Universitas Ahmad Dahlan
Bulletin of Applied Mathematics and Mathematics Education
Volume 3, Number 2, 2023 pp. 79-84
https://dx.doi.org/10.12928/bamme.v3i2.10040
e-ISSN 2776-1029, p-ISSN 2776-1002

Classification of student abilities in reducing students' drop-out rates

Uswatun Khasanah*, Dwi Astuti

Universitas Ahmad Dahlan, Jl. Jend. Ahmad Yani, Tamanan, Banguntapan, Bantul, DIY 55711, Indonesia

*Corresponding E-mail: uswatun.khasanah@pmat.uad.ac.id

ARTICLE INFO

Article History
Received 14 December 2023
Revised 29 December 2023
Accepted 29 December 2023

Keywords
Education data mining
Reducing drop-out rates
Student ability classification

How to cite this article:
Khasanah, U., & Astuti, D. (2023). Classification of student abilities in reducing student drop-out rates. *Bulletin of Applied Mathematics and Mathematics Education*, *3*(2), 79-84.

ABSTRACT

The number of students in the mathematics education study program has declined over the past four years. Besides a few students enrolled in the first year, there are those who have had a temporary leave. These two things are necessary to identify the cause of study leave and need to analyze education data mining (EDM) to obtain a model of classification of students' abilities so that students can improve their achievement. The impact will be a decrease in the number of students undergoing the study. The subject of this study is a student of mathematics education study program from 2008 to 2018. Based on this data set, each student has information about gender, date of birth, home address, postal code, parent's name, secondary school origin, parent job, parent income cell phone number, graduation/drop-out status. The aim of the study is to identify an efficient model between a decision tree, a random forest and a neural network, based on the accuracy of 80 percent decision tree methods, 87 percent random forest methods and 78 percent neural networks.

Introduction

The Institute of Education always wanted to know the level of success of its students in the learning process. Schools in Indonesia student success rate refers to the Minimum Graduation Score (KKM). If the student scores at least equal to KKM, then it is said that the student is grateful, otherwise if the score is less than KKM then students are said to have not gratefully so students are obliged to follow remedial activities. Therefore, a teacher or an educational institution, at the end of each year, holds an evaluation to find out the factors causing student incomes. Data Mining (EDM) is a science discipline that uses data mining techniques in education. EDM can be considered as an alternative in learning knowledge. The level of student achievement can be predicted with some techniques (Jalota & Agrawal, 2019). Jalota's research results show that perceptron multilayer algorithms have the best performance. These results can be used to identify unique patterns that help learning, predict and improve student academic performance, and even predict student failure.

Some methods of classification include decision tree, Naïve bayes, neural network, statistical analysis, genetic algorithms, rough sets, nearest k-neighbors, rule-based methods, memory-based reasoning and support vector machine (Larose, 2005). Classification with Neural Network has been discussed by (Bishop, 1995; Dreyfus, 2005; Ertel, 2017; Lintas et al., 2017), regarding the decision tree (Rokach & Maimon, 2015), also discusses it. Three efficient approaches—decision trees, neural

networks, and support machines—have been identified by prior research. Thus, we will compare the effectiveness of multilayer perceptrons and radial basis neural networks to decide which neural network technique is better in this study. Cartoz Pailo's data was utilized for this data analysis (Cortez & Silva, 2015) . The model that was developed can be used at Ahmad Dahlan University to reduce the number of student dropouts based on the data. To create a predictive model of the number of students who drop out, the study program should therefore determine the dropout cause factor and the necessary Educational Data Mining analysis.

During this time, there are two possibilities of the database being either unused or deleted. The database contains important information that can be used to make decisions or to develop science. Knowledge Discovery in Database (KDD) is a method for obtaining knowledge from existing databases. According to Bramer, the stages of KDD are data selection, pre-processing/ cleaning, transformation, data mining, interpretation/ evaluation (Bramer, 2020).

Educational Data Mining (EDM) is a science discipline that uses data mining techniques. EDM can be considered as an alternative in learning knowledge. The majority of scholars have focused on data mining, which is utilized in education to forecast students' performance. By using a previously labelled class training set as a basis, classification is a data mining technique that predicts a class or group for an instance of a new dataset. Numerous categorization techniques exist, including Support Vector Machine (SVM), Random Forest, K Nearest Neighbor, Naïve Bayes, Neural Networks, and Logistic Regression (Romero et al., 2011). In (Jalota & Agrawal, 2019), employ the five classification algorithms—random forest, Naive Bayes, Multilayer Perceptron, support vector machine, and J48—in data mining approaches. His research's findings indicate that the Perceptron Multilayer technique performs the best. ADT, ID3, C4.5, CART, and other algorithms are used in the decision-tree technique of the Drop-Out Student Data Classification Analysis in (Pal, 2012).

According to the study's findings, the machine learning algorithm ID3 is capable of creating a prediction model that is accurate. In order to assess the viability of applying machine learning in the field of education, (Salloum et al., 2020) examined how scholars have approached data mining in the past as well as current trends in data mining in educational research. Regarding his research findings, they gave the EDM authorities more confidence to incorporate the suggestions made into the real-world system so that administrators, teachers, and students can all benefit from the best possible outcome. The possibility of cooperation between scholars from constructivist and EDM schools was examined in (Berland et al., 2014). The differences and similarities between learning analysis and EDM were examined in (Baek & Doleck, 2021).

Educational data mining (EDM), a recently developed field that deals with the examination and analysis of data from academic databases, is defined by (Jacob et al., 2015) as learning science. By utilizing diverse data mining techniques to examine these large data sets, one might find distinctive patterns that aid in understanding, forecasting, and enhancing the academic achievement of students. To determine if a change in one variable affects another, correlation is utilized. In this study, the decision tree is utilized to predict student performance by offering potential outcomes. When building a model with multiple independent and dependent variables, regression analysis is employed. If the model is deemed satisfactory, the values of the free variables are utilized to calculate the bound variable's value. Data mining techniques like regression and decision trees are used to predict academic performance, and clustering is used to identify groups of objects that are more similar to each other than to objects in other clusters. These techniques are effectively used to predict student performance and to forecast academic failure. Students are effectively grouped into groups based on their. These strategies will be very helpful to teachers in identifying areas that require improvement, in helping weaker students learn more effectively, in identifying weak students who may struggle, and in helping their students have a better educational experience. Additionally, these strategies will assist students in identifying job profiles that they can apply for based on their areas of expertise.

Considering that a model will be created to forecast and enhance student academic progress using data mining techniques, based on background information and library studies. The classification will also reveal whether a student is in a strong or weak academic group, which will assist the lecturer in creating a lesson plan.

10.12928/bamme.v3i2.10040

Decision trees, Naïve Bayes, neural networks, statistical analysis, genetic algorithms, rough sets, K-nearest neighbors, rule-based techniques, memory-based reasoning, and support vector machines are a few examples of classification techniques (Larose, 2005). There is a discussion of Related Neural Network in (Bishop, 1995; Dreyfus, 2005; Ertel, 2017; Lintas et al., 2017). Within (Rokach & Maimon, 2015), relating decision trees are also covered. Previous research indicates that methods like decision trees, neural networks, and support machines likely to be effective. As a result, we shall compare the three techniques' levels of effectiveness in this study.

Method

This research is applied research. The data used in this research is secondary data that is the data of students of the study program of mathematics education in 2008-2018 obtained from the instance of academic information. As for the design and procedure of the research is as follows: (1) literature studies related to classification techniques, (2) study related to matters related to student performance, (3) predict student performance using decision tree and neural network, and (4) grouping students according to the strength of academic ability.

Data analysis using R programs related to classification and modeling. The steps in analysis are as follows: (1) Preparation of data sets that will then be used as training data and test data; (2) Compilation of algorithms for the classification process; (3) Evaluation of the model by looking at the accuracy values such that the baseline accurate values are possible i.e. by simulating the amount of test data taken, the number of iterations, for taking the activation function, number of neurons, amount of data taken; (4) Obtained model/classification of data with a high level of accuracy; (5) Advance the results, in order to improve the learning process so as to improve student performance. In other words, how do we reduce the dropout rate of students in the mathematics education study program of Universitas Ahmad Dahlan.

Results and discussion

Currently, the mathematics education study program of Universitas Ahmad Dahlan has entered the TS-1 (1 year before assessment) stage in accreditation, so it has begun to improve the quality of the study program. One of the indicators of performance success is measured by the success rate of the study. The success of this study is the number of students who entered the TS-7 (7 year before assessment) that graduated. Therefore, he did this research to anticipate the students who had the chance to do study program did some treatments so that the students could graduate even in the last year / semester 14. The study will then identify a group of students who have a chance to graduate and a group that has the chance to drop out (DO) by looking at some aspects such as gender, school origin, father's job, mother's work, parents' income.

Based on Hwang (Hwang et al., 2020), artificial intelligent (AI) advances have brought computer-supported education to a new era. By incorporating maneuverage intelligence, system computers can serve as intelligent tutors, tools, or tuttee as well as facilitate decision-making in educational settings. The integration of AI and Education will open up new opportunities to improve the quality of teaching and learning. Teachers can use it in assessment, data collection, improving learning progress and developing new learning strategies. Moreover, the integration of AI and Education is a transformation of human knowledge, cognition, and culture. To develop a model for forecasting academic performance, six steps are taken.

Data collection

The data used in the research is data of students of the study program of Mathematical Education in 2008 - 2018 of 1775 data taken from the academic system. Based on this data set, each student has information about gender, date of birth, home address, postal code, parents' name, high school origin, middle job, middle income cell phone number, graduation status/DO.

Initial insight

The data obtained was subsequently observed, empty sections, incorrect sections, duplication and unused data, then deleted, thus obtaining a final sample of 1250.

Statistical analysis

At this stage the analysis carried out is a descriptive statistical analysis of the variables used in this study. As for the ratio of respondents based on:

(1) Gender: 22% male and 78% female.
(2) School origin: from high school 86%, from SMK 7.5 % and from MA 6.4 %.
(3) Father's job: farmer 15 %, entrepreneur 36 %, PNS 34 % , TNI 2.4 %, and other 12.6 %.
(4) Mother' job: farmer 5.6 %, entrepreneur 17.4 %, PNS 23.1 %, TNI 0.08 %, housewives 47.36 %, and other 12.6 %.
(5) Average parental income in the range of IDR 1,000,000 – IDR 5,000,000.

Data preprocessing

The data processing was carried out on academic data of mathematics education study program's students. The 1250 data were taken from the academic system, grouped into group 1 (graduated) and group 2 (DO). The analysis was done with Decision Tree, Random Forest and Neural Network.

Decision tree

Based on the analysis obtained results presented in Table 1.

Table 1. Results of analysis using the decision tree method

	Precision	Recall	F1-score	Support
0	0.17	0.16	0.16	32
1	0.88	0.89	0.88	218
Accuracy			**0.80**	250
Macro avg	0.53	0.52	0.52	50
Weighted avg	0.79	0.89	0.79	250

Based on Table 1, it appears that the accuracy of the result using the decision tree is 80%.

Random Forest

Based on the analysis obtained results presented in Table 2.

Table 2. Results of analysis using the random forest method

	Precision	Recall	F1-score	Support
0	0.00	0.00	0.00	32
1	0.87	1.00	0.93	218
Accuracy			**0.87**	250
Macro avg	0.44	0.50	0.47	250
Weighted avg	0.76	0.87	0.81	250

Based on Table 2, it appears that the accuracy of the result using the decision tree is 87%.

Neural network

Using the sigoid activation function, iterations as many as 1000 times, training data 80% and test data 20% obtained the following results (See Figure 1, Figure 2, Table 3, and Table 4).

Figure 1. Training and validation loss

10.12928/bamme.v3i2.10040

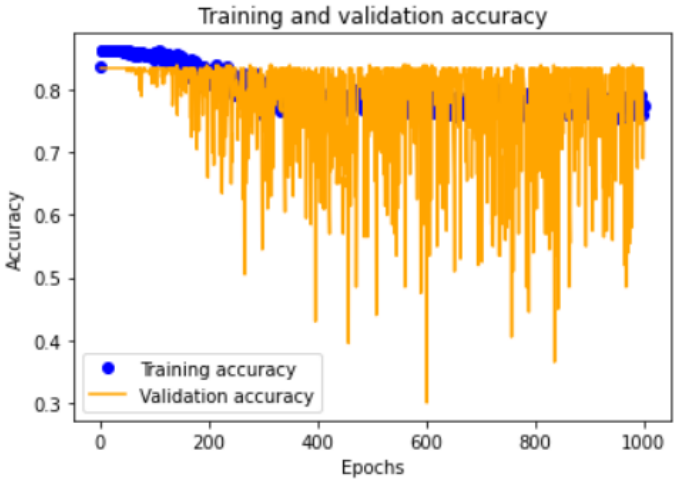

Figure 2. Training and validation accuracy

Table 3. Results of analysis using the neural network method

	Precision	Recall	F1-score	Support
0	0.15	0.16	0.15	32
1	0.88	0.87	0.87	218
Accuracy			0.78	250
Macro avg	0.51	0.51	0.51	250
Weighted avg	0.78	0.78	0.78	250

Model evaluation

Based on Table 1, 2 and 3 can be concluded as follows.

Table 4. Summary of accuracy results of the three methods

No	Methods	Accuracy
1	Decision tree	80%
2	Random Forest	87%
3	Neural Network	78%

Based on Table 4 above, it is then obtained that the best model based on accuracy values is by using the random forest method.

Conclusion

The subject of this study is a student of Mathematical Education from 2010 to 2018. Based on this data set, each student has information about gender, date of birth, home address, postal code, parent's name, secondary school origin, parent job, parent income cell phone number, graduation/DO status. The aim of the study is to identify an efficient model between a decision tree, a random forest and a neural network, based on the accuracy of 80 percent decision tree methods, 87 percent random forest methods and 78 percent neural networks. Thus, the best model based on accuracy values is by using the random forest method. However, the results have not been further analyzed to take policy at the prodi level so that it can reduce the level of DO in the study program.

Acknowledgement

We thank the Lembaga Penelitian dan Pengabdian kepada Masyarakat Universitas Ahmad Dahlan for providing the research grant under the scheme of Penelitian Dasar Number PD-234/SP3/LPPM-UAD/VII/2022.

References

Baek, C., & Doleck, T. (2021). Educational Data Mining versus Learning Analytics: A Review of Publications From 2015 to 2019. *Interactive Learning Environments*, *31*(6), 3828–3850. https://doi.org/10.1080/10494820.2021.1943689

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. In *Technology, Knowledge and Learning* (Vol. 19, Issues 1–2, pp. 205–220). Kluwer Academic Publishers. https://doi.org/10.1007/s10758-014-9223-7

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendo Press.

Bramer, M. (2020). *Principles of Data Mining*. Springer London. https://doi.org/10.1007/978-1-4471-7493-6

Cortez, P., & Silva, A. (2015). *Using Data Mining to Predict Secondary School Student Performance*.

Dreyfus, G. (2005). *Neural Networks: Methodology and Applications*. Springer-Verlag Berlin Heidelberg.

Ertel, W. (2017). *Introduction to Artificial Intelligence* (2nd ed.). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-58487-4

Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. In *Computers and Education: Artificial Intelligence* (Vol. 1). Elsevier B.V. https://doi.org/10.1016/j.caeai.2020.100001

Jacob, J., Jha, K., Kotak, P., & Puthran, S. (2015). Educational Data Mining Techniques and their Applications. *Proceedings of the 2015 International Conference on Green Computing and Internet of Things (ICGCIoT) 8-10 October 2015, Greater Noida, India : Venue: GCET, Greater Noida, Delhi*, 1344–1348.

Jalota, C., & Agrawal, R. (2019). Analysis of Educational Data Mining using Classification. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing : Trends, Prespectives and Prospects : COMITCON-2019 : 14th-16th February, 2019*, 243–247.

Larose, D. T. (2005). *Discovering knowledge in data: An Introduction to Data Mining*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Lintas, A., Rovetta, S., Verschure, P. F. M. J., & Villa, A. E. P. (2017). Artificial Neural Networks and Machine Learning-ICANN 2017. *26th International Conference on Artificial Neural Networks Alghero, Italy,*. https://doi.org/https://doi.org/10.1007/978-3-319-68612-7

Pal, S. (2012). Mining Educational Data to Reduce Dropout Rates of Engineering Students. *International Journal of Information Engineering and Electronic Business*, *4*(2), 1–7. https://doi.org/10.5815/ijieeb.2012.02.01

Rokach, L., & Maimon, O. (2015). *Data Mining With Decision Tree* (H. Bunke & P. Wang, Eds.; 2nd ed.). World Scientific Publishing Co. Pte. Ltd. http://www.worldscientific.com/series/smpai

Romero, C., Ventura, S., Pechenizkiy, M., & SJd Baker, R. (2011). *Handbook of Educational Data Mining*.

Salloum, S. A., Alshurideh, M., Elnagar, A., & Shaalan, K. (2020). Mining in Educational Data: Review and Future Directions. *Advances in Intelligent Systems and Computing*, *1153 AISC*, 92–102. https://doi.org/10.1007/978-3-030-44289-7_9

10.12928/bamme.v3i2.10040