資料科學 HW1

B10632026 四資工三甲 吳苡瑄

利用每日天氣觀測樣本做訓練，給入當天的觀測數據，預測隔天會不會降雨

執行環境

- Python 3
  - `python3 randomForest.py`
  - 讀進 `train.csv` 、 `test.csv`
  - 結果輸出到 `submit.csv`

程式架構

1.data preprocessing

- 利用 `pd.read_csv()` 讀進資料
- 只要有缺項的值就整row丟掉
- 按照Attribute23分成Yes和No兩群資料
- 因為train資料中No的資料遠比Yes多
  - 把No的資料random取跟Yes一樣多的值打亂合併，其餘丟掉
- 對方位的資料作one hot
  - 先把train中的Attribute23拿出來
  - 把train資料和test合在一起一起做onehot才不會做出不同結果
  - 做完one hot後再分開
- 把日期欄位丟掉

2.training

- RandomForest
  - 使用 RandomForestClassifier 建立 RandomForest model
  - fit training set
  - 印出score和oob_score
  - predict出test資料的結果

3.Output

- 把預測結果輸出成csv

Random Forest

- 眾人投票型
- n_estimators要創造幾個決策樹來投票
- n_jobs決定要用你電腦的幾個核心去算

Other

- 嘗試了DecisionTree、knn、ann效果都沒有RandomForest好
- 嘗試去補足缺失的值但效果不佳
- 嘗試把早上九點的資料丟棄，沒有顯著提升
- 嘗試去標記相關性較高的feature也還好