Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 588−598
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9024

Analisis Komparasi Algoritma Machine Learning Untuk Klasifikasi Kualitas Udara Indoor Berbasis Sensor Low-Cost

Stefanus Eko Prasetyo, Irvan Hansen*, Haeruddin

Fakultas Ilmu Komputer, Program Studi Teknologi Informasi, Universitas International Batam, Batam
Baloi-Sei Ladi, Jl. Gajah Mada, Tiban Indah, Kec. Sekupang, Kota Batam, Kepulauan Riau, Indonesia
Email: [1]stefanus@uib.ac.id, [2,*]2232040.irvan@uib.edu, [3]haeruddin@uib.ac.id
Email Penulis Korespondensi: 2232040.irvan@uib.edu
Submitted: **24/12/2025**; Accepted: **31/01/2026**; Published: **31/01/2026**

Abstrak−Kualitas udara dalam ruangan (Indoor Air Quality/IAQ) berpengaruh langsung terhadap kesehatan dan kenyamanan penghuni bangunan, sementara keterbatasan sistem pemantauan konvensional dan tingginya biaya perangkat komersial menyebabkan pemantauan kualitas udara indoor belum banyak diterapkan secara luas. Pemantauan IAQ berbasis sensor low-cost menawarkan solusi ekonomis, namun data yang dihasilkan cenderung bervariasi dan mengandung noise sehingga sulit diinterpretasikan secara langsung. Penelitian ini bertujuan melakukan analisis komparatif performa beberapa algoritma machine learning dalam mengklasifikasikan kualitas udara indoor berdasarkan data sensor. Dataset dibangun dari hasil pengukuran sensor DHT22 dan MQ-135 yang merepresentasikan parameter suhu, kelembapan, dan tingkat polutan udara, dengan total 18.000 data yang terbagi merata ke dalam tiga kelas kualitas udara, yaitu Good, Moderate, dan Poor. Proses penelitian meliputi tahap pra-pemrosesan data menggunakan imputasi median dan standardisasi fitur, pembagian dataset secara stratified dengan rasio 70% data latih, 15% data validasi, dan 15% data uji, serta pelatihan dan pengujian empat algoritma supervised learning, yaitu Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), dan Gaussian Naive Bayes. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil pengujian menunjukkan bahwa seluruh model mampu mencapai performa klasifikasi yang tinggi, dengan algoritma KNN menghasilkan performa terbaik dengan nilai F1-score sebesar 1,00 pada data uji, sedangkan model dengan performa terendah masih mencapai F1-score di atas 0,96, menunjukkan rentang performa antarmodel yang relatif sempit namun konsisten. Temuan ini menunjukkan bahwa pendekatan machine learning efektif digunakan untuk klasifikasi kualitas udara indoor berbasis sensor low-cost pada skenario pengujian terkontrol.

Kata Kunci: Kualitas Udara Indoor; Machine Learning; K-Nearest Neighbors (KNN); Sensor Berbiaya Rendah; DHT22; MQ-135

Abstract−Indoor Air Quality (IAQ) has a significant impact on occupants' health and comfort; however, limitations of conventional monitoring systems and the high cost of commercial devices have hindered the widespread implementation of indoor air quality monitoring. Sensor-based IAQ monitoring using low-cost devices provides an affordable solution; however, the resulting data often exhibit variability and noise, making direct interpretation challenging. This study presents a comparative analysis of several machine learning algorithms for indoor air quality classification using sensor data. The dataset was collected from DHT22 and MQ-135 sensors measuring temperature, humidity, and air pollutant levels, resulting in 18,000 samples evenly distributed across three air quality classes: Good, Moderate, and Poor. The proposed methodology includes data preprocessing through median imputation and feature standardization, stratified dataset splitting with a ratio of 70% training, 15% validation, and 15% testing data, and model training using four supervised learning algorithms: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Gaussian Naive Bayes. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. Experimental results indicate that all evaluated models achieved high classification performance, with KNN outperforming other algorithms by achieving an F1-score of 1.00 on the test dataset, while the lowest-performing model still achieved an F1-score above 0.96, indicating a relatively narrow yet consistent performance range among the evaluated algorithms. These findings demonstrate the effectiveness of machine learning approaches for indoor air quality classification using low-cost sensor data under controlled experimental conditions.

Keywords: Indoor Air Quality; Machine Learning; K-Nearest Neighbors (KNN); Low-Cost Sensor; DHT22; MQ-135

1. PENDAHULUAN

Kualitas udara dalam ruangan (Indoor Air Quality / IAQ) merupakan salah satu faktor lingkungan yang memiliki pengaruh signifikan terhadap kesehatan, kenyamanan, dan produktivitas manusia [1]. Dalam kehidupan modern, sebagian besar aktivitas manusia berlangsung di dalam ruangan, baik di lingkungan hunian, perkantoran, fasilitas pendidikan, maupun ruang publik tertutup lainnya [2]. Kondisi tersebut menyebabkan paparan terhadap polutan udara indoor menjadi lebih dominan dibandingkan paparan polusi udara luar, sehingga kualitas udara dalam ruangan menjadi isu penting yang perlu mendapat perhatian khusus [3].

Udara dalam ruangan dapat mengandung berbagai jenis polutan yang berasal dari aktivitas manusia maupun sumber lingkungan lainnya. Aktivitas memasak, penggunaan bahan kimia rumah tangga, asap rokok, peralatan elektronik, serta material bangunan tertentu dapat menghasilkan gas dan partikel yang berpotensi membahayakan kesehatan [4]. Selain itu, sistem ventilasi yang tidak memadai dan sirkulasi udara yang buruk dapat menyebabkan akumulasi polutan dalam jangka waktu yang lama [5]. Dampak paparan polutan udara indoor tidak hanya bersifat akut, seperti iritasi saluran pernapasan dan sakit kepala, tetapi juga dapat bersifat kronis, termasuk gangguan pernapasan, penurunan fungsi kognitif, serta peningkatan risiko penyakit kardiovaskular [6].

Pemantauan kualitas udara indoor secara kontinu menjadi langkah penting dalam upaya mitigasi risiko kesehatan tersebut. Sistem pemantauan yang baik diharapkan mampu memberikan informasi kondisi udara secara

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 588−598
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9024

akurat dan real-time, sehingga tindakan korektif dapat dilakukan sebelum kualitas udara mencapai tingkat yang membahayakan. Namun, perangkat pemantauan kualitas udara komersial umumnya memiliki biaya yang relatif tinggi dan kurang fleksibel untuk diterapkan secara luas, terutama pada skala rumah tangga atau bangunan kecil [7].

Perkembangan teknologi sensor lingkungan berbiaya rendah menawarkan alternatif solusi yang lebih ekonomis dan mudah diimplementasikan [7]. Sensor suhu dan kelembapan memungkinkan pengukuran kondisi termal ruangan, sedangkan sensor gas mampu memberikan indikasi keberadaan senyawa polutan udara [8]. Kombinasi sensor-sensor tersebut dapat digunakan untuk membentuk sistem pemantauan kualitas udara indoor yang sederhana namun fungsional [9]. Meskipun demikian, sensor berbiaya rendah memiliki keterbatasan inheren, seperti akurasi yang lebih rendah dibandingkan sensor industri, sensitivitas terhadap perubahan lingkungan, serta fluktuasi data yang cukup tinggi. Data yang dihasilkan oleh sensor low-cost sering kali mengandung noise dan variasi yang tidak konsisten, sehingga sulit untuk ditafsirkan secara langsung [10].

Pendekatan konvensional dalam menentukan kualitas udara umumnya menggunakan metode berbasis ambang batas (threshold-based), di mana nilai sensor dibandingkan dengan batas tertentu untuk menentukan kategori kualitas udara. Pendekatan ini memiliki keterbatasan karena tidak mampu menangkap hubungan kompleks antar parameter kualitas udara, terutama pada kondisi transisi di mana nilai sensor berada di antara dua kategori. Selain itu, metode berbasis ambang batas cenderung tidak adaptif terhadap variasi lingkungan dan karakteristik sensor yang berbeda [11].

Machine learning menawarkan pendekatan yang lebih fleksibel dan adaptif dalam mengolah data sensor kualitas udara. Dengan memanfaatkan data historis, algoritma machine learning mampu mempelajari pola dan hubungan antar fitur sensor secara otomatis, termasuk hubungan nonlinier yang sulit dimodelkan secara matematis menggunakan pendekatan konvensional [12]. Pendekatan ini memungkinkan sistem untuk melakukan klasifikasi kualitas udara berdasarkan pola data secara menyeluruh, bukan hanya berdasarkan satu parameter atau satu nilai ambang batas.

Sejumlah penelitian terdahulu telah menerapkan machine learning dalam analisis kualitas udara. Rahman et al. (2024) mengevaluasi berbagai algoritma machine learning untuk prediksi kualitas udara berbasis parameter lingkungan [13]. Alani et al. (2025) menunjukkan variasi performa antar algoritma dalam analisis kualitas udara perkotaan[14]. Kim (2025) menyoroti peran machine learning dalam meningkatkan kualitas data pada jaringan sensor low-cost [15]. sementara Laton et al. (2025) menunjukkan meningkatnya penggunaan machine learning dalam penelitian Indoor Air Quality pada berbagai skenario lingkungan [16].

Meskipun berbagai penelitian terdahulu telah menerapkan machine learning dalam analisis kualitas udara, sebagian besar studi masih berfokus pada prediksi kualitas udara, kalibrasi sensor, atau analisis lingkungan luar ruangan. Penelitian yang secara khusus melakukan analisis komparatif terstruktur antar algoritma machine learning untuk klasifikasi kualitas udara indoor berbasis sensor low-cost, dengan skema kelas yang konsisten dan dataset indoor terkontrol, masih relatif terbatas. Kondisi ini menunjukkan adanya celah penelitian dalam pemilihan algoritma yang paling sesuai untuk sistem pemantauan kualitas udara indoor berbasis sensor berbiaya rendah.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk melakukan analisis komparatif performa empat algoritma machine learning, yaitu Logistic Regression, K-Nearest Neighbors, Support Vector Machine, dan Gaussian Naïve Bayes dalam mengklasifikasikan kualitas udara indoor. Dataset dibangun dari hasil pengukuran sensor suhu, kelembapan, dan gas pada lingkungan ruangan terkontrol dengan tiga kategori kualitas udara, yaitu Good, Moderate, dan Poor. Ketiga parameter tersebut dipilih karena kombinasi kondisi termal (suhu dan kelembapan) serta konsentrasi gas mampu merepresentasikan kenyamanan dan potensi risiko kesehatan udara indoor, di mana peningkatan senyawa gas seperti CO_2 dan TVOC umumnya berkorelasi dengan kualitas udara yang menurun dan dampak fisiologis pada penghuni ruangan.

Dengan demikian, kontribusi utama penelitian ini adalah menyajikan analisis komparatif performa algoritma machine learning pada klasifikasi kualitas udara indoor berbasis sensor low-cost, sehingga dapat menjadi dasar pemilihan algoritma yang efektif dan efisien untuk sistem pemantauan kualitas udara indoor, khususnya pada skenario implementasi perangkat berbiaya rendah dan lingkungan terkontrol.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

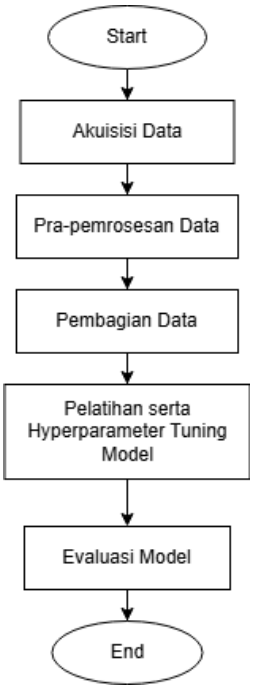
Penelitian ini menggunakan pendekatan eksperimen komparatif untuk menganalisis performa beberapa algoritma machine learning dalam mengklasifikasikan kualitas udara indoor berbasis data sensor low-cost. Tahapan penelitian yang dilakukan ditunjukkan pada Gambar 1, yang menggambarkan alur penelitian secara sistematis mulai dari proses awal hingga evaluasi akhir model.

Dari Gambar 1, tahapan penelitian diawali dengan proses akuisisi data, yaitu pengumpulan data kualitas udara indoor menggunakan sensor suhu, kelembapan, dan gas. Selanjutnya, data yang diperoleh melalui tahap pra-pemrosesan data untuk menangani nilai hilang dan perbedaan skala antar fitur. Data yang telah dipra-pemrosesan kemudian dibagi ke dalam data latih, data validasi, dan data uji menggunakan skema pembagian yang terkontrol. Setelah itu, dilakukan pelatihan model dan penyesuaian hyperparameter, di mana beberapa algoritma machine

Copyright © 2026 The Author, Page 589
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 588−598
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9024

learning dilatih dan dioptimalkan untuk memperoleh performa terbaik. Tahapan akhir adalah evaluasi model, yang dilakukan menggunakan metrik klasifikasi multikelas untuk menilai kemampuan model dalam mengklasifikasikan kualitas udara indoor.



Gambar 1. Flowchart Penelitian

2.2 Akuisisi Data

Data penelitian dikumpulkan menggunakan dua jenis sensor lingkungan, yaitu sensor DHT22 untuk mengukur suhu dan kelembapan udara, serta sensor MQ-135 untuk mendeteksi tingkat polutan udara dalam bentuk sinyal tegangan analog [9], [17]. Kedua sensor dihubungkan ke mikrokontroler ESP32 yang berfungsi sebagai unit pembacaan dan pengiriman data ke komputer melalui komunikasi serial.

Pengambilan data dilakukan pada lingkungan ruangan terkontrol dengan tiga skenario kondisi udara yang berbeda, yaitu kondisi udara baik (Good), sedang (Moderate), dan buruk (Poor). Kategori Good merepresentasikan kondisi udara normal tanpa penambahan sumber poluton, Moderate merepresentasikan kondisi dengan aktivitas manusia ringan dan ventilasi terbatas, sedangkan kategori Poor merepresentasikan kondisi udara dengan konsentrasi poluton yang lebih tinggi akibat penambahan sumber poluton seperti asap rokok, dan aktivitas pembakaran ringan di dalam ruangan.

Setiap skenario pengambilan data dilakukan dalam sesi terpisah dengan interval sampling dua detik. Secara keseluruhan, proses akuisisi data dilakukan selama kurang lebih 3,5 jam, sehingga diperoleh dataset yang merepresentasikan variasi kondisi kualitas udara indoor secara temporal. Data hasil pembacaan sensor disimpan dalam format Comma-Separated Values (CSV) menggunakan script Python untuk memastikan pencatatan data yang kontinu dan terstruktur.

Sensor MQ-135 diketahui memiliki sensitivitas terhadap perubahan suhu dan kelembapan serta memerlukan waktu pemanasan (burn-in) sebelum menghasilkan pembacaan yang stabil. Oleh karena itu, sensor MQ-135 dilakukan proses pre-heating selama kurang lebih 24 jam sebelum pengambilan data, dan data baru direkam setelah keluaran sensor menunjukkan kondisi stabil [18]. Penelitian ini tidak menerapkan kompensasi suhu dan kelembapan secara eksplisit pada pembacaan MQ-135, namun pengaruh variasi lingkungan ditangani pada tahap pra-pemrosesan data melalui normalisasi fitur.

Dataset yang dihasilkan terdiri dari tiga fitur numerik utama, yaitu suhu udara (temperature_C), kelembapan udara (humidity_%RH), dan tegangan keluaran sensor MQ-135 (mq135_V), serta satu label target berupa kelas kualitas udara. Total data yang dikumpulkan berjumlah 18.000 sampel dengan distribusi kelas yang seimbang, masing-masing 6.000 data untuk setiap kelas kualitas udara.

2.3 Pra-pemrosesan Data

Tahap pra-pemrosesan dilakukan untuk meningkatkan kualitas data sebelum digunakan pada proses pelatihan model machine learning. Penanganan nilai hilang pada data sensor dilakukan menggunakan metode median imputation, yang dipilih karena lebih robust terhadap keberadaan nilai ekstrem pada data lingkungan [19].

Selanjutnya, dilakukan proses normalisasi fitur menggunakan metode Standardization (Z-score normalization) yang mengubah setiap fitur numerik agar memiliki nilai rata-rata nol dan simpangan baku satu.

Copyright © 2026 The Author, Page 590
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 588−598
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9024

Proses ini penting karena algoritma berbasis jarak dan margin, seperti KNN dan SVM, sensitif terhadap perbedaan skala antar fitur [20], [21], [22]. Untuk mencegah terjadinya data leakage, seluruh parameter pra-pemrosesan dihitung hanya berdasarkan data latih dan kemudian diterapkan pada data validasi serta data uji. Seluruh tahapan pra-pemrosesan diimplementasikan dalam sebuah pipeline sehingga transformasi data dapat diterapkan secara konsisten pada seluruh model yang diuji. Proses standarisasi data dilakukan menggunakan Persamaan (1) sebagai berikut:

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Dimana x merupakan nilai asli fitur, μ adalah nilai rata-rata fitur, dan σ adalah simpangan baku fitur. Evaluasi performa model dilakukan menggunakan beberapa metrik klasifikasi, yaitu accuracy, precision, recall, dan F1-score, yang masing masing dihitung menggunakan Persamaan (2) – (5).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1-score = 2 \times \frac{Precision\ x\ Recall}{Precision+Recall} \quad (5)$$

Metrik-metrik tersebut digunakan untuk memberikawn evaluasi performa model secara komprehensif, khususnya pada permasalahan klasifikasi multikelas kualitas udara indoor.

2.4 Pembagian Dataset

Dataset dibagi menjadi tiga subset, yaitu data latih, data validasi, dan data uji dengan rasio masing-masing sebesar 70%, 15%, dan 15%. Pembagian data dilakukan menggunakan metode stratified sampling untuk memastikan proporsi kelas kualitas udara tetap seimbang pada setiap subset [23]. Data latih digunakan untuk melatih model, data validasi digunakan untuk pemilihan hyperparameter, sedangkan data uji digunakan secara eksklusif pada tahap evaluasi akhir guna mengukur kemampuan generalisasi model.

2.5 Pelatihan dan Penyesuaian Hyperparameter Model

Empat algoritma supervised learning digunakan dalam penelitian ini, yaitu Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), dan Gaussian Naive Bayes. Logistic Regression digunakan sebagai model baseline linear yang bersifat interpretatif [24], sedangkan KNN dipilih karena kemampuannya menangkap pola nonlinier berbasis kedekatan jarak antar data [25]. Gaussian Naive Bayes digunakan untuk mengevaluasi pendekatan probabilistik berbasis Teorema Bayes [26], dan SVM dipilih karena kemampuannya membentuk batas keputusan optimal dengan margin maksimum [27].

Pelatihan model dilakukan menggunakan data latih yang telah melalui tahap pra-pemrosesan. Untuk meningkatkan performa masing-masing algoritma, dilakukan proses penyesuaian hyperparameter menggunakan data validasi. Pada Logistic Regression, parameter regularisasi (C) dioptimalkan. Pada KNN, parameter jumlah tetangga terdekat (k), metrik jarak, dan metode pembobotan diuji. Untuk SVM, parameter C dan gamma disesuaikan, sedangkan pada Gaussian Naive Bayes, parameter var_smoothing digunakan untuk meningkatkan stabilitas perhitungan probabilitas. Model dengan performa terbaik pada data validasi dipilih sebagai model final untuk masing-masing algoritma dan selanjutnya digunakan pada tahap evaluasi data uji.

2.6 Evaluasi Model

Evaluasi performa model dilakukan menggunakan data uji yang belum pernah digunakan pada proses pelatihan maupun penyesuaian hyperparameter. Beberapa metrik evaluasi digunakan, yaitu accuracy, precision, recall, dan F1-score, untuk menilai performa klasifikasi secara komprehensif pada setiap kelas kualitas udara [28]. Selain itu, confusion matrix digunakan untuk menganalisis pola kesalahan klasifikasi yang terjadi pada masing-masing algoritma [29].

Karena permasalahan yang ditangani merupakan klasifikasi multikelas dengan tiga kelas kualitas udara (Good, Moderate, dan Poor), perhitungan nilai precision, recall, dan F1-score dilakukan menggunakan pendekatan macro-averaging dan weighted-averaging. Pendekatan macro-averaging memberikan bobot yang sama pada setiap kelas sehingga mencerminkan kemampuan model dalam mengklasifikasikan seluruh kelas secara seimbang. Sementara itu, weighted-averaging mempertimbangkan proporsi jumlah data pada masing-masing kelas, sehingga memberikan gambaran performa model secara keseluruhan dengan memperhitungkan distribusi data [30].

Pendekatan evaluasi ini memungkinkan perbandingan performa model secara objektif dan memberikan gambaran kemampuan generalisasi algoritma machine learning dalam mengklasifikasikan kualitas udara indoor berbasis sensor. Metodologi penelitian menjelaskan rancangan penelitian, prosedur penelitian (dapat dilengkapi dengan diagram), data penelitian serta pengujian atau eksperimen yang dilakukan.

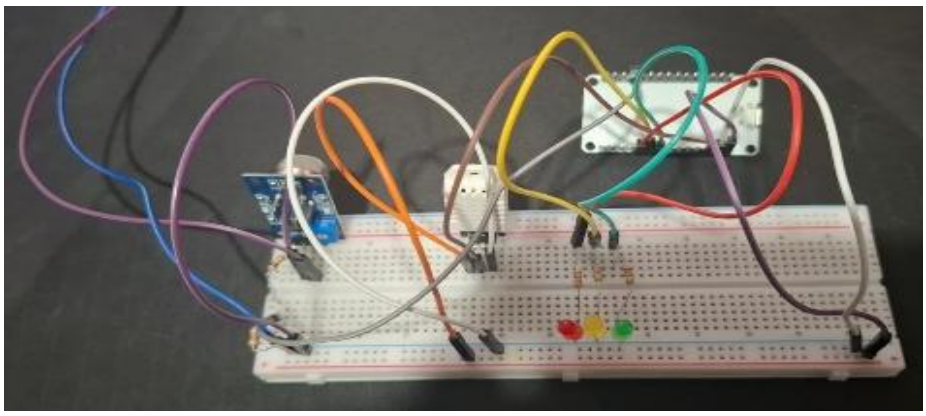
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 588−598
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9024

3. HASIL DAN PEMBAHASAN

3.1 Implementasi Sistem dan Akuisisi Data

Implementasi sistem dilakukan menggunakan mikrokontroler ESP32 yang terhubung dengan sensor DHT22 dan MQ-135 untuk membaca parameter suhu, kelembapan, dan tingkat polutan udara. Rangkaian perangkat yang digunakan pada penelitian ini ditunjukkan pada Gambar 2, yang memperlihatkan konfigurasi seluruh komponen sensor, mikrokontroler, serta rangkaian pendukung pada papan breadboard.



Gambar 2. Rangkaian Perangkat

Berdasarkan Gambar 2, sensor DHT22 digunakan untuk mengukur suhu dan kelembapan udara dan terhubung ke pin GPIO 15 digital ESP32, sedangkan sensor MQ-135 dihubungkan ke pin GPIO 34 analog ESP32 untuk membaca tegangan keluaran yang merepresentasikan tingkat polutan udara. Sensor DHT22 memperoleh catu daya dari ESP32, sedangkan sensor MQ-135 memperoleh catu daya dari power supply 2A eksternal. Konfigurasi ini memungkinkan sistem membaca parameter lingkungan secara simultan dalam satu platform pengukuran terintegrasi.

Proses akuisisi data dilakukan secara real-time dengan interval pengambilan sampel dua detik. Data hasil pembacaan sensor dikirimkan ke komputer melalui komunikasi serial. Mekanisme akuisisi dan pencatatan dataset ditunjukkan pada Gambar 3, yang menggambarkan alur perangkat lunak pada sisi komputer selama proses pengambilan data.

Gambar 3. Proses Akuisisi Dataset

Alur akuisisi dan pencatatan dataset ditunjukkan pada Gambar 3. Pada alur tersebut, data sensor yang diterima komputer selanjutnya disimpan secara otomatis dalam berkas CSV sebagai dataset terstruktur. Dataset inilah yang digunakan pada tahap pra-pemrosesan dan pelatihan model machine learning, sehingga proses pengambilan data dapat berlangsung secara kontinu tanpa intervensi manual.

3.2 Dataset dan Proses Pelabelan

Data yang diperoleh dari proses akuisisi awalnya tersimpan dalam bentuk dataset mentah (raw dataset) tanpa label kelas kualitas udara. Contoh dataset mentah ditunjukkan pada Gambar 4, yang menampilkan hasil pembacaan langsung sensor sebelum dilakukan proses pelabelan.

	A	B	C	D
1	temperature_C,humidity_%,mq135_V			
2	28.20,57.90,0.437			
3	28.20,58.00,0.443			
4	28.20,58.10,0.439			
5	28.20,58.10,0.440			
6	28.20,58.10,0.440			

Gambar 4. Raw Dataset

Berdasarkan Gambar 4, setiap baris data terdiri dari tiga atribut numerik utama, yaitu suhu udara (temperature_C), kelembapan udara (humidity_%), dan tegangan keluaran sensor MQ-135 (mq135_V). Dataset pada tahap ini belum memiliki informasi kelas kualitas udara dan masih merepresentasikan nilai sensor hasil pembacaan langsung dari sistem akuisisi data.

Copyright © 2026 The Author, Page 592
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 588−598
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9024

Selanjutnya, dilakukan proses pelabelan data berdasarkan skenario pengambilan data untuk menghasilkan dataset berlabel dengan tiga kelas kualitas udara, yaitu Good, Moderate, dan Poor. Proses pelabelan dilakukan dengan menambahkan satu atribut target berupa AQ_Label pada setiap baris data. Contoh dataset setelah proses pelabelan ditunjukkan pada Gambar 5.

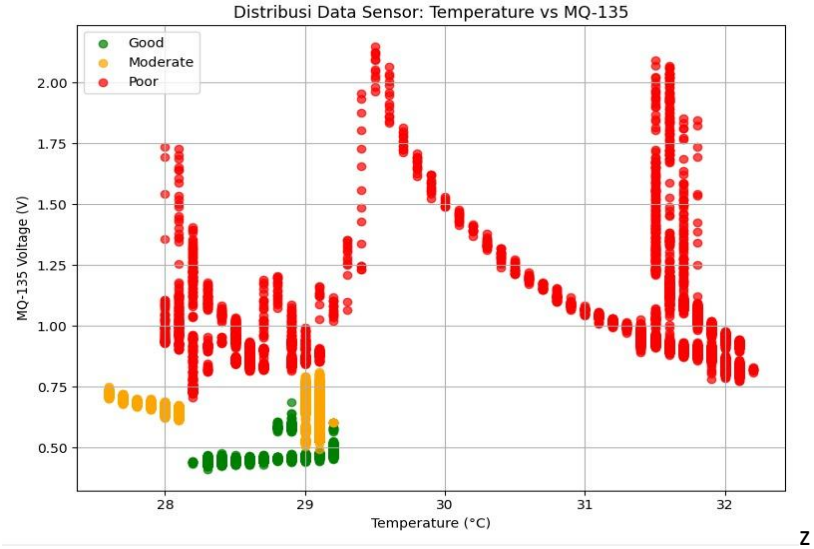
	A	B	C	D	E
1	temperature_C,humidity_%,mq135_V,AQ_Label				
2	28.5,61.9,0.447,Good				
3	28.7,63.6,0.454,Good				
4	31.6,59.2,1.098,Poor				
5	28.0,67.7,0.653,Moderate				
6	32.1,58.1,0.831,Poor				
7	28.5,61.8,0.452,Good				
8	28.9,66.7,0.457,Good				
9	29.1,77.6,0.598,Moderate				
10	29.1,77.2,0.572,Moderate				

Gambar 5. Labeled Dataset

Pada Gambar 5, terlihat bahwa setiap data sensor telah dilengkapi dengan label kelas kualitas udara yang sesuai dengan kondisi lingkungan saat data dikumpulkan. Dataset berlabel ini selanjutnya digunakan sebagai masukan pada tahap pra-pemrosesan data dan pelatihan model machine learning.

3.3 Visualisasi dan Pembagian Dataset

Untuk memastikan distribusi data yang seimbang antar kelas serta memahami pola sebaran data sensor, dilakukan visualisasi dataset menggunakan diagram scatter plot. Visualisasi distribusi data ditunjukkan pada Gambar 6, yang memperlihatkan hubungan antara nilai suhu udara (temperature_C) dan tegangan keluaran sensor MQ-135 (mq135_V) untuk setiap kelas kualitas udara.



Gambar 6. Diagram scatter plot pembagian dataset

Berdasarkan Gambar 6, terlihat bahwa data dari masing-masing kelas kualitas udara membentuk klaster yang relatif terpisah. Kelas Good cenderung berada pada rentang tegangan MQ-135 yang lebih rendah, yang merepresentasikan kondisi udara dengan konsentrasi polutan rendah. Kelas Moderate berada pada rentang tegangan menengah, sedangkan kelas Poor menunjukkan nilai tegangan MQ-135 yang lebih tinggi dan variasi yang lebih luas, yang mengindikasikan peningkatan konsentrasi polutan udara.

Selain itu, visualisasi pada Gambar 6 menunjukkan bahwa meskipun terdapat sedikit tumpang tindih antar kelas pada rentang suhu tertentu, pemisahan klaster secara umum masih dapat diamati dengan jelas. Kondisi ini menunjukkan bahwa kombinasi parameter suhu dan tegangan sensor MQ-135 memiliki kemampuan diskriminatif yang baik dalam membedakan kelas kualitas udara. Visualisasi ini sekaligus memberikan indikasi awal bahwa dataset yang digunakan memiliki separasi kelas yang cukup jelas, sehingga mendukung penggunaan algoritma machine learning untuk tugas klasifikasi kualitas udara indoor.

3.4 Hasil Pelatihan dan Hyperparameter Tuning Model

Pelatihan model dilakukan dengan terlebih dahulu melakukan penyesuaian hyperparameter menggunakan data validasi untuk memperoleh konfigurasi parameter terbaik dari masing-masing algoritma. Hasil proses hyperparameter tuning dan pelatihan untuk model Logistic Regression ditunjukkan pada Gambar 7.

Copyright © 2026 The Author, Page 593
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 588−598
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9024

```
================================================================================
📈 MODEL: Logistic Regression
================================================================================

   Best Hyperparameters:
      • model__C: 100
      • model__penalty: l2
      • model__solver: liblinear

   📊 Validation Accuracy: 0.9852
   📊 Test Set Metrics:
      • Accuracy:  0.9796
      • Precision: 0.9800
      • Recall:    0.9796
      • F1-score:  0.9796
      • Val-Test Gap: +0.0056 (✅ Excellent generalization)

   📋 Confusion Matrix:
[[858  42   0]
 [ 13 887   0]
 [  0   0 900]]

   📈 Classification Report (per class):
              precision    recall  f1-score   support

        Good     0.9851    0.9533    0.9689       900
    Moderate     0.9548    0.9856    0.9699       900
        Poor     1.0000    1.0000    1.0000       900
```

Gambar 7. Hasil tuning dan pelatihan Logistic Regression

Berdasarkan Gambar 7, model Logistic Regression menggunakan parameter terbaik dengan nilai regularisasi yang relatif besar, yang memungkinkan model membentuk batas keputusan linear yang stabil. Model ini menghasilkan nilai akurasi sebesar 0,9796 dan F1-score sebesar 0,9796 pada data uji. Confusion matrix menunjukkan bahwa sebagian kecil kesalahan klasifikasi terjadi antara kelas Good dan Moderate, sementara kelas Poor dapat diklasifikasikan secara sempurna. Hasil ini menunjukkan bahwa pendekatan linear masih cukup efektif dalam merepresentasikan hubungan antar fitur sensor dan kelas kualitas udara indoor.

Hasil tuning dan pelatihan model K-Nearest Neighbors (KNN) ditunjukkan pada Gambar 8, yang memperlihatkan parameter terbaik berupa jumlah tetangga terdekat dan metrik jarak yang digunakan pada tahap evaluasi akhir.

```
================================================================================
📈 MODEL: KNN
================================================================================

   Best Hyperparameters:
      • model__metric: manhattan
      • model__n_neighbors: 7
      • model__weights: uniform

   📊 Validation Accuracy: 1.0000
   📊 Test Set Metrics:
      • Accuracy:  1.0000
      • Precision: 1.0000
      • Recall:    1.0000
      • F1-score:  1.0000
      • Val-Test Gap: +0.0000 (✅ Excellent generalization)

   📋 Confusion Matrix:
[[900   0   0]
 [  0 900   0]
 [  0   0 900]]

   📈 Classification Report (per class):
              precision    recall  f1-score   support

        Good     1.0000    1.0000    1.0000       900
    Moderate     1.0000    1.0000    1.0000       900
        Poor     1.0000    1.0000    1.0000       900
```

Gambar 8. Hasil tuning dan pelatihan KNN

Berdasarkan Gambar 8, model KNN mencapai performa terbaik dibandingkan algoritma lainnya dengan nilai akurasi dan F1-score sebesar 1,0000 pada data uji. Confusion matrix menunjukkan bahwa seluruh data uji berhasil diklasifikasikan dengan benar ke dalam kelas Good, Moderate, dan Poor. Performa sempurna ini dipengaruhi oleh karakteristik dataset yang memiliki pemisahan klaster antar kelas yang sangat jelas, sehingga pendekatan berbasis jarak yang digunakan oleh KNN mampu menentukan kelas data uji secara akurat berdasarkan kedekatan terhadap data latih. Hasil tuning dan pelatihan model Support Vector Machine (SVM) ditunjukkan pada Gambar 9.

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 588−598
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9024

```
================================================================================
📈 MODEL: SVM (RBF)
================================================================================

   Best Hyperparameters:
      • model__C: 100
      • model__gamma: scale
      • model__kernel: rbf

   📊 Validation Accuracy: 0.9926
   📊 Test Set Metrics:
      • Accuracy:  0.9907
      • Precision: 0.9908
      • Recall:    0.9907
      • F1-score:  0.9907
      • Val-Test Gap: +0.0019 (✅ Excellent generalization)

   📋 Confusion Matrix:
[[890  10   0]
 [ 15 885   0]
 [  0   0 900]]

   📈 Classification Report (per class):
              precision    recall  f1-score   support

        Good     0.9834    0.9889    0.9861       900
    Moderate     0.9888    0.9833    0.9861       900
        Poor     1.0000    1.0000    1.0000       900
```

Gambar 9. Hasil tuning dan pelatihan SVM

Berdasarkan Gambar 9, model SVM dengan kernel Radial Basis Function (RBF) menghasilkan performa yang sangat tinggi dengan nilai akurasi sebesar 0,9907 dan F1-score sebesar 0,9907. Confusion matrix menunjukkan bahwa sebagian kecil kesalahan klasifikasi masih terjadi antara kelas Good dan Moderate. Hal ini mengindikasikan bahwa meskipun SVM mampu membentuk batas keputusan dengan margin maksimum, pendekatan berbasis margin ini sedikit kurang optimal dibandingkan KNN pada dataset dengan separasi klaster yang sangat jelas. Hasil tuning dan pelatihan model Gaussian Naive Bayes ditunjukkan pada Gambar 10.

```
================================================================================
📈 MODEL: Naive Bayes
================================================================================

   Best Hyperparameters:
      • model__var_smoothing: 1e-09

   📊 Validation Accuracy: 0.9770
   📊 Test Set Metrics:
      • Accuracy:  0.9696
      • Precision: 0.9716
      • Recall:    0.9696
      • F1-score:  0.9696
      • Val-Test Gap: +0.0074 (✅ Excellent generalization)

   📋 Confusion Matrix:
[[824  76   0]
 [  3 894   3]
 [  0   0 900]]

   📈 Classification Report (per class):
              precision    recall  f1-score   support

        Good     0.9964    0.9156    0.9543       900
    Moderate     0.9216    0.9933    0.9561       900
        Poor     0.9967    1.0000    0.9983       900
```

Gambar 10. Hasil tuning dan pelatihan Naive Bayes

Berdasarkan Gambar 10, model Gaussian Naive Bayes menghasilkan performa yang relatif lebih rendah dibandingkan algoritma lainnya dengan nilai akurasi sebesar 0,9696 dan F1-score sebesar 0,9696. Confusion matrix menunjukkan bahwa kesalahan klasifikasi masih terjadi terutama antara kelas Good dan Moderate. Hal ini disebabkan oleh asumsi independensi antar fitur pada Naive Bayes yang tidak sepenuhnya sesuai dengan karakteristik data sensor lingkungan. Meskipun demikian, keunggulan utama Naive Bayes terletak pada kesederhanaan model dan efisiensi komputasi, sehingga berpotensi digunakan pada sistem dengan keterbatasan sumber daya.

3.5 Analisis Performa Model

Evaluasi performa model dilakukan menggunakan data uji yang belum pernah digunakan pada proses pelatihan maupun penyesuaian hyperparameter. Hasil evaluasi menunjukkan bahwa seluruh algoritma mampu mencapai

Copyright © 2026 The Author, Page 595
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 588−598
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9024

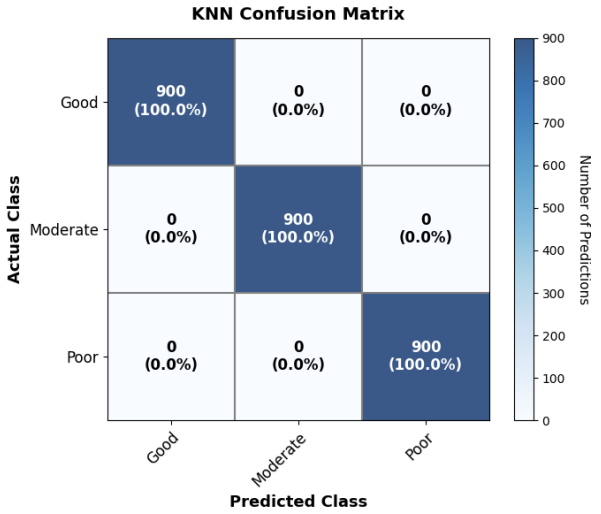
performa klasifikasi yang tinggi. Ringkasan perbandingan performa keempat model disajikan pada Tabel 1, yang menyajikan nilai accuracy, precision, recall, dan F1-score untuk masing-masing algoritma.

Tabel 1. Perbandingan performa model

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0,9796	0,9800	0,9796	0,9796
KNN	1,0000	1,0000	1,0000	1,0000
SVM	0,9907	0,9908	0,9907	0,9907
Naive Bayes	0,9696	0,9716	0,9696	0,9696

Berdasarkan Tabel 1, model K-Nearest Neighbors (KNN) menunjukkan performa terbaik dengan nilai akurasi dan F1-score sebesar 1,0000. Model Support Vector Machine (SVM) berada pada urutan berikutnya dengan nilai akurasi sebesar 0,9907, diikuti oleh Logistic Regression dengan akurasi 0,9796. Sementara itu, Gaussian Naive Bayes menghasilkan performa paling rendah di antara keempat model, meskipun tetap berada pada tingkat akurasi yang tinggi. Perbedaan performa ini menunjukkan bahwa karakteristik dataset lebih mendukung pendekatan berbasis jarak dibandingkan pendekatan linear maupun probabilistik.

Selain evaluasi metrik agregat, dilakukan pula analisis kesalahan klasifikasi menggunakan confusion matrix. Confusion matrix untuk model KNN ditunjukkan pada Gambar 11, yang digunakan untuk menganalisis pola prediksi model secara lebih rinci pada setiap kelas kualitas udara.



Gambar 11. Confusion matrix KNN

Berdasarkan Gambar 11, terlihat bahwa seluruh data uji pada kelas Good, Moderate, dan Poor berhasil diklasifikasikan dengan benar tanpa adanya kesalahan prediksi antar kelas. Setiap kelas memiliki nilai prediksi yang sepenuhnya berada pada diagonal utama confusion matrix, yang mengindikasikan bahwa model KNN mampu memisahkan ketiga kelas kualitas udara secara sempurna pada data uji. Hasil ini memperkuat temuan bahwa model KNN sangat efektif digunakan pada dataset dengan tingkat pemisahan klaster antar kelas yang jelas.

3.6 Pembahasan

Performa klasifikasi yang sangat tinggi pada seluruh model machine learning, khususnya pada KNN, sangat dipengaruhi oleh karakteristik dataset yang digunakan. Data dikumpulkan dalam skenario lingkungan terkontrol dengan perbedaan kondisi kualitas udara yang kontras antar kelas. Hal ini menyebabkan nilai tegangan keluaran sensor MQ-135 memiliki rentang nilai yang jauh lebih besar dan fitur yang lebih diskriminatif dibandingkan suhu dan kelembapan, sehingga menjadikan klaster antar kelas menjadi lebih terpisah secara jelas sebagaimana ditunjukkan pada visualisasi distribusi data.

Selain itu, analisis jarak antar klaster menunjukkan bahwa variasi nilai MQ-135 antara skenario Good dan Poor jauh lebih signifikan dibandingkan variasi suhu atau kelembapan. Dalam konteks klasifikasi berbasis jarak, seperti KNN, perbedaan jarak antar klaster yang besar meningkatkan kemampuan model untuk mengidentifikasi batas antar kelas tanpa tumpang tindih. Kondisi tersebut memperkuat bahwa performa sempurna yang dicapai oleh model KNN merupakan konsekuensi dari struktur natural dataset, bukan karena kesalahan metodologi seperti data leakage.

Algoritma berbasis jarak seperti KNN sangat diuntungkan dalam kondisi dataset dengan pemisahan klaster yang jelas karena proses klasifikasi dilakukan berdasarkan kedekatan data uji terhadap data latih. Ketika jarak antar kelas cukup besar dan tumpang tindih data minimal, KNN mampu menentukan kelas dengan tingkat

Copyright © 2026 The Author, Page 596
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 588−598
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9024

keyakinan yang tinggi. Hal ini menjelaskan mengapa model KNN mampu mencapai performa sempurna tanpa mengalami penurunan performa antara data validasi dan data uji.

Meskipun hasil yang diperoleh menunjukkan performa yang sangat baik, perlu dicatat bahwa hasil ini merepresentasikan kondisi pengujian yang bersifat terkontrol. Pada implementasi di lingkungan nyata, kualitas udara cenderung berubah secara gradual dan menghasilkan data dengan karakteristik yang lebih tumpang tindih antar kelas. Oleh karena itu, performa model pada kondisi dunia nyata berpotensi lebih rendah dibandingkan hasil yang diperoleh pada penelitian ini.

Jika dibandingkan dengan studi terdahulu, sebagian penelitian yang menggunakan algoritma serupa pada dataset kualitas udara yang tidak sepenuhnya terkontrol sering kali melaporkan akurasi yang lebih rendah atau variasi performa yang lebih besar antar kelas. Perbedaan ini sering disebabkan oleh data yang lebih berisik, variasi kondisi lingkungan yang luas, dan kompleksitas hubungan antar fitur yang lebih tinggi. Dalam konteks tersebut, hasil KNN yang sempurna pada penelitian ini mencerminkan efektivitas model pada dataset dengan distribusi klaster yang jelas, tetapi bukan jaminan performa pada skenario dunia nyata yang lebih bervariasi.

Dalam hal kontribusi fitur, meskipun semua fitur (suhu, kelembapan, dan tegangan MQ-135) berkontribusi pada klasifikasi, sensor MQ-135 menunjukkan kontribusi yang lebih dominan dalam memisahkan kelas kualitas udara. Rentang nilai MQ-135 yang lebih luas, serta perbedaan nilai yang lebih tajam antar skenario Good, Moderate, dan Poor, menjadi indikator bahwa informasi dari sensor ini sangat berperan dalam membentuk batas keputusan yang jelas antara kelas-kelas tersebut. Sebaliknya, variabilitas suhu dan kelembapan relatif lebih kecil dan memiliki kontribusi sekunder dalam membedakan kelas.

4. KESIMPULAN

Penelitian ini melakukan analisis komparatif terhadap empat algoritma machine learning—Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), dan Gaussian Naive Bayes—dalam mengklasifikasikan kualitas udara indoor berbasis data sensor suhu, kelembapan, dan gas pada lingkungan terkontrol dengan tiga kelas kualitas udara, yaitu Good, Moderate, dan Poor. Hasil evaluasi menunjukkan bahwa seluruh model mampu mencapai performa klasifikasi yang tinggi dengan rentang akurasi mulai dari 0,9696 (Gaussian Naive Bayes) hingga 1,0000 (KNN). Model KNN menunjukkan performa terbaik dengan nilai akurasi dan F1-score sempurna, yang dipengaruhi oleh karakteristik dataset dengan pemisahan klaster antar kelas yang jelas. Temuan ini mengindikasikan bahwa algoritma berbasis jarak sangat efektif digunakan pada dataset kualitas udara indoor dengan separasi kelas yang kuat. Meskipun demikian, karena pengujian dilakukan pada kondisi lingkungan yang bersifat terkontrol, penelitian selanjutnya disarankan untuk mengevaluasi performa model pada kondisi lingkungan yang lebih dinamis guna menguji kemampuan generalisasi. Secara keseluruhan, penelitian ini memberikan kontribusi sebagai referensi dalam pemilihan algoritma machine learning yang tepat untuk sistem klasifikasi kualitas udara indoor berbasis sensor low-cost.

REFERENCES

[1] H. Budianto and B. Sumanto, “Perancangan Sistem Monitoring Kualitas Udara dalam Ruangan Berbasis Internet of Things,” Jurnal Listrik, Instrumentasi, dan Elektronika Terapan, vol. 5, no. 1, 2024, doi: https://doi.org/10.22146/juliet.v5i1.87423.

[2] S. M. Nasri, A. D. Athari, L. R. Hastiti, and F. A. Putri, “Indoor Air Factors Affecting the Growth of Microorganism in an Indonesian Gas Company’s Dormitory,” Indonesian Journal of Occupational Safety and Health, vol. 11, no. 3, pp. 445–453, Nov. 2022, doi: 10.20473/ijosh.v11i3.2022.445-453.

[3] W. Indah, D. Aurora, F. Kedokteran, and D. I. Kesehatan, “EFEK INDOOR AIR POLLUTION TERHADAP KESEHATAN,” Scientific Of Environmental Health and Diseases, vol. 1, no. 2, 2021, doi: https://doi.org/10.22437/esehad.v2i1.13750.

[4] M. Mannan and S. G. Al-Ghamdi, “Indoor air quality in buildings: A comprehensive review on the factors influencing air pollution in residential and commercial structure,” Mar. 02, 2021, MDPI AG. doi: 10.3390/ijerph18063276.

[5] M. Gurung, S. Kaini, K. Bhandari, I. Panta, and N. Singh, “IoT Based Automatic Air Pollution Monitoring and Purification System,” International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 9, no. 9, pp. 2321–9653, 2021, doi: 10.22214/ijraset.2021.38039.

[6] T. Nie, G. Zhang, Y. Sun, W. Wang, T. Wang, and H. Duan, “Effects of Indoor Air Quality on Human Physiological Impact: A Review,” Buildings, vol. 15, no. 8, Apr. 2025, doi: 10.3390/buildings15081296.

[7] P. Yushananta, “Very Low-Cost, Internet of Things (IoT) Air Quality Monitoring Platform,” Jurnal Aisyah : Jurnal Ilmu Kesehatan, vol. 8, no. 2, Apr. 2023, doi: 10.30604/jika.v8i2.1919.

[8] M. T. Ilham Ashiddiq, matul Ma, and F. Mayanti, “Prototipe Sistem Pendeteksi Gas dan Api Berbasis Android Prototype of Android-Based Gas and Fire Detector System,” Semnas TE UIN Bandung 2019, pp. 23–24, 2019.

[9] R. Muttaqin, W. Sakti, W. Prayitno, N. E. Setyaningsih, and U. Nurbaiti, “Rancang Bangun Sistem Pemantauan Kualitas Udara Berbasis Iot (Internet Of Things) dengan Sensor DHT11 dan Sensor MQ135,” Jurnal Pengelolaan Laboratorium Pendidikan, vol. 6, no. 2, pp. 2654–251, 2024, doi: https://doi.org/10.14710/jplp.6.2.102-115.

[10] M. Vogt, P. Schneider, N. Castell, and P. Hamer, “Assessment of low-cost particulate matter sensor systems against optical and gravimetric methods in a field co-location in norway,” Atmosphere (Basel), vol. 12, no. 8, Aug. 2021, doi: 10.3390/atmos12080961.

Copyright © 2026 The Author, Page 597
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 588−598
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9024

[11] M. Kumar Gajendran, I. Fazil Syed Ahmed Kabir, S. Vadivelu, E. Yin-Kwee Ng, and R. Chandra Thota, "Machine Learning Techniques in Indoor Environmental Quality Assessment," Advancements in Indoor Environmental Quality and Health, 2023, doi: https://doi.org/10.5772/intechopen.114012.
[12] P. Mottahedin, B. Chahkandi, R. Moezzi, A. M. Fathollahi-Fard, M. Ghandali, and M. Gheibi, "Air quality prediction and control systems using machine learning and adaptive neuro-fuzzy inference system," Heliyon, vol. 10, no. 21, Nov. 2024, doi: 10.1016/j.heliyon.2024.e39783.
[13] M. M. Rahman et al., "AirNet: predictive machine learning model for air quality forecasting using web interface," Environmental Systems Research, vol. 13, no. 1, Dec. 2024, doi: 10.1186/s40068-024-00378-z.
[14] N. H. S. Alani, P. Chand, and M. Al-Rawi, "A Two-Stage Machine Learning Framework for Air Quality Prediction in Hamilton, New Zealand," Environments - MDPI, vol. 12, no. 9, Sep. 2025, doi: 10.3390/environments12090336.
[15] Y. H. Kim and S. H. Moon, "Machine Learning-Based Quality Control for Low-Cost Air Quality Monitoring: A Comprehensive Review of the Past Decade," Atmosphere (Basel), vol. 16, no. 10, Oct. 2025, doi: 10.3390/atmos16101136.
[16] D. Latoń, J. Grela, A. Ożadowicz, and L. Wisniewski, "Artificial Intelligence and Machine Learning Approaches for Indoor Air Quality Prediction: A Comprehensive Review of Methods and Applications," Energies (Basel), vol. 18, no. 19, Oct. 2025, doi: 10.3390/en18195194.
[17] R. B. Prototipe, D. Suhu, K. Dan, K. Udara Menggunakan Sensor, S. Surya, and J. Waworundeng, "Design Prototype Detector of Temperature, Humidity, and Air Quality using Sensors, Microcontrollers, Solar Cells, and IoT," Cogito Smart Journal |, vol. 9, no. 2, 2023.
[18] M. Kumar, G. Mishra, A. Sharma, A. Shaini, and S. Saxena, "Air Quality Monitoring Using MQ135 Gas Sensor and Arduino Uno," International Journal of Latest Technology in Engineering Management & Applied Science, vol. 14, no. 5, May 2025, doi: 10.51583/IJLTEMAS.
[19] H. Chun and J. Cho, "Missing-value Imputation of Environment Sensors Using Multilayer Stacking with Scoring Method," IEEE Access, vol. 13, Jun. 2023, doi: 10.21203/rs.3.rs-3050822/v1.
[20] J. M. H. Pinheiro et al., "The Impact of Feature Scaling In Machine Learning: Effects on Regression and Classification Tasks," IEEE Access, vol. 13, Nov. 2025, doi: 10.1109/ACCESS.2025.3635541.
[21] A. Dzaky and R. Hasudungan, "KLASIFIKASI KUALITAS UDARA BERBASIS IOT DAN ALGORITMA K-NEAREST NEIGHBORS (KNN)," Jurnal Ilmu Komputer dan Sistem Informasi (JIKSI), vol. 11, no. 9, pp. 54–66, 2025.
[22] A. S. Handayani, S. Soim, T. E. Agusdi, and N. L. Husni, "Air Quality Classification Using Support Vector Machine," Computer Engineering and Applications, vol. 10, no. 1, 2021.
[23] Analyttica Datalab, "What is meant by 'Stratified Split'? | by Analyttica Datalab | Medium," https://medium.com/@analyttica/what-is-meant-by-stratified-split-289a8a986a90. Accessed: Dec. 23, 2025. [Online]. Available: https://medium.com/@analyttica/what-is-meant-by-stratified-split-289a8a986a90
[24] A. M. Vélez-Pereira, N. Núñez-Magaña, D. Barreau, K. Bremer, and D. J. O'Connor, "Simplifying Air Quality Forecasting: Logistic Regression for Predicting Particulate Matter in Chile," Atmosphere (Basel), vol. 16, no. 12, p. 1377, Dec. 2025, doi: 10.3390/atmos16121377.
[25] K. Eldora, E. Fernando, and W. Winanti, "Comparative Analysis of KNN and Decision Tree Classification Algorithms for Early Stroke Prediction: A Machine Learning Approach," Journal of Information Systems and Informatics, vol. 6, no. 1, pp. 313–338, Mar. 2024, doi: 10.51519/journalisi.v6i1.664.
[26] Nimatul Mamuriyah, Haeruddin Haeruddin, and H. Hero, "PEMBANGUNAN CHATBOT INTERAKTIF DENGAN MENGGUNAKAN ALGORITMA NAIVE BAYES," Informatika: Jurnal Teknik Informatika dan Multimedia, vol. 4, no. 2, pp. 82–94, Dec. 2024, doi: 10.51903/informatika.v4i2.864.
[27] T. Tan, H. Sama, G. Wijaya, and O. E. Aboagye, "Studi Perbandingan Deteksi Intrusi Jaringan Menggunakan Machine Learning: (Metode SVM dan ANN) Comparative Study of Network Intrusion Detection Using Machine Learning: (SVM and ANN Method)," Jurnal Teknologi dan Informasi, vol. 13, no. 2, 2023, doi: 10.34010/jati.v13i2.
[28] "Memahami Confusion Matrix: Accuracy, Precision, Recall, Specificity, dan F1-Score untuk Evaluasi Model Klasifikasi | by Rina | Medium," https://esairina.medium.com/memahami-confusion-matrix-accuracy-precision-recall-specificity-dan-f1-score-610d4f0db7cf. Accessed: Dec. 23, 2025. [Online]. Available: https://esairina.medium.com/memahami-confusion-matrix-accuracy-precision-recall-specificity-dan-f1-score-610d4f0db7cf
[29] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," African Journal of Biomedical Research, vol. 27, no. 4, pp. 4023–4031, Nov. 2024, doi: 10.53555/ajbr.v27i4s.4345.
[30] C. Trois, L. D. Del Fabro, and V. A. Baulin, "Machine learning unveils large-scale impact of Posidonia oceanica on Mediterranean Sea water," Science of the Total Environment, vol. 968, Mar. 2025, doi: 10.1016/j.scitotenv.2025.178802.

Copyright © 2026 The Author, Page 598
This Journal is licensed under a Creative Commons Attribution 4.0 International License