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

2024-10-17　　By 刘震

基本思路：

1.分别采用Random Forest、MLP（Multilayer Perceptron）、KNN（K-Nearest Neighbors）对原始数据集进行有监督训练

2.对数据集进行PCA降维后再训练

3.利用SMOTE扩充异常数据点数量，再进行训练

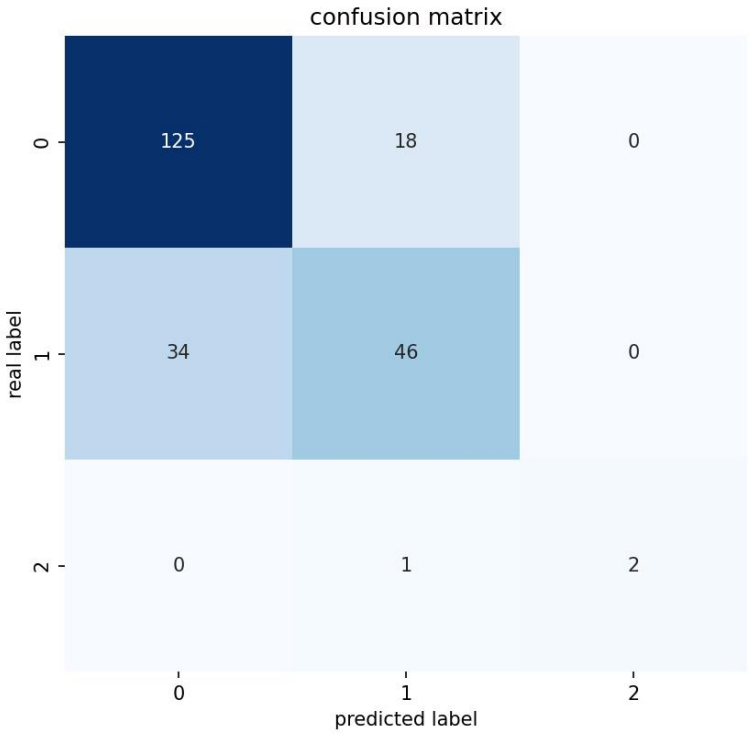
4.使用GAN模型对数据集进行扩充，利用扩充后的数据集进行有监督训练

5.使用堆叠模型（Stacking），将之前的几个预测算法作为基础模型，用它们预测的结果训练更高层次的模型

分别采用Random Forest、MLP、KNN对原始数据集进行有监督训练

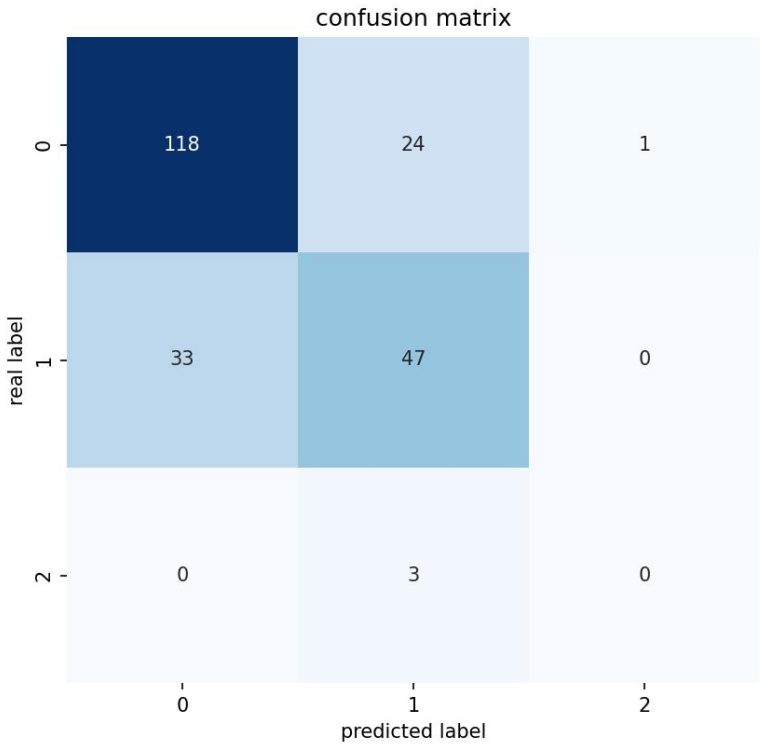
统一设定：训练数据占80% 测试数据占20%，precision、recall、F1score均以加权平均值给出（权为样本数量）

RF：



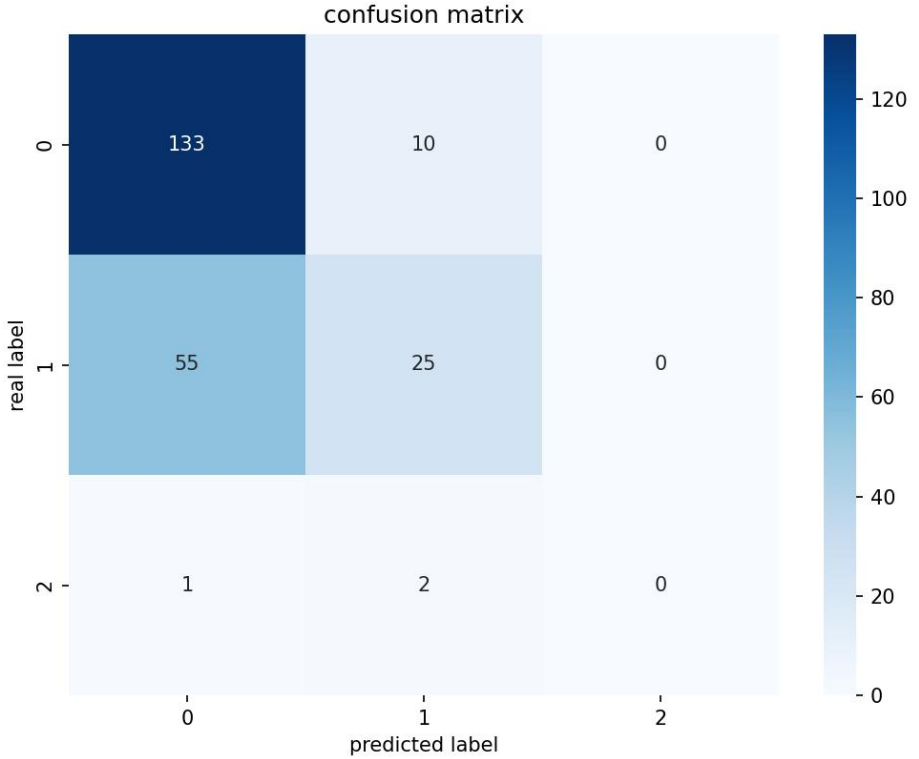
precision: 0.765487
recall: 0.765487
F1 score: 0.759009

MLP：



precision: 0.730088
recall: 0.730088
F1 score: 0.723984

KNN：

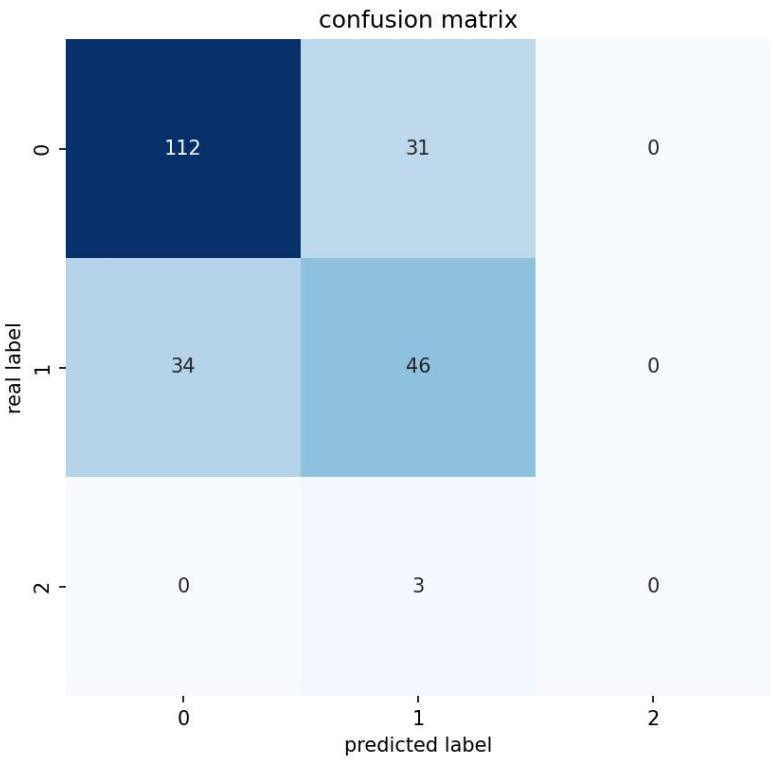


precision: 0.699115
recall: 0.699115
F1 score: 0.658232

对数据集进行PCA降维后再训练

采用降至7维后的数据

RF：

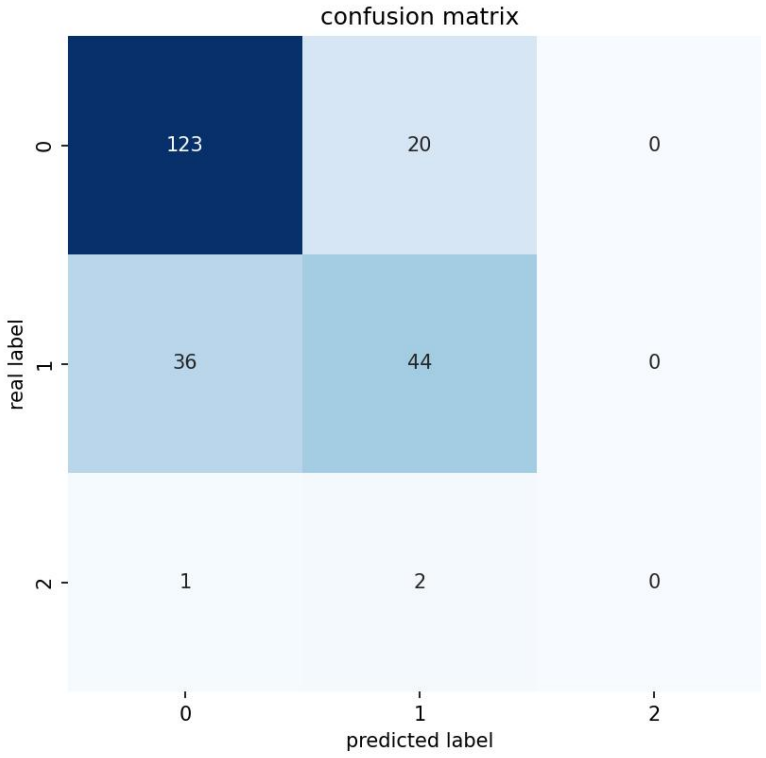


precision: 0.699115

recall: 0.699115

F1 score: 0.693971

MLP：

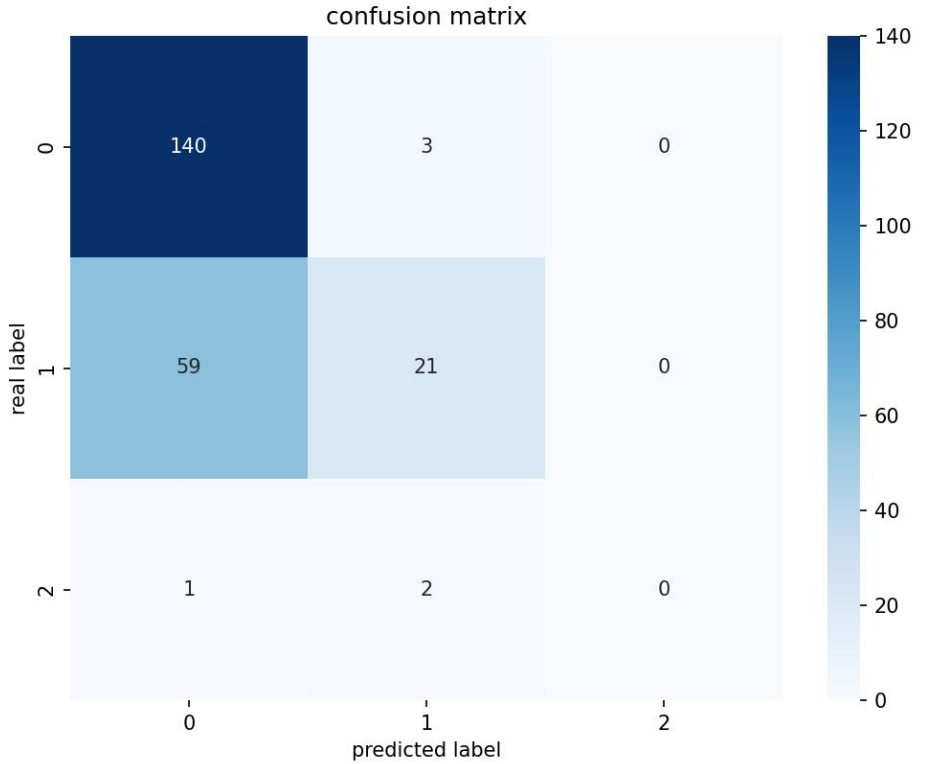


precision: 0.738938

recall: 0.738938

F1 score: 0.727072

KNN：



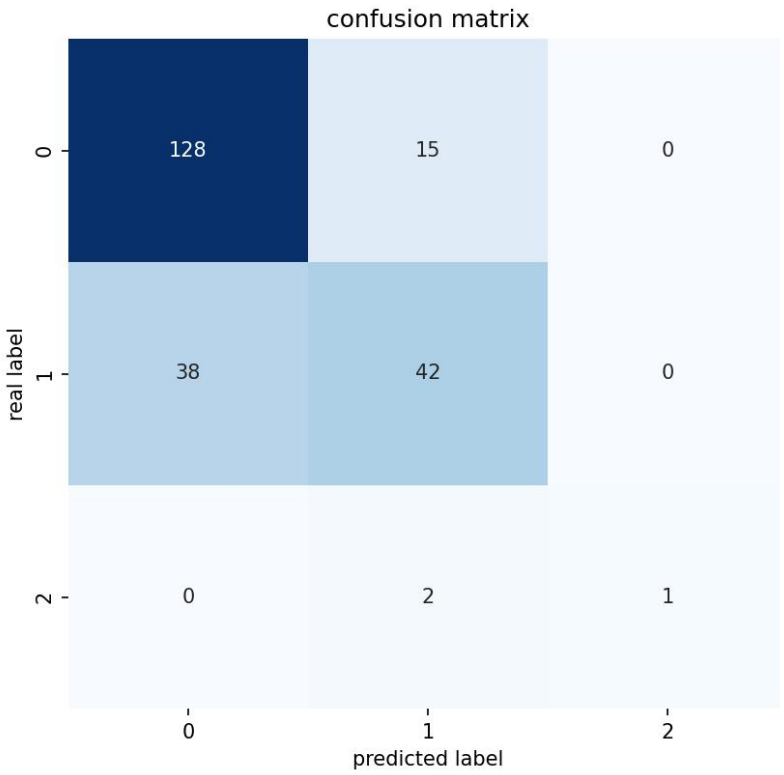
precision: 0.712389

recall: 0.712389

F1 score: 0.656782

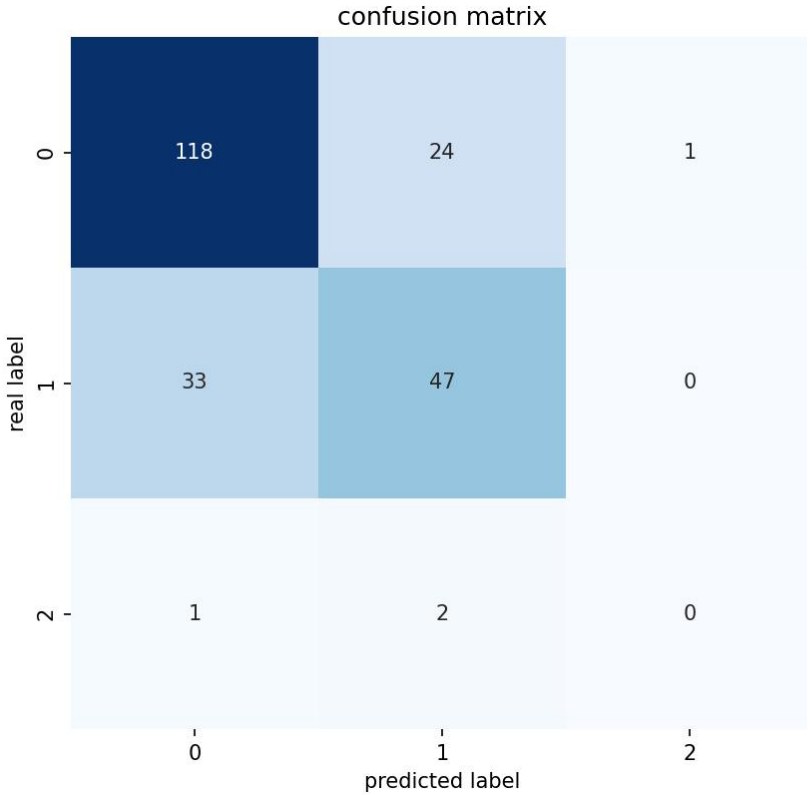
会不会是因为数据集中异常点数据的样本太少了？
改进：利用SMOTE扩充异常数据点数量（通过插值形成新的异常点样本），再进行训练

RF：



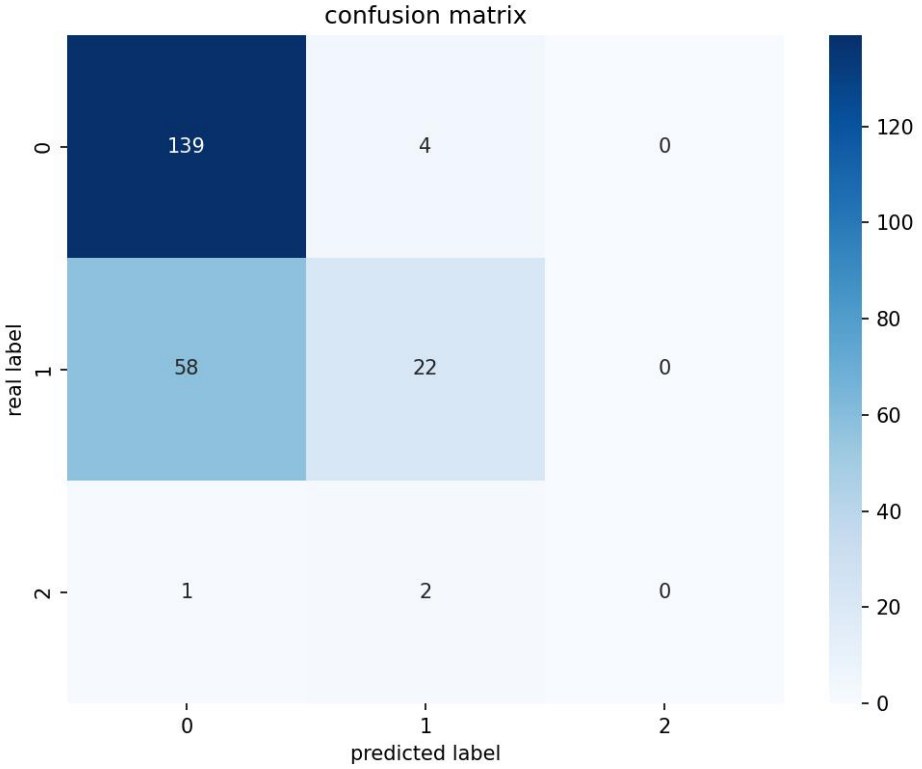
precision: 0.756637
recall: 0.756637
F1 score: 0.744769

MLP：



precision: 0.730088
recall: 0.730088
F1 score: 0.723674

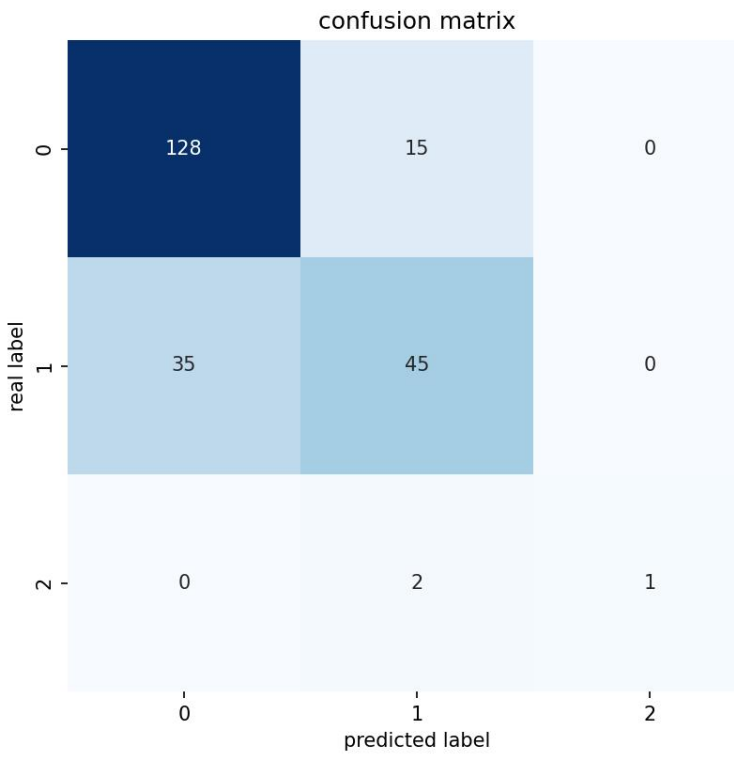
KNN：



precision: 0.712389
recall: 0.712389
F1 score: 0.660059

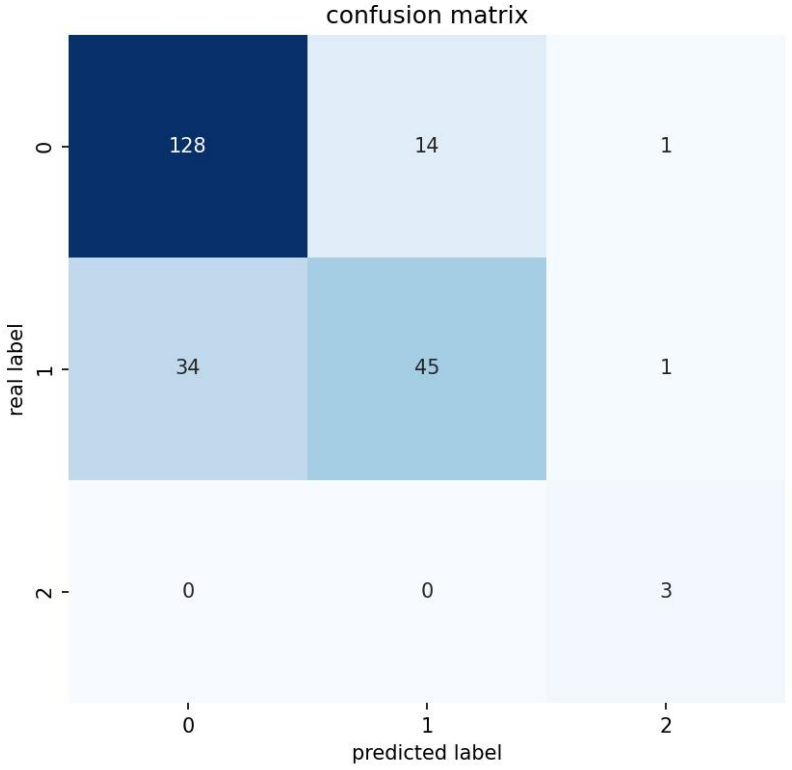
利用GAN模型生成更多的数据，然后再利用有监督训练的模型进行预测（严格分离用于训练的数据和用于测试的数据）

RF：



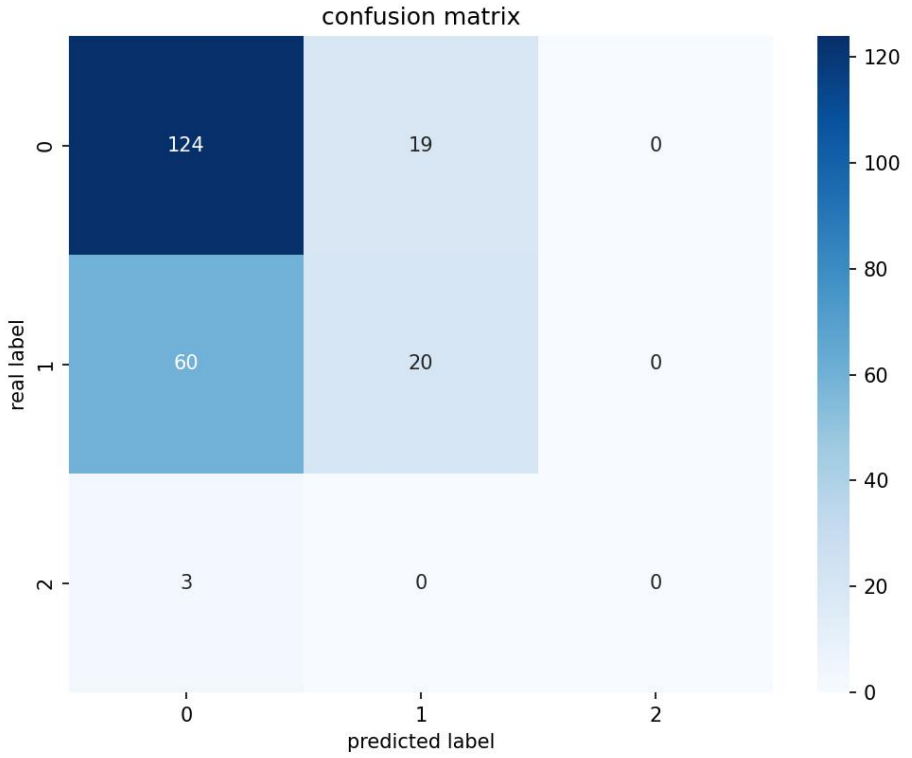

```
precision: 0.769912
recall: 0.769912
F1 score: 0.760346
```

MLP：



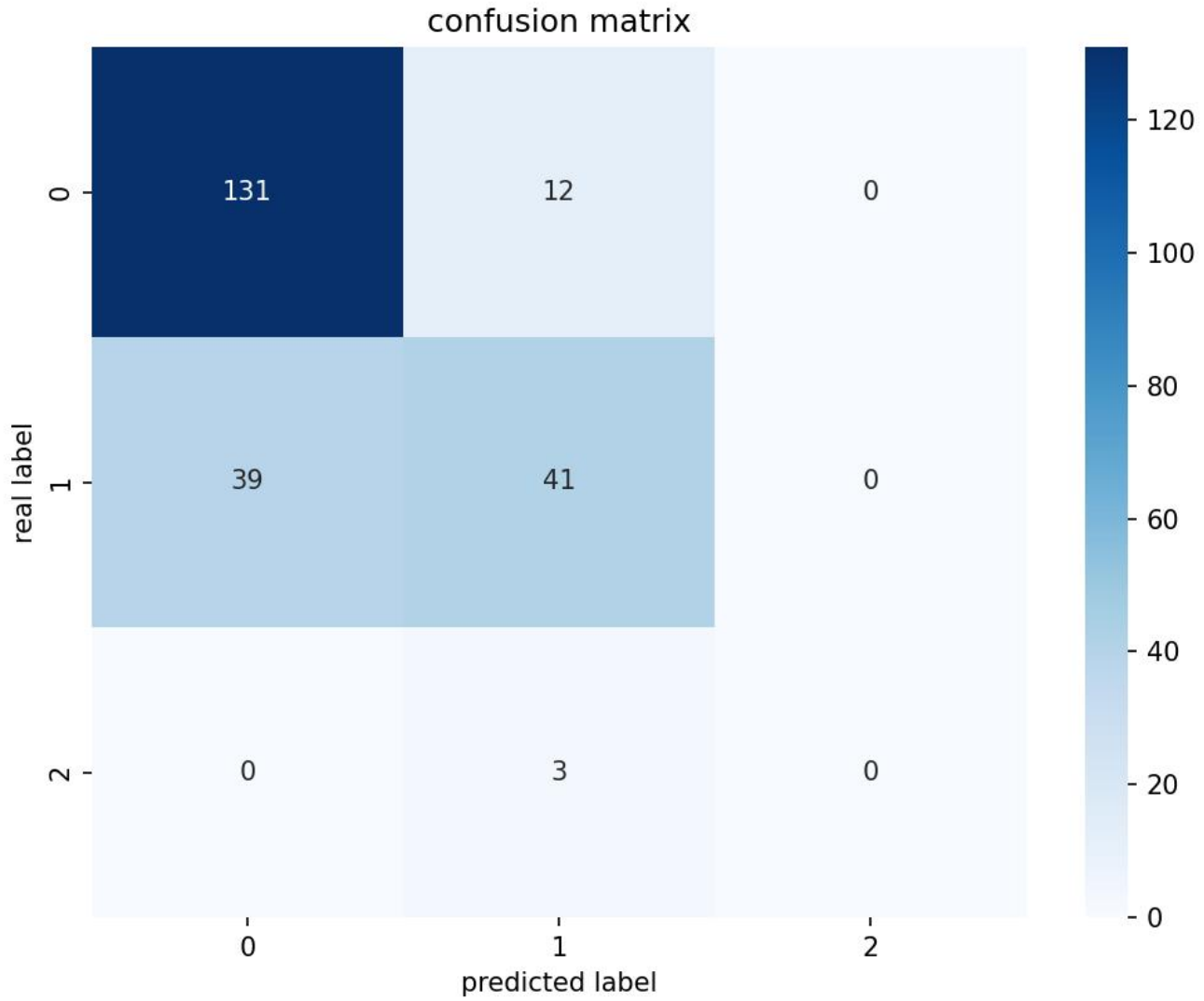

```
precision: 0.778761
recall: 0.778761
F1 score: 0.770242
```

KNN：




```
precision: 0.637168
recall: 0.637168
F1 score: 0.594502
```

使用堆叠模型，将之前的几个预测算法作为基础模型，用它们预测的结果训练更高层次的模型




```
precision: 0.7611
recall: 0.7611
F1 score: 0.7431
```

汇总 （F1score）

	RF	MLP	KNN
无特殊处理	0.7590	0.7240	0.6582
PCA降维后	0.6940	0.7271	0.6568
SMOTE扩充后	0.7448	0.7237	0.6601
GAN扩充后	0.7603	0.7702	0.5945
模型堆叠	0.7431		

PCA降维后，只有MLP算法的性能略微上升；采用SMOTE扩充异常数据点后，KNN算法性能小幅度上升；
采用GAN扩充数据后，RF、MLP算法F1score都得到了提升，但KNN反而下降

平均来看，RF算法较优（F1score较高）
F1score最高的处理方法：利用GAN扩充数据，然后采用MLP有监督训练