Student' s Academic Performance Prediction Based on Machine Learning Regression Models

Sijian Lyu

School of Mathematical Sciences, Beijing Normal University, Beijing, 100875, China
202011130136@mail.bnu.edu.cn

Abstract. The utilization of machine learning methods for predicting student grades has emerged as a valuable approach to assessing the educational advancement of academic institutions, driven by the rapid evolution of these techniques. While prior studies have predominantly encompassed data from various facets, this research specifically focuses on forecasting students' academic performance based solely on their past scores. The dataset employed in this study is obtained from Kaggle and comprises grades attained by college students majoring in computer science. Four distinct machine learning models, including linear regression, support vector regression, k-nearest neighbors, and random forest, are employed to predict the students' scores using regression techniques. Notably, the paper streamlines the problem through data preprocessing, initially eliminating missing data, and subsequently applies the aforementioned models to predict student performance. Furthermore, the parameters are adjusted to get the best performance. From the outcome in this study, random forest is proved as the best model to predict the student's grades in this dataset. This work finally shows predicting student's future progress using the four models just by its past scores was acceptable and reasonable, also gives some possible reasons for the different outcomes from diverge algorithms.

Keywords: Machine Learning, Student Performance Prediction, Random Forest Algorithm.

1 Introduction

The academic achievement of students holds significant importance for both educators and parents, commonly manifested through their grades. The significance of making comparisons, particularly with regards to one's own previous achievements, holds substantial value. Furthermore, analyzing student performance serve as a crucial means for teachers will provide education with higher qualities. Currently, the prediction of student's performance is frequently based on teachers' subjective perceptions of their students' past performance, which is less accurate and relies heavily on teachers' subjective impressions.

In the past few decades, machine learning technology has advanced at a rapid pace. According to Bengio et al. [1], Machine Learning is a part of Artificial Intelligence. It's a process to generalize the existing knowledge to new world that is unknown with

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_29

computers. The utilization of machine learning technologies in the field of student grade prediction has emerged as a promising avenue owing to the notable advantages exhibited by machines, including enhanced computational capabilities, heightened stability, and increased objectivity. By leveraging machine learning algorithms, it becomes feasible to evaluate students' academic performance by incorporating both dynamic and static data, thereby enabling a comprehensive assessment that takes into account various pertinent factors.

In the recent years, prior work mainly focused on using data on different factors to predict student's performance. In particular, intelligence quotient (IQ) can be significant predictors of excellent academic performance [2], which has been widely accepted. Children with high IQ can absorb knowledge more easily and achieve higher grades. Thousands of students in China proved that family background have a positive impact on students' academic achievement [3], because children from better family background have better education opportunities, and parents have higher expectations on children, so they are more involved in student's education, which leads to better academic performance. In [4], MacCann C. et al. suggested that emotional intelligence (EI), may benefit students' academic performance. Children with higher EI are expected to control their negative emotions better so they can more easily get through difficulties they encounter while studying. Also, students' grades can be predicted by their daily habits [5], for example, sleeping qualities, physical exercises, eating habits, mobile devices usage, cognitive control, etc. The results indicate that lifestyle habits are able to predict students' academic performance and that students with a healthier lifestyle gets a better grade. Besides, machine learning techniques have some great improvement and were used in some articles to predict student's academic performance. Xu et al. showed that support vector machine (SVM) can predict the student's academic performance by online habits [6]. Artificial neural network (ANN) was used to predict student's grade point average (GPA) in 2020 [7]. And also, in [8], Extreme Gradient Boosting (XGBoost) had given a classification solution based on a dataset with different features, reaching an outstanding result.

All of these previous articles have given some different aspects that can influence students' performance. However, an essential consideration in practical implementation is the cost associated with data collection, particularly the expenditure of time. Given that college professors are occupied with various responsibilities, including academic research and publication, their access to time-sensitive data may be limited. Consequently, the objective should center on predicting students' grades solely based on their previous academic performance, which constitutes the most readily available and directly accessible data for college professors.

In this regard, this study aims to predict the grades of college students based on machine learning techniques. Several typical regression algorithms are used, including linear regression (LR), support vector regression (SVR), k-nearest neighbors (KNN) and random forest (RF).

With these machine learning methods used in this study, R2 score were used to describe the quality of the regression, RF get the best result of 0.767, it's a pretty high

value. It leads to a conclusion that it's reasonable to predict students' academic performance only by their past grades.

2 Method

2.1 Dataset and Preprocessing

This study employed Grades of Students datasets from Kaggle [9]. comprising 571 instances and 43 attributes encompassing grades from 41 distinct courses. Following data preprocessing procedures, the dataset was refined to include records from 557 students and encompass grades solely from 20 computer science (CS) courses. To preprocess this data, the paper first involved identifying and addressing missing data, it was mainly from 14 students who dropped from their major. Those were dropped and also only CS major courses were left. This work then transformed the level data in text into GPA with GPA 4.0 algorithm. The target of this dataset is to predict student's average grade in the senior year with the data from the first three years by using regression methods. Therefore, this prediction task entailed 15 features, representing grades attained in 15 courses throughout the first three years of college, while the average score in the senior year was computed based on the remaining five courses. This study subsequently separated the training set and the test set based on the 70:30 ratio.

2.2 Machine Learning Algorithms

Introduction. To achieve the goal, the paper used several different regression models with machine learning including LR, SVR, KNN, RF.

Starting with LR, which serves as an introductory regression method in machine learning. LR used a linear function to predict the value of each data point. The resulting visualization depicts a straight line on a two-dimensional plane, aiming to minimize the mean error across the entirety of the dataset.

SVR is an extension from support vector classification. SVR involved an ε-tube that was the area around the function with the distance less than ε [10], and it was insensitive which means the model only need to consider the support vectors outside the tube and will get a new optimization problem. Solving this new problem needs to maximize the width of the tube(ε), and at the same time to minimize the loss between the values of the new function created and the original values. And achieving this is how SVR works.

KNN is also a widely used machine learning model to do regression. Assigning an integer value K at first, which decided how many nearest neighbors would be considered in the model. Then calculating the distance between the unknown point and others, to locate the k nearest points. Letting the k nearest points (which is the 'neighbor') decide what value the unknown point should have. While this is changed to a new optimization problem, it needs to minimize the error between the neighbors' value and the new one.

Another model used for regression is RF. RF started with building some decision trees, then combined them together. Each decision tree is a predictor, which depends

on the random vector independently with the same distribution [11]. During the construction of these decision trees, the discrepancy between the projected values and observed values is assessed at each node. The minimal error is then chosen for both individual nodes and the overall tree, thereby furnishing an optimal estimation of the desired value.

Implementation details. Scikit-learn was imported for implementing machine learning models in this study [12]. A pivotal aspect to consider is the determination of the error metric for comparing values, necessitating the adoption of appropriate evaluation metrics. In this study, the Euclidean distance, defined as the square root of the sum of squared differences, was employed as the prevailing norm for assessing the dissimilarity between values.

As in the introduction of these models, some important parameters were defined, for example: ε in SVR, k in KNN, and n for the total tree number in RF. LR had no special parameter, because everything is defined naturally.

In SVR, adjustment of the kernel function was the first thing to do, the default one is ‘rbf’, which is the radial basis function kernel. To the experimental life experience, rbf had the best result while doing regression for small features (i.e., in this work, 15) and not very large datasets. Then for the value of ε, 0.19 was picked as it got better performance than the default value 0.1.

The selection of an optimal value for the parameter 'K' in KNN algorithm was a particularly intriguing aspect of this investigation. This crucial step holds the key to enhancing accuracy in KNN-based analyses. It is worth noting that the influence of 'K' is limited to a small local neighborhood, as it governs the smoothness of the decision boundary. In instances where 'K' is excessively small, overfitting concerns arise, as the outcome becomes susceptible to the influence of noisy data points. While such a scenario may yield highly precise results on the training set, it proves detrimental to the generalizability of the model. Conversely, opting for an excessively large 'K' value can lead to underfitting, hindering the attainment of an ideal outcome. This paper used K-fold cross-validation to find the best K for the model. It randomly divides the dataset into K groups, choosing one group each time for test and repeat it for K times. Calculating the minimum error and get the best K. In this work, K is selected as 19.

In RF, how many trees are there in the forest needs to be determined. As the same with selecting k, n (for number) could not be too small or too big. The computational efficiency of model execution is a critical concern, as excessively large values can impede processing speed. In order to optimize performance, the experiment implemented a stopping criterion wherein the incremental increase in the quantity under consideration was evaluated for any significant improvement [12]. Consequently, it was determined that the optimal value for the parameter 'n' was 98, deviating from the default value of 100, based on the observed outcomes.

3 Results and Discussion

The performance of these models was assessed by three indicators in this paper as shown in Table 1. R2 described the effect when predicting the dependent variable using the independent one. In the context where the R2 attains a value of unity, the predictive performance achieves its optimal state. The error term is defined as the disparity between the value from prediction and the true value, thereby enabling the calculation of Mean Absolute Error (MAE) and Mean Squared Error (MSE) as measures of central tendency for the collective (squared) individual errors. Comparing the four different regression models, this study found the best model to predict the grades of student's dataset is RF, with the highest R2 score 0.7677 and the lowest mean error, although the other models had similar results.

Table 1. The performance of diverse models evaluated by various metrics.

Model	Performance		
	R2	MSE	MAE
LR	0.7390	0.0845	0.2199
SVR (kernel='rbf', ε=0.19)	0.7542	0.0796	0.2187
KNN (n_neighbors=19, weights='distance')	0.7417	0.0836	0.2170
RF (n_estimators=98)	0.7677	0.0752	0.2131

Each of the four employed models exhibited R2 scores surpassing 0.7, indicative of their commendable aptitude for predicting student academic performance within the present context. Nevertheless, subtle differentiations manifested within their respective outcomes. LR, owing to its simplistic nature, produced the least favorable results. KNN was not giving an impressive result, but it cost less time than other models since it did not actually train the training set. Also, it got the second lowest MAE which was from its principles. RF got the best score in predicting student's grades, primarily attributable to its superior performance albeit at the expense of increased computational time and algorithmic complexity when juxtaposed against alternative approaches.

Comparing with previous results in predicting student's performance, this paper shown a result that is not too powerful. Like in [13], Multiple Linear Regression (MLR) method was used and got a score 0.82, which is better than the one here. But their datasets have 97 attributes, more than 6 times of the one in this paper, while including attributes from different aspects. So, this work stood out because it didn't need too many attributes and all of them were just student's past grades, this study can still reach an acceptable result.

4 Conclusion

Student's academic performance is an important aspect to evaluate an educational institution, so predicting student's grade is an effective way to ensure learning quality. A model trained with LR, SVR, KNN and RF is developed in this study with the data from students majored in CS. Dataset is preprocessed that attributes are finally cut down to just 15 key factors, parameters in the model were adjusted to reach a better observation outcome. This study demonstrated that predicting student's academic performance just by student's past grade with machine learning algorithms made sense. The effectiveness and adequacy of the model is checked using three indicators. The adequacy of the constructed model was ascertained through the examination of three distinct indicators, ultimately yielding favorable outcomes. Remarkably, RF is the algorithm emerged as the top performer, having the most outstanding results; however, avenues for further improvement remain potential and apparent. So, further research will focus on harnessing more enriched datasets to achieve better precision and accuracy in predicting, or to make generalization and make predicting student's academic performance more applicable and convenient. Furthermore, other machine learning methods can be used for prediction in extending this study and provide additional insights, in particular neural networks.

References

1. Bengio, Y., Lecun, Y., Hinton, G.: Deep Learning for AI. Communications of the ACM, Vol. 64 No. 7, 58-65 (2021).
2. Quilez-Robres, A., González-Andrade, A., Ortega, Z., Santiago-Ramajo, S.: Intelligence quotient, short-term memory and study habits as academic achievement predictors of elementary school: A follow-up study,Studies in Educational Evaluation, Volume 70 (2021).
3. Li, Z., Qiu, Z.: How does family background affect children's educational achievement? Evidence from Contemporary China. The Journal of Chinese Sociology volume 5, Article number: 13 (2018).
4. MacCann, C., Jiang, Y., Brown, L E, Double, K. S, Bucich, M., and Minbashian, A.: Emotional intelligence predicts academic performance: A meta-analysis. Psychol. Bull. 146(2):150-186. Feb (2020).
5. Dubuc, M. M., Aubertin-Leheudre, M., Karelis, AD.: Lifestyle Habits Predict Academic Performance in High School Students: The Adolescent Student Academic Performance Longitudinal Study (ASAP). Int J Environ Res Public Health. Dec 29;17(1):243 (2019).
6. Xu, X., Wang, J., Peng, H., Wu, R.: Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Computers in Human Behavior, 98(January), 166–173 (2019).
7. Musso, M. F., Hernández, C. F. R., Cascallar, E. C.: Predicting key educational outcomes in academic trajectories: A machine-learning approach. Higher Education, 80(5), 875–894. (2020).
8. Ojajuni, O., et al.: Predicting Student Academic Performance Using Machine Learning. In: , et al. Computational Science and Its Applications – ICCSA 2021. ICCSA 2021. Lecture Notes in Computer Science(), vol 12957 (2021).

9. Kaggle.: Grades of students. https://www.kaggle.com/datasets/ssshayan/grades-of-students (2022)
10. Awad, M., Khanna, R.: Support Vector Regression. In: Efficient Learning Machines. Apress, Berkeley, CA (2015).
11. Pedregosa et al.: Scikit-learn: Machine Learning in Python. JMLR 12. pp. 2825-2830. (2011).
12. Oshiro, T. M., Perez, P. S., Baranauskas, J.A.: How Many Trees in a Random Forest?. In: Perner, P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2012. Lecture Notes in Computer Science(), vol 7376 (2012).
13. Kumar, A., Eldhose, K. K., Sridharan R., and Panicker, V. V.: Students' Academic Performance Prediction using Regression: A Case Study, 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, pp. 1-6. (2020).