# Penerapan Algoritma K-Means Clustering untuk Pola Kualitas Udara di Jakarta

**Muhammad Arya Fayyadh Razana, Nur Juzieatul Alifahb, Qurrata A’yunic**

*a,b,c Fakultas Teknik, Informatika, Universitas Mulawarman*

*Corresponding Author:*

*am.aryafayyadhrazan123@gmail.com*

## ABSTRAk

Pencemaran udara di daerah perkotaan, khususnya di Jakarta, menjadi isu krusial yang mempengaruhi kesehatan masyarakat dan kualitas lingkungan. Penelitian ini bertujuan untuk menganalisis pola kualitas udara di Jakarta dari tahun 2010 hingga 2023 dengan menggunakan metode K-Means *Clustering* berdasarkan data Indeks Standar Pencemaran Udara (ISPU). Tahapan pengolahan data berdasarkan metodologi CRISP-DM diterapkan untuk mengolah dan menganalisis data secara sistematis. Adapun beberapa tahapannya mencakup *business understanding*, *data understanding*, *data preparation*, *modeling*, dan *evaluation*. Hasil penelitian menunjukkan bahwa data terbagi menjadi tiga cluster: cluster sehat, sedang, dan tidak sehat. Cluster 0, yang mencakup stasiun DKI1 dan DKI2, menunjukkan kualitas udara yang lebih baik, sedangkan cluster 1, yang terdiri dari stasiun DKI3, DKI4, dan DKI5, menunjukkan tingkat polusi yang lebih tinggi. Temuan ini memberikan wawasan penting bagi pengambil kebijakan dalam merumuskan strategi pengendalian polusi udara yang lebih efektif. Dengan demikian, hasil penelitian ini tidak hanya berkontribusi pada pemahaman kualitas udara di Jakarta, tetapi juga menekankan perlunya tindakan mitigasi yang berbasis data untuk meningkatkan kesehatan masyarakat dan lingkungan.

### **Keywords:** Algoritma, AQI, *Clustering*, K-Means, Kualitas Udara

## ABSTRACT

Air pollution in urban areas, especially in Jakarta, is a crucial issue that affects public health and environmental quality. This study aims to analyze air quality patterns in Jakarta from 2010 to 2023 using the K-Means Clustering method based on Air Pollution Standard Index (ISPU) data. Data processing stages based on the CRISP-DM methodology are applied to process and analyze data systematically. The stages include business understanding, data understanding, data preparation, modeling, and evaluation. The results showed that the data were divided into three clusters: healthy, moderate, and unhealthy clusters. Cluster 0, which includes stations DKI1 and DKI2, shows better air quality, while cluster 1, which consists of stations DKI3, DKI4, and DKI5, shows higher pollution levels. These findings provide important insights for policy makers in formulating more effective air pollution control strategies. Thus, the results of this study not only contribute to the understanding of air quality in Jakarta, but also emphasize the need for data-driven mitigation actions to improve public health and the environment.

### **Keywords:** *Air quality*, Algorithm, AQI, *Clustering*, K-Means.

## PENDAhuluan

#### Lingkungan merupakan kesatuan ruang yang memuat seluruh daya, benda, keadaan, serta makhluk hidup, termasuk manusia dan perilakunya, yang mempengaruhi kelangsungan hidup dan alam. Udara bersih menjadi faktor utama dalam menentukan kualitas lingkungan (Adinda et al., 2021). Namun, pesatnya pertumbuhan kota dan aktivitas manusia, seperti penggunaan energi di industri, pembangkit listrik, dan transportasi, menyebabkan pembakaran bahan bakar yang mencemari udara (Mahendrasyah et al., 2024).

#### Pencemaran udara disebabkan oleh akumulasi polutan dari aktivitas industri, kendaraan, dan sumber emisi lainnya (Adityo et al., 2023). Polutan utama meliputi karbon monoksida (CO), sulfur dioksida (SO₂), nitrogen dioksida (NO₂), ozon permukaan (O₃), serta partikel debu seperti PM₁₀ dan PM₂.₅ (Pertiwi et al., 2024). Konsentrasi polutan yang tinggi berdampak buruk pada kesehatan manusia dan lingkungan. Jakarta merupakan salah satu kota yang sering mengalami kualitas udara tidak sehat berdasarkan data IQAir (2025) karena merupakan salah satu kota dengan aktivitas urbanisasi tertinggi di Indonesia, serta tingkat kepadatan penduduk dan kendaraan yang sangat besar. Kondisi ini menjadikan Jakarta rentan terhadap pencemaran udara yang kompleks dan bervariasi antar wilayah. Oleh karena itu, dilakukan analisis *clustering* untuk mengetahui pola penyebaran polusi udara di berbagai stasiun pemantauan, guna memberikan gambaran yang lebih jelas terhadap tingkat keparahan dan karakteristik polusi di masing-masing area. Pengelompokan ini penting sebagai dasar dalam perencanaan kebijakan lingkungan yang lebih terarah, seperti penentuan lokasi prioritas untuk pengendalian emisi, perbaikan tata ruang, atau kampanye kesadaran masyarakat.

## Untuk memahami pola pencemaran udara secara lebih sistematis dapat menggunakan teknik data mining *clustering*. Salah satu metode *clustering* yang efektif yaitu K-Means *Clustering*. Metode ini mengelompokkan wilayah berdasarkan karakteristik polusi tahunan berdasarkan ISPU (Salsabila et al., 2024). Berdasarkan jurnal Penerapan Algoritma K-Means untuk Klasterisasi Indeks Standar Pencemaran Udara (Mahendrasyah et al., 2024) dan Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM (Dhewayani et al, 2022), metode *elbow* digunakan dalam penelitian karena memiliki kemampuan untuk menentukan jumlah cluster optimal secara objektif melalui identifikasi titik di mana penambahan cluster tidak lagi memberikan pengurangan signifikan pada nilai inertia, sehingga menghindari over-clustering yang dapat mengaburkan pola pencemaran yang sebenarnya. Diharapkan hasil analisis dapat membantu pemerintah dalam merancang kebijakan pengelolaan kualitas udara berbasis data dan merumuskan solusi untuk mengurangi dampak pencemaran (Salsabila et al., 2024).

## LITERATUR REVIEW

1. *Data Mining*

*Data mining* adalah proses menganalisis data untuk menemukan pola dan hubungan yang tersembunyi dalam dataset besar. Teknik ini menggabungkan penggunaan metode statistik, kecerdasan buatan, dan pembelajaran mesin untuk mengambil informasi yang tidak dapat ditemukan secara manual, menurut (Kusrini & Luthfi., 2021). Ini mencakup langkah-langkah kritis seperti pembersihan data, pemilihan fitur, penciptaan model, dan evaluasi hasil. *Data mining* dapat digunakan di berbagai bidang termasuk bisnis untuk analisis perilaku konsumen, kesehatan untuk prediksi penyakit, dan keuangan untuk deteksi penipuan. Kekuatan utamanya terletak pada pengolahan data kompleks dan memberikan wawasan yang dapat ditindaklanjuti yang meningkatkan pengambilan keputusan strategis.

1. Algoritma K-Means

Algoritma K-Means adalah salah satu metode pengelompokan partisi yang paling populer karena kesederhanaan dan efisiensi yang ditawarkannya. Seperti yang dijelaskan oleh (Dharmawan., 2022), proses iteratif ini dimulai dengan menentukan jumlah *cluster* (K) dan menginisialisasi *centroid* secara acak. Kemudian, perhitungan jarak untuk setiap titik data ke *centroid*, pengelompokan data, dan pembaruan posisi centroid dilakukan hingga konvergensi. Keuntungan utama yang ditawarkan K-Means adalah kecepatan komputasi, membuatnya lebih disukai untuk dataset yang lebih besar, namun, algoritma ini menunjukkan beberapa kekurangan berdasarkan pilihan K pengguna dan kerentanan tinggi terhadap *outlier*. Metode inisialisasi *centroid* lainnya seperti K-Means++ telah mencoba memperbaiki keterbatasan ini dengan proses yang dikembangkan untuk inisialisasi *centroid*.

1. *Clustering*

Salah satu metode yang diterapkan dalam pengolahan data adalah *clustering*, yang berfungsi untuk mengelompokkan dataset berdasarkan kesamaan yang ada. Sebagaimana telah dijelaskan sebelumnya, teknik ini termasuk dalam kategori *unsupervised learning*, yang berarti tidak memerlukan label atau identitas yang ditentukan untuk data yang dianalisis. Dengan demikian, *clustering* dapat digunakan untuk mengidentifikasi pola dalam dataset yang belum terklasifikasi. (Hani et al., 2022) menjelaskan bahwa prinsip dasar teknik ini adalah untuk memaksimalkan keseragaman dalam setiap *cluster*, sekaligus memaksimalkan perbedaan antar *cluster*. Penerapan teknik ini sangat luas, mencakup segmentasi pasar, pengelompokan dokumen, dan analisis penyebaran penyakit. Di antara berbagai metode yang ada, terdapat *partitioning methods*, yang lebih dikenal sebagai K-Means, *hierarchical clustering*, dan *density-based clustering*. Setiap metode memiliki keunggulan dan keterbatasan yang bergantung pada karakteristik data yang sedang diproses.

1. Rapidminer

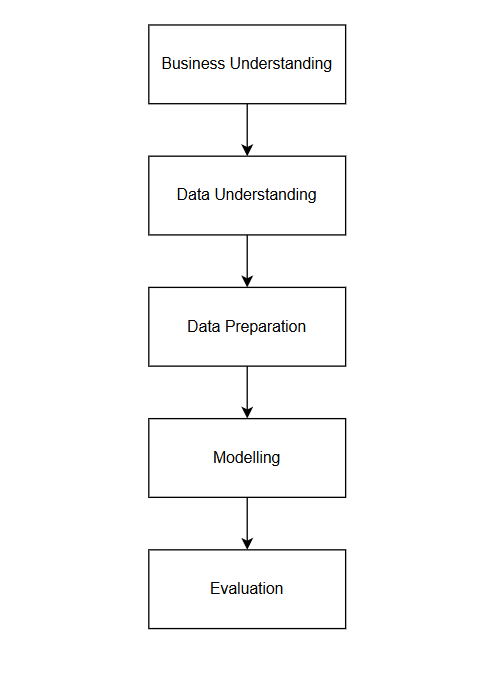
RapidMiner muncul sebagai platform analisis data yang powerful dengan antarmuka visual yang memudahkan proses *data mining* tanpa memerlukan keahlian pemrograman mendalam. (Afrahul., 2024) menyatakan alat ini memberikan lapangan terintegrasi untuk seluruh alur kerja analitik data termasuk pra-pemrosesan, pembuatan model, evaluasi dan hasil visualisasi. Ditekankan pula mengenai dukungan format-data seperti database dan file excel, library algoritma yang ekstensif terutama pada *clustering* yaitu K-Means serta kemampuan visualisasi data secara ramah pengguna. RapidMiner banyak digunakan di lingkungan perkuliahan ataupun industri, terutama karena kemudahan dalam mempercepat prototyping model meski dalam versi gratis, terhadap beberapa fitur dibandingkan dengan versi komersial. Meskipun dalam versi gratis terdapat lebih banyak fitur, RapidMiner tetap menjadi pilihan untuk implementasi *data mining* dan *machine learning* karena kemudahan dalam mengintegrasikan berbagai teknik analisis.

1. *Elbow Method*

Metode *Elbow* adalah pendekatan yang digunakan untuk menentukan jumlah klaster (k) yang optimal dalam algoritma klasterisasi, seperti K-Means. Prinsip dasar metode ini adalah dengan mengamati grafik hubungan antara jumlah klaster (k) dan nilai *Sum of Squared Errors* (SSE). Dalam grafik tersebut, akan tampak penurunan nilai SSE seiring dengan bertambahnya jumlah klaster. Pada titik tertentu, penurunan SSE mulai melambat dan membentuk sudut atau “siku” (*elbow*), yang menandakan titik optimal jumlah klaster. Titik siku ini dipilih sebagai nilai k yang optimal karena setelah titik tersebut, penambahan klaster tidak memberikan peningkatan signifikan terhadap penurunan SSE. Dengan demikian, metode *Elbow* membantu dalam memilih jumlah klaster yang paling efisien dengan mempertimbangkan keseimbangan antara kompleksitas model dan kualitas pemisahan data (Sari., 2021).

## METode penelitian

Pada penelitian ini, metode yang digunakan adalah CRISP-DM (*Cross Industry Standard Process for Data Mining*), yang merupakan pendekatan standar dalam proses *data mining* (Wardani., 2021). Metode ini diterapkan untuk mengolah dan menganalisis data secara sistematis guna menemukan pola atau informasi yang dapat digunakan dalam pengambilan keputusan. Dengan mengikuti tahapan CRISP-DM, penelitian ini memastikan bahwa proses eksplorasi data, pemodelan, dan evaluasi dilakukan secara optimal untuk memperoleh hasil yang akurat dan relevan. Pendekatan CRISP-DM juga terbukti efektif dalam berbagai penelitian terdahulu yang mengaplikasikan K-Means dalam analisis kualitas udara (Widuri., 2023; Mawaddah Anjelita., 2023).



**Gambar 1. Tahapan Metode Penelitian**

Pada Gambar 1 ditunjukkan bahwa Business Understanding, Data Understanding, Data Preparation, dan Modelling merupakan tahapan inti dalam analisis data, yang dimulai dari pemahaman terhadap tujuan bisnis, dilanjutkan dengan analisis serta persiapan data, hingga pembangunan model untuk menghasilkan solusi atas masalah yang diidentifikasi.

Algoritma K-Means adalah metode klastering yang mengelompokkan data berdasarkan kemiripan fitur. Proses dimulai dengan menentukan jumlah klaster (k) yang diinginkan, memilih *centroid* awal secara acak, dan menghitung jarak data ke setiap *centroid* menggunakan *Euclidean Distance*. Data kemudian dimasukkan ke dalam klaster terdekat, dan posisi *centroid* diperbarui berdasarkan rata-rata titik dalam klaster. Proses ini diulang hingga posisi *centroid* stabil atau jumlah iterasi tercapai (Adinda Amalia., 2023).

Kelebihan utama dari algoritma K-Means adalah kemampuannya untuk memproses data dalam jumlah besar dengan cepat, yang menjadikannya sangat sesuai untuk analisis kualitas udara. Dengan menggunakan K-Means, wilayah dapat dikelompokkan berdasarkan tingkat pencemaran udara, seperti konsentrasi PM10, CO, dan NO2. Hasil klasterisasi ini dapat memberikan wawasan yang lebih baik mengenai zona-zona dengan risiko tinggi terhadap polusi udara, yang pada akhirnya mendukung pembuatan kebijakan kesehatan yang lebih efektif dan berbasis data (Kartika Dian Pertiwi., 2023).

1. *Business Understanding (Pemahaman Bisnis)*

Tahap ini mencakup pemahaman terhadap kebutuhan dan tujuan dari perspektif bisnis. Selanjutnya, dilakukan proses penerjemahan informasi tersebut ke dalam bentuk perumusan masalah dalam data mining. Setelah itu, ditetapkan strategi serta perencanaan untuk mencapai hasil yang diinginkan dalam analisis data (Msy Aulia Hasanah, 2023).

1. *Data Understanding* (Pemahaman Data)

Pada tahap ini, proses dimulai dengan pengumpulan data, kemudian dilakukan deskripsi terhadap karakteristik data tersebut, serta evaluasi kualitasnya untuk memastikan kesesuaiannya dalam analisis lebih lanjut. Penelitian sebelumnya menyebutkan pentingnya pemahaman awal data sebagai dasar untuk menentukan parameter kualitas udara yang relevan seperti PM10, CO, dan NO2 (Belia Putri Salsabila., 2023; JSON., 2023).

1. *Data Preparation* (Persiapan Data)

Tahap ini bertujuan untuk menyusun dataset akhir dari data mentah. Proses yang dilakukan meliputi pembersihan data (*Data Cleaning*), pemilihan data (*Data Selection*) berdasarkan *record* dan atribut yang relevan, serta transformasi data (*Data Transformation*) agar sesuai untuk digunakan dalam pemodelan. Langkah ini penting dalam mencegah bias dan ketidaktepatan hasil analisis (Ni Wayan Wardani., 2023).

1. *Modeling* (Pemodelan dengan K-Means Clustering)

Sebelum melakukan klastering, langkah pertama adalah menentukan jumlah klaster yang optimal. Ini dilakukan dengan *Elbow Method*, di mana jumlah klaster dipetakan terhadap nilai *Within-Cluster Sum of Squares (WCSS)*. Titik optimal adalah saat grafik mulai melandai. Selain itu, *Silhouette Score* juga digunakan untuk mengevaluasi sejauh mana klaster yang terbentuk saling terpisah dan memiliki kepadatan yang baik. Biasanya, nilai K yang optimal berada dalam rentang 3 hingga 5, tergantung pada distribusi data (Adinda Inez Sang., 2023; Adityo Nugroho., 2023). Setiap titik data dihitung jaraknya terhadap setiap centroid menggunakan rumus *Euclidean Distance* berikut :

Dimana :

= Data ke-

= *Centroid* dari klaster ke-

= Jumlah fitur dalam dataset (misalnya PM 2.5, PM 10, CO, NO2)

Setelah data memiliki klaster, kita perlu memperbarui posisi *centroid* berdasarkan rata-rata semua titik dalam klaster tersebut :

Dimana :

=  *Centroid* baru dari klaster ke-

= Jumlah titik data dalam klaster ke-

= Semua titik data yang termasuk dalam klaster ke-

1. *Evaluation* (Evaluasi Data)

Sebelum melakukan klastering, langkah pertama adalah menentukan jumlah klaster yang optimal. Ini dilakukan dengan *Elbow Method*, di mana jumlah klaster dipetakan terhadap nilai *Within-Cluster Sum of Squares (WCSS)*. Titik optimal ditentukan pada saat grafik mulai melandai, yang menunjukkan bahwa penambahan jumlah klaster tidak lagi memberikan pengurangan signifikan terhadap WCSS. Evaluasi ini membantu memastikan pemilihan jumlah klaster yang efektif dalam membentuk pemodelan yang representatif (Adinda Inez Sang., 2023).

Dimana :

= Klaster ke-

= Centroid dari klaster ke-

= Data point dalam klaster

= J arak Euclidean kuadrat antara dan centroid

## HASIL

Pada bagian ini, diberikan dan dijelaskan hasil penelitian sesuai dengan metode penelitian yang telah disebutkan sebelumnya. Bagian ini menjelaskan hasil identifikasi pola kualitas udara di Jakarta dengan menggunakan metode K-Means *Clustering* dalam konteks *data mining*.

## *a. Business Understanding* (Pemahaman Data)

Dalam upaya memahami kondisi pencemaran udara di Jakarta, penelitian ini bertujuan untuk mengelompokkan stasiun pemantauan berdasarkan pola kualitas udara menggunakan metode *clustering* K-Means. Dengan menganalisis data ISPU dari tahun 2010 hingga 2023 tanpa kategori bawaan, penelitian ini berfokus pada identifikasi kelompok stasiun yang memiliki karakteristik pencemaran udara serupa. Hasil pengelompokan ini diharapkan dapat memberikan wawasan baru bagi pengambil kebijakan dan pihak terkait dalam merumuskan strategi pengendalian polusi udara secara lebih terarah dan berbasis data.

## *b. Data Understanding* (Pemahaman Data)

Penelitian ini menggunakan data publik mengenai pencemaran udara berdasarkan *Air Quality Index* (AQI) atau Indeks Standar Pencemar Udara (ISPU) yang diukur dari 5 Stasiun Pemantauan Kualitas Udara (SPKU) di Jakarta. Dataset yang digunakan diperoleh dari website datasets bernama Kaggle dan dapat diakses melalui tautan: https://www.kaggle.com/datasets/senadu34/air-quality-index-in-jakarta-2010-2021/data. Data ini mencakup periode tahun 2010 hingga 2023 dan digunakan untuk menganalisis tren kualitas udara di Jakarta. Dalam dataset, setiap stasiun berperan sebagai identifier yang membedakan setiap baris data, sehingga memungkinkan pelacakan dan analisis kualitas udara secara spesifik berdasarkan lokasi pemantauan.

**Tabel 1. Data Indeks Standar Pencemaran Udara DKI Jakarta Tahun 2010-2023**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No** | **Tanggal** | **Stasiun** | **pm10** | **pm25** | **so2** | **co** | **o3** | **no2** | ***max*** | ***critical*** |
| 1 | 2010-01-01 | DKI1 (Bunderan HI) | 60 | - | 4 | 73 | 27 | 14 | 73 | CO |
| 2 | 2010-01-02 | DKI1 (Bunderan HI) | 32 | - | 2 | 16 | 33 | 9 | 33 | O3 |
| 3 | 2010-01-03 | DKI1 (Bunderan HI) | 27 | - | 2 | 19 | 20 | 9 | 27 | PM10 |
| 4 | 2010-01-04 | DKI1 (Bunderan HI) | 22 | - | 2 | 16 | 15 | 6 | 22 | PM10 |
| 5 | 2010-01-05 | DKI1 (Bunderan HI) | 25 | - | 2 | 17 | 15 | 8 | 25 | PM10 |
| … | … | … | … | … | … | … | … | … | … | … |
| 4443 | 2023-11-29 | DKI4 (Lubang Buaya) | 71 | 105 | 30 | 19 | 22 | 14 | 105 | PM25 |
| 4444 | 2023-11-30 | DK1 (Bunderan HI) | 38 | 67 | 43 | 12 | 34 | 34 | 67 | PM25 |

Dengan rincian atribut data sebagai berikut:

1. tanggal : Tanggal pengukuran kualitas udara.
2. stasiun : Lokasi pengukuran di beberapa stasiun. DKI1 (Bunderan HI), DKI2 (Kelapa Gading), DKI3 (Jagakarsa), DKI4 (Lubang Buaya), dan DKI5 (Kebon Jeruk).
3. pm10 : Partikulat salah satu parameter yang diukur.
4. pm25 : Partikulat salah satu parameter yang diukur.
5. so2 : Sulfida salah satu parameter yang diukur.
6. co : Karbon Monoksida salah satu parameter yang diukur.
7. o3 : Ozon salah satu parameter yang diukur.
8. no2 : Nitrogen dioksida salah satu parameter yang diukur.
9. *max* : Nilai tertinggi dari seluruh parameter.
10. *critical* : Parameter yang hasil pengukurannya paling tinggi.

## *c. Data Preparation* (Persiapan Data)

Tahap data preparation merupakan langkah penting sebelum dilakukan proses pemodelan. Pada tahap ini, dilakukan beberapa proses utama yaitu pembersihan data (*data cleaning*), seleksi atribut, dan transformasi data. Tujuannya adalah untuk memastikan bahwa data yang digunakan bersih, relevan, dan siap untuk diolah secara analitik.

1. *Data Cleaning*

Total dataset dari tahun 2010-2023 berjumlah 4626 data, lalu setelah dilakukan *data cleaning* dengan menghapus data kosong (*missing value*) maka didapat data menjadi 4444 data. Kemudian beberapa atribut yang dihapus adalah atribut tanggal, pm25, *max,* dan *critical.*

1. *Attribute Selection*

Attribut lain yang tidak diperlukan untuk *modelling* nantinya seperti atribut *max* dan *critical* dihapus karena atribut tersebut merupakan nilai maksimum dari parameter lain. Atribut tanggal tidak diperlukan untuk *modelling.* Atribut pm25 juga dihapus karena banyak sekali terdeteksi *missing value* dan bisa diwakilkan dengan atribut partikulat lainnya yaitu pm10.

1. *Transformation Data*

Data numerik kemudian ditransformasi menggunakan metode *Z-transformation* (*standardization*). Proses ini mengubah nilai-nilai data menjadi skala standar dengan rata-rata nol dan simpangan baku satu, guna meningkatkan kinerja model dalam mendeteksi pola. Selain itu, atribut stasiun yang semula berbentuk label teks (DKI1 hingga DKI5) diubah menjadi angka integer mulai dari 0 hingga 4.

**Tabel 2. Hasil *Data Preparation***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No** | **stasiun** | **pm10** | **so2** | **co** | **o3** | **no2** |
| 1 | 0 | -0.269 | -1.716 | 3.098 | -1.249 | -0.495 |
| 2 | 0 | -1.725 | -1.860 | -1.070 | -1.128 | -1.079 |
| 3 | 0 | -1.986 | -1.860 | -0.850 | -1.391 | -1.079 |
| 4 | 0 | -2.246 | -1.860 | -1.070 | -1.492 | -1.430 |
| 5 | 0 | -2.090 | -1.860 | -0.996 | -1.492 | -1.196 |
| … | … | … | … | … | … | … |
| 4443 | 3 | 0.303 | 0.143 | -0.850 | -1.350 | -0.495 |
| 4444 | 0 | -1.413 | 1.073 | -1.362 | -1.108 | 1.842 |

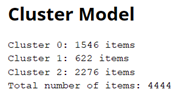
Tabel 22 di atas memperlihatkan hasil akhir setelah proses *data preparation* dilakukan. Tampak bahwa data telah dibersihkan dari *missing value*, atribut tidak relevan telah dihapus, data numerik telah dinormalisasi, dan label stasiun telah dikonversi ke bentuk numerik. Data hasil *data preparation* ini nantinya yang akan digunakan untuk modelling dengan menggunakan algoritma K-Means.

## *d. Modelling* (Pemodelan)

Setelah dilakukan tahap *preparation data*, tahap selanjutnya adalah melakukan *modelling* (pemodelan) *clustering* dengan menggunakan algoritma K-Means. Melakukan *clustering* dengan menggunakan algoritma K-Means memiliki beberapa tahapan diantaranya :

1. Menentukan jumlah *cluster* yang akan digunakan. Jumlah *cluster* yang digunakan untuk *clustering* ini adalah K=2 Sampai K=7.
2. Menentukan centroid awal dengan menginisialisasi tiga centroid secara acak.
3. Menghitung jarak antar data dan centroid menggunakan *euclidean distance* untuk mengukur kesamaan antar data dengan centroid yang terdekat.
4. Proses iterasi dilakukan hingga centroid tidak lagi berubah signifikan atau hingga batas maksimal iterasi sebanyak 100 langkah tercapai (*max optimization steps* = 100). Untuk memastikan hasil yang stabil, algoritma dijalankan sebanyak 10 kali dengan inisialisasi centroid yang berbeda-beda pada setiap percobaan (*max runs* = 10).

Setelah dilakukan proses pemodelan menggunakan algoritma K-Means dengan memanfaatkan perangkat lunak *rapidminer*, hasil clustering menunjukkan bahwa data terbagi ke dalam tiga cluster yang memiliki karakteristik berbeda. *cluster* 0 terdiri dari 1546 item, *cluster* 1 terdiri dari 622 item, dan *cluster* 2 terdiri dari 2276 item. Total keseluruhan data yang dikelompokkan berjumlah 4444 item yang dapat dilihat pada gambar .



**Gambar 2. Hasil Pemodelan *Clustering* K-Means*.***

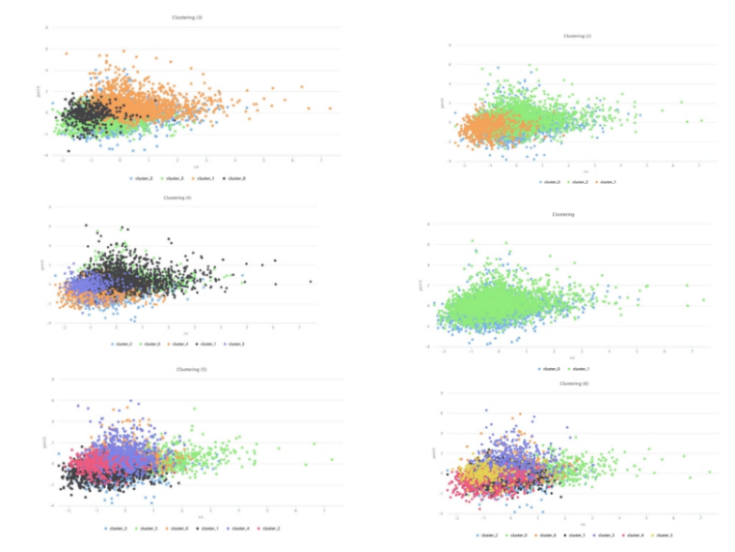
Kemudian untuk memahami lebih lanjut karakteristik masing-masing *cluster*, dapat dilakukan analisis terhadap tabel centroid yang berisi nilai rata-rata dari setiap atribut pada masing-masing *cluster* yang terbentuk setelah proses *clustering* menggunakan algoritma K-Means sebagai berikut:

**Tabel 3. Tabel Centroid**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Cluster*** | **stasiun** | **pm10** | **so2** | **co** | **o3** | **no2** |
| 0 | 0.961 | -0.327 | -0.455 | 0.146 | -0.159 | -0.233 |
| 1 | 2.437 | -0.474 | 1.626 | -1.014 | -0.818 | 1.006 |
| 2 | 3.204 | 0.352 | -0.136 | 0.178 | 0.331 | -0.117 |

Berdasarkan tabel 3 *centroid* tersebut maka dapat diketahui bahwa *Cluster* 0 memiliki nilai yang cenderung lebih rendah pada atribut-atribut polutan seperti PM10, SO2, dan NO2, sehingga dapat dikategorikan sebagai kelompok dengan kualitas udara yang lebih sehat (kategori sehat). Berdasarkan tabel 3, *centroid* tersebut maka dapat diketahui bahwa *Cluster* 0 memiliki nilai yang cenderung lebih rendah pada atribut-atribut polutan seperti PM10, SO2, dan NO2, sehingga dapat dikategorikan sebagai kelompok dengan kualitas udara yang lebih sehat (kategori sehat). *Cluster* 1 menunjukkan nilai yang lebih tinggi pada atribut SO2 dan NO2 dibandingkan *cluster* lainnya, mengindikasikan bahwa kelompok ini memiliki tingkat polusi udara yang lebih tinggi dan berpotensi berisiko terhadap kesehatan (kategori tidak sehat). *Cluster* 2 memiliki nilai yang bervariasi dengan beberapa atribut polutan yang lebih tinggi dibanding *Cluster* 0, namun lebih rendah dibanding *Cluster* 1. Hal ini menunjukkan bahwa kelompok ini memiliki kualitas udara yang sedang atau berada di antara kategori sehat dan berpolusi (kategori sedang).

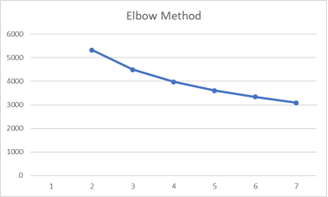
Untuk mengamati pola penyebaran data serta mengetahui kecenderungan pembentukan kelompok berdasarkan parameter pencemar udara, dilakukan proses clustering awal menggunakan kombinasi atribut pm10 dan co. Visualisasi hasil clustering tersebut dengan jumlah cluster yang divariasikan dengan K=2 hingga K=7 ditampilkan pada Gambar 3.



**Gambar 3. Clustering data berdasarkan pm10 dan co**

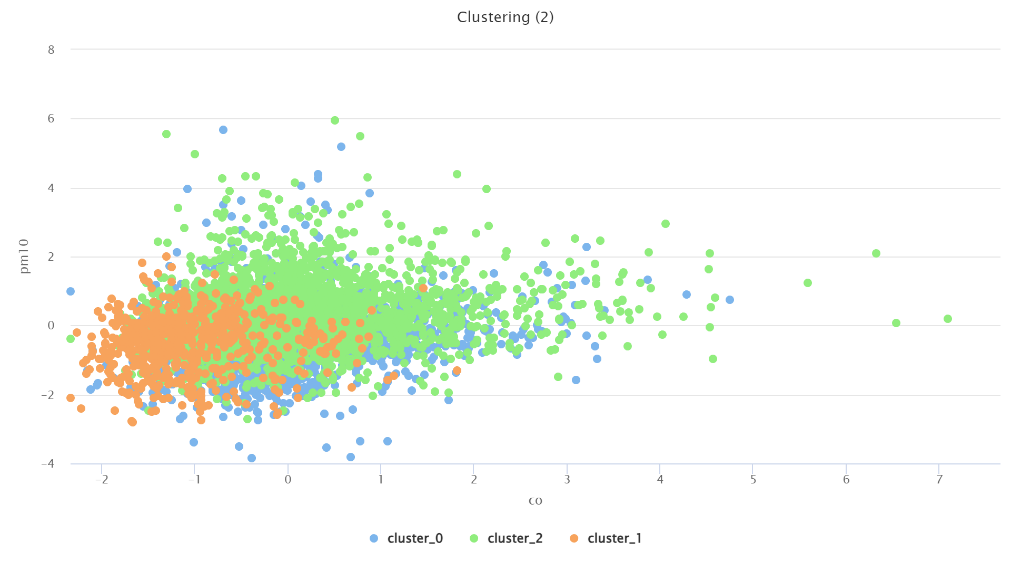
## *e. Evaluation* (Evaluasi)

Pada tahap evaluasi ini dilakukan penentuan jumlah *cluster* terbaik, dapat menggunakan salah satu metode yaitu metode Elbow (*Elbow method).* Dengan mengamati persentase hasil yang diperoleh saat membandingkan jumlah *cluster* yang membentuk siku. Kemudian dilakukan percobaan untuk menentukan jumlah *cluster* yang cocok dengan menjalankan algoritma K-Means dengan *cluster* K=2 hingga K=7 dan memanfaatkan rata-rata *Cluster Distance Perfomance* untuk membandingkan nilainya.



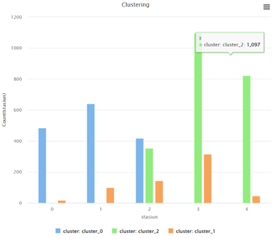
**Gambar 4. *Elbow Method***

Berdasarkan gambar 4 dengan mengamati persentase hasil yang diperoleh saat membandingkan jumlah *cluster* yang membentuk siku, *cluster* bernomor 3 menunjukkan nilai selisih tertinggi dan sudut siku tertajam pada grafik, sehingga dapat disimpulkan *cluster* K=3 yang merupakan jumlah *cluster* terbaik.Berdasarkan gambar 4, dengan mengamati persentase hasil yang diperoleh saat membandingkan jumlah *cluster* yang membentuk siku, *cluster* bernomor 3 menunjukkan nilai selisih tertinggi dan sudut siku tertajam pada grafik, sehingga dapat disimpulkan *cluster* K=3 yang merupakan jumlah *cluster* terbaik.



**Gambar 5. Visualisasi penyebaran data dengan K=3**

Berdasarkan metode Elbow yang sudah dilakukan, gambar 5 di atas merupakan visualisasi penyebaran data cluster menggunakan K=3. Pemilihan nilai K=3 ini didasarkan pada titik siku (elbow point) yang terlihat jelas pada grafik metode Elbow, yang menunjukkan bahwa tiga cluster merupakan jumlah yang optimal untuk mengelompokkan data secara efektif.



**Gambar 6. Visualisasi Distribusi Stasiun Berdasarkan Hasil *Clustering* K-Means**

Gambar 6 di atas menunjukkan distribusi stasiun pada masing-masing *cluster* hasil dari proses pemodelan menggunakan algoritma K-Means di perangkat lunak *rapidminer*. Stasiun 0 adalah stasiun DKI1 (Bunderan HI), stasiun 1 adalah stasiun DKI2 (Kelapa Gading), stasiun 2 adalah stasiun DKI3 (Jagakarsa), stasiun 3 adalah stasiun DKI4 (Lubang Buaya), dan stasiun 4 adalah stasiun DKI5 (Kebon Jeruk). Berdasarkan visualisasi gambar 3 menunjukkan bahwa:

1. Stasiun DKI1 (Bunderan HI) termasuk ke dalam kategori sehat.
2. Stasiun DKI2 (Kelapa Gading) termasuk ke dalam kategori sehat.
3. Stasiun DKI3 (Jagakarsa) termasuk ke dalam kategori sedang.
4. Stasiun DKI4 (Lubang Buaya) termasuk ke dalam kategori sedang.
5. Stasiun DKI5 (Kebon Jeruk)termasuk ke dalam kategori sedang.

## PEMBAHASAN

#### Penelitian ini bertujuan untuk mengelompokkan kualitas udara di wilayah DKI Jakarta berdasarkan parameter polusi udara menggunakan algoritma K-Means *Clustering*. Data yang digunakan merupakan data sekunder yang diperoleh dari situs resmi jakartasatu.jakarta.go.id, mencakup rentang waktu tahun 2010 hingga 2023. Parameter yang dianalisis meliputi PM₁₀, SO₂, CO, O₃, dan NO₂, yang merupakan indikator utama dalam penilaian kualitas udara menurut standar nasional dan internasional.

#### Tahapan awal dalam penelitian ini adalah melakukan praproses data (*data preprocessing*) untuk memastikan data bersih, konsisten, dan dapat diolah oleh algoritma. Proses ini meliputi penghapusan data kosong (*missing values*), normalisasi data agar memiliki skala yang seragam, serta pemilihan atribut yang relevan untuk proses klasterisasi.

#### Setelah data siap, dilakukan proses klasterisasi menggunakan algoritma K-Means *Clustering*. Penentuan jumlah klaster dilakukan dengan menggunakan metode *Elbow*, dan diperoleh bahwa jumlah klaster optimal adalah tiga (K=3). Klaster tersebut merepresentasikan tiga kategori kualitas udara, yaitu:

#### *Cluster* 0: Kualitas udara kategori buruk, dengan konsentrasi polutan yang tinggi pada semua parameter.

#### *Cluster* 1: Kualitas udara kategori sedang, dengan nilai parameter yang bervariasi namun tidak melebihi ambang batas berbahaya.

#### *Cluster* 2: Kualitas udara kategori baik, dengan konsentrasi polutan yang relatif rendah.

#### Distribusi data pada masing-masing klaster menunjukkan bahwa sebagian besar tahun dengan kualitas udara yang lebih baik berada di klaster 2, sementara beberapa tahun dengan nilai polusi tinggi, khususnya pada parameter PM₁₀ dan NO₂, tergolong ke dalam klaster 0. Hal ini menunjukkan bahwa meskipun terdapat peningkatan dalam pengelolaan lingkungan, kualitas udara di DKI Jakarta masih memerlukan perhatian, terutama pada tahun-tahun dengan aktivitas industri dan transportasi yang tinggi.

## KESIMPULAN

Dari hasil penelitian yang telah dilakukan, maka diperoleh kesimpulan sebagai berikut:

1. Berdasarkan metode *Elbow* diperoleh bahwa jumlah *cluster* terbaik adalah K=3 untuk pengelompokkan data kualitas udara menggunakan algoritma K-Means dengan rata-rata *Cluster Distance Performance*.
2. *Cluster* 0 terdiri dari wilayah DKI1 dan DKI2. Pengelompokan ini menunjukkan kualitas udara yang relatif lebih baik (sehat).
3. *Cluster* 1 terdiri dari wilayah DKI3, DKI4, dan DKI5. Pengelompokkan ini menunjukkan tingkat pencemaran udara yang lebih tinggi (tidak sehat).
4. *Cluster* 2 mencerminkan data dengan kategori kualitas udara sedang, namun tidak dominan dibandingkan dua cluster lainnya.
5. Hasil pengelompokan memberikan gambaran yang sistematis dan objektif mengenai kondisi kualitas udara di Jakarta. Informasi ini dapat dimanfaatkan oleh pembuat kebijakan sebagai dasar dalam merumuskan strategi pengendalian polusi udara yang lebih tepat sasaran.
6. Berdasarkan analisis, menunjukkan bahwa efektivitas metode K-Means *Clustering* sangat bergantung pada kelengkapan dan konsistensi data ISPU.

## BATASAN

Hasil implementasi algoritma K-Means *Clustering* dalam penelitian ini memberikan gambaran pengelompokan kualitas udara yang cukup baik, namun tetap memiliki beberapa keterbatasan yang perlu diperhatikan agar hasil analisis dapat diinterpretasikan secara tepat dan objektif. Beberapa batasan dalam penelitian ini antara lain:

1. Validasi hasil klaster belum melibatkan metrik evaluasi kuantitatif seperti Silhouette Coefficient atau Davies-Bouldin Index, sehingga penilaian kualitas klaster masih bergantung pada visualisasi dan interpretasi subjektif. Selain itu, penentuan jumlah klaster hanya mengandalkan metode Elbow, yang bersifat visual dan tidak selalu menjamin jumlah klaster yang optimal secara statistik maupun relevan dengan konteks domain. Keterbatasan ini menunjukkan perlunya pendekatan evaluasi dan penentuan jumlah klaster yang lebih komprehensif pada penelitian selanjutnya.
2. Keterbatasan Parameter Polusi: Penelitian hanya menggunakan lima parameter utama (PM₁₀, SO₂, CO, O₃, NO₂) yang tersedia pada sumber data. Parameter lain seperti PM₂.₅ atau nilai indeks kualitas udara (AQI) secara komprehensif tidak disertakan karena keterbatasan data historis.
3. Keterbatasan Dinamika Musiman: Analisis tidak memperhitungkan fluktuasi musiman atau perbedaan waktu (seperti musim kemarau dan hujan), yang dapat memengaruhi nilai polusi udara secara signifikan.

## DAFTAR PUSAKA

###### Adinda, R., et al. 2021. Penerapan data mining untuk clustering kualitas udara. Jurnal Teknologi 8 (2): 123–130. <https://journal.unipdu.ac.id/index.php/teknologi/article/view/4088>

###### Mahendrasyah, A., et al. 2024. Analisis pencemaran udara di Jakarta. Jurnal Sains dan Teknologi 12 (1): 45–56. <https://openjournal.unpam.ac.id/index.php/SNISIS/article/view/41824>

###### Adityo, R., et al. 2023. Penerapan algoritma K-Means untuk analisis kualitas udara. ResearchGate. <https://www.researchgate.net/publication/382630710_Penerapan_Data_Mining_Untuk_Clustering_Kualitas_Udara>

###### Pertiwi, D., et al. 2024. Dampak pencemaran udara terhadap kesehatan. Jurnal Ilmiah 10 (3): 78–85. <https://e-journal.stmiklombok.ac.id/index.php/jire/article/view/343>

###### Salsabila, N., et al. 2024. Analisis tren kualitas udara menggunakan K-Means clustering. International Journal of Computer Science and Information Technology 5 (2): 99–110. <https://iiast.iaic-publisher.org/ijcitsm/index.php/IJCITSM/article/view/122>

###### Mahendrasyah, I., Diana, A., Rusdah, & Mahdiana, D. (2024). Penerapan algoritma K-Means untuk klasterisasi Indeks Standar Pencemaran Udara. TEKNOLOGI: Jurnal Ilmiah Sistem Informasi, 14(2), 146–156. https://doi.org/10.26594/teknologi.v14i2.4088

###### Dhewayani, N., Amelia, D., Alifah, D. N., Sari, B. N., & Jajuli, M. (2020). Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM. JATI (Jurnal Teknik Informatika), 8(2), 97–104. https://ojs.unikom.ac.id/index.php/jati/article/view/6674/2967

###### Kusrini, & Luthfi, E. (2022). Algoritma data mining. Andi. Retrieved from <https://books.google.co.id/books?id=URZSEQAAQBAJ>

###### Yasa, I. M. D. N., & Ardana, I. M. (2022). Penerapan data mining untuk klasifikasi data menggunakan algoritma K-Means. Jurnal Ilmiah INSERT (Information System Research), 3(1), 48–54. <https://ejournal.undiksha.ac.id/index.php/insert/article/view/41673>

###### Prasetyo, F. H., & Risqi, A. H. (2020). Analisis kualitas udara menggunakan metode K-Means Clustering. Jurnal Komputer dan Informatika (JAKAKOM), 1(2), 10–15. https://ejournal.unama.ac.id/index.php/jakakom/article/view/34/58

###### Sihombing, A. M., & Sihombing, R. (2023). Implementasi algoritma K-Means menggunakan RapidMiner untuk klasterisasi data obat pada Rumah Sakit Royal Prima. Jurnal Teknologi Informasi dan Komputer (JUTIKOMP), 4(1), 45–52. Retrieved from https://jurnal.unprimdn.ac.id/index.php/JUTIKOMP/article/view/5537/3524

###### Iriani, A., & Rahmawati, R. (2023). Analisis Kualitas Udara Menggunakan Metode K-Means Clustering. KLIK - Kajian Ilmiah Informatika dan Komputer, 10(1), 17–22. https://djournals.com/klik/article/view/1525/877

###### Sihombing, A. M., & Sihombing, R. (2023). Implementasi algoritma K-Means menggunakan RapidMiner untuk klasterisasi data obat pada Rumah Sakit Royal Prima. Jurnal Teknologi Informasi dan Komputer (JUTIKOMP), 4(1), 45–52. Retrieved from https://jurnal.unprimdn.ac.id/index.php/JUTIKOMP/article/view/5537/3524

###### Wardani, N. W., and D. Darmansah. 2021. Analisis penyebaran penularan virus corona di Provinsi Jawa Tengah menggunakan metode K-Means clustering. JATISI (Jurnal Teknik Informatika dan Sistem Informasi) 8 (1): 105–117. https://jatisi.polibatam.ac.id/index.php/jatisi/article/view/6498

###### Widuri, A. 2023. Analisis kualitas udara dengan data mining. Jurnal Infoteh 9 (2): 56–62. http://www.jurnal.kampuswiduri.ac.id/index.php/infoteh/article/view/164/134

###### Mawaddah Anjelita. 2023. Penerapan algoritma K-Means dalam analisis kualitas udara. Prosiding Seminar Nasional 4 (1): 101–110. https://prosiding.seminar-id.com/index.php/sainteks/article/view/453

###### Adinda Amalia. 2023. Pengaruh polusi udara terhadap kesehatan masyarakat. Jurnal Pendidikan dan Ilmu Pengetahuan 11 (1): 22–30. https://www.jurnal.stkippgritulungagung.ac.id/index.php/jipi/article/view/2843/1185

###### Kartika Dian Pertiwi. 2023. Kualitas udara dan kesehatan. Jurnal Penelitian 9 (1): 50–60. https://jurnal.unw.ac.id/index.php/PJ/article/view/3351

###### Msy Aulia Hasanah. 2023. Analisis kualitas udara menggunakan K-Means. Jurnal Algoritma dan Informatika 5 (1): 30–40. https://jurnal.polibatam.ac.id/index.php/JAIC/article/download/3200/1497/

###### Belia Putri Salsabila. 2023. Penerapan K-Means clustering untuk kualitas udara. Jurnal Teknik Informatika 15 (1): 34–40. https://ejournal.itn.ac.id/index.php/jati/article/view/11592

###### JSON. 2023. Kualitas udara dan dampaknya. Jurnal Sains dan Teknologi 7 (2): 45–50. https://www.ejurnal.stmik-budidarma.ac.id/index.php/JSON/article/view/7145/

###### Ni Wayan Wardani. 2023. Penerapan data mining untuk kualitas udara. Jurnal Teknik Informatika 8 (1): 15–25. https://jurnal.undhirabali.ac.id/index.php/jutik/article/view/2081/pdf

###### Adinda Inez Sang. 2023. Analisis kualitas udara di Jakarta. Engineering Journal 6 (1): 88–95. https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15900

###### Adityo Nugroho. 2023. Model kualitas udara berbasis data. Engineering Journal 6 (2): 112–120. https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20030/0