DOI: https://doi.org/10.62017/tektonik

Penerapan K – Means Clustering dalam Analisis Kualitas Udara Jakarta

Birrham Efendi Lubis *[1]
Muhammad Zidan Fadillah [2]

[1,2] Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa, Indonesia
*e-mail: birrhamefendilubis@mhs.pelitabangsa.ac.id[1], muhammadzidanfadillah@mhs.pelitabangsa.ac.id[2]

Abstrak

Pencemaran udara adalah isu lingkungan yang serius yang berdampak pada kesehatan manusia secara global, terutama di kota-kota besar padat penduduk dan industri seperti Jakarta. Kenaikan tingkat polutan udara mengakibatkan berbagai penyakit pernapasan dan isu kesehatan lainnya, sehingga menjadi penyebab morbiditas dan mortalitas yang penting. Untuk mendukung pemantauan dan evaluasi yang efisien, penelitian ini menggunakan metode pengelompokan K-Means untuk menganalisis serta mengkategorikan data kualitas udara di Jakarta. Data kualitas udara, mencakup parameter seperti partikel materi (PM), karbon monoksida (CO), ozon (O_3), nitrogen dioksida (NO_2), dan sulfur dioksida (SO_2), diperoleh dari stasiun pengawas di Jakarta. Proses clustering K-Means digunakan untuk menganalisis pola sebaran polutan dan mengategorikan wilayah berdasarkan tingkat polusi. Proses yang dilakukan meliputi pra-pemrosesan data, penentuan jumlah kluster optimal dengan metode elbow, dan validasi kluster menggunakan Davies-Bouldin Index (DBI). Pelaksanaan analisis dengan pemrograman Python, mencakup standardisasi data serta penerapan Principal Component Analysis (PCA) untuk tujuan visualisasi. Analisis menunjukkan bahwa metode K-Means efektif dalam mengklasifikasikan data kualitas udara ke dalam kelompok-kelompok yang mencerminkan tingkat polusi yang berbeda, memberikan informasi berharga untuk perencanaan kota dan kebijakan pengurangan dampak.

***Kata kunci**: K -Means, Klastering, Kualitas Udara, Jakarta*

Abstract

Air pollution is a serious environmental issue that impacts human health globally, especially in densely populated and industrialized cities such as Jakarta. The increasing levels of air pollutants result in various respiratory diseases and other health issues, thus becoming an important cause of morbidity and mortality. To support efficient monitoring and evaluation, this study uses the K-Means clustering method to analyze and categorize air quality data in Jakarta. Air quality data, including parameters such as particulate matter (PM), carbon monoxide (CO), ozone (O_3), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2), were obtained from monitoring stations in Jakarta. The K-Means clustering process was used to analyze the distribution patterns of pollutants and categorize areas based on pollution levels. The processes carried out include data pre-processing, determining the optimal number of clusters using the elbow method, and cluster validation using the Davies-Bouldin Index (DBI). The analysis was carried out using Python programming, including data standardization and the application of Principal Component Analysis (PCA) for visualization purposes. The analysis shows that the K-Means method is effective in classifying air quality data into groups reflecting different pollution levels, providing valuable information for urban planning and impact reduction policies.

***Keywords**: K – Means, Clustering, Air Quality, Jakarta*

PENDAHULUAN

Salah satu masalah lingkungan paling mendesak di dunia saat ini adalah pencemaran udara, yang membahayakan masyarakat dan ekosistem secara langsung maupun tidak langsung(Aiello et al., 2025). Masalah ini semakin parah di Indonesia, terutama di kota-kota besar seperti Jakarta karena kepadatan penduduk yang meningkat dan aktivitas yang disebabkan oleh manusia, seperti industri dan peningkatan jumlah kendaraan bermotor(Alim et al., 2024). Transportasi menyumbang sekitar 60% polusi udara di dalam kota(Nanda et al., 2023). Polusi udara memiliki banyak efek negatif, mulai dari iritasi pada mata dan tenggorokan hingga penyakit pernapasan kronis dan kanker dalam jangka panjang. Menurut Organisasi Kesehatan Dunia (WHO), hampir setiap orang di seluruh dunia menghirup udara di atas batas pedoman WHO. Negara-negara berpenghasilan rendah dan menengah adalah yang paling terpapar.

Sangat penting untuk memantau kualitas udara secara terus menerus dan menganalisis data yang dihasilkan secara statistik. Data kualitas udara disimpan di berbagai lokasi dan waktu

DOI: https://doi.org/10.62017/tektonik

secara inheren bersifat spatiotemporal(Aiello et al., 2025) . Faktor utama yang memengaruhi kesehatan masyarakat adalah, *particulate matter* (PM) , karbon monoksida (CO) , ozon (O_3) , nitrogen dioksida (SO_2)(Choi et al., 2024). Di Malaysia, parameter pemantauan kualitas udara yang paling penting adalah PM (Rahman et al., 2022). Partikel dengan diameter kurang dari 2.5 µm yang dikenal sebagai PM2.5 sangat penting karena kemampuannya melewati paru-paru dan berdampak pada kesehatan manusia.

K-Means adalah salah satu algoritma clustering yang paling populer dan banyak digunakan (Mohd Aftar Abu Bakar et al., 2022). Keunggulannya termasuk kemudahan penggunaan dan kemampuan untuk memaksimalkan variasi antar-cluster dan intra-cluster. Metode ini juga dapat diterapkan pada dataset besar (Alim et al,. 2023).

Tujuan dari studi ini adalah untuk menggunakan algoritma clustering K-Means untuk menganalisis data kualitas udara di Jakarta. Tujuannya adalah untuk menemukan pola pencemaran dan mengklasifikasikan area menurut tingkat polutannya. Pemerintah daerah dapat dengan lebih mudah membuat kebijakan dan mencegah efek buruk polusi udara dengan memahami pola-pola ini (Annas et al., 2022).

METODE

Studi ini menggunakan pendekatan kuantitatif deskriptif dengan memanfaatkan metode Unsupervised Learning, khususnya algoritma K-Means Clustering. Metode ini memungkinkan pengelompokan data kualitas udara tanpa memerlukan data yang telah diberi label sebelumnya (Alim et al., 2023).

Data

Data kualitas udara yang digunakan dalam studi ini adalah data sekunder, yang diasumsikan berasal dari *Kaggle* yang memuat data Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta. Parameter kualitas udara yang dianalisis mencakup PM2.5, PM10, SO_2, CO, O_3, dan NO_2. Data ini dianggap tersedia dalam format harian, mencakup periode waktu yang relevan untuk menganalisis pola musiman dan tren polusi. Agar clustering berjalan dengan efektif, data bisa juga mencakup informasi temporal seperti tanggal dan nama stasiun.

Tahap Penelitian

Proses penelitian ini mengikuti langkah-langkah yang teratur:

- **Identifikasi Masalah**

  Tahap pertama ini mencakup pengertian mendalam mengenai permasalahan kualitas udara di Indonesia, terutama di Jakarta, serta pengaruhnya terhadap kesehatan masyarakat, termasuk peningkatan kasus ISPA (Alim et al., 2023).

- **Studi Literatur**

  Dilakukan pencarian literatur untuk mengumpulkan teori, metode, dan penelitian yang berkaitan dengan *Clustering* kualitas udara, serta algoritma *K-Means* beserta validasinya.

- **Pra – Pemrosesan Data (Normalisasi)**

  Data mentah dari stasiun pengamatan harus diproses terlebih dahulu. Hal ini mencakup penanganan nilai yang hilang yang dapat diselesaikan dengan menghapus baris yang memiliki nilai NaN pada fitur numerik. Selanjutnya, data dinormalisasi menggunakan StandardScaler agar berada pada rentang yang lebih sempit dan memastikan semua fitur memiliki skala yang konsisten, yang sangat penting untuk algoritma berbasis jarak seperti K-Means.

DOI: https://doi.org/10.62017/tektonik

- **Penentuan Jumlah K Optimal (*ELBOW METHOD)***

  Agar dapat menerapkan algoritma K-Means, penting untuk menetapkan jumlah cluster (K) yang paling sesuai. Dalam studi ini, nilai K diandaikan 3 berdasar hasil analisis awal (seperti yang diindikasikan oleh kode) serta kesimpulan dari sumber yang sejenis.(Alim et al., 2023)

- **Clustering K – Means dengan *Ecludean Distance***

  Setelah nilai K ditentukan, algoritma K-Means diterapkan pada data yang sudah dinormalisasi. Jarak Euclidean akan dipakai sebagai ukuran kesamaan untuk mengelompokkan setiap titik data ke dalam cluster yang paling dekat (Choi et al., 2023). Proses ini akan terus diulang sampai centroid cluster tidak mengalami perubahan signifikan lagi.

- **Visualisasi Hasil dengan *Principal Component Analysis* (PCA)**

  Untuk menggambarkan hasil clustering yang multidimensional, analisis komponen utama (PCA) diterapkan untuk mengurangi dimensi data menjadi 2 komponen utama (PCA1 dan PCA2). Hal ini memungkinkan visualisasi cluster dalam ruang 2D yang gampang dipahami (Choi et al., 2024).

- **Analisis Hasil**

  Analisis clustering akan dilakukan untuk mengenali pola-pola pencemaran udara, ciri-ciri setiap cluster, dan dampaknya terhadap keadaan kualitas udara di Jakarta. Visualisasi tambahan seperti boxplot untuk distribusi polutan per kluster dan heatmap untuk rata-rata nilai polutan per kluster akan diterapkan untuk analisis yang lebih mendetail.

HASIL DAN PEMBAHASAN

Implementasi Kode dan Pra – Pemrosesan Data

Analisis dimulai dengan mengimpor data dari file ispu_dki_all.csv melalui pustaka pandas. Fitur-fitur numerik yang berkaitan dengan analisis kualitas udara, seperti pm25,

Vol. 2, No. 4 Juli 2025, Hal. 314-323
DOI: https://doi.org/10.62017/tektonik

pm10, so2, co, o3, dan no2, telah dipilih. Baris-baris dengan nilai hilang pada fitur-fitur ini dihapus untuk menjaga kualitas data yang bersih sebelum tahap pemrosesan selanjutnya.

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import seaborn as sns

# 1. Load data
df = pd.read_csv('ispu_dki_all.csv')

# 2. Fitur numerik
features = ['pm25', 'pm10', 'so2', 'co', 'o3', 'no2']
df_clean = df[features].dropna()

# 3. Standardisasi
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df_clean)
```

Gambar 1. Kode Pra – Pemrosesan Data

Setelah melakukan pembersihan data, proses standarisasi dilaksanakan menggunakan StandardScaler. Langkah ini krusial karena algoritma K-Means sangat tergantung pada skala data, dan normalisasi memastikan bahwa setiap fitur memberikan kontribusi yang seimbang terhadap penghitungan jarak. Variabel X_scaled menyimpan data yang sudah distandardisasi.

Clustering K -Means dan Reduksi Dimensi dengan PCA

Algoritma K-Means selanjutnya diterapkan pada data yang sudah distandarisasi. Berdasarkan analisis sebelumnya (seperti yang ditunjukkan oleh kodenya dan sesuai dengan referensi), jumlah cluster (*n_clusters*) ditentukan sebanyak 3. Label cluster untuk setiap titik data hasil clustering ditambahkan sebagai kolom baru (*'cluster'*) ke dalam DataFrame *df_clean*.

```
# 4. Clustering KMeans
kmeans = KMeans(n_clusters=3, random_state=42)
df_clean['cluster'] = kmeans.fit_predict(X_scaled)

# 5. PCA untuk visualisasi
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
df_clean['pca1'] = X_pca[:, 0]
df_clean['pca2'] = X_pca[:, 1]
```

Gambar 2. Clustering K – Means

Untuk menggambarkan hasil pengelompokan yang awalnya berada dalam ruang 6 dimensi (dari 6 fitur pencemar), digunakan Analisis Komponen Utama (*PCA*). *PCA*

Vol. 2, No. 4 Juli 2025, Hal. 314-323
DOI: https://doi.org/10.62017/tektonik

mengurangi dimensi data menjadi 2 komponen utama (*pca1 dan pca2*) yang menangkap mayoritas variasi dalam data awal. Komponen-komponen PCA ini selanjutnya dimasukkan ke dalam *df_clean* untuk tujuan visualisasi. Selanjutnya, informasi penting seperti nama stasiun dan tanggal disatukan kembali dengan data yang sudah dikelompokkan untuk membangun DataFrame df_result. Output terakhir dari proses pengelompokan ini juga disimpan ke dalam file CSV yang diberi nama *hasil_klasterisasi.csv* untuk analisis lebih lanjut atau penggunaan di luar.

```
# 6. Gabungkan metadata
df_meta = df[['tanggal', 'stasiun']].iloc[df_clean.index].reset_index(drop=True)
df_result = pd.concat([df_meta, df_clean.reset_index(drop=True)], axis=1)

# 7. Simpan ke CSV
df_result.to_csv('hasil_klasterisasi.csv', index=False)
```

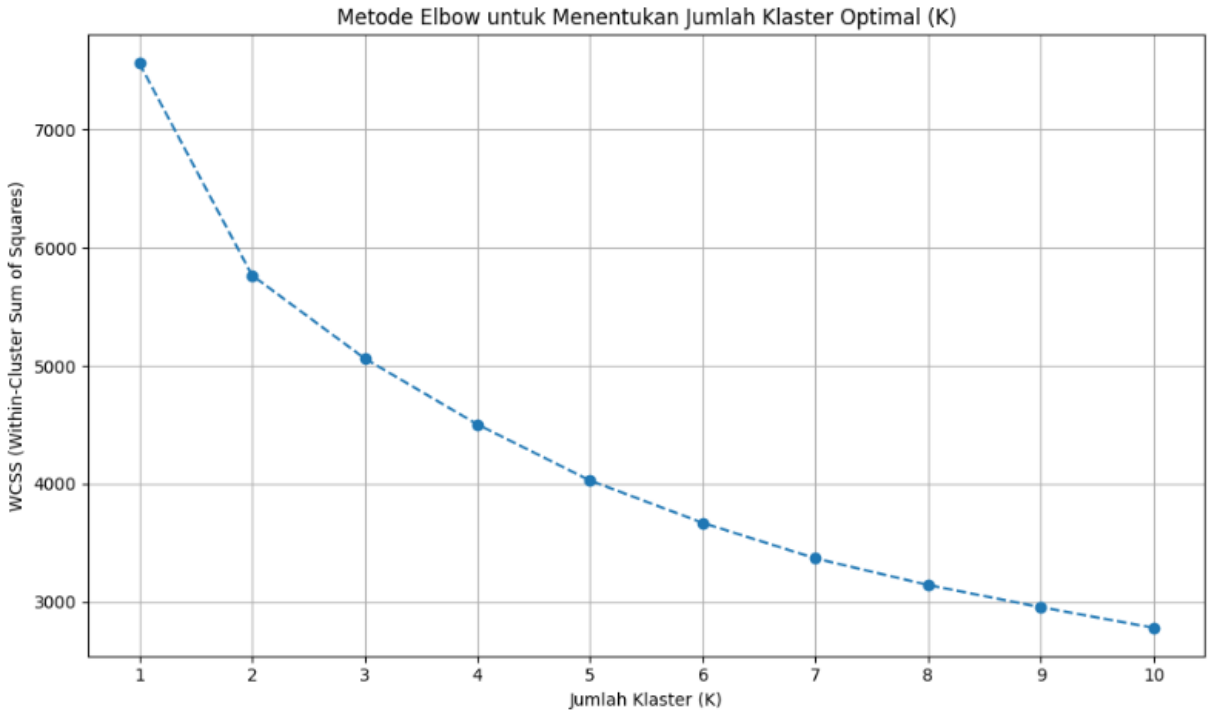
Gambar 3. Mengganbungkan Metadata

Visualisai Hasil Clustering

Penentuan Jumlah Klaster Optimal (Metode Elbow)

Sebelum melaksanakan clustering, menentukan jumlah klaster yang optimal merupakan langkah yang krusial. Metode elbow digunakan untuk menentukan nilai K yang paling tepat dengan memplot Within-Cluster Sum of Squares (WCSS) melawan jumlah klaster.

Vol. 2, No. 4 Juli 2025, Hal. 314-323
DOI: https://doi.org/10.62017/tektonik



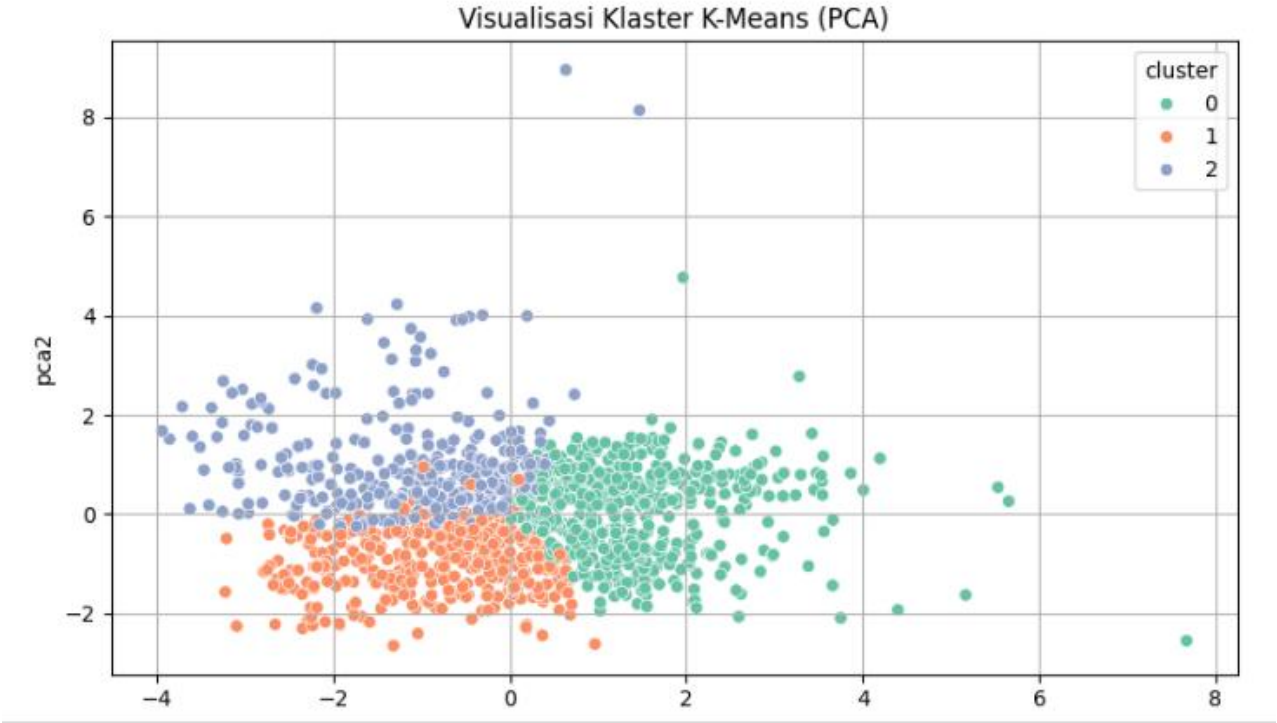
Gambar 4. Metode Elbow

Berdasarkan Gambar 4 (Metode Elbow), tampak jelas terdapat "siku" atau titik belok pada grafik WCSS saat jumlah klaster K=3. Setelah titik ini, penurunan WCSS menjadi kurang berarti. Ini menunjukkan bahwa 3 klaster merupakan jumlah yang ideal untuk pengelompokan data kualitas udara ini, karena memberikan keseimbangan yang baik antara kepadatan dalam klaster dan pemisahan di antara klaster. Penentuan ini menjadi alasan mengapa K-Means dioperasikan dengan n_clusters=3.

Vol. 2, No. 4 Juli 2025, Hal. 314-323
DOI: https://doi.org/10.62017/tektonik

Visualisasi Klaster K – Means (PCA Scatter Plot)

Visualisasi yang pertama merupakan scatter plot dari komponen-komponen PCA, di mana setiap titik data diberi warna sesuai dengan cluster yang telah ditentukan oleh algoritma K-Means.



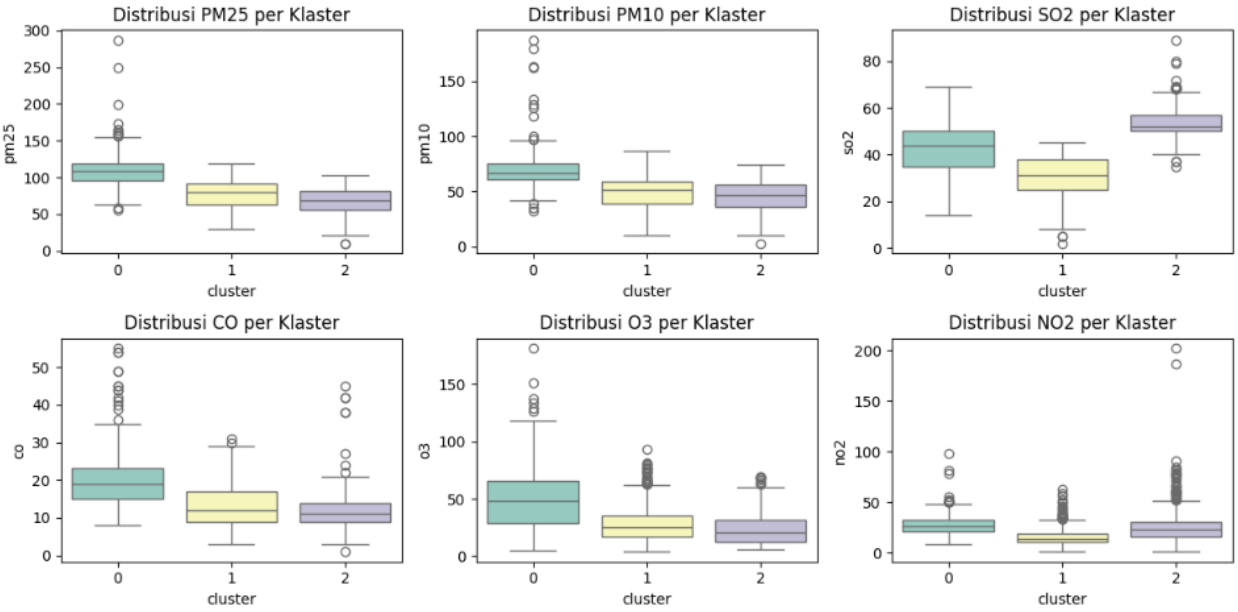
Gambar 5. Visualisasi Klaster K – Means (PCA Scatter Plot)

Pada Gambar 5, terlihat penyebaran titik-titik data dalam ruang dua dimensi yang dibentuk oleh dua komponen utama dari PCA. Setiap warna melambangkan satu kelompok. Sebaiknya, cluster akan menghasilkan kelompok yang terpisah dan padat, memperlihatkan perbedaan yang jelas antara pola kualitas udara yang berbeda. Nampak bahwa klaster 0 (hijau) dan klaster 1 (oranye) menunjukkan pemisahan yang cukup jelas, meskipun terdapat beberapa tumpang tindih. Klaster 2 (biru) terlihat lebih tersebar dan tumpang tindih dengan dua klaster lainnya, menunjukkan variabilitas yang lebih besar dalam kelompok ini atau batas yang kurang tegas pada dimensi PCA ini (Ma et al., 2019). Visualisasi ini memberikan pandangan umum tentang cara K-Means mengelompokkan data berdasarkan kombinasi fitur pencemar.

Vol. 2, No. 4 Juli 2025, Hal. 314-323
DOI: https://doi.org/10.62017/tektonik

Visualisasi Distribusi Polutan per Klaster (Boxplot)

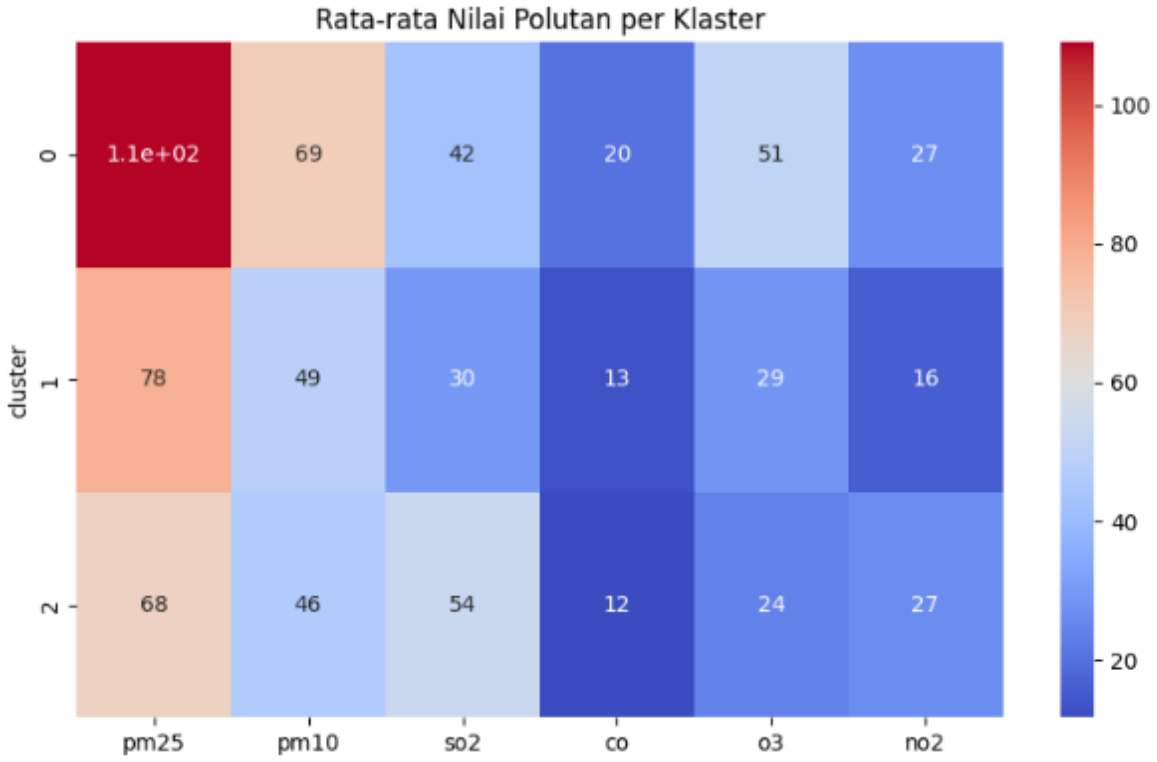
Visualisasi kedua menggunakan boxplot agar dapat menunjukkan distribusi nilai setiap polutan (pm25, pm10, so2, co, o3, no2) untuk setiap kluster.



Gambar 6. Distribusi Polutan per Klaster (Boxplot)

Visualisasi Rata – rata Nilai Polutan per Klaster (Heatmap)

Visualisasi ketiga adalah heatmap yang menampilkan nilai rata-rata setiap polutan untuk masing-masing cluster.



Gambar 7. Rata – rata Nilai Polutan per Klaster (Heatmap)

DOI: https://doi.org/10.62017/tektonik

Karakteristik Kluster yang Terbentuk

Berdasarkan tampilan grafis dan nilai rata-rata kontaminan per cluster (seperti yang terlihat pada Gambar 6 dan Gambar 7), analisis karakteristik tiap cluster dapat dilakukan**:**

- **Klaster 0 (Kualitas Udara Sedang)**
  Klaster ini cenderung menunjukkan nilai rata-rata polutan yang berada di antara klaster-klaster lainnya. Ini mungkin mencerminkan keadaan kualitas udara yang biasa, tidak terlalu buruk tetapi juga belum pada level ideal. Wilayah-wilayah dalam klaster ini membutuhkan pengawasan terus-menerus untuk menghindari kenaikan tingkat pencemaran.
- **Klaster 1 (Kualitas Udara Buruk / Tidak Sehat)**
  Klaster ini kemungkinan besar memiliki nilai rata-rata yang sangat tinggi untuk sebagian besar atau semua jenis polutan, terutama CO, PM10, dan PM2.5. Wilayah-wilayah dalam kluster ini sering kali menghadapi tingkat polusi udara yang tinggi, kemungkinan disebabkan oleh keberadaan industri berat, lalu lintas yang sangat padat, atau kejadian kebakaran. Tempat yang sering berada dalam kategori ini membutuhkan upaya mitigasi yang cepat dan intensif.
- **Klaster 2 (Kualitas Udara Baik)**
  Klaster ini akan ditandai dengan nilai rata-rata polutan terendah di antara semua klaster. Tempat-tempat dalam klaster ini mungkin terletak di wilayah yang tidak begitu padat, memiliki sirkulasi udara alami yang baik, atau terpisah dari sumber emisi utama. Ini menunjukkan bahwa udara memiliki kualitas yang cukup bersih dan sehat.

Umumnya, konsentrasi polutan udara mengalami variasi yang signifikan antara siang dan malam, di mana kadar CO biasanya lebih tinggi pada siang hari dan menurun pada malam hari di beberapa wilayah(Annas et al., 2022). Kepadatan lalu lintas, sektor industri, dan aktivitas konstruksi memberikan kontribusi besar terhadap penyebaran CO. Faktor cuaca seperti arah dan kecepatan angin juga sangat berpengaruh terhadap penyebaran polutan(Rahman et al., 2022). Angin yang kuat dapat mempermudah pencampuran polutan, sedangkan suhu tinggi dapat mempercepat reaksi kimia yang menghasilkan partikel halus. Sebaliknya, keadaan atmosfer yang tidak bergerak dapat mengakibatkan penumpukan polutan (Auliasari et al., 2021).

KESIMPULAN

Penggunaan K-Means clustering dalam analisis data kualitas udara di Jakarta telah berhasil menemukan pola-pola polusi yang beragam dan mengelompokkan wilayah-wilayah berdasarkan taraf kualitas udaranya. Dengan penerapan 3 kluster (menggunakan metode elbow), model dapat mengelompokkan kondisi kualitas udara ke dalam kategori yang dapat dipahami sebagai "Baik", "Sedang", dan "Tidak Sehat". PCA, boxplot, dan heatmap secara efektif memvisualisasikan perbedaan profil polutan antar cluster, memberikan pemahaman yang jelas mengenai karakteristik masing-masing kelompok.

Hasil *clustering* ini menyajikan data penting bagi pemerintah dalam menyusun kebijakan dan strategi penanggulangan polusi udara. Contohnya, wilayah yang tergolong dalam kluster "Tidak Sehat" membutuhkan perhatian utama serta langkah mitigasi yang lebih tegas, seperti pengurangan emisi kendaraan, peraturan industri yang lebih ketat, atau pemantauan yang lebih intensif.

DOI: https://doi.org/10.62017/tektonik

DAFTAR PUSTAKA

Aiello, L., Argiento, R., Legramanti, S., & Paci, L. (2025). *Bayesian nonparametric clustering for spatio-temporal data, with an application to air pollution*. http://arxiv.org/abs/2505.24694

ANALISIS DATA KUALITAS UDARA DI INDONESIA DENGAN K-MEANS CLUSTERING PADA BULAN AGUSTUS 2023 (2024).

Annas, S., Uca, U., Irwan, I., Safei, R. H., & Rais, Z. (2022). Using k-Means and Self Organizing Maps in Clustering Air Pollution Distribution in Makassar City, Indonesia. *Jambura Journal of Mathematics*, *4*(1), 167–176. https://doi.org/10.34312/jjom.v4i1.11883

Auliasari, K., Kertaningtyas, M., & Raya Karanglo Km, J. (2021). ANALISIS KUALITAS UDARA MENGGUNAKAN ALGORITMA K-MEANS. In *Jurnal Informatika & Rekayasa Elektronika)* (Vol. 4, Issue 2). http://e-journal.stmiklombok.ac.id/index.php/jireISSN.2620-6900

Choi, W., Ho, C. H., & Lee, Y. (2024). Temporal pattern classification of PM2.5 chemical compositions in Seoul, Korea using K-means clustering analysis. *Science of the Total Environment*, *927*. https://doi.org/10.1016/j.scitotenv.2024.172157

Choi, W., Song, M. Y., Kim, J. B., Kim, K., & Cho, C. (2023). Regional classification of high PM10 concentrations in the Seoul metropolitan and Chungcheongnam-do areas, Republic of Korea. *Environmental Monitoring and Assessment*, *195*(9). https://doi.org/10.1007/s10661-023-11732-6

Ma, J., Cheng, J. C. P., Lin, C., Tan, Y., & Zhang, J. (2019). Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. *Atmospheric Environment*, *214*. https://doi.org/10.1016/j.atmosenv.2019.116885

Nanda, A. P., Syarifuddin, A., & Mahdi, M. I. (2023). Sistem Pendukung Keputusan Mengukur Pencemaran Udara di Kabupaten Pringsewu Metode K-Means. *EXPERT: Jurnal Manajemen Sistem Informasi Dan Teknologi*, *13*(1), 17. https://doi.org/10.36448/expert.v13i1.3027

Rahman, E. A., Hamzah, F. M., Latif, M. T., & Dominick, D. (2022). Assessment of PM2.5 Patterns in Malaysia Using the Clustering Method. *Aerosol and Air Quality Research*, *22*(1). https://doi.org/10.4209/AAQR.210161

Mohd Aftar Abu Bakar, Fatin Nur Afiqah Suris, Noratiqah Mohd Ariff, Kamarulzaman Ibrahim, & Tan Zhen Jie. (2022). *Time series clustering of Malaysia Air quality time series data*.