Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5717895, 10 pages
https://doi.org/10.1155/2022/5717895

Research Article

Application of Decision Tree Algorithm Based on Data Mining in English Teaching Evaluation

Rui Hou

Department of Hotel Management, Zhengzhou Tourism College, Zhengzhou, Henan 450009, China

Correspondence should be addressed to Rui Hou; hourui@zztrc.edu.cn

Received 18 March 2022; Revised 15 April 2022; Accepted 27 April 2022; Published 7 June 2022

Academic Editor: Chia-Huei Wu

In order to bring new opportunities and challenges to teaching and learning English. This paper proposes a study of the use of a data tree-based decision tree algorithm in English language learning assessment. Teaching and learning through online English learning platforms using data mining technology to analyze student learning data, create relevant models, and explore the relationship between English exams and various elements is important for student learning and teacher teaching. Firstly, this paper introduces the function, process, and common machine learning models of data mining. Finally, by using data created by college English tutors and measurements to create and identify student images, through image users, you can learn about student behavior, learning behaviors, etc., provide a foundation after the design experiment, technology training technology in logistic regression model, wooden model making, and postfusion modeling of two models and analyze the factors affecting students' passing the exam according to the prediction results. The results show that the decision tree model predicts that the score of each question type occurs 72 times, the score of students' examination occurs 70 times, and the completion of homework occurs 70 times. Trees determined by this approach can have a profound impact on a wide range of indicators of academic excellence and provide a fundamental and useful basis for measuring future improvement.

1. Introduction

With Internet plus, Internet turn the world upside down. Internet plus education is becoming a trend. Online education platform is attracting more and more attention. More schools and students are willing to learn through online English learning platform. In teaching English at colleges and universities, educators teach English students to listen, speak, read, and write through online English platforms, which for students to learn more easily and efficiently [1, 2]. Data mining technology has been widely used in all walks of life and has achieved certain results. Teaching and learning through online English learning uses data mining technology to mine and analyze student curriculum and uses training models to explore relationships, quality of English language proficiency, and many other important factors for student learning and teachers. From the evolutionary process of data mining, the development of data mining has experienced four main processes: data collection, data access, data warehouse technology, and data analysis to realize decision-making knowledge [3, 4].

At present, there are many improved algorithms for decision tree algorithms, which are generally divided into two categories, one is based on the decision tree building process, and the other is based on the decision tree pruning process. In the optimization of decision tree building process, there are some optimizations of attribute selection, some optimization of split attribute calculation method, etc.; among them, there is more research on the selection standard of split attribute; in the optimization method of decision tree pruning process, it mainly focuses on the improvement of pruning algorithm and the application of new pruning strategy to the pruning process (see Figure 1). Secondly, in the improvement of decision tree algorithm, another kind is the research on the application of decision tree, such as the multivalue of decision tree variables, and the size of data set [5]. There are three kinds of methods used in dealing with the size of data set in decision tree: first,

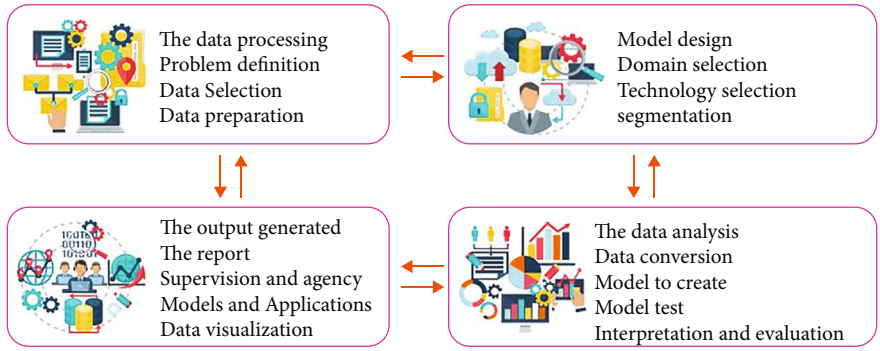

Figure 1: Application and research of data mining algorithm based on decision tree.

divide the data reasonably in the data preprocessing stage, process the big data into small data, and then, apply the algorithm; second, in the decision tree building stage, the decision tree building nodes are processed in parallel; the third is to improve the decision tree algorithm based on parallel computing algorithm. Among them, the parallelized decision tree algorithm has the most research potential. With the development and application of data mining technology, many university researchers have begun to study how to use data mining technology in English language assessment. English language usages assessment, such as its use in the management of student records, college and university assessments, student achievement. Monitoring and evaluation, in the system and enrollment, have played a positive role in improving the level of instructional management in schools [6, 7].

2. Literature Review

The application field of data mining technology is very broad. It can be widely used in various enterprises and institutions such as banking, finance, retail and wholesale, manufacturing, insurance, public facilities, government, education, remote communication, software development, transportation, and national defense scientific research. It can be said that where there is data accumulation, there is a place for data mining technology. Taranto Vera and others proposed two feature selection anomaly detection methods based on support vector data description (SVDD). One is the minimization description method, which uses the square of the partition radius as a standard function of the measured value of the normal observed boundary size [8]. Wang and others proposed a method to improve the accuracy of decision tree based on weight [9]. Li and others proposed a method to improve the performance of SLIQ algorithm. The algorithm improves the accuracy of decision tree by reducing the diversity of decision tree [10]. Shichkina and others believe that the evaluation index of teaching quality is an important basis for the teaching quality assurance system. For example, the primary indicators include teachers' teaching methods, teaching contents, teaching literacy, and teaching effects, while the secondary indicators are the concretization of the primary indicators [11]. Liu and others proposed a learning method to improve the information entropy of decision tree. The algorithm improves the calculation of information entropy by studying the uncertainty deviation of entropy [12]. Ramos and others proposed an algorithm to improve the performance of VFDT C very fast decision tree. The algorithm extends VFDT to two directions. One is to use the method of nonuniform interval numerical pruning to deal with numerical attributes; one is to use Bayesian classifier to detect peripheral node samples to reduce the size of decision tree [13]. Cui proposed two parallel methods of decision tree, one based on synchronous tree and the other based on split tree [14]. Virupaksha and Dondeti proposed the implementation of parallel algorithm of decision tree classification algorithm in PVM system and realized this method through PVM [15]. Radhika and Masood proposed a parallel algorithm of decision tree classification algorithm based on MPI. The data parallel strategy is adopted in the algorithm. The data is divided into an equal number of data blocks according to the number of processors. Each processor processes the local data and summarizes the statistics to a certain machine for the calculation of information entropy. However, the data segmentation is divided on each computing node, and the segmentation results are uploaded to a certain node for summary; after processing, the final result is the output [16].

Based on this research, this paper presents research on the use of logging algorithms based on data mining in English language testing. Data mining and analysis, how to, and identify student training programs, English or academic information, and my circumstances related to Test results based on the curriculum, this sentence uses student learning materials in English instruction and assessment for mine through portraits analysis and uses models that affect mine factors affecting the test.

3. Research Methods

3.1. Data Mining Theory

3.1.1. Data Mining Process. Knowledge Discovery in Database (KDD) creates schemas that describe data characteristics and relationships of data-based data analysis. Generally, this discovered knowledge can be used in two basic ways: one is to provide knowledge to guide business activities [17]. For example, through knowledge extraction,

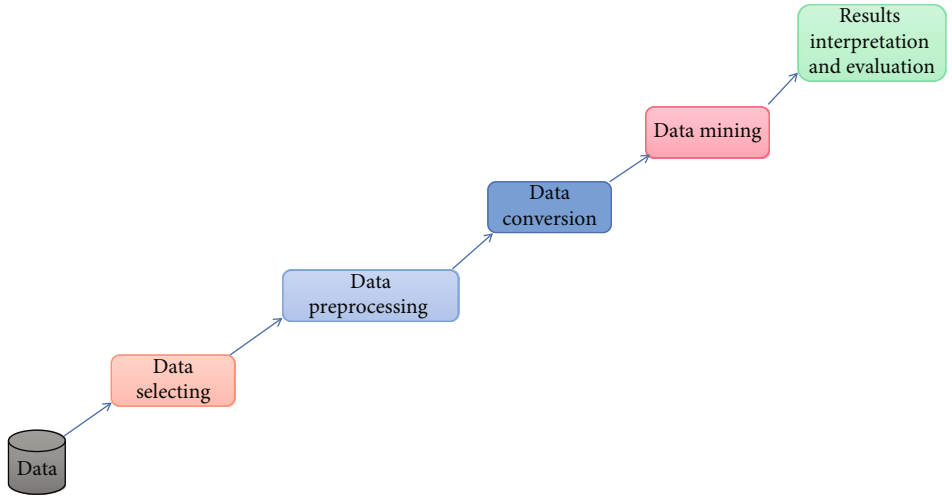

Figure 2: KDD process diagram.

we can get the associations between the goods sold in the supermarket. According to these associations, we can guide the delivery plan, and through the analysis of the sales situation of the goods, we can get the factors affecting the sales, to guide the delivery and reduce the possible backlog. Second, this model can be used for prediction. The (Knowledge Discovery in Database, KDD) process is shown in Figure 2.

3.1.2. Data Mining Process. Data mining consists of different processes in different applications and processes, and different processes are designed according to different goals and needs. In general, the data mining process can be combined into five stages: preparation stage, data preprocessing and analysis stage, model training stage, model validation and evaluation stage, and online use stage. Figure 3 shows the five steps of the data mining process.

3.1.3. Data Mining Environment. The data only refers to the complete process of excavation previously unknown and valid; data can be accessed from large data and use this information to determine or enhance knowledge. The data mining environment can be described according to the process in Figure 4.

3.2. Common Machine Learning Models of Data Mining

3.2.1. Logistic Regression Model

(1) Model Introduction. Linear regression (logistic regression, LR) is the process of multiplying each feature and attribute in a training sample by a parameter and adding it to the output. The general form of the model is shown in

$$h(x) = w_0x_0 + w_1x_1 + w_2x_2 + \cdots + w_mx_m. \quad (1)$$

Expressed by vector as

$$h(x) = w^Tx. \quad (2)$$

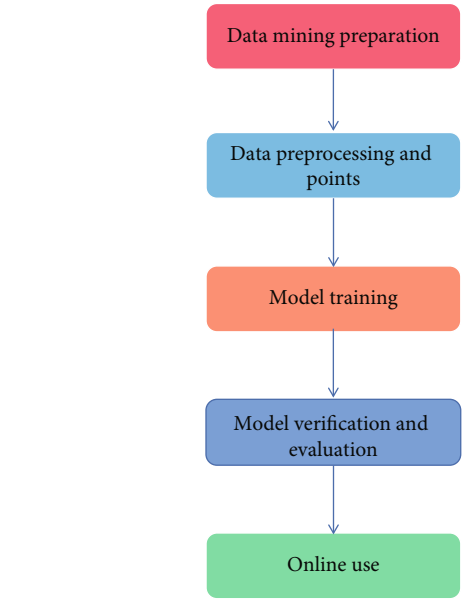

Figure 3: Five stages of data mining.

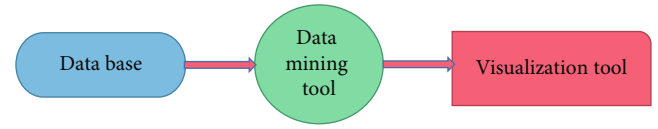

Figure 4: Data mining environment.

Logistic regression is the taking of a decimal (0.1) to include the sum of a linear regression in a Sigmoid function, as shown in

$$g(z) = \frac{1}{1 + e^{-wx}}. \quad (3)$$

Of these, W is the output parameter or regression coefficient resulting from the model training output, X is the

input sampling point, and the image of the sigmoid function is shown in Figure 5.

3.3. Model Training. Suppose there are n training models, and the result of each model follows the Bernoulli division, $p(y_i = 1|x_i)$ represents the probability of occurrence of positive class, and the probability of occurrence of negative class is $1 - p(y_i = 1|x_i)$. For each sample, the posterior probability is as shown in

$$p(y|x, w) = p(y_i = 1|x_i)^{y_1}(1 - p(y_i = 1|x_i))^{1-y_1}. \quad (4)$$

Thus, the maximum likelihood function of the sample is the posterior probability product of each sample, as shown in

$$L(w) = \prod_{i=1}^{m} p(y_i = 1|x_i)^{y_i}(1 - p(y_i = 1|x_i))^{1-y_i}. \quad (5)$$

The log likelihood function is shown in

$$\mathrm{l}(w) = \sum_{i=1}^{m} p(y_i = 1|x_i)^{y_i} + \log\,(1 - p(y_i = 1|x_i))^{1-y_i}. \quad (6)$$

Expand and solve it and derive W, as shown in

$$\frac{\partial \mathrm{l}(w)}{\partial w} = \sum_{i=1}^{m}(y_i - g(z))x_i. \quad (7)$$

Let the derivative be 0; we can see that w cannot be solved, so we need to optimize the algorithm to solve W.

3.3.1. Decision Tree Model

(1) Overview of Decision Tree. Decision tree model is a common machine learning algorithm. Its shape is a tree structure, which is a separate binary tree or multitree. Among various algorithms of data mining, decision tree model is the most intuitive in form. In the classification problem, the data set is classified according to specific characteristics, and each path to the leaf node in the decision tree is transformed into if then form. In this way, the decision tree can be simplified to make the model easier to read. It can also be considered as the conditional probability distributed in the feature space and category space. For example, the feature that can best distinguish the category is cut according to a certain condition in the feature space, so as to separate the feature space and cut other features by such inference [18–21]. The purpose of decision tree learning is to divide the sample data set according to some characteristics and finally generate a decision tree with generalization performance and strong generalization performance, that is, the generated decision tree can solve the problem well when dealing with unknown new sample data. Table 1 shows the relationship between the decision tree and the natural tree in terms of representation and classification.

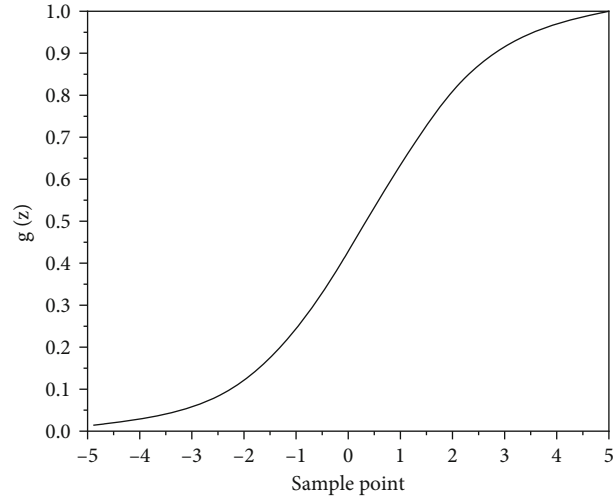

Figure 5: Sigmoid function image.

3.3.2. Common Algorithms of Decision Tree. Commonly used algorithms in decision tree design include the ID3 algorithm and the C4.5 algorithm. These commonly used algorithms can be used to divide data sets.

(1) ID3 Algorithm. Data entropy is one of the most commonly used parameters to measure the purity of data in a sample data set. The information entropy of a data set D is given by

$$H = -\sum_{k=1}^{|y|} P_k \log_2 P_k, \quad (8)$$

where H represents the information entropy of data set D, $|y|$ represents the number of categories in data set D, P represents the probability of each category in the data set, and a negative sign indicates that the information entropy is a positive number [22].

The idea of ID3 algorithm is the information gain criterion. The so-called data gain is the change in information before and after the data set is divided into the specified categories. Under the given characteristic conditions, the entropy of D is given by

$$H(D, a) = H - \sum_{v=1}^{v} \frac{|D^v|}{|D|} H(D^v), \quad (9)$$

where a represents a feature in the sample. $H\,(D, a)$ represents the information gain of feature a. V represents the number of values of feature a.

(2) C45 Algorithm. According to the shortcomings of ID3 algorithm, it has certain limitations on the splitting of nodes. In order to reduce the adverse impact of ID3 algorithm, its improved algorithm C45 appears. The optimal splitting point of C45 algorithm is selected by calculating the gain ratio. Its calculation attribute splitting formula is shown in

Table 1: The structure and representation of the decision tree.

Natural tree	The meaning of correspondence in decision tree	Representation meaning in classification problems
Tree root	Root node	Training instance entire data set space
Tree right	Internal (nonleaf) node and decision node	Attribute (set) of the object to be classified
Branch	Branch	A possible value of attribute
Leaf	Leaf node, state node	Data segmentation (classification results)

$$H(D,a)_{\text{ratio}} = \frac{H(D,a)}{H(D)}. \tag{10}$$

The process of splitting a data set using the C4.5 algorithm is similar to the ID3 algorithm. Before splitting each of the internal nodes in the decision tree, the data set must be split. The information gain ratio of each feature was calculated according to the formula above separately, and the features with the maximum information gain ratio were selected for division and calculated using the above formula.

3.4. Decision Tree Pruning. In the process of creating the decision tree, for example, when creating the classification tree, in order to divide the samples into the correct categories as much as possible, the node splitting process has been repeated, which may lead to too many branches of the decision tree. Therefore, the established decision tree learning model is prone to overfitting [23]. In order to prevent this phenomenon, it is necessary to prune the decision tree. Therefore, the risk of overfitting the decision tree can be reduced by pruning. There are two main methods of decision tree pruning, namely, "prepruning" and "postpruning."

3.5. Classification Regression Tree Algorithm. Classification regression tree belongs to a kind of decision tree and is a very important decision tree. It can be used to generate either a classification tree or a regression tree. Cart algorithm is a binary recursive segmentation technology. In the recursive process of creating classification tree, the principle of cart algorithm selecting splitting point is to calculate Gini index for the current data set, and select the feature with the smallest Gini index to select the splitting point. The Gini index formula for calculating data set D is as shown in

$$\text{Gini}(D) = 1 - \sum_{k=1}^{|y|} P_k^2, \tag{11}$$

where $|y|D$ represents the number of classes in the data set, P_k represents the probability of a category in the data set, and the implicit constraint in the formula is that the sum of the probability of each category is 1. The greater the Gini index, the greater the uncertainty. On the contrary, the smaller the Gini index, the greater the purity. Since cart algorithm creates a binary tree, there are only two values of K.

Calculate the Gini index on feature a in the training data set, and its formula is shown in

$$\text{Gini}_{\text{index}(D,a)} = \sum_{v=1}^{|v|} \frac{|D^v|}{D} \text{Gini}(D^v), \tag{12}$$

where v is the number of attribute values, $|D^v|/|D|$ represents the ratio of the number of samples contained in feature a when the value is V to $ijt\}$ training data set, and $\text{Gini}(D^v)$ represents the Gini index of sample data when the value of feature a is v.

The decision tree provides a way to show the rules, such as under what conditions and what value to obtain. For example, when applying for a loan, it is necessary to filter the risk of the application. Figure 6 lists examples of decision trees built to solve this problem; here, we can see the decision nodes, branches, and leaves, which are the main components of the decision tree [24].

3.6. Build User Portrait

3.6.1. Overview of User Portrait. English teaching is different from other subjects and needs to be really integrated into the situation. Learning English in context can mobilize students' enthusiasm for learning, let students gain experience in the learning process, enrich knowledge and stimulate students' interest in learning at the same time. English teaching and evaluation system is a platform for students to listen, speak, read, and write English in the scene. It is a language teaching environment suitable for English classroom teaching, proficiency testing, and independent training at the undergraduate and graduate levels. Through the platform, students can train in listening, speaking, reading, and writing and support a variety of teaching activities, which can be expanded according to different needs of English teaching, so as to improve students' English language skills. User profile is to describe and depict the user by using some tags based on the user's basic attributes, past historical behavior, and other data. These tags can objectively, accurately, and visually describe the target user, such as the portrait data of a user on a website. Table 2 is part of the data.

It can be seen from the data in Table 2 that the so-called user portrait is the labeling of user information. It is constructed by means of data statistics, classification, regression, and clustering in data mining. At present, it has been widely used in various fields such as the Internet and operators. The user's behavior is constantly changing, and its label data is correspondingly changing. Therefore, the user portrait has been constantly revised and increased.

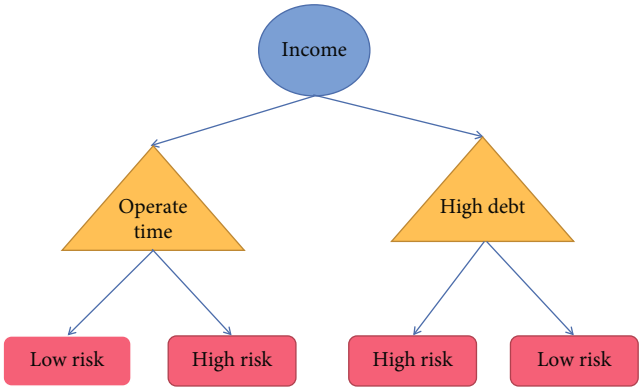

Figure 6: Example decision tree.

Table 2: Partial data of a user's portrait of a website.

Gender	Age	Marital status	Province	Income
Male	29	Unmarried	Beijing	10 thousand

3.7. User Group Portrait and Analysis. The data of user group portraits mainly comes from the learning data generated by the College English teaching and evaluation system. Students use the system for English practice and English pretest, midtest, and posttest. After answering the questions, they fill in the questionnaire data and the data in the system database to investigate boys, girls, and all students, respectively.

3.7.1. Main Purpose of Learning English. The goal of learning English is shown in Figure 7. These are: (A) likes English, (B) needs to communicate in English, (C) improves complex skills and quality, and (D) is competitive to work and travel abroad. Quantize the data and transform *A-I* to [1, 4] interval, that is, $a = 4$, and so on, $I = 1$. The analysis of the data revealed that the main goal of learning English is to improve comprehensive skills and quality, and the second reason is to be more competitive than others when working or traveling abroad. It can be seen from the data that students still agree with the importance of English in their future development, but they are not interested in English learning.

3.7.2. English Learning Attitude. According to the learning attitude, as shown in Figure 8. There are four situations: (A) often take the initiative to seek various opportunities to learn and speak English, (B) occasionally take the initiative to seek opportunities to learn and speak English, (C) never take the initiative to learn or speak English even in class, and (D) others. After analyzing the data, more than 70% of the students occasionally take the initiative to seek opportunities to learn and speak English. The proportion of girls taking the initiative to seek to learn English is higher than that of boys. Only 2% of the boys choose others. Some of the students who choose this part only read and do not speak, and some think there is no ideal environment. And less than 1% of girls choose this section. Statistics show that students are not motivated to learn English, and that most of them are actively seeking opportunities from time to time.

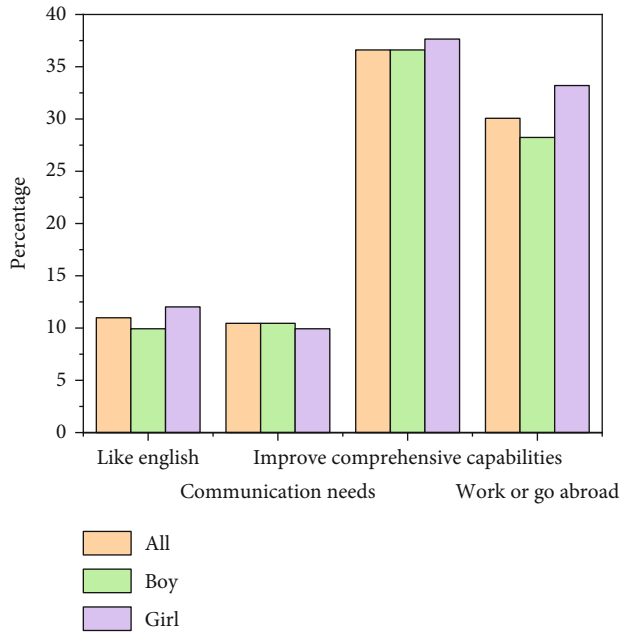

Figure 7: Main purpose of learning English.

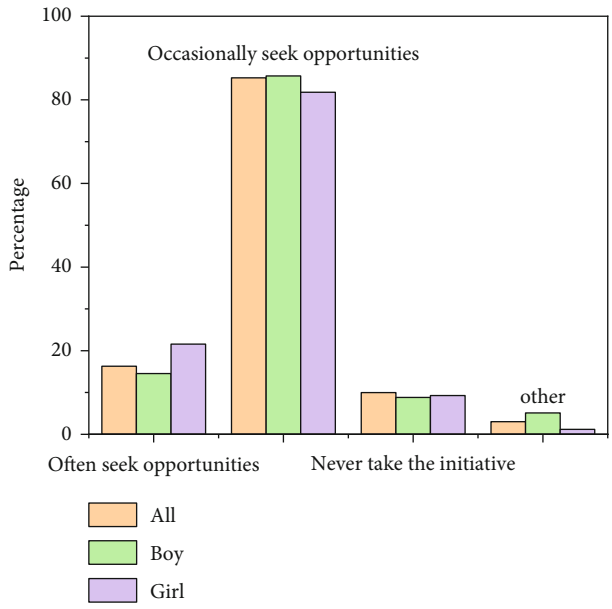

Figure 8: English learning attitude.

3.7.3. The Most Important Skills in English. According to the types of English listening, speaking, reading, and writing, there are three situations: listening, speaking, reading, and translation. As can be seen from the data in Figure 9, more than 80% of boys think listening and speaking are important, while more than 90% of girls think listening and speaking are important; in terms of reading, the proportion of boys is significantly higher than that of girls; few students think translation is important. This reflects some disadvantages in

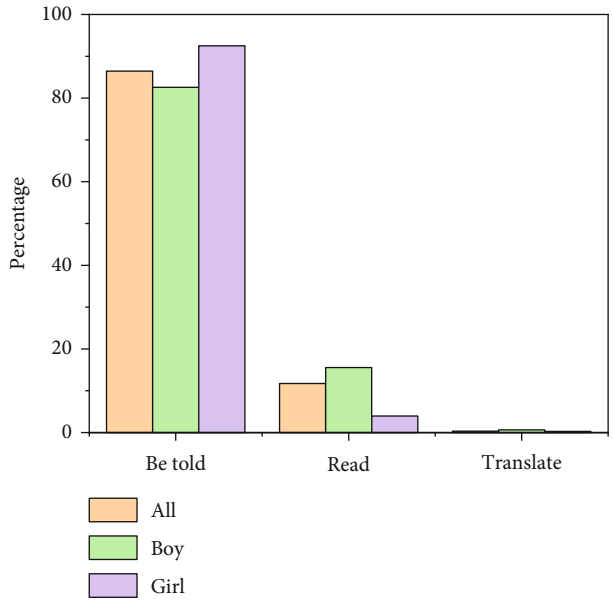

Figure 9: The most important skills in English.

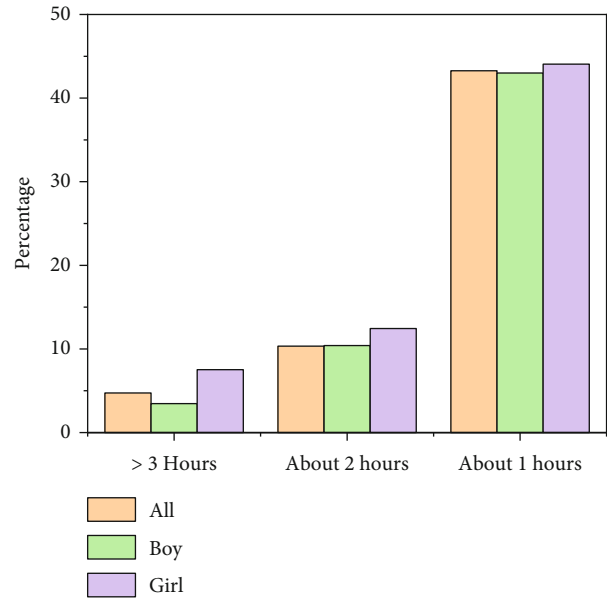

Figure 10: Time for learning English outside class.

English learning. The phenomenon of dumb English has always existed. How to improve listening and speaking skills is the key to learning English well.

3.7.4. Time for Learning English outside the Classroom. According to the time of learning English outside the classroom, it can be divided into three situations: (A) more than 3 hours, (B) about 2 hours, and (C) about 1 hour. As can be seen from Figure 10, most students study English for about 1 hour or just use class time to learn English. Few students study English for about 2 hours, and fewer students study English for more than 3 hours.

3.7.5. Reasons for Learning English Well. According to the reasons of student number English, there are three situations: (A) appropriate personal learning methods, (B) good learning environment and teaching conditions, and (C) good personal talent. It can be seen from the data that more than half of the students believe that proper personal learning methods account for the main reason, which is also the key factor to learn English well; more than 30% of students believe that learning environment and teaching conditions are the main factors for learning English well. As can be seen from Figure 11, students agree that personal subjective factors are the key to learning English well.

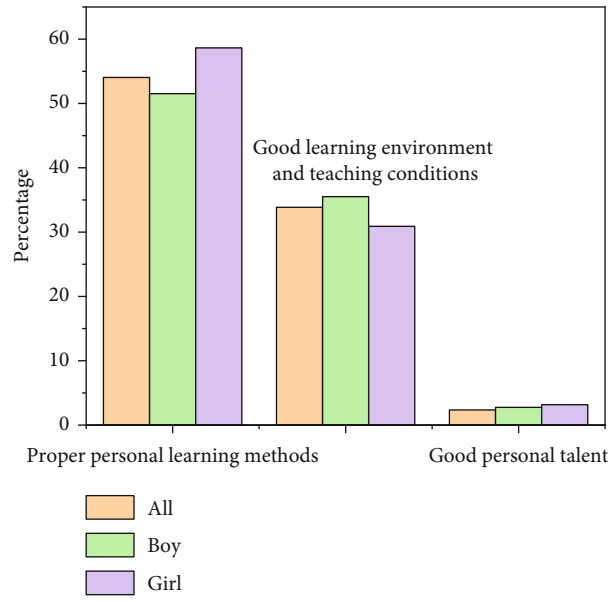

Figure 11: Reasons for learning English well.

4. Result Discussion

4.1. Data Set and Feature Processing

4.1.1. Data Extraction. Students use the English teaching and evaluation system for training to predict whether they will pass the exam. The research data comes from the teaching data generated by the system, and the information mainly includes the following three aspects:

(1) Student user portrait. Student portrait data is student individual portrait data, mainly including students' basic information, question answering behavior, and other data

(2) English training topic data English training topic data is the basic information of each topic and the data of students when doing the topic. Specific data mainly include test paper mode, question type, question type category 1, question type category 2, question type duration, unit number, course number, difficulty, and question weight

(3) Students' English training environment

The environmental data of students in English training include students' answer time, equipment information used in answer, Internet access, and IP address.

4.1.2. Data Preprocessing. There are many problems in the data obtained from the data extraction process, such as inconsistent data in the same dimension, incomplete data in a certain dimension, and noise in the data. That is, there are some "dirty data," and the quality of data is the prerequisite for using the model to predict the results. The purpose of data preprocessing is to remove those features or attributes that have little or no correlation with the target variables. By preprocessing the data, the correlation between each dimension of the data set and the target variables is high. This will provide cleaner and more efficient data for data mining, reduce the amount of data processing by machine learning model algorithms, increase the efficiency of algorithmic data processing, and increase the evaluation of the result and can be improved.

4.2. Experimental Design of Decision Tree Model to Predict Passing Grades. Test environment: Linux server, Python compilation environment, XGBoost open source tool. Test process: write a Python program, call the function of the open source tool XGBoost, design, and adjust the design parameters. Model evaluation in this study, the evaluation model used, the AUC indicator, and its AUC value, is shown in Table 3.

4.2.1. Experimental Design of Integrating Logistic Regression and Decision Tree Model to Predict Passing Grades. Fusion algorithm refers to the combination of multiple model algorithms in a certain way. The fusion algorithm plays an important role in the field of recommendation and advertising. In practical application, because the algorithm faces many application scenarios, the business data of each application scenario is also different. From the perspective of model algorithm theory, under the conditions of different business data, different algorithms are suitable for different application scenarios. There is no one algorithm, and the effect is better than other algorithms in all cases. Therefore, the fusion method can combine the better algorithms in their respective scenes to form a powerful algorithm. Linear weighting, cross fusion, and waterfall fusion are three common model fusion methods. In this experiment, the linear weighted model fusion algorithm is adopted, that is, the decision tree model and logistic regression model are linearly weighted. Different values of α are tested in the study. The specific AUC is shown in Table 4.

From the AUC value, it can be found that the method of model fusion is better than using logistic regression model and decision tree model alone.

4.3. Prediction Results and Analysis

4.3.1. Relationship between Prediction Results and Characteristics. In the experiment, the logistic regression model, the decision tree model, and the fusion method of the two models are used to predict the passing situation of students in the examination. According to the correlation between the prediction results and characteristics, it is as follows:

Table 3: AUC value of decision tree model training results.

Number of iterations	AUC
1	0.5512227181
2	0.5517231838
3	0.5510537201
4	0.5514020721
5	0.5520173803
6	0.5523072035
7	0.55281036174
8	0.55328667301
9	0.55332710731
10	0.5534047203

(1) Prediction results of logistic regression model

The relationship between the weight W of each feature output by the logistic regression model and the feature is shown in Table 5.

(2) Prediction results of decision tree model

The methods to measure the importance of features in the decision tree mainly include (1) the frequency of features, that is, the greater the frequency of features, the more important the features are; (2) when the features are close to the root node, that is, the closer the features are to the root node, the greater the Gini index, and the more important the corresponding features are. According to the prediction results of the decision tree model, the score of each question type occurs 72 times, the score of students' examination occurs 70 times, and the completion of homework occurs 70 times. The importance of features in the specific classifier model is shown in Table 6.

4.3.2. Result Analysis. Based on the sample data placed in the classifier, the variables are related to estimating the student's strength too, that is, the main factor affecting the student's test are scores of all questions, student test scores, completion of homework assignments, number of English language courses, and level of interest in English, grade 4 and grade 6, writing, and identifying everything, as follows.

The scores of each question type, students' examination scores, homework completion, English training times, English interest, and CET-4 and CET-6 are summarized and analyzed according to various factors, as follows:

(1) Students' completion of homework and the number of times of participating in English training have a great impact on passing the exam. The experimental middle school students have a high rate of completing homework or participate in English training

Table 4: AUC value of training results of logistic regression and decision tree fusion.

α	0.1	0.2	0.4	0.6	0.8	0.9
AUC	0.504213	0.510632	0.521003	0.513276	0.508263	0.507145

Table 5: Relationship between weight W and characteristics of logistic regression model.

Weight w	Features
0.404051006287	Score for each question type
0.100745136301	Student test scores
0.380654036301	Operation completion
0.314067383077	English training times
0.313051006284	Degree of interest in English
0.303458621255	Score of CET-4 and CET-6

Table 6: Importance of features in decision tree model.

Features	Importance (frequency)
Score for each question type	72
Student test scores	70
Operation completion	70
English training times	67
Degree of interest in English	63
Score of CET-4 and CET-6	62
Passing of CET-4 and CET-6	60

more times. The student has a high probability of passing the exam. Completing homework and participating in English training for many times reflect that the student spends a relatively long time in English learning and can reflect the degree of hard study, hard study, and students' attitude towards English learning. Therefore, students' high completion rate of homework and more training times have a great impact on passing the exam

(2) Students' interest in English is affected by the prediction of test scores. In the modern teaching of "student-class, teacher-oriented," the importance of promoting students' interest in learning is also important. Students' interest in learning can encourage students to participate in learning and promote students' self-directed learning experience. To some extent, an interest in English can affect students' emotional and psychological well-being in the process of learning English. For example, in speech, some English learners are influenced by localities, such as local voices, which make it difficult for anyone. If you understand it, in the long run, students will refuse to start learning English, and even hate learning English. Therefore, the level of English proficiency can be more relevant to English language learning

(3) It is found that the number of male students has a greater positive impact on the results of the survey than that of female students. Through the observation of the data, it is found that the number of male students has a greater positive impact on the results of English than that of female students. The variables "topic difficulty" and "topic weight" in the sample do not vary widely, the difficulty is moderate, and the weight of various types of topics fluctuates little, so the impact on the prediction results is not great. In the practice mode, watching video and listening repeatedly can improve the accuracy of question answers, which is very helpful for practice. However, in the examination mode, video and listening cannot be repeated. Therefore, the influence of these two factors on the examination results comes from the usual practice

5. Conclusion

In this paper, an application research method of decision tree algorithm based on data mining in English teaching evaluation is proposed and verified by experiments. Through the in-depth study of data mining (decision tree), decision tree method and other related theories and technologies, and a comprehensive analysis of the teaching data generated in the system, the main work is as follows: (1) according to the attribute and behavior data of students' practice or examination in College English teaching and evaluation, this paper constructs students' personal portrait and group portrait. Based on many reference materials and referring to the concept of Internet plus education and other user profiles in the Internet industry, the paper builds a picture of answering questions based on the existing source data. Through the establishment of group portraits and the analysis of various indicators according to the chart, students' learning habits can be obtained to a certain extent. (2) According to the experimental prediction results, the weight W value of the model parameters and the importance of the characteristics of each dimension, through the strong correlation between the relevant characteristics and the prediction goal, this paper analyzes the factors affecting students' passing English test scores, summarizes, and classifies each factor, to provide a basis for improving students' English language skills and to provide students with the highest level of examinations. This approach helps to develop targeted strategies to improve the quality of training. It is useful to observe the quality of English language teaching in different factors and environments (teaching methods, teaching methods, teacher research skills, etc.) and compare the advantages and disadvantages of different teaching activities, or compare the possibilities for different groups of students. Provide targeted improvement measures and Intuitive and sound rationale. The development of science and technology has

never stopped nor has the research on knowledge discovery. Although the decision tree algorithm and the parallelization of the algorithm have some effects, there are still some shortcomings. In the practical work application study of decision tree algorithm, although the prediction accuracy is acceptable and has certain guiding significance for the work, the model needs to be further evaluated and improved in the practical work.

Data Availability

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

Conflicts of Interest

The authors declare no competing interests.

Acknowledgments

This work is supported by the Zhengzhou Tourism College.

References

[1] H. Moayedi, M. M. Abdullahi, H. Nguyen, and A. Rashid, "Comparison of dragonfly algorithm and Harris hawks optimization evolutionary data mining techniques for the assessment of bearing capacity of footings over two-layer foundation soils," *Engineering with Computers*, vol. 37, no. 1, pp. 437–447, 2021.

[2] E. P. Booker and G. E. Jabbour, "Antiviral nanoparticle ligands identified with datamining and high-throughput virtual screening," *RSC Advances*, vol. 11, no. 37, pp. 23136–23143, 2021.

[3] Y. Li, R. K. Shyamasundar, and X. Wang, "Special issue on computational intelligence for social media data mining and knowledge discovery," *Computational Intelligence*, vol. 37, no. 2, pp. 658-659, 2021.

[4] Y. Liu, Z. Yu, and Y. Yang, "Diabetes risk data mining method based on electronic medical record analysis," *Journal of Healthcare Engineering*, vol. 2021, Article ID 6678526, 11 pages, 2021.

[5] K. Yu, W. Shi, and N. Santoro, "Designing a streaming algorithm for outlier detection in data mining—an incrementa approach," *Sensors*, vol. 20, no. 5, pp. 1261–1265, 2020.

[6] J. Zhang, "Interaction design research based on large data rule mining and blockchain communication technology," *Soft Computing*, vol. 24, no. 21, pp. 16593–16604, 2020.

[7] B. He and L. Yin, "Prediction modelling of cold chain logistics demand based on data mining algorithm," *Mathematical Problems in Engineering*, vol. 2021, Article ID 3421478, 9 pages, 2021.

[8] G. Taranto-Vera, P. Galindo-Villardón, J. Merchán-Sánchez-Jara, J. Salazar-Pozo, A. Moreno-Salazar, and V. Salazar-Villalva, "Algorithms and software for data mining and machine learning: a critical comparative view from a systematic review of the literature," *The Journal of Supercomputing*, vol. 77, no. 10, pp. 11481–11513, 2021.

[9] L. Wang, B. Lin, R. Chen, and K. H. Lu, "Using data mining methods to develop manufacturing production rule in iot environment," *The Journal of Supercomputing*, vol. 78, no. 3, pp. 4526–4549, 2022.

[10] P. Li, L. Yin, B. Zhao, and Y. Sun, "Virtual screening of drug proteins based on imbalance data mining," *Mathematical Problems in Engineering*, vol. 2021, Article ID 5585990, 10 pages, 2021.

[11] Y. Shichkina, Y. Irishina, E. Stanevich, and A. Salgueiro, "The main aspects of creating a system of data mining on the status of patients with Parkinson's disease," *Procedia Computer Science*, vol. 186, no. 9, pp. 161–168, 2021.

[12] J. Liu, H. Dong, and P. Wang, "Multi-fidelity global optimization using a data-mining strategy for computationally intensive black-box problems," *Knowledge-Based Systems*, vol. 227, no. 3, article 107212, 2021.

[13] S. Ramos, J. Soares, S. S. Cembranel, I. Tavares, and R. Fernandes, "Data mining techniques for electricity customer characterization," *Procedia Computer Science*, vol. 186, no. 3, pp. 475–488, 2021.

[14] Y. Cui, "Intelligent recommendation system based on mathematical modeling in personalized data mining," *Mathematical Problems in Engineering*, vol. 2021, Article ID 6672036, 11 pages, 2021.

[15] S. Virupaksha and V. Dondeti, "Anonymized noise addition in subspaces for privacy preserved data mining in high dimensional continuous data," *Peer-to-Peer Networking and Applications*, vol. 14, no. 3, pp. 1608–1628, 2021.

[16] A. Radhika and M. S. Masood, "Effective dimensionality reduction by using soft computing method in data mining techniques," *Soft Computing*, vol. 25, no. 6, pp. 4643–4651, 2021.

[17] W. G. Bond, M. A. Seale, and J. L. Hensley, "A dynamic hyperbolic surface model for responsive data mining," *Procedia Computer Science*, vol. 185, no. 8, pp. 170–176, 2021.

[18] A. V. Grishchenko, V. A. Kruchek, D. N. Kurilkin, and O. R. Khamidov, "Diagnostics of the technical condition of rolling bearings of asynchronous traction motors of locomotives based on data mining," *Russian Electrical Engineering*, vol. 91, no. 10, pp. 593–596, 2020.

[19] J. D. Campo-Ávila, A. Takilalte, A. Bifet, and L. Mora-López, "Binding data mining and expert knowledge for one-day-ahead prediction of hourly global solar radiation," *Expert Systems with Applications*, vol. 167, no. 8, article 114147, 2021.

[20] A. Abugabah, A. A. AlZubi, F. Al-Obeidat, A. Alarifi, and A. Alwadain, "Data mining techniques for analyzing healthcare conditions of urban space-person lung using meta-heuristic optimized neural networks," *Cluster Computing*, vol. 23, no. 3, pp. 1781–1794, 2020.

[21] M. Chang, G. D'Aniello, M. Gaeta, F. Orciuoli, and C. Simonelli, "Building ontology-driven tutoring models for intelligent tutoring systems using data mining," *IEEE Access*, vol. 8, no. 1, pp. 48151–48162, 2020.

[22] A. Yz, A. Sx, and W. B. Yu, "Performance improvement of centrifugal compressors for fuel cell vehicles using the aerodynamic optimization and data mining methods," *International Journal of Hydrogen Energy*, vol. 45, pp. 11276–11286, 2020.

[23] K. Li, T. Xu, J. Xi, H. Jia, and L. Leng, "Multi-factor analysis of algal blooms in gate-controlled urban water bodies by data mining," *Science of the Total Environment*, vol. 753, no. 1, article 141821, 2020.

[24] M. Ashraf, M. Zaman, and M. Ahmed, "An intelligent prediction system for educational data mining based on ensemble and filtering approaches," *Procedia Computer Science*, vol. 2, no. 167, pp. 1471–1483, 2020.