**Readme**

|  |
| --- |
| Plain Text  ## 主要功能\* \*\*数据管理\*\*: 支持加载原始 `.npy` 数据和预处理后的 `.xlsx` 数据。 \* \*\*有监督学习\*\*:  \* 支持多种分类模型（如 RandomForest, Logistic Regression, SVM, KNN, XGBoost, LightGBM）。  \* 内置 Grid Search 进行超参数优化。  \* 支持交叉验证评估模型性能。  \* 生成混淆矩阵、特征重要性、SHAP 值（摘要图、依赖图）等可视化图表。 \* \*\*无监督学习\*\*:  \* 支持基于 Autoencoder + UMAP + 聚类（KMeans, DBSCAN）的流程。  \* 支持肘部法则自动寻找 KMeans 最佳 K 值。  \* 生成聚类散点图、肘部法则图等可视化图表。 \* \*\*自动化报告\*\*:  \* 将所有模型性能指标汇总并保存为 CSV 文件。  \* 自动生成包含所有可视化图表的 PowerPoint 报告。  \* 终端实时打印美化后的模型性能摘要。 |

1. 安装所有必要的依赖：

为了更好地管理项目依赖并避免不同项目之间的库冲突，强烈建议为本项目创建一个独立的 Python 虚拟环境。

1.1、打开你的命令行或终端，导航到你的项目根目录 LabDataProcess/

1.2、然后执行以下命令来创建虚拟环境及安装依赖

|  |
| --- |
| python -m venv .venv  .venv\Scripts\activate pip install -r requirements.txt |

**注意**: requirements.txt 文件需要包含所有你在项目中使用的库，例如 pandas, numpy, scikit-learn, matplotlib, seaborn, shap, torch, lightgbm, xgboost, python-pptx, umap-learn, joblib 等。请确保这个文件是最新的。

1. 数据准备

本项目期望的数据结构如下：

|  |
| --- |
| Plain Text LabDataProcess/ ├── Data/ │ ├── raw/ │ ├── processed/ │ │ └── data.xlsx # 预处理后的数据 (用于有监督学习) │ └── external/ │ ├── split\_data.npy # 原始特征数据 (用于无监督学习) │ └── patient\_ids.npy # 对应的病人ID (用于无监督学习) ├── reports/ # 报告输出目录 (由程序自动创建子目录) │ └── 产妇抑郁/ # 你的指定输出子目录(可自行增加) ├── models/ # 保存训练好的模型 (由程序自动创建) ├── src/ │ ├── config.py # 核心配置文件 │ ├── data\_loader.py │ ├── supervised\_models.py │ ├── unsupervised\_models.py │ ├── visualizer.py │ └── \_\_init\_\_.py ├── main.py # 主运行脚本 └── requirements.txt # 项目依赖 |

**你需要将你的数据放置在对应位置：**

* **预处理后的数据 (data.xlsx)**: 放置在 Data/processed/ 目录下。
* 这个 Excel 文件应该包含所有特征和目标变量（标签）。
* 确保目标变量列名与 src/config.py 中 DEFAULT\_TARGET\_COLUMN 的设置一致（默认为 "Label"）。
* **原始特征数据 (split\_data.npy)**: 放置在 Data/external/ 目录下。
* 这是用于无监督学习（如 Autoencoder）的原始或初步处理过的特征数据（NumPy 数组）。
* **病人ID (patient\_ids.npy)**: 放置在 Data/external/ 目录下。
* 这是与 split\_data.npy 中样本对应的病人ID。

1. 配置项目 (src/config.py)

src/config.py 是项目的核心配置文件。你需要根据你的需求修改以下关键设置：

* **BASE\_DIR**: **重要！** 确保此路径正确指向你的项目根目录 LabDataProcess。
* Python

|  |
| --- |
| Plain Text BASE\_DIR = Path(\_\_file\_\_).resolve().parent.parent # 应该指向 LabDataProcess 目录 |

* **DATA\_DIR**: 项目的数据根目录。默认是 BASE\_DIR / "Data"，通常无需修改。
* **REPORTS\_OUTPUT\_DIR**: **重要！** 这是所有报告和图表输出的最终目录。目前设置为 REPORTS\_DIR / "产妇抑郁"，这意味着所有报告将保存到 LabDataProcess/reports/产妇抑郁/ 目录下。
* Python

|  |
| --- |
| Plain Text REPORTS\_OUTPUT\_DIR = REPORTS\_DIR / "产妇抑郁" # 根据你的偏好修改，例如 REPORTS\_DIR / "我的实验结果" |

* **PROCESSED\_DATA\_INPUT\_PATH**: 预处理后的 data.xlsx 文件路径。默认是 DATA\_PROCESSED\_DIR / "data.xlsx"。
* **DEFAULT\_TARGET\_COLUMN**: 有监督学习中目标变量（标签）的列名。如果你在 data.xlsx 中的标签列不是 "Label"，请修改此项。
* **SUPERVISED\_MODELS 和 UNSUPERVISED\_MODELS**:
* 你可以根据需要调整每个模型的参数网格 (param\_grid)、类型 (type) 和其他配置。
* 对于无监督学习的 Autoencoder，input\_dim 需要与 split\_data.npy 的特征维度匹配。

1. 运行分析 (main.py)

main.py 是程序的入口点。你可以在 if name == "\_\_main\_\_": 块中控制要运行哪些模型和实验。

Python

|  |
| --- |
| Plain Text if \_\_name\_\_ == "\_\_main\_\_":  main\_analysis\_pipeline(  experiment\_name="RandomForest\_SDS\_Classification", # 本次实验的名称，会创建同名子文件夹  target\_column\_supervised=config.DEFAULT\_TARGET\_COLUMN, # 指定有监督学习的目标列  supervised\_models\_to\_run=["RandomForestClassifier"], # 运行指定的有监督模型，或 None 运行所有  unsupervised\_models\_to\_run=None # 运行指定的无监督模型，或 None 运行所有  ) |

* **experiment\_name**: 为每次运行指定一个唯一的名称。这个名称将用于在 REPORTS\_OUTPUT\_DIR 下创建相应的子文件夹，并作为 PowerPoint 报告名称的一部分。
* **target\_column\_supervised**: 确保与 config.py 中的 DEFAULT\_TARGET\_COLUMN 一致，或者在函数调用时覆盖它。
* **supervised\_models\_to\_run**:
* 设置为 None 将运行 config.py 中定义的所有有监督模型。
* 设置为一个字符串列表（例如 ["RandomForestClassifier", "LogisticRegression"]）将只运行列表中指定的模型。
* **unsupervised\_models\_to\_run**:
* 设置为 None 将运行 config.py 中定义的所有无监督模型。
* 设置为一个字符串列表（例如 ["Autoencoder\_UMAP\_KMeans"]）将只运行列表中指定的模型。

运行main.py脚本

1. 查看结果

程序运行完成后：

* **控制台输出**: 会实时打印每个模型的性能摘要。
* **报告文件**: 所有生成的图表（PNG 格式）和性能汇总 CSV (model\_performance\_summary.csv) 将保存到 D:\MyProjects\LabDataProcess\reports\产妇抑郁\<experiment\_name> 目录下。
* **PowerPoint 报告**: 一个汇总所有图表的 PowerPoint 文件 (<experiment\_name>\_analysis\_report.pptx) 也将保存在 D:\MyProjects\LabDataProcess\reports\产妇抑郁\ 目录下。