1. 流程概述

本项目旨在对60岁及以上的中国社区老人是否抑郁进行预测。目前现有的全年龄样本量又20948个，过滤之后60岁及以上的样本量有一万多个。我们基于这一万多个的样本构建数据集，同时根据CES-D表的得分来区分样本抑郁与否（总分30分，大于等于12分为抑郁，标签为1，小于12分为非抑郁，标签为0）。

我们拟采用模型融合（Model Stacking）的集成方法来训练数据。模型融合的优势在于能集成不同模型对数据集内在关系的抓取，以期达到优于单一模型表现的效果。模型融合方法的基本思想是同时训练多个算法原理不尽相同的机器学习算法（诸如随机森林,支持向量机等），并在此基础上训练一个元模型来组合它们，然后基于这些模型返回的多个预测结果来输出最终的预测结果。同时，我们在融合算法的构建过程中采用K折交叉训练的方式来降低过拟合的可能性。

用训练好的模型对2015年的样本数据进行预测两年内的抑郁发病情况，并通过计算模型的曲线下面积（AUC）,准确率（Accuracy）,精确率（Precision）,灵敏度（Sensitivity）来综合评估模型表现。（具体定义及计算见）