Region-Educational Research and Reviews, 2025,Volume 7, Issue 2
DOI:10.12238/rerr.v7i2.3487
ISSN Online:2661-4634
ISSN Print:2661-4626

Analysis of College Students' Mental Health Based on XGBoost

Yuling Liu, Can Hou, Shaoyong Hong

School of Artificial Intelligence, Guangzhou Huashang College, Guangzhou, Guangdong, China

Abstract: With the rapid development of social economy and the change of students' growth environment, students' mental health problems have become increasingly prominent. In order to comprehensively and systematically understand the psychological stress of college students. In this paper, data mining technology and machine learning methods are used to intelligently evaluate the mental health of college students. By collecting relevant data and combining with SCL-90 mental health evaluation system, the machine learning algorithm is used to accurately simplify the questionnaire questions of the original evaluation system. And to determine the contribution of different indicators to the psychological stress of college students. Based on the simplified data, a psychological stress evaluation model for college students is established to realize the classification evaluation of mental health status. And provide a specific evaluation score or grade. Improve the accuracy of assessment through model training and optimization, and provide data support for follow-up mental health work. In addition, Kendall's Tau-B correlation analysis and XGBoost are used for feature selection, and key feature factors are retained. The method of this paper is helpful to accurately locate the psychological stress of college students, and provide a clear direction and basis for psychological intervention.

Keywords: student mental health, data mining, scl-90 scale, XGBoost model

Introduction

Promoting students' physical and mental health and all-round development is a major issue of social concern. However, with the rapid development of economy and society, the growth environment of students is constantly changing, and mental health problems are becoming more prominent. Mental health is an important part of human health. The pressure of employment, study and life of college students often seriously affects their mental health. It even leads to a series of negative emotions[1]. Through the evaluation and analysis of the psychological pressure of college students, we can have a comprehensive and systematic understanding of the current psychological pressure situation of college students, and make clear the psychological pressure problems of college students. Help eliminate psychological problems such as anxiety, depression and escape caused by stress[2].

Data mining is a technology that extracts valuable information from a large amount of data automatically or semi-automatically. Has been widely used in various fields[3]. In the field of psychology, data mining technology can be used to evaluate and analyze the psychological stress of college students. To help psychologists and doctors better understand and treat the psychological problems of college students[6]. Through mining and analyzing the data of college students' study, life, social interaction and other aspects, the psychological stress level and influencing factors of college students can be obtained. For example, the learning pressure of college students can be evaluated by analyzing their academic performance, homework completion, examination results and other data[4,5]; The evaluation and analysis of college students' psychological stress based on data mining can provide guidance for the prediction and intervention of

college students' mental health in the futureScientific basis, and then promote its all-round development.

1. Research methods

1.1 Data collection

This study collects data related to college students' mental health through questionnaires and school collection. Including SCL-90 scale scores, daily behavior data and so on.

Basic information questionnaire. The self-designed questionnaire was used to collect the basic background information of college students. Including the gender, grade and other basic content of the respondents. The collection of basic background information is mainly to facilitate the data analysis after the psychological census. As well as targeted tracking of students who are expected to have psychological crisis.

Symptom Checklist 90 (SCL-90). The Symptom Checklist 90 (SCL-90) used in this study is one of the most widely used measures of mental disorders and mental illness. The scale involves 90 items, including 10 factors, corresponding to 10 aspects of psychological symptoms. Evaluate the individual's recent mental state from daily life, emotional changes to thinking consciousness. Each item in the SCL-90 was graded from 1 to 5 (1 = "none"-5 = "severe"). It can clearly and multidimensionally assess whether an individual has certain psychological symptoms and their severity[7].

2667 copies of SCL-90 were collected by data collection personnel in the form of comprehensive survey. There are 2652 valid data. The evaluation data include the basic information of students after desensitization, as well as the answers to the 90-item evaluation scale and the scores of each dimension.

1.2 Data processing

There are some incomplete and invalid noise data in the data collected in this paper. The direct use of this data in the experiment will have a certain impact on the experimental results. And it will also affect the efficiency of the algorithm to a certain extent. In order to ensure ideal experimental results, the following preprocessing operations are performed on the data in this paper[8].

Determine the target of data mining, and keep only the data related to the target. For this study, it is necessary to screen out the data related to this goal from the massive original data. The psychological assessment data is mainly reflected in the specific data of SCL-90 symptom checklist[9].

At the same time, data cleaning is carried out, including missing filling, error detection, duplicate filtering, consistency checking, etc. Missing filling includes zero value filling, mean value filling, probability distribution filling, etc;Error detection refers to checking whether the data conforms to the specification through rules;Duplicate filtering refers to duplicate data generated during data export or system failure. Data need to be culled or consolidated to ensure data uniqueness;Consistency test includes whether the basic information of students is consistent with the data information of students'psychological assessment. Should be handled by rule amendment or manually[5].

After the above data processing, it is found that the data obtained in this paper are of high quality, and there are still 2617 data samples after processing. The original data set is divided into a training set and a test set, and the training data and the test data are divided according to the ratio of 7:There are 2617 sample data, so there are 1832 sample data in the training data set and 784 sample data in the test data set.

The training set undertakes the task of learning and adjusting the model parameters to ensure that the model can be trained effectively. The test set is used to finally evaluate the generalization ability and performance of the model. After the model is trained, the test set, as a data set not involved in the training, can objectively reflect the performance of the model on unknown data. By comparing the prediction results of the model on the test set with the actual data, the performance index of the model in the real application scenario can be obtained.

1.3 Methods and models

This paper is based on machine learning methods and data mining technology. Intelligent psychological stress assessment of college students' mental health was carried out[10], and modeling analysis was carried out with the help of big data. It combines machine learning related algorithms to realize intelligent mental health assessment and early warning,

and through machine learning methods. The Index weight of SCL-90 mental health evaluation system was analyzed. This paper explores the main sources, types and degrees of college students' psychological pressure. Combined with the data collection results of students, the psychological state of students is evaluated, and the risk factors affecting the psychological pressure of college students are identified. Predicting mental health status of college students.

1.3.1 Scale simplification

Through the machine learning algorithm, the original evaluation system is accurately simplified in the questionnaire group. The most representative topic and topic groups are effectively identify, in order to significantly improve the efficiency and accuracy of the evaluation, the machine learning algorithm is used to analyze the weight of the indicators in the SCL-90 evaluation system. To determine the contribution of different indicators to the psychological stress of college students.

Based on the simplified data, a psychological stress evaluation model for college students is established to evaluate the function of mental health status. The evaluation results shall include the classification of psychological state and specific evaluation scores or grades, as shown in Table 1. Through model training and optimization, the accuracy of evaluation is improved, which provides solid data support for the follow-up mental health work. It is helpful for us to locate the psychological pressure of college students more accurately. It provides a clear direction and basis for the follow-up psychological intervention work.

Table 1 Rating standard of average score of items in SCL-90

Classification of mental States	Normal	Mild pressure	Moderate pressure	Emphasis on pressure	Serious
Level	A	B	C	D	E
Limit range	(1,1. 5]	(1. 5,2. 5]	(2. 5,3. 5]	(3. 5,4. 5]	(4. 5,5]

The SCL-90 data table was selected by using Kendall's Tau B correlation analysis and XGBoost. The characteristics of the screening reservation are the same, that is to say, five characteristic factors are reserved: depression, anxiety, mental symptoms, interpersonal sensitivity and obsessive-compulsive symptoms.

For the simplified SCL-90 symptom checklist, the Kendall's Tau-B correlation analysis was used. The Kendall's Tau – B correlation analysis is a method of rank correlation analysis used to assess the correlation between two ordinal categorical variables. It is used to count the number of concordant pairs and discordant pairs to judge the positive or negative correlation between variables.

Correlation analysis was used to analyze the correlation of the ten factors and their corresponding inter-group items, and the correlation coefficient was calculated. A thermodynamic diagram is drawn to identify strong correlations between the variables. Finally, we select and reserve the topics whose correlation coefficient is greater than 50% for subsequent data analysis or modeling work. Here we only show the heat map of depressed mood factors, as shown in Figure 1.

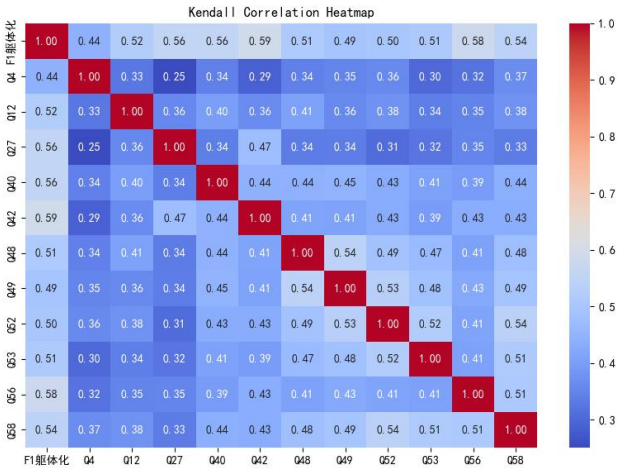

Fig. 1 Thermogram of depressed mood factors

According to the feature screening, the most influential factors in the SCL-90 table can be analyzed. These factors

account for a large proportion in the mental health of college students. Therefore, it is necessary to pay more attention to the performance of depression, anxiety, mental symptoms, interpersonal sensitivity, obsessive-compulsive disorder and so on.

1.3.2 XGBoost model

XGBoost model is a machine learning model based on gradient boosting decision tree algorithm. It optimizes the value of the objective function by integrating multiple decision trees[11], as shown in Figure 2.

In this study, the use of XGBoost to further improve the prediction accuracy optimizes the decision tree integration process[12]. The residual error of the previous model is corrected by iteratively adding a new decision tree, and the parameters of each tree are optimized based on the gradient information of the loss function. Column subsampling and row subsampling techniques are used to further improve the generalization ability of the model. Finally, the output values of all the passing trees are weighted and summed to obtain the final predicted value.

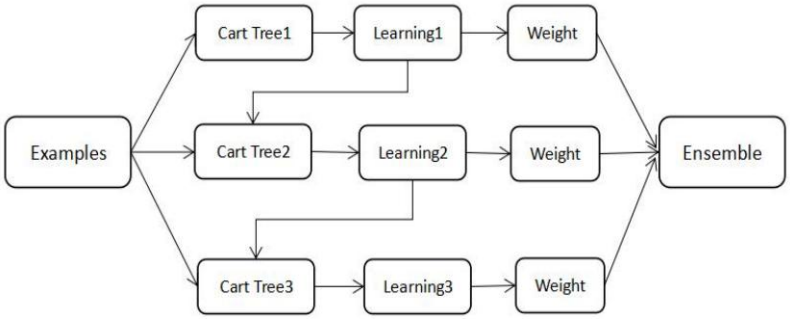

Figure 2 Structure diagram of XGBoost model

The XGBoost model contains several adjustable parameters, which have important effects on the performance and prediction results of the model. As Table 2 shows the parameters of the experimental model.

Table 2 Parameters of XGBoost model

Parameter name	Description	Default value
learning_rate	Learning rate	0. 1
max_depth	Maximum depth of the tree	3
n_estimators	Number of regression trees	100
min_child_w	Minimum sum of weights of child nodes	1
subsample	Proportion of subsamples used to train the model	1
reg_alpha	The parameters of the L1 regularization term	0
reg_lambda	The parameters of the L2 regularization term	1
objective	Define learning tasks and corresponding o	Return
random_state	Random number seed to control randomness	None

In this study, the XGBoost model was used to further improve the accuracy of mental health status prediction. Through the gradient lifting mechanism, XGBoost can predict the weight proportion of different characteristics of mental

health status (such as depression, anxiety, mental symptoms, etc.). situation[13]. Specifically, the model predicts that the harm of depression to mental health is dominant, followed by anxiety and psychiatric symptoms. This finding is of great significance for understanding the influencing factors of mental health status and formulating targeted interventions.

It is still speculated that depression is the most harmful to mental health, followed by anxiety and psychiatric symptoms. See Figure 3 for details.

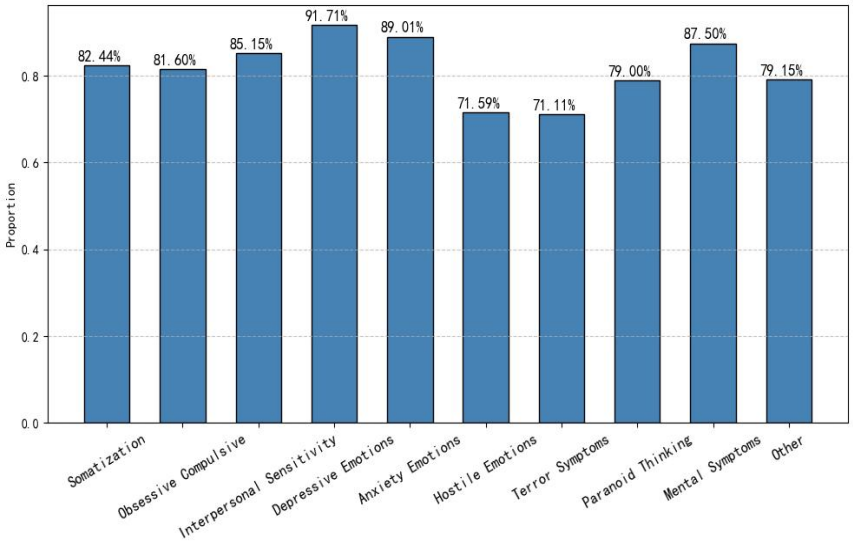

Fig. 3 Weight map of XGBoost feature index

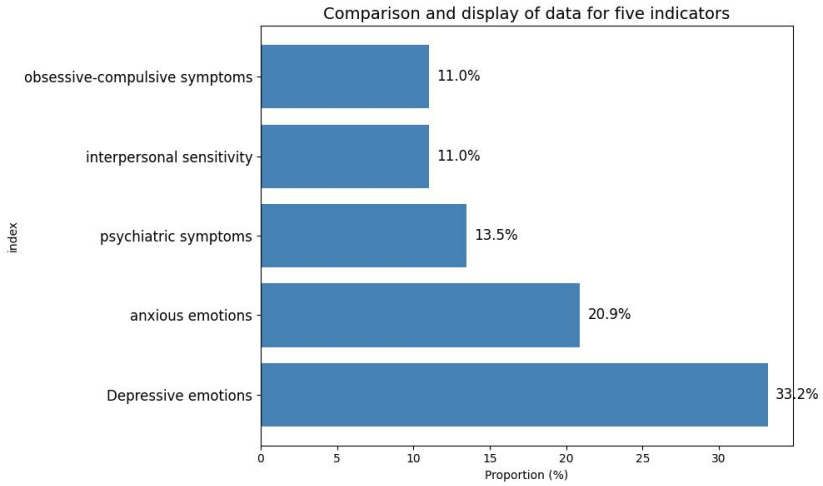

Fig. 4 Proportion of importance of five characteristic factors

The data of the five extracted indicators are compared and displayed. See Figure 4 for Fig. Fig. 4 Proportion of importance of five characteristic factors

details. In the figure, each horizontal column represents an indicator, and the width of the

column visually indicates the specific value of the indicator. By comparing the height of these bars, it can be clearly seen that depression accounts for 33. 2%. The following anxiety indicators accounted for 20. 9%, followed by mental symptoms accounted for 13. 5%. Interpersonal sensitivity and obsessive-compulsive symptoms were about 11%.

2. Results and conclusion

After in-depth training and rigorous verification of the XGBoost model, the XGBoost model shows excellent fitting ability in the training set. The accuracy is high, which fully proves that the model can accurately capture and learn the characteristics and rules of the training data. On the validation set, the performance of the model is also remarkable. Many

key evaluation indicators, such as accuracy, recall and F1 score, are maintained at a high level. This strongly shows that the model not only has excellent learning ability, but also has strong generalization ability. It can still maintain stable performance on unseen data.

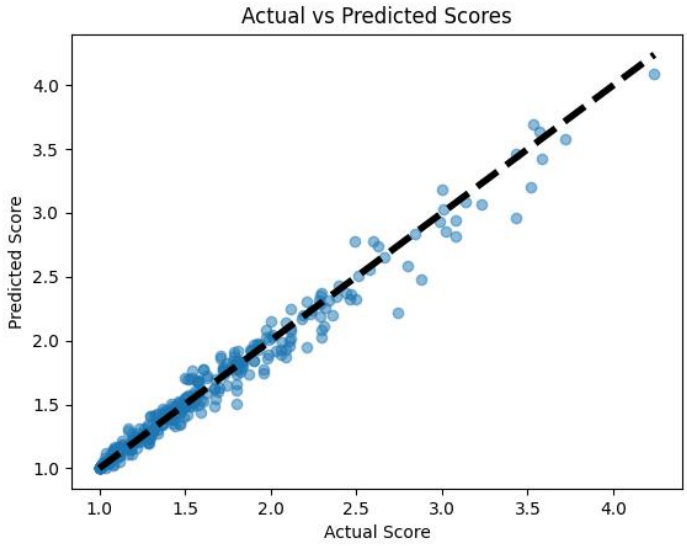

Fig. 5 Comparison between real and predicted XGBoost model

By further inputting the raw data of the SCL-90 table into the XGBoost model, the machine learning classification method is used to train and test many times. Through continuous optimization and adjustment, it finally ensures that the accuracy of the model in the evaluation of early warning is more than 95%. This result not only highlights the high accuracy of the model, but also greatly improves its credibility in practical applications.

Figure 5 Comparison between real and predicted XGBoost model

For the whole test process, we also analyze the performance of the model in detail. From data preprocessing to model training and verification, and then to the final evaluation and early warning. Every step has been strictly controlled and carefully optimized to ensure that the model can play the best performance in practical applications.

We compared the predicted results of the model with the actual values. The model was found to be able to accurately predict mental health status in most cases. The test performance is good in all indicators, the simulation effect is very good, and the expected goal is achieved. The result is shown in Figure 5, which visually demonstrates the relationship between the predicted value of the model and the actual value, further proving the effectiveness of the model.

In terms of accuracy, the Mean Squared Error on the test set of the XGBoost model is 0.0094, the decision Root Mean Squared Error is 0.0971, Mean Absolute Error is 0.0650 and R2 Score is 0.9722, indicating that the model has high prediction accuracy and explanatory power. In terms of effectiveness, the performance in different scenarios is relatively stable, indicating that it has strong adaptability. In terms of stability, the results of repeated tests show that the performance fluctuation is small, which indicates that the model has high stability.

Through the feature importance ranking of the XGBoost model, it was found that depressed mood, anxious mood and psychiatric symptoms dominated in predicting mental health status. Specifically, the weight of depression is the highest, followed by anxiety, followed by mental symptoms. This result is consistent with the expertise in the field of psychology, which further verifies the reliability of the model.

In the training process of the model, the key parameters such as the learning rate, the maximum depth of the tree and the number of regression trees are optimized. The prediction performance of the model is significantly improved, which indicates that when constructing the XGBoost model, Reasonable parameter selection and optimization are essential to improve the performance of the model.

At the same time, the results show that the XGBoost model performs well in the task of predicting mental health

status. It can accurately capture the key features in the data and effectively distinguish different mental health States. Through the analysis of feature importance, it is determined that depression, anxiety and psychiatric symptoms are the key features to predict mental health status. It is found that these characteristics dominate the prediction, which helps us better understand the influencing factors of mental health status. To provide a scientific basis for the development of targeted interventions.

XGBoost model has a wide range of applications, because the model has high efficiency and accuracy in dealing with large-scale data and complex models. Therefore, it can be widely used in mental health assessment, disease prediction, personalized treatment and other fields. In the future, the integration of XGBoost model and other machine learning algorithms can be further explored to improve the prediction performance and expand the application scenarios.

Conflicts of interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Xu Jie, Chen Yongyong, Hu Yongkang. Research on Automatic Rating Method of College Students' Mental Health Based on GA + BP Model [J]. Education Watch, 2022, 11(32): 18-22+37.

[2] Hao Wanlin. Research on Gray Correlation Algorithm of Factors for Mental Health Warning of College Students [J]. Electronic Design Engineering, 2022, 30(11): 12-16.

[3] Liu Jinming. Simplification and Application of SCL-90 Based on Machine Learning [D]. Qingdao University, 2020. 002011.

[4] Fan Wenrong. (2023) Research on Early Warning Model of College Students' Depression Tendency Based on Social Platform Data [D]. Nanjing University of Posts and Telecommunications.

[5] Zhu Lingyan. (2023) Data Analysis and Prediction of College Students' Mental Health Based on Data Mining [D]. Nanchang University.

[6] Gu Ronglong, Zhao Wenjie, Wang Lei. Application of data mining technology in the era of big data [J]. Science and Technology Innovation and Application, 2022, 12(5): 176-178.

[7] He Feng, Xing Lina, Ren Xuepu, et al. Investigation on mental health status of neurology practitioners in some areas of Hebei Province based on machine learning [J]. Journal of Brain and Neurological Diseases, 2024, 32(04): 240-246.

[8] Yang Juan. Research on the Application of Data Mining Technology in Predicting Mental Health Problems of Higher Vocational Students [J]. Science Consulting (Science and Technology · Management), 2021, (03): 161-162.

[9] Zhou Chengyi, Wang Yixin, Li Xiaoyuan. Intervention Model of College Students' Psychological Problems Based on Big Data [J]. Modern Communication, 2021, (17): 71-73.

[10] Hao Wanlin. Research on Gray Correlation Algorithm of Factors for Mental Health Warning of College Students [J]. Electronic Design Engineering, 2022, 30(11): 12-16.

[11] Tian Wei. Classification Algorithm Application of Big Data Mining — — Taking XGBoost as an Example [J]. Wireless Internet Technology, 2022, 19(19): 120-123.

[12] Mahendra A R P ,Irzam K R ,Sidharta S . Technique of Mental Health Issues Classification based on Machine Learning: Systematic Literature Review[J]. Procedia Computer Science, 2023, 227: 137-146.

[13] Wang Zhengli. (2023) Research and Application of Mental Health Prediction Method Based on Association Rule Mining [D]. Jinan University.

Fund project

Guangdong Provincial Education Science Planning Project under grant 2022GXJK378, Project of 2024 National University Student Innovation Training Program. (No. 202412621002).