UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 4 Nomor 2 Tahun 2026 Halaman 141-149.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Random Forest Algorithm Implementation for Air Quality Classification in DKI Jakarta Based on ISPU

Khairanisa Salsabila, Tessy Octavia Mukhti*
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: tessyoctaviam@fmipa.unp.ac.id

Submitted : 28 Februari 2026
Revised : 23 April 2026
Accepted : 18 Mei 2026

ABSTRACT

Air quality is an essential factor that has a direct impact on human health. High concentrations of air pollutants have the potential to cause various health impacts, across short-term and long-term horizons. This study aims to classify air quality in DKI Jakarta using the Air Pollution Standard Index (ISPU) data via the random forest algorithm. The dataset covers a timeframe from 2021 to 2025 and includes air pollutant parameters, namely PM10 and PM2.5 particulate matter, carbon monoxide (CO), nitrogen dioxide (NO2), sulfur dioxide (SO2), and ozone (O3). The research method employs a supervised learning approach, in which the data are stratified and evaluated through the implementation of K-Fold Cross Validation (k = 10) to ensure objective and stable model performance. Model performance was measured using Accuracy, Precision, Recall, and F1-Score metrics, along with Confusion Matrix and Feature Importance analyses. It can be seen from the results that the Random Forest model can classify air quality categories with excellent performance, reaching 100% Accuracy on training data and 98.44% on testing data. The Confusion Matrix analysis indicates that most data in each air quality are correctly classified. Furthermore, Feature Importance analysis reveals that PM2.5 is the most influential parameter in determining air quality categories. Therefore, this study indicates that the Random Forest algorithm proves effective for air quality classification and can function as a decision-support tool for air pollution control and management in DKI Jakarta.

Keywords: *Air Quality, ISPU, Random Forest, DKI Jakarta*

I. PENDAHULUAN

Kualitas udara merupakan faktor penting yang berdampak langsung pada kesehatan manusia dan perlu mendapatkan perhatian dalam jangka panjang. Berdasarkan *World Air Quality Report* 2024 yang dirilis oleh IQAir, Indonesia menduduki peringkat ke-15 sebagai negara yang memiliki tingkat pencemaran udara tertinggi di dunia, serta juga negara yang pencemaran udara paling tinggi di kawasan Asia Tenggara (IQAir, 2024). Selain itu Jakarta menempati peringkat ke-8 sebagai kota dengan rata-rata tahunan PM2.5 tertinggi pada tahun 2024, yaitu sebesar 41,7 µg/m³ (Hasan, 2024). Pencemaran udara menimbulkan berbagai dampak serius terhadap kesehatan, baik akibat paparan dalam jangka panjang maupun jangka pendek, terutama bagi seseorang yang rentan dan sensitif terhadap gangguan kesehatan.

Paparan jangka pendek berkaitan dengan kondisi saluran pernapasan seperti penyakit paru obstruktif kronik (PPOK), batuk, sesak napas, asma, serta peningkatan angka rawat inap. Sementara itu, paparan jangka panjang dikaitkan dengan asma kronis, penurunan fungsi paru, penyakit kardiovaskular, peningkatan angka kematian, serta risiko diabetes. Dampak lainnya yaitu terdapat hubungan antara konsentrasi partikulat (PM), khususnya partikel halus dan ultrahalus, dengan peningkatan morbiditas dan mortalitas. Hal ini dimungkinkan karena partikel tersebut mampu menembus saluran pernapasan terdalam hingga masuk ke dalam aliran darah (Manisalidis et al., 2020). Faktor risiko polusi udara terhadap penyakit PPOK memiliki risiko 36,6%, pneumonia 32%, asma 27,95%, kanker paru12,5%, dan tuberculosis 12,2% (Kemenkes, 2023).

Polusi udara memiliki dampak yang cukup signifikan di wilayah perkotaan, salah satunya yaitu DKI Jakarta yang merupakan pusat pemerintahan dan kegiatan ekonomi nasional dengan aktivitas padat menghadapi tekanan pencemaran udara yang cukup tinggi. Peningkatan aktivitas industri, transportasi, serta urbanisasi menyebabkan konsentrasi polutan seperti *particulate matter* (PM10, PM2.5), karbon monoksida (CO), nitrogen dioksida (NO2), sulfur dioksida (SO2), dan ozon (O3) meningkat secara signifikan. Pencemaran udara terbentuk saat zat atau energi

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/474

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

yang berasal dari beraneka macam aktivitas manusia masuk ke atmosfer sehingga menurunkan kualitas udara mencapai ambang batas tertentu, maka udara kehilangan kemampuannya untuk menjalankan perannya secara optimal. Seiring meningkatnya polusi udara, diperlukan pengembangan sistem prakiraan dan peringatan dini untuk memantau serta mengendalikan tingkat pencemaran.

Pengukuran kualitas udara penting dilakukan untuk mengetahui kondisi lingkungan suatu wilayah. Penentuan tingkat kualitas udara di wilayah perkotaan dapat dilihat dengan memanfaatkan pendekatan *machine learning*, yang berperan dalam mengklasifikasikan kondisi kualitas udara (Hamami & Dahlan, 2022). Indeks Standar Pencemaran Udara (ISPU) adalah nilai non satuan yang menunjukkan tingkat kualitas udara ambien di wilayah tertentu. Penentuan nilai ISPU didasarkan pada besarnya dampak pencemaran udara terhadap kesehatan manusia, kenyamanan lingkungan, serta makhluk hidup lainnya. Nilai ISPU ditetapkan sesuai dengan Peraturan Menteri Lingkungan Hidup dan Kehutanan Republik Indonesia Nomor P.14/MENLHK/SETJEN/KUM.1/7/2020. Parameter pencemaran yang ditetapkan adalah PM10, PM2.5, karbon monoksida (CO), nitrogen dioksida (NO2), sulfur dioksida (SO2), ozon (O3), dan hidrokarbon(HC). Selanjutnya nilai ISPU diklasifikasikan pada kategori kualitas udara seperti Baik, Sedang, Tidak Sehat, Sangat Tidak Sehat, dan Berbahaya yang memiliki implikasi langsung terhadap rekomendasi aktivitas masyrakat.

Salah satu algoritma yang banyak digunakan untuk klasifikasi adalah *Random Forest*, yaitu teknik *Machine Learning* yang mengumpulkan banyak pohon keputusan guna mengurangi korelasi antar fitur dan mencegah *overfitting*. Metode ini bekerja dengan memilih sampel data dan subset fitur secara acak dalam pembentukkan setiap pohon, sehingga meningkatkan akurasi model (Salman et al., 2024). Adapun penelitian terdahulu yang dilakukan oleh (Fei et al., 2022) di sepuluh kabupaten penghasil kapas di Xinjiang (China). Penelitian ini bertujuan untuk mengembangkan metode klasifikasi kapas menggunakan kombinasi fitur spectral, indeks vegetasi, dan fitur tekstur berbasis *Gray Level Co-occurrence Matrix* (GLCM) serta algoritma *Random Forest* (RF). Hasilnya menunjukkan bahwasanya RF memiliki akurasi dan stabilitas lebih baik dibandingkan *support vector machine* (SVM) dan *artificial neural network* (ANN). Klasifikasi dengan kombinasi multi-fitur menghasilkan *overall accuracy* sebesar 93,36%, lebih tinggi dibandingkan spektrum satu waktu maupun multiwaktu. Bahkan setelah seleksi fitur, akurasi RF tetap tinggi yaitu 92,12%.

Penelitian lain dilakukan oleh (Dou et al., 2019) yaitu membandingkan performa Decision Tree (DT) dan Random Forest (RF) untuk memodelkan kerentanan longsor di Pulau Vulkanik Izu-Oshima Jepang akibat hujan lebat pada skala regional menggunakan 12 faktor pemicu seperti elevasi, kemiringan, dan curah hujan kumulatif. Hasil menunjukkan kedua model sangat akurat (AUC > 0,9), namun RF lebih unggul dengan AUC = 0,956 dibandinkan DT = 0,928, serta RF lebih robust terhadap variasi data training. Oleh karena itu, RF direkomendasikan untuk pembuatan peta rawan longsor di area vulkanik agar dapat mencegah bencana serta pengambilan keputusan.

Penelitian yang dilakukan oleh (Balogun & Tella, 2022) di Malaysia. Penelitian tersebut memakai sepuluh stasiun pemantauan kualitas udara yang berada di kawasan perkotaan, industri, dan permukiman. Penelitian ini bertujuan untuk menganalisis pengaruh variabel iklim terhadap konsentrasi ozon permukaan. Analisis dilakukan pada periode monsun barat daya menggunakan pendekatan korelasi dan empat algoritma *Machine Learning*, yaitu *Random Forest, Decision Tree Regression, Linear Regression,* dan *Support Vector Regression*. Penelitian ini mengungkapkan adanya hubungan yang kuat antara suhu udara dan konsentrasi ozon, korelasi sedang hingga kuat dengan kecepatan angin, serta korelasi negarif dengan kelembapan relatif. Dari keempat algoritma yang diuji, *Random Forest* menunjukkan kinerja prediksi terbaik dengan nilai R^2 sebesar 0,970, RMSE sebesar 2,737, dan MAE sebesar 1,824. Sehingga mengindikasikan bahwa variabel iklim, khususunya suhu dan kecepatan angin, berperan penting dalam peningkatan konsentrasi ozon.

Penelitian dilakukan pada mahasiswa Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Lancang Kuning oleh (Sunaryanto et al., 2022). Penelitian ini bertujuan untuk mengklasifikasikan faktor-faktor yang berpengaruh terhadap durasi pendidikan mahasiswa menggunakan metode *Support Vector Machine* (SVM), *Iterative Dichotomiser* 3 (ID3), *Random Forest, dan K-Nearest Neighbor* (KNN). Hasil penelitian menunjukkan seluruh algoritma mampu melakukan klasifikasi data dengan baik, namun *Random Forest* menghasilkan tingkat akurasi paling tinggi mencapai hingga 100%, diikuti oleh SVM sebesar 94,4%, serta ID3 dan KNN masing-masing sebesar 90%.

Dari hasil penelitian sebelumnya, penelitian ini bertujuan untuk mengklasifikasikan tingkatan pencemaran kualitas udara di wilayah DKI Jakarta melalui penerapan algoritma *random forest*. Pengembangan penelitian difokuskan pada implementasi model klasifikasi dengan membagi data secara stratified, dilanjutkan oleh penilaian kinerja model menggunakan indikator seperti Akurasi, *Precision*, *Recall*, dan *F1-score*, serta eksplorasi parameter polutan udara yang paling signifikan dalam menentukan kategori kualitas udara.

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/474

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

II. METODE PENELITIAN

Penelitian kualitas udara menggunakan metode *random forest* memiliki empat prosedur tahapan. Seluruh proses analisis diimplementasikan menggunakan bahasa pemrograman *Python* serta didukung oleh pustaka (*library)* pendukung seperti *pandas*, *numpy*, *seaborn*, dan *matplotlib.pyplot*. Tahapan alur penelitain tersebut disajikan pada Gambar 1.

Gambar 1. *Flowchart* metode penelitian

A. Pengumpulan Data

Data yang digunakan pada penelitian ini merupakan data sekunder tentang kualitas udara di DKI Jakarta pada periode tahun 2021 hingga 2025. Data ini diperoleh melalui situs web Kaggle https://www.kaggle.com/datasets/senadu34/air-quality-index-in-jakarta-2010-2021 yang terdiri dari 1151 observasi dan 6 variabel. Variabel yang digunakan diringkas pada Tabel 1.

Tabel 1. Variabel Penelitian

Variabel	Deskripsi	Tipe Data
PM2.5	Partikular Mikrometer 2.5	Numerik
PM10	Partikular Mikrometer 10	Numerik
SO2	Sulfur Dioksida	Numerik
CO	Karbon Monoksida	Numerik
O3	Ozon	Numerik
NO2	Nitrogen Dioksida	Numerik
Kategori	Kategori	Kategorik

B. *Preprocessing* Data

Preprocessing data merupakan langkah awal penting dalam data mining karena mampu memperbaiki data, mengubah data mentah menjadi format analitis efektif (Bhargavi Konda, 2022). Dalam upaya meningkatkan validitas informasi data, dilakukan proses pembersihan data untuk menangani missing value dan *outlier* agar data akhir siap digunakan dalam proses pelatihan dan pemodelan. Nilai kosong dapat dibersihkan dengan mengelola data yang tidak lengkap dengan menggunakan teknik imputasi median. Untuk penanganan *outlier* dilakukan menggunakan teknik *capping* dengan metode *Interquartile Range* (IQR), dimana nilai ekstrem tidak dihapus melainkan dibatasi pada nilai ambang tertentu.

C. *Training* & *Testing* Data

Untuk menilai tingkat keakuratan model dalam mengklasifikasikan data dengan benar, digunakan metode *K-Fold Cross-Validation.* Parameter *k* menentukan jumlah lipatan (*fold*) yang digunakan dalam proses *K-Fold Cross-Validation*, dimana setiap *fold* dalam dataset memiliki kesempatan untuk digunakan sebanyak (k-1) kali sebagai data *training*, sehingga setiap bagaian data dapat terwakili secara merata. Hal ini memastikan bahwa seluruh perspektif data tercakup dan memugkinkan model untuk memahami kinerja sebenarnya secara lebih efisien. Pada umunya, nilai *k* yang digunakan berada pada rentang 5 hingga 10 (Prusty et al., 2022). Pada penelitian ini digunakan nilai $k = 10$ karena mampu memberikan keseimbangan yang baik antara bias dan varians serta efisiensi komputasi. Data dipartisi menjadi 10 bagian (*fold*) yang memiliki ukuran sama. Proses *training* dan *testing* dilakukan sebanyak 10 kali, di mana pada setiap iterasi satu *fold* digunakan sebagai data *testing*, sedangkan sembilan fold lainnya digunakan sebagai data *training*. Teknik ini yang dikenal sebagai 10-*Fold Cross Validation*. dengan begitu setiap bagian data memperoleh kesempatan yang sama untuk menjadi data pengujian, sehingga evaluasi model menjadi lebih objektif dan stabil.

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/474

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

D. Implementasi Random Forest

Penelitian ini mengimplementasikan teknik pendekatan *Supervised Learning* dengan algoritma *Random Forest* untuk mengklasifikasikan kategori kualiatas udara. *Random forest* merupakan penggabungan algoritma *Classification and Regression Trees* (CART) dan *boostrapping aggregation* (*bagging)*. Dalam CART, sebuah pohon klasifikasi tunggal disusun dengan simpul-simpul (nodes) yang terhubung (Sarica et al., 2017). Pohon keputusan dibangun dengan membagi dataset secara berulang melalui simpul dan cabang hingga memenuhi syarat tertentu. Langkah tahapanya dimulai dengan mengidentifikasikan fitur utama sebagai akar pohon menggunakan perhitungan *entropi*. Setalah itu, nilai *Information Gain* dihitung untuk mengukur efektivitas setiap fitur dalam mengklasifikasikan data. Tahapan ini diakhiri dengan pembentukkan struktur pohon keputusan. Rumus yang digunakan untuk menghitung entropi dan information gain disajikan pada persamaan 1 dan 2.

$$Entrophy\ (S) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{1}$$

Dimana : S = Kumpulan (himpunan) data
C = Jumlah kategori dalam dataset
p_i = Probabilitas kemunculan suatu kelas ke-i tertentu dalam data set

$$Gain(S, A) = E(S) - \sum_{\epsilon\ Values(A)} \frac{|S_i|}{|S|} E(S_i) \tag{2}$$

Dimana : A = Fitur yang diuji
S_i = bagian subset data yang dihasilkan setelah pembagian berdasarkan fitur A

Klasifikasi akhir dalam *random forest* diperoleh dari *majority voting* dari semua pohon yang ditunjukkan pada persamaan 3.

$$\hat{y} = mode\{T_1(X), T_2(X), \dots, T_N(X)\} \tag{3}$$

Dimana : $T_N(X)$ = Prediksi yang dihasilkan oleh pohon ke-N
$\hat{y}$ = Hasil akhir prediksi

Mekanisme voting merupakan inti dari kekuatan *random forest* karena menggabungkan hasil prediksi dari banyak pohon yang dilatih pada data berbeda., sehingga mampu mengurangi bias dan variansi serta meningkatkan kemampuan model untuk menggeneralisasi data yang tidak ada dilatih sebelumnya.

E. Akurasi Model

Evaluasi performa model klasifikasi dilakukan dengan cara melakukan perbandingan terhadap hasil prediksi model nilai sesungguhnya. *Confusion matrix* mennetukan rincian mengenai klasifikasi yang tepat dan untuk setiap kategori kelas. Matriks *Confusion matrix* di sajikan dalam Tabel 2.

Tabel 2. *Confusion Matrix*

		Kelas Pediksi		
		Positif	**Negatif**	**Total**
Kelas Aktual	***Positif***	*True Positive (TP)*	*False Negative (FN)*	*TP+FN*
	Negatif	*False Positive (FP)*	*True Negative (TN)*	*FP+TN*
	Total	*TP+FP*	*FN+TN*	*TP+FN+ FP+TN*

1) Selanjutnya adalah perhitungan Akurasi. Tujuannya Mengukur sejauh mana model memprediksi dengan benar. Perhitungan akurasi dilakukan dengan menggunakan rumus pada persamaan 4

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{4}$$

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/474

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

2) Setelah didapatkan nilai akurasi untuk masing-masing spesifikasi model, langkah berikutnya menghitung nilai precision. Tujuannya yaitu menunjukkan seberapa banyak prediksi positif yang benar, dengan menggunakan rumus pada persamaan 5

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

3) Recall (Sensitivity) menunjukkan seberapa banyak kasus positif sebenarnya yang dapat dideteksi, dengan menggunakan rumus pada persamaan 6

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

4) F1-Score yaitu rata-rata harmonik dari precision dan recall yang bertujuan untuk mengevaluasi keseimbangan antara ketepatan dan kelengkapan hasil klasifikasi, dengan menggunakan rumus pada persamaan 7

$$F1 = 2 \times \frac{Precision \times Recall}{Precision \quad Recall} \tag{7}$$

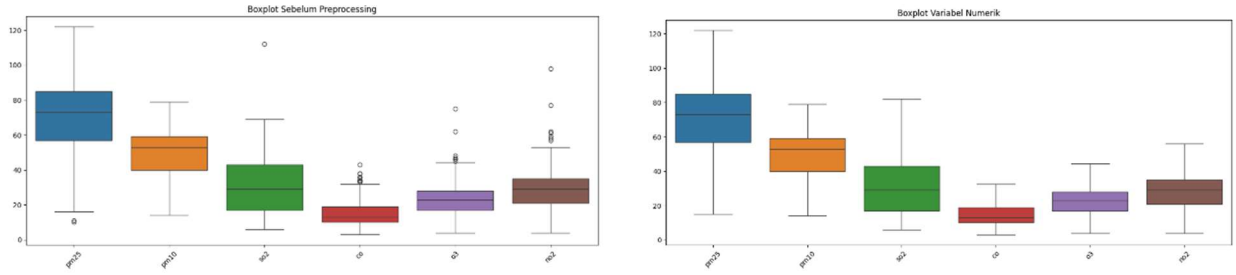
III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data kualitas udara di DKI Jakarta pada periode tahun 2021 hingga 2025 yang diperoleh dari website Kaggle akan diolah menggunakan bahasa pemrograman *Python.* Data yang digunakan terdiri dari 1.151 observasi dengan 6 variabel. Variabel bebas terdiri dari PM10, PM2.5, CO, NO2, SO2, dan O3, sedangkan variabel target yaitu variabel kategori yang hanya terbagi menjadi tiga, yakni kualitas udara dengan kategori Baik, Sedang, dan Tidak sehat. Hal ini disebabkan oleh rentang nilai ISPU yang relatif rendah selama periode tahun 2021 hingga 2025 sehingga tidak ada observasi yang masuk ke kategori Sangat Tidak Sehat dan Berbahaya.

B. Preprocessing Data

Tahap Preprocessing data berfungsi agar meningkatkan mutu data sebelum proses pemodelan dilakukan. Pada langkah ini, data dibersihkan dari elemen yang tidak relevan serta nilai yang hilang. Data yang memiliki kategori tidak valid atau kosong dihapus karena tidak dapat digunakan dalam proses klasifikasi. Untuk penanganan *missing value* pada variabel numerik dilakukan menggunakan metode imputasi median untuk mengurangi pengaruh nilai *ekstrem*. Selanjutnya *outlier* ditangani menggunakan metode *Interquartile Range* (IQR) untuk mendeteksi dan menghapus nilai *ekstrem* tanpa menghilangkan data penting, sehingga data menjadi lebih stabil dan siap digunakan dalam proses pemodelan. Perbandingan data sebelum dan sesudah dilakukan preprocessing disajikan pada Gambar 2.



(a) **(b)**

Gambar 2. Visualisasi Boxplot (a) Sebelum Preprocessing (b) Setelah Preprocessing

Gambar 2 menunjukkan perbandingan visualisasi boxplot dari variabel numerik PM2.5, PM10, SO2, CO, O3, dan NO2. Pada Gambar 2(a), variabel CO memiliki *outlier* paling banyak, menunjukkan sebaran data yang paling luas dengan adanya perbedaan signifikan kadar karbon monoksida antar wilayah. Variabel O3 dan NO2 juga menunjukkan sebaran yang cukup lebar dengan beberapa *outlier*, menggambarkan perbedaan konsentrasi ozon dan nitrogen dioksida di berbagai lokasi. Variabel PM2.5 dan SO2 hanya memiliki sedikit *outlier*, yang menunjukkan bahwa konsentrasi kedua polutan tersebut relatif lebih homogen antar wilayah dan tidak mengalami fluktuasi ekstream. Sementara itu, PM10 menunjukkan sebaran data yang relatif stabil tanpa adanya nilai ekstream, menandakan distribusi konsentrasi

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/474

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

yang lebih konsisten. Gambar 2(b) menunjukkan bahwa visualisasi boxplot telah berhasil dibersihkan menggunakan metode IQR dengan teknik *capping*. Metode ini efektif dalam membatasi nilai ekstrem tanpa menghilangkan data, sehingga distribusi data menjadi lebih stabil dan siap digunakan pada tahap pemodelan selanjutnya.

C. Implementasi Random Forest

Untuk menilai tingkat keakuratan model dalam mengklasifikasikan data, maka selanjutnya membagi data memakai teknik *K-Fold Cross Validation* (k = 10). Setiap fold secara bergantian berperan sebagai data *training* dan *testing*. Tabel akurasi data *training* dan *testing* berdasarkan *10-Fold Cross Validation* disajikan dalam Tabel 3.

Tabel 3. Nilai Akurasi *10-Fold Cross Validation*

Fold	Training Accuracy	Testing Accuracy
1	1,0000	0,9914
2	1,0000	0,9826
3	1,0000	0,9826
4	1,0000	0,9826
5	1,0000	0,9913
6	1,0000	0,9913
7	1,0000	1,0000
8	1,0000	0,9391
9	1,0000	0.9913
10	1,0000	0.9913
Rata-rata	1,0000	0.9844

Tabel 3 menunjukkan nilai akurasi menggunakan *10-Fold Cross Validation* menunjukan model *random forest* menggunakan *10-Fold Cross Validation* memperoleh Akurasi training sebesar 1,0000 atau 100% pada seluruh *fold*. Sementara itu, nilai akurasi testing pada setiap *fold* menunjukkan rata-rata akurasi sebesar 0,9844 atau 98,44%. Hasil tersebut mengindikasikan bahwa algoritma *random forest* berhasil mengklasifikasikan data dengan tingkat keunggulan yang tinggi dan menunjukkan stabilitas performa yang konsisten selama fase *training* dan *testing*. mampu mengklasifikasikan data dengan sangat baik serta memiliki kinerja yang stabil dalam proses *training* dan *testing*.

D. Akurasi Model

1) Evalusi model

Rincian evaluasi model dianalisis menggunakan klasifikasi standar, meliputi akurasi, *precision, recall, F1-Score, dan Support*. Rincian evaluasi model random forest disajikan dalam Tabel 4.

Tabel 4. Evaluasi Model Random Forest

Kelas	*Precision*	*Recall*	*F1-Score*	*Support*
Baik	0,96	0,96	0,96	192
Sedang	0,99	0,99	0,99	913
Tidak Sehat	0,96	0,98	0,97	46
Accuracy			0,98	1151
Macro Avg	0,97	0,98	0,97	1151
Weighted Avg	0,98	0,98	0,98	1151

Tabel 4 menunjukkan hasil penilaian menggunakan teknik *10-Fold Cross Vakidation*, algoritma *random forest* menampilkan hasil klasifikasi yang sangat bagus dengan akurasi yang diperoleh sebesar 98%. Pada kategori kualitas udara Baik, nilai *Precision* sebesar 0,96 menandakan 96% prediksi yang diberikan pada kategori Baik adalah benar, sementara *Recall* sebesar 0,96 yang berrati model berhasil mengenali 96% data kategori Baik dengan tepat, sehingga menghasilkan *F1-Score* sebesar 0,96 yang mencerminkan keseimbangan optimal antara ketepatan dan sensitivitas. Pada kategori kualitas udara Sedang, model menunjukkan performa sangat tinggi dengan *Precision* dan *Recall* masing-masing sebesar 0,99, yang berarti hampir seluruh data kategori kualitas udara Sedang berhasil diklasifikasikan secara benar, serta menghasilkan *F1-Score* sebesar 0,99 yang mempresentasikan keseimbangan optimal antara ketepatan dan sensitivitas. Pada kategori kualitas udara Tidak

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/474

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

sehat, meskipun jumlah data relatif sedikit, model tetap memberikan performa yang baik dengan *Precision* sebesar 0,96, *Recall* sebesar 0,98, dan *F1-Score* sebesar 0,97 menunjukkan kemampuan model dalam mendeteksi kelas minoritas. Nilai *Marco Avg* yang terdiri dari *Precision* 0,97, *Recall* 0,98, dan *F1-Score* 0,97 yang mengindikasikan model mempunyai performa yang stabil di seluruh kategori tanpa mempertimbangkan jumlah data tiap kategori. Sementara itu, *Weighted Avg* dengan nilai *Precision*, *Recall*, dan *F1-Score* masing-masing sebesar 0,98 menujukkan kinerja model tetap optimal ketika mempertimbangkan proporsi data pada setiap kategori, sehingga dapat disimpulkan bahwa algoritma *random forest* menunjukkan kemampuan generalisasi yang kuat dan stabil dalam melakukan klasifikasi.

2) Tabel Confusion Matrix

Confusion matrix pada model *random forest* menunjukkan kemampuan klasifikasi yang sangat baik dalam membedakan tiga kategori, yaitu baik, sedang, dan tidak sehat. Tabel confusion matrik disajikan pada Tabel 5.

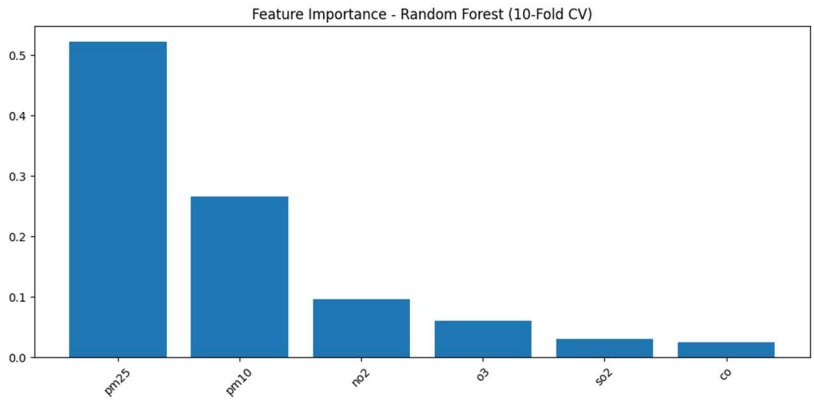
Tabel 5. Confusion Matrix

Actual / Predicted	Baik	Sedang	Tidak Sehat
Baik	184	8	0
Sedang	7	907	2
Tidak Sehat	0	0	45

Berdasarkan Tabel 5 terlihat bahwa hampir seluruh data pada kategori kualitas udara Baik dapat diprediksi dengan benar sebanyak 184 data yang diklasifikasikan ke kategori baik, walaupun terdapat 8 data yang salah diklasifikasikan ke kategori Sedang. Kelas Sedang memiliki performa paling tinggi yaitu dengan 907 data diklasifikasikan dengan benar ke kategori sedang, dengan 7 data salah ke kategori Baik dan 2 data salah ke kategori Tidak sehat, sementara kelas Tidak sehat diprediksi dengan benar sebanyak 45 data tanpa kesalahan klasifikasi.

3) Grafik *Feature Importance*

Analisis *feature importance* digunakna untuk mengetahui seberapa besar peran setiap variabel dalam memengaruhi hasil prediksi suatu model terutama pada model berbasis pohon. Grafik *feature importance* disajikan pada Gambar 3.



Gambar 3. Grafik *Feature Importance*

Terlihat gambar 3 menujukkan bahwa variabel PM2.5 pengaruh paling besar terhadap hasil prediksi, dengan nilai kepentingan lebih dari 0,5 yang berarti PM2.5 menjadi faktor utama yang menentukan keputusan model. Hal ini sesuai dengan sifat data dan menegaskan pentingnya pengendalian partikulat mikrometer di Jakarta.

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/474

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Sesuai hasil penelitian (Suleiman et al., 2020) bahwa algoritma *Random Forest* (RF), *Extreme Learning Machine* (ELM), dan *Deep Learning* (DL) mamou memodelkan konsentrasi partkulat PM10 dan PM2.5 di area tepi jalan dengan kinerja yang tinggi. Model machine learning menunjukkan performa rata-rata yang lebih baik dalam memprediksi PM2.5 dibandingkan PM10, dengan nilai koefisien korelasi (R) masaing-masing sebesar 0,90 untuk partikulat PM10 dan 0,94 untuk partikulat PM2.5. nilai FAC2 sebesar 99% dan 98%, RMSE sebesar 9,2 dan 4,5 serta Index of Agreement (IOA) sebesar 0,83 dan 0,84. Selain memiliki kaurasi yang tinggi, metode RF, ELM, dan DL juga menunjukkan kecepatan pelatihan dan skalabilitas yang lebih baik dibandingkan Artificial Neural Network (ANN) tradisional, sehingga lebih efektif untuk pemodelan kualitas udara berbasis data lalu lintas dan meterologi.

Hasil penelitian terdahulu sejalan dengan hasil penelitian ini, dimana random forest efektif dalam mengklasifikasikan kualiatas udara dan menghasilkan model dengan tingkat akurasi yang tinggi.

IV. KESIMPULAN

Algoritma *random forest* berhasil digunakan sebagai klasifikasi kualitas udara di DKI Jakarta tahun 2021 sampai 2025 secara akurat berdasarkan data ISPU. Model random forest menunjukkan kinerja tingkat akurasi yang sangat tinggi pada data *training* yaitu sebesar 100% dan akurasi *testing* sebesar 98,44%. Dalam *machine learning*, *overfitting* terjadi ketika model terlalu menyesuaikan data latih sehingga tidak mampu menggeneralisasi dengan baik akibat mempelajari *noise*, sehingga evaluasi dengan data uji diperlukan untuk mendeteksinya melalui perbedaan kinerja antara data *training* dan *testing*. Namun pada penelitia ini model tidak mengalami *overfitting* karena akurasi data *testing* tetap tinggi dan selisihnya kecil terhadap akurasi data t*raining*, yang menunjukkan kemampuan generalisasi yang baik. Evaluasi menggunakan confusion matrix menunjukkan bahwa sebagian besar data pada setiap kategori kualitas udara berhasil di klasifikasikan dengan benar, khusunya pada kategori kualitas udara Tidak Sehat yang tidak memiliki tingkat kesalahan. Analisis feature importance menunjukkan variabel PM2.5 merupakan faktor pencemar yang paling berpengaruh dalam penentuan kategori kualitas udara. Hal ini menegaskan bahwa partikulat mikrometer 25 memiliki peran penting terhadap kondisi kualitas udara di DKI Jakarta. Dengan demikian, penelitian ini menunjukkan bahwa algoritma *random forest* efektif sebagai pendekatan klasifikasi kulaitas udara dan dapat berfungsi sebagai fondasi untuk mendukung pembuatan kebijakan pengendalian polusi pencemaran udara.

Sebagai pengembangan lebih lanjut, untuk penelitian selanjutnya disarankan untuk memperhatikan adanya ketidakseimbangan jumlah data (*data imbalance*) antar kelas. Oleh karena itu, diperlukannya penerapan teknik penangan data tidak seimbang agar kinerja model dapat meningkat serta kesalahan klasifikasi antar kelas yang berdekatan dapat diminimalkan.

DAFTAR PUSTAKA

Balogun, A.-L., & Tella, A. (2022). Modelling and investigating the impacts of climatic variables on ozone concentration in Malaysia using correlation analysis with random forest, decision tree regression, linear regression, and support vector regression. *Chemosphere*, *299*, 134250. https://doi.org/10.1016/j.chemosphere.2022.134250.

Bhargavi Konda. (2022). The impact of data preprocessing on data mining outcomes. *World Journal of Advanced Research and Reviews*, *15*(3), 540–544. https://doi.org/10.30574/wjarr.2022.15.3.0931

Dou, J., Yunus, A. P., Tien Bui, D., Merghadi, A., Sahana, M., Zhu, Z., Chen, C.-W., Khosravi, K., Yang, Y., & Pham, B. T. (2019). Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Science of The Total Environment*, *662*, 332–346. https://doi.org/10.1016/j.scitotenv.2019.01.221

Fei, H., Fan, Z., Wang, C., Zhang, N., Wang, T., Chen, R., & Bai, T. (2022). Cotton Classification Method at the County Scale Based on Multi-Features and Random Forest Feature Selection Algorithm and Classifier. *Remote Sensing*, *14*(4), 829. https://doi.org/10.3390/rs14040829.

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/474

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Hamami, F., & Dahlan, I. A. (2022). Air Quality Classification in Urban Environment using Machine Learning Approach. *IOP Conference Series: Earth and Environmental Science*, *986*(1), 012004. https://doi.org/10.1088/1755-1315/986/1/012004.

Hasan, K. (2024). *Kualitas udara Indonesia 2024: Polusi Jabodetabek capai sepuluh kali lipat batas aman WHO, Pemerintah 'menunggu dan lihat nanti.'* https://kemkes.go.id/id/polusi-udara-sebabkan-angka-penyakit-respirasi-tinggi.

IQAir. (2024). *World's most polluted cities The most polluted cities according to data aggregated from over 80,000 data points.* https://www.iqair.com/in-en/world-most-polluted-cities?continent=&country=&state=&page=1&perPage=50&cities=

Kemenkes. (2023, April 4). *Polusi Udara Sebabkan Angka Penyakit Respirasi Tinggi*. Kementerian Kesehatan Republik Indonesia. https://kemkes.go.id/id/polusi-udara-sebabkan-angka-penyakit-respirasi-tinggi

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and Health Impacts of Air Pollution: A Review. In *Frontiers in Public Health* (Vol. 8). Frontiers Media S.A. https://doi.org/10.3389/fpubh.2020.00014.

Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, *4*. https://doi.org/10.3389/fnano.2022.972421.

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/BJML/2024/007.

Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Frontiers in Aging Neuroscience*, *9*. https://doi.org/10.3389/fnagi.2017.00329.

Suleiman, A., Tight, M. R., & Quinn, A. D. (2020). A comparative study of using Random Forests (RF), Extreme Learning Machine (ELM) and Deep Learning (DL) algorithms in modelling Roadside Particulate Matter (PM_{10} & $PM_{2.5}$). *IOP Conference Series: Earth and Environmental Science*, *476*, 012126. https://doi.org/10.1088/1755-1315/476/1/012126.

Sunaryanto, H., Hasan, M. A., & Guntoro, G. (2022). Classification Analysis of Unilak Informatics Engineering Students Using Support Vector Machine (SVM), Iterative Dichotomiser 3 (ID3), Random Forest and K-Nearest Neighbors (KNN). *IT Journal Research and Development*, *7*(1), 36–42. https://doi.org/10.25299/itjrd.2022.8912.

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/474