Mantik Journal, 7 (4) (2024), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

Modification of random forest method to predict student graduation data

Christine Dewi[1], Gerald Edgard Laukon[2], Henoch Juli Christanto[3], Stephen Aprius Sutresno[4]

[1,2]Department of Information Technology, Satya Wacana Christian University, Salatiga, Indonesia

[3,4] Department of Information System, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia

A R T I C L E I N F O

Article history:

Received Dec 08, 2023
Revised Dec 14, 2023
Accepted Dec 24, 2023

Keywords:

Machine learning;
Prediction;
Rendom forest
Student performance.

ABSTRACT

The graduation rate of students is an important measure of a school's success, as it indicates the school's ability to help students complete their education. Predicting student completion is crucial for schools to identify at-risk students and offer them early interventions to improve their academic performance. This can also assist policymakers in developing effective policies and programs to enhance graduation rates and reduce dropout rates. The dataset used in this study was obtained from the Kaggle website, and the best model proposed utilizes the Random Forest method with hyperparameter tuning. By adjusting the n_estimator parameter to 1000, our proposed method decreases the mean squared error (MSE) value from 0.5525155 to 0.5374983 and increases the R2 Score value from 0.9984039 to 0.9984873. The study also compares the performance of the proposed model with other datasets sourced from the University of California Irvine (UCI), demonstrating superior performance across all experiments. The results consistently show a decreasing trend in MSE value and an increasing trend in R2 value for all datasets.

Corresponding Author:

Henoch Juli Christanto,
Department of Information System,
Atma Jaya Catholic University of Indonesia,
Jl. Jenderal Sudirman No. 51, Jakarta,10220, Indonesia.
Email: henoch.christanto@atmajaya.ac.id

1. INTRODUCTION

Education today is one of the most important and fundamental aspects of society. Its function is to advance the development of individuals in both academic and financial terms. A person who has received an education should be able to make beneficial contributions not only to his or her own family but also to society and the community. All of this can be achieved through a proper learning process (Purwaningsih & Nurelasari, 2021). In creating skilled, knowledgeable, competitive, and innovative human resources, educational institutions are expected to organize a superior and quality education process for their students. Therefore, to realize this, it is necessary to implement school exams or national exams which aim to assess graduation at each level of education and become a quality standard in educational institutions

(Christanto, Sutresno, Denny, & Dewi, 2023). Participants can be considered graduates if they have completed their study period, obtained positive grades in the aspects of attitude or behaviours with a minimum of good categories, and took exams held by authorized agencies or education units. To assess exam results and student graduation, it is important to evaluate the quality as well as areas where students may face difficulties in the learning process Pandey & Taruna (2016) Therefore, we strive to always support students who are struggling with learning in the education unit/program (Altujjar et al., 2016).

Student achievement, development, and potential play a crucial role in evaluating learning outcomes, determining appropriate learning materials, and designing appropriate learning activities. Unfortunately, current work does not provide adequate analytical tools to assess student performance, identify factors that influence their performance, comprehend how students can advance and evaluate if they have the capacity for improvement Yang & Li (2018). Many factors contribute to a student's pass rate. One of the main factors is daily grades, which have an impact on exam results. However, in projecting a student's graduation progress, other factors also play an important role, including behaviors and discipline levels, which also influence the outcome. An increase in the student pass rate within an educational unit/program may increase the reputation and popularity of the unit/program (Altujjar et al., 2016). Conversely, if the pass rate within an educational unit/program decrease, this may have an impact on the attractiveness of potential new students who wish to join the unit/program. Therefore, it may cause concern for the education program" (Sianturi, 2018).

Since the importance of predicting student graduation is great, it is necessary to forecast student graduation rates for the next period in order to maintain and improve the success rate of students in completing their education. Forecasting is a technique that can be used to project future student graduation results. The application of machine learning can be utilized to forecast student graduation by using historical data as material to train the model. This model is created to accurately forecast students' future graduation outcomes. One technique that is often used in machine learning is Random Forest, which has various applications in tasks such as classification, regression, and many other related tasks. It belongs to the group of ensembles learning techniques, which combine multiple decision trees to improve prediction accuracy. In addition, Random Forest is a powerful machine learning algorithm that has a wide range of applications, such as forecasting stock prices, predicting student graduation, and performing disease diagnosis. The main points that can be drawn from this research are as follows: (1) Assessing the Mean Squared Error (MSE) levels from prior research involving various machine learning models such as Linear Regression, Decision Tree, Ridge Regression, Lasso Regression, KNN, and Random Forest. (2) Apply the Random Forest approach to forecast student graduation. Random Forest combines a number of decision trees by training the model on existing data. This approach is often used because it has a low error rate and provides high-accuracy results Louk & Tama (2022). (3) Perform hyperparameter tuning to find the most effective value in the Random Forest method. (4) Comparing the error rate and accuracy results of the hyperparameter tuning experiments that have been run on the Random Forest method.

Forecasting, Forecasting, often referred to as forecasting, is the result of projecting or predicting future values based on historical data (Sinaga et al., 2018). According to John E. Biegel (1999), forecasting is the act of projecting the anticipated level of demand for one or more products over a specified period of time in the future (Jungmeier, 2017). According to Buffa S. Elwood (1996), forecasting is the application of statistical techniques to create a picture of the future based on analysis of historical data (Putra et al., 2020). Prediction or forecasting can also be explained as the

systematic step of making estimates of what is likely to happen in the future, based on current information, with the aim of reducing the error between what is predicted and reality (Herdianto, 2013). But this does not mean that after understanding these techniques, we can predict everything with precision, but rather that we only understand certain techniques that are relevant for certain situations(Christanto, Sutresno, Simi, Dewi, & Dai, 2023). When selecting a forecasting approach, it is crucial to take into account the nature of the data pattern and opt for the method that best fits that pattern (Hosoe et al., 2021). In addition, prediction is the act of projecting the future value of a variable based on past data and relevant factors. It is a crucial element of decision-making in various industries, such as finance, marketing, and operations management Thakkar & Chaudhari (2021). The application of forecasting also covers a wide range of applications, such as sales forecasting, demand estimation, and stock management (Ren et al., 2020).

Random Forest, In 2001, Breiman introduced the Random Forest algorithm, which was able to address two types of problems, namely classification and regression (Hidayat et al., 2023). Random Forest is an evolution of the Classification and Regression Tree (CART) method, which utilizes bagging or bootstrap aggregation techniques and random feature selection (N. K. Dewi et al., 2012). The random forest algorithm process involves two stages: in the first stage, "k" trees are created to form a random forest, while in the second stage, this random forest is used to generate predictions (Suliztia, 2020). The Random Forest method can also improve accuracy by randomly generating child nodes for each parent node. This model, which is based on a tree structure, operates by iteratively classifying the initial dataset into two subgroups using specific criteria, until it reaches a predefined stopping condition, as illustrated in Figure 1 Schonlau & Zou, (2020). This method entails creating a decision tree structure comprising a root node, internal nodes, and leaf nodes through the random selection of attributes and data following relevant guidelines. The topmost node of this decision tree, also known as the root, is located at the top of the tree (Sarker et al., 2020). Internal nodes refer to nodes that are divided into branches, have one input, and at least two outputs. On the other hand, a leaf node, also called a terminal node, is the last node that receives a single input and produces no outputs" (Adebowale et al., 2013).

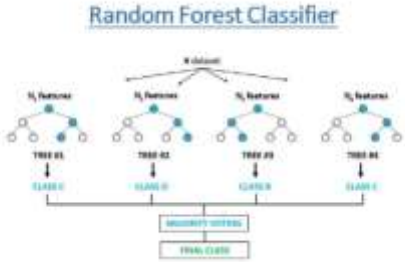


Figure 1: Random Forest Architecture

Predictions in the regression case are based on the average value of each tree. Formula (1) is used to calculate the average value of all tree predictions.

$$\hat{Y}_i = \frac{1}{N_{tree}} \sum_{n=1}^{N_{tree}} \hat{Y}_n \quad (1)$$

Description:
$\hat{Y}_i$ = Final prediction results

N_{tree} = The overall quantity of trees within the Random Forest
$\hat{Y}_n$ = n-tree prediction results

In Random Forest, the parameter "n_estimators" specifies the quantity of decision trees utilized in constructing the model. The more estimators used, the higher the accuracy of the model, although this may require more computational time and memory usage (Jun, 2021). The addition of estimators usually results in a model that better fits the training data as each decision tree is trained with a different subset of data and features. However, there is a limit where increasing the number of estimators may provide little benefit or even lead to overfitting (Tian et al., 2020).

The optimal value of the hyperparameter n_estimators depend on the dataset used and the specific task. Usually, a standard practice involves commencing with a modest number of estimators and progressively raising it until the model attains a desired level of performance during validation or testing, or until performance starts to decline (Probst et al., 2019). Cross-validation techniques can also be used to adjust the value of n_estimators as well as other hyperparameters. Overall, n_estimator is a very important hyperparameter in Random Forest that affects both the performance and complexity of the model (Contreras et al., 2021). In the context of this experiment, selecting an appropriate value for this hyperparameter is a key step in building a precise and efficient Random Forest model.

Mean Squared Error (MSE), MSE (Mean Square Error) is a metric that measures the average of squared errors. In this error calculation, larger values will be penalized more than smaller values, as the calculation involves squaring. MSE is a different evaluation method to measure the performance of forecasting methods. In this method, individual errors are taken, squared, and then calculated. This approach places more emphasis on large errors because they are squared. The outcome is the mean of the squared disparity between the predicted value and the actual value. One criticism of using MSE is that it tends to place greater emphasis on large differences due to the squaring process (Azmi et al., 2020). This can be seen in the following equation Lusiana & Yuliarty, (2020).

$$MSE = \sum \frac{(A_t - F_t)^2}{n} \quad (2)$$

Description :
A_t : Realized need during period-t
F_t : Predicted demand for period-t
n : Quantity of demand intervals included

1.4. R-2 Squared (R2)

The R-squared or R2 is a mathematical metric that quantifies the degree to which changes in the dependent variable can be clarified by the independent variables incorporated in the regression model. Frequently termed as the coefficient of determination, it signifies how closely the values forecasted by the regression model align with the actual data points situated on the regression line. The accuracy of the equation that has been constructed is assessed through a comparison between the value predicted by the model and the actual observed value (Cominotte et al., 2020). The R2 value spans from 0 to 1, where 0 signifies that the model doesn't align well with the provided data, and 1 signifies a perfect fit between the model and the data. The calculation of R2 is done with the following formula (Afzal et al., 2021):

$$R = \sqrt{1 - \frac{\sum_{x=1}^{n} (V_{ot} - V_{op})^2}{\sum_{x=1}^{n} (V_{ot(mean)} - V_{op})^2}} \quad (3)$$

Formula (3) above uses the symbol n to represent the number of data points, while V_{ot} and V_{op} refer to the expected estimates derived from the regression model and the actual measurements of the output.

2. RESEARCH METHOD

2.1. Research Design

This section will provide an explanation of the general framework within which this research was conducted. The flowchart of this research will be illustrated in Figure 2, which depicts a number of steps involving a series of experiments. The first step is the collection of a dataset from the website www.kaggle.com. This dataset encompasses information on the mathematics performance of high school students, encompassing their grades and demographic details. The data was gathered from three high schools situated in the United States. The data has eight attributes consisting of gender, race/ethnicity, parental education level, lunch, test preparation courses, math scores, reading scores, and writing scores with a total of 1000 data (Kaggle - Rattanaporn, 2023). Next, the modelling process involves implementing the Random Forest technique. Random Forest is a supervised machine learning algorithm that employs the concept of iteratively utilizing decision trees to create an ensemble or set of models. This algorithm combines predictions from various decision trees (Ganatra, 2020). Previously, there have been experimental efforts using various methods such as Linear Regression, Decision Tree, Ridge Regression, Lasso Regression, KNN, SVR, and Random Forest with similar datasets. The results of these experiments show that the Random Forest method produces the highest level of accuracy and has the lowest error rate. Third, the hyperparameter tuning stage is performed on the Random Forest algorithm. This tuning process involves finding the optimal hyperparameters for the learning algorithm on a specific dataset (Amusa et al., 2021). Hyperparameters may include determining the quantity of trees in the forest or the maximum number of nodes per tree. This requires iterative testing and adjustment to find the optimal configuration (Siji George & Sumathi, 2020). In the pursuit of the most suitable hyperparameter settings, a range of parameter combinations are empirically examined. Ultimately, the outcomes of this hyperparameter tuning process are assessed, and the best parameters are chosen based on the attained level of accuracy.

Figure 2: The Research Workflow

2.2. Dataset

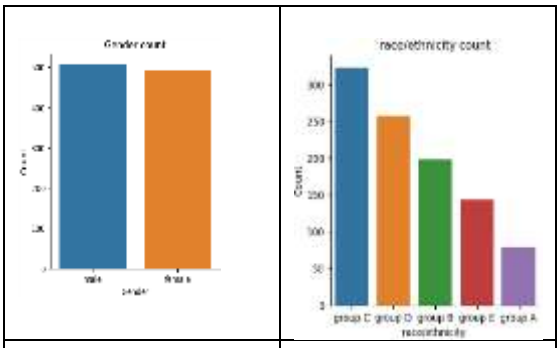
This research evaluates the performance of learning algorithms using the Student Performance Prediction dataset. The dataset was sourced from Kaggle and comprises information gathered from three American high schools. The dataset used consists of 1000 data and 8 features. An example of the Student Performance Prediction dataset used can be seen in Table 1.

Table 1*:* Student Performance Prediction Dataset

Gender	Race	Parental level of education	Lunch	Test Preparation Course	Math Score	Reading Score	Writing Score
Female	Group	Some	Standard	Completed	59	70	78

Male	Group D	College Associate's Degree	Standard	None	96	93	87
Female	Group D	Some College	Free/Reduced	None	57	76	77
Male	Group D	Some College	Free/Reduced	None	70	70	63
Female	Group B	College Associate's Degree	Standard	None	83	85	86

Columns from Table 1 can be explained as follows: (1). Gender : The student's gender (male/female). (2) Race/ethnicity : The student's cultural or ethnic heritage (Asian, African-American, Hispanic, etc.) (3) Parental level of education : The highest educational achievement reached by the student's parent(s) or guardian(s). (4.) Lunch : Does the student qualify for free or discounted lunch (yes/no)? (5). Test Preparation Course : Did the student finish a test preparation course (yes/no)? (6). Math Score : The student's result on a standardized math assessment. (7) Reading Score : The student's performance on a standardized reading assessment. (8) Writing Score : The student's result on a standardized writing assessment.



(a) Gender Count **(b)** Race/Ethnicity Count

Figure 3: Gender and Race/Ethinicity Count

Figure 3 (a) presents a count of the genders, which reveals that there are 508 male records and 492 female records among the collected data. In addition, the count of the races is presented in Figure 3(b), which shows that Group C has 323, Group D has 257, Group B has 198, Group E has 143, and Group A has 79.

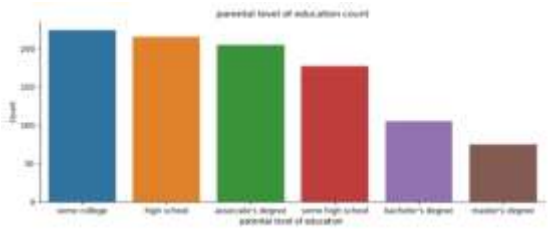
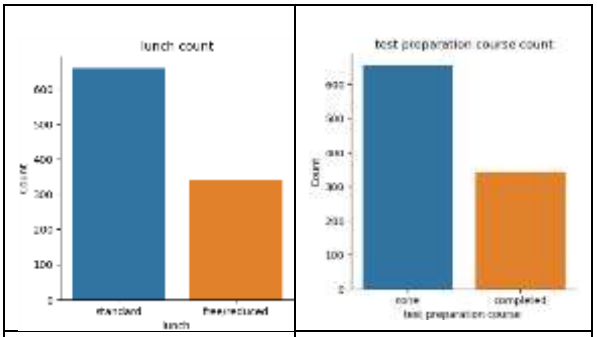


Figure 4: Parental Level of Education Count

Figure 4, which shows that Some College has 224, High School has 215, Associate's Degree has 204, Some High School has 177, Bachelor's Degree has 105, and Master's Degree has 75.



(a) Lunch Count **(b)** Test Preparation Count

Figure 5: Lunch and Test Preparation Count

Figure 5 (a) presents the lunch count, which shows that there are 660 standard and 340 free/reduced. Then, in figure 5(b) the number of test preparations, which shows that 656 were missing and 344 were completed.

3. RESULTS AND DISCUSSIONS

At this stage, the results of the experiments conducted are described. This study assesses the effectiveness of the learning algorithm utilizing the Student Performance Prediction dataset. The first experiment was conducted by comparing the error rate of previous research with several other methods. The model used in previous research by Hdevlet was Linear Regression with an MSE value of 1.38 (Kaggle - Rattanaporn, 2023).

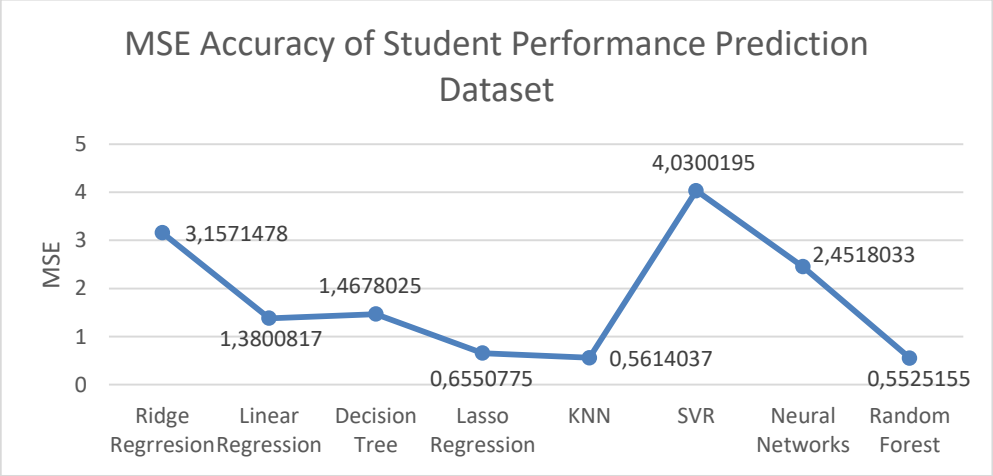


Figure 6: MSE results using Student Performance Prediction Dataset

Figure 6 shows the error rate results of the Student Performance Prediction dataset using various machine learning methods. The experimental results presented in Figure 6 use several machine learning methods such as Ridge Regression, Linear Regression, Decision Tree, Lasso Regression, K-Nearest Neighbors (KNN), Support Vector Regression (SVR) and Random Forest. Based on the experimental results, the smallest error value is obtained using the Random Forest method with an MSE value of 0.5525155. In conclusion, it can be concluded that the Random Forest method

outperforms other methods, based on the experimental results obtained and previous studies using the same dataset.

The next experiment is adding/replacing parameters. The methods to be compared are the same as the previous experiment, namely Ridge Regression, Linear Regression, Decision Tree, Lasso Regression, K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Neural Networks and Random Forest. The values to be compared are the accuracy value using R2 Score and the error value using MSE. The accuracy and error rate results are shown in Table 2.

Table 2: Accuracy and MSE results after parameter addition/replacement.

Method	MSE	R2 Score
Ridge Regression	2,9530154	0,9999999
Linear Regression	1,0132924	1,0
Lasso Regression	0,6377381	0,9999748
Decision Tree	1,3619781	0,9893406
KNN	0,5766570	0,9983844
SVR	4,0300195	0,8923635
Neural Networks	2,4518033	0,9783421
Random Forest	0,5525155	0,9984039

From the experimental results presented in Table 2, it can be concluded that Random Forest is the optimal method with the smallest MSE value of 0.5525155 and R2 Score value of 0.9984039. Therefore, it was decided in this study to model using the Random Forest method with the Student Performance Prediction dataset that had changed its parameters. After forming a model using the Random Forest method, the next experiment is to adjust the Random Forest hyperparameter value to improve accuracy results and reduce the error rate. The hyperparameter to be adjusted is the n_estimator parameter. In addition, n_estimator is the number of trees with increasing gradient (Ganatra, 2020). This experiment will try to change the value of the n_estimator parameter, and the results will be shown in Table 3.

Table 3: Evaluation of hyperparameter tuning using Student Performance Prediction Dataset

Method	MSE	R2 Score
Random Forest	0,5525155	0,9984039
Random Forest (n_estimator = 200)	0,5630736	0,9983352
Random Forest (n_estimator = 500)	0,5463999	0,9984363
Random Forest (n_estimator = 1500)	0,5464604	0,9983942
Random Forest (n_estimator = 2000)	0,5417357	0,9984624
Random Forest (n_estimator = 1000)	0,5374983	0,9984873

The experiments presented in Table 3 are hyperparameter tuning experiments for Random Forest by trying several values. The first row represents the Random Forest method without the n_estimator hyperparameter. Next, hyperparameter tuning experiments were conducted by entering n_estimator values of 200, 500, 1000, 1500, and 2000. From the results shown in Table 3, Random Forest with an n_estimator value of 1000 can minimize the MSE value from 0.5417357 to 0.5374983 and increase the R2 Score value from 0.9984624 to 0.9984873. Therefore, it can be concluded that Random Forest with an n_estimator value of 1000 is the best method when compared to other experiments. In other studies, several n_estimator values were also used, such as 50, 100, 200, 500, and 1000. The best model obtained also uses an

n_estimator of 1000, resulting in an accuracy value of 0.9984 (Agustina et al., 2022). The next stage involved experimenting on the Random Forest model with n_estimators of 1000 by utilizing a variety of different datasets. The aim of this experiment is to verify that the model created with Random Forest and an n_estimator value of 1000 maintains a consistently high level of accuracy across diverse datasets. Table 4 furnishes an in-depth overview of the datasets employed in our experiments.

Table 4: Dataset Descriptions

No	Dataset	Instance	Feature	Year
1	Student Performance Prediction (Kaggle - Rattanaporn, 2023)	1000	8	2023
2	Student Performance – Math Course (Paulo Cortez, 2014)	395	33	2014
3	Student Performance – Portuguese Language Course (Paulo Cortez, 2014)	649	33	2014

Table 5: Evaluation of hyperparameter tuning using Student Performance - Math Course Dataset

Method	MSE	R2 Score
Random Forest	0,3010254	0,9930548
Random Forest (n_estimator = 200)	0,2926249	0,9934225
Random Forest (n_estimator = 500)	0,2874371	0,9936465
Random Forest (n_estimator = 1500)	0,2913178	0,9934808
Random Forest (n_estimator = 2000)	0,2900142	0,9935404
Random Forest (n_estimator = 1000)	0,2870738	0,9936469

Table 5 shows the hyperparameter tuning evaluation results on the Student Performance - Math Course dataset. In addition, Random Forest with n_estimator 1000 successfully reduced the initial MSE value from 0.3010254 to 0.2870738 and increased the R2 Score value from 0.9930548 to 0.9936469.

Table 6: Evaluation of hyperparameter tuning using Student Performance - Portuguese Language Course Dataset

Method	MSE	R2 Score
Random Forest	0,4093905	0,9804961
Random Forest (n_estimator = 200)	0,4085096	0,9805879
Random Forest (n_estimator = 500)	0,4169781	0,9797218
Random Forest (n_estimator = 1500)	0,4082918	0,9806146
Random Forest (n_estimator = 2000)	0,4108604	0,9803311
Random Forest (n_estimator = 1000)	0,3989063	0,9814939

The evaluation results of hyperparameter tuning on the Student Performance - Portuguese Language Course dataset can be seen in Table 6. The dataset contains 649 instances and 33 features. From the results presented, it can be concluded that the Random Forest model with n_estimator 1000 is still the best compared to other models. The Random Forest model with n_estimator 1000 can minimize the initial MSE value of 0.4093905 to 0.3989063 and the R2 Score value increases from 0.9804961 to 0.9814939.

In general, a higher MSE level reflects poor performance, while a lower level is considered an indicator of better performance. Conversely, higher R values are usually considered a sign of better performance than lower ones. When the R value is 1, it indicates that the regression model fits the observed data very precisely Dewi & Chen, (2019). This study clearly shows a pattern of decreasing MSE values and increasing R2 values in each trial on all datasets. Based on the evaluation results illustrated in

Figure 7 the proposed method using Random Forest and n_estimator 1000 is able to outperform the performance of other methods on each dataset. Random Forest with n_estimator 1000 successfully minimizes the MSE value and maximizes the R2 Score value.

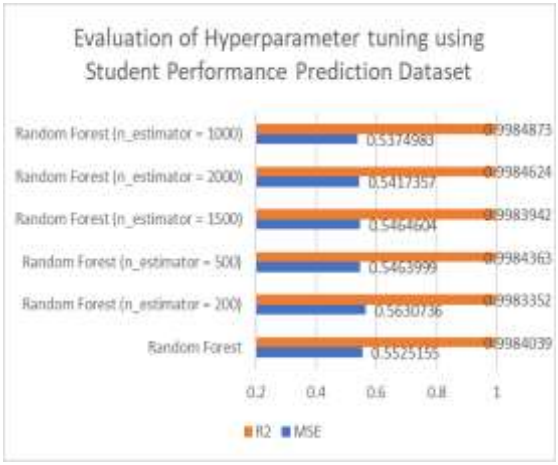


Figure 7: Student Performance Prediction Dataset

Figure 7 describes the Student Performance Prediction Dataset. The optimal R2 and MSE are achieved using Random Forest with n_estimator = 1000 with an R2 value of 0.9984873 and MSE of 0.5374983.

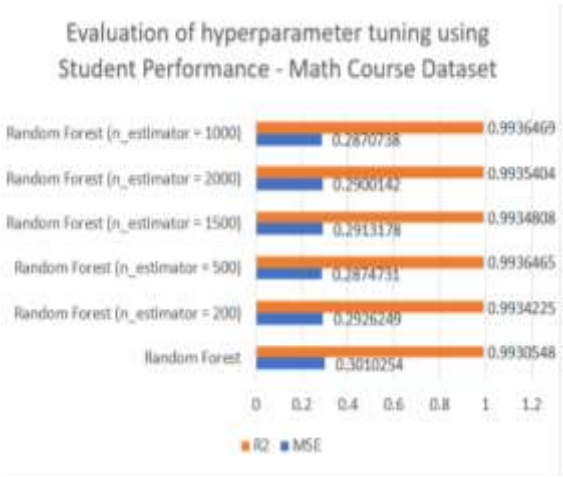


Figure 8: Student Performance - Math course Dataset

Figure 8 describes the Student Performance with Math Course Dataset. The optimal R2 and MSE were achieved using Random Forest with n_estimator = 1000 with an R2 value of 0.9936469 and MSE of 0.2870738. The R2 value which was originally 0.9930548 increased to 0.9936469 and the MSE value which was originally 0.3010254 decreased to 0.2870738.

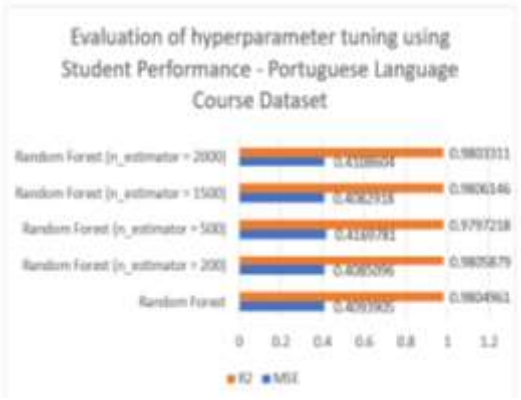


Figure 9: Student Performance - Portuguese Language

Figure 9 describes the Student Performance with Portuguese Language Course Dataset. In this dataset, the Random Forest method with n_estimator = 1000 is still the best. It can be seen in the figure that there is no drastic decrease. Initially, the MSE value was 0.4093905 to 0.3989069.

4. CONCLUSION

The contibution of this study is to improve the accuracy compared to previous studies by considering various alternative methods. Hyperparameters play a crucial role in training a Random Forest. They determine the structure and behavior of the model. Tuning hyperparameters can impact the model's performance, generalization ability, and overall effectiveness in handling different datasets. Key hyperparameters include the number of trees, tree depth, and the minimum samples required to split a node. Proper tuning helps optimize the Random Forest for specific tasks, balancing between underfitting and overfitting. The experimental results show that Random Forest has superior performance compared to other methods. After selecting Random Forest as the main method, the next step is to adjust the hyperparameters of this model to improve accuracy and reduce error rate.In this research, the optimized parameter is n_estimator. The hyperparameter tuning evaluation results, presented in Table 3, Table 5, and Table 6, show that the recommended model is Random Forest with n_estimator 1000, which successfully reduces the MSE value and increases the R2 Score value. For example, in Table 3, Random Forest with n_estimator 1000 successfully reduced the MSE value from 0.5525155 to 0.5374983 and increased the R2 Score value from 0.9984039 to 0.9984873. This model tuning process is very important because it can optimize the parameters to achieve the best accuracy. The results from the simulation experiments prove the validity and accuracy of the proposed algorithm. In future research, consideration of comparing different models and updating the dataset may also be relevant options.

REFERENCES

Adebowale, A., Idowu, S. a, & A, A. A. (2013). Comparative Study of Selected Data Mining Algorithms Used For Intrusion Detection. *International Journal of Soft Computing and Engineering (IJSCE)*, *3*(3).

Afzal, A., Alshahrani, S., Alrobaian, A., Buradi, A., & Khan, S. A. (2021). Power Plant Energy Predictions Based on Thermal Factors Using Ridge and Support Vector Regressor Algorithms. *Energies*, *14*(21). https://doi.org/10.3390/en14217254.

Agustina, I., Mulyani, Y., Septiana, T., & Mardiana, M. (2022). Analisis Pengembangan Model Prediksi Kesuksesan Kickstarter Menggunakan Algoritma Backpropagation dan Random Forest. *Jurnal Informatika Dan Teknik Elektro Terapan*, *10*(3).

https://doi.org/10.23960/jitet.v10i3.2742.
Altujjar, Y., Altamimi, W., Al-Turaiki, I., & Al-Razgan, M. (2016). Predicting Critical Courses Affecting Students Performance: A Case Study. *Procedia Computer Science*, *82*, 65–71. https://doi.org/10.1016/j.procs.2016.04.010.
Amusa, L., North, D., & Zewotir, T. (2021). Optimal Hyperparameter Tuning of Random Forests for Estimating Causal Treatment Effects. *Songklanakarin Journal of Science and Technology*, *43*(4).
Azmi, U., Hadi, Z. N., & Soraya, S. (2020). ARDL METHOD: Forecasting Data Curah Hujan Harian NTB. *Jurnal Varian*, *3*(2). https://doi.org/10.30812/varian.v3i2.627.
Christanto, H. J., Sutresno, S. A., Denny, A., & Dewi, C. (2023, August). Usability analysis of human computer interaction in google classroom and microsoft teams. Journal of Theoretical and Applied Information Technology, 101(16), 6425-6425.
Christanto, H. J., Sutresno, S. A., Simi, V. S., Dewi, C., & Dai, G. (2023). Analysis of Game Theory in Marketing Strategies of Tiktok and Instagram. Journal of Theoretical and Applied Information Technology, 101(22), 7100-7109.
Cominotte, A., Fernandes, A. F. A., Dorea, J. R. R., Rosa, G. J. M., Ladeira, M. M., van Cleef, E. H. C. B., Pereira, G. L., Baldassini, W. A., & Machado Neto, O. R. (2020). Automated Computer Vision System to Predict Body Weight and Average Daily Gain in Beef Cattle During Growing and Finishing Phases. *Livestock Science*, *232*. https://doi.org/10.1016/j.livsci.2019.103904.
Contreras, P., Orellana-Alvear, J., Muñoz, P., Bendix, J., & Célleri, R. (2021). Influence of Random Forest Hyperparameterization on Short-Term Runoff Forecasting in an Andean Mountain Catchment. *Atmosphere*, *12*(2). https://doi.org/10.3390/atmos12020238.
Dewi, C., & Chen, R. C. (2019). Random Forest and Support Vector Machine on Features Selection for Regression Analysis. *International Journal of Innovative Computing, Information and Control*, *15*(6). https://doi.org/10.24507/ijicic.15.06.2027.
Dewi, N. K., Mulyadi, S. Y., & Syafitri, U. D. (2012). Penerapan Metode Random Forest Dalam Driver Analysis. *Forum Statistika Dan Komputasi*, *16*(1).
Ganatra, D. (2020). Ensemble Methods to Improve Accuracy of a Classifier. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(3). https://doi.org/10.30534/ijatcse/2020/145932020.
Herdianto. (2013). Prediksi Keruskan Motor Induksi Menggunakan Metode Jaringan Saraf Tiruan Backprograption. *Fakultas Teknik, Universitas Sumatera Utara, Medan*.
Hidayat, Andi Sunyonto, & Hanif Al Fatta. (2023). *Klasifikasi Penyakit Jantung Menggunakan Random Forest Clasifier*.
Hosoe, M., Kuwano, M., & Moriyama, T. (2021). A Method for Extracting Travel Patterns Using Data Polishing. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-020-00402-w.
Jun, M. J. (2021). A Comparison of A Gradient Boosting Decision Tree, Random Forests, and Artificial Neural Networks to Model Urban Land Use Changes: The Case of The Seoul Metropolitan Area. *International Journal of Geographical Information Science*, *35*(11). https://doi.org/10.1080/13658816.2021.1887490.
Jungmeier, G. (2017). The Biorefinery Fact Sheet. *The International Journal of Life Cycle Assessment*, *23*(1).
Kaggle - KIATTISAK RATTANAPORN. (2023, February). *Student Performance Prediction*. https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics/data
Louk, M. H. L., & Tama, B. A. (2022). Tree-Based Classifier Ensembles for PE Malware Analysis: A Performance Revisit. *Algorithms*, *15*(9). https://doi.org/10.3390/a15090332.
Lusiana, A., & Yuliarty, P. (2020). Penerapan Metode Peramalan (FORECASTING) Pada Permintaan Atap di PT X. *Industri Inovatif : Jurnal Teknik Industri*, *10*(1). https://doi.org/10.36040/industri.v10i1.2530.
Pandey, M., & Taruna, S. (2016). Towards the Integration of Multiple Classifier Pertaining to the Student's Performance Prediction. *Perspectives in Science*, *8*. https://doi.org/10.1016/j.pisc.2016.04.076.
Paulo Cortez. (2014, November 26). *Student Performance*. https://doi.org/https://doi.org/10.24432/C5TG7T.
Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and Tuning Strategies for Random Forest. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 9, Issue 3). https://doi.org/10.1002/widm.1301.
Purwaningsih, E., & Nurelasari, E. (2021). Penerapan K-Nearest Neighbor Untuk Klasifikasi

Tingkat Kelulusan Pada Siswa. *Syntax: Jurnal Informatika*, *10*(01), 46.

Putra, P., Vinolia, & Novianty, H. (2020). *Implementation of Trend Moment Method in Egg Forecasting System in Sukamulia Farm*. https://doi.org/10.2991/aisr.k.200424.100.

Ren, S., Chan, H. L., & Siqin, T. (2020). Demand Forecasting in Retail Operations For Fashionable Products: Methods, Practices, and Real Case Study. *Annals of Operations Research*, *291*(1–2). https://doi.org/10.1007/s10479-019-03148-8.

Sarker, I. H., Colman, A., Han, J., Khan, A. I., Abushark, Y. B., & Salah, K. (2020). BehavDT: A Behavioral Decision Tree Learning to Build User-Centric Context-Aware Predictive Model. *Mobile Networks and Applications*, *25*(3). https://doi.org/10.1007/s11036-019-01443-z.

Schonlau, M., & Zou, R. Y. (2020). The Random Forest Algorithm for Statistical Learning. *Stata Journal*, *20*(1). https://doi.org/10.1177/1536867X20909688.

Sianturi, F. A. (2018). *Analisa Decision Tree Dalam Pengolahan Data Siswa*. *3*(2). http://ejournal.ust.ac.id/index.php/Jurnal_Means/

Siji George, C. G., & Sumathi, B. (2020). Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction. *International Journal of Advanced Computer Science and Applications*, *11*(9). https://doi.org/10.14569/IJACSA.2020.0110920.

Sinaga, H. D. E., Irawati, N., & Informasi, S. (2018). Perbandingan Double Moving Average Dengan Double Exponential Smoothing Pada Peramalan. *Jurteksi*, *IV*(2).

Suliztia, M. L. (2020). Penerapan Analisis Random Forest Pada Prototype Sistem Prediksi Harga Kamera Bekas Menggunakan Flask. *Fakultas Matematika Dan Ilmu Pengetahuan Alam*.

Thakkar, A., & Chaudhari, K. (2021). Fusion in Stock Market Prediction: A Decade Survey on the Necessity, Recent Developments, and Potential Future Directions. *Information Fusion*, *65*. https://doi.org/10.1016/j.inffus.2020.08.019.

Tian, Z., Xiao, J., Feng, H., & Wei, Y. (2020). Credit Risk Assessment based on Gradient Boosting Decision Tree. *Procedia Computer Science*, *174*. https://doi.org/10.1016/j.procs.2020.06.070.

Yang, F., & Li, F. W. B. (2018). Study on Student Performance Estimation, Student Progress Analysis, and Student Potential Prediction Based on Data Mining. *Computers and Education*, *123*. https://doi.org/10.1016/j.compedu.2018.04.006.