实验五

中国科大2025年春季学期"数据分析及实践"课程 - 实验五说明文档。

任务概述

本实验基于实验三和四所使用的PISA数据集子集`subdata.csv`，对部分属性进行分类预测，并尝试使用大语言模型进行辅助分析，请按要求完成实验任务。

任务列表

1.（80%）请选择若干合适的机器学习算法进行训练，对特征`TEACHBEHA`进行预测，并汇总分析结果。本实验的开放性较强，无固定任务列表，实验报告可参考以下主要步骤：

- S1：数据信息和预处理：请概述所使用数据集的基本信息，并处理缺失值和异常值。
- S2：数据集划分：按固定比例划分训练集/验证集/测试集，或者使用k折交叉验证法划分训练集/测试集（可调整随机种子以获得不同的划分方案）。
- S3：机器学习算法模型：选择至少两种机器学习算法模型，作为主实验的比较方法，汇报它们的模型/算法信息，并在报告中引用相应的参考文献。
- S4：特征选择与处理：根据数据集本身和选用机器学习算法模型的特点，参考实验三的数据分析结论，选取合适的特征并进行一定适应性处理后作为模型输入。
- S5：主实验：分别在选择的算法模型上进行训练（可直接使用机器学习库也可自行实现），确定评测指标（如MSE等），汇报并分析它们在测试集上的结果。
- S6：参数实验：对选用的模型，汇报并分析调整某些关键参数后所得到的实验结果。
- S7：结论分析：对使用方法的合理性和所得结论依据进行解释说明。

2.（20%）近年来，大语言模型已经成为人工智能领域重要的生产力，其在数据分析领域同样展现出强大的能力，请利用大语言模型辅助完成以下任务。

- Q1.（10%）登录DeepSeek（ https://chat.deepseek.com ）或腾讯元宝（ https://yuanbao.tencent.com/chat ）对话窗口，设计合适的提示词，要求DeepSeek大语言模型基于数据集特征定义的元信息推测可能存在关联的特征，并比较不同设置下的输出结论：

  （a）关闭"深度思考（R1）"，并要求模型直接输出结论；

  （b）关闭"深度思考（R1）"，并要求模型逐步思考输出答案；

  （c）打开"深度思考（R1）"，无额外输出要求。

  （d）简述（a）（b）（c）所得输出的特点，并比较相应推测结论与实验三的数据分析结论、T1-S4的特征选择分析依据是否有相似之处。

- Q2.（10%）参考DeepSeek接口文档（ https://api-docs.deepseek.com/zh-cn/ ），编写代码调用DeepSeek-V3的API接口，要求大语言模型对指定问题生成Python代码，代码生成的提示设计举例如下：

  ```
  [Question] 已知给定pandas.DataFrame实例df，请编写一段Python代码输出列A和列B的平均值。
  [Answer] ```print(df['A'].mean(), df['B'].mean())```
  ```

  现需要分别求特征`STUBEHA`与`TEACHBA`、`EDUSHORT`与`STAFFSHORT`的相关系数。

（a）请仿照上述示例设计输入提示词，调用API获取大语言模型的输出结果并展示。输出代码的内容和格式符合你的预期吗？代码能否正常执行（假设`df`数据集已经加载和预处理完成）？

（b）尝试在（a）的输入提示前补充一些样例（如上方展示的样例），重新调用API获取并展示大语言模型的输出。对比二者所得输出内容和格式的不同，并阐述你的发现。

格式和提交要求

1. 请按具体任务分步编写代码，存储于`.ipynb`格式文件中用于复现，必要时可增加注释。
2. 实验报告必须涵盖任务列表中的所有内容和相应结果，并请存储于`.pdf`格式文件中。
3. 提交时，请将实验源代码和实验报告保存至一个压缩包中，命名为“学号-姓名-实验五.zip”，并于2025年5月22日之前发送至USTC_AD2025@163.com。