基于树的玻璃文物的成分分析与鉴别

摘要

一句话概括整个赛题的要求和你要怎么思考，总之这里就是一个很大的话

针对问题一：摘要内容。

针对问题二：摘要内容。

针对问题三：摘要内容。

针对问题四：摘要内容。

关键词：随机森林；方差选择法；Voting Classifier；层次聚类分析；决策树

目录

1 问题重述

1.1 问题背景

问题背景内容

1.2 问题提出

现有一组由专家给出的关于玻璃的数据，要求通过分析与建模解决下面若干问题：

问题 1 问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容。

问题 2 问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容。

问题 3 问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容。

问题 4 问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容问题提出内容。

2 问题分析

2.1 问题一的分析

问题分析内容

问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问

题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容问题分析内容

2.2 问题二的分析

问题分析内容

问题分析内容

2.3 问题三的分析

问题分析内容

问题分析内容

2.4 问题四的分析

问题分析内容

问题分析内容

3 模型假设

1. 假设所给数据真实可靠。
2. 本文认为在对样本进行采样过程中没有破坏样本的完整性。
3. 玻璃统计规律可以代表玻璃的一般规律，不随其他无关因素而改变。
4. 从玻璃内部的所取化学物质与表中所给物质吻合，不发生化学反应。
5. 针对不同玻璃，其颜色特征不会因光照等因素而发生改变。
6. 针对不同的化学成分，在风化前后的化学物质的比例稳定。
7. 化学成分含量发生变化也适用于变化前的模型规律。

4 符号说明

符号	含义
$i=1, i=2$	分别表示高钾、铅钡玻璃
j	表示表中从二氧化硅 (SiO_2) 到二氧化硫 (SO_2) 中第 j 类化学物质
$z=1, z=2$	分别表示风化前和风化后
x_1, x_2, x_3, x_4	分别表示纹饰、类型、颜色、风化表面
y_j	表示第 j 类化学物质的含量
$\overline{y_j}$	表示第 j 类化学物质的平均含量

5 问题一建模与求解

为分析玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系以及文物样品表面有无风化化学成分含量的统计规律，并对风化前的化学含量进行分析，制作流程图如下：

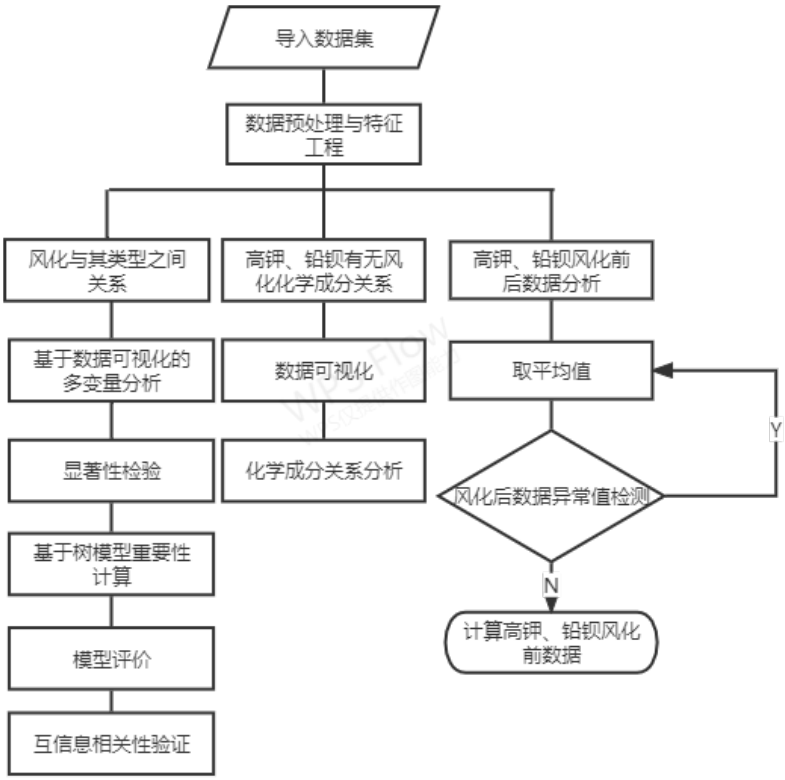

图 1: 问题一流程图

5.1 模型准备——数据预处理与特征工程

5.1.1 数据清洗

本文根据附件表 1 所给的数据，通过 Pandas 库进行数据读取，使用 isnull 函数查找数据中的缺失值，由此发现颜色特征中有数个缺失值，由数据可视化发现黑色文物必定被风化，而缺失值所在样本均为被风化对象，故采用黑色来补全缺失值。

对于附件表 2 所给的数据，表中空白部分为未检测到的成分将被 Pandas 视为缺失值，本文指定数值 0 填补空缺的值以避免缺失值对分析的干扰。

按照题意对其每一行进行求和，选取成分累加和介于 85% 105% 的有效数据并删除不在有效数据范围内的第 15 个和第 17 个样品采样点的样本构成新的数据集用于分析。

5.1.2 特征编码

为更准确的判断风化情况与三个变量相关关系，依据表 1 所给的纹饰、类型、颜色以及表面是否风化四个特征对其用特征编码进行分类，由于各个类别之间是相互独立的，故采用独热编码进行特征变换以消除编码后各个类别的不同取值差异对模型训练效果的影响。

然而事与愿违，颜色特征中有七个非序数类别，而独热编码最大的缺点即容易造成编码后数据集的“高维危机”，由此针对颜色特征本文放弃了进行独热编码，而是对其特征编码。

5.2 玻璃风化与其分类信息关系的分析模型

5.2.1 基于数据可视化的多变量分析

本问要求分析玻璃纹饰、类型、颜色与表面风化情况的关系，基于多变量分析的数据可视化容易分别得到这三个分类特征与风化情况之间的关系，通过 Pandas 库中的 crosstab 函数可以实现分类特征之间的多变量分析，图形呈现如下：

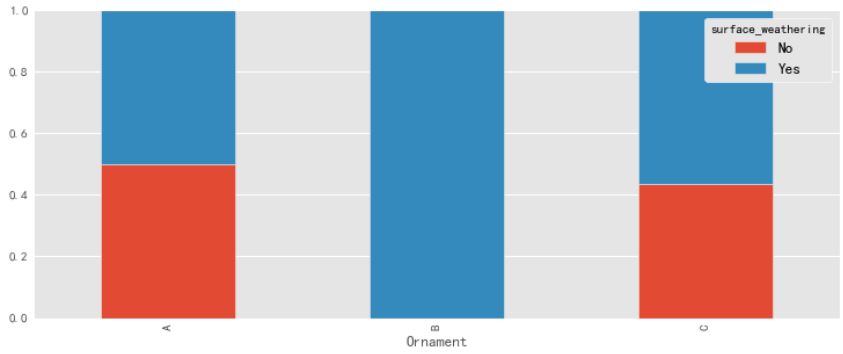

图 2: 玻璃纹饰与风化情况关系

由图 2 可知所提供的文物样品中 A 种纹饰的玻璃中近一半被风化，C 种纹饰中风化的玻璃超过一半，而 B 种纹饰的玻璃全部风化——由此可以推断 B 类的纹饰与风化情况有极大联系。下面再来考察玻璃类型与风化情况的关系：

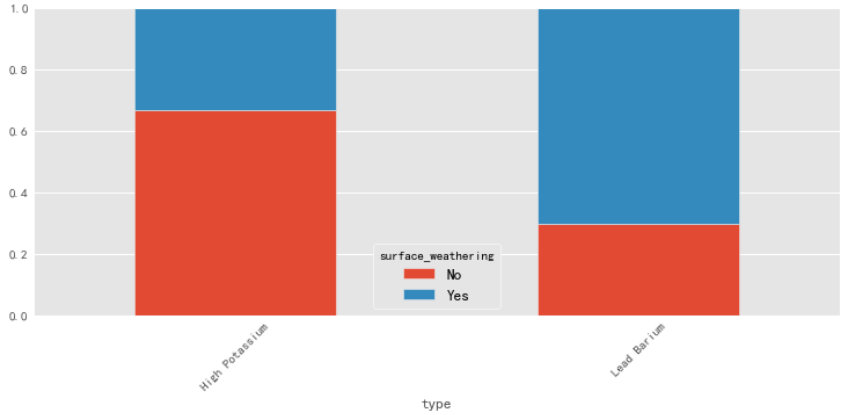

图 3: 玻璃类型与风化情况关系

由图 3 可知，对于样本所给的玻璃类型中高钾类玻璃风化的情况仅占百分之三十五，说明高钾类玻璃被风化的概率较小，而对于铅钡类的玻璃则是达到了百分之七十的风化比例。由此可见铅钡类玻璃更容易被风化。

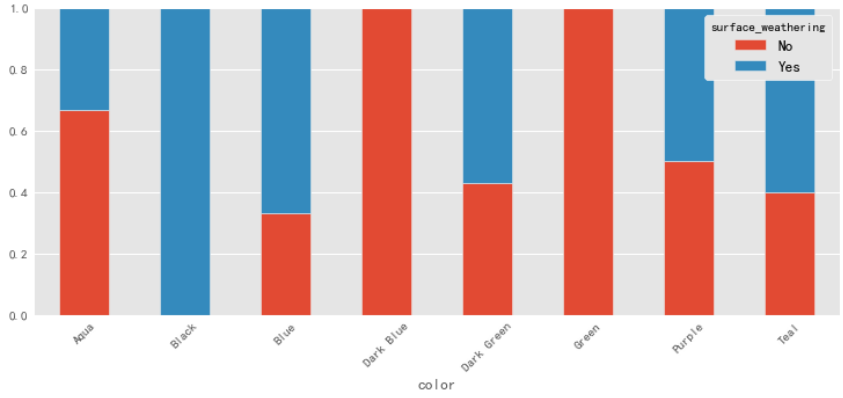

图 4: 玻璃颜色与风化情况关系

由图 4 可得，黑色的玻璃样本表面均被风化，而绿色和深蓝色的玻璃样本表面均为未被风化，其余颜色的样本表面被部分风化——由此可见黑色、绿色和深蓝色的玻璃与风化情况存在很大的关系，其他特征无法确定。

综上所述，可以总结出以下规律：

单个分类变量中玻璃纹饰和颜色都存在与表面风化情况确切相关的特征（如 B 种纹饰必定被风化，黑色文物必定风化，而深蓝和绿色文物均未被风化），而玻璃类型只能得到与风化情况的不同的频率，但不能确定是否风化，因此可以确定纹饰与颜色对玻璃风化的关系紧密，而玻璃类型可能相关性较弱。

下面通过集成学习中的树模型直接说明三种分类变量与玻璃风化情况的强弱关系。

5.2.2 基于树的模型重要性计算

通过模型准备中的特征工程本文已经实现了对数据分类特征的编码，经过模型准备的数据可以正式代入随机森林模型进行训练。

树模型作为一种集成学习算法，是一种效果优良的机器学习模型，且其独特的计算特征重要性的功能对于探索本问中多个变量的相关关系具用重要意义。本文基于树模型中最简单的 Bagging 算法——随机森林，对纹饰、类型、颜色与表面风化情况之间的关系进行分析，具体方法如下：

1. 为消除各类别之间的影响，便于对指标进行比较分析把分类特征编码数据标准化：进行标准化处理

$$z_i = \frac{x_i - \overline{x_i}}{\sigma_i} \tag{1}$$

2. 为找到最优模型，必须对树模型内部参数进行调整，找到最优参数，下面对最优参数进行探究。

（1）训练集的占比直接关系到模型的优劣，寻找训练集与测试集的占比最优参数，让模型达到较好的效果，利用 scikit-learn 库中的 learning curve 函数进行训练，并做交叉验证取平均值得到学习曲线图像如下：

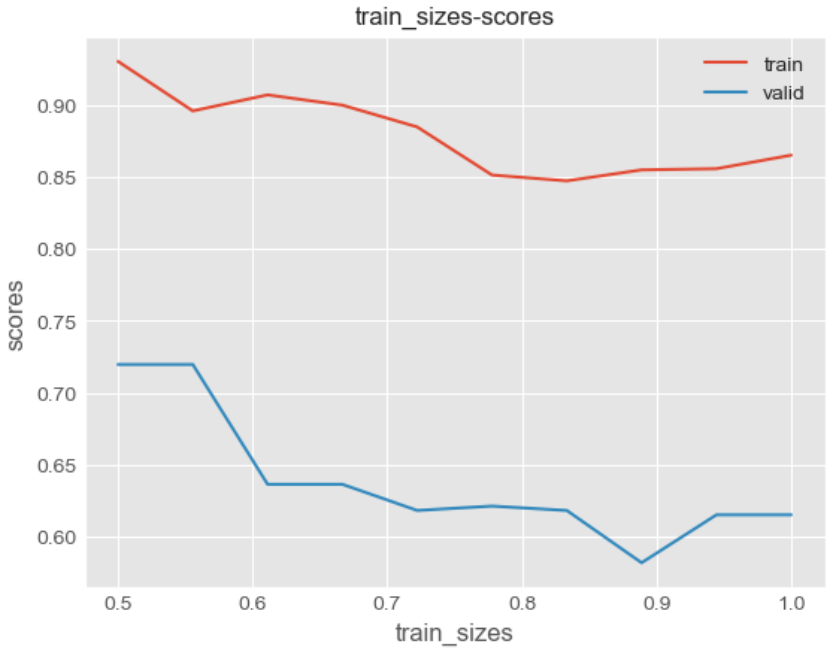

图 5: 学习曲线

图 5 可知，训练集占比 0.5 时，测试集效果最好，因此设置训练集占比为 0.55 是为最优训练集占比参数。

（2）对于树模型而言，大多数情况下其基学习器的数量是影响其性能的最重要的参数。通过 GridSearchCV 函数对随机森林模型中决策树的数量进行调整。结果如下：

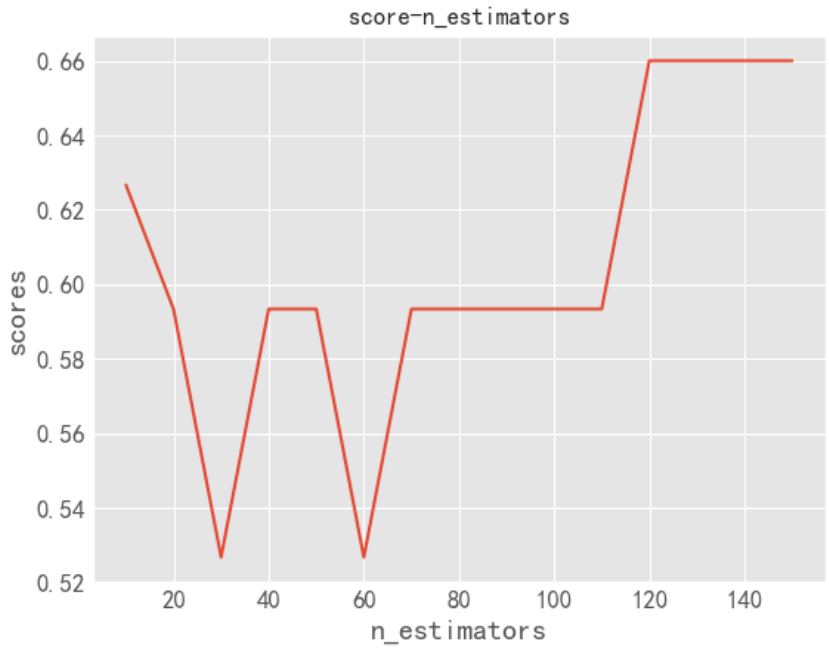

图 6: 决策树数量图

由图 6 知，在决策树的数量为 120 时，模型得分已经达到峰值，此时可以确定决策

树的数量最佳值为 120。

利用此模型对纹饰、类型、颜色三个特征进行特征重要性计算，结果如下：

变量	纹饰 A	纹饰 B	纹饰 C	高钾类型	铅钡类型	颜色
评分	0.0963	0.1014	0.0518	0.1708	0.1809	0.3987

表 1: 评分表

由表 1 可以看出玻璃纹饰 C, 纹饰 A, 纹饰 B 的评分低，玻璃类型次之，玻璃颜色最大；故本问得出玻璃颜色与玻璃表面风化的关系最强，玻璃的类型次之，而玻璃纹饰与玻璃表面风化关系最弱。结果表示与单个变量重要性之间有所差别，下面进一步说明树模型结果的可靠性以及两种结果差别的原因。

5.2.3 模型评价

树模型在本问中能够得到的结果是否可信，需要进一步对模型进行评价。通常是主要通过热力图和 ROC 图进行判断，同时利用准确率（0.78），精确率，召回率，F1 分数，支持率来综合判断模型。根据以上得到的模型参数利用 python 对模型进行调参后形成新的模型，得到热力图并绘制了 ROC 曲线测试模型，如下图所示：

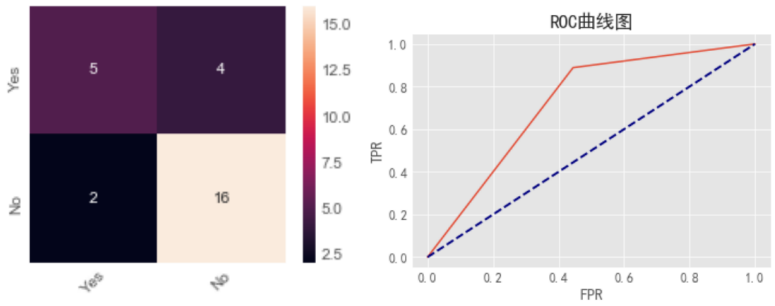

图 7: 分类结果混淆矩阵热力图

由图 7 可知，通过热力图知该模型中正确判断未风化的比例占 8/9, 正确判断风化的比列占 5/9；ROC 曲线图中可看出此模型的分类器性能极佳，由此认为这个模型可以用来分析批判纹饰、类型、颜色及风化表面之间关系。

对于分类模型的评价还有很多基于混淆矩阵的方法，结合对测试集的预测结果和测试集自身标签、通过计算得到模型的精确度得到如下指标：

	精确度	召回率	F1 分数	支持率
铅钡	0.71	0.56	0.63	9
高钾	0.80	0.89	0.84	18
加权平均	0.77	0.78	0.77	27

表 2: 评分数据表

由表 2 可知，利用该模型对玻璃类型中高钾和铅钡检验得到的精确度均大于百分之七十，召回率均大于百分之五十，其中高钾类召回率高达百分之八十九，由此可知该模型极为可信，说明该模型对于三种特征的特征重要性评分具有可信度。对于上面结果的不完全相同，可能由于单变量特征重要性没有量化评判标准，导致结果有部分不同，但颜色特征关联最大的特征是确定的，下面对于不同特征重要性进行验证。

5.2.4 互信息验证

寻找特征与标签的关系的手段并不是单一的，针对本题数据的具体情况，由于特征和标签均为分类型变量，本文也通过计算互信息以衡量特征与标签间的独立性来验证此题结果。

利用互信息计算风化表面与纹饰，类型，颜色的互信息进一步验证每种关联性的科学性，公式如下：

$$I_i\left(x_i; x_4\right) = \sum_{x_i, x_4} p(x_i, x_4) \log \frac{p(x_i, x_4)}{p(x_i)p(x_4)} \tag{2}$$

得到互信息表如下：

	纹饰	类型	颜色	表面风化情况
纹饰	0.943384	0.138290	0.383740	0.061379
类型	0.138290	0.619376	0.230124	0.059385
颜色	0.383740	0.230124	1.609943	0.077623
表面风化情况	0.061379	0.059385	0.077623	0.678209

表 3: 互信息表

由表 3 知由于纹饰和类型与玻璃表面风化情况的相关性值相差不大（约为 0.02），可忽略不计，造成在多个变量共同作用时，类型的占比增大，因此类型的评分增大；而颜色与表面风化情况的相关性评分比其余两个都强，故认为颜色与玻璃文物表面风化有很大联系；此结果与上述模型结果显示一致，由此认为树模型是一个较好的模型且适用于玻璃风化与不同特征重要性评价。

5.3 分析玻璃化学含量描述性模型

要结合玻璃类型分析文物样本有无风化化学成分含量的统计规律，本问分别对高钾类和铅钡类玻璃风化情况与各化学成分关系进行数据可视化分析以及显著性检验来探究风化情况不同时化学成分的差异。

5.3.1 数据可视化初探风化情况与化学成分浅层关系

用 Python 读取由附件表 2 手动处理后的有效数据并提取出高钾类玻璃的数据。利用 matplotlib 得到高钾类玻璃风化与各化学成分关系的直方图，如下所示：

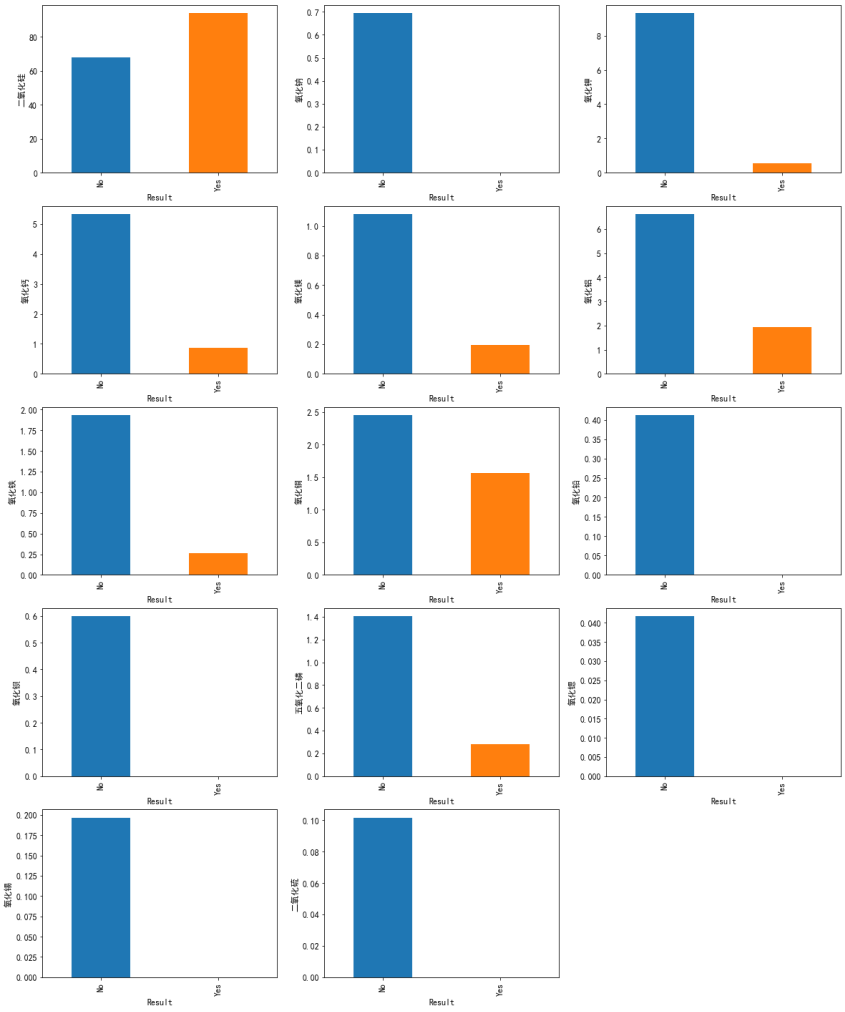

图 8: 高钾类玻璃风化与各化学成分关系直方图

由图 8 可知高钾玻璃中含氧化钠、氧化铅、氧化钡、氧化锶、氧化锡以及二氧化硫等化学分成表面未风化的均值较高，而表面风化的均值为 0；氧化钾、氧化钙、氧化镁、氧化铝、氧化铁、五氧化二磷等化学分成表面未风化的均值均远大于表面风化的均值；由此可知风化的玻璃中含有氧化钠、氧化铅、氧化钡、氧化锶、氧化锡以及二氧化硫等化学成分含量较少，而含氧化钾、氧化钙、氧化镁、氧化铝、氧化铁、五氧化二磷等成

分相对于多一些；另外，由于二氧化硅为玻璃制作的主要原材料，在化学成分中分布很广，在分析与玻璃风化是否有联系的结果并不是很明显。故玻璃风化联系最紧密的化学成分是氧化铜，所有化学成分与玻璃未风化联系都紧密。

同理本文可得到铅钡类玻璃风化与各化学成分关系的直方图如下：

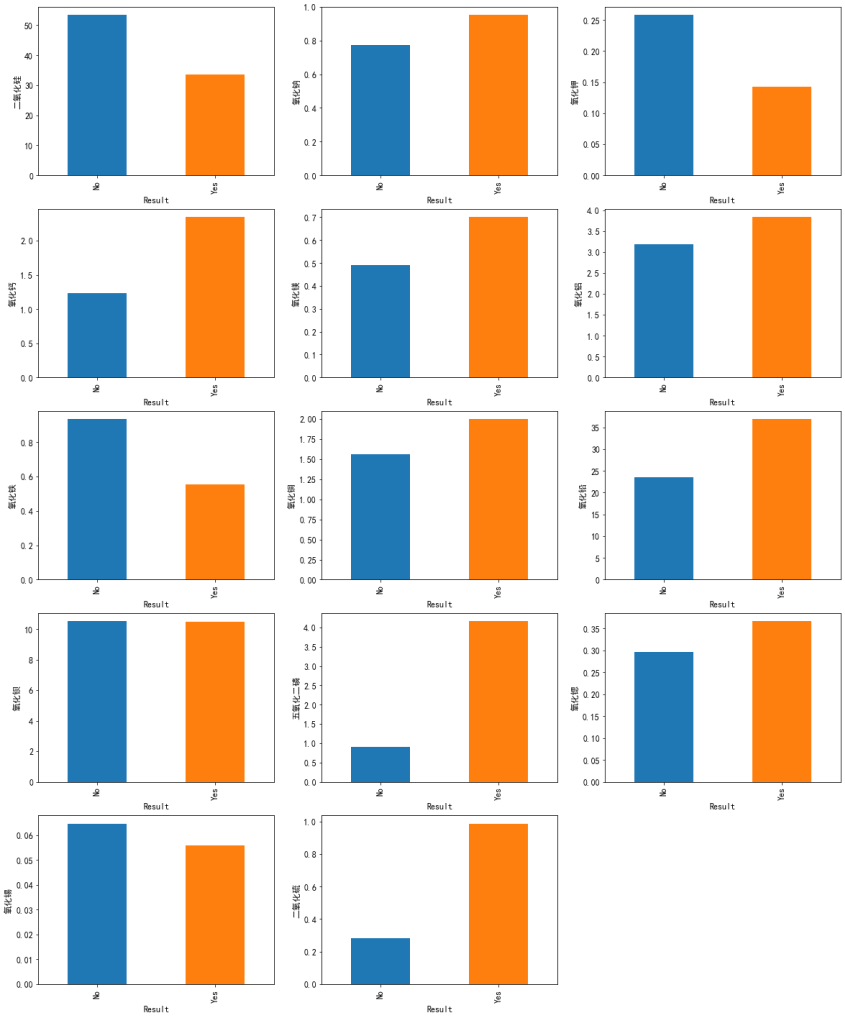

图 9: 铅钡类玻璃风化与各化学成分关系直方图

由图 9 知对于铅钡类玻璃中二氧化硫与五氧化二磷两种化学成分表面风化均值占比远高于表面未风化均值；而二氧化硅、氧化钾、氧化镁三个化学成分中玻璃表面风化均值低于表面未风化的均值；其余化学成分与表面是否风化占比相差不大。故二氧化硫与五氧化二磷两种化学成分绝大多数被风化，其余化学成分风化与被风化的程度相差不大。

5.3.2 显著性检验考察风化情况下各化学成分差异性

依据以上得到的数据可视化图，为了检验风化前后化学成分是否有明显变化，分别对高钾玻璃和铅钡玻璃风化前后对应的化学成分进行显著性检验。首先进行正态性检验得到下表：

二氧化硅	氧化钠	氧化钾	…	氧化锶	氧化锡	二氧化硫
0	0.002680	0	…	0.002680	0.002680	0.002680

表 4: 高钾玻璃化学成分正态检验部分表

由上表发现化学成分均未通过正态性检验。

假设风化前后的化学成分有显著性差异，对于通过正态性检验的成分进行配对样本 t 检验，结果如下：

$$p\_value = 0.10991037$$

故拒绝原假设，氧化铜在高钾玻璃风化前后不具有显著差异。

假设风化前后的化学成分有显著性差异，对于未通过正态性检验的成分进行配对 Milcoxon 检验，结果如下：

二氧化硅	氧化钠	氧化钾	…	氧化锶	氧化锡	二氧化硫
0.002526	0.181449	0.003263	…	0.5	1	0.181449

表 5: 配对检验表

氧化钠、氧化铜、氧化钡、氧化锶、二氧化硫五种成分拒绝原假设，在高钾玻璃风化前后不具有显著性差异；其他成分接受原假设，在风化前后具有显著性差异。同理可得铅钡玻璃化学成分正态性检验表：

二氧化硅	氧化钠	氧化钾	…	氧化锶	氧化锡	二氧化硫
0.279	0.001***	0.005***	…	0.585	0.000***	0.000***

表 6: 铅钡玻璃化学成分正态检验部分表

由上表得：化学成分氧化钠、氧化钾、氧化铁、氧化锡、二氧化硫未通过正态性检验，化学成分二氧化硅、氧化钙、氧化镁、氧化铝、氧化铜、氧化铅、氧化钡、五氧化二磷、氧化锶通过正态性检验。

假设风化前后的化学成分有显著性差异，对于未通过正态性检验的成分进行配对 Milcoxon 检验，结果如下：

二氧化硅	氧化钠	氧化钾	…	氧化锶	氧化锡	二氧化硫
0.000***	0.273	0.894	…	0.217	0.001***	0.225

表 7: 配对检验表

二氧化硅、氧化钙、氧化镁、氧化铝、氧化铜、氧化铅、氧化钡、五氧化二磷、氧化锶九种成分拒绝原假设，在铅钡玻璃风化前后不具有显著性差异；其余化学成分不拒绝原假设，其他成分接受原假设，在风化前后具有显著性差异。

5.4 预测风化前的化学成分含量

5.4.1 高钾、铅钡玻璃风化前后数据分析

把高钾和铅钡玻璃对表面是否风化的数据分别进行分析，得到部分结果如下表详情见附件。

	二氧化硅	氧化钠	氧化钡	二氧化硫	五氧化二磷
数量	12.00000	12.00000	12.00000	12.00000	12.00000
总和	67.984167	0.695000	0.598333	0.101667	1.402500
占比	8.755099	1.286917	0.982102	0.185513	1.433959
最小值	59.0100	0.0000	0.0000	0.0000	0.0000
$\frac{1}{4}$ 百分点	61.6775	0.0000	0.0000	0.0000	0.6900
$\frac{1}{2}$ 百分点	65.5300	0.0000	0.0000	0.0000	1.0200
$\frac{3}{4}$ 百分点	71.1675	0.5250	1.0725	0.0000	1.2925
最大值	87.050000	3.380000	2.860000	2.360000	4.500000

表 8: 高钾类玻璃风化前部分数据分析

接下来对高钾玻璃风化前后数据进行可视化，得到以下可视化分布图：

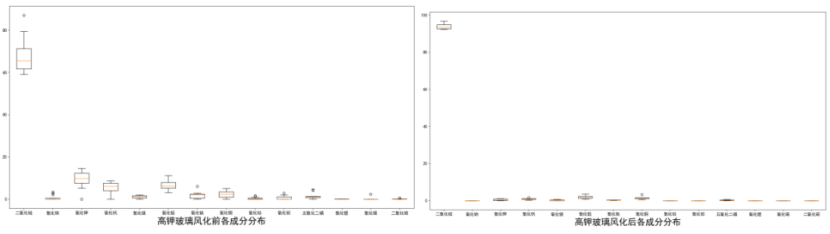

图 10: 左图为高钾玻璃风化前各成分分布图；右图为高钾玻璃风化后各成分分布图

调用 Matplotlib 的 boxplot 函数对高钾玻璃风化前后各特征的分布进行可视化，查看各特征离群值数量的多少。通过图 10 可以见高价玻璃风化前后各个化学成分的离群值较少，因此通过平均数可以反应数据各特征的一般水平。

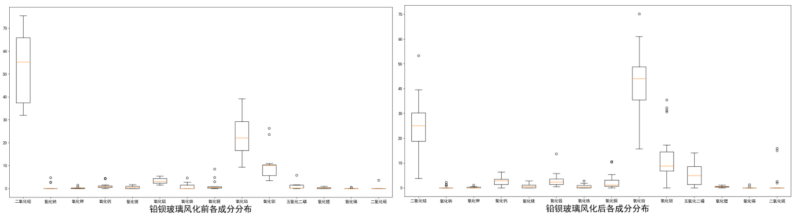

图 11: 左图为铅钡玻璃风化前各成分分布图；右图为铅钡玻璃风化后各成分分布图

通过图 11 可得铅钡玻璃风化前后各个化学成分的离群值依然较少，因此可以通过平均数来反映数据各个特征的一般水平。

5.4.2 求取平均值、预测风化前化学成分

由上把高钾和铅钡玻璃数据分析后可得数据异常值较少，为寻找一个风化前后具有代表性的数据，本问对两种玻璃风化前后的数据分别求平均值，具体方法如下：

1. 根据公式

$$\overline{y_{ijz}} = \frac{\sum y_{ijz}}{n_{ijz}}. \tag{3}$$

可分别求出高钾玻璃和铅钡玻璃风化前后每种化学成分的平均值。

2. 对风化后数据异常值进行检验，发现没有异常值，根据得到的风化前后的数据平均值求出对应化学成分的平均含量比率作为该化学含量的一般比率。设 H_j 为第 j 类化学物质的比例系数，利用公式

$$H_{ij} = \frac{\overline{y_{iy1}}}{\overline{y_{iy2}}}. \tag{4}$$

分别得到高钾和铅钡玻璃风化前后各成分比例系数部份表如下，具体详情见附件：

成分类型	高钾玻璃风化前后含量比值	铅钡玻璃风化前后含量比值
二氧化硅	0.723518	2.145246
五氧化二磷	5.008929	0.171270
氧化钾	6.129310	1.936599
氧化钙	5.487288	0.456906
氧化镁	3.430052	0.757396
氧化铝	7.289308	7 1.075628

表 9: 风化前后比例系数表

最后再通过公式

$$y'_{iy1} = H_{ij} \cdot y'_{iy2}. \tag{5}$$

求出风化后的玻璃对应的风化前的化学物质含量，得到对应的部分结果如下表所示，详情见附件：

样本检测位点成分	二氧化硅	氧化钠	...	氧化锡	二氧化硫
01	77.82951415	0	...	0	0
02	67.01947595	0.695	...	0.196666667	0.101666667
03	43.20524848	0	...	0	0.531689189
...	...	...	...	...	...
31	54.5321458	0	...	0	0
32	65.19401695	0	...	0	0

表 10: 风化前化学成分含量部分

6 问题二建模与求解

为分析高钾和铅钡两类玻璃的分类规律；对于每个类别选择合适的化学成分对其进行亚类划分得出分类结果，并对分类结果的合理性和敏感性进行分析，做流程图如下：

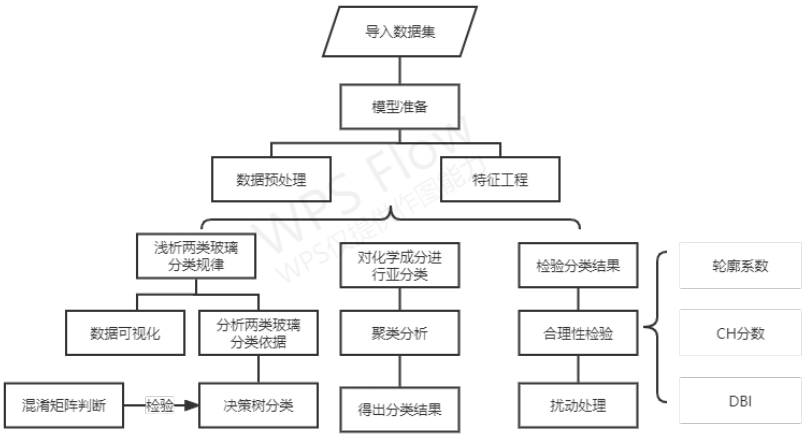

图 12: 问题二流程图

6.1 模型准备

6.1.1 数据预处理

用 Pandas 读取数据后用 isnull 函数查找缺失值，发现表中数据有缺失值，用众数来补全数据中的缺失值，再次检验未见异常值。

6.1.2 特征工程

要求分别在高钾玻璃和铅钡玻璃的化学成分中选择合适的特征进行亚类分析，聚类分析由于以欧几里得距离为组间、组内判别的衡量，往往对特征的要求极高。先对原来的 14 个化学成分指标进行标准化：再利用方差筛选法来对这几个特征进行过滤筛选，公式为：

$$\sigma_j^2 = \frac{\sum (y_j - \overline{y_j})^2}{n_j}, j = 1, \cdots, 14 \tag{6}$$

求出第 j 类化学成分的的方差值。为了筛选到最优方差范围，取不同方差值检验化学成分数据通过情况，得到方差通过如下表：

方差	二氧化硅	氧化钠	...	二氧化硫
$\sigma^2 = 0$	未通过	未通过	...	未通过
...	...	...	...	...
$\sigma^2 = 5$	通过	未通过	dots	未通过
...	...	...	...	...
$\sigma^2 = 10$	通过	未通过	...	未通过

表 11: 方差通过情况表

方差等于 10 已经能够较充分的分离特征明显的变量，故令数据

$${\sigma_j}^2 \leq 10 \tag{7}$$

利用方差法筛选选出特征明显的变量，用 VarianceThreshold 函数分别选出高钾类玻璃和铅钡类玻璃化学成分主要特征值，结果如下：

高钾类玻璃	二氧化硅	氧化钾	氧化钙	氧化铝
铅钡类玻璃	二氧化硅	氧化钡	氧化铅	五氧化二磷

表 12: 方差筛选后特征

6.2 数据可视化浅析高钾玻璃和铅钡玻璃分类规律

根据以上得到的数据，本文把 14 个化学成分作为数值特征与玻璃类型（高钾玻璃和铅钡玻璃）进行多变量数据可视化分析，利用 matplotlib 库得到数据可视化图如下：

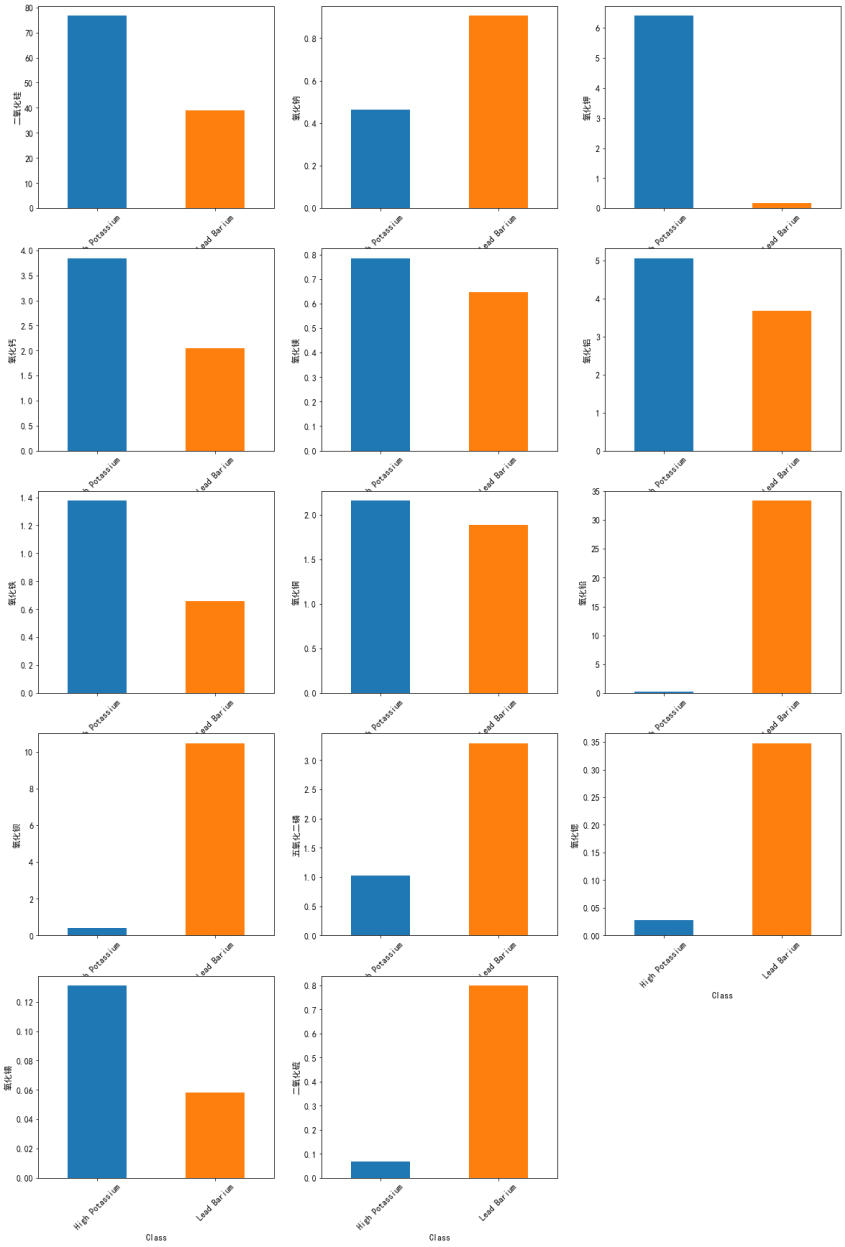

图 13: 数据可视化图

由图 13 可知氧化铅、氧化钡、氧化锶、二氧化硫、氧化钾五种化学成分在两种玻璃类型中占比相差极大，其中氧化铅、氧化钡、氧化锶、二氧化硫四种化学成分在高钾类玻璃的含量占比远大于铅钡类玻璃；而氧化钾在高钾类玻璃的含量占比远小于于铅钡类玻璃；氧化钙、氧化锡、氧化铁三种化学成分在高钾类玻璃的含量占比略大于铅钡类玻璃；氧化钠和五氧化二磷两种化学成分在高钾类玻璃的含量占比略于于铅钡类玻璃；而氧化镁、氧化铜两种化学成分在两类玻璃的占比相差不大。

接下来本文分析两个分类特征——纹饰和颜色与玻璃类型之间的关系，分别作多变量分析可视化进行分析，得到数据可视化图如下：

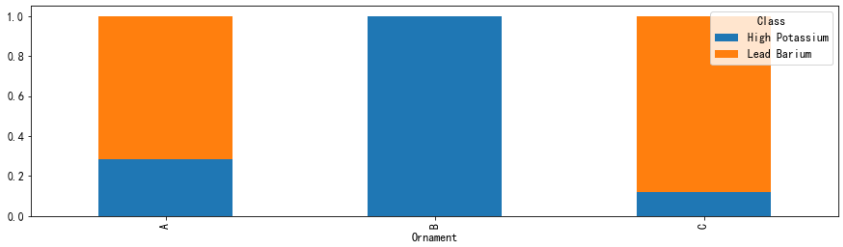

图 14: 纹饰与玻璃类型可视化图

图 14 可看出 A 类纹饰中高钾类玻璃占百分之七十左右，占比大于铅钡类玻璃占比，B 类纹饰全是高钾类玻璃，而 C 类纹饰中高钾类约占百分之九十以上，远大于铅钡类玻璃占比。

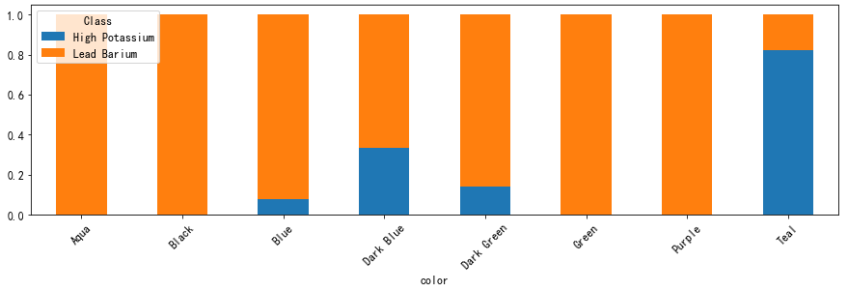

图 15: 纹饰与玻璃类型可视化图

图 15 可知浅蓝色样本玻璃中高钾类玻璃占比较大，蓝色、深蓝、深绿三种颜色玻璃中高钾类玻璃占比较小，而其余颜色玻璃中全为铅钡类玻璃。

综上所述，B 类纹饰以及蓝绿色、黑色、绿色、紫色的玻璃全为高钾类玻璃；浅蓝色玻璃大部分为高钾玻璃，C 类纹饰与蓝色、深绿色玻璃绝大多数多铅钡玻璃。可以得到如下显性分类结果表：

高钾玻璃	B 类纹饰、蓝绿色、黑色、绿色、紫色
铅钡玻璃	C 类纹饰、蓝色、深绿色

表 13: 显性分类规律表

上表可以直观得到两种玻璃的分类标准，下面为了更进一步的分析它们内部分类的本质性规律，决定采用决策树得到内部规律分类。

6.3 巧用决策树二叉树结构分析两类玻璃分类依据

决策树是一种经典的传统机器学习分类器，因其独特的树状分类机制，可以让操作者在建模完成后通过调用 plot_tree 函数了解每一次树枝分叉的条件，这一特点可以直接达成对玻璃分类依据的探究。

6.3.1 计算决策树分类结果与本身标签吻合度

附件早已给出各个样本的玻璃类别，本文通过决策树分类器寻找分类标准，一大前提就是当前决策树的分类结果和玻璃自身标签完全吻合，本文通过混淆矩阵计算决策树分类结果与样本本身标签的吻合度，结果如下：

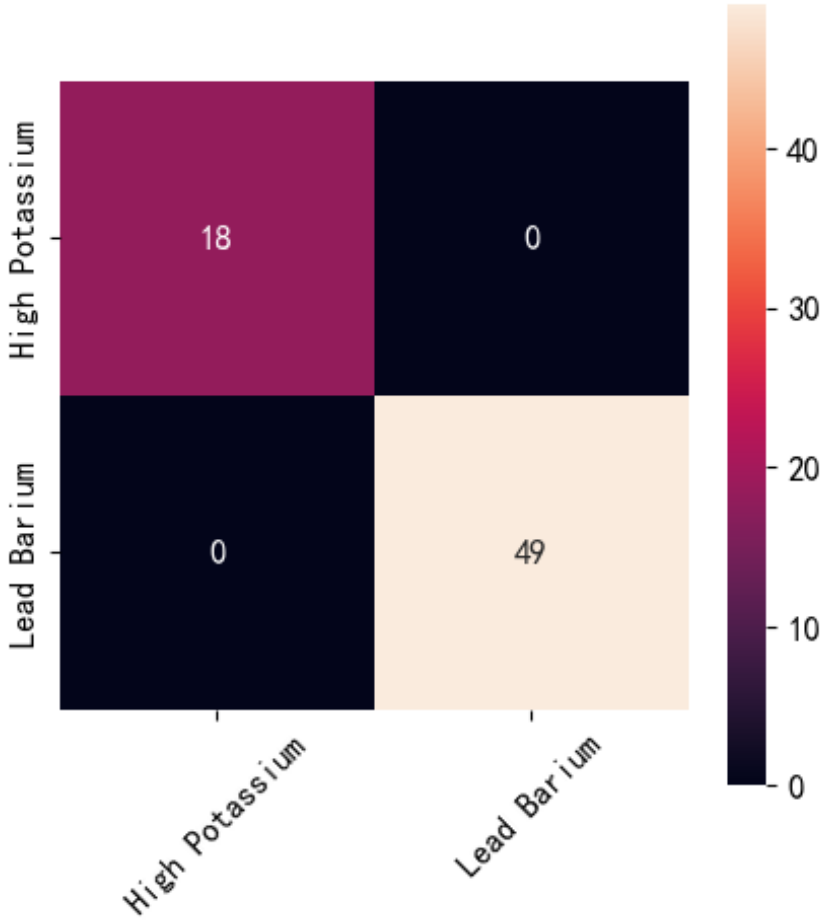

图 16: 决策树预测结果与样本标签混淆矩阵

由混淆矩阵热力图可知，使用决策树预测的结果与样本自身标签完全匹配。

6.3.2 决策树的树状结构导出

通过决策树独特的属性，从 sklearn.tree 中导入函数 plot_tree 作出决策树模型的树状结构，图形如下：

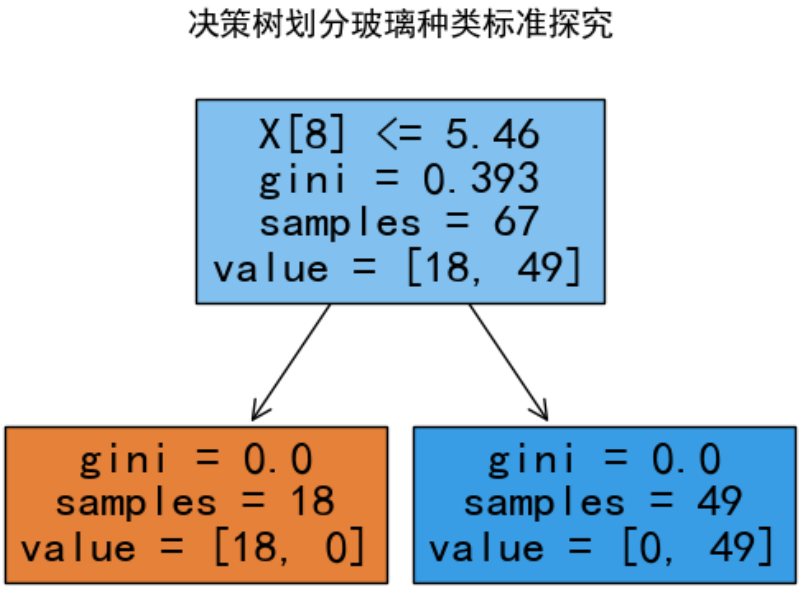

图 17: 决策树树状结构可视化

通过此图形，本文惊人地发现：决策树所挖掘的数据内在的分类标准与之前看似合理、充分的结果相差甚远——玻璃类比的划分居然只与氧化铅的含量有关！这也进一步告诉本文，简单的数据可视化分析往往只能摸索出浅层次的结论，想要挖掘数据深层次的联系必须通过合适的数据与算法。

6.4 聚类分析—层次聚类法

为了对化学成分进行更深层次的分析，又由表 8 可知每类玻璃分别选出 4 个变量进行类别划分，利用 python 中的层次聚类法函数选取三类指标进行划分，分类结果如下表：

	第一类	第二类	第三类
样本编号	03,10,12,13,14,18,21,22,27	07,09	01,04,05,06,15,16,17

表 14: 显性分类规律表

	第一类	第二类
样本编号	02,19,20,23,32,33,37,39,41,42,43,44,47,48,54,55,56,57,58	24,26,30,31,34,35,36,38,40,45,46,

表 15: 铅钡类划分结果表

下面对分类合理性进行检验。

6.4.1 合理性检验

本文的目的是为了对分类结果的合理性和敏感性，在对聚类分析结果进行检验时，仅用轮廓系数判断是有失可信度的，CH 分数、戴维森堡丁指数 (DBI) 是对聚类效果评估好坏的评价另外两种方式，下面利用 python 使用以上三种方法对结果进行检验，得到如下结果：

	轮廓系数	CH 分数	DBI
高钾玻璃	0.448406754	36.65655497	0.87395307
铅钡玻璃	0.446694394	45.85302210	0.82682893

表 16: 检验结果表

由上表可知，两类玻璃所得的轮廓系数接近 0.5，说明同类样本相距比较接近，聚类效果比较好；CH 分数是通过评估类之间方差和类内方差来计算得分，分值越大，表示聚类效果越好，DDB 值越小表示聚类结果同簇内部紧密，不同簇分离较远。即类内距离越小，类间距离越大，综上可得出聚类效果较好，该划分方法合理。

6.4.2 扰动处理

敏感性分析有很多方法可以实现，依题意所言对结果进行敏感性分析，即探究模型的鲁棒性。本文对被预测标签的数据集进行一定程度的扰动处理，再次预测后通过比较干扰前后的标签差异来衡量模型的稳定性。另外，本文不断对各个特征分别加大干扰，以求探究出影响模型划分类别的“阈值”，用表示高钾玻璃的扰动值，其中表示铅钡玻璃的扰动值，利用公式：

$$error1_j = \overline{y_{j1}} * d, d \in (0, +\infty) \tag{8}$$

$$error2_j = \overline{y_{j2}} * d, d \in (0, +\infty) \tag{9}$$

得到每种化学成分的扰动值，带到模型中进行扰动检验，可以得到敏感阈值，最后得出两类玻璃亚类分化的噪音百分比图：

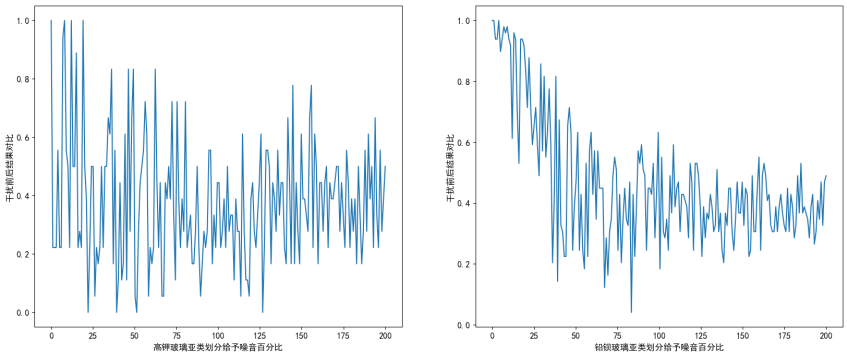

图 18: 左图为高钾玻璃亚类分化噪音百分比图；右图为铅钡玻璃亚类分化噪音百分比图

通过图 18（左）可看出该模型的敏感性阈值 d 在 (0,0.25) 之间，模型的准确率都在 20% 以上，当 d 大于 0.25 时模型的准确率波动性大，预测准确率变化大，因此最好将数据敏感性阈值控制在 0.25 以内。而对于右图可发现模型的敏感性阈值 d 在 (0,0.1) 之间模型的准确率都在 90% 以上，当 d 大于 0.10 时模型的准确率波动性大，准确率变化大，最好将数据敏感性阈值控制在 0.10 以内。

7 问题三建模与求解

7.1 数据预处理

通过 Pandas 库进行数据读取，使用 isnull 函数查找数据中的缺失值，对于附件表单 3 中缺失值进行补 0，并对于表中样本化学成分含量求和，均满足 85%—105%，无需对数据进行删减。

7.2 未知类别文物预测模型

7.2.1 模型准备

模型选择前文从附件表二处理后（补齐缺失值，删除不满足条件只值）的数据中提取的 14 个化学成分和风化情况这 15 个特征。以原数据为标准，将数据进行标准化将数据处理到 0 到 1 之间。对风化情况进行独热编码。

7.2.2 模型建立

投票（voting）在集成学习的分类算法中被广泛运用，投票主要运用软投票（soft voting）是一种对于各基分类器效果融合的模型。本文所选择的 4 个基分类器分别为线性判别分析、决策树、朴素贝叶斯以及支持向量机。并用 voting 进行基分类器融合得到一个强分类器。

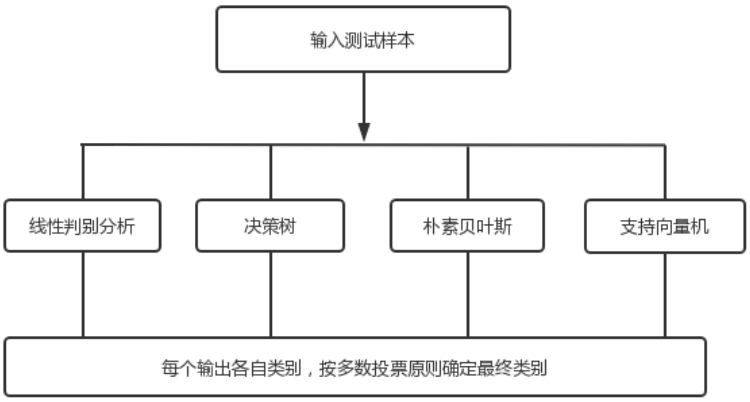

图 19: Voting 原理

7.2.3 模型求解

建立线性判别分析、决策树、朴素贝叶斯以及支持向量机四个性能优良的传统机器学习分类模型，并将训练集分别投入四个模型进行模型训练，通过比较交叉验证后的得分结果，发现四个模型性能均远超过随机分类（成功率 0.5）。

分类器	综合准确率
线性判别分析	0.971429
决策树	0.995382
朴素贝叶斯	0.985714
支持向量机	0.985714

表 17: 基分类器评分

从上表可以看出四种基分类器效果都不错，对于测试集的数据训练准确率都在百分之九十七以上，都是比较好的分类预测模型。

然而四个分类器的分类效果各不相同，所以本着训练性能最优良分类器的原则，本文使用基于结果进行模型融合的 VotingClassifier 投票算法对四个模型进行融合，公式为：

$$W_i = \frac{accuracy_i}{\sum_{i=1}^{5} accuracy_i} \tag{10}$$

软投票的分类结果如下表：

分类器	综合准确率
VotingClassifier	1

表 18: 分类结果表

最后得到了分类预测成功率为 1 的“完美模型”。下面对于该分类模型进行评价，利用混淆矩阵热力图如下：

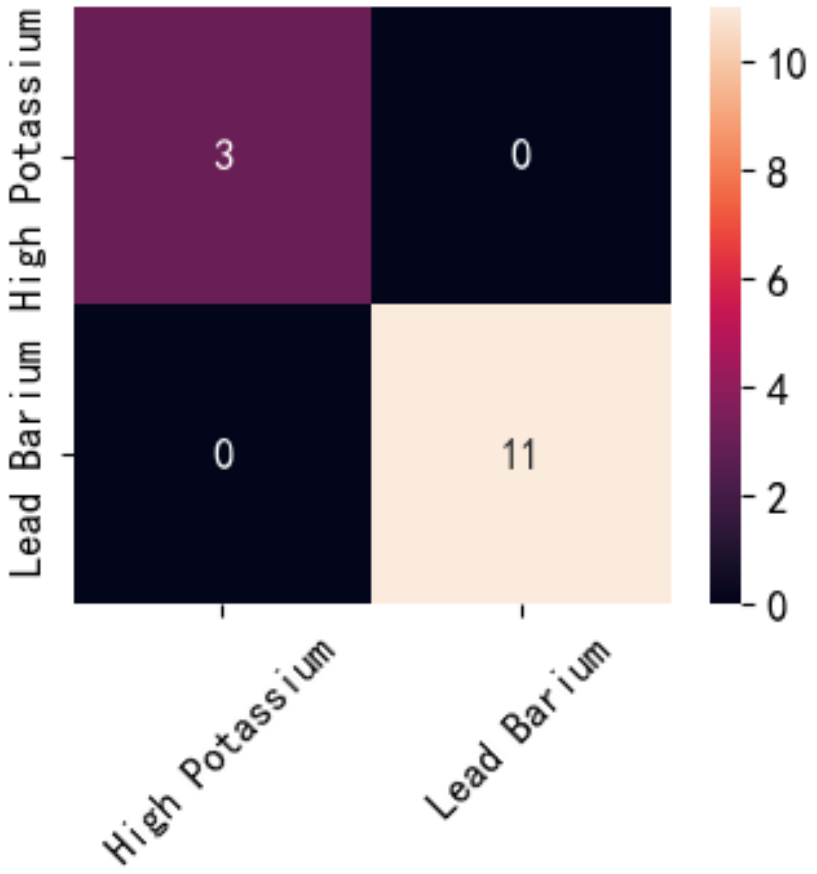

图 20: VotingClassifier 混淆矩阵热力图

可以看出 VotingClassifier 分类模型非常完美，是一个非常好的分类器。再基于混淆矩阵的方法对该分类模型进行评价，结合对测试集的预测结果和测试集自身标签、通过计算得到模型的精确度得到如下指标：

	精确度	召回率	F1 分数	支持率
铅钡	1	1	1	1
高钾	1	1	1	1
加权平均	1	1	1	1

表 19: 评分数据表

通过表 11 可以看到该分类器确实非常“完美”。利用该模型对预处理过后的附件表三进行分类预测，预测结果如下：

高钾玻璃	A_1, A_6, A_7
铅钡玻璃	A_2, A_3, A_4, A_5, A_8

表 20: 预测结果表

由上表可以知道成功预测 A_1, A_6, A_7 为高钾玻璃，A_2, A_3, A_4, A_5, A_8 为铅钡玻璃。

7.3 模型敏感性分析

利用公式（1）（2）得到每种化学成分的扰动值，带到模型中进行扰动检验，可以得到敏感阈值。对数据进行扰动，将扰动值重新带到模型中进行分类预测，得到干扰前后的结果图如下：

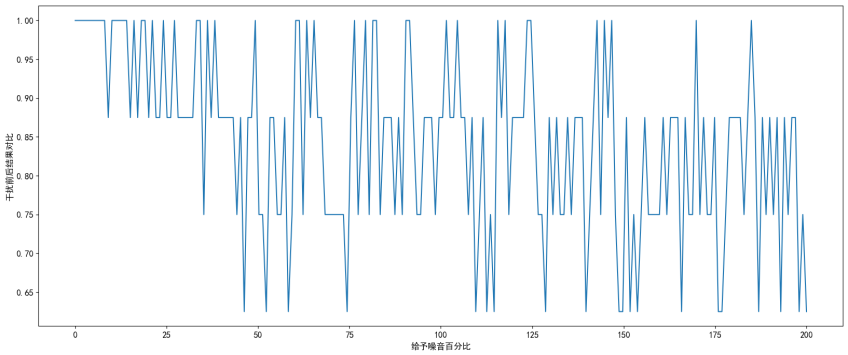

图 21: 敏感性阈值图

通过图 21 发现该模型的敏感性阈值 d 在 (0,0.3) 之间，模型的准确率都在 85% 以上，当 d 在 (0.3,0.45) 之间模型的准确率都在 75% 以上，而当 d 大于 0.45 时模型的准确率波动性大，预测准确率变化大，因此最好将数据的敏感性阈值控制在 0.45 以下。

8 问题四建模与求解

为分析不同类别的玻璃样本的化学成分的关联关系，并比较它们之间关联关系的差异性，利用灰色关联性分析模型并比较它们之间的差异性，做出流程图如下：

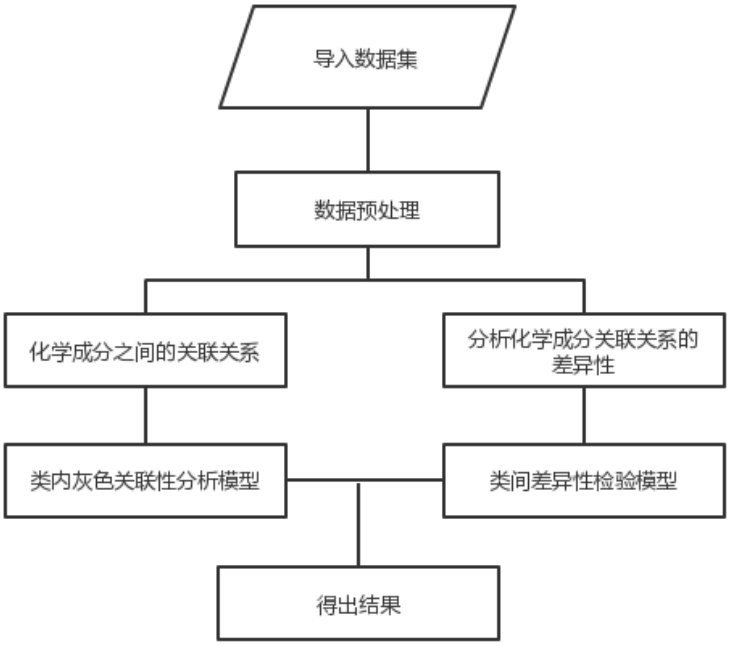

图 22: 问题四流程图

8.1 数据预处理

上述问题中处理后（补齐缺失值，删减异常值）的高钾玻璃数据表和铅钡玻璃数据表来进行类内的关联性分析，另外进行类间相关性分析。

8.2 类内灰色关联性分析模型

8.2.1 模型准备

对于类内各数据之间的关系，常见思路是进行相关系数矩求解，而求解相关性矩阵要进行正态性检验，本问对高钾玻璃和铅钡玻璃分别进行正态性检验，经过检验发现大部分特征都不满足正态分布（具体数据见附件）。因此，本问决定采用灰色关联预测进行关联性分析。

8.2.2 模型建立

对于灰色关联分析来分析类内的关联性，考虑到类内中每个化学成分之间都可能有关联性，因此用 14 个特征中随机抽取一个特征与其他特征进行灰色关联性分析。即有：

1. 将数据进行预处理:

$$\widetilde{e_{kj}} = \frac{e_{kj}}{\overline{e_j}}, \overline{e_j} = \frac{1}{n_z}\sum_{k=1}^{n} e_{kj}(k = 1, 2, \cdots, n_z) \tag{11}$$

$$\widetilde{f_k} = \frac{f_k}{\overline{f_j}}, \overline{f_j} = \frac{1}{n_z}\sum_{k=1}^{n} f_k. \tag{12}$$

2. 确定母序列和子序列:

$$E = [e_1, e_2, \cdots, e_{13}]^T. \tag{13}$$

$$F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1,13} \\ f_{21} & f_{22} & \cdots & f_{2,13} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n_z1} & f_{n_z2} & \cdots & f_{n_z13} \end{bmatrix} \tag{14}$$

3. 计算序列和母序列的关联系数:

$$r = \min_j \min_k |f_0(k) - f_j(k)| \tag{15}$$

$$b = \max_j \max_k |f_0(k) - x_j(k)| \tag{16}$$

4. 计算关联度:

(1) 构造:

$$a_{k,q} = \xi_q(k) = \frac{r + \rho b}{|e_{kq} - f_q| + \rho b}. \tag{17}$$

(2) 计算关联度:

$$R = \frac{1}{n_z}\sum_{k=1}^{n} \xi_q(k) = \frac{1}{n_z}\sum_{k=1}^{n} a_{kq}. \tag{18}$$

8.2.3 模型求解

根据公式可以分别算出高钾与铅钡的类内的灰色关联结果（下表为高钾玻璃灰色关联结果，铅钡结果见附件）

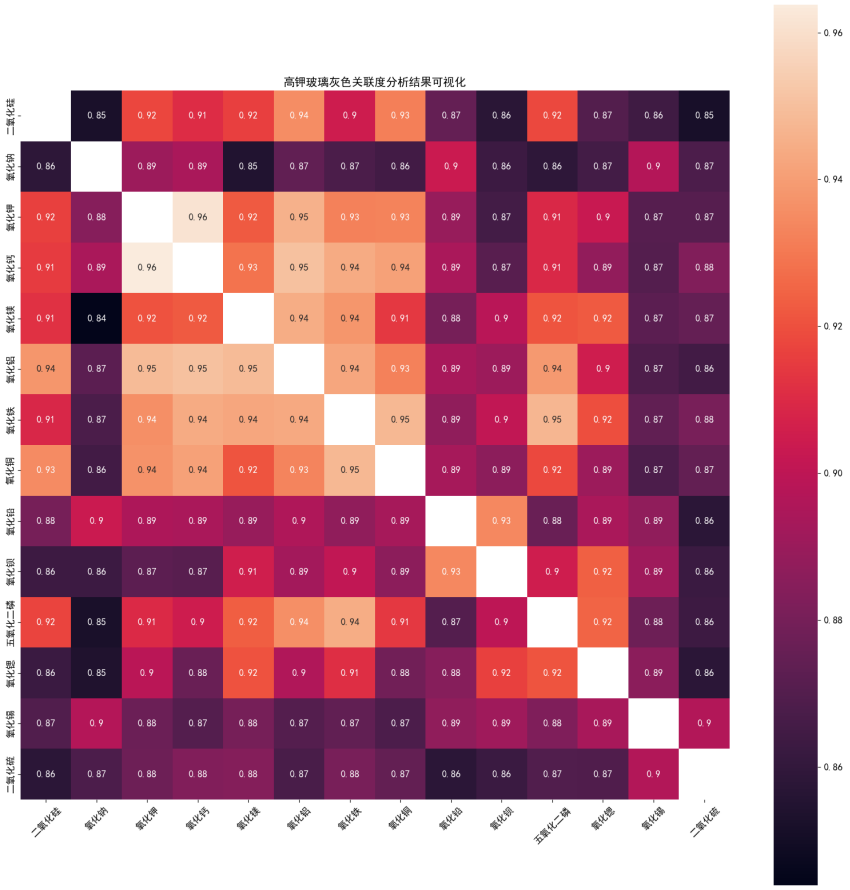

图 23: 高钾玻璃关联度分析结果可视化图

由图 20 可以看到每个化学成分之间的关联度结果最低为 0.84，即整个类内的化学物质之间都存在的显著关联性，可以判断出在同类化学成分之间每种成分之间都存在关联关系。

8.3 类间差异性检验模型

8.3.1 模型建立

本问基于类内灰色关联得到的高钾类各化学成分与铅钡类化学成分的关联系数组成新的关联的系数表，需要对类间的同一化学成分关联性的差异性进行分析，配对 t 检验是一种用于配对定量数据之间的差异对比关系的优良检验方法。首先需要对两类数据进行正态性检验，检验之后发现都不满足正态分布。需要使用其他检验方法，因此使用非参数检验，本问所求两种类型同一化学成分的差异性可以采用非参数配对样本 Wilcoxon 符号秩检验进行检验。

8.3.2 模型求解

利用 spasspro 中的非参数检验配对样本 Wilcoxon 符号秩检验对于新表数据进行差异性检验，得出两类玻璃的同一化学物质的差异性显著系数表。

化学成分	值（显著性）	Cohen's d 值（差异幅度）
二氧化硅	0.116	0.427
氧化钠	0.221	0.23
氧化钾	0.101	0.373
氧化钙	0.600	0.139
氧化镁	0.507	0.088
氧化铝	0.753	0.013
氧化铁	0.087*	0.359
氧化铜	0.807	0.059
氧化铅	0.013**	0.727
氧化钡	0.016**	0.625
五氧化二磷	0.507	0.033
氧化锶	0.064*	0.453
氧化锡	0.116	0.209
二氧化硫	0.002***	0.56

表 21: 差异性显著系数表

由 $p > 0.05$ 可以判定上表两种类间化学成分中二氧化硅、氧化钠、氧化钾、氧化钙、氧化镁、氧化铝、氧化铁、氧化铜、五氧化二磷、氧化锶、氧化锡不能拒绝显著性假设，故两个量之间不存在显著性差异。

由 $p \leq 0.5$ 可以判定上表两种类间化学成分中氧化铅、氧化钡、二氧化硫能拒绝显著性假设，故两个量之间存在显著性差异。

当两个变量之间具有显著性差异时，Cohen's d 值一般都较大，说明两个变量之间有较大的差异幅度。

9 模型评价和改进

9.1 模型优点

1. 在问题一中对文物表面是否风化与类型、颜色、纹饰关系分析过程中，不仅对于单变量之间进行了分析，还进一步用树模型进行了多变量与单变量的分析，同时利用互信息进一步对结果进行检验，提高模型的合理性。

2. 本文做了大量图表来统计分析数据特点，直观的对比出两类玻璃风化前后的化学成分的变化量以及各类玻璃化学成分之间的关系。

3. 在文本中多次对模型进行调参，利用混淆矩阵和多个指标检验，提高了模型的准确性。

4. 使用强分类器，构建 Voting 集成算法，得到一个完美模型，得到结果可信度高。

9.2 模型缺点

1. 做编码时由于颜色样本有 7 个非叙述类别，而数据集中存在大量的分类特征，没有找到合理高效的特征编码方式。

2. 本文在补充颜色缺失值时，鉴于分布分析补充黑色，而实际上在工业上常使用众数补充。

9.3 模型改进

在查阅文献时，涉猎了一种工业上常见新进的特征编码方式——目标编码，此方式通过计算类别出现的频率，对高维分类特征进行编码，避免出现维数灾难，同时也不会导致类别间的欧几里得距离过大。相信使用这种编码会让问题一中结果更具说服力。

10 参考文献

[1] 伏修锋, 干福熹. 基于多元统计分析方法对一批中国南方和西南地区的古玻璃成分的研究 [J]. 文物保护与考古学 2006（04）.

[2] 司守奎，孙玺菁. 数学建模算法与应用（第 3 版）——北京：国防工业出版社，2021.4.

[3] 周志华著；李楠译. 集成学习：基础与算法——北京：电子工业出版社，2020.8.

[4] 司守奎，孙玺菁.Python 数学建模算法与应用——北京：国防工业出版社，2022.1.

[5] 王贺，刘鹏，钱乾著. 机器学习算法竞赛实战——北京：人民邮电出版社，2021.9.

[6] 何道江，黄旭东，张琼编. 数学建模优秀论文选编——北京：科学出版社，2020.11.

[7] 何伟，张良均主编. 机器学习原理与实践——北京：人民邮电出版社，2021.7.

[8] 孙玉林，余本国著.Python 机器学习算法与实践——北京：电子工业出版社，2021,9.