**《商务智能》课程项目——预测糖尿病**

摘要和关键词

**本次实验的主要内容是使用回归分析和聚类分析来预测某人患糖尿病的可能性和身体的糖尿病指数。**

关键词：糖尿病；线性回归；聚类分析

目录

1使用说明

2背景

3需求分析

4解决方案

5关键代码实现

6背景

1.使用说明

数据来源：UCI机器学习库http://archive.ics.uci.edu/ml/datasets.html

配置相关： python3：pandas，numpy，matplotlib，seaborn

使用如图（1.1）：



图1.1输入身体指数

结果如图（1.2）：

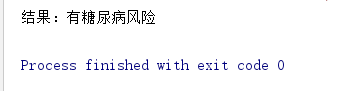


图1.2返回预测结果

2.背景

根据美国预防疾病中心，现在美国有1/7的成年人患有糖尿病。根据增长趋势，到了2050年患糖尿病的人数将高达三分之一。我的父亲就常年受到糖尿病的困扰。但是根据专家研究，只要早点发现糖尿病的趋势，控制好饮食，就能杜绝糖尿病的加重甚至根治。

3.需求分析

用户可以在家里定时测量血压等身体数据，或者定期去医院体检获得数据后。 由医生或者病人自行把数据输入软件中，随后软件会根据数据分析预测后输出结果，判断是否患有糖尿病的风险，提醒患者到医院进行深度检查。

4.解决方案

从网上获取了两个数据集，一个是age、 sex 、bmi、 map、 tc、 ldl、 hdl、 tch、 ltg、 glu几个糖尿病专用指标与糖尿病指数的关系，因为这个数据集的预测结果不是二值，所以可以使用这个数据集训练线性回归预测。第二个数据集如图2.1：

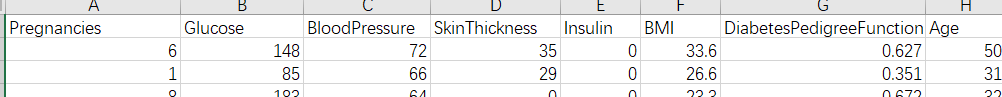


图2.1数据集部分

这几个指标，因为这个数据集结果是二值的——是否得了糖尿病，所以用聚类分析来预测。

* + - 1. 线性回归方法：
         1. 首先验证数据的完整性。然后计算出相关系数矩阵系数越接近1，数据集就越适合线性回归。

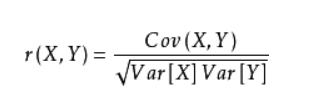


图2.2相关系数计算方法

* + - * 1. 接下来建立散点图来查看数据里的数据分析情况以及对相对应的线性情况，查看这些因素对 糖尿病指数有什么影响，将使用seaborn的pairplot来绘画。可以了解到不同的因素对糖尿病指数影响（置信度= 95 %），也可可以大致看出不同特征对于标签值的影响与相关关系在了解了数据的各种情况后需要对数据集建立模型
        2. 使用train\_test\_split函数来创建训练集和测试集，将训练集中的特征值与标签值放入LinearRegression()模型中且使用fit函数进行训练,在模型训练完成之后会得到所对应的方程式（线性回归方程式）。然后使用测试集验证回归结果，发现拟合地很好。
        3. 之后只需要把用户传来地数据放到回归方程中然后得出糖尿病指数
    1. 聚类分析法：
       1. 数据分析，画出各特征值的直方图，看看有没有异常的数据。去掉异常数据（如血压等于零）
       2. 聚类分析不止一种，先测试看看哪种方法最好，这次实验使用了7种分类器，分别为：K-Nearest Neighbors, Support Vector Classifier, Logistic Regression, Gaussian Naive Bayes, Random Forest and Gradient Boost。
       3. 划分训练集和测试集，放入分类器中训练。然后用K折叠交叉验证（K-Fold Cross Validation）获得各分类器的准确率。
       4. 选用一个准确率最高的分类器，用sklearn的GridSearchCV调整参数，获得最优参数。最后使用最优参数获得分类器，处理用户输入的数据得到预测结果。

5.关键代码实现

线性回归：

检测数据和相关系数，判断是否适合线性回归。

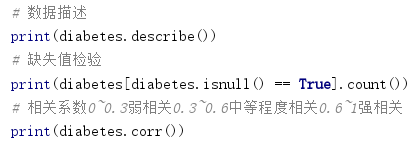


图5.1计算相关系数矩阵代码

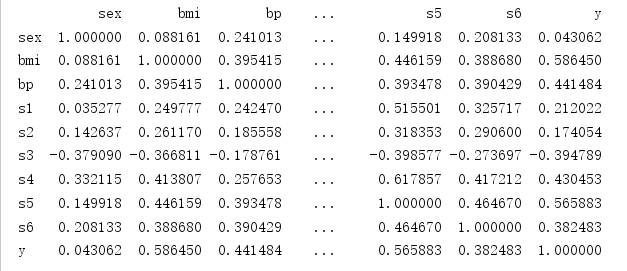
结果：

图5.2相关系数矩阵结果

训练集、测试集划分



图5.3训练集、测试集划分代码

做散点图看特征值对标签值的影响

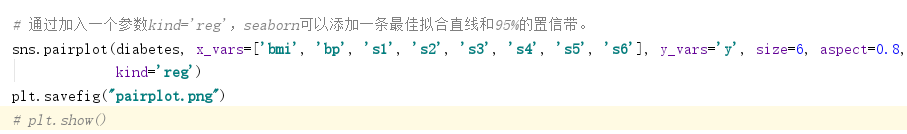


图5.4做散点图代码

结果（部分）：

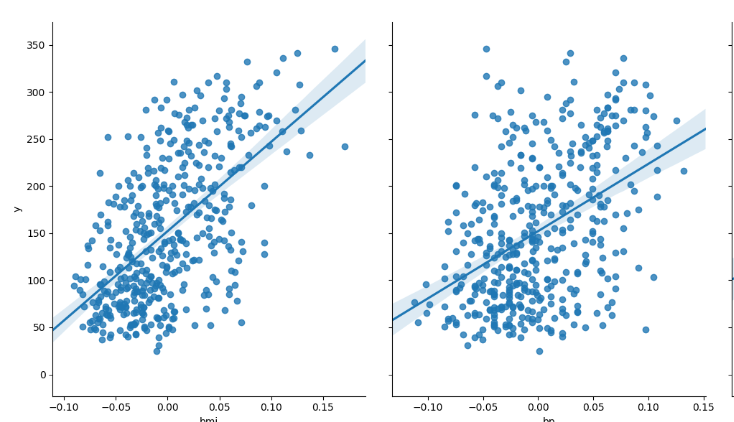


图5.5某两个标签的散点图

开始线性回归，测试准确率，并获得预测结果

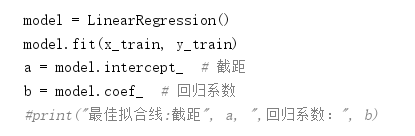


图5.6线性回归代码

结果：



图5.7线性回归参数

b ) 聚类分析：

查看直方图，检测异常值并且去除异常值

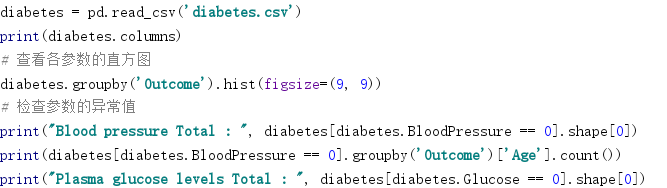




图5.8检查参数代码

使用7种分类器，并获得分类准确率：

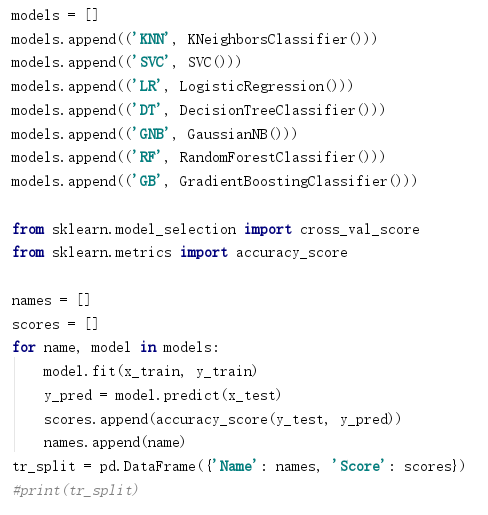


图5.9配置7种聚类分析器的代码

结果：GB准确率最高，用GB进行最终分类

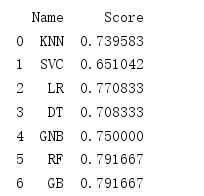


图5.10各类分类器的准确率

使用GridSearchCV调整GradientBoostingClassifier()的参数：

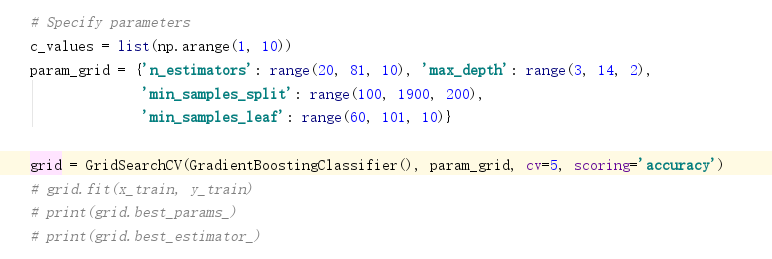


图5.11使用GridSearchCV关键代码

结果：返回最佳参数，用这些参数做最终的分类器分类

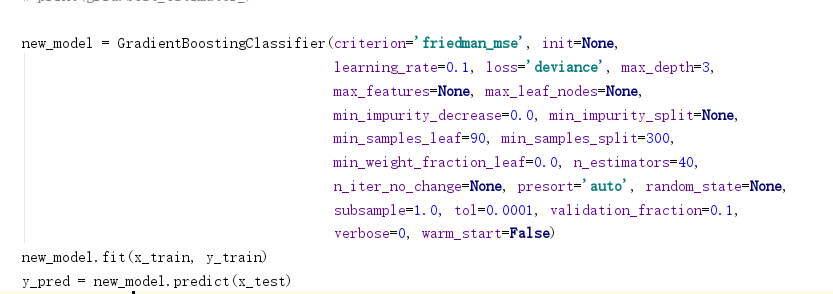


图5.12GridSearchCV返回的最佳参数