JUPITER
Teknologi Informatika & Komputer
Vol. 6 No. 2 Agustus 2025
Institut Bisnis Muhammadiyah Bekasi
P-ISSN: 2798-3129 , E-ISSN: 2798-6187

PREDIKSI KUALITAS UDARA DI DAERAH ISTIMEWA YOGYAKARTA MENGGUNAKAN ALGORITMA J48 DAN K-NN

Tri Fera Nofitasari[1,*], Jimly Assidqi Hardiansyah[1], Mu'afa Najmi Zain[1], Diya Ulhaq Jauhri Budiarto[1], Dhanar Intan Surya Saputra[1]

[1]Sistem Informasi, Universitas Amikom Purwokerto, chisakomikuyouku@gmail.com, jimlykuliah@gmail.com, nazmiefg400@gmail.com, ulhaqjb@gmail.com, dhanarsaputra@amikompurwokerto.ac.id

ABSTRAK

Pencemaran udara telah menjadi isu lingkungan yang signifikan di Indonesia, khususnya di wilayah Yogyakarta. Penelitian ini bertujuan untuk membandingkan performa dua algoritma klasifikasi, yaitu J48 dan *K-Nearest Neighbor* (K-NN), dalam mengklasifikasikan data kualitas udara ke dalam kategori "*Good*" dan "*Moderate*". *Dataset* yang digunakan berjumlah 5822 data yang diperoleh dari Kaggle. Tahapan *pre-processing* meliputi penghapusan data kosong, normalisasi, dan evaluasi menggunakan teknik 10-*fold cross-validation*. Hasil penelitian menunjkukan bahwa algoritma J48 memiliki akurasi sebesar 99,95% dengan nilai Kappa Statistic sebesar 0,9988, sedangkan K-NN memperoleh akurasi 98,57%. Implikasi dari penelitian ini menunjukkan bahwa J48 lebih andal digunakan dalam klasifikasi kualitas udara, terutama untuk sistem prediksi secara *real-time*. Penerapan sistem klasifikasi ini sangat penting dalam mendukung upaya pemantauan kualitas udara yang lebih cepat dan akurat, sehingga dapat membantu pengambilan keputusan yang responsif dalam menangani isu pencemaran lingkungan.
Kata Kunci: *Data Mining*, J48, K-NN, Kualitas Udara, Yogyakarta

ABSTRACT

Pollution has become a significant environmental issue in Indonesia, especially in the Yogyakarta region. This study aims to compare the performance of two classification algorithms, J48 and K-Nearest Neighbor (K-NN), in classifying air quality data into "Good" and "Moderate" categories. The dataset used contains 5822 records obtained from Kaggle. Pre-processing steps included removing missing values, normalizing data, and applying 10-fold cross-validation for evaluation. The findings show that the J48 algorithm outperforms K-NN with an accuracy of 99.95% and a Kappa Statistic of 0.9988, whereas K-NN achieved 98.57% accuracy. The implication of this study suggests that J48 is more reliable for air quality classification, particularly for real-time prediction systems. Implementing such a classification system is crucial for supporting faster and more accurate air quality monitoring, thereby enabling responsive decision-making in addressing environmental pollution issues.
Keywords: *Air classification, Data Mining, J48, K-NN, Yogyakarta*

PENDAHULUAN

Udara merupakan bagian dari komponen penting bagi kehidupan makhluk hidup, khususnya manusia, karena berfungsi sebagai sumber utama oksigen yang diperlukan dalam proses pernapasan (Deandra et al., 2024). Pencemaran udara menjadi penyebab utama menurunnya kualitas udara, dan permasalahan ini yang terjadi di berbagai wilayah di Indonesia, khususnya Daerah Istimewa Yogyakarta (DIY). Pencemaran udara terjadi ketika zat, energi, atau komponen lain masuk ke atmosfer akibat aktivitas manusia, yang kemudian menyebabkan menurunnya kualitas udara hingga dapat membahayakan atau berdampak pada kesehatan manusia (Sodiq & Sela, 2020). Kualitas udara memiliki

JUPITER
Teknologi Informatika & Komputer
Vol. 6 No. 2 Agustus 2025
Institut Bisnis Muhammadiyah Bekasi
P-ISSN: 2798-3129 , E-ISSN: 2798-6187

dampak tidak hanya terhadap ekosistem, tetapi juga terhadap kesehatan manusia, serta tingkat kualitas hidup secara keseluruhan (Maulana et al., 2024). Seiring meningkatnya kebutuhan akan pemantauan kualitas udara secara real-time, penggunaan algoritma machine learning semakin banyak diterapkan. Namun, proses klasifikasi kualitas udara masih menghadapi berbagai kendala umum. Tantangan tersebut meliputi keterbatasan data *real-time*, noise pada data sensor, serta ketidakseimbangan jumlah data pada masing-masing kategori kualitas udara (Maharajpet et al., 2024). Selain itu, kompleksitas dalam mengolah data dari berbagai parameter lingkungan seperti PM2.5, CO_2, dan suhu, menjadi tantangan tambahan dalam meningkatkan akurasi sistem klasifikasi. Oleh karena itu, pemilihan algoritma klasifikasi yang tepat sangat diperlukan agar sistem prediksi dapat memberikan hasil yang akurat dan efisien.

Teknologi *Internet of Things (IoT)* menjadi solusi dalam mengintegrasikan data sensor secara otomatis ke dalam sistem klasifikasi. Penggunaan IoT yang dikombinasikan dengan *machine learning* telah terbukti mampu meningkatkan efisiensi dan akurasi dalam prediksi kualitas udara secara *real-time* (Banciu et al., 2024). Sistem seperti ini dapat digunakan untuk mendeteksi kualitas udara pada waktu tetentu dan membantu proses pengambilan keputusan dalam mitigasi dampak polusi udara. Salah satu metode yang digunakan untuk menganalisis data kualitas udara adalah dengan menerapkan teknik *data mining*.

Data Mining merupakan proses pengolahan data yang mencakup berbagai sudut pandang, pengelompokkan data, serta pola-pola hubungan yang ditemukan (Astriyani et al., 2023). Klasifikasi merupakan proses untuk menemukan sekumpulan pola yang dapat menggambarkan, serta membedakan antar kelompok data (Kusuma, 2023). Metode klasifikasi seperti algoritma J48 dan *K-Nearest Neighbo*r (K-NN) digunakan untuk mengelompokkan data kualitas udara ke dalam kategori tertentu. Menurut (Alfian et al., 2024), algoritma *K-Nearest Neighbo*r (K-NN) bekerja dengan prinsip tetangga terdekat, di mana prediksi terhadap data baru dilakukan berdasarkan kesamaan nilai dengan data sebelumnya. Parameter K dalam algoritma ini menunjukkan jumlah tetangga yang dijadikan acuan dalam proses prediksi. Sementara itu, algoritma J48 merupakan implementasi dari metode keputusan pohon yang membentuk model berdasarkan atribut-atribut data.

Penelitian ini bertujuan untuk membandingkan kinerja algoritma J48 dan K-NN dalam mengklasifikasikan kualitas udara di wilayah Daerah Istimewa Yogyakarta (DIY), guna mengetahui algoritma mana yang lebih akurat dan efektif dalam mengidentifikasi kategori udara ”*Good*” dan ”*Moderate*”.

TINJAUAN PUSTAKA

Data Mining

Data mining merupakan teknik untuk menganalisis data dari berbagai sudut perspektif dan merangkumnya menjadi informasi yang bermanfaat (Astriyani et al., 2023). Mining adalah metode yang tergolong cepat dan praktis dalam mengidentifikasi pola maupun hubungan antar data secara otomatis (Suliman, 2021). Jadi, dapat disimpulkan bahwa melalui pendekatan ini, data yang sebelumnya tersebar dan tidak terstruktur dapat diolah menjadi pengetahuan baru yang mendukung proses pengambilan keputusan secara lebih tepat dan efisien.

Algoritma J48

Algoritma J48 adalah versi penyempurnaan dari algoritma C4.5 yang dikembangkan sebagai kelanjutan dari algoritma *decision tree* konvensional sebelumnya (Kusuma, 2023). Algoritma ini membentuk *decision tree* dari atribut dengan nilai *information gain* tertinggi dan akan terus membagi hingga kondisi klasifikasi paling optimal tercapai. Proses pembentukan pohon keputusan dimulai dari *node* akar, lalu atribut dengan nilai *information gain* tertinggi digunakan untuk membagi *node* tersebut. Prosedur ini diulangi secara rekursif hingga kondisi berhenti terpenuhi, yaitu ketika seluruh sampel dalam sebuah *node* memiliki kelas yang sama, atau tidak ada atribut yang memberikan gain signifikan (Srivastava et al., 2019). Kelebihan J48 meliputi kemampuannya untuk menangani data kontinu maupun kategorikal, toleransi pada nilai yang hilang, serta interpretasi model yang mudah melalui representasi pohon keputusan. Namun, seperti pohon keputusan lainnya, J48 rentan terhadap *overfitting* jika pohon terlalu kompleks, dan sifat *greedy*-nya bisa menyebabkan hasil lokal optimal, bukan global optimal. Beberapa penelitian juga mengombinasikan J48 dengan teknik ensemble seperti *AdaBost*, *Bagging*, atau *Rotation Forest* untuk meningkatkan akurasi.

Algoritma *K-Nearest Neighbor* (K-NN)

Algoritma *K-Nearest Neighbor* (K-NN) adalah algoritma klasifikasi yang mengklasifikasikan data baru berdasarkan kedekatannya dengan data yang telah memiliki label kelas (Sodiq & Sela, 2020). Pendekatan klasifikasi ini mengelompokkan data berdasarkan kategori yang paling umum di antara K tetangga terdekat (Maulana et al., 2024). Dengan kata lain, K-NN mengandalkan kedekatan data untuk membuat prediksi, sehingga akurasi algoritma ini sangat dipengaruhi oleh pemilihan nilai k dan kualitas data latih yang digunakan (Gunawan et al., 2024). Langkah utamanya dalam algoritma ini meliputi penentuan nilai k (jumlah tetangga terdekat), perhitungan jarak antara data baru dengan seluruh data latih (umumnya menggunakan jarak *Euclidean* atau *Manhattan*), kemudian menentukan kelas berdasarkan mayoritas kategori di antara tetangga tersebut. Untuk meningkatkan akurasi beberapa penelitian pemberian bobot pada kontribusi tetangga, misalnya skema bobot invers jarak atau metode probailistik seperti *Weighted* k-NN (W-kNN) dan *Proximity-Ratio* k-NN (PRkNN) yang secara signifikan membantu menangani ketidakseimbangan kelas dan noise (Kaunang, 2018).

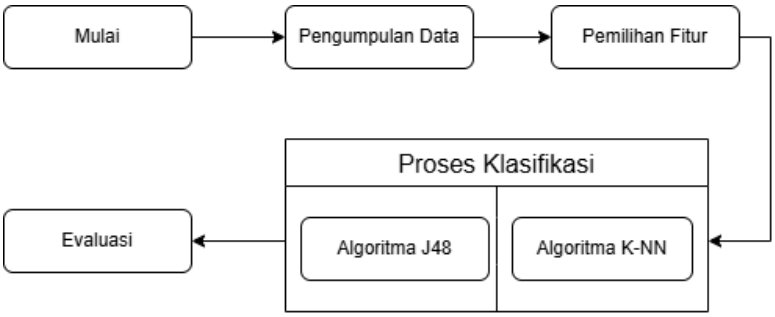
K-Fold Cross-Validation

K-fold cross-validation merupakan metode sampling yang digunakan untuk membagi *dataset* ke dalam beberapa subset (Maulana et al., 2024). Teknik ini sering digunakan untuk mengukur tingkat akurasi dan keandalan model yang dibangun berdasarkan *dataset* tertentu, dengan tujuan meminimalkan bias dan *overfitting* selama proses berlangsung. Pendekatan ini memastikan setiap titik data diuji satu kali, sehingga memberikan estimasi performa generalisasi yang lebih andal daripada pembagian sederhana seperti *train-test split* (Lumumba et al., 2024). Tujuan utama dari *K-fold cross-validation* adalah menurunkan bias dan *overfitting* selama pelatihan model, serta mendapatkan reliabilitas model secara konsisten di berbagai partisi data. Metode ini sangat berguna dalam proses pemilihan model dan penyetelan *hyperparameter* terutama ketika *dataset* terbatas atau heterogen karena memungkinkan evaluasi yang lebih stabil dan adil (Qiu, 2024).

JUPITER
Teknologi Informatika & Komputer
Vol. 6 No. 2 Agustus 2025
Institut Bisnis Muhammadiyah Bekasi
P-ISSN: 2798-3129 , E-ISSN: 2798-6187

METODE PENELITIAN

Urutan proses riset yang dilakukan ditampilkan pada Gambar 1, setiap bagian dijelaskan secara terstruktur untuk menunjukkan aliran kegiatan yang dilakukan dalam penelitian.



Gambar 1. Alur Metode Penelitian

Pengumpulan Data

Data yang digunakan dalam penelitian ini diambil dari *Kaggle*, berdasarkan data yang dilaporkan pada tahun 2021, dengan jumlah 5822 data dan dua kelas target, yaitu *Good* dan *Moderate*.

Pre-Processing Data

Dilakukan dengan tahapan menghapus nilai kosong (*missing value*) dan data duplikat, melakukan normalisasi data ke dalam rentang 0-1, mengimplementasikan algoritma J48 dan K-NN untuk klasifikasi menggunakan *software* Weka, melakukan evaluasi model menggunakan teknik 10-*fold cross-validation.*

Justifikasi Pemilihan Algortima

Algoritma J48 (C4.5) dipilih karena kemampuannya menghasilkan *decision tree* yang transparan, mudah diinterpretasikan, serta efektif dalam menangani atribut numerik dan kategorikal. Secara praktis, J48 memungkinkan visualisasi proses pengambilan keputusan dengan jelas, memudahkan analisis kesalahan dan interpretasi hasil (Kang & Michalak, 2018). Pemilihan J48 selaras dengan kebutuhan studi ini yang melibatkan dataset yang tidak terlalu besar (5.822 entri) dengan distribusi fitur yang beragam. J48 juga mampu menangani data hilang (*missing values*) dan menerapkan pruning untuk mencegah overfitting, hal yang sangat penting ketika berhadapan dengan data lingkungan yang berpotensi memuat noise (Pazhanivel et al., 2023). Dengan demikian, J48 menawarkan keseimbangan antara keandalan, interpretabilitas, dan performa yang robust.

Justifikasi Pemilihan *K-Nearest Neighbor (K-NN)*

Metode K-Nearest Neighbor (K-NN) dipilih karena kesederhanaannya dan fleksibilitasnya dalam memperhitungkan nilai sensor aktual tanpa memerlukan asumsi distribusi data (*non-parametric*). Karakter ini penting karena data kualitas udara sering bersifat non-linear dan multivariat. Sebuah telaah komprehensif oleh Halder et al. (2024) menyoroti bahwa peningkatan penyetelan jarak dan mekanisme pembobotan (weighted K-NN) dapat meningkatkan akurasi bahkan pada dataset yang kompleks, sangat relevan

dalam konteks kualitas udara (Evitania, 2023). Selain itu, K-NN tidak memerlukan fase pelatihan model yang kompleks, fitur yang sangat menguntungkan untuk aplikasi *real-time monitoring*, karena model dapat ditingkatkan saat data baru masuk tanpa perlu pelatihan ulang keseluruhan. Kombinasi kesederhanaan dan sensitivitas terhadap perubahan lokal menjadikan K-NN pilihan komparatif yang baik terhadap J48 dalam penelitian ini.

Pemilihan Parameter K=1 untuk K-NN

Nilai K=1 dipilih berdasarkan pengujian awal yang menunjukkan akurasi terbaik dibandingkan dengan nilai K lain (2–10), cocok dengan dataset yang relatif bersih dan seimbang. Pemilihan K terlalu besar dapat menyebabkan prediksi menjadi bias, karena mempertimbangkan terlalu banyak tetangga, sedangkan K terlalu kecil dapat memicu overfitting akibat sensitif terhadap noise (Halder et al., 2024). Pengaturan K=1 juga memudahkan interpretasi hasil: setiap datum diuji langsung terhadap satu tetangga terdekatnya. Strategi ini seringkali efektif dalam konteks data lingkungan yang memiliki klaster alami dan jarak magnetik antar sampel yang jelas, seperti yang ditekankan studi Halder et al. dalam domain kualitas udara. Selanjutnya, pemilihan nilai ini divalidasi melalui *cross-validation* untuk memastikan stabilitas performa.

Klasifikasi dengan Algoritma J48

J48 adalah implementasi dari algoritma C4.5, yang membentuk pohon keputusan berdasarkan *entropy* dan *informasi gain*.

1. *Entropy*

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

Berdasarkan rumus 1, S adalah himpunan data, n merupakan jumlah kelas, dan Pi merupakan proporsi jumlah data pada kelas ke-i terhadap total data.

2. *Information Gain*

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} . Entropy(S_v) \quad (2)$$

Berdasarkan rumus 2, *gain* adalah selisih antara nilai *entropy* sebelum dan sesudah data dibagi berdasarkan atribut tertentu. S merupakan himpunan data, A adalah atribut yang digunakan untuk memisahkan data, dan Sv adalah bagian dari data S yang memiliki nilai tertentu pada atribut A. Proporsi jumlah data pada setiap Sv terhadap total data dihitung, lalu dikalikan dengan *entropy* dari masing-masing Sv. Hasil perkalian tersebut dijumlahkan untuk seluruh nilai atribut A, kemudian dikurangkan dari *entropy* awal. Nilai gain menunjukkan seberapa besar atribut A dapat mengurangi ketidakpastian dalam data:

A: atribut yang sedang dihitung gainnya.
v: nilai-nilai unik dari atribut A.
S_v: subset dari S yang memiliki nilai atribut A = v
S: himpunan data.
|S|: jumlah data total dalam S.
$|S_v|$: jumlah data dalam subset S_v

JUPITER
Teknologi Informatika & Komputer
Vol. 6 No. 2 Agustus 2025
Institut Bisnis Muhammadiyah Bekasi
P-ISSN: 2798-3129 , E-ISSN: 2798-6187

Klasifikasi dengan Algoritma *K-Nearest Neighbor* (*K*-NN)

K-NN mengklasifikasikan data berdasarkan K tetangga terdekat menggunakan jarak. Beberapa metode pencarian jaraknya:

1. *Euclidean Distance*

$$distance(x, y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (3)$$

Berdasarkan rumus 3, *distance* atau jarak *Euclidean* digunakan untuk mengukur jarak lurus antara dua titik dalam ruang berdimensi n. X dan Y adalah dua titik atau vektor data, sedangkan xi dan yi masing-masing merupakan nilai pada dimensi ke-i dari titik X dan Y. Jarak dihitung dengan menjumlahkan kuadrat selisih antara xi dan yi untuk setiap dimensi, kemudian diakarkan. Rumus ini sering digunakan dalam pengukuran jarak dalam algoritma klasifikasi dan clustering.

2. *Manhattan Distance*

$$distance(x, y) = \sum_{i=1}^{n} |xi - yi| \quad (4)$$

Berdasarkan rumus 4, *distance* atau jarak *Manhattan* menghitung jumlah nilai mutlak selisih antara setiap pasangan elemen xi dan yi dari dua titik X dan Y. Metode ini mengukur jarak dengan cara menempuh jalur tegak lurus seperti pola jalan di kota *Manhattan*. Jarak ini cocok digunakan jika perpindahan antar titik tidak bisa dilakukan secara diagonal.

3. *Minkowski Distance*

$$distance(x, y) = (\sum_{i=1}^{n}|xi - yi|^A)^{\frac{1}{A}} \quad (5)$$

Berdasarkan rumus 5, *distance* atau jarak *Minkowski* adalah bentuk umum dari jarak *Euclidean* dan *Manhattan*. Rumus ini menghitung akar pangkat p dari jumlah pangkat p dari nilai mutlak selisih antara xi dan yi. Nilai p dapat disesuaikan; jika p = 1 maka menjadi jarak *Manhattan*, dan jika p = 2 menjadi jarak *Euclidean*. Rumus ini memberikan fleksibilitas dalam memilih jenis jarak sesuai kebutuhan analisis, dengan x data latih, y data uji, A lambda

Evaluasi

Evaluasi kinerja model dilakukan dengan menggukan metode 10-*fold cross-validation* untuk menghindari bias dalam pengujian. Teknik ini merupakan metode evaluasi yang digunakan untuk menilai tingkat akurasi algoritma dalam melakukan klasifikasi data (Deandra et al., 2024). Menurut (Qiu, 2024), pendekatan ini membagi *dataset* menjadi 10 bagian yang sama besar. Setiap bagian secara bergantian digunakan sebagai data uji, sementara sembilan bagian lainnya digunakan untuk melatih model. Proses ini diulang sebanyak 10 kali sehingga setiap bagian menjadi data uji satu kali. Rata-rata dari hasil evaluasi tiap iterasi digunakan untuk mengukur performa model secara keseluruhan. Teknik ini sangat efektif untuk mengurangi *overfitting* dan memberikan estimasi akurasi model yang lebih stabil. Matrik evaluasi yang digunakan meliputi: akurasi, *precision*, *recall*, F1-*score*, dan *confusion matrix*. Evaluasi dilakukan secara kuantitatif untuk mengetahui seberapa baik model mengenali kategori kualitas udara.

JUPITER
Teknologi Informatika & Komputer
Vol. 6 No. 2 Agustus 2025
Institut Bisnis Muhammadiyah Bekasi
P-ISSN: 2798-3129 , E-ISSN: 2798-6187

HASIL DAN PEMBAHASAN

Hasil Klasifikasi Menggunakan Algoritma J48

Setelah dilakukan klasifikasi menggunakan algoritma J48 pada *software* Weka, diperoleh hasil evaluasi model J48 dengan teknik 10-*fold cross-validation* menghasilkan performa seperti di Tabel 1 Evaluasi Klasifikasi J48 dan Tabel 1 *Confusion Matrix* J4.

Tabel 2. Evaluasi Klasifikasi J48

Metrik Evaluasi	**Nilai**
Jumlah Data	5822
Benar Diklasifikasikan	5819 (99,95%)
Salah Diklasifikasikan	3 (0,05%)
Kappa Statistic	0,9988
Mean Absolute Error	0,0048
Root Mean Squared Error	0,0335
Relative Absolute Error	1,11%
Root Relative Squared Error	7,15%

Tabel 3. Confusion Matrix J48

Aktual	*Good*	*Moderate*
Good	3943	0
Moderate	3	1876

Berdasarkan Tabel 1 Evaluasi Klasifikasi J48 dan Tabel 4 *Confusion Matrix J48,* algoritma J48 mampu mengklasifikasikan seluruh data pada kelas '*Good'* dengan benar, dan hanya salah mengklasifikasikan 3 data dari kelas '*Moderate'* menjadi '*Good'*. Ini menunjukkan stabilitas dan akurasi yang sangat tinggi, serta menunjukkan bahwa model J48 sangat tepat untuk data dengan struktur pola yang kuat. Gambar 2 hasil klasifikasi dan perhitungan pohon keputusan di *tools* Weka adalah hasil klasifikasi dan perhitungan pohon keputusan di *tools* Weka.

Classifier output

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

PM2.5 <= 50
|   O3 <= 46
|   |   SO2 <= 49: Good (3893.57/3.0)
|   |   SO2 > 49
|   |   |   SO2 <= 50: Good (12.17)
|   |   |   SO2 > 50: Moderate (17.23/0.23)
|   O3 > 46
|   |   O3 <= 50
|   |   |   SO2 <= 38: Good (29.22)
|   |   |   SO2 > 38: Moderate (18.1/0.1)
|   |   O3 > 50: Moderate (178.23/1.23)
PM2.5 > 50: Moderate (1673.49/9.49)

Number of Leaves  :     7

Size of the tree :      13
```

Gambar 3. Hasil Perhitungan Pohon Keputusan

JUPITER
Teknologi Informatika & Komputer
Vol. 6 No. 2 Agustus 2025
Institut Bisnis Muhammadiyah Bekasi
P-ISSN: 2798-3129 , E-ISSN: 2798-6187

Selanjutnya Gambar 3 menunjukkan representasi hasil dari proses klasifikasi yang telah dilakukan menggunakan model yang dibangun.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5819               99.9485 %
Incorrectly Classified Instances         3                0.0515 %
Kappa statistic                          0.9988
Mean absolute error                      0.0048
Root mean squared error                  0.0335
Relative absolute error                  1.1055 %
Root relative squared error              7.1552 %
Total Number of Instances             5822

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.002    0.999      1.000    1.000      0.999    0.999     0.998     Good
               0.998    0.000    1.000      0.998    0.999      0.999    0.999     0.999     Moderate
Weighted Avg.  0.999    0.001    0.999      0.999    0.999      0.999    0.999     0.999

=== Confusion Matrix ===

    a    b   <-- classified as
 3943    0 |    a = Good
    3 1876 |    b = Moderate
```

Gambar 4. *Output* Klasifikasi

Hasil Klasifikasi Menggunakan Algoritma K-NN

Setelah dilakukan klasifikasi menggunakan algoritma K-NN pada *software* Weka, diperoleh hasil evaluasi model J48 dengan teknik 10-*fold cross-validation*, dengan K=1 menghasilkan performa.

Tabel 5. Tabel Evaluasi Klasifikasi K-NN

Metrik Evaluasi	**Nilai**
Jumlah Data	5822
Data Benar	5739 (98,57%)
Data Salah	83 (1,43%)
Kappa Statistic	0,9679
Mean Absolute Error	0,0144
Root Mean Squared Error	0,1194
Relative Absolute Error	3,30%
Root Relative Squared Error	25,53%

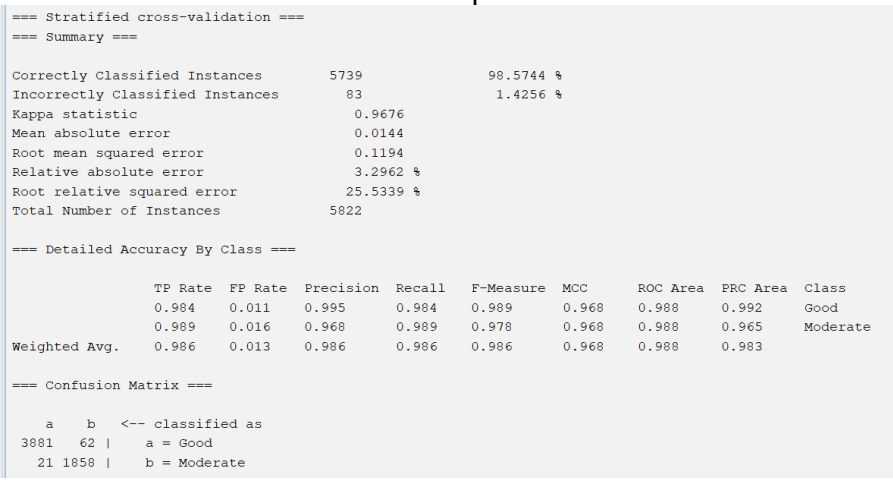
Tabel 6. Confusion Matrix K-NN

Aktual	***Good***	***Moderate***
Good	3881	62
Moderate	21	1858

Model K-NN dengan nilai K=1 pada mampu mengklasifikasikan data kualitas udara ke dalam dua kategori utama, yaitu Good dan Moderate. Pada Gambar 4 hasil klasifikasi menunjukkan bahwa mayoritas data dapat dipetakan dengan baik hal ini tercermin dari hasil evaluasi pada Tabel 3 di mana dari 5.822 data yang diuji, sebanyak 5.739 data berhasil diklasifikasikan dengan benar dan 83 data salah klasifikasi. Kesalahan tersebut terutama terjadi pada data yang berada di perbatasan antar kelas, karena sifat K-NN yang sangat bergantung pada jarak terdekat dengan tetangga data latih. Kondisi ini mengindikasikan bahwa K-NN lebih sensitif terhadap *noise* dan distribusi data yang tidak seimbang dibandingkan dengan algoritma berbasis pohon keputusan seperti J48. Meskipun begitu, dengan tingkat akurasi 98,57%, K-NN tetap dapat dikategorikan

JUPITER
Teknologi Informatika & Komputer
Vol. 6 No. 2 Agustus 2025
Institut Bisnis Muhammadiyah Bekasi
P-ISSN: 2798-3129 , E-ISSN: 2798-6187

sebagai model yang cukup andal, terutama untuk aplikasi prediksi kualitas udara yang membutuhkan fleksibilitas dan kemudahan implementasi secara *real-time*.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5739               98.5744 %
Incorrectly Classified Instances        83                1.4256 %
Kappa statistic                          0.9676
Mean absolute error                      0.0144
Root mean squared error                  0.1194
Relative absolute error                  3.2962 %
Root relative squared error             25.5339 %
Total Number of Instances             5822

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.984    0.011    0.995      0.984    0.989      0.968    0.988     0.992     Good
               0.989    0.016    0.968      0.989    0.978      0.968    0.988     0.965     Moderate
Weighted Avg.  0.986    0.013    0.986      0.986    0.986      0.968    0.988     0.983

=== Confusion Matrix ===

    a    b   <-- classified as
 3881   62 |    a = Good
   21 1858 |    b = Moderate
```

Gambar 5. Output Klasifikasi Algoritma K-NN

Perbandingan J48 dan K-NN

Pada Tabel 5 terlihat algoritma J48 memiliki performa lebih baik dibandingkan K-NN dalam mengklasifikasikan data kualitas udara. Hal ini terlihat dari akurasi yang lebih tinggi, kesalahan klasifikasi yang jauh lebih rendah dan nilai *Kappa* yang hampir sempurna.

Tabel 7. Perbandingan J48 vs K-NN

Metrik Evaluasi	J48	K-NN
Akurasi	99,95%	98,575%
Jumlah Kesalahan	3	83
Kappa Statistic	0,9988	0,9676
Mean Absolute Error	0,0048	0,0144
Root Mean Squared Error	0,0335	0,1194

Berdasarkan hasil evaluasi di atas, algoritma J48 menunjukkan performa yang lebih unggul dibandingkan algoritma K-NN, baik dari segi akurasi, tingkat kesalahan, hingga nilai *error*nya, hal ini disebabkan:

1. J48 membangun struktur pohon keputusan berdasarkan atribut dengan *information gain* tertinggi yang membuat modelnya lebih generalisasi dan tidak mudah *overfitting*.
2. Sementara K-NN sebagai algoritma *lazy learning*, hanya membuat keputusan berdasarkan jarak terdekat ke data latih yang membuatnya lebih sensitif terhadap *noise*, *outlier*, dan distribusi data yang tidak seimbang.

Meskipun K-NN juga menunjukkan hasil yang baik, namun secara konsisten J48 memberikan performa klasifikasi yang lebih stabil dan akurat dalam mengenali kategori kualitas udara (*Good* dan *Moderate*) di Daerah Istimewa Yogyakarta.

Manfaat dan Dampak Sosial Sistem Predikasi Kualitas Udara

Penerapan sistem kualitas udara berbasis algoritma J48 dan K-NN di Yogyakarta memiliki dampak signifikan terhadap kesehatan masyarakat. Sebuah studi yang menganalisis data peringatan kualitas udara menemukan bahwa pemberi sinyal peringatan kualitas udara dengan menurunkan pengeluaran untuk perawatan pernapasan

anak hingga 30% dan gangguan kardiovaskular orang dewasa hinga 23%. Hal ini menunjukkan bahwa prediksi kualitas udara secara *real-time* dan akurat dapat berkontribusi pada pengurangan beban ekonomi dan sosial akibat penyakit terkait polusi (Hilly et al., 2024).

Selain itu, sistem prediksi ini mendukung pembuatan kebijakan publik yang lebih responsif. Integrasi kecerdasan buatan dan IoT dalam pemantauan udara kota *smart-city* terbukti mampu mengoptimalkan manajemen polusi, memperkecil penyakit urban, serta mendorong pengembangan berkelanjutan. Dengan akurasi J48 mendekati 100%, hasil penelitian ini bisa menjadi alat pengambil keputusan *(decision-support)* yang kuat untuk perencanaan zona hijau, regulasi lalu lintas, atau kebijakan larangan tertentu berdasarkan data kondisi udara terkini (Neo et al., 2023).

Selanjutnya, sistem ini juga mendorong partisipasi masyarakat dalam pemantauan lingkungan. Riset telah menunjukkan bahwa keterlibatan publik melalui akses aplikasi atau platform interaktif meningkatkan kesadaran serta perubahan perilaku terhadap polusi udara, seperti penggunaan masker atau menghindari aktivitas luar ruangan saat kualitas udara buruk. Penggunaan data dan visualisasi terkait prediksi juga dapat memperkuat advokasi masyarakat dan mendorong tindakan kolektif dalam menjaga lingkungan yang lebih bersih dan sehat.

PENUTUP

Hasil penelitian menunjukkan bahwa algoritma J48 memberikan performa yang lebih unggul dibandingkan dengan algoritma *K-Nearest Neighbor* (K-NN) dalam klasifikasi kualitas udara Daerah Istimewa Yogyakarta ke dalam kategori "*Good*" dan "*Moderate*". Dengan akurasi hampir sempurna (99.95%) dan hanya 3 kesalahan klasifikasi dari 5822 data, serta nilai *Kappa Statistic* sebesar 0.9988, J48 terbukti sangat andal dan efisien dibandingkan K-NN yang memiliki akurasi 98,57% dengan 83 data salah klasifikasi. Atribut seperti PM2.5, PM10, CO, NO2, SO2, dan jadi faktor penting dalam proses klasifikasi. Berdasarkan hasil ini, J48 direkomendasikan untuk digunakan dalam sistem prediksi udara, khususnya yang berbasis *real-time*. Untuk penelitian selanjutnya, disarankan agar menggunakan data yang lebih beragam, baik dari sisi waktu maupun lokasi, serta mempertimbangkan penggunaan algoritma lain atau metode *ensemble* untuk meningkatkan akurasi model. Selain itu, penerapan *feature* selection juga dapat membantu menyaring atribut yang paling relevan, sehingga mempercepat proses klasifikasi dan meningkatkan akurasi prediksi.

Namun demikian, penelitian ini memiliki beberapa batasan yang perlu diperhatikan. Salah satunya adalah penggunaan data *historis* yang terbatas pada wilayah (DIY) dan periode waktu tertentu, yang mungkin belum sepenuhnya merepresentasikan variasi kualitas udara secara nasional atau musiman. Selain itu, sistem yang dikembangkan belum diuji dalam kondisi *real-time* berbasis sensor langsung, sehingga performa aktual dalam lingkungan dinamis masih perlu divalidasi lebih lanjut. Dengan mempertimbangkan hasil dan keterbatasan tersebut, penelitian ini memberikan kontribusi awal yang kuat untuk pengembangan sistem prediksi kualitas udara berbasis *machine learning*. Integrasi lebih lanjut dengan sensor IoT, visualiasi dashboard, dan penyebaran informasi kepada masyarakat melalui sistem notifikasi dapat menjadi arah pengembangan berikutnya yang menjanjikan baik secara teknis maupun dari sisi dampak sosial.

REFERENSI

Astriyani, M., Laela, I. N., Lestari, D. P., Anggraeni, L., & Astuti, T. (2023). Analisis Klasifikasi Data Kualitas Udara Dki Jakarta Menggunakan Algoritma C.45. JuSiTik : Jurnal Sistem Dan Teknologi Informasi Komunikasi, 6(1), 36–41. https://doi.org/10.32524/jusitik.v6i1.790

Banciu, C., Florea, A., & Bogdan, R. (2024). Monitoring and Predicting Air Quality with IoT Devices. Processes, 12(9). https://doi.org/10.3390/pr12091961

Evitania, C.G. (2023). Implementation of the K-Nearest Neighbor Algorithm to Predict Air Pollution. Information Technology and Systems, 1(1), 45–54. https://doi.org/10.58777/its.v1i1.123

Gunawan, M. N., Farhanah, T., Masruroh, S. U., Jundulloh, A. M., Raushanfikar, N. Z., & Amriza, R. N. S. (2024). Accuracy of K-Nearest Neighbors Algorithm Classification For Archiving Research Publications. MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer, 23(3), 593–602. https://doi.org/10.30812/matrik.v23i3.3915

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-024-00973-y

Hilly, J. J., Singh, K. R., Jagals, P., Mani, F. S., Turagabeci, A., Ashworth, M., Mataki, M., Morawska, L., Knibbs, L. D., Stuetz, R. M., & Dansie, A. P. (2024). Review of scientific research on air quality and environmental health risk and impact for PICTS. In Science of the Total Environment (Vol. 942). Elsevier B.V. https://doi.org/10.1016/j.scitotenv.2024.173628

Kang, K., & Michalak, J. (2018). Enhanced version of AdaBoostM1 with J48 Tree learning method.

Kaunang, F. J. (2018). Penerapan Algoritma J48 Decision Tree Untuk Analisis Tingkat Kemiskinan di Indonesia Application of J48 Decision Tree Algorithm For Analyzing Poverty Level in Indonesia. Cogito Smart Journal, 4(2). www.bps.go.id

Lumumba, V., Kiprotich, D., Mpaine, M., Makena, N., & Kavita, M. (2024). Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models. American Journal of Theoretical and Applied Statistics, 13(5), 127–137. https://doi.org/10.11648/j.ajtas.20241305.13

Maharajpet, S. S., Likhitha, S., & Kiran, T. (2024). Air Quality Prediction Using Machine Learning. Convergence of Machine Learning and IoT for Enabling the Future of Intelligent Systems, 97–103. https://doi.org/10.48001/978-81-966500-7-0-9

Neo, E. X., Hasikin, K., Lai, K. W., Mokhtar, M. I., Azizan, M. M., Hizaddin, H. F., Razak, S. A., & Yanto. (2023). Artificial intelligence-assisted air quality monitoring for smart city management. PeerJ Computer Science, 9. https://doi.org/10.7717/peerj-cs.1306

Pazhanivel. K., Kumar, U.D., Naveen, K., & Niranjan, M. (2023). Air Quality Prediction System using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology, 10–21. https://doi.org/10.48175/ijarsct-9254

Qiu, J. (2024). An Analysis of Model Evaluation with Cross-Validation: Techniques, Applications, and Recent Advances. Advances in Economics, Management and

Political Sciences, 99(1), 69–72. https://doi.org/10.54254/2754-1169/99/2024ox0213

Srivastava, A. K., Singh, D., Pandey, A. S., & Maini, T. (2019). A novel feature selection and short-term price forecasting based on a decision tree (J48) model. Energies, 12(19). https://doi.org/10.3390/en12193665

Suliman, S. (2021). Implementasi Data Mining Terhadap Prestasi Belajar Mahasiswa Berdasarkan Pergaulan dan Sosial Ekonomi Dengan Algoritma K-Means Clustering. Simkom, 6(1), 1–11. https://doi.org/10.51717/simkom.v6i1.48