Machine Learning based Wine Quality Evaluation

Xiaote Liu *

College of Mathematics and Physics, Beijing University of Chemical Technology, Beijing, 100029, China

* Corresponding author: 2020120062@buct.edu.cn

Abstract. In the wine industry, red wine is a kind of fruit wine that made from grapes, which is quite common in our life. With the gradual expansion of people who like red wine, the wine industry is getting more and more attention at the same time, the quality of red wine is also getting more and more attention. In order to better evaluate the quality of red wine, an evaluation model of red wine quality is established based on the collected 200 red wine samples and the corresponding 11 index data. Firstly, the dimensionality of the data index is decreased by using principal component analysis and factor analysis. Then, k-means clustering, sum of squares of deviation and class average algorithm are used to perform cluster analysis on the data processed by principal component analysis or factor analysis. Then, logistic regression analysis is used to test the accuracy of the data classification. Finally, Fisher discriminant method is used to perform discriminant analysis on the data and establish a model. The score function is obtained, and the score function is used to calculate the quality score of each group of wine, and the corresponding suggestions were given according to the quality evaluation results.

Keywords: principal component analysis, cluster analysis, logistic regression.

1. Introduction

With the continuous improvement of people's living standards, people's demand for food has gradually changed from being able to eat enough to eat well and eat healthily. The wine industry has a significant influence on the world economy, and it has also been developing rapidly in recent years, among which the red wine is gradually transforming into a high-end industry. With the gradual expansion of people who like red wine, the wine industry is getting more and more attention at the same time, the quality of red wine is also getting more and more attention. The method of manually judging the quality of wine is inefficient and expensive, which makes it difficult for producers to handle large quantities of wine samples. There is no specific evaluation standard for the quality evaluation system of red wine in the past, and it is difficult to find an evaluation system for the quality of unknown red wine samples. With the rise of machine learning, supervised machine learning can be well used in the wine evaluation system. By learning the quality evaluation of a large number of wine samples based on experience, a relatively perfect wine quality evaluation model can be established, and then the quality evaluation of subsequent wine samples can be directly conducted by the model.

Firstly, the sample data are used to analyze and compare the sample components of 200 red wine samples. Since 11 indicators were too many, it is not convenient to analyze. Therefore, dimensionality reduction of data indicators was carried out through principal component analysis and factor analysis of principal component method to compare the dimensionality reduction effects of the two methods.

Then, cluster classification is carried out according to the factor score data after dimension reduction, and quality evaluation is carried out based on the specific meaning of factors. Finally, based on the quality analysis of the red wine samples, targeted quality improvement suggestions can be provided for each group of red wine samples with different quality. The main contributions of this work are presented as follows:

- The factors composed of red wine composition indexes are given, through which the quality of red wine can be reflected.

• Through a series of analysis and processing of sample data, a set of wine quality evaluation system has been preliminarily established, which can evaluate the quality of a large number of wine samples and group them according to their quality.

• Through the model, specific suggestions for improving the quality of each group of red wine samples can be given to form a relatively complete wine quality evaluation model.

2. Method

2.1. Data preprocessing

This article's data are taken from: He whale community. There are a total of 4000 sample data in the original data. In order to facilitate analysis and mapping, the original data is reduced to 200 for further analysis. Next, Table 1 introduces the specific meanings of the eleven indicators of the data sample, as shown below:

Table 1. Data index interpretion

Index name	Indicator meaning
Alcohol	Generally, the higher the alcohol content of red wine, the better the quality.
Sulfites	Sulfites is very beneficial to red wine. It prevents oxidation and keeps the wine fresh.
Acidity	The higher the acidity in red wine, the better. Acidity can make red wine taste more refreshing.
Density	The quality and density of red wine are strongly inversely correlated.
Total sulfur dioxide/free sulfur dioxide	The role of sulfur dioxide in red wine is to kill harmful bacteria, prevent microbial pollution of red wine.
Chloride: An indicator of saltiness	Too much saltiness is considered undesirable, and the right proportion can make a red wine more palatable.
Sweetness	Any natural grape sugar left over after fermentation, intentionally or unintentionally.
Citric acid	It usually acts as an antiseptic. Adding citric acid to the finished wine increases acidity and compensates for specific flavors.
Volatile acids	Volatile acids can be a sign that the wine has gone rotten or gone wrong in the brewing process, producing unpleasant taste and smell.
Nonvolatile acidity	It is generally higher in red wine.

2.2. Factor analysis

For the existing 200 wine sample data and 11 indicators, firstly, principal component analysis was used to reduce the dimension of the data. The fundamental goal of principal component analysis is to attempt to replace the original numerous, correlated indicators with a new set of comprehensive, unrelated indicators.

Next, the main mathematical ideas of principal component analysis are introduced. Here, X_j ,$1 \ll j \ll p$ is an n×1 column vector, representing the value of the ith feature at each data point. For convenience, we assume that the mean value of each feature measured at n data points is 0 [1].

So, our first pivot vector $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^T$ is a linear mapping. $Z_1 = \phi_{11}X_1 + \phi_{21}X_1 + ... + \phi_{P1}X_P$ the vector with the largest variance. Write it as a matrix: $Z_1 = \phi_{11}X_1 + \phi_{21}X_1 + \ldots + \phi_{P1}X_P = X_{x\phi_1}$ So the sample variance of Z one is equal to $\frac{1}{n}\sum_{i=1}^{n}(Z_{i1} - E(Z_1))^2$. $E(Z_1)$ is the sample average and to maximize the variance of Z1, we need to find $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^T$ that could maximize $\frac{1}{n}\sum_{i=1}^{n}(Z_{i1} - E(Z_1))^2$. So, solving the first principal component becomes a process of maximizing the value of some function [2, 3].

$$\begin{aligned} E(Z_1) &= \frac{1}{n}\sum_{i=1}^{n} Z_{i1} \\ &= \frac{1}{n}\sum_{i=1}^{n}(\phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{P1}x_{ip}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\phi_{j1}\,x_{ij} \\ &= \frac{1}{n}\sum_{j=1}^{p}\phi_{j1}\sum_{i=1}^{n}x_{ij} \end{aligned} \tag{1}$$

However, according to our convention above, the mean value of each feature in n datapoints is 0. So $\sum_{i=1}^{n} x_{ij}\,, \forall j = 1,2,\ldots,9$. We have $E(Z_1) = 0$, and the maximization problem becomes:

$$\underset{\phi_{11},\phi_{21},\ldots,\phi_{p1}}{\text{maximize}}\ \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{j1}x_{ij})^2 \tag{2}$$

Write the above in matrix form, because $\frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{j1}x_{ij})^2 = \frac{1}{n}||X_{\phi_1}||^2$, X is n×p matrix ϕ_1 is p × 1 Column vector, $||X_{\phi_1}||^2$ can be write as $\phi_1^T X_T X_{\phi_1}$ [2,3]. Then the optimization problem becomes:

$$\underset{\phi_{11},\phi_{21},\ldots,\phi_{p1}}{\text{maximize}}\ \frac{1}{n}\phi_1^T X_T X_{\phi_1} = \underset{\phi_{11},\phi_{21},\ldots,\phi_{p1}}{\text{maximize}}\ \phi_1^T(\frac{1}{n}X^T X)\phi_1 \tag{3}$$

At the same time, the dimensionality of data index can also be reduced through the factor analysis of principal component method. The correlation coefficient matrix R and the number of factors m need to be provided through the factor analysis. In order to show the correlation between each index more directly and intuitively, the correlation heat map of each index is made:

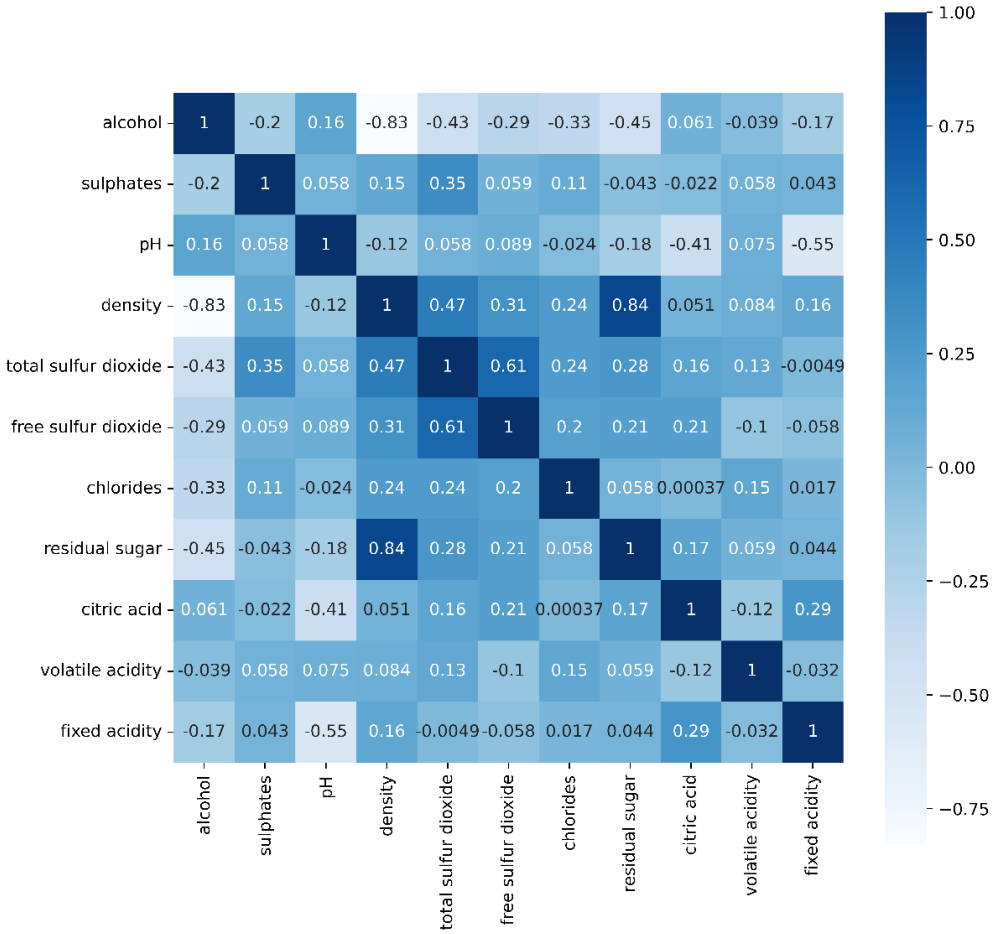


Figure 1. Heat map of correlation coefficient between indexes.

The results of factor analysis are as follows:

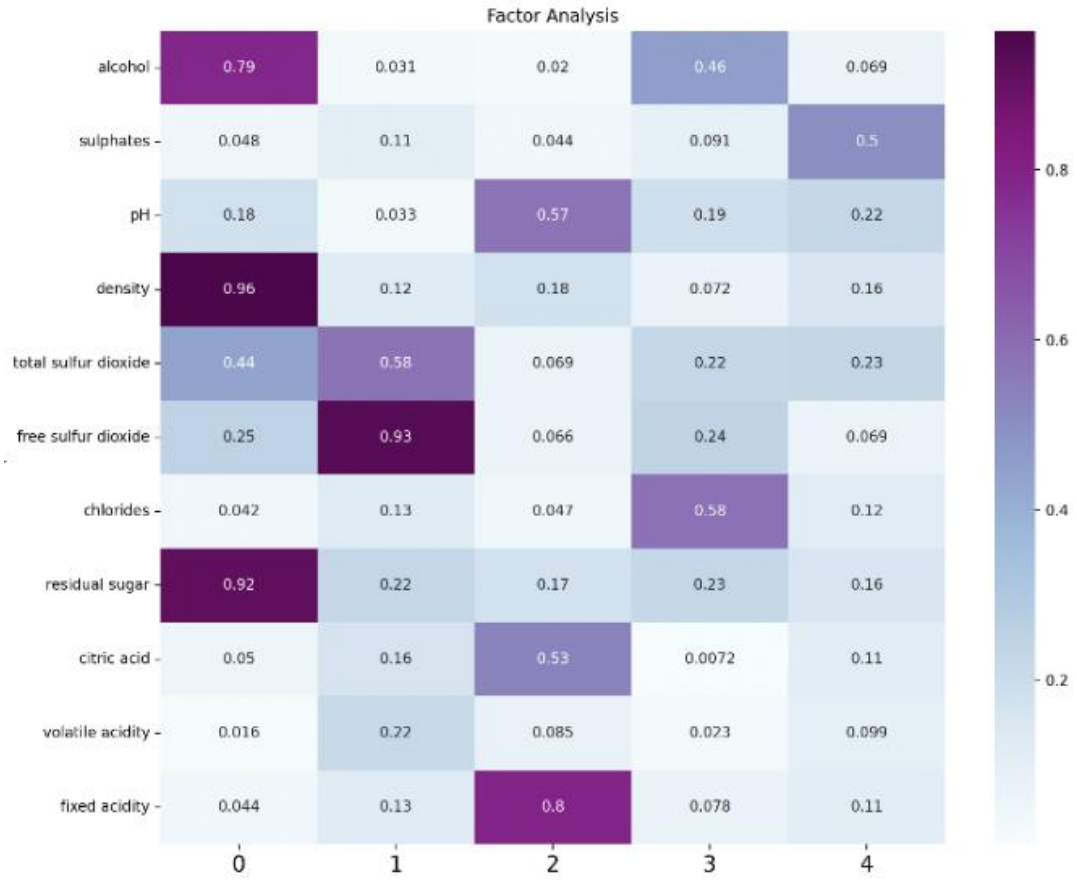


Figure 2. Figure of factor analysis result of principal component method.

According to the correlation coefficient in the figure, the indicators may be divided into five groups: {4,8}, {5, 6}, {3, 9, 11}, {1,7}, {2,10}. Based on the previous indicator interpretation, the analysis shows that:

Factor R1 is more closely related to indicators 1 and 3, which can be described as the taste factor of red wine; factor R2 is more closely related to indicators 4 and 8, which can be described as the sweetness factor of red wine; factor R3 is closely related to indicators 2, 9 and 11, which can be described as the antioxidant factor of red wine; factor R4 is closely correlated with index 5, 6 and 7, it can be described as the bactericidal factor of red wine; factor R5 is negatively correlated with index 10, it can be described as the odor factor. Therefore, five factors are selected to generate in factor analysis. Specifically, the principal component method was used for factor analysis and the final results were orthodontically rotated to facilitate interpretation of the meaning of the factors. Table 2 shows the specific meanings of each factor in the result of factor analysis:

Table 2. Factor meaning

	R1	R2	R3	R4	R5
Correlation	Indicators 4 and 8 are more closely related	Indicators 5, 6 are more closely correlated	ndicators 3, 9 and 11 are closely correlated	Indicator 1, 7 is closely correlated	Index 2, 10 has obvious negative correlation
Factor feature	Sweetness factor	Bactericidal factor	Antioxidant factor	Taste factor	Odor factor

By comparing the results of principal component analysis and factor analysis, the five major components of principal component analysis and factor analysis are shown to provide roughly 80% of the total variation. Because factor analysis can carry out orthogonalization rotation, each factor can be better explained, and the results of factor analysis are selected. Factor score is further obtained,

that is, the value of each factor on different individuals, so that we can check the performance of each factor on different individuals, and take the factor score as the new sample data for subsequent analysis.

2.3. Cluster analysis

Cluster analysis is a kind of statistical analysis techniques, the research object is divided into relatively homogeneous group through it [4].

2.3.1. Level cluster

The fundamental tenet of the hierarchical clustering approach is to calculate the similarity of nodes using a certain similarity metric, and then to connect the nodes progressively from high to low similarity. This strategy has the benefit of allowing the divide to be stopped at any time. his method has the advantage of stopping the division at whatever. The points that follow are the primary steps:

1. Get the network's starting state with n isolated nodes by eliminating all of the network's edges.
2. Determine how similar each network pair of nodes is.
3. To create a tree graph, connect the corresponding node pairs from strong to weak similarity.
4. The community structure is obtained by crosscutting the tree diagram according to the actual demand.

Specific steps:

1: Think of n samples as n different populations.

2: The matrix of distances between n populations should be calculated.

3: The two distinct categories with the least distances should be combined, and the matrix of distances between the two newly formed groups should be calculated [5].

Here, the choice of distance is particularly critical. There are various types of distance selection that can be considered. In this paper, the class average algorithm and the sum of deviation squares algorithm are selected to calculate distance respectively:

Class average algorithm [6]:

$$D(c_1, c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum_{x_1 \in c_1}\sum_{x_2 \in c_2} D(x_1, x_2) \tag{4}$$

Deviation sum of squares algorithm [6]:

$$D(c_1, c_2) = D((\frac{1}{|c_1|}\sum_{x_1 \in c_1} \vec{x}, \frac{1}{|c_2|}\sum_{x_2 \in c_2} \vec{x})) \tag{5}$$

4: If the population number is 1, go to step 5, otherwise go to step 3

5: Draw the system tree

6: Select the number of groups

Next, cluster analysis is carried out. Class average algorithm, sum of squares of deviation algorithm and k-means algorithm, are used to cluster the factor scores obtained from factor analysis. The number of clusters should be determined first before clustering so Rstudio is used to judge the number of clusters.

2.3.2. K-means cluster

The K-means clustering algorithm is as follows [7,8]:

1. Choosing k centers;
2. Every point of data should be assigned to the closest central spot;
3. Reconsider the average separation between the center of every category and the center of every class point;
4. Give each piece of information its closest central location;
5. Continue doing Steps 3 and 4 until either all observations have been allocated or the allocated number of iterations has been used.

2.4. Logistic regression

Logistic regression is a kind of generalized linear regression. When faced with a regression or classification issue, the cost function is determined, the ideal model parameters are then iteratively solved by the optimization approach, and finally the effectiveness of our solved model is confirmed by testing [9].

1. A good prediction function to use is the h function. We need this function, the sorting function, to forecast how the incoming data will turn out. This is a crucial procedure that calls for a thorough comprehension and analysis of the data, as well as knowledge of or an educated guess as to the overall form of the function, such as whether it is linear or nonlinear. [10].

$$y = \frac{1}{1+e^{-(x\omega^T+b)}} \tag{6}$$

2. Create a Cost function to indicate the difference between the training data category (y) and the forecasting function (h). The distinction (H-y) or the square of the difference (h-y) ^2 between the two can be calculated. Or in another way. Sum or mean cost, written as $J(\omega)$ function, represents the difference between the expected amount of every piece of training data and the actual category when taking the "loss" of all training data into account.

AThe logistic regression model's mathematical structure established; the only issue left is how to deal with the parameters that make up the model. Finding a collection of parameters where the likelihood of our data is greatest is a challenge that is frequently solved in statistics using maximum likelihood estimation [10].

Suppose:

$$P(Y = 1|x) = p(x) \tag{7}$$

$$P(Y = 0|x) = 1 - p(x) \tag{8}$$

Likelihood function:

$$L(\omega) = \prod[p(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i} \tag{9}$$

Take the logarithm of both sides and get:

$$L(\omega) = \sum[y_i lnp(x_i) + (1 - y_i)ln(1 - p(x_i))] \tag{10}$$

$$= \sum[y_i ln\frac{p(x_i)}{1-p(x_i)} + ln(1 - p(x_i))] \tag{11}$$

$$= \sum[y_i(\omega \cdot x_i) - ln(1 + e^{\omega \cdot x_i})] \tag{12}$$

3.Find ways to minimize $J(\omega)$ and get the best parameters. The estimate function is more accurate if you decrease the numerical value of the $J(\omega)$ function. Therefore, it is necessary to determine the $J(\omega)$ function's minimal value. There are several methods for determining a function's minimal value, and in this case, we employ gradient descent.

There are numerous approaches to solving logistic regression. In this article, we'll discuss gradient descent and Newton's method. Finding a direction in which the loss function's value can be decreased once a parameter is moved is the major goal of optimization. Typically, different combinations of first or second partial derivatives are used to create this orientation. The logistic regression's function of loss is:

$$J(\omega) = -\frac{1}{n}[\sum_{i=1}^{n}(y_i lnp(x_i) + (1 - y_i)ln(1 - p(x_i)))] \tag{13}$$

The descent of gradients is the process of iteratively updating parameters while determining the direction of descent using the first derivative of j with relationship to w. The update method is as follows:

$$g_i = \frac{\partial J(\omega)}{\partial \omega_i} = (p(x_i) - y_i) \cdot x_i \tag{14}$$

$$\omega_i^{k+1} = \omega_i^k - \alpha g_i \tag{15}$$

The sample data after dimensionality reduction was obtained through the factor analysis of the previous data.

2.5. Fisher discriminant analysis

Fisher's test is a projection method that projects a point in a higher dimensional space onto a lower dimensional space. In the original coordinate system, it may be difficult to separate the samples, but the difference may be obvious after projection. In general, it can be projected onto one dimension first, if it does not work well, onto another straight line, and so on. You can build a discriminant function for each projection [11].

In this paper, we hope to establish a new coordinate system through Fisher discriminant method, and project the data of red wine samples that have been grouped by cluster analysis into the new coordinate system for discriminant analysis, and obtain the score of each group through Fisher score function, which can be used as the basis for evaluating the quality of each group of red wine samples.

Since red wine samples are divided into five groups according to different quality in cluster analysis, multiple global Fisher discriminant functions need to be performed here. The main steps are as follows:

The Fisher test seeks to find a projection direction that reflects the differences between groups of combinations, that is, to find a discriminant function that maximizes the differences between populations and minimizes the sum of the squares of deviations within each population. There are five populations, G1, G2, G3, G4, G5, who's mean and covariance matrices are $\mu 1, \ldots, \mu 5$ and $\sum 1, \ldots, \sum 5$, respectively [12].

In the case of $X \in G_i$, there is [12]:

$$E(C'Y) = E(C'Y|G_i) = C'E(Y|G_i) = G'\mu_i, i = 1, \ldots, 5 \tag{16}$$

$$D(C'Y) = D(C'Y|G_i) = C'D(Y|G_i)C = C'\sum_i C, i = 1, \ldots, 5 \tag{17}$$

Make [12]:

$$B = \sum_{i=1}^{5} (C'\mu_i - C'\bar{\mu})^2 = C'\sum_{i=1}^{5} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})'C = C'B_0C \tag{18}$$

$$E = \sum_{i=1}^{5} C'\sum_i C = C'(\sum_{i=1}^{5} \sum i)C = C'E_0C \tag{19}$$

B is the difference between groups and E is the difference within groups. Using the idea of discriminant analysis, construction [12].

$$\triangle(C) = \frac{C'B_0C}{C'E_0C} \tag{20}$$

You get the maximum of $\triangle(C)$, you get the discriminant function. Obviously, B_0, E_0 are non-negative definite matrices. $\triangle(C)$ is the maximum characteristic root of the equation, and the coefficient vector C is the characteristic vector corresponding to the maximum characteristic root.

3. Result

3.1. Result of dimension reduction

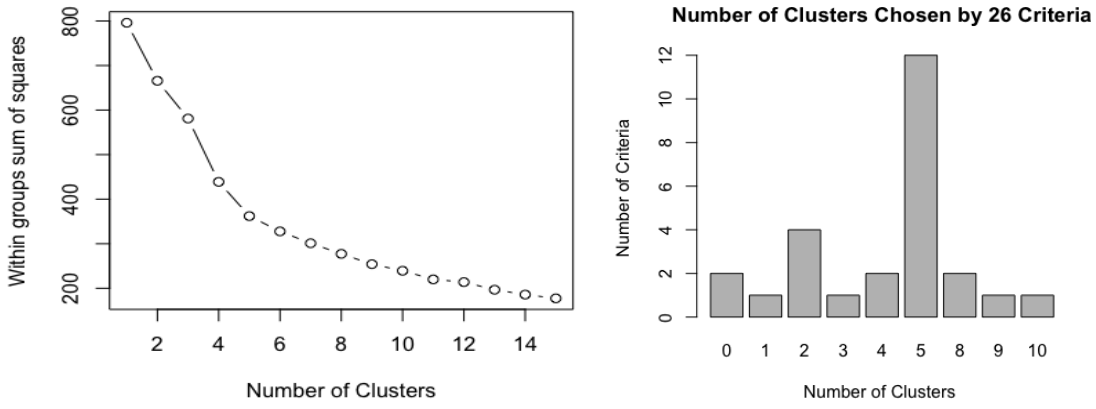
In the first step, the dimensionality of the original data was reduced by using principal component analysis and factor analysis respectively. In the obtained results, the cumulative variance share of the two methods was not much different. Since factor analysis can well reflect the load of each factor and facilitate further interpretation of the meaning of factors, factor analysis is chosen for data dimension reduction finally. The 11 indexes of each red wine sample were synthesized into 5 factors through the thermodynamic correlation matrix as the new indexes of the red wine sample. Through factor analysis, the load of each sample on the corresponding factor is obtained as the factor score of the sample. The factor analysis results obtained from part of the original sample data are shown in Table 3.

Table 3. Factor score of the sample (first six rows)

	R1	**R2**	**R3**	**R4**	**R5**
Sample1	2.149	0.458	-0.428	-0.43	-0.1
Sample2	-0.39	-0.547	-0.407	0.038	-0.41
Sample3	-0.127	0.84	-0.778	-0.543	0.221
Sample4	0.361	-0.036	0.504	-0.419	0.591
Sample5	-0.127	0.84	-0.778	-0.543	0.221
Sample6	0.275	-0.981	-0.32	-0.309	-0.729

3.2. Result of cluster analysis

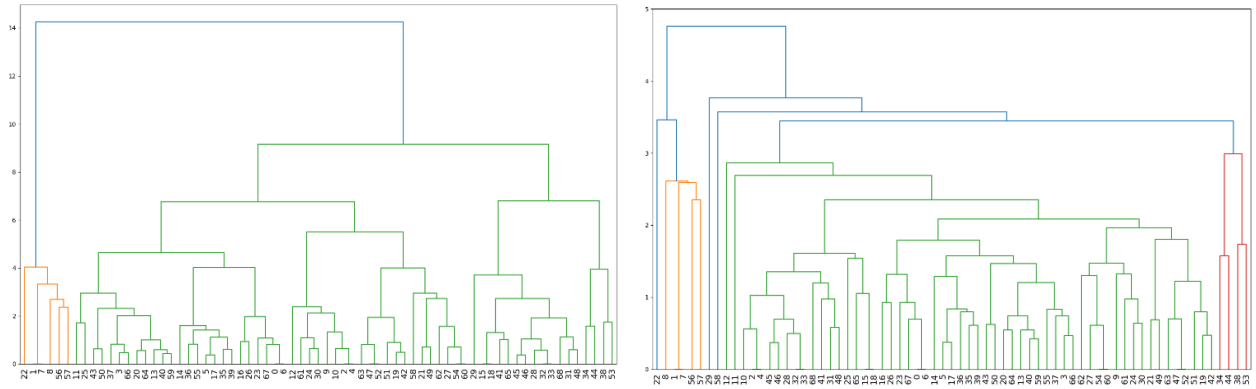
Then, cluster analysis is carried out. Before the clustering analysis, it is necessary to determine the best categories of data so Rstudio is used to determine cluster number, NbClust package is selected as a reference, and lithotripte map is drawn to determine cluster number. Compared between the sum of squares in the group and the number of clusters extracted, the rate of decline from the first category to the fifth category is fast and then slow, indicating that the recommended number of clusters is 5 categories. Then, 26 indicators recommended in the NbClust package are used for classification voting, and the class with high scores is the best number of recommended categories [13].



(a) (b)

Figure 3. (a) is cluster analysis of lithotripsy map. (b) is cluster analysis classification number scores chart.

Then, cluster analysis is performed on the findings of factor analysis using the class average method and the total of squares of deviance algorithm in the process of analytic hierarchy. These are the outcomes of cluster analysis(Because of the large number of sample data, it is not convenient to use the tree graph to show all data classification. The Figure 4 below shows the tree classification of only 70 samples):



(a) (b)

Figure 4. (a) is result of Class average algorithm tree (partial data). (b) is result of Sum of squares of deviation algorithm tree (partial data).

It is hard to properly and simply display the clustering findings of data using a tree graph due to the enormous amount of data samples. Therefore, K-means clustering was further used to perform cluster analysis on the data, and the clustering scatter diagram as shown in Figure 5 was obtained. Meanwhile, the specific clustering result data was shown in Table 4.

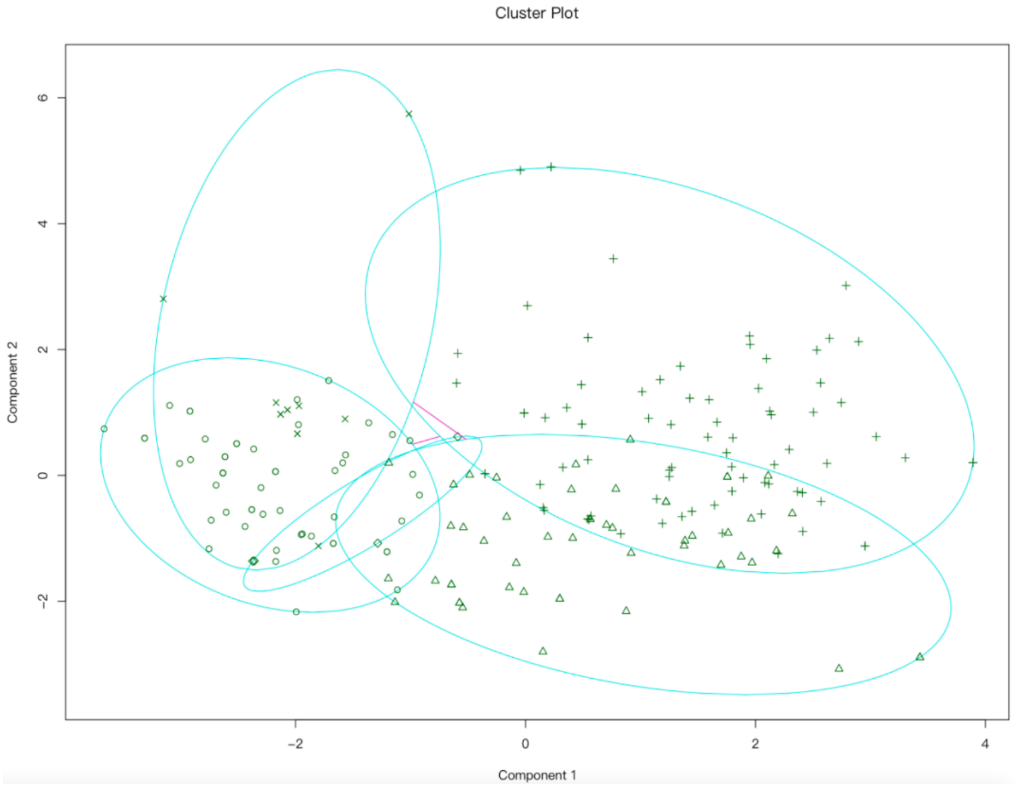


Figure 5. K-means clustering result graph.

Table 4. K-means clustering result table.

Clustering category	Number of clusters	R1	R2	R3	R4	R5
1	8	-0.622	1.398	0.076	2.256	-0.879
2	52	1.148	0.667	0.191	-0.423	-0.506
3	64	0.069	-0.392	-0.441	0.554	0.432
4	15	-0.1027	-0.087	0.173	-0.691	-0.591
5	8	0.163	-1.038	0.483	0.667	2.751

At the same time, the radar map of K-means clustering results is drawn. Through the radar map, we can more clearly and intuitively observe the characteristics of each group of data in the data after

clustering. Combined with the factor meaning of R1 -- R5 analyzed before, we can get the performance of the red wine samples of each group in each factor after clustering. Then the qualitative quality of each group of red wine samples was obtained.

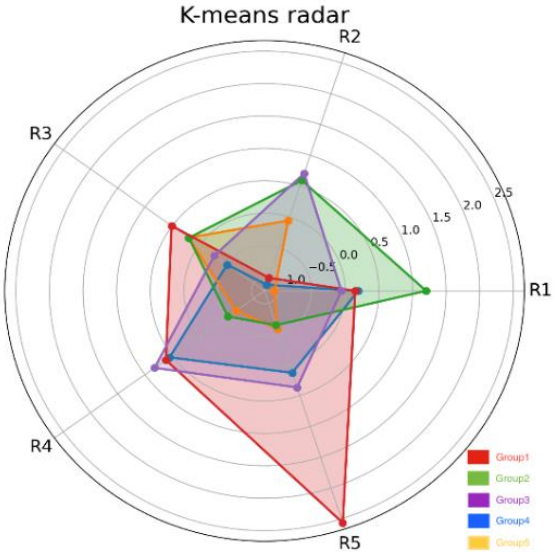


Figure 6. K-means algorithm radar map.

After cluster analysis, the clustering results using K-means clustering method are selected, and the grouping results combined with the meaning of R factor can reflect the quality of each group. Since the sample data were divided into five groups by K-means algorithm, the quality of each group was different, so the numbers 4-8 were selected to represent the quality of the five groups of red wine samples from low to high. Q quality column is added to the sample data to fill in the quality of each sample, which is also represented by numbers 4-8, representing the quality of red wine samples.

Before further analysis, factor score results of the original data and K-means clustering results are combined first, and then feature analysis is carried out to draw a boxplot of the data and observe the features of the new data as shown in the figure below:

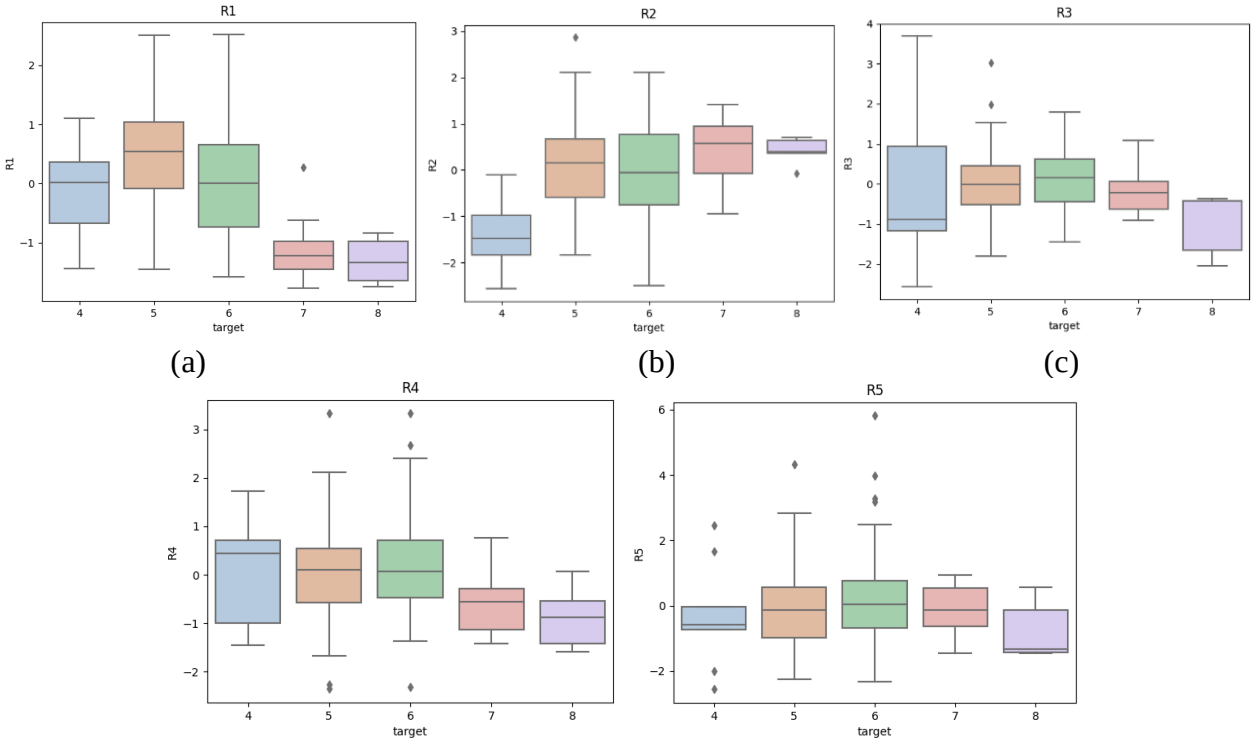


Figure 7. Factor score box plot. (a)-(e) is box plot for R1 to R5, respectively.

It can be seen from the boxplot of R1 -- R5 that there are not too many outliers in the quality column Q, indicating that the addition of the quality of the wine sample is reasonable. At the same time, it can be analyzed that in the sample group with high quality Q of 7 and 8, the score of R1 is relatively low and the score of R2 is relatively high. The samples of lower quality, Q 4 and 5, have higher scores on R1 and R5, and lower scores on R3, R4 and R5. The influence of R1 -- R5 factors on quality can be qualitatively analyzed by the boxplots.

3.3. Logistic regression

Then logical regression analysis is carried out on the data of added quality to determine the accuracy of such grouping. Randomly select 80% of the original data as the training set and 20% as the test set. This is done by using the train_test_split () function in the sklearn library. By comparing the differences between the grouping results of the test set and the original data, the accuracy of the prediction and the corresponding confusion matrix are obtained, as shown in the figure below:

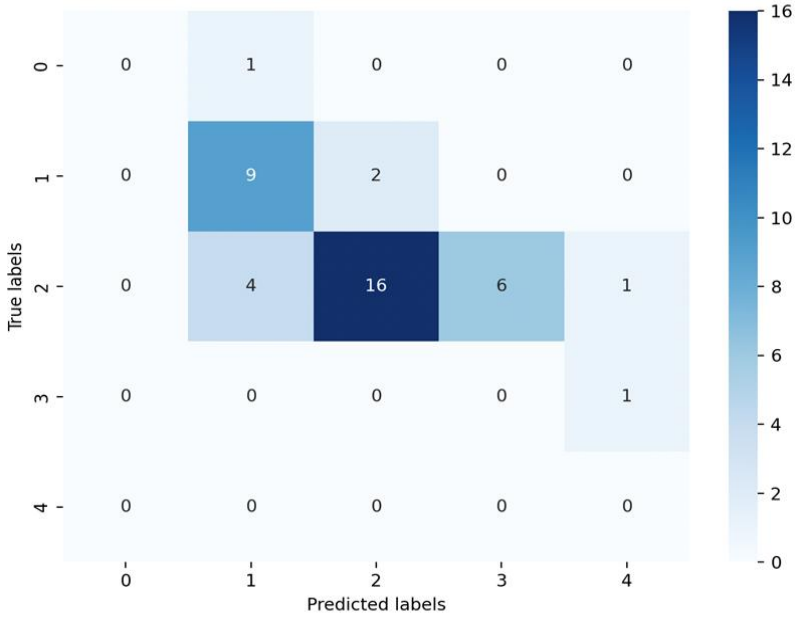


Figure 8. Classification confusion matrix graph.

Then, the ratio of training set to test set of sample data was changed to conduct repeated logistic regression analysis. The prediction accuracy under different training and testing ratios is obtained, which can more fully reflect the rationality and reliability of data grouping, as shown in Table 5:

Table 5. Result of logistic regression.

Training set ratio	Test set ratio	Prediction accuracy
0.6	0.4	0.70
0.65	0.35	0.66
0.7	0.3	0.65
0.75	0.25	0.67
0.8	0.2	0.67
0.85	0.15	0.65
0.9	0.1	0.66

3.4. Result of Fisher discriminant analysis

Finally, Fisher discriminant analysis is carried out. The same regression Fisher discriminant analysis can also make training prediction on the data set and judge whether the grouping is reasonable. At the same time, the discriminant function of red wine quality can be obtained. By combining the scoring function with the meaning of R factor, the quality of five groups of red wine

samples can be obtained. Each group of red wine samples can be scored quantitatively, and the difference of the quality of each group of red wine can be visually and clearly reflected by the number.

The score function obtained by Fisher discriminant analysis is: $S = 1.400 * \mathrm{R1} + 0.498 * \mathrm{R2} + 0.588 * \mathrm{R3} + 0.393 * \mathrm{R4} + 0.962 * \mathrm{R5}$

R1, R2, R3, R4 and R5 here represent the mean value of R1 -- R5 of each group of red wine samples, and S represents the final quality score of each group. The quality increases as the score rises. The final calculation results are summarized as shown in Table 6:

Table 6. Fischer discriminant analysis results table.

Category	Quantity	R1 mean	R2 mean	R3 mean	R4 mean	R5 mean	Total score
1	8	-0.45	-0.87	-0.30	-0.58	-1.97	-3.36
2	52	1.06	0.07	0.177	0.20	0.12	1.82
3	64	-0.76	-0.35	-0.04	0.86	00.31	-0.63
4	15	-1.05	0.92	5.05	-1.33	-0.22	1.22
5	8	-0.45	0.58	-0.49	-0.82	0.38	-0.59

The higher the total score is, the higher the overall quality of the wine is. That is, the quality of the wine samples in the second category is the best, while the quality of the wine samples in the first group is the worst, which is also consistent with the previous analysis results of the radar map.

4. Conclusion

Through the research, this paper analyzes the quality of 200 red wine samples by machine learning algorithm, and finally obtains five groups of red wine samples with different quality, and gives specific suggestions on improving the quality of red wine for each group of red wine samples. For the first type of red wine, the contents of sulfate, sulfur dioxide, non-volatile acid and citric acid can be appropriately increased to reduce the volatile content, so as to improve the taste and taste of red wine products, and improve the antioxidant capacity of red wine. For the second type of red wine can be appropriate to improve its sweetness. The third type of red wine needs to increase its alcohol content and acidity, so that the taste of red wine, and also conducive to bacterial growth. For the fourth type of red wine, the content of sulfur dioxide and chloride in red wine can be considered to improve the color of red wine and prolong the preservation time. The fifth type of red wine increases the content of sulfates, citric acid and non-volatile acids, thus inhibiting bacterial growth.

In this study, the quality of wine samples was evaluated scientifically and systematically, and a wine quality evaluation system model was initially established. It provides a method for wine manufacturers and suppliers to test the quality of wine, and can also assist professional wine tasters to evaluate the quality of wine.

Finally, due to the few numbers of samples surveyed in this study, all sample indexes of red wine and all factors affecting the quality of red wine were not fully considered. More professional judgment is also needed in the part of factor analysis interpretation, and more detailed wine quality system model can be further refined and improved in the above part, so as to facilitate the in-depth study of this topic.

References

[1] BarbaraPrice. "A First Course in Factor Analysis." Technometrics 35.4 (1992): 453 - 453.

[2] Heckler, C. E., "A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling." Technometrics 38.3 (1994): 296 - 297.

[3] Kaiser, and F. H., "The Application of Electronic Computers to Factor Analysis." educational & psychological measurement 20.1(1960): 141 - 151.

[4] Kaufman, L., and P. J. Rousseeuw. "Finding Groups in Data: An Introduction to Cluster Analysis." Wiley (2005).

[5] Hansen, Pierre, and B. Jaumard. "Cluster Analysis and Mathematical Programming." Mathematical Programming 79.1 (1997): 191 - 215.

[6] Hartigan, J. A., and M. A. Wong. "A K-Means Clustering Algorithm." Applied Statistics 28.1 (1979).

[7] Wagstaff, Kiri. "Constrained K-means Clustering with Background Knowledge." Proceedings of ICML-2001 (2001).

[8] Hamerly, G., and C. Elkan. "Learning the K in K-Means." Advances in neural information processing systems 17(2004).

[9] Cucchiara, Andrew. "Applied Logistic Regression." Technometrics 34.3 (2012): 358 - 359.

[10] Pregibon, D. "Logistic Regression Diagnostics." Annals of Statistics 9.4 (1981): 705 - 724.

[11] Song, F., S. Liu, and J. Yang. "Orthogonalized Fisher discriminant." Pattern Recognition 38. 2 (2005): 311 - 313.

[12] Jian, Y., J. Y. Yang, and D. Zhang. "What's wrong with Fisher criterion?" Pattern Recognition 35.11 (2002): 2665 - 2668.

[13] Robert I. Kabacoff. "R in action. "Post &Telecom Press (2013).