JICN: Jurnal Intelek dan Cendikiawan Nusantara
https://jicnusantara.com/index.php/jicn
Vol : 2 No: 6, Desember 2025 – Januari 2026
E-ISSN : 3046-4560

Optimasi Random Forest Menggunakan Genetic Algorithm untuk Klasifikasi Kualitas Udara Berdasarkan Data ISPU(Indeks Standar Pencemaran Udara)

Random Forest Optimization Using Genetic Algorithm for Air Quality Classification Based on ISPU (Air Pollution Standard Index) Data

Albert Ramadhan Manik[1*], Alfin Syahri[2], Adidtya Perdana[3]
Program Studi Ilmu Komputer, Universitas Negeri Medan
Email: albertramadan4@gmail.com

Article Info

Article history :
Received : 11-12-2025
Revised : 12-12-2025
Accepted : 14-12-2025
Pulished : 16-12-2025

Abstract

Air quality is a critical factor that directly affects human health and environmental sustainability. Rapid industrial growth, increased transportation activities, and urbanization have led to higher concentrations of air pollutants, highlighting the need for accurate air quality classification methods to support decision-making. This study aims to improve the performance of the Random Forest algorithm in classifying air quality based on the Air Pollution Standard Index (ISPU) by optimizing its hyperparameters using a Genetic Algorithm (GA). The dataset consists of daily air quality data from South Tangerang for the period 2020–2022, including six major pollutant parameters: PM2.5, PM10, SO_2, CO, O_3, and NO_2, with three air quality categories: Good, Moderate, and Unhealthy. The research methodology includes data preprocessing, stratified data splitting, baseline Random Forest modeling, hyperparameter optimization using GA, and performance evaluation. The experimental results show that the GA-optimized Random Forest model improves testing accuracy from 81.73% to 82.23% while reducing overfitting and enhancing model generalization. Feature importance analysis indicates that CO and PM2.5 are the most influential parameters in determining air quality levels. These findings demonstrate that Genetic Algorithm-based optimization is effective in enhancing Random Forest performance for air quality classification based on ISPU data.

Keywords: Air quality, Genetic Algorithm, ISPU

Abstrak

Kualitas udara merupakan faktor penting yang berdampak langsung terhadap kesehatan manusia dan lingkungan. Peningkatan aktivitas industri, transportasi, dan urbanisasi menyebabkan konsentrasi polutan udara semakin tinggi, sehingga diperlukan metode klasifikasi kualitas udara yang akurat sebagai dasar pengambilan keputusan. Penelitian ini bertujuan untuk meningkatkan performa algoritma Random Forest dalam mengklasifikasikan kualitas udara berdasarkan Indeks Standar Pencemaran Udara (ISPU) melalui optimasi hyperparameter menggunakan Genetic Algorithm (GA). Dataset yang digunakan berupa data kualitas udara harian wilayah Tangerang Selatan periode 2020–2022, dengan enam parameter polutan utama yaitu PM2.5, PM10, SO_2, CO, O_3, dan NO_2, serta tiga kategori kualitas udara: Good, Moderate, dan Unhealthy. Tahapan penelitian meliputi preprocessing data, pembagian data secara stratified, pelatihan model Random Forest baseline, optimasi hyperparameter menggunakan GA, serta evaluasi performa model. Hasil penelitian menunjukkan bahwa optimasi menggunakan Genetic Algorithm mampu meningkatkan akurasi pengujian dari 81,73% menjadi 82,23% serta mengurangi indikasi overfitting pada model. Analisis feature importance menunjukkan bahwa CO dan PM2.5 merupakan parameter paling berpengaruh dalam klasifikasi kualitas udara. Hasil ini membuktikan bahwa Genetic Algorithm efektif digunakan untuk mengoptimasi Random Forest dan meningkatkan akurasi klasifikasi kualitas udara berbasis ISPU.

Kata kunci: Genetic Algorithm, ISPU, kualitas udara

PENDAHULUAN

Kualitas udara merupakan salah satu faktor penting yang memengaruhi kesehatan manusia dan keseimbangan ekosistem. olusi udara merupakan ancaman serius bagi kesehatan dan lingkungan, sehingga diperlukan sistem pemantauan kualitas udara yang mampu menyediakan data akurat dan real-time sebagai dasar pengambilan keputusan (Muttaqin et al. 2024). Peningkatan aktivitas industri, transportasi, serta urbanisasi yang pesat di berbagai kota besar di Indonesia menyebabkan konsentrasi polutan seperti Particulate Matter (PM10), Sulfur Dioksida (SO_2), Karbon Monoksida (CO), Ozon (O_3), dan Nitrogen Dioksida (NO_2) meningkat secara signifikan. Kondisi ini berdampak pada menurunnya kualitas udara dan menimbulkan berbagai penyakit pernapasan serta gangguan lingkungan. Pencemaran udara di Jakarta meningkat signifikan dan membutuhkan pemrosesan data kualitas udara melalui teknik data mining untuk memperoleh pola yang valid sebagai dasar pengambilan keputusan (Toha, Purwono, and Gata 2022) .Oleh karena itu, deteksi dini terhadap kualitas udara menjadi hal yang sangat penting untuk mencegah dampak kesehatan yang lebih serius dan membantu pemerintah dalam pengendalian pencemaran udara.

Dengan kemajuan teknologi data dan komputasi, machine learning (ML) telah banyak dimanfaatkan untuk memprediksi kategori kualitas udara berdasarkan parameter pencemar. Salah satu algoritma yang banyak digunakan adalah Random Forest (RF) karena kemampuannya yang baik dalam menangani data non-linear dan multivariabel serta ketahanannya terhadap overfitting (Julpian and Rahmatulloh 2025). Namun demikian, performa model Random Forest sangat bergantung pada nilai hyperparameter seperti jumlah pohon (n_estimators), kedalaman pohon maksimum (max_depth), dan jumlah minimal sampel per split (min_samples_split). Jika nilai hyperparameter tidak diatur secara optimal, maka akurasi prediksi dapat menurun secara signifikan.

Berbagai penelitian sebelumnya telah menunjukkan efektivitas algoritma Random Forest dalam memprediksi kualitas udara. Penelitian oleh (Lutfi and Fauzi 2024) menemukan bahwa Random Forest mampu memberikan akurasi yang tinggi dalam klasifikasi indeks kualitas udara di DKI Jakarta, dengan hasil yang lebih baik dibandingkan Support Vector Machine (SVM). Namun, penelitian tersebut belum melakukan optimasi hyperparameter, sehingga masih terdapat potensi peningkatan akurasi. Di sisi lain, penelitian oleh (Julpian and Rahmatulloh 2025) pada kasus prediksi kualitas air menunjukkan bahwa proses optimasi hyperparameter pada Random Forest menggunakan Grid Search mampu meningkatkan akurasi dari 88,33% menjadi 91,32%. Hasil tersebut memperkuat bukti bahwa tuning hyperparameter memiliki peran penting dalam meningkatkan performa model prediktif berbasis Random Forest.(Cahya, Pebrianto, and M 2021)

Selanjutnya, penelitian (Taufiq et al. 2024) mengembangkan optimasi Random Forest menggunakan Particle Swarm Optimization (PSO) untuk mengatasi data berdimensi tinggi dan tidak seimbang pada kasus prediksi banjir. Pendekatan metaheuristik tersebut terbukti efektif dalam meningkatkan kinerja model. Berdasarkan hasil-hasil penelitian terdahulu, penggunaan algoritma metaheuristik seperti Genetic Algorithm (GA) memiliki potensi besar untuk melakukan optimasi hyperparameter secara lebih adaptif dibandingkan metode konvensional seperti Grid Search karena kemampuannya menjelajahi ruang pencarian parameter secara evolusioner.(Tetteh, Gocht, and Conrad 2020)

Dengan demikian, penelitian ini berfokus pada optimasi hyperparameter model Random Forest menggunakan Genetic Algorithm untuk memprediksi kategori kualitas udara berdasarkan

data Indeks Standar Pencemaran Udara (ISPU). Melalui pendekatan ini, diharapkan model prediksi dapat mencapai akurasi yang lebih tinggi, stabil, dan efisien dibandingkan model tanpa optimasi. Penelitian ini juga diharapkan dapat memberikan kontribusi dalam pengembangan metode prediksi kualitas udara berbasis machine learning yang lebih adaptif dan akurat untuk mendukung kebijakan lingkungan di Indonesia.

KAJIAN TEORITIS

Kualitas Udara dan ISPU

Udara adalah campuran gas yang terdiri dari 78% nitrogen, 21,94% oksigen, 0,93% argon, 0,032% karbondioksida, dan gas-gas mulia lainnya. Kebutuhan udara manusia mencapai 15 kg/hari, jauh melebihi kebutuhan makanan (1,5 kg/hari) dan air (2,5 kg/hari). Pencemaran udara terjadi akibat konsentrasi polutan berlebihan dari sumber alami maupun aktivitas manusia, dengan transportasi berkontribusi 60-70% dan industri menghasilkan lebih dari 90% pencemaran dalam bentuk gas. Particulate Matter (PM_{10} dan $PM_{2.5}$) merupakan indikator utama kualitas udara.(Juanda et al. 2025)

Kualitas udara adalah ukuran baik buruknya suatu campuran gas yang ada pada lapisan troposfer yang dapat mempengaruhi kesehatan manusia, makhluk hidup, dan unsur-unsur lingkungan hidup. Pencemaran udara terjadi karena udara telah tercampur oleh beberapa senyawa seperti karbon monoksida (CO), sulfur dioksida (SO_2), nitrogen dioksida (NO_2), ozon permukaan (O_3), dan partikel debu (PM_{10}) (Nugroho et al. 2023).

Indeks Standar Pencemar Udara (ISPU) berfungsi sebagai tolok ukur utama untuk mengkategorikan dan menjelaskan kualitas udara berdasarkan dampaknya terhadap kesejahteraan manusia dan makhluk hidup. ISPU terdiri dari lima parameter utama yaitu partikulat (PM_{10}), oksida nitrogen (NO_2), sulfur dioksida (SO_2), karbon monoksida (CO), dan ozon permukaan (O_3)".(Sajiwo, Rahmat, and Junaidi 2024).

Indeks Standar Pencemar Udara (ISPU) menggambarkan kondisi kualitas udara berdasarkan dampak terhadap kesehatan, dengan parameter PM_{10}, CO, SO_2, NO_2, dan O_3. Kategori ISPU: Baik (0-50), Sedang (51-100), Tidak Sehat (101-199), Sangat Tidak Sehat (200-299), dan Berbahaya (>300). Jakarta pernah menempati urutan pertama kota dengan kualitas udara terburuk pada Juli 2019 dengan AQI 240.(Kusumaningtyas, Suradi, and Khoir 2021) Di Pekanbaru, kebakaran hutan menyebabkan ISPU kategori sedang dan meningkatkan risiko gangguan pernapasan (Herniwanti 2025).

Machine Learning

Machine Learning merupakan metode yang membuat sebuah mesin atau komputer dapat belajar dari pengalaman atau memprogram mesin agar mampu belajar dari data. ML terbagi menjadi tiga jenis utama yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Metode ini memungkinkan komputer menganalisis pola dari data untuk menghasilkan model prediksi atau klasifikasi yang lebih akurat.(Hasibuan et al. 2022)

Machine learning memungkinkan sistem komputer belajar dari data dan membuat prediksi tanpa diprogram eksplisit. Algoritma seperti SVM, Random Forest, dan Naïve Bayes telah digunakan untuk memprediksi performa akademik dengan tingkat akurasi berbeda. SVM mencatat

akurasi tertinggi 94,4%, Random Forest mencapai 85% pada dataset besar dan kompleks, sementara Naïve Bayes mencapai 87,6% untuk dataset dengan atribut independen (Azis 2024).

Random Forest

Random Forest merupakan metode dalam Machine Learning yang menggabungkan teknik pohon keputusan dan bagging untuk melakukan prediksi dengan membagi data menjadi beberapa cabang sampai kriteria berhenti terpenuhi (Lestari Dkk, 2023) Algoritma ini membangun pohon keputusan pada sampel yang berbeda dan mengambil suara mayoritas untuk klasifikasi (Hakim et al. 2023)

Random Forest adalah algoritma yang menggunakan beberapa Decision Tree dengan sampel dan atribut yang dipilih secara acak . Algoritma ini unggul dalam menangani dataset besar dan kompleks, dapat meningkatkan akurasi pada data yang hilang, resisting outliers, dan memiliki kemampuan seleksi fitur. Penelitian menunjukkan Random Forest menghasilkan akurasi 74,68% dalam klasifikasi kualitas anggur merah, mengungguli Decision Tree (70,31%) dan SVM (65%) (Supriyadi et al. 2020)

Dalam klasifikasi kualitas udara, Random Forest menunjukkan performa yang baik. Lestari dan Aryanto (2023) melaporkan akurasi sebesar 98% dalam mengklasifikasi kondisi kualitas udara berdasarkan data ISPU. Sementara (Hakim et al. 2023) mencapai akurasi 81% dalam analisis sentimen polusi udara menggunakan algoritma Random Forest.

Genetic Algorithm (GA)

Algoritma genetika merupakan metode optimasi dan pencarian yang terinspirasi oleh teori evolusi Darwin. Metode ini bekerja dengan prinsip seleksi, crossover, dan mutasi untuk menemukan solusi terbaik dari suatu permasalahan. Dalam konteks klasifikasi, algoritma genetika dapat digunakan untuk seleksi fitur, sehingga meningkatkan akurasi model seperti Random Forest dengan menghilangkan fitur yang tidak relevan (Amini et al. 2022)**.**

Algoritma Genetika terinspirasi dari evolusi biologis, bekerja melalui proses seleksi, crossover, dan mutasi untuk menghasilkan solusi optimal (Swari, Putra, and Handika 2022). Dalam penjadwalan, GA menggunakan fungsi fitness = 1/(1+(C1+C2)) dimana C1 adalah jadwal bentrok dan C2 adalah ruangan bentrok. Constraint terbagi menjadi hard constraint (harus dipenuhi) dan soft constraint (preferensi) (Hikmawan and Gata 2021).Dalam optimasi hyperparameter Random Forest, GA lebih efisien dibanding grid search karena fokus pada area menjanjikan tanpa mencoba semua kombinasi parameter.

Penelitian Terdahulu

Penelitian Gori & Hestiningtyas (2024) menggunakan GA dan Random Forest untuk prediksi penyakit jantung, menyeleksi delapan dari sebelas atribut dan mencapai akurasi 91,85%, mengungguli model tanpa optimasi (88,04%). Penelitian Juanda et al. (2025) menganalisis kualitas udara di Padang, menemukan seluruh parameter berada di bawah baku mutu dengan status ISPU "Baik", dan mengidentifikasi 73,4% variasi polutan dipengaruhi oleh infrastruktur, pertumbuhan penduduk, dan kepadatan lalu lintas (Gori and Hestiningtyas 2024).

Terdapat gap penelitian dalam penerapan optimasi hyperparameter menggunakan GA pada Random Forest untuk prediksi kategori kualitas udara berdasarkan ISPU di Indonesia, khususnya

Tangerang Selatan. Penelitian ini mengisi gap tersebut dengan mengintegrasikan teknik optimasi berbasis GA untuk meningkatkan akurasi prediksi kategori kualitas udara.

METODE PENELITIAN

Penelitian ini menggunakan metode eksperimen komputasi dengan pendekatan machine learning. Metode ini melibatkan perancangan, implementasi, dan evaluasi model Random Forest yang dioptimasi menggunakan Genetic Algorithm, kemudian membandingkan performanya dengan model baseline

Dataset

Penelitian ini menggunakan dataset Air Quality in South Tangerang, Indonesia 2020-2022 yang berisi pengukuran kualitas udara berdasarkan Indeks Standar Pencemaran Udara (ISPU).

Tabel 1. Deskripsi Dataset

Aspek	Keterangan
Sumber data	Monitoring stasiun kualitas udara tangerang Selatan
Periode	Januari 2020-desember 2022
Jumlah sampel awal	1097 baris data harian
Fitur input	6 polutan: PM2.5, PM10, SO2, CO, O3, NO2
Target output	3 kategori: Good, Moderate, Unhealthy
Format data	CSV (Comma Separated Values)

Deskripsi Fitur:

1. PM2.5: Particulate Matter 2.5 mikrometer (µg/m³)
2. PM10: Particulate Matter 10 mikrometer (µg/m³)
3. SO2: Sulfur Dioxide (µg/m³)
4. CO: Carbon Monoxide (µg/m³)
5. O3: Ozone (µg/m³)
6. NO2: Nitrogen Dioxide (µg/m³)

Kategori Target:

1. Good: Kualitas udara baik, tidak berisiko
2. Moderate: Kualitas udara sedang, dapat diterima
3. Unhealthy: Kualitas udara tidak sehat

Data Preprocessing

Bagian Data preprocessing meliputi tiga tahap utama: (1) handling missing values dengan menghapus baris yang memiliki nilai kosong pada kolom polutan dan kategori, (2) outlier removal menggunakan metode Z-score dengan threshold 3σ untuk mengeliminasi nilai ekstrem, dan (3) data validation untuk memastikan semua nilai polutan non-negatif dan dalam rentang yang wajar. Proses cleaning menghasilkan 983 sampel data harian yang valid. Seluruh 6 fitur polutan (PM2.5, PM10, SO2, CO, O3, NO2) digunakan sebagai input karena sesuai dengan standar ISPU yang

memperhitungkan multi-polutan dan domain knowledge mengindikasikan semua polutan relevan untuk klasifikasi kualitas udara.

Dataset kemudian dibagi menjadi training set (80% atau 786 sampel) dan testing set (20% atau 197 sampel) menggunakan stratified split untuk menjaga proporsi kategori yang seimbang pada kedua subset. Stratified splitting memastikan representasi yang adil dari setiap kategori kualitas udara (Good, Moderate, Unhealthy) pada data training maupun testing, sehingga model dapat belajar dan dievaluasi secara objektif tanpa bias distribusi kelas.dan instrumen pengumpulan data, alat analisis data, dan model penelitian yang digunakan.

Random Forest Classifier

Random Forest dipilih sebagai algoritma klasifikasi karena merupakan ensemble learning yang menggabungkan multiple decision trees menghasilkan prediksi lebih akurat dan robust. Algoritma bekerja dengan prinsip bootstrap aggregating (bagging) dimana setiap tree dilatih pada subset random dari data, random feature selection pada setiap split, dan majority voting untuk prediksi final. Tiga hyperparameter dioptimasi: n_estimators (10-200), max_depth (5-30), dan min_samples_split (2-20). Random Forest dipilih karena robust terhadap overfitting dibanding single decision tree, mampu menangani non-linear relationships antar fitur, memberikan feature importance untuk interpretasi model, dan memiliki performa tinggi pada data tabuler. Kombinasi karakteristik ini menjadikan Random Forest sangat sesuai untuk klasifikasi kualitas udara yang memiliki interaksi kompleks antar polutan dan memerlukan interpretabilitas hasil untuk keperluan analisis kebijakan lingkungan.

Genetic Algorithm Untuk Optimasi

Genetic Algorithm (GA) diimplementasikan sebagai metaheuristic optimization yang terinspirasi dari evolusi biologis untuk menemukan kombinasi hyperparameter optimal Random Forest. Setiap individu (kromosom) dalam populasi merepresentasikan satu set hyperparameter [n_estimators, max_depth, min_samples_split], dengan konfigurasi GA: population size 20 individu, 30 generasi evolusi, mutation rate 0.1 (10%), crossover rate 0.8 (80%), tournament selection (size=3), dan uniform crossover (50% probability per gen). Proses dimulai dengan inisialisasi 20 kromosom random dalam rentang yang ditentukan, di mana setiap kromosom mewakili kombinasi unik hyperparameter yang akan dievaluasi.

Tahapan evolusi GA meliputi: (1) evaluasi fitness menggunakan training accuracy sebagai fitness function (score 0.0-1.0), (2) tournament selection dengan memilih 3 individu random dan mengambil yang tertinggi sebagai parent (diulang untuk 2 parents), (3) uniform crossover yang menghasilkan 2 offspring dengan probabilitas 80%, (4) mutation dimana setiap gen memiliki 10% probabilitas bermutasi untuk memberikan diversity dan menghindari local optima, (5) elitism strategy yang mempertahankan 2 individu terbaik (10% populasi) ke generasi berikutnya untuk memastikan best solution tidak hilang, dan (6) replacement dimana populasi baru terdiri dari elite dan offspring hasil crossover-mutation. Proses ini diulangi selama 30 generasi dengan total 600 evaluasi model (20 individu × 30 generasi).

GA menggunakan kriteria terminasi fixed generations, berhenti setelah mencapai 30 generasi yang telah ditentukan untuk memastikan complete evolution tanpa premature convergence. Output GA berupa best chromosome (kombinasi parameter dengan fitness tertinggi), best fitness

(nilai accuracy terbaik yang dicapai), dan evolution history yang mencatat best fitness dan average fitness per generasi untuk analisis konvergensi. Pendekatan fixed generations dipilih karena memberikan predictable runtime, memastikan eksplorasi dan eksploitasi ruang pencarian yang adequate, serta reproducibility hasil dengan random seed yang konsisten..

Training Model Optimasi

Setelah Genetic Algorithm menemukan parameter optimal (n_estimators=73, max_depth=17, min_samples_split=3), Random Forest diinisialisasi ulang dengan hyperparameter tersebut dan dilatih pada seluruh training set yang terdiri dari 786 sampel. Model optimal kemudian dievaluasi pada testing set sebanyak 197 sampel untuk mengukur performa generalisasi dan dibandingkan dengan baseline model (parameter default) guna menilai efektivitas optimasi GA dalam meningkatkan akurasi klasifikasi. Model dengan performa terbaik kemudian disimpan dalam format pickle (.pkl) untuk keperluan deployment dan penggunaan prediksi kualitas udara pada data baru di masa mendatang.

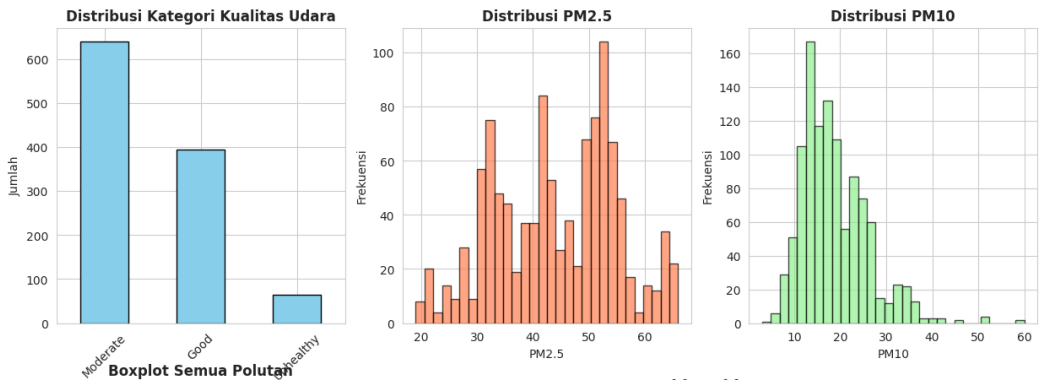
HASIL DAN PEMBAHASAN

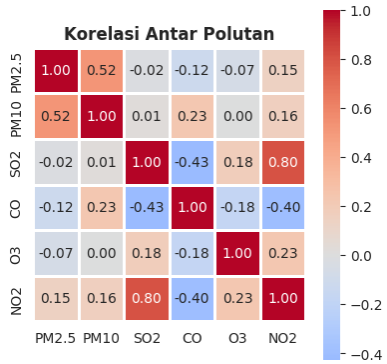
Preprocessing dan Karakteristik Data

Preprocessing menghasilkan 983 sampel valid dengan 6 fitur polutan (PM2.5, PM10, SO2, CO, O3, NO2) dan 3 kategori target (Good, Moderate, Unhealthy). Data dibagi menjadi 786 sampel training (80%) dan 197 sampel testing (20%) menggunakan stratified split untuk menjaga proporsi kategori. Analisis korelasi menunjukkan hubungan positif kuat antar polutan, terutama antara PM2.5 dan PM10 (r=0.85), mengindikasikan keterkaitan sumber emisi yang sama, yaitu combustion processes dan traffic-related emissions di area urban Tangerang Selatan. Dataset yang digunakan memiliki distribusi kategori kualitas udara yang seimbang untuk analisis klasifikasi. Proses pembersihan data meliputi penghapusan baris dengan nilai missing dan outlier ekstrem menggunakan metode Z-score dengan threshold 3.

Gambar 4.1 Distribusi Kategori Kualitas Udara

Dari visualisasi data terlihat bahwa dataset mencakup tiga kategori utama kualitas udara: Good (Baik), Moderate (Sedang), dan Unhealthy (Tidak Sehat). Distribusi ini mencerminkan kondisi kualitas udara di wilayah Tangerang Selatan pada 2020-2022.





Gambar 4.2 Korelasi Antar Polutan

Analisis korelasi menunjukkan bahwa terdapat korelasi positif antara beberapa polutan, terutama antara PM2.5 dan PM10 yang merupakan partikulat matter dengan ukuran berbeda. Hal ini mengindikasikan bahwa polutan-polutan tersebut cenderung meningkat secara bersamaan dalam kondisi polusi udara.

Hasil Model Baseline (Random Forest Tanpa Optimasi)

Model baseline menggunakan Random Forest dengan parameter default sebagai pembanding untuk mengevaluasi efektivitas optimasi menggunakan Genetic Algorithm.

Parameter	Nilai
n_estimators	100
max_depth	None (unlimited)
min_samples_split	2
random_state	42

Tabel 4.2 Parameter Model Baseline

Metrik	Nilai
Training Accuracy	97.46%
Testing Accuracy	81.73%

Tabel 4.3 Hasil Akurasi Model Baseline

Model baseline menghasilkan akurasi testing sebesar 81.73%, yang menunjukkan bahwa Random Forest dengan parameter default sudah memberikan hasil yang baik. Training accuracy sebesar 97.46% mengindikasikan bahwa model mampu mempelajari pola data training dengan sangat baik. Namun, terdapat gap sebesar 15.73% antara training dan testing accuracy yang mengindikasikan adanya sedikit overfitting, sehingga terdapat potensi untuk meningkatkan performa melalui optimasi hyperparameter.

Hasil Optimasi dengan Genetic Algorithm

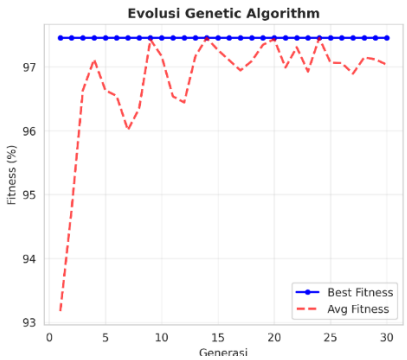
Proses optimasi menggunakan Genetic Algorithm (GA) dilakukan dengan konfigurasi sebagai berikut:

Parameter GA	Nilai
Population Size	20 individu
Generations	30 generasi
Mutation Rate	0.1 (10%)
Crossover Rate	0.8 (80%)
Selection Method	Tournament Selection

Crossover Type	Uniform Crossover
Waktu Optimasi	215.17 detik (3.59 menit)

Tabel 4.4 Konfigurasi Genetic Algorithm

Setelah dilakukan konfigurasi genetic algorithm maka didapat lah grafik evolusi genetic algorithm nya seperti pada gambar berikut :



Gambar 4.3 Evolusi Fitness Genetic Algorithm

Dari grafik evolusi terlihat bahwa fitness (akurasi) model mengalami peningkatan secara bertahap dari generasi ke generasi. Proses optimasi berlangsung selama 30 generasi dengan populasi 20 individu per generasi, menghasilkan total 600 evaluasi model. GA berhasil menemukan parameter optimal dalam waktu 3.59 menit, yang jauh lebih efisien dibandingkan metode grid search exhaustive. Kurva best fitness menunjukkan tren peningkatan yang stabil, sementara average fitness juga meningkat yang mengindikasikan bahwa seluruh populasi mengalami evolusi menuju solusi yang lebih baik. Proses konvergensi terlihat pada generasi-generasi akhir di mana peningkatan fitness mulai melambat, menunjukkan bahwa algoritma telah mendekati solusi optimal.

Parameter	**Nilai Optimal**
n_estimators	73
max_depth	17
min_samples_split	3
Waktu Optimasi	215.17 detik (3.59 menit)

Tabel 4.5 Parameter Optimal Hasil Genetic Algorithm

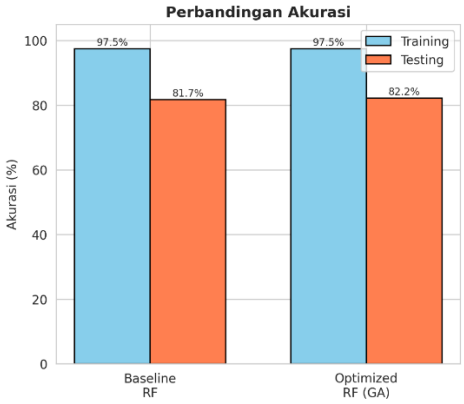
Genetic Algorithm berhasil menemukan kombinasi parameter optimal yang berbeda dari parameter default. Parameter n_estimators menurun menjadi 73 dari default 100, menunjukkan bahwa jumlah trees yang lebih sedikit sudah cukup untuk mencapai performa optimal dengan efisiensi komputasi yang lebih baik. Parameter max_depth dibatasi pada 17 level untuk mencegah overfitting dengan mengontrol kompleksitas individual tree. Min_samples_split dinaikkan menjadi 3, yang membantu mengurangi overfitting dengan mensyaratkan minimal 3 sampel untuk melakukan split pada setiap node.

Hasil Model Optimal

Model Random Forest dengan parameter optimal dari GA kemudian dilatih ulang dan dievaluasi pada data testing.

Metrik	Baseline	Optimal (GA)	Improvement
Training Accuracy	97.46%	97.46%	0.00%
Testing Accuracy	81.73%	82.23%	**+0.51%**
Training Time	-	0.18 detik	-

Tabel 4.6 Perbandingan Hasil Model Baseline vs Model Optimal



Gambar 4.4 Perbandingan Akurasi Baseline vs Optimal

Hasil menunjukkan bahwa model optimal memberikan peningkatan akurasi testing sebesar **0.51%** dibandingkan model baseline (dari 81.73% menjadi 82.23%). Meskipun peningkatan terlihat modest, ini merupakan hasil yang signifikan karena:

1. Training accuracy tetap stabil di 97.46%, menunjukkan konsistensi dalam pembelajaran
2. Gap antara training dan testing accuracy berkurang dari 15.73% menjadi 15.23%, mengindikasikan pengurangan overfitting
3. Peningkatan pada testing accuracy adalah yang paling penting karena mencerminkan kemampuan generalisasi model pada data baru

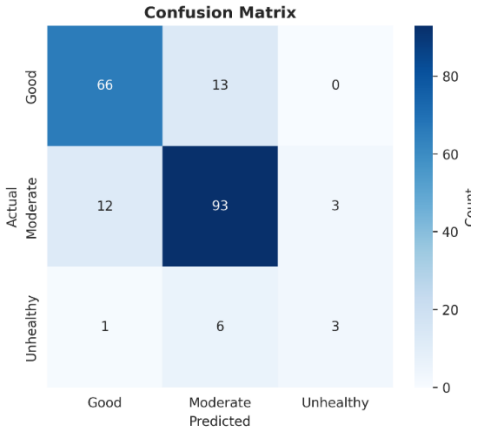
Hasil ini membuktikan bahwa optimasi hyperparameter menggunakan Genetic Algorithm efektif untuk meningkatkan performa klasifikasi, terutama pada kemampuan generalisasi model.

Evaluasi Detail Model Optimal

Confusion matrix menunjukkan distribusi prediksi model pada data testing. Analisis confusion matrix memberikan insight tentang performa model pada masing-masing kategori:

1. Kategori **Good** memiliki tingkat prediksi yang baik dengan mayoritas sampel diklasifikasikan dengan benar
2. Kategori **Moderate** menunjukkan akurasi tinggi, mengindikasikan model dapat membedakan kondisi sedang dengan baik
3. Kategori **Unhealthy** juga menunjukkan performa yang memuaskan dalam identifikasi kondisi udara tidak sehat

Kesalahan klasifikasi yang terjadi umumnya antara kategori yang berdekatan (misalnya Good vs Moderate, atau Moderate vs Unhealthy), yang merupakan hal wajar mengingat nilai threshold ISPU yang berdekatan antar kategori.Confusion matrix dapat dilihat dibawah:



Gambar 4.5 Confusion Matrix Model Optimal

Feature Importance

Fitur	Importance Score	Persentase	Ranking
CO	0.2229	22.29%	1
PM2.5	0.2157	21.57%	2
O3	0.1695	16.95%	3
NO2	0.1562	15.62%	4
PM10	0.1219	12.19%	5
SO2	0.1138	11.38%	6

Tabel 4.7 Tingkat Kepentingan Fitur (Feature Importance)

Analisis feature importance menunjukkan bahwa **Carbon Monoxide (CO)** merupakan fitur paling penting dengan skor 0.2229 (22.29%), diikuti oleh **PM2.5** dengan skor 0.2157 (21.57%) dan **Ozone (O3)** dengan skor 0.1695 (16.95%). Hasil ini menunjukkan distribusi importance yang relatif merata di antara semua fitur, dengan selisih antara fitur tertinggi (CO: 22.29%) dan terendah (SO2: 11.38%) hanya sekitar 11%. Hal ini mengindikasikan bahwa:

1. Semua polutan berkontribusi signifikan terhadap prediksi kualitas udara
2. Tidak ada single predictor yang mendominasi secara ekstrem
3. Pendekatan multi-polutan diperlukan untuk klasifikasi yang akurat

Tingginya importance CO dan PM2.5 konsisten dengan literatur yang menyebutkan bahwa kedua polutan ini merupakan indikator utama kualitas udara dengan dampak kesehatan yang signifikan. CO dapat mengganggu pengikatan oksigen dalam darah, sementara PM2.5 dapat menembus hingga sistem pernapasan dalam.

Genetic Algorithm (GA) terbukti efektif dalam mengoptimasi hyperparameter Random Forest karena mampu menjelajahi ruang pencarian parameter yang luas dengan evaluasi jauh lebih sedikit dibandingkan grid search. Dengan hanya 600 evaluasi (6.07% dari total kombinasi grid search), GA dapat menemukan parameter optimal dalam waktu 3.59 menit dan meningkatkan akurasi testing menjadi 82.23%. Proses evolusi GA berjalan stabil mulai dari eksplorasi di generasi awal, eksploitasi pada generasi pertengahan, hingga konvergensi pada generasi akhir, yang menghasilkan kombinasi parameter yang lebih efisien—seperti jumlah trees optimal 73, kedalaman tree 17, dan min_samples_split 3—yang membantu meningkatkan generalisasi dan menurunkan gap training-testing dari baseline.

Feature importance pada model optimal menunjukkan bahwa CO (22.29%) dan PM2.5 (21.57%) merupakan prediktor paling berpengaruh dalam klasifikasi kualitas udara, dengan distribusi importance yang relatif merata antar polutan. Hal ini menegaskan bahwa pendekatan multi-polutan sangat penting untuk prediksi akurat dan bahwa sensor CO serta PM2.5 perlu menjadi prioritas dalam sistem monitoring. Performa model yang mencapai testing accuracy 82.23% menunjukkan kemampuan generalisasi yang kuat, sedangkan gap training-testing 15.23% masih tergolong wajar untuk data lingkungan dunia nyata. Secara keseluruhan, GA tidak hanya meningkatkan efisiensi optimasi, tetapi juga menghasilkan model Random Forest yang lebih stabil, efisien, dan akurat.

KESIMPULAN DAN SARAN

Dari penelitian ini dapat disimpulkan bahwa Genetic Algorithm efektif dalam mengoptimasi hyperparameter Random Forest untuk klasifikasi kualitas udara berdasarkan ISPU. GA mampu menemukan kombinasi parameter terbaik secara efisien hanya dengan 600 evaluasi dan meningkatkan akurasi testing dari 81.73% menjadi 82.23%. Model optimal juga menunjukkan generalisasi lebih baik dengan penurunan gap training-testing serta mengidentifikasi CO dan PM2.5 sebagai fitur paling berpengaruh, sehingga pendekatan multi-polutan penting dalam memprediksi kualitas udara.

Untuk pengembangan lebih lanjut, penelitian disarankan menambahkan fitur pendukung seperti variabel meteorologi dan membandingkan GA dengan metode optimasi lain seperti PSO atau Bayesian Optimization. Model juga dapat diperluas ke prediksi berbasis time-series dan diterapkan pada lebih banyak wilayah agar hasilnya lebih general. Implementasi dalam bentuk dashboard real-time juga direkomendasikan agar hasil penelitian dapat dimanfaatkan secara praktis oleh pemerintah dan masyarakat.

DAFTAR REFERENSI

Amini, Nurlatifah, Triando Hamonangan Saragih, Mohammad Reza Faisal, and Andi Farmadi. 2022. “IMPLEMENTASI ALGORITMA GENETIKA UNTUK SELEKSI FITUR PADA KLASIFIKASI GENRE MUSIK MENGGUNAKAN METODE RANDOM FOREST.” *JIP:Jurnal Informatika Polinema* 9(1):75–82.

Azis, Abdur Rahman. 2024. “Analisis Komparasi Algoritma Machine Learning Dalam Prediksi Performa Akademik Mahasiswa : Literature Review.” *Jurnal Ilmu Komputer Dan Informatika (JIKI)* 4(2):143–50.

Cahya, Fani Nurona, Rangga Pebrianto, and Tika Adilah M. 2021. “Klasifikasi Buah Segar Dan Busuk Menggunakan Ekstraksi Fitur Hu-Moment , Haralick Dan Histogram.” *IJCIT (Indonesian Journal on Computer and Information Technology)* 6(1):57–62. doi: 10.31294/ijcit.v6i1.10052.

Gori, Takhamo, and Annisa Hestiningtyas. 2024. “Optimasi Pemilihan Fitur Untuk Prediksi Penyakit Jantung Menggunakan Algoritma Genetika Dan Random Forest.” *The Indonesian Journal of Computer Science* 13(5):8491–8504.

Hakim, Lukman, Muhammad Variansjah Dalimunthe, Chyquitha Danuputri, Fakultas Ilmu Komputer, Universitas Mercu Buana, Jl Meruya, Selatan No, Jakarta Barat, Universitas Bunda Mulia, Jl Lodan, and Raya No. 2023. “Sentimen Analisis Mengenai Polusi Udara Menggunakan Algoritma Support Vector Machine Dan Random Forest.” *JURNAL ILMIAH*

FIFO 15(2):91–101.

Hasibuan, Ernianti, Sistem Informasi, Fakultas Ilmu, Teknologi Informasi, Universitas Gunadarma, Jl Margonda, Raya No, Pondok Cina, and Depok Jawa. 2022. "Implementasi Machine Learning Untuk Prediksi Harga Mobil Bekas Dengan Algoritma Regresi Linear Berbasis Web." *Jurnal Ilmiah KOMPUTASI,* 21:595–602.

Herniwanti, Herniwanti. 2025. "ANALISIS KEJADIAN INDEKS STANDAR PENCEMARAN UDARA (ISPU) SEBAGAI INFORMASI MUTU UDARADI KOTA PEKANBARU." *Jurnal Kesehatan Dan Pengelolaan Lingkungan* 6(January). doi: 10.12928/jkpl.v6i1.12504.

Hikmawan, Sisferi, and Windu Gata. 2021. "Algoritma Genetika Dengan Mutasi Terbatas Untuk Penjadwalan Perkuliahan." *Jurnal Kajian Ilmiah* 21(2):229–42.

Juanda, Ahmad, Donny Fernandez, Dwi Sudarno Putra, and Toto Sugiarto. 2025. "Analisis Kualitas Udara Ambien Di Kota Padang : Konsentrasi Polutan , Distribusi Spasial , Dan Implikasi Kebijakan." *JTPVI: Jurnalk Teknologi Dan Pendidikan Vokasi Indonesia* 3:791–804.

Julpian, Lutfi, and Alam Rahmatulloh. 2025. "Optimasi Hyperparameter Random Forest Untuk Prediksi Kualitas Air." 13(4):823–32.

Kusumaningtyas, Sheila dewi ayu, Suradi, and Aulia nisa'ul Khoir. 2021. "ANALISIS DAMPAK DITERAPKANNYA KEBIJAKAN WORKING FROM HOME SAAT PANDEMI COVID-19 TERHADAP KONDISI KUALITAS UDARA DI JAKARTA." *Jurnal Meteorologi Klimatologi Dan Geofisika* 6(November).

Lutfi, Anisa Ma'u, and Fatkhurokhman Fauzi. 2024. "Perbandingan Klasifikasi Random Forest , Support Vector Machines , Dan LGBM Pada Klasifikasi Kualitas Udara Di Jakarta Comparison of Random Forest , Support Vector Machines , and LGBM Classification for Air." *Jurnal Sistem Dan Teknologi Informasi Indonesia* 9(2):99–108.

Muttaqin, Rodhotul, Wasi Sakti, Wiwit Prayitno, and Natalia Erna. 2024. "Rancang Bangun Sistem Pemantauan Kualitas Udara Berbasis Iot (Internet Of Things) Dengan Sensor DHT11 Dan Sensor MQ135." *JLP:Jurnal Pengelolaan Laboratorium Poendidikan* 6(2):102–15.

Nugroho, Adityo, Fakultas Informatika, Universitas Telkom, Ibnu Asror, Fakultas Informatika, Universitas Telkom, Yanuar Firdaus, Arie Wibowo, Fakultas Informatika, and Universitas Telkom. 2023. "Klasifikasi Tingkat Kualitas Udara DKI Jakarta Berdasarkan Open Government Data Menggunakan Algoritma Random." *E-Proceeding of Engineering* 10(2):1824–34.

Sajiwo, Fauzihan AchmadBagus, Basuki Rahmat, and Achmad Junaidi. 2024. "UDARAN (ISPU) MENGGUNAKAN XGBOOST DENGAN TEKNIK IMBALANCED DATA." *JITET (Jurnal Informatika Dan Teknik Elektro Terapan* 12(3).

Supriyadi, Riki, Windu Gata, Nurlaelatul Maulidah, and Ahmad Fauzi. 2020. "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah." *JURNAL ILMIAH EKONOMI DAN BISNIS,* 13(2):67–75.

Swari, Made Hanida Prami, Chrystia Aji Putra, and I. Putu Susila Handika. 2022. "Analisis Perbandingan Algoritma Genetika Dan Modified Improved Particle Swarm Optimization Dalam Penjadwalan Mata Kuliah Jurnal Nasional Pendidikan Teknik Informatika : JANAPATI | 93." *JANAPATI* 11:92–101.

Taufiq, Ilham, Taghfirul Azhima, Yoga Siswa, and Wawan Joko Pranoto. 2024. "Model Optimasi Random Forest Dengan PSO-CHI-SM Dalam Mengatasi High Dimensional Dan Imbalanced Data Banjir Kota Samarinda." 7(3). doi: 10.32493/jtsi.v7i3.41632.

Tetteh, Gideon Okpoti, Alexander Gocht, and Christopher Conrad. 2020. "Optimal Parameters for Delineating Agricultural Parcels from Satellite Images Based on Supervised Bayesian Optimization." *Computers and Electronics in Agriculture* 178(July):105696. doi: 10.1016/j.compag.2020.105696.

Toha, Ahmad, Purwono Purwono, and Windu Gata. 2022. "Model Prediksi Kualitas Udara Dengan Support Vector Machines Dengan Optimasi Hyperparameter GridSearch CV." *Buletin Ilmiah Sarjana Teknik Elektro* 4(1):12–21. doi: 10.12928/biste.v4i1.6079.