Comparison and Analysis of Multiple Machine Learning Algorithms for Predicting Student Adaptation Levels in Online Education

Yucong Li[1,a,*]

[1]*Professional Applied and Continuing Education, The University of Winnipeg, Winnipeg, Manitoba, R3C 0E8, Canada*
a. li-y88@webmail.uwinnipeg.ca
**corresponding author*

Abstract: With the rapid development and popularization of Internet technology, online education has become a new way of education. Compared with traditional classroom teaching, online education has a more flexible learning mode, a more convenient learning environment and a wider range of learning resources. However, at the same time, online education also faces some challenges, one of the most important challenges is the adaptability of students to online education. In this paper, we use machine learning techniques to predict students' adaptability in online classrooms. After using logistic regression model, k-neighborhood algorithm model, random forest model, XGBoost model and Cat Boost model to make predictions, it is found that random forest model is the best in predicting students' adaptability to online classroom, with a prediction accuracy of 89.6%. The XGBoost model and CatBoost model were also better in prediction, with prediction accuracies of 89.1% and 88.6%, respectively. In contrast, the logistic regression and KNN models have poorer prediction accuracy with 68.8% and 77.1%, respectively. The research in this article has important implications for the online education industry. By using machine learning techniques to predict students' adaptability in an online classroom, it can help educational institutions better understand students' learning and improve teaching effectiveness. Meanwhile, for students, knowing their adaptive ability in online classroom also helps them to better plan their study programs and improve their learning efficiency. This study uses machine learning techniques to predict students' adaptive ability in online classrooms, and the results show that the random forest model performs the best in terms of predictive effectiveness. This study provides a useful reference for the online education industry and also provides some ideas for future research.

Keywords: XGBoost, Machine Learning Algorithms, CatBoost

1. Introduction

With the rapid development and popularization of Internet technology, online education has become a new way of education [1]. Compared with traditional classroom teaching, online education has more flexible learning methods, more convenient learning environment and more extensive learning resources [2]. However, at the same time, online education also faces some challenges, one of the most important challenges is the adaptability of students to online education [3].

First of all, students need to have certain computer skills and network skills to be able to carry out online learning smoothly. For some students who lack computer skills and internet skills, they may encounter some difficulties, which will affect their learning [4]. Secondly, students need to have some self-management and self-study skills to be able to achieve good results in online learning [5]. The learning mode of online learning is very different from the traditional classroom teaching, students need to make their own study plan, manage their study time, evaluate their learning effect and so on, which will also cause some difficulties for some students who lack self-management and self-learning ability [6]. Finally, students need to have certain motivation and interest in learning in order to really enjoy the fun of online learning [7]. Compared with traditional classroom teaching, online learning is freer and more flexible, but at the same time more lonely and boring [8]. If students lack motivation and interest in learning, they may quickly lose interest in learning, which may affect learning outcomes [9].

In order to solve the problem of students' adaptation to online education, researchers have conducted a lot of research work. Among them, machine learning techniques play an important role in the research. Machine learning technique is an automated data-based learning method, which can automatically learn some laws and patterns from a large amount of data and be used for tasks such as prediction, classification, clustering, etc [10].

Machine learning techniques can be used to predict students' academic performance and learning outcomes. By analyzing and modeling the data of students' personal information, learning behaviors, and learning achievements, researchers can use machine learning techniques to predict students' learning achievements and learning effects, so as to provide educators with more accurate teaching strategies and help students improve their learning effects. Machine learning techniques can be used to personalize recommended learning resources. Online learning platforms usually have a large number of learning resources, and students need to choose the learning resources suitable for them according to their own learning needs and interests. By analyzing and modeling students' learning behaviors and learning interests, researchers can use machine learning technology to recommend learning resources that are suitable for students, thus improving students' learning effects and learning interests.

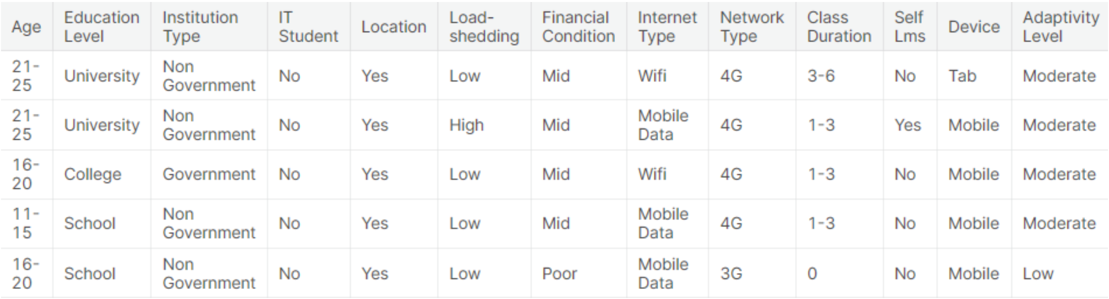
The problem of students' adaptability to online education is a very important issue, and solving it requires a lot of research work from researchers. Machine learning technology, as a powerful tool, can help researchers to discover patterns from a large amount of data, so as to provide educators with more accurate teaching strategies and help students improve their learning effectiveness.

2. Data set overview and feature calculation

2.1. Source of data sets

The Online Education Dataset is a very important educational dataset that is a collection of information about students at three different levels (schools, colleges, and universities) collected through online surveys and actual surveys. The purpose of this dataset is to help researchers understand student learning and the factors that affect student achievement, so that they can provide educators with better teaching strategies and help students improve their academic performance.

This dataset contains 13 features and 1 target column, the features include students' gender, age, family background, study time, study style, academic performance and so on, and the target column is students' level of adjustment. All these features are very important because they help us to understand the students' personal situation, study habits, academic level, and so on, which are important factors affecting the students' academic performance. This online education dataset is a very important dataset that can help us to understand how students learn and the factors that affect their performance so that we can provide educators with better teaching strategies and help students to improve their academic performance. Some of the data are shown in Figure 1.



Age	Education Level	Institution Type	IT Student	Location	Load-shedding	Financial Condition	Internet Type	Network Type	Class Duration	Self Lms	Device	Adaptivity Level
21-25	University	Non Government	No	Yes	Low	Mid	Wifi	4G	3-6	No	Tab	Moderate
21-25	University	Non Government	No	Yes	High	Mid	Mobile Data	4G	1-3	Yes	Mobile	Moderate
16-20	College	Government	No	Yes	Low	Mid	Wifi	4G	1-3	No	Mobile	Moderate
11-15	School	Non Government	No	Yes	Low	Mid	Mobile Data	4G	1-3	No	Mobile	Moderate
16-20	School	Non Government	No	Yes	Low	Poor	Mobile Data	3G	0	No	Mobile	Low

Figure 1: Data overview.

(Photo credit: Original)

2.2. Statistical analysis of data

Age, level of education, economic situation, type of network and gender all have an impact on students' adaptation in online education, and the data were analyzed statistically separately to observe the changes of each factor on students' adaptation in online education.

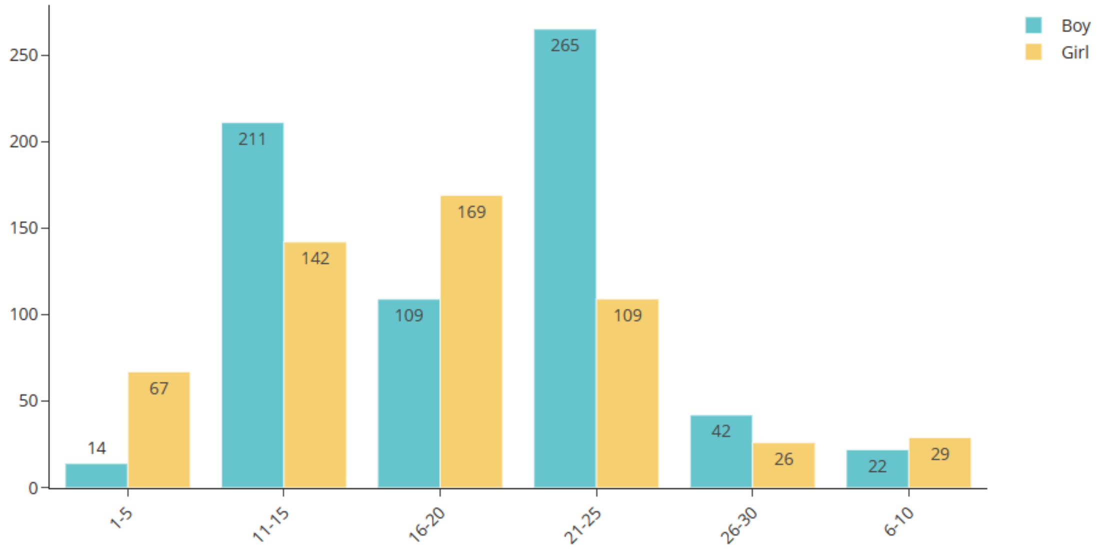


Figure 2: Statistical analysis of data.

(Photo credit: Original)

First of all, the distribution of male and female ratio in each age group is counted, as shown in Figure 2. As can be seen from Figure 2, the 11-25 age group accounts for the majority of the data, while the number of people in other age groups is relatively small, and at the same time, boys account for the majority of the 11-25 age group, while girls account for a minority.

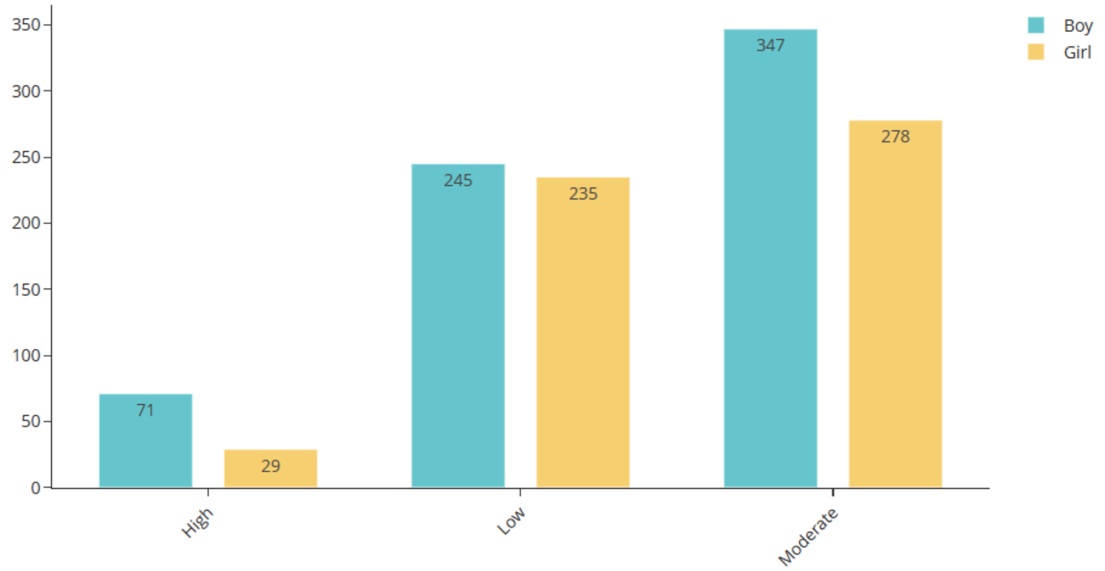


Figure 3: Statistical analysis of data.

(Photo credit: Original)

The students' adaptability to online classroom is categorized into high, medium and low in total, and it can be seen from Fig. 3 that most of the students' adaptability to online classroom is low and medium, and only a small number of them are very adaptable to online classroom, and the number of male students is greater than the number of female students in all the three adaptable abilities.

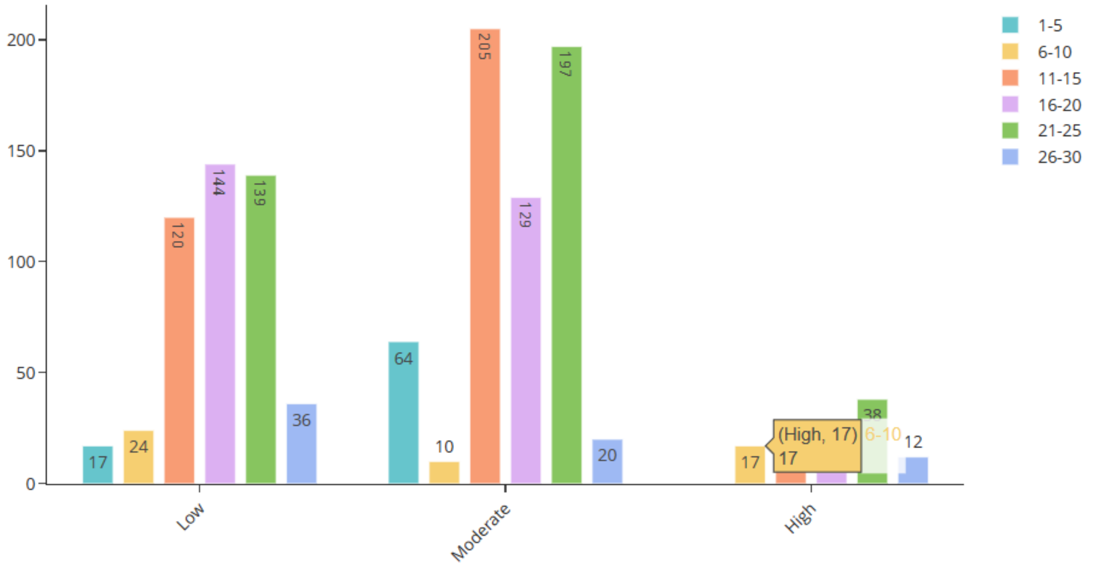


Figure 4: Statistical analysis of data.

(Photo credit: Original)

The number of students with various adaptive abilities is shown by age group in Figure 4, and it can be seen that the strongest age group for each adaptive ability is the 21-25 year olds, while the weakest adaptive ability is the 16-20 year olds.

3. Various machine learning algorithm models

3.1. Logistic regression model

Logistic regression model is a widely used model for classification problems, and its basic idea is to use an S-shaped function to establish a relationship between input and output variables. The advantage of the logistic regression model is that it is computationally fast and the results of the model can be interpreted easily. The disadvantage of the logistic regression model is that it is weak at modeling nonlinear relationships, and therefore more complex models may need to be used when dealing with nonlinear problems.

3.2. KNN

The KNN model is a similarity-based classification model, and its basic idea is to use the distance between samples to determine the similarity between them and to predict the label of a new sample based on the labels of the nearest K neighbors.The advantage of the KNN model is that it is simple to implement and has a strong modeling ability for nonlinear problems.The disadvantage of the KNN model is that it is weak for high-dimensional data, as the In high dimensional spaces, the distance between samples tends to become very large, causing the KNN model to become less effective.

3.3. Random forest model

Random forest model is an integrated learning model based on decision trees, and its basic idea is to use multiple decision trees for classification or regression, and use voting or averaging to get the final prediction results. The advantages of the Random Forest Model are that it is more capable of modeling nonlinear problems and can handle high dimensional data. The disadvantage of the Random Forest model is that its model structure is more complex and requires a large amount of computational resources.

3.4. XGBoost model

XGBoost model is a gradient boosting model based on decision trees, its basic idea is to use multiple decision trees to classify or regress, and gradually optimize the prediction effect of the model by gradient descent.The advantage of XGBoost model is that its prediction effect is better, and it can deal with high-dimensional data.The disadvantage of XGBoost model is that it is slower in computation speed and requires a large amount of computational resources.

3.5. CatBoost model

CatBoost model is a gradient boosting model based on decision tree, its basic idea is similar to XGBoost model, but it adopts some new techniques to improve the performance of the model. For example, CatBoost model can automatically handle category-based features and missing values, which reduces the workload of feature engineering.The advantage of CatBoost model is that its prediction effect is better and it can automatically handle category-based features and missing values.The disadvantage of CatBoost model is that it is slower and requires a lot of computational resources.

4. Experiments and Results

In this paper, logistic regression model, KNN model, random forest model, XGBoost model and Cat Boost model were used to predict students' adaptation to online classroom, respectively, with each parameter of students as inputs and students' adaptability level to the classroom as target variables. The dataset was divided into training set, validation set and test set according to 6:2:2, and each model was predicted five times and the average prediction accuracy was calculated. The results are shown in Table 1 and Figure 5.

Table 1: Model evaluation parameter.

Model	Mean Accurcy
Logisitic Regression	0.68809
K Nearest	0.77112
Random Forest Classifier	0.89568
XGB Classifier	0.89094
Cat Boost Classifier	0.88619

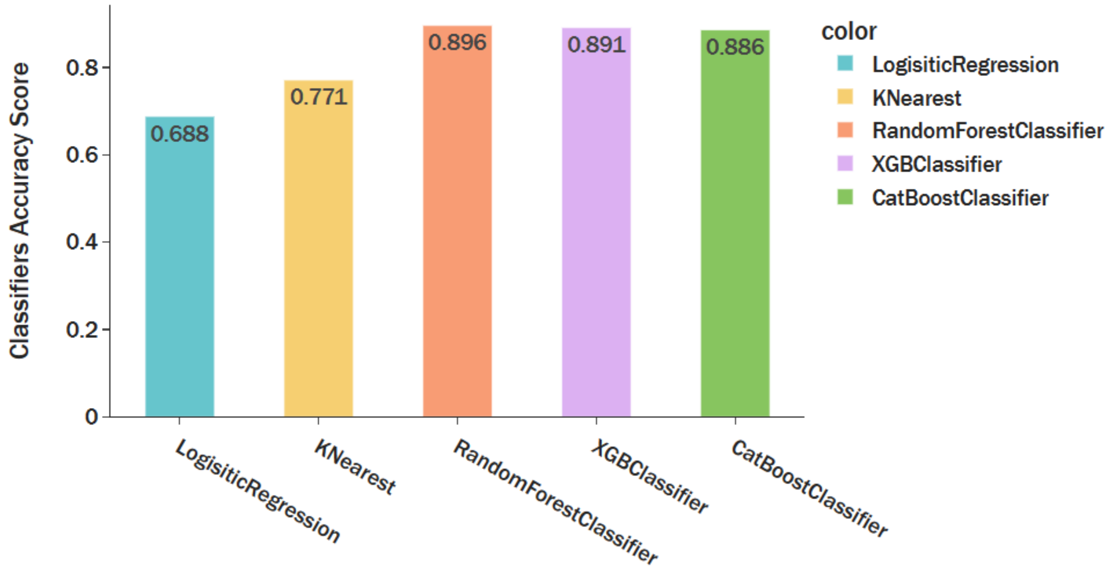


Figure 5: Model evaluation parameter.

(Photo credit: Original)

From the results, it can be seen that Random Forest is the best in predicting students' adaptability to the online classroom, with a prediction accuracy of 89.6%, while XGBoost and CatBoost models are also better in prediction, with prediction accuracies of 89.1% and 88.6%, respectively. While logistic regression and KNN model had poorer prediction accuracy with 68.8% and 77.1% respectively.

5. Conclusion

In this paper, logistic regression model, KNN model, random forest model, XGBoost model and CatBoost model were used to predict students' adaptability level to online classroom, and the prediction effect of each model was compared. The results showed that the random forest model was the most effective in predicting students' adaptive ability to online classrooms, with a prediction accuracy of 89.6%.The XGBoost model and the CatBoost model were also better in predicting the results, with prediction accuracies of 89.1% and 88.6%, respectively. The logistic regression and KNN models, on the other hand, have poorer prediction accuracies of 68.8% and 77.1%, respectively.

First of all, logistic regression model is a linear model widely used in classification problems, and its basic idea is to transform the relationship between the independent variable and the dependent variable into a probability value by transforming the independent variable into a logistic function, so as to carry out classification. In this paper, the logistic regression model has a low prediction accuracy, probably because the adaptability rating of online classrooms is affected by several factors, and the logistic regression model cannot deal with nonlinear relationships and cannot capture the complex relationship between these factors well.

Second, the KNN model is an instance-based classification method, and its basic idea is to compare the similarity of new samples with known samples to determine the category to which the new samples belong. In this paper, the prediction accuracy of the KNN model is low, probably because the KNN model is sensitive to the number of samples, and a large number of samples are needed to ensure the prediction effect. And in this paper, the number of samples may be insufficient, resulting in the poor prediction effect of KNN model.

Random forest model is an integrated learning method, and its basic idea is to improve the prediction accuracy by constructing multiple decision tree models and integrating their results. In this paper, the Random Forest model has the highest prediction accuracy, probably because the Random Forest model is able to deal with nonlinear relationships, and at the same time has a good generalization ability, which can adapt well to new data.

Both XGBoost model and CatBoost model are integrated learning methods based on gradient boosted trees, and the basic idea is to continuously improve the prediction accuracy by multiple iterations, each of which constructs a new decision tree model and adds it to the integrated model. In this paper, the XGBoost model and CatBoost model have higher prediction accuracies, probably because they are able to handle nonlinear relationships and have better generalization ability.

In summary, this paper used logistic regression model, KNN model, random forest model, XGBoost model and CatBoost model to predict students' adaptability level to online classroom, and compared the prediction effect of each model. The results showed that the random forest model was the most effective in predicting students' adaptive ability to online classrooms, with a prediction accuracy of 89.6%.The XGBoost model and CatBoost model were also better in predicting effectiveness, with prediction accuracies of 89.1% and 88.6%, respectively. While the logistic regression and KNN models have poorer prediction accuracy with 68.8% and 77.1% respectively. This indicates that the integrated learning approach can improve the prediction accuracy better than a single model when dealing with problems with nonlinear and complex relationships. Meanwhile, the size of the data volume also affects the prediction effect of the model, and the appropriate model and data volume need to be selected according to the specific situation.

References

[1] *Rodić D L ,Stančić I ,Čoko D , et al.Towards a Machine Learning Smart Toy Design for Early Childhood Geometry Education: Usability and Performance[J].Electronics,2023,12(8):*

[2] *A. J G ,Young P ,Ricardo S , et al.Data Analytics and Machine Learning in Education[J].Applied Sciences,2023,13(3):1418-1418.*

[3] *Kumar P U ,Vishal D ,P.S. N , et al.Predicting Global Ranking of Universities Across the World Using Machine Learning Regression Technique[J].SHS Web of Conferences,2023,156.*

[4] *Muhammad I Z ,G. A C .AGILEST approach: Using machine learning agents to facilitate kinesthetic learning in STEM education through real-time touchless hand interaction[J].Telematics and Informatics Reports,2023,9.*

[5] *Kwasi D D ,Charles A B .Using Machine Learning Techniques to Predict Learner Drop-out Rate in Higher Educational Institutions[J].Mobile Information Systems,2022,2022.*

[6] *Rui Z ,Yong L .Improve the Assessment Model of Personnel Develop Level in Higher Education Based on Machine Learning[J].Wireless Communications and Mobile Computing,2022,2022.*

[7] *S C M ,Fangzhou M ,S R R , et al.An approachable, flexible and practical machine learning workshop for biologists.[J].Bioinformatics (Oxford, England),2022,38(Supplement_1):i10-i18.*

[8] *Yongkang X ,Zhenfeng Z ,Jianbo X , et al. Renewable energy time series regulation strategy considering grid flexible load and N-1 faults[J]. Energy,2023,284.*

[9] *S M R ,Svetlana S ,Janez K , et al. Time resolved study of temperature sensing using GdO:Er,Yb: deep learning approach[J]. Physica Scripta,2023,98(11).*

[10] *Daniel L ,G. D A . Topological data analysis and machine learning[J]. Advances in Physics: X,2023,8(1).*