Jurnal Ilmiah Informatika Komputer, 29(1): 45-59.
Doi: https://doi.org/10.35760/ik.2024.v29i1.10069

PREDIKSI TINGKAT KUALITAS UDARA DENGAN PENDEKATAN ALGORITMA K-NEAREST NEIGHBOR

[1]Nesya Syahira, [2]Dede Brahma Arianto*
[1]Program Studi Kesehatan Lingkungan, Fakultas Kesehatan Masyarakat, Universitas Indonesia, Jalan Margonda Raya, Pondok Cina, Kecamatan Beji, Kota Depok, Jawa Barat 16424, Indonesia
[2]Magister Informatika, Fakultas Teknologi Industri, Universitas Islam Indonesia, Yogyakarta, 55584, Indonesia
[1]nesya.syahira@ui.ac.id, [2]dede.brahma2@gmail.com
**) Penulis Korespondensi*

Abstrak

Udara merupakan sumber kehidupan bagi makhluk hidup. Namun, dalam beberapa tahun terakhir penurunan kualitas udara menjadi permasalahan serius yang mendesak untuk diatasi. Indeks Standar Pencemar Udara (ISPU) menjadi indikator yang dapat digunakan untuk mengetahui kondisi kualitas udara ambien di lokasi tertentu. Kota Yogyakarta merupakan salah satu kota besar di Indonesia dengan permasalahan pencemaran udara yang serius dalam beberapa tahun terakhir. Penelitian ini bertujuan untuk memprediksi kualitas udara di Kota Yogyakarta berdasarkan data ISPU dengan menggunakan teknik data mining dan metode klasifikasi. Algoritma yang digunakan dalam prediksi adalah K-Nearest Neighbor (K-NN), yang mengklasifikasikan objek baru berdasarkan tetangga terdekatnya. Evaluasi model algoritma dilakukan dengan mengukur akurasi, presisi, recall, dan f-measure untuk nilai K = 5. Hasil pengujian menunjukkan bahwa nilai K = 5 memberikan performa yang baik dengan akurasi sebesar 99% dimana untuk kategori "Good" menghasilkan precision 100%, recall 99%, dan f-measure 100%, sedangkan kategori "Moderate" menghasilkan precision 98%, recall 100%, dan f-measure 99%.

Kata Kunci: *Data Mining, KNN, Kualitas Udara, Prediksi.*

Abstract

Air is the source of life for living things. However, in recent years the decline in air quality has become a serious problem that urgently needs to be addressed. The Air Pollution Standard Index (ISPU) is an indicator that can be used to determine the condition of ambient air quality in a particular location. Yogyakarta is one of the major cities in Indonesia with serious air pollution problems in recent years. This research aims to predict air quality in Yogyakarta City based on ISPU data using data mining techniques and classification methods. The algorithm used in prediction is K-Nearest Neighbor (K-NN), which classifies new objects based on their nearest neighbors. Evaluation of the algorithm model is done by measuring accuracy, precision, recall, and f-measure for the value of K = 5. The test results show that the value of K = 5 provides good performance with an accuracy of 99% which is for the "Good" category producing 100% precision, 99% recall, and 100% f-measure, while the "Moderate" category produces 98% precision, 100% recall, and 99% f-measure.

Keywords: *Data Mining, KNN, Air Quality, Prediction.*

PENDAHULUAN

Udara merupakan salah satu elemen lingkungan yang vital dalam kehidupan manusia. Pada udara bersih dan kering, rata-rata persentase gas per volume di dalamnya yaitu Nitrogen 78.08%, Oksigen 20.95%, Argon 0.934%, Karbon Dioksida 0.03%, dan gas lainnya 0.27%. Pencemaran udara merupakan suatu kondisi di mana adanya kontaminasi pada atmosfer oleh zat kimia, fisik, atau biologis yang mampu mengubah karakteristik atmosfer yang dapat mengakibatkan terjadinya kerusakan lingkungan hingga gangguan pada kesehatan manusia. Saat ini, aktivitas manusia yang semakin mengedepankan industrialisasi sangat mempengaruhi kualitas udara baik di tingkat perkotaan, regional, dan bahkan skala global karena terjadinya peningkatan jumlah gas dan partikel berbahaya yang berpotensi merusak udara [1]. Pada akhirnya, pencemaran udara dapat mengakibatkan masalah kesehatan pada masyarakat karena terekspos polutan pada kadar yang tinggi, seperti infeksi saluran pernafasan hingga kanker paru-paru [2]. Pencemaran udara telah menjadi permasalahan yang mendesak untuk diatasi dalam beberapa tahun terakhir di kota-kota besar di Indonesia, termasuk Kota Yogyakarta. Meskipun luas wilayahnya relatif kecil, yakni sebesar 32,5 kilometer persegi, namun kompleksitas infrastrukturnya yang cukup besar telah berdampak signifikan terhadap tingkat pencemaran di kota ini. Sumber utama penyebab pencemaran udara di Kota Yogyakarta teridentifikasi berasal dari dua sumber, yakni sumber bergerak, seperti pertumbuhan jumlah kendaraan bermotor yang sangat pesat, dan sumber tidak bergerak, berupa emisi gas buang dari pabrik [3].

Kementerian Lingkungan Hidup dan Kehutanan RI berkomitmen untuk menyediakan informasi yang akurat tentang kualitas udara kepada masyarakat sebagai bagian dari usahanya dalam mengendalikan pencemaran udara. KLHK telah menunjukkan komitmennya dengan terus meningkatkan jumlah Stasiun Pemantauan Kualitas Udara Ambien (SPKUA) kontinu yang dimilikinya dan menyampaikan hasil dari pemantauan stasiun-stasiun tersebut ke dalam bentuk Indeks Standar Pencemar Udara (ISPU) agar informasi mengenai kualitas udara dapat dipantau secara langsung dan lebih mudah dipahami oleh masyarakat [4]. Hadirnya ISPU dapat dimanfaatkan untuk memberikan informasi mengenai kualitas udara di suatu tempat, termasuk di Kota Yogyakarta, hingga dapat dijadikan sebagai dasar untuk membuat langkah dalam mengurangi pencemaran udara [5].

Data mining merupakan sebuah proses eksplorasi untuk mencari pola dalam sekumpulan data yang bertujuan untuk mendapatkan informasi yang berguna [6]. Salah satu metode yang digunakan dalam implementasi *data mining* adalah dengan *machine learning* metode K-Nearest Neighbor. *Machine learning* adalah disiplin ilmu yang memfokuskan pada cara

mengembangkan program yang yang mampu menghasilkan pengetahuan baru berdasarkan pengetahuan yang telah ada [7]. Sedangkan metode K-Nearest Neighbor merupakan salah satu model dari *machine learning* yang dapat digunakan untuk memprediksi dan melakukan mengklasifikasikan *dataset* dengan memanfaatkan pembelajaran data dan termasuk ke dalam *supervised learning*. Dalam *supervised learning*, algoritma belajar dari *dataset* yang diberikan dan melakukan klasifikasi berdasarkan kedekatan jarak suatu data dengan data lain dalam kategori yangsama [8]. Penelitian serupa telah dilakukan oleh Nugroho et al. (2023) yang bertujuan memprediksi kualitas udara yang ada di DKI Jakarta berdasarkan data ISPU menggunakan algoritma Random Forest [9]. Temuan mereka memberikan pemahaman bahwa klasifikasi dengan algoritma Random Forest menghasilkan akurasi sebesar 90%. Penelitian lainnya yang serupa untuk memprediksi kualitas udara yang ada di DKI Jakarta juga dilakukan oleh Kirono et al. (2022) namun menggunakan algoritma Naive Bayes. Hasil klasifikasi yang diberikan oleh algoritma Naive Bayes menghasilkan akurasi 88% [10]. Kemudian Amalia et al. (2022) juga melakukan penelitian untuk memprediksi kualitas udara yang ada di DKI Jakarta berdasarkan data ISPU menggunakan algoritma K-Nearest Neighbor dan didapatkan nilai K = 7 memberikan akurasi 96% [11]. Berdasarkan hal tersebut, maka penelitian dalam artikel ini akan dilakukan untuk memperoleh performa yang lebih baik dalam memprediksi kualitas udara di Kota Yogyakarta menggunakan *machine learning* dengan algoritma K-Nearest Neighbor dengan nilai parameter yang digunakan adalah 5.

METODE PENELITIAN

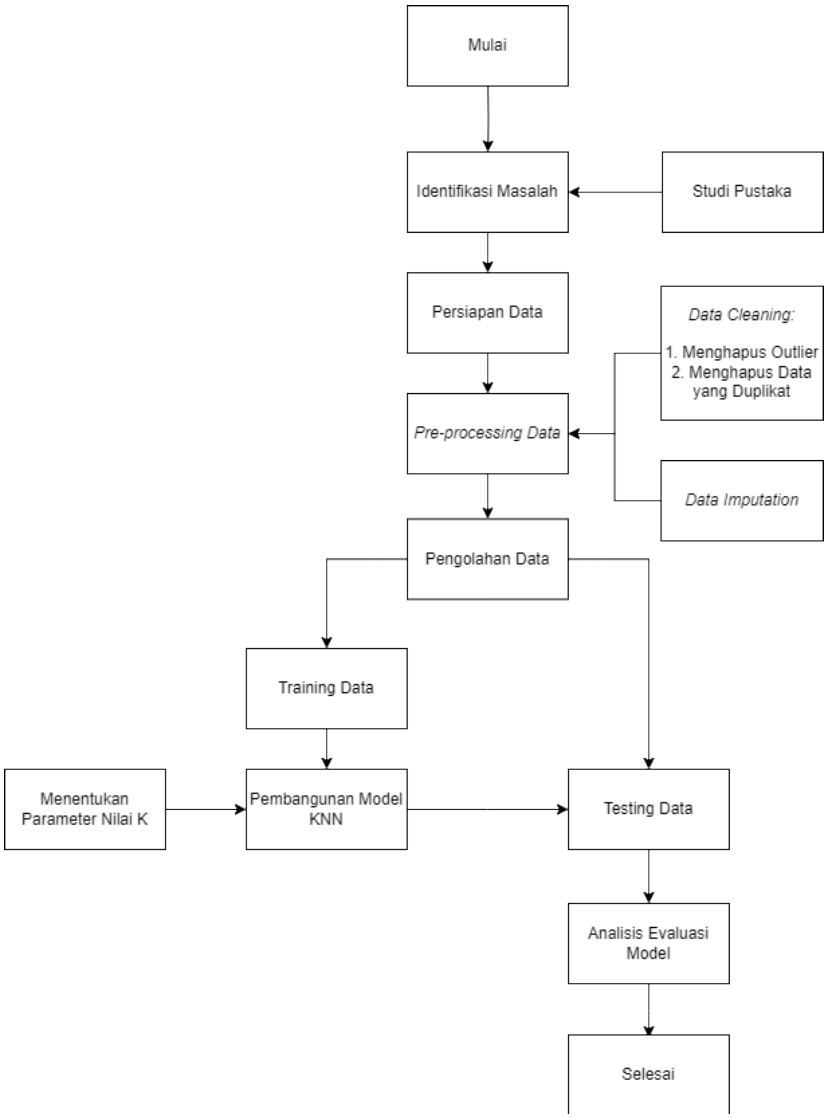
Alur yang dilakukan pada penelitian ini dapat dilihat pada Gambar 1.

A. Identifikasi Masalah

Penelitian ini dilakukan dengan menilik permasalahan penurunan kualitas udara yang terjadi di Kota Yogyakarta berdasarkan informasi yang didapatkan melalui studi literatur.

B. Persiapan Data

Dataset yang digunakan dalam penelitian ini didapatkan dari *website* Kaggle dengan laman URL https://www.kaggle.com/ berupa file berekstensi *.csv* dengan sumber data utama didapatkan dari Dinas Lingkungan Hidup Yogyakarta. Pada *dataset* tersebut, memuat informasi kualitas udara Kota Yogyakarta berupa hasil pengukuran polusi udara seperti Partikulat (PM_{10}, $PM_{2.5}$), Sulfur Dioksida (SO_2), Karbon Monoksida (CO), Ozon (O_3), dan Natrium Dioksida (NO_2) pada bulan Januari hingga Desember tahun 2021 yang telah dikonversi ke Indeks Standar Pencemaran Udara (ISPU). *Dataset* ini memuat total sebanyak 8.760 *records data* dengan 10 atribut dan 1 kelas.



Gambar 1. Alur Penelitian

C. Pre-processing Data

Proses pre-processing data menjadi langkah yang dilakukan untuk memastikan kualitas data yang digunakan. *Data cleansing* merupakan aspek penting dalam proses ini yang dilakukan dengan menghapus *outliers* dan data yang duplikat, dimana *outliers* atau pencilan merupakan data yang menyimpang secara ekstrim sedangkan data duplikat merupakan data yang dapat muncul akibat adanya kesalahan dalam pengumpulan data. Dua hal tersebut tentunya mampu mempengaruhi hasil analisis sehingga penting untuk dihapus. Selanjutnya, dilakukan *data imputation* yang bertujuan untuk mengisi nilai-nilai pada data yang hilang dengan perkiraan atau estimasi yang sesuai seperti menggunakan rata-rata, median, atau modus dari kolom data yang sesuai.

D. Eksplorasi Data

Proses eksplorasi data dilakukan dengan melibatkan analisis korelasi untuk memberikan wawasan tentang hubungan antar variabel dan membantu dalam memahami pola-pola dalam data. Secara lebih spesifik, analisis korelasi dilakukan untuk melihat hubungan linier antara variabel-variabel yang tersedia di dalam data, sehingga langkah ini membantu untuk mengidentifikasipola dan tren, memahami bagaimana perubahan dalam satu variabel berkaitan dengan perubahan dalam variabel lainnya, dan memberikan petunjuk untuk analisis lebih lanjut.

E. Perancangan Model

Algoritma K-Nearest Neighbor digunakan untuk membantu dalam mengklasifikasikan objek data baru berdasarkan atributnya dan sampel data yang ada dalam *dataset* latihan [8]. Pada tahap pengolahan data, proses yang dilakukan melibatkan pembagian *dataset* hasil *pre-processing* menjadi dua bagian, yaitu data latih dan data uji. Selain itu, tahapan ini juga mencakup implementasi algoritma K-Nearest Neighbor (K-NN) untuk melakukan klasifikasi terhadap objek baru dengan menghitung jarak terdekat dari objek tersebut yang disimbolkan dengannilai parameter K menggunakan perhitungan *euclidean distance* [11]. Nilai K adalah parameter jumlah tetangga atau data terdekat yang akan dilibatkan untuk menentukan label kelas dari objek baru, dimana metode KNN akan melakukan klasifikasi berdasarkan label kelas yang mendominasi dari data tetangga terdekat tersebut [12].

Berdasarkan hal tersebut, data akan terbagi ke dalam beberapa kategori dan algoritma K-NN akan membantu dalam menentukan kategori yang paling sesuai untuk data baru. Algoritma K-NN sangat dipengaruhi oleh nilai K yang digunakan. Oleh karena itu, dibutuhkan pemilihan nilai K yang optimal agar bisa mendapatkan hasil yang baik dan memberikan akurasi tertinggi terhadap klasifikasi. Untuk mendapatkan nilai K yang optimal, maka harus mencari nilai K dengan melakukan perbandingan hasil dari berbagai nilai K tersebut.

Euclidean distance adalah salah satu metode perhitungan jarak yang lebih optimal jika dibandingkan dengan metode lainnya, dimana *euclidean distance* digunakan untuk mengukur jarak dari dua titik, yaitu titik data dan tetangga terdekat dalam algoritma K-NN [13]. Berikut adalah langkah-langkah yang digunakan untuk melakukan perhitungan manual

menggunakan algoritma K-Nearest Neighbor.

1. Menentukan nilai K sebagai parameter yang menunjukkan jumlah tetangga terdekat yang akan digunakan untuk mengklasifikasi objek pengujian. Pada perhitungan ini, akan digunakan nilai K minimum untuk evaluasi model, yaitu K = 5. Nilai K tersebut didapatkan berdasarkan hasil dari penggunaan teknik optimasi parameter, yaitu *Grid Search*. *Grid Search* adalah sebuah metode untuk menentukan parameter yang optimal bagi suatu model, dimana algoritma ini melakukan uji coba seluruh kombinasi nilai *hyperparameter* dan memilih parameter yang memberikan kinerja terbaik sesuai dengan metrik evaluasi yang telah ditentukan [14]. Melalui metode ini, dapat diketahui bahwa model K-NN akan memberikan kinerja terbaik ketika jumlah tetangga terdekat yang digunakan untuk mengklasifikasi objek pengujian, yaitu pada nilai 5.
2. Menghitung jarak antara objek baru dan semua objek dalam data latih. Perhitungan jarak dilakukan pada setiap baris data dengan memasukkan nilai-nilai dari data latih dan data uji ke dalam rumus. Berikut adalah rumus yang digunakan untuk menghitung *euclidean distance* pada Persamaan (1).

   Euclidean distance (d) = $\sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}$ (1)

   Berdasarkan Persamaan (1), d adalah jarak Euclidean antara dua titik dalam ruang n-dimensi. Variabel a_1, a_2, a_n mewakili koordinat dari titik pertama, sedangkan b_1, b_2, b_n mewakili koordinat dari titik kedua. Untuk setiap dimensi (i), perbedaan antara koordinat titik pertama (a_i) dan koordinat titik kedua (b_i) di kuadratkan, kemudian semua hasilnya dijumlahkan. Akhirnya, akar kuadrat dari jumlah tersebut memberikan jarak Euclidean antara kedua titik tersebut (d).
3. Melakukan pengurutan hasil perhitungan jarak dari yang paling kecil hingga yang terbesar.
4. Menentukan tetangga terdekat berdasarkan nilai K yang telah ditentukan sebelumnya.
5. Menetapkan kategori dari tetangga terdekat objek pengujian.

F. Analisis Evaluasi Model

Metode analisis evaluasi model yang digunakan yaitu dengan *confusion matrix*. Metode ini umum

digunakan untuk menggambarkan kinerja dari model klasifikasi pada kumpulan data uji yang dilakukan dan kondisi asli yang diketahui.*Confusion matrix* digambarkan dalambentuk dua dimensi yang terdiri dari

kelas sebenarnya dan kelas prediksi sehingga terdapat empat kombinasi nilai prediksi (Tabel 1) [15].

Pada *confusion matrix*, berdasarkan nilai *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), dan *False Negatives* (FN) dapat diperoleh nilai akurasi, presisi, *recall*, dan *f-measure*.

1. Akurasi

   Akurasi merupakan nilai yang menunjukkan seberapa dekat nilai prediksi yang dihitung dengan nilai aktual atau nilai sebenarnya. Akurasi dapat dihitung menggunakan rumus pada Persamaan (2).

$$\text{Akurasi} = \frac{(TP+TN)}{TP+TN+FP+FN} \quad (2)$$

2. Presisi

   Presisi merupakan nilai yang menunjukkan seberapa dekat nilai yang diukur satu sama lain. Presisi dapat dihitung menggunakan rumus pada Persamaan (3).

$$\text{Presisi} = \frac{TP}{TP+FP} \quad (3)$$

3. *Recall*

   Recall merupakan nilai yang menunjukkan rasio dari semua prediksi positif yang terprediksi dengan benar. *Recall* dapat dihitung menggunakan rumus pada Persamaan (4).

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

4. *F-Measure*

   F-Measure merupakan nilai yang menggambarkan perbandingan rata-rata antara presisi dan recall. *F-Measure* dapat dihitung menggunakan rumus pada Persamaan (5).

$$F\text{-}Measure = 2\frac{Presisi\ x\ Recall}{Presisi+Recall} \quad (5)$$

Tabel 1. Evaluasi Model *Confusion Matrix*

Kelas Sebenarnya	**Kelas Prediksi**	
	-	+
-	*True Negatives* (TN)	*False Positives* (FP)
+	*False Negatives* (FN)	*True Positives* (TP)

Keterangan:
True Positives (TP) : Jumlah data positif yang terklasifikasi benar oleh sistem
True Negatives (TN) : Jumlah data negatif yang terklasifikasi benar oleh sistem
False Positives (FP) : Jumlah data positif namun terklasifikasi salah oleh sistem
False Negatives (FN) : Jumlah data negatif namun terklasifikasi salah oleh sistem

HASIL DAN PEMBAHASAN

A. Persiapan Data

Data ISPU pada penelitian ini disusun oleh Dinas Lingkungan Hidup Yogyakarta. Secara keseluruhan, *dataset* ini memuat 8.760 *record* data dengan 10 atribut dan 1 kelas. Atribut-atribut yang terdapat dalam *dataset* ini antara lain Tanggal, Waktu, PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, NO_2, Max, *Critical Component*, dan *Category* (Tabel 2).

B. Tahap *Pre-Processing* Data

Dataset yang digunakan dalam penelitian ini terdiri dari beberapa kolom dengan nilai yang hilang (*missing value*) yang ditandai dengan kode NaN, dan terdapat baris data yang tidak memiliki pengukuran. Selain itu, terdapat beberapa atribut yang dihapus, seperti tanggal, stasiun, dan *critical component* karena atribut-atribut ini dianggap tidak memberikan kontribusi yang signifikan terhadap pengolahan data kualitas udara selanjutnya. Oleh karena itu, penentuan kualitas udara dalam penelitian ini didasarkan pada nilai ukuran dari parameter-parameter seperti PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, NO_2, Max, dan variabel *Category* sebagai kelas yang dituju. Untuk memberikan gambaran lebih jelas, contoh data yang perlu melewati tahap *pre-processing* (Tabel 3).

Tabel 2. Data Contoh ISPU Januari – Desember 2021

Date	Time	PM_{10}	$PM_{2.5}$	SO_2	CO	O_3	NO_2	*Max*	*Critical Component*	*Category*
4/1/2021	10:00:00	12	31	19	11	7	3	19	PM2.5	Good
4/2/2021	10:00:00	9	23	15	10	19	3	19	PM2.5	Good
4/3/2021	10:00:00	20	50	20	17	13	4	20	PM2.5	Good
4/4/2021	10:00:00	26	55	22	22	22	6	26	PM2.5	Good
4/5/2021	10:00:00	9	22	13	11	21	6	21	PM2.5	Good
4/6/2021	10:00:00	22	51	24	23	17	7	24	PM2.5	Good
4/7/2021	10:00:00	22	51	21	16	15	4	22	PM2.5	Good
4/8/2021	10:00:00	21	51	25	20	20	5	25	PM2.5	Good
4/9/2021	10:00:00	27	58	16	7	24	4	27	PM2.5	Good
4/10/2021	10:00:00	36	65	13	13	44	6	44	PM2.5	Good

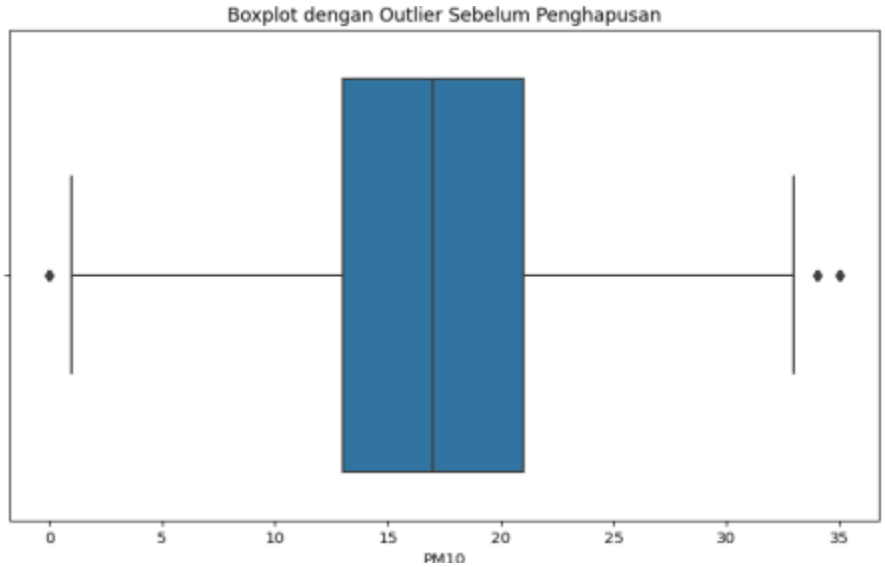
Tabel 3. Contoh *Dataset Missing Value*

Date	Time	PM_{10}	$PM_{2.5}$	SO_2	CO	O_3	NO_2	*Max*	*Critical Component*	*Category*
11/18/2021	15:00:00	8	17	0	19	40	3	40	O3	Good
11/19/2021	15:00:00	10	18	NaN	19	41	1	41	O3	Good
11/20/2021	15:00:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
11/21/2021	15:00:00	19	36	NaN	24	21	2	36	PM2.5	Good
11/22/2021	15:00:00	38	57	0	21	25	2	57	PM2.5	Moderate

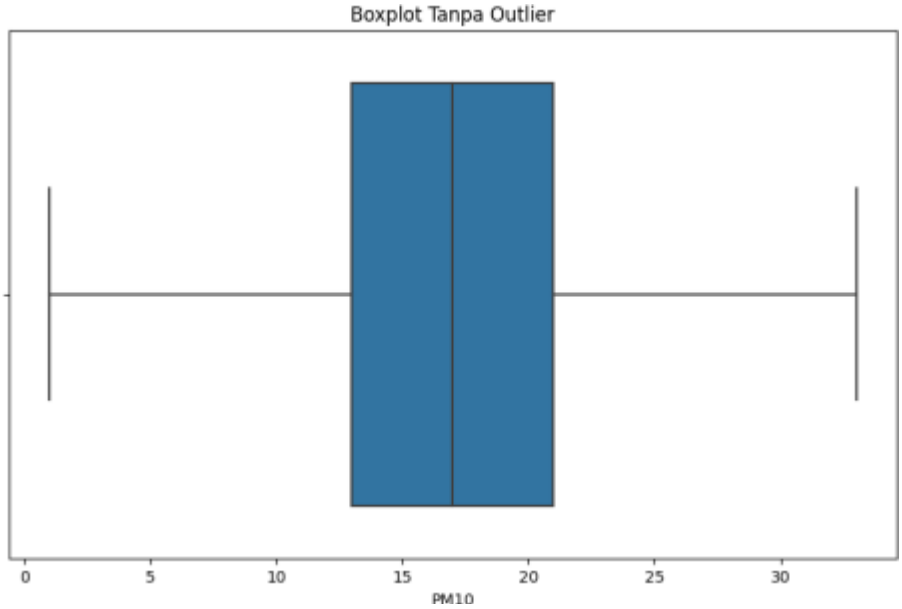
Pembersihan data dilakukan dalam dua proses, yaitu pembersihan data (*data cleaning*) dan pengisian data (*data imputation*) menggunakan *Python*. Pada proses *data cleaning*, hal pertama yang dilakukan adalah mendeteksi dan mengatasi data hilang (*missing value*) yang ditandai dengan kode NaN dan baris yang tidak memiliki data pengukuran dalam *dataset*. Data yang terdeteksi hilang selanjutnya memasuki proses *data imputation* dengan melakukan pengisian terhadap nilai yang hilang menggunakan nilai rata-rata (*mean*) dari setiap kolom. Melalui langkah ini, setiap nilai yang hilang (NaN) akan digantikan dengan nilai rata-rata dari kolom yang sesuai.

Proses *data cleaning* selanjutnya dilakukan dengan mendeteksi duplikasi baris dalam data frame. Hasilnya adalah seri boolean yang menandai baris-baris yang merupakan duplikasi, dimana seri yang berisi nilai *true* pada indeks yang sesuai dengan baris-baris yang merupakan duplikasi dan *false* untuk baris yang tidak duplikat. Setelah hasil dicetak, diketahui bahwa tidak ada data yang duplikat dalam *data frame*.

Proses *data cleaning* berikutnya adalah penanganan *outliers* pada setiap variabel, dan didapatkan bahwa *outliers* hanya terdapat pada variabel PM_{10} (Gambar 2). *Outliers* dideteksi dengan menghitung kuartil dan rentang interkuartil, kemudian menentukan batas atas (*max*) dan batas bawah (*min*). Nilai yang terdeteksi kurang dari batas bawah (*min*) dan batas atas (*max*) selanjutnya diganti dengan kode NaN (nilai yang hilang). Ini dilakukan untuk menghapus *outliers* dari data. Kemudian menghitung jumlah nilai yang hilang dalam kolom PM_{10} setelah *outliers* dihapus, dimana terdapat 266 baris dan dilanjutkan dengan menghapus baris yang memiliki nilai yang hilang tersebut. Setelah dilakukan penghapusan baris yang memiliki nilai yang hilang, dapat dilihat bahwa *outliers* dalam kolom PM_{10} telah teratasi (Gambar 3).



Gambar 2. Boxplot sebelum Penghapusan *Outliers*



Gambar 3. Boxplot setelah Penghapusan *Outliers*

Tabel 4. Hasil Analisis Korelasi

Atribut	Peringkat Korelasi
Max	0.790719
$PM_{2.5}$	0.651176
PM_{10}	0.522489
O_3	0.434482
SO_2	0.302397
CO	0.291442
NO_2	0.144981
Critical Component	-0.137747

C. Eksplorasi Data

Proses eksplorasi data dalam penelitian ini terdiri dari analisis korelasi antara atribut yaitu PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, NO_2, Max, dengan variabel *Category* sebagai kelas. Analisis korelasi dimulai dengan pemberian label angka pada kolom kategorikal, karena dalam beberapa algoritma pembelajaran mesin memerlukan input numerik dan tidak dapat bekerja langsung dengan data kategorik. Kemudian melakukan perhitungan jumlah kelas dalam kolom *Category* untuk memahami sebaran kategori atau kelas dalam data, didapatkan bahwa jumlah kelas yang tersedia adalah "Good" dan "Moderate". Perhitungan matriks korelasi juga dilakukan untuk memberikan informasi tentang sejauh mana satu variabel berkaitan dengan variabel lainnya. Kemudian, korelasi antara setiap atribut dan kolom target *Category* diekstraksi. Korelasi ini kemudian diurutkan dari yang tertinggi ke yang terendah, dan hasilnya dicetak untuk memberikan pemahaman tentang sejauh mana setiap fitur berkorelasi dengan variabel target (Tabel 4). Berdasarkan hasil dari analisis korelasi, didapatkan

peringkat korelasi setiap atribut numerik dengan variabel target *Category*. Hasilkorelasi ini memberikan gambaran tentang sejauh mana setiap atribut berkorelasi dengan variabel target. Atribut dengan korelasi positif yang tinggi cenderung memiliki hubungan yang lebih kuat dengan variabel target, sementara atribut dengan korelasi negatifdapat dianggap sebagai faktor yang berlawanan dengan variabel target.

D. Perhitungan Manual Algoritma K-Nearest Neighbor

Perhitungan manual dilakukan untuk mengetahui cara kerja dari algoritma K-Nearest Neighbor dalam melakukan prediksi. Data pelatihan yang digunakan terdiri dari 10 data contoh yang telah di-*split* menggunakan *Python* (Tabel 5). Selanjutnya akan dilakukan penentuan kelas dari 1 contoh data uji yang belum diketahui kelasnya. Berikut ini data uji yang akan digunakan untuk perhitungan manual (Tabel 6). Berikut adalah hasil perhitungan manual algoritma K-Nearest Neighbor dengan menggunakan Persamaan (1) (Tabel 7). Berdasarkan Tabel 7, dapat disimpulkan bahwa kelas dari data uji yang digunakan termasuk dalam kategori "Good" karena lima tetangga terdekatnya memiliki kelas mayoritas yaitu "Good".

Tabel 5. Data Latih Perhitungan Manual

PM_{10}	$PM_{2.5}$	SO_2	CO	O_3	NO_2	*Max*	*Category*
13	40	0	25	0	0	40	Good
11	52	0	23	10	0	52	Moderate
13	53	0	23	11	0	53	Moderate
10	50	0	22	9	0	50	Good
16	40	21	12	1	3	40	Good
22	51	22	18	8	3	51	Moderate
27	56	12	12	32	8	32	Good
13	0	10	20	3	10	20	Good
51	0	16	22	1	10	51	Moderate
14	28	51	23	21	1	51	Moderate

Tabel 6. Data Uji Perhitungan Manual

PM_{10}	$PM_{2.5}$	SO_2	CO	O_3	NO_2	*Max*	*Category*
21	38	1	25	20	1	38	n/a

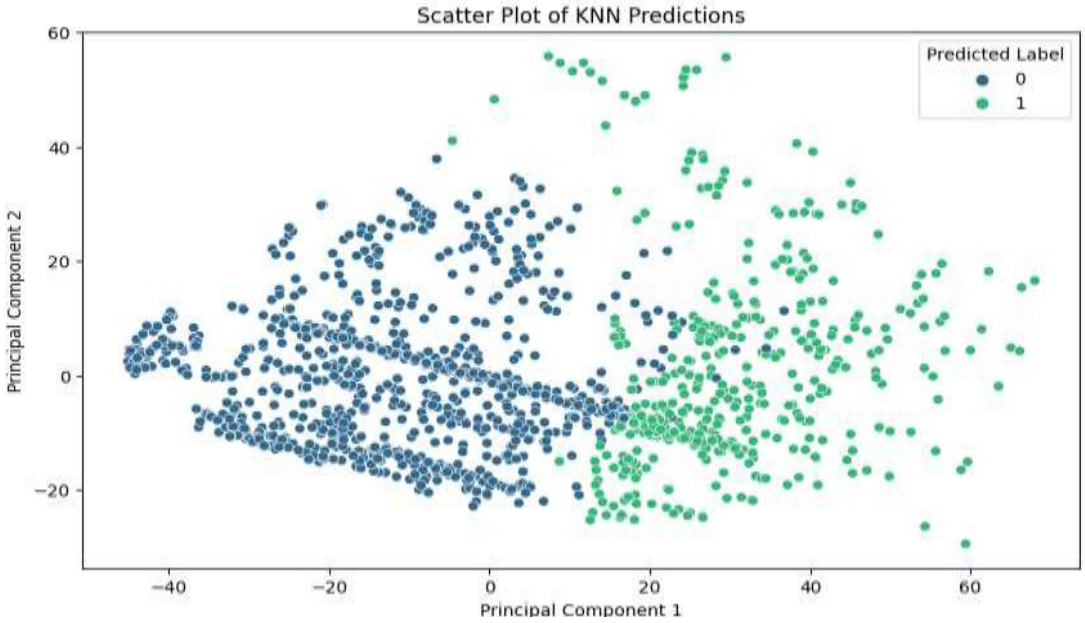
Tabel 7. Penentuan Jarak Terdekat

Urutan ke-	d	K=5	*Category*
1	21,83	Ya	Good
2	23,21	Ya	Good
3	24,49	Ya	Moderate
4	24,53	Ya	Moderate
5	29,65	Ya	Good
6	31,05	Tidak	Good
7	31,24	Tidak	Moderate
8	48,04	Tidak	Good
9	56,10	Tidak	Moderate
10	53,12	Tidak	Moderate

E. Hasil Penerapan *Modelling* Algoritma K-Nearest Neighbor

Jumlah data yang telah melalui proses pembersihan pada tahap sebelumnya mencapai 8.494 data. Selanjutnya, *dataset* akan dibagi menjadi dua bagian, yaitu data latih dan data uji. Algoritma akan mempelajari pola-pola yang terdapat dalam data latih, sementara data uji akan digunakan untuk menguji prediksi dengan menggunakan model yang telah dibangun oleh algoritma. Pembagian data yang digunakan untuk membentuk model algoritma dalam penelitian ini dilakukan dengan rasio 20% untuk data uji dan 80% untuk data latih. Atribut yang digunakan didasarkan pada *dataset* yang telah menjalani tahap *pre-processing* dan terdiri dari 7 kolom sebagai atribut dan 1 kolom sebagai kelas. Atribut adalah variabel-variabel yang digunakan untuk membuat prediksi atau analisis, sedangkan kelas adalah label atau kategori yang ingin diprediksi. Proses pembagian data dilakukan menggunakan perintah *Python*.

Algoritma K-Nearest Neighbor memerlukan suatu parameter untuk menentukan jumlah tetangga terdekat dengan objek data yang baru, yaitu parameter K. Nilai K yang digunakan dalam penerapan algoritma ini adalah K=5. Artinya, saat algoritma memprediksi kategori atau nilai untuk suatu data baru, ia akan mempertimbangkan lima tetangga terdekat dari data tersebut untuk membuat prediksi. Pemilihan nilai K ini akan dievaluasi berdasarkan akurasi, presisi, *recall*, dan *f-measure* dari setiap nilai parameter yang diujikan. Berdasarkan hasil penerapan *modelling* algoritma K-Nearest Neighbor, diketahui bahwa model yang dibangunberhasil memprediksi dan mengklasifikasikankategori ke dalam kelas “Good” yang ditandaidengan angka 0 dan kelas “Moderate” yang ditandai dengan angka 1 (Gambar 4).



Gambar 4. Scatterplot hasil Prediksi dengan K-NN

Tabel 8. Hasil Evaluasi Model

Nilai K		Kelas	Presisi	*Recall*	*F-Measure*
5	0	Good	1.00	0.99	1.00
	1	Moderate	0.99	1.00	0.99
Accuracy					0.99
Macro Avg			0.99	0.99	0.99
Weighted Avg			0.99	0.99	0.99

F. Evaluasi Model

Evaluasi kinerja model dilakukan melalui penggunaan *confusion matrix*. Penilaian model ini dipertimbangkan berdasarkan tingkat akurasi, presisi, *recall*, dan *f-measure* dalam memprediksi kelas "Good" dan "Moderate" pada target *Category*(Tabel 8). Hasil *confusion matrix* dari penerapan algoritma ini direpresentasikan menggunakan perintah *Python*.

Berdasarkan tabe;menunjukkan hasil *accuracy* dari model adalah 0.99 atau 99%. Hal ini menunjukkan bahwa model ini benar dalam memprediksi lebih dari 99% dari seluruh data pengujian. Kemudian *score* dari kategori "Good" menghasilkan *precision* 100%, *recall* 99%, dan *f-measure* 100%, sedangkan *score* dari kategori "Moderate" menghasilkan *precision* 99%, *recall* 100%, dan *f-measure* 99%. Menghasilkan nilai rata-rata dengan *macro average* menghasilkan *precision* 99%, *recall* 99%, *f-measure* 99%, sedangkan *weighted average* menghasilkan *precision* 99%, *recall* 99%, *f-measure* 99%.

Model yang dievaluasi memiliki kinerja yang sangat baik, dengan akurasi yang tinggi dan metrik presisi, *recall*, dan *f-measure* yang baik untuk kedua kelas. Model ini memiliki kemampuan yang baik untuk memprediksi data dengan benar, sehingga jika kedepannya dalam *dataset* terdapat klasifikasi kategori baru selain kategori *Good* dan *Moderate*, maka model K-NN akan melakukan prediksi data dengan baik.

KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilaksanakan, dapat disimpulkan bahwa algoritma K-Nearest Neighbor (K-NN) dapat diterapkan untuk memprediksi kualitas udara berdasarkan Indeks Standar Pencemar Udara (ISPU). Prediksi ini dilakukan dengan memanfaatkan 7 (tujuh) fitur yang mencakup parameter-parameter udara seperti PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, NO, dan nilai maksimum untuk menentukan kualitas udara. Saat melakukan evaluasi model algoritma menggunakan *confusion matrix*, hasilnya menunjukkan bahwa nilai K optimal yang didapatkan melalui metode *Grid Search*, yaitu K = 5, terbukti mampu memberikan performa yang baik dengan tingkat akurasi mencapai 99%. Berdasarkan hal tersebut, dapat disimpulkan bahwa menggunakan algoritma K-Nearest Neighbor dengan nilai parameter 5 memberikan performa yang baik dalam memprediksi kualitas udara di Kota Yogyakarta. Dari penerapan algoritma K-NN ini, prediksi kualitas udara dapat menjadi lebih akurat dan dapat memberikan kontribusi

positif dalam upaya pemantauan dan pengelolaan lingkungan. Performa yang tinggi pada nilai akurasi juga menunjukkan bahwa algoritma K-NN dapat menjadi pilihan yang efektif dalam memprediksi kualitas udara berdasarkan parameter-parameter yang diukur. Berdasarkan hasil dari penelitian ini, disarankan untuk melanjutkan penelitian dengan beberapa pertimbangan. Pertama, penting untuk melakukan eksperimen lebih lanjut guna mengoptimalkan parameter K pada algoritma K-Nearest Neighbors (K-NN). Penyesuaian nilai K dapat signifikan dalam mempengaruhi performa model. Selanjutnya, penelitian dapat diperluas dengan melibatkan data ISPU di kota lainnya untuk validasi model, sehingga dapat memastikan bahwa model ini dapat diterapkan dengan baik pada lingkungan yang berbeda. Selain itu, dapat juga mempertimbangkan penambahan fitur atau data tambahan untuk memperkaya model, serta analisis lebih mendalam terhadap kontribusi masing-masing fitur terhadap prediksi kualitas udara. Berdasarkan pertimbangan-pertimbangan ini, penelitian selanjutnya dapat lebih mendalam dan berkontribusi pada pemahaman lebih lanjut tentang aplikasi K-NN dalam prediksi kualitas udara.

DAFTAR PUSTAKA

[1] World Health Organization, "Air Pollution," 2023. https://www.who.int/health-topics/air-pollution#tab=tab_1 (accessed Sep. 20, 2023).

[2] Kementerian Lingkungan Hidup dan Kehutanan, *Pengendalian Pencemaran Udara*. 2016.

[3] Pemerintah Kota Yogyakarta, "Program Langit Biru Kendalikan Pencemaran Udara di Kota Yogya," 2016. https://warta.jogjakota.go.id/detail/index/4669 (accessed Sep. 28, 2023).

[4] Kementerian Lingkungan Hidup dan Kehutanan, "Indeks Standar Pencemar Udara (ISPU) sebagai Informasi Mutu Udara Ambien di Indonesia," 2020. https://ditppu.menlhk.go.id/portal/read/indeks-standar-pencemar-udara-ispu-sebagai-informasi-mutu-udara-ambien-di-indonesia (accessed Sep. 22, 2023).

[5] Menteri Lingkungan Hidup dan Kehutanan RI, "Permen LHK Nomor 14 Tahun 2020 Tentang Indeks Standar Pencemar Udara (ISPU)," 2020. https://ditppu.menlhk.go.id/portal/uploads/news/1600940556_P_14_2020_ISPU_menlhk_07302020074834.pdf

[6] Universitas Pembangunan Jaya, *Konsep Data Mining*. 2008. [Online]. Available: https://ocw.upj.ac.id/files/Handout-TIF311-DM-1.pdf

[7] N. Hasdyna and R. K. Dinata, "Machine Learning." 2020. [Online]. Available: http://repository.unimal.ac.id/id/eprint/6707

[8] Lembaga Penelitian dan Pengabdian Masyarakat Universitas Medan Area,

"Algoritma K-Nearest Neighbors (KNN) – Pengertian dan Penerapan," 2023. https://lp2m.uma.ac.id/2023/02/16/algoritma-k-nearest-neighbors-knn-pengertian-dan-penerapan/ (accessed Oct. 02, 2023).

[9] A. Nugroho, I. Asror, and Y. F. A. Wibowo, "Klasifikasi Tingkat Kualitas Udara DKI Jakarta Berdasarkan Open Government Data Menggunakan Algoritma Random Forest," *eProceedings Eng.*, vol. 10, No. 2, no. 2, pp. 1824–1834, 2023, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20030%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20030/19395

[10] A. A. H. Kirono, I. Asror, and ..., "Klasifikasi Tingkat Kualitas Udara Dki Jakara Dengan Algoritma NaiveBayes," *eProceedings ...*, vol. Vol. 9, No, no. 3, pp. 1962–1969, 2022, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/18002%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/18002/17631

[11] A. Amalia, A. Zaidiah, and I. N. Isnainiyah, "Prediksi Kualitas Udara Menggunakan Algoritma K-Nearest Neighbor," *J. Ilm. Penelit. dan Pembelajaran Inform.*, vol. 7, no. 2, pp. 496–507, 2022, doi: 10.33387/jiko.v4i2.2871.

[12] R. Aljabar and D. Kusumaningsih, "Penerapan Algoritma Klasifikasi K-Nearest Dengan Menggunakan Hasil Nilai Try Out Siswa Sekolah Menengah Kejuruan Berbasis Desktop Pada SMK Bina Informatika Bintaro," *Skanika*,vol. 1, no. 1, pp. 136–142, 2018.

[13] W. Wahyu Pribadi, A. Yunus, and A. S. Wiguna, "Perbandingan Metode K-Means Euclidean Distance Dan Manhattan Distance Pada Penentuan Zonasi Covid-19 Di Kabupaten Malang," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 493–500, 2022, doi: 10.36040/jati.v6i2.4808.

[14] M. Fajri and A. Primajaya, "Komparasi Teknik Hyperparameter Optimization pada SVM untuk Permasalahan Klasifikasi dengan Menggunakan Grid Search dan Random Search," *J. Appl. Informatics Comput.*, vol. 7, no. 1, pp. 14–19, 2023, doi: 10.30871/jaic.v7i1.5004.

[15] P. Chaurasia, "Confusion Matrix," 2016. https://mgcub.ac.in/pdf/material/20200429020322e5dac20f58.pdf (accessed Oct. 02, 2023).