DIGINTEL-AI: DIGITAL INNOVATION AND INTELLIGENCE – AI
e-ISSN: 3123-8076
Journal homepage: https://journal.ajirapublisher.id/index.php/digintelai

Klasifikasi Kualitas Udara Berdasarkan Parameter Lingkungan Menggunakan Algoritma *Decision Tree* C4.5

Air Quality Classification Based on Environmental Parameters Using the C4.5 Decision Tree Algorithm

Galang Putra Wicaksana[1], Irwansyah[2*], Ahmad Taufik[3], Faldy Irwiensyah[4]

[1,2]Teknik Informatika, Fakultas Teknologi Industri dan Informatika, Universitas Muhammadiyah Prof. Dr. Hamka, Jakarta, Indonesia.
[3]Sistem Informasi, Fakultas Teknologi, Institut Sosial dan Teknologi Widuri, Jakarta, Indonesia.
[4]Sistem dan Teknologi Informasi, Fakultas Teknologi Industri dan Informatika, Universitas Muhammadiyah Prof. Dr. Hamka, Jakarta, Indonesia.
*Corresponding Author: irwansyah@uhamka.ac.id

ARTICLE INFO

Vol. 2, No. 1, October 2026, pp. 11-21.
DOI: https://doi.org/10.66217/digintel-ai.v2i1.22

Article history:
Revised: 12 October 2026
Accepted: 200 October 2026
Published: 30 October 2026

Abstract

Air quality directly affects public health. In Indonesia it is communicated through the Air Pollution Standard Index (ISPU), which is calculated from six pollutants: PM2.5, PM10, SO2, CO, O3, and NO2. This study aims to build an accurate and interpretable model that classifies ISPU categories using the C4.5 Decision Tree algorithm. The secondary dataset from Kaggle contains 1,416 air quality monitoring records from DKI Jakarta for 2021–2025. Preprocessing consisted of median and mode imputation for missing values, removal of one unlabeled record, merging the very small "Very Unhealthy" class (4 records) into "Unhealthy", and label encoding, which resulted in 1,415 records in three classes (Good, Moderate, Unhealthy). The data were split 70:30 with stratification into 990 training and 425 testing records. The tree was built from entropy and information gain, pruned from 39 to 19 nodes (depth from nine to five levels), and converted into ten IF-THEN rules. PM2.5 had the highest information gain (0.8717) and became the root node, with a feature importance of 0.9048. The model correctly classified 416 of 425 test records, an accuracy of 97.88%. The F1-scores for the Unhealthy, Moderate, and Good classes were 99.35%, 98.27%, and 69.57%, respectively; the lower score for Good reflects only 11 test records. SMOTE and ADASYN did not improve accuracy (95.06% and 95.53%). The results show that C4.5 can classify ISPU categories accurately using transparent rules.

Keywords: *Air Quality; Classification; C4.5 Decision Tree; ISPU; Machine Learning.*

Abstrak

Kualitas udara berpengaruh langsung terhadap kesehatan manusia. Di Indonesia, kondisinya dikomunikasikan melalui Indeks Standar Pencemar Udara (ISPU) yang dihitung dari enam polutan, yaitu PM2.5, PM10, SO2, CO, O3, dan NO2. Penelitian ini bertujuan membangun model klasifikasi kategori ISPU yang akurat dan mudah ditafsirkan menggunakan algoritma *Decision Tree* C4.5. Data sekunder dari Kaggle berisi 1.416 catatan pemantauan kualitas udara DKI Jakarta periode 2021–2025. Prapemrosesan meliputi imputasi median dan modus untuk nilai hilang, penghapusan satu data tanpa label, penggabungan kelas SANGAT TIDAK SEHAT (4 data) ke TIDAK SEHAT, dan *label encoding* sehingga diperoleh 1.415 data dengan tiga kelas (BAIK,

Copyright © 2025 by the authors. Licensee DIGINTEL-AI, Jakarta, Indonesia. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International License (CC BY-NC-SA 4.0) (https://creativecommons.org/licenses/by-nc-sa/4.0/)

SEDANG, TIDAK SEHAT). Data dibagi 70:30 secara stratifikasi menjadi 990 data latih dan 425 data uji. Pohon keputusan dibangun berdasarkan *entropy* dan *information gain*, dipangkas dari 39 menjadi 19 simpul (kedalaman 9 menjadi 5 level), lalu diekstraksi menjadi 10 aturan *IF-THEN*. PM2.5 memiliki *information gain* tertinggi (0,8717) sehingga menjadi simpul akar, dengan *feature importance* 0,9048. Model mengklasifikasikan 416 dari 425 data uji dengan benar (akurasi 97,88%). *F1-score* kelas TIDAK SEHAT, SEDANG, dan BAIK berturut-turut 99,35%, 98,27%, dan 69,57%; nilai kelas BAIK yang lebih rendah berkaitan dengan hanya 11 data uji. SMOTE dan ADASYN tidak meningkatkan akurasi (95,06% dan 95,53%). Hasil ini menunjukkan bahwa C4.5 mampu mengklasifikasikan kategori ISPU secara akurat dengan aturan yang transparan.

Kata Kunci: *Decision Tree* C4.5; ISPU; Klasifikasi; Kualitas Udara; *Machine Learning.*

1. Pendahuluan

Kualitas udara merupakan salah satu faktor lingkungan yang menentukan kesehatan manusia. Di Indonesia, kondisi udara ambien dinyatakan melalui Indeks Standar Pencemar Udara (ISPU) yang dihitung dari enam polutan, yaitu partikulat (PM2.5 dan PM10), nitrogen dioksida (NO_2), sulfur dioksida (SO_2), karbon monoksida (CO), dan ozon permukaan (O_3) (Maharani & Aryanta, 2023). Kenaikan konsentrasi polutan tersebut menurunkan kualitas udara dan dapat membahayakan kesehatan (Ahmad et al., 2025; Harpad et al., 2022). Karena itu, nilai ISPU dikelompokkan menjadi kategori Baik, Sedang, Tidak Sehat, Sangat Tidak Sehat, dan Berbahaya sebagai informasi bagi masyarakat (Fenny & Herniwanti, 2025).

Pemantauan kualitas udara dilakukan melalui *Air Quality Monitoring System* (AQMS), dan saat ini tercatat 72 unit AQMS resmi di Indonesia. Namun, sensor kadang gagal mencatat data, pola polutan spesifik sulit dikenali, dan biaya operasionalnya relatif tinggi. Kondisi ini menghasilkan data yang tidak lengkap dan mendorong pemanfaatan *machine learning* untuk menganalisis serta mengklasifikasikan kualitas udara secara otomatis (Akbar et al., 2024; Alfian et al., 2024). Teknik penambangan data semacam ini telah dimanfaatkan di berbagai bidang, misalnya penggalian pola dengan algoritma Apriori (Aushofi et al., 2025; Firmansyah et al., 2025) dan klasifikasi sentimen dengan K-NN serta SVM (Kurniawan et al., 2025; Novryadi et al., 2025; Putri et al., 2025).

Decision tree adalah metode *supervised learning* yang memecah data secara rekursif menjadi simpul akar, simpul internal, dan simpul daun sehingga menghasilkan aturan keputusan yang mudah dibaca (Irwansyah et al., 2023; Tangkelayuk & Mailoa, 2022). Algoritma C4.5 dipilih dalam banyak studi karena mampu menangani atribut numerik dan kategorikal serta nilai yang hilang (Nazifah & Prianto, 2023; Romli & Zy, 2020). Pada kualitas udara, C4.5 telah digunakan untuk memetakan pencemaran udara di Samarinda dan mengenali polutan dominan (Prafanto et al., 2024), mengelompokkan status udara Tangerang Selatan ke dalam tiga kategori beserta visualisasi pohonnya (Ragga et al., 2022), membandingkannya dengan *Support Vector Machine* pada data Jakarta (Sang et al., 2021), serta membandingkannya dengan algoritma lain pada data Jakarta tahun 2023 (Kannajmi & Saputra, 2025). Saputra dan Firmansyah (2025) memprediksi kualitas udara dengan RapidMiner dan memperoleh akurasi 89,52%. Algoritma ini juga dipakai pada data meteorologi (Najah et al., 2025; Prasetiyo, 2024).

Meskipun demikian, penelitian-penelitian tersebut umumnya berfokus pada akurasi akhir dan belum membahas secara memadai penanganan nilai hilang, ketidakseimbangan kelas, serta pemangkasan pohon (*pruning*). Sebagian studi juga memakai data sampai tahun 2023 sehingga belum menggambarkan kondisi DKI Jakarta pada periode yang lebih baru. Pada studi yang ditinjau, kontribusi masing-masing polutan terhadap keputusan model dan aturan klasifikasi yang dihasilkan juga belum disajikan secara kuantitatif.

Model yang akurat, ringkas, dan transparan penting bagi instansi lingkungan yang membutuhkan dasar klasifikasi yang dapat dijelaskan, serta bagi masyarakat yang membutuhkan informasi kualitas udara yang mudah dipahami. Penelitian ini bertujuan (1) membangun model klasifikasi kategori ISPU berdasarkan enam parameter polutan dengan *Decision Tree* C4.5 pada data DKI Jakarta periode 2021–2025; (2) menghasilkan struktur pohon yang telah dipangkas beserta aturan *IF-THEN* dan menganalisis kontribusi tiap polutan; dan (3) mengukur kinerja model dengan *confusion matrix*, *accuracy*, *precision*, *recall*, dan *F1-score*. Kontribusi penelitian ini adalah prapemrosesan data yang terdokumentasi (imputasi, penggabungan kelas, dan uji penyeimbangan kelas), pemangkasan pohon yang menyederhanakan struktur model, serta aturan keputusan eksplisit yang dapat langsung dipakai.

2. Metodologi

Penelitian ini bersifat kuantitatif dengan data sekunder. Tahapannya adalah pengumpulan data, eksplorasi data, prapemrosesan, pembagian data, pembangunan model *Decision Tree* C4.5, dan evaluasi model. Seluruh tahapan dijalankan dengan Python 3.13 di Jupyter Notebook menggunakan pustaka Pandas, NumPy, Scikit-learn, dan Matplotlib, pada laptop dengan prosesor AMD Ryzen 3 3300U, memori 4 GB, dan sistem operasi Windows 11. Kode program dan hasil prediksi tersedia di https://github.com/galang-after/Hasil-Penelitian-ISPU-Dengan-Algoritma-Decision-Tree-C4.5.

2.1 Pengumpulan Data

Dataset sekunder diunduh dari Kaggle dalam format *Comma-Separated Values* (CSV). *Dataset* berisi 1.416 baris dan 11 atribut hasil pemantauan kualitas udara DKI Jakarta pada 15 April 2021 sampai 28 Februari 2025, yaitu tanggal, stasiun, PM2.5, PM10, SO_2, CO, O_3, NO_2, *max*, *critical*, dan *categori*. Tujuh atribut bertipe numerik (enam polutan dan *max*), sedangkan empat atribut lainnya bertipe teks. Kolom *categori* (kategori ISPU) menjadi variabel target, dan konsentrasi enam polutan menjadi fitur.

2.2 Eksplorasi Data

Eksplorasi data mencakup pemeriksaan baris awal dan akhir, struktur dan tipe atribut, statistik deskriptif (rerata, simpangan baku, kuartil, nilai minimum dan maksimum), jumlah nilai hilang pada tiap atribut, serta distribusi kelas target.

2.3 Prapemrosesan Data

Nilai hilang pada atribut numerik diisi dengan median, dan pada atribut kategorikal dengan modus (Istiana & Mustafiril, 2023). Satu data berkategori TIDAK ADA KETERANGAN dihapus karena bukan kategori ISPU. Kelas SANGAT TIDAK SEHAT hanya berisi 4 data sehingga digabungkan ke kelas TIDAK SEHAT, dan diperoleh tiga kelas target. Kelas target kemudian diubah menjadi angka dengan *label encoding* (BAIK = 0, SEDANG = 1, TIDAK SEHAT = 2) (Daniswara & Nuryana, 2023). Angka tersebut hanya penanda kelas dan tidak menunjukkan tingkatan nilai.

2.4 Pembagian Data

Data dibagi dengan rasio 70% data latih dan 30% data uji secara stratifikasi berdasarkan kelas target agar setiap kelas terwakili pada kedua subset (Farisi et al., 2024). Data latih dipakai untuk membangun model dan menguji penyeimbangan kelas dengan SMOTE dan ADASYN, sedangkan data uji hanya dipakai untuk evaluasi.

2.5 Pembangunan Model *Decision Tree* C4.5

C4.5 memilih atribut pemisah berdasarkan *information gain* yang dihitung dari *entropy* (Islam et al., 2022). Pada atribut numerik, algoritma mencari nilai ambang (*threshold*) yang membagi data menjadi dua kelompok. *Entropy* dan *information gain* dihitung dengan Persamaan (1) dan (2).

$$Entropy(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (1)$$

$$Gain(S, A) = Entropy(S) - \sum_v \frac{|S_v|}{|S|} \times Entropy(S_v) \quad (2)$$

dengan S himpunan data, p_i proporsi data pada kelas ke-i, A atribut yang diuji, dan S_v subset data hasil pemisahan atribut A. Atribut dengan *information gain* tertinggi menjadi simpul akar, lalu proses diulang secara rekursif pada setiap cabang (Lumbanbatu et al., 2026). Pohon yang terbentuk kemudian dipangkas (*pruning*) untuk mengurangi kompleksitas dan risiko *overfitting* (Kahfi et al., 2024; Syahputri & Hasibuan, 2024). Setiap jalur dari simpul akar ke simpul daun selanjutnya diubah menjadi aturan *IF-THEN* (Iqbal et al., 2026).

2.6 Evaluasi Model

Kinerja model diukur pada data uji menggunakan *confusion matrix* yang memuat *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) (Valerian et al., 2025). Dari matriks tersebut dihitung *accuracy*, *precision*, *recall*, dan *F1-score* (Hanum et al., 2024) dengan Persamaan (3) sampai (6). Untuk tiga kelas, akurasi total dihitung sebagai jumlah prediksi benar dibagi jumlah data uji, sedangkan *precision*, *recall*, dan *F1-score* dihitung untuk setiap kelas.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (6)$$

3. Hasil dan Pembahasan

3.1 Hasil Eksplorasi Data

Pemeriksaan struktur data menunjukkan beberapa atribut belum lengkap. PM10 memiliki nilai hilang terbanyak, yaitu 199 data atau sekitar 14% dari 1.416 data, diikuti O_3 (20 data), CO (19), SO_2 (18), NO_2 (14), dan PM2.5 (4), sedangkan atribut *max* dan *critical* masing-masing kehilangan 1 data. Statistik deskriptif menunjukkan PM2.5 memiliki rerata tertinggi (90,85 µg/m³) dan simpangan baku terbesar (26,54), sedangkan CO memiliki rerata terendah (16,27 µg/m³) dengan simpangan baku terkecil (11,92).

Distribusi kelas awal terdiri atas SEDANG 864 data, TIDAK SEHAT 511 data, BAIK 36 data, SANGAT TIDAK SEHAT 4 data, dan TIDAK ADA KETERANGAN 1 data (Tabel 1). Dominasi kelas SEDANG dan TIDAK SEHAT menunjukkan bahwa kualitas udara DKI Jakarta selama periode pengamatan umumnya berada pada kategori sedang dengan sejumlah hari tidak sehat, sekaligus mengindikasikan adanya ketidakseimbangan kelas.

3.2 Hasil Prapemrosesan dan Pembagian Data

Imputasi median menghasilkan nilai PM2.5 = 91, PM10 = 58, SO_2 = 41, CO = 15, O_3 = 28, NO_2 = 20, dan *max* = 92, sedangkan imputasi modus menghasilkan stasiun DKI4 (Lubang Buaya) dan *critical* = PM25. Setelah imputasi, seluruh atribut lengkap (1.416 data). Penghapusan satu data TIDAK ADA KETERANGAN menyisakan 1.415 data, dan penggabungan kelas SANGAT TIDAK SEHAT menghasilkan 36 data BAIK, 864 data SEDANG, dan 515 data TIDAK SEHAT. Pembagian 70:30 menghasilkan 990 data latih dan 425 data uji (Tabel 1).

Tabel 1. Distribusi kelas pada setiap tahap prapemrosesan

Kelas	Data awal	Setelah prapemrosesan	Data latih	Data uji
BAIK (0)	36	36	25	11
SEDANG (1)	864	864	605	259
TIDAK SEHAT (2)	511	515	360	155
SANGAT TIDAK SEHAT	4	digabung ke TIDAK SEHAT	–	–
TIDAK ADA KETERANGAN	1	dihapus	–	–
Total	**1.416**	**1.415**	**990**	**425**

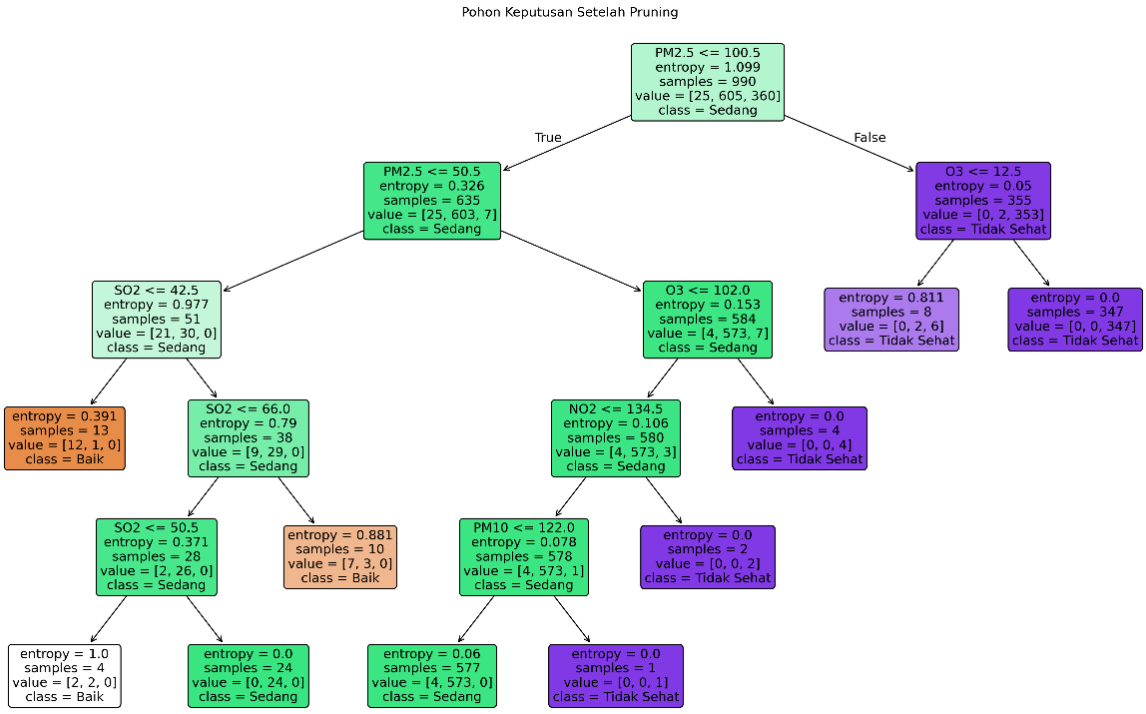
3.3 Pembentukan Model dan Penentuan Simpul Akar

Entropy himpunan data latih (990 data: 25 BAIK, 605 SEDANG, 360 TIDAK SEHAT) adalah 1,0989. Untuk setiap polutan dicari nilai ambang pemisah, kemudian dihitung *entropy* kedua cabang dan *information gain* (Tabel 2). PM2.5 dengan ambang 100,5 memberikan gain tertinggi, yaitu 0,8717. Cabang ≤ 100,5 berisi 635 data (25 BAIK, 603 SEDANG, 7 TIDAK SEHAT) dengan *entropy* 0,3263, sedangkan cabang > 100,5 berisi 355 data (0 BAIK, 2 SEDANG, 353 TIDAK SEHAT) dengan *entropy* 0,0502. Karena itu, PM2.5 ditetapkan sebagai simpul akar. Hasil perhitungan manual ini identik dengan keluaran program.

Tabel 2. Rekapitulasi *entropy* dan *information gain* pada data latih

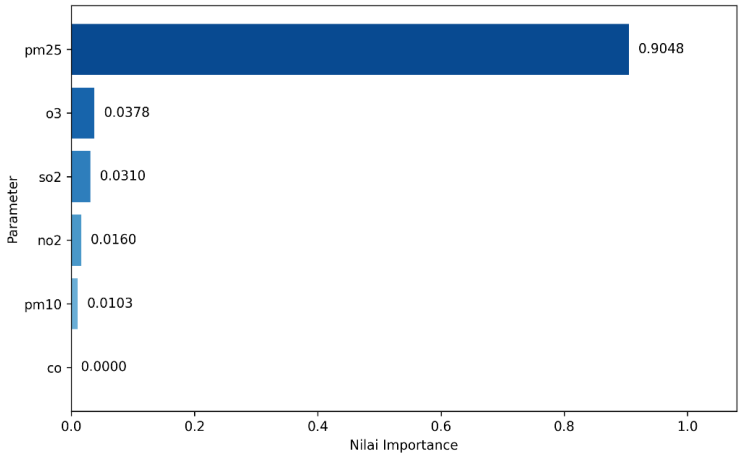
Atribut	Ambang	*Entropy* (≤ ambang)	*Entropy* (> ambang)	*Information gain*
PM2.5	100,5	0,3263	0,0502	0,8717
PM10	50,5	0,6652	1,0122	0,1761
CO	15,5	1,0055	0,9951	0,0981
O_3	12,5	0,8377	1,0504	0,0745
NO_2	20,5	1,0437	1,1206	0,0185
SO_2	35,5	1,1702	1,0647	0,0032

Model awal berisi 39 simpul dengan kedalaman 9 level. Struktur sedalam itu menandakan partisi data yang sangat rinci dan berisiko *overfitting*. Pemangkasan menyusutkan pohon menjadi 19 simpul dengan kedalaman 5 level (Gambar 1). Pada pohon hasil pemangkasan, *entropy* simpul akar sebesar 1,099 turun menjadi 0,326 pada cabang PM2.5 ≤ 100,5 dan 0,050 pada cabang PM2.5 > 100,5. Cabang kanan hampir seluruhnya berisi kelas TIDAK SEHAT (353 dari 355 data latih), sedangkan cabang kiri dipisahkan lagi oleh PM2.5 pada ambang 50,5, kemudian oleh SO_2, O_3, NO_2, dan PM10.



Gambar 1. Struktur *decision tree* setelah pemangkasan (*pruning*)

Skor *feature importance* PM2.5 adalah 0,9048, jauh di atas O_3 (0,0378), SO_2 (0,0310), NO_2 (0,0160), PM10 (0,0103), dan CO (0,0000) (Gambar 2). Dengan demikian, sebagian besar keputusan model bergantung pada PM2.5.



Gambar 2. *Feature importance* parameter polutan

3.4 Aturan *IF-THEN*

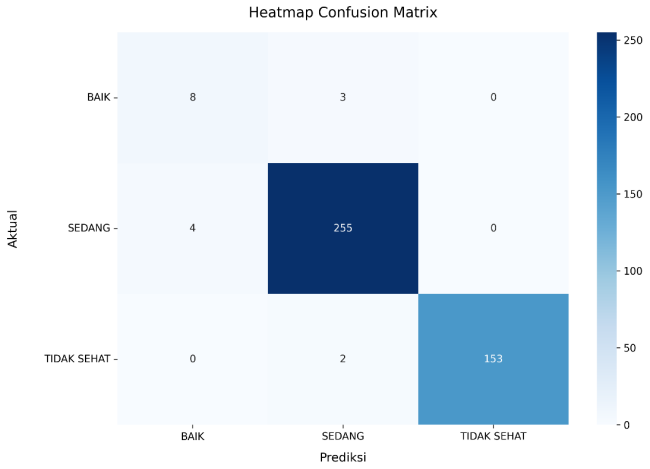
Pohon hasil pemangkasan diekstraksi menjadi 10 aturan *IF-THEN* yang merepresentasikan jalur dari simpul akar ke simpul daun (Tabel 3). Satuan konsentrasi mengikuti *dataset*.

Tabel 3. Aturan *IF-THEN* hasil ekstraksi *decision tree* C4.5

No	Kondisi (*IF*)	Kelas (*THEN*)
1	PM2.5 ≤ 50,5 dan SO_2 ≤ 42,5	BAIK
2	PM2.5 ≤ 50,5 dan 42,5 < SO_2 ≤ 50,5	BAIK
3	PM2.5 ≤ 50,5 dan 50,5 < SO_2 ≤ 66,0	SEDANG
4	PM2.5 ≤ 50,5 dan SO_2 > 66,0	BAIK
5	50,5 < PM2.5 ≤ 100,5 dan O_3 ≤ 102,0 dan NO_2 ≤ 134,5 dan PM10 ≤ 122,0	SEDANG
6	50,5 < PM2.5 ≤ 100,5 dan O_3 ≤ 102,0 dan NO_2 ≤ 134,5 dan PM10 > 122,0	TIDAK SEHAT
7	50,5 < PM2.5 ≤ 100,5 dan O_3 ≤ 102,0 dan NO_2 > 134,5	TIDAK SEHAT
8	50,5 < PM2.5 ≤ 100,5 dan O_3 > 102,0	TIDAK SEHAT
9	PM2.5 > 100,5 dan O_3 ≤ 12,5	TIDAK SEHAT
10	PM2.5 > 100,5 dan O_3 > 12,5	TIDAK SEHAT

3.5 Hasil Evaluasi Model

Dari 425 data uji, model mengklasifikasikan 416 data dengan benar sehingga akurasinya 416/425 = 97,88%. Sembilan data diklasifikasikan salah: 3 data BAIK diprediksi SEDANG, 4 data SEDANG diprediksi BAIK, dan 2 data TIDAK SEHAT diprediksi SEDANG (Gambar 3). Tidak ada kesalahan antara kelas BAIK dan TIDAK SEHAT. Hasil perhitungan manual dari *confusion matrix* sama dengan keluaran program. Pemeriksaan sepuluh sampel representatif data uji juga menunjukkan seluruhnya diprediksi sesuai kelas aktual.



Gambar 3. *Confusion matrix* pada data uji

Kelas TIDAK SEHAT memperoleh kinerja terbaik (*precision* 100%, *recall* 98,71%, *F1-score* 99,35%), diikuti kelas SEDANG dengan *F1-score* 98,27%. Kelas BAIK memperoleh *F1-score* terendah, yaitu 69,57% (Tabel 4). Nilai *macro-average* dihitung dari ketiga kelas dan lebih rendah daripada akurasi karena memberi bobot yang sama pada kelas BAIK.

Tabel 4. Nilai evaluasi setiap kategori ISPU pada data uji

Kelas	TP	FP	FN	*Precision* (%)	*Recall* (%)	*F1-score* (%)	*Support*
BAIK	8	4	3	66,67	72,73	69,57	11
SEDANG	255	5	4	98,08	98,46	98,27	259
TIDAK SEHAT	153	0	2	100,00	98,71	99,35	155
Macro-average				88,25	89,96	89,06	425

3.6 Pengaruh Penyeimbangan Kelas dan Perbandingan Hasil

Untuk mengurangi pengaruh ketidakseimbangan kelas, data latih diseimbangkan dengan SMOTE dan ADASYN, lalu hasilnya dibandingkan dengan model tanpa penyeimbangan (Tabel 5). Model tanpa penyeimbangan memperoleh akurasi 97,88%, sedangkan SMOTE dan ADASYN menghasilkan 95,06% dan 95,53%.

Tabel 5. Perbandingan akurasi

Metode	Akurasi (%)	Keterangan
C4.5 tanpa penyeimbangan kelas	97,88	Penelitian ini (model utama)
C4.5 + ADASYN	95,53	Penelitian ini
C4.5 + SMOTE	95,06	Penelitian ini
C4.5 (RapidMiner), Saputra & Firmansyah (2025)	89,52	*Dataset* dan prapemrosesan berbeda

Pembahasan

Akurasi 97,88% menunjukkan bahwa batas antarkategori ISPU pada data ini dapat dipisahkan dengan baik oleh ambang konsentrasi polutan. Dominasi PM2.5 (information gain 0,8717 dan feature importance 0,9048) masuk akal karena ISPU ditentukan dari polutan dominan (Fenny & Herniwanti, 2025), dan pada data ini atribut critical paling sering bernilai PM25 (modus). Artinya, PM2.5 paling sering menjadi polutan penentu kategori udara di DKI Jakarta, sehingga PM2.5 layak menjadi prioritas pemantauan. Polutan lain berperan sebagai pemisah lanjutan pada subkelompok data, sedangkan CO tidak dipakai sebagai pemisah pada pohon yang dipangkas. Kinerja tinggi pada kelas SEDANG dan TIDAK SEHAT didukung jumlah data yang besar (864 dan 515 data). Kelas BAIK hanya memiliki 36 data (25 data latih dan 11 data uji), sehingga model memiliki lebih sedikit contoh untuk mempelajari polanya. Akibatnya, precision BAIK hanya 66,67% (4 data SEDANG diprediksi BAIK) dan recall 72,73% (3 data BAIK diprediksi SEDANG). Dengan 11 data uji, selisih satu data saja mengubah recall BAIK sekitar 9 poin persen, sehingga nilai kelas ini sebaiknya dibaca sebagai indikasi, bukan estimasi yang stabil. Seluruh kesalahan terjadi antara kategori yang bersebelahan (BAIK–SEDANG dan SEDANG–TIDAK SEHAT), yang sejalan dengan kemiripan konsentrasi polutan di dekat batas kategori. Akurasi total juga cenderung menutupi kelemahan pada kelas minoritas: macro F1-score sebesar 89,06% lebih rendah daripada akurasi 97,88%. Aturan hasil ekstraksi juga perlu ditafsirkan dengan hati-hati. Aturan 2 dan 4 (Tabel 3) berasal dari simpul daun yang kecil (4 dan 10 data latih).

Aturan 4, yang mengategorikan $SO_2 > 66{,}0$ sebagai BAIK, mencerminkan pola pada sampel kecil dan tidak berarti SO_2 yang tinggi menyebabkan udara baik. Akurasi penelitian ini (97,88%) lebih tinggi daripada akurasi 89,52% yang dilaporkan Saputra dan Firmansyah (2025) untuk C4.5. Perbandingan ini tidak sepenuhnya setara karena perbedaan dataset, wilayah, dan prapemrosesan, sehingga angkanya dipakai sebagai gambaran, bukan bukti keunggulan metode. Temuan bahwa C4.5 menghasilkan aturan yang mudah ditafsirkan sejalan dengan Sang et al. (2021) yang membandingkannya dengan SVM, dan identifikasi polutan dominan sejalan dengan Prafanto et al. (2024). Penelitian ini menambahkan ukuran kontribusi polutan secara kuantitatif, pemangkasan pohon, dan 10 aturan eksplisit pada data periode 2021–2025, yang lebih baru daripada data tahun 2023 pada Kannajmi dan Saputra (2025). Beberapa faktor memengaruhi hasil. Pertama, kualitas data: sekitar 14% nilai PM10 hilang dan diisi dengan median, yang dapat mengurangi variasi data. Dampaknya pada hasil kemungkinan kecil karena PM10 bukan penentu utama, tetapi hal ini tidak diuji terpisah. Kedua, ketidakseimbangan kelas, yang membuat kelas BAIK sulit dipelajari. Ketiga, penyeimbangan kelas: SMOTE dan ADASYN menurunkan akurasi total menjadi 95,06% dan 95,53%. Dampaknya pada precision dan recall tiap kelas tidak dilaporkan, sehingga belum dapat disimpulkan apakah kelas BAIK terbantu. Keempat,

pemangkasan, yang menyederhanakan pohon dari 39 menjadi 19 simpul dan membuat aturan lebih mudah dibaca. Evaluasi dalam artikel ini dilakukan pada model hasil pemangkasan, dan perbandingan metrik sebelum dan sesudah pemangkasan tidak disajikan. Kelebihan penelitian ini adalah modelnya transparan (10 aturan IF-THEN), ringan sehingga dapat dijalankan pada laptop dengan memori 4 GB, penanganan datanya terdokumentasi, serta hasil perhitungan manual yang diverifikasi dengan keluaran program.

Keterbatasannya adalah (1) kelas BAIK sangat sedikit sehingga evaluasinya tidak stabil; (2) evaluasi memakai satu kali pembagian 70:30 tanpa validasi silang; (3) data berasal dari satu wilayah dan periode sehingga generalisasi ke daerah lain belum teruji; (4) imputasi median dan modus dilakukan sebelum pembagian data sehingga statistiknya dihitung dari seluruh data, dan penelitian lanjutan sebaiknya menghitungnya hanya dari data latih; (5) belum ada perbandingan dengan algoritma lain maupun penyetelan hiperparameter; dan (6) parameter meteorologi tidak dilibatkan. Secara praktis, aturan IF-THEN dapat ditanamkan pada dasbor atau sistem peringatan dini untuk memperkirakan kategori ISPU dari konsentrasi polutan, misalnya ketika data sensor tidak lengkap, dan dapat menjadi dasar prioritas pemantauan PM2.5. Secara akademik, hasil ini menjadi acuan awal (baseline) bagi penelitian yang membandingkan C4.5 dengan metode machine learning lain (Novryadi et al., 2025; Putri et al., 2025) pada data kualitas udara DKI Jakarta.

4. Kesimpulan

Algoritma *Decision Tree* C4.5 berhasil membangun model klasifikasi kategori ISPU berdasarkan enam parameter polutan pada 1.415 data DKI Jakarta periode 2021–2025. PM2.5 menjadi simpul akar dengan *information gain* 0,8717 dan *feature importance* 0,9048. Pemangkasan mengurangi pohon dari 39 menjadi 19 simpul (kedalaman 9 menjadi 5 level) dan menghasilkan 10 aturan *IF-THEN*. Pada 425 data uji, model mengklasifikasikan 416 data dengan benar (akurasi 97,88%), dengan *F1-score* 99,35% untuk TIDAK SEHAT, 98,27% untuk SEDANG, dan 69,57% untuk BAIK. Dengan demikian, ketiga tujuan penelitian tercapai. Kelemahan penelitian ini adalah jumlah data kelas BAIK yang sangat sedikit, penggunaan satu kali pembagian data, dan cakupan satu wilayah. Penelitian selanjutnya dapat memperbesar data dan cakupan wilayah, menambahkan parameter meteorologi, memakai validasi silang, membandingkan C4.5 dengan *Random Forest*, *Naïve Bayes*, SVM, dan XGBoost (Kurniawan et al., 2025; Novryadi et al., 2025), serta mengembangkan hasilnya menjadi aplikasi web atau seluler.

5. Pernyataan Etik

Penelitian ini tidak melibatkan subjek manusia maupun hewan sehingga tidak memerlukan persetujuan etik (*ethical clearance*). Data yang digunakan adalah data sekunder kualitas udara yang diperoleh dari Kaggle.

6. Kontribusi Peneliti

Galang Putra Wicaksana: konseptualisasi, metodologi, perangkat lunak, analisis formal, investigasi, kurasi data, visualisasi, dan penulisan draf awal. Irwansyah: supervisi, validasi, peninjauan, dan penyuntingan naskah. Ahmad Taufik: peninjauan dan penyuntingan naskah. Faldy Irwiensyah: peninjauan dan penyuntingan naskah.

7. Pendanaan

Penelitian ini tidak menerima pendanaan khusus dari lembaga pemberi dana mana pun, baik di sektor publik, komersial, maupun nirlaba.

8. Pernyataan Konflik Kepentingan

Peneliti menyatakan tidak terdapat konflik kepentingan terkait penelitian, kepenulisan, dan/atau publikasi artikel ini.

9. Pernyataan Penggunaan AI Generatif

Peneliti menggunakan Claude (Anthropic), perangkat AI generatif, untuk membantu menyusun format sesuai templat jurnal. Data, hasil analisis, dan angka bersumber dari skripsi peneliti. Peneliti telah meninjau seluruh isi naskah dan bertanggung jawab penuh atasnya.

Daftar Pustaka

Ahmad, J., Donny, F., Toto, S., & Putra, D. S. (2025). Analisis kualitas udara ambien di Kota Padang: Konsentrasi polutan, distribusi spasial, dan implikasi kebijakan. *JTPVI: Jurnal Teknologi dan Pendidikan Vokasi Indonesia, 3*(3). https://doi.org/10.24036/jtpvi.v3i3.328

Akbar, I., Supriadi, F., & Junaedi, D. I. (2024). Pemanfaatan *machine learning* di bidang kesehatan. *JATI (Jurnal Mahasiswa Teknik Informatika), 9*.

Alfian, H., Wahyuni, S., Revalino, A., Mirano, M. F., Rahmayana, E., & Mukhtar, H. (2024). Teknik *machine learning* untuk analisa klasifikasi kualitas udara: *A review*. *Journal of Software Engineering and Information System (SEIS), 4*(2). https://doi.org/10.37859/seis.v4i2.7617

Aushofi, M., Irwansyah, I., & Shidqon, M. (2025). Implementasi *data mining* untuk menganalisis pola penimbangan sampah menggunakan algoritma Apriori. *DIGINTEL-AI: DIGital INnovation and inTELligence – AI, 1*(1), 1–10. https://doi.org/10.66217/digintel-ai.v1i1.1

Daniswara, A. A. A., & Nuryana, I. K. D. (2023). *Data preprocessing* pola pada penilaian mahasiswa Program Profesi Guru. *Journal of Informatics and Computer Science (JINACS), 5*. https://doi.org/10.26740/jinacs.v5n01.p97-100

Farisi, I., Shadiq, J., Priyadi, W., Maulana, D., Acep, A., & Gusril, S. F. (2024). Penerapan model *recurrent neural network* (RNN) untuk prediksi curah hujan berbasis data historis. *Information System for Educators and Professionals: Journal of Information System, 9*. https://doi.org/10.51211/isbi.v9i2.3280

Fenny, M. R., & Herniwanti. (2025). Analisis kejadian indeks standar pencemaran udara (ISPU) sebagai informasi mutu udara di Kota Pekanbaru. *Jurnal Kesehatan dan Pengelolaan Lingkungan, 6*, 1–11. https://doi.org/10.12928/jkpl.v6i1.12504

Firmansyah, F., Irwansyah, I., & Budiyantara, A. (2025). Analisis pola pembelian konsumen di Rumah Makan Tepi Laut Baubau menggunakan algoritma Apriori. *DIGINTEL-AI: DIGital INnovation and inTELligence – AI, 1*(1), 25–36. https://doi.org/10.66217/digintel-ai.v1i1.3

Hanum, A. R., Zetha, I. A., Putri, S. C., Wulandari, R. A., Andina, S. P., Fajrina, J. N., & Yudistira, N. (2024). Analisis kinerja algoritma klasifikasi teks BERT dalam mendeteksi berita hoaks. *Jurnal Teknologi Informasi dan Ilmu Komputer, 11*(3), 537–546. https://doi.org/10.25126/jtiik.938093

Harpad, B., Salmon, & Saputra, R. M. (2022). Sistem monitoring kualitas udara di kawasan NodeMCU ESP32 berbasis IoT. *Jurnal Informatika Wicida, 12*(2), 39–47. https://doi.org/10.46984/inf-wcd.1955

Iqbal, M., Ali, I., Nurhakim, B., & Marta, P. P. (2026). Penerapan algoritma *decision tree* untuk klasifikasi persetujuan pinjaman sebagai pendukung pengambilan keputusan risiko kredit. *Journal of Computer Science and Artificial Intelligence (JCSAI), 3*. https://ruangjurnal.or.id/index.php/jcsai/article/view/295

Irwansyah, I., Wiranata, A. D., & TT, M. (2023). Komparasi algoritma *decision tree*, *naive bayes* dan *K-nearest neighbor* untuk menentukan kualitas udara di Provinsi DKI Jakarta. *Infotech: Journal of Technology Information, 9*(2), 193–198.

Islam, H. I., Mulyadien, M. K., & Enri, U. (2022). Penerapan algoritma C4.5 dalam klasifikasi status gizi balita. *Jurnal Ilmiah Wahana Pendidikan, 8*(10), 116–125.

Istiana, N., & Mustafiril, A. (2023). Perbandingan metode klasifikasi pada data dengan *imbalance class* dan *missing value*. *Jurnal Informatika, 10*, 101–108. https://doi.org/10.31294/inf.v10i2.15540

Kahfi, M., Umbara, F. R., & Ashaury, H. (2024). Prediksi pengangguran menggunakan *decision tree* dengan algoritma C5.0 pada data penduduk Kecamatan Caringin Kabupaten Bogor. *Informatics and Digital Expert (INDEX), 4*, 75–80. https://doi.org/10.36423/index.v4i2.913

Kannajmi, A. R., & Saputra, D. (2025). Penentuan model algoritma klasifikasi terbaik untuk klasifikasi kualitas udara di Jakarta 2023. *Jurnal Informatika dan Teknik Elektro Terapan, 13*(1).

Kurniawan, D. I., Irwansyah, I., & Taufik, A. (2025). Analisis sentimen terhadap komentar video IShowSpeed Tour Indonesia pada YouTube menggunakan metode SVM. *DIGINTEL-AI: DIGital INnovation and inTELligence – AI, 1*(1), 50–62. https://doi.org/10.66217/digintel-ai.v1i1.5

Lumbanbatu, N. A., Maha, Y. L., & Sipayung, S. P. (2026). Penerapan metode *decision tree* untuk memprediksi dampak kesehatan akibat penggunaan gadget. *RIGGS: Journal of Artificial Intelligence and Digital Business, 4*(4). https://doi.org/10.31004/riggs.v4i4.5787

Maharani, S., & Aryanta, W. R. (2023). Dampak buruk polusi udara bagi kesehatan dan cara meminimalkan risikonya. *Jurnal Ecocentrism, 3*(2), 47–58. https://doi.org/10.36733/jeco.v3i2.7035

Najah, A. N., Amalina, A. N., & Hidayat, R. (2025). Penerapan algoritma *decision tree* C4.5 untuk prediksi cuaca di Kota Semarang. *Indexia, 7*(1), 45–52.

Nazifah, N., & Prianto, C. (2023). Analisis perbandingan *decision tree* algoritma C4.5 dengan algoritma lainnya: *Systematic literature review*. *Jurnal Informatika dan Teknologi Komputer (J-ICOM), 4*, 57–64. https://doi.org/10.55377/j-icom.v4i2.7719

Novryadi, A. R., Irwansyah, I., & Shidqon, M. (2025). Analisis sentimen masyarakat terhadap kinerja Presiden Indonesia Joko Widodo periode kedua menggunakan metode *Naïve Bayes* dan SVM. *DIGINTEL-AI: DIGital INnovation and inTELligence – AI, 1*(1), 11–24. https://doi.org/10.66217/digintel-ai.v1i1.2

Prafanto, A., Astuti, I. F., Salamah, U., Agus, F., Kridalaksana, A. H., & Kamila, V. Z. (2024). Air pollution assessment of Samarinda using the C4.5 algorithm. *Buletin Poltanesa, 24*(2), 235–241. https://doi.org/10.51967/tanesa.v24i2.2946

Prasetiyo, A. (2024). Pemanfaatan algoritma *decision tree* C4.5 dalam memprakirakan hujan di Stasiun Meteorologi Kelas I Sultan Hasanuddin. *Buletin GAW Bariri (BGB), 5*(1), 47–55.

Putri, S. B. P., Irwansyah, I., & Tri M., T. (2025). Implementasi algoritma K-NN pada sosial media X untuk analisis sentimen pengalaman warganet tinggal di luar negeri. *DIGINTEL-AI: DIGital INnovation and inTELligence – AI, 1*(1), 37–49. https://doi.org/10.66217/digintel-ai.v1i1.4

Ragga, W. A., Naufal, I. F., Fayza, A. R., Juwita, P., & Heri, K. (2022). Klasifikasi data kualitas udara Kota Tangerang Selatan menggunakan algoritma C4.5. *Jurnal Sistem Informasi*, 1–11.

Romli, I., & Zy, A. T. (2020). Penentuan jadwal *overtime* dengan klasifikasi data karyawan menggunakan algoritma C4.5. *J-SAKTI (Jurnal Sains Komputer dan Informatika), 4*(2), 694–702.

Sang, A. I., Sutoyo, E., & Darmawan, I. (2021). Analisis *data mining* untuk klasifikasi data kualitas udara DKI Jakarta menggunakan algoritma *decision tree* dan *support vector machine*. *e-Proceedings of Engineering, 8*(5), 8954–8963.

Saputra, A. D. M., & Firmansyah, H. (2025). Implementasi algoritma *decision tree* untuk memprediksi kualitas udara dan polusi dengan RapidMiner. *Sudo Jurnal Teknik Informatika, 4*(3), 212–219.

Syahputri, C. N., & Hasibuan, M. S. (2024). Optimasi klasifikasi *decision tree* dengan teknik *pruning* untuk mengurangi *overfitting*. *JSiI (Jurnal Sistem Informasi), 11*, 87–96. https://doi.org/10.30656/jsii.v11i2.9161

Tangkelayuk, A., & Mailoa, E. (2022). Klasifikasi kualitas air menggunakan metode KNN, *Naïve Bayes* dan *decision tree*. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi), 9*. https://doi.org/10.35957/jatisi.v9i2.2048

Valerian, F. R., Syarief, M., & Fatah, D. A. (2025). Klasifikasi tingkat obesitas menggunakan metode GBM dan *confusion matrix*. *JATI (Jurnal Mahasiswa Teknik Informatika), 9*. https://doi.org/10.36040/jati.v9i2.13062