2025 Fall Data Mining

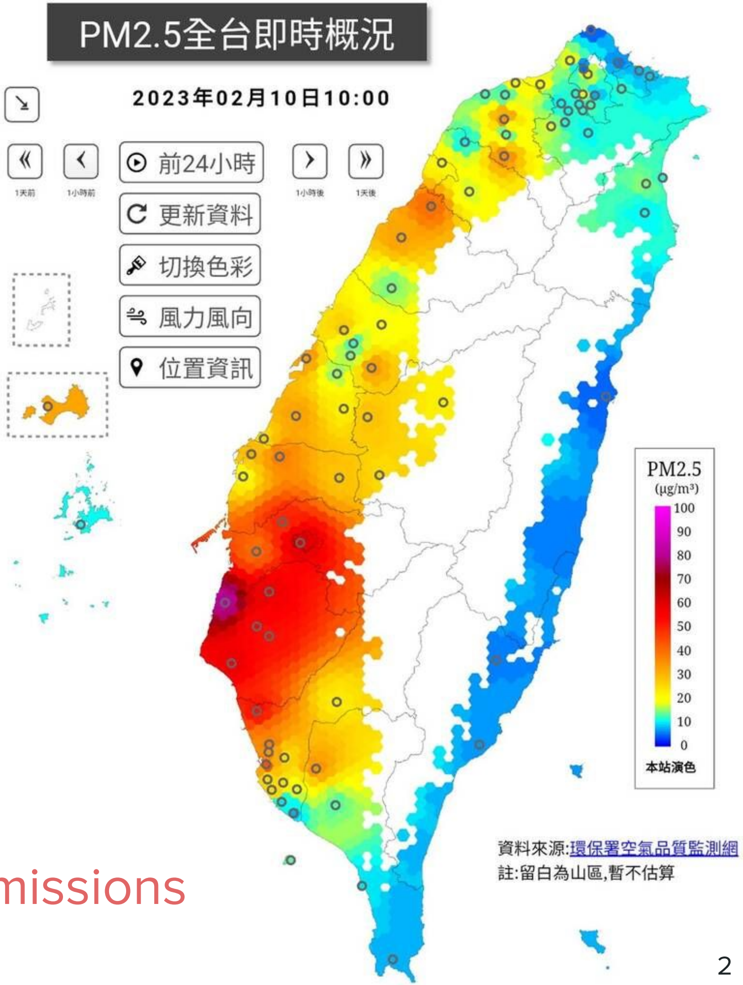
HW 1

Task introduction

- PM2.5 prediction
  - Implement linear regression using only Numpy to predict PM2.5 values
  - You may use pandas, CSV and matplotlib for data analysis and preprocessing

- Requirement
  - Upload your submission to Kaggle
  - Submit a report and your source code to E3

- Deadine is 10/8 (Wed.) 23:59, no late submissions



PM2.5全台即時概況

2023年02月10日10:00

資料來源:環保署空氣品質監測網

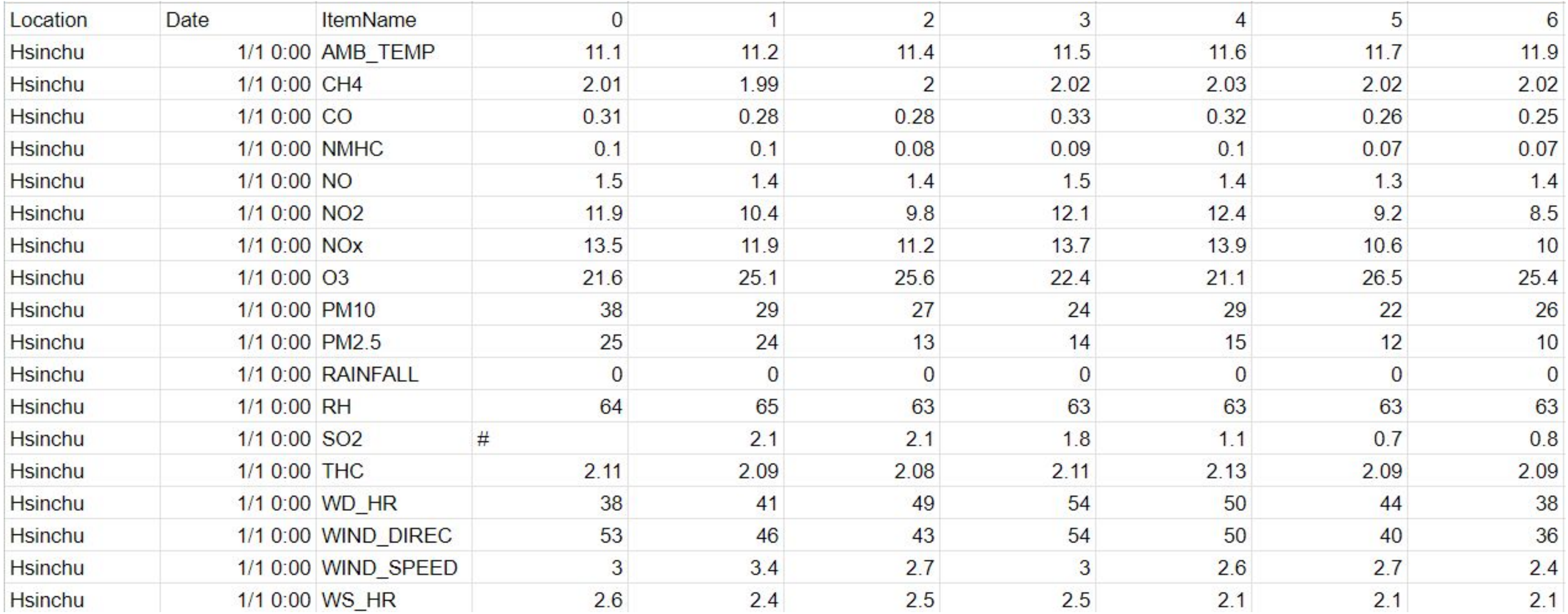
註:留白為山區,暫不估算

Dataset

Hsinchu meteorological observation data form Central Weather Bureau.

- train.csv
  - Climate data for the first 20 days of each month.
  - link
- test.csv
  - Consists of continuous 10-hours samples from the remaining 10 days of each month.
  - For each sample, the first 9 hours are used as features, and the PM 2.5 value in the 10th hour is used as the target.
  - link

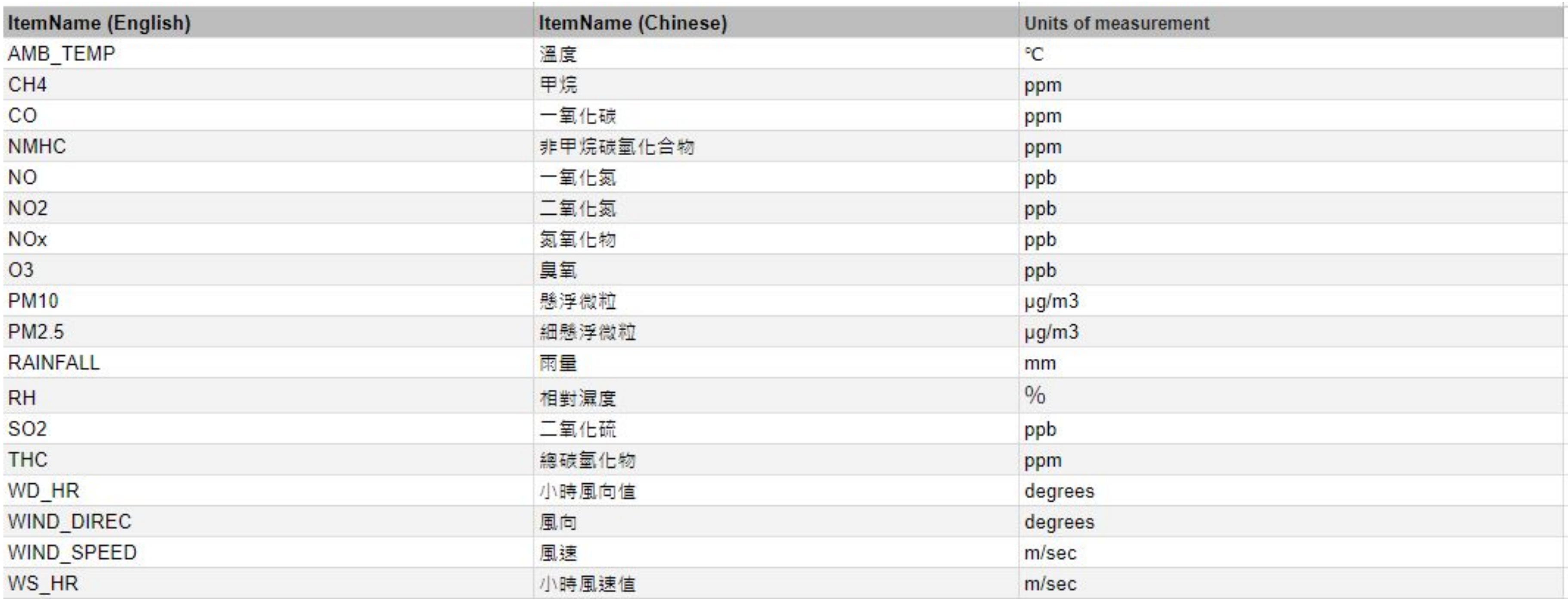
Training Data



Location	Date	ItemName	0	1	2	3	4	5	6
Hsinchu	1/1 0:00	AMB_TEMP	11.1	11.2	11.4	11.5	11.6	11.7	11.9
Hsinchu	1/1 0:00	CH4	2.01	1.99	2	2.02	2.03	2.02	2.02
Hsinchu	1/1 0:00	CO	0.31	0.28	0.28	0.33	0.32	0.26	0.25
Hsinchu	1/1 0:00	NMHC	0.1	0.1	0.08	0.09	0.1	0.07	0.07
Hsinchu	1/1 0:00	NO	1.5	1.4	1.4	1.5	1.4	1.3	1.4
Hsinchu	1/1 0:00	NO2	11.9	10.4	9.8	12.1	12.4	9.2	8.5
Hsinchu	1/1 0:00	NOx	13.5	11.9	11.2	13.7	13.9	10.6	10
Hsinchu	1/1 0:00	O3	21.6	25.1	25.6	22.4	21.1	26.5	25.4
Hsinchu	1/1 0:00	PM10	38	29	27	24	29	22	26
Hsinchu	1/1 0:00	PM2.5	25	24	13	14	15	12	10
Hsinchu	1/1 0:00	RAINFALL	0	0	0	0	0	0	0
Hsinchu	1/1 0:00	RH	64	65	63	63	63	63	63
Hsinchu	1/1 0:00	SO2	#	2.1	2.1	1.8	1.1	0.7	0.8
Hsinchu	1/1 0:00	THC	2.11	2.09	2.08	2.11	2.13	2.09	2.09
Hsinchu	1/1 0:00	WD_HR	38	41	49	54	50	44	38
Hsinchu	1/1 0:00	WIND_DIREC	53	46	43	54	50	40	36
Hsinchu	1/1 0:00	WIND_SPEED	3	3.4	2.7	3	2.6	2.7	2.4
Hsinchu	1/1 0:00	WS_HR	2.6	2.4	2.5	2.5	2.1	2.1	2.1

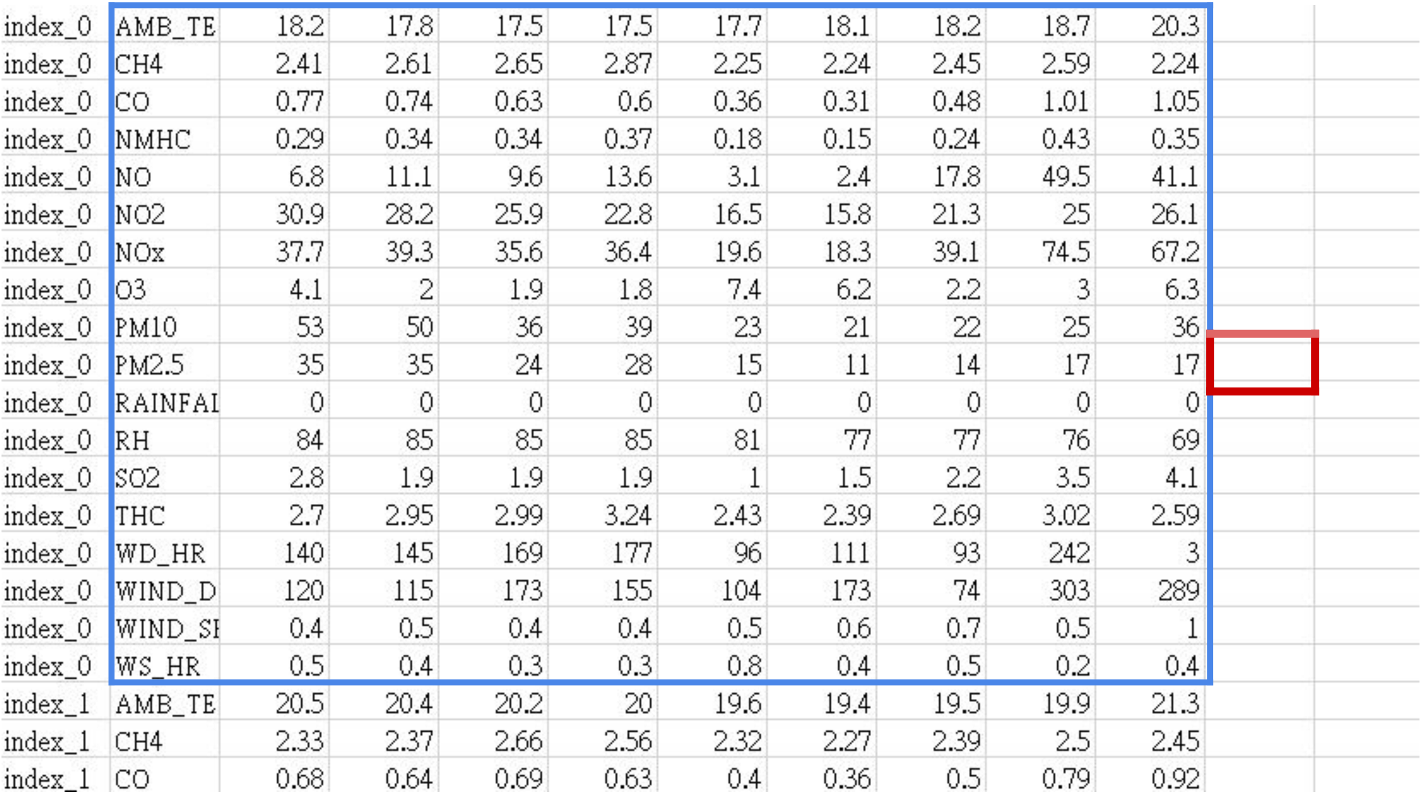
- There will be some invalid values, such as #, *, x, A.

ItemName



ItemName (English)	ItemName (Chinese)	Units of measurement
AMB_TEMP	溫度	°C
CH4	甲烷	ppm
CO	一氧化碳	ppm
NMHC	非甲烷碳氫化合物	ppm
NO	一氧化氮	ppb
NO2	二氧化氮	ppb
NOx	氮氧化物	ppb
O3	臭氧	ppb
PM10	懸浮微粒	μg/m3
PM2.5	細懸浮微粒	μg/m3
RAINFALL	雨量	mm
RH	相對濕度	%
SO2	二氧化硫	ppb
THC	總碳氫化合物	ppm
WD_HR	小時風向值	degrees
WIND_DIREC	風向	degrees
WIND_SPEED	風速	m/sec
WS_HR	小時風速值	m/sec

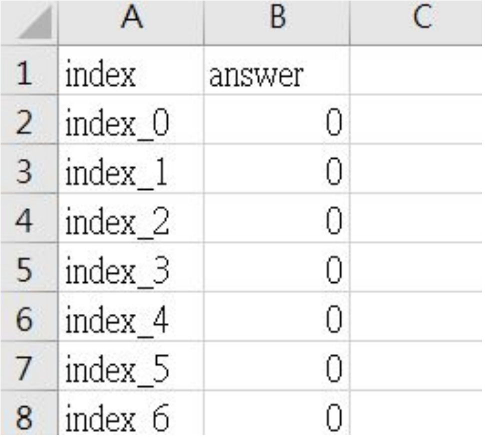
Testing Data



index_0	AMB_TE	18.2	17.8	17.5	17.5	17.7	18.1	18.2	18.7	20.3
index_0	CH4	2.41	2.61	2.65	2.87	2.25	2.24	2.45	2.59	2.24
index_0	CO	0.77	0.74	0.63	0.6	0.36	0.31	0.48	1.01	1.05
index_0	NMHC	0.29	0.34	0.34	0.37	0.18	0.15	0.24	0.43	0.35
index_0	NO	6.8	11.1	9.6	13.6	3.1	2.4	17.8	49.5	41.1
index_0	NO2	30.9	28.2	25.9	22.8	16.5	15.8	21.3	25	26.1
index_0	NOx	37.7	39.3	35.6	36.4	19.6	18.3	39.1	74.5	67.2
index_0	O3	4.1	2	1.9	1.8	7.4	6.2	2.2	3	6.3
index_0	PM10	53	50	36	39	23	21	22	25	36
index_0	PM2.5	35	35	24	28	15	11	14	17	17
index_0	RAINFAI	0	0	0	0	0	0	0	0	0
index_0	RH	84	85	85	85	81	77	77	76	69
index_0	SO2	2.8	1.9	1.9	1.9	1	1.5	2.2	3.5	4.1
index_0	THC	2.7	2.95	2.99	3.24	2.43	2.39	2.69	3.02	2.59
index_0	WD_HR	140	145	169	177	96	111	93	242	3
index_0	WIND_D	120	115	173	155	104	173	74	303	289
index_0	WIND_SI	0.4	0.5	0.4	0.4	0.5	0.6	0.7	0.5	1
index_0	WS_HR	0.5	0.4	0.3	0.3	0.8	0.4	0.5	0.2	0.4
index_1	AMB_TE	20.5	20.4	20.2	20	19.6	19.4	19.5	19.9	21.3
index_1	CH4	2.33	2.37	2.66	2.56	2.32	2.27	2.39	2.5	2.45
index_1	CO	0.68	0.64	0.69	0.63	0.4	0.36	0.5	0.79	0.92

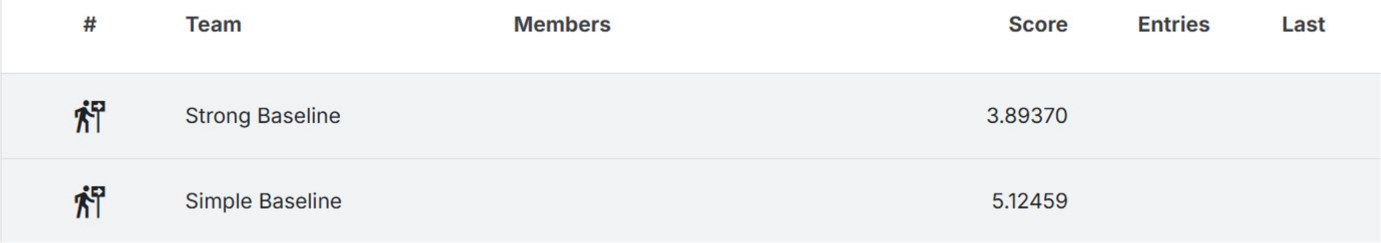
Kaggle Submission

- Kaggle link
- Display team name : <student ID>
- Submission format
  - A CSV file with size 245 x 2
  - The first row contains the column names: index and answer
  - The next 244 rows contains your predicted results
  - sample submission



	A	B	C
1	index	answer	
2	index_0	0	
3	index_1	0	
4	index_2	0	
5	index_3	0	
6	index_4	0	
7	index_5	0	
8	index 6	0	

- One simple bassline and one strong bassline are provided. Outperform the baselines to achieve a higher score



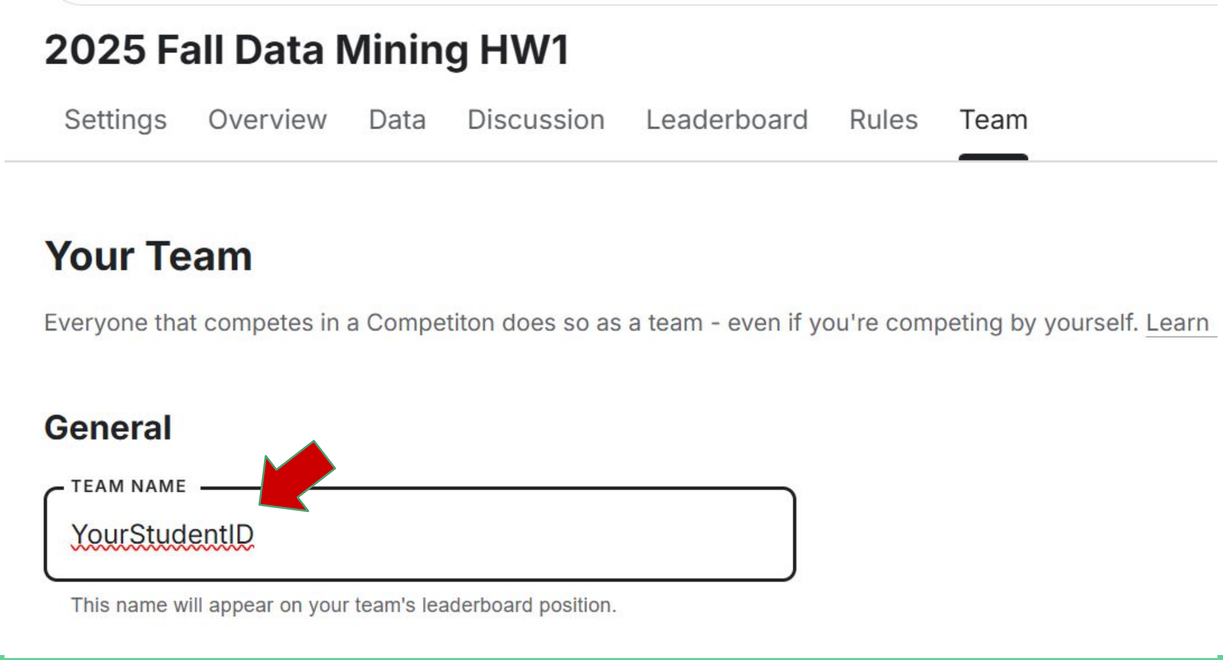
#	Team	Members	Score	Entries	Last
	Strong Baseline		3.89370		
	Simple Baseline		5.12459		

Kaggle Submission

- The scoring metric is RMSE.
- You can submit at most 5 times per day.
- You can choose 2 submissions to count toward the private leaderboard. If not specified, your best public submissions will be used by default.
- The private leaderboard will only be revealed after the competition ends.
- Public leaderboard is calculated using 50% of the test data, and private leaderboard is calculated with the other 50% of the test data, so the final standings may be different from the public leaderboard.
- Please tune your model parameters using your own validation set instead of adjusting parameters based on the public leaderboard. Otherwise, it's easy to overfit, leading to poor performance on the private leaderboard.

Change your team name

Remember to change the team name to <student ID>, or there will be a deduction of 5 points for HW 1.



2025 Fall Data Mining HW1

Your Team

Everyone that competes in a Competiton does so as a team - even if you're competing by yourself. Learn

General

TEAM NAME

YourStudentID

This name will appear on your team's leaderboard position.

Report Submission

Answer the following 3 questions:

1. How do you select features for your model input, and what preprocessing did you perform?
2. Compare the impact of different amounts of training data on the PM2.5 prediction accuracy. Visualize the results and explain them.
3. Discuss the impact of regularization on PM2.5 prediction accuracy.

Please answer the questions in detail to receive full points for each question.

Grading policy

- Kaggle (70%)
  - 30% based on the public leaderboard score and 70% based on the private leaderboard score
  - Leaderboard score consists of basic score and ranking score
    - Basic score :

      Over strong baseline : 55

      Over simple bassline : 40

      Under simple baseline : 25
    - Ranking score:

      15-(15/N)*(ranking-1), N=numbers of people in the interval

- Report (30%)
  - 10 for each quesiton

E3 Submission

Submit your source code and report to E3 before 10/8 (Wed.) 23:59.

No late submission !

- Format
    - source code : HW1_<student ID>.py or HW1_<student ID>.ipynb
    - report : HW1_<student ID>.pdf

If you have any question about HW 1, please feel free to contact with TA : Jun-Han Chen through email jhchen.cs12@nycu.edu.tw

Have Fun !