Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 12, No. 1, 2025

Research on Wine Evaluation based on Principal Component Analysis and Linear Regression

Mengting Liu [1, †, *], Chongrun Wang [2, †], Yumeng Yang [3, †]

[1] School of Economics and Management, Xi'an Mingde Institute of Technology, Xi'an, Shaanxi, China
[2] Higher vocational and technical colleges, Xi'an Mingde Institute of Technology, Xi'an, Shaanxi, China
[3] School of Information Engineering, Xi'an Mingde Institute of Technology, Xi'an, Shaanxi, China
[†] These authors also contributed equally to this work

*** Corresponding author:** Mengting Liu (Email: 290079416@qq.com)

Abstract: This study integrates computer technology and mathematical modeling methods to deeply explore the grading of wine-making grapes and the internal relationships between physicochemical indicators and wine quality. By using the computer software SPSS, hierarchical clustering, principal component analysis, step-by-step regression models, and linear regression models are applied to process and analyze wine evaluation data, as well as the physicochemical indicator data of grapes and wines. Through model construction, and solution, a reasonable grading of wine-making grapes is achieved. It is found that there is a linear correlation between the physicochemical indicators of grapes and wines, and both have a linear impact on wine quality. This indicates that it is feasible to evaluate wine quality using the physicochemical indicators of grapes and wines, providing a scientific basis for quality control and production decision-making in the wine industry. It also demonstrates the powerful effectiveness of computer-assisted mathematical modeling in solving complex economic management problems.

Keywords: Principal Component Analysis; Step-by-Step Regression Model; Linear Regression Model.

1. Introduction

In today's digital age, computer technology has deeply integrated into various fields, and the economic management field has also benefited greatly from it [1]. For the wine industry, accurately grasping the grading of wine-making grapes and the relationships between physicochemical indicators and wine quality is the key to enhancing economic benefits and product competitiveness [2]. This research utilizes the powerful data processing and analysis capabilities of the computer software SPSS and combines mathematical methods such as hierarchical clustering, principal component analysis, step-by-step regression models, and linear regression models for in-depth exploration. The computer software is used to efficiently pre-process a large amount of data. Hierarchical clustering is employed to initially explore the internal structure of the data, principal component analysis is used to extract key features, step-by-step regression models are applied to analyze the relationships between the physicochemical indicators of wine-making grapes and wines, and linear regression models are relied on to clarify the impact of these indicators on wine quality. By establishing and solving these mathematical models, the correlations between various factors are quantified, providing data-driven decision-making support for the economic management of the wine industry. This not only helps to optimize the allocation of industry resources but also provides a reference example for the application of computer-assisted mathematical modeling in other similar fields.

2. Grading of Wine-Making Grapes Based on Hierarchical Clustering and Principal Component Analysis

Considering the differences in the absolute values and dispersion degrees among the indicator values of various physicochemical indicators, in order to avoid the influence of these factors on the analysis results, the physicochemical indicator values of wine-making grapes and wines are respectively mean-standardized.

Hierarchical clustering is performed using SPSS [3], and the evaluation index results are shown in Table. 1.

Table 1. Evaluation indexes

Silhouette coefficient	DBI	CH
0.624	0.39	48.665

For a sample set, its silhouette coefficient is the average of the silhouette coefficients of all samples. The value range of the silhouette coefficient is [-1, 1]. The closer the distances of samples in the same category and the farther the distances of samples in different categories, the higher the score and the better the clustering effect. DBI (Davies - Bouldin Index) is used to measure the ratio of the sum of the within-cluster distances of any two clusters to the between-cluster distance [4]. The smaller this index is, the better the clustering effect. CH (Calinski-Harabasz Score) measures the compactness within a class (the denominator) by calculating the sum of the squared distances between each point in the class and the class center, and measures the separation degree of the data set (the numerator) by calculating the sum of the squared distances between the center points of different classes and the center point of the data set [5]. The CH index is obtained by the ratio of the separation degree to the compactness. The larger the CH value, the better the clustering effect.

It can be seen from Table. 1 that the silhouette coefficient is 0.624, the DBI is 0.39, and the CH is 48.665, indicating a relatively good clustering effect.

For each indicator in red wine, the average value of multiple experiments is taken as the reference value. Then, the data is subjected to principal component analysis using SPSS [3], and the comprehensive scores of each sample are obtained. Some of the data are shown in Table.2.

Table 2. Comprehensive score table of red wine

Ranking	Row index	Comprehensive score	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7
1	Grape sample 9	0.754	0.637	1.714	1.814	-0.156	0.442	-0.84	-1.16
2	Grape sample 1	0.709	0.398	2.052	-0.689	-0.632	0.339	2.994	0.705
3	Grape sample 8	0.646	0.217	1.409	0.198	0.456	-0.319	1.735	2.303
4	Grape sample 23	0.516	0.566	0.301	2.07	-0.998	1.029	-0.557	0.659
5	Grape sample 2	0.5	0.594	1.708	0.175	-0.371	0.183	-0.331	-2.018
6	Grape sample 3	0.472	0.664	1.635	-0.572	0.093	-0.042	-2.156	0.562
7	Grape sample 21	0.292	0.171	0.096	0.477	2.393	0.142	-1.287	0.05

The comprehensive scores in the above table and the average scores of the second group's red wine tasting are averaged, and then the grading process is carried out through the score differences. The specific grading ranking is shown in Figure 1.

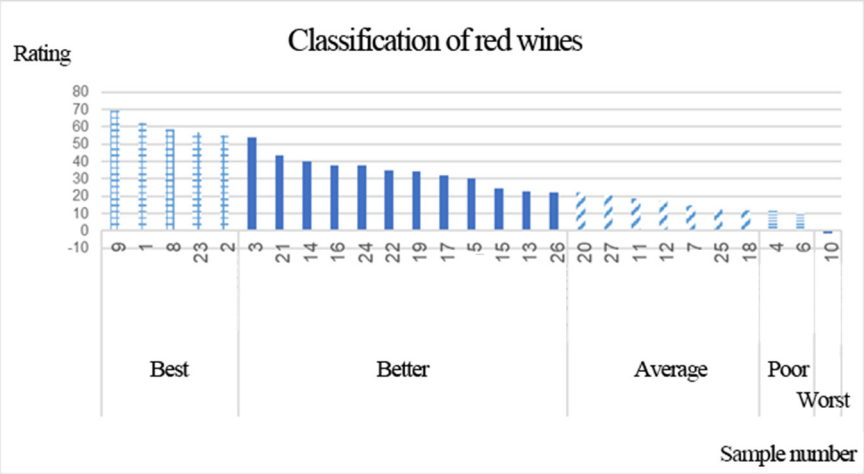

Figure 1. Grading ranking of red wine

From Figure 1, the grading ranking of red wine can be seen. There are 8 grape samples with the highest grade, which are Grape Samples 9, 1, 8, 23, and 2. One grape sample has the lowest grade, which is Grape sample 10.

For each indicator in white wine, the average value of multiple experiments is taken as the reference value. Then, the data is subjected to principal component analysis using SPSS, and the comprehensive scores of each sample are obtained. Some of the data are shown in Table. 3.

Table 3. Comprehensive score table of white wine

Ranking	Row index	Comprehensive score	PCA1	PCA 2	PCA 3	PCA 4	PCA 5	PCA 6	PCA 7	PCA 8	PCA 9	PCA 10
1	Grape sample 27	1.18	1.298	1.589	2.461	-2.173	2.173	1.037	0.815	0.869	0.912	0.463
2	Grape sample 5	0.441	1.099	-0.042	-0.612	1.457	0.502	0.481	1.324	-0.636	0.772	0.127
3	Grape sample 24	0.44	1.158	1.352	0.857	-0.357	-1.933	0.94	-0.849	-0.395	0.097	-0.979
4	Grape sample 28	0.396	1.386	-0.303	-0.424	1.219	0.413	0.43	-0.007	0.665	-0.937	1.176
5	Grape sample 3	0.362	0.44	-0.519	0.851	2.428	0.963	0.73	-0.68	-1.556	0.269	0.582
6	Grape sample 7	0.32	-0.196	1.654	-0.205	0.683	0.296	-0.494	0.205	1.733	-2.189	0.768
7	Grape sample 15	0.274	-0.923	1.403	0.866	1.51	-0.216	-1.886	-0.963	0.414	1.576	1.556

The comprehensive scores in the above table and the average scores of the second group's white wine tasting are averaged, and then the grading process is carried out. The specific grading ranking is shown in Figure 2.

From Figure 2, the grading ranking of white wine can be seen. There are 9 grape samples with the highest grade, which are Grape Samples 27, 5, 24, 28, 3, 7, 15, 20, and 6. One grape sample has the lowest grade, which is Grape Sample 8.

3. Analysis of the Relationship Between Physicochemical Indicators of Wine-Making Grapes and Wines Based on the Stepwise Regression Model

3.1. Model Establishment

A stepwise regression model is established [6]:

$$Y_1 = \beta_0 + \beta_1 \times 1 + \beta_2 \times 2 + \cdots + \beta_k \times k \quad (1)$$

$$Y_2 = \beta_0 + \beta_1 \times 1 + \beta_2 \times 2 + \cdots + \beta_m \times m \quad (2)$$

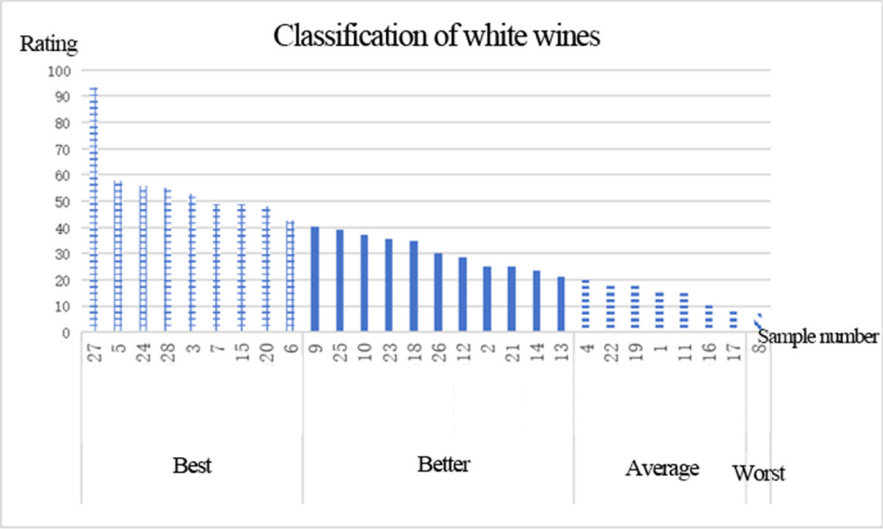

Figure 2. Grading ranking of white wine

Where k represents anthocyanins (mg / L), m represents color, and β_i represents the coefficient of the i-th indicator.

3.2. Model Solution

Determine the variables selected and retained through stepwise regression according to the results table of the stepwise regression model. Analyze the model fitting situation through R^2. At the same time, analyze the VIF value to check for collinearity (VIF > 10 or 5, strictly 10). Analyze the significance of X. If it is significant (P<0.05), it is used to explore the influence relationship of X on Y. Combine with the regression coefficient B value to comparatively analyze the influence degree of X on Y.

Table 4. Results table of the stepwise regression model for red wine

Linear regression analysis results (n=27)									
	Unstandardized coefficients		Standardized coefficient	t	P	VIF	R^2	Adjusted R^2	F
	B	Std. error	Beta						
Constant	19.344	2.485	0	7.783	0.000***	-	0.211	0.18	F=6.698, P=0.016**
Anthocyanins (mg/L)_missing value treatment	-0.02	0.008	-0.46	-2.588	0.016**	1			
Dependent variable: sample number									
Note: ***, **, *represent significance levels of 1%, 5%, and 10% respectively.									

From the analysis of the F-test results in Table. 4, it can be obtained that the significance P-value is 0.016**, indicating a significant level. Thus, the null hypothesis that the regression coefficient is 0 is rejected. Regarding the collinearity of variables, all VIF values are less than 10. Therefore, the model has no multicollinearity problem and is well - constructed.

The formula of the solved model is as follows:

$$Y_1 = 19.344 - 0.02 \times k \quad (3)$$

Where k represents anthocyanins (mg / L)

The equation fitting graph is shown in Figure 3.

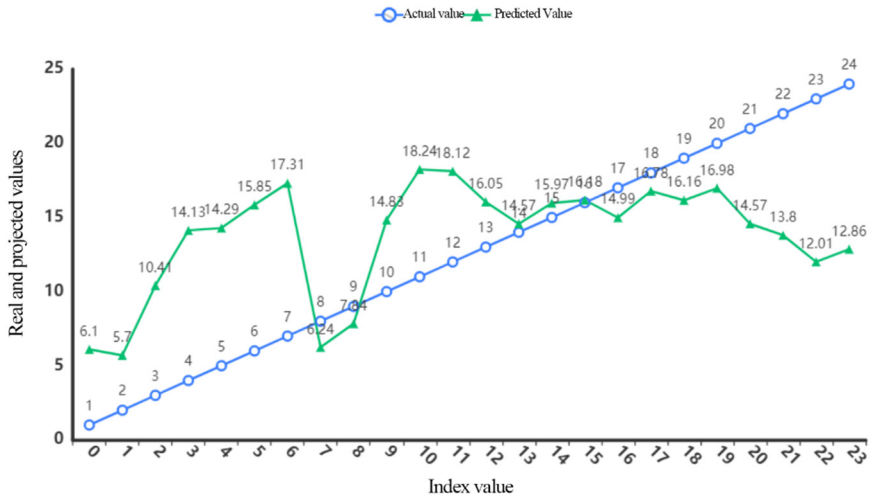

Figure 3. Fitting graph of the red wine equation

Similarly, the results of the stepwise regression model for white wine obtained by using SPSS are shown in Table. 5.

Table 5. Results table of the stepwise regression model for white wine

Linear regression analysis results (n=28)									
	Unstandardized coefficients		Standardized coefficient	t	P	VIF	R^2	Adjusted R^2	F
	B	Std. error	Beta						
Constant	-785.321	228.531	0	-3.436	0.002***	-	0.32	0.294	F=12.249, P=0.002***
Color_missing value treatment	22.955	6.559	0.566	3.5	0.002***	1			
Dependent variable: sample number									
Note: ***, **, *represent significance levels of 1%, 5%, and 10% respectively.									

From the analysis of the F-test results in Table. 5, it can be found that the significance P-value is 0.002***, indicating a significant level. Thus, the null hypothesis that the regression coefficient is 0 is rejected. Regarding the collinearity of variables, all VIF values are less than 10. Therefore, the model has no multicollinearity problem and is well-constructed.

The formula of the solved model is as follows:

$$Y_2 = -785.321 + 22.955 \times m \tag{4}$$

Where m represents color.

The equation fitting graph is shown in Figure 4.

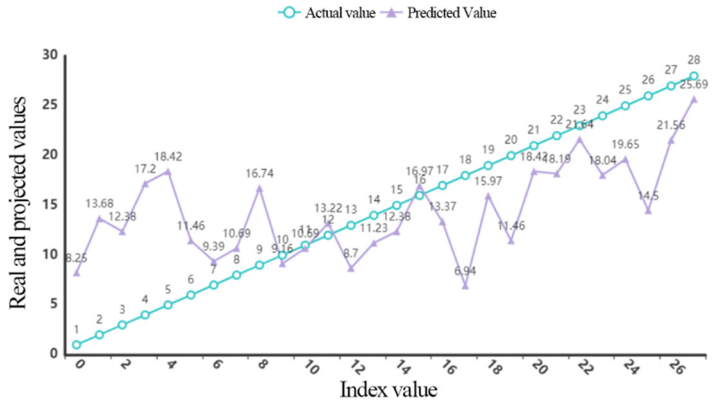

Figure 4. Fitting graph of the white wine equation

In conclusion, the physicochemical indicators of wine-making grapes and wines are linearly correlated.

4. Analysis of the Impact of Physicochemical Indicators on Wine Quality based on the Linear Regression Model

4.1. Model Establishment

A linear regression model is established [7]:

$$Y_1 = \omega_1\beta_1 + \omega_2\beta_2 + \cdots + \omega_n\beta_n \tag{5}$$

Where β_i represents the i-th physicochemical indicator of wine-making grapes, and ω_i represents the coefficient of the i-th indicator.

$$Y_2 = \omega_1\beta_1 + \omega_2\beta_2 + \cdots + \omega_n\beta_n + b \tag{6}$$

Where β_i represents the i-th physicochemical indicator of wine, and ω_i represents the coefficient of the i-th indicator.

4.2. Model Solution

The results of the linear regression model for white wine [4] solved by SPSS are shown in Table. 6.

Table 6. Prediction results of the wine - making grape model

Variables	Coefficients	Test Values
Constant	-37.083271780025946	1
Anthocyanins (mg/L)_missing value treatment	0.059062383687369546	
Tannins (mmol/L)_missing value treatment	-4.681082628562012	
Total Phenols (mmol/L)_missing value treatment	1.3636646368711853	
Total flavonoids in wine (mmol/l)_missing value treatment	0.6394689305455599	
Resveratrol (mg/L)_missing value treatment	-1.106575906932564	
Reducing sugar (g/L)_missing value treatment	-0.09970177473285376	
Solid - acid ratio_missing value treatment	0.040026366022056184	
Dry matter content (g/100g)_missing value treatment	0.05989427670837938	
Cluster mass (g)_missing value treatment	1.3876964381650831	
100 - grain mass (g)_missing value treatment	0.20400180367706633	
Pedicel ratio (%)_missing value treatment	-51.37386420194571	
Soluble solids (g/l)_missing value treatment	-0.22761832176145402	
pH value_missing value treatment	-0.4057860603092651	
Titratable Acid (G/L)_Missing Value Treatment	6.054747989733821	
Prediction results	37.083271780025946	

The formula of the solved model is as follows:

$$Y_1 = -37.083 + 0.059\beta_1 - 4.681\beta_2 + \cdots + 6.055\beta_{14} \quad (7)$$

Where β_i represents the i-th physicochemical indicator of wine-making grapes.

The equation fitting curve graph is shown in Figure 5.

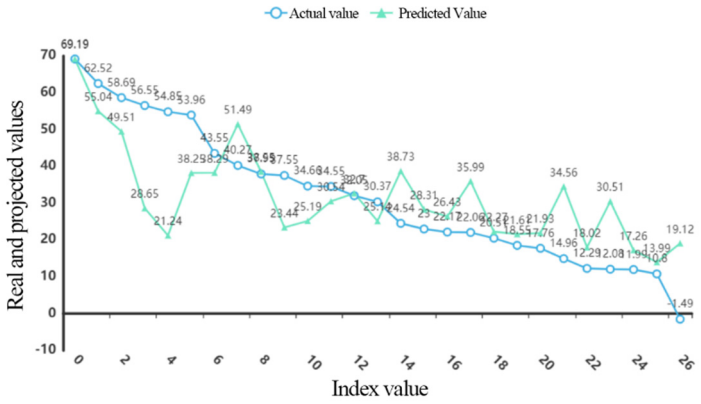

Figure 5. Fitting curve graph of the wine - making grape equation

Similarly, the results of the linear regression model for wine are shown in Table. 7.

Table 7. Prediction Results of the Wine Model

Variables	Coefficients	Test Values
Constant	567.3488190355799	1
Total amino acids_missing value treatment	-0.007451052193115104	
Protein (mg/100g)_missing value treatment	0.5602250929878388	
VC content (mg/L)_missing value treatment	-0.0606280108682864	
Anthocyanins (mg/100g fresh weight)_missing value treatment	-0.6690405983751857	
Tartaric acid (g/l)_missing value treatment	1.1204501856210767	
Malic acid (g/l)_missing value treatment	-0.12125602185125214	
Citric acid (g/l)_missing value treatment	-17.738671054456642	
Polyphenol oxidase activity_missing value treatment	236.61302966626437	
Browning degree_missing value treatment	-0.7959071541178319	
DPPH radical 1/ic50 (g/l)_missing value treatment	0.26757469321249244	
Total phenols (mmol/kg)_missing value treatment	-12.028787811452652	
Total grape flavonoids (mmol/kg)_missing value treatment	-0.05188408914333145	
Resveratrol (mg/kg)_missing value treatment	-3.8780916613675975	
Flavonol (mg/kg)_missing value treatment	-0.1556522676231852	
Total sugar (g/L)_missing value treatment	-7.949923638729981	
Reducing sugar (g/l)_missing value treatment	-0.5325567765308781	
Soluble solids (g/l)_missing value treatment	-18.115560957539337	
pH value_missing value treatment	0.04658564650004515	
Titratable acid (g/l)_missing value treatment	15.458154631685211	
Solid - acid ratio_missing value treatment	0.038465299319630365	
Dry matter content (g/100g)_missing value treatment	-0.20494598449657525	
Cluster mass (g)_missing value treatment	14.61675526393696	
100 - grain mass (g)_missing value treatment	-0.5668516090466018	
Pedicel ratio (%)_missing value treatment	179.60646770357494	
Juice yield (%)_missing value treatment	-29.398455246731892	
Prediction results	567.3488190355799	

The formula of the solved model is as follows:

$$Y_1 = 567.349 - 0.007\beta_1 + 0.56\beta_2 + \cdots - 29.398\beta_{25} \quad (8)$$

Where β_i represents the i-th physicochemical indicator of wine.

The equation fitting curve graph is shown in Figure 6.

Based on the equations, we can infer that the physicochemical indicators of wine-making grapes have a linear relationship with wine quality, and the physicochemical indicators of wines also have a linear relationship with wine quality.

In conclusion, the physicochemical indicators of grapes and wines can be used to evaluate the quality of wines.

5. Conclusion

This study, with the assistance of computer-aided mathematical modeling, has successfully achieved the grading of wine-making grapes and an in-depth analysis of the relationships between physicochemical indicators and wine quality. Hierarchical clustering and principal component analysis, supported by computer software, effectively integrate multi-dimensional data and realize the scientific grading of wine-making grapes. The step-by-step regression

model and the linear regression model clearly reveal the linear correlation between the physicochemical indicators of grapes and wines, as well as their linear influence mechanism on wine quality. This achievement verifies the effectiveness of using physicochemical indicators to evaluate wine quality, providing a strong theoretical basis for quality control, production planning, and resource allocation in the wine industry. From the perspective of computer science, this study fully demonstrates the advantages of computer-aided mathematical modeling in processing complex data and solving practical economic management problems. Future research can further expand the application of computer algorithms. For example, machine-learning algorithms can be adopted to explore more potential influencing factors, improve the accuracy of model prediction, and provide more precise and efficient decision-making support for the wine industry and even broader economic management fields.

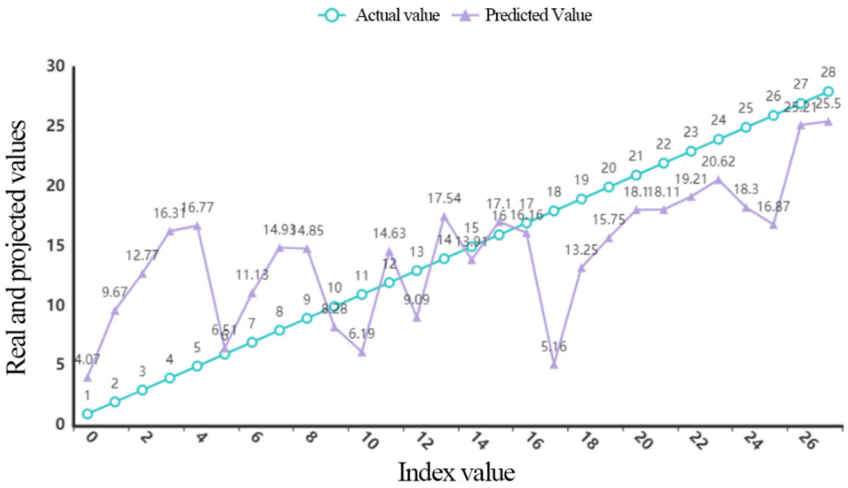

Figure 6. Fitting curve graph of the wine equation

References

[1] Legner C, Eymann T, Hess T, et al. Digitalization: opportunity and challenge for the business and information systems engineering community[J]. Business & information systems engineering, 2017, 59: 301-308.

[2] Kulasiri D, Somin S, Kumara Pathirannahalage S. A Machine Learning Pipeline for Predicting Pinot Noir Wine Quality from Viticulture Data: Development and Implementation[J]. Foods, 2024, 13(19): 3091.

[3] Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview, II[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2017, 7(6): e1219.

[4] Ros F, Riad R, Guillaume S. PDBI: A partitioning Davies-Bouldin index for clustering evaluation[J]. Neurocomputing, 2023, 528: 178-199.

[5] El Khattabi M Z, El Jai M, Lahmadi Y, et al. Understanding the interplay between metrics, normalization forms, and data distribution in K-means clustering: A comparative simulation study[J]. Arabian Journal for Science and Engineering, 2024, 49(3): 2987-3007.

[6] Żogała-Siudem B, Jaroszewicz S. Fast stepwise regression based on multidimensional indexes[J]. Information Sciences, 2021, 549: 288-309.

[7] Ferraro M B, Coppi R, Rodríguez G G, et al. A linear regression model for imprecise response[J]. International Journal of Approximate Reasoning, 2010, 51(7): 759-770.