IMPLEMENTASI MODEL *RANDOM FOREST* UNTUK KLASIFIKASI KUALITAS UDARA DI JAKARTA PERIODE 2019-2024

Ernawati Prahesta Kurnia Sari[1] , Yesy Diah Rosita[2] , Siti Khomsah[3]
[1,2,3] Universitas Telkom
ernawatiphs@student.telkomuniversity.ac.id[1], yesydr@telkomuniversity.ac.id[2], sitijk@telkomuniversity.ac.id[3]

***Abstrak* — Peningkatan polusi udara di Jakarta menjadi isu lingkungan yang krusial karena dampaknya terhadap kesehatan masyarakat, terutama pada kelompok rentan seperti anak-anak dan lansia. Indeks Standar Pencemar Udara (ISPU) digunakan sebagai indikator kualitas udara di Indonesia, dengan kategori penilaian mulai dari "baik" hingga "berbahaya". Untuk meningkatkan akurasi pemantauan dan prediksi kualitas udara, penelitian ini menerapkan algoritma *Random Forest* dalam tugas klasifikasi berdasarkan data ISPU. Metode ini dipilih karena keandalannya dalam menangani data berukuran besar dan kompleks. Evaluasi dilakukan menggunakan berbagai skenario proporsi pembagian data (70:30, 80:20, dan 90:10). Hasil menunjukkan bahwa *Random Forest* mampu mencapai akurasi tinggi secara konsisten, dengan nilai tertinggi sebesar 0,9217 pada skenario 80:20. Nilai precision, recall, dan F1-score juga menunjukkan stabilitas performa di semua skenario. Temuan ini membuktikan bahwa Random Forest merupakan metode yang efektif dan andal dalam klasifikasi kualitas udara, serta dapat mendukung pengambilan keputusan dalam mitigasi dampak polusi udara di wilayah perkotaan.**

***Kata kunci*— klasifikasi, *machine learning, random forest*, udara**

I. PENDAHULUAN

Kualitas udara merupakan salah satu faktor lingkungan yang sangat memengaruhi kesehatan dan kesejahteraan manusia. Udara bersih berperan penting dalam menekan risiko penyakit pernapasan, gangguan kardiovaskular, hingga kanker paru-paru yang dapat timbul akibat paparan polutan udara jangka panjang [1]. Selain dampak kesehatan, kualitas udara yang baik juga berdampak positif terhadap produktivitas, kenyamanan aktivitas sehari-hari, dan kualitas hidup secara keseluruhan [2].

Di kota metropolitan besar seperti Jakarta, kualitas udara menunjukkan tren penurunan signifikan. Berdasarkan laporan IQAir pada Agustus 2023, Jakarta menempati peringkat kedua kota dengan kualitas udara terburuk di dunia dengan skor US Air Quality Index (AQI) sebesar 154, yang masuk kategori Tidak Sehat [3]. Kondisi ini menyebabkan peningkatan kasus penyakit pernapasan, iritasi mata, dan penyakit kronis yang berhubungan dengan sistem pernapasan serta kardiovaskular [4]. Data menunjukkan lebih dari 10.000 kematian per tahun di Jakarta terkait langsung dengan paparan polusi udara, dan lebih dari 5.000 kasus rawat inap disebabkan oleh gangguan pernapasan akut [5]. Sumber utama polusi meliputi emisi kendaraan bermotor, aktivitas industri, pembakaran sampah terbuka, dan kebakaran hutan, dengan partikel halus seperti PM2.5 dan PM10 sebagai ancaman terbesar karena kemampuannya menembus jauh ke saluran pernapasan dan aliran darah [6].

Populasi rentan seperti anak-anak dan lansia memiliki risiko lebih tinggi mengalami penurunan fungsi paru-paru permanen serta memburuknya penyakit kronis akibat paparan udara tercemar dalam jangka panjang [7]. Di Indonesia, kualitas udara dipantau menggunakan Indeks Standar Pencemar Udara (ISPU) yang mengklasifikasikan kondisi udara ke dalam lima kategori, mulai dari Baik, Sedang, Tidak Sehat, dan Sangat Tidak Sehat [8]. Ketika nilai ISPU melebihi ambang batas aman, risiko kesehatan masyarakat meningkat drastis sehingga dibutuhkan sistem pemantauan kualitas udara yang akurat dan berkelanjutan [9].

II. KAJIAN TEORI

Penelitian sebelumnya telah mengkaji berbagai model klasifikasi untuk menilai kualitas udara di Jakarta. Penelitian [10] yang menggunakan Random Forest pada data ISPU harian dari Februari hingga Oktober 2021 melaporkan akurasi klasifikasi sebesar 90%, dengan nilai presisi dan recall yang seimbang untuk kategori sedang dan tidak sehat. Penelitian lain [11] melaporkan metrik kinerja yang hampir sempurna dengan *Random Forest*, yaitu akurasi 100% pada data pelatihan dan 99,95% pada data pengujian, yang menunjukkan kemampuan generalisasi dan keandalan model yang sangat tinggi. Selain itu, penelitian komparatif [12] antara *Random Forest, Support Vector Machines* (SVM), dan Light Gradient Boosting Machine (LGBM) menunjukkan bahwa Random Forest mencapai akurasi klasifikasi tertinggi sebesar 98%, mengungguli SVM (95%) dan LGBM (88%). Temuan-temuan ini semakin memperkuat efektivitas Random Forest sebagai algoritma yang andal untuk klasifikasi kualitas udara, sehingga menjadi dasar pemilihan algoritma ini dalam penelitian ini.

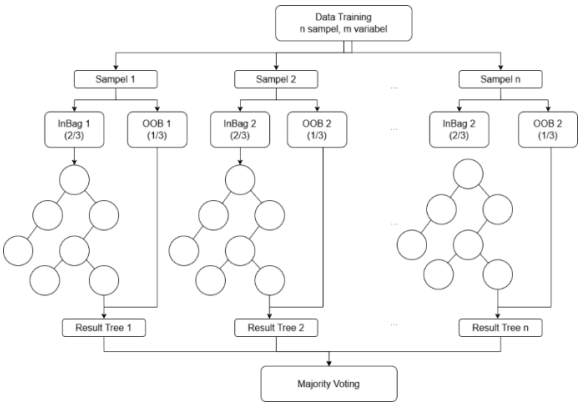
A. Pencemaran Udara

Pencemaran udara memiliki dampak yang signifikan terhadap kualitas udara. Pencemaran udara merupakan masuknya zat, energi, maupun komponen lain ke dalam udara karena kegiatan manusia, sehingga menyebabkan penurunan mutu udara. Semakin tinggi tingkat pencemaran udara, maka semakin besar pula potensi terjadinya gangguan kesehatan seperti asma, penyakit jantung, hingga gangguan fungsi paru [13].

Partikel pencemar udara meliputi partikulat (PM2.5 dan PM10) yang dapat menembus saluran pernapasan hingga ke aliran darah, nitrogen dioksida (NO_2) dari pembakaran bahan bakar fosil yang mengiritasi paru-paru, karbon monoksida (CO) yang mengurangi pasokan oksigen dalam darah, sulfur oksida (SO_x) dari pembakaran batu bara dan minyak yang memicu gangguan pernapasan, serta ozon (O_3) di permukaan bumi yang terbentuk dari reaksi kimia polutan dan sinar matahari, yang dapat memperburuk penyakit paru kronis [13].

B. Random Forest

Dalam klasifikasi, Random Forest menentukan prediksi berdasarkan voting mayoritas dari seluruh pohon. Random Forest menggunakan teknik bootstrap sampling dan pemilihan fitur secara acak untuk menciptakan variasi di antara pohon-pohonnya, yang membantu meningkatkan akurasi dan mengurangi overfitting [14].



GAMBAR 1
CARA KERJA RANDOM FOREST [15]

Proses Random Forest dimulai dengan bootstrap sampling, kemudian model akan memilih subset acak dari fitur untuk menentukan fitur terbaik yang akan digunakan dalam pembagian data untuk setiap node dalam pohon. Setelah semua pohon selesai dilatih, Random Forest membuat prediksi dengan menggabungkan hasil dari setiap pohon. Hasil akhirnya ditentukan melalui voting mayoritas dari semua pohon [15].

C. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE adalah salah satu metode penyeimbangan data yang digunakan untuk menangani permasalahan ketidakseimbangan kelas karena jumlah data pada satu kelas jauh lebih besar dibandingkan kelas lainnya sehingga dapat menyebabkan model cenderung bias terhadap kelas mayoritas. SMOTE bekerja dengan cara melakukan oversampling terhadap kelas minoritas dengan mensintesis data baru secara artificial. Teknik ini membuat sampel-sampel baru dengan cara mengambil titik data dari kelas minoritas, lalu mencari beberapa tetangga terdekatnya, kemudian menghasilkan data baru dengan interpolasi linear antara titik-titik tersebut [16]. Secara matematis, data sintetik dapat dilihat pada Persamaan 1.

$$x_{synthetic} = x_i + \delta \,.\, (x_{zi} - x_i) \quad (1)$$

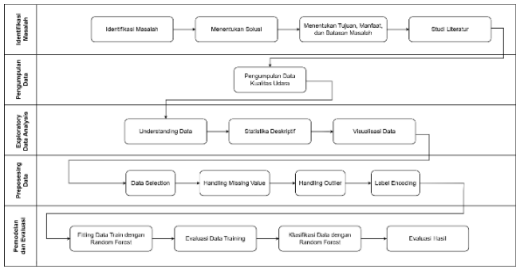
Pada Persamaan 1, nilai x_i merupakan sampel dari kelas minoritas, kemudian x_{zi} adalah salah satu dari k tetangga terdekat x_i yang juga merupakan kelas minoritas, nilai $\delta \in [0,1]$ adalah bilangan acak yang diambil dari distribusi uniform.

D. Evaluasi Sistem

Confusion matrix merupakan alat untuk mengukur kinerja model klasifikasi dalam machine learning dengan menampilkan jumlah prediksi yang benar dan salah pada setiap kelas. Matriks ini terdiri dari empat komponen utama, yaitu True Positive (TP) yang menunjukkan data positif yang diprediksi benar, True Negative (TN) yaitu data negatif yang diprediksi benar, False Positive (FP) yaitu data negatif yang salah diprediksi sebagai positif , dan False Negative (FN) yaitu data positif yang salah diprediksi sebagai negatif. Dari nilai-nilai ini, dapat dihitung metrik evaluasi seperti akurasi, presisi, recall, dan skor F1 [17].

III. METODE

Penelitian ini terdiri dari beberapa tahap utama, yaitu identifikasi masalah, pengumpulan data, analisis data eksploratori (EDA), prapemrosesan data, pemodelan menggunakan algoritma Random Forest, dan evaluasi hasil klasifikasi. Alur keseluruhan proses penelitian ditunjukkan pada Gambar 2.



GAMBAR 2
DIAGRAM ALIR PENELITIAN

A. Identifikasi Masalah

Tahap awal penelitian adalah mengidentifikasi masalah, yaitu isu lingkungan berupa polusi udara di Jakarta sebagai topik yang relevan dan signifikan untuk dikaji. Solusi yang diajukan adalah penerapan teknik klasifikasi data mining. Peneliti merumuskan tujuan penelitian, manfaat (akademis dan praktis), serta batasan penelitian. Tinjauan pustaka dilakukan untuk mengetahui penelitian terdahulu dan research gap yang menjadi dasar pemilihan metode dan perumusan hipotesis.

B. Pengumpulan Data

Dataset yang digunakan adalah Indeks Standar Pencemar Udara (ISPU) periode 2019–2024 dari lima Stasiun

Pemantauan Kualitas Udara (SPKU) di Jakarta: DKI1 (Jakarta Pusat), DKI2 (Jakarta Utara), DKI3 (Jakarta Selatan), DKI4 (Jakarta Timur), dan DKI5 (Jakarta Barat). Variabel polutan yang dicatat meliputi PM_{10}, SO_2, CO, O_3, dan NO_2, beserta kategori ISPU (Baik, Sedang, Tidak Sehat, Sangat Tidak Sehat).

TABEL 1
CONTOH DATASET ISPU

pm10	so2	co	o3	no2	categori
29	15	5	NaN	13	Baik
24	17	5	NaN	6	Baik
59	53	21	36	20	Sedang
78	53	29	43	30	Sedang

C. Exploratory Data Analysis (EDA)

Statistik deskriptif digunakan untuk merangkum karakteristik utama dari dataset pencemaran udara sebelum dilakukan analisis dan pemodelan lebih lanjut. Tujuan dari tahap ini adalah untuk memperoleh pemahaman awal mengenai distribusi, ukuran pemusatan, dan variabilitas dari setiap variabel polutan dalam dataset. Tabel 2 menyajikan ringkasan statistik untuk lima polutan udara utama, yaitu PM_{10}, SO_2, CO, O_3, dan NO_2. Untuk setiap variabel, tabel memuat nilai minimum, maksimum, rata-rata, median, dan simpangan baku.

TABEL 2
STATISTIKA DESKRIPTIF

	pm10	so2	co	o3	no2
Count	8867	9081	9146	9150	9102
Mean	53.57	31.21	16.45	44.48	19.28
Std	15.007	16.57	13.08	30.27	16.50
Min	3	1	1	1	1
25%	44	18	9	23	9
50%	55	27	14	35	15
75%	63	45	20	60	24
Max	187	112	213	243	213

D. Data Preprocessing

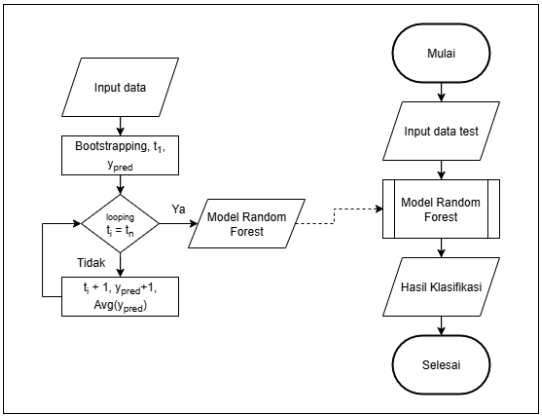
Tahap preprocessing merupakan langkah penting dalam menyiapkan data mentah untuk analisis dan pemodelan. Tahap ini memastikan bahwa dataset bersih, konsisten, dan sesuai untuk dimasukkan ke dalam algoritma machine learning. Pada penelitian ini, tahap preprocessing mencakup beberapa langkah: pemilihan data, penanganan nilai hilang, penanganan outlier, dan label encoding. Tabel 3 menunjukkan struktur dataset setelah seluruh langkah preprocessing diselesaikan. Setiap baris kini merepresentasikan observasi yang bersih, siap dalam bentuk numerik, serta hanya memuat fitur polutan yang relevan beserta label kualitas udara yang telah dikodekan. Dataset hasil preprocessing ini menjadi masukan untuk tahap selanjutnya, yaitu pelatihan model dan klasifikasi menggunakan Random Forest.

TABEL 3
CONTOH DATASET SETELAH PREPROCESSING

pm10	so2	co	o3	no2	categori
29	15	5	44.48	13	0
24	17	5	44.48	6	0
59	53	21	36	20	1

E. Pemodelan dan Evaluasi

Tahap pemodelan dalam penelitian ini menggunakan algoritma Random Forest, yaitu metode machine learning berbasis ensemble yang kuat dan sesuai untuk tugas klasifikasi yang melibatkan data lingkungan terstruktur. Random Forest bekerja dengan membangun sejumlah decision tree selama proses pelatihan dan menghasilkan kelas akhir berdasarkan suara mayoritas dari setiap pohon. Algoritma ini dikenal memiliki ketangguhan, ketahanan terhadap overfitting, serta kemampuan menangani data berdimensi tinggi. Untuk menilai kemampuan generalisasi model, dataset hasil preprocessing dibagi menjadi data latih dan data uji dengan tiga rasio pembagian berbeda: 90:10, 80:20, dan 70:30. Variasi ini digunakan untuk mengamati pengaruh ukuran data latih terhadap akurasi dan kestabilan model.



GAMBAR 3
FLOWCHART RANDOM FOREST

Pelatihan model dilakukan pada setiap skenario pembagian data, dan evaluasi dilakukan menggunakan beberapa metrik klasifikasi standar seperti akurasi, presisi, recall, dan F1-score.

1) Akurasi, digunakan untuk mengukur tingkat keseluruhan kebenaran prediksi.

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

dengan TP = prediksi positif yang benar, TN = prediksi negatif yang benar, FP = prediksi positif yang salah, dan FN = prediksi negatif yang salah.

2) Presisi, digunakan untuk mengukur proporsi prediksi positif yang benar-benar tepat.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} Presisi = \frac{TP}{TP+FP} \quad (3)$$

dengan TP = prediksi positif yang benar dan FP = prediksi positif yang salah.

3) Recall, digunakan untuk menilai kemampuan model dalam mendeteksi semua kasus positif yang relevan.

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

dengan TP = prediksi positif yang benar dan FN = prediksi negatif yang salah.

4) F1-score, yaitu metrik yang menggabungkan presisi dan recall menjadi satu ukuran.

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

IV. HASIL DAN PEMBAHASAN

Sebelum dilakukan evaluasi akhir, serangkaian percobaan dilakukan untuk menentukan hyperparameter optimal pada

Random Forest classifier. Beberapa kombinasi parameter diuji untuk meningkatkan kinerja model.

TABEL 4
KOMBINASI PARAMETER PENGUKURAN

Parameter	Nilai
N estimator	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200
Max depth	None, 5, 10, 15, 20
Max feature	None, sqrt, log2
Min sample leaf	1, 2, 3, 4, 5
Min sample split	2, 3, 4, 5, 6, 7, 8, 9, 10
Criterion	Gini, Entropy, Log Loss

Setelah mengevaluasi berbagai kombinasi, konfigurasi dengan kinerja terbaik diperoleh pada nilai parameter dengan n_estimators = 120, max_depth = 20, max_features = 'log2', min_samples_leaf = 1, min_samples_split = 9, dan criterion = 'log_loss'. Parameter ini kemudian diterapkan pada model akhir yang dievaluasi menggunakan skenario pembagian data berbeda sebagaimana dijelaskan pada bagian berikutnya.

Model klasifikasi dilatih dan diuji dengan tiga rasio pembagian data berbeda, yaitu 90:10, 80:20, dan 70:30, untuk mengamati pengaruh proporsi data latih terhadap kinerja model. Nilai akurasi untuk setiap rasio pembagian ditunjukkan pada Tabel 5.

TABEL 5
PERBANDINGAN AKURASI RF BERDASARKAN UKURAN SPLIT DATASET

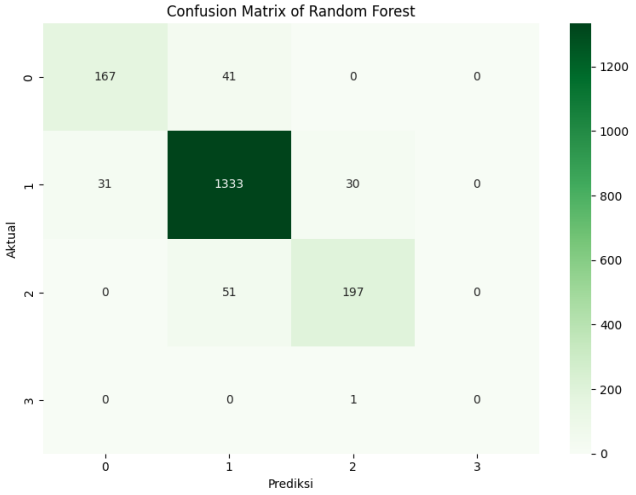
Ukuran Split Train dan Test	Akurasi
90:10	0.9168
80:20	0.9217
70:30	0.9204

Berdasarkan Tabel 5, model mencapai akurasi tertinggi sebesar 0,9217 pada pembagian data 80:20, yang menunjukkan keseimbangan optimal antara data latih dan data uji. Ketiga konfigurasi menghasilkan nilai akurasi yang tinggi, sehingga menunjukkan konsistensi dan kemampuan generalisasi model pada berbagai proporsi data latih. Evaluasi lebih lanjut dilakukan menggunakan classification report yang disajikan pada Tabel 6, yang memuat nilai presisi, recall, dan F1-score untuk setiap kategori kualitas udara.

TABEL 6
CLASSIFICATION REPORT RF

Kelas	Presisi	Recall	F1-Score	Support
0	0.85	0.80	0.82	208
1	0.93	0.96	0.95	1394
2	0.86	0.79	0.82	248
3	0.00	0.00	0.00	1

Berdasarkan Tabel 6, model Random Forest memperoleh akurasi keseluruhan sebesar 0,92 dari total 1.851 data uji. Kelas 0 (Baik) mencatat presisi 0,85, recall 0,80, dan F1-score 0,82 pada 208 data uji, menunjukkan kinerja yang cukup baik meskipun recall sedikit lebih rendah daripada presisi. Kelas 1 (Sedang) memiliki performa terbaik dengan presisi 0,93, recall 0,96, dan F1-score 0,95 pada 1.394 data uji, mencerminkan kemampuan model yang sangat tinggi dalam mengenali kelas ini. Kelas 2 (Tidak Sehat) juga menunjukkan hasil yang baik dengan presisi 0,86, recall 0,79, dan F1-score 0,82 pada 248 data uji, walaupun recall masih di bawah presisi. Sementara itu, kelas 3 (Sangat Tidak Sehat) tidak terdeteksi sama sekali (presisi, recall, dan F1-score = 0,00) karena hanya terdapat satu data uji, sehingga model tidak dapat mempelajari pola yang representatif. Nilai macro average sebesar presisi 0,66, recall 0,64, dan F1-score 0,65 menunjukkan bahwa meskipun akurasi model tinggi, kinerja antar kelas masih belum merata, terutama pada kelas dengan jumlah data yang sangat sedikit.



GAMBAR 4
CONFUSION MATRIX RF

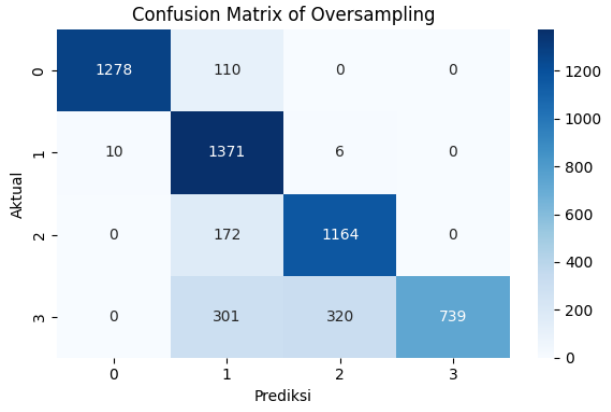
Confusion matrix Random Forest menunjukkan bahwa model memiliki kinerja sangat baik pada kelas mayoritas, khususnya kelas 1 dengan 1.333 prediksi benar dan sedikit kesalahan ke kelas 0 dan kelas 2. Kelas 0 terdeteksi dengan 167 prediksi benar namun terdapat 41 kesalahan klasifikasi ke kelas 1. Kelas 2 hanya memiliki 197 prediksi benar, dengan 51 data salah diklasifikasikan ke kelas 1, menandakan kesulitan model dalam mengenali kelas ini. Kelas 3 hampir tidak terdeteksi sama sekali, hanya ada 1 prediksi benar dari jumlah data yang sangat sedikit. Hasil ini menunjukkan bahwa meskipun akurasi keseluruhan tinggi, model cenderung bias terhadap kelas mayoritas dan kurang optimal dalam mendeteksi kelas minoritas.

Selanjutnya, dilakukan penerapan metode Synthetic Minority Over-sampling Technique (SMOTE) untuk mengatasi masalah ketidakseimbangan kelas pada dataset. SMOTE bekerja dengan cara menghasilkan data sintetis pada kelas minoritas melalui interpolasi antara data yang ada, sehingga distribusi data antar kelas menjadi lebih seimbang. Proses ini diharapkan dapat meningkatkan kemampuan model dalam mengenali dan memprediksi kelas dengan jumlah data yang sedikit, khususnya kategori Tidak Sehat dan Sangat Tidak Sehat yang sebelumnya menunjukkan performa rendah. Dataset hasil SMOTE kemudian digunakan kembali dalam proses pelatihan dan pengujian model Random Forest dengan prosedur evaluasi yang sama seperti sebelumnya, guna membandingkan perubahan kinerja model setelah dilakukan penyeimbangan data.

TABEL 7.
CLASSIFICATION REPORT SMOTE

Kelas	Presisi	Recall	F1-Score	Support
0	0.99	0.92	0.95	1388
1	0.67	0.98	0.80	1387
2	0.81	0.84	0.82	1336
3	1.00	0.56	0.72	1360

Berdasarkan hasil classification report setelah penerapan SMOTE, model Random Forest memperoleh akurasi keseluruhan sebesar 0,83 dari total 5.471 data uji. Kelas 0 (Baik) memiliki presisi 0,99, recall 0,92, dan F1-score 0,95 pada 1.388 data uji, menunjukkan kinerja yang sangat tinggi. Kelas 1 (Sedang) mencatat presisi 0,67 dan recall 0,98 dengan F1-score 0,80 pada 1.387 data uji, menandakan kemampuan deteksi yang kuat meskipun presisinya lebih rendah. Kelas 2 (Tidak Sehat) menunjukkan hasil seimbang dengan presisi 0,81, recall 0,84, dan F1-score 0,82 pada 1.336 data uji. Kelas 3 (Sangat Tidak Sehat) memiliki presisi sempurna 1,00, recall 0,56, dan F1-score 0,72 pada 1.360 data uji, mengindikasikan semua prediksi positif benar namun masih banyak data kelas ini yang tidak terdeteksi. Nilai macro average sebesar presisi 0,87, recall 0,83, dan F1-score 0,82 menunjukkan kinerja yang lebih merata antar kelas dibandingkan sebelum penerapan SMOTE, dengan peningkatan signifikan pada kelas minoritas walaupun akurasi keseluruhan menurun.



GAMBAR 5
CONFUSION MATRIX SMOTE

Confusion matrix hasil oversampling menggunakan SMOTE menunjukkan bahwa model Random Forest mampu mengklasifikasikan sebagian besar data dengan benar pada setiap kelas. Kelas 0 dan kelas 1 memiliki jumlah prediksi benar yang sangat tinggi (1.278 dan 1.371), dengan sedikit kesalahan prediksi ke kelas lain. Kelas 2 juga teridentifikasi dengan baik (1.164 benar) meskipun ada 172 data yang salah diklasifikasikan ke kelas 1. Kelas 3 memiliki prediksi benar sebanyak 739, namun masih terdapat cukup banyak kesalahan prediksi ke kelas 1 (301) dan kelas 2 (320), yang menjadi tantangan utama dalam klasifikasi. Secara keseluruhan, SMOTE membantu meningkatkan deteksi kelas minoritas, meskipun masih ada kebingungan antar kelas dengan distribusi fitur yang mirip.

V. KESIMPULAN

Berdasarkan hasil pengujian, model *Random Forest* tanpa penerapan SMOTE menghasilkan akurasi tertinggi sebesar 92%, namun nilai *precision, recall*, dan F1-*score* masing-masing hanya sebesar 66%, 64%, dan 65%, yang menunjukkan adanya ketidakseimbangan kinerja antar kelas. Penerapan SMOTE menurunkan akurasi menjadi 83%, tetapi secara signifikan meningkatkan *precision*, *recall*, dan F1-*score* menjadi 87%, 83%, dan 82%. Hal ini membuktikan bahwa SMOTE mampu meningkatkan pemerataan performa model dalam mengklasifikasikan data pada setiap kelas, terutama pada kelas minoritas, meskipun harus mengorbankan sebagian tingkat akurasi keseluruhan.

Secara keseluruhan, penerapan SMOTE dapat menjadi strategi yang efektif untuk mengatasi ketidakseimbangan kelas pada data kualitas udara. Penelitian selanjutnya disarankan untuk mengombinasikan SMOTE dengan metode optimasi parameter atau algoritma lain yang fokus pada peningkatan deteksi kelas minoritas agar diperoleh kinerja model yang lebih seimbang tanpa penurunan akurasi yang signifikan.

REFERENSI

[1] A. Hidayat, “Dampak Polusi Udara Pada Kesehatan Jantung,” *Univ. Medan Area*, no. November, pp. 1–12, 2019.

[2] S. Rahmawati and I. N. Pratama, “Pengaruh Penggunaan Transportasi Berkelanjutan Terhadap Kualitas Udara Dan Kesejahteraan Masyarakat,” *JEPTEC J. Enviromental Policy Technol.*, vol. 1, no. 2, pp. 90–99, 2023.

[3] *IQAir, “Air Quality Index.” Accessed*: Dec. 15, 2023. [*Online*]. *Available*: https://www.iqair.com/indonesia

[4] A. P. Marpaung, “Evaluasi Kualitas Udara dan Dampaknya Terhadap Kesehatan Pernafasan Penduduk Kota Medan,” *J. Ilm. Maksitek*, vol. 8, no. 2, pp. 105–111, 2023.

[5] Y. Abdurrohman, Y. Sriharyani, M. Syaiful, and C. D. A. Sembiring, “Pengaruh Terpaan Media Sosial Terhadap Persepsi Risiko Kesehatan (Survei pada Isu Polusi Udara Jakarta),” *J. Stud. Komun. dan Media*, vol. 28, no. 1, pp. 89–104, 2024, doi: 10.17933/jskm.2023.5620.

[6] M. A. Fath, “Pengaruh Kualitas Udara dan Kondisi Iklim terhadap Perekonomian Masyarakat,” *Media Gizi Kesmas*, vol. 10, no. 2, p. 329, 2021, doi: 10.20473/mgk.v10i2.2021.329-342.

[7] R. Umah and E. Gusmira, “Dampak Pencemaran Udara terhadap Kesehatan Masyarakat di Perkotaan,” *Profit J. Manajemen, Bisnis dan Akunt.*, vol. 3, no. 3, pp. 103–112, 2024, doi: 10.58192/profit.v3i3.2246.

[8] Menhlk, “Indeks Standar Pencemar Udara.” *Accessed:* Dec. 15, 2023. [*Online*]. *Available*: https://ispu.menlhk.go.id/webv4/

[9] Q. Xu, L. Ning, Y. Tianmeng, and H. Wu, “*Application of Data Mining Combined with Power Data in Assessment and Prevention of Regional Atmospheric Pollution*,” *Elsevier*, vol. 9, pp. 3397–

3405, 2023, doi: 10.1016/j.egyr.2023.02.016.

[10] A. Nugroho, I. Asror, and Y. A. Wibowo, "Klasifikasi Tingkat Kualitas Udara DKI Jakarta Berdasarkan Open Government Data Menggunakan Algoritma *Random Forest*," *eProceedings Eng.*, vol. 10, no. 2, pp. 1824–1834, 2023, [*Online]. Available*: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20030%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20030/19395

[11] R. Firdaus *et al.*, "Implementasi Algoritma Random Forest Untuk Klasifikasi Pencemaran Udara di Wilayah Jakarta Berdasarkan Jakarta Open Data," vol. 14, no. 2, pp. 520–525, 2021.

[12] A. M. Luthfi and F. Fauzi, "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*," vol. 9, no. 2, pp. 99–108, 2024.

[13] D. Natalia, "Laporan Pemantauan Kualitas Udara Tahun 2013," 2013.

[14] Yoga Religia, Agung Nugroho, and Wahyu Hadikristanto, "Klasifikasi Analisis Perbandingan Algoritma Optimasi pada *Random Forest* untuk Klasifikasi Data Bank Marketing," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 187–192, 2021, doi: 10.29207/resti.v5i1.2813.

[15] R. I. Arumnisaa and A. W. Wijayanto, "*Comparison of Ensemble Learning Method: Random Forest, Support Vector Machine, AdaBoost for Classification Human Development Index* (HDI)," *Sistemasi*, vol. 12, no. 1, p. 206, 2023, doi: 10.32520/stmsi.v12i1.2501.

[16] A. Syukron, S. Sardiarinto, E. Saputro, and P. Widodo, "Penerapan Metode *Smote* Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung," *J. Teknol. Inf. dan Terap.*, vol. 10, no. 1, pp. 47–50, 2023, doi: 10.25047/jtit.v10i1.313.

[17] D. Marutho, "Perbandingan Metode Naïve Bayes, KNN, *Decision Tree* Pada Laporan Water Level Jakarta," *Infokam*, vol. 15, no. 2, pp. 90–97, 2019.