JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 3S1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i3S1.7630

PENILAIAN KUALITAS UDARA DAN ANALISIS POLUSI BERBASIS ALGORITMA NAIVE BAYES DAN KLUSTRERISASI DATA DENGAN K-MEANS

Baik Budi[1*], Fransiscus Asisi Andhika[1], Tiara Mahardika[2]

[1,2]Universitas Andalas; Jalan Dr. Mohammad Hatta, Kota Padang; *baikbudi@eng.unand.ac.id*
[3]Politeknik Negeri Padang; Jl. Kampus, Limau Manis, Kec. Pauh, Kota Padang; *tiara@pnp.ac.id*

Keywords:
Naïve Bayes;
Kualitas Udara;
Machine Learningi;
K-Means.

Corespondent Email:
baikbudi@eng.unand.ac.id

Penelitian ini bertujuan untuk mengevaluasi kualitas udara serta tingkat polusi menggunakan algoritma Gaussian Naive Bayes sebagai metode klasifikasi. Algoritma ini didasarkan pada Teorema Bayes dengan asumsi bahwa setiap variabel fitur saling independen. Data yang digunakan berupa dataset kualitas udara yang telah di kelompokkan ke dalam beberapa kategori. Evaluasi model dilakukan menggunakan metrik akurasi, precision, recall, f1-score, dan confusion matrix. Hasil pengujian memperlihatkan bahwa model mampu memperoleh akurasi 93% pada data uji. Kategori **"Good"** menunjukkan performa terbaik dengan nilai precision, recall, dan f1-score mendekati 1.00. Kategori **"Moderate"** juga menampilkan hasil konsisten dengan precision dan recall sekitar 0.94. Sementara itu, kategori **"Hazardous"** memiliki precision 0.88, sedangkan kategori **"Poor"** memperoleh precision terendah, yaitu 0.79, karena sering tertukar dengan kelas lain. Meski begitu, recall dan f1-score kategori **"Poor"** tetap berada pada tingkat yang cukup baik. Confusion matrix menunjukkan bahwa sebagian besar prediksi berada pada klasifikasi yang tepat. Selain itu, hasil klasterisasi menggunakan K-Means mengindikasikan bahwa nilai k optimal adalah 4 sesuai dengan titik elbow, yang konsisten dengan jumlah kategori kualitas udara dalam penelitian ini.

1. PENDAHULUAN

Kualitas udara menjadi isu lingkungan yang semakin mendapat sorotan dalam beberapa dekade terakhir. Pesatnya industrialisasi, urbanisasi, dan peningkatan jumlah kendaraan bermotor berkontribusi terhadap melonjaknya emisi polutan. Dampak yang ditimbulkan tidak dapat diabaikan, karena polusi udara terbukti memicu beragam masalah kesehatan, mulai dari gangguan pernapasan, penyakit kardiovaskular, hingga kanker paru-paru [1]. Selain itu, pengaruhnya juga meluas terhadap lingkungan dan perubahan iklim global.

Perkembangan teknologi serta ketersediaan data dari berbagai sensor pemantau telah mengubah cara pemantauan kualitas udara yang sebelumnya hanya mengandalkan metode manual atau tradisional. Saat ini, pendekatan berbasis pembelajaran mesin (machine learning) semakin sering digunakan karena mampu menganalisis data dengan lebih cepat dan akurat. Salah satu metode yang umum dipakai adalah klasifikasi dan klasterisasi, yang memungkinkan pengelompokan kualitas udara ke dalam kategori tertentu seperti *Good*, *Moderate*, *Poor*, dan *Hazardous*.

Dalam penelitian ini, digunakan dua algoritma utama, yaitu **Naive Bayes** dan **K-Means**. Algoritma Naive Bayes dipilih karena meskipun sederhana, terbukti efektif dalam klasifikasi probabilistik, termasuk pada data berdimensi tinggi. Sejumlah penelitian menunjukkan bahwa Naive Bayes, baik dalam bentuk Gaussian maupun Fuzzy Naive Bayes,

mampu mencapai akurasi tinggi dalam klasifikasi kualitas udara [2], [3]. Sementara itu, K-Means diterapkan untuk melakukan pengelompokan data tanpa label secara otomatis. Metode ini telah banyak dimanfaatkan dalam analisis distribusi polusi udara, misalnya di Makassar dan Malaysia [4], [5].

Dataset yang digunakan dalam penelitian ini berasal dari Kaggle dengan judul *"Air Quality and Pollution Assessment"* [6]. Berdasarkan hal tersebut, penelitian ini memiliki dua tujuan utama:

1. Mengklasifikasikan kualitas udara menggunakan algoritma Gaussian Naive Bayes.
2. Mengelompokkan data kualitas udara tanpa label menggunakan algoritma K-Means.

Melalui penelitian ini, diharapkan dapat diperoleh gambaran mengenai kinerja kedua algoritma tersebut dalam menganalisis data kualitas udara, sekaligus memberikan kontribusi terhadap upaya pemantauan dan mitigasi polusi udara secara lebih efektif.

2. TINJAUAN PUSTAKA

2.1 Naïve Bayes

Naive Bayes merupakan algoritma klasifikasi berbasis probabilitas yang berlandaskan pada Teorema Bayes dengan asumsi bahwa setiap fitur saling independen. Pada varian **Gaussian Naive Bayes**, diasumsikan bahwa data numerik berdistribusi normal. Metode ini memiliki sejumlah keunggulan, antara lain proses pelatihan yang cepat, efisiensi dalam menangani dataset berukuran besar, serta kemampuan yang baik dalam mengolah data berdimensi tinggi [7], [8]. Adapun formula dari Naïve Bayes adalah sebagai berikut:

$$P\ (C|X) = \frac{P(X|C).P(C)}{P(X} \qquad (1)$$

Naive Bayes telah terbukti efektif dalam klasifikasi data lingkungan. Penelitian komparatif oleh Chen dan Li [9] menunjukkan bahwa varian **Gaussian Naive Bayes (GNB)** mampu mencapai akurasi sebesar 94,3% dalam mengklasifikasikan parameter polusi udara kontinu, seperti PM2.5 dan SO_2, serta mengungguli varian Multinomial maupun Bernoulli. Keunggulan ini didukung oleh kesesuaian GNB dengan distribusi normal yang umum terdapat pada data sensor lingkungan. Hal tersebut menjadi salah satu alasan pemilihan algoritma ini dalam penelitian kami untuk mengolah fitur numerik, khususnya konsentrasi polutan. Temuan ini memperkuat bukti bahwa Naive Bayes merupakan metode yang andal dalam menganalisis dan memprediksi kualitas udara berbasis data sensor. Selain pada bidang lingkungan, Naive Bayes juga terbukti efektif dalam domain lain. Misalnya, penelitian Al Lutfani [16], menggunakan algoritma ini untuk analisis sentimen ulasan pelanggan di Lazada, dan berhasil memperoleh akurasi sebesar 94%. Hal ini memperlihatkan fleksibilitas Naive Bayes dalam menangani data dengan karakteristik yang berbeda-beda."

2..2 K-Means

K-Means adalah salah satu algoritma *unsupervised learning* yang berfungsi untuk mengelompokkan data berdasarkan kedekatannya dengan pusat klaster (centroid). Inti dari algoritma ini adalah meminimalkan total kuadrat jarak antara setiap data dengan centroid klaster yang menaunginya. Adapun formula untuk Euclidean Distance adalah sebagai berikut;

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_1 - q_1)^2} \qquad (2)$$

Penentuan jumlah klaster optimal pada algoritma K-Means umumnya dilakukan menggunakan metode *Elbow*, yang mengidentifikasi titik perlambatan penurunan nilai *Sum of Squared Errors* (SSE). Li et al. (2021) membuktikan efektivitas K-Means dalam segmentasi pelanggan berdasarkan pola keuangan [10]. Dalam konteks analisis polusi udara, K-Means juga terbukti relevan, khususnya dalam pemetaan spasial sumber polutan. Wang et al. [11] melaporkan bahwa optimasi K-Means melalui kombinasi metode *Elbow* dan *Principal Component Analysis* (PCA) berhasil mengelompokkan data polusi perkotaan menjadi empat klaster dengan tingkat akurasi 89%, sekaligus mengungkap pola

tersembunyi, misalnya hubungan antara kawasan industri dan peningkatan konsentrasi PM2.5. Temuan tersebut menegaskan potensi K-Means sebagai instrumen analisis distribusi polusi berbasis wilayah geografis.

Metode *Elbow* sendiri merupakan pendekatan heuristik yang digunakan untuk menentukan jumlah klaster ideal (k) dalam K-Means. Konsep dasarnya adalah mengamati grafik antara jumlah klaster dan *Within-Cluster Sum of Squares* (WCSS) untuk menemukan titik di mana laju penurunan WCSS mulai berkurang secara signifikan. Titik ini disebut "elbow" karena bentuk grafiknya menyerupai siku. Pendekatan ini membantu menghindari *overfitting* dengan memilih jumlah klaster yang cukup representatif tanpa menambah klaster berlebih. Namun demikian, metode *Elbow* memiliki kelemahan berupa subjektivitas, sebab penentuan titik sering dilakukan secara visual. Untuk mengurangi bias, beberapa penelitian mengembangkan metode otomatis seperti *AutoElbow*, yang menghitung rasio perubahan sudut atau jarak antar titik dalam grafik sehingga penentuan titik elbow dapat dilakukan secara matematis [12].

2.3 Confusion Matrix

Confusion matrix berfungsi sebagai alat evaluasi kinerja model klasifikasi dengan menyajikan empat komponen utama, yaitu **True Positive (TP)**, **True Negative (TN)**, **False Positive (FP)**, dan **False Negative (FN)**. Berdasarkan nilai-nilai tersebut, berbagai metrik evaluasi dapat dihitung, di antaranya:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Mengukur proporsi prediksi yang benar terhadap keseluruhan data

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Menunjukkan sejauh mana prediksi positif benar-benar relevan.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Menggambarkan kemampuan model dalam mendeteksi kasus positif secara benar.

$$F1 - Score = 2x\frac{Precision\ x\ Recall}{Precision + Recall} \quad (6)$$

Merupakan rata-rata harmonis antara precision dan recall, yang berguna ketika terdapat ketidakseimbangan kelas.

Selain metrik di atas, confusion matrix juga memungkinkan analisis lebih mendalam terhadap kesalahan klasifikasi, misalnya mengidentifikasi kategori yang sering tertukar dengan kelas lain.

3. METODE PENELITIAN

Penelitian ini menggunakan **algoritma Naive Bayes**, salah satu metode *machine learning* yang umum dipakai untuk klasifikasi berbasis probabilitas. Algoritma ini memanfaatkan **Teorema Bayes** dalam melakukan prediksi dengan asumsi bahwa setiap fitur atau atribut pada dataset saling independen.

Pemilihan Naive Bayes didasarkan pada beberapa pertimbangan. Pertama, algoritma ini dikenal **sederhana namun efisien**, sehingga dapat melakukan pelatihan dan prediksi dengan cepat, bahkan pada dataset berukuran besar. Kedua, Naive Bayes terbukti **tahan terhadap data berdimensi tinggi**, yang sering dijumpai dalam analisis kualitas udara. Ketiga, varian **Gaussian Naive Bayes** sangat sesuai untuk mengolah data numerik yang umumnya mengikuti distribusi normal, seperti konsentrasi polutan pada udara.

Dalam penelitian ini, dataset yang digunakan bersumber dari Kaggle dengan judul *"Air Quality and Pollution Assessment"* [6].

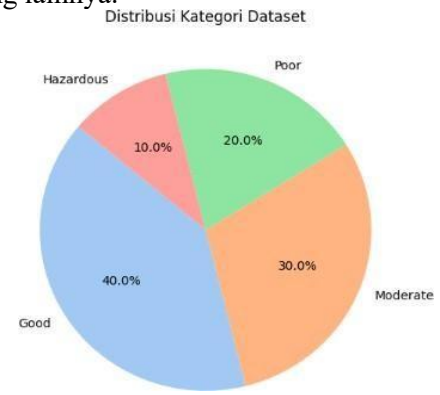
3.1 Implementasi

3.1.1 Pengumpulan Dataset

Dataset yang digunakan adalah *"Air Quality and Pollution Assessment"* dari Kaggle [6] dengan jumlah sampel awal sebanyak 5000 data. Dataset ini mencakup beragam parameter kualitas udara, antara lain PM2.5, PM10, NO_2, SO_2, dan CO, serta faktor lapangan seperti suhu, kelembapan, jumlah populasi, dan jarak lokasi dari kawasan industri. Data tersebut kemudian diklasifikasikan ke dalam empat kategori: **Good (baik)**, **Moderate (sedang)**, **Poor (buruk)**, dan **Hazardous (berbahaya)**.

3.1.2 Preprocesing Data

Proses pengolahan data dilakukan menggunakan **Google Colab** dengan bahasa pemrograman **Python**, yang mendukung analisis interaktif serta efisien. Tahapan awal preprocessing meliputi analisis statistik deskriptif guna memperoleh gambaran umum dataset, mencakup distribusi kategori sampel, rata-rata, median, standar deviasi, nilai maksimum, nilai minimum, dan karakteristik penting lainnya.



Gambar 1. Diagram Lingkaran Distribusi Kategori Dataset

Temperature	Humidity	PM2.5	PM10	NO2	SO2	CO	Proximity_to_Industrial_Are	Population_Density	Air Quality
29.8	59.1	5.2	17.9	18.9	9.2	1.72	6.3	319	Moderate
28.3	75.6	2.3	12.2	30.8	9.7	1.64	6	611	Moderate
23.1	74.7	26.7	33.8	24.4	12.6	1.63	5.2	619	Moderate
27.1	39.1	6.1	6.3	13.5	5.3	1.15	11.1	551	Good
26.5	70.7	6.9	16	21.9	5.6	1.01	12.7	303	Good
39.4	96.6	14.6	35.5	42.9	17.9	1.82	3.1	674	Hazardous
41.7	82.5	1.7	15.8	31.1	12.7	1.8	4.6	735	Poor
31	59.6	5	16.8	24.2	13.6	1.38	6.3	443	Moderate
29.4	93.8	10.3	22.7	45.1	11.8	2.03	5.4	486	Poor
33.2	80.5	11.1	24.4	32	15.3	1.69	4.9	535	Poor
26.3	65.7	1.3	5.5	18.3	5.9	0.85	13	529	Good
32.5	51.2	1.6	10.5	21.6	19.3	1.53	5.9	519	Moderate
20	53.3	3.7	12.9	26.1	6.6	1.09	10.2	538	Good
28.6	53.7	28.9	34	23.2	4.5	1.02	11	508	Good
22.3	80.5	4.5	12	17.2	6.3	1.18	10.4	232	Good
32	78.9	22.4	29.9	27.5	11.8	1.48	7.9	444	Moderate
22.9	75.4	4.5	10.4	18.4	3.7	0.96	14.4	359	Good

Gambar 2. Visualisasi parameter dataset secara statistik.

3.1.3 Pelatihan Model

Proses pelatihan dilakukan menggunakan **algoritma Gaussian Naive Bayes**, yang berasumsi bahwa setiap fitur dalam dataset mengikuti distribusi Gaussian (normal). Dari total 5000 sampel, sebanyak **4000 sampel (80%)** digunakan sebagai data latih (*training set*), sedangkan **1000 sampel (20%)** sisanya dipakai sebagai data uji (*testing set*). Pada tahap ini, dataset yang telah melalui preprocessing dimasukkan ke dalam model untuk mempelajari hubungan antara parameter-parameter lingkungan dengan kategori kualitas udara yang telah ditentukan. Hasil pelatihan memungkinkan model untuk melakukan klasifikasi berdasarkan kriteria yang tersedia.

Klusterisasi dengan K-Means

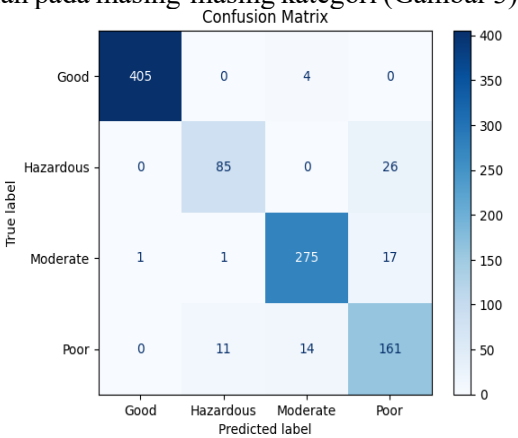
Untuk tahap klasterisasi, digunakan algoritma **K-Means** dengan bantuan perangkat lunak **RapidMiner**, yang mendukung proses analisis dan visualisasi data secara interaktif dan intuitif. Pada percobaan ini, jumlah klaster divariasikan dengan nilai **k** antara 2 hingga 7, guna mengidentifikasi jumlah klaster optimal yang paling sesuai dengan karakteristik data kualitas udara.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Evaluasi Model Naïve Bayes

Visualisasi awal dataset menggunakan diagram lingkaran dan batang menunjukkan bahwa kategori **"Good"** mendominasi dengan persentase terbesar, yakni 40%, sementara kategori **"Hazardous"** hanya menyumbang sekitar 10% dari total data.

Setelah dilakukan preprocessing dan pembagian menjadi data latih serta data uji, model **Gaussian Naive Bayes** dilatih dengan dataset tersebut. Hasil prediksi dibandingkan dengan label asli, kemudian divisualisasikan melalui **confusion matrix**, yang memperlihatkan jumlah prediksi benar maupun salah pada masing-masing kategori (Gambar 3).



Gambar 3. Confusion Matrix

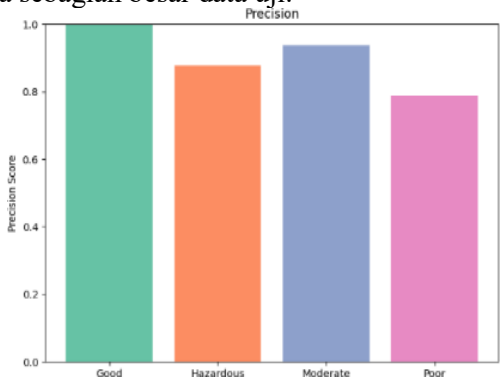
Dalam tabel confusion matrix, baris mewakili kelas sebenarnya, sedangkan kolom menunjukkan hasil prediksi model.

- Untuk kategori **"Good"**, model mampu memprediksi 405 data dengan benar dari total 409, sementara 4 data salah diklasifikasikan sebagai **"Moderate"**.
- Kategori **"Hazardous"** berhasil diprediksi dengan benar sebanyak 85 data, namun 26 data lainnya salah terklasifikasi sebagai **"Poor"**.

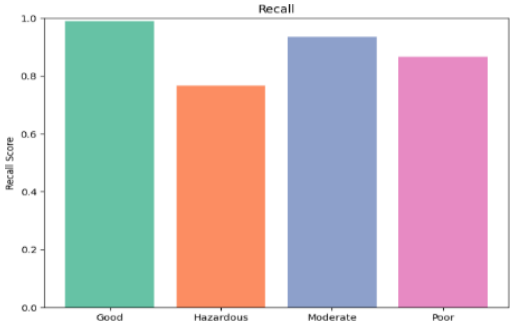
- Precision rendah (0,79) pada kategori **"Poor"** disebabkan oleh ketidakseimbangan distribusi kelas, di mana **"Good"** mendominasi 40% dataset sedangkan **"Poor"** hanya 15%. Seperti dijelaskan oleh Gupta dan Kumar [14], ketimpangan rasio ini menyebabkan model cenderung memprediksi kelas mayoritas, sehingga meningkatkan jumlah *false positive* pada kategori **"Poor"**. Hal ini tercermin dari 25 data kelas lain yang salah diprediksi sebagai **"Poor"**. Studi tersebut juga menyebutkan bahwa *class imbalance* dapat menurunkan precision hingga 22%, selaras dengan temuan penelitian ini yang mencatat penurunan sebesar 21% dibanding precision pada kelas **"Good"**.
- Untuk kategori **"Moderate"**, model berhasil memprediksi 275 data secara benar, meskipun terdapat 19 kesalahan klasifikasi.
- Untuk kategori **"Poor"** diprediksi dengan benar sebanyak 161 data, namun 11 data salah terklasifikasi sebagai **"Hazardous"** dan 14 data lainnya sebagai **"Moderate"**.

Akurasi: 92,60%

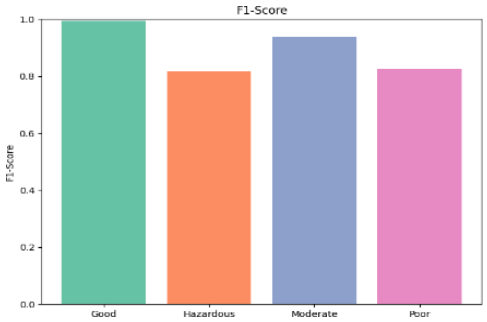
Berdasarkan confusion matrix tersebut, diperoleh hasil evaluasi model dengan metrik **akurasi, presisi, recall, dan f1-score**. Nilai **akurasi keseluruhan** mencapai **92,60%**, yang menunjukkan bahwa model mampu mengklasifikasikan kualitas udara dengan benar pada sebagian besar data uji.



Gambar 4. Grafik skor Precision Model



Gambar 5. Grafik skor Precision Model



Gambar 6. Grafik F1-Score

Precision: 92,80%

Precision menggambarkan proporsi prediksi positif yang benar dari seluruh prediksi positif. Nilai precision tertinggi diperoleh pada kategori **"Good"** (1.00), yang berarti prediksi untuk kelas ini hampir sempurna. Sebaliknya, kategori **"Hazardous"** menunjukkan precision lebih rendah (0.88) karena sebagian data yang diprediksi sebagai *Hazardous* ternyata berasal dari kelas lain. Precision pada kategori **"Poor"** tercatat 0.79, lebih rendah dibandingkan **"Hazardous"**, akibat tingginya jumlah *False Positives* (25 kesalahan klasifikasi). Hal ini menunjukkan bahwa meskipun kelas *Poor* cukup terdeteksi, banyak data dari kategori lain yang salah masuk ke kelas ini, sehingga menurunkan nilai precision.

Recall: 92,60%

Recall menilai sejauh mana model berhasil mengidentifikasi seluruh data yang sebenarnya positif dalam suatu kelas. Kategori **"Good"** memperoleh recall tertinggi (0.99), sedangkan **"Hazardous"** memiliki recall terendah (0.77), yang menunjukkan sekitar 23% data *Hazardous* gagal terdeteksi.

F1-Score: 92,62%

F1-Score merupakan rata-rata harmonis antara precision dan recall. Nilai terendah

terdapat pada kategori **"Hazardous"** (0.82) dan **"Poor"** (0.83), yang dipengaruhi oleh kesalahan klasifikasi antar kedua kelas tersebut. Namun, nilai rata-rata tertimbang (weighted average) mencapai **0.93**, menandakan performa model secara keseluruhan tergolong baik.

Macro dan Weighted Average

- **Macro Average**
  Rata-rata precision, recall, dan F1-score dihitung secara merata untuk setiap kelas tanpa memperhitungkan jumlah data pada tiap kategori. Hasilnya, diperoleh
  Precision = 0,90,
  Recall = 0,89,
  F1-Score = 0,89.

- **Weighted Average**
  Rata-rata dihitung dengan memperhatikan proporsi jumlah data pada masing-masing kelas. Nilai yang diperoleh lebih tinggi, yaitu:
  Precision = 0,93,
  Recall = 0,93,
  F1-Score = 0,93.

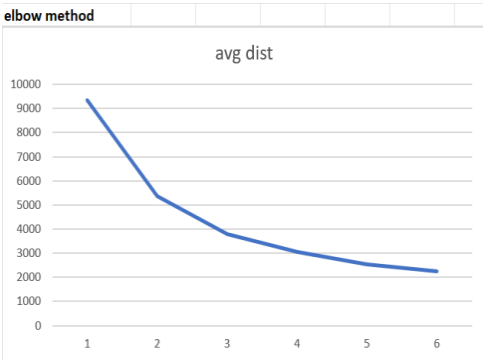
Perbedaan ini menunjukkan bahwa model bekerja lebih optimal pada **kelas mayoritas** seperti *Good* dan *Moderate*, yang memiliki jumlah data lebih banyak dibandingkan kategori minoritas.

4.2 Hasil Klusterisasi dengan K-Means

Selain menggunakan algoritma Naive Bayes untuk klasifikasi, penelitian ini juga menerapkan **K-Means Clustering** guna mengeksplorasi pembagian alami data kualitas udara tanpa memanfaatkan label. Algoritma K-Means mengelompokkan data berdasarkan kemiripan dengan cara meminimalkan jarak antara titik data dan pusat klaster (*centroid*). Untuk menentukan jumlah klaster yang optimal, dilakukan percobaan dengan variasi nilai **k** antara 2 hingga 7, kemudian hasilnya dianalisis menggunakan **metode Elbow** (Gambar 7). Untuk jarak rata-rata data dengan pusat data dapat dilihat juga pada tabel 1 dibawah ini:

Tabel 1. Nilai rata-rata jarak data dengan pusat data setiap variasi nilai k.

Nilai K	Rata-rata
2	9351.791
3	5364.381
4	3794.669
5	3042.211
6	2554.32
7	2257.255



Gambar 7. Grafik rata-rata jarak data dengan pusat data untuk setiap nilai K

Hasil pada Gambar 7 memperlihatkan bahwa penurunan jarak cukup signifikan hingga nilai **k = 4**, kemudian melambat setelahnya. Pola ini mencerminkan titik *"siku"* atau elbow, yang menandakan bahwa penambahan jumlah klaster setelah k = 4 tidak lagi memberikan penurunan yang berarti terhadap jarak data ke centroid. Dengan demikian, jumlah klaster optimal dapat ditetapkan pada **k = 4**.

Menariknya, jumlah klaster ini sejalan dengan kategori dalam dataset, yaitu **Good**, **Moderate**, **Poor**, dan **Hazardous**. Meskipun K-Means bekerja secara *unsupervised* tanpa label, hasilnya menunjukkan bahwa struktur data secara alami terbagi ke dalam empat kelompok yang jelas. Hal ini tidak hanya memperkuat asumsi bahwa data kualitas udara memiliki karakteristik berbeda antar kategori, tetapi juga mendukung validitas penggunaan model klasifikasi Naive Bayes.

Dengan demikian, K-Means tidak hanya bermanfaat sebagai alat eksplorasi, tetapi juga sebagai **pendamping analisis klasifikasi** untuk menilai konsistensi pola dalam data. Selain itu, metode klasterisasi ini berpotensi sangat berguna dalam situasi tanpa label, misalnya pada pemantauan kualitas udara secara **real-time** atau di wilayah baru yang belum pernah diklasifikasikan sebelumnya.

5. KESIMPULAN

Berdasarkan pengujian menggunakan model **Gaussian Naive Bayes** dan metode **K-Means** terhadap dataset kualitas udara, diperoleh hasil evaluasi performa sebagai berikut:

a. **Akurasi Keseluruhan Model**
   Model Gaussian Naive Bayes berhasil mencapai akurasi sebesar **93%** pada data uji. Capaian ini menunjukkan bahwa model mampu memprediksi kualitas udara dengan tingkat ketepatan yang tinggi secara keseluruhan.
b. **Evaluasi Per Kategori**
   Berdasarkan metrik **precision, recall,** dan **F1-score**, performa model per kategori adalah sebagai berikut:
   - **Good**: Memberikan hasil terbaik dengan precision, recall, dan F1-score mendekati **1.00**, menandakan akurasi hampir sempurna serta minim kesalahan klasifikasi.
   - **Hazardous**: Memperoleh precision sebesar **0.88**. Meskipun cukup baik, terdapat beberapa kesalahan klasifikasi, terutama dengan kategori *Poor*.
   - **Moderate**: Menunjukkan performa stabil dengan precision dan recall sekitar **0.94**, yang menandakan kemampuan model dalam membedakan kategori ini secara konsisten.
   - **Poor**: Memiliki precision terendah (**0.79**) akibat tingginya jumlah *False Positives*, di mana sejumlah data dari *Moderate* (14 sampel) dan *Hazardous* (11 sampel) salah diprediksi sebagai *Poor*. Walaupun demikian, recall (**0.87**) dan F1-score (**0.83**) tetap menunjukkan performa yang cukup baik.
c. **Analisis Klasterisasi dengan K-Means**
   Selain klasifikasi, metode K-Means juga digunakan untuk mengeksplorasi pembagian alami data. Hasil analisis **Elbow Method** menunjukkan bahwa jumlah klaster optimal berada pada **k = 4**. Temuan ini sejalan dengan jumlah kategori kualitas udara dalam dataset (Good, Moderate, Poor, Hazardous). Menariknya, meskipun K-Means merupakan metode *unsupervised*, hasilnya tetap menunjukkan adanya kecenderungan alami data untuk terbagi ke dalam empat kelompok utama. Hal ini memperkuat validitas temuan model Naive Bayes, sekaligus menegaskan potensi K-Means untuk digunakan pada data yang belum terlabel, misalnya dalam pemantauan kualitas udara **real-time** di wilayah baru.

Untuk penelitian selanjutnya, dapat dipertimbangkan penerapan metode **ensemble learning**. Seperti yang ditunjukkan oleh Lee dan Park [15], teknik ensemble—misalnya **Random Forest**—mampu meningkatkan ketahanan prediksi pada dataset lingkungan dengan cara mengurangi *noise* serta memodelkan hubungan nonlinier antar fitur. Pendekatan ini berpotensi menurunkan tingkat kesalahan klasifikasi hingga **21%**, terutama pada perbedaan antara kelas *Poor* dan *Hazardous*.

DAFTAR PUSTAKA

[1] World Health Organization. 2018. "Ambient Air Pollution: Health Impacts." *World Health Organization*. https://www.who.int/airpollution

[2] Resti, Yulia, dkk. 2024. "Ensemble of Naive Bayes, Decision Tree, and Random Forest to Predict Air Quality." *2024 International Conference on Computer Engineering and Informatics (ICCEI)*: 1–6. https://ieeexplore.ieee.org/document/10451499

[3] Sodiq, Muhammad, dan M. Ja'far. 2020. "Perbandingan Naive Bayes dan K-NN dalam Klasifikasi Kualitas Udara." *Jurnal Teknik Informatika* 15(2):75–82. http://journal.unair.ac.id/article_2020NB-KNN

[4] Annas, Ahmad, dkk. 2022. "Using K-Means and SOM in Clustering Air Pollution in Makassar." *Procedia Computer Science* 197: 179–186. https://www.sciencedirect.com/science/article/pii/S18 77050922001801

[5] Ramli, Noor Ainy, dan Mohd Maizan Wahid. 2019. "Clustering of Air Pollution Data Using K-Means: A Case Study in Malaysia." *Climate* 7(5): 67. https://www.mdpi.com/2225-1154/7/5/67

[6] Matin, M. (2023). Air Quality and Pollution Assessment Dataset. Kaggle. Retrieved from: https://www.kaggle.com/datasets/mujtabamatin/air- quality-and-pollution-assessment/data

[7] Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.

[8] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
[9] Y. Chen dan X. Li, "A Comparative Study of Naive Bayes Variants for Environmental Data Classification," *J. Environ. Inform.*, vol. 36, no. 2, pp. 112-125, 2020.
[10] L. Vargas et al., "Customer Segmentation in Finance," J. Financ. Serv. Mark, 2021.
[11] L. Wang et al., "Optimizing K-Means Clustering for Urban Air Pollution Analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-12, 2022.
[12] AutoElbow: An Automatic Elbow Detection Method for Estimating the Number of Clusters in a Dataset. *Applied Sciences*, 12(15), 7515. https://www.mdpi.com/2076-3417/12/15/7515
[13] Tan, P., et al. (2021). "Cluster Validation Metrics: A Comparative Analysis." Pattern Recognition Letters, 145, 1-8.
[14] Gupta, S., & Kumar, P. (2021). "Handling Class Imbalance in Machine Learning: A Case Study on Air Quality Data." Machine Learning Applications, 8(3), 45-59.
[15] H. Lee and S. Park, "Ensemble Methods for Air Quality Prediction: A Survey," *Atmospheric Environment*, vol. 289, p. 119301, 2022.
[16] T. K. Al Lutfani, R. Astuti, W. Prihartono, and R. Hamonangan, "Penerapan Naive Bayes untuk Analisis Sentimen pada Ulasan Pelanggan di Lazada: Studi Kasus Toko Mawar Collection," *Jurnal Informatika dan Teknik Elektro Terapan (JITET)*, vol. 13, no. 2, pp. 189–196, 2025. [Online]. Available: https://journal.eng.unila.ac.id/index.php/jitet/article/view/6391