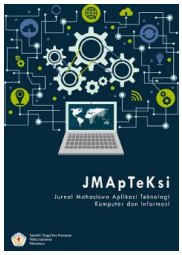


Jurnal Mahasiswa Aplikasi Teknologi Komputer dan Informasi

Vol. 7 No. 1 Thn 2025 Hal 46-52
E ISSN: 2685-6565

Klasifikasi Kualitas Udara Menggunakan PCA Dan KNN

Aldy Syahputra Harahap[a], Dewi Nasien[b], Deny Jollyta[c], Erlin[d]

[a] *Institut Bisnis Dan Tenologi Pelita Indonesia, aldy.syahputra.h@student.pelitaindonesia.ac.id*
[b] *Institut Bisnis Dan Tenologi Pelita Indonesia, dewi.nasien@lecturer.pelitaindonesia.ac.id*
[c] *Institut Bisnis Dan Tenologi Pelita Indonesia, deny.jollyta@lecturer.pelitaindonesia.ac.id*
[d] *Institut Bisnis Dan Tenologi Pelita Indonesia, erlin@lecturer.pelitaindonesia.ac.id*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 25 April 2025
Revisi Akhir: 29 April 2025
Diterbitkan *Online*: 30 April 2025

KATA KUNCI

Klasifikasi Kualitas Udara, Principal Component Analysis (PCA), K-Nearest Neighbor (KNN)

KORESPONDENSI

E-mail:
aldy.syahputra.h@student.pelitaindonesia.ac.id

A B S T R A C T

Penelitian ini bertujuan untuk membangun model klasifikasi kualitas udara yang akurat dan efisien menggunakan ekstraksi fitur Principal Component Analysis (PCA) dan metode K-Nearest Neighbor (KNN). Kualitas udara menjadi isu penting bagi kesehatan manusia dan lingkungan, sehingga diperlukan sistem klasifikasi yang efektif untuk membantu masyarakat dan pemangku kepentingan dalam mengambil langkah-langkah pencegahan. Metode PCA digunakan untuk mereduksi dimensi data kualitas udara yang berdimensi tinggi tanpa kehilangan informasi penting. Hal ini meningkatkan efisiensi dan akurasi model klasifikasi. Selanjutnya, metode KNN digunakan untuk mengklasifikasikan kualitas udara berdasarkan fitur-fitur yang telah diekstraksi oleh PCA. KNN merupakan algoritma klasifikasi yang sederhana, mudah dipahami, dan adaptif terhadap data yang tidak terdistribusi normal. Penelitian ini menggunakan dataset kualitas udara dari stasiun pemantauan di beberapa wilayah. Data dipre-proses dan dianalisis untuk memastikan kesesuaian dengan metode PCA dan KNN. Model klasifikasi kemudian dievaluasi berdasarkan akurasi dan presisi klasifikasinya. Hasil penelitian menunjukkan bahwa model klasifikasi yang menggunakan kombinasi PCA dan KNN mampu menghasilkan akurasi dan presisi yang tinggi. Berdasarkan pengujian yang telah dilakukan didapatkan hasil akurasi mencapai 98,6% dengan presisi sebesar 87%, recall sebesar 75%, f-measure sebesar 80%. Model ini terbukti efektif dalam mengklasifikasikan kualitas udara dan dapat memberikan informasi yang bermanfaat bagi masyarakat dan pemangku kepentingan terkait.

1. PENDAHULUAN

Dalam kehidupan sehari-hari, udara digunakan untuk bernafas oleh makhluk hidup. Udara yang bersih mengandung banyak manfaat bagi kehidupan. Namun, pada kenyataannya udara yang ada di alam tidak selalu dalam keadaan bersih, sehingga dapat menyebabkan penurunan kualitas udara. Kualitas udara seperti ini dapat memberikan dampak terhadap kesehatan manusia serta lingkungan sekitarnya[1].

Kualitas udara merupakan faktor penting bagi kesehatan manusia dan merupakan perhatian jangka panjang. Terutama di daerah perkotaan, karena akan berpengaruh langsung terhadap kesehatan masyarakat maupun kenyamanan kota. Timbulnya kualitas udara ini berasal dari aktivitas alam maupun dari aktivitas manusia. Kualitas udara cukup memprihatinkan dalam kondisi sekarang ini. Banyak sekali kegiatan manusia yang dibuat sehingga menghasilkan pencemaran udara. Berbagai pencemaran udara dapat dibilang berasal dari kegiatan antara lain, transportasi, pabrik industri, perkantoran, dan perumahan[2].

Kualitas udara yang buruk dapat menyebabkan berbagai masalah kesehatan, termasuk gangguan pernapasan, penyakit jantung, dan dampak negatif lainnya. Oleh karena itu, identifikasi dan pemantauan polutan utama dalam udara menjadi sangat penting untuk mengambil tindakan yang tepat dalam menjaga kualitas udara yang baik.

Data merupakan suatu nilai, hasil, angka yang dapat diolah menjadi sesuatu. Data preprocessing merupakan hal yang penting karena dapat mengurangi data spike karena nilai yang diberikan sensor kadangkala dapat berubah terlalu terlalu rendah sepersekian detik akibat dari pengaruh internal maupun ekternal dari komponen. Analisis komponen utama (PCA) merupakan suatu teknik analisis data multivariable yang digunakan untuk mengurangi dimensi data set[3].

K-Nearest Neighbor (KNN) yaitu suatu algoritma yang mempunyai fungsi dalam memprediksi dan mengidentifikasi. Algoritma ini termasuk jenis algoritma supervised learning, di mana algoritma ini mengklasifikasikan objek data baru menurut data-data yang mempunyai jarak terdekat. Berdasarkan penelitian ini mengenai pemanfaatan metode algoritma K-Nearest Neighbor

untuk memprediksi curah hujan dalam kurun waktu 1 tahun dengan nilai persentase terbaik 82,46%[4].

Latar belakang ini mencerminkan urgensi pengembangan model klasifikasi kualitas udara menggunakan ekstraksi fitur Principal Component Analysis dan pendekatan K-Nearest Neighbors. Metode ini dapat membantu pihak berwenang, peneliti, dan masyarakat untuk lebih memahami polutan udara yang dominan dalam wilayah tertentu, mengidentifikasi sumber polusi, dan mengambil tindakan yang diperlukan untuk mengurangi dampak negatifnya. Dengan memanfaatkan data pengukuran udara yang tersedia dan teknik supervised learning seperti KNN, akan menciptakan alat yang lebih efektif dalam memantau dan menjaga kualitas udara yang baik, serta mengurangi risiko terkait polutan udara terhadap kesehatan dan lingkungan.

2. TINJAUAN PUSTAKA

2.1. Pengenalan Pola

Pengenalan pola adalah proses pengidentifikasian, pemodelan, dan pengklasifikasian pola dalam data. Tujuannya adalah untuk menemukan hubungan atau struktur dalam data yang dapat digunakan untuk memahami, mengklasifikasikan, atau membuat prediksi tentang data yang baru atau tidak terlabel[5].

Pola adalah intensitas yang terdefinisi dan dapat didefinisikan melalui ciri-cirinya (*feature*). Ciri-ciri tersebut digunakan untuk membedakan suatu pola dengan pola lainnya[6].

2.1.1. Preprocessing

Preprocessing data penting dalam analisis data mining untuk membersihkan, mengubah format, dan mempersiapkan data agar lebih mudah dan akurat[7].

Data *preprocessing* adalah tahap untuk melakukan sebuah proses awal dalam pengolahan data. Pada tahap ini data yang akan diolah bertujuan untuk menghindarkan dari data yang mengganggu (noise) atau data yang tidak konsisten[8].

2.1.2. Ekstraksi Fitur (PCA)

Principal Component Analysis adalah proses reduksi dimensi yang umumnya diterapkan pada tahap *pre-processing* data bertujuan untuk mengurangi jumlah fitur (dimensi) tanpa menghilangkan informasi penting dari suatu data[9].

2.1.3. Klasifikasi (KNN)

Klasifikasi adalah salah satu teknik yang dapat digunakan untuk dapat membedakan antar sebuah objek. Klasifikasi harus menggunakan metode yang tepat agar tercapai hasil akurasi yang maksimal[10].

K-Nearest Neighbor adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus yang lama, yaitu berdasarkan pada pencocokan bobot dari fitur yang ada. Algoritma ini merupakan salah satu teknik untuk mencari jarak terdekat dari tiap-tiap kasus yang ada didalam database, dan seberapa mirip ukuran kemiripan setiap source case yang ada didalam database dengan target case[11].

2.2. Kualitas Udara

Kualitas udara merupakan faktor penting bagi kesehatan manusia dan merupakan perhatian jangka panjang. Terutama di daerah perkotaan, karena akan berpengaruh langsung terhadap kesehatan masyarakat maupun kenyamanan kota[12].

Indeks Standar Pencemaran Udara (ISPU) merupakan angka yang tidak mempunyai satuan dalam menggambarkan kondisi kualitas udara di lokasi dan waktu tertentu untuk dapat melihat dampak terhadap kesehatan manusia. Berikut adalah tabel indeks standar pencemaran udara[13].

Tabel 1 : Indeks Standar Pencemaran Udara[13].

ISPU	Pencemaran Udara Level	Dampak Kesehatan
0 - 50	Baik	Tidak memberikan dampak bagi Kesehatan Manusia atau hewan
51-100	Sedang	Tidak berpengaruh pada Kesehatan manusia Ataupun hewan tetapi berpengaruh pada Tumbuhan yang peka
101-199	Tidak Sehat	Bersifat merugikan pada manusia ataupun kelompok hewan yang peka atau dapat menimbulkan kerusakan pada tumbuhan dan nilai estetika
200-299	Sangat Tidak	Kualitas udara yang dapat merugikan Kesehatan pada sejumlah segmen populasi yang terpapar
300-500	Sehat	Kualitas udara berbahaya yang secara umum dapat merugikan Kesehatan yang serius pada populasi (misalnya iritasi mata, batuk, dahak, dan sakit tenggorokan)

2.3. Bahasa Pemrograman (Python)

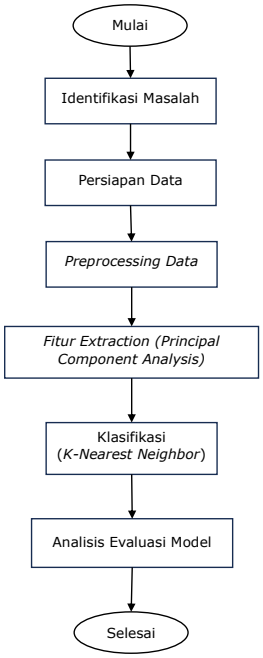
Python adalah bahasa pemrograman tingkat tinggi yang berorientasi objek dan dibuat oleh Guido van Rossum. *Python* adalah bahasa lintas fungsi yang ditafsirkan secara maksimal yang memiliki banyak fungsi. Bahasa pemrograman berorientasi objek ini biasanya digunakan untuk merampingkan kumpulan data kompleks yang besar. Fitur lain yang membuat *Python* ini terkenal di kalangan ilmuwan dan penganalisis data adalah fleksibilitasnya. Karena fleksibilitasnya, bahasa ini memungkinkan untuk membuat data model, mensistemasikan kumpulan data, membuat algoritma yang didukung ML (Machine Learning), layanan web, dan menerapkan penambangan data untuk menyelesaikan berbagai tugas dalam waktu singkat[14].

Python adalah bahasa pemrograman open-source, efisien, dan mudah. Bagian terbaik dari bahasa *python* adalah pengguna tidak perlu menulis baris kode yang besar untuk beroperasi. Bahkan masalah yang sangat kompleks dapat diselesaikan hanya dengan menulis beberapa baris kode[15].

3. METODOLOGI

3.1. Kerangka Penelitian

Flowchart kerangka penelitian yang disajikan pada **Gambar 1** ini merupakan gambaran visual dari langkah-langkah penelitian yang akan dilakukan.



Gambar 1. Kerangka Penelitian

3.1.1. Identifikasi Masalah

Penelitian ini melakukan pengidentifikasian masalah yang terjadi yaitu permasalahan mengenai penurunan kualitas udara yang terjadi dengan membaca beberapa artikel dan jurnal terkait.

3.1.2. Persiapan Data

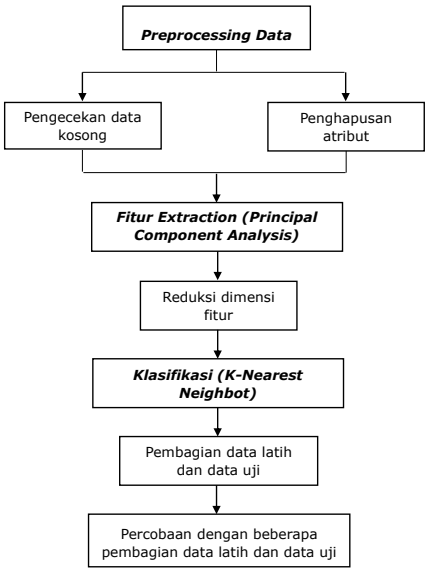
Penelitian ini memperoleh dataset penelitian dari website Kaggle.com dengan laman URL https://www.kaggle.com/code/moneytadholah/analisis-desember-2021/input. Dataset yang digunakan memuat informasi kualitas udara di Yogyakarta pada bulan April 2021. Dataset memuat total sebanyak 720 record data dengan 11 atribut.

Terdapat 6 variabel yang terdiri dari date, time, pm2.5, pm10, so2, co, o3, dan no2. Penjelasan lebih lanjut dapat dilihat dibawah ini :

1. Date dan Time
   Merupakan waktu perekaman keadaan udara setiap jam di setiap harinya selama sebulan tersebut.
2. PM2.5 dan PM10
   PM adalah singkatan dari particulate matter atau partikulat atmosfer. PM 2.5 adalah polutan udara yang berukuran sangat kecil, sekitar 2,5 mikron (mikrometer). Begitu pula PM10 adalah polutan udara yang berukuran 10 mikron.
   PM 2.5 dibentuk dari emisi pembakaran bensin, minyak, bahan bakar, dan kayu. Sementara itu, PM 10 ditemukan pada tempat pembangunan, pembuangan sampah, pertanian, kebakaran hutan, debu, serbuk sari, dan fragmen bakteri.
3. SO2
   Sulfur dioksida adalah salah satu spesies dari gas-gas oksida sulfur (SO2). Gas ini sangat mudah terlarut dalam air, memiliki bau namun tidak berwarna, SO2 dan gas-gas oksida sulfur lainnya terbentuk saat terjadi pembakaran bahan bakar fosil yang mengandung sulfur.
4. CO
   Karbon monoksida (CO) merupakan salah satu gas hasil pembakaran tidak sempurna kendaraan bermotor yang dapat mencemari udara dan mengganggu kesehatan manusia.
5. O3
   Ozon Permukaan (O3) merupakan salah satu polutan sekunder yang terbentuk dari hasil reaksi fotokimia. Ozon ini merupakan polutan udara yang berdampak negatif bagi lingkungan, alam maupun manusia.
6. NO2
   Nitrogen dioksida (NO2) adalah kontributor utama pembentukan kabut asap dan prekursor dari banyak polutan sekunder yang berbahaya, termasuk ozon dan partikel. Nitrogen dioksida sangat reaktif dengan bahan kimia lain dan merupakan agen pengoksidasi yang kuat. Nitrogen dioksida adalah gas berwarna merah-oranye tua

3.2. Metode Penelitian

Pada penelitian ini terdapat beberapa proses pengolahan data, untuk lebih spesifik dapat dilihat pada **Gambar 2** dibawah ini



Gambar 2. Gambar kerangka metode penelitian

3.2.1. Preprocessing Data

Tahap *preprocessing* yang akan dilakukan yaitu pembersihan data atau *data cleansing*. Pembersihan data yang akan dilakukan yaitu dengan melakukan pencarian data yang masih kosong, menghapusan baris atau kolom yang kosong, dan kemudian mengisi kembali data pada kolom yang masih kosong.

3.2.2. *Fitur Extraction (PCA)*

PCA digunakan untuk mengurangi dimensi data dengan mempertahankan informasi yang relevan. Ini membantu dalam mengatasi masalah dimensionalitas tinggi dan mengidentifikasi pola yang tersembunyi dalam data. Dengan mengurangi dimensi, *PCA* mempermudah analisis data dan memungkinkan kita untuk fokus pada fitur-fitur yang paling penting.

3.2.3. *Klasifikasi (K-Nearest Neighbor)*

Pengolahan data dilakukan dengan membagi dataset yang telah melewati tahap sebelumnya menjadi 2 (dua) yakni data uji serta data latih lalu diterapkan algoritma *K-Nearest Neighbor*. Cara kerja algoritma tersebut yakni melakukan klasifikasi terhadap objek baru dengan melakukan perhitungan jarak terdekat objek tersebut terhadap data-data yang sudah ada. Jarak yang terdapat diantara kedua objek data tersebut dapat dihitung dengan menggunakan perhitungan *euclidean distance.*

3.3. *Analisis Evaluasi Model*

Metode analisis evaluasi model yang digunakan yaitu menggunakan *confusion matrix*. Analisis performa ini melakukan perbandingan hasil *akurasi*, *presisi*, *recall*, dan *f-measure* menurut parameter nilai K yang telah digunakan. *Confusion matrix* merupakan salah satu metode evaluasi model yang dapat merepresentasikan informasi dari perbandingan hasil prediksi yang dilakukan dan kondisi yang sebenarnya.

1. *Akurasi*
   Akurasi adalah tingkatan dekatnya nilai prediksi serta nilai aktual atau nilai sesungguhnya. Seperti yang terdapat pada rumus (3.1)
   $$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.1)$$
2. *Presisi*
   Presisi merupakan dipilihnya rasio item relevan berdasarkan seluruh item terpilih. Seperti yang terdapat pada rumus (3.2)
   $$Presisi = \frac{TP}{TP+FP} \quad (3.2)$$
3. *Recall*
   Recall merupakan dipilihnya rasioitem relevan berdasarkan total item relevan yang ada. Seperti yang terdapat pada rumus (3.3)
   $$Recall = \frac{TP}{TP+FN} \quad (3.3)$$
4. *F-measure*
   F-measure atau yang juga disebut F1-Score merupakan harmonic mean dari *presisi* serta *recall*. Seperti yang terdapat pada rumus (3.4)
   $$\text{F1-Score} = 2\frac{Presisi \text{ x } Recall}{Presisi+Recall} \quad (3.4)$$

3.4. *Teknik Pengujian Manual*

Perhitungan manual ini bertujuan untuk menjelaskan cara kerja algoritma K-Nearest Neighbor. Data latih yang digunakan diambil dari dataset yang digunakan dalam penelitian, dengan mengambil 10 data secara acak sebagai contoh perhitungan. Berikut ini adalah data latih yang digunakan :

Tabel 2 : Data latih perhitungan manual

NO	PM2.5	PM10	SO2	CO	O3	NO2	CATEGORY
1	45	19	21	15	8	3	Baik
2	62	32	12	14	52	7	Sedang
3	61	31	12	14	51	7	Sedang
4	19	8	15	10	20	3	Baik
5	44	18	20	14	8	3	Baik
6	35	14	19	12	7	3	Baik
7	25	10	16	11	18	25	Baik
8	34	13	15	8	51	34	Sedang
9	34	13	14	8	65	34	Sedang
10	34	13	16	9	59	34	Sedang

Selanjutnya akan dilakukan penentuan *category* dari 1 contoh data uji yang belum diketahui *category*-nya. Berikut ini data uji yang akan digunakan untuk perhitungan manual yakni pada **Tabel 3.**

Tabel 3 : Data uji perhitungan manual

PM2.5	PM10	SO2	CO	O3	NO2	CATEGORY
37	15	16	11	20	3	n/a

Berikut ini langkah-langkah perhitungan manual algoritma K-Nearest Neighbor:

1. Menentukan nilai K sebagai parameter banyaknya jumlah tetangga terdekat dengan objek yang akan diuji. Dalam perhitungan ini, akan menggunakan nilai K minimal yang akan digunakan untuk evaluasi model pada penelitian ini, yaitu K = 3.
2. Menghitung jarak antara objek yang baru terhadap semua objek data yang ada di data latih. Perhitungan jarak dilakukan pada setiap baris data dengan memasukan nilai-nilai yang ada di data latih dan data uji ke dalam rumus
   $$euclidean\ distance = \sqrt{(a_1-b_1)^2+\cdots+(a_n-b_n)^2}$$

$$d1 = \sqrt{(a_1-b_1)^2+\cdots+(a_n-b_n)^2}$$
$$d1 = \sqrt{(37-45)^2+(15-19)^2+(16-21)^2+(11-15)^2+(20-8)^2+(3-3)^2}$$
$$d1 = \sqrt{(-8)^2+(-4)^2+(-5)^2+(-4)^2+(12)^2+(0)^2}$$
$$d1 = \sqrt{64+16+25+16+144+0}$$
$$d1 = \sqrt{265}$$
$$d1 = 16{,}27882$$

Dan seterusnya hingga perhitungan ke-10

3. Melakukan pengurutan hasil perhitungan jarak dari yang paling terkecil ke terbesar. Dapat dilihat pada **Tabel 4**

Tabel 4 : Hasil perhitungan manual

Data latih ke -	d	K=3	Category
d6	13,56466	YA	BAIK
d5	15,0665192	YA	BAIK
d1	16,2788206	YA	BAIK
d4	19,3649167	TIDAK	BAIK
d7	25,6320112	TIDAK	BAIK
d3	42,8252262	TIDAK	SEDANG
d8	44,1021541	TIDAK	SEDANG
d2	44,4859528	TIDAK	SEDANG
d10	49,989999	TIDAK	SEDANG
d9	54,8816909	TIDAK	SEDANG

4. Menentukan kategori dari tetangga terdekat dari objek baru yang diuji. Berdasarkan **Tabel 4**, maka dapat disimpulkan bahwa kelas dari data uji yang digunakan termasuk dalam *category* BAIK karena 3 (tiga) tetangga terdekatnya memiliki kelas mayoritas yaitu BAIK.

4. HASIL DAN PEMBAHASAN

4.1. Persiapan Data

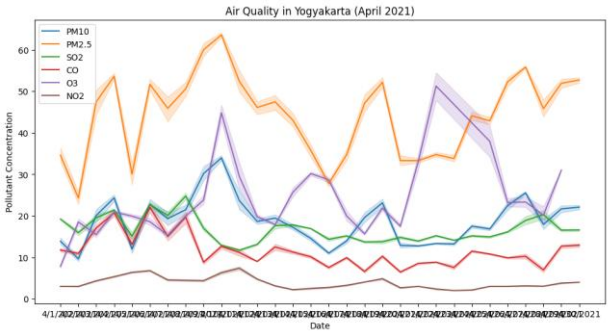
Dataset yang digunakan pada penelitian ini memuat informasi kualitas udara di Yogyakarta pada bulan April 2021. Dataset memuat total sebanyak 720 record data dengan 11 atribut, dan data yang diperoleh berformat *csv*. Seperti pada **Gambar 3.**

	Date	Time	PM2.5	PM10	SO2	CO	O3	NO2	Max	Critical Component	Category
0	4/1/2021	00:00:00	45	19	21	15	8.0	3	21	PM2.5	Good
1	4/1/2021	01:00:00	44	18	20	14	8.0	3	20	PM2.5	Good
2	4/1/2021	02:00:00	43	17	20	14	7.0	3	20	PM2.5	Good
3	4/1/2021	03:00:00	40	17	20	13	7.0	3	20	PM2.5	Good
4	4/1/2021	04:00:00	38	16	19	12	7.0	3	19	PM2.5	Good
...	...	...	...	...	...	...	...	...	...	...	...
715	4/30/2021	19:00:00	54	23	17	14	NaN	4	23	PM2.5	Good
716	4/30/2021	20:00:00	54	23	17	14	NaN	4	23	PM2.5	Good
717	4/30/2021	21:00:00	54	23	17	13	NaN	4	23	PM2.5	Good
718	4/30/2021	22:00:00	54	23	17	13	NaN	4	23	PM2.5	Good
719	4/30/2021	23:00:00	54	23	17	12	NaN	4	23	PM2.5	Good

720 rows × 11 columns

Gambar 3. Tampilan dataset

Berikut gambaran dataset yang diperoleh jika divisualisasikan, sehingga memudahkan pembaca untuk melihat data secara sekilas, seperti pada **Gambar 4.**



Gambar 4. Visualisasi dataset

4.2. Preprocessing Data

Pada penelitian ini dilakukan pembersihan data atau *data cleansing*. Jika ada data yang kosong akan dihapus dan dilakukan pengisian ulang kolom yang kosong, serta menghapus atribut yang tidak diperlukan dalam pengolahan data nantinya. Bertujuan untuk mengurangi kompleksitas atribut saat penerapan algoritma.

4.2.1. Pengecekan data kosong

Lakukan pengecekan pada dataset apakah terdapat nilai kosong memakai kodingan seperti pada **Gambar 5.**

```
Date                  0
Time                  0
PM2.5                 0
PM10                  0
SO2                   0
CO                    0
O3                   99
NO2                   0
Max                   0
Critical Component    0
Category              0
dtype: int64
```

Gambar 5. Pengecekan data kosong

Terdapat 99 data kosong pada atribut ‘O3’, maka kita perlu mengganti data tersebut agar tidak terjadi kesalahan pada pengolahan dataset nantinya. Dan lakukan pengecekan kembali apakah masih terdapat nilai kosong, seperti pada **Gambar 6.**

```
Date                  0
Time                  0
PM2.5                 0
PM10                  0
SO2                   0
CO                    0
O3                    0
NO2                   0
Max                   0
Critical Component    0
Category              0
dtype: int64
```

Gambar 6. Setelah data diperbaiki

Setelah data diperbaiki dapat dilihat sudah tidak ada data yang null atau kosong, maka proses dapat dilanjutkan.

4.2.2. Penghapusan Atribut

Pada tahap ini atribut yang tidak diperlukan dihapus untuk pengolahan dataset nantinya, agar mempermudah penerapan algoritma di tahap selanjutnya. Atribut yang akan dihapus, yaitu ‘*Date* (tanggal)’, ‘*Time* (waktu)’, dan ‘*Critical Component’*. Seperti pada **Gambar 7.**

	PM2.5	PM10	SO2	CO	O3	NO2	Max	Category
0	45	19	21	15	8.0	3	21	Good
1	44	18	20	14	8.0	3	20	Good
2	43	17	20	14	7.0	3	20	Good
3	40	17	20	13	7.0	3	20	Good
4	38	16	19	12	7.0	3	19	Good

Gambar 7. Penghapusan atribut tidak perlu

4.3. Fitur Extraction (PCA)

Penerapan *PCA* adalah untuk mereduksi dimensi fitur sebelum diterapkan ke dalam algoritma *KNN.* Mereduksi dimensi fitur menjadi 2 komponen, yakni ‘PC1’ dan ‘PC2’. Yang mana dari hasil diatas menunjukkan 50,2% dapat dijelaskan dengan komponen ‘PC1’ dan 39,3% dapat dijelaskan dengan komponen ‘PC2’.

4.4. Klasifikasi (K-Nearest Neighbor)

Dalam klasifikasi ini data dibagi menjadi data uji dan data latih untuk *KNN* sebanyak 3 percobaan, dan akan dicari hasil terbaik dari ketiga percobaan tersebut.

1. Pada percobaan pertama membagi data latih dan data uji sebanyak 80% untuk latih dan 20% untuk uji. Setelah dibagi menjadi 80:20 didapat data latih sebanyak 576 data dan data uji sebanyak 144 data.
2. Pada percobaan kedua membagi data latih dan data uji sebanyak 70% untuk latih dan 30% untuk uji. Setelah dibagi menjadi 70:30 didapat data latih sebanyak 504 data dan data uji sebanyak 216 data.
3. Pada percobaan ketiga membagi data latih dan data uji sebanyak 60% untuk latih dan 40% untuk uji. Setelah dibagi menjadi 60:40 didapat data latih sebanyak 432 data dan data uji sebanyak 288 data.

Hasil dari ketiga percobaan tersebut dapat dilihat pada **Tabel 5.**

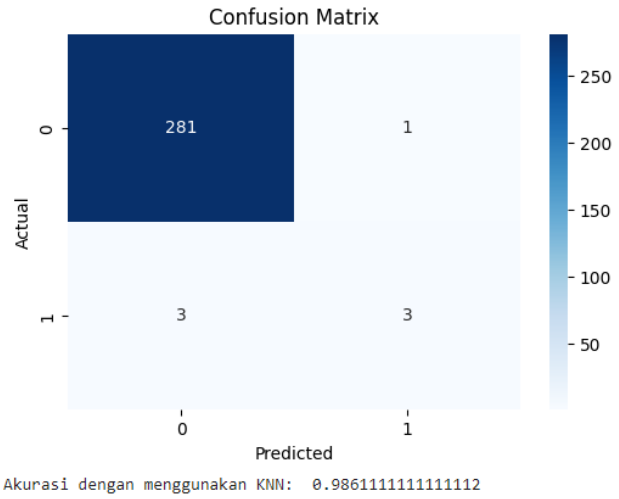
Tabel 5. Perbandingan hasil ketiga percobaan

Perbandingan Data Latih dan Data Uji	Akurasi
80 : 20	97,222%
70 : 30	98,148%
60 : 40	98,611%

Berdasarkan hasil yang didapatkan dari ketiga percobaan diatas, maka penelitian ini mengambil pembagian data latih dan data uji sebesar 60:40. Karena pembagian data ini yang mendapatkan tingkat akurasi tertinggi dengan nilai 98,6%.

4.5. Analisis Evaluasi Model

Pada tahap ini akan dicari *confusion matrix* sebagai sebagai perbandingan hasil *akurasi*, *presisi*, *recall*, dan *f-measure* menurut pembagian data dan parameter tertangga terdekat (nilai K) yang telah digunakan, seperti pada **Gambar 8.**



Gambar 8. Confusion matrix

Setelah di dapat *confusion matrix* maka bisa didapatkan hasil *akurasi*, *presisi*, *recall*, dan *f-measure,* seperti pada **Gambar 9.**

```
Akurasi: 0.9861111111111112
Precision: [0.98943662 0.75      ]
Recall: [0.9964539 0.5      ]
F1 Score: [0.99293286 0.6       ]
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       282
           1       0.75      0.50      0.60         6

    accuracy                           0.99       288
   macro avg       0.87      0.75      0.80       288
weighted avg       0.98      0.99      0.98       288
```

Gambar 9. Hasil analisis evaluasi model

Berdasarkan keseluruhan hasil penelitian dapat dilihat bahwa hasil terbaik diperoleh dari pembagian data perbandingan 60:40 menggunakan nilai K=3 (tetangga terdekat), dengan hasil akurasi sebesar 98,6%, presisi sebesar 87%, recall sebesar 75%, f-measure sebesar 80%.

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan penelitian yang telah dilakukan untuk mengklasifikasi kualitas udara menggunakan ekstraksi fitur principal component analysis dan metode k-nearest neighbor, langkah yang dilakukan meliputi mengidentifikasi masalah, persiapan dataset, preprocessing data, mereduksi fitur menggunakan PCA, klasifikasi menggunakan algoritma KNN, dan terakhir analisis evaluasi model.

Berdasarkan hasil pengujian pada bab sebelumnya, dilakukan 3 kali percobaan pada pembagian data latih dan uji yang berbeda-beda, yakni dengan perbandingan 80:20, 70:30, dan 60:40. Dan didapatkan hasil terbaik dengan perbandingan data latih dan uji pada perbandingan 60:40 dengan nilai K=3. Tingkat akurasi mencapai 98,6%, presisi 87%, recall 75%, dan f-measure 80%.

5.2. Saran

Perlu dilakukan pengembangan dengan menerapkan algoritma lain dalam pengklasifikasian masalah ini. Agar dapat dilakukan perbandingan algoritma mana yang terbaik dalam mengklasifikasikan masalah kualitas udara. Dan juga diharapkan dilakukan pengembangan aplikasi berdasarkan model yang telah dihasilkan, dengan harapan agar penelitian ini dapat berguna bagi semua orang, baik orang awam sebagai pembaca saja atau bagi pemangku kepentingan dalam mengambil Keputusan.

DAFTAR PUSTAKA

[1] Syihabuddin Azmil Umri, S. (2021). Analisis Dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara Di Dki Jakarta. JIKO (Jurnal Informatika Dan Komputer), 4(2), 98–104. https://doi.org/10.33387/jiko.v4i2.2871

[2] Pratiwi, B. P., Handayani, A. S., & Sarjana, S. (2021). Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi Wsn Menggunakan Confusion Matrix. Jurnal Informatika Upgris, 6(2), 66–75. https://doi.org/10.26877/jiu.v6i2.6552

[3] Putra, G. T., Kallista, M., & ... (2023). Pembelajaran Mesin Pada Data Preprocessing Dengan Metode Principal Component Analysis Dan Smote. EProceedings …, 10(5), 4002–4010. https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/21085

[4] Reza Noviansyah, M., Rismawan, T., Marisa Midyanti, D., Sistem Komputer, J., & MIPA Universitas Tanjungpura Jl Hadari Nawawi, F. H. (2018). Penerapan Data Mining Menggunakan Metode K-Nearest Neighbor Untuk Klasifikasi Indeks Cuaca Kebakaran Berdasarkan Data Aws (Automatic Weather Station) (Studi Kasus: Kabupaten Kubu Raya). Jurnal Coding, Sistem Komputer Untan, 06(2), 48–56.

[5] Wahyu, E., Nasution, A. Y., & Rosnelly, R. (2024). Pengenalan Pola Aksara Batak menggunakan Backpropagation Recognition of Batak Script Patterns using Backpropagation. 14(1), 57–67.

[6] Fadlisyah, F., Ula, M., & Nasriah, N. (2020). Implementasi Pengenalan Pola Alif Lam Qamariah Pada Huruf Hijaiyah Menggunakan Metode Cosineimplementasi Pengenalan Pola Alif Lam Qamariah Pada Huruf Hijaiyah Menggunakan Metode Cosine. TECHSI - Jurnal Teknik Informatika, 12(1), 52.

https://doi.org/10.29103/techsi.v12i1.1285

[7] Agung, A., Daniswara, A., Kadek, I., & Nuryana, D. (2023). Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru. Journal of Informatics and Computer Science, 05, 97–100.

[8] Alghifari, F., & Juardi, D. (2021). Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes. Jurnal Ilmiah Informatika, 9(02), 75–81. https://doi.org/10.33884/jif.v9i02.3755

[9] Ferliadi, H. Sulistiani, and F. Hamidy, "Sistem Informasi Manajemen Aset Dan Keuangan," *J. Ilm. Sist. Inf. Akunt.*, vol. 1, no. 2, pp. 7–15, 2021, doi: 10.33365/jimasia.v1i2.1103.

[10] Suwitono, Y. A., & Kaunang, F. J. (2022). Implementasi Algoritma Convolutional Neural Network (CNN) Untuk Klasifikasi Daun Dengan Metode Data Mining SEMMA Menggunakan Keras. Jurnal Komtika (Komputasi Dan Informatika), 6(2), 109–121. https://doi.org/10.31603/komtika.v6i2.8054

[11] Anderio, J., & Johan. (2019). Sistem Pendukung Keputusan Pemilihan Material Bangunan Berdasarkan Kesesuain Budget Konsumen Menggunakan K-Nearest Neighbor (KNN). (Studi Kasus : Toko Bangunan AJJ). Jurnal Mahasiswa Aplikasi Teknologi Komputer Dan Informasi, 1(Thn), 12–19.

[12] Pratiwi, B. P., Handayani, A. S., & Sarjana, S. (2021). Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi Wsn Menggunakan Confusion Matrix. Jurnal Informatika Upgris, 6(2), 66–75. https://doi.org/10.26877/jiu.v6i2.6552

[13] Jayadi, B. V., Handhayani, T., Lauro, M. D., Informatika, T., & Tarumanagara, U. (2023). Perbandingan Knn Dan Svm Untuk Klasifikasi Kualitas Udara Di Jakarta. Jurnal Ilmu Komputer Dan Sistem Informasi, Vol. 11(No. 2).

[14] Junaidi, S., Devegi, M., & Kurniawan, H. (2023). Pelatihan Pengolahan dan Visualisasi Data Penduduk menggunakan Python. ADMA : Jurnal Pengabdian Dan Pemberdayaan Masyarakat, 4(1), 151–162. https://doi.org/10.30812/adma.v4i1.2963

[15] Pankaj Dumka, Kritik Rana, Surya Pratap Singh Tomar, Parth Singh Pawar, D. R. M. (2022). Modelling air standard thermodynamic cycles using python. Advances in Engineering Software, Volume 172(103186).