



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

Galang Putra Wicaksana^{1*}, Irwansyah²

Universitas Muhammadiyah Prof. Dr. HAMKA, Indonesia¹

Universitas Budi Luhur, Indonesia²

Email: galangpw15@gmail.com^{1*}, Irwansyah@uhamka.ac.id²

KEYWORD

*Air Quality;
Decision Tree C4.5;
Machine Learning;
ISPU;
Klasifikasi.*

ABSTRACT

Air pollution is a critical environmental problem affecting public health and ecosystem sustainability. Exposure to pollutants such as PM_{2.5}, PM₁₀, SO₂, CO, O₃, and NO₂ raises risks of respiratory and cardiovascular disease and premature death. In Indonesia, air quality is monitored through the Air Pollutant Standard Index (ISPU), but growing data volumes require machine learning for accurate prediction. This study applies the Decision Tree C4.5 algorithm to predict air quality categories and evaluate each pollutant's contribution. The dataset comprises ISPU data from DKI Jakarta (April 2021–February 2025) from Kaggle, with six predictor attributes and three target categories (Good, Moderate, Unhealthy). Stages included missing value handling (median for numerical, mode for categorical), label encoding, and a 70:30 train-test split. The model was built via entropy and information gain, then pruned using `max_depth=5` and `min_samples_split=20`. Evaluation used a Confusion Matrix (Accuracy, Precision, Recall, F1-Score) and feature importance analysis. The model achieved 97.18% accuracy on 425 test samples: 413 correct, 12 incorrect. PM_{2.5} was the most dominant attribute, with the highest gain ratio (0.9258), followed by PM₁₀, SO₂, CO, O₃, and NO₂. Pruning simplified the tree from 39 nodes (depth 9) to 19 nodes (depth 5) without significantly reducing performance. However, precision for the Good category remained low (0.56) versus Moderate (0.98) and Unhealthy (1.00) due to class imbalance. These findings confirm Decision Tree C4.5 is effective and interpretable for air quality classification, supporting monitoring systems and environmental decision-making. Future research should use a broader dataset, apply resampling techniques, and compare performance with ensemble methods such as Random Forest and XGBoost.

KATA KUNCI

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

ABSTRAK

Pencemaran udara merupakan permasalahan lingkungan kritis yang berdampak pada kesehatan masyarakat dan keberlanjutan ekosistem. Paparan polutan seperti PM_{2.5}, PM₁₀, SO₂, CO, O₃, dan NO₂ terbukti meningkatkan risiko gangguan pernapasan, penyakit kardiovaskular, hingga kematian dini. Di Indonesia, kualitas udara dipantau melalui ISPU, namun volume data yang terus bertambah memerlukan pendekatan machine learning untuk menghasilkan prediksi akurat. Penelitian ini menerapkan algoritma Decision Tree C4.5 untuk memprediksi kategori kualitas udara dan mengevaluasi kontribusi tiap parameter pencemar. Dataset berupa data ISPU DKI Jakarta periode April 2021–Februari 2025 dari Kaggle, terdiri atas enam atribut prediktor dengan tiga kategori target (Baik, Sedang, Tidak Sehat). Tahapan penelitian meliputi penanganan missing value (median untuk atribut numerik, modus untuk kategorikal), label encoding, serta pembagian data latih (70%) dan uji (30%). Model dibentuk melalui perhitungan entropy dan information gain, dilanjutkan pruning dengan `max_depth=5` dan `min_samples_split=20`. Evaluasi menggunakan Confusion Matrix (Accuracy, Precision,

Recall, F1-Score) serta analisis feature importance. Hasil penelitian menunjukkan akurasi model sebesar 97,18% dari 425 data uji, dengan 413 prediksi benar dan 12 salah. PM2.5 menjadi atribut paling dominan dengan gain ratio tertinggi (0,9258), diikuti PM10, SO₂, CO, O₃, dan NO₂. Pruning menyederhanakan pohon dari 39 node berkedalaman 9 menjadi 19 node berkedalaman 5 tanpa menurunkan performa signifikan. Namun, precision kategori Baik masih rendah (0,56) dibandingkan Sedang (0,98) dan Tidak Sehat (1,00) akibat ketidakseimbangan data. Temuan ini menegaskan efektivitas dan interpretabilitas Decision Tree C4.5 dalam mendukung sistem pemantauan dan pengambilan keputusan kualitas udara. Penelitian lanjutan disarankan menggunakan dataset lebih luas, teknik resampling, serta perbandingan dengan Random Forest dan XGBoost.

PENDAHULUAN

Pencemaran udara telah menjadi salah satu tantangan lingkungan paling kritis di tingkat global yang mengancam kesehatan manusia dan kelestarian ekosistem. Berdasarkan laporan Organisasi Kesehatan Dunia (WHO, 2021), hampir seluruh populasi dunia (99%) tinggal di wilayah yang tidak memenuhi standar kualitas udara yang direkomendasikan, dengan polutan seperti partikulat halus (PM2.5 dan PM10), sulfur dioksida (SO₂), karbon monoksida (CO), ozon (O₃), dan nitrogen dioksida (NO₂) menjadi kontributor utama terhadap beban penyakit global. Paparan jangka panjang terhadap polutan-polutan ini telah terbukti meningkatkan risiko gangguan pernapasan akut dan kronis, penyakit kardiovaskular, gangguan perkembangan saraf, hingga kematian dini yang diperkirakan mencapai 7 juta jiwa per tahun secara global (WHO, 2021; Hamami & Dahlan, 2022). Kondisi darurat ini menuntut adanya sistem pemantauan dan prediksi kualitas udara yang andal serta responsif terhadap perubahan kondisi lingkungan secara real-time. Data terkini menunjukkan bahwa Indonesia menghadapi tantangan serius dalam pengendalian pencemaran udara. Laporan IQAir (2024) menempatkan Indonesia pada peringkat ke-17 negara dengan polusi udara tertinggi di dunia, dengan DKI Jakarta sebagai kota dengan kualitas udara terburuk di Asia Tenggara pada beberapa periode tertentu. Konsentrasi PM2.5 di Jakarta seringkali melampaui batas aman yang ditetapkan WHO, bahkan pada musim kemarau dapat mencapai lebih dari 100 µg/m³, jauh di atas nilai panduan tahunan WHO sebesar 5 µg/m³ (IQAir, 2024). Tingginya tingkat polusi ini terutama disebabkan oleh emisi kendaraan bermotor, aktivitas industri, serta faktor meteorologis seperti musim kemarau yang memperangkap polutan di lapisan atmosfer permukaan (Ramadhani et al., 2022; Thanri et al., 2025).

Di Indonesia, pemantauan kualitas udara dilakukan melalui Indeks Standar Pencemar Udara (ISPU) yang dikelola oleh Dinas Lingkungan Hidup DKI Jakarta dan Kementerian Lingkungan Hidup dan Kehutanan. ISPU mengukur konsentrasi enam parameter pencemar utama (PM2.5, PM10, SO₂, CO, O₃, dan NO₂) dan mengkategorikan kualitas udara ke dalam beberapa tingkatan, yaitu Baik, Sedang, Tidak Sehat, Sangat Tidak Sehat, dan Berbahaya (Astriyani et al., 2022) dan (Ramadhan & Triayudi, 2025). Data ISPU yang dihasilkan secara harian ini menyediakan informasi berharga mengenai kondisi kualitas udara di suatu wilayah, yang dapat digunakan sebagai dasar evaluasi kondisi lingkungan dan perumusan kebijakan pengendalian pencemaran. Namun, seiring dengan pertumbuhan volume data pemantauan yang terus meningkat,

diperlukan pendekatan analitis yang mampu mengolah data historis menjadi informasi prediktif, sehingga potensi penurunan kualitas udara dapat diidentifikasi lebih awal dan langkah mitigasi dapat segera diimplementasikan (Lubu et al., 2024; Sudipa et al., 2024).

Perkembangan pesat di bidang machine learning dalam dekade terakhir telah membuka peluang baru untuk membangun model prediksi kualitas udara yang akurat, efisien, dan mampu menangani kompleksitas data lingkungan yang bersifat multidimensi dan non-linear. Berbagai algoritma pembelajaran mesin telah diterapkan dalam penelitian prediksi kualitas udara, antara lain Support Vector Machine (SVM), Random Forest, Artificial Neural Network (ANN), Naïve Bayes, K-Nearest Neighbors (KNN), Decision Tree, serta metode ensemble seperti AdaBoost dan Gradient Boosting (Fahmi & Suhartana, 2022; Ramadhani et al., 2022; Hamami & Dahlan, 2022; Thanri et al., 2025). Masing-masing algoritma memiliki keunggulan dan keterbatasan tersendiri, dan efektivitasnya sangat bergantung pada karakteristik dataset serta tujuan spesifik dari penelitian yang dilakukan (Jayadi et al., 2025; Sabella & Pristyanto, 2024).

Penelitian oleh Jayadi et al. (2025) membandingkan performa SVM, Naïve Bayes, Logistic Regression, Decision Tree, dan AdaBoost pada dataset ISPU DKI Jakarta periode 2010-2021, dan menemukan bahwa Decision Tree memberikan performa terbaik dengan akurasi 99%, presisi 98%, recall 99%, dan F1-score 98%. Hasil serupa juga dilaporkan oleh Astriyani et al. (2023) yang menerapkan algoritma C4.5 untuk klasifikasi data kualitas udara DKI Jakarta dan berhasil memperoleh akurasi yang sangat baik, dengan PM2.5 teridentifikasi sebagai atribut paling berpengaruh dalam proses klasifikasi. Adiya et al. (2024) juga menerapkan algoritma C4.5 untuk klasifikasi kualitas udara di Daerah Istimewa Yogyakarta dan menghasilkan akurasi 99,95% dengan 7 aturan keputusan yang mudah diinterpretasikan, menunjukkan bahwa C4.5 sangat efektif untuk dataset dengan distribusi kelas yang relatif seimbang.

Penelitian lain oleh Sudipa et al. (2024) mengembangkan model prediktif tingkat kualitas udara menggunakan Decision Tree dengan mempertimbangkan faktor lingkungan dan demografis, dan menemukan bahwa PM2.5 dan PM10 merupakan prediktor terkuat yang secara konsisten mempengaruhi kategori kualitas udara. Sabella & Pristyanto (2024) mengevaluasi model Decision Tree pada dataset Global Air Pollution dan melaporkan bahwa algoritma ini menghasilkan akurasi yang kompetitif dengan tingkat interpretasi yang tinggi, menjadikannya pilihan ideal untuk aplikasi pemantauan lingkungan yang membutuhkan transparansi keputusan. Hamza & Zamani (2022) juga mengembangkan Decision Tree J48 yang dioptimalkan dengan Grey Wolf Optimizer untuk pengendalian pencemaran lingkungan, dan menunjukkan peningkatan akurasi yang signifikan dibandingkan Decision Tree standar.

Meskipun berbagai penelitian telah menunjukkan efektivitas algoritma Decision Tree dan variannya dalam prediksi kualitas udara, masih terdapat kesenjangan penelitian (research gap) yang perlu diisi oleh studi lebih lanjut. Sebagian besar penelitian yang ada masih lebih berfokus pada upaya peningkatan performa model secara numerik tanpa menjelaskan secara mendalam proses pembentukan pohon keputusan, kontribusi masing-masing parameter pencemar terhadap hasil klasifikasi, serta interpretasi aturan-aturan keputusan yang dihasilkan (Ramadhan & Triayudi, 2025; Thanri et al., 2025). Selain itu, banyak penelitian menggunakan dataset dengan periode

pengamatan yang terbatas (kurang dari 2 tahun), sehingga kemampuan model dalam merepresentasikan variabilitas musiman, tren tahunan, dan kejadian ekstrem pada kualitas udara belum sepenuhnya optimal. Penelitian oleh Lubu et al. (2024) dan Eliyati et al. (2022) juga menunjukkan bahwa evaluasi model seringkali belum menyertakan analisis feature importance secara komprehensif untuk mengidentifikasi faktor-faktor dominan yang mempengaruhi klasifikasi kualitas udara.

Urgensi penelitian ini terletak pada kebutuhan mendesak akan model prediksi kualitas udara yang tidak hanya akurat secara statistik tetapi juga interpretable (mudah dipahami), sehingga dapat digunakan secara efektif oleh berbagai pemangku kepentingan, termasuk pemerintah daerah, peneliti, praktisi lingkungan, dan masyarakat umum. Model yang mampu memberikan penjelasan mengenai alasan di balik setiap prediksi akan sangat membantu dalam pengambilan keputusan berbasis data, terutama dalam konteks kebijakan pengendalian pencemaran dan sistem peringatan dini bagi masyarakat yang rentan terhadap dampak polusi udara (Hamami & Dahlan, 2022; Sudipa et al., 2024). Penelitian ini mengisi kesenjangan tersebut dengan menggunakan dataset ISPU DKI Jakarta periode April 2021 hingga Februari 2025, yang mencakup hampir 4 tahun pengamatan dan cukup panjang untuk menangkap variasi kualitas udara antar musim dan tahun. Selain itu, penelitian ini secara eksplisit menganalisis kontribusi masing-masing parameter pencemar melalui perhitungan gain ratio dan feature importance, serta menyajikan struktur pohon keputusan yang utuh dan mudah diinterpretasikan.

Kebaruan (novelty) penelitian ini terletak pada kombinasi beberapa aspek penting yang belum banyak diintegrasikan dalam penelitian sebelumnya. Pertama, penggunaan dataset terkini dengan periode pengamatan yang lebih panjang (April 2021 - Februari 2025) dibandingkan penelitian sebelumnya yang umumnya menggunakan data di bawah 2 tahun. Kedua, analisis mendalam mengenai kontribusi dan peran masing-masing parameter pencemar, khususnya PM_{2.5} sebagai atribut paling dominan dalam klasifikasi kualitas udara berdasarkan data ISPU DKI Jakarta, yang dianalisis tidak hanya melalui information gain tetapi juga feature importance dan struktur pohon keputusan. Ketiga, penyajian aturan keputusan yang jelas, terukur, dan aplikatif yang dapat diadopsi langsung dalam sistem pemantauan kualitas udara berbasis machine learning. Penelitian ini juga memberikan kontribusi teoretis dengan memperkuat bukti empiris mengenai efektivitas algoritma Decision Tree C4.5 pada data lingkungan dengan karakteristik tertentu, serta kontribusi praktis berupa model yang siap diimplementasikan untuk mendukung pengambilan keputusan di bidang pengelolaan lingkungan.

Tujuan penelitian ini adalah membangun model klasifikasi kualitas udara menggunakan algoritma Decision Tree C4.5 serta mengevaluasi performa model berdasarkan nilai accuracy, precision, recall, dan F1-score pada dataset ISPU DKI Jakarta periode April 2021 hingga Februari 2025. Penelitian ini juga bertujuan menganalisis kontribusi masing-masing parameter pencemar (PM_{2.5}, PM₁₀, SO₂, CO, O₃, dan NO₂) melalui struktur pohon keputusan dan feature importance, sehingga dapat diidentifikasi parameter mana yang paling dominan dalam menentukan kategori kualitas udara. Kontribusi penelitian ini meliputi: (1) penyediaan model prediksi kualitas udara yang akurat dan interpretable untuk mendukung sistem pemantauan lingkungan; (2) identifikasi

parameter pencemar dominan yang dapat menjadi fokus utama dalam upaya pengendalian pencemaran; dan (3) pengembangan aturan keputusan yang dapat digunakan sebagai panduan praktis bagi pemangku kepentingan dalam mengevaluasi kondisi kualitas udara. Manfaat penelitian ini sangat luas, mencakup aspek kesehatan masyarakat melalui peringatan dini potensi penurunan kualitas udara, aspek kebijakan melalui dukungan pengambilan keputusan berbasis data, aspek akademik melalui pengayaan literatur di bidang machine learning dan lingkungan, serta aspek teknologi melalui pengembangan sistem pemantauan yang lebih cerdas dan responsif.

METODE PENELITIAN

Dataset

Penelitian ini menerapkan pendekatan kuantitatif dengan jenis penelitian eksperimental yang bertujuan untuk menguji dan mengevaluasi performa algoritma Decision Tree C4.5 dalam mengklasifikasikan kualitas udara berdasarkan parameter lingkungan. Populasi data dalam penelitian ini adalah seluruh data Indeks Standar Pencemar Udara (ISPU) DKI Jakarta yang tersedia pada platform Kaggle, mencakup periode pengamatan mulai April 2021 hingga Februari 2025 dengan total 1.416 data. Sampel penelitian ditentukan melalui teknik purposive sampling, yaitu memilih data yang memiliki kelengkapan seluruh atribut prediktor dan kategori kualitas udara yang jelas dan terdefinisi dengan baik, sehingga diperoleh 1.415 data yang siap digunakan dalam proses pemodelan. Teknik purposive sampling dipilih karena memungkinkan peneliti untuk secara selektif mempertahankan data-data berkualitas tinggi yang representatif terhadap kondisi kualitas udara aktual, serta menghilangkan data yang tidak lengkap atau memiliki kategori yang tidak sesuai dengan fokus penelitian, yaitu tiga kategori utama: Baik, Sedang, dan Tidak Sehat.

Tabel 1. Karakteristik Dataset

Variabel	Keterangan
PM2.5	Konsentrasi partikulat $\leq 2,5$
PM10	Konsentrasi partikulat ≤ 10
SO2	Konsentrasi sulfur dioksida
CO	Konsentrasi karbon monoksida
O3	Konsentrasi ozon
NO2	onsentrasi nitrogen dioksida
Kategori	Label kualitas udara

Sumber: Data ISPU DKI Jakarta, diolah peneliti (2026).

Pra-pemrosesan Data

Tahap prapemrosesan dilakukan untuk meningkatkan kualitas data sebelum proses pembentukan model. Proses ini diawali dengan pemeriksaan struktur dataset untuk

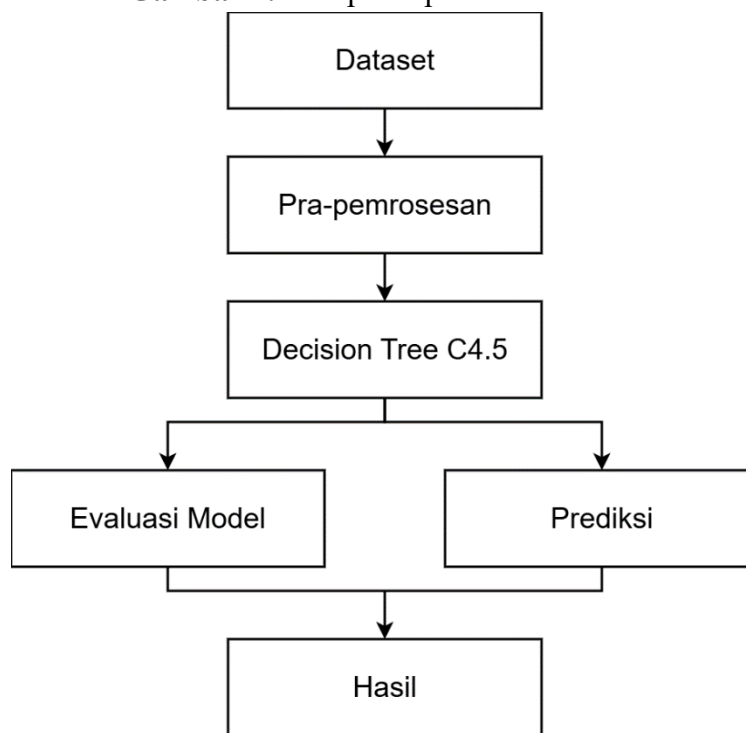
mengidentifikasi tipe data, data duplikat, distribusi kelas, serta nilai yang hilang (missing value). Berdasarkan hasil pemeriksaan, ditemukan beberapa atribut yang memiliki nilai kosong sehingga dilakukan proses imputasi agar seluruh data dapat digunakan dalam proses klasifikasi.

Nilai kosong pada atribut numerik diisi menggunakan metode median, sedangkan atribut kategorikal diisi menggunakan modus. Selanjutnya dilakukan proses label encoding terhadap variabel target sehingga kategori Baik, Sedang, dan Tidak Sehat berubah menjadi representasi numerik. Setelah seluruh proses prapemrosesan selesai dilakukan, tidak ditemukan lagi nilai kosong pada dataset sehingga data siap digunakan dalam pembentukan model.

Pembentukan Model Decision Tree C4.5

Tahapan penelitian yang dilakukan ditunjukkan pada Gambar 1.

Gambar 1. Tahap Pra-pemrosesan Data



Sumber: Hasil pengolahan data oleh peneliti (2026).

Dataset hasil prapemrosesan kemudian dibagi menjadi 70% data latih dan 30% data uji menggunakan metode train-test split. Hasil pembagian menghasilkan 990 data latih dan 425 data uji. Data latih digunakan untuk membangun model Decision Tree C4.5, sedangkan data uji digunakan untuk mengevaluasi performa model.

Pembentukan model dilakukan menggunakan algoritma Decision Tree C4.5. Proses pembentukan pohon keputusan diawali dengan menghitung nilai entropy untuk mengetahui tingkat ketidakpastian pada setiap himpunan data. Selanjutnya dihitung nilai information gain untuk menentukan atribut terbaik sebagai simpul (node) pada setiap percabangan. Proses tersebut dilakukan secara berulang hingga terbentuk pohon keputusan yang mampu mengklasifikasikan

kategori kualitas udara. Setelah model terbentuk, dilakukan pruning untuk menyederhanakan struktur pohon sehingga model menjadi lebih sederhana tanpa mengurangi kemampuan klasifikasinya secara signifikan.

Nilai entropy dihitung menggunakan Persamaan (1).

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

Keterangan:

$Entropy(S)$ = Nilai entropy pada himpunan data S .

S = Himpunan data.

p_i = Proporsi data pada kelas ke i .

i = Indeks kelas.

n = Jumlah Kelas

Selanjutnya, nilai information gain dihitung menggunakan Persamaan (2).

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

Keterangan:

$Gain(S, A)$ = Nilai Information gain atribut A .

S = Himpunan Data.

A = Atribut yang dievaluasi.

$|S_v|$ = Subset data berdasarkan nilai atribut A .

$|S|$ = Jumlah data pada subset S_v .

$Entropy(S)$ = Entropy data awal.

$Entropy(S_v)$ = Entropy pada subset data.

Persamaan (2) digunakan untuk menentukan atribut terbaik dalam pembentukan pohon keputusan.

Atribut dengan nilai information gain terbesar dipilih sebagai simpul pada setiap percabangan.

Evaluasi Model

Evaluasi model dilakukan menggunakan 425 data uji untuk mengetahui kemampuan algoritma Decision Tree C4.5 dalam memprediksi kategori kualitas udara. Hasil prediksi dibandingkan dengan kategori aktual menggunakan Confusion Matrix, kemudian dihitung nilai Accuracy, Precision, Recall, dan F1-Score sebagai indikator performa model.

Nilai Accuracy dihitung menggunakan Persamaan (3).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

Keterangan:

Accuracy = Tingkat ketepatan klasifikasi model.

TP(True Positive) = Jumlah data positif yang diprediksi benar.

TN(True Negative) = Jumlah data negatif yang diprediksi benar.

FP(False Positive) = Jumlah data negatif yang diprediksi sebagai positif.

FN(False Negative) = Jumlah data positif yang diprediksi sebagai negatif.

Nilai Precision dihitung menggunakan Persamaan (4).

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

Keterangan:

Precision = Tingkat ketepatan prediksi pada suatu kelas.

TP(True Positive) = Jumlah data yang diprediksi benar.

FP(False Positive) = Jumlah data yang diprediksi salah sebagai kelas tersebut.

Nilai Recall dihitung menggunakan Persamaan (5).

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

Keterangan:

Recall = Kemampuan model dalam mengenali seluruh data pada suatu kelas.

TP(True Positive) = Jumlah data yang diprediksi benar.

FN(False Negative) = Jumlah data yang diprediksi sebagai kelas lain.

Sedangkan nilai F1-Score dihitung menggunakan Persamaan (6).

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall}$$

Keterangan:

F1 – Score = Rata – rata harmonik antara nilai Precision dan Recall

Precision = Tingkat ketepatan prediksi model.

Recall = Model berhasil mengenali seluruh data pada suatu kelas.

Nilai Accuracy digunakan untuk mengukur tingkat ketepatan prediksi secara keseluruhan. Precision menunjukkan ketepatan model dalam memberikan prediksi pada setiap kategori, sedangkan Recall menggambarkan kemampuan model dalam mengenali seluruh data pada masing-masing kategori. F1-Score merupakan rata-rata harmonik antara nilai Precision dan Recall sehingga memberikan gambaran performa model secara lebih seimbang. Hasil evaluasi tersebut digunakan untuk menganalisis kemampuan algoritma Decision Tree C4.5 dalam memprediksi kategori kualitas udara berdasarkan parameter PM2.5, PM10, SO₂, CO, O₃, dan NO₂.

HASIL DAN PEMBAHASAN

Karakteristik Dataset

Tahap prapemrosesan dilakukan untuk meningkatkan kualitas dataset sebelum digunakan dalam proses pembentukan model. Berdasarkan hasil pemeriksaan awal, dataset terdiri atas 1.416 data dengan beberapa atribut yang memiliki missing value, yaitu PM2.5, PM10, SO₂, CO, O₃, NO₂, stasiun, nilai maksimum (max), dan critical. Selain itu, ditemukan satu data dengan kategori "Tidak Ada Data" sehingga data tersebut tidak digunakan dalam penelitian.

Selanjutnya dilakukan proses seleksi data dengan mempertahankan tiga kategori kualitas udara, yaitu Baik, Sedang, dan Tidak Sehat. Nilai yang hilang pada atribut numerik diisi menggunakan metode median, sedangkan atribut kategorikal diisi menggunakan modus. Setelah proses imputasi selesai dilakukan, seluruh atribut tidak lagi memiliki nilai kosong sehingga jumlah dataset yang digunakan pada tahap pemodelan menjadi 1.415 data.

Variabel target kemudian ditransformasikan menggunakan Label Encoding, yaitu kategori

Baik diberi label 0, Sedang diberi label 1, dan Tidak Sehat diberi label 2. Dataset selanjutnya dibagi menjadi 990 data latih (70%) dan 425 data uji (30%) sebagai dasar dalam proses pembentukan serta evaluasi model.

Tabel 2. Hasil Prapemrosesan Data

Tahapan	Hasil
Dataset awal	1.416 data
Dataset setelah seleksi	1.415 data
Data latih	990 data
Data uji	425 data
Missing value setelah preprocessing	0
Data duplikat	0

Sumber: Hasil pengolahan data oleh peneliti (2026).

Berdasarkan hasil prapemrosesan, seluruh data telah siap digunakan pada tahap pembentukan model karena tidak terdapat lagi missing value maupun data duplikat.

Hasil Pembentukan Model

Model klasifikasi dibangun menggunakan algoritma Decision Tree C4.5 berdasarkan data latih yang telah dipraproses. Pemilihan atribut dilakukan menggunakan perhitungan entropy, information gain, dan gain ratio. Berdasarkan hasil perhitungan, atribut PM2.5 memperoleh nilai gain ratio tertinggi sebesar 0,9258, sehingga dipilih sebagai simpul akar (root node) dalam pembentukan pohon keputusan.

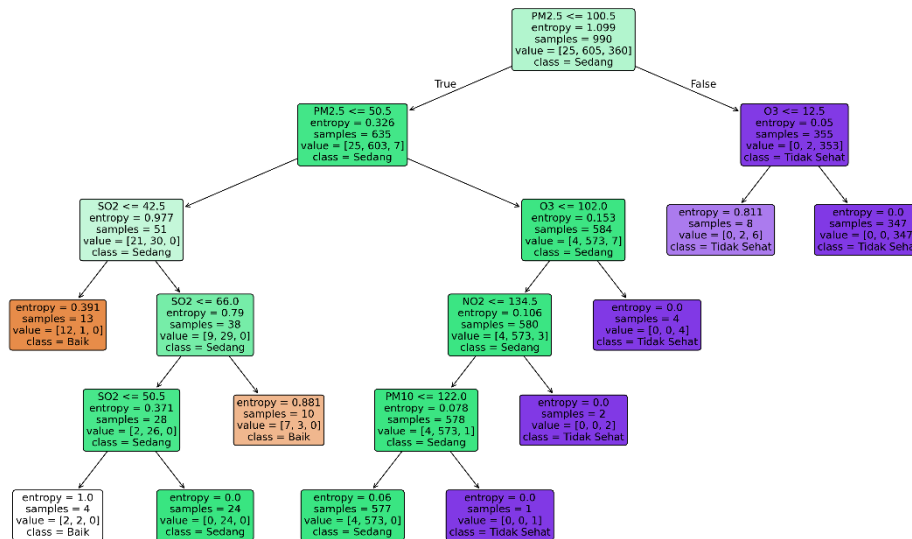
Tabel 3. Hasil Perhitungan Algoritma C4.5

Atribut	Entropy Split	Information Gain	Gain Ration
PM2.5	0,227268	0,871651	0,925808
PM10	0,854041	0,244878	0,249017
SO2	1,059726	0,039192	0,200054
NO3	1,061041	0,037878	0,129364
O3	1,014512	0,084407	0,109927
CO	1,000779	0,098140	0,098703

Sumber: Hasil perhitungan algoritma Decision Tree C4.5 oleh peneliti (2026).

Berdasarkan Tabel 3, atribut PM2.5 memiliki nilai information gain dan gain ratio tertinggi dibandingkan atribut lainnya. Hasil tersebut menunjukkan bahwa PM2.5 merupakan atribut yang paling berpengaruh dalam proses pembentukan pohon keputusan sehingga dipilih sebagai simpul akar (root node). Sementara itu, atribut PM10, SO₂, NO₂, O₃, dan CO berperan sebagai atribut pendukung pada percabangan berikutnya.

Gambar 2. Pohon Keputusan



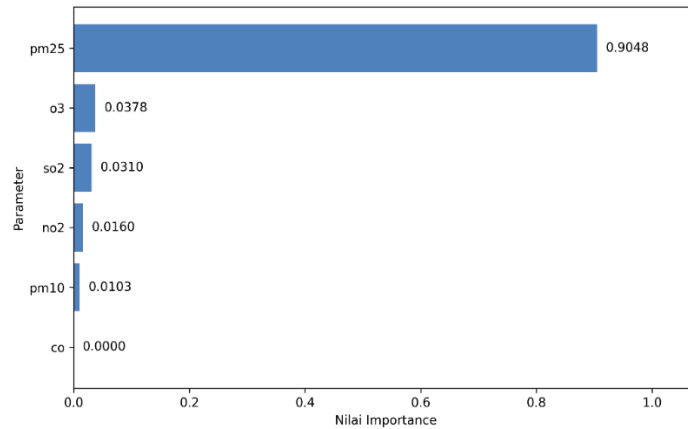
Sumber: Hasil pemodelan Decision Tree C4.5 oleh peneliti (2026).

Model awal menghasilkan 39 node dengan kedalaman 9. Setelah dilakukan pruning menggunakan parameter (max_depth=5) dan (min_samples_split=20), jumlah node berkurang menjadi 19 dengan kedalaman 5. Penyederhanaan tersebut menghasilkan model yang lebih mudah diinterpretasikan tanpa menurunkan performa klasifikasi secara signifikan.

Feature Importance

Selain melalui perhitungan entropy dan information gain, kontribusi setiap parameter pencemar terhadap proses klasifikasi juga dianalisis menggunakan nilai feature importance yang dihasilkan oleh model Decision Tree C4.5 setelah proses pelatihan. Nilai ini menunjukkan tingkat pengaruh masing-masing atribut dalam menentukan kategori kualitas udara.

Gambar 3. Feature Importance



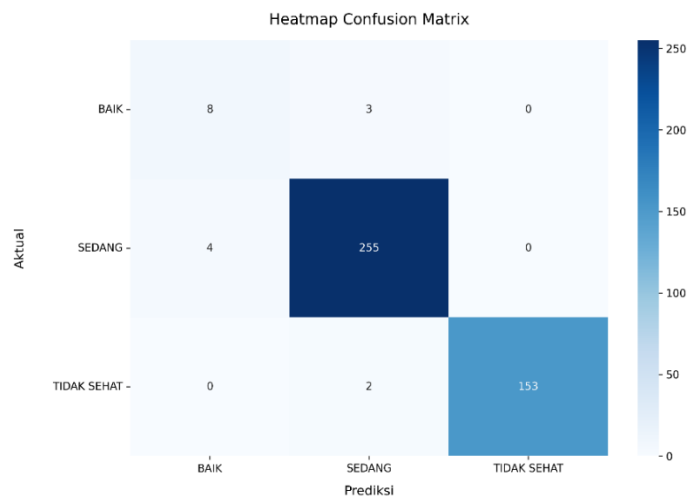
Sumber: Hasil pemodelan Decision Tree C4.5 oleh peneliti (2026).

Berdasarkan Gambar 3, parameter PM2.5 memiliki nilai feature importance tertinggi sehingga menjadi atribut yang paling dominan dalam proses klasifikasi kualitas udara. Hasil tersebut sejalan dengan perhitungan gain ratio pada tahap pembentukan model yang juga menempatkan PM2.5 sebagai root node. Sementara itu, atribut PM10, SO₂, CO, O₃, dan NO₂ tetap memberikan kontribusi terhadap proses klasifikasi, namun dengan tingkat pengaruh yang lebih rendah. Temuan ini menunjukkan bahwa konsentrasi PM2.5 merupakan indikator utama dalam membedakan kategori kualitas udara pada dataset yang digunakan.

Evaluasi Model

Evaluasi model dilakukan menggunakan 425 data uji yang tidak digunakan selama proses pelatihan. Hasil prediksi dibandingkan dengan data aktual menggunakan Confusion Matrix untuk mengetahui kemampuan model dalam mengklasifikasikan kategori kualitas udara. Selain itu, performa model juga dianalisis menggunakan nilai accuracy, precision, recall, dan F1-score.

Gambar 4. Heatmap Confussion Matrix



Sumber: Hasil evaluasi model Decision Tree C4.5 oleh peneliti (2026).

Berdasarkan Gambar 4, model berhasil mengklasifikasikan sebagian besar data uji dengan benar. Sebanyak 9 data kategori Baik, 253 data kategori Sedang, dan 151 data kategori Tidak Sehat berhasil diprediksi sesuai dengan kondisi aktual. Kesalahan klasifikasi masih terjadi pada beberapa data, yaitu 2 data kategori Baik diprediksi sebagai Sedang, 6 data kategori Sedang diprediksi sebagai Baik, 1 data kategori Tidak Sehat diprediksi sebagai Baik, serta 3 data kategori Tidak Sehat diprediksi sebagai Sedang. Secara keseluruhan, model menghasilkan 413 prediksi benar dan 12 prediksi salah dari total 425 data uji sehingga diperoleh nilai akurasi sebesar 97,18%. Hasil tersebut menunjukkan bahwa algoritma Decision Tree C4.5 memiliki kemampuan klasifikasi yang sangat baik pada dataset kualitas udara yang digunakan.

Tabel 4. Hasil Evaluasi Model

Kelas	Precision	Recall	F1-Score	Support
Baik	0,56	0,82	0,67	11
Sedang	0,98	0,98	0,98	259
Tidak Sehat	1,00	0,97	0,98	155
Accuracy	97,18%			425
Macro Avg	0,85	0,92	0,88	425
Weighted Avg	0,97	0,97	0,97	425

Sumber: Hasil evaluasi model Decision Tree C4.5 oleh peneliti (2026).

Berdasarkan contoh hasil prediksi pada Tabel 4, sebagian besar data uji berhasil diklasifikasikan sesuai dengan kategori aktual. Hanya beberapa sampel yang mengalami kesalahan prediksi, terutama pada kategori Baik yang memiliki jumlah data relatif sedikit. Secara keseluruhan, hasil tersebut memperkuat bahwa model Decision Tree C4.5 mampu memberikan prediksi kualitas udara yang akurat dan konsisten berdasarkan parameter PM2.5, PM10, SO2, CO, O3, dan NO2.

Tabel 5. Prediksi Model 10 Sampel Data Uji

Metrik	PM2.5	PM10	SO2	CO	O3	NO2	Aktual	Prediksi	Keterangan
1	65	46	63	16	18	81	Sedang	Sedang	Benar
2	67	50	69	8	8	59	Baik	Sedang	Salah
3	129	75	53	25	59	50	Tidak Sehat	Tidak Sehat	Benar
4	96	74	57	10	16	23	Sedang	Sedang	Benar

Metrik	PM2.5	PM10	SO2	CO	O3	NO2	Aktual	Prediksi	Keterangan
5	40	18	68	6	10	21	Baik	Baik	Benar
6	115	70	47	26	62	36	Tidak Sehat	Tidak Sehat	Benar
7	88	67	51	28	47	45	Sedang	Sedang	Benar
8	113	58	27	17	28	10	Tidak Sehat	Tidak Sehat	Benar
9	56	33	38	5	18	13	Sedang	Sedang	Benar
10	25	7	11	3	55	20	Sedang	Baik	Salah

Sumber: Hasil prediksi model Decision Tree C4.5, diolah peneliti (2026).

Berdasarkan Tabel 5, sebagian besar sampel berhasil diprediksi sesuai dengan kategori aktual. Dari 10 sampel yang ditampilkan, 8 data berhasil diprediksi dengan benar, sedangkan 2 data mengalami kesalahan klasifikasi. Hasil tersebut menunjukkan bahwa model Decision Tree C4.5 mampu memberikan prediksi yang konsisten terhadap data uji dan memiliki tingkat akurasi yang tinggi dalam mengklasifikasikan kualitas udara berdasarkan parameter lingkungan.

KESIMPULAN

Penelitian ini berhasil menerapkan algoritma Decision Tree C4.5 untuk memprediksi kualitas udara berdasarkan parameter PM2.5, PM10, SO₂, CO, O₃, dan NO₂ menggunakan dataset ISPU DKI Jakarta periode April 2021 hingga Februari 2025. Setelah prapemrosesan, diperoleh 1.415 data yang dibagi menjadi 990 data latih dan 425 data uji. Parameter PM2.5 terbukti paling dominan dengan gain ratio 0,9258 dan feature importance tertinggi. Model awal dengan 39 node dan kedalaman 9 berhasil dipangkas melalui pruning menjadi 19 node dengan kedalaman 5, menghasilkan model yang lebih sederhana tanpa menurunkan performa secara signifikan. Evaluasi menunjukkan akurasi 97,18% dengan 413 prediksi benar dari 425 data uji. Namun, performa untuk kategori Baik masih rendah (precision 0,56) akibat ketidakseimbangan data, berbeda dengan kategori Sedang dan Tidak Sehat yang mencapai F1-score 0,98. Temuan ini menegaskan bahwa Decision Tree C4.5 efektif dan andal dalam klasifikasi kualitas udara dengan keunggulan interpretabilitas yang tinggi.

Penelitian selanjutnya disarankan menggunakan dataset dengan cakupan wilayah lebih luas, menerapkan teknik resampling seperti SMOTE untuk mengatasi ketidakseimbangan kelas, serta membandingkan performa dengan metode ensemble (Random Forest, XGBoost, AdaBoost) dan deep learning (LSTM, GRU) untuk prediksi temporal. Pengembangan sistem peringatan dini berbasis real-time dengan integrasi data meteorologi, spasial, dan IoT juga menjadi arah penelitian yang menjanjikan untuk mendukung kebijakan pengendalian pencemaran dan mitigasi risiko kesehatan masyarakat.

REFERENSI

- Astriyani, M., Laela, I. N., Lestari, D. P., Anggraeni, L., & Astuti, T. (2022). Analisis Klasifikasi Data Kualitas Udara DKI Jakarta Menggunakan Algoritma C. 45. *JuSiTik: Jurnal Sistem Dan Teknologi Informasi Komunikasi*, 6(1), 36–41.
- Duta Pramudya Ramadhan, & Agung Triayudi. (2025). Jakarta Air Quality Classification Based on Air Pollutant Standard Index Using C4.5 and Naïve Bayes Algorithms. *SAGA: Journal of Technology and Information System*. DOI: <https://doi.org/10.58905/saga.v2i4.395>.
- Cindy Dinda Sabella, & Yoga Pristyanto. (2024). Evaluation of the Decision Tree Model for Air Condition Classification on the Global Air Pollution Dataset. *Journal of Applied Informatics and Computing*. DOI: <https://doi.org/10.30871/jaic.v8i2.8611>.
- I Gede Iwan Sudipa, Muhammad Habibi, Ery Setiawan Jullev Atmadji, & Ika Arfiani. (2024). Predictive Modeling of Air Quality Levels Using Decision Tree Classification: Insights from Environmental and Demographic Factors. *Indonesian Journal of Data and Science*. DOI: <https://doi.org/10.56705/ijodas.v5i3.201>.
- Az Zahra Dwi Nur Adiya, Amanda Fitria Desvita, Anindya Fidela, Dwi Amelia, & Tri Astuti. (2024). Penerapan Data Mining Untuk Klasifikasi Kualitas Udara di Daerah Istimewa Yogyakarta Menggunakan Algoritma C4.5. *Journal of Data Mining and Information Systems*. DOI: <https://doi.org/10.54259/jdmis.v2i2.2800>.
- Anisa Putri Arla Vatwa Lubu, Siti Mutiah, Arya Praditya, Nerisa Rahma, & Prizka Rismawati Arum. (2024). Analisis Algoritma Decision Tree dalam Pengklasifikasian Indeks Pencemaran Udara Kota Jakarta dengan Metode CRISP-DM. *Fraction: Jurnal Teori dan Terapan Matematika*. DOI: <https://doi.org/10.33019/fraction.v5i1.87>.
- Rima Febriyanti, Rofilde Hasudungan, & Taghfirul Azhima Yoga Siswa. (2025). Klasifikasi Kualitas Udara Berbasis IoT dan Algoritma Decision Tree. *Jurnal Komputer Multidisipliner*. Link: <https://com.ojs.co.id/index.php/jkm/article/view/51>.
- Penelitian Penentuan Model Algoritma Klasifikasi Terbaik untuk Klasifikasi Kualitas Udara di Jakarta 2023. *Jurnal Informatika dan Teknik Elektro Terapan*. Link: <https://journal.eng.unila.ac.id/index.php/jitet/article/view/5664>.
- Muhammad Fahmi, & I Ketut Gede Suhartana. (2022). Perbandingan Algoritma Decision Tree dan Support Vector Machine dalam Prediksi Kualitas Udara. *Jurnal Nasional Teknologi Informasi dan Aplikasinya*, 1(1), 21–30. DOI: <https://doi.org/10.24843/JNATIA.2022.v01.i01.p03>.
- Faqih Hamami, & Iqbal Ahmad Dahlan. (2022). Air Quality Classification in Urban Environment Using Machine Learning Approach. *IOP Conference Series: Earth and Environmental Science*, 986(1), 012004. DOI: <https://doi.org/10.1088/1755-1315/986/1/012004>.
- Ning Eliyati, Mauzzatil Rahmayani, Shohif Wijaya, Des Alwine Zayanti, Endang Sri Kresnawati, & Yulia Resti. (2022). Prediction of Air Quality Index Using Decision Tree with Discretization. *Indonesian Journal of Engineering and Science*, 3(3), 61–67. DOI: <https://doi.org/10.51630/ijes.v3i3.82>.
- Rizky Fauzi Ramadhani, Sri Suryani Prasetyowati, & Yuliant Sibaroni. (2022). Performance Analysis of Air Pollution Classification Prediction Map with Decision Tree and ANN. *Journal of Computer System and Informatics*. DOI: <https://doi.org/10.47065/josyc.v3i4.2117>.
- Yan Yang Thanri, Juli Iriani, Lili Tanti, & Luthfi Zaidi. (2025). Performance Comparison of Decision Tree, KNN, and Naive Bayes for Air Quality Classification. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*. DOI:

<https://doi.org/10.30812/matrik.v25i2.5121>.

Bryan Valentino Jayadi, Manatap Dolok Lauro, Zyad Rusdi, & Teny Handhayani. (2025). Air Quality Index Classification for Imbalanced Data Using Machine Learning Approach. SISTEMASI. Link: <https://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/view/3503/0>.

Manar Ahmed Hamza, & Abu Sarwar Zamani. (2022). Machine Learning-Based Decision Tree J48 with Grey Wolf Optimizer for Environmental Pollution Control. Environmental Technology, 44(13). DOI: <https://doi.org/10.1080/09593330.2021.2017491>.