International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 8075-8084

https,//doi.org/10.70517/ijhsa463690

Research on students' physical fitness assessment based on logistic regression and clustering algorithm

Bin Ge[1,*]
[1] Department of Physical Education, Nanjing Medical University, Nanjing, Jiangsu, 211166, China
Corresponding authors: (e-mail: gebinnjykdx@163.com).

Abstract Physical health is not only related to students' personal health, but also directly affects the quality of education and the future development of the country. In this paper, a method for assessing college students' physical fitness and health based on cluster analysis and logistic regression is proposed. First, the Relief algorithm is used to select features for students' physical health data, and the improved K-means clustering algorithm is used to classify the data and analyze the physical characteristics of different classes. The results show that the improved K-means algorithm is significantly better than the original K-means algorithm in clustering effect, and the profile coefficient and Dunn's index are 0.396658 and 0.043811, respectively, both of which are improved compared with the original algorithm. Then, the influencing factors of students' physical health assessment were further analyzed based on the logistic regression model. The results showed that dynamic behavioral time, sleep duration and quality, and dietary and nutritional status had a significant effect on students' physical health status, while static behavioral time and health knowledge did not significantly affect students' physical health. The AUC values of the ROC curves of the model were 0.77, 0.87, and 0.83, respectively, indicating that the model has a good assessment performance. Eventually, a series of recommendations to improve the physical health of college students were proposed based on the model.

Index Terms college students, physical health, cluster analysis, K-means algorithm, logistic regression model, health assessment

I. Introduction

In modern society, students' physical health has always been in the spotlight, and physical health plays an important role in students' learning and growth. In order to better understand students' physical condition, physical health assessment has become a necessary task [1]-[4]. The importance of students' physical health assessment is reflected in several levels [5]. First of all, physical fitness and health assessment can provide students with scientific and personalized physical activity suggestions [6]. Each student's physical condition is different, and targeted exercise can better help them develop their physical fitness and improve their physical function [7], [8]. Secondly, the assessment results can be provided, to students, parents and teachers as a basis for understanding the students' physical fitness and health status and making relevant plans [9], [10].

Finally, students' physical health is closely related to their academic performance [11]. By understanding students' physical condition through assessment, schools can provide appropriate health care measures and create a better learning environment for students [12]-[14]. Students' physical health assessment can be performed by a variety of methods, the most common of which are height, weight and body fat percentage measurements [15], [16]. Height and weight are important indicators for assessing students' growth and development, while body fat percentage can help to understand students' body fat content [17], [18]. And with the development of artificial intelligence, intelligent algorithms such as logistic regression and clustering algorithms are used in order to get a comprehensive understanding of students' fitness level [19], [20].

In recent years, the issue of college students' physical fitness and health has become a hot issue widely concerned by the society. With the increase of study pressure and bad life habits, there is a general trend of decline in the physical fitness level of college student groups. Physical health directly affects students' learning efficiency and quality of life, and may even have a far-reaching impact on their future development. However, current health assessment methods mainly rely on traditional physical examination data and lack a comprehensive analysis of students' physical fitness. In order to monitor and improve students' physical health more effectively, there is an urgent need to develop new assessment methods by combining modern data mining techniques. In this study, clustering analysis was first used to extract features from college students' physical fitness and health data. The feature selection was optimized by Relief algorithm and combined with the improved K-means clustering algorithm

to classify the data. The results of cluster analysis can reveal the physical characteristics and weaknesses of different groups of students, and then provide scientific basis for health intervention. Next, the study introduces the logistic regression model to assess the impact of students' behavior, living habits and environmental factors on physical fitness and health by analyzing them. By combining cluster analysis and logistic regression model, the study was able to more accurately assess the physical fitness and health status of college students, and provide a theoretical basis for the development of personalized health intervention programs.

II. Extraction of college students' physical fitness and health data based on cluster analysis

In this study, the extraction method of college students' physical fitness and health data based on cluster analysis is used to adjust the distance of features by Relief algorithm, and K-means algorithm is utilized to obtain the extracted college students' physical fitness and health data [21].

II. A.Relevant feature screening based on Relief algorithm

The two parts of data were studied by examining the normal data of college students' physical fitness as well as the physical health data among them. The two kinds of data are constructed as links according to time, and data cleaning is carried out, after which data mining is carried out. In the process of data cleaning, it mainly contains two parts of checking, which are the checking of null value and invalid value, and deleting the data with the existence of both. When the link is built, the information redundant values are effectively removed, this is due to the fact that there are various forms of differences in both the quantity and dimensions of the data, which are described in normalized form through equation (1):

$$q = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

In Equation (1), the sample value of the physical health data is described by x, the maximum value of the sample is described by $x_{\max}$, the minimum value of the sample is described by $x_{\min}$, and the final normalized form of the data is described by q.

Feature selection is performed on the processed data, i.e., “dimensionality reduction” of the normalized data. The Relief algorithm is selected to adjust the distance of features, separate dissimilar samples, and bring similar samples closer, i.e., when the classification results are influenced by the features, the weight of the features will be increased [22]. The algorithm is able to improve the speed and reduce the overhead of feature selection based on the statistical characteristics of the data when performing feature selection, which is more suitable for large datasets. In the final calculation, the corresponding weight of each feature can be obtained, and the relevance of the feature is related to the weight. In order to obtain a subset of relevant features, features with weights less than a given weight threshold can be filtered. The algorithm gives various feature weight thresholds to obtain various feature subsets to make the next clustering analysis more convenient, and analyzes the final clustering subject to the change of various feature weight thresholds.

II. B.Cluster analysis

The data were divided into two major categories, normal data and abnormal data, and three subsets of data with smaller ranges were also divided in the abnormal data because of the existence of different forms of abnormalities in college students' physical fitness. And the normal data are described by N_1, and the data subsets in abnormal data are described by A_1, A_2, and A_3, respectively. Since A_1 is closer to normal data, it can find the difference between normal data and abnormal data faster, therefore, the class cluster A_1 is collected firstly when profiling the threshold, through which the form can not only obviously divide the two classes of data, but also mine out the smaller class clusters and realize more refined clustering in abnormal data. K-means algorithm is used for clustering, which is calculated based on the distance, and the calculation process is simple and faster, and it is also able to give various k values, so that the final clustering results are different. In order to mine smaller class clusters, it can take the form of modifying the k value, but the algorithm is more meticulous in selecting the initial point with the disturbing data, and the algorithm ends when the objective function does not change or is not greater than some set threshold. The objective function is described through equation (2):

$$\min \sum_{i=1}^{k} \sum_{x \in c_i} dist(c_i, x)^2 \tag{2}$$

In Equation (2), the center of mass of the i th cluster is described by c_i; the interval between the center of mass and the sample x within the cluster c_i is described by $dist(c_i, x)$; and the number of clusters given is described by k.

II. C.Improved K-means clustering algorithm

Since the selection of K value is affected by the user's subjective intention and possesses randomness, for this reason, the selected K value of the K-means algorithm is improved.

II. C. 1) Improved selection of K values

Based on the clustering results of K-means, the following two parts of the presidential measures are calculated, namely the total X value and the V value. In this case, the sum of the sum of squared deviations of all clustered variables, denoted by X, is calculated using equation (3):

$$X = \sum_{i=1}^{k} \sum_{x \in c_i} dist(c_i, x)^2 \tag{3}$$

In Equation (3), the i th cluster is described by c_i; the points in c_i are described by x; the mean of the i th cluster is represented through c_i; and the spacing between two objects is represented by $dist$.

Between categories, the sum of the squared sums of the clustering variables' deviations is described by V and is calculated using equation (4):

$$V = \sum_{i=1}^{k} m_i dist(c_i, c)^2 \tag{4}$$

In Equation (4), the size of the cluster is denoted by m_i, the mean value of the i th cluster is denoted by c_i, the total mean value is described by c, and the meaning of $dist$ is consistent with the above.

When the value of K is known (the value represents the number of clusters), the clustering algorithm envisions having a smaller total X value with a larger total V value, which indicates that its intra-group data possesses higher aggregation ability and the inter-group data possesses better segmentation performance, i.e., the higher the value of $Total\ V / Total\ X$ is possessed, the stronger the performance is.

In order to keep the final calculation result from being changed by the sample n as well as the number of clusters K, the form of calculation of $Total\ V / Total\ X$ is adapted to Equation (5):

$$\frac{Total\ V}{Total\ X} \Big/ \frac{n-K}{K-1} \tag{5}$$

In Equation (5), the complexity is described by $(n-k)/(k-1)$, and its ratio is more excellent the higher it is, which is the Calinski-Harabasz formula, which has the characteristics of high arithmetic efficiency, and is therefore used to determine the final value of adaptation K.

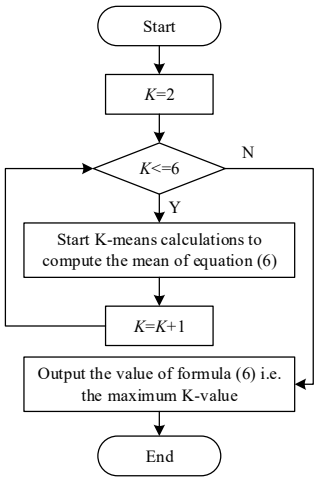

Figure 1: The process of selecting K value

II. C. 2) Process for selecting K values

The process of selecting K values is shown in Fig. 1. Based on the enumeration method, the K values are set in order, respectively 2~6, and the operation is repeated for 1000 times to prevent the phenomenon of local optimal solution from occurring, and the Calinski-Harabasz form of the K values is calculated, and finally the K value corresponding to the largest form of the Calinski-Harabasz value is taken as the final selected K value.

III. Data processing of students' physical fitness assessment based on cluster analysis

In this chapter, on the basis of clustering algorithm, we will use the extraction method of college students' abnormal physical fitness data based on cluster analysis proposed in this paper to cluster and analyze the physical fitness test data of freshmen, sophomores, and juniors in College A from 2022 to 2024, and complete the processing of students' physical fitness assessment data. The test data mainly covered five tests, including 50m, standing long jump, seated forward bending, 800m, and sit-ups.

III. A. Clustering sample size

In order to test the reasonableness of the data processed by the extraction method of college students' physical abnormality data based on cluster analysis proposed in this paper, the improved K-Means algorithm utilized in it is compared with the original K-Mean algorithm in terms of quality and effectiveness. The classification sample size of the improved K-Means algorithm and the original K-Mean algorithm is shown in Table 1. It can be seen that both the improved K-Means algorithm and the original K-Mean algorithm in this paper cluster the data into 6 classes, but the original K-Mean algorithm has the most sample size in class 1, which reaches 1176, while the improved K-Means algorithm in this paper has the most sample size in class 5, which is 1235.

Table 1: The sample size of k-Means clustering

Algorithm	Cluster	Sample size
K-Means algorithm	1	1176
	2	125
	3	224
	4	257
	5	464
	6	191
Improved K-Means algorithm	1	283
	2	169
	3	163
	4	530
	5	1235
	6	57

III. B. Comparative analysis of clustering result parameters

The results of the two algorithms, the improved K-Means algorithm and the original K-Mean algorithm in this paper, are compared in terms of parameters, as shown in Table 2. This time, 2 parameter indexes are introduced: contour coefficient and Dunn's index. The larger the value of contour coefficient, the more separated the clusters are from each other, and when its value ranges from -1 to 1, the larger the value is, the better the clustering effect is. And the larger Dunn's index represents the better clustering effect. It can be seen that the contour coefficient and Dunn's index of the improved K-Means algorithm in this paper are 0.396658 and 0.043811, respectively, and the contour coefficient value is in the range of -1~1 and is higher than that of the original K-Mean algorithm, and the Dunn's index value is higher than that of the original K-Means algorithm, which is 0.008199. Obviously, the improved K-Means algorithm of this paper has significant improvement compared with the K-Means algorithm. Means algorithm.

Table 2: Algorithm comparison

Algorithm	Contour coefficient	Dunn index
K-Means algorithm	0.370825	0.035612
Improved K-Means algorithm	0.396658	0.043811

III. C. Characterization of cluster structure

In the above analysis, it is known that the improved K-Means algorithm in this paper is more superior in clustering effect. In this section, the clustering structure characteristics of the improved K-Means algorithm of this paper will be analyzed, first of all, according to the improved K-Means algorithm to draw the 6 classes of change folding charts, specifically as shown in Figure 2. The main and secondary weak points of the 6-class grouping can be seen by observing the chart:

Class 1 college students: the main weak point is 800m, and the secondary weak point is sit-ups/pull-ups.

College students in category 2: the primary weakness is seated forward bends and the secondary weakness is sit-ups/pull-ups.

Category 3 students: primary weakness is sit-ups, no secondary weaknesses.

Category 4 students: primary weakness is sit-ups/pull-ups, no secondary weaknesses.

Category 5 students: all physical fitness indicators are not low, no major or minor weaknesses.

Category 6 students: primary weakness is standing long jump, secondary weakness is sit-ups/pull-ups.

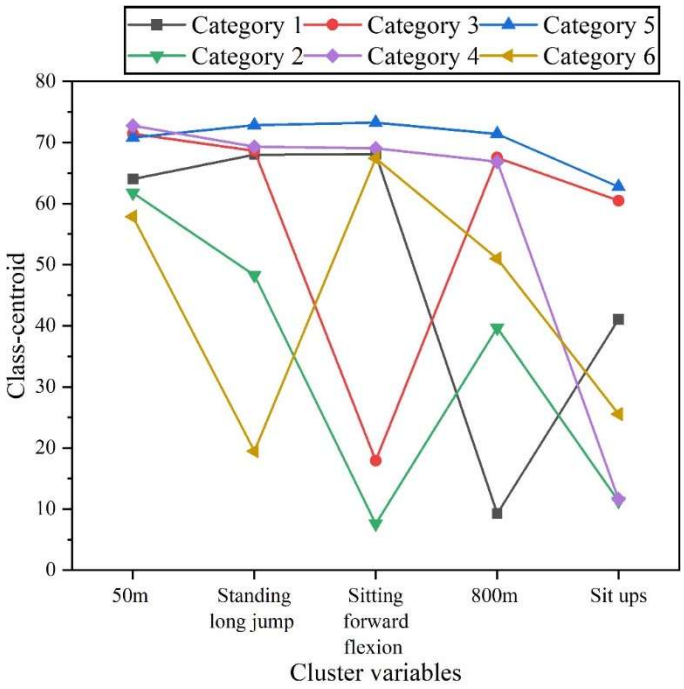

Figure 2: The mean change of various clustering variables

IV. Analysis of students' physical fitness assessment based on logistic regression modeling

In the above paper, this paper proposes the extraction method of college students' physical fitness and health data based on cluster analysis, improves the original K-means algorithm, and uses this method to cluster analyze the data related to the physical fitness and health assessment of the students in College A, and completes the processing of the students' physical fitness and health assessment data. In this chapter, the logistic regression model will be further used to analyze the students' physical fitness and health assessment on this basis. Next, the relevant principles of the logistic regression model will be presented to provide a methodological basis for the following regression analysis.

IV. A. Principles of logistic regression modeling

Logistic regression model is a common statistical analysis method to analyze the dependent variable as a qualitative variable, which is one of the more commonly used machine learning methods at present, and it can be used to make predictions about the likelihood of the occurrence of an event, as well as to classify it, belonging to a kind of probabilistic nonlinear regression method [23]. Because the logistic regression model does not require data normality, variance alignment and independent variable type, and has the advantages of coefficient interpretability, which makes it widely used in the fields of quantitative psychology, biostatistics, sociology, econometrics, and clinical medicine.

IV. A. 1) Linear regression

Linear regression is a method of statistical analysis that utilizes regression analysis in mathematical statistics to determine the interdependent quantitative relationship between two or more variables, which is widely used.

In regression analysis, only one independent variable and one dependent variable are included, and the relationship between the two can be approximated by a straight line, this regression analysis is called univariate linear regression analysis. It can be simply expressed as:

$$y = a + bx + c \tag{6}$$

where, y is the dependent variable, x is the independent variable, a is the constant term (intercept term), b is the regression coefficient, and c is the random error term.

If two or more independent variables are included in the regression analysis and there is a linear relationship between the dependent and independent variables, it is called multiple linear regression analysis. Suppose that a certain dependent variable y is affected by k independent variables $x_1, x_2, \cdots, x_k$ with n sets of observations $(y_a, x_{1a}, x_{2a}, \cdots, x_{ka})$ with $a = 1, 2, \cdots, n$. Then, the multiple linear regression model can be expressed in the following form:

$$y_a = \beta_0 + \beta_1 x_{1a} + \beta_2 x_{2a} + \cdots + \beta_k x_{ka} + \varepsilon_a \tag{7}$$

where, β_0 is a constant term, $\beta_1, \beta_2, \cdots, \beta_k$ are parameters to be determined, i.e., regression coefficients, and ε_a is a random variable.

IV. A. 2) Logistic function

The value of the function grows roughly exponentially at the beginning; then the increase slows down as it starts to become saturated; and finally, the increase stops when it reaches maturity. The function can be expressed as the following equation.

$$P(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

where x is the independent variable, $P(x)$ is the dependent variable, and e is the base of the natural logistic function.

The logistic function has a distinct S-type distribution, with $P(x) = 0$ when $x = -\infty$ and $P(x) = 1$ when $x = +\infty$, i.e., regardless of the value of x, the range of values of $P(x)$ is between 0 and 1.

IV. A. 3) Logistic regression models

(1) Formula for logistic regression model

Consider a vector $x = (x_1, x_2, \cdots, x_n)$ with n variables, and let the conditional probability $P(Y = 1 \mid x) = p$ be the probability of occurrence of an event according to the observations relative to an event, then the logistic regression model can be expressed as:

$$P(Y = 1 \mid x) = \pi(x) = \frac{1}{1 + e^{-g(x)}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \tag{9}$$

where β_0 is the intercept term, $\beta = (\beta_1, \beta_2, \cdots, \beta_n)$ is the regression coefficients of the independent variables, and x_n is the different independent variables. If $g(x)$ also contains nominal variables (categorical variables) it should be turned into a dummy variable, and a nominal variable with k values will be turned into $k-1$ dummy variables, so $g(x)$ becomes:

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{i=1}^{k-1} \beta_{ji} D_{ji} + \beta_n x_n \tag{10}$$

It can be seen that a logistic regression model is actually a probabilistic nonlinear regression model in which a linear regression is normalized by a Logistic equation, and $g(x)$ is often referred to as a linear function of a set of factors that affect the probability of an event occurring. Since $\pi(x)$ has a value domain of $[0,1]$, we can estimate the probability of occurrence when the dependent variable $Y = 1$ based on its value.

(2) Maximum Likelihood Estimation

Maximum likelihood estimation, also known as great likelihood estimation, the basic idea of this method is: when the model from the overall randomly selected m group of sample observations, the most reasonable parameter estimates should make the model from the m group of sample observations drawn from the probability of the largest. This is an iterative algorithm that takes a predictive estimate as the initial value of the parameter, determines the direction and variation of the parameter that increases the log-likelihood value according to the algorithm, and after estimating that initial function, tests the residuals and re-estimates them with an updated and improved function until the log-likelihood value no longer changes significantly.

Suppose there are m samples of observations with $y_1, y_2, \cdots, y_m$, and let $p_i = P(y_i = 1 \mid x_i)$ be the probability of obtaining $y_i = 1$ under the given conditions. The conditional probability of getting $y_i = 0$ under the same conditions is $P(y_i = 0 \mid x_i) = 1 - p_i$. Thus, the probability of an observation can be obtained as:

$$P(y_i) = p_i^{y_i}(1 - p_i)^{(1-y_i)} \tag{11}$$

Because the observations in Eq. (11) are independent, their joint distribution can be expressed as the product of the marginal distributions.

$$l(\beta) = \prod_{i=1}^{m} P(y_i) = \prod_{i=1}^{m} P(x)^{y_i}[1 - P(x)]^{1-y_i} \tag{12}$$

The above equation is commonly referred to as the likelihood function for m observations, and combined with equation (9), the regression coefficients in the logistic regression model can be obtained by finding the parameter estimates that maximize the value of equation (12). The log likelihood function can be obtained by finding the natural logarithm on both sides of equation (12):

$$\begin{aligned} L(\beta) &= In[l(\beta)] \\ &= \sum_{i=1}^{m}\left[y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}\right) - In\left(1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}}\right)\right] \end{aligned} \tag{13}$$

To solve for the values of the parameters $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$ when $L(\beta)$ obtains its maximum value, the above equation is usually derived, and the Newton-Raphson iterative method is applied to the nonlinear equations obtained by the derivation. Due to the complexity of the solution method, it is not described here, and in practice, it is usually calculated using SPSS software. Finally, the obtained parameters can be substituted into the (9) formula to establish the prediction model of logistic regression.

(3) Significance test

After estimating the parameters of the model, a significance test must be performed to determine whether the independent variable x_j in the logistic regression model is significantly correlated with the dependent variable.

IV. B. Construction of students' physical health assessment index system

Before formally carrying out the logistic regression analysis of students' physical health assessment, academic journals and Chinese and foreign language databases were consulted to conduct a literature review, screen the indicators of students' physical health assessment, and construct a system of students' physical health assessment indicators, as shown in Table 3. There are two first-level indicators, namely, physical form and physical quality, among which there are nine indicators under the first-level indicators of individual factors, including dynamic behavior time, static behavior time, sleep duration and quality, diet and nutritional status, health knowledge, mastery of motor skills, mental health status, physical exercise habits, self-efficacy, etc., and the first-level indicators of environmental factors are the quality of physical education courses, the configuration of physical education facilities, the organization of extracurricular physical exercise, the quality of physical education courses, and the quality of physical education facilities. configuration, frequency of extracurricular physical activity organization, family sports consumption input, demonstration of family sports habits, accessibility of sports resources on campus, mode of transportation, food safety on campus, supervision of sports activities, quality of the natural environment, and medical care, among other 11 indicators.

Table 3: Student physical health assessment index system

System	First-level indicators	Secondary indicators
Student system health evaluation index system	Individual factors	Dynamic behavior time
		Static behavior time
		Sleep duration and quality
		Nutritional status
		Health knowledge
		Mastery of motor skills
		Mental health status
		Physical exercise habits
		Self-efficacy
	Environmental factors	The quality of physical education curriculum
		Sports venues and facilities configuration
		Traffic mode of travel
		Family sports consumption investment
		Family exercise habits demonstration
		Extracurricular physical exercise organization frequency
		Accessibility of campus sports resources
		Campus food safety
		Supervision of sports activities
		Quality of natural environment
		Medical security

IV. C. Logistic regression analysis of students' physical fitness assessment

Based on the index system of students' physical fitness assessment constructed above, the logistic regression-based students' physical fitness assessment model was established by combining the principle of logistic regression model. The ROC curve is used to monitor the assessment effect of the assessment model, as shown in Figure 3. It can be seen that in the three tests, the AUC area under the ROC curve of the logistic regression model reaches 0.77, 0.87 and 0.83 respectively, and these values are larger than 0.5, which indicates that the model predicts and evaluates well.

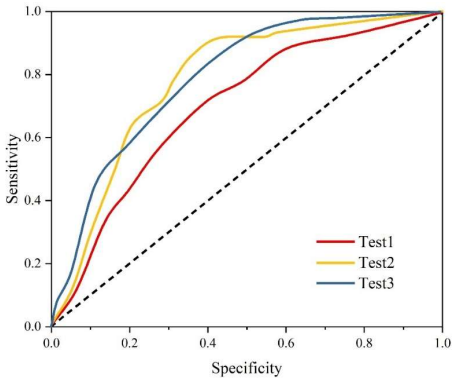

Figure 3: ROC curve

The results of the parameter assessment of the logistic regression model conducted are shown in Table 4. The results of the logistic regression analysis showed that the four variables such as static behavior time, health knowledge, physical activity facilities and campus food safety did not have a significant effect on students' physical fitness ($P>0.05$). In addition to these four variables, the remaining other variables such as dynamic behavior time, sleep duration and quality, and dietary and nutritional status all have a significant effect on students' physical fitness ($P<0.05$).

Table 4: Maximum likelihood estimation analysis table

Variable	Estimate	Std.Error	Z value	P-value
Dynamic behavior time	-0.01382	0.09	0.000131	0.02
Static behavior time	-0.0137	0.03	0.000347	0.539
Sleep duration and quality	6.110999	0.44	0.000225	0.021
Nutritional status	0.876142	0.26	0.005129	0.033
Health knowledge	0.876571	0.03	0.006244	0.586
Mastery of motor skills	0.790613	0.38	0.009756	0.01
Mental health status	0.788576	0.52	0.006774	0.012
Physical exercise habits	2.865611	0.3	0.004458	0.014
Self-efficacy	0.603681	0.66	0.009015	0.026
The quality of physical education curriculum	0.778692	0.06	0.00223	0.033
Sports venues and facilities configuration	0.781751	0.12	0.000569	0.089
Extracurricular physical exercise organization frequency	0.532328	0.12	0.001127	0.019
Family sports consumption investment	-1.12413	0.14	0.000125	0.024
Family exercise habits demonstration	-2.20245	0.26	0.000154	0.022
Accessibility of campus sports resources	2.203796	0.18	0.00423	0.02
Traffic mode of travel	-1.33965	0.15	0.000178	0.018
Campus food safety	0.000178	0.1	0.000486	0.051
Supervision of sports activities	-1.56051	0.23	0.00004	0.023
Quality of natural environment	-1.88961	0.33	0.00028	0.029
Medical security	2.667811	0.29	0.000392	0.03

Based on the results of the above regression of student characteristics, it can be seen that the following characteristics of students' physical fitness assessment were derived:

(1) In terms of physical form

Dynamic behavior time, dietary and nutritional status, sleep duration and quality, motor skill mastery, mental health status and other student morphological information have a significant impact on students' physical health assessment results. This indicates that in the cultivation of students' physical health, priority should be given to monitoring and guiding students' dynamic behavior time, diet and nutritional status, sleep duration and quality, mastery of motor skills, mental health status, etc., so as to improve students' physical health from the aspect of students' self-body shape.

(2) In terms of physical quality

Family sports consumption investment and family sports habit demonstration will have a greater impact on the assessment of students' physical fitness and health quality. As the above information indicates that in the cultivation of students' physical fitness and health, schools can cooperate with students' families to stimulate students' enthusiasm for exercise from the perspectives of family accompaniment and encouragement. The influence of physical activity supervision and medical care on students' physical fitness assessment results is very significant, and the frequency of organizing extracurricular physical activity also has a certain impact on students' physical fitness assessment results. This is because proper supervision of physical activity and medical care can provide a solid foundation for the improvement of students' satisfaction with physical activity, and to a certain extent, it can promote students' willingness to do physical activity.

V. Conclusion

The college students' physical fitness and health assessment method based on cluster analysis and logistic regression model proposed in this study can effectively and accurately assess students' physical fitness and health. The results of cluster analysis showed that the improved K-means algorithm was able to classify the data into different physical fitness groups more accurately and significantly improved the clustering effect, especially in data classification, the profile coefficient reached 0.396658 and the Dunn index was 0.043811, which was a significant advantage over the original algorithm. The analysis results of the logistic regression model further revealed the main influencing factors of students' physical health. Dynamic behavioral time, sleep duration and quality, and dietary and nutritional status significantly affected students' physical health, in which dynamic behavioral time and sleep quality had a negative effect on physical health, while dietary and nutritional status had a positive effect. In addition, family factors such as parents' investment in sports consumption and exercise habits, as well as the configuration of sports facilities in schools, also had a significant impact on students' physical fitness. In the future management

of students' physical fitness and health, multiple factors should be combined to promote individualized interventions to enhance the overall physical fitness level of college students.

References

[1] Prontenko, K., Griban, G. P., Dovgan, N., Loiko, O., Andreychuk, V., Tkachenko, P., ... & Bloshchynskyi, I. (2019). Students' health and its interrelation with physical fitness level. Sport Mont. International Scientifi c Journal, (17 (3)), 41-46.

[2] Bogomolova, E. S., Shaposhnikova, M. V., Kotova, N. V., Badeeva, T. V., Maksimenko, E. O., Kiseleva, A. S., ... & Olyushina, E. O. (2019). Characteristics of physical health of students of modern educational institutions. Hygiene and sanitation, 98(9), 956-961.

[3] Murphy, M. H., Carlin, A., Woods, C., Nevill, A., MacDonncha, C., Ferguson, K., & Murphy, N. (2018). Active students are healthier and happier than their inactive peers: the results of a large representative cross-sectional study of university students in Ireland. Journal of Physical Activity and Health, 15(10), 737-746.

[4] Shankar, N. L., & Park, C. L. (2016). Effects of stress on students' physical and mental health and academic success. International Journal of School & Educational Psychology, 4(1), 5-9.

[5] Bakiko, I., Savchuk, S., Dmitruk, V., Radchenko, 0., & Nikolaev, S. (2020). Assessment of the physical health of students of middle and upper grades. Journal of Physical Education and Sport, 20(1).

[6] Korol, S. A. (2014). Assessment of physical health and physical fitness of students of technical specialties of I course. Pedagogics, psychology, medical-biological problems of physical training and sports, 18(11), 23-29.

[7] Haas, J., Baber, M., Byrom, N., Meade, L., & Nouri-Aria, K. (2018). Changes in student physical health behaviour: an opportunity to turn the concept of a Healthy University into a reality. Perspectives in public health, 138(6), 316-324.

[8] Edwards, B., Traylor, A., & Froehle, A. (2022). Mental health symptoms, diagnoses, treatment-seeking, and academic impacts in student-athletes and non-athlete college students using the National College Health Assessment. Journal of Issues in Intercollegiate Athletics, 15(1), 20.

[9] Vihos, J., Chute, A., Carlson, S., Buro, K., Velupillai, N., & Currie, T. (2022). Virtual health assessment laboratory course delivery and nursing student clinical judgment: a mixed-methods exploratory study. Nurse Educator, 47(3), E51-E56.

[10] Roche, M. K., & Strobach, K. V. (2019). Nine Elements of Effective School Community Partnerships to Address Student Mental Health, Physical Health, and Overall Wellness. Coalition for Community Schools.

[11] Shaw, S. R., Gomes, P., Polotskaia, A., & Jankowska, A. M. (2015). The relationship between student health and academic performance: Implications for school psychologists. School Psychology International, 36(2), 115-134.

[12] Bass, R. W., Brown, D. D., Laurson, K. R., & Coleman, M. M. (2013). Physical fitness and academic performance in middle school students. Acta paediatrica, 102(8), 832-837.

[13] Del Pozo, F. J. F., Alonso, J. V., Álvarez, M. V., Orr, S., & Cantarero, F. J. L. (2017). Physical fitness as an indicator of health status and its relationship to academic performance during the prepubertal period. Health promotion perspectives, 7(4), 197.

[14] Santana, C. D. A., Azevedo, L. D., Cattuzzo, M. T., Hill, J. O., Andrade, L. P., & Prado, W. D. (2017). Physical fitness and academic performance in youth: A systematic review. Scandinavian journal of medicine & science in sports, 27(6), 579-603.

[15] Bodnar, I. R., Stefanyshyn, M. V., & Petryshyn, Y. V. (2016). Assessment of senior pupils' physical fitness considering physical condition indicators. Pedagogics, psychology, medical-biological problems of physical training and sports, 20(6), 9-17.

[16] Chen, S. L., & Liu, C. C. (2021). Development and evaluation of a physical examination and health assessment course. Nurse Education Today, 107, 105116.

[17] Huang, Y. (2021). THE EVALUATION OF STUDENTS'PHYSICAL HEALTH BASED ON THE INTEGRATION OF FAMILY AND SCHOOL PHYSICAL EDUCATION. Revista Brasileira de Medicina do Esporte, 27, 80-82.

[18] Fotynyuk, V. G. (2017). Determination of first year students' physical condition and physical fitness level. Physical education of students, 21(3), 116-120.

[19] Liu, D., Peng, L., & Zhao, Z. (2023). A review of intelligent methods of health assessment technology. Intelligence & Robotics, 3(3), 355-373.

[20] Tang, X., Li, F., Seetharam, T. G., & Vignesh, C. C. (2021). Internet of Things-assisted intelligent monitoring model to analyze the physical health condition. Technology and Health Care, 29(6), 1355-1369.

[21] Gyeong tae Gwak,Ui jae Hwang & Jun hee Kim. (2025). Clustering of shoulder movement patterns using K-means algorithm based on the shoulder range of motion. Journal of Bodywork & Movement Therapies,41,164-170.

[22] Supaporn, Lonapalawong & Jun, Zhang Le. (2017). Applying Relief Algorithm for Feature Selection in Sentiment Classification for Movie Reviews. Journal of Computational and Theoretical Nanoscience,14(11),5418-5423(6).

[23] M. Cherifi,M.N. El Korso,S. Fortunati,A. Mesloub & L. Ferro Famil. (2025). Robust inference with incompleteness for logistic regression model. Signal Processing,236,110027-110027.