Hindawi
Complexity
Volume 2023, Article ID 9824174, 1 page
https://doi.org/10.1155/2023/9824174

*Retraction*

# Retracted: Application of Intelligent Fuzzy Decision Tree Algorithm in English Teaching Model Improvement

**Complexity**

Received 15 August 2023; Accepted 15 August 2023; Published 16 August 2023



This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

## References

[1] J. Li, "Application of Intelligent Fuzzy Decision Tree Algorithm in English Teaching Model Improvement," *Complexity*, vol. 2021, Article ID 8631019, 10 pages, 2021.

Hindawi
Complexity
Volume 2021, Article ID 8631019, 10 pages
https://doi.org/10.1155/2021/8631019



*Research Article*

# Application of Intelligent Fuzzy Decision Tree Algorithm in English Teaching Model Improvement

**Jingjing Li**

*Department of Foreign Languages, Xijing University, Shaanxi, Xi'an 710123, China*

Correspondence should be addressed to Jingjing Li; 20140121@xijing.edu.cn

Received 26 April 2021; Revised 13 May 2021; Accepted 6 August 2021; Published 16 August 2021

Academic Editor: Zhihan Lv



As the number of students in universities continues to grow, the university academic management system has a large amount of data on student performance. However, the utilization of these data is only limited to simple query and statistical work, and there is no precedent of using these data for improving English teaching mode. With the application of fuzzy theory in machine learning and artificial intelligence, the fuzzy decision tree algorithm was born by integrating fuzzy set theory with decision tree algorithm. In this paper, we propose a way to obtain the centroids of continuous attribute clustering by *K*-means algorithm and combine the triangular fuzzy number to fuzzy the continuous data. In addition, this paper analyzes the influence of nearest neighbor distance on classification, introduces Gaussian weight function, gives different voting weights to the neighborhood according to the distance, and establishes a weighted *K*-nearest neighbor classification algorithm. To address the problem of low classification efficiency of *K*-nearest neighbor algorithm when the dataset is large, this paper further improves the algorithm and establishes the partitioned weighted *K*-nearest neighbor algorithm. The classification time was shortened from 11.39 seconds to 5.22 seconds, and the classification efficiency greatly improved.



## 1. Introduction

It is well known that English is the most widely spoken language in the world and its importance cannot be overstated [1]. In China, there are more and more people devoted to learning English [2]. Reading not only enriches vocabulary and grammar, improves writing skills, but also broadens the horizons and access to information [3]. The provision of appropriate reading materials is the prerequisite and key to effective reading activities [4]. Therefore, how to provide learners with personalized reading materials that match their ability level and meet their learning needs has gradually become the focus of educational technology research. In order to provide learners with reading materials that meet their ability level, the first step is to assess the difficulty of the reading materials [5]. The right level of difficulty is the only way to improve learners' learning efficiency. Teaching improvement is the optimization and upgrading of the original teaching model to better carry out relevant teaching activities. The core of teaching improvement is to change the original purely receptive learning style and to establish a learning style that aims to fully mobilize and bring into play the subjectivity of students [6]. The five elements of the classroom include clear and measurable learning objectives; sufficient time for students' independent, cooperative, and inquiry learning activities; scientific guidance on learning strategies; authentic learning feedback that points to learning objectives; and democratic and equal learning with active feedback [7].

Since the birth of decision tree algorithm, it has been developed continuously and has been applied in many fields with good results [8–11]. By integrating the decision tree algorithm with fuzzy theory, the adaptability of the algorithm to different datasets has been enhanced. Nowadays, the decision tree algorithm has an irreplaceable position in the field of data mining, and its improvement and optimization work have never stopped. Data mining is the process of mining valuable, potentially meaningful, and regular information from a large amount of data, and whose data are valuable. The implementation process of data

mining involves machine learning, pattern recognition, statistical analysis, distributed storage, distributed computing, visualization, etc., in which knowledge of specialized fields is also required. In other words, data mining is a process of extracting potential, comprehensible, and valuable information from a large amount of fuzzy, incomplete, and random data of various types using various methods [12]. Currently, data mining techniques are gaining attention and importance from both academic and business communities and are widely used in various fields. Data mining tasks can be generally classified into predictive tasks and descriptive tasks. The goal of predictive tasks is to predict the value of a particular attribute based on the value of other attributes, and the goal of descriptive tasks is to derive patterns (correlations, trends, clusters, and anomalies) that summarize potential connections in the data [13]. Data mining is a process of mining information from data that are useful to users. The final data of data mining has different expressions, which are essentially the external characteristics of data mining results, while for the internal characteristics of data mining, data mining has the following expressions, i.e., potential, understandable, and valuable [14]. The value of data mining is whether the investigated data have a useful meaning and the process of data mining consists in finding it from large amounts of data. This can be more difficult than the process of searching for a small dataset [15]. Among these are many new and unforeseeable problems encountered in the processing of large amounts of data, and these problems that occur when large amounts of data are processed are not present in small records [16]. In particular, the end result presented in the data mining process may be the appearance of the data and not the nature of the data, or it may be a meaningless representation of a random process [17–23].

By applying the fuzzy theory in mechanical learning and artificial intelligence, the fuzzy set theory was integrated into the decision tree algorithm, and the fuzzy decision tree algorithm was born. In this article, we propose a method for the misuse of continuous data using the *K*-means algorithm to maintain the focus of continuous attribute clustering and to combine triangular fuzzy numbers. In addition, the paper analyzes the impact of the nearest neighbor distance on the classification, introduces gauche weight function, gives neighbors different weights by distance, and sets weighted *K*-next neighbor classification algorithm. The time required for classification was reduced from 11.39 seconds to 5.22 seconds, and the classification efficiency was greatly improved.

## 2. Intelligent Fuzzy Decision Tree Modeling

*2.1. Intelligent Decision Tree Algorithm Based on Fuzzy Theory.* Unlike classical mathematics, which is deterministic, the introduction of fuzzy theory brings the discipline of mathematics into a new field, which expands the range of applications of mathematics, bringing it from the exact space to the fuzzy space. Like all disciplines, fuzzy theory was born from the needs of production and practice. From time to time, we encounter fuzzy concepts in many areas of real life. For example, we use fast and slow when describing speed, expensive, and cheap when describing price, and it is difficult to separate these adjectives with a precise boundary. For classical mathematics, it is impossible to express the state of "both here and there" of objective things in the transition stage. With the development of science and technology, people are pursuing mathematical expressions in various fields, so the pursuit of mathematization of fuzzy concepts has led to the development of fuzzy theory. Integrating the decision tree algorithm with fuzzy theory enhanced the adaptability of algorithms to various datasets.

While classical mathematics is distinguished by its precision, in specific application scenarios, fuzziness can have an advantage over precision. For example, for the concept of "very old," a 50-year-old or younger person may consider 75 to be in the very old range, but a 90-year-old person may not consider a 75-year-old person to be very old, so the concept of "very old" does not lend itself to an exact interval. Therefore, there are times when absolute precision does not lend itself to practical application. Also, ambiguity is different from randomness, for example, if we toss a coin, it will either be tails or heads; their occurrence is random and uncertain, but the event itself is certain.

Also, vagueness indicates the uncertainty of the event itself in terms of definition (e.g., beautiful and ugly). In a sense, the development of computer technology has also prompted the study of fuzzy theory. Professor Zadeh is keen on the study of the conflict between intelligent systems and computers, analyzing the differences between computers compared to the human brain. He believes that although computers are far superior to the human brain in terms of both computational speed and storage capacity, they are far inferior to humans in processing, recognizing, perceiving, and reasoning about fuzzy information. For example, for the user login verification code, the human eye can easily identify, while the computer to identify is quite difficult. Therefore, in order to make computers more intelligent and to make computer-based complex systems capable of active recognition and fuzzy control, the key to the problem is how to convert real-world fuzziness into computationally understandable instructions, which are the areas to be studied by fuzzy theory.

The distinctive feature of fuzzy decision tree algorithm is the integration of fuzzy theory and classical decision tree algorithm, which expands the application field of decision tree algorithm. The fuzzification improvement of the classical decision tree algorithm mainly includes the following:

(1) Preprocessing of continuous attributes, i.e., how to fuzzify continuous attributes: for most fuzzy decision tree algorithms, fuzzification of continuous attributes before modeling is necessary, and a few algorithms fuzzify the data in the process of modeling.

(2) Selection rules for splitting attributes: compared with the rules for split attribute selection of clear decision tree, the fuzzy decision tree algorithm extends them to adapt to fuzzy data.

(3) Matching rules for decision trees: for a fuzzy decision tree, it will give the degree to which the test data belong to a certain classification, i.e., a reflection of the attribute affiliation, rather than an absolute classification like a clear decision tree.

Compared with the decision rules of the clear decision tree algorithm, the fuzzy rules generated by the fuzzy decision tree are more realistic, and the set composed of these fuzzy rules is called the fuzzy rule set, as shown in Figure 1. If the basis for splitting attribute selection of decision tree algorithm is collectively called clear heuristic, and that of fuzzy decision tree is called fuzzy heuristic, then the differences between these two types of algorithms are mainly in the differences of heuristics, leaf node selection criteria or branch ending criteria, and the final generated rules. Although the fuzzy decision tree algorithm is an improvement of the decision tree algorithm, it does not mean that this fuzzy algorithm is better than the clear algorithm in all aspects, and different algorithms should be chosen according to the actual application area.

*2.2. ELT Data Acquisition.* In this work, in order to figure out how to use these data more effectively in a large amount of data stored in a database of academic systems in a university, it is the key to influence the English IV test of the university. So, we need to consider the factors that affect the passing of the College English Test Band 4, as well as more effective ways to predict students' passing the exam.

In this paper, 360 texts were selected from the original texts of four English textbooks (including the Junior High School Textbook of the Renminbi edition, the Koran Middle School Textbook of the Renminbi edition, and the second edition of the 21st Century College English and New Vision College English Textbooks), including 120 texts at each of the three levels of junior high school (two difficulty levels), Koran middle school (two difficulty levels), and college (two difficulty levels), and 60 texts at each difficulty level of each level. Each level has 60 texts for each difficulty level. The six defined difficulty levels are junior-middle, junior-high, senior-middle, senior-high, and college-1. The junior-middle level of difficulty is for students in the seventh and first semester of eighth grade; the junior-high level of difficulty is for students in the second semester of eighth-grade and ninth grade. The junior-high level is for students in the second semester of their sophomore year and their junior year; the college-1 level is for students in the second semester of their junior year and their freshman year; and the college-2 level is for students in their sophomore year and beyond. The level of difficulty is suitable for sophomore and later students. The quantitative indicators of the training dataset are total words, families, PETS1, baseword l, PETS2, average sentence length, and PETS2. Average sentence length, PETS3 (number of vocabulary in national English proficiency test level 3), number of clauses (number of subordinate clauses), and some data are shown in Table 1.

*2.3. Decision Tree Data Crawling Based on Improved K-Nearest Neighbor Algorithm.* With the common application of fuzzy set theory in decision tree algorithms, many excellent algorithms have been proposed one after another, and two representative fuzzy decision tree algorithms are introduced in this section. Regardless of the decision tree algorithm, the structure of the generated decision tree is generally similar, with nodes consisting of individual attribute names and edges consisting of fuzzy subsets whose attribute values have been fuzzified.

In order to obtain accurate and reliable experimental results to complete the analysis of the algorithm, the first step is to select a suitable experimental dataset; the UCI dataset, as a standard test dataset with high frequency, has been applied to data mining research by many experts and scholars, and seven datasets are selected as experimental data in this paper. In actual English teaching, some students are able to pass the university English IV exam even though they have lower scores on the entrance English test, and some even pass the university English IV exam earlier. The reason for this is that the student did not understand or did not pay attention to the entrance English test when he or she enrolled so that he or she entered with a low English score. The decision tree algorithm occupies an unprecedented position in the field of data mining and does not stop the work of improvement and optimization.

For these students, the English scores on the entrance exam do not represent their true English proficiency. These data are anomalies, and if such data are used for prediction and classification, the accuracy of the classification will be greatly affected. In order to use the data more effectively for classification and prediction, the data samples with scores below 30 on the entrance English test were removed from this paper. The English scores of the students at this university include 20 points from their regular grades. Normally, most students have a perfect or close to perfect score, and a very small number of students have a score of less than 15. It was found that some students' "English 1" or "English 2" scores were lower than 20 because they did not take the English exam, so they only had a regular grade and no paper grade. The scores without paper scores do not reflect the true level of students and are not suitable for classification prediction. Therefore, the data of "English score 1" and "English score 2" with scores less than 20 were removed from this paper.

In order to make better use of the students' entrance English scores, the missing values of the entrance English scores need to be filled in order to make better predictions. The common methods to fill the missing values are as follows: if the dataset has a large number of samples and the missing values are relatively small, then the samples with missing values can be deleted directly; if the feature is not very important in the dataset, then the missing values can be filled with the mean or plurality of the feature; for some important data features, models such as logistic regression or random forest can be used to predict the missing values of the feature. According to the previous analysis, the entrance English scores represent students' English foundation, which is an important and useful feature. Although

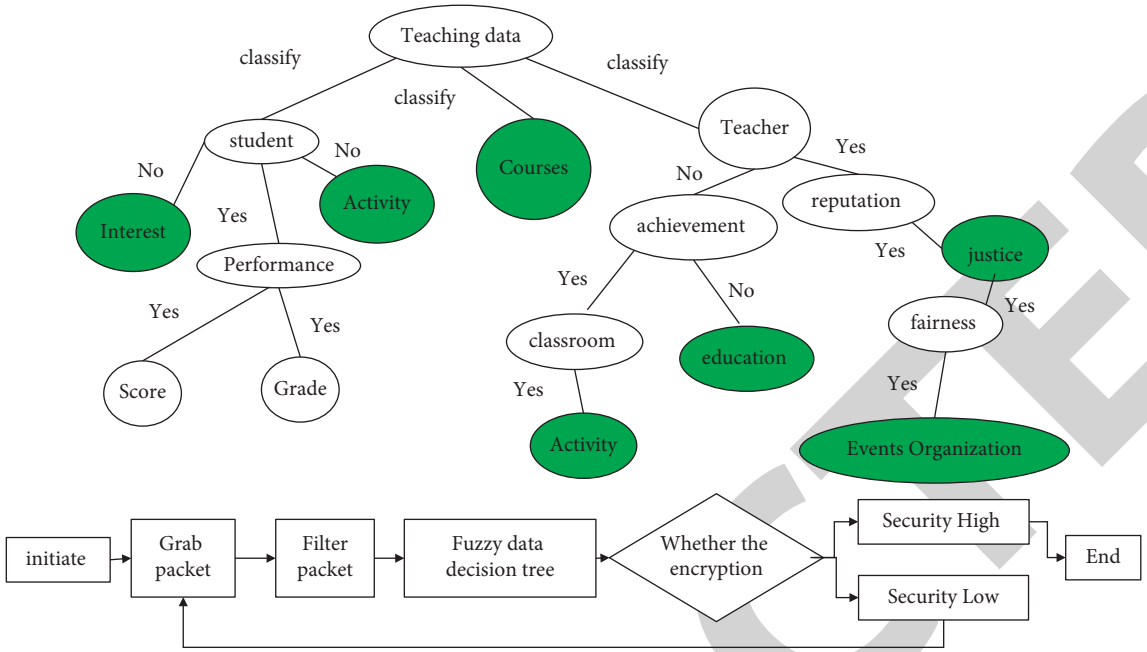


Figure 1: Decision tree model generated based on fuzzy algorithm.

Table 1: Training dataset.

| Total words | Families | PETS | Average sentence length | Level |
|---|---|---|---|---|
| 0.48 | 0.26 | 0.25 | 0.01 | Junior-middle |
| 0.39 | 0.35 | 0.24 | 0.02 | Junior-high |
| 0.38 | 0.31 | 0.27 | 0.03 | Senior-middle |
| 0.41 | 0.37 | 0.17 | 0.04 | Senior-high |
| 0.42 | 0.39 | 0.14 | 0.05 | College |

the missing values are small, it is found by statistics that most of these missing data belong to students' scores of the same grade, and in order to avoid the impact on data balance caused by directly deleting the data, this paper uses the random forest model to predict the missing entrance English scores, and its prediction process is as follows. First, appropriate features are selected as the input variables. The improvement of pupils' English skills is continuous and stable, and there is no sudden increase or decrease in the short term, and the correlation between "English one point" and "English two points" and "entrance English ability" is strong. In addition, according to the analysis in Chapter 3, gender and semester have a strong influence on whether students can pass the level 4 exam, so gender and semester are also used as input features for predicting "entrance English score." Thus, in this paper, four features were selected to predict the missing values of entry English scores.

In the $K$-nearest neighbor algorithm, only numerical features can be used to calculate the distance between samples, so it is necessary to convert the categorical features such as gender, semester, college, and major. The most common method of conversion is to convert the value domain of each categorical feature from a multicolumn representation to a multicolumn representation with only true values. In this paper, we take semester as an example to illustrate, and there are three values for semester, 2, 3, and 4, as shown in Table 2.

In order to eliminate the adverse effects of order of magnitude on prediction, this paper uses a standardization method to normalize the scores to 110 points. In order to eliminate the negative effect of order of magnitude on the prediction, the normalization method is used to normalize the score so that the different ranges of values are all in the interval of [−1, 1]. The data were normalized to [−1, 1]. The standardized data follow a normal distribution with a mean of 0 and a variance of 1. If the original score is greater than the mean, the standardized score will be greater than 0, which is a positive value, and if the original score is less than the mean, the standardized score will be less than 0, which is a negative value.

The predictive classification with the $K$-next algorithm requires the determination of the corresponding number of the nearest neighbors $K$ and the selection of important features. To improve classification accuracy and make classification performance more stable, this document uses an iterative approach to select $K$ and important features. As shown in Figure 2, first use all features to the next neighbor to reduce the feature to select important features according to the prediction if the $K$ value does not change. Finally, an important feature is used as input variable to determine the corresponding $K$ value.

Following the cleaning and processing of the data, 2674 data elements and the results of all English IV tests contained in the data were identified. Features have been removed. Although the missing values are small, it is found by statistics that most of these missing data belong to students' scores of the same grade, and in order to avoid the impact on

Table 2: Variable conversion table.

| Student ID | Semester | Significance | Score |
|---|---|---|---|
| 1 | 1 | 0.2541 | 100 |
| 1 | 2 | 0.1984 | 110 |
| 2 | 1 | 0.1271 | 100 |
| 3 | 1 | 0.0447 | 110 |
| 3 | 2 | 0.1361 | 100 |

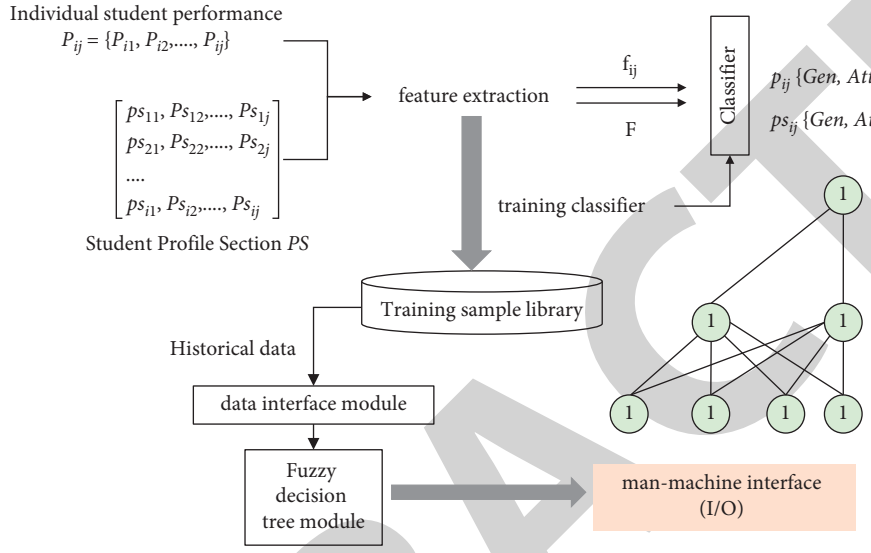


Figure 2: Generation of dataset numbers.

data balance caused by directly deleting the data, this paper uses the random forest model to predict the missing entrance English scores, and its prediction process is as follows. The results of measured pollutant concentration changes are as shown in Table 3.

*2.4. Evaluation of Algorithm Improvements.* The commonly used distances in $K$-nearest neighbor algorithm are Minkowski distance, Euclidean distance, absolute distance, etc. Suppose $X_1, X_2, \ldots, X_n$ are observation points, then the abovementioned distances between $X_i$ and $X_j$ are defined as follows:

Minkowski distance:

$$\mathrm{Dis}(x_i, x_j) = \left( \sum_{i=1}^{n} \left( X_i - X_j \right)^q \right)^{1/q}, \quad (1)$$

European distance:

$$\mathrm{Dis}(x_i, x_j) = \left( \sum_{i=1}^{n} \left( X_i - X_j \right)^2 \right)^{1/2}. \quad (2)$$

The absolute distance is the sum of the absolute values of the differences of the $m$ components of the two sample points and is a special case of the Minkowski distance $p = 1$. It should be noted that the abovementioned distances have a common shortcoming, they are all easily affected by the size of each component, that is, when each component is a quantity of different nature, the magnitude of the distance is highly dependent on the units of the component, and the component of larger magnitude will contribute more to it than the component of smaller magnitude. Therefore, in order to make each component have equal influence on the distance, the attribute values need to be processed before calculating the distance, and the two common processing methods for continuous variables are the normalization method and the polarization method:

$$X = \frac{X_i - X_j}{\left( \sum_{i=1}^{n} \left( X_i - X_j \right)^2 \right)}, \qquad X = \frac{X_i - \overline{X}_j}{\sigma}, \quad (3)$$

where $X_i$ is the mean value of $X$ and $\sigma$ is the standard deviation of $X$. For the categorical variables, the following functions need to be defined:

$$Z\left(u_i(x), v_i(x)\right) = \begin{cases} 0, & u = v, \\ 1, & u \neq v, \end{cases} \quad (4)$$

so that it can be used to compare the magnitude of the $i$th attribute value of a pair of records, where $u_i$ and $v_i$ are categorical values, and the $i$th term in the distance formula

Table 3: Measured pollutant concentration changes.

| Students | The input results | Gender | College | Professional |
|---|---|---|---|---|
| 1 | A++ | Male | Language | English |
| 2 | A+ | Female | Language | English |
| 3 | A++ | Male | / | / |
| 4 | A | Female | / | / |
| 5 | B | Male | / | / |

can be replaced by the $Z(v_i, u_i)$ function when calculating the Euclidean distance containing the categorical variables.

The growth process of decision tree is a continuous grouping process of the training sample set, and the core algorithm of decision tree growth is to determine the branching criterion of the decision tree. The branching criterion involves two problems: first, how to select a current best grouping feature from many input features; second, how to find an optimal splitting point from many values of the grouping features. Different decision tree algorithms use different strategies, and the common algorithms for decision trees are ID3 and C4.5. These two algorithms use the information gain and information gain ratio as the strategy for the selection of the most valuable feature, respectively.

Information entropy is a measure of the uncertainty of a random variable $X$. Its information entropy is defined as

$$Q(x) = \frac{\sum_{i=1}^{n} p(X = x_i)}{\log_2 (p(x))}, \tag{5}$$

where $x_1, x_2, \ldots, x_n$ are the possible values of the random variable $X$ and specify $0 \log_2 0 = 0$. The greater the information entropy, the greater the uncertainty of the random variable. The conditional entropy $Q(Y[X])$ denotes the uncertainty of the random variable $Y$ conditional on the random variable $X$ and is defined as

$$Q(Y[X]) = \frac{\sum_{i=1}^{n} p(X = x_i)}{q(x)}. \tag{6}$$

The information gain represents the degree of information uncertainty reduction of class $y$ by learning information about feature $x$. It is defined as

$$p(x, y) = \frac{\sum_{i}^{n} q_i(x) + p_i(x)}{q(x)}. \tag{7}$$

Python programming is used to obtain accuracy, pass, reproduction, and $F$-value plots for the abovementioned eight scenarios. The accuracy reflects the ability of classification models (algorithms) to correctly classify predictions. Accuracy reflects the number of actual positive cases from predicted positive cases, while reproduction rates reflect the ability to correctly classify and predict many positive cases. In precision graphics, if $K$ is small, the higher the $K$, the higher the accuracy of the eight cases. Scenario 1 has the highest input function, but the accuracy diagram has the least predicted accuracy and only three features, and for scenario 8, the input function for scenario 8, which is the English input score, has the lowest function, But on average, scenario 11 has the highest prediction accuracy, and the error is relatively small. This indicates that the accuracy and reproducibility of the predictions are more influenced by the next neighbor number $K$ and the amount of feature. Of the eight cases, case 8 has the lowest fit, reproduction, and $F$-values, but case 8 has a relatively high fit, re-reproduction, and $F$-values. We have only adopted entrance English score and English score 1, taking into account the purpose of this paper, to predict whether the fourth-year student will pass the examination with high accuracy. The exam notes of the fourth-grade students are predicted from three characteristics: English grade 1 and English grade 2. The accuracy rate is as shown in Figure 3.

## 3. Results and Analysis

The data adoption aspect is based on the specificity of the data for data compilation. On the one hand, some students passed the university English level 4 exam for the first time; although these students did not take all of their university English courses, it would be less reliable and unreasonable to still use future English grades in analyzing the data and predicting the already passed university English level 4 grades; on the other hand, some students may need to take several university English level 4 exams to pass. On the other hand, some students may need to take the college English level 4 exam several times to pass it, and their time of taking the level 4 exam is far from the time of studying college English classes, which is not too valuable to study the prediction of college English level 4 exam. Considering the above two aspects, this paper only uses the two most recent college English scores from the level 4 exam. For students who repeatedly took the Level 4 exam, because they were in different semesters when they took the exam, they had different college English scores from the two closest times to level 4 so that the data for students who repeatedly took the exam are considered as independent data in this paper.

Figure 4 shows that classification accuracy, fit accuracy, reproducibility, and $F$-values in case 8 basically reach their maximum $K = 15$, with relatively small variations and high algorithm accuracy and stability. Since $K = 15$ is reliable, the number of nearest neighbors can be determined as $K = 15$. This paper therefore contains the number of nearest neighbors $K = 15$, "English entry grade," "English grade 1," and "English grade 2." By default, the $K$-adjacent algorithm contributes to the prediction results for the next neighborhood of $K$ for classification purposes. However, the closer a classified sample approaches a known classified sample, the higher the characteristics of a known classified sample are. The central idea of weighting is to define weights as a nonlinear function of the distance between known and classified samples. The closer the distance, the greater the weight and the greater the impact on the results of the classification prediction.

In this article, input features are still used for classification purposes with weighted $K$ next-door neighbors. The next neighbor number $K$ is selected. In the same way as the $K$-adjacent algorithm, the value of $K$ is determined by running 18 odd 1, 3, . . ., 35s according to the same principle with high classification accuracy. The evaluation indices of the weighted $K$-next algorithm, the weighted $K$-next

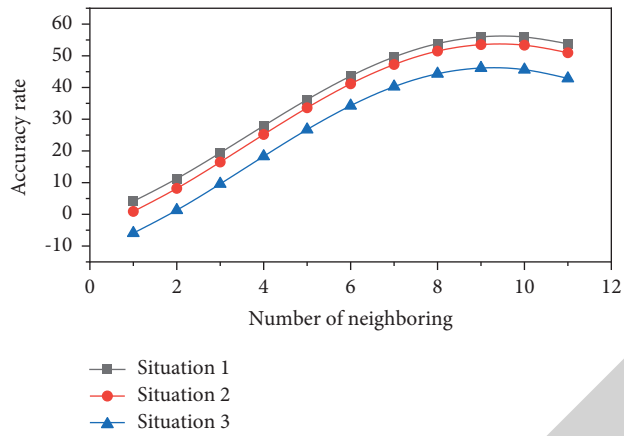


Figure 3: Accuracy rate.

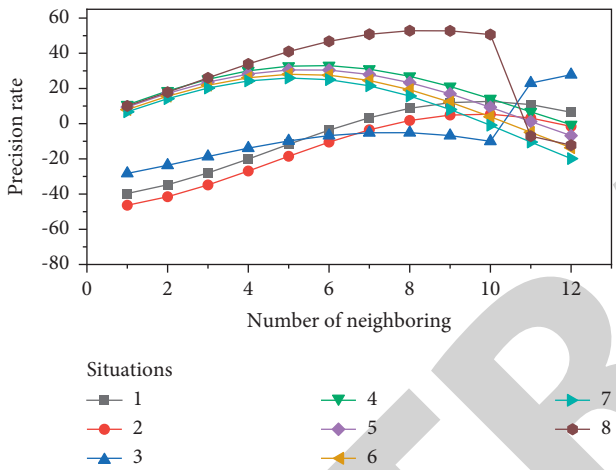


Figure 4: Classification precision rate.

algorithm, and the *K*-next algorithm are shown in the diagram by Python programming for comparison and analysis. In Figure 5, the horizontal coordinates of each subplot are the number of nearest neighbors *K* and the vertical coordinates are the scores of each evaluation index.

The classification curve of the weighted *K*-next algorithm is flatter than the classification curve of *K*-next neighbor algorithm. In the reproduced curve, the weighted *K*-next neighbor algorithm has a higher reproducibility than the *K*-next neighbor algorithm, and the curve variation is relatively small. The weighted K-Nachbar algorithm reproduces more than the K-next algorithm, and the curve variation is relatively small. The overall accuracy of the weighted *K*-next algorithm is not as high as that of the *K*-next algorithm, but more stable. Therefore, taking into account the accuracy and stability of the classification, you can use the weighted *K*-method algorithm for the next neighbor to predict the results of a level 4 examination.

As shown in Figure 6, the *K*-nearest neighbor algorithm has a low classification effect and takes longer to complete the classification prediction when the dataset is large. Split the entire dataset based on certain features into several feature subsets, and let the samples you classify find the most similar neighbors in the corresponding feature subset based on these features. This significantly reduces search time, improves classification, and predicts efficiency.

When comparing the calculation time of the three algorithms, the efficiency of the weighted *K*-neighborhood algorithm and the next *K*-neighborhood algorithm is basically similar, and there is no significant change. The average calculation time for the split-weighted *K*-next algorithm is 6, 17 seconds shorter. As the classification time with 11.39 is significantly shorter than that of the next *K*-category at 11.39 and the classification efficiency 118%, you can therefore select the partition-weighted *K*-next category if you want to predict the test of a fourth-grade student quickly. As shown in Figure 7, the partition-weighted *K*-nearest neighbor algorithm has higher classification accuracy and better classification stability at the number of nearest neighbors $K = 15$. There are many reasons why students cannot graduate normally, but the number of students failing and repeating too many times is an important reason. If the rules between the number of students not graduating normally and the number of early make-up exams and repeats can be tapped, students can be warned in time so that they can avoid excessive repeats, which may affect their studies and graduation. To this end, this paper uses a decision tree model to predict students' graduation using the number of required make-up exams in the first four semesters and the number of required repeats in the second through fifth semesters as input variables and the graduation situation (normal graduation and failure to graduate normally) as the predicted outcome. Since the dataset of students processed by balancing is not large, and there are only 8 input features, all of which are integers, a decision tree with maximum depth max_depth = 3 and feature selection criterion "entropy" is chosen. The specific classification method and procedure are the same as above, and the dataset is divided into training and testing sets according to 7 : 3, one for training the decision tree and one for graduation prediction and evaluation of the classification effect of the decision tree.

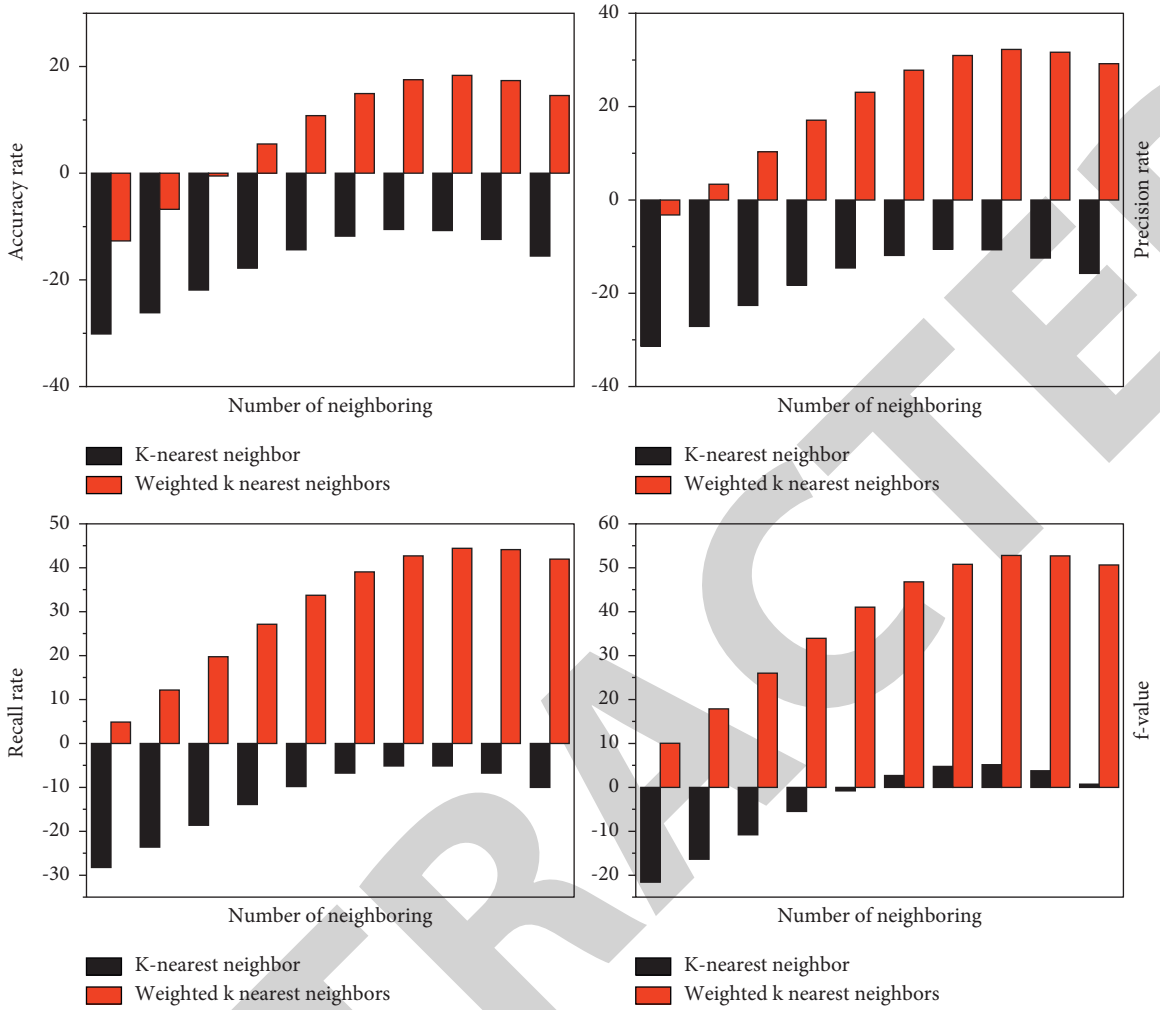


Figure 5: Comparison of weighted *K*-nearest neighbor and *K*-nearest neighbor classification results.

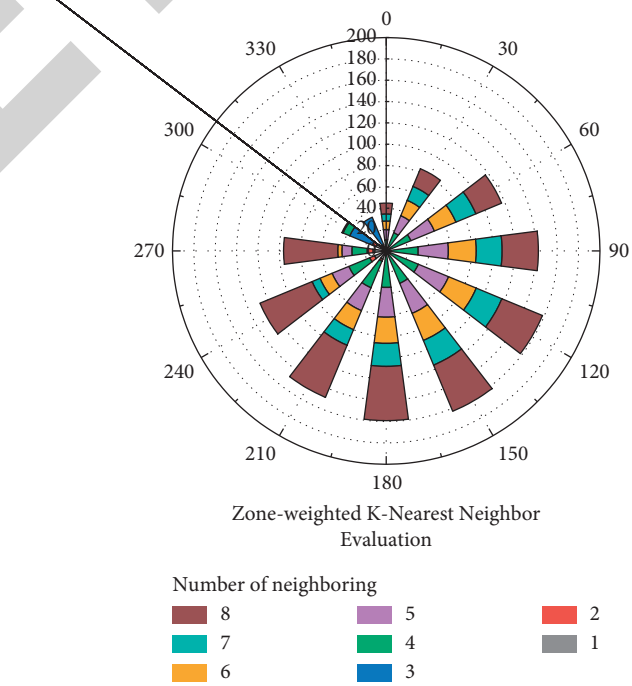


Figure 6: Zone-weighted *K*-nearest neighbor evaluation index curve.

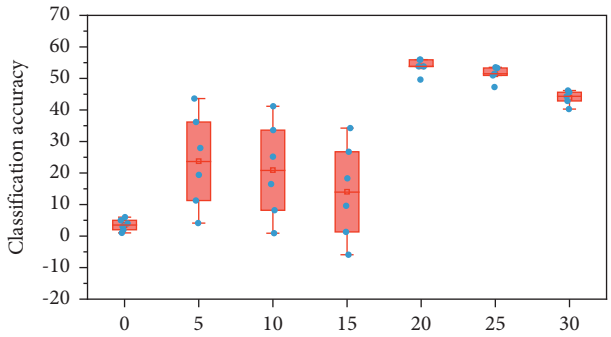


Figure 7: Classification accuracy and stability.

## 4. Conclusion

In order to improve the English teaching model and predict the results of the college English IV exam, this paper firstly preprocesses the data to fill in the missing values of the entrance English scores, factorizes the categorical variables, and normalizes the numerical variables; then, it selects the prediction input features and the number of nearest neighbors $K$ based on various indicators and establishes the $K$-nearest neighbor classification model and the weighted $K$-neighbor classification model.

The $K$-nearest neighbor algorithm and the split-weighted $K$-nearest neighbor algorithm are applied to the classification and prediction of university English IV test results, and by filtering the input using statistical techniques, relevant factors that affect the English IV test results are investigated. The neighborhood algorithm can obtain enough features for classification and prediction. The neighborhood algorithm and its classification efficiency are drastically improved, and the classification time is improved by 118%.

Finally, to remedy the low efficiency deficiencies of the $K$ next-door neighbor classification model, a split-weighted $K$ next-door neighbor model is developed to achieve a quick classification of the university of English IV test results. This paper uses the next neighbor. Algorithms are used to predict whether college students will pass the fourth-grade exam.

Also, in the future, the methods we used to predict students passing the exam more effectively would be the research direction and focus on the decision tree algorithm with fuzzy theory; the adaptability of the algorithm to different datasets is supposed to be enhanced.

## Data Availability

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

## Consent

Informed consent was obtained from all individual participants included in the study references.

## Conflicts of Interest

The author declares that there are no conflicts of interest.

## References

[1] Y.-W. Li and K. Cao, "Establishment and application of intelligent city building information model based on BP neural network model," *Computer Communications*, vol. 153, pp. 382–389, 2020.

[2] M. Hasanipanah, W. Zhang, D. J. Armaghani, and H. Nikafshan Rad, "The potential application of a new intelligent based approach in predicting the tensile strength of rock," *IEEE Access*, vol. 8, pp. 57148–57157, 2020.

[3] X. Huang, D. Zhuang, and W. Li, "Research on application of intelligent assistant decision-making in the five-level triage of pediatric emergency department," *Chinese Pediatric Emergency Medicine*, vol. 26, no. 12, pp. 907–911, 2019.

[4] C. Liu, Y. Zhou, Y. Cen, and D. Lin, "Integrated application in intelligent production and logistics management: technical architectures concepts and business model analyses for the customised facial masks manufacturing," *International Journal of Computer Integrated Manufacturing*, vol. 32, no. 4-5, pp. 522–532, 2019.

[5] X. Wang, Y. Zhang, and M. Mingming, "Design and application of intelligent cloud disease-diagnostic system based on non-coding amino acid mass spectrometry," *Biomedical Journal of Scientific and Technical Research*, vol. 24, no. 1, pp. 17990–17993, 2020.

[6] H. Zhu and L. Gu, "Design and application of intelligent information system for comprehensive management of obstetrics and gynecology health care," *Journal of Medical Imaging and Health Informatics*, vol. 10, no. 8, pp. 1834–1840, 2020.

[7] C.-C. Kuo and W.-J. Chen, "Development and application of intelligent monitoring system for rapid tooling applied in low-pressure injection molding," *International Journal of Advanced Manufacturing Technology*, vol. 111, pp. 1–15, 2020.

[8] Y. Kong, Y. Jiang, and J. Lou, "Terminal computing for sylvester equations solving with application to intelligent control of redundant manipulators," *Neurocomputing*, vol. 335, pp. 119–130, 2019.

[9] Z. An, S. Li, X. Jiang, Y. Xin, and J. Wang, "Adaptive cross-domain feature extraction method and its application on machinery intelligent fault diagnosis under different working conditions," *IEEE Access*, vol. 8, pp. 535–546, 2020.

[10] S. Li, X. Ma, and C. Yang, "A combined thermal power plant investment decision-making model based on intelligent fuzzy grey model and ito stochastic process and its application," *Energy*, vol. 159, pp. 1102–1117, 2018.

[11] K.-S. Chen, "Fuzzy testing decision-making model for intelligent manufacturing process with taguchi capability index," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 2, pp. 2129–2139, 2020.

[12] A. Andrew and S. Kumanan, "Development of an intelligent decision making tool for maintenance planning using fuzzy logic and dynamic scheduling," *International Journal of Information Technology*, vol. 12, no. 1, pp. 27–36, 2020.

[13] M. Lu, G. Wei, F. E. Alsaadi, T. Hayat, and A. Alsaedi, "IOS press has retracted the following publication from its online content," *Journal of Intelligent & Fuzzy Systems*, vol. 37, no. 6, p. 8613, 2019.

[14] S. Pariserum Perumal, G. Sannasi, and K. Arputharaj, "An intelligent fuzzy rule-based e-learning recommendation system for dynamic user interests," *The Journal of Supercomputing*, vol. 75, no. 8, pp. 5145–5160, 2019.

[15] A. Boubellouta, F. Zouari, and A. Boulkroune, "Intelligent fuzzy controller for chaos synchronization of uncertain

fractional-order chaotic systems with input nonlinearities," *International Journal of General Systems*, vol. 48, no. 3, pp. 211–234, 2019.

[16] B. Xu and F. Sun, "Composite intelligent learning control of strict-feedback systems with disturbance," *IEEE Transactions on Cybernetics*, vol. 48, no. 2, pp. 730–741, 2018.

[17] G. Wei, M. Lu, F. E. Alsaadi, T. Hayat, and A. Alsaedi, "Pythagorean 2-tuple linguistic aggregation operators in multiple attribute decision making," *Journal of Intelligent and Fuzzy Systems*, vol. 37, no. 6, p. 8617, 2019.

[18] F. Yang, Y. Qiao, W. Wei et al., "Ddtree: a hybrid deep learning model for real-time waterway depth prediction and smart navigation," *Applied Sciences*, vol. 10, no. 8, p. 2770, 2020.

[19] B. Wang and L. L. Chen, "New results on fuzzy synchronization for a kind of disturbed memristive chaotic system," *Complexity*, vol. 2018, Article ID 3079108, 9 pages, 2018.

[20] S. Yang, T. Gao, J. Wang, B. Deng, B. Lansdell, and B. Linares-Barranco, "Efficient spike-driven learning with dendritic event-based processing," *Frontiers in Neuroscience*, vol. 15, Article ID 601109, 2021.

[21] F. Orujov, R. Maskeliūnas, R. Damaševičius, and W. Wei, "Fuzzy based image edge detection algorithm for blood vessel detection in retinal images," *Applied Soft Computing*, vol. 94, Article ID 106452, 2020.

[22] K. Sim, J. Yang, W. Lu, and X. Gao, "Blind stereoscopic image quality evaluator based on binocular semantic and quality channels," *IEEE Transactions on Multimedia*, 2021.

[23] G. Dong, W. Wei, X. Xia, M. Woźniak, and R. Damaševičius, "Safety risk assessment of a pb-zn mine based on fuzzy-grey correlation analysis," *Electronics*, vol. 9, no. 1, p. 130, 2020.