J. COMBIN. MATH. COMBIN. COMPUT. 127b (2025) 1013--1034

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Decision Tree-Based Modeling in Mental Health Early Warning System for Higher Education Students

Ying Jin[1, ✉]

[1] Art Department, Fushun Vocational Technology College, Fushun, Liaoning, 113122, China

ABSTRACT

Students' mental health problems are increasingly becoming an important part of the educational and teaching process in colleges and universities. In this paper, we collect students' psychological data through the students' mental health early warning system and preprocess the data through data cleaning and other data. The features of the processed mental health data are extracted using Global Chaos Bat Based Algorithm (GCBA). Construct a mental health early warning system for college students and build a decision tree model into the system for categorizing students' mental health status. The performance of the decision tree model in this paper is verified by evaluating the finger with other models and comparing the actual classification prediction results, constructing the decision tree model with the psychological condition of interpersonal relationship of college students as an example, and conducting the visualization analysis of the decision tree. Independent sample t-test is conducted on three measures such as using the mental health early warning system constructed in this paper, and according to the results, the application of the system in this paper highlights the role of the enhancement of the level of students' mental health and the significant improvement of depression and other psychological conditions.

Keywords: decision tree, GCBA, categorical prediction, independent sample test, mental health early warning

1. Introduction

Psychological problems are a common phenomenon, which may be caused by childhood living environment, excessive expectations and evaluations, worries about the future and life, conflicts in interpersonal relationships, excessive competition and pressure, and sudden calamities [1-4]. Therefore, research on mental health problems of Chinese college students has gradually increased.

The emergence of mental health problems among college students is mainly closely related to a variety of factors. First of all, academic pressure is one of the important factors leading to the mental

✉ Corresponding author.
E-mail address: 15242746595@163.com (Y. Jin).

Received 15 July 2024; Revised 05 October 2024; Accepted 15 February 2025; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-057

health problems of college students. In recent years, with the intensification of employment competition, many college students are facing the double pressure of employment pressure and academic burden, which leads to a decrease in their psychological tolerance [5-6]. Secondly, interpersonal relationship problems are also important factors affecting the psychological health of college students. During their college years, students need to adapt to new social environments and face various pressures and expectations from classmates, teachers and family members, which may lead to loneliness, social anxiety, depression and other psychological problems [7-10]. In addition, social changes and technological development have also had a certain impact on college students' mental health. For example, the popularization of the Internet has made college students more susceptible to virtual socialization, cyber violence, fraud, etc., increasing the incidence of mental health problems [11-12]. Further, individual physical health conditions, changes in emotional state, bullying on campus, and being assaulted may also cause psychological problems, and in severe cases, there may be self-injurious or suicidal behaviors [13-16]. Therefore, conducting in-depth research and providing scientific and effective mental health services are crucial to maintaining the physical and mental health of college students. And the establishment of an early warning system for students' psychological problems in colleges and universities is an extremely effective method.

The early warning system for students' psychological problems aims to popularize psychological knowledge among students, set up early warning mechanisms for psychological problems, analyze and identify students' behaviors, discover students' psychological problems in time, and strengthen psychological crisis prevention, intervention, and health education [17-20]. The early warning system can detect potential psychological problems early, provide students with necessary psychological education, psychological counseling and psychotherapy and other help in time, and reduce the burden of students and their parents. At the same time, some serious psychological problems can be avoided to a certain extent to ensure that students can successfully complete their studies and move towards a successful life.

As the intersection of two circles of school and society, college students' mental health has a great impact on their personal development. The freedom, empathy and privacy of today's social media make most college students like to express their opinions and inner feelings on the platform. Especially, when people with depressive tendencies or other psychological problems have no one to talk to and no one to understand, social media becomes their secret garden. Therefore, by analyzing and monitoring social media data, it is possible to understand the psychological state of users and whether they are depressed or suicidal tendencies [21]. Therefore, literature [22] created a suicide early warning system based on Google Trends and media coverage data. And literature [23] integrated and analyzed multiple types of information such as social media and bad moods to design a mental health early warning service platform for college students. When it comes to suicide, it is the biggest worry of family members of depressed patients. Therefore, early warning signals or systems for the prediction of mental illnesses such as depression can be considered to alleviate their concerns. Some literature [24] mentioned that early warning signals of critical deceleration phenomenon have the potential to predict mental illnesses such as depression. And literature [25] designed a machine learning model to predict suicide among children and adolescents through electronic health records. And literature [26] built an early warning system for high-risk suicidal patients directly through machine learning algorithms and deep neural networks on electronic health record data. This study deepens the science of early warning. The study mentioned machine learning algorithms, which are widely used with deep learning, data mining, and other big data analytics to monitor [27], predict [28], analyze and evaluate [29], and prevent and warn [30] individuals' mental health conditions. Among them, fuzzy clustering and decision tree association law are data mining and machine learning algorithms, which are also applied to the prediction and analysis of mental health. For example, literature [31] fuzzy clustering algorithm analyzes the mental health of college students, and thus proposes a method for early warning of psychological crisis. And literature [32] used a heuristic fuzzy clustering algorithm to analyze college students' stress, mental health, and academic performance to predict students' mental health status. Literature [33] established a decision tree model for depression in college students and utilized an association rule algorithm to data mine the links between psychological parameters to assess students' mental health status. In addition, literature [34] designed an individual psychological crisis early warning instrument, which can effectively screen high-risk psychological crisis populations and realize the dynamic monitoring of mental health, contributing to the early warning of mental health.

In this paper, the mental health early warning system is utilized to collect the mental health data of college and university students over the age of 18, and the Big Five Personality Inventory (TIPI) and Depression-Anxiety-Stress Scale (DASS-42) are used to obtain information about the students' psychological conditions. In order to prevent the interference of abnormal data on the final results, this paper carries out preprocessing operations such as data screening, data cleaning, data transformation and normalization on the collected students' mental health data. The features of students' psychological conditions exhibited in the mental health data are extracted using the global chaotic bat algorithm, and an improved logical mapping method is used to ensure that the bat population explores the entire solution space and enriches the population diversity. The OM-70 early warning data recorder and AT24C16N memory chip are used as the hardware facilities of the mental health early warning system in this paper, and the decision tree model based on C4.5 algorithm is built in to construct the mental health early warning system for college students, which includes the functions of basic information management, psychological assessment, and assessment result management. Taking the accuracy rate and AUC value as evaluation indexes, we compare and analyze the index scores of this paper's model with those of LR, LightGBM, Random Forest and other models, as well as the performance of classification prediction experiments, so as to highlight the good performance of this paper's decision tree model. The decision tree model is constructed for the psychological condition of college students' interpersonal relationships, and the information gain rate of each split attribute in the sample data set is calculated. According to the classification results of the decision tree, the system constructed by this paper issues an early warning to students who may have interpersonal relationship sensitivity symptoms. In order to deeply analyze the application effect of this paper's system, this paper's system, the traditional mental health early warning system, and no psychological warning as a student mental health intervention, according to the results of independent samples t-test and one-way ANOVA to demonstrate that this paper's decision-tree-model-based mental health early warning system for college students is significant for the students' mental health.

2. Psychological data processing and feature selection

2.1. Psychological data acquisition

This experimental study was conducted on college and university students over the age of 18, and the dataset used was obtained from a questionnaire designed to be recycled by the Mental Health Early Warning System over a period of six months, with a deadline of January 2023, and a total of 1,693 pieces of completed data were eventually recycled. The dataset mainly consists of three parts: personality trait information, demographic information and psychological state information, of which the psychological state information was obtained through the Big Five Personality Inventory (TIPI) and the Depression-Anxiety-Stress Scale (DASS-42).

2.2. Pre-processing of psychological data

2.2.1. Pre-processing of data. When questionnaires are assessed, there are always various reasons that affect the integrity of the final scale data obtained. In this project, the following steps are taken to preprocess the raw data for the data with anomalies:

1) Data screening. When analyzing massive data, it is generally not necessary to use every attribute of the original dataset. For example, in the study of this paper, the demographic information filled in by the testers such as age and gender are high-quality and valid data, while the completion time, login location, and browser type recorded in the background of the network are some of the data that are not useful for the study.

2) Data Cleaning. In this paper, the abnormal data collected from the demographic information table and the two large number of tables are processed as follows: firstly, if a piece of raw data is missing more than three relevant attributes, it is deleted. Secondly, raw data with missing type of numerical type are filled with mean value, and raw data with missing type of string type are filled with plural. After the above processing, the final obtained experimental valid data 1485.

3) Data integration. Since this paper mainly uses the demographic information and personality trait information of the psychological tester base as input, it is necessary to perform data integration on them, examples are as follows:

(1) Basic demographic information is recorded as A, e.g., (gender, age, whether living in town, habitual right hand) is recorded as (a,b,c,d).

(2) Personality trait information is recorded as B, e.g., record (Extraversion, Neuroticism, Openness, Likability, Dutifulness) as ($\alpha_1,\beta_1,\gamma_1,\eta_1,\mu_1$).

Then one complete piece of input data from the psychometric tester is noted as C, at this point, $A=(a,b,c,d)$, $B=\alpha_1,\beta_1,\gamma_1,\eta_1,\mu_1$, $C=(A,B)$.

According to the above description, one complete piece of data comes from the dataset integrating demographic information, personality trait information, and psychological state information, the details of which are shown in Table 1, where E_1, E_2, and E_3 indicate the presence of depression, the presence of anxiety, and the presence of stress.

Table 1. Integrated data

Serial number	0001	0002	0003	0004	0005	0006	0007
a	Male	Female	Male	Male	Female	Male	Male
b	22	24	19	21	23	20	19
c	Yes	Yes	No	Yes	Yes	No	Yes
d	Yes	Yes	Yes	Yes	Yes	Yes	Yes
…	…	…	…	…	…	…	…
α_1	6.4	5.3	4.1	4.2	4.8	3.3	6.7
β_1	4.4	2.8	6.2	3.7	6.1	6.3	2.7
γ_1	6.6	5.7	3.8	4.1	3.2	4.8	4.2
η_1	6.3	6.1	4.7	3.6	4.8	4.3	4.7
μ_1	4.3	5.2	6.1	6.7	5.8	6.2	4.3
E_1	9	3	11	5	8	7	2
E_2	12	8	7	3	4	5	7
E_3	10	8	12	9	6	4	8

After the data preprocessing process, the study obtained 1485 complete and valid experimental data collected by January 2023. The dataset consisted of demographic information about the testers, data from the TIPI Ten-Item Personality Questionnaire, and data from the DASS-42 scale. During the design of the questionnaire, as much as possible, basic information about the testers and information about variables that may affect psychological changes in the population was collected.

2.2.2. Normalization of data. Normalization of data, also known as normalization, is an operation that involves scaling the data range to a specific range, usually [0,1] or [-1,1]. This processing method can make different features have the same feature scale, avoiding problems due to different scales of the data itself. Currently there are a variety of normalization methods, the commonly used method is the minimum - maximum (Min-max) normalization method, the conversion function is as follows:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x' is the scaled value, x is the original data value, x_{min} is the minimum value of the data and x_{max} is the maximum value of the main sentence. The Min-max (Min-max) normalization method scales the data to a formulated range by subtracting the minimum value from each data point and dividing by the difference between the maximum and minimum values.

2.3. Feature selection based on global chaotic bat algorithm

The Global Chaotic Bat Algorithm, as a heuristic algorithm [35], applies feature selection in the following steps: first, a chaotic mapping method is introduced in the initial phase to ensure that the bat population explores the entire solution space and enriches the population diversity. In addition, a fitness function based on accuracy and feature subset length is used to calculate the score of the feature subset after each update. The Global Chaotic Bat Algorithm (GCBA) then selects the highest scoring feature subset by calculating the score, thus effectively eliminating irrelevant and redundant features from all feature variables.

In this paper, an improved logistic mapping method for population initialization is used, which is mathematically modeled as:

$$u_i^w = |1 - 2 \times (u_i^{w-1})^2| \tag{2}$$

where $u_i^w (i = 1, 2, \cdots, n, w = 1, 2, \cdots m)(u_i^w \in [0,1])$ is a chaotic variable, n represents the number of bat populations, and w represents the initial population dimension. Next, the inverse mapping of u_i^w can be performed to obtain the position of individual bats in the solution space x_i^w, and x_i^w is calculated as:

$$x_i^w = k_i + (p_i - k_i) u_i^w \tag{3}$$

where p_i and k_i are the maximum and minimum values within the range of the variable, respectively. When the position of individual bats is updated, the local optimal position of individual bats and the global optimal position of the population are recorded. The position of the ith bat at the $t+1$rd iteration can be calculated by the following equation:

$$x_i^{t+1} = x_i^t + v_i^{t+1} F_1 c_1 (O_i - x_i^t) + F_2 c_2 (O_g - x_i^t) \tag{4}$$

where $F_1 = F_2 = 1.496$, O_i are the locally optimal positions of the ird bat and O_g is the globally optimal position of the bat population. c_1 and c_2 are two random values within [0, 1].

When the global chaotic bat algorithm is applied to demographics and feature selection for personality trait attributes, the bat population is initialized with a matrix of size $n \times m$. Where n is the number of bat populations and m is the number of traits. The binary discretization of bat locations is then performed using the transfer equation. The transfer equation is:

$$S(x_i^w(t)) = \frac{1}{1 + e^{-x_i^w(t)}} \tag{5}$$

where $x_i^w(t)$ represents the position of the ith bat in the wrd dimension at the tnd iteration. The update equation for the position of individual bats is:

$$x_i^w(t) = \begin{cases} 0, rand < S(x_i^w(t)) \\ 1, rand \geq S(x_i^w(t)) \end{cases} \tag{6}$$

where rand represents a random number within [0, 1]. When the position of the ind bat in the wrd dimension is 0 at the tst iteration, the bat is not selected; when the position of the ith bat in the wth dimension is 1, the bat is selected.

After feature extraction of mental health data by the above model, the optimal features are selected sequentially. Next, these features are combined and scored, and the optimal feature combination is selected. Figure 1 shows the optimal feature combination scoring graph obtained through continuous iteration of the global chaotic bat algorithm, where the horizontal axis represents the number of feature combinations and the vertical axis represents the value of the evaluation function. As can be seen from the figure, a total of five psychological state features of students' mental health, obsessive-compulsive, anxiety, interpersonal relationship and depression are extracted in this paper.

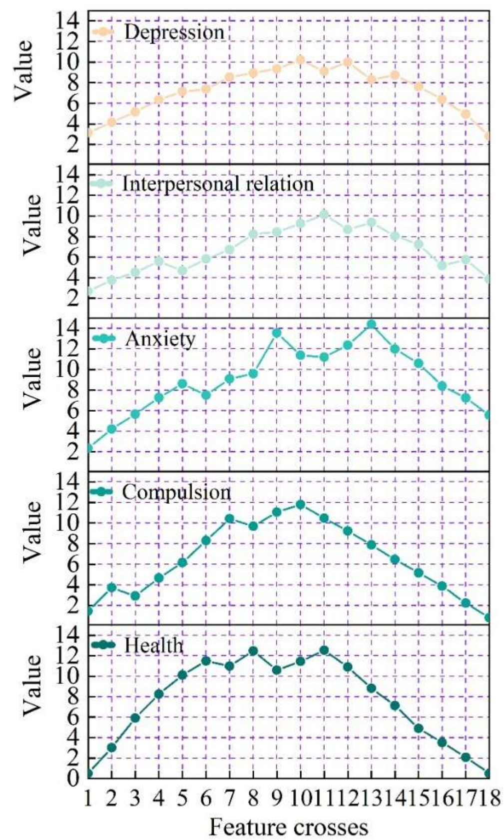

Fig. 1. Global chaotic bat algorithm iteration evaluation

3. Mental health early warning system based on decision tree modeling

With the surge of learning pressure, the emotional attitudes and learning psychology of students in colleges and universities have gradually undergone a transformation. In order to ensure the mental health of students and solve the psychological conflict problems faced by students in a timely manner, it is necessary to actively carry out mental health early warning. The use of conventional mental health early warning system for early warning often exists more influencing factors, resulting in low prediction accuracy, the need to design a new mental health early warning system for college students. Therefore, this paper designs a new mental health early warning system for college students based on the decision tree model.

3.1. Hardware Design

3.1.1. OM-70 early warning data logger. In the process of mental health early warning, often involves a variety of early warning data and early warning information, in order to improve the comprehensive performance of the crisis early warning system, the need to centralize the processing of these early warning information, this paper selects the OM-70 early warning data recorder real-time recording of early warning data. The early warning data logger is suitable for dealing with a variety of types of information, the internal set of data logging library to store early warning data, once the relevant warning information, will immediately start the transmission channel, the data will be transmitted to the early warning center to complete the early warning. The warning data logger comes with a USB interface, which can quickly transfer data and record the warning limit values, and the relevant processors can download the warning data from the data logger and import it into the warning analysis center for further analysis.

OM-70 early warning data logger output data can be quickly summarized in the form of a report to facilitate further processing of early warning information, with excellent graphical features, you can use graphics to show the changes and fluctuations in early warning data. In the process of early warning data recording, once the recorded values do not meet the subsequent needs of early warning analysis, can be immediately adjusted to make it meet the subsequent needs of early warning analysis. Therefore, the use of this early warning data recorder can reduce the difficulty of early warning and improve the accuracy of prediction.

3.1.2. AT24C16N memory chip. Early warning data and warning signals from the data logger are numerous and cannot be analyzed in a unified way, these early warning data and warning signals must be stored in the storage chip, and then extracted and analyzed, so in this paper, the AT24C16N storage chip is selected as the core storage chip for designing the early warning system, which is packaged in a SOP-8 package, has good storage performance, contains 8 storage interfaces, and a single storage interface structure is shown in Figure 2. The memory chip has a wealth of storage interfaces, good storage performance, the maximum supply voltage of 5.36V, in line with the actual use of the system designed in this paper, can effectively improve the storage performance of the early warning system.

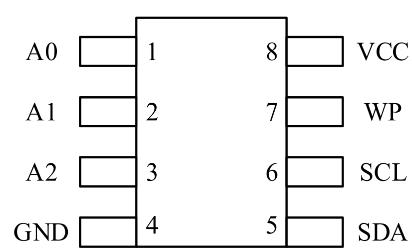

Fig. 2. Storage interface

3.2. Software design

3.2.1. Classification of psychological conditions based on decision tree modeling. The idea of decision tree is very simple, the essence is a tree consisting of multiple judgment nodes, where each internal node represents a judgment on an attribute, each branch represents an output of a judgment result, and finally each leaf node represents a classification result. In the decision tree algorithm, information gain, information gain rate, and Gini index are its important classification basis [36-38].

1) Information Gain and Information Gain Rate. Information gain is the difference of entropy before and after dividing the dataset with a certain feature, which can know how to decide efficiently and

process the sequence of features. The information gain rate is jointly defined by the ratio of the information gain and the corresponding “intrinsic value” of its attributes.

The ID3 algorithm uses information gain as the evaluation criterion. For example, in this project, five characteristics of students are known, and the psychological loneliness of each individual is predicted, and it is necessary to know which characteristic can most directly reflect the results or have the greatest influence on the results, and it is necessary to pay attention to this characteristic first in the judgment.

The formula for information gain is:

$$Gain(D,a)=Ent(D)-Ent(D\mid a)=Ent(D)-\sum_{v=1}^{v}\frac{D^{v}}{D}Ent(D^{v}) \tag{7}$$

Among them, information entropy is one of the most commonly used metrics to measure the purity of the sample set, which is calculated as:

$$Ent(D)=-\sum_{k=1}^{n}\frac{C^{k}}{D}\log\frac{C^{k}}{D} \tag{8}$$

The conditional entropy is calculated as:

$$Ent(D,a)=\sum_{v=1}^{v}\frac{D^{v}}{D}Ent(D^{v})=-\sum_{v=1}^{v}\frac{D^{v}}{D}\sum_{i=1}^{\kappa}\frac{C^{iv}}{D_{v}}\log\frac{C^{iv}}{D_{v}} \tag{9}$$

For example, 15 pieces of data in this project are randomly taken out, and the information gain of each feature is calculated in this range, and the calculation process is:

$$Ent(D)=-\left[\left(\frac{10}{15}\right)\times\log\left(\frac{10}{15}\right)+\left(\frac{4}{15}\right)\times\log\left(\frac{4}{15}\right)+\left(\frac{1}{15}\right)\times\log\left(\frac{1}{15}\right)\right]=0.348 \tag{10}$$

$$Ent(family-0)=-\left[\left(\frac{8}{10}\right)\times\log\left(\frac{8}{10}\right)+\left(\frac{2}{10}\right)\times\log\left(\frac{2}{10}\right)\right]=0.217 \tag{11}$$

$$Ent(family-1)=-\left[\left(\frac{1}{3}\right)\times\log\left(\frac{1}{3}\right)+\left(\frac{2}{3}\right)\times\log\left(\frac{2}{3}\right)\right]=0.27 \tag{12}$$

$$Ent(family-2)=-g\left[\left(\frac{1}{1}\right)\times\log\left(\frac{1}{1}\right)\right]=0 \tag{13}$$

$$\begin{aligned}Ent(D,jamuy)=&\frac{1}{3}Ent(jamuy-0)\\&+\frac{1}{3}Ent(jamuy-1)\\&+\frac{1}{3}Ent(jamuy-2)\\&=0.162\end{aligned} \tag{14}$$

Then the information gain of the FAMILY feature is:

$$Gain(D,family)=Ent(D)-Ent(D,family)=0.186 \tag{15}$$

Similarly, the information gain of the remaining four features is calculated to know the feature that has the most influence on the students' psychological loneliness status among these five features.

Using the information gain rate as the optimal division attribute is the C4.5 decision tree algorithm. It overcomes the shortcomings of using information gain to select attributes with a bias toward selecting flexibilities with more values, produces classification rules that are easy to understand, and can also handle continuous numerical attributes. Each feature attribute in this project is discretized, so the specific calculation of the information gain rate is not demonstrated here. It is calculated by the formula:

$$Gain\_ratio(D,a)=\frac{Gain(D,a)}{IV(a)} \tag{16}$$

In the formula:

$$IV(a)=-\sum_{v=1}^{V}\frac{D^v}{D}\log\frac{D^v}{D} \tag{17}$$

2) Gini value and Gini coefficient. The Gini value is the probability that two samples are taken at random and their categories are different. The smaller the Gini value, the higher the purity of the data set. Generally, the flexural that minimizes the Gini coefficient after division is chosen as the optimal partition attribute, i.e., the CART algorithm:

$$Gini\left(D\right)=\sum_{k=1}^{|y|}\sum_{k=k}p_k p_{k'}=1-\sum_{k=1}^{|y|}p_{k'}^2 \tag{18}$$

$$Gini\_index(D,a)=\sum_{r=1}^{r}\frac{D^r}{D}Gain(D^r) \tag{19}$$

3.2.2. Functional modules. From the user's point of view, the system divides users into student users and system management users, and sets up two major categories of functional modules respectively.

1) Student users. Students mainly use the front-end functions of the system to log in, participate in mental health assessment activities, and obtain assessment results.

2) System Management Users. System administration users mainly manage the background of the system. For the system management user, there are three functional modules: basic information management, psychological assessment settings, and assessment result management.

Basic Information Management: This module is mainly used to collect students' basic information, screen and clean the data, and obtain standardized data information.

Psychological Assessment Setting: It is mainly used for designing the questions of psychological questionnaires, generating and distributing questionnaires, and automating the statistics of assessment results.

Evaluation result management: It mainly realizes the analysis of various evaluation results. The module uses the built-in decision tree classification model to classify students' psychological conditions according to different students' psychological characteristics, screen out students with psychological problems, and then generate a list of students and issue an early warning.

4. Analysis of the application of mental health early warning systems

4.1. Decision Tree Modeling Analysis

4.1.1. Model performance analysis. The data processed in the previous section was used as the source of research data, and the students' psychological conditions were categorized into mental health, compulsion, anxiety, interpersonal relationship and depression, and the accuracy rate, precision rate and AUC value were selected as the evaluation indexes.The AUC value refers to the area under the ROC curve, and the closer it is to 1, it means that the model performs better, and it can distinguish the positive and negative samples more accurately. In order to verify the effectiveness of the decision tree model for the classification of students' psychological conditions, the decision tree model was compared with the LR model, LightGBM model, LR+LightGBM fusion model and random forest model, and the results are shown in Table 2. After parameter optimization, the above models are compared and analyzed. From the table, it can be seen that the decision tree model constructed for students' mental health data has the best overall performance among all the models.

Table 2. Model test assessment indicators

Models	A	P	R	F1	AUC
LR	0.724	0.702	0.672	0.648	0.872
LightGBM	0.738	0.733	0.688	0.697	0.897
LR+LightGBM	0.762	0.779	0.714	0.733	0.903
RF	0.803	0.827	0.739	0.779	0.914
DT	0.926	0.942	0.857	0.908	0.928

Figure 3 shows the ROC curves of the five models, combined with Table 2, it can be seen that the AUC value of the decision tree model in this paper reaches 0.928, which is higher than that of the random forest model and higher than that of the LR+LightGBM fusion model.The F1 value reacts to the combined performance of the classification model in terms of prediction accuracy and recall for the positive samples.In terms of the F1 value performance, the decision tree model scores the highest, and the LR model scores the lowest. The logistic regression model (LR) performs poorly in terms of precision rate values. In terms of the performance of the recall rate value, the decision tree model in this paper has the best performance of 0.857. Taken together, the decision tree model in this paper has an excellent performance in all the metrics.

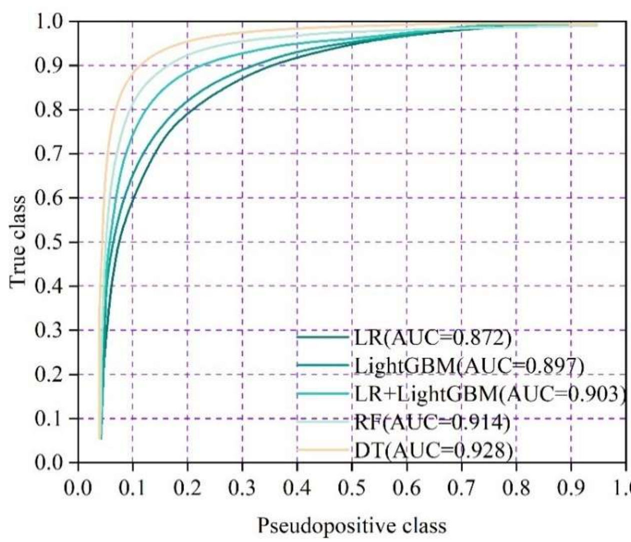

Fig. 3. ROC curve of five models

4.1.2. Classification Forecast Analysis. To ensure the reliability of the results, 100 pieces of data were randomly selected as training samples and 50 pieces of data as test sample set, and the decision tree model was compared with the random forest model, LightGBM model, and LR+LightGBM model, and the test results are shown in Fig. 4, and (a)-(d) denote the classification prediction results of the decision tree model, the random forest model, the LightGBM model, and the LR +LightGBM model's classification prediction results. In the figure, "+" represents the prediction category of students' psychological status, and "●" represents the actual category of students' psychological status, which can intuitively display the classification results of college students' psychological status and the actual students' psychological status categories through comparative display, where 1, 2, 3, 4, and 5 represent mental health, obsessive-compulsive, anxiety, interpersonal relationship and depression, respectively. When “+” and “●” coincide, the predicted and actual categories of students' psychological status are the same, indicating correct categorization. When “+” and “●” do not coincide, the predicted categories of students' psychological status and the actual categories do not coincide, and then the students' psychological status is misclassified. By comparison, the decision tree model used in this paper is more effective in the prediction scenario of students' psychological condition classification.

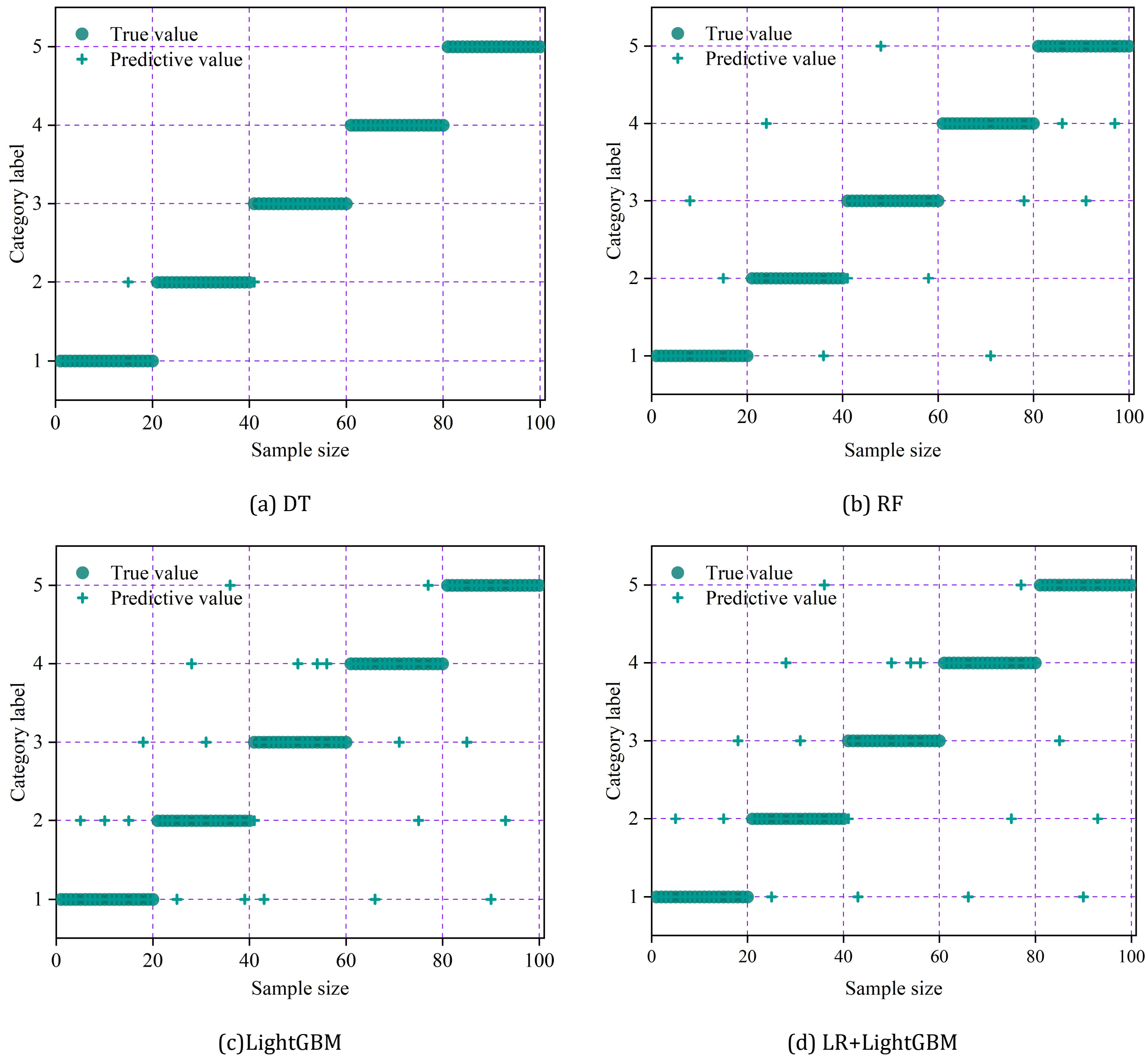

Fig. 4. Model classification prediction results

4.2. Decision Tree Visualization and Analysis

4.2.1. Construction of decision trees. For the sake of research convenience, this section only constructs the decision tree model for the psychological status of college students' interpersonal relationships. Two-thirds of the data in all the processed college students' mental health datasets are randomly selected as training samples, totaling 891 records, and the remaining one-third of the data are used as test samples. The following is the process of constructing the decision tree model based on the C4.5 algorithm, the specific steps are as follows:

1) Calculate the information gain rate of each split attribute in the set of divided training samples.

2) Select the classification attribute with the maximum information gain rate as the root node of the decision tree, and split the dataset into corresponding sub-datasets according to the number of values of this attribute.

3) Execute steps 1) and 2) sequentially and recursively in the sub-data set.

The information gain rate of attributes such as gender (XB), place of birth (SY), faculty (YX), whether only child (DS), whether single parent (DQ), and family economic status (JJ) are calculated according to the formula as shown in Table 3. According to the calculation results in the table, it can be learned that the attribute with the largest information gain rate is the attribute of whether it is an only child or not, which is taken as the root node, and two branches are generated according to its two values, corresponding to two subdatasets. Then the same steps are performed on the subdatasets on the branches respectively, and so on, the generated decision tree model of college students about the sensitive symptoms of interpersonal relationships is shown in Fig. 5.

Table 3. Information gain rate calculation results

Attribute	Information gain rate
XB	0.078
SY	0.032
YX	0.024
DS	0.183
DQ	0.126
JJ	0.118

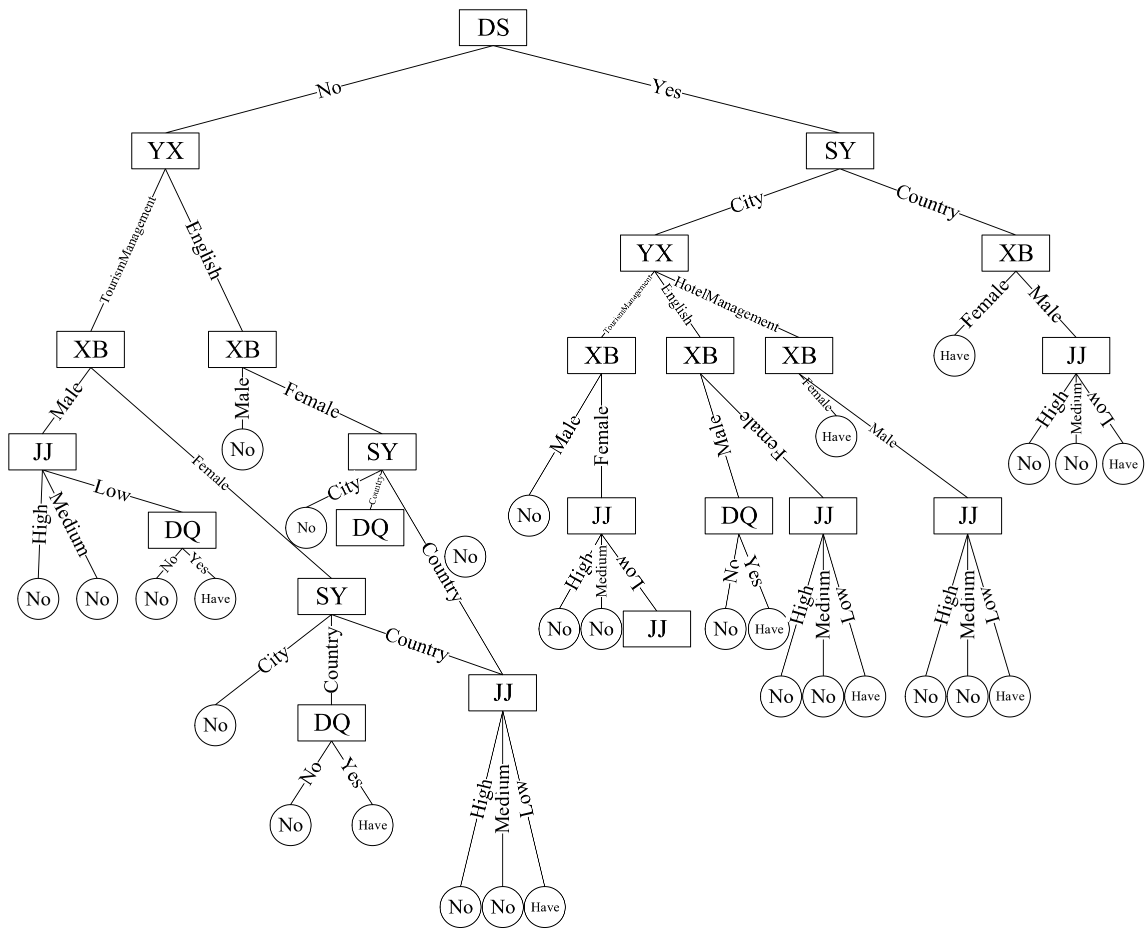

Fig. 5. Unpruned decision tree

4.2.2. Tree pruning. As the process of constructing the decision tree is completely in accordance with the training sample set, and does not process the data with noise data and outlier data, many branches in the generated decision tree are useless branches constructed from these abnormal data, which makes the generated decision tree too cumbersome, and reduces the practicability and readability of the decision tree. Pruning is an effective way to simplify the decision tree. Based on some kind of pruning algorithm, some abnormal branches in the decision tree are subtracted and replaced by leaves, which can make the decision tree smaller, reduce the complexity, and improve the ability and speed of classification prediction of the decision tree.

Pruning is divided into first pruning and post pruning. First pruning is mainly to limit the full growth of the decision tree by establishing some rules, which is carried out before the node splits. Post-pruning is to build the decision tree model first, then prune it and replace it with leaf nodes. Usually, post pruning is better than first pruning and has been applied in many places. In this study, PEP post pruning algorithm is used to prune the decision tree.

1) PEP pruning algorithm. The PEP pruning algorithm, also known as the pessimistic pruning algorithm, first calculates the accuracy of the rule on the training dataset to which it is applied, and when the accuracy has no effect before or after the pruning then the pruning is performed. In order to improve prediction reliability, the PEP algorithm estimates the error using the idea of continuity correction in statistics, which follows a binomial distribution when calculating the standard error rate.

The PEP algorithm uses a top-down approach, when the following inequality is satisfied, then Tt should be pruned and replaced with the corresponding leaf node:

$$e^{'}(t) \le e^{'}(T_t) + S_e(e^{'}(T_t)) \tag{20}$$

where T_t denotes the subtree at the root node with node t, and $Se(e'(T_t))$ is the standard error, defined as:

$$S_e(e'(T_t)) = \sqrt{\frac{e'(Tt)[n(t) - e'(Tt)]}{n(t)}} \tag{21}$$

where $e(t)$ denotes the number of misclassified instances at node t and $n(t)$ denotes the number of training sets at node t.

PEP algorithm also has some defects in the process of pruning, for example, sometimes the node that meets the pruning conditions is cut off at the same time, the node under the node may exist under the child nodes that do not need to be cut off will also be cut off. Nevertheless, the PEP algorithm shows high accuracy in practical applications, in the pruning process, each sub-tree in the tree is accessed at most once, which is more efficient than other algorithms, and the algorithm does not need to separate the training dataset.

2) Pruning decision tree based on PEP algorithm. Using the PEP algorithm to prune the previously generated decision tree, each non-leaf node in the tree is calculated once, and the non-leaf nodes and their corresponding branches that conform to the judgment formula are subtracted and replaced with leaf nodes. The class to which most of the samples in the subtracted branches belong is used to identify that leaf node, thus constructing the decision tree after pruning based on the PEP algorithm, and the results are shown in Figure 6. By comparing with the decision tree generated for the first time, it can be found that the size of the decision tree after being pruned by the PEP algorithm becomes smaller, which will improve the readability and classification speed of the decision tree model.

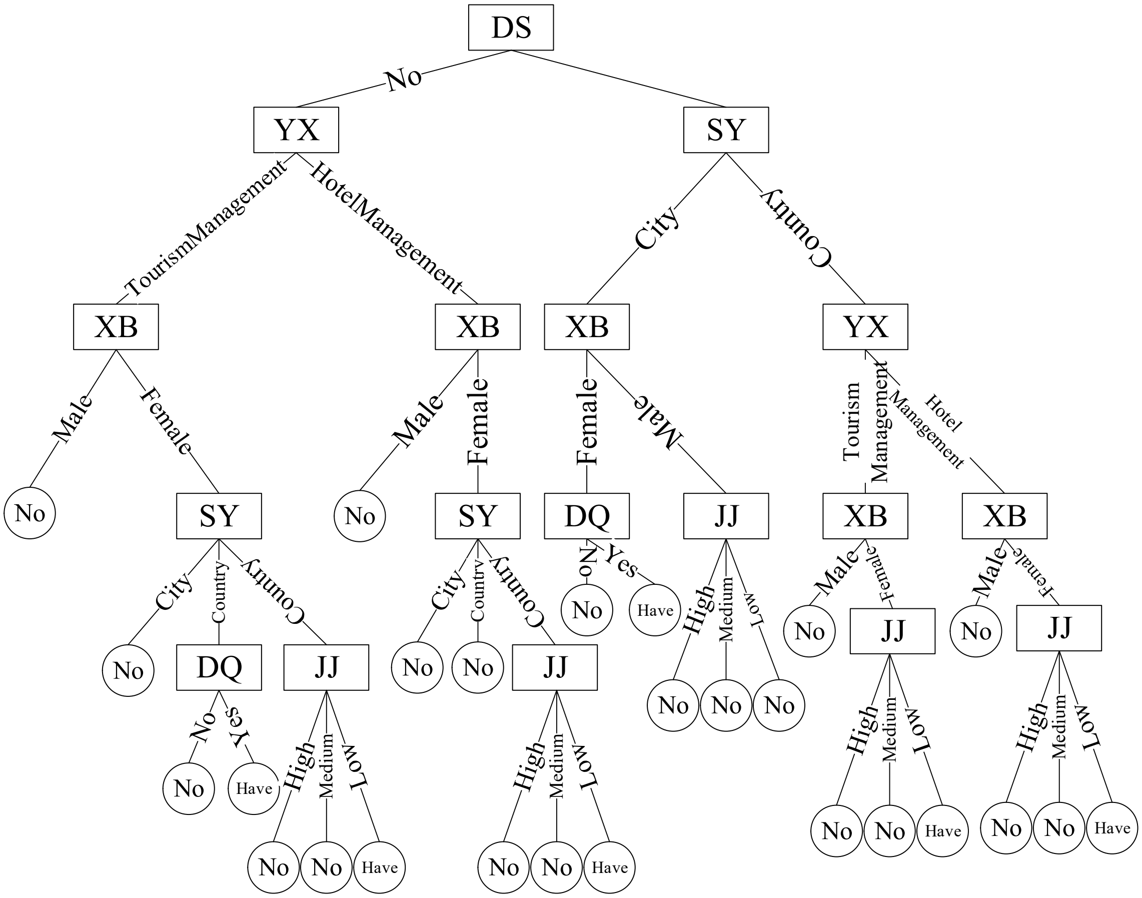

Fig. 6. Decision tree after pruning

4.2.3. Access to classification rules. The decision tree algorithm is able to interpret the classification rules directly using the expression IF...THEN. According to Fig. 5 the decision tree gets more entries of association rules and some of the rules are longer, such as:

IF DS=No and YX=HotelManagement and XB=Male and SY
=Country and DQ=No and JJ=Low then Have
IF DS=Yes and YX=HotelManagement and XB=Female and SY
=City and DQ=No and JJ=Low then Have

The rules extracted from the decision tree model based on interpersonal sensitivity symptoms obtained after pruning according to PEP to categorize students regarding interpersonal sensitivity symptoms are shown below (partial rules):

IF DS=No and YX=TourismManagement and XB=Male
then No
IF DS=No and YX=TourismManagement and XB=Female and SY=City
then No
IF DS=No and YX=TourismManagement and XB=Female and SY=City and DQ=No
then No
IF DS=No and YX=TourismManagement and XB=Female and SY=City and DQ=Yes
then Have
IF DS=No and YX=TourismManagement and XB=Female and SY=Country and JJ=High
then No
IF DS=No and YX=TourismManagement and XB=Female and SY=Country and JJ=Medium
then No
IF DS=No and YX=TourismManagement and XB=Female and SY=Country and JJ= Low
then No
IF DS=No and YX=HotelManagement and XB=Male
then No
IF DS=No and YX=HotelManagement and XB=Female and SY=City
then No
IF DS=No and YX=HotelManagement and XB=Female and SY=Country
then No
IF DS=No and YX=HotelManagement and XB=Female and SY=Country and JJ=High
then No
IF DS=No and YX=HotelManagement and XB=Female and SY=Country and JJ=Medium
then No
IF DS=No and YX=HotelManagement and XB=Female and SY=COountry and JJ=Low
then No
……

4.2.4. Analysis of results. From the generated rules, it can be found that there is a relationship between the attribute of being an only child or not and the psychological status of students' interpersonal relationships, among the population of interpersonal relationship sensitivity symptoms, only children have a greater chance of experiencing such symptoms, from the point of view of gender, girls have a greater chance of experiencing such symptoms, from the point of view of faculties, students from the department of hospitality management are higher than those from other faculties, and from the point of view of whether or not being a single parent is a single parent, children from single-parent families are more likely to be sensitive to interpersonal relationships. It can also be seen from the rules that among the one-child population, female students from rural areas are more likely to be sensitive to interpersonal relationships. It is also more common for students with poor family economic status. The results show that the mental health early warning system for college students constructed in this paper is able to predict the mental health status of students based on the relationship between these similar attributes and the mental health status of students, and provide early warning hints to students who are prone to psychological problems.

4.3. System application effect analysis

In order to analyze the effects of the three measures of this paper's Student Mental Health Alert System, the traditional Student Mental Health Alert System, and No Mental Alert in the five psychological conditions of students, firstly, the paired samples t test was carried out before the application of this paper's system (T1 moments) and after the application of this paper's system for one academic year (T2 moments) and the calculation of effect sizes t was carried out by calculating the effect size. Finally, a one-way analysis of variance (ANOVA) was used to compare the between-group changes in the different groups. If the ANOVA results showed significant differences, post hoc multiple comparisons were performed using LSD to determine which measure was the main cause of the differences. A value of p less than 0.05 on all tests was considered statistically significant.

4.3.1. Initial examination scores. All participants were randomly assigned to three groups (test group, control group, and blank group, denoted as G1, G2, and G3), where the test group used the mental health warning system for students designed in this paper, the control group used the traditional mental health warning system, and the blank group did not have psychological warnings. The results of conducting paired t test samples are shown in Table 4, and there is no significant difference in the scores among the three groups to ensure that there is no significant difference in any of the groups before the application of the system in this paper.

Table 4. T test results

Test	G1		G2		G3	
	t	p	t	p	t	p
Health	0.872	0.492	0.828	0.573	0.863	0.674
Compulsion	0.306	0.904	0.348	0.428	0.429	0.592
Anxiety	1.542	0.378	1.847	0.393	1.239	0.472
Interpersonal relation	0.944	0.652	0.871	0.471	0.772	0.396
Depression	0.536	0.474	0.705	0.398	0.638	0.235

4.3.2. Pre- and post-system application ratings. In order to assess the effects of each of the three measures (the psychological warning system in this paper, the traditional psychological warning system, and no psychological warning), participants in the test group, the control group, and the blank group were asked to fill out the TIPI and DASS-42 scales at the T1 and T2 moments, respectively, and the scores were tallied. The measurements of the scales are summarized in Table 5.

Table 5. T test results under different measures

Group	Test	T1		T2		t	p	Cohen'D
		M	SD	M	SD			
G1	Health	10.86	4.376	32.48	2.387	5.348	0.000	1.636
	Compulsion	33.28	3.287	12.39	1.363	6.457	0.000	1.997
	Anxiety	47.62	1.342	10.87	1.042	4.592	0.000	1.435
	Interpersonal relation	63.46	2.874	12.04	0.973	5.773	0.000	1.648
	Depression	72.93	3.208	11.49	0.389	6.089	0.000	1.706
G2	Health	8.74	3.592	9.62	4.386	1.231	0.572	0.748
	Compulsion	32.09	1.643	30.36	3.297	1.034	0.431	0.632
	Anxiety	41.38	2.795	38.47	3.374	1.852	0.396	0.539
	Interpersonal relation	59.36	3.287	52.74	2.853	1.463	0.507	0.607
	Depression	64.87	4.396	63.91	3.291	1.036	0.422	0.705
G3	Health	9.72	5.734	8.43	3.974	-0.352	0.047	-0.487
	Compulsion	32.48	2.887	33.07	2.863	-0.341	0.036	-0.433
	Anxiety	45.73	4.896	46.89	4.732	-0.674	0.031	-0.376
	Interpersonal relation	62.19	1.303	63.94	3.998	-0.743	0.025	-0.325
	Depression	73.98	3.371	75.68	2.969	-0.539	0.027	-0.371

1) Test group. Mental health: the scores of the test group between T1 (M=10.86, SD=4.376) and T2 (M=32.48, SD=2.387), there is a significant difference ($t=5.348$, $p<0.001$), and the value of this effect was calculated to be 1.636 using Cohen's D. It indicates that the mental health of the students of the higher education institutions of this paper has been improved after the use of the Early Warning System improved the mental health of students.

Compulsion: student's M=33.28, SD=3.287 at T1 moment. m=12.39,SD=1.363 at T2 moment, $t=6.457$,$p<0.001$ indicates significant difference in the score results between the two moments, Cohen's D was calculated as 1.997. It is shown that the use of this early warning system for the mental health of students in higher education significantly reduces the “obsessive-compulsive” mentality of students.

Anxiety: between T1 (M=47.62,SD=1.342) and T2 (M=10.87,SD=1.042), the scores of the students in the test group produced a significant difference ($t=4.592$,$p<0.001$), and the value of this effect was calculated to be 1.435 using Cohen's D. This result indicates that the use of this paper's college students' Mental Health Early Warning System is effective in reducing the anxiety value of students.

Relationships: after using this paper's mental health early warning system for college students interpersonal sensitivity symptom score at the T2 moment M=12.04, SD=0.973, and $t=5.773$, $p<0.001$. It indicates that there is a significant difference between the students' interpersonal sensitivity symptoms at the moment of T2 and T1, and that the students' interpersonal sensitivity symptoms improved significantly at the moment of T2.

Depression: between T1 (M=72.93,SD=3.208) and T2 (M=111.49,SD=0.389), there was a significant difference in the scores of the test groups ($t=6.089$,$p<0.001$), and the value of this effect was calculated to be 1.706 using Cohen's D. This result indicates that the use of this paper's college students' Mental Health Early Warning System is effective in reducing the psychological state of depression in students.

2) Control group. The mean values M of the five psychological states of mental health, obsessive-compulsive, anxiety, interpersonal relations and depression of the students in the control group at the moment of T1 were 8.74, 32.09, 41.38, 59.36 and 64.87, respectively, and at the moment of T2 were 9.62, 30.36, 38.47, 52.74 and 63.91, respectively. Although the students' psychological symptoms improved in T2 compared to T1, the results were not significant, with p-values of 0.572, 0.431, 0.396,

0.507, and 0.422, respectively, indicating that the traditional early warning system of students' mental health has only a slight effect on the students' mental health level, and it cannot significantly improve the students' own psychological problems, such as anxiety, depression, and so on.

3) Blank group. From the results in the table, it can be clearly seen that the psychological status of the students in the blank group produces a significant deterioration at the moment of T2. The mean value of mental health score decreased from 9.72 at the moment of T1 to 8.43 at the moment of T2, and the mean value of the remaining four scores of the state of mind increased from 32.48, 45.73, 62.19, and 73.98 at the moment of T1 to 33.07, 46.89, 63.94, and 75.68 at the moment of T2, respectively, with p-values of 0.047, 0.036, 0.031, 0.025, and 0.027, which are statistically significant. It indicates that in the absence of any psychological early warning measures for students, the lack of psychological counseling may result in further exacerbation of students' psychological problems, and this result is not conducive to the personal development of students or to efficient daily student management.

4.3.3. Inter-measure scoring. The results of one-way ANOVA with different measures between groups are shown in Table 6. From the table, it can be seen that the mental health early warning system for college students designed in this paper has a significant effect on the five psychological states of students, with F-values of 5.946, 5.873, 5.325, 6.984, and 5.738, and p-values of 0.002, 0.007, 0.001, 0.004, and 0.009, respectively. According to the results of LSD test, compared with the traditional student mental health early warning system and no mental health early warning measures, the use of this system caused significant differences in the psychological status of students in the test group, the control group and the blank group, with p values of 0.014, 0.011, 0.008, 0.013, 0.007, 0.002, 0.009, 0.012, 0.006 and 0.001, respectively. In contrast, there was no significant difference in the psychological status of the students in the test and blank groups, with p-values of 0.574, 0.396, 0.297, 0.384, and 0.273, respectively. Combined with the results of the previous analysis, it can be concluded that the use of decision tree-based mental health early warning system for students in colleges and universities can significantly improve the level of students' mental health, effectively improve students' psychological state of compulsion, anxiety, interpersonal relationship sensitivity and depression. The traditional student mental health early warning system and no psychological warning measures can not achieve this effect.

Table 6. Analysis results of single factor variance

Test	F	p	G1/G2	G1/G3	G2/G3
Health	5.946	0.002	0.014	0.002	0.574
Compulsion	5.873	0.007	0.011	0.009	0.396
Anxiety	5.325	0.001	0.008	0.012	0.297
Interpersonal relation	6.984	0.004	0.013	0.006	0.384
Depression	5.738	0.009	0.007	0.001	0.273

5. Conclusion

In this paper, a decision tree model based on the C4.5 algorithm is used to construct a mental health early warning system for college students. The AUC value of the decision tree model is as high as 0.928, and the precision rate, recall rate and F1 value are the best compared with LR, LightGBM, LR+LightGBM and Random Forest models, and the accuracy rate is the highest among all the models, which is 0.926. In the randomly selected samples, this paper's decision tree model has the best performance in the classification and prediction of students' five psychological conditions, which fully demonstrates the good performance of this paper's decision tree model based on C4.5 algorithm in classifying and predicting students' mental health conditions. In the randomly selected samples, the decision tree model based on C4.5 algorithm in this paper has the best performance in the classification prediction of students' mental health conditions. The decision tree model after pruning can clearly show the relationship between different attributes and students' interpersonal psychological status, and the system can use such attributes to give timely warning to the students who may have psychological problems, and help colleges and universities accurately master the psychological counseling objects. After using the system in this paper, the mental health level of the test group was significantly improved, while the psychological conditions of "obsessive-compulsive", "anxiety", "interpersonal relationship" and "depression" were significantly decreased. It shows that the early warning system of students' mental health constructed in this paper can effectively improve the level of students' mental health and significantly improve the psychological problems of students.

References

[1] Anniko, M. K., Boersma, K., & Tillfors, M. (2019). Sources of stress and worry in the development of stress-related mental health problems: A longitudinal investigation from early-to mid-adolescence. Anxiety, Stress & Coping: An International Journal.

[2] Sheek-Hussein, M., Abu-Zidan, F. M., & Stip, E. (2021). Disaster management of the psychological impact of the COVID-19 pandemic. International Journal of Emergency Medicine, 14(1), 19.

[3] Safianov, V., Sokolovskaya, I., Shcherbakova, O., Belobragin, V., & Tkach, R. (2021). Ethical and psychological features of conflict prevention in interpersonal communication contacts. In E3S Web of Conferences (Vol. 291, p. 05021). EDP Sciences.

[4] Karatekin, C., & Ahluwalia, R. (2020). Effects of Adverse Childhood Experiences, Stress, and Social Support on the Health of College Students. Journal of Interpersonal Violence, 35.

[5] Barry, K. M., Woods, M., Warnecke, E., Stirling, C., & Martin, A. (2018). Psychological health of doctoral candidates, study-related challenges and perceived performance. Higher Education Research & Development.

[6] Bibi, A. (2022). RELATIONSHIP BETWEEN ACADEMIC EXPECTATION STRESS, CAREER ADAPTABILITY AND PSYCHOLOGICAL WELL-BEING AMONG POSTGRADUATE STUDENTS. FJWU (Economics) Working Paper Series, 2(1).

[7] Mofatteh, M. (2020). Risk factors associated with stress, anxiety, and depression among university undergraduate students. AIMS public health, 8(1), 36-65.

[8] Ma, J., & Lin, B. (2022). The impact of college students' social anxiety on interpersonal communication skills: A moderated mediation model. Scientific and Social Research, 4(6), 101-108.

[9] Acharya, L., Jin, L., & Collins, W. (2018). College life is stressful today—Emerging stressors and depressive symptoms in college students. Journal of American College Health.

[10] Richardson, T., Elliott, P., & Roberts, R. (2017). Relationship between loneliness and mental health in students. Journal of Public Mental Health.

[11] Kumar, M., & Mondal, A. (2018). A study on Internet addiction and its relation to psychopathology and self-esteem among college students. Industrial Psychiatry Journal, 27(1), 61.

[12] Mishna, F., Regehr, C., Lacombe-Duncan, A., Daciuk, J., Fearing, G., & Van Wert, M. (2018). Social media, cyber-aggression and student mental health on a university campus. Journal of Mental Health, 27(3).

[13] Carey, K. B., Norris, A. L., Durney, S. E., Shepardson, R. L., & Carey, M. P. (2018). Mental health consequences of sexual assault among first-year college women. Journal of American College Health.

[14] Patalay, P., & Fitzsimons, E. (2021). Psychological distress, self-harm and attempted suicide in UK 17-year olds: prevalence and sociodemographic inequalities. The British Journal of Psychiatry, 219, 437-439.

[15] Ghrouz, A. K., Noohu, M. M., BaHammam, A. S., & Pandi-Perumal, S. R. (2019). Physical activity and sleep quality in relation to mental health among college students. Sleep & Breathing, 23(2), 627-634.

[16] Lin, M., Wolke, D., Schneider, S., & Margraf, J. (2020). Bullying history and mental health in university students: The mediator roles of social support, personal resilience, and self-efficacy. Frontiers in Psychiatry.

[17] Salimi, N., Gere, B., Talley, W., & Irioogbe, B. (2023). College Students Mental Health Challenges: Concerns and Considerations in the COVID-19 Pandemic. Journal of College Student Psychotherapy, 37(1).

[18] Zhang, Z. (2024). Early warning model of adolescent mental health based on big data and machine learning. Soft Computing, 28(1), 811-828.

[19] Fang, L. (2021). The Prevention and Analysis of College Students' Psychological Crisis Based on Data Mining Technology. In Big Data Analytics for Cyber-Physical System in Smart City: BDCPS 2020, 28-29 December 2020, Shanghai, China (pp. 1505-1509). Springer Singapore.

[20] Cai, L. (2022). A Novel Recognition Model of University Students' Psychological Crisis Based on DM. Scientific Programming.

[21] Skaik, R., & Inkpen, D. (2020). Using social media for mental health surveillance: a review. ACM Computing Surveys (CSUR), (6), 1-31.

[22] Chai, Y., Luo, H., Zhang, Q., Cheng, Q., Lui, C. S. M., & Yip, P. S. F. (2019). Developing an early warning system of suicide using Google Trends and media reporting. Journal of Affective Disorders.

[23] Schislyaeva, E. R., & Saychenko, O. A. (2023). The design of a mental health service platform for college students based on multi-modal information. Microprocessors and Microsystems, 96, 104748.

[24] Dablander, F., Pichler, A., Cika, A., & Bacilieri, A. (2022). Anticipating critical transitions in psychological systems using early warning signals: Theoretical and practical considerations. Psychological Methods.

[25] Su, C., Aseltine, R., Doshi, R., Chen, K., Rogers, S. C., & Wang, F. (2020). Machine learning for suicide risk prediction in children and adolescents with electronic health records. Translational Psychiatry, 10(1).

[26] Zheng, L., Wang, O., Hao, S., Ye, C., Liu, M., Xia, M., ... & Ling, X. B. (2020). Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. Translational Psychiatry, 10.

[27] Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K. J., & Tørresen, J. (2018). Mental health monitoring with multimodal sensing and machine learning: A survey. Pervasive and Mobile Computing, 51, 1-26.

[28] Srividya, M., Mohanavalli, S., & Bhalaji, N. (2018). Behavioral Modeling for Mental Health using Machine Learning Algorithms. Journal of Medical Systems, 42(5), 1-12.

[29] Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: a scoping review. Translational Psychiatry, 10.

[30] Wei, H., Wang, L., He, D., & Li, Y. (2022, September). Research on the Early Warning System of College Students' Mental Health Based on Big Data Analysis. In 2022 International Conference on Computer Network, Electronic and Automation (ICCNEA) (pp. 50-54). IEEE Computer Society.

[31] Zhang, Y. (2020, October). Research on College Students' Psychological Crisis Early Warning Method Based on Fuzzy Clustering Algorithm. In 2020 International Conference on Computers, Information Processing and Advanced Education (CIPAE) (pp. 26-30). IEEE Computer Society.

[32] Han, H. (2023). Fuzzy clustering algorithm for university students' psychological fitness and performance detection. Heliyon, 9, e18550.

[33] Zhang, W., & Liang, Y. (2022). Construction of an Intelligent Evaluation Model of Mental Health Based on Big Data. Journal of Sensors, 2022.

[34] YANG, X., HU, Q., ZHAO, J., ZHAO, J., & ZHANG, X. (2019). Development of an Evaluation Instrument for Warning Signs of Individual Psychological Crisis. Chinese General Practice, 22(4), 484.

[35] Li Ying,Cui Xueting,Fan Jiahao & Wang Tan. (2022). Global chaotic bat algorithm for feature selection. The Journal of Supercomputing(17),18754-18776.

[36] Xinting Du & Hejin Wang. (2024). Efficient estimation of expected information gain in Bayesian experimental design with multi-index Monte Carlo. Statistics and Computing(6),200-200.

[37] Jinrui Gao,Ziqian Wang,Ting Jin,Jiujun Cheng,Zhenyu Lei & Shangce Gao. (2024). Information gain ratio-based subfeature grouping empowers particle swarm optimization for feature selection. Knowledge-Based Systems111380-.

[38] Hassan Arabshahi & Hamed Fazlollahtabar. (2018). Classifying Innovative Activities Using Decision Tree and Gini Index. International Journal of Innovation and Technology Management(3).