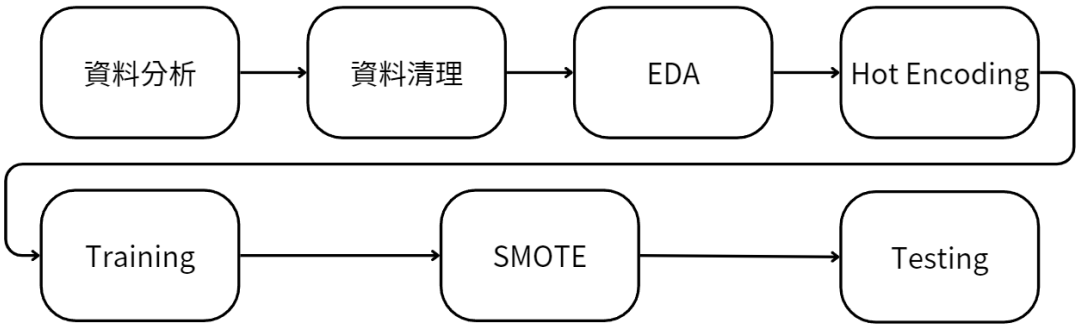
HW1 報告

B11015030 四資工三乙 張睿麟

實作流程

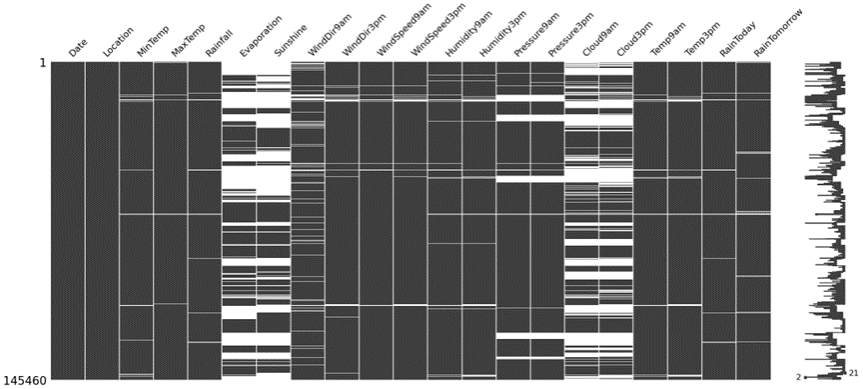


資料分析

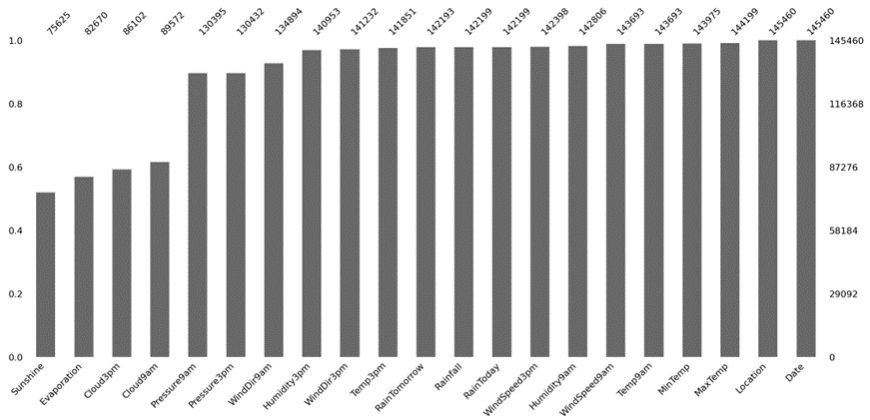
- 查看訓練資料以及測試資料

	Training Data	Testing Data
shape	(17109, 21)	(823, 20)
缺失值	有	無

使用 missingno 套件，將缺失值視覺化，



白色空缺部分即為資料集缺失的內容。Date 和Location欄位並無缺失值，剩下的欄位零星皆有缺失值。



排序缺失值的比例，Sunshine、Evaporation、Cloud3pm、Cloud9am有大比例的缺失值。

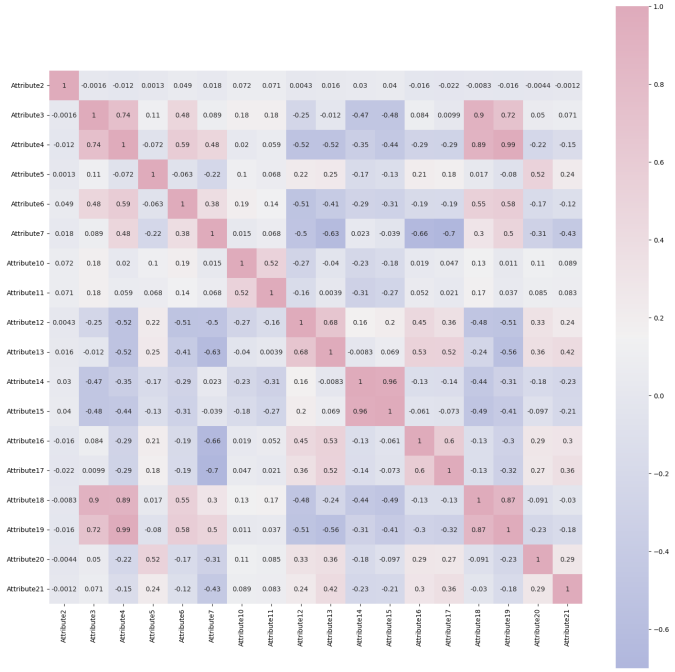
資料清理

- 處理缺失值
  - 不可直接捨棄有缺失值的資料，會使資料從 17109 → 7276 筆資料
  - 數值型欄位 ⇒ 中位數補值
  - 類別型欄位 ⇒ 眾數補值

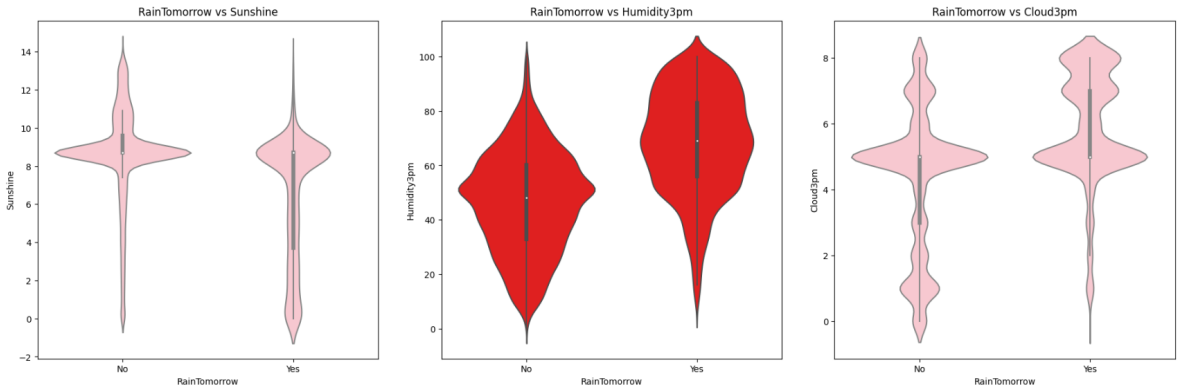
  有使用過其他方式補值（補零、補平均值），中位數的補值有比較好的分數。

EDA

- 捨棄 Attribute 1 (Date) 欄位，因為時間及地區提供的資訊沒有太有幫助
  - 每年、每月、每日的降雨沒有特殊的、突出的特徵
- Correlation Matrix
  - Attribute 3, 4, 18, 19 有高度正相關，所以把 Attribute 18, 19 給 drop。



- **univariate analysis**
  - **利用 histogram查看 數值型 欄位 的分布狀況**
- **Bivariate analysis**
  - **查看 數值型 欄位和明天是否降雨的關係**
    - Sunshine 的欄位中可以發現，值越低，明天價與的可能性就越高，反之亦然。
    - 在 Humidity3pm 欄位中，發現值越高，隔天就更容易下雨，反之亦然。所以，我想要新增一個欄位是「Humidity3pm_70」，大於等於 70 填 1，其餘填 0。
    - Cloud9am 跟 Cloud3pm 分布蠻類似的，而且 大於 6 的情況比較容易隔天下雨。



處理離群值

去除三個標準差的離群值

Hot Encoding

針對 ['Location', 'WindDir9am', 'WindDir3pm'] 這三個類別型的欄位，做 hot encoding，把 training data 的 shape 變成
⇒ (15911, 100)

Scaling

使用 Standardization 會導致成效降低，故不做任何的 Standardization

Training

- 切分 8 : 2 的 training data 跟 validation data

SMOTE

- 使用 SMOTE 做 oversampling

Model

- 使用 5 個 Model 集合成一個 Voting Classifier
  - Random Forest
  - Logistic Regression

- SVM
- Gradient Boosting Classfier
- Xgboost

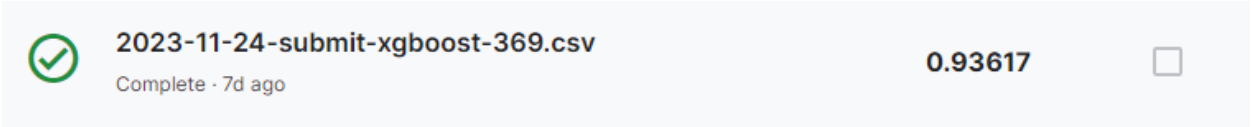
參數請參考程式碼

Testing

- 評估模型的表現，使用 accuracy、precision、recall 和 F1-score 進行評估
- 計算並繪製混淆矩陣，以了解模型的預測情況
- 根據模型的性能，進行必要的調整和改進，以提高模型的準確性和穩定性

最終成績

補充



我的成績在 Kaggle 上有 0.93617 的 public 成績，是因為我使用了類似的資料集(Rain in Australia: https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package)

會這麼高的成績是因為模型 overfitting 了，使用的資料集有 public set 的資料，所以模型基本上是在看答案作答，所以才會這麼高。