赛区评阅编号（由赛区组委会填写）：

# 2022 年高教社杯全国大学生数学建模竞赛

# 承　诺　书

我们仔细阅读了《全国大学生数学建模竞赛章程》和《全国大学生数学建模竞赛参赛规则》（以下简称 “竞赛章程和参赛规则”，可从 http://www.mcm.edu.cn 下载）。

我们完全清楚，在竞赛开始后参赛队员不能以任何方式，包括电话、电子邮件、“贴吧”、QQ 群、微信群等，与队外的任何人（包括指导教师）交流、讨论与赛题有关的问题；无论主动参与讨论还是被动接收讨论信息都是严重违反竞赛纪律的行为。

**我们以中国大学生名誉和诚信郑重承诺，严格遵守竞赛章程和参赛规则，以保证竞赛的公正、公平性。如有违反竞赛章程和参赛规则的行为，我们将受到严肃处理。**

我们授权全国大学生数学建模竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

我们参赛选择的题号（从 A/B/C/D/E 中选择一项填写）：C

我们的报名参赛队号（12 位数字全国统一编号）：90

参赛学校（完整的学校全称，不含院系名）：浙江工商大学

参赛队员 (打印并签名) ：1.周洋

2.胡佳伟

3.吴金波

指导教师或指导教师组负责人 (打印并签名)：杨晓蓉

（指导教师签名意味着对参赛队的行为和论文的真实性负责）

日期：2022 年 9 月 24 日

**（请勿改动此页内容和格式。此承诺书打印签名后作为纸质论文的封面，注意电子版论文中不得出现此页。以上内容请仔细核对，如填写错误，论文可能被取消评奖资格。）**

赛区评阅编号：
（由赛区填写）

全国评阅编号：
（全国组委会填写）

# 2022 年高教社杯全国大学生数学建模竞赛

## 编 号 专 用 页

赛区评阅记录（可供赛区评阅时使用）：

| | | | | | | |
|---|---|---|---|---|---|---|
| 评阅人 | | | | | | |
| 备注 | | | | | | |

送全国评阅统一编号:
（赛区组委会填写）

**（请勿改动此页内容和格式。此编号专用页仅供赛区和全国评阅使用，参赛队打印后装订到纸质论文的第二页上。注意电子版论文中不得出现此页。）**

# 题目：古代玻璃制品的成分分析与鉴别


**摘　　要：**

古代玻璃制品作为丝绸之路中常见的贸易货物，见证了中西方交流的盛世，但经过长年累月的风化侵蚀，其内部化学成分逐渐发生了转变，不同玻璃类型和不同风化程度的玻璃制品均有着不同的外部与内部表现，因此对其内部化学成分的研究，对于探讨我国玻璃技术的起源和发展，以及对其它玻璃制品进行判别有着重要的历史意义和现实意义。本文针对已采集的玻璃制品的相关数据，建立相应模型，通过可视化手段，以研究玻璃制品的化学成分的影响因素以及不同玻璃类型的划分依据，从而对未知玻璃类型进行预测，从而解答问题。

**针对问题一**，对玻璃制品表面风化与玻璃制品、纹饰和颜色分别进行相关性分析，并采用皮尔逊卡方检验对对其相关性检验，其中，玻璃类型与表面风化相关性显著。根据题意，将样本数据分为:风化高钾、无风化高钾、风化铅钡、无风化铅钡，通过数据可视化，分析各类数据样本分布。为预测风化前样本化学成分含量，先用逻辑回归筛选出主要成分与次要成分，**主要成分为二氧化硅(SiO2)，氧化钾(K2O)，氧化铁(Fe2O3)，氧化铅(PbO)，氧化钡(BaO)，五氧化二磷(P2O5)。其中主要成分采用逐步回归进行预测，次要成分采用均值法进行预测。**

**针对问题二**，将表单 1 与表单 2 根据文物采样点进行合并，并根据合并得到的数据集分别筛选出高钾玻璃和铅钡玻璃的数据集，从而建立**随机森林分类模型**，模型预测精确度为 100%，预测精度优异。根据分类模型的变量重要性进行**降维**，选取重要性前六的变量进行**聚类**，分别得到其聚类结果，并根据聚类结果的合理性进行分析。

**针对问题三**，根据问题二所建立的随机森林分类模型，对未知玻璃文物化学成分进行玻璃性质的分析预测，得到预测结果并进行敏感性分析，**其中准确率（precision）为 1，召回率（recall）为 1，f1-score 为 1，模型预测结果机器优异。预测结果如下：高钾、铅钡、铅钡、铅钡、铅钡、高钾、高钾、铅钡。**

**针对问题四**，本文针对高钾玻璃数据集和铅钡玻璃数据集作相关性分析，根据化学成分变量的性质，采用**皮尔逊相关系数**来进行相关性分析，观察各类分布情况，并通过绘制**热力图**观察两类关联关系的显著差异。



**关键词：**相关性；分类模型；逻辑斯蒂模型；随机森林；皮尔逊相关系数

目　　录

# 1 问题重述

## 1.1 背景知识

玻璃是人类最早发明的人工材料之一，其发展历史源远流长。于我国而言，玻璃出现的时间晚于世界对玻璃的发现时间。在很长的一段时间内，我国古代先民将其置于宝物的位置，如琉璃。缘于我国古代人民对玻璃的喜爱，与玻璃相关的工艺技术也陆续得到了快速发展，同时玻璃也呈现出了繁荣发展的状态。

玻璃的主要原料是石英砂，主要化学成分是二氧化硅（SiO2）。由于纯石英砂的熔点较高，为了降低熔化温度，在炼制时需要添加助熔剂。古代常用的助熔剂有草木灰、天然泡碱、硝石和铅矿石等，并添加石灰石作为稳定剂，石灰石煅烧以后转化为氧化钙（CaO）。添加的助熔剂不同，其主要化学成分也不同。例如，铅钡玻璃在烧制过程中加入铅矿石作为助熔剂，其氧化铅（PbO）、氧化钡（BaO）的含量较高，通常被认为是我国自己发明的玻璃品种，楚文化的玻璃就是以铅钡玻璃为主。钾玻璃是以含钾量高的物质如草木灰作为助熔剂烧制而成的，主要流行于我国岭南以及东南亚和印度等区域。

古代玻璃极易受埋藏环境的影响而风化。在风化过程中，内部元素与环境元素进行大量交换，导致其成分比例发生变化，从而影响对其类别的正确判断。

## 1.2 要解决的问题。

现需根据题目给出的背景和数据，需要解决以下四个问题。

问题一：对这些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。

# 2 模型假设和符号说明。

## 2.1 模型假设

本文做出如下假定：

一、玻璃文物在短时间内化学成分含量的变化忽略不计。

二、在预测风化前的化学成分含量中，不考虑玻璃文物风化前在不同时点下的成分含量，即不考虑其受时间变化的影响，而预测的是风化前的一般水平。

三、同一文物上的不同部位可能存在差异，但其性质不存在差异。

四、文物采样点能代表该文物的大部分水平。

## 2.2 符号说明

| 序号 | 变量名称 | 变量描述 |
|---|---|---|
| 1 | $r$ | 相关系数 |
| 2 | $x_{ij}$ | 样本变量值 |
| 3 | $v_i$ | 剩余误差 |
| 4 | $\sigma$ | 标准误差 |
| 5 | $\overline{x_i}$ | 样本均值 |
| 6 | $R^2$ | 相关系数 |
| 7 | $w$ | 权值向量 |
| 8 | $\varepsilon$ | 随机误差 |

# 3 问题一的模型建立与求解

## 3.1 问题一分析

问题一由三小问组成，第一问考察的是表面风化与玻璃类型、纹饰和颜色之间的关系，因此需要对表面风化与其它三个玻璃特征分别建立交叉表，通过可视化来研究其中的数量关系，由于玻璃特征为定类变量，因此需要进行相关性分析以及卡方检验。第二问考察不同玻璃类型和不同表面风化的文物之间的化学成分含量的统计规律，初步根据两类特征建立四张表，作探索性分析观察统计规律，并分别建立分类模型，从而观察不同分类模型下，化学成分变量的重要性。第三问要预测风化前的化学成分含量，初步判断要筛选出两部分变量来进行预测。其中一部分通过均值比对来进行预测，另一部分通过逐步回归来进行预测。

## 3.2 理论基础

**定义 1**：皮尔逊（Pearson）相关系数 $\rho$：

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

其中 $\sigma_x$ 为随机变量 $X$ 的标准差，$\sigma_y$ 为随机变量 $Y$ 的标准差，$\sigma_{xy}$ 为 $X$、$Y$ 的协方差。

**定义 2**：皮尔逊样本相关系数 $r$：

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

其中 $x_i, y_i$ 分别是 $X$、$Y$ 的样本数据，$\bar{x}, \bar{y}$ 分别是 $X$、$Y$ 的样本均值。

**皮尔逊卡方检验**：是由 Karl Pearson 提出，被广泛用于分类变量独立性检验的卡方检验。该检验主要包含以下四个步骤：

（1）给定原假设 $H_0$。

（2）计算卡方检验统计量 $\chi^2$，与自由度。

（3）设定置信水平。

（4）通过卡方自由度的临界值与卡方统计值比较，判断是否拒绝原假设 $H_0$

逻辑回归（logistic regression）是统计学习中的经典分类方法，常用于解决二分类问题。

**定义 3**：逻辑分布（logistic distribution）：随机变量 $X$ 服从逻辑分布，则有以下分布函数与密度函数：

$$F(x) = P(X \le x) = \frac{1}{1 + e^{-(x-\mu)/\gamma}}$$

$$f(x) = F'(x) = \frac{e^{-(x-\mu)/\gamma}}{\gamma(1 + e^{-(x-\mu)/\gamma})^2}$$

**定义 4：**二项逻辑回归模型基本形式如下：

$$P(Y=1|x)=\frac{e^{(w\bullet x+b)}}{1+e^{(w\bullet x+b)}}$$

$$P(Y=0|x)=\frac{1}{1+e^{(w\bullet x+b)}}$$

其中 $x\in R^n$ 是输入，　$Y\in\{0,1\}$ 是输出，$w,b\in R^n$ 是参数，$w$ 为权值向量，$b$ 为偏置，$w\bullet x$ 为 $w$ 和 $x$ 的内积。

**多元回归模型基本形式：**

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon$$

$x$ 为自变量，$y$ 为因变量，$\beta_0$ 为回归常数，$\beta_p$ 为回归系数，$\varepsilon$ 为随机误差。

玻璃表面风化影响因素分析

探求玻璃文物的表面风化与玻璃类型、纹饰和颜色的关系流程图如下：

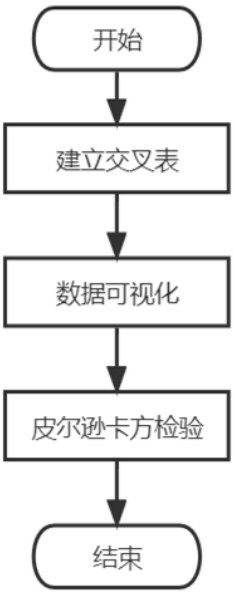


**图 1-1：“表面风化-属性”分析**

通过 SPSS 统计软件得到以下结果：

**表 1-1：“纹饰-表面风化”交叉表**

| 纹饰 | 风化 | 无风化 | 总计 |
|---|---|---|---|
| 纹饰 A | 11 | 11 | 22 |
| 纹饰 B | 6 | 0 | 6 |
| 纹饰 C | 17 | 13 | 30 |
| 总计 | 34 | 24 | 58 |

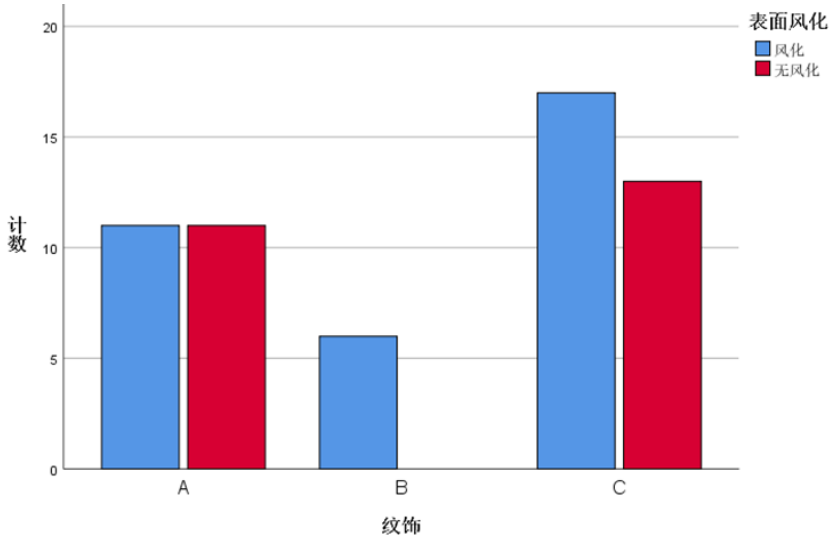


**图 1-2：纹饰-表面风化**

由表 1-1 和图 1-2 可以发现，A 纹饰玻璃制品有无表面风化数量相等，C 纹饰的玻璃制品表面风化的数量相较无风化的多 4 个，B 纹饰的玻璃制品有 6 个表面风化而没有无风化的，初步了解到纹饰 B 与纹饰 C 的玻璃制品表面更容易风化。

**表 1-2：“类型-表面风化”交叉表**

| 类型 | 风化 | 无风化 | 总计 |
|---|---|---|---|
| 高钾 | 6 | 12 | 18 |
| 铅钡 | 28 | 12 | 40 |
| 总计 | 34 | 24 | 58 |

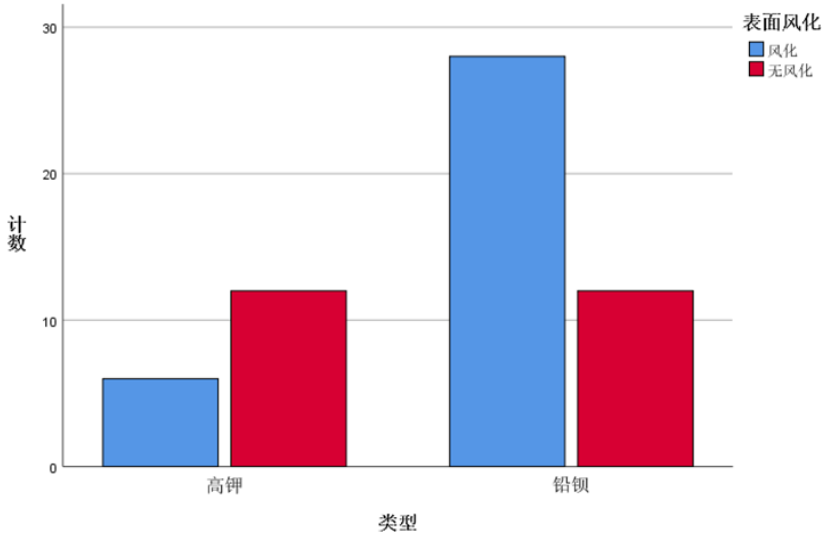


**图 1-3：类型-表面风化**

由表 1-2 和图 1-3 可以发现，高钾类无风化的玻璃制品比有风化的多，相差 6 个，而铅钡类有风化玻璃制品比无风化的多，相差 14 个，两者之间有明显差异，铅钡类的玻璃制品相较高钾类的表面更加容易风化。

**表 1-3：“颜色-表面风化”交叉表**

| 颜色 | 风化 | 无风化 | 总计 |
|---|---|---|---|
| 黑 | 2 | 0 | 2 |
| 蓝绿 | 9 | 6 | 15 |
| 绿 | 0 | 1 | 1 |
| 浅蓝 | 16 | 8 | 24 |
| 浅绿 | 1 | 2 | 3 |
| 深蓝 | 0 | 2 | 2 |
| 深绿 | 4 | 3 | 7 |
| 紫 | 2 | 2 | 4 |
| 总计 | 34 | 24 | 58 |

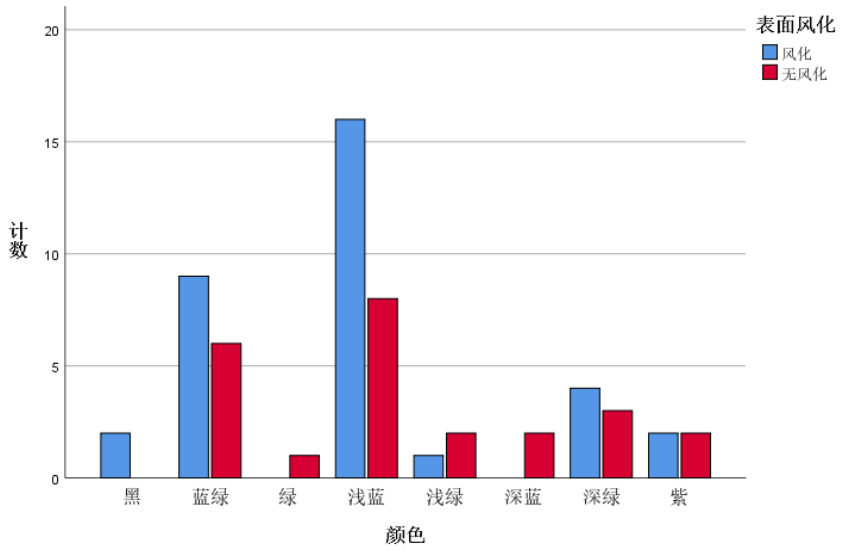


**图 1-4：颜色-表面风化**

由表 1-3 和图 1-4 可以发现，黑色、蓝绿色、浅蓝色和深绿色的玻璃制品都是风化的数量多，而绿色、浅绿色以及深蓝色的玻璃制品都是无风化的数量多，其中黑色、蓝绿色、浅蓝色、深绿色的玻璃制品表面更加容易风化。

**表 1-4：皮尔逊卡方检验**

| 属性 | 值（皮尔逊卡方） | 自由度 | 显著性水平 |
|---|---|---|---|
| 纹饰 | 4.957 | 2 | 0.084 |
| 类型 | 6.88 | 1 | 0.009 |
| 颜色 | 7.234 | 7 | 0.405 |

通过对纹饰、类型、颜色进行皮尔逊卡方检验，发现其中玻璃表面有无风化与纹饰的相关性并不显著，其 p-值为 0.084（>0.05）；玻璃表面有无风化与类型的相关性显著，其 p-值为 0.009（<0.05）；玻璃表面有无风化与类型的相关性不显著，其 p-值为 0.405（>0.05）。

综上所述，纹饰 B 与纹饰 C 的玻璃制品和颜色为黑色、蓝绿色、浅蓝色、深绿色的玻璃制品，和铅钡类的玻璃制品表面相较更容易被风化。另外玻璃制品表面是否风化与其类型相关性显著，在之后的分析中需要着重关注玻璃制品的类型。

### 3.3 化学成分含量统计规律分析

根据题意将样本数据分成四类：风化高钾、无风化高钾、风化铅钡、无风化铅钡。利用数据可视化对该四类分析。

**表 1-5：风化高钾玻璃化学成分含量表**

| | count | mean | std | min | max |
|---|---|---|---|---|---|
| 文物编号 | 6 | 14.50 | 8.07 | 7.00 | 27.00 |
| 二氧化硅(SiO2) | 6 | 93.96 | 1.73 | 92.35 | 96.77 |
| 氧化铝(Al2O3) | 6 | 1.93 | 0.96 | 0.81 | 3.50 |
| 氧化铜(CuO) | 6 | 1.56 | 0.93 | 0.55 | 3.24 |
| 氧化钙(CaO) | 6 | 0.87 | 0.49 | 0.21 | 1.66 |
| 氧化钾(K2O) | 6 | 0.54 | 0.45 | 0.00 | 1.01 |
| 氧化镁(MgO) | 6 | 0.20 | 0.31 | 0.00 | 0.64 |
| 五氧化二磷(P2O5) | 6 | 0.28 | 0.21 | 0.00 | 0.61 |
| 氧化铁(Fe2O3) | 6 | 0.27 | 0.07 | 0.17 | 0.35 |
| 氧化钠(Na2O) | 6 | 0.00 | 0.00 | 0.00 | 0.00 |
| 氧化铅(PbO) | 6 | 0.00 | 0.00 | 0.00 | 0.00 |
| 氧化钡(BaO) | 6 | 0.00 | 0.00 | 0.00 | 0.00 |
| 氧化锶(SrO) | 6 | 0.00 | 0.00 | 0.00 | 0.00 |
| 氧化锡(SnO2) | 6 | 0.00 | 0.00 | 0.00 | 0.00 |
| 二氧化硫(SO2) | 6 | 0.00 | 0.00 | 0.00 | 0.00 |

**表 1-6：无风化高钾玻璃化学成分含量表**

| | count | mean | std | min | max |
|---|---|---|---|---|---|
| 二氧化硅(SiO2) | 12 | 68.13 | 8.61 | 59.81 | 87.05 |
| 文物编号 | 12 | 9.50 | 7.10 | 1.00 | 21.00 |
| 氧化钾(K2O) | 12 | 8.76 | 3.91 | 0.00 | 14.52 |
| 氧化钙(CaO) | 12 | 4.61 | 3.25 | 0.00 | 8.27 |
| 氧化铝(Al2O3) | 12 | 6.11 | 3.14 | 0.00 | 11.15 |
| 氧化铁(Fe2O3) | 12 | 1.78 | 1.66 | 0.00 | 6.04 |
| 氧化铜(CuO) | 12 | 2.15 | 1.53 | 0.00 | 5.09 |
| 五氧化二磷(P2O5) | 12 | 1.31 | 1.48 | 0.00 | 4.50 |
| 氧化钠(Na2O) | 12 | 0.63 | 1.19 | 0.00 | 3.38 |
| 氧化钡(BaO) | 12 | 0.60 | 0.98 | 0.00 | 2.86 |
| 氧化锡(SnO2) | 12 | 0.20 | 0.68 | 0.00 | 2.36 |
| 氧化镁(MgO) | 12 | 1.15 | 0.59 | 0.00 | 1.98 |
| 氧化铅(PbO) | 12 | 0.43 | 0.58 | 0.00 | 1.62 |
| 二氧化硫(SO2) | 12 | 0.10 | 0.19 | 0.00 | 0.47 |
| 氧化锶(SrO) | 12 | 0.04 | 0.05 | 0.00 | 0.12 |

表 1-7：风化铅钡玻璃化学成分含量表

| | count | mean | std | min | max |
|---|---|---|---|---|---|
| 二氧化硅(SiO2) | 36.00 | 33.61 | 17.22 | 3.72 | 68.08 |
| 文物编号 | 36.00 | 38.31 | 15.29 | 2.00 | 58.00 |
| 氧化铅(PbO) | 36.00 | 36.87 | 15.16 | 12.31 | 70.21 |
| 氧化钡(BaO) | 36.00 | 10.49 | 8.87 | 0.00 | 35.45 |
| 五氧化二磷(P2O5) | 36.00 | 4.16 | 4.15 | 0.00 | 14.13 |
| 二氧化硫(SO2) | 36.00 | 0.99 | 3.61 | 0.00 | 15.95 |
| 氧化铝(Al2O3) | 36.00 | 3.84 | 3.41 | 0.45 | 14.34 |
| 氧化铜(CuO) | 36.00 | 2.00 | 2.49 | 0.00 | 10.57 |
| 氧化钠(Na2O) | 36.00 | 0.95 | 1.92 | 0.00 | 7.92 |
| 氧化钙(CaO) | 36.00 | 2.35 | 1.61 | 0.00 | 6.40 |
| 氧化铁(Fe2O3) | 36.00 | 0.56 | 0.69 | 0.00 | 2.74 |
| 氧化镁(MgO) | 36.00 | 0.70 | 0.66 | 0.00 | 2.73 |
| 氧化锶(SrO) | 36.00 | 0.37 | 0.25 | 0.00 | 1.12 |
| 氧化锡(SnO2) | 36.00 | 0.06 | 0.23 | 0.00 | 1.31 |
| 氧化钾(K2O) | 36.00 | 0.14 | 0.21 | 0.00 | 1.05 |

表 1-8：无风化铅钡玻璃化学成分含量表

| | count | mean | std | min | max |
|---|---|---|---|---|---|
| 二氧化硅(SiO2) | 13.00 | 53.44 | 14.59 | 31.94 | 75.51 |
| 文物编号 | 13.00 | 35.77 | 9.94 | 20.00 | 55.00 |
| 氧化铅(PbO) | 13.00 | 23.59 | 9.09 | 9.30 | 39.22 |
| 氧化钡(BaO) | 13.00 | 10.50 | 6.95 | 3.42 | 26.23 |
| 氧化铜(CuO) | 13.00 | 1.56 | 2.49 | 0.00 | 8.46 |
| 五氧化二磷(P2O5) | 13.00 | 0.90 | 1.57 | 0.00 | 5.75 |
| 氧化钠(Na2O) | 13.00 | 0.77 | 1.54 | 0.00 | 4.66 |
| 氧化钙(CaO) | 13.00 | 1.23 | 1.46 | 0.00 | 4.49 |
| 氧化铁(Fe2O3) | 13.00 | 0.93 | 1.45 | 0.00 | 4.59 |
| 氧化铝(Al2O3) | 13.00 | 3.19 | 1.39 | 1.44 | 5.45 |
| 二氧化硫(SO2) | 13.00 | 0.28 | 1.02 | 0.00 | 3.66 |
| 氧化镁(MgO) | 13.00 | 0.49 | 0.55 | 0.00 | 1.67 |
| 氧化钾(K2O) | 13.00 | 0.26 | 0.40 | 0.00 | 1.41 |
| 氧化锶(SrO) | 13.00 | 0.30 | 0.31 | 0.00 | 0.91 |
| 氧化锡(SnO2) | 13.00 | 0.06 | 0.16 | 0.00 | 0.44 |

在风化高钾玻璃中，二氧化硅所占的化学成分含量最高，平均值占比有 93.96，其中最大占比有 96.77%，最小占比有 92.35%，标准差为 1.73，氧化铝所占的化学含量第二，平均值为 1.93%，最大值为 3.50%，最小值为 0.81%，标准差为 0.96，其中不含有氧化钠，氧化铅，氧化钡，氧化锶，氧化锡，氧化硫成分。

在无风化高钾玻璃中，二氧化硅所占的化学成分含量最高，平均值占比有 68.13，其中最大占比有 87.05%，最小占比有 59.81%，标准差为 8.61，氧化钾所占的化学含量第二，平均值为 8.76%，最大值为 14.52%，最小值为 0，标准差为 3.91，所占化学成分

含量最少的是氧化锶，平均值为 0.04%，最大值为 0.12%，最小值为 0，标准差为 0.05。

在风化铅钡玻璃中，氧化铅所占的化学成分含量最高，平均值占比有 36.87，其中最大占比有 70.21%，最小占比有 12.31%，标准差为 15.16，二氧化硅所占的化学含量第二，平均值为 33.61%，最大值为 68.08%，最小值为 3.72%，标准差为 17.22，所占化学成分含量最少的是氧化锡，平均值为 0.06%，最大值为 1.31%，最小值为 0，标准差为 0.23。

在无风化铅钡玻璃中，二氧化硅所占的化学成分含量最高，平均值占比有 53.44，其中最大占比有 75.51%，最小占比有 31.94%，标准差为 14.59，氧化铅所占的化学含量第二，平均值为 23.59%，最大值为 39.22%，最小值为 9.30%，标准差为 9.09，所占化学成分含量最少的是氧化锡，平均值为 0.06%，最大值为 0.44%，最小值为 0，标准差为 0.16。

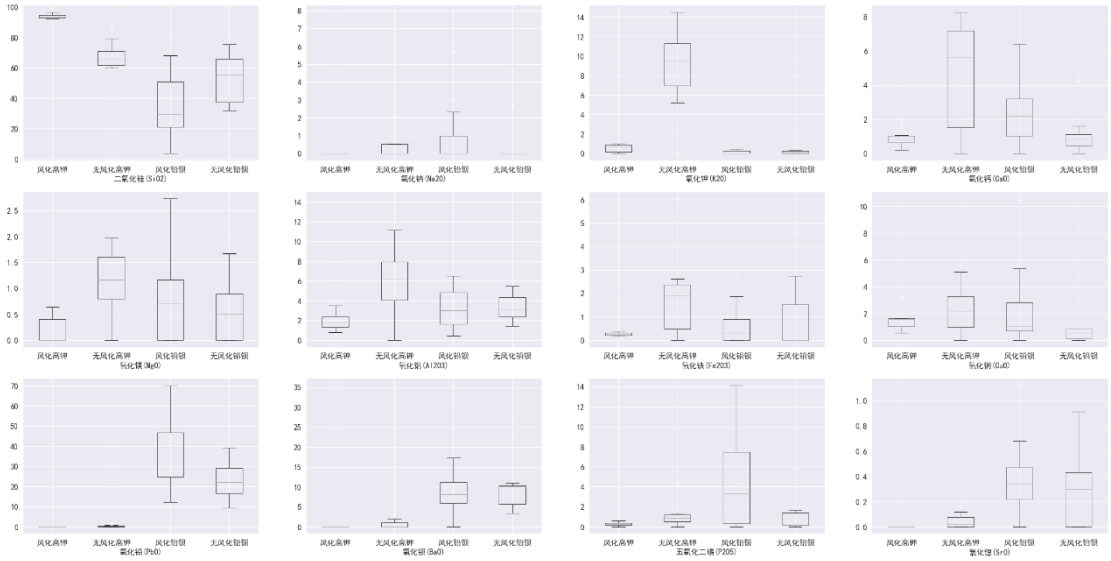

**图 1-5：玻璃样本化学元素箱型图**

根据图 1-5，风化高钾类的玻璃制品样本量较少，各类化学成分含量分布都比较集中；在无风化高钾中，二氧化硅（SiO2）、氧化钠(Na2O)、氧化铅(PbO)、氧化钡(BaO)、五氧化二磷(P2O5)、氧化锶(SrO)分布比较集中，其他化学元素较为分散；在风化铅钡类中，氧化钠(Na2O)、氧化钾(K2O)分布比较集中，其他化学元素分布比较分散；在无风化铅钡类中，氧化钠(Na2O)、氧化钾(K2O)、氧化钡(BaO)、五氧化二磷(P2O5)、氧化钙(CaO)分布比较集中，其他化学元素分布比较分散。

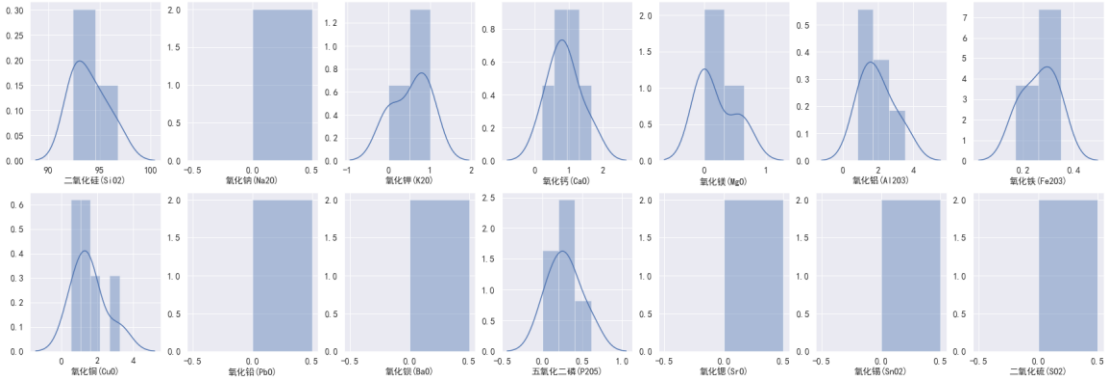

**图 1-6：风化高钾玻璃化学成分方差分布图**

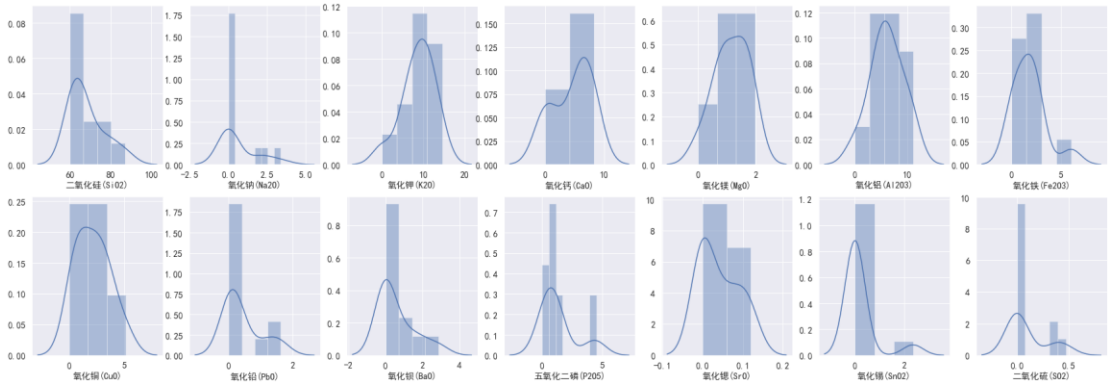


**图 1-7：无风化高钾玻璃化学成分方差分布图**

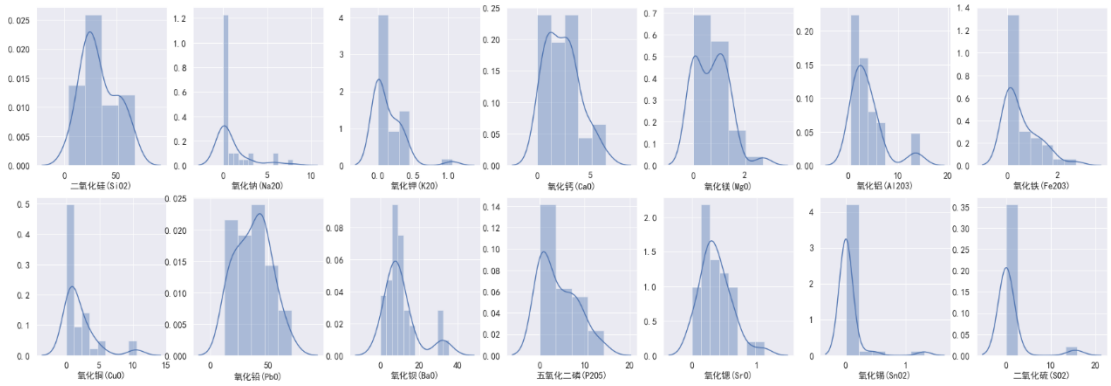


**图 1-8：风化铅钡玻璃化学成分方差分布图**

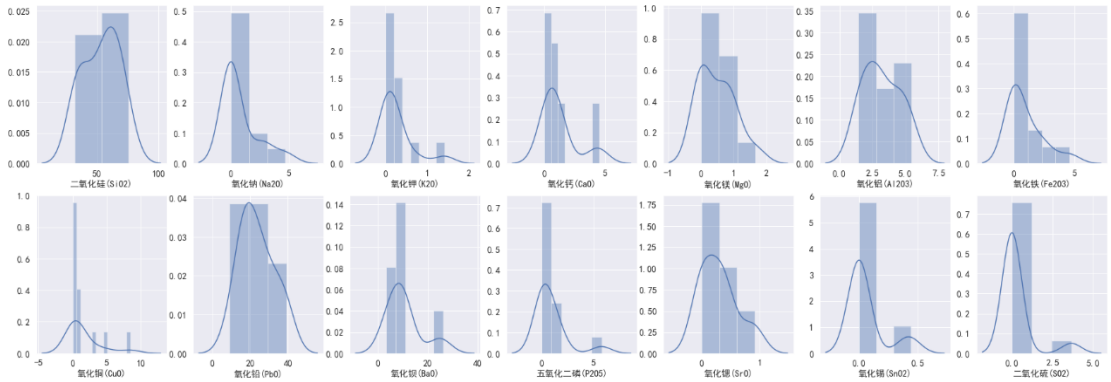


**图 1-9：无风化铅钡玻璃化学成分方差分布图**

结合方差分布图，综合分析发现，风化高钾玻璃与无风化高钾玻璃中二氧化硅的含量都呈现出轻微的右偏分布，在风化高钾玻璃中，氧化铝的含量近似正态分布，且不含有氧化钠，氧化铅，氧化钡，氧化锶，氧化锡，氧化硫成分，在无风化高钾玻璃中，氧化钾含量近似正态分布；风化铅钡玻璃与无风化铅钡玻璃对比发现，风化铅钡玻璃中二氧化硅含量呈现出轻微的右偏分布，氧化铅近似正态分布，且含量最高，无风化铅钡玻璃中，二氧化硅含量轻微左偏分布，氧化铅含量近似正态分布。

### 3.4 风化前化学成分含量预测

建立该预测模型，首先根据不同化学成分的玻璃制品风化过程的“重要性”对化学成分分类，采用逻辑回归模型分出主要成分与次要成分，对主要变量采用“逐步回归法”，建立多元回归模型进行预测，对次要变量使用“均值法”预测估计。预测流程图如下：

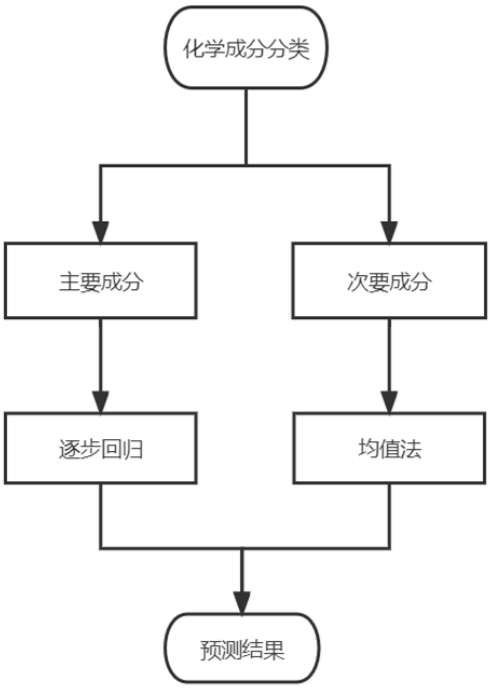


**图 1-10：预测流程**

利用 python 对逻辑回归分类模型求解，得到以下结果：

**表 1-9：主次成分**

| 主要成分 | 次要成分 |
|---|---|
| 二氧化硅(SiO2) | 氧化钠(Na2O) |
| 氧化钾(K2O) | 氧化钙(CaO) |
| 氧化铁(Fe2O3) | 氧化镁(MgO) |
| 氧化铅(PbO) | 氧化铝(Al2O3) |
| 氧化钡(BaO) | 氧化铜(CuO) |
| 五氧化二磷(P2O5) | 氧化锶(SrO) |
| | 氧化锡(SnO2) |
| | 二氧化硫(SO2) |

对次要成分使用均值法进行预测，模型形式如下：

$$\text{风化（预测）}=\frac{\text{无风化（均值）}}{\text{风化前（均值）}}\times\text{风化前（初始）}$$

预测结果如下表：

**表 1-10：次要成分预测结果**

| 氧化钠 (Na2O) | 氧化钙 (CaO) | 氧化镁 (MgO) | 氧化铝 (Al2O3) | 氧化铜 (CuO) | 氧化锶 (SrO) | 氧化锡 (SnO2) | 二氧化硫 (SO2) |
|---|---|---|---|---|---|---|---|
| 0.00 | 6.32 | 0.87 | 3.93 | 3.87 | 0.00 | 0.00 | 0.39 |
| 0.00 | 2.01 | 0.00 | 4.06 | 0.78 | 0.00 | 0.00 | 0.00 |
| 0.00 | 5.87 | 1.11 | 5.50 | 5.09 | 0.10 | 0.00 | 0.00 |
| 0.00 | 7.12 | 1.56 | 6.44 | 2.18 | 0.00 | 0.00 | 0.36 |
| 0.00 | 7.35 | 1.77 | 7.50 | 3.27 | 0.06 | 0.00 | 0.47 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 1.98 | 11.15 | 2.51 | 0.11 | 0.00 | 0.00 |
| 0.00 | 5.41 | 1.73 | 10.05 | 2.18 | 0.12 | 0.00 | 0.00 |
| 3.38 | 8.23 | 0.66 | 9.23 | 0.47 | 0.00 | 0.00 | 0.00 |
| 2.10 | 8.27 | 0.52 | 6.18 | 1.07 | 0.04 | 0.00 | 0.00 |
| 2.12 | 0.00 | 0.85 | 0.00 | 1.09 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 1.53 | 3.05 | 0.00 | 0.07 | 2.36 | 0.00 |
| 0.00 | 0.00 | 0.00 | 5.45 | 4.78 | 0.00 | 0.00 | 0.00 |
| 0.00 | 4.71 | 1.22 | 6.19 | 3.28 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.47 | 0.00 | 1.59 | 8.46 | 0.91 | 0.00 | 0.00 |
| 0.00 | 1.34 | 1.00 | 4.70 | 0.33 | 0.12 | 0.23 | 0.00 |
| 0.92 | 2.98 | 1.49 | 14.34 | 0.74 | 0.25 | 0.00 | 0.00 |
| 0.00 | 4.49 | 0.98 | 4.35 | 0.00 | 0.35 | 0.40 | 0.00 |
| 0.00 | 4.24 | 0.51 | 3.86 | 0.00 | 0.48 | 0.44 | 0.00 |
| 0.00 | 1.60 | 0.89 | 3.11 | 0.44 | 0.30 | 0.00 | 0.00 |
| 0.00 | 0.46 | 0.00 | 2.36 | 0.11 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.64 | 1.00 | 2.35 | 0.47 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.38 | 0.00 | 1.44 | 0.16 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.89 | 0.00 | 2.72 | 3.01 | 0.31 | 0.00 | 3.66 |
| 5.74 | 0.79 | 1.09 | 3.53 | 2.67 | 0.35 | 0.00 | 0.00 |
| 5.68 | 0.00 | 1.16 | 5.66 | 2.72 | 0.00 | 0.00 | 0.00 |
| 3.06 | 2.14 | 0.00 | 12.69 | 0.43 | 0.26 | 0.00 | 0.00 |
| 2.66 | 0.84 | 0.74 | 5.00 | 0.53 | 0.23 | 0.00 | 0.00 |
| 0.00 | 0.00 | 1.67 | 4.79 | 0.77 | 0.43 | 0.00 | 0.00 |
| 4.66 | 0.87 | 0.61 | 3.06 | 0.65 | 0.85 | 0.00 | 0.00 |
| 0.00 | 2.08 | 1.20 | 6.50 | 0.45 | 0.30 | 0.00 | 0.00 |
| 0.00 | 3.12 | 0.54 | 4.16 | 0.70 | 0.23 | 0.00 | 0.00 |
| 3.04 | 0.78 | 1.14 | 6.06 | 0.54 | 0.27 | 0.00 | 0.00 |
| 2.71 | 1.13 | 0.00 | 1.45 | 0.86 | 0.00 | 0.00 | 0.00 |

我们对主要成分进行逐步回归，首先把次要成分作为自变量，以一个主要成分为因变量，建立如下形式的多元回归模型。之后再将该主要成分加入到自变量中，以下一个主要成分为因变量，以此类推，直到所有成分都被添加到回归模型当中。逐步回归模型流程图如下：

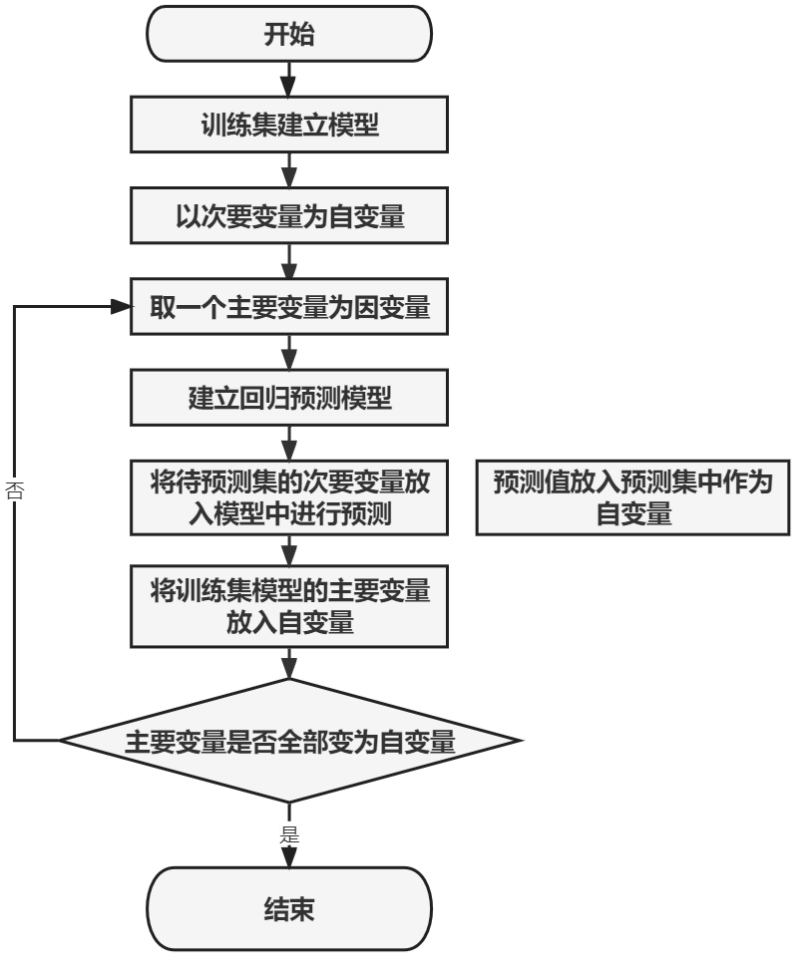


**图 1-11 逐步回归模型流程图**

**表 1-11：主要成分预测结果**

| 二氧化硅 (Si02) | 氧化钾 (K20) | 氧化铁 (Fe203) | 氧化铁 (Fe203) | 氧化钡 (Ba0) | 五氧化二磷 (P205) |
|---|---|---|---|---|---|
| 61.68 | 9.71 | 1.96 | 2.55 | 3.50 | 1.29 |
| 77.64 | 6.53 | 0.97 | 0.00 | 0.00 | 1.37 |
| 57.35 | 9.43 | 2.25 | 1.85 | 6.32 | 1.71 |
| 65.49 | 9.89 | 2.58 | 2.23 | 0.00 | 1.40 |
| 61.91 | 10.14 | 2.79 | 1.50 | 0.50 | 1.63 |
| 77.71 | 5.31 | 2.34 | 0.00 | 0.16 | 2.96 |
| 67.51 | 8.61 | 2.75 | 0.00 | 0.00 | 2.07 |
| 65.75 | 11.16 | 1.26 | 1.03 | 0.00 | 0.00 |
| 63.31 | 10.94 | 1.47 | 4.62 | 0.00 | 0.09 |
| 77.82 | 5.93 | 0.49 | 0.27 | 0.00 | 0.24 |
| 75.20 | 8.98 | 0.30 | 2.60 | 1.02 | 1.22 |
| 52.70 | 0.00 | 0.57 | 14.75 | 16.33 | 2.64 |
| 66.14 | 8.52 | 2.16 | 0.00 | 2.21 | 1.80 |
| 28.37 | 0.00 | 0.89 | 28.51 | 26.98 | 1.75 |

| | | | | | |
|---|---|---|---|---|---|
| 59.67 | 0.06 | 1.21 | 21.61 | 6.07 | 1.73 |
| 57.01 | 0.77 | 1.99 | 15.08 | 5.65 | 2.55 |
| 50.74 | 2.05 | 1.54 | 29.03 | 6.14 | 1.05 |
| 49.26 | 1.86 | 1.22 | 30.72 | 7.58 | 0.85 |
| 56.64 | 0.00 | 1.32 | 25.35 | 6.61 | 1.34 |
| 63.54 | 0.00 | 0.47 | 20.84 | 6.74 | 1.50 |
| 62.12 | 0.00 | 1.13 | 21.22 | 5.61 | 1.62 |
| 63.24 | 0.00 | 0.42 | 21.63 | 6.87 | 1.39 |
| 60.93 | 0.34 | 0.00 | 16.56 | 10.04 | 1.30 |
| 47.70 | 1.08 | 0.00 | 21.65 | 12.02 | 0.00 |
| 54.49 | 0.86 | 0.00 | 14.25 | 10.87 | 0.08 |
| 57.47 | 0.84 | 0.22 | 14.79 | 8.30 | 1.32 |
| 57.67 | 0.04 | 0.37 | 20.75 | 7.33 | 0.64 |
| 57.69 | 0.00 | 1.76 | 23.46 | 6.48 | 1.78 |
| 47.11 | 0.10 | 0.00 | 29.68 | 10.27 | 0.00 |
| 56.83 | 0.00 | 1.71 | 22.69 | 5.91 | 1.81 |
| 54.26 | 0.76 | 1.28 | 25.03 | 7.31 | 1.38 |
| 57.32 | 0.05 | 0.56 | 20.15 | 6.79 | 0.64 |
| 58.01 | 0.55 | 0.00 | 21.30 | 8.64 | 0.24 |

# 4 问题二的模型建立与求解

## 4.1 问题二分析

问题二由两问组成。第一问考察高钾玻璃和铅钡玻璃的分类规律，考虑到样本数量较少，初步选择随机森林模型建立分类模型，并根据不同分类模型下，分析不同变量对模型的解释程度。第二问要分别对两类玻璃进行亚类划分，根据前一问建立的随机森林模型，通过模型导出的变量重要性筛选出重要性程度最高的几个变量进行降维，并根据筛选出的变量建立聚类模型，并根据聚类结果来分析模型的稳定性。

## 4.2 随机森林理论和建模过程

随机森林（Random Forest，简称 RF）是一种新兴起、高度灵活的机器学习算法，拥有广泛的应用前景，在大量分类以及回归问题中具有极好的准确率。并且，随机森林算法自带特征筛选机制，即随机森林能够评估各个特征在相应问题上的重要性。基于随机森林的操作变量特征筛选模型中随机森林训练过程包含以下步骤：

Stepl：原始训练集为 N，应用 bootstrap（有放回抽样）法有放回地随机抽取 k 个新的自助样本集，并由此构建 k 棵分类树，每次术被抽到的样本组成了 k 个袋外数据：

Step2：假如特征空间共有口个特征，则在每一轮生成决策树的过程中，从 D 个特征中随机选择 d 个特征（d<D）组成一个新的特征集，通过使用新的特征集来生成决策树，在 k 轮中共生成 k 个相互独立的决策树：

Step3：将生成的多棵树组成随机森林，相互独立的若干棵决策树的重要性是相等的，无需考虑它们的权值。随机森林算法的流程图如图 2-1 所示。

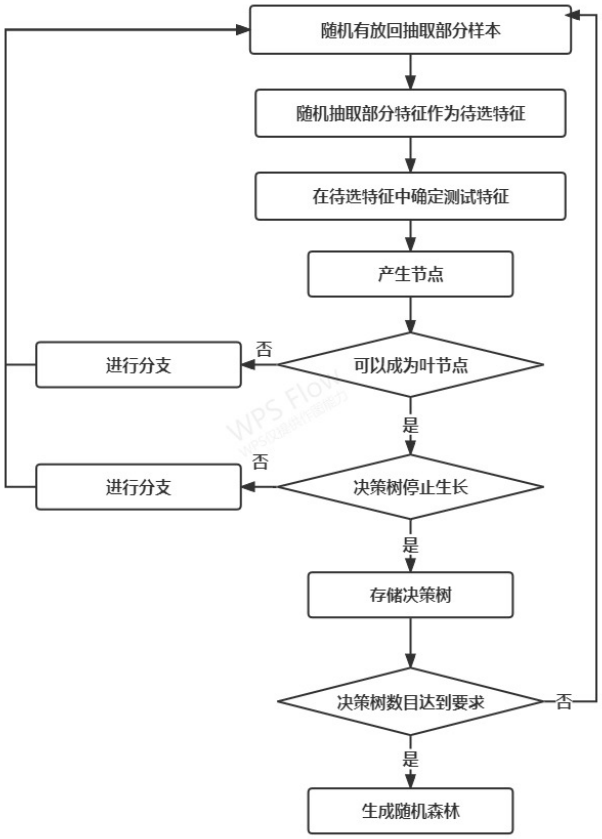

**图 2-1:随机森林实现流程图**

### 4.3 玻璃类型的分类规律

在风化高钾玻璃中二氧化硅所占的化学成分含量最高，氧化铝所占的化学含量第二，其中不含有氧化钠，氧化铅，氧化钡，氧化锶，氧化锡，氧化硫成分。在无风化高钾玻璃中，二氧化硅所占的化学成分含量最高，氧化钾所占的化学含量第二。在风化铅钡玻璃中，氧化铅所占的化学成分含量最高，二氧化硅所占的化学含量第二。在无风化铅钡玻璃中，二氧化硅所占的化学成分含量最高，氧化铅所占的化学含量第二。根据每种类型玻璃的主要化学成分，和各成分重要性排名对玻璃类型进行分类。

### 4.4 基于随机森林模型的降维结果

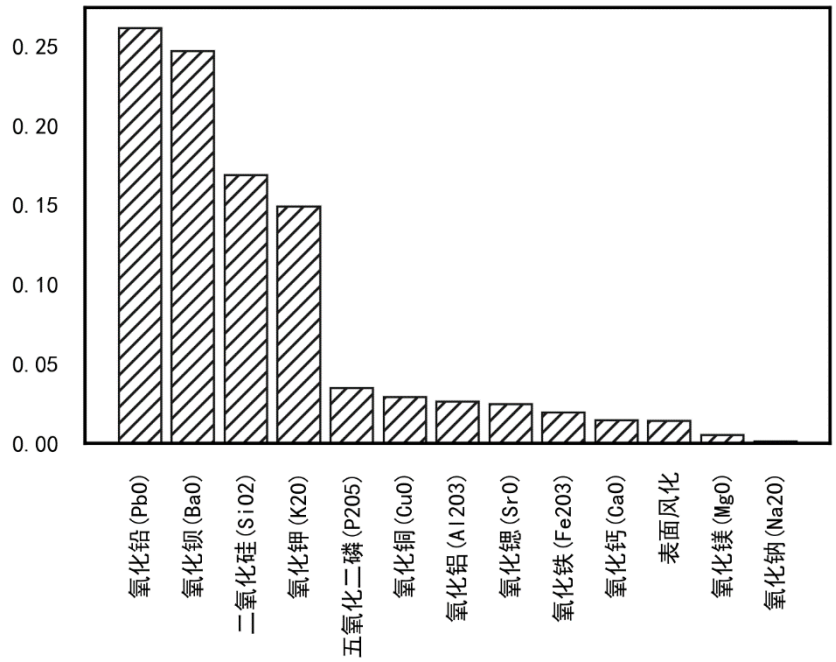


**图 2-2：随机森林变量重要性排名图**

**表 2-1：各成分重要性排名得分表**

| 排序 | 名称 | 得分 |
|---|---|---|
| 1 | 氧化铅(PbO) | 0.261471 |
| 2 | 氧化钡(BaO) | 0.247121 |
| 3 | 二氧化硅(SiO2) | 0.169023 |
| 4 | 氧化钾(K2O) | 0.149290 |
| 5 | 五氧化二磷(P2O5) | 0.034905 |
| 6 | 氧化铜(CuO) | 0.029424 |
| 7 | 氧化铝(Al2O3) | 0.026606 |
| 8 | 氧化锶(SrO) | 0.024891 |
| 9 | 氧化铁(Fe2O3) | 0.019397 |
| 10 | 氧化钙(CaO) | 0.014640 |
| 11 | 表面风化 | 0.014497 |
| 12 | 氧化镁(MgO) | 0.005394 |
| 13 | 氧化钠(Na2O) | 0.001542 |

| | | |
|---|---|---|
| 14 | 二氧化硫(SO2) | 0.001396 |
| 15 | 氧化锡(SnO2) | 0.000403 |

建立随机森林模型进行降维，通过建立随机森林模型得到变量重要性的排序，根据表 1 和图 2 的结果，我们选取前 6 个作为重要性最大的变量，依次为：氧化铅，氧化钡，二氧化硅，氧化钾，五氧化二磷和氧化铜。

### 4.5 基于 k 均值聚类模型的亚类划分

#### 4.5.1 k 均值聚类的含义

k 均值算法属于聚类技术中一种基本的划分方法。具有简单、快速的优点。其基本思想是选取 k 个数据对象作为初始聚类中心，通过迭代把数据对象划分到不同的簇中，使簇内部对象之间的相似度很大，而簇之间对象的相似度很小。算法中参数 k 的值是事先给定的并在数据对象集中随机选取 k 个数据对象作为初始聚类中心。

#### 4.5.2 k 均值聚类的步骤

输入：聚类个数 k 以及包含 n 个数据对象的数据样本集；输出：满足方差最小标准的 k 个聚类；

步骤：

（1）从 n 个数据对象中任意选择 k 个对象作为初始聚类中心；

（2）循环执行（3）到（4）直到每个聚类不再发生变化为止；

（3）根据每个聚类中所有对象的均值（中心对象）计算样本集中每个对象与这些中心对象的距离，并根据最小距离重新对相应对象进行划分；

（4）重新计算每个（有变化）聚类的均值（中心对象）。

#### 4.5.3 聚类个数的选取

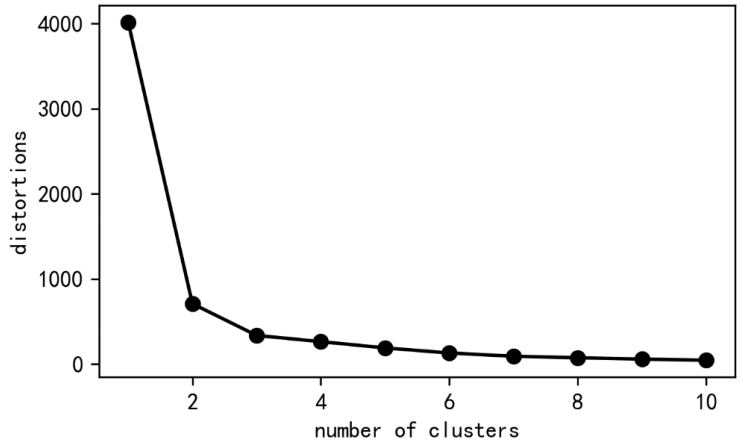


**图 2-3：高钾玻璃手肘图**

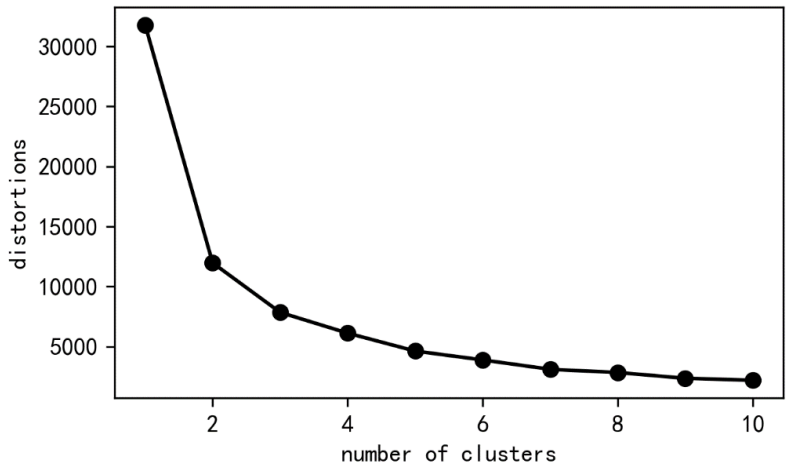


**图 2-4：铅钡玻璃手肘图**

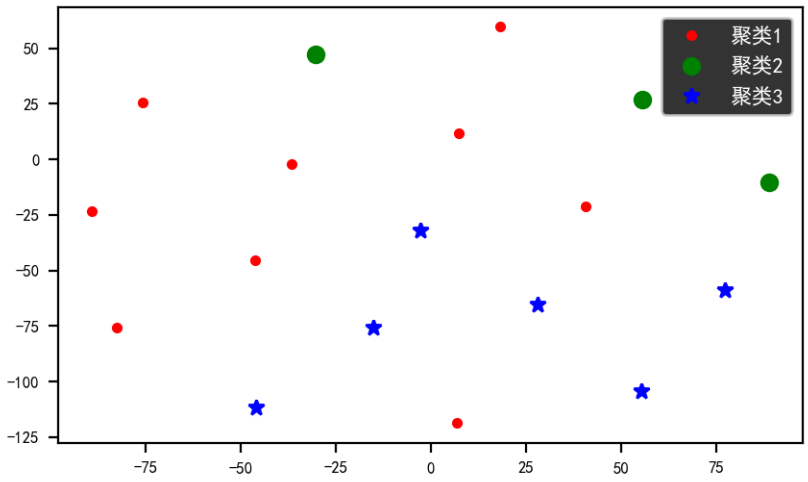


**图 2-5：高钾玻璃聚类图**

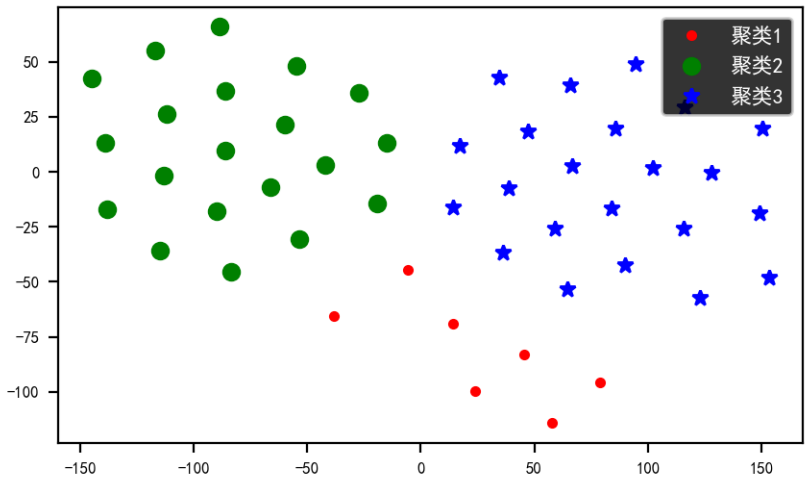


**图 2-6：铅钡玻璃聚类图**

根据高钾玻璃和铅钡玻璃的手肘图和聚类图，可以看出，当把聚类变量选取 3 个时可以达到最好的效果。

### 4.5.4 聚类结果

使用 k 均值聚类和利用手肘图以及聚类图，对“高钾玻璃”“铅钡玻璃”进行分别聚类，聚类结果如表 2-2 和表 2-3 所示。

**表 2-2：高钾玻璃分类结果**

| | 氧化铅 | 氧化钡 | 二氧化硅 | 氧化钾 | 五氧化二磷 | 氧化铜 | 分类值 |
|---|---|---|---|---|---|---|---|
| 0 | 47.43 | 0.00 | 36.28 | 1.05 | 3.57 | 0.27 | 2 |
| 1 | 28.68 | 31.23 | 20.14 | 0.00 | 3.59 | 10.82 | 0 |
| 2 | 32.45 | 30.62 | 4.61 | 0.00 | 7.56 | 3.26 | 0 |
| 3 | 25.39 | 14.61 | 33.59 | 0.21 | 9.38 | 5.12 | 0 |
| 4 | 42.82 | 5.35 | 29.64 | 0.00 | 8.83 | 3.65 | 2 |
| 5 | 9.30 | 23.55 | 37.36 | 0.71 | 5.75 | 4.78 | 0 |
| 6 | 16.98 | 11.86 | 53.79 | 0.00 | 0.00 | 2.99 | 1 |
| 7 | 29.14 | 26.23 | 31.94 | 0.00 | 0.14 | 8.46 | 0 |
| 8 | 31.90 | 6.65 | 50.61 | 0.00 | 0.19 | 1.12 | 1 |
| 9 | 29.53 | 32.25 | 19.79 | 0.00 | 3.13 | 10.98 | 0 |
| 10 | 29.92 | 35.45 | 3.72 | 0.40 | 6.04 | 3.74 | 0 |
| 11 | 17.14 | 4.04 | 68.08 | 0.26 | 1.04 | 0.33 | 1 |
| 12 | 12.31 | 2.03 | 63.30 | 0.30 | 0.41 | 0.74 | 1 |
| 13 | 39.22 | 10.29 | 34.34 | 1.41 | 0.00 | 0.00 | 2 |
| 14 | 37.74 | 10.35 | 36.93 | 0.00 | 1.41 | 0.00 | 2 |
| 15 | 16.55 | 3.42 | 65.91 | 0.00 | 1.62 | 0.44 | 1 |
| 16 | 19.76 | 4.88 | 69.71 | 0.21 | 0.17 | 0.11 | 1 |
| 17 | 16.16 | 3.55 | 75.51 | 0.15 | 0.13 | 0.47 | 1 |
| 18 | 46.55 | 10.00 | 35.78 | 0.25 | 0.34 | 1.57 | 2 |
| 19 | 22.05 | 5.68 | 65.91 | 0.00 | 0.42 | 0.16 | 1 |
| 20 | 41.61 | 10.83 | 39.57 | 0.14 | 0.07 | 0.71 | 2 |
| 21 | 17.24 | 10.34 | 60.12 | 0.23 | 1.46 | 3.01 | 1 |
| 22 | 49.31 | 9.79 | 32.93 | 0.00 | 0.48 | 0.76 | 2 |
| 23 | 61.03 | 7.22 | 26.25 | 0.00 | 1.16 | 0.91 | 2 |
| 24 | 70.21 | 6.69 | 16.71 | 0.00 | 1.77 | 0.00 | 2 |
| 25 | 44.12 | 9.76 | 18.46 | 0.44 | 7.46 | 0.20 | 2 |
| 26 | 21.88 | 10.47 | 51.26 | 0.15 | 0.08 | 2.67 | 1 |
| 27 | 20.12 | 10.88 | 51.33 | 0.35 | 0.00 | 2.72 | 1 |
| 28 | 59.85 | 7.29 | 12.41 | 0.00 | 0.00 | 5.56 | 2 |
| 29 | 44.75 | 3.26 | 21.70 | 0.00 | 12.83 | 1.57 | 2 |
| 30 | 13.61 | 5.22 | 60.74 | 0.20 | 0.00 | 0.43 | 1 |
| 31 | 15.99 | 10.96 | 61.28 | 0.11 | 0.00 | 0.53 | 1 |
| 32 | 25.25 | 10.06 | 55.21 | 0.25 | 0.20 | 0.77 | 1 |
| 33 | 25.40 | 9.23 | 51.54 | 0.29 | 0.10 | 0.65 | 1 |
| 34 | 15.71 | 7.31 | 53.33 | 0.32 | 1.10 | 0.00 | 1 |
| 35 | 34.18 | 6.10 | 28.79 | 0.00 | 11.10 | 0.73 | 2 |
| 36 | 23.02 | 4.19 | 54.61 | 0.30 | 4.32 | 0.45 | 1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **37** | 44.00 | 14.20 | 17.98 | 0.00 | 6.34 | 1.17 | 2 |
| **38** | 30.61 | 6.22 | 45.02 | 0.00 | 6.34 | 0.70 | 1 |
| **39** | 40.24 | 8.94 | 24.61 | 0.00 | 8.10 | 1.42 | 2 |
| **40** | 51.34 | 0.00 | 21.35 | 0.00 | 8.75 | 0.78 | 2 |
| **41** | 47.42 | 8.64 | 25.74 | 0.00 | 5.71 | 0.73 | 2 |
| **42** | 13.66 | 8.99 | 63.66 | 0.11 | 0.00 | 0.54 | 1 |
| **43** | 55.46 | 7.04 | 22.28 | 0.32 | 4.24 | 0.86 | 2 |
| **44** | 58.46 | 0.00 | 17.11 | 0.00 | 14.13 | 1.39 | 2 |
| **45** | 32.92 | 7.95 | 49.01 | 0.00 | 0.35 | 0.86 | 1 |
| **46** | 41.25 | 15.45 | 29.15 | 0.00 | 2.54 | 0.82 | 2 |
| **47** | 45.10 | 17.30 | 25.42 | 0.00 | 0.00 | 1.21 | 2 |
| **48** | 39.35 | 7.66 | 30.39 | 0.34 | 8.99 | 3.25 | 2 |

**表 2-3：铅钡玻璃分类结果**

| | 氧化铅 | 氧化钡 | 二氧化硅 | 氧化钾 | 五氧化二磷 | 氧化铜 | 分类值 |
|---|---|---|---|---|---|---|---|
| **0** | 0.00 | 0.00 | 69.33 | 9.99 | 1.17 | 3.87 | 0 |
| **1** | 0.25 | 0.00 | 87.05 | 5.19 | 0.66 | 0.78 | 1 |
| **2** | 1.41 | 2.86 | 61.71 | 12.37 | 0.70 | 5.09 | 0 |
| **3** | 0.00 | 0.00 | 65.88 | 9.67 | 0.79 | 2.18 | 0 |
| **4** | 0.00 | 0.00 | 61.58 | 10.95 | 0.94 | 3.27 | 0 |
| **5** | 0.20 | 1.38 | 67.65 | 7.37 | 4.18 | 2.51 | 0 |
| **6** | 0.35 | 0.97 | 59.81 | 7.68 | 4.50 | 2.18 | 0 |
| **7** | 0.00 | 0.00 | 92.63 | 0.00 | 0.61 | 3.37 | 2 |
| **8** | 0.00 | 0.00 | 95.02 | 0.59 | 0.35 | 1.61 | 2 |
| **9** | 0.00 | 0.00 | 96.77 | 0.92 | 0.00 | 0.87 | 2 |
| **10** | 0.00 | 0.00 | 94.29 | 1.01 | 0.15 | 1.71 | 2 |
| **11** | 1.62 | 0.00 | 62.47 | 12.28 | 0.16 | 0.47 | 0 |
| **12** | 0.11 | 0.00 | 65.18 | 14.52 | 0.00 | 1.07 | 0 |
| **13** | 0.19 | 0.00 | 60.71 | 5.71 | 0.18 | 1.09 | 0 |
| **14** | 0.00 | 0.00 | 79.46 | 9.42 | 1.36 | 0.00 | 1 |
| **15** | 1.00 | 1.97 | 76.68 | 0.00 | 1.10 | 3.28 | 1 |
| **16** | 0.00 | 0.00 | 92.35 | 0.74 | 0.21 | 0.57 | 2 |
| **17** | 0.00 | 0.00 | 92.72 | 0.00 | 0.36 | 1.60 | 2 |

# 5 问题三的模型建立与求解

## 5.1 问题三分析

问题三主要是根据化学成分预测玻璃类型。首先需要对化学成分的数据进行探索性分析，观察其数据分布，其次根据问题二所建立的随机森林分类模型，将化学成分输入模型中进行分类，得到模型结果，这里鉴别的所属类型主要是指玻璃类型。

## 5.2 玻璃类型预测结果

根据问题二所建立的随机森林分类模型，进行预测，并分析其敏感性。随机森林模型预测有以下优点：

(1)随机森林模型采用了集成算法，其精度要更好。

(2)可以处理非线性数据进行预测分类。

(3)可以直接处理缺省值。

(4)能够处理高维数据，适用于各种数据集。

(5)模型预测结果如下表：

**表 3-1：类型判别**

| 文物编号 | 表面风化 | 预测玻璃类型 |
|---|---|---|
| A1 | 无风化 | 高钾 |
| A2 | 风化 | 铅钡 |
| A3 | 无风化 | 铅钡 |
| A4 | 无风化 | 铅钡 |
| A5 | 风化 | 铅钡 |
| A6 | 风化 | 高钾 |
| A7 | 风化 | 高钾 |
| A8 | 无风化 | 铅钡 |

## 5.3 敏感性分析

令$\{A_{i,j}\}$为预测结果，则有以下四种不同的情况：

TP（真正）:真正的分类结果属于$i$预测的结果也属于$i$，此时对于$A_{i,j}$而言$i=j$。

FN（假负）:真正的分类结果不属于分类$i$，预测的分类结果属于分类$i$。

TN（真负）:真正的分类结果属于分类$i$，预测的结果不属于分类$i$。

FP（假正）:真正的分类结果不属于分类$i$，预测的结果属于分类$i$。

则有准确率（precision）和召回率（recall）表达式如下：

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

$$recall=\frac{TP}{TP+TN}$$

f1-score 可以看作精准率（precision）与召回率（recall）的调和平均数，用于衡量二分类模型的精确度，其表达式如下：

$$f1-score=2\times\frac{precision\times recall}{precision+recall}$$

精确度分析结果如下：

**表 3-2：敏感性分析**

| | precision | recall | f1-score |
|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 |
| 1 | 1.00 | 1.00 | 1.00 |

精确率与召回率越高，预测结果越准确。在表 3-2 当中，准确率与召回率都为 100%，即模型预测效果好，准确率高。

# 6 问题四的模型建立与求解

## 6.1 问题四分析

问题四是针对不同玻璃类别的化学成分之间的关联关系，并分析关联关系的差异性。初步判断这是一道相关性分析的问题，选择相关性分类来分析化学成分的关联关系。由于化学成分均为连续型变量，因此这里采用皮尔逊相关系数来进行相关性分析，并通过绘制热力图观察两类关联关系有无显著差异。

## 6.2 常用的相关性分析方法

（1）Pearson 简单相关系数

Karl Pearson 在 20 世纪初提出了相关系数的。概念。假设 X、Y 为两个随机变量，$\sigma_x$、$\sigma_y$分别表示 X 和 Y 的标准差，$\sigma_{xy}$表示 X、Y 的协方差，Pearson 相关系数定义为$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$。若 X、Y 的样本数据分别为$x_1$，$x_2$，…，$x_n$；$y_1$，$y_2$，…，$y_n$，则 Pearson 样本相关系数 r 为：

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

其中$\bar{x}$、$\bar{y}$分别表示 X、Y 的样本均值。

Pearson 相关系数通常用来衡量两个连续变量之间的相关关系。其取值在$[-1，1]$之间，其绝对值$|r|$表示两变量间相关关系程度的强弱，越接近 1，表明两变量间相关程度越高，它们之间的关系越密切。r＞0 表示正相关，r＜0 表示负相关，r＝0 表示不相关，$|r| = 1$表示完全相关。r 是样本统计量，其 取值会受到抽样波动性的影响，因此一般可通过对 r 进行统计检验判断两个变量之间线性关系是否显著。此外，需要注意的是，Pearson 相关系数只是衡量两变量间线性相关程度的大小，r＝0 只表示两变量间无线性相关关系，但不能确定是否存在其他非线性相关关系。

（2）Spearman 等级相关系数

Spearman 相关系数又称秩相关系数，是利用两变量的秩次大小作线性相关分析，对原始变量的分布不作要求，属于非参数统计方法，适用范围相对来说比较广。在 Pearson 提出简单相关系数之后，Spearman 指出对于“大”“中”“小”这样的定序变量间的相关性，Pearson 相关系数不再适用，进而提出了适用于定序数据的等级相关系数方法。其基本定义为：2 个定序随机变量 X 和 Y 的秩之间的 Pearson 相关系数。若变量 X、Y 的样本数据分别为$x_1$，$x_2$，…，$x_n$；$y_1$，$y_2$，…，$y_n$，则 X、Y 之间的 Spearman 等级相关系数可以表示为：

$$r_s = \frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}}$$

式中，$r_i$和$s_i$分别表示$x_i$和$y_i$的秩，当变量里出现相等值的时候，该值对应的秩为这几个值对应的秩的平均值。然而，实际应用中，为计算简便，可以通过被观测两个变量间的等级差值简化计算$r_s$，计算公式为：

$$r_s = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

式中，$r_s$表示等级相关系数；D 为每对观测值的等级差；n 为样本容量。$r_s$的取值范

围为[−1,1]，当一个变量随另一个变量单调递增的时候，$r_s$＝１；反之，当一个变量随另一个变量单调递减的时候，$r_s$=-1。Spearman 相关系数对数据条件的要求没有 Pearson 相关系数要求严格，只要两个变量的观测值是成对的等级评定资料，或者是由连续变量观测资料转化得到的等级资料，不论两个变量的总体分布形态、样本容量大小如何，都可以用 Spearman 相关系数进行研究。

（3）Kendall 等级相关系数

１９３８年，Kendall 提出了计算等级相关系数的新方法。Kendall 相关系数的计算也是以变量 X 和 Y 的等级数据进行，根据配对等级顺序排列的位置是否颠倒或换位，得出等级换位的次数，进而进行计算。其计算公式为：

$$r_k = 1 - \frac{4\sum_i}{n(n-1)}$$

其中 n 为样本容量，$\sum_i$为换位总次数。在测量等级相关方面，与Ｓｐｅａｒｍａｎ相关系数相比，Kendall 相关系数有更大的优势，Kendall 相关系数的置信区间更容易解释，可靠性也更高。两种相关系数在样本量小的时候，均不服从正态分布；当样本量较大的时候，均渐近服从正态分布。这里需要注意的是，Kendall 相关系数只是说明两组数据相关性高低，并不反映线性相关程度的大小。

## 6.3 相关性分析方法的选取和结果分析

因为本题中的变量是连续性变量，所以我们采用 Pearson 的方法来经行相关性分析。分析的结果如下图所示。

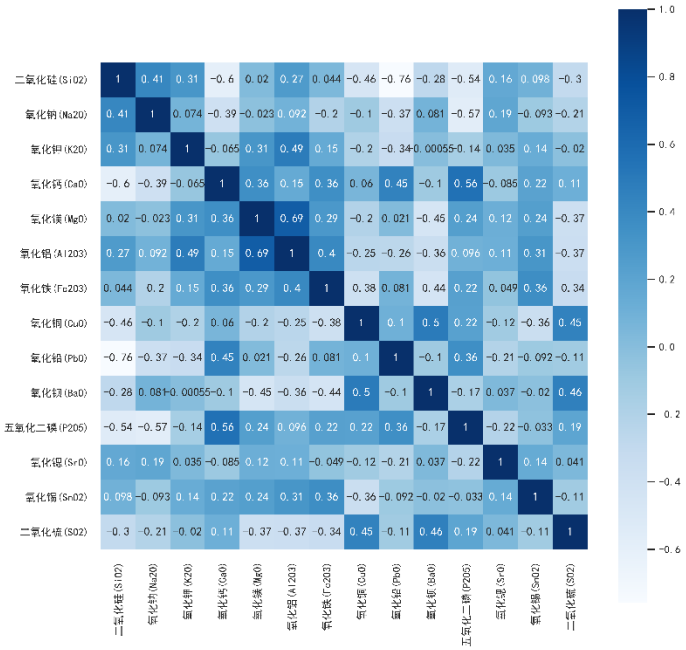


**图 4-1：铅钡玻璃相关热力图**

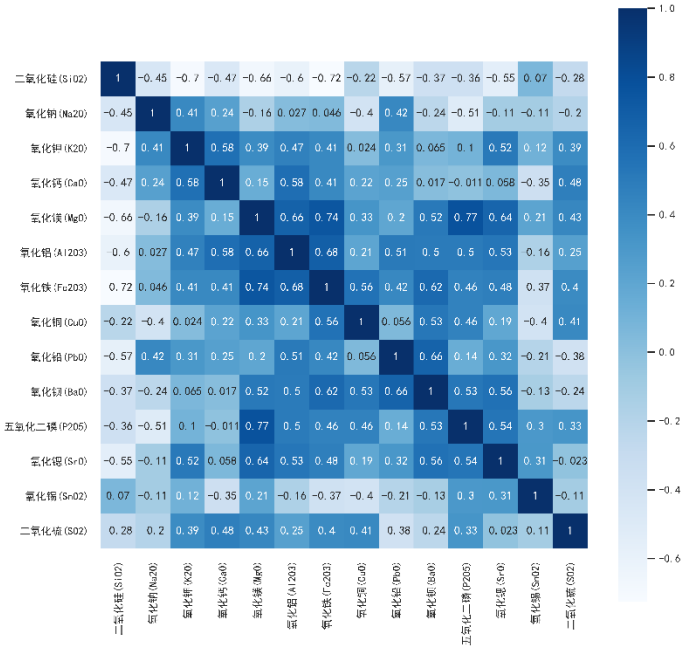


**图 4-2：高钾玻璃相关热力图**

根据铅钡玻璃相关热力图，可以发现二氧化硅与氧化铅呈高度负相关，与氧化钙呈中度负相关，与氧化钠和氧化钾呈弱的正相关；氧化钠与五氧化二磷呈中度负相关；氧化钾与氧化铝呈中度正相关。氧化钡与氧化铜，二氧化硫呈中度正相关，与氧化铁呈中度负相关；五氧化二磷与氧化钙呈中度正相关，与二氧化硅和氧化钠呈中度负相关。

根据高钾玻璃相关热力图，氧化钡与氧化镁，氧化铝，氧化铁，氧化铜，氧化铅，五氧化二磷和氧化锶都呈中度的正相关，与二氧化硅呈中度的负相关；氧化钠与氧化钾和氧化铅呈中度的正相关，与二氧化硅，氧化铜，五氧化二磷呈中度负相关；五氧化二磷与氧化镁呈高度正相关，与氧化铝，氧化铁，氧化铜，氧化钡和氧化锶呈中度正相关，与二氧化硅和氧化钠呈中度负相关；氧化钙与氧化钾，氧化铝，二氧化硫呈中度正相关，与二氧化硅，氧化锡呈中度负相关；氧化铁与氧化镁呈高度正相关，与二氧化硅呈高度负相关。

## 6.4 化学成分差异性分析

通过对比高钾玻璃和铅钡玻璃的化学成分相关性分析，可以发现，在铅钡玻璃中，氧化钡与氧化铜，二氧化硫呈中度正相关，与氧化铁呈中度负相关，而在高钾玻璃中，氧化钡与氧化镁，氧化铝，氧化铁，氧化铜，氧化铅，五氧化二磷和氧化锶都呈中度的正相关，与二氧化硅呈中度的负相关。在铅钡玻璃中，氧化钠与五氧化二磷呈中度负相关，而在高钾玻璃中，氧化钠与氧化钾和氧化铅呈中度的正相关，与二氧化硅，氧化铜，五氧化二磷呈中度负相关。在铅钡玻璃中，五氧化二磷与氧化钙呈中度正相关，与二氧化硅和氧化钠呈中度负相关，而在高钾玻璃中，五氧化二磷与氧化镁呈高度正相关，与氧化铝，氧化铁，氧化铜，氧化钡和氧化锶呈中度正相关，与二氧化硅和氧化钠呈中度负相关。

# 7 模型的评价与推广

## 7.1 模型的评价

### 7.1.1 模型的优点

1. 本文在建立风化的分类模型上，将玻璃类型通过重新编码，能抵消分类型变量在建模上的缺点。
2. 本文大量使用可视化的手段，来分析数据的分布规律，并且能为建模提供更好的训练手段。
3. 本题采用的模型均为简单模型，过程清晰易懂，模型复杂度较低，属于高效算法。

### 7.1.2 模型的缺点

1. 本文在第一问的预测模型建立上，仅使用了多元线性回归模型，但由于数据量不够多，绝大部分的机器学习模型不能在此数据上有很好的效果，因此这里可以通过训练更多的模型来获取更好的预测精度。
2. 针对化学成分关联关系的异常性，本文仅从数据特点的分布和大小出发，也许有更好的分析方法，来分析其中的差异性。

## 7.2 展望

本题所建立的模型既可以用在对玻璃文物的分析与鉴别，对其它文物的分析同样适用，很好地体现了统计学在考古领域的适用性，未来也许能在文物鉴别上起到一定的作用与效果。

但目前题目所给的数据量较少，也许分析得到的结果未能完全呈现实际现象，因此当未来数据量越来越多后，我们也许可以训练更多的模型以得到更具体真实的结果。

# 参考文献


[1] 李青会,董俊卿,干福熹.中国早期釉砂和玻璃制品的化学成分和工艺特点探讨[J].广西民族大学学报(自然科学版),2009,15(04):31-41.DOI:10.16177/j.cnki.gxmzzk.2009.04.014.

[2] 陈姝聿.我国古代玻璃的起源和发展[J].文物鉴定与鉴赏,2019(04):44-45.

[3] 陶莹,杨锋,刘洋,戴兵.K 均值聚类算法的研究与优化[J].计算机技术与发展,2018,28(06):90-92.

[4] Bao Chong. K-means clustering algorithm: a brief review[J]. Academic Journal of Computing & Information Science,2021,4.0(5.0).

[5] Mei Kai,Tan Meifang,Yang Zhihui,Shi Shaoyue. Modeling of Feature Selection Based on Random Forest Algorithm and Pearson Correlation Coefficient[J]. Journal of Physics: Conference Series,2022,2219(1).

[6] 王晓燕,李美洲.浅谈等级相关系数与斯皮尔曼等级相关系数[J].广东轻工职业技术学院学报,2006(04):26-27.

[7] Komarek P. Logistic regression for data mining and high-dimensional classification[M]. Carnegie Mellon University, 2004.

[8] Cheng Q, Varshney P K, Arora M K. Logistic regression for feature selection and soft classification of remote sensing data[J]. IEEE Geoscience and Remote Sensing Letters, 2006, 3(4): 491-494.

[9] 张虎,刘强.问卷调查分析中的 Logistic 回归与自变量筛选问题研究[J].中南财经政法大学学报,2003(05):128-132+144.

[10] 许汝福.Logistic 回归变量筛选及回归方法选择实例分析[J].中国循证医学杂志,2016,16(11):1360-1364.

[11] 孟杰,李春林.基于随机森林模型的分类数据缺失值插补[J].统计与信息论坛,2014,29(09):86-90.

[12] 李欣海.随机森林模型在分类与回归分析中的应用[J].应用昆虫学报,2013,50(04):1190-1197.
雍凯. 随机森林的特征选择和模型优化算法研究[D].哈尔滨工业大学,2008.

[13] Speiser J L, Miller M E, Tooze J, et al. A comparison of random forest variable selection methods for classification prediction modeling[J]. Expert systems with applications, 2019, 134: 93-101.

# 附录

| python 程序 | 数据预处理 |
| --- | --- |

```
#对表单 1 进行缺失值处理
import pandas as pd
dataset1 = pd.read_excel("附件.xlsx",sheet_name = "合并")
data = pd.DataFrame(dataset1)
data = data.fillna(0)
ych = list()
for i in data.index:
    b = 0
    for j in range(6,19):
        b += data.iloc[i,j]
    if b >= 105 or b <= 85:
        ych.append(i)
data = data.drop(ych)
```

| 问题 1　python 程序 | 绘图、分类模型 |
| --- | --- |

```
#绘制箱型图
plt.figure(figsize=(40,20))
a = 0
for i in range(6,18):
    xlab = data_lx1.columns[i]
     hxcf_x = pd.DataFrame({"风化高钾":data_lx1.iloc[:,i],"无风化高钾":data_lx2.iloc[:,i],"风化铅钡":data_lx3.iloc[:,i],"无风化铅钡":data_lx4.iloc[:,i]})
    plt.subplot(3,4,a+1)
    a += 1
    #plt.subplots(dpi=1080,facecolor='w')# 设置画布大小，分辨率，和底色
    hxcf_x.boxplot()
    plt.xlabel(xlab,fontsize = 16) # 我们设置横纵坐标的标题。
    plt.tick_params(labelsize = 16)

#绘制分布图
plt.figure(figsize=(30,10))
for n,i in enumerate(["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)"]):
    plt.subplot(2,7,n+1)
#       plt.title(i)
    sns.distplot(data_lx1[i])
```

```
    plt.xlabel(i, fontsize = 14)
    plt.ylabel('')
    plt.tick_params(labelsize = 14)

#建立逻辑回归模型
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression ()
lr.fit(X_train,y_train)
lr.predict(X_test)

#输出模型系数
print('训练模型自变量参数为：',lr.coef_)
print('训练模型截距为：',lr.intercept_)
#模型评价
print('模型的平均正确率为：',lr.score(X_test,y_test))

#预测精度
from sklearn.metrics import accuracy_score
y_predict=lr.predict(X_test)
accuracy_score(y_test,y_predict)

#逐步回归
from sklearn.linear_model import LinearRegression as LR
from sklearn.metrics import r2_score

model_sio2=LR().fit(data_2_x1,data_2_si02)

perfomance_reg(model_sio2,data_2_x1,data_2_si02,'训练集')
data1_sio2 = model_sio2.predict(data_1_x1)
data1_sio2 = pd.DataFrame(data1_sio2,columns=['二氧化硅(SiO2)'])
data_1_x2 = pd.concat([data_1_x1,data1_sio2],axis=1)
```

| 问题 2　python 程序 | 绘图、分类模型、聚类模型 |
|---|---|

```
#建立随机森林分类模型
model_rf = RandomForestClassifier(random_state=9)
model_rf.fit(X_train,y_train)

#模型评价
expected = y_test
predicted = model_rf.predict(X_test)
```

```
print(metrics.classification_report(expected,predicted))
print(metrics.confusion_matrix(expected,predicted))

#绘制聚类手肘图
d=[]
for i in range(1,11):    #k 取值 1~11，做 kmeans 聚类，看不同 k 值对应的簇内误差平方和
    km=KMeans(n_clusters=i,init='k-means++',n_init=10,max_iter=300,random_state=0)
    km.fit(data_gj_jl)
    d.append(km.inertia_)  #inertia 簇内误差平方和

#绘图参数设置
fig = plt.figure(dpi=300,figsize=(5,3)) #设置分辨率，画布大小
ax = fig.add_subplot(111)
#设置背景色
ax.patch.set_facecolor('white')#设置画布外颜色
fig.patch.set_facecolor('white')#设置画布内颜色
#设置画框的颜色
ax.spines['bottom'].set_color('black')
ax.spines['top'].set_color('black')
ax.spines['left'].set_color('black')
ax.spines['right'].set_color('black')
#作图
plt.plot(range(1,11),d,marker='o',color = 'black')
plt.xlabel('number of clusters',color='black')
plt.ylabel('distortions',color='black')
#设置横纵坐标轴的颜色
plt.tick_params(axis='x',colors='black')
plt.tick_params(axis='y',colors='black')

#聚类
Kmean1= KMeans(n_clusters=3,random_state=0)
Kmean1.fit(data_qb_jl)
#获取聚类结果
qb_yc = Kmean1.labels_
```

| 问题 4　python 程序 | 绘图、相关性分析模型 |
|---|---|

```
corr_gj = data_gj.corr(method="pearson")
corr_qb = data_qb.corr(method="pearson")
plt.subplots(figsize=(9,9),dpi=1080,facecolor='w')#  设置画布大小，分辨率，和底色
p1 = sns.heatmap(corr_gj ,annot=True, vmax=1, square=True, cmap="Blues", fmt='.2g')
p2 = sns.heatmap(corr_qb ,annot=True, vmax=1, square=True, cmap="Blues", fmt='.2g')
```