# 人工智能大作业

# 糖尿病预测

开源地址：https://github.com/ChLiQiYu/New-Diabetes-detection.git

## 1 介绍

### 1.1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

### 1.2任务背景

随着人工智能技术的迅速发展，其在医疗健康领域的应用越来越广泛。糖尿病作为一种全球性的慢性疾病，对人类社会的健康和经济都造成了巨大的负担。因此，利用人工智能技术进行糖尿病的早期预测和干预，对于提高患者的生活质量和减轻社会经济压力具有重要意义。本报告旨在介绍人工智能在糖尿病预测中的应用背景，并概述研究的主要内容和目的。

### 1.3任务目标

本次项目是根据健康指标相关信息数据，然后选择合适的人工智能算法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。

## 2 数据分析与处理

### 2.1 数据描述

|  |  |
| --- | --- |
| **特征名称** | **特征解释** |
| **Id** | 样本标识ID |
| **Hig**hBP | 是否高血压 |
| **HighChol** | 高胆固醇 |
| **CholCheck** | 胆固醇检查情况 |
| **BMI** | 体重指数 |
| **Smoker** | 是否吸烟 |
| **Str**oke | 是否有过中风 |
| **HeartD**isea**seorAttack** | 心脏疾病或心脏病发作历史 |
| **PhysActi**vity | 身体活动频率 |
| **Fru**its | 水果摄入量 |
| **Veggies** | 蔬菜摄入量 |
| **HvyA**lcoh**olConsump** | 是否酗酒 |
| AnyHealthcare | 任何医疗保健 |
| NoDocbcCost | 就医成本问题 |
| GenHlth | 一般健康状况 |
| MentHlth | 心理健康状况 |
| PhysHlth | 身体健康状况 |
| DiffWalk | 走路/爬楼困难 |
| Sex | 性别 |
| Age | 年龄 |
| Education | 教育程度 |
| Income | 收入水平 |
| target | 0为⾮糖尿病，1为糖尿病前期，2为糖尿病 |

利用数据分析库pandas对data数据集进行数据处理及分析，得出上表，可以看到样本一共有300000份，且一共有22个特征值，1个目标值。

### 2.2 数据预处理

**缺失值处理**：利用数据分析库pandas 检查并处理缺失值，通过运行发现数据集无缺失值，所以不需要对缺失值进行补充

**分离无关特征**：通过对数据的分析我们可以发现id是无关数据，所以当我们对数据进行模拟的时候，需要将id这个特征值剔除

**分离目标变量**：通过对数据的分析我们可以发现target是目标列，所以当我们对数据进行模拟的时候，需要将target这个特征值剔除，转换为目标变量

**特征标准化**:利用StandardScaler特征标准化库对data进行特征标准化，使特征数据不易受异常值的影响

**过采样处理**：利用通过随机复制少数类样本来增加其数量，防止某一类数量过大

**分割数据集**：利用train\_test\_split将数据集被划分为训练集和测试集，比例为80%训练集和20%测试集，同时保证测试集与整个数据集里target的数据分类比例一致。模型运行之后输出准确率，最终模型的性能通过f1\_score进行评估。

## 3 算法选择与训练

为了解决糖尿病预测问题，我们选择了以下三种算法进行实验：

**K-近邻算法（k-nearest neighbor classification）**：

**原因**：因为其简单性和解释性强，KNN是一个基本而简单的分类算法。作为监督学习， KNN模型需要的是有标签的训练数据，对于新样本的类别由与新样本距离最近的k个训练样本点按照分类决策规则决定。

**实验**：对于本次任务，我们从sklearn中导入KNeighborsClassifier，通过多次实验，观察准确率和f1\_score的得分，最终确定n\_neighbors=5，即取邻近点的个数5，再利用 KneighborsClassifier中所封装的方法来对数据集进行预测。

**随机森林（Random Forest）**：

**原因**：随机森林是一个一个强大的集成学习方法，将多个决策树结合在一起，每次数据集是随机有放回的选出，同时随机选出部分特征作为输入，输出通常能够提供比单一模型更好的分类效果。由于集成了多个模型，随机森林能够容忍个别决策树的过度拟合或错误，因为最终结果是由多数树的预测综合得出的。

**实验**：对于本次任务，我们从sklearn中导入RandomForestClassifier，通过多次实验，观察准确率和f1\_score的得分，最终确定n\_estimators=100，即一共构造了100个决策树，再利用 RandomForestClassifier中所封装的方法来对数据集进行预测，最后还绘制了RandomForestClassifier模型下各个特征的重要性，给实验带来了更直观的感受。

**梯度提升决策树（Gradient Boosting Decision Tree）**

**原因**：梯度提升决策树通过构造一组弱的学习器（树），并把多颗决策树的结果累加起来作为最终的预测输出。梯度提升树能够对特征的重要性进行评分，有助于理解模型的决策过程。通过控制树的深度和学习率等参数，可以有效避免模型过拟合。

**实验**：对于本次任务，我们基于lightgbm来构造梯度提升决策树，通过多次实验，观察准确率和f1\_score的得分，最终确定参数（见下表）,并且设置模型训练1000轮，取最好的一轮实验数据来输出。之后用lightgbm中所封装的方法来对数据集进行预测，最后还绘制了梯度提升决策树模型下各个特征的重要性，给实验带来了更直观的感受。

|  |  |  |
| --- | --- | --- |
| **参数名称** | **参数设置** | **参数解释** |
| boosting\_type | gbdt | 使用 GBDT 算法 |
| objective | multiclass | 指定为多分类问题 |
| metric | multi\_logloss | 使用交叉熵损失函数 |
| max\_depth | 5 | 限制每棵树的最大深度为 5 |
| num\_leaves | 3 | 随机森林中每个决策树的最大叶子节点数为 3 |
| learning\_rate | 0.01 | 设置学习率为 0.01 |
| feature\_fraction | 0.7 | 每次训练时随机选择 70% 的特征进行训练 |
| bagging\_fraction | 0.7 | 设置每次取出 70% 的样本数进行训练 |
| bagging\_freq | 15 | 每 15 次迭代进行一次 bagging，从原始数据集中随机抽取部分样本进行训练 |
| verbose | -1 | 不输出训练过程的详细信息 |
| num\_class | 3 | 指定类别的数量为 3 |

## 4. 结果与分析

### 4.1 结果展示

|  |  |  |
| --- | --- | --- |
| **算法名称** | **算法准确率** | **算法f1\_score** |
| K-近邻算法 | 0.9073 | 0.9048 |
| 随机森林 | 0.9459 | 0.9460 |
| 梯度提升决策树 | 0.6422 | 0.6422 |

### 4.2 结果分析

通过结果展示，我们可以清楚的看到，随机森林模型在准确率和f1\_score上相较于K-近邻算法和梯度提升决策树明显更好。而梯度提升决策树准确率和f1\_score上相较于其他两者很低。

**分析**：这可能是由于随机森林是一种基于多个决策树的集成学习方法，通过组合多个决策树的预测结果，来提高模型的准确率和泛化能力。而这种集成方法可以减少单个模型的过拟合问题，从而提高模型的稳定性和准确率。而梯度提升决策树通过迭代地构建多个弱分类器来提高模型的性能，但这种方法容易导致过拟合，尤其是在没有适当正则化的情况下，且梯度提升决策树的训练过程涉及多次迭代，计算成本较高。同时梯度提升决策树涉及大量的参数调整，需要通过多次测试，仔细调优才能达到最佳性能

而通过绘制的随机森林特征重要性以及梯度提升决策树重要性，我们可以很直观的看到在两个模型当中BMI age 以及GenHlth特征重要性都很高，这代表着BMI age 以及GenHlth的特征权重很高，即对结果有着重要影响。

## 5 讨论与改进方向

尽管我们经过了第一次和第二次优化，模型已经取得了不错的预测效果，但仍有改进空间。比如可以还可以绘制出学习率和损失函数来进一步展示分类数据，更为直观。此外，我们还可以尝试其他先进的机器学习模型如支持向量机或深度学习模型可能进一步提高预测性能。

同时在测试过程中，我们发现由于数据集过大，模型预测时需要花费的时间很长，后续改进我们将考虑如何在保证准确率和f1\_score下，使模型更快地输出预测结果。

## 6 结论

本报告展示了使用K-近邻算法,随机森林,以及梯度提升决策树模型对糖尿病进行预测的过程和结果。通过对数据集的深入分析和合理的模型选择，我们成功实现了对糖尿病及其前期状态的有效预测。