Research on the Composition Analysis and Identification of Ancient Glass Products Based on Decision Tree and K-means Clustering Algorithm

Wenxuan Hu [1, *], Yuping Ke [1], Yueyang Liu [2], Xingzhi Dong [2], Chaohao Hu [2]

[1] School of Internet, Anhui University, HeFei, AnHui, 230039

[2] Stony Brook Institute, Anhui University, Hefei, Anhui, 230039

* Corresponding author:huwenxuan03@163.com

Abstract. Ancient glass is susceptible to weathering by environmental influences, resulting in changes in its chemical composition and proportions, which affects the correct judgment of its category by archaeologists. In order to correctly analyze and identify the composition of ancient glass artifacts, firstly, this paper establishes a decision tree model to derive the basis for determining the type of glass artifacts. Secondly, the aggregation coefficients of high potassium glass and lead-barium glass were plotted separately in this paper, and the K-value of clustering was determined to be 2 according to the elbow rule. After using principal component analysis to reduce the dimensionality of the chemical composition content data, the K-means clustering algorithm was used to classify the subclasses of high potassium glass and lead-barium glass respectively. Finally, the established model was used to identify the chemical composition and its type of unknown class of glass artifacts.

Keywords: Ancient glass, decision tree, K-means clustering.

1. Introduction

Glass products from early West Asia and Egypt were introduced to China through the Silk Road and were valuable physical evidence of trade exchanges. After learning their technology, China took and made them locally, and they were similar in appearance and different in chemical composition from the foreign glass products. Therefore, it is important to analyze and identify the composition of ancient glass products. It should be noted that ancient glass is susceptible to weathering due to environmental influences, resulting in changes in its chemical composition and proportions, which affects the correct judgment of archaeologists on its category. Therefore, in order to correctly analyze and identify the composition of ancient glass objects, machine learning algorithms are used for classification and category delineation. Machine learning methods have been applied to the classification and identification of objects many times in recent years. Xie Zhaoxian,Zou Xingmin, and Zhang Wenjing (2023) proposed a novel parametric pruning decision tree algorithm for large data sets, which finds the optimal parameters in the process of decision tree generation, thus completing pruning from top down, reducing the risk of overfitting and improving the accuracy and efficiency of the model. [1] Weijia Gong (2022) established a hybrid expert model for task division based on k-means to divide the data into different task ranges, thus improving the accuracy of lithology identification. [2] Liheng Zhang (2022) proposed an adaptive k-means clustering algorithm, which was optimized in three areas of similarity, determination of initial clustering K-values, and selection of initial clustering centers, respectively, to perform complete segmentation of coke optical tissue images, resulting in significant improvement of classification results. [3].

The data on the proportion of chemical composition of classified glass artifacts used in this paper were obtained from the C problem of the 2022 National Student Mathematical Modeling Competition. Firstly, this paper used a supervised learning decision tree algorithm to classify ancient glass artifacts, drew a decision tree and obtained the classification criteria of high potassium glass and lead-barium glass, and obtained the classification law of ancient glass artifacts. Secondly, we further applied the unsupervised learning algorithm K-means to cluster the data, so as to select the appropriate chemical composition to classify the subclasses of high-potassium glass and lead-barium glass. Finally, we

conducted the reasonableness and sensitivity analysis of the proposed model, and the results showed that the obtained classification model was reasonable and accurate.

2. Model establishment and solution

2.1. Classification law of glass types

- **Pre-processing of data**

Since the data in this paper have the characteristic of compositionality, that is, the cumulative sum of the proportion of each component is 100%, but the cumulative sum of the proportion of each component may not be equal to 100% due to the detection means and other reasons, therefore, the data with the cumulative sum of the proportion of components between 85% and 105% are regarded as valid data in this problem. By observing and analyzing the data, it can be seen that some of the original data are abnormal, and here the outliers are processed to remove the abnormal data. In addition, for the components that were not detected, the missing values are processed here to facilitate subsequent statistics, i.e., the corresponding null values are set to zero.

- **Building a decision tree model**

a. The basic structure of decision tree

Decision trees were first proposed to deal with decision problems, and they have the advantages of simple structure, clear logic and good interpretability, and the best decision tree is constructed to predict unknown data categories by known "a priori data" [4].

Its basic structure is roughly divided into: decision nodes, branch nodes, and leaf nodes, as shown in Figure 1:

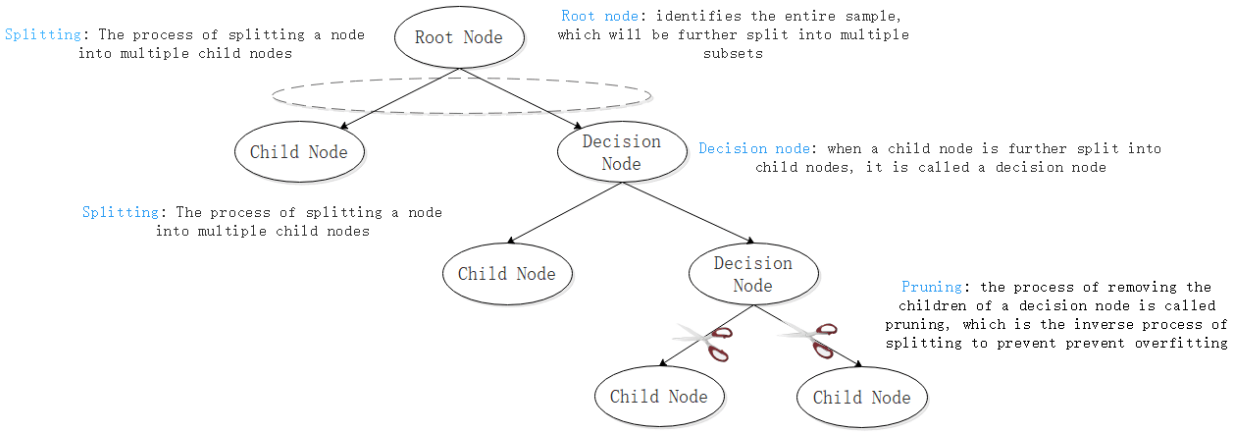

Figure 1. Schematic diagram of the basic structure of the decision tree

b. Decision tree generation

In this paper, the glass type is used as the target variable of the decision tree, and each chemical composition given in the question is used as the feature variable to divide the training set and the test set in the ratio of 7:3.

- **Learning objective**

To train the data set with different types of glass according to the given chemical components to obtain the classification laws for different glass types while ensuring the minimum loss function.

- **Basis of tree construction**

The main index used is the Gini coefficient (gini).

The Gini coefficient is used to calculate the disorder phenomenon in the system, and the higher the Gini coefficient, the higher the degree of system disorder. [5] Therefore, it is necessary to find a suitable classification to reduce the value of this indicator when building a decision tree model. Its calculation formula is as follows:

$$gini(T) = 1 - \sum p_i^2 \tag{1}$$

Where is the probability of occurrence of category i in the re-sampled data T, i.e., the ratio of samples with category i to the total number of samples.

- **Construction process**

Step1.Construct the root node, put all the training data in the root node, select an optimal feature, and partition the training data set into subsets according to this feature, so that each subset has a classification that is the best under the current conditions.

Step2. If these subsets can already be classified basically correctly, then construct leaf nodes and divide these subsets into the corresponding leaf nodes.

Step3. If there are still subsets that cannot be classified correctly, then select new optimal features for these subsets, continue to segment them and build the corresponding nodes, and so on recursively until all the subsets of training data are basically correctly classified or there are no suitable features.

Step4. Each subset is divided into leaf nodes, i.e., all have a clear class, so that a decision tree is generated.

According to the above method, the given data is input and solved using python, and the classification law obtained is as follows figure.2.

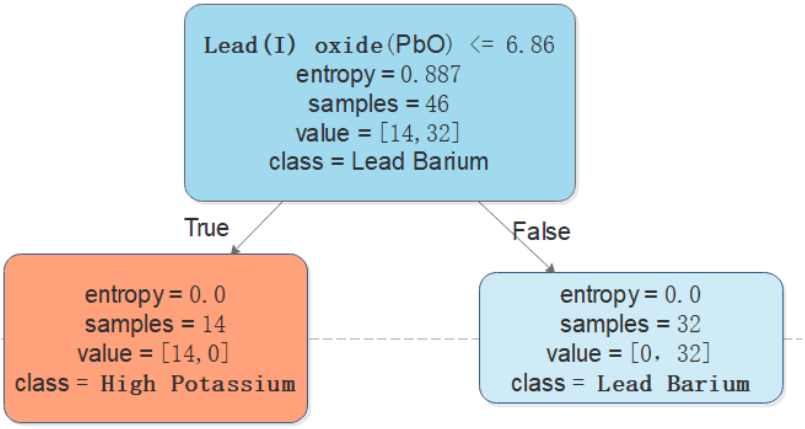

Figure 2. Glass type division decision tree

As can be seen from the above chart, high potassium glass and lead-barium glass can be divided according to the content of lead oxide. When the content of lead oxide is less than or equal to 6.86, the glass artifact can be judged as high potassium glass, and vice versa for lead-barium glass.

2.2. Subclass division

In this paper, K-means was used to cluster the data.

- **Determination of the number of clusters K-value**

In this paper, the optimal number of clusters K can be roughly estimated by drawing a graph using the elbow rule, which is calculated by finding the location (called the elbow) where the cost function, i.e., the category distortion degree, has the greatest improvement effect. [6] where the degree of aberration of each class is equal to the sum of the squares of the distances between that center of gravity and its internal member positions:

Suppose a total of n samples are divided into K classes (, i.e., at least two elements in one class), denoted by the kth class (),and the position of the center of gravity of that class is noted as, then the degree of aberration of the kth class is (here the absolute value sign denotes the distance):

$$\sum_{i \in C_k} \left| x_i - u_k \right|^2 \tag{2}$$

Therefore, we can define the total aberration degree (aggregation coefficient) of all classes as follows:

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} |x_i - u_k|^2 \tag{3}$$

Based on the above knowledge a line graph of aggregation coefficients is drawn as shown in Figure 3.

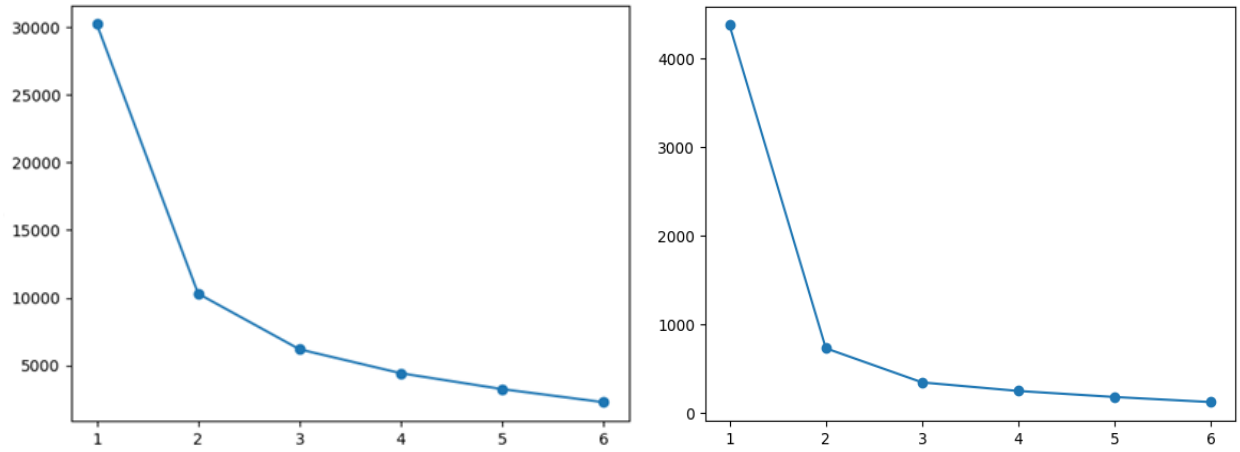

Figure 3. Folding graph of polymerization coefficient for high potassium glass (left) and lead-barium glass (right)

From the above figure, it can be seen that the change in the degree of distortion is larger when the K value is from 1 to 2. After the K value reaches 2, the change in the degree of distortion decreases significantly. That is, the rate of decline suddenly slows down at K=2. Therefore, the elbow is K=2, so the number of categories of clusters can be set to 2.

- **Principal component analysis**

In order to facilitate the cluster analysis of the subsequent glass product types, this paper reduced the dimensionality of the chemical composition content data given in the question.

The spacing between sample data was defined using variance:

$$Var\left(x_j\right) = \frac{1}{n} \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2 \tag{4}$$

The sample data is first processed by subtracting the sample mean from all samples so that)the)sample mean goes to zero, at which point:

$$Var\left(x_j\right) = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 , \quad \overline{x}_j = 0 \tag{5}$$

Find the unit vector, with the largest variance after mapping the sample points such that the variance after mapping the sample to:

$$Var\left(X\right) = \frac{1}{n} \sum_{i=1}^{n} \left(X_1^{(i)} \omega_1 + X_2^{(i)} \omega_2\right)^2 \tag{6}$$

Where is the sample data matrix.

From this, the chemical composition content data can be dimensionalized, and two sets of characteristic variables can be obtained. A scatter plot of the relationship between the two sets of data is plotted using Python (the horizontal and vertical axes indicate features 1 and 2 obtained after dimensionality reduction, respectively), as shown in Figure.4.

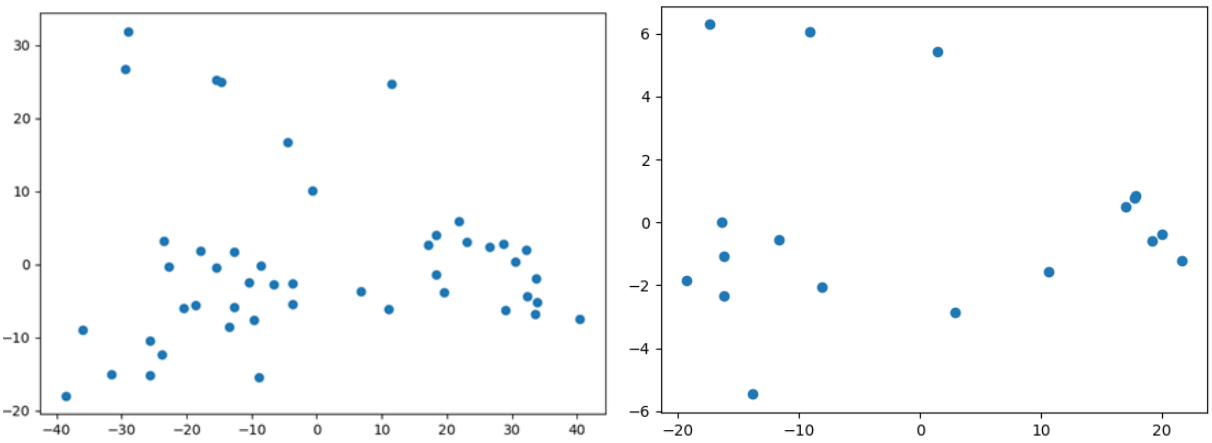

Figure 4. Scatter plots of characteristic variables for lead-barium glass (left) and high-potassium glass (right)

- **K-means clustering**

In order to subclassify the ancient glassware types, the K-means algorithm was chosen to cluster the data for analysis in this paper, and the aggregation results are shown in Figures.5. and Figures.6.

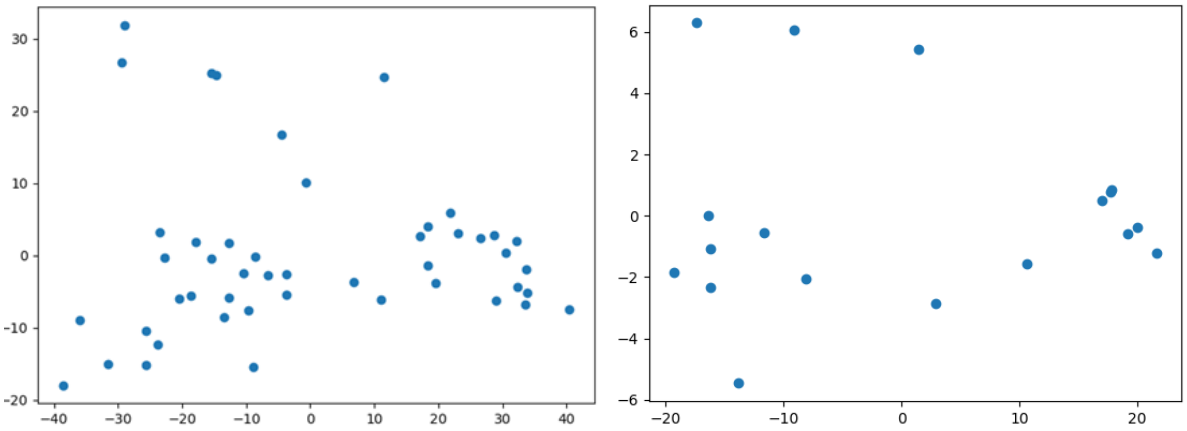

Figure 5. The effect of lead barium glass before and after clustering

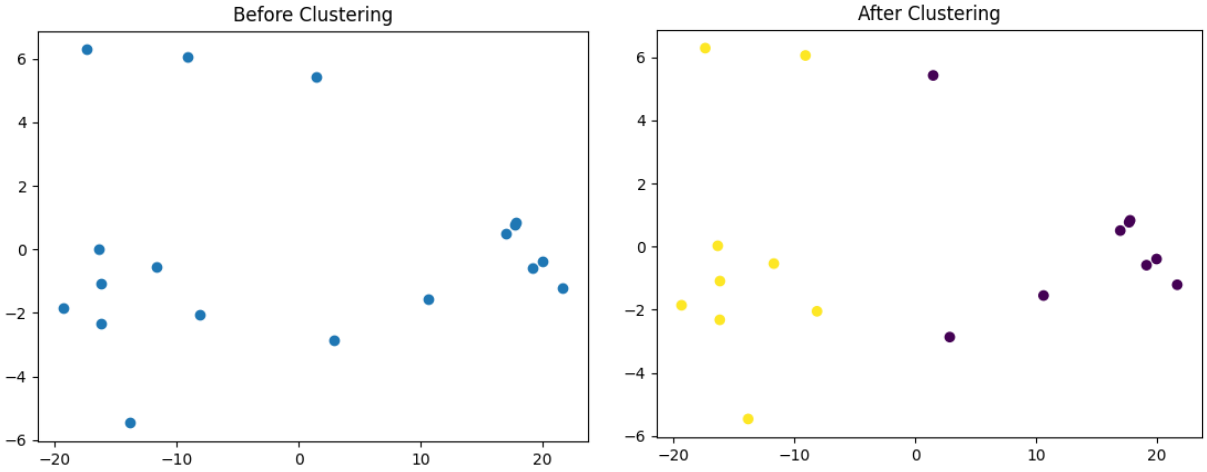

Figure 6. Before and after effect of high potassium glass clustering

In the above figure, the horizontal and vertical axes indicate the two sets of characteristic variables obtained after performing principal component analysis, and the right figure shows the results obtained after clustering analysis. It can be seen that K-means divides the above data into two categories, which are indicated by purple and yellow, respectively.

In addition, this paper also establishes a hierarchical clustering algorithm for comparative analysis, i.e., a hierarchical nested clustering tree is created by the similarity between data points of different

categories. It can be seen that both distance-based K-means clustering and hierarchical clustering work best when the number of clusters is 2, and the clustering results obtained by the two clustering methods are basically the same. The hierarchical tree obtained by hierarchical clustering is shown in Figure 7.

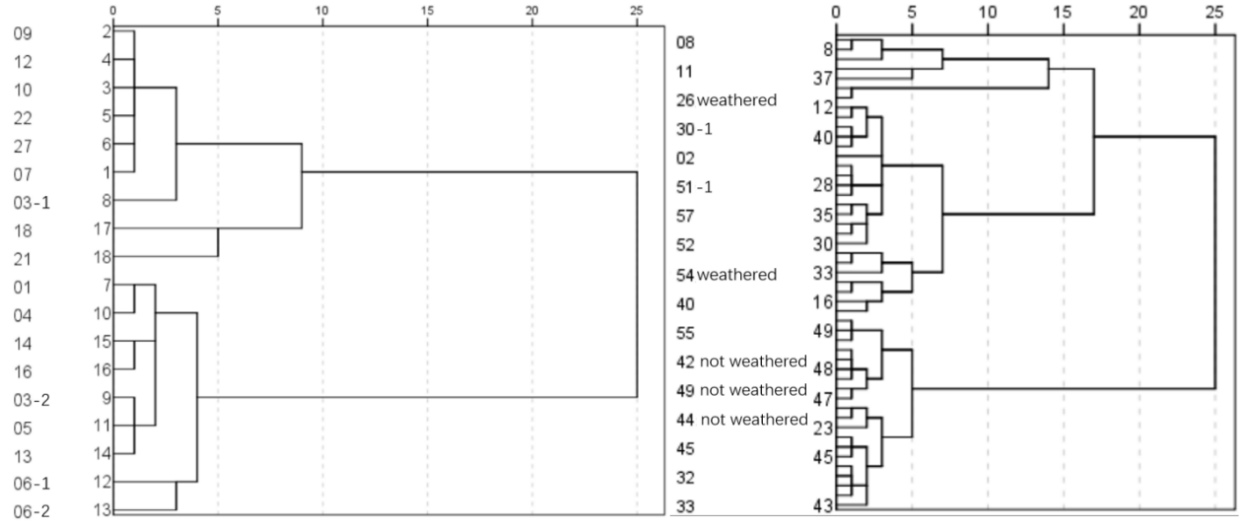

Figure 7. Hierarchical clustering tree for high potassium glass (left) and lead-barium glass (right)

2.3. Reasonableness and sensitivity analysis

- **Reasonableness Analysis**

First define the following concepts:

TP (True Positive): predicting the answer correctly;

FP (False Positive): wrong prediction of other classes as this class;

FN (False Negative): wrong prediction of the label of this class as other class;

Accuracy (precision): the proportion of positive samples in positive cases determined by the classifier [7];

$$precision = \frac{TP}{TP + FP} \tag{7}$$

Recall: refers to the proportion of predicted positive cases to the total number of positive cases;

$$recall = \frac{TP}{TP + FN} \tag{8}$$

Accuracy: indicates the proportion of classifier's correct judgments for the whole sample

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

The F1-score is an important measure of the classification problem and is the summed average of the precision and recall rates.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{10}$$

Using Python, the evaluation metrics of the clustering model for this question were calculated as follows:

F1-score:1.0
acc-score:1.0

That is, the accuracy and F1-score of the model have reached 100%, and the model fit is excellent.

- **Sensitivity analysis**

Sensitivity analysis of the experimental results was performed using K-fold cross-validation. The original chemical content data were divided into K groups (set in this problem K=10), and each subset of data was used as a validation set, while the remaining K-1 subsets were used as the training set of the clustering model, so that K models could be obtained, and the classification accuracy evaluation index of the final validation set of these K models was used as the sensitivity analysis performance index of the model.

The results obtained using Python are shown in Figure 8:

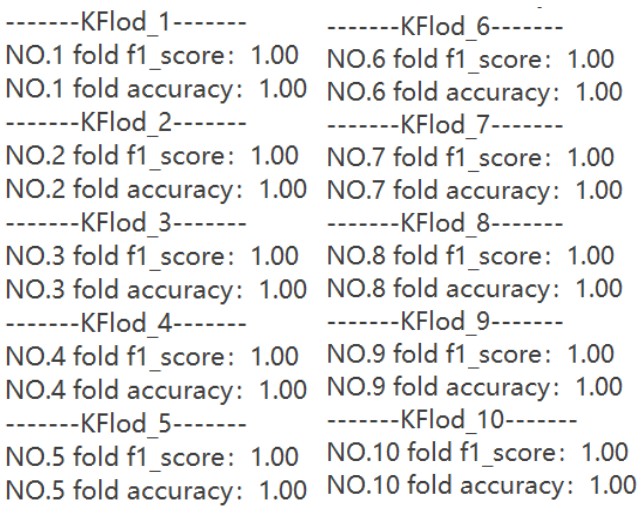

Figure 8. K-fold cross-validation test results

This shows that the model has 100% accuracy and F1 score at cross-validation, i.e., the model classifies better.

3. Type identification of glass products

3.1. Type identification

The following is the content of chemical components of 8 glass cultural relics.

Table 1. The chemical composition data of the unknown category of glass artifacts.

Number	A1	A2	A3	A4	A5	A6	A7	A8
Surface weathering	No	Yes	No	No	Yes	Yes	Yes	No
SiO_2	78.45	37.75	31.95	35.47	64.29	93.17	90.83	51.12
Na_2O					1.2			0.00
K_2O			1.36	0.79	0.37	1.35	0.98	0.23
CaO	6.08	7.63	7.19	2.89	1.64	0.64	1.12	0.89
MgO	1.86		0.81	1.05	2.34	0.21		0.00
Al_2O_3	7.23	2.33	2.93	7.07	12.75	1.52	5.06	2.12
Fe_2O_3	2.15		7.06	6.45	0.81	0.27	0.24	0.00
CuO	2.11		0.21	0.96	0.94	1.73	1.17	9.01
PbO		34.3	39.58	24.28	12.23			21.24
BaO			4.69	8.31	2.16			11.34
P_2O_5	1.06	14.27	2.68	8.45	0.19	0.21	0.13	1.46
SrO	0.03		0.52	0.28	0.21			0.31
SnO_2					0.49			0.00
SO_2	0.51						0.11	2.26

According to the above model, the chemical composition data of the unknown category of glass artifacts are substituted into the model to obtain.

Table 2. Results of classification

Number	A1	A2	A3	A4	A6	A7	A8
Type	1	2	2	2	1	1	2

3.2. Sensitivity analysis

The traditional statistical model is model-driven, while the machine learning model (K-means) used in this question is data-driven. When using the machine learning algorithm for prediction, the model parameters have been determined, and generally only sensitivity analysis is required for the selection of model parameters during the training process, and no sensitivity analysis is required for prediction. And in the training process, it is known from the elbow rule that the model training effect is poor when K=1 or 3, and the model training effect is best when K=2. The precision, recall, accuracy and F1 score all reach a very high level, and the model classification effect is good, so the classification effect of the model can be modeled has reached a high level.

4. Conclusion

In this paper, the classification laws of high potassium glass and lead-barium glass were obtained by analyzing the relevant data, and subclassing the two categories and analyzing the rationality and sensitivity of the classification results. Firstly, a decision tree model was established in this paper, and the pre-processed data were substituted for the training: when the lead oxide content was less than or equal to 6.86, the glass artifacts could be judged as high potassium glass, and the opposite was lead-barium glass. Next, the aggregation coefficients of high potassium glass and lead-barium glass were plotted separately in this paper, and the K value of clustering could be determined as 2 according to the elbow rule. after using principal component analysis to reduce the dimensionality of the chemical composition content data, the K-means clustering algorithm was used to divide the high potassium glass and lead-barium glass into subclasses respectively, and the division results are shown in Figs. 16-17. next, the model was tested in this paper, and the model The accuracy and F1 score of the model reached 100%, therefore, the model is reasonable. Finally, the K-fold cross-validation method was used to perform sensitivity analysis on the experimental results, and the results of cross-validation showed that the accuracy and F1 scores of the model were still at extremely high levels, i.e., the classification effect of the model was good. Secondly, this paper identifies the types of unknown categories of glass artifacts based on their chemical composition given and performs sensitivity analysis on the classification results. The data given in the question can be directly substituted into the established clustering model for training, and the results obtained are: artifacts A1, A6, and A7 belong to category 1, and artifacts A2, A3, A4, A5, and A8 belong to category 2. Since the model parameters have been determined when the K-means algorithm is used to predict the type of artifacts in this paper, it is meaningless to conduct sensitivity analysis on the classification results, and can be replaced by a sensitivity analysis on the model sensitivity analysis on the choice of parameter K during the training process: from the elbow rule, it is known that the model training effect is poor when K=1 or 3, and the model training effect is best when K=2. The precision, recall, accuracy and F1 score all reach a very high level, and the model classification effect is good.

References

[1] Xie Zhaoxian, Zou Xingmin, and Zhang Wenjing. Research on efficient parameter pruning decision tree algorithm for large datasets[J/OL]. Computer Engineering: 1 - 11 [2023-04-08]. https://doi.org/10.19678/j.issn.1000 - 3428.0066519.

[2] Gong Weijia. Research and application of lithology recognition based on ensemble learning[D]. Northeast Petroleum University, 2022. DOI: 10.26995/d.cnki.gdqsc.2022.000332.

[3] Zhang Liheng. Research on segmentation and recognition algorithm of optical microstructure images of coke [D]. Wuhan University of Science and Technology, 2022. DOI: 10.27380/d.cnki.gwkju.2022.000579.

[4] Lu Yanyi. SMT loose material sorting and mounting based on machine vision[D]. Hangzhou Dianzi University, 2020. DOI: 10.27075/d.cnki.ghzdc.2020.000556.

[5] Zhang Junqi, Zhu Bin, and Zhang Youwei. Research on intelligent ventilation control of mines based on decision tree [J]. Modern Machinery, 2023 (01): 90 - 94. DOI: 10.13667/j.cnki.52-1046/th.2023.01.015.

[6] Yin Zhong Dong, Chen Junye, and Shen Zilun, et al. Identification algorithm for winding material of distribution transformers based on Kmeans clustering[J/OL]. Journal of North China Electric Power University: 1-9 [2023-04-08]. http://kns.cnki.net/kcms/detail/13.1212.TM.20221226.0855.001.html.

[7] Du Yang. Research on Chinese short text classification technology based on semi-supervised clustering [D]. Jiangsu University, 2020. DOI: 10.27170/d.cnki.gjsuu.2020.000402.

[8] Ding Mengya. Research on adaptive classification and identification method for bamboo species papermaking[D]. Anhui Agricultural University, 2022. DOI: 10.26919/d.cnki.gannu.2022.000073.

[9] Wan Yongjing and Lin Jiajun. A fuzzy classification system modeling method for microstructure image recognition of fiber-reinforced composite materials [J]. East China University of Science and Technology (Natural Science Edition), 2008 (03): 417 - 421. DOI: 10.14135/j.cnki.1006-3080.2008.03.026.

[10] Wang Quande, Wen Biyang, and Wang Xianpei. Insulation material hydrophobicity level determination based on image recognition and classification [J]. Electric Machines and Control, 2008, No.51 (01): 93 - 98.

[11] Cheng Guojian, Yang Jing, Huang Quanzhou, et al. Research on rock thin section image classification and recognition based on probabilistic neural network [J]. Science Technology and Engineering, 2013, 13 (31): 9231 - 9235.