**PEKAN ANALISIS STATISTIK (PAS) JAMBORE STATISTIKA XIII 2024**

Optimasi Model Prediktif Kualitas Udara dan Indentifikasi Zat Pencemar di Jakarta: Integrasi *Machine Learning* dan Analisis Statistik yang Komprehensif.



Disusun oleh:

Ishak Bintang Dikaputra NIM A11.2021.13763 Angkatan 2021

Noval Ariyanto NIM A11.2021.13789 Angkatan 2021

**UNIVERSITAS DIAN NUSWANTORO**

**SEMARANG**

**2024**

# **Latar Belakang Masalah**

Tantangan yang berkaitan dengan polusi udara di suatu wilayah dapat memberikan dampak negatif terhadap kesehatan. Hal ini memerlukan perhatian serius, terutama dalam upaya melindungi kesehatan masyarakat. Polusi udara dapat berasal dari berbagai sumber, seperti emisi kendaraan bermotor, industri, dan aktivitas perkotaan lainnya. Apabila partikel-partikel berbahaya serta gas pencemar lainnya melebihi standar yang ditetapkan oleh badan kesehatan, maka hal tersebut akan merugikan kesehatan penduduk. Ini menjadi suatu permasalahan yang memiliki dampak secara signifikan pada berbagai aspek kehidupan, terutama pada kawasan perkotaan seperti Provinsi DKI Jakarta. Peningkatan emisi polutan dari kendaraan bermotor, industri, dan faktor-faktor lainnya dapat menyebabkan kualitas udara yang tidak sehat, dengan potensi dampak yang merugikan. Pemerintah berusaha untuk mengatasi masalah ini melalui penggunaan Indeks Kualitas Udara (AQI) untuk melakukan pemantauan dan memberitahu masyarakat tentang tingkat polusi udara. AQI mengukur konsentrasi karbon monoksida (CO), sulfur dioksida (SO2), nitrogen dioksida (NO2), ozon permukaan (O3), dan partikel debu (PM2,5 dan PM10). Data AQI diperbarui secara real time, dan menempatkan tingkat polusi udara Provinsi DKI Jakarta di kategori sedang. Pada kategori ini kualitas udara masih dapat diterima. Namun, di terdapat beberapa polutan yang kemungkinan dapat memunculkan masalah kesehatan moderat bagi sejumlah orang yang sangat sensitif terhadap polusi udara.

Meskipun sudah terdapat pemantauan kualitas udara, hal itu tidak cukup untuk mengatasi penyebab utama polusi udara. Masih diperlukan suatu pendekatan yang lebih proaktif untuk mengatasi akar penyebab dari polusi udara. Oleh karena itu, penting untuk memiliki sistem prediksi kualitas udara yang dapat memberikan informasi akurat. Pengembangan sistem untuk memprediksi kualitas udara menjadi suatu langkah yang strategis dalam rangka mengambil tindakan preventif dan mengelola dampak buruk polusi udara. Dalam konteks ini, penelitian ini mencoba untuk membangun kerangka evaluasi menggunakan algoritma Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), dan Rocchio Classifier untuk memperkirakan tingkat AQI dengan pendekatan pembelajaran mesin. Algoritma ini diharapkan dapat memberikan respons yang lebih rinci dan komprehensif terhadap prediksi kualitas udara. Penting untuk memahami bahwa prediksi ini dapat memberikan wawasan berharga bagi pemerintah, lembaga kesehatan, dan masyarakat umum untuk mengambil tindakan preventif dan mendukung upaya mitigasi dampak polusi udara.

# **Tujuan Penelitian**

Berdasarkan latar belakang masalah diatas, maka tujuan dari penelitian yang dilakukan adalah sebagai berikut.

1. Membangun model prediktif menggunakan metode statistik dan pembelajaran mesing berupa teknik *Linear Discriminant Analysis* (LDA), *Support Vector Machine* (SVM), dan *Rocchio Classifier* untuk melakukan prediksi kualitas udara di Provinsi DKI Jakarta dengan informasi yang akurat.
2. Mengidentifikasi faktor-faktor yang paling berpengaruh dalam prediksi kualitas udara, dengan demikian memberikan wawasan tentang aspek-aspek yang dapat diperhatikan dalam upaya pengendalian polusi udara di wilayah Provinsi DKI Jakarta.

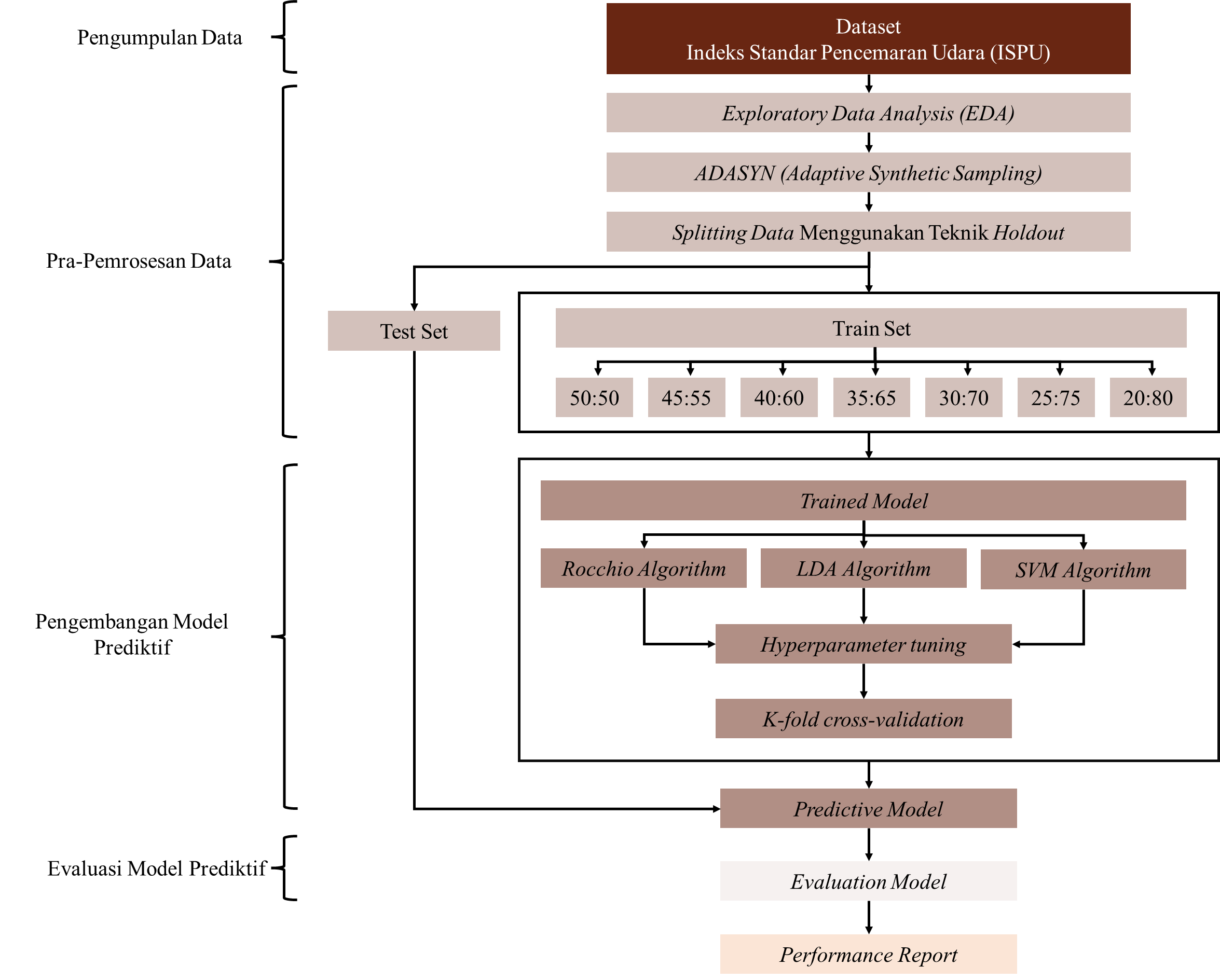
# **Metodologi Penelitian**

Penelitian ini menggunakan kerangka kerja penelitian yang terdiri dari empat langkah utama untuk menganalisis kualitas udara di wilayah Provinsi DKI Jakarta. Pendekatan metodologis ini sengaja dibuat untuk memahami komponen-komponen yang mempengaruhi kualitas udara dan membuat model prediksi yang akurat. Setiap tahap metodologi memiliki tujuan khusus yang menambah keberhasilan penelitian secara keseluruhan. Secara detail kerangka kerja penelitian ini disajikan pada Gambar 1 berikut.

## **Sumber Data dan Variabel Penelitian**

Data untuk penelitian ini berasal dari Jakarta Open Data, yang dapat diakses melalui situs web resmi [https://data.jakarta.go.id](https://data.jakarta.go.id/). Dataset yang digunakan adalah Indeks Standar Pencemaran Udara (ISPU), yang mencakup data dari 5 stasiun pemantau kualitas udara (SPKU) di Provinsi DKI Jakarta sepanjang tahun 2021. Dataset ini mengandung 1.517 pengamatan kualitas udara yang tercatat. Variabel yang digunakan terdiri dari variabel target berupa kategori kualitas udara yaitu merupakan hasil perhitungan Indeks Standar Pencemaran Udara (ISPU). Kategori ini memberikan gambaran tingkat polusi udara dengan memberikan label seperti "baik" kategori (0), "sedang" kategori (1), atau "tidak sehat" kategori (2), sehingga dapat memberikan informasi yang mudah dimengerti kepada masyarakat umum. Selain itu terdapat variabel atribut yang terdiri dari enam zat polutan yang merupakan zat yang diduga mempengaruhi kualitas udara di DKI Jakarta, enam zat tersebut antara lain.

1. **PM10 (Partikulat 10):** Merupakan partikulat matter dengan diameter kurang dari 10 mikrometer. PM10 dapat mencakup debu, asap, dan partikel-partikel kecil lainnya. Tingginya konsentrasi PM10 dapat berdampak negatif pada kesehatan manusia, terutama sistem pernapasan.
2. **PM2.5 (Partikulat 2,5):** Partikulat matter dengan diameter kurang dari 2,5 mikrometer. PM2.5 lebih kecil dan dapat lebih dalam masuk ke paru-paru manusia, menyebabkan risiko kesehatan yang lebih besar. Pemantauan PM2.5 penting untuk mengevaluasi dampak partikel halus terhadap udara.
3. **SO2 (Sulfur Dioksida):** SO2 dapat berasal dari aktivitas industri dan pembakaran bahan bakar fosil. Peningkatan SO2 dapat menyebabkan iritasi pada saluran pernapasan dan masalah kesehatan lainnya.
4. **CO (Carbon Monoksida):** CO adalah gas beracun yang dapat diproduksi oleh kendaraan bermotor dan pembakaran bahan bakar. Peningkatan konsentrasi CO dapat membahayakan kesehatan manusia dan hewan.
5. **O3 (Ozon):** Ozon di lapisan atas atmosfera melindungi Bumi dari sinar UV, tetapi ozon di permukaan bumi dapat menyebabkan masalah kesehatan. Tingkat ozon yang tinggi dapat merugikan sistem pernapasan.
6. **NO2 (Nitrogen Dioksida):** NO2 dapat berasal dari emisi kendaraan dan industri. Tingginya konsentrasi NO2 dapat menyebabkan iritasi pada saluran pernapasan dan masalah kesehatan lainnya.



**Gambar 1.** Kerangka Penelitian Model Prediksi Kualitas Udara

## **Pra-Pemrosesan Data**

Langkah-langkah pada Gambar 1 adalah tahap kritis dalam persiapan data sebelum penerapan model klasifikasi. Proses pra-pemrosesan data melibatkan beberapa teknik penting untuk memastikan kualitas dan kehandalan data yang digunakan dalam model klasifikasi. Berikut adalah rincian langkah-langkah pra-pemrosesan data yang dijelaskan:

## ***Exploratory Data Analysis (EDA)***

Proses ini melibatkan analisis eksploratif data untuk memahami distribusi variabel, deteksi outlier, dan memahami hubungan antar variabel. EDA membantu mengidentifikasi pola atau tren dalam data yang dapat memengaruhi kinerja model.

## ***ADASYN (Adaptive Synthetic Sampling)***

ADASYN adalah metode oversampling yang digunakan untuk menangani ketidakseimbangan kelas dalam dataset. Dalam konteks ini, ketidakseimbangan kelas mungkin terjadi jika jumlah sampel dari kategori polusi udara "baik", "sedang", atau "tidak sehat" tidak seimbang. ADASYN menciptakan sampel sintetis untuk kelas minoritas, sehingga mendukung model untuk melatih dengan lebih baik pada kelas yang kurang representatif.

## ***Splitting Data* Menggunakan Teknik *Holdout***

Data dibagi menjadi set pelatihan (train set) dan set pengujian (test set) menggunakan teknik *holdout*, pada penelitian ini menggunakan tujuh kombinasi rasio data yaitu (50%:50%), (55%:45%), (60%:40%), (65%:35%), (70%:30%), (75%:25%), (80%:20%) sesuai yang disajikan pada Gambar 1. Set pelatihan digunakan untuk melatih model, sedangkan set pengujian digunakan untuk menguji kinerja model pada data yang belum pernah dilihat sebelumnya. Hal ini membantu memastikan bahwa model dapat menggeneralisasi dengan baik pada data baru.

## **Pengembangan Model Prediktif (*Model Development*)**

Dalam penelitian ini, tiga model klasifikasi digunakan, yaitu *Linear Discriminant Analysis* (LDA), *Support Vector Machine* (SVM), dan *Rocchio Classifier*. Proses pengembangan model melibatkan langkah-langkah berikut.

## **Hiperparameter Tuning**

Untuk setiap model (LDA, SVM, dan Rocchio Classifier), dilakukan penyetelan hiperparameter. Hiperparameter adalah parameter-model yang tidak diajarkan oleh model itu sendiri tetapi harus diatur sebelum pelatihan. Penyetelan ini dilakukan untuk memastikan bahwa model memiliki kinerja optimal.

## **Validasi dengan Teknik *Cross-Validation***

Penggunaan teknik *cross-validation* dengan 10-fold. Teknik ini membagi dataset menjadi 10 bagian (fold) dan menggunakan 9 bagian untuk pelatihan model, sementara 1 bagian digunakan untuk pengujian. Proses ini diulangi sebanyak 10 kali dengan fold yang berbeda untuk memastikan setiap bagian data digunakan sebagai set pengujian. Hal ini membantu mengevaluasi kinerja model secara lebih *robust*.

## **Evaluasi Model Prediktif (*Model Evaluation*)**

Dalam evaluasi model prediktif, dilibatkan pengukuran waktu komputasi, akurasi, presisi, recall, dan F1-Score. Waktu komputasi mencerminkan efisiensi model, sedangkan akurasi memberikan gambaran kemampuan model mengklasifikasikan data secara benar. Presisi dan recall menilai kemampuan model dalam memprediksi positif dengan benar, dan F1-Score memberikan gambaran keseluruhan kinerja model dengan mempertimbangkan presisi dan recall. Evaluasi ini bertujuan untuk memilih model yang optimal dalam mengatasi berbagai aspek, termasuk *false positives* dan *false negatives*.

# **Analisis dan Pembahasan**

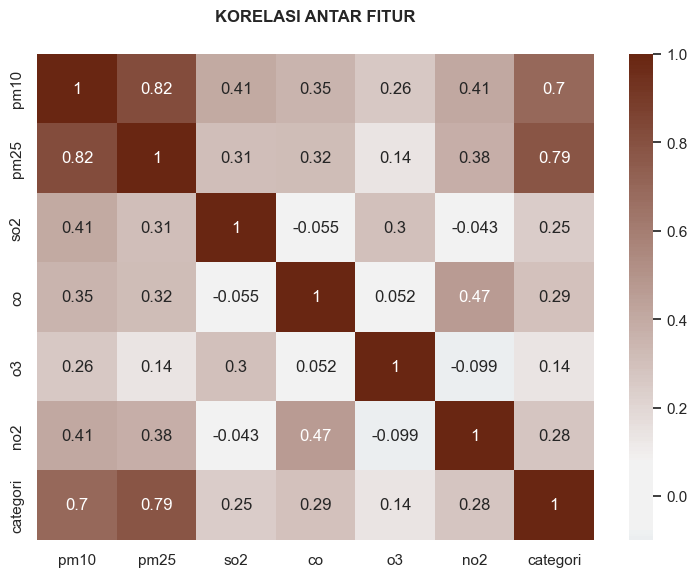
Dalam menggali lebih dalam pada penelitian ini, peneliti melakukan dua pendekatan analisis, yakni analisis statistika deskriptif dan analisis statistika inferensial. Kedua pendekatan ini memberikan wawasan yang berbeda untuk memahami data kualitas udara di Provinsi DKI Jakarta dan memilih model klasifikasi yang optimal.

# **Analisis Statistika Deskriptif**

Analisis statistika deskriptif menjadi langkah awal untuk merinci karakteristik dataset kualitas udara.

# **Hasil *Exploratory Data Analysis***

Dalam upaya mendapatkan pemahaman yang lebih mendalam tentang dataset Indeks Kualitas Udara, bagian ini membahas hasil dari Analisis Data Eksploratif (EDA). EDA menjadi langkah awal yang krusial dalam proses analisis data, memungkinkan peneliti untuk mengidentifikasi pola, hubungan, dan karakteristik utama dari variabel-variabel yang terlibat. Pada bagian ini diawali dengan analisis korelasi antar variabel yang merupakan Langkah awal dalam pemahaman struktur dan hubungan dalam sebuah dataset. Pada konteks kualitas udara di Jakarta, penelitia akan mengeksplore Correlogram (Gambar 2) yang memetakan hubungan antara beberapa parameter utama, seperti Particulate Matter (PM10 dan PM2.5), gas-gas polutan (SO2, CO, O3, NO2), dan kategori kualitas udara.

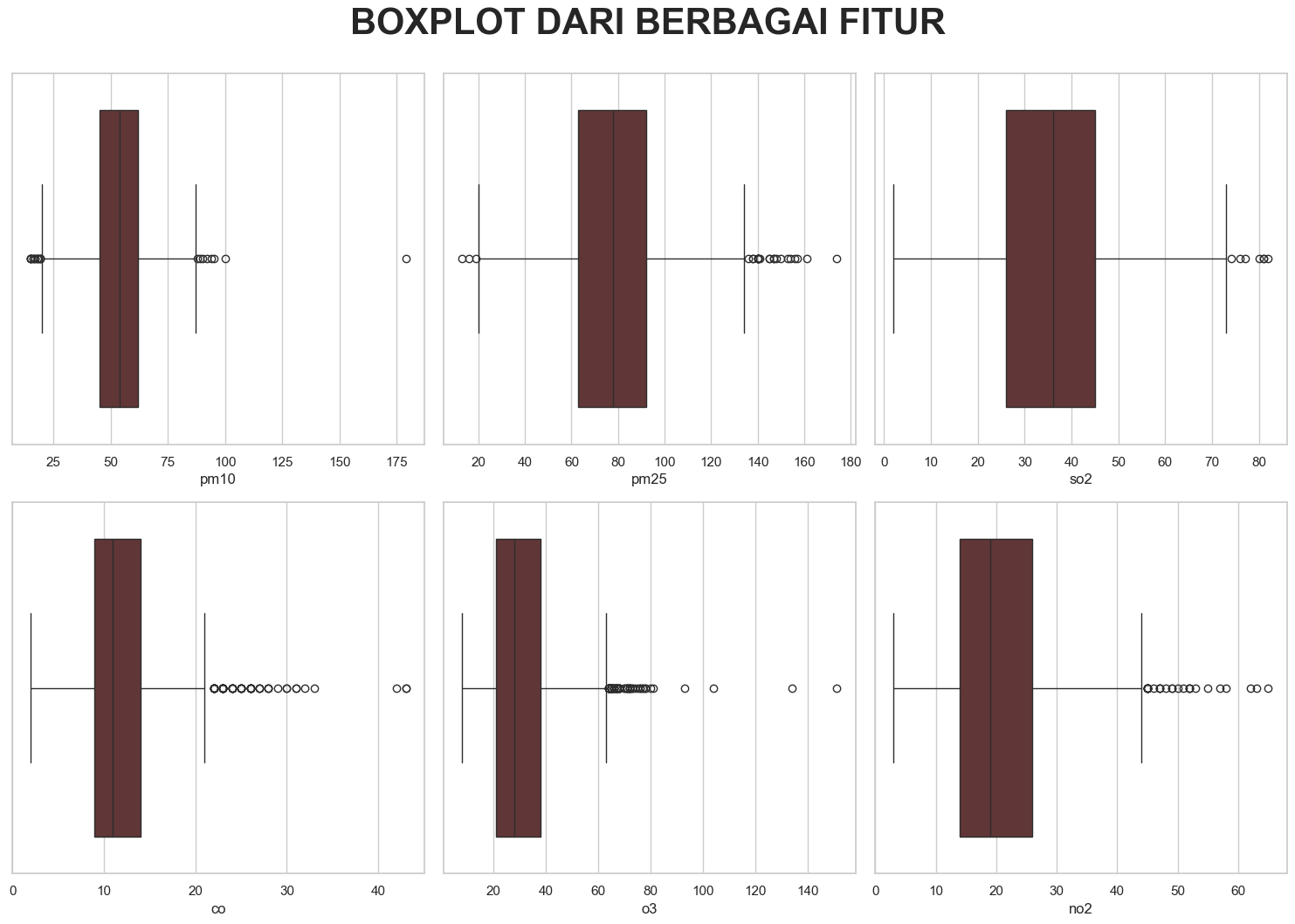


**Gambar 2.** Correlogram Hubungan Antar Variabel dalam Kualitas Udara Jakarta

Berdasarakan Correlogram Hubungan Antar Variabel dalam Kualitas Udara Jakarta pada Gambar 2 Berikut adalah beberapa poin yang dapat dianalisa:

1. **PM10 dan PM2.5.** Terdapat korelasi yang tinggi antara pm10 dan pm25 sebesar 0.819901. Hal ini menunjukkan bahwa kadar Particulate Matter (PM) dengan ukuran 10 mikrometer dan 2.5 mikrometer cenderung berCOrak serupa.
2. **PM10 dan kualitas udara.** Korelasi antara pm10 dan variabel target (categori) adalah 0.695886. Ini menunjukkan adanya keterkaitan yang cukup signifikan antara tingkat PM10 dan kategori kualitas udara.
3. **PM25 dan kualitas udara.** Korelasi antara pm25 dan variabel target (categori) adalah 0.789274. Ini menandakan adanya korelasi positif yang kuat antara tingkat PM2.5 dan kategori kualitas udara.
4. **SO2 dan CO,** Terdapat korelasi positif antara tingkat SO2 (sulfur dioksida) dan CO (carbon monoxide) sebesar 0.054524. Meskipun nilai ini rendah, namun menunjukkan adanya hubungan positif antara kedua gas tersebut.
5. **O3 dan NO2,** terdapat korelasi negatif antara tingkat O3 (ozon) dan NO2 (nitrogen dioksida) sebesar -0.098685. Ini menunjukkan bahwa kadar ozon cenderung berkurang ketika kadar nitrogen dioksida meningkat, dan sebaliknya.

Dari hasil analisis Correlogram sebelumnya, untuk memahami sejauh mana hubungan antar variabel dalam dataset kualitas udara di Jakarta. Selanjutnya, akan dilakukan analisa lebih jauh dengan menganalisis boxplot untuk melihat penyebaran dan pola distribusi data untuk masing-masing parameter kualitas udara.

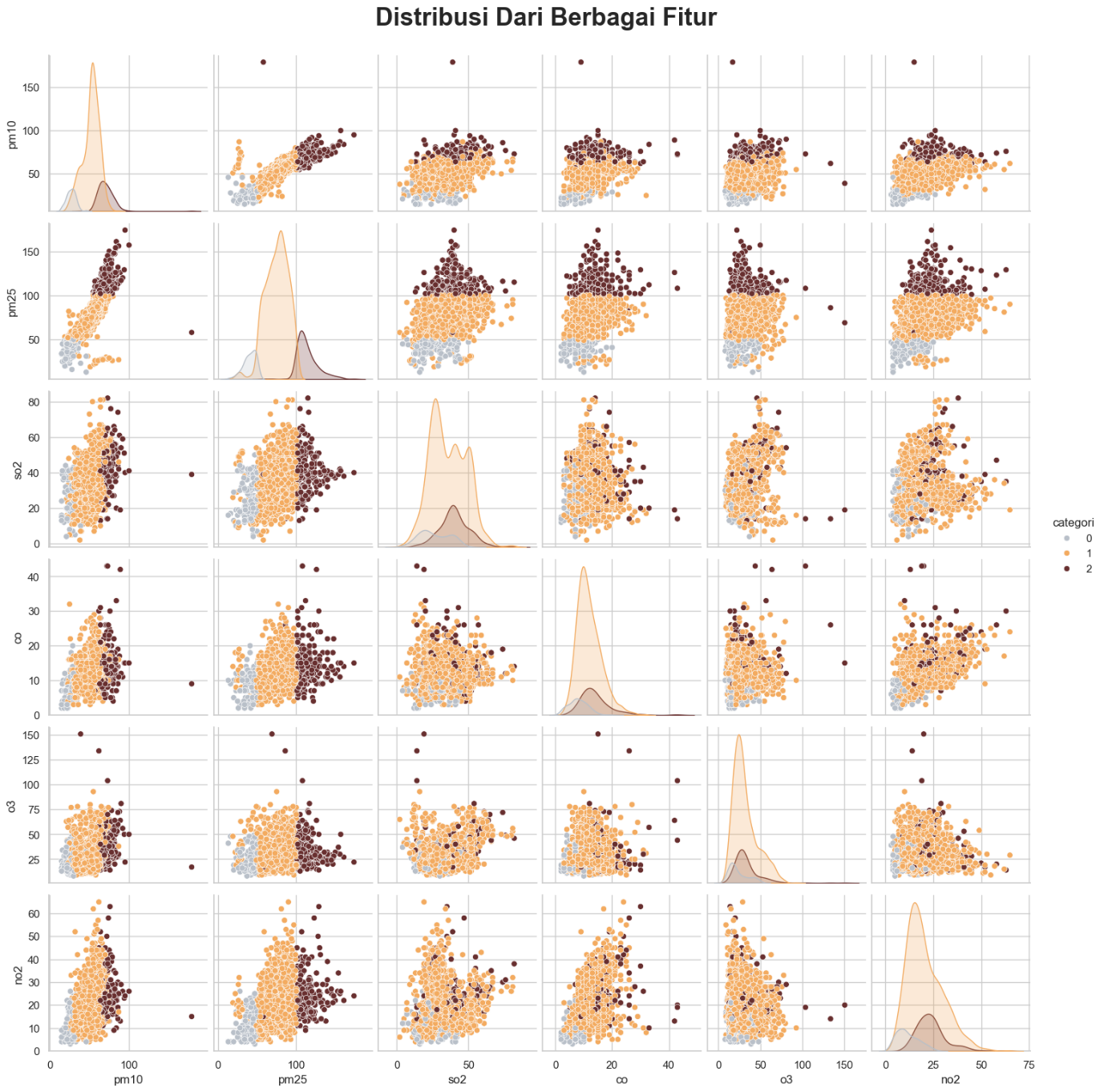


**Gambar 3.** Boxplot Antar Variabel dalam Kualitas Udara Jakarta

Pada boxplot, dapat melihat distribusi nilai dan kecenderungan outliers pada setiap parameter. Terdapat beberapa pola menarik yang dapat diidentifikasi dari boxplot ini, yang memberikan konteks lebih lanjut tentang karakteristik dataset:

1. **PM10 dan PM25:** Meskipun korelasi antara PM10 dan PM25 tinggi, boxplot mengungkapkan variasi nilai yang signifikan, dengan beberapa nilai yang melampaui nilai kuartil ketiga (Q3). Ini menunjukkan adanya situasi di mana tingkat partikulat PM10 dan PM25 dapat meningkat secara signifikan.
2. **SO2 dan CO:** Penyebaran outliers yang cenderung melebihi nilai Q3 pada SO2 dan CO menandakan adanya kondisi tertentu di mana tingkat SO2 dan CO dapat mencapai nilai yang lebih tinggi dari nilai median. Hal ini dapat berhubungan dengan faktor-faktor seperti polusi industri atau sumber emisi lokal.
3. **O3 dan NO2:** Outliers yang menonjol pada O3 dan NO2 menunjukkan variasi tinggi dalam tingkat ozon dan nitrogen dioksida pada beberapa periode tertentu. Ini bisa terkait dengan perubahan musiman atau faktor-faktor lingkungan lainnya.

Boxplot memberikan informasi tentang sebaran dan kecenderungan nilai-nilai pada setiap fitur kualitas udara. Dari analisis boxplot, dapat dilihat beberapa temuan signifikan seperti penyebaran nilai yang melampaui kuartil atas (Q3) pada fitur tertentu, menandakan adanya outliers atau nilai ekstrem. Selanjutnya analisis pairplot, untuk menyelidiki lebih lanjut tentang keterkaitan dan pola distribusi antar fitur tersebut. Pairplot memberikan visualisasi hubungan dua per dua antar fitur, dan dari sini dapat diamati apakah nilai ekstrem pada boxplot juga tercermin dalam hubungan antar fitur pada pairplot.



**Gambar 4.** Analisis Pairplot Pola Distribusi Kualitas Udara dan Faktor-faktor Pengaruh

Dengan melihat pairplot kualitas udara di wilayah tersebut seperti pada Gambar 3, peneliti dapat mengamati beberapa insight penting terkait pola distribusi dan faktor-faktor pengaruh. Beberapa temuan utama termasuk:

1. **Kategori Kualitas Udara**

Distribusi data menunjukkan bahwa kualitas udara umumnya berada pada kategori sedang. Kategori sedang memiliki frekuensi yang lebih tinggi dibandingkan kategori baik dan tidak sehat.

1. **Pola Distribusi Fitur SO2, NO2, O3, dan CO**

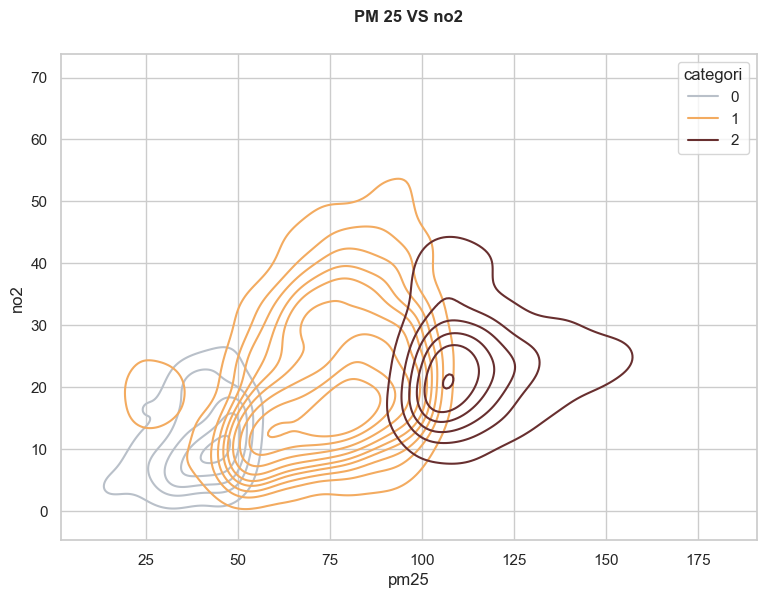
Pada fitur SO2, NO2, O3, dan CO, terdapat beberapa data yang memiliki nilai yang sama dan mengerucut pada satu posisi. Faktor-faktor yang dapat menyebabkan pola ini melibatkan pengaruh cuaca dan iklim serta aktivitas manusia. COntohnya, cuaca yang cerah dapat meningkatkan konsentrasi ozon (O3), dan aktivitas manusia seperti pembakaran bahan bakar fosil dapat memengaruhi konsentrasi CO dan NO2.

1. **Variasi Kualitas Udara dari Waktu ke Waktu**

Sebaran data untuk kategori baik, sedang, dan tidak sehat menunjukkan tingkatan yang berbeda, menandakan bahwa kualitas udara di wilayah tersebut dapat bervariasi dari waktu ke waktu. Faktor-faktor lingkungan dan manusia dapat menyebabkan fluktuasi dalam konsentrasi polutan udara.

1. **Perlunya Pemantauan Terus-menerus**

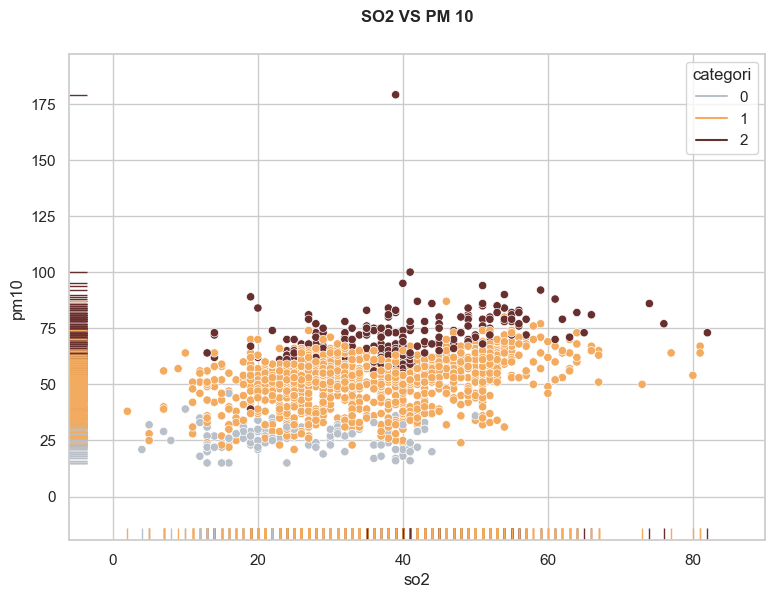
Dengan adanya variasi dalam kualitas udara, terutama yang berkaitan dengan faktor-faktor eksternal seperti cuaca dan aktivitas manusia, penting untuk terus memantau kualitas udara secara terus-menerus. Hal ini bertujuan untuk memastikan bahwa kualitas udara tetap berada dalam kondisi yang aman bagi kesehatan masyarakat.



**Gambar 5.** Statistika Deskriptif Berdasarkan PM2.5 dan NO2

Dari hasil analisis PM2.5 vs NO2 peneliti dapat mengidentifikasi beberapa temuan yang mungkin hubungan antar PM2.5 vs NO2:

1. Pada kategori baik, penyebaran data konsentrasi NO2 dan PM2.5 cenderung terkonsentrasi pada nilai yang lebih rendah. Hal ini menunjukkan bahwa kualitas udara di kategori baik cenderung berada dalam kondisi yang baik.
2. Pada kategori sedang, penyebaran data konsentrasi NO2 dan PM2.5 cenderung terkonsentrasi pada nilai yang lebih tinggi dibandingkan kategori baik, tetapi masih berada dalam batas yang aman. Hal ini menunjukkan bahwa kualitas udara di kategori sedang cenderung berada dalam kondisi yang sedang.
3. Pada kategori tidak sehat, penyebaran data konsentrasi NO2 dan PM2.5 cenderung terkonsentrasi pada nilai yang lebih tinggi dibandingkan kategori sedang. Hal ini menunjukkan bahwa kualitas udara di kategori tidak sehat cenderung berada dalam kondisi yang tidak sehat.

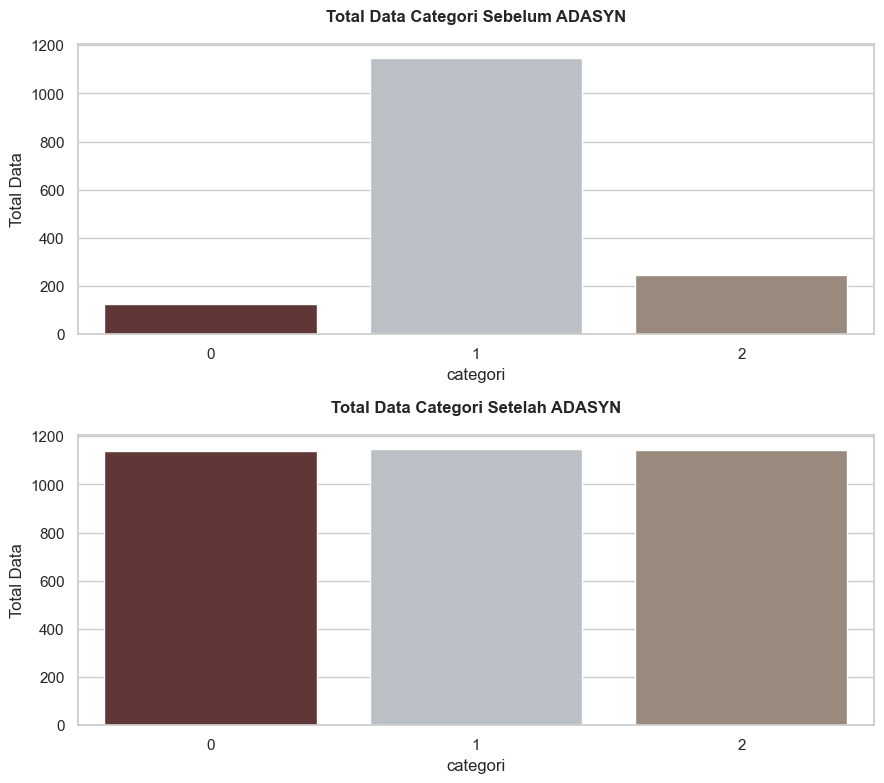


**Gambar 6.** Statistika Deskriptif Berdasarkan SO2 dan PM10

Penyebaran data antara kategori sedang dan tidak sehat memiliki kecenderungan yang sama, yaitu berbentuk pola kurva parabola. Hal ini menunjukkan bahwa terdapat hubungan positif antara konsentrasi SO2 dan konsentrasi PM10. Artinya, konsentrasi SO2 cenderung meningkat seiring dengan meningkatnya konsentrasi PM10.

# **Hasil Penanganan Data Tidak Seimbang menggunakan *ADAYSN***

Bab ini memfokuskan pada tahap handling imbalanced data menggunakan *Adaysn* adalah untuk meningkatkan akurasi dan sensitivitas dari model klasifikasi yang digunakan. Imbalanced data adalah data yang memiliki distribusi kelas yang tidak seimbang, di mana salah satu kelas memiliki jumlah data yang jauh lebih banyak daripada kelas lainnya. Hal ini dapat menyebabkan model klasifikasi menjadi bias terhadap kelas yang memiliki jumlah data lebih banyak, sehingga dapat menyebabkan penurunan akurasi dan sensitivitas model. *Adaysn* adalah metode resampling yang termasuk ke dalam kategori oversampling. Metode ini bekerja dengan cara mensintesis data baru dari kelas minoritas. Data sintetis ini dibuat dengan memperhatikan jarak antara data minoritas asli dengan data mayoritas. Hal ini dilakukan agar data sintetis yang dihasilkan dapat mewakili data minoritas asli dengan lebih baik.



**Gambar 7.** Hasil Deteksi dan Penanganan ketidakseimbangan data

Sebelum menerapkan ADASYN seperti yang ditunjukkan Gambar 2, dataset secara total menunjukkan ketidakseimbangan dengan 1517 observasi pada kelas kualitas udara baik sebanyak 125 (kelas 0),pada kelas kualitas udara sedang sebanyak 1147(kelas 1) dan hanya 245 observasi pada kelas kualitas udara tidak sehat (kelas 2) . Setelah menerapkan ADASYN pada dataset total, distribusi kelas kualitas udara baik (kelas 0) disamakan dengan kelas udara sedang (kelas 1) dan kelas kualitas udara buruk (kelas 2), yaitu masing-masing 1140 observasi. Hal ini bertujuan untuk mengatasi ketidakseimbangan data dan meningkatkan kinerja model pada kelas minoritas.

# **Statistika Inferensia**

Pada bab ini, fokus utama terletak pada statistika inferensi, yang melibatkan pengembangan, validasi, dan evaluasi model klasifikasi untuk prediksi kualitas udara.

Dalam konteks model klasifikasi, analisis statistika inferensial menjadi kunci untuk mengambil kesimpulan lebih luas dari sampel data. Melalui uji statistik, seperti uji ANOVA, peneliti dapat menilai apakah terdapat perbedaan yang signifikan antara akurasi model Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), dan Rocchio Classifier. Uji Tukey digunakan untuk mengidentifikasi pasangan model yang menunjukkan perbedaan signifikan dalam performa.

# **Model Development**

Dalam tahap Pembangunan model, peneliti akan membandingkan performa tiga algoritma klasifikasi, yaitu Rocchio Classifier, Linear Discriminant Analysis (LDA), dan Support Vector Machine (SVM), menggunakan teknik Holdout dengan variasi rasio split data (Training:Testing). Tahap ini diarahkan untuk menentukan algoritma yang mampu memberikan akurasi terbaik, menjadi langkah awal dalam membangun model klasifikasi yang dapat diandalkan.

**Tabel 1.** Performa Algoritma Klasifikasi dengan Variasi Rasio Split Data menggunakan Teknik *Holdout*

|  |  |  |  |
| --- | --- | --- | --- |
| Rasio Split Data (Training:Testing) | Rocchio Classifier | Linear Discriminant Analysis (LDA) | Support Vector Machine (SVM) |
| (50%:50%) | 0.8756 | 0.8857 | 0.9558 |
| (55%:45%) | 0.8790 | 0.8894 | 0.9559 |
| (60%:40%) | 0.8776 | 0.8921 | 0.9582 |
| (65%:35%) | 0.8844\* | 0.8911 | 0.9589\* |
| (70%:30%) | 0.8834 | 0.8990 | 0.9573 |
| (75%:25%) | 0.8787 | 0.8943 | 0.9565 |
| (80%:20%) | 0.8796 | 0.8990\* | 0.9554 |

Keterangan : \*) Rasio split dengan akurasi tertinggi.

Berdasarkan hasil eksperimen pada Tabel 1, akurasi masing-masing algoritma pada berbagai rasio split data. Dari hasil pengamatan, terlihat bahwa Linear Discriminant Analysis (LDA) menunjukkan konsistensi dalam memberikan performa akurasi, terutama ketika lebih banyak data digunakan untuk pelatihan. Rocchio Classifier menunjukkan variabilitas dalam performanya, namun pada beberapa rasio split data, mampu bersaing secara kompetitif. Paling menCOlok adalah Support Vector Machine (SVM) yang mencapai akurasi tertinggi pada beberapa skenario, terutama pada rasio 65%:35%. Berdasarkan akurasi tertinggi pada rasio split data (65%:35%), algoritma yang dipilih adalah Support Vector Machine (SVM). Hasil ini menggambarkan bahwa SVM memiliki kemampuan yang signifikan untuk menangani tugas klasifikasi pada konteks dataset yang digunakan menggunakan teknik Holdout.

**Tabel 2.** Perbandingan Akurasi Model Prediksi Kualitas Udara Menggunakan One-way ANOVA pada Rasio Split Data

|  |  |
| --- | --- |
| **F-statistic** | **p-value** |
| 997.52 | 3.65e-19\* |

Keterangan : \*) Signifikan pada =5%.

Untuk memperjelas perbedaan antara ketiga algoritma tersebut, hasil analisis One-way ANOVA (Tabel 2) menunjukkan bahwa terdapat perbedaan signifikan dalam akurasi prediksi antara model SVM, LDA, dan Rocchio Classifier pada berbagai rasio split data. F-statistik sebesar 997.52 dengan p-value sepenelitir 3.65e-19, menunjukkan bahwa setidaknya satu pasang mean kelompok memiliki perbedaan yang signifikan. Sebelumnya, analisis One-way ANOVA telah dilakukan untuk mengevaluasi apakah terdapat perbedaan signifikan dalam akurasi prediksi antara model-model SVM, LDA, dan Rocchio Classifier pada berbagai rasio split data. Hasilnya menunjukkan bahwa terdapat perbedaan yang signifikan antara setidaknya dua kelompok model, mengindikasikan adanya variasi performa di antara algoritma.

Berikutnya, analisis Multiple COmparison of Means - Tukey HSD dilakukan untuk melakukan perbandingan lebih rinci antar kelompok model tersebut. Hasilnya disajikan dalam tabel berikut:

**Tabel 3.** Perbandingan Rata-Rata Akurasi Model dalam Prediksi Kualitas Udara menggunakan Analisis Tukey HSD

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algoritma 1** | **Algoritma 2** | ***Mean diff*** | **p-value** | **lower** | **upper** | **Keputusan** |
| LDA | Rocchio | -0.0132 | 0.0\* | -0.0179 | -0.0085 | Signifikan Berbeda |
| LDA | SVM | 0.0639 | 0.0\* | 0.0592 | 0.0686 | Signifikan Berbeda |
| Rocchio | SVM | 0.0771 | 0.0\* | 0.0724 | 0.0818 | Signifikan Berbeda |

Keterangan : \*) Signifikan pada =5%.

Tabel 3 menggunakan Analisis Tukey HSD menyajikan hasil perbandingan antara model LDA, Rocchio, dan SVM. Hasilnya menunjukkan bahwa terdapat perbedaan signifikan dalam akurasi prediksi kualitas udara antara ketiga model tersebut. Secara khusus, rata-rata akurasi SVM lebih tinggi secara signifikan dibandingkan dengan LDA dan Rocchio. Demikian pula, terdapat perbedaan yang signifikan antara LDA dan Rocchio, di mana rata-rata akurasi LDA lebih unggul. Temuan ini memberikan informasi mengenai kinerja relatif antar model dan menyoroti keunggulan SVM sebagai model yang lebih optimal dalam konteks prediksi kualitas udara berdasarkan analisis Tukey HSD ini.

Berdasarkan eksperimen yang dilakukan diantaranya Holdout, uji One-way ANOVA, dan analisis Tukey HSD menunjukkan bahwa model SVM secara konsisten unggul dengan akurasi tertinggi dalam prediksi kualitas udara. Uji statistik mengonfirmasi perbedaan signifikan antara model, dengan SVM secara khusus lebih baik daripada LDA dan Rocchio. Hasil ini memberikan rekomendasi jelas untuk memilih SVM sebagai pilihan utama dalam upaya prediksi kualitas udara.

Langkah selanjutnya dalam pemilihan dan pengoptimalan model adalah melakukan hiperparameter tuning pada model SVM. Eksperimen ini fokus pada peningkatan performa model SVM dalam prakiraan kualitas udara di Jakarta. Proses kritis ini memungkinkan peneliti menyesuaikan konfigurasi model secara mendalam dengan menguji berbagai kombinasi nilai parameter. Tujuannya adalah menemukan kombinasi terbaik yang dapat mengoptimalkan kinerja model SVM. Dalam konfigurasi hiperparameter yang dipakai pada penelitian ini, terdapat beberapa hiperparameter kunci yang akan dioptimalkan selama proses pencarian menggunakan teknik optimasi Optuna. Adapun parameter yang digunakan adalah sebagai berikut:

1. **C:** Hiperparameter regulasi yang mengontrol trade-off antara penalti terhadap kesalahan klasifikasi dan kompleksitas model SVM. Rentang nilai yang dugunakan adalah 0.1 hingga 10.0.
2. **Kernel:** Hiperparameter yang menentukan jenis fungsi dasar untuk transformasi ruang fitur, kernel yang digunakan adalah linear, rbf, poly, dan sigmoid.
3. **Degree:** Hiperparameter yang digunakan untuk kernel polynomial, menentukan derajat polinomial. Rentang nilai yang digunakan pada penelitian ini adalah 2 hingga 5.
4. **Gamma:** Hiperparameter yang mengontrol sejauh mana pengaruh satu data point terhadap yang lain pada model. Rentang nilai: dari 1e-4 hingga 10.0 (logaritmik).

Setelah penyetelan selesai, parameter ideal diperoleh untuk membangun model SVM dengan performa terbaik, dan hasil eksperimen disajikan pada tabel berikut.

**Tabel 4.** Top 10 Konfigurasi Hiperparameter Terbaik untuk Model SVM pada Prakiraan Kualitas Udara Jakarta

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **C** | **Kernel** | **Degree** | **Gamma** | **Akurasi** |
| 8.42 | polynomial | 5 | 0.000421 | 99.33%\* |
| 8.66 | polynomial | 5 | 0.009893 | 99.11% |
| 6.29 | polynomial | 4 | 0.678306 | 99.11% |
| 7.35 | polynomial | 5 | 0.074372 | 99.11% |
| 5.81 | polynomial | 4 | 0.074372 | 99.11% |
| 4.41 | polynomial | 5 | 0.026177 | 99.11% |
| 6.44 | polynomial | 4 | 0.010016 | 99.11% |
| 7.62 | polynomial | 5 | 0.054299 | 99.11% |
| 6.96 | polynomial | 5 | 0.160122 | 99.11% |
| 8.22 | polynomial | 5 | 1.016131 | 99.11% |

Keterangan : \*) Konfigurasi Hiperparameter Terbaik dengan akurasi tertinggi.

Dari hasil eksperimen, terlihat bahwa performa terbaik pada model SVM untuk prakiraan kualitas udara di Jakarta diperoleh ketika menggunakan kernel polinomial (polynomial). Semua konfigurasi hiperparameter terbaik, yang memiliki akurasi tertinggi, memiliki kernel polinomial dengan derajat polinomial 5. Meskipun terdapat variasi dalam nilai C dan gamma, konsistensi performa terbaik tetap terjadi pada kernel polinomial. Kombinasi hiperparameter optimal untuk model ini melibatkan penggunaan kernel polinomial dengan derajat 5, nilai C sepenelitir 8.42, dan gamma yang sangat rendah, yaitu 0.00042 yang memiliki akurasi sebesar 99.33%.

# **Model Validation**

Langkah selanjutnya dalam pengembangan model adalah validasi silang (cross-validation) dengan 10 fold. Proses ini memberikan pemahaman yang lebih komprehensif mengenai kemampuan generalisasi model pada berbagai subset data validasi. Hasil model evaluasi dengan Teknik *cross-validation* menggunakan fold sebanyak 10 adalah sebagai berikut.

**Tabel 5.** Hasil Model Validasi menggunakan Teknik K-Fold.

|  |  |
| --- | --- |
| Fold | Akurasi |
| Fold 1\* | 0.999 |
| Fold 2 | 0.982 |
| Fold 3 | 0.994 |
| Fold 4 | 0.982 |
| Fold 5 | 0.994 |
| Fold 6 | 0.988 |
| Fold 7 | 0.994 |
| Fold 8 | 0.982 |
| Fold 9 | 0.982 |
| Fold 10 | 0.998 |

Note: \*) Fold Terpilih

Proses evaluasi model menggunakan teknik cross-validation dengan 10-fold menghasilkan variasi akurasi pada setiap lipatan (fold). Hasil menunjukkan bahwa Fold 1 memiliki tingkat akurasi tertinggi sebesar 99.9%, dan lipatan ini dipilih sebagai fold terbaik. Performa model juga menunjukkan konsistensi pada beberapa lipatan lain, menandakan kemampuan model dalam menggeneralisasi dengan baik pada data validasi.

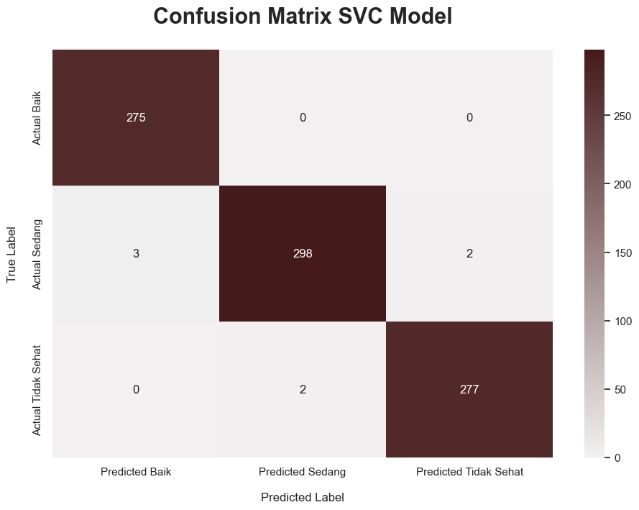
# **Model Evaluation**

Hasil dari model validasi kemudian diuji pada dataset pengujian untuk mendapatkan gambaran lebih lengkap mengenai performa model prediksi kualitas udara di Jakarta. Hasil model prediksi pada dataset pengujian adalah sebagai berikut.

**Tabel 6.** Hasil Evaluasi Model pada Dataset Pengujian

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Waktu Komputasi** | **Akurasi** | **Presisi** | **Recall** | **F1-Score** |
| 0.8088 detik | 99.18% | 99.18% | 99.18% | 99.18% |

Hasil Tabel 6, dalam uji pada dataset pengujian, waktu komputasi yang efisien sepenelitir 0.8088 detik dicapai, sementara metrik performa lainnya seperti akurasi, presisi, recall, dan F1-SCOre, mencapai tingkat yang sangat tinggi, yaitu sepenelitir 99.18%. Ini menunjukkan bahwa model SVM mampu memberikan prediksi kualitas udara dengan tingkat akurasi yang sangat tinggi dan waktu komputasi yang cepat, memberikan kesimpulan bahwa model ini dapat diandalkan untuk aplikasi prediksi kualitas udara di Jakarta.



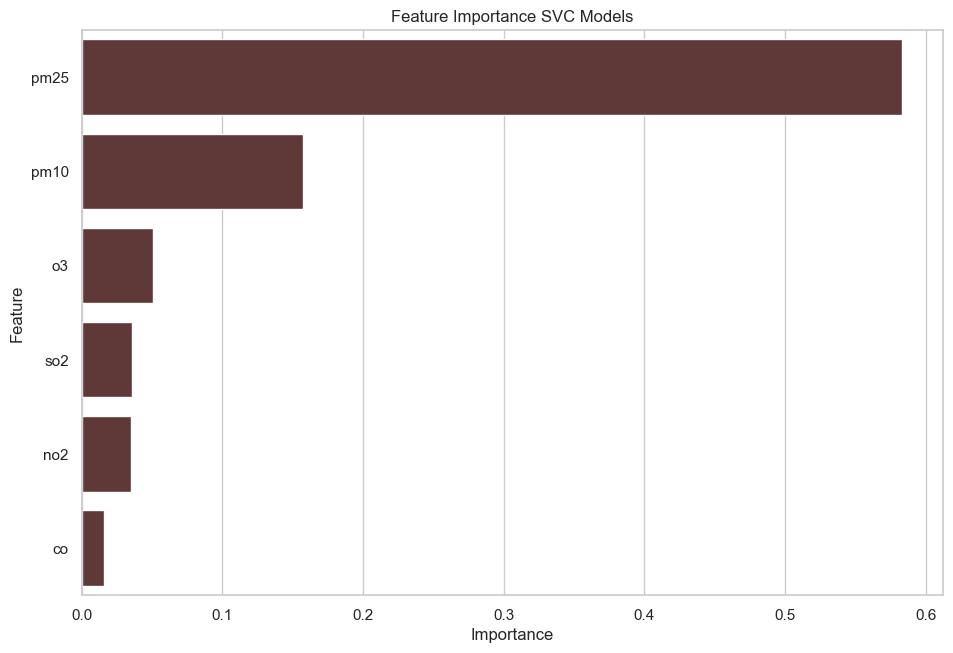
**Gambar 8.** Evaluasi Kinerja Model SVM: COnfusion Matrix Kualitas Udara

Confusion matrix atau matriks kebingungan adalah alat evaluasi kinerja model klasifikasi yang memberikan gambaran tentang seberapa baik model dapat mengklasifikasikan instans ke dalam kategori yang benar. Dalam konteks prediksi kualitas udara di Jakarta menggunakan model Support Vector Machine (SVM), COnfusion matrix memberikan wawasan tentang sejauh mana model dapat membedakan antara kualitas udara baik, sedang, dan tidak sehat.

Model Support Vector Machine (SVM) menunjukkan kinerja yang sangat baik dalam mengklasifikasikan kualitas udara baik dan tidak sehat. Dengan angka spesifik, terdapat 275 True Positives untuk kelas 1 (Kualitas Udara Baik) dan 277 True Positives untuk kelas 3 (Kualitas Udara Tidak Sehat), menandakan keberhasilan model dalam mengidentifikasi kedua kelas ini. Namun, pada kelas 2 (Kualitas Udara Sedang), meskipun terdapat 298 True Positives, terdapat 3 False Negatives (kasus yang seharusnya sedang tetapi diprediksi baik) dan 2 False Positives (kasus yang seharusnya sedang tetapi diprediksi tidak sehat). Ini menunjukkan bahwa meskipun model cenderung akurat dalam mengenali kualitas udara sedang, masih ada beberapa kasus yang memerlukan perhatian lebih lanjut. Secara keseluruhan, model SVM mencapai akurasi yang tinggi dengan total True Positives sebanyak 850 dari 855 sampel. Evaluasi lebih lanjut dan penyesuaian pada kelas tertentu dapat diterapkan untuk meningkatkan performa model dalam mengklasifikasikan kualitas udara.

# ***Feature Importance***

Analisis fitur atau feature importance memberikan informasi mendalam tentang variable atau faktor yang paling berpengaruh dalam model prediksi kualitas udara menggunakan Support Vector Machine (SVM). Fitur atau variabel-variabel ini membantu memahami aspek-aspek apa yang paling signifikan dalam memprediksi kualitas udara di Jakarta.



**Gambar 9.** Kontribusi Fitur terhadap Prediksi Kualitas Udara di Jakarta

Dari hasil analisis fitur atau feature importance pada model SVM seperti yang ditunjukkan pada Gambar 4, peneliti dapat mengidentifikasi lima fitur teratas yang paling signifikan dalam memprediksi kualitas udara di Jakarta. Berikut adalah penjelasan lebih lanjut mengenai masing-masing fitur:

1. PM2.5 (Particulate Matter 2.5):

Tingginya tingkat importance pada pm25 (0.583431) menunjukkan bahwa partikel dengan diameter kurang dari 2.5 mikrometer memainkan peran kunci dalam memprediksi kualitas udara. Peningkatan konsentrasi pm25 dapat memberikan informasi signifikan terkait dampak polusi udara terhadap kesehatan masyarakat.

1. PM10 (Particulate Matter 10):

Pentingnya pm10 (0.157176) sebagai fitur kedua menegaskan bahwa partikel dengan diameter kurang dari 10 mikrometer juga mempengaruhi penilaian kualitas udara. Pemantauan dan pengendalian pm10 menjadi kritis untuk menjaga kualitas udara di kota.

1. O3 (Ozon):

Walaupun tingkat importance O3 (0.050525) relatif lebih rendah, namun tetap memberikan kontribusi penting dalam prediksi. O3 yang tinggi dapat menjadi indikator polusi udara yang berpotensi berdampak negatif terhadap kesehatan manusia.

1. SO2 (Sulfur Dioxide):

Tingkat importance SO2 (0.035356) menunjukkan bahwa kandungan sulfur dioksida juga perlu dipertimbangkan. Peningkatan SO2 dapat menjadi sinyal polusi udara dari sumber-sumber tertentu, seperti industri.

1. NO2 (Nitrogen Dioxide):

Importance NO2 (0.034656) menyoroti peran nitrogen dioksida dalam memprediksi kualitas udara. Tingkat NO2 yang tinggi dapat menjadi indikasi polusi udara dari kendaraan bermotor dan sumber-sumber lainnya.

1. CO (Carbon Monoxide):

CO memiliki dampak yang relatif lebih rendah dibandingkan dengan fitur lainnya. Meskipun kontribusinya lebih rendah, tingkat CO dalam udara tetap relevan sebagai indikator polusi udara, terutama dari sumber-sumber pembakaran. Peningkatan pemantauan dan kontrol terhadap CO dapat membantu dalam memahami dan mengurangi dampak polusi udara yang berasal dari aktivitas pembakaran.

# **Kesimpulan Dan Rekomendasi**

# **Kesimpulan**

Penelitian ini berhasil mencapai tujuan-tujuan yang telah dirumuskan, menghasilkan pemahaman mendalam mengenai kualitas udara di Provinsi DKI Jakarta. Berikut adalah kesimpulan yang terkait dengan tujuan penelitian.

* 1. Model prediktif yang dikembangkan menggunakan teknik Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), dan Rocchio Classifier berhasil memberikan hasil yang sangat baik dalam memprediksi kualitas udara di wilayah Jakarta. Hasil analisis ANOVA menunjukkan adanya perbedaan signifikan antara setidaknya satu pasang grup model klasifikasi. Uji Tukey memperkuat temuan ini dengan menegaskan perbedaan signifikan antara model SVM dengan model LDA dan Rocchio Classifier pada tingkat kepercayaan 95%. SVM menonjol sebagai model dengan performa tertinggi, mencapai akurasi sekitar 99.18% dalam evaluasi dengan teknik cross-validation.
  2. Analisis fitur importance pada model SVM mengidentifikasi PM2.5 (Particulate Matter 2.5) dan PM10 (Particulate Matter 10) menjadi faktor paling signifikan dalam prediksi kualitas udara, diikuti oleh O3 (Ozon), SO2 (Sulfur Dioxide), NO2 (Nitrogen Dioxide), dan CO (Carbon Monoxide). Pemantauan dan pengendalian PM2.5 dan PM10 perlu ditingkatkan sebagai langkah utama dalam menjaga kualitas udara.

# **Rekomendasi**

Penelitian ini memberikan landasan untuk merumuskan rekomendasi yang dapat memperbaiki dan meningkatkan manajemen kualitas udara di Indonesia, khususnya di Provinsi DKI Jakarta. Dengan mengintegrasikan hasil analisis model prediktif dan temuan statistika inferensial, peneliti menyusun rekomendasi **PRIMA: Penyelarasan Responsif, Integratif, dan Mitigatif.**

1. **Penyelarasan Responsif**

* Mengintegrasikan model prediktif, terutama SVM, dalam sistem pemantauan kualitas udara secara real-time.
* Memperbarui dan meningkatkan jaringan stasiun pemantauan udara untuk mencakup lebih banyak wilayah, terutama di titik-titik yang cenderung memiliki tingkat polusi tinggi.

1. **Integratif**

* Meluncurkan kampanye edukasi publik yang lebih intensif mengenai dampak polusi udara pada kesehatan masyarakat.
* Menyoroti peran PM2.5 dan PM10 sebagai faktor utama yang memengaruhi kualitas udara, serta memberikan informasi tentang cara mengurangi paparan terhadap partikel-partikel berbahaya tersebut.

1. **Mitigatif**

* Menerapkan sistem peringatan dini berbasis model prediktif untuk memberikan informasi lebih cepat kepada pemerintah, lembaga kesehatan, dan masyarakat terkait potensi peningkatan polusi udara.
* Mengembangkan aplikasi seluler atau platform online untuk memberikan akses mudah dan cepat kepada masyarakat terkait kondisi kualitas udara.