- 資料前處理

```
In [1]: #coding = utf-8
        #Loading package
        import pandas as pd
        import math
        import numpy as np
        import datetime

        #Loading data
        data = pd.read_csv("新竹_2020.csv",encoding = "utf-8")
        data
```

Out[1]:

	測站	日期	測項	0	1	2	3	4	5	6	...	14	15	16	17	18	19	20	21	22	23
0	新竹	2020/1/1 00:00	AMB_TEMP	15.2	15.2	15.3	15.3	15.3	15.4	15.5	...	18.1	18.2	17.9	17.3	16.7	16.4	16.2	16.1	16	15.8
1	新竹	2020/1/1 00:00	CH4	1.74	1.74	1.77	1.78	1.77	1.77	1.77	...	1.78	1.78	1.77	1.8	1.81	1.82	1.85	1.83	1.92	1.94
2	新竹	2020/1/1 00:00	CO	0.28	0.25	0.24	0.22	0.2	0.19	0.2	...	0.28	0.29	0.28	0.34	0.39	0.41	0.46	0.49	0.58	0.52
3	新竹	2020/1/1 00:00	NMHC	0.06	0.07	0.05	0.05	0.05	0.05	0.07	...	0.09	0.09	0.07	0.08	0.12	0.12	0.16	0.14	0.17	0.2
4	新竹	2020/1/1 00:00	NO	0.3	0.6	0.6	0.6	0.3	0.3	0.5	...	1.6	1.6	1.2	0.7	0.9	1.1	1.1	1.7	1.8	1.4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6583	新竹	2020/12/31 00:00	THC	2.01	2.02	2	2	1.99	2	1.98	...	2.03	2.07	2.07	2.1	2.1	2.07	2.07	2.05	2.04	2.07
6584	新竹	2020/12/31 00:00	WD_HR	54	55	54	53	58	52	52	...	54	50	52	45	47	42	42	47	45	44
6585	新竹	2020/12/31 00:00	WIND_DIREC	53	52	57	58	49	54	36	...	48	43	44	33	50	40	46	46	51	38
6586	新竹	2020/12/31 00:00	WIND_SPEED	4.7	4.6	4.7	4.9	4.1	5.3	5.5	...	4.5	4.4	4.2	3.8	3.7	4.7	4.5	4.4	3.9	3.9
6587	新竹	2020/12/31 00:00	WS_HR	3.7	3.6	3.6	3.5	3.5	3.3	3.8	...	3.7	3.1	3.3	3.1	2.9	3.3	3.1	2.9	2.8	2.6

6588 rows × 27 columns

把 csv 檔用 utf-8 編碼後便可以進到 DataFrame 裡面且將 column name 保持中文不亂碼。

```
In [3]: data["日期"]= pd.to_datetime(data["日期"],format ="%Y/%m/%d  %H:%M")
        data["日期"]
```

```
Out[3]: 0       2020-01-01
        1       2020-01-01
        2       2020-01-01
        3       2020-01-01
        4       2020-01-01
                   ...
        6583    2020-12-31
        6584    2020-12-31
        6585    2020-12-31
        6586    2020-12-31
        6587    2020-12-31
        Name: 日期, Length: 6588, dtype: datetime64[ns]
```

將日期照著格式轉換後並透過篩選(10/1-12/31)把需要的一千多筆資料轉存。

```
In [4]: date_data = data[(data["日期"]>="2020-10-01")&(data["日期"]<="2020-12-31")]
        date_data
```

Out[4]:

	測站	日期	測項	0	1	2	3	4	5	6	...	14	15	16	17	18	19	20	21	22	23
4932	新竹	2020-10-01	AMB_TEMP	23.7	23.8	23.8	23.9	23.9	23.8	24.1	...	29.9	29.6	28.7	27.5	26.4	25.7	25.5	25.3	24.9	24.5
4933	新竹	2020-10-01	CH4	1.97	1.95	1.96	1.96	1.95	1.96	1.97	...	1.97	1.98	1.97	2	2.03	2.04	2.05	2.02	2.1	2.14
4934	新竹	2020-10-01	CO	0.23	0.22	0.21	0.2	0.2	0.22	0.24	...	0.29	0.3	0.33	0.38	0.46	0.5	0.45	0.39	0.46	0.45
4935	新竹	2020-10-01	NMHC	0.06	0.05	0.03	0.03	0.03	0.04	0.04	...	0.06	0.07	0.09	0.11	0.13	0.15	0.1	0.07	0.12	0.18
4936	新竹	2020-10-01	NO	1.2	0.7	0.5	0.7	0.5	0.3	0.7	...	1.3	1	0.9	0.8	0.5	0.9	0.9	0.3	0.7	0.9
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6583	新竹	2020-12-31	THC	2.01	2.02	2	2	1.99	2	1.98	...	2.03	2.07	2.07	2.1	2.1	2.07	2.07	2.05	2.04	2.07
6584	新竹	2020-12-31	WD_HR	54	55	54	53	58	52	52	...	54	50	52	45	47	42	42	47	45	44
6585	新竹	2020-12-31	WIND_DIREC	53	52	57	58	49	54	36	...	48	43	44	33	50	40	46	46	51	38
6586	新竹	2020-12-31	WIND_SPEED	4.7	4.6	4.7	4.9	4.1	5.3	5.5	...	4.5	4.4	4.2	3.8	3.7	4.7	4.5	4.4	3.9	3.9

```
In [8]: filled_data = fill_missing(nan_data)
        filled_data
```

Out[8]:

	測站	日期	測項	0	1	2	3	4	5	6	...	14	15	16	17	18	19	20	21	22	23
4932	新竹	2020-10-01	AMB_TEMP	23.70	23.80	23.80	23.90	23.90	23.80	24.10	...	29.90	29.60	28.70	27.50	26.40	25.70	25.50	25.30	24.90	24.50
4933	新竹	2020-10-01	CH4	1.97	1.95	1.96	1.96	1.95	1.96	1.97	...	1.97	1.98	1.97	2.00	2.03	2.04	2.05	2.02	2.10	2.14
4934	新竹	2020-10-01	CO	0.23	0.22	0.21	0.20	0.20	0.22	0.24	...	0.29	0.30	0.33	0.38	0.46	0.50	0.45	0.39	0.46	0.45
4935	新竹	2020-10-01	NMHC	0.06	0.05	0.03	0.03	0.03	0.04	0.04	...	0.06	0.07	0.09	0.11	0.13	0.15	0.10	0.07	0.12	0.18
4936	新竹	2020-10-01	NO	1.20	0.70	0.50	0.70	0.50	0.30	0.70	...	1.30	1.00	0.90	0.80	0.50	0.90	0.90	0.30	0.70	0.90
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6583	新竹	2020-12-31	THC	2.01	2.02	2.00	2.00	1.99	2.00	1.98	...	2.03	2.07	2.07	2.10	2.10	2.07	2.07	2.05	2.04	2.07
6584	新竹	2020-12-31	WD_HR	54.00	55.00	54.00	53.00	58.00	52.00	52.00	...	54.00	50.00	52.00	45.00	47.00	42.00	42.00	47.00	45.00	44.00
6585	新竹	2020-12-31	WIND_DIREC	53.00	52.00	57.00	58.00	49.00	54.00	36.00	...	48.00	43.00	44.00	33.00	50.00	40.00	46.00	46.00	51.00	38.00
6586	新竹	2020-12-31	WIND_SPEED	4.70	4.60	4.70	4.90	4.10	5.30	5.50	...	4.50	4.40	4.20	3.80	3.70	4.70	4.50	4.40	3.90	3.90
6587	新竹	2020-12-31	WS_HR	3.70	3.60	3.60	3.50	3.50	3.30	3.80	...	3.70	3.10	3.30	3.10	2.90	3.30	3.10	2.90	2.80	2.60

1656 rows × 27 columns

寫了一個 function(fill_missing)，功能大致就是取 missing value 的前後(直到有值)做平均。但在執行的過程中發現後面某一項會往後找到沒有值(到 12/31 23:00)，因此多設了一個限制，超過限定範圍則取前或後有值的為其值。

- 時間序列

  比較下列 MAE

```
In [15]: from sklearn.linear_model import LinearRegression
         from sklearn.metrics import mean_absolute_error
         regressor = LinearRegression()
         regressor.fit(x_train, y_train)
         y_pred = regressor.predict(x_test)
         mae = mean_absolute_error(y_test,y_pred)
         #regression,預測一小時後，只看pm2.5特徵
         mae

Out[15]: 2.5223536456517683
```

特徵採用僅 pm2.5，使用 Linear Regression，預測一小時後的 pm2.5 值的 MAE。

```
In [16]: from xgboost import XGBClassifier
         xgbc = XGBClassifier()
         xgbc.fit(x_train,y_train)
         y_pred = xgbc.predict(x_test)
         #xgboost,預測一小時後，只看pm2.5
         mae = mean_absolute_error(y_test,y_pred)
         mae

         C:\Users\Home\anaconda3\lib\site-packages\xgboos
         recated and will be removed in a future release.
         se when constructing XGBClassifier object; and 2
         lass - 1].
           warnings.warn(label_encoder_deprecation_msg, U

         [07:22:14] WARNING: ..\src\learner.cc:1061: Star
         ulti:softprob' was changed from 'merror' to 'mlo

Out[16]: 3.176151761517615
```

特徵採用僅 pm2.5，使用 XGBoost，預測一小時後的 pm2.5 值的 MAE。

(不知道為甚麼好像 xgboost 在 jupyter notebook 上說即將被刪除)

```
In [19]: regressor = LinearRegression()
         regressor.fit(x_train, y_train)
         y_pred = regressor.predict(x_test)
         mae = mean_absolute_error(y_test,y_pred)
         #regression,預測六小時後，只看pm2.5特徵
         mae

Out[19]: 4.579414220758536
```

特徵採用僅 pm2.5，使用 Linear Regression，預測六小時後的 pm2.5 值的 MAE。

In [20]:
```
xgbc = XGBClassifier()
xgbc.fit(x_train,y_train)
y_pred = xgbc.predict(x_test)
#xgboost,預測六小時後，只看pm2.5
mae = mean_absolute_error(y_test,y_pred)
mae
```

```
[07:22:18] WARNING: ..\src\learner.cc:1061: S
ulti:softprob' was changed from 'merror' to '
```

```
C:\Users\Home\anaconda3\lib\site-packages\xgb
recated and will be removed in a future relea
se when constructing XGBClassifier object; an
lass - 1].
  warnings.warn(label_encoder_deprecation_msg
```

Out[20]:
```
5.34174624829468
```

特徵採用僅 pm2.5，使用 XGBoost，預測六小時後的 pm2.5 值的 MAE。

In [23]:
```
regressor = LinearRegression()
regressor.fit(x_train, y_train)
y_pred = regressor.predict(x_test)
mae = mean_absolute_error(y_test,y_pred)
#regression,預測一小時後，看全部特徵
mae
```

Out[23]:
```
2.718420294073232
```

全部特徵皆採用，使用 Linear Regression，預測一小時後的 pm2.5 值的 MAE。

```
In [24]: xgbc = XGBClassifier()
         xgbc.fit(x_train,y_train)
         y_pred = xgbc.predict(x_test)
         #xgboost,預測一小時後，看全部
         mae = mean_absolute_error(y_test,y_pred)
         mae
```

```
[07:22:38] WARNING: ..\src\learner.cc:1061: Star
ulti:softprob' was changed from 'merror' to 'mlo

C:\Users\Home\anaconda3\lib\site-packages\xgboos
recated and will be removed in a future release.
se when constructing XGBClassifier object; and 2
lass - 1].
  warnings.warn(label_encoder_deprecation_msg, U
```

```
Out[24]: 3.292682926829268
```

全部特徵皆採用，使用 XGBoost，預測一小時後的 pm2.5 值的 MAE。

```
In [27]: regressor = LinearRegression()
         regressor.fit(x_train, y_train)
         y_pred = regressor.predict(x_test)
         mae = mean_absolute_error(y_test,y_pred)
         #regression,預測六小時後，看全部特徵
         mae
```

```
Out[27]: 5.749477319528851
```

全部特徵皆採用，使用 Linear Regression，預測六小時後的 pm2.5 值的 MAE。

```
mae = mean_absolute_error(y_test,y_pred)
mae
```

```
[07:23:00] WARNING: ..\src\learner.cc:1061: Starting
ulti:softprob' was changed from 'merror' to 'mloglos

C:\Users\Home\anaconda3\lib\site-packages\xgboost\sk
recated and will be removed in a future release. To
se when constructing XGBClassifier object; and 2) En
lass - 1].
  warnings.warn(label_encoder_deprecation_msg, UserW
```

```
Out[28]: 5.066848567530696
```

全部特徵皆採用，使用 XGBoost，預測六小時後的 pm2.5 值的 MAE。

當預測時間拉長，MAE 皆會變大；但在 XGBoost 跟 Linear Regression 的比較上可能參數調整或資料內容有關，Linear Regression 大致表現的都比較好，XGBoost 只有在預測六小時、採用全特徵時贏過 Linear Regression。