ANALYSIS OF MACHINE LEARNING ALGORITHM PERFORMANCE IN PREDICTING ULTISOL SOIL NUTRIENTS BASED ON IMPEDANCE VALUES

Dwi Rahmah Amanda, Samsidar, Jesi Pebralia*

Physics Study Program, Faculty of Science and Technology, University of Jambi, Street. Jambi-Muaro Bulian No.KM. 15, Muaro Jambi, 3636 Indonesia

*email: jesipebralia@unja.ac.id

ABSTRACT

A study comparing the performance of machine learning algorithms to predict soil nutrient values based on soil impedance has been conducted. The algorithm models used include Linear Model, K-Nearest Neighbors (K-NN) with n-neighbors 3, 18, 21, 24, 27, and 30, Decision Tree with max depth 3, and Random Forest with n-estimators 6 and 21. During the training phase, 10 model variations with the best performance were found, including Linear Model, K-NN (n-neighbors), Decision Tree (max depth 3), and Random Forest (n-estimators 6 and 21). In the testing phase, Random Forest (n-estimator 21) showed the best performance with MAE = 0.15%, MSE = 0.09%, RMSE = 0.31%, and accuracy = 99.85%. Regression analysis indicated an R-squared value of 0.924, indicating that most of the variations in soil impedance values can be explained by variations in soil nutrient values. A regression value approaching 1 indicates that the regression model used has a very good ability to explain the variations observed in the data. This indicates that most of the variations in the dependent variable (the variable being predicted, which is the nutrient values) can be explained by the independent variable (the predictor variable, which is the soil impedance values) in the model. Correlation analysis resulted in a strong negative correlation between impedance and Al, Fe, K, Ca, Zn, Ni, Ta, V, Cr, and Mn (values -0.81 to -0.99), while a positive correlation occurred with Mg, Si, S, Cl, Ti, Zr, and Ga (values 0.65 to 0.99). This indicates that an increase in impedance values is generally followed by an increase in nutrient values.

Keywords: *Soil Nutrients; Soil Impedance; Machine Learning; Linear Model; K-Nearest Neighbors; Decision Tree; Random Forest*

INTRODUCTION

The success of agricultural production depends not only on the type of crops and cultivation techniques but also on the fertility conditions of the soil that support plant nutrition (Ifadah et al., 2021). One type of soil commonly used as a planting medium is Ultisol, typically found in tropical or subtropical forests (Subardja et al., 2014). Ultisols naturally have relatively low fertility, but with proper handling such as fertilizer addition, organic matter, or lime, this soil can become more productive and fertile (Kasno, 2019). Soil fertility is closely related to the availability of nutrients, both macro (N, P, K) and micro (Cu, Fe, Mn, Zn) (Hou et al., 2020). The proper mineral composition supports plant growth and health, and further measurements in the laboratory are needed to determine the nutrient content present in the soil.

Laboratory analysis conducted involves extraction using chemical solvents to extract nutrients from soil samples. The extraction method may vary depending on the nutrient to be measured (Umaternate, 2014). Soil nutrient measurements through laboratory testing currently require a relatively long time, necessitating the design of a more efficient, reliable, and practical system for use. To increase productivity in agriculture, the application of technology is needed to determine the nutrient content in the soil. One way is by utilizing machine learning to predict nutrient values. Machine learning has the ability to identify patterns, make decisions, and improve performance as more data and training sets are accumulated (Ambarwari et al., 2020).

One study related to the application of machine learning in agriculture was conducted by Bouslihim et al. in 2021, comparing the performance of two machine learning algorithms, namely Multiple Linear Regression and Random Forest. The modeling was done to predict the Mean Weight Diameter (MWD) value as an index of soil aggregate fertility. Model performance was assessed by calculating the coefficient of determination (R^2)

and Root Mean Squared Error (RMSE) values for each model. The results showed that Multiple Linear Regression had R^2 values ranging from 0.52 to 0.59 and RMSE values ranging from 0.277 to 0.401, while Random Forest had R^2 values ranging from 0.57 to 0.6 and RMSE values ranging from 0.261 to 0.410. The study concluded that the Random Forest algorithm modeling had better performance compared to Multiple Linear Regression algorithm.

State of the Art

Referring to previous studies, the author will utilize machine learning algorithms to predict nutrient values. Soil impedance measurements are conducted using an Earth Resistance Tester, a specialized tool for measuring soil resistance to electric current (Azzyati et al., 2019). The soil impedance values will be identified as potential indicators correlated with the physical and chemical properties of soil and nutrient content. Unlike previous studies, the author will compare the performance of machine learning algorithms consisting of five different algorithms (Linear Model, K-Nearest Neighbors, Decision Tree, and Random Forest) to determine the most effective algorithm by examining the smallest error values and the most accurate predictions of soil nutrient values based on measured soil impedance values. Another difference is that if previous studies only predicted two or three elements related to soil fertility values, the author will predict soil nutrient values consisting of 17 nutrient elements (Mg, Al, Si, Fe, S, Cl, K, Ca, Ti, Zn, Zr, Ni, Ga, Ta, V, Cr, Mn) based on soil impedance values.

METHOD

1. Equipment and Mateials

In this research, materials are required as the objects of study. The materials processed in this study are datasets of soil nutrient values and soil impedance values on ultisol soil samples. The equipment used in this research includes; Python software, Jupyter Notebook, Libraries (Pandas, Numpy, Seaborn, Statsmodel, Matplotlib, Scikit Learn), and a PC.

2. Object and Research Variables

The research object used in this study is the dataset of soil nutrient values and soil impedance values from ultisol soil testing conducted at the Soil Laboratory of the Faculty of Agriculture, Universitas Jambi. Variables are crucial points in a study, consisting of dependent variables and independent variables. Dependent variables are variables influenced by other variables, while independent variables are not dependent on other variables. The dependent variable in this study refers to the percentage values of nutrients (Mg, Al, Si, Fe, S, Cl, K, Ca, Ti, Zn, Zr, Ni, Ga, Ta, V, Cr, Mn) contained in the ultisol soil dataset. These nutrient values will be predicted using machine learning algorithm modeling. Meanwhile, the independent variable in this study is the impedance values found in the ultisol soil dataset that has been tested.

3. Data Analysis Technique

The data analysis process in this study is conducted using machine learning algorithm modeling designed on the Jupyter Notebook platform using the Python language. The modeling begins with preparing two datasets, namely the train data and the test data. The process starts with regression analysis on the train data to determine the extent of the cause-and-effect relationship and the relationship between independent variables (explanatory variables) and dependent variables (variables to be predicted or explained). The output produced in regression analysis is the R-Squared value. The R-squared value (coefficient of determination) is a measure indicating the extent to which the variability of the dependent variable (output) can be explained by the independent variables (input) in the linear regression model. After the data passes through the regression analysis process, the next step is to perform train-test-split, which involves dividing the data into a training set and a testing set. Research conducted by Fashoto et al. (2021) explains how the use of 30% for testing data and 70% for training data empirically produces the best results. The author divides the dataset into 70% for the training set and 30% for the testing set. The goal is to measure the performance of the model trained with unseen data. By separating this data, the system can identify the extent to which the model can apply the patterns learned from the training data to new situations. The next step is model training. Model training will be done using four modeling algorithms (Linear Regression, K-Nearest Neighbors (K-NN), Decision Tree, and Random Forest). In the modeling algorithms used, variations in values are performed for n-neighbors (number of nearest neighbors in the K-NN algorithm), Random state (number of classes or features in the Decision Tree algorithm), and n-estimators (number of decision trees in the Random forest algorithm). From the dataset available, 10 variations of values (n-neighbors, Random state, and n-estimators) are conducted to see the modeling performance produced by each algorithm used.

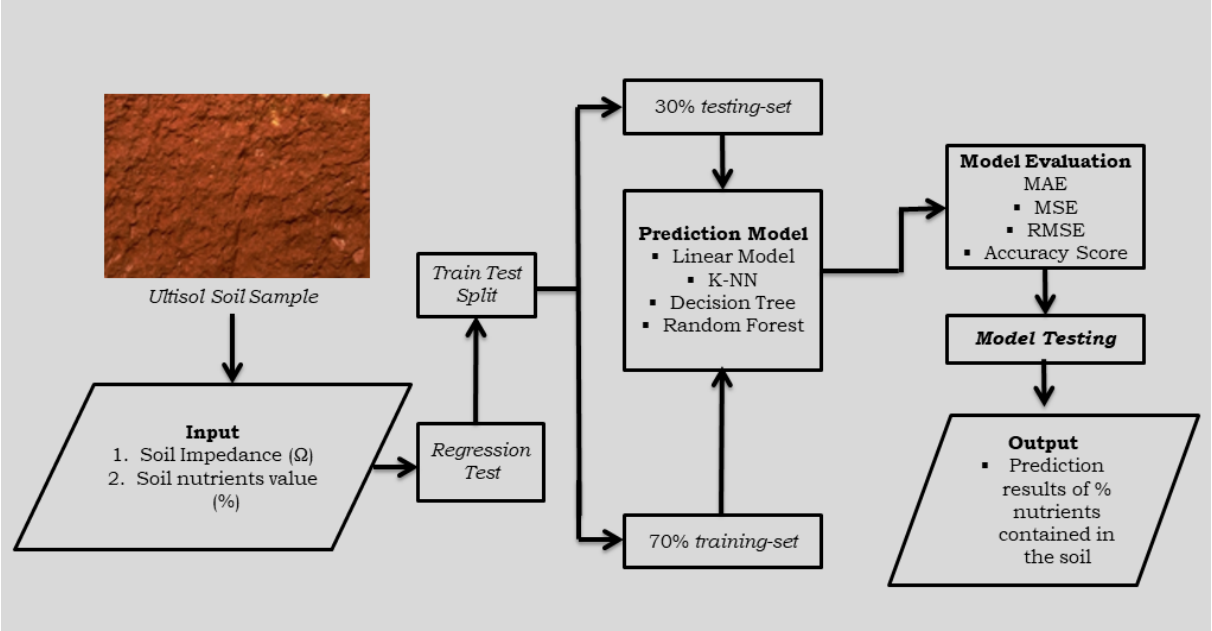

Figure 1. Flowchart of Machine Learning Prediction Modeling Workflow.

4. Model Evaluation

The performance of Machine Learning prediction is assessed by calculating the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the model accuracy score. Model accuracy is a common evaluation method used in machine learning to measure how well the model predicts the actual target or output values. The MAE, MSE, and RMSE values closer to 0 indicate better performance of the model in predicting data. Meanwhile, an accuracy score approaching 1 in machine learning models indicates that the model is closer to perfect performance in classification or prediction.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

dimana n menyatakan jumlah data, y_i sebagai nilai aktual data ke-I, and $\hat{y}_i$menyatakan nilai prediksi dari model untuk data ke-i.

(9)

Table 1. Performance Classification Based on Accuracy Values

Accuracy Range	Performance Classification
90%-100%	Excellent
80%-90%	Good
70%-80%	Fair
60%-70%	Poor
≤60%	Vary Poor

(Source : Sang et al, 2021)

RESULT AND DISCUSSION

1. Preprocessing Data

Data preprocessing aims to prepare raw data into a suitable and useful format for machine learning algorithms. This process involves several actions, such as data transformation, categorical variable encoding, and splitting the data into training data (train data) and test data (test data). The dataset used in this study consists of 45 sets of train data and 3 sets of test data. The following are the train data and test data used in this study:

Table 2. Train Data Samples for the First 5 Data and Last 5 Data

Impeandsi (Ω)	Mg (%)	Al (%)	Si (%)	Fe (%)	S (%)	Cl (%)	K (%)	Ca (%)	Ti (%)	Zn (%)	Zr (%)	Ni (%)	Ga (%)	Ta (%)	V (%)	Cr (%)	Mn (%)
114.28	4.36	19.03	61.23	10.81	0.10	0.08	0.61	0.06	3.26	0.01	0.31	0.01	0.01	0	0.06	0.31	0.02
78.56	2.33	21.53	59.57	12.04	0.07	0.09	0.62	0	3.24	0.01	0.32	0.01	0.01	0.02	0.07	0.05	0.02
134.5	2.84	19.98	61.14	11.35	0.09	0.08	0.67	0.05	3.31	0.01	0.32	0.01	0.01	0	0.07	0.05	0.02
102.13	3.34	21.98	56.91	12.99	0.07	0.08	0.67	0	3.43	0.01	0.36	0.01	0.01	0	0.08	0.04	0.02
93.17	3.09	18.95	60.17	12.71	0.12	0.08	0.71	0.12	3.49	0.01	0.36	0.01	0.01	0	0.08	0.05	0.03
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
82.63	4.3	20.52	59.62	11.26	0.08	0.06	0.58	0.04	3.09	0.01	0.31	0	0.01	0	0.07	0.04	0.01
79.35	5.72	21.82	54.45	13.43	0.05	0.05	0.61	0	3.37	0.01	0.32	0	0.01	0.02	0.08	0.04	0.02
87.52	4.35	16.08	67.22	8.16	0.09	0.06	0.46	0.36	2.75	0.01	0.28	0.01	0	0.02	0.06	0.05	0.02
58.32	4.35	20.3	62.16	9.19	0.08	0.07	0.43	0.13	2.84	0.01	0.29	0.01	0.01	0.02	0.07	0.05	0.01
90.9	4.08	19.23	61.15	11.71	0.11	0.08	0.43	0.04	2.72	0.01	0.31	0.01	0	0.02	0.06	0.03	0.02

Table 3. Sampel Data Test

Sample_Code	Impeandsi (Ω)
C. CS 3 (20-40)	83.11
C. DT 4 (0-20)	33.2
C. DS 4 (20-40)	35.9

2. Regression Test

The main purpose of conducting regression analysis is to understand the cause-and-effect relationship between one or more independent variables (soil nutrient values: Mg, Al, Si, Fe, S, Cl, K, Ca, Ti, Zn, Zr, Ni, Ga, Ta, V, Cr, and Mn) and the dependent variable (soil impedance). Here are the results of the regression analysis conducted:

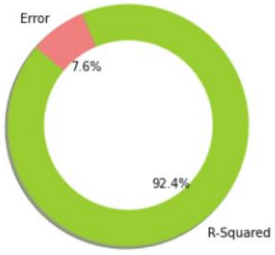

Figure 2. Regression Test Results using Python

The regression equation obtained is :

$$y = 568.41 - 5.84X_1 - 7.38X_2 - 4.13X_3 - 5.92X_4 - 239.26X_5 - 37.33X_6 - 0.97X_7 - 42.00X_8 - 1.70X_9 - 1621.58X_{10} + 226.80X_{11} + 33.87X_{12} + 1313.15X_{13} - 85.69X_{14} - 727.11X_{15} + 52.11X_{16} + 1272.91X_{17}$$

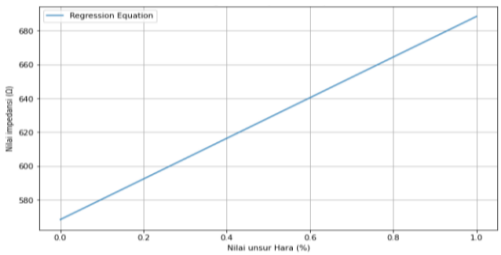

Figure 3. Regression Equation Graph

Based on the regression test results, the R-Squared value (coefficient of determination) obtained is 0.924. This indicates that the regression model built fits the data very well and shows a strong relationship between the dependent and independent variables.

3. Training Data

In this stage, the data will be used to train the model or machine learning algorithm. This data serves as examples that provide information to the model on how it should behave or make predictions. The author splits the training data into 70% training set and 30% testing set. The training data is used to train the classification model, and the testing data is used to test the model's performance with data that has never been "seen" by the model before. The scikit-learn library is used to perform the train-test-split.

Modeling is performed using four algorithms (Linear Model, K-Nearest Neighbors, Decision Tree, and Random Forest) using the scikit-learn library. These algorithms are used to predict nutrient values. In the modeling algorithms used, 10 variations of values are performed with a spacing of 3 digits for each number of neighbors, number of max depth, and number of estimators ranging from 3, 6, 9, 12, 15, 18, 21, 24, 27, 30. These variations are intended to observe the best performance (based on error and accuracy values) of the modeling produced by each algorithm used.

Table 4. Value Variations for n-neighbors, Max Depth, and n-estimators.

Algorithm Model	Value Variations	
Linear Model	-	-
K-Nearest Neighbors	Number of neighbors :	3, 6, 9, 12, 15, 18, 21, 24, 27, 30
Decision Tree	Number of max depth :	3, 6, 9, 12, 15, 18, 21, 24, 27, 30
Random Forest	Number of estimator :	3, 6, 9, 12, 15, 18, 21, 24, 27, 30

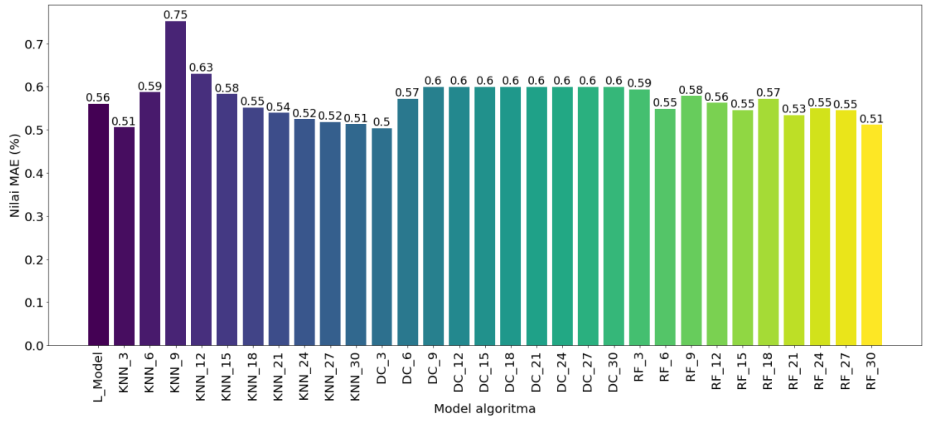

Figure 4. Comparison of Mean Absolute Error (MAE) Values of Algorithm Models

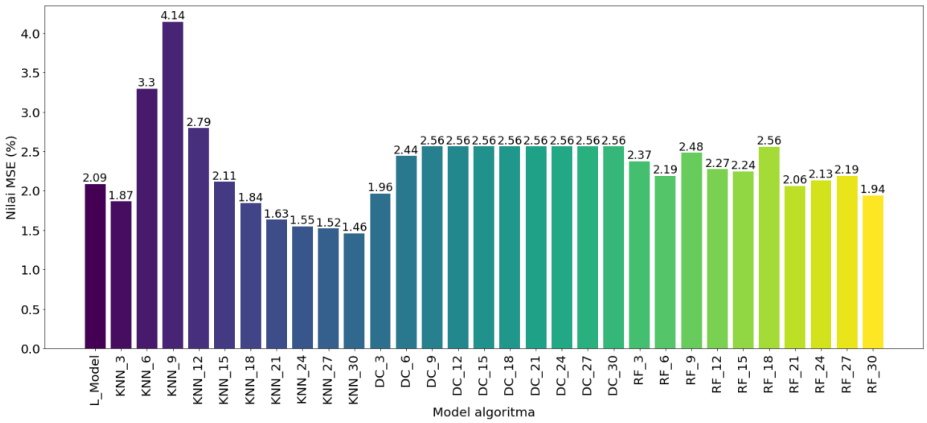

Figure 5. Comparison of Mean Squared Error (MSE) Values of Algorithm Models

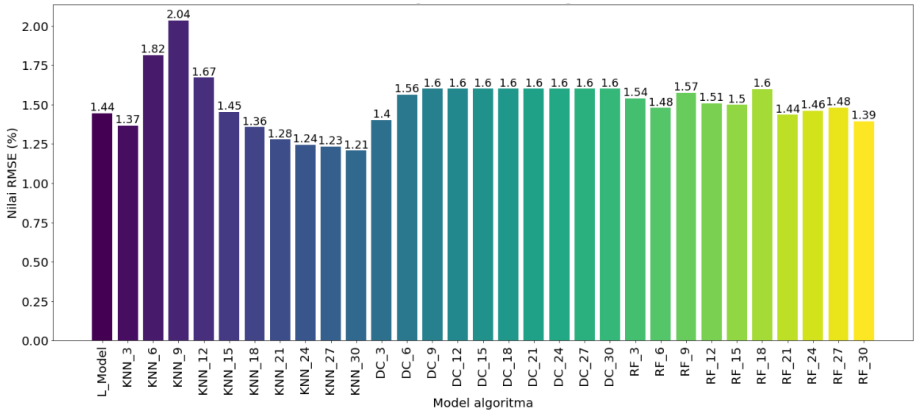

Figure 6. Comparison of Root Mean Squared Error (RMSE) Values of Algorithm Models

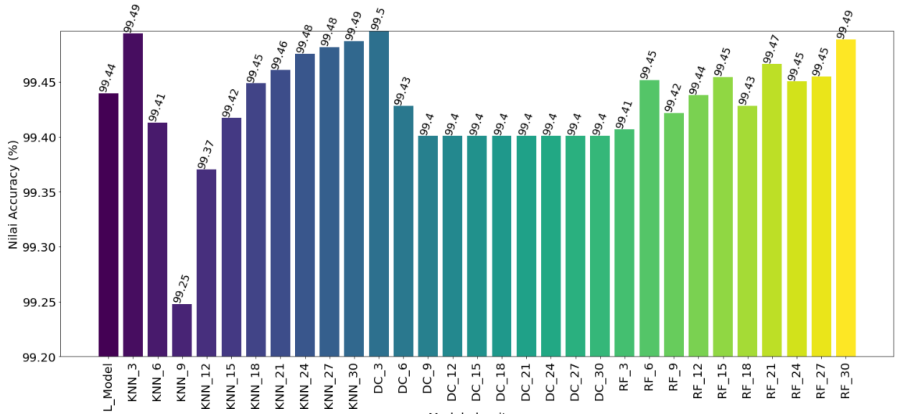

Figure 7. Comparison of Accuracy Score Values of Algorithm Models

Table 5. Ranking of Performance of the Top 10 Model Variations

Model	MAE (%)	MAE Score	MSE (%)	MSE Score	RMSE (%)	RMSE Score	Accuracy (%)	ACC Score (%)	Score Total
KNN_30	0.51	28.00	1.46	31.00	1.21	31.00	99.49	28.00	90.00
KNN_27	0.52	27.00	1.52	30.00	1.23	30.00	99.48	27.00	87.00
KNN_24	0.52	26.00	1.55	29.00	1.24	29.00	99.48	26.00	84.00
KNN_3	0.51	30.00	1.87	26.00	1.37	26.00	99.49	30.00	82.00
KNN_21	0.54	24.00	1.63	28.00	1.28	28.00	99.46	24.00	80.00
RF_30	0.51	29.00	1.94	25.00	1.39	25.00	99.49	29.00	79.00
DC_3	0.50	31.00	1.96	24.00	1.40	24.00	99.50	31.00	79.00
KNN_18	0.55	19.00	1.84	27.00	1.36	27.00	99.45	19.00	73.00
RF_21	0.53	25.00	2.06	23.00	1.44	23.00	99.47	25.00	71.00
L_Model	0.56	18.00	2.09	22.00	1.44	22.00	99.44	18.00	62.00

The top-performing 10 algorithm model variations were selected based on the ranking matrix. These 10 models include Linear Model, K-NN (with the number of n-neighbors being 3, 18, 21, 24, 27, 30), Decision Tree with max depth of 3, and Random Forest (with the number of n-estimators being 21 and 30).

4. Prediction of Soil Nutrient Values

Predictions are made using the 10 best algorithm variations from the training section to compare which algorithm performs better when predicting on new data (test data). Here is the data for which nutrient values will be predicted based on impedance values:

	Sample_Code	Impedance
0	C. CS 3 (20-40)	83.11
1	C. DT 4 (0-20)	33.20
2	C. DS 4 (20-40)	35.90

Figure 8. Data Test

The performance of the models (MAE, MSE, RMSE values, and accuracy) of the algorithms used was measured in the testing section. The models include Linear Model, K-NN (with n-neighbors values of 3, 18, 21, 24, 27, 30), Decision Tree with max depth value of 3, and Random Forest (with n-estimator values of 21 and 30).

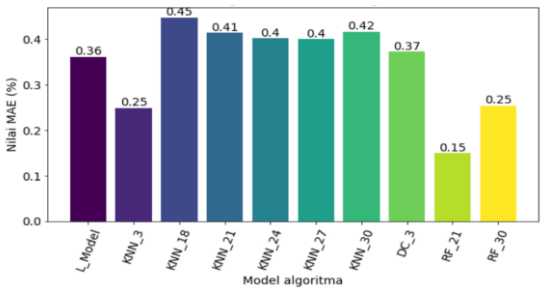

Figure 9. Comparison of MAE Values during Testing Section

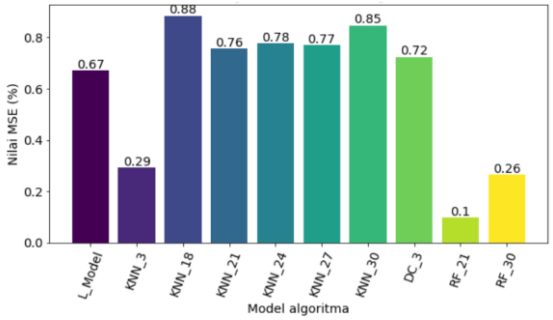

Figure 10. Comparison of MSE Values during Testing Section

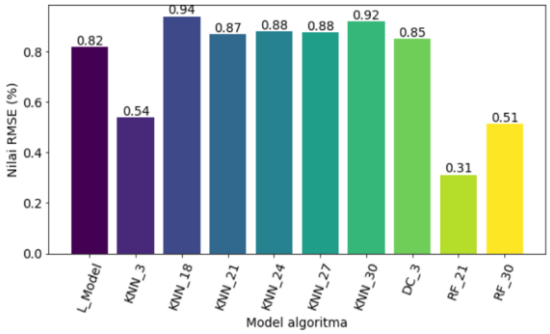

Figure 11. Comparison of RMSE Values during Testing Section

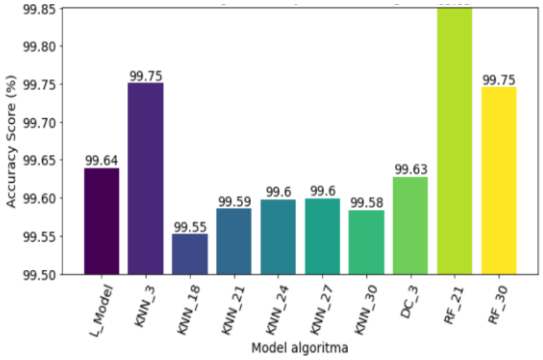

Figure 12. Comparison of Accuracy Values during Testing Section

Based on the performance measurements conducted in the testing section, the algorithm model with the best performance, having the smallest error values and the most accurate predictions, is the Random Forest model (with n-estimator 21) with the following values: MAE = 0.15%, MSE = 0.10%, RMSE = 0.31%, and accuracy = 99.85%.

5. Prediction Results of Ultisol Soil Nutrient Values

Below are the prediction results of soil nutrient values based on soil impedance values using the Random Forest algorithm (with n-estimator 21):

Table 6. Prediction Results of Nutrients Values

Impedance (Ω)	Mg (%)	Al (%)	Si (%)	Fe (%)	S (%)	Cl (%)	K (%)	Ca (%)	Ti (%)	Zn (%)	Zr (%)	Ni (%)	Ga (%)	Ta (%)	V (%)	Cr (%)	Mn (%)
83.11	5.35	20.45	58.68	11.81	0.07	0.06	0.43	0.00	2.56	0.01	0.30	0.01	0.01	0.02	0.06	0.03	0.01
33.2	3.37	22.58	56.28	12.62	0.07	0.03	0.54	1.15	1.01	0.02	0.26	0.01	0.00	0.02	0.08	0.04	0.04
35.9	3.49	23.08	55.39	13.08	0.06	0.03	0.47	0.97	1.88	0.01	0.27	0.01	0.01	0.02	0.08	0.04	0.03

The results provided by the Random Forest algorithm model (with n-estimators=21) in predicting soil nutrient values are excellent. The MAE value of 0.14% indicates a very low average error rate in predicting soil nutrient values. This means that the model has an average error rate of only about 0.14% from the actual values, demonstrating very high accuracy. Additionally, the MSE value of 0.09% and RMSE value of 0.31% also depict excellent prediction quality. The model accuracy of 99.85% is very close to perfection. This implies that the model almost accurately predicts 100% of the soil nutrient values, with a relatively small error rate. The high accuracy level indicates that the Random Forest model with 21 estimators is an effective modeling tool for predicting soil nutrient values based on soil impedance values. The results obtained in this study are consistent with the findings of previous research conducted by Bouslihin et al. in 2021, where the random forest algorithm exhibited the highest accuracy and the smallest error values.With these results, the model can be confidently used in various applications related to determining soil nutrient values.

CONCLUSION AND RECOMENDATIONS

Conclusion

Based on the conducted research, there is a significant correlation between soil nutrient values and soil impedance. Regression analysis yielded an R-squared value of 0.924. This indicates that almost all variations in the dependent variable (soil impedance values) can be explained by variations in the independent variables (soil nutrient values).

Furthermore, comparing machine learning algorithms, the best performance was achieved by the Random Forest model (n-estimator=21) with MAE = 0.14%, MSE = 0.09%, RMSE = 0.31%, and accuracy = 99.85%.

REFERENCES

Ambarwari, A., Adrian, Q.J. and Herdiyeni, Y. 2020. Analysis of the effect of data scaling on the performance of the *machine learning* algorithm for plant identification. *Jurnal RESTI (Rekayasa Sistem And Teknologi Informasi)*. 4(1) : 117-122.

Azzyati, F., Seniari, N.M. and Citarsa, I.B.F. 2019. Analisis Perbandingan Nilai Impeandsi Pentanahan Berdasarkan Panjang Elektroda Grounding Dengan Three Point Metho. *Dielektrika*. 6(1) : 45-54.

Bouslihim, Y., Rochdi, A. and Paaza, N.E.A. 2021. *Machine learning* approaches for the prediction of soil aggregate stability. *Heliyon*. 7(3).

Fashoto, S.G., Mbunge, E., Ogunleye, G. and den Burg, J.V. 2021. Implementation of *machine learning* for predicting maize crop yields using multiple linear regression and backward elimination. *Malaysian Journal of Computing (MJoC)*. 6(1) : 679-697.

Hou, D., Guo, K., and Liu, C. 2020. Asymmetric effects of grazing intensity on macroelements and microelements in grassland soil and plants in Inner Mongolia Grazing alters nutrient dynamics of grasslands. *Ecology and Evolution*. 10(16) : 8916–8926.

Ifadah, N.F., Syarof, Z.N., Al Jauhary, M.R. and Musyaffa, H.J. 2021. *Dasar-Dasar Manajemen Kesuburan Tanah*. Malang : Universitas Brawijaya Press.

Kasno, A. 2019. Perbaikan tanah untuk meningkatkan efektivitas and efisiensi pemupukan berimbang and produktivitas lahan kering masam. *Jurnal Sumberdaya Lahan*. 13(1) : 27-40.

Sang, A.I., Sutoyo, E. and Darmawan, I. 2021. Analisis Data Mining Untuk Klasifikasi Data Kualitas Udara DKI Jakarta Menggunakan Algoritma *Decision Tree* And Support Vector Machine. *Journal of Engineering*. 8(5) : 8954-8963.

Subardja. D.S., S. ritung, M. Anda, Sukarman, E. Suryani, and R.E. Subandiono. 2014. *Petunjuk Teknis Klasifikasi Tanah Nasional*. Bogor : Balai Besar Litbang Sumberdaya Lahan Pertanian

Umaternate, G.R., Abidjulu, J. and Wuntu, A.D. 2014. Uji metode Olsen and Bray dalam menganalisis kandungan fosfat tersedia pada tanah sawah di Desa Konarom Barat Kecamatan Dumoga Utara. *Jurnal MIPA*. 3(1) : 6-10.