International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-16

https://doi.org/10.70917/ijcisim-2026-0005
Article

Multivariate Statistical Analysis of Athletes' Physical Fitness Data and Optimization of Sports Teaching and Training Strategies

Hui Zhang [1] and Chunhua Li [2,*]

[1] Department of Physical Education, Donghua University, Shanghai 201620, China
[2] Department of Physical Education, TongJi University, Shanghai 200092, China; coulinran@163.com

Abstract: The development of physical quality is of great significance in enhancing one's physical health, mastering sports skills, and improving sports level. In this study, the main factors affecting athletes' physical fitness were investigated by principal component analysis, a multiple regression model was constructed to realize the prediction of athletes' physical fitness, and the validity of the model was verified by partial least squares. The 17 indicators of athletes' physical fitness test were analyzed by principal component analysis, and the important factors affecting athletes' physical fitness were categorized into physical characteristics, morphological development, physiological functions and flexibility and endurance. After analyzing the correlation between each factor and verifying that each factor affects athletes' physical fitness, a regression model was constructed to predict athletes' physical fitness. The partial least squares regression equation was used to predict the four quantitative indexes of physical quality, and the scatter plots of the four indexes were basically symmetrically distributed on the diagonal, indicating that the multiple regression model constructed in this paper is able to predict the physical quality of athletes. In order to help athletes achieve higher quality development, this paper adopts the stratified training method to further optimize the training strategy of physical education.

Keywords: multiple regression model; principal component analysis; partial least squares; physical fitness

1. Introduction

As sports events become increasingly intense, the physical fitness of athletes has attracted much attention. Athletes' physical fitness is the basis for their excellent performance on the field of play. Good physical fitness before and during the competition not only allows athletes to maintain abundant energy during the competition, but also enhances their endurance, speed, strength and agility, which gives them an advantage in the competition [1–3].

Physical fitness generally includes the speed, strength, endurance, agility, coordination, flexibility and other related functions embodied in sports or daily physical activities. People who are often engaged in physical labor will have stronger physical fitness than those who do not work regularly, and at the same time, insisting on physical fitness can also better improve people's physical fitness, which is determined by genetic factors on the one hand, and hard training also plays a great role on the other hand, and the impact of different training contents and methods on physical fitness is also significantly different [4–7].

For athletes, physical quality is not only affected by innate genetic factors, but also acquired scientific training, reasonable nutritional intake and the cultivation of the corresponding psychological aspects are important ways to improve athletes' physical quality [8,9]. Therefore, the results of the analysis of athletes' physical fitness data can be used as an important reference for athletes' training. Coaches can accurately understand the advantages and deficiencies of each athlete according to the physical quality data, and then carry out targeted training to help athletes improve their performance. At the same time, due to the existence of a certain link between physical fitness and sports injuries, through the analysis of

physical fitness data can prevent injuries and make athletes' physical training more scientific and effective [10-13].

In the traditional analysis of physical quality data, the evaluation of a single indicator, the lack of specialization, and the feedback lag phenomenon is serious, which is not conducive to the adjustment of training strategies. In addition, because the physical quality data is a combination of speed, strength, endurance, sensitivity, coordination, flexibility, etc., the integration and utilization of each other is low, and it is difficult to develop a scientific training plan for athletes [14]. And multivariate statistical analysis through the analysis of multiple variables (or multiple factors) in the objective things, to explore the statistical regularity of the interdependence between the variables, in the high-dimensional data processing, nonlinear modeling and individualized diagnosis has great advantages, for the analysis of athletes' physical fitness data to open up new paths [15-17].

In order to construct a multiple regression model to realize the prediction of athletes' physical quality, the study firstly uses principal component analysis to extract the main factors affecting athletes' physical quality from 17 variables. Then the correlation between the factors was analyzed to verify the correlation between the predictors and the athletes' physical quality, and then after regression analysis, the regression model was constructed to analyze the significance level of each influence factor and explore the degree of influence of each factor on the athletes' physical quality. Then the partial least squares regression equation was applied to test the predictive effect of the model. Finally, the stratified training method is proposed to optimize the training strategy of physical education.

2. Subject and Methodology of the Study

2.1. Subjects of Study

A total of 1,350 current Chinese college athletes with an average age of 22 years were used as the study population.

2.2. Data Acquisition

Athletes were tested for physical fitness in the following categories: height, weight, lung capacity, endurance category (pull-ups), endurance category (1000 meter run), flexibility and strength category (seated forward bending), and speed and dexterity category (standing long jump). The physical fitness of the athletes in each event was used as the source of data for the study.

2.3. Research Methodology

2.3.1. Principal Component Analysis Methods

The principal component analysis method [18] is used, and its basic steps are as follows:

(1) Normalized collection of raw indicator data P -dimensional random vector $x=(x_1,x_2,\ldots,x_p)^T$, n samples, $x_i=(x_{i1},x_{i2},\ldots,x_{ip})^T$, $i=1,2,\ldots,n$. p, construct the sample array, and apply the following normalization transformation to the sample array elements:

$$Z_{ij}=\frac{x_{ij}-\overline{x_j}}{s_j},i=1,2,\ldots,n;j=1,2,\ldots,p \tag{1}$$

where $\overline{x_j}=\frac{\sum_{i=1}^{n}x_{ij}}{n}$, and $s_j^2=\frac{\sum_{i=1}^{n}(x_{ij}-\overline{x_j})^2}{n-1}$ to obtain the normalized array Z.

(2) Find the correlation coefficient matrix for the normalized array:

$$R=[r_{ij}]_{p\times p}=\frac{Z^TZ}{n-1} \tag{2}$$

where $r_{ij}=\frac{\sum Z_{kj}gZ_{ki}}{n-1}$, $i,j=1,2,\ldots,p$.

(3) Solve the characteristic equation of the sample correlation matrix: $\left|R-\lambda I_p\right|=0$ to get p eigenroots and determine the principal components. Then the m values are determined by

$\frac{\sum_{j=1}^{m}\lambda_j}{\sum_{j=1}^{p}\lambda_j} \geq 0.85$, which makes the utilization rate of information to be 85% or more, and for every λ_j, $j=1,2,\ldots,m$. Solve the system of equations $Rb=\lambda_j b$ to obtain the unit eigenvector b_j^0.

(4) Convert the standardized indicator variables into principal components:

$$U_{ij}=Z_i^T b_j^0,\ j=1,2,\ldots,m \tag{3}$$

where U_1 is called the first principal component, U_2 is called the second principal component, ... and U_p is called the p th principal component.

(5) Finally, the m principal components are synthesized and evaluated. The m principal components are weighted and summed to obtain the final evaluation value, and the weights are the variance contribution ratio of each principal component.

2.3.2. Multiple Regression Models

The multiple regression model can be used to study the relationship between an explanatory variable and multiple explanatory variables with the following expression:

$$y=\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+\mu \tag{4}$$

where y is the explanatory variable, α is the intercept term, which represents the average effect of the variables not included in the model on the explanatory variables, x_1 to x_k is the explanatory variable, β_1 to β_k is the partial regression coefficient, taking β_1 as an example, β_1 represents the change in the mean of y by changing one unit x_1 under the condition that all other variables remain constant, and μ is the error term.

The commonly used parameter estimation method is ordinary least squares (OLS), and the basic idea of OLS solution is to minimize the sum of squares of the differences between the true observed and fitted values of the explanatory variables [19].

The goodness of fit (R^2) measures how well the multiple regression model fits the sample data, and the higher the R^2 the better the fit. However, if the explanatory variables in the regression equation are increased, R^2 will only increase but not decrease. Therefore most of the studies use corrected R^2 (adjusted R^2), which penalizes too many explanatory variables by adjusting the degrees of freedom.

Conducting multiple regression analysis requires attention to the following aspects:

(1) Multicollinearity problem

Strict multicollinearity exists if an explanatory variable can be linearly tabulated by other explanatory variables. Strict multicollinearity leads to unrecognizable regression coefficients, and results cannot be derived. In addition, there may be approximate (non-strict) multicollinearity among the explanatory variables. If the k th explanatory variable x_k is regressed on the rest of the explanatory variables and a high value of goodness of fit R_k^2 is obtained, there is an approximate multicollinearity in the explanatory variable x_k. The OLS estimator under approximate multicollinearity is still the best linear unbiased estimate, but the variance increases, making the coefficients inaccurate, and also leading to irrationality in some of the coefficient estimates. The high degree of correlation between the explanatory variables also does not make it easy to distinguish the extent of their respective effects on the explanatory variables.

Therefore, a multicollinearity test is needed when performing multiple regression, and the Variance Inflation Factor (VIF) is able to measure multicollinearity, which is calculated by the formula:

$$VIF=\frac{1}{1-R_j^2} \tag{5}$$

where R_j^2 is the goodness of fit of the regression of the j th variable x_j against all other variables.

If the value of VIF is larger then it proves that the correlation between the variables is larger. When VIF is greater than 10 then the equation is considered to have serious multicollinearity.

(2) Form of model construction

In addition to the form of multiple regression model shown in Equation (4), the explanatory variables and the explained variables in multiple regression can be expressed in logarithmic terms, which means that the elasticity of the original explained variables to the explanatory variables is a percentage change rather than a numerical change. A model in which both explanatory and interpreted variables take logarithmic values is called a double logarithmic model, as shown in Equation (6), where x increases by 1% for every 1% increase in y, the average change in $b\%$. Models in which only one of the explanatory and interpreted variables takes logarithm are called semi-logarithmic models, as shown in equations (7) and (8). Equation (7) is the explanatory variable taking logarithms, when x increases by 1%, y changes on average by $b/100$ units. Equation (8) takes logarithms for the explanatory variables and when x increases by 1 unit, y changes by (100b)% on average, i.e.,:

$$\ln y = \alpha + \beta \ln x + \mu \tag{6}$$

$$y = \alpha + \beta \ln x + \mu \tag{7}$$

$$\ln y = \alpha + \beta x + \mu \tag{8}$$

In addition, in order to compare the relative importance of the influencing factors and remove the effect of the scale, standardized regression coefficients can be obtained using standardized regression equations. The standardized regression equation is solved in the same way as the original multiple regression, but the data are standardized. The standardized data is calculated as:

$$x_i' = \frac{x_i - \overline{x}}{s} \tag{9}$$

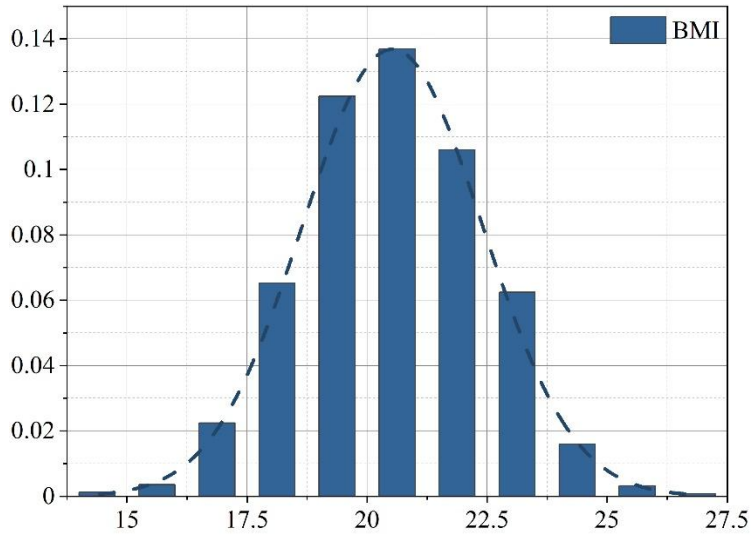
where x_i is the raw data, $\overline{x}$ is the mean of the variable of that type, and s is the standard deviation.

The standardization process does not affect the standard error and significance of the regression coefficients, and the larger the absolute value of the standardized coefficient, the greater the effect of the variable on the explanatory variables.

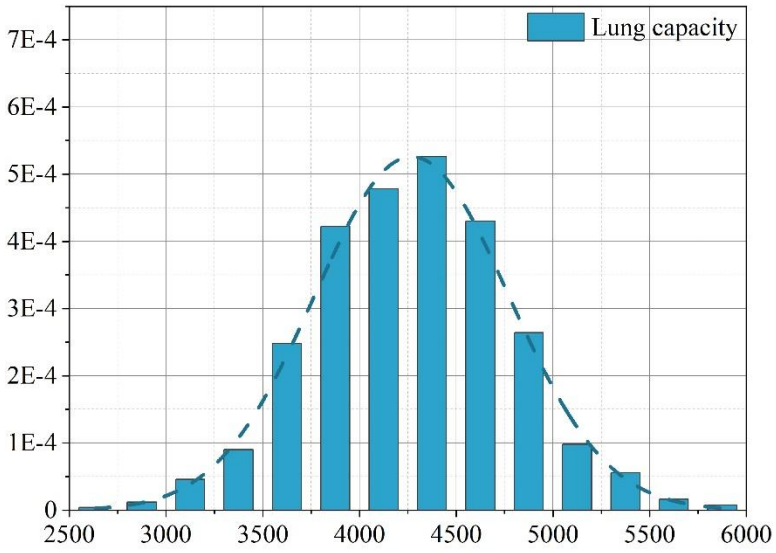
3. Empirical analysis

3.1. Data Visualization

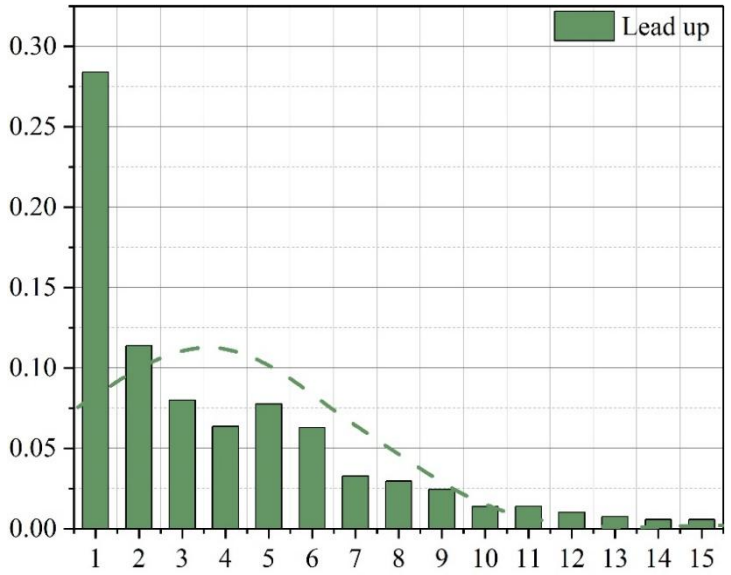
First of all, through the descriptive analysis of IBM SPSS statistics 25 software, the maximum value, minimum value and quartile value of the data were obtained, and based on the sports professional knowledge and experience of physical fitness test to find the obvious abnormal data for deletion, such as 1000 meters running more than 10,000 seconds, the height of 40cm or less and the record of the negative value of the situation of the index data. Summarize the athlete's physical health data summary table and analyze the data processing situation, according to the results of the analysis to choose the next step of data processing methods, data processing situation is shown in Figure 1, (a) ~ (f) represent the BMI, lung capacity, pull-ups, 1000 m running, seated forward bending, and standing long jump data statistics of the six items, respectively. The mean values of each item were 20.5, 4250mL, 3.5, 265s, 15cm, and 205m, respectively.



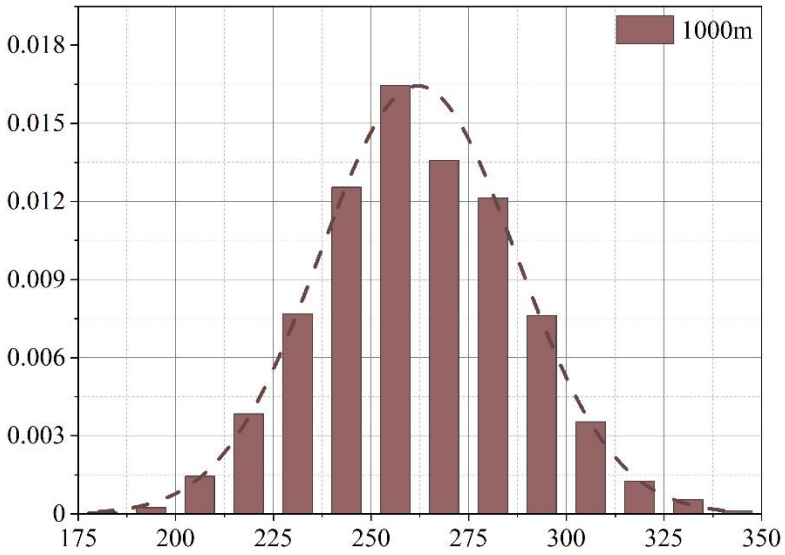
(a) BMI



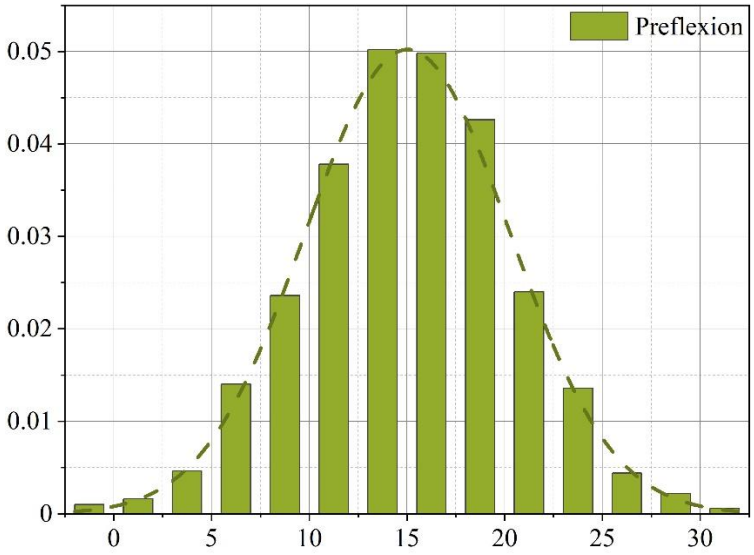
(b) Lung capacity



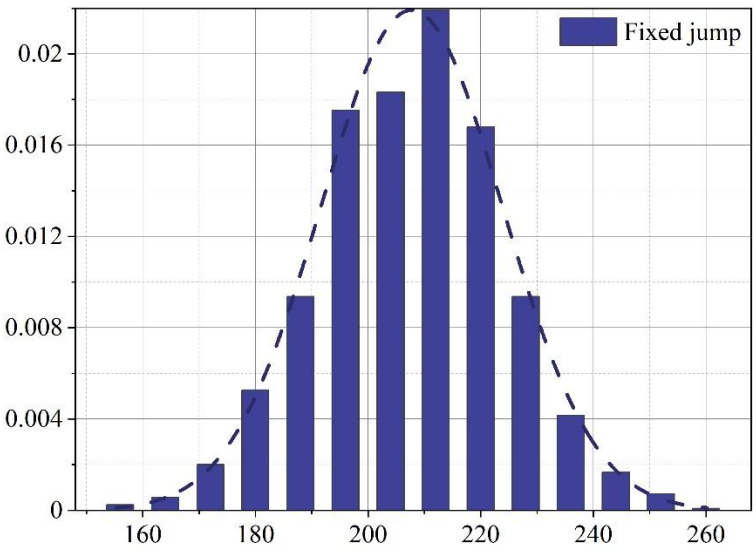
(c) Lead up



(d) 1000m



(e) Preflexion



(f) Fixed jump

Figure 1. Physical health data distribution of athletes.

3.2. Principal Component Analysis

3.2.1. KMO Test, Bartlett's Test of Sphericity

Principal component analysis is used to summarize the variables with strong correlations from multiple original variables to obtain a small number of representative factor variables. If the correlation between the original variables is low, the public factor variables cannot be extracted from them. Therefore, before conducting principal component analysis, KMO test and Bartlett's spherical test are needed to determine whether it is suitable for principal component analysis. When the KMO value is closer to 1, the stronger the partial correlation between the variables, the better the analysis effect is.The KMO value is above 0.7, the analysis result is acceptable, and the analysis effect is the best when it is greater than 0.9.The results of the KMO and Bartlett's test are shown in Table 1. The KMO value of this study is 0.903, which indicates that the analytical results are better, and the Sig. of Bartlett's test statistic for spherical test is <0.01, thus rejecting the null hypothesis that the correlation matrix is a unit array, i.e., there will be a significant correlation between the variables. In conclusion, the variables selected for this study are very suitable for principal component analysis.

Table 1. Inspection of KMO and Bartlett.

The Kaiser-Meyer-Olkin metric of the sampling is sufficient		0.903
Bartlett’s Test of Sphericity	Approximate card	18800.362
	DF	169
	Sig.	0.000

3.2.2. Common Factor Extraction and Rotation Analysis

The variance of the common factor is shown in Table 2. It indicates the degree of extraction of information from the original variables by the public factors the degree of information extraction of the 17 variables is basically above 60%. Therefore, it can be considered that the extracted public factors can explain the original variables better.

Table 2. Common factor variance.

Project	Initial	Extraction
Height (X1)	1.000	0.702
Weight (X2)	1.000	0.675
Altitude (X3)	1.000	0.738
Chest circumference (X4)	1.000	0.623
Waistline (X5)	1.000	0.745
Hip circumference (X6)	1.000	0.649
Pulse (X7)	1.000	0.636
Systolic pressure (X8)	1.000	0.678
Diastolic pressure (X9)	1.000	0.581
Upper arm (X10)	1.000	0.662
The thickness of the skin of the shoulder blade (X11)	1.000	0.594
Abdominal thickness (X12)	1.000	0.754
Lung capacity (X13)	1.000	0.664
Lead up (X14)	1.000	0.655
1000 meters run (X15)	1.000	0.696

Preflexion (X16)	1.000	0.615
Fixed jump (X17)	1.000	0.593

Scatterplot is used to show the importance of each factor in principal component analysis, and its horizontal axis is the serial number of the factor, and the vertical axis indicates the size of the eigenroot. In this study, the public factors were extracted according to the default criterion that the eigenroot is greater than 1. The extraction is shown in Fig. 2, the scatters of the first four factors are located on steep slopes and have eigenvalues greater than 1, and the scatters of the remaining 13 factors form a platform, so from the fragmentation diagram, it is concluded that the first four public factors should be considered in this study.

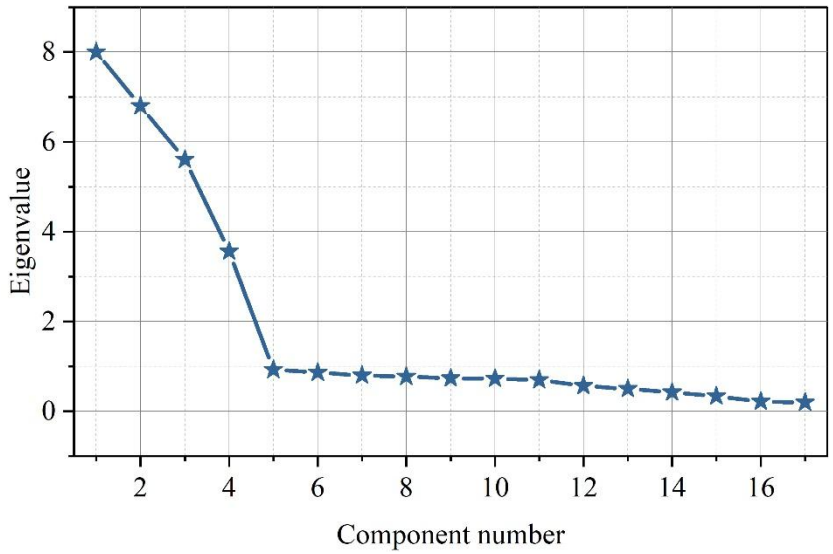


Figure 2. Gravel figure.

The data were subjected to automatic extraction of principal components, rotated by the maximum variance method, and the total variance explained is shown in Table 3. As can be seen from the initial eigenvalues, there are four factors with eigenvalues greater than 1, and the rest of the factors with initial eigenvalues less than 1 can be ignored. The cumulative contribution rate of variance is 85.041%, indicating that the four extracted principal components can better respond to most of the information of the 17 indicators in the original variables. In the later study, the 17 indicators can be replaced by 4 principal components for the study.

Table 3. The total variance of the explanation.

Con.	Initial eigenvalue			Extract the sum of squares and load			Rotate the squares and load		
	Tot	Variance %	Cumulative %	Tot	Variance %	Cumulative %	Tot	Variance %	Cumulative %
1	7.997	47.041	47.041	7.997	47.041	47.041	7.126	41.918	41.918
2	3.293	19.371	66.412	3.293	19.371	66.412	4.021	23.653	65.571
3	1.605	9.441	75.853	1.605	9.441	75.853	1.694	9.965	75.536
4	1.562	9.188	85.041	1.562	9.188	85.041	1.616	9.505	85.041
5	0.322	1.894	86.935						
6	0.318	1.871	88.806						
7	0.297	1.747	90.553						
8	0.297	1.747	92.3						

9	0.295	1.735	94.035						
10	0.261	1.535	95.57						
11	0.233	1.371	96.941						
12	0.227	1.335	98.276						
13	0.096	0.565	98.841						
14	0.071	0.418	99.259						
15	0.066	0.388	99.647						
16	0.034	0.2	99.847						
17	0.026	0.153	100						

The rotated component matrix is shown in Table 4. It can be seen that the factor loadings of each indicator in the four principal components, the higher the loading coefficient and the closer the absolute value to 1, it means that the public factor contains more information about the variable, the closer the relationship between the factor and the original variable, and the indicator is categorized to the public factor.X2 weight variable in the first principal component and the second principal component factor loadings are greater than 0.6, in order to better carry out the analysis, this study will X2 indicator excluded from the principal component categorization. Combined with Table 3 and Table 4, it can be seen that the maximum variance contribution rate of the first principal component is 41.918%, and the indicators with the highest correlation with the first principal component are X1, X3, X13, X14, X17 which are five original variables, namely height, sitting height, lung capacity, pull-up, and standing long jump respectively, and based on the characteristics of the original variables, the first principal component is categorized as the physical characteristics factor. The variance contribution rate of the second principal component was 23.653%, in which the six variables X4, X5, X6, X10, X11, and X12 had a high correlation with it, which were chest circumference, waist circumference, hip circumference, sebaceous thickness of the upper arms, sebaceous thickness of the scapulae, and cortical thickness of the abdomen, which were grouped into the factor of morphology and development based on the characteristics of the original variables. The third principal component had a variance contribution of 9.965%, with three variables with high correlation, X7, X8, and X9, which were pulse, systolic blood pressure, and diastolic blood pressure, and it was named as physiological function factor. The variance contribution rate of the fourth principal component was 9.505%, among the variables, X15 and X16 had higher factor loading values, which were seated forward bending and 1000 meter endurance running, respectively, naming them as flexibility and endurance factors.

Table 4. Rotational composition matrix.

Project	Con.			
	1	2	3	4
X17	0.933			
X1	0.873			
X13	0.844			
X14	-0.825			
X3	0.821			
X2	0.713	0.653		
X12		0.811		
X11		0.806		
X6		0.801		
X5		0.779		
X4		0.716		

X10		0.689		
X9			0.806	
X7			0.669	
X8			0.625	
X15				0.714
X16				-0.701

The matrix of component score coefficients is shown in Table 5. It reflects the scores of each variable on the four principal components. From Table 5, the four principal component function expressions are shown:

$$\begin{cases} F_1 = 0.134X1 + 0.056X2 + 0.037X3 - 0.005X4 + \cdots - 0.013X17 \\ F_2 = -0.018X1 + 0.166X2 + 0.194X3 + 0.203X4 + \cdots - 0.041X17 \\ F_3 = -0.047X1 - 0.018X2 - 0.023X3 + 0.003X4 + \cdots + 0.02X17 \\ F_4 = 0.073X1 - 0.063X2 - 0.017X3 - 0.058X4 + \cdots + 0.609X17 \end{cases} \tag{10}$$

Table 5. Component score coefficient matrix.

Project	Con.			
	1	2	3	4
X1	0.134	-0.018	-0.047	0.073
X2	0.056	0.166	-0.018	-0.063
X3	0.037	0.194	-0.023	-0.017
X4	-0.005	0.203	0.003	-0.058
X5	-0.084	-0.047	0.473	0.17
X6	0.042	0.006	0.393	-0.062
X7	-0.035	0.006	0.573	-0.138
X8	-0.112	0.187	0.007	0.043
X9	-0.028	0.212	-0.012	0.059
X10	-0.08	0.235	0.025	0.036
X11	0.13	-0.003	-0.019	-0.066
X12	-0.158	0.047	0.045	-0.013
X13	-0.024	0.009	0.072	-0.577
X14	0.147	-0.061	-0.011	-0.026
X15	0.123	-0.024	-0.011	-0.042
X16	-0.147	0.023	0.047	-0.14
X17	-0.013	-0.011	0.02	0.609

The variance contribution rates of the four principal components were used as weights to construct a comprehensive evaluation function of the factors affecting athletes' physical fitness and health:

$$F = 0.5365F_1 + 0.3426F_2 + 0.1455F_3 + 0.1485F_4 \tag{11}$$

By bringing the principal component factor scores of each tester into this composite rating function, a comprehensive evaluation of the tester's physical fitness level can be analyzed.

3.3. Correlation Analysis of Influencing Factors

The correlation coefficient is the basis of the regression equation and the factors are correlated and analyzed to prepare the follow-up accordingly. The regression model made is meaningful when the correlation between variables is high. The correlation coefficient test of factors affecting athletes' physical fitness is shown in Figure 3. It can be seen that there is a certain degree of correlation between the factors, which indicates that the predictive factors are correlated with the physical quality of athletes, and regression analysis can be done.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17
X17	-0.006	0.033	0.42	-0.894	0.257	0.6	-0.363	-0.492	-0.729	0.018	0.442	-0.303	-0.211	0.093	0.496	-0.731	1
X16	0.057	-0.684	0.226	-0.565	0.362	0.345	-0.37	-0.513	-0.55	0.094	0.298	-0.788	-0.656	-0.369	0.485	1	-0.731
X15	0.129	-0.258	0.022	-0.239	0.466	0.256	-0.375	-0.493	-0.369	0.22	0.169	-0.317	-0.111	0.527	1	0.485	0.496
X14	0.209	0.136	-0.18	0.058	0.357	0.871	-0.358	-0.507	-0.178	0.356	0.783	0.202	0.45	1	0.527	-0.369	0.093
X13	0.319	0.541	-0.383	0.402	-0.135	0.619	-0.374	-0.512	0.063	0.454	-0.14	0.694	1	0.45	-0.111	-0.656	-0.211
X12	-0.281	-0.51	-0.164	-0.306	-0.808	0.365	-0.375	-0.501	-0.761	-0.422	0.063	1	0.694	0.202	-0.317	-0.788	-0.303
X11	-0.291	-0.51	-0.157	-0.323	-0.791	-0.674	-0.352	-0.496	-0.754	-0.421	1	0.063	-0.14	0.783	0.169	0.298	0.442
X10	0.176	0.343	-0.192	0.32	0.073	0.556	0.669	0.276	0.054	1	-0.421	-0.422	0.454	0.356	0.22	0.094	0.018
X9	0.425	0.387	0.445	0.764	0.354	0.042	0.236	0.798	1	0.054	-0.754	-0.761	0.063	-0.178	-0.369	-0.55	-0.729
X8	0.23	0.781	-0.242	0.399	0.252	0.317	0.151	1	0.798	0.276	-0.496	-0.501	-0.512	-0.507	-0.493	-0.513	-0.492
X7	0.298	0.322	0.052	-0.191	0.806	0.41	1	0.151	0.236	0.669	-0.352	-0.375	-0.374	-0.358	-0.375	-0.37	-0.363
X6	0.466	0.369	-0.381	0.358	0.343	1	0.41	0.317	0.042	0.556	-0.674	0.365	0.619	0.871	0.256	0.345	0.6
X5	0.414	0.42	0.511	0.411	1	0.343	0.806	0.252	0.354	0.073	-0.791	-0.808	-0.135	0.357	0.466	0.362	0.257
X4	0.737	0.691	-0.082	1	0.411	0.358	-0.191	0.399	0.764	0.32	-0.323	-0.306	0.402	0.058	-0.239	-0.565	-0.894
X3	-0.255	-0.299	1	-0.082	0.511	-0.381	0.052	-0.242	0.445	-0.192	-0.157	-0.164	-0.383	-0.18	0.022	0.226	0.42
X2	0.61	1	-0.299	0.691	0.42	0.369	0.322	0.781	0.387	0.343	-0.51	-0.51	0.541	0.136	-0.258	-0.684	0.033
X1	1	0.61	-0.255	0.737	0.414	0.466	0.298	0.23	0.425	0.176	-0.291	-0.281	0.319	0.209	0.129	0.057	-0.006

Figure 3. Test of correlation coefficient of athletes' physical quality factors.

3.4. Regression Modeling and Analysis

Multiple comparative analysis of the factors affecting the physical quality of athletes was carried out, and the results of the analysis are shown in Table 6. It can be seen that physical characteristics, morphological development, physiological functions and flexibility and endurance all have a significant effect on athletes' physical quality. In order to explore the corresponding influence rate of athletes' physical quality influencing factors, regression analysis was used for interpretation.

Table 6. The physical quality influences multiple comparisons.

	Mean difference	Standard error	Sig.	95% confidence interval	
				Upper limit	Lower limit
Physical characteristics	-0.026	0.063	0.000**	-0.23	-0.16
Morphological development	-0.355*	0.076	0.023**	-0.051	0.011
Physiological function	0.362*	0.077	0.001**	0.14	0.58
Flexural endurance	0.328*	0.071	0.002***	0.11	0.52

The 4 equation model is shown in Table 7. Among the factors affecting athletes' physical fitness, the

physical characteristics factor can explain 2.1% of the physical test scores, the explanatory power is unchanged by adding the morphological development factor, the explanatory power rises to 3.5% by adding the physiological function factor, and the explanatory power is 3.9% by adding the flexibility and endurance factor. It reflects that physical characteristics, physiological functions, and flexibility and endurance have a greater influence on athletes' physical fitness.

Table 7. Equation model.

Model	R	R^2	Adj_R^2
1	0.155[a]	0.021	0.02
2	0.016[b]	0.021	0.02
3	0.187[c]	0.035	0.034
4	0.193[d]	0.039	0.038

After interpreting the equation model, the overall significance of the fitted equation also needs to be verified and the test results are shown in Table 8. The F value is 6.053 and the significance level of the model is 0.000, which indicates that the regression model is meaningful.

Table 8. The overall significance of the fitting equation.

Model	Sum of squares	Freedom	Mean square	F	Sig.
Regression	44.635	18	0.411	6.053	0.000
Residual error	7.632	563	0.076	-	-
Total	51.635	623	-	-	-

Physical fitness test scores were used as the dependent variable and physical characteristics, morphological development, physiological functions and flexibility and endurance were used as predictor variables. The significance levels of the influencing factors are shown in Table 9. Each factor was entered into the regression model.

Table 9. Significance level of influencing factors.

Model	Unnormalized coefficient		Normalization factor	T	Sig.
	B	Standard error	Beta		
Constants	1.996	0.677	-	2.942	-0.02
X1	-0.02	0.021	-0.101	-1.093	0.022
X2	-0.097	0.09	-0.118	-0.961	0.027
X3	0.045	0.02	0.139	1.743	0.082
X4	0.017	0.029	0.067	0.557	0.059
X5	0.002	0.025	0.028	0.179	0.099
X6	-0.014	0.029	-0.054	-0.599	0.065
X7	-0.055	0.016	-0.171	-1.842	0.072
X8	-0.015	0.02	-0.037	-0.509	0.064
X9	0.052	0.059	0.103	1.586	0.059
X10	0.026	0.074	0.088	2.431	0.061
X11	0.125	0.064	0.19	1.378	0.016
X12	-0.15	0.083	-0.205	-1.982	0.045
X13	0.043	0.065	0.04	0.42	0.08

X14	0.027	0.045	0.02	0.341	0.066
X15	0.142	0.089	0.402	1.888	0.06
X16	-0.188	0.064	-0.399	-2.381	0.01
X17	0.02	0.04	0.032	0.222	0.082

3.5. Partial Least Squares Regression Analysis

In order to visualize the influence of each independent variable on the dependent variable, the histogram of regression coefficients was plotted using Matlab software, and the results were plotted as shown in Figure 4. The histogram of regression coefficients reflects the degree to which each dependent variable (physical fitness index) is influenced by each independent variable.

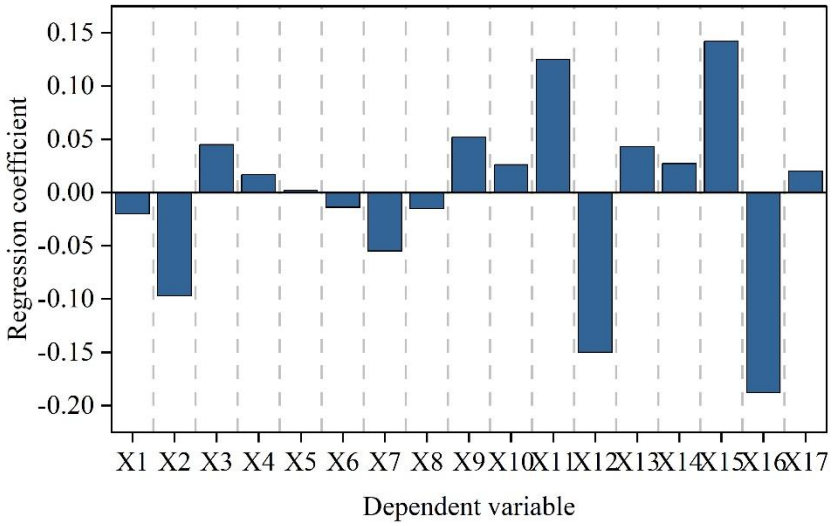


Figure 4. Histogram of regression coefficients.

In order to test the reliability of the regression model, the processed original data are substituted into the regression equation to obtain the predicted values of the dependent variable, and the prediction plots are plotted for all the sample points. On the prediction graph, if all the points can present a uniform distribution near the diagonal of the graph, the fitted value of the equation is less different from the original value, and then the fitting effect of the equation is more satisfactory.The prediction drawn by MATLAB software is shown in Fig. 5, and (a)~(d) represent the prediction of physical characteristics, morphological development, physiological function and flexibility and endurance on the athletes' physical fitness, respectively.The prediction of the four indicators The scatter plot is basically distributed symmetrically about the diagonal line, and the prediction effect is more satisfactory.

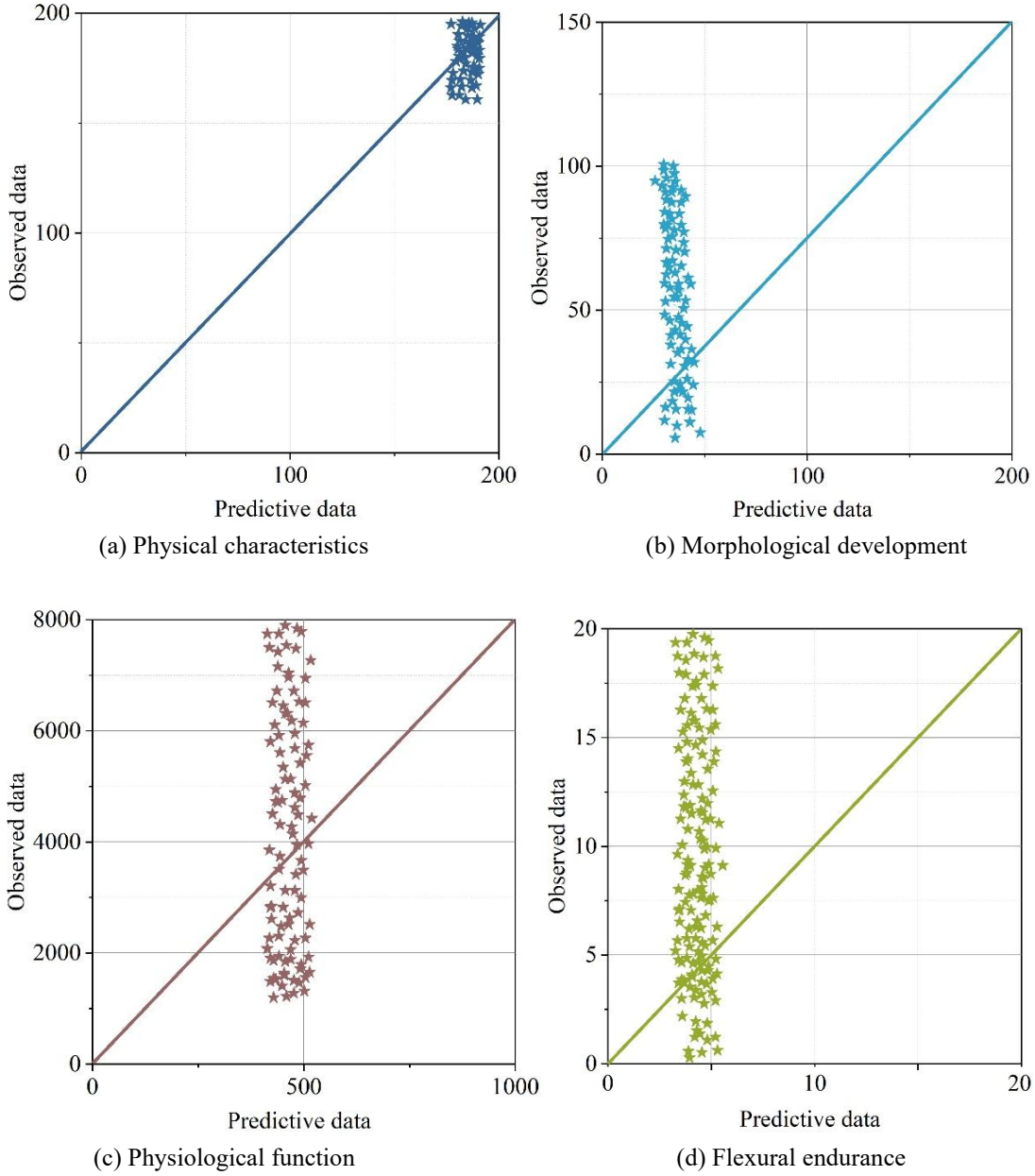


Figure 5. Data forecast graph.

4. Optimization of Training Strategies for Teaching Physical Education

Based on the above multivariate statistical analysis of athletes' physical fitness data, this chapter proposes the use of stratified training method to optimize the training strategy of physical education.

4.1. Scientific Classification of Athletes

First of all, the coaches know the physical fitness of the athletes by organizing physical testing activities. The contents of the physical test include height, weight, lung capacity, 50-meter sprint, 800-meter (female)/1000-meter (male) long-distance running, sit-ups, sit-ups (female)/pull-ups (male), standing long jump, etc. The coach records the physical test data in detail. The coaches recorded the athletes' physical test data in detail and scored the athletes' physical fitness level according to the athletes' physical test data. Secondly, the coach learns about the athletes' interest in sports and exercise frequency through one-on-one conversations or questionnaires. For example, the coach understands the athletes' physical exercise situation by distributing questionnaires, which include the number of times they participate in physical exercise per week, the time they participate in physical exercise per week, their attitudes towards physical exercise, and their favorite physical exercise programs. Finally, the coach tests

the learning effect and knowledge mastery of the athletes before organizing training activities, and fine-tunes the stratification results according to the test results to further enhance the scientificity and accuracy of the athlete's stratification [20].

4.2. Innovative Tiered Training Models

Coaches innovate the layered training mode to bring novel and interesting training experience for athletes and stimulate athletes' interest in participating in sports training activities. Athletes are the main body of teaching, stimulating athletes' interest in training is conducive to enhancing the enthusiasm of athletes to participate in sports training activities, and improving the efficiency and quality of layered training. The implementation of the current sports stratified training method can be divided into five steps, respectively, according to the training objectives to analyze the athletes' motor skills, and the macro training objectives are decomposed into specific and clear micro training objectives. According to the athlete's learning situation to match the different difficulty of the training program. Grading the training content and scientifically determining the number of training sessions and training time. Integrate and analyze the training results of the athletes, and make adjustments and guidance according to the analysis results. Evaluate the training results of athletes through competitions or examinations, and provide guidance for the improvement of subsequent hierarchical training and teaching.

4.3. Optimizing Tiered Training Evaluation

The coach optimizes the evaluation of stratified training, constantly improves and upgrades the evaluation system of stratified training teaching in the light of the actual application of the stratified training method, and gives full play to the diagnostic and guiding roles of educational evaluation. Firstly, the coach evaluates the athletes according to their sports attitude, sports ability and comprehensive sports level in sports training, so as to enhance the scientificity, fairness and objectivity of the evaluation. Secondly, the coach follows the principle of teaching according to the ability of the athlete, designs differentiated teaching evaluation programs for different levels of athletes, carries out a comprehensive evaluation of the athletic ability, cognitive level and physical quality of the athletes at all levels, and puts forward specific and practical improvement suggestions for the athletes according to the evaluation results to help athletes achieve better development. Finally, when the coach carries out the evaluation of stratified training, he pays attention to exploring the development potential of the athletes, and uses the teaching evaluation to stimulate the athletes' motivation and sense of challenge, and strives to make each athlete more capable, stronger and healthier.

5. Conclusion

This paper uses principal component analysis to study the test data of 17 indicators of athletes' physical fitness, establishes a regression model, analyzes the key factors affecting athletes' physical fitness, and then seeks to optimize the strategy of sports teaching and training. The conclusion of the study shows that:

(1) In the research related to each influencing factor and physical health, physical characteristics, morphological development, physiological functions and flexibility and endurance all have a significant impact on athletes' physical quality.

(2) Using physical fitness test scores as the dependent variable and physical characteristics, morphological development, physiological functions and flexibility endurance as the predictor variables, each factor entered the regression model.

(3) The partial least squares method was used to test the prediction effect of the model, and the predicted scatter plots of the four indexes were basically symmetrically distributed about the diagonal line, which made the prediction effect more satisfactory.

Finally, this paper puts forward the solution ways to improve and enhance athletes' physical quality from three aspects, namely, scientific division of athletes' level, innovation of stratified training mode and optimization of stratified training evaluation.

References

1. KOHMURA, Y. (2020). The Effects of Physical Fitness and Competition Experience on the Performance and Health of Athletes. Juntendo Medical Journal, 66(Suppl. 1), 101-107.
2. Ibáñez, S. J., Piñar, M. I., García, D., & Mancha-Triguero, D. (2023). Physical fitness as a predictor of performance during competition in professional women's basketball players. International journal of environmental research and public Health, 20(2), 988.
3. Supriadi, D., Friskawati, G. F., & Karisman, V. A. (2023). Physical fitness of futsal athletes in competition preparation. International Journal of Human Movement and Sports Sciences, 11(1), 71-76.

4. Sung, B. J., & Ko, B. G. (2017). Differences of physique and physical fitness among male south korean elite national track and field athletes. International Journal of Human Movement and Sports Sciences, 5(2), 17-26.
5. Simon, J. E., & Docherty, C. L. (2017). The impact of previous athletic experience on current physical fitness in former collegiate athletes and noncollegiate athletes. Sports health, 9(5), 462-468.
6. Cerit, M., Dalip, M., & Yildirim, D. S. (2020). Genetics and athletic performance. Research in Physical Education, Sport & Health, 9(2).
7. Gebel, A., Prieske, O., Behm, D. G., & Granacher, U. (2020). Effects of balance training on physical fitness in youth and young athletes: a narrative review. Strength & Conditioning Journal, 42(6), 35-44.
8. Kokarev, B., Kokareva, S., Atamanuk, S., Terehina, O., & Putrov, S. (2023). Effectiveness of innovative methods in improving the special physical fitness of qualified athletes in aerobic gymnastics. Journal of Physical Education and Sport, 23(3), 622-630.
9. Yerzhanova, Y., Madiyeva, G., Sabyrbek, Z., Dilmakhambetov, E., & Milašius, K. (2020). Can a high-energy diet affect the physical fitness of elite athletes?. Pedagogika/Pedagogy, 139(3), 239-252.
10. Till, K., Morris, R., Emmonds, S., Jones, B., & Cobley, S. (2018). Enhancing the evaluation and interpretation of fitness testing data within youth athletes. Strength & Conditioning Journal, 40(5), 24-33.
11. Yu, Q. (2024). Performance assessment and fitness analysis of athletes using decision tree and data mining techniques. Soft Computing, 28(2), 1055-1072.
12. Farley, J. B., Barrett, L. M., Keogh, J. W., Woods, C. T., & Milne, N. (2020). The relationship between physical fitness attributes and sports injury in female, team ball sport players: a systematic review. Sports medicine-open, 6, 1-24.
13. Li, X., Chen, X., Guo, L., & Rochester, C. A. (2022). Application of big data analysis techniques in sports training and physical fitness analysis. Wireless Communications and Mobile Computing, 2022(1), 3741087.
14. Jiang, B., Sun, H., Bai, W., Li, H., Wang, Y., Xiong, H., & Wang, N. (2018, July). Data analysis of soccer athletes' physical fitness test based on multi-view clustering. In Journal of Physics: Conference Series (Vol. 1060, No. 1, p. 012024). IOP Publishing.
15. Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining, 2(1), 20-30.
16. Koerich, A. C. C., Borszcz, F. K., Thives Mello, A., de Lucas, R. D., & Hansen, F. (2023). Effects of the ketogenic diet on performance and body composition in athletes and trained adults: A systematic review and Bayesian multivariate multilevel meta-analysis and meta-regression. Critical reviews in food science and nutrition, 63(32), 11399-11424.
17. Burdukiewicz, A., Pietraszewska, J., Stachoń, A., & Andrzejewska, J. (2017). Anthropometric profile of combat athletes via multivariate analysis. The Journal of sports medicine and physical fitness, 58(11), 1657-1665.
18. Xiaoping Xie. (2014). Based on principal component analysis (PCA) of the influence factors of sports dance development research. BioTechnology: An Indian Journal,10(11),
19. Richard A. Ashley & Christopher F. Parmeter. (2020). Sensitivity Analysis of an OLS Multiple Regression Inference with Respect to Possible Linear Endogeneity in the Explanatory Variables, for Both Modest and for Extremely Large Samples. Econometrics,8(1),11-11.
20. Binbin Zhu & Zeng Li. (2022). Research on the Development Strategy of Sports Training and Physical Education Teaching in Junior Middle Schools. Advances in Educational Technology and Psychology,6(7),