产后糖尿病预测

**一、数据预处理**

1、先手动筛选一些不予考虑的变量（比如明显不相关或者有效样本数过少等）

2、缺失值处理：对于定性变量，用众数填补；对于定量变量，用随机森林回归填补

**二、特征选择**

1、首先将两个snp以及GDM指标归入模型。

2、剩下的特征利用随机森林法中的特征重要度这一指标进行筛选，共选出如下排名前七的特征：

10min

OGTT血糖曲线下面积

2h血糖

胰岛素分泌指数HOMA-β

HOMA-β/HOMA-IR

0h血糖

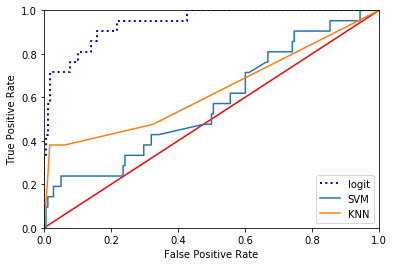
诊断时增加BMI（即OGTT时BMI-孕前BMI）

3、整合1和2的所有特征，共十个

**三、建模预测**

仍采用五折检验来构建模型并进行预测，结果如下

|  |  |  |  |
| --- | --- | --- | --- |
|  | Knn | SVM | 逻辑斯蒂 |
| auc | 0.6310 | 0.5732 | 0.9419 |
| accuracy | 0.9195 | 0.8944 | 0.9447 |
| f1-score | 0.5 | 0.0 | 0.6857 |



ROC曲线