Hindawi
Mobile Information Systems
Volume 2022, Article ID 3879482, 8 pages
https://doi.org/10.1155/2022/3879482

Research Article

Relationship Analysis between Psychological State of College Students and Epidemic Situation Based on Big Data Mining

Jian Xiang[1] **and Yanjun Zhang**[2]

[1]*Training Centre, The United Front Work Department of CPC Central Committee, Beijing 100037, China*
[2]*College of Chinese Language and Culture, Jinan University, Guangzhou 510610, Guangdong, China*

Correspondence should be addressed to Yanjun Zhang; zhangyanjun@hwy.jnu.edu.cn

Received 13 June 2022; Accepted 22 August 2022; Published 5 September 2022

Academic Editor: Le Sun

COVID-19 is a sudden and highly contagious infectious disease, which has a very bad impact on the psychology of college students in early adulthood. In order to grasp the psychological state of college students in real-time, this work studies the psychological state of college students during COVID-19. First, this study introduces the relevant theories of data mining, and the research object and method are determined. Then, the features of the model are analyzed and constructed from two aspects which are static features and dynamic features, and the characteristics related to the psychological state are excavated. Finally, the GA is selected to build the model and the model is evaluated; the results show that the model can accurately predict the psychological state of students during COVID-19.

1. Introduction

COVID-19 is a sudden and highly contagious infectious disease, which brings great inconvenience to the daily life of people [1, 2]. Based on the current situation, COVID-19 will persist for a long time. Compared with other people, college students are more likely to feel lonely and depressed, so the emotions of students are complex and changeable. At the same time, college students lack experience in dealing with emergencies and their emotions are more likely to be unstable [3, 4]. As a major public health emergency in the world, COVID-19 also had a very bad impact on the psychology of college students in early adulthood. Therefore, it is important to study the psychological status of different students, which can provide a scientific basis for further improving psychological health education.

Researchers have carried out a lot of research studies on the psychological health of college students during the epidemic. The psychological status of college students during the COVID-19 was studied in references [5, 6], in which the results showed that higher-grade college students were more likely to have anxiety and hostility, those who had major life events were more likely to have somatization symptoms, and students without close friends were more likely to have depression and phobic symptoms than students with close friends [5, 6]. The anxiety status of returning college students during COVID-19 was investigated in reference [7], in which the results showed that 20% of returning college students had mild anxiety and 6.5% had moderate or severe anxiety [7]. Sutton and Barto [8] investigated the psychological health status of college students in two stages which are emergency state and normalization state. The results showed that college students had obvious negative emotions such as anxiety, self-blame, and boredom during the epidemic, which decreased with the stabilization of the epidemic [8]. The abovementioned studies show that COVID-19 has a very bad impact on the psychology of students, but there are two problems in the studies, one is that the research sample is too small and the other is that the sample is not representative enough. Most studies focus on the overall psychological health of college students, and the psychological health status of key groups is not paid enough attention [9, 10].

First, this study introduces the relevant theories of data mining; considering the numerous impacts that COVID-19 may have on college students, the research object and research method are determined. Then, the characteristics of the model are analyzed and constructed from two aspects which are static model characteristics and dynamic model characteristics. Finally, a model of the relationship between the psychological state of college students and the epidemic situation is constructed based on GA, and the accuracy of the model is evaluated.

2. Data Mining Theory

With the development of the times, data in various fields are intertwined, which means that the era of big data has arrived. Big data has five characteristics which are massive data scale, rapid data circulation, dynamic data system, various data types, and huge data value, and its relationship is shown in Figure 1. The value hidden in the data does not float on the surface, so the great value inside the data often needs to be mined by various methods [11].

Knowledge discovery is used to identify effective data from the dataset, whose process can be divided into five parts which are data screening, data preprocessing, data conversion, data mining, and data interpretation and evaluation [12]. First, knowledge discovery needs to filter the target data from the most original data for data preprocessing, and the preprocessed data are transformed. Then, the data mining method is applied to form patterns or rules, the knowledge is generated through interpretation and evaluation, and the basic process is shown in Figure 2.

3. Research Objects and Methods

3.1. Research Object. Through the convenient sampling method, an online questionnaire survey was conducted on college students in a university in Beijing, and 12018 valid questionnaires were collected. According to gender and educational background, the research objects can be divided into two parts and the proportion of different research objects in total number is shown in Figure 3. In terms of gender, the number of boys is far less than that of girls, boys account for 26.19% of the total number while girls account for 73.81%. In terms of educational background, the proportion of college students is the highest, accounting for 89.91% of the total number, followed by junior college students, accounting for 6.83%, and the proportion of junior college to undergraduate is the lowest, accounting for 3.26%.

3.2. Research Methods. The questionnaire mainly considered the influence of COVID-19 on the growth experience, personality neuroticism, negative life events, social support, degree of depression, concealment, and the potential risk of suicide of college students, as shown in Figure 4.

Growth experience refers to the impact of things experienced in the process of personal growth by the research object, such as parental emotional incompatibility and maltreatment experience. Research shows that the divorce rate has greatly increased during the epidemic, which will have a certain impact on children and have a far-reaching impact on the future interpersonal relationships of children. At the same time, the economic pressure caused by the epidemic will make parents lose control of their emotions, and children may be abused [13].

Personality trait neuroticism is a basic personality trait in psychological research, which is used to measure the stability of emotional changes. People with high scores of personality trait neuroticism have poor pressure resistance and are more likely to have a sense of hostility. They are usually self-centered and have difficulty controlling their impulses and desires, and a small setback often leads them to despair.

Negative life events refer to the negative events encountered by the subjects in their recent life, such as excessive academic pressure and high family expectations. These events may happen to most of their peers, causing physical or social or psychological trauma to them, but different people may show different reactions when they encounter the same thing because of the difference in gender, age, and cultural background.

Social support refers to the concern and assistance that the research object can feel from the society. When an individual suffers from difficulties, the greater the degree of social support he can get, it will help the individual get out of trouble to a great extent.

Depression is a state of depression and aversion to activities, which has a certain impact on the thought, behavior, feeling, and physical health of people. The higher the degree of depression, the stronger the feeling of depression, helplessness, irritability, and other negative emotions.

The masked psychology refers to the psychology in which the research object deceives itself and others in order to avoid responsibility which is a psychological defense mechanism and psychological diseases are usually accompanied by an excessive psychological defense mechanism.

The potential risk of suicide is used to measure the possibility of subjects to have suicidal behavior. The higher the potential risk of suicide, the stronger their suicidal consciousness, and most of them suffer from mental diseases, especially depression.

The research shows that there are mainly two kinds of negative emotions during the outbreak of a new type of coronary pneumonia, one is anxiety and the other is depression, so the Depression Anxiety Stress scale is selected as a survey tool and is suitable for college students. As shown in Figure 5, negative emotions are divided into five levels which are normal, mild, moderate, severe, and extremely severe. In this study, SPSS software is used to analyze the data, and the significance level of data difference is less than 0.05.

4. Model Feature Analysis and Construction

The factors that affect the psychological state of college students can be divided into two aspects, one is the innate static factor that will not change greatly over time, such as gender, place of birth, and nationality, and these are called inherent attributes. The other is the instantaneous or staged psychological impact of the acquired environment that will change with the passage of time or external intervention,

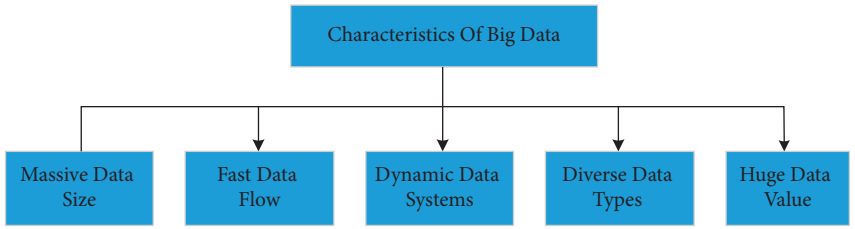

Figure 1: Features of big data.

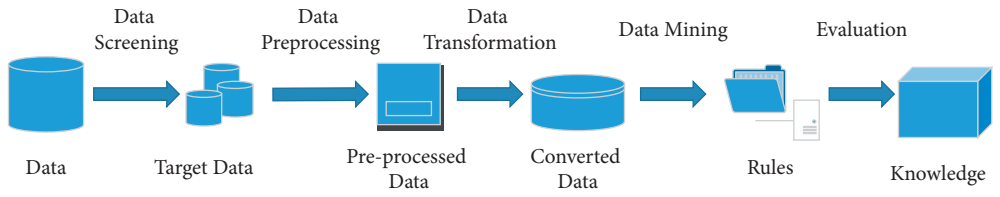

Figure 2: Basic process.

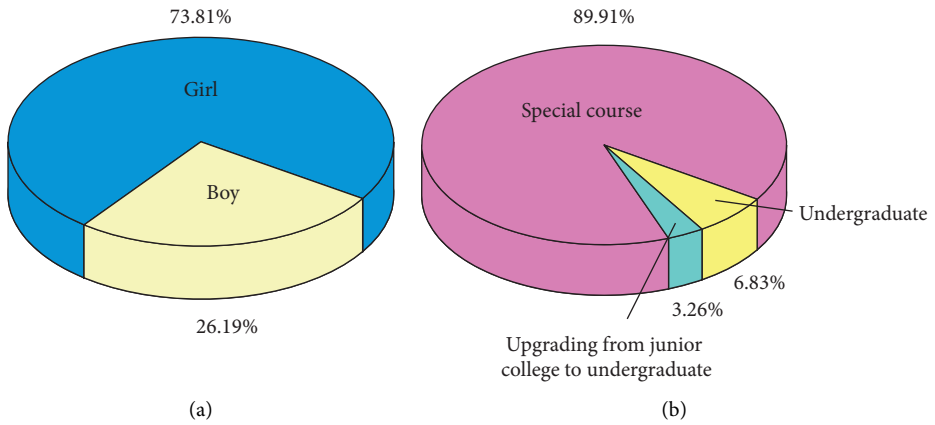

Figure 3: Proportion of research objects. (a) Gender. (b) Academic qualifications.

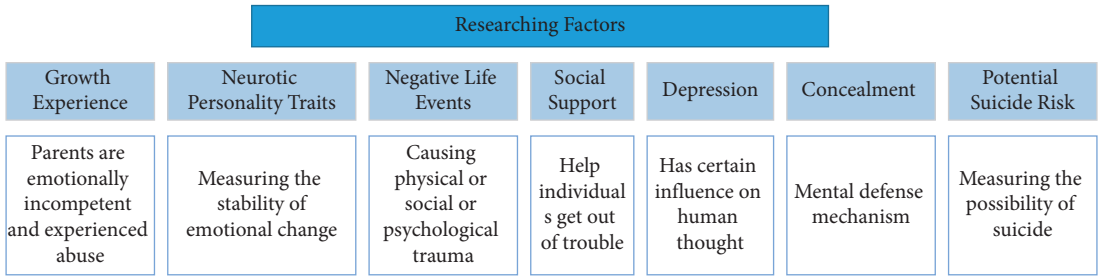

Figure 4: Influencing factors.

such as diet change and behavior change, and these are called dynamic characteristics. This work mainly studies the psychological state of college students during the COVID-19 epidemic, and it is obvious that COVID-19 is a dynamic feature, so this study will mainly construct the characteristics of the psychological state perception model through dynamic features.

4.1. Data Cleaning. Before specific feature analysis and construction, the data need to be cleaned first. For the questionnaire with missing data, it is necessary to complete the data through specific filling strategies. For the problem of partial missing fields, there are two ways to fill in the data, one is zero value filling and the other is mean filling. For example, for the data with a missing age field, the mean

	Depression	Anxiety	Stress
Normal	0~9	0~7	0~14
Mild	10~13	8~9	15~18
Moderate	14~20	10~14	19~25
Severe	21~27	15~19	26~33
Extreme	>27	>19	>33

Figure 5: Depression Anxiety Stress scale.

filling method should be used to fill it, and data whose time field do not meet the requirements shall be filtered. For the data with duplicate data, the key fields of the data shall be counted, and the duplicate data shall be eliminated. There are relevant measurement standards for judging whether the attributes in the data source are redundant, as shown in the following formulas:

$$\gamma_{A,B} = \frac{\sum(A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A\sigma_B}, \tag{1}$$

$$\sigma_A = \sqrt{\frac{\sum(A-\overline{A})^2}{n-1}}, \tag{2}$$

$$\sigma_B = \sqrt{\frac{\sum(B-\overline{B})^2}{n-1}}, \tag{3}$$

where n represents the number of tuples, $\overline{A}$ represents the average of A, $\overline{B}$ represents the average of B, and σ_A and σ_B are the standard deviation of A and B.

4.2. Static Characteristic Analysis. From the perspective of data, static characteristics include three parts which are gender, age, and nationality information in the basic information of students. Considering that they are new college students, static characteristics also include information such as college entrance examination scores, candidate types, and admission batches. This section analyzes the correlation between the inherent attributes and the mental state of people through regression analysis; we can judge whether there is a statistical significance between this attribute and mental state according to the size of the regression coefficient, and the static features in the mental state perception model can be constructed.

For the discrete attribute of gender, the data should be numerically processed first, and the male and female are mapped to 1 and 0, respectively. The crossover frequency analysis is carried out and the results are shown in Figure 6. It can be seen that there is no significant relationship between gender and depression.

The unit of age is year; when the month is less than 12 months, the decimal part is obtained by dividing the number

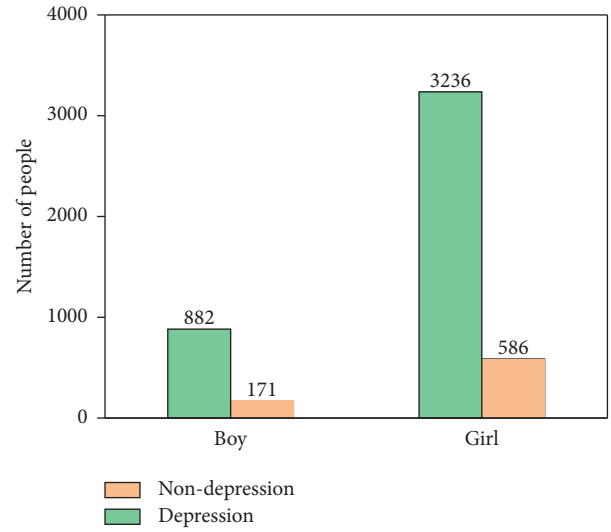

Figure 6: Crossover frequency analysis.

of months by 12. For example, when the time from the birth date to the evaluation date is 20 years and 3 months, the age attribute of the participant is 20.25. For continuous attributes, data need not be specially processed, and the output results can be obtained by adding a constant term into the data.

Similarly, for the discrete attribute of nation, it is necessary to do numerical processing first. Due to the small number of ethnic minorities, it is not meaningful to classify each ethnic group independently, so only the ethnic attribute is divided into two categories, one is the Han nationality and the other is the ethnic minorities, which are mapped to 1 and 0, respectively. Among them, the proportion of depression in ethnic minorities is 24.34% and that in Han nationality is 18.72%; it can be seen that the proportion of depression in ethnic minorities is significantly higher than that in Han nationality.

4.3. Dynamic Characteristic Analysis. The ultimate goal of the psychological state perception model is to predict the psychology of students in real-time during COVID-19. Based on student behavior data, this study constructs three

dynamic characteristics which are student consumption characteristics, student behavior characteristics, and social relationship characteristics.

The consumption data of students during COVID-19 mainly include three parts which are daily consumption data, library access control data, and course selection data. The daily consumption data include the time, place, and amount of consumption, the access control data of the library include the state and time of entry and exit, and the course selection data include the semester and course type. In this study, only the behavioral data of college students from enrollment to test time range are selected, which is also beneficial to the demand of data time span when the model is applied.

In addition to consumption data, college students will also generate behavioral data during their school days which include access control records of dormitories, access control records of libraries, and course selection records. Behavior of people can reflect the psychological state of people to a great extent; this study characterizes the self-discipline ability of people through their behavioral characteristics, which mainly include three aspects, namely, library learning regularity, course selection behavior characteristics, and lost goods behavior characteristics.

The active degree of social relations is closely related to introversion and extroversion of personality. When the social relations of students are more active, the students are more active and the number of their friends is relatively large. When the social relations of students are not active, it shows that the students are not good at or unwilling to socialize. People with relatively isolated personalities are not good at expressing themselves and they tend to accumulate pressure and negative emotions in their hearts. In the long run, when the psychological pressure reaches the threshold they cannot bear, it will lead to mental collapse and eventually form depression. There is no scientific quantitative method to assess the active degree of social relations; this study will characterize the active degree of social relations by the number of friends, that is, the more friends the students have, the more active they are in social relations.

5. Model Building and Verification

The purpose of the model is to predict the psychological state of students through the behavior data during COVID-19. Based on the construction of data mining algorithms and model features introduced above, this section will start from specific data to construct the model of the relationship between the psychological state of college students and the epidemic situation.

5.1. Model Construction. This study is used to construct the model of the relationship between the psychological state of college students and epidemic situations, and various types of classification model algorithms are needed in the preselection stage. Selecting from a variety of data mining models, the GA is finally selected as the model input optimization algorithm, which can find the optimal subset in a wide range of datasets.

The basic principle of GA is to find an optimal binary code through the genetic algorithm in which every bit in the code corresponds to a feature in the feature vector table. If the i-bit is "1," the corresponding feature is selected. If it is "0," it means that the feature is not selected. The basic steps are mainly divided into five parts which are encoding, calculating fitness, selecting the individuals with the largest fitness, crossover and mutation operations, and reproduction, and the basic process is shown in Figure 7.

The training method of the genetic algorithm mainly has two bases, one is to calculate the fitness and the other is the result of the selection strategy. First, it is necessary to select appropriate training methods; then, the labeled supervised samples should be subjected to limited iterative operations, so as to select the most suitable feature combination for distinguishing labels. The genetic algorithm in this study is based on the fitness function of the distance criterion which directly depends on the data of the sample itself. The algorithm has two advantages, one is intuitive and simple and the other is clear physical concept, and the separability of samples can be judged by calculating the distance between similar samples and the distance between different samples. The corresponding calculation contents are given in the following equations:

$$s_b = \sum_{i=1}^{c} \left(P(\omega_i) E\left\{ (M_i - M_0)(M_i - M_0)^T \right\} \right), \tag{4}$$

$$J = \frac{tr(S_b)}{tr(S_w)}, \tag{5}$$

where C represents the number of categories, M_i represents the mean vector, and M represents the mean vector of the sample set.

While searching for the variable combination of the optimal proportioning factors, the value range of the proportioning factors should be determined first, and then, the general form of inversion based on a genetic algorithm can be obtained, as shown in the following equations:

$$\min f(x) = \sum_{i=1}^{n} \frac{\left| y_i^l - y_i^* \right|}{y_i^l}, \tag{6}$$

$$a_j \le X_j \le b_j \, (1 \le j \le k), \tag{7}$$

where X_j represents the combination of variables, a_j represents the lower limit of the value interval, and b_j represents the upper limit of the value interval.

The main purpose of the model is to classify and judge the mental state information of students according to their data during the epidemic, and the construction of the model is mainly divided into four parts which are the preparation of the training sample dataset, the division of the sample

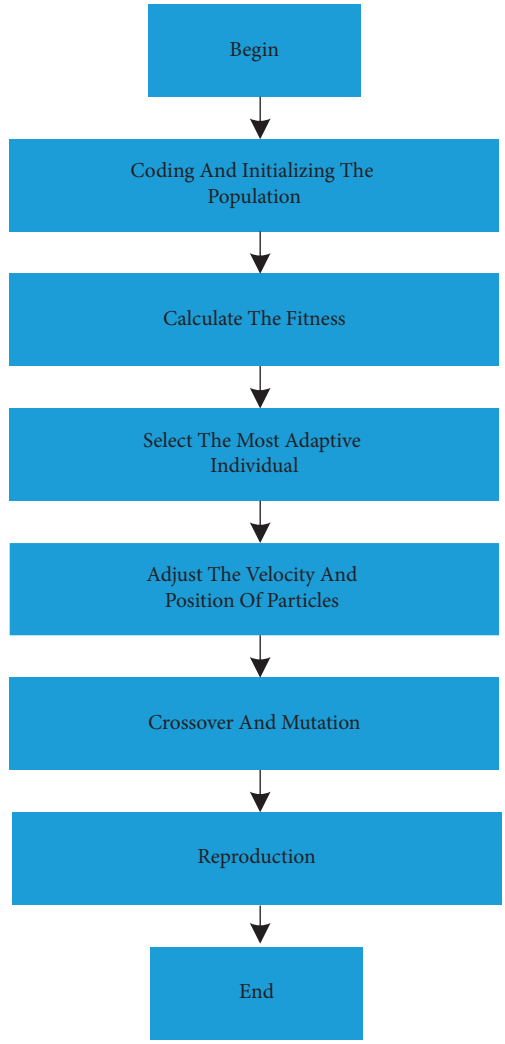

Figure 7: Flowchart.

dataset, the preselection model algorithm, and the training model, as shown in Figure 8.

In the construction of the model, the label data should be obtained first. Second, according to the feature selection method based on a genetic algorithm, different label data should be matched to select the optimal feature dimension. Finally, to compare the influence of the selection of features by the genetic algorithm on the output of the model, the model algorithm of the control group should be selected for comparison and the model should be trained.

5.2. Model Evaluation. For the model, there are three most common evaluation indexes, namely, accuracy rate, recall rate, and F_1 score which combines the accuracy rate and recall rate. The calculation of these evaluation indexes is inseparable from the existence of a confusion matrix, which is often used in classification algorithms. The accuracy reflects the proportion of correct classification in the classification results of each category, that is, the accuracy of each category judged by the model. The recall rate can reflect the sensitivity of the classification model to each category dataset. F_1 score is a new index designed to represent the comprehensive performance of the model which is the harmonic mean of the accuracy and recall rate. The value range of the F_1 score is 0-1, and the greater the value, the better the effect of the model. The calculation methods of accuracy, recall, and F_1 score are given in the following equations:

$$\text{Precision} = \frac{TP}{(TP + FP)}, \tag{8}$$

$$\text{Recall} = \frac{TP}{(TP + FN)} = \frac{TP}{P} = \text{sensitivity}, \tag{9}$$

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{10}$$

where TP represents the positive case when the actual result is positive, FN represents the negative case when the actual result is positive, FP represents the positive case when the actual result is negative, and TN represents the negative case when the actual result is negative.

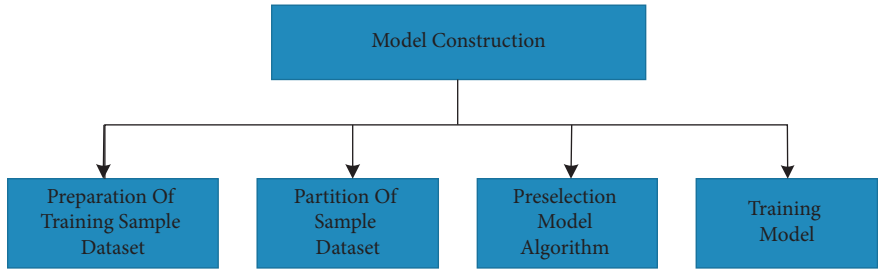

Figure 8: Construction of the model.

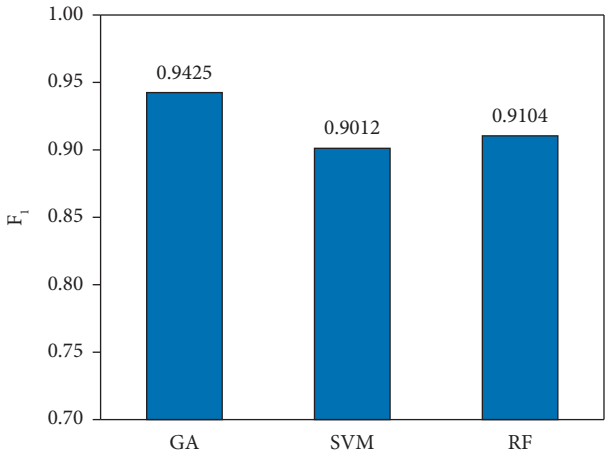

Figure 9: Model verification.

In order to verify the model, the model without the genetic algorithm is set as the control group and the model with the genetic algorithm is set as the experimental group. It can be seen from Figure 9 that the F_1 scores of multiple algorithm models can reach more than 0.9, which indicate that the psychological state of students during COVID-19 is indeed predictable, and there is a relatively large correlation between the psychological state and the behavioral data of students. Through the change of behavioral data, we can accurately predict the psychological state of students during COVID-19.

6. Conclusion

First, this study introduces the relevant theories of data mining; considering the numerous impacts that the COVID-19 epidemic may have on college students, the research object and research method of this paper is determined. Then, the characteristics of the model are analyzed and constructed from two aspects, one is the static model characteristics which include the basic information, such as gender, age, and nationality, and the other is the dynamic model characteristics which include the characteristics of student consumption, student behavior, and social relations. Finally, the model of the relationship between the psychological state of college students and the epidemic situation is constructed based on GA; the results show that the feature dimension data extracted by the genetic algorithm are more representative, and the model can accurately predict the psychological state of students during COVID-19 [14, 15].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. Qian, Y. Cao, and S. Huang, "Relationship between depressive symptoms, neuroticism and online social activities of college students," *Chinese Mental Health Journal*, no. 12, pp. 932–937, 2018, in Chinese.

[2] P. Cohen, S. G. West, and L. S. Aiken, *Applied Multiple Regression/correlation Analysis for the Behavioral Sciences*, Psychology Press, England, UK, 2014.

[3] M. I. Jordan and D. E. Rumelhart, "Forward models: supervised learning with a distal teacher," *Cognitive Science*, vol. 16, no. 3, pp. 307–354, 1992.

[4] D. Liu, J. Qiu, and C. Wan, "Feasibility analysis of using text data of quasi private social network to detect depression users," *Chinese Journal of information technology*, vol. 32, no. 9, 2018, in Chinese.

[5] L. Y. He, "A study on depression and its influencing factors among college students in Beijing," *Modern Preventive Medicine*, vol. 42, no. 7, pp. 1261–1264, 2015, in Chinese.

[6] Z. Zhang and B. Guo, "Peer relationship and mental health from the perspective of social network," *Progress in psychological science*, vol. 188, no. 4, pp. 121–132, 2016, in Chinese.

[7] S. D. Amarasuriya, A. F. Jorm, and N. J. Reavley, "Predicting intentions to seek help for depression among undergraduates in Sri Lanka," *BMC Psychiatry*, vol. 18, no. 1, p. 122, 2018.

[8] R. S. Sutton and A. G. Barto, "Reinforcement learning," *A bradford book*, vol. 15, no. 7, pp. 665–685, 1998.

[9] T. Joachims, "Making large-scale SVM learning practical," *Advances in Kernel Methods Support Vector Learning*, MIT Press, Cambridge, MA, USA, 1998.

[10] Y. Liu, X. Tan, and L. Yang, "Analysis on the current situation and influencing factors of College Students' depression," *Chinese general practice*, vol. 13, no. 1, pp. 91–93, 2015, in Chinese.

[11] M. Collins, R. E. Schapire, and Y. Singer, "Logistic regression, adaBoost and bregman distances," *Machine Learning*, vol. 48, no. 1/3, pp. 253–285, 2002.

[12] Y. Qi, H. Ma, and H. Yan, "Analysis of individual behavior of social network users from the perspective of psychology," *Progress in psychological science*, vol. 22, no. 10, pp. 1647–1659, 2014, in Chinese.
[13] Y. Sasaki, "The truth of the f-measure," *Teach Tutor mater*, vol. 1, no. 5, pp. 1–5, 2007.
[14] S. Ma, C. Jiao, and M. Zhang, "The application of social network analysis in psychological research," *Progress in psychological science*, vol. 19, no. 5, pp. 755–764, 2011, in Chinese.
[15] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, pp. 1189–1232, 2001.