Data Science

四資工三甲 B10632024 林奕辰

演算法流程：

1. Pandas讀入csv file
2. dropna()刪除缺漏值的data

(嘗試塞入mean value of column，但結果不佳)

1. 將y\_train從train\_data分離，並將y\_train的text encode，並且轉成DataFrame格式
2. 合併test\_data和train\_data一起處理
3. 將時間欄位刪除
4. 將categorical data(文字)用labelEncoder和oneHotEncoder轉成number
5. 分離test\_data和train\_data
6. MinMaxScaler做正規化
7. (SelectKBest做feature selection，但要配合AdaBoost和outlier remove結果才好)
8. (Density base cluster(DBSCAN)做outlier detection)
9. Downsample majority修正imbalancing data

(嘗試upsample minority，但效果不彰)

1. AdaBoost(n\_estimators=165,criterion="entropy")目前是配合downsample效果最好

(嘗試RandomForest、GradientBoost、Voting，但效果不彰)

1. 輸出output.csv檔案

目前最佳配方：AdaBoost(ne=165) + SelectKBest(10) + downsample

其他較佳配方：Random Forest(ne=150,c="entropy") + downsample

核心突破：downsample with random number 533

如何執行程式：

在cmd上輸入：python classifier.py

即可取得output5.csv

和當前的random值存在random.csv(Downsample majority使用)