





LDM-EDBME: Leveraging Data Mining for Enhancing Development of Basic Mathematics Education at Middle School in Chinese Rural Region

Mingcai Liu¹(✉) , Weixia Lu² , and Yingbiao Hu³ 

¹ Xindong No.1 Middle School, Gaozhou, Guangdong, China
274381445@qq.com

² Wenming Road Primary School, Gaozhou, Guangdong, China

³ Nanjing University of Science and Technology, Nanjing, Jiangshu, China

Abstract. We explore the use of data mining to predict math scores and improve education quality in rural Chinese primary schools. Traditional academic performance evaluation has limitations, and data mining can provide more accurate insights into students' learning situations. By analyzing classification prediction theory, this research sheds light on the factors affecting math performance and provides effective solutions for teachers to help students achieve better academic performance.

Keywords: Education Data Mining · Mathematics · Big Data · Classification

1 Introduction

With the rapid development of the Internet and online education, educational data mining has become a strategic tool in modern educational research. China's education industry has great potential in terms of data resources, but data mining should focus on understanding students' personalities and academic motivations rather than simply developing education. However, transforming diverse education data into academic motivation remains a pressing issue. In rural areas, where educational resources are scarce, data mining can assist teachers in understanding students' learning situations and improving education quality. It is important to continue to expand the scope of educational data mining and strengthen relevant laws and regulations to fully realize its potential.

Wang et al. [1] used data mining and deep learning to explore spatiotemporal information. They tested five classification algorithms on nine datasets. They used big data to analyze data security issues. The team's research shows that classification algorithms can be effectively used for analysis and prediction. Cristobal et al. [2] investigated related conferences in the field of EDM in recent years. Through data mining, they found that educators must have a certain amount of professional knowledge to solve students' learning problems. However, the team did not propose a solution to students' learning

problems. In response to the above problems, Cristobal et al.[3] further studied the use of data mining in application and education. They found that data mining is more widely used in applications. Based on the conclusions of the above team, we began to study the application of data mining in the field of education.

Based on Hanan et al.'s [4] study and analysis of educational data mining in the 21st century, we believe that predicting students' performance requires the continuous use of new methods. Therefore, we will use Logistic Regression (LR) [5], Decision Trees (DT) [6], Random Forest (RF) [7], and other methods to explore the methods of predicting student achievement.

The research and investigation of the above teams have given us great inspiration. Although their research did not propose effective performance prediction methods, they then the possibility of data mining applications in student math performance prediction. Therefore, this paper focuses on the following two points to carry out the design of a performance prediction scheme:

- 1) We use artificial intelligence to analyze the factors that affect students' academic performance. We provide some suggestions for rural education in China from the analysis conclusion.
- 2) We use LR, DT, and RF to analyze the student achievement data set and mine the factors that affect the student's mathematics achievement from the analysis results.

Section 1 introduces the development of data mining in the field of education. Section 2 introduces the theory of LR, DT, and RF. Section 3 analyzes the factors that affect students' performance. Section 4 is the conclusion of this article.

2 Proposed Approach

2.1 Decision Trees

As shown in Fig. 1, Decision tree inducers are algorithms that automatically construct decision trees from a given dataset. The usual goal is to find the optimal decision tree by minimizing the generalization error. However, other objective functions can also be defined, such as minimizing the number of nodes or the average depth.

Inducing an optimal decision tree from a given data is considered to be a difficult task. It has been shown that finding a minimal decision tree consistent with the training set is np-hard [9]. Furthermore, it has been shown that constructing a minimal binary tree is NP-complete in terms of the expected number of tests required to classify an unseen instance [10]. Finding a minimal equivalent decision tree for a given decision tree [11] or constructing an optimal decision tree from a decision table is np-hard [12].

2.2 Random Forest

Mathematically, the estimation of the jth tree is in the form:

$$m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n) = \sum_{i \in \mathcal{D}_n^*(\Theta_j)} \frac{1_{\mathbf{x}_i \in A_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)}^{Y_i}}{N_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)} \quad (1)$$

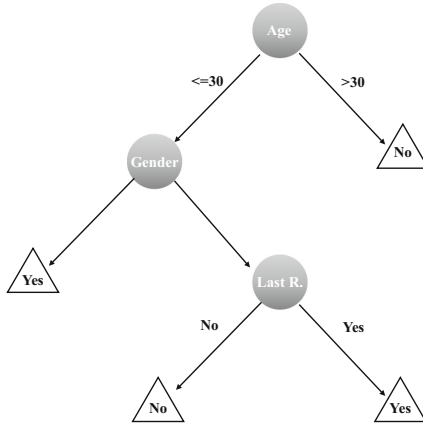


Fig. 1. Decision Tree Presenting Response to Direct Mailing [8].

where $D_n^*(\Theta_j)$ is the set of data points selected before the tree construction, $A_n(\mathbf{x}; \Theta_j, D_n)$ is the cell containing \mathbf{x} , and $N_n(\mathbf{x}; \Theta_j, D_n)$ is the number of preselected points belonging to $A_n(\mathbf{x}; \Theta_j, D_n)$.

3 Experiments

This section mainly explores the factors affecting students’ mathematics performance by analyzing students’ mathematics performance data sets. We use LR, DT, and RF three analysis methods and get the corresponding experimental results. We perform our experiments on one large-scale dataset of student performance: Student Performance Data Set [13].

3.1 Experimental Result Analysis

The experiment found a relationship between long absence time and the child’s age of 17–18. Parental education level also affects a child’s performance, with stable results when parents have received education and unstable results when they haven’t. The distance from home to school and alcohol consumption were also found to negatively impact student achievement.

Figure 2 further explores the factors affecting students’ mathematics performance through thermal images. We can see from the thermal image analysis that students’ math scores are unrelated to the number of failures. However, there is an excellent correlation between students’ mathematics achievement and learning time, service education level, etc. G1 represents foreign language achievement, and the results presented in the figure indicate a strong correlation between foreign language achievement and mathematics achievement.

After analyzing Fig. 3, we can conclude that the number of student failures is related to gender. The number of failures of male students will increase with age. For female students between 15–17 years old, the number of failures gradually reduced. Nevertheless, female students will fail more often with the increase in age.

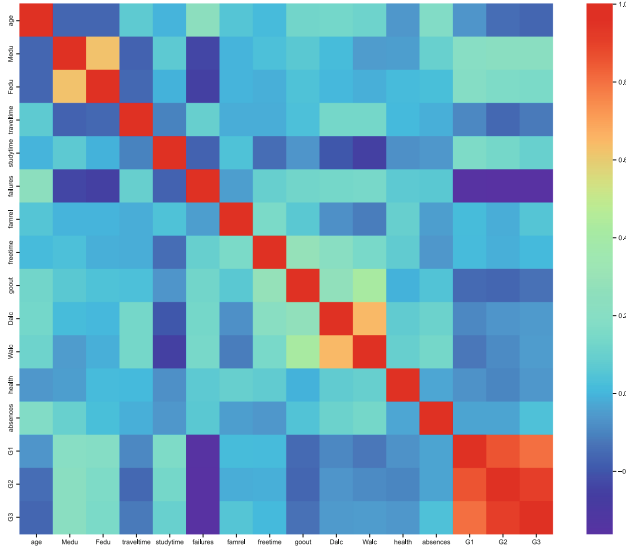


Fig. 2. Thermal image analysis of influencing factors of students mathematics achievement

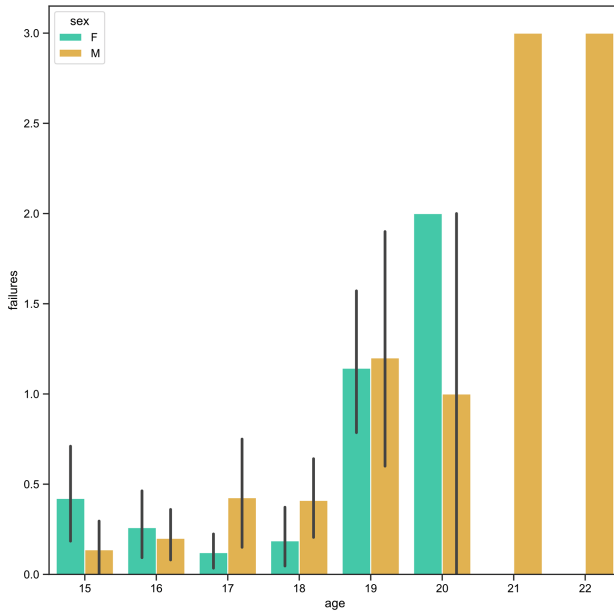


Fig. 3. The relationship between the number of student failures and gender

3.2 Performance Prediction Model Analysis

We use logistic regression, decision tree and random forest to predict students' math scores. The figure of logistic regression prediction is as follows:

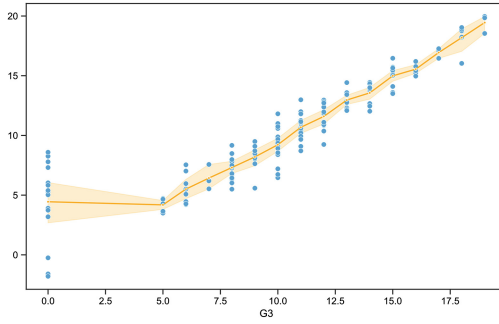


Fig. 4. Logistic Regression Analysis

Finally, a random forest model was employed to predict students' math scores, and the prediction results are displayed in Fig. 6.

In Fig. 4 ormanance and actual performance of the model on the test dataset.

To predict students' math scores, a decision tree was used and the results are shown in Fig. 5.

We use logistic regression, decision trees, and random forest to predict students math scores. The result of machine learning prediction is as Table 1:

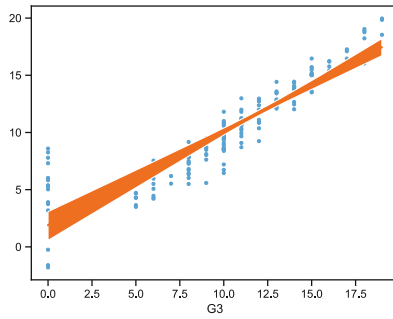


Fig. 5. Decision Trees Analysis

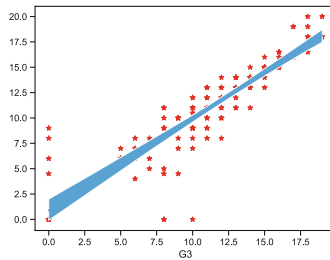


Fig. 6. Random Forest Analysis

Table 1. The result of machine learning prediction

Model	MSE	R2 Score
LR	4.01186155181831	0.8172756829811947
DT	4.319620253164557	0.8032584997400133
RF	2.713249363737478	0.876422296610426

4 Conclusion

This study utilized visual analysis methods such as heat maps and scatter plots to investigate the effects of parental education, distance between home and school, alcohol consumption, and language performance on student performance. The results showed that the performance in the previous exam often predicted the performance in the subsequent exam, and that language teaching played an important role in improving student performance. Academic achievement and language teaching were found to be closely related. Finally, the effectiveness of three prediction methods, logistic regression (LR), decision tree (DT), and random forest (RF), was compared using metrics such as mean squared error (MSE) and R2 score. The research results indicated that the prediction accuracy of the random forest model was the highest.

References

1. Wang, S., Cao, J., Yu, P.: Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering* (2020)
2. Romero, C., Ventura, S.: Data mining in education. *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery* 3, 12-27 (2013)
3. Romero, C., Ventura, S.: Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, e1355 (2020)
4. Aldowah, H., Al-Samarraie, H., Fauzy, W.M.: Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics* 37, 13-49 (2019)
5. LaValley, M.P.: Logistic regression. *Circulation* 117, 2395-2399 (2008)
6. Kotsiantis, S.B.: Decision trees: a recent overview. *Artificial Intelligence Review* 39, 261-283 (2013)
7. Breiman, L.: Random forests. *Machine learning* 45, 5-32 (2001)
8. Rokach, L., Maimon, O.: Decision trees. *Data mining and knowledge discovery handbook* 165-192 (2005)
9. Hancock, T., Jiang, T., Li, M., Tromp, J.: Lower bounds on learning decision lists and trees. *Information and Computation* 126, 114-122 (1996)
10. Laurent, H., Rivest, R.L.: Constructing optimal binary decision trees is NP-complete. *Information processing letters* 5, 15-17 (1976)
11. Zantema, H., Bodlaender, H.L.: Finding small equivalent decision trees is hard. (1999)
12. Naumov, G.: NP-completeness of problems of construction of optimal decision trees. In: *Soviet Physics Doklady*, p. 270. (Year)
13. Cortez, P., Silva, A.M.G.: Using data mining to predict secondary school student performance. (2008)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

