

## 1. 项目背景

按照ADA标准，我国成年人中糖尿病前期患病率超过35%，在糖尿病前期进行控糖能够有效降低II型糖尿病和心血管疾病的风险，具有重大意义。然而，由于成本体验等原因，糖尿病前期人群中通过有创设备监测控糖水平的人很少，智能穿戴技术的巨大进步为解决糖尿病前期人群的控糖水平监测提供了前所未有的机会。随着智能医疗的发展，将机器学习模型应用于高糖预测可以使健康管理更加智能化和个性化。通过实时监测患者的生物指标和生活数据，结合机器学习算法进行实时预测，可以为患者提供个性化的健康建议和风险评估。因此，通过开展这样一个机器学习与模式识别课程设计项目，可以加深对高糖预测问题的理解，掌握相关的数据处理和机器学习技术，并将其应用于实际的健康管理和医疗领域。

## 2. 项目目标

通过记录用户的葡萄糖浓度 (mg/dl) ,血容量脉冲 (BVP) 信号、皮肤电活动 (EDA) 、皮肤温度 (TEMP) 、心率 (HR) 、搏动间隔 (IBI) 和三轴加速度计 (ACC) 。以血糖监测设备得到的血糖值是否处于正常范围 (小于7.8mmol/L, 即140mg/dl) , 把血糖值分类为正常值和异常高值 (高糖) 共两类。以两分类作为金标, 从穿戴设备记录的数据种提取特征, 训练机器学习模型, 测试人群在一天内的不同阶段是否会发生高糖。此项目可以为高糖预测和健康管理领域带来实际的应用和益处。

## 3. 数据预处理

本项目包含一份血糖监测设备记录用户测量值数据集，由16个糖前（或者接近糖前）用户组成，血糖监测设备记录每个用户的葡萄糖浓度 (mg/dl) ,穿戴记录了每个用户的血容量脉冲 (BVP) 信号、皮肤电活动 (EDA) 等多个信号数据。由于这些数据分散在多个表中且数据量较大，需要考虑如何采样和数据整合问题。

## 4. 数据探索分析

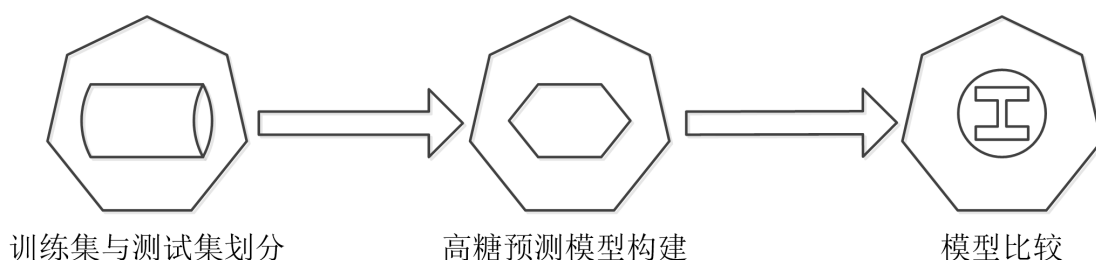
原始数据中，不同数据存在不同的缺失率。根据特征类型、特征意义我们需要采用不同的缺失值处理方法，缺失过于严重的特征直接删去，其他根据意义进行填充，缺失值填补好之后我们还需对入模特征进行进一步处理如离散化、数字编码等，同时对异常值进行排查和处理。

- 缺失值处理：依据不同特征属性值含义进行不同的缺失值填补处理，如使用默认值；和使用众数填补等。
- 特征编码：对于某些离散型的特征，由于特征中各个取值之间没有大小关系，我们将其进行One-Hot编码。
- 异常值处理：包括检验离散型特征的唯一取值，看是否为特殊标记值；检验连续型特征的最大最小值，以及通过盒图排除异常值；检验数据取值与业务逻辑不符合的数据等。

同时还将利用Python中绘制图形的库(如Seaborn、Matplotlib)进行一些可视化操作，辅助对数据的预处理工作，同时探究各个特征各自的统计信息及其与预测目标之间的关系。

## 5. 构建高糖预测分类模型

经过前两天的学习与实验，已经完成了数据预处理以及特征选择工作。数据准备完毕后就要完成模型的选择工作。构建高糖预测分类模型的基本流程为：



运用两种以上分类算法，构建高糖预测分类模型，并使用正则化、交叉验证和网格搜索等技术优化模型效果。

## 6. 模型预测及分析

在模型训练完成后，使用测试样本对模型进行评估。本项目通过预测高糖的准确度accuracy、精确率precision、召回率recall和F1分数f1score来评估模型。然后，分析现有模型，从数据获取与预处理、特征提取、任务建模等多方面进行分析与改进。

## 7. 项目总结

请从项目目的、项目流程、数据处理方法、高糖预测模型构建方法、效果分析和初步结论等方面对本项目进行总结。项目目的：解决了什么样的实际问题？项目流程：项目的总体执行流程是怎样的？数据处理方法：对项目数据进行了怎样的处理？高糖预测模型构建方法：使用了哪些方法构建高糖预测模型？效果分析：使用哪些方法对高糖预测模型的效果进行了分析与评估？初步结论：通过本项目，得到了哪些初步的结论？