銘傳大學

資訊傳播工程學系

多媒體通訊期中專題報告

題目:空汙小幫手

班級：資傳二甲

組員：06160485 曾宏鈞

06160114 黃旭雲

06160282 彭俐嘉

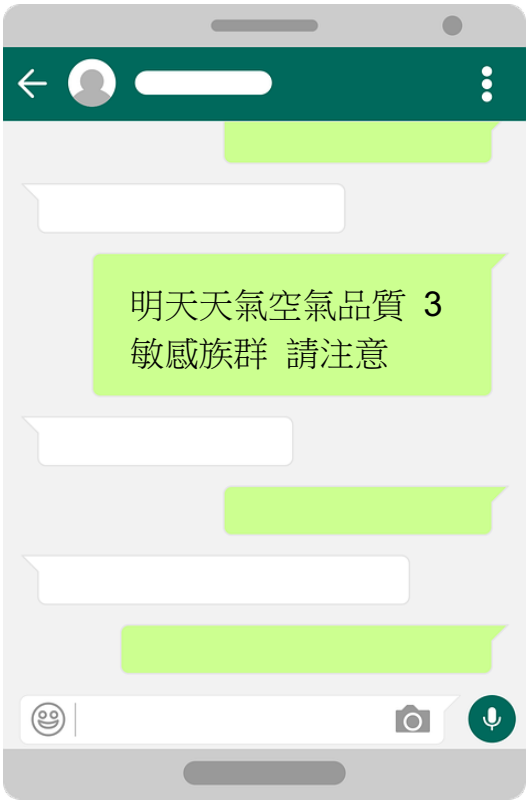
中 華 民 國 一 〇 八 年 五 月 十 五 日

目錄

一.動機

一早醒來，急忙著出門，外出後看見瀰漫的天空才驚覺空氣品質有點不好，就會再跑回宿舍拿口罩。然而，如果我們可以設計一個方法，更便捷的提醒使用者有關於空氣品質的資訊，早上一醒來就可以**主動提醒** 使用者空氣品質的資訊，就可以造福更多人，使一般民眾能更早的得知有關於當日的空氣品質資訊。

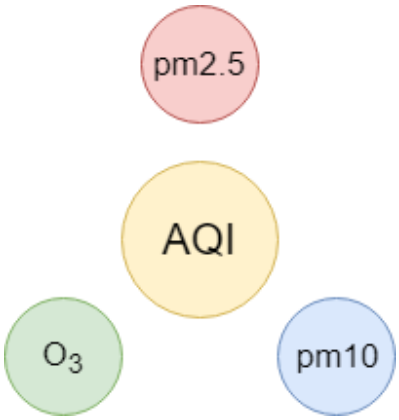
長期暴露在空氣污染下會對人體造成一系列的健康問題，主要是對呼吸系統的影響，導致健康的惡化，死亡率提升。因此我們要對 PM2.5 進行預測，提前採取防護措施，減少傷害。



二.研究方法

2.1 研究資料

一開始，我們先去找尋有關於空氣品質的資料，我們得知，目前我國空氣品質的指標為 AQI，包含了 pm10、pm2.5、0_3等測量的因子來得出一個衡量的數值。然而，因其每項測項的影響因子皆不同，所以我們著重在 pm2.5 上面的研究。



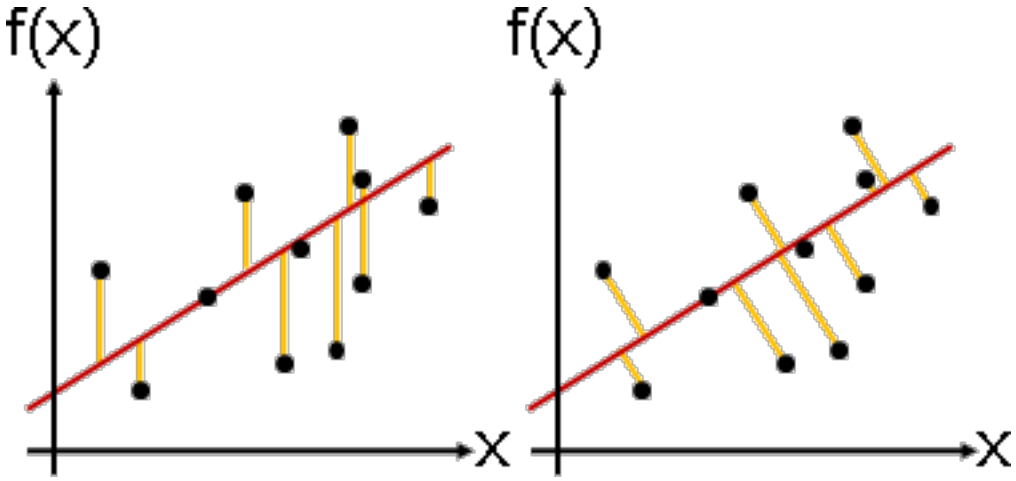
根據參考資料，pm2.5 的成因與雨水及風有關，因此我們篩選出所需的資料，來作為我們的特徵。

我們使用了線性回歸法，把台北市各區的 **pm2.5 濃度**(1.μg/m3)、**風向**(degree)、**風速**(m/sec)及**雨量**(mm)作為我們使用線性回歸法的**特徵**，使用**松山區下一個小時 pm2.5 濃度**來當作我們的**答案**，來做監督式學習的訓練。

2.2 研究資料

由於 pm2.5 的數值為連續性的資料。因此，我們使用線性回歸法來當作我們的模型，藉由前一段的特徵及答案，來達成預測未來空氣品質的目標。

線性回歸



基本數學模型：$Y=\alpha X+\beta$

Y = 答案

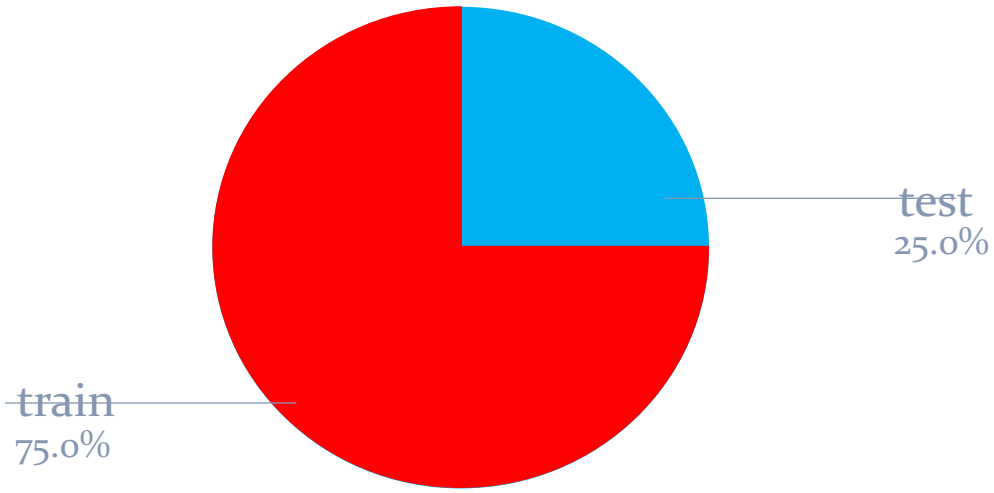
X = 每一項的特徵

α = x 特徵的係數

β = y 截距

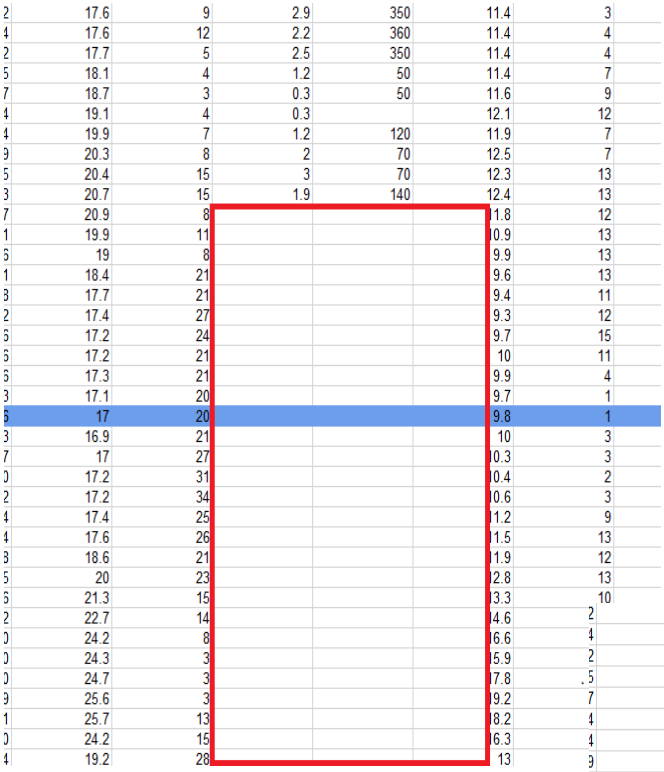
資料的介紹

我們將資料拆成兩個部分，訓練集及測試集。各為 75%及 25%，使用隨機拆分的方法。

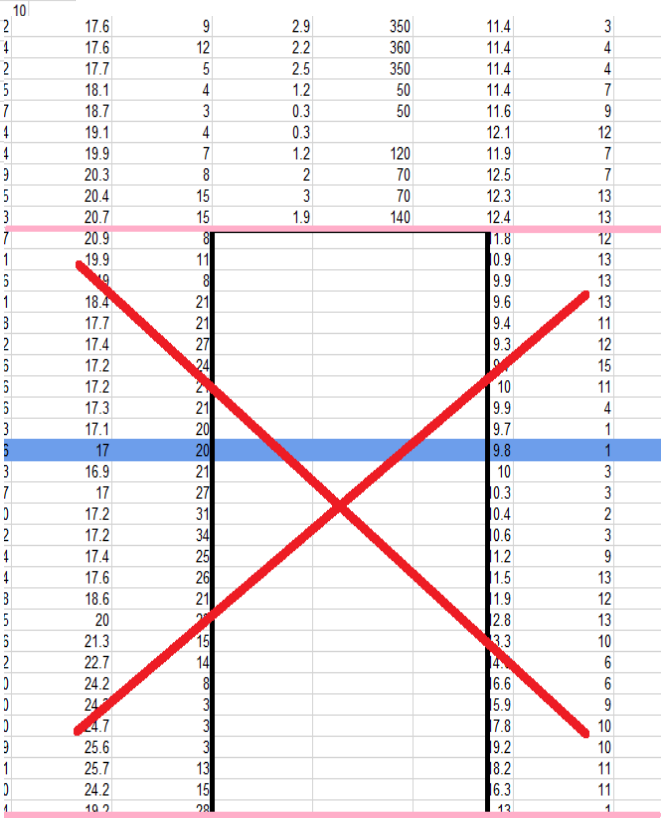


空值問題

由於測站的資料(雨量)有可能會有空值，因此我們針對資料做兩種不同的處理，並且輸出成 2 個模型。



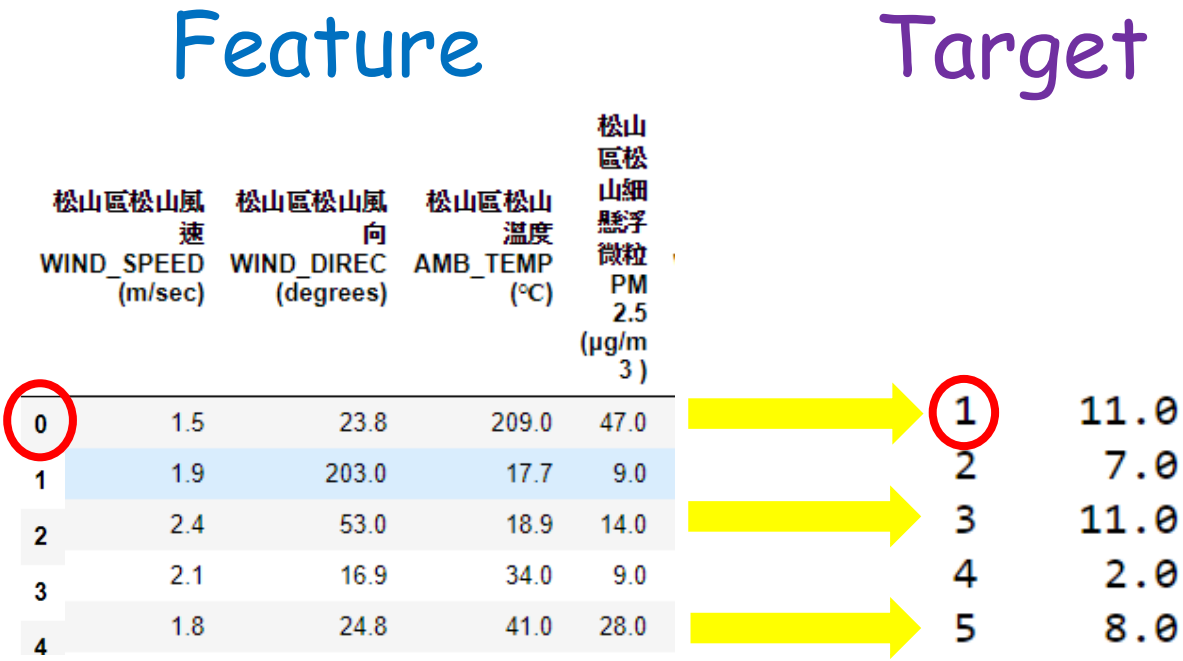
全部補 0 及丟掉那一行，並且輸出成 model1 和 model2。



後續會根據此做進一步的分析。

資料前處理

由於我們是使用前一筆去預測後一筆，因此必須要將資料做處理，如下圖：



正規化

由於資料有不同的範圍及離散程度，因此我們使用了正規化，即為減掉該項特徵平均再除以 L2 範數。

評估模型

模型評估我們使用兩個算法來評估我們的模型，分別是 MSE(mean square error)以及 $\boldsymbol{R^2}$，公式如下：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

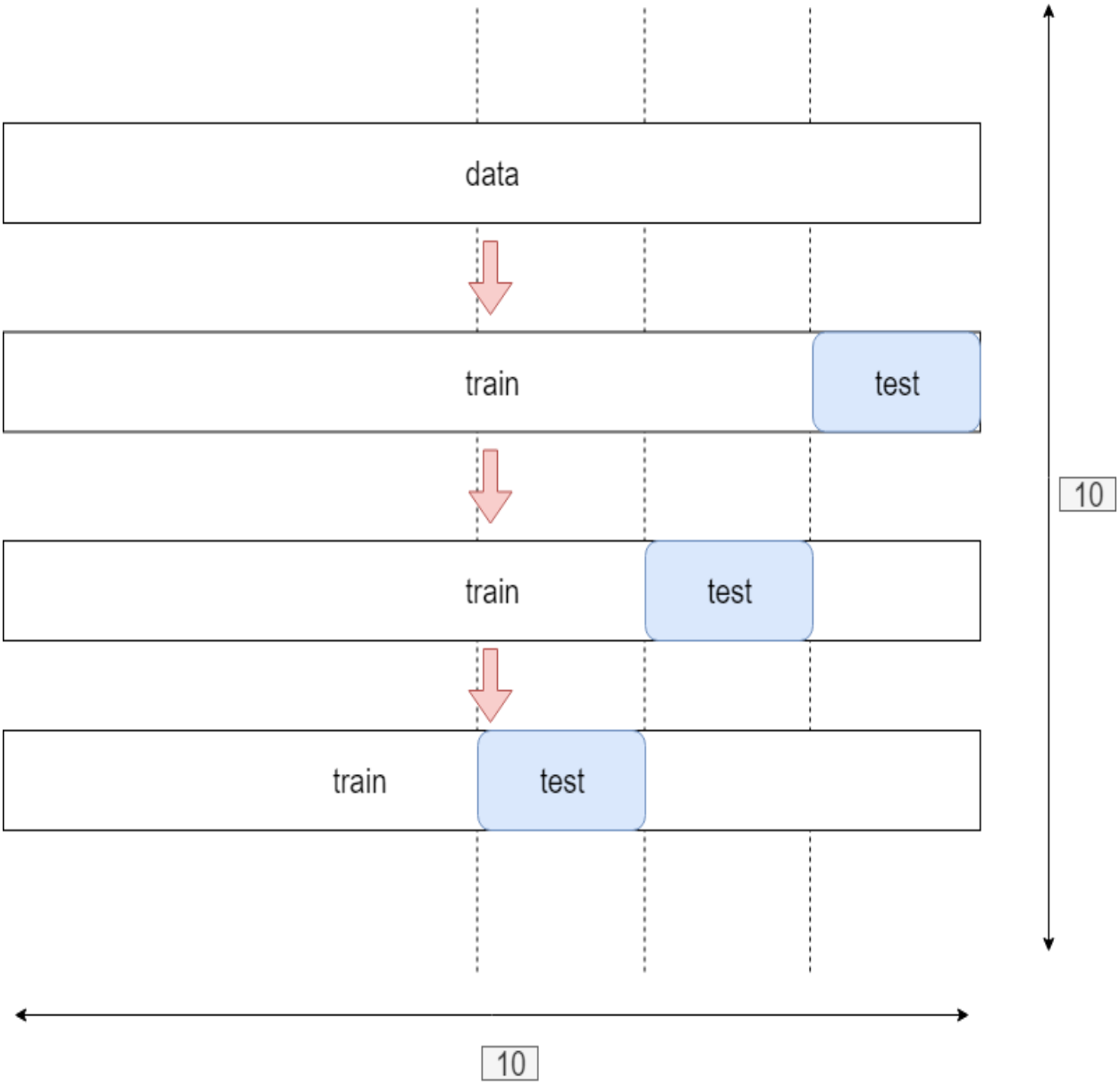
$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

MSE 方法，誤差最小是 0

$\boldsymbol{R^2}$ 最好是 1 最差是 0，介於 0-1 中間

K-Fold Cross Validation

由於我們希望找出訓練集的哪一段的資料拿去訓練準確率會最高，因此我們使用 K-Fold Cross Validation 來找出此區間。



三.實驗

3-1 介紹

Anaconda 虛擬環境管理器，內含許多的機器學習及數據處理的函式，並使用 conda 來做管理。

我們使用 Python 是一種直譯器的程式語言，因其許多人開發套件，再加上物件導向語言的特性，因此是目前在機器學習領域熱門的程式語言之一。

scikit learn 有許多機器學習的函式，方便我們使用。

Pandas 處理數據很方便的套件；

Request 使用其他 API 的服務

3-2 程式碼講解

匯入模組

```
from sklearn.linear_model import LinearRegression
```

建立線性回歸模型

```
lm = LinearRegression()
X = feature
Y = target
lm.fit(X,Y)
```

預測

```
lm.predict(X_test)
```

係數

```
print('coef is:',lm.coef_)
```

畫圖

```
plt.scatter(X 軸方向數字, Y 軸方向數字)
plt.xlabel('X 軸標籤')
plt.ylabel(' X 軸標籤')
plt.title('標題')
plt.show()
```

提醒

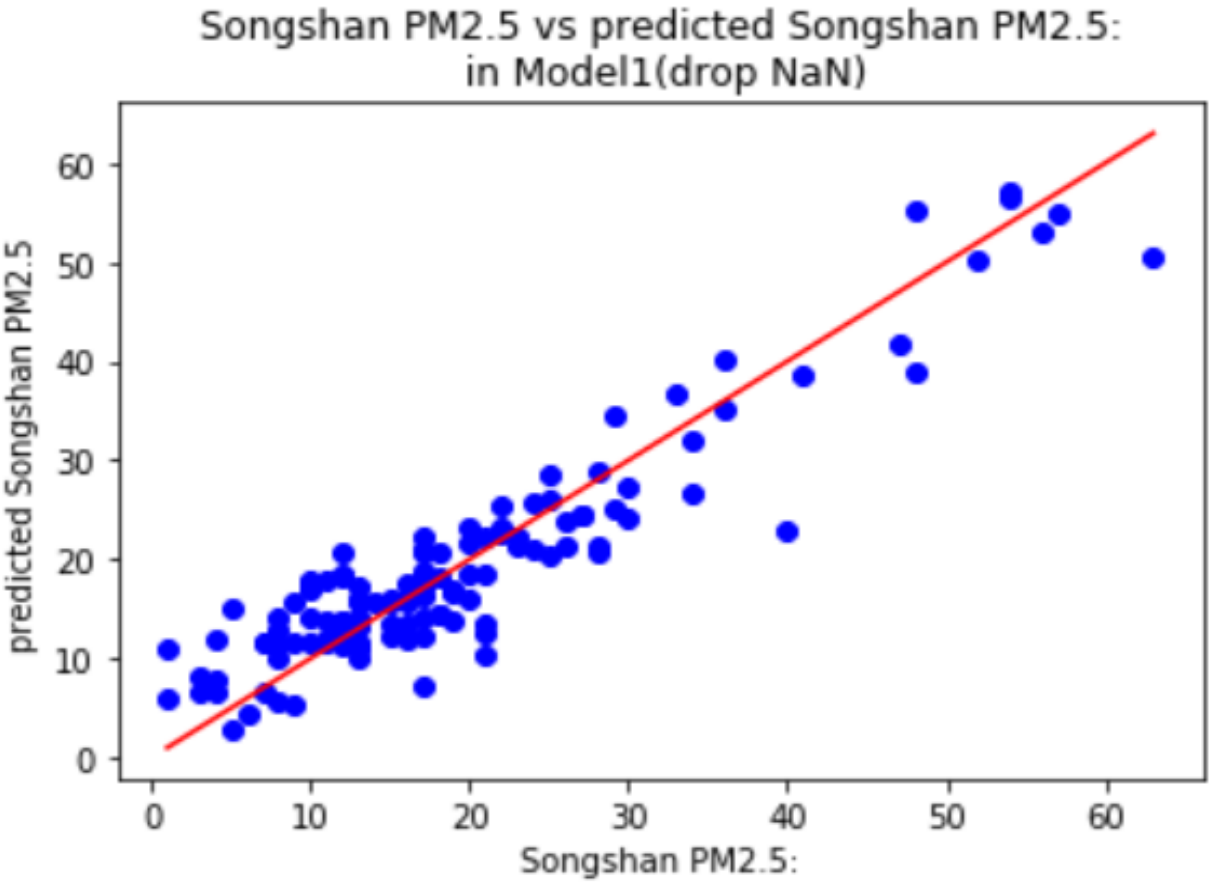
```
path='https://maker.ifttt.com/trigger/noticePM25/with/key/{your_key}
body={
    'value1':' 松山區 ',
    'value2':answer,
    'value3':message
}
r = requests.post(path, data = body)
print(r.text)
```

*更多完整程式碼請參考附錄

3-3 改進的部分

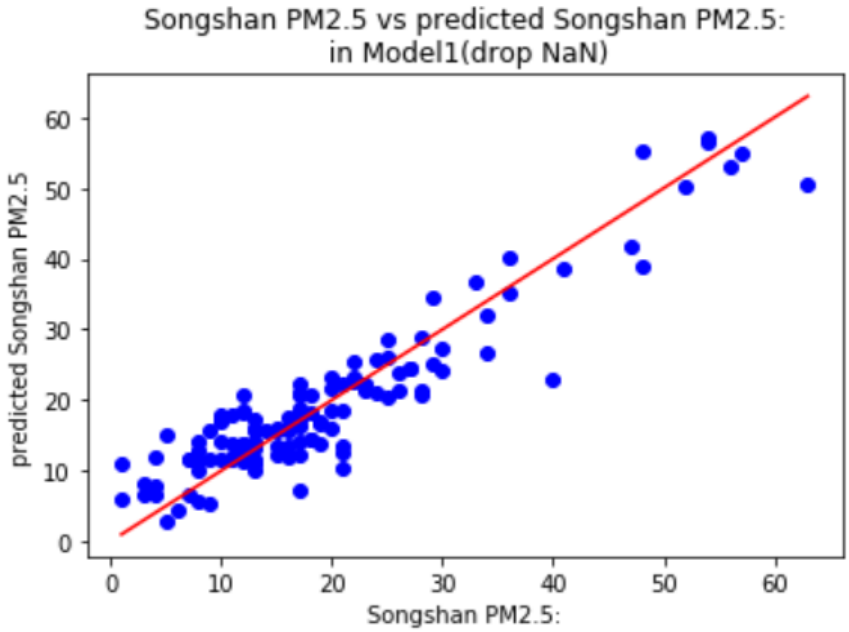
Q 資料視覺化

A 已補足資料視覺化的部分



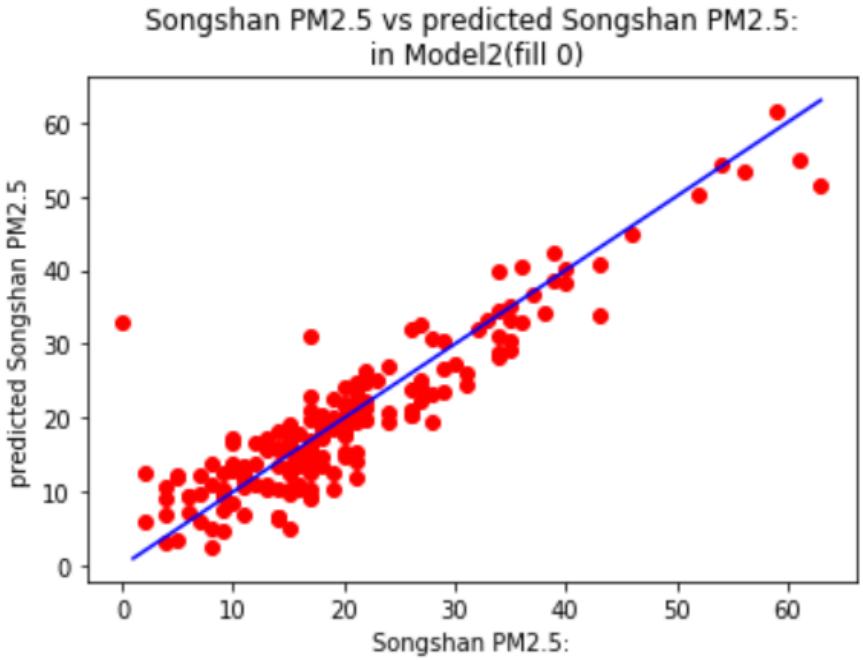
3-4 結果分析

我們的資料在 0～30 比較多，因此若在此範圍會預測較準確。



Model**1**：

丟掉空值



Model2：

補 0

	Model 1 **(Drop NaN)**	**Model 2** **(Fill 0)**
drop	**O**	**X**
MSE	**21.34**	**22.76**
$\boldsymbol{R^2}$	**0.85215**	**0.83438**

若直接算 MSE、 $\boldsymbol{R^2}$，model1 的準確率比較高

	feature	estimatedCoefficients
0	中山區中山風速 WIND_SPEED (m/sec)	0.425141
1	中山區中山風向 WIND_DIREC (degrees)	0.006312
2	中山區中山溫度 AMB_TEMP (°C)	-0.008180
3	中山區中山細懸浮微粒 PM 2.5 (μg/m 3)	0.065534
4	北投區陽明風速 WIND_SPEED (m/sec)	-0.057287
5	北投區陽明風向 WIND_DIREC (degrees)	0.001974
6	北投區陽明溫度 AMB_TEMP (°C)	0.001105
7	北投區陽明細懸浮微粒 PM 2.5 (μg/m 3)	0.108967
8	北投區士林風速 WIND_SPEED (m/sec)	-0.176550
9	北投區士林風向 WIND_DIREC (degrees)	0.002480
10	北投區士林溫度 AMB_TEMP (°C)	0.971418
11	北投區士林細懸浮微粒 PM 2.5 (μg/m 3)	0.026034
12	萬華區萬華風速 WIND_SPEED (m/sec)	0.003119
13	萬華區萬華風向 WIND_DIREC (degrees)	0.004734
14	萬華區萬華溫度 AMB_TEMP (°C)	-0.886350
15	萬華區萬華細懸浮微粒 PM 2.5 (μg/m 3)	0.132703
16	松山區松山風速 WIND_SPEED (m/sec)	-0.150628
17	松山區松山風向 WIND_DIREC (degrees)	0.001378
18	松山區松山溫度 AMB_TEMP (°C)	0.001760
19	松山區松山細懸浮微粒 PM 2.5 (μg/m 3)	0.457837
20	大安區古亭風速 WIND_SPEED (m/sec)	-0.616736
21	大安區古亭風向 WIND_DIREC (degrees)	0.001128
22	大安區古亭溫度 AMB_TEMP (°C)	0.000335
23	大安區古亭細懸浮微粒 PM 2.5 (μg/m 3)	0.092640
24	大同區大同細懸浮微粒 PM 2.5 (μg/m 3)	0.044637

	feature	estimatedCoefficients
0	中山區中山風速 WIND_SPEED (m/sec)	0.370677
1	中山區中山風向 WIND_DIREC (degrees)	0.004366
2	中山區中山溫度 AMB_TEMP (°C)	0.021780
3	中山區中山細懸浮微粒 PM 2.5 (μg/m 3)	0.024032
4	北投區陽明風速 WIND_SPEED (m/sec)	0.023676
5	北投區陽明風向 WIND_DIREC (degrees)	0.000258
6	北投區陽明溫度 AMB_TEMP (°C)	-0.010177
7	北投區陽明細懸浮微粒 PM 2.5 (μg/m 3)	0.015562
8	北投區士林風速 WIND_SPEED (m/sec)	0.170713
9	北投區士林風向 WIND_DIREC (degrees)	0.001730
10	北投區士林溫度 AMB_TEMP (°C)	0.059548
11	北投區士林細懸浮微粒 PM 2.5 (μg/m 3)	0.077631
12	萬華區萬華風速 WIND_SPEED (m/sec)	-0.007025
13	萬華區萬華風向 WIND_DIREC (degrees)	-0.001136
14	萬華區萬華溫度 AMB_TEMP (°C)	-0.059012
15	萬華區萬華細懸浮微粒 PM 2.5 (μg/m 3)	0.074166
16	松山區松山風速 WIND_SPEED (m/sec)	-0.164906
17	松山區松山風向 WIND_DIREC (degrees)	0.005609
18	松山區松山溫度 AMB_TEMP (°C)	0.011365
19	松山區松山細懸浮微粒 PM 2.5 (μg/m 3)	0.583124
20	大安區古亭風速 WIND_SPEED (m/sec)	-0.005864
21	大安區古亭風向 WIND_DIREC (degrees)	-0.002120
22	大安區古亭溫度 AMB_TEMP (°C)	0.007566
23	大安區古亭細懸浮微粒 PM 2.5 (μg/m 3)	0.127641
24	大同區大同細懸浮微粒 PM 2.5 (μg/m 3)	0.018937

Model1 的相關係數最高的是士林區的溫度，而 Model2 的相關係數最高的是松山區的細懸浮微粒，根據常識，我們會先採用 model2 來做預測，但為何 model1 會造成這種結果?我們在繼續使用 **K-Fold Cross Validation NMSE**

K-Fold Cross Validation NMSE

Model1

```
[-26.71379162 -12.4015139  -25.3515253  -16.95141041 -24.96383944
 -26.49745983 -22.75309958 -21.60589266 -13.94001151 -24.24012578]
mean mse:-21.54186700327143
```

Model2

```
[-19.90431352 -31.85432236 -19.29704058 -22.12155725 -18.02283207
 -25.26631013 -25.47245394 -22.7819572  -18.13430009 -18.74802375]
```

Model1 的數值約在-12~-26 之間

Model2 的數值約在-18~-31 之間

在使用 NMSE 時是 Model1 的評分比較好

R^2

Model1

```
[0.85271729 0.87944734 0.83117563 0.83656947 0.4958474  0.78801908
 0.79915394 0.79306628 0.88326503 0.85163242]
```

Model2

```
[0.82163238 0.8151866  0.91533627 0.77052979 0.86005427 0.78947641
 0.78665755 0.80214003 0.82490262 0.85717761]
```

Model1 的數值約在 0.49~0.88 之間

Model2 的數值約在 0.77~0.91 之間

在使用R^2時反而是 Model2 的評分比較好

實驗成果


搜尋聊天和訊息
LINE Notify
下午 7:17
【IFTTT】下一個小時 松山區
的pm2.5濃度: 15 PM 2.5 (μ...
昨天
昨天
LINE Notify
今天
以下為尚未閱讀的訊息
【IFTTT】下一個小時 松山區 的pm2.5濃度: 15 PM 2.5 (μg/m 3)
空氣品質良好喔~
下午 7:17

四. 結論與未來工作

這次的作業讓我們瞭解收集資料、整理資料、參考資料的重要性，理論與實務的重要性，勇於發問和求證不懈的毅力。

老師認真負責對待學生的態度，使我們對這次的作業加倍加倍加倍用心，或許我們仍有許多的不足，但我們仍會秉持初心，認真對待每一次的作業，努力解決每一個遇到的困難。

五. 附錄

上台報告實作程式碼

https://github.com/alanhc/Taipei_pm2_5

參考文獻

台北市政府資料開放平台

https://data.taipei/dataset/detail/metadata?id=4ba06157-3854-4111-9383-3b8a188c962a

空氣品質預估方法(AQI)

https://taqm.epa.gov.tw/taqm/tw/b0203.aspx

台灣空氣汙染

https://zh.wikipedia.org/wiki/%E8%87%BA%E7%81%A3%E7%A9%BA%E6%B0%A3%E6%B1%A1%E6%9F%93#%E5%8F%B0%E7%81%A3%E7%A9%BA%E6%B0%A3%E6%B1%99%E6%9F%93%E7%9A%84%E5%9C%B0%E7%90%86%E5%9B%A0%E7%B4%A0

氣象資料開放平台

https://opendata.cwb.gov.tw/index

linear regression

http://www.csie.ntnu.edu.tw/~u91029/Regression.html#1
https://blog.csdn.net/zhaoyuxia517/article/details/78108805
https://bigdata-madesimple.com/how-to-run-linear-regression-in-python-scikit-learn/
https://www.dataquest.io/blog/learning-curves-machine-learning/
https://www.ycc.idv.tw/ml-course-foundations_4.html
https://scikit-learn.org/stable/modules/sgd.html
https://morvanzhou.github.io/tutorials/machine-learning/sklearn/3-5-save/
https://machine-learning-python.kspax.io/general_examplesmd/ex1_plotting_cross-validated_predictions
https://scikit-learn.org/stable/modules/cross_validation.html

數學

https://medium.com/@ken90242/machine-learning%E5%AD%B8%E7%BF%92%E6%97%A5%E8%A8%98-coursera%E7%AF%87-week-3-4-the-c05b8ba3b36f
https://statisticsbyjim.com/glossary/regression-coefficient/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5356327/
https://medium.com/@chih.sheng.huang821/%E7%B7%9A%E6%80%A7%E5%9B%9E%E6%AD%B8-linear-regression-3a271a7453e

line

https://www.oxxostudio.tw/articles/201804/line-bot-apps-script.html
https://ithelp.ithome.com.tw/articles/10193441
https://stackoverflow.com/questions/4828214/is-it-possible-to-start-a-timer-in-google-app-engine
https://medium.com/@earlg3/google-cloud-functions-scheduled-trigger-915b5fb8310f