

澳 門 科 技 大 學

Macao University of Science and Technology

数据科学常用工具期末报告

Final Report on Spatial Database Systems

組 長： 吴百恒

組 員： 黄熙 李杰相 彭嘉男

學 院： 太空科學研究所

課 程： 数据科学常用工具

完成日期： 27/5/2025

目 录

[目 录 2](#_Toc199497982)

[1 前言 4](#_Toc199497983)

[**1.1** 作业概述 4](#_Toc199497984)

[1.2 环境适配 4](#_Toc199497985)

[**1.3** 报告撰写格式 4](#_Toc199497986)

[2 任务实现流程 5](#_Toc199497987)

[**2.1** 数据预处理 5](#_Toc199497988)

[**2.1.1** 读取数据 5](#_Toc199497989)

[**2.1.2** 定义单位映射 5](#_Toc199497990)

[**2.1.3** 缺失值处理 6](#_Toc199497991)

[**2.1.4** CLR变换 7](#_Toc199497992)

[**2.1.5** 异常值标记 8](#_Toc199497993)

[2.2 探索性分析 8](#_Toc199497994)

[**2.2.1** 散点矩阵图 8](#_Toc199497995)

[**2.2.2** 相关性热力图 9](#_Toc199497996)

[**2.2.3** PCA双标图 11](#_Toc199497997)

[**2.2.4** 地球化学比值图 11](#_Toc199497998)

[2.3 模型训练准备 12](#_Toc199497999)

[**2.3.1** 单位均同 12](#_Toc199498000)

[**2.3.2** 对数变换 12](#_Toc199498001)

[**2.3.3** 异常值影响分析 13](#_Toc199498002)

[2.4 模型训练-随机森林 14](#_Toc199498003)

[**2.4.1** 最佳树数量确定 14](#_Toc199498004)

[**2.4.2** 随机森林最佳模型预测结果展示 14](#_Toc199498005)

[2.5 模型训练-XGBoost 17](#_Toc199498006)

[**2.5.1** 最佳确定 17](#_Toc199498007)

[**2.5.2** 最佳模型保存与预测 18](#_Toc199498008)

[2.6 模型训练-SVM 21](#_Toc199498009)

[**2.6.1** 模型实现 21](#_Toc199498010)

[**2.6.2** 保存SVM最佳模型与预测 21](#_Toc199498011)

[2.7 模型训练-Keras神经网络 23](#_Toc199498012)

[**2.7.1** 模型训练 23](#_Toc199498013)

[**2.7.2** 保存最佳模型与预测 23](#_Toc199498014)

[2.8 SHAP分析-XGBoost 26](#_Toc199498015)

[**2.8.1 SHAP**分析流程 26](#_Toc199498016)

[**2.8.2 SHAP**分析结果 26](#_Toc199498017)

[2.9 SHAP分析-神经网络 27](#_Toc199498018)

[**2.9.1 SHAP**分析流程 28](#_Toc199498019)

[**2.9.2 SHAP**分析结果 28](#_Toc199498020)

[2.10 SHAP分析结果解释 29](#_Toc199498021)

[**2.10.1** XGBoost和神经网络的相似结果 30](#_Toc199498022)

[**2.10.2** 地质意义解释 30](#_Toc199498023)

[2.11 模型对比分析 30](#_Toc199498024)

[2.12 可执行程序实现 31](#_Toc199498025)

[**2.12.1** 程序逻辑介绍 31](#_Toc199498026)

[**2.12.2** 程序打包逻辑介绍 32](#_Toc199498027)

[**2.12.3** cli\_predict.exe程序运行展示 33](#_Toc199498028)

[**2.12.4** 网页版本运行展示图 36](#_Toc199498029)

[3 小组任务分工 38](#_Toc199498030)

[**3.1** 小组成员介绍 38](#_Toc199498031)

[**3.2** 具体任务分工 38](#_Toc199498032)

[**3.2.1** 团队协作方式概述 38](#_Toc199498033)

[**3.2.2** 数据预处理阶段分工 39](#_Toc199498034)

[**3.2.3** 探索性分析阶段分工 39](#_Toc199498035)

[**3.2.4** 模型训练准备阶段分工 40](#_Toc199498036)

[**3.2.5** 模型训练-随机森林阶段分工 40](#_Toc199498037)

[**3.2.6** 模型训练-XGBoost阶段分工 41](#_Toc199498038)

[**3.2.7** 模型训练-SVM阶段分工 41](#_Toc199498039)

[**3.2.8** 模型训练-Keras神经网络阶段分工 42](#_Toc199498040)

[**3.2.9** SHAP分析实现阶段分工 42](#_Toc199498041)

[**3.2.10** SHAP结果分析与模型对比阶段分工 43](#_Toc199498042)

[**3.2.11**可执行程序阶段分工 43](#_Toc199498043)

[参 考 文 献 45](#_Toc199498044)

[总 结 46](#_Toc199498045)

1 前言

**1.1** 作业概述

使用包含445个斑岩岩石样品数据集（每个样品包含36个主、微量元素特征）构建一个端到端工作流，该流程需实现：区分未知样品为Cu-rich或Au-rich； 解释主要控矿地球化学特征；交付一个可离线运行的预测工具。

1.2 环境适配

PC：Microsoft Win10/11

运行环境：Python3.10.16

Server：Ubuntu 24.04.1 LTS

运行环境：Python3.11.11

**1.3** 报告撰写格式

标题：

字体使用黑体，字号由小三递减

1.5倍行距，大标题段后1行，其余段前0.5行

正文：

汉字：宋体小四 1.25倍行距

英文、数字：Times New Roman

参考文献：

APA格式，用方括号“[]”和数字按顺序以右上角标形式标注在引用处

2 任务实现流程

**2.1** 数据预处理

**2.1.1** 读取数据

构建本地模块data\_preprocessing.loader，借助os以及pandas两个库文件，实现根据文件后缀加载数据功能模块。功能见表2.1，试运行展示见图2.1。

|  |  |
| --- | --- |
| 功能 | 详细描述 |
| 1.多方式输入支持 | 支持文件流、本地路径 |
| 2.文件扩展名识别 | 提取扩展名并转为小写 |
| 3.主体读取、异常处理 | 根据扩展名选择读取方式，限制支持Excel和CSV文件 |
| 4.内容初判 | 判断文件是否为空，并设定状态码值 |
| 5.异常捕获 | 针对不同异常返回专属状态码 |

表2.1 本地模块data\_preprocessing.loader功能介绍

***亮点自述：***

1.支持本地路径和流式对象

2.设定多种异常分析以及专属状态码，方便问题查找

3.设定标准化返回，返回执行信息，成功或错误描述

日历

AI 生成的内容可能不正确。

图2.1 读数据模块试运行结果图

**2.1.2** 定义单位映射

设置专属映射字典，定义每个化学元素/氧化物的计量单位。试运行展示见图2.2、2.3。

***亮点自述：***

1. 为数据的每一列指定计量单位，便于后续数据解释和处理。
2. 指定后调用专属校验模块验证单位信息、数据类型等是否设置成功
3. 设定基础以及专属异常分析以及对应状态码
4. 设定标准化返回

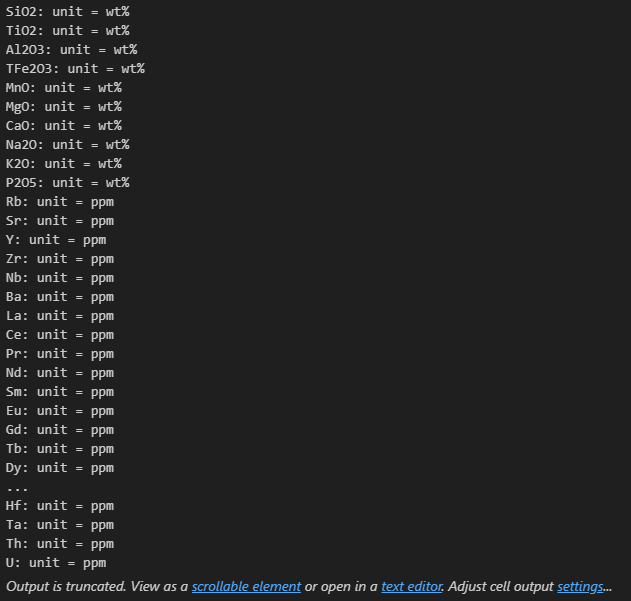


图2.2 单位映射试运行结果图

文本

AI 生成的内容可能不正确。

图2.3 校验验试运行结果图

**2.1.3** 缺失值处理

设置专属缺失值处理模块，自动检测所有数值列，并支持检测指定列；对缺失值支持“填0”“填中位数”处理。试运行展示见图2.4。

***亮点自述：***

1. 支持自动检测所有列以及指定列
2. 支持多种方式处理缺失值
3. 准确反馈状态和日志

文本

AI 生成的内容可能不正确。

图2.4 缺失值处理试运行结果图

**2.1.4** CLR变换

针对数据进行CLR变换，消除比例型数据的闭合效应。试运行展示见图2.5。

***亮点自述：***

1.在功能模块中实现单位自动统一、CLR变换、常规对数变换（任意底数）

2.不改变原始数据，可对针对专属列进行操作

3.支持元数据跟随转换，不会“换算了数值却忘了改单位”

4.增加非正值处理，防止脏数据报错（参数可兼容科研中零/极小异常值）

5. 同步索引，适合后续直接拼接

图形用户界面

AI 生成的内容可能不正确。

图2.5 CLR变换试运行结果图

**2.1.5** 异常值标记

针对数据进行检测标记，按Z分数阈值3.0检测异常值。试运行展示见图2.6。

***亮点自述：***

1. 检测双侧异常（取绝对值）
2. 定义专属列标记异常值

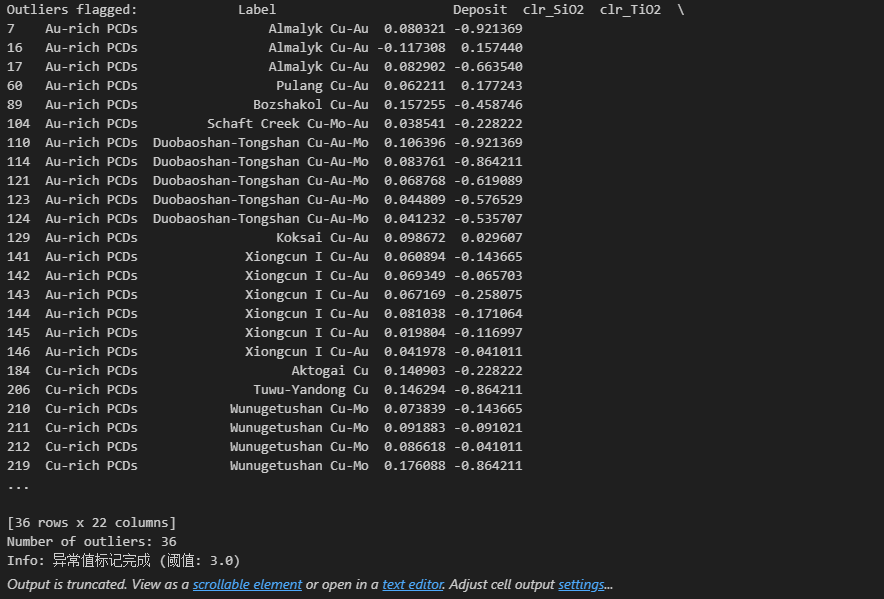


图2.6 异常值标记试运行结果图

2.2 探索性分析

**2.2.1** 散点矩阵图

为分析元素对之间的线性、非线性关系、不同类别样品分布规律、区分度，以及快速检查异常，选择了SiO2,TiO2,Al2O3,TFe2O3,MnO,MgO,CaO,Na2O,K2O,P2O5，十种元素生成散点矩阵，见图2.7。

由图可以看出：Au-rich和Cu-rich PCDs在部分组合上可以明显区分，MgO、CaO、SiO2与其它变量组合的散点关系较强，可能主控样品的分组，并且可以看到部分子图中有孤立的点。



图2.7 散点矩阵

**2.2.2** 相关性热力图

为分析元素之间整体相关性结构，尽可能找出可能的矿物组合、分带、迁移趋势，分别使用皮尔逊和斯皮尔曼法对上述元素生成相关性热力图，见图2.8、2.9。

根据两张相关性热力图，可以发现：大多数主量氧化物之间存在明显相关性，但相关性强弱、正负方向有差异。其中TiO2与MgO、TFe2O3，以及MgO与TFe2O3三者之间相关性最强（皮尔逊0.83/0.79，斯皮尔曼0.85/0.82），表明这些元素（及其组合）在样品中共变性强，可能源于同一地球化学过程、同源分带或矿物共生；SiO2与TiO2、TFe2O3、MgO等，都有很强的负相关（皮尔逊-0.78/-0.78/-0.75，斯皮尔曼-0.75/-0.81/-0.78），这显示SiO₂的增多通常伴随这些元素的减少，可能反映了地球化学分异、岩浆演化或不同矿床成因。

图表, 条形图, 树状图

AI 生成的内容可能不正确。

图2.8 皮尔逊法-相关性热力图

图表, 条形图, 树状图

AI 生成的内容可能不正确。

图2.9 斯皮尔曼法-相关性热力图

**2.2.3** PCA双标图

为查看样品是否有天然分群、查看哪些元素主导主要变异，分析可疑异常样品，生成PCA双标图，见图2.10。

由图可知，Au与Cu两类样品在主成分空间可见明显分布趋势，PC1方向上有判别能力。

图表, 散点图

AI 生成的内容可能不正确。

图2.10 PCA双标图

**2.2.4** 地球化学比值图

为判断分异趋势、成因路径，查看组分分布是否有分段、突变等情况，生成地球化学比值图，见图2.11。

左图可知：大部分数据点的K2O/Na2O分布在0到3之间，仅有少量离群值；SiO2含量从50%到75%不等，K2O/Na2O的变化主要集中在SiO2 60%到70%区间。说明这两类PCDs在K2O/Na2O与SiO2的关系上区分度有限，即K/Na比不能单独很好地区分这两类矿床类型。

右图可知：Sr/Y的值分布很广，大部分样品集中在0到120之间，也有少数极高离群值；样品点在SiO2 60%到70%范围内分布最密集，Sr/Y在这个区间内有较大的变异。总体来说，Sr/Y和SiO2的联合分布也未表现出特别明显的类型分带。

综上所述，两张比值图揭示了K2O/Na2O和Sr/Y这两个常用地球化学比值随SiO2含量的变化特征。

图表, 散点图

AI 生成的内容可能不正确。

图2.11 地球化学比值图

2.3 模型训练准备

**2.3.1** 单位均同

为保证所有后续数学分析、统计、机器学习和深度学习处理前提成立。使不同来源、数量级的元素数据可以公平、准确地参与模型训练和解释，必须要进行单位均同，运行展示见图2.12。

屏幕上有字

AI 生成的内容可能不正确。

图2.12 单位均同运行结果图

**2.3.2** 对数变换

为消除数量级影响，减小极差，避免训练时出现“强信号被噪声淹没”，使模型训练和特征评估更可靠，对数据进行对数变换，运行展示见图2.13。

***亮点自述：***

1.在变换过后，对异常值进行二次检测与去除

图形用户界面, 文本

AI 生成的内容可能不正确。

图2.13 对数变换运行结果图

**2.3.3** 异常值影响分析

为查看异常值对结果的影响，对异常值清理前后的数据结果，同时进行随机森林分类（使用相同的树数量）。评估异常值处理对模型分类性能的影响（见图2.14）。

日历

AI 生成的内容可能不正确。

图2.14 异常值清理前后对比图

2.4 模型训练-随机森林

**2.4.1** 最佳树数量确定

树越多，模型越稳定，但计算成本越高。树太少可能欠拟合，过多又可能过拟合，影响模型性能，故模型训练过程见表2.2。

|  |
| --- |
| 模型训练过程表 |
| 1.设定随机森林树的数量从1到100，步长为1。 |
| 2.选择清理异常值后的数据为样本，将数据80%作为训练集，20%作为测试集。 |
| 3.为适应分类器，将类别标签转为0/1。 |
| 4.循环测试不同树数量模型。 |
| 5.计算训练集上的分类误差，用5折交叉验证评估模型泛化误差，并记录。 |

表2.2 模型训练过程表

根据验证误差曲线（图2.15）选择38棵树最合适，因为此时训练集和验证集的误差都不再显著下降，同时防止过拟合，是合理的选择。

图表

AI 生成的内容可能不正确。

图2.15 随机森林模型误差曲线

**2.4.2** 随机森林最佳模型预测结果展示

在选择好最佳树数量后，将最好模型保存（图2.16），计算并绘制测试集分类结果的混淆矩阵（图2.17）、AUC&ROC曲线（图2.18）、精确率-召回率曲线（图2.19）、特征重要性柱状图（图2.20）。

日历

AI 生成的内容可能不正确。

图2.16 保存最好模型

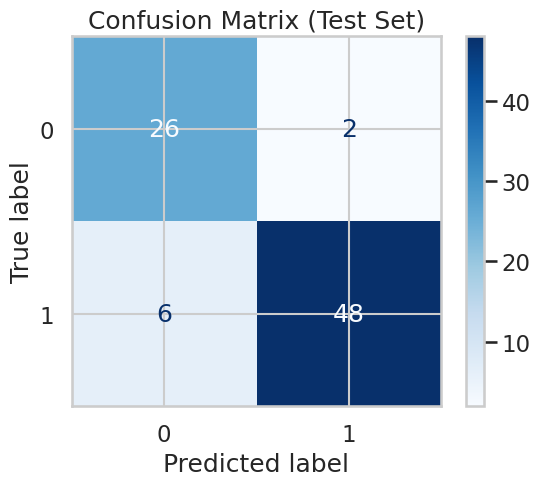


图2.17 混淆矩阵

图表

AI 生成的内容可能不正确。

图2.18 AUC&ROC曲线

图表

AI 生成的内容可能不正确。

图2.19 精确率-召回率曲线

图表, 条形图

AI 生成的内容可能不正确。

图2.20 特征重要性柱状图

2.5 模型训练-XGBoost

**2.5.1** 最佳确定

与随机森林相似，选择测试1到100棵树的模型，将数据80%为训练集20%为测试集，循环遍历不同树数量训练XGBoost分类器，记录训练误差，5折交叉验证记录验证误差，由可视化结果可知（图2.21）：当n在30左右时，训练集和验证集的误差都不再显著下降，模型的性能也不再提升。此时，是最佳参数。图表

AI 生成的内容可能不正确。

图2.21 XGBoost模型误差曲线

**2.5.2** 最佳模型保存与预测

在选择好最佳模型后，将最好模型保存（图2.22），计算并绘制测试集分类结果的混淆矩阵（图2.23）、AUC&ROC曲线（图2.24）、精确率-召回率曲线（图2.25）、特征重要性柱状图（图2.26）。

日历

AI 生成的内容可能不正确。

图2.22 最佳模型保存

日历

AI 生成的内容可能不正确。

图2.23 混淆矩阵

图表

AI 生成的内容可能不正确。

图2.24 AUC&ROC曲线

图表

AI 生成的内容可能不正确。

图2.25 精确率-召回率曲线

图表, 条形图

AI 生成的内容可能不正确。

图2.26 特征重要性柱状图

2.6 模型训练-SVM

**2.6.1** 模型实现

设计SVM专属自动调参、训练、并输出详细评估的SVM分类器训练函数模块,详细过程见表2.3。

|  |  |
| --- | --- |
| 详细过程 | 详细描述 |
| 1. 数据集分割 | 将数据随机分为训练集和测试集。 |
| 2. 数据标准化 | 对特征做零均值单位方差标准化，防止SVM受变量量纲影响。 |
| 3. 参数网格设置 | 只搜索RBF核，遍历C、gamma多个参数组合。 |
| 4. 网格搜索自动调参并训练 | 5折交叉验证自动寻找最佳超参数（全核并行）。 |
| 5. 模型预测与评估 | 用测试集预测标签，输出最优参数、最佳交叉验证分数、详细分类报告、混淆矩阵、测试集准确率。 |

表2.3 SVM模型实现过程

**2.6.2** 保存SVM最佳模型与预测

在选择好SVM最佳模型后，将最好模型保存（图2.27），计算并绘制测试集分类结果的混淆矩阵（图2.28）、AUC&ROC曲线（图2.29）、精确率-召回率曲线（图2.30）。

日历

AI 生成的内容可能不正确。

图2.27 最佳模型保存

图示

AI 生成的内容可能不正确。

图2.28 混淆矩阵

图表

AI 生成的内容可能不正确。

图2.29 AUC&ROC曲线

图表

AI 生成的内容可能不正确。

图2.30 精确率-召回率曲线

2.7 模型训练-Keras神经网络

**2.7.1** 模型训练

使用Keras构建深度学习二分类模型，并自动调参，具体过程见表2.4。

|  |  |
| --- | --- |
| 详细过程 | 详细描述 |
| 1. 数据准备 | 取特征和标签，将标签转为0/1（二分类），并切分80%训练/20%测试集。 |
| 2. KerasTuner超参数搜索 | 随机搜索隐藏层数、每层神经元数、激活函数、dropout、学习率等超参数。 |
| 3. 防止过拟合 | 只搜索RBF核，遍历C、gamma多个参数组合。 |
| 4. 网格搜索自动调参并训练 | EarlyStopping防止过拟合。 |
| 5. 保存最佳模型 | 用joblib保存最佳模型到本地，方便后续调用或部署。 |

表2.4 深度神经网络实现过程

**2.7.2** 保存最佳模型与预测

在选择好最佳模型后，将最好模型保存（图2.31），计算并绘制测试集分类结果的混淆矩阵（图2.32）、AUC&ROC曲线（图2.33）、精确率-召回率曲线（图2.34）。

文本

AI 生成的内容可能不正确。

图2.31 最佳模型保存

图片包含 日历

AI 生成的内容可能不正确。

图2.32 混淆矩阵

图表

AI 生成的内容可能不正确。

图2.33 AUC&ROC曲线

图表

AI 生成的内容可能不正确。

图2.34 精确率-召回率曲线

2.8 SHAP分析-XGBoost

**2.8.1 SHAP**分析流程

首先加载数据和训练好的XGBoost模型，然后创建一个SHAP解释器，用训练好的模型解释全体数据，得到SHAP值，从而分析每个样本、每个特征的“对模型输出的贡献”

**2.8.2 SHAP**分析结果

根据上述流程，输出蜜蜂群图（图2.35）以及SHAP依赖图（图2.36、2.37）

图表

AI 生成的内容可能不正确。

图2.35 蜜蜂群图

在蜜蜂群图中，每一行为一个变量，每个点表示一条样本记录，颜色表示该变量在该样本中的值高低（红=高，蓝=低），X轴为对模型输出的影响（SHAP值），后续不再做赘述。

首先可以直观地发现以下几点：

1. TiO₂的高值（红色）向右偏移，表示对“富金”判断贡献更大；说明高TiO₂倾向于富金，但相对而分界并不清晰不算太大。
2. MgO和TFe₂O₃：低值倾向于负向影响（即富铜），说明富铜样本在这些元素上MgO/TFe₂O₃含量较低。
3. Al2O3与TiO₂的贡献相反，但是分界更清晰。
4. SiO₂的但是影响力相对较少，但仍能观察出高值为正（富金），低值为负（富铜），说明SiO₂升高与金富集相关。

接下来，我们对其中的两两组合进行分析。

图表, 散点图

AI 生成的内容可能不正确。

图2.36 SHAP依赖图

上图中，当 clr\_SiO₂ > 0 时，SHAP值迅速上升，即SiO₂升高对模型“富金”判断有正面贡献。此趋势在 clr\_P₂O₅ 较高（红色）时尤为明显，说明高SiO₂ + 高P₂O₅ 的组合更容易指示富金矿体。

图表, 散点图

AI 生成的内容可能不正确。

图2.37 SHAP依赖图

上图中， TFe₂O₃ 含量增加时（右移），SHAP值下降，表明高铁倾向于富铜矿而非富金。颜色为 TiO₂，可以看到TiO₂高时更集中于低SHAP值，即Fe高+Ti高倾向于富铜。

2.9 SHAP分析-神经网络

**2.9.1 SHAP**分析流程

首先加载数据和训练好的深度学习模型，用前500个样本作为背景分布，为加快解释速度，否则运行时间极长。取前50个样本做SHAP解释，得到每个样本每个特征对预测概率的贡献。

**2.9.2 SHAP**分析结果

根据上述流程，输出蜜蜂群图（图2.38）以及SHAP依赖图（图2.39、2.40）

示意图

AI 生成的内容可能不正确。

图2.38 蜜蜂群图

此图中展现了与XGBoost模型相似的结果。首先前三大关键因子还是为：MgO, TFe2O3, TiO2。SiO2的SHAP值接近零但略有偏移，但仍然说明其对结果有微弱贡献，或许在某些极端值中能发挥作用。

接下来，我们对其中的两两组合进行分析。

图表

AI 生成的内容可能不正确。

图2.39 SHAP依赖图

上图中，整体呈现线性上升趋势，说明 SiO₂ 含量越高，样本越倾向于被模型判为“富金”。

图表

AI 生成的内容可能不正确。

图2.40 SHAP依赖图

上图中，呈现负相关关系：Fe含量越高，样本越倾向被模型判为“富铜”。

2.10 SHAP分析结果解释

通过对这两个模型的Shap结果进行分析，不难发现，两个模型中都有着非常大的Shap影响力相似度，但是仍有部分值不相似（例如P2O5的相对影响力在两个模型中，差距巨大）。这是由于两个模型的算法并不一致，各自找到了符合相同的训练集验证集的最优解。如果要消除这种不确定性，那么尽可能丰富多样的数据集，能有助于使得两个模型的Shap分析解释趋于一致。

我们保留他们不一致的部分，主要探讨一致的部分。

**2.10.1** XGBoost和神经网络的相似结果

我们主要讨论前四大一致的组分， MgO,TFe2O3,TiO2和CaO。这几大组分影响力巨大，说明其可能直接对富铜和富金的结果产生了直接影响。

其中SiO2和Cao的SHAP依赖图在两个模型中都保持着正相关，而TFe2O3和Na2O的SHAP依赖图也保持着负相关，说明两个模型都学习到了这一联系，有可能这两个组分和前文提到的主要影响组分存在着直接关系，间接影响了富铜和富金；抑或是间接影响了铜和金的生成。

**2.10.2** 地质意义解释

首先，研究表明，二氧化钛（TiO₂）在岩浆和热液过程中表现为相对不易迁移的元素。在中基性岩浆早期结晶过程中。Ti作为不相容元素会在熔体中富集，使得较原始的玄武质岩浆往往TiO₂含量较高。当岩浆演化至中酸性阶段，磁铁矿或钛铁矿等含钛矿物开始晶出，TiO₂含量随之急剧下降，因此，原始岩浆（如基性玄武质）或未经历充分分异的岩浆源区常残留较高TiO₂。对于二氧化硅，也是类似过程的关键性指标。

换句话说，前四大指标中的两项，都和岩浆冷却的过程高度相关。那么不难猜想，冷却的过程会导致富金或者富铜。

根据猜想，我们查阅文献可以得到，高度氧化、富钾的演化岩浆有利于金随流体高效沉淀（富金型），而巨量中性岩浆长时间演化则偏向铜的规模富集但金相对匮乏（富铜型）。结论和数据能相互佐证。

而另外两项中，我们猜想是影响到了熔浆中的Cu/Au比值，较少含量的Cu熔浆，即使在更易于形成Au的条件下，由于元素匮乏，也难以形成。由于时间限制，我们未能查阅资料验证我们的猜想。

2.11 模型对比分析

对比我们训练出的四个模型，神经网络以微弱的优势取得在整个数据的正确性率，因此神经网络的优势不言而喻。首先，其能自动提取特征，能自动从原始数据中学习出复杂的、高阶的抽象特征，而其他三个模型依赖于前期的特征工程，即使取了自然对数，也未必所有自然界的特征分布都符合自然对数，难以重新线性映射。

说到线性，神经网络能通过多层网络结构和非线性激活函数（如ReLU、Sigmoid）拟合极其复杂的非线性函数，有效处理数据中的复杂非线性关系。SVM通常依靠核函数处理非线性问题，但受限于核函数选择，且对高维数据的性能会下降；XGBoost和随机森林的变种基于树模型，虽然能捕获非线性关系，但树模型的组合方式对高度复杂非线性模式的拟合能力相对受限。

但是神经网络也有不足，由于数据维度少，数据量也受限，我们可能容易过拟合或者欠拟合。同时维度较低的小型网络，更容易陷入局部最优解内。此外，神经网络的可解释性较差，无法从权重中提取出每个维度之间易于理解的关系。

因此，对于科学目标，基于树模型和SVM模型都具有良好的可解释性，也有较高的准确度，非常适合用于研究复杂的科学问题，研究隐藏在数据背后的科学规律。但是如果更偏向工程问题，即富铜富金的准确性，那么神经网络无疑是更合适的。

2.12 可执行程序实现

**2.12.1** 程序逻辑介绍

在完成上述功能后，主程序逻辑见表2.5。

|  |  |
| --- | --- |
| 主程序逻辑 | 详细介绍 |
| 1. 模型加载 | 加载四种模型：随机森林、深度学习、SVM、XGBoost。  支持 PyInstaller 打包后路径和源码运行的相对路径自动适配。 |
| 2. 数据上传 | 限制文件格式为CSV和Excel。 |
| 3. 数据预处理 | 对数据依次执行：单元类型验证、单位验证、缺失值处理、CLR变换、离群值检测与标记、单位统一、对数变换、离群值丢弃。 |
| 4. 模型选择与预测 | 设置用户界面选择模型，根据模型类型分别处理，预测完成后，将预测类别写入新列。 |
| 5. 性能评估可视化 | 将各类图都以高分辨率PNG横向三栏排布展示。 |
| 6. 结果展示与导出 | 结果表格以固定高度、内部滚动，支持网页端一键CSV下载，并自动保存CSV到本地result\_csv文件夹。 |

表2.5 主程序逻辑介绍

***亮点自述：***

1. 打包后路径和源码运行的相对路径自动适配。
2. 用 st.cache\_resource 缓存，减少重复I/O。
3. 对不同模型输出进行定制化处理。
4. 表格特质化处理展示，避免页面卡顿。
5. 下载和保存专属化提示。

**2.12.2** 程序打包逻辑介绍

主要借助PyInstaller实现打包，具体逻辑如下：

**1. 依赖分析**

首先根据spec文件中的Analysis对象配置，从指定的主入口脚本cli\_predict.py出发，递归分析所有直接间接依赖的Python模块、包和资源文件。

然后在pathex参数中动态指定项目根目录，确保分析路径与运行环境一致。

通过datas参数，将本地自定义的数据处理、可视化、模型文件夹整体纳入打包范围。同时，为解决常见第三方库的动态链接库、初始化脚本及元数据在部署环境中缺失问题，显式指定其核心文件和元信息，确保运行时依赖完整。

此外，显式列举所有可能在运行期被动态导入的库，防止静态分析遗漏依赖，提升打包后稳定性和可移植性。

**2. 资源整理与代码归档**

分析完成后，将所有纯Python源码与字节码、资源文件打包成.pyz格式归档，实现代码压缩和快速加载，保证Python逻辑、辅助脚本及所有项目内部依赖能被统一调度管理。

**3. 生成可执行文件**

结合阶段一整理的依赖、资源与阶段二的代码归档，根据EXE对象配置生成最终可执行文件。

***亮点自述：***

1.为便于在不同环境下便捷分发和部署。采用onefile=True，所有内容均被打入一个独立的EXE文件，便于在不同环境下便捷分发和部署。

2.为保证程序体积和启动速度，开启了upx压缩。并使用sole=True 让打包程序默认以命令行模式启动，便于批量预测、测试和日志调试。

3.使用bootloader\_ignore\_signals=True允许用户用Ctrl+C等信号安全中断程序运行，提升用户体验。

4.在打包完成后，所有用户只需分发单一EXE，无需手动配置Python环境和依赖安装，支持科学软件在多平台、多用户环境下的交付使用。

**2.12.3** cli\_predict.exe程序运行展示

1. 启动运行控制终端展示（图2.41）

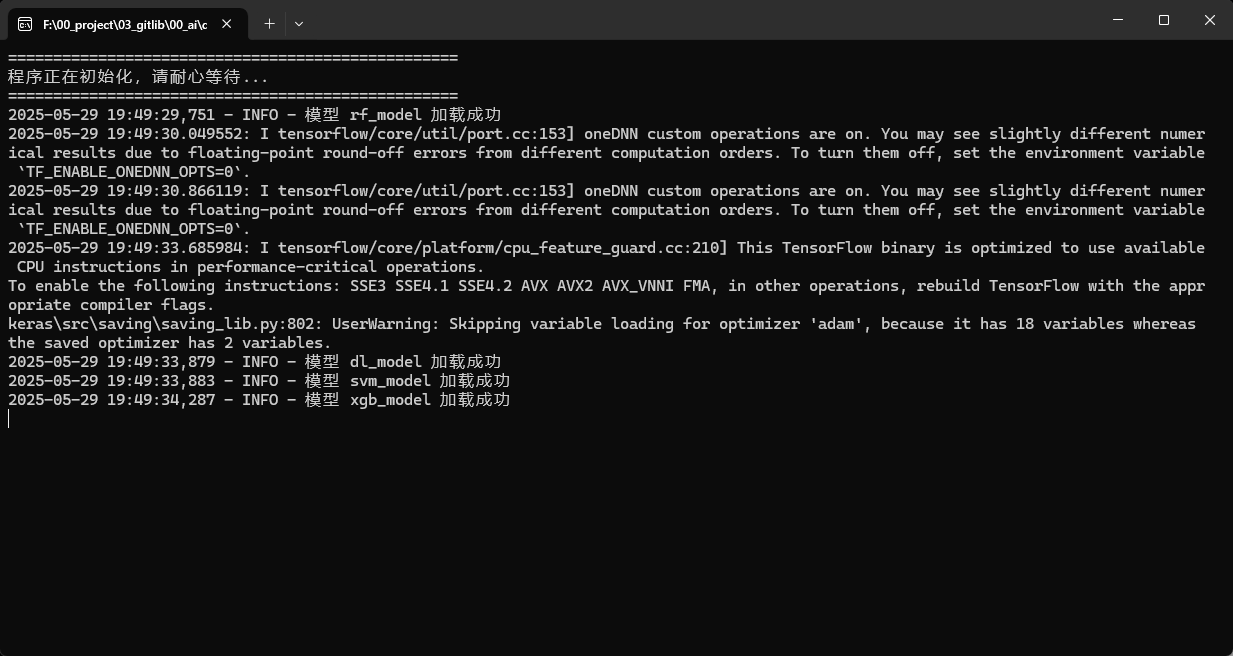


图2.41 启动运行控制终端展示图

2. 程序界面基本样式展示（图2.42）

图形用户界面, 应用程序

AI 生成的内容可能不正确。

图2.42 界面基本样式展示图

3. 数据预处理结果样式展示（图2.43）

图形用户界面, 应用程序, 表格

AI 生成的内容可能不正确。

图2.43 数据预处理结果样式展示图

4. 随机森林预测结果展示（图2.44）

图形用户界面

AI 生成的内容可能不正确。

图2.44 随机森林预测结果展示图

1. 深度学习预测结果展示（图2.45）

图形用户界面

AI 生成的内容可能不正确。

图2.45 深度学习预测结果展示图

6. 支持向量机预测结果展示（图2.46）

图形用户界面

AI 生成的内容可能不正确。

图2.46 支持向量机预测结果展示图

7. XGBoost预测结果展示（图2.47）

图形用户界面

AI 生成的内容可能不正确。

图2.47 XGBoost预测结果展示图

**2.12.4** 网页版本运行展示图

1. 网页版本效果展示（图2.48）

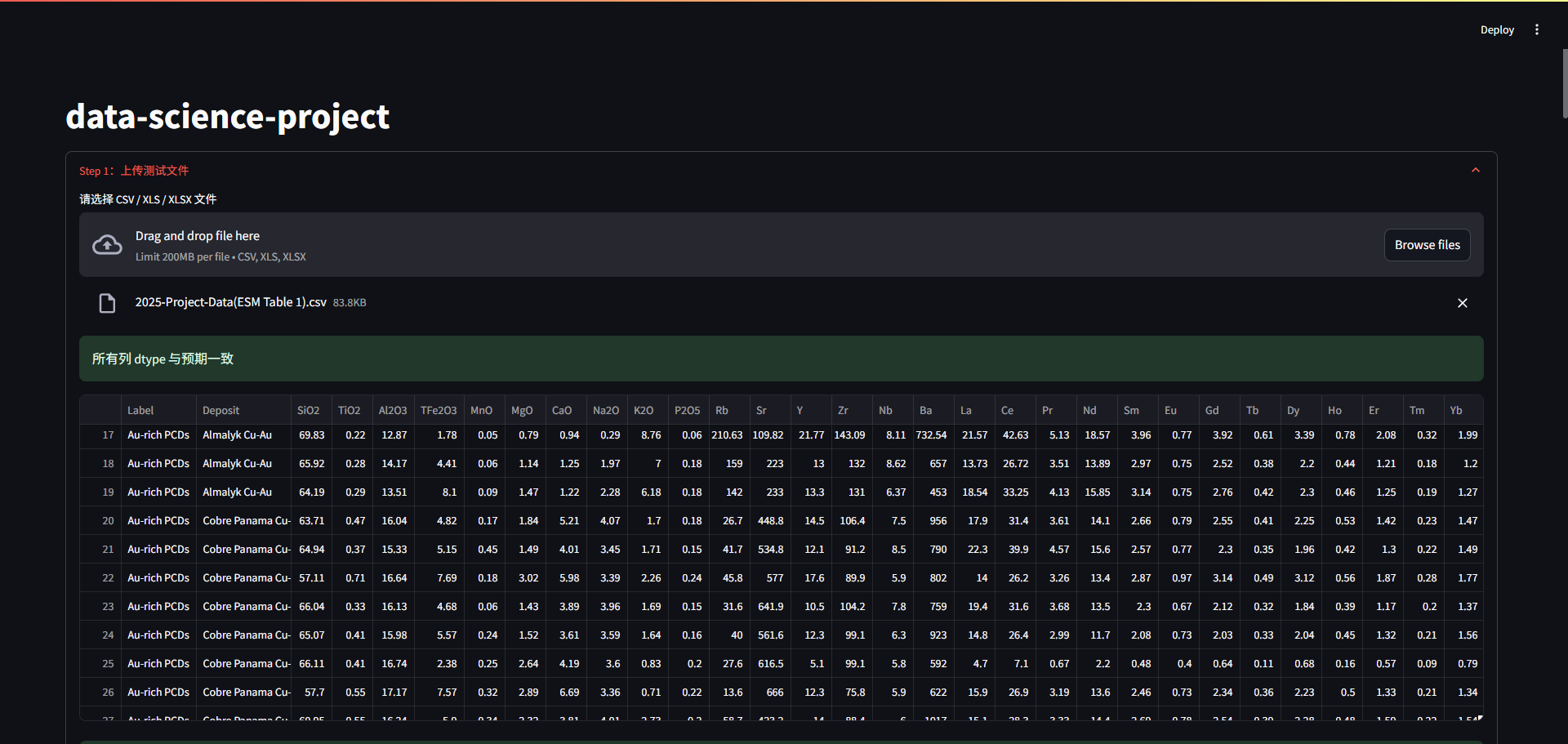


图2.48 网页版本效果展示图

2. 网页的预测效果图（图2.49）

图形用户界面, 应用程序

AI 生成的内容可能不正确。

图2.49 数据预处理结果样式展示图

3 小组任务分工

**3.1** 小组成员介绍

|  |  |
| --- | --- |
| 成员及GitHubID | 学号 |
| 吴百恒（Henry-Baiheng） | 2240007514 |
| 黄熙（hx101700） | 2240003603 |
| 李杰相（sethome2） | 2240023371 |
| 彭嘉男（PJNICer） | 2240022663 |

表3.1 小组成员介绍

**3.2** 具体任务分工

**3.2.1** 团队协作方式概述

本项目团队采用微信线上沟通与GitHub分支协作（图3.1、3.2）相结合的方式，配合软件工程中的敏捷开发流程开展项目管理与开发。

日常需求讨论、任务分配、进度同步主要通过微信群进行高效即时交流，保证信息畅通与快速响应。

代码开发和版本管理则依托GitHub仓库，团队成员分别在独立分支上开发、提交和测试功能，定期通过Pull Request方式进行代码合并和评审，确保主分支稳定性。

整个开发周期采用敏捷迭代方式，分阶段完成需求梳理、模块设计、开发实现与集成测试，提升团队协作效率和项目交付质量。

亮着屏幕的电脑截图

AI 生成的内容可能不正确。

图3.1 小组GitHub协作展示图

图3.2 小组GitHub部分提交展示图

**3.2.2** 数据预处理阶段分工

|  |  |
| --- | --- |
| 成员名 | 任务分工 |
| 吴百恒 | 1. 任务划分，数据预处理阶段的协调与进度把控 2. 数据读取限制、文件流操作、文件初查等工作 |
| 黄熙 | 1. 单位映射、单位标准化 2. 异常捕获、处理 3. 撰写相应的文档说明 |
| 李杰相 | 1. 缺失值处理模块，实现缺失值的自动检测与填充策略 2. 负责部分测试与单元测试用例撰写 |
| 彭嘉男 | 1. 承担CLR变换与异常值标记实现，编写对应的数据转换与异常检测算法 2. 相关流程的代码复查 |

表3.2 数据预处理阶段分工表

**3.2.3** 探索性分析阶段分工

|  |  |
| --- | --- |
| 成员名 | 任务分工 |
| 吴百恒 | 1. 任务划分，探索性分析阶段的协调与进度把控 2. 散点矩阵图的编写和实现，包括相关数据可视化代码开发与文档说明 |
| 黄熙 | 1. 相关性热力图实现，完成相关性分析与可视化脚本的编写 2. 参与部分测试和方法优化 |
| 李杰相 | 1. 承担PCA双标图编写，实现主成分分析的绘制 2. 负责相关代码复查与单元测试用例撰写 |
| 彭嘉男 | 1. 负责地球化学比值图开发，实现比值图的绘制 2. 承担所有结果图讨论分析 |

表3.3 探索性分析阶段分工表

**3.2.4** 模型训练准备阶段分工

|  |  |
| --- | --- |
| 成员名 | 任务分工 |
| 吴百恒 | 1. 任务划分，模型训练准备阶段的协调与进度把控 2. 负责部分代码复查，编写相关数据处理脚本与文档说明，完成各模块验证和相关测试 |
| 黄熙 | 负责单位均同（单位统一），实现特征单位的规范化与一致性校验 |
| 李杰相 | 负责对数变换实现，包括数据特征的对数转换、验证和相关测试 |
| 彭嘉男 | 承担异常值影响分析模块，编写并优化异常值对模型训练影响的分析，辅助代码复查与文档整理 |

表3.4 模型训练准备阶段分工表

**3.2.5** 模型训练-随机森林阶段分工

|  |  |
| --- | --- |
| 成员名 | 任务分工 |
| 吴百恒 | 1. 负责整体模型训练与验证流程的统筹与进度把控 2. 最优树数量确定的实验设计与模型训练 |
| 黄熙 | 1. 最优树数量确定的误差分析及相关结果可视化 2. 模型输出和结果文档进行复查与整理 |
| 李杰相 | 1. 随机森林最佳模型预测结果展示 2. 混淆矩阵、AUC&ROC曲线绘制 |
| 彭嘉男 | 1. 精确率-召回率曲线、特征重要性等结果的生成与可视化 2. 模型测试与指标计算，实现性能评估脚本 |

表3.5 模型训练-随机森林阶段分工表

**3.2.6** 模型训练-XGBoost阶段分工

|  |  |
| --- | --- |
| 成员名 | 任务分工 |
| 吴百恒 | 1. 负责整体XGBoost模型训练流程的协调与进度把控 2. 最佳确定环节的实验设计与XGBoost不同树数量下模型训练 |
| 黄熙 | 1. 最佳确定环节交叉验证和误差曲线的生成与可视化 2. 模型输出和结果文档进行复查与整理 |
| 李杰相 | 1. 最佳模型保存与预测，完成最佳模型的保存、加载 2. 模型在测试集上的预测与分类评估结果输出以及混淆矩阵、AUC&ROC曲线绘制 |
| 彭嘉男 | 1. 精确率-召回率曲线、特征重要性等结果的生成与可视化 2. 模型性能评估分析 |

表3.6 模型训练-XGBoost阶段分工表

**3.2.7** 模型训练-SVM阶段分工

|  |  |
| --- | --- |
| 成员名 | 任务分工 |
| 吴百恒 | 1. 负责整体SVM模型训练流程的协调与进度把控 2. 数据归一化、交叉验证自动调参代码实现 |
| 黄熙 | 1. 参数网格搜索、模型训练代码实现 2. 代码开发、相关结果文档整理 |
| 李杰相 | 1. 保存SVM最佳模型与预测，实现最佳模型的保存与加载 2. 负责测试集预测与分类评估结果输出 3. SVM模型性能评估 |
| 彭嘉男 | 1. 混淆矩阵、AUC&ROC曲线、精确率-召回率曲线等指标的可视化生成 2. 文档复查、代码格式复查 |

表3.7 模型训练-SVM阶段分工表

**3.2.8** 模型训练-Keras神经网络阶段分工

|  |  |
| --- | --- |
| 成员名 | 任务分工 |
| 吴百恒 | 1. 整体深度神经网络训练流程的协调与进度把控 2. KerasTuner超参数搜索、网络结构设计 |
| 黄熙 | 1. 最佳模型保存与预测 2. 模型在测试集上的预测与分类评估结果输出 3. 混淆矩阵绘制 |
| 李杰相 | 1. 数据准备与模型测试 2. 防止欠拟合、过拟合实现 3. 代码开发与文档说明整理 |
| 彭嘉男 | 1. AUC&ROC曲线、精确率-召回率曲线等指标的生成 2. 相关结果文档整理与复查 3. 代码Debug |

表3.8 模型训练-Keras神经网络阶段分工表

**3.2.9** SHAP分析实现阶段分工

|  |  |
| --- | --- |
| 成员名 | 任务分工 |
| 吴百恒 | 1. 负责整体SHAP分析流程的协调与进度把控 2. SHAP\_XGBoost依赖图的生成 3. 代码测试与复查 |
| 黄熙 | 1. 加载XGBoost模型、构建SHAP\_XGBoost解释器 2. XGBoost模型特征重要性贡献度计算、蜜蜂群图绘制实现 |
| 李杰相 | 1. 神经网络依赖图的绘制 2. 数据准备、结果展示 3. 关键图表排版与文档说明 |
| 彭嘉男 | 1. 加载神经网络模型、构建神经网络SHAP解释器 2. 神经网络模型特征重要性贡献度计算、蜜蜂群图绘制实现 |

表3.9 SHAP分析实现阶段分工表

**3.2.10** SHAP结果分析与模型对比阶段分工

|  |  |
| --- | --- |
| 成员名 | 任务分工 |
| 吴百恒 | 1. 负责整体解释与对比分析阶段的协调与进度把控 2. Boost、神经网络及地质意义解释等具体内容的撰写 3. 最终结果目标设立与之前所有代码检查 |
| 黄熙 | XGBoost相关的SHAP解释与模型特征贡献度分析，参与模型对比部分的数据整理和结果展示 |
| 李杰相 | 神经网络相关的SHAP解释结果分析，协助撰写神经网络模型的特征贡献讨论，并参与模型对比 |
| 彭嘉男 | 地质意义解释部分，将模型特征重要性与地质知识结合，深入分析关键地球化学指标的实际意义，同时参与模型对比分析文本与结果总结 |

表3.10 SHAP结果分析与模型对阶段分工表

**3.2.11**可执行程序阶段分工

|  |  |
| --- | --- |
| 成员名 | 任务分工 |
| 吴百恒 | 1. 负责可执行程序开发整体统筹与进度把控，主导需求汇总与实现流程优化。 2. 程序逻辑梳理、运行展示及打包流程代码实现，程序最终审核。 3. 本地程序预测展示实现、文档说明 |
| 黄熙 | 1. 本地程序各功能模块流程梳理 2. 命令行窗口运行设计代码实现 3. 本地端程序界面及数据处理 4. 打包结果Windows系统测试运行 |
| 李杰相 | 1. 打包逻辑梳理撰写 2. 依赖、数据、模型文件的收集整理与打包测试 3. 打包结果Linux系统测试运行 |
| 彭嘉男 | 1. 网页端主要界面、页面端数据处理和预测结果呈现 2. 参与各类运行截图和效果说明整理 3. 协助完成整体程序实现过程的文档总结与测试反馈 |

表3.11 可执行程序阶段分工表

参 考 文 献

[1] Roncoroni, G., Forte, E., Santin, I., Černok, A., Rajšić, A., Frigeri, A. L. E. S. S. A. N. D. R. O., & Pipan, M. (2024). High frequency Lunar Penetrating Radar quality control, editing and processing of Chang’E-4 lunar mission. Scientific Data, 11(1), 118.

[2] Cao, H., Xu, Y., Xu, L., Zhang, L., Bugiolacchi, R., & Zhang, F. (2023). From Schrödinger to Von Kármán: An Intriguing New Geological Structure Revealed by the Chang'e‐4 Lunar Penetrating Radar. Geophysical Research Letters, 50(2), e2022GL101413.

[3] Hearst M A, Dumais S T, Osuna E, et al. Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4): 18-28.

[4] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20: 273-297.

[5] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. nature, 1986, 323(6088): 533-536.

[6] Breiman L. Random forests[J]. Machine learning, 2001, 45: 5-32.

[7] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[8] Khashgerel, B.-E., Kavalieris, I., & Hayashi, K. (2008). Mineralogy, textures, and whole-rock geochemistry of advanced argillic alteration: Hugo Dummett porphyry Cu–Au deposit, Oyu Tolgoi mineral district, Mongolia. Mineralium Deposita, 43(8), 913–932. DOI: 10.1007/s00126-008-0199-8

[9] Jenner, F. E., O’Neill, H. S. C., Arculus, R. J., Mavrogenes, J. A., & Claoué-Long, J. (2010). The magnetite crisis in the evolution of arc-related magmas and the initial concentration of Au, Ag and Cu. Journal of Petrology, 51(12), 2445–2469. DOI: 10.1093/petrology/egq063

[10] Chiaradia, M. (2020). *Gold endowments of porphyry deposits controlled by precipitation efficiency*. Nature Communications, 11(1), 248. DOI: 10.1038/s41467-019-14113-1

[11] Liu, Y., Zhao, X., Xue, C., Nurtaev, B., & Chen, J. (2023). *Contrasting apatite geochemistry between ore-bearing and ore-barren intrusions of the giant Kalmakyr gold-rich porphyry Cu deposit, Tien Shan, Uzbekistan*. Frontiers in Earth Science, 11, 1162994. DOI: 10.3389/feart.2023.1162994

[12] Sillitoe, R. H. (2010). Porphyry copper systems. Economic Geology, 105(1), 3–41. DOI: 10.2113/gsecongeo.105.1.3

总 结

通过本次期末报告，完成了样品区分、相关特征解释、以及可离线运行预测工具。深刻理解了敏捷开发流程在小项目中的重要性，真正的将平时所学各类模型以及各类图像曲线绘制运用到具体项目。

在项目实现过程中，组内各个成员也发现了很多平时发现不了的技术短板，通过积极沟通以及相互学习，将每个成员技术短板进行了补足，深刻理解了团队协作共同进步的重要性。

感恩老师一学期的辛苦授课与课上的答疑解惑，老师上课让我们分享的实用软件也对我们有非常大的帮助，特别是Snipaste让每个组员钉图完成任务代码实现以及便利的窗口捕捉截图分享。希望老师后续课程的小组任务可以更加复杂、困难一些，真正展示出我们准硕士二年级学生与本科毕业生的区别。

最后祝老师心想事成、科研顺利、论文秒中无修，也提前祝老师假期愉快！