PENERAPAN METODE CRISP-DM UNTUK PERBANDINGAN ALGORITMA KLASIFIKASI INDEKS STANDAR PENCEMAR UDARA (ISPU) DI SPKU GONDOKUSUMAN YOGYAKARTA

Lydia Nur Sa'adah[1*], Heppy Nur Asavia Ginasputri[2], Yolan Triky[3], Fatkhurokhman Fauzi[4]
[1,2,3,4]Program Studi Statistika, Universitas Muhammadiyah Semarang, Kota Semarang, Jawa Tengah, Indonesia
**Email: lydianursaadah@gmail.com*

ABSTRAK

Kepadatan aktivitas di wilayah perkotaan seringkali menimbulkan peningkatan kadar polusi udara, yang berpotensi mengancam kesehatan masyarakat dan menurunkan kualitas lingkungan. Kondisi ini juga terjadi di Yogyakarta, di mana tren kenaikan konsentrasi polutan menuntut adanya sistem pemantauan kualitas udara yang lebih akurat dan responsif. Penelitian ini bertujuan mengembangkan model klasifikasi kualitas udara berbasis data mining menggunakan pendekatan CRISP-DM serta membandingkan performa empat algoritma klasifikasi, yaitu Decision Tree, Random Forest, K-Nearest Neighbor, dan Gradient-Boosted Trees. Dataset yang digunakan berjumlah 366 observasi yang berasal dari Stasiun Pemantauan Kualitas Udara Gondokusuman yang dikelola oleh Dinas Lingkungan Hidup Kota Yogyakarta, dengan parameter utama yaitu PM10, SO_2, CO, O_3, NO_2, dan nilai maksimum polutan (Max). Pembagian data dilakukan menggunakan skema cross-validation, dan hasil evaluasi menunjukkan bahwa algoritma Random Forest memberikan performa terbaik dalam mengklasifikasikan tingkat kualitas udara, dengan akurasi mencapai 99,73%, precision tertinggi sebesar 100% pada kategori "Good" dan "Unhealthy", serta recall tertinggi sebesar 100% pada kategori "Moderate" dan "Good". Temuan ini diharapkan dapat menjadi landasan dalam pengembangan sistem pemantauan kualitas udara serta untuk pengelolaan dan pengendalian pencemaran udara di wilayah perkotaan Yogyakarta.

Kata kunci: *CRISP-DM, Klasifikasi Kualitas Udara, Random Forest, ISPU*

1 PENDAHULUAN

Udara sebagai unsur utama kehidupan sering kali menjadi perhatian, terutama ketika dikaitkan dengan pencemaran udara yang turut menimbulkan kekhawatiran dalam kehidupan sehari-hari. Udara berperan sebagai salah satu zat utama penyusun atmosfer yang mengelilingi bumi, terdiri atas campuran berbagai gas yang tersebar dalam lapisan atmosfer (Andini et al., 2025). Udara bersih dan kering mengandung gas-gas dengan persentase rata-rata per volume, yaitu nitrogen sebesar 78,08%, oksigen 20,95%, argon 0,934%, karbon dioksida 0,03%, serta gas-gas lainnya sebanyak 0,27% (Syahira & Arianto, 2024). Namun kehadiran zat kimia, fisik, maupun biologis yang mencemari atmosfer dapat mengubah karakteristik udara, sehingga memicu kerusakan lingkungan serta menimbulkan gangguan kesehatan pada manusia. Aktivitas manusia yang semakin mengutamakan industrialisasi turut memperburuk kualitas udara di wilayah perkotaan, regional, hingga global, karena meningkatkan jumlah gas dan partikel berbahaya di atmosfer (Kumar & K S, 2024). Pencemaran udara dapat berdampak serius terhadap kesehatan manusia, seperti memicu infeksi saluran pernapasan, gangguan paru-paru, penyakit jantung,

bahkan kanker. Oleh karena itu, masyarakat yang tinggal di daerah perkotaan perlu meningkatkan kesadaran akan bahaya polusi udara demi menjaga kualitas hidup.

Di Indonesia, pencemaran udara menjadi salah satu permasalahan lingkungan yang mendesak, terutama di kota-kota besar seperti Yogyakarta. Meskipun memiliki luas wilayah hanya 32,5 kilometer persegi, Kota Yogyakarta menghadapi peningkatan pencemaran udara yang signifikan akibat kompleksitas infrastrukturnya. Sumber utama pencemaran di kota ini berasal dari dua jenis, yaitu sumber bergerak seperti pertumbuhan pesat kendaraan bermotor, serta sumber tidak bergerak berupa emisi gas buang dari pabrik (adminwarta, 2025). Untuk mengatasi permasalahan tersebut Kementerian Lingkungan Hidup dan Kehutanan (KLHK) berkomitmen menyediakan informasi yang akurat mengenai kualitas udara kepada masyarakat. Salah satu upayanya adalah dengan meningkatkan jumlah Stasiun Pemantauan Kualitas Udara Ambien (SPKUA) kontinu. Di wilayah DIY, pengukuran kualitas udara dilakukan oleh Dinas Lingkungan Hidup Kota Yogyakarta dan hasilnya dicatat dalam Indeks Standar Pencemar Udara (ISPU).

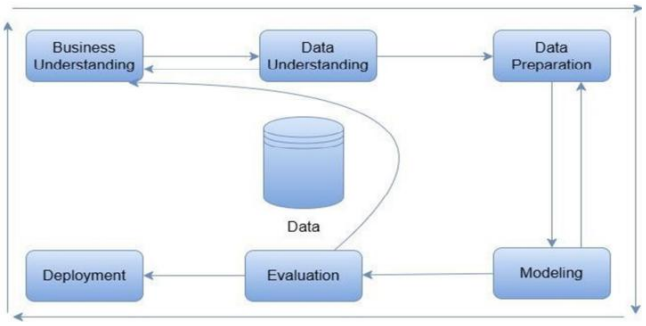
Di wilayah Daerah Istimewa Yogyakarta DIY, pengukuran kualitas udara dilakukan oleh Dinas Lingkungan Hidup Kota Yogyakarta, salah satunya melalui Stasiun Pemantau Udara Gondokusuman yang berlokasi di Kampus 3 Institut Sains & Teknologi AKPRIND Yogyakarta, Jalan Bimasakti, Demangan, Gondokusuman, Yogyakarta. Hasil pengukuran tersebut dicatat dalam Indeks Standar Pencemar Udara (ISPU). ISPU secara internasional dikenal sebagai Air Quality Index (AQI), digunakan untuk mengukur tingkat kualitas udara di suatu wilayah (Irjayana et al., 2025). Pengukuran ini didasarkan pada enam parameter, yaitu PM10, NO_2, SO_2, CO, O_3, dan Max. Nilai ISPU yang dihasilkan menjadi dasar dalam pengambilan kebijakan guna mengurangi tingkat pencemaran udara (Dasrul et al., 2020). Penelitian mengenai klasifikasi dan prediksi kualitas udara telah banyak dilakukan dengan menggunakan berbagai pendekatan algoritma data mining. Salah satu Penelitian oleh Yuliska & Syaliman (2020) menggunakan algoritma *K-Nearest Neighbor* (KNN) pada data ISPU Kota Pekanbaru. Hasil penelitian tersebut menunjukkan bahwa KNN mampu mencapai tingkat akurasi sebesar 97,09%, yang menunjukkan kinerja tinggi dalam klasifikasi kualitas udara berdasarkan parameter pencemar yang tersedia.

Penelitian lain oleh Kirono et al (2022) menggunakan algoritma *Naive Bayes* untuk mengklasifikasikan kualitas udara di DKI Jakarta dan memperoleh hasil akurasi 88%, *precision* 85%, *recall* 96%, dan f1-score 90%. Meskipun berbagai algoritma telah diterapkan dalam klasifikasi kualitas udara, masih terdapat keterbatasan pada model klasifikasi dalam hal akurasi, sensitivitas terhadap variasi data, serta kemampuan generalisasi, terutama pada wilayah dengan karakteristik pencemaran yang unik seperti Yogyakarta. Salah satu pendekatan yang menjanjikan untuk menangani data berskala besar dan variabel interaktif adalah *Random Forest*, namun belum banyak digunakan secara spesifik untuk data ISPU di Kota Yogyakarta. Oleh karena itu, penelitian ini berfokus pada perbandingan empat algoritma klasifikasi yaitu *Decision Tree*, *Random Forest, K-Nearest Neighbor,* dan *Gradient-Boosted Trees* dengan menerapkan kerangka kerja CRISP-DM (*Cross Industry Standard Process for Data Mining*) guna menghasilkan model prediksi yang akurat. Tujuannya adalah untuk mendukung pengambilan keputusan berbasis data dalam pengendalian pencemaran udara di Kota Yogyakarta secara lebih efektif dan ilmiah.

2 METODE

Dalam penelitian ini, terdapat tahapan-tahapan kerja yang dilakukan untuk memahami hubungan antara data dan atribut yang terdapat dalam dataset kualitas udara. Metode penelitian yang digunakan adalah CRISP-DM (*Cross Industry Standard Process for Data Mining*) dengan algoritma *Random Forest.* Metode ini menerapkan 6 tahapan metode CRISP-DM, yaitu (1)

Business Understanding; (2) *Data Understanding*; (3) *Data Preparation*; (4) *Modelling*; (5) Evaluation; (6) *Deployment*. Adapun 6 tahapan tersebut disajikan dalam flowchart pada Gambar 1. berikut.



Gambar 1. *Workflow* Metode CRISP-DM
Source: (Fitrianti et al., 2023)

Gambar 1 menampilkan tahapa atau alur kerja keseluruhan dari metode CRISP-DM yang digunakan dalam menganalisis data indeks standar pencemaran udara Kota Yogyakarta. Tujuan dari analisis ini adalah untuk untuk mengidentifikasi kondisi dan pola kualitas udara berdasarkan parameter pencemar. Hasil dari proses pengolahan dan optimalisasi data ini nantinya dapat dijadikan sebagai dasar rekomendasi dalam upaya pengendalian pencemaran udara di wilayah tersebut.

2.1 Business Understanding

Tahap ini dimulai dengan memahami masalah bisnis yang ingin diselesaikan atau tujuan yang ingin dicapai dengan proyek data mining.

2.2 Data Understanding

Pada tahap ini, data yang tersedia dieksplorasi dan dipahami secara mendalam. Hal ini mencakup pemeriksaan dan pemahaman data yang tersedia, struktur data, kualitas data, serta kemungkinan masalah.

2.3 Data Preparation

Pada tahap ini, data disiapkan untuk digunakan dalam proses data mining. Hal ini meliputi pemilihan atribut yang relevan, transformasi data, dan eliminasi data yang tidak relevan atau tidak valid.

2.4 Modelling

Tahap ini mencakup pemilihan model data mining yang sesuai untuk menyelesaikan masalah bisnis yang telah diidentifikasi. Sebelum pemodelan dilakukan, digunakan analisis *correlation matrix* untuk melihat hubungan antar atribut. Pada tahap klasifikasi dibandingkan empat algoritma yaitu : *Decision Tree, Random Forest, K-Nearest Neighbor,* dan *Gradient-Boosted Trees*. Untuk diketahui algoritma yang terbaik.

2.5 Evaluation

Tahap ini mengevaluasi semua tahapan yang telah dilakukan berdasarkan tujuan semula.

2.6 Deployment

Pada tahap ini, model data mining yang telah dikembangkan diimplementasikan ke dalam lingkungan produksi untuk digunakan dalam proses bisnis.

Melalui tahapan ini, struktur data dikenali, karakteristik setiap atribut dipahami, potensi masalah pada data dideteksi, serta pola-pola tersembunyi yang relevan untuk proses klasifikasi dapat diidentifikasi.

3 HASIL DAN PEMBAHASAN

3.1 Business Understanding

Pemahaman bisnis merupakan fase awal dalam tahapan CRISP-DM yang bertujuan untuk memahami permasalahan utama yang akan diselesaikan melalui data mining. Dalam penelitian ini, permasalahan yang dihadapi adalah meningkatnya pencemaran udara di Gondokusuman Kota Yogyakarta yang disebabkan oleh berbagai sumber, seperti pertumbuhan kendaraan bermotor dan emisi dari aktivitas industri. Permasalahan tersebut menuntut adanya sistem yang mampu mengklasifikasikan tingkat kualitas udara secara akurat berdasarkan parameter Indeks Standar Pencemar Udara (ISPU). Oleh karena itu, penelitian ini memanfaatkan metode perbandingan algoritma klasifikasi untuk membangun model prediktif.

3.2 Data Understanding

Tahapan pemahaman data dilakukan dengan mengumpulkan dan meninjau dataset kualitas udara harian pada Stasiun Pemantau Kualitas Udara (SPKU) Gondokusuman Kota Yogyakarta selama rentang waktu satu tahun dari Januari hingga Desember 2020 yang yang telah dikonversi ke Indeks Standar Pencemaran Udara (ISPU).

a) Pemahaman fitur (atribut)

Dataset diperoleh dari sumber resmi Dinas Lingkungan Hidup Kota Yogyakarta berupa file berekstensi .csv sebanyak 366 observasi. Satuan dari nilai kosentrasi adalah ($\mu g/m^3$) yang dapat diartikan banyaknya molekul yang tersebar disuatu daerah. Data kualitas udara di Stasiun Pemantau Kualitas Udara (SPKU) Gondokusuman Yogyakarta disajikan pada Gambar 2 berikut.

Row No.	PM10	SO2	CO	O3	NO2	Max	Category
1	20	1	0	?	?	20	Good
2	22	1	2	?	?	22	Good
3	10	0	2	11	?	11	Good
4	16	0	3	21	?	21	Good
5	13	1	5	11	?	13	Good
6	13	0	6	15	0	15	Good
7	10	0	6	12	2	12	Good
8	11	0	6	14	2	14	Good
9	18	1	7	7	3	18	Good
10	13	0	7	14	5	14	Good
11	13	0	7	18	6	18	Good
12	13	0	8	22	6	22	Good
13	12	0	8	16	6	16	Good
14	11	0	8	14	7	14	Good
15	23	2	9	0	7	23	Good
16	24	1	9	0	7	24	Good

ExampleSet (366 examples,0 special attributes,7 regular attributes)

Gambar 2 Dataset Kualitas Udara di SPKU Gondokusuman

Melalui tahapan ini, struktur data dikenali, karakteristik setiap atribut dipahami, potensi masalah pada data dideteksi, serta pola-pola tersembunyi yang relevan untuk proses klasifikasi dapat diidentifikasi. Dari dataset ISPU tersebut diperoleh beberapa atribut, antara lain:

a. PM_{10}

Particulate Matter 10 berisi konsentrasi partikel udara berukuran ≤10 mikron ($\mu g/m^3$). PM_{10} merupakan salah satu parameter utama dalam penentuan kualitas udara karena dapat masuk ke sistem pernapasan dan berdampak langsung terhadap kesehatan manusia.

b. SO_2

Berisi data kadar gas sulfur dioksida ($\mu g/m^3$) yang dihasilkan dari pembakaran bahan bakar fosil. Gas ini dapat menyebabkan iritasi saluran pernapasan, terutama pada penderita asma

dan penyakit paru-paru.

c. CO

Berisi data konsentrasi karbon monoksida (μg/m³), yaitu gas beracun yang dihasilkan dari pembakaran tidak sempurna. Paparan CO dalam kadar tinggi dapat mengganggu proses pengangkutan oksigen dalam tubuh.

d. O_3

Berisi data kadar ozon troposferik (μg/m³) yang terbentuk akibat reaksi kimia antara sinar matahari dan emisi kendaraan atau industri. Dalam konsentrasi tinggi, O_3 dapat menyebabkan gangguan pada saluran pernapasan.

e. NO_2

Berisi data kadar nitrogen dioksida (NO_2) dalam satuan μg/m³ yang dihasilkan dari proses pembakaran bahan bakar, terutama oleh kendaraan bermotor dan aktivitas industri. Dalam konsentrasi tinggi, NO_2 dapat menyebabkan iritasi pada saluran pernapasan dan memperburuk penyakit paru-paru seperti asma.

f. *Max*

Berisi nilai maksimum dari parameter pencemar pada waktu pengamatan tertentu. Nilai ini mencerminkan parameter dominan yang menentukan kategori kualitas udara (ISPU) saat itu.

g. *Category*

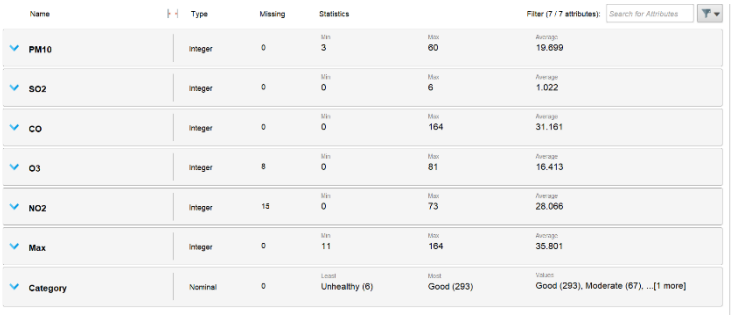
Berisi label kategori kualitas udara yang menjadi target klasifikasi. Berdasarkan peraturan yang telah ditetapkan oleh pemerintah (Peraturan Menteri LHK RI.P.14/MENLHK/SETJEN/KUM.1/7/2020), penentuan kategorinya adalah sebagai berikut:

Tabel 1. Indeks Standar Pencemaran Udara (ISPU)

ISPU	**Kategori**
0 - 50	Baik
51 –100	Sedang
101 - 200	Tidak Sehat
201 - 300	Sangat Tidak Sehat
≥300	Berbahaya

b) Deteksi *Missing Value*

Sebelum dilakukan proses pemodelan lebih lanjut, langkah penting yang harus dilakukan adalah mendeteksi apakah terdapat *missing values* pada atribut-atribut dalam dataset. *Missing value* dapat memengaruhi hasil dari proses klasifikasi, karena sebagian besar algoritma pembelajaran mesin memerlukan data yang lengkap untuk dapat bekerja secara optimal. Gambar 3 berikut memperlihatkan hasil deteksi nilai hilang yang dilakukan pada atribut-atribut dalam dataset menggunakan tool RapidMiner:



Gambar 3. *Descriptive Statistics* Dataset

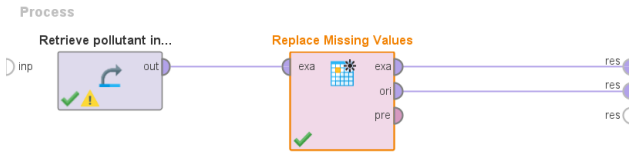
Missing Value pada atribut O_3 dan NO_2 perlu ditangani agar data dapat digunakan untuk proses pemodelan. Setelah nilai hilang teridentifikasi, langkah berikutnya adalah melakukan penanganan agar data dapat digunakan secara utuh dalam proses analisis. Dalam metode CRISP-DM, tahapan preprocessing data merupakan bagian dari tahap ketiga, yaitu *data preparation*,

3.3 Data Preparation

Tahapan *Data Preparation* merupakan langkah krusial dalam metode CRISP-DM, yang berfungsi menjembatani data mentah menuju bentuk yang optimal untuk proses pemodelan. Tahap ini mencakup berbagai aktivitas seperti seleksi atribut, penggabungan data, transformasi, pembersihan (*cleaning*), termasuk penanganan nilai yang hilang (*missing value*).

a) Penanganan *Missing Value*

Dalam penelitian ini, penanganan *missing value* dilakukan menggunakan operator *replace missing values* pada RapidMiner. Operator ini memungkinkan pengguna untuk secara otomatis mengganti nilai-nilai yang hilang pada setiap atribut. Penggunaan operator *Replace Missing Values* juga memastikan bahwa semua atribut dalam dataset sudah lengkap dan siap untuk digunakan dalam proses pemodelan lebih lanjut, seperti klasifikasi model. Berikut proses penanganan *missing value* menggunakan operator *replace missing value* pada Gambar 4.

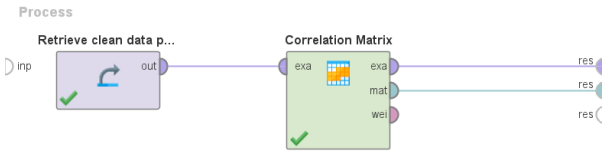


Gambar 4. Penanganan *Missing Values*

b) Analisis Korelasi Antar Atribut

Setelah proses penanganan *missing value* dilakukan, langkah selanjutnya dalam tahap *data preparation* adalah mengevaluasi hubungan antar atribut numerik dalam dataset. Hubungan atau korelasi antar atribut penting untuk diketahui agar dapat memahami pola keterkaitan antar variabel, serta membantu dalam proses seleksi fitur pada tahapan modeling selanjutnya. Untuk menganalisis korelasi, digunakan operator *correlation matrix* yang tersedia di RapidMiner. Gambar 5 berikut menunjukkan proses analisis korelasi antar atribut

menggunakan operator c*orrelation matrix* di RapidMiner.



Gambar 5. Proses *Correlation Matrix*

Operator *correlation matrix* menghitung nilai korelasi Pearson antar pasangan atribut numerik dalam data dan menyajikannya dalam bentuk matriks. Hasil *correlation matrix* disajikan pada Gambar 6 berikut.

Attribut...	Catego...	PM10	SO2	CO	O3	NO2	Max
Categor...	1	0.186	0.368	0.676	0.371	0.616	0.792
PM10	0.186	1	0.211	0.311	-0.072	0.351	0.276
SO2	0.368	0.211	1	0.591	-0.165	0.536	0.485
CO	0.676	0.311	0.591	1	0.008	0.732	0.905
O3	0.371	-0.072	-0.165	0.008	1	-0.041	0.336
NO2	0.616	0.351	0.536	0.732	-0.041	1	0.620
Max	0.792	0.276	0.485	0.905	0.336	0.620	1

Gambar 6. Hasil *Correlation Matrix*

Berdasarkan Gambar 6 visualisasi *correlation matrix*, dapat dilihat bahwa atribut target *Category* memiliki tingkat korelasi yang bervariasi dengan atribut-atribut lainnya dalam dataset kualitas udara. Tabel matriks korelasi menunjukkan bahwa atribut Max, CO, dan NO_2 memiliki korelasi yang cukup kuat terhadap atribut target.

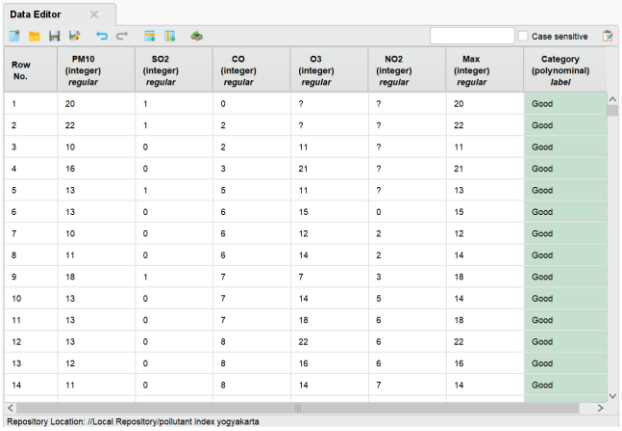
Atribut Max menunjukkan nilai korelasi tertinggi dengan *Category*, yaitu sebesar 0.792. Hal ini menunjukkan bahwa semakin tinggi nilai maksimum dari parameter kualitas udara yang tercatat, semakin besar kemungkinan kategori kualitas udara berada pada tingkat yang lebih buruk. Korelasi ini bersifat positif kuat, sehingga atribut Max dapat menjadi indikator yang signifikan dalam menentukan kelas kategori. Selanjutnya, atribut CO juga memiliki nilai korelasi yang tinggi dengan *Category*, yakni sebesar 0.676. Ini menunjukkan bahwa konsentrasi gas CO di udara memiliki hubungan erat dengan klasifikasi kualitas udara. Atribut NO_2 juga menunjukkan korelasi yang cukup signifikan terhadap *Category*, yaitu sebesar 0.616. Nilai ini menunjukkan bahwa semakin tinggi kadar NO_2, semakin besar kecenderungan kualitas udara diklasifikasikan dalam kategori buruk.

Setelah memastikan nilai korelasi tidak terlalu rendah dan tidak terlalu tinggi, Sebelum memasuki tahap pemodelan, langkah selanjutnya pada *data preparation* dalam metode CRISP-DM adalah menetapkan label pada atribut target yang akan digunakan untuk proses klasifikasi.

c) Proses Penetapan Label

Tahapan persiapan data, dalam penelitian ini digunakan seluruh atribut yang tersedia

pada dataset kualitas udara Kota Yogyakarta. Atribut target diklasifikasikan ke dalam bentuk label polinomial berdasarkan kategori Indeks Standar Pencemar Udara (ISPU), yaitu "*Good*", "*Moderate*", dan "*Unhealthy*" dapat dilihat pada Gambar 7 berikut.


Data Editor

Case sensitive

Row No.	PM10 (integer) regular	SO2 (integer) regular	CO (integer) regular	O3 (integer) regular	NO2 (integer) regular	Max (integer) regular	Category (polynominal) label
1	20	1	0	?	?	20	Good
2	22	1	2	?	?	22	Good
3	10	0	2	11	?	11	Good
4	16	0	3	21	?	21	Good
5	13	1	5	11	?	13	Good
6	13	0	6	15	0	15	Good
7	10	0	6	12	2	12	Good
8	11	0	6	14	2	14	Good
9	18	1	7	7	3	18	Good
10	13	0	7	14	5	14	Good
11	13	0	7	18	6	18	Good
12	13	0	8	22	6	22	Good
13	12	0	8	16	6	16	Good
14	11	0	8	14	7	14	Good

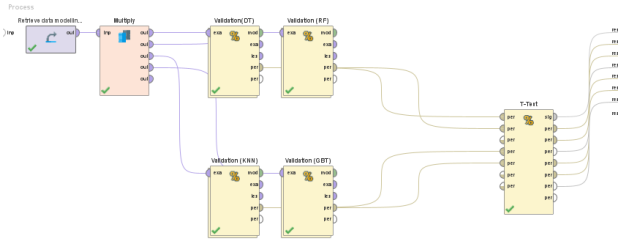
Repository Location: //Local Repository/pollutant index yogyakarta

Gambar 7. Penetapan Label Pada Atribut Target

Gambar 7 menunjukkan proses penetapan label pada atribut target dalam dataset kualitas udara yang ditampilkan melalui Data Editor. Pada tabel tersebut terlihat beberapa variabel input berupa konsentrasi polutan, yaitu PM10, SO2, CO, O3, dan NO2 yang bertipe numerik (integer). Selain itu, terdapat kolom "*Max*" yang merepresentasikan nilai maksimum dari variasi polutan pada setiap baris data. Berdasarkan nilai maksimum tersebut, kemudian ditetapkan kolom "*Category*" sebagai atribut target dengan tipe nominal (polynomial), yang dalam tampilan ini didominasi oleh label "*Good*". Artinya, data pada baris-baris yang ditampilkan diklasifikasikan ke dalam kategori kualitas udara "baik" berdasarkan nilai ambang tertentu yang digunakan sebagai dasar pelabelan.

3.4 Modelling

Tahap pemodelan dalam proses data mining bertujuan untuk mengevaluasi hubungan antar variabel serta memilih algoritma yang paling sesuai untuk digunakan. Pada tahap ini, dilakukan perbandingan terhadap empat algoritma klasifikasi, yaitu *Decision Tree*, *Random Forest*, *K-Nearest Neighbor*, dan *Gradient-Boosted Trees*. Untuk menilai kinerja masing-masing algoritma, digunakan uji statistik berupa *T-Test*. Melalui uji tersebut, dapat diketahui tingkat akurasi dari setiap model yang dihasilkan. Berikut Gambar 8 merupakan *modelling* keempat algoritma klasifikasi yang diterapkan.



Gambar 8. *Modelling* Keempat Metode Klasifikasi

Gambar 8 memperlihatkan proses pemodelan klasifikasi menggunakan empat algoritma berbeda, yaitu *Decision Tree* (DT), *Random Forest* (RF), *K-Nearest Neighbor* (K-NN), dan *Gradient Boosting Trees* (GBT) dengan metode validasi silang (*Cross Validation*) pada *software* RapidMiner. Operator Multiply digunakan untuk mendistribusikan data yang sama ke masing-masing alur pemodelan agar dapat dibandingkan secara adil dengan data input yang identik.

Setiap algoritma dimasukkan ke dalam operator Validation, yang bertujuan melakukan *training* dan *testing* model secara otomatis menggunakan metode *k-fold cross validation* berjumlah 10 lipatan (*number of folds*) untuk menghindari overfitting dan memastikan bahwa hasil evaluasi lebih representatif terhadap performa model secara umum. Output dari masing-masing proses validasi berupa metrik evaluasi kemudian dikumpulkan oleh operator T-Test, yang bertujuan untuk membandingkan hasil performa keempat model secara statistik.

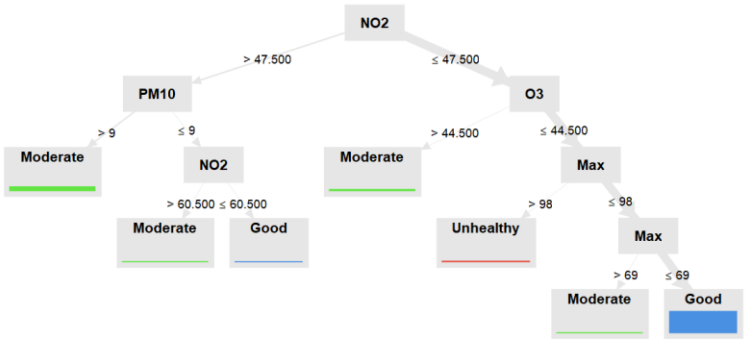
Adapun perbandingan matriks evaluasi (*performance accuracy*) keempat algoritma disajikan pada Tabel 2 yang memuat metrik-metrik utama hasil pengujian masing-masing algoritma.

Tabel 2 Perbandingan *Performance Accuracy*

	Category	**Algoritma**			
		RF	**GBT**	**KNN**	**DT**
Class Precision	Good	100.00%	99.32%	99.32%	90.15%
	Moderate	98.53%	93.34%	91.55%	92.31%
	Unhealthy	100.00%	20.00%	100.00%	100.00%
Class Recall	Good	100.00%	100.00%	98.98%	100.00%
	Moderate	100.00%	92.54%	97.01%	53.73%
	Unhealthy	83.33%	16.67%	50.00%	33.33%

Berdasarkan Tabel 4.2, terlihat bahwa hasil evaluasi metrik *class precision* dan *class recall*, di mana model *Random Forest* menunjukkan akurasi yang konsisten tinggi di semua kategori kelas (*Good*, *Moderate*, dan *Unhealthy*). Precision dan *recall* untuk kategori "*Good*" dan "*Moderate*" mencapai angka 100%, serta *recall* untuk kategori "*Unhealthy*" juga lebih tinggi dibandingkan metode lainnya (83.33%). Hal ini menunjukkan bahwa model *Random Forest* mampu mengidentifikasi dengan sangat baik data yang termasuk dalam masing-masing kategori, terutama dalam kasus kategori minoritas seperti "*Unhealthy*" yang seringkali menjadi tantangan dalam pemodelan klasifikasi. Oleh karena itu, model *Random Forest* dipilih sebagai model dengan performa terbaik dalam mengklasifikasikan data kualitas udara SPKU Gondokusuman Kota Yogyakarta.

Algoritma *Random Forest* terdiri atas sejumlah pohon Keputusan yang bekerja secara kolektif untuk mengklasifikasikan data ke dalam kelas tertentu. Algoritma ini digunakan untuk membangun pohon keputusan yang terdiri dari *root node*, *internal node*, dan *leaf node* dengan mengambil atribut dan data secara acak sesuai ketentuan yang diberlakukan. *root node* merupakan simpul yang terletak paling atas, atau biasa disebut sebagai akar dari pohon keputusan. *Internal node* adalah simpul percabangan, dimana node ini mempunyai output minimal dua dan hanya ada satu input (Siburian et al., 2019). Sedangkan *leaf node* atau terminal node merupakan simpul terakhir yang hanya memiliki satu input dan tidak mempunyai output. Berikut ini disajikan salah satu model *Random Forest* yang terbentuk.



Gambar 9. Model Algoritma Random Forest

Gambar 9 memperlihatkan salah satu pohon keputusan dari algoritma *Random Forest* yang digunakan untuk mengklasifikasikan kualitas udara di SPKU Gondokusuman Kota Yogya berdasarkan beberapa parameter pencemar, seperti NO_2, PM10, O_3, dan Max. Model ini bekerja dengan menetapkan kondisi-kondisi berupa nilai ambang pada setiap node untuk menentukan jalur klasifikasi yang akan diambil hingga mencapai suatu kelas pada *leaf node*, seperti "*Moderate*", "*Good*", atau "*Unhealthy*".

Cara membaca model ini dimulai dari *root node*, yaitu parameter NO_2. Jika nilai NO_2 lebih dari 47.500 maka proses klasifikasi berlanjut ke atribut PM_{10}. Jika PM_{10} lebih dari 9, maka hasil klasifikasi adalah "*Moderate*". Jika tidak, maka dilanjutkan ke node NO_2 dengan ambang batas 60.500 untuk menentukan apakah klasifikasi adalah "*Moderate*" atau "*Good*". Sebaliknya, jika nilai awal NO_2 kurang dari atau sama dengan 47.500, maka proses klasifikasi bergerak ke atribut O_3. Jika nilai O_3 kurang dari atau sama dengan 44.500, maka diproses lebih lanjut dengan atribut "*Max*". Jika nilai Max lebih dari 98, maka hasil klasifikasi adalah "*Unhealthy*", dan jika tidak, maka proses klasifikasi masuk lagi ke parameter yang sama, yaitu Max, namun dengan ambang batas berbeda.

Jika nilai Max kurang dari atau sama dengan 69, maka diklasifikasikan sebagai "*Good*", sedangkan jika Max > 69, maka hasil klasifikasinya adalah "*Moderate*". Namun, jika O_3 lebih dari atau sama dengan 4.,500, maka kualitas udara diklasifikasikan sebagai "*Moderate*". Tahap selanjutnya adalah mengevaluasi hasil dari model terbaik.

3.5 Evaluation

Pada tahap evaluasi, dilakukan uji statistik menggunakan operator *T-Test* untuk membandingkan kinerja empat algoritma klasifikasi, yaitu *Decision Tree*, *Random Forest*, *K-Nearest Neighbor*, dan *Gradient Boosted Tree*. Gambar 10 berikut disajikan hasil uji beda ke-4 model algoritma klasifikasi yang diterapkan.

A	B	C	D	E
	0.905 +/- 0.087	0.997 +/- 0.009	0.978 +/- 0.025	0.973 +/- 0.029
0.905 +/- 0.087		0.003	0.019	0.029
0.997 +/- 0.009			0.035	0.018
0.978 +/- 0.025				0.674
0.973 +/- 0.029				

Gambar 10. Hasil Uji Beda 4 Algoritma

Gambar 4.9 menunjukkan hasil pengujian menggunakan uji beda (*T-Test*) pada empat algoritma klasifikasi, diketahui bahwa *Random Forest* memberikan hasil paling optimal

dalam mengklasifikasikan kualitas udara di SPKU Gondokusuman Kota Yogyakarta dengan akurasi sebesar 99.7% dan *p-value* signifikan yaitu 0.018 (<0.05). Untuk melihat akurasi pengukuran Metode klasifikasi digunakan *Confution Matrix* dan *ROC Curve/ Area Under Curve* (AUC). Penelitian ini tidak menyajikan kurva ROC-AUC (*Area Under Curve*) karena dataset yang digunakan memiliki label bertipe polinomial, sehingga keluaran dari pemodelan hanya menampilkan nilai akurasi dari algoritma yang diterapkan. Setelah proses pelatihan dan pengujian model dilakukan, tahap selanjutnya adalah mengevaluasi performa model klasifikasi melalui *confusion matrix* yang ditampilkan pada Gambar 11 berikut ini.

accuracy: 99.73% +/- 0.85% (micro average: 99.73%)

	true Good	true Moderate	true Unhealthy	class precision
pred. Good	293	0	0	100.00%
pred. Moderate	0	67	1	98.53%
pred. Unhealthy	0	0	5	100.00%
class recall	100.00%	100.00%	83.33%	

Gambar 11. *Confusion Matrix* Algoritma Random Forest

Berdasarkan Gambar 4.10, hasil evaluasi terhadap model *Random Forest* yang diterapkan pada data kualitas udara di SPKU Gondokusuman Kota Yogyakarta menghasilkan nilai akurasi sebesar 99,73% ± 0,88%, dengan *class precision* tertinggi sebesar 100% pada kategori Good dan Unhealthy, serta *class recall* tertinggi sebesar 100% pada kategori Moderate dan Good. Hal ini menunjukkan bahwa model ini sangat andal dalam mengklasifikasikan kondisi udara di wilayah tersebut.

Jika dilihat dari *confusion matrix* pada tabel, model berhasil mengklasifikasikan sebanyak 293 data kualitas udara yang diklasifikasikan sebagai "Good", yang seluruhnya berhasil diidentifikasi secara tepat. Ini mencerminkan bahwa sebagian besar wilayah di Yogyakarta selama periode pengamatan memiliki kualitas udara yang tergolong baik.

Sementara itu, 67 data diklasifikasikan sebagai "*Moderate*", dengan hanya satu kasus salah klasifikasi, menunjukkan bahwa kondisi udara sedang juga cukup umum terjadi. Untuk kategori "*Unhealthy*", dari 6 data yang sebenarnya masuk dalam kategori ini, 5 berhasil diklasifikasikan dengan benar, dengan *recall* sebesar 83.33%, yang mengindikasikan masih terdapat sedikit tantangan dalam mengidentifikasi secara tepat kondisi udara yang memburuk.

3.6 Deployment

Pola pengetahuan yang diperoleh melalui proses standar *data mining* perlu ditindaklanjuti agar hasil analisis dapat dimanfaatkan secara nyata, khususnya dalam mendukung pengambilan keputusan terkait pengelolaan dan pemantauan kualitas udara. Berdasarkan hasil penelitian ini, terdapat beberapa rekomendasi implementatif yang dapat dilakukan sebagai berikut:

a. Pemanfaatan algoritma *Random Forest*
   Algoritma *Random Forest* terbukti memberikan akurasi tertinggi dalam klasifikasi kualitas udara, sehingga direkomendasikan untuk digunakan dalam sistem prediksi kualitas udara di Yogyakarta.
b. Fokus pada Atribut CO, Max, dan NO_2
   Atribut CO, NO_2, dan nilai maksimum polutan (Max) menunjukkan korelasi paling kuat terhadap kategori *Unhealthy*, sehingga perlu menjadi prioritas dalam pemantauan kualitas udara.

c. Penyederhanaan atribut-atribut dengan pengaruh rendah seperti PM_{10}, SO_2, dan O_3 dapat dipertimbangkan untuk disederhanakan dalam model, tujuannya untuk meningkatkan efisiensi prediksi.
d. Visualisasi *Confusion Matrix,* dapat digunakan sebagai alat evaluasi hasil prediksi model dan perlu diintegrasikan dalam sistem pemantauan kualitas udara.
e. Pengembangan Sistem Peringatan Dini
   Hasil klasifikasi dapat dimanfaatkan untuk membangun sistem peringatan dini yang memberi notifikasi otomatis saat kualitas udara diprediksi memburuk.

4 KESIMPULAN

Berdasarkan hasil evaluasi model menggunakan *confusion matrix* dan Uji *T-test,* algoritma *Random Forest* terbukti memberikan hasil klasifikasi paling optimal. Dengan tingkat akurasi mencapai 99,73% ± 0,88%, *class precision* hingga 100%, dan *class recall* tertinggi pada kategori "*Good*" dan "*Moderate*". Model ini juga menunjukkan korelasi tinggi terhadap parameter CO, NO_2 dan *Max,* yang terbukti paling berpengaruh dalam penentuan kategori kualitas udara (ISPU).

Dengan demikian, *Random Forest* direkomendasikan sebagai algoritma klasifikasi yang andal untuk sistem prediksi kualitas udara di SPKU Gondokusuman Kota Yogyakarta. Temuan ini juga membuka peluang pengembangan sistem peringatan dini berbasis data mining yang dapat digunakan oleh instansi terkait dalam memantau dan mengendalikan pencemaran udara secara efisien.

UCAPAN TERIMAKASIH

Ucapan terima kasih penulis sampaikan kepada seluruh pihak yang telah berkontribusi dalam penyusunan artikel ini. Secara khusus, penulis mengucapkan terima kasih kepada dosen pembimbing yang telah memberikan arahan, bimbingan, serta masukan yang sangat berharga. Penulis juga mengucapkan terima kasih kepada Dinas Lingkungan Hidup Kota Yogyakarta sebagai pihak resmi penyedia dataset yang digunakan dalam penelitian ini. Selain itu, penulis turut mengapresiasi dukungan dari keluarga serta rekan-rekan yang telah memberikan motivasi dan bantuan selama proses penyusunan artikel ini.

DAFTAR PUSTAKA

adminwarta. (2025). Program Langit Biru Kendalikan Pencemaran Udara di Kota Yogya. *Portal Berita Pemerintah Kota Yogyakarta*. https://warta.jogjakota.go.id/detail/index/4669

Andini, M., Arif Budianto, Laili Mardiana, & Susi Rahayu. (2025). Pengukuran Konsentrasi Emisi Partikulat di Ruang Tertutup Menggunakan Kit Pemantauan Kualitas Udara. *Lambda Jurnal Ilmiah Pendidikan MIPA Dan Aplikasinya*, *5*(1), 133–139. https://doi.org/10.58218/lambda.v5i1.1227

Dasrul Chaniago, Annisa Zahara, I. S. R. (2020). Indeks Standar Pencemar Udara (ISPU) Sebagai Informasi Mutu Udara Ambien di Indonesia. *Kementerian Lingkungan Hidup Dan Kehutanan; Direktorat Pengendalian Pencemaran Udara*. https://ditppu.menlhk.go.id/portal/

Fitrianti, I., Voutama, A., & Umaidah, Y. (2023). Clustering Film Populer Pada Aplikasi Netflix Dengan Menggunakan Algoritma K-Means Dan Metode CRISP-DM Clustering Popular Movies on Netflix App Using K-Means Algorithm and CRISP-DM Method. *Jtsi*, *4*(2), 301–311.

Irjayana, R. C., Fadlil, A., & Umar, R. (2025). Pengaruh Seleksi Fitur Terhadap Klasifikasi Indeks Standar Pencemar Udara Menggunakan Naïve Bayes. *Insect (Informatics and Security): Jurnal Teknik Informatika*, *11*(1), 67–78. https://doi.org/10.33506/insect.v11i1.4303

Kirono, Abdul Aziiz Hendrie, Ibnu Asror, and Y. F. A. W. (2022). *Klasifikasi Tingkat Kualitas Udara Dki Jakara Dengan Algoritma Naive Bayes*.

Kumar, R., & K S, S. (2024). Impact of sentimental factors on stock portfolio returns –an empirical analysis. *Heliyon*, *10*(9), e30217. https://doi.org/10.1016/j.heliyon.2024.e30217

Siburian, Vanissa Wanika, and I. E. M. (2019). Prediksi Harga Ponsel Menggunakan Metode Random Forest. *Annual Research Seminar (ARS)*, *4*(1), 144–147.

Syahira, N., & Arianto, D. B. (2024). Prediksi Tingkat Kualitas Udara dengan Pendekataan Algoritma K-Nearest Neighbor. *Jurnal Ilmiah Informatika Komputer*, *29*(1), 45–59. https://doi.org/10.35760/ik.2024.v29i1.10069

Yuliska, Y., & Syaliman, K. U. (2020). Peningkatan Akurasi K-Nearest Neighbor Pada Data Index Standar Pencemaran Udara Kota Pekanbaru. *IT Journal Research and Development*, *5*(1), 11–18. https://doi.org/10.25299/itjrd.2020.vol5(1).4680

adminwarta. (2025). Program Langit Biru Kendalikan Pencemaran Udara di Kota Yogya. *Portal Berita Pemerintah Kota Yogyakarta*. https://warta.jogjakota.go.id/detail/index/4669

Andini, M., Arif Budianto, Laili Mardiana, & Susi Rahayu. (2025). Pengukuran Konsentrasi Emisi Partikulat di Ruang Tertutup Menggunakan Kit Pemantauan Kualitas Udara. *Lambda Jurnal Ilmiah Pendidikan MIPA Dan Aplikasinya*, *5*(1), 133–139. https://doi.org/10.58218/lambda.v5i1.1227

Dasrul Chaniago, Annisa Zahara, I. S. R. (2020). Indeks Standar Pencemar Udara (ISPU) Sebagai Informasi Mutu Udara Ambien di Indonesia. *Kementerian Lingkungan Hidup Dan Kehutanan; Direktorat Pengendalian Pencemaran Udara*. https://ditppu.menlhk.go.id/portal/

Fitrianti, I., Voutama, A., & Umaidah, Y. (2023). Clustering Film Populer Pada Aplikasi Netflix Dengan Menggunakan Algoritma K-Means Dan Metode CRISP-DM Clustering Popular Movies on Netflix App Using K-Means Algorithm and CRISP-DM Method. *Jtsi*, *4*(2), 301–311.

Irjayana, R. C., Fadlil, A., & Umar, R. (2025). Pengaruh Seleksi Fitur Terhadap Klasifikasi Indeks Standar Pencemar Udara Menggunakan Naïve Bayes. *Insect (Informatics and Security): Jurnal Teknik Informatika*, *11*(1), 67–78. https://doi.org/10.33506/insect.v11i1.4303

Kirono, Abdul Aziiz Hendrie, Ibnu Asror, and Y. F. A. W. (2022). *Klasifikasi Tingkat Kualitas Udara Dki Jakara Dengan Algoritma Naive Bayes*.

Kumar, R., & K S, S. (2024). Impact of sentimental factors on stock portfolio returns –an empirical analysis. *Heliyon*, *10*(9), e30217. https://doi.org/10.1016/j.heliyon.2024.e30217

Siburian, Vanissa Wanika, and I. E. M. (2019). Prediksi Harga Ponsel Menggunakan Metode Random Forest. *Annual Research Seminar (ARS)*, *4*(1), 144–147.

Syahira, N., & Arianto, D. B. (2024). Prediksi Tingkat Kualitas Udara dengan Pendekataan Algoritma K-Nearest Neighbor. *Jurnal Ilmiah Informatika Komputer*, *29*(1), 45–59. https://doi.org/10.35760/ik.2024.v29i1.10069

Yuliska, Y., & Syaliman, K. U. (2020). Peningkatan Akurasi K-Nearest Neighbor Pada Data Index Standar Pencemaran Udara Kota Pekanbaru. *IT Journal Research and Development*, *5*(1), 11–18. https://doi.org/10.25299/itjrd.2020.vol5(1).4680

.