ASTEEC Conference Proceeding: Computer Science
3rd International Conference on Information Science and Technology Innovation (ICoSTEC)
July 27, 2024, Yogyakarta, Indonesia

OPTIMISATION OF HYPERPARAMETERS IN REGRESSION ALGORITHM FOR PREDICTIONS OF STUDENT ACADEMIC PERFORMANCE

Muh Nur Aslam[1], Annida Rizki Luthfi Astuti[2], Shofwatul 'Uyun[3]
Magister of Informatics, Faculty of Science and Technology, UIN Sunan Kalijaga, Yogyakarta, Indonesia
[1]23206051031@student.uin-suka.ac.id
[2]23206051021@student.uin-suka.ac.id
[3]Shofwatul.uyun@uin-suka.ac.id

Abstract: **Students' academic achievement is measured by test scores, knowledge, and skills gained from formal education. The importance of identifying potential academic failures motivates this research to find out the factors that affect student academics. This study aims to predict student achievement based on several factors in the internal scope and exam results by using random forest regression, decision tree, and Gradient Boosting methods. The results show that the Ensemble model dominates, with high R-squared values indicating its ability to explain variations in student academic performance and low average MAE, MSE, and RMSE values indicating better performance. The results of the model identify factors that affect variations in student performance based on the tested dataset. This research provides insights for teachers and other stakeholders to improve education by better understanding the factors that influence student academic performance.**

Keywords: student academic performance, machine learning regression, hyperparameter tuning, gridsearchcv.

I. INTRODUCTION

Higher education is essential for the survival of a nation. The number of young people who are highly educated produces quality human resources. Academic achievement is a measure of knowledge gained through formal education and demonstrated through test scores. Goods define student achievement as knowledge and skills acquired and then developed in various school subjects, usually determined by test scores and teacher performance [1]. Early prediction of student performance is necessary to determine quality education, reduce dropout rates, increase graduation rates, and improve educational outcomes [2].

Educators must predict student performance to improve it. Predicting student performance is used to assess student learning and recognize students who excel academically and those who are likely to fail [3].

1. The problems that will be known in this study are: What is the average error in each model?
2. what is the percentage of variation and what factors influence student academic performance?

To predict student academic performance, researchers use machine learning to make predictions that require labels for the data[4]. This research uses gradient-boosting regression, random forest regression, and decision tree regression methods to predict the factors that affect student academic performance. The Gradient Boosting method works by sequentially adding previous predictors that do not match the prediction to the ensemble, to ensure that errors that occurred previously have been corrected [5]. The random forest was chosen because it can improve the accuracy of the results by creating child nodes for each node (the one above) and selecting them randomly [6]. Decision tree regression is a predictive model that can be used to represent classification and regression models in operations, making decisions with the most likely strategy to achieve a goal[7]. This method was chosen because it has the advantage that the results can be described in the form of a decision tree, and it allows direct observation of the results [8].

This research aims to create a student performance prediction model using the Decision Tree, Gradient Boosting, and Random Forest methods. Gradient boost method by applying hyperparameter optimization to get optimal parameters[9]. The hyperparameter optimization used is GridsearchCV.

GridSearchCV is a Python library function that provides certain parameters and implements them into the model to find the optimal parameters [10]. This study uses RMSE, MAE, MSE, and R-squared evaluation metrics that are used to measure the expected difference and evaluation of a model [11].

II. RESEARCH METHOD

The process in our research can be seen in Figure 1, namely the block diagram, this research begins by taking the UCI Student Performance public dataset, then doing preprocessing by doing data cleaning, feature selection, and one hot encoding, after the data are clean. Then the data is split into training and tests after that cross-validation to get the best hyperparameter that is put into the regression algorithm and the results of the algorithm are evaluated on MAE, MSE, RMSE, and R-squared. The following is an explanation of each stage of our research:

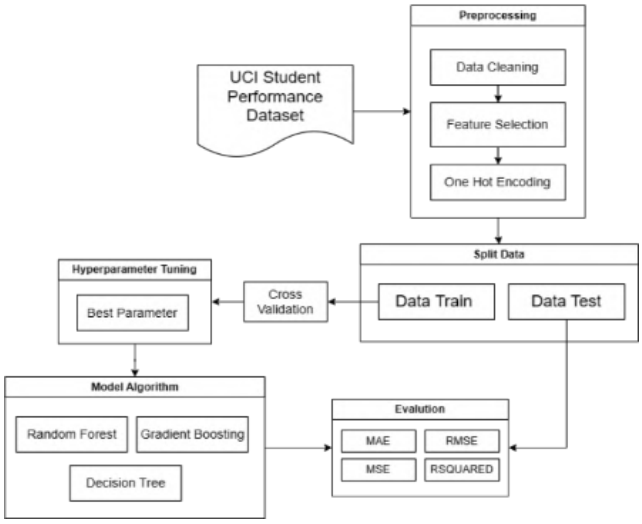

Figure 1 Block diagram

A. *Data Collection*

The dataset is taken from the UCI Student Performance Dataset which has various features, including demographic information, family conditions, student behavior, and student academic performance. Table 1 is a list and type of dataset that provide insight into factors that influence student academic outcomes, including learning patterns and family conditions.

Table 1 Dataset UCI Student Performance Data Set

Name	Mean	Type
School	Student's place in school	Binary
Sex	Student Gender	Binary
Age	Student Age	Numeric
Address	Residential address	Binary
Famsize	Student Family	Binary
PStatus	Living status with parents	Binary
Medu	Mother's Education Level	Numeric
Fedu	Father's Education Level	Numeric
Mjob	Student's Mother's Occupation	Nominal
FJob	Student's Father's Occupation	Nominal
Reason	Reasons for choosing a school	Nominal
Guardian	Student Guardian	Nominal
Traveltime	The journey from home to school	Numeric
Study time	Weekly study time	Numeric
Failures	failed to make the grade	Numeric
Schoolsip	Extra education support	Binary
Famsup	Educational support from family	Binary
Paid	Additional paid classes	Binary
Activities	Extracurricular Activities	Binary
Nursery	have attended kindergarten	Binary
Higher	Want to continue higher education	Binary
Internet	Internet access	Binary
Romantic	Student romantic relationship	Binary
Famrel	Quality of family relationships	Numeric
Freetime	Free time after school	Numeric
Goout	Hangout Time	Numeric
Dalc	Weekday alcohol consumption	Numeric
Walc	Weekend alcohol consumption	Numeric
Health	Health Status	Numeric
Absences	Number of Absences	Numeric
G1	First Period Value	Numeric
G2	Second Period Value	Numeric
G3	Final grade	Numeric

B. *Preprocessing Data*

a) *Data Cleaning*

Data cleaning helps eliminate duplication of missing data [12], ensuring that data used in analysis or processing are more accurate, consistent, and reliable[13]. Data cleaning involves the process of identifying, selecting, and transforming incomplete data, data accuracy, or relevance so that it can be used for further data analysis or processing purposes [13].

b) *Feature Selection*

Feature selection removes irrelevant or redundant features [14] that can improve the accuracy of classification, reduce the difficulty of an algorithm [15], and remove unnecessary and redundant features to reduce the dimension of the feature subspace. This can improve the performance and classification accuracy of the built model [15].

C. *Hyperparameter Tuning*

Hyperparameter tuning works by finding the optimal parameters. To do this, the hyperparameters are tested by trial and error until an optimal value is found[16]. Hyperparameter tuning refers to finding the optimal hyperparameters of an algorithm during the learning process. The hyperparameter method used is GridsearchCV. GridsearchCV is a hyperparameter optimization method that allows us to scan a selected number of hyperparameters. GridsearchCV applies a set of hyperparameter combinations to the model and evaluates the performance of each combination using cross-validation. The combination with the best performance is selected as the optimal hyperparameter for the model [16] and has a low error value [17].

Table 2 Hyperparameter Tuning

Model	Best Param	Value
Gradient Boosting	learning_rate	0.01,0.1,0.2
	max_depth	1, 100
	max_features	Sqrt, log2
	min_samples_leaf	1,2,4
	min_samples_split	2,5,10
	n_estimator	0,50
	subsample	0.8,0.9,0.10
Random Forest	max_depth	1,100
	max_features	Sqrt,log2
	min_samples_leaf	1,2,4
	min_samples_split	2,5,10
Decision Tree	n_estimators	50
	max_depth	1,100
	max_features	Sqrt,log2
	min_samples_leaf	1,2,4
	min_samples_split	2,5,10

From the Hyperparameter Tuning experiment using GridSearchCV, the best parameter information from the performance evaluation on each combination and cross-validation is described in Table 2.

D. Machine Learning

a) *Random Forest*

Random Forest is a machine learning algorithm that is often used to solve problems related to classification and

regression. This algorithm consists of many decision trees that are combined to produce more accurate predictions[3]. Random forest is a method that is sensitive to hyperparameter values and can significantly improve accuracy and prediction [17].

b) Gradient Boosting

Gradient-boosting regression (GBR) is a machine learning technique that is also frequently used to create accurate prediction models [18]. GBR is an ensemble learning method that combines predictions from several basic estimators, usually decision trees, to improve the accuracy and robustness of the model [19].

c) Decision Tree

Decision Tree is one of the machine learning models that uses a treelike structure to make decisions [20]. Decision Tree also has a powerful and easy-to-understand structure [21].

E. Metric Evaluation

a) Mean Absolute Error (MAE)

The mean absolute error (MAE) is an evaluation metric calculated by taking the difference between the predicted value and the actual value, then taking the absolute value of the difference and then taking the average of all the absolute values. The lower the MAE value, the better the performance[22].

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | Y_i - \hat{Y}_i |$$

b) RMSE (Root Mean Square Error)

Root mean square error (RMSE) is a metric used to measure how much error there is between the predicted value and the true value. RMSE has an advantage over MAE in describing the distribution of errors [23].

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(\hat{r}_n - r_n)^2}{N}}$$

c) Mean Square Error (MSE)

The mean square error (MSE) is an evaluation metric that is used to measure how close the model predictions are to the actual values in the data set. A smaller MSE value indicates better model performance in predicting actual values[22].

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}(X_{obsi} - X_{model,i})^2$$

d) R-Squared

R squared is an evaluation metric used to evaluate the degree to which the regression model fits the observed data. The R-squared values range from 0 to 1, where higher values indicate that the variability in the data can be better explained by the regression model[22].

$$R - Squared = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

III. Result and Discussion

Comparison of RandomForest, Gradient Boosting, and Decision Tree regression algorithm models to predict student learning performance taken from the Kaggle Student Performance Kaggle dataset. The average error for each model is obtained from parameter evaluation using MAE (mean absolute error), MSE (Mean Squared Error), and RMSE (root mean squared error) to obtain the average error value for each model. To answer the first question, what is the average error for each model, we will look at the comparison metrics in Table 3 below:

Table 3 Table MAE, MSE, RMSE

Model	MAE	MSE	RMSE
Random Forest	1.20	2.73	1.65
Gradient Boosting	1.28	3.49	1.87
Decision Tree	2.56	14.11	3.76

Table 3 shows that the Random Forest model is the best in measuring how close the model prediction is to the actual value because the Random Forest model has the lowest error rate. Where the average error of the value is average (MAE) 1.20%, the average square error value (MSE) is 2.73%, and the root mean square error value (RMSE) is 1.65%.

Then Gradient Boosting shows a slightly higher error than Random Forest and Decision Tree which shows the highest error among the three models.

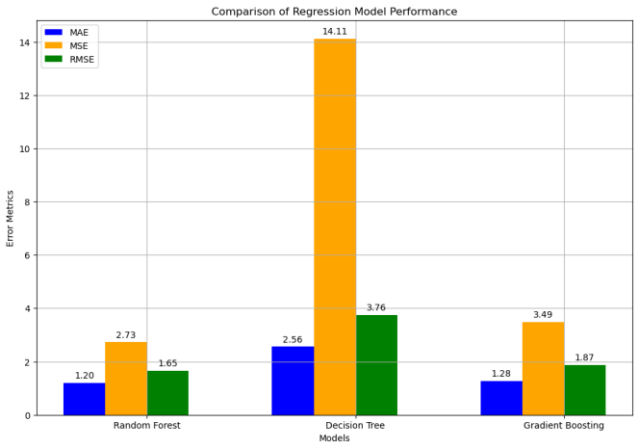

Figure 2 MAE, MSE, RMSE, Comparison

Figure 2 shows a comparative visualization of the three regression models in identifying the average error MAE, MSE, and RMSE values.

The next question is what percentage of variation and what factors affect student academic performance?

Table 4 R-squared comparison

Model	R Squared
Random Forest	0.867
Gradient Boosting	0.837
Decision Tree	0.312

table 4 shows the results of the R-squared formula 3.4 which shows the Random Forest model gets a high value close to 1 from the other three models such as Random Forest with a value of 0.867. this model explains the excellent prediction performance with high R-squared, the ensemble nature of random forest which combines predictions from many decision trees that can capture complex patterns in the data.

Table 5 Top 10 Feature Importance

Failure	Random Forest	Decision Tree	Gradient Boosting
G3	0.566	0.418	0.409
Absences	0.065	0.207	0.110
Failures	0.046	0.020	0.077
Age	0.026	0.068	0.021
Freetime	0.023	0.000	0.013
Medu	0.022	0.010	0.043
Goout	0.022	0.007	0.039
Health	0.021	0.021	0.042
Study time	0.019	0.025	0.0.23
Fedu	0.018	0.063	0.017

The table above is based on the weights of the top 10 features in evaluating the contribution of key features in three different models. from the table it can be seen that Random Forest can handle less important features more efficiently as it relies on key features such as G3, Decision Tree provides interpretable results and shows clarity in the importance of features. however, it tends to give very high weights to some features. Gradient Boosting offers a more balanced approach and can capture feature interactions better making it a robust model for different types of data although it is more complex and requires a lot of time for training. then we will compare the correlation matrix between each model

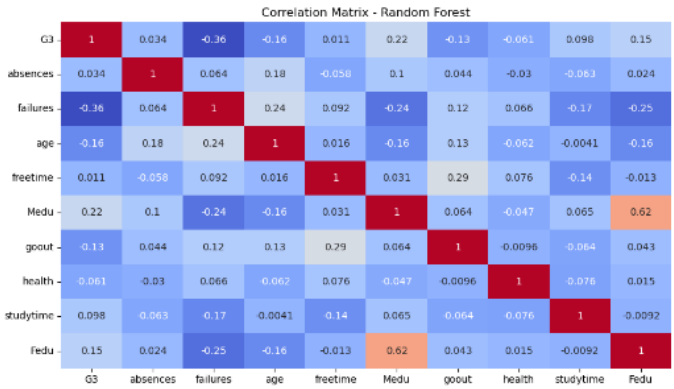

Figure 3 Correlation Matrix Random Forest

Random Forest can handle non-linear features and complex interactions between features. Figure 3 some examples show G3 has a strong negative correlation with failures which means students with more failures tend to have lower final grades, a positive correlation between Medu and G3 indicates mother's education plays a role in students' academic performance and absenteeism shows a low correlation so absenteeism does not affect the final grade.

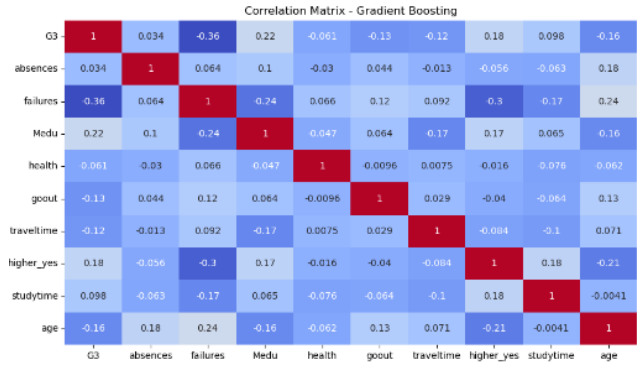

Figure 4 Correlation Matrix Gradient Boosting

Figure 6 explains the correlation of the Gradient Boosting performs well also because of its ability to handle non-linear relationships iteratively. Just like Random Forest, G3 features have a strong negative correlation with failure, and a positive correlation between Medu and G3, and a low correlation between absences and G3 which indicates that it does not influence the prediction of the final grade too much.

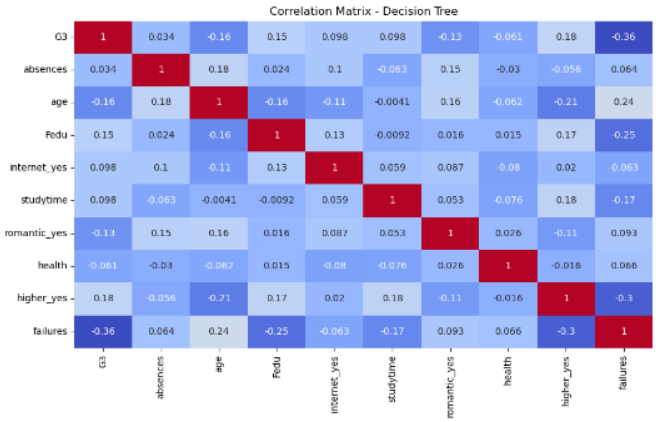

Figure 5 Correlation Matrix Decision Tree

Figure 7 shows the correlation of the decision tree model also has a negative correlation with failure just like Random Forest and Gradient Boosting, with a positive correlation between G3 and higher_yes indicating that students who intend to pursue higher education tend to have better final grades and a low correlation between absences and G3 indicating absenteeism does not affect final grades in the model.

From the correlation matrix explanation, Random Forest can handle less influential variables such as absences better and the Gradient Boosting Model also provides more consistent and accurate results because it can handle non-linear relationships iteratively while the Decision Tree is more prone to overfitting, seen from the unbalanced correlation with some variables. Random Forest and Gradient Boosting models tend to provide more stable and generalizable results compared to decision trees due to the consistent correlation factors between G3 and

failures, as well as between Medu and Fedu in predicting students' academic performance.

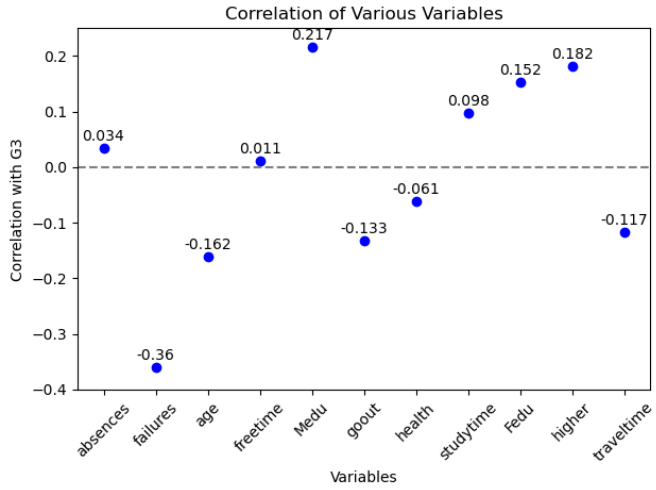

Figure 6 Correlation of Variubs Variables

The correlation results between various variables and students' final grades (G3) show that the mother's education (Medu) has the highest positive correlation (0.217), followed by intention to pursue higher education (Higher, 0.182) and the father's education (Fedu, 0.152), indicating that parents' educational background and students' motivation to pursue higher education are associated with better final grades. Study time (Studytime, 0.098) also showed a positive correlation, albeit a lower one, while absences (Absences, 0.034) and free time (Freetime, 0.011) showed very low positive correlations. In contrast, the number of previous failures (Failures, -0.36) has a strong negative correlation, suggesting that more failures are associated with lower final scores. Age (Age, -0.162), going out, -0.133, and travel time (Traveltime, -0.117) also show negative correlations, which may indicate the negative impact these factors have on students' final grades. Health (Health, -0.061) showed a low negative correlation, indicating that health may not significantly affect academic performance. Overall, parental education and motivation to pursue higher education seem to be important factors in determining students' final grades.

IV. Conclusions

From the analysis of student academic performance using hyperparameters from three regression models, namely Random Forest Regression, Decision Tree Regression, and Gradient Boosting Regression, it is concluded that the ensemble models of Gradient Boosting Regression and Random Forest Regression show better performance in predicting student final grades. This can be seen from the lower Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) values compared to the other two models. The ensemble model can capture the complexity of the data more effectively, providing more accurate and stable predictions in the context of student academic performance.

Acknowledgment

The author would like to thank the following parties for their contributions the following parties for their contributions to this work, UIN Sunan Kalijaga Yogyakarta for the support provided. The authors are also grateful to all participants for their time and cooperation.

References

[1] H. Hasanah, A. Farida, and P. P. Yoga, "Implementation of Simple Linear Regression for Predicting Students' Academic Performance in Mathematics," *J. Pendidik. Mat.*, vol. 5, no. 1, p. 38, 2022, doi: 10.21043/jpmk.v5i1.14430.

[2] I. El Guabassi, Z. Bousalem, R. Marah, and A. Qazdar, "Comparative Analysis of Supervised Machine Learning Algorithms to Build a Predictive Model for Evaluating Students' Performance," *Int. J. online Biomed. Eng.*, vol. 17, no. 2, pp. 90–105, 2021, doi: 10.3991/ijoe.v17i02.20025.

[3] M. R. Apriyadi, Ermatita, and D. P. Rini, "Hyperparameter Optimization of Support Vector Regression Algorithm using Metaheuristic Algorithm for Student Performance Prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, pp. 144–150, 2023, doi: 10.14569/IJACSA.2023.0140218.

[4] I. A. Firdaus, "Deteksi Infeksi Mycoplasma Pneumoniae Pneumonia Menggunakan Komparasi Algoritma Klasifikasi Machine Learning," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 7, no. 1, p. 35, 2022, doi: 10.31328/jointecs.v7i1.3242.

[5] S. E. Suryana, B. Warsito, and S. Suparti, "Penerapan Gradient Boosting Dengan Hyperopt Untuk Memprediksi Keberhasilan Telemarketing Bank," *J. Gaussian*, vol. 10, no. 4, pp. 617–623, 2021, doi: 10.14710/j.gauss.v10i4.31335.

[6] G. A. Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest," *CogITo Smart J.*, vol. 6, no. 2, pp. 167–178, 2020, doi: 10.31154/cogito.v6i2.270.167-178.

[7] Patlisan and Rusdah, "OPTIMASI AKURASI MODEL DECISION TREE MENGGUNAKAN RANDOM FOREST REGRESSION UNTUK PREDIKSI KUANTITAS," vol. 14, no. 2, pp. 217–228, 2023.

[8] D. Nike Aria Kurniawan, "Implementasi Metode Decision Tree pada Sistem Prediksi Status Gizi Balita," *J. Sains Komput. Inform. (J-SAKTI*, vol. 7, no. 2, pp. 731–739, 2023.

[9] U. L. Yuhana, A. Purwarianti, and I. Imamah, "Tuning Hyperparameter pada Gradient Boosting untuk Klasifikasi Soal Cerita Otomatis," *J. Edukasi dan Penelit. Inform.*, vol. 8, no. 1, p. 134, 2022, doi: 10.26418/jp.v8i1.50506.

[10] Andriana *et al.*, "Prediksi Gelombang Corona Dengan Metode Neural Network," *JIKOMSI (Jurnal Ilmu Komput. dan Sist. Inf.*, vol. 3, no. 2, pp. 102–107, 2020.

[11] K. Das, R. Kumar, and A. Krishna, "Analyzing electric vehicle battery health performance using supervised machine learning," *Renewable and Sustainable Energy Reviews*, vol. 189. 2024. doi: 10.1016/j.rser.2023.113967.

[12] V. Tziovara, P. Vassiliadis, and A. Simitsis, "Deciding the physical implementation of ETL workflows," *Dol. Proc. ACM Int. Work. Data Warehous. Ol.*, pp. 49–56, 2007, doi: 10.1145/1317331.1317341.

[13] F. Zou, "Research on data cleaning in big data environment," *Proc. - 2022 Int. Conf. Cloud Comput. Big Data Internet Things, 3CBIT 2022*, pp. 145–148, 2022, doi: 10.1109/3CBIT57391.2022.00037.

[14] H. SabbaghGol, H. Saadatfar, and M. Khazaiepoor, "Evolution of the random subset feature selection algorithm for classification problem," *Knowledge-Based Syst.*, vol. 285, no. December 2023, p. 111352, 2024, doi: 10.1016/j.knosys.2023.111352.

[15] D. Doreswamy and M. Nigus, "Feature Selection Methods for Household Food Insecurity Classification," *2020 Int. Conf. Comput. Sci. Eng. Appl. ICCSEA 2020*, 2020, doi: 10.1109/ICCSEA49143.2020.9132945.

[16] I. Muhamad Malik Matin, "Hyperparameter Tuning Menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware," *Multinetics*, vol. 9, no. 1, pp. 43–50, 2023, doi:

10.32722/multinetics.v9i1.5578.
[17] A. Baita, I. A. Prasetyo, and N. Cahyono, “Hyperparameter Tuning on Random Forest for Diagnose Covid-19,” *JIKO (Jurnal Inform. dan Komputer)*, vol. 6, no. 2, pp. 138–143, 2023, doi: 10.33387/jiko.v6i2.6389.
[18] M. I. Khan and Y. M. Abbas, “Robust extreme gradient boosting regression model for compressive strength prediction of blast furnace slag and fly ash concrete,” *Mater. Today Commun.*, vol. 35, no. March, p. 105793, 2023, doi: 10.1016/j.mtcomm.2023.105793.
[19] J. Cai, K. Xu, Y. Zhu, F. Hu, and L. Li, “Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest,” *Appl. Energy*, vol. 262, no. November 2019, p. 114566, 2020, doi: 10.1016/j.apenergy.2020.114566.
[20] E. Novianto, A. Hermawan, and D. Avianto, “Klasifikasi algoritma k-nearest neighbor, naive bayes, decision tree untuk prediksi status kelulusan mahasiswa s1 1) 1,2,3),” vol. 8, no. 2, pp. 146–154, 2023.
[21] J. Pal, “Decision Tree Psychological Risk Assessment in Currency Trading,” 2023.
[22] A. G. Priya Varshini, K. Anitha Kumari, D. Janani, and S. Soundariya, “Comparative analysis of Machine learning and Deep learning algorithms for Software Effort Estimation,” *J. Phys. Conf. Ser.*, vol. 1767, no. 1, 2021, doi: 10.1088/1742-6596/1767/1/012019.
[23] D. S. K. Karunasingha, “Root mean square error or mean absolute error? Use their ratio as well,” *Inf. Sci. (Ny).*, vol. 585, pp. 609–629, 2022, doi: 10.1016/j.ins.2021.11.036.