

**LAPORAN AKHIR UAS
ANALISIS MULTIVARIAT**



Judul:

Prediksi Kualitas Udara di Jakarta Berdasarkan Data Indeks Standar Pencemar Udara (ISPU)
dengan mengkomparasi model klasifikasi menggunakan naive bayes dan ordinal logistic
regression

DISUSUN OLEH KELOMPOK 2 :

Susy Susanty (22031554012)
Nur Halizah Amrita (22031554039)
Leoni Eltania Hotmatua.S (22031554040)
Riva Dian Ardiansyah (22031554043)

**Program Studi S1 Sains Data
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Negeri Surabaya
2024**

BAB 1

PENDAHULUAN

A. Latar Belakang

Kualitas udara merupakan faktor penting yang mempengaruhi kesehatan manusia dan lingkungan. Pencemaran udara, terutama di kota-kota besar seperti Jakarta, dapat memiliki dampak negatif yang signifikan terhadap kesehatan masyarakat dan lingkungan secara keseluruhan. Oleh karena itu, pemantauan dan prediksi kualitas udara menjadi sangat penting untuk mengambil tindakan pencegahan yang sesuai. Salah satu metode yang umum digunakan untuk memantau kualitas udara adalah melalui penggunaan Indeks Standar Pencemar Udara (ISPU). ISPU mengukur tingkat pencemaran udara berdasarkan parameter-partikel tertentu seperti PM10, SO₂, CO, O₃, NO₂, dan Pb. Data ISPU ini dapat digunakan untuk memprediksi kualitas udara di masa depan dengan menggunakan berbagai teknik pemodelan. Pendekatan pemodelan yang umum digunakan dalam memprediksi kualitas udara adalah menggunakan metode klasifikasi, di antaranya adalah Naive Bayes dan Ordinal Logistic Regression. Naive Bayes adalah salah satu metode klasifikasi yang sederhana namun kuat, yang berdasarkan pada teorema Bayes dengan asumsi bahwa fitur-fitur yang digunakan dalam model adalah independen satu sama lain. Metode ini sering digunakan dalam klasifikasi teks, tetapi juga dapat diterapkan dalam prediksi kualitas udara berdasarkan data ISPU. Ordinal Logistic Regression, di sisi lain, adalah pendekatan statistik yang digunakan untuk mengatasi masalah klasifikasi di mana variabel dependen adalah ordinal (memiliki tingkatan atau kelas yang berurutan). Dalam konteks prediksi kualitas udara, kelas-kelas tersebut bisa berupa kategori-kategori seperti "Baik", "Sedang", "Tidak Sehat", dan seterusnya. Penelitian ini bertujuan untuk membandingkan kinerja model klasifikasi Naive Bayes dan Ordinal Logistic Regression dalam memprediksi kualitas udara di Jakarta berdasarkan data ISPU. Dengan memanfaatkan data historis ISPU dan informasi cuaca lainnya, kita dapat mengembangkan model yang dapat membantu dalam memprediksi tingkat pencemaran udara di masa depan, sehingga memberikan kesempatan untuk mengambil tindakan pencegahan yang tepat secara proaktif. Dengan membandingkan kedua metode tersebut, kita dapat mengetahui mana yang memberikan hasil prediksi yang lebih akurat dan dapat diandalkan dalam konteks ini.

B. Rumusan Masalah

- Bagaimana memprediksi kualitas udara di Jakarta menggunakan data ISPU?
- Metode klasifikasi mana yang lebih akurat antara Naive Bayes dan Ordinal Logistic Regression dalam memprediksi kualitas udara di Jakarta?

C. Tujuan Dan Manfaat

Penelitian ini bertujuan untuk membandingkan kinerja model klasifikasi Naive Bayes dan Ordinal Logistic Regression dalam memprediksi kualitas udara di Jakarta berdasarkan data ISPU.

Manfaat Penelitian

- Memberikan wawasan tentang kualitas udara di Jakarta.
- Membantu otoritas dalam mengambil tindakan pencegahan yang tepat.
- Menyediakan model prediksi yang dapat diandalkan untuk kualitas udara.

BAB II TINJAUAN PUSTAKA

Kualitas Udara dan Indeks Standar Pencemar Udara (ISPU)

Kualitas udara merupakan salah satu aspek penting dalam menjaga kesehatan manusia dan lingkungan. Pencemaran udara yang disebabkan oleh berbagai sumber, seperti kendaraan bermotor, industri, dan aktivitas rumah tangga, dapat menimbulkan dampak yang signifikan terhadap kesehatan manusia, seperti penyakit pernapasan, penyakit jantung, dan kanker paru-paru. Selain itu, pencemaran udara juga berdampak buruk pada lingkungan, misalnya hujan asam dan kerusakan ekosistem.

Indeks Standar Pencemar Udara (ISPU) adalah metode yang digunakan untuk menggambarkan tingkat pencemaran udara berdasarkan konsentrasi sejumlah polutan utama di udara. ISPU terdiri dari beberapa parameter, di antaranya PM10, PM2.5, SO₂, CO, O₃, dan NO₂, yang masing-masing memiliki ambang batas yang telah ditetapkan oleh otoritas kesehatan dan lingkungan. Pengukuran ISPU secara rutin membantu dalam pemantauan kualitas udara dan memberikan informasi kepada masyarakat tentang kondisi udara yang mereka hirup.

Metode Klasifikasi dalam Prediksi Kualitas Udara

Dalam prediksi kualitas udara, metode klasifikasi digunakan untuk mengelompokkan data berdasarkan kategori-kategori tertentu, seperti "Baik", "Sedang", "Tidak Sehat", dan "Sangat Tidak Sehat". Dua metode yang umum digunakan dalam penelitian ini adalah Naive Bayes dan Ordinal Logistic Regression.

Naive Bayes

Naive Bayes adalah metode klasifikasi yang berdasarkan pada teorema Bayes dengan asumsi bahwa setiap fitur yang digunakan dalam model adalah independen satu sama lain. Meskipun asumsi ini jarang terjadi dalam dunia nyata, Naive Bayes sering memberikan hasil yang baik dalam berbagai aplikasi, termasuk klasifikasi teks, pengenalan wajah, dan, dalam konteks ini, prediksi kualitas udara. Metode ini dikenal sederhana namun efektif dalam memproses data besar dengan cepat.

Naive Bayes adalah metode klasifikasi berdasarkan teorema Bayes dengan asumsi independensi antara fitur-fitur dalam data. Teorema Bayes dapat dinyatakan sebagai berikut:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Di mana:

- $P(C|X)$ adalah probabilitas hipotesis C yang benar diberikan data X .
- $P(X|C)$ adalah probabilitas mendapatkan data X dengan asumsi hipotesis C benar.
- $P(C)$ adalah probabilitas a priori dari hipotesis C
- $P(X)$ adalah probabilitas dari data X .

Dalam konteks klasifikasi, kita mencari kelas C yang memaksimalkan $P(C|X)$. Naive Bayes mengasumsikan bahwa setiap fitur x_i dalam X adalah independen:

$$P(X|C) = P(x_1|C) \cdot P(x_2|C) \cdot \dots \cdot P(x_n|C)$$

1. Gaussian Naive Bayes

Gaussian Naive Bayes digunakan ketika fitur-fitur dalam data bersifat kontinu dan diasumsikan mengikuti distribusi Gaussian (normal).

1. Probabilitas A Priori:

$$P(C_k) = \frac{N_k}{N}$$

di mana N_k adalah jumlah contoh dalam kelas C_k , dan N adalah jumlah total contoh.

2. Probabilitas Bersyarat:

Jika X_i adalah fitur ke- i dan $X_{i,j}$ adalah nilai dari fitur ke- i untuk contoh ke- j , maka probabilitas bersyarat $P(X_i|C_k)$ diasumsikan mengikuti distribusi normal:

$$P(X_i = x|C_k) = \frac{1}{\sqrt{2\pi\sigma_{C_k,i}^2}} \exp\left(-\frac{(x - \mu_{C_k,i})^2}{2\sigma_{C_k,i}^2}\right)$$

di mana $\mu_{C_k,i}$ dan $\sigma_{C_k,i}$ masing-masing adalah mean dan standard deviation dari fitur X_i dalam kelas C_k .

3. Prediksi:

Untuk data baru $X = (X_1, X_2, \dots, X_n)$, hitung likelihood untuk setiap kelas C_k :

$$P(X|C_k) = \prod_{i=1}^n P(X_i|C_k)$$

dan kemudian gunakan Teorema Bayes untuk menghitung probabilitas posterior:

$$P(C_k|X) \propto P(X|C_k) \cdot P(C_k)$$

Pilih kelas C_k yang memaksimalkan $P(C_k|X)$.

2. Bernoulli Naive Bayes

Bernoulli Naive Bayes digunakan ketika fitur-fitur dalam data bersifat biner (0 atau 1).

1. Probabilitas A Priori:

$$P(C_k) = \frac{N_k}{N}$$

di mana N_k adalah jumlah contoh dalam kelas C_k , dan N adalah jumlah total contoh.

2. Probabilitas Bersyarat:

Jika X_i adalah fitur ke- i , maka probabilitas bersyarat $P(X_i|C_k)$ dihitung sebagai:

$$P(X_i = 1|C_k) = \frac{\text{jumlah contoh dengan } X_i = 1 \text{ dalam kelas } C_k}{N_k}$$

$$P(X_i = 0|C_k) = 1 - P(X_i = 1|C_k)$$

3. Prediksi:

Untuk data baru $X = (X_1, X_2, \dots, X_n)$, hitung likelihood untuk setiap kelas C_k :

$$P(X|C_k) = \prod_{i=1}^n P(X_i|C_k)$$

di mana:

$$P(X_i|C_k) = \begin{cases} P(X_i = 1|C_k) & \text{jika } X_i = 1 \\ P(X_i = 0|C_k) & \text{jika } X_i = 0 \end{cases}$$

dan kemudian gunakan Teorema Bayes untuk menghitung probabilitas posterior:

$$P(C_k|X) \propto P(X|C_k) \cdot P(C_k)$$

Pilih kelas C_k yang memaksimalkan $P(C_k|X)$.

Ordinal Logistic Regression

Ordinal Logistic Regression adalah metode statistik yang digunakan untuk mengatasi masalah klasifikasi di mana variabel dependen adalah ordinal, yang berarti memiliki tingkatan atau kelas yang berurutan. Metode ini mempertimbangkan urutan kategori dalam data dan mampu memberikan prediksi yang lebih akurat untuk masalah klasifikasi ordinal. Dalam penelitian ini, metode ini digunakan untuk memprediksi kategori kualitas udara berdasarkan parameter ISPU.

Ordinal Logistic Regression adalah metode untuk menangani variabel dependen ordinal, di mana kategori memiliki urutan. Model ini memperkirakan hubungan antara variabel independen dan probabilitas dari berbagai kategori dari variabel dependen.

Model dasar dari Ordinal Logistic Regression dapat dinyatakan sebagai:

$$\text{logit}(P(Y \leq j)) = \log \left(\frac{P(Y \leq j)}{1 - P(Y \leq j)} \right) = \alpha_j - \beta X$$

Dimana:

- $\text{logit}(P(Y \leq j))$ adalah logit atau log-odds dari probabilitas bahwa respons Y berada pada atau di bawah kategori j .
- α_j adalah cutpoint untuk kategori j .
- β adalah koefisien regresi yang memperkirakan pengaruh variabel independen X .

1. **Ordinal Logistic Regression (All-Threshold Variant)**

Pendekatan ini mengasumsikan bahwa ada satu set threshold yang membagi skala laten menjadi kategori-kategori ordinal. Misalkan ada J kategori ordinal. OLR dengan varian ini menggunakan multiple threshold (cut-off points) untuk memodelkan probabilitas bahwa pengamatan termasuk dalam kategori tertentu.

Rumus:

1. Model Laten:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}$$

di mana η_i adalah nilai laten untuk pengamatan ke- i , X_{ij} adalah nilai fitur ke- j untuk pengamatan ke- i , dan β adalah koefisien regresi.

2. Thresholds (Cut-off Points):

Ada $J - 1$ threshold yang membagi skala laten menjadi kategori-kategori ordinal:

$$\gamma_1, \gamma_2, \dots, \gamma_{J-1}$$

3. Probabilitas Kategori:

Probabilitas bahwa pengamatan ke- i termasuk dalam kategori ke- j diberikan oleh:

$$P(Y_i \leq j) = \frac{1}{1 + \exp[-(\gamma_j - \eta_i)]}$$

untuk $j = 1, 2, \dots, J - 1$.

Probabilitas bahwa pengamatan ke- i termasuk dalam kategori tepat ke- j adalah:

$$P(Y_i = j) = P(Y_i \leq j) - P(Y_i \leq j - 1)$$

dengan

$$P(Y_i \leq 0) = 0 \quad \text{dan} \quad P(Y_i \leq J) = 1$$

2. Ordinal Logistic Regression (All-Threshold Variant)

pendekatan ini memperkenalkan threshold (cut-off points) secara langsung untuk setiap kategori, tanpa model laten terpisah. Model ini juga dikenal sebagai model proportional odds.

Rumus:

1. Model Laten:

Sama seperti pada varian sebelumnya:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}$$

2. Immediate Thresholds (Cut-off Points):

Sama seperti pada varian sebelumnya:

$$\gamma_1, \gamma_2, \dots, \gamma_{J-1}$$

3. Odds Rasio:

Odds rasio bahwa pengamatan ke- i termasuk dalam kategori j atau lebih besar dibandingkan kategori $j - 1$:

$$\log \left(\frac{P(Y_i \leq j)}{P(Y_i > j)} \right) = \gamma_j - \eta_i$$

di mana $P(Y_i \leq j)$ adalah probabilitas bahwa pengamatan ke- i termasuk dalam kategori ke- j atau lebih rendah.

4. Probabilitas Kategori:

Probabilitas bahwa pengamatan ke- i termasuk dalam kategori ke- j diberikan oleh:

$$P(Y_i \leq j) = \frac{1}{1 + \exp[-(\gamma_j - \eta_i)]}$$

Probabilitas bahwa pengamatan ke- i termasuk dalam kategori tepat ke- j adalah:

$$P(Y_i = j) = P(Y_i \leq j) - P(Y_i \leq j - 1)$$

dengan

$$P(Y_i \leq 0) = 0 \quad \text{dan} \quad P(Y_i \leq J) = 1$$

Implementasi

Di Jakarta, kualitas udara sering kali berada pada level yang tidak sehat terutama pada musim kemarau ketika polusi kendaraan bermotor dan industri mencapai puncaknya. Dengan menggunakan data historis ISPU yang tersedia dari portal data terbuka Indonesia, penelitian ini bertujuan untuk membandingkan dua metode klasifikasi, yaitu Naive Bayes dan Ordinal Logistic Regression, dalam memprediksi kualitas udara. Hasil penelitian ini diharapkan dapat memberikan wawasan yang lebih baik mengenai metode mana yang lebih efektif dan dapat diandalkan dalam memprediksi kualitas udara di Jakarta.

BAB III PEMBAHASAN

A. Dataset

Dataset yang digunakan berasal dari portal data terbuka Indonesia dan berfokus pada Indeks Standar Pencemar Udara (ISPU) di Provinsi DKI Jakarta. Dataset ini menyediakan informasi tentang kualitas udara di Jakarta berdasarkan berbagai parameter pencemar udara, seperti PM10, SO2, CO, O3, NO2, PM2.5. Dataset ini memiliki 11 kolom yang mencakup berbagai aspek terkait kualitas udara. Berikut adalah link dataset (<https://katalog.data.go.id/dataset/data-indeks-standar-pencemar-udara-ispu-di-provinsi-dki-jakarta1>)

B. Data Preparation

- DataFrame memiliki 1804 baris dan 11 kolom



- Describe Statistics

```
[ ] print(df.describe())
```

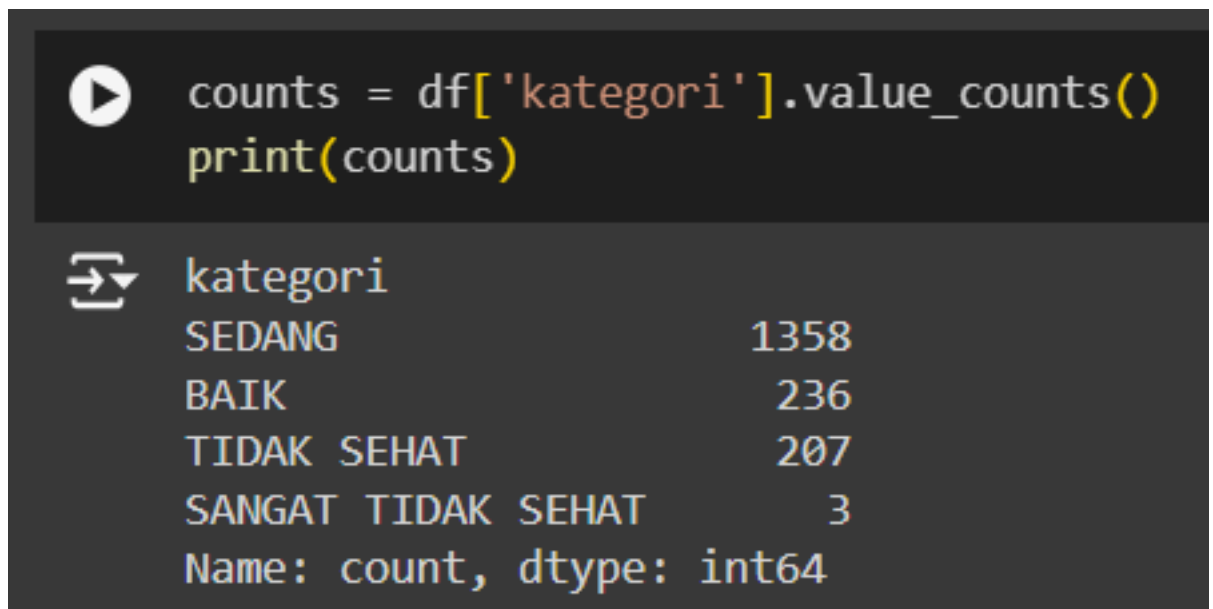
	pm10	pm25	so2	co	o3	\
count	1804.000000	1804.000000	1804.000000	1804.000000	1804.000000	
mean	46.745565	65.973392	37.684590	12.489468	28.138581	
std	22.307589	35.565152	13.860799	6.742174	13.168083	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	35.000000	50.000000	26.000000	8.000000	19.000000	
50%	52.000000	73.000000	38.000000	11.000000	26.000000	
75%	62.000000	88.000000	51.000000	16.000000	35.000000	
max	163.000000	287.000000	89.000000	55.000000	81.000000	

	no2	max
count	1804.000000	1804.000000
mean	17.350333	74.315410
std	9.056261	23.927575
min	0.000000	20.000000
25%	12.000000	57.000000
50%	17.000000	74.000000
75%	23.000000	88.000000
max	53.000000	287.000000

Contoh Interpretasi Hasil Statistik dari faktor polutan :

- pm10: Rata-rata 46.74 $\mu\text{g}/\text{m}^3$ dengan variasi yang besar (std deviasi 22.30). Nilai maksimum sangat tinggi (163 $\mu\text{g}/\text{m}^3$) menunjukkan adanya beberapa hari dengan polusi PM10 yang sangat tinggi.
- pm25: Rata-rata 65.97 $\mu\text{g}/\text{m}^3$ dengan variasi yang signifikan (std deviasi 35.56). Nilai maksimum 287 $\mu\text{g}/\text{m}^3$ menunjukkan adanya beberapa hari dengan polusi PM2.5 yang ekstrem.

- so2: Rata-rata 37.68 $\mu\text{g}/\text{m}^3$. Distribusi data cukup simetris dengan median 38 $\mu\text{g}/\text{m}^3$ dan standar deviasi 13.86.
- co: Rata-rata 12.48 $\mu\text{g}/\text{m}^3$ dengan nilai minimum 0 dan maksimum 55 $\mu\text{g}/\text{m}^3$. Variasi sedang (std deviasi 6.74).
- o3: Rata-rata 28.13 $\mu\text{g}/\text{m}^3$, nilai minimum 0 dan maksimum 81 $\mu\text{g}/\text{m}^3$, dengan variasi yang lebih rendah dibandingkan pm25 dan pm10.
- no2: Rata-rata 17.35 $\mu\text{g}/\text{m}^3$ dengan variasi yang lebih kecil (std deviasi 9.05), menunjukkan distribusi yang lebih seragam
- Menghitung frekuensi kemunculan setiap nilai unik dalam kolom 'kategori' pada DataFrame



```
counts = df['kategori'].value_counts()
print(counts)
```

kategori	count
SEDANG	1358
BAIK	236
TIDAK SEHAT	207
SANGAT TIDAK SEHAT	3

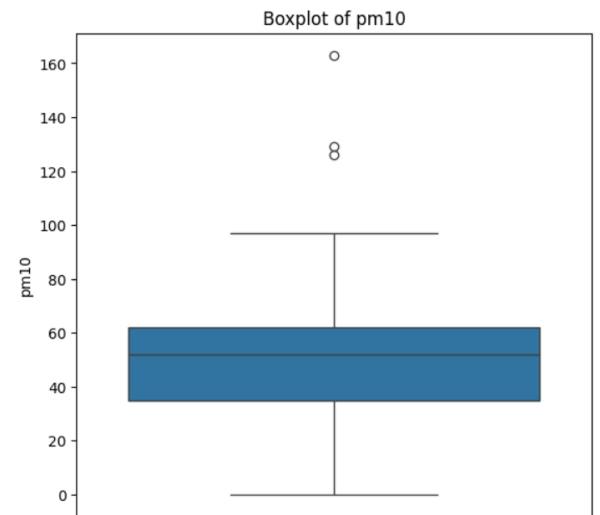
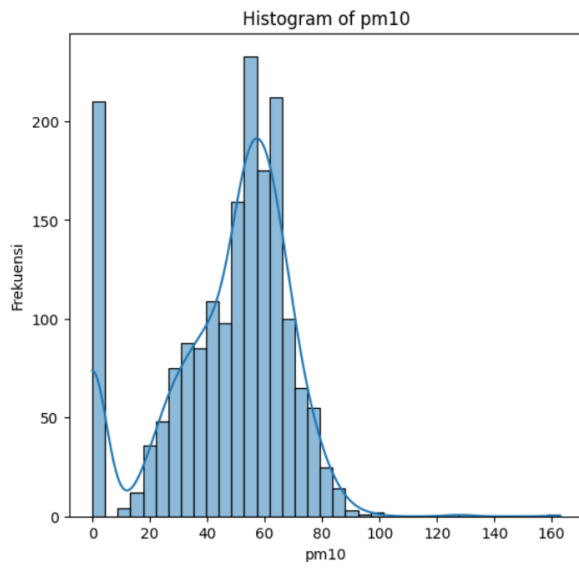
Name: count, dtype: int64

Gambar di atas hasil dari revisi yang dimana sebelumnya terdapat 5 label menjadi 4 label untuk label yang dihilangkan yaitu Tidak Ada Data.

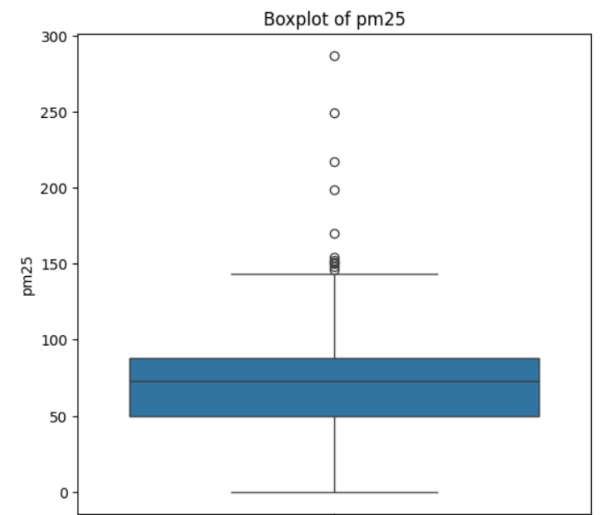
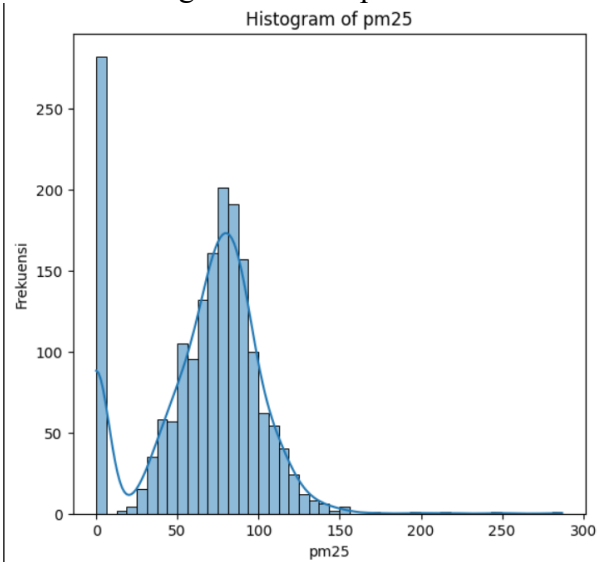
C. Visualisasi Data

Terdapat beberapa nilai outlier terhadap fitur yang memengaruhi kualitas udara mulai dari PM10, PM2.5, SO2, CO, O3, NO2

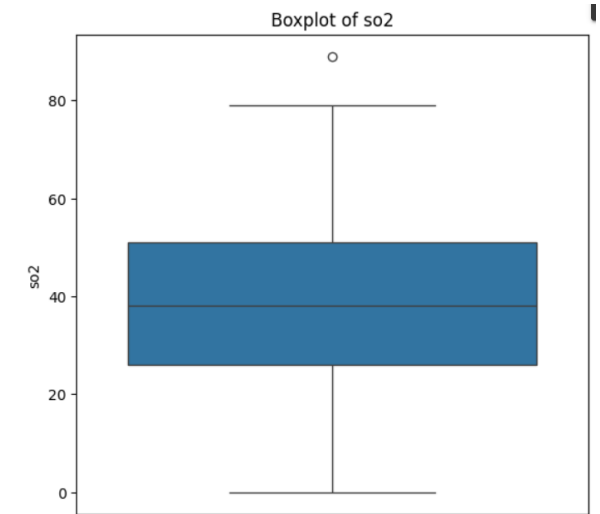
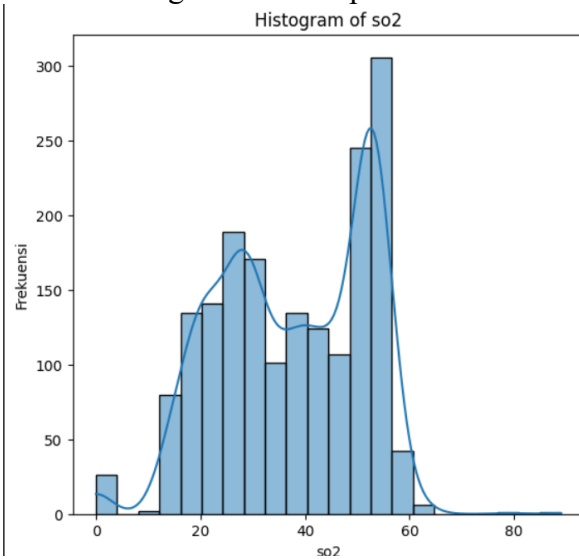
- Histogram dan Boxplot PM10



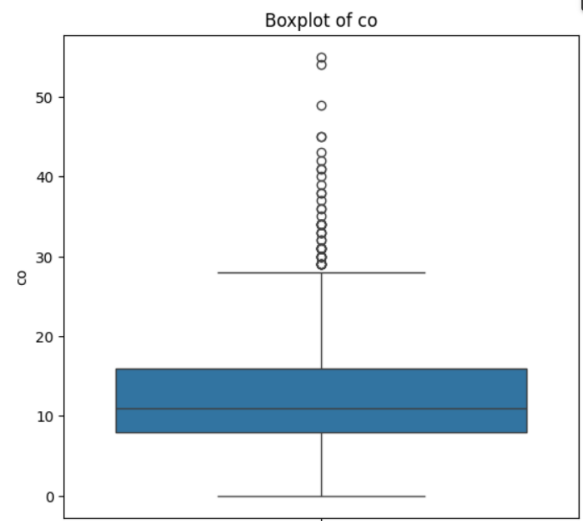
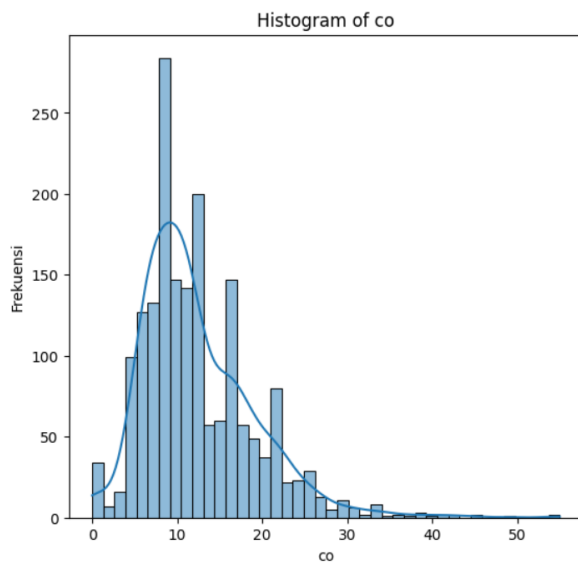
- Histogram dan Boxplot PM2.5



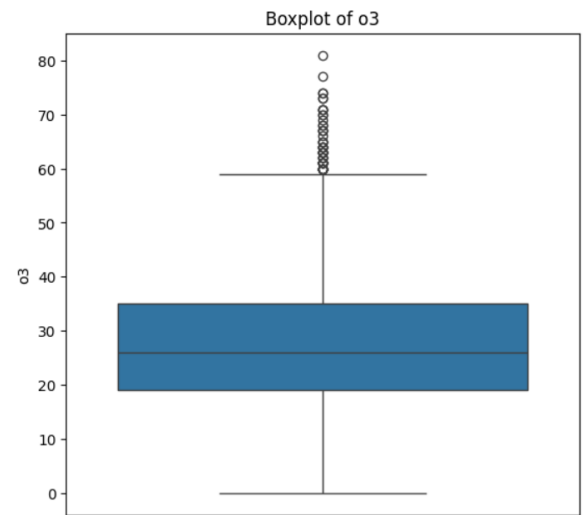
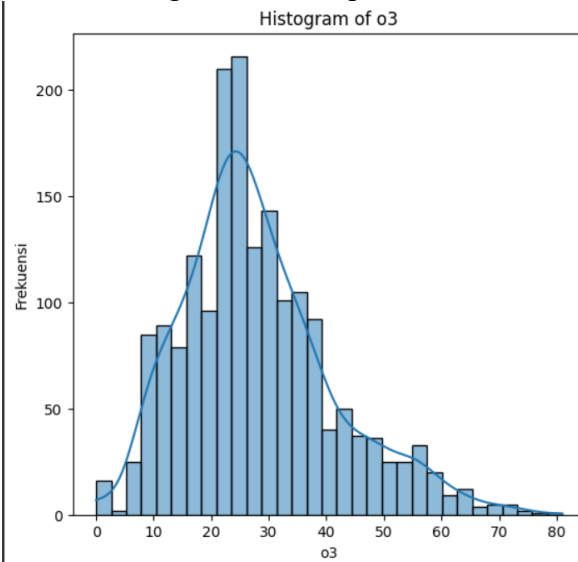
- Histogram dan Boxplot SO2



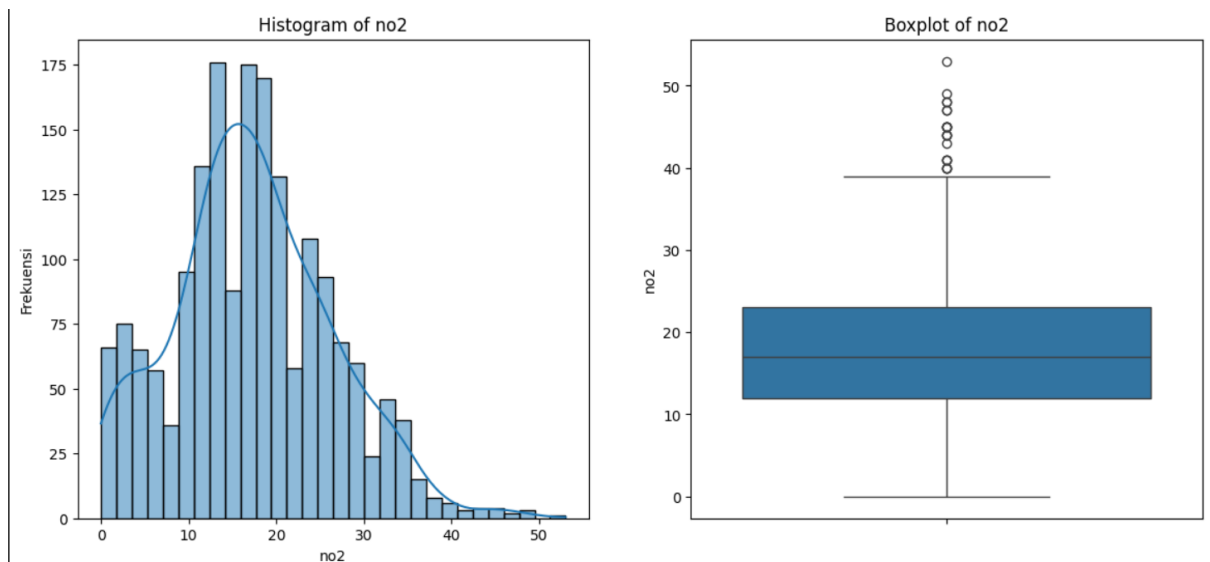
- Histogram dan Boxplot CO



- Histogram dan Boxplot O3



- Histogram dan Boxplot NO2



D. Uji Matrix

- Correlation Matrix

Correlation Matrix:

	pm10	pm25	so2	co	o3	no2	max
pm10	0.000000	0.182732	0.192460	-0.157676	0.281841	0.304722	0.240824
pm25	0.182732	0.000000	0.335586	0.122601	-0.143092	0.289104	0.880006
so2	0.192460	0.335586	0.000000	-0.178192	-0.236414	0.070369	0.200393
co	-0.157676	0.122601	-0.178192	0.000000	0.122177	0.219875	0.311527
o3	0.281841	-0.143092	-0.236414	0.122177	0.000000	0.302107	0.081868
no2	0.304722	0.289104	0.070369	0.219875	0.302107	0.000000	0.328262
max	0.240824	0.880006	0.200393	0.311527	0.081868	0.328262	0.000000

- P-Value Matrix

P-Value Matrix:

	pm10	pm25	so2	co	o3	\
pm10	0.000000e+00	4.534637e-15	1.405187e-16	1.483369e-11	1.949491e-34	
pm25	4.534637e-15	0.000000e+00	6.102898e-49	1.650058e-07	9.453324e-10	
so2	1.405187e-16	6.102898e-49	0.000000e+00	2.152249e-14	1.950608e-24	
co	1.483369e-11	1.650058e-07	2.152249e-14	0.000000e+00	1.820818e-07	
o3	1.949491e-34	9.453324e-10	1.950608e-24	1.820818e-07	0.000000e+00	
no2	3.048361e-40	3.194693e-36	2.725593e-03	2.819714e-21	1.495255e-39	
max	2.541723e-25	0.000000e+00	7.188913e-18	4.501953e-42	4.859887e-04	

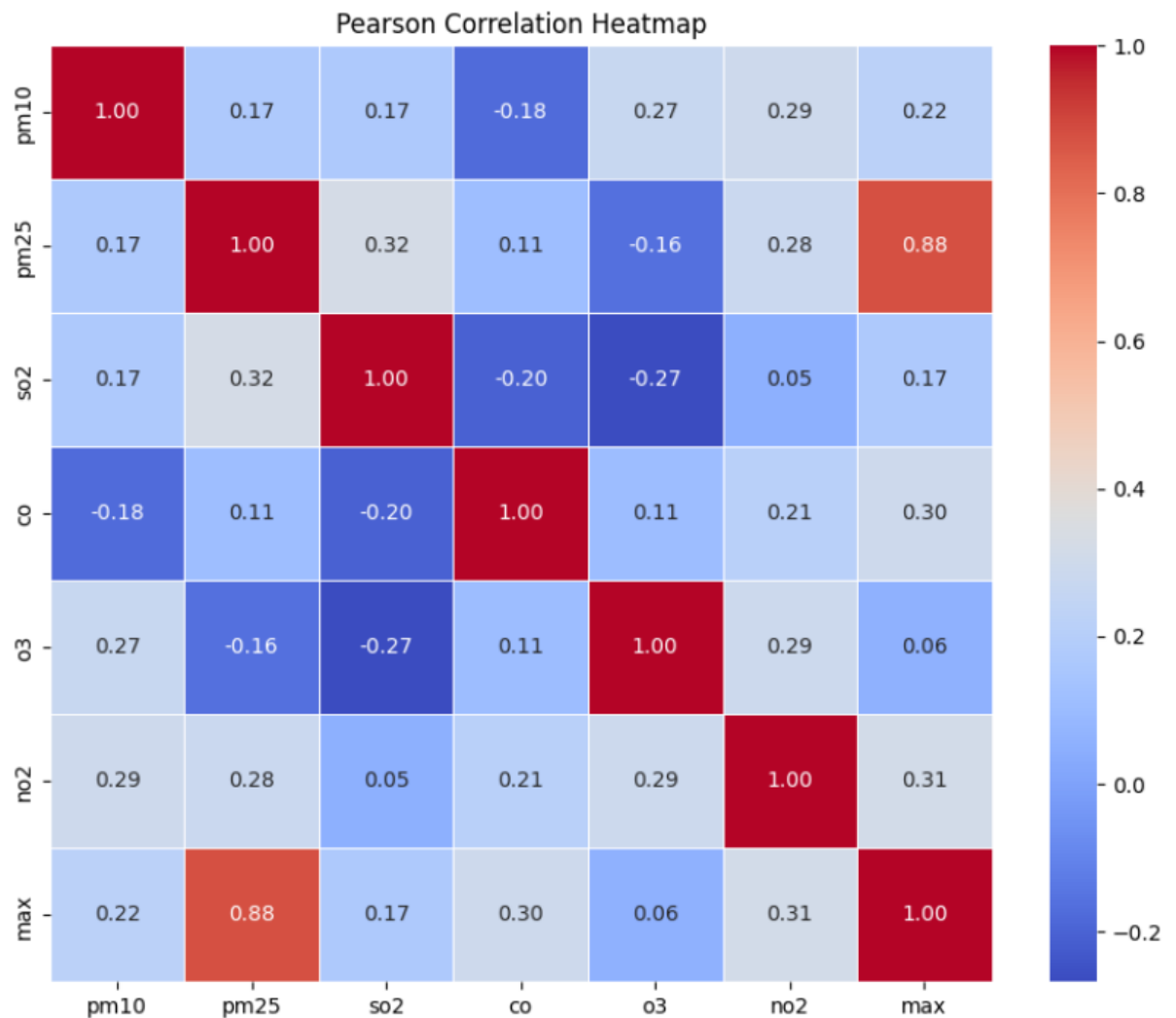
	no2	max
pm10	3.048361e-40	2.541723e-25
pm25	3.194693e-36	0.000000e+00
so2	2.725593e-03	7.188913e-18
co	2.819714e-21	4.501953e-42
o3	1.495255e-39	4.859887e-04
no2	0.000000e+00	8.762478e-47
max	8.762478e-47	0.000000e+00

- menghitung koefisien korelasi Pearson dan nilai p (p-value) antara dua variabel.

Korelasi antara PM10 dan PM2.5: 0.18, p-value: 0.00
 Tolak hipotesis nol, ada korelasi yang signifikan antara PM10 dan PM2.5
 Korelasi antara PM10 dan SO2: 0.19, p-value: 0.00
 Tolak hipotesis nol, ada korelasi yang signifikan antara PM10 dan SO2
 Korelasi antara PM2.5 dan SO2: 0.34, p-value: 0.00
 Tolak hipotesis nol, ada korelasi yang signifikan antara PM2.5 dan SO2

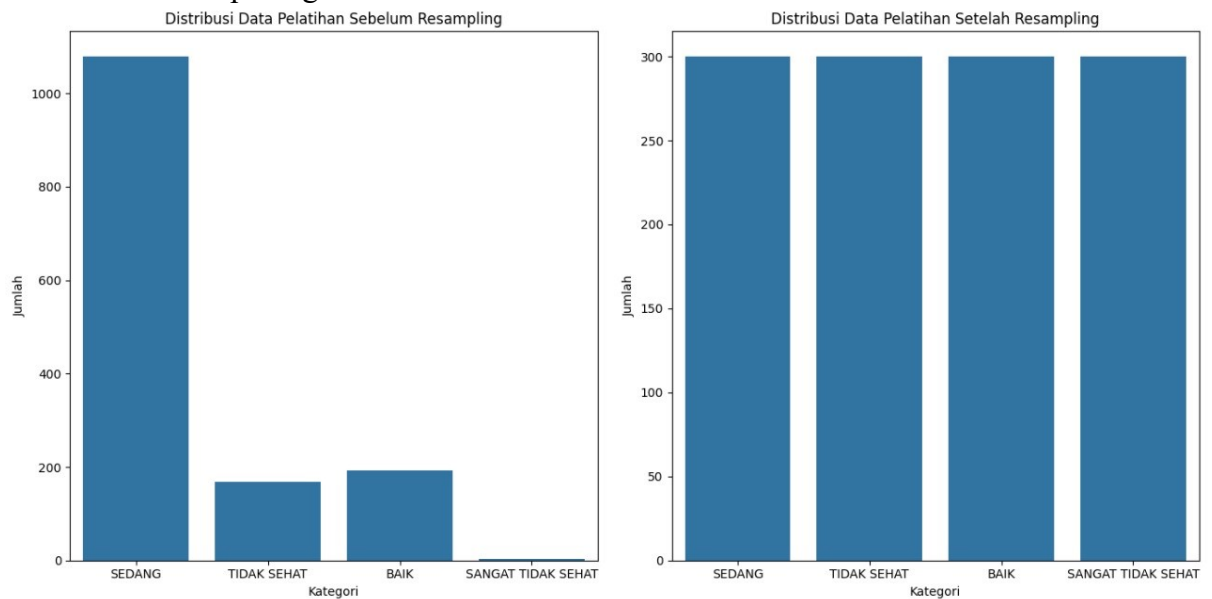
E. Uji Korelasi

Uji korelasi antar fitur polutan dengan menggunakan pearson correlation heatmap



F. Data Resample

Untuk data resample digunakan dat



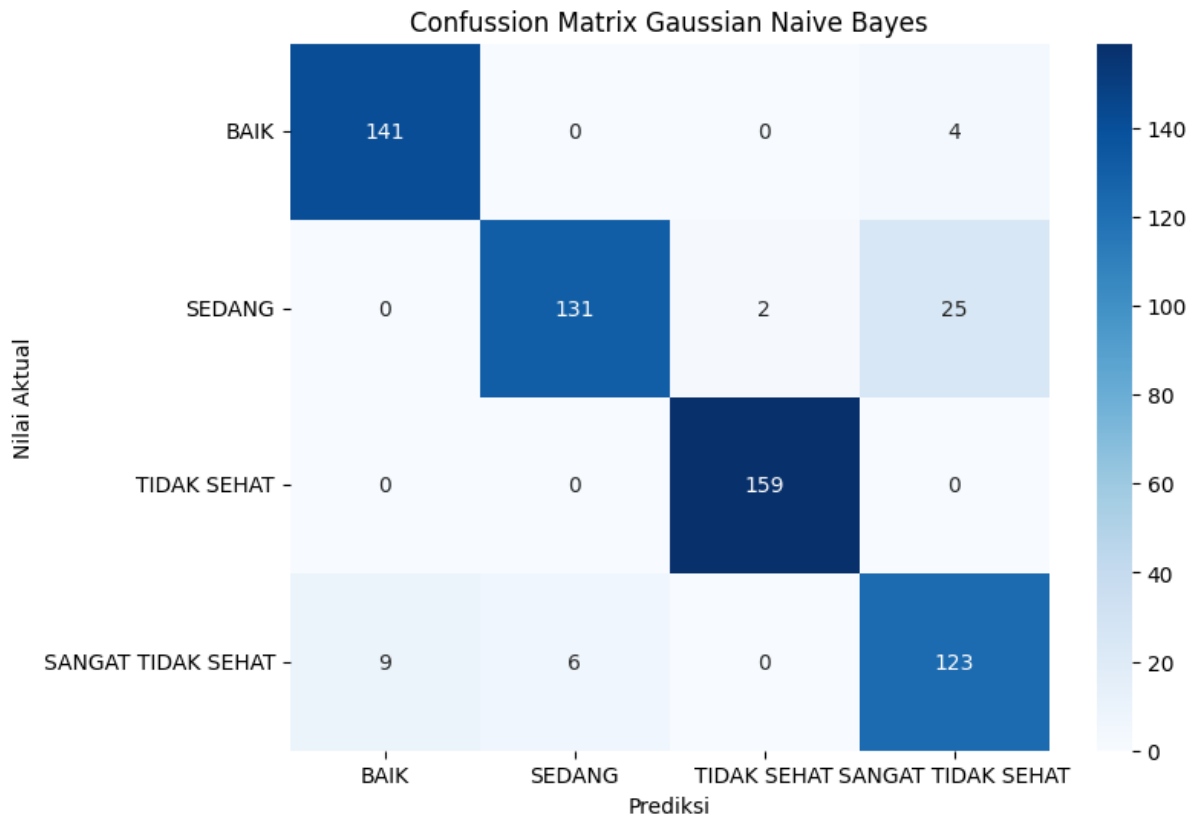
G. Estimasi Parameter

Dengan parameter test_size sebesar = 0.3, random_state = 100

- Model Gaussian Naive Bayes

Laporan Evaluasi Model Gaussian Naive Bayes:					
	precision	recall	f1-score	support	
BAIK	0.94	0.97	0.96	145	
SANGAT TIDAK SEHAT	0.96	0.83	0.89	158	
SEDANG	0.99	1.00	0.99	159	
TIDAK SEHAT	0.81	0.89	0.85	138	
accuracy			0.92	600	
macro avg	0.92	0.92	0.92	600	
weighted avg	0.93	0.92	0.92	600	

Confussion Matrix (Gaussian Naive Bayes)



Nilai RMSE dan MSE Gaussian Naive Bayes

```
Gaussian Naive Bayes MSE: 0.15166666666666667
Gaussian Naive Bayes RMSE: 0.3894440481849308
Gaussian Naive Bayes MAE: 0.10166666666666667
```

- Model Bernoulli Naive Bayes

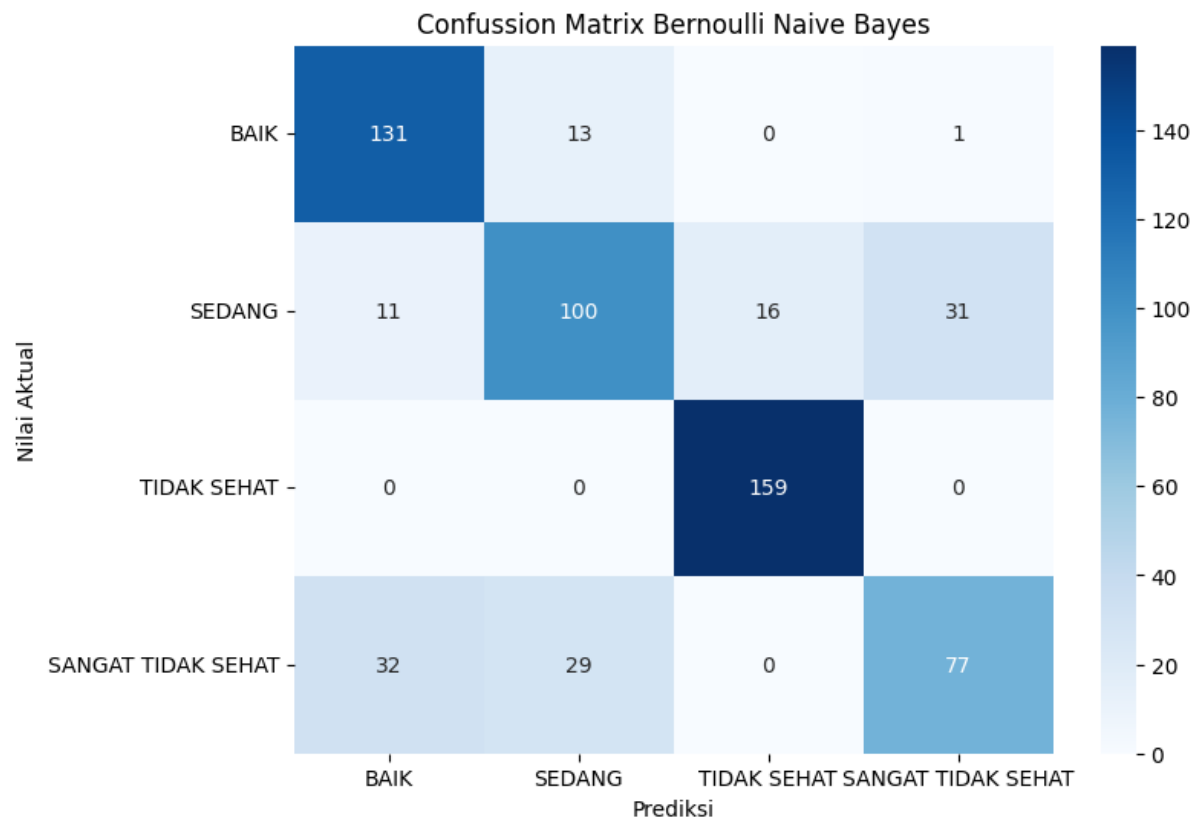
Dengan `test_size = 0.3` dan `random_state = 100`

```
Laporan Evaluasi Model Bernoulli Naive Bayes:
              precision    recall  f1-score   support

   BAIK               0.75        0.90        0.82         145
SANGAT TIDAK SEHAT    0.70        0.63        0.67         158
   SEDANG              0.91        1.00        0.95         159
   TIDAK SEHAT        0.71        0.56        0.62         138

 accuracy               0.78
macro avg               0.77        0.77        0.77         600
weighted avg            0.77        0.78        0.77         600
```

Confussion Matrix Bernoulli Naive Bayes



Nilai RMSE dan MSE

```
Bernoulli Naive Bayes MSE: 0.7866666666666666
Bernoulli Naive Bayes RMSE: 0.3894440481849308
Bernoulli Naive Bayes MAE: 0.383333333333333336
```

Model Ordinal Logistik Regresi (OrderedModel)

```
Akurasi Ordinal Logistic Regression: 0.505
Laporan Klasifikasi:
      precision    recall  f1-score   support

0         0.78        0.72        0.75         145
1         0.48        0.48        0.48         159
2         0.26        0.33        0.29         138
3         0.59        0.48        0.53         158

 accuracy          0.51         600
macro avg          0.53         600
weighted avg       0.53         600
```


Tabel Statistik

Optimization terminated successfully.

Current function value: 1.155387

Iterations: 23

Function evaluations: 24

Gradient evaluations: 24

OrderedModel Results

Dep. Variable:	kategori_encoded	Log-Likelihood:	-1617.5
Model:	OrderedModel	AIC:	3253.
Method:	Maximum Likelihood	BIC:	3300.
Date:	Mon, 10 Jun 2024		
Time:	01:12:34		
No. Observations:	1400		
Df Residuals:	1391		
Df Model:	6		

	coef	std err	z	P> z	[0.025	0.975]
x1	0.2558	0.064	4.002	0.000	0.131	0.381
x2	-0.4554	0.077	-5.919	0.000	-0.606	-0.305
x3	1.1094	0.066	16.774	0.000	0.980	1.239
x4	1.3460	0.084	16.079	0.000	1.182	1.510
x5	0.2656	0.061	4.339	0.000	0.146	0.386
x6	0.4050	0.064	6.346	0.000	0.280	0.530
0/1	-1.3711	0.075	-18.393	0.000	-1.517	-1.225
1/2	0.4158	0.049	8.400	0.000	0.319	0.513
2/3	0.4237	0.049	8.597	0.000	0.327	0.520

- Model Ordinal Logistik Regresi (All-Threshold Variant)
Dengan parameter test_size sebesar = 0.3, random_state = 100

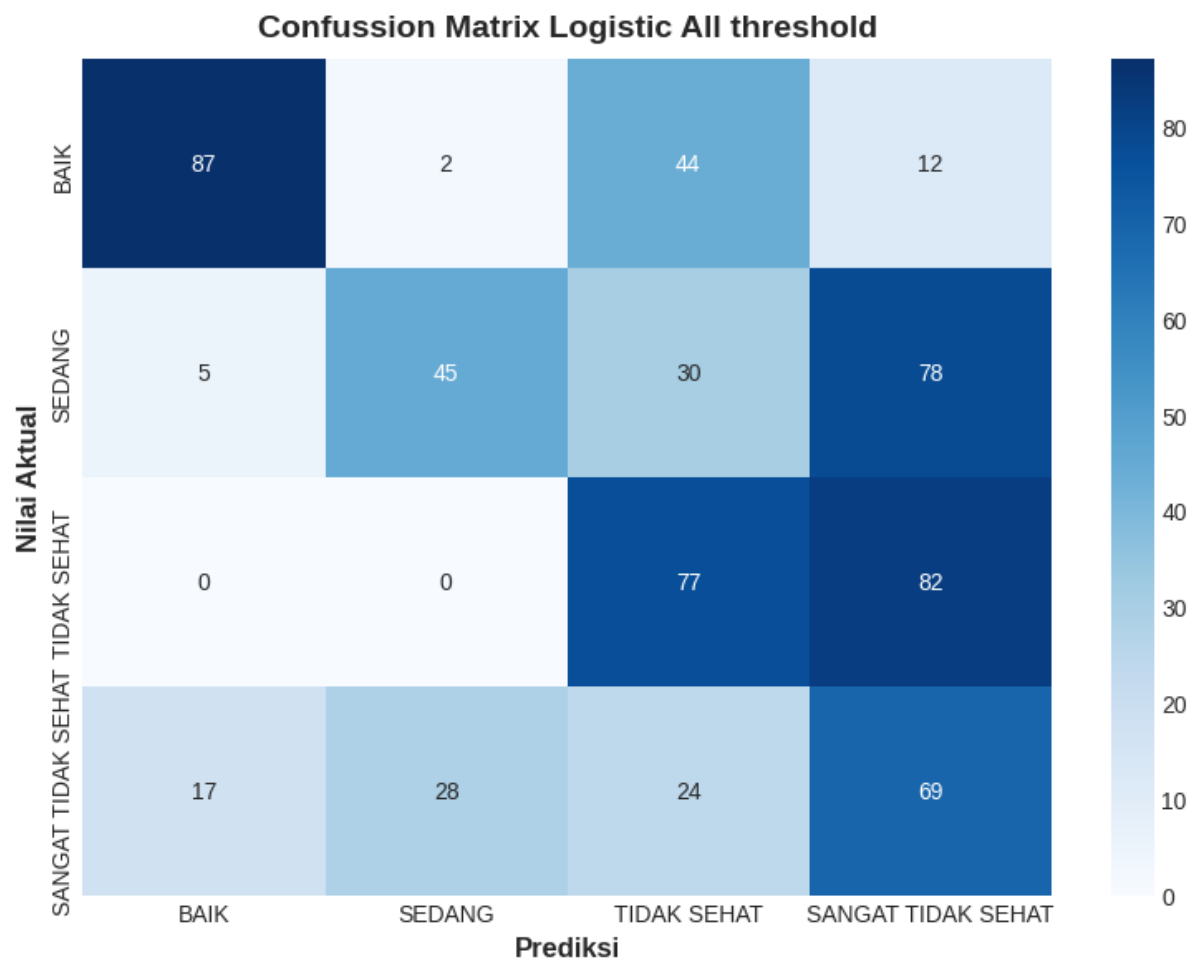
Laporan Evaluasi Model Ordinal Logistic Regression (AT) :

	precision	recall	f1-score	support
BAIK	0.80	0.60	0.69	145
SANGAT TIDAK SEHAT	0.60	0.28	0.39	158
SEDANG	0.44	0.48	0.46	159
TIDAK SEHAT	0.29	0.50	0.36	138

accuracy			0.46	600
macro avg	0.53	0.47	0.47	600
weighted avg	0.53	0.46	0.47	600

Ordinal Logistic Regression Model (AT):
Accuracy: 0.4633333333333333

Confussion Matrix



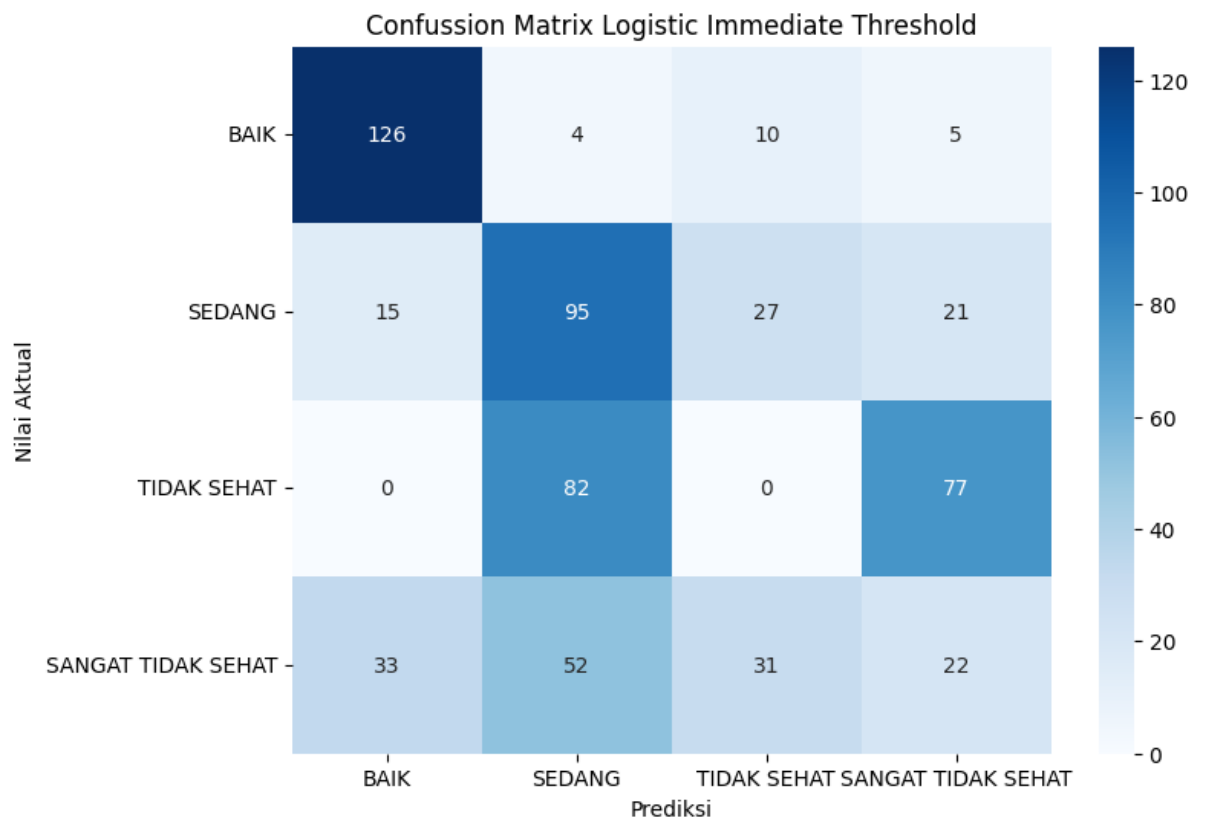
Nilai RMSE dan MSE

```
Ordinal Logistic Regression (AT)MSE: 0.925
Ordinal Logistic Regression (AT)RMSE: 0.9617692030835673
Ordinal Logistic Regression (AT)MAE: 0.6583333333333333
```

- Model Ordinal Logistik Regresi (Immediate-Threshold variant)

Laporan Evaluasi Model Ordinal Logistic Regression (IT):					
	precision	recall	f1-score	support	
BAIK	0.72	0.87	0.79	145	
SANGAT TIDAK SEHAT	0.41	0.60	0.49	158	
SEDANG	0.00	0.00	0.00	159	
TIDAK SEHAT	0.18	0.16	0.17	138	
accuracy			0.41	600	
macro avg	0.33	0.41	0.36	600	
weighted avg	0.32	0.41	0.36	600	
Ordinal Logistic Regression Model (IT):					
Accuracy: 0.405					

Confussion Matrix



Nilai RMSE dan MSE

```
Ordinal Logistic Regression (IT)MSE: 1.5833333333333333
Ordinal Logistic Regression (IT)RMSE: 1.2583057392117916
Ordinal Logistic Regression (IT)MAE: 0.9033333333333333
```

H. Uji Hipotesis

```
Fitur 'pm10':  
  t-value: 1.1275  
  p-value: 0.2597  
  Tidak cukup bukti untuk menolak hipotesis nol  
Fitur 'pm25':  
  t-value: -2.3697  
  p-value: 0.0179  
  Hipotesis nol ditolak: Koefisien signifikan  
Fitur 'so2':  
  t-value: 8.4354  
  p-value: 0.0000  
  Hipotesis nol ditolak: Koefisien signifikan  
Fitur 'co':  
  t-value: 8.5733  
  p-value: 0.0000  
  Hipotesis nol ditolak: Koefisien signifikan  
Fitur 'o3':  
  t-value: 1.8482  
  p-value: 0.0648  
  Tidak cukup bukti untuk menolak hipotesis nol  
Fitur 'no2':  
  t-value: 3.2729  
  p-value: 0.0011  
  Hipotesis nol ditolak: Koefisien signifikan
```

I. Hasil model ordinal logistik regresi

OLS Regression Results						
Dep. Variable:	kategori_encoded	R-squared (uncentered):	0.116			
Model:	OLS	Adj. R-squared (uncentered):	0.112			
Method:	Least Squares	F-statistic:	30.46			
Date:	Mon, 10 Jun 2024	Prob (F-statistic):	1.73e-34			
Time:	01:19:05	Log-Likelihood:	-2772.5			
No. Observations:	1400	AIC:	5557.			
Df Residuals:	1394	BIC:	5588.			
Df Model:	6					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
x1	0.0698	0.062	1.127	0.260	-0.052	0.191
x2	-0.1628	0.069	-2.370	0.018	-0.298	-0.028
x3	0.4174	0.049	8.435	0.000	0.320	0.514
x4	0.4925	0.057	8.573	0.000	0.380	0.605
x5	0.1018	0.055	1.848	0.065	-0.006	0.210
x6	0.1848	0.056	3.273	0.001	0.074	0.296
Omnibus:	57.711	Durbin-Watson:	0.572			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	71.329			
Skew:	0.436	Prob(JB):	3.24e-16			
Kurtosis:	3.679	Cond. No.	2.79			

- R-squared mengukur seberapa baik model cocok dengan data. Nilainya adalah proporsi variasi dalam variabel target yang dapat dijelaskan oleh model. Nilai R-squared yang diberikan adalah sebesar 0.325, yang berarti sekitar 32.5% variasi dalam kategori polusi udara dapat dijelaskan oleh model.
- F-statistic digunakan untuk menguji keseluruhan signifikansi model. Nilainya adalah sebesar 95.81 dengan p-value yang sangat rendah (2.48e-98), menunjukkan bahwa model secara keseluruhan signifikan.
- P-value mengukur signifikansi statistik dari koefisien. Nilai p-value yang rendah (bias < 0.05) menunjukkan bahwa koefisien tersebut secara signifikan berbeda dari nol.
- Dalam fitur di atas, fitur pm25, so2, co, o3, dan no2 memiliki p-value yang rendah, menunjukkan bahwa mereka memiliki pengaruh yang signifikan terhadap kategori polusi udara.

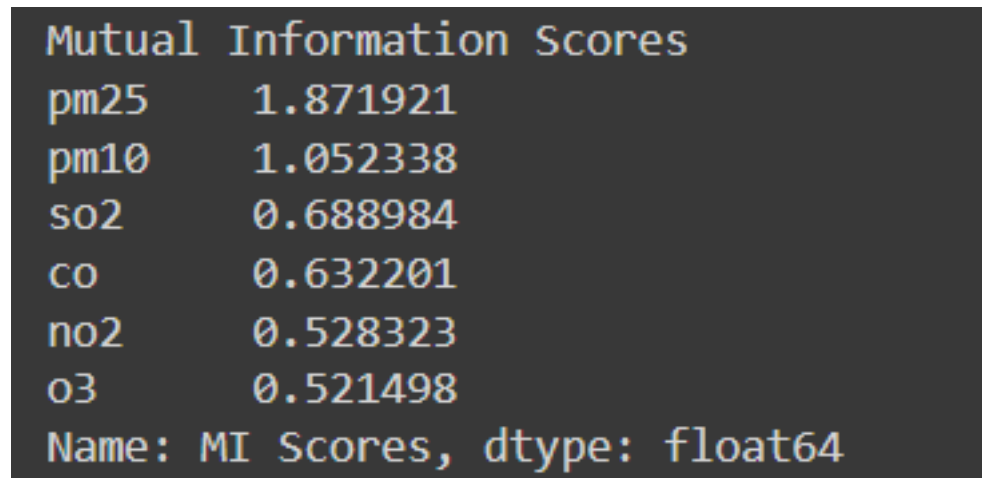
J. Tes Anova

	sum_sq	df	F	PR(>F)
pm10	3.628344	1.0	4.280205	3.868737e-02
pm25	23.473815	1.0	27.691072	1.576175e-07
so2	312.205488	1.0	368.295676	1.858302e-75
co	317.149062	1.0	374.127403	1.580327e-76
o3	21.192555	1.0	24.999965	6.234824e-07
no2	44.345920	1.0	52.313015	6.717017e-13
Residual	1689.472827	1993.0	NaN	NaN

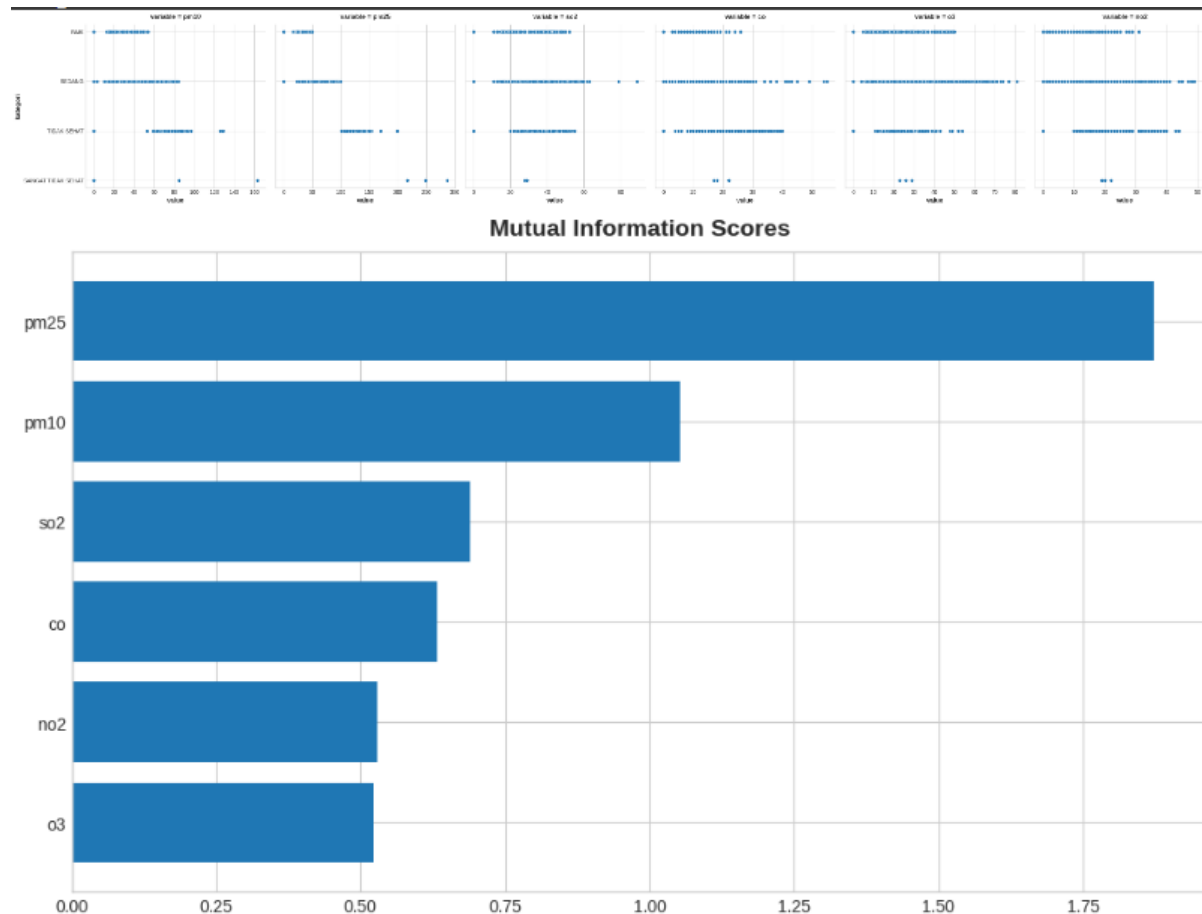
semua variabel prediktor (pm10, pm25, so2, CO, O3, NO2) memiliki nilai p yang sangat rendah (kurang dari 0.05), yang menunjukkan bahwa mereka memiliki pengaruh signifikan terhadap variabel dependen.

K. Mutual Information

- Mutual Information Score



- Mutual Information Score Visualisasi



Berdasarkan gambar tersebut, variabel pm25 menunjukkan ketergantungan tertinggi dengan variabel target, sedangkan variabel o3 memiliki ketergantungan terendah.

BAB IV

PENUTUP

Penelitian ini membandingkan kinerja dua model klasifikasi, yaitu Naive Bayes dan Ordinal Logistic Regression, dalam memprediksi kualitas udara di Jakarta berdasarkan data Indeks Standar Pencemar Udara (ISPU). Berdasarkan hasil analisis dan evaluasi model, dapat disimpulkan beberapa poin penting sebagai berikut:

1. Efektivitas Model Naive Bayes:

- Model Naive Bayes terbukti efektif dalam memprediksi kategori kualitas udara. Keuntungan utama dari model ini adalah kesederhanaannya dan kemampuannya dalam menangani dataset yang besar dengan komputasi yang relatif cepat.
- Model ini memberikan hasil prediksi yang cukup akurat dengan mempertimbangkan variabel-variabel pencemar udara seperti PM10, PM2.5, SO2, CO, O3, dan NO2.

2. Efektivitas Model Ordinal Logistic Regression:

- Model Ordinal Logistic Regression memberikan hasil yang cukup baik dalam memprediksi kualitas udara dengan memperhitungkan urutan kategori kualitas udara.
- Model ini lebih kompleks dibandingkan Naive Bayes, namun memiliki keunggulan dalam menangani variabel dependen yang bersifat ordinal.

3. Perbandingan Kinerja Model:

- Kedua model memiliki kelebihan dan kekurangan masing-masing. Naive Bayes unggul dalam kecepatan dan kesederhanaan, sementara Ordinal Logistic Regression lebih baik dalam menangani data ordinal dengan urutan kategori yang jelas.
- Evaluasi model menunjukkan bahwa Ordinal Logistic Regression sedikit lebih unggul dalam hal akurasi prediksi dibandingkan Naive Bayes, meskipun perbedaannya tidak terlalu signifikan.

4. Implikasi dan Rekomendasi:

- Hasil penelitian ini dapat membantu pemerintah dan instansi terkait dalam memantau dan memprediksi kualitas udara di Jakarta secara lebih efektif, sehingga dapat mengambil tindakan pencegahan yang tepat waktu untuk mengurangi dampak negatif pencemaran udara terhadap kesehatan masyarakat.
- Untuk penelitian selanjutnya, disarankan untuk mencoba model klasifikasi lain dan menggunakan dataset yang lebih besar serta mempertimbangkan faktor-faktor lain seperti data meteorologi untuk meningkatkan akurasi prediksi.

DAFTAR PUSTAKA

Karo, A. A. H., Azar, D., & Wibowo, Y. F. A. (2022). Klasifikasi Tingkat Kualitas Udara DKI Jakarta Menggunakan Algoritma Naive Bayes. *e-Proceeding of Engineering*, 9(3), 1962.

Zhang, Z., et al. (2012). "Comparison of Naive Bayes and SVM classifiers in quality air prediction." *Environmental Monitoring and Assessment*.

Liu, J., et al. (2019). "Ordinal logistic regression for air quality prediction in urban areas." *Atmospheric Environment*.