台南空氣品質指標預測

第十七組

統計 112 劉恩兆

統計 112 宋穎恩

摘要

空氣污染在近年來在世界各地已成為嚴重的環境污染問題，是影響健康的環境風險之一，世界衛生組織更聲稱，估計每年有 700 萬人死於空氣污染有關的疾病，並將其與吸菸和不健康飲食相提並論。本研究使用機器學習之方法去預測台南空氣品質指標（AQI），文中共使用五種機器學習模型，分別為 Decision Tree、Time Series Random Forest、LightGBM、XGBoost 及 LSTM，並利用不同機器學習模型來比較空氣品質指標預測效力，搜集 2018 年 6 月至 2023 年 5 月間的台南測站空氣品質指標歷史資料與天氣歷史資料，以過去四年的資料來預測未來一年空氣品質指標的走向，使用 2018 年至 2022 年對模型進行訓練，並用 2022 年至 2023 年來進行驗證，來評估模型的效果。

壹、前言

一、研究背景與動機

世界衛生組織於 2014 年指出室內與室外空氣污染會對呼吸道造成危害，包含急性呼吸道感染與慢性阻塞性肺病，且與心血管疾病與癌症有強烈的相關性。同時，世界衛生組織於日內瓦舉辦的世界衛生大會也提出各國衛生當局須擔任關鍵性角色並喚起公眾意識，有效改善空氣污染便可減少醫療支出與解救生命。

隨著科技發展與經濟水平提升，空氣污染議題受到廣泛的關注以及討論，台灣環保署於 1933 年完成全國空氣品質監測網的設置，藉此達到監督空氣品質與促進國民健康的目的，根據政府資料顯示，台灣南部地區的空氣品質較北部地區差，舉例來說，2023 年 1 月由於境外空污移入，全台共有 15 測站空氣品質亮紅燈，多集中在彰化以南等地，其中又以台南 PM2.5 濃度最高。而台南地區為成功大學學生的主要活動地區，因此，本研究選擇台南測站作為主要分析對象，結合空氣指標與天氣歷史資料，使用機器學習模型進行預測，進而為社會大眾提供準確且有效的資訊，以降低健康風險。

二、研究目的

本研究針對 2018 年 6 月至 2023 年 5 月間台南測站的空氣品質指標與天氣歷史資料，來進行分析以及研究，以期望達成以下目標：

1. 根據空氣品質之預測結果，以提醒高危險地區之居民外出需注意的事項，可以事先調整其活動。
2. 找出最能有效預測空氣品質指標之模型，使民眾可提前得知資訊並加以防範。

三、資料來源

1. 空氣品質指標（AQI）歷史資料：

於行政院環保署開放資料平台下載，有 25 個變數，包含 SiteName（測站名稱）、County（縣市）、AQI（空氣品質指標）、Pollutant（空氣污染指標物）、Status（狀態）、SO2（二氧化硫）[ppb]、CO（一氧化碳）[ppm]、O3（臭氧）[ppb]、O3_8hr（臭氧 8 小時移動平均）[ppb]、PM10（懸浮微粒）[μg/m3]、PM2.5（細懸浮微粒）[μg/m3]、NO2（二氧化氮）[ppb]、Nox（氮氧化物）[ppb]、NO（一氧化氮）[ppb]、WindSpeed（風速）[m/sec]、WindDirec（風向）[degrees]、DataCreationDate（資料發布時間）、Unit（單位）、CO_8hr（一氧化碳 8 小時移動平均）[ppm]、PM2.5_AVG（細懸浮微粒移動平均值）[μg/m3]、PM10_AVG（懸浮微粒移動平均值）[μg/m3]、SO2_AVG（二氧化硫移動平均值）[ppb]、Longitude（經度）、Latitude（緯度）及 SiteId（測站編號），每小時回傳一次，資料涵蓋時間為 2018 年 6 月 1 日至 2023 年 5 月 31 日，共五年。

AQI_202105

	"sitename"	"county"	"aqi"	"pollutant"	"status"	"so2"	"co"	"o3"	"o3_8hr"	"pm10"	...	"winddirec"	"datacreationdate"	"unit"	"co_8hr"	"pm2.5_avg"	"pm10_avg"	"so2_avg"	"longitude"	"latitude"	"siteid"
0	屏東(枋寮)	屏東縣	68.0	細懸浮微粒	普通	0.5	0.3	15.2	45	48.0	...	65	2021-05-01 00:00	NaN	0.3	22.0	41.0	4.0	120.591167	22.370947	313
1	馬祖	連江縣	129.0	臭氧八小時	對敏感族群不健康	3.3	0.33	59.3	79	29.0	...	207	2021-05-01 00:00	NaN	0.3	31.0	44.0	3.0	119.949875	26.160469	75
2	埔里	南投縣	85.0	細懸浮微粒	普通	2.0	0.45	NaN	65	55.0	...	351	2021-05-01 00:00	NaN	0.4	29.0	43.0	1.0	120.967903	23.968842	72
3	復興	高雄市	74.0	臭氧八小時	普通	1.6	0.5	50	62	23.0	...	238	2021-05-01 00:00	NaN	0.5	14.0	24.0	2.0	120.312017	22.608711	71
4	永和	新北市	77.0	細懸浮微粒	普通	1.4	0.94	18.8	35	35.0	...	142	2021-05-01 00:00	NaN	1.2	26.0	44.0	2.0	121.516306	25.017000	70
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
120379	苗栗	苗栗縣	20.0	NaN	良好	2.1	0.24	12.8	13	5.0	...	220	2021-05-31 23:00	NaN	0.4	6.0	5.0	2.0	120.820200	24.565269	26
120380	臺南	臺南市	29.0	NaN	良好	1.8	0.18	31.2	31	13.0	...	217	2021-05-31 23:00	NaN	0.2	6.0	17.0	2.0	120.202617	22.984581	46
120381	安南	臺南市	40.0	NaN	良好	1.9	0.24	29.2	31	24.0	...	245	2021-05-31 23:00	NaN	0.2	12.0	24.0	2.0	120.217500	23.048197	45
120382	淡水	新北市	20.0	NaN	良好	1.8	0.37	13.5	22	5.0	...	NaN	2021-05-31 23:00	NaN	0.3	5.0	8.0	2.0	121.449239	25.164500	10
120383	善化	臺南市	23.0	NaN	良好	1.3	0.17	9.6	25	17.0	...	79	2021-05-31 23:00	NaN	0.2	7.0	15.0	1.0	120.297142	23.115097	44

圖 1、2021 年 5 月空氣品質指標（AQI）歷史資料概覽

2. 歷史天氣資料：

使用 Python 爬蟲抓取中央氣象局台南站歷史天氣資料，有 17 個變數，包含觀測時間[hour]、經度[hPa]、海平面氣壓[hPa]、氣溫[℃]、露點溫度[℃]、相對溼度

[%]、風速[m/s]、風向[360degree]、最大陣風[m/s]、最大陣風風向[360degree]、降水量[mm]、降水時數[h]、日照時數[h]、全天空日射量[MJ/㎡]、能見度[km]、紫外線指數及總雲量[0～10]，每小時回傳一次，資料涵蓋時間為 2018 年 6 月 1 日至 2023 年 5 月 31 日，共五年。

	觀測時間(hour)	測站氣壓(hPa)	海平面氣壓(hPa)	氣溫(℃)	露點溫度(℃)	相對濕度(%)	風速(m/s)	風向(360degree)	最大陣風(m/s)	最大陣風風向(360degree)	降水量(mm)	降水時數(h)	日照時數(h)	全天空日射量(MJ/㎡)	能見度(km)	紫外線指數	總雲量(0~10)
0	1	1007.4	1010.5	29.2	24.5	76	3.4	330	7.5	330	0.0	0.0	...	0.00	...	0	...
1	2	1007.2	1010.3	28.8	24.5	78	3.5	350	7.1	310	0.0	0.0	...	0.00	20.0	0	2.0
2	3	1007.0	1010.1	29.0	24.5	77	2.7	350	5.1	330	0.0	0.0	...	0.00	...	0	...
3	4	1006.3	1009.4	28.6	24.6	79	2.4	10	5.3	360	0.0	0.0	...	0.00	...	0	...
4	5	1006.5	1009.6	28.5	24.5	79	2.4	340	5.9	30	0.0	0.0	...	0.00	15.0	0	7.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
43819	20	995.0	998.0	28.7	23.6	74	5.7	20	9.3	30	0.0	0.0	...	0.00	...	0	...
43820	21	995.5	998.5	28.4	24.0	77	4.9	10	8.2	20	0.0	0.0	...	0.00	...	0	...
43821	22	996.0	999.0	28.0	24.0	79	4.2	20	9.3	20	0.0	0.0	...	0.00	...	0	...
43822	23	995.9	998.9	27.8	24.2	81	3.2	350	6.5	350	0.0	0.0	...	0.00	...	0	...
43823	24	995.7	998.7	27.6	24.1	81	2.6	10	6.1	10	0.0	0.0	...	0.00	...	0	...

43824 rows × 17 columns

圖 2、2018 年 6 月至 2023 年 5 月台南歷史天氣資料概覽

貳、敘述統計

一、空氣品質指標（AQI）分級介紹

根據行政院環保署空氣品質監測網，空氣品質指標為依據監測資料將當日空氣中 O3（臭氧）、PM2.5（細懸浮微粒）、PM10（懸浮微粒）、CO（一氧化碳）、SO2（二氧化硫）及 NO2（二氧化氮）濃度等數值，以其對人體健康的影響程度，分別換算出不同污染物之副指標值，再以當日各副指標之最大值為該測站當日之空氣品質指標值（AQI）。其中 AQI 指標 0 – 50 為良好 (Good)、51 – 100 為普通 (Moderate)、101 – 150 為對敏感族群不健康 (Unhealthy for Sensitive Groups)及 151 – 200 為對所有族群不健康 (Unhealthy)。

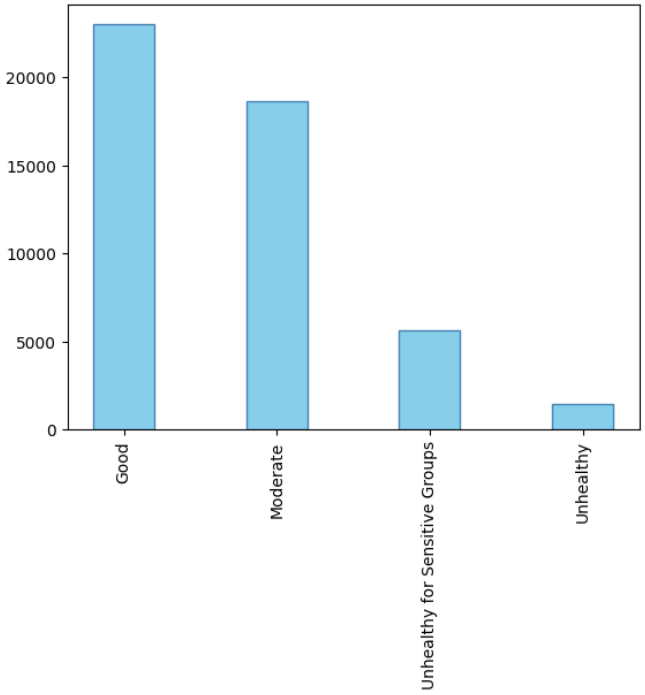
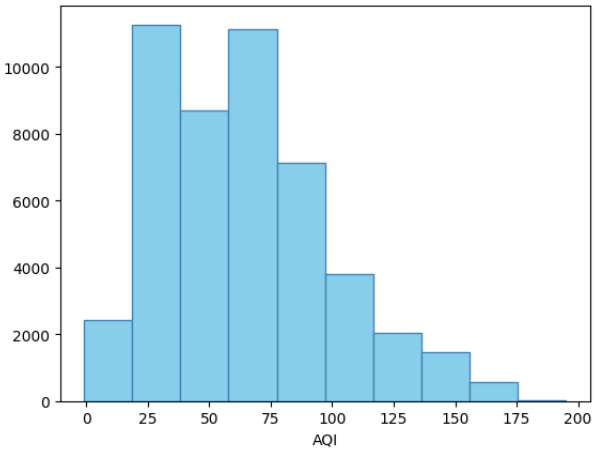
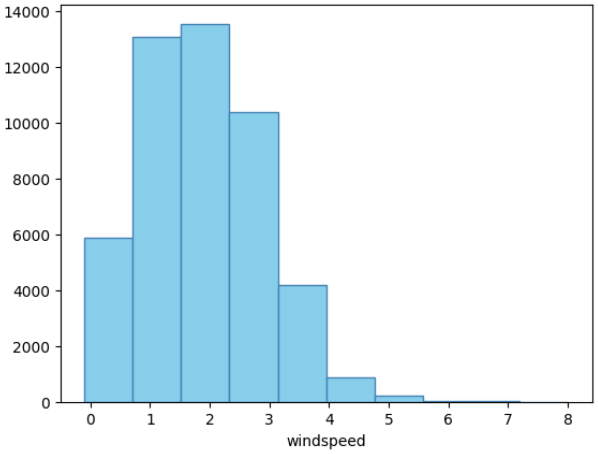


圖 3、2018 年 6 月至 2023 年 5 月 AQI 等級分佈情形

二、特徵分佈圖

根據資料集欄位 AQI、windspeed、winddirec、氣溫、相對濕度、最大陣風、降水量以及降水時數依序繪製分布圖。





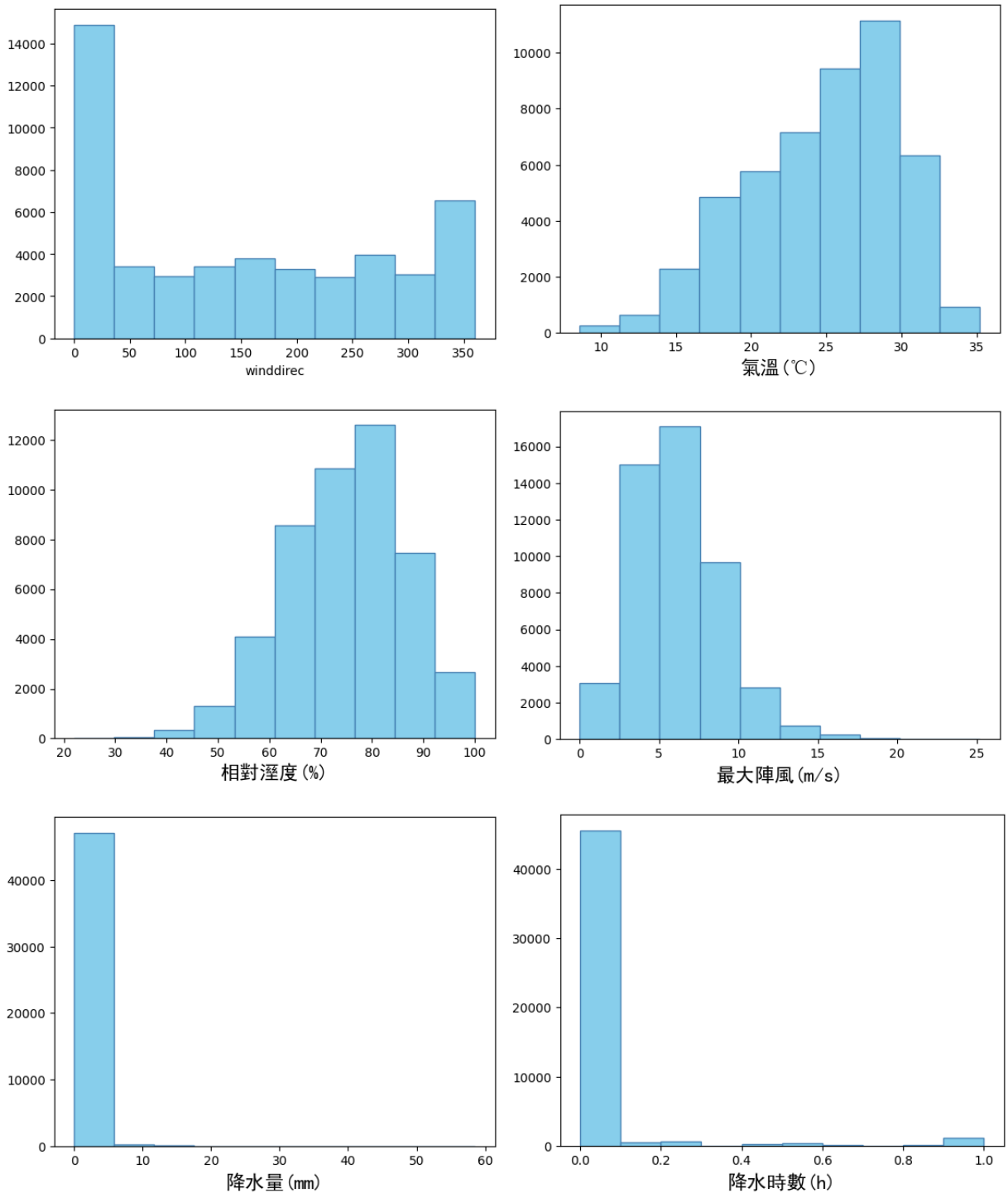


圖 4 – 11、2018 年 6 月至 2023 年 5 月 AQI、windspeed、winddirec、氣溫、相對濕度、最大陣風、降水量以及降水時數分佈情形

參、研究方法

一、資料預處理

1. 篩選測站

由於本研究是以台南地區為主要分析對象，須先從空氣品質指標歷史資料篩選出測站名稱為台南的資料。

2. 篩選空氣品質指標（AQI）相關欄位

由於兩份資料中有許多與空氣品質指標（AQI）不直接相關之欄位，因此需將不必要的欄位去除。根據行政院環保署空氣品質監測網，在無特殊異常污染排放情形下，空氣品質受到氣象條件變化影響較大，影響空氣品質的氣象因素包含水平風、垂直風、混合層高的、降雨等等，因此，我們在空氣品質指標歷史資料欄位中選擇空氣品質指標（AQI）、記錄時間、風速以及風向，共四個欄位；而在台南歷史天氣資料欄位中選擇觀測時間、氣溫、相對濕度、最大陣風、降水量以及降水時數，共六個欄位。

"aqi"	"datacreationdate"	"windspeed"	"winddirec"
22.0	2018-06-01 00:00	2.1	359.0
24.0	2018-06-01 01:00	2.4	8.4
24.0	2018-06-01 02:00	2.0	1.4
23.0	2018-06-01 03:00	1.9	358.0
22.0	2018-06-01 04:00	2.4	46.0
...	...	...	...
24.0	2022-05-31 19:00	1	219
23.0	2022-05-31 20:00	1.4	219
22.0	2022-05-31 21:00	1.1	207
32.0	2022-05-31 22:00	0.7	187
19.0	2022-05-31 23:00	0.8	160

圖 12、空氣品質歷史資料欄位篩選

觀測時間(hour)	氣溫(°C)	相對溼度(%)	最大陣風(m/s)	降水量(mm)	降水時數(h)
2018-06-01 00:00	29.2	76	7.5	0.0	0.0
2018-06-01 01:00	28.8	78	7.1	0.0	0.0
2018-06-01 02:00	29.0	77	5.1	0.0	0.0
2018-06-01 03:00	28.6	79	5.3	0.0	0.0
2018-06-01 04:00	28.5	79	5.9	0.0	0.0
...	...	...	...	...	...
2023-05-31 19:00	28.7	74	9.3	0.0	0.0
2023-05-31 20:00	28.4	77	8.2	0.0	0.0
2023-05-31 21:00	28.0	79	9.3	0.0	0.0
2023-05-31 22:00	27.8	81	6.5	0.0	0.0
2023-05-31 23:00	27.6	81	6.1	0.0	0.0

圖 13、台南歷史天氣資料欄位篩選

3. 合併資料

藉由空氣品質指標歷史資料的紀錄時間與台南天氣歷史資料中的觀測時間來合併資料。

	"aqi"	"datacreationdate"	"windspeed"	"winddirec"	氣溫(℃)	相對溼度(%)	最大陣風(m/s)	降水量(mm)	降水時數(h)
0	22.0	2018-06-01 00:00	2.1	359.0	29.2	76.0	7.5	0.0	0.0
1	24.0	2018-06-01 01:00	2.4	8.4	28.8	78.0	7.1	0.0	0.0
2	24.0	2018-06-01 02:00	2.0	1.4	29.0	77.0	5.1	0.0	0.0
3	23.0	2018-06-01 03:00	1.9	358.0	28.6	79.0	5.3	0.0	0.0
4	22.0	2018-06-01 04:00	2.4	46.0	28.5	79.0	5.9	0.0	0.0
...	...	...	...	...	...	...	...	...	...
40089	24.0	2022-05-31 19:00	1.0	219.0	29.3	77.0	4.0	0.0	0.0
40090	23.0	2022-05-31 20:00	1.4	219.0	29.2	76.0	4.5	0.0	0.0
40091	22.0	2022-05-31 21:00	1.1	207.0	29.2	78.0	4.3	0.0	0.0
40092	32.0	2022-05-31 22:00	0.7	187.0	29.1	74.0	3.3	0.0	0.0
40093	19.0	2022-05-31 23:00	0.8	160.0	28.5	81.0	3.8	0.0	0.0

圖 14、使用共同欄位時間合併的最終資料概覽

4. 分割訓練集與測試集

總資料集為 2018 年 6 月 1 日至 2023 年 5 月 31 日台南空氣品質指標與天氣歷史資料，一共有 48, 794 筆及 8 個變數，依照訓練集 80%與測試集 20%比例分割，因此，訓練集資料為 2018 年 6 月 1 日至 2022 年 5 月 31 日台南空氣品質指標與天氣歷史資料，一共有 40, 094 筆及 8 個變數，測試集資料則為 2022 年 6 月 1 日至

2023 年 5 月 31 日台南空氣品質指標與天氣歷史資料，一共有 8, 700 筆及 8 個變數。

5. 遺失值填補

首先，將觀測日期時間按先後順序排序並設為 index，接下來，由於此資料集為時間序列資料，我們認為使用平均數或是中位數補值較為偏頗，且資料會每小時更新一次，所以我們選擇使用 ffill 函數向前填補缺失值，fillna(method = 'ffill') 用同一欄位的前一筆資料進行遺失值填補。

選擇使用 ffill 函數向前填補缺失值的原因有以下：

a. 時間序列資料若是使用平均數、中位數等，即無法反映時間資料隱含的資訊
b. 因此資料集每筆資料間隔時間僅為一小時，在一小時中天氣資料應不會有劇烈變化，因此用前一筆資料進行填補較符合所需

欄位	遺失值個數
AQI	2
風速	387
風向	387
氣溫	0
相對濕度	0
最大陣風	0
降水量	282
降水時數	0

表 1、訓練集資料遺失值統計

欄位	遺失值個數

AQI	227
風速	238
風向	227
氣溫	3
相對濕度	811
最大陣風	13
降水量	945
降水時數	0

表 2、測試集資料遺失值統計

6. MinMaxScaler

使用 sklearn.preprocessing 中的 MinMaxScaler 模塊，將原始數據按照最小值和最大值進行線性轉換，使數據的範圍落在[0, 1]之間，其優點為：

a. 提升收斂速度：對特徵進行縮放可以幫助算法更快地收斂。舉例來說，XGBoost 使用梯度提升算法，在訓練過程中需要對特徵進行加權和調整。如果特徵具有不同的範圍和尺度，這可能會導致模型收斂速度變慢。通過將數據縮放到一個統一的範圍，可以減少這種不同尺度帶來的影響。

b. 避免權重偏置：如果某些特徵具有較大的範圍和尺度，則他們的權重也可能會較大，這可能導致模型在訓練過程中偏向於這些特徵。舉例來說，XGBoost 的梯度提升算法依賴於特徵的權重和梯度來進行模型的訓練，通過使用 MinMaxScaler 縮放特徵，可以將所有特徵的範圍調整到相似的尺度，從而避免權重偏置，使得模型更公平地對待各個特徵。

二、分析方法

本文使用的分析方法為使用不同的機器學習模型進行回歸預測 AQI 值，再將其值對應分級區間轉換為類別型標籤，最後進行準確率與混淆矩陣評估。

1. Decision Tree

決策樹 (Decision Tree) 是一種常用的監督式機器學習算法，用於解決分類和回歸問題。決策樹會根據訓練資料產生一棵樹，依據訓練出來的規則來對新樣本進行預測，其結構由節點和分支組成，每個節點代表一個特徵，每個分支代表特徵的不同取直，而每個葉子節點代表一個分類標籤或回歸值，從根節點開始，通過選擇適當的特徵和取值來分裂數據，直到達到葉子節點為止。

2. Time Series Random Forest

隨機森林 (Random Forest) 常被用於分類與回歸的問題，而對於時間序列預測，可以使用時間序列隨機森林 (Time Series Random Forest, TSF)，是一種隨機森林的變體，使用滾動窗口來建構決策樹，每棵決策樹只預測下一個時間的輸出，此方法可以捕捉時間序列中的動態模式，避免了過擬合與洩漏的問題。

3. LightGBM

LightGBM 全名為 Light Gradient Boosting Machine，由微軟團隊於 2017 所開發的，是一種基於決策樹算法的分布式梯度提升機器學習框架，用於處理大規模數據集，其主要優點為：

a. 梯度提升算法：通過迭代新增新的模型來最小化損失函數，以提高預測性能。
b. 輕量級和高效：使用基於直方圖的算法來有效地處理特徵值，減少了內存使用量和計算時間。
c. 分布式訓練：可將數據集的特徵劃分為多個子集，並在多個計算節點上進行並行計算。
d. 高效的特徵工程：具有內置特徵工程功能，可以處理缺失值、類別特徵和高維稀疏特徵。

4. XGBoost (eXtreme Gradient Boosting)

XGBoost 全名為 eXtreme Gradient Boosting，由華盛頓大學博士生陳天奇於 2014 年所提出的，是一種 Gradient Boosted Tree(GBDT)，每一棵樹都是互相關聯的，每一次保留原來的模型不變，並加入一個新的函數至模型中，目的為修正上一棵樹錯誤的地方，以提升整體模型。XGBoost 除了可以做分類也能進行回歸連續性數值的預測，其主要優點為：

a. 利用了二階梯度來對節點進行劃分。
b. 使用局部近似算法對分裂節點進行優化。
c. 在損失函數中加入 L1/L2 項，使用泰勒展開式展開，控制模型複雜度，加快模型優化速度。
d. 採用特徵隨機採樣的技巧，和隨機森林一樣在生成每一棵樹的時候隨機抽取特徵，不需要使用全部的特徵進行訓練，能避免過擬合。
e. 提供 GPU 平行化運算

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{1}$$

XGBoost 的目標函數 = 損失函數 + 正則項，如上述數學式(1)，損失函數用於衡量預測值與實際值的誤差，正則項可以說是懲罰函數，用於限制模型避免過擬合，其演算法步驟如下：

a. 先建立一個簡單一層的回歸樹預測模型$F(x)$，我們可以得到$F(x)$預測出來的值$(\hat{Y})$與觀察值y的殘差$(Residual)$。
b. 為了提升$F(x)$的預測能力，再透過 Input 為x、Output 為$Residual$，建立一個$h(x)$。
c. 因此得到了新的模型為$H(x) = F(x) + h(x)$，並希望預測能力能比$F(x)$好，因為有修正殘差$(bias)$。
d. 重複 a - c 的過程。
e. 最後可以得到$F(x) + h(x) + n(x) + g(x)$ ...的預測模型。

5. LSTM (Long Short-Term Memory)

長短期記憶 (Long Short-Term Memory, LSTM) 是一種循環神經網路(RNN)的變體，由 Sepp Hochreiter 和于爾根・施密德胡伯 (Jürgen Schmidhuber) 於 1997 年所提出，專門設計於處理序列數據，特別是具有長期相依性的數據，利用三個控制閥 (Gate) 來決定記憶的儲存與使用，分別為遺忘閥 (Forget Gate, 以 ft 表示)、輸入閥 (Input Gate, 以 it 表示) 以及輸出閥 (Output Gate)，也就是多加一個長期記憶 (Long Term Memory) 的變數 (Ct)，因此，LSTM 可以有效地捕捉與記憶長期的相關信息。

三、主要研究結果

1. Decision Tree

先從訓練集與測試集中分離特徵和目標函數 AQI，建立 DecisionTreeClassifier 模型，其參數設定如下：

- criterion：entropy

最後，將資料集放入模型進行訓練，並驗證模型與將其結果視覺化。

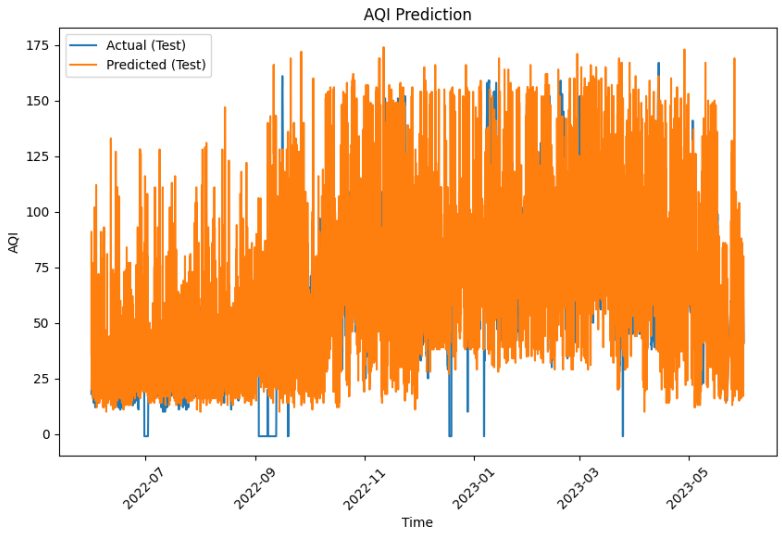


圖 15、Decision Tree 預測結果與實際 AQI 比較圖

橘色：預測結果 / 藍色：實際 AQI

2167	868	165	20
1032	**2467**	716	126
112	418	**134**	29
216	169	48	**13**

表 3、Decision Tree 混淆矩陣

若是對應至 AQI 分級，其測試集準確率為 0.5495。

2. Time Series Random Forest

先從訓練集與測試集中分離特徵和目標函數 AQI，使用 tsfresh 庫建立 TimeSeriesForestRegressor 模型，其參數設定如下：

- n_estimators：100
- random_state：0

最後，將資料集放入模型進行訓練，並驗證模型與將其結果視覺化。

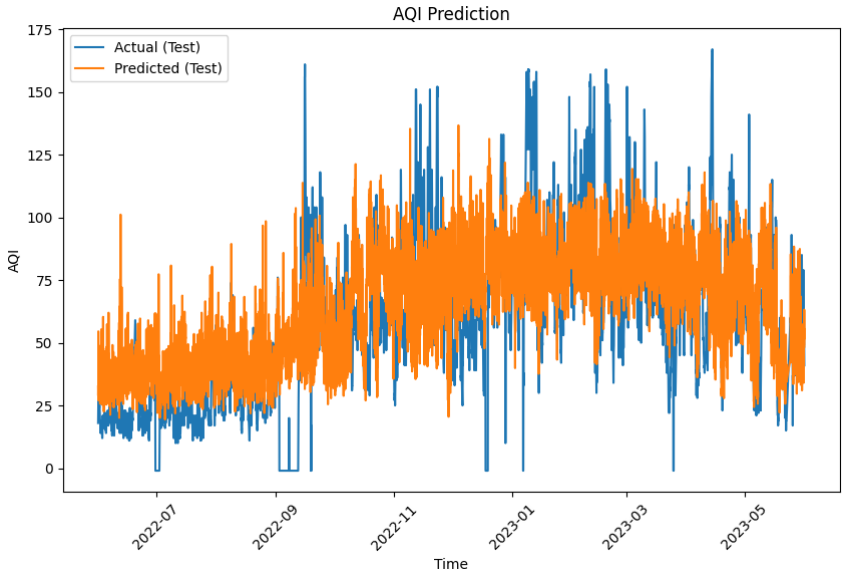


圖 16、Time Series Random Forest 預測結果與實際 AQI 比較圖

橘色：預測結果 / 藍色：實際 AQI

2141	1043	36	0
493	**3661**	187	0
26	606	**61**	0
166	259	21	**0**

表 4、Time Series Random Forest 混淆矩陣

若是對應至 AQI 分級，其測試集準確率為 0.6739。

3. LightGBM

在 LightGBM 回歸預測中，我們使用天氣特徵預測 AQI 值，先從訓練集與測試集中分離特徵和目標函數 AQI，並建構 lgb 中的 Dataset 格式，建立 LightGBM 模型，其參數設定如下：

- num_leaves：31
- num_trees：100
- objective：regression
- metric：rmse，使用均方根誤差（RMSE）作為評估誤差

最後，將資料集放入模型進行訓練，並驗證模型與將其結果視覺化。

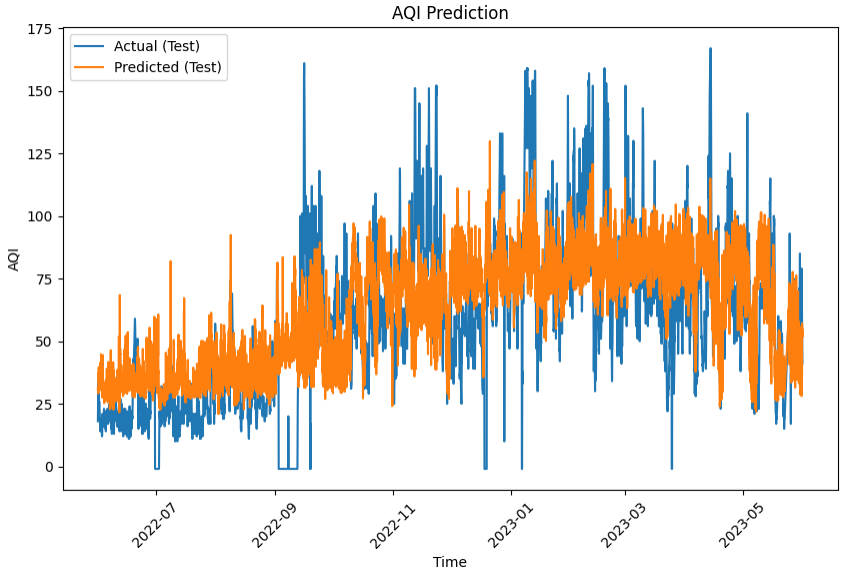


圖 17、LightGBM 預測結果與實際 AQI 比較圖

橘色：預測結果 / 藍色：實際 AQI

2341	878	1	0
544	**3738**	59	0
23	634	**36**	0
204	229	13	**0**

表 5、LightGBM 混淆矩陣

若是對應至 AQI 分級，其測試集準確率為 0.7029。

4. XGBoost (eXtreme Gradient Boosting)

在 XGBoost 回歸預測中，我們使用天氣特徵預測 AQI 值，先從訓練集與測試集中分離特徵和目標函數 AQI，並建立 DMatrix 資料結構，DMatrix 是 XGBoost 中特有的資料結構，用於有效地存儲和操作訓練數據，他將數據組織為稠密或稀疏矩陣的形式，並提供了一個高效的接口，使得 XGBoost 可以快速地從 DMatrix 中獲取數據進行模型訓練和預測。下一步，我們需定義模型參數：

- objective：reg:squarederror，使用平方誤差作為目標函數
- eta：0.1，學習率為 0.1
- max_depth：6，樹的最大深度為 6
- colsample_bytree：0.8，每棵樹使用的特徵比例為 0.8
- subsample：0.8，每棵樹使用的樣本比例為 0.8
- eval_metric：rmse，使用均方根誤差（RMSE）作為評估誤差

最後，將資料集放入模型進行訓練，並驗證模型與將其結果視覺化。

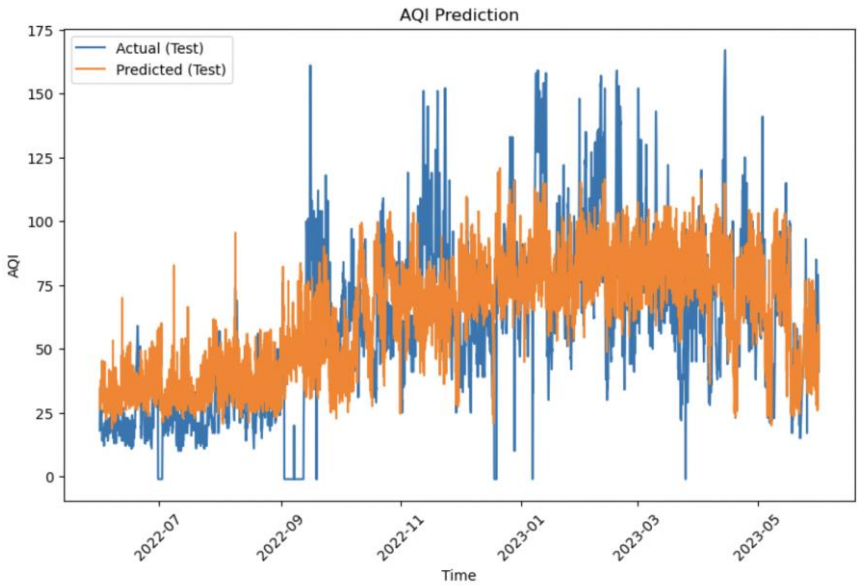


圖 18、XGBoost 預測結果與實際 AQI 比較圖

橘色：預測結果 / 藍色：實際 AQI

2301	911	8	0
532	**3728**	81	0
26	631	**36**	0
181	249	16	**0**

表 6、XGBoost 混淆矩陣

若是對應至 AQI 分級，其測試集準確率為 0.6971。

由圖 18 與表 6 可見，模型對於預測較極端之 AQI 的效果不盡理想。

5. LSTM (Long Short-Term Memory)

使用 PyTorch 框架建立 LSTM 模型，其參數設定如下：

- input_size ： 7
- hidden_size ： 64
- num_layers ： 2
- output_size ： 1
- num_epochs ： 80
- batch_size ： 7*24，由過去 7 天資料預測下一個時間點的值
- learning_rate ： 0.01

最後，將資料集放入模型進行訓練，並驗證模型與將其結果視覺化。

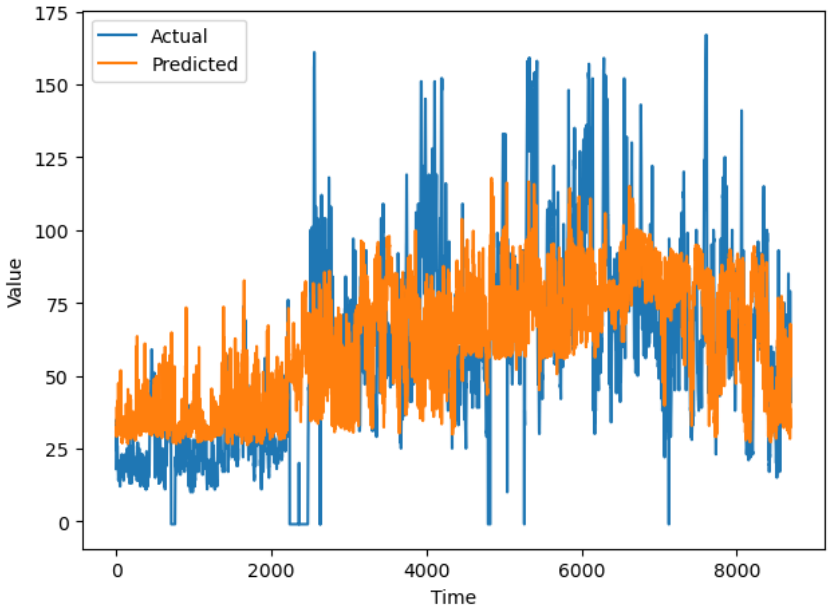


圖 19、LSTM 預測結果與實際 AQI 比較圖

橘色：預測結果 / 藍色：實際 AQI

2308	912	0	0
587	**3728**	33	0
57	631	**9**	0
146	296	4	**0**

表 7、LSTM 混淆矩陣

若是對應至 AQI 分級，其測試集準確率為 0.6940。

由圖 19 與表 7 可見，模型對於預測較極端之 AQI 的效果不盡理想。

6. 分析結果比較

統整上述模型的分析結果：

模型	測試集準確率
Decision Tree	0.5495
Time Series Random Forest	0.6739
LightGBM	0.7029
XGBoost	0.6971
LSTM	0.6940

表 8、Decision Tree、Time Series Random Forest、LightGBM、XGBoost 及 LSTM 的準確率比較

註：所有模型皆有使用 GridSearchCV 挑選出最佳模型，詳細說明可見第 7 點「模型參數選擇」

7. 模型參數選擇

為了能找出模型的最佳效能，我們使用了 GridSearchCV 套件，此套件可以幫助我們系統地探索不同參數組合的模型，並找到最佳的參數配置。

GridSearchCV 是一種參數調優的方法，它通過遍歷所有可能的參數組合，並對每個組合進行交叉驗證來評估模型的性能。這樣可以找到在給定參數空間中性能最好的模型參數。以下將使用 XGBoost 為例，此模型可調整的參數有：

- max_depth： [6, 8, 10]
- learning_rate： [0.1, 0.01, 0.001]
- colsample_bytree： [0.8, 1.0]
- subsample： [0.8, 1.0]

此套件會遍歷參數網格中的所有參數組合，對每個組合執行交叉驗證，並找到最佳的參數組合。在經歷 5 次交叉驗證後，最佳參數組合為：

- max_depth： 6
- learning_rate：0.1
- colsample_bytree： 0.8
- subsample：1.0

在挑選出最佳模型參數設定後，模型準確率從 0.65432 顯著提升至 0.69333。

8. 特徵縮放方法選擇

特徵縮放對模型有以下優點：

- 消除數值間的量級差異：當特徵的數值範圍差異很大時，機器學習模型可能會受到影響。例如，某個特徵的數值範圍在 0 到 1000 之間，而另一個特徵的數值範圍在 0 到 1 之間，這將導致模型在進行學習和預測時對於數值較大的特徵更加敏感。特徵縮放可以將所有特徵縮放到相同的範圍，消除數值間的量級差異，從而使模型更加穩定。
- 提高模型的收斂速度：在某些機器學習算法中，例如梯度下降法，特徵的縮放程度會影響模型的收斂速度。如果特徵的範圍差異很大，梯度下降法可能需要更多的迭代次數才能找到最優解。通過特徵縮放，可以加速模型的收斂速度，提高訓練效率。
- 避免特徵間的相對權重差異：某些機器學習算法，如支持向量機（SVM）和最近鄰算法（KNN），對特徵間的相對權重非常敏感。如果

某個特徵的數值範圍很大，它在模型中的權重可能比其他特徵更大，這將影響模型的性能和結果。通過特徵縮放，可以使所有特徵在模型中具有相似的重要性，避免特徵間的相對權重差異。

因此，除了模型參數外，我們也挑選了兩種特徵縮放進行比較。

1. StandardScaler

   StandardScaler 通常用於將特徵縮放為平均值為 0，標準差為 1 的分佈。它的運作方式是對每個特徵進行獨立的縮放，使得數據符合標準常態分佈。

   優點：

   - 不受異常值的影響
   - 對於大部分機器學習算法，特別是基於距離計算的算法（如 KNN、SVM 等），標準化可以提升模型的性能

   缺點：

   - 如果數據分佈不接近正態分佈，則可能不適用
   - 縮放後的數據不保留原始數據的最小值和最大值的信息

2. MinMaxScaler

   MinMaxScaler 會將特徵縮放到一個固定的範圍，通常是[0, 1]或[-1, 1]。它的運作方式是通過線性變換將特徵縮放到指定範圍內。

   優點：

   - 易於理解和實現，計算簡單
   - 縮放後的數據保留了原始數據的分佈範圍，不改變數據的分佈形狀

   缺點：

   - 對異常值敏感，異常值會影響縮放結果
   - 如果數據中存在極端值，縮放後的數據可能集中在較小的範圍內

   實驗結果：以 StandardScaler 進行縮放丟入 XGBoost 得到 0.70023 的準確率，相較使用 MinMaxScaler 得到 0.69333 的準確率有微幅的提升，推測是因為此資料含有些許極端值，使 MinMaxScaler 無法得到良好的結果。

肆、結果與未來展望

一、結論

本研究共搜集 2018 年至 2023 年五年間台南測站的空氣品質與天氣歷史資料，使用五種模型來預測台南的空氣品質，其中以 LightGBM 的預測效力最高，達到 70.29% 的準確率，其次為 XGBoost 與 LSTM，不過我們發現目前使用的五個機器學習與時間序列的模型方法皆無法很準確地預測 AQI 分級，準確率都落在 70% 上下，而由視覺化及混淆矩陣的結果可以看出模型對於預測較極端之 AQI 的效果不盡理想，猜測表現不佳的原因為特徵變量太過單一（只含有天氣資訊），不包含其他重要因素如：車流量、工業排放廢氣量等，因此期待後續能夠併入更多與 AQI 相關的資料來協助模型達到更良好的預測效果。

二、後續研究建議

根據上述研究結果，我們認為後續可藉由以下建議來分改善模型預測結果：

1. 可以併入與空氣品質相關且更多元的資料進行更深入的分析與預測，如：車流數據、工業排放數據等等
2. 嘗試使用其他時間序列模型或是機器學習、深度學習模型進行測試，如：神經網路、ARIMA 等

伍、參考文獻

[1] Avan Chowdary Gogineni, Vamsi Sri Naga Manikanta Murukonda (2022). Prediction of Air Quality Index Using Supervised Machine Learning.

[2] Samayan Bhattacharya, Sk Shahnawaz (2021). Using Machine Learning to Predict Air Quality Index in New Delhi.

[3] Cyuan Heng Luo, Hsuan Yang, Li-Pang Huang, Sachit Mahajan, Ling-Jyh Chen (2018). A first PM2.5 Forecast approach based on tome series data analysis, regression, and regularization.