

# Data Science HW1

四資工三甲 B10632024 林奕辰

## 演算法流程：

讀檔：

1. Pandas 讀入 csv file

MissingValue：

2. dropna()刪除缺漏值的 data  
(嘗試塞入 mean value of column，但結果不佳)

Categorical Data：

3. 將 y\_train 從 train\_data 分離，並將 y\_train 的 text encode，最後轉成 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

Preprocessing(Normalize, FeatureSelect, ImbalanceData)：

8. MinMaxScaler()正規化
9. SelectKBest 做 feature selection(降維)，但要配合 AdaBoost 結果才好
10. (DBSCAN、zscore 做 outlier detection，但效果不彰)
11. Downsample majority 修正 imbalancing data  
(嘗試 upsample minority、SMOTE，但效果不彰)

Algorithm(Bagging or Boosting)：

12. AdaBoost(n\_estimators=165, criterion="entropy")  
(嘗試 RandomForest(不差)、GradientBoost、Voting，但效果不彰)

Output(Visualize, CSV file) :

13. plt.hist() 視覺化結果

14. 輸出 output.csv 檔案

### **嘗試與搭配：**

目前最佳配方：AdaBoost(ne=165) + SelectKBest(23 取 18) + 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 output21.csv