Jurnal Informatika: Jurnal pengembangan IT

Vol. 10, No. 4, 2025
ISSN : 2477-5126, DOI:10.30591/jpit.v10i4.9466

Implementasi Random Oversampling pada Algoritma Random Forest untuk Klasifikasi Indeks Standar Udara di DKI Jakarta 2024

Harsena Argretya [1], Kusnawi [2]

[1] Program Studi Informatika, Universitas Amikom Yogyakarta, Kaliurang, Yogayakrta, 55281, Indonesia

Info Artikel

Riwayat Artikel:

Received 2025-07-27
Revised 2025-10-16
Accepted 2025-10-21

Abstract *– Air pollution remains one of the most pressing global environmental issues, posing serious risks to public health and ecosystems, especially in densely populated metropolitan areas such as DKI Jakarta. The difficulty in analyzing air quality data arises not only from high pollutant concentrations but also from the inherent class imbalance within the Air Pollution Standard Index (ISPU) dataset. Most samples fall within the "Moderate" category, while the "Good" and "Unhealthy" classes are underrepresented, leading machine learning models to develop a bias toward the majority class. This study applies the Random Oversampling (ROS) technique in combination with the Random Forest algorithm to improve the classification performance of ISPU categories using 2024 air quality data. The workflow consists of data acquisition from the Satu Data Jakarta platform, preprocessing using KNN imputation, class balancing through ROS, model training using Random Forest, and performance evaluation using accuracy, precision, recall, and F1-score. Experimental results indicate that ROS enhances the overall classification accuracy from 96.71% to 97.07%. An improvement is also observed in the recall for the "Unhealthy" class, increasing from 0.93 to 0.94, while the recall for the "Good" class remains at 0.88. Although the performance gain is modest, the implementation of ROS leads to a more balanced learning process and improved model stability. The study also highlights a limitation of ROS, namely its tendency to duplicate existing samples without generating additional variation, suggesting that more advanced oversampling techniques such as SMOTE or ADASYN may be considered in future research. Overall, the integration of Random Forest and ROS strengthens data-driven air quality monitoring and supports environmental decision-making in urban settings.*

Keywords: *Air Quality, Class Imbalance, ISPU Classification, Random Forest, Random Oversampling*

Corresponding Author:

Harsena Argretya

Email: harsenaargrtya1@gmail.com

Abstrak *– Degradasi kualitas udara telah menjadi tantangan lingkungan yang signifikan dengan implikasi serius bagi kesehatan publik dan keberlanjutan ekosistem. Fenomena ini sangat krusial di kawasan metropolitan seperti DKI Jakarta, di mana konsentrasi penduduk yang padat memperburuk risiko paparan polutan secara luas. Tantangan utama dalam analisis kualitas udara tidak hanya terletak pada tingginya konsentrasi polutan, tetapi juga pada ketidakseimbangan kelas dalam data ISPU yang menyebabkan hasil klasifikasi menjadi kurang akurat. Data ISPU umumnya didominasi oleh kategori "Sedang", sedangkan kategori "Baik" dan "Tidak Sehat" tercatat jauh lebih sedikit sehingga model cenderung bias terhadap kelas mayoritas. Penelitian ini menerapkan teknik Random Oversampling (ROS) yang dikombinasikan dengan algoritma Random Forest untuk meningkatkan akurasi klasifikasi ISPU menggunakan data tahun 2024. Tahapan penelitian meliputi pengumpulan data dari portal Satu Data Jakarta, preprocessing dengan metode imputasi KNN, penyeimbangan kelas menggunakan ROS, pelatihan model Random Forest, serta evaluasi performa melalui metrik akurasi, presisi, recall, dan F1-score. Hasil eksperimen menunjukkan bahwa penggunaan ROS mampu meningkatkan akurasi model dari 96,71% menjadi 97,07%. Peningkatan juga terlihat pada nilai recall kelas "Tidak Sehat", yang naik dari 0,93 menjadi 0,94, sementara recall untuk kelas "Baik" tetap berada pada nilai 0,88. Meskipun peningkatannya relatif kecil, ROS memberikan proses pembelajaran yang lebih seimbang dan meningkatkan stabilitas model. Penelitian ini juga menegaskan keterbatasan ROS yang hanya menggandakan data tanpa menghasilkan variasi baru, sehingga metode lain seperti SMOTE atau ADASYN direkomendasikan untuk penelitian selanjutnya. Secara keseluruhan, integrasi Random Forest dan ROS mendukung pengembangan sistem pemantauan kualitas udara berbasis data untuk pengambilan keputusan lingkungan di wilayah perkotaan.*

Kata Kunci*: Kualitas Udara, Klasifikasi ISPU, Ketidakseimbangan Data, Random Forest, Random Oversampling*

I. PENDAHULUAN

Di seluruh dunia, polusi udara kini dipandang sebagai tantangan lingkungan yang memerlukan perhatian khusus, terutama pada pusat-pusat kota yang didominasi oleh kepadatan lalu lintas dan kegiatan industri. Kondisi atmosfer yang terkontaminasi ini berkorelasi langsung dengan memburuknya status kesehatan masyarakat serta terganggunya kenyamanan hidup di lingkungan perkotaan [1]. WHO melaporkan bahwa lebih dari 90%

penduduk dunia tinggal di area yang tidak memenuhi standar kualitas udara yang ditetapkan, sehingga miliaran orang berada dalam risiko tinggi terhadap berbagai penyakit akibat polusi udara [2]. Indonesia tercatat berada pada urutan ke-17 secara global dan menjadi negara dengan tingkat polusi udara tertinggi di kawasan Asia Tenggara [3]. Kondisi ini juga sangat terlihat di DKI Jakarta sebagai ibu kota negara, di mana pada Agustus 2023 wilayah tersebut dilaporkan menempati posisi ketiga sebagai kota dengan kualitas udara paling parah di dunia [4].

Indeks Standar Pencemar Udara (ISPU) ditentukan berdasarkan enam jenis polutan utama, yaitu partikulat PM10, partikulat halus PM2.5, nitrogen dioksida (NO_2), sulfur dioksida (SO_2), karbon monoksida (CO), serta ozon (O_3) [5]. Jika kadar polutan melebihi ambang batas yang ditentukan, dampaknya bisa sangat merugikan, seperti menyebabkan iritasi saluran pernapasan, penyakit kronis, bahkan kematian dini [6]. Oleh karena itu, pemantauan ISPU secara akurat dan berkala sangat diperlukan sebagai dasar pengambilan keputusan kebijakan lingkungan dan kesehatan publik.

Metode machine learning banyak digunakan untuk analisis dan klasifikasi kualitas udara karena mampu mempelajari pola kompleks antar variabel polutan secara otomatis. Pendekatan ini dapat dibedakan menjadi model parametrik dan non-parametrik [7], di mana metode non-parametrik seperti *Random Forest* lebih fleksibel karena tidak mengasumsikan bentuk distribusi data tertentu [8]. Beberapa penelitian telah menunjukkan keunggulan Random Forest untuk klasifikasi ISPU. Penelitian Firdaus mengimplementasikan algoritma *Random Forest* untuk memprediksi ISPU di Jakarta Open Data dan memperoleh akurasi mencapai 99% [1]. Penelitian Irwansyah membandingkan algoritma *Decision Tree*, *Naïve Bayes*, dan *K-Nearest Neighbor*, dan menemukan bahwa *Random Forest* memberikan hasil paling stabil terhadap variasi polutan [6]. Penelitian Jayadi juga menjelaskan bahwa algoritma *ensemble learning* seperti RF dan *Extreme Gradient Boosting* (XGBoost) memberikan kinerja yang baik untuk klasifikasi kualitas udara, namun masih terpengaruh oleh distribusi data yang tidak seimbang [9].

Permasalahan utama pada penelitian-penelitian tersebut adalah ketidakseimbangan distribusi kelas (*class imbalance*) dalam dataset kualitas udara. Sebagian besar data cenderung didominasi oleh kategori "SEDANG", sedangkan kategori "BAIK" dan "TIDAK SEHAT" jauh lebih sedikit. Kondisi ini menyebabkan model cenderung bias terhadap kelas mayoritas dan kurang akurat dalam mengenali kelas minoritas. Sajiwo menerapkan algoritma *XGBoost* dengan teknik *Synthetic Minority Oversampling Technique* (SMOTE) untuk mengatasi ketimpangan data, tetapi penelitian tersebut hanya dilakukan pada dataset tahun 2021 sehingga belum mewakili kondisi polusi terbaru di DKI Jakarta [3]. Penelitian lain membandingkan efektivitas metode *oversampling* dan *undersampling* pada prediksi kualitas udara di wilayah metropolitan, dan hasilnya menunjukkan bahwa efektivitas kedua metode tersebut tergantung pada algoritma yang digunakan, dengan *oversampling* (SMOTE+NCR) menunjukkan performa terbaik pada model KNN, sementara *undersampling* (NCR) lebih baik untuk *Naive Bayes* [10]. Putra dan Salam membandingkan teknik SMOTE, ADASYN, dan *Random Oversampling* pada algoritma *Random Forest*, *Decision Tree*, dan *LightGBM*, dan menemukan bahwa *Random Oversampling* memberikan performa terbaik untuk *Random Forest* dengan akurasi 0,85, recall 0,888, dan F1-score 0,879, menunjukkan pentingnya pemilihan teknik resampling yang sesuai dengan algoritma klasifikasi [11].

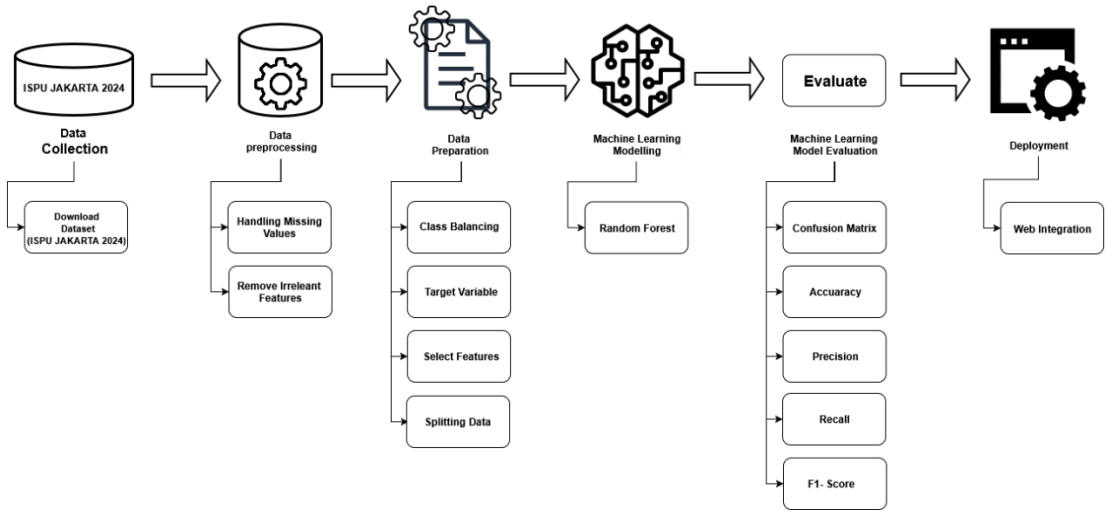
Selain penyeimbangan data, penanganan terhadap data hilang (*missing value*) juga berpengaruh pada akurasi model. Pendekatan gabungan seperti *Median–KNN Regressor–SMOTE–Tomek Links* terbukti efektif dalam mengatasi data hilang dan distribusi kelas tidak seimbang pada prediksi kualitas udara [12]. Hingga saat ini, belum ditemukan penelitian yang secara khusus menguji penerapan teknik Random Oversampling bersama algoritma Random Forest dalam mengklasifikasikan ISPU di DKI Jakarta dengan menggunakan data terbaru tahun 2024. Dataset tersebut memiliki ciri khas tersendiri, yaitu tingginya proporsi kelas "SEDANG" dan adanya ketimpangan yang cukup besar antara kelas mayoritas dan minoritas.

Dengan mempertimbangkan permasalahan tersebut, penelitian ini dilakukan untuk menganalisis dan menilai efektivitas penerapan Random Oversampling pada algoritma Random Forest dalam tugas klasifikasi Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta untuk tahun 2024. Data yang digunakan bersumber dari portal Satu Data Jakarta dan mencakup seluruh parameter polutan utama sepanjang Januari hingga Desember 2024. Melalui penelitian ini, diharapkan diperoleh kontribusi ilmiah terkait penanganan ketidakseimbangan kelas pada klasifikasi ISPU serta diperoleh model yang lebih akurat dan dapat diandalkan sebagai dasar dalam sistem pemantauan kualitas udara di kawasan perkotaan.

II. METODE

Penelitian ini menerapkan pendekatan analisis kuantitatif eksploratif dengan memanfaatkan algoritma machine learning sebagai bagian dari kecerdasan buatan *(Artificial Intelligence/AI*). Penelitian ini mengadopsi pendekatan hibrida yang mengintegrasikan algoritma Random Forest (RF) dengan teknik Random Oversampling (ROS) guna mengoptimalkan klasifikasi data Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta. Langkah

ini diambil sebagai strategi mitigasi terhadap kerentanan RF dalam menghadapi dataset yang tidak seimbang, di mana algoritma tersebut cenderung bias pada kelas mayoritas meskipun dikenal memiliki akurasi tinggi pada domain data lingkungan. Di sisi lain, *Random Oversampling* diketahui efektif dalam menyeimbangkan distribusi kelas pada dataset yang tidak seimbang. Dataset yang digunakan berasal dari portal Satu Data Jakarta, berisi data pemantauan kualitas udara tahun 2024 dengan enam parameter pencemar utama, yaitu PM10, PM2.5, CO, SO_2, NO_2, dan O_3. Alur dan tahapan penelitian yang dilakukan tersaji pada Gambar 1.



Gambar 1. Tahapan Penelitian

A. Identifikasi Masalah

Pencemaran udara adalah salah satu isu lingkungan yang cukup serius, terutama di kawasan perkotaan seperti DKI Jakarta. Berbagai laporan dari lembaga pemantau lingkungan menunjukkan bahwa kualitas udara di Jakarta sering berada pada kategori tidak sehat dan bahkan termasuk dalam deretan kota dengan kualitas udara terburuk di tingkat global. Situasi ini semakin memburuk dengan meningkatnya jumlah kendaraan bermotor, aktivitas industri, serta pengaruh kondisi meteorologis yang berdampak pada tingginya konsentrasi polutan di atmosfer. Oleh sebab itu, diperlukan strategi analisis berbasis data untuk melakukan klasifikasi dan prediksi Indeks Standar Pencemar Udara (ISPU), sehingga pemerintah dan masyarakat dapat mengambil tindakan pencegahan secara lebih tepat dan cepat.

B. Data Collection

Data penelitian ini diperoleh melalui portal resmi Satu Data Jakarta yang dapat diakses pada alamat https://satudata.jakarta.go.id Dataset tersebut berisi informasi pemantauan kualitas udara di wilayah DKI Jakarta untuk periode Januari hingga Desember 2024. Data disajikan dalam format CSV dan mencakup total 1.830 entri. Detail struktur dataset secara lengkap ditampilkan pada Tabel 1.

TABEL 1
INFORMASI DATASET

No.	Column	Count	Data Type
1	periode_data	1830	int64
2	bulan	1830	int64
3	tanggal	1830	int64
4	stasiun	1830	object
5	pm_sepuluh	1718	float64
6	pm_duakomalima	1810	float64
7	sulfur_dioksida	1821	float64
8	karbon_monoksida	1818	float64
9	ozon	1821	float64
10	nitrogen_dioksida	1804	float64
11	max	1825	float64
12	parameter_pencemaran_kritis	1784	object
13	kategori	1830	object

C. Data Pre-processing

Proses *pre-processing* dilakukan untuk menjamin bahwa data yang akan digunakan dalam pelatihan model berada dalam keadaan terstruktur, bebas dari kesalahan, dan siap diolah lebih lanjut. Proses ini diawali dengan pembersihan data (data cleansing), yang mencakup identifikasi serta penanganan nilai hilang (missing values) pada dataset. Dalam penelitian ini, penanganan nilai hilang dilakukan menggunakan metode *K-Nearest Neighbors* (KNN) Imputer, yang menggantikan nilai kosong berdasarkan rata-rata nilai dari beberapa tetangga terdekat (*nearest neighbors*) dalam ruang fitur. Metode ini dipilih karena mampu mempertahankan hubungan antarvariabel dan mengurangi potensi hilangnya informasi penting dibandingkan metode penghapusan data. Penelitian oleh Fandi Yulian menerangkan bahwasanya metode KNN Imputer menunjukan output yang lebih akurat dibandingkan dengan imputasi sederhana seperti mean imputation, terutama pada dataset dengan karakteristik non-linear dan multivariable [13]. Oleh karena itu, penggunaan KNN Imputer dalam penelitian ini diharapkan mampu memperbaiki kualitas data dan meningkatkan stabilitas model dalam tahap pelatihan.

Selain menangani nilai yang hilang, tahap *pre-processing* juga mencakup penghapusan fitur atau kolom yang dianggap tidak relevan terhadap tujuan klasifikasi. Fitur-fitur yang tidak memiliki kontribusi signifikan atau mengandung informasi yang redundan dapat memperbesar kompleksitas model dan memperlambat proses pelatihan. Oleh karena itu, penghapusan fitur-fitur yang tidak relevan penting dilakukan untuk meningkatkan efisiensi komputasi dan fokus model pada informasi yang benar-benar penting. penghapusan fitur yang tidak relevan dapat membantu dalam menyederhanakan proses klasifikasi serta meningkatkan akurasi model secara signifikan [14].

D. Data Preparation

Tahapan data preparation merupakan proses dalam sistem machine learning yang bertujuan untuk memastikan bahwa data siap digunakan dalam pelatihan dan pengujian model. Pada tahap ini, dilakukan penentuan variabel target yang akan diklasifikasikan. Dalam penelitian ini, kolom "kategori" dari dataset ISPU digunakan sebagai variabel target karena merepresentasikan tingkat kualitas udara, yaitu BAIK, SEDANG, dan TIDAK SEHAT. Selanjutnya dilakukan proses feature selection untuk memilih kolom input yang relevan dengan prediksi target. Fitur-fitur yang digunakan meliputi parameter pencemar udara PM10, PM2.5, CO, SO_2, NO_2, dan O_3, yang secara ilmiah diketahui memiliki pengaruh langsung terhadap nilai ISPU. Pemilihan fitur dilakukan untuk mengurangi kompleksitas model serta meningkatkan efisiensi dan akurasi prediksi.

Distribusi awal kelas pada dataset menunjukkan ketidakseimbangan yang cukup signifikan, di mana kategori SEDANG memiliki jumlah data terbanyak sebanyak 1.447 sampel, sementara kategori TIDAK SEHAT berjumlah 178 sampel, dan kategori BAIK hanya 168 sampel. Ketimpangan ini menyebabkan model cenderung bias terhadap kelas mayoritas dan sulit mengenali kelas minoritas dengan baik. Oleh karena itu, dilakukan proses penyeimbangan data (*class balancing*) menggunakan teknik *Random Oversampling* (ROS).

Secara teoretis, *Random Oversampling* merupakan metode resampling sederhana yang bekerja dengan cara menduplikasi sampel dari kelas minoritas secara acak hingga proporsinya seimbang dengan kelas mayoritas. Pendekatan ini bertujuan untuk memberikan bobot pembelajaran yang setara antar kelas sehingga model tidak bias terhadap kelas dominan. Berbeda dengan metode sintetis seperti SMOTE atau ADASYN yang menghasilkan data baru berdasarkan interpolasi antar titik, ROS mempertahankan struktur asli data sehingga risiko distorsi terhadap distribusi fitur relatif lebih kecil [15]. Dengan penerapan ROS pada penelitian ini, jumlah setiap kelas berhasil diseimbangkan menjadi 1.034 sampel per kelas, sehingga data yang digunakan dalam proses pelatihan menjadi representatif untuk seluruh kategori ISPU.

Pada tahap akhir, dataset dibagi menjadi dua bagian, yaitu data latih sebesar 70% dan data uji sebesar 30%, dengan menggunakan metode *stratified splitting* agar proporsi setiap kelas tetap terjaga pada kedua subset. Melalui proses ini, tahap data preparation menghasilkan dataset yang lebih bersih, memiliki distribusi kelas yang seimbang, dan siap digunakan dalam proses pelatihan serta evaluasi model klasifikasi ISPU.

E. Machine Learning Modelling

Pemodelan *machine learning* dilakukan menggunakan dataset yang telah melalui tahap *data preparation*, meliputi proses pembersihan data, penanganan nilai hilang, penyeimbangan kelas, dan pemilihan fitur yang relevan. Pada tahap ini, algoritma *Random Forest* diterapkan sebagai metode utama untuk melakukan klasifikasi. *Random Forest* merupakan bagian dari pendekatan *ensemble learning* yang membangun sejumlah *decision tree* dari sampel data yang diambil secara acak melalui teknik *bootstrapping*. Setiap pohon memberikan prediksi, dan keputusan akhir diperoleh melalui mekanisme voting mayoritas, yang membuat model lebih stabil serta mengurangi risiko *overfitting*.

Dalam proses pembentukan pohon, algoritma *Random Forest* menggunakan teknik pemilihan atribut terbaik di setiap percabangan pohon. Pemilihan ini dilakukan dengan menghitung nilai *information gain*, yaitu selisih antara nilai *entropy* sebelum dan sesudah suatu atribut digunakan untuk memisahkan data. *Entropy* sendiri merupakan ukuran ketidakpastian atau *impurity* dalam suatu himpunan data semakin kecil nilai *entropy*, maka

semakin homogen data pada node tersebut. Penggunaan *information gain* memastikan bahwa atribut yang memberikan pemisahan data paling informatif akan dipilih terlebih dahulu dalam pembentukan struktur pohon. Secara matematis, konsep ini dapat dijelaskan melalui rumus entropy dan information gain yang digunakan untuk mengukur efektivitas pemisahan kelas pada setiap node pohon keputusan. Secara matematis, konsep ini dapat dijelaskan melalui Persamaan (1) dan Persamaan (2) berikut [16] :

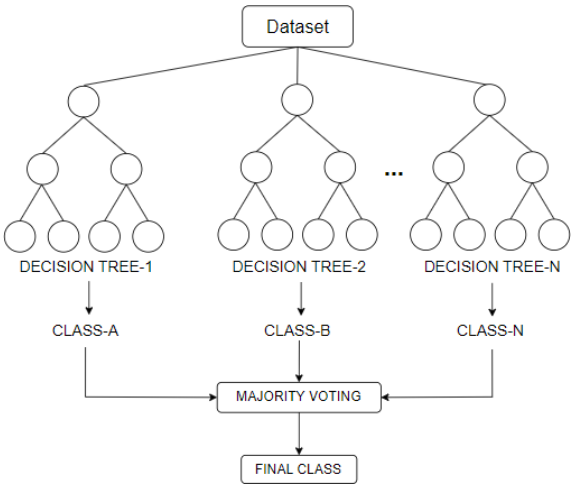
$$Entropy\ (s) = \sum_{i=1}^{n} -p_i log_2 p_i \quad (1)$$

Keterangan :
S : Himpunan dataset
n : Banyaknya jumlah kelas
pi: Probabilitas kelas ke-i dalam output S

$$Information\ Gain\ (A) = Entropy\ (S) = \sum_{i=1}^{n} \frac{S_i}{S} x\ Entropy\ (S_i) \quad (2)$$

Keterangan
A : Atribut
S : Himpunan dataset
|Si| : Jumlah sampel untuk nilai ke-i
|S| : Banyaknya jumlah data

Berdasarkan ilustrasi alur kerja Random Forest, setiap pohon keputusan dibangun secara terpisah menggunakan subset data yang berbeda, kemudian hasil prediksi dari masing-masing pohon digabungkan melalui mekanisme voting mayoritas untuk memperoleh keluaran akhir. Pendekatan ini membuat model menjadi lebih general dan tidak mudah terfokus pada pola tertentu yang dapat menyebabkan overfitting. Selain itu, metode ini membantu model mengenali pola pada setiap kategori ISPU dengan tingkat akurasi yang tinggi serta memberikan konsistensi performa antar kelas. Visualisasi mengenai proses kerja algoritma Random Forest ditunjukkan pada Gambar 2 [17].



Gambar 2. Visualisasi algoritma RF

F. Machine Learning Model Evaluation

Kualitas dan konsistensi prediksi model divalidasi melalui serangkaian metrik evaluasi yang komprehensif, mencakup *Accuracy*, *Precision*, *Recall*, *F1-Score*, serta *Confusion Matrix*. Sebagai instrumen diagnostik utama, *Confusion Matrix* memetakan perbandingan antara label aktual dan hasil prediksi ke dalam format tabel yang terdiri dari empat parameter fundamental: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Melalui representasi ini, peneliti dapat melakukan bedah kasus terhadap pola galat serta mengidentifikasi kategori yang memiliki tingkat misklasifikasi paling tinggi, sebagaimana diperlihatkan pada Tabel 2 *[18]*.

TABEL 2
CONFUSION MATRIX

Prediksi	Aktual	
	Positif	**Negatif**
Positif	TP (True Positif)	FP (False Negatif)
Negatif	FN (False Negatif)	TN (True Negatif)

Berdasarkan Tabel 2 dapat dijabarkan bahwa:

1) *True Positif*(TP), merujuk pada jumlah data yang bernilai positif dan telah diklasifikasikan dengan benar
2) *True Negatif* (TN), mengacu pada data yang bernilai negatif dan telah diklasifikasikan dengan benar
3) *False Negatif* (FN), merujuk pada data yang bernilai negatif tetapi diklasifikasikan sebagai positif oleh system
4) *False Positif* (FP), mengacu pada data yang bernilai positif tetapi diklasifikasikan secara salah oleh system

Melalui *confusion matrix*, dapat dilakukan analisis mendalam terhadap kekuatan dan kelemahan model dalam mengklasifikasikan setiap kelas. Nilai *precision* menunjukkan ketepatan prediksi, *recall* menggambarkan sensitivitas model terhadap kelas tertentu, sementara *f1-score* merupakan harmonisasi antara *precision* dan *recall* [19].

1. Akurasi

Akurasi digunakan untuk mengukur seberapa besar proporsi prediksi yang tepat yang dihasilkan model dibandingkan dengan seluruh prediksi yang dibuat. Nilai akurasi dihitung menggunakan persamaan berikut:

$$Accuracy = \frac{TP+TN}{TP+\ TN+FP+FN} \quad (1)$$

2. Presisi

Presisi digunakan untuk menilai sejauh mana model mampu memberikan prediksi positif yang benar, yaitu dengan membandingkan jumlah true positive terhadap seluruh prediksi yang diklasifikasikan sebagai positif. Rumus presisi dituliskan sebagai berikut:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

3. Recall

Recall berperan dalam mengevaluasi sensitivitas model dalam mengidentifikasi seluruh instansi yang secara faktual tergolong dalam kategori positif. Dengan kata lain, metrik ini menjadi tolok ukur sejauh mana algoritma mampu meminimalisir data positif yang terlewatkan atau salah diklasifikasikan sebagai negatif (false negative). Rumus recall dituliskan sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

4. F1-Score

F1-Score merupakan rata-rata harmonik yang mensinergikan metrik precision dan recall secara simultan. Parameter ini berfungsi untuk menyajikan representasi kinerja model yang lebih objektif dan seimbang, sehingga peneliti dapat memperoleh gambaran utuh mengenai efektivitas algoritma dalam mengklasifikasikan kelas positif. Adapun formulasi matematis untuk mengalkulasi F1-Score dijabarkan sebagai berikut:

$$f1\ score = 2 * \frac{precision*recall}{precision+recall} \quad (4)$$

G. Deployment

Deployment dilakukan dengan membangun arsitektur web yang mengintegrasikan model *machine learning* ke dalam sebuah sistem aplikasi yang dapat diakses secara langsung oleh pengguna. Arsitektur tersebut terdiri dari backend berbasis Flask dan frontend menggunakan *Next.js*. Model klasifikasi kualitas udara yang telah dilatih sebelumnya dengan algoritma *Random Forest* disimpan dalam format .pkl menggunakan *library Pickle*. File model ini kemudian di-load ke dalam aplikasi *Flask* yang bertugas sebagai penyedia layanan *API*. Melalui antarmuka web yang dirancang dengan *Next.js*, pengguna dapat memasukkan nilai-nilai parameter pencemarn udara, seperti PM10, PM2.5, SO_2, NO_2, dan O_3, secara manual. Nilai-nilai ini dikirimkan ke *backend* melalui permintaan *HTTP* (*API request*), lalu diproses oleh model klasifikasi untuk menghasilkan prediksi kategori ISPU (Indeks Standar Pencemaran Udara). Hasil prediksi tersebut dikembalikan dalam bentuk respons dan langsung ditampilkan secara interaktif melalui tampilan web. Pendekatan ini berhasil mengintegrasikan

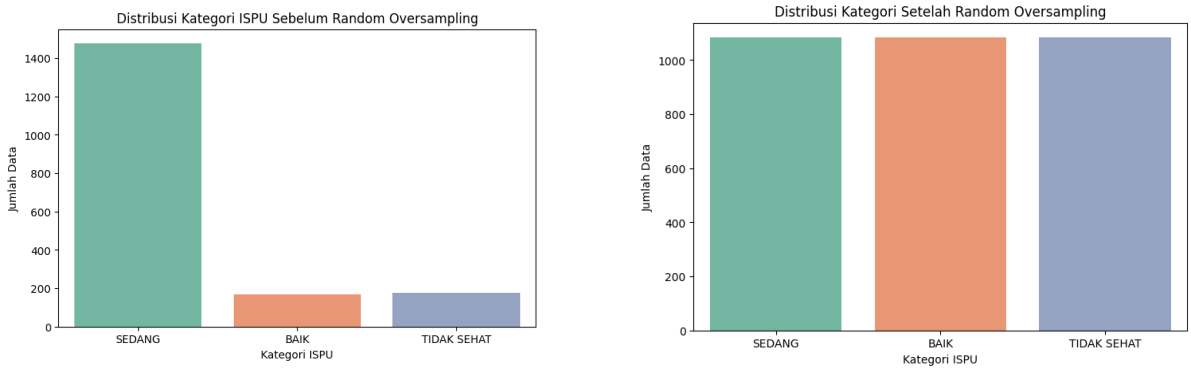
kecepatan akses informasi dengan akurasi klasifikasi data untuk kepentingan pengguna. Lebih jauh lagi, pengembangan ini mencerminkan potensi besar penggunaan model cerdas dalam mendukung sistem monitoring lingkungan yang bersifat proaktif dan digital. Visualisasi sistematis dari arsitektur yang diusulkan disajikan dalam Gambar 3.

Gambar 3. Arsitektur Sistem berbasis Next js dan Flask

III. HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini bersumber dari portal resmi Satu Data Jakarta, yang menyediakan data pemantauan kualitas udara harian untuk wilayah Provinsi DKI Jakarta sepanjang tahun 2024. Dataset tersebut terdiri atas 1.830 baris data dengan 13 kolom awal yang memuat berbagai parameter pencemar seperti PM10, PM2.5, CO, NO_2, SO_2, dan O_3, serta informasi pendukung berupa tanggal, waktu, dan lokasi pengukuran. Beberapa atribut, seperti timestamp dan lokasi, dihilangkan karena tidak memberikan kontribusi langsung terhadap proses klasifikasi. Setelah tahap pembersihan data, analisis difokuskan pada enam parameter utama pencemar udara dan satu variabel target berupa kategori ISPU.

Distribusi kategori ISPU sebelum dilakukan proses *oversampling* menunjukkan ketidakseimbangan yang cukup tinggi, di mana kategori SEDANG memiliki 1.477 data, sedangkan TIDAK SEHAT hanya 178, dan BAIK 168 data. Ketimpangan ini berpotensi membuat model *machine learning* bias terhadap kelas mayoritas. Untuk mengatasi hal tersebut, dilakukan teknik *Random Oversampling* (ROS) pada data latih, yang menyeimbangkan distribusi setiap kelas menjadi 1.034 sampel per kategori. Perbandingan distribusi kelas sebelum dan sesudah ROS ditunjukkan pada Gambar 4, yang memperlihatkan bahwa penerapan ROS berhasil menghilangkan dominasi kelas mayoritas dan memberikan proporsi yang seimbang antar kelas.



Gambar 4. Ilustrasi perbedaan distribusi kelas (a) distribusi kelas sebelum penerapan metode ROS dan (b) distribusi kelas setelah penerapan metode ROS

Pada percobaan pertama, model dikembangkan menggunakan algoritma *Random Forest* tanpa menerapkan teknik *Random Oversampling* (ROS). Hasil evaluasi menunjukkan bahwa model ini mampu menghasilkan akurasi sebesar 96.71% dengan nilai *macro average F1-score* sebesar 0.93. Namun demikian, performa model dalam mengklasifikasikan kelas minoritas masih kurang optimal. Sebagai contoh, pada kategori BAIK, nilai recall hanya mencapai 0.84, yang mengindikasikan bahwa model masih cenderung bias terhadap kelas mayoritas (SEDANG). Hasil evaluasi lengkap model tanpa *oversampling* dapat dilihat pada Tabel 3.

TABEL 3
HASIL PERFORRMA RANDOM FOREST TANPA RANDOM OVERSAMPLING

Kategori	Precision	Recall	F1-Score	Support
BAIK	0.88	0.84	0.86	50
SEDANG	0.97	0.99	0.98	443
TIDAK SEHAT	1.00	0.93	0.96	54
Accuracy	-	-	0.97	547
Macro Avg	0.95	0.92	0.93	547
Weighted Avg	0.97	0.97	0.97	547

Setelah dilakukan *Random Oversampling* (ROS) pada data latih, performa model *Random Forest* menunjukkan adanya peningkatan meskipun relatif kecil. Akurasi model naik dari 96.71% menjadi 97.07%, dengan nilai macro average F1-score meningkat dari 0.93 menjadi 0.94. Selain itu, recall untuk kelas TIDAK SEHAT meningkat dari 0.93 menjadi 0.94, sementara kelas BAIK tetap berada pada nilai recall 0.86. Hasil lengkap model setelah penerapan ROS ditampilkan pada Tabel 4.

TABEL 4
HASIL PERFORMA RANDOM FOREST DENGAN RANDOM OVERSAMPLING

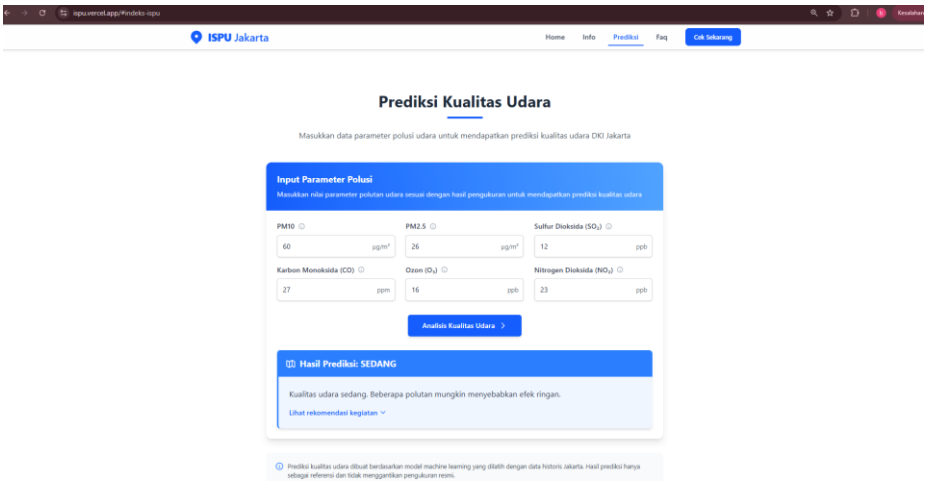
Kategori	Precision	Recall	F1-Score	Support
BAIK	0.88	0.86	0.87	50
SEDANG	0.98	0.99	0.98	443
TIDAK SEHAT	1.00	9.4	0.97	54
Accuracy	-	-	0.97	547
Macro Avg	0.95	0.93	0.94	547
Weighted Avg	0.97	0.97	0.97	547

Peningkatan yang diperoleh setelah penerapan ROS tergolong moderat, namun tetap positif secara metodologis. Hal ini disebabkan oleh dua faktor utama. Pertama, model *RandSom Forest* secara alami sudah cukup kuat terhadap ketidakseimbangan data karena sifatnya yang berbasis *bootstrap aggregation* (bagging), sehingga dampak dari penyeimbangan data tidak terlalu besar. Kedua, karakteristik dataset ISPU tahun 2024 menunjukkan distribusi polutan yang relatif seragam antar kelas, sehingga batas pemisahan antar kategori (*decision boundary*) tidak banyak berubah setelah dilakukan *oversampling*.

Dengan demikian, meskipun *Random Oversampling* tidak memberikan peningkatan drastis pada metrik performa, metode ini tetap berkontribusi dalam meningkatkan generalisasi model dan mengurangi potensi bias terhadap kelas mayoritas, sehingga hasil prediksi lebih adil dan seimbang antar kategori kualitas udara.

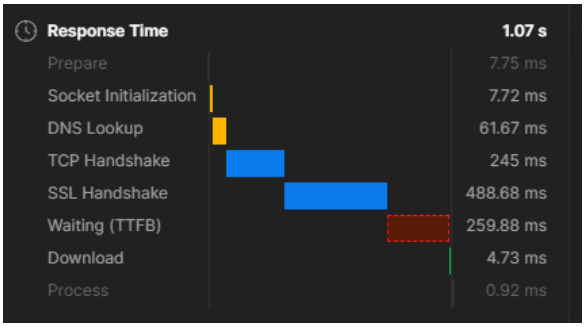
Setelah model *Random Forest* dengan performa terbaik diperoleh melalui penerapan teknik *Random Oversampling* (ROS), langkah selanjutnya adalah melakukan *deployment* agar model dapat digunakan secara praktis oleh pengguna. *Deployment* bertujuan untuk mengintegrasikan model ke dalam sebuah aplikasi berbasis web yang mudah diakses, sehingga hasil klasifikasi kualitas udara berdasarkan data ISPU dapat ditampilkan secara *real-time*. Dalam penelitian ini, proses *deployment* dilakukan menggunakan *framework Flask* pada sisi *backend* dan *Next.js* pada sisi *frontend*. Model yang telah dilatih disimpan dalam format .pkl (Pickle) dan di-load ke dalam API Flask yang bertugas memproses permintaan prediksi dari pengguna. Melalui antarmuka web, pengguna dapat memasukkan nilai parameter pencemar seperti PM10, PM2.5, SO_2, NO_2, dan O_3, kemudian sistem akan memberikan hasil prediksi kategori ISPU secara instan.

Tampilan antarmuka website dari sistem klasifikasi ini dapat dilihat pada Gambar 5, yang menunjukkan form input parameter kualitas udara dan hasil prediksi kategori ISPU yang ditampilkan secara interaktif melalui halaman web.



Gambar 5. Tampilan Antarmuka Website Sistem Prediksi ISPU

Untuk memastikan sistem berjalan secara efisien dan adaptif, dilakukan evaluasi performa API menggunakan Postman dengan pengujian kecepatan respons terhadap permintaan prediksi. Hasil pengujian menunjukkan waktu respons total sebesar 1.07 detik, dengan rincian tahapan proses sebagaimana ditunjukkan pada Gambar 6. Nilai *Time to First Byte* (TTFB) yang relatif kecil menunjukkan bahwa sistem mampu memberikan hasil prediksi dalam waktu kurang dari satu detik sejak permintaan dikirimkan.



Gambar 6. Visualisasi Waktu Respons API Hasil Pengujian Menggunakan Postman

IV. SIMPULAN

Dalam penelitian ini, kualitas udara di DKI Jakarta diklasifikasikan berdasarkan data ISPU tanpa menuliskan istilah Indeks Standar Pencemar Udara (ISPU) secara langsung. Proses pemodelan dilakukan dengan memanfaatkan algoritma Random Forest yang dipadukan dengan teknik penyeimbangan data melalui Random Oversampling. Sumber datanya berasal dari platform Satu Data Jakarta, kemudian setelah tahap pra-proses selesai, seluruh entri tersebut dipisahkan ke dalam tiga kelas utama: kategori baik, kategori sedang, serta kategori tidak sehat.

Pada percobaan awal tanpa penerapan ROS, model *Random Forest* menghasilkan akurasi sebesar 96.71% dengan nilai macro average F1-score 0.93, namun masih menunjukkan kecenderungan bias terhadap kelas mayoritas, khususnya pada kategori BAIK dengan nilai recall sebesar 0.88. Hal ini menandakan bahwa model cenderung lebih akurat dalam mengenali kelas dominan dibandingkan kelas minoritas.

Setelah dilakukan penerapan *Random Oversampling* pada data latih, performa model meningkat menjadi akurasi 97.07% dengan *macro average F1-score* 0.94. Meskipun peningkatannya hanya sebesar 0.3%, teknik ini tetap membantu memperbaiki proporsi data antar kelas, sehingga model dapat belajar lebih seimbang. Namun, recall pada kelas BAIK tetap konstan di angka 0.88, yang menunjukkan bahwa ROS belum sepenuhnya efektif dalam meningkatkan sensitivitas terhadap kelas minoritas. Hal ini dapat disebabkan oleh distribusi data yang masih memiliki noise, atau karena karakteristik kelas BAIK yang secara alami memiliki variasi nilai polutan yang lebih kecil, sehingga sulit dibedakan oleh model meskipun data sudah diseimbangkan.

Keterbatasan lain dari ROS adalah kemampuannya yang hanya menyalin data minoritas tanpa menghasilkan variasi baru, sehingga model berpotensi mengalami overfitting apabila proporsi data hasil duplikasi terlalu besar. Oleh karena itu, metode balancing lain seperti *Synthetic Minority Over-sampling Technique* (SMOTE) atau *Adaptive Synthetic Sampling* (ADASYN) dapat menjadi alternatif yang potensial untuk penelitian selanjutnya karena keduanya mampu menghasilkan sampel sintetis yang lebih representatif daripada sekadar duplikasi data.

Kontribusi utama dari penelitian ini terletak pada evaluasi empiris penerapan *Random Oversampling* pada algoritma *Random Forest* untuk klasifikasi ISPU tahun 2024 di DKI Jakarta, serta implementasi model ke dalam sistem web interaktif berbasis *Flask* dan *Next.js* yang memungkinkan pengguna melakukan prediksi kualitas udara secara langsung. Pendekatan ini menunjukkan bahwa integrasi *machine learning* dengan sistem berbasis web dapat mendukung pemantauan kualitas udara yang lebih adaptif, efisien, dan dapat dimanfaatkan sebagai dasar penyusunan kebijakan berbasis data dalam mitigasi pencemaran udara perkotaan.

DAFTAR PUSTAKA

[1] 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.

[2] Rosatul Umah and Eva 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.

[3] A. F. B. Sajiwo, B. Rahmat, and A. Junaidi, "Klasifikasi Indeks Standar Pencemaran Udaran (Ispu) Menggunakan Algoritma Xgboost Dengan Teknik Imbalanced Data (Smote)," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4699.

[4] J. S. Dyana *et al.*, "Dampak Bahaya Pencemaran Udara Terhadap Kesehatan Masyarakat Di Perkotaan," vol. 11, pp. 132–140, 2025.

[5] F. D. C.W., R. Emilia, G. G. P., and F. Indrayatna, "Klasifikasi Tingkat Pencemaran Udara Kota Jakarta Tahun 2021 Menggunakan Algoritma Decision Tree," *Pros. Nas. SNSA 2*, pp. 127–131, 2023, [Online]. Available: https://www.data.jakarta.go.id/

[6] I. Irwansyah, A. D. Wiranata, and T. T. M, "Komparasi Algoritma Decision Tree, Naive Bayes Dan K-Nearest Neighbor Untuk Menentukan Kualitas Udara Di Provinsi Dki Jakarta," *Infotech J. Technol. Inf.*, vol. 9, no. 2, pp. 193–198, 2023, doi: 10.37365/jti.v9i2.203.

[7] I. M. Faiza, G. Gunawan, and W. Andriani, "Tinjauan Pustaka Sistematis: Penerapan Metode Machine Learning untuk Deteksi Bencana Banjir," *J. Minfo Polgan*, vol. 11, no. 2, pp. 59–63, 2022, doi: 10.33395/jmp.v11i2.11657.

[8] L. F. Syarifa, "Studi Komparasi Model Parametrik Dan Non-Parametrik Dalam Estimasi Efisiensi Teknis Perkebunan Karet Rakyat Di Sumatera Selatan," *J. Penelit. Karet*, vol. 38, no. 2, pp. 189–196, 2020, doi: 10.22302/ppk.jpk.v2i38.736.

[9] B. V. Jayadi, M. D. Lauro, Z. Rusdi, T. Handhayani, and F. T. Informasi, "Klasifikasi Indeks Standar Pencemaran Udara untuk Data Tidak Seimbang menggunakan Pendekatan Pembelajaran Mesin," *J. Sist. Inf.*, vol. 13, no. 3, pp. 951–958, 2024, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[10] D. Javale, P. Pillai, P. Patel, and S. Jagtap, "A Comparison of Oversampling and Undersampling Methods for Predicting Air Quality in Metropolitan Region," *Proc. - Int. Conf. Appl. Artif. Intell. Comput. ICAAIC 2022*, no. January, pp. 1615–1620, 2022, doi: 10.1109/ICAAIC53929.2022.9793084.

[11] A. H. Putra and A. Salam, "A Comparative Performance of SMOTE, ADASYN and Random Oversampling in Machine Learning Models on Prostate Cancer Dataset," *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 603–610, 2025, doi: 10.30871/jaic.v9i3.9308.

[12] W. Chandra, B. Suprihatin, and Y. Resti, "Median-KNN Regressor-SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction," *Symmetry (Basel).*, vol. 15, no. 4, 2023, doi: 10.3390/sym15040887.

[13] F. Yulian Pamuji, Ahmad Rofiqul Muslikh, Rizza Muhammad Arief, and Delviana Muti, "Komparasi Metode Mean dan KNN Imputation dalam Mengatasi Missing Value pada Dataset Kecil," *J. Inform. Polinema*, vol. 10, no. 2, pp. 257–264, 2024, doi: 10.33795/jip.v10i2.5031.

[14] G. Li, C. Wang, D. Zhang, and G. Yang, "An improved feature selection method based on random forest algorithm for wind turbine condition monitoring," *Sensors*, vol. 21, no. 16, pp. 1–11, 2021, doi: 10.3390/s21165654.

[15] N. A. Prakoso Indaryono, "Analisa Perbandingan Algoritma Random Forest Dan Naïve Bayes Untuk Klasifikasi Curah Hujan Berdasarkan Iklim Di Indonesia," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, no. 1, pp. 158–167, 2024, doi: 10.29100/jipi.v9i1.4421.

[16] F. Diba, M. S. Lydia, and P. Sihombing, "Analisis Random Forest Menggunakan Principal Component Analysis Pada Data Berdimensi Tinggi," *Indones. J. Comput. Sci.*, vol. 12, no. 4, pp. 2152–2160, 2023, doi: 10.33022/ijcs.v12i4.3329.

[17] M. A. N. Anargya, W. Ghozi, and F. A. Rafrastara, "Random Under Sampling for Performance Improvement in Attack Detection on Internet of Vehicles Using Machine Learning," *J. Inform. J. Pengemb. IT*, vol. 10, no. 1, pp. 11–19, 2025, doi: 10.30591/jpit.v10i1.8034.

[18] F. Hanifatun and L. Zahrotun, "Jurnal Informatika : Jurnal pengembangan IT Penerapan Data Mining Dalam Pemberian Kelayakan Kredit Nasabah Pada Badan Usaha Milik Desa Gedong Gincu Dengan Metode Naïve Bayes," vol. 10, no. 1, pp. 226–236, 2025, doi: 10.30591/jpit.v9ix.xxx.

[19] S. Syihabuddin Azmil Umri, "Prediksi Kualitas Udara Menggunakan Algoritma K-Nearest Neighbor," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 2, pp. 98–104, 2021, doi: 10.33387/jiko.v4i2.2871.