
	Perak	0.1123		0.3339	
	Pahang	0.1241		0.3127	
	Johor	0.4155		0.6180	
Linear Regression	Melaka	0.0089	0.0569	0.0880	0.1711
	Perak	0.0058		0.0758	
	Pahang	0.0455		0.1761	
	Johor	0.1672		0.3446	
Neural Network	Melaka	0.0446	0.0528	0.1571	0.1757
	Perak	0.0104		0.0693	
	Pahang	0.1007		0.3077	
	Johor	0.0556		0.1688	

CONCLUSIONS

The prediction of rubber crop yield using machine learning has been successfully implemented and analysed. Random Forest, Decision Tree, Linear Regression, and Neural Network were applied in this study, with Linear Regression showing the most consistent performance due to its lowest MAE values. Hyperparameter optimisation further improved model accuracy, demonstrating the importance of parameter tuning in predictive modelling. However, this study was limited to climatic and plantation attributes such as temperature, rainfall, humidity, and planted area. The findings revealed state-level variations in accuracy, particularly in Johor, which consistently recorded higher error margins compared to other states. This may be due to larger plantation heterogeneity, diverse soil conditions, or variability in cultivation practices. Such discrepancies

suggest that state-specific calibration and the inclusion of more contextual parameters are necessary to enhance the robustness of predictions. Future improvements to this study should focus on the integration of broader agricultural and economic parameters to make the prediction models more realistic. These include soil fertility indices, soil pH, nutrient content, pest and disease outbreaks, fertilizer application patterns, and production costs. Incorporating such multidimensional data would provide a more holistic view of the factors influencing rubber yield. An accurate prediction crop yield system can help farmers as well as industry stakeholders in financial decisions, marketing and insurance (Zhao et al., 2020). In addition, real-time data collection using farm-based sensors (e.g., soil moisture, salinity, and pH sensors) and satellite imagery can significantly improve data quality and enable more adaptive prediction systems. Beyond individual algorithms, ensemble and hybrid approaches such as stacking, boosting, or integrating machine learning with crop growth models could further enhance predictive accuracy and resilience. By advancing toward these directions, future prediction systems can better support farmers, policymakers, and industry stakeholders in financial planning, risk management, and sustainable rubber production strategies. Therefore, this kind of system also can be suggested the fertiliser composition according to the soil nutrients level. Future research could explore ensemble learning or hybrid models that integrate machine learning with crop simulation or econometric approaches. An option such multi label classification also believe may improve predictive robustness beyond the performance of individual algorithms (Mohamed et al, 2017).

ACKNOWLEDGMENTS

The authors would like to thank Centre for Research and Innovation Management (CRIM), Universiti Teknikal Malaysia Melaka (UTeM) for sponsoring this work.

REFERENCES

1. “Media Statement Gross Domestic Product (GDP) By State”, Department of Statistics Malaysia, 2024.A. A. Khin, R. L. L. Bin, S. B. Kai, K. L. L. Teng, and F. Y. Chiun. “Challenges of the Export for Natural Rubber Latex in the ASEAN Market,” iIOP Conference Series: Materials Science and Engineering, vol. 548, no. 1, 2019.
2. “Methodology for Estimation of Crop Area and Crop Yield under Mixed and Continuous Cropping Publication prepared in the framework of the Global Strategy to improve Agricultural and Rural Statistics,” 2017.
3. B. Fulkerson, D. Michie, D. J. Spiegelhalter, and C. C. Taylor, “Machine Learning, Neural and Statistical Classification,” Technometrics, vol. 37, no. 4, p. 459, Nov. 1995. Shah, A. Dubey, V. Hemnani, D. Gala, and D. R. Kalbande, “Smart farming system: Crop yield prediction using regression techniques,” in Lecture Notes on Data Engineering and Communications Technologies, vol. 19, Springer, pp. 49–56, 2018.
4. C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” CLIMATE RESEARCH, vol. 30, pp. 79–82, Dec. 2005.
5. el Naqa and M. J. Murphy, “What Is Machine Learning?,” in Machine Learning in Radiation Oncology, Springer International Publishing, pp. 3–11, 2015.
6. J. K. Jaiswal and R. Samikannu, “Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression,” in Proceedings - 2nd World Congress on Computing and Communication Technologies, WCCCT 2017, Oct. 2017.
7. Kavitha. S, Varuna. S, and Ramya. R, “A Comparative Analysis on Linear Regression and Support Vector Regression,” Online International Conference on Green Engineering and Technologies (IC-GET, 2016.
8. Mohamed, R., Perumal, T., Sulaiman, M.N., Mustapha, N., Zainudin, M.N.S., “Modelling activity recognition of multi resident using label combination of multi label classification in smart home”, AIP Conference Proceeding, ICAST 2017.
9. N. da Silva, D. Hernane Spatti, R. Andrade Flauzino, L. H. B. Liboni, and S. F. dos Reis Alves, “Artificial Neural Network Architectures and Training Processes,” in Artificial Neural Networks, Springer International Publishing, 2017, pp. 21–28.

10. P. Charoen-Ung and P. Mittrapiyanuruk, “Sugarcane yield grade prediction using random forest with forward feature selection and hyper-parameter tuning,” in Advances in Intelligent Systems and Computing, vol. 769, pp. 33–42, 2019.
11. T. Osman, S. Shahjahan Psyche, M. Rafik Kamal, F. Tamanna, F. Haque, and R. M.Rahman, “Predicting Early Crop Production by Analysing Prior Environment Factors,” Advances in Intelligent Systems and Computing, vol. 538, 2017.
12. Y. Y. Song and Y. Lu, “Decision tree methods: applications for classification and prediction,” Shanghai Archives of Psychiatry, vol. 27, no. 2, pp. 130–135, Apr. 2015.L. Breiman, Random Forests, vol. 45. 2001.
13. Y. Zhao, A. B. Potgieter, M. Zhang, B. Wu, and G. L. Hammer, “Predicting wheat yield at the field scale by combining high-resolution Sentinel-2 satellite imagery and crop modelling,” Remote Sensing, vol. 12, no. 6, Mar. 2020.