Eng Int Syst (2022) 6: 447–452


**Engineering**
***Intelligent Systems***

# Research on Online Creativity Education for College Students Using Data Mining

**Kun Zhou***

*Institute of Education, Xiamen University, Xiamen, Fujian 361000, China*

Creativity education plays a very important role in the future development of students. This paper analyzed the effectiveness of two decision tree methods: C5.0 and classification and regression tree (CART) algorithms, using data mining, in predicting the performance of students enrolled in online education, discussed the data collection and processing methods, selected 12 features as input, and conducted an example analysis of the data set. It was found that the decision tree method achieved better performance and higher prediction precision in comparison with the nearest neighbor algorithm, and the accuracy and F1-measure value of the C5.0 algorithm was 92.59% and 92.53% respectively, which was better than the CART algorithm. The analysis of the decision tree demonstrated that the factors which had the greatest impact on students' grades were the fifth and sixth chapter test scores, the number of questions answered correctly, etc. The experiment results verify the effectiveness of the decision tree method in predicting students' performance in online creativity education, which can be further applied in practice.

Keywords: data mining, colleges, online education, creativity education, decision tree

## 1. INTRODUCTION

With the development of information technology, people's living and working styles have changed greatly, and education has entered the information era. With the expansion of enrollments, educational methods have become more diverse (Zheng and Zhou, 2021) and online education has become increasingly popular in colleges and universities. Unlike offline education, the teacher and students are in different spaces during online education, and the number of students in online classes tends to be very large. Therefore, research on online education has received a lot of attention from researchers. Educational data mining (EDM) is to (Hegazi and Abugroon, 2016) study the data in the educational process and find valuable information using data mining to help teachers better plan their teaching. In recent times, an increasing number of techniques and methods have been applied in EDM (Agaoglu, 2016). Black et al. (2021) used a random forest approach to predict the status of students in physician assistant education and performed leave-one-out cross-validation and bootstrap aggregation on the samples. They found that the method obtained a positive predictive value of 63.3%, i.e., it was able to identify learners who were likely to encounter academic challenges. Neto et al. (2016) conducted a study on students' programming learning and identified learners with learning difficulties using data mining and Bloom's Taxonomy to improve the learning process of these students. Thangakumar et al. (2020) selected the features of students involved in learning tasks using a group of algorithms and used logistic regression methods to classify the data, finding that the method had a 94.91% accuracy, 97.02% F-measure value and 79.57% kappa value. Wang et al. (2021) mined relevant patterns from student performance data using the Apriori algorithm and established an early warning mechanism for student performance based on decision trees. The results found that C language courses were dependent on other courses such as higher mathematics and linear algebra and the teaching

*Corresponding address: Room 1502, Building 134, Hongwenyili, Siming District, Xiamen, Fujian 361000, China. Email: k2868q@yeah.net.

**Table 1** Feature selection.

| Features | Description |
|---|---|
| Chapter1_score$_i$(C1) | Chapter 1 test results |
| Chapter2_score$_i$(C2) | Chapter 2 test results |
| Chapter3_score$_i$(C3) | Chapter 3 test results |
| Chapter4_score$_i$(C4) | Chapter 4 test results |
| Chapter5_score$_i$(C5) | Chapter 5 test results |
| Chapter6_score$_i$(C6) | Chapter 6 test results |
| coursetime | Course online hours |
| loginnum | Total number of logins |
| Total_correct_answers | Number of correct answers to questions |
| First_got_correct | Number of questions answered correctly on the first attempt |
| Total_time_spent | Total time spent answering questions |
| Communication | Time to interact with other students |

method of C language $\rightarrow$ C++ $\rightarrow$ Java was more consistent with the learning mechanism. This paper studies online creativity education for college students and analyzes the effect of the decision tree algorithm for the purpose of grade prediction to contribute to improving online creativity education and promote the better development of online education.

## 2. ONLINE CREATIVITY EDUCATION FOR COLLEGE STUDENTS

Education shapes the character and intelligence of individuals (Aithal, 2016). Creativity plays a very important role in the development of society, and creativity education in colleges and universities has been highly valued by the state and schools (Shkabarina et al., 2020). Strengthening creativity education and cultivating the creative spirit of college students is not only beneficial to improving the overall quality of college students, it also has an important relationship with economic and social development (Gulicheva et al., 2017). With the development of network technology, online education is becoming increasingly popular, which also provides a new channel for the implementation of creativity education.

Online education breaks through the limitations of time and space, enabling learners to study anytime and anywhere and it also provides a wider range of educational resources to better meet the learning needs of learners. Moreover, the special nature of online education enables students to arrange their learning without being limited by the number of credit hours and credits, maximizing their initiative and motivation. Finally, online education can also aggregate all relevant data on the web platform, which is more conducive to EDM. Therefore, this paper focuses on the prediction of student achievement using a decision tree-based approach for students enrolled in online education.

## 3. DECISION TREE-BASED PERFORMANCE PREDICTION METHOD

### 3.1 Data Collection and Processing

A large amount of useful data exists in online education systems. This paper investigates whether it is possible to predict if students will fail an online course or not using the data in online education systems. The online course under investigation is divided into six chapters. After the students have completed their study of each chapter, they are required to complete an online test. During their study of each chapter, the students are required to answer questions which appear in pop-up windows. The selected data features are shown in Table 1.

For missing data, upsampling is used to fill in the data using the closest data. For duplicate data, simple de-duplication is performed, and the last submitted data are reserved. For erroneous data, such as null values and abnormal values, iteration deletion is used. Upsampling is used again to fill in the data.

### 3.2 Decision Tree Algorithm

A decision tree algorithm is a classification algorithm, each branch of which is a decision process. It has high accuracy and a relatively simple computational process, with a very wide range of applications in data mining (Decaestecker et al., 2015). The most commonly used decision tree algorithms include ID3, C4.5, C5.0 and classification and regression tree (CART) algorithms (Ngoc et al., 2017). This paper compares two algorithms, C5.0 and CART algorithms.

The C5.0 algorithm is a further improvement of the C4.5 algorithm (Siknun and Sitanggang, 2016), which combines the information gain rate and boosting algorithm to further improve both the computational speed and efficiency of the model, and to be able to process multiple types of data. Suppose there is a data set T, containing the following classes, $\{C_1,\ C_2,\ \cdots,\ C_k\}$. The data set $T$ is divided into multiple subsets using attribute $V$.

Suppose there is $\{v_1, v_2, \cdots, v_k\}$. $T$ is divided into $n$ subsets, $T_1,\ T_2,\ \cdots, T_n$ . Let $|T|$ be the number of examples of $T$, $|T_i|$ be the number of examples of $V = v_i$, $|C_j| = freq(C_j, T)$ be the number of examples of $C_j$, and $|C_j v|$ be the number of examples whose class is $C_j$ in $V = v_i$. Then, the occurrence probability of class $C_j$ can be written as:

$$\mathrm{P}(C_j) = \frac{freq(C_j, T)}{|T|}.$$

The occurrence probability of attribute $V = v_i$ can be written as:

$$\mathrm{P}(v_i) = \frac{|T_i|}{|T|}.$$

The conditional probability when the class in $\mathrm{V} = v_i$ is $C_j$ can be written as:

$$\mathrm{P}(C_j|v_i) = \frac{|C_j v|}{|T_i|}.$$

The information entropy of a class can be written as:

$$\mathrm{H}(C) = -\sum_j P(C_j)\log_2(P(C_j)) = \mathrm{info}(T).$$

The conditional entropy of a class can be written as:

$$\mathrm{H}(\frac{C}{V}) = -\sum P(v_i)\sum P\left(\frac{C_j}{v_i}\right)\log_2\left(P\left(\frac{C_j}{v_i}\right)\right) = \mathrm{infov}(T). \quad (1)$$

The information gain can be written as:

$$I(C,\ V) = H(C) - H(\frac{C}{V}) = info(T) - infov(T) = gain(V).$$

The information entropy of attribute $V$ can be written as:

$$H(V) = \sum_i P(v_i)\log_2(P(v_i)) = split\_info(V).$$

The information gain rate can be written as:

$$gain\_ratio = \frac{I(C,V)}{H(V)} = \frac{gain(V)}{split\_info(V)}.$$

Before predicting the online students' results using the C5.0 algorithm, the collected data set is preprocessed to form a training set and a test set. Then, the gain_ratio of each attribute in the training set is calculated and the attributes that are maximum and not lower than the average value are selected as the main nodes to build a decision tree branch and generate child nodes. The operation is repeated until the initial decision tree is built. The initial decision tree is trimmed. The rules from the root node to the leaf node are extracted to form a rule set, and the established rule set can be used to classify the test set and predict students' performance.

The main difference between the CART algorithm (Breiman et al., 2015) and the C5.0 algorithm is that the CART algorithm uses the Gini index for feature selection, and the feature with the smallest Gini partition index is used as the current node to construct the decision tree.

It is assumed that sample $D$ has $m$ data and $n$ decision attribute values. Then, the calculation method of the Gini coefficient can be written as:

$$Gini(D) = 1 - \sum_{i=0}^{n} p_i^2,$$

where $p_i$ refers to the relative probability of class $i$ in sample set $D$. If the sample set is divided into two sub-samples, $D_1$ and $D_2$, then its Gini split index can be written as:

$$Gini_{split}(D) = \frac{m_1}{m}Gini(D_1) + \frac{m_2}{m}Gini(D_2),$$

where $m_1$ and $m_2$ refer to the number of samples in sub-samples $D_1$ and $D_2$, respectively.

The other calculations in the CART algorithm are the same as the C5.0 algorithm.

**Table 2** Confusion matrix.

| | Positive sample | Negative sample |
|---|---|---|
| Positive sample | TP | FN |
| Negative sample | FP | TN |

## 4. EXPERIMENT ANALYSIS

Experiments were conducted using a Windows $10 \times 64$ operating system with an 8 GB memory. The integrated development environment was PyCharm3.1+Python3.7. The algorithm implementation language was Python. The experiment data were obtained from the online education course system offered by Xiamen University. After data cleaning and processing, a total of 1260 data were obtained. The predicted results have two categories: higher than 60 points (not failed) and lower than 60 points (failed). The performance of C5.0 and CART algorithms on grade prediction was compared. The nearest neighbor algorithm (Adeniyi et al., 2016) was used as a comparison.

Based on the confusion matrix (Table 2), the evaluation indexes of the algorithm are:

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

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

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

$$\mathrm{F1-measure} = \frac{2\times\mathrm{Recall}\times\mathrm{Precision}}{\mathrm{Recall}+\mathrm{Precision}}.$$

To improve the accuracy, a 10-fold cross-validation method was used, which means that the data set was divided equally into 10 parts, one of which was used as the test sample and the remaining nine parts were used as the training samples. The test was repeated 10 times. The results of each repetition were averaged to obtain the final result.

The accuracy of the three algorithms is shown in Figure 1.

It can be seen from Figure 1 that the accuracy of the nearest neighbor algorithm was the lowest (72.36%), while the accuracy of both C5.0 and the CART algorithms was better than the nearest neighbor algorithm, indicating that the decision tree method achieved better performance in predicting students' grades. The accuracy of the CART algorithm was 87.24%, which was 14.88% higher than the nearest neighbor algorithm. The accuracy of the C5.0 algorithm was 95.59%, which was 20.23% higher than the nearest neighbor algorithm and 5.35% higher than the CART algorithm.

The results of precision, recall rate, and F1-measure value are shown in Table 3.

It can be seen from Table 3 that the precision of the nearest neighbor algorithm was the lowest (85.64%), followed by the CART algorithm (87.32%) and the C5.0 algorithm (91.89%), the precision of the CART algorithm was 1.68% higher than that of the nearest neighbor algorithm, and the precision of the C5.0 algorithm was 6.25% higher than the nearest neighbor algorithm and 4.57% higher than the CART

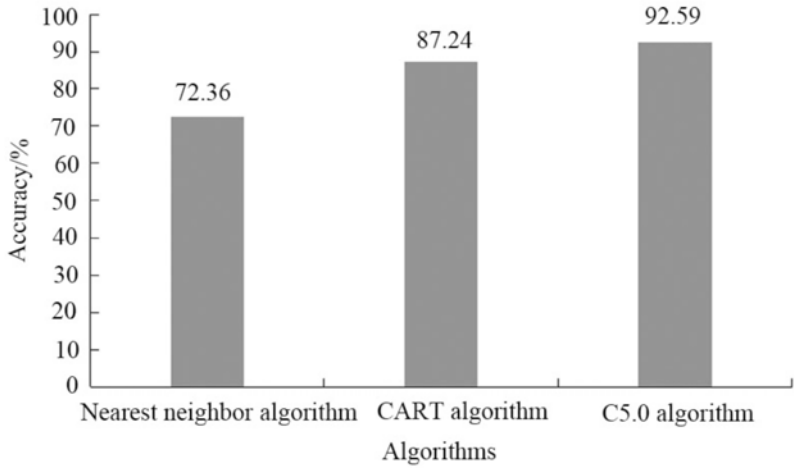


**Figure 1** Comparison of accuracy.

**Table 3** Comparison of precision, recall rate, and F1-measure value.

| | Nearest neighbor algorithm | CART algorithm | C5.0 algorithm |
|---|---|---|---|
| Accuracy/% | 85.64 | 87.32 | 91.89 |
| Recall rate/% | 59.64 | 78.21 | 93.18 |
| F1-measure value/% | 70.31 | 82.51 | 92.53 |

**Table 4** Info and gain values for each feature.

| Feature | info | gain |
|---|---|---|
| Chapter1_score$_i$(C1) | 0.803 | 0.025 |
| Chapter2_score$_i$(C2) | 0.805 | 0.021 |
| Chapter3_score$_i$(C3) | 0.812 | 0.027 |
| Chapter4_score$_i$(C4) | 0.808 | 0.035 |
| Chapter5_score$_i$(C5) | 0.901 | 0.147 |
| Chapter6_score$_i$(C6) | 0.899 | 0.142 |
| coursetime | 0.879 | 0.035 |
| loginnum | 0.822 | 0.091 |
| Total_correct_answers | 0.886 | 0.127 |
| First_got_correct | 0.872 | 0.119 |
| Total_time_spent | 0.831 | 0.027 |
| Commucation | 0.854 | 0.108 |

algorithm; the recall rate of the C5.0 algorithm was 33.54% higher than the nearest neighbor algorithm and 14.97 % higher than the CART algorithm. The F1-measure value of the CART algorithm (82.51%) was 12.2% higher than the nearest neighbor algorithm (70.31%), and the F1-measure value of the C5.0 algorithm (92.53%) was 22.22% higher than the nearest neighbor algorithm and 10.02% higher than the CART algorithm. In a comprehensive view, it was found that the C5.0 algorithm achieved better performance in predicting students' online performance.

Therefore, the researchers chose the C5.0 algorithm to predict students' online performance. The information and gain of the 12 features selected in this paper were calculated and the results are shown in Table 4.

It can be seen from Table 4 that Chapter5_score$_i$(C5) and Chapter6_score$_i$(C6) have the largest gain values of the 12 features, which indicates that the test scores of these two chapters had the greatest influence on whether students failed the final exam or not in online courses. Therefore, in the learning process, both students and teachers should pay more attention to the study of these two chapters. The features Total_correct_answers and First_got_correct also had large gain values, indicating that the solution of pop-up questions could, to a certain extent, reflect the students' learning level in online education. The more questions answered correctly and quickly, the better the mastery of the content, and the more likely the student is to pass the final exam.

## 5. DISCUSSION

Data mining is a method that enables the processing and analysis of data (Chamizo-Gonzalez et al., 2015). In the process of online education for college students, a large volume of data exist, such as students' attendance, the number of answers to questions, the correctness of submitted assignments, etc. After collecting, processing, counting and analyzing these data, a lot of useful information can

be obtained. With the continuous development of online education, the amount of data related to education is also growing (Huda et al., 2016), and the diversity and accessibility of data have promoted the development of EDM (Río and Insuasti, 2016). Currently, the research focus of EDM broadly includes learner models (Shrestha and Pokharel, 2020), grade prediction (Kumar et al., 2017), factor analysis (Zhang and Luo, 2021), teaching evaluation (Wang, 2019) and emotional analysis (Shi, 2019), and research on EDM can provide suggestions for the improvement and innovation of education, help students find better learning styles, and give early warnings for particular student behaviors such as retaking courses and dropping out, which are very important for the development of the whole education field. Especially in online education, a large amount of basic data is stored in the academic system, which is conducive to the organization and analysis of these data.

This study takes the online education course as an example to study the performance of two decision tree algorithms, C5.0 and CART algorithms, for student achievement prediction. Firstly, the decision tree algorithms achieved significantly better performance than the nearest neighbor algorithm. As seen from Figure 1, both C5.0 and CART algorithms had higher accuracy than the nearest neighbor algorithm, which means that the decision tree algorithms achieved better performance in data prediction and obtained more accurate results. A comparison of the other algorithm indicators shows that the nearest neighbor algorithm had poorer performance, and its F1-measure value was only 70.31%, which does not meet the requirements of performance prediction. The accuracy, precision and recall rate of the C5.0 algorithm were all higher than the CART algorithm, indicating that the performance of the C5.0 algorithm was better than that of the CART algorithm. Therefore, the C5.0 algorithm was chosen to be applied to the performance prediction of students enrolled in online education. The calculation and comparison of the information gain values shows that the test scores of chapters 5 and 6, the total number of questions answered correctly and the number of questions answered correctly on the first attempt had a great impact on the final grade. Therefore, in future study and teaching, we can start from these aspects to further improve students' experience of online education.

Although some results on EDM have been achieved in this paper, there are several shortcomings. In future research, more data mining methods will be analyzed to further improve the prediction performance, and data mining will be performed on more aspects of online education to further improve the level of online education.

## 6. CONCLUSION

In this paper based on data mining, the decision tree method was used to study the performance prediction of online education courses for college students and an analysis was carried out. It was found that the C5.0 algorithm achieved good performance in terms of accuracy and recall rate and was able to accurately predict students' performance. In addition, the analysis of information gain values found the key indicators that affected the results of the final exams, making some contributions to the adjustment of teachers' teaching programs and students' learning styles.

## REFERENCES


1. Adeniyi, D.A., Wei, Z. & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing & Informatics*, *12*(1), 90–108.
2. Agaoglu, M. (2016). Predicting Instructor Performance Using Data Mining Techniques in Higher Education. *IEEE Access*, *4*, 2379–2387.
3. Aithal, S. (2016). An Innovative Education Model to realize Ideal Education System. *International Journal of Scientific Research & Management Studies*, *3*(3), 2464–2469.
4. Black, E.W., Buchs, S.R. & Garbas, B. (2021). Using Data Mining for the Early Identification of Struggling Learners in Physician Assistant Education. *Journal of Physician Assistant Education*, *32*(1), 38–42.
5. Breiman, L., Friedman, J., Olshen, R. & Stone, C. (2015). Classification and Regression Trees. Belmont, CA: Wadsworth International Group. *Encyclopedia of Ecology*, *57*(1), 582–588.
6. Chamizo-Gonzalez, J., Cano-Montero, E.I., Urquia-Grande, E. and Muñoz-Colomina, C.I. (2015). Educational data mining for improving learning outcomes in teaching accounting within higher education. *International Journal of Information & Learning Technology*, *32*(5), 272–285.
7. Decaestecker, C., van Velthoven, R., Petein, M., Janssen, T., Salmon, I., Pasteels, J., van Ham, P., Schulman, C. & Kiss, R. (2015). The use of the decision tree technique and image cytometry to characterize aggressiveness in World Health Organization (WHO) grade II superficial transitional cell carcinomas of the bladder. *Journal of Pathology*, *178*(3), 274–283.
8. Gulicheva, E., Lisin, E., Osipova, M. & Khabdullin, A. (2017). Leading factors in the formation of innovative education environment. *Journal of International Studies*, *10*(2), 129–137.
9. Hegazi, M.O. & Abugroon, M.A. (2016). The State of the Art on Educational Data Mining in Higher Education. *International Journal of Emerging Trends & Technology in Computer Science*, *31*(1), 46–56.
10. Huda, M., Anshari, M., Almunawar, M.N., Shahrill, M., Tan, A., Jaidin, J.H., Daud, S. & Masri, M. (2016). Innovative Teaching In Higher Education: The Big Data Approach. *Turkish Online Journal of Educational Technology*, *15*(Special issue), 1210–1216.
11. Kumar, M., Singh, A.J. & Handa, D. (2017). Literature Survey on Student's Performance Prediction in Education using Data Mining Techniques. *International Journal of Education and Management Engineering*, *6*(6), 40–49.
12. Neto, V. (2016). Apprentices Identifying Groups with Difficulties in Programming Education Using Data Mining. *The International Journal of E-Learning and Educational Technologies in the Digital Media*, *2*(2), 59–72.
13. Ngoc, P.V., Ngoc, C., Ngoc, T. & Dat, N.D. (2017). A C4.5 algorithm for English emotional classification. *Evolving Systems*, 1–27.
14. Río, C.A.D. & Insuasti, J.A.P. (2016). Predicting academic performance in traditional environments at higher-education institutions using data mining: A review. *Ecos de la Academia*, *4*(Diciembre), 2016.
15. Shi, X. (2019). Emotional Data Mining and Machine Learning in College Students Psychological Cognitive Education. *Engineering Intelligent Systems*, *27(*4), 167–175.

16. Shkabarina, M.A., Verbytska, K., Vitiuk, V., Shemchuk, V. & Saleychuk, E. (2020). Development of Pedagogical Creativity of Future Teachers of Primary School by Means of Innovative Education Technologies. *Revista Romaneasca pentru Educatie Multidimensionala*, *12*(4), 137–155.
17. Shrestha, S. & Pokharel, M. (2020). Data Mining Applications Used in Education Sector. *Journal of Education and Research*, *10*(2), 27–51.
18. Siknun, G.P. & Sitanggang, I.S. (2016). Web-based Classification Application for Forest Fire Data Using the Shiny Framework and the C5.0 Algorithm. *Procedia Environmental Sciences*, *33*, 332–339.
19. Thangakumar, J. & Kommina, S.B. (2020). Ant Colony Optimization Based Feature Subset Selection with Logistic Regression Classification Model for Education Data Mining. *International Journal of Advanced Science and Technology*, *29*(3), 5821–5834.
20. Wang, L. & Chung, S.J. (2021). Sustainable Development of College and University Education by use of Data Mining Methods. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(5), 102.
21. Wang, L. (2019). Evaluation of Web-Based Teaching Based on Machine Learning and Text Emotion. *Engineering Intelligent Systems*, *27*(3), 111–119.
22. Zhang, X. & Luo, P. (2021). Analysis of psychological education factors based on computer software and hardware collaboration and data mining. *Microprocessors and Microsystems*, *81*, 103744.
23. Zheng, C. & Zhou, W. (2021). Research on Information Construction and Management of Education Management Based on Data Mining. *Journal of Physics: Conference Series*, *1881*(4), 042073 (6pp).