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）。

我们对数据集分别训练不同的机器学习分类模型（诸如逻辑回归，随机森林，支持向量机等），并且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）等指标来评估各个模型的表现。在此基础上，我们列出表现较好的前5个模型，采用五折交叉训练的方式来构建模型融合（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，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。最后，我们将模型融合算法与选出的5个表现较好的模型放在一起再做一轮评估来找出最优模型。

用选出的模型对2015年的样本数据进行预测两年内的抑郁发病情况，得到发生抑郁的风险预测概率。

另外，在我们训练基于树模型（tree-based model）的机器学习算法时（诸如随机森林，GBDT等），通过计算基尼重要性来排列出具有较高重要性的多个特征变量，在此基础上，我们通过多次迭代将排名靠前的特征变量从原始数据集中筛选出来重新带入我们选出的最优模型框架，并将训练结果与原模型进行比较，在曲线下面积（AUC），准确率(Accuracy),精确率(Precision),灵敏度(Sensitivity)等指标相差不大的前提下得出最优的特征选择组合，并以此来知道我们在设计问卷过程中需要注意的重要问题。

1. 数据清洗与预处理

具体包括：

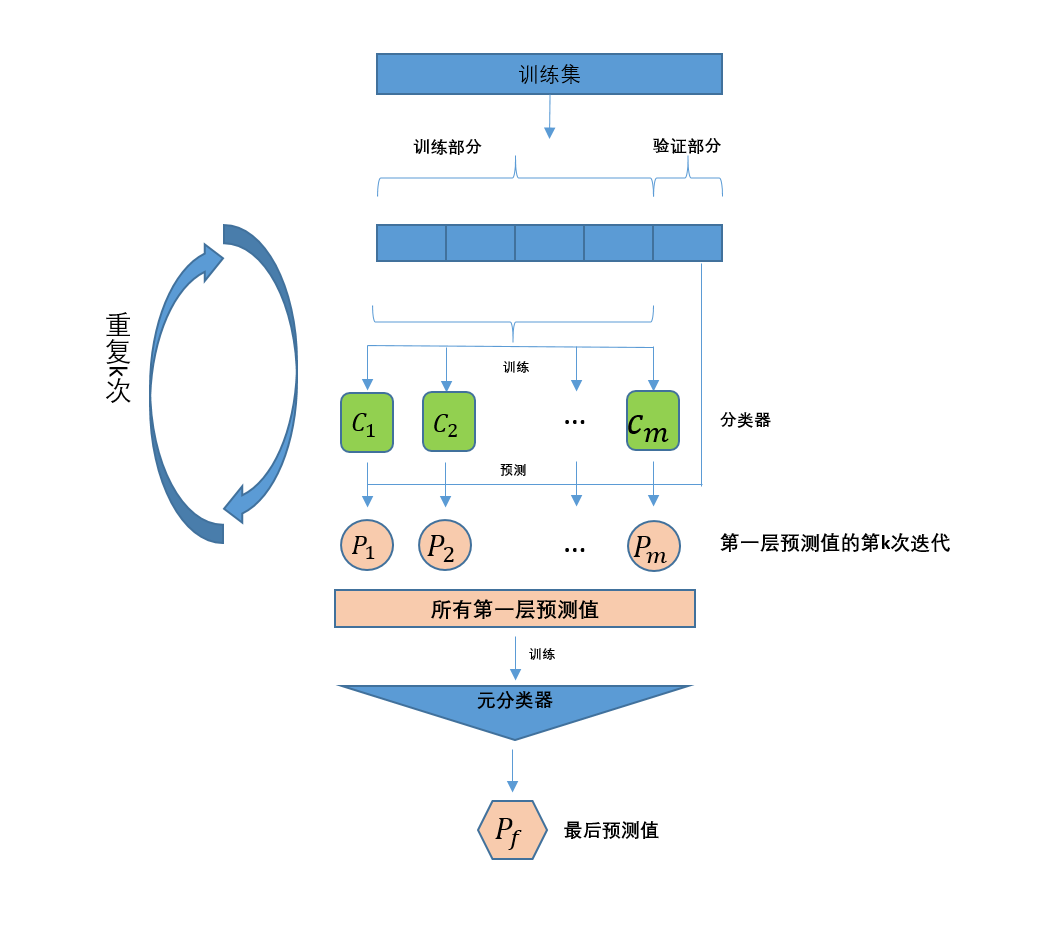
* 去除重复样本
* 去除缺失率较高的特征变量
* 特征变量的逻辑筛选（如身高5米）
* 独立热编码（one-hot encoding），用于处理类别变量
* 数据标准化，无量纲化且加快算法收敛速度

其中在缺失值填充阶段，我们考虑采用K-近邻算法（KNN）进行缺失值填充，它通过计算 高维空间上样本之间的距离来识别相邻点，并利用相邻观测值的完整值来估计缺失值（Berretta and Santaniello,2016）。KNN的优点有：a)插补值是数据集中实际出现过的值，并非是二次计算得来的。b)根据变量信息保留了原始数据的分布结构。c)KNN插补完全是非参数的。

1. 类别不平衡问题的处理

对于在实际建模过程中存在的目标变量类别不平衡问题（既样本中非抑郁人群数量远超抑郁人群数量，导致二个类别比例极不均衡）,我们拟采用合成少数类过采样技术（Synthetic Minority Over-sampling Technique）来扩充少数类样本。它是基于随机过采样算法的一种改进方案，其基本思想是对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中来达到均衡样本类别的目的。

1. 模型融合算法原理示意图



1. 模型评估指标说明