同济大学 TONGJI UNIVERSITY

面向磁浮轨道异常检测的大数据分析框架研究

小组汇报

2025-3-14　　By 刘震

利用 GridSearch 找出模型参数最优值

Grid Search 是一种超参数优化方法，通过遍历所有可能的超参数组合，寻找最优模型配置

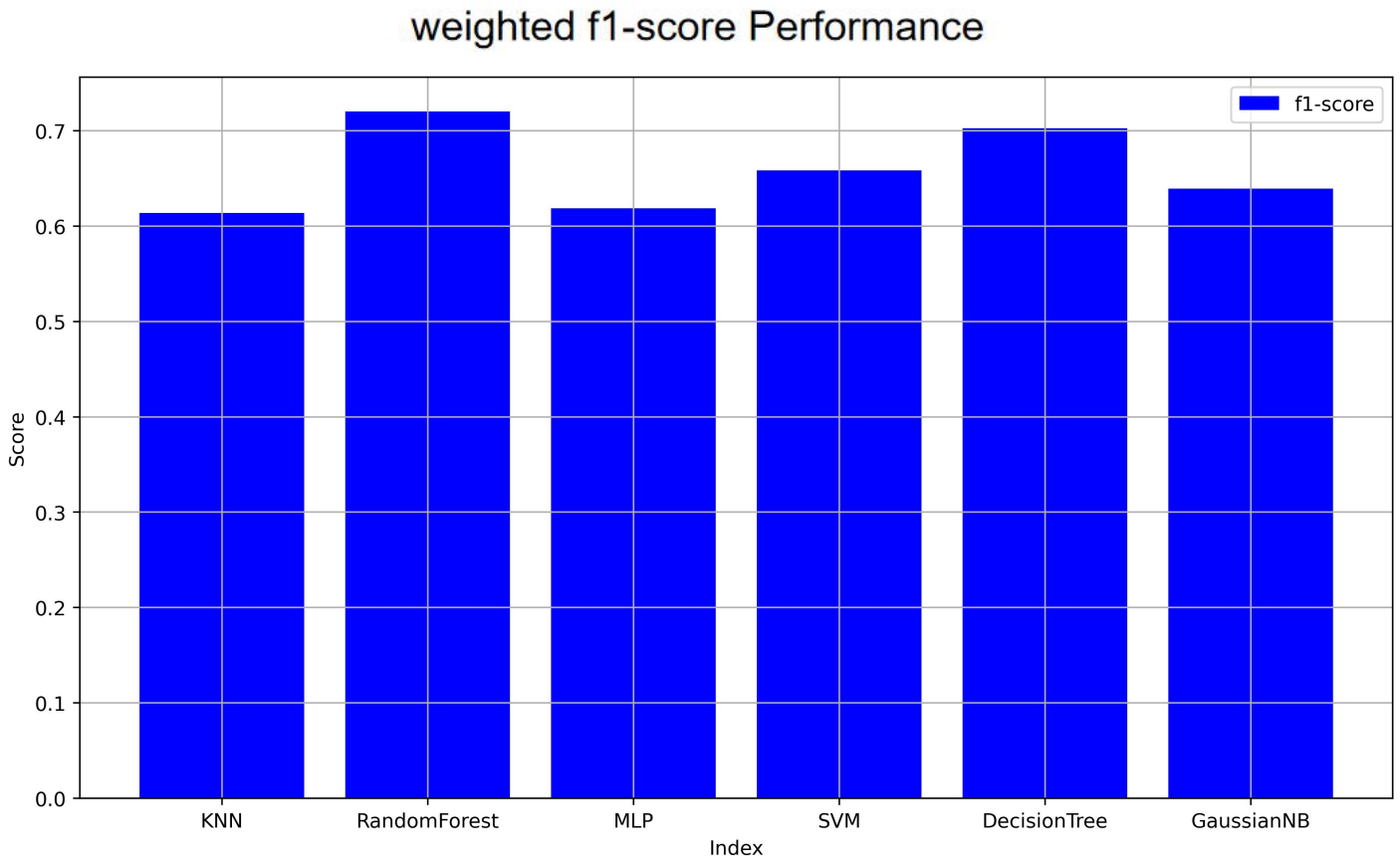
给定优化目标 (如 macro-f1 score, weighted f1-score, recall), 可以得到相应模型的最优参数

下面展示了将优化目标设定为 macro f1-score 的时候，各类参数的优化情况：

model	best_macro_f1_score	best_params
KNN	0.542143439	{'metric': 'euclidean', 'n_neighbors': 15, 'weights': 'distance', 'smote__k_neighbors': 3}
RandomForest	0.631262333	{'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 300, 'smote__k_neighbors': 3}
MLP	0.53567211	{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (100,), 'learning_rate_init': 0.0001, 'smote__sampling_strategy': 'auto'}
SVM	0.562062487	{'C': 1, 'gamma': 'auto', 'kernel': 'rbf', 'smote__k_neighbors': 3}
DecisionTree	0.600285471	{'criterion': 'entropy', 'max_depth': 10, 'min_impurity_decrease': 0.0, 'smote__k_neighbors': 5}
GaussianNB	0.532100485	{'var_smoothing': 1e-9, 'smote__k_neighbors': 3}

在尝试了优化目标分别为 weighted f1 score, recall, precision 后，确定了这些模型最终的参数，以及最终的检测结果（具体数值将会在后面展示）

右图展示了在以 weighted f1-score 为优化目标时最终的结果



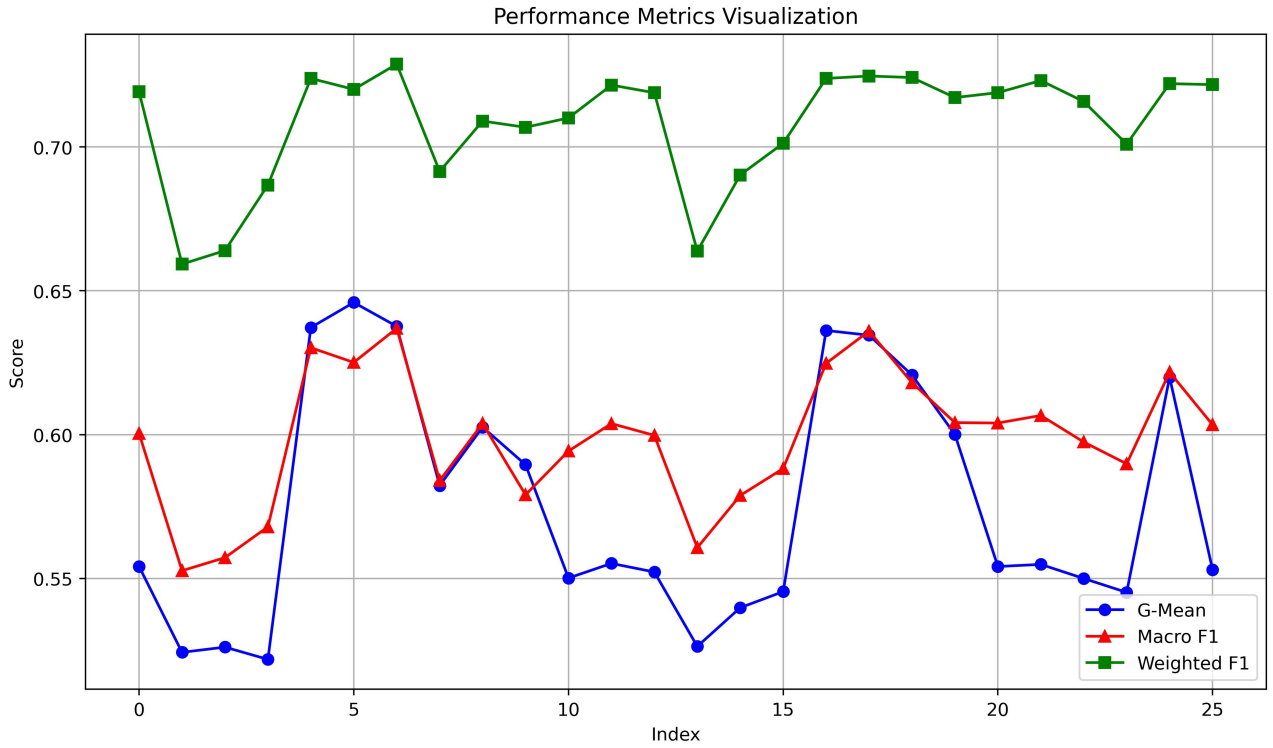
在参数调优后，结合上面的柱状图可以看出，RandomForest 模型胜出
模型表现可以排序为：
RandomForest > DecisionTree > SVM > Naive Bayes > KNN = MLP

Stack: 模型堆叠

模型堆叠是一种集成学习方法，通过结合多个基模型（Base Models）的预测提取特征，再通过一个元模型（Meta Model）进行最终预测。

All Models : [MLP，RF，KNN，SVM，DecisionTree]
Base Models ∈ All Models
Meta Model : Logistic Regression

一共遍历了所有可能的 26 种 Base Models 的组合：



表现较好的组合有：
MLP + KNN,
MLP + KNN + DecisionTree,
MLP + KNN + DecisionTree + SVM
MLP + DecisionTree + SVM
MLP + SVM

这个结果说明在模型堆叠中，MLP 具有较大的作用

下图展示了各种组合中明显异常类的 F1-score

其中效果较好的 Base models 是:
MLP + DecisionTree + SVM
MLP + KNN + DecisionTree + SVM

对比可知，利用模型堆叠，可以提高对明显异常类的检测效果

在单一算法模型中不如 RF 算法优秀的 MLP, KNN, DecisionTree, 在模型堆叠后，表现出比RF更好的效果

Compare : RF算法单一模型检测中，明显异常类的 F1-score 为 **0.4519**
小于上图 F1-score 的峰值

基于多模型投票机制的异常检测

All Models : [MLP，RF，KNN，SVM，DecisionTree，NaiveBayes]
Vote Models ∈ All Models

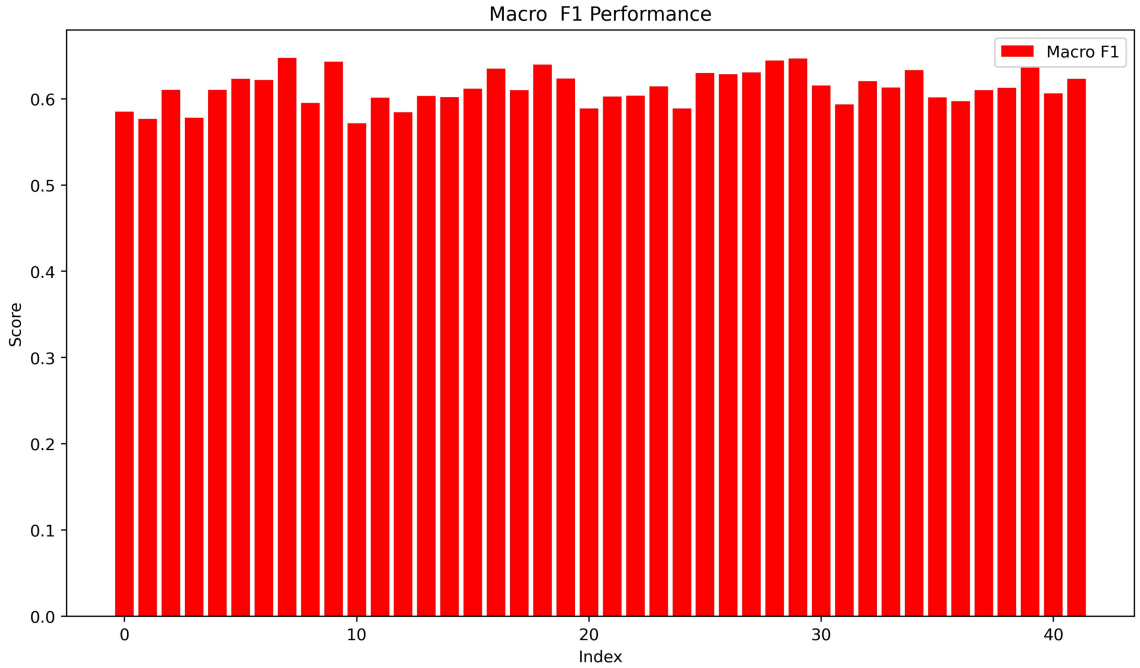
为了找出最佳的用于投票的模型组合

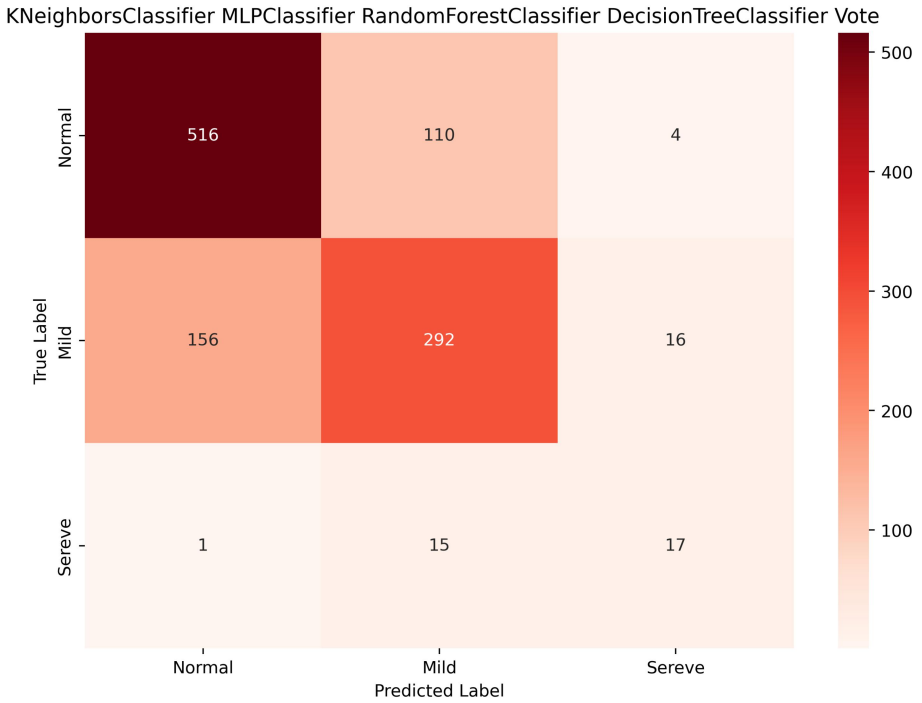
可以遍历 All Models 的所有子集，最终的结果如右图所示

（为了处理票数相等的情况，给不同的类赋予相应的权重，将其权重相加，然后再按总分数进行分类）

最佳组合为： **KNN + MLP + RF + DecisionTree**

对应的 confusion matrix 如右图所示





项目总结部分

1. 数据预处理

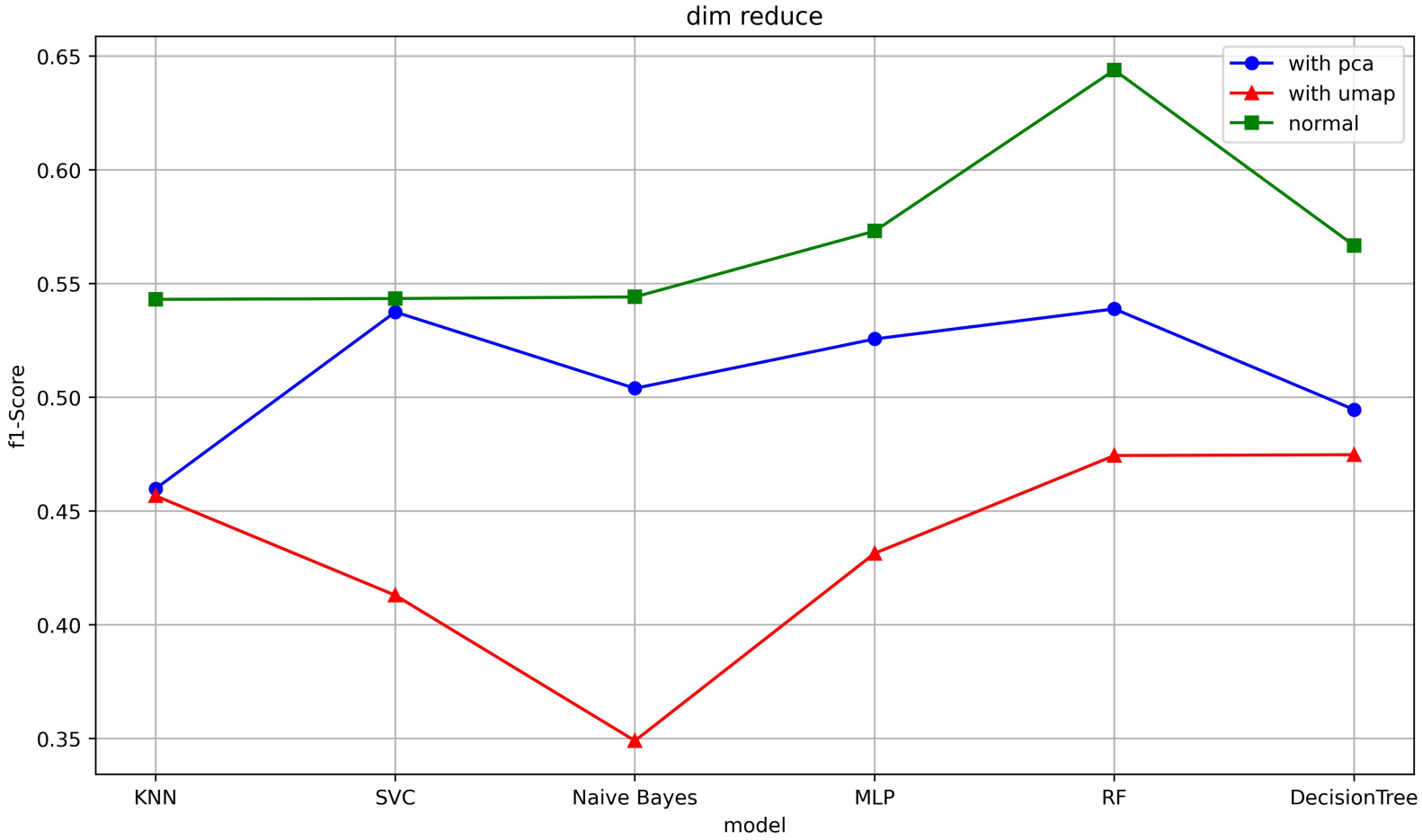
2. 降维（UMAP, PCA, T-SNE）

3. 利用各种有监督方法训练（RF, KNN, MLP, SVM, DecisionTree, NaiveBayes）
期间也尝试了部分无监督方法（如 GAN）

4. 结果评估

降维的局限：
对于有监督方法，对部分模型可能具有有效性，但是大部分情况下会使效果更差

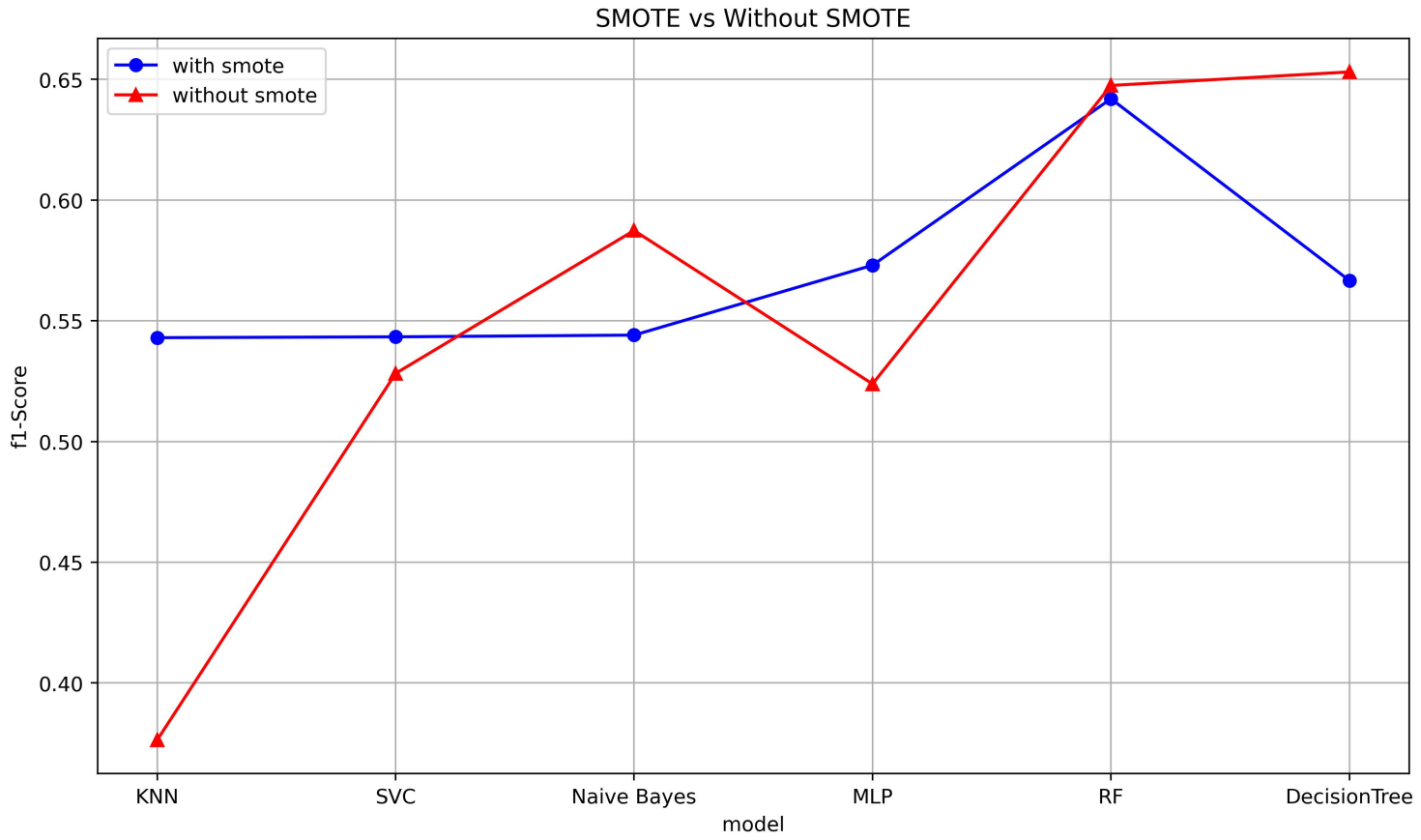
右图分别展现了：
不降维、pca降维至 7 维、umap降维至 4 维 后在各个模型上的表现
明显可以看出的是，不降维的效果是要明显优于其他两种的



SMOTE 策略对某些算法的必要性

本数据集标签具有不平衡性，而在有监督训练的很多算法中，若标签不平衡，会导致模型对少数类的忽视、对多数类过于重视等
SMOTE 方法可以较好的改善这个问题

如右图所示，SMOTE 策略对 **KNN、SVM、MLP** 算法的检测均有帮助



Final Result

单模型检测（前面的Grid Search 中得到的最佳参数下）：

Algorithm	weighted f1-score	macro f1-score
RF	0.7296	0.6438
KNN	0.6279	0.5430
MLP	0.6817	0.5660
DecisionTree	0.7014	0.5667
SVM	0.6705	0.5434
Naive Bayes	0.6496	0.5441

以最大化 weighted f1-score 为目标：

配置：RF单一算法，SMOTE + kFold,
n_estimators = 300，
max_depth = 10，
min_samples_split = 2
结果：weighted f1-score = 0.728

以最大化 严重异常类 的 f1-score 为目标：

配置：多模型投票， KNN + MLP + RF + decisionTree
结果：严重异常类的 f1-score = 0.5074