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：

1. dropna()刪除缺漏值的data

(嘗試塞入mean value of column，但結果不佳)

Categorical Data：

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

Preprocessing(Normalize, FeatureSelect, ImbalanceData)：

1. MinMaxScaler()正規化
2. SelectKBest做feature selection(降維)，但要配合AdaBoost結果才好
3. (DBSCAN、zscore做outlier detection，但效果不彰)
4. Downsample majority修正imbalancing data

(嘗試upsample minority、SMOTE，但效果不彰)

Algorithm(Bagging or Boosting)：

1. AdaBoost(n\_estimators=165,criterion="entropy")

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

Output(Visualize, CSV file)：

1. plt.hist()視覺化結果
2. 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