# Educational visualization application based on machine learning algorithm to predict student learning

Jiafu Liu[1], Jiangtao Huang[1*], Ying Xie[1]
1. College of Computer and Information Engineering, Nanning Normal University, Nanning, Guangxi, China
jiafuliu@163.com, Jiangtao@nnnu.edu.cn, 472634310@qq.com
Corresponding Author: Jiangtao Huang Email: Jiangtao@nnnu.edu.cn

***Abstract*—With the development of computer applications, machine learning has become more widely used in the field of education, machine learning can analyze the learning situation of a large number of students more efficiently, and timely understand the learning effectiveness of students and assist in improving learning efficiency. This paper first uses a variety of machine learning models to analyze the relevant factors of student learning, extract the potential relevance of various factors in student learning, and then promote the development of student learning in a positive direction. Then, based on Logistic Regression(LR)、Support Vector Classification(SVC)、Decision Tree(DT)、eXtreme Gradient Boosting(XBG) and Random Forest(RF) predict student learning. Finally, based on the comparison of prediction performance, the prediction performance of random forest algorithm on abnormal data and its reasons are further analyzed. It is clear that educational visualization application based on machine learning algorithm can better provide choices for students' learning and development. And it can get students' learning feedback efficiently, and then improve the quality of education.**

***Keywords—machine learning; learning analysis; random forest; educational visualization***

## I. RESEARCH BACKGROUND

With the continuous development of machine learning, the application of machine learning in the field of education has become more extensive and popular. Based on machine learning and data analysis, we can better understand the degree of knowledge and suitable learning methods for students. And then students can efficiently conduct learning and improve learning effectiveness.

The analysis and prediction of education data can better discover the potential value in the data, explore the potential teaching models that are conducive to the healthy development of education and enhance the learning experience and learning effectiveness of students, and then promote the development of education in a positive direction. Machine learning can provide richer information in education analysis and prediction, and play a positive role in promoting the development of students. At the same time, the prediction method of machine learning in the education field can be used to better analyze the abnormal values of student learning. The reasons of the outliers and relevant solutions can be found by analysing outlier information, so as to help improving the learning effect of students.

Studies have shown that the use of predictive education analysis in student education can actively improve student learning and better promote students to develop in a positive direction[1]. Using machine learning methods to analyze and predict students' learning conditions can more effectively grasp the students' real-time learning dynamics in a more timely and effective manner, thereby improving students' learning effects. The educational prediction based on machine learning can help to analyze the relevant factors used to improve the learning process of students, so that students can develop and improve in a learning direction that is more suitable for them. The learning effect of students has a certain relationship with many factors. Machine learning and educational visualization application can better discover this phenomenon, and can effectively assist students in learning and development.

Analyzing and predicting the learning situation of students can promptly enable students to better understand the direction they need to improve, and then help students to adjust their learning progress or learning modules in time. It can positive effect on improving students' learning performance.

## II. RELATED WORK

Vladimir L. Uskov et al.[2] use machine learning methods to predict and analyze students' academic performance in STEM education, and analyze the characteristics of machine learning algorithms such as linear regression, logistic regression, and K-Nearest Neighbor Classification, and finally analyze the reasons for the different prediction results of STEM education.

Sarfraz Nawaz Brohi et al.[3] compare the accuracy of machine learning algorithms used in higher education predictive analysis, the bagging random forest outperformed other methods with the accuracy value of 0.7959. Through the analysis of the results of random forest and naive Bayes method, determine the characteristics of factors that affect student performance.

Sheshadri Chatterjee et al.[4] provide suitable

2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA) | 978-1-6654-2877-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIBA52610.2021.9688161

individual learning methods through a quantitative analysis using structural equation modelling, and the research emphasizes the identification of factors that influence the adoption of artificial intelligence in higher education. They analyze the application and development of artificial intelligence in higher education through questionnaire surveys, and provide a new solution for the teaching and learning of artificial intelligence in higher education institutions in India.

Yakir Wang et al.[5] analyzed and processed the data of the students through the visualization method of multi-dimensional data. They studied the data information of the art examination scores in different regions, classified the data and analyzed the number of students who took the art examinations in different regions of China.

Sujith Jayaprakash et al.[6] propose a technical classifier for improving the random forest. The random forest algorithm is applied to improve the accuracy and efficiency of the classifier in this research work. The main objective is to introduce an early intervention mechanism to identify the students at risk and the factors influencing the performance. To improve the classification accuracy, various mining techniques are used and discussed.

## III. Analysis of Influencing Factors of Learning

In the analysis of the relevant characteristic values of students' learning, it is easy to find that the performance of students' learning is related to many factors, and different factors have certain influences on students' learning. Through the correlation analysis of the influencing factors of students' learning, the characteristics related to the students' academic performance can be discovered. By analyzing the various factors that affect the learning effectiveness of students in learning, combining the students' personal background and learning situation can give more appropriate learning suggestions, so that the teaching will develop in a more personalized and positive direction.

Prediction and research on student education data, and analysis of students' mastery and adaptive development direction of learning subjects can help students allocate learning time better. In the early learning process, through data mining and visual analysis can better help teachers understand students' learning situation and learning effectiveness. Using the correlation analysis of education data, we can better understand the degree of students' influence on the teaching situation of various elements in learning.

The dataset was obtained from the Kaggle[7]. It contains 1000 student records in rows and 8 features in the columns. The features are classified into gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score, writing score. First, determine the influence of the correlation of the features in the data by analyzing the correlation of the student's feature values. And then by analyzing the correlation between different feature values, the correlation of the feature values of the student education data is shown in Figure 1.

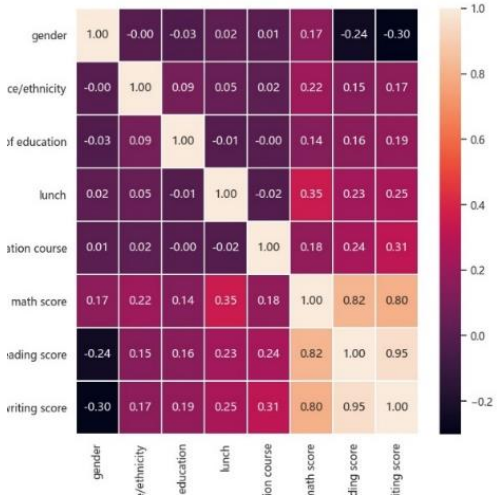
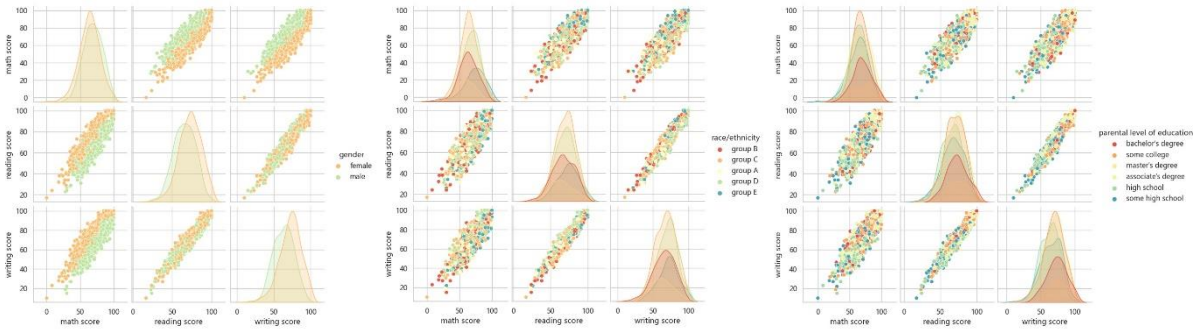



Fig. 1. Correlation of eigenvalues of student education data

In order to understand the degree of influence of each characteristic value on student performance, a scatter diagram was drawn according to gender, race/ethnicity, parent's education level, lunch situation, and preparation for the exam, as shown in Figure 2. From the scatter diagram, it is easy to find that different factors have different effects on students' learning. For example, boys are more suitable for logic learning in mathematics, and girls have higher scores in reading and writing, and so on.



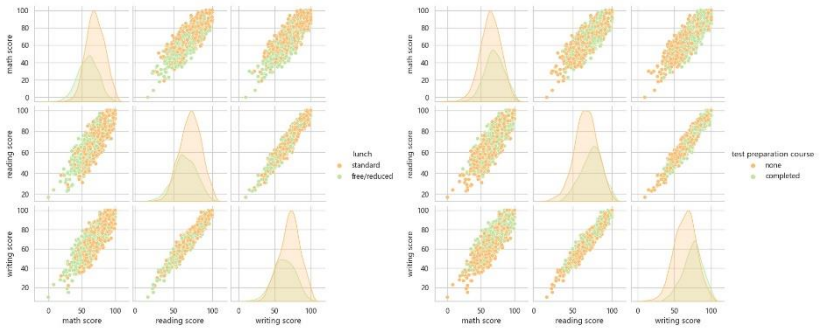



Fig. 2. Scatter plot of student performance based on the characteristic values ofgender, race/ethnicity, parental level of education, lunch, and test preparation course

The students' preparations for the exam are better focused on the children of the parents with the high school education level. The parents and children with the high school education level perform better in the preparation and review before the test. Children pay more attention to review and preparation before exams. At the same time, combined with the data of other groups, it can be found that the students who have done the pre-test preparation have higher mathematics scores. This data shows that the pre-test preparation is helpful for students to study and can better improve students' knowledge. Therefore, the preparation before the test is a relatively important factor in the study of students, which will have a certain impact on the study of the students. The histogram model of the relationship between the preparation for the test and the parents' education is shown in the Figure 3.

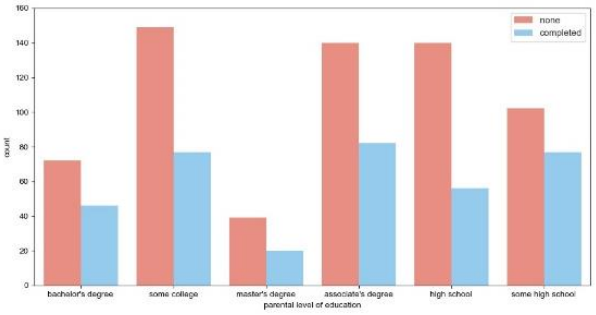



Fig. 3. Histogram of the relationship between parents' educational background and students' preparation for the exam

By observing the scatter plot distribution of reading scores and math scores of students' educational data, it can be found that boys' math scores will be significantly better than girls' scores when reading scores are the same, and there is a significant difference between the scores of boys and girls. The difference in the degree of mastery of learning between boys and girls reflects that boys are more inclined to learn related to logical and logical analysis of mathematics, and girls are more inclined to learn related to perceptual comprehension in reading. The distribution of reading scores and mathematics scores by gender is shown in Figure 4.

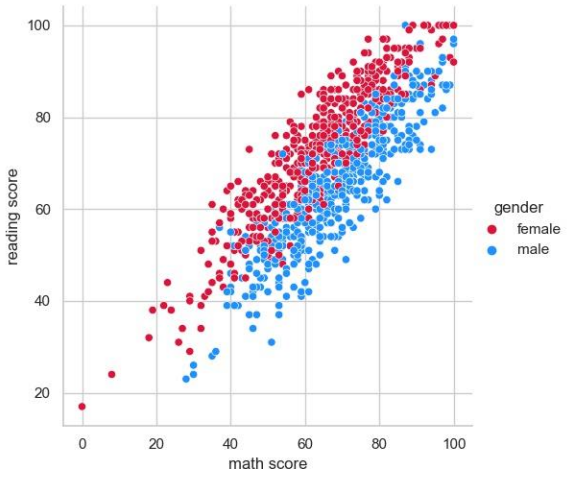



Fig. 4. Scatter plot distribution of reading scores and mathematics scores by gender

Based on the analysis of student scores, it can be found that most of the students' scores are concentrated in the 50-85 interval. Reading scores and writing scores are closer. From the distribution of discrete points, students with better reading scores can also achieve writing scores. Relatively good academic performance, but the distribution of mathematics scores, reading scores, and writing scores is relatively discrete, indicating that the correlation between mathematics scores, reading scores, and writing scores is relatively small. Students with good math scores may not necessarily be in reading and writing. This is so prominent, so the correlation between mathematical rational logic analysis and perceptual comprehension of text is relatively small. The scatter diagram distribution shows a certain correlation between different subjects. Point distribution of students' mathematics, reading, and writing scores is shown in Figure 5.

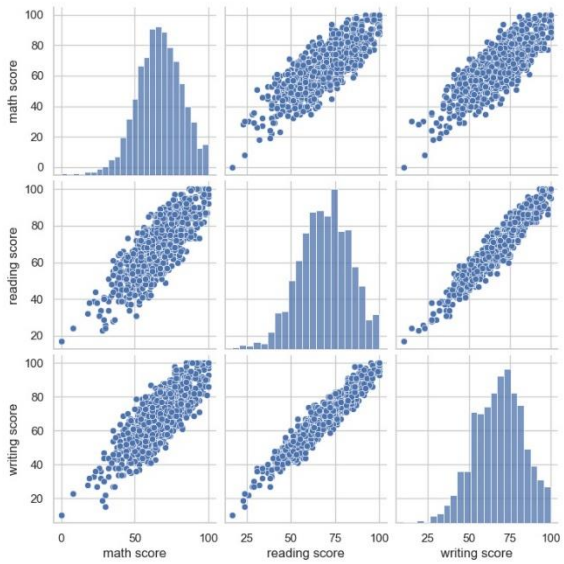



Fig. 5. Scattered distribution of students' math scores, reading scores and writing scores

The three-dimensional scatter model between student achievement and gender is shown in Figure 6. In the three-dimensional coordinate scatter plot, the students' scores of different gender are evenly distributed. It is clear that the math scores of boys are significantly higher than those of girls. At the same time, writing scores and reading scores have a certain correlation.

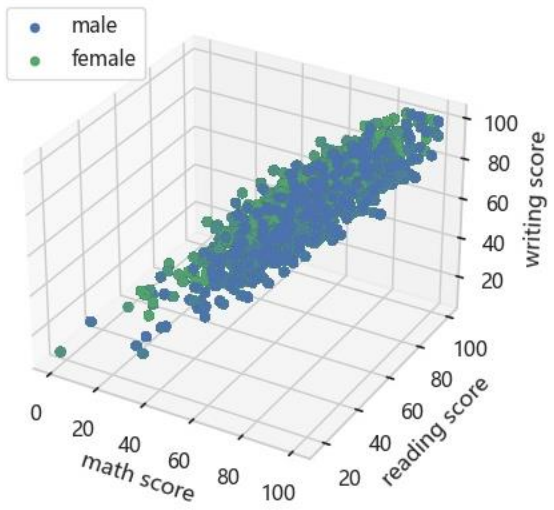



Fig. 6. Three-dimensional coordinate graphics of students' mathematics, reading, and writing performance

## IV. Results and Discussion

This article uses Accuracy, Precision, Recall, F-score, AUC and other measurement indicators to analyze different machine learning methods in predicting the predictive performance of students' mathematics performance.

### *A. Evaluation Indicators*

Confusion matrix is widely used to evaluate the relevant performance indicators of machine learning. Through the confusion matrix, we can analyze the classification of the model, The confusion matrix consists of four elements: true positive(TP), true negative(TN), false positive(FP) and false negative(FN). In the research of this article, accuracy, precision, recall, F-score, AUC are used to evaluate the performance of machine learning algorithms. The description of evaluation indicators is as follows.

Accuracy means the sample that predicted the correct(TP and TN) in the proportion of all samples(TP, FP, FN, TN). Formula to measure the accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision represents the proportion of samples whose true category is positive in the samples predicted to be positive. Precision is related to samples that are truly positive and samples that are truly negative. Formula to measure the precision:

$$Precision = \frac{TP}{TP + FP}$$

Recall represents the proportion of samples successfully predicted by the model in the samples that are truly positive. The recall rate is only related to samples that are truly positive. Formula to measure the recall:

$$Recall = \frac{TP}{TP + FN}$$

The F-score value is the harmonic average of the precision rate and the recall rate. The F-score value is considered as important as the precision rate and the recall rate in the measurement of the model. Formula to measure the F-score:

$$F - score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

AUC is the most commonly used evaluation indicator in model evaluation. The full name is Area Under Curve, which represents the area under the characteristic curve. The AUC value of the ROC characteristic curve can be used to evaluate the relevant model algorithms of machine learning. The value of AUC is [0, 1], the higher the AUC value of the metric in machine learning, the better the performance.

### *B. Experimental Results Analysis*

The research in this article is carried out using python in PyCharm. The data set is divided into 70% and 30%, of which 70% is used as training data and 30% is used as test data. This section will compare and analyze the experimental results of different machine

learning algorithms in student learning analysis.

This article uses machine learning methods such as LR, SVC, DT, RF and XGB to predict student performance. The experimental results are shown in Figure 7. It is easy to find from Figure 7 that the ROC value of the logistic regression method(LR) shows better results than other machine learning methods. By comparing the predictions of different machine learning methods for students' mathematics performance, it can be found that the random forest algorithm is relatively prominent in predicting student performance.

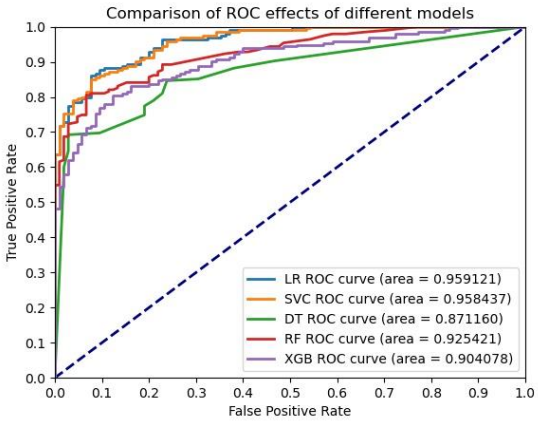



Fig. 7. Test results of machine learning models (LR, SVC, DT, RF, XGB)

Random Forest's ROC value reached 92.5% in the process of predicting students' mathematics scores. In the experiment, Local Outlier Factor(LOF) algorithm was introduced to detect outliers in the random forest, and the related reasons for the low prediction of random forest in machine learning were studied. In the test, it is found that there are many outliers in the process of algorithmic prediction by random forest, and the distance between the data is relatively large. The prediction effect of random forest on students' mathematics performance is shown in Figure 8.

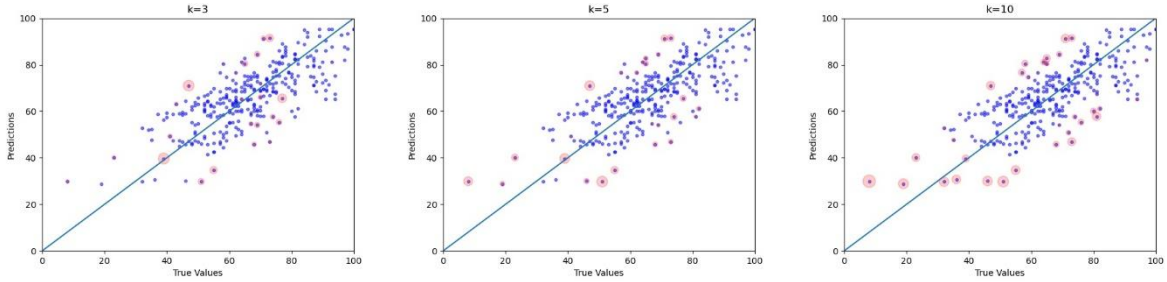



Fig. 8. Random forest predicts abnormal value detection of students' mathematics performance

By comparing the experimental results of different machine learning models, RF algorithm has good prediction performance in predicting students' learning outcome, and it is better than DT and XGB algorithms. But its prediction performance is worse than the LR and SVC algorithms. The reason is that part of the outlier data appeared in the process of RF prediction, and at the same time, the feature value of the data was less in the process of data training using RF. The students' mathematics performance classification performance of different machine learning methods are shown in Table I.

Table I. Classification performance of different machine learning models

| Classifier | Accuracy | Precision | Recall | F-score | AUC |
|---|---|---|---|---|---|
| LR | 0.857 | 0.844 | 0.872 | 0.850 | 0.959 |
| SVC | 0.850 | 0.843 | 0.876 | 0.845 | 0.958 |
| DT | 0.790 | 0.804 | 0.832 | 0.787 | 0.871 |
| RF | 0.817 | 0.813 | 0.844 | 0.812 | 0.925 |
| XGB | 0.790 | 0.794 | 0.823 | 0.786 | 0.904 |

## V. CONCLUSIONS

This paper analyzes the impact of relevant factors in the students learning process through data visualizing the situation of students' education and learning. At the same time, different machine learning methods are used to predict the learning performance, and LOF outlier analysis is used to detect outliers in RF. The related reasons for the appearance of outliers in the prediction process provide more exploration prospects for the application of machine learning in the field of education.

## ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (62067007), the Natural Science Foundation of Guangxi, China (2020GX NSFAA159078).

## REFERENCES

[1] G. Al-Tameemi, J. Xue, S. Ajit, T. Kanakis and I. Hadi, "Predictive Learning Analytics in Higher Education: Factors, Methods and Challenges," 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2020, pp. 1-9, doi: 10.1109/ICACCE49060.2020.9154946.

[2] V. L. Uskov, J. P. Bakken, A. Byerly and A. Shah, "Machine Learning-based Predictive Analytics of Student Academic Performance in STEM Education," 2019 IEEE Global Engineering Education Conference (EDUCON), 2019, pp. 1370-1376, doi: 10.1109/EDUCON.2019.8725237.

[3] Brohi S.N., Pillai T.R., Kaur S., Kaur H., Sukumaran S., Asirvatham D. (2019) Accuracy Comparison of Machine Learning Algorithms for Predictive Analytics in Higher Education. In: Miraz M., Excell P., Ware A., Soomro S., Ali M. (eds) Emerging Technologies in Computing. iCETiC 2019. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 285. Springer, Cham. https://doi.org/10.1007/978-3-030-23943-5_19.

[4] Chatterjee, S., Bhattacharjee, K.K. Adoption of artificial intelligence in higher education: a quantitative analysis using structural equation modelling. Educ Inf Technol 25, 3443–3463 (2020). https://doi.org/10.1007/s10639-020-10159-7.

[5] Y. Wang, M. Shi, C. Li and W. Lin, "Art Exam Scores Analysis and Visualization," 2016 4th Intl Conf on Applied Computing and Information Technology/3rd Intl Conf on Computational Science/Intelligence and Applied Informatics/1st Intl Conf on Big Data, Cloud Computing, Data Science & Engineering (ACIT-CSII-BCD),2016, pp. 384-388, doi: 10.1109/ACIT-CSII-BCD.2016.080.

[6] S. Jayaprakash, S. Krishnan and V. Jaiganesh, "Predicting Students Academic Performance using an Improved Random Forest Classifier," 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), 2020, pp. 238-243, doi: 10.1109/ESCI48226.2020.9167547.

[7] Jakki Seshapanpu: To understand the influence of the parents background, test preparation etc on students performance. Students Performance in Exams. 2018.