糖尿病预测项目介绍书

策划人：赵思祺 202100203052 丘碧阳 202100203043

一、 清晰的问题陈述/定义

糖尿病是一种以高血糖为特征的代谢型疾病。本项目目的是探究糖尿病与年龄等因素间潜在关系，并且根据这些因素对皮马印第安女性进行糖尿病预测。

二、 拟使用的数据介绍

糖尿病数据集源自 kaggle 网站，统计于国家糖尿病、消化和肾脏疾病研究所，该数据集共有八个特征（分别为怀孕次数、血压、皮肤厚度、胰岛素、BMI、糖尿病遗传函数、年龄）和一个对应标签（糖尿病标签，0 为无糖尿病，1 为有糖尿病）。数据所统计的均为年龄为 21 岁以上的皮马印第安人血统的女性患者。

三、 实现计划+拟运用的工具、方法、模型等

实现计划:

- 找到相关的数据集，针对数据进行数据处理，对数据的特征进行特征选择。
- 建立模型，训练模型，进行调参。输入最优参数，提高模型精确率。
- 得到各因素与糖尿病间的潜在关系，预测患者是否得糖尿病。

工具:

- python、 Excel

方法:

- 使用 python 相关模块，查看数据完整性，将数据可视化（绘制散点图、直方图等）分析得到各因素与糖尿病潜在关系，再通过相关性筛选特征，所得特征将输入后续模型（分为 Logistic、SVM 和 XGBoost）。然后对模型进行训练，再逐步调参以提高精确度，利用混淆矩阵对模型进行评价及精确度的计算，当精确度达到一定高度时，将数据输入模型中即可得预测的用户糖尿病标签。

- 使用 Excel 或 python 对所得数据进行可视化展示，方便直观的对数据分析的结果进行可视化展示。

模型:

- 使用对比分析、用户分群、用户画像分析有针对性的对主题进行分析。
- 绘制热力图观察各个属性之间的相关性，处理多重共线性的变量。
- 将选择好的属性值输入对糖尿病风险预警模型进行训练。
- 参考 Logistic、SVM 和 XGBoost 模型，进行后续研究。

四、 研究计划（学期里程碑）

第九周 完成数据的查找和收集，得到 .csv 文件

第十周到第十二周 完成数据处理、数据分析、选择特征

第十三周到第十六周 完成模型训练、优化、对比、数据结果可视化

第十六周到第十八周 总结、海报、论文、ppt 整理

五、 小组成员分工

丘碧阳：数据处理、模型训练、模型总结、海报、实践报告

赵思祺：收集数据、数据结果可视化、模型对比、ppt、实践报告