1. 数据预处理

在建模前，我们对样本数据进行缺失值和异常值的检查工作。其中在表单1中出现部分样本的“颜色“列出现缺失，应对其进行缺失值处理。在表单2中，由于部分样本的化学成分未观察到，因此在表格当中呈现为缺失状态，因此需对其进行处理。同时，根据题目的定义，有效数据是指成分比例累加和介于85%~105%之间的数据，但我们发现在样本中部分出现化学成分的比例累计和不在此范围中，因此需要对其进行异常样本的处理。

* 1. 缺失值处理

缺失值处理的方式通常分为填充与剔除。在本文所提供的数据中，由于样本数量较少，因此采取填充的方式进行处理。

在表单1中，我们对缺失值进行处理。在样本中筛选出与出现缺失值的样本同属性的样本，包括同纹饰、同类型、同表面风化状态的其它样本。从中取出现频率最大的颜色属性作为该样本的颜色属性，将其进行填充，填充结果如下：

表单1填充结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 文物编号 | 纹饰 | 类型 | 颜色 | 表面风化 |
| 19 | A | 铅钡 | 浅蓝 | 风化 |
| 40 | C | 铅钡 | 浅蓝 | 风化 |
| 48 | A | 铅钡 | 浅蓝 | 风化 |
| 58 | C | 铅钡 | 浅蓝 | 风化 |

在表单2中，我们对全部出现空缺的地方，填充为0，以方便后面的计算与建模。

* 1. 异常样本处理

在本文中，我们根据有效数据的定义对异常样本进行查找，其中发现，15号和17号样本出现化学成分累计和小于85%的情况，因此认为这两个样本属于无效样本，因此对其进行剔除。

异常样本

|  |  |
| --- | --- |
| 文物采样点 | 化学成分比例累计和 |
| 15 | 79.47 |
| 17 | 71.89 |

* 1. 合并样本

对表单1和表单2进行合并处理，我们根据表单2的文物采样点作为连接列进行合并，合并得到的表格可以为后续的建模提供方便，合并效果如下：

表单1和表单2合并

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 文物编号 | 纹饰 | 类型 | 颜色 | 表面风化 | 文物采样点 | 二氧化硅(SiO2) | … | 二氧化硫(SO2) |
| 01 | C | 高钾 | 蓝绿 | 无风化 | 01 | 69.33 | … | 0.39 |
| 02 | A | 铅钡 | 浅蓝 | 风化 | 02 | 36.28 | … | 0 |
| 03 | A | 高钾 | 蓝绿 | 无风化 | 03部位1 | 87.05 | … | 0 |
| 03 | A | 高钾 | 蓝绿 | 无风化 | 03部位2 | 61.71 | … | 0 |
| … | … | … | … | … | … | … | … | … |
| 55 | C | 铅钡 | 绿 | 无风化 | 55 | 49.01 | … | 0 |
| 56 | C | 铅钡 | 蓝绿 | 风化 | 56 | 29.15 | … | 0 |
| 57 | C | 铅钡 | 蓝绿 | 风化 | 57 | 25.42 | … | 0 |
| 58 | C | 铅钡 | 浅蓝 | 风化 | 58 | 30.39 | … | 0 |