Education and Information Technologies (2023) 28:12039–12057
https://doi.org/10.1007/s10639-023-11672-1

A comparative study on student performance prediction using machine learning

Yawen Chen[1] · Linbo Zhai[1]

Received: 10 October 2022 / Accepted: 14 February 2023 / Published online: 4 March 2023

Abstract

Accompanied with the development of storage and processing capacity of modern technology, educational data increases sharply. It is difficult for educational researchers to derive useful information from much educational data. Therefore, educational data mining techniques are important for the development of modern education field. Recently, researches have demonstrated that machine learning, as an important tool for data mining, has shown promising performance in educational applications, especially in student performance prediction. However, few studies comprehensively compare existing machine learning methods in educational data. Moreover, most current studies only focus on a single type of educational data for student performance prediction. In this paper, three different types of task-oriented educational data are employed to investigate the performance of machine learning methods in different application scenarios. Specifically, seven parameter-optimized machine learning methods are implemented to study multiple types of performance prediction, including binary and multi-classification prediction tasks. In the experimental section, four evaluation metrics and visualizations are presented for a comparative study of different methods on three tasks, and an elaborated discussion of the experimental results is provided. The experimental results demonstrate that Random Forest has achieved superior generality on all selected datasets. In addition, the performance of Decision Tree and Artificial Neural Network models on the selected datasets indicates that they are also potential candidates to solve student performance prediction tasks.

Keywords Educational data mining · Student performance prediction · Machine learning · Classification tasks

Extended author information available on the last page of the article

1 Introduction

Educational data mining (EDM) has recently received growing interest in modern educational community. Student performance analysis and prediction, which are widely studied by researchers, have shown theoretical value and practical significance. Recently, the development of modern technology and the improvement of storage capacity have provided necessary guarantees for diversified educational application scenarios, such as dropout rate prediction (Prenkaj et al., 2020), exam pass prediction (Al-Shehri et al. 2017; Bujang et al. 2021) and admission rate prediction (Maltz et al., 2007), etc. Accordingly, the prediction results help teachers to develop more appropriate educational programs and facilitate targeted learning for students.

Nowadays, various studies have shown the superiority of machine learning techniques in the field of student performance prediction. With the technique guarantees of machine learning, copious methods are provided, such as K Nearest Neighbors (KNN) (Amra & Maghari, 2017), Support Vector Machine (SVM) (Al-Shehri et al. 2017), Random Forest (Ali et al., 2012), Decision Tree (Priyanka & Kumar, 2020) and Artificial Neural Network (ANN) (Zador, 2019), etc. These machine learning methods have provided novel insights for education researchers to generate diverse prediction tasks. Existing researches on student performance prediction analysis mainly focus on the exploration of educational data and application of machine learning methods to specific scenarios.

In particularly, student score prediction is the most prevalent research task in student performance prediction. For instance, Beaulac and Rosenthal (2019) collected students' course grades in the first two semesters to predict whether they would successfully obtain an undergraduate degree and demonstrated that Random Forest method could achieve better results than baseline models. Sekeroglu et al. (2019) used three common machine learning algorithms to predict the grades of students, and the analysis results confirmed the effectiveness of machine learning algorithms in predicting grades data. Moreover, this work also illustrated that the student grades were considerably influenced by the factors of social environment and family. Student employment prediction is an extensively studied task in student performance prediction. For example, Sharma and Uttam (2021) utilized KNN method to analyze and explore student characteristic data for a school to help university administrators predict the employment situation of graduates, where the experimental results obtained an 82% accuracy rate. In addition, dropout rate prediction is also a meaningful task in student performance prediction. Particularly, Basnet et al. (2022) investigated the application of machine learning algorithms for student dropout rates in MOOCs, which used machine learning algorithms to optimize dropout prediction models. Haiyang et al. (2018) proposed a time series forest classification algorithm to predict the dropout rates of students, where the early prediction results could help instructors design interventions to encourage course completion for students.

Although a wide range of methods have been proposed for some specific educational situations, there are still some challenges. At first, only a few machine learning methods are selected to predict student performance. Thus, extensive machine learning methods are not analyzed and compared systematically. Furthermore, most existing studies have used only one dataset to predict student performance at a certain education level, which is not comprehensive in student performance prediction.

In this work, the aforementioned challenges motivate us to perform a comparative study on student performance prediction using machine learning methods, and broaden the application in educational scenarios oriented to multiple specific tasks. Specifically, three datasets from different educational sources are selected, which include student admission dataset, campus placement dataset and students' grade dataset. Based on these datasets, binary and multi-classification prediction tasks can be established to investigate the effectiveness of machine learning methods for student performance prediction. Wherein, the binary classification prediction tasks involve the possibility of admission to graduate school and the chance of obtaining an internship for graduates, while the multi-classification prediction task is to predict the assessment of students in a middle school on a particular exam. Moreover, it is worth mentioning that the three datasets cover students at different levels of education, including secondary school students, college students, and graduate students, and this study addresses the prevalent tasks in student performance prediction. Furthermore, seven machine learning methods are extensively compared on these datasets, and the corresponding discussion and analysis are conducted based on the experimental results.

This paper is organized as follows. Related work is presented in Sect. 2. The method used in this research is described in Sect. 3. The results of the experiments and the discussion are presented in Sect. 4. Finally, conclusions are given in Sect. 5.

2 Related work

A series of studies using machine learning techniques have emerged to predict student performance. We summarize the methods, data, limitations and specific educational tasks addressed by the existing studies in Table 1.

In the real educational problem, existing works typically treat the problem as a classification or regression task, and employ machine learning methods for training and prediction. Specifically, Coussement et al. (2020) utilized Decision Tree to predict the list of students at risk of dropping out of school by their cumulative grade point averages at graduation, and provided targeted instructions for teachers or school administrators based on the results of the experiment. Sa'ad and Mustafa (2020) applied Extreme Learning Machine (ELM) algorithm, ANN and SVM to predict the dropout of a PhD class respectively, and illustrated that ELM achieved highest prediction accuracy. Bujang et al. (2021) compared the performance of common machine learning techniques such as Naive Bayes, KNN, and Logistic Regression (Hosmer et al., 1997) in predicting students' first semester grades based on the synthetic minority oversampling technique. Liang, Li and Zheng (2016) collected data from 39 courses on a MOOC platform and utilized the data on students' learning activities over a certain period to predict the dropout rate of students in the coming days, where the gradient boosting decision tree method achieved 89% accuracy in the task. Gray and Perkins (2019) proposed a prediction model based on Decision Tree method to accurately predict whether a student would be retained in the third week of a new semester based on their attendance.

In reality, there are abundant types of educational data, such as grades data, behavioral data, and demographic data, which reflect the characteristics of students from different perspectives. Specifically, grades data represents the scores which students have earned on exams. Thus, grades data are the most obvious factors in judging student performance. Behavioral data refers to students' own attributes and behavioral characteristics, such as gender, attendance and number of placements. Demographic data is used to describe information about students' social relationships, such as parents' occupation, education level, etc. Based on these diverse educational data, various studies have been conducted to facilitate the modeling of machine learning models in educational data mining. Kuzilek et al. (2021) mined and analyzed behavioral data during student exams to predict the probability of students failing the exam, and the prediction results showed that the dataset consisting of various patterns of student behavior had a significant effect on the classification evaluation. Bydžovská and Popelínský (2014) used several data mining techniques to obtain the grades and

Table 1 Summary of methods, data, limitations and tasks solved in related articles

Articles	Methods	Data types	Prediction tasks	limitations
Tomasevic et al. (2020)	ANN, SVM, KNN, Decision Tree	Grades, Behavioral, Demographic	Grades	Only student performance prediction task was implemented
Sa'ad and Mustafa (2020)	ELM, SVM, ANN	Behavioral	Dropout rate of PhD	Only behavioral data were utilized to predict dropout rate
Bujang et al. (2021)	Naive Bayes, KNN, Logistic Regression	Grades, Demographic	Semester-end grades	Limited methods were applied for make comparison
Liang, Li and Zheng (2016)	Gradient boosting Decision Tree	Behavioral	Dropout rate of MOOC	Limited comparison methods were applied and only behavioral data were used for analysis
Gray and Perkins (2019)	KNN, Decision Tree	Behavioral	Possibility to pass the exam	Only behavioral data were used to predict the possibility of passing the exam
Sharma and Uttam (2021)	KNN	Grades, Behavioral	Employment Status	Sufficient comparisons were not implemented
Beaulac and Rosenthal (2019)	Decision Tree, Random Forest	Grades	Possibility of obtaining a degree	Limited comparison methods and only grades data were used for modeling
Kuzilek et al. (2021)	KNN, SVM, Decision Tree, Random Forest	Behavioral	Grades	Only behavioral data were utilized for student performance prediction
Kardan et al. (2013)	ANN	Behavioral	Course selection behavior	No comparison methods and only behavioral data were analyzed
Eashwar et al. (2017)	SVM	Behavioral, Grades	Academic levels	Limited samples and no comparison methods were applied
Guarín et al. (2015)	Naive Bayes, Decision Tree	Grades, Demographic	Possibility of dropping out	Limited methods for comparison

behavioral data from university information system and confirmed that behavioral data could have a significant impact on student performance prediction tasks. Tomasevic et al. (2020) mined and analyzed the data of student grades and demographics to illustrate that student pass rate and class participation were important factors for the prediction of machine learning models. Iam-On and Boongoen (2017) collected mixed types of empirical data related to student demographics, academic performance, and enrollment records, which were used for early prediction of students' risk of dropping out. The results showed that multiple types of data were helpful for model prediction enhancement.

Overall, although existing works have been proposed to reflect the effectiveness of machine learning for student performance prediction, there are still some challenges that need to be addressed. As shown in Table 1, we can see that only a few machine learning methods are discussed in each study. However, the same machine learning technique may show distinct prediction performance for different classification prediction tasks. Thus, only a few machine learning methods are not sufficient for an adequate analysis. Furthermore, each study uses only one dataset to focus on one specific task, and most of the data types are not comprehensive, which would not facilitate the full exploitation of the role of different data types in machine learning-based prediction of student performance. Therefore, aforementioned challenges motivate us to perform a comprehensive analysis and comparison of large-scale machine learning methods on different types of tasks.

3 Method

The research framework diagram of this work is shown in Fig. 1, which consists of dataset collection and preprocessing, feature selection, modeling and prediction, analysis and discussion for experimental results. As shown in Fig. 1, in the phase of data collection and preprocessing, three tables represent the instances of three real educational datasets, where each piece of data in the table represents a student sample containing several features. In the phase of feature selection, the correlations between different features and the target variable are compared, and the features with stronger correlations are selected as input variables for the models. Subsequently, the performance of seven classifier instances is compared in tabular form, and the bar charts are utilized for the visualization of the tables to illustrate a more intuitive presentation of the results.

In this study, we first select three datasets oriented to different classification prediction tasks and preprocess them separately. Furthermore, feature selection based on mutual information and other filtering methods are implemented to improve the prediction accuracy of the models. Subsequently, seven classical machine learning methods are selected and the predictions are implemented on three datasets after parameter optimization. Finally, the prediction results are compared and discussed based on different model evaluation metrics.

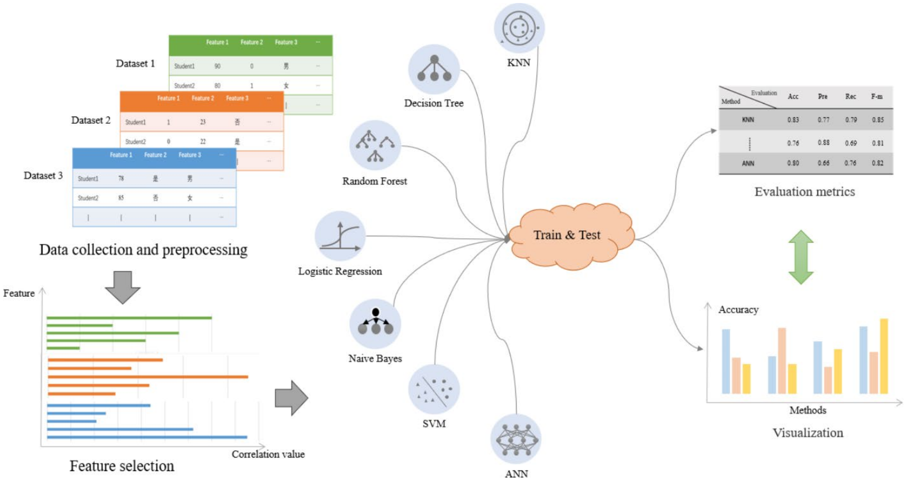

Fig. 1 Research framework diagram of this work

3.1 Dataset collection and description

In this section, three educational datasets oriented to different tasks are selected, including the description of the features and the range of the feature values.

3.1.1 Student admission dataset

The Student Admission Dataset (SAD) is obtained from Acharya et al. (2019), where the data are collected from graduates in the National Institute of Engineering. The dataset can be found on the website (https://www.kaggle.com/datasets/mohansacharya/graduate-admissions). The purpose of constructing this dataset is to predict whether a student will be admitted to the graduate exam based on student-related characteristics, which is a binary prediction task. The dataset includes a total of 400

Table 2 Description and value range of features on the SAD

Feature	Description	Value range
GRE Scores	The score of Graduate Record Examination	290 to 340
TOEFL Scores	Test of English as a Foreign Language	92 to 120
UR	University Rating	1 to 5
SOP	Statement of Purpose	1 to 5
LOR	Letter of Recommendation Strength	1 to 5
CGPA	Undergraduate CGPA	6.8 to 9.92
Research	Research Experience	0 or 1
Chance of Admit	Label	0 for < 70% and 1 for >= 70%

samples with 7 features that are considered as important indicators for master's programs. In the current education field, it is informative to measure students' chances to accept master degrees based on their relevant features. Table 2 shows the specific description of the seven features of this dataset and their value ranges. The GRE Scores are test scores required by graduate schools for admission, and the TOEFL Scores represent the English level of students. CGPA indicates cumulative grade point average in college. Chance of Admit is the label of this dataset, the value less than 0.7 is set to 0, meaning not admitted, while the value greater than 0.7 is set to 1, meaning admitted.

3.1.2 Engineering placements prediction dataset

The Engineering Placements Prediction Dataset (EPPD) is obtained from a university's published campus placement records for their engineering programs in 2013–2014 and includes a total of 2966 sample records. The EPPD dataset can be accessed through the website (https://www.kaggle.com/datasets/tejashvi14/engineering-placements-prediction/versions/3). The objective is to perform a binary classification prediction based on students' characteristics to predict whether a graduating engineering student will receive an internship. Campus placement activities can help universities increase employment rates and students get jobs. Therefore, the classification task is meaningful.

The dataset is shown in Table 3, where Age is the graduation age of the sample students, ranging from 19 to 30 years old. Gender is the gender of the internship candidate, and Stream is the major of the sample students, including Computer Science, Information Technology, Electronics and Communication, Mechanical, Electrical and Civil six directions. Internships refer to the number of internships during university, ranging from 0 to 3. CGPA refers to the average GPA of six semesters ranging from 5 to 9. Hostel indicates whether the student lived in a university residence during the school year, and history of backlogs indicates whether the student had any incomplete subjects. Both of them take the value of 0 or 1. Placed or not is the label of this dataset, where 1 means one is accepted to the internship, and 0 means not.

Table 3 Description of features on the EPPD

Feature	Description
Age	Age at the time of final year
Gender	Gender of candidate
Stream	Engineering stream that the candidate belongs to
Internships	Number of internships undertaken during the course of studies
CGPA	CGPA till 6th semester
Hostel	Whether student lives in college accommodation
History of backlogs	Whether student ever had any backlogs in any subjects
Placed or not	Label

Table 4 Description and value range of features on the SPD

Feature	Description	Value range
Reason	Reason to choose this school	Four values such as 'close to home'
Guardian	Student's guardian	['mother', 'father', 'other']
Medu	mother's education	[0,1,2,3,4]
Fedu	father's education	[0,1,2,3,4]
Travel time	Home to school travel time	[1,2,3,4]
Study time	Weekly study time	[1,2,3,4]
Famrel	Quality of family relationships	1 to 5
Free time	Free time after school	1 to 5
Go out	Going out with friends	1 to 5
Dalc	Workday alcohol consumption	1 to 5
Walc	Weekend alcohol consumption	1 to 5
Health	Current health status	1 to 5
Failures	Number of past class failures	[1,2,3,4]
G1	First period grade	0 to 20
G2	Second period grade	0 to 20
G3	Label: final grade	0 to 20

3.1.3 Student performance dataset

The Student Performance Dataset (SPD) is designed to study the classification of student performance in two Portuguese secondary schools with 649 samples and 32 relevant features such as student performance, demographics, society and school. The SPD dataset is available in the website (https://www.kaggle.com/datasets/larsen0966/student-performance-data-set). To facilitate the representation, the 14 remaining features after data preprocessing and feature selection with the strongest correlation to label G3 are presented in Table 4. Reason is the explanation for choosing this school, with four values of 'close to home', 'school reputation', 'course preference' or 'other'. Medu and Fedu represent the education level of the parents, with 0 indicating no education, 1 indicating only primary education up to grade 4, 2 indicating the highest level of education from grade 5–9, 3 indicating secondary education, and 4 indicating higher education. Travel time is used to measure the distance from home to school, and Study time is used to measure the number of hours of study per week. 1 to 5 measures the degree of six features (Famrel to Health) from low to high. Failures equals n if $1<n<3$, otherwise 4. G1 and G2 represent the final grades for the first two semesters of secondary school, respectively, and take values between 0 and 20. G3 is the Label of this dataset, which represents the students' grades in the last semester of secondary school. We classify students' grades less than 10 as failing, 10–15 as passing, and no less than 15 as excellent. Unlike the first two datasets, SPD is used for tri-classification prediction.

3.2 Data preprocessing and feature selection

Machine learning methods are applied to student performance prediction based on the selection of data, and the accuracy of the prediction depends heavily on the

Table 5 The samples of the datasets and the size of each subset

Dataset	Total	Train	Test	Feature numbers
SAD	400	300	100	7
EPPD	2966	2224	742	7
SPD	649	486	163	14

quality of the input data. Therefore, the selection of features with a high degree of relevance to the labels plays a key role in the classification task (Tomasevic et al., 2020). To improve prediction accuracy using basic machine learning techniques, data preprocessing is necessary. In this study, the three selected datasets contain different feature types, thus data preprocessing and feature engineering are initially performed for the three original datasets.

In the data pre-processing phase, checking for missing values is the first step. For numerical missing values, considering the missing values only occupy a small percentage, mean imputation as a simple and effective way, is adopted for filling the missing values. The mode value is used to fill in the values for categorical types, so as to avoid unnecessary information wastage. In the second step, the categorical data and continuous data are processed in different ways. First, for categorical data, we perform Ordinal Encoding (Potdar et al., 2017) to convert it into an array of integers. However, the obtained integers cannot be fed directly into the model, because the machine learning model automatically interprets them as ordered values. They only represent different categories without size distinction. Thus, OneHot Encoding (Bagui et al., 2019) is needed. Secondly, for continuous data, the split-box method is adopted, which can reduce the noise of the data by discretizing the continuous data, and reduce the risk of model overfitting. For categorical labels, such as binary classification prediction where labels are often classified as Yes or No, a Label Encoding (Bisong, 2019) is required to process them.

In the feature engineering phase, feature selection is a critical step that can directly determine the final prediction of the model. Specifically, variance filtering and mutual information filtering methods are selected to retain only features with strong correlation with the labels as input variables to the model, thus effectively improving prediction accuracy.

In addition, 75% of the processed data samples are classified as training set and 25% are classified as test set. The 10-fold cross-validation method is used to train the machine learning model on the training set. Parameter optimization is performed by plotting learning curves and grid search to find the parameters that can achieve the best results on the validation set. Finally, classification predictions are implemented on the test set. Specifically, the specific datasets are divided as shown in Table 5.

3.3 Modeling and prediction

In this subsection, seven popular machine learning algorithms are selected, including KNN, Decision Tree, Random Forest, Logistic Regression, SVM, Naive Bayes, and ANN. The principle and important parameters of each algorithm are described below.

KNN. KNN algorithm is one of the classification algorithms in supervised learning and is commonly used to solve binary and multi-classification problems because of its simple theoretical form and easy implementation. When it predicts a new sample, the category of the sample can be determined based on the class of the nearest *K* points to it. A smaller *K* value makes the prediction error rate higher, while a larger *K* value reduces the effect of noise but makes the boundaries between categories blurred. Thus, the selection of *K* is crucial. In specific experiments, grid search is used to determine the optimal *K* value.

Decision tree. Decision Tree is a common machine learning prediction model used to build classification systems. The method can be efficiently used for modeling complex data by constructing a tree diagram with root, inner and leaf nodes to describe possible decisions without introducing complex parameters, where each inner node represents a test of a feature and each leaf node represents a classification label (Song & Ying, 2015). Therefore, the parameter that most obviously affects the prediction effect of Decision Tree classification is *max_depth*, which indicates the maximum depth of the built tree. In practical applications, the most appropriate *max_depth* is found by plotting the learning curve of this parameter to obtain the best value on the validation set.

Random forest. For the classification problem, Random Forest is a classifier containing multiple mutually independent Decision Trees, whose outputs are determined by output categories of each tree. For an input sample, *N* Decision Trees output *N* classification results, and Random Forest integrates all the classification voting results and designates the category with the most votes as the final output, which is a simple Bagging idea. The classification result of Random Forest integrated Decision Trees will be judged incorrectly only when more than half of the trees are wrong. Hence, it can effectively prevent overfitting and it is more robust.

Logistic regression. Logistic regression is a common classifier that belongs to supervised learning in machine learning. The algorithm is actually mainly used to solve the binary classification problem, but can also be used to solve the multi-classification problem. In the process of training the model, to avoid overfitting and improve the generalization ability of the model, the L2 regular term is chosen as the value of the parameter penalty. The Logistic Regression algorithm is easy to implement, and it can be converted into multiple binary tasks in a certain way when it deals with multiple classification tasks. The principle is intuitive and easy to understand.

SVM. SVM is first created by Vapnik (Cortes & Vapnik, 1995) and has been successfully used to solve many classification problems. The technique tries to find a hyperplane as a decision boundary in a set of labeled data distributions so that the classification error (generalization error) of the model on the unknown data set is as small as possible. Specifically, the SVM model treats feature vectors as points in a multidimensional space so that vectors belonging to different classes are divided as widely as possible using hyperplanes. Usually, since the data in most realistic problems are nonlinear, the technique utilizes the Kernel Trick to implicitly map the input feature vectors to a higher dimensional space for effective nonlinear classification. For this purpose, it uses a kernel function that includes linear, polynomial, sigmoid and radial basis functions (RBF). In this study, RBF is used, which is a function that determines the mapping method based on the distance between sample points.

Naive bayes. Naive Bayes simplifies the Bayesian algorithm accordingly, and it assumes that the attributes are independent of each other for a given target value. There are three common model forms for this method, including Gaussian, Bernoulli and polynomial. It is experimentally validated that Gaussian Naive Bayes classification works best on the three selected datasets. Naive Bayes performs well for small-scale data and can handle each classification task effectively with stable classification efficiency. There are few parameters to be considered for estimation during the training of this model, so its training process is relatively simple.

ANN. ANN is a modeling method that simulates the structure of the nervous system and the working process of the human brain (Da Silva et al., 2017), with the advantages of effectively modeling the relationship between variables and implementing inference. In the ANN learning process, the network consisting of neurons is first built using the training data and the initial weights among the neurons are set. Then, the calculated values are compared with the actual values, and the neuron weights are adjusted using the gradient descent algorithm. Specifically, the number of neurons and network layers are selected by the performance on the validation set.

4 Results and discussion

The performance results of the seven mentioned classifiers are shown in Tables 6, 7 and 8, respectively. Figures 2, 3 and 4 represent the visualization results of Tables 6, 7 and 8, respectively. The bar chart allows a more visual comparison of the performance of the seven machine learning methods on the same classification task. Furthermore, in order to evaluate and compare the performance of the models, four evaluation metrics are selected, namely Accuracy, Precision, Recall and F-measure.

Table 6 reports the classification results of the seven machine learning methods on SAD, and Fig. 2 is a visualization of Table 6. From Fig. 2, we can see more intuitively that Decision Tree and Random Forest perform better in the binary classification prediction task. The accuracy rate of Random Forest reaches 87%, the recall rate reaches 86.75%, the precision rate and the F-measure rate of Decision Tree are 88.07%, and the above four values are the highest values of each of the four metrics. Random Forest in Fig. 2 shows better uniformity than Decision Tree in the four metrics, indicating that Random Forest has better robustness in performing SAD classification. KNN with the accuracy rate of 78% shows the worst overall performance in this task. Naive Bayes performs a little better than KNN but the overall performance is still poor. In addition, the prediction results of Logistic Regression, SVM and ANN are in the middle level.

Table 7 reports the classification results of seven machine learning methods on EPPD, and Fig. 3 is a visualization of Table 7. Figure 3 shows that Decision Tree, Random Forest, Logistic Regression and SVM have significantly better prediction

Table 6 The comparative performance of seven models on SAD

Method	Accuracy	Precision	Recall	F-measure
KNN	78.00	77.26	77.26	77.26
Decision Tree	86.00	**88.07**	83.67	**88.07**
Random Forest	**87.00**	86.49	**86.75**	86.61
Logistic Regression	84.00	83.46	83.46	83.46
SVM	85.00	85.78	83.19	84.00
Naive Bayes	81.00	80.55	76.70	78.58
ANN	83.00	83.07	81.50	82.05

results than the other three models. The accuracy rate of Random Forest can reach 89.08%, the precision rate reaches 90.07%, the recall rate is 89.72% and the F-measure rate reaches 89.07%, which are better than the other techniques in all four metrics and still achieve the best performance. Secondly, KNN performs the worst on the binary classification task of predicting whether graduates would successfully obtain an internship, with none of the four model evaluation metric results reaching 75%. The results of all four metrics of Naive Bayes are between 75% and 80%, performing slightly better than KNN, but the overall performance is still poor. Finally, we can find that ANN performs at a medium level in this task, with all four metrics around 85%.

Table 8 reports the classification results of seven machine learning methods on SPD, and Fig. 4 is a visualization of Table 8. At first, from the overall view of the four model evaluation metrics, ANN shows the best results with an overall range of 70–80%, and the accuracy rate even reaches 80.36%. Naive Bayes is inferior to ANN. Although neither the recall rate nor the F-measure rate of Random Forest reaches 70%, the accuracy reaches 80.98% and the precision rate reaches 85.35%, which clearly surpasses the other metrics. In addition, Logistic Regression has the worst overall performance in this task.

It is worth mentioning that SPD is used to predict the grades of students in a secondary school, which is oriented to a triple classification prediction task. Thus, these seven machine learning models may perform differently. From Fig. 4, we can find that the accuracy rate and the precision rate tend to be higher than the recall rate and the F-measure rate among the four selected model evaluation metrics. After studying the data, the uneven distribution of data samples of different categories in SPD leads to bias in the classification performance on different categories, which affects the calculation of evaluation metrics.

Table 7 The comparative performance of seven models on EPPD

Method	Accuracy	Precision	Recall	F-measure
KNN	73.98	73.91	73.64	73.72
Decision Tree	88.54	89.26	89.10	88.54
Random Forest	**89.08**	**90.07**	**89.72**	**89.07**
Logistic Regression	87.46	87.81	87.88	87.46
SVM	88.94	89.58	89.48	88.94
Naive Bayes	78.84	78.79	78.58	78.65
ANN	84.90	85.56	85.43	84.90

Table 8 The comparative performance of seven models on SPD

Method	Accuracy	Precision	Recall	F-measure
KNN	74.23	62.26	59.25	59.05
Decision Tree	76.07	73.74	61.68	65.38
Random Forest	**80.98**	**85.35**	69.36	69.17
Logistic Regression	71.16	61.68	50.15	51.84
SVM	71.77	70.80	59.46	62.36
Naive Bayes	77.91	70.66	74.51	71.56
ANN	80.36	76.53	**71.97**	**72.22**

Table 9 is summarized to demonstrate the differences between the datasets and the performance of methods from an intuitive perspective, where the data type and oriented tasks of datasets are listed. Moreover, the top two favorable and unfavorable methods on these datasets are also provided. First, unlike existing studies which discuss only one type of dataset, the applied datasets cover all three data types as shown in Table 9. The prediction task includes binary or multi-classification. Hence, it is more comprehensive and comparable. In addition, Random Forest model has achieved promising performance on all three data sets while KNN and Naive Bayes have realized relatively worst results on binary classification tasks. Subsequently, SVM and Logistic Regression have showed the worst results in the multi-classification task.

By comparing the performance of the seven machine learning techniques on the above three different datasets, it is summarized as follows. At first, Decision Tree and Random Forest are suitable for both binary and multi-classification prediction, among which Random Forest is more stable and has higher accuracy of prediction results than Decision Tree. Since Random Forest is an integration of a series of Decision Trees, it can further improve the classification performance based on Decision Trees, and the voting mechanism it uses can further avoid overfitting. Thus, Random Forest exhibits the most effective performance in predicting student performance. Secondly, Logistic Regression and SVM are more suitable for binary classification prediction. Among them, Logistic Regression is essentially used to solve binary classification problems, and the idea is to transform a multi-classification problem into multiple binary classification problems when predicting a multi-classification task. For example, the transformation idea of Logistic Regression in solving the triple classification problem is to consider one of the categories as positive and the other two categories as negative, and repeat the process three times to finally get three binary classifiers to achieve the prediction. Similarly, SVM implements the triple classification task by constructing decision boundaries three times to achieve the outcome prediction. Therefore, these two models are susceptible to imbalance in the number of samples in the training set when implementing multi-classification prediction, which can result in bias and instability of the classifier. Finally, it is also clear from the experimental results that KNN, Naive Bayes and ANN are more suitable for multi-classification prediction tasks. It is worth noting that ANN is more suitable for modeling complex data with a large number of features and a large variety of classifications.

In order to explore the effect of different dataset sizes on model performance, we use SAD dataset for discussion. Figure 5 shows the performance changes of three representative machine learning methods as the SAD dataset size increases. Overall, the accuracy rate shows an increasing trend as the dataset increases. Specifically, the accuracy of the three methods improves rapidly when the dataset size increases from 50 to 200 at the beginning. However, the improvement in accuracy levels off when the dataset size increases from 200 to 400. The change in accuracy largely results

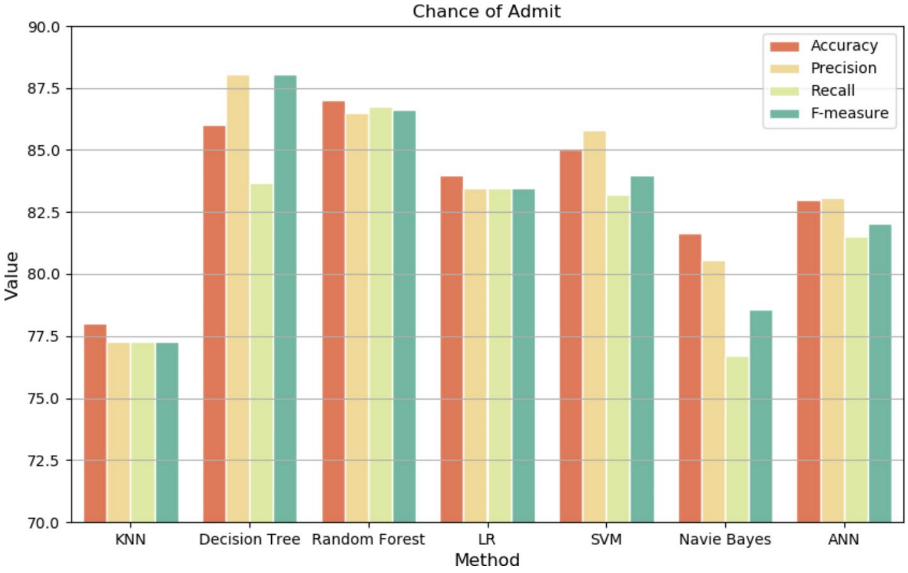

Fig. 2 Visualization of classification performance of seven classifiers on SAD.

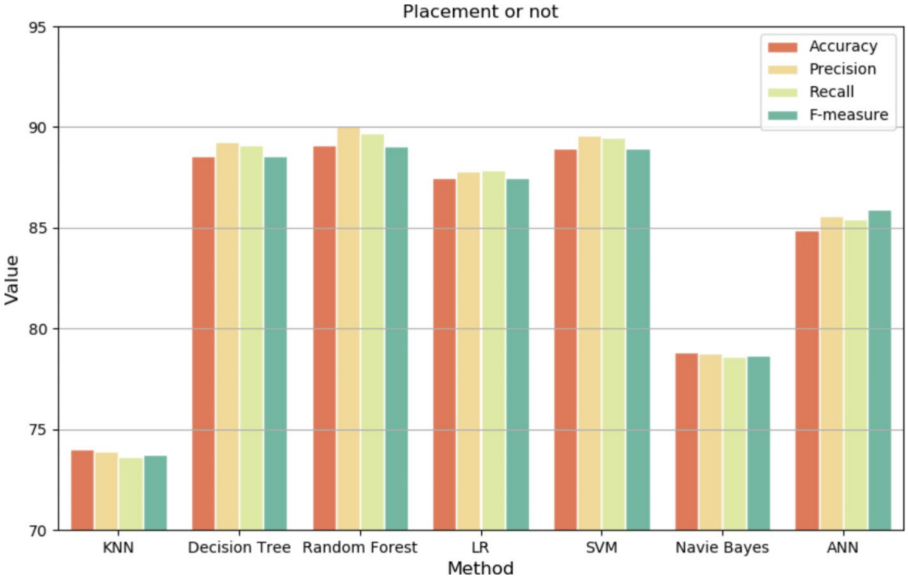

Fig. 3 Visualization of classification performance of seven classifiers on EPPD

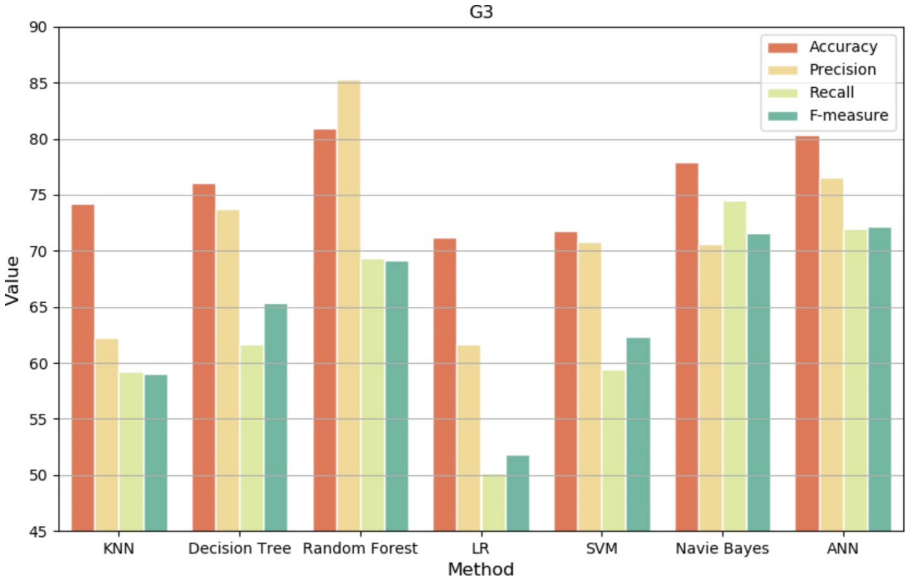

Fig. 4 Visualization of classification performance of seven classifiers on SPD

Table 9 The Comparison between datasets and favorable methods for each dataset

Dataset	Data type	Task-oriented	Favorable methods	Unfavorable methods
SAD	Grades	binary	Decision Tree, Random Forest	KNN, Naive Bayes
EPPD	Grades and behavioral	binary	Random Forest	KNN, Naive Bayes
SPD	Grades, behavioral and demographics	tri-classification	Random Forest, ANN	SVM, Logistic Regression

from the fact that machine learning models rely on a certain amount of training data to find the optimal parameters for student performance prediction tasks. In addition, overfitting may occur when the training data is too small, which would also lead to ineffective prediction performance.

In summary, this work considerably has extended the existing studies on machine learning based student performance prediction, which only applied a few machine learning methods on one type of dataset to solve a specific task. For example, as mentioned in the related work section, Kuzilek et al. (2021) and Tomasevic et al. (2020) studied the performance of several machine learning methods in predicting student performance. However, they only use four methods, while our work uses seven methods for a systematic comparison. Sa'ad and Mustafa (2020) and Liang, Li and Zheng (2016) applied machine learning models for predicting the dropout

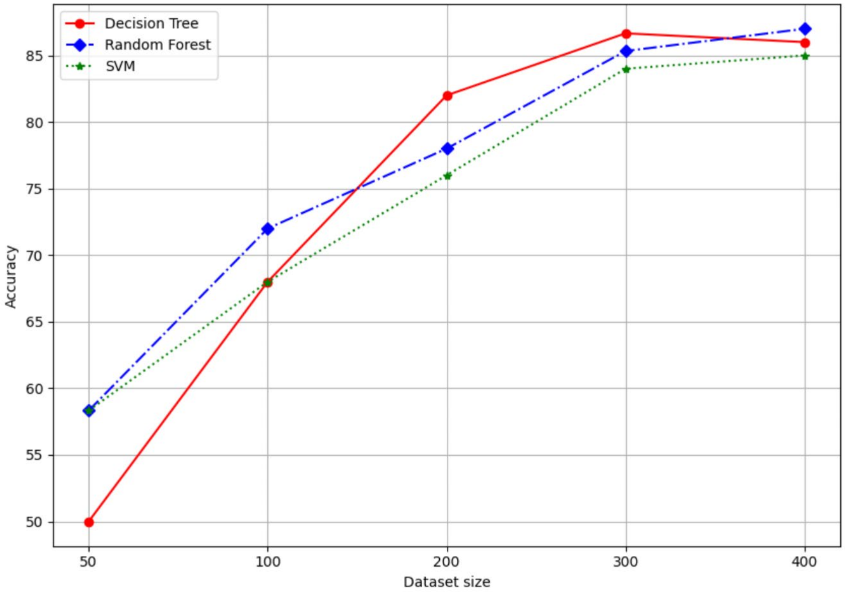

Fig. 5 The changes in accuracy as the size of the SAD dataset increases

rate of students, where only one dataset with behavior type was used for analysis, which was not enough for investigation. In this work, we select three datasets from different educational sources, covering student admission dataset, campus placement dataset and students' grades dataset. Moreover, based on these datasets, binary and multi-classification prediction tasks are established to investigate the effectiveness of machine learning methods in student performance prediction. Thus, compared other studies, this work is capable of facilitating the full exploitation of the role of different data types in machine learning-based prediction of student performance.

Moreover, based on the experimental results some helpful suggestions can be provided. Firstly, after specifying a student performance prediction task, when it is categorized as a binary or multi-classification task, Random Forest model can be selected in preference. Random Forest is an ensemble algorithm consisting of multiple decision tree models, which introduces randomness in the training procedure. Thus, it is less susceptible to the presence of noise and not prone to overfitting, achieving better prediction results. Furthermore, Random Forest is capable of demonstrating strong generalization over different types of data categories and can be a primary candidate for student performance prediction tasks. In addition, we have found that Decision Tree and ANN are capable of achieving relatively promising performance of on the selected datasets, which indicates that they are also potential candidates to solve student performance prediction tasks. Secondly, in the procedure of modeling educational data, an excessive number of features tend to increase the training samples of the model, which leads to expanding the search space of the model significantly. In order to simplify the model and make it easier to understand, it is necessary to remove irrelevant features using feature selection, which would enhance the stability of the model and reduce the risk of overfitting. Lastly, when exploring the performance of a machine learning method on specific educational data set, we can first conduct experiments with certain types of data separately, and then construct combinations of different types of data to explore the impact on prediction results. Thus, empirical validation is a necessary procedure to identify the optimal combination of input data types.

5 Conclusion

In this paper, student performance datasets from three different educational scenarios are selected to validate the performance of machine learning models for three prevalent prediction tasks, including the possibility of admission to graduate school exams, the chance of obtaining an internship for graduates and the assessment of students in a middle school on a particular exam. Moreover, these datasets contain students at different levels of education, covering secondary school students, college students, and graduate students. Furthermore, seven selected machine learning methods are implemented for experimental performance comparisons, and the results show that different machine learning methods are suitable for different classification tasks. Specifically, Random Forest model has achieved promising performance on all three datasets. KNN and Naive Bayes have realized relatively worst results on binary classification tasks. SVM and Logistic Regression have showed the worst results in the multi-classification task. Thus, in the view of the promising performance of Random Forest, it can be considered as a primary candidate for student performance prediction tasks.

In the future, we would like to apply more advanced techniques like deep learning in the student performance prediction tasks to further improve the prediction of the model. Furthermore, it would be a meaningful topic to consider richer educational data to leverage the role of advanced technologies in the field of modern educational data mining.

Acknowledgements All errors remain mine. The author is thankful for all reviewers' comments, recommendations, and suggestions.

Funding The research contained in this paper was supported by the Key Research and Development Program of Shandong Province, China, under Grant 2017GGX10142.

Data Availability The links of the data used have been provided in the paper.

Declarations

Conflict of Interest The authors have no relevant financial or non-financial interests to disclose.

References

Acharya, M. S., Armaan, A., & Antony, A. S. (2019, February). A comparison of regression models for prediction of graduate admissions. In *2019 international conference on computational intelligence in data science (ICCIDS)* (pp. 1–5). IEEE.

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), 272.

Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., & Olatunji, S. O. (2017, April). Student performance prediction using support vector machine and k-nearest neighbor. In *2017 IEEE 30th canadian conference on electrical and computer engineering (CCECE)* (pp. 1–4). IEEE.

Amra, I. A. A., & Maghari, A. Y. (2017, May). Students performance prediction using KNN and Naïve Bayesian. In *2017 8th International Conference on Information Technology (ICIT)* (pp. 909–913). IEEE.

Bagui, S., Nandi, D., Bagui, S., & White, R. J. (2021). Machine learning and deep learning for phishing email classification using one-hot encoding. *Journal of Computer Science*, *17*, 610–623.

Basnet, R. B., Johnson, C., & Doleck, T. (2022). Dropout prediction in Moocs using deep learning and machine learning. *Education and Information Technologies*, 1–15.

Beaulac, C., & Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education*, *60*(7), 1048–1064.

Bisong, E. (2019). Introduction to Scikit-learn. In *Building machine learning and deep learning models on Google cloud platform* (pp. 215–229). Apress, Berkeley, CA.

Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N. A. M. (2021). Multiclass prediction model for student grade prediction using machine learning. *Ieee Access : Practical Innovations, Open Solutions*, *9*, 95608–95621.

Bydžovská, H., & Popelínský, L. (2014, July). The influence of social data on student success prediction. In *Proceedings of the 18th International Database Engineering & Applications Symposium* (pp. 374–375).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: the beneficial impact of the logit leaf model. *Decision Support Systems*, *135*, 113325.

Da Silva, I. N., Spatti, D. H., Flauzino, R. A., & Liboni, L. H. (2017). Artificial neural networks. *B., & dos Reis Alves* (p. 39). Cham: Springer International Publishing.

Eashwar, K. B., Venkatesan, R., & Ganesh, D. (2017). Student performance prediction using SVM. *International Journal of Mechanical Engineering and Technology*, *8*(11), 649–662.

Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education*, *131*, 22–32.

Guarín, C. E. L., Guzmán, E. L., & González, F. A. (2015). A model to predict low academic performance at a specific enrollment using data mining. *IEEE Revista Iberoamericana de tecnologias del Aprendizaje*, *10*(3), 119–125.

Haiyang, L., Wang, Z., Benachour, P., & Tubman, P. (2018, July). A time series classification method for behaviour-based dropout prediction. In *2018 IEEE 18th international conference on advanced learning technologies (ICALT)* (pp. 191–195). IEEE.

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, *16*(9), 965–980.

Iam-On, N., & Boongoen, T. (2017). Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics*, *8*(2), 497–510.

Kardan, A. A., Sadeghi, H., Ghidary, S. S., & Sani, M. R. F. (2013). Prediction of student course selection in online higher education institutes using neural network. *Computers & Education*, *65*, 1–11.

Kuzilek, J., Zdrahal, Z., & Fuglik, V. (2021). Student success prediction using student exam behaviour. *Future Generation Computer Systems*, *125*, 661–671.

Liang, J., Li, C., & Zheng, L. (2016, August). Machine learning application in MOOCs: Dropout prediction. In *2016 11th International Conference on Computer Science & Education (ICCSE)* (pp. 52–57). IEEE.

Maltz, E. N., Murphy, K. E., & Hand, M. L. (2007). Decision support for university enrollment management: implementation and experience. *Decision Support Systems*, *44*(1), 106–123.

Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, *175*(4), 7–9.

Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A survey of machine learning approaches for student dropout prediction in online courses. *ACM Computing Surveys (CSUR)*, *53*(3), 1–34.

Priyanka, & Kumar, D. (2020). Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences*, *12*(3), 246–269.

Sa'ad, M. I., & Mustafa, M. S. (2020, October). Student Prediction of Drop Out Using Extreme Learning Machine (ELM) Algorithm. In *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1–6). IEEE.

Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019, March). Student performance prediction and classification using machine learning algorithms. In *Proceedings of the 2019 8th International Conference on Educational and Information Technology* (pp. 7–11).

Sharma, G., & Uttam, A. K. (2021). Preparing application of K-Nearest Neighbor (KNN): a supervised machine learning based model in placement prediction for graduate course students. *Recent Trends in Communication and Electronics* (pp. 425–429). CRC Press.

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & education*, *143*, 103676.

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, *10*(1), 1–7.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yawen Chen[1] · Linbo Zhai[1]

✉ Linbo Zhai
zhai@mail.sdu.edu.cn

Yawen Chen
chenyawen111@163.com

[1] Present address: School of Information Science and Engineering, Shandong Normal University, 250014 Jinan, China