

Pemetaan Zona Kualitas Udara di Provinsi Sumatera Utara Menggunakan Algoritma K-Means Berbasis Data CAMS

Callysa Elistia¹, Chelsea Febiola br Sebayang², Muhammad Rois Lukman Damanik^{3*}

^{1,2,3}Fakultas Matematika dan Ilmu Pengetahuan Alam, Ilmu Komputer, Universitas Negeri Medan, Medan, Indonesia

Email: ¹callysaelistia@gmail.com, ²chelsea.sby04@email.com, ^{3*}roislukman08@gmail.com

(*Email Corresponding Author: roislukman08@gmail.com)

Received: 24 Maret 2026 | Revision: 26 Juni 2026 | Accepted: 30 September 2026

Abstrak

Kualitas udara merupakan indikator penting dalam menilai kondisi lingkungan dan dampaknya terhadap kesehatan masyarakat. Penelitian ini bertujuan untuk mengklasifikasikan tingkat kualitas udara di Provinsi Sumatera Utara menggunakan pendekatan data mining berbasis algoritma K-Means Clustering. Data yang digunakan berasal dari layanan Copernicus Atmosphere Monitoring Service (CAMS) dalam format NetCDF dengan parameter utama PM_{2.5}, PM₁₀, CO, NO₂, dan SO₂ selama periode Januari 2026. Tahapan penelitian meliputi pra-pemrosesan data melalui rata-rata spasial, agregasi temporal, konversi satuan, normalisasi Min-Max, serta ekstraksi nilai maksimum harian. Penentuan jumlah kluster optimal dilakukan menggunakan metode Elbow yang menghasilkan tiga kluster sebagai representasi zona kualitas udara. Hasil penelitian menunjukkan bahwa wilayah penelitian terbagi menjadi tiga kategori, yaitu zona rendah, sedang, dan tinggi dengan karakteristik polutan yang berbeda. Zona tinggi didominasi oleh konsentrasi gas buang di wilayah pesisir timur, zona sedang menunjukkan tingginya partikulat di wilayah dataran tinggi, sedangkan zona rendah tersebar di wilayah pedalaman dengan tingkat polusi relatif rendah. Visualisasi spasial memperlihatkan pola distribusi polutan yang berkorelasi dengan aktivitas manusia dan kondisi geografis. Penelitian ini memberikan kontribusi dalam penyediaan informasi kualitas udara berbasis data spasial yang dapat mendukung pengambilan kebijakan lingkungan secara lebih tepat sasaran.

Kata Kunci: Kualitas udara, K-Means Clustering, CAMS, Polusi udara, Analisis spasial.

Abstract

Air quality is a critical indicator for assessing environmental conditions and their impact on public health. This study aims to classify air quality levels in North Sumatra Province using a data mining approach based on the K-Means Clustering algorithm. The dataset was obtained from the Copernicus Atmosphere Monitoring Service (CAMS) in NetCDF format, including key parameters such as PM_{2.5}, PM₁₀, CO, NO₂, and SO₂ during January 2026. The research stages consist of data preprocessing through spatial averaging, temporal aggregation, unit conversion, Min-Max normalization, and extraction of daily maximum values. The optimal number of clusters was determined using the Elbow method, resulting in three clusters representing air quality zones. The findings indicate that the study area is categorized into low, medium, and high pollution zones with distinct pollutant characteristics. High pollution zones are dominated by gaseous emissions in eastern coastal areas, medium zones are characterized by elevated particulate matter in highland regions, while low pollution zones are distributed in inland areas with relatively clean air conditions. Spatial visualization reveals pollutant distribution patterns influenced by human activities and geographical factors. This study contributes to spatial-based air quality information to support more effective environmental policy decision-making.

Keywords: Air quality, K-Means Clustering, CAMS, Air pollution, Spatial analysis.

1. PENDAHULUAN

Kualitas udara di lingkungan terbuka merupakan salah satu indikator fundamental dalam menilai tingkat kesehatan lingkungan sekaligus kesejahteraan masyarakat. Di Provinsi Sumatera Utara, dinamika pembangunan yang ditandai oleh pertumbuhan sektor industri, ekspansi lahan, serta peningkatan mobilitas transportasi secara bertahap berkontribusi terhadap peningkatan beban pencemaran udara [1]. Kondisi ini tidak dapat dipandang sebagai fenomena temporer, melainkan sebagai akumulasi tekanan lingkungan yang berpotensi menimbulkan dampak serius dalam jangka panjang. Paparan polutan udara secara kontinu diketahui berkorelasi dengan meningkatnya kasus penyakit infeksi saluran pernapasan (ISPA), gangguan paru-paru kronis, hingga komplikasi kardiovaskular. Oleh karena itu, pemantauan kualitas udara tidak lagi sekadar bersifat administratif, melainkan menjadi kebutuhan strategis dalam upaya perlindungan kesehatan publik.

Dalam kerangka pengukuran kualitas udara, terdapat sejumlah parameter yang telah diakui secara internasional sebagai indikator utama pencemaran. Penelitian ini menitikberatkan pada lima parameter kunci, yaitu partikulat berukuran halus PM_{2.5} dan PM₁₀, serta gas karbon monoksida (CO), nitrogen dioksida (NO₂), dan sulfur dioksida (SO₂) [2]. Partikel PM_{2.5} dan PM₁₀ memiliki karakteristik berbahaya karena ukurannya yang sangat kecil memungkinkan penetrasi hingga ke alveoli paru-paru, sehingga meningkatkan risiko gangguan respirasi. Di sisi lain, gas CO, NO₂, dan SO₂ umumnya berasal dari proses pembakaran bahan bakar fosil, aktivitas industri, serta emisi kendaraan bermotor.

Konsentrasi yang tinggi dari gas-gas tersebut tidak hanya berdampak pada kesehatan manusia, tetapi juga berkontribusi terhadap degradasi lingkungan, seperti pembentukan hujan asam dan penurunan kualitas ekosistem [3].

Selama ini, sistem pemantauan kualitas udara di Indonesia masih didominasi oleh penggunaan stasiun pengukuran berbasis darat (ground-based monitoring). Meskipun metode ini memiliki tingkat akurasi yang tinggi, cakupan spasialnya cenderung terbatas dan terfokus pada wilayah perkotaan tertentu. Keterbatasan jumlah stasiun dan distribusinya yang tidak merata menyebabkan banyak wilayah, khususnya daerah pedalaman, belum terpantau secara optimal [4]. Kondisi tersebut mengakibatkan adanya kesenjangan informasi yang cukup signifikan dalam memahami distribusi polusi udara secara menyeluruh.

Sebagai alternatif, pemanfaatan data penginderaan jauh melalui layanan Copernicus Atmosphere Monitoring Service (CAMS) menawarkan pendekatan yang lebih komprehensif. Data berbasis satelit memungkinkan pengamatan atmosfer secara luas tanpa dibatasi oleh batas administratif wilayah. Informasi yang dihasilkan disajikan dalam bentuk grid koordinat spasial yang mencakup lintang dan bujur, sehingga mampu menggambarkan kondisi kualitas udara secara kontinu di seluruh wilayah penelitian [5]. Dengan pendekatan ini, distribusi polutan dapat dianalisis tidak hanya secara temporal, tetapi juga secara spasial, sehingga memberikan gambaran yang lebih utuh terhadap dinamika pencemaran udara di Provinsi Sumatera Utara selama periode pengamatan Januari 2026.

Namun demikian, karakteristik data satelit yang bersifat multidimensi dan berukuran besar menuntut adanya metode analisis yang mampu mengelola kompleksitas tersebut secara efisien. Pendekatan konvensional tidak lagi memadai untuk mengekstraksi pola dari data dalam jumlah besar. Oleh karena itu, teknik data mining menjadi solusi yang relevan dalam mengidentifikasi struktur tersembunyi dalam dataset [6]. Salah satu metode yang banyak digunakan dalam analisis pola adalah algoritma K-Means Clustering, yang termasuk dalam kategori unsupervised learning. Algoritma ini bekerja dengan mengelompokkan data berdasarkan tingkat kemiripan karakteristik, tanpa memerlukan label awal. Dalam konteks penelitian ini, K-Means digunakan untuk mengelompokkan wilayah berdasarkan tingkat kemiripan konsentrasi polutan udara.

Untuk memastikan hasil pengelompokan yang optimal, diperlukan proses penentuan jumlah cluster yang tepat. Penelitian ini menggunakan metode Elbow sebagai pendekatan evaluasi berbasis variansi, yang bertujuan untuk mengidentifikasi jumlah cluster yang paling representatif [7]. Pendekatan ini membantu menghindari penentuan jumlah cluster secara subjektif, sehingga hasil analisis menjadi lebih objektif dan dapat dipertanggungjawabkan secara ilmiah.

Sejumlah penelitian sebelumnya telah mengkaji penerapan algoritma K-Means dalam analisis kualitas udara. Namun demikian, sebagian besar studi tersebut masih terbatas pada data stasiun pengukuran darat dengan cakupan wilayah yang relatif sempit, serta penyajian hasil yang cenderung bersifat deskriptif dalam bentuk tabel [8]. Integrasi antara data satelit berskala luas dengan metode clustering berbasis kecerdasan buatan, yang kemudian divisualisasikan dalam bentuk peta spasial seperti choropleth, masih jarang dilakukan, khususnya pada skala provinsi.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengklasifikasikan kualitas udara di wilayah Provinsi Sumatera Utara ke dalam beberapa kategori zona, yaitu zona rendah, sedang, dan tinggi berdasarkan tingkat konsentrasi polutan. Hasil klasifikasi ini tidak hanya disajikan dalam bentuk numerik, tetapi juga divisualisasikan secara spasial untuk memberikan gambaran yang lebih informatif. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi nyata dalam penyediaan informasi berbasis data yang dapat dimanfaatkan oleh pemangku kebijakan dalam merumuskan strategi pengendalian pencemaran udara yang lebih efektif dan tepat sasaran.

2. METODOLOGI PENELITIAN

2.1. Sumber Data dan Karakteristik Dataset

Penelitian ini memanfaatkan data sekunder yang diperoleh dari layanan Copernicus Atmosphere Monitoring Service (CAMS) dalam format NetCDF (Network Common Data Form). Format NetCDF digunakan karena kemampuannya dalam menyimpan data multidimensi yang mencakup variabel waktu, lintang (latitude), dan bujur (longitude) secara terstruktur. Hal ini menjadikan format tersebut sangat sesuai untuk analisis fenomena atmosfer yang bersifat dinamis dan spasial [9]. Wilayah kajian difokuskan pada Provinsi Sumatera Utara dengan rentang koordinat sekitar 0°–4° Lintang Utara dan 97°–100° Bujur Timur. Data yang digunakan memiliki resolusi spasial berbentuk grid, di mana setiap titik grid merepresentasikan kondisi atmosfer pada lokasi tertentu. Dengan struktur seperti ini, data memungkinkan analisis yang tidak hanya bersifat temporal, tetapi juga mampu menggambarkan distribusi spasial polutan udara secara komprehensif.

Parameter yang dianalisis dalam penelitian ini meliputi lima indikator utama kualitas udara, yaitu PM_{2.5}, PM₁₀, karbon monoksida (CO), nitrogen dioksida (NO₂), dan sulfur dioksida (SO₂). Pemilihan parameter ini didasarkan pada standar internasional yang ditetapkan oleh organisasi kesehatan dunia (WHO), yang menyatakan bahwa partikel halus seperti PM_{2.5} dan PM₁₀ memiliki keterkaitan langsung dengan gangguan sistem pernapasan, sedangkan gas seperti CO, NO₂, dan SO₂ umumnya dihasilkan dari aktivitas pembakaran bahan bakar fosil, industri, dan transportasi [10]. Penggunaan data berbasis satelit seperti CAMS memberikan keuntungan dalam hal cakupan wilayah yang luas dan kontinuitas pengamatan. Berbeda dengan stasiun pengukuran darat yang terbatas secara geografis, data CAMS mampu merepresentasikan kondisi atmosfer secara regional. Namun demikian, terdapat keterbatasan berupa resolusi yang relatif

lebih kasar, sehingga diperlukan tahapan pengolahan data yang sistematis agar hasil analisis tetap akurat dan representatif [11].

Tabel 1. Jenis-jenis Polutan Udara

Nama Variabel	Satuan	Simbol	Deskripsi
Particulate Matter 2.5	$\mu\text{g}/\text{m}^3$	PM2.5	Partikel debu sangat halus ($\leq 2.5 \mu\text{m}$) yang dapat masuk hingga ke paru-paru dan berbahaya bagi kesehatan
Particulate Matter 10	$\mu\text{g}/\text{m}^3$	PM10	Partikel debu kasar ($\leq 10 \mu\text{m}$) yang dapat terhirup dan mengganggu sistem pernapasan
Karbon Monoksida	$\mu\text{g}/\text{m}^3$ (ppm)	CO	Gas beracun tidak berwarna dan tidak berbau hasil pembakaran tidak sempurna
Nitrogen Dioksida	$\mu\text{g}/\text{m}^3$ (ppm)	NO ₂	Komponen angin zonal pada ketinggian 10 meter Gas polutan dari kendaraan dan industri yang dapat menyebabkan iritasi paru-paru
Sulfur Dioksida	$\mu\text{g}/\text{m}^3$ (ppm)	SO ₂	Gas hasil pembakaran bahan bakar fosil yang dapat menyebabkan gangguan pernapasan dan hujan asam

Pada kondisi aslinya, matriks observasi bawaan dari satelit CAMS masih direpresentasikan dalam bentuk notasi ilmiah (scientific notation) dengan nilai ukur yang terlampau kecil. Apabila data tersebut langsung diumpankan ke dalam algoritma, komputasi mesin akan mengalami kesulitan dalam mengukur jarak antar-titik datanya secara presisi. Menyiasati kondisi tersebut, serangkaian prosedur prakomputasi diterapkan pada data mentah. Langkah awal dimulai dengan perhitungan rata-rata spasial (spatial mean) guna meleburkan angka-angka dari belasan titik grid koordinat menjadi sebuah nilai representasi tunggal untuk keseluruhan wilayah Sumatera Utara. Selanjutnya, fluktuasi perekaman data setiap 12 jam diringkas melalui proses agregasi temporal (daily mean) sehingga menghasilkan satu angka mutlak per harinya. Menjelang tahap akhir pemrosesan, 31 baris data harian yang terbentuk dikalikan dengan besaran tertentu untuk menormalkan skala ukurnya. Nilai konsentrasi untuk debu partikulat dikalikan dengan faktor satu miliar, sementara kelompok gas dikalikan satu juta. Rangkaian instrumen ini menghasilkan sebuah dataset matang berupa tabel terstruktur dengan angka bulat yang siap didistribusikan ke tahapan algoritma machine learning.

2.2 Tahapan Pra-Pemrosesan Data

Tahapan pra-pemrosesan merupakan langkah krusial dalam penelitian ini karena data mentah yang diperoleh masih mengandung kompleksitas tinggi baik dari sisi dimensi maupun skala [12]. Proses ini dilakukan menggunakan bahasa pemrograman Python dengan bantuan pustaka seperti xarray dan pandas yang dirancang khusus untuk pengolahan data multidimensi dan tabular.



Gambar 1. Alur Tahapan Penelitian

a. Perhitungan Rata-rata Spasial (Spatial Averaging)

Data awal terdiri dari banyak titik grid yang merepresentasikan kondisi atmosfer pada berbagai lokasi dalam satu waktu. Untuk menyederhanakan analisis, dilakukan perhitungan rata-rata spasial dengan rumus:

$$X_{spatial} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

di mana x_i adalah nilai pada tiap titik koordinat dan n adalah jumlah titik grid.

Proses rata-rata spasial bertujuan untuk mereduksi kompleksitas data tanpa menghilangkan informasi utama yang terkandung di dalamnya. Dengan menggabungkan nilai dari berbagai titik koordinat, diperoleh satu nilai representatif yang mencerminkan kondisi rata-rata wilayah secara keseluruhan. Hal ini penting terutama pada tahap awal analisis untuk memahami tren umum kualitas udara.

Namun demikian, pendekatan ini juga memiliki konsekuensi berupa hilangnya detail variasi spasial lokal. Misalnya, wilayah perkotaan dengan tingkat polusi tinggi dapat tertutupi oleh wilayah pedesaan yang relatif bersih. Oleh karena itu, dalam penelitian ini, rata-rata spasial hanya digunakan pada tahap awal, sementara analisis clustering tetap mempertahankan data berbasis koordinat untuk menjaga informasi spasial tetap utuh.

b. Agregasi Temporal (Daily Mean)

Data CAMS memiliki resolusi waktu yang tinggi, seperti per jam atau per 12 jam. Untuk menyederhanakan analisis, dilakukan agregasi menjadi rata-rata harian menggunakan rumus:

$$X_{daily} = \frac{1}{m} \sum_{j=1}^m x_j \quad (2)$$

di mana m adalah jumlah observasi dalam satu hari.

Agregasi temporal dilakukan untuk mengurangi fluktuasi jangka pendek yang sering kali bersifat noise atau dipengaruhi oleh faktor sementara, seperti kondisi cuaca harian. Dengan mengubah data menjadi rata-rata harian, tren kualitas udara menjadi lebih stabil dan mudah dianalisis. Selain itu, dalam konteks kebijakan lingkungan, data harian lebih relevan dibandingkan data per jam. Hal ini karena standar kualitas udara umumnya ditetapkan dalam skala harian. Oleh karena itu, agregasi temporal tidak hanya berfungsi sebagai penyederhanaan data, tetapi juga sebagai penyesuaian terhadap kebutuhan analisis kebijakan.

c. Konversi Satuan dan Penyesuaian Skala

Data dari CAMS disajikan dalam bentuk scientific notation dengan nilai yang sangat kecil. Oleh karena itu dilakukan konversi satuan sebagai berikut:

- PM2.5 dan PM10 dikalikan 10^9
- CO, NO₂, dan SO₂ dikalikan 10^6

Konversi satuan bertujuan untuk meningkatkan keterbacaan data serta mempermudah interpretasi hasil analisis. Nilai yang terlalu kecil dapat menyebabkan kesulitan dalam memahami pola data, terutama ketika ditampilkan dalam bentuk tabel atau grafik. Selain itu, skala data yang terlalu kecil dapat menyebabkan ketidakstabilan dalam perhitungan jarak pada algoritma K-Means. Oleh karena itu, konversi ini menjadi langkah penting untuk memastikan bahwa data berada dalam rentang yang lebih realistis dan dapat diolah secara numerik dengan lebih stabil.

2.2.1 Seleksi Data dan Ekstraksi Fitur

Penelitian ini menggunakan subset data selama 10 hari, yaitu:

- 1–5 Januari 2026
- 27–31 Januari 2026

Selanjutnya dilakukan ekstraksi nilai maksimum harian menggunakan fungsi:

$$X_{max} = \max(x_1, x_2, \dots, x_n) \quad (3)$$

Pemilihan subset data dilakukan untuk menangkap variasi kondisi atmosfer pada awal dan akhir bulan. Hal ini penting karena pola polusi udara dapat berubah seiring waktu akibat faktor meteorologi maupun aktivitas manusia. Penggunaan nilai maksimum harian merupakan pendekatan yang berfokus pada kondisi polusi terburuk. Dalam konteks kesehatan lingkungan, paparan puncak memiliki dampak yang lebih signifikan dibandingkan nilai rata-rata. Oleh karena itu, pendekatan ini dianggap lebih representatif dalam menggambarkan risiko kualitas udara terhadap kesehatan manusia.

2.2.2 Normalisasi Data (Min-Max Scaling)

Sebelum dilakukan clustering, data dinormalisasi menggunakan metode Min-Max:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (4)$$

Normalisasi merupakan langkah krusial dalam algoritma K-Means karena metode ini berbasis jarak Euclidean. Tanpa normalisasi, variabel dengan skala besar akan mendominasi perhitungan jarak, sehingga hasil clustering menjadi bias [13]. Dengan menggunakan metode Min-Max, seluruh variabel diubah ke dalam rentang 0 hingga 1. Hal ini memastikan bahwa setiap parameter memiliki kontribusi yang seimbang dalam proses pengelompokan. Selain itu, normalisasi juga membantu mempercepat konvergensi algoritma karena data berada dalam skala yang seragam.

2.2.3 Metode Clustering K-Means

Algoritma K-Means digunakan untuk mengelompokkan data ke dalam K cluster berdasarkan kemiripan karakteristik. Jarak antar data dihitung menggunakan Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (4)$$

K-Means dipilih karena memiliki keunggulan dalam hal kesederhanaan dan efisiensi komputasi. Algoritma ini mampu menangani dataset berukuran besar dengan waktu proses yang relatif cepat. Selain itu, hasil clustering yang dihasilkan mudah diinterpretasikan karena setiap cluster memiliki centroid yang merepresentasikan karakteristik kelompok tersebut. Namun, K-Means juga memiliki beberapa kelemahan, seperti sensitivitas terhadap outlier dan ketergantungan pada jumlah cluster yang ditentukan di awal. Selain itu, K-Means cenderung menghasilkan cluster berbentuk bulat (spherical), sehingga kurang optimal jika data memiliki distribusi yang kompleks.

2.2.4 Penentuan Jumlah Cluster (Metode Elbow)

Jumlah cluster optimal ditentukan menggunakan metode Elbow dengan menghitung nilai WCSS:

$$WCSS = \sum_{i=1}^k \sum_{x \in C_i} (x - \mu_i)^2 \quad (5)$$

Metode Elbow digunakan untuk menentukan jumlah cluster yang paling optimal dengan cara mengamati perubahan nilai WCSS terhadap jumlah cluster. Pada penelitian ini, nilai WCSS mengalami penurunan signifikan hingga K=3, kemudian penurunannya menjadi lebih landai. Fenomena ini menunjukkan bahwa penambahan cluster setelah K=3 tidak memberikan peningkatan yang signifikan dalam kualitas clustering. Oleh karena itu, K=3 dipilih sebagai jumlah cluster optimal yang mewakili tiga kategori kualitas udara, yaitu rendah, sedang, dan tinggi.

2.2.5 Evaluasi dan Visualisasi Data

Hasil clustering divisualisasikan menggunakan berbagai metode, antara lain:

- Heatmap korelasi
- Grafik Elbow
- Grafik batang (bar chart)
- Scatter plot
- Peta spasial (scatter dan choropleth)

Visualisasi memiliki peran penting dalam penelitian ini karena tidak hanya berfungsi sebagai alat penyajian data, tetapi juga sebagai sarana validasi hasil analisis. Melalui visualisasi, pola hubungan antar variabel dapat diamati dengan lebih jelas, sehingga membantu dalam interpretasi hasil clustering.

Penggunaan peta spasial menjadi nilai tambah utama dalam penelitian ini, karena mampu menghubungkan hasil analisis numerik dengan kondisi geografis secara langsung. Hal ini sangat penting dalam konteks pengambilan keputusan, karena pemangku kebijakan dapat dengan mudah mengidentifikasi wilayah yang membutuhkan perhatian khusus dalam pengendalian polusi udara.

3. HASIL DAN PEMBAHASAN

Bagian ini akan mengupas tuntas hasil dari seluruh proses pengolahan data spasial yang telah dirancang sebelumnya, mulai dari wujud data mentahnya hingga menjelma menjadi peta persebaran polusi di Provinsi Sumatera Utara.

3.1 Ekstraksi Data Mentah dan Konversi Satuan Polutan

Langkah pertama yang membuahkan hasil nyata dalam penelitian ini adalah pembongkaran arsip data mentah berformat NetCDF yang baru saja diunduh dari satelit CAMS. Setelah data koordinat dan waktu berhasil diekstrak ke dalam wujud tabel (dataframe), ditemukan sebuah kendala teknis: angka-angka bawaan rekaman satelit masih menggunakan format notasi ilmiah dengan nominal desimal yang sangat kecil. Jika dibiarkan, angka-angka ini tidak hanya membingungkan saat dibaca, tetapi juga rentan menimbulkan eror saat mesin melakukan perhitungan jarak kluster [14]. Untuk mengatasi hal tersebut, dilakukan operasi konversi satuan secara matematis agar deretan angka tersebut kembali normal dan merepresentasikan kondisi udara yang sebenarnya di dunia nyata. Materi partikulat debu (PM2.5 dan PM10) dikonversi menggunakan faktor pengali satu miliar ($\times 1.000.000.000$) sehingga satuannya berubah menjadi mikrogram (μg). Sementara itu, kelompok gas emisi yakni Karbon Monoksida (CO), Nitrogen Dioksida (NO₂), dan Sulfur Dioksida (SO₂) dikalikan dengan angka satu juta ($\times 1.000.000$) agar wujudnya berubah menjadi miligram (mg).

Hasil dari konversi tahap awal ini membuahkan sebuah tabel data historis yang jauh lebih rapi dan siap untuk dianalisis lebih lanjut.

3.2 Ekstraksi Data Ekstrem Harian dan Normalisasi Skala

Setelah format satuan polutan disesuaikan, data yang berjumlah sangat masif ini kembali disaring secara ketat guna mempertajam fokus analisis. Penelitian ini sengaja menyoroti kondisi kualitas udara pada momen paling fluktuatif, yakni pada 10 hari yang terbagi menjadi lima hari di awal tahun (1-5 Januari 2026) dan lima hari di akhir bulan (27-31 Januari 2026). Selain itu, untuk menghindari bias fluktuasi minor yang terjadi setiap jamnya, algoritma diprogram untuk hanya menarik satu nilai puncak (*maximum value*) harian dari masing-masing polutan di setiap titik observasi. Alhasil, data yang digunakan benar-benar merepresentasikan momen polusi paling pekat dan ekstrem yang menyelimuti wilayah Sumatera Utara pada hari tersebut.

Tantangan selanjutnya muncul dari rentang nominal kelima variabel yang sangat timpang. Sebagai contoh, fluktuasi kadar debu PM_{2.5} di udara Sumatera Utara mungkin hanya berada di rentang puluhan mikrogram, namun sebaran gas Karbon Monoksida (CO) hasil pembakaran mesin dan industri angkanya bisa meroket melampaui seribu miligram. Ketimpangan skala ekstrem ini sangat berbahaya jika langsung diumpankan ke dalam mesin perhitungan *Euclidean Distance* milik K-Means, karena sistem secara otomatis akan memprioritaskan variabel yang memiliki nominal terbesar (CO) dan mengabaikan variabel dengan nominal kecil.

Untuk menetralkan ketimpangan tersebut, dilakukan operasi penyesuaian skala menggunakan metode *Min-Max Scaler*. Perhitungan matematis ini memaksa seluruh ragam rentang angka polutan, baik yang besar maupun kecil, untuk ditekan ke dalam skala tunggal yang proporsional, yakni di antara rasio \$0\$ (terendah) hingga \$1\$ (tertinggi). Berkat penyesuaian ini, kelima indikator polusi akhirnya memiliki hak suara dan pembobotan yang setara saat algoritma mencari kemiripan data antar-koordinat wilayah.

Cuplikan hasil akhir tabel normalisasi untuk sampel salah satu titik observasi koordinat (Latitude -0.8000, Longitude 97.2000) selama periode 10 hari ekstrem dapat diamati pada Tabel 2.

Tabel 2. Cuplikan Hasil Ekstraksi Maksimal Harian dan Normalisasi Min-Max

Tanggal	latitude	longitude	PM 2.5	PM 10	Carbon Monoxide (CO)	Nitrogen Dioxide (NO ₂)	Sulphur Dioxide (SO ₂)
2026-01-01	-0.8000	97.2000	0.0427	0.0820	0.2650	0.0525	0.0034
2026-01-02	-0.8000	97.2000	0.0911	0.1451	0.5028	0.0460	0.0011
2026-01-03	-0.8000	97.2000	0.1079	0.1550	0.4450	0.0434	0.0022
2026-01-04	-0.8000	97.2000	0.1531	0.2014	0.4559	0.0388	0.0052
2026-01-05	-0.8000	97.2000	0.1874	0.2549	0.3748	0.0466	0.0112
2026-01-27	-0.8000	97.2000	0.1616	0.2078	0.1858	0.0092	0.0098
2026-01-28	-0.8000	97.2000	0.1038	0.1684	0.1899	0.0094	0.0070
2026-01-29	-0.8000	97.2000	0.0789	0.1261	0.2364	0.0030	0.0048
2026-01-30	-0.8000	97.2000	0.0612	0.0952	0.2760	0.0017	0.0082
2026-01-31	-0.8000	97.2000	0.0587	0.0967	0.2004	0.0000	0.0080

Sebagaimana yang terlihat pada Tabel 2, kolom tanggal sudah menampilkan rekam jejak polusi dengan frekuensi yang presisi, yakni satu angka observasi per hari. Di saat yang bersamaan, seluruh angka polutan tidak lagi menggunakan

satuan aslinya yang berjauhan, melainkan telah berwujud rasio desimal di bawah angka 1. Dengan struktur data yang telah terkalibrasi secara sempurna inilah, proses *Clustering* siap untuk dieksekusi ke tahap selanjutnya.

3.3 Analisis Keterkaitan Antar-Polutan (Uji Korelasi Pearson)

Sebelum membiarkan algoritma K-Means membagi wilayah ke dalam beberapa klaster, ada satu analisis pendahuluan yang sangat menarik untuk dikupas, yakni melihat seberapa erat "persahabatan" atau keterkaitan antar-polutan di udara Sumatera Utara. Tujuannya sederhana: kita ingin memastikan apakah kemunculan suatu gas beracun selalu diikuti oleh gas lainnya, atau justru mereka bergerak secara independen akibat sumber emisi yang berbeda-beda.

Untuk menguji hubungan ini, diterapkan perhitungan statistik **Korelasi Pearson**. Rumus matematis ini bekerja dengan mengukur kekuatan hubungan linear antara dua variabel polutan (*X dan Y*), dengan persamaan sebagai berikut:

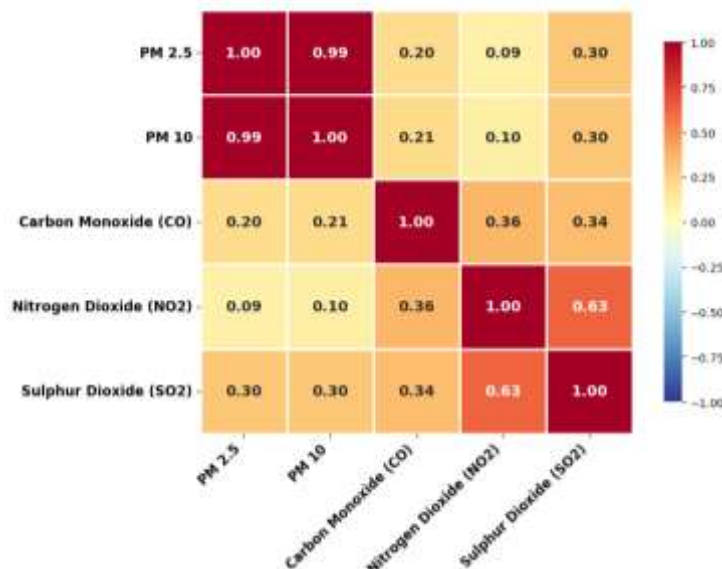
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (6)$$

Hasil dari perhitungan ini berupa angka (koefisien *r*) yang bergerak dari -1 hingga 1. Jika angkanya semakin mendekati 1, itu tandanya kedua polutan tersebut sangat serasi dan selalu muncul bersamaan. Berdasarkan kalkulasi terhadap seluruh rekaman data bulan Januari 2026, nilai korelasi antar-kelima polutan berhasil dirangkum ke dalam wujud matriks pada Tabel 3.

Tabel 3. Matriks Nilai Korelasi Pearson Antar-Polutan

Polutan Udara	PM 2.5	PM 10	CO	NO2	SO2
PM 2.5	1.00	0.99	0.20	0.09	0.30
PM 10	0.99	1.00	0.21	0.10	0.30
Carbon Monoxide (CO)	0.20	0.21	1.00	0.36	0.34
Nitrogen Dioxide (NO2)	0.09	0.10	0.36	1.00	0.63
Sulphur Dioxide (SO2)	0.30	0.30	0.34	0.63	1.00

Agar hubungan angka-angka di atas lebih mudah dicerna secara visual, matriks tersebut juga diproyeksikan ke dalam bentuk *Heatmap* atau peta panas korelasi, yang dapat dilihat pada Gambar 2.



Gambar 2. Visualisasi Heatmap Korelasi Polutan Udara

Membaca Tabel 3 dan visualisasi *Heatmap* tersebut memberikan *insight* yang sangat berharga. Temuan paling mencolok adalah hubungan ibarat "dua sisi mata uang" antara debu PM_{2.5} dan PM₁₀. Keduanya mencetak angka korelasi nyaris sempurna, yakni **0.99**. Fakta ini membuktikan bahwa jika suatu daerah di Sumatera Utara diserang oleh polusi debu kasar (PM₁₀), sudah bisa dipastikan udara di sana juga tercemar oleh debu halus (PM_{2.5}) yang mematikan. Keduanya dipastikan lahir dari rahim sumber emisi fisik yang sama, seperti debu jalanan, proyek konstruksi, atau kebakaran lahan.

Temuan menarik lainnya terlihat pada hubungan antara gas Nitrogen Dioksida (NO₂) dan Sulfur Dioksida (SO₂) yang mencatatkan tingkat korelasi cukup kuat di angka **0.63**. Mengingat kedua gas ini merupakan "langganan" hasil residu pabrik dan asap knalpot kendaraan bermotor, tingginya angka korelasi ini menjadi sinyal kuat bahwa kawasan industri dan jalur lintas padat di Sumatera Utara merupakan penyumbang utama dari kedua gas beracun tersebut secara bersamaan. Di sisi lain, gas Karbon Monoksida (CO) terlihat cukup independen dan memiliki korelasi yang lemah dengan polutan lainnya (berada di kisaran 0.20 hingga 0.36). Dinamika polusi yang saling tumpang tindih inilah yang membuat udara Sumatera Utara sangat kompleks, sehingga penggunaan mesin K-Means untuk memetakan zonanya menjadi sangat krusial dan tak bisa ditawar lagi.

3.4 Penentuan Jumlah Klaster Optimal (Metode *Elbow*)

Berbeda dengan cara kerja otak manusia yang bisa mengelompokkan sesuatu secara instingtif, algoritma K-Means memiliki satu kelemahan mendasar: mesin ini terlahir "buta" mengenai berapa jumlah kelompok (klaster) yang harus ia bentuk. Jika kita tidak memberinya instruksi angka K (jumlah zona), algoritma ini tidak akan bisa bekerja. Namun, menentukan angka K secara asal-asalan tentu akan mencederai nilai objektivitas penelitian. Oleh karena itu, diterapkanlah sebuah evaluasi matematis yang dikenal dengan sebutan Metode *Elbow*.

Metode ini bekerja dengan melakukan simulasi pengelompokan berulang kali (mulai dari K=1 hingga K=10). Pada setiap simulasi, mesin akan menghitung seberapa padat dan rapat titik-titik data di dalam kelompok tersebut menggunakan rumus *Within-Cluster Sum of Squares* (WCSS). Secara sederhana, rumus WCSS mengukur total jarak kuadrat antara setiap titik data observasi (x_i) dengan titik pusat klasternya (c_j), dengan persamaan:

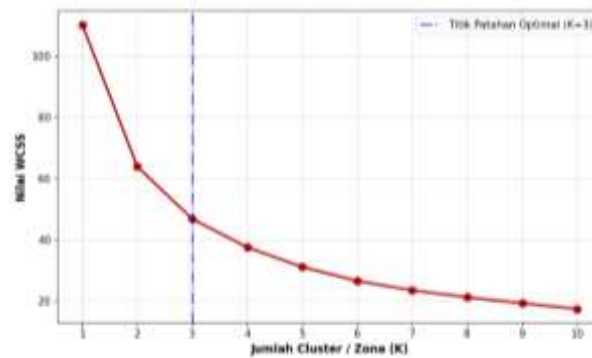
$$WCSS = \sum_{j=1}^K \sum_{i \in C_j} (x_i - c_j)^2$$

Semakin kecil nilai WCSS, berarti titik-titik polusi di dalam zona tersebut semakin mirip dan mengelompok dengan rapat. Hasil perhitungan mesin untuk 10 kali simulasi uji coba pembagian zona dapat dilihat secara rinci pada Tabel 4.

Tabel 4. Hasil Evaluasi WCSS Berdasarkan Jumlah Klaster (K)

Jumlah Klaster / Zona (K)	Nilai Variansi Jarak (WCSS)	Penurunan Jarak
1	110.05	-
2	63.86	Turun drastis
3	46.80	Turun tajam (Titik Siku)
4	37.55	Mulai melandai
5	31.16	Melandai
6	26.59	Melandai
7	23.53	Melandai
8	21.27	Melandai
9	19.31	Melandai
10	17.47	Sangat landai

Untuk memudahkan mata menangkap momen perubahan angka tersebut, data pada Tabel 4 juga diproyeksikan ke dalam bentuk grafik garis dua dimensi yang tersaji pada Gambar 3.



Gambar 3. Grafik Metode Elbow untuk Penentuan Kluster Optimal

Membaca pergerakan grafik pada Gambar Z memberikan kepastian matematis yang sangat solid. Perhatikan bagaimana garis merah terjun bebas secara ekstrem dari K=1 (110.05) menuju K=2 (63.86) dan masih berlanjut menurik tajam hingga menyentuh angka K=3 (46.80). Namun, keanehan terjadi tepat setelah melewati angka 3. Laju penurunan garis tiba-tiba melambat drastis, seolah-olah mesin sudah mencapai "titik jenuh", sehingga garisnya tampak patah dan melandai (membentuk visualisasi seperti siku lengan manusia).

Dalam kaidah analisis *Data Mining*, titik patahan siku inilah yang dinobatkan sebagai angka optimal. Hal ini membuktikan bahwa membagi kualitas udara Sumatera Utara menjadi **3 zona (K=3)** adalah keputusan yang paling pas secara alami, tidak terlalu sedikit sehingga datanya bias, dan tidak terlalu banyak sehingga datanya tumpang tindih. Berbekal angka K=3 inilah, algoritma K-Means akhirnya diberi "lampu hijau" untuk mengeksekusi pembagian wilayah ke dalam kelompok Zona Rendah (Hijau), Zona Sedang (Kuning), dan Zona Tinggi (Merah).

3.5 Eksekusi K-Means dan Karakteristik Pusat Klaster (*Centroid*)

Berbekal nilai K=3 yang diperoleh dari evaluasi metode *Elbow*, tahap inti dari pengelompokan data pun mulai dijalankan. Algoritma K-Means bertugas mempartisi ribuan titik data observasi harian ke dalam tiga zona yang berbeda. Pembagian ini tidak dilakukan secara acak, melainkan dihitung secara matematis menggunakan pendekatan Jarak *Euclidean* (*Euclidean Distance*) [15]. Rumus ini berfungsi untuk mengukur seberapa dekat jarak antara suatu titik observasi aktual (x_i) dengan titik pusat dari sebuah klaster atau *centroid* (c_j). Persamaan tersebut dirumuskan sebagai berikut:

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^n (x_{ik} - c_{jk})^2} \quad (7)$$

Titik-titik koordinat yang memiliki jarak terpendek (paling mirip) ke salah satu *centroid* akan langsung ditarik dan dikunci menjadi anggota dari klaster tersebut. Secara bawaan, mesin komputer hanya mengenali ketiga kelompok ini dengan label angka numerik, yakni Klaster 0, Klaster 1, dan Klaster 2.

Setelah seluruh data berhasil dipilah, tugas krusial selanjutnya adalah menerjemahkan nilai dari ketiga *centroid* tersebut ke dalam bahasa yang dipahami manusia. Mengingat data yang masuk ke dalam mesin sebelumnya sudah berwujud rasio 0 hingga 1 (akibat proses normalisasi), maka angka ini wajib dikembalikan ke wujud aslinya melalui teknik *Inverse Transform* (Denormalisasi). Lewat proses inilah, label numerik dari mesin akhirnya bisa kita petakan menjadi tingkatan zona bahaya.

Sebagai catatan penting sebelum membedah angka-angka polusi tersebut, kita perlu memahami terlebih dahulu ancaman nyata dari dua parameter utamanya, yakni **PM 10** dan **PM 2.5** (*Particulate Matter*):

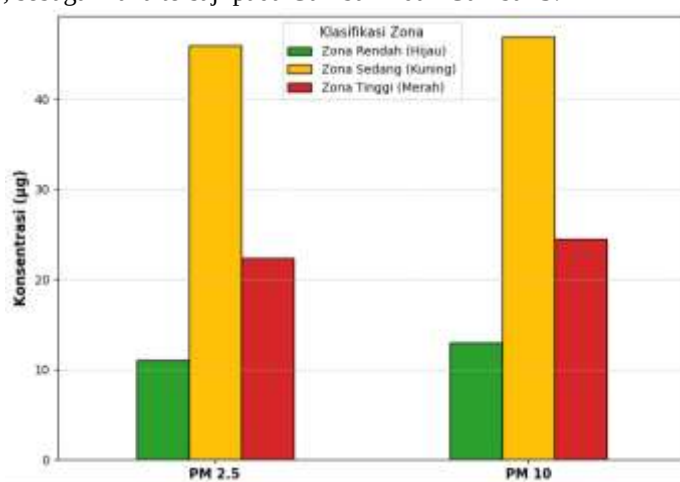
- **PM 10:** Merupakan debu partikulat kasar berukuran 10 mikrometer atau lebih kecil. Biasanya berasal dari debu jalanan, material konstruksi, atau serbuk sari. Meski kotor, partikel ini umumnya masih bisa disaring oleh bulu hidung atau tenggorokan manusia sehingga jarang mencapai paru-paru dalam.
- **PM 2.5:** Ini adalah polutan pembunuh yang sesungguhnya. Ukurannya sangat mikroskopis, yakni 2.5 mikrometer (sekitar 30 kali lebih tipis dari sehelai rambut manusia). Sumber utamanya adalah asap knalpot, emisi pabrik, dan pembakaran lahan. Saking kecilnya, PM 2.5 bisa menembus masker kain biasa, masuk langsung ke kantung udara paru-paru (alveolus), bahkan menyusup masuk ke dalam pembuluh darah.

Dengan membawa pemahaman tersebut, mari kita bongkar hasil terjemahan nilai pusat (*centroid*) dari masing-masing klaster pada Tabel 5.

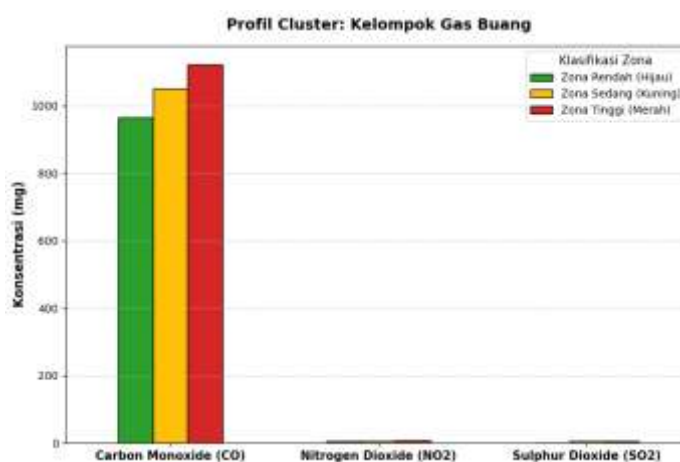
Tabel 5. Nilai Pusat Klaster (*Centroid*) Berdasarkan Parameter Polutan Asli

Label Mesin	Kategori Zona	PM 2.5 (µg)	PM 10 (µg)	CO (mg)	NO2 (mg)	SO2 (mg)
Klaster 0	Zona Rendah (Hijau)	11.07	13.00	964.14	4.59	1.03
Klaster 1	Zona Sedang (Kuning)	45.95	46.87	1050.43	6.15	4.43
Klaster 2	Zona Tinggi (Merah)	22.39	24.43	1121.28	6.82	5.72

Agar perbedaan karakteristik di antara ketiga klaster tersebut tidak sekadar menjadi deretan angka yang membingungkan, profil *centroid* pada Tabel 5 juga diproyeksikan ke dalam dua grafik batang (*Bar Chart*) yang terpisah antara materi debu dan gas, sebagaimana tersaji pada Gambar 4 dan Gambar 5.



Gambar 4. Profil Klaster untuk Kategori Partikulat Debu



Gambar 5. Profil Klaster untuk Kategori Gas Buang

Apabila kita bedah tabel dan kedua grafik di atas, masing-masing klaster ternyata menyimpan karakteristik polusi yang sangat spesifik dan berbeda satu sama lain:

1. **Klaster 2 / Zona Tinggi (Merah):** Klaster ini memegang rekor terburuk untuk kategori gas buang beracun. Kadar Karbon Monoksida (CO) di klaster ini menembus puncak tertinggi di angka 1121.28 mg, disusul oleh lonjakan NO2 (6.82 mg) dan SO2 (5.72 mg). Tingginya kadar ketiga gas pembuangan ini memberikan indikasi kuat bahwa wilayah di Klaster 2 adalah episentrum dari padatnya mobilitas kendaraan bermotor dan cerobong asap kawasan industri.
2. **Klaster 1 / Zona Sedang (Kuning):** Terdapat sebuah anomali lingkungan yang sangat mencolok di sini. Meski status gas buangnya berada di level menengah, Klaster 1 justru mencatatkan rekor paling mematikan untuk paparan partikulat debu. Angka PM2.5 di zona ini menembus batas kewajaran hingga menyentuh 45.95 µg/m³.

g\$. Temuan ini mengisyaratkan bahwa wilayah-wilayah yang tergabung dalam Klaster 1 kemungkinan besar sedang dikepong oleh aktivitas yang merusak tanah secara fisik, seperti pembukaan hutan skala besar, perkebunan sawit yang mengering, atau proyek infrastruktur masif yang menerbangkan debu-debu halus mematikan.

3. **Klaster 0 / Zona Rendah (Hijau):** Dari semua kelompok yang dibentuk mesin, ini adalah klaster dengan kondisi udara yang paling bisa bernapas lega. Seluruh parameter polutannya, baik partikel pembunuh PM 2.5 maupun sisa gas emisi industri, terkunci pada tingkat konsentrasi yang paling rendah dan aman bagi kesehatan penduduknya.

3.6 Distribusi Populasi Klaster dan Visualisasi Scatter Plot

3.7

Setelah algoritma K-Means berhasil merumuskan titik pusat (*centroid*) untuk ketiga kelompok, mesin beralih ke tahapan "pelabelan massal" (*data labeling*). Pada tahap ini, seluruh riwayat data observasi yang berjumlah 1.260 baris (merupakan akumulasi dari ratusan titik koordinat satelit selama 10 hari pengamatan ekstrem) dieksekusi satu per satu. Setiap titik observasi dilekatkan pada satu label zona tertentu berdasarkan jarak terdekatnya dengan *centroid*.

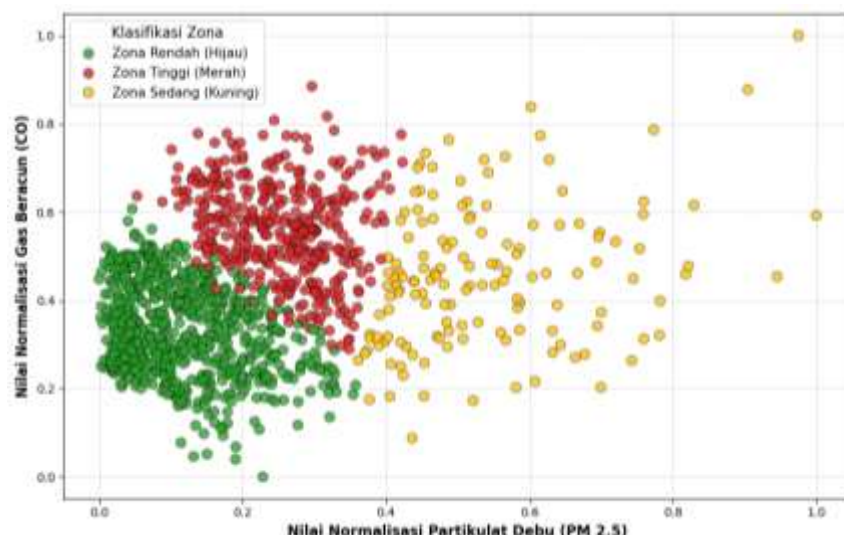
Hasil kalkulasi pembagian populasi untuk masing-masing zona kualitas udara di Sumatera Utara disajikan secara rinci pada Tabel 6.

Tabel 6. Distribusi Jumlah Titik Observasi per Kategori Zona

Kategori Zona	Jumlah Titik Observasi	Persentase Populasi (%)
Zona Rendah (Hijau)	727	57.70
Zona Sedang (Kuning)	143	11.35
Zona Tinggi (Merah)	390	30.95

Jika kita membedah lebih dalam postur data pada Tabel 6, terdapat sebuah realitas lingkungan yang cukup melegakan sekaligus mengkhawatirkan. Di satu sisi, lebih dari separuh total pengamatan (57.70%) masih didominasi oleh Zona Rendah (Hijau), yang mengartikan bahwa secara umum daratan Sumatera Utara masih memiliki ruang udara yang relatif bersih. Namun di sisi lain, volume observasi yang terjerumus ke dalam Zona Tinggi (Merah) menempati urutan kedua dengan persentase yang cukup mencolok, yakni menyentuh angka 30.95%.

Guna membuktikan secara empiris bahwa algoritma K-Means telah bekerja dengan baik dalam memisahkan ketiga kelompok populasi tersebut tanpa adanya batas yang tumpang tindih (*overlapping*), dilakukan proyeksi visual menggunakan diagram sebar (*Scatter Plot*) [16]. Dari kelima parameter yang ada, dipilih dua polutan yang memiliki sifat paling kontras sebagai sumbu pengamatan, yakni partikulat debu (PM 2.5) pada sumbu X dan gas Karbon Monoksida (CO) pada sumbu Y. Hasil pemetaan titik sebar ini dapat dilihat pada Gambar 6.



Gambar 6. Scatter Plot Pemisahan Zona Polusi Berdasarkan PM 2.5 dan CO

Visualisasi *Scatter Plot* pada Gambar 6 memberikan konfirmasi grafis yang sangat memuaskan. Mesin komputasi terbukti berhasil membelah lautan data menjadi tiga "pulau" yang sangat terstruktur:

1. **Gugusan Hijau (Kiri Bawah):** Titik-titik observasi pada zona rendah terlihat menumpuk rapat di sudut kiri bawah, membuktikan konsistensinya yang memiliki kadar debu (PM 2.5) dan gas beracun (CO) di level paling minimal.
2. **Gugusan Merah (Tengah Atas):** Sesuai dengan profil *centroid*-nya, titik-titik zona merah tampak mengambang dan mendominasi area atas grafik. Hal ini mempertegas karakteristik utamanya sebagai area yang sangat pekat oleh gas buang beracun (CO), meski tingkat debunya berada di batas wajar.
3. **Gugusan Kuning (Menyebar ke Kanan):** Visualisasi ini memberikan pembuktian telak atas temuan anomali sebelumnya. Titik-titik zona kuning secara agresif menyebar dan memanjang ke arah sisi kanan grafik (mendekati angka 1.0 pada sumbu X). Pola sebaran ini mengunci fakta bahwa wilayah-wilayah di zona kuning memang merupakan episentrum dari krisis paparan debu halus (PM 2.5) ekstrem di Sumatera Utara.

3.8 Identifikasi Spasial Administratif dan Tabulasi Zona Dominan

Setelah algoritma K-Means sukses melabeli 1.260 baris data observasi, penelitian ini dihadapkan pada satu tantangan spasial terakhir: menerjemahkan deretan angka koordinat satelit (Garis Lintang dan Bujur) menjadi nama wilayah administratif Kabupaten/Kota di daratan Sumatera Utara. Satelit CAMS secara bawaan merekam data murni berdasarkan titik koordinat geometris, tanpa menyertakan label batas wilayah kenegaraan.

Untuk mengatasi kekosongan label tersebut, diterapkanlah metode *Spatial Bounding Box* atau penyekatan batas koordinat. Melalui bahasa pemrograman Python, sistem diinstruksikan untuk memagari rentang koordinat tertentu dan mengikatnya dengan nama wilayah yang ada di bawahnya. Sebagai contoh, seluruh titik observasi yang jatuh di atas koordinat Lintang ≥ 3.0 dan Bujur ≥ 98.5 secara otomatis diidentifikasi sebagai kawasan metropolitan yang mencakup Kota Medan, Binjai, Deli Serdang, dan Langkat.

Masalah berikutnya muncul akibat lamanya durasi observasi. Mengingat pengamatan dilakukan selama 10 hari ekstrem, satu kabupaten pasti dinaungi oleh banyak titik observasi yang status zonanya bisa saja berubah-ubah (misalnya, Medan hari ini berstatus merah, namun besok menjadi kuning). Untuk menentukan status mutlak dan eksklusif dari suatu kabupaten, penelitian ini menggunakan pendekatan probabilitas statistik yakni **Nilai Modus (Mo)** atau *Majority Voting*. Secara konseptual, Modus adalah nilai yang paling sering muncul dalam sebuah himpunan data. Rumus logikanya diterapkan dengan mengevaluasi seluruh riwayat label klaster (Z) yang jatuh di suatu daerah (D), lalu mengunci status akhir daerah tersebut berdasarkan frekuensi label terbanyak:

$$Status_{Final}(D) = Mode(Z_{D_1}, Z_{D_2}, \dots, Z_{D_n})$$

Melalui perumusan probabilitas ini, apabila mayoritas titik observasi harian yang melayang di atas Kabupaten Asahan menunjukkan indikator Zona Tinggi (Merah), maka secara eksklusif seluruh daratan Kabupaten Asahan akan divonis masuk ke dalam Zona Merah. Hasil rekapitulasi akhir yang mengikat nama wilayah dominan beserta rata-rata konsentrasi polutannya disajikan secara komprehensif pada Tabel 7.

Tabel 7. Persebaran Administratif Wilayah Dominan Berdasarkan Zona Polusi

Klasifikasi Zona	Cakupan Wilayah Dominan di Sumatera Utara	PM 2.5 (µg)	PM 10 (µg)	CO (mg)	SO2 (mg)	Titik Observasi
Zona Tinggi (Merah)	Medan, Binjai, Deli Serdang, Langkat, Batu Bara, Asahan, Tanjung Balai, Serdang Bedagai	25.77	28.36	1145.53	8.14	240
Zona Sedang (Kuning)	Karo, Dairi, Pakpak Bharat	33.62	34.21	983.10	6.00	40
Zona Rendah (Hijau)	Humbang Hasundutan, Simalungun, Toba, Samosir, Tapanuli Utara, Tapanuli Selatan, Padang Lawas, Padangsidempuan, dkk.	16.16	17.87	994.23	1.46	980

Tabulasi pada Tabel 7 akhirnya menyibak tabir misteri mengenai letak geografis dari masing-masing klaster, sekaligus mengonfirmasi anomali data yang ditemukan pada sub-bab sebelumnya:

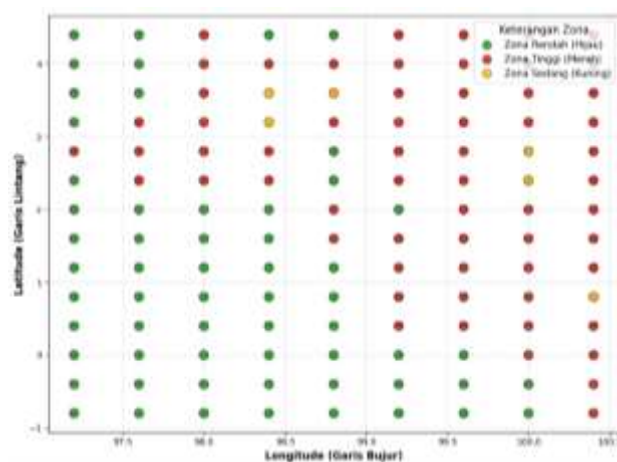
1. **Episentrum Emisi Gas (Zona Merah):** Sangat masuk akal jika klaster dengan paparan gas buang CO dan SO₂ tertinggi jatuh pada deretan wilayah pesisir timur Sumatera Utara (Medan, Deli Serdang, Batu Bara, hingga Asahan). Wilayah-wilayah ini merupakan urat nadi perekonomian, pusat padat penduduk, gerbang tol lintas daerah, serta rumah bagi kawasan industri kelas berat. Beban emisi dari jutaan knalpot kendaraan dan cerobong pabrik terakumulasi secara masif di langit kawasan ini.
2. **Episentrum Partikulat Debu (Zona Kuning):** Misteri melonjaknya angka debu mematikan (PM 2.5) di Zona Sedang akhirnya terpecahkan setelah melihat daftar wilayahnya. Kabupaten Karo, Dairi, dan Pakpak Bharat merupakan kawasan dataran tinggi. Tingginya angka partikulat debu di wilayah ini sangat berkorelasi erat dengan karakteristik alam dan aktivitas warganya, seperti intensitas pembukaan lahan pertanian masif, serta yang paling utama: paparan abu vulkanik dari aktivitas Gunung Sinabung di Kabupaten Karo yang kerap menebarkan partikel silika halus ke udara sekitarnya.
3. **Kawasan Udara Bersih (Zona Hijau):** Mendominasi dengan 980 titik observasi, zona ini merangkum kawasan pedalaman, area pegunungan, hingga bentang alam Danau Toba (Samosir, Toba, Humbang Hasundutan, hingga Tapanuli). Minimnya aktivitas industri berat dan masih luasnya tutupan vegetasi hutan di kawasan selatan ini menjadikannya sebagai paru-paru utama bagi Provinsi Sumatera Utara dengan tingkat polusi paling minim.

3.9 Visualisasi Pemetaan Spasial Berbasis Grid Koordinat

Sebagai muara dari seluruh rangkaian analisis komputasi, penelitian yang menyoroti krisis kualitas udara di **Provinsi Sumatera Utara selama periode Januari 2026** ini diakhiri dengan sebuah rekonstruksi visual. Jika pada tahapan sebelumnya kita telah menerjemahkan klaster ke dalam wujud tabel administratif, pada tahap pamungkas ini kita mengembalikan data tersebut ke habitat aslinya, yakni ruang spasial geografis.

Metode yang digunakan untuk membangun visualisasi ini adalah proyeksi pemetaan Kartesius dua dimensi (*2D Cartesian Mapping*). Setiap titik data polusi (P_i) diplotkan kembali ke atas kanvas digital menggunakan parameter koordinat aslinya, di mana Garis Bujur (*Longitude*) bertindak sebagai sumbu mendatar (X) dan Garis Lintang (*Latitude*) bertindak sebagai sumbu vertikal (Y). Secara matematis, penempatan setiap titik diwakili oleh fungsi spasial berikut: $P_i = (\text{Longitude}_i, \text{Latitude}_i, K_i)$

Di mana K_i merupakan identitas klaster (warna zona) yang telah diputuskan oleh algoritma K-Means untuk titik tersebut. Melalui teknik ini, ribuan deret angka yang terkurung di dalam tabel Excel akhirnya menjelma menjadi sebuah Peta Distribusi Spasial yang menampilkan konfigurasi rona lingkungan udara secara utuh, sebagaimana direpresentasikan pada Gambar 7.



Gambar 7. Peta Distribusi Spasial Polusi Atmosfer Sumatera Utara (Januari 2026)

Memperhatikan susunan titik koordinat pada Gambar 7 bagaikan melihat potret rontgen dari kondisi paru-paru Provinsi Sumatera Utara. Formasi titik-titik tersebut tidak menyebar secara acak, melainkan membentuk pola bentang alam yang sangat selaras dengan realitas geografis di lapangan:

1. **Sabuk Merah di Sisi Timur:** Perhatikan rentetan titik berwarna merah yang mendominasi sisi kanan peta (berada di atas garis *Longitude* 98.5 hingga 100.5). Formasi ini mencetak cetak biru yang sangat presisi dari letak geografis Pesisir Timur Sumatera Utara. Wilayah ini memang dikenal sebagai jalur sutra perekonomian, membentang dari Langkat, Medan, Deli Serdang, hingga Asahan. Warna merah solid di area ini menjadi bukti

visual yang tak terbantahkan bahwa konsentrasi industri dan laju urbanisasi telah menciptakan lorong emisi gas buang (CO dan SO₂) terpanjang di provinsi ini.

2. **Bintik Kuning di Titik Kritis:** Menariknya, titik-titik kuning (Zona Sedang) tampak terjepit dan berkumpul di titik koordinat spesifik (sekitar *Latitude* 3.0 dan *Longitude* 98.5). Letak koordinat ini sangat identik dengan lanskap dataran tinggi Kabupaten Karo dan sekitarnya. Sebagaimana yang telah dianalisis sebelumnya, kehadiran titik kuning di area ini secara visual mengonfirmasi lokus terjadinya krisis paparan partikulat debu (PM 2.5), yang sangat rentan dipengaruhi oleh faktor topografi dan aktivitas vulkanik setempat.
3. **Benteng Hijau di Sisi Barat dan Selatan:** Sepanjang area kiri hingga ke bawah peta didominasi oleh kerapatan titik hijau yang mewakili Zona Rendah. Gugusan spasial ini mencerminkan topografi sabuk hijau pegunungan Bukit Barisan, bentang alam Danau Toba, hingga wilayah Tapanuli di bagian selatan. Visualisasi ini memberikan angin segar sekaligus pesan konservasi yang tegas: selama wilayah-wilayah ini bebas dari intervensi industri berat, ia akan terus berfungsi sebagai benteng pertahanan terakhir yang menyuplai udara bersih bagi Provinsi Sumatera Utara.

Secara keseluruhan, pemetaan spasial ini membuktikan bahwa algoritma *K-Means* tidak hanya cerdas dalam mengenali pola angka matematis, tetapi juga sangat akurat dalam memotret kondisi geografis dan demografis di dunia nyata tanpa campur tangan manusia.

3.10 Pemetaan Spasial Tematik (*Choropleth Map*) Wilayah Administratif

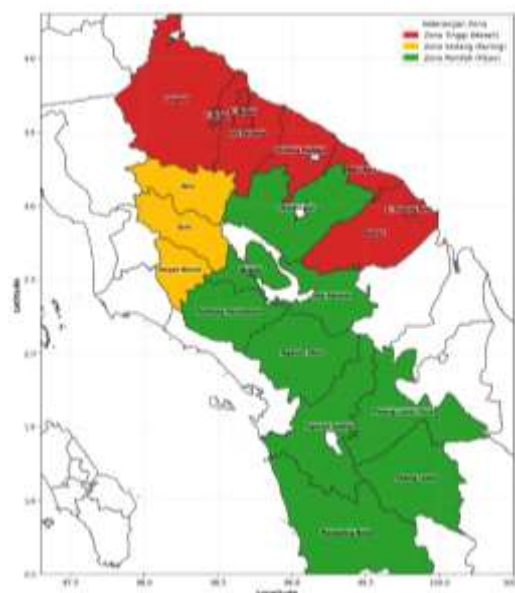
Sebagai puncak dari keseluruhan analisis komputasi di dalam penelitian ini, visualisasi grid koordinat yang sebelumnya masih berupa titik-titik sebar kini dikonversi menjadi sebuah pemetaan wilayah administratif yang utuh. Teknik pemetaan yang digunakan adalah *Choropleth Map*, yakni jenis peta tematik di mana setiap area atau poligon wilayah diwarnai secara penuh berdasarkan kategori data statistiknya.

Agar bentuk poligon kabupaten di peta selaras dengan data yang kita miliki, sistem melakukan integrasi data menggunakan metode Pencocokan Atribut (*Attribute/Database Join*). Algoritma diprogram untuk mencari kesamaan nama daerah antara tabel data hasil komputasi *K-Means* dengan pangkalan data peta geometris global (GeoJSON). Secara konseptual, operasi penggabungan ini dapat direpresentasikan melalui fungsi relasional berikut:

$$Wilayah_{Spasial} = Geometri(D_i) \bowtie Status_Klaster(D_i)$$

Di mana D_i melambangkan identitas nama Kabupaten/Kota. Melalui proses relasi ($\bowtie$) ini, ketika sistem mendeteksi nama "Kota Medan" pada data klaster berstatus Zona Merah, maka secara otomatis seluruh luasan poligon tanah Kota Medan di atas peta akan terarsir dengan warna merah pekat.

Secara geografis, lensa pengamatan pada peta ini dikalibrasi secara spesifik untuk membidik daratan utama Provinsi Sumatera Utara. Pemotongan area (*Bounding Box*) dikunci dengan presisi pada rentang koordinat Garis Bujur (*Longitude*) antara 97.3° hingga 100.5° Bujur Timur, serta Garis Lintang (*Latitude*) antara 0.5° hingga 4.3° Lintang Utara. Area-area di luar batas observasi tersebut (seperti Kepulauan Nias) secara sengaja dianulir dan diberi warna putih (Tidak Terobservasi) agar tidak mendistorsi fokus analisis wilayah daratan utama. Hasil visualisasi akhir dari pemetaan ini terpampang dengan sangat jelas pada Gambar 8.



Gambar 8. Peta Choropleth Persebaran Zona Polusi Wilayah Sumatera Utara

Jika kita menyanggahkan formasi titik koordinat dengan blok warna kabupaten pada peta di atas, kita bisa mendedah anatomi kualitas udara di Sumatera Utara secara sangat presisi. Pola pembagian zona ini terbukti tidak acak, melainkan terikat kuat pada kondisi topografi, aktivitas manusia, dan batas koordinat berikut:

1. Lorong Emisi Pesisir Timur (Zona Tinggi / Merah)

Visualisasi peta menunjukkan dominasi warna merah yang merajai secara solid di sisi timur laut. Jika ditelusuri dari kode logikanya, kawasan ini membentang secara presisi mulai dari koordinat Latitude $\geq 3.0^\circ$ LU dan Longitude $\geq 98.5^\circ$ BT, berlanjut menyusuri pesisir ke arah selatan di Latitude $2.5^\circ - 3.0^\circ$ LU dan Longitude $\geq 99.0^\circ$ BT. Rentang koordinat ini merupakan cetak biru dari urat nadi perekonomian Sumatera Utara, yang merangkum wilayah Kota Medan, Binjai, Deli Serdang, Langkat, hingga ke Batu Bara, Asahan, dan Tanjung Balai. Terkuncinya seluruh kawasan pesisir timur ini ke dalam Zona Merah memberikan konfirmasi empiris yang tak terbantahkan. Masifnya kawasan industri kelas berat, padatnya arus transportasi di Jalan Lintas Sumatera, serta tingginya laju urbanisasi di koridor ini telah menumpuk emisi gas buang (khususnya CO, NO₂, dan SO₂) hingga mencapai level tertinggi di provinsi ini.

2. Kantong Partikulat Dataran Tinggi (Zona Sedang / Kuning)

Tepat bersebelahan dengan zona merah, terdapat anomali wilayah berwarna kuning yang terjepit di sebuah blok koordinat sempit. Kawasan ini dibatasi secara ketat pada Latitude 2.5° hingga 3.5° LU dan Longitude 98.0° hingga 98.5° BT. Blok spasial ini merupakan kawasan dataran tinggi yang menaungi Kabupaten Karo, Dairi, dan Pakpak Bharat. Kehadiran zona kuning di wilayah pegunungan ini menjawab temuan janggal pada analisis *centroid* sebelumnya, di mana zona sedang justru memiliki kadar debu halus (PM 2.5 dan PM 10) yang sangat ekstrem. Secara geografis, tingginya angka partikulat debu di koordinat ini sangat masuk akal dan valid. Wilayah ini tidak hanya didominasi oleh aktivitas pembukaan lahan pertanian masif yang mengering, tetapi secara khusus wilayah Kabupaten Karo merupakan rumah bagi Gunung Sinabung. Tingginya paparan debu di udara kawasan ini merupakan imbas langsung dari sebaran abu vulkanik berskala mikroskopis yang secara periodik menyelimuti langit sekitarnya.

3. Sabuk Hijau Penyeimbang Ekosistem (Zona Rendah / Hijau)

Sebagai penawar dari pekatnya polusi di kedua zona sebelumnya, hamparan warna hijau mendominasi separuh lebih luas wilayah di bagian tengah hingga selatan peta. Wilayah udara bersih ini menguasai bentangan alam di sebelah barat Longitude 98.0° BT serta merajai area selatan di bawah Latitude 2.5° LU. Gugusan hijau ini merepresentasikan kawasan Kabupaten Simalungun, Toba, Samosir, Humbang Hasundutan, meluas hingga ke Tapanuli Utara, Tapanuli Selatan, dan Padang Lawas. Rona hijau yang tebal ini menjadi cerminan dari minimnya aktivitas industri berat dan masih luasnya ruang terbuka hijau di bentang alam Danau Toba serta deretan Bukit Barisan. Dari kacamata lingkungan, pemetaan ini mengirimkan pesan konservasi yang tegas: kawasan selatan dan pedalaman ini adalah tameng ekologis dan paru-paru raksasa yang masih berhasil menjaga kualitas udara Sumatera Utara di ambang batas aman.

3.11 Perbandingan dengan Penelitian Sebelumnya (Novelty)

Untuk mengukur sejauh mana kebaruan (*novelty*) dan kontribusi dari penelitian ini, temuan yang dihasilkan perlu disandingkan dengan kajian-kajian terdahulu pada bidang yang sama. Berdasarkan tinjauan literatur, mayoritas penelitian terkait klasterisasi kualitas udara menggunakan algoritma K-Means masih bertumpu pada data Indeks Standar Pencemar Udara (ISPU) yang diperoleh dari stasiun pemantau darat (*ground-based stations*). Pendekatan konvensional tersebut memang memiliki tingkat presisi yang baik untuk skala lokal, namun memiliki kelemahan spasial yang cukup mendasar. Stasiun pemantau umumnya hanya berpusat di kawasan perkotaan atau kawasan industri, sehingga menciptakan *blind spot* (area buta data) yang sangat luas untuk wilayah pedalaman, pesisir, dan pegunungan.

Penelitian ini hadir untuk menutupi celah kelemahan tersebut. Pembeda paling signifikan dari riset ini terletak pada transisi sumber data, yakni pemanfaatan teknologi penginderaan jauh dari satelit CAMS. Dengan mengolah jaringan koordinat spasial, metode ini mampu merekam fluktuasi polusi di seluruh daratan Provinsi Sumatera Utara tanpa terpotong oleh batas administratif.

Keunggulan dari pendekatan ini terbukti secara empiris melalui hasil pemetaan yang berhasil mengidentifikasi anomali lingkungan di Zona Sedang (Kuning). Melalui data satelit, algoritma ini mampu mendeteksi krisis partikulat debu mematikan (PM 2.5 dan PM 10) di kawasan dataran tinggi Kabupaten Karo dan sekitarnya. Apabila pengukuran hanya mengandalkan stasiun darat yang berada di pusat kota, paparan debu akibat aktivitas vulkanik dan pembukaan lahan di area pegunungan tersebut hampir pasti luput dari pantauan. Oleh karena itu, kerangka kerja komputasi spasial yang ditawarkan dalam riset ini terbukti jauh lebih komprehensif, objektif, dan sangat relevan untuk dijadikan rujukan oleh pemerintah daerah dalam menyusun kebijakan mitigasi tata ruang dan lingkungan hidup.

4. KESIMPULAN

Penelitian ini secara komprehensif menyimpulkan bahwa integrasi data penginderaan jauh dari satelit *Copernicus Atmosphere Monitoring Service* (CAMS) dengan algoritma *K-Means Clustering* merupakan instrumen yang sangat efektif dan objektif dalam membedah pola distribusi polusi udara di tingkat provinsi. Berdasarkan hasil pemrosesan data pada periode pengamatan Januari 2026, ditemukan fakta bahwa kualitas udara di Provinsi Sumatera Utara secara spasial terbagi ke dalam tiga zona dengan karakteristik polutan yang sangat heterogen. Zona Tinggi yang ditandai dengan rona merah teridentifikasi mendominasi koridor pesisir timur, khususnya di kawasan metropolitan Medan, Deli Serdang, hingga Asahan. Wilayah ini terkonfirmasi sebagai episentrum polusi gas buang (CO, NO₂, dan SO₂) yang bersumber dari tingginya intensitas aktivitas industri dan kepadatan arus transportasi di jalur lintas timur. Di sisi lain, penelitian ini berhasil mengungkap anomali lingkungan yang signifikan pada Zona Sedang yang mencakup wilayah dataran tinggi seperti Kabupaten Karo, Dairi, dan Pakpak Bharat. Meskipun konsentrasi gas emisi di wilayah tersebut berada pada level moderat, kawasan ini justru menghadapi krisis paparan partikulat debu halus (PM 2.5 dan PM 10) yang melampaui batas zona lainnya. Temuan ini memberikan bukti empiris bahwa polusi di kawasan pegunungan Sumatera Utara lebih dipicu oleh faktor topografi, aktivitas agrikultur masif, serta residu abu vulkanik dari aktivitas Gunung Sinabung. Sebagai penyeimbang ekosistem, Zona Rendah (Hijau) masih mendominasi lebih dari separuh wilayah provinsi, terutama di sektor selatan dan kawasan konservasi Danau Toba, di mana vegetasi hutan primer masih mampu menekan angka polutan ke titik terendah. Secara keseluruhan, pemetaan spasial ini memberikan kesimpulan definitif bahwa karakteristik polusi atmosfer di Sumatera Utara memiliki korelasi linear yang kuat dengan tata letak geografis, beban demografis, dan pusat pertumbuhan ekonomi masing-masing wilayah.

REFERENCES

- [1] R. A. Pratama and S. Wibowo, "Penerapan Algoritma K-Means untuk Klasterisasi Indeks Standar Pencemar Udara (ISPU) di Wilayah Metropolitan," *Jurnal Ilmu Komputer dan Informatika*, vol. 12, no. 2, pp. 145–156, 2023.
- [2] D. Rahmawati and B. Santoso, "Pemanfaatan Penginderaan Jauh Satelit untuk Pemantauan Polutan PM_{2.5} dan Emisi Karbon di Indonesia," *Jurnal Pengelolaan Lingkungan Berkelanjutan*, vol. 8, no. 1, pp. 33–47, 2024.
- [3] K. Khairunnisa and M. Afif, "Analisis Konsentrasi Particulate Matter (PM₁₀ dan PM_{2.5}) di Wilayah Perkotaan Medan," *Jurnal Teknik Lingkungan*, vol. 7, no. 1, pp. 45–52, 2021.
- [4] M. Sitorus and H. Simanjuntak, "Analisis Spasial Sebaran Gas Karbon Monoksida (CO) Akibat Aktivitas Transportasi Darat," *Jurnal Sains dan Teknologi*, vol. 11, no. 3, pp. 210–218, 2022.
- [5] A. S. Ginting and T. Karo-Karo, "Kajian Paparan Debu dan Partikel Atmosfer di Kawasan Dataran Tinggi Sumatera Utara," *Jurnal Meteorologi dan Geofisika*, vol. 22, no. 2, pp. 89–101, 2021.
- [6] T. Hidayat and A. Saputra, "Implementasi Data Mining Menggunakan K-Means Clustering untuk Pemetaan Kualitas Udara," *Jurnal Informatika dan Sistem Informasi*, vol. 4, no. 1, pp. 12–23, 2023.
- [7] D. Santoso and K. Wijaya, "Evaluasi Kualitas Udara Ambien Berdasarkan Parameter Gas NO₂ dan SO₂ di Kawasan Industri," *Jurnal Rekayasa Lingkungan*, vol. 15, no. 2, pp. 134–142, 2022.
- [8] P. Lestari and N. Fitriani, "Studi Perbandingan Kualitas Udara Antara Wilayah Urban dan Rural Menggunakan Pemetaan Spasial," *Jurnal Geografi dan Lingkungan*, vol. 10, no. 1, pp. 56–68, 2024.
- [9] A. Putra dan D. Wibowo, "Penerapan K-Means Clustering untuk Analisis Kualitas Udara di Indonesia," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 8, no. 2, pp. 123–130, 2021.
- [10] S. Mishra, P. Sahu, and R. Kumar, "Air Quality Analysis Using K-Means Clustering Approach," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 234–240, 2019.
- [11] B. Santoso dan R. Nugroho, "Analisis Data Lingkungan Menggunakan Metode Clustering," *Jurnal Sains Data Indonesia*, vol. 5, no. 1, pp. 45–53, 2022.
- [12] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2011.



- [13] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [14] J. Flemming *et al.*, “The CAMS interim reanalysis of carbon monoxide, ozone and nitrogen dioxide for 2003–2015,” *Atmospheric Chemistry and Physics*, vol. 15, no. 4, pp. 1995–2017, 2015.
- [15] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.
- [16] A. Inness *et al.*, “The CAMS reanalysis of atmospheric composition,” *Atmospheric Chemistry and Physics*, vol. 19, no. 6, pp. 3515–3556, 2019.