JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 12 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v12i3.4699

KLASIFIKASI INDEKS STANDAR PENCEMARAN UDARAN (ISPU) MENGGUNAKAN ALGORITMA XGBOOST DENGAN TEKNIK IMBALANCED DATA (SMOTE)

Achmad Fauzihan Bagus Sajiwo[1*], Basuki Rahmat[2], Achmad Junaidi[3]

[1,2,3]Universitas Pembangunan Nasional "Veteran" Jawa Timur, Jl. Rungkut Madya No.1, Gn. Anyar, Surabaya

Received: 30 Juni 2024
Accepted: 31 Juli 2024
Published: 7 Agustus 2024

Keywords:
ISPU, Klasifikasi, XGBoost, Imbalanced Data, SMOTE

Corespondent Email:
20081010069@student.upnjatim.ac.id

Abstrak. Polusi udara adalah masuknya zat-zat berbahaya ke atmosfer, yang dapat disebabkan oleh tindakan manusia, serta oleh peristiwa alam. Menurut Air Quality Live Index (AQLI) pada bulan April 2021, DKI Jakarta sebagai ibu kota negara, menempati posisi keenam di dunia dengan kota tingkat kualitas udara yang paling buruk. Untuk menghadapi masalah polusi udara yang terus memburuk, perlu diambil tindakan yang tepat, satu diantaranya adalah melakukan penelitian klasifikasi indeks standar pencemaran udara (ISPU). Penerapan klasifikasi ISPU membutuhkan metode yang dapat mengolah dan menganalisis pola data dari sensor-sensor yang mengukur tingkat polutan udara. Metode yang digunakan pada penelitian ini adalah eXtreme Gradient Boosting (XGBoost). Untuk membantu menyeimbangkan data, pada penelitian ini menggunakan Synthetic Minority Over-sampling Technique (SMOTE). Data yang digunakan adalah ISPU DKI Jakarta tahun 2022-2023. Hasil klasifikasi indeks standar pencemaran udara menggunakan algoritma XGBoost dengan teknik SMOTE, didapatkan akurasi sebesar 99.63%.

Abstract. *Air pollution is the introduction of harmful substances into the atmosphere, which can be caused by human actions, as well as by natural events. According to the Air Quality Live Index (AQLI) in April 2021, DKI Jakarta, as the nation's capital, is in sixth place in the world with the city with the worst air quality level. To face the problem of air pollution which continues to worsen, appropriate action needs to be taken, one of which is conducting research on the classification of the air pollution standard index (ISPU). Implementing the ISPU classification requires a method that can process and analyse data patterns from sensors that measure air pollutant levels. The method used in this research is eXtreme Gradient Boosting (XGBoost). To help balance the data, this research used Synthetic Minority Over-sampling Technique (SMOTE). The data used is the DKI Jakarta ISPU for 2022-2023. The results of the standard air pollution index classification using the XGBoost algorithm with the SMOTE technique, obtained an accuracy of 99.63%.*

1. PENDAHULUAN

Kesejahteraan makhluk hidup sangat dipengaruhi oleh adanya lingkungan yang sehat. Kualitas udara yang memenuhi standar kesehatan merupakan elemen penting dalam menciptakan lingkungan yang sehat. Udara yang kita hirup sangat penting bagi keberadaan kita, karena menyediakan unsur oksigen yang

penting bagi kita [1]. Polusi udara menimbulkan bahaya paling signifikan terhadap kesenjangan Lingkungan [2]. Polusi udara adalah masuknya berbagai zat secara berbahaya ke dalam atmosfer kita, baik yang disebabkan oleh tindakan manusia yang disengaja maupun tidak, atau yang timbul dari kejadian alam [3].

Menurut Organisasi Kesehatan Dunia (WHO), dampak buruk dari polusi udara mengakibatkan kematian dini pada 2 juta orang setiap tahunnya [4]. Hanya 5% negara yang mampu mematuhi pedoman polusi udara yang ditetapkan oleh WHO [5]. Indonesia menempati peringkat ke-17 secara global dalam hal tingkat polusi udara dan peringkat ke-1 secara wilayah Asia Tenggara [6]. Berdasarkan Air Quality Live Index (AQLI) pada bulan April 2021, DKI Jakarta sebagai ibukota negara menempati posisi kota ke-6 di dunia dengan kualitas udara paling buruk [7]. Meningkatnya aktivitas manusia telah memicu kekhawatiran yang mendesak terhadap masalah polusi udara [8]. Untuk menanggulangi masalah polusi udara yang hari demi hari terus memburuk, diperlukan adanya tindakan seperti menggunakan kendaraan umum sebagai alat transportasi sehari-hari, menerapkan konsep 3R yaitu reduce, reuse, recycle, memanfaatkan sumber energi yang berkelanjutan dan ramah lingkungan dan rutin mengecek kualitas standar udara [9]. Di Indonesia pengukuran standar kualitas udara yang resmi yaitu Indeks Standar Pencemar Udara (ISPU), hal ini sesuai dengan Keputusan Menteri Negara Lingkungan Hidup Nomor : KEP 45/MENLH/1997 Tentang Indeks Standar Pencemar Udara (ISPU) [10].

ISPU berfungsi sebagai tolok ukur utama untuk mengkategorikan dan menjelaskan kualitas udara yang didasarkan pada dampak besar terhadap kesejahteraan individu dan vitalitas semua makhluk hidup. Terdapat lima parameter dalam ISPU sebagai pengamatan kualitas udara seperti : Tingkat Partikulat (PM10), Oksida Nitrogen (NO2), Sulfur Dioksida (SO2), Karbon Monoksida (CO), dan ozon permukaan (O3) [11]. Jika senyawa-senyawa tersebut melampaui ambang batas yang dapat diterima, maka berpotensi membahayakan kesejahteraan seseorang, bahkan menyebabkan kematian, terutama pada sistem pernafasan [12]. ISPU dibagi menjadi lima kategori berbeda, yaitu baik, sedang, tidak sehat, sangat tidak sehat, dan berbahaya [13].

Penerapan klasifikasi ISPU memerlukan metode yang mampu menganalisis dan menguraikan pola data yang berasal dari sensor yang mengukur tingkat polutan udara. Metode umum yang sering digunakan adalah metode pembelajaran mesin (machine learning), sebuah bidang kecerdasan buatan yang memungkinkan komputer memperoleh pengetahuan dari data secara mandiri, sehingga tidak memerlukan pemrograman eksplisit [14]. Machine learning yang digunakan pada penelitian ini adalah eXtreme Gradient Boosting (XGBoost).

XGBoost adalah algoritma pembelajaran mesin yang terkenal digunakan dalam klasifikasi dan regresi. Algoritma ini mengadopsi pendekatan ensemble dari pohon keputusan, di mana sejumlah pohon keputusan dibangun secara berurutan untuk meningkatkan kinerja keseluruhan model. Teknik yang digunakan pada penelitian ini adalah Synthetic Minority Over-sampling Technique (SMOTE). SMOTE adalah teknik yang umum digunakan dalam pemrosesan data yang tidak seimbang (imbalanced data). Teknik ini bekerja dengan membuat sampel sintetis baru dalam kelas minoritas.

Pada tahun 2023 [15] menerapkan algoritma XGBoost untuk klasifikasi kualitas udara. Parameter yang digunakan dalam klasifikasi tersebut antara lain Particulate Matter 10 (PM10), Particulate Matter 2.5 (PM2.5), Sulfur Dioksida (SO2), Karbon Monoksida (CO), Ozon (O3) dan Nitrogen Dioksida (NO2). Dalam penelitian tersebut didapatkan nilai precision sebesar 97%, nilai recall sebesar 100%, nilai F-1 Score sebesar 98% dan nilai accuracy sebesar 98.61%.

Penelitian selanjutnya tentang model prediksi kepadatan lalu lintas : Perbandingan antara algoritma Random Forest dan XGBoost pernah dilakukan oleh [16]. Data yang digunakan dari Metro Interstate Traffic Volume, yang disediakan oleh repositori data UCI. Hasil accuracy yang didapatkan untuk algoritma XGBoost sebesar 95.92%, sedangkan algoritma Random Forest sebesar 95.53% dengan algoritma XGBoost memiliki waktu pemrosesan lebih cepat 532% dibanding algoritma Random Forest.

Selanjutnya penelitian tentang penerapan teknik SMOTE pada klasifikasi objektivitas

berita online dengan algoritma KNN [17]. Nilai k tetangga yang digunakan bervariasi yaitu 1, 3, 5, 7 dan 9. Dari hasil pengujian teknik SMOTE dapat meningkatkan nilai accuracy. Dengan nilai accuracy k=1 meningkat 5,00 dan untuk nilai k=3 meningkat 1,72. SMOTE meningkatkan nilai precision untuk semua nilai k tetangga. Untuk nilai recall dan F-measure SMOTE dapat menurunkan nilai keduanya untuk semua nilai k tetangga.

Berdasarakan penelitian sebelumnya, peneliti ingin melakukan perbandingan algoritma eXtreme Gradient Boosting dengan teknik imbalance data pada klasifikasi indeks standar pencemaran udara (ISPU) dengan Synthetic Minority Over-sampling Technique.

2. TINJAUAN PUSTAKA

2.1. Pencemaran Udara

Pencemaran udara adalah kondisi di mana terjadi keberadaan satu atau beberapa zat kimia, fisik, atau biologis dalam jumlah yang berpotensi membahayakan, yang dapat membahayakan kesehatan manusia, hewan, dan tumbuhan, mengganggu estetika dan kenyamanan, atau menyebabkan kerusakan pada property [18].

Sumber pencemaran udara dapat dibagi menjadi tiga kategori, yaitu sumber perkotaan dan industri, sumber pedesaan atau pertanian, Pencemaran udara di wilayah perkotaan dan industri disebabkan oleh kemajuan teknologi yang menghasilkan banyak pabrik industri, pembangkit listrik, dan kendaraan bermotor. dan sumber alami. Menurut UU No. 32 Tahun 2009 tentang Perlindungan dan Pengelolaan Lingkungan Hidup, tindakan untuk mengendalikan pencemaran udara meliputi langkah-langkah pencegahan, penanggulangan, dan pemulihan kualitas udara. Penyelenggaraan pengendalian pencemaran dilakukan oleh pemerintah pusat, pemerintah daerah, dan pelaku usaha atau kegiatan, sesuai dengan kewenangan, peran, dan tanggung jawab mereka masing-masing [19].

2.2. Indeks Standar Pencemaran Udara

Indeks Standar Pencemar Udara (ISPU) merupakan suatu parameter yang digunakan untuk mengukur kualitas udara dengan mengindikasikan tingkat pencemaran udara yang disebabkan oleh zat kimia dan partikel yang ada di udara [20]. Indeks Standar Pencemaran Udara (ISPU) merupakan salah satu cara untuk mengukur seberapa tercemarnya udara. Penghitungannya menggunakan metode tertentu yang telah ditetapkan dalam Keputusan Kepala Badan Pengendalian Dampak Lingkungan Nomor KEP-107/KABAPEDAL/11/1997 [21].

2.3. Klasifikasi

Klasifikasi adalah proses mendasar yang melibatkan identifikasi dan kategorisasi kelas atau konsep data. Tujuan utama klasifikasi adalah untuk mengembangkan model atau fungsi yang dapat membedakan kelas-kelas tersebut secara akurat untuk memprediksi klasifikasi objek yang tidak diketahui. Dengan memahami pola dan hubungan yang ada dalam data, klasifikasi memungkinkan kita membuat keputusan dan prediksi berdasarkan hasil atau observasi di masa depan [22].

2.4. eXtreme Gradient Boosting

XGBoost merupakan perkembangan dari metode gradient boosting yang diusulkan oleh Dr. Tianqi Chen dari University of Washington pada tahun 2014. Gradien boosting adalah teknik serbaguna yang memiliki kemampuan untuk secara efektif mengatasi beragam masalah seperti regresi, klasifikasi, dan pemeringkatan. Prinsip dasarnya melibatkan penyesuaian parameter pembelajaran secara berulang untuk meminimalkan fungsi kerugian secara bertahap, yang berfungsi sebagai metrik untuk mengevaluasi performa model [23]. XGBoost berupaya mengatasi masalah overfitting dan meningkatkan efisiensi komputasi dengan menyederhanakan fungsi tujuan. Dengan menggabungkan istilah prediksi dan regularisasi, XGBoost bertujuan untuk mencapai keseimbangan antara mengendalikan kompleksitas model dan mencegah overfitting, sambil mempertahankan kecepatan komputasi yang optimal [24].

2.5. Imbalance Data

Ketidakseimbangan data terjadi ketika salah satu kelas memiliki jumlah yang jauh lebih banyak dibandingkan dengan kelas lainnya, yang mengakibatkan penurunan kinerja klasifikasi pada kelas yang lebih sedikit jumlahnya [25]. Ini menjadi masalah dalam klasifikasi karena algoritma cenderung lebih

sering memprediksi kelas dengan data yang lebih banyak (mayoritas) dan mengabaikan kelas dengan data yang lebih sedikit (minoritas). Akibatnya, akurasi prediksi untuk kelas mayoritas akan tinggi, sedangkan untuk kelas minoritas akan rendah

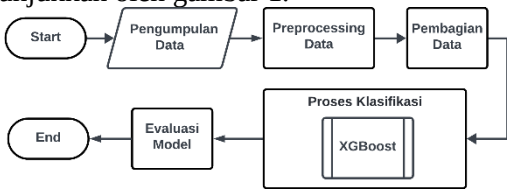
2.6. Synthetic Minority Over-sampling Techniqueasis

Teknik Pengambilan Sampel Minoritas Sintetis (SMOTE) adalah algoritma efektif yang dirancang untuk mengatasi masalah ketidakseimbangan data dalam tugas klasifikasi. Dengan meningkatkan representasi kelas minoritas secara artifisial melalui proses oversampling, SMOTE berpotensi meningkatkan kinerja model klasifikasi secara signifikan.

Prosedur formalnya dimulai dengan menetapkan parameter *oversampling* N, yang dapat disesuaikan untuk mencapai distribusi kelas yang seimbang atau ditentukan melalui proses pembungkusan. Prosedur ini melibatkan beberapa langkah yang dilakukan secara iteratif. Pertama, *instance* dari kelas minoritas dipilih secara acak dari dataset pelatihan. Kemudian, K tetangga terdekat (biasanya 5) diidentifikasi. Terakhir, N *instance* dipilih secara acak dari K tetangga ini untuk membuat instance baru melalui interpolasi.

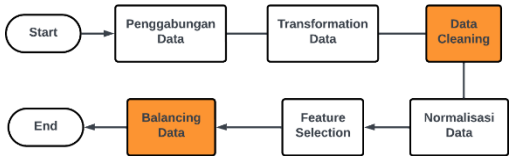
3. METODE PENELITIAN

Metodologi mencakup kumpulan teknik, prinsip, dan prosedur komprehensif yang digunakan dalam bidang atau disiplin ilmu tertentu untuk memfasilitasi penelitian yang cermat dan sistematis. Metodologi penelitian ini ditunjukkan oleh gambar 1.



Gambar 1. Diagram Alur Penelitian

Pada diagram alur penelitian dapat dilihat, penelitian ini dimulai dari pengumpulan data. Selanjutnya yaitu *preprocessing* data. Pada *preprocessing* data, terdapat beberapa tahapan untuk membuat data siap untuk dilakukan klasifikasi.



Gambar 2. Diagram Alur Preprocessing Data

Setelah *preprocessing* data, data dibagi menjadi data *train* dan data *test*. Lalu masuk ke proses klasifikasi. Terakhir adalah evaluasi model.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data Indeks Standar Pencemaran Udara (ISPU) DKI Jakarta tahun 2021 sampai tahun 2023. Dataset didapat dari website Satu Data Jakarta yang beralamatkan (url : https://satudata.jakarta.go.id/home). Dari data 3 tahun tersebut terkumpul total dataset berjumlah 4015 data. Data tersebut diambil dari stasiun pemantauan kualitas udara (SPKU) yang tersebar di lima titik.

periode	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	kategori	stasiun
202101	2021-01-01	38	53	29	6	31	13	53	PM25	Sedang	DKI1
202101	2021-01-01	38	58	2	11	65	6	65	O3	Sedang	DKI2
202101	2021-01-01	43	56	15	10	33	5	56	PM25	Sedang	DKI3
202101	2021-01-01	41	-	37	14	35	4	41	PM25	Sedang	DKI4
202101	2021-01-01	37	-	20	12	25	4	37	PM10	Baik	DKI5
---	---	---	---	---	---	---	---	---	---	---	---

Gambar 3. Data ISPU

Data ISPU tahun 2021 berjumlah 1825 data dengan kategori sedang berjumlah 1349 data, kategori tidak sehat berjumlah 272 data dan kategori baik berjumlah 188 data. Data ISPU tahun 2022 berjumlah 365 data dengan kategori sedang berjumlah 225 data, kategori tidak sehat berjumlah 137 data dan kategori baik berjumlah 3 data. Data ISPU tahun 2023 berjumlah 1825 data dengan kategori sedang berjumlah 1358 data, kategori tidak sehat berjumlah 207 data dan kategori baik berjumlah 236 data.

4.2. Preprocessing Data

4.2.1. Penggabungan Data

Tahap proses preprocessing data pada penelitian ini dimulai dari penggabungan data.

Data yang digunakan adalah data indeks standar pencemaran udara (ISPU) DKI Jakarta tahun 2021 – 2023 dalam bentuk per satu tahun. Untuk mendapatkan data yang mencakup semuanya, data digabungkan menjadi satu.

periode	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	kategori	stasiun
202101	2021-01-01	43	56	15	10	33	5	56	PM25	Sedang	DKI3
202101	2021-01-01	41	-	37	14	35	4	41	PM25	Sedang	DKI4
202101	2021-01-01	37	-	20	12	25	4	37	PM10	Baik	DKI5
---	---	---	---	---	---	---	---	---	---	---	---
202209	2022-13-09	59	81	52	18	35	27	79	PM25	Sedang	DKI1
202202	2022-28-02	27	50	40	12	30	11	50	PM25	Baik	DKI5
202201	2022-20-01	40	57	55	14	68	24	68	O3	Sedang	DKI2
202310	2023-07-10	72	107	46	5	30	38	107	PM25	Tidak Sehat	DKI1
202304	2023-11-04	24	32	22	6	19	13	32	PM25	Baik	DKI2
202301	2023-13-01	-	135	55	15	28	15	135	PM25	Tidak Sehat	DKI4

Gambar 4. Hasil Penggabungan Data ISPU

4.2.2. Transformation Data

Pada tahap *transformation* data, nilai dalam parameter kategori yang awalnya bertipe data string diubah menjadi tipe data numerik. Pada penelitian ini parameter kategori adalah label kategorikal.

periode	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	kategori	stasiun
202101	2021-01-01	43	56	15	10	33	5	56	PM25	1	DKI3
202202	2022-28-02	27	50	40	12	30	11	50	PM25	0	DKI5
202202	2022-15-02	74	119	52	33	61	38	119	PM25	2	DKI3
202304	2023-01-04	31	40	22	6	24	14	40	PM25	0	DKI2

Gambar 5. Hasil Transformation Data

Parameter kategori yang sebelumnya berisi sedang, tidak sehat dan baik berubah menjadi 0, 1 dan 2.

4.2.3. Data Cleaning

Pada tahap data cleaning dilakukan pembersihan data dan penghapusan parameter yang tidak digunakan dalam klasifikasi ini. Terdapat data yang bernilai "-", yang menjadi fokus utama dalam pembersihan data

periode	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	kategori	stasiun
202101	2021-01-01	43	56	15	10	33	5	56	PM25	1	DKI3
202101	2021-01-01	37	-	20	12	25	4	37	PM10	0	DKI5
202101	2021-01-01	41	-	37	14	35	4	41	PM25	1	DKI4
202201	2022-20-01	40	57	55	14	68	24	68	O3	1	DKI2
202202	2020-06-02	39	51	58	15	57	21	58	SO2	1	DKI1
202301	2023-13-01	-	135	55	15	28	15	135	PM25	2	DKI4
202311	2023-06-11	67	92	30	43	-	19	92	PM25	1	DKI4
202311	2023-16-11	-	82	-	-	-	-	82	PM25	1	DKI2

Gambar 6. Data ISPU Kotor

Pada baris periode 202101 kolom pm2.5 terdapat data yang bernilai "-". Begitu juga pada baris 202311 terdapat data yang bernilai "-" pada kolom pm10, so2, co, o3 dan no2. Data yang memiliki nilai "-'' tersebut diubah menjadi NaN untuk memudahkan dalam perhitungan atau pengecekan data yang tidak memiliki nilai (*missing value*). Peneliti melakukan perbandingan dalam mengatasi *missing value*, yaitu dengan menghapus data yang memiliki *missing value*, mengisi data yang memiliki *missing value* dengan mean dan median pada masing-masing parameter tersebut.

periode	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	kategori	stasiun
202101	2021-01-01	43	56	15	10	33	5	56	PM25	1	DKI3
202101	2021-01-01	37	79	20	12	25	4	37	PM10	0	DKI5
202101	2021-01-01	41	79	37	14	35	4	41	PM25	1	DKI4
202201	2022-20-01	40	57	55	14	68	24	68	O3	1	DKI2
202202	2020-06-02	39	51	58	15	57	21	58	SO2	1	DKI1
202301	2023-13-01	55	135	55	15	28	15	135	PM25	2	DKI4
202311	2023-06-11	67	92	30	43	28	19	92	PM25	1	DKI4
202311	2023-16-11	55	82	39	12	28	18	82	PM25	1	DKI2

Gambar 7. Missing Value diisi Median

Pada gambar 7, dapat dilihat tidak ada data yang bernilai "-" karena data yang bernilai "-" telah diisi dengan median dari setiap parameter masing-masing. Parameter pm10 memiliki nilai median 55, parameter pm2.5 memiliki nilai median 79, parameter so2 memiliki nilai median 39, parameter co memiliki nilai median 12, parameter o3 memiliki nilai median 28 dan parameter no2 memiliki nilai median 18.

periode	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	kategori	stasiun
202101	2021-01-01	43	56	15	10	33	5	56	PM25	1	DKI3
202101	2021-01-01	37	79	20	12	25	4	37	PM10	0	DKI5
202101	2021-01-01	41	79	37	14	35	4	41	PM25	1	DKI4
202201	2022-20-01	40	57	55	14	68	24	68	O3	1	DKI2
202202	2020-06-02	39	51	58	15	57	21	58	SO2	1	DKI1
202301	2023-13-01	53	135	55	15	28	15	135	PM25	2	DKI4
202311	2023-06-11	67	92	30	43	32	19	92	PM25	1	DKI4
202311	2023-16-11	53	82	37	12	32	19	82	PM25	1	DKI2

Gambar 8. Missing Value diisi Mean

Pada gambar 8, dapat dilihat tidak ada data yang bernilai "-" karena data yang bernilai "-" telah diisi dengan mean dari setiap parameter masing-masing. Parameter pm10 memiliki nilai mean 53, parameter pm2.5 memiliki nilai mean 79, parameter so2 memiliki nilai mean 37, parameter co memiliki nilai mean 12, parameter o3 memiliki nilai mean 32 dan parameter no2 memiliki nilai mean 19.

Setelah mengatasi data *missing value*, peneliti menghapus variable atau kolom periode data, tanggal, max, critical dan stasiun. Variable yang nantinya digunakan dalam klasifikasi adalah pm10, pm25, so2, co, no2 dan kategori.

	pm10	pm25	so2	co	o3	no2	kategori
10	59	79	21	26	15	31	1
11	30	46	21	14	16	24	0
12	23	33	19	11	14	19	0
13	36	53	23	14	12	24	1
14	29	36	20	14	12	21	0

Gambar 9. Hasil Penghapusan Kolom

4.2.4. Normalisasi Data

Proses normalisasi data bertujuan untuk mengubah variable numerik sehingga memiliki rentang nilai yang konsisten atau setara. Tujuannya adalah untuk memastikan bahwa semua variable memberikan pengaruh yang seimbang dalam analisis statistik atau pemodelan, serta untuk mempermudah perbandingan antara variable yang memiliki rentang nilai berbeda. Dengan normalisasi, variable-variable dengan skala yang berbeda dapat dianalisis bersama-sama tanpa menyebabkan bias atau distorsi dalam hasil analisis.

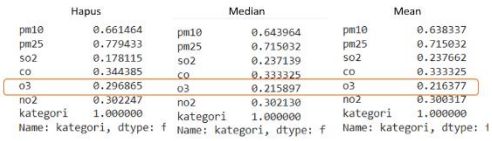
	pm10	pm25	so2	co	o3	no2	kategori
10	0.318182	0.354839	0.218391	0.462963	0.062147	0.476923	1
11	0.153409	0.177419	0.218391	0.240741	0.067797	0.369231	0
12	0.113636	0.107527	0.195402	0.185185	0.056497	0.292308	0
13	0.187500	0.215054	0.241379	0.240741	0.045198	0.369231	1
14	0.147727	0.123656	0.206897	0.240741	0.045198	0.323077	0

Gambar 10. Hasil Normalisasi Data

Peneliti menggunakan metode min-max scaling dalam menormalisasi data. Metode normalisasi min-max mengkonversi setiap nilai setiap fitur ke rentang [0, 1]. Prosedur ini menetapkan nilai minimum karakteristik menjadi 0 dan nilai maksimum menjadi 1, serta mengubah nilai lainnya menjadi angka desimal antara 0 dan 1.

4.2.5. Feature Selection

Pada feature selection dilakukan proses memilih subset fitur yang paling relevan dari dataset yang tersedia. Dalam pembuatan model prediktif, variable target atau dependen harus dikeluarkan dari dataset karena variable tersebut adalah yang diprediksi oleh model.



Gambar 11. Correlation Matrix

Sebelum memilih subset, variable atau parameter o3 dihapus, karena setelah dilakukan analisis korelasi, parameter o3 memiliki korelasi yang rendah dengan variable target atau variable lain dalam dataset. Parameter o3 dianggap kurang memberikan informasi yang signifikan untuk tujuan analisis atau pemodelan.

Langkah selanjutnya yaitu kolom kategori dihapus. Dalam proses pembuatan model prediktif, fitur target atau variable dependen harus dihapus dari dataset karena variable ini adalah variable yang ingin diprediksi oleh model. Dengan kata lain, fitur target tidak boleh dimasukkan sebagai fitur input ke dalam model, karena hal itu mengakibatkan bocornya informasi target ke dalam proses pembuatan model, yang dapat menyebabkan overfitting. Oleh karena itu kolom kategori dihapus, karena kolom kategori termasuk dalam fitur target yang ingin diprediksi.

4.2.6. Balancing Data

Tahap balancing data merupakan tahap untuk menyeimbangkan distribusi kelas dalam

dataset dengan memperoleh jumlah sampel yang seragam atau setidaknya lebih seimbang di antara kelas-kelas yang ada. Teknik yang digunakan dalam balancing data adalah Synthetic Minority Over-sampling Technique (SMOTE) yang merupakan teknik imbalance data.

Tabel 1. Hasil Balancing Data

Kelas	Original	SMOTE
0	230	2449
1	2449	2449
2	483	2449
Jumlah		7347

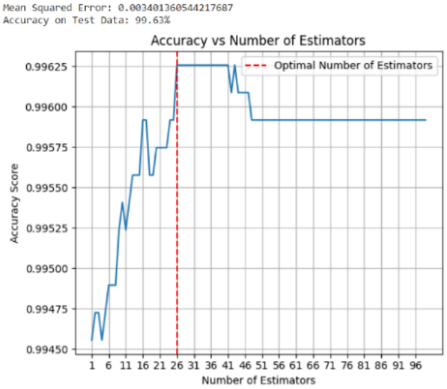
Kategori 0 adalah kelas minoritas dan kategori 1 adalah kelas mayoritas. Teknik SMOTE menyeimbangkan data dengan membuat sampel sintetis baru sesuai dengan jumlah kelas mayoritas yaitu sebanyak 2449 data.

4.3. Pembagian Data

Data yang diperoleh dari hasil balancing data sejumlah 2449 data per kategori. Peneliti membagi atau mengalokasikan 80% data untuk pelatihan dan 20% data untuk pengujian. Hasil dari pembagian data pelatihan dan data pengujian pada teknik SMOTE yaitu terdapat 5877 data pelatihan dan 1470 data pengujian

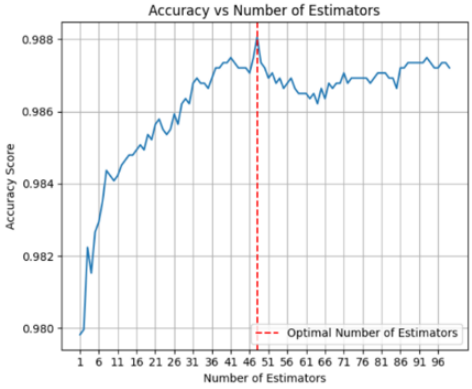
4.4. Klasifikasi

Langkah pertama pada klasifikasi XGBoost yaitu dengan pembuatan model. Fungsi yang digunakan dari modul library XGBoost yaitu XGBClassifier. Model dibuat dengan pohon keputusan awal yaitu 50, learning_rate=0.1 dan max_depth=3.



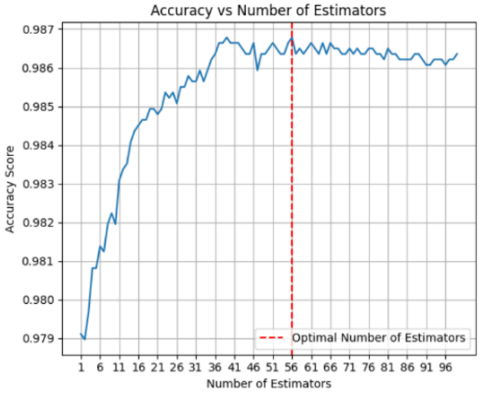
Gambar 12. Grafik Akurasi Missing Value dihapus

Dari grafik akurasi menunjukkan bahwa akurasi model berubah-ubah seiring dengan peningkatan jumlah estimator. Akurasi mencapai nilai tertinggi berada pada titik 26 yaitu dengan nilai accuracy sebesar 99.63%. Titik tersebut menunjukkan jumlah n_estimator optimal untuk model tersebut. Setelah titik 26 tidak terjadi peningkatan akurasi secara signifikan karena model telah mencapai batas kemampuan.



Gambar 13. Grafik Akurasi Missing Value diisi Mean

Dari grafik akurasi menunjukkan bahwa akurasi model berubah-ubah seiring dengan peningkatan jumlah estimator. Akurasi mencapai nilai tertinggi berada pada titik 48 yaitu dengan nilai akurasi sebesar 98.81%. Titik tersebut menunjukkan jumlah n_estimator optimal untuk model tersebut.

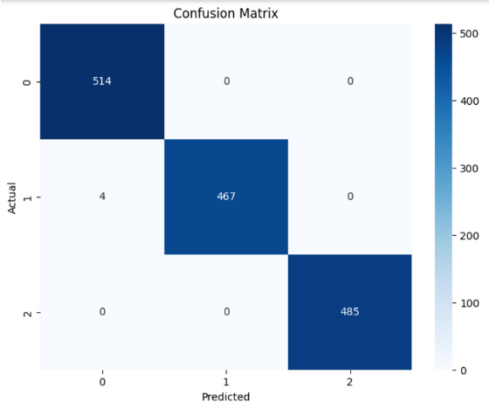


Gambar 14. Grafik Akurasi Missing Value dihapus

Dari grafik akurasi menunjukkan bahwa akurasi model berubah-ubah seiring dengan peningkatan jumlah estimator. Akurasi mencapai nilai tertinggi berada pada titik 56 yaitu dengan nilai akurasi sebesar 98.68%. Titik tersebut menunjukkan jumlah n_estimator optimal untuk model tersebut. Setelah titik 56 tidak terjadi peningkatan akurasi secara signifikan karena model telah mencapai batas kemampuan.

4.5. Evaluasi Model

Pada tahap ini dilakukan evaluasi dan perbandingan menyeluruh untuk menilai kinerja model klasifikasi algoritma XGBoost.



Gambar 15. Confusion Matrix Missing Value dihapus

Gambar 15 merupakan *confusion matrix* dari hasil pengujian dari *missing value* yang dihapus. Nilai *precision, recall* dan *f1-score* didapatkan dari perhitungan *confusion matrix.* Berikut adalah perhitungan manual untuk mencari nilai *precision, recall* dan *f1-score* dari *confusion matrix*.

- Kelas 0
  TP = 514
  FP = 4+0 = 4
  FN = 0+0 = 0

$$Precision = \frac{TP}{TP+FP} = \frac{514}{514+4} = 0.99 \quad (1)$$

$$Recall = \frac{TP}{TP+FN} = \frac{514}{514+0} = 1.00 \quad (2)$$

$$F1-score = \frac{2xPrecisionxRecall}{Precision+Recall} = \frac{2x0.99x1.00}{0.99x1.00} = 1.00 \quad (3)$$

- Kelas 1
  TP = 467
  FP = 0+0 = 0
  FN = 4+0 = 4

$$Precision = \frac{TP}{TP+FP} = \frac{467}{467+0} = 1.00 \quad (4)$$

$$Recall = \frac{TP}{TP+FN} = \frac{467}{467+4} = 0.99 \quad (5)$$

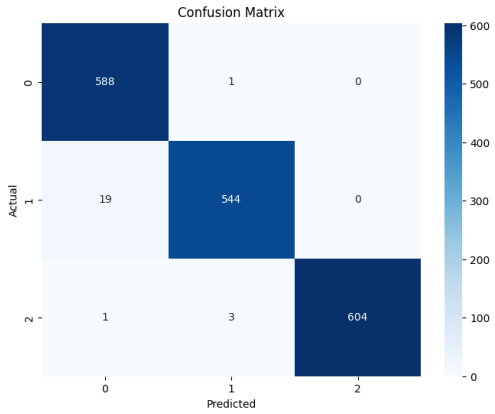
$$F1-score = \frac{2xPrecisionxRecall}{Precision+Recall} = \frac{2x1.00x0.99}{1.00x0.99} = 1.00 \quad (6)$$

- Kelas 2
  TP = 485
  FP = 0+0 = 0
  FN = 0+0 = 0

$$Precision = \frac{TP}{TP+FP} = \frac{485}{485+0} = 1.00 \quad (7)$$

$$Recall = \frac{TP}{TP+FN} = \frac{485}{485+0} = 1.00 \quad (8)$$

$$F1-score = \frac{2xPrecisionxRecall}{Precision+Recall} = \frac{2x1.00x1.00}{1.00x1.00} = 1.00 \quad (9)$$



Gambar 16. Confusion Matrix Missing Value diisi Mean

Gambar 16 merupakan *confusion matrix* dari hasil pengujian. Nilai *precision, recall* dan *f1-score* didapatkan dari perhitungan *confusion matrix.* Berikut adalah perhitungan manual untuk mencari nilai *precision, recall* dan *f1-score* dari *confusion matrix*.

- Kelas 0
  TP = 588
  FP = 19+1 = 20
  FN = 1+0 = 1

$$Precision = \frac{TP}{TP+FP} = \frac{588}{588+20} = 0.97 \quad (10)$$

$$Recall = \frac{TP}{TP+FN} = \frac{588}{588+1} = 1.00 \quad (11)$$

$$F1-score = \frac{2xPrecisionxRecall}{Precision+Recall} = \frac{2x0.97x1.00}{0.97x1.00} = 0.98 \quad (13)$$

- Kelas 1
  TP = 544
  FP = 1+3 = 4
  FN = 19+0 = 19

$$Precision = \frac{TP}{TP+FP} = \frac{544}{544+4} = 0.99 \quad (14)$$

$$Recall = \frac{TP}{TP+FN} = \frac{544}{544+19} = 0.97 \quad (15)$$

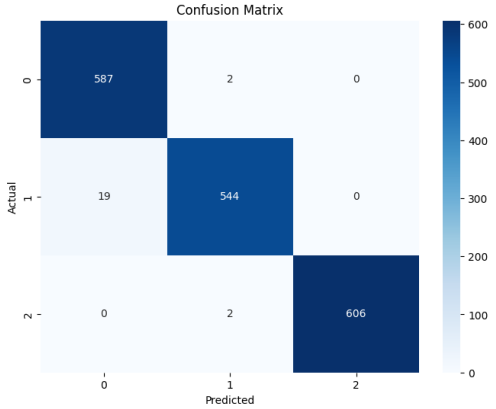
$$F1-score = \frac{2xPrecisionxRecall}{Precision+Recall} = \frac{2x0.99x0.97}{0.99x0.97} = 0.98 \quad (16)$$

- Kelas 2
  TP = 604
  FP = 0+0 = 0
  FN = 1+3 = 4

$$Precision = \frac{TP}{TP+FP} = \frac{604}{604+0} = 1.00 \quad (17)$$

$$Recall = \frac{TP}{TP+FN} = \frac{604}{604+4} = 0.99 \quad (18)$$

$$F1-score = \frac{2xPrecisionxRecall}{Precision+Recall} = \frac{2x1.00x0.99}{1.00x0.99} = 1.00 \quad (19)$$



Gambar 17. Confusion Matrix Missing Value diisi Median

Gambar 15 merupakan *confusion matrix* dari hasil pengujian dari *missing value* diisi median. Nilai *precision, recall* dan *f1-score* didapatkan dari perhitungan *confusion matrix.* Berikut adalah perhitungan manual untuk mencari nilai *precision, recall* dan *f1-score* dari *confusion matrix*.

- Kelas 0
  TP = 587
  FP = 19+0 = 19
  FN = 2+0 = 2

$$Precision = \frac{TP}{TP+FP} = \frac{587}{587+19} = 0.97 \quad (20)$$

$$Recall = \frac{TP}{TP+FN} = \frac{587}{587+2} = 1.00 \quad (21)$$

$$F1-score = \frac{2xPrecisionxRecall}{Precision+Recall} = \frac{2x0.97x1.00}{0.97x1.00} = 0.98 \quad (22)$$

- Kelas 1
  TP = 544
  FP = 2+2 = 4
  FN = 19+0 = 19

$$Precision = \frac{TP}{TP+FP} = \frac{544}{544+4} = 0.99 \quad (23)$$

$$Recall = \frac{TP}{TP+FN} = \frac{544}{544+19} = 0.97 \quad (24)$$

$$F1-score = \frac{2xPrecisionxRecall}{Precision+Recall} = \frac{2x0.99x0.97}{0.99x0.97} = 0.98 \quad (6)$$

- Kelas 2
  TP = 606
  FP = 0+0 = 0
  FN = 0+2 = 2

$$Precision = \frac{TP}{TP+FP} = \frac{606}{606+0} = 1.00 \quad (25)$$

$$Recall = \frac{TP}{TP+FN} = \frac{606}{606+2} = 1.00 \quad (26)$$

$$F1-score = \frac{2xPrecisionxRecall}{Precision+Recall} = \frac{2x1.00x1.00}{1.00x1.00} = 1.00 \quad (27)$$

Tabel 2. Hasil Validasi K-fold

Balancing Data	Akurasi					Rata-Rata
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	
Hapus	98.97 %	99.40 %	99.57 %	99.31 %	99.74 %	99.40 %
Mean	99.07 %	98.43 %	99.00 %	99.07 %	98.00 %	98.72 %
Median	98.79 %	98.43 %	98.57 %	99.07 %	98.22 %	98.62 %

Hasil dari validasi tidak terlalu berbeda dengan hasil akurasi. Hasil validasi terbaik ditunjukkan oleh balancing data yang dihapus dengan rata-rata sebesar 99.40%. Hal ini menunjukkan bahwa model memiliki kinerja yang sangat bagus.

5. KESIMPULAN

a. Teknik SMOTE menyeimbangkan data dengan membuat sampel sintetis baru sesuai dengan jumlah kelas mayoritas.
b. Hasil akurasi terbaik diperoleh dari balancing data dihapus dengan pohon keputusan yang berjumlah 26 dengan hasil akurasi sebesar 99.63%.
c. Rata-rata hasil validasi menggunakan k-fold cross validation adalah 99.40%. Hasil tersebut menunjukkan bahwa model yang dihasilkan memiliki tingkat akurasi yang sangat tinggi dalam mengklasifikasikan data ISPU, membuktikan bahwa model ini mampu memprediksi dengan tepat sebagian besar sampel dalam dataset.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak terkait yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

[1] M. S. Novelan, “Sistem Monitoring Kualitas Udara Dalam Ruangan Menggunakan Mikrokontroler dan Aplikasi Android,” InfoTekJar J. Nas. Inform. dan Teknol. Jar., vol. 4, no. 2, 2020.
[2] A. D. Wiranata, S. Soleman, I. Irwansyah, I. K. Sudaryana, and R. Rizal, “Klasifikasi Data Mining Untuk Menentukan Kualitas Udara Di

Provinsi Dki Jakarta Menggunakan Algoritma K-Nearest Neighbors (K-NN)," Infotech J. Technol. Inf., vol. 9, no. 1, 2023, doi: 10.37365/jti.v9i1.164.

[3] D. Kurniawan, S. R. Sulistiyanti, and U. Murdika, "Sistem Pemantau Gas Karbon Monoksida (Co) Dan Karbon Dioksida (Co2) Menggunakan Sensor Mq7 Dan Mq-135 Terintegrasi Telegram," J. Inform. dan Tek. Elektro Terap., vol. 11, no. 2, pp. 200–206, 2023, doi: 10.23960/jitet.v11i2.2963.

[4] A. Riyanto, A. Maheswara, R. Zulianty, V. Mayer Alegra, and A. Nur Muhammad, "Tanggung Jawab Pemerintah dalam Penyelesaian Masalah Polusi Udara di DKI Jakarta," J. Pendidik. Tambusai, vol. 7, no. 3, 2023, [Online]. Available: https://www.jptam.org/index.php/jptam/article/view/11232

[5] Iqa. S. Writers, "Laporan Kualitas Udara Dunia IQAir 2022 Menemukan Hanya 5% Negara yang Memenuhi Pedoman Polusi Udara PM2.5 WHO," IQAir, 2023. https://www.iqair.com/id/newsroom/world-air-quality-report-press-release-2022 (accessed Feb. 09, 2023).

[6] E. K. N. S. H. Pranita, "Polusi Udara di Indonesia Peringkat 1 di Asia Tenggara dan Peringkat 17 Negara Paling Berpolusi di Dunia," Kompas.com, 2022. https://www.kompas.com/sains/read/2022/04/07/123100123/polusi-udara-di-indonesia-peringkat-1-di-asia-tenggara-dan-peringkat-17?page=all (accessed Feb. 09, 2023).

[7] A. Amalia, A. Zaidiah, and I. N. Isnainiyah, "Prediksi Kualitas Udara Menggunakan Algoritma K-Nearest Neighbor," JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform., vol. 7, no. 2, 2022, doi: 10.29100/jipi.v7i2.2843.

[8] N. P. Decy Arwini, "Dampak Pencemaran Udara Terhadap Kualitas Udara Di Provinsi Bali," J. Ilm. Vastuwidya, vol. 2, no. 2, 2020, doi: 10.47532/jiv.v2i2.86.

[9] K. Therin and J. M. J. P. Santosa, "BANGUNAN UNTUK BERNAFAS SOLUSI POLUSI UDARA DI JAKARTA," J. Sains, Teknol. Urban, Perancangan, Arsit., vol. 3, no. 2, 2022, doi: 10.24912/stupa.v3i2.12442.

[10] F. Insani and S. I. Darlianti, "Pembentukan Model Regresi Linier Menggunakan Algoritma Genetika untuk Prediksi Parameter Indeks Standar Pencemar Udara (ISPU)," J. CoreIT J. Has. Penelit. …, vol. 5, no. 2, 2019.

[11] B. K. Hidayatullah, M. Kallista, C. Setianingsih, P. S1, and T. Komputer, "PREDIKSI INDEKS STANDAR PENCEMAR UDARA MENGGUNAKAN METODE LONG SHORT-TERM MEMORY BERBASIS WEB (STUDI KASUS PADA KOTA JAKARTA)," in e-Proceeding of Engineering , 2022.

[12] M. A. Fath, "Pengaruh Kualitas Udara dan Kondisi Iklim terhadap Perekonomian Masyarakat (Literature Review)," Media Gizi Kesmas, vol. 10, no. 2, 2021, doi: 10.20473/mgk.v10i2.2021.329-342.

[13] H. Aljuaid and N. Alwabel, "Air pollution prediction using machine learning algorithms," Int. J. Eng. Adv. Technol., vol. 8, no. 6 Special Issue 3, 2019, doi: 10.35940/ijeat.F1026.0986S319.

[14] D. Rahman Sya'ban, A. Hamzah, and E. Susanti, "Klasifikasi Buah Segar Dan Busuk Menggunakan Algoritma Convolutional Neural Network Dengan Tflite Sebagai Media Penerapan Model Machine Learning," Pros. SNAST, 2022, doi: 10.34151/prosidingsnast.v8i1.4180.

[15] A. Mulyadi Sapari, A. Id Hadiana, and F. Rakhmat Umbara, "Air Quality Classification Using Extreme Gradient Boosting (XGBOOST) Algorithm," Innov. Res. Informatics, vol. 5, no. 2, pp. 44–51, 2023, [Online]. Available: https://jurnal.unsil.ac.id/index.php/innovatics/article/view/8444

[16] F. Indah Sari, E. Leonie Gunawan, C. Ayu Adhigiadany, and A. Lisanthoni, "Model Prediksi Kepadatan Lalu Lintas: Perbandingan Algoritma Random Forest dan XGBoost," Pros. SENADA, vol. 3, no. 1, 2023, [Online]. Available: https://prosiding-senada.upnjatim.ac.id/index.php/senada/article/view/126

[17] A. N. Kasanah, M. Muladi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 3, no. 2, 2019, doi: 10.29207/resti.v3i2.945.

[18] Y. Primasanti and E. Indriastiningsih, "Analisis Dampak Pencemaran Udara Pt Delta Dunia Textile Terhadap Kondisi Masyarakat," J. Ilm. Keperawatan Indones., vol. 12 (1), no. 1, 2021.

[19] M. Nurwita, M. Maesaroh, and N. Widowati, "Upaya Dinas Lingkungan Hidup Dalam Pengendalian Pencemaran Udara di Kota Tangerang," J. Public Policy Manag. Rev., vol. 10, no. 2, 2021.

[20] D. Perdana and A. Muklason, "Machine Learning untuk Peramalan Kualitas Indeks Standar Pencemar Udara DKI Jakarta dengan Metode Hibrid ARIMAX-LSTM," J. Comput. Sci. Appl. Informatics, vol. 5, no. 3, 2023, [Online]. Available:

https://journal.unublitar.ac.id/ilkomnika/index.php/ilkomnika/article/view/588

[21] C. Afrilla, S. Suharwanto, and W. A. D. Kristanto, "Analisis Particulate Matter 10 μm (PM10) yang Ditimbulkan oleh Kegiatan Penambangan Andesit di Kabupaten Kulon Progo, DIY," Pros. Semin. Nas. Tek. Lingkung. Kebumian SATU BUMI, vol. 4, no. 1, 2023, doi: 10.31315/psb.v4i1.8874.

[22] Y. Sarwiyah, N. Rahaningsih, F. M. Basysyar, and E. Penulis Korespondensi, "Klasifikasi Data Nasabah Produk Asuransi Kendaraan Menggunakan Algoritma Naive Bayes Pada PT. Jasaraharja Putera," KOPERTIP Sci. J. Informatics Manag. Comput., vol. 4, no. 3, 2020.

[23] S. E. Herni Yulianti, Oni Soesanto, and Yuana Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," J. Math. Theory Appl., 2022, doi: 10.31605/jomta.v4i1.1792.

[24] J. Melvin Ayu Soraya Dachi and P. Sitompul, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," J. Ris. Rumpun Mat. dan Ilmu Pengetah. Alam, vol. 2, no. 2, 2023.

[25] R. D. Fitriani, H. Yasin, and T. Tarno, "Penanganan Klasifikasi Kelas Data Tidak Seimbang Dengan Random Oversampling Pada Naive Bayes (Studi Kasus: Status Peserta KB IUD di Kabupaten Kendal)," J. Gaussian, vol. 10, no. 1, pp. 11–20, 2021, doi: 10.14710/j.gauss.v10i1.30243.