

对古代玻璃制品成分的分析与鉴别

摘 要

本文分析与鉴别古代玻璃制品的成分，探究不同分类的玻璃和其所含的不同化学成分之间的关系，对其类别的正确判断、不同成分的玻璃选择合适的保护方法，对古代玻璃文物的考古研究有着极其重要的意义。

针对问题一，首先整合附件表单 1 中数据，通过**卡方检验**，计算玻璃表面是否风化与其玻璃类型、纹饰和颜色间的卡方值、自由度、**P-value** 等，得到它们之间的相关关系：玻璃文物表面风化与纹饰和颜色不会表现出显著性差异，与类型呈现出显著性差异。其次由于表单 2 中化学成分为定量变量，所以对玻璃类型有无风化与化学成分分析，采用**单因素方差分析**，得到文物样品表面有无风化化学成分含量的统计规律：对二氧化硅(SiO_2)，氧化钾(K_2O) 含量，无风化>风化；高钾>铅钡等结果。最后应用风化点检测数据，根据第二小问的统计规律预测其风化前的化学成分含量。

针对问题二，求解化学成分与分类的关系，采用 **K-Means 聚类与层次聚类**，通过对聚类中心点的分析，得出：区分高钾玻璃与铅钡玻璃的主要化学成分为氧化钾(K_2O)，氧化铅(PbO)，氧化钡(BaO)。在高钾玻璃中含量差距较大为氧化钾(K_2O)，氧化钙(CaO)，氧化铝(Al_2O_3)，铅钡玻璃中含量差距较大为氧化铅(PbO)，氧化钡(BaO)。

针对问题三，首先需要对附件表单 3 中的未知类别玻璃文物根据其化学成分鉴定类别，并对分类结果进行**敏感性分析**。通过对附件表单 2 数据进行训练，再将附件表单 3 未知数据代入进行预测分类。对于未知类别鉴定，使用**决策树**进行分析预测，通过氧化铝的阈值 5.46，得到分类结果为：高钾玻璃,铅钡玻璃,铅钡玻璃,铅钡玻璃,铅钡玻璃,高钾玻璃,高钾玻璃,铅钡玻璃。其次，分类模型的评价指标均为 1，在排除模型过拟合后，发现分类结果准确。

针对问题四，本题探究不同类别玻璃的化学成分之间的关系和关联关系的差异性，通过表单 2 中的数据，应用 **Pearson 相关性**分析多变量间的相关关系，并通过**偏相关分析**得到任意两化学成分之间的相关关系：高钾玻璃中氧化铝的增加，二氧化硅(SiO_2)、氧化锡(SnO_2)成负相关，其他成分成正相关，其中与五氧化二磷(P_2O_5)相关性最大，相关性为 0.6834；铅钡玻璃中随着氧化铝的增加，氧化铜(CuO)、氧化铅(PbO)等成分和二氧化硫(SO_2)成负相关，其他成分如二氧化硅(SiO_2)、氧化钠(Na_2O)等成分能是成正相关，其中与五氧化二磷(P_2O_5)相关性最大，相关性为 -0.0743。最后，通过偏相关系数的比较，即可得到各化学成分间的差异性。

关键词：卡方检验；层次聚类；K-Means 聚类；决策树；Pearson 相关性；

一、问题重述

我国古代遗留下了大量玻璃制品，它们作为历史的物质遗存，是我国悠久历史文化的见证和重要载体，但留存至今都遭受了不同程度的破坏，其中风化是极其重要的一个因素。因此分析与鉴别古代玻璃制品的成分，探究不同类别和化学成分之间的规律和联系，这对不同成分的玻璃选择合适的保护方法，正确判断其类别，有着极其重要的意义。

基于以上背景，通过附件中已知文物的分类信息和相应的主要成分所占比例，我们需要建立数学模型解决以下问题：

问题一：对玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系进行分析；结合玻璃的类型，分析文物样品表面有无风化化学成分含量的统计规律，并根据风化点检测数据，预测其风化前的化学成分含量。

问题二：依据附件数据分析高钾玻璃、铅钡玻璃的分类规律；对于每个类别选择合适的化学成分对其进行亚类划分，给出具体的划分方法及划分结果，并对分类结果的合理性和敏感性进行分析。

问题三：对附件表单 3 中未知类别玻璃文物的化学成分进行分析，鉴别其所属类型，并对分类结果的敏感性进行分析。

问题四：针对不同类别的玻璃文物样品，分析其化学成分之间的关联关系，并比较不同类别之间的化学成分关联关系的差异性。

二、问题分析

2.1 问题一的分析

根据题意可将问题一分为三个小问。第一小问对玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系进行分析，根据表单 1 中的数据可以得到：文物纹饰、类型、颜色、表面风化四个变量均为定类变量，因此对定类变量之间的相关性进行分析采用卡方检验；第二小问分析文物样品表面有无风化化学成分含量的统计规律，利用方差分析去研究玻璃与有无风化对于化学成分的相关性，并得出玻璃类型与有无风化和化学成分的关系；第三小问根据风化点检测数据，结合第二小问的统计规律，预测其风化前的化学成分含量。

2.2 问题二的分析

根据题意可将问题二分为三个小问。第一小问分析高钾玻璃、铅钡玻璃的分类规律，使用 K-Means 算法对高钾与铅钡类型进行分类，并得出化学成分与玻璃类型的相关性；第二小问对于每个类别选择合适的化学成分对其进行亚类划分，我们使用层次聚类决定聚类个数，再使用 K-Means 算法分出亚类；第三小问对分类算法与结果的合理性和敏感性进行分析。

2.3 问题三的分析

根据题意可将问题三分为两个小问，第一小问对附件表单 3 中的未知类别玻璃文物根据其化学成分鉴定类别；第二小问对分类结果进行敏感性分析。针对第一个问题，本文对附件表单 1、2 里面的数据预处理之后，利用处理后有标签的附件表单 2 数据使用决策树算法进行训练，再将附件表单 3 未知数据代入进行分类。其次，使用分类模型的评价标准准确率（Accuracy）、精确率（Precision）、召回率（Recall）分析具体情况。

2.4 问题四的分析

根据题意可将问题三分为两个小问。第一小问针对不同类别的玻璃文物样品，分析其化学成分之间的关联关系，即探究不同类玻璃化学成分间的相关程度，因表单 2 中的数据为连续变量，因此选择 Pearson 相关性分析多变量间的关系，使用偏相关分析更进一步分析两两变量之间的相关关系，并通过偏相关系数判断化学成分关联关系的差异性。

三、模型假设与约定

1. 假设所给数据均真实可靠。
2. 假设所有玻璃文物均处于相同的风化环境。
3. 假设未知玻璃的分类仅以附件数据作为影响因素。

四、符号说明与名词定义

变量符号	含义说明
χ^2	卡方，体现数据的相关性
A	观测值，为观测到的频次
T	期望值，是期望得到的频次
k	组别数量,样本组别数

n	样本量，样本总量
x	组别内平均值
x^i	化学成分训练集， $x^i \in R^n$
μ^j	聚类质心点， $\mu^j \in R^n$

五、模型的建立与求解

5.1 问题一的建模与求解

根据题意可将问题分为三个小问。第一小问对玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系进行分析，纹饰、类型、颜色和表面风化四个变量均为定类变量，因此对定类变量之间的相关性进行分析采用卡方检验。第二小问根据玻璃的类型，分析文物样品表面有无风化化学成分含量的统计规律，化学成分为定量变量，因此对玻璃类型有无风化与化学成分分析，采用单因素方差分析。第三小问根据风化点检测数据，预测其风化前的化学成分含量，根据第二小问的统计规律进行预测。

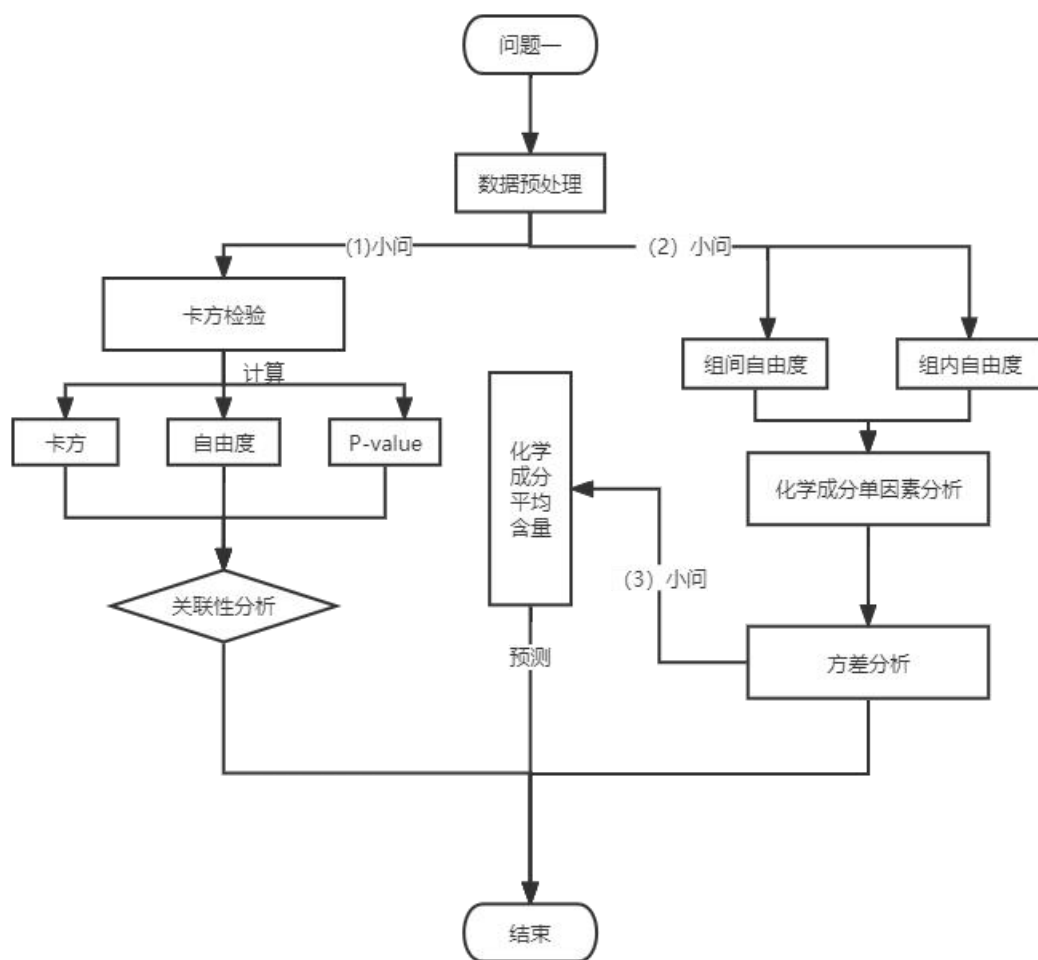


图 5- 1 问题一解题流程图

5.1.1 数据预处理

对附件的表单 1，其中 19、40、48、58 号文物缺失颜色信息，由于颜色是非连续的数据信息，因此在问题一的解决中，将去掉这几个样本数据。

对附件的表单 2，根据题目描述可知，成分比例累加和介于 85%~105%之间的数据视为有效数据。而编号 15 和编号 17 的总成分小于 85%，属于无效数据，因此在后续引用数据时将剔除这两组数据。

此外，表单 2 空白处表示未检测到该成分，即该成分的含量为 0，所以对于表单 2 中所用的控制用 0 进行填充。

5.1.2 表面风化与类型、纹饰、颜色的关系分析

1. 数据统计

由于附件表单 1 中纹饰、类型、颜色和表面风化四个变量均为定类变量，因

此对定类变量之间的相关性进行分析采用卡方检验。

通过统计附件中表单一的数据得到表 5-1:

表 5- 1：表面风化与其玻璃类型、纹饰和颜色数量统计

名称	纹饰			总计	类型		总计	颜色								总计
	A	B	C		高钾	铅钡		蓝绿	浅蓝	紫	深绿	深蓝	浅绿	黑	绿	
表面风化	11	6	17	34	6	28	34	9	12	2	4	0	1	2	0	30
表面未风化	11	0	13	24	12	12	24	6	8	2	3	2	2	0	1	24
总计	22	6	30	58	18	40	58	15	20	4	7	2	3	2	1	54

2. 卡方的计算:

$$\chi^2 = \sum \frac{(A-T)^2}{T} \quad (5.1)$$

其中，A 表示观测值，为观测到的频次;T 表示期望值，是期望得到的频次。

上述公式，分子代表了实际值与期望值的偏差，而分母则是标准化过程。因此卡方的值越小，表示观测值与期望值(理论值) 越接近，说明两个变量之间越符合卡方分布。而卡方分布的前提是变量之间相互独立，因此卡方值越小，越表示两个变量之间独立。

通过计算的到表 5-2:

表 5- 2：卡方检验分析结果

名称		表面风化	表面未风化	χ^2	P
纹饰	A	11	11	4.90	0.084
	B	6	0		
	C	17	13		
类型	高钾	6	12	6.880	0.009**
	铅钡	28	12		
	蓝绿	9	6		
颜色	浅蓝	12	8	6.287	0.507
	紫	2	2		

深绿	4	3
深蓝	0	2
浅绿	1	2
黑	2	0
绿	0	1

注:***、**、*分别代表 1%、5%、10%的显著性水平

3. 自由度的计算:

$$K = (R - 1) - (C - 1) \quad (5.2)$$

其中 R 为单元格的行数, C 为单元格的列数, 即(行数-1)(列数-1), 所以表面风化与纹饰的自由度为 2, 表面风化与类型的自由度为 1, 表面风化与颜色的自由度为 7。

4. 关系分析:

表面是否风化与玻璃纹饰不会表现出显著性($p>0.05$), 意味着不同纹饰对于是否风化均表现出一致性, 并没有差异性, 即玻璃文物表面风化与纹饰不会表现出显著性差异。

表面是否风化与玻璃类型呈现出显著性($p<0.05$), 意味着不同类型对于是否风化均表现出差异性, 即玻璃文物表面风化与类型呈现出显著性差异。

表面是否风化与玻璃颜色不会表现出显著性($p>0.05$), 意味着不同颜色对于是否风化均表现出一致性, 并没有差异性, 即玻璃文物表面风化与颜色不会表现出显著性差异。

5.1.3 有无风化化学成分含量的统计规律

根据玻璃类型与有无风化情况不同, 结合表单 1 与表单 2, 将样本类型分为铅钡无风化, 铅钡风化, 高钾无风化, 高钾风化(见附录)。对化学成分进行单因素方差分析。

其中组间自由度与组内自由度:

$$df1 = k - 1 \quad (5.3)$$

$$df2 = n - k \quad (5.4)$$

F 的计算:

$$F = \frac{\sum n_i (x_i - \bar{x})^2 / df2}{\sum \sum (x_{ij} - x_i)^2 / df1} \quad (5.5)$$

其中 k 表示组别数量, n 表示样本量, \bar{x} 表示平均值, \bar{x} 表示总平均值, p 值

是结合 F 值，df1 和 df2 计算得到。

通过表单 1 与表单 2，根据单因素方差分析可得表 5-3

表 5-3：玻璃类型有无风化化学成分含量的方差分析表

	类型表面风化(平均值±标准差)				F	p
	铅钡无风化 (n=13)	铅钡风化 (n=27)	高钾无风化 (n=12)	高钾风化(n=7)		
二氧化硅 (SiO ₂)	53.44±14.59	32.75±21.97	67.98±8.76	59.90±35.08	9.782	0.000**
氧化钠 (Na ₂ O)	0.77±1.54	0.21±0.55	0.69±1.29	0.00±0.00	1.668	0.184
氧化钾 (K ₂ O)	0.26±0.40	0.14±0.22	9.33±3.92	3.58±4.33	50.453	0.000**
氧化钙 (CaO)	1.23±1.46	2.49±1.82	5.33±3.09	1.68±1.97	9.145	0.000**
氧化镁 (MgO)	0.49±0.55	0.64±0.69	1.08±0.68	0.75±0.94	1.647	0.189
氧化铝 (Al ₂ O ₃)	3.19±1.39	3.17±2.61	6.62±2.49	4.29±4.37	5.256	0.003**
氧化铁 (Fe ₂ O ₃)	0.93±1.45	0.66±0.77	1.93±1.67	1.27±2.27	2.519	0.067
氧化铜 (CuO)	1.56±2.49	2.19±2.55	2.45±1.66	3.29±3.33	0.774	0.513
氧化铅 (PbO)	23.59±9.09	38.21±18.88	0.41±0.59	8.81±14.90	21.731	0.000**
氧化钡 (BaO)	10.50±6.95	10.45±9.37	0.60±0.98	9.17±14.87	3.952	0.013*
五氧化二磷(P ₂ O ₅)	0.90±1.57	4.47±4.34	1.40±1.43	3.16±2.58	4.597	0.006**
氧化锶 (SrO)	0.30±0.31	0.38±0.32	0.04±0.05	0.17±0.20	4.741	0.005**
氧化锡 (SnO ₂)	0.06±0.16	0.07±0.26	0.20±0.68	0.34±0.89	0.792	0.504
二氧化硫 (SO ₂)	0.28±1.02	0.66±3.08	0.10±0.19	2.52±5.60	1.216	0.313

类型表面风化(平均值±标准差)				F	p
铅钡无风化 (n=13)	铅钡风化 (n=27)	高钾无风化 (n=12)	高钾风化(n=7)		

* $p < 0.05$ ** $p < 0.01$

从上表可知,利用方差分析去研究玻璃类型与有无风化对于化学成分的差异性,并可以看出:不同类型表面风化样本对于氧化钠(Na_2O),氧化镁(MgO),氧化铁(Fe_2O_3),氧化铜(CuO),氧化锡(SnO_2),二氧化硫(SO_2)共6项不会表现出显著性($p > 0.05$),意味着玻璃类型与有无风化样本对于氧化钠(Na_2O),氧化镁(MgO),氧化铁(Fe_2O_3),氧化铜(CuO),氧化锡(SnO_2),二氧化硫(SO_2)全部均表现出一致性,并没有差异性。另外玻璃类型与有无风化样本对于二氧化硅(SiO_2),氧化钾(K_2O),氧化钙(CaO),氧化铝(Al_2O_3),氧化铅(PbO),氧化钡(BaO),五氧化二磷(P_2O_5),氧化锶(SrO)共8项呈现出显著性($p < 0.05$),意味着不同的玻璃类型与有无风化样本对于二氧化硅(SiO_2),氧化钾(K_2O),氧化钙(CaO),氧化铝(Al_2O_3),氧化铅(PbO),氧化钡(BaO),五氧化二磷(P_2O_5),氧化锶(SrO)有着差异性。具体分析可知:

对于二氧化硅(SiO_2),氧化钾(K_2O)含量,无风化>风化;高钾>铅钡。

对于氧化钙(CaO)含量,铅钡风化>铅钡无风化;高钾无风化>高钾风化。

对于氧化铝(Al_2O_3)含量,高钾无风化>高钾风化。

对于氧化铅(PbO)含量,风化>无风化;高钾>铅钡。

对于氧化钡(BaO),五氧化二磷(P_2O_5),氧化锶(SrO)含量,高钾风化>高钾无风化。

5.1.4 风化前化学成分含量的预测

根据问题一第(2)问的结果,可得不同玻璃类型风化前后的化学成分平均含量。鉴于风化反应在长时间的过程中,反应物资充足,我们认为反应时化学成分按比例变化,即风化前化学成分为:

$$y_j = \frac{x_i}{y_i} \times x_j \quad (5.6)$$

x_i 指风化前化学成分平均含量, y_i 指风化后化学成分平均含量, x_j 指风化后化学成分, y_j 指预测风化前化学成分。

根据此公式,可根据风化点检测数据,预测其风化前的部分化学成分含量见表(详细见附录)。

表 5- 4 预测风化前的化学成分含量

类型	高钾	高钾	高钾	高钾	高钾	高钾	高钾
文物采 样点	06 部位 1	06 部位 2	07	08	08 严重 风化点	09	18
二氧化 硅 (SiO ₂)	62.15	49.93	91.80	53.86	19.82	93.63	72.93
氧化钠 (Na ₂ O)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
氧化钾 (K ₂ O)	15.55	14.72	0.00	0.00	0.00	1.34	19.86
氧化钙 (CaO)	0.00	12.62	2.96	11.06	38.33	1.71	0.00
氧化镁 (MgO)	2.31	1.83	0.00	0.00	0.00	0.00	1.78
氧化铝 (Al ₂ O ₃)	13.94	11.42	2.67	4.88	6.50	1.77	3.81
氧化铁 (Fe ₂ O ₃)	2.93	6.73	0.23	0.00	0.00	0.42	0.00
氧化铜 (CuO)	1.51	1.20	2.11	18.29	8.87	1.00	0.00
氧化铅 (PbO)	0.01	0.01	0.00	3.16	5.74	0.00	0.00
氧化钡 (BaO)	0.07	0.05	0.00	4.80	7.57	0.00	0.00
五氧化 二磷 (P ₂ O ₅)	1.50	1.47	0.24	3.75	12.69	0.13	0.49
氧化锶 (SrO)	0.02	0.02	0.00	0.21	0.49	0.00	0.01
氧化锡 (SnO ₂)	0.00	0.00	0.00	0.00	0.00	0.00	1.11
二氧化 硫 (S O ₂)	0.00	0.00	0.00	0.25	2.30	0.00	0.00

5.2 问题二的建模与求解

第一小问依据附件数据分析高钾玻璃、铅钡玻璃的分类规律，我们采用 K-Means 算法聚类分析寻找规律，第二小问对于每个类别选择合适的化学成分对其进行亚类划分，给出具体的划分方法及划分结果，对此我们先行使用层次聚类分析，再使用 K-Means 算法聚类分析寻找规律；第三小问对分类结果的合理性和敏感性进行分析，对此我们使用分类模型评价指标，准确率、精确率、召回率进行分析。

5.2.1 数据预处理

对表单 1、表单 2 的数据整理，分为风化与无风化类型。

5.2.2 分析高钾玻璃、铅钡玻璃的分类规律

Step 1: 首先将表单 1 中玻璃的纹饰和颜色作为分类规律的标准，并进行可视化处理，分析其可行性。

通过 origin 作图，得到图 5-1 和图 5-2

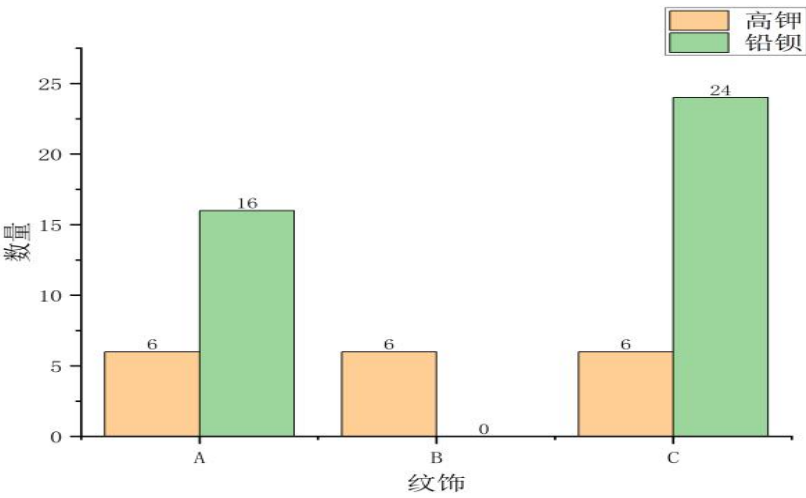


图 5- 2 高钾和铅钡玻璃的不同纹饰数量情况

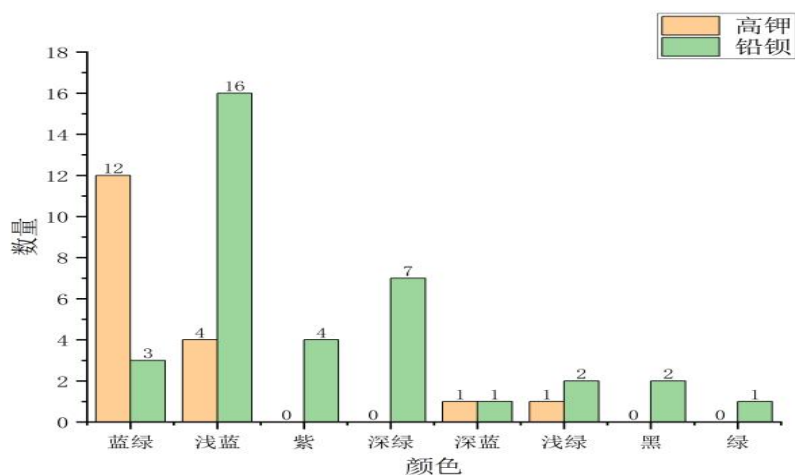


图 5-3 高钾和铅钡玻璃的不同颜色数量情况

由图 5-1 可以看出，高钾玻璃存在 A、B、C 三种纹饰，而铅钡玻璃只有 A 和 C 两种纹饰

由图 5-2 可以看出高钾玻璃只有蓝绿、浅蓝、深蓝、浅绿四种颜色，并且蓝绿所占比例较大；铅钡玻璃颜色分布更广，浅蓝色和深绿色最多，其余颜色数量分布较均匀

Step 2: 其次通过表单 2 中数据进行可视化，得到高钾和铅钡的成分占比图。如图 5-4 所示。

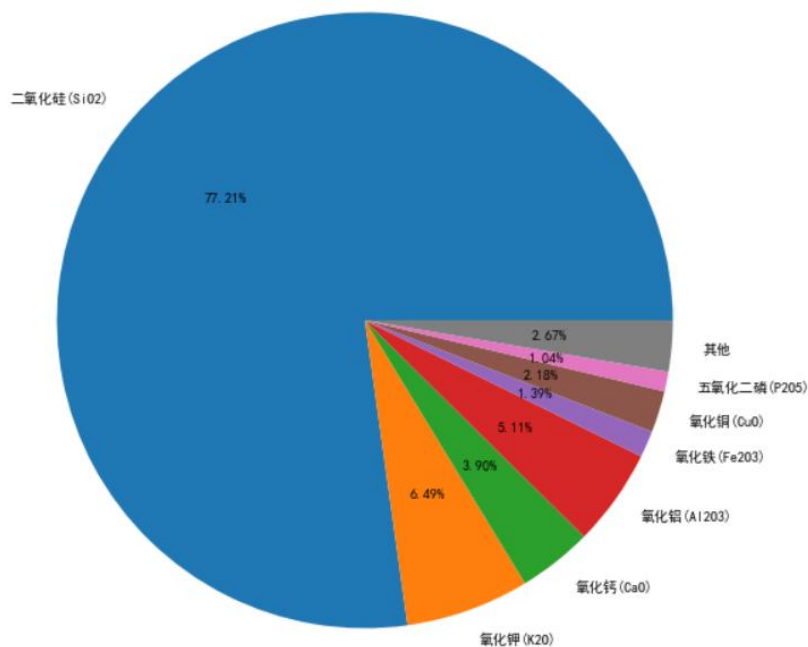


图 5-4 高钾玻璃成分占比图

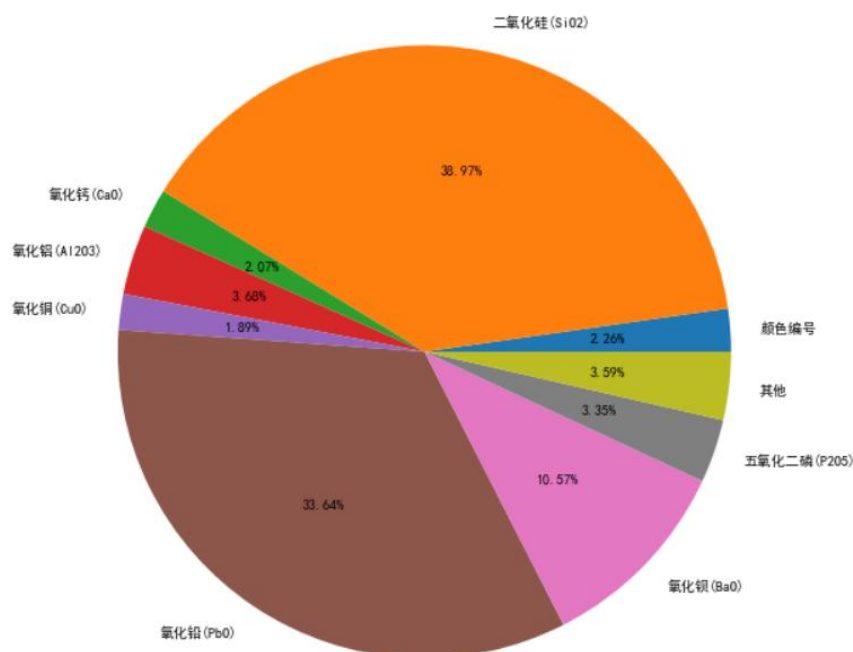


图 5-5 铅钡玻璃成分占比图

由图 5-4 和图 5-5 可以看出，高钾玻璃中的二氧化硅的成分占比远高于其他成分，铅钡玻璃二氧化硅和氧化铅远高于其他化学成分。

Step 3: K-Means 算法应用：

1、 随机选取 k 个聚类质心点 (cluster centroids) 为 $\mu^j \in R^n$ 。

2、 重复下面过程直到收敛。

对于每一个样例 i ，计算其应该属于的类

$$c^{(i)} = \arg \min_j \|x^{(i)} - \mu_j\|^2 \quad (5.7)$$

对于每一个类 j ，重新计算该类的质心

$$\mu_j = \frac{\sum_{i=1}^m 1\{c^j = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^j = j\}} \quad (5.8)$$

首先应用 SPSS 通过 K-Means 算法对风化与无风化的进行二分类，以化学成分作为分类变量，分类结果准确率仅有 76%，我们注意到玻璃的主要化学成分是二氧化硅 (SiO2)，在 K-Means 算法不应将二氧化硅 (SiO2) 作为分类变量，为此，我们将除二氧化硅 (SiO2) 外的 13 种化学成分作为分类变量，将得到无风化数据集聚类标注 (附录问题二) 与风化数据集聚类标注 (问题二) (SPSS 操作步骤见附录)。

对数据对比可得，K-Means 算法对高钾铅钡分类中，无风化准确率 100%，

风化准确率 94.12%，根据 K-Means 算法可得风化与无风化聚类中心点。

表 5-5 风化聚类中心点

聚类种类	1	2	方差
氧化钠 (Na ₂ O)	0.20	0.08	0.01
氧化钾 (K ₂ O)	0.08	2.68	3.39
氧化钙 (CaO)	2.58	1.72	0.37
氧化镁 (MgO)	0.56	0.92	0.06
氧化铝 (Al ₂ O ₃)	2.36	5.89	6.20
氧化铁 (Fe ₂ O ₃)	0.51	1.44	0.43
氧化铜 (CuO)	2.60	1.96	0.20
氧化铅 (PbO)	44.45	2.66	873.32
氧化钡 (BaO)	12.97	3.52	44.68
五氧化二磷 (P ₂ O ₅)	5.14	1.95	5.07
氧化锶 (SrO)	0.46	0.05	0.08
氧化锡 (SnO ₂)	0.02	0.37	0.06
二氧化硫 (SO ₂)	1.48	0.00	1.10

由表 5-5 分析可知对于风化玻璃，分类高钾玻璃与铅钡玻璃主要化学成分为氧化铅(PbO)，其次为氧化钾(K₂O)，氧化铝(Al₂O₃)，氧化钡(BaO)，五氧化二磷(P₂O₅)。

表 5-6 无风化聚类中心点

聚类种类	1	2	方差
氧化钠 (Na ₂ O)	0.70	0.77	0.00
氧化钾 (K ₂ O)	9.33	0.26	41.15
氧化钙 (CaO)	5.33	1.23	8.41
氧化镁 (MgO)	1.08	0.49	0.17
氧化铝 (Al ₂ O ₃)	6.62	3.19	5.87
氧化铁 (Fe ₂ O ₃)	1.93	0.93	0.50
氧化铜 (CuO)	2.45	1.56	0.40
氧化铅 (PbO)	0.41	23.59	268.71
氧化钡 (BaO)	0.60	10.50	49.01
五氧化二磷 (P ₂ O ₅)	1.40	0.90	0.12
氧化锶 (SrO)	0.04	0.30	0.03
氧化锡 (SnO ₂)	0.20	0.06	0.01

二氧化硫 (SO ₂)	0.10	0.28	0.02
-------------------------	------	------	------

由表 5-6 对于无风化玻璃，分类高钾玻璃与铅钡玻璃的主要化学成分为氧化钾(K₂O)，氧化铅(PbO)，氧化钡(BaO)。

5.2.3 亚类划分方法及划分结果

对于第二小问，我们先使用层次聚类进行分析。

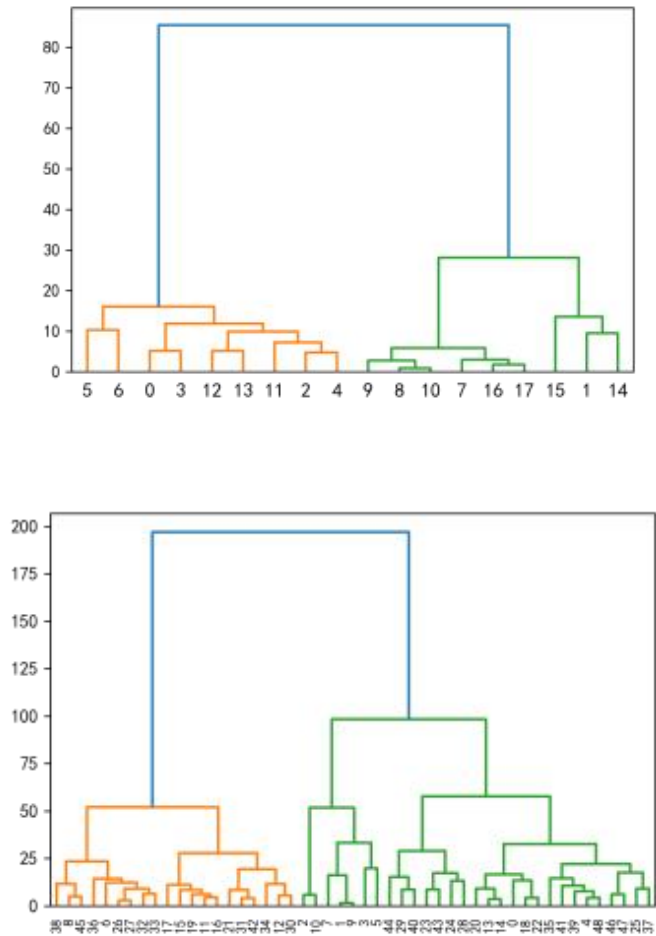


图 5-6 上图为高钾玻璃谱系图，下图为铅钡玻璃谱系图

由图 5-6 分析，我们选择 K-Means 算法聚类个数为 3。再行使用第一小问中的 K-Means 算法对其进行分类，将数据分为高钾与铅钡两种情况分别使用 K-Means 算法分类。可得到聚类个数 3 的分类情况聚类中心点。

表 5-7 高钾聚类中心点

聚类种类	1	2	3	方差
氧化钠 (Na ₂ O)	0.00	0.00	1.19	0.47
氧化钾 (K ₂ O)	2.82	7.53	11.76	20.00
氧化钙 (CaO)	1.75	2.71	7.41	9.18
氧化镁 (MgO)	0.74	1.86	0.93	0.36
氧化铝 (Al ₂ O ₃)	3.54	10.60	6.42	12.61
氧化铁 (Fe ₂ O ₃)	0.64	4.22	1.77	3.34
氧化铜 (CuO)	1.58	2.34	2.95	0.47
氧化铅 (PbO)	0.25	0.28	0.45	0.01
氧化钡 (BaO)	0.44	1.17	0.41	0.19
五氧化二磷 (P ₂ O ₅)	0.79	4.34	0.72	4.29
氧化锶 (SrO)	0.02	0.11	0.03	0.00
氧化锡 (SnO ₂)	0.52	0.00	0.00	0.09
二氧化硫 (SO ₂)	0.00	0.00	0.17	0.01

由表 5-7, 分析可知在高钾玻璃中氧化钾(K₂O), 氧化钙(CaO), 氧化铝(Al₂O₃)含量差距较大, 对此, 我们将高钾玻璃分为高钾低钾玻璃, 高钾高铝玻璃, 高钾高钙玻璃, 分析结果如表 5-8 所示。

表 5-8 高钾亚类

聚类类别	频数	百分比%
高钾低钾玻璃	9	45.0%
高钾高铝玻璃	4	20.0%
高钾高钙玻璃	7	35.0%
合计	20	100.0%

表 5-9 铅钡聚类中心点

聚类种类	1	2	3	方差
氧化钠 (Na ₂ O)	0.34	0.00	0.74	0.14
氧化钾 (K ₂ O)	0.13	0.07	0.27	0.01
氧化钙 (CaO)	2.83	1.68	0.77	1.06
氧化镁 (MgO)	0.68	0.00	0.59	0.14
氧化铝 (Al ₂ O ₃)	2.75	1.25	4.49	2.63

氧化铁 (Fe ₂ O ₃)	0.78	0.00	0.89	0.24
氧化铜 (CuO)	1.23	7.44	1.43	12.45
氧化铅 (PbO)	46.68	29.81	17.52	214.32
氧化钡 (BaO)	8.37	30.33	10.23	148.37
五氧化二磷 (P ₂ O ₅)	4.75	3.43	1.52	2.65
氧化锶 (SrO)	0.41	0.63	0.22	0.04
氧化锡 (SnO ₂)	0.06	0.00	0.12	0.00
二氧化硫 (SO ₂)	0.00	5.92	0.33	11.06

由表 5-9，分析可知在铅钡玻璃中氧化铅(PbO)，氧化钡(BaO)含量差距较大，对此，我们将铅钡玻璃分为铅钡高铅玻璃，铅钡高钡玻璃，铅钡低铅钡玻璃。分析结果如表 5-10 所示。

表 5-10 铅钡亚类

聚类类别	频数	百分比%
铅钡高铅玻璃	22	56.41%
铅钡高钡玻璃	6	15.385%
铅钡低铅钡玻璃	11	28.205%
合计	39	100.0%

5.2.4 模型分析

合理性分析：K-Means 算法要事先确定簇数 K，对此我们先行采用层次凝聚算法决定结果粗的数目，再使用 K-Means 算法对数据进行聚类分析，算法合理性良好。

敏感性分析：K-Means 算法对“噪声”和孤立点敏感，如出现化学成分与数据差别较大的数据，分类效果不佳。

5.3 问题三的建模与求解

5.3.1 问题三的分析思路

问题三需要根据附件表单 3 中的未知类别玻璃文物的化学成分鉴定其类别，并且对分类结果进行敏感性分析。

解决未知类别玻璃的分类问题，有两种思路。一是直接根据问题二中得到的高钾玻璃、铅钡玻璃分类规律，进行分类得到结果。但是由于 K-Means 聚类算法是一种无监督学习算法，在初始化聚类中心时是在最小值和最大值之间随机取

一个值作为其聚类中心，这样的随机取值会导致聚类中心可能选择的不好，最终对结果会产生很大的影响。由于本文样本类别区分度不太明显，按照 K-Means 初始化聚类中心，对结果的影响很大，因此问题三不适合使用问题二中的规律得到结果。二是通过有监督的决策树算法，先对已知类别数据进行训练搭建模型，最后将未知类别数据代入分类。使用决策树算法作为一个有监督的，并且更能挖掘出准确性高的分类规则的算法来说，决策树算法适合本题的使用。

本文对附件表单 1、2 里面的数据预处理之后，分为高钾玻璃和铅钡玻璃两种类别数据，再将其作为标签数据来训练决策树模型。但是由于当样本数量少并且样本特征多时，决策树容易出现过拟合情况，因此在训练模型之前，先进行维度规约即主成分分析来降低特征维度。由于本文的分类回归树划分不是很多时，不会对噪声数据产生过拟合，因此不需要通过剪枝来解决。最后将未知类别数据代入模型进行预测得到分类结果。

解决对分类结果的敏感性分析，需要从分类模型的评价指标：准确率（Accuracy）、精确率（Precision）、召回率（Recall）进行具体分析。

5.3.2 使用决策树算法预测分类结果

需要解决未知类型玻璃文物数据分类问题，并且已知两种玻璃文物的数据情况，因此在处理本问题时本文选择采用准确性高、易理解，方便处理分类问题的决策树算法进行求解。

决策树算法的根本目的在于使用层层推理来实现最终的分类。他的运行机制就是预测时，在树的内部节点处用某一属性值进行判断，根据判断结果决定进入哪个分支节点，直到到达叶节点处，得到分类结果。

本题采用 CART 分类决策树，使用基尼系数(GiniIndex)来选择划分属性。基尼系数代表了模型的不纯度，基尼系数越小，不纯度越低，特征越好。基尼系数的表示为：

$$\text{Gini}(D) = \sum_{i=1}^n P(X_i) * (1 - P(X_i)) = 1 - \sum_{i=1}^n P(X_i)^2 \quad (5.9)$$

其中, $P(X_i)$ 是分类 X_i 出现的概率, n 是分类的数目。Gini(D)反映了从数据集 D 中随机抽取两个样本，其类别标记不一致的概率。因此，Gini(D)越小，则数据集 D 的纯度越高。

对于样本 D，个数为 $|D|$ ，根据特征 A 是否取某一可能值 a，把样本 D 分成两部分 D_1 和 D_2 。所以 CART 分类树算法建立起来的是二叉树，而不是多叉树。

$$D_1 = (x, y) \in D | A(X) = a, D_2 = D - D_1 \quad (5.10)$$

在属性 A 的条件下，样本 D 的基尼系数定义为：

$$\text{GiniIndex}(D|A = a) = \frac{|D_1|}{|D|} \text{Gini}(D_1) + \frac{|D_2|}{|D|} \text{Gini}(D_2) \quad (5.11)$$

CART 分类决策树的形成。首先根据训练标注数据，从根节点开始，递归的对每个节点进行操作，构成一个二叉决策树。

- (1) 对于当前节点的数据集为 D ，如果样本个数小于阈值或没有特征，则返回决策子树，当前节点停止递归；
- (2) 计算样本集 D 的基尼系数，如果基尼系数小于阈值，则返回决策子树，当前节点停止递归；
- (3) 计算当前节点现有各个特征的各个值的基尼指数；
- (4) 在计算出来的各个特征的各个值的基尼系数中，选择基尼系数最小的特征 A 及其对应的取值 a 作为最优特征和最优切分点。然后根据最优特征和最优切分点，将本节点的数据集划分成两部分 D_1 和 D_2 ，同时生成当前节点的两个子节点，左节点的数据集为 D_1 ，右节点的数据集为 D_2 。
- (5) 对左右的子节点递归调用 1-4 步，生成 CART 分类树。

对生成的 CART 分类树做预测时，假如测试集里的样本落到了某个叶子节点，而该节点里有多个训练样本。则该测试样本的类别为这个叶子节点里概率最大的类别。

算法停止的条件有三个：节点中的样本个数小于设定阈值、样本集的基尼系数小于设定阈值和没有更多特征可分。

本文首先将已知类别的数据使用 python 来训练 CART 分类模型，再将附件表单 3 未知数据代入模型进行预测分类，最后得到分类结果。

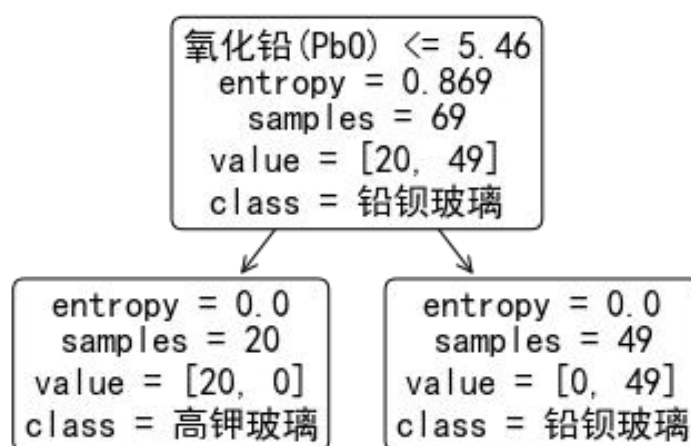


图 5-7 决策树可视化

上面图 1 是该决策树模型结果可视化，由图可以看出来当氧化铅的化学成分

含量 ≤ 5.46 的数据样本为高钾玻璃；当氧化铅的化学成分含量 ≤ 5.46 的数据样本为铅钡玻璃。因此，未知类型的玻璃文物依次可以分为：高钾玻璃，铅钡玻璃，铅钡玻璃，铅钡玻璃，铅钡玻璃，高钾玻璃，高钾玻璃，铅钡玻璃。

5.3.3 分类结果的敏感性分析

需要对分类结果进行敏感性分析，实际上是对使用的分类模型进行效果评估。因此本文将通过分类模型的评价指标来进行分析。一般选取其中的准确率（Accuracy）、精确率（Precision）、召回率（Recall）来进行效果评估。

- (1) 准确率（Accuracy）：预测正确的结果占总样本的百分比。python 中的具体实现。

```
6 print(accuracy_score(y_test, pred_model))  
1.0
```

图 5-8 准确率结果

图 5-8 里面的 `y_test` 为验证集，`pred_model` 为分类器的返回值。结果显示该模型预测全部正确。

- (2) 精确率（Precision）：所有预测值为 1 的样本中，被正确预测的比例。python 中的具体实现。

```
1 precision_score(y_test, pred_model)  
1.0
```

图 5-9 精确率结果

图 5-9 结果显示，在所有预测值为 1 的样本中，全部被预测正确。

- (3) 召回率（Recall）：也表示为检测率。在所有实际值为 1 的样本中，被正确预测为 1 的比例。python 的具体实现。

```
1 recall_score(y_test, pred_model)  
1.0
```

图 5-10 召回率结果

图 5-10 结果显示，在所有实际值为 1 的样本中，全部被预测为 1。

对 CART 分类树模型进行了效果评估后发现，该模型与本文的数据非常契合，预测分类的结果很好。由于之前在训练模型中为了防止出现结果过拟合现象，就进行了数据维度规约，并且在训练模型过程中进行了剪枝操作。因此，该模型很稳定，分类结果准确。

5.4 问题四的建模与求解

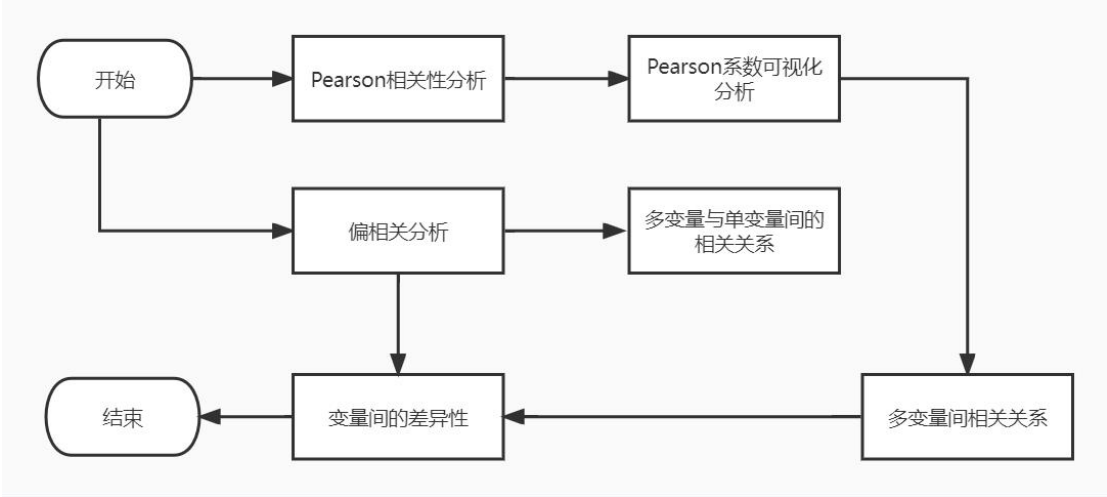


图 5-11 问题四解题流程

探究不同类别的玻璃文物样品之间化学成分之间的关联关系，即探究不同类玻璃化学成分间的相关程度，因表单 2 中的数据为连续变量，因此选择 Pearson 相关性分析多变量间的关系，使用偏相关分析更进一步分析两两变量之间的相关关系并判断化学成分关联关系的差异性的差异性。

5.4.1 对不同类别玻璃，分析其化学成分之间的关联关系

1. Pearson 相关性分析

Step1: 首先通过 Pearson 相关性分析，计算高钾玻璃中化学成分之间的相关性，得到结果部分见表 5-11（详见附录）

表 5-11 高钾玻璃各化学成分间 Pearson 系数

	二氧化硅 (SiO ₂)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al ₂ O ₃)
二氧化硅(SiO ₂)	1	-0.457198	-0.876793	-0.822889	-0.602978	-0.830732
氧化钠(Na ₂ O)	-0.457198	1	0.55422	0.620628	-0.244731	0.344563

氧化钾 (K ₂ O)	-0.876793	0.55422	1	0.759379	0.386963	0.590856
氧化钙 (CaO)	-0.822889	0.620628	0.759379	1	0.20506	0.536369
氧化镁 (MgO)	-0.602978	-0.244731	0.386963	0.20506	1	0.697763
氧化铝 (Al ₂ O ₃)	-0.830732	0.344563	0.590856	0.536369	0.697763	1

将上表进行可视化，得到图 5-12：

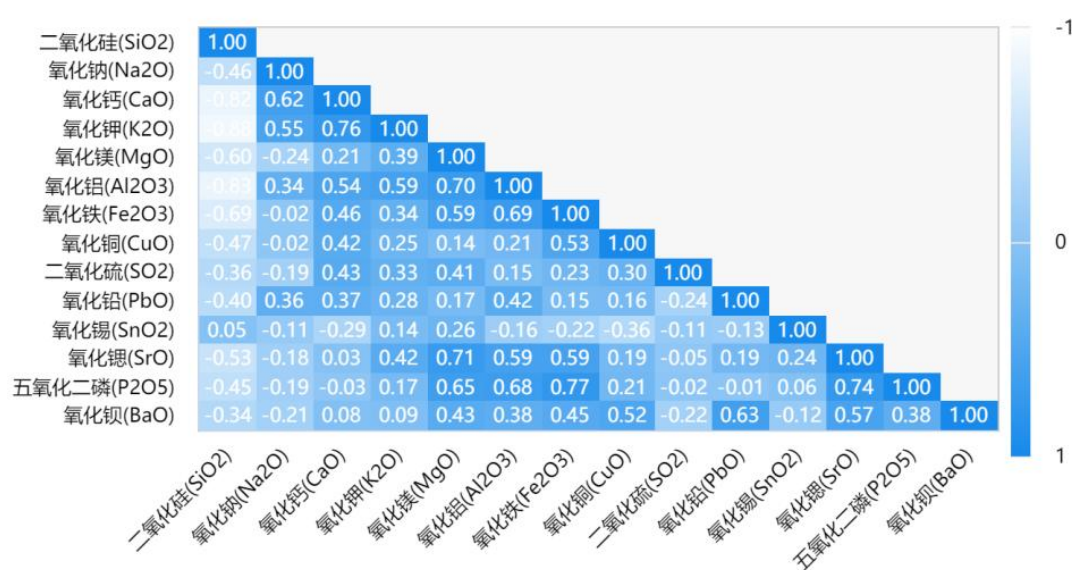


图 5-12 高钾玻璃各成分间 Pearson 系数可视化

由表 5-11 和图 5-12 分析可知：

二氧化硅(SiO₂)与氧化钾(K₂O)，氧化钙(CaO)，氧化镁(MgO)，氧化铝(Al₂O₃)，氧化铁(Fe₂O₃)，氧化锶(SrO)共 6 项之间全部均呈现出显著性，相关系数值分别是-0.877, -0.823, -0.603, -0.831, -0.687, -0.525，全部均小于 0，意味着二氧化硅(SiO₂)与氧化钾(K₂O)，氧化钙(CaO)，氧化镁(MgO)，氧化铝(Al₂O₃)，氧化铁(Fe₂O₃)，氧化锶(SrO)共 6 项之间有着负相关关系。同时，二氧化硅(SiO₂)与氧化钠(Na₂O)，氧化铜(CuO)，氧化铅(PbO)，氧化钡(BaO)，五氧化二磷(P₂O₅)，氧化锡(SnO₂)，二氧化硫(SO₂)共 7 项之间并不会呈现出显著性，相关系数值接近于 0，说明二氧化硅(SiO₂)与氧化钠(Na₂O)，氧化铜(CuO)，氧化铅(PbO)，氧化钡(BaO)，五氧化二磷(P₂O₅)，氧化锡(SnO₂)，二氧化硫(SO₂)共 7 项之间并没有相关关系。

Step2: 计算铅钡玻璃中化学成分之间的相关性，得到结果部分见表 5-12（详见附录）

表 5-12 铅钡玻璃各化学成分间 Pearson 系数

	二氧化硅 (SiO ₂)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al ₂ O ₃)
二氧化硅 (SiO ₂)	1	0.358253	0.080851	-0.485343	0.087003	0.401822
氧化钠 (Na ₂ O)	0.358253	1	-0.065782	-0.369802	0.024852	0.102557
氧化钾 (K ₂ O)	0.080851	-0.065782	1	0.104365	0.269183	0.306631
氧化钙 (CaO)	-0.485343	-0.369802	0.104365	1	0.419708	0.136033
氧化镁 (MgO)	0.087003	0.024852	0.269183	0.419708	1	0.449899
氧化铝 (Al ₂ O ₃)	0.401822	0.102557	0.306631	0.136033	0.449899	1

将上表进行可视化，得到图 5-13：



图 5-13 铅钡玻璃各成分间 Pearson 系数可视化

由表 5-12 和图 5-13 分析可知：

二氧化硅(SiO₂)与氧化钠(Na₂O), 氧化钙(CaO), 氧化铝(Al₂O₃), 氧化铜(CuO), 氧化铅(PbO), 氧化钡(BaO), 五氧化二磷(P₂O₅), 氧化锶(SrO), 二氧化硫(SO₂)共 9 项之间的相关关系系数值呈现出显著性, 除此之外, 二氧化硅(SiO₂)与氧化钾(K₂O), 氧化镁(MgO), 氧化铁(Fe₂O₃), 氧化锡(SnO₂)共 4 项之间的相关关系系数值并不会呈现出显著性(p>0.05), 意味着二氧化硅(SiO₂)与氧化钾(K₂O), 氧化镁(MgO), 氧化铁(Fe₂O₃), 氧化锡(SnO₂)共 4 项之间并没有相关关系。

Step3: 在问题三中, 根据决策树分析得到, 氧化铝作为区别高钾玻璃和铅钡玻璃中最关键化学成分。首先高钾玻璃中氧化铝为主要分析对象, 分析氧化铝与其他化学成分的相关性。

在公式 (5.12) 中, x 代表氧化铝, y 分别代表二氧化硅、氧化钠、氧化钾等其他化学成分

$$P_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{E[(x-x_i)(y-y_i)]}{\sigma_x \sigma_y} \quad (5.12)$$

其中 E 是数学期望, cov(x,y)是 x,y 的协方差, σ_x 、 σ_y 分别是 x 和 y 的标准差, x_i 、 y_i 分别是 x 和 y 的期望。

首先以高钾玻璃中氧化铝为主要分析对象, 分析氧化铝与其他化学成分的相关性如下表 5-13 所示。

表 5-13 高钾玻璃中氧化铝与其他化学成分间 Pearson 系数

	氧化镁 (MgO)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化锶 (SrO)	五氧化二磷 (P ₂ O ₅)	氧化铁 (Fe ₂ O ₃)	氧化钡 (BaO)	氧化锡 (SnO ₂)	氧化钙 (CaO)	二氧化硅 (SiO ₂)	二氧化硫 (SO ₂)	氧化铜 (CuO)	氧化铅 (PbO)
氧化铝 (Al ₂ O ₃)	0.698**	0.345	0.591**	0.595*	0.676**	0.689**	0.381	-0.163	0.536*	-0.831**	0.147	0.21	0.422

*p<0.05 **p<0.01

应用偏相关分析可在排除了其他变量的影响情况下, 再计算两变量的相关关系,



图 5-14 高钾玻璃中氧化铝与其他化学成分间 Pearson 系数可视化

从表 5-13 和图 5-14 中可知，随着氧化铝的增加，二氧化硅、氧化锡成负相关，其他成分如氧化钠(Na₂O)、氧化钾(K₂O)、氧化钙(CaO)、氧化镁(MgO)等成分能是成正相关，其中与五氧化二磷(P₂O₅)相关性最大，相关性为 0.6834。

Step4: 以铅钡玻璃中氧化铝为主要分析对象，分析氧化铝与其他化学成分的相关性如下表 xx 所示。

表 5-14 铅钡玻璃中氧化铝与其他化学成分间 Pearson 系数

	二氧化硅(SiO ₂)	氧化钠(Na ₂ O)	氧化钾(K ₂ O)	氧化钙(CaO)	氧化镁(MgO)	氧化铁(Fe ₂ O ₃)	氧化铜(CuO)	氧化铅(PbO)	氧化锡(SnO ₂)	二氧化硫(SO ₂)	氧化钡(BaO)	氧化锶(SrO)	五氧化二磷(P ₂ O ₅)
氧化铝(Al ₂ O ₃)	0.402**	0.103	0.307*	0.136	0.450**	0.232	-0.282	-0.443*	0.469**	-0.206	-0.341*	-0.156	-0.075



图 5-15 高钾玻璃中氧化铝与其他化学成分间 Pearson 系数可视化

从表 5-14 和图 5-15 中可知，随着氧化铝的增加，氧化铜(CuO)、氧化铅(PbO)、氧化钡(BaO)、五氧化二磷(P₂O₅)、氧化锶(SrO)和二氧化硫(SO₂)成负相关，其他成分如二氧化硅(SiO₂)、氧化钠(Na₂O)、氧化钙(CaO)、氧化镁(MgO)、氧化铁(Fe₂O₃)等成分能是成正相关，其中与五氧化二磷(P₂O₅)相关性最大，相关性为-0.0743。

2. 偏相关分析

为更深入探究不同类别的玻璃的化学成分之间的关联关系，通过偏相关分析，排除其他变量的影响，只分析两个化学成分之间的关联性。

设有 14 个变量 x_1, x_2, \dots, x_{14} ，则任意两个变量 x_i 和 x_j 的 g 阶 ($g \leq 12$) 阶样本偏相关系数公式为：

$$r_{ij \cdot l_1 \cdot l_2 \cdots l_g} = \frac{r_{ij \cdot l_1 \cdot l_2 \cdots l_g} - r_{il_g \cdot l_2 \cdots l_{g-1}} r_{jl_g \cdot l_2 \cdots l_{g-1}}}{\sqrt{(1 - r_{il_g \cdot l_2 \cdots l_{g-1}}^2)(1 - r_{jl_g \cdot l_2 \cdots l_{g-1}}^2)}} \quad (5.13)$$

由氧化铝作为其他 13 个化学成分的“桥梁”，进行偏相关分析。

高钾玻璃中化学成分偏相关分析部分结果见表 5-15（详见附录）

表 5-15 高钾玻璃中化学成分偏相关分析结果

		二氧化硅 (SiO ₂)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)
二氧化硅 (SiO ₂)	相关性	1	-0.327	-0.859	-0.803	-0.058
	显著性 (双尾)	.	0.2	0	0	0.824
	自由度	0	15	15	15	15
氧化钠 (Na ₂ O)	相关性	-0.327	1	0.463	0.55	-0.721
	显著性 (双尾)	0.2	.	0.061	0.022	0.001
	自由度	15	0	15	15	15
氧化钾 (K ₂ O)	相关性	-0.859	0.463	1	0.65	-0.044
	显著性 (双尾)	0	0.061	.	0.005	0.867
	自由度	15	15	0	15	15
氧化钙 (CaO)	相关性	-0.803	0.55	0.65	1	-0.28
	显著性 (双尾)	0	0.022	0.005	.	0.277
	自由度	15	15	15	0	15
氧化镁 (MgO)	相关性	-0.058	-0.721	-0.044	-0.28	1
	显著性 (双尾)	0.824	0.001	0.867	0.277	.
	自由度	15	15	15	15	0

通过上式计算铅钡玻璃中化学成分偏相关，部分结果见表 5-16（详见目录）

表 5-16 铅钡玻璃中化学成分偏相关分析结果

		二氧化硅 (SiO ₂)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)
二氧化硅 (SiO ₂)	相关性	1	0.348	-0.049	-0.595	-0.115
	显著性 (双尾)	.	0.017	0.746	0	0.443
	自由度	0	45	45	45	45

氧化钠 (Na ₂ O)	相关性	0.348	1	-0.103	-0.389	-0.024
	显著性 (双尾)	0.017	.	0.492	0.007	0.873
	自由度	45	0	45	45	45
氧化钾 (K ₂ O)	相关性	-0.049	-0.103	1	0.066	0.154
	显著性 (双尾)	0.746	0.492	.	0.657	0.3
	自由度	45	45	0	45	45
氧化钙 (CaO)	相关性	-0.595	-0.389	0.066	1	0.405
	显著性 (双尾)	0	0.007	0.657	.	0.005
	自由度	45	45	45	0	45
氧化镁 (MgO)	相关性	-0.115	-0.024	0.154	0.405	1
	显著性 (双尾)	0.443	0.873	0.3	0.005	.
	自由度	45	45	45	45	0

通过表 5-15 和表 16 可以看出，它们之间仍然具有相关性，高钾玻璃中随着氧化铝的增加，二氧化硅、氧化锡等化学成分的相关性减小，氧化钠 (Na₂O)、氧化钾 (K₂O) 等成分相关性增大；铅钡玻璃中随着氧化铝的增加，氧化铜 (CuO)、氧化铅 (PbO) 等化学成分的相关性减小，二氧化硅 (SiO₂)、氧化钠 (Na₂O) 等成分相关性增大；

5.4.2 不同类别间的化学成分关联关系的差异性

应用偏相关分析的结果，若两化学成分间的偏相关系数大，则这两个化学成分关联关系的差异性也大；反之，若两化学成分间的偏相关系数小，则这两个化学成分关联关系的差异性小；通过分析得到：

在高钾玻璃中，二氧化硅(SiO₂)与氧化钾(K₂O)，氧化钙(CaO)，氧化镁(MgO)，氧化铝(Al₂O₃)，氧化铁(Fe₂O₃)，氧化锶(SrO)共 6 项之间全部均呈现出显著性，相关系数值分别是-0.877, -0.823, -0.603, -0.831, -0.687, -0.525，全部均小于 0，意味着二氧化硅(SiO₂)与氧化钾(K₂O)，氧化钙(CaO)，氧化镁(MgO)，氧化铝(Al₂O₃)，氧化铁(Fe₂O₃)，氧化锶(SrO)共 6 项之间有着负相关关系。同时，二氧化硅(SiO₂)与氧化钠(Na₂O)，氧化铜(CuO)，氧化铅(PbO)，氧化钡(BaO)，五氧化二磷(P₂O₅)，氧化锡(SnO₂)，二氧化硫(SO₂)共 7 项之间并不会呈现出显著性，相

关系数值接近于 0，说明二氧化硅(SiO_2)与氧化钠(Na_2O)，氧化铜(CuO)，氧化铅(PbO)，氧化钡(BaO)，五氧化二磷(P_2O_5)，氧化锡(SnO_2)，二氧化硫(SO_2)共 7 项之间并没有相关关系。

在铅钡玻璃中，二氧化硅(SiO_2)与氧化钠(Na_2O)，氧化铝(Al_2O_3)，氧化铜(CuO)，氧化铅(PbO)，氧化钡(BaO)，五氧化二磷(P_2O_5)，氧化钙(CaO)，氧化锶(SrO)，二氧化硫(SO_2)共 9 项之间的相关关系系数值呈现出显著性，具体来看，二氧化硅(SiO_2)和氧化钠(Na_2O)之间的相关系数值为 0.358，并且呈现出 0.05 水平的显著性，因而说明二氧化硅(SiO_2)和氧化钠(Na_2O)之间有着显著的正相关关系。二氧化硅(SiO_2)和氧化铝(Al_2O_3)之间的相关系数值为 0.402，并且呈现出 0.01 水平的显著性，因而说明二氧化硅(SiO_2)和氧化铝(Al_2O_3)之间有着显著的正相关关系。二氧化硅(SiO_2)和氧化铜(CuO)之间的相关系数值为-0.350，并且呈现出 0.05 水平的显著性，因而说明二氧化硅(SiO_2)和氧化铜(CuO)之间有着显著的负相关关系。二氧化硅(SiO_2)和氧化铅(PbO)之间的相关系数值为-0.736，并且呈现出 0.01 水平的显著性，因而说明二氧化硅(SiO_2)和氧化铅(PbO)之间有着显著的负相关关系。二氧化硅(SiO_2)和氧化钡(BaO)之间的相关系数值为-0.446，并且呈现出 0.01 水平的显著性，因而说明二氧化硅(SiO_2)和氧化钡(BaO)之间有着显著的负相关关系。二氧化硅(SiO_2)和五氧化二磷(P_2O_5)之间的相关系数值为-0.564，并且呈现出 0.01 水平的显著性，因而说明二氧化硅(SiO_2)和五氧化二磷(P_2O_5)之间有着显著的负相关关系。二氧化硅(SiO_2)和氧化钙(CaO)之间的相关系数值为-0.485，并且呈现出 0.01 水平的显著性，因而说明二氧化硅(SiO_2)和氧化钙(CaO)之间有着显著的负相关关系。二氧化硅(SiO_2)和氧化锶(SrO)之间的相关系数值为-0.513，并且呈现出 0.01 水平的显著性，因而说明二氧化硅(SiO_2)和氧化锶(SrO)之间有着显著的负相关关系。二氧化硅(SiO_2)和二氧化硫(SO_2)之间的相关系数值为-0.390，并且呈现出 0.01 水平的显著性，因而说明二氧化硅(SiO_2)和二氧化硫(SO_2)之间有着显著的负相关关系。除此之外，二氧化硅(SiO_2)与氧化镁(MgO)，氧化钾(K_2O)，氧化铁(Fe_2O_3)，氧化锡(SnO_2)共 4 项之间的相关关系数值并不会呈现出显著性($p>0.05$)，意味着二氧化硅(SiO_2)与氧化镁(MgO)，氧化钾(K_2O)，氧化铁(Fe_2O_3)，氧化锡(SnO_2)共 4 项之间并没有相关关系。

六、模型的评价与推广

6.1 模型优点

- (1) 模型在一定程度上准确且巧妙的描述了问题，并做出了一些简化，易于理解和解答。
- (2) 采用 Agglomerative Clustering 方法，这是一种自底而上的层次聚类方法，它能够根据指定的相似度或距离定义计算出类之间的距离，可以一次性得到聚类树，后期再分类无需重新计算，能够发现类别的层次关系。
- (3) 决策树算法更适应与这样的小数据集，可以处理不相关特征数据。并且效率高，便于解释，决策树只需要一次构建，反复使用，每一次预测的最大计算次数不超过决策树的深度。

6.2 模型缺点

- (1) K-means 算法中，一开始的 k 个中心点是随机选定的，在后面的迭代中再进行重算，直到收敛。但是根据算法的步骤不难看出，这样一来最后所生成的结果往往很大程度上取决于一开始 K 个中心点的位置。这样一来，也就意味着结果具备很大的随机性，每次计算结果都会因为初始随机选择的中心质点不一样而导致结果不一样。
- (2) 在该题数据中出现了较多缺失，对于各类别样本数量不一致的数据，在决策树当中，信息增益的结果会偏向于那些具有更多数值的特征，可能会使结果不稳定。

6.3 模型改进

- (1) K-means 算法中的 K 值的选取不好把握，可以通过在一开始给定一个适合的数值给 k ，通过一次 K-means 算法得到一次聚类中心。对于得到的聚类中心，根据得到的 k 个聚类的距离情况，合并距离最近的类，因此聚类中心数减小，当将其用于下次聚类时，相应的聚类数目也减小了，最终得到合适数目的聚类数。可以通过一个评判值 E 来确定聚类数得到一个合适的位置停下来，而不继续合并聚类中心。重复上述循环，直至评判函数收敛为止，最终得到较优聚类数的聚类结果。

6.4 模型推广

- (1) K-Means 收敛速度快，聚类效果较优，在分类、预测、规则提取等领域有着

广泛的应用。利用决策树算法对银行的信用卡客户进行分类,构建客户分类模型,给出分类步骤中信用卡的客户类型特征,可以为银行的信用卡推广业务提供相应策略,为银行的发展提供有力支撑。

参考文献

- [1] 于飞,赵岩红.基于卡方检验法的河北省农村电商发展影响因素分析[J].河北软件职业技术学院学报,2021,23(02):17-20.DOI:10.13314/j.cnki.jhbsi.2021.02.006.
- [2] 周世杰. 层次聚类的算法研究[J]. 课程教育研究, 2018, (40):240-241.
- [3] 栾丽华,吉根林.决策树分类技术研究[J].计算机工程,2004,30(9):94-96+105.
- [4] 邢涛,雍毅,侯江,吴怡,吴迪,刘恒博.基于主成分分析与层次聚类分析的水质综合评价[J].四川环境,2022,41(04):131-139.DOI:10.14034/j.cnki.schj.2022.04.019.

附录

问题一：

卡方分布临界值表：

df/a	1.00	0.99	0.98	0.95	0.90	0.10	0.05	0.03	0.01	0.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99

29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67

样本类型:

类型	表面风化	二氧化硅 (SiO ₂)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P ₂ O ₅)	氧化锶 (SrO)	氧化锡 (SnO ₂)	二氧化硫 (SO ₂)
高钾	无风化	69.33	0	9.99	6.32	0.87	3.93	1.74	3.87	0	0	1.17	0	0	0.39
高钾	无风化	87.05	0	5.19	2.01	0	4.06	0	0.78	0.25	0	0.66	0	0	0
高钾	无风化	61.71	0	12.37	5.87	1.11	5.5	2.16	5.09	1.41	2.86	0.7	0.1	0	0
高钾	无风化	65.88	0	9.67	7.12	1.56	6.44	2.06	2.18	0	0	0.79	0	0	0.36
高钾		61.58	0	10.95	7.35	1.77	7.5	2.62	3.27	0	0	0.94	0.06	0	0.47

无 风 化														
高 钾 无 风 化	67. 65	0	7.3 7	0	1.9 8	11.1 5	2.39	2.5 1	0.2	1.3 8	4.1 8	0.1 1	0	0
高 钾 无 风 化	59. 81	0	7.6 8	5.4 1	1.7 3	10.0 5	6.04	2.1 8	0.3 5	0.9 7	4.5	0.1 2	0	0
高 钾 无 风 化	59. 01	2.8 6	12. 53	8.7	0	6.16	2.88	4.7 3	0	0	1.2 7	0	0	0
高 钾 无 风 化	62. 47	3.3 8	12. 28	8.2 3	0.6 6	9.23	0.5	0.4 7	1.6 2	0	0.1 6	0	0	0
高 钾 无 风 化	65. 18	2.1	14. 52	8.2 7	0.5 2	6.18	0.42	1.0 7	0.1 1	0	0	0.0 4	0	0
高 钾 无 风 化	79. 46	0	9.4 2	0	1.5 3	3.05	0	0	0	0	1.3 6	0.0 7	2.3 6	0
高 钾	76. 68	0	0	4.7 1	1.2 2	6.19	2.37	3.2 8	1	1.9 7	1.1	0	0	0

无 风 化														
高 钾 风 化	67.6 5	0	7.3 7	0	1.98	11.1 5	2.39	2.51	0.2	1.3 8	4.18	0.1 1	0	0
高 钾 风 化	59.8 1	0	7.6 8	5.4 1	1.73	10.0 5	6.04	2.18	0.3 5	0.9 7	4.5	0.1 2	0	0
高 钾 风 化	92.6 3	0	0	1.0 7	0	1.98	0.17	3.24	0	0	0.61	0	0	0
高 钾 风 化	95.0 2	0	0.5 9	0.6 2	0	1.32	0.32	1.55	0	0	0.35	0	0	0
高 钾 风 化	79.4 6	0	9.4 2	0	1.53	3.05	0	0	0	0	1.36	0.0 7	2.36	0
高 钾 风 化	76.6 8	0	0	4.7 1	1.22	6.19	2.37	3.28	1	1.9 7	1.1	0	0	0
高 钾 风 化	92.3 5	0	0.7 4	1.6 6	0.64	3.5	0.35	0.55	0	0	0.21	0	0	0
高 钾 风 化	92.7 2	0	0	0.9 4	0.54	2.51	0.2	1.54	0	0	0.36	0	0	0

铅 钡 无 风 化	37. 36	0	0.7 1	0	0	5.45	1.51	4.7 8	9.3	23. 55	5.7 5	0	0	0
铅 钡 无 风 化	31. 94	0	0	0.4 7	0	1.59	0	8.4 6	29. 14	26. 23	0.1 4	0.9 1	0	0
铅 钡 无 风 化	34. 34	0	1.4 1	4.4 9	0.9 8	4.35	2.12	0	39. 22	10. 29	0	0.3 5	0.4	0
铅 钡 无 风 化	36. 93	0	0	4.2 4	0.5 1	3.86	2.74	0	37. 74	10. 35	1.4 1	0.4 8	0.4 4	0
铅 钡 无 风 化	65. 91	0	0	1.6	0.8 9	3.11	4.59	0.4 4	16. 55	3.4 2	1.6 2	0.3	0	0
铅 钡 无 风 化	69. 71	0	0.2 1	0.4 6	0	2.36	1	0.1 1	19. 76	4.8 8	0.1 7	0	0	0
铅 钡 无 风 化	75. 51	0	0.1 5	0.6 4	1	2.35	0	0.4 7	16. 16	3.5 5	0.1 3	0	0	0

铅 钡 无 风 化	65. 91	0	0	0.3 8	0	1.44	0.17	0.1 6	22. 05	5.6 8	0.4 2	0	0	0
铅 钡 无 风 化	60. 12	0	0.2 3	0.8 9	0	2.72	0	3.0 1	17. 24	10. 34	1.4 6	0.3 1	0	3.6 6
铅 钡 无 风 化	61. 28	2.6 6	0.1 1	0.8 4	0.7 4	5	0	0.5 3	15. 99	10. 96	0	0.2 3	0	0
铅 钡 无 风 化	55. 21	0	0.2 5	0	1.6 7	4.79	0	0.7 7	25. 25	10. 06	0.2	0.4 3	0	0
铅 钡 无 风 化	51. 54	4.6 6	0.2 9	0.8 7	0.6 1	3.06	0	0.6 5	25. 4	9.2 3	0.1	0.8 5	0	0
铅 钡 无 风 化	49. 01	2.7 1	0	1.1 3	0	1.45	0	0.8 6	32. 92	7.9 5	0.3 5	0	0	0
铅 钡 风 化	20.1 4	0	0	1.4 8	0	1.34	0	10.4 1	28. 68	31. 23	3.59	0.3 7	0	2.5 8
铅	4.61	0	0	3.1	0	1.11	0	3.14	32.	30.	7.56	0.5	0	15.

钡 风 化				9					45	62		3		03
铅 钡 风 化	29.6 4	0	0	2.9 3	0.59	3.57	1.33	3.51	42. 82	5.3 5	8.83	0.1 9	0	0
铅 钡 风 化	37.3 6	0	0.7 1	0	0	5.45	1.51	4.78	9.3	23. 55	5.75	0	0	0
铅 钡 风 化	31.9 4	0	0	0.4 7	0	1.59	0	8.46	29. 14	26. 23	0.14	0.9 1	0	0
铅 钡 风 化	19.7 9	0	0	1.4 4	0	0.7	0	10.5 7	29. 53	32. 25	3.13	0.4 5	0	1.9 6
铅 钡 风 化	3.72	0	0.4	3.0 1	0	1.18	0	3.6	29. 92	35. 45	6.04	0.6 2	0	15. 95
铅 钡 风 化	35.7 8	0	0.2 5	0.7 8	0	1.62	0.47	1.51	46. 55	10	0.34	0.2 2	0	0
铅 钡 风 化	39.5 7	2.22	0.1 4	0.3 7	0	1.6	0.32	0.68	41. 61	10. 83	0.07	0.2 2	0	0
铅 钡 风 化	32.9 3	1.38	0	0.6 8	0	2.57	0.29	0.73	49. 31	9.7 9	0.48	0.4 1	0	0

铅 钡 风 化	26.2 5	0	0	1.1 1	0	0.5	0	0.88	61. 03	7.2 2	1.16	0.6 1	0	0
铅 钡 风 化	16.7 1	0	0	1.8 7	0	0.45	0.19	0	70. 21	6.6 9	1.77	0.6 8	0	0
铅 钡 风 化	18.4 6	0	0.4 4	4.9 6	2.73	3.33	1.79	0.19	44. 12	9.7 6	7.46	0.4 7	0	0
铅 钡 风 化	12.4 1	0	0	5.2 4	0.89	2.25	0.76	5.35	59. 85	7.2 9	0	0.6 4	0	0
铅 钡 风 化	21.7	0	0	6.4	0.95	3.41	1.39	1.51	44. 75	3.2 6	12.8 3	0.4 7	0	0
铅 钡 风 化	53.3 3	0.8	0.3 2	2.8 2	1.54	13.6 5	1.03	0	15. 71	7.3 1	1.1	0.2 5	1.31	0
铅 钡 风 化	28.7 9	0	0	4.5 8	1.47	5.38	2.74	0.7	34. 18	6.1	11.1	0.4 6	0	0
铅 钡 风 化	17.9 8	0	0	3.1 9	0.47	1.87	0.33	1.13	44	14. 2	6.34	0.6 6	0	0
铅 钡 风	24.6 1	0	0	3.5 8	1.19	5.25	1.19	1.37	40. 24	8.9 4	8.1	0.3 9	0.47	0

化														
铅 钡 风 化	21.3 5	0	0	5.1 3	1.45	2.51	0.42	0.75	51. 34	0	8.75	0	0	0
铅 钡 风 化	25.7 4	1.22	0	2.2 7	0.55	1.16	0.23	0.7	47. 42	8.6 4	5.71	0.4 4	0	0
铅 钡 风 化	22.2 8	0	0.3 2	3.1 9	1.28	4.15	0	0.83	55. 46	7.0 4	4.24	0.8 8	0	0
铅 钡 风 化	17.1 1	0	0	0	1.11	3.65	0	1.34	58. 46	0	14.1 3	1.1 2	0	0
铅 钡 风 化	29.1 5	0	0	1.2 1	0	1.85	0	0.79	41. 25	15. 45	2.54	0	0	0
铅 钡 风 化	25.4 2	0	0	1.3 1	0	2.18	0	1.16	45. 1	17. 3	0	0	0	0
铅 钡 风 化	30.3 9	0	0.3 4	3.4 9	0.79	3.52	0.86	3.13	39. 35	7.6 6	8.99	0.2 4	0	0

预测风化前的化学成分含量：

类型	二氧化硅 (SiO ₂)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P ₂ O ₅)	氧化锶 (SrO)	氧化锡 (SnO ₂)	二氧化硫 (SO ₂)
铅钡风化	43.24		0.00		0.00		0.00		16.11		0.65		0.00	
铅钡风化	9.90		0.00		0.00		0.00		18.23		1.37		0.00	
铅钡风化	63.64		0.00		0.50		2.17		24.06		1.59		0.00	
铅钡风化	80.22		1.63		0.00		2.47		5.23		1.04		0.00	
铅钡风化	68.58		0.00		0.00		0.00		16.37		0.03		0.00	
铅钡风化	42.49		0.00		0.00		0.00		16.59		0.57		0.00	
铅钡风	7.99		0.92		0.00		0.00		16.81		1.09		0.00	

化							
铅 钡 风化	76.82	0.58	0.00	0.77	26.16	0.06	0.00
铅 钡 风化	84.96	0.32	0.00	0.52	23.38	0.01	0.00
铅 钡 风化	70.71	0.00	0.00	0.47	27.71	0.09	0.00
铅 钡 风化	56.36	0.00	0.00	0.00	34.29	0.21	0.00
铅 钡 风化	35.88	0.00	0.00	0.31	39.45	0.32	0.00
铅 钡 风化	39.64	1.01	2.33	2.92	24.79	1.35	0.00
铅 钡 风化	26.65	0.00	0.76	1.24	33.63	0.00	0.00
铅 钡 风化	46.59	0.00	0.81	2.27	25.14	2.32	0.00
铅 钡	114.51	0.74	1.31	1.68	8.83	0.20	1.24

风 化							
铅 钡 风 化	61.82	0.00	1.25	4.48	19.20	2.00	0.00
铅 钡 风 化	38.61	0.00	0.40	0.54	24.72	1.14	0.00
铅 钡 风 化	52.84	0.00	1.01	1.94	22.61	1.46	0.44
铅 钡 风 化	45.84	0.00	1.24	0.69	28.85	1.58	0.00
铅 钡 风 化	55.27	0.00	0.47	0.38	26.64	1.03	0.00
铅 钡 风 化	47.84	0.74	1.09	0.00	31.16	0.77	0.00
铅 钡 风 化	36.74	0.00	0.95	0.00	32.85	2.55	0.00
铅 钡 风 化	62.59	0.00	0.00	0.00	23.18	0.46	0.00
铅	54.58	0.00	0.00	0.00	25.34	0.00	0.00

钡 风 化							
铅 钡 风 化	65.25	0.78	0.67	1.40	22.11	1.62	0.00
高 钾 风 化	56.06	21.32	2.24	3.12	0.42	3.70	0.00
高 钾 风 化	49.56	22.22	1.95	7.88	0.74	3.99	0.00
高 钾 风 化	76.76	0.00	0.00	0.22	0.00	0.54	0.00
高 钾 风 化	78.74	1.71	0.00	0.42	0.00	0.31	0.00
高 钾 风 化	65.85	27.25	1.73	0.00	0.00	1.20	1.57
高 钾 风 化	63.54	0.00	1.38	3.09	2.12	0.97	0.00
高 钾 风 化	76.53	2.14	0.72	0.46	0.00	0.19	0.00

高 钾 风 化	76.83	0.00	0.61	0.26	0.00	0.32	0.00
------------------	-------	------	------	------	------	------	------

问题二：

无风化数据集聚类标注：

	聚 类 种 类	氧 化 钠 (Na 2O)	氧 化 钾 (K2 O)	氧 化 钙 (Ca O)	氧 化 镁 (Mg O)	氧 化 铝 (Al 2O3)	氧 化 铁 (Fe 2O3)	氧 化 铜 (Cu O)	氧 化 铅 (Pb O)	氧 化 钡 (Ba O)	五 氧 化 二 磷 (P2O 5)	氧 化 锶 (Sr O)	氧 化 锡 (Sn O2)	二 氧 化 硫 (SO 2)
0	1	0	9.9 9	6.3 2	0.8 7	3.9 3	1.7 4	3.8 7	0	0	1.17	0	0	0.3 9
1	1	0	5.1 9	2.0 1	0	4.0 6	0	0.7 8	0.2 5	0	0.66	0	0	0
2	1	0	12. 37	5.8 7	1.1 1	5.5	2.1 6	5.0 9	1.4 1	2.8 6	0.7	0.1	0	0
3	1	0	9.6 7	7.1 2	1.5 6	6.4 4	2.0 6	2.1 8	0	0	0.79	0	0	0.3 6
4	1	0	10. 95	7.3 5	1.7 7	7.5	2.6 2	3.2 7	0	0	0.94	0.0 6	0	0.4 7
5	1	0	7.3 7	0	1.9 8	11. 15	2.3 9	2.5 1	0.2	1.3 8	4.18	0.1 1	0	0
6	1	0	7.6 8	5.4 1	1.7 3	10. 05	6.0 4	2.1 8	0.3 5	0.9 7	4.5	0.1 2	0	0
7	1	2.8 6	12. 53	8.7	0	6.1 6	2.8 8	4.7 3	0	0	1.27	0	0	0
8	1	3.3 8	12. 28	8.2 3	0.6 6	9.2 3	0.5	0.4 7	1.6 2	0	0.16	0	0	0
9	1	2.1	14. 52	8.2 7	0.5 2	6.1 8	0.4 2	1.0 7	0.1 1	0	0	0.0 4	0	0
10	1	0	9.4 2	0	1.5 3	3.0 5	0	0	0	0	1.36	0.0 7	2.3 6	0

11	2	0	0.71	0	0	5.45	1.51	4.78	9.3	23.55	5.75	0	0	0
12	1	0	0	4.71	1.22	6.19	2.37	3.28	1	1.97	1.1	0	0	0
13	2	0	0	0.47	0	1.59	0	8.46	29.14	26.23	0.14	0.91	0	0
14	2	0	1.41	4.49	0.98	4.35	2.12	0	39.22	10.29	0	0.35	0.4	0
15	2	0	0	4.24	0.51	3.86	2.74	0	37.74	10.35	1.41	0.48	0.44	0
16	2	0	0	1.6	0.89	3.11	4.59	0.44	16.55	3.42	1.62	0.3	0	0
17	2	0	0.21	0.46	0	2.36	1	0.11	19.76	4.88	0.17	0	0	0
18	2	0	0.15	0.64	1	2.35	0	0.47	16.16	3.55	0.13	0	0	0
19	2	0	0	0.38	0	1.44	0.17	0.16	22.05	5.68	0.42	0	0	0
20	2	0	0.23	0.89	0	2.72	0	3.01	17.24	10.34	1.46	0.31	0	3.66
21	2	2.66	0.11	0.84	0.74	5	0	0.53	15.99	10.96	0	0.23	0	0
22	2	0	0.25	0	1.67	4.79	0	0.77	25.25	10.06	0.2	0.43	0	0
23	2	4.66	0.29	0.87	0.61	3.06	0	0.65	25.4	9.23	0.1	0.85	0	0
24	2	2.71	0	1.13	0	1.45	0	0.86	32.92	7.95	0.35	0	0	0

风化数据集聚类标注：

聚 类 种 类	氧 化 钠 (Na	氧 化 钾 (K2	氧 化 钙 (Ca	氧 化 镁 (Mg	氧 化 铝 (Al	氧 化 铁 (Fe	氧 化 铜 (Cu	氧 化 铅 (Pb	氧 化 钡 (Ba	五 氧 化二 磷 (P2O	氧 化 锶 (Sr	氧 化 锡 (Sn	二 氧 化 硫
------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	---------------------------	--------------------	--------------------	------------------

		20)	0)	0)	0)	203)	203)	0)	0)	0)	5)	0)	02)	(S0 2)
0	2	0	7.3 7	0	1.9 8	11. 15	2.3 9	2.5 1	0.2	1.3 8	4.18	0.1 1	0	0
1	2	0	7.6 8	5.4 1	1.7 3	10. 05	6.0 4	2.1 8	0.3 5	0.9 7	4.5	0.1 2	0	0
2	2	0	0	1.0 7	0	1.9 8	0.1 7	3.2 4	0	0	0.61	0	0	0
3	1	0	0	1.4 8	0	1.3 4	0	10. 41	28. 68	31. 23	3.59	0.3 7	0	2.5 8
4	1	0	0	3.1 9	0	1.1 1	0	3.1 4	32. 45	30. 62	7.56	0.5 3	0	15. 03
5	2	0	0.5 9	0.6 2	0	1.3 2	0.3 2	1.5 5	0	0	0.35	0	0	0
6	2	0	9.4 2	0	1.5 3	3.0 5	0	0	0	0	1.36	0.0 7	2.3 6	0
7	1	0	0	2.9 3	0.5 9	3.5 7	1.3 3	3.5 1	42. 82	5.3 5	8.83	0.1 9	0	0
8	2	0	0.7 1	0	0	5.4 5	1.5 1	4.7 8	9.3	23. 55	5.75	0	0	0
9	2	0	0	4.7 1	1.2 2	6.1 9	2.3 7	3.2 8	1	1.9 7	1.1	0	0	0
10	2	0	0.7 4	1.6 6	0.6 4	3.5	0.3 5	0.5 5	0	0	0.21	0	0	0
11	1	0	0	0.4 7	0	1.5 9	0	8.4 6	29. 14	26. 23	0.14	0.9 1	0	0
12	1	0	0	1.4 4	0	0.7	0	10. 57	29. 53	32. 25	3.13	0.4 5	0	1.9 6
13	1	0	0.4	3.0 1	0	1.1 8	0	3.6	29. 92	35. 45	6.04	0.6 2	0	15. 95
14	2	0	0	0.9 4	0.5 4	2.5 1	0.2	1.5 4	0	0	0.36	0	0	0
15	1	0	0.2 5	0.7 8	0	1.6 2	0.4 7	1.5 1	46. 55	10	0.34	0.2 2	0	0

1 6	1	2.2 2	0.1 4	0.3 7	0	1.6	0.3 2	0.6 8	41. 61	10. 83	0.07	0.2 2	0	0
1 7	1	1.3 8	0	0.6 8	0	2.5 7	0.2 9	0.7 3	49. 31	9.7 9	0.48	0.4 1	0	0
1 8	1	0	0	1.1 1	0	0.5	0	0.8 8	61. 03	7.2 2	1.16	0.6 1	0	0
1 9	1	0	0	1.8 7	0	0.4 5	0.1 9	0	70. 21	6.6 9	1.77	0.6 8	0	0
2 0	1	0	0.4 4	4.9 6	2.7 3	3.3 3	1.7 9	0.1 9	44. 12	9.7 6	7.46	0.4 7	0	0
2 1	1	0	0	5.2 4	0.8 9	2.2 5	0.7 6	5.3 5	59. 85	7.2 9	0	0.6 4	0	0
2 2	1	0	0	6.4	0.9 5	3.4 1	1.3 9	1.5 1	44. 75	3.2 6	12.8 3	0.4 7	0	0
2 3	2	0.8	0.3 2	2.8 2	1.5 4	13. 65	1.0 3	0	15. 71	7.3 1	1.1	0.2 5	1.3 1	0
2 4	1	0	0	4.5 8	1.4 7	5.3 8	2.7 4	0.7	34. 18	6.1	11.1	0.4 6	0	0
2 5	1	0	0	3.1 9	0.4 7	1.8 7	0.3 3	1.1 3	44	14. 2	6.34	0.6 6	0	0
2 6	1	0	0	3.5 8	1.1 9	5.2 5	1.1 9	1.3 7	40. 24	8.9 4	8.1	0.3 9	0.4 7	0
2 7	1	0	0	5.1 3	1.4 5	2.5 1	0.4 2	0.7 5	51. 34	0	8.75	0	0	0
2 8	1	1.2 2	0	2.2 7	0.5 5	1.1 6	0.2 3	0.7	47. 42	8.6 4	5.71	0.4 4	0	0
2 9	1	0	0.3 2	3.1 9	1.2 8	4.1 5	0	0.8 3	55. 46	7.0 4	4.24	0.8 8	0	0
3 0	1	0	0	0	1.1 1	3.6 5	0	1.3 4	58. 46	0	14.1 3	1.1 2	0	0
3 1	1	0	0	1.2 1	0	1.8 5	0	0.7 9	41. 25	15. 45	2.54	0	0	0
3 2	1	0	0	1.3 1	0	2.1 8	0	1.1 6	45. 1	17. 3	0	0	0	0

3	1	0	0.3	3.4	0.7	3.5	0.8	3.1	39.	7.6	8.99	0.2	0	0
3			4	9	9	2	6	3	35	6		4		

(1) 层次聚类模型代码:

```
import pandas as pd
file1 = pd.read_excel('data/附件.xlsx', sheet_name=' 表单 1')
file2 = pd.read_excel('data/附件.xlsx', sheet_name=' 表单 2')
file3 = pd.read_excel('data/附件.xlsx', sheet_name=' 表单 3')
from tqdm import tqdm
import re
tqdm.pandas()
def clear_id(s):
    ss = str(s)
    n = int(re.findall(r"\d+", ss)[0])
    return n
file2['id'] = file2['文物采样点'].progress_apply(clear_id)
file1['id'] = file1['文物编号'].progress_apply(clear_id)

cols = ['纹饰', '类型', '颜色', '表面风化', '二氧化硅(SiO2)',
        '氧化钠(Na2O)', '氧化钾(K2O)', '氧化钙(CaO)', '氧化镁(MgO)', '氧化铝(Al2O3)',
        '氧化铁(Fe2O3)', '氧化铜(CuO)', '氧化铅(PbO)', '氧化钡(BaO)', '五氧化二磷(P2O5)',
        '氧化锶(SrO)', '氧化锡(SnO2)', '二氧化硫(SO2)']
cols1 = ['二氧化硅(SiO2)',
        '氧化钠(Na2O)', '氧化钾(K2O)', '氧化钙(CaO)', '氧化镁(MgO)', '氧化铝(Al2O3)',
        '氧化铁(Fe2O3)', '氧化铜(CuO)', '氧化铅(PbO)', '氧化钡(BaO)', '五氧化二磷(P2O5)',
        '氧化锶(SrO)', '氧化锡(SnO2)', '二氧化硫(SO2)']
cols2 = ['纹饰', '类型', '颜色', '表面风化',]
file_data = pd.merge(file1, file2, on='id')
Gao_data = file_data[file_data['类型']=='高钾'][cols]
Qian_data = file_data[file_data['类型']=='铅钡'][cols]
```

```

Gao_data.isnull().any()
Qian_data.isnull().any()
Qian_data['颜色'].value_counts()
Qian_data['颜色'] = Qian_data['颜色'].fillna('浅蓝')
Qian_data = Qian_data.fillna(0)
Gao_data = Gao_data.fillna(0)
Gao_data.info()
print(len(Gao_data))
Gao_data
Gao_data[cols1].sum(axis=1)
# 存在两个无效数据
(Gao_data[cols1].sum(axis=1)<85).value_counts()
# 删除两个无效数据
Gao_data = Gao_data.drop([17, 19])

from sklearn.preprocessing import LabelEncoder

Gao_train = pd.DataFrame() # 用于存放编码后的训练数据
label = LabelEncoder() # 标号编码器
X= Gao_data
for c in X.columns: # 对每个特征列
    if X[c].dtype=='object': # 如果是字符串形式的(字符串读到 pandas 里 dtype 是 object)
        Gao_train[c] = label.fit_transform(X[c]) # 将整个这一列进行标号编码, 写到新的 dataframe 里
    else: # 其它类型的特征(数值, 布尔)保持原样写入
        Gao_train[c] = list(X[c])

Qian_train = pd.DataFrame() # 用于存放编码后的训练数据
label = LabelEncoder() # 标号编码器
X= Qian_data
for c in X.columns: # 对每个特征列
    if X[c].dtype=='object': # 如果是字符串形式的(字符串读到 pandas 里 dtype 是 object)
        Qian_train[c] = label.fit_transform(X[c]) # 将整个这一列进行标号编码, 写

```

到新的 dataframe 里

```
else: # 其它类型的特征(数值, 布尔)保持原样写入
    Qian_train[c] = list(X[c])
```

```
import pandas as pd
import numpy as np
import matplotlib

from sklearn.cluster import AgglomerativeClustering #导入 sklearn 的层次聚类函数
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt
plt.rcParams['axes.unicode_minus'] = False
plt.rcParams['font.sans-serif'] = ['SimHei']
data = np.array(Gao_train)

Scores = [] # 存放轮廓系数
SSE = [] # 存放每次结果的误差平方和
for k in range(2, 9):
    estimator = AgglomerativeClustering(n_clusters=k, linkage='ward')
    estimator.fit(data)
    Scores.append(silhouette_score(
        np.array(data), estimator.labels_, metric='euclidean'))
X = range(2, 9)
plt.xlabel('k 值', fontsize=20)
plt.ylabel('轮廓系数', fontsize=20)
plt.plot(X, Scores, 'o-')
plt.savefig('img2/AGG 聚类——高钾数据手肘法.png', dpi=300)
plt.show()

from sklearn.cluster import AgglomerativeClustering #导入 sklearn 的层次聚类函数
import matplotlib.pyplot as plt
k = 6
model = AgglomerativeClustering(n_clusters=k, linkage='ward')
data = Gao_train
model.fit(data) #训练模型
```

```

# 详细输出原始数据及其类别
r = pd.concat([data, pd.Series(model.labels_, index = data.index)], axis = 1) #
详细输出每个样本对应的类别

from scipy.cluster.hierarchy import linkage, dendrogram
#这里使用 scipy 的层次聚类函数
Z = linkage(data, method = 'ward', metric = 'euclidean') #谱系聚类图
P = dendrogram(Z, 0) #画谱系聚类图
plt.savefig('img2/高钾分为 6 类-AGG 层次聚类.png', dpi=300)
plt.show()
clf1.fit_predict(data1)

from sklearn.cluster import AgglomerativeClustering #导入 sklearn 的层次聚类函数
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import linkage, dendrogram
k = 5
model = AgglomerativeClustering(n_clusters = k, linkage = 'ward')
data = Qian_train
model.fit(data) #训练模型
#详细输出原始数据及其类别
r = pd.concat([data, pd.Series(model.labels_, index = data.index)], axis = 1) #
详细输出每个样本对应的类别

#这里使用 scipy 的层次聚类函数
Z = linkage(data, method = 'ward', metric = 'euclidean') #谱系聚类图
P = dendrogram(Z, 0) #画谱系聚类图
plt.savefig('img2/铅钡分为 5 类-AGG 层次聚类.png', dpi=300)
plt.show()
clf1.fit_predict(data1)

```

问题三：

(1) 数据预处理：

```

import pandas as pd
file1 = pd.read_excel('data/附件.xlsx', sheet_name=' 表单 1')
file2 = pd.read_excel('data/附件.xlsx', sheet_name=' 表单 2')

```

```

file3 = pd.read_excel('data/附件.xlsx', sheet_name=' 表单 3')
from tqdm import tqdm
import re
tqdm.pandas()
def clear_id(s):
    ss = str(s)
    n = int(re.findall(r"\d+", ss)[0])
    return n
file2['id'] = file2['文物采样点'].progress_apply(clear_id)
file1['id'] = file1['文物编号'].progress_apply(clear_id)
cols = ['表面风化', '二氧化硅(SiO2)',
        '氧化钠(Na2O)', '氧化钾(K2O)', '氧化钙(CaO)', '氧化镁(MgO)', '氧化铝(Al2O3)',
        '氧化铁(Fe2O3)', '氧化铜(CuO)', '氧化铅(PbO)', '氧化钡(BaO)', '五氧化二磷(P2O5)',
        '氧化锶(SrO)', '氧化锡(SnO2)', '二氧化硫(SO2)']
file_data = pd.merge(file1, file2, on='id')
Gao_data = file_data[file_data['类型']=='高钾'][cols]
Qian_data = file_data[file_data['类型']=='铅钡'][cols]
test_all = file3[cols]
Qian_data = Qian_data.fillna(0)
Gao_data = Gao_data.fillna(0)
test_all = test_all.fillna(0)
train_all = pd.concat([Gao_data, Qian_data])
encode_data = pd.concat([train_all, test_all])
len(train_all)
from sklearn.preprocessing import LabelEncoder

all_data = pd.DataFrame() # 用于存放编码后的训练数据
label = LabelEncoder() # 标号编码器
X = encode_data
for c in X.columns: # 对每个特征列
    if X[c].dtype=='object': # 如果是字符串形式的(字符串读到 pandas 里 dtype 是 object)
        all_data[c] = label.fit_transform(X[c]) # 将整个这一列进行标号编码, 写到

```

新的 dataframe 里

```
else: # 其它类型的特征(数值, 布尔)保持原样写入
    all_data[c] = list(X[c])
train = all_data[0:len(train_all)]
test = all_data[len(train_all):]
from sklearn.model_selection import train_test_split # 拆分数据
import numpy as np
X_var, y_var = train, [0]*len(Gao_data)+[1]*len(Qian_data)
X_train, X_test, y_train, y_test = train_test_split(X_var, y_var, test_size=0.2, random_state=123)
```

(2) 决策树模型训练代码:

```
from sklearn.tree import DecisionTreeClassifier as dtc # 树算法
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
model = dtc(criterion='entropy', max_depth=4)
model.fit(X_train, y_train)
pred_model = model.predict(X_test)
print(accuracy_score(y_test, pred_model))
precision_score(y_test, pred_model)
recall_score(y_test, pred_model)
from sklearn.metrics import precision_recall_curve, average_precision_score
# 求 precision 和 recall 值
precision, recall, _ = precision_recall_curve(y_test, pred_model)
# 求 AP 值
print('AP 值: {:.3f}'.format(average_precision_score(y_test, pred_model)))
```

(3) 决策树分类可视化:

```
from sklearn.tree import plot_tree # 树图
import matplotlib.pyplot as plt
plt.rcParams['axes.unicode_minus'] = False
plt.rcParams['font.sans-serif'] = ['SimHei']
model = dtc(criterion='entropy', max_depth=4)
model.fit(X_var, y_var)
feature_names = X_var.columns
```

```

target_names = ['高钾玻璃', '铅钡玻璃']

plot_tree(model,
           feature_names = feature_names,
           class_names = target_names,
           filled = False,
           rounded = True)

plt.savefig('./img3/决策树可视化.png')
plt.show()
from sklearn.tree import plot_tree # 树图
import matplotlib.pyplot as plt
plt.rcParams['axes.unicode_minus'] = False
plt.rcParams['font.sans-serif'] = ['SimHei']
model = dtc(criterion = 'gini', max_depth = 4)
model.fit(X_var, y_var)
feature_names = X_var.columns
target_names = ['高钾玻璃', '铅钡玻璃']

plot_tree(model,
           feature_names = feature_names,
           class_names = target_names)
plt.show()

```

(4) 结果预测:

```
model.predict(test)
```

问题四:

高钾玻璃各化学成分间 Pearson 系数:

	二 氧 化 硅 (Si	氧 化 钠 (Na	氧 化 钾 (K ₂	氧 化 钙 (Ca	氧 化 镁 (Mg	氧 化 铝 (Al ₂	氧 化 铁 (Fe ₂	氧 化 铜 (Cu	氧 化 铅 (Pb	氧 化 钡 (Ba	五 氧 化 二 磷	氧 化 锶 (Sr	氧 化 锡 (Sn	二 氧 化 硫 (SO
--	-------------------------	--------------------	--------------------------------	--------------------	--------------------	---------------------------------	---------------------------------	--------------------	--------------------	--------------------	-----------------------	--------------------	--------------------	-------------------------

	O ₂)										(P ₂				2)
	O ₅)														
二 氧 化 硅 (S iO ₂) 氧 化 钠 (N a ₂ O) 氧 化 钾 (K ₂ O) 氧 化 钙 (C a O) 氧 化 镁 (M g O)		-0. 1 457 198	-0. 876 793	-0. 822 889	-0. 602 978	-0. 830 732	-0. 687 378	-0. 465 189	-0. 398 057	-0. 344 856	-0. 447 796	-0. 525 401	0.0 485 81	-0. 357 223	
	-0. 457 198		0.5 542 2	0.6 206 28	-0. 244 731	0.3 445 63	-0. 023 046	-0. 022 229	0.3 591 65	-0. 213 569	-0. 185 782	-0. 181 886	-0. 106 211	-0. 194 33	
	-0. 876 793	0.5 542 2		0.7 593 79	0.3 869 63	0.5 908 56	0.3 357 81	0.2 506 7	0.2 752 14	0.0 926 29	0.1 711 73	0.4 208 29	0.1 419 2	0.3 323 95	
	-0. 822 889	0.6 206 28	0.7 593 79		0.2 050 6	0.5 363 69	0.4 570 84	0.4 234 72	0.3 676 6	0.0 783 3	-0. 031 251	0.0 346 59	-0. 290 069	0.4 289 8	
	-0. 602 978	-0. 244 731	0.3 869 63	0.2 050 6		0.6 977 63	0.5 947 84	0.1 406 93	0.1 663 76	0.4 273 46	0.6 504 44	0.7 121 32	0.2 612 07	0.4 075 64	

氧化铝	-0.	0.3	0.5	0.5	0.6		0.6	0.2	0.4	0.3	0.6	0.5	-0.	0.1
(Al ₂ O ₃)	830	445	908	363	977	1	894	096	224	805	760	948	162	468
	732	63	56	69	63		65	59	39	76	68	92	773	29
氧化铁	-0.	-0.	0.3	0.4	0.5	0.6		0.5	0.1	0.4	0.7	0.5	-0.	0.2
(Fe ₂ O ₃)	687	023	357	570	947	894	1	256	509	517	651	905	219	324
	378	046	81	84	84	65		86	23	93	34	44	301	85
氧化铜	0.4	-0.	0.2	0.4	0.1	0.2	0.5		0.1	0.5	0.2	0.1	-0.	0.3
(CuO)	651	022	506	234	406	096	256	1	646	208	133	868	360	012
	89	229	7	72	93	59	86		54	56	44	49	503	89
氧化铅	-0.	0.3	0.2	0.3	0.1	0.4	0.1	0.1		0.6	-0.	0.1	-0.	-0.
(PbO)	398	591	752	676	663	224	509	646	1	317	014	882	133	243
	057	65	14	6	76	39	23	54		91	095	55	216	74
氧化钡	-0.	-0.	0.0	0.0	0.4	0.3	0.4	0.5	0.6		0.3	0.5	-0.	-0.
(BaO)	344	213	926	783	273	805	517	208	317	1	776	652	118	216
	856	569	29	3	46	76	93	56	91		22	44	282	415
五氧化	-0.	-0.	0.1	-0.	0.6	0.6	0.7	0.2	-0.	0.3		0.7	0.0	-0.
	447	185	711	031	504	760	651	133	014	776	1	405	646	021
化	796	782	73	251	44	68	34	44	095	22		87	36	063

二 磷 (P 2O 5) 氧 化 铈 (S rO) 氧 化 锡 (S n O2) 二 氧 化 硫 (S O2)	-0. 525 401	-0. 181 886	0.4 208 29	0.0 346 59	0.7 121 32	0.5 948 92	0.5 905 44	0.1 868 49	0.1 882 55	0.5 652 44	0.7 405 87	0.2 1 402 18	-0. 048 54	
	0.0 485 81	-0. 106 211	0.1 419 2	-0. 290 069	0.2 612 07	-0. 162 773	-0. 219 301	-0. 360 503	-0. 133 216	-0. 118 282	0.0 646 36	0.2 402 18	-0. 1 627	
	-0. 357 223	-0. 194 33	0.3 323 95	0.4 289 8	0.4 075 64	0.1 468 29	0.2 324 85	0.3 012 89	-0. 243 74	-0. 216 415	-0. 021 063	-0. 048 54	-0. 107 627	1

铅钡玻璃各化学成分间 Pearson 系数:

二 氧 化 硅 (Si O2)	氧 化 钠 (Na 2O)	氧 化 钾 (K2 O)	氧 化 钙 (Ca O)	氧 化 镁 (Mg O)	氧 化 铝 (Al2 O3)	氧 化 铁 (Fe2 O3)	氧 化 铜 (Cu O)	氧 化 铅 (Pb O)	氧 化 钡 (Ba O)	五 氧 化 磷 (P2 O5)	氧 化 铈 (Sr O)	氧 化 锡 (Sn O2)	二 氧 化 硫 (SO 2)
--------------------------------	---------------------------	--------------------------	--------------------------	--------------------------	----------------------------	----------------------------	--------------------------	--------------------------	--------------------------	--------------------------------	--------------------------	---------------------------	-------------------------------

二 氧 化 硅 (S iO ₂)		0.3	0.0	-0.	0.0	0.4	0.0	-0.	-0.	-0.	-0.	-0.	0.0	-0.
	1	582	808	485	870	018	906	350	736	445	564	512	827	389
		53	51	343	03	22	01	002	001	654	051	811	87	745
氧 化 钠 (N a ₂ O)	0.3		-0.	-0.	0.0	0.1	-0.	-0.	-0.	-0.	-0.	-0.	-0.	-0.
	582	1	065	369	248	025	238	051	345	063	390	099	085	132
	53		782	802	52	57	818	436	148	424	028	764	734	906
氧 化 钾 (K ₂ O)	0.0	-0.		0.1	0.2	0.3	0.2	-0.	-0.	-0.	-0.	-0.	0.1	0.0
	808	065	1	043	691	066	596	138	123	049	088	130	920	068
	51	782		65	83	31	05	779	359	729	87	222	73	68
氧 化 钙 (C a O)	-0.	-0.	0.1		0.4	0.1	0.3	-0.	0.3	-0.	0.5	0.1	0.2	0.1
	485	369	043	1	197	360	825	076	514	117	319	790	136	070
	343	802	65		08	33	87	315	88	905	97	92	27	07
氧 化 镁 (M g O)	0.0	0.0	0.2	0.4		0.4	0.2	-0.	-0.	-0.	0.2	0.0	0.2	-0.
	870	248	691	197	1	498	964	309	035	455	810	859	455	267
	03	52	83	08		99	17	181	109	692	42	06	1	127
氧 化 铝	0.4	0.1	0.3	0.1	0.4		0.2	-0.	-0.	-0.	-0.	-0.	0.4	-0.
	018	025	066	360	498	1	315	282	442	341	074	156	687	206
	22	57	31	33	99		47	491	726	212	941	277	82	365

(A l ₂ O ₃) 氧化 铁 (F e ₂ O ₃) 氧化 铜 (C u O) 氧化 铅 (P b O) 氧化 钡 (B a O) 五 氧 化 二 磷 (P	0.0 906 01	-0. 238 818	0.2 596 05	0.3 825 87	0.2 964 17	0.2 315 47		-0. 259 552	-0. 057 742	-0. 288 307	0.1 348 38	-0. 082 554	0.2 311 02	-0. 177 074
		-0. 350 002	-0. 051 436	-0. 138 779	-0. 076 315	-0. 309 181	-0. 282 491	-0. 259 552		-0. 094 587	0.7 337 78	0.0 785 12	0.2 061 4	-0. 180 737
	-0. 736 001	-0. 345 148	-0. 123 359	0.3 514 88	-0. 035 109	-0. 442 726	-0. 057 742	-0. 094 587		-0. 132 954	0.3 204 91	0.4 452	-0. 123 344	-0. 062 02
	-0. 445 654	-0. 063 424	-0. 049 729	-0. 117 905	-0. 455 692	-0. 341 212	-0. 288 307	0.7 337 78	-0. 132 954		-0. 001 935	0.1 569 09	-0. 080 524	0.6 340 43
	-0. 564 051	-0. 390 028	-0. 088 87	0.5 319 97	0.2 810 42	-0. 074 941	0.1 348 38	0.0 785 12	0.3 204 91	-0. 001 935		0.3 053 4	-0. 075 436	0.1 844 55

2O														
s)														
氧														
化														
锶	-0.	-0.	-0.	0.1	0.0	-0.	-0.	0.2		0.1	0.3		-0.	0.1
(S	512	099	130	790	859	156	082	061	0.4	569	053	1	041	787
rO	811	764	222	92	06	277	554	4	452	09	4		565	51
)														
氧														
化														
锡	0.0	-0.	0.1	0.2	0.2	0.4	0.2	-0.	-0.	-0.	-0.	-0.		-0.
(S	827	085	920	136	455	687	311	180	123	080	075	041	1	072
n	87	734	73	27	1	82	02	737	344	524	436	565		601
O ₂														
)														
二														
氧														
化	-0.	-0.	0.0	0.1	-0.	-0.	-0.	0.2	-0.	0.6	0.1	0.1	-0.	
硫	389	132	068	070	267	206	177	455	062	340	844	787	072	1
(S	745	906	68	07	127	365	074	23	02	43	55	51	601	
O ₂														
)														

Pearson 相关分析代码:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
plt.rcParams['font.sans-serif']='SimHei'
plt.rcParams['axes.unicode_minus']=False
data1 = pd.read_excel('gao.xlsx')
data2 = pd.read_excel('qian.xlsx')
print(data1.corr())
print(data1.corr())
```

