Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 2, 2024

Performance Analysis of Universities Based on Combination Weighting Method and Grey Correlation Analysis

Yimeng Liang, Hui Wang

Henan University of Science and Technology, China

Abstract: The performance evaluation of universities is of great significance for improving their educational level and allocating funding resources. This article constructs a performance evaluation model for universities based on the combination weighting method and grey correlation analysis. The model first uses principal component analysis and average mutual information method to select indicator features and select indicators with high importance; On this basis, the grey correlation analysis method is used to obtain the relationship between input indicators and output indicators; Finally, develop a visual performance analysis platform for universities using Python language and validate it with examples to conduct research. The research results indicate that the model and platform can effectively and intuitively display evaluation results, and can assist education management departments in identifying shortcomings and formulating targeted improvement strategies based on evaluation results.

Keywords: Performance evaluation of universities; Principal component analysis; Average mutual information; Grey correlation analysis.

1. Introduction

Performance management in universities provides a new means for evaluating the input and output of public investment in higher education. It plays an irreplaceable role in the rational allocation of university resources and promoting the healthy development of universities. At the same time, it also provides important guidance for the government to allocate financial funds. Performance evaluation, as a key link in performance management, the scientific nature of evaluation methods determines the reliability of evaluation results. In traditional performance analysis models, there are redundant indicators in the evaluation index system of universities, and the evaluation index data of universities presents characteristics of high dimensionality and complexity. In order to reduce the dimension of indicators, important features can be selected based on the size of the indicator weight coefficient. The indicator weight reflects the position of each indicator in the multi-attribute evaluation process and directly affects the reliability and objectivity of the evaluation results. The methods for determining indicator weights are generally divided into two categories: subjective weighting method and objective weighting method. Common methods include Analytic Hierarchy Process, Least Squared Method, Entropy Method, Standard Deviation Method, Critical Method, Complex Correlation Coefficient Method, Principal Component Analysis Method, etc. Although the process of subjective weighting is simple, the weights obtained often have strong individual subjectivity; The objective weighting method is based on various mathematical models and methods. In previous studies, many researchers have chosen one method for research, but for the same problem, there will be significant differences in the results obtained by different methods. Therefore, based on existing performance data of universities, a model combining principal component analysis and average mutual information method is proposed to determine indicator weights and perform feature selection. The principal component analysis method transforms the original indicators into a small number of new indicators to replace the original information. When determining weights, it does not rely on subjective judgment by humans, but objectively determines weights based on the correlation or variance contribution rate between various indicators. This helps to avoid the influence of subjective factors on the results, making the analysis results more objective and reliable. However, there may also be limitations in certain situations, such as when there are extreme or missing values in the data, which may affect the accuracy of the analysis results. Mutual information is a method that measures the correlation between indicators for both discrete and continuous data. It can not only measure the linear relationship between indicators, but also their nonlinear and non functional relationships. In addition, mutual information has strong generalization and fairness, and is suitable for any form of function between variables. For the same level of noise in different forms of functions, the same results can be obtained. By combining the above methods, we can more objectively seek the relationship between indicators and achieve optimal decision-making, thereby obtaining the combination weights. Further explore the relationship between input indicators and output indicators using the grey correlation method. Based on the results, identify the main factors that affect output, and enable decision-makers to adjust funding planning, organizational structure, and optimize resource allocation according to the results, in order to achieve the goal of improving performance. In the past, researchers had certain limitations in improving and applying algorithms, often only analyzing specific datasets and presenting research results in the form of text charts. If there is a change in data, it will be difficult for non researchers to reproduce the process. To address this issue, a performance analysis platform for universities has been developed using Python language. Users only need to organize the data in the format and upload it, and can intuitively obtain the weight of each indicator calculated by the above model and the relationship between indicators. Facilitate decision-makers to obtain performance evaluation results, grasp insights into development trends, and reasonably determine future development directions.

2. Technical Methods and Models

2.1. Principal Component Analysis

Principal component analysis is an unsupervised dimensionality reduction method that uses dimensionality reduction techniques to transform multiple variables into a few principal components in statistical analysis. The advantage of principal component analysis is that the calculated principal components can reflect the vast majority of information about the original variables. In the calculation process, it is not affected by factors outside the dataset, and the results are more objective. But the obtained principal components are brand new variables with fuzzy characteristics, which are not as clear and precise as the original variables. To solve the above problems, we can use the eigenvalues and eigenvectors of the principal components, as well as their respective variance contribution rates, to calculate the weights of the original variables. The process is as follows:

(1) Obtain data

Construct an m * n performance indicator data matrix X, where m represents the number of data, n represents the number of evaluation indicators included in each data, and the corresponding data for each variable is denoted as $X_1, X_2, X_3 ... X_n$.

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix} = (X_1, X_2 ..., X_n) \tag{1}$$

(2) data standardization

Different evaluation indicators often have different dimensions and units, which can affect the results of data analysis. In order to eliminate the dimensional influence between indicators, data standardization is necessary. The processed data for each indicator is in the same order of magnitude, making it suitable for comprehensive comparative evaluation. Here, we use the zero mean method (z-score) to process the data and obtain data that follows a standard normal distribution with a mean of 0 and a standard deviation of 1.

$$S - Core_{ij} = \frac{x_{ij} - \bar{x_j}}{s_j} \tag{2}$$

$$\bar{x_j} = \frac{1}{m} \sum_{i=1}^{m} x_{ij} \tag{3}$$

$$s_j = \sqrt{\frac{\sum_{i=1}^{m} (x_{ij} - \bar{x_j})^2}{m}} \tag{4}$$

(3) Find the covariance matrix

The covariance matrix can help us understand the relationship between the distribution and features of data. For two-dimensional random variables, the covariance matrix can tell us the correlation between two variables, including positive correlation, negative correlation, and their strength. If the covariance between two variables is positive, then they are positively correlated; If the covariance is negative, then they are negatively correlated; If the covariance is 0, then they are independent.

$$r_{ij} = cov(x_{ki}, x_{kj}) = \frac{\sum_{k=1}^{m} x_{ki} x_{kj}}{m-1} \tag{5}$$

$$C = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) \\ cov(x_2, x_1) & cov(x_2, x_2) \end{bmatrix} \tag{6}$$

(4) Calculate the eigenvalues and eigenvectors of the covariance matrix

The characteristic equation for constructing the covariance matrix can be used to calculate the eigenvalues and eigenvectors. The characteristic equation is: $C - \lambda E = 0$. Solve its characteristic values: $\lambda_1 \geq \lambda_2 \geq \lambda_3 ... \geq \lambda_n > 0$. The feature vector is:

$$\beta_1 = \begin{bmatrix} \beta_{11} \\ \cdots \\ \beta_{n1} \end{bmatrix}, \ \beta_2 = \begin{bmatrix} \beta_{12} \\ \cdots \\ \beta_{n1} \end{bmatrix}, \ ... \beta_n = \begin{bmatrix} \beta_{1n} \\ \cdots \\ \beta_{nn} \end{bmatrix} \tag{7}$$

(5) Determine the number of principal components

Determine the contribution threshold and use the principle of obtaining enough original information with fewer principal components to determine the number of principal components. Generally, values such as 0.8, 0.85, 0.9, 0.95, 0.99, etc. are taken. When the cumulative contribution rate of the current p principal component eigenvalues is higher than this value, it can be considered that these p principal components can represent the original n variables.

$$\frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq \alpha \tag{8}$$

(6) Calculate principal components

After obtaining the number of principal components, the eigenvectors corresponding to the first p eigenvalues can be used to calculate the principal components.

$$F_i = \beta_{1i}{}^T X_1 + \beta_{2i}{}^T X_2 + ... + \beta_{ni}{}^T X_n, i = 1,2, ..., p \tag{9}$$

(7) Determine the coefficients of indicators in linear combinations of principal components

Using the eigenvalues and eigenvectors corresponding to the first P principal components, calculate the index coefficients corresponding to each principal component.

$$\sigma_i = \frac{\beta^i}{\sqrt{\lambda_i}} = \begin{bmatrix} \frac{\beta_1^i}{\sqrt{\lambda_i}} \\ \frac{\beta_2^i}{\sqrt{\lambda_i}} \\ \frac{\beta_3^i}{\sqrt{\lambda_i}} \\ \vdots \\ \frac{\beta_n^i}{\sqrt{\lambda_i}} \end{bmatrix}, \quad \sigma = (\sigma_1, \sigma_2, \cdots, \sigma_p) = \begin{bmatrix} \frac{\beta_1^1}{\sqrt{\lambda_1}} & \frac{\beta_1^2}{\sqrt{\lambda_2}} & \cdots \frac{\beta_1^p}{\sqrt{\lambda_i}} \\ \frac{\beta_1^1}{\sqrt{\lambda_1}} & \frac{\beta_1^2}{\sqrt{\lambda_2}} & \cdots \frac{\beta_1^p}{\sqrt{\lambda_i}} \\ \vdots & \vdots & \vdots \\ \frac{\beta_n^1}{\sqrt{\lambda_1}} & \frac{\beta_n^2}{\sqrt{\lambda_2}} & \cdots \frac{\beta_n^p}{\sqrt{\lambda_i}} \end{bmatrix} \tag{10}$$

(8) Using the variance contribution rate of principal components to determine the coefficients of the comprehensive score model

$$\gamma_i = \frac{\sum_{j=1}^{n} \varphi_i \sigma_{ij}}{\sum_{k=1}^{p} \varphi_k} \tag{11}$$

$$Y = \gamma_1 X_1 + \gamma_2 X_2 + \cdots + \gamma_n X_n \tag{12}$$

(9) Normalization of indicator weights

$$\omega_i = \frac{\gamma_i}{\sum_{i=1}^{n} \gamma_i} \tag{13}$$

2.2. Average maximum information coefficient

The maximum information coefficient is a method used to measure the nonlinear correlation between two variables. Its main idea is that if there is a certain correlation between the two variables, a grid of a specific scale is used in the scatter plot of the joint sample of the two variables. Based on the marginal probability density function and the joint probability density function in the grid, the mutual information value of the two variables can be calculated. The normalized result can detect the correlation between the two variables. Mutual information is an indicator that measures the relationship between two variables, indicating the degree of information sharing between the two variables.

(1) Mutual information

P (x, y) is the joint probability density function, p (x), p (y) are the marginal probability density functions of two variables X and Y, I (X, Y) is the joint probability density function, p (x, y) is the relative entropy of the product probability p (x) p (y) of the two variables. The calculation formula is as follows:

$$I(X,Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y) log \frac{p(x,y)}{p(y)p(x)} \tag{14}$$

(2) Formula for calculating the maximum information coefficient

In the following equations, a and b are the number of grids divided in the x and y directions, and B is a constant, usually set to the power of 0.6 of the sample size n.

$$\mathrm{MIC}(X,Y) = \max_{a*b<B} \frac{I(X,Y)}{\log_2 \min(a,b)} \tag{15}$$

(3) Average maximum information coefficient

The average maximum information coefficient reflects the importance of a variable in the indicator set. For any random variable indicator X, the mutual information between it and other variable indicators in the same dataset can be calculated. Let Z be the mean of the maximum information coefficient sum between any variable and other variables. The larger the value of Z, the higher the importance of X; The smaller the Z value, the lower the importance of X.

$$Z = \frac{\sum_{i=1}^{n} \mathrm{MIC}(X,Y_i)}{n} \tag{16}$$

2.3. Grey correlation analysis

Grey relational analysis is a quantitative method for describing the future development trend of a system. The essence of grey relational analysis is to compare the degree of similarity between the curves of the reference sequence and the comparison sequence in the time series, and the comparison sequence that is closer to the reference sequence curve has a greater degree of correlation. The importance of evaluation indicators in the entire evaluation system can be horizontally seen through the magnitude of the correlation degree.

(1) Determine the reference sequence and comparison sequence

A data sequence that reflects the behavioral characteristics of a system is called a reference sequence. A data sequence composed of factors that affect system behavior is called a comparative sequence.

(2) Data processing

Dimensionless processing of reference and comparison sequences may result in different dimensions of data due to the different physical meanings of various factors in the system, making it difficult to compare or draw correct conclusions during comparison. Therefore, when conducting grey correlation analysis, dimensionless data processing is generally required.

(3) Calculate the grey correlation coefficient

The degree of correlation is essentially the degree of difference in geometric shapes between curves. Therefore, the difference between curves can be used as a measure of the degree of correlation.

$$\delta_i(k) = \frac{\min\min|x_0(k) - x_s(k)| + \rho \max\max|x_0(k) - x_s(k)|}{|x_0(k) - x_s(k)| + \rho \max\max|x_0(k) - x_s(k)|} \tag{17}$$

(4) Determine the grey correlation degree

$$\gamma_i = \frac{1}{n} \sum_{k=1}^{n} \delta_i(k) \tag{18}$$

3. Data Visualization Analysis Platform

3.1. Overall System Architecture Design

The overall architecture of the data visualization analysis platform based on combination weighting method and grey correlation analysis is shown in the figure. It consists of terminal layer, visualization analysis layer, service layer, computing architecture layer, and data layer from top to bottom.

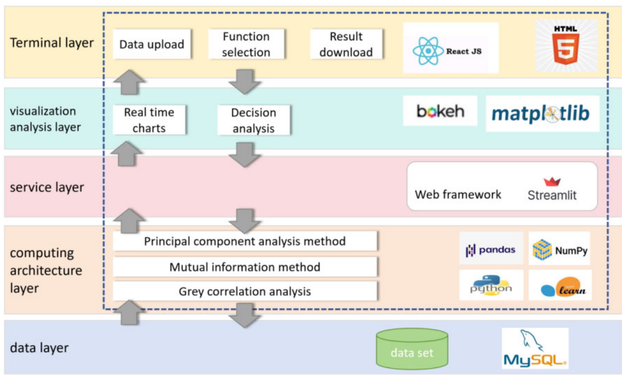

Figure 1. System architecture diagram

3.2. Functional design of the system

This data visualization analysis platform provides data and visualization support for identifying the correlation between input indicators and output indicators. Not only can it simplify the operational process of data analysis, but it also provides a scientific basis for the performance evaluation of universities, which is of great significance to decision-makers. The data visualization platform system based on combination weighting method and grey correlation analysis is divided into upload data module, function selection module, principal component analysis weight module, mutual information weight module, comprehensive weight module, and grey correlation analysis module.

(1) Upload Data Module: This module will temporarily save the preprocessed original data template to the cloud for subsequent calculation and processing, and the data file format will be in Excel format.

(2) Principal Component Analysis Module: This module implements the cumulative principal component analysis cumulative variance contribution rate graph module and the principal component analysis weight result bar graph module. In this module, the retention degree and weight importance ranking of the calculated data features can be intuitively seen. The module interface is shown in Figure 2.

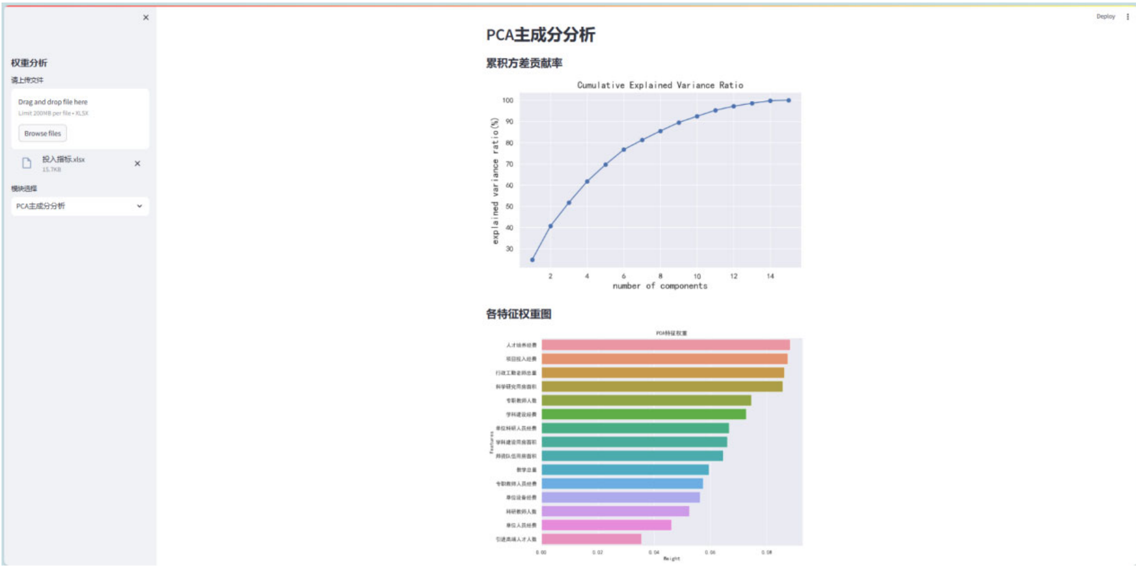

Figure 2. Principal Component Analysis Interface

(3) Mutual Information Weight Module: This module implements heatmaps and mutual information weight result maps between various indicators. The heatmaps can visually show the correlation between any two indicator variables. For ease of differentiation, correlation weights between 0.5 and 1 are represented in blue, with darker blue indicating greater correlation between the two indicators; The correlation weight between 0 and 0.5 is represented in red, with darker values indicating smaller correlation between the two indicators; The weight bar chart displays the calculated ranking of weight importance.

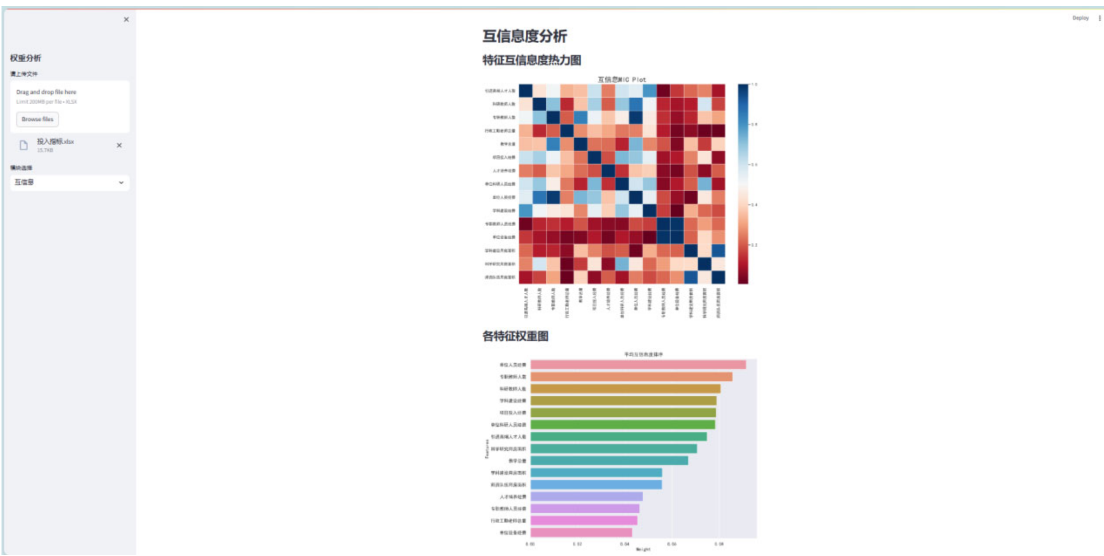

Figure 3. Mutual Information Interface

(4) Comprehensive weight module: This module visualizes the combined weights obtained by combining the above two methods. The pie chart visually displays the proportion and importance of each indicator, and the table displays the comprehensive weight values of each indicator.

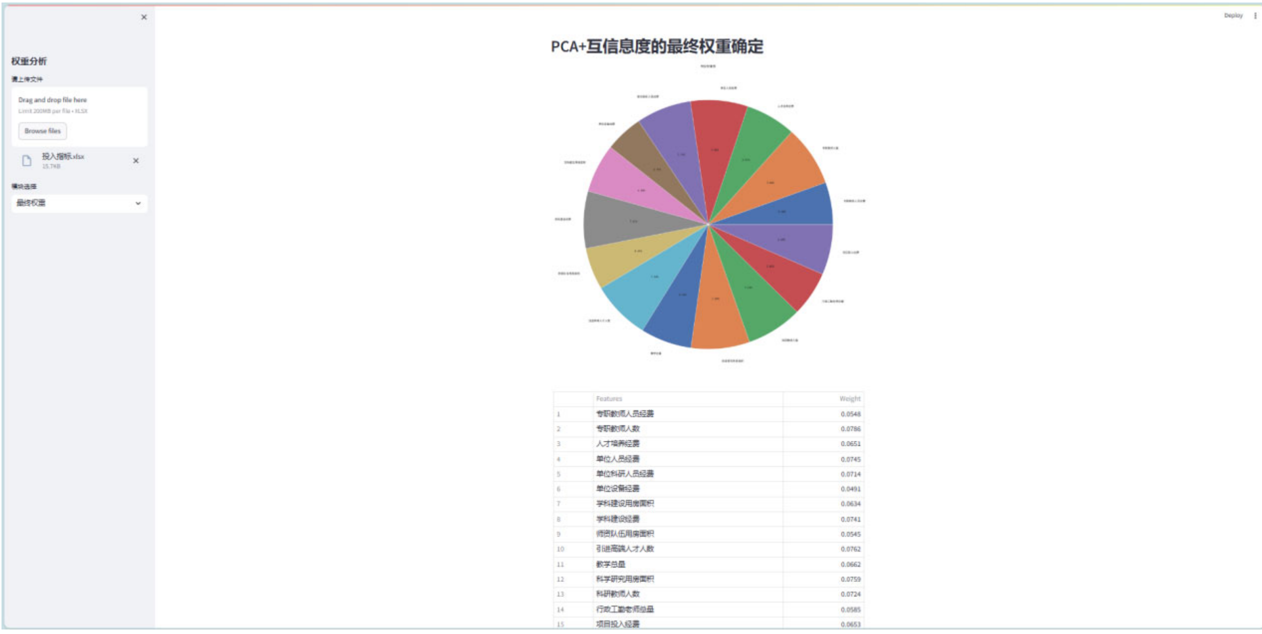

Figure 4. Comprehensive weight interface

(5) Grey correlation module: This module allows users to choose the output indicators and input indicators they want to perform correlation analysis on, and the line chart visually displays the results of correlation analysis.

Figure 5. Grey correlation interface

4. Example Application and Analysis

In order to verify the feasibility of the model and the practicality of the system, a complete and scientifically reasonable evaluation index system was first constructed, and the evaluation index data of all colleges in a certain university in 2020 were selected for experiments.

4.1. Establishing an indicator system

1.Performance investment indicators for universities

In a broad sense, input refers to the various resources and elements required for the operation of a system. For the higher education system, this kind of investment is extremely extensive, generally in three forms: personnel, finance, and materials, each of which can be divided into multiple types. In this article, investment indicators are divided into three secondary indicators: teacher team investment, funding and financial investment, and equipment and material resources investment, as well as 15 tertiary indicators.

Table 1. Input indicators

Level 1 indicator	Secondary indicators	Level 3 indicators
Input indicators	Teachers' input	Number of high-end talents introduced Total number of full-time teachers Number of scientific research teachers Number of administrative workers Total amount of teaching
	Funds and financial input	Funds for full-time teachers Administrative and logistic personnel expenses Funds for scientific research personnel Funds for discipline construction Personnel training funds Project investment funds
	Input of equipment and material resources	Input of scientific research equipment Room area for discipline construction Area of scientific research room Room area for teaching staff

2. Output indicators of universities

The performance evaluation of universities promotes the firm establishment of quality and efficiency awareness, expansion of funding channels, rational allocation of internal resources, and optimization of management and operation mechanisms in teaching and research units through statistics and measurement of input-output ratios, thereby improving the level and efficiency of teaching and research units. At the same time, it also provides scientific basis for schools to scientifically judge the development performance of various teaching and research units, and to reasonably allocate educational resources. Higher education institutions are institutions with multiple outputs. In order to facilitate analysis, we can divide the output of universities into four primary indicators: discipline construction output, talent cultivation output, scientific research output, and social impact output, and 17 secondary indicators.

Table 2. Output indicators

Level 1 indicator	Secondary indicators	Level 3 indicators
Output indicators	Discipline construction	Number of quality disciplines Number of first-class majors Number of support platforms
	Personnel training	Employment rate Number of master's degree students Number of undergraduates trained Number of student modeling contest winners Number of winners in the Student Challenge Cup
	Scientific research output	Number of teaching achievement awards Number of property rights works Number of scientific research awards Number of papers in academic journals Number of teaching and research projects
	Social impact	International or national conferences Distinguished alumni Academic status Alumni donations

4.2. Data Collection and Preprocessing

Organize and summarize the data used in the experiment, and preprocess it. During the data collection phase, data duplication, loss, and anomalies often occur. This requires pre-processing of data that is prone to problems before data analysis. For example, data duplication can be deleted, and missing data can be filled or deleted. Based on in-depth analysis of the causes of abnormal data information, timely adjustments and processing are made to comprehensively ensure the accuracy and authenticity of data information.

4.3. Importance of investment indicator weights

Upload the organized data to the system, and Figure 6 shows the weight sorting bar chart calculated by the principal component analysis method; Figure 7 shows the weight ranking bar chart calculated from the average mutual information degree; Table 3 shows the weighted average values of the above methods. Select important features based on the magnitude of their importance coefficients. The higher the importance coefficient of a feature, the higher its importance; The lower the feature importance coefficient, the lower the importance of this feature.

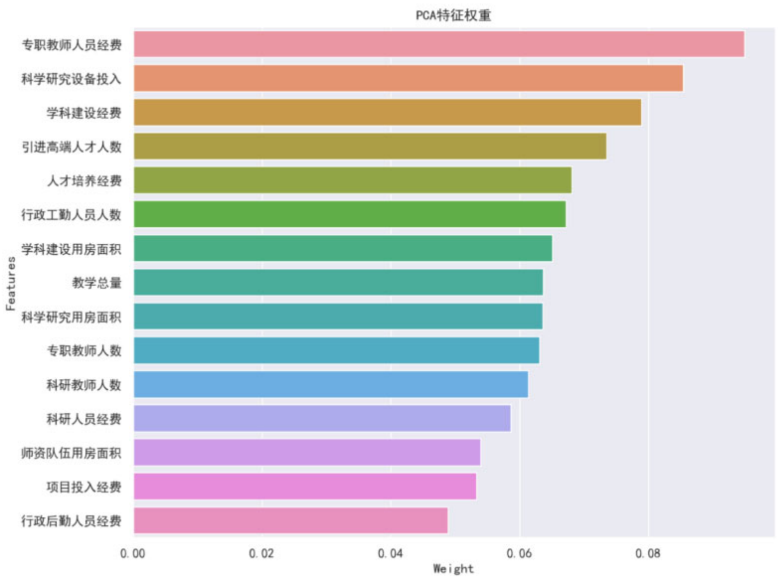

Figure 6. Input indicator PCA weight

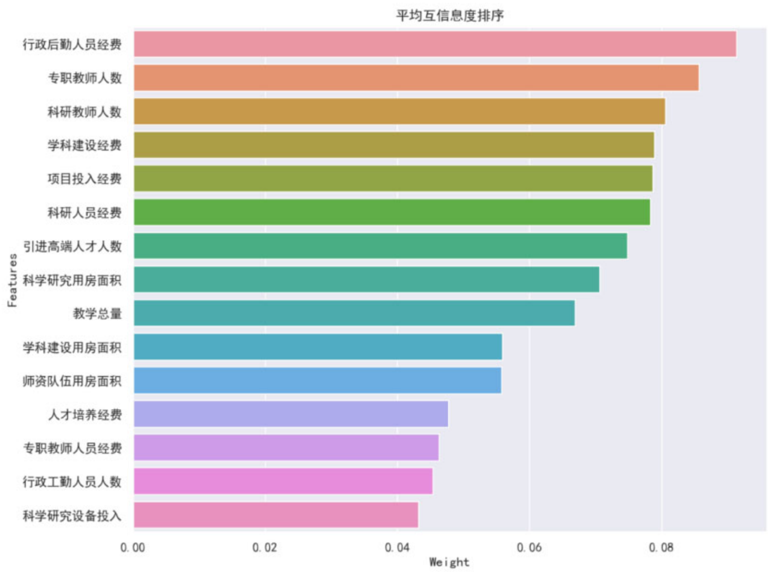

Figure 7. Input indicator MIC weight

Table 3. Comprehensive weight of investment indicators

Ranking of weight importance	Indicator name	Weight
1	Funds for discipline construction	0.0789
2	Number of full-time teachers	0.0744
3	Number of high-end talents introduced	0.0742
4	Number of scientific research teachers	0.071
5	Funds for full-time teachers	0.0706
6	Administrative and logistic personnel expenses	0.0701
7	Funds for scientific research personnel	0.0685
8	Area of scientific research room	0.0671
9	Project investment funds	0.066
10	Total amount of teaching	0.0653
11	Input of scientific research equipment	0.0643
12	Room area for discipline construction	0.0605
13	Personnel training funds	0.0579
14	Number of administrative workers	0.0563
15	Room area for teaching staff	0.0549

According to the above chart, the combination weighting method can comprehensively optimize the weights obtained by the two weighting methods, maximize the advantages of various methods, and obtain the comprehensive weights of investment indicators. Table 3 shows that the weights of discipline construction funds, number of full-time teachers, number of introduced high-end talents, number of scientific research teachers, funding for full-time teachers, and administrative and logistics personnel are all above 0.07.

4.4. Importance of output indicator weights

Similarly to the previous summary, upload the organized output indicator data to the system. Figure 8 shows the weight sorting bar chart calculated by principal component analysis method; Figure 9 shows the weight ranking bar chart calculated from the average mutual information degree; Table 4 shows the weighted average values of the above methods. Select important features based on the magnitude of their importance coefficients. The higher the importance coefficient of a feature, the higher its importance; The lower the feature importance coefficient, the lower the importance of this feature.

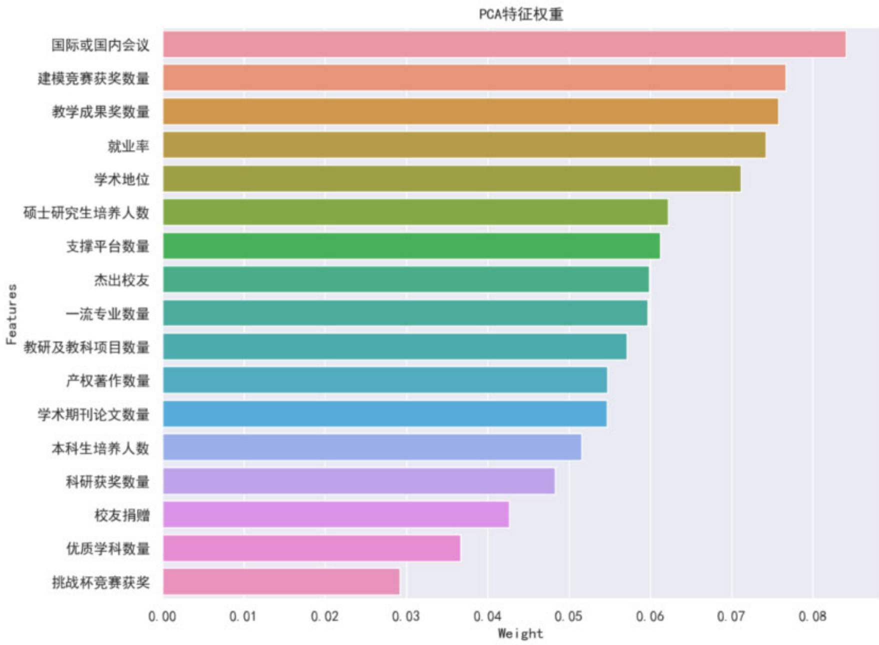

Figure 9. Output indicator PCA weight

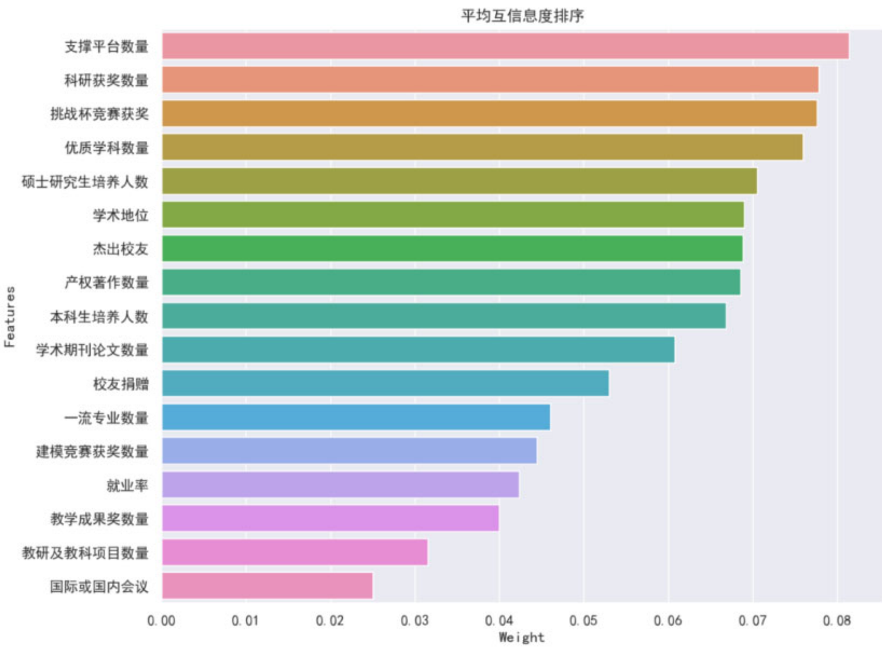

Figure 10. Output indicator MIC weight

Table 4. Comprehensive weight of output indicators

Ranking of weight importance	Indicator name	Weight
1	Number of support platforms	0.0713
2	Academic status	0.0701
3	Number of postgraduates trained	0.0664
4	Distinguished alumni	0.0643
5	Number of scientific research awards	0.0631
6	Number of property rights works	0.0617
7	Number of winners of modeling contest	0.0606
8	Number of undergraduates trained	0.0592
9	Employment rate	0.0583
10	Number of teaching achievement awards	0.0579
11	Number of papers in academic journals	0.0578
12	Number of quality disciplines	0.0563
13	International or national conferences	0.0546
14	The Challenge Cup competition won the prize	0.0534
15	Number of first-class majors	0.0529
16	Alumni donations	0.0478
17	Number of teaching and research projects	0.0444

According to the above chart, the combination weighting method can comprehensively optimize the weights obtained by the two weighting methods, maximize the advantages of various methods, and obtain the comprehensive weights of investment indicators. Table 4 shows that the weights of the number of supporting platforms, academic status, number of master's students trained, outstanding alumni, number of scientific research awards, number of property rights works, and number of modeling competition awards are all above 0.06.

4.5. Grey correlation analysis between output indicators and input indicators

Based on the feature importance coefficient, select indicators with a feature importance coefficient higher than 0.07 as input indicators; Select indicators with feature importance coefficients higher than 0.06 for correlation analysis of output indicators. The system selection interface is shown in Figure 11.

Figure 11. Indicator selection interface

After selection, the grey correlation analysis results can be obtained, as shown in Figure 12. From the analysis of the 6 selected input indicators and 7 output indicators, it can be seen that the correlation between discipline construction funds and support platforms is the highest; The number of research awards for output indicators is most closely related to the number of high-end talents introduced and the funding for full-time teachers. The line chart results can visually show the correlation between input indicators and output indicators.

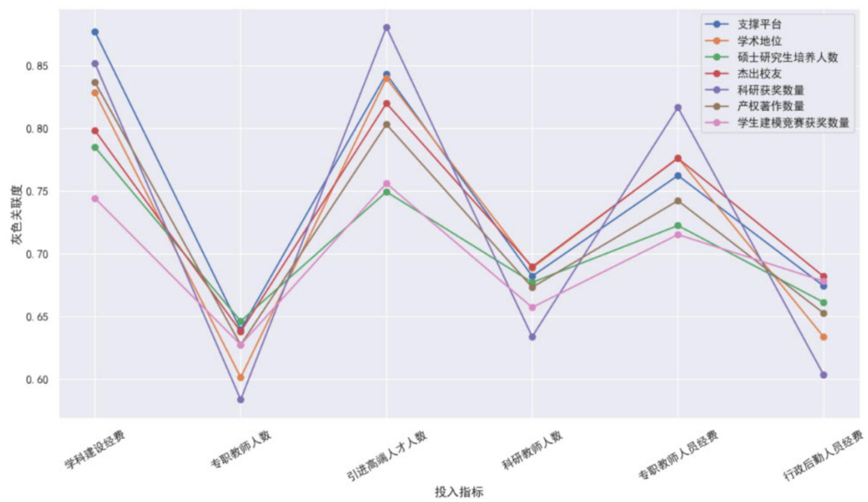

Figure 12. Indicator selection interface

5. Conclusion

This summary proposes the construction of a university performance evaluation model based on combination weighting method and grey correlation analysis. The model first uses principal component analysis and average mutual information method to select indicator features and select indicators with high importance; On this basis, the grey correlation analysis method is used to obtain the relationship between input indicators and output indicators; Finally, develop a visual performance analysis platform for universities using Python language and validate it with examples to conduct research. The research results indicate that the model and platform can effectively and intuitively display evaluation results, and can assist education management departments in identifying shortcomings and formulating targeted improvement strategies based on evaluation results.

References

[1] Liu Xinmin.Design and implementation of performance appraisal system in universities [J].Journal of Tianjin University of Technology, 2022, 38 (2): 52-57 DOI: 10.3969/j.issn.1673-095X.2022.02-010.

[2] Ying Wangjiang, Li Quanying.Index system of university's performance evaluation: design and application[J].Journal of National Academy of Education Administration, 2010 (2): 45-50 DOI: 10.3969/j.issn.1672-4038.2010.02.010.

[3] Meng Yanxia. The maximal information coefficient algorithm research[D]. Shanxi: Taiyuan University of Technology,2019.

[4] Wang Yue.Algorithm analysis and Improvement of Maximum Information Coefficient [D]. Shaanxi: Xidian University,2019. DOI:10.7666/d.D01906373.

[5] Wen Xueping.Dimensionality Reduction Based on Feature Selection [D]. Hubei: Huazhong University of Science and Technology,2015. DOI:10.7666/d.D735608.

[6] Li Changjuan. A Research of Key Technology of Dimensionality Reduction of High Dimensional Data[D]. Sichuan: University of Electronic Science and Technology of China,2017.

[7] Ling Wen, Tang Genfu, Ying Xiaowen, etc. Performance evaluation of surgical inpatient work based on principal component analysis and rank sum ratio method [J].Chinese Journal of Health Statistics, 2022, 39 (2): 275-277 DOI: 10.3969/j.issn.1002-3674.2022.02-029.

[8] Yan Huiqin, Niu Wanhong, Han Huili. Objective weight method based on principal component analysis to establish index weight [J].Journal of University of Jinan(Science and Technology), 2017,31 (06): 519-523. DOI: 10.13349/j.cnki. jdxbn.2017.06.09.

[9] Dong Shaoyang, Xia Qingguo, Li Ning. Research on software defect list： prediction method based on grey relational analysis. Computer Engineering and Applications, 2013, 49 (2): 67-71.