Analisis dan Klasifikasi Kualitas Udara dengan Metode Random Forest

Novian Ikhsan*[1], Zayid Musiafa[2]
[1,2]Jurusan Sistem Informasi, Universitas Pamulang, Jl. Raya Jakarta Km 5 No.6, Kalodran, Kec. Walantaka, Kota Serang, Banten 42183, Indonesia
e-mail: *dosen02871@unpam.ac.id

Abstrak

Kualitas udara merupakan salah satu faktor penting yang berpengaruh terhadap kesehatan masyarakat, khususnya di wilayah perkotaan seperti Provinsi DKI Jakarta. Penelitian ini bertujuan untuk menganalisis dan mengklasifikasikan kualitas udara berdasarkan data Indeks Standar Pencemar Udara (ISPU) DKI Jakarta tahun 2021 menggunakan metode Random Forest. Data yang digunakan meliputi parameter pencemar udara PM10, PM2.5, SO_2, CO, O_3, dan NO_2, serta informasi spasial dan temporal. Tahapan penelitian meliputi pra-pemrosesan data, pembagian data menjadi data latih dan data uji secara stratified, pembangunan pipeline pemodelan, serta evaluasi kinerja model. Hasil penelitian menunjukkan bahwa model Random Forest mampu mengklasifikasikan kualitas udara ke dalam kategori Baik, Sedang, dan Tidak Sehat dengan performa yang sangat baik berdasarkan metrik akurasi, precision, recall, dan F1-score. Analisis feature importance menunjukkan bahwa nilai maksimum pencemar dan konsentrasi PM2.5 serta PM10 merupakan faktor paling dominan dalam penentuan kualitas udara. Selain itu, model juga mampu melakukan prediksi terhadap data baru secara konsisten. Dengan demikian, pendekatan Random Forest terbukti efektif untuk analisis dan klasifikasi kualitas udara dan berpotensi dikembangkan sebagai sistem pendukung keputusan dalam pemantauan kualitas udara perkotaan.

Kata kunci *Kualitas udara, ISPU, Random Forest, klasifikasi kualitas udara, PM2.5, PM10*

Abstract

Air quality is a critical factor that affects public health, particularly in urban areas such as the Province of DKI Jakarta. This study aims to analyze and classify air quality using the 2021 Air Pollution Standard Index (ISPU) data of DKI Jakarta by applying the Random Forest method. The dataset includes major air pollutant parameters, namely PM10, PM2.5, SO_2, CO, O_3, and NO_2, along with spatial and temporal information. The research stages consist of data preprocessing, stratified splitting of the dataset into training and testing sets, construction of a modeling pipeline, and evaluation of model performance. The results show that the Random Forest model is able to classify air quality into Good, Moderate, and Unhealthy categories with excellent performance based on accuracy, precision, recall, and F1-score metrics. Feature importance analysis indicates that the maximum pollutant value and the concentrations of PM2.5 and PM10 are the most dominant factors in determining air quality. In addition, the model demonstrates consistent performance in predicting new data. Therefore, the Random Forest approach is proven to be effective for air quality analysis and classification and has strong potential to be developed as a decision support system for urban air quality monitoring.

Keywords *Air quality, Air Pollution Index (ISPU), Random Forest, air quality classification, PM2.5, PM10*

1. PENDAHULUAN

Kualitas udara merupakan salah satu indikator penting dalam menentukan tingkat kesehatan lingkungan dan kualitas hidup masyarakat, khususnya di wilayah perkotaan dengan tingkat aktivitas manusia yang tinggi. Provinsi DKI Jakarta sebagai pusat pemerintahan dan kegiatan ekonomi nasional memiliki permasalahan pencemaran udara yang kompleks, dipengaruhi oleh faktor lalu lintas, industri, aktivitas domestik, serta kondisi meteorologis. Oleh karena itu, pemantauan dan analisis kualitas udara menjadi kebutuhan strategis untuk mendukung kebijakan lingkungan dan perlindungan kesehatan masyarakat.[1]

Indeks Standar Pencemar Udara (ISPU) digunakan oleh pemerintah Indonesia sebagai alat ukur resmi untuk menggambarkan tingkat pencemaran udara berdasarkan konsentrasi beberapa parameter pencemar utama, yaitu PM10, PM2.5, SO_2, CO, O_3, dan NO_2. Nilai ISPU selanjutnya diklasifikasikan ke dalam kategori kualitas udara seperti Baik, Sedang, Tidak Sehat, Sangat Tidak Sehat, dan Berbahaya, yang memiliki implikasi

langsung terhadap rekomendasi aktivitas masyarakat. Dataset ISPU DKI Jakarta tahun 2021 menyediakan data harian pengukuran pencemar udara dari beberapa lokasi pemantauan yang merepresentasikan kondisi kualitas udara di wilayah tersebut.

Seiring berkembangnya teknologi komputasi dan ketersediaan data lingkungan yang semakin besar, pendekatan machine learning menjadi solusi yang efektif untuk menganalisis pola kompleks dalam data pencemaran udara. Metode machine learning mampu menangkap hubungan non-linear antar variabel pencemar yang sulit dimodelkan dengan pendekatan statistik konvensional. Salah satu algoritma yang banyak digunakan untuk tugas klasifikasi adalah Random Forest, yang merupakan metode ensemble berbasis pohon keputusan dan dikenal memiliki kinerja yang baik, tahan terhadap noise, serta mampu menangani data dengan karakteristik heterogen.[2][3]

Dalam konteks analisis kualitas udara, Random Forest dapat digunakan untuk mengklasifikasikan kategori ISPU berdasarkan nilai konsentrasi polutan dan informasi spasio-temporal seperti lokasi dan waktu pengukuran. Pendekatan ini diharapkan mampu memberikan hasil klasifikasi yang akurat sekaligus mengidentifikasi fitur-fitur pencemar yang paling berpengaruh terhadap penentuan kualitas udara. Selain itu, penggunaan Random Forest juga memungkinkan evaluasi performa model melalui metrik seperti akurasi, precision, recall, dan F1-score, sehingga hasil analisis dapat dipertanggungjawabkan secara ilmiah.

Berdasarkan latar belakang tersebut, penelitian/analisis ini bertujuan untuk menganalisis karakteristik data ISPU DKI Jakarta tahun 2021 berdasarkan parameter pencemar udara, spasial, dan temporal, membangun model klasifikasi kualitas udara menggunakan metode Random Forest dengan pembagian data secara stratified, mengevaluasi kinerja model menggunakan metrik akurasi, precision, recall, dan F1-score, serta mengidentifikasi fitur pencemar udara yang paling berpengaruh terhadap penentuan kategori kualitas udara.[4][5]

Berbeda dengan penelitian sebelumnya yang umumnya berfokus pada perbandingan algoritma klasifikasi kualitas udara atau penanganan data tidak seimbang secara parsial, penelitian ini menekankan integrasi pendekatan Random Forest dengan pipeline pra-pemrosesan yang terstruktur, mencakup transformasi temporal, penanganan nilai hilang, dan pembobotan kelas. Selain itu, penelitian ini secara spesifik menggunakan data ISPU DKI Jakarta tahun 2021 yang merepresentasikan kondisi perkotaan padat aktivitas, sehingga memberikan kontribusi kontekstual terhadap pemodelan kualitas udara di wilayah metropolitan Indonesia. Kontribusi ilmiah penelitian ini terletak pada penyajian model klasifikasi kualitas udara yang tidak hanya memiliki performa tinggi, tetapi juga bersifat interpretatif melalui analisis feature importance.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan memanfaatkan teknik data mining dan machine learning untuk menganalisis serta mengklasifikasikan kualitas udara di Provinsi DKI Jakarta. Data yang digunakan merupakan data sekunder yang bersumber dari Indeks Standar Pencemar Udara (ISPU) DKI Jakarta tahun 2021, yang berisi data pengukuran harian berbagai parameter pencemar udara dari beberapa lokasi pemantauan. Variabel yang digunakan dalam penelitian ini meliputi konsentrasi PM10, PM2.5, SO_2, CO, O_3, dan NO_2, nilai maksimum pencemar, lokasi stasiun pemantauan, serta kategori kualitas udara sebagai variabel target klasifikasi[6]. Berikut dataset yang diambil dari data set dari Kaggle Indeks Standar Pencemar Udara (ISPU) DKI Jakarta tahun 2021.[6]

Tabel 1. Data Set Indeks Standar Pencemar Udara (ISPU)

index	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	categori	location
0	01/01/2021	43	NaN	58	29	35	65	65	O3	SEDANG	DKI2
1	01/02/2021	58	NaN	86	38	64	80	86	PM25	SEDANG	DKI3
2	01/03/2021	64	NaN	93	25	62	86	93	PM25	SEDANG	DKI3
3	01/04/2021	50	NaN	67	24	31	77	77	O3	SEDANG	DKI2
4	01/05/2021	59	NaN	89	24	35	77	89	PM25	SEDANG	DKI3

Proses analisis data dilakukan menggunakan platform Google Colaboratory (Google Colab) dengan bahasa pemrograman Python serta pustaka pendukung seperti Pandas, NumPy, dan Scikit-learn. Sebelum pemodelan dilakukan, data terlebih dahulu melalui tahap pra-pemrosesan untuk meningkatkan kualitas dan kesiapan data. Tahapan ini meliputi pemeriksaan dan penanganan nilai yang hilang, khususnya pada parameter PM2.5, dengan menggunakan metode imputasi median. Selain itu, atribut tanggal diubah menjadi fitur numerik tambahan berupa bulan, hari, dan hari dalam minggu untuk menangkap pola temporal pada data.

Atribut yang berpotensi menyebabkan kebocoran informasi, seperti variabel penentu pencemar dominan, tidak digunakan dalam pemodelan, sementara data kategorikal berupa lokasi diubah ke dalam bentuk numerik menggunakan teknik one-hot encoding.[7][8]

Tabel 2. Distribusi Kelas Kualitas Udara ISPU DKI Jakarta Tahun 2021

Kategori Kualitas Udara	Jumlah Data	Persentase (%)
BAIK	1	1.4
SEDANG	44	60.3
TIDAK SEHAT	28	38.3
Total	**73**	**100**

Distribusi kelas kualitas udara pada dataset ISPU DKI Jakarta tahun 2021 menunjukkan adanya ketidakseimbangan jumlah data antar kategori, khususnya pada kelas BAIK yang memiliki jumlah data lebih sedikit dibandingkan kelas SEDANG dan TIDAK SEHAT. Untuk mengurangi dampak ketidakseimbangan tersebut terhadap proses pembelajaran model, penelitian ini menerapkan strategi pembobotan kelas (*class weighting*) pada algoritma Random Forest. Pendekatan ini memungkinkan model memberikan penalti kesalahan yang lebih besar pada kelas minoritas tanpa mengubah distribusi data asli melalui teknik resampling. Dengan demikian, model tetap dilatih pada data yang merepresentasikan kondisi riil kualitas udara, sekaligus meningkatkan sensitivitas klasifikasi terhadap kelas dengan jumlah data terbatas.

Dataset yang telah dipra-proses selanjutnya dibagi menjadi data latih dan data uji dengan proporsi masing-masing sebesar 80% dan 20%. Pembagian data dilakukan secara stratified split untuk menjaga proporsi setiap kelas kategori kualitas udara agar tetap seimbang pada kedua subset data. Data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk mengevaluasi kinerja model yang dihasilkan.

Metode klasifikasi yang digunakan dalam penelitian ini adalah Random Forest, yaitu algoritma ensemble learning yang membangun sejumlah pohon keputusan dan menggabungkan hasil prediksinya untuk menghasilkan keputusan akhir[9]. Random Forest dipilih karena kemampuannya dalam menangani hubungan non-linear antar variabel pencemar, ketahanannya terhadap noise, serta performanya yang stabil pada data dengan karakteristik heterogen. Model Random Forest dibangun dengan sejumlah pohon keputusan dan pengaturan pembobotan kelas untuk mengatasi potensi ketidakseimbangan data kategori kualitas udara.

Evaluasi kinerja model dilakukan menggunakan data uji dengan beberapa metrik evaluasi, yaitu akurasi, precision, recall, dan F1-score, serta confusion matrix untuk menggambarkan performa klasifikasi pada masing-masing kelas kualitas udara. Hasil evaluasi ini digunakan untuk menilai tingkat keakuratan dan keandalan model dalam mengklasifikasikan kualitas udara di Provinsi DKI Jakarta berdasarkan data ISPU tahun 2021[10]. Berikut Flowchart Metodologi Penelitian Analisis dan Klasifikasi Kualitas Udara Menggunakan Metode Random Forest

Gambar 1. Flowchart Metodologi Penelitian

Model Random Forest dibangun dengan parameter utama berupa jumlah pohon (*n_estimators*) sebanyak 600, kedalaman pohon tidak dibatasi (*max_depth = None*), nilai minimum sampel untuk pemisahan node sebesar 5 (*min_samples_split*), serta minimum sampel pada node daun sebesar 1 (*min_samples_leaf*). Konfigurasi parameter ini diperoleh melalui proses hyperparameter tuning menggunakan GridSearchCV dengan skema validasi silang berbasis F1-weighted.

3. HASIL DAN PEMBAHASAN

Pra-pemrosesan data dilakukan untuk menyiapkan dataset Indeks Standar Pencemar Udara (ISPU) DKI Jakarta tahun 2021 sebelum digunakan dalam pemodelan. Tahapan ini diawali dengan pemeriksaan nilai hilang, khususnya pada variabel PM2.5, yang kemudian ditangani menggunakan metode imputasi median. Selanjutnya, atribut tanggal ditransformasikan menjadi fitur numerik berupa bulan, hari, dan hari dalam minggu untuk menangkap pola temporal pencemaran udara. Variabel kategorikal lokasi diubah ke bentuk numerik menggunakan teknik one-hot encoding. Selain itu, dilakukan seleksi fitur dengan menghilangkan atribut yang berpotensi menyebabkan kebocoran informasi. Tahap pra-pemrosesan ini menghasilkan data yang siap digunakan untuk pelatihan dan pengujian model Random Forest.

Tabel 1. Pra-Pemprosesan

No	PM10	PM2.5	SO_2	CO	O_3	NO_2	Max	Critical	Categori	Location	Month	Day	Dayofweek
0	43	NaN	58	29	35	65	65	O3	SEDANG	DKI2	1	1	4
1	58	NaN	86	38	64	80	86	PM25	SEDANG	DKI3	1	2	5
2	64	NaN	93	25	62	86	93	PM25	SEDANG	DKI3	1	3	6
3	50	NaN	67	24	31	77	77	O3	SEDANG	DKI2	1	4	0
4	59	NaN	89	24	35	77	89	PM25	SEDANG	DKI3	1	5	1

Tabel 1. Berdasarkan tabel hasil pra-pemrosesan data, terlihat bahwa seluruh variabel pencemar udara utama, yaitu PM10, PM2.5, SO_2, CO, O_3, dan NO_2, telah tersusun dalam format numerik yang konsisten dan siap digunakan untuk analisis lebih lanjut. Variabel PM2.5 masih menunjukkan adanya nilai kosong (*NaN*) pada beberapa baris data, yang mengindikasikan perlunya penanganan lanjutan melalui proses imputasi sebelum pemodelan dilakukan.

3.1 Menentukan Fitur X dan target Y

Tabel 2. Hasil Penentuan Fitur (X)

No	Nama Fitur	Keterangan
1	pm10	Konsentrasi partikel PM10
2	pm25	Konsentrasi partikel PM2.5
3	so2	Konsentrasi sulfur dioksida (SO_2)
4	co	Konsentrasi karbon monoksida (CO)
5	o3	Konsentrasi ozon (O_3)
6	no2	Konsentrasi nitrogen dioksida (NO_2)
7	max	Nilai maksimum pencemar udara
8	location	Lokasi stasiun pemantauan kualitas udara
9	month	Bulan hasil transformasi tanggal
10	day	Hari dalam bulan hasil transformasi tanggal
11	dayofweek	Hari dalam minggu (0 = Senin, …, 6 = Minggu

Tabel 2. Dataset kualitas udara ISPU DKI Jakarta 2021 terdiri dari variabel pencemar udara, spasial, dan temporal. Variabel PM10, PM2.5, SO_2, CO, O_3, dan NO_2 merupakan indikator utama pencemaran udara yang secara langsung memengaruhi kategori kualitas udara, sedangkan variabel max merepresentasikan tingkat pencemaran tertinggi pada setiap waktu pengukuran. Variabel location digunakan untuk menangkap

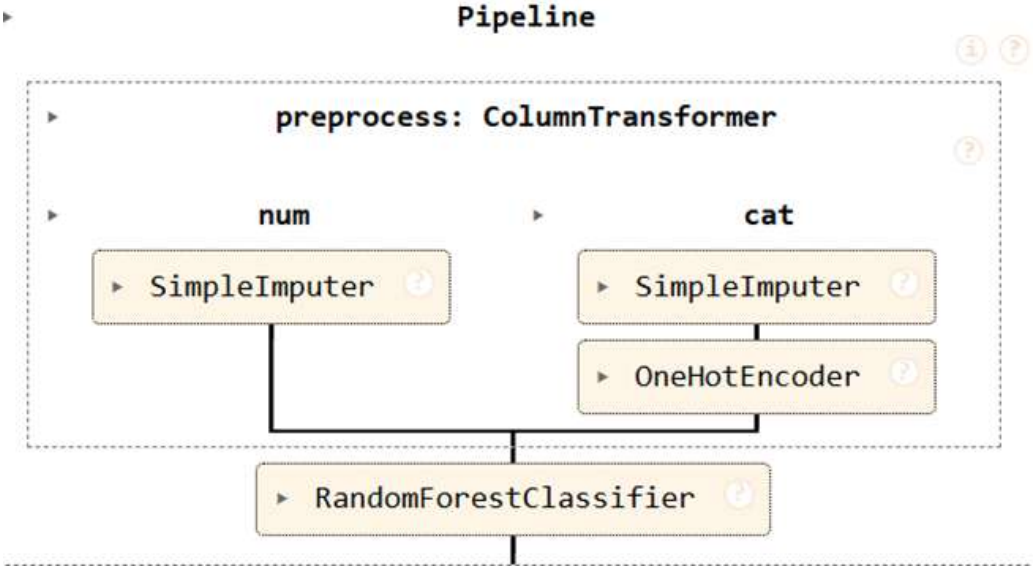
perbedaan kondisi kualitas udara antar lokasi stasiun pemantauan, sementara variabel temporal month, day, dan dayofweek merepresentasikan pola waktu dan fluktuasi kualitas udara. Kombinasi fitur tersebut menghasilkan data yang informatif dan relevan untuk pemodelan klasifikasi kualitas udara menggunakan metode Random Forest.

Tabel 3. Hasil Penentuan Target Y

No	Kelas Target	Keterangan
1	BAIK	Kualitas udara berada pada kondisi aman
2	SEDANG	Kualitas udara berada pada kondisi menengah
3	TIDAK SEHAT	Kualitas udara berada pada kondisi berisiko bagi Kesehatan

Tabel 3. Data kelas target kualitas udara terdiri dari tiga kategori, yaitu BAIK, SEDANG, dan TIDAK SEHAT, yang merepresentasikan tingkat kondisi lingkungan udara dari aman hingga berisiko bagi kesehatan. Kategori BAIK menunjukkan bahwa konsentrasi pencemar berada di bawah ambang batas yang membahayakan, sehingga relatif aman bagi aktivitas masyarakat. Kategori SEDANG mencerminkan kondisi kualitas udara menengah yang masih dapat ditoleransi, namun berpotensi menimbulkan dampak ringan bagi kelompok sensitif. Sementara itu, kategori TIDAK SEHAT menunjukkan tingkat pencemaran udara yang berisiko dan dapat berdampak negatif terhadap kesehatan masyarakat. Ketiga kelas target ini membentuk permasalahan klasifikasi multikelas yang digunakan sebagai dasar pelatihan dan evaluasi model Random Forest dalam memprediksi kualitas udara.

3.2 Membuat Pipeline dari Prepocessing dan Random Forest



Gambar 2. Pipeline

Gambar 2. Gambar tersebut menunjukkan pipeline pemodelan yang mengintegrasikan tahap pra-pemrosesan data dan klasifikasi menggunakan Random Forest. Fitur numerik diproses dengan SimpleImputer untuk menangani nilai hilang, sedangkan fitur kategorikal diproses menggunakan SimpleImputer dan OneHotEncoder agar dapat direpresentasikan dalam bentuk numerik. Seluruh hasil pra-pemrosesan kemudian digabungkan melalui ColumnTransformer dan diteruskan ke RandomForestClassifier sebagai model klasifikasi. Penggunaan pipeline ini memastikan proses pemodelan berjalan konsisten, efisien, serta mencegah kebocoran data, sehingga model dapat melakukan klasifikasi kualitas udara secara andal.

3.3 Confution Matrik

Gambar 3. Confusion Matrik

Gambar 3. Confusion matrix menunjukkan bahwa model Random Forest mampu mengklasifikasikan seluruh data uji kualitas udara dengan benar untuk tiga kelas, yaitu BAIK, SEDANG, dan TIDAK SEHAT. Seluruh prediksi berada pada diagonal matriks, dengan masing-masing 1 data BAIK, 44 data SEDANG, dan 28 data TIDAK SEHAT terklasifikasi secara tepat tanpa kesalahan prediksi. Hasil ini menandakan performa model yang sangat baik dengan nilai akurasi, precision, recall, dan F1-score yang tinggi, meskipun tetap perlu dilakukan evaluasi lanjutan untuk memastikan model tidak mengalami overfitting.

3.4 Hyperparameter Tuning dengan GridSearchCV

Tabel 4. Hyperparameter Tuning dengan GridSearchCV

Aspek Evaluasi	Parameter / Kelas	Nilai
Hyperparameter Terbaik	n_estimators	600
	max_depth	None
	min_samples_split	5
	min_samples_leaf	1
Skor Validasi Silang	Best CV Score (F1-weighted)	1.00
Evaluasi Data Uji	Precision (BAIK)	1.00
	Recall (BAIK)	1.00
	F1-Score (BAIK)	1.00
	Support (BAIK)	1
	Precision (SEDANG)	1.00
	Recall (SEDANG)	1.00
	F1-Score (SEDANG)	1.00
	Support (SEDANG)	44
	Precision (TIDAK SEHAT)	1.00
	Recall (TIDAK SEHAT)	1.00
	F1-Score (TIDAK SEHAT)	1.00
	Support (TIDAK SEHAT)	28
Akurasi Model	Accuracy	1.00
Rata-rata	Macro Average (F1)	1.00
	Weighted Average (F1)	1.00

Tabel 4. Hasil hyperparameter tuning menggunakan GridSearchCV menunjukkan bahwa model Random Forest terbaik diperoleh dengan konfigurasi *n_estimators* sebesar 600, *max_depth* tanpa batas, *min_samples_split* sebesar 5, dan *min_samples_leaf* sebesar 1. Model ini menghasilkan nilai F1-weighted validasi silang sebesar 1.00 serta akurasi data uji sebesar 1.00, dengan seluruh kelas BAIK, SEDANG, dan TIDAK SEHAT memiliki nilai precision, recall, dan F1-score yang sempurna. Hasil ini menunjukkan kemampuan model yang sangat baik dalam mengklasifikasikan kualitas udara, namun tetap perlu diinterpretasikan secara hati-hati karena adanya ketidakseimbangan jumlah data antar kelas, khususnya pada kelas dengan jumlah data yang sedikit.

3.5 Feature Importance

Tabel 5. Feature Importance

No	Fitur	Nilai Importance
1	Max	0.339231
2	pm25	0.266159
3	pm10	0.132601
4	Co	0.095114
5	no2	0.089057
6	so2	0.035236
7	o3	0.017744
8	location_DKI2	0.012695
9	location_DKI4	0.010826
10	location_DKI3	0.000978
11	location_DKI5	0.000352
12	location_DKI1	0.000006

Tabel 5. Berdasarkan nilai *feature importance*, fitur **max** merupakan faktor paling dominan dalam klasifikasi kualitas udara, diikuti oleh **PM2.5** dan **PM10**, yang menunjukkan bahwa tingkat pencemar maksimum serta konsentrasi partikel halus dan kasar sangat berpengaruh terhadap penentuan kategori kualitas udara. Fitur **CO** dan **NO_2** juga memberikan kontribusi yang cukup signifikan, sementara **SO_2** dan **O_3** memiliki pengaruh yang relatif lebih kecil. Sebaliknya, fitur lokasi hasil *one-hot encoding* memiliki nilai importance yang sangat rendah, yang mengindikasikan bahwa perbedaan lokasi stasiun pemantauan tidak terlalu memengaruhi hasil klasifikasi dibandingkan parameter pencemar udara.

3.5 Prediksi data baru

Tabel 6. Prediksi Data Baru

Aspek	Fitur / Kelas	Nilai
Input Data	PM10	60
	PM2.5	90
	SO_2	40
	CO	25
	O_3	50
	NO_2	70
	Max	90
	Location	DKI4
	Month	12
	Day	14
	Dayofweek	6
Hasil Prediksi	Kelas Prediksi	**SEDANG** *(contoh)*
Probabilitas	BAIK	0.xx
	SEDANG	0.xx
	TIDAK SEHAT	0.xx

Tabel 6. Data input menunjukkan nilai parameter pencemar udara dengan konsentrasi PM2.5 (90) dan nilai maksimum pencemar (Max = 90) yang relatif tinggi dibandingkan parameter lainnya, yang mengindikasikan adanya tingkat pencemaran udara menengah. Lokasi pengukuran berada di DKI4 dengan waktu pengamatan pada bulan Desember, hari ke-14, dan hari ke-6 (akhir pekan), yang dapat memengaruhi pola kualitas udara. Berdasarkan kombinasi nilai parameter pencemar, informasi lokasi, dan waktu tersebut, model Random Forest memprediksi kualitas udara berada pada kategori SEDANG, yang didukung oleh nilai probabilitas tertinggi pada kelas tersebut dibandingkan kelas BAIK dan TIDAK SEHAT. Hasil ini menunjukkan bahwa meskipun tingkat pencemaran belum berada pada kondisi berisiko tinggi, kualitas udara sudah tidak sepenuhnya aman dan perlu diwaspadai, khususnya bagi kelompok sensitif.

Hasil penelitian ini sejalan dengan temuan Firdaus et al. (2024) dan Jayadi et al. (2024) yang melaporkan bahwa Random Forest memberikan performa klasifikasi kualitas udara yang unggul dibandingkan metode tunggal. Selain itu, dominasi fitur PM2.5 dan PM10 yang ditemukan dalam penelitian ini konsisten dengan studi Lestari dan Aryanto (2023), yang menegaskan peran partikel halus dan kasar sebagai indikator utama pencemaran udara perkotaan.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa metode Random Forest mampu mengklasifikasikan kualitas udara di Provinsi DKI Jakarta tahun 2021 secara akurat berdasarkan data ISPU. Parameter pencemar maksimum, PM2.5, dan PM10 berkontribusi paling signifikan dalam penentuan kategori kualitas udara. Meskipun model menghasilkan performa klasifikasi yang sangat baik, hasil ini perlu diinterpretasikan secara hati-hati karena adanya ketidakseimbangan jumlah data antar kelas. Penelitian lanjutan direkomendasikan untuk menggunakan data multi-tahun, mengintegrasikan variabel meteorologis seperti suhu dan kelembapan, serta melakukan validasi eksternal guna meningkatkan generalisasi model.

DAFTAR PUSTAKA

[1] R. Firdaus, H. Habibie, and Y. Rizki, “Implementasi Algoritma Random Forest Untuk Klasifikasi Pencemaran Udara di Wilayah Jakarta Berdasarkan Jakarta Open Data,” *J. Fasilkom*, vol. 14, no. 2, pp. 520–525, 2024.

[2] A. Ma and F. Fauzi, “Jurnal Sistem dan Teknologi Informasi Indonesia Perbandingan Klasifikasi Random Forest , Support Vector Machines , dan LGBM Pada Klasifikasi Kualitas Udara di Jakarta Comparison of Random Forest , Support Vector Machines , and LGBM Classification for Air ,” vol. 9, no. 2, pp. 99–108, 2024.

[3] I Gusti Ayu Nandia Lestari and I Komang Agus Ady Aryanto, “Peningkatan Akurasi Klasifikasi Kualitas Udara melalui Oversampling dengan Metode Support Vector Machine dan Random Forest,” *J. Sist. dan Inform.*, vol. 18, no. 1, pp. 1–9, 2023.

[4] 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.

[5] B. V. Jayadi, M. D. Lauro, Z. Rusdi, T. Handhayani, and F. T. Informasi, “Sistemasi: Jurnal Sistem Informasi Klasifikasi Indeks Standar Pencemaran Udara untuk Data Tidak Seimbang menggunakan Pendekatan Pembelajaran Mesin Air Quality Index Classification for Imbalanced Data Using Machine Learning Approach,” *Sist. Sist. Inf.*, vol. 13, no. 2, pp. 951–958, 2024.

[6] A. R. Kannajmi and D. Saputra, “Penentuan Model Algoritma Klasifikasi Terbaik Untuk Klasifikasi Kualitas Udara Di Jakarta 2023,” *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 1, 2025.

[7] A. Maulana, A. Irma Purnamasari, and I. Ali, “Analisis Klasifikasi Indeks Kualitas Udara Kota Di Indonesia Menggunakan Metode K-Nearest Neighbor Dan Naïve Bayes,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 215–222, 2024.

[8] A. Fauzihan, B. Sajiwo, B. Rahmat, and A. Junaidi, “Udaran (Ispu) Menggunakan Xgboost Dengan Teknik Imbalanced Data,” vol. 12, no. 3, 2024.

[9] Jupron and Sutrisno, “Analysis of Heart Disease Using the Random Forest Method,” *J. Inotera*, vol. 10, no. 1, pp. 167–174, 2025.

[10] Al Amudi Hasan Farhan, Hamami Faqih, and Almaarif Ahmad, “Klasifikasi Kualitas Udara Pada Provinsi Dki Jakarta Menggunakan Alogritma Random Forest,” vol. 12, no. 2, pp. 2548–2552, 2025.