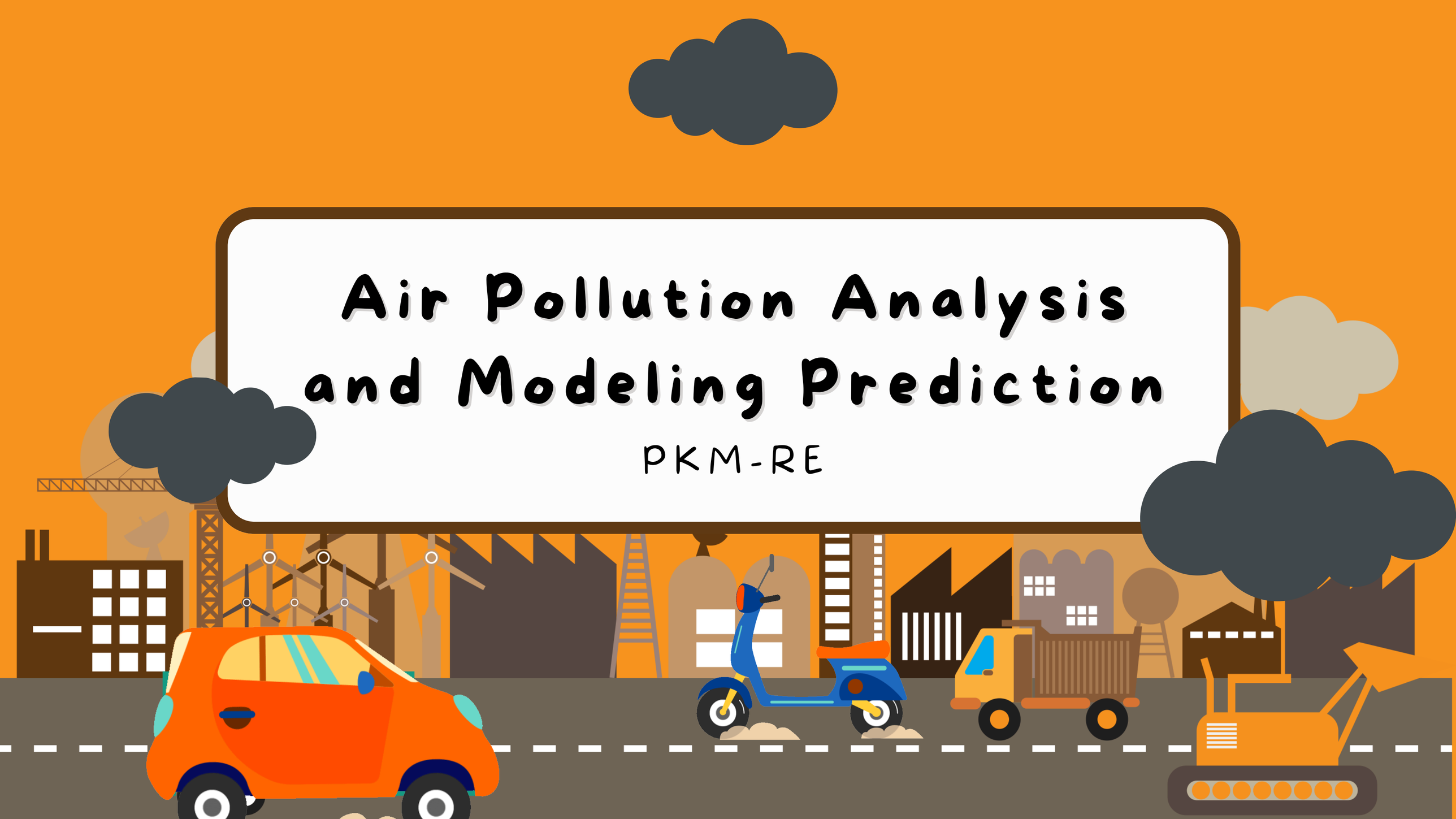
Air Pollution Analysis and Modeling Prediction

PKM-RE



LB09 - Kelompok Menolak Nilai Jelek

Kimberly Kayla Dewi - 2602190816

Jennifer Patricia - 2602186472

Wella Priscillia - 2602135963

David Paul Ong - 2602164433

Fiona Maharani Nugraha - 2602199582

Problem

Kesehatan Masyarakat

Pencemaran udara berdampak serius pada kesehatan masyarakat, termasuk peningkatan kasus penyakit pernapasan, risiko kematian dini, dan dampak sosial-ekonomi yang signifikan.

Masalah Lingkungan

Kerusakan ekosistem, penurunan kualitas air dan tanah, keanekaragaman hayati.

Solution

Mengembangkan model prediktif, penggunaan machine learning pada dataset kualitas udara dapat memberikan landasan untuk mengembangkan model yang dapat memprediksi tingkat pencemaran udara di masa depan berdasarkan tren historis.

Dataset

Berisi Indeks Standar Pencemar Udara (ISPU) yang diukur dari 5 stasiun pemantau kualitas udara (SPKU) yang ada di provinsi DKI Jakarta Tahun 2021.

Indeks Pencemaran Udara

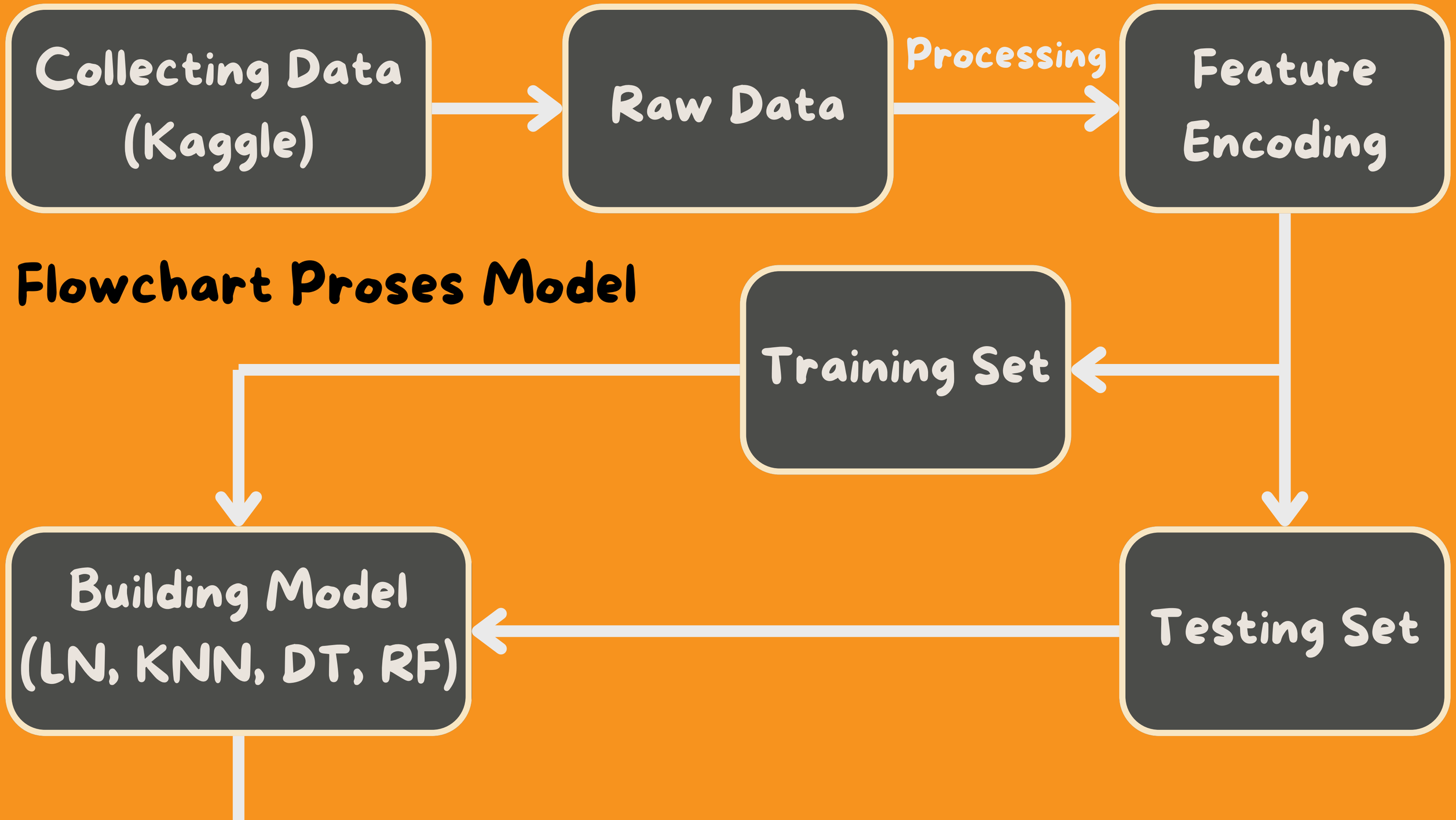
TThis dataset contains Indeks Standar Pencemar Udara (ISPU)

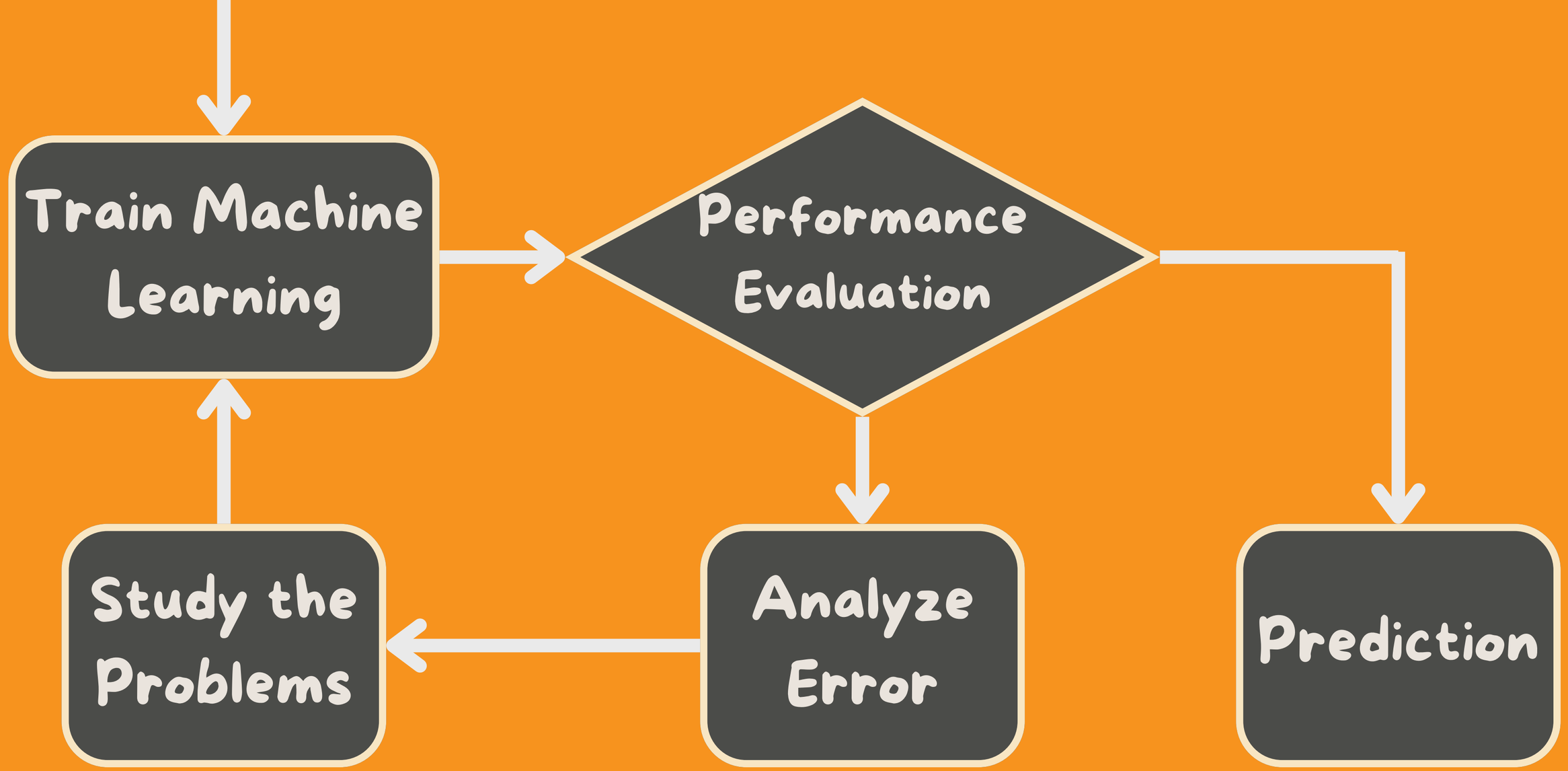
k kaggle.com

tanggal	# pm10	# pm25	# so2	# co

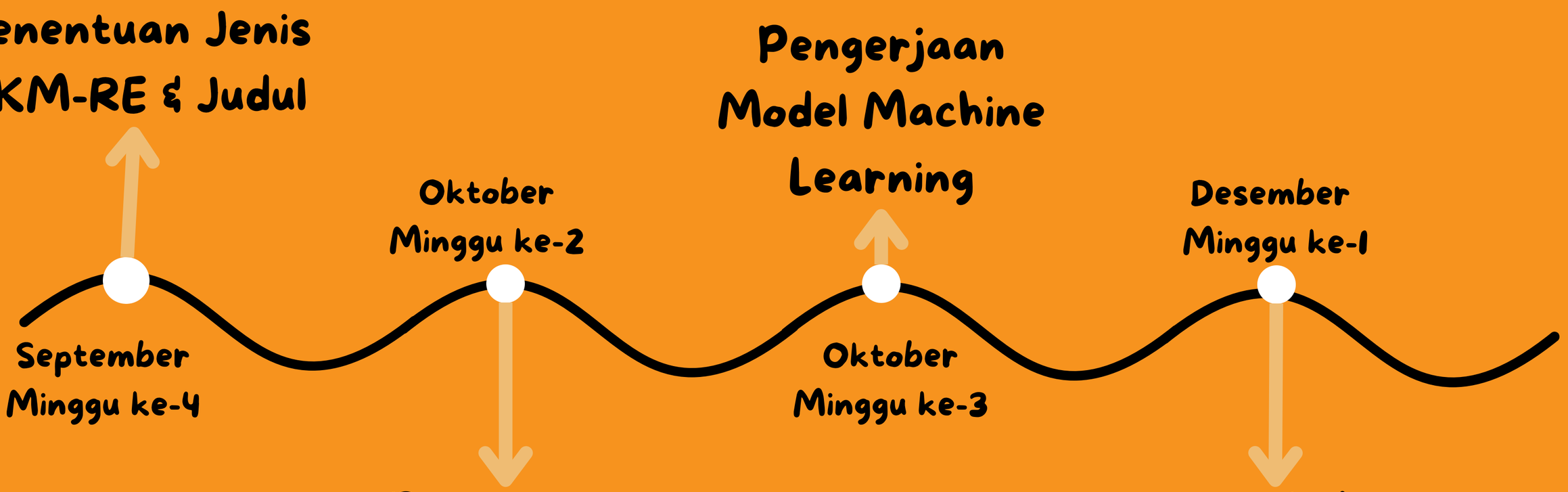
# o3	# no2	# max

A critical		A categori		A location	
PM25	92%	SEDANG	61%	DKI4	62%
O3	7%	TIDAK SEHAT	38%	DKI3	15%
Other (4)	1%	Other (3)	1%	Other (84)	23%


Collecting Data (Kaggle)
Raw Data
Processing
Feature Encoding
Flowchart Proses Model
Training Set
Building Model (LN, KNN, DT, RF)
Testing Set


Train Machine Learning
Performance Evaluation
Prediction
Analyze Error
Study the Problems

Timeline Pengerjaan

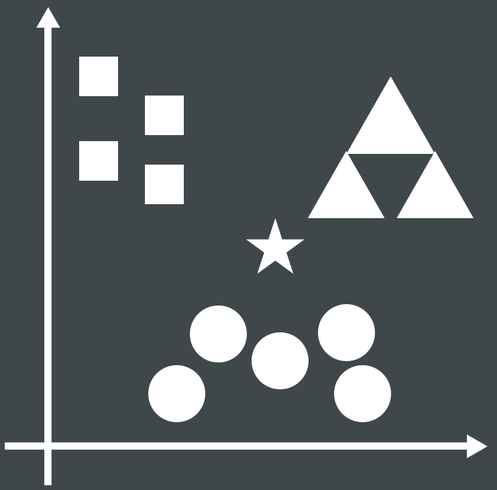


Machine Learning Model

Linear Regression

Memahami bagaimana parameter akan berkontribusi terhadap nilai ISPU, sehingga dapat mengukur sejauh mana hubungan antara parameter dengan ISPU dan membuat prediksi.

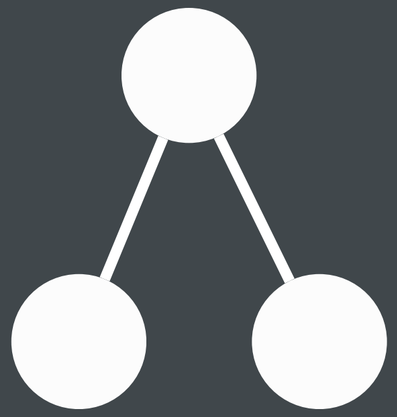
Methods



KNN

Menganalisis data ini dengan memanfaatkan informasi dari stasiun pemantau yang terdekat untuk memprediksi nilai ISPU di lokasi tertentu berdasarkan data yang ada.

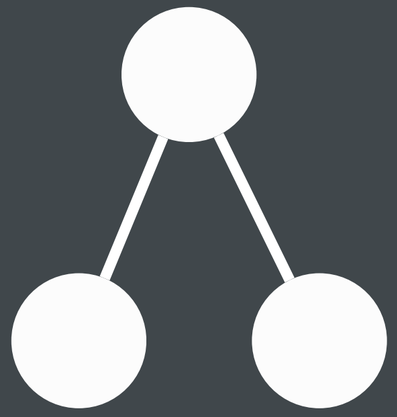
Machine Learning Model



Random Forest

Memanfaatkan data dari berbagai stasiun pemantau untuk membuat model yang dapat memprediksi atau mengevaluasi ISPU di lokasi tertentu dengan mempertimbangkan berbagai variabel dan pengaruhnya.

Methods



Decision Tree

Memahami bagaimana parameter mempengaruhi kategori ISPU, memecah data menjadi kelompok-kelompok yang lebih kecil berdasarkan nilai-nilai parameter dan memutuskan kategori ISPU.

Data Cleaning

Final Dataframe After Being Cleaned

	tanggal	pm10	pm25	so2	co	o3	no2	categori	bulan
0	2021-01-01	43	95	58	29	35	65	SEDANG	January
1	2021-01-02	58	95	86	38	64	80	SEDANG	January
2	2021-01-03	64	95	93	25	62	86	SEDANG	January
3	2021-01-04	50	95	67	24	31	77	SEDANG	January
4	2021-01-05	59	95	89	24	35	77	SEDANG	January
...	...	...	...	...	...	...	...	...	...
360	2021-12-27	75	121	61	23	40	47	TIDAK SEHAT	December
361	2021-12-28	59	89	53	16	34	33	SEDANG	December
362	2021-12-29	61	98	54	15	37	29	SEDANG	December
363	2021-12-30	60	102	53	17	38	44	TIDAK SEHAT	December
364	2021-12-31	64	90	52	44	37	53	SEDANG	December

365 rows × 9 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 365 entries, 0 to 364
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   tanggal   365 non-null    datetime64[ns]
 1   pm10      365 non-null    int64
 2   pm25      365 non-null    int64
 3   so2       365 non-null    int64
 4   co        365 non-null    int64
 5   o3        365 non-null    int64
 6   no2       365 non-null    int64
 7   categori  365 non-null    object
 8   bulan     365 non-null    object
dtypes: datetime64[ns](1), int64(6), object(2)
memory usage: 25.8+ KB
```

Removing unnecessary columns ('critical','location','max')
Filling empty data (column "pm25" with its mean)
Adding new column: "bulan"

Numeric Dataframe for Creating Prediction Model

	pm10	pm25	so2	co	o3	no2
0	43	95	58	29	35	65
1	58	95	86	38	64	80
2	64	95	93	25	62	86
3	50	95	67	24	31	77
4	59	95	89	24	35	77
...	...	...	...	...	...	...
360	75	121	61	23	40	47
361	59	89	53	16	34	33
362	61	98	54	15	37	29
363	60	102	53	17	38	44
364	64	90	52	44	37	53

Summary Report

Versi 1

nilai count setiap partikel untuk setiap bulan

bulan	categori	pm10	pm25	so2	co	o3	no2
January	BAIK	0	0	0	0	0	0
	SEDANG	28	28	28	28	28	28
	TIDAK SEHAT	3	3	3	3	3	3
February	BAIK	0	0	0	0	0	0
	SEDANG	25	25	25	25	25	25
	TIDAK SEHAT	3	3	3	3	3	3
March	BAIK	0	0	0	0	0	0
	SEDANG	24	24	24	24	24	24
	TIDAK SEHAT	7	7	7	7	7	7
April	BAIK	0	0	0	0	0	0
	SEDANG	19	19	19	19	19	19
	TIDAK SEHAT	11	11	11	11	11	11
May	BAIK	0	0	0	0	0	0
	SEDANG	21	21	21	21	21	21
	TIDAK SEHAT	10	10	10	10	10	10
June	BAIK	0	0	0	0	0	0
	SEDANG	12	12	12	12	12	12
	TIDAK SEHAT	18	18	18	18	18	18
July	BAIK	0	0	0	0	0	0
	SEDANG	2	2	2	2	2	2
	TIDAK SEHAT	29	29	29	29	29	29
August	BAIK	0	0	0	0	0	0
	SEDANG	11	11	11	11	11	11
	TIDAK SEHAT	20	20	20	20	20	20
September	BAIK	0	0	0	0	0	0
	SEDANG	14	14	14	14	14	14
	TIDAK SEHAT	16	16	16	16	16	16
October	BAIK	0	0	0	0	0	0
	SEDANG	17	17	17	17	17	17
	TIDAK SEHAT	14	14	14	14	14	14
November	BAIK	3	3	3	3	3	3
	SEDANG	25	25	25	25	25	25
	TIDAK SEHAT	2	2	2	2	2	2
December	BAIK	0	0	0	0	0	0
	SEDANG	25	25	25	25	25	25
	TIDAK SEHAT	6	6	6	6	6	6

Versi 2

nilai tertinggi setiap partikel untuk setiap bulan

bulan	so2	co	o3	no2
January	126.0	47.0	67.0	134.0
February	44.0	26.0	60.0	42.0
March	59.0	25.0	74.0	45.0
April	63.0	30.0	73.0	44.0
May	74.0	31.0	151.0	45.0
June	66.0	30.0	76.0	63.0
July	62.0	24.0	81.0	51.0
August	58.0	20.0	67.0	52.0
September	63.0	20.0	68.0	45.0
October	82.0	28.0	72.0	48.0
November	77.0	23.0	73.0	47.0
Desember	NaN	NaN	NaN	NaN

Versi 3

Dibagi lebih jelas berdasarkan tingkat kategori untuk setiap bulan

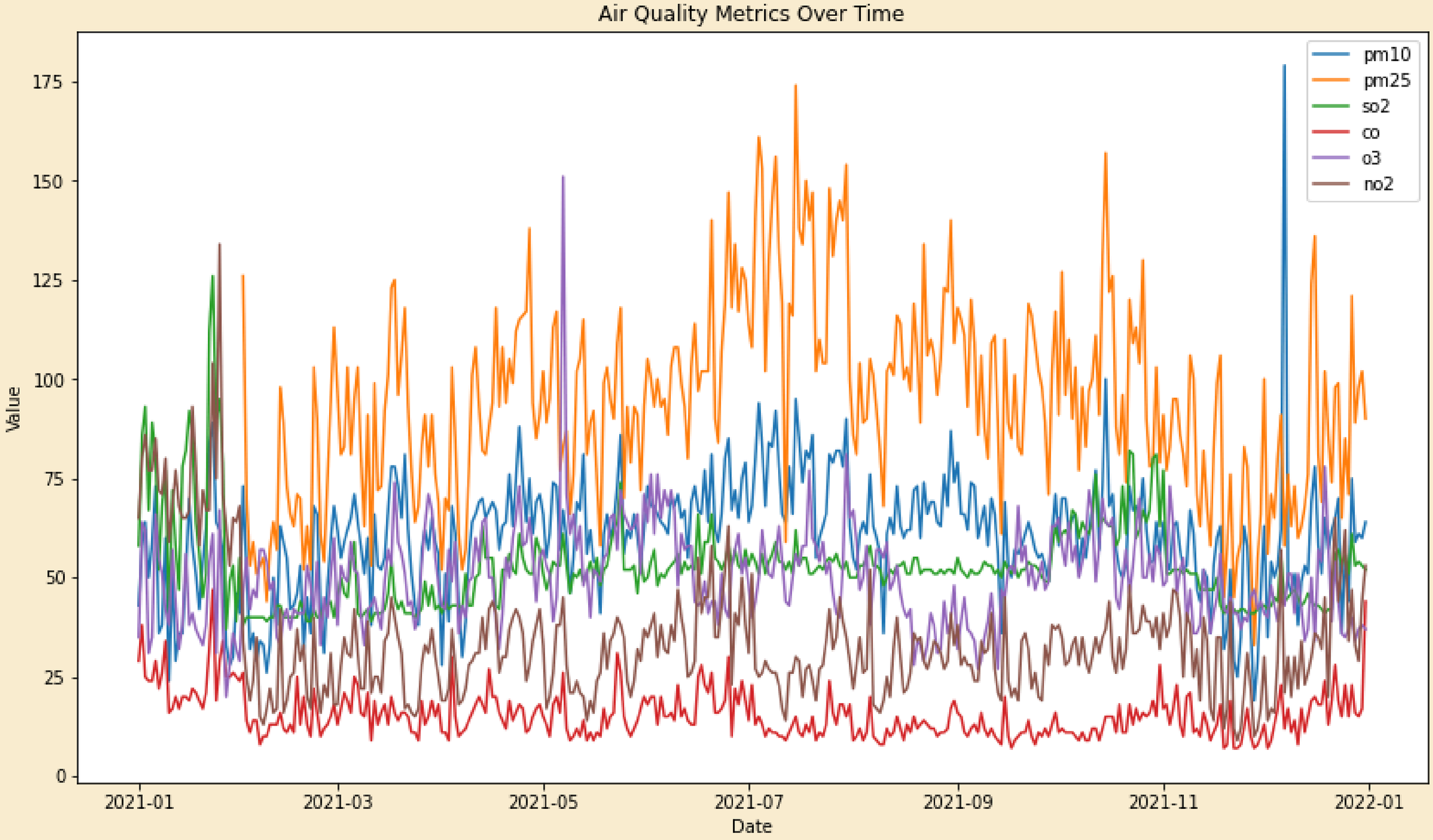
bulan	categori	pm10 max	pm25 max	so2 max	co max	o3 max	no2 max
January	SEDANG	73.0	95.0	93.0	38.0	66.0	93.0
	TIDAK SEHAT	89.0	95.0	126.0	47.0	67.0	134.0
February	SEDANG	66.0	98.0	44.0	25.0	57.0	42.0
	TIDAK SEHAT	73.0	126.0	42.0	26.0	60.0	40.0
March	SEDANG	73.0	99.0	59.0	25.0	71.0	42.0
	TIDAK SEHAT	81.0	125.0	55.0	23.0	74.0	45.0
April	SEDANG	71.0	99.0	63.0	27.0	65.0	44.0
	TIDAK SEHAT	88.0	138.0	61.0	30.0	73.0	42.0
May	SEDANG	66.0	99.0	62.0	20.0	71.0	38.0
	TIDAK SEHAT	86.0	118.0	74.0	31.0	151.0	45.0
June	SEDANG	68.0	100.0	66.0	25.0	76.0	55.0
	TIDAK SEHAT	85.0	147.0	66.0	30.0	72.0	63.0
July	SEDANG	64.0	100.0	52.0	18.0	65.0	29.0
	TIDAK SEHAT	95.0	174.0	62.0	24.0	81.0	51.0
August	SEDANG	64.0	100.0	55.0	13.0	67.0	35.0
	TIDAK SEHAT	87.0	140.0	58.0	20.0	59.0	52.0
September	SEDANG	68.0	99.0	54.0	12.0	68.0	36.0
	TIDAK SEHAT	79.0	120.0	63.0	20.0	63.0	45.0
October	SEDANG	70.0	100.0	80.0	28.0	72.0	39.0
	TIDAK SEHAT	100.0	157.0	82.0	18.0	70.0	48.0
November	BAIK	32.0	45.0	42.0	7.0	47.0	12.0
	SEDANG	64.0	100.0	77.0	23.0	73.0	47.0
	TIDAK SEHAT	67.0	106.0	51.0	21.0	56.0	38.0
December	SEDANG	70.0	100.0	57.0	44.0	58.0	65.0
	TIDAK SEHAT	179.0	136.0	61.0	24.0	78.0	47.0

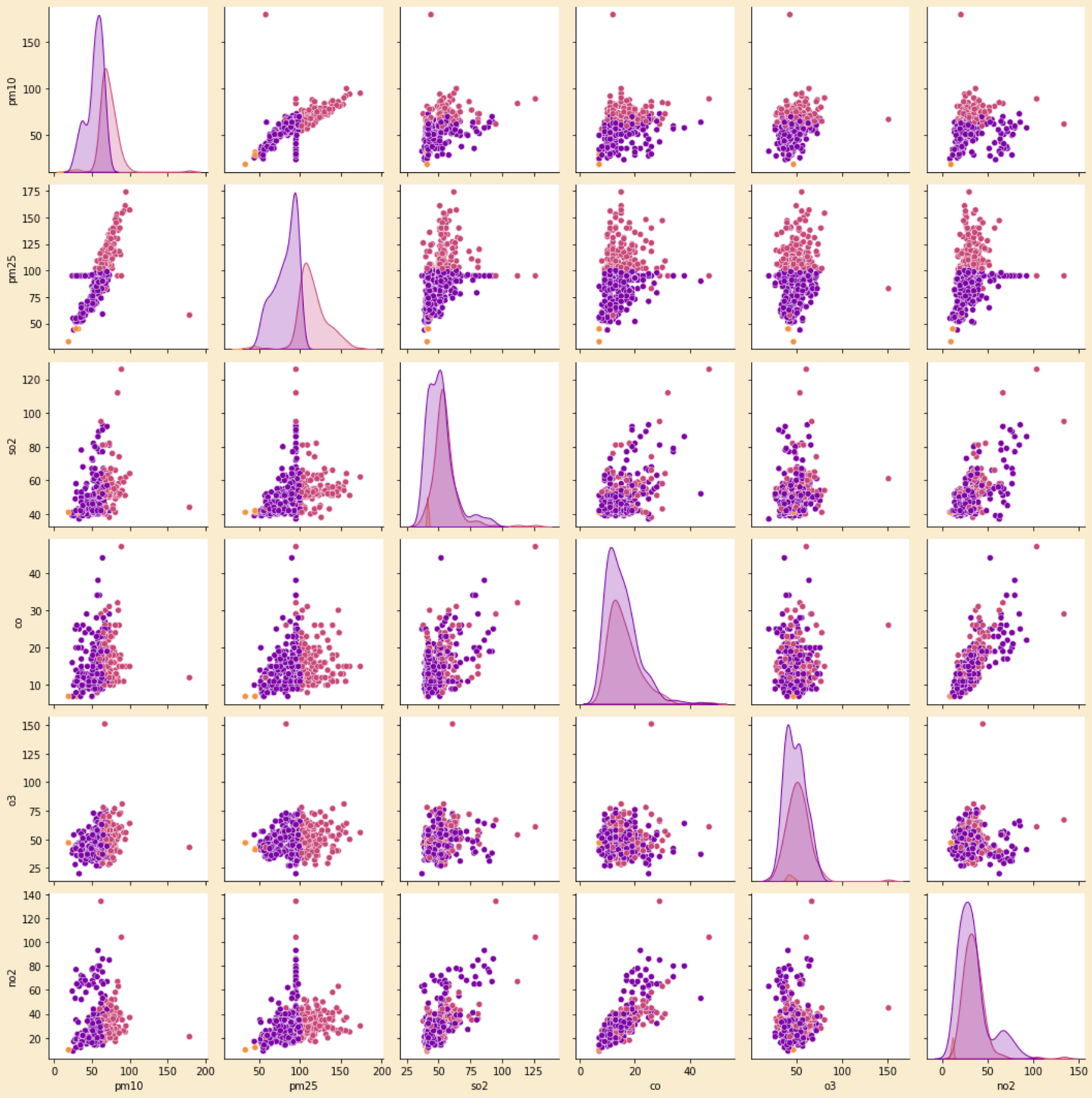
Versi 4

	co			no2			o3			pm10			pm25			so2		
	max	min	mean	max	min	mean	max	min	mean	max	min	mean	max	min	mean	max	min	mean
bulan																		
January	47.0	16.0	24.71	134.0	53.0	74.00	67.0	20.0	42.74	89.0	24.0	51.00	95.0	95.0	95.00	126.0	37.0	70.19
February	26.0	8.0	14.14	42.0	13.0	24.00	60.0	33.0	45.54	73.0	26.0	46.18	126.0	44.0	70.57	44.0	38.0	40.68
March	25.0	9.0	16.16	45.0	15.0	28.32	74.0	33.0	50.16	81.0	38.0	60.10	125.0	53.0	87.94	59.0	39.0	44.19
April	30.0	9.0	15.90	44.0	18.0	31.63	73.0	32.0	52.10	88.0	28.0	61.70	138.0	52.0	91.87	63.0	41.0	49.97
May	31.0	9.0	14.58	45.0	14.0	27.84	151.0	39.0	60.84	86.0	41.0	62.16	118.0	58.0	91.19	74.0	46.0	53.23
June	30.0	10.0	18.67	63.0	23.0	39.37	76.0	38.0	55.33	85.0	55.0	68.87	147.0	80.0	106.40	66.0	48.0	54.43
July	24.0	9.0	13.39	51.0	14.0	28.94	81.0	40.0	54.00	95.0	55.0	76.42	174.0	59.0	129.55	62.0	50.0	53.77
August	20.0	8.0	11.94	52.0	16.0	28.87	67.0	28.0	44.52	87.0	36.0	64.71	140.0	68.0	103.23	58.0	48.0	52.16
September	20.0	7.0	11.57	45.0	19.0	28.67	68.0	27.0	45.47	79.0	36.0	60.97	120.0	61.0	98.87	63.0	49.0	52.37
October	28.0	9.0	13.42	48.0	26.0	34.29	72.0	42.0	57.00	100.0	50.0	63.97	157.0	74.0	101.74	82.0	58.0	65.77
November	23.0	7.0	12.83	47.0	9.0	27.37	73.0	36.0	45.40	67.0	19.0	48.47	106.0	33.0	74.17	77.0	41.0	47.30
Desember	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Data Visualization

Data Visualization

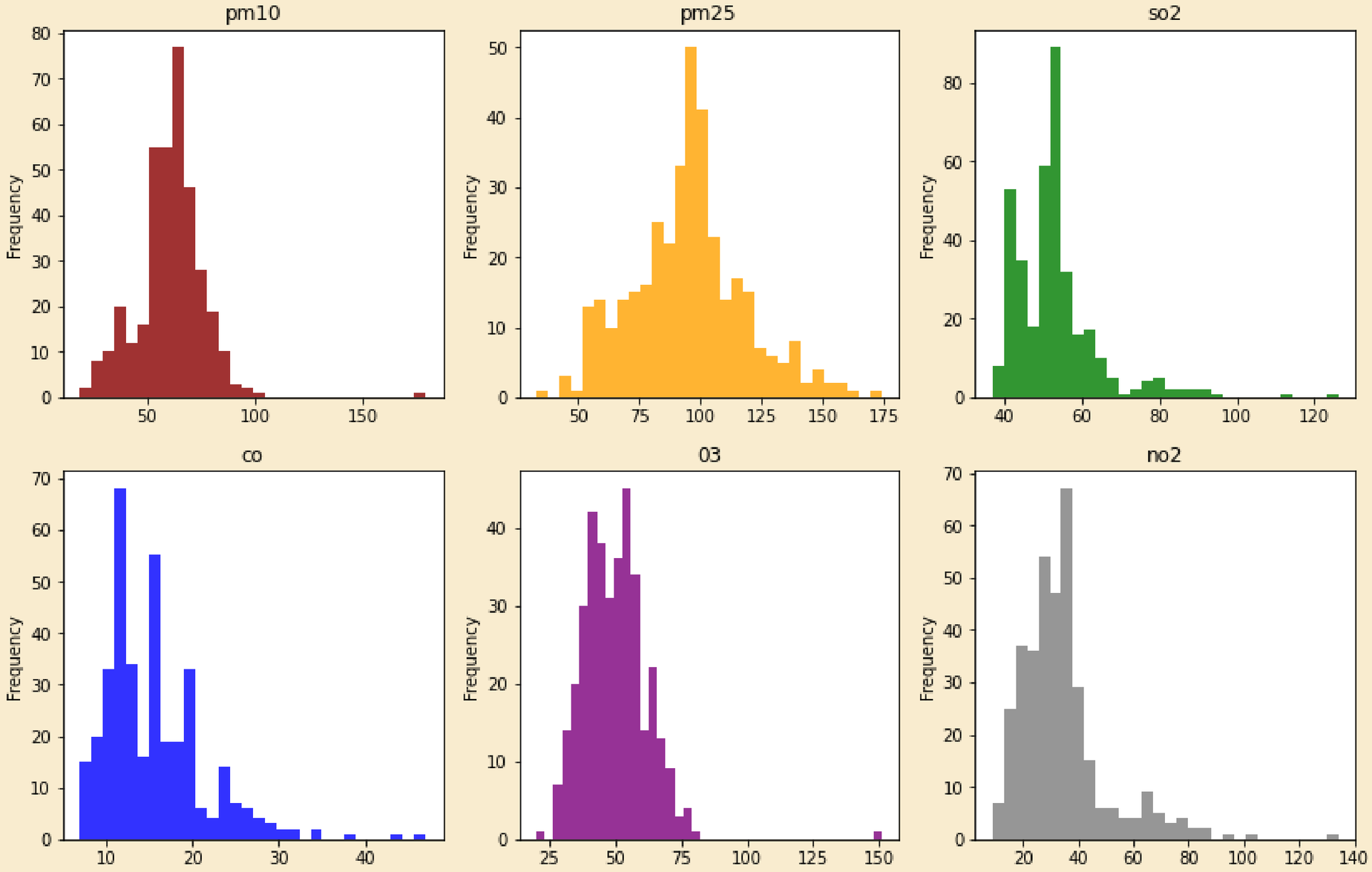




Data Visualization

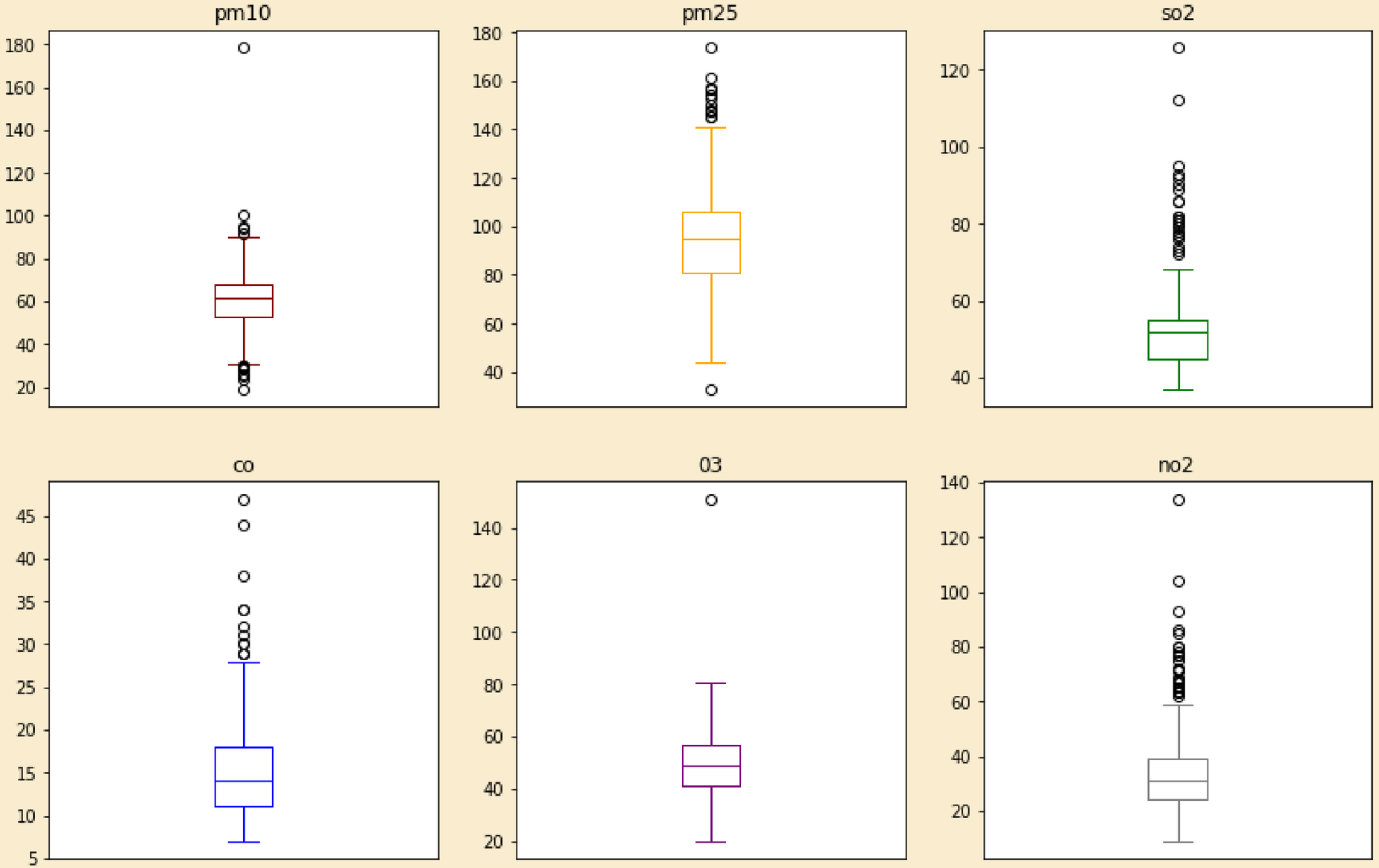
Data Visualization

Distribution of Each Particle



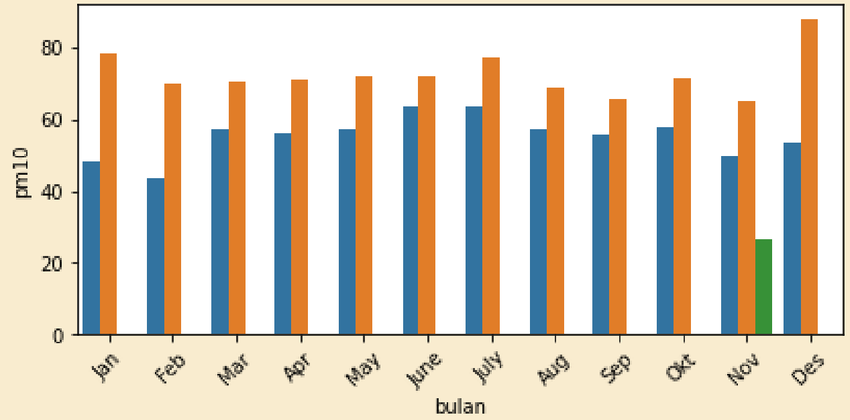
Data Visualization

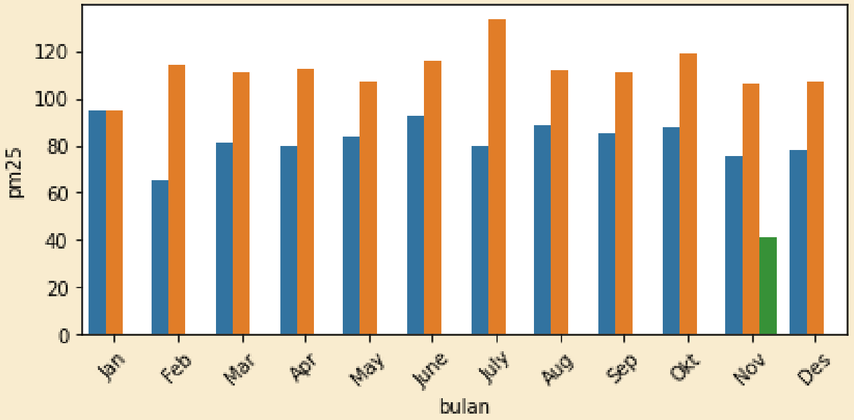
BoxPlot of Each Particle

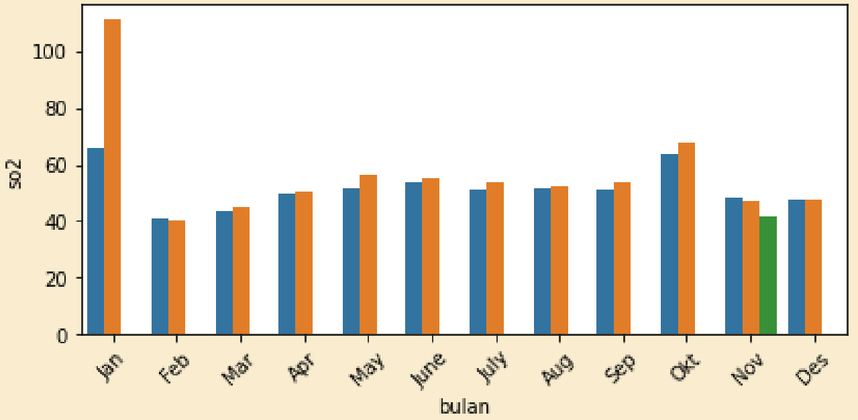


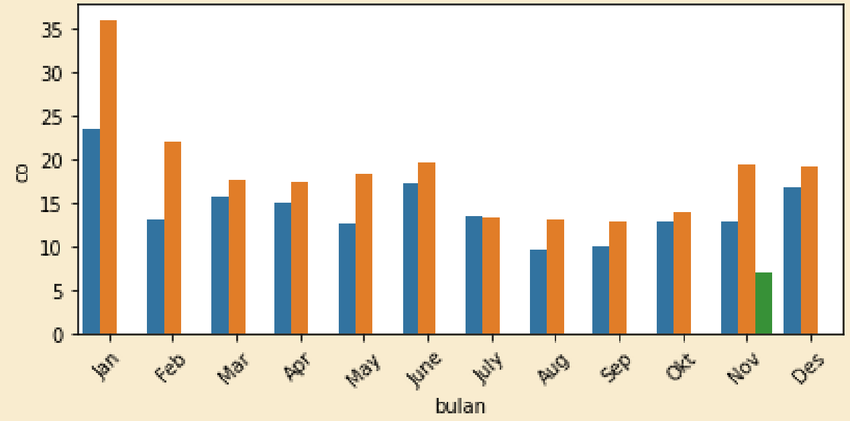
Data Visualization

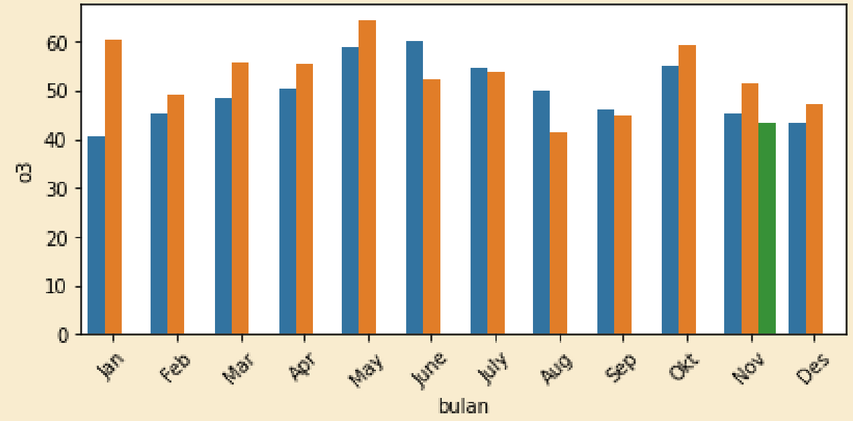
Particle Amount In Each Month

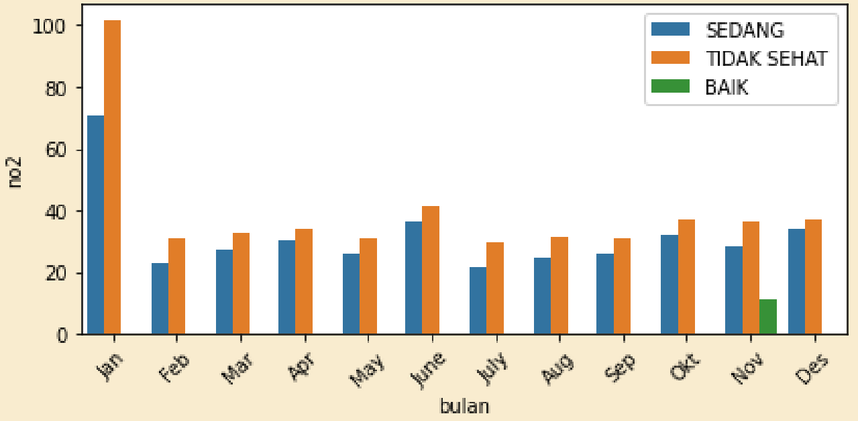












KNN

Linear Regression

Prediction Model

Random Forest

Decision Tree

Linear Regression

```
Column: pm10
Mean Squared Error: 209.333598315352
Root Mean Squared Error: 14.468365433432762

Column: pm25
Mean Squared Error: 561.2925628472929
Root Mean Squared Error: 23.69161376621046

Column: so2
Mean Squared Error: 157.37033114209817
Root Mean Squared Error: 12.54473320330481

Column: co
Mean Squared Error: 16.631166103903112
Root Mean Squared Error: 4.0781326736514

Column: o3
Mean Squared Error: 124.3866221839405
Root Mean Squared Error: 11.152875063585197

Column: no2
Mean Squared Error: 117.67095271959492
Root Mean Squared Error: 10.847624289197839
```

Decision Tree

```
Column: pm10
Mean Squared Error: 232.36363636363637
Root Mean Squared Error: 15.24347848634413

Column: pm25
Mean Squared Error: 235.46818181818182
Root Mean Squared Error: 15.344972525820365

Column: so2
Mean Squared Error: 105.99090909090908
Root Mean Squared Error: 10.29518863794681

Column: co
Mean Squared Error: 32.679545454545455
Root Mean Squared Error: 5.71660261471317

Column: o3
Mean Squared Error: 485.27272727272725
Root Mean Squared Error: 22.028906629080055

Column: no2
Mean Squared Error: 119.5909090909091
Root Mean Squared Error: 10.93576284906129
```

Random Forest

```
Column: pm10
Mean Squared Error: 227.5614958565117
Root Mean Squared Error: 15.08514155904782

Column: pm25
Mean Squared Error: 217.5913693605715
Root Mean Squared Error: 14.75097858992994

Column: so2
Mean Squared Error: 101.83072278744675
Root Mean Squared Error: 10.091120987652797

Column: co
Mean Squared Error: 22.74051056796782
Root Mean Squared Error: 4.768701140558907

Column: o3
Mean Squared Error: 118.46973223916233
Root Mean Squared Error: 10.884380195452671

Column: no2
Mean Squared Error: 70.45284026165982
Root Mean Squared Error: 8.39361902052147
```

KNN

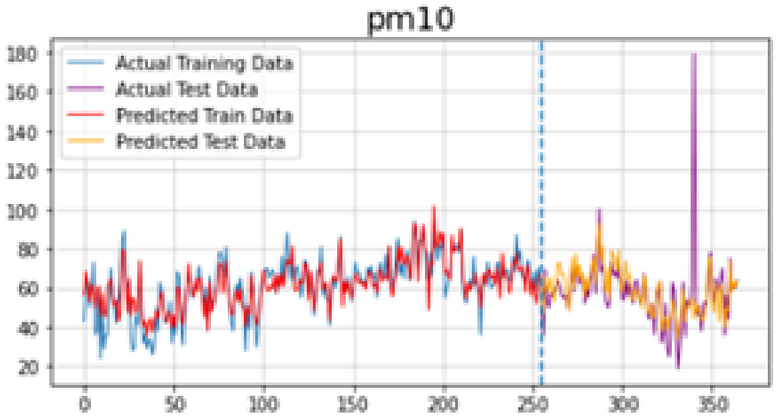
```
Column: pm10
Mean Squared Error: 227.54954545454535
Root Mean Squared Error: 15.084745455411085

Column: pm25
Mean Squared Error: 196.05136363636362
Root Mean Squared Error: 14.001834295418712

Column: so2
Mean Squared Error: 103.56481818181817
Root Mean Squared Error: 10.176680115922784

Column: co
Mean Squared Error: 18.496909090909092
Root Mean Squared Error: 4.300803307628598

Column: o3
Mean Squared Error: 115.50399999999999
Root Mean Squared Error: 10.747278725333217

Column: no2
Mean Squared Error: 88.95745454545454
Root Mean Squared Error: 9.431725957928089
```

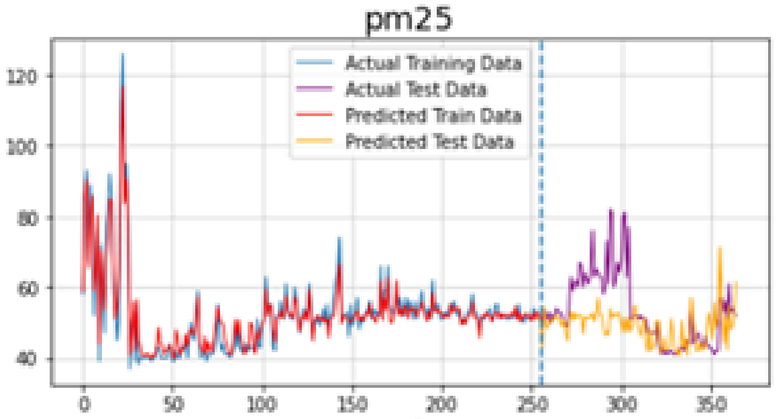
Conclusion

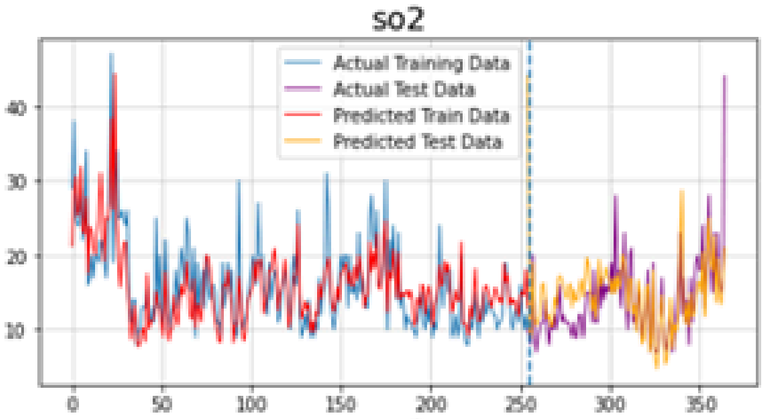
Setiap fitur memiliki metode prediction yang berbeda-beda untuk menghasilkan nilai error yang terendah. Berdasarkan perbandingan dari semua nilai error antar keempat metode, berikut adalah hasil yang terbaik menurut setiap fitur:

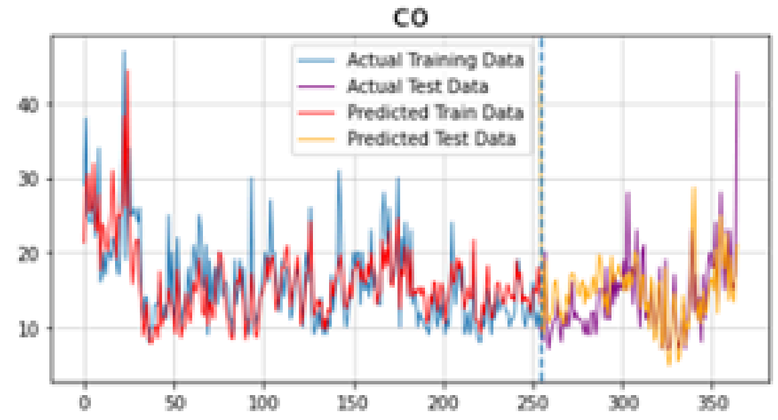
No.	Nama Kolom	Metode	RMSE
1.	pm10	Linear Regression	14.468365433432762
2.	pm25	KNN	14.001834295418712
3.	so2	Random Forest	9.995679087397505
4.	co	Linear Regression	4.07813267365 14
5.	o3	KNN	10.747278725333217
6.	no2	Random Forest	8.637747802697493

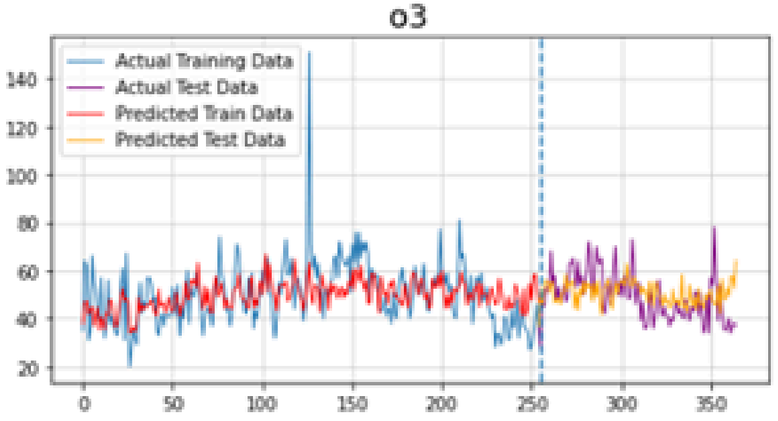
Predictions Using The Most Effective Method

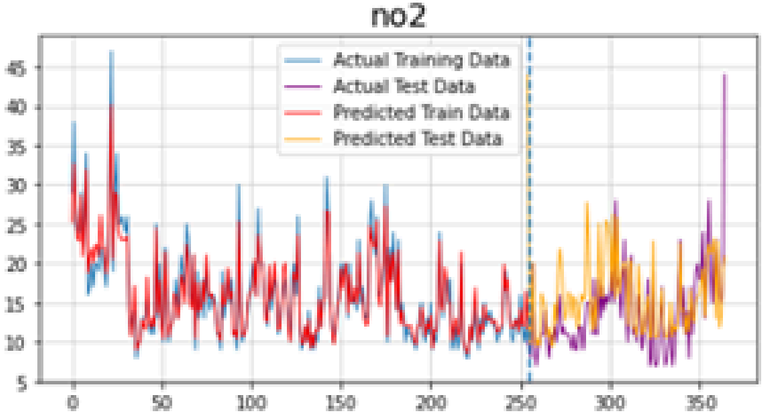













THANKYOU