Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 216 (2023) 597–603

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Computer Science and Computational Intelligence 2022

Determining factors that affect student performance using various machine learning methods

Nicholas Robert Beckham[a], Limas Jaya Akeh[a], Giodio Nathanael Pratama Mitaart[a], Jurike V Moniaga[a,]*

[a]*Computer Science Department,School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia*

Abstract

Students face problems that might hinder their academic pursuit toward success, problems ranging from trivial matters such as class condition, feeling of the student to severe matters such as family breakdown, economic reasons, and many more. This is a major problem because students shape the future of a nation – which will affect many things in the future. Teachers are looking for an effective way to find what might generally be the best solution for solving certain problems, as each student may face different problems, solving one at a time is not possible with the number of students each year. In this paper, we will try to find factors that might hinder or improve student performance using Pearson correlation between each feature toward the student G3 result. Based on the result, past failures will negatively impact student grades with -0.360415 correlation, and then Mother's Education will positively impact student grades with 0.217147. After finding out which factor affects student grade, we try to predict student grade using ML models to prove whether that factor actually affects student grade. Our MLP 12-Neuron model performs the best with RMSE value of 4.32, followed by Random Forest with RMSE value of 4.52, and finally Decision Tree with RMSE value of 5.69.

Keywords: *Student Performance; Machine Learning; Multilayer Perceptron; Random Forest; Decision Tree*

* Corresponding author.
E-mail address: jurike@binus.edu

10.1016/j.procs.2022.12.174

1. Introduction

Academic performance is one component that every student wants to achieve. This can be seen from the performance of students in doing the tasks given in class and at home. [1] However, not all students perform similarly. Poor academic performance in students has been a subject of concern to many people including parents, administrators, educators, psychologists and counsellors. [2]

The main objective of this paper is to determine factors that affect student performance using a combination of Machine Learning methods, inclusive but not limited to using Pearson correlations, using models such as Multilayer Perceptron, Decision Tree, and Random Forest to predict student grades. This method is hopefully to find any patterns or insights that help educators decide what factors they need to focus on to ensure better student performance in schools.

Student Performance (usually abbreviated as "Academic Performance" or "Academic Achievement"), is a measurement of how students (the one who are pursuing toward their goal) fare to their long-term educational goal. Completions of such benchmark result in diplomas and degrees representing the academy. Measurements can be calculated in many standards, such as KPI (Key Performance Index), GPA (Grade Point Average); in Indonesia alone, GPA are the most used standard ranging from 0-100 or 0-10 for high school students (with 100 and 10 for the highest attainable score, respectively), and 0 to 4.0 for undergraduate and master students (with 4.0 for the highest attainable score), note that many academies can include their own standards, since measurement can be different from each school or university. ITB is one of the most popular universities in Indonesia that adopts multiple half-grading systems (such as AB, BC, and CD).

Students with high Student Performance may be given rewards in many forms, examples including, scholarship that relieves some of the economic burden to the student family, and Latin titles for exceptional students in Universities, in Binus University alone, Summa Cum Laude are given for students with a minimum of 3.90 GPA, and Cum Laude are given for students with a minimum of 3.50 GPA and must attain at least B across all course.

However, the problems are focused on students with low Student Performance, students that failed the standard may result in repeating the same course, take supplementary lessons, or in the worst case, may result in drop-out of the student. In 2012 alone, the drop-out rate for primary education students was 18.2% [3], meaning that for every 1000 students, 182 of them drop-out for some reason. Note that drop-out might happen not because of student performance, but from other reasons that result in the student discontinuing their study.

With the recent rise of Artificial Intelligence, almost any task can be automated. Task that may seem redundant at first, and considered repetitive, can be easily maintained by robots thanks to Artificial Intelligence, in recent years, many researchers are trying to find the most effective way to build the most advanced Artificial Intelligence; hence the birth of Machine Learning, a subset studying how machine learns and improving statistical models to carry out tasks without explicit instruction.

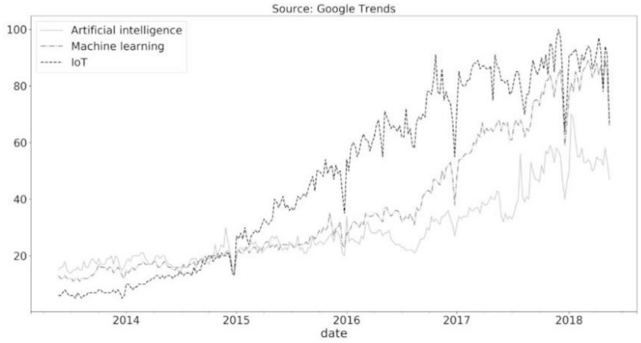

Fig. 1. Machine Learning search trend over the year according to Google Trends. (source: Salamone et.al. [4])

According to the statistics above, specific keywords related to Machine Learning are increasing in search trends, and published research papers related to it are increasing steadily, since Machine Learning can be developed for many reasons, including for business, economical growth, science, industry, education, and many more. IoT

(Internet of Things) is also one of key reasons why Machine Learning is popular, for example in Education, especially in pandemics, teachers can use web cameras to connect students with other students to study together without having to actually meet at all.

Machine Learning can be approached from many types that have their own strengths and weaknesses. Most common approaches include supervised learning, where machines are given data with labels, and try to study the characteristics of the data, and try to correctly predict whenever a new data belongs to which group. This is the most traditional learning method and is quite effective on building models, however, Machine Learning requires a lot of data, so supervised learning is usually much more expensive due to lack of labelled data. Another approach, which is unsupervised learning, combats the lack of data by just allowing the machine to consume the huge data, however without any label. This is usually used for clustering since the machine only identifies common points between data. Reinforced Learning is one area concerning how the machine takes action in an environment to maximize reward; the machine will start off randomly, and will be given reward for any correct or win, and punishment for any incorrect answer or lose, most common usage including Mind Games, where Machine are given reward for winning in chess, and punishment otherwise. In fact, the machine was able to beat the World Grandmaster of Chess.

2. Goal

The goal of this study is to find what affects student performance the most, by using several Machine Learning methods. We can find out the weight of each factor that affects Student Performance using Pearson correlation coefficient, with stronger weight meaning that the factor strongly affects student performance.

Analysis will then be performed on the factors that are selected by the Machine learning models to try and break down why those factors affect student performance to that degree. We will also look into how much each factor influences student performance, in other words, the intensity of correlation between the factors and the target. Furthermore, this will validate that whether factors chosen by the models are able to predict a student's performance, if the model is able to predict correctly most of the time with the factors then that factors greatly affects student performance.

3. Methodology and Dataset

The data used in this paper is quantitative student data collected from kaggle. The data consists of 395 students with 33 features which includes School, School Support, Sex, Family Support, Age, Paid Class, Address, Activities, Family size, Nursery, Parent status, Higher Education, Mother Education, Internet, Father Education, Romantic, Mother Job, Family Relationship, Father Job, Free time, Reason for school, Go out with friends, Guardian, Workday Alcohol, Travel time, Weekend Alcohol, Study time, Health, Past Failures, and Absences, that might have an effect on student performance along with the target student grade values in secondary education of two Portuguese schools in the Mathematic subject.

Various machine learning models are used to determine that our finding would be less biased and gather a general conclusion, the models used are Multilayer Perceptron, Decision tree, and Random Forest. These three models were chosen primarily because the task being done is regression of a student grade given a set of specific features, this is for verification purposes whether the features chosen are the greatest factor for student performance, since if it does, the model should be able to predict student grade correctly, or at least close to the student actual grade.

In the Multilayer Perceptron (MLP) there can be more than one layer (the combinations of neurons). The first layer is the input layer, in the middle there will be a hidden layer and the last layer will be the output layer. We can modify the number of hidden layers in order to make the model do a more complex task to the dataset. Decision Tree is also a supervised model for classification, this algorithm built by splitting the dataset to root and node. The Decision Tree is used because it can perform well with little data preparation to large datasets and Random Forest, supervised classification algorithm. The difference between Random Forest and Decision Tree is in the process of finding the root, Random Forest's process of finding the root and splitting the dataset will run randomly.

In this study, about 80% of the total data have been used as the training sets and 20% used to test and evaluate the model. All the data has been cleaned first, handling the missing data or NaN values. Then, the dataset is used in Backpropagation Algorithm. The first step of the learning process is model initialization and move to forward propagate, in this step the model performance is checked. Next step, loss function, is a performance metric to calculate how good the model generates the input to target.

These models are developed based on several input factors and significantly fit with students' performance factors. Students' performance factors related to student's experience and performance in their academics can make the prediction more accurate. In conclusion, the impact of this work is to help the students in developing their academic performance and help the education system to identify students who need more attention in their study.

4. Previous Works

There are various papers that preceded or have similar topic with this paper, according to "Re" (Amirah Mohamed Shahiri, 2015), Decision tree are simpler than most methods, however fall short from Neural Network when predicting new dataset, while Neural Network are more effective because of its ability to detect all possible interactions between variables. This paper also warns readers that precautions are needed because data might be biased, and filters are required before feeding it into the model.

Another paper from Tarik A. Rashid, titled "Student Academic Performance using Artificial Intelligence" [5], supports that NN with 4 neurons is good enough to build a model to calculate weight of each factor, with the most important factor that affects student performance are student department and the tutor for that course, followed by school backgrounds, student address, and score of English, and then Father and Mother's degree and GPA does not affect student performance at all. However, this paper is rather biased because data collected only represent a certain group. These factors may differ from our finding because demographic is one of the factors that affects student performance. These findings about NN being "good enough" for most A.I. tasks is supported by another paper from Etebong Isong in 2018, where the team argues that ANN offers more robust technology, an easy tool for prediction, forecast, and modelling. Data obtained in this paper are from samples from students in the Department of Computer Science, Akwa Ibom State University, with almost 99% accuracy, one thing to consider is that the model maybe overfitted for that dataset, the authors had only tested the model using only 5 datasets, which is insufficient for model testing.

Another study titled by Ruangsak Trakunphutthirak in 2019 [6], argues that Random Forest has the highest accuracy between other Machine Learning methods such as Logistic Regression, Neural Network, Decision Tree, with 77.8% accuracy in 10-fold cross-validation dataset. It is not mentioned, however, the amount of nodes in the Neural Network model, so it could not be certain that Neural Network performs worse.

In another study, Mashael Al Iuhaybi et.al. in their paper (2018), argues that classification algorithms evaluate student's performance and identify key features affecting the prediction process based on a combination of three major attributes categories: Admission information, module-related data, and 1st year final grades. Two classification algorithms used to develop the prediction method are Naïve Bayes and C4.5 Decision Tree. Naive Bayes performs better than Decision Tree algorithm with 88.48% accuracy for Naive Bayes and 84.29% for Decision Tree, and the finding is that student qualification on entry has a high impact on students' academic performance. Fadhilah Ahmad et.al. in their paper, however, mentioned that Rule-based Learning has higher classification accuracy, when compared with Decision Tree, and Naive Bayes. There are several issues though, as the dataset size is relatively small due to incomplete or missing value, and the result may be overfitted or biased.

Alireza Ahadi et.al. in 2015 on their paper [7] states that previous paper work was based on somewhat static factors such as students' educational background and results from various questionnaires, while more recently, constantly accumulating data such as progress with course assignments and behavior in lectures has gained attention. This paper will result in early detection of students in need of assistance and provide a starting point for using machine learning techniques on naturally accumulating programming process data, and concludes that Random Forest has the highest accuracy evaluation, with example benefit is that safe students perform well on challenging harder tasks, while students that perform poorly benefit from rehearsal tasks.

"Performance of Student Prediction " by Mohammed Afzal Ahamed et.al. in 2019 [8], focused on data mining methods, which is one of preprocessing steps required before feeding the dataset into the machine. The aim for great data mining, stated, is that it can derive conclusions much easier and faster with good filtering. Supported by another paper by Yahya M. AlMurtadha et.al. in 2016 [9], using two data mining methods such as PCA (Principal Component Analysis) and Statistical Method using STATISTICA. It is worth noting that with great data mining methods, even finding the conclusion without using Machine Learning is possible, albeit doing the calculation manually, and a further study using Machine Learning may bring more efficient methods to combine both Machine Learning methods such as ANN or Naïve Bayes and PCA or STATISTICA Analysis.

A paper by Arto Hellas et.al. in 2018 [10] summarized many research papers around predicting academic performance by analyzing thousands of research paper, and concludes that there is indeed an increase in predicting easily attainable metrics such as individual course grade (38%), individual exam grade (14.7%), program retention or dropout (13.4%), GPA (12.2%), and assignment performance (11.4%), Machine Learning methods are often used but usually insignificant or not compared with other Machine Learning methods.

5. Result and Discussion

We used Pearson Correlation to obtain which factors affect student performance, and experimented with several Machine Learning methods to ensure that results obtained actually affects the model greatly, we tried using Multi-Layered Perceptron, Decision Tree, and Random Forest.

First, we determine correlation between each feature on the data that factors against G3 (which is the result of test score), If the feature has higher correlation toward the student performance, this means that every time the value changes, the student performance will be affected more significantly than changing other features, According to Friday Zinzendoff Okwonu et al [11], Pearson correlation formula can be derived as:

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \tag{1}$$

where rxy is correlation between the two variables, xi is the value of each x variable in sample to the i-th observation, yi is the value of each y variable in sample to the i-th observation, x̄ is the mean value of variable x, and Ȳ is the mean value of variable y. Formula 1 will help us obtain the correlation between each feature toward G3, as following:

Table 1. Correlation Between Each Feature Toward G3 Student Performance

Feature	Correlation	Feature	Correlation
School	-0.045017	School Support	-0.082788
Sex	0.103456	Family Support	-0.039157
Age	-0.161579	Paid Class	0.101996
Address	-0.105756	Activities	0.016100
Family size	-0.081407	Nursery	0.051568
Parent status	0.058009	Higher Education	0.182465
Mother Education	0.217147	Internet	0.098483
Father Education	0.152457	Romantic	-0.129970
Mother Job	-0.053363	Family Relationship	0.051363
Father Job	-0.013496	Free time	0.011307
Reason for school	0.120454	Go out with friends	-0.132791
Guardian	-0.054193	Workday Alcohol	-0.054660
Travel time	-0.117142	Weekend Alcohol	-0.051939
Study time	0.097820	Health	-0.061335
Past Failures	-0.360415	Absences	0.034247

From the result above, we can conclude that past failures have a really high correlation (-0.360415) toward G3 which means that students' grades are affected negatively from that student past failures, in such that if a student has ever failed a class before, their grade will be impacted negatively. The second-most highest correlation toward G3 is mother education (0.217147) which means that if the student's mother has better education, their performance will improve accordingly.

From the data above, we feed the data and train a model using 3 methods, which is Multi-layer Perceptron with 12 Neurons, Random Forest, and Decision Tree. From these models, the model will try and predict students' grade based on the features given. By using 9 of the highest-correlation features, which is Mother's Education, Higher Education, Father's Education, Failures, Age, Go out with Friends, Reason for school, Romantic, and Travel time, we then predict and calculate the RMSE (Root Mean Square Error), which derived from the following formula:

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

In formula 2, ŷ are the predicted values (the value from outcome predicted by the model), while y values are the observed values (the actual value given in the dataset), and n is the number of observations.

Each model will then output RMSE, which determines how far the result is between prediction and actual data, note that the lower RMSE value, the better the model on predicting the result. RMSE penalizes outliers (such that models predict too far away from the actual result) so models that are able to predict close to the actual target consistently are scored higher than models that usually predict very close to the target, but sometimes predict way off. The range of RMSE values is between 0 and 20, for the Normalized value, we divide all values by 20 to obtain a 0-1 range. Based on our own assumption, each model should not output a very different RMSE result, as all models are fed using the same feature dataset. The RMSE of each model is shown on the following table:

Table 2. RMSE Value for Each Model

Model	RMSE Value	Normalized RMSE Value
Multi-Layer Perceptron (12 Neurons)	4.32	0.216
Decision Tree	5.69	0.2845
Random Forest	4.52	0.226

The RMSE Result varies, which means overall, there is a deviation of around 4.32 to 5.69 between the predicted student grade and the actual student grade, on a scale of 0-20. which is pretty good overall with more than 300 datas, with Multi-Layer Perceptron performing the best with 4.32 RMSE value, and Decision Tree performing the worst with 5.69 RMSE value, this generally proves that the 9 selected features are affecting student performance greatly.

6. Conclusion and Future Works

From the Pearson correlation above, student performance is affected heavily based on several factors, such as past failures with a negative correlation of -0.360415, implying that students are more likely to fail when they have a past failure class record, and other negative factor which is age with a negative correlation of -0.161579, with older students tend to fail more than those who are younger, and several positive factors such as mother education with a positive correlation of 0.217147, and whether the student wants to pursue higher education with a positive correlation of 0.182465. The models in this paper (Multi-Layer Perceptron, Decision Tree, and Random Forest) are also able to predict student performance with excellent performance, with MLP 12-Neurons as the best model between them with 4.32 RMSE value, followed by Random Forest with 4.52 RMSE value, and Decision Tree with 5.69 RMSE value.

Future works including obtaining more data from around the world, since this data variation is still average and are limited to only students in two Portuguese schools, and using other parametric or advanced Machine Learning methods such as Deep Learning to explore more options for finding consistency, especially with the advancement of Artificial Intelligence, other statistical methods such as using MAE for calculating error might be helpful on finding other hidden factors within a huge dataset.

References

[1] Saputra, W. N. E., Supriyanto, A., Astuti, B., Ayriza, Y., & Adiputra, S. (2020). The effect of student perception of negative school climate on poor academic performance of students in Indonesia. International Journal of Learning, Teaching and Educational Research, 19(2), 279-291.https://www.researchgate.net/profile/Wahyu-Saputra-s2/publication/340293569_The_Effect_of_Student_Perception_of_Negative_School_Climate_on_Poor_Academic_Performance_of_Students_in_Indonesia/links/60f68d2916f9f3130095b471/The-Effect-of-Student-Perception-of-Negative-School-Climate-on-Poor-Academic-Performance-of-Students-in-Indonesia.pdf

[2] AH, K., Oldayo, A. A., & Fakai, A. A. (2020). Factors and Effects of Poor Background on the Students Academic Performance in Physics at Senior Secondary School in Birnin Kebbi Metropolis. https://www.saudijournals.com/media/articles/SJEAT_53_114-119.pdf

[3] Abu-Naser, S. S., Zaqout, I. S., Abu Ghosh, M., Atallah, R. R., & Alajrami, E. (2015). Predicting student performance using artificial neural network: In the faculty of engineering and information technology. http://dstore.alazhar.edu.ps/xmlui/bitstream/handle/123456789/391/27-05-2019-12.pdf

[4] Salamone, F., Belussi, L., Currò, C., Danza, L., Ghellere, M., Guazzi, G., ... & Meroni, I. (2018). Application of IoT and Machine Learning techniques for the assessment of thermal comfort perception. Energy Procedia, 148, 798-805. https://www.researchgate.net/publication/328590706_Application_of_IoT_and_Machine_Learning_techniques_for_the_assessment_of_thermal_comfort_perception

[5] Rashid, T. A., & Aziz, N. K. (2016). Student Academic Performance Using Artificial Intelligence. ZANCO Journal of Pure and Applied Sciences, 28(2), 56-69. https://www.researchgate.net/publication/291262353_Student_Academic_Performance_Using_Artificial_Intelligence

[6] Trakunphutthirak, R., Cheung, Y., & Vincent (2019). A Study of Educational Data Mining: Evidence from a Thai University. AAAI-19, 734-741. https://researchmgt.monash.edu/ws/portalfiles/portal/286376593/262329453_oa.pdf

[7] Ahadi, A., Lister, R., Haapala, H., & Vihavainen, A. (2015). Exploring Machine Learning Methods to Automatically Identify Students in Need of Assistance. ICER '15: Proceedings of the eleventh annual International Conference on International Computing Education Research, 121-130. https://dl.acm.org/doi/10.1145/2787622.2787717

[8] Ahamed, M. A., Chaisanit, P., & R., M. T. (2019). Performance of Student Prediction. IJCSMC, 8(6), 45-50. https://ijcsmc.com/docs/papers/June2019/V8I6201909.pdf

[9] AlMurtadha, Y. M., Alhawiti, K. M., Elfaki, A.O., & Abdalla, O. A. (2016). Factors Influencing Academic Achievement of Undergraduate Computing Students. International Journal of Computer Applications, 146(3), 23-28. https://www.ijcaonline.org/archives/volume146/number3/almurtadha-2016-ijca-910658.pdf

[10] Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting Academic Performance: A Systematic Literature Review. ITiCSE 18 Companion, 175-199. https://circuit.bcit.ca/repository/islandora/object/repository%3A864/datastream/PDF/download/citation.pdf

[11] Okwonu, F. Z., Asaju, B. L., & Arunaye, F. I. (2020, September). Breakdown analysis of pearson correlation coefficient and robust correlation methods. In IOP Conference Series: Materials Science and Engineering (Vol. 917, No. 1, p. 012065). IOP Publishing. https://iopscience.iop.org/article/10.1088/1757-899X/917/1/012065/pdf