sudo
Jurnal Teknik Informatika
https://jurnal.ilmubersama.com/index.php/sudo

Data Mining

Implementasi Algoritma *Decision Tree* untuk Memprediksi Kualitas Udara dan Polusi dengan *RapidMiner*

Abiestia Dika Mulya Saputra [*], *Hasbi Firmansyah*

Teknik dan Ilmu Komputer, Program Studi Informatika, Universitas Pancasakti Tegal, Tegal, Indonesia

INFORMASI ARTIKEL

Diterima Redaksi: 07 Agustus 2025
Revisi Akhir: 20 September 2025
Diterbitkan *Online*: 27 September 2025

KATA KUNCI

Decision Tree
RapidMiner
Kualitas udara
Pemodelan

KORESPONDENSI (*)

Phone: +62 852-2500-4924
E-mail: dikamulyasaputra45@gmail.com

A B S T R A K

Kualitas udara menjadi faktor utama yang dapat mempengaruhi kesehatan masyarakat dan lingkungan. Dalam studi ini, Data yang digunakan berisi berbagai variabel yang memiliki hubungan dengan kualitas udara dan polusi. Studi ini juga memiliki tujuan untuk mengembangkan model prediksi akurat metode Decision Tree di pilih karena diyakini dapat Mengelola data kompleks dan menghasilkan model baru yang mudah dipahami. *RapidMiner* digunakan dalam proses Analisis guna memfasilitasi visualisasi dan pemodelan data yang efisien. Hasil pengujian menunjukan bahwasannya model *Decision Tree* dapat mencapai tingkat akurasi yang cukup tinggi dalam memprediksi kualitas udara dan polusi, dengan tingkat nilai akurasi mencapai 89.52%. Tingkat akurasi ini tergolong tinggi karena mendekati 90%, menunjukkan bahwa model ini mampu melakukan klasifikasi dengan tingkat kesalahan yang relatif rendah. Oleh karena itu, model ini berpotensi diandalkan sebagai alat untuk pemantauan kualitas udara secara real-time maupun untuk merumuskan kebijakan lingkungan. Studi ini di harapkan mampu memberikan kontribusi atau keikutsertaan dalam upaya pemantauan dan pengendalian kualitas udara dan polusi, dan juga dapat menjadi acuan atau refrensi untuk studi yang lebih lanjut dalam bidang kualitas lingkungan.

PENDAHULUAN

Pencemaran udara dan polusi menjadi masalah lingkungan secara global, tidak hanya di Indonesia tetapi juga di negara-negara lain. Terdapat dampak negatif apabila kualitas udara dan polusi di suatu wilayah memiliki tingkat polutan yang buruk, seperti kesehatan manusia, ekosistem, dan pertumbuhan ekonomi[1]. Kualitas udara dan polusi yang buruk di suatu wilayah dapat berasal dari padatnya populasi masyarakat di suatu wilayah tersebut, aktivitas lalu lintas yang padat, serta adanya aktivitas industri. Sedangkan suatu wilayah dapat dikatakan memiliki kualitas udara dan polusi yang baik apabila wilayah tersebut memiliki tingkat polutan yang baik yakni di rentang angka 0-50[2]. Dalam data pengukururan kualitas udara dan polusi menggunakan beberapa parameter seperti suhu, kelembapan, tingkat partikel halus, tingkat partikel kasar, tingkat nitrogen dioksida, tingkat sulfur dioksida, tingkat karbon monoksida, kedekatan dengan kawasan industri, kepadatan penduduk[3]. Oleh karena itu, pemantauan kualitas udara dan polusi Sangat penting untuk melakukan hal ini untuk menjadi acuan dalam pengambilan kebijakan atau aturan-aturan yang tepat.

Dengan adannya permasalahan kondisi tersebut, penulis membuat studi mengenai implementasi algoritma *decision tree* untuk memprediksi kualitas udara dan polusi dengan *RapidMiner*. *RapidMiner* adalah alat analisis yang dapat digunakan untuk memproses data dengan cepat dan menampilkan hasil[4]. Dengan memeriksa nonlinieritas variabel masukan dan keluaran, peneliti dapat membuat prediksi kualitas udara dan polusi yang tepat menggunakan teknik ini, yang dimana penulis ingin mengembangkan model prediksi akurat metode *decision tree* di pilih untuk dapat mengelola data yang membuat dan memproduksi model baru membuat mudah untuk di pahami. Berdasarkan uraian tersebut, maka terdapat identifikasi masalah dari studi ini, yaitu memprediksi kategori kualitas udara dan polusi dan memprediksi tingkat akurasi pada algoritma *decision tree.*

https://doi.org/10.56211/sudo.v4i3.1145

TINJAUAN PUSTAKA

Kualitas Udara Dan Polusi

Kualitas udara dan polusi adalah topik yang sangat penting, terutama karena dampaknya terhadap kesehatan manusia dan lingkungan kita. Menurut data dari *World Health Organization* (WHO), Konsentrasi polusi udara yang lebih tinggi meningkatkan risiko terjadinya penyakit kardiovaskular dan pernapasan, kanker, serta hasil kelahiran yang buruk, dan juga berhubungan dengan angka kematian yang lebih tinggi [5]. Untuk mengukur kualitas udara, kita biasanya melihat beberapa parameter kunci, seperti *Particulate Matter* (PM10, PM2.5), *Nitrogen Dioksida* (NO_2), *Sulfur Dioksida* (SO_2), *Karbon Monoksida* (CO), dan *Ozon* (O_3). Oleh sebab itu, kualitas udara sangat penting dipantau untuk mengurangi risiko kesehatan dan membantu merumuskan kebijakan lingkungan yang lebih efektif[6].

Prediksi Kualitas Udara

Dalam beberapa tahun terakhir, teknologi untuk memprediksi kualitas udara telah Berkembang pesat sejalan dengan kemajuan dalam ilmu komputer[7]. Tujuan dari prediksi ini adalah untuk mengantisipasi polusi dan memberikan peringatan dini kepada masyarakat. Berbagai pendekatan telah digunakan, mulai dari metode statistik klasik hingga *machine learning.* Metode *machine learning*, khususnya, mampu mengolah data yang kompleks dan memberikan prediksi yang lebih akurat daripada metode konvensional[8].

Algoritma Decision Tree

*Decision Tree*merupakan salah satu algoritma pembelajaran mesin yang paling terkenal untuk klasifikasi dan prediksi. Algoritma ini bekerja dengan membagi dataset menjadi beberapa cabang berdasarkan fitur-fitur tertentu, sampai akhirnya mencapai keputusan yang berupa kelas prediksi[9]. Salah satu keunggulan dari *Decision Tree* adalah kemudahan dalam interpretasinya, serta kemampuannya untuk menjelaskan proses pengambilan keputusan. Berbagai penelitian telah menunjukkan bahwa *Decision Tree* sangat efektif dalam memprediksi kualitas udara dan parameter lingkungan lainnya[10].

RapidMiner untuk Data Mining

RapidMiner adalah salah satu platform yang sangat berguna untuk Data Mining dan Pemrosesan Analitis. Ia mendukung berbagai teknik pembelajaran mesin, termasuk Pohon Keputusan. Dengan *RapidMiner*, pengguna dapat dengan mudah melaksanakan seluruh proses penambangan data, mulai dari pra-pemrosesan data, pelatihan model, hingga evaluasi hasil prediksi. Banyak penelitian yang memanfaatkan *RapidMiner* sebagai alat utama dalam proyek-proyek prediksi lingkungan, berkat antarmukanya yang ramah pengguna dan kemampuan pemrosesan data yang luas[11].

METODOLOGI

Metode penelitian yang digunakan dalam studi ini dapat dijelaskan sebagai berikut:

Pengumpulan Data Dan Persiapan Data

Sumber Data

Data set ini berisi 5000 setdata tentang penilaian kualitas udara dan polusi, yang bersumber dari situs web kaggle.

Parameter Yang Tersedia

Data set ini memiliki 10 parameter yang di ukur , yaitu :

1. *Temperature*: yang berisi informasi suhu di berbagai wilayah.
2. *Humidity*: yang berisi informasi kelembapan relatif yang tercatat di wilayah tersebut.
3. *PM2.5 Concentration*: yang berisi informasi tingkat partikel halus.
4. *PM10 Concentration*: yang berisi informasi tingkat partikel kasar.
5. *NO2 Concentration*: yang berisi informasi tingkat nitrogen dioksida.
6. *SO2 Concentration* : yang berisi informasi tingkat sulfur dioksida.
7. *CO Concentration:* yang berisi informasi tingkat karbon monoksida.
8. *Proximity to Industrial Areas* (Kedekatan dengan Kawasan Industri): yang berisi informasi jarak ke kawasan industri terdekat.

9. *Population Density*: yang berisi informasi jumlah orang per kilometer persegi di wilayah tersebut.
10. *Air Quality*: yang berisi informasi kualitas udara.

Persiapan Data

1. Data Cleaning (Pembersihan Data)
   Langkah pembersihan dilakukan untuk menjaga kualitas dataset:
   a. Menghapus duplikasi baris yang muncul lebih dari sekali.
   b. Menangani missing value:
      i. Jika sebuah baris memiliki lebih dari 30% data kosong, baris tersebut dihapus.
      ii. Nilai kosong pada atribut numerik diisi dengan median (misalnya suhu, PM2.5, PM10).
      iii. Nilai kosong pada atribut kategorikal diisi dengan mode (misalnya label Air Quality).
   c. Memeriksa konsistensi data: format angka diseragamkan, kategori label distandarkan (Good, Moderate, Poor, Hazardous).
   d. Mengatasi outlier: nilai yang tidak wajar (misalnya kelembapan >100% atau nilai negatif pada PM) dihapus dari dataset.
2. Pemilihan Data (Feature Selection)
   Agar model lebih fokus, hanya fitur relevan yang digunakan:
   a. Menghapus fitur tidak penting, seperti ID data.
   b. Mempertahankan fitur utama: Suhu (Temperature), Kelembapan (Humidity), PM2.5, PM10, NO2, SO2, CO, Kedekatan dengan area industri (Proximity to Industrial Areas), populasi penduduk (Population Density).
   c. Menentukan label target: Air Quality, yang diklasifikasikan menjadi Good, Moderate, Poor, dan Hazardous.

Tabel 1. Data Set Penilaian kualitas udara Dan polusi

Temperature	Humidity	PM2.5	PM10	NO2	SO2	CO	Proximity_to_Industrial_Areas	Population_Density	Air Quality
29.8	59.1	5.2	17.9	18.9	9.2	1.72	6.3	319	Moderate
28.3	75.6	2.3	12.2	30.8	9.7	1.64	6	611	Moderate
23.1	74.7	26.7	33.8	24.4	12.6	1.63	5.2	619	Moderate
27.1	39.1	6.1	6.3	13.5	5.3	1.15	11.1	551	Good
26.5	70.7	6.9	16	21.9	5.6	1.01	12.7	303	Good
39.4	96.6	14.6	35.5	42.9	17.9	1.82	3.1	674	Hazardous
41.7	82.5	1.7	15.8	31.1	12.7	1.8	4.6	735	Poor
31	59.6	5	16.8	24.2	13.6	1.38	6.3	443	Moderate
29.4	93.8	10.3	22.7	45.1	11.8	2.03	5.4	486	Poor
33.2	80.5	11.1	24.4	32	15.3	1.69	4.9	535	Poor
26.3	65.7	1.3	5.5	18.3	5.9	0.85	13	529	Good
32.5	51.2	1.6	10.5	21.6	19.3	1.53	5.9	519	Moderate
20	53.3	3.7	12.9	26.1	6.6	1.09	10.2	538	Good
28.6	53.7	28.9	34	23.2	4.5	1.02	11	508	Good
22.3	80.5	4.5	12	17.2	6.3	1.18	10.4	232	Good
32	78.9	22.4	29.9	27.5	11.8	1.48	7.9	444	Moderate
22.9	75.4	4.5	10.4	18.4	3.7	0.96	14.4	359	Good
37.6	81.2	28.1	56.6	46.7	13.7	1.85	4.1	560	Poor
34.7	59.3	9	15.7	28.5	7.1	1.52	6.1	437	Moderate
37.8	97.2	0.6	24.6	37.1	11.7	1.13	7.7	803	Poor
27.6	44.1	3.5	14.4	30.7	9.4	0.97	8	338	Moderate
27.6	77.5	3.8	10.9	9.1	1.7	1.04	14.4	520	Good

 https://doi.org/10.56211/sudo.v4i3.1145

25.6	58.3	0.4	0.2	25.3	4.5	0.98	10	536	Good
24.6	48.4	8.3	15.4	23.3	4.6	1.03	11.2	461	Good

Tranformasi Data

Data seringkali memiliki format yang berbeda – beda. Fase transformasi ini mencakup normalisasi untuk menyesuaikan skala data agar konsisten, pengodean ulang variabel mengubah variabel kategorikal menjadi format numerik untuk analisis lebih lanjut, atau mengubah format data menjadi format konsisten yang memungkinkan analisis efektif.

Analisis Dan Pemodelan

Penelitian ini menggunakan *RapidMiner* untuk melakukan analisis data dan pemodelan. *RapidMiner* memfasilitasi visualisasi data dan membantu membuat model pohon keputusan.

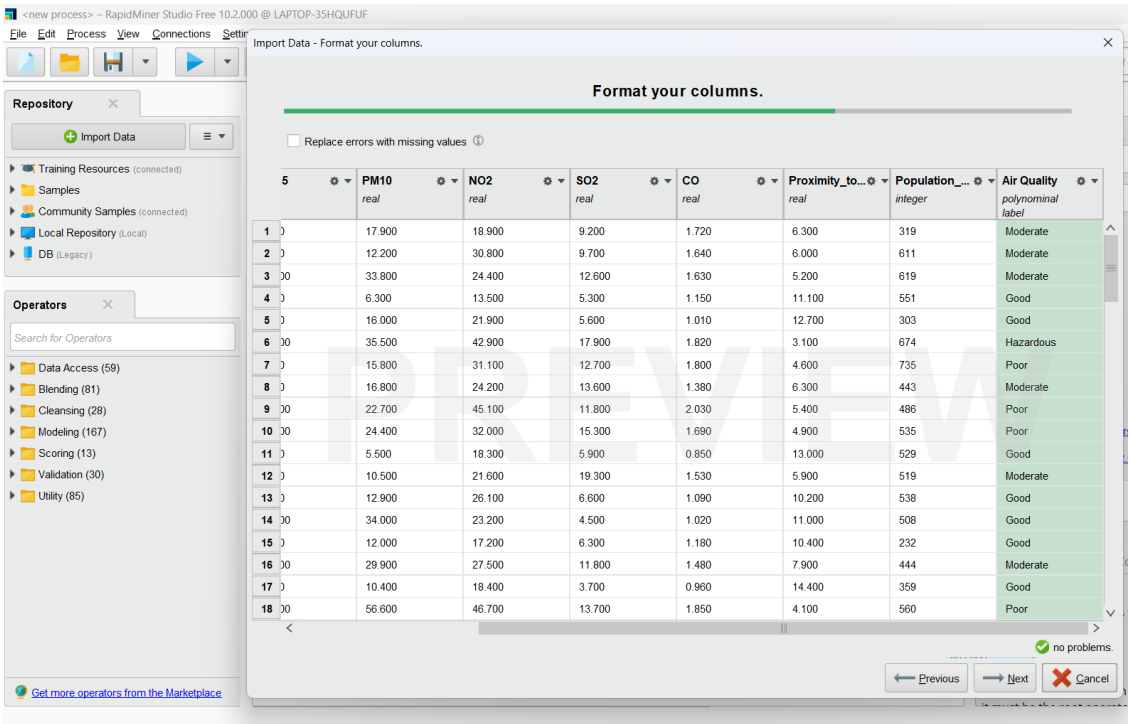
Evaluasi

Model *decision tree* di evaluasi berdasarkan tingkat akurasi prediksi yang di capai, dengan hasil menunjukkan akurasi sebesar 89.52%. Hal ini mengindikasikan bahwa algoritma *decision tree* mampu secara efektif memprediksi kualitas udara dan polusi.

HASIL DAN PEMBAHASAN

Pengolahan Data Dan Tranformasi Data

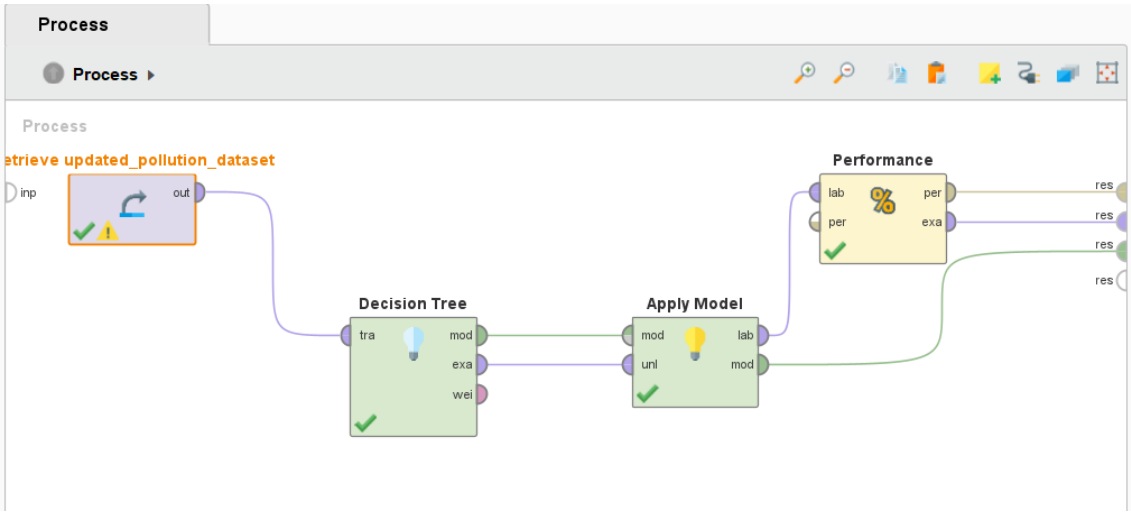
Pada gambar 1, di tampilkan langkah pertama dalam menginput dataset yang akan digunakan untuk menganalisis dalam sistem *RapidMiner.* Proses ini merupakan prasyarat penting untuk dapat melanjutkan ke tahap selanjutnya, yakni tahap pengolahan data. Selama tahap pengolahan data, semua atribut di gunakan untuk memprediksi kualitas udara dan polusi, sehingga tidak ada data yang di hapus selama tahap pengolahan data. Tranformasi atribut *Air Quality* sebagai label.



Gambar 1. Pengolahan Data Dan Tranformasi Data

Pemodelan

Setelah tahap pengolahan data selesai, langkah selanjutnya adalah tahap pemodelan. Pada tahap ini, algoritma *decision tree* (pohon keputusan) seringkali digunakan sebagai metode pemodelan dan alat bantu analisis pada aplikasi *RapidMiner.* Pada Gambar 2, di tampilkan metode pemodelan dari *preprosesing decision tree.*



Gambar 2. Pemodelan

1. *Decision Tree*: Metode yang digunakan untuk memperoleh informasi untuk tujuan pembuatan keputusan.
2. *Apply Model*: Model yang digunakan untuk pembelajaran mesin yang telah dilatih, guna memprediksi atau mengklasifikasikan data baru.
3. *Performance*: Untuk menilai seberapa sukses model pembelajaran mesin atau proses analisis data dalam mencapai tujuannya. Dalam penelitian ini, kinerja model yang dievaluasi diukur berdasarkan akurasi.

Menerapkan kumpulan data ke algoritma decision tree memerlukan serangkaian perhitungan yang mengelompokkan data. Algoritma decision tree yang diterapkan pada situasi ini adalah:
Perhitungan Gain:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} * \text{Entropy}(Si) \quad (1)$$

Keterangan:
S : himpunan
A : atribut
N : jumlah partisi atribut A
| Si | : jumlah kasus pada partisi ke-i
| S | : jumlah kasus dalam S

Menghitung Nilai Entropy:

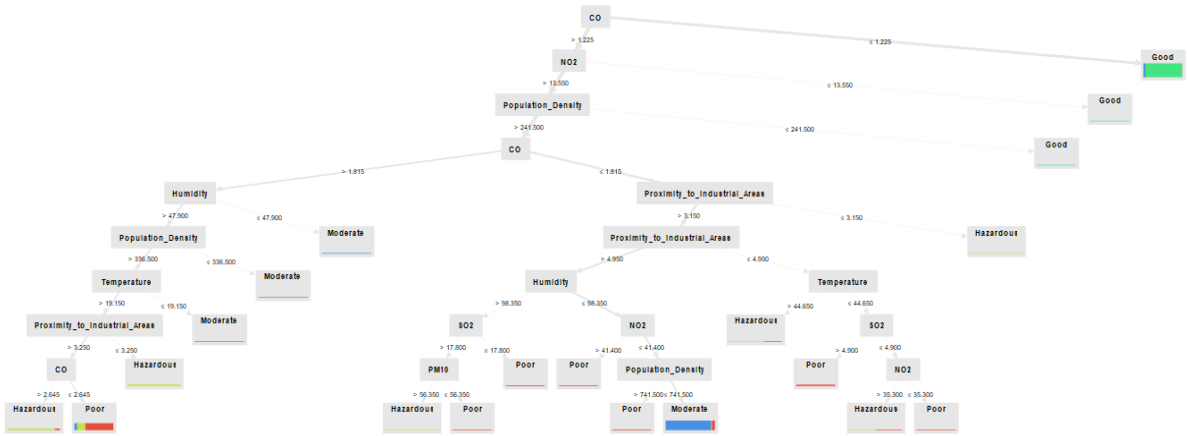
$$\text{Entropy}(S) - \sum_{i=1}^{n} - pi * \log_2 pi \quad (2)$$

Keterangan:
S : himpunan kasus
A : fitur
N : jumlah partisi S
Pi : proporsi dari Si terhadap S

 https://doi.org/10.56211/sudo.v4i3.1145

Evaluasi

Setelah melalui proses pengolahan dan pemodelan, langkah terakhir adalah menjalankan data menggunakan algoritma *decision tree* pada *platform RapidMiner*. Setelah menyelesaikan tahap ini, akan menghasilkan algoritma *decision tree* (pohon keputusan) seperti pada gambar 3, dan akurasi seperti tabel 2, yang menggambarkan kualitas dan polusi udara. tahap pengolahan data selesai, langkah selanjutnya adalah tahap pemodelan. Pada tahap ini, algoritma *decision tree* (pohon keputusan) seringkali digunakan sebagai metode pemodelan dan alat bantu analisis pada aplikasi *RapidMiner*[12]. Pada Gambar 2, di tampilkan metode pemodelan dari *preprosesing decision tree*.



Gambar 3. Decision Tree

Model yang dihasilkan dengan algoritma *decision tree* menunjukkan klasifikasi kualitas udara berdasarkan berbagai parameter lingkungan. Pada tahap awal, parameter utama yang diperiksa adalah CO Jika nilai CO rendah (≤ 1.225), kualitas udara langsung diklasifikasikan sebagai *Good*. Namun, jika nilainya lebih tinggi, analisis dilanjutkan ke parameter NO2. Nilai NO2 yang rendah (≤ 13.550) tetap menunjukkan kualitas udara *Good*, tetapi jika nilainya lebih tinggi, maka parameter *Population Density* menjadi penentu berikutnya. Kepadatan populasi yang rendah (≤ 241.500) tetap menghasilkan kualitas udara Good, sementara kepadatan yang tinggi memerlukan analisis lebih lanjut terhadap nilai CO. Jika nilai CO kedua lebih rendah (≤ 1.815), parameter *Proximity to Industrial Area* diperiksa. Dekatnya lokasi ke kawasan industri (≤ 3.150) menghasilkan kualitas *Moderate*, sedangkan jarak yang lebih jauh menghasilkan kualitas *Hazardous*. Sebaliknya, jika nilai CO lebih tinggi, parameter *Humidity* menjadi faktor berikutnya. Pada kelembaban rendah (≤ 98.350), analisis berlanjut ke SO2 dan PM10, di mana tingkat polusi yang tinggi pada parameter ini menghasilkan kualitas *Poor* atau *Hazardous*. Sementara itu, pada kelembaban tinggi (> 98.350), analisis dilanjutkan ke *NO2* dan *Population Density*, dengan tingkat polusi yang tinggi cenderung menghasilkan kualitas *Poor,* sedangkan tingkat polusi yang lebih rendah menghasilkan kualitas *Moderate*. Pohon keputusan ini memberikan pemahaman yang mendalam mengenai pengaruh berbagai parameter lingkungan terhadap kualitas udara. Pemilihan atribut CO (karbon monoksida) sebagai pemisah utama dalam pohon keputusan menunjukkan bahwa parameter ini memiliki kontribusi paling besar dalam menentukan kualitas udara. Secara teori, karbon monoksida (CO) dikenal sebagai salah satu polutan utama yang berbahaya bagi kesehatan. Gas ini dapat berikatan dengan hemoglobin dalam darah, yang mengurangi kemampuan darah untuk mengangkut oksigen. Ini sejalan dengan tinjauan pustaka yang menunjukkan bahwa konsentrasi CO yang tinggi sangat berkaitan dengan penurunan kualitas udara. Dengan demikian, hasil model mendukung teori yang ada dan memberikan implikasi ilmiah bahwa pengendalian emisi CO, misalnya dari sektor transportasi dan industri, menjadi kunci utama dalam menjaga kualitas udara. Selain itu, munculnya atribut lain seperti NO_2, kepadatan penduduk, dan kelembaban dalam level keputusan berikutnya juga memperkuat bahwa kualitas udara merupakan fenomena multifaktor, di mana faktor antropogenik (aktivitas manusia) dan faktor lingkungan saling memengaruhi.

Tabel 2. Hasil Akurasi *Confusion Matrix*

accuracy: 89.52%

	true Moderate	**true Good**	**true Hazardous**	**true Poor**	**Class precicision**
pred. Moderate	1314	21	0	86	92.47%
pred. Good	114	1979	1	5	94.28%
pred. Hazardous	0	0	290	16	94.77%
pred. Poor	72	0	209	893	76.06%
class recall	87.60%	98.95%	58.00%	89.30%	

Berdasarkan confusion matrix, terlihat adanya misklasifikasi pada beberapa kategori, khususnya pada kelas *Poor* dan *Moderate*. Hal ini dapat dijelaskan dengan adanya tumpang tindih karakteristik data antar kategori. Misalnya, kelas *Poor* sering terprediksi sebagai *Hazardous* karena keduanya sama-sama ditandai oleh tingginya konsentrasi polutan, sehingga batas antara keduanya sulit dipisahkan dengan jelas oleh model. Demikian pula, kasus *Moderate* yang terklasifikasi sebagai *Good* dapat terjadi karena nilai parameter lingkungannya berada sangat dekat dengan ambang batas kualitas udara yang digunakan dalam pohon keputusan. Fenomena ini sesuai dengan karakteristik algoritma decision tree, yang menggunakan ambang nilai (threshold) untuk memisahkan kelas. Jika distribusi data antar kelas saling berdekatan, maka kemungkinan misklasifikasi meningkat. Di sinilah keunggulan *RapidMiner* terlihat, karena selain memberikan hasil klasifikasi, ia juga memvisualisasikan jalur keputusan secara jelas, sehingga peneliti dapat mengidentifikasi titik-titik kritis di mana misklasifikasi terjadi. Temuan ini memperlihatkan bahwa meskipun akurasi keseluruhan tinggi (89,52%), optimasi model tetap diperlukan, misalnya dengan tuning parameter atau mengombinasikan decision tree dengan metode lain seperti Random Forest atau ensemble learning untuk meningkatkan ketepatan klasifikasi. Berdasarkan evaluasi *confusion matrix*, kinerja model klasifikasi menunjukkan tingkat akurasi sebesar 89.52%. Klasifikasi yang akurat dicapai pada 1314 sampel kategori *Moderate*, 1979 sampel *Good*, 290 sampel *Hazardous*, dan 893 sampel *Poor*. Analisis kesalahan prediksi mengungkapkan beberapa misklasifikasi penting, termasuk 114 kasus *Moderate* yang terkategorikan sebagai *Good*, 209 kasus *Poor* yang teridentifikasi sebagai *Hazardous,* serta 86 kasus *Moderate* yang terprediksi sebagai *Poor*. Meski model memperlihatkan kehandalan dalam deteksi kelas *Hazardous* dengan tingkat *false* positive minimal, optimasi lebih lanjut diperlukan untuk meningkatkan presisi klasifikasi pada kategori *Poor* dan *Moderate*.

KESIMPULAN DAN SARAN

Penerapan algoritma decision tree pada *RapidMiner* untuk klasifikasi kualitas udara menunjukkan bahwa metode ini efektif dalam menganalisis pengaruh berbagai parameter lingkungan terhadap kualitas udara. Dengan akurasi mencapai 89,52%, model berhasil melakukan klasifikasi yang baik pada kategori *Moderate, Good, Hazardous,* dan *Poor*. Pohon keputusan yang dihasilkan memberikan pemahaman mendalam tentang hubungan parameter seperti CO, NO2, kepadatan penduduk, kelembapan, dan jarak ke kawasan industri terhadap kualitas udara. Model ini juga menunjukkan performa yang sangat baik dalam mendeteksi kategori risiko tinggi *Hazardous* dengan tingkat kesalahan minimal. Namun, terdapat beberapa kesalahan klasifikasi yang perlu diperhatikan, khususnya pada kategori *Poor dan Moderate*, yang menunjukkan bahwa model masih memerlukan optimasi lebih lanjut. Langkah-langkah seperti penyempurnaan parameter algoritma atau penggunaan pendekatan pembelajaran yang lebih canggih dapat membantu meningkatkan akurasi. Secara keseluruhan, penelitian ini memberikan kontribusi penting dalam analisis kualitas udara menggunakan algoritma *decision tree.* Dalam penelitian selanjutnya sebaiknya fokus pada penerapan dan perbandingan efektivitas berbagai algoritma lain, seperti *Random Forest, Support Vector Machine,* atau metode *ensemble*. Tujuannya adalah untuk mendapatkan model prediksi yang lebih akurat dan dapat diandalkan. Selain itu, optimalisasi parameter pada algoritma *Decision Tree* juga sangat krusial untuk mengurangi kesalahan klasifikasi, terutama pada kategori yang masih menunjukkan tingkat misklasifikasi tinggi, seperti kelas *Poor* dan *Moderate*.

DAFTAR PUSTAKA

[1] Z. Fang, P. Wu, Y. Lin, T. Chang, dan Y. Chiu, “Air Pollution ’ s Impact on the Economic , Social , Medical , and Industrial Injury Environments in China,” 2021.

[2] J. I. Lingkungan, A. Pradifan, dan A. Suprihanto, "Pemantauan Kualitas Udara Kota Tegal (Studi Kasus : Kecamatan Tegal Selatan , Kecamatan Tegal Barat , Kecamatan Tegal Timur)," vol. 19, no. 1, hal. 73–82, 2021, doi: 10.14710/jil.19.1.73-82.

[3] B. Mutu, "Analisis Kualitas Udara (Nilai Parameter PM 2 , 5 dan Karbon Monoksida) di Sekitar Kampus Universitas Bosowa Makassar," vol. 23, no. April, hal. 164–171, 2023, doi: 10.35965/eco.v23i1.2514.

[4] A. S. Ramadhantya, "Penggunaan Rapidminer Untuk Memprediksi Kelulusan Mahasiswa Dengan Algorithm Naive Bayes," vol. 10, no. 1, hal. 52–60, 2024.

[5] WHO, "Air pollution." Diakses: 31 Juli 2025. [Daring]. Tersedia pada: https://www.who.int/teams/environment-climate-change-and-health/healthy-urban-environments/transport/health-risks

[6] F. Haya, K. Nisa, R. F. Ladipasa, A. Suriani, dan A. Media, "Dampak Polusi Udara terhadap Kesehatan Manusia," vol. 3, 2025.

[7] C. Rosca dan M. Carbureanu, "Data-Driven Approaches for Predicting and Forecasting Air Quality in Urban Areas," no. 2, 2025.

[8] B. Dutta, "Comparative Analysis of Machine Learning and Deep Learning Models for Lung Cancer Prediction Based on Symptomatic and Lifestyle Features," 2025.

[9] U. Putra dan I. Yptk, "Penerapan algoritma klasifikasi untuk prediksi tingkat kelulusan mahasiswa menggunakan rappidminer," vol. 6, no. 1, hal. 376–388, 2025, doi: 10.46576/djtechno.

[10] D. B. Olawade, O. Z. Wada, A. O. Ige, B. I. Egbewole, A. Olojo, dan B. I. Oladapo, "Artificial intelligence in environmental monitoring : Advancements , challenges , and future directions," *Hyg. Environ. Heal. Adv.*, vol. 12, no. November 2023, hal. 100114, 2024, doi: 10.1016/j.heha.2024.100114.

[11] J. Nasional, S. Informasi, V. Riandaru, H. Lazuardi, A. Adhi, dan C. Lauw, "Penerapan Aplikasi RapidMiner Untuk Prediksi Nilai Tukar Rupiah Terhadap US Dollar Dengan Metode Regresi Linier," vol. 01, hal. 8–17, 2021.

[12] S. Informasi, "Jurnal Advance Research Informatika PENERAPAN DATA MINING MENGGUNAKAN ALGORITMA DECISION TREE C4 . 5 UNTUK MEMPREDIKSI MAHASISWA DROP OUT DI UNIVERSITAS WIRARAJA," vol. 1, no. Juni, hal. 1–7, 2023.

NOMENKLATUR

Perhitungan Gain:

Gain (S, A) = Entropy (S) - ∑n i = 1 * Entropy (Si)

S artinya himpunan
A artinya atribut
N artinya jumlah partisi atribut A
| Si | artinya jumlah kasus pada partisi ke-i
| S | artinya jumlah kasus dalam S

Menghitung Nilai Entropy:

Entropy (S) - ∑an i = 1 – pi * log2 pi

S artinya himpunan kasus
A artinya fitur
N artinya jumlah partisi S
Pi artinya proporsi dari Si terhadap S