JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 3S1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i3S1.8141

ANALISIS POLA DAN TREN KUALITAS UDARA DKI JAKARTA 2022–2025 MENGGUNAKAN K-MEANS *CLUSTERING*

Nadhia Ayu Rahmadenti[1*], Rachmat Adi Purnama[2], Tomi Alfian Armawan Sandi[3]

[1,2,3] Teknik Informatika, Universitas Bina Sarana Informatika; Jl. Margonda No.8, Pondok Cina, Kota Depok, Jawa Barat 16424

Keywords:
Kualitas udara, ISPU, K-Means Clustering, Data Mining, DKI Jakarta

Corespondent Email:
nadhiaayurahmadenti@gmail.com

Abstrak. Kualitas udara di DKI Jakarta menjadi salah satu permasalahan lingkungan yang signifikan akibat padatnya populasi, lalu lintas, dan aktivitas industri. Penelitian ini bertujuan untuk menemukan pola dan tren kualitas udara melalui pengelompokan data Indeks Standar Pencemar Udara (ISPU) periode 2022–2025 menggunakan algoritma K-Means Clustering. Metode penelitian meliputi pengumpulan data dari situs resmi Pemerintah Provinsi DKI Jakarta, pra-pemrosesan data, penentuan jumlah cluster dengan metode Elbow dan Silhouette Score, implementasi algoritma K-Means, serta evaluasi hasil clustering. Hasil penelitian menunjukkan bahwa algoritma K-Means mampu membentuk tiga cluster optimal yang mewakili kategori kualitas udara: baik, sedang, dan tidak sehat. Visualisasi menunjukkan adanya tren perbaikan kualitas udara dari tahun 2022 hingga 2025. Hasil ini dapat menjadi dasar bagi pemerintah dalam menetapkan kebijakan pengendalian polusi udara di DKI Jakarta.

Abstract. *Air quality in DKI Jakarta is a significant environmental problem due to dense population, traffic, and industrial activity. This study aims to identify air quality patterns and trends by clustering Air Pollutant Standard Index (ISPU) data for the 2022–2025 period using the K-Means Clustering algorithm. The research method includes data collection from the official website of the DKI Jakarta Provincial Government, data pre-processing, determining the number of clusters using the Elbow and Silhouette Score methods, implementing the K-Means algorithm, and evaluating the clustering results. The results show that the K-Means algorithm is able to form three optimal clusters representing the air quality categories: good, moderate, and unhealthy. The visualization shows a trend of improving air quality from 2022 to 2025. These results can be used as a basis for the government in establishing air pollution control policies in DKI Jakarta.*

1. PENDAHULUAN

Umumnya, semua bentuk kehidupan di Bumi sangat tergantung pada udara yang ada di sekeliling kita. Oksigen dari udara itu menjadi bahan pokok untuk proses metabolisme di tubuh manusia, hewan, bahkan tumbuhan [1]. Udara yang sehat adalah udara yang bersih, bebas dari polutan, dan nyaman untuk dihirup. Seluruh makhluk hidup membutuhkan udara yang bersih karena berperan langsung dalam menjaga kesehatan fisik dan psikologis. Oleh karena itu, tingkat polusi atau pencemaran udara yang rendah menjamin udara yang bersih dan sehat, yang dibutuhkan bagi semua manusia demi kelangsungan hidup [2]. Pencemaran udara menjadi isu utama di kota besar seperti DKI Jakarta akibat tingginya emisi kendaraan bermotor, industri, dan aktivitas penduduk

[3]. Menurut laporan IQAir 2025, Jakarta termasuk kota dengan kualitas udara tidak sehat dengan konsentrasi $PM_{2.5}$ mencapai 35,5 µg/m³. Pemantauan kualitas udara rutin dilakukan oleh Pemerintah DKI Jakarta bersama lembaga terkait untuk memahami tren kualitas udara dan mengambil tindakan untuk mengendalikan atau mengurangi polusi udara [4]. Beberapa jenis polutan masih dapat diterima oleh tubuh dalam jumlah yang masih di batas wajar, namun jika melampaui batas wajar dapat menyebabkan masalah serius pada kesehatan manusia seperti penyakit jantung, infeksi pada saluran pernapasan, stroke, asma hingga penyakit paru [5]. Untuk memahami kondisi udara secara komprehensif, diperlukan analisis berbasis data melalui pendekatan data mining, khususnya metode clustering. Penelitian ini menggunakan algoritma K-Means untuk mengelompokkan data ISPU berdasarkan tingkat pencemaran udara. Kebaruan penelitian ini terletak pada analisis pola temporal 2022–2025 dengan pendekatan Elbow dan Silhouette yang menghasilkan jumlah cluster optimal. Tujuan utama penelitian ini adalah untuk mengidentifikasi pola dan tren kualitas udara di DKI Jakarta yang dapat digunakan sebagai dasar kebijakan pengendalian polusi.

2. TINJAUAN PUSTAKA

2.1 Kualitas Udara

Udara merupakan kombinasi campuran berbagai jenis gas yang ditemukan di bawah lapisan bumi dan sangat penting bagi kehidupan makluk hidup, terutama manusia. Sekitar 78% nitrogen, 21,94% oksigen, 0,93% argon, dan 0,032% karbon dioksida adalah komponen udara, selain gas mulia lainnya[6]. Kualitas udara ditentukan oleh konsentrasi polutan seperti PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, dan O_3 yang diukur menggunakan Indeks Standar Pencemar Udara (ISPU) [7]. Kualitas udara di suatu tempat selalu berubah dari hari ke hari, terutama di era perkembangan teknologi dan aktivitas industri yang semakin pesat. Dampak kualitas udara di setiap lokasi terhadap kesehatan dan lingkungan ditentukan oleh kualitas udara di lokasi tersebut (Indeks Standar Pencemaran Udara). ISPU digunakan sebagai tolak ukur untuk menghitung tingkat kadar polutan di udara [8].

2.2 Indeks Standar Pencemar Udara (ISPU)

Indeks ini tidak hanya menunjukkan seberapa bersih atau kotor udara di suatu tempat, tetapi juga memberi masyarakat gambaran tentang risiko pencemaran udara bagi kesehatan. Particulate Matter (PM_{10} dan $PM_{2.5}$), Ozon, Nitrogen Dioksida, dan Sulfur Dioksida adalah beberapa zat pencemar yang biasanya dicatat dalam penilaian ISPU. Nilai ISPU sendiri didasarkan pada skala 0-500, dengan nilai yang lebih tinggi menunjukkan bahwa kualitas udara lebih buruk dan risiko yang dihadapi manusia lebih besar. Nilai-nilai ini secara teratur diperiksa oleh lembaga lingkungan dan kesehatan publik untuk memastikan bahwa informasi tersebut tetap relevan [9].

2.3 Data Mining

Data mining didefinisikan sebagai proses penemuan pola dalam data. Berdasarkan tugasnya, data mining dikelompokkan menjadi deskripsi, estiminasi prediksi, klasifikasi, klastering, dan asosiasi. Suatu proses penambangan informasi penting dari suatu data. Informasi penting ini didapat dari suatu proses yang amat rumit seperti menggunakan artificial intelligence, teknik statistik, ilmu matematika, machine learning dan lain sebagainya [10].

2.4 Clustering

Clustering merupakan data yang sangat besar sulit unuk dianalisis dan dipahami, oleh karena itu perlu adanya penggelompokkan atau clustering. Dalam hal ini penggelompokkan bertujuan untuk meningkatkan pemahaman terhadap data, untuk menilai kualitas dari data [11].

Clustering adalah teknik pengumpulan data yang digunakan untuk menganalisis dan memecahkan masalah dalam pengelompokan data menjadi partisi yang lebih efisien dari dataset ke subset. Dalam proses ini, tujuannya adalah untuk mendistribusikan kasus (benda, orang, peristiwa, dll.) ke dalam kelompok, sehingga tingkat hubungan antara anggota cluster yang sama kuat dan lemah antara anggota cluster yang berbeda [12]. Pengelompokan atau clustering merupakan salah satu metode dari data mining, yang tidak membutuhkan tujuan hasil dari pengawasan yang diberikan. Metode ini dilaksanakan tanpa bimbingan ataupun seorang pelatih. Pengelompokan hierarkis dan non-hierarkis adalah dua jenis teknik pemisahan data yang tersedia dalam pengolahan data [13].

2.5 Algoritma K-Means

Algoritma K-Means merupakan salah satu metode pengelompokan data non-hierarki (sekatan) yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih. Metode ini mempartisi data ke dalam kelompok sehingga data berkarakteristik sama dimasukkan ke dalam kelompok yang lain. Adapun tujuan pengelompokan data ini adalah untuk meminimalkan fungsi objektif yang diset dalam proses pengelompokan, yang pada umumnya berusaha meminimalkan variasi di dalam suatu kelompok dan memaksimalkan variasi antar kelompok [14]. Algoritma K-Means adalah algoritma pengelompokan interaktif yang membagi set data ke dalam sejumlah cluster K yang telah ditetapkan sebelumnya. Algoritma K-Means mudah digunakan dan dijalankan. Sangat praktis, cepat, dan dapat disesuaikan. K-Means telah lama menjadi salah satu algoritma yang paling penting dalam data mining [15].

Algoritma K-Means memiliki kelebihan cepat dan efisien untuk dataset dengan jumlah yang besar dan mudah diimplementasikan, namun terdapat kekurangan nya juga yaitu harus menentukan jumlah cluster terlebih dahulu, sensitive terhadap data outlier, dan dapat menghasilkan hasil yang berbeda tergantung inisialisasi awal.

Langkah-langkah analisis metode K-Means adalah sebagai berikut:

1. Tentukan k sebagai pusat massa awal.
2. Hitung jarak Euclidean.

   $dik = 2\sqrt{\sum_{j=1}^{m}\left(x_{ij-}c_{jk}\right)}$

   dengan:

xij = pusat cluster

cjk= data indikator tingkat pengangguran

dik = Jarak tiap benda

3. Menghitung anggota cluster berdasarkan jarak terpendek.

   $Min = 2\ \sum_{k=1}^{k} d_{ik}\sqrt{\sum_{j=1}^{m}\left(x_{ij-}c_{jk}\right)}$

4. Hitung centroid baru untuk iterasi selanjutnya.

   $c = \frac{\sum_{i=1}^{p} x_{ij}}{p}$

   dengan:

   p = banyaknya anggota cluster k.
5. Ulangi langkah ketiga dan keempat, jika tidak ada perubahan anggota cluster maka iterasi dihentikan dan diperoleh hasil cluster.

2.6 Metode Elbow

Metode elbow biasanya digunakan untuk menentukan jumlah klaster yang akan dibuat untuk proses clustering. Metode elbow menghitung kohesi dan pemisah klister. Kohesi klaster menunjukkan seberapa dekat data terhubung dalam satu klaster, sedangkan pemisah menunjukkan seberapa jauh satu klaster dari yang lain [16].

2.7 Silhouette Coefficient

Metode Silhouette menggunakan koefisien siluet yang menggabungkan pemisahan dan kohesi. Semakin besar koefisien Silhouette, semakin baik cluster tersebut.

$$SC = maks_k\ SI\ (k)$$

Keterangan:

SC = *Silhouette Coefficient*

SI = *Silhouette* Index Global

k = jumlah cluster

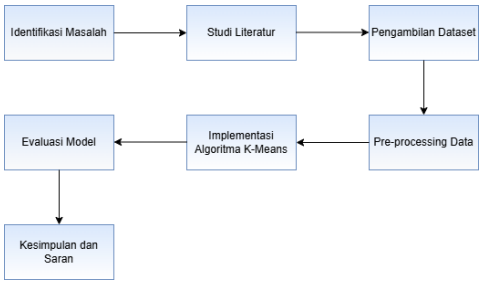
2.8 Heatmap

Heatmap adalah visualisasi atau pemetaan yang menampilkan kelompok data dengan representasi warna yang berbeda-beda. Biasanya, semakin besar kelompok data, warnanya akan semakin gelap, dan biasanya ditandai dengan warna merah.

3. METODE PENELITIAN

3.1 Kerangka Penelitian

Kerangka penelitian ini mendeskripsikan alur kerja atau tahapan-tahapan yang dilakukan secara berurutan dan sistematis untuk memastikan penelitian berjalan sesuai dengan rencana dan tujuan yang telah ditetapkan.



Gambar 1. Kerangka Penelitian

1. Identifikasi Masalah

Langkah pertama yang dilakukan penulis yaitu mengidentifikasi masalah dalam penelitian. Adanya permasalahan pencemaran udara di daerah DKI Jakarta yang disebabkan oleh kendaraan ataupun pabrik-pabrik di Kawasan industri membuat pemerintah cukup sulit untuk menangani masalah udara tersebut. Oleh karena itu pentingnya mengklasifikasi kualitas udara menggunakan data ISPU dapat mempermudah pemerintah DKI Jakarta untuk mengatasi permasalahan dan melihat daerah mana saja yang polusi udara nya sangat tinggi sehingga bisa menjadi dasar kebijakan lingkungan dan pemerintah.

2. Studi Literatur

Setelah dilakukan identifikasi masalah, selanjutnya adalah mencari teori-teori yang relevan. Studi literatur dilakukan untuk mengumpulkan teori penelitian dari sumber-sumber yang resmi dan dapat dipertanggung jawabkan, seperti buku, e-book, jurnal, artikel internet, dan lainnya. Studi literatur yang dicari dalam penelitian ini yaitu kualitas udara, ISPU, data mining, dan algoritma K-Means adalah subjek penelitian ini, bersama dengan tinjauan penelitian sebelumnya yang relevan.

3. Pengambilan Dataset

Langkah selanjutnya yaitu mengambil data ISPU dari situs resmi pemerintah DKI Jakarta (https://satudata.jakarta.go.id/) dengan rentang tahun 2022–2025 dan jumlah data awal yaitu 4105 yang mencakup lima stasiun pemantauan kualitas udara, yaitu:

a. DKI1 Bundaran HI (Jakarta Pusat)
b. DKI2 Kelapa Gading (Jakarta Utara)
c. DKI3 Jagakarsa (Jakarta Selatan)
d. DKI4 Lubang Buaya (Jakarta Timur)
e. DKI5 Kebon Jeruk (Jakarta Barat)

Data tersebut akan digunakan untuk pelatihan model klasifikasi dengan Algoritma K-Means Clustering.

4. Pra-Pemrosesan Data (Preprocessing)
   a. Melakukan pembersihan data agar siap untuk diolah dan dianalisis.
   b. Tahap preprocessing meliputi:
      1) Pembersihan Data (Data *Cleaning*)
      2) Normalisasi Data
      3) Transformasi Data

   Berdasarkan hasil preprocessing jumlah data awal yaitu 4105 data. Namun setelah dilakukan proses cleaning jumlah data menjadi 3020 data, hal ini berarti sebanyak 1085 data tidak valid atau hilang sudah dibersihkan.

5. Implementasi K-Means Clustering

Setelah data siap diolah, langkah selanjutnya adalah menerapkan algoritma K-Means pada data ISPU untuk membentuk kelompok-kelompok (cluster) berdasarkan tingkat kualitas udara menggunakan tools Google Colab.

6. Evaluasi dan Visualisasi

Tahap selanjutnya yaitu mengevaluasi hasil clustering dengan interpretasi visual seperti grafik scatter plot, grafik batang, dan grafik garis.

7. Penarikan Kesimpulan

Setelah semua proses dilakukan, maka penulis akan melakukan peringkasan. Penulis akan menyimpulkan proses dan pola atau tren kualitas udara berdasarkan hasil clustering yang telah dilakukan secara ringkas. Penulis juga akan menulis saran-saran berdasarkan kekurangan yang terjadi pada penelitian ini agar pada penelitian selanjutnya akan mendapatkan hasil yang maksimal.

4. HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini berasal dari portal resmi Satu Data DKI Jakarta (https://satudata.jakarta.go.id/) dengan periode waktu tahun 2022–2025. Dataset terdiri atas 4105 baris data pengamatan harian yang dikumpulkan dari lima stasiun pemantauan kualitas udara, yaitu: DKI1 Bundaran HI, DKI2 Kelapa Gading, DKI3 Jagakarsa, DKI4 Lubang Buaya, dan DKI5 Kebon Jeruk. Setiap data berisi parameter PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, dan O_3 serta nilai Indeks Standar Pencemar Udara (ISPU).

4.2 Preprocessing Data

Sebelum masuk ke tahap pelatihan model, data harus dipastikan layak untuk dilatih. Hal pertama yang peneliti lakukan adalah melakukan tahap preprocessing, tahap preprocessing dilakukan untuk meningkatkan kualitas data sebelum dilakukan clustering. Langkah-langkahnya meliputi pembersihan data, transformasi, dan normalisasi. Nilai kosong atau tanda '-' pada dataset dihapus. Jumlah data dapat dilihat pada gambar 2.

```
Jumlah data kosong (NaN) per kolom:
periode_data                0
tanggal                     0
stasiun                     0
PM_10                       0
PM_2.5                      0
Sulfur_Dioksida(SO2)        0
Karbon_Monoksida(CO)        0
Ozon(O3)                    0
Nitrogen_Dioksida(NO2)      0
max                         0
parameter_pencemar_kritis   20
kategori                    0
dtype: int64


Jumlah nilai '-'atau '---' per kolom:
PM_10: 355
PM_2.5: 326
Sulfur_Dioksida(SO2): 74
Karbon_Monoksida(CO): 78
Ozon(O3): 58
Nitrogen_Dioksida(NO2): 119
max: 8
parameter_pencemar_kritis: 56
```

Gambar 3. Hasil Pengecekan Dataset

Pada gambar 2 menunjukan hasil saat dilakukan pengecekan dataset perkolom untuk melihat apakah terdapat data yang kosong, atau yang bernilai "-". Pada kolom parameter_pencemar_kritis terdapat data yang kosong dengan jumlah 20, pada kolom dengan nilai "-" yaitu PM_{10} berjumlah 355, $PM_{2.5}$ berjumlah 326, Sulfur Dioksida(SO_2) berjumlah 74, Karbon Monoksida(CO_2) berjumlah 78, Ozon(O_3) berjumlah 58, Nitrogen Dioksida(NO_2) berjumlah 119.

```
data_cleaned.replace(['-', '---'], np.nan, inplace=True)
numeric_cols = ['PM_10', 'PM_2.5','Sulfur_Dioksida(SO2)',
                'Karbon_Monoksida(CO)', 'Ozon(O3)', 'Nitrogen_Dioksida(NO2)']
for col in numeric_cols:
    data_cleaned[col] = pd.to_numeric(data_cleaned[col], errors='coerce')

data_cleaned.dropna(subset=numeric_cols, inplace=True)
print("Jumlah data kosong (NaN) atau '-' per kolom setelah cleaning:")
print(data_cleaned.isna().sum())
```

Gambar 2. Mengganti Nilai Kosong (NaN) atau "-"

Selanjutnya mengganti nilai "-" yang tidak valid dengan NaN, menkonversi kolom polutan menjadi tipe data numerik. Jika ada nilai yang tidak dapat dikonversi seperti teks atau symbol yang tidak relevan, maka nilai akan diubah menjadi NaN, kemudian menghapus baris-baris yang memiliki nilai NaN atau "-".

```
Jumlah data kosong (NaN) atau '-' per kolom setelah cleaning:
tanggal                 0
stasiun                 0
PM_10                   0
PM_2.5                  0
Sulfur_Dioksida(SO2)    0
Karbon_Monoksida(CO)    0
Ozon(O3)                0
Nitrogen_Dioksida(NO2)  0
kategori                0
dtype: int64
```

Gambar 4. Hasil Cleaning Dataset

Pada gambar 4 menunjukan hasil cleaning dataset. Setelah mengkonversi ke tipe numerik dan penghapusan baris data kosong, jumlah data tersisa 3020 baris dari 4105 data.

4.3 Normalisasi Data

Pada tahap ini normalisasi data bertujuan untuk mengatur skala data setiap atribut agar tidak ada variable yang memiliki skala lebih besar serta mendominasi hasil analisis dan memastikan bahwa semua fitur memiliki skala yang sama.

```
scaler = MinMaxScaler()
data_scaled = scaler.fit_transform(data_cleaned[numeric_cols])
```

Gambar 5. Normalisasi Data

Pada gambar 5 menunjukan tahapan normalisasi data yang menggunakan MinMaxScaler untuk menstandarisasi data polutan, yang memastikan bahwa setiap kolom memiliki distribusi dengan mean 0 dan standar deviasi 1.

4.4 Transformasi Data

Pada tahap ini Transformasi data dilakukan untuk memastikan bahwa data memiliki format yang sesuai untuk dianalisis.

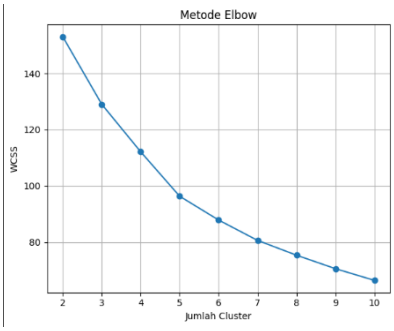
```
[ ] data_cleaned['tanggal'] = pd.to_datetime(data_cleaned['tanggal'], format='%m/%d/%Y', errors='coerce')

    data_cleaned['hari'] = data_cleaned['tanggal'].dt.day
    data_cleaned['bulan'] = data_cleaned['tanggal'].dt.month
    data_cleaned['tahun'] = data_cleaned['tanggal'].dt.year

    data_cleaned.dropna(subset=['tanggal'], inplace=True)
```

Gambar 6. Transformasi Data

Pada gambar 6 menunjukan tahap transformasi data. Pada tahap ini dilakukan perubahan pada kolom tanggal menjadi tipe datetime, dan menambahkan tiga kolom baru yaitu hari, bulan, dan tahun yang diekstrasi dari kolom tanggal agar dataset memiliki format yang sesuai saat dianalisis.

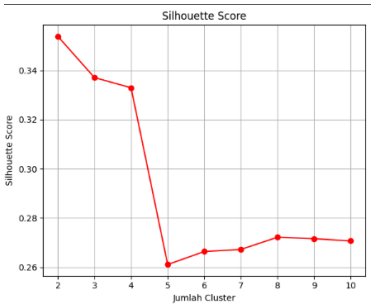
4.5 Penentuan Jumlah Cluster

Tahap ini akan menentukan berapa jumlah cluster (k) yang terbentuk menggunakan metode Elbow dan Silhouette Score. Pada metode Elbow, kita menghitung nilai inertia (sum of squared distances between samples and their centroids) untuk berbagai nilai K, kemudian memilih nilai K pada titik di mana penurunan inertia mulai melambat.



Gambar 7. Grafik Metode Elbow

Berdasarkan gambar grafik metode Elbow, sumbu x menunjukan jumlah cluster yang bervariasi antara 2 sampai 10. Sumbu y menunjukan nilai Within-Cluster Sum of Squares (WCSS), yang menggambarkan total jarak antar titik dalam cluster terhadap pusat cluster (centroid). Pada grafik ini terjadi penurunan yang cukup tajam pada WCSS antara jumlah cluster 2 dan 4, tapi setelah jumlah cluster 4 penurunan mulai melambat secara signifikan. Titik siku terlihat jelas terjadi pada jumlah cluster 4, yang menunjukan bahwa jika menambah jumlah cluster lebih dari 4 tidak akan memberikan penurunan WCSS yang signifikan lagi.

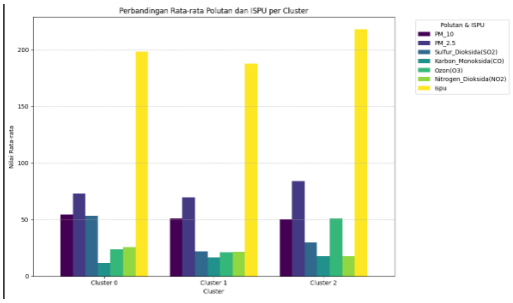


Gambar 8. *Silhouette Score*

Pada gambar grafik Silhouette Score, sumbu x menunjukan jumlah cluster yang diuji mulai dari 2 sampai 10, sumbu y menunjukan Silhouette Score, yang mengukur kualitas clustering. Nilai Silhouette Score bernilai antar -1 sampai 1:

a. Nilai lebih tinggi mendekati 1 menunjukkan clustering yang baik, di mana titik data dalam satu cluster lebih dekat satu sama lain dan lebih jauh dari cluster lain.
a. Nilai yang lebih rendah menunjukkan bahwa titik data mungkin lebih cocok berada di cluster lain. Grafik Silhouette Score mencapai nilai tertinggi sekitar jumlah cluster 3 dengan nilai 0.337, maka itu jumlah cluster yang optimal adalah 3, karena setelah jumlah cluster 3 Silhouette Score mulai menurun, yang menunjukan bahwa clustering lebih dari 3 cluster kurang optimal.

Berdasarkan hasil metode Elbow dan *Silhouette Score* penulis memutuskan untuk menggunakan 3 cluster sebagai jumlah cluster optimal dalam penelitian ini, karena *Silhouette* menggabungkan aspek kohesi dan pemisahan serta memberikan penilaian kualitas cluster, jika menggunakan 4 cluster sesuai dengan hasil metode Elbow hasil pada Silhouette Score tidak optimal dan scorenya rendah yaitu dengan nilai 0.333, sedangkan *Silhouette Score* menunjukan hasil yang lebih baik dalam hal memisahkan cluster.



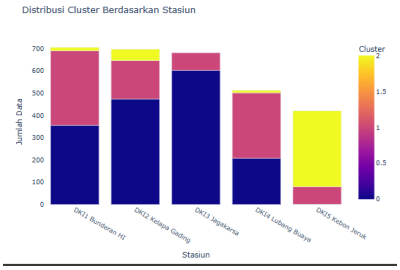
Gambar 9. Grafik Perbandingan Rata-rata Polutan per Cluster

Cluster 0 merupakan cluster dengan kualitas udara yang buruk dengan nilai $PM_{2.5}$ rata-rata 72.90 µg/m3, SO_2 rata-rata 53.16 µg/m3, dan PM_{10} rata-rata 54.40 µg/m3 yang merupakan nilai SO_2 tertinggi dibandingkan klaster lainnya, dan nilai ISPU rata-rata 198.48. Nilai ini menunjukkan bahwa gas beracun dan polutan partikulat halus yang paling banyak dihasilkan dari pembakaran biomassa, aktivitas kendaraan bermotor, dan emisi dari aktivitas industry atau pembangkit energi. Cluster ini masuk ke dalam kategori Tidak Sehat (ringan) karena tingkat polutan yang tinggi, yang berarti udara di dalamnya dapat berdampak negatif pada kesehatan masyarakat umum, terutama kelompok rentan seperti anak-anak, orang tua, dan penderita penyakit pernapasan.

Cluster 1 merupakan cluster dengan kualitas udara yang buruk juga. Nilai $PM_{2.5}$ adalah 69.48 µg/m3, CO adalah 16.54 µg/m3, PM_{10} adalah 50.98 µg/m3, dan nilai rata-rata ISPU adalah 187.91. Kadar O_3 dan SO_2, dan NO_2 relatif cukup rendah, namun $PM_{2.5}$, PM_{10}, dan CO cukup tinggi. Hal tersebut menunjukkan bahwa cluster ini dikategorikan sebagai Tidak Sehat (sedang), yang berarti memiliki pengaruh terhadap kesehatan masyarakat. Kualitas udara pada cluster ini biasanya menunjukkan daerah perkotaan dengan aktivitas lalu lintas yang padat dan tidak terkontrol atau pembakaran terbuka di area permukiman padat penduduk.

Cluster 2 memiliki nilai rata-rata $PM_{2.5}$ sebesar 84.03 µg/m³, O_3 sebesar 51.16 µg/m³, dan CO sebesar 17.80 µg/m³. Angka tersebut menunjukkan tingkat pencemaran yang cukup tinggi, terutama dari polusi debu/partikulat emisi, pembakaran bahan bakar fosil dari sektor industry, dan O_3 yang tinggi menunjukan reaksi fotokimia di atmosfer yang umumnya terjadi saat cuaca panas dan paparan sinar matahari kuat. Kualitas udara di cluster ini masuk dalam kategori Sangat Tidak Sehat, yang berarti tidak berdampak pada masyarakat umum tetapi berisiko bagi kelompok sensitif. Cluster ini memiliki polutan $PM_{2.5}$ dan O_3

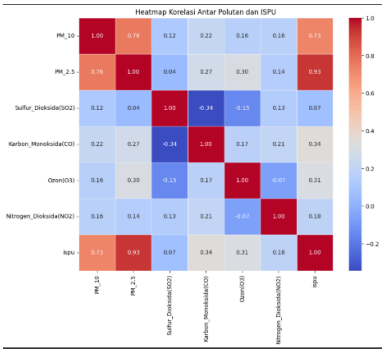
yang dominan, menunjukkan bahwa emisi dari sumber yang tidak bergerak seperti pabrik dan pembangkit listrik harus dikendalikan.



Gambar 11. Distribusi Cluster Berdasarkan Stasiun

Gambar 10 merupakan distribusi cluster berdasarkan lokasi stasiun pemantauan kualitas udara di DKI Jakarta.

Stasiun DKI1 Bunderan HI dan DKI2 Kelapa Gading didominasi oleh cluster 0 (biru) dan cluster 1 (merah muda), yang termasuk ke dalam kategori tidak sehat sedang hingga berat. Stasiun DKI3 Jagakarsa menunjukan proporsi campuran antara cluster 0 (biru) dan cluster 1 (merah muda) yang berarti kualitas udara relative lebih baik dibanding wilayah lain, meskipun masih masuk ke dalam kategori tidak sehat. Stasiun DKI4 Lubang Buaya di dominasi oleh cluster 1 (merah muda), cluster 0 (biru), dan sedikit cluster 2 (kuning) yang berarti daerah ini tergolong tidak sehat sedang, kadar CO dan NO_2 cukup tinggi akibat aktivitas kendaraan dan permukiman padat, fluktuasi kualitas udara daerah lubang buaya sebagian besar juga dipengaruhi oleh arah angin dan aktivitas pembakaran terbuka. Stasiun DKI5 Kebon Jeruk didominasi oleh cluster 2 (kuning) yang merupakan kategori sangat tidak sehat, disebabkan oleh konsentrasi $PM_{2.5}$ dan CO yang tinggi. Kondisi ini menunjukan bahwa area barat Jakarta memiliki tingkat penecemaran udara paling berat.



Gambar 10. *Heatmap* Korelasi Antar Polutan dan ISPU

5. KESIMPULAN

Berdasarkan hasil analisis data kualitas udara DKI Jakarta menggunakan metode K-Means Clustering, dapat ditarik beberapa kesimpulan sebagai berikut:

1. Penerapan algoritma K-Means Clustering dalam mengelompokkan data ISPU DKI Jakarta tahun 2022–2025 berhasil dilakukan di lima stasiun pemantauan melalui tahapan yang sistematis, dimulai dari pengumpulan 4105 dataset dari situs resmi pemerintah DKI Jakarta, dilanjutkan dengan pra-pemrosesan data yang meliputi pembersihan data dari nilai kosong atau tidak valid sehingga menghasilkan 3020 data siap diolah, normalisasi data menggunakan *MinMaxScaler*, dan transformasi kolom tanggal menjadi format yang sesuai untuk analisis.
2. Hasil pengelompokan kualitas udara berdasarkan data ISPU menggunakan metode K-Means menunjukkan bahwa jumlah cluster optimal adalah 3. Penentuan ini didasarkan pada hasil Silhouette Score dan Metode Elbow yang menunjukkan titik "siku" pada cluster ke-3 di mana penurunan WCSS mulai melambat secara signifikan, serta Silhouette Score yang mencapai nilai tertinggi sekitar 0.337 pada jumlah cluster 3. Ini mengindikasikan bahwa

data berhasil dikelompokkan dengan baik ke dalam 3 kategori kualitas udara.

3. Pemetaan hasil clustering menunjukkan perbedaan distribusi cluster pada setiap stasiun:
   a. DKI1 (Bundaran HI) dan DKI2 (Kelapa Gading) didominasi oleh cluster 0 (tidak sehat sedang–berat).
   b. DKI3 (Jagakarsa) relatif lebih baik dengan dominasi cluster 0 (tidak sehat ringan).
   c. DKI4 (Lubang Buaya) didominasi cluster 1 (tidak sehat sedang, polutan utama CO dan NO_2).
   d. DKI5 (Kebon Jeruk) menunjukkan kondisi paling berat (cluster 2, sangat tidak sehat).

Hasil pemetaan ini menunjukkan bahwa wilayah Jakarta Barat dan Pusat memiliki tingkat pencemaran tertinggi, sedangkan Jakarta Selatan relative lebih baik.

4. Berdasarkan hasil clustering, terbentuk tiga kelompok kualitas udara yang termasuk dalam kategori tidak sehat, namun berbeda tingkat keparahan polutannya, yaitu:
   a. Cluster 0: Tidak sehat ringan (rata-rata ISPU 198, polutan yang dominan $PM_{2.5}$ dan SO_2).
   b. Cluster 1: Tidak sehat sedang (rata-rata ISPU 187, polutan yang dominan CO dan PM_{10}).
   c. Cluster 2: Sangat tidak sehat (rata-rata ISPU 218, polutan dominan $PM_{2.5}$ dan CO).
5. Berdasarkan hasil heatmap korelasi, diperoleh bahwa polutan $PM_{2.5}$ dan CO memiliki korelasi paling kuat dengan peningkatan nilai ISPU. Hal ini menunjukkan bahwa kedua polutan tersebut merupakan indikator utama penurunan kualitas udara di DKI Jakarta. Sementara itu, polutan O_3 menunjukkan korelasi negatif terhadap $PM_{2.5}$ dan CO, menandakan adanya reaksi fotokimia di atmosfer yang berbeda arah terhadap pembentukan polutan partikulat.

DAFTAR PUSTAKA

[1] A. D. Wiranata, S. Soleman, I. Irwansyah, I. K. Sudaryana, and R. Rizal, "Klasifikasi Data Mining Untuk Menentukan Kualitas Udara Di Provinsi Dki Jakarta Menggunakan Algoritma K-Nearest Neighbors (K-Nn)," *Infotech: Journal of Technology Information*, vol. 9, no. 1, pp. 95–100, Jun. 2023, doi: 10.37365/jti.v9i1.164.

[2] M. Astriyani, I. N. Laela, D. P. Lestari, L. Anggraeni, and T. Astuti, "Analisis Klasifikasi Data Kualitas Udara Dki Jakarta Menggunakan Algoritma C.45," *JuSiTik : Jurnal Sistem dan Teknologi Informasi Komunikasi*, vol. 6, no. 1, pp. 36–41, Feb. 2023, doi: 10.32524/jusitik.v6i1.790.

[3] A. I. Sang, E. Sutoyo, and I. Darmawan, "Analisis Data Mining Untuk Klasifikasi Data Kualitas Udara Dki Jakarta Menggunakan Algoritma Decision Tree Dan Support Vector Machine Data Minning Analysis For Classification Of Air Quality Data Dki Jakarta Using Decision Tree Algorthm And Support Vector Machiner Algorithm," *e-Proceeding of Engineering*, vol. 8, Oct. 2021.

[4] I. Irwansyah, A. D. Wiranata, and T. T. M, "Komparasi Algoritma Decision Tree, Naive Bayes Dan K-Nearest Neighbor Untuk Menentukan Kualitas Udara Di Provinsi Dki Jakarta," *Infotech: Journal of Technology Information*, vol. 9, no. 2, pp. 193–198, Nov. 2023, doi: 10.37365/jti.v9i2.203.

[5] S. Syihabuddin Azmil Umri, "Analisis Dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara Di Dki Jakarta," *JIKO (Jurnal Informatika dan Komputer)*, vol. 4, no. 2, pp. 98–104, Aug. 2021, doi: 10.33387/jiko.v4i2.2871.

[6] Kirono Hendrie Aziz Abdul, Asror Ibnu, and Wibowo Arie Firdaus Yanuar, "Klasifikasi Tingkat Kualitas Udara Dki Jakarta Menggunakan Algoritma Naive Bayes," *e-Proceeding of Engineering*, vol. 9, pp. 1962–1969, Jun. 2022.

[7] B. Nakulo, I. D. Sari, and D. Hariyadi, "Pemantauan Sistem Kualitas Udara Menggunakan Openhab," *Indonesian Journal of Business Intelligence (IJUBI)*, vol. 3, no. 1, p. 14, Jul. 2020, doi: 10.21927/ijubi.v3i1.1203.

[8] R. C. Irjayana, A. Fadlil, and R. Umar, "Pengaruh Seleksi Fitur Terhadap Akurasi Klasifikasi Indeks Standar Pencemar Udara Menggunakan Naïve Bayes," *Informatics and security*, vol. 11, no. 1, 2025, [Online]. Available: https://satudata.jakarta.go.id/home

[9] T. Lidia Putri and R. Danar Dana, "Penerapan Data Mining Pada Clustering Data Harga Rumah Dki Jakarta Menggunakan Algoritmak-Means," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 1, pp. 1174–1179, Mar. 2024, doi: 10.36040/jati.v8i1.8957.

[10] A. Rifqi and R. T. Aldisa, "Penerapan Data Mining Untuk Clustering Kualitas Udara," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 5, no. 2, p. 289, Dec. 2023, doi: 10.30865/json.v5i2.7145.

[11] F. Febriansyah, "Penerapan Algoritma K-Means Clustering Data Gizi Balita Pada Uptd Puskesmas Bumi Agung," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4923.

[12] M. Mujibulloh, M. Martanto, and U. Hayati, "Clustering Produk Ekspor Indonesia Berdasarkan Tingkat Permintaan Menggunakan Metode K-Means Tahun 2020-2022," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 6, pp. 3580–3586, Feb. 2024, doi: 10.36040/jati.v7i6.8254.

[13] N. Wisna and M. Cahaya Rani, "Algoritma K-Means Clustering Analisis Rasio Aktivitas Menggunakan Python," *Jurnal Ekonomi, Bisnis dan Manajemen*, vol. 10, no. 2, Jun. 2023, doi: 10.36987/ecobi.v10i2.

[14] A. Yahya and R. Kurniawan, "Implementasi Algoritma K-Means untuk Pengelompokan Data Penjualan Berdasarkan Pola Penjualan," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 1, pp. 350–358, Jan. 2025, doi: 10.57152/malcom.v5i1.1773.

[15] A. Wulandari, Irmayansyah, and L. T. Ningrum, "Penerapan Metode K-Means Clustering Untuk Pemetaan Perbaikan Jalan Di Kota Bogor," *Jurnal Informatika Teknologi dan Sains (Jinteks)*, vol. 7, no. 1, pp. 107–116, Feb. 2025, doi: 10.51401/jinteks.v7i1.5323.

[16] P. Vania and B. Nurina Sari, "Perbandingan Metode Elbow dan Silhouette untuk Penentuan Jumlah Klaster yang Optimal pada Clustering Produksi Padi menggunakan Algoritma K-Means," *Jurnal Ilmiah Wahana Pendidikan*, vol. 9, no. 21, pp. 547–558, 2023, doi: 10.5281/zenodo.10081332.