

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](#_Toc197354514)

[1 前言 3](#_Toc197354515)

[**1.1** 内容简介 3](#_Toc197354516)

[**1.2** 完成环境 3](#_Toc197354517)

[2 文章复现及优化 4](#_Toc197354518)

[**2.1** 复现文章选择 4](#_Toc197354519)

[**2.2** 文章复现思路 4](#_Toc197354520)

[**2.3** 文章具体复现 5](#_Toc197354521)

[**2.3.1** 雷达数据预处理-去重复道以及数据拼接 5](#_Toc197354522)

[**2.3.2** 雷达数据后处理 7](#_Toc197354523)

[**2.4** 个人优化 8](#_Toc197354524)

[**2.4.1** 读取数据优化 8](#_Toc197354525)

[**2.4.2** 引入早期人工智能算法优化数据预处理 8](#_Toc197354526)

[**2.4.3** 使用更精准的零时矫正 9](#_Toc197354527)

[**2.4.4** 设置新型带通滤波 10](#_Toc197354528)

[**2.4.5** 额外使用新的方法进行背景噪声去除 10](#_Toc197354529)

[**2.5** 未来工作 11](#_Toc197354530)

[参 考 文 献 13](#_Toc197354531)

[总 结 14](#_Toc197354532)

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

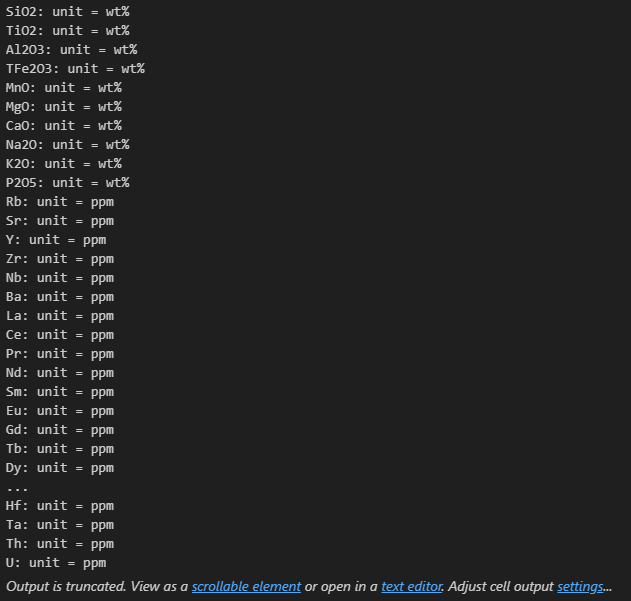


图2.2 单位映射试运行结果图

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

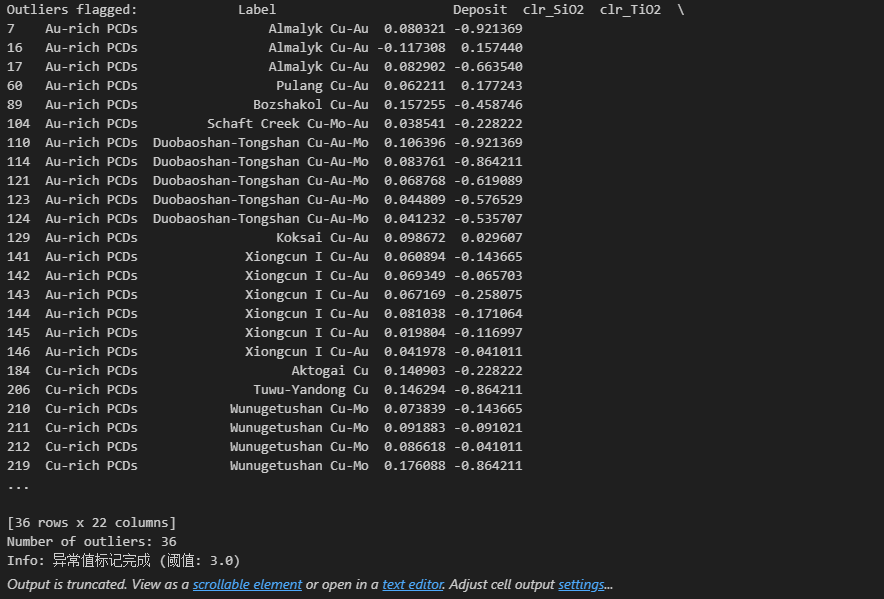


图2.6 异常值标记试运行结果图

**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

日历

AI 生成的内容可能不正确。

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

图表, 散点图

AI 生成的内容可能不正确。

图2.36 SHAP依赖图

图表, 散点图

AI 生成的内容可能不正确。

图2.37 SHAP依赖图

**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，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

图表

AI 生成的内容可能不正确。

图2.39 SHAP依赖图

图表

AI 生成的内容可能不正确。

图2.40 SHAP依赖图

**2.10** SHAP分析结果解释

**2.10.1** Boost

**2.10.2** 神经网络

**2.10.3** 地质意义解释

**2.11** 模型对比分析

**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）

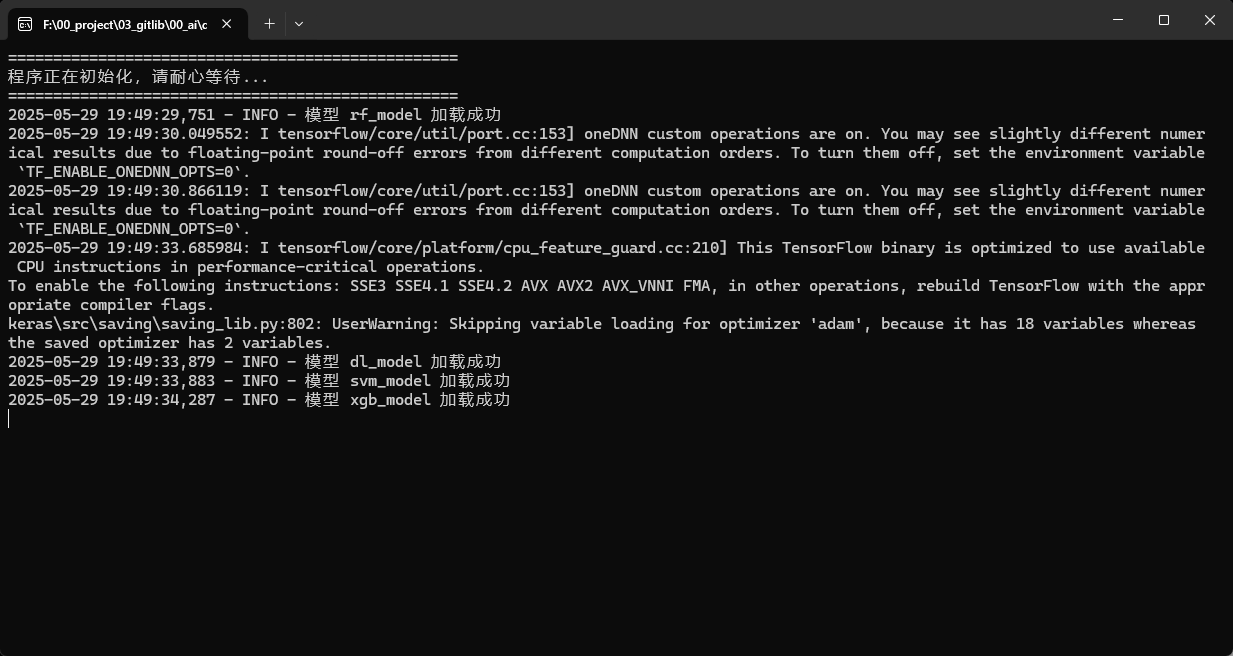


图2.41 启动运行控制终端展示图

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）

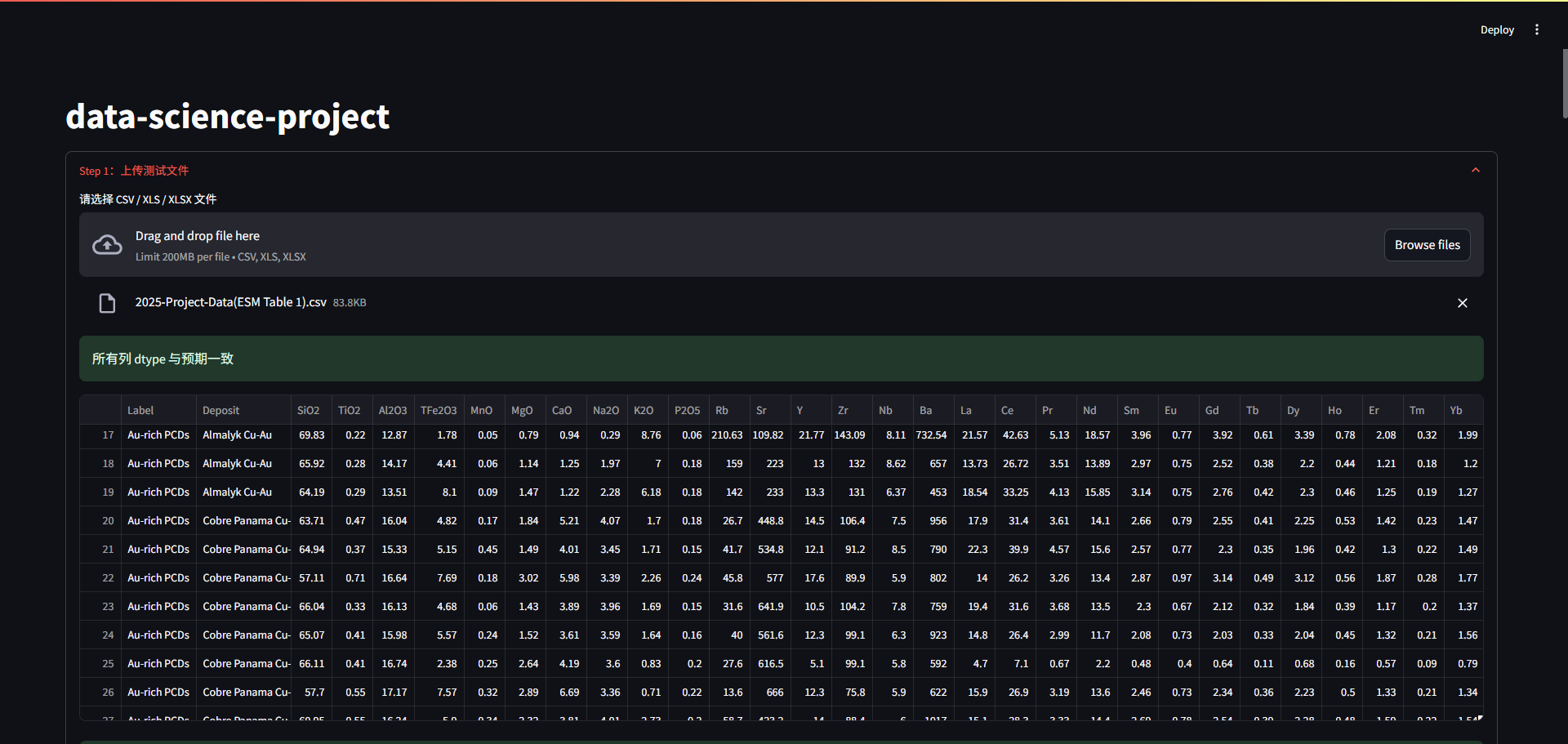


图2.48 网页版本效果展示图

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

# 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.

# 总 结

通过本次期末报告，完成了雷达数据处理领域的入门及优化，在准备课上Pre过程也将许多雷达专业术语以及领域内不足的知识查明弄清，也养成了查文献读文献等相关习惯的培养。

感恩老师一学期的辛苦授课与课上的答疑解惑，祝老师心想事成、诸事顺利、论文顺利，也提前祝老师假期愉快！