IOAI Vietnam

National Round

Group 3

An overview of our ideas, solutions and results

Members

Pham Dinh Hieu

12A1 Math - HUS High School for Gifted Students

Bui Quang Nguyen

11A1 Math - HUS High School for Gifted Students

Nguyen Trong Viet

12A1 Math - HUS High School for Gifted Students

Phan Viet Hoang

11A2 English - Hanoi - Amsterdam High school for the Gifted

Table of Tasks

01 Prediction

Predict PM2.5 parameter

02 Classificatio n

Classify pill images

03 Summarization

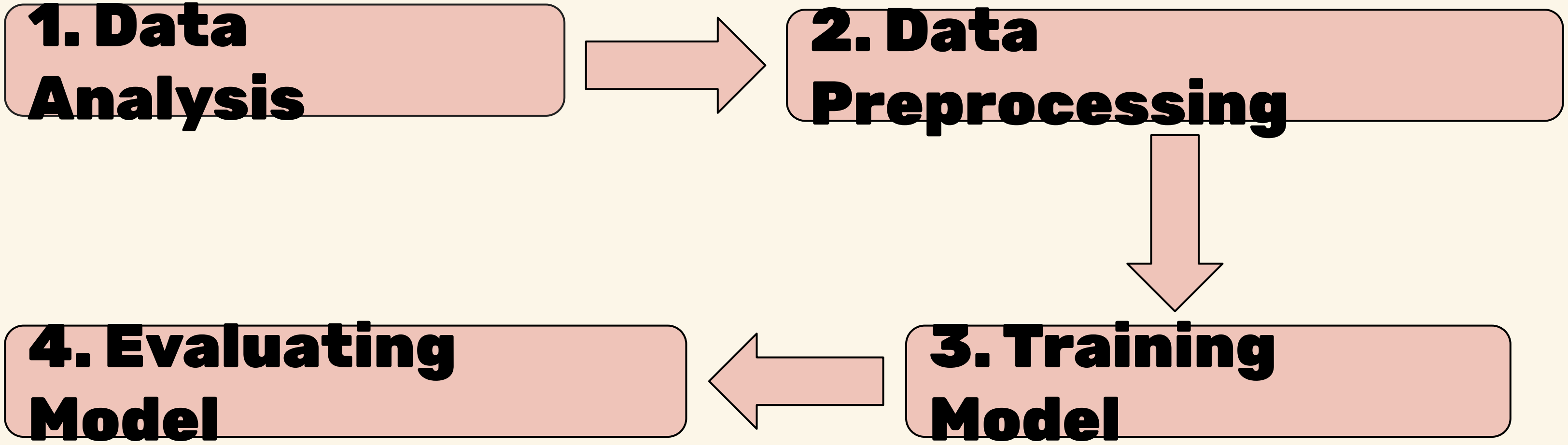
Summarize biomedical texts

01

Task 1

Predict PM2.5 level

General Idea



1. Data Analysis

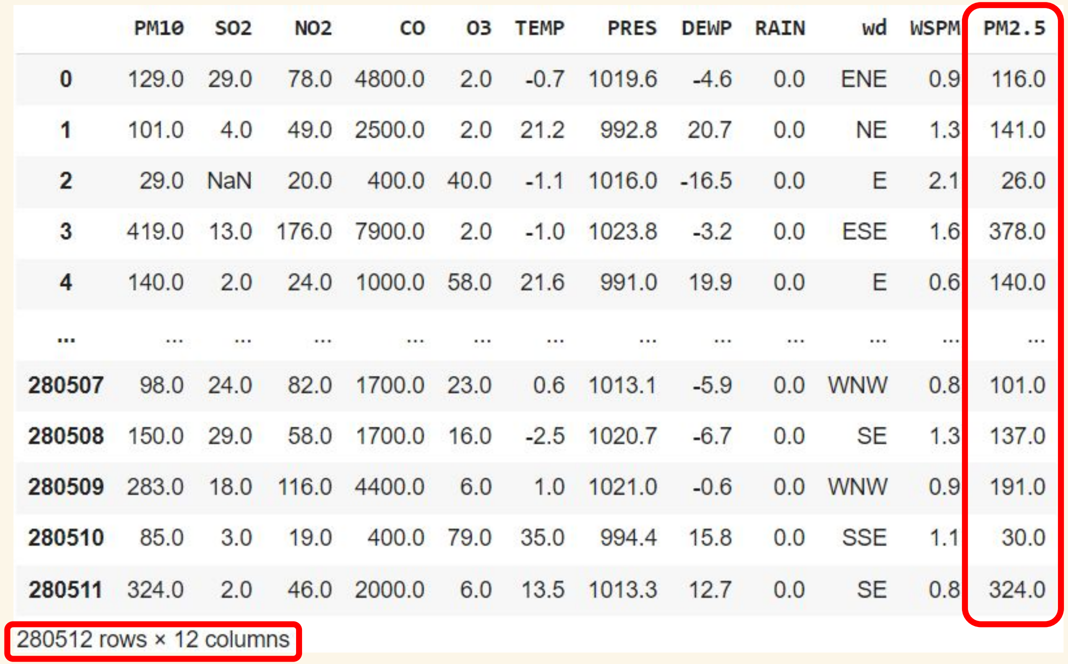
2. Data Preprocessing

3. Training Model

4. Evaluating Model

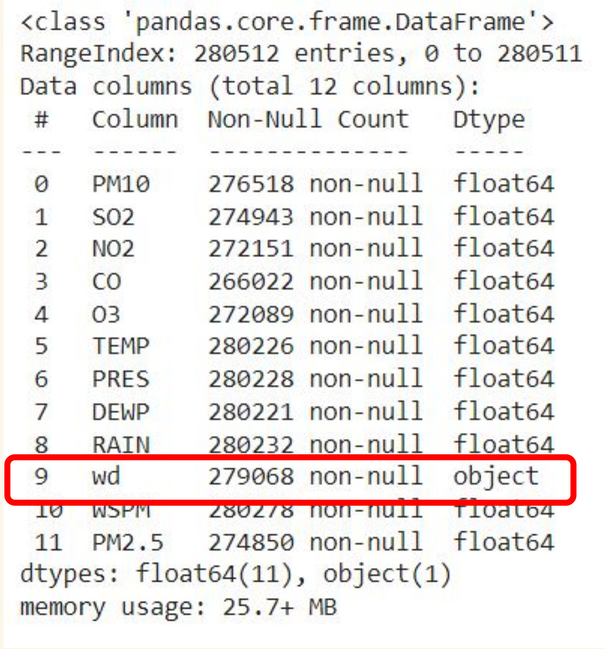
Data Analysis

Checking Data statistics



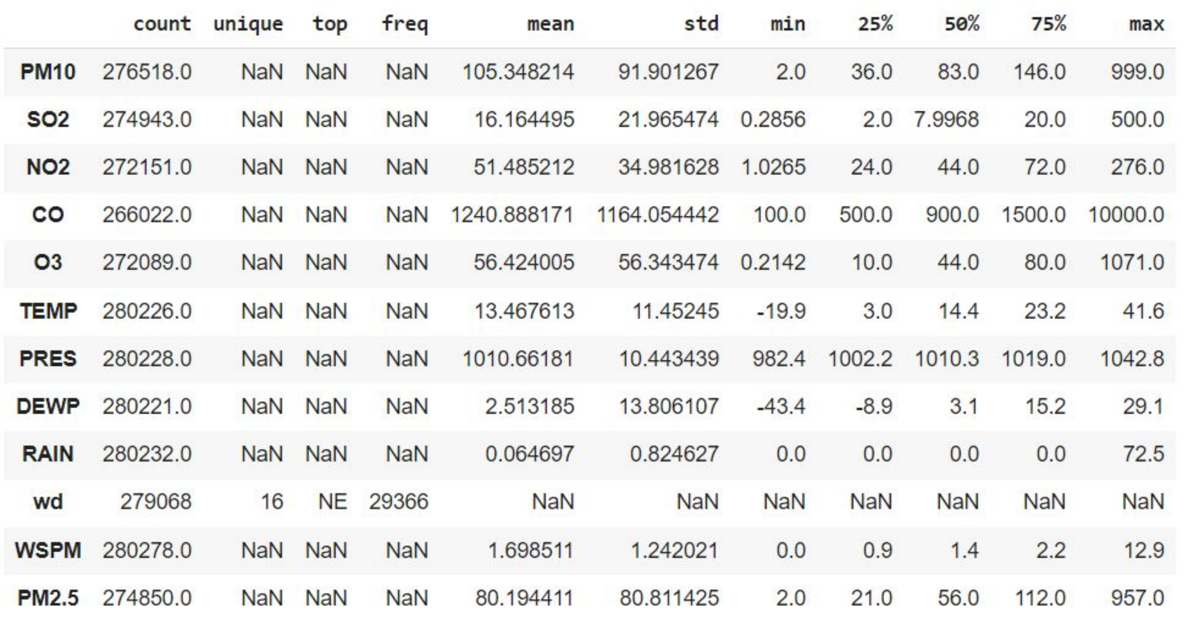
	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	wd	WSPM	PM2.5
0	129.0	29.0	78.0	4800.0	2.0	-0.7	1019.6	-4.6	0.0	ENE	0.9	116.0
1	101.0	4.0	49.0	2500.0	2.0	21.2	992.8	20.7	0.0	NE	1.3	141.0
2	29.0	NaN	20.0	400.0	40.0	-1.1	1016.0	-16.5	0.0	E	2.1	26.0
3	419.0	13.0	176.0	7900.0	2.0	-1.0	1023.8	-3.2	0.0	ESE	1.6	378.0
4	140.0	2.0	24.0	1000.0	58.0	21.6	991.0	19.9	0.0	E	0.6	140.0
...	...	...	...	...	...	...	...	...	...	...	...	...
280507	98.0	24.0	82.0	1700.0	23.0	0.6	1013.1	-5.9	0.0	WNW	0.8	101.0
280508	150.0	29.0	58.0	1700.0	16.0	-2.5	1020.7	-6.7	0.0	SE	1.3	137.0
280509	283.0	18.0	116.0	4400.0	6.0	1.0	1021.0	-0.6	0.0	WNW	0.9	191.0
280510	85.0	3.0	19.0	400.0	79.0	35.0	994.4	15.8	0.0	SSE	1.1	30.0
280511	324.0	2.0	46.0	2000.0	6.0	13.5	1013.3	12.7	0.0	SE	0.8	324.0

280512 rows × 12 columns



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 280512 entries, 0 to 280511
Data columns (total 12 columns):
 #   Column  Non-Null Count   Dtype  
---  ------  --------------   -----  
 0   PM10    276518 non-null  float64
 1   SO2     274943 non-null  float64
 2   NO2     272151 non-null  float64
 3   CO      266022 non-null  float64
 4   O3      272089 non-null  float64
 5   TEMP    280226 non-null  float64
 6   PRES    280228 non-null  float64
 7   DEWP    280221 non-null  float64
 8   RAIN    280232 non-null  float64
 9   wd      279068 non-null  object 
 10  WSPM    280278 non-null  float64
 11  PM2.5   274850 non-null  float64
dtypes: float64(11), object(1)
memory usage: 25.7+ MB
```

Checking Data statistics



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
PM10	276518.0	NaN	NaN	NaN	105.348214	91.901267	2.0	36.0	83.0	146.0	999.0
SO2	274943.0	NaN	NaN	NaN	16.164495	21.965474	0.2856	2.0	7.9968	20.0	500.0
NO2	272151.0	NaN	NaN	NaN	51.485212	34.981628	1.0265	24.0	44.0	72.0	276.0
CO	266022.0	NaN	NaN	NaN	1240.888171	1164.054442	100.0	500.0	900.0	1500.0	10000.0
O3	272089.0	NaN	NaN	NaN	56.424005	56.343474	0.2142	10.0	44.0	80.0	1071.0
TEMP	280226.0	NaN	NaN	NaN	13.467613	11.45245	-19.9	3.0	14.4	23.2	41.6
PRES	280228.0	NaN	NaN	NaN	1010.66181	10.443439	982.4	1002.2	1010.3	1019.0	1042.8
DEWP	280221.0	NaN	NaN	NaN	2.513185	13.806107	-43.4	-8.9	3.1	15.2	29.1
RAIN	280232.0	NaN	NaN	NaN	0.064697	0.824627	0.0	0.0	0.0	0.0	72.5
wd	279068	16	NE	29366	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WSPM	280278.0	NaN	NaN	NaN	1.698511	1.242021	0.0	0.9	1.4	2.2	12.9
PM2.5	274850.0	NaN	NaN	NaN	80.194411	80.811425	2.0	21.0	56.0	112.0	957.0

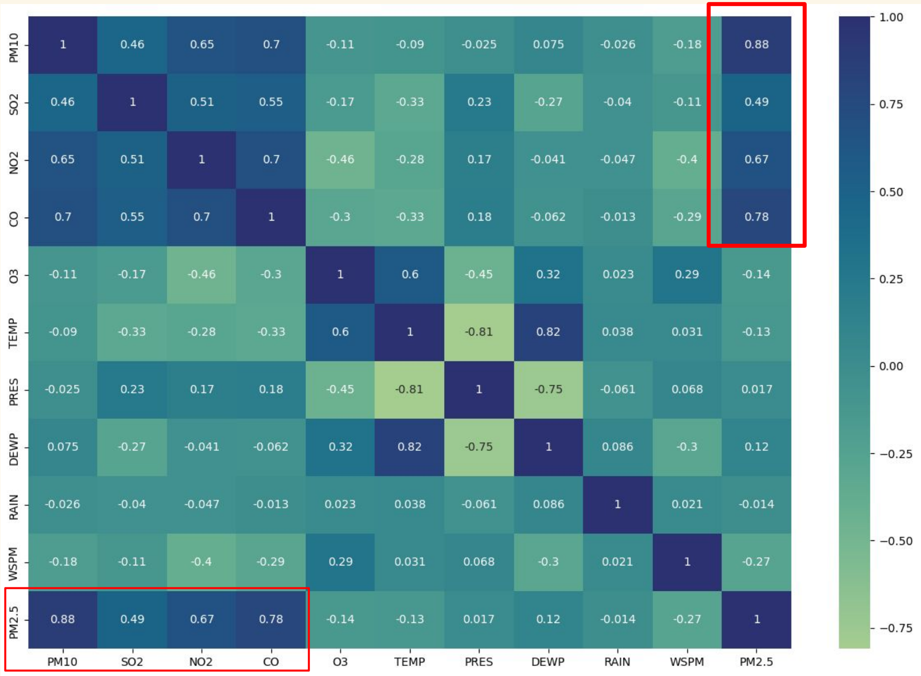
Missing values

Variability

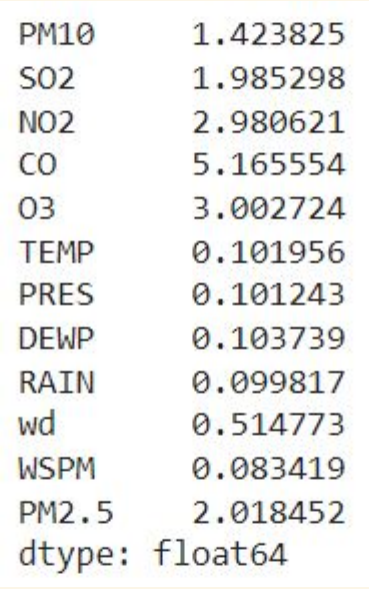
Outliers

Categorical Data

Correlation and Missing

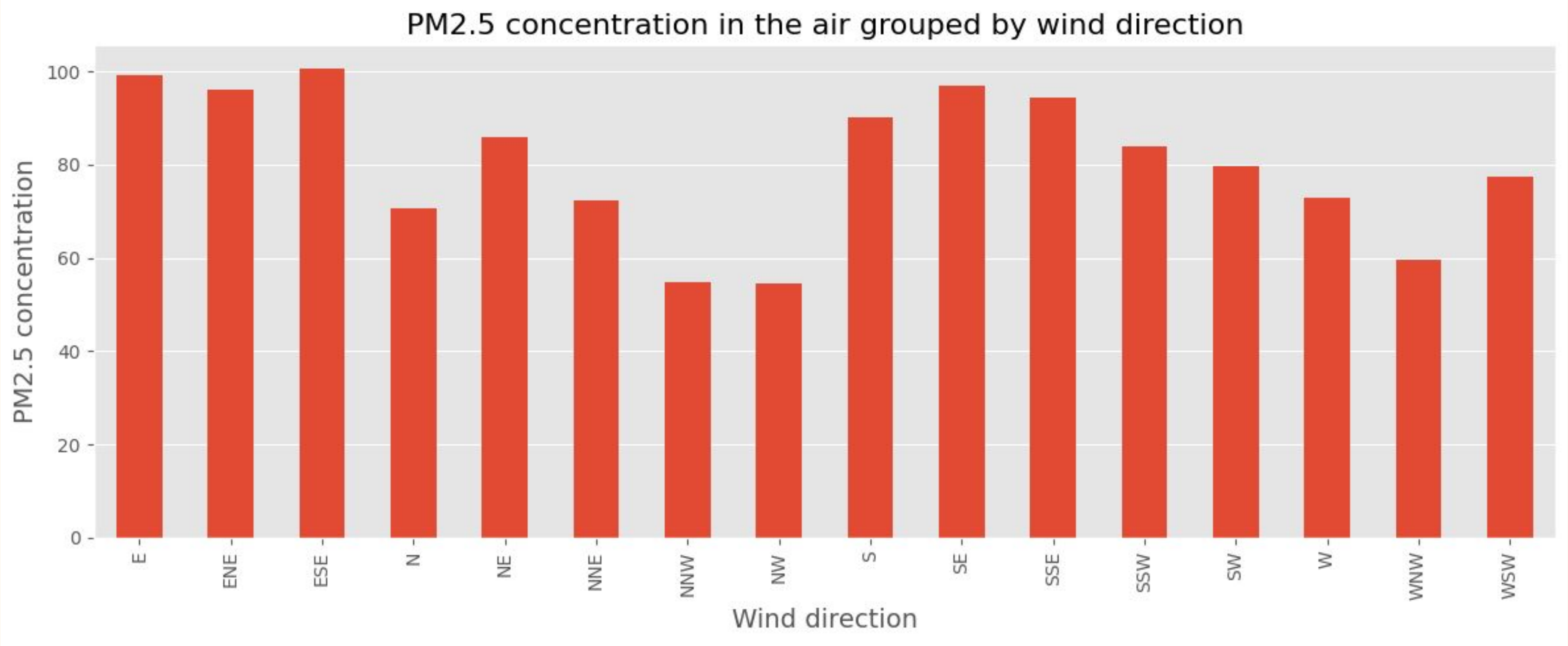


	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	WSPM	PM2.5
PM10	1	0.46	0.65	0.7	-0.11	-0.09	-0.025	0.075	-0.026	-0.18	0.88
SO2	0.46	1	0.51	0.55	-0.17	-0.33	0.23	-0.27	-0.04	-0.11	0.49
NO2	0.65	0.51	1	0.7	-0.46	-0.28	0.17	-0.041	-0.047	-0.4	0.67
CO	0.7	0.55	0.7	1	-0.3	-0.33	0.18	-0.062	-0.013	-0.29	0.78
O3	-0.11	-0.17	-0.46	-0.3	1	0.6	-0.45	0.32	0.023	0.29	-0.14
TEMP	-0.09	-0.33	-0.28	-0.33	0.6	1	-0.81	0.82	0.038	0.031	-0.13
PRES	-0.025	0.23	0.17	0.18	-0.45	-0.81	1	-0.75	-0.061	0.068	0.017
DEWP	0.075	-0.27	-0.041	-0.062	0.32	0.82	-0.75	1	0.086	-0.3	0.12
RAIN	-0.026	-0.04	-0.047	-0.013	0.023	0.038	-0.061	0.086	1	0.021	-0.014
WSPM	-0.18	-0.11	-0.4	-0.29	0.29	0.031	0.068	-0.3	0.021	1	-0.27
PM2.5	0.88	0.49	0.67	0.78	-0.14	-0.13	0.017	0.12	-0.014	-0.27	1



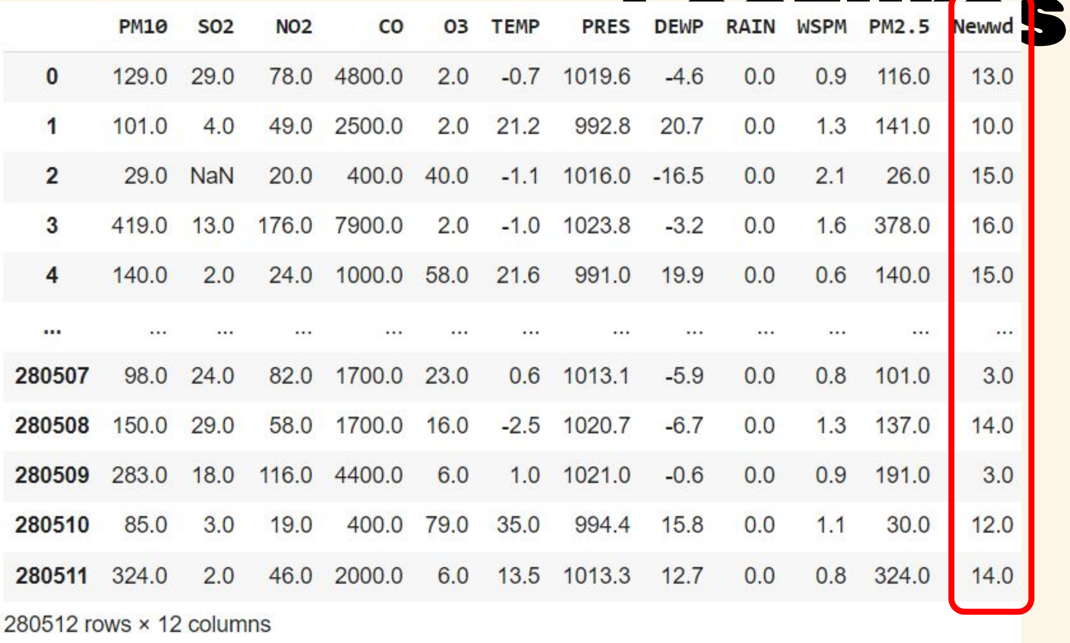
```
PM10      1.423825
SO2       1.985298
NO2       2.980621
CO        5.165554
O3        3.002724
TEMP      0.101956
PRES      0.101243
DEWP      0.103739
RAIN      0.099817
wd        0.514773
WSPM      0.083419
PM2.5     2.018452
dtype: float64
```

Checking Categorical



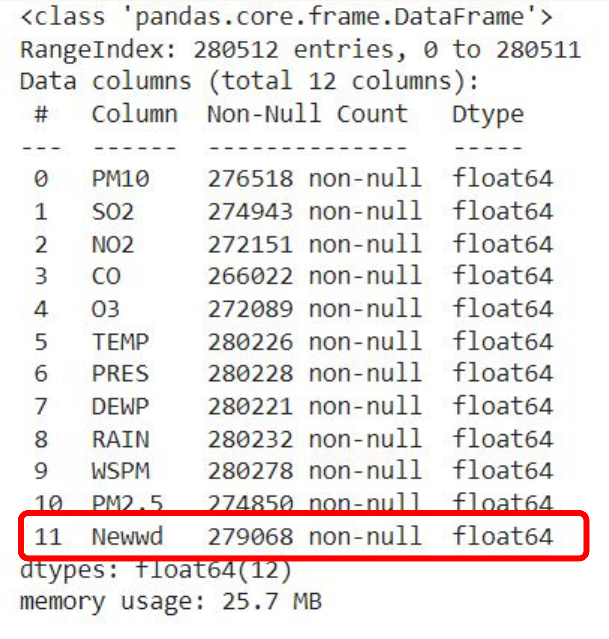
Data Preprocessing

Encoding Categorical Features



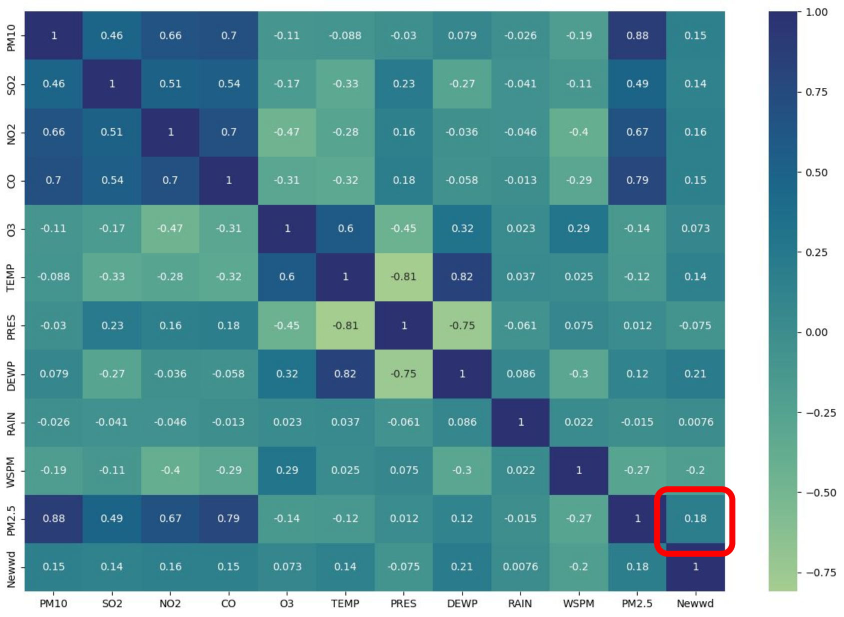
	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	WSPM	PM2.5	Newwd
0	129.0	29.0	78.0	4800.0	2.0	-0.7	1019.6	-4.6	0.0	0.9	116.0	13.0
1	101.0	4.0	49.0	2500.0	2.0	21.2	992.8	20.7	0.0	1.3	141.0	10.0
2	29.0	NaN	20.0	400.0	40.0	-1.1	1016.0	-16.5	0.0	2.1	26.0	15.0
3	419.0	13.0	176.0	7900.0	2.0	-1.0	1023.8	-3.2	0.0	1.6	378.0	16.0
4	140.0	2.0	24.0	1000.0	58.0	21.6	991.0	19.9	0.0	0.6	140.0	15.0
...	...	...	...	...	...	...	...	...	...	...	...	...
280507	98.0	24.0	82.0	1700.0	23.0	0.6	1013.1	-5.9	0.0	0.8	101.0	3.0
280508	150.0	29.0	58.0	1700.0	16.0	-2.5	1020.7	-6.7	0.0	1.3	137.0	14.0
280509	283.0	18.0	116.0	4400.0	6.0	1.0	1021.0	-0.6	0.0	0.9	191.0	3.0
280510	85.0	3.0	19.0	400.0	79.0	35.0	994.4	15.8	0.0	1.1	30.0	12.0
280511	324.0	2.0	46.0	2000.0	6.0	13.5	1013.3	12.7	0.0	0.8	324.0	14.0

280512 rows × 12 columns



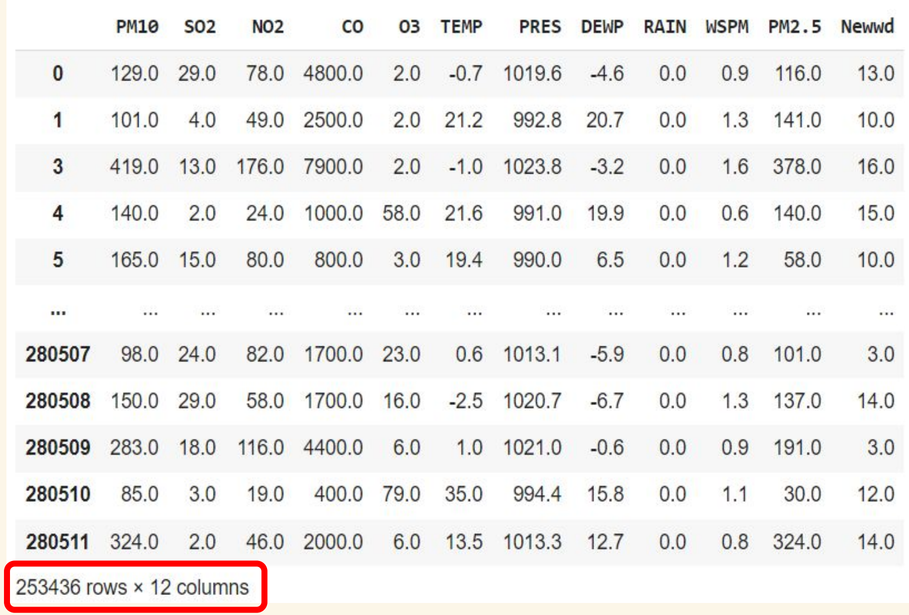
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 280512 entries, 0 to 280511
Data columns (total 12 columns):
 #   Column  Non-Null Count   Dtype  
---  ------  --------------   -----  
 0   PM10    276518 non-null  float64
 1   SO2     274943 non-null  float64
 2   NO2     272151 non-null  float64
 3   CO      266022 non-null  float64
 4   O3      272089 non-null  float64
 5   TEMP    280226 non-null  float64
 6   PRES    280228 non-null  float64
 7   DEWP    280221 non-null  float64
 8   RAIN    280232 non-null  float64
 9   WSPM    280278 non-null  float64
 10  PM2.5   274850 non-null  float64
 11  Newwd   279068 non-null  float64
dtypes: float64(12)
memory usage: 25.7 MB
```

Handling Missing Values



	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	WSPM	PM2.5	Newwd
PM10	1	0.46	0.66	0.7	-0.11	-0.088	-0.03	0.079	-0.026	-0.19	0.88	0.15
SO2	0.46	1	0.51	0.54	-0.17	-0.33	0.23	-0.27	-0.041	-0.11	0.49	0.14
NO2	0.66	0.51	1	0.7	-0.47	-0.28	0.16	-0.036	-0.046	-0.4	0.67	0.16
CO	0.7	0.54	0.7	1	-0.31	-0.32	0.18	-0.058	-0.013	-0.29	0.79	0.15
O3	-0.11	-0.17	-0.47	-0.31	1	0.6	-0.45	0.32	0.023	0.29	-0.14	0.073
TEMP	-0.088	-0.33	-0.28	-0.32	0.6	1	-0.81	0.82	0.037	0.025	-0.12	0.14
PRES	-0.03	0.23	0.16	0.18	-0.45	-0.81	1	-0.75	-0.061	0.075	0.012	-0.075
DEWP	0.079	-0.27	-0.036	-0.058	0.32	0.82	-0.75	1	0.086	-0.3	0.12	0.21
RAIN	-0.026	-0.041	-0.046	-0.013	0.023	0.037	-0.061	0.086	1	0.022	-0.015	0.0076
WSPM	-0.19	-0.11	-0.4	-0.29	0.29	0.025	0.075	-0.3	0.022	1	-0.27	-0.2
PM2.5	0.88	0.49	0.67	0.79	-0.14	-0.12	0.012	0.12	-0.015	-0.27	1	0.18
Newwd	0.15	0.14	0.16	0.15	0.073	0.14	-0.075	0.21	0.0076	-0.2	0.18	1

Correlation of Preprocessed Data



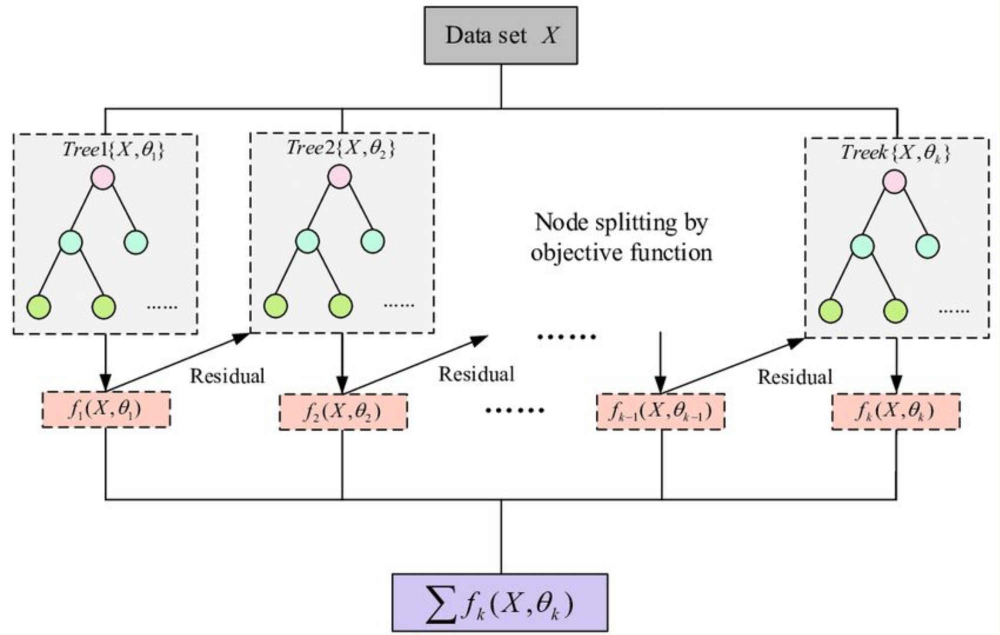
	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	WSPM	PM2.5	Newwd
0	129.0	29.0	78.0	4800.0	2.0	-0.7	1019.6	-4.6	0.0	0.9	116.0	13.0
1	101.0	4.0	49.0	2500.0	2.0	21.2	992.8	20.7	0.0	1.3	141.0	10.0
3	419.0	13.0	176.0	7900.0	2.0	-1.0	1023.8	-3.2	0.0	1.6	378.0	16.0
4	140.0	2.0	24.0	1000.0	58.0	21.6	991.0	19.9	0.0	0.6	140.0	15.0
5	165.0	15.0	80.0	800.0	3.0	19.4	990.0	6.5	0.0	1.2	58.0	10.0
...	...	...	...	...	...	...	...	...	...	...	...	...
280507	98.0	24.0	82.0	1700.0	23.0	0.6	1013.1	-5.9	0.0	0.8	101.0	3.0
280508	150.0	29.0	58.0	1700.0	16.0	-2.5	1020.7	-6.7	0.0	1.3	137.0	14.0
280509	283.0	18.0	116.0	4400.0	6.0	1.0	1021.0	-0.6	0.0	0.9	191.0	3.0
280510	85.0	3.0	19.0	400.0	79.0	35.0	994.4	15.8	0.0	1.1	30.0	12.0
280511	324.0	2.0	46.0	2000.0	6.0	13.5	1013.3	12.7	0.0	0.8	324.0	14.0

253436 rows × 12 columns

Train data after dropping all missing values

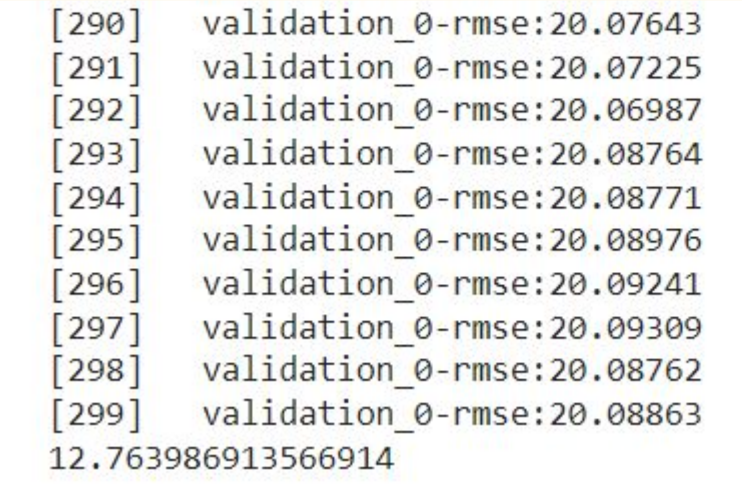
Training Model

XGBoost (Extreme Gradient Boosting)



A popular library that implements the Gradient Boosting algorithm, with a lot of optimisation.

Training Process



```
[290]	validation_0-rmse:20.07643
[291]	validation_0-rmse:20.07225
[292]	validation_0-rmse:20.06987
[293]	validation_0-rmse:20.08764
[294]	validation_0-rmse:20.08771
[295]	validation_0-rmse:20.08976
[296]	validation_0-rmse:20.09241
[297]	validation_0-rmse:20.09309
[298]	validation_0-rmse:20.08762
[299]	validation_0-rmse:20.08863
12.763986913566914
```

- Data Split: The data is split into 2 parts: training set, validation set.

- Hyperparameter values:
  + n_estimators = 300
  + learning_rate = 0.1
  + max_depth = 14

Evaluating Model

Result

15.5612

Result

15.5612

Potential ways to improve:

- Further optimise hyperparameters
- Try different methods for handling missing data and transforming category data

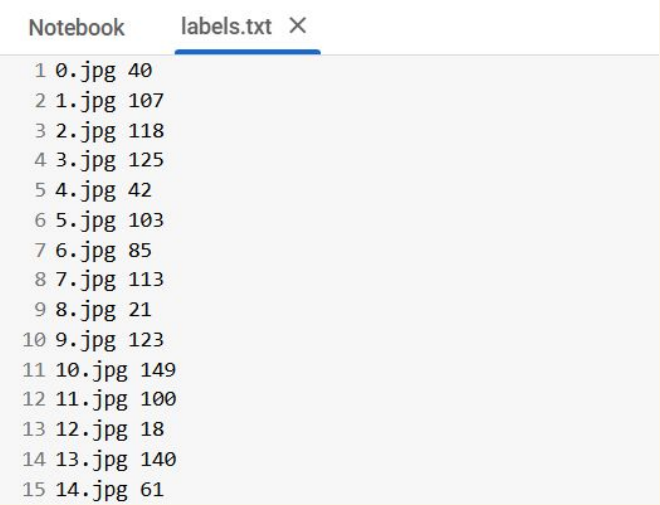
02

Classification

Classify Pill Images

Preprocessing

Preprocessing

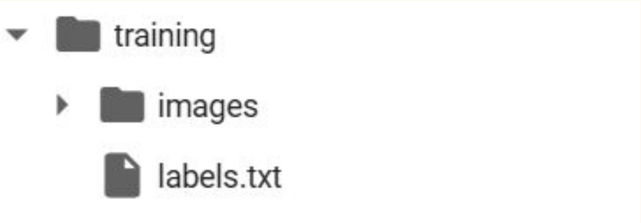


```
0.jpg 40
1.jpg 107
2.jpg 118
3.jpg 125
4.jpg 42
5.jpg 103
6.jpg 85
7.jpg 113
8.jpg 21
9.jpg 123
10.jpg 149
11.jpg 100
12.jpg 18
13.jpg 140
14.jpg 61
```

8693 images

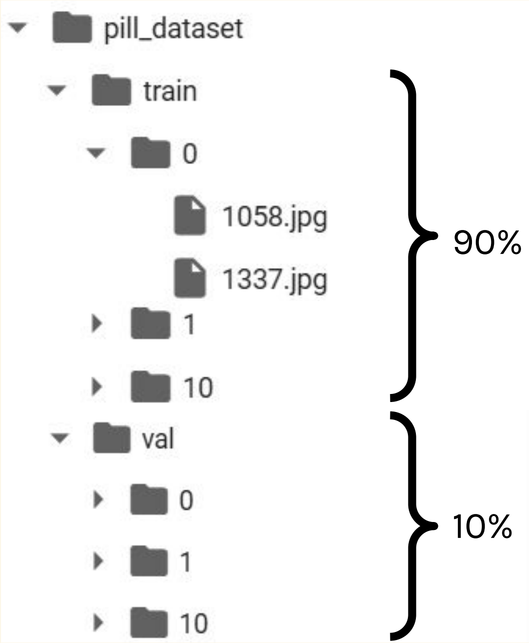
150 types of pills

Preprocessing



```
▾ training
   ▸ images
     labels.txt
```

⟹



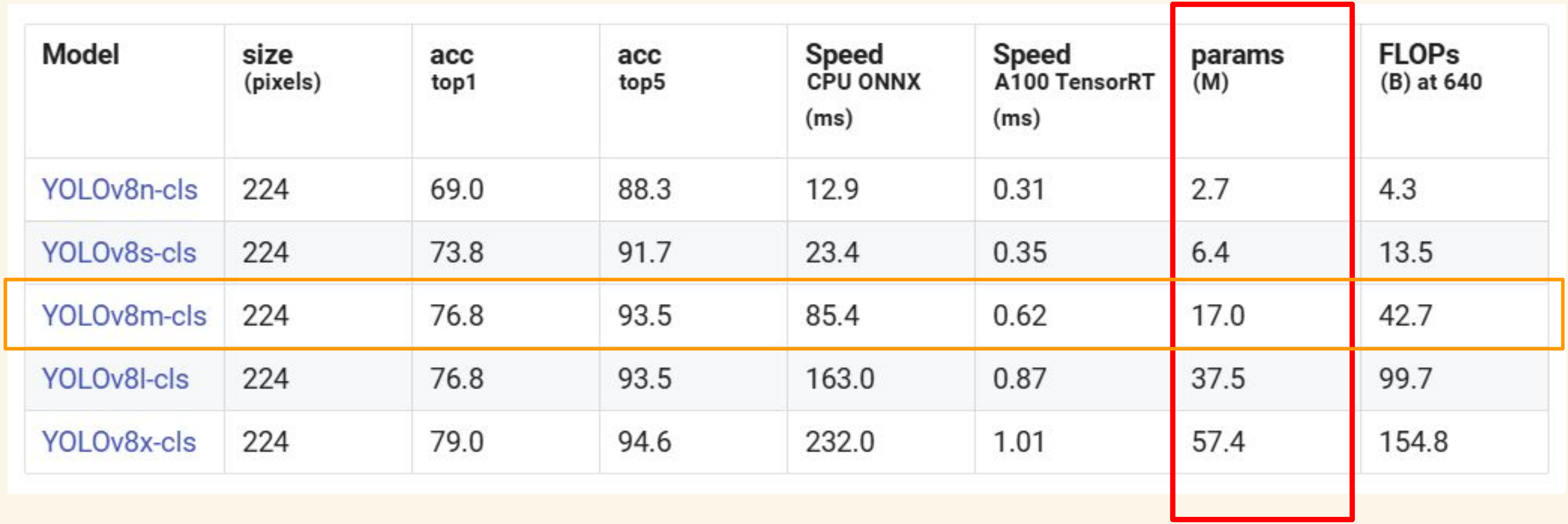
```
▾ pill_dataset
   ▾ train
      ▾ 0
           1058.jpg
           1337.jpg
      ▸ 1
      ▸ 10
   ▾ val
      ▸ 0
      ▸ 1
      ▸ 10
```

90%

10%

Building Model

Building Model

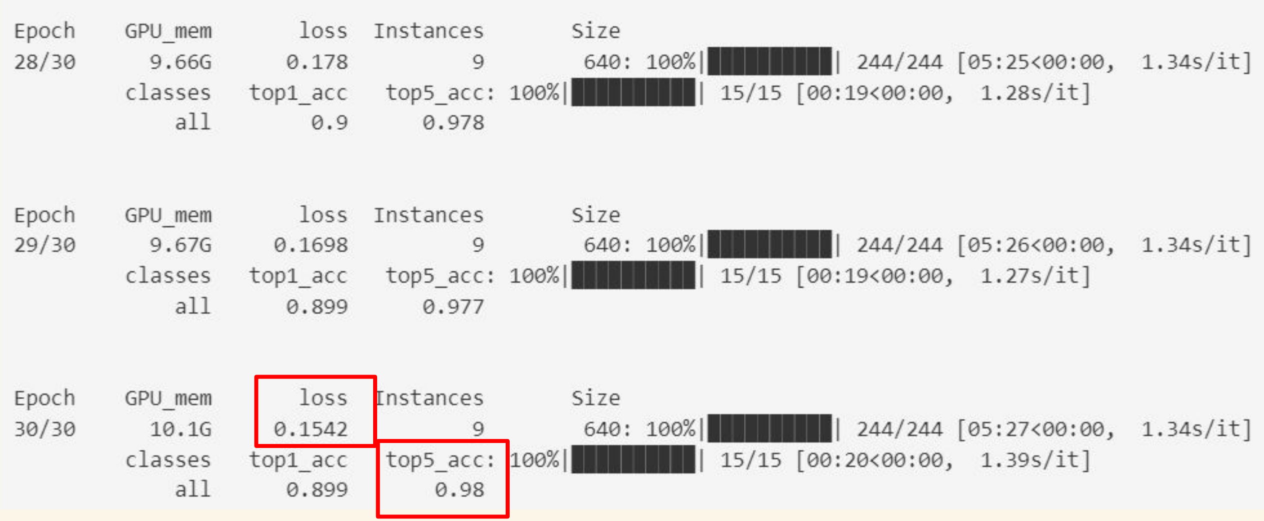


Model	size (pixels)	acc top1	acc top5	Speed CPU ONNX (ms)	Speed A100 TensorRT (ms)	params (M)	FLOPs (B) at 640
YOLOv8n-cls	224	69.0	88.3	12.9	0.31	2.7	4.3
YOLOv8s-cls	224	73.8	91.7	23.4	0.35	6.4	13.5
YOLOv8m-cls	224	76.8	93.5	85.4	0.62	17.0	42.7
YOLOv8l-cls	224	76.8	93.5	163.0	0.87	37.5	99.7
YOLOv8x-cls	224	79.0	94.6	232.0	1.01	57.4	154.8

Building Model

YOLOv8m-cls:

- epochs = 30
- imgsz = 640
- patience = 10
- batch = 32



```
Epoch    GPU_mem       loss  Instances       Size
28/30      9.66G      0.178          9        640: 100%|██████████| 244/244 [05:25<00:00,  1.34s/it]
         classes   top1_acc   top5_acc: 100%|██████████| 15/15 [00:19<00:00,  1.28s/it]
             all        0.9      0.978

Epoch    GPU_mem       loss  Instances       Size
29/30      9.67G     0.1698          9        640: 100%|██████████| 244/244 [05:26<00:00,  1.34s/it]
         classes   top1_acc   top5_acc: 100%|██████████| 15/15 [00:19<00:00,  1.27s/it]
             all      0.899      0.977

Epoch    GPU_mem       loss  Instances       Size
30/30      10.1G     0.1542          9        640: 100%|██████████| 244/244 [05:27<00:00,  1.34s/it]
         classes   top1_acc   top5_acc: 100%|██████████| 15/15 [00:20<00:00,  1.39s/it]
             all      0.899       0.98
```

Result

0.8947

Result

0.8947

For reference:

- Score of previous round with this approach: **0.8761**
- Score of previous round with data augmentation (fine-tune from the best model) : **0.8698**

0
3

Summarisatio
n

Extract relevant sentences from a given text

General idea

This text is important. (1)

This text is not as important. (2)

This text is very important. (3)

This text is not important. (4)

This text is very important. (5)

General idea

This text is important. (1)

This text is not as important. (2)

This text is very important. (3)

This text is not important. (4)

This text is very important. (5)

⟹

This text is very important. (3)

This text is very important. (5)

This text is important. (1)

This text is not as important. (2)

This text is not important. (4)

Level of importance decreases

General idea

Will be in the summary

This text is very important. (3)

This text is very important. (5)

This text is important. (1)

This text is not as important. (2)

This text is not important. (4)

Preprocessing

Preprocessing

the quick brown fox ? he jumps over the lazy dog .

Preprocessing

the quick brown fox ? he jumps over the lazy dog .

⇩

the quick brown fox he jumps over the lazy dog

Preprocessing

the quick brown fox he jumps over the lazy dog

Preprocessing

the quick brown fox he jumps over the lazy dog

⇩

quick brown fox jumps lazy dog

Preprocessing

quick brown fox jumps lazy dog

Preprocessing

quick brown fox jumps lazy dog

⇩

quick brown fox jump lazi dog

Preprocessing

the quick brown fox ? he jumps over the lazy dog .

⇩

quick brown fox jump lazi dog

Punctuation and stopwords removed
Stemmed with Porter stemmer

TF-IDF score

TF-IDF score

$$tfidf(w) = tf(w) \times idf(w)$$

$$tf(w) = \frac{\text{count of word } w \text{ in text}}{\text{number of words in text}}$$

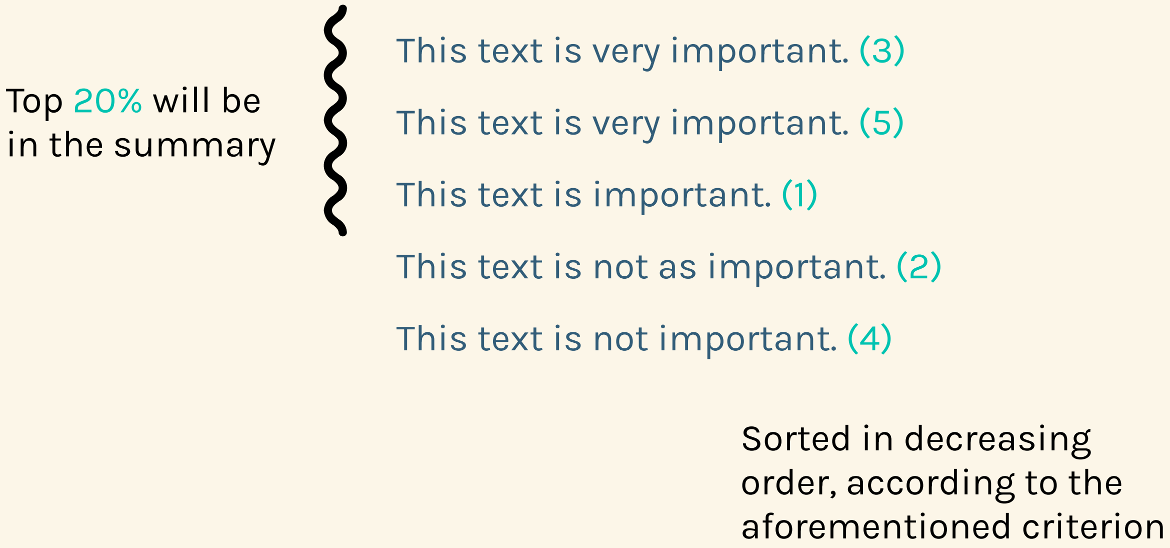
$$idf(w) = -\log \frac{\text{count of sentence with word } w \text{ in text}}{\text{number of sentences in text}}$$

TF-IDF score

$$tfidf(w) = tf(w) \times idf(w)$$

$$score(s) = \frac{\sum_{w \text{ in } s} tfidf(w)}{\text{number of words in s}}$$

General idea



Top 20% will be in the summary

This text is very important. (3)

This text is very important. (5)

This text is important. (1)

This text is not as important. (2)

This text is not important. (4)

Sorted in decreasing order, according to the aforementioned criterion

Result

Result

0.2871

Result

0.2871

For reference:

- Score of previous round with this approach: **0.3767**
- Score of previous round with a random approach (25% of all sentences are randomly selected) : **0.2482**

04

Q & A

Ask us anything!

Thanks for listening!