产后血糖研究

一、首先研究rs2297508突变情况、rs11868035突变情况、GDM对产后血糖的影响

1、控制基因rs2297508突变情况和GDM

|  |  |  |
| --- | --- | --- |
| 研究对象 | 逻辑回归 OR | 卡方检验 pvalue |
| 突变+GDM对无突变+GDM | 2.670642209 | 0.00973125 |
| 突变+无GDM对无突变+无GDM | 0.42970095 | 0.604858626 |
| GDM+突变对无GDM+突变 | 4.752024388 | 0.00013828 |
| GDM+无突变对无GDM+无突变 | 1.859378744 | 0.047570197 |
| 突变+GDM对无突变+无GDM | 5.713234227 | 5.43E-06 |

2、控制rs11868035突变情况和GDM

|  |  |  |
| --- | --- | --- |
| 研究对象 | 逻辑回归 OR | 卡方检验 pvalue |
| 突变+GDM对无突变+GDM | 3.165240404 | 0.006189882 |
| 突变+无GDM对无突变+无GDM | 0.454205749 | 0.648258926 |
| GDM+突变对无GDM+突变 | 4.870645041 | 0.000149919 |
| GDM+无突变对无GDM+无突变 | 2.004756352 | 0.029466901 |
| 突变+GDM对无突变+无GDM | 6.415561272 | 1.29E-06 |

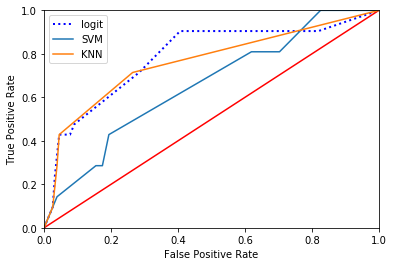
3、患病与不患病人群的突变和GDM水平是否显著不同

|  |  |
| --- | --- |
|  | t-test pvalue |
| GDM | 1.16E-05 |
| rs2297508 | 0.062984487 |
| rs11868035 | 0.063707225 |

二、仅用rs2297508突变情况、rs11868035突变情况、GDM对产后血糖情况进行建模预测

采用五折检验来构建模型并进行预测，结果如下：

|  |  |  |  |
| --- | --- | --- | --- |
|  | Knn | SVM | 逻辑回归 |
| accuracy | 0.8693 | 0.8409 | 0.8806 |
| auc | 0.7617 | 0.6566 | 0.7863 |



ROC 曲线

三、使用产前所有的调查数据对产后血糖情况进行建模预测

1、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

2、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，并检验患病与不患病人群的这些特征是否显著不同：

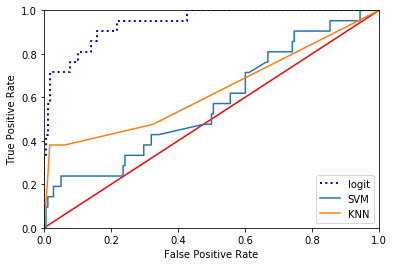
|  |  |
| --- | --- |
| Features | t-test pvalue |
| 10min | 5.289537790901964e-25 |
| 0h血糖 | 5.199452696712105e-10 |
| OGTT血糖曲线下面积 | 1.2465419885015587e-09 |
| HOMA-β/HOMA-IR | 0.07579129294738247 |
| 胰岛素分泌指数HOMA-β | 0.06719931289752244 |
| 2h血糖 | 9.981006191982438e-09 |
| 1h血糖 | 5.104999071608184e-07 |

3）整合1和2的所有特征，共十个

3、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 | 逻辑回归 |
| 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