Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 763−773
DOI 10.47065/bits.v8i2.10602
https://ejurnal.seminar-id.com/index.php/bits

Implementasi Teknik SMOTE Menggunakan Random Forest dan XGBoost pada Klasifikasi Tingkat Kualitas Udara

Implementation of the SMOTE Technique Using Random Forest and XGBoost for Air Quality Level Classification

Nanda Putri Karizki*, Heni Sulistiani

[1] Fakultas Teknik dan Ilmu Komputer, Sistem Informasi, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
[2] Fakultas Teknik dan Ilmu Komputer, Magister Ilmu Komputer, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
Email: [1,*]nanda_putri_karizki@teknokrat.ac.id, [2]henisulistiani@teknokrat.ac.id
Email Penulis Korespondensi: nanda_putri_karizki@teknokrat.ac.id
Received: **04/07/2026**, *Revised*: **17/07/2026**, *Accepted*: **01/09/2026**, *Published*: **08/09/2026**

Abstrak−Polusi udara merupakan isu lingkungan global yang berdampak signifikan terhadap kesehatan masyarakat dan keberlanjutan lingkungan. Permasalahan ini menjadi semakin mendesak karena distribusi kelas pada data kualitas udara umumnya tidak seimbang, sehingga kelas dengan tingkat bahaya tertinggi seperti Hazardous justru menjadi kelas minoritas yang paling sulit dideteksi oleh model klasifikasi konvensional. Penelitian ini mengimplementasikan pendekatan *machine learning* menggunakan algoritma Random Forest dan XGBoost yang dikombinasikan dengan teknik Synthetic Minority Oversampling Technique (SMOTE) untuk mengklasifikasikan tingkat kualitas udara. Dataset terdiri dari 5.000 sampel dengan 9 fitur lingkungan dan demografi. Kualitas udara dikategorikan ke dalam empat kelas: Good, Moderate, Poor, dan Hazardous, dengan distribusi kelas asal yang tidak seimbang (Good 40%, Moderate 30%, Poor 20%, Hazardous 10%). SMOTE diterapkan hanya pada data *training* untuk menyeimbangkan distribusi kelas. Hasil menunjukkan XGBoost dengan SMOTE mencapai performa terbaik dengan akurasi 0,952, *F1-Score* 0,952, dan rata-rata *cross-validation* 0,969. Dibandingkan penelitian sebelumnya menggunakan Decision Tree dengan akurasi 0,930 tanpa SMOTE, penelitian ini menunjukkan peningkatan sebesar 0,022. Pada kelas minoritas Hazardous, XGBoost dengan SMOTE meningkatkan recall dari 0,80 menjadi 0,87 dibandingkan Random Forest tanpa SMOTE, dengan precision dan F1-Score yang juga lebih seimbang (0,87). CO dan Proximity to Industrial Areas menjadi fitur dominan, berbeda dengan PM2.5 pada penelitian sebelumnya. Temuan ini mengonfirmasi bahwa kombinasi *ensemble method* dengan SMOTE efektif menangani ketidakseimbangan kelas. Evaluasi dilakukan menggunakan metrik *precision*, *recall*, *F1-Score*, serta validasi 5-fold *stratified cross-validation* untuk memastikan kestabilan model pada seluruh kelas, termasuk kelas minoritas Hazardous yang paling kritis bagi kesehatan masyarakat. Hasil ini mengindikasikan bahwa kombinasi algoritma ensemble dengan teknik penyeimbangan data dapat menjadi acuan praktis dalam pengembangan sistem pemantauan kualitas udara berbasis *machine learning*.

Kata Kunci: Ensemble Learning; Ketidakseimbangan Kelas; Kualitas Udara; Random Forest; SMOTE; XGBoost

Abstract−Air pollution is a global environmental issue with significant impacts on public health and environmental sustainability. The problem is compounded by the generally imbalanced class distribution in air quality data; consequently, the highest-risk category "Hazardous" often constitutes the minority class, making it the most difficult to detect using conventional classification models. This study implements a machine learning approach using Random Forest and XGBoost algorithms combined with the Synthetic Minority Oversampling Technique (SMOTE) to classify air quality levels. The dataset comprises 5,000 samples featuring nine environmental and demographic variables. Air quality is categorized into four classes Good, Moderate, Poor, and Hazardous with an initial imbalanced distribution (Good: 40%, Moderate: 30%, Poor: 20%, Hazardous: 10%). SMOTE was applied exclusively to the training data to balance the class distribution. Results indicate that XGBoost combined with SMOTE achieved the best performance, yielding an accuracy of 0.952, an F1-Score of 0.952, and an average cross-validation score of 0.969. This represents a 0.022 improvement over a previous study that utilized a Decision Tree model without SMOTE (achieving 0.930 accuracy). For the "Hazardous" minority class, the XGBoost-SMOTE combination improved recall from 0.80 (Random Forest without SMOTE) to 0.87, while also achieving more balanced precision and F1-Score values (0.87). CO levels and proximity to industrial areas emerged as the dominant features, contrasting with PM2.5 in the earlier study. These findings confirm that combining ensemble methods with SMOTE effectively addresses class imbalance. Evaluation was conducted using precision, recall, and F1-Score metrics, alongside 5-fold stratified cross-validation to ensure model stability across all classes including the "Hazardous" minority class, which is most critical for public health. These results suggest that combining ensemble algorithms with data balancing techniques can serve as a practical reference for developing machine learning-based air quality monitoring systems.

Keywords: Air Quality; Class Imbalance; Ensemble Learning; Random Forest; SMOTE; XGBoost

1. PENDAHULUAN

Kualitas udara merupakan salah satu indikator utama kesehatan lingkungan yang berdampak langsung terhadap kesehatan masyarakat dan keberlanjutan ekosistem. Paparan polutan udara dalam jangka panjang terbukti meningkatkan risiko penyakit pernapasan, kardiovaskular, bahkan kematian dini [1]. Seiring meningkatnya urbanisasi dan industrialisasi, pemantauan dan prediksi kualitas udara menjadi semakin kritis untuk mendukung pengambilan keputusan kebijakan lingkungan yang tepat sasaran. Organisasi Kesehatan Dunia (WHO) bahkan menempatkan polusi udara sebagai salah satu ancaman lingkungan terbesar bagi kesehatan global, dengan jutaan kematian dini setiap tahun yang berkaitan dengan paparan polutan di udara terbuka maupun dalam ruangan [1]. Kondisi ini menegaskan perlunya

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 763−773
DOI 10.47065/bits.v8i2.10602
https://ejurnal.seminar-id.com/index.php/bits

sistem pemantauan yang tidak hanya akurat, tetapi juga mampu memberikan klasifikasi tingkat bahaya secara cepat agar kebijakan mitigasi dapat diambil sebelum dampak kesehatan meluas.

Berbagai polutan seperti particulate matter (PM2.5, PM10), nitrogen dioksida (NO2), sulfur dioksida (SO2), dan karbon monoksida (CO) berinteraksi dengan kondisi lingkungan dan demografis membentuk pola kompleks yang sulit dianalisis dengan metode statistik konvensional [2]. Pendekatan *machine learning* (ML) menawarkan solusi adaptif dan akurat dalam mengklasifikasikan tingkat kualitas udara secara otomatis berdasarkan pola-pola tersebut [3], [4]. Tantangan ini muncul karena data lingkungan bersifat multidimensi dan memiliki hubungan non-linear antar polutan yang sulit diasumsikan oleh metode statistik konvensional, sehingga machine learning yang mampu mempelajari pola secara adaptif dari data menjadi solusi yang lebih tepat.

Berbagai pendekatan algoritma telah diuji untuk tugas klasifikasi kualitas udara, mulai dari model pohon keputusan tunggal hingga metode ensemble dan deep learning yang lebih kompleks [5]. Metode ensemble seperti Random Forest dan XGBoost menggabungkan banyak model dasar untuk menghasilkan prediksi yang lebih stabil dan tahan terhadap *overfitting* dibandingkan model tunggal [6], [7], sehingga banyak dipilih pada studi-studi terbaru di bidang lingkungan. Namun demikian, sebagian besar studi tersebut masih berfokus pada peningkatan akurasi prediksi secara umum dan belum secara eksplisit menangani ketidakseimbangan distribusi kelas pada data kualitas udara, padahal kelas dengan tingkat bahaya tertinggi pada umumnya justru merupakan kelas minoritas dalam dataset.

Penelitian sebelumnya [8] menggunakan algoritma Decision Tree untuk mengklasifikasikan kualitas udara ke dalam empat kategori: Good, Moderate, Poor, dan Hazardous menggunakan dataset Air Quality and Pollution Assessment dari Kaggle yang terdiri dari 5.000 sampel dengan 9 fitur. Penelitian tersebut berhasil mencapai akurasi sebesar 93% dengan fitur paling berpengaruh adalah PM2.5. Namun, penelitian tersebut memiliki beberapa keterbatasan: hanya menggunakan satu algoritma tanpa perbandingan, tidak menerapkan *cross-validation*, dan yang paling krusial adalah tidak menangani ketidakseimbangan kelas (*class imbalance*) pada dataset. Keterbatasan-keterbatasan tersebut berimplikasi pada keandalan model ketika diterapkan pada data baru, karena evaluasi tanpa *cross-validation* cenderung memberikan estimasi performa yang kurang stabil, sementara ketiadaan penanganan *class imbalance* berisiko menyebabkan model bias terhadap kelas mayoritas sehingga kurang sensitif dalam mendeteksi kelas Hazardous yang justru paling penting untuk diwaspadai.

Ketidakseimbangan kelas merupakan permasalahan umum dalam dataset kualitas udara, di mana kelas Good mendominasi distribusi data (40%) sementara kelas Hazardous hanya 10%. Kondisi ini menyebabkan model cenderung bias terhadap kelas mayoritas sehingga performa prediksi pada kelas minoritas yang justru paling kritis menjadi rendah [9], [10]. Teknik Synthetic Minority *Oversampling* Technique (SMOTE) telah terbukti efektif mengatasi masalah ini dengan membangkitkan sampel sintetik pada kelas minoritas [11]. Permasalahan ini telah menjadi perhatian khusus dalam literatur pembelajaran mesin, sebagaimana dibahas secara komprehensif dalam berbagai kajian tentang *imbalanced learning* yang merangkum strategi penanganan mulai dari pendekatan berbasis data seperti *oversampling* dan *undersampling* hingga pendekatan berbasis algoritma seperti *cost-sensitive learning* [12]. SMOTE dipilih dibandingkan random oversampling karena membangkitkan sampel sintetik melalui interpolasi antar tetangga terdekat sehingga lebih variatif dan tidak sekadar menduplikasi data mentah, serta lebih dipilih dibandingkan undersampling karena tidak menghilangkan informasi dari kelas mayoritas yang tetap penting bagi generalisasi model [13].

Pemilihan Random Forest dan XGBoost didasarkan pada temuan penelitian sebelumnya yang menunjukkan bahwa algoritma ensemble secara konsisten mengungguli algoritma tunggal seperti Decision Tree dalam pemodelan kualitas udara [2], [3], [14]. Random Forest dipilih karena ketahanannya terhadap *overfitting* melalui mekanisme *bagging*, sedangkan XGBoost dipilih karena kemampuan *boosting*-nya yang sensitif terhadap pola *residual* dan efektif menangani data kompleks [6], [7]. Kedua karakteristik ini relevan untuk mengatasi keterbatasan penelitian sebelumnya yang tidak menangani *class imbalance* dan hanya mengandalkan satu algoritma. Penelitian ini mengusulkan pendekatan yang lebih komprehensif dengan mengimplementasikan dua algoritma *ensemble learning*, yaitu Random Forest dan XGBoost, dikombinasikan dengan SMOTE. Kontribusi utama penelitian ini meliputi: (1) perbandingan performa Random Forest dan XGBoost dengan dan tanpa SMOTE, (2) analisis pengaruh SMOTE terhadap kelas minoritas Hazardous, dan (3) identifikasi fitur-fitur paling berpengaruh dalam prediksi kualitas udara serta perbandingannya dengan penelitian sebelumnya. Pendekatan ini diharapkan tidak hanya meningkatkan akurasi prediksi secara teknis, tetapi juga memberikan dasar yang lebih andal bagi pengambil kebijakan dalam mendeteksi kondisi udara berbahaya secara dini, mengingat kesalahan klasifikasi pada kelas minoritas Hazardous berpotensi menimbulkan risiko kesehatan yang signifikan bagi masyarakat. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi baik dari sisi akademis melalui perbandingan metodologis yang lebih lengkap, maupun dari sisi praktis melalui model yang lebih siap diadopsi dalam sistem pemantauan lingkungan berbasis data.

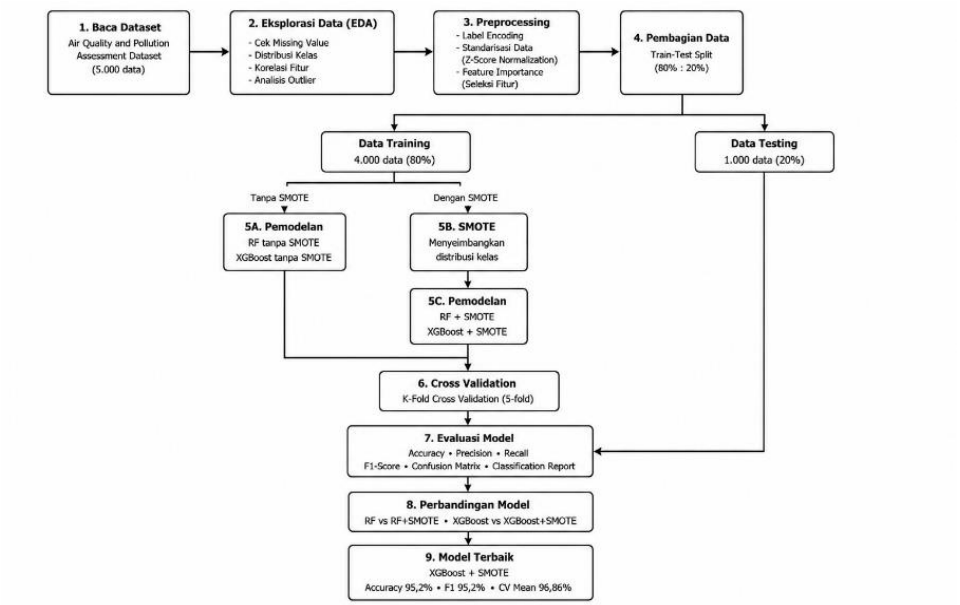
2. METODOLOGI PENELITIAN

2.1 Tahapan penelitian

Penelitian ini mengadopsi pendekatan klasifikasi *supervised learning* untuk mengklasifikasikan tingkat kualitas udara. Gambar 1 menampilkan alur kerja penelitian secara keseluruhan, mulai dari pembacaan dataset hingga diperoleh model dengan performa terbaik.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 763−773
DOI 10.47065/bits.v8i2.10602
https://ejurnal.seminar-id.com/index.php/bits

Berdasarkan Gambar 1, tahapan penelitian diawali dengan pembacaan dataset Air Quality and Pollution Assessment yang berisi 5.000 data (tahap 1). Tahap selanjutnya adalah eksplorasi data (*Exploratory Data Analysis*/EDA) (tahap 2) yang meliputi pengecekan *missing value*, analisis distribusi kelas, korelasi antar fitur, dan deteksi *outlier* untuk memahami karakteristik data secara menyeluruh. Data kemudian melalui tahap *preprocessing* (tahap 3), yang mencakup *label encoding*, standardisasi data menggunakan *Z-Score normalization*, serta seleksi fitur berdasarkan *feature importance*. Setelah *preprocessing*, data dibagi menggunakan skema *train-test split* dengan proporsi 80% data *training* (4.000 data) dan 20% data *testing* (1.000 data) (tahap 4). Pada data *training*, penelitian ini membangun dua jalur pemodelan secara paralel (tahap 5), yaitu pemodelan tanpa SMOTE menggunakan algoritma Random Forest dan XGBoost pada data asli (tahap 5A), serta pemodelan dengan SMOTE yang terlebih dahulu menyeimbangkan distribusi kelas (tahap 5B) sebelum dilatih dengan algoritma yang sama (tahap 5C). Kedua jalur pemodelan tersebut selanjutnya divalidasi menggunakan *K-Fold Cross Validation* dengan 5-fold untuk memastikan kestabilan performa model (tahap 6).



Gambar 1. Alur Kerja Membangun Model Klasifikasi Kualitas Udara

Tahap evaluasi model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, *F1-Score*, *confusion matrix*, dan *classification report* pada data *testing* (tahap 7). Hasil evaluasi keempat skenario tersebut kemudian dibandingkan, yaitu RF vs RF+SMOTE dan XGBoost vs XGBoost+SMOTE (tahap 8), untuk menentukan model dengan performa terbaik di antara keempat skenario tersebut (tahap 9).

2.2 Machine Learning untuk Klasifikasi Lingkungan

Machine learning (ML) merupakan metode komputasi yang memungkinkan sistem belajar dari data tanpa diprogram secara eksplisit [15], [16], dan telah terbukti mampu mengklasifikasikan kualitas udara dengan akurasi lebih baik dibandingkan metode statistik konvensional [2], [17] berkat kemampuannya menangani data multivariat berdimensi tinggi dan hubungan non-linear antar polutan [3]. Berbagai algoritma ML, mulai dari model pohon tunggal hingga metode ensemble, telah diterapkan pada dataset Air Quality and Pollution Assessment yang sama dengan penelitian ini [8], namun dengan pendekatan metodologis yang berbeda–beda. Penelitian ini memposisikan diri sebagai kelanjutan metodologis dari studi tersebut dengan menerapkan tiga penyesuaian utama pada rancangan eksperimen. Pertama, evaluasi dilakukan menggunakan 5-fold *stratified cross-validation*, bukan hanya *single train-test split*, untuk memperoleh estimasi performa yang lebih stabil dan representatif. Kedua, dua algoritma ensemble, yaitu Random Forest dan XGBoost, diuji secara paralel dan dibandingkan langsung, bukan mengandalkan satu algoritma tunggal, sehingga karakteristik masing–masing metode terhadap dataset yang sama dapat dianalisis secara komparatif. Ketiga, SMOTE diterapkan secara eksplisit dan konsisten hanya pada data *training* untuk menyeimbangkan distribusi kelas Hazardous, sejalan dengan rekomendasi literatur *imbalanced learning* bahwa penanganan ketidakseimbangan kelas krusial untuk membangun model yang andal pada kasus deteksi kondisi berbahaya [9], [10].

Ketiga penyesuaian metodologis ini secara langsung menjawab kesenjangan yang telah diuraikan pada bagian Pendahuluan. Penjelasan rinci mengenai masing–masing algoritma, teknik SMOTE secara keseluruhan disajikan berturut–turut pada Subbab 2.3 hingga 2.5.

2.3 Random Forest

Random Forest adalah metode ensemble berbasis *bagging* yang membangun sejumlah besar pohon keputusan secara acak dan paralel dari subset data dan fitur [6]. Prediksi akhir ditentukan melalui *voting* mayoritas dari seluruh pohon.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 763−773
DOI 10.47065/bits.v8i2.10602
https://ejurnal.seminar-id.com/index.php/bits

Pemilihan atribut terbaik pada setiap percabangan menggunakan pengukuran *Gini Impurity* atau Entropy. Nilai Entropy dihitung menggunakan Persamaan (1):

$$Entropy(Y) = -\sum p(c|Y) log_2 p(c|Y) \quad (1)$$

dengan Y adalah himpunan kasus, dan p(c|Y) adalah proporsi nilai Y terhadap kelas c. Pemilihan atribut terbaik menggunakan *Information Gain* yang diformulasikan pada Persamaan (2):

$$IG(Y, a) = Entropy(Y) - \sum(\frac{|Y_v|}{|Y_a|}) \times Entropy(Y_v) \quad (2)$$

dengan Y_v adalah subkelas dari Y yang berhubungan dengan kelas a, dan Y_a adalah semua nilai yang sesuai dengan atribut a. Keunggulan Random Forest mencakup *robustness* terhadap *overfitting*, kemampuan menangani *missing value*, dan kemudahan interpretasi melalui *feature importance*. Pada penelitian ini, Random Forest diimplementasikan dengan n_estimators=100 dan random_state=42. Nilai n_estimators=100 dipilih karena telah mampu menghasilkan estimasi yang stabil pada penelitian sejenis tanpa menambah beban komputasi yang signifikan, sedangkan random_state=42 ditetapkan sebagai nilai baku untuk menjamin reprodusibilitas hasil pada seluruh proses pelatihan model dalam penelitian ini, termasuk pada XGBoost [18].

2.4 XGBoost (Extreme Gradient Boosting)

XGBoost adalah implementasi *gradient boosting* yang dioptimalkan untuk performa tinggi dan efisiensi komputasi [7]. Berbeda dengan Random Forest yang menggunakan *bagging* (pohon paralel), XGBoost menggunakan pendekatan *boosting* di mana pohon dibangun secara sekuensial, setiap pohon baru fokus memperbaiki *residual* kesalahan pohon sebelumnya. Fungsi objektif yang diminimalkan adalah Persamaan (3):

$$Obj(\theta) = \sum l(y_i, \hat{y}_i) + \sum \Omega(f_k) \quad (3)$$

dengan l adalah fungsi *loss*, y_i adalah nilai aktual, $\hat{y}_i$ adalah nilai prediksi, dan $\Omega(f_k)$ adalah regularisasi untuk mencegah *overfitting*. Regularisasi dalam XGBoost didefinisikan pada Persamaan (4):

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum w_j^2 \quad (4)$$

dengan T adalah jumlah daun (*leaves*), w_j adalah skor daun ke-j, γ adalah parameter minimum gain, dan λ adalah koefisien L2 *regularization*.

Mekanisme *boosting* membuat XGBoost lebih sensitif terhadap pola kompleks dan *residual* yang terlewat oleh pohon sebelumnya. XGBoost juga dilengkapi regularisasi L1 dan L2 bawaan serta penanganan *missing value* secara otomatis [7]. Pada penelitian ini, XGBoost diimplementasikan dengan n_estimators=100, random_state=42, dan eval_metric='mlogloss'.

2.5 SMOTE (Synthetic Minority Oversampling Technique)

SMOTE adalah teknik *oversampling* yang membangkitkan sampel sintetik baru pada kelas minoritas melalui interpolasi antara sampel yang ada, bukan sekadar menduplikasi data [11]. Untuk setiap sampel minoritas x, SMOTE memilih *k-nearest neighbor* kemudian membangkitkan sampel baru menggunakan Persamaan (5):

$$x_{new} = x + \lambda \times (\tilde{x} - x) \quad (5)$$

dengan x_new adalah sampel sintetik baru, x adalah sampel minoritas asal, x̃ adalah tetangga terdekat yang dipilih secara acak, dan λ adalah bilangan acak dalam rentang [0, 1]. Metode ini membantu mengurangi ketidakseimbangan kelas sehingga model machine learning dapat mempelajari pola pada kelas minoritas dengan lebih baik.

SMOTE diterapkan hanya pada data *training* untuk menghindari *data leakage*. Penelitian sebelumnya menunjukkan kombinasi SMOTE dengan *ensemble learning* secara konsisten meningkatkan performa pada kelas minoritas tanpa mengorbankan performa kelas mayoritas [19], [10], [20], [21]. SMOTE tetap menjadi metode yang paling banyak digunakan karena kesederhanaan dan efektivitasnya [22].

2.6 Dataset

Dataset yang digunakan adalah Air Quality and Pollution Assessment Dataset yang tersedia di platform Kaggle [23]. Dataset ini terdiri dari 5.000 sampel dengan 9 fitur masukan dan 1 variabel target. Hasil eksplorasi menunjukkan tidak terdapat *missing value* pada seluruh atribut. Tabel 1 mendeskripsikan fitur-fitur yang digunakan.

Pemilihan kesembilan fitur pada dataset ini didasarkan pada relevansinya terhadap faktor-faktor yang secara umum memengaruhi tingkat pencemaran udara, yaitu kombinasi antara variabel meteorologis (*temperature* dan *humidity*), konsentrasi polutan utama (PM2.5, PM10, NO2, SO2, dan CO), serta variabel demografis-spasial (*proximity to industrial areas* dan *population density*). Kombinasi fitur lingkungan dan demografis ini memungkinkan model *machine learning* mempelajari pola kualitas udara secara lebih komprehensif dibandingkan hanya mengandalkan data konsentrasi polutan semata, karena faktor jarak terhadap kawasan industri dan kepadatan penduduk turut berkontribusi terhadap tingkat pencemaran suatu wilayah. Selain itu, sumber dataset yang berasal dari

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 763−773
DOI 10.47065/bits.v8i2.10602
https://ejurnal.seminar-id.com/index.php/bits

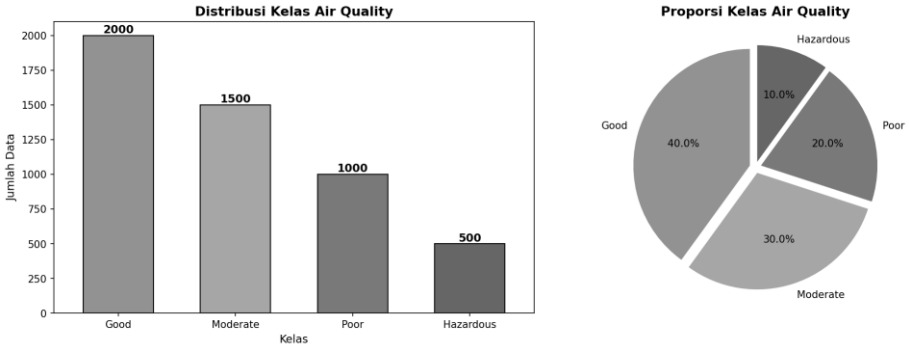
platform Kaggle dengan dokumentasi yang jelas turut mendukung reprodusibilitas penelitian ini bagi studi-studi selanjutnya.

Tabel 1. Deskripsi Fitur Dataset

Fitur	Deskripsi	Tipe
Temperature (°C)	Suhu rata-rata wilayah	Numerik
Humidity (%)	Kelembapan relatif wilayah	Numerik
PM2.5 (µg/m³)	Konsentrasi partikulat halus	Numerik
PM10 (µg/m³)	Konsentrasi partikulat kasar	Numerik
NO2 (ppb)	Konsentrasi nitrogen dioksida	Numerik
SO2 (ppb)	Konsentrasi sulfur dioksida	Numerik
CO (ppm)	Konsentrasi karbon monoksida	Numerik
Proximity to Industrial Areas (km)	Jarak ke kawasan industri terdekat	Numerik
Population Density (people/km²)	Kepadatan penduduk per km²	Numerik
Air Quality Levels	Variabel target (Good/Moderate/Poor/Hazardous)	Kategorik

Tabel 1 memperlihatkan bahwa dataset terdiri dari sembilan fitur numerik yang merepresentasikan kondisi lingkungan (suhu, kelembapan, konsentrasi polutan PM2.5, PM10, NO2, SO2, CO) dan kondisi demografis-spasial (jarak ke kawasan industri dan kepadatan penduduk), dengan Air Quality Levels sebagai satu variabel target kategorik berisi empat kelas. Karena seluruh fitur masukan bertipe numerik namun memiliki satuan dan rentang nilai yang berbeda-beda (misalnya CO dalam ppm dan Population Density dalam ribuan orang/km²), standardisasi data pada tahap *preprocessing* menjadi penting agar setiap fitur berkontribusi secara proporsional saat pelatihan model.

Distribusi kelas target menunjukkan ketidakseimbangan signifikan: Good 2.000 (40%), Moderate 1.500 (30%), Poor 1.000 (20%), dan Hazardous 500 (10%). Gambar 2 menampilkan visualisasi distribusi tersebut.



Gambar 2. Distribusi Kelas Target Air Quality

Berdasarkan Gambar 2, kelas Good mendominasi dataset dengan 2.000 sampel (40%), diikuti Moderate sebanyak 1.500 sampel (30%) dan Poor sebanyak 1.000 sampel (20%), sedangkan kelas Hazardous yang paling kritis bagi kesehatan hanya berjumlah 500 sampel (10%). Rasio ketidakseimbangan antara kelas mayoritas Good dan kelas minoritas Hazardous mencapai 4:1, kondisi yang berisiko membuat model bias dalam memprediksi kelas mayoritas dan kurang sensitif dalam mengenali kelas Hazardous. Ketidakseimbangan inilah yang menjadi dasar penerapan SMOTE pada tahap pemodelan selanjutnya.

2.7 Preprocessing

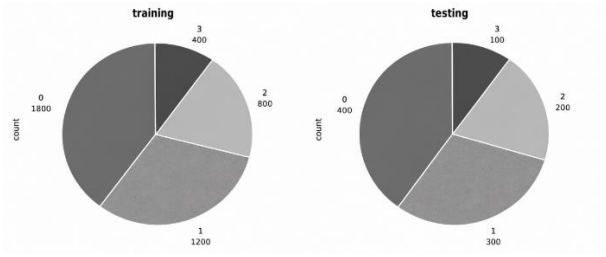
Tahap *preprocessing* meliputi tiga langkah. Pertama, *label encoding*: variabel target dikodekan menjadi nilai numerik menggunakan LabelEncoder scikit-learn, yaitu Good=0, Hazardous=1, Moderate=2, Poor=3 (urutan alfabetis). Kedua, normalisasi *Z-Score* yang diterapkan pada seluruh fitur numerik menggunakan StandardScaler dengan formula pada Persamaan (6):

$$z = \frac{(x-\mu)}{\sigma} \tag{6}$$

dengan z adalah nilai ternormalisasi, x adalah nilai asli, µ adalah rata-rata fitur, dan σ adalah standar deviasi fitur. Ketiga, pembagian data: dataset dibagi 80% *training* (4.000 sampel) dan 20% *testing* (1.000 sampel) menggunakan *stratified sampling*. Gambar 3 menunjukkan distribusi kelas pada data *training* dan *testing*.

Pendekatan pembagian data dengan proporsi 80:20 ini dipilih karena telah menjadi praktik umum dalam penelitian *machine learning* untuk menyediakan porsi data *training* yang cukup besar bagi model dalam mempelajari pola, sekaligus menyisakan data *testing* yang memadai untuk evaluasi performa secara objektif. Penggunaan *stratified sampling* pada tahap ini memastikan proporsi setiap kelas tetap terjaga pada kedua subset, sehingga evaluasi model pada tahap selanjutnya mencerminkan kondisi distribusi data yang sebenarnya sebelum SMOTE diterapkan pada data *training*.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 763−773
DOI 10.47065/bits.v8i2.10602
https://ejurnal.seminar-id.com/index.php/bits



Gambar 3. Distribusi Kelas pada Data Training dan Testing (80:20)

Gambar 3 memperlihatkan bahwa proporsi keempat kelas pada data training dan data testing tetap konsisten mengikuti distribusi 80:20, sebagai hasil dari penerapan stratified sampling pada tahap pembagian data, sehingga karakteristik ketidakseimbangan kelas yang sama tetap terjaga pada kedua subset sebelum SMOTE diterapkan.

2.8 Penerapan SMOTE

SMOTE diterapkan hanya pada data *training* setelah proses *split* untuk menghindari *data leakage*. Tabel 2 menampilkan perbandingan distribusi kelas sebelum dan sesudah SMOTE.

Tabel 2. Distribusi Kelas Sebelum dan Sesudah SMOTE pada Data Training

Kelas	Sebelum SMOTE	Sesudah SMOTE	Selisih	Kenaikan (%)
Good	1.600	1.600	0	0%
Moderate	1.200	1.600	+400	33,3%
Poor	800	1.600	+800	100%
Hazardous	400	1.600	+1.200	300%
Total	4.000	6.400	+2.400	60%

Berdasarkan Tabel 2 menunjukkan bahwa SMOTE menyeimbangkan seluruh kelas minoritas ke jumlah yang sama dengan kelas mayoritas Good, yaitu 1.600 sampel per kelas, sehingga total data training meningkat dari 4.000 menjadi 6.400 sampel (+60%). Kelas Hazardous mengalami peningkatan terbesar sebesar 300%, dari 400 menjadi 1.600 sampel, diikuti kelas Poor sebesar 100% (dari 800 menjadi 1.600 sampel) dan kelas Moderate sebesar 33,3% (dari 1.200 menjadi 1.600 sampel), sedangkan kelas Good tidak berubah karena telah menjadi acuan jumlah sampel maksimum, sebagaimana dirujuk pada Tabel 2.

2.9 Evaluasi Model

Evaluasi menggunakan metrik *Precision*, *Recall*, *F1-Score*, *Accuracy*, dan 5-fold *Stratified Cross-Validation*. Formula metrik evaluasi ditunjukkan pada Persamaan (7) hingga (9):

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

dengan TP adalah *True Positive*, FP adalah *False Positive*, dan FN adalah *False Negative*. Keempat skenario yang dievaluasi adalah: Random Forest tanpa SMOTE, Random Forest dengan SMOTE, XGBoost tanpa SMOTE, dan XGBoost dengan SMOTE.

3. HASIL DAN PEMBAHASAN

3.1 Feature Importance

Analisis *feature importance* dilakukan untuk mengidentifikasi faktor paling berpengaruh dalam klasifikasi kualitas udara pada masing-masing model. Tabel 3 menyajikan peringkat lima fitur dengan importance tertinggi pada Random Forest dan XGBoost.

Tabel 3. Peringkat Feature Importance Top 5 per Model

Rank	Fitur (RF)	Importance	Fitur (XGBoost)	Importance
1	CO	0,344	CO	0,595
2	Proximity to Industrial Areas	0,283	Proximity to Industrial Areas	0,219
3	NO2	0,096	NO2	0,047
4	SO2	0,094	Temperature	0,031
5	Temperature	0,070	SO2	0,028

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 763−773
DOI 10.47065/bits.v8i2.10602
https://ejurnal.seminar-id.com/index.php/bits

Berdasarkan Tabel 3, terdapat perbedaan menarik antara penelitian ini dengan penelitian sebelumnya [8]. Penelitian sebelumnya mengidentifikasi PM2.5 sebagai fitur paling berpengaruh, sedangkan penelitian ini menemukan CO sebagai fitur dominan pada kedua model. Perbedaan ini dapat dijelaskan dari dua aspek. Pertama, penggunaan algoritma yang berbeda: Decision Tree menggunakan *information gain* dalam satu level percabangan, sehingga fitur dengan variabilitas tinggi seperti PM2.5 cenderung terpilih di *node* awal. Sebaliknya, Random Forest dan XGBoost mengevaluasi kontribusi fitur secara agregat dari ratusan pohon keputusan, sehingga fitur seperti CO yang memiliki pola diskriminatif konsisten lintas semua kelas lebih teridentifikasi sebagai fitur kritis. Kedua, CO memiliki korelasi yang kuat dan konsisten dengan seluruh kategori kualitas udara dari Good hingga Hazardous karena CO merupakan produk langsung pembakaran tidak sempurna dari kendaraan dan industri yang berpengaruh pada seluruh spektrum kualitas udara [17], [5]. PM2.5 memang berbahaya, namun kontribusinya dalam dataset ini lebih bersifat kolinear dengan CO dan Proximity to Industrial Areas, sehingga bobot pentingnya terdistribusi ke fitur lain saat dievaluasi oleh ensemble model.

3.2 Perbandingan Performa Model

Tabel 4 menampilkan perbandingan komprehensif performa keempat model pada data *testing* (1.000 sampel) beserta hasil *cross-validation*.

Tabel 4. Tabel Perbandingan Performa Semua Model

Model	Accuracy	Precision	Recall	F1-Score	CV Mean	CV Std
RF (Tanpa SMOTE)	0,951	0,951	0,951	0,951	0,953	0,007
RF (Dengan SMOTE)	0,948	0,948	0,948	0,948	0,964	0,004
XGBoost (Tanpa SMOTE)	0,951	0,951	0,951	0,951	0,955	0,004
XGBoost (Dengan SMOTE)*	0,952	0,952	0,952	0,952	0,969	0,002
Decision Tree [8]	0,930	-	-	0,930	-	-

**Model terbaik. Baris terakhir merupakan hasil penelitian sebelumnya [8] sebagai pembanding.*

Berdasarkan Tabel 4 menunjukkan bahwa keempat skenario model yang diusulkan dalam penelitian ini memiliki akurasi yang lebih tinggi dibandingkan penelitian sebelumnya yang menggunakan Decision Tree [8], yaitu sebesar 93%. XGBoost dengan SMOTE secara konsisten unggul di semua metrik evaluasi.

Hasil penelitian menunjukkan SMOTE memberikan dampak yang berbeda pada kedua algoritma. Pada Random Forest, penerapan SMOTE justru sedikit menurunkan *F1-Score* sebesar 0,003 (dari 0,951 menjadi 0,948). Hal ini terjadi karena Random Forest memiliki mekanisme *bootstrap sampling* internal yang secara alami memberikan eksposur berulang pada sampel minoritas selama proses *bagging* [6]. Penambahan sampel sintetik SMOTE pada Random Forest berpotensi menambah *noise* dan redundansi yang mengganggu proses pemilihan fitur acak (*feature randomness*) pada setiap pohon, sehingga sedikit menurunkan performa keseluruhan. Meski demikian, CV Mean Random Forest justru meningkat dari 0,953 menjadi 0,964 setelah SMOTE, menunjukkan generalisasi yang lebih baik.

Perbedaan arah perubahan antara *F1-Score* dan CV Mean ini dapat dipahami karena keduanya mengukur hal yang berbeda. Perlu dicatat bahwa SMOTE hanya diterapkan pada data *training*, bukan pada data *testing*. Pada evaluasi *test set*, model yang telah dilatih dengan data ber-SMOTE diuji langsung pada data asli yang tidak mengandung sampel sintetik, akibatnya, sedikit *noise* yang dihasilkan dari sampel sintetik selama pelatihan memengaruhi performa pada distribusi data asli tersebut. Sebaliknya, pada proses *cross-validation*, SMOTE diterapkan ulang secara independen di setiap fold hanya pada bagian *training*-nya, sehingga proses ini lebih terkontrol dan tidak mencemari data validasi. Hal ini membuat distribusi kelas pada setiap iterasi lebih seimbang dan menghasilkan generalisasi yang lebih konsisten. Dengan demikian, peningkatan CV Mean ini mengindikasikan bahwa SMOTE tetap berkontribusi positif terhadap stabilitas generalisasi Random Forest secara keseluruhan, meskipun dampaknya pada data *testing* bersih sedikit kurang optimal.

Sebaliknya, XGBoost dengan SMOTE menunjukkan peningkatan *F1-Score* sebesar 0,001 (dari 0,951 menjadi 0,952). XGBoost menggunakan optimasi *gradient* yang lebih sensitif terhadap distribusi kelas karena setiap iterasi *boosting* menghitung *gradient loss* secara langsung dari distribusi data. Sampel SMOTE memberikan sinyal *gradient* yang lebih seimbang, membantu model mengoptimalkan *loss function* pada kelas minoritas secara lebih efektif. Selain itu, CV Mean XGBoost meningkat signifikan dari 0,955 menjadi 0,969 dengan standar deviasi terendah (0,002), mengindikasikan model paling stabil dan mampu generalisasi terbaik.

3.3 Classification Report per Kelas

Tabel 5 menampilkan *classification report* lengkap untuk model terbaik (XGBoost dengan SMOTE) dibandingkan dengan Random Forest tanpa SMOTE.

Tabel 5. Classification Report per Kelas

Kelas	Precision RF	Recall RF	F1 RF	Precision XGB+S	Recall XGB+S	F1 XGB+S
Good	1,00	1,00	1,00	1,00	1,00	1,00
Moderate	0,96	0,97	0,97	0,97	0,96	0,96

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 763−773
DOI 10.47065/bits.v8i2.10602
https://ejurnal.seminar-id.com/index.php/bits

Kelas	Precision RF	Recall RF	F1 RF	Precision XGB+S	Recall XGB+S	F1 XGB+S
Poor	0,87	0,90	0,88	0,88	0,89	0,88
Hazardous	0,90	0,80	0,85	0,87	0,87	0,87
Accuracy			0,951			0,952

RF = Random Forest (tanpa SMOTE); XGB+S = XGBoost dengan SMOTE.

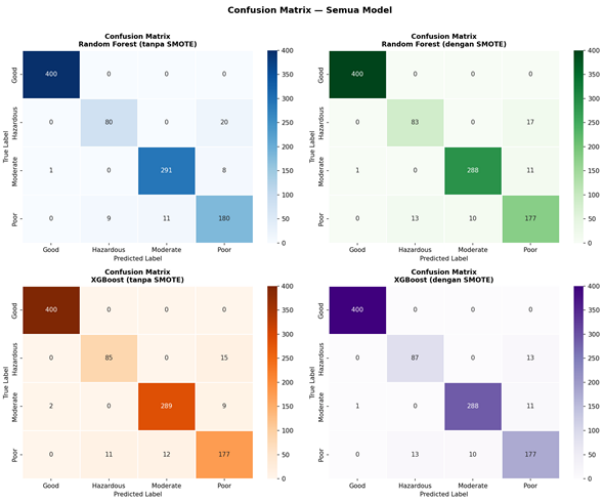
Berdasarkan Tabel 5, kelas Good memperoleh *precision*, *recall*, dan *F1-Score* sempurna (1,00) pada kedua model, sedangkan kelas Moderate dan Poor menunjukkan performa yang relatif setara antara RF dan XGB+S, dengan selisih *F1-Score* tidak lebih dari 0,01. Perbedaan paling signifikan terlihat pada kelas Hazardous, yaitu kelas minoritas dengan proporsi data asli paling kecil (10%). Pada RF tanpa SMOTE, kelas Hazardous memperoleh *precision* 0,90 namun *recall* hanya 0,80, sehingga *F1-Score*-nya 0,85. *Gap* antara *precision* dan *recall* pada RF ini menunjukkan bahwa model cenderung bersikap konservatif dalam memprediksi kelas Hazardous: ketika model memprediksi suatu sampel sebagai Hazardous, prediksi tersebut sebagian besar benar (*precision* tinggi), tetapi banyak sampel Hazardous yang sebenarnya justru tidak terdeteksi dan salah diklasifikasikan ke kelas lain (*recall* rendah). Kondisi ini merupakan dampak langsung dari ketidakseimbangan kelas pada data *training* RF tanpa SMOTE, di mana sampel Hazardous yang minim menyebabkan model kurang terekspos pada pola kelas tersebut.

Setelah penerapan SMOTE pada XGBoost, kelas Hazardous memperoleh *precision* 0,87 dan *recall* 0,87, sehingga *F1-Score*-nya naik menjadi 0,87. Dibandingkan RF, *recall* Hazardous pada XGB+S meningkat dari 0,80 menjadi 0,87, sementara *precision*-nya sedikit menurun dari 0,90 menjadi 0,87. Pergeseran ini menunjukkan bahwa SMOTE menggeser keseimbangan model dari yang sebelumnya konservatif (*precision* tinggi, *recall* rendah) menjadi lebih seimbang antara *precision* dan *recall*, sehingga *F1-Score* Hazardous naik dari 0,85 menjadi 0,87. Peningkatan *recall* ini penting secara praktis karena kelas Hazardous merepresentasikan kondisi udara paling berbahaya bagi kesehatan, sehingga model yang mampu mengenali lebih banyak kasus Hazardous (*recall* lebih tinggi) lebih bernilai meskipun harus berbagi sedikit penurunan pada *precision*-nya. Hal ini sejalan dengan tujuan utama penerapan SMOTE, yaitu memperbaiki kemampuan model dalam mendeteksi kelas minoritas yang justru paling kritis untuk dikenali dengan akurat [9], [10].

Perbandingan dengan penelitian sebelumnya [8] memperkuat temuan ini. Penelitian sebelumnya sebenarnya telah melaporkan *classification report* per kelas, dengan *precision*, *recall*, dan *F1-Score* kelas Hazardous masing-masing sebesar 0,82. Menariknya, *recall* Hazardous pada RF tanpa SMOTE dalam penelitian ini (0,80) justru sedikit lebih rendah dibandingkan *recall* Decision Tree pada penelitian sebelumnya (0,82), menunjukkan bahwa tanpa penanganan *class imbalance*, model Decision Tree pada penelitian sebelumnya tetap mampu mengenali sebagian besar sampel Hazardous dengan cukup baik meskipun datasetnya juga tidak seimbang. Setelah penerapan SMOTE pada XGBoost, *recall* Hazardous meningkat menjadi 0,87, melampaui baik RF tanpa SMOTE (0,80) maupun Decision Tree pada penelitian sebelumnya (0,82). Hal ini menunjukkan bahwa kombinasi *ensemble learning* dan SMOTE memberikan perbaikan nyata dalam mendeteksi kelas minoritas dibandingkan model tunggal tanpa penanganan *class imbalance*. Tabel 5 dalam penelitian ini menunjukkan secara rinci bagaimana RF tanpa SMOTE masih mengalami *gap precision-recall* pada kelas Hazardous, dan bagaimana penerapan SMOTE pada XGBoost berhasil menutup *gap* tersebut sehingga *recall* dan *precision* menjadi seimbang pada 0,87. Hal ini menjadi nilai tambah penelitian ini dibandingkan penelitian sebelumnya, karena evaluasi per kelas memberikan gambaran yang lebih lengkap mengenai kemampuan model dalam menangani kelas minoritas yang secara praktis paling penting untuk dikenali dengan tepat.

3.4 Confusion Matrix

Gambar 4 menampilkan *confusion matrix* keempat model. Visualisasi ini memperlihatkan pola kesalahan klasifikasi antar kelas pada setiap skenario.



Gambar 4. Confusion Matrix Keempat Model

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 763−773
DOI 10.47065/bits.v8i2.10602
https://ejurnal.seminar-id.com/index.php/bits

Berdasarkan Gambar 4 terlihat bahwa kelas Hazardous dan Poor masih sering tertukar pada seluruh model. Hal ini wajar karena kedua kelas berbagi karakteristik nilai CO dan Proximity to Industrial Areas yang berdekatan, sebagaimana ditunjukkan pada analisis *feature importance* sebelumnya.

Pada keempat *confusion matrix*, kelas Good selalu terklasifikasi sempurna (400 dari 400 sampel), sehingga seluruh kesalahan prediksi terkonsentrasi pada tiga kelas lainnya, terutama pada pasangan Hazardous-Poor. Pada RF tanpa SMOTE, dari 100 sampel Hazardous, hanya 80 yang terklasifikasi benar, sedangkan 20 sampel salah diklasifikasikan sebagai Poor; tidak ada satu pun yang tertukar dengan Good atau Moderate. Pola serupa terlihat pada kelas Poor, di mana 11 dari 200 sampel salah diklasifikasikan sebagai Moderate dan 9 sampel sebagai Hazardous. Pola kesalahan yang terkonsentrasi hanya pada pasangan Hazardous-Poor ini konsisten dengan temuan *feature importance*, di mana kedua kelas tersebut memiliki rentang nilai CO dan Proximity to Industrial Areas yang saling berdekatan, sehingga lebih mudah tertukar dibandingkan dengan kelas Good atau Moderate yang karakteristiknya lebih berbeda.

Setelah penerapan SMOTE, jumlah sampel Hazardous yang berhasil terklasifikasi benar meningkat pada kedua algoritma. Pada Random Forest, jumlah prediksi benar untuk kelas Hazardous naik dari 80 (tanpa SMOTE) menjadi 83 (dengan SMOTE), sementara pada XGBoost naik dari 85 (tanpa SMOTE) menjadi 87 (dengan SMOTE), dengan sampel yang salah klasifikasi terus berkurang dan seluruhnya teralihkan ke kelas Poor. Peningkatan ini selaras dengan kenaikan *recall* kelas Hazardous pada Tabel 5, yaitu dari 0,80 (RF tanpa SMOTE) menjadi 0,87 (XGBoost dengan SMOTE), yang menunjukkan model semakin mampu mendeteksi kondisi udara berbahaya yang sebelumnya terlewat. Meskipun kebocoran antara Hazardous dan Poor tidak hilang sepenuhnya pada model manapun, *confusion matrix* XGBoost dengan SMOTE menunjukkan jumlah kesalahan klasifikasi paling sedikit secara keseluruhan, sejalan dengan posisinya sebagai model dengan performa terbaik pada Tabel 4 dan Tabel 5. Secara praktis, performa XGBoost dengan SMOTE yang konsisten unggul pada kelas Hazardous menunjukkan potensi penerapan model ini sebagai komponen pendukung sistem peringatan dini kualitas udara, khususnya pada wilayah dengan tingkat pencemaran tinggi. Kemampuan model dalam mengenali kondisi udara berbahaya dengan *recall* yang lebih baik dibandingkan Random Forest tanpa SMOTE maupun Decision Tree pada penelitian sebelumnya [8] menjadikannya kandidat yang relevan untuk diintegrasikan ke dalam sistem pemantauan operasional secara bertahap dan terukur, meskipun validasi lebih lanjut pada data lapangan dan periode waktu yang berbeda tetap diperlukan sebelum implementasi skala luas dilakukan.

Meskipun demikian, kebocoran yang tersisa antara Hazardous dan Poor pada seluruh model memiliki implikasi praktis yang perlu diperhatikan. Berdasarkan Gambar 4, kesalahan klasifikasi kelas Hazardous pada seluruh skenario selalu teralihkan ke kelas Poor, tidak pernah ke kelas Good maupun Moderate yang secara konseptual jauh lebih aman. Karakteristik ini menunjukkan bahwa meskipun model belum sepenuhnya bebas dari kesalahan, kesalahan yang terjadi bersifat konservatif karena sampel udara berbahaya tidak pernah diklasifikasikan sebagai udara bersih, melainkan hanya bergeser ke kategori tercemar berat yang berdekatan. Dari sudut pandang penerapan sistem peringatan dini, pola ini relatif lebih dapat ditoleransi dibandingkan skenario di mana kelas Hazardous justru tertukar dengan kelas Good, karena kesalahan semacam itu berpotensi menimbulkan risiko kesehatan yang jauh lebih besar akibat kegagalan total dalam mendeteksi kondisi udara berbahaya. Pola serupa juga konsisten terlihat pada kelas Poor, di mana kesalahan klasifikasi hanya mengarah ke Hazardous dan Moderate yang berdekatan, tanpa pernah tertukar dengan kelas Good. Konsistensi pola kesalahan yang selalu terjadi antar kelas yang berdekatan ini memperkuat temuan pada analisis *feature importance*, sekaligus menunjukkan bahwa batas keputusan (*decision boundary*) yang dibentuk baik oleh Random Forest maupun XGBoost telah merepresentasikan urutan tingkat keparahan kualitas udara (Good–Moderate–Poor–Hazardous) secara logis dan konsisten. Dengan mempertimbangkan pola *confusion matrix* pada Gambar 4 secara keseluruhan, XGBoost dengan SMOTE tetap menjadi pilihan paling rasional untuk diadopsi pada sistem pemantauan kualitas udara skala nyata, karena selain menghasilkan jumlah kesalahan klasifikasi paling sedikit secara keseluruhan, model ini juga mempertahankan pola kesalahan yang konservatif dan tidak pernah mengklasifikasikan kondisi udara berbahaya sebagai kondisi aman.

3.5 Pembahasan

Secara keseluruhan, penelitian ini menunjukkan peningkatan performa yang konsisten dibandingkan penelitian sebelumnya [8] yang menggunakan Decision Tree tunggal tanpa penanganan *class imbalance* maupun *cross-validation*. Model terbaik pada penelitian ini, XGBoost dengan SMOTE, mencapai akurasi 0,952, melampaui akurasi penelitian sebelumnya (0,930) dengan peningkatan 2,2%, sejalan dengan literatur yang menunjukkan keunggulan metode *ensemble* atas model pohon tunggal [6], [7]. Perbedaan *feature importance* juga teramati: penelitian sebelumnya mengidentifikasi PM2.5 sebagai fitur paling berpengaruh, sedangkan penelitian ini menemukan CO dan Proximity to Industrial Areas sebagai fitur dominan, sebagaimana telah dibahas pada Subbab 3.1.

Perbedaan paling substansial terletak pada penanganan ketidakseimbangan kelas. Penelitian sebelumnya [8] tidak menerapkan teknik apa pun untuk mengatasi *class imbalance*, dengan *recall* kelas Hazardous sebesar 0,82, sementara penelitian ini menunjukkan bahwa penerapan SMOTE pada XGBoost mampu meningkatkan *recall* Hazardous menjadi 0,87, melampaui capaian tersebut. Hasil ini konsisten dengan literatur *imbalanced learning* yang menunjukkan efektivitas kombinasi SMOTE dan *ensemble learning* dalam meningkatkan sensitivitas model terhadap kelas minoritas, sekaligus menegaskan bahwa penelitian ini berhasil menjawab keterbatasan utama penelitian sebelumnya dalam mendeteksi kelas yang paling kritis bagi kesehatan masyarakat.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 763−773
DOI 10.47065/bits.v8i2.10602
https://ejurnal.seminar-id.com/index.php/bits

Pergeseran fitur dominan dari PM2.5 menjadi CO dan *Proximity to Industrial Areas* juga memberikan implikasi praktis, karena CO memiliki pola diskriminatif yang lebih konsisten dibandingkan PM2.5 [17], [5]. Sistem pemantauan yang dikembangkan pun berpotensi menitikberatkan pengukuran CO dan jarak kawasan industri sebagai indikator dini risiko udara berbahaya, alih-alih hanya berfokus pada PM2.5 seperti penelitian sebelumnya [8].

Keunggulan lain penelitian ini terletak pada penggunaan 5-fold *stratified cross-validation*, yang tidak dilaporkan pada penelitian sebelumnya [8]. XGBoost dengan SMOTE memperoleh CV Mean tertinggi (0,969) dengan CV Std terendah (0,002), menunjukkan performa yang konsisten pada berbagai pembagian data.

4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan teknik SMOTE dalam klasifikasi kualitas udara menggunakan algoritma Random Forest dan XGBoost, dengan seluruh model melampaui akurasi penelitian sebelumnya yang menggunakan Decision Tree sebesar 93%. Model terbaik adalah XGBoost dengan SMOTE yang mencapai akurasi 0,952, *F1-Score* 0,952, dan CV Mean 0,969 dengan stabilitas tertinggi (CV Std 0,002). Penerapan SMOTE terbukti memberikan dampak yang berbeda pada kedua algoritma. Pada XGBoost, SMOTE meningkatkan *recall* kelas Hazardous dari 0,80 (RF tanpa SMOTE) menjadi 0,87 (XGBoost dengan SMOTE), sekaligus meningkatkan CV Mean secara signifikan dari 0,955 menjadi 0,969. Sebaliknya, pada Random Forest, SMOTE sedikit menurunkan *F1-Score* pada data *testing* (dari 0,951 menjadi 0,948), meskipun CV Mean-nya turut meningkat dari 0,953 menjadi 0,964, yang mengindikasikan generalisasi model yang lebih baik. Selain itu, CO dan Proximity to Industrial Areas teridentifikasi sebagai fitur paling dominan pada kedua model, berbeda dengan penelitian sebelumnya yang menemukan PM2.5 sebagai fitur utama. Penelitian selanjutnya disarankan untuk menerapkan *hyperparameter tuning* melalui *Grid Search* atau *Bayesian Optimization* guna meningkatkan performa model lebih lanjut. Selain itu, eksplorasi algoritma ensemble lain seperti LightGBM dan CatBoost serta implementasi metode *Explainable AI* seperti SHAP perlu dipertimbangkan untuk memperluas perbandingan dan meningkatkan transparansi model. Terakhir, kombinasi SMOTE dengan teknik *undersampling* seperti Tomek Links atau ENN diharapkan dapat menghasilkan distribusi data yang lebih representatif.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: N.P.K., dan H.S.;
Metodologi: N.P.K., dan H.S.;
Perangkat Lunak: N.P.K.;
Validasi: N.P.K., dan H.S.;
Analisis Formal: N.P.K., dan H.S.;
Investigasi: N.P.K., dan H.S.;
Sumber Daya: N.P.K.;
Kurasi Data: N.P.K.;
Penulisan Draf Awal: N.P.K., dan H.S.;
Penulisan, Peninjauan, dan Penyuntingan: N.P.K., dan H.S.;
Visualisasi: N.P.K.;
Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis korespondensi.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan eksternal

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

REFERENCES

[1] W. H. Organization, "WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide," Geneva, 2021. [Online]. Available: https://iris.who.int/handle/10665/345329

[2] K. Kumar and A. Pande, "Air pollution prediction with machine learning: a case study of Indian cities," *International Journal of Environmental Science and Technology*, vol. 20, pp. 5333–5348, 2023, doi: 10.1007/s13762-022-04241-5.

[3] M. Karmoude, "Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks," *Science of the Total Environment*, vol. 1002, p. 180593, 2025, doi: 10.1016/j.scitotenv.2025.180593.

[4] I. E. Agbehadji and I. C. Obagbuwa, "Systematic Review of Machine Learning and Deep Learning Techniques for Spatiotemporal Air Quality Prediction," *Atmosphere (Basel).*, vol. 15, no. 11, p. 1352, 2024, doi: 10.3390/atmos15111352.

[5] W. Peng, X. Liu, and F. Taghizadeh-Hesary, "Carbon Monoxide Emission and Air Quality Analysis Based on an Improved Double-Weight Fuzzy Comprehensive Evaluation," *Front. Public Health*, vol. 9, p. 790383, 2022, doi: 10.3389/fpubh.2021.790383.

[6] V. Ignatenko, A. Surkov, and S. Koltcov, "Random forests with parametric entropy-based information gains for classification and regression problems," *PeerJ Comput. Sci.*, vol. 10, p. e1775, 2024, doi: 10.7717/peerj-cs.1775.

[7] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, 2021, doi: 10.1007/s10462-020-09896-5.

[8] I. G. I. Sudipa, M. Habibi, E. S. J. Atmadji, and I. Arfiani, "Predictive Modeling of Air Quality Levels Using Decision Tree Classification: Insights from Environmental and Demographic Factors," *Indonesian Journal of Data and Science*, vol. 4, no. 3, pp. 251–258, 2024, doi: 10.56705/ijodas.v5i3.201.

[9] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," *Appl. Soft Comput.*, vol. 143, p. 110415, 2023, doi: 10.1016/j.asoc.2023.110415.

[10] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artif. Intell. Rev.*, vol. 57, no. 6, p. 137, 2024, doi: 10.1007/s10462-024-10759-6.

[11] X. Yi, Y. Xu, Q. Hu, S. Krishnamoorthy, W. Li, and Z. Tang, "ASN-SMOTE: a synthetic minority oversampling method with adaptive qualified synthesizer selection," *Complex & Intelligent Systems*, vol. 8, no. 3, pp. 2247–2272, 2022, doi: 10.1007/s40747-021-00638-w.

[12] I. Araf, A. Idri, and I. Chairi, "Cost-sensitive learning for imbalanced medical data: a review," *Artif. Intell. Rev.*, vol. 57, no. 4, p. 80, 2024, doi: 10.1007/s10462-023-10652-8.

[13] Y. Li, Y. Yang, P. Song, L. Duan, and R. Ren, "An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space," *Sci. Rep.*, vol. 15, p. 23521, 2025, doi: 10.1038/s41598-025-09506-w.

[14] A. F. B. Sajiwo, B. Rahmat, and A. Junaidi, "Klasifikasi Indeks Standar Pencemaran Udara (ISPU) Menggunakan Algoritma XGBoost dengan Teknik Imbalanced Data (SMOTE)," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, pp. 234–242, 2024, doi: 10.23960/jitet.v12i3.4699.

[15] B. Bala and S. Behal, "A Brief Survey of Data Preprocessing in Machine Learning and Deep Learning Techniques," in *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2024, pp. 1755–1762. doi: 10.1109/I-SMAC61858.2024.10714767.

[16] E. Barbierato and A. Gatti, "The Challenges of Machine Learning: A Critical Review," *Electronics (Basel).*, vol. 13, no. 2, p. 416, 2024, doi: 10.3390/electronics13020416.

[17] J. Zhu, Z. Zhang, W. Gu, C. Zhang, J. Xu, and P. Li, "Air Quality Prediction Using Neural Networks with Improved Particle Swarm Optimization," *Atmosphere (Basel).*, vol. 16, no. 7, p. 870, 2025, doi: 10.3390/atmos16070870.

[18] Y. Y. Ghadi *et al.*, "Explainable AI analysis for smog rating prediction," *Sci. Rep.*, vol. 15, p. 8070, 2025, doi: 10.1038/s41598-025-92788-x.

[19] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Syst. Appl.*, vol. 244, p. 122778, 2024, doi: 10.1016/j.eswa.2023.122778.

[20] Y. Han, Z. Wei, and G. Huang, "An imbalance data quality monitoring based on SMOTE-XGBOOST supported by edge computing," *Sci. Rep.*, vol. 14, p. 10151, 2024, doi: 10.1038/s41598-024-60600-x.

[21] A. R. Salehi and M. Khedmati, "A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data," *Sci. Rep.*, vol. 14, p. 5152, 2024, doi: 10.1038/s41598-024-55598-1.

[22] B. Nemade, V. Bharadi, S. S. Alegavi, and B. Marakarkandy, "A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 9s, pp. 790–803, 2023, [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/3268

[23] M. Mujtaba, "Air Quality and Pollution Assessment Dataset," 2023. [Online]. Available: https://www.kaggle.com/datasets/mujtabamatin/air-quality-and-pollution-assessment

