Hindawi
Complexity
Volume 2021, Article ID 5577868, 11 pages
https://doi.org/10.1155/2021/5577868

Research Article

Analysis and Prediction of CET4 Scores Based on Data Mining Algorithm

Hongyan Wang

School of Foreign Languages, Xi'an University of Finance and Economics, Xi'an 710000, China

Correspondence should be addressed to Hongyan Wang; 2004010030@xaufe.edu.cn

Received 8 February 2021; Revised 2 March 2021; Accepted 10 March 2021; Published 20 March 2021

Academic Editor: Wei Wang

This paper presents the concept and algorithm of data mining and focuses on the linear regression algorithm. Based on the multiple linear regression algorithm, many factors affecting CET4 are analyzed. Ideas based on data mining, collecting history data and appropriate to transform, using statistical analysis techniques to the many factors influencing the CET-4 test were analyzed, and we have obtained the CET-4 test result and its influencing factors. It was found that the linear regression relationship between the degrees of fit was relatively high. We further improve the algorithm and establish a partition-weighted *K*-nearest neighbor algorithm. The K-weighted K nearest neighbor algorithm and the partition algorithm are used in the CET-4 test score classification prediction, and the statistical method is used to study the relevant factors that affect the CET-4 test score, and screen classification is performed to predict when the comparison verification will pass. The weight K of the input feature and the adjacent feature are weighted, although the allocation algorithm of the adjacent classification effect has not been significantly improved, but the stability classification is better than *K*-nearest neighbor algorithm, its classification efficiency is greatly improved, classification time is greatly reduced, and classification efficiency is increased by 119%. In order to detect potential risk graduating students earlier, this paper proposes an appropriate and timely early warning and preschool K-nearest neighbor algorithm classification model. Taking test scores or make-up exams and re-learning as input features, the classification model can effectively predict ordinary students who have not graduated.

1. Introduction

The value of data mining refers to whether the data studied has instructive significance. The process of data mining is to search for massive data, which is difficultly compared with the process of searching for small data sets. Among them, in the process of processing large amounts of data, there will be many unpredictable new problems, which will not exist in small data sets. Particularly in the process of data mining, the final result may be the representation of the data rather than the essence of the data, or it may be representative of meaningless in the random process.

With the accumulation of a large number of data in the database of the educational administration system of colleges and universities, how can teachers and educational administrators use these data more effectively? For example, what are the key factors affecting the passing of CET-4, what are the relevant factors related to the results of CET-4, how to more effectively predict the passing of CET-4, how to find out the students who could not graduate in time, and so on. The application of data mining technology can solve these problems more effectively. At the same time, it will also assist teachers to guide students to prepare for CET-4 and help students graduate.

CET 4 has become a selection condition for many employers. How to pass CET-4 smoothly? Many college students do not know how to pass CET-4, and many teachers do not know how to manage to help students pass CET-4. What factors are closely related to students' CET-4 scores? Therefore, the use of data mining technology to explore and research and solve these problems has become a part of the university that attaches great importance to the educational issues. This has certain theoretical significance and practical value for improving students' own quality and

competitiveness, enhancing employment quality, and promoting the development of some colleges and universities. By predicting whether students can pass CET-4, students can better and more effectively improve CET-4 scores and prepare for the exam. Through the academic performance of the students in the previous semesters, we can predict whether they can graduate or not, find out the important courses or key factors that affect the graduation of the students, strengthen the corresponding teaching management, and give appropriate learning suggestions or warnings to the students in time to help them graduate normally. The first chapter is the introduction, the second chapter is the introduction of related work, the third part is the analysis and prediction of CET4 performance based on multiple regression and *K*-nearest neighbor data mining, the fourth chapter is the result and analysis, and the fifth chapter is the conclusion.

2. Related Work

Data mining technology has been widely used in business, financial industry, enterprise production, marketing, and other aspects. With the maturity of data mining technology and the continuous expansion of its application fields, many university researchers have begun to study the application of data mining technology in the analysis of college students' performance. Based on the classification mining method of the K-nearest neighbor algorithm, we analyze the student performance database data, combine the SLIQ algorithm to analyze the student's performance, and establish a K-nearest neighbor algorithm model of professional ability for teachers and school education decision-makers to understand the existing problems in teaching, in order to use the performance information provided by the optimized teaching plan and decision-making [1, 2]. This paper studies the application of the principal component analysis method and the Bayesian *K*-nearest neighbor algorithm in data mining. The principal component analysis method is adopted in the comprehensive grade evaluation of graduate students. By removing the correlation between analysis factors, the analysis index is reduced while maintaining the information content, so as to reduce the information content [3]. Principal component analysis and Bayesian *K*-nearest neighbor classification method are used in the prediction of graduates' employment direction. Graduates' performance is used as the characteristic data. The principal component analysis is used to reduce the dimension of the characteristic data, and the Bayesian *K*-nearest neighbor algorithm is used to classify the career direction [4]. After the study and analysis of ID3 (Iterative Dichotomiser 3) algorithm, an improved algorithm is put forward to mine and analyze the data stored in the educational administration management system, so as to find out the gaps and gaps between curriculum settings and provide some data basis for the statistical decision-making in colleges and universities [5].

This paper introduces the classic Apriori algorithm and the famous Decision ID3 algorithm of association rules and uses the Apriori algorithm to mine the influence of excellence in a certain course on other courses: nirvana. The ID3 algorithm is used to generate *K*-nearest neighbor algorithm to analyze the factors related to students' excellent performance, and the postpruning method is used to prune the *K*-nearest neighbor algorithm. Finally, the *K*-nearest neighbor algorithm generates classification rules [6] to complete the establishment of the *K*-nearest neighbor algorithm for performance analysis. The improved Apriori algorithm is used to realize the application of association rules in the analysis of students' grades, and the clustering algorithm is used to further analyze the results. By introducing the information dimension of teachers, the teaching effects of teachers with different titles in different courses are analyzed [7]. By introducing the dimensions of class time and examination information, and comparing the quality of students' proficiency in the same subject under different class time arrangements, we can find out which kind of teacher titles is suitable to be assigned to teach in different majors and different examination requirements. On this basis, the introduction of time reflects the different time periods and different teaching levels of each teacher. Through internal improvement and title improvement, the teaching effect is enhanced; the age of introduced teachers can reflect which age group of teachers has better teaching effect [9]. We put forward the idea of multistrategy design, combined with data mining technology and statistical analysis, based on the classification of the K-nearest neighbor algorithm mining method, analyze the student scores in the library data, and generate the K-nearest neighbor algorithm. Student scores can directly display the calculation position of the results at different levels and provide evaluation information for the teaching department. At the same time, the statistical analysis method based on summary rules is adopted to complete the query, prediction, and comparative analysis of scores under different circumstances [10, 11] and realize the automatic generation of student score analysis report, test paper quality report, and quality analysis table. The improved ID3 algorithm is applied to analyze students' performance of different courses, to find out the potential factors affecting students' performance, so that students can maintain a good learning state, so as to provide decision-making support information for teaching departments, promote better teaching work, and improve teaching quality [12]. The rough set theory is applied to analyze the English performance of a class to find out the most important factors affecting students' overall performance, so as to provide a basis for foreign language teachers to improve teaching methods and methods and improve teaching quality [13]. In recent ten years, experts and scholars have realized the importance of using data mining technology in CET-4. At the same time, we should also be aware of the necessity, urgency, and responsibility of collecting and mining the information in CET-4. The *K*-nearest neighbor algorithm is used to analyze the results of CET-4. It is suggested to strengthen the teaching of College English (III) [14, 15] and strengthen the English teaching of single enrollment classes in higher vocational colleges to improve the learning of English scores in school [16]. Data mining is carried out with the ID3 algorithm. Passing CET-4 is related to English foundation, effort level, and other factors [17]. Data mining is conducted, respectively. It is suggested that attention should be paid to the study of English (I) and the passing of CET-4, which are closely related to gender and learning attitude [18]. Design college

student achievement management analysis system [19]; C5.0 algorithm is used to analyze CET-4 scores, and it is concluded that grade, gender, English learning foundation, learning interest, and attitude towards examination have a high influence on passing CET-4 [20]. The conclusion is given that the results of the college entrance examination determine students' CET-4 scores, and the relationship between attendance, teaching evaluation scores, and CET-4 scores is also analyzed [21]. In the same year, he emphasized the need to strengthen listening training [22]. Using data mining technology, he analyzed and studied the ID3 algorithm of CET-4 through the *K*-nearest neighbor algorithm that has the greatest influence with gender factors and found the dominant factors affecting students' performance in the performance analysis [23]. The decision method is also used to analyze the relevant factors affecting students' performance [24]. The improved Apriori algorithm of association rules is used to carry out correlation analysis on all courses of computer application technology major and quantify the correlation degree among the scores of various courses of this major, thus optimizing the curriculum setting [25]. Using the improved clustering and performance analysis of association rule mining method, the relevant analysis of the students' course performance during the school period is carried out. The results provide a reference for the design of teaching plans and information to improve the quality of students' learning [26]. Use the *K*-nearest neighbor algorithm classification method to construct the student performance analysis system and use this system for course grade analysis, so as to promote the education quality promotion [9]. Previous studies have the following shortcomings: first, the decision tree classifier is not perfect enough, and the data preprocessing needs to be completed manually by other database tools; second, the pruning operation of the decision tree is not controlled automatically by the program. Third, in the classification results, there are still relatively large errors in the classification rules, which may be due to the selection of attribute fields in the data set, and it did not consider many factors that affect student performance, and the content of student information survey is not comprehensive enough.

These studies all have the following shortcomings: first, the *K*-nearest neighbor algorithm classifier is not perfect enough, and other database tools are needed to complete the data preprocessing manually; the second is that the pruning operation of the *K*-nearest neighbor algorithm is not controlled automatically by the program. Third, in the classification results, there are still relatively large errors in the classification rules, which may be due to the selection of attribute fields in the data set.

3. Analysis and Prediction of CET4 Performance Based on Multiple Regression and *K*-Nearest Neighbor Data Mining

3.1. Analysis and Prediction Theory of CET-4 Achievement Based on Data Mining. Data association is a kind of important knowledge that can be found in the database. If there is some regularity between the values of two or more variables, it is called association. Association rules reflect dependencies or associations between one event and other events. If there is a correlation between two or more items, then the attribute value of one item can be analyzed and predicted based on other attribute values to find out the relationship between them and give a reasonable explanation. Data mining algorithms include heuristics and calculation functions to create data mining models from data. It analyzes user-supplied data and looks for specific types of patterns and trends, and the algorithm uses the results of this analysis to define the best parameters for creating a mining model, which are applied to the entire data set to extract viable patterns and detailed statistics. Most data mining algorithms use one or several objective functions and use several search methods (such as heuristic algorithm, maximum and minimum value method, gradient descent method, and network deduction method) to obtain a point or a small region in the data body or in the data space where distance relationship has been established. Data mining algorithms can be divided into teacher-oriented and unteacher-oriented, also known as supervised learning and unsupervised learning. In supervised learning, a teacher signal is first given, category markers and classification costs are provided for each input instinct in the training sample set, and the direction that can reduce the overall cost is sought. There is no explicit teacher in the unsupervised learning algorithm, so the system automatically forms clustering for input samples. The four-level achievement analysis and prediction frame diagram based on multiple linear regression algorithm are shown in Figure 1.

Linear regression is a kind of regression algorithm, in which data are modeled with a straight line. Bivariate regression treats one random variable X (called the response variable) as a linear function of another random variable X (called the predictor variable). That is,

$$X = \delta + \chi Y, \tag{1}$$

where the variance of X is a constant and δ, χ are the regression coefficient, respectively, representing the intercept and slope of the line on the Y-axis. These coefficients can be solved using the least square method to minimize the error between the actual data and the estimate of the line. Given the data points of S samples, the regression coefficient can be calculated by the following formula:

$$\chi = \frac{(x_1 - \overline{x})(y_1 - y)}{(x_1 - x) + (x_2 - x)}. \tag{2}$$

Multivariate regression is an extension of linear regression by designing multiple predictive variables. The corresponding variable Y can be a linear function of a multidimensional eigenvector. Multiple regression based on the two predictive variables is as follows:

$$X = \delta + \chi_1 Y + \chi_2 Y. \tag{3}$$

In fact, there are many factors that affect CET-4 scores. Based on years of work experience, four factors considered very important are selected in this data mining and then

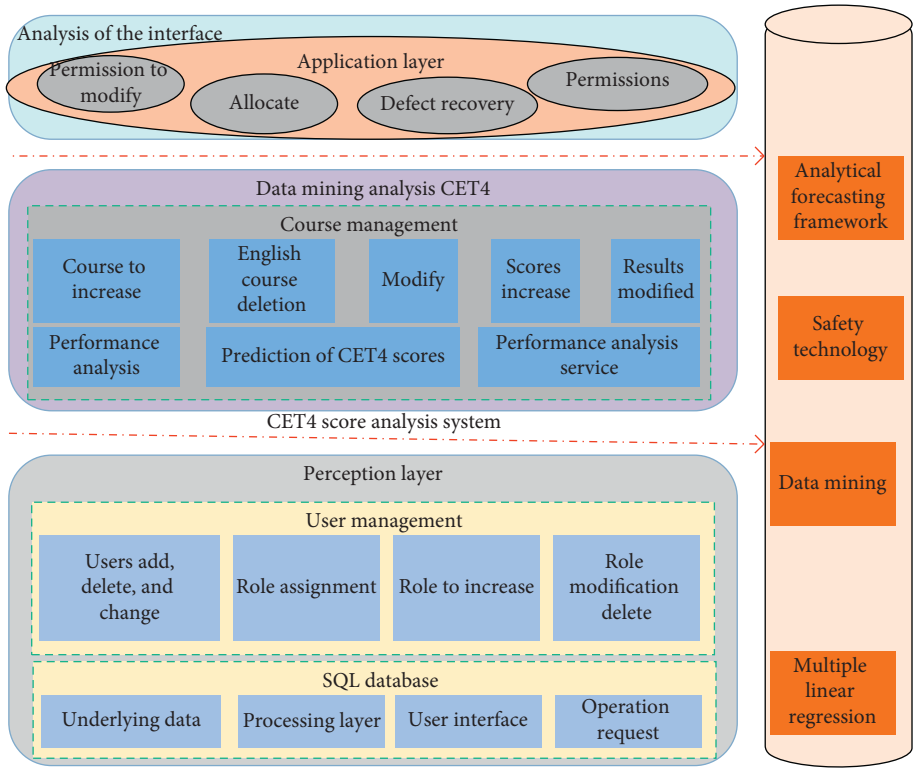

Figure 1: Block diagram of four-level achievement analysis and prediction based on multiple linear regression algorithm.

analyzed. The data of 75 non-English major students in a school are collected. The data table is as follows (Table 1).

As shown in Figure 2, some professional level 4 pass rate is very high, and some professional level 4 pass rate is very low. In the CET-4, the passing rate of two majors is more than 71%, the passing rate of one major is just over 51%, and the passing rate of the CET-4 of "Major 3" is not more than 36%. There are five majors with a failing rate of between 41% and 51%, and two majors with a passing rate of no more than 41%. It can be seen that whether students can pass CET-4 has a great deal to do with their major.

There are four College English scores of students, namely, College English (I), College English (II), College English (III), and College English (IV). However, the two College English scores closest to CET-4 are more closely related to CET-4 scores, which has been verified in the previous correlation coefficient figure. Therefore, this paper only discusses the relationship between the two recent College English scores from CET-4 and students' CET-4 results and presents them with bar Figures 3 and 4. "English score 2" refers to the score of the College English course that the student is studying when taking CET-4, while "English score 1" refers to the score of the College English course in the previous semester relative to "English score 2."

As can be seen from Figure 3, the higher the score of English 1 is, the more likely it is to pass CET-4. If the score of English 1 is over 75, the possibility of passing CET-4 will be over 60%, while if the score of English 1 is less than 50, the possibility of passing CET-4 will be less than 20%. In Figure 4, the overall trend of passing rate of CET-4 is basically the same as that in Figure 3, and the possibility of passing CET-4 increases with the increase of "English score 2." In addition, some students with a low score of "English score 2" also passed CET-4. After verification, these students all took the CET-4 in the fourth semester, but they did not take the CET-3 in the third semester. They only had the usual results instead of the paper results, so the "English score 2" was very low, but they could pass the CET-4. If this situation is removed, it can be said that the trend in Figure 4 is the same as that in Figure 4. This content can be verified by the correlation coefficients of "English score 1," "English score 2," "and the CET-4 results," and their correlation coefficients are 0.58 and 0.62, respectively.

3.2. A Prediction Model of Grade 4 Based on Improved K-Nearest Neighbor Algorithm. Because of the noise and outliers in the data, many points of the initial generated *K*-nearest neighbor algorithm reflect more anomalies in the training data. The pruning method uses statistical metrics to cut out the least reliable branches, which results in faster classification and improves the ability of the *K*-nearest neighbor algorithm to correctly classify foreign data. Figure 5 describes the operation flow of the *K*-nearest neighbor algorithm.

3.2.1. Step 1: Determine the Object and Target. Understanding the purpose of our data mining is the primary process in the process of data mining. Only when the

Table 1: Student data table.

Serial number	English score in college entrance examination	Sex	Learning attitude	Simulation results	Grade 4 (Y)
FX001	108	Male	Seriously	88	559
FX002	102	Female	General	77	483
FX003	105	Female	Poor	61	385

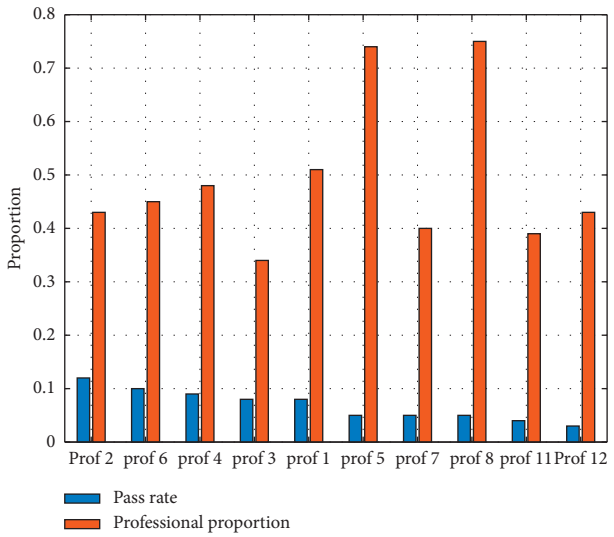

Figure 2: The relationship between CET 4 and major.

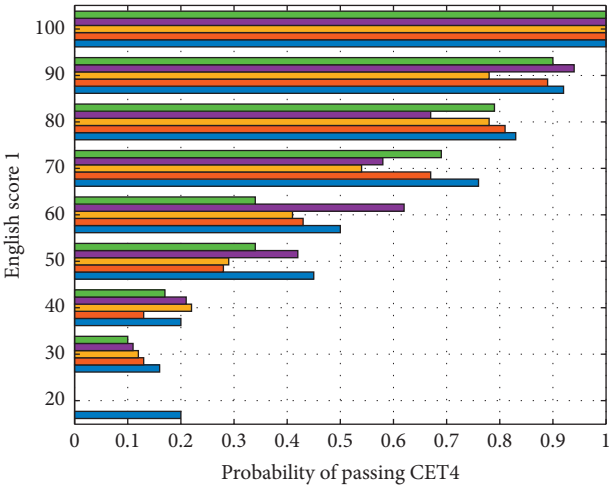

Figure 3: The relationship between CET-4 and College English 1.

purpose of data mining is established, data mining will not be blind and correct conclusions can be reached.

Data mining to student achievement data warehouse is designed in the previous chapter as analysis object, according to the university computer professional students over the course of information and each student achievement information, mining analysis of the various courses, the connection between the mining influence degree between the curriculum and curriculum, and the influence of the order to student achievement of the course, sit is hoped that the final analysis result can help the school's teaching plan and guidance.

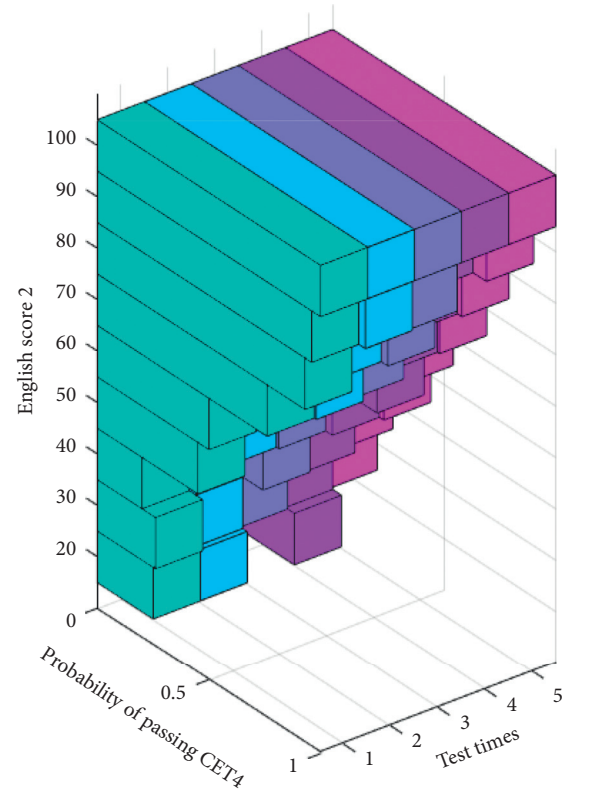

Figure 4: The relationship between CET-4 and College English 2.

3.2.2. Step 2: Select the Model. This step is the key to how to analyze the expansion after the whole mining process. The choice of algorithm will directly affect the quality of the subsequent mining analysis results.

Through the comprehensive analysis above, the Apriori algorithm will be adopted in this data mining. In order to obtain the interconnection between students and courses, the following two basic concepts should be paid attention to during the use of this model algorithm.

At the same time, the main idea of the algorithm is to find and analyze the frequent itemsets in the data that meet the set minimum support and then generate strong association rules that meet the preset minimum support and confidence from the above frequent itemsets.

3.2.3. Step 3: Data Acquisition. Data collection is a process of heavy workload and time-consuming. In this process, workers need to use various data extraction and collection methods to obtain the data needed for mining and analysis.

3.2.4. Step 4: Data Preprocessing. This process is the most workload in the whole data mining process, and it is also a

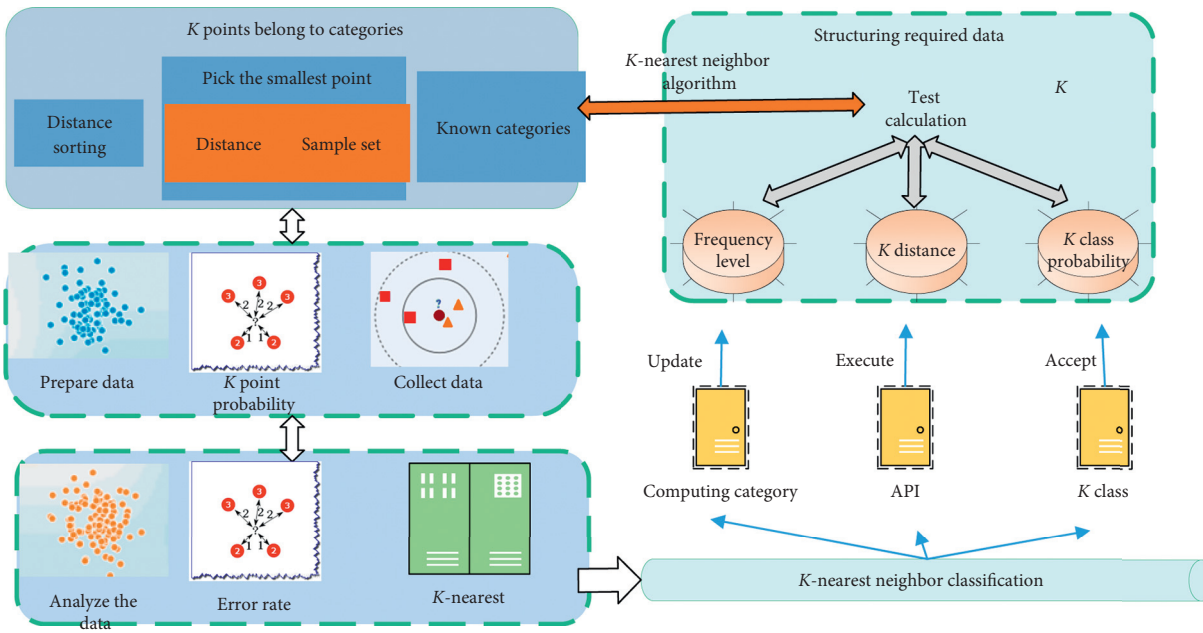

Figure 5: Operation flowchart of *K*-nearest neighbor algorithm.

very key link. The detailed and accurate completion of this process is the fundamental guarantee for the accurate and efficient acquisition of the final mining results.

3.2.5. Step 5: Data Mining. According to the type of data function and the characteristics of the data, the corresponding algorithm is selected for mining. Here, the classic association rule mining algorithm, Apriori algorithm, is selected, and then the preprocessed data are mined.

After completing the above process, the next step is to use the association rules algorithm to carry out data mining on students' course results. This paper takes part of the data of computer major of our university as the basic data, carries on the data mining processing, and does the research and analysis to the mining results obtained.

3.2.6. Step 6: Result Analysis. Result analysis is to make further analysis and research on the information obtained from data mining and interpret the mining results as theoretical results that are easy to be understood by everyone. By analyzing the mining results, we select and extract some strong association rules and then analyze and study the above strong association rules.

3.2.7. Step 7: Apply Your Knowledge. Knowledge application is the last step in the process of mining and analysis and also the step to realize the true meaning of database knowledge discovery. The knowledge obtained from mining processing is applied to solve real problems. This excavation analysis, through the analysis and research of the students in the course of the value of the law and information, is integrated into the teaching guidance of colleges and universities, to provide a scientific guarantee and important basis for the formulation of teaching plans.

Analysis nodes can be used to analyze the model used in *K*-nearest neighbor CET4 performance analysis and prediction. The algorithm flow is shown in Table 2.

4. Results and Analysis

When the *K*-nearest neighbor algorithm is used for prediction classification, it is necessary to determine the appropriate nearest neighbor number *K* and select important features. In order to achieve higher classification accuracy and more stable classification performance, this paper selects *K* and important features in an iterative way. Firstly, all the features are taken as input variables to determine the nearest neighbors of several numbers *K*. Then, under the condition of constant *K* value, the features are successively reduced, and the important features are selected according to the criteria of prediction effect. Finally, the appropriate *K* value is determined by taking the important features as input variables.

According to the correlation between students' "Entrance English score," "English score 1," "English score 2," and CET-4 scores, they are relatively important features and should be input variables for predictive classification. Considering the classification effect based on the gender of students, the college, and major of students, the importance of features is investigated from the perspective of the accuracy rate of classification effect. The selection idea is to remove the features one by one. If the effect of classification prediction is significantly reduced after removing a feature, it indicates that the feature is relatively important and should be retained. If the classification effect is not significantly reduced or the classification effect is better after the removal of the feature, it indicates that the feature is not important or even has a negative effect on the classification effect and should be removed.

Table 2: *K*-nearest neighbor CET4 performance analysis and prediction algorithm flow.

Algorithm flow: *K*-nearest neighbor CET4 performance analysis and prediction
Step 1. Coincidence matrix. A table displays rows defined by actual values and columns defined by predicted values, as well as the number of records in each cell that conforms to the schema. If more than one field related to the same output segment is generated, but these fields are generated by different models, the sum of the cases where these fields are the same but different is counted and displayed.
Step 2. Performance evaluation. This statistic is a measure used to predict the average information content of bits in the model for records belonging to that category. Considering the different difficulty of different categories in the classification problem, the accurate prediction of rare categories will get a higher performance evaluation than that of common categories. If the model does not perform random guesses for a category, the category's performance evaluation index will be 0.
Step 3. Number of letters. For models that generate confidence fields, this option reports statistics about the confidence values and their relationship to the predictions. This option has two settings: one is a threshold. The accuracy of the report meets a specified percentage of confidence. The second is to improve accuracy. Report the confidence level of accuracy improved by the specified coefficient.
Step 4. Divide by partition. If you use the partitioning field to split the record into a training example 8, a test example, and a validation example, selecting this option will display the results for each partition separately.
Step 5. User-defined analysis. The CLEM expression is used to specify what should be evaluated for each record in order to combine the score values at the record level into an overall score value.

There are 2674 pieces of data after data cleaning and processing, and all the CET-4 results contained in the data have been marked. If all the data are used to train the model and use it to predict unlabeled data, there is no way to evaluate the prediction effectiveness of the algorithm. To solve this problem, this paper divides the data into a training set and a test set according to 7 : 3. One part is used to train the nearest neighbor number and features, that is, to train the *K*-nearest neighbor model, and the other part is used to evaluate the prediction classification effect of the *K*-nearest neighbor model. To avoid a tie vote, let K take an odd number, 18 from 1 to 35. Table 3 lists eight cases of input features. The first case is the case with the most input features, and the other cases are the cases with some features removed.

As shown in Figure 6, the prediction accuracy is not the highest. "Situation 4" has only three features, namely, "Enrollment English score", "English score 1," and "English score 2." Although the input features of "Situation 4" are the least, as shown in Figure 6, when the abscissa is 11, the average prediction accuracy is the highest; when $K = 15$, the classification accuracy of "Case 8" basically reaches the maximum and the fluctuation is relatively small, which indicates that the accuracy and stability of the algorithm classification are reliable, when $K = 15$. Therefore, the nearest neighbor number can be determined as $K = 15$. In this way, the nearest neighbor number $K = 15$ is selected at the end of this paper, and the *K*-nearest neighbor classification algorithm with input features of "English score for admission," "English score 1," and "English score 2" is selected. The test will be predicted on the test set of the algorithm, and the test results are shown in Table 4.

When the *K*-nearest neighbor algorithm classifies the prediction, the default *K*-nearest neighbors make the same contribution to the prediction result. However, in general, the closer the sample to be classified is to the known classification sample, the more the characteristics of the known classification sample will be. Therefore, the contribution of the nearest neighbor to the prediction should be greater. The core idea of weighting is to define the weight as a nonlinear function of the distance between the known sample and the sample to be classified, and the closer the distance is, the higher the weight will be and the greater the influence on the result of classification prediction.

When the *K*-nearest neighbor algorithm is used for classification, the distance between each sample is classified and the sample in the training set needs to be calculated. If the data set is relatively large, the classification effect of the *K*-nearest neighbor algorithm is relatively low, and it takes a long time to complete the classification prediction. To improve this defect, the thinking of this paper is on the premise of guaranteeing the classification effect, first of all, according to some features to divide the whole data set into several subsets, and classification on the basis of these characteristics makes samples in the corresponding feature subset to find the most similar neighbors; this will greatly reduce the search time, so as to improve the efficiency of classification prediction. Based on this idea and considering the advantages of the weighted *K*-nearest neighbor algorithm, a partitioned weighted *K*-nearest neighbor algorithm is proposed in this paper.

The specific steps are as follows:

(i) Step 1. Initialize the value of K.

(ii) Step 2. Load the data set and divide the data set into several feature subsets according to a certain feature.

(iii) Step 3. The Euclidean distance between the sample to be classified and the training sample in the corresponding feature subset was calculated.

(iv) Step 4. Sort the distance values in ascending order.

(v) Step 5. Extract the first K distances from the sorted array.

(vi) Step 6. Gaussian function is used to calculate the voting weights of the first K distances.

(vii) Step 7. The categories of samples corresponding to the first K distances were obtained, and the weighted total score of each category was calculated.

(viii) Step 8. Return to the predicted results.

Table 3: Input characteristics.

Situation	Input features
1	Entrance English score, English score 1, English score 2, semester, gender, school, and major
2	Entrance English score, English score 1, English score 2, semester, gender, and school
3	Entrance English score, English score 1, English score 2, semester, gender, and major
4	Entrance English score, English score 1, English score 2, semester, school, and major
5	Entrance English score, English score 1, English score 2, gender, school, and major
6	Entrance English score, English score 1, English score 2, semester, and gender
7	Entrance English score, English score 1, English score 2, and semester
8	Entrance English, English 1, and English 2

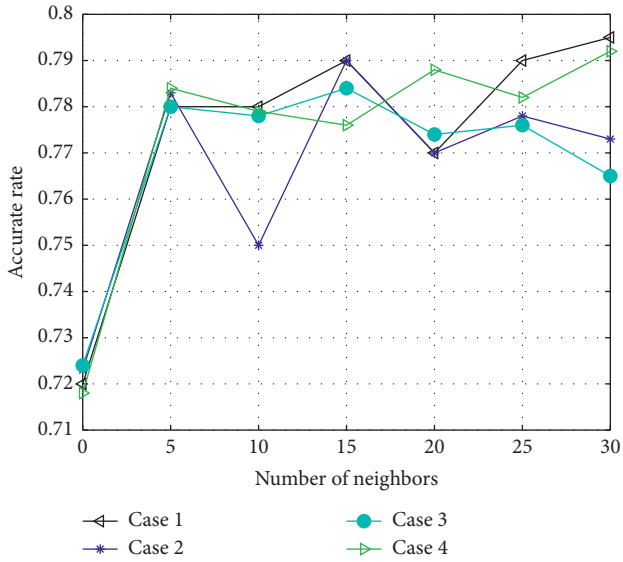

Figure 6: Precise rate diagram.

Table 4: Evaluation index table of K-nearest neighbor algorithm.

Accuracy	Precision	Recall	F1
0. 7718	0. 798	0. 7639	0. 7785

In the analysis of the factors affecting CET-4 scores, it is found that semester is a feature closely related to the test results, and the CET-4 pass rate varies greatly from semester to semester. Moreover, the semester is a classification variable, which is suitable for the classification of the data set. Therefore, this paper divides the data set into three feature subsets based on terms.

According to the above ideas and the steps of the partition-weighted K-nearest neighbor algorithm, the partition-weighted K-nearest neighbor classification model can be established to realize the prediction of the results of CET-4. In order to investigate the classification of the model prediction effect, still in the "admission English," "English 1," and "English 2" as the input characteristics, traverse the odd number K, calculation under the condition of different neighbor number K scores of each index and the prediction model using a python program to realize the classification accuracy and recall rate, precision rate and f values of visualization, and get the graph 7 evaluation indexes. In Figure 7, the abscissa of each subgraph is the nearest neighbor number K, and the ordinate is the score of each evaluation index. The curve of each index increases with the increase of the nearest neighbor number K, and the curve gradually flattens out, in order to compare the classification prediction effect of partition-weighted K-nearest neighbor algorithm with weighted K-nearest neighbor algorithm and K-nearest neighbor algorithm. Table 5 shows the scores of each evaluation index of the three algorithms with different nearest neighbor numbers. A comparative analysis of the scores of each evaluation index of the three algorithms shows that the classification and prediction effect of the partition-weighted K-nearest neighbor algorithm are not significantly reduced because of the partition. Of course, the accuracy of its classification has not improved significantly. However, the prediction time of the partition-weighted K-nearest neighbor algorithm is effectively reduced.

By comparing the operation time of the three algorithms, the efficiency of the weighted K-neighbor algorithm and K-nearest neighbor algorithm is basically the same, and there is no significant change. The average operation time of the partitioned weighted K-nearest neighbor algorithm is reduced by 6.17 seconds compared with the 11.39 of the K-nearest neighbor algorithm, and the classification time is greatly reduced compared with the K-nearest neighbor algorithm, and the classification efficiency is improved by

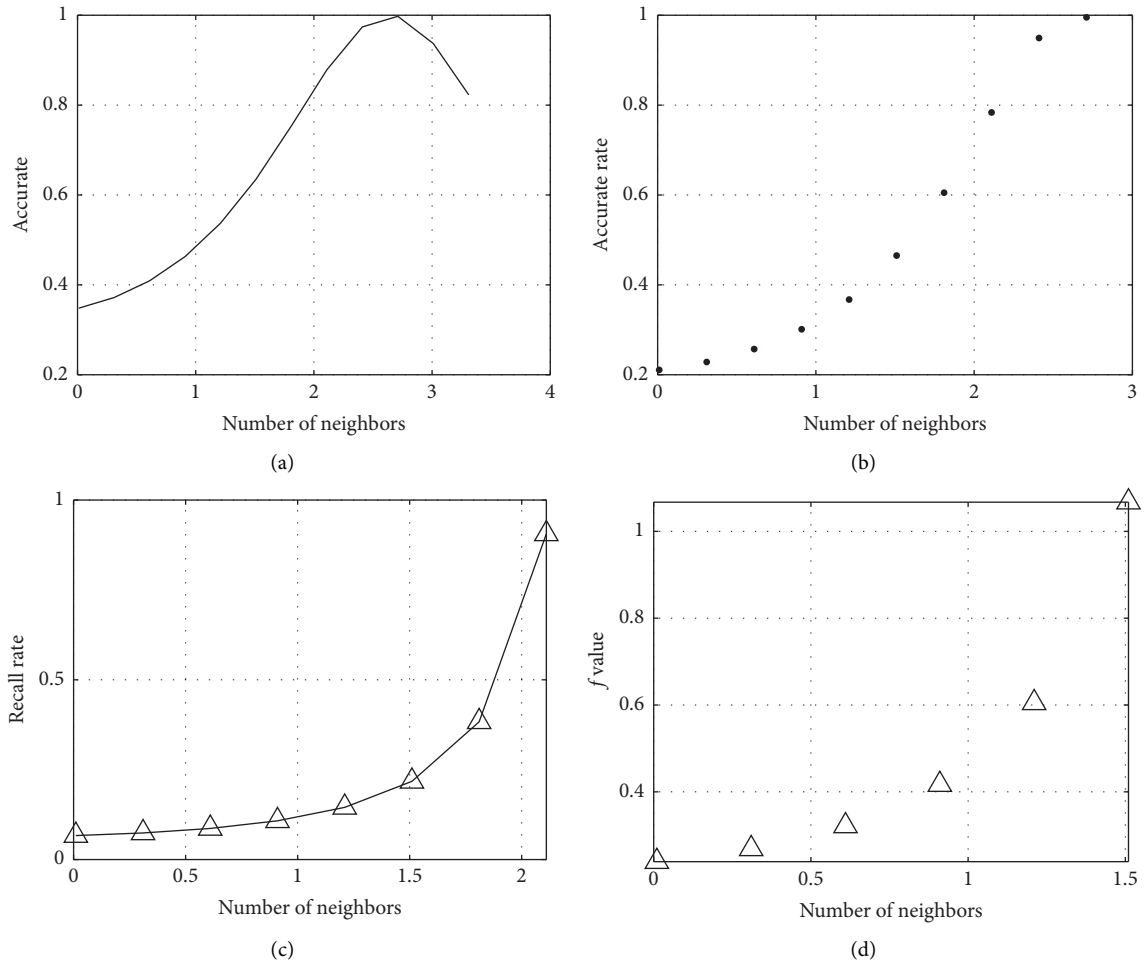

Figure 7: Partition-weighted K-nearest neighbor evaluation index curve.

Table 5: Scores of each evaluation index.

K	Precision rate			Accurate rate		
	K-nearest neighbor	Weighted K- nearest neighbor	Partition-weighted K-nearest neighbor	K-nearest neighbor	Weighted K- nearest neighbor	Partition-weighted K-nearest neighbor
1	0. 703	0. 702	0. 691	0. 716	0. 716	0.698
2	0.732	0.764	0. 733	0.738	0.738	0.728
3	0.741	0. 726	0.752	0. 725	0. 729	0. 727
5	0.756	0. 765	0. 776	0. 729	0. 756	0. 761
7	0.765	0. 761	0.755	0. 717	0. 739	0. 755
9	0. 767	0. 763	0. 753	0. 768	0. 769	0. 739
11	0.774	0.756	0.768	0. 773	0.773	0.766
12	0. 772	0. 771	0. 761	0. 765	0.764	0. 762
13	0.762	0.66	0. 772	0. 768	0. 771	0.766
19	0.764	0. 772	0.766	0. 763	0. 771	0.766
21	0. 764	0. 768	0. 757	0. 729	0. 769	0. 773
23	0.761	0. 773	0.774	0. 767	0.773	0. 772
25	0. 712	0. 771	0.773	0. 762	0. 778	0. 772
27	0.734	0.774	0. 765	0. 768	0. 778	0.774
29	0.765	0. 772	0. 771	0.723	0. 769	0.773
31	0.772	0.766	0.773	0.752	0. 769	0.776
33	0.775	0.768	0.772	0.756	0. 771	0.763
33	0. 768	0. 77	0.773	0. 727	0.763	0. 773
Average	0.762	0. 765	0. 763	0.767	0. 768	0. 764

118%. Therefore, if you want to quickly predict students' CET-4 results, you can choose the partition-weighted *K*-nearest neighbor algorithm.

5. Conclusion

In this paper, the principle based on one-time record is used to improve the algorithm of association rules. The improved optimization algorithm is used to mine students' CET4 scores in the database, and the correlation between CET4 scores is mined out. The scientific analysis of these relationships provides good decisions for education and teaching administrators and teachers, which can better guide the teaching work. The *K*-nearest neighbor algorithm is used to predict whether the college students can pass the CET-4 examination. Considering the stability of classification performance, the *K*-nearest neighbor algorithm is improved and the weighted *K*-nearest neighbor algorithm is established. It is found that although the classification accuracy is not significantly improved, the stability of classification is improved. Considering the efficiency of classification, the algorithm is further improved, and the partition-weighted *K*-nearest neighbor algorithm is established. The time required for classification is greatly reduced, and the classification efficiency is greatly improved. The application of data mining technology to school teaching management decision-making, students' psychological analysis, teaching quality evaluation, students' moral education evaluation, and so on is a new subject to be studied. Although there are still many imperfections in this system, I believe that, with the continuous advancement of the research, the functions of the system will be more abundant and practical. In this paper, only three English scores were used to predict the passing of CET-4 without other features, so the prediction accuracy was not very high. Therefore, it is necessary to further study other related factors and features in future work to further improve the prediction effect. When predicting whether students can graduate normally, the data set is relatively small and the data are unbalanced. Although the balance processing has been done, it has an impact on the reliability of the prediction. Therefore, future work is needed to collect and accumulate more data for further prediction. The model application in this paper is not quite perfect, and the data set used in the study of classification results may not be the best data set, which needs further study in the future. If we want to extend the application of the model to other courses, we need to put forward a better method, which needs to be studied.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. Zhou, Z. Zhang, and J. Li, "Analysis on course scores of learners of online teaching platforms based on data mining," *Ingénierie des systèmes d information*, vol. 25, no. 5, pp. 609–617, 2020.

[2] Z. liang, "Analysis and prediction of influencing factors of fiscal revenue in wenzhou based on data mining technology," *Advances in Social Sciences*, vol. 6, no. 12, pp. 1510–1519, 2017.

[3] J. Xu and Y. Liu, "CET-4 score analysis based on data mining technology," *Cluster Computing*, no. 5, pp. 1–11, 2018.

[4] D. S. Chezganov, M. A. Borovykh, and O. A. Chikova, "Prediction of steel corrosion resistance based on EBSD-data analysis," *IOP Conference Series: Materials Science and Engineering*, vol. 192, no. 1, pp. 012031–012043, 2017.

[5] B. Ju Lee and J. Yeol Kim, "Indicators of hypertriglyceridemia from anthropometric measures based on data mining," *Computers in Biology and Medicine*, vol. 57, pp. 201–211, 2015.

[6] H. J. Zhang, K. Wei, and A. P. Tchameni, "Methodology of uncertainty analysis prediction based on multi-well data fusion," *Geosystem Engineering*, vol. 21, no. 3, pp. 142–150, 2017.

[7] G. P. Diller, S. Orwat, and J. Vahle, "Prediction of prognosis in patients with tetralogy of Fallot based on deep learning imaging analysis," *Heart (British Cardiac Society)*, vol. 106, no. 13, pp. 20193–20212, 2020.

[8] N. Jain, J. L. Brock, A. T. Malik, F. M. Phillips, and S. N. Khan, "Prediction of complications, readmission, and revision surgery based on duration of preoperative opioid use," *Journal of Bone and Joint Surgery*, vol. 101, no. 5, pp. 384–391, 2019.

[9] S. W. Kim and S. Won, "Prediction of product distribution in fine biomass pyrolysis in fluidized beds based on proximate analysis," *Bioresource Technology*, vol. 175, pp. 275–283, 2015.

[10] T. R. Kumar, P. Yasaswini, and G. Rafi, "Comparative analysis on job prediction of students based on resume using data mining techniques," *International Journal of Engineering and Technology*, vol. 7, no. 2.7, pp. 1100–1121, 2018.

[11] J. Yoon, "Cycling winner prediction model by using match information: application of decision tree analysis based on data mining," *The Korean Journal of Measurement and Evaluation in Physical Education and Sports Science*, vol. 19, no. 4, pp. 15–26, 2017.

[12] D. Ramesh, "Analysis of crop yield prediction using data mining techniques," *International Journal of Research in Engineering and Technology*, vol. 4, no. 1, pp. 470–473, 2015.

[13] Y. H. Pang, H. B. Wang, and J. J. Zhao, "Analysis and prediction of hydraulic support load based on time series data modeling," *Geofluids*, vol. 2020, no. 1, 45 pages, Article ID 8851475, 2020.

[14] M. Zhang, L. Shen, and B. Liao, "Research on the quantitative indicators for 3D prospectivity prediction based on spatial data mining: a case study of Zhonggu ore field in Ningwu Basin," *Scientia Geologica Sinica*, vol. 53, no. 4, pp. 1300–1313, 2018.

[15] Z. Li, "Analysis and prediction of yunnan CPI series—based on SARIMA model[J]," *Statistics and Applications*, vol. 5, no. 2, pp. 155–162, 2016.

[16] L. Matrosova and E. Semenov, "Prediction of road safety indicators based on statistical analysis," *Central Russian Journal of Social Sciences*, vol. 11, no. 3, pp. 55–60, 2016.

[17] T. Zhang, "Analysis and prediction of building settlement deformation based on AR model," *Geomatics Science and Technology*, vol. 7, no. 2, pp. 83–89, 2019.

[18] M. Praveena, R. A. Deepika, and C. S. Raghavendhar, "Analysis on prediction of heart disease using data mining

techniques," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 10, no. 2, pp. 126–136, 2018.

[19] G. Saltos and M. Cocea, "An exploration of crime prediction using data mining on open data," *International Journal of Information Technology and Decision Making*, vol. 16, no. 5, pp. 256–274, 2017.

[20] L. Huafeng, "Analysis of computer teaching pattern based on outlier data mining and machine learning," *Journal of Intelligent and Fuzzy Systems*, pp. 1–11, 2020.

[21] G. Yu and W. Fu, "Analysis of distributed database access path prediction based on recurrent neural network in internet of things," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 1, pp. 1–9, 2020.

[22] Z. Wei, Z. Zhang, W. Gu, and N. Fang, "Visualization classification and prediction based on data mining," *Journal of Physics: Conference Series*, vol. 1550, pp. 032122–032135, 2020.

[23] C. Beyene and P. Kamat, "Survey on prediction and analysis the occurrence of heart disease using data mining techniques," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 8, pp. 165–173, 2018.

[24] P. Sagar, P. Pranima, and I. Indu, "Analysis of prediction techniques based on classification and regression," *International Journal of Computer Applications*, vol. 163, no. 7, pp. 47–51, 2017.

[25] C. Li, "Research on the prediction model of shrin*K*ing cities based on data envelopment analysis," *Advances in Applied Mathematics*, vol. 9, no. 5, pp. 722–727, 2020.

[26] M. A. Shaik, D. Verma, and P. Praveen, "RNN based prediction of spatiotemporal data mining RNN based prediction of spatiotemporal data mining," *IOP Conference Series Materials Science and Engineering*, pp. 981–998, 2020.