# SIC项目交接文档

**目的：通过联合多种机器学习模型来预测患者（早期）是否发展DIC，并进行特征筛选**

GitHub地址：<https://github.com/ChinaJianNengCloud/SIC-DIC>

项目背景：

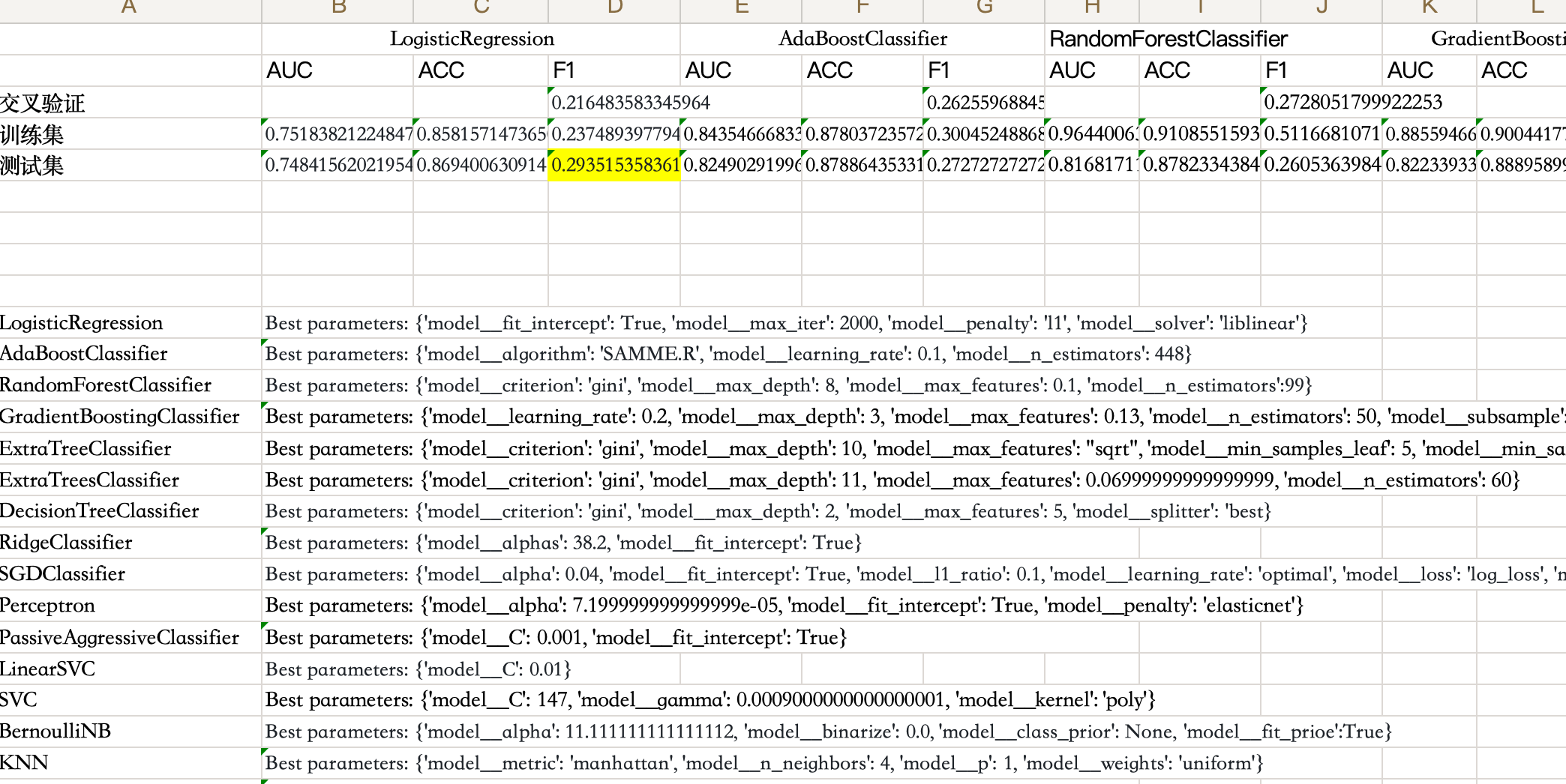
DIC（弥散性血管内凝血）：是一种以全身性凝血系统激活为特征的病理状态，表现为广泛的微血栓形成、凝血因子大量消耗以及继发性纤溶亢进，常导致全身出血、微循环衰竭和多器官功能障碍。

SIC（脓毒症性凝血病）：是脓毒症引起的血管内皮细胞损伤和凝血紊乱，表现为免疫血栓的过度形成和血栓炎症的失控，是脓毒症相关DIC的早期阶段

**脓毒症可能进展为 SIC，SIC 可能进展为 DIC**

目前已经完成的步骤：数据预处理（缺失值填补、标准化），利用网格搜索确定最佳超参数，输出在训练集和测试集上的性能，结果见表格。

**现有结果：多种机器学习模型在最佳超参数下，在训练集和测试集上的表现**，详见表格



为什么进展比较缓慢？

1. 项目思路比较混乱，**脓毒症可能进展为 SIC，SIC 可能进展为 DIC，**所以建模方向其实大致有两个，一个是基于脓毒症患者（可能有SIC，可能无）的数据来预测DIC，另一个是基于SIC患者的数据来预测DIC，代码类似，但是临床意义有所不同，但是一开始并没有下定论说做哪个，**两种都做了，数据结果都有**（分别保存在data和data/脓毒症患者）**，但是目前主要是在做SIC-->DIC的思路**
2. 模型性能较差，由于**标签严重的不平衡，导致F1很低**，大约0.3，即使已经使用SMOTENC法进行过采样，仍然不理想，**后续要输出混淆矩阵情况进一步查看**。还有就是与DIC预测最相关的几个指标，**比如D二聚体并没有被纳入**，因为海大蓉说缺失太多（一万多人里面只有五百多人有），直接删掉了，这可能也是效果不理想的原因，**目前海大蓉正在整理这五百多人的数据，后续可供使用，混淆矩阵输出和海大荣新数据的处理应该是下一步要做的事情**

最后，海大蓉新给的数据里面可能存在入院时间、死亡情况这种和建模无关的变量，需要留意一下。超参数搜索的过程中我有使用过采样和不使用过采样两个版本的代码，不过目前我是倾向于过采样（SMOTENC），所以另一个版本代码可以直接忽略

最后再讲一下GitHub上文件的存放