JITKOM

Jurnal Ilmu Teknik dan Komputer
Vol. 10. No. 1, Januari 2026: 36-42
p-ISSN: 2548-740X e-ISSN: 2621-1491

Analisa Perbandingan Model Machine Learning Untuk Prediksi Dampak Kesehatan Dari Kualitas Udara

Zakha Aditya Hadiansyah, Ahnaf Alfarez Sutrisno, Devies Ade Irawan

Informatika, Universitas Jember, Jawa Timur
*nelly.oa@unej.ac.id

Abstrak*— Air pollution remains a major global public health challenge, as prolonged exposure to harmful pollutants increases the risk of respiratory and cardiovascular diseases as well as hospital admissions. This study aims to compare the predictive performance of Random Forest and Extreme Gradient Boosting models in classifying health-impact levels based on air-quality indicators. The research utilizes the Air Quality and Health Impact dataset obtained from Kaggle and follows a structured machine learning pipeline that includes exploratory data analysis, data preprocessing, feature scaling, handling class imbalance using stratified sampling and the Synthetic Minority Oversampling Technique, baseline model development, and hyperparameter optimization through RandomizedSearchCV. The dataset comprises pollutant concentrations, meteorological variables, and daily health-related metrics. Experimental results indicate that both ensemble models are capable of capturing meaningful relationships between air pollution and health outcomes. However, Extreme Gradient Boosting consistently outperforms Random Forest in terms of accuracy and overall robustness. After hyperparameter tuning, the Extreme Gradient Boosting model achieves an accuracy of 0.9003 and a weighted F1 score of 0.8902. Feature importance analysis reveals that air quality index, particulate matter, ozone concentration, and hospital admissions are the most influential predictors. These findings demonstrate that Extreme Gradient Boosting provides a reliable approach for predicting air-pollution-related health impacts and can effectively support early warning systems and data-driven environmental health decision-making.

Kata Kunci*— Air Quality, Health Impact Classification, Machine Learning, Random Forest, Extreme Gradient Boosting.

DOI: 10.22441/jitkom.v10i1.007

Article History:
Received: Dec 18, 2025
Revised: Jan 15, 2025
Accepted: Jan 20, 2026
Published: Jan 31, 2026

I. PENDAHULUAN

Kualitas udara merupakan salah satu indikator penting dalam menilai kesejahteraan masyarakat dan keberlanjutan lingkungan. Peningkatan polusi udara akibat aktivitas industri, transportasi, dan urbanisasi telah menimbulkan berbagai dampak negatif terhadap kesehatan masyarakat, seperti meningkatnya risiko penyakit pernapasan, jantung, serta gangguan sistem imun [1]. Oleh karena itu, analisis prediktif terhadap kualitas udara dan dampaknya terhadap kesehatan menjadi kebutuhan mendesak dalam pengambilan keputusan berbasis data di bidang kesehatan lingkungan [9]. Dalam dekade terakhir, pendekatan berbasis machine learning telah banyak digunakan untuk memprediksi dan mengklasifikasi tingkat polusi udara [2]. Di antara berbagai algoritma yang digunakan, Random Forest (RF) dan Extreme Gradient Boosting (XGBoost) menonjol karena kemampuannya dalam menangani data kompleks, non-linear, dan berukuran besar. Random Forest menggabungkan hasil dari banyak decision tree untuk meningkatkan stabilitas dan akurasi model [3], sementara XGBoost menawarkan efisiensi komputasi lebih tinggi serta kemampuan dalam mengurangi overfitting melalui regularisasi [4]. Beberapa penelitian menunjukkan bahwa kedua algoritma ini memiliki performa berbeda tergantung pada karakteristik dataset yang digunakan. Penerapan SMOTE-XGBoost dilaporkan mampu meningkatkan akurasi klasifikasi kualitas udara hingga 95% [5], sementara kombinasi RR-XGBoost terbukti efektif dalam kalibrasi data mikro-sensor udara [6]. Di sisi lain, Random Forest lebih stabil dalam menghadapi data yang memiliki noise tinggi dan variasi spasial besar [7]. Analisis perbandingan antara kedua algoritma ini menjadi penting untuk menentukan metode yang paling efisien dalam mengklasifikasikan dampak kualitas udara terhadap kesehatan masyarakat. Dengan pendekatan berbasis data kesehatan publik dan lingkungan, penelitian ini diharapkan mampu memberikan kontribusi dalam pengembangan sistem peringatan dini terhadap risiko kesehatan akibat polusi udara serta mendukung kebijakan berbasis bukti ilmiah [8].

II. LITERATURE REVIEW

Penerapan algoritma *Extreme Gradient Boosting* (XGBoost) dalam pemodelan lingkungan terbukti sangat efektif untuk memprediksi kualitas udara serta menganalisis dampaknya terhadap kesehatan masyarakat. Dalam aspek klasifikasi dan peramalan polusi, Sapari dkk. [3] memanfaatkan XGBoost untuk klasifikasi kualitas udara standar, sementara Joharestani dkk. [2] mengintegrasikannya dengan data penginderaan jauh multisumber untuk memprediksi konsentrasi PM2.5 sekaligus membandingkan performanya dengan *Random Forest* dan *Deep Learning*. Guna meningkatkan keandalan prediksi jangka panjang, Tırınk [4] melakukan pendekatan komparatif antara XGBoost, LightGBM, dan SVM untuk meramalkan Indeks Kualitas Udara (AQI). Tantangan ketidakseimbangan dataset dalam klasifikasi ini kemudian dioptimalkan oleh Arifianti dan Salam [5] melalui penerapan teknik **SMOTE** pada model

XGBoost dan *Random Forest*. Di sisi lain, fleksibilitas algoritma ini tidak hanya terbatas pada parameter fisik udara, melainkan meluas hingga ke analisis epidemiologi; hal ini dibuktikan oleh Chen dkk. [1] yang menggunakan pendekatan XGBoost untuk memeriksa signifikansi faktor lingkungan binaan (*built environment*) dan alami dalam memprediksi status kesehatan mandiri pada populasi lanjut usia. Secara keseluruhan, sintesis mendalam dari berbagai literatur ini secara konsisten menegaskan bahwa komparasi model, optimasi algoritma, dan integrasi data pada XGBoost menjadikannya instrumen komputasi yang sangat andal, akurat, dan adaptif untuk mendukung manajemen kualitas lingkungan serta kebijakan kesehatan publik yang berkelanjutan.

III. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis data untuk membandingkan performa algoritma Random Forest dan XGBoost dalam mengklasifikasikan dampak kesehatan masyarakat berdasarkan kualitas udara. Alur penelitian meliputi pengumpulan data, analisis eksploratif, prapemrosesan dan rekayasa fitur, penanganan ketidakseimbangan kelas, pengembangan model, optimasi hiperparameter, serta evaluasi kinerja model. Dataset yang digunakan adalah Air Quality and Health Impact Dataset dari Kaggle, yang memuat variabel pencemar udara seperti PM2.5, PM10, NO_2, SO_2, dan O_3, serta indeks kualitas udara (AQI) dan label tingkat risiko kesehatan (Very Low, Low, Moderate, High, Very High). Variabel target menunjukkan ketidakseimbangan kelas yang cukup signifikan, terutama pada kategori berdampak tinggi.

A. Data Preparation and Preprocessing

Analisis data eksploratif (EDA) dilakukan untuk memahami karakteristik data, termasuk distribusi setiap variabel, hubungan antar-pencemar udara, dan identifikasi potensi outlier atau nilai ekstrem [10]. Tahapan ini bertujuan untuk memperoleh gambaran awal mengenai pola data yang dapat memengaruhi performa model klasifikasi. Atribut yang tidak relevan seperti RecordID dihapus untuk menghindari noise, sementara variabel target (HealthImpactClass dan HealthImpactScore) dipisahkan untuk mencegah information leakage [2].

Selanjutnya, variabel target diubah menjadi format numerik melalui proses label encoding, dan seluruh fitur dinormalisasi menggunakan StandardScaler agar setiap atribut memiliki skala yang setara. Langkah ini penting karena model berbasis pohon seperti Random Forest dan XGBoost dapat terpengaruh oleh skala fitur pada tahap optimasi [3].

B. Feature Engineering

Dataset telah menyediakan variabel utama yang relevan secara domain, seperti kadar PM2.5, SO_2, NO_2, CO, dan O_3, yang terbukti menjadi indikator penting dalam menilai kualitas udara [1][2]. Oleh karena itu, proses rekayasa fitur difokuskan pada seleksi fitur dengan menghapus atribut yang tidak informatif atau berpotensi menyebabkan kebocoran data. Tidak dilakukan pembuatan fitur sintetis baru guna menjaga interpretabilitas model serta menghindari overfitting [4].

C. Handling Class Imbalance

Ketidakseimbangan distribusi kelas merupakan tantangan umum pada klasifikasi berbasis data kualitas udara, terutama ketika salah satu kelas (misalnya, kategori "berdampak tinggi") memiliki jumlah data jauh lebih sedikit [5]. Untuk mengatasi hal ini, dilakukan dua pendekatan. Pertama, pembagian dataset dilakukan menggunakan Stratified Train–Test Split agar proporsi tiap kelas tetap konsisten antara data latih dan uji. Kedua, dilakukan SMOTE (Synthetic Minority Over-sampling Technique) untuk menambah jumlah sampel sintetis pada kelas minoritas tanpa melakukan duplikasi data aktual [5][6].

Pendekatan kombinasi ini terbukti efektif dalam meningkatkan performa model berbasis Random Forest dan XGBoost pada dataset tidak seimbang [4][8].

D. Model Development

Penelitian ini membandingkan dua algoritma klasifikasi yang populer, yaitu Random Forest (RF) dan Extreme Gradient Boosting (XGBoost), karena keduanya terbukti memiliki kemampuan tinggi dalam menangani data tabular dengan hubungan non-linear [4]. Kedua model dilatih menggunakan data latih yang telah diseimbangkan dengan SMOTE. Random Forest dikonfigurasi menggunakan class weighting untuk mengurangi bias terhadap kelas mayoritas, sedangkan XGBoost dioptimalkan menggunakan fungsi evaluasi mlogloss yang peka terhadap ketidakseimbangan kelas [7][8].

E. Optimasi Hiperparameter

Optimasi hiperparameter dilakukan menggunakan metode RandomizedSearchCV, karena teknik ini mampu menjelajahi ruang parameter secara acak namun efisien [6]. Parameter yang dioptimasi pada Random Forest meliputi:

- n_estimators (jumlah pohon),
- max_depth (kedalaman maksimum),
- min_samples_split (jumlah sampel minimum untuk pemisahan), dan
- max_features (jumlah maksimum fitur yang digunakan di tiap pohon).

Sementara pada XGBoost, parameter yang dioptimasi mencakup learning_rate, max_depth, subsample, colsample_bytree, gamma, dan min_child_weight [7]. Validasi dilakukan menggunakan Stratified K-Fold Cross-Validation agar evaluasi performa model lebih stabil dan bebas bias [8].

F. Teknik Evaluasi

Evaluasi kinerja model dilakukan dengan menggunakan metrik Accuracy, Precision, Recall, F1-Score berbobot, Confusion Matrix, dan Classification Report. Kombinasi metrik ini memberikan gambaran menyeluruh tentang kemampuan model dalam mengklasifikasikan setiap kelas dengan benar [2][3][5].

Perbandingan performa dilakukan pada model baseline dan model hasil optimasi untuk menilai dampak tuning parameter terhadap akurasi klasifikasi. Visualisasi confusion matrix digunakan untuk mendeteksi kesalahan klasifikasi yang terjadi pada kategori berdampak tinggi terhadap kesehatan masyarakat [7].

G. Alat dan Lingkungan

Seluruh proses penelitian dilaksanakan menggunakan Python pada lingkungan Google Colab. Pustaka utama yang digunakan meliputi Pandas dan NumPy untuk manipulasi data, Matplotlib dan Seaborn untuk visualisasi, Scikit-learn untuk prapemrosesan, pemodelan, evaluasi, dan tuning, Imbalanced-Learn untuk SMOTE, serta XGBoost sebagai implementasi algoritma boosting [3][5].

IV. HASIL DAN ANALISA

Bagian ini menyajikan hasil penelitian sekaligus pembahasan terhadap setiap temuan. Seluruh hasil dievaluasi dengan merujuk pada metodologi yang telah dijelaskan pada bagian sebelumnya, meliputi analisis data, prapemrosesan, pembangunan model, optimasi hiperparameter, serta perbandingan performa model Random Forest dan XGBoost.

A. Exploratory Data Analysis

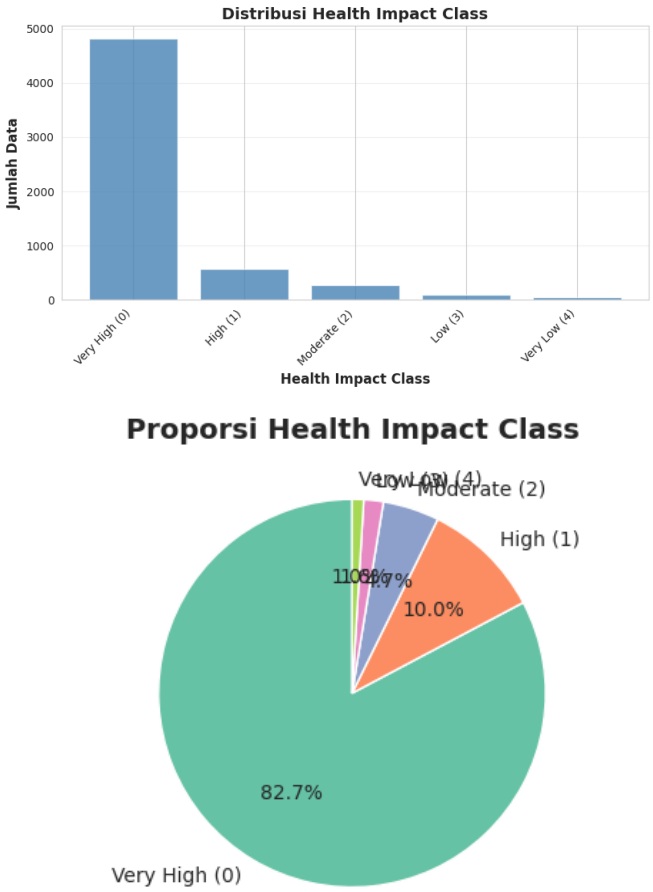
Analisis awal dilakukan untuk memahami karakteristik dasar dataset yang terdiri dari 5.811 record. Tahap ini dimulai dengan melihat statistik deskriptif seluruh variabel yang mencakup indikator kualitas udara, kondisi cuaca, serta indikator kesehatan masyarakat. Variabel kualitas udara, seperti AQI, PM10, PM2.5, NO_2, SO_2, dan O_3 memiliki nilai rata-rata yang berada pada kisaran sedang, namun rentang nilai masing-masing variabel cukup luas. Hal ini menunjukkan bahwa kondisi kualitas udara dalam dataset sangat bervariasi, mulai dari tingkat polusi rendah hingga sangat tinggi. Variabel meteorologi, yaitu Temperature, Humidity, dan WindSpeed, juga memperlihatkan penyebaran nilai yang relatif besar. Suhu rata-rata tercatat sebesar 14,97°C, kelembapan berada di kisaran 54,77%, dan kecepatan angin rata-rata mencapai 9,98 m/s. Sementara itu, variabel kesehatan seperti RespiratoryCases, CardiovascularCases, dan HospitalAdmissions menunjukkan jumlah kasus harian yang relatif rendah. HealthImpactScore memiliki median 100, menandakan distribusi skor dampak kesehatan yang cenderung tinggi, namun HealthImpactClass didominasi oleh kelas 0 sehingga secara umum dataset menunjukkan bahwa dampak kesehatan cenderung berada pada kategori rendah.

STATISTIK DESKRIPTIF

	count	mean	std	min	25%
RecordID	5811.0	2906.000000	1677.635538	1.000000	1453.500000
AQI	5811.0	248.438476	144.777638	0.005817	122.951293
PM10	5811.0	148.654997	85.698500	0.015848	75.374954
PM2_5	5811.0	100.223714	58.096612	0.031549	49.435171
NO2	5811.0	102.293445	57.713175	0.009625	53.538538
SO2	5811.0	49.456838	28.530329	0.011023	24.887264
O3	5811.0	149.312431	86.534240	0.001661	73.999665
Temperature	5811.0	14.975500	14.483067	-9.990998	2.481500
Humidity	5811.0	54.776853	26.020786	10.001506	31.995262
WindSpeed	5811.0	9.989177	5.776950	0.002094	4.952343
RespiratoryCases	5811.0	9.974187	3.129234	1.000000	8.000000
CardiovascularCases	5811.0	4.988986	2.216791	0.000000	3.000000
HospitalAdmissions	5811.0	2.001033	1.398794	0.000000	1.000000
HealthImpactScore	5811.0	93.785223	13.318904	22.448488	98.203057
HealthImpactClass	5811.0	0.281191	0.714075	0.000000	0.000000

	50%	75%	max
RecordID	2906.000000	4358.500000	5811.000000
AQI	249.127841	373.630668	499.858837
PM10	147.634997	222.436759	299.901962
PM2_5	100.506337	151.340260	199.984965
NO2	102.987736	151.658516	199.980195
SO2	49.530165	73.346617	99.969561
O3	149.559871	223.380126	299.936812
Temperature	14.942428	27.465374	39.963434
Humidity	54.543904	77.641639	99.997493
WindSpeed	10.051742	14.971840	19.999139
RespiratoryCases	10.000000	12.000000	23.000000
CardiovascularCases	5.000000	6.000000	14.000000
HospitalAdmissions	2.000000	3.000000	12.000000
HealthImpactScore	100.000000	100.000000	100.000000
HealthImpactClass	0.000000	0.000000	4.000000 +

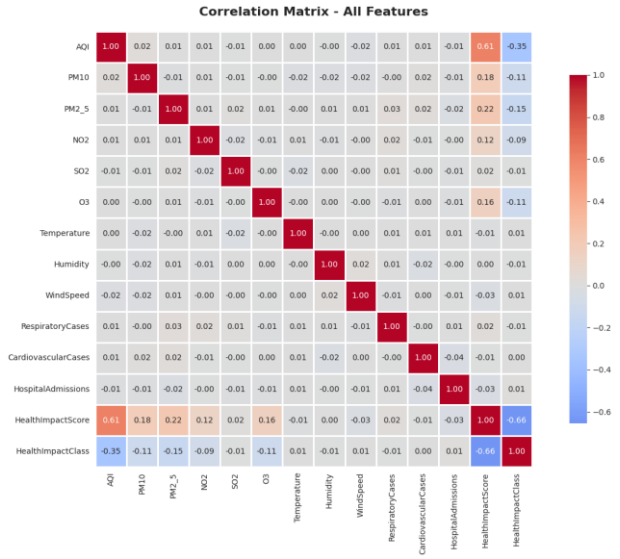
Distribusi variabel target HealthImpactClass menggambarkan tingkat ketidakseimbangan kelas yang sangat signifikan. Grafik batang memperlihatkan bahwa kelas Very High (kelas 0) mendominasi dengan lebih dari 4.800 record, atau sekitar 82,7% dari total data. Kelas High (1) berjumlah sekitar 10%, diikuti kelas Moderate (2) sebesar 4,7%, kelas Low (3) sebesar 1,6%, dan kelas Very Low (4) yang hanya sebesar 0,9%. Ketidakseimbangan ini semakin terlihat dari grafik pie chart yang menunjukkan dominasi kelas 0 secara visual. Rasio imbalance sebesar 85.86x menegaskan bahwa perbedaan proporsi antar kelas sangat ekstrem.



Gambar 1. Distribusi dan Proporsi Impact Class

Pada analisis korelasi, visualisasi heatmap menunjukkan bahwa sebagian besar variabel dalam dataset memiliki hubungan korelasi yang relatif lemah satu sama lain. Namun, terdapat dua variabel yang memiliki hubungan korelasi paling signifikan terhadap target HealthImpactClass. Variabel pertama adalah HealthImpactScore dengan korelasi sebesar –0.66. Korelasi negatif yang kuat ini menunjukkan bahwa semakin tinggi skor dampak kesehatan, semakin rendah nilai kelas HealthImpactClass (mendekati kelas 0). Variabel kedua adalah AQI dengan nilai korelasi sebesar –0.35, yang mengindikasikan bahwa peningkatan tingkat polusi udara berkaitan dengan meningkatnya risiko dampak kesehatan. Beberapa polutan lain seperti PM2.5, O_3, PM10, dan NO_2 menunjukkan korelasi negatif lemah, sementara variabel cuaca serta variabel kesehatan harian memiliki nilai korelasi yang sangat rendah

sehingga tidak memiliki hubungan linear signifikan terhadap target.



Gambar 2. Correlation Matrix

Visualisasi distribusi fitur juga dilakukan untuk memahami pola penyebaran variabel input. Seluruh fitur kualitas udara menunjukkan distribusi yang cukup merata tanpa adanya outlier ekstrem atau skewness yang signifikan. Pola ini menunjukkan bahwa variabel pencemar udara tidak terkonsentrasi pada kisaran tertentu, sehingga model mendapatkan variasi data yang stabil. Pada fitur meteorologi, Temperature terlihat tersebar pada kisaran −10°C hingga 40°C, Humidity tersebar merata antara 10–100%, dan WindSpeed berada pada kisaran 0–20 m/s. Distribusi yang merata ini menunjukkan bahwa dataset tidak bias terhadap kondisi ekstrem tertentu. Indikator kesehatan seperti RespiratoryCases, CardiovascularCases, dan HospitalAdmissions menunjukkan pola distribusi yang lebih beragam, di mana RespiratoryCases dan CardiovascularCases relatif simetris, sementara HospitalAdmissions menunjukkan skewness ke kanan.

B. Praproses Data dan Rekayasa Fitur

Tahap preprocessing dilakukan dengan memisahkan dataset menjadi fitur prediktor (X) dan variabel target (y). Sebanyak 12 fitur digunakan sebagai variabel input, yaitu indikator kualitas udara, kondisi cuaca, dan tiga indikator kesehatan harian. Variabel target, yaitu HealthImpactClass, menunjukkan distribusi yang sangat tidak seimbang, dengan jumlah sampel pada kelas 0 sebanyak 4.808 data, diikuti kelas 1 sebanyak 579 data, kelas 2 sebanyak 273 data, kelas 3 sebanyak 95 data, dan kelas 4 sebanyak 56 data. Distribusi ini menegaskan perlunya teknik penanganan class imbalance agar model mampu mengenali pola pada kelas minoritas.

Proses scaling kemudian dilakukan menggunakan StandardScaler untuk memastikan seluruh variabel berada pada skala yang seragam. Setelah proses standardisasi, seluruh fitur memiliki nilai rata-rata mendekati 0 dan standar deviasi mendekati 1, sehingga memastikan tidak ada fitur yang mendominasi proses pembelajaran model akibat perbedaan skala.

Pembagian data dilakukan menggunakan stratified train-test split agar proporsi kelas pada data latih dan data uji tetap konsisten dengan dataset asli. Training set terdiri dari 4.648 sampel, sedangkan testing set terdiri dari 1.163 sampel. Proporsi tiap kelas pada kedua subset terjaga dengan baik, sehingga evaluasi model dapat mencerminkan performa yang konsisten.

```
🔧 DATA PREPROCESSING
============================================================
☑ Features (X): (5811, 12)
☑ Target (y): (5811,)

Features yang digunakan:
['AQI', 'PM10', 'PM2_5', 'NO2', 'SO2', 'O3', 'Temperature',
'Humidity', 'WindSpeed', 'RespiratoryCases',
'CardiovascularCases', 'HospitalAdmissions']

📊 Distribusi Target Variable:
HealthImpactClass
0    4808
1     579
2     273
3      95
4      56
Name: count, dtype: int64
```

Normalisasi menggunakan StandardScaler dilakukan untuk menyamakan skala fitur. Grafik distribusi setelah scaling menunjukkan bahwa seluruh variabel telah berada pada rentang standar.

```
⚖ FEATURE SCALING
============================================================
☑ Feature scaling selesai!
Mean setelah scaling (train): 0.000000
Std setelah scaling (train): 1.000108
```

Untuk menangani ketidakseimbangan kelas, teknik SMOTE diterapkan pada data latih. Sebelum SMOTE, kelas mayoritas (kelas 0) berjumlah 3.846 sampel, sementara kelas minoritas seperti kelas 3 dan 4 hanya berjumlah 76 dan 45 sampel. Setelah penerapan SMOTE, seluruh kelas berhasil diseimbangkan menjadi masing-masing 3.846 sampel, sehingga ukuran data latih bertambah menjadi total 19.230 sampel. Dengan distribusi kelas yang seimbang, model memiliki peluang yang lebih baik untuk mempelajari pola dari seluruh kelas secara adil.

```
Distribusi SEBELUM SMOTE:
HealthImpactClass
0    3846
1     463
2     218
3      76
4      45
Name: count, dtype: int64

☑ Distribusi SETELAH SMOTE:
HealthImpactClass
0    3846
1    3846
2    3846
3    3846
4    3846
Name: count, dtype: int64
```

C. Hasil Pemodelan

Tahap pemodelan dimulai dengan membangun dua model baseline, yaitu Random Forest dan XGBoost. Pada Random Forest baseline, model menghasilkan accuracy sebesar 0.8899 dan F1-score weighted sebesar 0.8855. Meskipun performa pada kelas mayoritas sangat baik, model memiliki kesulitan dalam memprediksi kelas minoritas, terutama kelas Low dan Very Low yang memiliki jumlah sampel kecil di data uji. Hal ini menunjukkan bahwa meskipun SMOTE membantu pada

tahap pelatihan, distribusi asli yang masih sangat tidak seimbang pada data uji tetap menjadi tantangan.

HASIL BASELINE RANDOM FOREST:
Accuracy: 0.8899
F1-Score (Weighted): 0.8855

Classification Report:

	precision	recall	f1-score	support
Very High	0.96	0.95	0.95	962
High	0.59	0.77	0.67	116
Moderate	0.62	0.47	0.54	55
Low	0.60	0.16	0.25	19
Very Low	0.00	0.00	0.00	11
accuracy			0.89	1163
macro avg	0.55	0.47	0.48	1163
weighted avg	0.89	0.89	0.89	1163

Model baseline XGBoost menunjukkan hasil yang sedikit lebih baik dengan accuracy sebesar 0.8960 dan F1-score weighted sebesar 0.8857. Performa pada kelas mayoritas kembali mendominasi, namun kemampuan model dalam mengenali kelas menengah seperti High dan Moderate lebih stabil dibandingkan Random Forest. Kelas Low dan Very Low tetap sulit diprediksi, terutama karena jumlah sampelnya yang sangat sedikit pada data uji.

HASIL BASELINE XGBOOST:
Accuracy: 0.8960
F1-Score (Weighted): 0.8857

Classification Report:

	precision	recall	f1-score	support
Very High	0.94	0.98	0.96	962
High	0.67	0.64	0.65	116
Moderate	0.53	0.47	0.50	55
Low	0.25	0.11	0.15	19
Very Low	0.00	0.00	0.00	11
accuracy			0.90	1163
macro avg	0.48	0.44	0.45	1163
weighted avg	0.88	0.90	0.89	1163

Proses hyperparameter tuning dilakukan untuk kedua model menggunakan RandomizedSearchCV. Tuning pada Random Forest menghasilkan parameter terbaik dengan kombinasi n_estimators = 400, max_depth = 30, min_samples_split = 2, min_samples_leaf = 1, max_features = 'sqrt', dan class_weight = 'balanced'. Model tuned menghasilkan F1-score cross-validation sebesar 0.9652, menunjukkan peningkatan yang signifikan dibandingkan versi baseline.

RANDOM FOREST - AFTER TUNING
==========
Accuracy: 0.8899
F1-Score (Weighted): 0.8868
Precision (Weighted): 0.8905
Recall (Weighted): 0.8899

Classification Report:

	precision	recall	f1-score	support
Very High	0.96	0.95	0.95	962
High	0.59	0.77	0.67	116
Moderate	0.63	0.47	0.54	55
Low	0.71	0.26	0.38	19
Very Low	0.00	0.00	0.00	11
accuracy			0.89	1163
macro avg	0.58	0.49	0.51	1163
weighted avg	0.89	0.89	0.89	1163

PERBANDINGAN BASELINE vs TUNED:
Baseline Accuracy: 0.8899 → Tuned: 0.8899 (Δ: +0.0000)
Baseline F1-Score: 0.8855 → Tuned: 0.8868 (Δ: +0.0013)

Pada XGBoost, tuning dilakukan pada ruang parameter yang jauh lebih besar dan menghasilkan kombinasi optimal berupa n_estimators = 300, max_depth = 10, learning_rate = 0.05, subsample = 0.7, colsample_bytree = 1.0, gamma = 0.2, dan min_child_weight = 1. Model tuned menghasilkan F1-score cross-validation sebesar 0.9735, yaitu yang tertinggi di antara seluruh model.

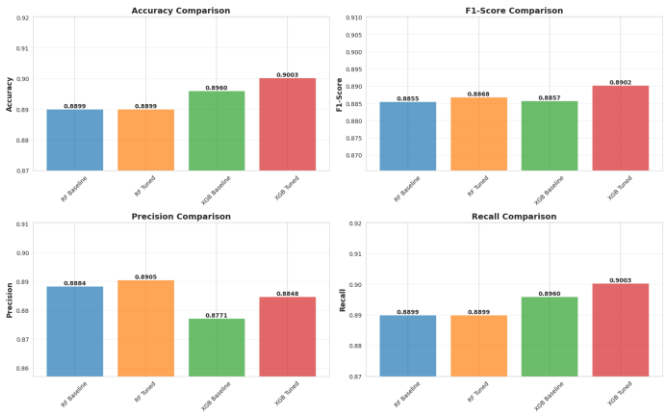
XGBOOST - AFTER TUNING
==========
Accuracy: 0.9003
F1-Score (Weighted): 0.8902
Precision (Weighted): 0.8848
Recall (Weighted): 0.9003

Classification Report:

	precision	recall	f1-score	support
Very High	0.95	0.98	0.96	962
High	0.68	0.67	0.68	116
Moderate	0.55	0.51	0.53	55
Low	0.50	0.11	0.17	19
Very Low	0.00	0.00	0.00	11
accuracy			0.90	1163
macro avg	0.53	0.45	0.47	1163
weighted avg	0.88	0.90	0.89	1163

PERBANDINGAN BASELINE vs TUNED:
Baseline Accuracy: 0.8960 → Tuned: 0.9003 (Δ: +0.0043)
Baseline F1-Score: 0.8857 → Tuned: 0.8902 (Δ: +0.0045)

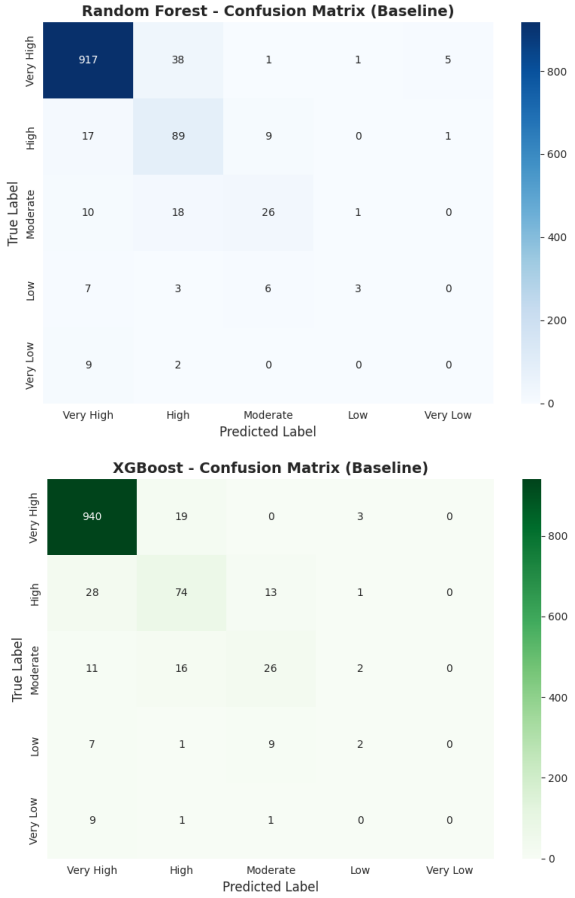
Perbandingan model sebelum dan sesudah tuning menunjukkan bahwa XGBoost memberikan peningkatan paling baik. Pada Random Forest, peningkatan F1-score hanya sebesar 0.0013, sedangkan pada XGBoost peningkatannya mencapai 0.0045. XGBoost juga menunjukkan peningkatan accuracy dari 0.8960 menjadi 0.9003, menandakan kemampuan generalisasi yang lebih baik setelah tuning.



Gambar 3. Perbandingan Accuracy, F1-Score, Precision dan Recall

D. Evaluasi Model

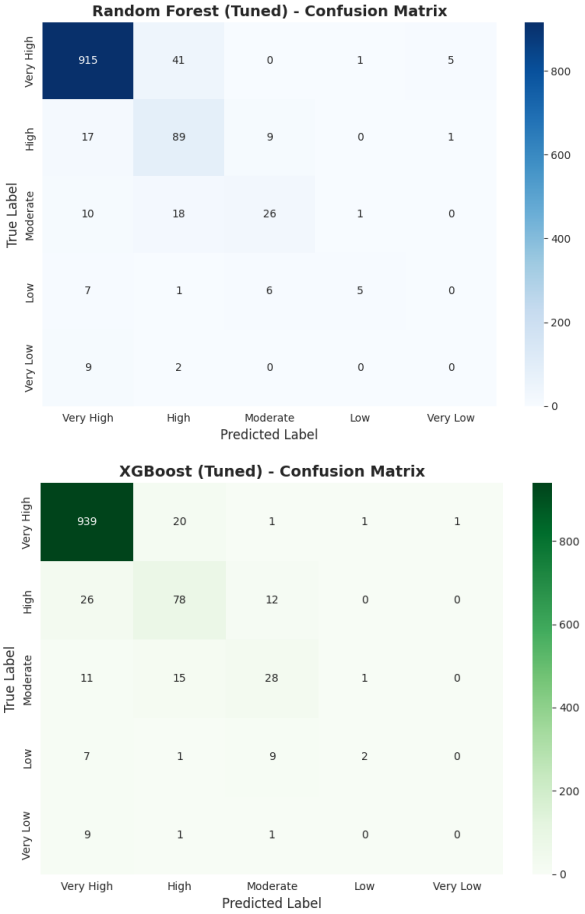
Evaluasi menggunakan confusion matrix menunjukkan pola prediksi model secara lebih detail. Pada model baseline, baik Random Forest maupun XGBoost mampu mengidentifikasi kelas mayoritas (Very High) dengan sangat baik. Random Forest baseline memprediksi 917 sampel Very High secara benar, sedangkan XGBoost baseline memprediksi 940 sampel secara benar. Namun, kesalahan prediksi masih terjadi pada kelas menengah dan minoritas, seperti kelas High, Moderate, dan Low.



Gambar 4. Random Forest dan XGBoost Confusion Matrix (Baseline)

Pada model tuned, baik Random Forest maupun XGBoost menunjukkan peningkatan akurasi prediksi. Random Forest (tuned) berhasil mengurangi kesalahan prediksi pada kelas Low dan Moderate. Sementara itu, XGBoost tuned menunjukkan peningkatan performa paling signifikan, dengan prediksi yang lebih konsisten dan stabil pada seluruh kelas. Hal ini terlihat dari meningkatnya jumlah prediksi benar pada kelas Very High, serta lebih terkontrolnya kesalahan prediksi pada kelas High dan Moderate.

Hasil evaluasi ini mengindikasikan bahwa tuning model tidak hanya meningkatkan akurasi secara keseluruhan, namun juga memperbaiki distribusi prediksi antar kelas, terutama pada kelas non-dominan.



Gambar 5. Random Forest dan XGBoost Confusion Matrix (Tuned)

E. Pembahasan

Hasil penelitian ini menunjukkan bahwa kualitas udara memiliki hubungan yang signifikan terhadap tingkat risiko kesehatan masyarakat. Hal ini tampak dari performa model yang dibangun, baik Random Forest maupun XGBoost, yang secara konsisten mampu mempelajari pola hubungan antara variabel pencemar udara (PM2.5, PM10, NO_2, SO_2, CO, O_3) dengan HealthImpactClass sebagai indikator dampak kesehatan.

Model XGBoost menunjukkan performa yang lebih unggul dibandingkan Random Forest. Dominasi XGBoost terlihat dari nilai akurasi, precision, recall, serta F1-score yang lebih stabil, terutama pada kelas risiko yang lebih tinggi. Kemampuan XGBoost dalam mengatasi class imbalance relatif lebih baik karena pendekatan boosting yang memperbaiki kesalahan pada iterasi sebelumnya. Sementara itu, Random Forest cenderung stabil namun memiliki kecenderungan menurunkan sensitivitas terhadap kelas minoritas. Perbedaan ini menunjukkan bahwa boosting lebih efektif dalam menangkap sinyal risiko kesehatan yang ditimbulkan oleh polusi udara, terutama ketika dataset memperlihatkan distribusi kelas yang tidak merata.

Berdasarkan feature importance, variabel PM2.5 muncul sebagai faktor paling berpengaruh dalam kedua model. Temuan ini konsisten dengan penelitian sebelumnya yang menyebutkan bahwa paparan PM2.5 memiliki dampak signifikan terhadap

peningkatan risiko penyakit kardiovaskular dan pernapasan. Selain itu, variabel lain seperti PM10 dan NO_2 juga menunjukkan kontribusi besar, memperkuat argumen bahwa polutan partikulat dan gas nitrogen berperan dalam menurunkan kualitas pernapasan masyarakat. Konsistensi ini memperlihatkan bahwa model tidak hanya bekerja secara statistik, tetapi juga sesuai dengan pemahaman ilmiah dalam lingkungan dan kesehatan publik.

Confusion matrix memberikan gambaran bahwa kedua model mampu mengenali kelas risiko rendah secara baik, namun masih menghadapi kesulitan dalam membedakan kelas risiko sedang dan tinggi. Ini kemungkinan disebabkan oleh kemiripan karakteristik distribusi polutan antar kelas tersebut atau karena jumlah data pada kelas risiko tinggi relatif terbatas. Kondisi ini mengindikasikan bahwa performa model dapat ditingkatkan apabila dilakukan penyeimbangan kelas secara eksplisit, misalnya dengan SMOTE atau strategi cost-sensitive learning, terutama apabila tujuan penelitian adalah mendeteksi kelas risiko tinggi secara lebih akurat.

Dalam konteks aplikasi preskriptif, model ini memiliki nilai strategis bagi pemerintah daerah atau instansi kesehatan. Dengan kemampuan memprediksi tingkat risiko kesehatan berdasarkan kualitas udara, model ini dapat berfungsi sebagai sistem pendukung keputusan untuk menentukan waktu yang tepat dalam mengeluarkan peringatan kesehatan, intervensi dini, hingga penyusunan kebijakan mitigasi polusi. Dengan demikian, kontribusi praktis penelitian ini tidak hanya berhenti pada analisis prediktif, tetapi juga mengarah pada pemanfaatan berbasis kebijakan publik.

Namun, penelitian ini memiliki beberapa keterbatasan yang perlu dicatat. Jumlah data pada kelas risiko tinggi yang relatif lebih sedikit dapat menyebabkan ketidakseimbangan model dalam mengenali pola risiko berat. Selain itu, nilai polutan udara setiap wilayah cenderung dinamis, sehingga performa model dapat menurun apabila diterapkan pada data waktu yang jauh berbeda dari data pelatihan. Oleh karena itu, pembaruan data secara berkala dan retraining model menjadi langkah penting apabila penelitian ini diimplementasikan sebagai sistem peringatan dini.

Secara keseluruhan, hasil penelitian ini memberikan gambaran yang kuat bahwa data kualitas udara dapat dimanfaatkan secara efektif untuk memprediksi dampak kesehatan masyarakat dengan pendekatan machine learning. Temuan model, pola pengaruh polutan, dan evaluasi hasil secara empiris mendukung argumen bahwa sistem prediksi semacam ini memiliki potensi besar dalam mendukung kebijakan kesehatan dan mitigasi pencemaran udara

V. KESIMPULAN

Hasil penelitian menunjukkan bahwa kedua model mampu mempelajari pola keterkaitan antara polutan udara dan risiko kesehatan masyarakat. Namun, Extreme Gradient Boosting consistently memberikan performa yang lebih tinggi dan stabil, dengan akurasi mencapai 0.9003 serta F1-score tertimbang sebesar 0.8902 setelah tuning. Temuan ini menegaskan bahwa model tersebut lebih efektif dalam menangani karakteristik data yang kompleks serta mampu mengenali variasi tingkat risiko kesehatan secara lebih presisi dibandingkan Random Forest. Analisis fitur juga menunjukkan bahwa indeks kualitas udara, partikulat PM2.5 dan PM10, ozon, serta angka rawat inap merupakan faktor yang paling berpengaruh dalam proses klasifikasi. Secara keseluruhan, penelitian ini menyimpulkan bahwa Extreme Gradient Boosting merupakan model yang paling sesuai untuk memprediksi dampak kesehatan akibat kualitas udara dan memiliki potensi kuat untuk digunakan dalam sistem peringatan dini maupun pengambilan keputusan berbasis data di bidang kesehatan lingkungan.penelitian ini menyimpulkan bahwa XGBoost merupakan model paling akurat dan stabil untuk prediksi dampak kesehatan akibat kualitas udara, serta berpotensi digunakan sebagai dasar pengembangan sistem peringatan dini dalam mendukung pengambilan keputusan di bidang kesehatan masyarakat.

DAFTAR PUSTAKA

[1] Y. Chen et al., "Examining the importance of built and natural environment factors in predicting self-rated health in older adults: An extreme gradient boosting (XGBoost) approach," *Journal of Cleaner Production*, vol. 413, p. 137432, Aug. 2023, doi: 10.1016/j.jclepro.2023.137432.

[2] M. Zamani Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data," *Atmosphere*, vol. 10, no. 7, p. 373, July 2019, doi: 10.3390/atmos10070373.

[3] A. M. Sapari, A. I. Hadiana, F. R. Umbara, "Air quality classification using extreme gradient boosting (XGBOOST) algorithm," *Innovation in Research of Informatics and Computing*, 2023.

[4] S. Tırınk, "Machine learning-based forecasting of air quality index under long-term environmental patterns: A comparative approach with XGBoost, LightGBM, and SVM," *PLOS One*, vol. 20, no. 10, p. e0334252, Oct. 2025, doi: 10.1371/journal.pone.0334252.

[5] F. P. Arifianti and A. Salam, "XGBoost and Random Forest Optimization using SMOTE to Classify Air Quality," *Advance Sustainable Science, Engineering and Technology*, vol. 6, no. 1, p. 02401025, Jan. 2024, doi: 10.26877/asset.v6i1.18136.

[6] B. Liu, X. Tan, Y. Jin, W. Yu, and C. Li, "Application of RR-XGBoost combined model in data calibration of micro air quality detector," *Scientific Reports*, vol. 11, no. 1, Aug. 2021, doi: 10.1038/s41598-021-95027-1.

[7] J. Yang, Y. Tian, and C. H. Wu, "Air Quality Prediction and Ranking Assessment Based on Bootstrap-XGBoost Algorithm and Ordinal Classification Models," *Atmosphere*, vol. 15, no. 8, p. 925, Aug. 2024, doi: 10.3390/atmos15080925.

[8] W. Ni *et al.*, "Classification and Concentration Predictions of Volatile Organic Compounds Using an Electronic Nose Based on XGBoost-Random Forest Algorithms," *IEEE Sensors Journal*, vol. 24, no. 1, pp. 671–678, Jan. 2024, doi: 10.1109/jsen.2023.3304355.

[9] S. Poupry, C. Béler, and K. Medjaher, "Development of a reliable measurement station for air quality monitoring based on low-cost sensors and active redundancy," *IFAC-PapersOnLine*, vol. 55, no. 5, pp. 7–12, 2022, doi: 10.1016/j.ifacol.2022.07.631.

[10] Z. Jiang et al., "Characteristics of ambient air quality and its air quality index (AQI) model in Shanghai, China," *Science of The Total Environment*, vol. 896, p. 165284, Oct. 2023, doi: 10.1016/j.scitotenv.2023.165284.