Eng Int Syst (2025) 5: 591–595


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

# Student Management and Career Guidance in Schools Through Data Analysis

**Hongshuo Chen**[1,*] **and Xueli Zhang**[2]

[1] *Student Affairs Office, North China University of Science and Technology, Tangshan, Hebei 063210, China*
[2] *Jitang College, North China University of Science and Technology, Tangshan, Hebei 063210, China*

This paper analyzed the learning status of sophomore students at the North China University of Science and Technology using the random forest (RF) algorithm, and the employment direction of graduates from the same university using the association rule algorithm. The results indicated that the random forest (RF) algorithm performed best when the number of decision trees was set to 70. Among the factors influencing learning status, the factor with the highest relative importance was "final exam score", followed by "midterm exam score" and "online learning time". The association rule algorithm was effective in mining the rules that impact the employment direction of graduates.

Keywords: random forest, association rule, learning status, career guidance

## 1. INTRODUCTION

Due to the more sophisticated development of education informatization, schools have gathered vast data on students' lives, behaviors, and learning [1,2]. This data contains valuable information that, if effectively mined and analyzed, can significantly impact student management and career guidance [3,4]. In the information age, schools can record students' daily learning and living behaviors, and these records can reveal their behavioral patterns [5]. Additionally, data on students' academic performance, completion of daily coursework, and online learning activities can be used to construct a model for evaluating the effectiveness of students' learning. By utilizing this model, schools can accurately assess students' learning outcomes and identify areas of weakness to provide personalized teaching recommendations. Moreover, choosing an employment direction has always been a significant concern for universities and students [6]. A well-chosen employment direction can help graduates integrate into society quickly and contribute to the workforce, while a high employment rate can reflect universities' teaching quality, and strengthen their appeal. Chen [7] proposed a text clustering method based on a convolutional neural network (CNN) to develop a Chinese teaching data mining and analysis system, and optimized it for more comprehensive and in-depth mining of Chinese character data. Agus [8] applied the ID3 algorithm to analyze students and determine the factors influencing academic performance to improve students' learning outcomes. Similarly, Josephng [9] utilized an instructional data mining strategy to analyze factors influencing student performance and support effective teaching practices. This paper analyzed the learning status of sophomore students at the North China University of Science and Technology using the random forest (RF) algorithm and the employment direction of graduates from the same university using the association rule algorithm.

*Corresponding address: No. 21, Bohai Road, Caofeidian Xincheng, Tangshan, Hebei 063210, China. Email: chs_chen@outlook.com

# 2. DATA MINING TECHNOLOGY

## 2.1 Random Forest Algorithm

The management of students in schools primarily focuses on improving their learning outcomes. However, each student possesses unique learning abilities, and a standardized management approach, while efficient, may not provide personalized teaching for students [10]. Analyzing students' academic performance, completion of daily homework, and other learning activities can help schools understand their students' learning status.

This paper employs the RF algorithm, a data mining technology, to classify students' learning status. The fundamental principle of building a classification model using the RF algorithm involves utilizing bootstrap sampling to create multiple decision trees, and the final classification result is determined by a majority vote among the trees [11]. In addition to classifying students' learning status, the RF classification model can also evaluate the impact of input features, i.e., the factors used to determine learning status. The construction process of the RF algorithm is shown below.

① In this paper, the data related to students' academic performance may not be comprehensive, although it can reflect students' learning status. Therefore, the K-means clustering algorithm [12] initially clusters the dataset. Subsequently, based on the data distribution within the clusters, the label of 'good' or 'bad' learning status is assigned to each cluster, serving as the classification labels of samples for training the RF algorithm.

② The training sample set is sampled using replacements to create multiple sub-training sets, the number of which depends on the number of decision trees in the RF algorithm. Simultaneously, the feature set of the training sample set is also sampled with replacements to generate feature subsets with the same number as the sub-training sets. These feature subsets are randomly assigned to the sub-training sets one by one [13].

③ A decision tree is built for each sub-training set. The required features are selected from the corresponding feature subset. The decision tree is constructed using the features in the feature subset as branch nodes. During feature selection, the Gini index is calculated when a feature serves as the branch node, and the feature with the smallest Gini index [14] is taken as the branch node. The decision tree branch grows. The feature is continuously to be selected based on the Gini index as the subsequent branch node until no further division is possible. After the growth of the decision tree, it is pruned to prevent overfitting. The pruning operation [15] converts part of the branch nodes in the decision tree to leaf nodes starting from the leaf nodes to the root nodes.

④ The decision trees from each sub-training set are combined to create an RF classification model. During model utilization, input samples are classified by each decision tree in the model, and the final classification result is based on the majority output of all decision trees.

## 2.2 Association Rule Algorithm

The RF algorithm can predict students' learning status in order to understand their learning effectiveness and provide guidance to targeted teaching interventions for students with poor learning outcomes. Besides using the RF algorithm for student management, career guidance is also crucial for colleges and universities. Teachers need to provide career guidance for students based on their performance in school. By analyzing past graduates' academic performance and employment outcomes, and extracting patterns, valuable employment recommendations can be offered to current students. The association rule algorithm can mine hidden patterns within student data samples [16]. The process of mining rules using the association rule algorithm is as follows.

① The sample dataset comprising data from past graduates is scanned. The performance of students in various aspects such as academic performance and employment direction can be considered as an item. For example, academic performance includes features such as students' majors and learning outcomes in majors. The manifestation of the feature "student's major" consists of various items associated with the major. The manifestation of the feature "performance in majors" is evaluated based on the students' academic performance. Employment direction can be manifested as pursuing postgraduate studies, taking civil service examinations, or applying for positions in state-owned enterprises.

② Based on the scanning results of feature manifestation items within the dataset, each feature manifestation item is denoted as a candidate item set. Then, the support of each candidate item set is calculated.

③ The candidate itemsets are pruned to derive frequent itemsets.

④ A new candidate set is constructed using the frequent itemsets, and step

② is repeated.

⑤ Return to step

③ for further pruning until no additional candidate itemsets can be generated from frequent itemsets [17].

⑥ An association itemset is constructed for each frequent itemset. The elements within an association itemset represent association rules, which denote relationships between non-void proper subsets of the itemset and the corresponding set of remaining elements.

# 3. CASE STUDY

## 3.1 Subjects

This paper utilized data mining technology to support student learning management and offer career guidance. For student learning management, the RF algorithm was applied to

$$MDA(j) = \frac{\sum_t^T \left( \frac{1}{|D_t|} \left( \sum_{X_i \subset D_t} I(P(X_i) = y_i) - \sum_{X_i^j \subset D_i^j} I(P(X_i^j) = y_i) \right) \right)}{T}, \quad (1)$$

**Table 1** Sample of sophomores' relevant data.

| Student number | 202122****23 | 202123****12 | 202132 ****25 | …… |
|---|---|---|---|---|
| Name | Li** | Zhou** | Wu** | …… |
| Gender | Female | Female | Male | …… |
| Class attendance | Average | Average | Poor | …… |
| Classroom performance | Average | Excellent | Average | …… |
| Online learning time | Poor | Excellent | Average | …… |
| Homework after school | Average | Excellent | Average | …… |
| Mid-term exam result | Average | Excellent | Poor | …… |
| Final-term exam result | Average | Average | Poor | …… |

**Table 2** Sample of graduates' relevant data.

| Name | Wang*** | He** | Ma*** | … |
|---|---|---|---|---|
| Gender | Male | Female | Female | … |
| Major | Food engineering | Food safety | Civil engineering | … |
| Performance in major | Excellent | Moderate | Poor | … |
| English level | Moderate | Moderate | Poor | … |
| Awards | None | Have | Have | … |
| Status of participation in associations | Non-participation | Active participation | Little participation | … |
| Status of participation in competitions | Active participation | Non-participation | Little participation | … |
| Penalties for violations | None | Few | Many | … |
| Employment direction | Taking a civil service entrance exam | Applying for positions in state-owned enterprises | Waiting for employment | … |

assess students' learning status. Furthermore, the models constructed by the algorithm was utilized to analyze the importance of factors influencing learning status. The association rule algorithm was applied for career guidance to predict students' employment directions. Given that these two data mining techniques were applied in different contexts, the sample sets targeted for mining also differed. In the case of the RF algorithm, sophomore students at the North China University of Science and Technology were selected as the subjects (Table 1). On the other hand, graduates of the North China University of Technology were the focus (Table 2) of the association rule algorithm. Due to space limitations, the tables present only some major types and employment directions of the graduates. Finally, 9,752 items of relevant data were collected from the sophomore students, and 9,785items were collected from the graduates.

## 3.2 Analysis Method

Before applying the RF algorithm to analyze the performance of sophomore students. Because the midterm and final grades give a one-sided reflection of the student's learning status, the K-means clustering algorithm was utilized to segment the samples initially. The samples were divided into two cluster classes; i.e., the k value of the clustering algorithm was set to 2.

When the RF algorithm constructs a classification model, the sample set and the feature set of samples are sampled with a replacement. This paper set the number of random sub-features to 3, and the number of decision trees within the algorithm was set to 10, 30, 50, 70, and 90, respectively. The RF algorithm was evaluated using the evaluation formula under different parameters:

Above Equation 1

where $T$ is the number of decision trees in the RF, $D_t$ is the set of out-of-bag samples of decision tree $t$, $D_i^j$ is the set of samples after $D_t$ exchanges the $j$-th dimensional feature, $X_i$ is the $i$-th sample, $X_i^j$ is the sample after $X_i$ exchanges the $j$-th dimensional feature, $P(\ )$ is the prediction function of the decision tree, and $I(\ )$ is an indicative function, which is 1 when the prediction of the decision tree is correct and 0 when the opposite is true.

The association rule algorithm was applied to extract association rules from the sample data about graduates' employment directions. The support threshold was set to 0.2, while the confidence threshold was set to 0.5 for the mining process.

## 3.3 Analysis Results

The clustering results from the K-means clustering algorithm are presented in Table 3. The clustering algorithm divided the sample set of sophomore students into two clusters. A comparison of the two clusters was conducted using an independent samples t-test. A P value below 0.05 indicated a significant difference. As observed in the table, the category

**Table 3** Clustering results for sophomore students.

| Feature indicator | | Cluster 1 | Cluster 2 | P value |
|---|---|---|---|---|
| Number of people | | 4,858 | 4,894 | 0.697 |
| Gender | Male:female | 1:0.996 | 1:0.997 | 0.686 |
| Class attendance | Excellent:average:poor | 1:0.031:0.013 | 0.011:0.021:1 | 0.001 |
| Classroom performance | | 1:0.032:0.036 | 0.018:0.037:1 | 0.000 |
| Online learning time | | 1:0.038:0.022 | 0.041:0.037:1 | 0.001 |
| Homework after school | | 1:0.042:0.058 | 0.038:0.035:1 | 0.002 |
| Mid-term exam result | | 1:0.031:0.054 | 0.028:0.057:1 | 0.000 |
| Final-term exam result | | 1:0.057:0.011 | 0.031:0.034:1 | 0.001 |

**Table 4** Comparison of the RF algorithm performance.

| Number of decision trees | 10 | 30 | 50 | 70 | 90 |
|---|---|---|---|---|---|
| *P* | 0.635 | 0.732 | 0.811 | 0.976 | 0.975 |
| *R* | 0.633 | 0.729 | 0.808 | 0.975 | 0.976 |
| *F* | 0.636 | 0.720 | 0.809 | 0.976 | 0.974 |
| Calculation time/s | 1.34 | 1.93 | 2.42 | 5.95 | 10.21 |

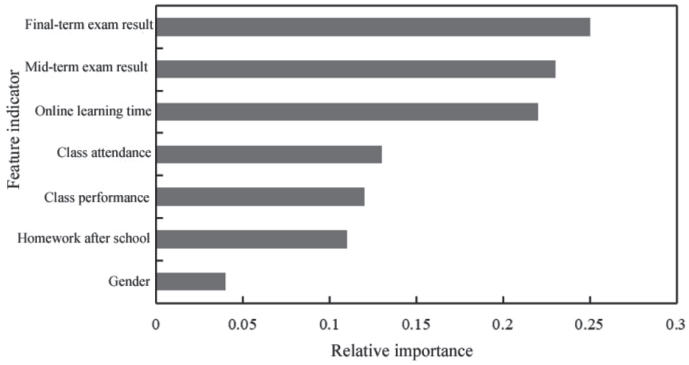


**Figure 1** Relative importance of feature indicators when analyzing the learning status.

label for Cluster 1 was a favorable learning status, while the label for Cluster 2 was a poor learning status.

Subsequently, the learning status labels of sophomore students were used to train and test the RF algorithm. The performance of the RF algorithm using varying numbers of decision trees is detailed in Table 4. It was seen that increasing the number of decision trees could enhance the algorithm's performance. However, after reaching a threshold, the increase in classification accuracy became insignificant while the computational time continued to increase.

Then, the RF algorithm with optimal performance was employed to assess the importance of the feature indicators. The outcomes are depicted in Figure 1. It can be seen that the "final-term exam result" indicator exhibited the highest relative importance in the students' learning status, followed by the "mid-term exam result" and "online learning time". The indicator with the most insignificant impact on learning status was gender.

The association rule algorithm was utilized for data mining on the employment directions of graduates. Due to space constraints, partial mining results are presented in Table 5.

## 4. CONCLUSION

This study analyzed the learning status of sophomore students at the North China University of Science and Technology using the random forest (RF) algorithm, and the association rule algorithm was applied to determine the employment direction of graduates from the same university. The K-means clustering algorithm was employed to ascertain the learning status of sophomore students. Cluster 1 indicated a superior learning status compared to Cluster 2. The RF algorithm achieved optimal performance with 70 decision trees. The analysis of the importance of feature indicators using the RF algorithm showed that the "final-term exam result" was the best indicator of students' learning status, followed by the "mid-term exam result" and "online learning time", with gender having the least influence. Those with poor academic performance were likely to find it difficult to secure a job immediately after graduation, and had to wait longer for employment opportunities. Graduates with excellent performance tended to undertake civil service examinations or postgraduate studies. Graduates with moderate academic performance were inclined towards foreign companies if they were proficient in English, while those with moderate English proficiency were more likely to choose state-owned enterprises.

## REFERENCES

1. Jahan N, Shahariar R. Predicting Fertilizer Treatment of Maize Using Decision Tree Algorithm[J]. Indonesian Journal of

**Table 5** Partial mining results of association rules for graduates' employment direction.

| Serial number of association rule | Concrete content | Confidence level |
|---|---|---|
| 1 | Poor performance in major, poor English proficiency, and many penalties for violations ⇒ wait for employment | 0.875 |
| 2 | Excellent performance in major, excellent English proficiency, active participation in associations, and no penalty for violations ⇒ take the civil service examination | 0.846 |
| 3 | Excellent performance in major, excellent English proficiency, active participation in competitions, and have awards ⇒ pursue postgraduate studies | 0.856 |
| 4 | Excellent performance in major, moderate English proficiency, no penalty for violations, active participation in associations ⇒ apply for positions in state-owned enterprises | 0.824 |
| 5 | Moderate performance in major, excellent English proficiency, no penalty for violations, and have awards⇒ choose foreign enterprises | 0.847 |

Electrical Engineering and Computer Science, 2020, 20(3): 1427–1434.
2. Li H, Yu X. Application and Evaluation of Intelligent Management Accounting Platform Based on Association Rule Algorithm and PS-DR-DP Model [J]. Engineering Intelligent Systems, 2025, 33(2):121–130.
3. Cheng Y F, Hsu Y S. Decision tree for investigating the factors affecting graduate salaries[J]. Journal of Research in Education Sciences, 2017, 62(2):125–151.
4. Fang X. Association Rule Mining for English Digital Archive System Based on Improved Apriori Algorithm [J]. Engineering Intelligent Systems, 2025, 33(2): 131–140.
5. Nguyen H L, Jung J E. Statistical approach for figurative sentiment analysis on Social Networking Services: a case study on Twitter[J]. Multimedia Tools & Applications, 2016, 76(6): 1–14.
6. Li R. A data mining-based approach to integrating multimedia English teaching resources[J]. International Journal of Computational Systems Engineering, 2024, 8(1/2):1–9.
7. Chen X. Optimization of Data Mining and Analysis System for Chinese Language Teaching Based on Convolutional Neural Network[J]. Computational Intelligence and Neuroscience, 2021, 2021:1148954.
8. Agus N, Lusi M, Deasy P. Implementation ID3 Algorithm to Predict Children Achievement in Response (Case Study Children Playgroup School)[J]. Journal of Engineering & Applied Sciences, 2017, 12(2):204–207.
9. Prayogo D, Susanto Y T T. Optimizing the prediction accuracy of load-settlement behavior of single pile using a self-learning data mining approach[J]. MATEC Web of Conferences, 2019, 258:1–6.
10. Sun J, Sun T. Research on the Effectiveness of KMV Model in China's Bond Credit Rating Market[J]. Journal of Financial Studies, 2020, 4(1):59–62.
11. Mihm B. Mispricing of Risk in Sovereign Bond Markets with Asymmetric Information[J]. German Economic Review, 2016, 17(4):491–511.
12. Robinson D. Actavis to test risk appetite with jumbo bond[J]. International Financing Review, 2015(2071):20–21.
13. Wijayanti E, Yuliana I. Risk Profile, Secure Bond, and Bond Rating in Banking Industry[J]. The Winners, 2020, 21(1):49–57.
14. Lakkakula NP, Naidu MM, Reddy KK. An entropy based elegant decision tree classifier to predict precipitation [C]. 2014 European Modelling Symposium, 2015: 11–19.
15. Prasad N, Naidu MM. Gain ratio as attribute selection measurein elegant decision tree to predict precipitation [C]. 2013 8th EUROSIM Congress on Modelling and Simulation, 2013: 141–150.
16. Cleghern Z, Lahiri S, Özaltin O, Roberts DL. Predicting future states in DotA 2 using value-split models of time series attribute data[C]// Proceedings of the 12th International Conference on the Foundations of Digital Games, 2017:1–10.
17. Malik A J, Khan F A. A hybrid technique using binary particle swarm optimization and decision tree pruning for network intrusion detection[J]. Cluster Computing, 2017, 21(3): 1–14.