JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 160-168
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Aplikasi Monitor dan Prediksi Tingkat Polusi Udara Berbasis Web dengan Streamlit

Muhammad Ikmal Tawakal[1*], Dony Marulitua Simanungkalit[2], Yohanes Septian Hildegardis[3], Reno Nurhadi[4], Lisnawanty[5]
[1,2,3,4,5]Informatika, Unversitas Bina Sarana Informatika, Pontianak, Indonesia
Email: [1]15230786@bsi.ac.id, [2]15230727@bsi.ac.id, [3]15230781@bsi.ac.id, [4]15230733@bsi.ac.id, [5]lisnawanty.lsy@bsi.ac.id
Email Penulis Korespondensi: [1]15230786@bsi.ac.id

Abstrak- Penelitian ini mengembangkan model *machine learning* untuk memprediksi konsentrasi Benzena C6H6(GT) dan merancang aplikasi web berbasis *Streamlit* untuk monitoring serta prediksi kualitas udara secara *real-time*. Dataset *Air Quality* dari UCI (9.357 rekaman per jam, Maret 2004–Februari 2005) digunakan dengan pembersihan nilai hilang, penghapusan NMHC(GT) yang dominan hilang, interpolasi berbasis waktu, serta rekayasa fitur waktu dan meteorologi. Pemodelan regresi membandingkan *Linear Regression*, *Random Forest*, dan XGBoost menggunakan pembagian data 80/20, standardisasi fitur, dan evaluasi dengan MAE, MSE, serta *R-squared*. Hasil menunjukkan *Random Forest* sebagai model terbaik dengan MAE 1.1444, MSE 3.5262, dan R^2 0.8900, mengungguli XGBoost dan *Linear Regression*. Model terbaik diintegrasikan ke aplikasi *Streamlit* sehingga pengguna dapat memasukkan parameter seperti suhu dan kelembapan dan memperoleh prediksi konsentrasi Benzena secara langsung melalui antarmuka yang ramah pengguna. Temuan ini menegaskan bahwa kombinasi fitur waktu dan cuaca efektif untuk memodelkan polutan dan implementasi web memungkinkan pemanfaatan praktis untuk pemantauan serta pengambilan keputusan. Secara keseluruhan, solusi ini valid secara prediktif dan layak diterapkan sebagai alat bantu fungsional untuk memonitor tingkat polusi udara.

Kata Kunci: Kualitas udara; Benzena (C6H6); Random Forest; XGBoost; Streamlit.

Abstract− This research develops a machine learning model to predict the concentration of Benzene C6H6(GT) and designs a Streamlit-based web application for real-time air quality monitoring and prediction. The UCI Air Quality dataset (9,357 hourly records, March 2004–February 2005) was used with missing value cleaning, removal of dominant missing NMHC(GT) values, time-based interpolation, and time and meteorological feature engineering. Regression modeling compared Linear Regression, Random Forest, and XGBoost using 80/20 data splitting, feature standardization, and evaluation with MAE, MSE, and R-squared. Results show Random Forest as the best model with MAE 1.1444, MSE 3.5262, and R^2 0.8900, outperforming XGBoost and Linear Regression. The best model is integrated into a Streamlit application so users can input parameters such as temperature and humidity and obtain Benzene concentration predictions directly through a user-friendly interface. These findings confirm that the combination of time and weather features is effective for modeling pollutants, and web implementation enables practical use for monitoring and decision making. Overall, this solution is predictively valid and feasible as a functional tool for monitoring air pollution levels.

Keywords: Air quality; Benzene (C6H6); Random Forest; XGBoost; Streamlit.

1. PENDAHULUAN

Polusi udara merupakan ancaman serius terhadap kesehatan masyarakat global. Menurut *World Health Organization* (WHO), sekitar 7 juta kematian prematur setiap tahun disebabkan oleh paparan polusi udara. Di Indonesia, kualitas udara di kota-kota besar seperti Jakarta, Surabaya, dan Bandung sering melampaui standar baku mutu, terutama pada musim kemarau dan saat kebakaran hutan [1], [2], [3], [4]. Faktor meteorologi seperti suhu, kelembaban, dan kecepatan angin turut berkontribusi terhadap fluktuasi tingkat polusi, sehingga pemantauan real-time dan prediksi polusi sangat penting untuk mendukung kebijakan publik dan peringatan dini [5], [6], [7].

Pendekatan tradisional dalam pemodelan kualitas udara, seperti model regresi statistik dan model transportasi kimia atmosfer, seringkali memiliki keterbatasan dalam menangkap hubungan nonlinear yang kompleks antara polutan dan faktor meteorologi. Selain itu, model-model tersebut membutuhkan sumber daya komputasi yang besar dan data input yang sangat detail, sehingga kurang efisien untuk implementasi *real-time* [1].

Tiga algoritma *machine learning* digunakan dalam penelitian ini. **Random Forest**, metode pengajaran kelompok yang menggabungkan banyak pohon keputusan, penelitian sebelumnya menunjukkan prediksi yang akurat dan tahan terhadap *overfitting* dengan akurasi hingga 98,2% [1], [5]; **XGBoost**, algoritma gradient boosting yang efisien dengan teknik regularisasi untuk mencegah overfitting [2]; dan **Regresi Linear**, model statistik klasik sebagai baseline untuk perbandingan [1].

Implementasi model melalui **Streamlit**, *framework* Python untuk pengembangan aplikasi web interaktif tanpa pengetahuan mendalam tentang *web development*, memungkinkan integrasi model prediksi ke antarmuka ramah pengguna untuk monitoring dan prediksi *real-time* [7].

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah Bagaimana model *machine learning* (Random Forest, XGBoost, dan Regresi Linear) dapat digunakan untuk memprediksi tingkat polusi udara

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **18/01/2026**; Published: **27/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 160-168
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

(PM2.5, PM10, CO, NO_2, dll.) di wilayah tertentu, Bagaimana model prediksi tersebut dapat diimplementasikan menjadi aplikasi web interaktif menggunakan Streamlit sebagai antarmuka monitoring dan prediksi real-time.

Tujuan dari penelitian ini adalah Menganalisis dan membandingkan kinerja algoritma Random Forest, XGBoost, dan Regresi Linear dalam memprediksi tingkat polusi udara berdasarkan data ***Air Quality Dataset***, Mengembangkan dan mengimplementasikan aplikasi berbasis web menggunakan Streamlit sebagai antarmuka sistem monitoring dan prediksi kualitas udara secara real-time.

2. METODOLOGI PENELITIAN

2.1 Persiapan Data

Data yang digunakan dalam penelitian ini adalah dataset ***Air Quality Data Set***. Dataset ini merupakan data sekunder yang diperoleh dari repositori *online* Kaggle (*https://www.kaggle.com/datasets/fedesoriano/air-quality-data-set*), yang diunggah oleh pengguna Federico Soriano.

Dataset ini aslinya berasal dari ***UCI Machine Learning Repository*** dan berisi 9.357 rekaman data. Data ini merupakan respons rata-rata per jam (*hourly averaged responses*) dari 5 sensor kimia *metal oxide* yang ditempatkan di lapangan (area dengan polusi tinggi di sebuah kota di Italia). Periode pengambilan data berlangsung selama satu tahun, yaitu dari Maret 2004 hingga Februari 2005.

Tabel 1. Deskripsi Atribut Dataset Kualitas Udara

Nama Atribut	Tipe Data	Deskripsi
Date	Teks	Tanggal pencatatan data (format DD/MM/YYYY)
Time	Teks	Waktu pencatatan data (format HH.MM.SS)
CO(GT)	Numerik	Konsentrasi Karbon Monoksida (CO) rata-rata per jam (mg/m³)
PT08.S1(CO)	Numerik	Respons sensor PT08.S1 (sensor TGS2602, target CO)
NMHC(GT)	Numerik	Konsentrasi Hidrokarbon Non-Metana (NMHC) rata-rata per jam (µg/m³)
C6H6(GT)	Numerik	Konsentrasi Benzena (C6H6) rata-rata per jam (µg/m³)
PT08.S2(NMHC)	Numerik	Respons sensor PT08.S2 (sensor TGS2602, target NMHC)
NOx(GT)	Numerik	Konsentrasi Oksida Nitrogen (NOx) rata-rata per jam (ppb)
PT08.S3(NOx)	Numerik	Respons sensor PT08.S3 (sensor TGS2600, target NOx)
NO2(GT)	Numerik	Konsentrasi Nitrogen Dioksida (NO2) rata-rata per jam (µg/m³)
PT08.S4(NO2)	Numerik	Respons sensor PT08.S4 (sensor TGS2602, target NO2)
PT08.S5(O3)	Numerik	Respons sensor PT08.S5 (sensor TGS2602, target Ozon, O3)
T	Numerik	Temperatur / Suhu (dalam °C)
RH	Numerik	Kelembapan Relatif (Relative Humidity) (dalam %)
AH	Numerik	Kelembapan Absolut (Absolute Humidity)

2.2 Pemodelan Data Mining

Tahap pemodelan *data mining* merupakan inti dari penelitian ini, di mana model prediktif dibangun dan dievaluasi. Proses ini bertujuan untuk menemukan algoritma terbaik yang mampu memprediksi konsentrasi Benzena (C6H6(GT)) berdasarkan fitur-fitur lainnya (seperti sensor, temperatur, dan kelembapan).

2.2.1. Pemilihan Algoritma

Penelitian ini menggunakan pendekatan *supervised learning* untuk tugas regresi, yaitu memprediksi nilai konsentrasi Benzena (C6H6(GT)) yang bersifat kontinu. Untuk menemukan model dengan kinerja terbaik dalam memprediksi kualitas udara [2], [4], [5], tiga algoritma regresi yang berbeda akan diuji dan dibandingkan. Algoritma-algoritma ini dipilih berdasarkan rekam jejak keberhasilan mereka dalam menangani masalah serupa [2].

a. **Regresi Linear (Linear Regression):** Untuk perbandingan, regresi linear digunakan sebagai model dasar. Untuk memodelkan hubungan linier sederhana antara variabel input dan output, algoritma ini digunakan sebagai dasar statistik. Ini adalah praktik umum dalam studi yang mengevaluasi kinerja berbagai model regresi [8].
b. **Random Forest (RF):** adalah algoritma pengajaran kelompok yang kuat. Algoritma ini bekerja dengan membangun banyak pohon keputusan selama pelatihan dan mengumpulkan prediksi rata-rata dari setiap pohon.. Keunggulan RF terletak pada kemampuannya menangani hubungan non-linier dan ketahanannya terhadap *overfitting*. Kinerjanya dalam tugas-tugas regresi telah dievaluasi secara luas bersama model lain [8], dan strategi *ensemble* (seperti RF) dikenal efektif untuk peramalan [2].
c. **XGBoost (Extreme Gradient Boosting):** XGBoost adalah versi yang sangat efektif dari algoritma peningkatan gradient dan juga metode ensemble yang sangat baik. Dipilihnya algoritma ini karena kinerjanya yang luar biasa

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **18/01/2026**; Published: **27/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 160-168
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

dalam banyak kompetisi *data science* dan kemampuan untuk menangani jumlah data yang sangat besar dengan kecepatan komputasi yang tinggi [2]. XGBoost telah terbukti sangat efektif dalam peramalan kualitas udara dan telah digunakan secara luas [2], [9].

2.2.2. Skenario Pengujian

Untuk memastikan model dapat digeneralisasi pada data baru, skenario pengujian standar diterapkan:

a. **Pembagian Data (*Train-Test Split*):** Dataset dibagi menjadi dua bagian secara proporsional:
   1. **Data Latih (*Training Data*):** Sebesar 80% dari total data, digunakan untuk melatih setiap algoritma.
   2. **Data Uji (*Testing Data*):** Sebesar 20% dari total data, digunakan untuk menguji performa akhir model pada data yang belum pernah dilihat.

b. **Standardisasi Fitur (*Feature Scaling*):** Sebelum dimasukkan ke dalam model, fitur-fitur numerik (seperti T, RH, PT08.S1(CO), dll.) memiliki rentang nilai yang sangat berbeda. Untuk mencegah fitur dengan nilai besar (misal: ribuan) mendominasi fitur dengan nilai kecil (misal: puluhan), proses **Standardisasi** (*StandardScaler*) diterapkan. Proses ini mengubah skala data sehingga memiliki rata-rata (*mean*) 0 dan standar deviasi 1.

2.2.3. Metode Evaluasi

Untuk mengevaluasi dan memvalidasi kinerja model regresi yang dikembangkan, serangkaian metrik evaluasi standar digunakan [10], [9],[11]. Pemilihan metrik yang tepat merupakan langkah krusial dalam menilai keandalan dan akurasi model *machine learning* [12], [13]. Dalam penelitian ini, tiga metrik utama *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), dan Koefisien Determinasi (R^2) dipilih untuk mengukur kinerja model secara komprehensif, sejalan dengan praktik umum dalam studi evaluasi model prediktif [8], [12].

a. ***Mean Absolute Error* (MAE):** menghitung rata-rata dari selisih absolut (nilai mutlak) antara nilai yang diprediksi (yi) dan nilai aktual (^yi) [10], [11]. MAE sering dianggap sebagai ukuran yang lebih alami dan tidak ambigu untuk kesalahan rata-rata model [14], [15]. Keunggulan utamanya adalah kemudahan interpretasi, karena MAE direpresentasikan dalam unit yang sama dengan variabel target . Selain itu, MAE tidak memberikan bobot yang berlebihan pada kesalahan besar atau *outlier*, sehingga memberikan representasi yang kuat dari besaran kesalahan rata-rata [16], [14].

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |yi - \hat{y}i| \quad (1)$$

b. ***Root Mean Squared Error* (RMSE):** adalah akar kuadrat dari rata-rata selisih kuadrat antara nilai prediksi dan aktual [10], [12]. RMSE merupakan salah satu metrik yang paling umum digunakan dalam evaluasi model regresi [17], [8], [12]. Berbeda dengan MAE, proses pengkuadratan selisih dalam RMSE memberikan bobot yang jauh lebih besar pada kesalahan prediksi yang besar (outlier) [16], [14]. Hal ini membuat RMSE menjadi metrik pilihan ketika kesalahan prediksi yang besar dianggap sangat tidak diinginkan dan harus dihindari, karena metrik ini akan memberikan "hukuman" yang lebih berat pada model yang menghasilkan *outlier* tersebut [16].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (yi - \hat{y}i)^2} \quad (2)$$

c. ***R-squared* (R^2) atau Koefisien Determinasi:** adalah metrik statistik yang mengukur proporsi varians dalam variabel dependen (target) yang dapat dijelaskan atau diprediksi oleh variabel independen (fitur) [17], [11]. R^2 memberikan gambaran tentang seberapa baik model yang dipilih cocok dengan data, dibandingkan dengan model rata-rata sederhana [17], [10]. Nilainya berkisar dari 0 hingga 1 (atau ∞ hingga 1), di mana nilai yang lebih tinggi (mendekati 1) menunjukkan bahwa model mampu menjelaskan sebagian besar variabilitas data [10], [11]. R^2 dianggap sebagai metrik yang sangat informatif karena tidak memiliki unit (skala-bebas) dan memberikan ukuran relatif tentang kecocokan model, yang oleh beberapa peneliti dianggap lebih informatif daripada metrik berbasis kesalahan seperti MAE atau RMSE [17].

$$R^2 = 1 - \frac{\Sigma_{i=1}^{n}(yi - \hat{y}i)^2}{\Sigma_{i=1}^{n}(yi - \overline{y})^2} \quad (3)$$

2.3. Rancang Bangun Aplikasi Streamlit

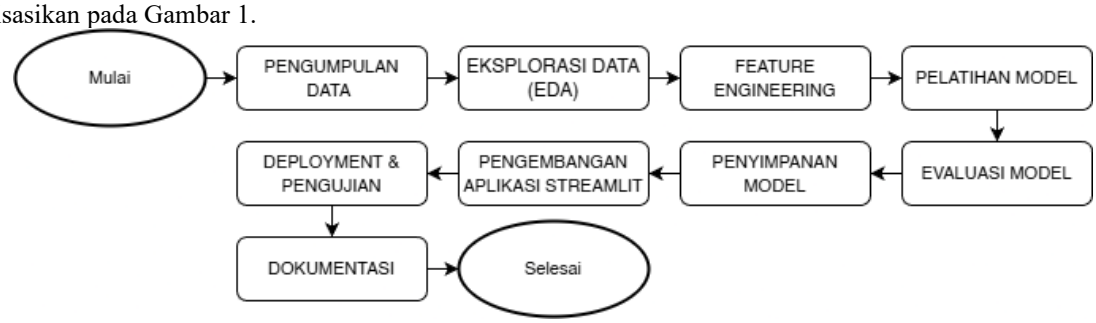
Setelah mendapatkan model *machine learning* terbaik, tahap selanjutnya adalah merancang dan membangun sebuah prototipe aplikasi web. Aplikasi ini bertujuan untuk mengimplementasikan model tersebut sehingga dapat diakses dan digunakan dengan mudah. Platform yang dipilih adalah Streamlit.

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **18/01/2026**; Published: **27/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 160-168
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

2.3.1. Arsitektur Sistem

Arsitektur sistem dirancang untuk mengelola alur data dari input pengguna hingga output prediksi, seperti yang divisualisasikan pada Gambar 1.



Gambar 1. Diagram Alir

Alur kerja sistem dapat dijelaskan dalam beberapa langkah utama:

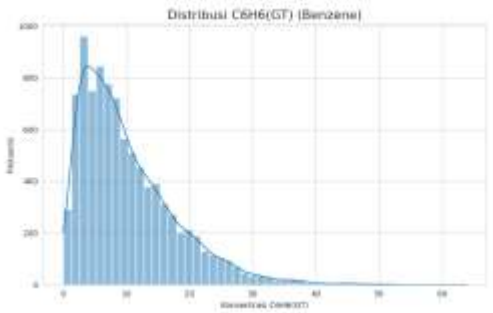
a. **Pemuatan Aset (Model & Scaler):** Saat aplikasi pertama kali dijalankan, sistem secara otomatis memuat (load) dua aset penting yang telah disimpan dari tahap pemodelan:
   1. **File Model (model.pkl):** Berisi model *machine learning* Random Forest yang sudah dilatih.
   2. **File Scaler (scaler.pkl):** Berisi objek StandardScaler yang digunakan untuk mentransformasi data latih.
b. **Input Pengguna:** Pengguna **berinteraksi** dengan antarmuka (UI) aplikasi dan memasukkan data parameter kualitas udara (seperti T, RH, PT08.S1(CO), dll.) melalui komponen *slider* atau *input box*.
c. Proses Prediksi (Backend):
   1. Ketika tombol "Prediksi" ditekan, aplikasi Streamlit menerima data input dari pengguna.
   2. Data input tersebut kemudian **ditransformasi** menggunakan scaler.pkl yang telah dimuat, agar skalanya sesuai dengan data yang digunakan saat melatih model.
   3. Data yang telah ditransformasi kemudian dimasukkan ke dalam model.pkl untuk **melakukan prediksi**.
d. **Tampilan Output:** Model mengembalikan hasil prediksi (estimasi nilai C6H6(GT)), dan aplikasi Streamlit akan **menampilkan hasil** tersebut secara jelas di antarmuka pengguna.

3. HASIL DAN PEMBAHASAN

3.1 Analisis Data Eksploratif (EDA)

Sebelum tahap pemodelan, data mentah **AirQuality.csv** melalui proses pra-pemrosesan. Nilai yang hilang (ditandai sebagai -200) diidentifikasi. Fitur NMHC(GT) dihapus karena memiliki data hilang yang signifikan, dan nilai hilang pada fitur lainnya diisi menggunakan interpolasi berbasis waktu (interpolate(method='time')).

Setelah data bersih, analisis eksploratif dilakukan pada variabel target, yaitu **C6H6(GT)** (konsentrasi Benzena).

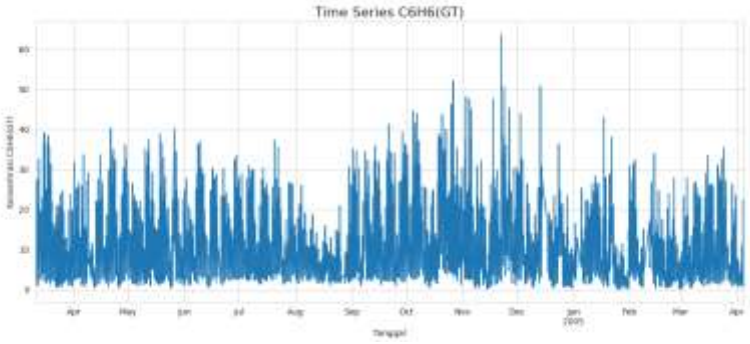


Gambar 2. Distribusi Konsentrasi C6H6(GT)

Gambar 2 di atas menunjukkan histogram dari variabel target C6H6(GT). Terlihat jelas bahwa distribusi data **condong ke kanan (*right-skewed*)**, yang mengindikasikan bahwa sebagian besar pencatatan menunjukkan konsentrasi Benzena yang rendah, dengan beberapa kasus pencatatan bernilai sangat tinggi.

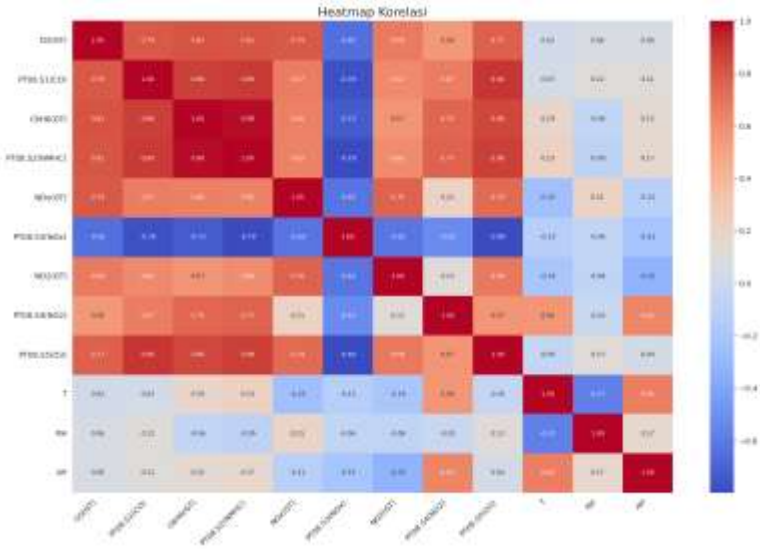
Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **18/01/2026**; Published: **27/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 160-168
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin



Gambar 3. Time Series Konsentrasi C6H6(GT)

Grafik *time series* di atas memvisualisasikan data C6H6(GT) terhadap waktu. Grafik ini menunjukkan adanya **fluktuasi dan pola musiman (*seasonality*)** yang jelas, di mana konsentrasi polutan cenderung naik dan turun secara periodik. Pola ini memperkuat keputusan (di Bab 2) untuk menggunakan fitur berbasis waktu (jam, hari, bulan) dalam pemodelan.



Gambar 4. Heatmap Korelasi Antar Fitur

Heatmap korelasi di atas menunjukkan hubungan linier antar variabel. Terlihat korelasi positif yang sangat kuat antara sensor-sensor (misalnya PT08.S1(CO) dan C6H6(GT)). Menariknya, fitur meteorologi seperti Temperatur (T) memiliki korelasi negatif yang kuat dengan hampir semua sensor polutan, menunjukkan bahwa suhu yang lebih tinggi cenderung berkaitan dengan pembacaan sensor yang lebih rendah.

3.2 Hasil Pemodelan Data Mining

Tiga model regresi diuji menggunakan data latih yang telah di-*scale*, dan dievaluasi kinerjanya pada data uji (20% data terakhir). Model yang diuji adalah *Linear Regression*, Random *Forest*, dan XGBoost.

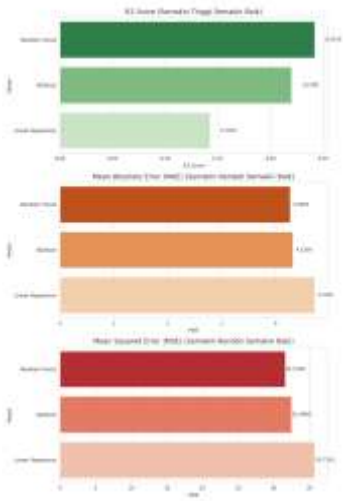
Tabel 2. Perbandingan Kinerja Model

Model	MAE (Mean Absolute Error)	MSE (Mean Squared Error)	R2 (R-squared)
Random Forest	**1.1444**	**3.5262**	**0.8900 (89.0%)**
XGBoost	1.1645	3.6190	0.8871 (88.7%)
Linear Regression	2.5085	11.6833	0.6358 (63.6%)

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **18/01/2026**; Published: **27/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 160-168
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Berdasarkan Tabel 2, **Random Forest Regressor** dipilih sebagai model terbaik karena menghasilkan nilai *error* (MAE dan MSE) terendah dan nilai R-squared (R^2) tertinggi, yaitu **0.89**. Ini berarti model *Random Forest* mampu menjelaskan 89% variabilitas data konsentrasi Benzena menggunakan fitur waktu dan cuaca yang diberikan.



Gambar 5. Visualisasi Perbandingan Metrik Model

Grafik di atas mengkonfirmasi keunggulan Random Forest (bar hijau tua), yang secara konsisten menunjukkan R2 Score tertinggi (semakin tinggi semakin baik) serta MAE dan MSE terendah (semakin rendah semakin baik).

Gambar 6. Perbandingan Nilai Aktual vs. Prediksi (Model Random Forest)

Untuk menganalisis kinerja model terbaik lebih lanjut, dibuat *scatter plot* yang membandingkan nilai aktual C6H6(GT) (sumbu Y) dengan nilai yang diprediksi oleh model (sumbu X) pada data uji. Hasilnya menunjukkan bahwa titik-titik data **berkumpul rapat di sekitar garis diagonal**, yang mengindikasikan bahwa nilai prediksi model sangat dekat dengan nilai kenyataannya.

3.3. Implementasi Aplikasi Streamlit

Sebagai tahap akhir, model Random Forest terbaik dan objek scaler yang telah disimpan, diimplementasikan ke dalam sebuah prototipe aplikasi web fungsional menggunakan *framework* Streamlit.

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **18/01/2026**; Published: **27/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 160-168
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Gambar 7. Tampilan Aplikasi Monitor dan Prediksi Tingkat Polusi Udara

Gambar 7 menunjukkan antarmuka input utama aplikasi. Pengguna berinteraksi melalui *sidebar* (panel samping) berjudul "Form Input Prediksi". Di sini, pengguna dapat memasukkan data parameter yang dibutuhkan oleh model, yang terbatas hanya pada data meteorologi dan waktu. Input ini mencakup pemilihan Tanggal dan Waktu, serta penyesuaian nilai Suhu (T), Kelembapan Relatif (RH), dan Kelembapan Absolut (AH) menggunakan komponen *slider* yang intuitif.

Gambar 8. Tampilan Hasil Aplikasi Monitor dan Prediksi Tingkat Polusi Udara

Setelah pengguna menekan tombol "Prediksi Kualitas Udara", aplikasi akan memproses data input (melakukan *scaling* dan prediksi) dan menampilkan hasilnya di halaman utama, seperti terlihat pada Gambar 8. Hasil prediksi disajikan dalam dua format: (1) nilai numerik spesifik untuk konsentrasi C6H6(GT) (Benzena), dan (2) visualisasi *gauge chart* (meteran) yang memberikan interpretasi langsung terhadap level kualitas udara (misalnya "BAIK"). Aplikasi juga menampilkan kembali tabel "Data Input yang Digunakan" untuk konfirmasi, membuktikan bahwa model dapat memberikan prediksi *real-time* berdasarkan input pengguna.

4. KESIMPULAN

Penelitian ini telah berhasil dilaksanakan untuk mencapai dua tujuan utama: pertama, menganalisis dan membangun model *machine learning* yang akurat untuk prediksi kualitas udara, dan kedua, mengimplementasikan model tersebut ke dalam sebuah prototipe aplikasi web yang fungsional. Berdasarkan dataset *Air Quality* UCI, penelitian ini berfokus pada prediksi konsentrasi Benzena (C6H6(GT)) dengan menggunakan pendekatan inovatif, yaitu hanya memanfaatkan fitur meteorologi (Temperatur, Kelembapan Relatif, Kelembapan Absolut) dan fitur turunan waktu (jam, hari dalam seminggu, dan bulan), tanpa bergantung pada pembacaan sensor polutan lainnya. Dalam tahap pemodelan, dilakukan perbandingan kinerja antara tiga algoritma regresi: Regresi Linear, Random Forest, dan XGBoost. Evaluasi dilakukan menggunakan metrik standar Mean Absolute Error (MAE), Mean Squared Error (MSE), dan R-Squared (R^2). Hasil komputasi menunjukkan bahwa model **Random Forest Regressor** memberikan kinerja

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **18/01/2026**; Published: **27/01/2026**

terbaik secara signifikan dibandingkan model lainnya. Model ini berhasil mencapai skor **R^2 sebesar 0.8900 (89.0%)**, dengan nilai MAE terendah (1.1444) dan MSE (3.5262). Hasil ini membuktikan bahwa model Random Forest sangat efektif dan mampu menjelaskan 89% variabilitas konsentrasi Benzena hanya berdasarkan kondisi cuaca dan waktu. Selanjutnya, model Random Forest terbaik beserta objek *scaler* yang telah dilatih, berhasil diimplementasikan ke dalam sebuah aplikasi web interaktif menggunakan framework **Streamlit**. Aplikasi ini dirancang dengan antarmuka pengguna yang intuitif, memungkinkan pengguna memasukkan data parameter cuaca dan waktu secara manual melalui *sidebar*. Aplikasi kemudian memproses input ini secara *real-time* untuk menghasilkan prediksi konsentrasi C6H6(GT). Hasil prediksi ditampilkan secara jelas kepada pengguna, tidak hanya dalam bentuk nilai numerik tetapi juga melalui visualisasi *gauge chart* yang memberikan indikator level kualitas udara (Baik, Sedang, atau Berbahaya). Dengan demikian, penelitian ini berhasil mentransformasi sebuah model prediktif teknis menjadi alat bantu praktis yang dapat diakses untuk pemantauan kualitas udara.

UCAPAN TERIMAKASIH

Terima kasih kepada semua pihak yang telah memberikan dukungan dalam penyusunan dan penyelesaian penelitian ini. Ucapan terima kasih disampaikan khususnya kepada Universitas Bina Sarana Informatika atas fasilitas dan dukungan yang diberikan selama proses penelitian berlangsung.

Penulis juga berterima kasih kepada rekan-rekan dosen, pembimbing, serta semua pihak yang telah memberikan masukan, motivasi, dan bantuan, baik secara langsung maupun tidak langsung, sehingga penelitian ini dapat terselesaikan dengan baik.

REFERENCES

[1] 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, p. e39783, 2024, doi: 10.1016/j.heliyon.2024.e39783.

[2] Y. Özüpak, F. Alpsalaz, and E. Aslan, "Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction," *Water. Air. Soil Pollut.*, vol. 236, no. 7, pp. 1–17, 2025, doi: 10.1007/s11270-025-08122-8.

[3] C. Praveen, R. V. V. S. V Prasad, A. Guna, and S. Krishna, "International Journal of Research Publication and Reviews Deep Learning-Based Real-Time Air Quality Prediction and Pollution Monitoring System," vol. 6, no. 4, pp. 3358–3364, 2025.

[4] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, "Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam," *Chemosphere*, vol. 338, no. May, p. 139518, 2023, doi: 10.1016/j.chemosphere.2023.139518.

[5] S. Chadalavada *et al.*, "Application of artificial intelligence in air pollution monitoring and forecasting : A systematic review," *Environ. Model. Softw.*, vol. 185, no. September 2024, p. 106312, 2025, doi: 10.1016/j.envsoft.2024.106312.

[6] L. Naizabayeva, G. Sembina, A. Aliman, and M. Satymbekov, "Air Pollution Forecasting in Almaty using Ensemble Machine Learning Models," vol. 6, no. 4, pp. 2461–2476, 2025.

[7] R. K. Sapkota, N. Adhikari, C. Subedi, R. Jha, and M. Subedi, "Web Application Prototype on Air Quality Index Prediction , Monitoring , and Information Dissemination System : Machine Learning and Python-Streamlit-based Application Tailored for Kathmandu Metropolitan City," 2025.

[8] J. Kaliappan, K. Srinivasan, and S. M. Qaisar, "Performance Evaluation of Regression Models for the Prediction of the COVID-19 Reproduction Rate," vol. 9, no. September, pp. 1–12, 2021, doi: 10.3389/fpubh.2021.729795.

[9] S. Tırınk, "Machine learning-based forecasting of air quality index under long-term environmental patterns : A comparative approach with XGBoost , LightGBM , and SVM," pp. 1–21, 2025, doi: 10.1371/journal.pone.0334252.

[10] O. Rainio, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, pp. 1–14, 2024, doi: 10.1038/s41598-024-56706-x.

[11] J. A. S. Cenita and P. R. F. Asuncion, "Performance Evaluation of Regression Models in Predicting the Cost of Medical Insurance," vol. 7, pp. 2052–2065, 2023, doi: 10.25147/ijcsr.2017.001.1.146.

[12] H. Khoshvaght, R. Ramyad, A. Razmjou, and M. Khiadani, "Journal of Environmental Chemical Engineering A critical review on selecting performance evaluation metrics for supervised machine learning models in wastewater quality prediction," *J. Environ. Chem. Eng.*, vol. 13, no. 6, p. 119675, 2025, doi: 10.1016/j.jece.2025.119675.

[13] M. R. Salmanpour, M. Alizadeh, G. Mousavi, S. Sadeghi, S. Amiri, and M. Oveisi, "Machine Learning Evaluation Metric Discrepancies across Programming Languages and Their Components : Need for

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **18/01/2026**; Published: **27/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 160-168
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Standardization," 2025, doi: 10.1109/ACCESS.2025.3549702.

[14] C. Res, C. J. Willmott, and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," vol. 30, pp. 79–82, 2005.

[15] S. M. R. Id and C. J. Willmott, "Decomposition of the mean absolute error (MAE) into systematic and unsystematic components," pp. 1–8, 2023, doi: 10.1371/journal.pone.0279774.

[16] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," no. 2, pp. 5481–5487, 2022.

[17] D. Chicco, M. J. Warrens, and G. Jurman, "The coef fi cient of determination R-squared is more informative than SMAPE , MAE , MAPE , MSE and RMSE in regression analysis evaluation," pp. 1–24, 2021, doi: 10.7717/peerj-cs.623.

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **18/01/2026**; Published: **27/01/2026**