Predict student’s performance based on machine learning algorithms

Yuqi Han

Macau University of Science and Technology, Macau SAR, 999078, China

hanyuqi49777@163.com

Abstract. Several studies have used models for students’ academic performance prediction to improve teaching quality. The aim of this study is to use machine learning algorithms to forecast students’ performances using their daily study behaviors and the extent of parents’ concerns about their children’s studies to generate synchronous predictions with daily teaching activities. The data includes study attitudes, behaviors, demographic features, and parents’ concerns about the students. The preprocessed dataset after feature engineering was used to train the models (i.e Support Vector Machine, Decision Tree, Random Forest, And K Nearest Neighbor). Random Forest has the best performance among the algorithms applied. The impact of students’ daily study behavior is highly related to academic achievement, and parents’ impact is also an influencing factor on children’s performances. This study could encourage and motivate parents to care more about their children’s studies with their favorable actions and behaviors. Besides, this study would help students realize the importance of their daily performance and realize it is essential to their final exam grades.

Keywords: student’s performance, education data mining, machine learning.

1. Introduction

A student's primary goal is to succeed academically. Academic performance is whether a student could achieve his or her short and long-term study goals.

The degree to which a student has effectively attained both short- and long-term educational goals is referred to as academic performance [1].

According to students’ different academic performance levels, schools, and teachers could adopt different teaching approaches and strategies to improve teaching quality and activation students’ potential. As a result, realizing which student belongs to which academic achievement group is essential for this purpose. In the past, teachers used to predict students’ performance based on the teacher’s teaching experience, while the predictive method is subjective. So the application of data mining used in predicting students’ performance prospers recently. Student performance prediction is an integral part of Education Data Ming (EDM).

Educational data mining is to use data mining approaches and algorithms to solve problems in the education field [2]. It is helpful for institutions to discover hidden educational information to provide efficient advice for students [3]. The four basic objectives of educational data mining are to advance scientific knowledge of learning and learners, construct or improve domain models, investigate the effects of educational support, and forecast future learning behavior in students [2].

Some studies have indicated that several variables significantly affect students' academic advancement, making them useful for forecasting students' success. Individual differences like academic aptitude and personality, non-cognitive traits such as academic aptitude and personality, non-cognitive traits such as perceptions, behavior patterns, and individual tactics, inspiration, self-control, and participants' involvement in extracurricular activities are some of the factors that have been proven to have an impact on student's performance at different levels in schools. Using their prior academic success, students' performances have been predicted in several studies with a high degree of accuracy.

But it uses information based on the past, which means there is a lag between the prediction result and real academic performance at present. Besides, it could not accurately predict the subject that students first take in a semester. For example, in China, some senior high school students are first exposed to biology in their first year but they are not finished in their first year of high school, which means that they do not know whether they perform well in this course and whether they are suitable for this course based on the grades, and they have to face the problem of choosing liberal arts which do not include biology as an exam subject to take in College Entrance Examination or science subjects to study. At that time, predicting students using their prior academic performance is not realistic. So in this study, I use the information of students' daily instant study behavior and performance to predict to reduce the lagging problem through machine learning algorithms such as SVM, Decision Tree, Random Forest and KNN. This analysis could assists teachers in dynamically changing their teaching strategies based on the academic group that students belong to improve the quality of its instruction and activate students' potential.

2. Literature review

Some studies are related to educational data mining, some of them focus on predicting students' academic achievements such as GPA, dropout rate, etc, while some of them concentrate on analyzing the factors that may influence the student's performance.

Asif et al. conducted a study on predicting students' performance using the marks that students get for all courses that are taught, which could be divided into two parts [4]. The first part is to predict the student's performance when a four-year study program is accomplished. Besides, this study also identifies the student's progress in the study and tries to combine the progressions with predictions. Ahmad, Z., & Shahzadi also uses previous degree marks as variables, but they added some students' behavior in a study such as study habits, learning skills, hardworking, and academic interaction, etc as features to identify whether the student is in the risk group [5]. He proposed a Message Passing Neural Network model, which reached a maximum accuracy of 95. Dabhade et al. used the previous semester's GPA and attributes based on questionnaire survey to predict the sixth semester's GPA in the final year [6]. In this study support vector regression linear, the algorithm performed best. Olabanjo et al. built a Radical Basis Function Neural Network model to predict students' performance based on their past academic behaviors, and their cognitive and psychomotor abilities. This model has an overall accuracy of 86.59%, and an AUC score of 94% [7].

Hoffait and Schyns identified the students most likely to fail by combining environmental factors and the student information requested during enrollment. They discovered that applying DM techniques allowed for more accurate classification of kids who might have issues [8].

Fernandes et al. predicted student performance with demographic characteristics, student average grade, and absenteeism using Gradient Boosting Machine. They discovered that other demographic characteristics, such as "neighborhood," "school," and "age," might also indicate how successful or unsuccessful a student is [9]. Additionally, Cruz-Jesus et al [10]. examined several demographic characteristics to forecast students' academic success. The whole forecast was based on the person's academic standing and the academic environment in which they are positioned. To categorize levels of grade point average, academic retention, and degree completion outcomes, Musso et al. developed multiplier perceptron artificial neural network models with a backpropagation algorithm. They discovered that a student's learning practices could have a significant impact on their grade point average [11]. Coping mechanisms are the best predictors of degree completion, and background knowledge has the biggest influence on determining whether a student would drop out or not. Rahman et al. tried to find out the effect of co-curricular on students' performance [3]. The author experimented with several algorithms, voting perceptrons, logistic regression, multilayer perceptrons, and the random forest classifier. He chose Logistic Regression to do the task which has the highest accuracy of 99.5294%. By using Logistic Regression, the research result showed that the extracurricular activities that could improve the academic curriculum positively influenced the pupil's academic performance [3]. Xu et al. paid attention to the impact of students' online usage behavior and condition on their academic achievements with the Decision Tree, Neural Network, and Support Vector Machine algorithms. The study concluded that the frequency of getting access to the Internet has a positive effect on the student's academic performance, while the Internet traffic volume plays a negative role in influencing a student's academic achievements. So behavior discipline is essential to a student's academic success [12]. Liao and wu researched the impact of social media distraction and peer learning engagement on social media on students' academic achievements. Some machine learning model (Random Forest, Support Vector Machine, Fully Connected Neural Network, and Long Short Term Memory algorithms) applied has identified that peer learning engagement predicted academic achievements [13].

3. Methodology

This section describes the dataset, data preprocessing, and some supervised learning algorithms applied.

3.1. Dataset

In this study, the Tianchi Big Data Competition student performance prediction dataset is used. The dataset has 17 attributes and 481 instances. Among the attributes, there are some demographic features such as gender, and nationality included. Besides, students' behavior in the study and the frequency of some co-curricular activities is collected. In addition, their parent's attitudes and behavior are also included in this dataset.

3.2. Data preprocessing

Data preprocessing is an essential upper stage for data mining since most of the data is not suitable and clean enough for an algorithm to execute, which means most of the data has noises, missing values, and some features types that are difficult for the computer to process. This dataset, fortunately, does not have missing values so it is no need to fill them in. Typically, the average of that attribute could be used when filling in the null values but it also depends on the feature type. This dataset includes several object-type features, which are hard for a machine to handle. So The author conducted feature engineering, using one-hot coding to transform discrete categorical features into dummy variables to make the computer could read and process the data.

3.3. Machine learning algorithms employed

Data Ming has two objectives, one is the predictive model, to generate predictions based on existing data, and the other is the descriptive model, to explain the relationships between variables. Since the task of this study is to predict students' performance in which group (low, medium, and high), some supervised learning algorithms which could perform classification tasks are employed. When a class, also known as a label or a discrete value, is predicted using a classifier, this is referred to as classification [14]. A classifier builds a classification model using training data that includes objects defined by the values of various attributes, one of which is designated as the class. The created model should closely resemble the training data and accurately predict the class or label of unknown data, i.e., the test data, which is a distinct set of data not utilized to generate the classifier [14].

When the focus is on analyzing the factor that impacts mainly the prediction, statistical methods can be employed [15]. However, the focus of this study is to make predictions, we care more about prediction accuracy, and support vector machines, decision trees, and random forests are more efficient and give more accurate results [16]. These algorithms are like black boxes, which means the outcome that they deliver is not interpretable [4]. So the supervised learning algorithms to make predictions that I used are Support Vector Machine, Decision Tree, Random Forest, and K Nearest Neighbor.

Support Vector Machines are a useful and well-liked method for learning classification [17-18]. As a constrained quadratic program-mining issue, a learning support vector machine is modeled. However, it is an unrestricted empirical loss minimization with a penalty term for the learned classifier norm in its natural form.

A Decision Tree is a process of classifying data through a series of rules. It adopts a top-down recursive approach, compares attribute values at nodes inside the decision tree, and branches down according to different attribute values, to achieve the effect of classification

In essence, random forests are a subset of ensemble learning, which integrates numerous decision trees into a single forest and uses them to predict the outcome. The CART decision tree algorithm, developed by Breiman et al. significantly decreased the amount of processing required to repeat dichotomous data for classification or regression [19]. Breiman merged classification trees into random forests by randomly using variables (columns) and data (rows), creating numerous classification trees, and then summarizing the classification tree findings [20].

The KNN algorithm categorizes the sample to be classified by comparing the category of K neighbors that are closest to the sample to be classified. The basic idea is to calculate the distance between the sample to be classified and the training sample, choose the K training samples that are closest to the sample to be classified and choose the category of the sample to be classified that most closely matches the majority of the K samples.

To improve the accuracy and avoid overfitting, K-fold cross-validation is applied. In this study, k = 5. By iteratively detecting if the test performance differs depending on the samples you used for training and testing, the K-Fold cross-validation enhances machine learning models. We can better assess the model performance by running the train/test comparison numerous times [7].

4. Results

The dimension of the dataset was 481 × 17. After the feature engineering of the dataset, the preprocessed dataset has 481 × 73 dimensions, as some object features were transformed into dummy variables using the one-hot technique. The input variables were some information about students' behavior in school, their demographic information, and information about parents' attitudes and attentiveness toward their student's studies and school. These features were used to predict which performance class a student belongs to (Low, High, and Medium). The algorithms applied and two experiments are given in Table 1. The performance of the models applied showed significant differences. Table 2 shows the average accuracy of employed algorithms in both experiment A and experiment B. And Table 3 and Table 4, illustrate the performance of algorithms in experiment A and experiment B respectively.

Table 1. Experiments in model building.

Experiment	Description	Technique
A	Full features	All algorithms employed
B	Full features without parents intervention	All algorithms employed

Table 2. Accuracy of employed algorithms(continue).

Algorithms	Accuracy
Support Vector Machine	68.25%
Decision Tree	82.30%
Random Forest	83.86%
K Nearest Neighbor	60.42%

When the full features were used with the parent's intervention in students' study included, among these four models, Random Forest gave the highest accuracy of 85.42%, 86.33% precision, 84.33% recall, and 85.33% F1-score, even if the accuracy of Decision Tree (84.38%) is very close to the Random Forest, while K Nearest Neighbor had the lowest accuracy with just 60.42% of accuracy. When extracting the parent's intervention features, and using other attributes to do the prediction, Random Forest also had the best performance with 82.29% accuracy, 82.33% precision, 81% recall, and 81% F1-score. The K Nearest Neighbor had the worst performance just like the results in experiment A displayed. In this experiment, the accuracy of the Support Vector Machine rose to 69.29% from that of experiment A, which scored 66.71% in accuracy.

Table 3. Performance of algorithms in Experiment A.

	Accuracy	Precision	Recall	F1-score
Support Vector Machine	66.71%	68%	68.33%	67.67%
Decision Tree	84.38%	83.33%	85.67%	84%
Random Forest	85.42%	86.33%	84.33%	85.33%
K Nearest Neighbor	60.42%	61.33%	63.33%	61%

Table 4. Performance of algorithms in Experiment B.

	Accuracy	Precision	Recall	F1-score
Support Vector Machine	69.79%	69%	70%%	70%
Decision Tree	80.21%	79.67%	80.67%	80%
Random Forest	82.29%	82.33%	81%	81%
K Nearest Neighbor	60.42%	61.33%	63.33%	61%

5. Discussion

The goal of this study was to assess how well the algorithms used to forecast students' academic success performed utilizing their demographic data and regular learning habits as predictors. The inclusion of parents' attitudes and levels of interest in their kids' education as well as their feelings of satisfaction with the school takes the investigation of whether parents' attitudes and actions have an impact on their children's academic achievement one step further. Experiment A used all the features. Experiment B evaluated the impact of parents' intervention on students' academic performance production. By comparing the results of Experiment A and Experiment B in Table 3 and Table 4, the Random Forest algorithm in Experiment A gave the most accurate result. And among the algorithms applied, by reducing the parent's intervention, the accuracy of Random Forest and Decision Tree algorithms declined, the accuracy of the Support Vector Machine increased, and K Nearest Neighbor accuracy kept unchanged.

Table 5. Confusion matrices in Experiment A.

SVM		Actual			Class Precision	DT	Actual			Class Precision
		H	L	M			H	L	M	
Predicted	H	16	1	11	57.1%	H	18	0	4	81.8%
	L	0	17	5	77.3%	L	0	22	3	88%
	M	6	8	32	69.6%	M	4	4	41	83.7%
RF		Actual			Class Precision	KNN	Actual			Class Precision
		H	L	M			H	L	M	
Predicted	H	17	0	4	81%	H	16	2	16	47.1%
	L	0	25	4	86.2%	L	0	17	7	70.8%
	M	5	1	40	87%	M	6	7	25	65.8%

Table 6. Confusion matrices in Experiment B.

SVM		Actual			Class Precision	DT	Actual			Class Precision
		H	L	M			H	L	M	
Predicted	H	15	1	6	68.2%	H	18	1	6	72%
	L	0	19	9	67.9%	L	0	23	6	79.3%
	M	7	6	33	71.7%	M	4	2	36	85.7%
RF		Actual			Class Precision	KNN	Actual			Class Precision
		H	L	M			H	L	M	
Predicted	H	17	0	5	77.3%	H	16	2	16	47.1%
	L	0	24	5	82.8%	L	0	17	7	70.8%
	M	5	2	38	84.4%	M	6	7	25	65.8%

Tables 5 and 6 display the resulting confusion matrices. Use the algorithm with the best performance in Table 5 as an example to comprehend these confusion matrices. It is discussed how the classifier "RF" handles confusion. Out of the 22 (i.e., 17 + 5) actual class H (i.e., high-level) students in the first column, the classifier correctly identified 17 as belonging to that category; the recall for class "H" is 72.7% (i.e., 17/22). 28 (16 + 1 + 11) were projected as being in class "H" in the first line; the accuracy for class "B" is 16/28, or 57.1%. For the other classes, the remaining columns and lines are comparable. Each matrix's diagonals hold all of the accurate predictions.

6. Conclusion

This analysis revealed that a machine learning classifier can identify student performance. This suggests that a student's regular study attitude and behavior can predict whether they would succeed or fail on their exams, which suggests that parents, schools, or other institutions may be able to deduce whether a student will succeed or fail academically based on their regular study actions. Besides, this study showed that parents should pay more attention to their children's studies, and just urge their children to achieve high academic performance. Parents who are more attentive to their children's study, the children are more likely to achieve the academic grade that they supposed to be. The reduction of accuracy of Experiment B (Full features without parent's intervention) could support that argument.

One impact of this study is that student's academic success or failure could be predicted before they have examinations, their academic achievements could be estimated by instructors and academic organizations through their daily study behaviors and attitudes, and parent's attitudes toward student's study instead of waiting for the scores to predict next semester or future performance. This will also help parents and teachers know the direction of how to teach students, provoke students' interest to study, or improve students' class engagement.

Another impact of this study is that even if the academic achievement is highly related to the student himself, parents should be more attentive and care more about their children's studies, Even if parent's personal information would not influence children's performance, their concern of it could help children to get the grade which is competent to their capability. This study could raise parents' awareness of children's education. Besides, students, instructors, and academic organizations could also derive benefits from it since the prediction could help identify which teaching strategies are more suitable for a different student group to improve teaching quality, and prevent future academic failures.

More deep learning and machine learning models will be used in the future to further enhance the outcomes. Deep learning is also a direction like RNN, which could adaptably iterate the hyperparameters to utilize the input features. And I hope to get more input attributes to make it more precise and robust for the prediction as Domingos concluded that the essential for a machine learning project to be successful is the features chosen and used [21].

References

[1] Camacho-Morles J, Slemp G R, Pekrun R, Loderer K, Hou H and Oades L G 2021 Activity achievement emotions and academic performance: a meta-analysis Educ. Psychol. Rev. 33 1051–95

[2] Baker R S J d and Yacef K 2009 The state of educational data mining in 2009: a review and future visions J. Educ. Data Min. 1 3–17

[3] Rahman S R, Islam M A, Akash P P, Parvin M, Moon N N and Nur F N 2021 Effects of co-curricular activities on student’s academic performance by machine learning Curr. Res. Behav. Sci. 2 100057

[4] Asif R, Merceron A, Ali S A and Haider N G 2017 Analyzing undergraduate students’ performance using educational data mining Comput. Educ. 113 177–94

[5] Ahmad Z and Shahzadi E 2018 Prediction of students’ academic performance using artificial neural network Bull. Educ. Res. 40 157–64

[6] Dabhade P, Agarwal R, Alameen K P, Fathima A T, Sridharan R and Gopakumar G 2021 Educational data mining for predicting students’ academic performance using machine learning algorithms Mater. Today Proc. 47 5260–7

[7] Olabanjo O A, Wusu A S and Manuel M 2022 A machine learning prediction of academic performance of secondary school students using radial basis function neural network Trends Neurosci. Educ. 29 100190

[8] Hoffait A S and Schyns M 2017 Early detection of university students with potential difficulties Decis. Support Syste. 101 1–11

[9] Fernandes E, Holanda M, Victorino M, Borges V, Carvalho R and Van Erven G 2019 Educational data mining: predictive analysis of academic performance of public school students in the capital of Brazil J. Bus. Res. 94 335–43

[10] Cruz-Jesus F, Castelli M, Oliveira T, Mendes R, Nunes C, Sa-Velho M and Rosa-Louro A 2020 Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country Heliyon 6 e04081

[11] Musso M F, Hernández C F R and Cascallar E C 2020 Predicting key educational outcomes in academic trajectories: a machine-learning approach High. Educ. 80 875–94

[12] Xu X, Wang J, Peng H and Wu R 2019 Prediction of academic performance associated with internet usage behaviors using machine learning algorithms Comput. Hum. Behav. 98 166–73

[13] Liao C H and Wu J Y 2022 Deploying multimodal learning analytics models to explore the impact of digital distraction and peer learning on student performance Comput. Educ. 190 104599Sarkar S, Agrawal S, Baker T, Maddikunta P K and Gadekallu T R 2022 Catalysis of neural activation functions: adaptive feed-forward training for big data applications Applied Intelligence vol 52.12 p 13364-83

[14] Han J and Kamber M 2006 Data Mining Concepts and Techniques 2nd edn (San Francisco, CA: Morgan Kaufmann)

[15] Arias Ortiz E and Dehon C 2013 Roads to success in the Belgian French community's higher education system: predictors of dropout and degree completion at the Université Libre de Bruxelles Res. High. Educ. 54 693–723

[16] Huang S and Fang N 2013 Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models Comput. Educ. 61 133–45

[17] Cristianini N and Shawe-Taylor J 2000 An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods (Cambridge: Cambridge University Press)

[18] Vapnik V N 1998 Statistical Learning Theory (Hoboken, NJ: John Wiley & Sons)

[19] Breiman L 2001 Random forests Mach. Learn. 45 5–32

[20] Breiman L, Friedman J H, Olshen R A and Stone C J 1984 Classification and Regression Trees (New York, NY: CRC Press) pp 246–280

[21] Domingos P 2012 A few useful things to know about machine learning Commun. ACM 55 78-87