Analisis Klasifikasi Kualitas Udara Menggunakan Metode Algoritma *K-Nearest Neighbor* Pada Provinsi Dki Jakarta

1st Valen Deandra
Fakultas Rekayasa Industri
Universitas Telkom
Bandung, Indonesia
valendeandra@student.telkomuniversity.ac.id

2nd Faqih Hamami.
Fakultas Rekayasa Industri
Universitas Telkom
Bandung, Indonesia
faqihhamami@telkomuniversity.ac.id

3rd Irfan Darmawan
Fakultas Rekayasa Industri
Universitas Telkom
Bandung, Indonesia
irfandarmawan@telkomuniversity.ac.id

Abstrak : Udara sangat penting bagi keberlangsungan makhluk hidup, udara membuat makhluk hidup bisa beraktifitas dengan baik. Namun dengan meningkatnya pencemaran udara seiring waktu karena besar nya pertumbuhan pada bidang industri dan banyaknya masyarakat memiliki kendaraan bermotor. Pada tahun 2019 indonesia termasuk pada titik pencemaran udara terburuk yang sudah mencapai titik merah yang menandakan tidak sehat nya udara yang ada pada DKI Jakarta serta memburuknya udara membuat Provinsi DKI Jakarta menduduki posisi ke 5 pencemaran udara terburuk pada IQair dunia. Untuk mengetahui dan monitoring serta mendeteksi informasi kualitas udara maka yang dapat dilakukan klasifikasi. Klasifikasi digunakan karena dapat memonitor informasi kualitas udara berdasarkan pengolahan data ISPU yang sudah memiliki label target. Klasifikasi dilakukan dengan menggunakan dataset ISPU pencemaran udara Provinsi DKI Jakarta dari tahun 2019 sampai 2022. Pada penelitian ini akan mengklasifikasikan data ISPU menggunakan algoritma K-Nearest Neighbor. Dengan menggunakan 5 atribut PM10, SO2, NO, O3, dan CO2 serta kategori sebagai target label dalam penelitian ini. Hasil dari penelitian menunjukan algoritma KNN mendapatkan akurasi tertinggi pada pengujian awal dengan rasio 80:20 dengan ketetanggaan K = 5 dengan nilai akurasi sebesar 90.98%. pengujian kedua dengan tuning hyperparameter yang menghasilkan akurasi tertinggi pada rasio 80:20 dengan ketetanggan k = 7 dengan kombinasi parameter weight "distance", p "1" sebesar 91,37%, presisi 82,87%, recall 85,22% dan f1-score 84.03%. dan validasi algoritma menggunakan K-Fold Cross Validation dengan jumlah fold 10 menghasilkan rata rata sebesar 89,43%.

Kata kunci— Udara, klasifikasi, KNN, DKI Jakarta

I. PENDAHULUAN

Udara adalah salah satu unsur penting dalam kehidupan makhluk hidup, termasuk manusia, karena berperan sebagai sumber oksigen yang diperlukan untuk bernapas. Udara yang baik bagi kesehatan manusia adalah udara bersih, tidak berwarna, tanpa bau yang tidak sedap, serta segar dan sejuk untuk dihirup (Rizi et al., 2019). Selain itu, udara yang bersih juga memiliki manfaat yang signifikan, seperti mengurangi stres dan mencegah berbagai penyakit.

Namun, sayangnya, masalah kualitas udara masih menjadi perhatian serius di banyak negara, terutama karena tingginya tingkat polusi udara. Gambaran mengenai kualitas udara yang diberikan dalam Gambar I.1 menunjukkan bahwa sebagian besar daerah di berbagai negara masih mengalami masalah kualitas udara yang buruk. Warna-warna yang ditunjukkan dalam gambar tersebut, mulai dari merah (tidak sehat) hingga biru (sangat sehat), mencerminkan tingkat kualitas udara yang berbeda-beda (IQAir, 2019).

Buruknya kualitas udara seringkali disebabkan oleh pencemaran udara, dan ini adalah masalah yang terjadi di banyak negara, termasuk Indonesia, khususnya di wilayah Provinsi DKI Jakarta. Pencemaran udara di DKI Jakarta disebabkan oleh berbagai faktor, seperti tingginya jumlah kendaraan bermotor, aktivitas industri, pembakaran batu bara, debu jalanan, dan debu dari aktivitas konstruksi. Namun, salah satu penyumbang utama adalah peningkatan jumlah kendaraan bermotor, yang menyumbang sekitar 31 hingga 40% dari pencemaran udara di Provinsi DKI Jakarta (Identifying the Main Sources of Air Pollution in Jakarta: A Source Apportionment Study - Vital Strategies, n.d.).

Melihat permasalahan yang ada, penelitian ini akan berfokus pada identifikasi dan pemantauan kualitas udara dengan menggunakan data pencemaran udara di Provinsi DKI Jakarta. Hasil dari penelitian ini diharapkan dapat menjadi dasar bagi pemerintah untuk mengembangkan kebijakan dan praktik pengelolaan yang lebih baik dalam mengatasi masalah kualitas udara di wilayah ini.

Untuk mencapai tujuan ini, penelitian akan mengadopsi metode data mining, yang memungkinkan kami untuk mengidentifikasi pola dan hubungan dalam data pencemaran udara. Teknik klasifikasi akan menjadi pendekatan utama dalam analisis data mining ini, dan salah satu algoritma yang akan digunakan adalah algoritma K-Nearest Neighbor (KNN).

K-Nearest Neighbor adalah algoritma yang mengklasifikasikan objek berdasarkan jarak terdekatnya

dengan objek lainnya. Algoritma ini telah digunakan dalam penelitian sebelumnya dan berhasil menghasilkan akurasi yang tinggi dalam berbagai konteks, seperti dalam klasifikasi jenis-jenis sapi (Wijaya et al., 2022) dan identifikasi masyarakat pra sejahtera (Khairi, 2021).

Melalui penelitian ini, kami berharap dapat memberikan kontribusi dalam pemahaman dan penanganan masalah kualitas udara di Provinsi DKI Jakarta. Selain itu, kami juga akan membandingkan kinerja algoritma KNN dengan teknik lain dalam analisis data pencemaran udara, untuk memastikan bahwa hasil yang kami peroleh dapat diandalkan dan relevan.

II. KAJIAN TEORI

A. Pencemaran Udara

Pencemaran udara merupakan peristiwa ketika zat-zat, energi, dan komponen yang berasal dari aktivitas manusia atau alam bercampur dalam atmosfer, yang mengakibatkan kerusakan lingkungan, potensi bahaya bagi kesehatan manusia, dan penurunan kualitas lingkungan secara umum (Riska, 2023). Pencemaran udara dapat terjadi baik secara alami maupun sebagai dampak dari aktivitas manusia. Faktor-faktor penyebab utama pencemaran udara yang dihasilkan oleh manusia mencakup emisi kendaraan bermotor, pembangkit listrik, limbah industri, limbah pertanian, pertambangan, dan kebakaran hutan yang semuanya berkontribusi terhadap polusi udara.

Pencemaran udara memiliki dampak serius pada kualitas udara dan kesehatan manusia. Dalam konteks ini, beberapa jenis polutan pencemaran udara yang umumnya diidentifikasi adalah sebagai berikut:

1. Partikulat Matter (PM): Partikulat matter, yang sering disebut sebagai asap, adalah polutan yang sangat berbahaya. Asal mula partikulat matter umumnya berasal dari gas buangan kendaraan bermotor, cerobong industri yang mengeluarkan asap hitam tebal, panasnya fasilitas pembangkit listrik, dan reaksi polusi gas dalam atmosfer. Partikulat matter mencakup partikel-partikel halus yang berukuran sangat kecil, yang dapat menembus paru-paru manusia dan berdampak negatif pada kesehatan.
2. Sulfur Dioksida (SO2): SO2 berasal dari pembakaran bahan bakar fosil yang mengandung sulfur, terutama ketika batu bara dibakar di pembangkit listrik dan pabrik asam sulfat. SO2 adalah gas berbau tajam yang tidak berwarna, tetapi dapat menyebabkan masalah pernapasan dan menjadi pemicu pembentukan partikel halus yang bercampur dengan zat asam.
3. Karbon Monoksida (CO): Karbon monoksida biasanya dilepaskan melalui pembuangan asap kendaraan bermotor dan beberapa proses industri. Konsentrasi tinggi karbon monoksida dapat terjadi di kota-kota dengan lalu lintas padat. Upaya pengendalian emisi seperti penggunaan katalis telah membantu mengurangi kadar CO di beberapa kota.
4. Nitrogen Oksida (NO2): NO2 berasal dari asap kendaraan bermotor, fasilitas pembangkit listrik, bahan peledak, dan pabrik pupuk. NO2 dapat menyebabkan kerusakan pada paru-paru dan berkontribusi pada pembentukan kabut asap yang berbahaya.
5. Ozon (O3): Ozon terbentuk melalui reaksi antara nitrogen oksida, hidrokarbon, dan sinar matahari. Ozon adalah gas beracun yang berbau tajam. Paparan terhadap ozon dapat menyebabkan masalah pernapasan dan berdampak negatif pada kesehatan manusia serta lingkungan.

Pengetahuan tentang jenis-jenis polutan pencemaran udara ini penting untuk pemahaman yang lebih baik tentang masalah kualitas udara dan upaya pengendaliannya. Dalam konteks penelitian ini, analisis data akan dilakukan untuk memantau dan mengidentifikasi pola pencemaran udara, dengan menggunakan algoritma seperti K-Nearest Neighbor (KNN) dan teknik lainnya untuk menghasilkan hasil yang relevan dan akurat dalam pemantauan dan pengelolaan kualitas udara.

B. Indeks Standar Pencemaran Udara (ISPU)

ISPU, atau Indeks Standar Pencemar Udara, merupakan suatu angka yang tidak memiliki satuan dan digunakan untuk menggambarkan kualitas udara ambien di suatu lokasi tertentu. Angka ISPU didasarkan pada dampak pencemaran udara terhadap kesehatan manusia, nilai estetika, dan makhluk hidup lainnya. Di daerah yang rawan terhadap kebakaran hutan dan lahan, informasi ISPU dapat digunakan sebagai sistem peringatan dini untuk melindungi masyarakat sekitar.

Tujuan utama pembuatan ISPU adalah untuk menyediakan informasi yang konsisten tentang kualitas udara ambien kepada masyarakat pada lokasi dan waktu tertentu. Selain itu, informasi ini juga menjadi pertimbangan penting bagi pemerintah pusat dan daerah dalam upaya pengendalian pencemaran udara.

Menurut (Apriawati & Kiswandono, 2017). ISPU atau Indeks Standar Pencemar Udara merupakan nilai rata-rata yang dihitung dari kombinasi beberapa unsur pencemar udara, yaitu CO, PM10, SO2, NO2, dan O3. Setiap unsur tersebut dihitung berdasarkan kadar tertimbangnya, kemudian nilai standarnya dihitung. Di Indonesia, ISPU merupakan indeks standar kualitas udara yang digunakan secara resmi sesuai dengan Keputusan Menteri Lingkungan Hidup Nomor : KEP 45/MENLH/10/1997 dan KEP-107/KABAPEDAL/11/1997 tentang Indeks Standar Pencemaran Udara..

Tabel di bawah ini menunjukkan kategori ISPU yang digunakan di Indonesia:

TABEL 1
(Kategori ISPU)

Kategori	Rentang angka
Baik	1-50
Sedanng	51-100
Tidak sehat	101-200
Sangat tidak sehat	201-300
Berbahaya	>301

C. Data Mining

Menurut (Firdaus, 2017) Data Mining merupakan suatu Langkah dalam Knowledge Discovery in Databases(KDD) yang terdiri dari pembersihan data (cleaning data), integrasi data(data integration), pemilihan data (data selection), transformasi data(data transformation), data mining, evaluasi pola (pattern evaluation), dan penyajian pengetahuan

(knowledge presentation). Menurut (Yuli Mardi, 2019) data mining memiliki dibagi menjadi beberapa kelompok tugas:

1. Deskripsi pada data mining merupakan analisis untuk mencari cara mengidentifikasi pola dan tren yang terdapat pada data. Deskripsi tentang pola dan kecenderungan sering memberikan penjelasan untuk menjelaskan suatu pola atau tren yang ada didalam data tersebut.
2. Estimasi hampir memiliki kesamaan dengan fungsi klasifikasi, tetapi ada perbedaan dalam variabel target dan tujuan model. Pada estimasi variabel target 11 lebih bersifat numerik. Proses pembangunan model menggunakan data yang sudah menyediakan nilai variabel target sebagai nilai prediksi. Setelahnya estimasi nilai variabel target dibuat berdasar nilai prediksi yang dihasilkan model. jadi yang menjadi pembeda antara klasifikasi dan estimasi terdapat pada variabel target yang digunakan, numerik untuk estimasi dan kategori untuk klasifikasi.
3. Prediksi juga memiliki kesamaan dengan klasifikasi dan estimasi namun perbedaannya berfokus pada hasil pada masa mendatang.
4. Clustering merupakan pengelompokan rekaman, pengamatan, objek yang berdasarkan kemiripan karakteristik yang bertujuan membentuk kelas kelas yang memiliki kesamaan diantaranya. Setiap kelas disebut cluster yang terdiri dari objek yang mirip satu sama lain dan memiliki perbedaan dengan rekaman kluster lain. Perbedaan utama klister dan klasifikasi adalah tidak adanya keterlibatan variabel target yang dipakai, pada klasterisasi ini berusaha membagi data menjadi kelompok yang memiliki kesamaan antara rekaman dalam satu kelompok dalam mencapai nilai maksimal, sedangkan kemiripan dengan rekaman dalam kelompok lain akan bernilai minimal.
5. Klasifikasi memiliki fungsi untuk mencapai target variabel kategori.
6. Asosiasi bertujuan untuk mengidentifikasi atribut yang cenderung muncul bersamaan dalam satu waktu.

D. Klasifikasi

Klasifikasi merupakan teknik melihat label dan atribut dari kelompok yang sudah didefinisikan. Teknik ini dapat mengklasifikasikan data baru dengan memproses data yang diklasifikasikan dan menggunakan hasilnya untuk menghasilkan beberapa aturan atau label saat ada data baru. (Anggada Maulana, 2018).

E. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) adalah metode klasifikasi yang digunakan untuk mengklasifikasikan data berdasarkan mayoritas kedekatan jarak antara data. Secara umum, KNN menggunakan metode jarak Euclidean untuk mengukur jarak antara dua objek. Formula jarak Euclidean dapat didefinisikan sebagai berikut:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (xi\ training - yi\ testing)^2}$$

F. Confusion Matrix

Merupakan evaluasi nilai yang dilakukan setelah proses klasifikasi. Nilai evaluasi berdasarkan menganalisis confusion matrix.

TABEL 2
(Confusion Matrix)

Klasifikasi	*Predictive positive*	*Predictive negative*
Actual Positive	*True Positif(TP)*	*False Negative(FN)*
Actual Negative	*False Positive(FP)*	*True Negative(TN)*

Menurut (Nurdalia et al., 2023) Untuk mengukur kemampuan model dalam klasifikasi dalam *confusion matrix*, beberapa metrik evaluasi utama digunakan, yaitu:

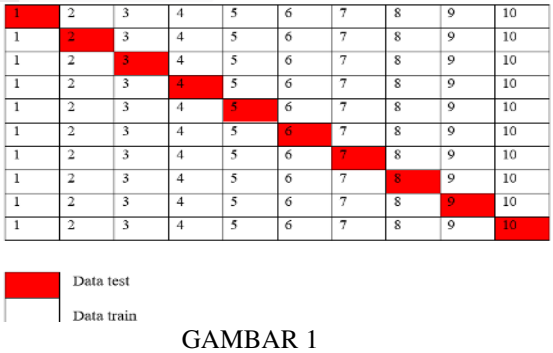
1. Akurasi (Accuracy): Merupakan rasio data yang akuratnya terdeteksi dalam data prediksi. Ini mengukur sejauh mana model klasifikasi mencocokkan hasilnya dengan nilai sebenarnya.
2. Presisi (Precision): Merupakan nilai yang mengukur sejauh mana model memberikan hasil positif yang benar.
3. Recall (Sensitivitas): Merupakan nilai yang mengukur tingkat keberhasilan model dalam mendeteksi hasil positif dengan benar.
4. F1-Measure: Merupakan metrik yang menggabungkan Presisi dan Recall untuk memberikan penilaian yang seimbang tentang kinerja model.

G. Gridsearch CV

Menurut (Belete & Huchaiah, 2022) *grid search* menggambarkan bagaimana sebuah pencarian menyeluruh yang menguji semua hasil hyperparameter yang diberikan pada konfigurasi grid. Metode ini beroperasi dalam menilai dari sekumpulan nilai terbatas yang ditentukan oleh peneliti.

H. K-fold Cross Validation

Menurut (Azis et al., 2020) Teknik ini digunakan untuk memvalidasi model menilai dengan melakukan prediksi model dan memperkirakan seberapa akurat model Ketika dijalankan. menurut (Tempola et al., 2018) k fold cross validation merupakan teknik menilai algoritma dengan melakukan prediksi model dan memperkirakan akuratnya model Ketika digunakan. k fold cross validation memecah data menjadi k dengan ukuran yang sama.



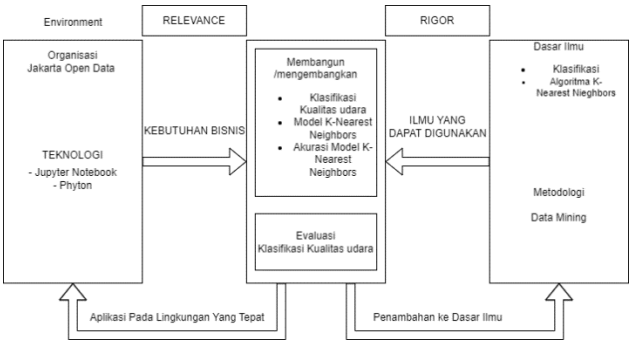
GAMBAR 1
(*K fold validation*)

III. METODE

A. Model Konseptual

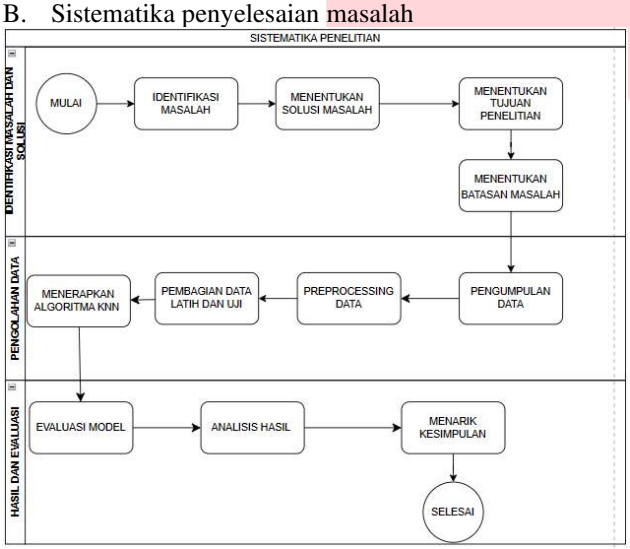
Model konseptual merupakan tahap memastikan peneliti membangun model sesuai dengan keperluan, pengetahuan sebelumnya serta pengalaman. Model ini bukan representasi

struktur dasar sistem melakukan model yang membantu dalam membaca penggunaan sistem secara efektif.



GAMBAR 2
(Model konseptual)

B. Sistematika penyelesaian masalah



GAMBAR 3
(Sistematika penyelesaiain masalah)

1. Identifikasi Masalah

Penulis melakukan pencarian studi kasus dan mengangkat tentang kualitas udara DKI Jakarta. Pada tahap ini dilakukan studi literatur, lalu mementukan pemecahan masalah yang dibahas menentukan tujuan penelitian , menentukan batasan masalah dan diakhiri dengan penulis.

2. Pengolahan Data

Selanjutnya penulis melakukan pencarian dataset serta mengumpulkan melalui website Jakarta open data. penulis mengumpulkan data dan melakukan preprocesing terhadap dataset tersebut , data dibersihkan dari nilai null , variabel yang tidak sesuai dan menyeleksi variabel apa saja yang akan digunakan dalam penelitian ini. Selanjutnya penulis akan melakukan splitting data menjadi data training dan data testing Setelah itu data yang sudah diolah tersebut akan diimplementasikan menggunakan model K-Nearest Neighbor.

3. Hasil dan evaluasi

Penulis akan melakukan *tuninghyperparamater* untuk mendapatkan parameter terbaik menggunakan GridsearchCV selanjutnya uji evaluasi performansi algoritma menggunakan *confusion matrix* , selanjutnya akan dilakukan perhitungan akurasi , recall, dan presisi dari algoritma yang digunakan. Selanjutnya penulis menggunakan *K-Fold Cross Validation* untuk mengetahui nilai K mana yang optimal.

4. Pengumpulan Data

Pada tahap ini pengumpulan data berasal dari situs resmi Jakarta Open Data Pemerintah wilayah DKI Jakarta dan dinas Lingkungan Hidup DKI Jakarta yang menyediakan data tentang informasi Indeks Standar Kualitas Udara DKI Jakarta. Data yang dipakai sebagai data training memiliki atribut “pm10”, “So2”, “Co”, “o3”, dan “No2”. Dengan menggunakan target label empat kelas yaitu “BAIK”, “SEDANG”, “TIDAK SEHAT”, dan “SANGAT TIDAK SEHAT”.

5. Pengolahan data

Setelah mengumpulkan data selanjut data mentah diolah melalui proses *preprocessing* menjadi data yang bisa diolah. Pada pengolahannya dilakukan *splitting data* dengan membagi data set menjadi tiga skenario 80:20, 70:30, 60:40. Selanjutnya dilakukan pemodelan menggunakan algoritma *K-Nearest Neighbor*. Setelah pemodelan selanjutnya akan dilakukan evaluasi pada model.

6. Metode evaluasi

Evaluasi dilakukan dengan tiga metode, pertama menggunakan GridsearchCV pada tahap ini dilakukan pencarian parameter terbaik yang digunakan dalam proses pemodelan klasifikasi, K-Fold Cross Validation yang mana untuk mengetahui K mana yang lebih optimal dengan uji sebanyak 10 fold. KNN confusion matrix yang mana mencari akurasi, presisi, dan recall dari setiap nilai yang didapatkan.

IV. HASIL DAN PEMBAHASAN

A. Implmentasi Algoritma

1. Pengujian awal sebelum tuning

peneliti melakukan pengujian dengan K 1 sampai 10 pada tiga rasio, rasio 60 : 40 menghasilkan nilai akurasi tertinggi pada K = 5 mendapakatkan skor 89,79%, pada rasio 70 : 30 nilai akurasi tertinggi pada K = 7 mendapatkan skor sebesar 90.48%, terakhir pada rasio 80:20 mendapatkan nilai akurasi tertinggi pada nilai K = 5 dengan skor sebesar 90.98%.

2. Tuning Hypermeter

Tunning Hyperparameter ini bertujuan mencari kombinasi parameter terbaik untuk mendapat nilai akurasi optimal yang lebih tinggi dari pengujian awal guna meningkatkan kinerja algoritma dalam klasifikasi. Peneliti menggunakan metode *GridsearchCV*. Pengujian dilakukan pada tiga rasio dengan pengujian sebanyak 10 fold. Parameter yang digunakan adalah grid dari rentang K bernilai 3,5,7,9 yang termasuk angka ganjil agar mencegah terjadinya hasil voting dengan jumlah suara yang sama untuk berbagai kelas, selanjutnya menggunakan fungsi weight yang digunakan untuk mengontrol bagaimana bobot kontribusi tetangga digunakan saat melakukan prediksi untuk titik data baru. Parameter weight yang digunakan adalah uniform adalah semua tetangga diberi bobot yang sama dan menganggap tetangga sama pentingnya dalam prediksi dan distance memberi titik dalam lingkungan diberi bobot yang berbanding terbalik dari jarak mereka yang memungkinkan tetangga yang lebih dekat mempunyai pengaruh yang besar dalam prediksi dari pada titik ketetanggan yang terjauh,

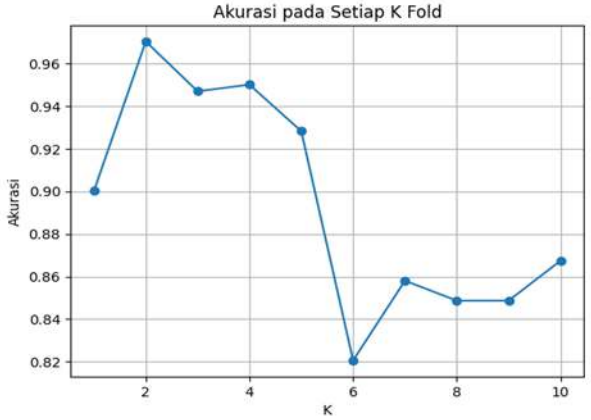
selanjutnya p merupakan parameter yang digunakan untuk mengontrol metrik yang memungkinkan untuk mengubah bagaimana jarak antara titik akan dihitung parameter p bernilai 1 ini jarak antara dua titik dihitung sebagai jumlah mutlak perbedaan antara koordinat parameter ini cocok untuk situasi perpindahan dua titik mengikuti jalur sejajar dengan sumbu dan parameter bernilai 2 jarak dihitung sebagai jarak garis lurus antara mereka dalam ruang berdimensi n, metrik ini merupakan pilihan baik untuk mengukur jarak antara dua titik dalam bentuk jarak Euclidean. Berdasarkan penjelasan yang dipaparkan peneliti melakukan penelitian dan menghasilkan kombinasi parameter seperti pada tabel berikut.

TABEL 3
(Tuning Hyperparameter tiga rasio)

Grid Search CV	K	Metric	Weight	P	Akurasi
60 : 40	7	Euclidean	distance	1	89.44%
70 : 30	9	Euclidean	distance	1	90.42%
80 : 20	7	Manhattan	distance	1	91.37%

3. K fold cross validation

K fold merupakan pengujian untuk memvalidasi keakuratan algoritma dalam mengklasifikasikan data. Peneliti melakukan pengujian dengan fold sebanyak 10.



GAMBAR 4
(K Fold Cross Validation)

TABEL 4
(Hasil K Fold Cross Validation)

Fold 1	90.12%
Fold 2	97.64%
Fold 3	94.81%
Fold 4	95.13%
Fold 5	92.77%
Fold 6	81.94%
Fold 7	85.40%
Fold 8	84.77%
Fold 9	85.08%
Fold 10	86.65%
Rata-rata akurasi	89.43%

B. Evaluasi Performansi

Berdasarkan hasil pengujian yang telah dilakukan didapatkan bahwa pada rasio 80:20 mendapatkan nilai akurasi terbaik selanjutnya akan dilakukan evaluasi performa klasifikasi pengujian nilai awal dan sesudah *tuninghyperparameter* dengan menggunakan *Confussion Matrix*.

TABEL 5
(Confusion Matrix GridsearchCV)

	Prediksi Baik	Prediksi tidak sehat	Prediksi Sedang	Prediksi tidak sehat
True Baik	150	0	25	1
True sangat tidak sehat	0	0	0	0
True Sedang	31	0	885	37
True tidak Sehat	0	1	15	130

1. Akurasi

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Akurasi} = \frac{149+0+883+128}{149+32+1+27+883+15+1+39+128}$$

$$\text{Akurasi} = \frac{1160}{1275}$$

$$\text{Akurasi} = 90{,}98\%$$

2. Precision

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Precision} = \frac{149}{149+32}$$

$$\text{Precision} = 82{,}32\%$$

3. Recall

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Recall} = \frac{149}{149+27+1}$$

$$\text{Recall} = 84.18\%$$

4. F1 Score

$$\text{F1 score} = 2 \text{ x } \frac{Recall*Precision}{Recall+Precision}$$

$$\text{F1 score} = 2 \text{ x } \frac{0{,}8418 * 0.8232}{0.8418+0.8232}$$

$$\text{F1 score} = 2 \text{ x } \frac{0.6929}{1.665}$$

$$\text{F1 score} = 2 \text{ x } 0.4162$$

$$\text{F1 Score} = 83{,}24\%$$

V. KESIMPULAN

A. Penelitian ini melakukan klasifikasi kualitas udara data ISPU yang didapat dari menggunakan algoritma KNN dengan menggunakan parameter pm10, so2, co, o3, no2 sebagai data training dan kategori sebagai output label yang akan diklasifikasi. Simulasi yang digunakan adalah melakukan splitting data menjadi tiga rasio 60:40, 70:30,

80:20. Kelas yang digunakan yaitu baik, sedang , tidak sehat dan sangat tidak sehat.

B. Setelah melakukan pengujian hasil yang didapatkan model terbaik pada didapatkan pada rasio 80:20 pada ke K = 5 dengan akurasi sebesar 90.98%. pengujian kedua menggunakan tuninghyperparameter yang menghasilkan akurasi terbaik pada rasio 80:20 k = 7 dengan parameter weight "distance", p "1" sebesar 91,37%. Saat menerapkan K-Fold Cross Validation dengan jumlah fold 10 menghasilkan rata rata sebesar 89.43%.

REFERENSI

Abidin, J., Artauli Hasibuan, F., kunci, K., Udara, P., & Gauss, D. (2019). Pengaruh Dampak Pencemaran Udara Terhadap Kesehatan Untuk Menambah Pemahaman Masyarakat Awam Tentang Bahaya Dari Polusi Udara. In Prosiding SNFUR-4.

Apriawati, E., & Kiswandono, A. A. (2017). Kajian Indeks Standar Polusi Udara (ISPU) Nitrogen Dioksida (NO2) di Tiga Lokasi Kota Bandar Lampung. Analytical and Environmental Chemistry, 2(01), 42–51.

Firdaus, D. (2017). Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer. In Jurnal (Vol. 6).

Gilabert, P. L., Gadringer, M. E., Montoro, G., Mayer, M. L., Silveira, D. D., predistortion and ofdm clipping for power amplifiers. International Journal of RF and Microwave Computer-Aided Engineering, 19(5), 583–591. https://doi.org/10.1002/mmce.20381

Identifying the Main Sources of Air Pollution in Jakarta: A Source Apportionment Study - Vital Strategies. (n.d.). Retrieved March 17, 2023, from https://www.vitalstrategies.org/resources/identifying-the-main-sources-of-air-pollution-in-jakarta-a-source-apportionment-study/

Rizi, U. F., Suradi, Sunaryo, Agus, A., Ahmad, M., Kusumaningtyas, S. D. A., Nurhayati, H., Khoir, A. N., Sucianingsih, C., & W, N. F. P. (2019). Analisis Dampak Diterapkannya Kebijakan Working From Home Saat Pandemi Covid-19 Terhadap Kondisi Kualitas Udara Di Jakarta. Jurnal Meteorologi Klimatologi Dan Geofisika, 6(3), 6–14. https://jurnal.stmkg.ac.id/index.php/jmkg/article/view/141

Roihan, A., Sunarya, P. A.4, & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. IJCIT (Indonesian Journal on Computer and Information Technology), 5(1), 75–82. https://doi.org/10.31294/ijcit.v5i1.7951

Silaparasetty, N., & Silaparasetty, N. (2020). Machine Learning vs. Deep Learning. In Machine Learning Concepts with Python and the Jupyter Notebook Environment. https://doi.org/10.1007/978-1-4842-5967-2_4

Sodiq, M. J., & Sela, E. I. (2019). Perbandingan Metode Naive Bayes Dan K-Nearest Neighbor Pada Klasifikasi Kualitas Udara Di Dki Jakarta.

Wahyono, T. (2018). Fundamental of Python for Machine Learning: Dasar-Dasar Pemrograman Python untuk Machine Learning dan Kecerdasan Buatan. SeAmbarwari, A., Jafar Adrian, Q., & Herdiyeni, Y. (2020). Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 4(1), 117–122. https://doi.org/10.29207/resti.v4i1.1517

Anggada Maulana. (2018). Konsep Dasar Data Mining. Konsep Data Mining, 1, 1–16.

Azis, H., Purnawansyah, P., Fattah, F., & Putri, I. P. (2020). Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung. ILKOM Jurnal Ilmiah, 12(2), 81–86. https://doi.org/10.33096/ilkom.v12i2.507.81-86

Belete, D. M., & Huchaiah, M. D. (2022). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. International Journal of Computers and Applications, 44(9), 875–886. https://doi.org/10.1080/1206212X.2021.1974663

BPS Provinsi DKI Jakarta. (n.d.). Retrieved March 16, 2023, from https://jakarta.bps.go.id/indicator/17/786/1/jumlah-kendaraan-bermotor-menurut-jenis-kendaraan-unit-di-provinsi-dki-jakarta.html

Firdaus, D. (2017). Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer. In Jurnal (Vol. 6).

Khairi, A. (2021). Implementasi K-Nearest Neighbor (KNN) untuk Klasifikasi Masyarakat Pra Sejahtera Desa Sapikerap Kecamatan Sukarapu. Jurnal TRILOGI, 2(3), 319–323. https://ejournal.unuja.ac.id/index.php/trilogi/article/view/2878

Klein, R. H., Klein, D. B., & Luciano, E. M. (2018). Open Government Data: Concepts, Approaches and Dimensions Over Time. Revista Economia & Gestão, 18(49), 4–24. https://doi.org/10.5752/p.1984-6606.2018v18n49p4-24

Kusnandar, M. (2020). Permen LHK Nomor 14 Tahun 2020. Permen LHK Nomor 14 Tahun 2020 Tentang Indeks Standar Pencemar Udara (ISPU), 1–16.

Ndaumanu, R. I., & Arief, Kusrini, M. R. (2014). Analisis Prediksi Tingkat Pengunduran Diri Mahasiswa dengan Metode K-Nearest Neighbor. Jatisi, 1(1), 1–15. http://www.mdp.ac.id/jatisi/vol-1-no-1/JATISI_Vol_1_No_1_September_2014_1.pdf

Normawati, D., & Ismi, D. P. (2019). K-Fold Cross Validation for Selection of Cardiovascular Disease Diagnosis Features by Applying Rule-Based Datamining. Signal and Image Processing Letters, 1(2), 23–35. https://doi.org/10.31763/simple.v1i2.3

Nurdalia, Zilrahmi, Permana, D., & Salma, A. (2023). Comparison of Naïve Bayes and K-Nearest Neighbor for DKI Jakarta Air Pollution Standard Index Classification. UNP Journal of Statistics and Data Science, 1(2), 67–73. https://doi.org/10.24036/ujsds/vol1-iss2/29

Riska, V. (2023). Pencemaran Polusi Udara. May. https://www.researchgate.net/publication/370816936

Tempola, F., Muhammad, M., & Khairan, A. (2018). Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation. Jurnal Teknologi Informasi Dan Ilmu Komputer, 5(5), 577. https://doi.org/10.25126/jtiik.201855983

Tuntun, R., Kusrini, K., & Kusnawi, K. (2022). Analisis Perbandingan Kinerja Algoritma Klasifikasi dengan Menggunakan Metode K-Fold Cross Validation. Jurnal Media Informatika Budidarma, 6(4), 2111. https://doi.org/10.30865/mib.v6i4.4681

Wahyono, W., Trisna, I. N. P., Sariwening, S. L., Fajar, M., & Wijayanto, D. (2020). Comparison of distance measurement on k-nearest neighbour in textual data classification. Jurnal Teknologi Dan Sistem Komputer, 8(1), 54–58. https://doi.org/10.14710/jtsiskom.8.1.2020.54-58

Wijaya, S. F. A., Koredianto, K., & Saidah, S. (2022). Analisis Perbandingan K-Nearest Neighbor dan Support Vector Machine pada Klasifikasi Jenis Sapi dengan Metode Gray Level Coocurrence Matrix. Jurnal Ilmu Komputer Dan Informatika, 2(2), 93–102. https://doi.org/10.54082/jiki.27

Yuli Mardi. (2019). Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika. Jurnal Edik Informatika, 2(2), 213–219.