Journal of Student Research Exploration

https://shmpublisher.com/index.php/josre
p-ISSN 2964-1691 | e-ISSN 2964-8246

Classification to predict student academic performance using a Random Forest

Aditya Fajar Mulyana[1], Wiyanda Puspita[2], Jumanto[3]
[1, 2, 3] Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received June 12, 2023
Revised July 09, 2023
Accepted July 09, 2023

Keywords:

Classify
Student
Academic
Performance
Random Forest

ABSTRACT

This research aims to classify the academic performance of students who are successful and who have dropped out of school with high accuracy so that these matters can be addressed quickly. Things like this need fast handling to find out what factors influence it. In addition, this research was conducted to test how good the random forest algorithm is in classifying a problem. Random forest, which includes an algorithm that is commonly used for classifying a problem. By using the random forest algorithm, the accuracy results will be better than a single decision tree. This algorithm is quite good at handling and managing large datasets. From this study it can be concluded that this method can provide good prediction accuracy with a fairly high level of accuracy, namely 89%. Utilization of this random forest can be an alternative in classifying student academic achievement. This algorithm can work well in handling large datasets. This study discusses how the use of Random Forest can work to classify students' academic performance.

1. Introduction

The trend of dropping out of school and student success has always been a complex issue in the world of education. Although many efforts have been made to address the problem of dropping out of school, until now, the dropout rate in many

[1] ***Corresponding Author:***
Aditya Fajar Mulyana,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang, Indonesia.
Email: mulyanaaditya2082@students.unnes.ac.id
DOI: https://doi.org/10.52465/josre.v1i2.169

countries is still quite high [1]. Students who drop out of school are likely to face greater challenges in achieving future success, such as difficulty finding decent jobs and earning enough income to meet their living needs. Against, students who successfully complete their education have greater opportunities for future success, such as having difficulty finding decent jobs and earning enough income to meet their living needs. Conversely, students who successfully complete their education have a greater chance of success in the future and have a positive impact on society as a whole [2]. One of the main factors that can affect the dropout rate and success of students is their academic performance. Students who have difficulty understanding material and achieve low scores tend to be more prone to dropping out of school [3]. However, students who have good academic performance and get high scores in certain subjects tend to have a greater chance of successfully completing their education and achieving success in the future [4].

In addition, economic factors also play an important role in determining whether students will complete their education or not [5]. Families with low income levels may not be able to meet the financial needs necessary to support their children's education, such as school fees and educational supplies [6]. As a result, students may be forced to drop out of school due to the economic pressures they face. On the other hand, students from families with higher income levels may have access to greater resources and support that can help them achieve success in their education [7]. However, not all students from low-income families drop out of school, and not all students from high-income families succeed in their education. Individual factors such as motivation, aptitude, and academic ability also play a role in determining the success or failure of students in their education.

In some cases, out-of-school students can find success outside of formal education through the skills and experience they gain from work or training. However, this is not always the case and can limit future career and earning possibilities. This is because the orientation of society will always have the view that someone with a higher education is better than someone who dropped out of school in the middle of their education. Such statements cannot be condemned or justified. It only needs a decrease in dropout students. Therefore, efforts to reduce dropout rates and improve student academic achievement must be carried out by considering various factors, including individual, social, and economic factors [8]. Quality education and the right support from family, teachers and educational institutions can help students overcome obstacles and reach their potential in education and life.

There are many algorithms that can be applied to solve this problem where each algorithm has its own characteristics [9]. In this case, artificial intelligence technology can help educational institutions predict student academic achievement and identify students who are at risk of dropping out of school early [10]. In particular, there has been increasing interest in adopting Machine Learning

to predict student performance and identify at-risk students based on preliminary data collected during their studies [11]. For example, deep learning algorithm which is part of machine learning. The majority of techniques used are Deep Neural Network (DNN), Recurrent Neural Network (RNN) [12]. Thus, educational institutions can provide appropriate interventions and provide the necessary support to students to improve their academic performance and reduce the risk of dropping out. In the long term, this can help increase the effectiveness of the education system and provide greater benefits to society as a whole. However, to achieve future success, student academic achievement alone is not enough. Many other factors can affect student success, such as the ability to adapt to change, the ability to communicate well, and social skills. In addition, environmental factors such as family support and social environment can also affect student success. Therefore, educational institutions need to pay attention to these factors in supporting students to achieve success in the future.

To increase student success rates and reduce dropout rates, educational institutions need to continue to develop appropriate strategies and programs. The use of artificial intelligence technology can be one way to help educational institutions predict student academic achievement and identify students who are at risk of dropping out of school [13]. Several previous studies have been conducted to predict academic performance. The study [14] has conducted research related to predicting academic performance. It predicts student academic performance based on student learning activities in the e learning management system. This research produces an accuracy value of 76.92% using the C4.5 algorithm combined with correlation-based feature selection. In [15] a study was conducted related to predicting the academic success of architecture students based on previous academic performance (also referred to as pre-enrollment requirements) using K-nearest neighbor (k-NN). This study focused entirely on using previous academic performance as a predictor of academic success for undergraduate architecture students.

In this study using random forest to classify students' academic performance. Random Forest is one of the algorithms that uses bagging techniques [16]. Random forest has many advantages, such as the ability to process large datasets with incomplete attributes and the ability to process large datasets [17]. However, this technology can only be an effective tool if it is used correctly and in the right context. Therefore, educational institutions need to establish cooperation with various parties. Using algorithms such as the Decision Tree [18] or Random Forest can help educational institutions predict student academic performance and provide appropriate interventions to help them overcome the problems they face. In this way, it is expected to reduce dropout rates and increase student academic achievement.

2. Method

The stages of the research conducted in this article are described in Figure 1.

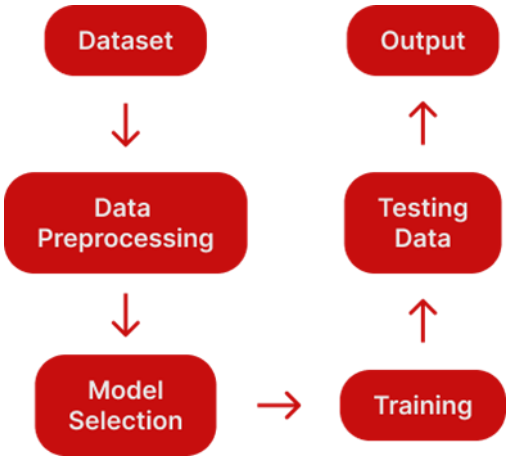

Figure 1. Research stage

2.1.Dataset

To obtain student information data, the authors use a dataset from Kaggle entitled "Student Academic Achievement Dataset" with multivariate characteristics. It is an educational data set collected from a learning management system (LMS) called Kalboard 360 [19]. The data is collected using a student activity tracking tool, called the experience API (xAPI). XAPI is one of the components of the training and learning architecture (TLA) which makes it possible to monitor learning progress and student actions such as reading articles or watching training videos. The experiences API helps learning activity providers define the students, activities, and objects that describe the learning experience. The dataset consists of 480 student records and 16 features. Features are classified into three main categories: (1) Demographic features such as gender and nationality. (2) Features of academic background such as educational stage, grade level, and section. (3) Behavioral characteristics such as raising hands in class, opening resources, answering surveys by parents, and school satisfaction.

2.2.Data exploration

In the data exploration stage, an in-depth analysis of the "Student Academic Performance" dataset was carried out. The purpose of this data exploration is to understand the structure of the dataset, identify patterns or trends, and obtain relevant insights to improve accuracy in predicting student academic achievement using the Random Forest Classifier. In data exploration, the writer visualizes a plot that displays the number of students in each class from the dataset. The purpose of this visualization is to see the distribution of the number of students in each

class. By using a countplot, we can easily see the difference in the number of students between classes.

2.3.Data processing

Data preparation is how the data is processed before it is used by the author to get results. In data processing, feature engineering and feature selection are carried out. Feature engineering is the process of creating new features or transforming existing features into more informative and relevant representations. The main goal of feature engineering is to improve understanding of the data, uncover hidden information, and improve model performance. In feature engineering, the writer changes the value of the student's gender category to be numeric, for example, 'M' (male) becomes 0, and 'F' (female) becomes 1. This is done to change the category value into a numerical representation that can be used in data processing and model building.

Meanwhile, feature selection is the process of selecting the most relevant and informative subset of features from the dataset. The purpose of feature selection is to reduce dimensional data, increase modeling speed and efficiency, and eliminate features that do not make a significant contribution to model creation. In selecting features the author checks all the columns in the dataset. After all the columns are checked, then you can select certain columns to do the modeling you want.

2.4.Building models and implementing algorithms

Determining and applying modelling techniques that are appropriate to the data conditions is essential to obtain optimal results [20]. After reviewing the widely used prediction methods, it is important to reemphasize the value of automation to select the optimal prediction model, given the complexity of such a task. Given the complexity of selecting the optimal predictive model for a given data set from a broad set of prediction methods and the different hyper parameter values per model, automating this process can help improve prediction accuracy. Random forest algorithm can be used for both Classification and Regression problems, it is considered to be the strongest algorithm. An algorithm model with many decision trees is like a forest with random value attributes. The level of accuracy is based on the number of trees made [21]. This random forest algorithm is used to divide the training data and test data. Random Forest involves combining multiple trees for training on a given set of sample data [22], then the training data is used to train the model. This parameter is used as much as possible to achieve good performance. After that, an evaluation is carried out using test data. Starting from an existing dataset, then the dataset is processed. After that, model training is carried out with training data. After the training data produces good performance, an evaluation of the test data is also carried out which then produces output.

3. Results and Discussion

The authors experimented with the random forest classifier method to predict student performance from a data sourced from a learning management system (LMS) called Kalboard 360. Among other things, the authors obtained an accuracy of 89% in the random forest classifier method.

In this visualization there is no separation based on other factors such as gender or other variables. The focus of this visualization is to display the total number of students in each class. Next, a visualization is carried out to display the number of students in each class based on gender. The purpose of this visualization is to see the distribution of students in each class based on gender. By disaggregating the data by gender, we can see the difference in the number of male and female students in each class. Figure 2 and Figure 3 illustrate the results of the visualization for displaying the number of students per class and displaying the number of students per class based on gender.

Figure 2. Visualization displays the number of students in each class

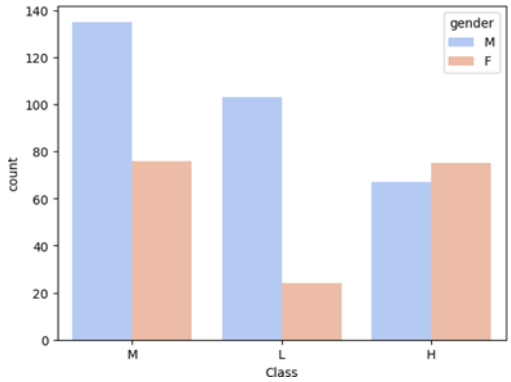

Figure 3. The visualization displays the number of students in each class by gender

Finally, the pair plot displays the relationship between the variables in the dataset based on the class of students. Pairplot is a grid that displays the scatter plot of each pair of variables in the dataset. In this case, the pair plot is used to see the

relationship between variables in the Student Academic Achievement Dataset, with separation based on student class. In each scatter plot, the x and y axes show the values of the two variables being compared. The color of the dots on the scatter plot is differentiated based on the class of students. This helps us to see the pattern of relationships between variables in the dataset and understand how these variables affect the classification of students into certain classes. The purpose of this pair plot is to provide an initial understanding of the relationship between the variables in the dataset and the class of students. By looking at the scatter plot pattern, we can see if there is a clear relationship between the variables and student classes, or whether there are significant differences between classes in terms of the values of certain variables. Figure 4 illustrates a visualization that shows the relationship between variables in a dataset based on student classes.

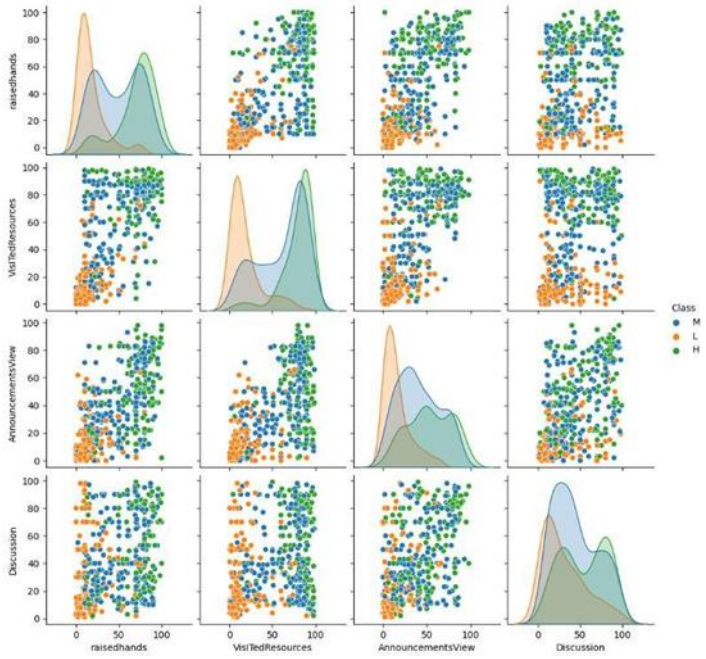

Figure 4. The visualization displays the relationship between the variables in the dataset based on the class of students

After going through the data exploration process, the next stage is data processing and model building. Figure 5 is an illustration for the existing class confusion matrix, namely class M, class L, and class H. This illustration provides information about how well the model can distinguish the existing classes. Here 0 represents class M, 1 represents class L, and 2 represents class H. Illustrations of the models that have been built can be seen in Figure 5 and Figure 6.

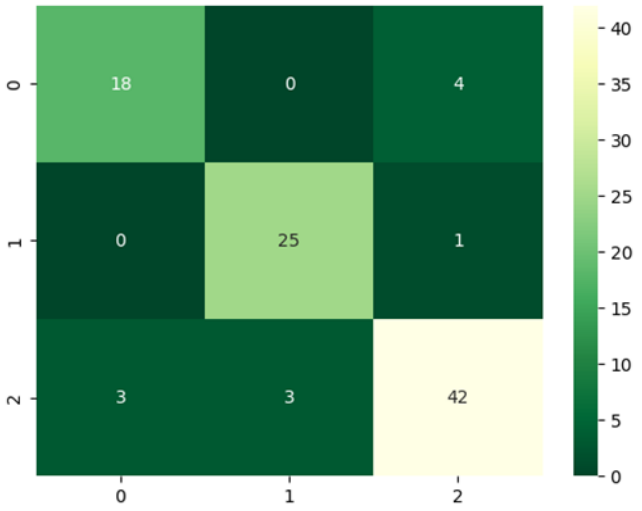

Figure 5. Confusion matrix class M, L, and H

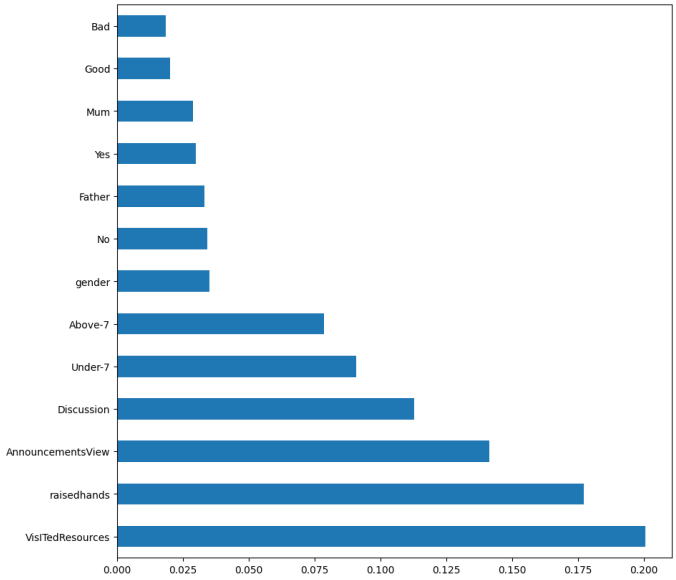

Figure 6. Influential features

From the illustration, class H gets a high rating compared to other classes. From Figure 6 Visited Resource feature gets the highest rating after the raisehands feature. The Visited Resource feature is the number of times students visit course content.

4. Conclusion

Student performance is one of the most significant criteria for any college. The Random Forest algorithm model can be used to classify dropout rates and student success. This classification method applies a visualization system to produce the

best accuracy. The best accuracy results are obtained as a temporary accuracy of 89%. This accuracy is obtained through calculation results that utilize the Random Forest algorithm with the application of a visualization system. The final results are obtained by taking the average results from the test data and training data.

REFERENCES

[1] A. M. Mariano, A. B. de M. L. Ferreira, M. R. Santos, M. L. Castilho, and A. C. F. L. C. Bastos, "Decision trees for predicting dropout in Engineering Course students in Brazil," *Procedia Comput. Sci.*, vol. 214, pp. 1113–1120, 2022, doi: 10.1016/j.procs.2022.11.285.

[2] I. A. Abu Amra and A. Y. A. Maghari, "Students performance prediction using KNN and Naïve Bayesian," in *2017 8th International Conference on Information Technology (ICIT)*, IEEE, May 2017, pp. 909–913. doi: 10.1109/ICITECH.2017.8079967.

[3] S. Sarwat *et al.*, "Predicting Students' Academic Performance with Conditional Generative Adversarial Network and Deep SVM," *Sensors*, vol. 22, no. 13, p. 4834, Jun. 2022, doi: 10.3390/s22134834.

[4] Y.-H. Lo, D.-F. Chang, and A. Chang, "Exploring Concurrent Relationships between Economic Factors and Student Mobility in Expanding Higher Education Achieving 2030," *Sustainability*, vol. 14, no. 21, p. 14612, Nov. 2022, doi: 10.3390/su142114612.

[5] S. C. Gewalt, S. Berger, R. Krisam, J. Krisam, and M. Breuer, ""University students' economic situation during the COVID-19 pandemic: A cross-sectional study in Germany"," *PLoS One*, vol. 17, no. 10, p. e0275055, Oct. 2022, doi: 10.1371/journal.pone.0275055.

[6] D. Novitasari, J. Juliana, M. Asbari, and A. Purwanto, "The Effect of Financial Literacy, Parents' Social Economic and Student Lifestyle on Students Personal Financial Management," *Econ. Educ. Anal. J.*, vol. 10, no. 3, pp. 522–531, Oct. 2021, doi: 10.15294/eeaj.v10i3.50721.

[7] H. Li, "How to Retain Global Talent? Economic and Social Integration of Chinese Students in Finland," *Sustainability*, vol. 12, no. 10, p. 4161, May 2020, doi: 10.3390/su12104161.

[8] O. Taylan and B. Karagözoğlu, "An adaptive neuro-fuzzy model for prediction of student's academic performance," *Comput. Ind. Eng.*, vol. 57, no. 3, pp. 732–741, Oct. 2009, doi: 10.1016/j.cie.2009.01.019.

[9] R. Muzayanah and E. A. Tama, "Application of the Greedy Algorithm to Maximize Advantages of Cutting Steel Bars in the Factory Construction," *J. Student Res. Explor.*, vol. 1, no. 1, pp. 41–50, Dec. 2022, doi: 10.52465/josre.v1i1.112.

[10] Z. Chen *et al.*, "Education 4.0 using artificial intelligence for students performance analysis," *Intel. Artif.*, vol. 23, no. 66, 2020, doi: 10.4114/intartif.vol23iss66pp124-137.

[11] N. N. Hamadneh, S. Atawneh, W. A. Khan, K. A. Almejalli, and A. Alhomoud, "Using Artificial Intelligence to Predict Students' Academic Performance in Blended Learning," *Sustainability*, vol. 14, no. 18, p. 11642, Sep. 2022, doi: 10.3390/su141811642.

[12] M. H. Diponegoro, S. S. Kusumawardani, and I. Hidayah, "Tinjauan Pustaka Sistematis: Implementasi Metode Deep Learning pada Prediksi Kinerja Murid," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 10, no. 2, pp. 131–138, 2021.

[13] S. R. Rahman, M. A. Islam, P. P. Akash, M. Parvin, N. N. Moon, and F. N. Nur, "Effects of co-curricular activities on student's academic performance by machine learning," *Curr. Res. Behav. Sci.*, vol. 2, p. 100057, Nov. 2021, doi: 10.1016/j.crbeha.2021.100057.

[14] A. S. B. Asmoro, W. S. G. Irianto, and U. Pujianto, "Perbandingan Kinerja Hasil Seleksi Fitur pada Prediksi Kinerja Akademik Siswa Berbasis Pohon Keputusan," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, p. 84, Dec. 2018, doi: 10.26418/jp.v4i2.29294.

[15] R. O. Aluko, O. A. Adenuga, P. O. Kukoyi, A. A. Soyingbe, and J. O. Oyedeji, "Predicting the academic success of architecture students by pre-enrolment requirement: using machine-learning techniques," *Constr. Econ. Build.*, vol. 16, no. 4, pp. 86–98, 2016.

[16] T. Lailatul Nikmah, R. M. Syafei, R. Muzayanah, A. Salsabila, and A. A. Nurdin, "Prediction of Used Car Prices Using K-Nearest Neighbour, Random Forest, and Adaptive Boosting Algorithm," *Int. Conf. Optim. Comput. Appl.*, vol. 1, no. 1 SE-Articles, pp. 17–22, Dec. 2022, [Online]. Available: https://e-conference.ptti.web.id/index.php/icoca/article/view/15

[17] J. Jumanto, M. A. Muslim, Y. Dasril, and T. Mustaqim, "Accuracy of Malaysia Public Response to Economic Factors During the Covid-19 Pandemic Using Vader and Random Forest," *J. Inf. Syst. Explor. Res.*, vol. 1, no. 1, pp. 49–70, 2022.

[18] H. Guruler, A. Istanbullu, and M. Karahasan, "A new student performance analysing system using knowledge discovery in higher educational databases," *Comput. Educ.*, vol. 55, no. 1, pp. 247–254, Aug. 2010, doi: 10.1016/j.compedu.2010.01.010.

[19] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm," *Procedia Technol.*, vol. 25, pp. 326–332, 2016, doi: 10.1016/j.protcy.2016.08.114.

[20] R. Ruswati, A. I. Gufroni, and R. Rianto, "Associative Analysis Data Mining Pattern Against Traffic Accidents Using Apriori Algorithm," *Sci. J. Informatics*, vol. 5, no. 2, pp. 91–104, Nov. 2018, doi: 10.15294/sji.v5i2.16199.

[21] S. Jayaprakash, S. Krishnan, and V. Jaiganesh, "Predicting Students Academic Performance using an Improved Random Forest Classifier," in *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*, IEEE, Mar. 2020, pp. 238–243. doi: 10.1109/ESCI48226.2020.9167547.

[22] J. Jumanto, M. F. Mardiansyah, R. N. Pratama, M. F. Al Hakim, and B. Rawat, "Optimization of breast cancer classification using feature selection on neural network," *J. Soft Comput. Explor.*, vol. 3, no. 2, pp. 105–110, Sep. 2022, doi: 10.52465/joscex.v3i2.78.