“数据仓库与数据挖掘”大作业

任务一、患者肺损伤疾病辅助决策

任务描述

根据数据集中提供的患者生命体征、血气、生化等检测指标，构建分类模型，预测患者肺损伤严重程度并给出推荐治疗方案，辅助医生进行治疗决策。

数据：dataset/classification.csv

作业要求

数据：

1. 采用合适的数据预处理方法，进行数据清洗，对缺失数据进行填补、规范化等，对比预处理前后的结果差异。

2. 采用合适的特征选择方法，降低模型训练的参数量。对比特征选择前后的结果差异。

模型与算法：

1. 选择至少3种分类算法进行实现，其中至少1种分类算法为自己实现，并比较自己的实现效果和所使用的库（如 scikit-learn 或 matlab 等）中提供的实现之间的差异。

2. 使用合适的方法对分类算法中的超参数进行选择，对比各个算法在该任务上的效果差异并分析原因。

评价：

1. 使用K折交叉验证进行模型评估。

2. 使用合理的评价指标对结果进行评估，并说明使用该指标的原因。

3. 对分类结果绘制混淆矩阵和ROC曲线。

任务二、地理位置数据聚类分析

任务描述

根据数据集中提供的地理位置信息，将地点进行聚类划分。每一行的数据表示为(latitude, longitude, location)。前两个数据表示为地理位置信息(经纬度)，最后一个location代表这个地方的真实编号。对于聚类的过程中采用空间点之间的距离作为对象间距离即可。完成聚类再针对真实标号location来评定聚类结果的效果。

数据：dataset/clustering.csv

作业要求

模型与算法：

1. 选择至少3种聚类模型或算法进行实现，其中至少1种模型或算法为自己实现，并比较自己的实现效果和所使用的库（如 scikit-learn 或 matlab 等）中提供的实现之间的差异。

2. 使用合适的方法对聚类算法中的超参数进行选择，如 K-means 中的 K 或者 DBSCAN 中的密度阈值和覆盖范围半径 等。对于每个实现算法都要根据参数进行至少5次试验来进行对比分析。例如对于k-means方法中的参数k，可以选取5个不同的k值进行对比实验。

评价：

1. 使用常见的聚类算法指标对结果进行评估，不少于两种
2. 使用可视化的方法对聚类结果进行展示

提交要求

1. 本次作业分组完成，每组2人。提交作业时由一人提交即可。任务一和任务二需要小组成员共同完成，不能一人做一题，在小组成员分工中需列举每个人再每个任务中的工作。

2. 提交内容：

   (1) 所有实现作业要求所需的源代码，语言不限

   (2) 文档需覆盖作业要求中的所有点，中英文不限，不超过6页，提交 PDF 格式的文件。

   (3) 作业分工说明。分别列举小组成员在本次大作业中的分工内容。

   (4) 以上内容打包成一个压缩文件上传。文件名格式:姓名1_姓名2_大作业.zip，如：张小明_李小华_大作业.zip。

3. 作业提交截止时间：2022年6月2日 23:59:59（含）。本次作业占课程总成绩的 30%。