DOI: 10.53469/jsshl.2025.08(02).03

Research on College Students' Mental Health Based on Logistic Regression Model

Xinyu Pang[1], Tianning Liu[2], Hang Zuo[3]

[1,2,3]School of Mathematics, Chengdu Normal University, Chengdu, Sichuan, China
[1]3056547923@qq.com, [2]2025500782@qq.com, [3]zuohang2021@163.com

Abstract: ***Objective:*** ***College students are at a transitional stage where their psychological development is gradually maturing, and they are moving from a single environment to a more diverse one. Therefore, studying the factors that affect their mental health is of great significance.*** ***Methods:*** ***This paper first predicts the impact of various factors on college students' mental health based on a logistic regression model. Next, it classifies college students using a k-means clustering model. Finally, a multiple linear regression model is established to predict the influence of different factors.*** ***Conclusions:*** ***1) The impact of family education methods on the mental health of college students is the most significant. 2) The structure is most stable when the optimal number of clusters for college students is k=3. 3) The primary factor affecting the mental health of college students is family-related, followed by academic factors, and lastly, social factors.***

Keywords: College Students' Mental Health; Logistic Regression Model; K-Means Clustering Model; Multiple Linear Regression Model.

1. INTRODUCTION

In recent years, the mental health issues of college students in China have developed rapidly. Huang Xinyi [1] conducted a systematic and comprehensive study on psychological positivity through four sub-studies. Liangqun Y [2] applied the fuzzy comprehensive evaluation method to analyze and evaluate students' mental health and its influencing factors. Li Yi and others [3] developed a practical tool to evaluate the satisfaction status and influencing factors of college mental health work to understand the current state of satisfaction in this area. Wang Zhenjie et al. [4] used the random forest algorithm to analyze the factors affecting the mental state of college students during the COVID-19 pandemic. Qian Yimian [5] proposed that there is no significant correlation between anxiety self-rating scores and the gender, family structure, or family income of the participants. Pan Miao et al. [6] analyzed the factors influencing the mental health of college students under stress, taking the COVID-19 pandemic as a stress factor. Wu Haipeng [7] focused on the "main influencing factors" that trigger mental health issues among college students and analyzed the current state and causes of college mental health education. Lu Ping [8], through literature analysis combined with interviews and semi-open questionnaires, explored the structural dimensions of college students' academic mental health. Zhong Jinyuan et al. [9] suggested integrating the impacts of physical exercise, lifestyle, and health cognition on mental health into research to achieve more comprehensive results. Gu Dacheng [10] proposed that lifestyle positively promotes mental health and that this promotion process is gender-dependent. Zhang Bin et al. [11] detailed that economic factors are significant contributors to mental health issues among college students. Pei Xuejin [12] believed that establishing the discipline of college mental health education and developing standards for evaluating students' mental health status could enhance the scientific level of such evaluations. Xu Tonghai [13] proposed that colleges should provide personalized attention and education to every student to promote their overall positive changes and healthy growth. Pan Shubo [14] identified significant and non-significant factors affecting college students' mental health through an investigation and analysis of various factors related to college students' mental health and their interrelationships. Cao Rong [15] argued that the factors influencing mental health are multifaceted and that life events can have both direct and indirect impacts on mental health.

2. METHODS

2.1 The Establishment of the Logistic Regression Model

2.1.1 The selection of indicators

Based on the collected data, the following definitions are provided for indicators strongly associated with influencing factors.

Definition 1 $X_i(i=1,2,3)$ represents the *i*-th broad category of mental health influencing factors, where X_1 represents family factors, X_2 represents social influences, and X_3 represents school factors.

Definition 2 $X_j = \{X_i(k)\}$, $k = 1,2,3,4; j=1,2,3$ represents the *j*-th subcategory of mental health influencing factors, where $X_i(k)$ denotes the *k*-th indicator of the *i*-th influencing factor. $X_1(1)$ represents "Parenting Style", $X_1(2)$ represents "Family Atmosphere", $X_1(3)$ represents "Family Structure", $X_1(4)$ represents "Family Economic Conditions"; $X_2(1)$ represents "Social Competition Pressure", $X_2(2)$ represents "Social Expectations", $X_2(3)$ represents "Cultural Differences in Society", $X_2(4)$ represents "Influence of Social Media"; $X_3(1)$ represents "Academic Pressure", $X_3(2)$ represents "Academic Competition", $X_3(3)$ represents "Interpersonal Relationships", $X_3(4)$ represents "Career Guidance and Support".

(1) Proportion of Key Family Factors Influencing College Students' Mental Health (PTF_k)

The formula for calculating the proportion of the *k*-th key influencing factor within family factors is:

$$PTF_k = \frac{X_1(k)}{X_1} \tag{1}$$

This indicator shows the proportion of the four main influencing factors in family factors and can reflect the ranking deviation of their influence levels to a certain extent.

(2) Proportion of Major Social Factors Affecting College Students' Mental Health (ERI_k)

The formula for calculating the proportion of the *k*-th major influencing factor among social factors is:

$$ERI_k = \frac{X_2(k)}{X_2} \tag{2}$$

This indicator reflects the degree of influence of the four major social factors on college students' mental health, thereby exhibiting a strong correlation with the classification of influence levels.

(3) Proportion of Major School Factors Affecting College Students' Mental Health (GZS_k)

The formula for calculating the proportion of the *k*-th major influencing factor among school factors is:

$$GZS_k = \frac{X_3(k)}{X_3} \tag{3}$$

This indicator can describe the degree to which the four main influencing factors among school factors tend to influence the mental health of college students, and it has a strong correlation with the classification of the severity of the influence.

2.1.2 Establish a logistic regression model

*STEP*1: Define the logistic regression model

In logistic regression, the *j*-th indicator of the *i*-th object is denoted as x_{ij}, and the classification corresponding to the *i*-th object is denoted as y_i. Then the logistic regression model can be expressed as:

$$P(\hat{y}_i = 1; B) = \frac{1}{1+e^{-X_i^T B}} \tag{4}$$

Here, B is the parameter vector, X_i represents the various features of the *i*-th object, y_i is the label (0 or 1) of the *i*-th sample, $P(\hat{y}_i=1; B)$ represents the probability that the *i*-th object is classified as 1 with B. The expressions of X_i and B are as follows:

$$X_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{im} \end{pmatrix}, B = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{pmatrix} \tag{5}$$

Here, x_{ik} represents the *k*-th indicator of the *i*-th object, and b_k represents the parameter vector for the *k*-th indicator.

*STEP*2: Estimate the parameter B using the maximum likelihood estimation

Use the maximum likelihood estimation to estimate the parameter B, and its formula is as follows:

$$max\, L\,(B) = \prod_{i=1}^{m} P(\hat{y}_i = 1; B) \tag{6}$$

Among them, max$L(B)$ represents the maximum value B that the likelihood function $L(B)$ reaches on the given data set.

The log-likelihood function obtained after taking the logarithm of the likelihood function is shown as follows:

$$maxln\, L\,(B) = \sum_{i=1}^{m} ln\,[\,P(\hat{y}_i = 1; B)] \tag{7}$$

By maximizing this log-likelihood function, the parameter estimation of the logistic regression model can be obtained.

*STEP*3: Use the Newton's tangent method to find the parameter *B*

Iterate according to the following formula:

$$B_n = B_{n-1} \frac{[\,ln\, L\,(B_{n-1})]^{'}}{[\,ln\, L\,(B_{n-1})]^{''}} \tag{8}$$

Set the initial value of the parameter as B_0; B_n is the current iteration value; $B_{n\text{-}1}$ is the value of the previous iteration.

*STEP*4: Draw the *ROC* curve, calculate the value of *AUC* and test the accuracy of the model

Substitute the obtained parameter *B* into the logistic regression model to predict the test data set, and the predicted probability $P(\hat{y}_i{=}1;\, B)$ of each sample can be obtained. Then, according to the true label and predicted probability of each sample, divide them into positive or negative classes and calculate the confusion matrix *CM*. Then use different probability thresholds *D* to calculate the corresponding proportion of the true positive rate (*TPR*) and the false positive rate (*FPR*). The calculation formulas are as follows:

$$CM = \begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix}, TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{FP+TN} \tag{9}$$

Among them, *TP* represents the true positive, *FN* represents the false negative, *FP* represents the false positive, and *TN* represents the true negative.

The *ROC* curve takes *TPR* as the y-axis and *FPR* as the x-axis, depicting the relationship between the true positive rate and the false positive rate of the model under all possible thresholds, as shown in Figure 1.

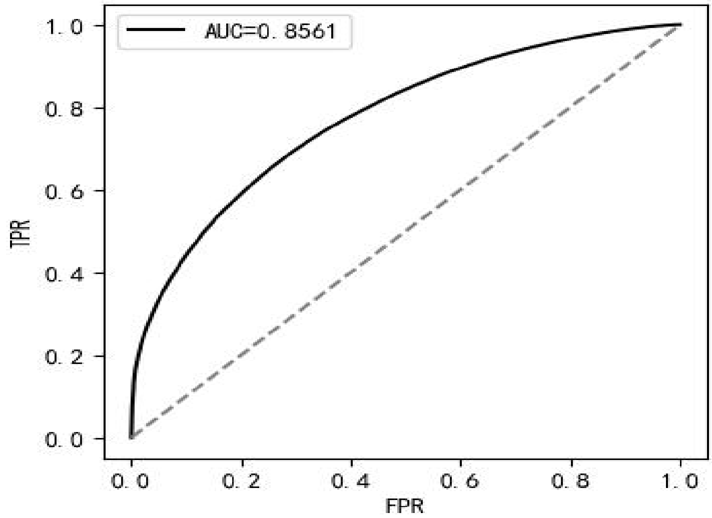

Figure 1: ROC Curve Graph

The value of *AUC* is the area between the *ROC* curve and the x-axis, reflecting the overall performance of the model under all possible probability thresholds. Its range is between 0 and 1, and the closer the value of *AUC* is to

1, the better the performance of the model is. As can be seen from the figure, the AUC value in this paper is 0.8561.

2.2 Establishment of the *k-means* Clustering Model

*STEP*1: Establish the basic *k-means* algorithm

Build a *k-means* model for clustering on the feature indicators, and take the data of each feature indicator as an object.

Definition 3 $Yi= (x_{i1}, x_{i2},..., x_{im})$ is the ith object to be clustered, where xik represents the data corresponding to the kth indicator of the ith object. i=1, 2, 3; k = 1, 2, 3,4.

Definition 4 $L_i \in \{1, 2, . . . , k\}$, i = 1, 2, 3 is the classification label of the i-th object.

Definition 5 $C_i = (c_{i1}, c_{i2}, . . . , c_{im})$ represents the ith clustering center, where c_{ik} represents the k-th indicator of the i-th clustering center. $i = 1, 2, . . . , k$.

*STEP*2: Determine the value of k, the number of clusters

Randomly select (sampling without replacement) k samples from the sample set $H=\{X_1, X_2, . . . , X_n\}$ as the clustering centers.

*STEP*3: Allocate labels to the samples

Calculate the distance between each sample and the k clustering centers, and assign the sample to the clustering center with the smallest distance. The calculation formula is as follows:

$$D_{ij} = \sqrt{\sum_{i=1}^{m} (x_{im} - c_{jm})^2} \tag{10}$$

The smaller the distance, the greater the possibility of classifying the sample into that category, so select the closest clustering center, and classify the sample into its class.

Definition 6 $Ui = X_j^{(i)}$ represents the i-th cluster, where $X_j^{(i)}$ represents the j-th object assigned to the i-th cluster.

*STEP*4: Optimize the clustering centers

Move each clustering center to the geometric center of the cluster to optimize the clustering centers.

$$c_{ij} = \frac{1}{|U_i|} \sum_{t=1}^{|U_i|} x_{tj}^{(i)} \tag{11}$$

Among them, c_{ij} represents the j-th indicator data of the i-th clustering center, $x_{tj}^{(i)}$ represents the jth indicator data of the t-th object in the cluster corresponding to the i-th clustering center, and $|U_i|$ represents the number of objects in the cluster U_i. Here, i= 1, 2, . . . , k; j = 1, 2, 3, 4; t = 1, 2, 3.

*STEP*5: Judge whether the clustering centers have moved

If there is a clustering center that has moved, then return to *STEP*2; otherwise, it indicates that the algorithm has stabilized, so exit the algorithm.

2.3 Establishment of the Multiple Linear Regression Model

*STEP*1: Construct the multiple linear regression equation

The multiple linear regression equation can be used for prediction, analyzing the relationships among variables and explaining the influence of variables. In multiple linear regression, assuming there are p variables x_1, x_2, x_3, ..., x_p, the formula is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{12}$$

Among them, β_0 is the intercept, β_1-β_p represent the regression coefficients of each random variable, y represents the predicted variable; x_1-x_p represent the input variables used to predict the dependent variable; ε represents the error term.

Assuming there are n samples, the expression can be written in the following form:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} \end{cases} \tag{13}$$

*STEP*2: Construct matrices

Establish the X matrix, Y matrix and the regression coefficient matrix β for the calculation and solution of the model, as well as the error matrix. The matrix representations are as follows:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{14}$$

The relationship among the four matrices can be expressed as:

$$Y = X \cdot \beta + \varepsilon \tag{15}$$

Among them, ε is the random error term. Generally, it is assumed to follow a normal distribution $\varepsilon \sim N(0, \sigma^2)$, that is, the multiple linear regression equation with an expected value of 0 can be written in the following expression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{16}$$

Among them, β_1-β_p can also represent the parameters of the multiple linear regression equation.

*STEP*3: Calculate the values of the regression coefficients β using the least squares method

After obtaining the matrix of Y through $Y = X \cdot \beta$, the values of β_0-β_p in β can be solved by using the least squares method. The solution formula is as follows:

$$\beta = (X^T X)^{-1} X^T Y \tag{17}$$

From this formula, the values of the intercept β_0 and the regression coefficients β_1-β_p can be calculated. From the 10 feature indicators obtained after dimension reduction, it can be known that $p = 10$. Therefore, the values of β_0 - β_p, that is, β_0-β_{10}, are respectively: -2.1619×10^{-2}, 3.6862×10^{-5}, 2.5578×10^{-5}, -2.9108×10^{-6}, -2.8920×10^{-2}, -2.5408×10^{-6}, 1.8527, -1.8238, 6.5842×10^{-7}, 4.9497×10^{-6}, -1.430.

3. RESULTS AND ANALYSIS

3.1 Establish the Logistic Regression Model for Prediction

By substituting the above-obtained parameter B into the logistic regression model to predict the test data set, the predicted probability $P(\hat{y}_i=1;\, B)$ of each sample can be obtained. Some of the predicted values for behavioral decision-making are shown in Table 1 as follows:

Table 1: Partial Predicted Result Values

Influencing Factors	Family Education Methods	Social and Cultural Differences	Family Economic Conditions	Social Competition Pressure	Employment Guidance and Support
Predicted Probability P of Behavioral Decision-making	4	1	3	2	1

The above table shows the predicted probability P of behavioral decision-making obtained by substituting the parameter B into the logistic regression model to predict the test data set.

Based on the obtained values of the regression coefficient B, the degree of influence can be judged by the magnitude of its absolute value. The larger the absolute value is, the more significant the influence will be. Thus, it can be determined which specific feature indicators have a greater impact on the predicted results of behavioral decision-making.

3.2 Establish the *k-means* Clustering Model for Clustering

Use the elbow method to determine the optimal value of the number of clusters *k*. Iteratively calculate the *SSE* (Sum of Squared Errors) values for the data set (with *k* ranging from 1 to 102). Obtain the within-cluster error *SSE* graph. The graph of the change of *SSE* with the value of *k* is shown in Figure 2 as follows:

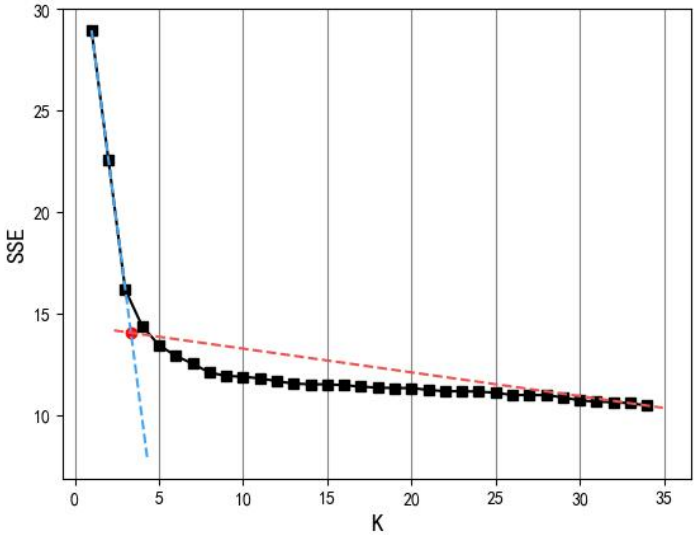

Figure 2: Graph of the Change of SSE with the Value of k

From the analysis of Figure 2, it can be seen that the *SSE* continuously decreases as the number of clusters *k* increases. When *k* < 3, the error decreases rapidly; when *k* > 3, the error decreases slowly. Therefore, when *k* = 3, the data structure is relatively stable and it is an inflection point of the graph. Since the number of clusters is an integer, the optimal number of clusters *k* = 3 is taken. Through calculation, when the number of clusters *k* = 3, the error *SSE* = 16.1759.

Then substitute the data and the optimal number of clusters *k* = 3 into the *k-means* model established in this paper, and the categories of each object and each clustering center can be obtained. The categories are represented by the numbers 0, 1, and 2 respectively.

3.3 Establish the Multiple Linear Regression Model

Substitute the data values of the 12 feature indicators obtained after dimension reduction into the function with an expected value of 0 in formula (12), that is, $y = \beta_0 + \beta_{1x1} + \beta_{2x2} + ... + \beta_{pxp}$, and then the predicted values of the influence degrees of different factors can be obtained. All the predicted result values are shown in Table 2 as follows:

Table 2: All Predicted Result Values

Influencing Factors	Predicted Values	Influencing Factors	Predicted Values	Influencing Factors	Predicted Values
Family Education Methods	4	Social Competition Pressure	3	Academic Pressure	4
Family Atmosphere	3	Social Expectations	1	Academic Competition	2
Family Structure	3	Social and Cultural Differences	1	Interpersonal Relationships	3
Family Economic Conditions	2	Social Media Influence	2	Employment Guidance and Support	2

The above table shows the predicted values of the influence degrees of different factors obtained by substituting the 12 feature indicator values obtained after reducing the dimensionality of the influencing factor data in a small range into the multiple linear regression equation.

4. CONCLUSION

Firstly, this paper used the logistic regression model to determine the predicted probabilities of the influence degrees of different factors on the mental health of college students, and verified the model. The accuracy rate of the test results is relatively high and the persuasiveness is relatively strong.

Secondly, the *k-means* clustering model was used to cluster the selected indicators, and the clustering centers were optimized. The obtained clustering results are better and the effect is more excellent.

Finally, through the 12 selected feature indicators after dimensionality reduction, a multiple linear regression model was constructed to predict the influence degrees of all factors again. It is applicable to a variety of different data sets and analysis requirements and can better fit complex data patterns.

FUND PROJECT

(1) This research is funded by the 2024 Digital Economy Special Research Project of Sichuan Association of Higher Education "Grey Clustering Correlation Analysis of the Regional Economic Development Level in Sichuan Province under the Trend of Digital Economy" (Project Number: SZJJ2024YB - 016).

(2) This research is funded by the 2024 Innovation and Entrepreneurship Training Program for College Students "Research on College Students' Mental Health Based on Logistic Regression Model" (Project Number: 202414389024).

(3) This research is funded by the 2024 Innovation and Entrepreneurship Training Program for College Students "Hong'an Under Heaven, Agriculture Ensures the Country - Application Research of Grey Prediction Model in the Development of Rural Revitalization Cross-border E-commerce Platform" (Project Number: 202414389012).

REFERENCES

[1] Huang Xinyi. Research on the Structure, Development Characteristics and Influencing Factors of College Students' Positive Psychological Energy and Its Relationship with Mental Health [D]. Guangzhou University, 2023.

[2] Liangqun Y. Mental Health Status and Influencing Factors of College Students [J]. International Journal of Fuzzy System Applications (IJFSA), 2023, 13(1).

[3] Li Yihe, Li Xueying, Liang Qianrong, et al. Research on the Satisfaction Degree of College Students' Mental Health Education and Its Influencing Factors [J]. Medicine and Philosophy, 2022, 43(16).

[4] Wang Zhenjie, Peng Qiushi, Chen Yujiang, et al. Analysis of Influencing Factors of College Students' Mental Health under the COVID-19 Epidemic Based on Random Forest Model [J]. Chinese Health Service Management, 2022, 39(03).

[5] Qian Yimian. Research on the Mental Health Status and Influencing Factors of College Students in Chengdu [J]. Heilongjiang Science, 2022, 13(23).

[6] Pan Miao, Zhang Sanqiang, Zhou Shengsheng, et al. Related Influencing Factors and Coping Styles of College Students' Mental Health under Stress [J]. Chinese Journal of Health Psychology, 2021, 29(02).

[7] Wu Haipeng. Analysis of the Main Influencing Factors of College Students' Mental Health in Universities [J]. Journal of Changchun University, 2021, 31(02).

[8] Lu Ping. Discussion on College Students' Academic Mental Health and Its Influencing Factors [J]. Heilongjiang Researches on Higher Education, 2015(04).

[9] Zhong Yuanjin, Qiu Yuan. Investigation and Analysis of the Relationship between College Students' Health Cognition and Mental Health [J]. Journal of Physical Education, 2012, 19(04).

[10] Gu Dacheng. Research on the Promotion Path of Lifestyle for College Students' Mental Health [J]. Journal of Shandong Sport University, 2011, 27(04).

[11] Zhang Bin, Zhang Xiaodong. Discussion on the Problem of the Impact of Economic Factors on College Students' Mental Health [J]. Technology and Innovation Management, 2009, 30(01).

[12] Pei Xuejin. Influencing Factors and Improvement Strategies for the Evaluation of College Students' Mental Health Status [J]. Jiangsu Higher Education, 2006(04).

[13] Xu Tonghai. Social Factors Affecting College Students' Mental Health [J]. Journal of Wuhan Institute of Physical Education, 2005(12).

[14] Pan Shubo. Research and Analysis on Various Factors Affecting College Students' Mental Health [J]. Journal of Chengdu Sport University, 2002, (01).

[15] Cao Rong. Analysis of Social and Psychological Factors Affecting College Students' Mental Health [J]. Journal of Northwest University (Philosophy and Social Sciences Edition), 2000(01).