

机器学习实践

— 实验报告

|  |  |
| --- | --- |
| 题目： | 糖尿病预测系统 |
| 学号： | 20201111414 |
| 姓名： | 韩天路 |
| 班级： | 20数据A1 |
| 教师： | 陈方疏 |
| 日期： | 2023-09-17 |

**1.项目背景**

糖尿病是一种广泛影响全球各年龄段的慢性代谢性疾病，其在全球范围内的传播速度令人担忧。据世界卫生组织（WHO）的统计数据，糖尿病已成为全球第七大致命疾病，每年导致数百万人的死亡。不仅如此，糖尿病还对个人的生活质量产生了巨大影响，常常伴随着严重的并发症，如心血管疾病、视力丧失、肾病、神经病变等。在这一背景下，早期糖尿病检测和有效的管理至关重要，以减轻其潜在的危害。

机器学习和人工智能技术在医疗领域的应用已经取得了显著进展，糖尿病的早期检测和管理也不例外。通过收集和分析患者的生理特征和健康数据，机器学习模型可以识别患有潜在糖尿病风险的个体，从而帮助医疗专业人员更加精确地进行干预和治疗。因此，本项目旨在充分发挥机器学习的潜力，构建一个高效的糖尿病预测模型，以便更好地为患者提供个性化的医疗建议和支持。

**2 数据探索**

**2.1数据描述**

Diabetes数据集包含了3536条患者的患病标签Outcome(1患病0不患病)和8个属性特征，包括Pregnancies（怀孕次数）、Glucose（葡萄糖浓度）、BloodPressure（血压）、SkinThickness（皮肤厚度）、Insulin（胰岛素水平）、BMI（体质指数）、DiabetesPedigreeFunction（糖尿病谱系功能）、Age（年龄）如表1所示。

表1 数据样例

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
| 0 | 109 | 88 | 30 | 155 | 32.5 | 0.855 | 38 | 1 |
| 2 | 109 | 92 | 29 | 155 | 42.7 | 0.845 | 54 | 0 |
| 1 | 95 | 66 | 13 | 38 | 19.6 | 0.334 | 25 | 0 |
| 4 | 146 | 85 | 27 | 100 | 28.9 | 0.189 | 27 | 0 |
| 2 | 100 | 66 | 20 | 90 | 32.9 | 0.867 | 28 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

**2.2 标签分布**

根据统计，不患病与患病的年龄分布大致呈现如下，不患病平均年龄为31.10，大部分的鲍鱼年龄集中在25-50岁之间，约占75.50%。患病平均年龄为36.95，大部分的年龄分布在22-50之间，约占80%左右

患者年龄分布 非患者年龄分布

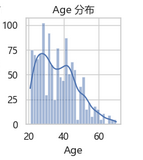
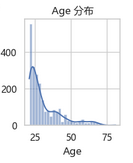
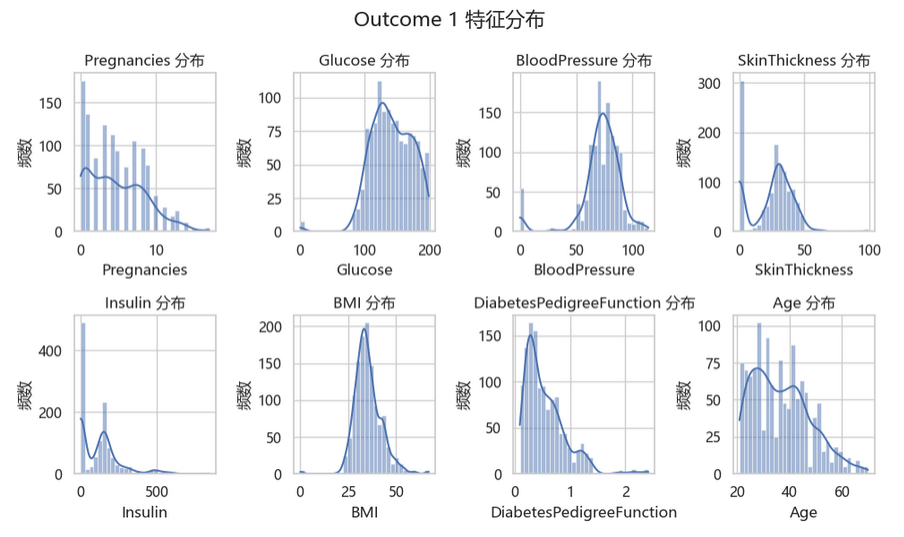
 

图1患者年龄分布图

**2.3特征分布**

患者的怀孕次数在平均值约为3.76次，最少为0次，最多为17次。口服葡萄糖后，2小时血浆葡萄糖浓度的平均值约为121.23毫克/分升。舒张压（BloodPressure）的平均值约为69.84毫米汞柱，但存在最小值为0的异常情况，可能需要进一步检查和处理。三头肌皮肤褶皱厚度（SkinThickness）的平均值约为22.62毫米，同样存在最小值为0的异常情况。2小时血浆胰岛素浓度（Insulin）的平均值约为96.45微国际单位/毫升，也有最小值为0的异常情况。患者的平均体质指数（BMI）约为32.21，最小值为0，可能需要进一步检查和处理。糖尿病家族史的函数得分（DiabetesPedigreeFunction）的平均值约为0.47，用于衡量遗传糖尿病风险。患者的平均年龄约为33.16岁，最小值为21岁，最大值为81岁。标签列指示患者是否患有糖尿病，其中1表示患病，0表示不患病，大约34.5%的样本患有糖尿病。

以下是各个特征的在患者和非患者中的分布



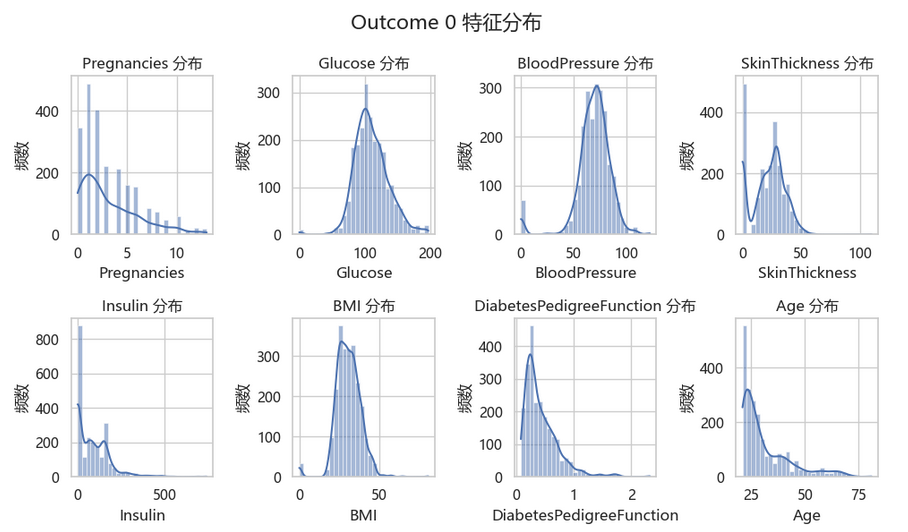
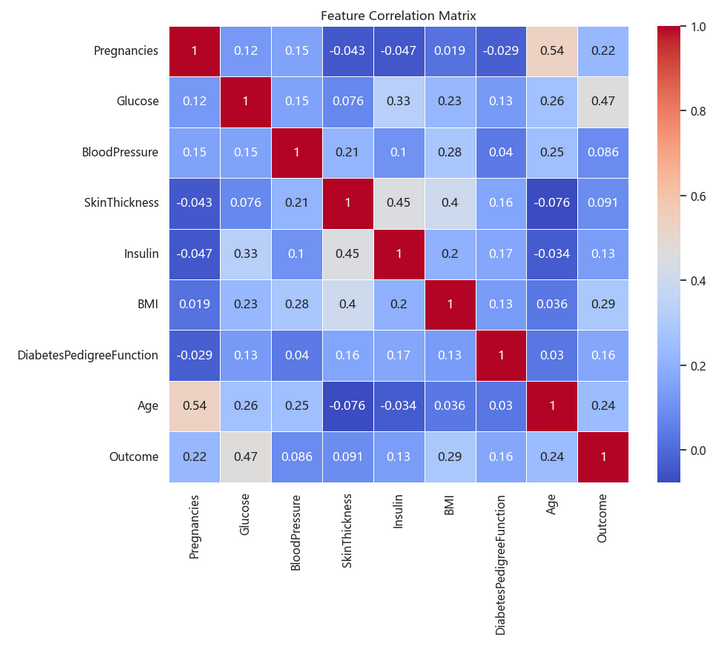


图2特征分布

**2.4特征相关性**

"Glucose"（葡萄糖浓度）特征与"糖尿病"之间存在较高的正相关性，相关系数约为0.467，这表明患者的葡萄糖浓度升高可能会增加患糖尿病的风险。"BMI"（体质指数）特征也与"糖尿病"之间存在一定的正相关性，相关系数约为0.286，这意味着患者的体质指数增加可能会略微提高患糖尿病的可能性。此外，"Age"（年龄）和"Pregnancies"（怀孕次数）也与"糖尿病"之间存在一定程度的正相关性，相关系数分别约为0.237和0.223，这表明年龄和怀孕次数增加可能会稍微提高患糖尿病的风险。"DiabetesPedigreeFunction"（糖尿病谱系功能）也呈现一定程度的正相关性，相关系数约为0.163，表明患糖尿病家族史对患糖尿病的风险有一定影响。然而，"BloodPressure"（血压）、"SkinThickness"（皮肤厚度）和"Insulin"（胰岛素水平）特征与"糖尿病"之间的相关性较低，相关系数分别约为0.085、0.090和0.134，这些特征可能对糖尿病的预测影响较小。

根据这些相关性结果，初步推测葡萄糖浓度、体质指数、年龄和怀孕次数可能是预测患糖尿病的重要特征，而血压、皮肤厚度和胰岛素水平可能对预测的贡献较小。



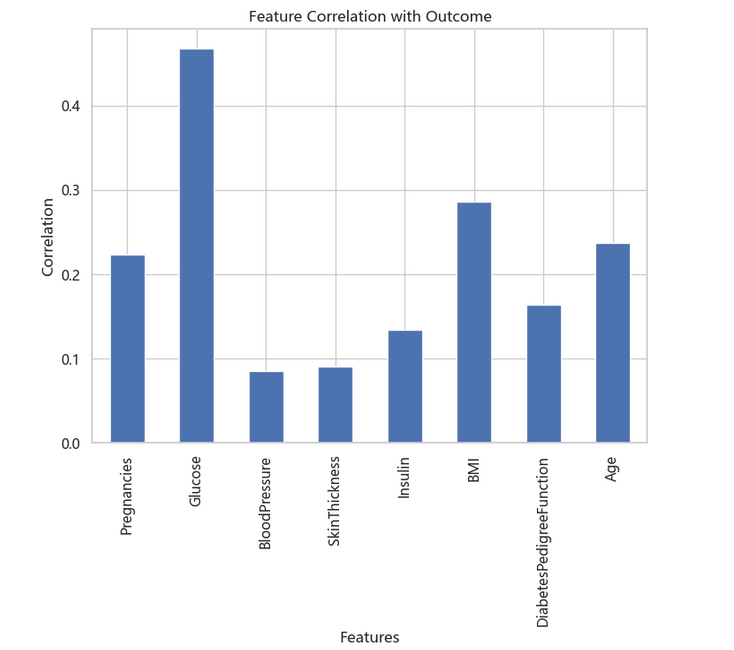


图2特征与标签相关性分析

**2.5共线性分析**

本实验中的分类问题可以抽象为回归问题，其核心在于构造预测标签与8个属性特征的映射函数。在该过程中，如果特征,之间存在很显著的共线性，在预测过程中则为无效信息，甚至将对模型造成一定的干扰。选择一个

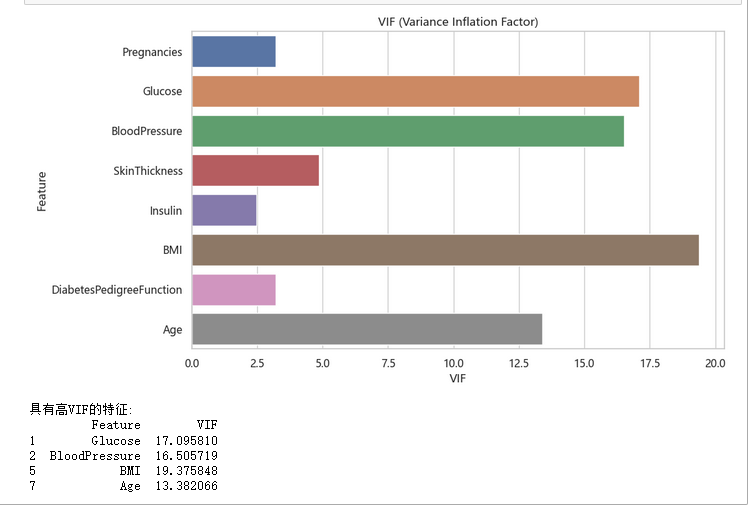
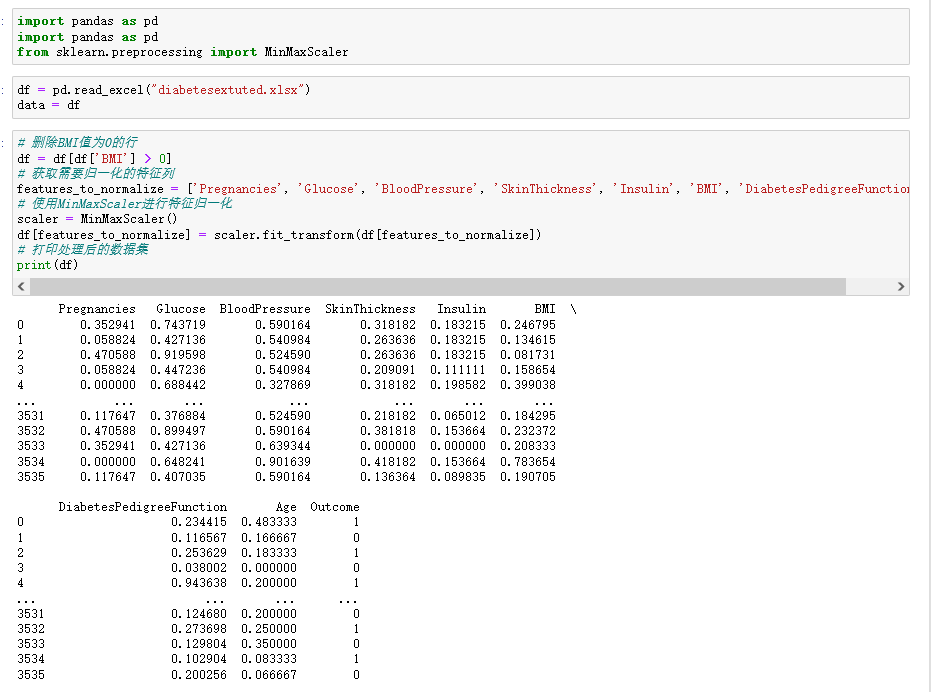


图2特征共线性分析

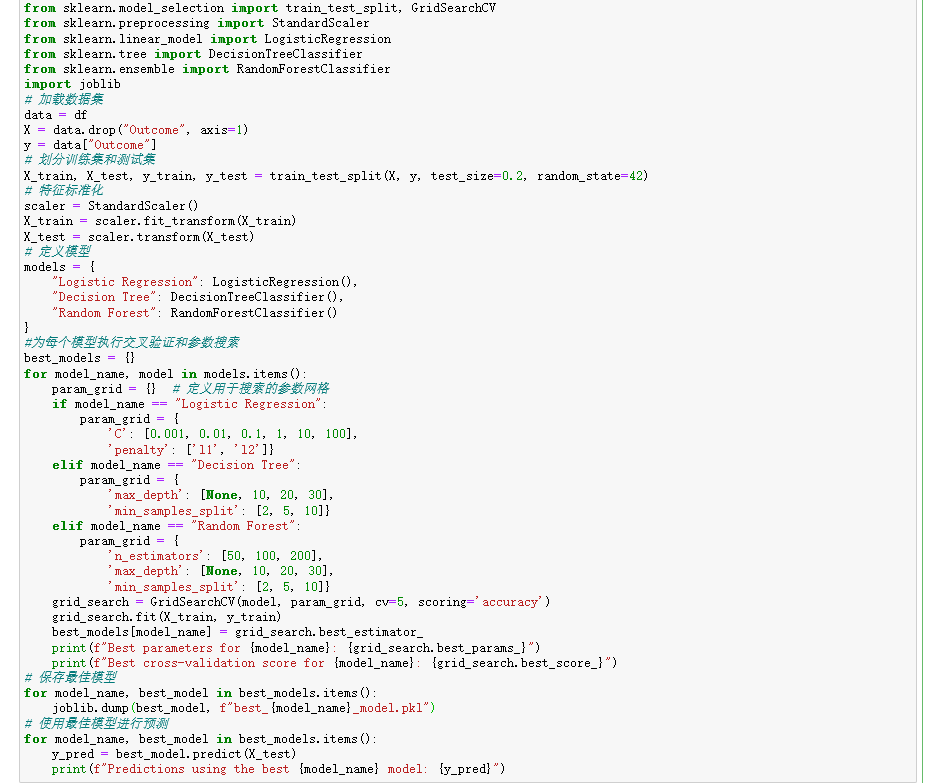
**3 模型设计**

**首先进行数据处理，然后进行模型参数调优，选择最优参数，然后将最优参数写在web系统中**

1. **数据处理**



1. **模型调优**



* 线性回归（Linear Regression，LR）

使用LogisticRegression()创建了一个逻辑回归分类器，并将其存储在models字典中，以便后续的参数搜索。

使用网格搜索（GridSearchCV）方法进行参数搜索。为了找到最佳的逻辑回归模型，我搜索了两个关键参数：'C'：正则化参数，控制模型的复杂度。在代码中，我定义了一系列可能的C值（0.001, 0.01, 0.1, 1, 10, 100）。'penalty'：惩罚项，可以是'l1'或'l2'。这些惩罚项有助于控制模型的过拟合。网格搜索通过交叉验证来评估不同参数组合的性能，并选择具有最佳交叉验证分数的参数组合。

保存最佳模型：最佳的逻辑回归模型（即具有最佳参数组合的模型）被保存到best\_models字典中，以备后续使用。

预测：

对于最佳逻辑回归模型，使用测试集数据进行预测，并输出预测结果。

* 决策树（Decision Tree，DT）

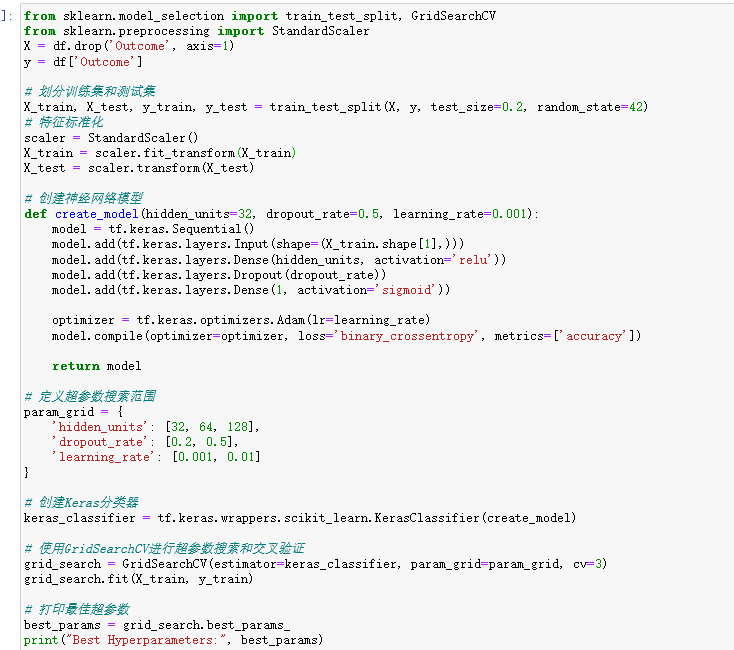
使用DecisionTreeClassifier()创建了一个决策树分类器，并将其存储在models字典中。

通过网格搜索（GridSearchCV）方法进行参数搜索，以找到最佳的决策树模型。搜索了以下两个关键参数：'max\_depth'：决策树的最大深度。在代码中，定义了一系列可能的最大深度值，包括不限制深度（None）、深度为10、20和30。'min\_samples\_split'：内部节点再划分所需的最小样本数。搜索了包括2、5和10个最小样本数的不同情况。网格搜索通过交叉验证来评估不同参数组合的性能，并选择具有最佳交叉验证分数的参数组合。最佳的决策树模型（即具有最佳参数组合的模型）被保存到best\_models字典中。

* 随机森林（Random Forest，RF）

使用RandomForestClassifier()创建了一个随机森林分类器，并将其存储在models字典中，通过网格搜索（GridSearchCV）方法进行参数搜索，搜索了以下三个关键参数：'n\_estimators'：随机森林中树的数量。定义了一系列可能的树的数量，包括50、100和200。'max\_depth'：决策树的最大深度。定义了一系列可能的最大深度值，包括不限制深度（None）、深度为10、20和30。'min\_samples\_split'：内部节点再划分所需的最小样本数。搜索了包括2、5和10个最小样本数的不同情况。保存最佳模型，输出预测结果。

* 神经网络(Neural Network)



从DataFrame df 中分离出特征矩阵 X 和目标向量 y，其中 X 包含了特征列，y 包含了二分类的目标标签。使用 train\_test\_split 函数将数据划分为训练集（X\_train 和 y\_train）和测试集（X\_test 和 y\_test），测试集占总数据的20%。使用 StandardScaler 对特征进行标准化，以确保不同特征的尺度不会影响神经网络模型的性能。定义了一个函数 create\_model，该函数用于创建神经网络模型。模型包括：输入层：与特征数相匹配。隐藏层：包含指定数量的神经元（hidden\_units），使用 ReLU 激活函数。Dropout 层：用于减少过拟合，随机丢弃一部分神经元的输出。输出层：包含单个神经元，使用 sigmoid 激活函数，适用于二分类问题。定义了优化器（Adam）和损失函（binary\_crossentropy）。

定义了超参数网格，包括隐藏层神经元数量、Dropout 率和学习率的不同取值。

使用 tf.keras.wrappers.scikit\_learn.KerasClassifier 将 Keras 模型包装成一个Scikit-Learn兼容的分类器，以便与 GridSearchCV 结合使用。

创建 GridSearchCV 对象，传入模型（keras\_classifier）、超参数网（param\_grid）和交叉验证折数（cv=3，本例中为3折交叉验证）。调用 fit 方法执行超参数搜索和交叉验证。

**4 实验分析与讨论**

**4.1 实验设置**

所有模型采用Python程序语言实现，运行环境为Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz , 16.0 GB RAM，GPU为NVIDIA GeForce GTX1050。

实验过程中将整个数据集分为测试集20%，训练集80%

针对各模型的预测结果，本文采取了平均绝对误差（MAE）,均方误差(MSE),均方根误差（RMSE）,相关系数（）指标进行评价，计算方式如公式1-4，其中表示测试集的样本数量， 分别表示模型的预测值和真实值，表示预测的均值。

|  |  |  |
| --- | --- | --- |
|  |  | (1) |
|  |  | (2) |
|  |  | (3) |
|  |  | (4) |

**4.2 模型预测效果对比**

针对分类问题，本文采用了多种算法进行预测，包括线性回归（Linear Regression，LR）,决策树（Decision Tree，DT），随机森林（Random Forest，RF），神经网络(Neural Network)，实验结果如表2所示。

表2模型预测效果对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Metric***  ***Model*** | ***MSE*** | ***RMSE*** | ***MAE*** |  |
| **LR** | 0.2314 | 0.4811 | 0.2314 | -0.0280 |
| **NN** | **0.2442** | 0.4943 | 0.4669 | -0.0413 |
| **DT** | 0.0649 | 0.0649 | 0.0042 | 0.9812 |
| **RF** | 0.0112 | 0.1059 | 0.0112 | **0.9510** |

**4.3 训练集规模对预测效果的影响**

将训练集的占比设置为0.1-0.9，并分别记录了LR、RF、DT、NN 4个模型的预测效果。随着训练集规模的增长，预测效果较为平缓，说明仅需要很少量的标注数据，模型即可学到充足的信息，可以降低数据标注的成本。

**4.4 运行时间分析**

如图所示LR和DT的训练时间最少,RF的时间居中，NN的训练时间最长



**5 系统设计与实现**

**5.1 开发环境**

Python3.8

Intel(R) Core(TM) i7-7700HQ CPU

16.0 GB RAM

GTX 1050

**5.2 总体设计**

使用web.py编写系统分为6个页面，分别是[主页](http://localhost:8080/index.html),[查询](http://localhost:8080/tnb.html),[预测](http://localhost:8080/pred.html),[训练](http://localhost:8080/dashboard.html),[大屏](http://localhost:8080/screen.html),[历史日志](http://localhost:8080/log.html)，实现数据可视化，数据查询，数据预测的功能

**5.3 详细设计**

[主页](http://localhost:8080/index.html)--展示数据，展示数据集中标签占比

查询--输入编号即可查询到指定信息，并返回真实值和预测值，预测值使用决策树模型

预测--输入属性值，选择不同的模型，点击提交即可得到预测值

训练--实现新增,删除,修改数据,根据CSV文件批量导入数据。也可以选择不同的模型训练，训练出的模型将会更新在系统中

大屏--一些数据展示，模型准确度的展示，各指标不同维度的数学指标

历史日志--显示用户的操作日志和用户训练模型的日志

**6 总结**

糖尿病是一种全球性的健康问题，早期检测和有效管理对患者的健康至关重要。机器学习和人工智能技术在医疗领域的应用，是未来的必然趋势。

在数据探索阶段，详细分析了提供的糖尿病数据集，包括标签分布、特征分布和特征相关性。这些分析可以更好地理解数据，为后续的建模工作提供了基础。

在模型设计阶段，尝试了多个机器学习算法，包括线性回归、决策树、随机森林和神经网络。使用不同的评估指标来衡量模型的性能。

在实验分析与讨论中，使用了多个指标来比较不同模型的性能，并对模型在不同训练集规模下的表现进行了分析。这有助于更好地理解模型的优势和局限性。

最后，在系统设计与实现阶段，设计了一个Web应用程序，将模型应用于实际情境中。这个应用程序允许用户查询数据、进行预测，并提供了数据可视化和历史日志功能。

**7 心得体会**

首先，我认识到了糖尿病是一种严重的健康问题，它可以对人们的生活产生巨大的负面影响。因此，早期的糖尿病检测和管理至关重要。机器学习和人工智能可以在医疗领域中发挥关键作用。通过分析患者的数据，机器学习模型可以帮助医生更好地了解患者的健康状况。同时在机器学习项目中，数据是关键。通过仔细分析数据，能够更好地理解问题，并准备数据以供模型使用，同时数据处理对模型的准确度也至关重要。最后不同的机器学习算法，有不同的效果，需要不断地尝试，才能找到一个最佳的模型。