Artificial Intelligence, Medical Engineering and Education
Z.B. Hu et al. (Eds.)

doi:10.3233/ATDE231332

Analysis of Psychological Health Data for College Students Based on Data Mining

Qin JIN[1]
College of Resources and Environment, Wuhan University of Technology, Wuhan, Hubei 430070, China
ORCiD ID: Qin Jin https: //orcid.org/0009-0004-9330-1482

Abstract. The backlog of pressure from multiple parties has led to a variety of psychological problems and has also led to many irreparable tragedies. It is very important for college student management to pay attention to the mental health problems of college students, and to discover and guide the mental health development of college students in time. This paper uses data mining technology to set up a questionnaire with 90 questions reflecting mental health as a sample. Using the "SCL-90 Symptom Self-Rating Scale" as the evaluation standard, data were collected, a data set was established, and data analysis was carried out, so as to detect and adjust psychological problems in time and avoid the aggravation of problems and lead to suicide and other irreversible problems.

Keywords. Mental health of college students; SCL-90 symptom self-rating scale; Data mining; Prediction model

1. Introduction

For a long time, people did everything they could for their physical health, but the low level of economy and consumption made people unconcerned about mental health. At the same time, due to limited awareness, people did not realise the importance of mental health, and physical health was basically all people knew about health. This view has long been embedded in people's minds, and even now many people still disdain mental health and consider psychological problems to be insignificant. While physical health is essential, psychological health is also vital [1]. Physical health and psychological health are indispensable, and neither can be called healthy without the other.

There are a variety of definitions of mental health, and specific measures vary. The American scholar Cambs analyzed and researched from the perspective of personality traits and believed that mental health should include traits such as positive perceptions, identification with others, open acceptance and confrontation with reality, clear understanding of oneself and one's surroundings and environment, rich experience and the ability to call on experience to solve problems when encountering them. Orbolt, on the other hand, has studied from the perspective of mature human nature and believes that a healthy personality should have the ability to extend itself, good interpersonal relationships, emotional security and identity, perceptual objectivity, and a unified and correct outlook on life [2].

[1] Corresponding Author: Qin JIN, E-mail: txcf2000@163.com.

In general, psychological health is defined as a positive state of all aspects and processes of human psychological activity, the ability to harmonise relationships, and the ability to adapt to the environment [3]. In the context of positive psychology, the main criteria for the mental health of university students include the following: self-acceptance, open-mindedness, the ability to accept both strengths and weaknesses and face them objectively, constantly developing self-awareness and facing people and events in life positively, positive and optimistic, maintaining enthusiasm for life, not being negative in the face of setbacks, being able to summarize the reasons for failure in time and solve problems calmly, good empathy, being able to feel hope for a better future and find a sense of meaning, always feeling a sense of happiness in life, forming a perfect personality, good empathy, being able to feel a sense of happiness in life, being able to feel a sense of meaning in life [4]. They are hopeful, full of hope for a better future and able to find a sense of meaning, always feeling a sense of happiness in life and forming a perfect personality, good empathy, able to feel the emotions of others and care about their feelings, with the ability to think differently, able to agree with the views and attitudes of others and accept their advice, good interpersonal relationships the student should be able to live in harmony with others, not to malign and hate others, and to live in harmony with others [5]. University students should be aware of their own state of being, in their daily lives.

Studies on prevalence reveal that a considerable proportion of college students experience psychological distress, with common disorders being anxiety, depression, and substance abuse. The research on causes reproduces a wide range of contributing factors from academic performance pressures, social isolation, to financial difficulties.

While notable work has been done to map the extent and underpinnings of psychological health issues in college students, there remain gaps in understanding how to effectively interpret the data obtained and adapt it into effective strategies for prevention and intervention. Studies on management and intervention have largely focused on individual counseling, group therapies, and helpline services, but there is a dearth of research analyzing the efficacy of these strategies based on collected psychological health data.

In light of the existing research, several important questions emerge: How can psychological health data for college students be effectively utilized to improve prevention and intervention strategies? What are the key indicators in the data that could help facilitate early detection of these disorders? How can digital technology improve psychological health data collection and analysis efficiency?

These questions underscore the need for more empirical research focused on the meaningful interpretation of psychological health data, which can provide insights on tailoring more effective mental health programs for college students. This review suggests that further research is necessary to address these pressing concerns and bridge the gap in our understanding of college students' psychological health.

2. Factors Influencing the Mental Health of University Students

At present, the mental health problems of university students are emerging and showing an increasing trend of seriousness, and their influencing factors are various, the main factors include the following aspects.

(1)Physiological factors. Mental health problems can have not only psychological but also physiological causes. Research has shown that abnormalities in the functioning of the neuro-endocrine-immune network system can lead to abnormal moods in the body,

with pentazocine and norepinephrine being the most representative substances in this system.

(2)Self-factor. Many students may have the wrong perception of appearance, taking appearance as an important criterion for judging a person, and there is no lack of people who combine beauty and wisdom among university students, which may devalue themselves in comparison [6]. Secondly, university students begin to face emotional problems, they are all new to relationships and may have misconceptions in their cognition in the face of complex emotional problems, and their immature view of love may lead to falling out of love and emotional problems [7].

(3)Family factors. Many students' families have more or less problems. Some parents may be divorced, or their parents may be away for a long time, making the child a left-behind child, and the lack of affection may make the child have low self-esteem, be more sensitive and withdrawn, and may have some problems in interacting with others. In addition, for some parents, they devote all their efforts to their children, hoping that they will become the best of the best, and that they will become outstanding in the future, so there may be excessive emphasis on academic achievement, making their children attend a variety of tutorial classes at the expense of their inner needs, and completely arranging their children's lives, making them unable to take charge of their own lives, which in the long run will lead to a lack of self-confidence, lack of initiative and timidity. In the long run, the child will become unconfident, lacking in initiative, timid, etc.

(4)The school factor. The university life tests students' ability in all aspects and is no longer limited to academics. This requires students to be self-monitoring, which can lead to academic anxiety for those who have poor self-control, over-indulgence or neglect of their studies, as they may experience poor academic performance and may have a significant psychological gap. Secondly, university is no longer judged on academic criteria alone, but also on social skills, practical skills, and the ability to work in the classroom.

This can be a big shock to students who are known for their grades in secondary school, and other deficiencies such as personal strengths can gradually become apparent, leading to an inferiority complex. The changes in all areas may not be a quick adjustment for university students who have just left high school life.

(5)Social factors. For students nearing graduation, the influence of social factors may be more pronounced [8]. The increasing competitiveness of employment has placed higher demands on the qualifications and knowledge level of university students, especially It is the uninterrupted emergence of the new crown pneumonia epidemic in recent years that has put many small and medium-sized companies at risk of bankruptcy, greatly reducing employment opportunities and leading to a more difficult employment situation, with many university students unable to find suitable jobs and facing the risk of unemployment upon graduation.

(6)The internet factor. The rapid development of the Internet has brought about a variety of problems. Social software has become more mature, and basically all university students are used to surfing online, so it is highly likely that they will face the problem of online violence, and the collision of ideas may lead to strong online arguments or even online attacks, which will seriously damage the psychological health of university students. Also, although college students are already adults, many of them are not yet mentally mature and cannot perfectly handle the problem between online games and studies, and may waste their studies to the extent that they face failure and difficulties in graduation [9].

3. Psychological Data Collection and Processing

The dataset for this paper is derived from online mental health questionnaires for undergraduate students at a number of universities in 2018-2020, all using the SCL-90 symptom self-assessment scale, and integrating data from each university to obtain a total of 11, 897 mental health data.

3.1. Questionnaire Design Principles

Questionnaires are about providing valuable data to the investigator by understanding targeted, purposeful questions, analysing and researching the data to derive the basis for what is needed. Therefore, the design of the questionnaire is particularly important. A good questionnaire can both communicate the questions clearly and understandably to the respondents, and at the same time make the respondents enjoy themselves and gain their support. Therefore, the questionnaire set up both to follow certain principles, but also need to be skillful. The principles of questionnaire design are as follows [10] .

(1) Clarity of purpose. The main purpose of the questionnaire is to obtain the data needed by the investigator, and the questions set must be indispensable, with no irrelevant questions. The subject is clear, the objective is clear and focused.

(2) Be logical. Questions should not be arranged randomly, but in a logical order, easy first, then difficult, set in a regular sequence, with difficult questions placed later.

(3) Easy to understand. The questionnaire is aimed at the general population, the questions should be set clearly and unambiguously, do not beat around the bush and do not use professional vocabulary, if the questions are too advanced and difficult to understand, it will reduce the interest of the respondents, the respondents can easily refuse or give up, so use a simple and easy to understand presentation.

(4) Reasonable length. The questionnaire should not be too long, too much length will cause respondents to lose patience and may give up answering, or they may answer indiscriminately to save time, resulting in less credible results.

3.2. Data Optimization

Once the data has been obtained, it needs to be optimized. A total of 11897 data were obtained in this paper, and the data were processed using Python, using data.isnull().any() to see if there were missing values in the mental health data obtained in this paper for university students, and using data.duplicated().sum() to see if there were duplicate values in the mental health data obtained in this paper for university students, and the results of the run showed that The obtained data is perfect and there are no missing values or duplicate values, this is because the score of the survey questions in this paper is limited and within a manageable range, therefore the possibility of outliers is very small. After data processing, it was found that the data set in this paper was of high quality and the sample size of the data after processing was still 11897 data items. Some of the data obtained are shown in Table 1 below.

Table 1. Selected sample data sheets

F1	F2	F3	F4	F5	F6	F7	F8	F9	Other
1.33	1.9	1	1.08	1	1.33	1.14	1	1.2	1.57
1.42	2	1.67	1.85	1.8	1.33	1.14	1.5	1.4	1.14
1	1	1	1	1	1.17	1	1	1	1
1	1.1	1	1.08	1	1	1	1	1	1
1	1.2	1.44	1.08	1.3	1	1.29	1.33	1.2	1
1.5	1.9	1.78	1.38	1.3	1.33	1.29	1.17	1.3	1.86
1.25	1.5	1.33	1.15	1.3	1	1	1.17	1.1	1.14
1.17	2	1.89	1.69	1.6	1.33	1.43	1.67	1.5	1.43
1.08	1.6	1.11	1.15	1.2	1.17	1	1.17	1	1.14
1.25	1.6	1.44	1.23	1.7	1.17	1	1.5	1.7	1.86
1.92	2.9	2.56	2.23	2.2	1.83	3.14	2	1.4	1.29
1.17	1.2	1.56	1.31	1.3	1	1.43	1.33	1.3	1.57
1	1.2	1.89	1	1.2	1.17	1.29	1.67	1.1	1
1.08	1.3	1.56	1.38	1.5	1.17	1.43	1	1.4	1.14
1	2.1	1.89	1	1.2	1.17	1.29	1.33	1.1	1
1.5	2	1.56	1.46	1.6	1.17	1.29	1.67	1.3	1.86
1.08	1.5	1.67	1.15	1.2	1.17	1.29	1.5	1	1
1.08	1.3	1.33	1.23	1.1	1.33	1	1.33	1.6	1.57

Relational sensitivity with 9 items, F4 for depression with 13 items, F5 for anxiety with 10 items, F6 for hostility with 6 items, F7 for terror with 7 items, F8 for paranoia with 6 items, F9 for psychoticism with 10 items, and other with 7 items.

4. Correlation Analysis of Variables

In this paper, there are 14 variables affecting the mental health of college students, namely total score, total mean score, mean score of positive items, number of positive items, depression, anxiety, interpersonal sensitivity, psychoticism, obsessive-compulsive disorder, paranoia, hostility, somatization, phobia, and others. In order to analyze the degree of correlation between the variables, this paper obtains a heat map based on person's correlation coefficient, as shown in Table 2. From the table we can see that the correlation coefficients for all variables were above 0.4, with strong correlations for total score, total mean score, and suppressed the correlation coefficients for depression, anxiety, number of positive items, interpersonal sensitivity, and psychoticism reached 0.8 or more, with stronger correlations relative to the other variables.

Table 2. Table of predictive model variables

Variable name	Meaning of variables	Variable type	Variable length
TOTAL_SCORE	Total score for psychological survey questions	Int	64
DEPRESSION	Depression	Float	64
ANXIETY	Anxiety	Float	64
NUMBER_OF_POSITIVE ITEMS	Number of positive items	Int	64
INTERPERSONAL_SENSITIVITY	Interpersonal sensitivity	Float	64
PSYCHOTIC	Psychogenic	Float	64
OBSESSIVE-COMPULSIVE_DISORDER	Obsessive Compulsive Disorder (OCD)	Float	64
PARANOID	Paranoia	Float	64
OTHER	Other	Float	64
HOSTILITY	Hostile	Float	64
SOMATIZATION	Somatization	Float	64
FEAR	Horror	Float	64

5. Statistical Analysis of Psychological Data

In terms of the psychological status of university students, of the 11897 data obtained from the sample of university students' mental health in this paper.The sample data for those without psychological problems was 8537, representing a proportion of the total student mental health sample of the number of data samples with psychological problems was 3360, accounting for 28.24% of the total number of students' mental health samples, but this aspect alone cannot clearly and concretely show the distribution of psychological problems. In this section, we will classify and count the mental health data obtained from university students, obtain the overall variables such as total score, total mean score, number of positive items and mean score of positive items, analyse the data from different perspectives such as local and overall, and draw the data distribution of each influence factor, total score, total mean score, number of positive items and mean score of positive items to show the distribution of mental health level and problems of university students from multiple perspectives. The data distribution of each impact factor, score, total mean score, number of positive items and mean score of positive items are plotted to show the distribution of mental health and problems among university students from multiple perspectives [11].

5.1. Analysis of the Distribution of Factors Influencing Psychological Problems

The data mining methods used in this article include:

(1) Pearson Correlation Coefficients: Measures the degree of linear correlation between two variables.

$$r = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 * (y_i - \bar{y})^2}} \tag{1}$$

(2) Decision Tree: An algorithm used for classification or regression problems.

$$\mathrm{F}(x) = \mathrm{argmax}_c \sum_{y[i]=c} \left[h(x,c) - y[i]\right] + \alpha \cdot \mathrm{Count}(C) \quad (2)$$

(3) K-Nearest Neighbor (KNN): An algorithm used for classification or regression problems.

$$\mathrm{P(x)} = \mathrm{argmax}_c \sum [\mathrm{kNN}(x, \mathrm{train}[c]) < \mathrm{d}(\mathrm{train}[c], x)] \cdot (y[\mathrm{train}[c] == c) \quad (3)$$

(4) Neural Network: An algorithm used for classification or regression problems.

$$y = \mathrm{Sigmoid}(\omega x + \mathrm{b}) \quad (4)$$

These formulas are only a small part involved in data mining, and when used, appropriate algorithms and formulas need to be selected based on specific problems and datasets.

5.2. Analysis of Total Score Distribution

The higher the total score is, the worse the mental health level is, and the more teachers and counsellors need to pay attention to it. When the total score does not exceed 160, it means that the student is less likely to have mental health problems, when the total score exceeds 160 but is not higher than 200, it means that the student may have mild psychological problems, but not serious, when the total score exceeds 200 but is not higher than 300 When the total score is above 160 but not above 200, the student is likely to have a moderate psychological problem, and when the total score is above 300, the student is highly likely to have a serious psychological disorder.

Among the 11897 students in the mental health sample, 10243 had a total score of 160 or less, accounting for 86.10% of the total sample. The number of students with a total score of more than 160 but not more than 200 is 1114, accounting for 9.36% of the total sample. These students have mild psychological problems, so counsellors need to communicate with these students from time to time to understand their inner thoughts, find out where the problems lie, and regularly understand their status. Compared to students with mild psychological problems, these students need the help of teachers and counsellors to help them solve their psychological problems and regulate their psychological state, and to inform their parents about their psychological condition, so that parents and schools can work together to help students and relieve their psychological pressure. There are 30 people with a total score of over 300, accounting for 0.25% of the total sample. These people are suffering from very serious psychological problems and are in an unstable and extreme state of psychological emotions, and may have thoughts of light-heartedness, and need psychological assistance from a professional counselling teacher or more systematic and professional treatment at a psychological clinic. For students who may be mentally ill, counsellors should contact parents in a timely manner and send students home for medical treatment to avoid accidents.

5.3. Analysis of the Total Mean Score Distribution

The total mean score is an extension of the total score and is an objective description of the average level of psychological problems of students. In 3.5.2, the total score is divided into four levels, no, mild, moderate and severe, so the corresponding total mean score will also be divided into four levels, no, mild, moderate and severe, which are no more than 1.78, no more than 1.78 but no more than 2.22, no more than 2.22 but no more than 3.22, no more than 1.78 but no more than 2.22 and no more than 3.22. over 1.78 but no higher than 2.22, over 2.22 but no higher than 3.33, over 3.33, and Although the total score and the mean score represent different meanings, they are consistent in terms of the level of mental health of the students and the number of students in each level remains the same, so this section does not analyse the data on the total mean score in detail.

5.4. Analysis of the Distribution of the Number of Positive Items

Although the data on the mental health of university students have been analysed from the perspective of total and mean scores, the analysis from the overall perspective is still rather thin as there is a certain degree of consistency between the total and mean scores, and they overlap to a considerable extent. The number of positive items refers to the total number of items with a score of more than 2. The questionnaire in this paper contains 90 items, and according to the results of the SCL-90 Symptom Self-Rating Inventory, we can see that when the number of positive items exceeds 43, the student has a high probability of having mental health problems. Therefore, in this paper, the number of positive items obtained from the psychological survey in the mental health data set was categorised using 43 items as the classification boundary.The majority of these students (more than 47 items) scored below 2, indicating that in the majority of the 90 survey items in this paper did not show any psychological problems, so on the whole, these people are less likely to have mental health problems, but it should not be ignored that there are still some university students whose number of positive items is at the border of 43, which is within the normal range but still has a certain risk. Therefore, we should not ignore this group of university students and should conduct psychological surveys from time to time to pay attention to their psychological health and prevent them from developing psychological problems due to the occurrence of other problems [14].

The number of positive items exceeded 43, accounting for 19.79% of the total sample. The number of positive items indicates that most of them have mental health problems and need the help of teachers and counsellors. For those students who are positive for most of the items, they need not only psychological counselling, but also professional psychological staff to diagnose and analyse their problems and seek medical treatment in time.

5.5. Analysis of the Mean Score Distribution of Positive Items

The mean score of positive items is the average score of all items with a positive test result, i.e. those with a score of 2 or more. The value is the ratio of the total score of positive items to the number of positive items, which is an objective description of the average level of positive items, as it is only for items that present a positive result. As there are no positive items for those with a mean positive score of no more than 2, and all items are in a normal state, this paper does not examine them, but only those with a mean positive score of more than 2. This paper examines the mean positive item scores

in the mental health dataset.The classification is based on a scale of 3 and 4, and is divided into those with a score of more than 2 but not more than 3, those with a score of more than 3, and those with a score of more than the three levels of scoring, not higher than 4 and more than 4, are categorised and counted. The number of people who scored more than 2 but not more than 3 was 8180, which is the percentage of the total sample 68.76%, those scoring more than 3 but not more than 4 were 353, or 2.97% of the total sample.The number of people with a score of more than 4 was 35, representing 0.29% of the total sample. From the above data we can see that the majority of people with a positive score are in the mild stage and not yet severe enough to warrant intervention. This is because the mean score of positive items is a measure of the severity of positive items, not a measure of the overall distribution of psychological problems among university students. The mean score of positive items is higher than 2, so that those who do not score above 2 are those who do not have any of the 90 items.For those with a positive score of more than 2, we can only judge the presence of positive items but not the number of positive items, and therefore we cannot judge whether the university student has a mental health problem. If this item is used as an indicator to evaluate whether a university student has a psychological problem, there is a high risk of misjudgment for those students who have 43 positive items or less, so it can be used to determine whether the average level of positive items is mild, moderate or severe, but it cannot be used as a criterion to evaluate whether a university student has a psychological problem.

Therefore, in order to prevent overfitting, we can select one of the predictor variables and discard one, in this paper, we choose to discard the total mean score. In addition, the mean score of positive items is the average score of all positive items. When the number of positive items does not exceed 43 and the mean score of each factor does not exceed 2, there may be cases where the mean score of positive items is high. Therefore, this indicator is not used as a predictor variable in this paper. As each of the other factors exceeds the corresponding value, it is possible that there is a mental health problem, and the purpose of psychological prediction is to identify and improve the psychological condition of students with mental health problems as much as possible, so each variable has an important value.

6. Conclusion

In recent years, the proportion of college students suffering from depression has increased. Especially since the COVID-19 in 2020, many smaller enterprises are facing the crisis of development difficulties or even bankruptcy, which has led to a sharp reduction in the number of social recruitments, increasing employment pressure on college students, which has greatly increased the living and psychological pressure on college students. Students with insufficient psychological endurance are prone to depression, Therefore, it is very important for universities to promptly identify students with psychological problems and provide professional psychological guidance. This article obtained psychological health data of some college students through a questionnaire survey, and fully utilized data mining technology to analyze the sample data. Multiple prediction models were used to predict the psychological health of college students.

Reference

[1] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[2] Brathwaite R, Rocha B M, Kieling C, et al. Predicting the risk of future depression among school-attending adolescents in Nigeria using a model developed in Brazil[J]. Psychiatry Research, 2020, 294: 113511.

[3] Cho S E, Geem Z W, Na K S. Predicting depression in community dwellers using a machine learning algorithm[J]. Diagnostics, 2021, 11(8): 1429.

[4] Y Bengü, üzer Ahmet. The relationship between internet addiction, social anxiety, impulsivity, self-esteem, and depression in a sample of turkish undergraduate medical students[J]. Psychiatry Research, 2018, 267: 313.

[5] A, Herran-Boix, I, et al. The role of personality in the prediction of hospitalization duration in mental health[J]. Personality and Individual Differences, 2014, 60(Suppl): S75.

[6] Mohammed J. Zaki, Wagner Meira, Jr. Data mining and analysis: fundamental concepts and algorithms[M]. Cambridge University Press, 2014.

[7] Burnap P, Colombo G, Amery R, et al. Multiclass machine classification of suicide-related communication on Twitter[J]. Online Social Networks and Media, 2017, 2: 32-44.

[8] Gaur M, Kursuncu U, Alambo A, et al. Let me tell you about your mental health! Contextualized classification of Reddit post to DSM-5 for web-based intervention[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, Italy: ACM, 2018: 753-762.

[9] Wei Xinjuan. A Review of the Relationship between Psychological Capital and Mental Health of College Students[J]. Research on Communication Power, 2019, 3(29): 260 (in Chinese).

[10] Lu Liming. Psychological health education for college students under the guidance of the "Healthy Personality Theory". Journal of Higher Education, 2015, (09): 230-231 (in Chinese).

[11] Tianye. Psychological health standards and educational strategies for college students based on positive psychology[J]. International Education, 2020, 10 (11): 89-90 (in Chinese).

[12] Akhrorov Voris Yunusovich, Farruh Ahmedov, Komiljon Norboyev, Farrukh Zakirov. Analysis of experimental research results focused on improving student psychological Health[J]. International Journal of Modern Education and Computer Science, 2022, 14(2): 14-30.

[13] Han Xiyang. A visual analysis of the research on the use of mobile phones by college students based on VOSviewer[J]. International Journal of Education and Management Engineering, 2020, 10(6): 10-16.

[14] Chuanmei Wang. An Investigation and Structure Model Study on College Students' Studying-interest[J]. International Journal of Modern Education and Computer Science, 2011, 3(3): 33-39.