Advanced and Sustainable Technologies
ASET
E-ISSN 2976-2294
Volume 4, No 2, December 2025 [371-384]

Boosting Algorithm Comparisons for the Prediction of Added Amount Macronutrients (NPK) of Harumanis Mango Tree Penology Stage

Erdy Sulino bin Mohd Muslim Tan[1]*, Marni Azira Binti Markom[2], Abu Hassan Abdullah[1], Norasmadi Abdul Rahim[1], Fathinul Syahir Ahmad Saad[1], Imaduddin Helmi Wan Nordin[3], Mohd Amri Zainol Abidin[3], and Yogesh CK[4]

[1]Faculty of Electrical Engineering and Technology, Universiti Malaysia Perlis, Malaysia.
[2]Faculty of Electronic Engineering and Technology, Universiti Malaysia Perlis, Malaysia.
[3]Faculty of Mechanical Engineering and Technology, Universiti Malaysia Perlis, Malaysia.
[4]Vellore Institute of Technology (VIT), India.

Received 16 August 2025, Revised 31 August 2025, Accepted 13 September 2025

ABSTRACT

The Harumanis mango, a prized cultivar grown in Perlis, Malaysia, requires meticulous nutrient management to enhance yield and fruit quality. Conventional soil nutrient analysis techniques are often expensive and time-consuming, highlighting the need for efficient predictive methods. This study explores the application of boosting algorithms to predict the added amount of NPK fertilizer macronutrient nitrogen (N), phosphorus (P), and potassium (K) critical for mango cultivation. The predictive models were developed based on soil nutrient data collected via TDR sensors throughout different Harumanis mango phenology stages. These data-driven models provide a cost-effective alternative to traditional soil testing, facilitating timely and precise nutrient management. To evaluate model performance, multiple boosting algorithms, including XGBoost, LightGBM, Gradient Boosting Regressor (GBR), and AdaBoost, were fine-tuned and assessed using performance metrics such as MAE, RMSE, R^2, RMSLE, and MAPE. Among these, the XGBoost model exhibited the highest predictive accuracy, achieving an MAE of 38.4046, RMSE of 51.6798, R^2 of 0.8278, RMSLE of 0.4507, and MAPE of 0.5739. The results indicate that the XGBoost model effectively forecasts soil nutrient levels, outperforming other evaluated models. Accurately predicting macronutrient concentrations enables targeted fertilization strategies, reducing costs and environmental impact while optimizing Harumanis mango production. However, the model relies on soil nutrient data and is highly dependent on accurate sensor readings. Future studies should focus on expanding the dataset and incorporating additional environmental parameters to further enhance model precision and applicability across diverse agricultural regions.

Keywords: NPK, Machine Learning, Prediction, CSTV, Boosting.

1. INTRODUCTION

The Harumanis mango, a highly esteemed variety originating from Perlis, Malaysia, is renowned for its exceptional organoleptic properties and economic significance [1, 2]. The optimal yield and fruit quality of this cultivar are contingent on various factors, with soil nutrient management playing a pivotal role [1, 3, 4]. Specifically, the availability of essential macronutrients (N, P, K) in the soil directly influences mango tree branch growth, vegetative flush, generative flush, flowering, fruit set, and ultimately, the overall success of cultivation [1, 3, 5].

Traditional methods for determining soil nutrient levels involve labor-intensive processes such as soil sampling, chemical extraction, and laboratory analyses using spectrophotometry or

*Corresponding author: erdysulino@unimap.edu.my

chromatography. These procedures require specialized equipment, skilled personnel, and extended processing times, making them costly and impractical for frequent use. This underscores the necessity for efficient and cost-effective methods to forecast soil macronutrient levels, which can facilitate prompt and well-informed decisions about fertilizer application and soil management strategies.

There is a significant gap in the research on the nutritional requirements of Harumanis trees. Previous studies have not comprehensively investigated the specific nutrient needs of Harumanis trees throughout their growth stages, compelling farmers to rely on empirical knowledge rather than data-driven methods. Furthermore, current research predominantly focuses on predicting nutrient addition based on fertilizer type and quantity, while overlooking soil NPK levels. This dependence on personal knowledge results in suboptimal nutrient management practices, where fertilizers are often applied based on general guidelines without considering the specific requirements of the trees during each growth stage [6].

In recent years, machine learning (ML) techniques have emerged as powerful tools for predicting complex phenomena in various domains, including agriculture [7–11]. Boosting algorithms have gained significant attention owing to their ability to handle high-dimensional data and achieve high levels of predictive accuracy. These algorithms combine multiple weak learners to create a strong predictive model, effectively capturing complex relationships within the data [12–14].

This study investigates the potential of four prominent boosting algorithms, XGBoost, LightGBM, Gradient Boosting Regressor (GBR), and AdaBoost, to predict the added amount of soil macronutrients in the context of Harumanis mango cultivation [13-16]. By evaluating and comparing the performance of these algorithms, this study aims to identify a robust and accurate predictive model that can assist farmers and agricultural practitioners in optimizing soil nutrient management practices, ultimately contributing to improved Harumanis mango yield and quality [4, 15].

2. METHODS AND METHODOLOGY

The research began with problem formulation and initial soil sample analysis. Data were collected from the first location, Muzium Mempelam Kuala Sala, Alor Star, and underwent data cleaning and labelling to ensure dataset quality. This step is crucial for removing noise and inconsistencies. The next step involved feature engineering based on target features, utilizing the Mitscherlich Method [22,23] and the modified arsine log calibration curve (MALCC) method [21] to derive important soil attributes. Subsequently, exploratory data analysis (EDA) was conducted to gain insights into the dataset and identify patterns or correlations that may influence model performance (Figure 1).

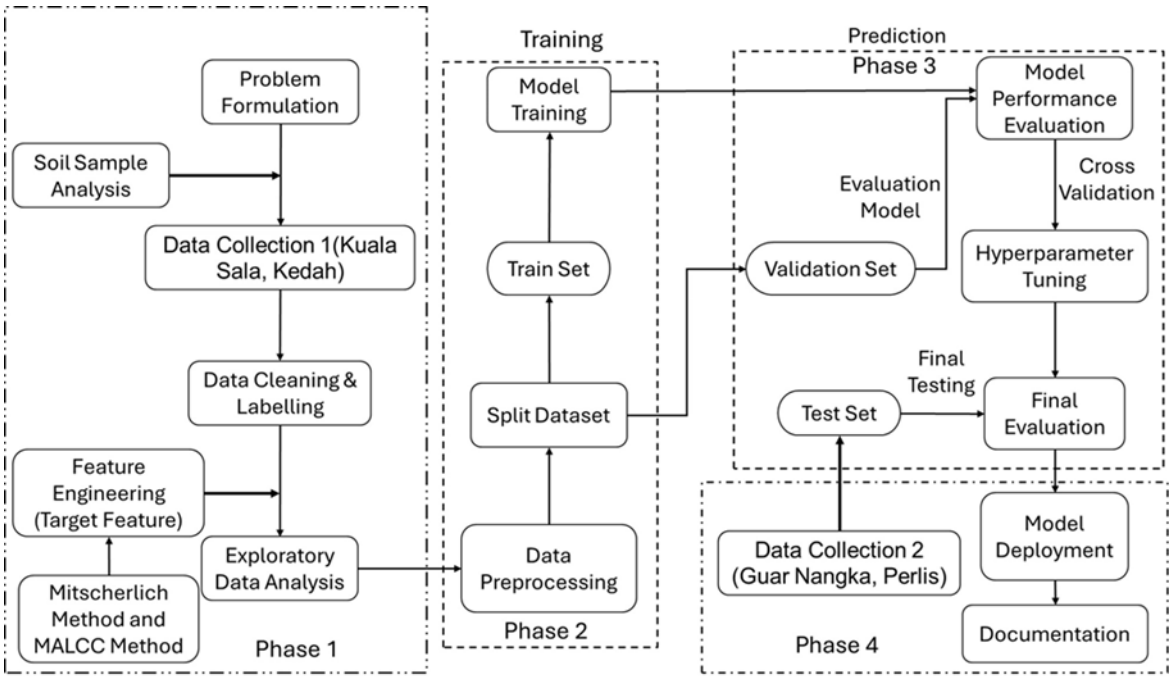

Figure 1: Methodology outlines a structured approach for predicting soil nutrient levels using machine learning and is divided into four phases.

In the model training phase, the cleaned and labelled data were split into training and validation sets. The training set was used for data preprocessing, which involved normalizing or standardizing the data, handling missing values, and transforming variables as needed. The preprocessed training data were then fed into the model training step, where machine learning boosting algorithms were applied.

After training, model performance was assessed using a validation set. Evaluation metrics such as MAE, MSE, RMSE, R^2, RMSLE, and MAPE were calculated to determine accuracy and robustness. Cross-validation ensured model generalization [25]. The model underwent hyperparameter tuning to further optimize performance. The best-performing model was then tested using data from a new location in Guar Nangka, Perlis.

Finally, the model was evaluated using the test set to ensure it performed well on unseen data. If the model met the performance criteria, it was deployed. The methodology, model, and findings have been documented for publication and future reference.

2.1 Data Collection

To conduct comprehensive data collection on Harumanis mango orchards, it is crucial to carefully select a representative orchard with a consistent production history and typical agricultural practice. For this data collection, we chose Muzium Mempelam Kuala Sala, Alor Star (Government Orchard, Department of Agriculture, Kedah, Loc: 5°58'05.2 "N 100°24'05.3" E) and Kampung Guar Nangka, Perlis (farmer orchard, Loc: 6°28'34.9 "N 100°17'06.0" E) as the sites for this study from 1 May 2020 to 30 May 2022. To ensure a proper study, permission was obtained from the Head of the Agriculture Department, Kedah, to collect data and the owner of the harumanis orchard, Kampung Guar Nangka, Perlis. The selected orchard was divided into distinct sampling zones, considering factors such as tree growth, soil texture, and management practices. This zoning approach allowed for a more accurate representation of the overall soil conditions in the orchard.

2.1.1 Data Collection 1: Train and Validation dataset

Initial soil samples were collected from Muzium Mempelam Kuala Sala, and analysed to determine the baseline levels of essential macronutrients, namely, N, P, and K. The data collection process involved standard soil sampling techniques to ensure representativeness across different parts of the Harumanis mango orchards in Figure 2.

Figure 2: Location of Data Collection 1: Muzium Mempelam Kuala Sala. Kedah. Malaysia.

Table 1 summarizes the soil and environmental data collected from Muzium Mempelam Kuala Sala, focusing on key parameters, such as N, P, K, Soil temperature, Soil Moisture, pH, electrical conductivity (EC), rainfall, and phenology stage. With 30,318 samples, the table provides statistical measures, including the mean, standard deviation, and quartiles, revealing the distribution and variability of each parameter. For instance, the N levels ranged from 2 to 143 units, with an average of 55.27 units and a standard deviation of 31.37 units, indicating moderate variability. These data are crucial for understanding soil nutrient dynamics in Harumanis mango orchards.

Table 1: Measurements and conversions for the refractive index of water.

Variable	count	mean	std	min	25%	50%	75%	max
N	30318	55.27	31.37	2	29	49	82	143
P	30318	57.11	33.20	0	34	54	73	178
K	30318	65.18	63.63	5	28	45	72	372
Temperature	30318	27.03	4.31	16	23.83	28.1	30.2	34.8
Moisture	30318	70.48	16.61	14.26	60.6	78.07	82.81	99.98
pH	30318	6.51	0.68	3.5	6.09	6.49	6.94	8.99
EC	30318	1.62	0.48	0.23	1.3	1.6	1.93	3.31
Stage	30318	2.35	1.89	0	0	3	4	5
Rainfall	30318	6.94	9.64	0	0.4	3.1	9.9	84.9

2.1.2 Data Collection 2: Test Dataset

Soil samples were collected from Kampung Guar Nangka, Perlis, to validate the developed predictive models in Figure 3. This secondary dataset was used to test models in a different but relevant geographical context to ensure generalization.

Figure 3: Location Data Collection 2: Harumanis orchard, Guar Nangka, and Perlis, Malaysia.

Table 2 provides descriptive statistics for soil and environmental parameters from 4,628 samples, highlighting essential factors such as nutrient levels, soil temperature, and soil moisture. The N levels showed an average of 73.0 units with moderate variability (std = 32.3), ranging from 20–134 units. P and K had averages of 72.9 and 106.7 units, respectively, with K exhibiting higher variability (std = 85.5). The average temperature and humidity were 28.3°C and 64.4%, respectively, indicating consistent environmental conditions. The average pH was 6.9, reflecting slightly acidic to neutral soil conditions. The EC and rainfall averages were 1.5 and 8.3, respectively.

Table 2: Data collected from the Harumanis orchards in Guar Nangka and Perlis. Malaysia.

Variable	count	mean	std	min	25%	50%	75%	max
N	4628	73.02	32.28	20	47	65	102	134
P	4628	72.85	37.61	14	46	70	91	153
K	4628	106.68	85.52	29	58	79	109	332
Temperature	4628	28.01	4.05	17.9	25.5	29.1	30.9	35.9
Moisture	4628	72.17	10.97	44.3	62.9	76.4	81	87.4
pH	4628	6.84	0.63	3.88	6.42	6.82	7.27	8.22
EC	4628	1.25	0.65	0.1	0.7	1.3	1.7	2.9
Stage	4628	2.31	1.92	0	0	3	4	5
Rainfall	4628	13.29	18.67	0	1.1	1.6	19.8	85

2.2. Data Preprocesing

Data preprocessing involves cleaning and transforming raw data by handling missing values, normalizing or standardizing the data, and encoding categorical variables for model training. Feature engineering was performed to enhance the predictive capabilities of these models. This step involved creating new variables or modifying existing ones to better capture the underlying relationships between soil properties and macronutrient levels. The target features included the levels of N, P, and K in the soil. To predict the amount of each nutrient (N, P, K) remaining in the soil after 14 days, we used a decay model that accounts for these losses:

Step Process:

1. Determination of the initial nutrient content
    - For an NPK fertilizer with a ratio of 15-15-15, each component (N, P, and K) comprises 15% of the fertilizer by weight.
    - When 200 g of fertilizer was applied, the initial amount of each nutrient was 30 g (200 g × 15%).

2. Convert to Elemental Form:
    - Nutrient forms in fertilizers (such as P_2O_5 and K_2O) are converted to their elemental forms:
        - Elemental N = 30 grams (N is already elemental)
        - Elemental P = 30 grams * 0.4364 = 13.092 grams
        - Elemental K = 30 grams * 0.8301 = 24.903 grams
3. Apply Decay Model:
    - The amount of each nutrient remaining after 14days was calculated using a decay model:
      $Nt = N_0 \times e^{-kt}$ (1)
    - Here, Nt is the remaining amount after t days, N_0 is the initial amount, k is the decay constant (which varies by nutrient and environment), and e is the base of the natural logarithm.
      Equations for Each Nutrient:
      For N $N_{14} = N_{\text{initial}} \times e^{-k_N \times 14}$ (2)
      For P $P_{14} = P_{\text{initial}} \times e^{-k_P \times 14}$ (3)
      For K $K_{14} = K_{\text{initial}} \times e^{-k_K \times 14}$ (4)

2.3. Model Development

Model development involves selecting appropriate machine learning algorithms, splitting the dataset into training and testing sets, training the model on the training data, tuning the hyperparameters, and evaluating its performance using the metric accuracy of R^2. This process ensures that the model is well-suited for accurately predicting outcomes based on the input data.

2.4. Model Training

Four different boosting algorithms, XGBoost, LightGBM, GBR, and AdaBoost, were used to predict soil macronutrient levels. Each algorithm was trained using the training set, and the hyperparameters were tuned based on the performance of the validation set.

1. Data Splitting: The dataset was divided into training and testing sets using an appropriate ratio (e.g., 80:20 or 70:30) to ensure model generalization [16,17]
2. Hyperparameter Tuning: Grid search or other optimization techniques were employed to determine the optimal hyperparameters for each boosting algorithm, maximizing the predictive performance of the training data [24]
3. Model Training: Each boosting algorithm was trained on the training data using optimized hyperparameters.

3. RESULTS AND DISCUSSION

Table 3 shows that the dataset underwent preprocessing to improve model performance. Yeo-Johnson transformation normalized skewed data, while Robust Scaling handled outliers. Features with high multicollinearity (>0.9) were removed. One categorical variable was encoded, and 10-fold K-Fold cross-validation ensured reliability. The final dataset had 15 features after iterative imputation and feature engineering.

Table 3: Data Preprocessing and Transformation Summary.

Description	Value	Explanation
Target Variable	Target_Feature	The dependent variable (what the model predicts). In this case, it refers to the amount of Target_Feature fertilizer needed.
Target Type	Regression	The problem type is regression, meaning the model predicts a continuous value rather than a category.
Original Data Shape	(30,318, 10)	The raw dataset has 30,318 samples (rows) and 10 features (columns) before preprocessing.
Transformed Data Shape	(29,256, 15)	After preprocessing, the dataset was reduced to 29,256 samples and expanded to 15 features (possibly due to feature engineering or encoding categorical variables).
Transformed Train Set Shape	(20,160, 15)	The training dataset consists of 20,160 samples with 15 features.
Transformed Test Set Shape	(9,096, 15)	The test dataset consists of 9,096 samples with 15 features.
Multicollinearity Threshold	0.9	Features with a correlation above 0.9 were removed to avoid redundancy and overfitting.
Outliers Threshold	0.05	The most extreme 5% of outliers were removed to improve model performance.
Transformation Method	yeo-johnson	The Yeo-Johnson transformation was applied to normalize skewed numerical features. Unlike log transformation, it works for both positive and negative values.
Normalize Method	robust	The Robust Scaler was used to normalize data based on the interquartile range (IQR), making it more resistant to outliers.
Transform Target Method	yeo-johnson	The Target_Feature was also transformed using Yeo-Johnson to improve normality and model performance.
Fold Generator	KFold	K-fold cross-validation was used to validate model performance by splitting the data into multiple folds.
Fold Number	10	10-fold cross-validation was performed, meaning the data was split into 10 subsets, with each subset used as a test set once while the others were used for training.

3.1 Performance Evaluation

The performance of the models was evaluated using a training set. Key metrics, such as MAE, RMSE, and R^2, were calculated to assess the accuracy and robustness of each model.

Table 4: Performance Metrics.

Model Hyperparameter	MAE	MSE	RMSE	R2	RMSLE	MAPE
XGBoost	11.5327	286.7185	16.8774	0.9793	0.1044	0.0583
LightGBM	12.2258	409.9217	20.1082	0.9704	0.1364	0.073
GBR	11.9667	313.139	17.6516	0.9773	0.1189	0.0631
AdaBoost	32.1306	1632.444	40.3736	0.8818	0.2204	0.1648

3.2 Prediction Performance Evaluation

The predictive accuracy of each model was evaluated using the coefficient of determination, which quantifies the proportion of variance in the dependent variable (soil macronutrient levels) explained by the independent variables (soil properties). Higher R^2 from Table 4 values indicates a better model fit and predictive power.

3.2.2 Analysis of the Models

The performance of the models showed some changes after hyperparameter tuning. The Extreme Gradient Boosting model improved further, with an R^2 increasing to 0.8278, meaning the model now explains 82.78% of the variance in the soil macronutrient levels, making it an even more reliable predictor. The error metrics for XGBoost also improved slightly, reinforcing its position as the best-performing model.

Table 5: Prediction Tuned Model using Guar Nangka, Perlis Orchard.

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
XGBoost	38.4046	2670.7991	51.6798	0.8278	0.4507	0.5739
LightGBM	41.6924	3453.6349	58.7676	0.7773	0.5003	0.7189
GBR	52.2665	3935.0174	62.7297	0.7463	0.6692	1.0237
AdaBoost	61.9511	5826.416	76.331	0.6244	0.7271	1.5183

3.3 Evaluation and Model Selection

Based on the performance metrics, the best-performing model was XGBoost, with a score R^2 of 82.78% for the final evaluation. This model was then tested on a secondary dataset (Data Collection 2) to verify its generalizability and reliability in different environments. Table 6 compares the actual Target_feature values with their predicted labels, highlighting the percentage error for each prediction. Target_feature values ranged from approximately 100 to 450, with predictions closely matching the actual values, demonstrating the model's accuracy. The percentage error varied from as low as 1% to as high as 16%, indicating instances in which the predictions were either highly accurate or slightly off. For example, an actual Target_feature value of 98.9 has predictions of 96.6 and 99.7, resulting in percentage errors of 2% and 1%, respectively. Overall, the table illustrates the effectiveness of the model in predicting Target_feature values with reasonable accuracy.

Table 6: Comparison of the actual Target_feature values with their predicted labels.

Target_feature	Prediction_label	Percentage Error
103.7	90.5	13%
103.7	89.7	13%
100.5	91.1	9%
98.9	96.6	2%
98.9	99.7	1%
438.6	408.7	7%
438.5	380.4	13%
444.9	372.2	16%
446.5	389.9	13%
452.9	423.4	7%

3.3.1 Model Analysis

The residual plot for the XGBoost model in Figure 4 shows the residuals (differences between the predicted and actual values) against the predicted values for both the training and test sets. The training set had an R^2 of 0.989, indicating that the model explained 98.9% of the variance, while the test set had an R^2 of 0.982, showing strong generalization with 98.2% variance explained. The residuals were mostly centered around zero, suggesting minimal bias and good model performance. The distribution plot on the right shows a roughly normal distribution of residuals, reinforcing that the errors are randomly distributed without obvious patterns.

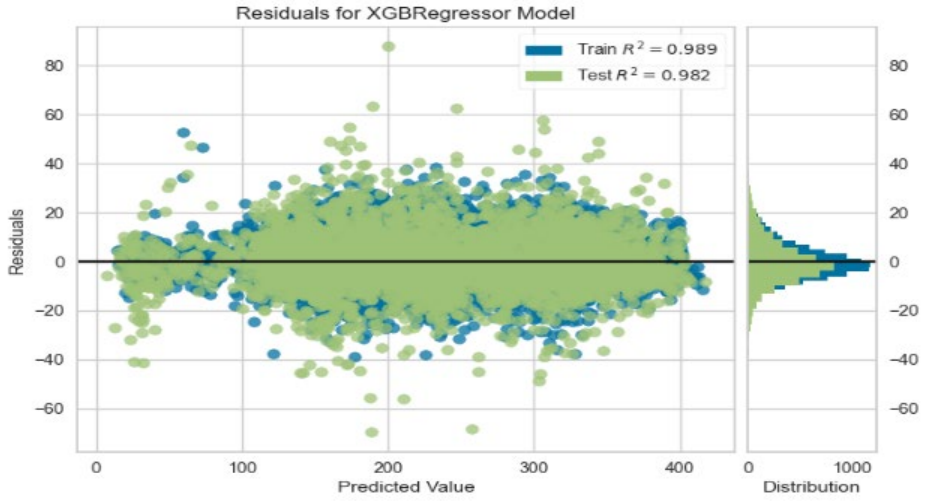

Figure 4: The residual plot for the XGBoost model.

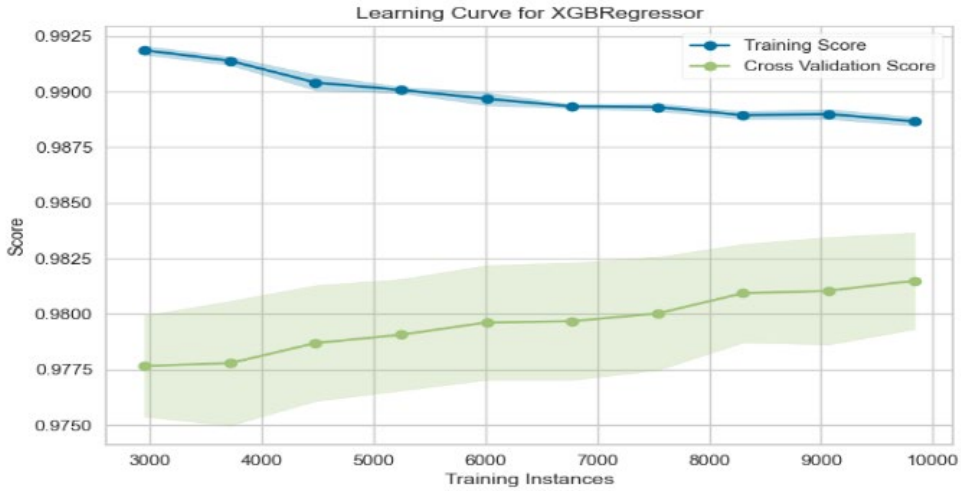

Figure 5: The learning curve for XGBoost.

The learning curve for XGBoost in Figure 5 shows the relationship between the model performance and the number of training instances. The Training Score (blue line) started high, close to 0.9925, but gradually decreased as the number of training instances increased, stabilizing at approximately 0.990. This slight decline suggests that the model becomes less overfitted as more data are included.

The Cross Validation score (green line) started lower, at approximately 0.9775, and increased steadily, reaching approximately 0.980 as more data were added, indicating improved generalization. The gap between the training and cross-validation scores narrowed slightly, suggesting that the model was learning and generalizing well with the additional data. The shaded

area around the cross-validation score represents the variability in performance, which remained relatively consistent, indicating stable model behavior across different subsets of the data.

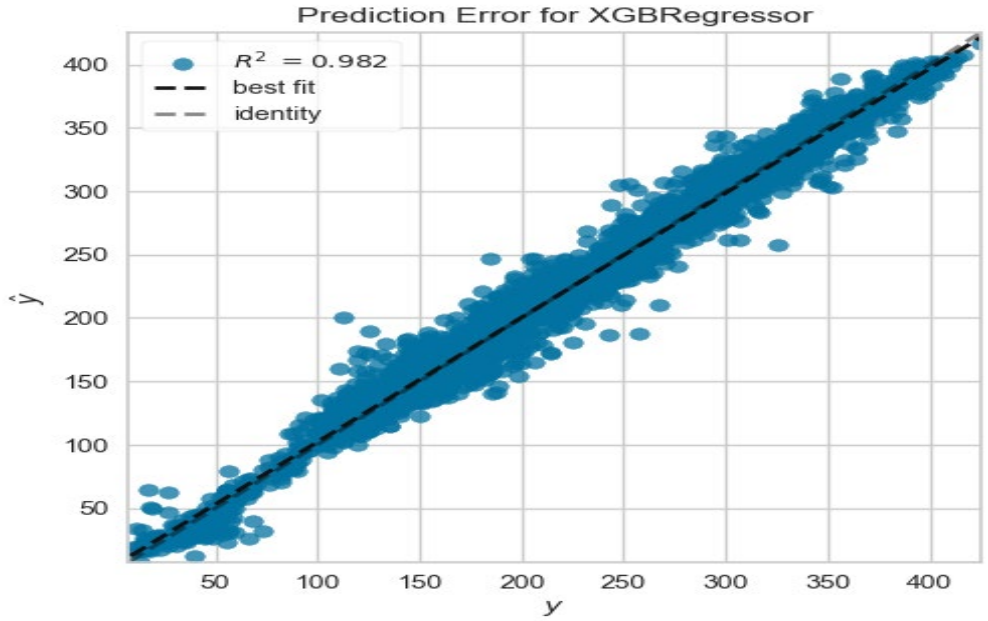

Figure 6: The prediction error plot for the XGBoost.

The prediction error plot for the XGBoost in Figure 5 Regressor shows the relationship between the actual values (y) and the predicted values $\left(y^{\hat{y}}\right)$. The R^2 value of 0.982 indicates that the model explains 98.2% of the variance in the data, suggesting a strong predictive accuracy.

The plot features two key lines: the identity line (dashed, grey), where $y^{y} = \hat{y}$, representing perfect predictions, and the best fit line (solid, black), which represents the actual relationship between predicted and true values. The points are closely clustered around the identity line, indicating that the predictions are close to the actual values. The proximity of the best-fit line to the identity line further confirms the accuracy of the model. The outliers were minimal, demonstrating that the model performed consistently well across a range of data points.

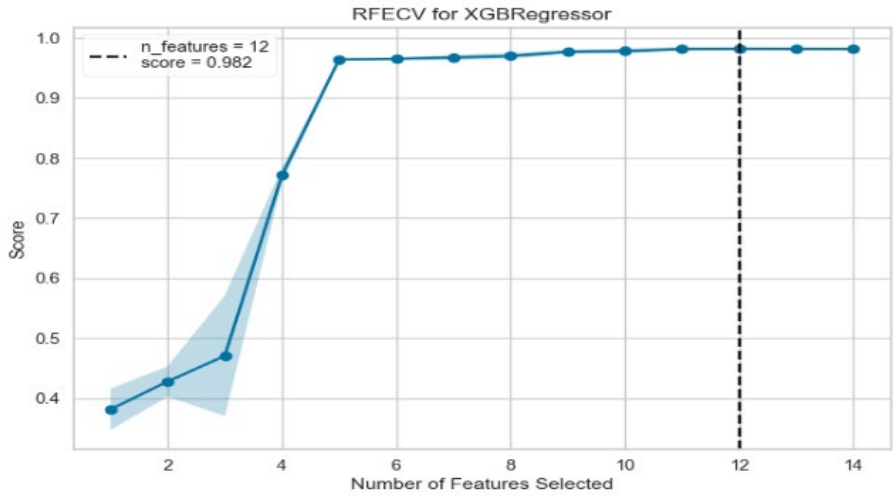

Figure 7: The recursive feature elimination with cross-validation (RFECV) plot for XGBoost.

The recursive feature elimination with cross-validation (RFECV) plot for XGBoost in Figure 7 shows how the model performance changes as different numbers of features are selected. The x-

axis represents the number of features selected, and the y-axis represents the cross-validation score, which indicates the model's performance.

As the number of selected features increased from 1 to 4, the score improved significantly, indicating that these initial features were highly important for the model performance. After reaching 12 features, the score stabilized at a high value of 0.982, suggesting that adding more features beyond this point did not significantly improve the model performance. The vertical dashed line marks the point at which 12 features are selected, indicating that this is the optimal number of features for maximizing the model accuracy. The shaded area around the line represents the variability in the cross-validation score, which narrows as more features are added, indicating an increased model stability.

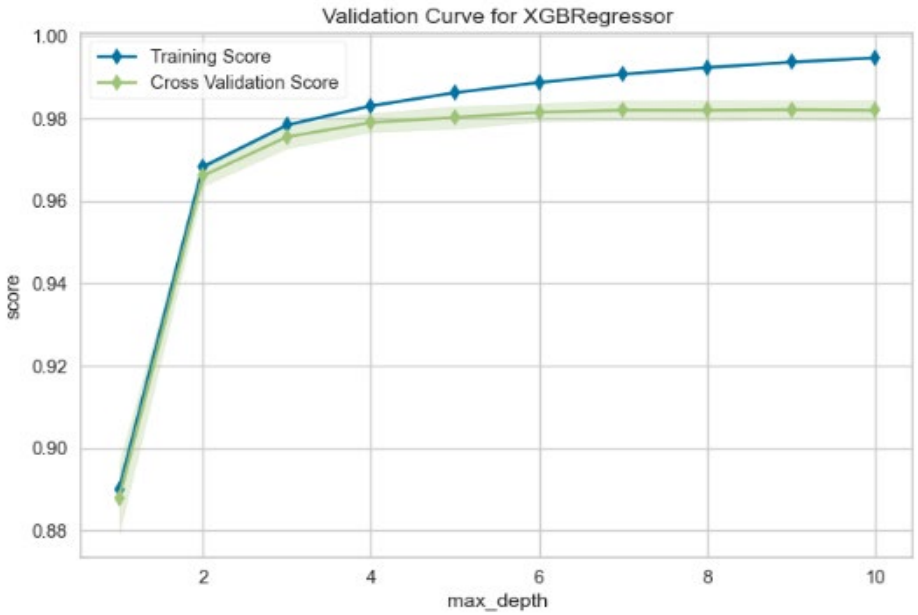

Figure 8: Validation curve for the XGBoost.

The validation curve for the XGBoost in Figure 8 illustrates how the model performance varies with changes in the max_depth parameter, which controls the maximum depth of each tree in the ensemble. X-axis: Represents the max_depth values ranging from 1 to 10. Y-axis: Represents the model performance score, with separate lines for the training (blue) and cross-validation (green) scores.

As max_depth increased from 1 to 4, both the training and cross-validation scores increased sharply, indicating that deeper trees improved the model's ability to capture patterns in the data. Beyond a *max_depth* of 4, the training score continues to increase, approaching nearly 1.0, suggesting that the model fits the training data very well. However, the cross-validation score levels off at approximately 0.98 and shows a slight decline beyond a max_depth of 6. This divergence indicates potential overfitting, where the model performs well on the training data but does not generalize as effectively to unseen data.

The plot suggests that a *max_depth* of approximately 4 to 6 provides a good balance between model complexity and generalization, where the model achieves high accuracy without significant overfitting.

4. CONCLUSION

In this study, the application of various boosting algorithms, particularly XGBoost, for predicting soil macronutrient levels in the context of Harumanis mango cultivation was investigated. The results of the different analyses, including residual plots, learning curves, and feature importance, offer insights into the model performance and the critical factors influencing soil nutrient prediction.

XGBoost demonstrated exceptional predictive capability, with an R^2 value of 0.982 on the test set, indicating that the model accurately explained 98.2% of the variance in soil nutrient levels. The minimal residuals and alignment of the predicted values with the actual values confirmed the model's reliability and robustness. The learning curve analysis further supported this, showing a good balance between the training and cross-validation scores, which suggests that the model generalizes well to unseen data without significant overfitting.

Feature importance analysis revealed that the phonological stages of Harumanis mango, particularly the flowering stage, are the most influential predictors of soil macronutrient levels. This emphasizes the critical role of plant growth stages in nutrient uptake and availability. The high importance of nitrogen and potassium also aligns with the agronomic understanding that these nutrients are vital for plant growth and for fruit development.

ACKNOWLEDGEMENTS

This work was supported by Yayasan Inovasi Malaysia (YIM) and the Ministry of Science, Technology, and Innovation (MOSTI), Malaysia, under the MyGRiS grant scheme [Grant amount: RM50,000.00, 2023–2024]. The authors would like to express their sincere gratitude for the financial support that made this research possible.

REFERENCES

[1] Nasron, N., Ghazali, N. S., Shahidin, N. M., Mohamad, A., Pugi, S. A., & Razi, N. M. Soil suitability assessment for Harumanis mango cultivation in UiTM Arau, Perlis. IOP Conference Series: Earth and Environmental Science, vol. 620, issue 1 (2021) p.012007. https://doi.org/10.1088/1755-1315/620/1/012007

[2] Uda, M. N. A., Gopinath, S. C. B., Hashim, U., Bakar, A. H. A., Anuar, A., Bakar, M. A. A., Sulaiman, M. K., & Azizah, N. Harumanis mango: Perspectives in disease management and advancement using interdigitated electrodes (IDE) nano-biosensor. IOP Conference Series: Materials Science and Engineering, vol. 864, issue 1 (2020) p.012180. https://doi.org/10.1088/1757-899X/864/1/012180

[3] Hazis, N. H., Aznan, A. A., Jaafar, M. S., Azizan, F. A., Ruslan, R., & Rukunudin, I. H. Assessment of carbohydrate contents in Perlis Harumanis mango leaves during vegetative and productive growth. IOP Conference Series: Materials Science and Engineering, vol. 429 (2018) p.012025. https://doi.org/10.1088/1757-899X/429/1/012025

[4] Yusuf, S., Wahab, Z., Zakaria, Z., Subbiah, V. K., Masnan, M. J., & Wahab, Z. Morphological variability identification of Harumanis mango (Mangifera indica L.) harvested from different locations and tree age. Tropical Life Sciences Research, vol. 31, issue 2 (2020) pp.107–143. https://doi.org/10.21315/tlsr2020.31.2.6

[5] Razi, N. M., Zakaria, S. N. S., Shahidin, N. M., & Nasron, N. Assessments of nutrients content in soil and leaves of Harumanis mangoes and its relationship with the yield. IOP Conference Series: Earth and Environmental Science, vol. 1051, issue 1 (2022) p.012017. https://doi.org/10.1088/1755-1315/1051/1/012017

[6] Sharif, M. Fertilizer management for sustainable agriculture. Agricultural Review Journal, vol. 18 (2019) pp.55–61.
[7] Fink, R. K., Hoskinson, R. L., & Hess, J. R. From prediction to prescription: Intelligent decision support for variable rate fertilization. ASAE Annual International Meeting (2001) Paper No. 01-5527. https://doi.org/10.13031/2013.5527
[8] Motia, S., & Reddy, S. R. N. Exploration of machine learning methods for prediction and assessment of soil properties for agricultural soil management: A quantitative evaluation. Journal of Physics: Conference Series, vol. 1950, issue 1 (2021) p.012037. https://doi.org/10.1088/1742-6596/1950/1/012037
[9] Khanal, S., Fulton, J. P., Klopfenstein, A., Douridas, N., & Shearer, S. A. Integration of high-resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. Computers and Electronics in Agriculture, vol. 153 (2018) pp.213–225. https://doi.org/10.1016/j.compag.2018.07.016
[10] Chlingaryan, A., Sukkarieh, S., & Whelan, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture, vol. 151 (2018) pp.61–69. https://doi.org/10.1016/j.compag.2018.05.012
[11] Qin, Z., Myers, D. B., Ransom, C. J., Kitchen, N. R., Liang, S., Camberato, J. J., Carter, P. R., Ferguson, R. B., Fernández, F. G., Franzen, D. W., Laboski, C. A. M., Malone, B. D., Nafziger, E. D., Sawyer, J. E., & Shanahan, J. F. Application of machine learning methodologies for predicting corn economic optimal nitrogen rate. Agronomy Journal, vol. 110, issue 6 (2018) pp.2596–2607. https://doi.org/10.2134/agronj2018.03.0222
[12] Schapire, R. E. The boosting approach to machine learning: An overview. In Nonlinear Estimation and Classification, Springer (2003) pp.149–171.
[13] Zemel, R. S., & Pitassi, T. A gradient-based boosting algorithm for regression problems. Advances in Neural Information Processing Systems, vol. 13 (2000) pp.696–702. http://papers.nips.cc/paper/1797-a-gradient-based-boosting-algorithm-for-regression-problems.pdf
[14] Friedman, J. H. Greedy function approximation: A gradient boosting machine. Annals of Statistics, vol. 29, issue 5 (2001) pp.1189–1232.
[15] Daoud, E. A. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. International Journal of Computer and Information Engineering, vol. 13, issue 1 (2019) pp.6–10. https://doi.org/10.5281/zenodo.3607805
[16] Avnimelech, R., & Intrator, N. Boosting regression estimators. Neural Computation, vol. 11, issue 2 (1999) pp.499–520. https://doi.org/10.1162/089976699300016746
[17] Bühlmann, P., & Hothorn, T. Boosting algorithms: Regularization, prediction and model fitting. Statistical Science, vol. 22, issue 4 (2007) pp.477–505. https://doi.org/10.1214/07-STS242
[18] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. CatBoost: Unbiased boosting with categorical features. arXiv preprint arXiv:1706.09516 (2017). https://doi.org/10.48550/arXiv.1706.09516
[19] Ramesh, D. Adaboost.RT-based soil N–P–K prediction model for soil and crop-specific data: A predictive modelling approach. In Proceedings of International Conference on Advances in Computing and Data Sciences (2018) pp.229–239. https://doi.org/10.1007/978-3-030-04780-1_22
[20] Ducamp, M. J. L., et al. An integrated approach for mango production and quality management. Acta Horticulturae, vol. 820 (2009) pp.225–232. https://doi.org/10.17660/ActaHortic.2009.820.26
[21] Correndo, A. A., Salvagiotti, F., García, F. O., & Gutiérrez-Boem, F. H. A modification of the arcsine log calibration curve for analysing soil test value–relative yield relationships. Crop and Pasture Science, vol. 68, issue 3 (2017) pp.297–304.
[22] Ferreira, I. E., Zocchi, S. S., & Baron, D. Reconciling the Mitscherlich's law of diminishing returns with Liebig's law of the minimum: Some results on crop modeling. Mathematical Biosciences, vol. 293 (2017) pp.29–37.

[23] Correndo, A. A., Pearce, A., Bolster, C. H., Spargo, J. T., Osmond, D., & Ciampitti, I. A. The soiltestcorr R package: An accessible framework for reproducible correlation analysis of crop yield and soil test data. SoftwareX, vol. 21 (2023) pp.101275.
[24] Thorson, J., Collier-Oxandale, A., & Hannigan, M. Using a low-cost sensor array and machine learning techniques to detect complex pollutant mixtures and identify likely sources. Sensors, vol. 19, issue 17 (2019) pp.3723. https://doi.org/10.3390/s19173723
[25] Zhang, Y., Ma, J., Liang, S., Li, X., & Li, M. An evaluation of eight machine learning regression algorithms for forest aboveground biomass estimation from multiple satellite data products. Remote Sensing, vol. 12, issue 24 (2020) p.4015. https://doi.org/10.3390/rs12244015

Conflict of interest statement: The authors declare no conflict of interest. The funding sponsor, Yayasan Inovasi Malaysia (YIM) under the Ministry of Science, Technology and Innovation (MOSTI), Malaysia, had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Author contributions statement: Conceptualization, E.S.M.M.T.; Methodology, E.S.M.M.T.; Data Collection, E.S.M.M.T. & M.A.Z.A.; Model Development, E.S.M.M.T.; Result Interpretation, E.S.M.M.T. & M.A.B.M.; Visualization, A.M.A. & P.N.; Literature Review, A.M.A.; Writing – Original Draft Preparation, E.S.M.M.T. & M.A.B.M.; Writing – Review & Editing, I.H.W.M, A.H.A., N.A.R., & F.S.A.S.; Supervision, A.H.A., N.A.R., & F.S.A.S.; Project Administration, A.H.A., N.A.R., & F.S.A.S.