Study and methodology exploration of high-potassium glass and lead-barium glass classification patterns based on cluster analysis and decision tree

Ruixuan Xu[1,2,7], **Xin Xiong**[1,3], **Yingying Hou**[1,4], **Yilan Wu**[1,5], **Yuhua Wang**[1,6]

[1]Tianjin University of Science and Technology

[2]xuruixuan@mail.tust.edu.cn
[3]tust xiongxin@mail.tust.edu.cn
[4]1791152334@qq.com
[5]wuyilan5555@126.com
[6]2218396056@qq.com
[7]Corresponding author

Abstract. This study aims to analyze the classification patterns of high-potassium glass and lead-barium glass. It involves preprocessing the data through cluster analysis and employing machine learning decision trees and K-means clustering models for in-depth analysis. In the clustering analysis model, we start by performing a cluster difference analysis on the feature attributes of high-potassium glass and lead-barium glass to obtain initial classification results. Subsequently, we use machine learning decision tree models to partition the data into training and testing sets to explore the classification patterns.

Construct a K-means clustering model by iteratively determining the initial cluster centroids' positions using an algorithm and calculating the silhouette coefficient for different numbers of clusters.This process aids in determining the appropriate number of clusters and the initial cluster centroids' positions. The results indicate significant differences in various features between high-potassium glass and lead-barium glass, and the employment of decision trees and K-means models successfully classifies artifacts. In conclusion, this study provides a robust method for the classification of high-potassium glass and lead-barium glass, offering valuable insights for research and application in related fields.

Keywords: CART (Classification And Regression Trees) Decision Tree, K-means Clustering Model, Data Mining Techniques

1. Introduction

The establishment, growth, and prosperity of the ancient Silk Road in China significantly propelled economic and cultural exchanges between the East and West.[1] With the expansion of the Silk Road, foreign glass products made their way into China, broadening the horizons of the Chinese and introducing advanced production techniques. While adopting foreign technologies, China began to manufacture glass using local resources.

Consequently, although ancient Chinese glass products resemble foreign counterparts in appearance, their chemical compositions differ markedly.[2] In the study of ancient Chinese glass, defining its chemical composition is pivotal. The types of ancient glass are diverse, including lead-barium glass with high amounts of PbO and BaO, numerous sodium-calcium glass types, and high-potassium glass with K_2O as the primary flux. Due to prolonged burial, ancient glasses often underwent environmental erosion and weathering, altering their appearance and chemical compositions. Delving into the origin and evolution of glass not only deepens our understanding of it but also rekindles interest in traditional craftsmanship.

Furthermore, elucidating the techniques and components of ancient Chinese glass is crucial for studying the ancient cultural and technical exchanges between China and other civilizations. Therefore, we aim to predict and analyze the components of different ancient glass products, hoping to contribute to the developme nt of ancient Chinese glass technology.

2. Analyzing the classification patterns of high-potassium glass and lead-barium glass.

This study employed cluster analysis to process data related to high-potassium glass and lead-barium glass. Using MATLAB, we established a machine learning decision tree analysis model to determine the classification criteria for the two types of glass. Furthermore, a K-means clustering model was developed to differentiate subcategories, and the resulting classifications were subjected to a rationality and sensitivity analysis. The data utilized in this research originated from the 2022 Mathematical Contest in Modeling Problem C.

2.1. Establishing a clustering analysis model to filter data.

Given the multitude of characteristic variables present in high-potassium glass and lead-barium glass, this study initiated by conducting a cluster analysis on these two glass categories based on their respective attributes, as illustrated in Table 1. Subsequently, building upon the summarized results of the clustering process, an analysis of the frequencies within each clustering category was performed. The results of this analysis are presented in Table 2.

Table 1. Table for Clustering Field Differentiation Analysis of High-Potassium Glass and Lead-Barium Glass.

	Cluster Categories (Mean ± Standard Deviation)		F-test results	Significance P-values
	Category 1 (n=35)	Category 2 (n=32)		
Decoration or Ornamentation	1.0±0.0	2.0±0.0	-	**0.000*****
Color	1.914±0.658	1.25±0.44	23.115	**0.000*****
SiO_2	2.429±1.975	3.844±2.288	7.382	**0.008*****
Na_2O	68.711±14.083	27.488±11.18	173.895	**0.000*****
K_2O	1.223±2.09	0.308±0.761	5.477	**0.022****
CaO	3.385±4.9	0.165±0.332	13.74	**0.000*****
MgO	2.523±2.786	2.542±1.73	0.001	**0.974**
Al_2O_3	0.811±0.632	0.543±0.651	2.901	**0.093***
Fe_2O_3	5.269±3.588	2.698±1.531	14.068	**0.000*****
CuO	0.995±1.407	0.69±0.858	1.118	**0.294**
PbO	1.595±1.415	2.347±2.864	1.906	**0.172**
BaO	9.079±9.601	41.29±12.25	144.8	**0.000*****
P_2O_5	3.723±4.172	12.216±9.65	22.535	**0.000*****
SrO	0.845±1.189	4.697±4.146	27.755	**0.000*****
SnO_2	0.136±0.184	0.4±0.278	21.394	**0.000*****
SO_2	0.111±0.451	0.041±0.13	0.727	**0.397**

[a] ***、**、* representing significance levels of 1%, 5%, and 10% respectively.

Table 2. Results of Clustering Models for High-Potassium Glass and Lead-Barium Glass.

Cluster Categories	Frequency	Percentage (%)
Category 1	35	52.239%
Category 2	32	47.761%
Total	67	100.0%

The tables above presents the outcomes of quantitative field differentiation analysis, including results for mean ± standard deviation, F-test results, and significance P-values. In this context, we determine field differentiation by assessing whether each analysis item is < 0.05. Significance of differentiation is indicated by rejecting the null hypothesis, suggesting substantial differences between the two data sets. Analysis of these differences is performed using the mean ± standard deviation method. Conversely, if differentiation is not significant, it indicates that the data does not exhibit notable disparities.

Considering the P-values for the five variables, MgO, $\mathrm{Al_2O_3}$, CuO, PbO, and $\mathrm{SO_2}$, are all > 0.05, for the purpose of subsequent machine learning decision tree classification analysis, these variables are excluded. Subsequently, the remaining variables are categorized into high-potassium glass and lead-barium glass, and MATLAB programming is employed to perform machine learning decision tree model analysis.

2.2. Establishing a machine learning decision tree analysis model to derive classification patterns

2.2.1. Algorithmic Approach

For the remaining variables, categorized as high-potassium and lead-barium, we employed a machine learning decision tree model for analysis. Utilizing the if-then tree structure, we established a classification tree, wherein each leaf node represents a classification label.[3]

The logic behind constructing the decision tree model in this study is as follows: starting from the root node, for each sample of high-potassium glass and lead-barium glass, we evaluate the filtered features. Based on the evaluation outcome, instances are allocated to their respective child nodes. Each node corresponds to a value of a specific feature. Through recursive assessment and allocation, instances are progressively assigned to leaf nodes. Every path from the root node to a leaf node forms a rule. Each feature-based criterion rule exhibits a significant attribute: they are mutually exclusive and collectively exhaustive.

$$\begin{cases}\text{Completeness: Each instance is covered by at least one rule path.}\\ \text{Exclusivity: Each instance is covered by only one rule path.}\end{cases}$$

The construction of decision tree rule paths is based on conditional probability, where this conditional probability distribution is defined over partitions of the feature space.[4] The path of a decision tree corresponds to a partition unit. During the classification process of the decision tree, instances belonging to that node are forcibly allocated to the category with a higher conditional probability.

2.2.2. CART (Classification And Regression Trees) Decision Tree Learning Process

During the process of decision tree learning, we induce a set of classification rules from the training dataset. The number of these decision rules can range from zero to multiple. In such cases, it becomes imperative to choose decision tree rules that exhibit minimal contradiction with the dataset while maintaining robust generalization capabilities.[5]

In our scenario, we use high-potassium glass and lead-barium glass as the leaf nodes and calculate the Gini coefficient:

$$Gini(D) = 1 - \sum_{k=1}^{K} (\frac{|C_k|}{|D|})^2$$

Algorithm Principles:

The learning process of the decision tree involves constructing a root node and placing all training data sets into it. The optimal feature is selected at each step to divide the training data into subsets, achieving the best classification under the current conditions.[6]

As the decision tree is a supervised learning technique, the testing set comprises the last five data entries from the EXCEL spreadsheet. The training set is composed of the remaining data. Commencing from the root node and based on the training data set, a binary decision tree is recursively built by performing operations on each node.

For a given node's training data set, denoted as D, the Gini coefficient of the existing features with respect to the dataset is computed. During this calculation, each possible value a of every feature A is used to partition the data into D^1 and D^2 subsets, followed by the calculation of the Gini coefficient for A=a. From all possible features A and potential split points a, the feature with the smallest Gini coefficient, along with its corresponding split point, is selected as the optimal feature and split point. Using this optimal feature and split point, two child nodes are generated from the current node, and the training data set D is distributed to these child nodes based on the feature conditions. This recursive process continues until the stopping criteria are met, resulting in the creation of the CART classification decision tree.[7]

2.2.3. Establishing the CART (Classification And Regression Trees) Decision Tree Analysis Model

Based on the aforementioned study and analysis, this paper employed MATLAB programming to engage in the analysis, utilizing the selected variables as presented in Table 3:

Table 3. List of Symbols.

Symbols	Explanation of Symbols
X_1	Decoration or Ornamentation
X_2	Color
X_3	SiO_2
X_4	Na_2O
X_5	K_2O
X_6	CaO
X_7	Fe_2O_3
X_8	BaO
X_9	P_2O_5
X_{10}	SrO
X_{11}	SnO_{2}
K	The Number of Cluster Centers

The obtained decision tree model is as presented in Figure 1:

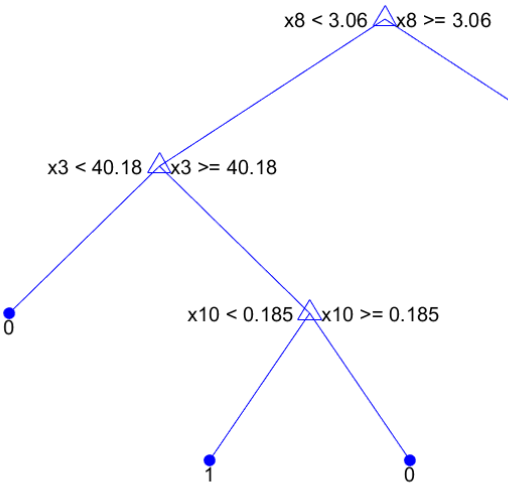

Figure 1. CART (Classification And Regression Trees) Decision Tree Model.

In summary, the classification patterns for high-potassium glass and lead-barium glass are as presented in Table 4:

Table 4. Classification Patterns of High-Potassium Glass and Lead-Barium Glass.

Lead-Barium Glass	(1)X_8(BaO)>=3.06
	(2)X_8(BaO)<=3.06 and X_3 (SiO_2)<=40.18
	(3)X_8 (BaO) <=3.06、X_3 (SiO_2)>= 40.18 and X_{10} (SrO)>=0.185
High-Potassium Glass	X_8 (BaO)<=3.06、X_3 (SiO_2)>=40.18 and X_{10} (SrO)<=0.185

Based on comprehensive chart data analysis, we have extensively investigated the classification patterns of lead-barium glass and high-potassium glass, providing a robust foundation for artifact classification and identification. For lead-barium glass, we have identified three distinct classification criteria as follows: (1) X_8(BaO)>=3.06, (2) X_8(BaO)<=3.06 and X_3 (SiO2)<=40.18, (3)X_8 (BaO) <=3.06、X_3 (SiO2)>= 40.18 and X_{10} (SrO)>=0.185.

In the case of high-potassium glass, the classification criteria are: X_8 (BaO)<=3.06、 X_3 (SiO2)>=40.18 and X_{10} (SrO)<=0.185.

These precise classification rules not only contribute to artifact identification but also offer substantial support for further research endeavors.

Table 5. Classification Pattern Model Evaluation Results for High-Potassium Glass and Lead-Barium Glass.

	Accuracy	Recall	Precision	F1
Training Set	1	1	1	1
Cross-Validation Set	0.883	0.883	0.945	0.888

In the decision tree classification model, we conducted a comprehensive assessment of the testing data, and the results are presented in Table 5. The model exhibits outstanding performance on the training set, achieving 100% accuracy, recall, precision, and F1 score. This remarkable performance indicates that the model excels in capturing the nuances of different glass types within the training set. On the cross-validation set, although there is a slight decrease in accuracy, it still maintains a high level of 88.3%. The recall stands at 88.3%, precision at 94.5%, and F1 score at 88.8% on the cross-validation set. These metrics underscore the model's robustness and strong classification performance across distinct datasets.

3. Analysis of Subclass Chemical Composition

Building upon the decision tree classification outcomes for high-potassium glass and lead-barium glass, we developed a K-means clustering model to facilitate subcategory division. The K-means model is primarily challenged by two aspects:[8]

(1) Determining the Initial Number of Cluster Centers (K)

(2) Selection of Initial Centroid Positions

Due to the random nature of initializing cluster centers in the MATLAB K-means function and the presence of fourteen dimensions in this context, directly determining the initial cluster center positions is not feasible. Utilizing a traversal algorithm, we computed all possible silhouette coefficients associated with different initial cluster center positions for each K value. The results are presented in Figure 2, illustrating average silhouette coefficients concerning distinct K values and various initial cluster center positions:

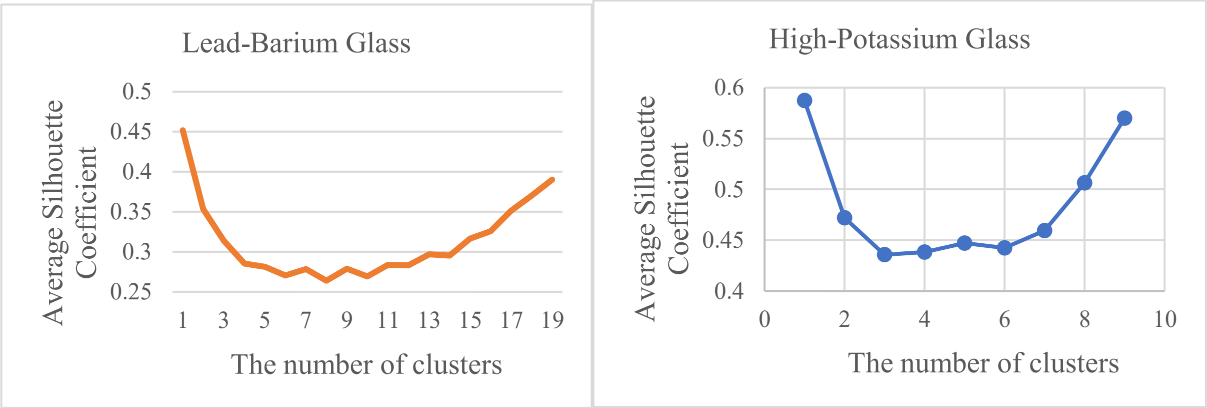

Figure 2. Average Silhouette Coefficient Graph for K-means Clustering Models of High-Potassium Glass and Lead-Barium Glass.

From the aforementioned two graphs, it becomes apparent that the average silhouette coefficient tends to approach 1 as the value of K approaches either 1 or the dimension of the samples. Nevertheless, considering the potential bias when adopting extreme K values [9], the machine learning outcomes might deviate significantly from actual results. Hence, we determined the initial cluster center positions and relevant model parameters by selecting the maximum average silhouette value for the same category of artifacts under identical K values. This approach ensures a more accurate final outcome.

3.1. Lead-Barium Glass Clustering Model

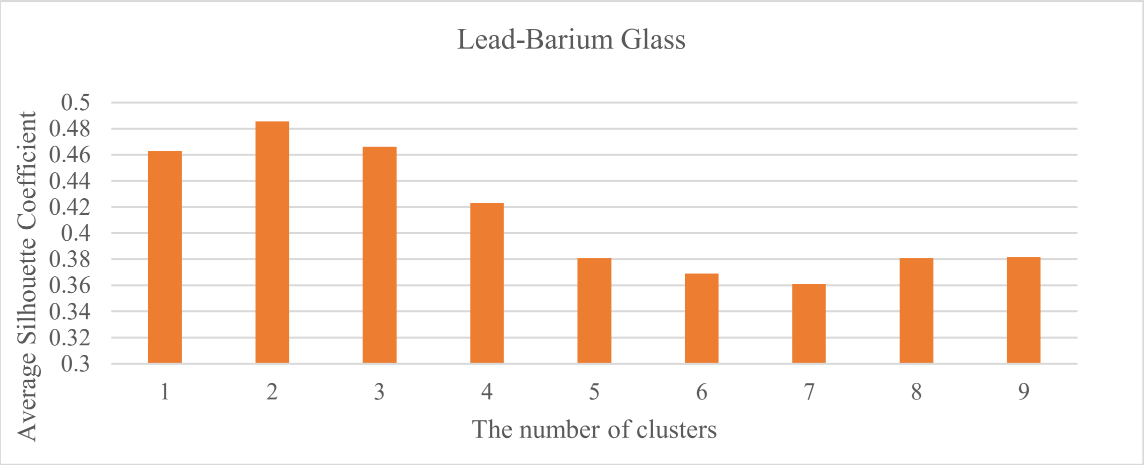

Figure 3. Maximum Silhouette Coefficient Graph for Lead-Barium Glass.

As inferred from the schematic representation of Lead-Barium Glass in Figure 3, it can be observed that when k=3, the silhouette coefficient reaches its maximum value of 0.4856. Under this specific scenario, the relevant model results fitted using MATLAB are depicted in Figure 4 as follows:

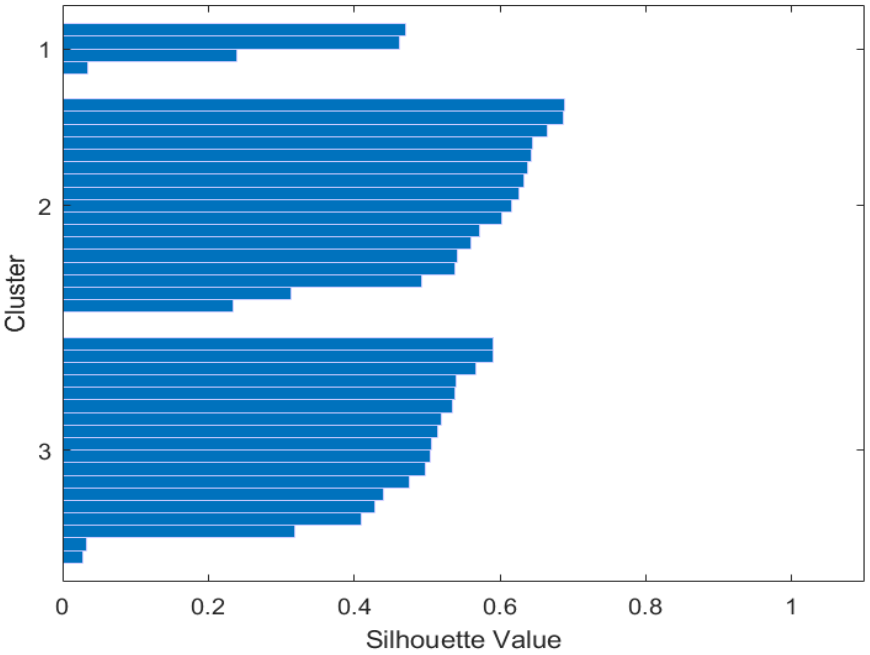

Figure 4. Silhouette Coefficient Graph for Lead-Barium Glass with Optimal Parameters at K=3.

As observed above, the silhouette coefficient values for each category of artifacts are all greater than 0, indicating favorable classification performance for these data.

3.2. High-Potassium Glass Clustering Model

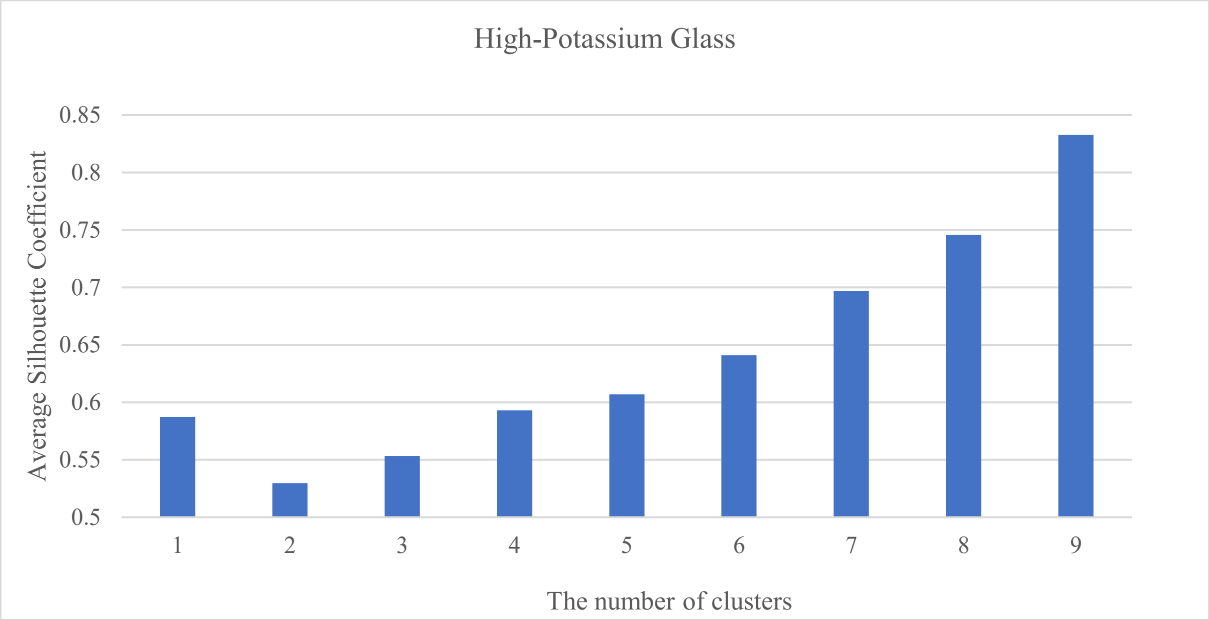

Figure 5. Maximum Silhouette Coefficient Graph for High-Potassium Glass.

As evident from Figure 5, within a reasonable range, when k values are 2, 3, and 4, there are instances where some artifacts exhibit silhouette coefficients less than zero. Consequently, there are disparities between certain artifacts and the clustering outcomes, leading to poorer classification outcomes. However, when K=5, as depicted in Figure 6, all silhouette coefficients are greater than 0, indicating a better classification scenario.[10] At this point, the maximum average silhouette value is measured at 0.5929.

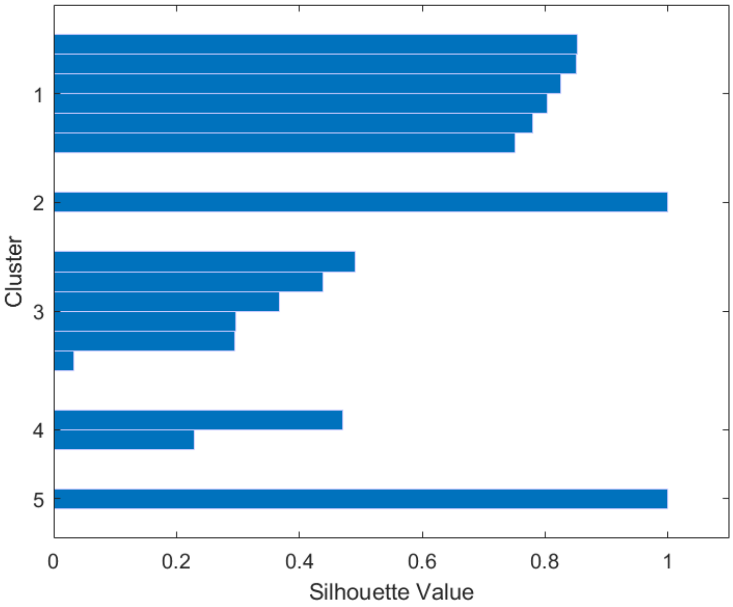

Figure 6. Silhouette Coefficient Graph for High-Potassium Glass with Optimal Parameters at K=5.

In conclusion, the subcategory division results, as illustrated in Figure 6 and Figure 7, have been achieved after determining the optimal initial cluster centroids and K value.

Table 6. Subcategory Classification Results for High-Potassium Glass.

Artifact IDs	Categories	Subcategory Types
01	High-Potassium Glass	3
03	High-Potassium Glass	4
04	High-Potassium Glass	3
05	High-Potassium Glass	3
06	High-Potassium Glass	2
07	High-Potassium Glass	1
09	High-Potassium Glass	1
10	High-Potassium Glass	1
12	High-Potassium Glass	1
13	High-Potassium Glass	3
14	High-Potassium Glass	3
16	High-Potassium Glass	3
18	High-Potassium Glass	5
21	High-Potassium Glass	4
22	High-Potassium Glass	1
27	High-Potassium Glass	1

Table 7. Subcategory Classification Results for Lead-Barium Glass.

Artifact IDs	Categories	Subcategory Types
2	Lead-Barium Glass	3
8	Lead-Barium Glass	1
11	Lead-Barium Glass	3
19	Lead-Barium Glass	3
20	Lead-Barium Glass	1
23	Lead-Barium Glass	2
24	Lead-Barium Glass	1
25	Lead-Barium Glass	2
26	Lead-Barium Glass	1
28	Lead-Barium Glass	2
29	Lead-Barium Glass	2
30	Lead-Barium Glass	3
31	Lead-Barium Glass	2
32	Lead-Barium Glass	2
33	Lead-Barium Glass	2
34	Lead-Barium Glass	3
35	Lead-Barium Glass	2
36	Lead-Barium Glass	3
37	Lead-Barium Glass	2
38	Lead-Barium Glass	3
39	Lead-Barium Glass	3
40	Lead-Barium Glass	3

Table 7. (continued).

41	Lead-Barium Glass	3
42	Lead-Barium Glass	2
43	Lead-Barium Glass	3
44	Lead-Barium Glass	2
45	Lead-Barium Glass	2
46	Lead-Barium Glass	2
47	Lead-Barium Glass	2
48	Lead-Barium Glass	2
49	Lead-Barium Glass	3
50	Lead-Barium Glass	3
51	Lead-Barium Glass	3
52	Lead-Barium Glass	3
53	Lead-Barium Glass	2
54	Lead-Barium Glass	3
55	Lead-Barium Glass	2
56	Lead-Barium Glass	3
57	Lead-Barium Glass	3

Based on the information provided in the two tables, we can derive valuable insights regarding the categorization of artifacts into different subcategory types for both "High-Potassium Glass" and "Lead-Barium Glass" categories. The subcategory types assigned to each artifact ID offer significant context for understanding the attributes and characteristics associated with these glass types.

For the "High-Potassium Glass" category, the subcategory types span a range from 1 to 5, indicating the existence of multiple distinct subgroups within this category. This suggests a certain degree of variation or differentiation in the characteristics of high-potassium glass artifacts. The distribution of subcategory types across different artifact IDs provides a foundation for further analysis and interpretation.

Conversely, the "Lead-Barium Glass" category also exhibits a diversity of subcategory types, including 1, 2, and 3. This variability in subcategory types within the "Lead-Barium Glass" category might reflect distinct compositional and structural attributes present in these artifacts. The distribution of subcategory types across different artifact IDs helps in understanding the underlying patterns and differences among lead-barium glass artifacts.

In both categories, the allocation of subcategory types allows for a more nuanced and detailed classification of artifacts based on specific features and characteristics. This categorization provides researchers and experts with a systematic framework to comprehend the complexities and diversities within the studied glass artifacts.

Overall, the subcategory types assigned to the artifact IDs provide a structured approach for organizing and categorizing the artifacts, enhancing the depth of analysis and interpretation within the context of glass composition and typology. This classification scheme contributes to the scholarly understanding of the artifacts and their inherent variations, supporting further research and preservation efforts in the realm of cultural heritage.

4. Conclusion

Through meticulous data preprocessing and a multi-step analysis, this study successfully revealed the classification patterns of high-potassium glass and lead-barium glass. In the analytical process, we initiated with cluster analysis, conducting a differential analysis of the characteristic fields of these glass types to obtain preliminary classification outcomes. Subsequently, we delved deeper into the classification patterns using machine learning decision tree models, effectively segmenting the dataset into training and testing sets and successfully constructing a classification tree model capable of categorizing various types of artifacts.

In the application of the K-means clustering model, we surmounted two critical challenges: the determination of the optimal K value for initial centroids and the selection of initial centroid positions. Employing an exhaustive search algorithm, we explored different K values and calculated average silhouette coefficients for various initial centroid positions under different K values. Through graphical analyses, we identified the optimal K value and initial centroid positions, ensuring the rationality and stability of the clustering model.

In the clustering analysis of lead-barium glass, we identified that the maximum silhouette coefficient occurred at K=3, indicating that this model is best suited for categorizing artifacts at this K value. In the clustering analysis of high-potassium glass, we found that when K=5, all silhouette coefficients were greater than 0, signifying effective classification of high-potassium glass artifacts by this model.

In conclusion, this study, through comprehensive analysis and the application of multiple models, successfully uncovered the classification patterns of high-potassium glass and lead-barium glass. Utilizing decision trees and K-means models, we achieved accurate artifact classification and provided an effective approach for the subdivision of similar artifacts. This not only supports the research and application of high-potassium glass and lead-barium glass but also offers valuable insights for decision-making in related fields.

Acknowledgments

This work was supported by College Student Innovation and Entrepreneurship Training Program:

easy-to-check cloud——multi-scenario integrated intelligent health detection platform from pathogenic microorganism to small molecule (No.202310057382)

References

[1] Chin T. The invention of the Silk Road, 1877[J]. Critical Inquiry, 2013, 40(1): 194-219.

[2] Zhou X, Lv H, Cui J, et al. Fluorite used in ancient Chinese glassmaking during the 10th to 12th centuries: Evidence from glass products excavated in the capital city site of the Liao dynasty[J]. Archaeometry, 2022, 64(5): 1138-1147.

[3] Steinberg D, Colla P. CART: classification and regression trees[J]. The top ten algorithms in data mining, 2009, 9: 179.

[4] Timofeev R. Classification and regression trees (CART) theory and applications[J]. Humboldt University, Berlin, 2004, 54.

[5] Lewis R J. An introduction to classification and regression tree (CART) analysis[C]//Annual meeting of the society for academic emergency medicine in San Francisco, California. San Francisco, CA, USA: Department of Emergency Medicine Harbor-UCLA Medical Center Torrance, 2000, 14.

[6] Daniya T, Geetha M, Kumar K S. Classification and regression trees with gini index[J]. Advances in Mathematics: Scientific Journal, 2020, 9(10): 8237-8247.

[7] Bittencourt H R, Clarke R T. Feature selection by using classification and regression trees (CART)[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2004.

[8] Pham D T, Dimov S S, Nguyen C D. Selection of K in K-means clustering[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2005, 219(1): 103-119.

[9] Hamerly G, Elkan C. Learning the k in k-means[J]. Advances in neural information processing systems, 2003, 16.

[10] Pham D T, Dimov S S, Nguyen C D. Selection of K in K-means clustering[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2005, 219(1): 103-119.