ANALISIS KUALITAS UDARA DAN PREDIKSI POLUSI MENGGUNAKAN K-NEAREST NEIGHBORS (KNN)

ANISA HAYATULLAH

ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024

DAFTAR ISI

DAFTAR GAMBAR

I. PENDAHULUAN

1.1 Latar Belakang

Kualitas udara yang buruk telah menjadi perhatian utama di seluruh dunia karena dampaknya yang merugikan terhadap kesehatan manusia dan lingkungan. Salah satu dampak yang dapat dirasakan dari kualitas udara yang buruk yaitu polusi atau pencemaran udara. Pencemaran udara ini dapat menyebabkan berbagai masalah kesehatan, mulai dari gangguan pernapasan hingga penyakit jantung dan kanker, serta dapat memperburuk kondisi lingkungan seperti pemanasan global dan kerusakan ekosistem (Maharani S dan Aryanta WR).

Menurut Dinas Lingkungan Hidup dan Kebersihan Provinsi Banten (2023), polusi udara memiliki dampak yang signifikan terhadap kesehatan masyarakat karena udara kotor tersebut mengandung berbagai zat yang berbahaya. Di Indonesia, tingkat darurat polusi udara masih di tahap mengkhawatirkan. Mengutip dari IQAir, Indonesia per 2023 berada di peringkat 14 di dunia sebagai negara dengan tingkat polusi udara terparah di dunia. Peringkat ini sejalan dengan nilai Air Quality Index (AQI) Indonesia yang sangat tinggi, yaitu sebesar 105 mikrogram per meter kubik. Dengan kata lain, konsentrasi rata-rata PM2.5 di Indonesia pada tahun 2023 adalah 7,4 kali lipat dari nilai panduan kualitas udara tahunan yang ditetapkan oleh Organisasi Kesehatan Dunia (WHO).

Berbagai Usaha dilakukan pemerintah Indonesia dalam menangani permasalahan ini. Koalisi IBUKOTA (Koalisi Inisiatif Bersihkan Udara Koalisi Semesta) mengajukan permintaan untuk melakukan revisi baku mutu udara ambien (BMUA) nasional yang sesuai dengan standar WHO dan meminta pemerintah untuk memberikan upaya konkret dalam menangani polusi udata seperti alat ukur polusi sesuai dengan standar ahli dan memberikan informasi tentang kualitas udara Jakarta secara real-time kepada warga.

Selain koalisi IBUKOTA, Indeks Standar Pencemar Udara (ISPU) juga salah satu tindakan pemerintah melalui Kementerian Lingkungan Hidup dan Kehutanan (KLHK) untuk menghasilkan informasi mutu udara yang tepat dan akurat. ISPU disampaikan dalam bentuk hasil pemantauan mutu udara dari stasiun pemantauan otomatis kontinu yang dimiliki KLHK.

Dalam upaya untuk memantau, memahami dan menyampaikan informasi mengenai kualitas udara, berbagai metode dan model analisis telah dikembangkan. Salah satu pendekatan yang banyak digunakan adalah menggunakan Indeks Kualitas Udara (Air Quality Index/AQI) yang mengukur tingkat pencemaran udara dan memberikan penilaian terhadap risiko kesehatan yang terkait. Namun, masalah utama yang dihadapi adalah keterbatasan stasiun pemantauan udara, terutama di area yang luas dan terpencil. Untuk mengatasi tantangan ini, pengembangan model prediksi kualitas udara menjadi penting.

Salah satu model yang bisa digunakan adalah K-Nearest Neighbors (KNN), yang merupakan metode pembelajaran mesin yang dapat digunakan untuk melakukan klasifikasi berdasarkan data terdekat. Metode KNN sering digunakan karena sederhana, cepat, mudah dimengerti dan efektif diaplikasikan untuk dataset yang memiliki data

training yang berukuran besar. Dengan lebih banyak data training, metode ini dapat menghasilkan data perhitungan lebih akurat (Afandie, M. N., Cholissodin, I., Supianto, 2014)

Dengan menggunakan model KNN, diharapkan dapat dibangun sistem yang efektif untuk memprediksi kualitas udara dengan akurasi tinggi, sehingga membantu pemerintah dan masyarakat untuk mengambil langkah-langkah preventif yang tepat dalam menjaga kesehatan dan lingkungan dari dampak buruk polusi udara. Oleh karena itu, penelitian mengenai analisis kualitas udara menggunakan model KNN memiliki relevansi yang tinggi dalam konteks perlindungan lingkungan dan kesehatan masyarakat.

1.2 Rumusan masalah

a. Bagaimana mengembangkan model prediksi kualitas udara menggunakan metode K-Nearest Neighbors (KNN) untuk memperkirakan tingkat polusi udara?
b. Seberapa efektif model KNN dalam memprediksi kualitas udara berdasarkan data yang diberikan?

1.3 Tujuan

a. Meningkatkan pemahaman tentang kualitas udara

   Penelitian ini bertujuan untuk memberikan pemahaman yang lebih baik tentang kualitas udara, sehingga membantu mengambil keputusan mengenai langkah preventif berkaitan dengan polusi udata.

b. Mengembangkan sistem prediksi yang efektif

   Tujuan utama penelitian ini adalah untuk mengembangkan sistem prediksi kualitas udara yang efektif menggunakan model KNN.

c. Mendukung pengambilan keputusan

   Penelitian ini bertujuan untuk memberikan dukungan kepada pemerintah, lembaga lingkungan, dan masyarakat umum dalam mengambil keputusan yang berkaitan dengan perlindungan lingkungan dan kesehatan masyarakat berdasarkan informasi yang akurat tentang kualitas udara.

1.4 Manfaat

a. Pengembangan kebijakan lingkungan yang lebih efektif

   Informasi yang dihasilkan dari penelitian ini dapat menjadi dasar untuk pengembangan kebijakan lingkungan yang lebih efektif dalam menangani masalah polusi udara, termasuk dalam revisi baku mutu udara ambien dan strategi pengurangan emisi polutan udara.

b. Peningkatan kesadaran masyarakat

   Penelitian ini juga dapat meningkatkan kesadaran masyarakat tentang pentingnya menjaga kualitas udara dan dampak buruk pencemaran udara terhadap kesehatan dan lingkungan, sehingga masyarakat dapat berperan aktif dalam menjaga lingkungan hidup mereka.

II. METODE PENELITIAN

1.1 Data

Dataset yang digunakan merupakan sebuah dataset yang diimpor dari Kaggle. Adapun link data yaitu World Air Quality Index by City (kaggle.com). Dataset ini terdiri dari informasi mengenai kota-kota di berbagai wilayah dan data tentang tingkat polusi udara di negara-negara di seluruh dunia. Dataset mengandung data tentang tingkat polusi udara di berbagai negara di seluruh dunia. Data ini meliputi berbagai parameter yang digunakan untuk mengukur kualitas udara, seperti tingkat partikulat PM2.5 atau PM10, kadar oksida nitrogen (NO2), karbon monoksida (CO), sulfur dioksida (SO2), dan lain-lain. Informasi ini diperoleh dari berbagai sumber, termasuk badan pemerintah, lembaga penelitian, atau organisasi lingkungan.

Selanjutnya, data polusi udara ini mencakup parameter-parameter penting seperti PM2.5, NO2, CO, dan Ozone. PM2.5 merujuk pada partikulat berukuran sangat kecil yang dapat masuk ke dalam saluran pernapasan manusia dan berpotensi merusak kesehatan. NO2 merupakan gas beracun yang terutama berasal dari aktivitas pembakaran, yang dapat menyebabkan iritasi pada saluran pernapasan dan masalah kesehatan lainnya. CO adalah gas beracun yang dihasilkan dari pembakaran bahan bakar fosil dan dapat mengganggu kemampuan darah untuk membawa oksigen, berpotensi menyebabkan keracunan. Sedangkan Ozone (O3) adalah gas reaktif yang dapat menyebabkan iritasi pada saluran pernapasan dan dapat menyebabkan masalah kesehatan serius ketika terhirup dalam kadar tinggi di permukaan bumi. Integrasi data ini memberikan pemahaman yang lebih lengkap tentang kualitas udara dan potensi dampaknya terhadap kesehatan manusia dan lingkungan. Masing-masing parameter ini diberi nilai untuk setiap kota yang dicantumkan**.**

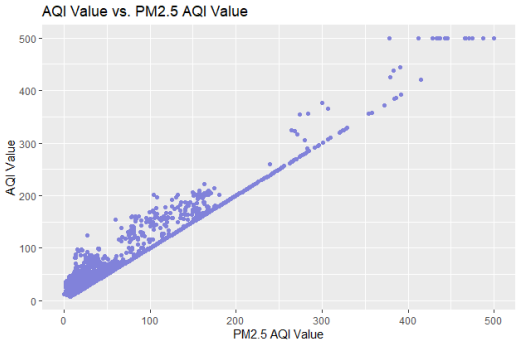
Adapun tingkatan atau kategori dari AQI yang dicantumkan pada data ini terdiri atas 6 tingkatan yaitu:

- Good: menandakan bahwa kualitas udara memuaskan, dan tidak ada atau sedikit risiko kesehatan yang terkait.
- Moderate: menunjukkan bahwa kualitas udara dapat diterima, tetapi mungkin ada kekhawatiran kesehatan sedang bagi sejumlah kecil individu, terutama mereka yang sangat sensitif terhadap polusi udara.
- Unhealthy for Sensitive Groups:Kategori ini menandakan bahwa kualitas udara menjadi perhatian bagi individu yang lebih rentan terhadap efek polusi udara, seperti anak-anak, lansia, dan orang-orang dengan kondisi pernapasan atau kardiovaskular.
- Unhealthy: menunjukkan bahwa populasi umum mungkin mulai mengalami efek kesehatan akibat kualitas udara yang buruk. Disarankan untuk membatasi aktivitas di luar ruangan dan mengambil tindakan pencegahan yang diperlukan.
- Very Unhealthy: menandakan risiko kesehatan yang signifikan akibat polusi udara. Disarankan untuk menghindari aktivitas di luar ruangan dan meminimalkan paparan udara tercemar.
- Hazardous: menunjukkan tingkat polusi udara tertinggi, yang mengancam kesehatan secara serius. Sangat penting untuk tetap berada di dalam ruangan,

menggunakan sistem filtrasi udara, dan mengikuti panduan otoritas lokal untuk melindungi kesehatan.

Adapun efek dari atribut ini pada AQI sebagai berikut:

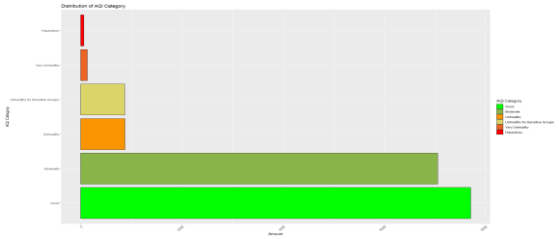
a. Perubahan AQI Value diikuti oleh perubahan PM2.5



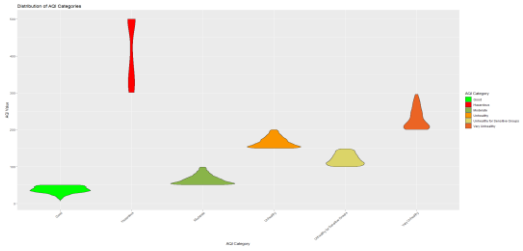
Gambar 1 *AQI Value VS PM2.5 AQI Value*

Gambar 1 menunjukkan hubungan erat antara nilai Indeks Kualitas Udara (AQI) dan konsentrasi partikulat halus (PM2.5). Hubungan positif antara AQI dan PM2.5 ini berarti bahwa setiap kenaikan konsentrasi PM2.5 berpotensi meningkatkan nilai AQI, yang menunjukkan kualitas udara yang semakin tidak sehat. Partikel PM2.5 yang berukuran sangat kecil, sekitar 2.5 mikrometer atau sepersepuluh diameter rambut manusia, dapat menembus jauh ke dalam paru-paru dan aliran darah, membawa berbagai konsekuensi kesehatan yang serius (Panuju AYT dan Usman M).

b. Distribusi dari Kategori AQI



Gambar 2 Distribusi kelas

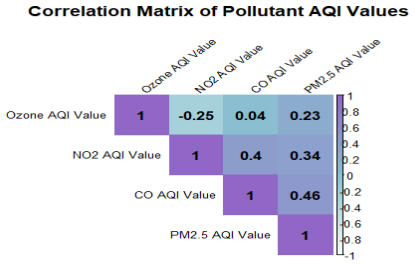


Gambar 3 Distribusi kelas menggunakan violin plot

Meskipun mayoritas nilai AQI yang tercatat menunjukkan kualitas udara yang "Baik" hingga "Sedang", hal ini bukan berarti kita boleh lengah. Masih terdapat beberapa lokasi dengan nilai AQI yang lebih tinggi, menandakan kualitas udara yang "Tidak Sehat" bagi kelompok sensitif dan "Berbahaya" bagi

semua orang. Area-area dengan kualitas udara yang buruk ini perlu mendapatkan perhatian dan tindakan segera.

c. Polutan Udara yang paling signifikan mempengaruhi nilai AQI



Gambar 4 Korelasi matriks heatmap

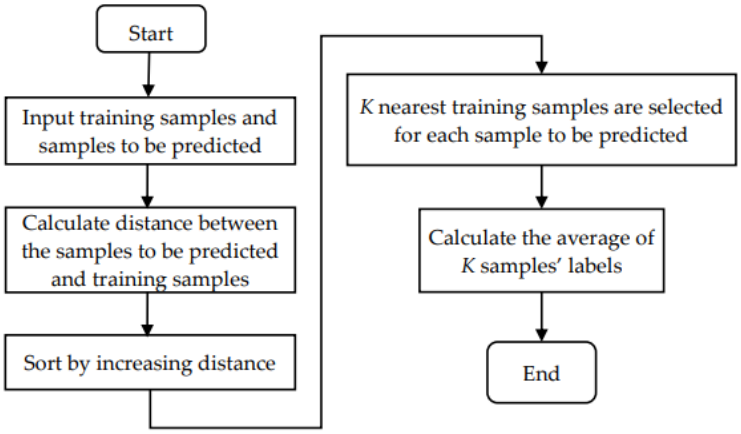
Di antara polutan yang diukur, PM2.5 (partikel halus berdiameter 2.5 mikron atau kurang), NO2 (nitrogen dioksida), CO (karbon monoksida), dan Ozon (O3) merupakan penyumbang utama polusi udara. Menariknya, polutan-polutan ini menunjukkan tingkat korelasi yang berbeda-beda, yang mengindikasikan adanya kemungkinan interaksi dan sumber bersama di antara mereka.

1.2 Metode dan Prinsipnya

Metode yang digunakan dalam penelitian ini adalah pendekatan kuantitatif dengan penerapan model K-Nearest Neighbors (KNN) dalam analisis kualitas udara. Model KNN digunakan untuk memprediksi kualitas udara di lokasi yang tersedia. KNN bekerja dengan cara mencari k-neighbors terdekat dari titik yang akan diprediksi berdasarkan jarak euclidean atau metrik lainnya, lalu memberikan prediksi berdasarkan mayoritas kelas tetangga tersebut (Rahmani MA *et al*).

Algoritma K-Nearest Neighbors (KNN) merupakan metode klasifikasi yang terbilang sederhana dan mudah diterapkan. Algoritma ini bekerja dengan memanfaatkan data latih yang telah memiliki label kelas untuk memprediksi kelas data baru. Dalam penerapannya, KNN pertama-tama mendefinisikan data uji dan data latih. Langkah selanjutnya adalah menghitung jarak antar data uji dan data latih. Jarak ini dihitung menggunakan berbagai metrik, seperti jarak Euclidean atau Manhattan. Setelah memperoleh jarak antar data, langkah berikutnya adalah mengurutkannya dari yang terkecil hingga terbesar. Hasil pengurutan ini menghasilkan daftar tetangga terdekat, di mana tetangga terdekat adalah data latih yang memiliki jarak terkecil dengan data uji. Penentuan kelas data uji dilakukan berdasarkan mayoritas kelas dari k tetangga terdekat. Jumlah tetangga terdekat yang digunakan untuk klasifikasi ditentukan oleh nilai k. Kelas dengan jumlah tetangga terdekat terbanyak akan menjadi kelas yang diprediksikan untuk data uji. Algoritma KNN memiliki beberapa kelebihan, seperti mudah dipahami dan diimplementasikan, serta dapat digunakan untuk berbagai jenis data. Namun, algoritma ini juga memiliki beberapa kekurangan, seperti dapat menjadi lambat untuk data latih yang besar dan sensitif terhadap nilai k.(Siringoringo, 2017) .

Alur algoritma ini ditunjukkan oleh gambar berikut:



Kelebihan utama KNN adalah kesederhanaannya. Algoritma ini mudah dipahami dan diimplementasikan, bahkan untuk orang yang baru memulai dengan machine learning. KNN juga tidak memerlukan asumsi tentang bentuk hubungan antara fitur dan label, sehingga dapat digunakan untuk berbagai jenis data.

Dengan demikian, model ini memungkinkan untuk memperoleh estimasi kualitas udara dengan tingkat akurasi yang tinggi bahkan di daerah-daerah yang minim stasiun pemantauan. Metode ini diharapkan dapat memberikan pemahaman yang lebih baik tentang kualitas udara dan mendukung pengambilan keputusan yang tepat terkait perlindungan lingkungan dan kesehatan masyarakat.

III. HASIL PEMBAHASAN

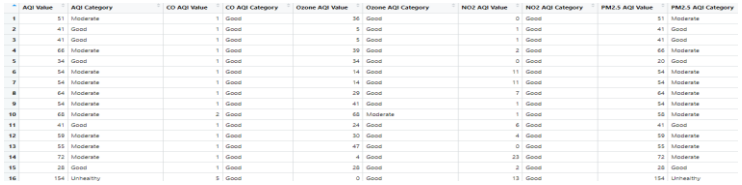
1.3 Preprocessing Data

Studi ini memanfaatkan dataset yang mencakup data Kualitas Udara Indeks (AQI) serta koordinat lintang dan bujur dari berbagai negara di seluruh dunia. Dataset ini diperoleh dari file CSV yang disebut "AQI and Lat Long of Countries.csv". Terdiri dari 12 kolom, dataset ini meliputi informasi penting seperti nama negara, nilai AQI, dan kategori AQI untuk polutan CO, Ozon, NO2, dan PM2.5.

Langkah pertama yang dilakukan adalah membaca data. Pembacaan data dilakukan dengan memanfaatkan library readr. Setelah pembacaan data, dilakukan pemilihan kolom yang dianggap relevan untuk analisis selanjutnya. Hal ini memastikan fokus pada fitur-fitur yang paling berdampak dalam menentukan kualitas udara.

Selanjutnya, kolom kategori AQI dikonversi menjadi representasi numerik. Ini dilakukan menggunakan fungsi factor dan as.numeric untuk mengubah kategori menjadi nilai numerik yang dapat diproses lebih lanjut oleh algoritma K-Nearest Neighbors (KNN). berikut cara dan data sebelum dan sesudah encode:

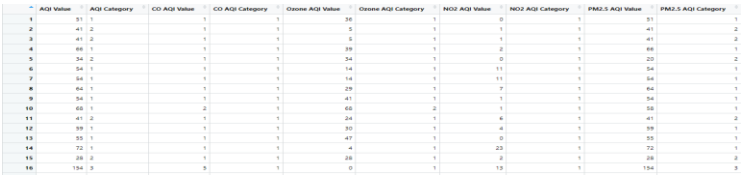
Sebelum encoding:



	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
1	51	Moderate	1	Good	36	Good	0	Good	51	Moderate
2	41	Good	1	Good	5	Good	1	Good	41	Good
3	41	Good	1	Good	5	Good	1	Good	41	Good
4	66	Moderate	1	Good	39	Good	2	Good	66	Moderate
5	34	Good	1	Good	34	Good	0	Good	20	Good
6	54	Moderate	1	Good	14	Good	11	Good	54	Moderate
7	54	Moderate	1	Good	14	Good	11	Good	54	Moderate
8	64	Moderate	1	Good	29	Good	7	Good	64	Moderate
9	54	Moderate	1	Good	41	Good	1	Good	54	Moderate
10	68	Moderate	2	Good	68	Moderate	1	Good	58	Moderate
11	41	Good	1	Good	24	Good	6	Good	41	Good
12	59	Moderate	1	Good	30	Good	4	Good	59	Moderate
13	55	Moderate	1	Good	47	Good	0	Good	55	Moderate
14	72	Moderate	1	Good	4	Good	23	Good	72	Moderate
15	28	Good	1	Good	28	Good	2	Good	28	Good
16	154	Unhealthy	5	Good	0	Good	13	Good	154	Unhealthy

Gambar 5 Data asli

Setelah encoding:



	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
1	51	1	1	1	36	1	0	1	51	1
2	41	2	1	1	5	1	1	1	41	2
3	41	2	1	1	5	1	1	1	41	2
4	66	1	1	1	39	1	2	1	66	1
5	34	2	1	1	34	1	0	1	20	2
6	54	1	1	1	14	1	11	1	54	1
7	54	1	1	1	14	1	11	1	54	1
8	64	1	1	1	29	1	7	1	64	1
9	54	1	1	1	41	1	1	1	54	1
10	68	1	2	1	68	2	1	1	58	1
11	41	2	1	1	24	1	6	1	41	2
12	59	1	1	1	30	1	4	1	59	1
13	55	1	1	1	47	1	0	1	55	1
14	72	1	1	1	4	1	23	1	72	1
15	28	2	1	1	28	1	2	1	28	2
16	154	3	5	1	0	1	13	1	154	3

Gambar 6 Data setelah encoding

Terlihat dengan jelas bahwa masing-masing kategori diubah menjadi sebuah nilai numerik. Pada data ini terdapat 6 kategori seperti yang telah disebutkan sebelumnya, sehingga rentang nilai berada pada 1 hingga 6. Nilai 1 untuk kategori “Good” hingga nilai 6 untuk kategori “Hazardous”.

Terakhir, dilakukan penskalaan fitur menggunakan fungsi scale. Hal ini penting untuk memastikan bahwa semua fitur memiliki rentang yang seragam dan berkontribusi seimbang dalam perhitungan jarak dalam algoritma KNN. Dengan melakukan penskalaan ini, perbedaan magnitudo antar fitur dapat dinormalisasi, menghindari dominasi oleh fitur dengan skala yang lebih besar.

Langkah-langkah preprocessing ini membantu mempersiapkan data untuk analisis lebih lanjut menggunakan algoritma KNN. Dengan demikian, memastikan bahwa model yang dihasilkan mampu memberikan hasil klasifikasi yang akurat terkait dengan kategori AQI berdasarkan informasi yang tersedia dalam dataset.

1.4 Model KNN

Dalam penelitian ini, algoritma K-Nearest Neighbors (KNN) digunakan untuk mengklasifikasikan kategori AQI berdasarkan data yang tersedia. Proses pelatihan model dilakukan dengan membagi dataset menjadi data latih dan data uji, dengan proporsi 80% dan 20% secara berturut-turut. Pembagian dataset ini penting untuk memvalidasi kinerja model. Data latih digunakan untuk melatih model, sementara data uji digunakan untuk mengevaluasi kinerja model yang telah dilatih. Proporsi 80% untuk data latih dan 20% untuk data uji merupakan nilai yang umum digunakan dalam praktik machine learning karena memberikan keseimbangan yang baik antara pelatihan yang cukup untuk model dan evaluasi yang dapat diandalkan terhadap kinerjanya. Mengacu pada buku "Introduction to Machine Learning with Python" yang ditulis oleh Andreas C. Müller dan Sarah Guido strategi pembagian data latih dan data uji seperti proporsi 80% dan 20% cukup efektif untuk melatih model.

Dalam penggunaan algoritma K-Nearest Neighbors (KNN), parameter penting yang perlu ditetapkan adalah nilai k, yang merupakan jumlah tetangga terdekat yang akan dipertimbangkan saat mengklasifikasikan suatu titik data. Menentukan nilai k yang optimal merupakan langkah penting dalam pengembangan model KNN, karena akan mempengaruhi kinerja dan generalisasi model tersebut.

Pada tahap pemilihan nilai k, dilakukan serangkaian eksperimen dan observasi terhadap kinerja model KNN dengan berbagai nilai k. Metode yang dipilih untuk menemukan nilai k terbaik pada riset ini adalah metode validasi silang menggunakan tuneGrid. Melalui langkah-langkah yang terdiri dari definisi rentang nilai k yang ingin dieksplorasi serta pelatihan model KNN untuk setiap nilai k dalam rentang tersebut, proses tuneGrid memungkinkan evaluasi performa model menggunakan metrik yang relevan seperti akurasi atau presisi pada data validasi. Rentang nilai k yang dipilih adalah 1 hingga 40. Ini memberikan kerangka kerja untuk mengeksplorasi berbagai nilai k dalam pencarian nilai optimal. Selanjutnya, model KNN dilatih menggunakan fungsi `train` dari paket `caret`, dengan argumen `tuneGrid` yang diatur dengan nilai `k_grid`. Proses ini memandu pelatihan model untuk mencoba setiap nilai k dalam grid tuning, sehingga memungkinkan penilaian performa model untuk setiap k. Hasil akhirnya adalah model `knn_model` yang telah disesuaikan dengan nilai k optimal berdasarkan evaluasi performa yang dilakukan.

Setelah melatih model KNN dengan menggunakan data latih dan labelnya , langkah selanjutnya adalah menggunakan model yang sudah dilatih untuk membuat prediksi pada data uji .

```
predictions <- knn(train = data_train,
                   test = data_test,
                   cl = train_labels,
                     k= best_k)
```

Dalam proses ini, setiap titik data pada `data_test` akan diklasifikasikan ke dalam kelas yang sesuai berdasarkan label tetangga terdekatnya dalam ruang fitur. Untuk melakukan ini, kita menggunakan fungsi `knn` dengan memberikan argumen `train` yang

berisi data latih, `test` yang berisi data uji, `cl` yang berisi label dari data latih, dan nilai k yang telah ditentukan sebelumnya, yaitu 7. Ketika fungsi `knn` dijalankan, model KNN akan mencari 7 tetangga terdekat dari setiap titik data uji dalam data latih, dan kemudian menentukan kelas mayoritas dari tetangga-tetangga tersebut. Hasil prediksi ini kemudian disimpan dalam variabel `predictions`, yang berisi kelas prediksi untuk setiap titik data uji dalam `data_test`. Dengan demikian, dengan menggunakan nilai k yang telah ditentukan sebelumnya, kita dapat menghasilkan prediksi yang akurat untuk data uji menggunakan model KNN yang sudah dilatih.

1.5 Evaluasi Model

```
> # Print the metrics
> cat("Recall (Sensitivity):", recall, "\n")
Recall (Sensitivity): 0.8666667
> cat("Precision:", precision, "\n")
Precision: 0.8125
> cat("F1 Score:", f1_score, "\n")
F1 Score: 0.8387097
```

Interpretasi hasil evaluasi model KNN untuk klasifikasi kategori AQI menyoroti kinerja model dalam mengidentifikasi dan mengklasifikasikan kualitas udara. Secara umum, evaluasi tersebut menggambarkan seberapa baik model mampu membedakan antara berbagai kategori AQI dengan memperhitungkan recall, precision, dan F1-score.

Recall, yang merupakan proporsi dari instance yang termasuk dalam kategori tertentu yang berhasil diidentifikasi dengan benar oleh model, mencerminkan kemampuan model dalam mengenali sebagian besar kasus positif. Dalam konteks ini, recall sebesar 86.7% menunjukkan bahwa model berhasil mengidentifikasi 86.7% dari semua kasus yang sebenarnya masuk ke dalam kategori AQI tertentu. Namun, ini juga berarti terdapat 13.3% dari kasus yang tidak terdeteksi oleh model.

Precision, di sisi lain, menunjukkan proporsi dari instance yang diklasifikasikan oleh model sebagai kategori AQI tertentu yang benar-benar masuk ke dalam kategori tersebut. Dengan precision sebesar 81.25%, model mampu memberikan label kategori AQI dengan tingkat kepercayaan yang cukup tinggi, di mana sebagian besar instance yang diklasifikasikan sebagai kategori AQI tertentu memang secara faktual masuk ke dalam kategori tersebut.

Nilai F1-score, yang merupakan rata-rata harmonik dari recall dan precision, memberikan gambaran keseluruhan tentang keseimbangan antara kedua matriks tersebut. Dengan nilai F1-score yang moderat sebesar 0.8387, model menunjukkan keseimbangan yang wajar antara kemampuan mengidentifikasi kasus positif dan kemampuan menghindari kesalahan klasifikasi.

Nilai Recall yang diutamakan pada kasus ini didasarkan pada alasan:

1. Prioritaskan Deteksi Kesalahan Negatif (False Negatives): Dalam konteks kualitas udara, kesalahan yang paling berbahaya adalah ketika kondisi yang berpotensi berbahaya tidak terdeteksi. Misalnya, jika kondisi udara benar-benar buruk (tingkat polusi tinggi) dan model gagal mengklasifikasikannya sebagai buruk (False Negative), ini dapat mengakibatkan tindakan yang tidak tepat dari pemerintah atau masyarakat, seperti tidak mengambil langkah-langkah pengendalian polusi yang diperlukan untuk melindungi kesehatan.

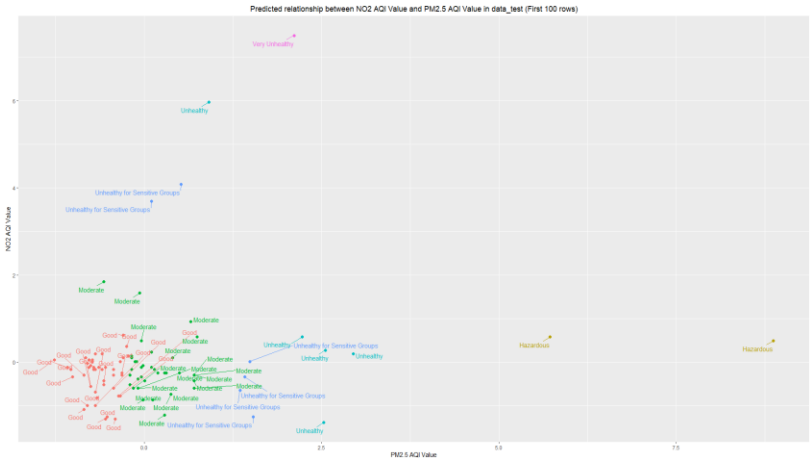
2. Kesehatan Masyarakat dan Keselamatan: Prediksi yang lebih baik dalam mendeteksi situasi berpotensi berbahaya (recall yang tinggi) penting untuk melindungi kesehatan masyarakat dan keselamatan individu. Mengabaikan kondisi berbahaya karena kesalahan prediksi dapat berakibat fatal.
3. Mengurangi Risiko Kesalahan Interpretasi: Ketika kesalahan dapat memiliki dampak kesehatan dan lingkungan yang signifikan, meminimalkan kemungkinan kesalahan mendeteksi kondisi berbahaya menjadi prioritas utama, bahkan jika itu berarti peningkatan dalam false positive (yang kemudian bisa ditangani melalui langkah-langkah mitigasi lebih lanjut).

1.6 Visualisasi

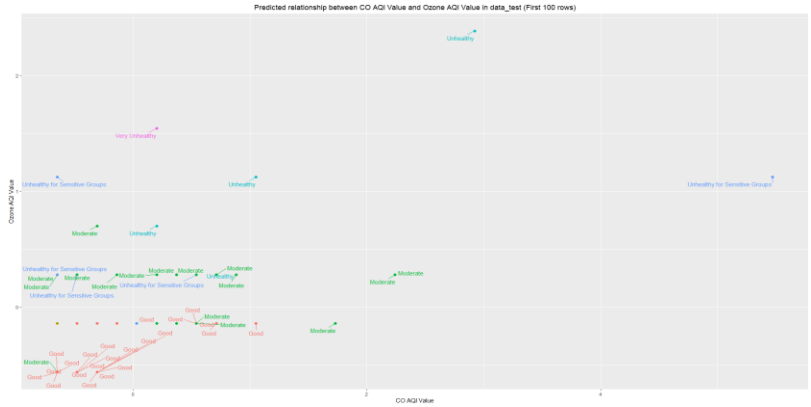
Setelah membuat prediksi, maka berikut visualisasi untuk pemodelan KNN pada data World Air Quality Index by City and Coord. Visualisasi ini hanya dilakukan pada 100 data pertama dari data test. Hal ini dilakukan untuk menjaga keterbacaan visualisasi dan data tetap dapat dipahami dengan baik.

Gambar 7 Visualisasi prediksi AQI Categori dengan beberapa variabel

Dalam penelitian ini, analisis distribusi kelas prediksi dari model KNN diimplementasikan untuk memahami seimbangnya prediksi antara kelas yang berbeda dalam dataset. Hasil visualisasi menunjukkan bahwa terdapat ketidakseimbangan yang signifikan dalam distribusi prediksi, di mana kelas “Good” dan “Moderate” memiliki frekuensi yang jauh lebih rendah dibandingkan dengan kelas lain. Hal ini terjadi karena pada data utama frekuensi masing-masing kelas yang tidak seimbang, sehingga berpengaruh pada hasil prediksi.



Gambar 8 Visualisasi prediksi berdasarkan variabel NO2 dan PM2.5



Gambar 9 Visualisasi prediksi berdasarkan variabel CO dan Ozone

Selanjutnya, analisis kesalahan dilakukan untuk mengidentifikasi pola atau tren tertentu dalam kesalahan yang dibuat oleh model KNN. Temuan dari analisis ini mengungkapkan bahwa terdapat pola yang konsisten dalam kesalahan klasifikasi, di mana sejumlah fitur tertentu cenderung menyebabkan kesalahan dalam prediksi. Sebagai contoh, pada prediksi hubungan antara CO dan Ozone AQI Value, terjadi kesalahan dalam klasifikasi Moderate. Data yang seharusnya dilabeli sebagai Moderate oleh model malah diklasifikasikan sebagai kelas Good. Hal ini berkaitan erat dengan metrik evaluasi dari model yang tidak sempurna. Dimana terdapat 19% data diklasifikasikan ke dalam kelas yang salah.

IV. SIMPULAN DAN SARAN

Model K-Nearest Neighbors (KNN) digunakan untuk memprediksi kualitas udara pada data uji dengan cara mengklasifikasikan setiap titik data ke dalam kelas yang sesuai berdasarkan tetangga terdekatnya pada fitur ruang. Performa model dievaluasi menggunakan metrik seperti perolehan, presisi, dan skor F1. Model ini mencapai skor F1 moderat sebesar 0,8387, yang menunjukkan keseimbangan antara mengidentifikasi kasus positif dan menghindari kesalahan klasifikasi. Penarikan kembali diprioritaskan untuk mendeteksi negatif palsu, meningkatkan kesehatan dan keselamatan masyarakat, serta mengurangi kesalahan interpretasi. Visualisasi dan analisis kesalahan menunjukkan ketidakseimbangan yang signifikan dalam distribusi prediksi karena frekuensi kelas yang tidak merata pada data utama. Kesalahan dalam mengklasifikasikan nilai AQI Sedang ke Baik menunjukkan 19% data salah klasifikasi ke dalam kelas yang salah.

Saran untuk penelitian selanjutnya adalah melakukan pengembangan model K-Nearest Neighbors (KNN) dengan memperhatikan ketidakseimbangan dalam distribusi prediksi. Selain itu, disarankan untuk mengembangkan aplikasi informasi kualitas udara real-time, melakukan kolaborasi dengan pihak terkait, seperti pemerintah dan lembaga lingkungan, serta mengadakan kampanye edukasi tentang polusi udara untuk meningkatkan kesadaran masyarakat.

V. DAFTAR PUSTAKA

Afandie MN, Cholissodin I, Supianto AA. 2014. Implementasi Metode KNearest Neighbor Untuk Pendukung Keputusan Pemilihan Menu Makanan Sehat dan Bergizi. *DORO: Repository Jurnal Mahasiswa FILKOM Universitas Brawijaya*. 3(1).

Fadhlurrahman I. 2023 Des 25. Polusi Udara Senin Sore: Banten Terparah, Bagaimana Daerah Lain? (Senin, 25 Desember 2023). Katadata.co.id. [diakses 2024 Mei]. Polusi Udara Senin Sore: Banten Terparah, Bagaimana Daerah Lain? (Senin, 25 Desember 2023) (katadata.co.id).

Hastie T, Tibshirani R, Friedman J. 2009. The Elements of Statistical Learning. In Springer series in statistics. https://doi.org/10.1007/978-0-387-84858-7.

[KLHK] Kementerian Lingkungan Hidup dan Kehutanan. 2020. Indeks Pencemaran Udara (ISPU) Sebagai Informasi Mutu Udara Ambien di Indonesia. Jakarta: KLHK.

Maharani S, Aryanta WR. 2023. Dampak Buruk Polusi Udara Bagi Kesehatan dan Cara Meminimalkan Risikonya. *Jurnal Ecocentrism*. 3(2): 47–58. https://doi.org/10.36733/jeco.v3i2.7035.

Müller AC, Guido S. 2016. Introduction to Machine Learning with Python: A Guide for Data Scientists. http://cds.cern.ch/record/2229831.

Panuju AYT, Usman M. 2023. PM2.5 Concentration Pattern in ASEAN Countries Based on Population Density. *Procedia of Engineering and Life Science*. 1(4). 10.21070/pels.v4i0.1385.

Rahmani MA, Ratnawati DE, Hanggara BT. 2021. K-Nearest Neighbor untuk Memprediksi Pergantian Komputer di Bank X. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. 5(7): 3198 - 3207.

Siringoringo, R. 2016. Kajian Kinerja Metode Fuzzy K-Nearest Neighbor pada Prediksi Cacat Software. *Jurnal METHODIKA*. 2(2): 211–218.

Wakang AA. 2023 Nov 17. MA Tolak Kasasi, Koalisi Ibukota Desak Jokowi Segera Jalankan Putusan Pengadilan Soal Udara Bersih. Tempo. [diakses 2024 Mei]. MA Tolak Kasasi, Koalisi Ibukota Desak Jokowi Segera Jalankan Putusan Pengadilan Soal Udara Bersih - Metro Tempo.co