1.研究提案

1.1研究背景

糖尿病对身体最大的危害不是它本身的疾病，而是它所引起的并发症。其中当糖尿病患者体内酮体轻度增加时，机体通过自身调节可以维持血液的酸碱度在正常范围，此时机体仅有酮症而无酸中毒，称为糖尿病酮症。

1.2研究目的

本项目旨在通过糖尿病相关指标来预测糖尿病患者酮症的发病情况，找出与发生酮症相关的因素，采取适当的措施，从而更好的控制酮症的发生。

1.3研究对象

南京逸夫医院所提供的224例糖尿病患者。

1.4研究方案

对收集到的224例患者的检查指标信息进行数据预处理后，采用递归特征消除法和模型选择法分别进行特征选择，对选择后的数据集按照7:3的比例构建训练集和测试集，分别采用GBDT,XGBOOST,RF进行建模，准确率作为评价指标，对不同数量的特征数据集进行对比，选出特征相关性最高且准确率最高的结果。

处理流程

2.数据分析

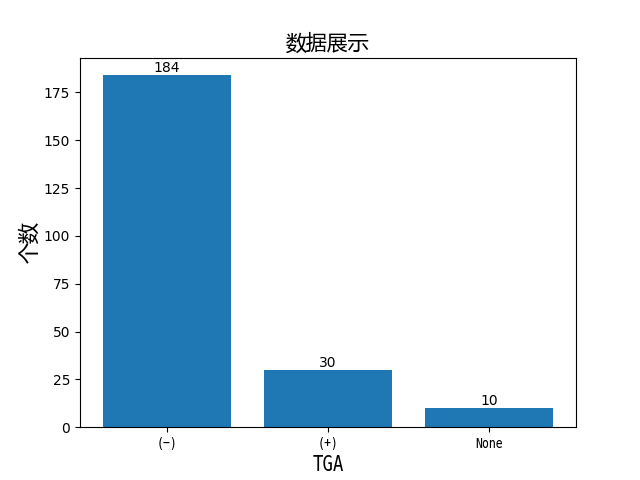
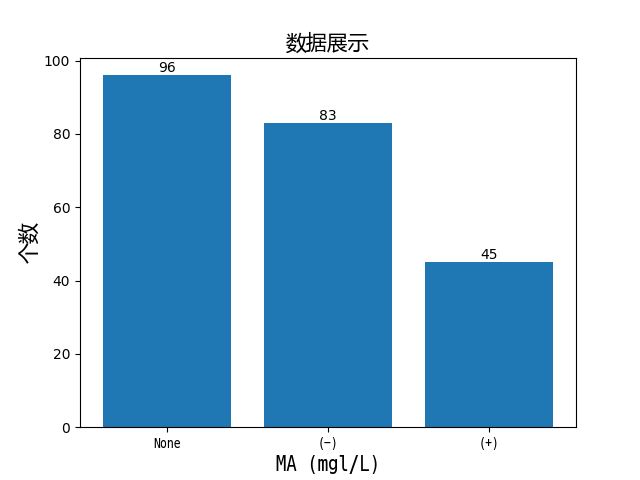
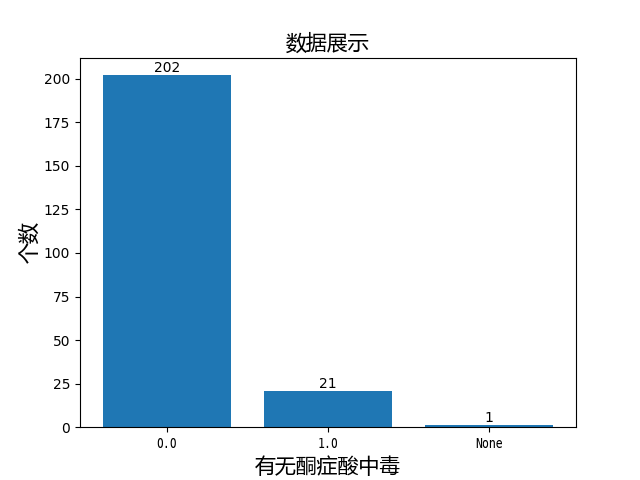
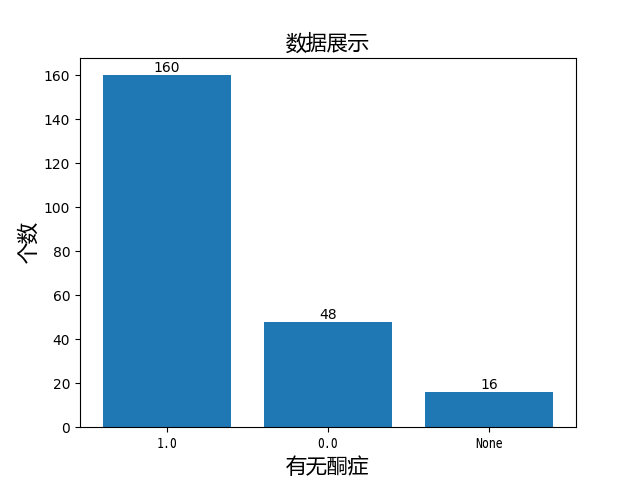
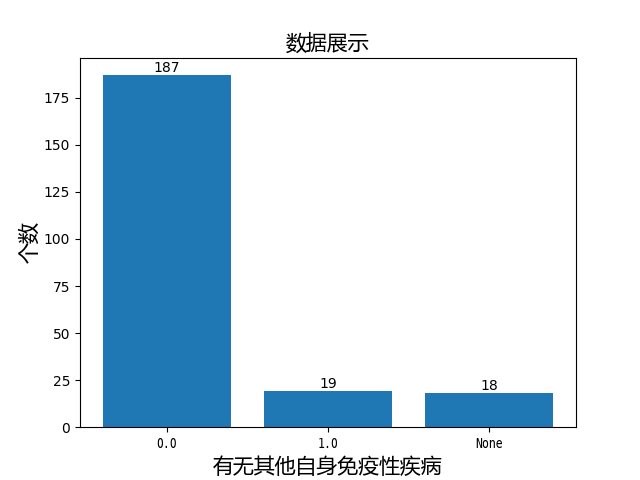
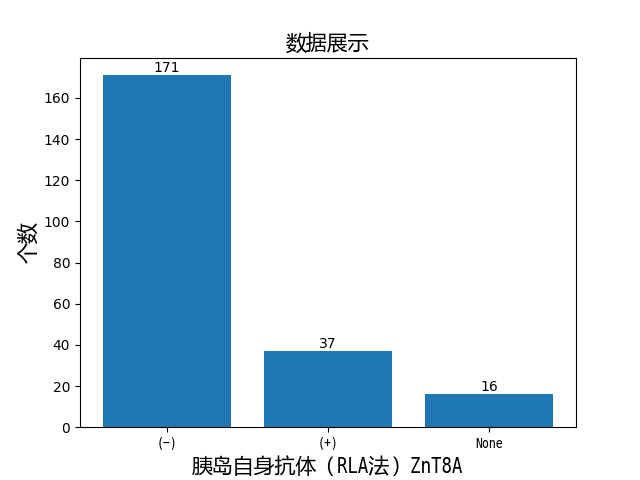
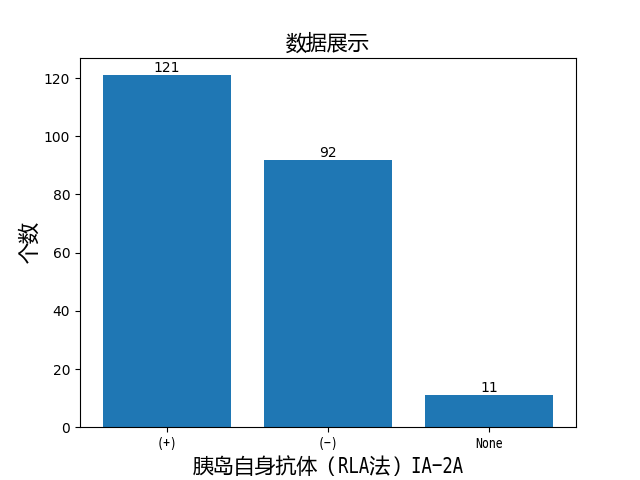
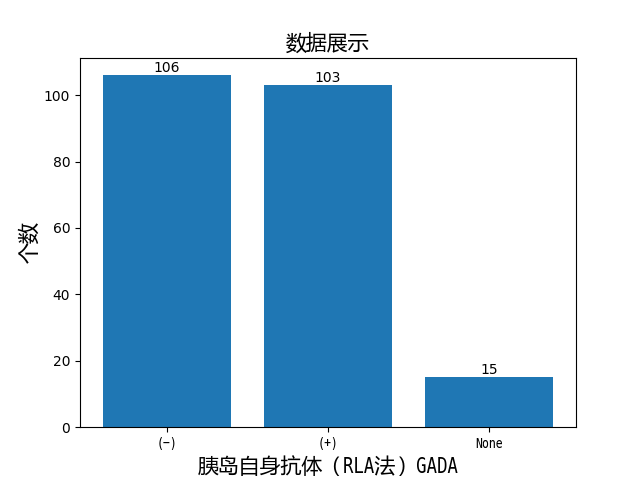
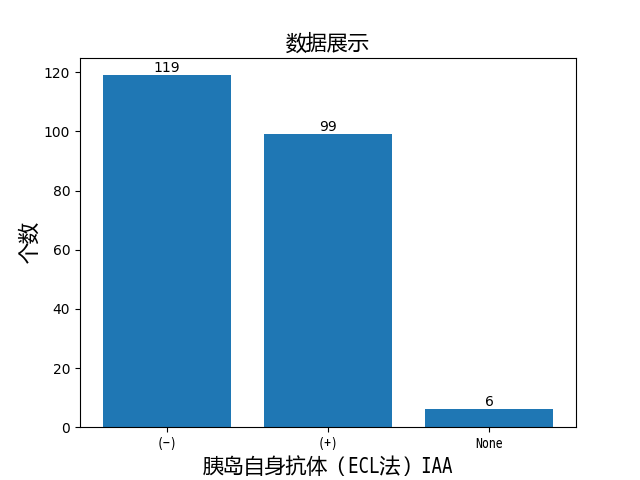
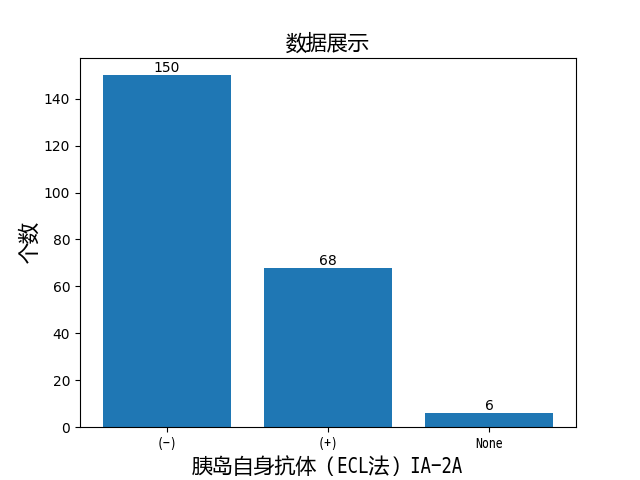
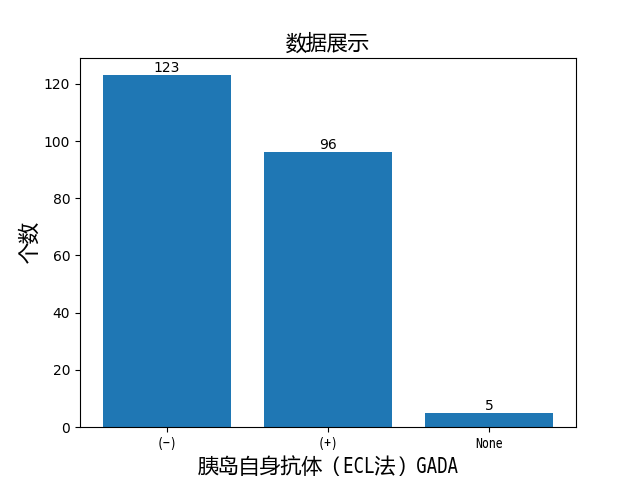
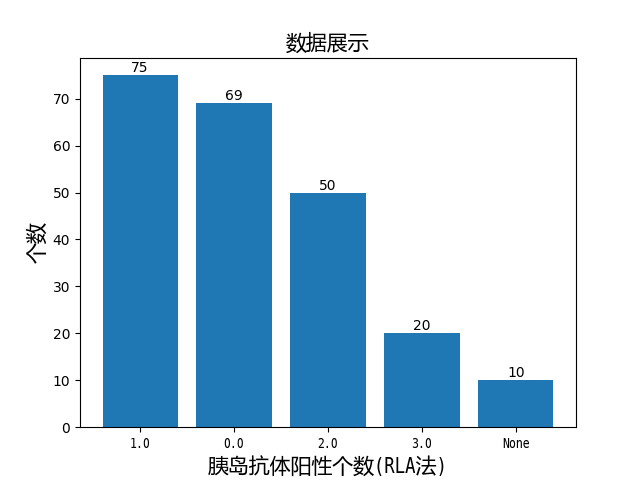
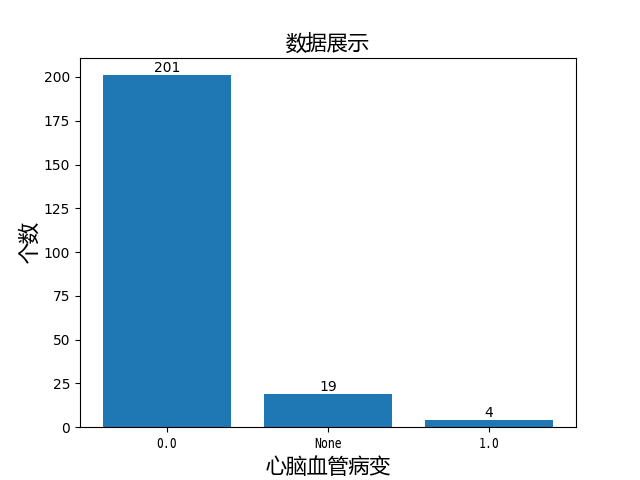
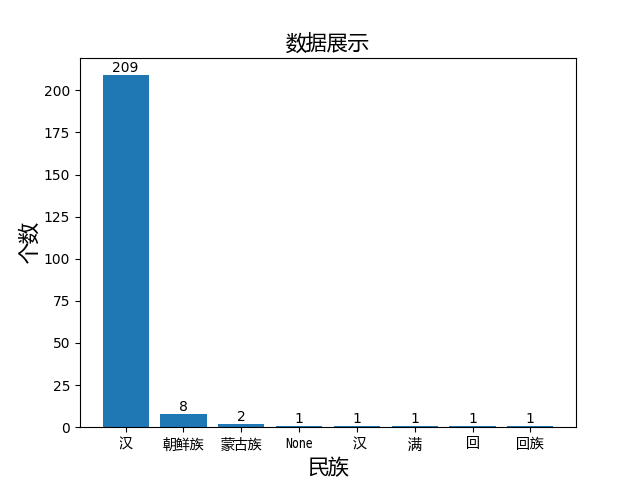
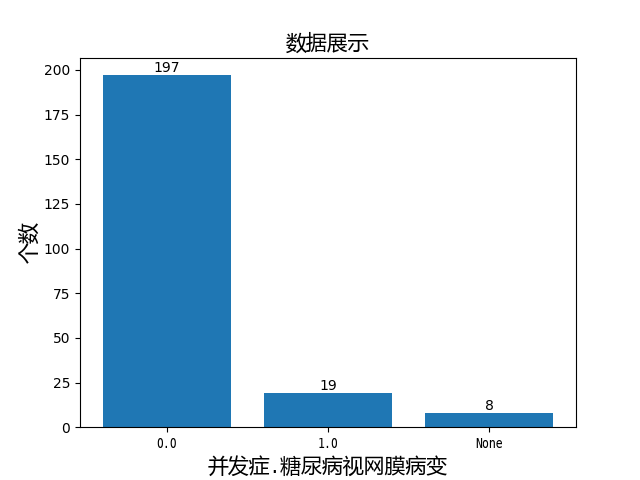
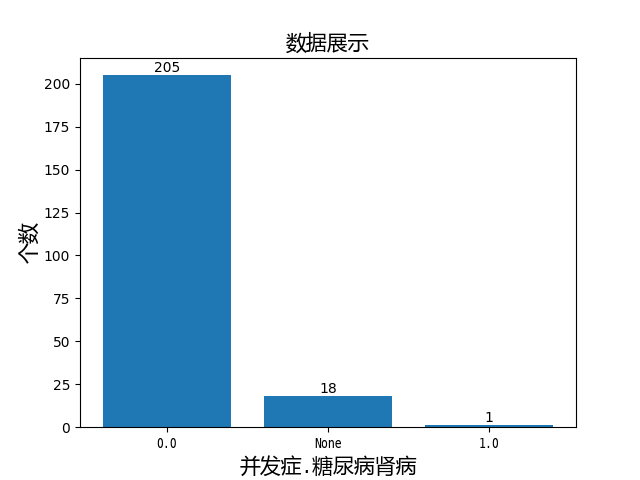
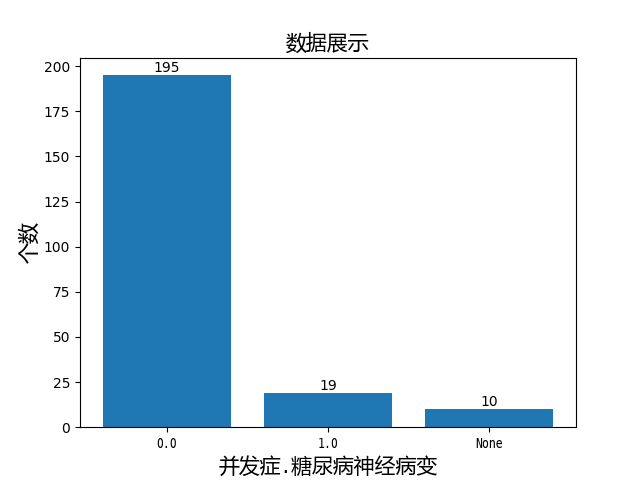
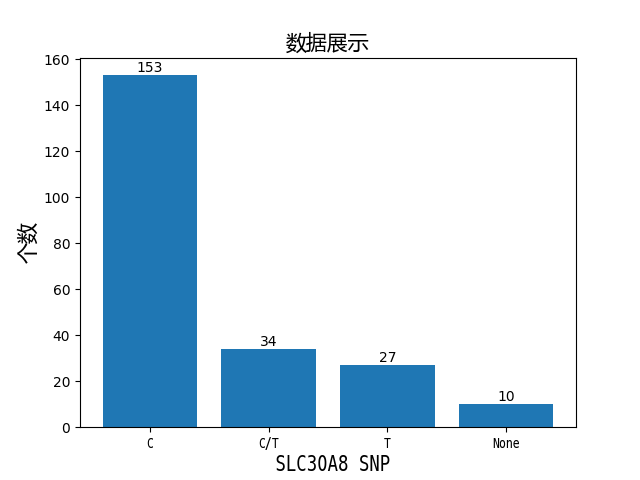
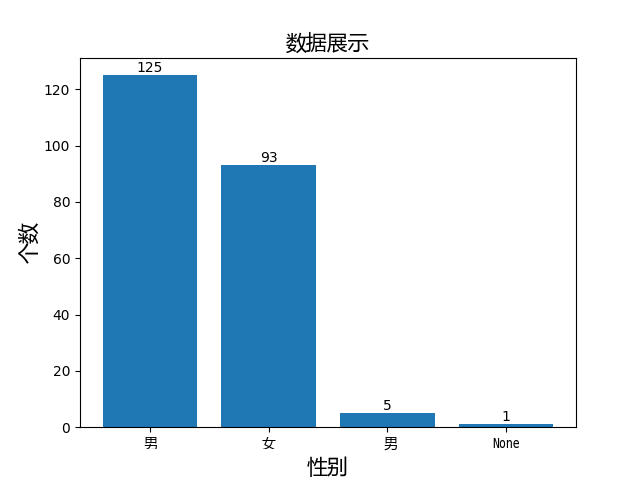
2.1数据概况

样本数：224 维度：45



2.2样本标签分布

2.3类别特征展示



2.4数值特征展示





3.数据预处理

3.1特征分离

将数据分为数值特征，分类特征，文本特征分别进行处理

3.2去除部分特征

去除缺失值大于90%

3.3异常值处理

对于数值型特征将明显偏离范围的数据进行修改

对分类特征的值域进行归一

3.4One-hot编码

对分类型进行编码，编码前将空值设为分类将少的一方

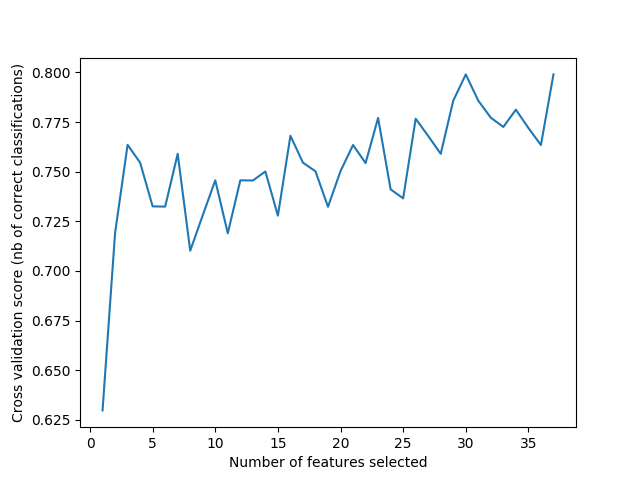
3.5填充空值

数值型使用平均值进行填充，文本型和基因型数据填充null字符

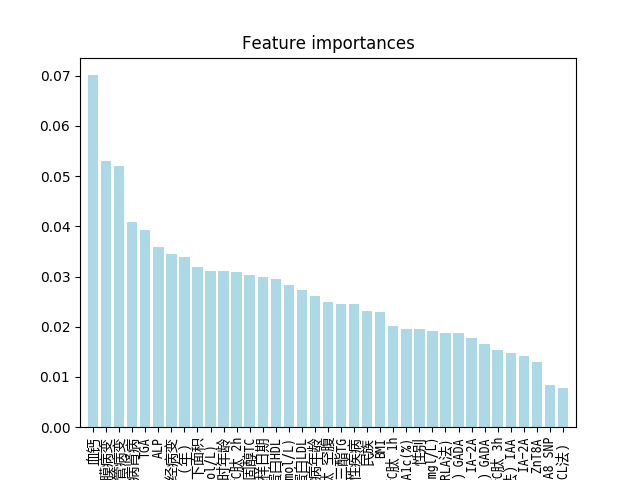
3.6特征选择

分别使用sklearn库中的RFECV和selectmodel进行特征选择。最终选择出来的特征数如下：

递归特征消除：30



模型选择： 17



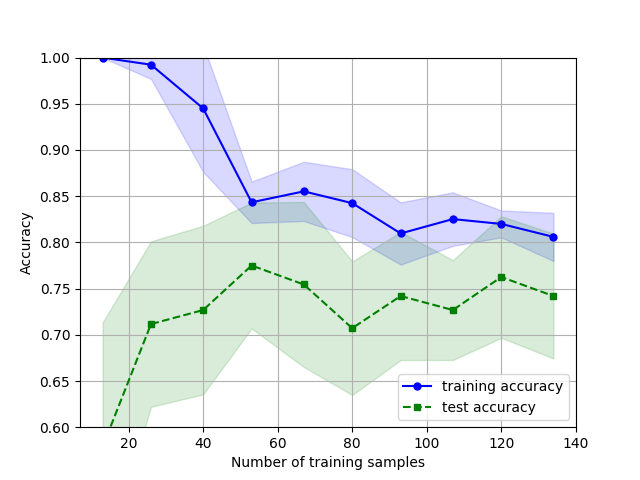
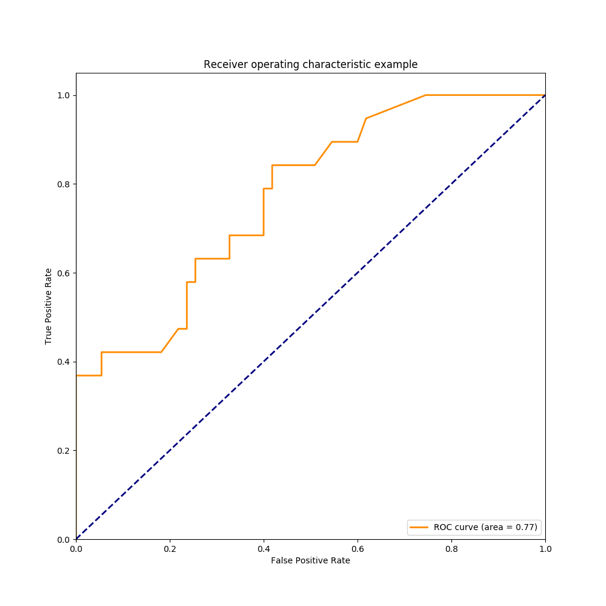
4.模型构建

选择accuracy作为评价指标（输出评价指标说明）

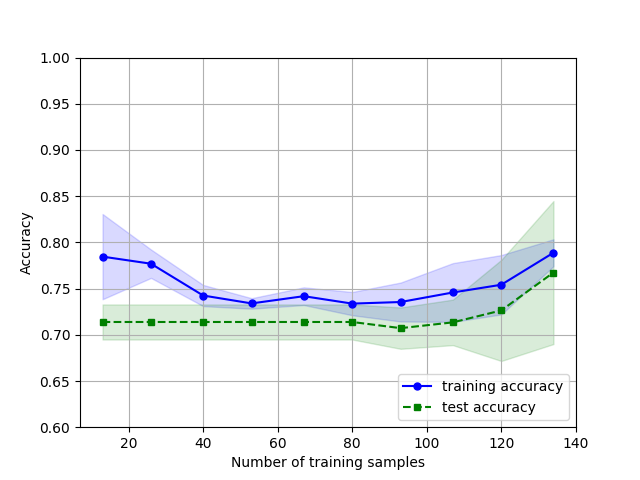
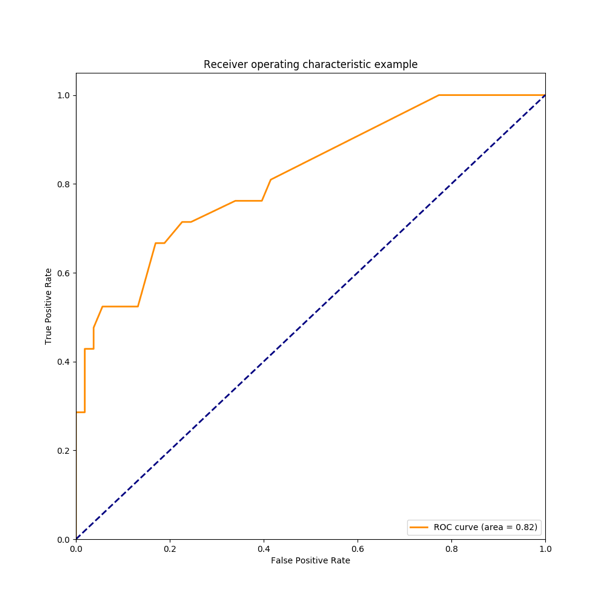
GBDT,XGBOOST,RF



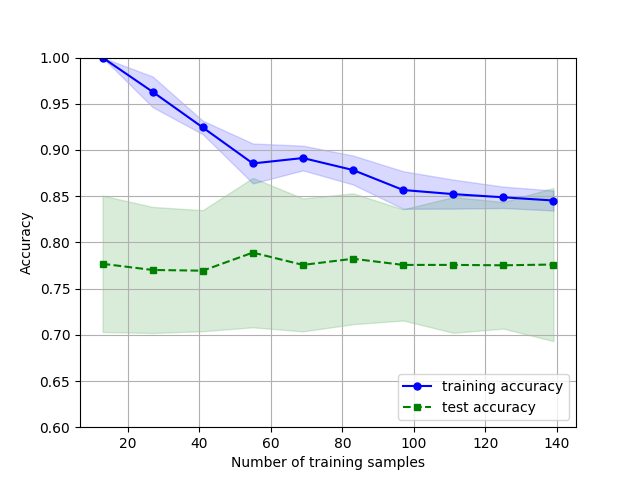
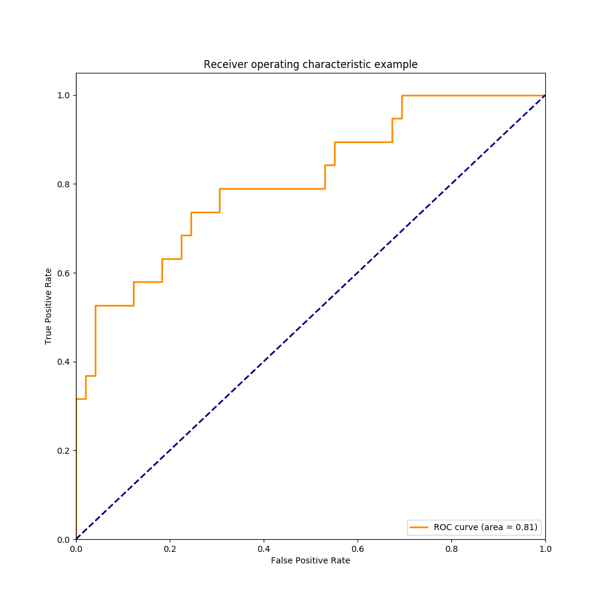
4.1GBDT（一句话解释）



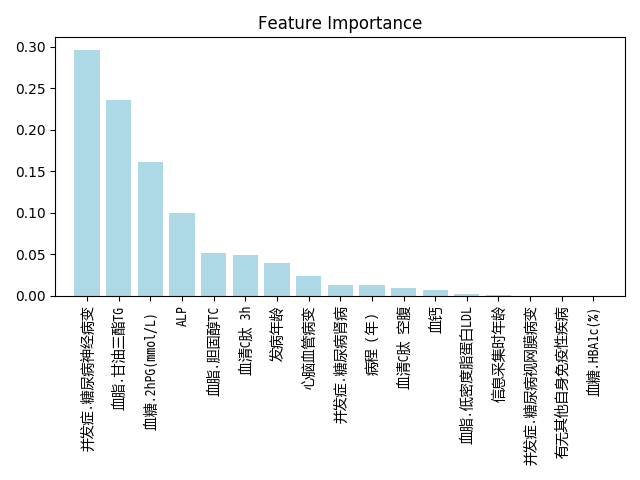
4.2XGBOOST

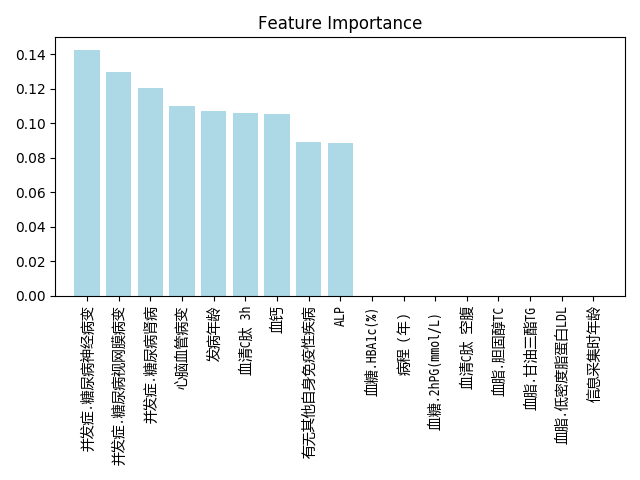


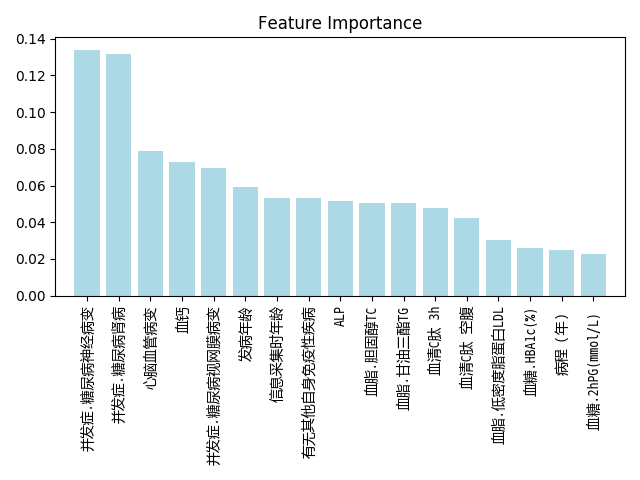
4.3RF



5.特征重要度（特征重要度展示）







6.总结

6.1特征重要性总结

具有糖尿病其他并发症的患者更易患酮症

血糖血脂异常更易患酮症

6.2改进方案

增加样本数量

增加患者其他信息作为新的特征

对文本数据进行结构化处理

补充缺失数据