# 人工智能小组作业

软件工程系
提交日期：第 15 周随堂提交
题目：糖尿病预测

## 目标

本作业旨在巩固理论知识，通过使用 python 编程环境实现相应的人工智能解决方案对现实生活中的问题进行探究并解决。本次作业能更好地帮助学习者熟悉掌握对数据集的操作、相关AI python 库的使用、对实际问题中数据特性的目标的思考、以及不同模型间的比较和分析。该作用分组完成，每队严格限制在 7~10 人(重修的同学统一为一组)，组长将分组名单于 10.18 日 20 时 59 分前发送至邮箱 jy.lin@fjnu.edu.cn。

## 问题背景

糖尿病是全球最普遍的慢性疾病之一，每年影响全球数千万人，给经济带来沉重的负担。糖尿病对人体的健康危害也不容小觑，当人失去有效调节血液中葡萄糖水平的能力，需要长期药物干预治疗，并可能导致生活质量和预期寿命下降。
在消化过程中，不同的食物被分解成糖后，糖就会被释放到血液中。这向胰腺发出释放胰岛素的信号。胰岛素有助于体内的细胞利用血液中的这些糖来获取能量。糖尿病通常的特点是身体没有产生足够的胰岛素，或者无法根据需要有效地使用胰岛素。


## 任务目标

本次作业任务是根据健康指标相关信息数据，然后通过训练数据训练模型，预测测试集所属类别。目标分为三类，0 为非糖尿病，1 为糖尿病前期，2 为糖尿病。

## 数据说明

本次作业数据包含 22 个字段，其中 target 字段为预测目标。**数据集已上传云平台，文件名为 data.csv**。具体字段说明见下表：

| 特征字段 | 字段描述 |
| --- | --- |
| Id | 样本标识id |
| HighBP | 高血压 |
| HighChol | 高酒精 |
| CholCheck | 胆固醇检查 |

| | |
|---|---|
| BMI | 体重指数 |
| Smoker | 吸烟者 |
| Stroke | 中风 |
| HeartDiseaseorAttack | 心脏疾病 |
| PhysActivity | 身体活动 |
| Fruits | 水果 |
| Veggies | 蔬菜 |
| HvyAlcoholConsump | 酗酒者 |
| AnyHealthcare | 任何医疗保健 |
| NoDocbcCost | 是否就医成本 |
| GenHlth | 健康状况 |
| MentHlth | 心理健康 |
| PhysHlth | 身体健康 |
| DiffWalk | 走路/爬楼困难 |
| Sex | 性别 |
| Age | 年龄 |
| Education | 教育 |
| Income | 收入 |
| target | 0为非糖尿病，1为糖尿病前期，2为糖尿病 |

## 评估指标

本次作业的评价标准采用 f1-score，即分数越高，效果越好。评估代码参考：

```
from sklearn.metrics import f1_score
y_pred = [1, 0, 2, 0]
y_true = [0, 0, 1, 0]
f1_score(y_true, y_pred, average='macro')
```

## 要求

1. 实现**不少于两种**人工智能算法，解决上述问题。
2. 撰写相应分析报告（**双面打印，不多于 6 页**）。

## 作业提交要求

1. 每个小组打包好的源码、电子版报告、成员分工与小组自评表（1 份），发送至邮箱：jy.lin@fjnu.edu.cn

2. 源码压缩包应至少包含一份 jupyter notebook 文件,该文件展示大致建模过程和实验结果，其余**自定义**文件可以以普通 python（.py）文件包含在压缩包内。
3. 报告纸质版文件于第 15 周随堂提交。

## 提示

1. 报告中可包含对任务的思考、数据的分析、模型选择的原因、对查阅相关资料的思考和总结、不同模型的比较、结果的分析、以及可改进思路的思考等等。
2. 报告中**不能**包含任何实验过程的源码，如有需要可以包含**简洁**的伪代码。
3. 撰写报告时建议图文并茂，在分析时可适当加入图表。
4. 可以使用未学过的人工智能模型或算法，但必须在报告中给出模型原理、选择原因、结果分析。
5. 可使用任意实验课中未涉及的 python 人工智能库（library），不必给出详细说明。

## 加分项

小组作业上传 GitHub 开源，需要撰写相关**英文** ReadMe 文件,并进行必要**英文**注释。作业总成绩最多加 5 分，依据组员分工不同，小组成员所加分数会有差异（会参考成员分工与小组自评表）。**开源地址请于实验报告中标明。**

## 小组作业阶段性 ppt 汇报

**第 12 周实验课（最后一节实验课），组长汇报展示（每组 10~15 分钟）**

**后续可继续完善小组作业**