JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i1.5664

PENENTUAN MODEL ALGORITMA KLASIFIKASI TERBAIK UNTUK KLASIFIKASI KUALITAS UDARA DI JAKARTA 2023

Ahmad Rafi Kannajmi[1*], Dicky Saputra[2]

[1,2]Jurusan Informatika, Universitas Sultan Ageng Tirtayasa; Fakultas Teknik: Jln. Jend. Sudirman KM. 3, Cilegon, Banten 42435.

Received: 11 Desember 2024
Accepted: 14 Januari 2025
Published: 20 Januari 2025

Keywords:
Air Quality, Classification, Machine Learning, Data Mining.

Corespondent Email:
3337220031@untirta.ac.id

Abstrak. Pesatnya urbanisasi dan kemajuan sektor manufaktur telah mengakibatkan banyaknya pencemaran terhadap lingkungan, terutama pencemaran udara. Jakarta adalah kota dengan penduduk terpadat di Indonesia. Hal-hal seperti transportasi, pembakaran hutan untuk wilayah industri, serta pabrik-pabrik yang aktif setiap hari, serta ditambah bencana alam seperti kemarau, menjadikan kualitas udara di Jakarta memburuk setiap harinya. Dampak yang ditimbulkan berpotensi merugikan kondisi kesehatan penduduk. Mengingat banyaknya penduduk di Jakarta, maka diperlukan perhatian tinggi terhadap hal ini. Dengan demikian, sangat mendesak untuk mengembangkan pendekatan yang mampu mengategorikan tingkat polusi udara, sehingga masyarakat dan otoritas terkait dapat memperoleh informasi akurat tentang kondisi lingkungan mereka. Dalam riset ini, data kualitas udara atau Indeks Standar Pencemar Udara (ISPU) dikumpulkan melalui berbagai pos Sistem Pemantauan Kualitas Udara (SPKU) di Jakarta. Data tersebut kemudian diolah dan digunakan untuk melatih beberapa model algoritma klasifikasi, seperti Naïve Bayes, Random Forest, Neural Network, K-Nearest Neighbors (KNN), Decision Tree and Support Vector Machine (SVM). Performa model-model tersebut akan dievaluasi berdasarkan akurasi, recall, F1-score, dan presisi. Hasil riset memperlihatkan model Random Forest memiliki performa terbaik di antara model lainnya dengan akurasi 95,3%, recall 95,3%, F1-score 95,1%, dan presisi 95,2%. Dapat dikatakan model ini bekerja dengan baik dalam mengklasifikasikan kualitas udara.

Abstract. *The growth of cities and the industrial development that follows has resulted in pollution to the environment, especially air pollution. Jakarta by far the most crowded city in Indonesia. Transportation, burning forests for industrial areas, and active factories, along with natural disasters like droughts, make Jakarta's air quality deteriorate. This can harm public health. Considering the large population in Jakarta, there is a need for attention to this matter. Therefore, a method is needed that can classify air quality into certain classes so that the community and government knows how good and polluted the air around. In this study, air quality data or ISPU was collected from various SPKU stations in Jakarta. The data is processed to train several classification algorithm models, such as Naïve Bayes, Random Forest, Neural Network, K-Nearest Neighbors (KNN), Decision Tree and Support Vector Machine (SVM). These model's performance evaluated according to the accuracy, recall, F1-score, and precision of each model's. The outcome showed Random Forest has the best performance among the rest of the models*

with 95.3% recall, 95.2% precision, 95.1% F1-score, and 95.3% accuracy It is safe to say that the model performs very effectively in classifying air quality.

1. PENDAHULUAN

Semakin berkembangnya kawasan perkotaan dan industri telah membangkitkan kesadaran kritis mengenai hubungan saling memengaruhi antara kontaminasi lingkungan, derajat kesehatan publik, dan sistem ekologis [1]. Konsekuensi dari proses transformasi masyarakat agraris ke masyarakat industri, terutama di daerah perkotaan dan subperkotaan, di mana terjadinya peningkatan pencemaran atau polusi lingkungan akibat dari adanya aktivitas industri, transportasi, dan pembakaran lahan atau hutan. Dampak pencemaran udara bagi kesehatan dapat menyebabkan pernapasan menjadi terganggu, mengganggu jalannya oksigen yang ada dalam darah, memicu keguguran dan autisme, meningkatkan risiko penyakit kardiovaskular, peningkatan risiko kanker, dan bahkan dapat menyebabkan kematian [2]. Lembaga kesehatan internasional WHO memperkirakan bahwa sebanyak 23 persen kematian di dunia berkaitan dengan risiko lingkungan, termasuk di dalamnya polusi udara, pencemaran air, dan pemaparan zat kimia berbahaya [3]. Tentunya hal ini perlu sekali diperhatikan, bahwasanya akibat dari buruknya kualitas udara dapat menyebabkan penyakit atau bahkan kematian seseorang. Sehingga informasi terkait kualitas udara pada suatu daerah pusat perkotaan yang padat akan penduduk, sangatlah krusial dalam menghindari dampak negatif akibat dari kualitas udara yang buruk.

DKI Jakarta merupakan daerah perkotaan pusat dan sekaligus ibu kota dari Indonesia, dengan jumlah warga sebesar 9,041 juta dan kepadatan penduduk yang sangat tinggi, yakni 13.667,01 orang di setiap kilometer perseginya [4]. Hal tersebut menjadikan Jakarta sebagai kota terpadat dibandingkan dengan kota lainnya di Indonesia. DKI Jakarta sangat cocok dijadikan objek penelitian terkait kualitas udara, karena sebagai kota metropolitan dan jantung Indonesia, Jakarta tidak terhindar dari masalah polusi udara. Di kota Jakarta, transportasi yang berupa kendaraan bermotor saja telah menjadi penyumbang emisi karbon monoksida (CO) terbesar di Jakarta, sekaligus pemicu dari meningkatnya polusi udara dari tahun ke tahun. Belum lagi pabrik-pabrik industri yang beroperasi setiap hari, termasuk Pembangkit Listrik Tenaga Uap (PLTU) yang memanfaatkan batu bara sebagai sumber energi, juga merupakan sektor terbesar yang menghasilkan sulfur dioksida (SO_2). Selain itu, faktor alam, seperti musim kemarau, dapat membuat kualitas udara di Jakarta menurun [5].

Dari hasil pengamatan langsung melalui situs IQAir yang menyediakan informasi kualitas udara secara real-time, tercatat pada awal Januari 2024. Berdasarkan data tersebut, Jakarta menempati posisi ke-15 dari kota-kota dengan kualitas udara terburuk di dunia. Indeks Kualitas Udara (AQI) sebesar 165 mengonfirmasi status udara yang tidak sehat, dengan konsentrasi polusi PM2,5 mencapai 82 mikrogram per meter kubik, jauh melampaui ambang batas aman.. Sedangkan kategori baik atau sehat memiliki rentang nilai konsentrasi PM2,5 sebesar 0-50 [6].

Pemantauan dan analisis kualitas udara di DKI Jakarta merujuk pada protokol dan standar baku yang telah ditentukan. Dinas Lingkungan Hidup Provinsi DKI Jakarta bertanggung jawab menerapkan standar tersebut, dengan pemantauan sistematis menggunakan infrastruktur Stasiun Pengendalian Kualitas Udara (SPKU). SPKU sendiri merupakan alat atau teknologi pengukur kualitas udara yang tersebar di beberapa wilayah Jakarta. Stasiun Pengendalian Kualitas Udara (SPKU) melakukan pengukuran parameter lingkungan yang selanjutnya dikompilasi dalam Indeks Standar Pencemar Udara (ISPU). Indeks ini merupakan instrumen pemantauan komprehensif yang menghasilkan laporan sistematis tentang kondisi kualitas atmosfer, mengindikasikan tingkat pencemaran dan potensi dampaknya terhadap kesehatan masyarakat. ISPU disajikan dalam format numerik tanpa satuan spesifik. [7]. ISPU berperan sebagai sistem informasi yang menjelaskan karakteristik kualitas udara di wilayah tertentu. Dengan indeks ini, pemerintah DKI Jakarta dapat memetakan tingkat pencemaran dan mengembangkan langkah-

langkah preventif serta memberikan pemahaman kepada masyarakat.

Knowledge Discovery in Database (KDD) atau yang lebih dikenal dengan *data mining* adalah teknik menganalisis dan mengekstrak informasi penting dari kumpulan data. [8]. Data mining untuk tipe prediksi dibagi menjadi tiga model analisis: klasifikasi, regresi, dan *time series* [9]. Dalam riset ini, peneliti menerapkan metode klasifikasi *data mining* untuk menganalisis kualitas udara di DKI Jakarta dengan menggunakan beragam algoritma, mencakup *Random Forest*, *Naive Bayes*, *Decision Tree*, *SVM*, *KNN*, dan *Neural Network*.

Penelitian ini bertujuan untuk memeriksa dan membandingkan akurasi atau performa model algoritma klasifikasi yang dipilih dalam mengklasifikasikan nilai polutan pada data, Apakah masuk ke dalam klasifikasi lima kategori, yaitu "Berbahaya", "Sangat Tidak Sehat", "Tidak Sehat", "Sedang", "Baik". Hasil perbandingan tersebut nantinya akan menjadi penentu algoritma mana yang paling akurat untuk dijadikan model *machine learning.*

2. TINJAUAN PUSTAKA

Penelitian sebelumnya yang melakukan prediksi kualitas udara menggunakan algoritma klasifikasi, di antaranya adalah penelitian [7]. Pada penelitian tersebut, algoritma *K-Nearest Neighbor* dipilih sebagai teknik klasifikasi. Klasifikasi dilakukan untuk membuat sebuah sistem masukan yang mana akan memberi tahu kualitas udara yang dimasukkan serta memberikan peringatan kepada pengguna.

Penelitian lainnya adalah penelitian [10], di mana pada penelitian tersebut algoritma klasifikasi yang digunakan adalah *Random Forest*. Klasifikasi dilakukan untuk melihat pengaruh perubahan pada nilai *interval tree* dan perbandingan rasio pada data latih dan data uji terhadap akurasi dari algoritma *Random Forest* pada data kualitas udara.

2.1. Klasifikasi

Klasifikasi adalah teknik di mana model atau pengklasifikasi dibangun untuk memprediksi label kategorikal, seperti “aman” atau “berisiko”. Untuk data kualitas udara yang digunakan oleh peneliti, terdapat lima kategori klasifikasi, yaitu "Berbahaya", "Sangat Tidak Sehat", "Tidak Sehat", "Sedang", "Baik". Sebagai metode *supervised learning*, klasifikasi ini menganalisis hubungan antara masukan dan target atribut. Metodenya difokuskan untuk memprediksi kelas pada objek yang label-nya belum teridentifikasi [11].

2.2. Indeks Standar Pencemaran Udara (ISPU)

Indeks Standar Pencemar Udara (ISPU) adalah parameter numerik yang menjelaskan mutu udara lingkungan di area tertentu. Perhitungannya mempertimbangkan konsekuensi terhadap organisme hidup, nilai estetis, dan kesehatan manusia. Di kawasan rentan kebakaran, ISPU dapat difungsikan sebagai sistem peringatan dini bagi masyarakat dikawasan tersebut. Tujuan pembentukan ISPU adalah menghasilkan informasi terstandarisasi terkait kondisi udara di berbagai lokasi dan waktu. Indeks ini juga berperan sebagai bahan evaluasi dan pertimbangan kebijakan penanganan pencemaran udara bagi pemerintah di tingkat nasional dan regional.

Kementerian Lingkungan Hidup dan Kehutanan menerbitkan regulasi baru Nomor 14 Tahun 2020 tentang ISPU, menggantikan ketentuan sebelumnya dari tahun 1997. Perhitungan ISPU kini mencakup tujuh parameter: PM10, PM2.5, NO2, SO2, CO, O3, dan HC. Penambahan HC dan PM2.5 dilakukan setelah menganalisis potensi dampaknya terhadap kesehatan manusia [12].

Kategori ISPU sendiri terbagi menjadi lima kategori, dengan awal berstatus baik hingga akhir yaitu berbahaya dengan rentang masing-masing. Rentang sendiri diukur dengan melihat nilai konsentrasi polutan tertinggi yang ada pada data. Kategori ISPU dapat diamati dalam Tabel 1 berikut.

Tabel 1. Kategori Indeks Standar Pencemar Udara dan Nilai Rentang

Kategori	Rentang
Baik	1 – 50
Sedang	51 – 100
Tidak Sehat	101 – 200
Sangat Tidak Sehat	201 – 300
Berbahaya	≥ 310

2.3. Decision Tree

Decision Tree merupakan algoritma machine learning untuk pemodelan prediksi yang menggunakan struktur pohon biner. Setiap node dalam algoritma ini mewakili variabel input (x), cabang menggambarkan nilai variabelnya, dan simpul (leaf) menunjukkan variabel output (y) atau klasifikasi [13]. Node teratas disebut root, dengan nama 'pohon keputusan' karena pola aturannya menyerupai struktur pohon [14].

2.4. K-Nearest Neighbors (KNN)

Algoritma K-Nearest Neighbor merupakan pendekatan klasifikasi yang mengelompokkan data baru berdasarkan kedekatan dengan data latih terdekatnya. Jumlah tetangga terdekat (K) menjadi penentu, dengan klasifikasi akhir diputuskan melalui mayoritas jarak dan karakteristik tetangga terdekat [15].

2.5. Random Forest

Random Forest termasuk algoritma machine learning yang sangat populer dan powerful, merupakan pengembangan lanjutan dari pohon keputusan atau Decision Tree. Dinamakan 'hutan acak' karena menggunakan sejumlah pohon keputusan yang dibentuk secara acak. Setiap node dalam struktur pohon beroperasi pada subset fitur random, tidak menggunakan pendekatan greedy. Algoritma ini mengombinasikan output dari masing-masing pohon keputusan untuk menghasilkan prediksi akhir [14].

2.6. Support Vector Machine

Support Vector Machine (SVM) adalah algoritma machine learning dalam supervised learning yang digunakan untuk klasifikasi dan regresi, mampu menangani data linear dan non-linear. Metode ini mengklasifikasikan data dengan menciptakan batas keputusan yang mengoptimalkan jarak dari titik data terdekat di setiap kelas, yang dikenal sebagai Maximum Margin Classifier. Batas keputusan yang dihasilkan dinamakan maximum margin hyperplane [14].

2.7. Naive Bayes

Algoritma Naive Bayes adalah metode klasifikasi probabilistik yang menganggap setiap fitur bersifat independen satu sama lain dalam suatu kelas. Berbasis Teorema Bayes, algoritma ini memungkinkan penyesuaian probabilitas secara subjektif saat menerima fakta baru. Terdapat tiga jenis Naive Bayes: Bernoulli, Gaussian, dan Multinomial, yang masing-masing dirancang untuk konteks data yang lebih spesifik [15].

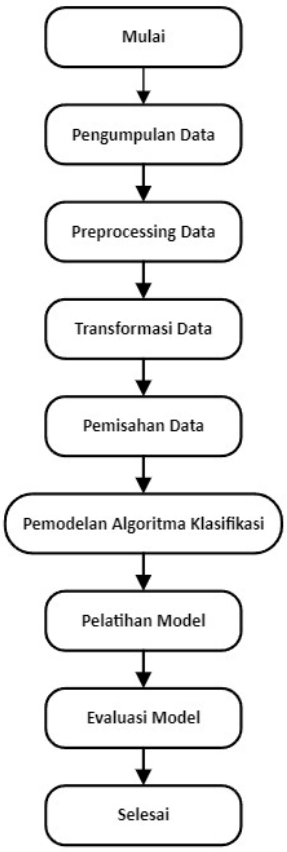
2.8. Neural Network

Neural networks atau Jaringan Syaraf Tiruan adalah sebuah metode dalam kecerdasan buatan yang bertujuan untuk mengajarkan komputer memproses data dengan cara kerja yang terinspirasi dari otak manusia. Neural network merupakan salah satu jenis proses machine learning, yang juga disebut deep learning, yang menggunakan node atau neuron yang saling berhubungan dalam struktur berlapis yang menyerupai struktur otak manusia. Neural network menciptakan sistem adaptif yang digunakan oleh komputer untuk belajar dari kesalahannya dan memperbaikinya secara terus-menerus. Oleh karena itu, neural network buatan berusaha untuk memecahkan masalah yang rumit, seperti meringkas dokumen atau mengenali wajah, dengan akurasi yang lebih tinggi [16]. Neural Network memiliki struktur dasar yang terdiri dari tiga lapisan utama: input layer yang berperan menerima data masukan, output layer yang menghasilkan prediksi akhir, serta hidden layer di antaranya yang melaksanakan proses komputasi kompleks dalam jaringan [14].

3. METODE PENELITIAN

3.1. Diagram Alir

Diagram alir metodologi penelitian ini menjelaskan tahapan penelitian yang terorganisir secara logis untuk menganalisis dan mengkategorikan kondisi kualitas udara di wilayah DKI Jakarta.



Gambar 1. Diagram Alir

3.2. Pendekatan Penelitian

Penelitian ini mengadopsi pendekatan kuantitatif dan memanfaatkan teknik data mining guna melakukan klasifikasi komprehensif terhadap kualitas udara di DKI Jakarta. Data mining dipilih karena mampu mengolah data dalam jumlah besar dan kompleks untuk menemukan pola serta hubungan yang relevan. Fokus utama penelitian ini adalah penentuan dari berbagai algoritma klasifikasi untuk mengklasifikasi kualitas udara berdasarkan data polutan yang ada.

3.3. Data dan Sumber Data

Penelitian menggunakan dataset kualitas udara yang diperoleh dari Stasiun Pengendalian Kualitas Udara (SPKU) di wilayah DKI Jakarta, dengan mengakses sumber resmi di website satudata.jakarta.go.id. Dataset komprehensif ini mencakup parameter polutan utama seperti PM2.5, SO2, NO2, CO, dan O3, yang dikumpulkan secara sistematis oleh Dinas Lingkungan Hidup Provinsi DKI Jakarta.

3.4. Pengolahan Data

Data yang dikumpulkan akan melalui beberapa tahap pengolahan sebelum dianalisis. Tahap pertama adalah pengolahan data yang mencakup pembersihan data dari anomali dan data hilang, serta normalisasi data untuk memastikan konsistensi. Selanjutnya, data akan ditransformasikan menjadi format yang dapat digunakan oleh algoritma klasifikasi. Terakhir, dataset akan dipisahkan menjadi subset pelatihan dan pengujian dengan komposisi 80% untuk pelatihan dan 20% untuk pengujian, guna menjamin validitas dan performa model klasifikasi.

3.5. Metode dan Algoritma Klasifikasi

Penelitian ini menggunakan enam algoritma klasifikasi untuk mengklasifikasi kualitas udara: Decision Tree, Naive Bayes, Random Forest, Neural Network, K-Nearest Neighbors (KNN), dan Support Vector Machine (SVM).

3.6. Evaluasi Model

Evaluasi model klasifikasi akan dilakukan terhadap set pengujian melalui analisis komprehensif yang meliputi akurasi, presisi, recall, F1-score, dan confusion matrix, guna mengukur kemampuan prediktif model secara menyeluruh.

3.7. Prosedur Penelitian

Prosedur penelitian dimulai dengan pengumpulan data kualitas udara dari sumber-sumber yang telah disebutkan. Kemudian, data tersebut akan melalui proses pengolahan dan analisis, termasuk pengolahan data dan transformasi data, serta implementasi algoritma klasifikasi. Selanjutnya melatih model menggunakan dataset pelatihan, kemudian melakukan validasi performa melalui evaluasi pada set pengujian. Hasil klasifikasi akan ditafsirkan dan efektivitas masing-masing algoritma akan dibandingkan. Akhirnya, laporan hasil penelitian akan disusun, termasuk rekomendasi berdasarkan temuan yang didapat.

3.8. Alat dan Perangkat Lunak

Penelitian ini akan menggunakan perangkat lunak serta bahasa pemrograman seperti Python dengan library pendukung seperti scikit-learn, TensorFlow, dan pandas untuk analisis data dan implementasi algoritma klasifikasi.

Alat lain yang digunakan meliputi Jupyter Notebook untuk pengembangan dan visualisasi model. Dengan pendekatan dan metode ini, Diharapkan penelitian ini dapat memberikan solusi inovatif dalam mengatasi permasalahan polusi udara di DKI Jakarta melalui pengembangan model klasifikasi kualitas udara yang presisi, yang nantinya dapat menjadi instrumen strategis bagi masyarakat dan pengambil kebijakan dalam merancang intervensi lingkungan.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Penelitian menggunakan dataset resmi yang diperoleh dari Dinas Lingkungan Hidup Provinsi DKI Jakarta, terdiri dari 1.804 entri data dengan struktur 11 variabel prediktor dan satu variabel target kategorisasi kualitas udara. Kumpulan atribut yang ada dalam dataset ini antara lain sebagai berikut: PM10, PM2.5, SO2, CO, O3, NO2, periode_data, tanggal, stasiun, max, parameter_pencemar_kritis, dan kategori.

4.2. Preprocessing Data

Sebelum masuk ke tahap pelatihan model, data harus dipastikan layak untuk dilatih. Maka hal pertama yang peneliti lakukan ialah pengecekan *missing data*. Sebelumnya, peneliti sudah melakukan pembersihan data manual pada file CSV melalui Excel, dengan mengubah *missing data* yang tadinya berbentuk '-' diubah menjadi *Null*. Jumlah data yang hilang terlihat di Gambar 2.

```
dataset.isnull().sum()
```

```
periode_data                0
tanggal                     0
stasiun                     0
PM10                      201
PM2.5                     274
SO2                        18
CO                         23
O3                          8
NO2                        45
max                         0
parameter_pencemar_kritis   0
kategori                    0
dtype: int64
```

Gambar 2. Hitung banyaknya Missing Data

Untuk menangani ketidaklengkapan data pada parameter PM10, PM2.5, SO2, CO, O3, dan NO2, penelitian akan menggunakan metode linear interpolation guna mengisi data yang hilang dengan estimasi berbasis trend linear.

```
dataset['PM10'] = dataset['PM10'].interpolate(method='linear')
dataset['PM2.5'] = dataset['PM2.5'].interpolate(method='linear')
dataset['SO2'] = dataset['SO2'].interpolate(method='linear')
dataset['CO'] = dataset['CO'].interpolate(method='linear')
dataset['O3'] = dataset['O3'].interpolate(method='linear')
dataset['NO2'] = dataset['NO2'].interpolate(method='linear')
```

Gambar 3. Isi Missing Data dengan Linear Interpolation

Selanjutnya peneliti mengubah nama dari beberapa kolom seperti "parameter_pencemar_kritis" menjadi "critical" dan "kategori" menjadi "category". Setelah data sudah bersih dan tidak ada *missing data*, selanjutnya peneliti memisahkan data untuk data latih. Dalam proses persiapan data latih, peneliti melakukan seleksi fitur dengan mengeliminasi variabel periode_data, tanggal, stasiun, dan critical, dengan fokus pada parameter polutan utama sebagai prediktor klasifikasi kualitas udara. Lalu pada data latih, jumlah nilai "SANGAT TIDAK SEHAT" pada fitur kategori hanya memiliki 3 saja. Maka dari itu kategori untuk "SANGAT TIDAK SEHAT" akan dihapus karena jumlah data yang kecil.

```
Frekuensi kategori:
 category
SEDANG                1358
BAIK                   236
TIDAK SEHAT            207
SANGAT TIDAK SEHAT       3
```

Gambar 4. Jumlah Data pada Fitur Kategori

Setelahnya kategori akan diubah tipe datanya menjadi numerik menggunakan *labelEncoder*, dan sekarang data sudah layak untuk masuk ke tahap transformasi data.

4.3. Pemisahan dan Transformasi Data

Tahapan selanjutnya melakukan splitting dataset, memisahkan fitur-fitur input dari label klasifikasi untuk mempersiapkan struktur data yang siap digunakan dalam pemodelan machine learning. Data fitur akan berisi 'PM10', 'PM2.5', 'SO2', 'CO', 'O3', 'NO2', dan data label akan berisi "category".

Setelah data dipisah menjadi data fitur(x) dan label(y), langkah selanjutnya ialah melakukan transformasi atau normalisasi pada data fitur(x). Pada tahap ini, peneliti

menggunakan metode normalisasi *StandardScaler*.

```
# Pertama kita pisahkan data label dengan data train
y = df_train_filtered["category"]
x = df_train_filtered[['PM10','PM2.5','SO2','CO','O3','NO2']]

# Normalisasi data fitur(x)
STscaler = StandardScaler()
x_scaled = STscaler.fit_transform(x)
x_scaled = STscaler.transform(x)
print(x_scaled)
```

Gambar 5. Pemisahan Data dan Normalisasi Data

Selanjutnya proses pembagian dataset dilakukan dengan menggunakan metode *train_test_split*, mengalokasikan 80% data untuk keperluan pelatihan model dan 20% sisanya sebagai dataset evaluasi/testing. Pengaturan nilai *seed*-nya adalah 42 untuk generator angka acak.

```
# Untuk Validasi model menggunakan train_test_split dengan rasio data split 80:20
X_train, X_test, y_train, y_test = train_test_split(x_normalized,
                                                    y, test_size = 0.2,
                                                    random_state = 42)
print('Classes and number of values in trainset',Counter(y_train))
```

```
Classes and number of values in trainset Counter({1: 1083, 0: 182, 2: 175})
```

Gambar 6. Membagi Data dengan train_test_split

Penggunaan *random_state* dengan nilai tertentu memungkinkan data untuk mendapatkan pembagian yang sama setiap kali menjalankan kode tersebut. Hal ini penting untuk menjaga konsistensi dalam penelitian dan evaluasi model. Jika *random_state* tidak ditentukan, maka pembagian data akan bergantung pada keadaan acak pada saat kode dijalankan, dan hasilnya mungkin berbeda setiap kali kode dijalankan.

4.4. Pemodelan Algoritma Klasifikasi

Pada tahap ini, peneliti membuat pemodelan algoritma klasifikasi yang nantinya akan dibandingkan kinerjanya. Parameter untuk setiap model tidak ditambahkan atau dengan kata lain *default*. Karena fokus penelitian ini adalah pada masing-masing model, maka untuk mendapatkan hasil perbandingan yang adil, semua parameter model akan di-*default*.

```
# Pemodelan
svm = SVC() # SVM
dt = DecisionTreeClassifier() # Decision Tree
knn = KNeighborsClassifier() # KNN
rf = RandomForestClassifier() # Random Forest
nb = GaussianNB() # Naive Bayes
nn = MLPClassifier() # Neural Network Classifier
```

Gambar 7. Pemodelan Algoritma Klasifikasi

4.5. Pelatihan Model

Setelah dilakukan pemodelan maka tahap selanjutnya adalah pelatihan model pada data latih yang sudah dibagi sebelumnya.

```
# Fit Model / Train model
svm.fit(X_train,y_train)
dt.fit(X_train, y_train)
knn.fit(X_train, y_train)
rf.fit(X_train, y_train)
nb.fit(X_train, y_train)
nn.fit(X_train, y_train)
```

Gambar 8. Fit/Pelatihan tiap Model

Setelah model dilatih selanjutnya peneliti akan melakukan prediksi pada model yang telah dilatih dengan data uji untuk melihat hasil pelatihan modelnya.

```
# Lakukan Predict test data ke model yang telah dilatih
y_pred_svm = svm.predict(X_test)
y_pred_dt = dt.predict(X_test)
y_pred_knn = knn.predict(X_test)
y_pred_rf = rf.predict(X_test)
y_pred_nb = nb.predict(X_test)
y_pred_nn = nn.predict(X_test)
```

Gambar 9. Prediksi Data Uji ke Model

4.6. Evaluasi Model

Proses evaluasi model dilaksanakan menggunakan classification report dengan metode weighted average untuk menghasilkan metrik komprehensif pada klasifikasi multiclass, mencakup accuracy, precision, recall, dan F1-score. Evaluasi model dilakukan dengan cara membandingkan tiap model algoritma klasifikasi untuk mengetahui perbedaan performa dari setiap model. Karena label memiliki *multiclass*, maka setiap nilai akurasi, presisi, recall, dan *F1-score* akan dipakai *method weighted average* untuk mendapatkan nilai rata rata tiap hasil *multiclass*. Lalu nilai yang didapat dari tiap model akan dijabarkan dalam bentuk tabel.

Tabel 2. Evaluasi Model

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	95.2%	95.3%	95.1%	95.3%
Decision Tree	94.6%	94.7%	94.6%	94.7%
Support Vector Machine	92.5%	92.5%	92.0%	92.5%
Neural Network	92.3%	92.5%	92.2%	92.5%
K-Nearest Neighbor	92.1%	92.2%	92.1%	92.2%
Naïve Bayes	87.0%	86.1%	86.5%	86.1%

Berdasarkan hasil evaluasi pada tabel 2 di atas, model algoritma dari *Random Forest* ke *KNN* memiliki performa klasifikasi yang sangat baik dengan nilai akurasi 90% ke atas sedangkan untuk *Naive Bayes* sendiri dapat dikatakan cukup baik karena masih berada pada akurasi 80% ke atas. Model algoritma klasifikasi *Random Forest* menjadi model algoritma dengan performa terbaik dalam penelitian ini, baik dari performa nilai akurasi, presisi, *recall* dan *F1-score*.

5. KESIMPULAN

a. Pemodelan yang telah dilakukan dengan algoritma klasifikasi untuk mengidentifikasi kualitas udara di wilayah Jakarta periode 2023 menghasilkan temuan bahwa, algoritma Random Forest merupakan algoritma terbaik di antara algoritma Neural Network, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Naive Bayes. Random Forest memiliki kinerja paling unggul dengan akurasi 95%, diikuti oleh algoritma Decision Tree dengan akurasi 94,7%. Algoritma Naive Bayes memiliki performa terendah dengan akurasi 86,1%.

b. Hasil uji penelitian menunjukkan bahwa, model yang digunakan masih menggunakan parameter default, memungkinkan akurasi dapat berubah jika dilakukan penyesuaian parameter. Model Random Forest dapat digunakan sebagai alat efektif untuk memantau dan mengklasifikasikan kualitas udara di wilayah Jakarta.

c. Penelitian ini diharapkan dapat Memberikan pemahaman dalam pengembangan algoritma klasifikasi, dan membantu masyarakat dan pemerintah dalam upaya penanggulangan polusi udara.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada bapak Achmad Bahauddin, ST., MT., PH.D selaku dosen mata kuliah data mining, yang telah memberi wawasan, dan dukungannya dalam membimbing penulis dalam melakukan dan meyelesaikan penelitian ini.

DAFTAR PUSTAKA

[1] E. Fikri, *PENCEMARAN UDARA DAN DAMPAKNYA BAGI KESEHATAN*. CV.EUREKA MEDIA AKSARA, 2021.

[2] Prudential Indonesia, "Dampak Polusi Udara Bagi Kesehatan." Accessed: May 24, 2024. [Online]. Available: https://www.prudential.co.id/id/pulse/article/tinggal-di-perkotaan-ketahui-dulu-dampak-pencemaran-udara-ini/#:~:text=Dampak%20pencemaran%20udara%20dari%20asap,tubuh%20manusia%2C%20sehingga%20membahayakan%20kesehatan.

[3] S. Basri K. *et al.*, *TEORI KESEHATAN LINGKUNGAN*. Yayasan Penerbit Muhammad Zaini, 2023.

[4] Badan Pemeriksa Keuangan Jakarta, "Sekilas Tentang Provinsi DKI Jakarta." Accessed: May 24, 2024. [Online]. Available: https://jakarta.bpk.go.id/pemerintah-provinsi-dki-jakarta/

[5] IBLAM, "Polusi Udara Jakarta, Penyebabnya dan Bahaya Kesehatan." Accessed: May 24, 2024. [Online]. Available: https://iblam.ac.id/2023/11/13/polusi-udara-jakarta-penyebabnya-dan-bahaya-kesehatan/#:~:text=Selain%20emisi%20kendaraan%2C%20penggunaan%20pembangkit,menghasilkan%20Sulfur%20Dioksida%20(SO2).

[6] S. Nurhaliza, "Awal 2024, kualitas udara Jakarta tak sehat." Accessed: May 24, 2024. [Online]. Available:

https://www.antaranews.com/berita/3894927/awal-2024-kualitas-udara-jakarta-tak-sehat

[7] A. Amalia, A. Zaidiah, and I. N. Isnainiyah, "PREDIKSI KUALITAS UDARA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR," *JIPI: Jurnal Ilmiah Penelitian dan Pembelajaran Informatika*, vol. 7, no. 2, Jun. 2022, doi: https://doi.org/10.29100/jipi.v7i2.2843.

[8] M. A. Muslim *et al.*, *Data Mining Algoritma C4.5*. ILKOM UNNES, 2019.

[9] A. Fransiska, "METODE DATA MINING CLASSIFICATION," Binus University: School of Information Systems. Accessed: May 24, 2024. [Online]. Available: https://sis.binus.ac.id/2021/10/22/metode-data-mining-classification/

[10] A. Nugroho, ibnu Asror, and Y. F. A. Wibowo, "Klasifikasi Tingkat Kualitas Udara DKI Jakarta Berdasarkan Open Government Data Menggunakan Algoritma Random Forest," *eProceedings of Engineering*, vol. 10, no. 2, 2023.

[11] Populix, "Data Mining: Pengertian, Tahapan, Contoh, dan Manfaat." Accessed: May 24, 2024. [Online]. Available: https://info.populix.co/articles/data-mining-adalah/#:~:text=Klasifikasi%20data%20mining%20adalah%20sebuah,objek%20yang%20labelnya%20belum%20diketahui.

[12] Dinas Lingkungan Hidup DKI Jakarta, "Tentang ISPU." Accessed: May 26, 2024. [Online]. Available: https://e-llhd.dinaslhdki.id/tentang#:~:text=Login-,Tentang%20ISPU,estetika%20dan%20makhluk%20hidup%20lainnya.

[13] S. S. Elfaretta, A. A. Arifiyanti, and A. S. Fitri, "KLASIFIKASI CALON PENDONOR DARAH POTENSIAL MENGGUNAKAN ALGORITMA DECISION TREE DI UTD PMI KOTA SURABAYA," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4957.

[14] Kristiawan and A. Widjaja, "Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 1, Apr. 2021, doi: https://doi.org/10.28932/jutisi.v7i1.3182.

[15] G. D. Nursyafitri, "4 Contoh Algoritma Classification Data Science," DQLab. Accessed: May 24, 2024. [Online]. Available: https://dqlab.id/4-contoh-algoritma-classification-data-science

[16] Amazon Web Services, "Apa itu Jaringan Neural?," AWS Amazon. Accessed: May 24, 2024. [Online]. Available: https://aws.amazon.com/id/what-is/neural-network/