Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 14 (2026) pp. 1-16

Article

Research on the optimization algorithm of college mental health education strategy based on big data in the new quality productivity environment

Wei Zhang [1,*], Qianru Shi [2] and Duanhe Li [1]

[1] Shijiazhuang institute of railway technology Shijianzhuang ,Hebei,China,050041
[2] Hebei Vocational College of Rail Transportation, Shijiazhuang, Hebei Province, 050000
* Correspondence author: peacelove_wei@163.com

Abstract: This paper proposes an improved RS-ID3 algorithm based on rough set theory and continuous attribute discretization method. Taking the students of 2022 class in a higher education institution as the research object, the mental health data were collected. Data mining was performed on the mental health test data, and a mental health prediction model was built. The predictive performance of the model was evaluated, and an obsessive-compulsive symptom decision-making model was built using the optimized RS-ID3 algorithm to explore the students' mental health. The results showed that the decision tree prediction accuracy using the RS-ID3 algorithm was 8.5 percentage points higher than that using the ID3 algorithm, and the mining rules indicated that students who were not from two-parent families were more likely to have obsessive-compulsive symptoms. The mining results of the combined two algorithms were fed back to the counselors, providing a theoretical basis for the evaluation and intervention of college students' mental health status.

Keywords: mental health; decision tree; RS-ID3 algorithm; data mining

1. Introduction

Psychological health is the foundation for the growth and success of college students, including four aspects, namely, emotional, intellectual and physical harmony, harmonious interpersonal relationships, the ability to adapt to the environment, the ability to give full play to their own abilities, and the ability to feel a sense of well-being in their studies, work and life [1-3]. However, from the current situation, college students' psychology is generally in a subhealthy state, that is, although there is no mental disease or psychological disorder, but often appear helplessness, anxiety and other negative psychological emotions, these negative emotions if not promptly resolved are very likely to cause mental illness [4-6]. Therefore, in order to better promote the mental health of college students, it is of great practical significance to explore the problems and improvement measures of college students' mental health education in the context of big data. Big data has a strong ability in acquiring, storing, managing, and analyzing information data, and has various characteristics such as massive, high-speed, diversity, and low value density, which can satisfy a variety of data needs [7-8]. The development of big data has brought a series of development opportunities for college students' mental health education, which can help mental health education timely and accurately find out the needs of college students' subjects at the psychological level, and can also obtain more and richer resources for mental health education, and use a variety of educational methods to enhance the efficiency of education, and help colleges and universities and educators to effectively lead the development of college students' mental health in an accurate way [9-11]. At the same time, in the era of big data, there are certain challenges in college students' mental health education, which affect the effectiveness of education and the continuous promotion of college students' mental health [12].

Under the background of big data, college students are generally accustomed to using the Internet to obtain information and learn knowledge, and have developed the habit of independence and autonomy, and also show individualized needs in their daily study and life, coupled with the fact that big data has accelerated the rapid dissemination of a variety of ideological viewpoints and various types of information, which makes college students susceptible to the impact of accepting this kind of information, and shows certain problems in the level of mental health, affecting their mental health development. [13-15]. Moreover, under the background of big data, how colleges and universities and educators can effectively use big data technology to analyze the information and data related to college students' mental health education, and how to grasp data privacy in the process of precision education, etc., all of which bring challenges to college education [16-18]. In view of the opportunities for precision development of mental health education in the era of big data, colleges and universities and educators should grasp the characteristics of the era of big data, combine the actual schooling with the mental health status of college students, and explore the feasible path of precision development of education.

Literature [19] verified that k-mean clustering played a positive role in predicting college students' mental health data based on a pooled assessment test of students' mental health data. Literature [20] describes the dissatisfaction of teachers and students with the process of mental health education MHE, and proposes that the MHE strategy has been optimized based on big data technology, which has improved the effectiveness of mental health education. Literature [21] proposed a computer technology-based mental health teaching platform for colleges and universities, and combined with big data analysis technology to analyze students' mental health and promote the information construction of mental health education. Literature [22], based on the perspective of big data analysis, discusses the current teaching of mental health education is not standardized, the degree of informatization is not high, and proposes that the construction of a big data analysis platform has improved and sound mental health education system and guarantee mechanism in colleges and universities. Literature [23] in-depth understanding of the advantages and characteristics of big data technology on the basis of the introduction of big data analysis technology in mental health education, in order to simplify the workflow of mental health education and efficient and high-quality solution to student mental health problems, to provide technical support for student mental health education. Literature [24] envisioned a prediction framework for college students' mental health with C4.5 decision tree algorithm as the underlying logic and a negative psychological structure model based on the decision tree algorithm, and finally used a clustering method to discover students' potential depression, which provided valuable references for students' psychological counseling. Literature [25] conceptualized an analytical framework combining the Internet of Things and big data to assess how music education affects students' mental health, and the method accurately explains the correlation and mechanism of action between music education and students' mental health.

In this paper, ID3 algorithm is elaborated, rough set theory is introduced, and improved RS-ID3 algorithm is proposed. Based on the background of the general environment of new quality productivity, the decision tree model is introduced into the research of mental health education in colleges and universities. The questionnaire survey method is used to collect the mental health data of students in a university, and the mental health prediction model is constructed. The performance of the model is determined by three indicators: gain assessment, response assessment, and enhancement assessment. The ID3 algorithm is used to construct the decision tree model and classification rules for college students' mental health prediction, and the RS-ID3 algorithm is used to optimize the decision tree model, and the results of the study are analyzed in general.

2. ID3 algorithm and its improvements

2.1. ID3 algorithm

2.1.1. Theoretical foundations

Information theory, i.e., measuring and studying information mathematically, measures the amount of information through the elimination of uncertainty about the appearance of various symbols in the source after communication, and the related concepts are as follows:

(1) Self-information. Let $X_1,\cdots X_n$ be the signal sent by the source, and before receiving X_i, the uncertainty of the recipient about the signal sent by the source is defined as the amount of self-information of the information symbols $I(X_i)$. i.e., $I(X_i)=-\log_2 P(Xi)$, where $P(X_i)$ is the probability that the source sends X_i.

(2) Information entropy. The self-information can only reflect the uncertainty of the symbol, while the information entropy can be used to measure the uncertainty of the whole source X as a whole, defined as follows:

$$H(X) = -\sum_{i=1}^{n} P(Xi) \log_2 P(Xi) n \in [1, n] \tag{1}$$

where n is the number of all possible symbols for source X, i.e., the information entropy (average amount of information) is defined in terms of the average amount of self-information provided by the source per symbol sent.

(3) Conditional entropy. If source X and random variable Y are not independent of each other, and the recipient receives message Y, then conditional entropy $H(X/Y)$ is used to measure the uncertainty that the recipient still has about random variable X after receiving random variable Y. Let X corresponds to the source symbol X_i, Y to the sex source symbol Y_i, $P(X_i / Y_i)$ is the probability that X is X_i when Y is Y_i, then we have:

$$\begin{aligned} &H(X) = -\sum_{i=1}^{n} \sum_{j=1}^{m} H(X) - H(X/Y) \\ &\log_2 P(Xi/Yj) i \in [1, n], j \in [1, m] \end{aligned} \tag{2}$$

(4) Average information content. It is used to indicate the amount of information about X that signal Y can provide, and is denoted by $I(X,Y)$:

$$I(X,Y) = H(X) - H(X/Y) \tag{3}$$

2.1.2. Basic Ideas

The ID3 algorithm is an algorithm that selects the class of examples based on the values taken from the set of attributes, and its basic idea is:

Let $E = F_1 * F_3 * \cdots * F_n$ be a n-dimensional exhaustive vector space, where F_j is a set of exhaustive discrete symbols, and the elements $e = \langle V_1, V_2, \cdots V_n \rangle$ in E are called examples. where $Vn \in Fj, j = 1, 2, \cdots, n$. Let PE, NE be the two sets of examples in E, the positive example set and the negative example set, respectively.

Assuming that the sizes of the positive and negative example sets in the vector space are p and n, respectively, ID3 is based on the following two assumptions.

(1) A correct decision tree on vector space E has the same classification probability for any example as the probability of positive and negative examples in E.

(2) The amount of information required for a decision tree to make a correct category judgment for an example is:

$$I(p,n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n} \tag{4}$$

If attribute A is used as the root of the decision tree, A has v values, $\{V_1, V_2, \ldots, V_V\}$ they divide E into v subsets $\{E_1, E_2, \ldots, E_V\}$ and assuming that E_i contains p_i positive and n_i negative examples, then the desired information required for subset E_i is $I(p_i, n_i)$ and the desired entropy with attribute A as the root is:

$$E(A) = -\sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i) \tag{5}$$

The information gain rooted at A is:

$$gain(A) = I(p,n) - E(A) \tag{6}$$

ID3 chooses the attribute that maximizes $gain(A)$, i.e., minimizes $E(A)$, as the root node, and different values of A correspond to subsets of E. E_i recursively calls the above procedure to

generate subnodes $B_1, B_2, \ldots, B_V$ of A.

The ID3 classification algorithm is basically based on two-class classification problems, where we can easily extend it to multi-class classification problems.

Suppose the sample problem set S has C classes of samples, each with a number of samples of $P_i\left(i=1,2,\cdots C\right)$. If attribute A is used as the root of the decision tree, A has v values $V_1, V_2, \cdots, V_v$, which divides E into v subsets $\left\{E_1, E_2, \ldots, E_v\right\}$. Suppose E_i contains j classes of samples with a number of samples of $p_{ij}\left(j=1,2,\cdots C\right)$. Then the information content $E\left(E_i\right)$ of subset E_i is:

$$E\left(E_i\right)=\sum_{j=1}^{v}\frac{p_{ij}}{\left|E_i\right|}\log_2\frac{p_{ij}}{\left|E_i\right|} \tag{7}$$

The information entropy with A as the root is:

$$E\left(A\right)=\sum_{i=1}^{v}\frac{\left|E_i\right|}{\left|E\right|}E\left(E_i\right) \tag{8}$$

Attribute A is chosen to minimize Eq. (8) $E\left(A\right)$ and maximize the information gain.

ID3 is a typical decision learning system. It uses information entropy as an evaluation function for separating the objectives and searches out a part of the full space using a top-down non-returnable strategy, which ensures that the decision tree is the simplest to build and the least amount of test data is made at a time.

2.2. Improved ID3 algorithm - RS-ID3 algorithm

For the shortcomings of ID3 algorithm that cannot deal with discretized data, this paper adopts a continuous attribute discretization method based on rough set theory. It is a method to deal with uncertainty, which can analyze the data, reason, and discover knowledge and reveal laws from it, and it can also effectively analyze various incomplete information, such as some incomplete, imprecise, and inconsistent data. As an effective tool to deal with data discretization, rough set theory can divide the conditional attributes in the decision system without destroying the discriminative relationship of the decision system. The following introduces the concepts related to decision table in rough set theory.

A decision table is a special knowledge system, let $S=\left(U,A,V,F\right)$ be an information system. Where $U=\left\{X_1X_2\ldots\ldots X_n\right\}$ is the domain; A is the set of attributes; V is the set of attribute values; and F is the mapping from $U\times A$ to V. If $A=C\cup D$, $C\cap D=\varnothing$, and C are called sets of conditional attributes and D is called set of decision attributes, then the information system is called a decision table.

A decision table usually consists of the following four parts:

(1) Condition Pile - lists all the conditions of the problem.

(2) Conditional Items - lists all possible values for all conditions given in the condition stakes.

(3) Action Stake - lists the possible actions that can be taken as specified by the problem.

(4) Action item - indicates the action taken for each value of the condition item.

The decision table structure is shown in Figure 1.

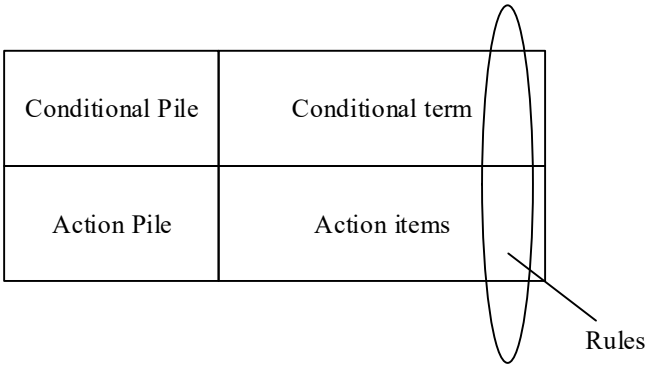


Figure 1. The structure of decision table

A decision table lists all the possible cases and clearly indicates the corresponding treatment. Users can see the relationship between actions and cases according to the decision table, which is much better than the complex nesting in a hierarchical language. And the flat listing of all possible situations can take into account all situations as much as possible without omissions due to logical nesting, especially in the case of if-then-else structures where the else part is optional. Because of the importance of logical control in programming, decision tables are a very important tool in designing logical control.

After understanding the concepts related to decision tables, the principle of discretization is first determined. For a certain conditional attribute, for the same attribute value if divided into different intervals, there will be contradictions in the prediction, which will affect the prediction. For attribute values that are different are determined based on the decision attributes. Based on this the principle of classification of intervals in the discretization process is:

(1) Firstly, the set of C must be divided into the same interval for a condition attribute with the same value;

(2) Secondly, those with different values of the condition attributes merge intervals according to the values of the decision attributes in D, and finally the discrete intervals are determined to complete the discretization process.

The specific steps of discretization are described as follows.

(1) Select the conditional attributes to be discretized in Set C, and arrange its attribute values in the order from smallest to largest.

(2) Find the first breakpoint where the values of the conditional attributes are different and the first occurrence of the decision attribute is different to divide the conditional attributes into two intervals. For example, the whole data set has 50 instances, which are first sorted according to the size of the value of the attribute to be discretized, and after sorting, assume that the decision attributes of the first 12 instances are all 1, the decision attribute of the 13th instance is non-1, and the conditional attributes of the 12th and 13th samples have different values, in which case the first breakpoint is inserted between the 12th and 13th instances.

(3) By analogy, the rest of the samples are divided into intervals, and the neighboring ones with the same decision attribute are divided into the same interval, and it is important to note that the conditional attributes are divided into the same interval when they take the same value. After this step has been carried out then the conditional attributes are divided into a number of intervals.

(4) Combination of intervals:

First of all, the definition gives the concept of interval category and secondary intervals.

Interval category: We define the decision attribute with the largest proportion in an interval as the interval category of this interval.

Secondary interval: when the number of instances in an interval is less than a certain value (which can depend on the specifics of the dataset), such as 3, then this interval is called a secondary interval.

Then the merging rules are: (1) Under the premise of keeping the ordering of continuous attributes unchanged, if the categories of two neighboring intervals are the same then two intervals can be merged. (2) When the interval categories of the left and right neighboring intervals of the secondary interval are the same, the three intervals can be merged, and the category of new intervals will be determined by the decision-making attribute with the largest number of instances (which is obviously the category of the interval of the original neighboring intervals).

(5) Merge consecutive attributes as described above.

(6) Assign interval numbers to all intervals to produce the final discretized interval.

The flowchart of the discretization algorithm is shown in Fig. 2.

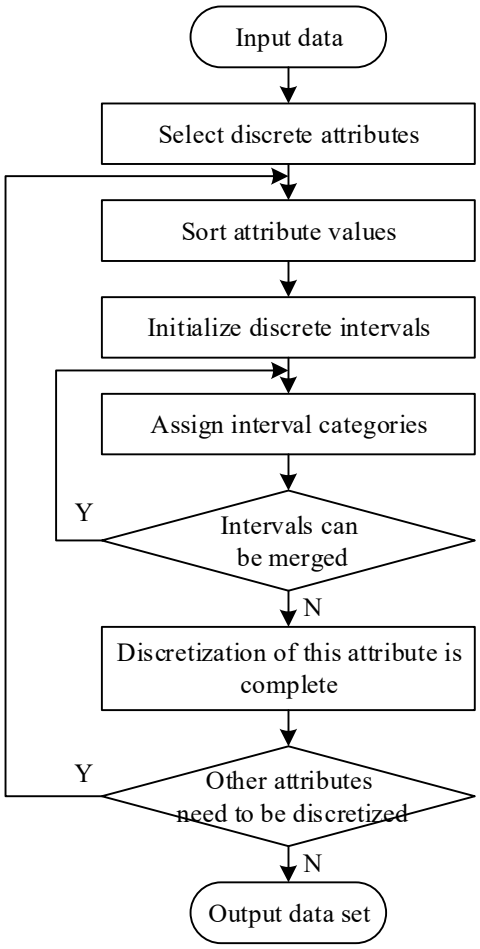


Figure 2. The flowchart of discrete algorithm

This algorithm takes into account the essence of the discretization algorithm while being simple and easy to understand, and divides the discrete intervals under the premise of ensuring that the decision attributes remain basically unchanged, and the number of intervals is moderate, so that neither the category information is lost too much, nor the decision tree is too large.

3. Application of RS-ID3 algorithm in mental health education in colleges and universities

The essence of the new quality of productivity is the enhancement of man's ability to recognize and transform the reality of nature, so it is particularly important to strengthen the thinking of big data and update the concept of learning in a timely manner, to reshape the thinking pattern to cope with the changes in the social situation, and to make education keep pace with the times. Technological change and innovation should be closely centered on the implementation of the main principle of better service to students, the traditional mental health education for college students due to excessive reliance on experience there is a great lag, which leads to sudden psychological crisis can not be effectively curbed, the physical and mental college students caused by a number of adverse consequences, awareness is the precursor to behavior, the only way to establish a big data thinking, college students' mental health education can better Utilizing big data empowers students to grow healthily during their school years. Establishing the big data thinking of college students' mental health workers mainly includes two aspects: on the one hand, colleges and universities should vigorously advocate and strengthen the publicity and education of big data, and utilize a variety of communication channels to impart the knowledge of big data to relevant educators, so that they can update their cognitive structure of data and form the inertia thinking of big data, and can better grasp the technology of big data; on the other hand, the concept of applying big data should be actively promoted and take root, so that relevant staff can better utilize big data to empower students' health growth during their school years. On the other hand, the concept of big data application should be actively promoted to take root, so that relevant staff

put it into practice and constantly summarize the lessons learned, formulate diversified safeguards to promote the active use of big data technology by mental health educators in colleges and universities to innovate the education and teaching methods, and gradually create the environment and atmosphere of big data application, so as to guard against the old-fashioned and convenience-oriented ideas.

Based on the new quality productivity environment, this paper introduces the ID3 algorithm into the study of mental health education in colleges and universities.

3.1. Data preparation

3.1.1. Data selection

This paper adopts the Symptom Self-assessment Scale (SCL-90), the data obtained by testing the mental health status of students in the class of 2022 in a university, 692 questionnaires were sent out, of which 648 were valid questionnaires, with a validity rate of 93.6%. The questionnaire was divided into two parts, the first part set 11 questions and items, the student's number, name, gender, and grade majors and other information was collected, aimed at investigating the basic information of the research object, as well as screening the effective questionnaire. The second part is the main content of the questionnaire, which aims to investigate the students' mental health status, classify the symptoms of psychological problems into 10 dimensions such as somatization, obsessive-compulsive symptoms, interpersonal sensitivity, depression, anxiety, hostility, horror, paranoia, psychoticism, and other, as well as to collect diagnostic results.

3.1.2. Data pre-processing

The purpose of data mining is to extract some valuable knowledge or information from daily business data, but the actual database is too large and is very susceptible to the intrusion of noisy data, vacant data and inconsistent data, which brings great inconvenience to the subsequent data analysis and data mining, and even leads to wrong conclusions, so it is necessary to preprocess the data. Experience has shown that only through careful data preparation in the early stage can we save the time of mining, improve the efficiency of mining, and get high-quality mining results in the process of data digging and shaking.

There are a variety of common data preprocessing methods, including data extraction, data cleansing, data integration, data transformation and so on.

(1) Data Extraction

In the process of data mining generally do not need to use all the data, some data on the construction of data models do not interfere much, some data will reduce the efficiency of mining calculations, and may even lead to fallacies, and these data do not have any benefit for the final data analysis, and will not affect the correct conclusions obtained. Therefore, according to the defined project tasks, the required data sources are identified, data are collected and extracted from them, data attribute features are found, and data size is reduced, so that the amount of data can be streamlined as much as possible without affecting the data analysis, and the implicit laws and inner connections between the data can be easily mined out.

There are a lot of attributes in the collected mental health test data of college students, and some of these attributes do not have much relationship with the digging task, or the data itself does not have the significance of digging. For example, attributes such as student's academic number, name, ID number, etc., the values of these attributes are unique and the amount of data is large, which will only increase the time and space for digging calculation, and can be deleted directly. In addition, attributes such as ethnicity, year of birth, student category, because more than 80% of the selected data are Han Chinese students, all of them are class 2022, the category is too centralized, there is no significance of classification, and it does not have much impact on the results of digging, so in the extraction of the data, also Lei has to clear this part of the attribute value, so that the data size can be reduced. According to the characteristics of the mental health data of college students, after data extraction, the basic attributes of students related to the digging task were determined to be gender, only child, whether two parents, family location, and ten psychological symptoms such as somatization, obsessive-compulsive symptoms, interpersonal sensitivity, depression, anxiety, hostility, terror, paranoia, psychoticism, and other psychological symptoms were mined, respectively.

(2) Data Cleaning

The purpose of data cleaning is to clean up the data by detecting the errors and inconsistencies present in the data, including null value processing, noise processing and inconsistent data processing. The dataset used in this paper has been preliminarily screened for the validity of the data when the symptom self-assessment scale was retrieved, and the test data with large deficiencies were eliminated.

Therefore, this part of data cleaning mainly focuses on further screening for incompleteness and inconsistency in the data to ensure the validity of the data. In the process of psychological testing, a large amount of noise data can be generated due to incomplete data caused by students' irregular filling, wrong filling, or other factors interfering with the data. These erroneous data and null data, duplicate data need to be data cleaned by preprocessing methods. After the cleaning of data null values and inconsistencies, the quality of the data obtained has been greatly improved to ensure the accuracy and effectiveness of the data mining results.

(3) Data conversion

The basic principle of data conversion is to discretize continuous data and categorize discrete data. Some data in the source data is continuous type, there are some data is discrete type, but its classification is too much, such as family location attribute value dispersed widely, professional attribute value has more than one, etc., is not conducive to data mining, must be converted. In addition, in order to improve the efficiency of mining, it is necessary to standardize the data format, Chinese attribute values can be replaced by some English characters or digital numbering. Some of the data converted codes are shown in Table 1.

Table 1. Mental mining data conversion code

Attribute	Attribute value	Code
Gender	Male	XB1
	Female	XB2
Only child	Yes	DS1
	No	DS2
Parents or not	Yes	SQ1
	No	SQ2
Location of family	City	JT1
	Town	JT2
Somatization	Symptomatic(factor≥3)	QT1
	Asymptomatic (factor<3)	QT2
Obsessive-compulsive symptoms	Symptomatic(factor≥3)	QP1
	Asymptomatic (factor<3)	QP2
...	...	...

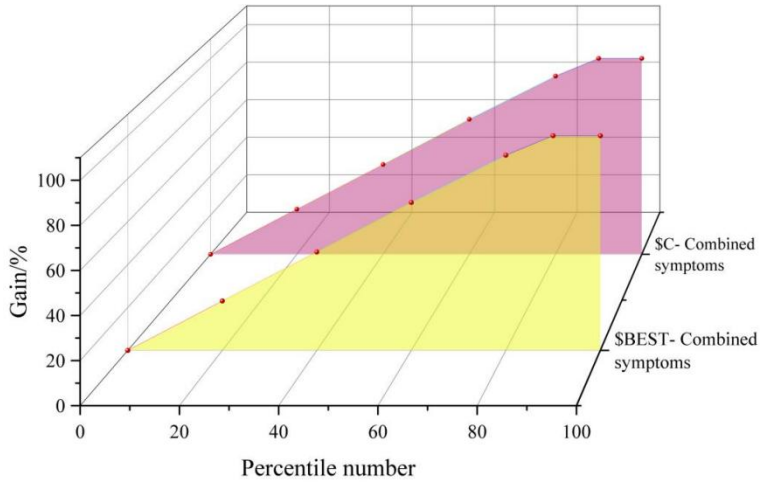
3.2. Constructing mental health prediction model based on RS-ID3 algorithm

In this study, the RS-ID3 algorithm is used to data mine the mental health test data of college students, and by constructing a decision tree and forming classification rules, it provides decision support information for college counseling centers or student administrators, which makes college students' mental health education and counseling work more targeted and directional.

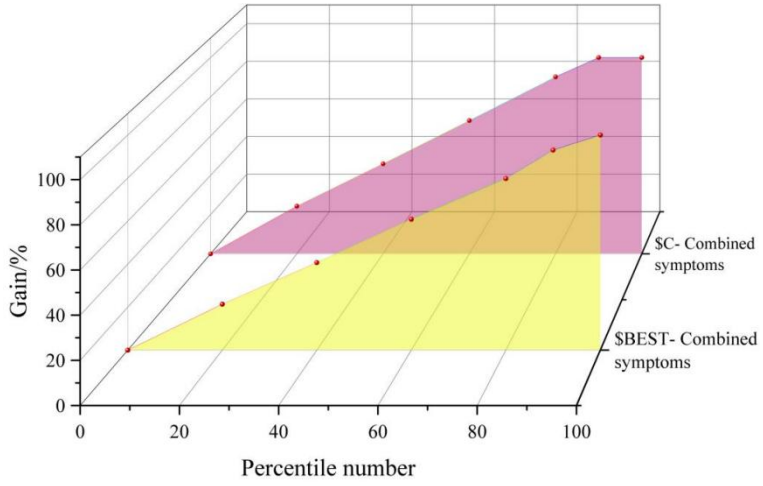
By randomly selecting two-thirds of the data volume in the data set after data preprocessing, roughly 432 records are used as the training sample set, and the remaining 216 records are used as the test sample set.

The assessment plot demonstrates how the college student mental health prediction model performs in predicting certain outcomes by categorizing records based on predictive value and confidence in the prediction, dividing the records into equally sized groups, and then plotting business-standard values of the variables for each quartile from highest to lowest. This is specified below:

(1) Gain evaluation. The gain evaluation of the college students' mental health prediction model is shown in Figure 3, and Figure 3(a) and (b) are the gain trends on the training set and the test set, respectively, and the gain is defined as the percentage of the number of successful records at each quantile to the total number of successful records. According to the gain evaluation results of the mental health prediction model of college students, the trend of "$BEST-syndrome" and "$C-syndrome" in the training set were completely consistent, both of which increased sharply from 0% to 100% in the former part, and remained at the same level in the latter part. The trend of "$BEST-syndrome" and "$C-syndrome" in the test set were basically consistent, which was mainly reflected in the fact that the "$C-syndrome" in the first part reached 100% faster than the "$BEST-syndrome", and the latter part remained at the level. Therefore, the gain trends of the training set and the test set are consistent with the ideal trend, and on the whole, the prediction model of college students' mental health has achieved a large gain effect.



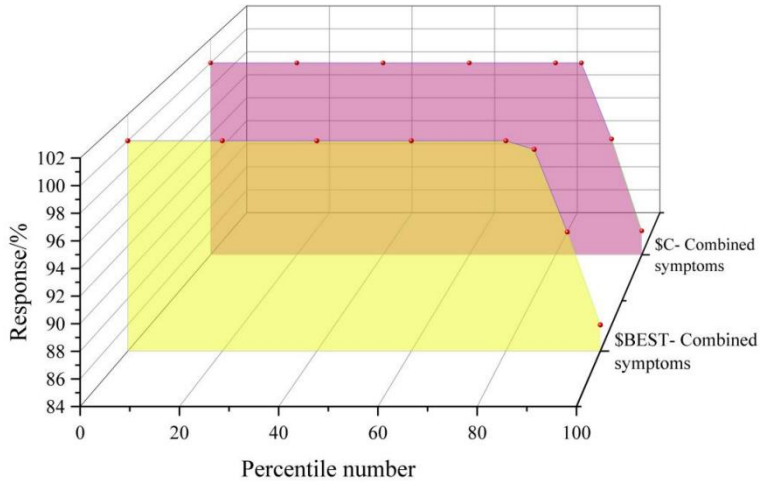
(a)Training set



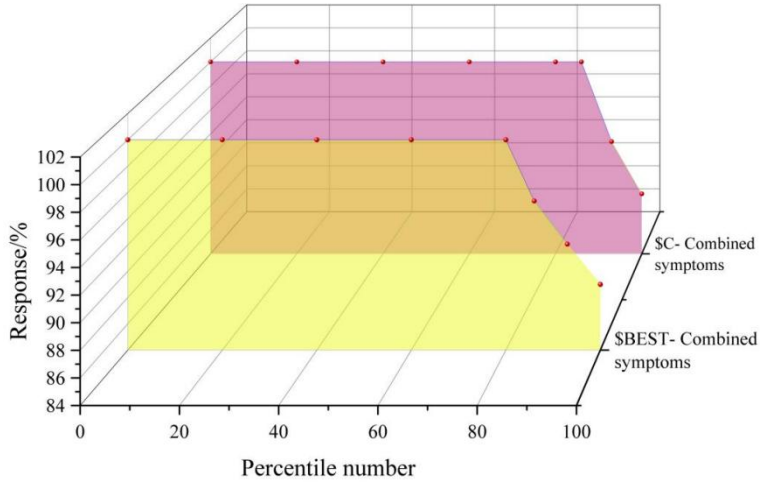
(b)Test set

Figure 3. Gain assessment of mental health prediction model for college students

(2) Response evaluation. The response evaluation of the college students' mental health prediction model is shown in Figure 4, and Figure 4 (a) and (b) are the response trends on the training set and the test set, respectively, and the response is the percentage of the number of successes at the quantile to the number of records at the quantile level. According to the response evaluation results of the mental health prediction model of college students, the trend of "$BEST-syndrome" and "$C-syndrome" in the training set are basically consistent, and they both start to maintain the level from 100% of the first part, and then drop sharply to a certain percentage of the second part. Although the percentile at the beginning of the decline of the "$C-syndrome" was larger than the percentile at the beginning of the decline of the "$BEST-syndrome", it quickly coincided with the response trend of the ideal model. The trend of "$BEST-syndrome" and "$C-syndrome" in the test set were also basically consistent, and the main difference was that the percentile of "$BEST-syndrome" when it began to decline was smaller than that of "$C-syndrome" when it began to decline, and there was a rapid decline. In addition, the percentage of responses corresponding to the test set's "$BEST-syndrome" and "$C-syndrome" when they were at their lowest point was higher than that of the training set. Therefore, the gain trends of the training set and the test set are basically consistent with the ideal trend, and on the whole, the prediction model of college students' mental health has achieved a large response effect.



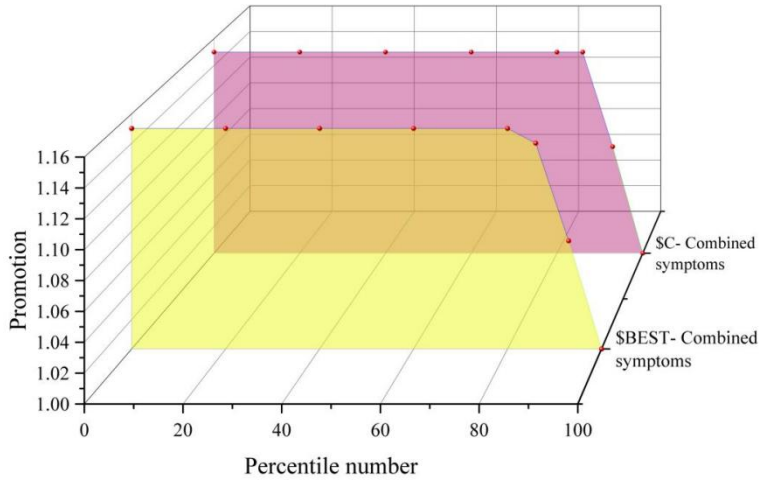
(a)Training set



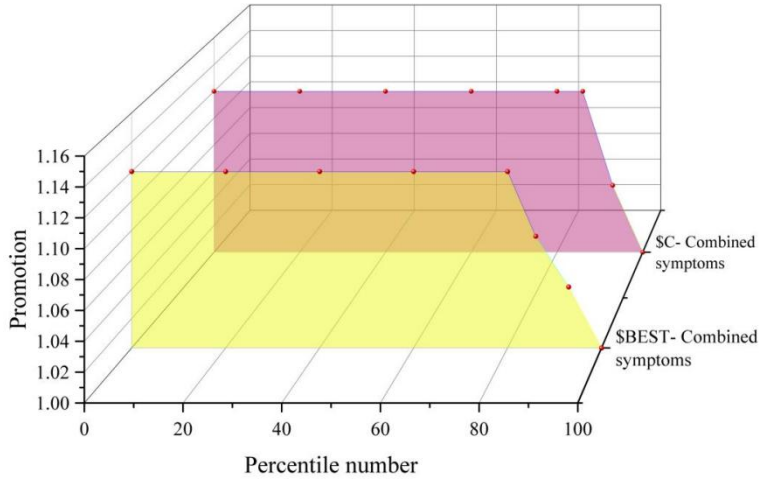
(b)Test set

Figure 4. Response evaluation of mental health prediction model for college students

(3) Improvement assessment. Figure 5 (a) and (b) show the improvement trend of the training set and the test set, respectively, and compare the percentage of successes at each quantile with the percentage of successes in the training data. According to the improvement evaluation results of the mental health prediction model of college students, the trend of "$BEST-syndrome" and "$C-syndrome" in the training set are basically consistent, and the starting point is about 1.15, which is higher than 1.0 as the starting point to maintain the level first, and then rapidly decrease to 1.0. Although the percentile at the beginning of the decline of the "$C-syndrome" was larger than the percentile at the beginning of the decline of the "$BEST-syndrome", it quickly coincided with the response trend of the ideal model. The test set "$BEST-syndrome" and "$C-syndrome" were also basically consistent, and the main difference was that the starting point was lower than that of the training set, with a starting value of about 1.12, and the percentile of "$BEST-syndrome" when it began to decline was smaller than that of "$C-syndrome", and there was a rapid decline. Therefore, the improvement trend of the training set and the test set is basically consistent with the ideal trend, and the evaluated model conforms to the ideal improvement trend of "$BEST-syndrome" in the two partitioned datasets.



(a)Training set



(b)Test set

Figure 5. Improving the prediction model of college students' mental health

3.3. Comparative analysis of different decision trees

3.3.1. Constructing decision trees

The attributes in the dataset selected in this paper all have only two attribute values, so it circumvents the shortcomings of ID3 algorithm that tends to select multi-attribute-valued attributes, ID3 algorithm is a traditional decision tree construction algorithm, which selects the optimal splitting node by calculating the information gain of each attribute and then comparing the size.

The steps for constructing a decision tree model of a student with or without obsessive-compulsive symptoms using the ID3 algorithm are as follows:

First the information gain of each split attribute in the training sample set is calculated using equations (6), (7), (8).

Then the information gains generated according to different split attributes are compared, and the attribute with the largest information gain is set as the root node of the decision tree, and then according to the number of values of this attribute in this root node it is determined how many branches are generated downward, and it is also determined how many subdatasets to divide the total training sample into.

Finally, the first two steps are performed recursively in a loop on each of the split subdatasets until the leaf nodes of the spanning tree or no attributes are available to continue the split, then it ends.

According to Table 1 we can learn that the attribute QP (obsessive-compulsive symptom) has two different values: 0 (no symptom), 1 (symptom). Therefore, the training sample set can be divided into two categories, in which there are 116 samples with the value of “1” and 316 samples with the value of

"0". The expected information i.e. entropy of the training sample D categorization is calculated according to equation (5) as:

$$E(A) = -\frac{116}{432}\log_2\frac{116}{432} - \frac{316}{432}\log_2\frac{316}{432} = 0.83933076 \tag{9}$$

Next, the entropy of each split attribute needs to be calculated. Taking the gender attribute as an example, there are 184 entries for gender male (XB1), of which 46 are symptomatic and 138 are asymptomatic, and 248 entries for gender female (XB0), of which 70 are symptomatic and 178 are asymptomatic. The expected information calculated by dividing the samples in D based on their gender is:

$$\begin{aligned} E_{XB}(A) &= \frac{184}{432}\left(-\frac{46}{184}\log_2\frac{46}{184} - \frac{138}{184}\log_2\frac{138}{184}\right) \\ &+ \frac{248}{432}\left(-\frac{70}{248}\log_2\frac{70}{248} - \frac{178}{248}\log_2\frac{178}{248}\right) = 0.83839235 \end{aligned} \tag{10}$$

The information gain of XB (gender) can be calculated based on equation (6) as:

$$gain(XB) = E(A) - E_{XB}(A) = 0.00093841 \tag{11}$$

By analogy, the information gain results for the other attributes are calculated:

$$gain(DS) = E(A) - E_{DS}(A) = 0.00091772 \tag{12}$$

$$gain(SQ) = E(A) - E_{SQ}(A) = 0.00094565 \tag{13}$$

$$gain(JT) = E(A) - E_{JT}(A) = 0.00090546 \tag{14}$$

At this point, the information gains of all attributes are computed, and the attribute with the largest information gain is whether or not it is biparental (SQ), so it is used as the root node for splitting.

The calculation of the second level of split nodes was performed next:

There are 379 who are biparental, 92 who are symptomatic, and 287 who are asymptomatic, and the expected information i.e. entropy of the training sample D classification is calculated according to Eq. (5) as:

$$E(A) = -\frac{92}{379}\log_2\frac{92}{379} - \frac{287}{379}\log_2\frac{287}{379} = 0.79957386 \tag{15}$$

There were 154 entries for gender male (XB1), of which 35 were symptomatic and 119 asymptomatic, and 225 entries for gender female (XB0), of which 57 were symptomatic and 168 asymptomatic. The expected information calculated by dividing the samples in D based on their gender is:

$$\begin{aligned} E_{XB}(A) &= \frac{154}{379}\left(-\frac{35}{154}\log_2\frac{35}{154} - \frac{119}{154}\log_2\frac{119}{154}\right) \\ &+ \frac{225}{379}\left(-\frac{57}{225}\log_2\frac{57}{225} - \frac{168}{225}\log_2\frac{168}{225}\right) = 0.79867895 \end{aligned} \tag{16}$$

The information gain of XB (gender) can be calculated based on equation (6) as:

$$gain(XB) = E(A) - E_{XB}(A) = 0.00089491 \tag{17}$$

By analogy, the information gain results for the other attributes are calculated:

$$gain(DS) = E(A) - E_{DS}(A) = 0.00088761 \tag{18}$$

$$gain(JT) = E(A) - E_{JT}(A) = 0.00090216 \tag{19}$$

One of the attributes with the largest information gain is the home location (JT), so it is used as a split node for the second layer with the samples that are two parents.

There are 53 people who are not two parents, 24 people who have symptoms and 29 people who are asymptomatic, and the expected information i.e. entropy for the classification of training sample D is

calculated according to equation (5):

$$E(A) = -\frac{24}{53}\log_2\frac{24}{53} - \frac{29}{53}\log_2\frac{29}{53} = 0.99357048 \tag{20}$$

There were 30 entries for gender male (XB1), of which 11 were symptomatic and 19 asymptomatic, and 23 entries for gender female (XB0), of which 13 were symptomatic and 10 asymptomatic. The expected information calculated by dividing the samples in D based on their gender is:

$$\begin{aligned} E_{XB}(A) &= \frac{30}{53}\left(-\frac{11}{30}\log_2\frac{11}{30} - \frac{19}{30}\log_2\frac{19}{30}\right) \\ &+ \frac{23}{53}\left(-\frac{13}{23}\log_2\frac{13}{23} - \frac{10}{23}\log_2\frac{10}{23}\right) = 0.99267544 \end{aligned} \tag{21}$$

The information gain of XB (gender) can be calculated based on equation (6) as:

$$gain(XB) = E(A) - E_{XB}(A) = 0.00089504 \tag{22}$$

By analogy, the information gain results for the other attributes are calculated:

$$gain(DS) = E(A) - E_{DS}(A) = 0.00088853 \tag{23}$$

$$gain(JT) = E(A) - E_{JT}(A) = 0.00089012 \tag{24}$$

Among them, the attribute with the largest information gain is gender (XB), so it is used as the split node for the second layer of samples that are not biparental.

By the same token, the remaining samples from the root node are computed layer by layer to generate the corresponding split node is also the same method step, which will not be repeated here, after the calculation, the ID3 classification decision tree for obsessive-compulsive symptoms, for example, is shown in Figure 6.

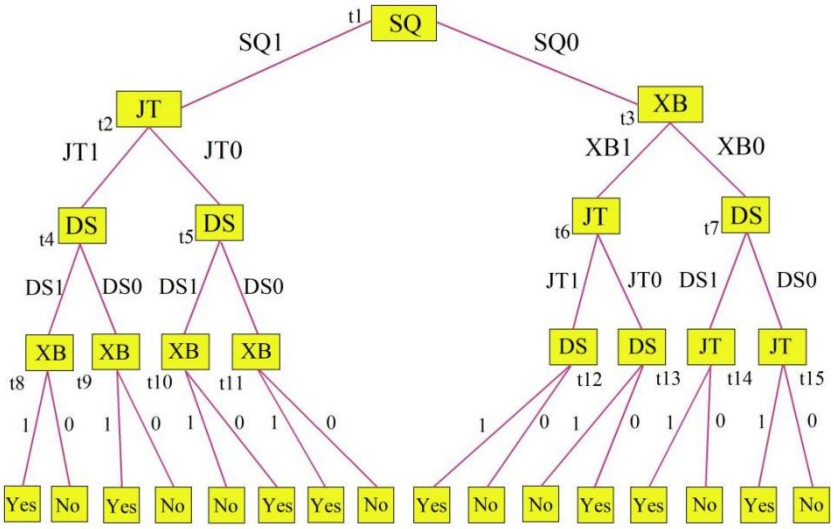


Figure 6. ID3 decision tree based on obsessive-compulsive symptoms

In the process of constructing the decision tree is to use the training set data, the decision tree constructed is more complex and luxuriant, in order to improve the classification efficiency and readability of the decision tree, this paper adopts the RS-ID3 algorithm for optimization. Also taking the obsessive-compulsive symptoms as an example, the decision tree model based on the RS-ID3 algorithm is shown in Figure 7.

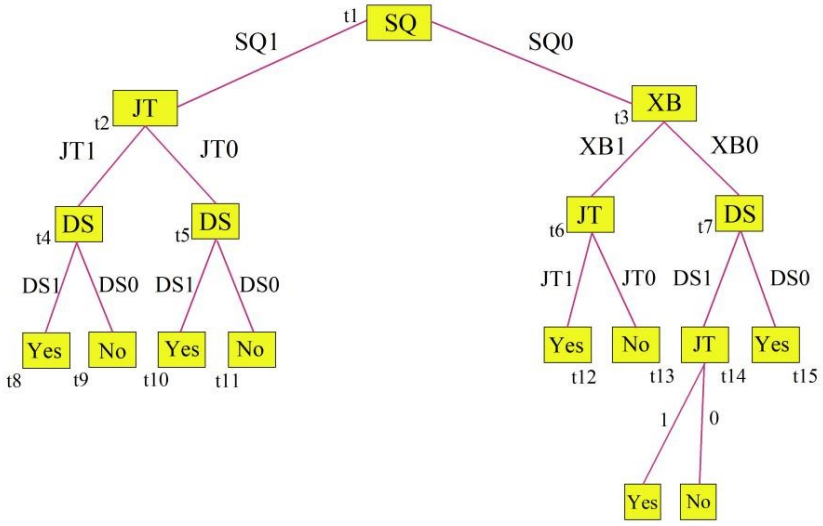


Figure 7. RS-ID3 decision tree based on obsessive-compulsive symptoms

3.3.2. Classification rule extraction

Some of the rules generated by the decision tree model using the ID3 algorithm are.

IF SQ=SQ1 AND JT=JT1 AND DS=DS1 XB=XB1 THEN yes;

IF SQ=SQ1 AND JT=JT0 AND DS=DS1 XB=XB0 THEN no;

.....

The rule corresponding to reducing the code to the original field meaning is.

IF is two parents AND household location is urban AND is an only child AND sex male THEN yes;

IF is two parents AND household location is urban AND is an only child AND sex female THEN none;

.....

The decision tree model using RS-ID3 algorithm produces rules that are:

IF SQ=SQ1 AND JT=JT1 AND DS=DS1 THEN yes;

IF SQ=SQ1 AND JT=JT1 AND DS=DS0 THEN no;

IF SQ=SQ1 AND JT=JT0 AND DS=DS1 THEN yes;

IF SQ=SQ1 AND JT=JT0 AND DS=DS0 THEN none;

IF SQ=SQ0 AND XB=XB1 AND JT=JT1 THEN yes;

IF SQ=SQ0 AND XB=XB1 AND JT=JT0 THEN none;

IF SQ=SQ0 AND XB=XB0 AND DS=DS0 THEN yes;

IF SQ=SQ0 AND XB=XB0 AND DS=DS1 AND JT=JT1 THEN yes;

IF SQ=SQ0 AND XB=XB0 AND DS=DS1 AND JT=JT0 THEN no;

.....

Reduce the above “with symptoms” classification rule code to the corresponding text field information as:

IF two parents AND household location is urban AND only child THEN yes;

IF both parents AND household location is rural AND not an only child THEN none;

.....

3.3.3. Analysis of mining results

From the above generated rules, it can be analyzed and concluded that the attribute of whether the students are two parents has some relationship with the emergence of obsessive-compulsive symptoms; children who are not two parents are more likely to have obsessive-compulsive symptoms, so that special attention should be given to those children from single-parent families. In terms of gender attribute, students whose gender is female are more likely to develop obsessive-compulsive symptoms; students who are not two parents and only children are more likely to develop symptoms.

In this study, the remaining one-third of the data was applied to the resulting decision tree classification model, and its accuracy was verified; the status of obsessive-compulsive symptoms was already known in the remaining data, and the classification prediction of the test set using the decision tree of the ID3 algorithm compared the already existing categories with the predicted classification results, with an accuracy of 72.1%; and the classification prediction of the test data using the decision

tree of the RS-ID3 algorithm set is compared with the known categories with an accuracy of 80.6%. The accuracy of the classification prediction of the decision tree model using the ID3 algorithm is found to be lower than that of the classification model using the RS-ID3 algorithm, so it can be seen that using the RS-ID3 decision tree algorithm to construct decision trees for classification of psychological assessment data in this way of classification and mining the results of psychological prevention and intervention is of some reference value.

4. Conclusion

(1) The results of gain evaluation showed that the trend of "$BEST-syndrome" and "$C-syndrome" in the training set were completely consistent, while the "$C-syndrome" in the first part of the test set reached 100% faster than that in "$BEST-syndrome", and the latter part remained at the level. The response evaluation results show that the trend of "$BEST-syndrome" and "$C-syndrome" in the training set are basically consistent, while the corresponding response percentages of "$BEST-syndrome" and "$C-syndrome" in the test set are higher than those in the training set when they fall to the lowest point. The improvement evaluation results show that the trend of "$BEST-syndrome" and "$C-syndrome" in the training set are basically consistent, with a starting point of about 1.15, while the "$BEST-syndrome" and "$C-syndrome" in the test set are also basically consistent, with a starting value of about 1.12.

(2) Respectively using ID3 algorithm, RS-ID3 algorithm decision tree for classification prediction of the test set, the already existing categories and the predicted classification results are compared, the accuracy rate of 72.1%, 80.6% respectively. It can be seen that the use of RS-ID3 decision tree algorithm for classification to build decision trees this way of psychological assessment data classification mining results for psychological prevention and intervention is a certain reference value.

(3) From the above rules of decision tree generation, it can be analyzed that the attribute of whether students have two parents has a certain relationship with the emergence of obsessive-compulsive symptoms, and children who are not parents are more likely to have obsessive-compulsive symptoms, so that children from single-parent families should be given special attention. From the gender attribute, students whose gender is female are more likely to develop obsessive-compulsive symptoms; students who are not two parents and only children are more likely to develop symptoms.

References

1. Classen, B., Tudor, K., Johnson, F., & McKenna, B. (2021). Embedding lived experience expertise across the mental health tertiary education sector: An integrative review in the context of Aotearoa New Zealand. Journal of Psychiatric and Mental Health Nursing, 28(6), 1140-1152.
2. Barrable, A., Papadatou-Pastou, M., & Tzotzoli, P. (2018). Supporting mental health, wellbeing and study skills in Higher Education: an online intervention system. International Journal of Mental Health Systems, 12, 1-9.
3. Nurunnabi, M., Almusharraf, N., & Aldeghaither, D. (2020). Mental health and well-being during the COVID-19 pandemic in higher education: Evidence from G20 countries. Journal of Public Health Research, 9(1_suppl), jphr-2020.
4. Yang, X. H., Yu, H. J., Liu, M. W., Zhang, J., Tang, B. W., Yuan, S., ... & He, Q. Q. (2020). The impact of a health education intervention on health behaviors and mental health among Chinese college students. Journal of American College Health, 68(6), 587-592.
5. Wang, T., & Park, J. (2021). Design and implementation of intelligent sports training system for college students' mental health education. Frontiers in psychology, 12, 634978.
6. Guo, T., Zhao, W., Alrashoud, M., Tolba, A., Firmin, S., & Xia, F. (2022). Multimodal educational data fusion for students' mental health detection. IEEE Access, 10, 70370-70382.
7. Kamran Ul haq, A., Khattak, A., Jamil, N., Naeem, M. A., & Mirza, F. (2020). Data analytics in mental healthcare. Scientific Programming, 2020(1), 2024160.
8. Guo, Y., Zhong, N., Li, X., & Shi, Y. (2024). Mental Health and Perceived Social Support of Local College Teachers in the Context of Big Data. International Journal of Multiphysics, 18(3).
9. Long, J., & Lin, J. (2024). Empowering English language learning and mental health using AI and Big data. Education and Information Technologies, 29(10), 12703-12734.
10. Liang, Y., Zheng, X., & Zeng, D. D. (2019). A survey on big data-driven digital phenotyping of mental health. Information Fusion, 52, 290-307.
11. Hidalgo-Mazzei, D., Murru, A., Reinares, M., Vieta, E., & Colom, F. (2016). Big data in mental health: a challenging fragmented future. World Psychiatry, 15(2), 186.
12. Conglin, C., Lin, L., Yi, L., & Lei, T. (2020, April). A Study on College Students' Mental Health Education and Early Warning Mechanism Based on Big Data. In Proceedings of the 2020 3rd International Conference on Big Data and Education (pp. 1-4).
13. Zhang, Z. (2024). Early warning model of adolescent mental health based on big data and machine learning. Soft Computing, 28(1), 811-828.

14. Simon, G. E. (2019). Big data from health records in mental health care: hardly clairvoyant but already useful. JAMA psychiatry, 76(4), 349-350.
15. Rosenfeld, A., Benrimoh, D., Armstrong, C., Mirchi, N., Langlois-Therrien, T., Rollins, C., ... & Yaniv-Rosenfeld, A. (2021). Big Data analytics and artificial intelligence in mental healthcare. In Applications of big data in healthcare (pp. 137-171). Academic Press.
16. Rubeis, G. (2022). iHealth: The ethics of artificial intelligence and big data in mental healthcare. Internet Interventions, 28, 100518.
17. Chekroud, A. M. (2017). Bigger data, harder questions—opportunities throughout mental health care. JAMA psychiatry, 74(12), 1183-1184.
18. Liang, L., Zheng, Y., Ge, Q., & Zhang, F. (2022). Exploration and strategy analysis of mental health education for students in sports majors in the era of artificial intelligence. Frontiers in Psychology, 12, 762725.
19. Ouyang, X. (2024). Application and Effectiveness Assessment of Big Data Analysis Algorithm in College Students' Mental Health Education. Journal of Electrical Systems, 20(9s), 491-497.
20. Zhang, X., & Jia, S. (2021, April). The ways of college mental health education based on big data. In Journal of Physics: Conference Series (Vol. 1852, No. 3, p. 032030). IOP Publishing.
21. Liu, Q., & Liao, X. (2021, April). Research on university mental health education based on computer big data statistical analysis. In 2021 2nd International Conference on Big Data and Informatization Education (ICBDIE) (pp. 29-34). IEEE.
22. Lun, G., & Meng, Q. (2018). Problems and countermeasures of mental health education of college students under the background of big data. Educational Sciences: Theory & Practice, 18(6).
23. Li, W. (2020, October). Application of big data technology in college students' mental health education innovation. In Journal of Physics: Conference Series (Vol. 1648, No. 4, p. 042069). IOP Publishing.
24. Xie, W. (2021, January). Big data Analysis on The Management Content of College Students' Mental Health Education. In 2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) (pp. 662-665). IEEE.
25. Jia, Y. (2024). Impact of Music Teaching on Student Mental Health Using IoT, Recurrent Neural Networks, and Big Data Analytics. Mobile Networks and Applications, 1-20.