Eng Int Syst (2024) 3: 127–132

Engineering
Intelligent Systems

Data Mining: Analysis of the Influence of College Students' Daily Behaviors on Their Academic Performance

Yujia Zhang*

Chengdu Polytechnic, Chengdu, Sichuan 610041, China

Students' academic performance reflects the quality of a university. It is of vital practical significance to find the daily behaviors that affect students' academic performance and improve them in a targeted manner. This paper mined and analyzed the academic performance and daily behaviors of college students by using the decision tree algorithm in data mining technology to find the correlation between the students' daily behaviors and their academic performance. By means of a decision tree model, it was found that students' class attendance had the greatest influence on their academic performance, followed by the time spent daily on the Internet. Moreover, the decision tree model achieved a correct rate of 82.07%. The study shows that the decision tree algorithm can identify the daily behaviors that affect college students' performance, help them study better, and have a positive impact on improving college students' academic performance.

Keywords: data mining, performance analysis, decision tree

1. INTRODUCTION

Learning is of the utmost importance, and academic performance is the best way to test students' learning outcomes. Students' daily behaviors have an influence on their academic performance (Ajai and Shiaki, 2020), and good learning habits can promote students' learning. We often find that a high-achieving student tends to have excellent daily habits. Scholars often use data mining techniques to determine those behaviors that have the greatest impact on academic performance. Data mining techniques (Zhou, 2022) include clustering analysis, association rules, neural networks (Bokadiya et al., 2018), and decision trees (Wang et al., 2018), to name a few. Related studies are reviewed below.

Lin analyzed the collected data using data mining title association models and found that the paper-based algorithm based on the diagnostic evaluation model could provide students with better practice guidance and test question recommendations that could improve their learning (Lin, 2020). Using data mining techniques, Saini et al. extracted useful information and found that Indian students with greater social media engagement had higher final grades (Saini and Sood, 2019). Chen et al., by applying cluster analysis used in data mining technology, found that the system they proposed for automatically tracking students' learning progress could affect their final grades (Chen et al., 2020). Xiao et al. constructed a predictive model and found that data mining analysis based on xAPI data had better predictive accuracy in terms of online learning data (Xiao et al., 2020). Pandiangan et al. compared the performance of the decision tree and Naive Bayes algorithms when applied to the assessment of learning achievement and found that the decision tree algorithm obtained more accurate data (Pandiangan et al., 2020). In this study, data mining technology was used

*Corresponding address: No. 56, Dashi West Road, Qingyang District, Chengdu, Sichuan 610041, China. Email: jia30855330@163.com

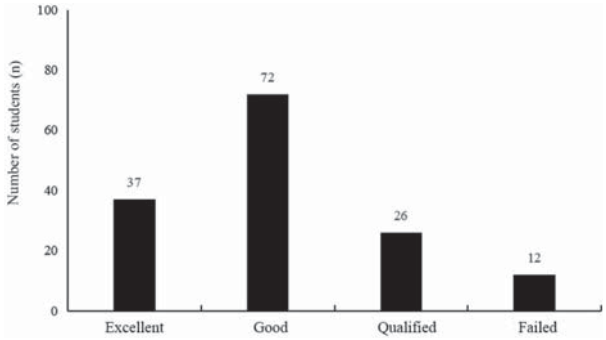

Figure 1 Distribution of students at different performance levels.

to determine how the daily behaviors of college students influence their grades. The decision tree algorithm in data mining technology was applied to explore the correlation between students' academic performance and their daily behaviors such as their class attendance, daily online time, library study frequency, pre/post-class pre/review of course material, homework completion rate. The aim of this study was to provide, through decision tree analysis, a theoretical basis for determining the influence of college students' daily behaviors on their learning process, targeting those behaviors that should be encouraged so as to improve students' academic outcomes.

2. DECISION TREE ALGORITHM

In this paper, a data mining approach, decision tree classification (Tang, 2022), was used to analyze the effect of college students' daily behaviors on their academic performance. To classify based on the decision tree, it is necessary to find an optimal divided feature attribute among all the unclassified feature attributes, which is the key to the success of the decision tree classification. Nowadays, the methods most commonly used to select the optimal segmentation features are information gain (Suryakanthi, 2020), information gain ratio, etc. The method chosen for this paper is information gain. Entropy (Arellano et al., 2018) refers to the uncertainty of the value of random variables. If the uncertainty of the variable is higher, then the value of entropy is higher, and vice versa. Information gain is measured by how much the entropy value decreases; if the entropy value is 10 originally and then changed to 4, the information gain is 6.

Suppose A is an original data set containing m categories of samples. The calculation formula for information entropy is:

$$H(A) = -\sum_{i=1}^{m} p(x_i)\log_2 p(x_i),$$

where $p(x_i)$ is the occurrence probability of the x_i-th category.

It is assumed that attribute B divides A into n parts, so the following equation is derived from the entropy of the subsets divided by B:

$$H(A, B) = \sum_{i=1}^{n} \frac{|A_i|}{|A|} H(A_i),$$

where A_i is the i-th subset of A divided by attribute B and $|A_i|$ and $|A|$ are the number of samples in A_i and A. Then, the formula for calculating the information gain is obtained:

$$g(A, B) = H(A) - H(A, B),$$

where $g(A, B)$ is the decreasing value of the entropy derived after knowing the value of B attribute.

3. ANALYSIS OF COLLEGE STUDENTS' OVERALL ACADEMIC PERFORMANCE AND THEIR DAILY BEHAVIORS

(1) Analysis of performance

The objective classification of the grades of every student becomes problematic due to the diversity of courses in colleges and the differences between different majors and optional units. In this paper, the 4.5 grade point average (GPA) system was used to discretize students' total grades (Tsai and Chen, 2019). The end-of-semester total grades of 147 students in three classes who were enrolled at Chengdu Polytechnic in 2021 were taken as the initial data. The GPAs of these students were classified thus: 4.0–4.5 = excellent, 3.0–4.0 = good, 2.5–3.0 = qualified, and below 2.5 = failed, as shown in Figure 1.

(2) Analysis of daily behaviors

The data on the daily behaviors of the 147 college students were collected by means of an electronic questionnaire. The questionnaire collected students' demographic information as well as data on their daily behaviors, such as class attendance, daily online time, library study frequency, and pre/review of course material before and after class. The students were ranked according to their student number and numbered from 1 to 147. After completion, the questionnaires were collected and sorted.

The newly-collected data were confusing, and also contained some invalid and incomplete information. Before data mining technology was applied to build a decision tree model, the data were cleaned. As two students did not attend all the final exams or missed an exam, some data were blank or missing in the grades column, so the questionnaires and grades of these two students were excluded from analysis moreover, the initial data collected by the questionnaire

were not completely consistent, so discretization was used. Therefore, for the purpose of constructing the decision tree model, the data were processed as follows.

(1) Number of class attendances: perfect attendance = high, absence [0–5] = medium, absence [5 and above] = low

(2) Daily online time (h): [0–5] = low, [5–10] = medium, [10 and above] = high

(3) Frequency of library study: often = high, rarely = medium, never = low

(4) Completion rate of homework assignments: [95%–100%] = high, [80%–95%] = medium, [80% and below] = low

A summary of the cleaned data is presented in Table 2.

4. EXPERIMENTAL ANALYSIS

4.1 Experimental Design

By means of data mining, this study aimed to determine those daily behaviors of students that affect their college grades. A decision tree model was constructed, with the final grades of college students as the data mining subject. The main daily behaviors studied were: class attendance, daily online time, frequency of library study, pre/review of course material before and after class, and completion rate of homework assignments. The information gain values of these five daily behaviors were calculated, and the root node of the decision tree was obtained by comparing these values. Finally, the decision tree was built. After that, the number of students who were correctly classified by the decision tree was compared with the number of students who were correctly classified by the support vector machine (Jalal and Jalal, 2020) and random forest (Nurwulan and Selamaj, 2020) to confirm whether the decision tree model is valid and applicable.

4.2 Process and Results Analysis

First of all, a decision tree model was established. After the pre-processing of the grade data set, the test grades were divided into four categories according to the GPA, namely excellent, good, qualified, and failed, i.e., the output results. Among them, there were 37 students with excellent results, 72 with good results, 26 with qualified results, and 10 with failed results.

According to the calculation formula of entropy, the following is obtained:

$$H(A) = -\left(\frac{37}{145}\log_2\frac{37}{145} + \frac{72}{145}\log_2\frac{72}{145} + \frac{26}{145}\log_2\frac{26}{145} + \frac{10}{145}\log_2\frac{10}{145}\right) = 1.7150.$$

Next, the information gain was calculated. Taking the attribute of "number of class attendances" as an example, its feature values were A_1 (number of class attendances = high), A_2 (number of class attendances = medium), and A_3 (number of class attendances = low). Taking the above three feature values as the root nodes, 108 students had high feature values, among which 35 were excellent, 50 were good, 19 were qualified, and 5 were failed; the number of students with medium feature values was 32, of which 2 were excellent, 22 were good, 4 were qualified, and 3 were failed; the number of students with low feature values were 5, of which 0 were excellent, 0 were good, 3 were qualified, and 2 were failed.

The corresponding entropy values were calculated:

$$H(A_1) = -\left(\frac{35}{108}\log_2\frac{35}{108} + \frac{50}{108}\log_2\frac{50}{108} + \frac{19}{108}\log_2\frac{19}{108} + \frac{5}{108}\log_2\frac{5}{108}\right) = 1.6875,$$

$$H(A_2) = -\left(\frac{2}{32}\log_2\frac{2}{32} + \frac{22}{32}\log_2\frac{22}{32} + \frac{4}{32}\log_2\frac{4}{32} + \frac{3}{32}\log_2\frac{3}{32}\right) = 1.2466,$$

$$H(A_3) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.9710.$$

The corresponding information gain is calculated:

$$g(A,\ B) = H(A) - \frac{108}{145}H(A_1) - \frac{32}{145}H(A_2) - \frac{5}{145}H(A_3) = 0.1495.$$

Thus, the information gain value of the number of class attendances was 0.1495. Similarly, the information gain values of the other four factors -daily online time, frequency of library study, pre/review before and after class, and completion rate of homework assignments- were also calculated, and the results were 0.1264, 0.1198, 0.0963, and 0.0934, respectively.

The information gain value of the five factors was compared. It was found that the number of class attendances had the largest information gain value, and therefore it was used as the root node of the decision tree. The data set was divided into four subsets, i.e., excellent, good, qualified, and failed, according to the GPA, and the calculation was continued to obtain the decision tree model of student performance distribution shown in Figure 2.

The IF-THEN classification rules were used to analyze and summarize the decision tree (see Figure 2), and the following rules were obtained.

(1) IF the number of class attendances = high + daily online time = low + frequency of library study = high THEN excellent performance

(2) IF the number of class attendances = high + daily online time = low + frequency of library study = medium + pre/review before and after class = yes THEN good performance

(3) IF the number of class attendances = high + daily online time = low + frequency of library study = medium + pre/review before and after class = no THEN qualified performance

Table 1 Statistics for students' daily behaviors.

	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	...	No. 147
Class attendance	Perfect attendance	Perfect attendance	Truant once	Perfect attendance	Ask for leave five times	Late twice	...	Perfect attendance
Daily online time	5 h	6 h	8 h	7 h	10 h	6 h	...	9 h
Frequency of library study	Often	Rarely	Never	Rarely	Rarely	Often	...	Rarely
Pre/review before and after class	Yes	No	No	Yes	No	Yes	...	No
Completion rate of homework assignments	99%	100%	85%	100%	74%	95%	...	96%

Table 2 Summary of students' daily behaviors.

	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	...	No. 147
Class attendance	High	High	Medium	High	Low	Medium	...	High
Daily online time	Low	Medium	Medium	Medium	High	Medium	...	Medium
Frequency of library study	High	Medium	Low	Low	Low	High	...	Medium
Pre/review before and after class	Yes	No	No	Yes	No	Yes	...	No
Completion rate of after-class assignments	High	High	Medium	High	Low	High	...	High
Grade obtained	Excellent	Good	Pass	Pass	Failed	Excellent		Good

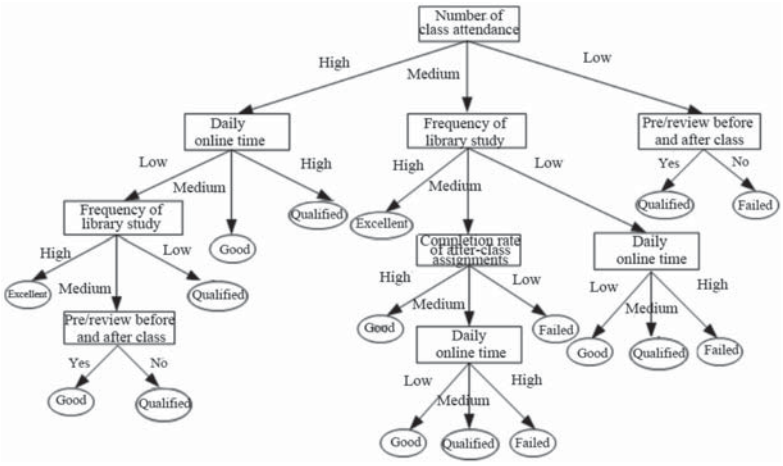

Figure 2 The decision tree model of student performance distribution.

(4) IF the number of class attendances = high + daily online time = low + frequency of library study = low THEN qualified performance

(5) IF the number of class attendances = high + daily online time = medium THEN good performance

(6) IF the number of class attendances = high + daily online time = high THEN qualified performance

(7) IF the number of class attendances = medium + frequency of library study = high THEN excellent performance

(8) IF the number of class attendances = medium + frequency of library study = medium + completion rate of after-class assignments = high THEN good performance

(9) IF the number of class attendances = medium + frequency of library study = medium + completion rate of after-class assignments = medium + daily online time = low THEN good performance

(10) IF the number of class attendances = medium + frequency of library study = medium + completion rate of after-class assignments = medium + daily online time = medium THEN qualified performance

(11) IF the number of class attendances = medium + frequency of library study = medium + completion rate

Table 3 Comparison of test results.

	Decision tree	Support vector machine	Random forest
Number of students correctly classified (n)	119	107	103
Correct rate (%)	82.07%	73.79%	71.03%

of after-class assignments = medium + daily online time = high THEN failed performance

(12) IF the number of class attendances = medium + frequency of library study = medium + completion rate of after-class assignments = low THEN failed performance

(13) IF the number of class attendances = medium + frequency of library study = low + daily online time = low THEN good performance

(14) IF the number of class attendances = medium + frequency of library study = low + daily online time = medium THEN qualified performance

(15) IF the number of class attendances = medium + frequency of library study = low + daily online time = high THEN failed performance

(16) IF the number of class attendances = low + pre/review before and after class = yes THEN passed performance

(17) IF the number of class attendances = low + pre/review before and after class = no THEN failed performance

A summary of the rules generated above leads to the following conclusions.

(1) Students with high number of class attendances did not have failed performance at the end of the semester; students with low number of class attendances did not have an overall performance above qualified at the end of the semester.

(2) Students with long daily online time did not receive overall performance above qualified at the end of the semester.

(3) Students with high frequency of library study received excellent overall performance at the end of the semester.

(4) In the case of a low number of class attendances, pre/review of course material before and after class was the factor that directly determines a student's overall performance at the end of the semester.

Therefore, it was found that the number of class attendances had the greatest impact on the academic performance of college students. A low number of class attendances not only means that students miss the information given by the teacher during each lesson, and which is vital for final exams; attendance can also affect their academic performance throughout the semester. The total final scores of college students are made up of usual performance and the final exam results. If a student wants to achieve excellent academic results, then the frequency of library study is very important. The college library is the place where students can access important resources. Visiting the library and acquiring information from academic books will enhance students' mastery of professional knowledge and expand their horizons. Secondly, through the decision tree model, it was found that the time spent on the Internet every day was also very important. Teachers in colleges no longer strictly manage students, so students have more free time to play with their cell phones. However, if students increase the amount of time they spend on the Internet, this will decrease the amount of time devoted to study, resulting in poor academic performance. Therefore, it is recommended that college students should control the amount of time they spend on the Internet and, instead, participate in competitions related to their majors and club activities to enrich their after-school life and improve their academic performance, or they should spend their time on acquiring knowledge related to their courses instead of surfing online. Moreover, the pre/review of course material before and after class should not be neglected. Pre-review before class can help students to have an understanding of the knowledge points before class; review after class can help students to consolidate the information given by the teacher in class, which improves their understanding and memory.

As shown in Table 3, of the 145 students, 119 in the decision tree classification had the same grade level as reality, and the correct rate was 82.07%; the correct rates of both the support vector machine and random forest approaches were below 80%. This indicated that the classification rules obtained from the decision tree model were relatively accurate and this model could be used to analyze the effect of daily behaviors on college students' performance.

5. CONCLUSION

This paper examined the daily behaviors of college students and their impact on academic performance. Based on data mining, the decision tree algorithm was used to build a decision tree model to analyze the relationship between students' academic performance and their daily behaviors such as class attendance, daily online time, frequency of library study, pre/review of course material before and after class, the completion rate of homework assignments. It was found that the decision tree model could analyze the daily behaviors affecting college students' academic performance. It was that class attendance had the greatest influence on the students' academic performance, because it affected not only their final exam results, but also their usual performance. Secondly, the time that students spent daily on the Internet also had an influence on academic performance. It was also found that decision tree model classification achieved a higher accuracy rate than the support vector machine and random forest. This study demonstrates that the decision tree

model in data mining technology can effectively analyze those daily behaviors of college students that affect their academic performance.

REFERENCES

1. Ajai, J., & Shiaki, B.O. (2020). Study Habits and Academic Achievement: A Case Study of Secondary School Science Students in the Jalingo Metropolis, Taraba State, Nigeria. *American Journal of Educational Research*, *8*(5), 282–285.
2. Arellano, A.R., Bory-Reyes, J., & Hernandez-Simon, L.M. (2018). Statistical Entropy Measures in C4.5 Trees. *International Journal of Data Warehousing and Mining (IJDWM)*, *14*(1), 1–14.
3. Bokadiya, L., Kumar, A., & Chauhan, S. (2018). Data Mining Using Neural Network. *IJERT-International Journal of Engineering Research & Technology*, (10).
4. Chen, H.M., Nguyen, B.A., Yan, Y.X., & Dow, C.R. (2020). Analysis of Learning Behavior in an Automated Programming Assessment Environment: A Code Quality Perspective. *IEEE Access*, *8*, 167341–167354.
5. Jalal, M., & Jalal, H. (2020). Behavior assessment, regression analysis and support vector machine (SVM) modeling of waste tire rubberized concrete. *Journal of Cleaner Production*, *273*, 1–15.
6. Lin, S. (2020). Data mining artificial intelligence technology for college English test framework and performance analysis system. *Journal of Intelligent and Fuzzy Systems*, *40*(2), 1–11.
7. Nurwulan, N.R., & Selamaj, G. (2020). Random Forest for Human Daily Activity Recognition. *Journal of Physics: Conference Series*, *1655*, 1–6.
8. Pandiangan, N., Buono, M., & Loppies, S. (2020). Implementation of Decision Tree and Nave Bayes Classification Method for Predicting Study Period. *Journal of Physics: Conference Series*, *1569*, 1–6.
9. Saini, M., & Sood, S. (2019). Learning Analytics: A Comprehensive Analysis of Methods for Student Performance Prediction. *International Journal of Innovative Technology and Exploring Engineering*, *9*(2), 1509–1514.
10. Suryakanthi, T. (2020). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm. *International Journal of Advanced Computer Science and Applications*, *11*(2).
11. Tang, C. (2022). A study of the Influence of Various Characteristic Factors on the Employment Choice of College Graduates Using Data Mining. *Engineering Intelligent Systems*, *30*(6), 417–422.
12. Tsai, C.F., & Chen, Y.C. (2019). The optimal combination of feature selection and data discretization: An empirical study. *Information Sciences*, *505*, 282–293.
13. Wang, S.V., Maro, J.C., Baro, E., Izem, R., Dashevsky, I., Rogers, J., Nguyen, M., Gagne, J.J., Patorno, E., Huybrechts, K.F., Major, J.M., Zhou, E., Reidy, M., Cosgrove, A., Schneeweiss, S., & Kulldorff, M. (2018). Data Mining for Adverse Drug Events with a Propensity Score-matched Tree-based Scan Statistic. *Epidemiology*, *29*(6), 895–903.
14. Xiao, J., Wang, L., Zhao, J., & Fu, A. (2020). Research on Adaptive Learning Prediction Based on XAPI. *International Journal of Information and Education Technology*, *10*(9), 679–684.
15. Zhou, K. (2022). Research on Online Creativity Education for College Students Using Data Mining. *International Journal of Engineering Intelligent Systems*, *30*(6), 447–452.