Data Science

四資工三甲 B10632024 林奕辰

演算法流程：

1. Pandas 讀入 csv file
2. 刪除缺漏值的 data
   （嘗試塞入 mean value of column，但結果不佳）
3. 將 y_train 從 train_data 分離，並將 y_train 的 text encode，並且轉成 DataFrame 格式
4. 合併 test_data 和 train_data 一起處理
5. 將時間欄位刪除
6. 將 categorical data(文字)用 labelEncoder 和 oneHotEncoder 轉成 number
7. 分離 test_data 和 train_data
8. MinMaxScaler 做正規化
9. (SelectKBest 做 feature selection，但要配合 AdaBoost 和 outlier remove 結果才好)
10. Density base cluster(DBSCAN)做 outlier detection
11. Downsample majority 修正 imbalancing data
    （嘗試 upsample minority，但效果不彰）
12. Random Forest(n_estimators=250,criterion="entropy")目前是配合 downsample 效果最好
    （嘗試 AdaBoost、GradientBoost、Voting，但效果不彰）
13. 輸出 output.csv 檔案

目前最佳配方：Random Forest(ne=250,c="entropy") + downsample
其他較佳配方：AdaBoost(ne=200) + SelectKBest(10) + DBSCAN +downsample
核心突破：downsample with random number

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)