任务背景

根据 CDC 的说法，心脏病是美国大多数种族（非裔美国人、美洲印第安人和阿拉斯加原住民以及白人）的主要死因之一。大约一半的美国人（47%）至少有三种主要的心脏病危险因素之一：高血压、高胆固醇和吸烟。其他关键指标包括糖尿病状态、肥胖（高 BMI）、缺乏足够的体力活动或饮酒过量。检测和预防心脏病在医疗保健中非常重要。

任务说明

1. 实现 Random Forest 算法学习一个分类器，预测患者是否患有心脏病。
2. 需要数据预处理，如处理数据缺失、数值差异大、非数值数据，自行分析数据、判断需要使用哪些预处理方法。
3. 需要处理正负样本不平衡的问题 。
4. 使用 out-of-bag error 进行模型选择。
5. 使用 Area Under Precision-Recall Curve (AUPRC) 作为评价指标。

数据集

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

要求

1. 自行实现算法，数据预处理过程可以使用 pandas 等工具，决策树和随机森林算法请自行实现，禁止直接调用工具包中的算法模块。
2. 使用 Python 编程语言。
3. 提交实验报告和代码文件，实验报告中需要包含算法实现方式、**任务说明**中各项的完成情况以及对结果的分析，代码要有注释说明。
4. 给出随机森林算法参数设置，如决策树个数，最小分类样本、最小不纯度、最大深度等。