Jurnal Teknologi Maritim Volume 8 No 2 Tahun 2025
19 Agustus 2025 / 12 Oktober 2025 / 13 Oktober 2025

Jurnal Teknologi Maritim

http://jtm.ppns.ac.id

Klasifikasi Indeks Kualitas Udara DKI Jakarta dengan Multinomial Logistic Regression Berbasis Machine Learning

Alma Vita Sophia[1*], Farhah Izzah Dinillah[2], Imam Mahfudzi[3]

[1,2] *Jurusan Teknik Bangunan Kapal, Politeknik Perkapalan Negeri Surabaya, Jl Teknik Kimia ITS Surabaya*

[3] *Jurusan Teknik Permesinan Kapal, Politeknik Perkapalan Negeri Surabaya, Jl Teknik Kimia ITS Surabaya*

Abstrak. Penyusunan strategi pengendalian pencemaran udara yang tepat sasaran memerlukan data pengukuran kualitas udara dan analisis hasilnya yang akurat. Dalam pelaksanaannya diperlukan penentuan indeks kualitas udara yang terpantau terus menerus dengan mengolah data hasil pengukuran Stasiun Pemantauan Kualitas Udara (SPKU) tahun 2024 yang tersebar di DKI Jakarta. Penelitian ini diusulkan dengan tujuan menganalisis data Indeks Kualitas Udara (PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, dan NO_2) di DKI Jakarta dan menentukan klasifikasi indeks kualitas udara dengan metode regresi logistik multinomial berbasis machine learning.

Kata kunci: Indeks Kualitas Udara (IKU), Klasifikasi, Regresi Logistik Multinomial, Pembelajaran Mesin

Abstract. Effective air pollution control strategy creation needs air quality measurement data and accurate its result analysis. In practice, simultaneously air quality index monitoring is necessary by means of Stasiun Pemantauan Kualitas Udara (SPKU) measurement data analysis on year 2024 which is spread in DKI Jakarta. This research is proposed in order to analyzing air quality index (PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, and NO_2) data in DKI Jakarta and to decide air quality index classification through machine learning-based multinomial logistic regression method.

Keywords: Air Quality Index (AQI), Classification, Multinomial Logistic Regression, Machine Learning

1. Pendahuluan

Situs pemantau udara IQAir menunjukkan kondisi udara DKI Jakarta tidak sehat dengan mencatat indeks kualitas udara di angka 183 dan menempatkannya di peringkat kedua terburuk di dunia pada Rabu, 23 Juli 2024

Email Korespondensi: alma@ppns.ac.id
doi: 10.35991/jtm.v8i2.80

(https://www.tempo.co/lingkungan/kualitas-udara-jakarta-terburuk-kedua-di-dunia-2050384). Saat tingkat polusi udara mencapai ambang yang membahayakan kesehatan, maka indeks kualitas udara (*air quality index*, AQI) dirancang sebagai sistem peringatan bagi masyarakat.

Emisi kendaraan bermotor, industri, dan pembakaran sampah menjadi penyebab utama polusi udara yang bisa mengancam kesehatan manusia, hewan dan lingkungan. Sektor transportasi menjadi penyumbang polusi udara tertinggi di DKI Jakarta sekitar 67,04% dari kendaraan bermotor berbahan bakar fosil, disusul sektor industri manufaktur dan energi sekitar 32,5% dari pabrik dan fasilitas industri lainnya lewat pembakaran dan produksi, dan terakhir sektor rumah tangga sekitar 0,43% dari pembakaran sampah dan pemakaian bahan bakar fosil.

Polusi udara mengurangi 2,3 tahun dari harapan hidup rata-rata individu, ditunjukkan University of Chicago dalam penelitiannya. Greenpeace Indonesia memperkirakan 7.390 penduduk Jakarta meninggal lebih awal karena polusi, sedangkan 2000 bayi lahir dengan berat badan rendah karena penyebab yang sama. Populasi yang paling terdampak oleh polusi udara di antaranya adalah anak-anak, lansia, dan individu dengan penyakit komorbid. Polusi udara mengakibatkan infeksi paru maupun iritasi membran mukosa pada hidung, mulut, kulit, dan mata. Ukuran PM2,5 yang kecil dapat memasuki sirkulasi darah dan dapat mengakibatkan kerusakan organ dalam, termasuk gangguan pada kesehatan jantung serta mengganggu kesehatan janin di dalam kandungan. (https://ohce.wg.ugm.ac.id/polusi-jakarta-peringkat-1-di-dunia-bagaimana-dampaknya-pada-kesehatan/)

Polusi udara memiliki dampak yang signifikan terhadap kesehatan manusia. Partikel PM2.5 yang terdapat dalam polusi udara dapat menyebabkan berbagai penyakit pernapasan, kardiovaskular, dan bahkan kanker. Selain itu, polusi udara juga berdampak negatif pada kesehatan hewan, terutama hewan peliharaan yang sering terpapar udara luar. Ekosistem juga terpengaruh, dengan polusi udara yang merusak tanaman dan mengurangi kualitas tanah dan air.

Penyusunan strategi pengendalian pencemaran udara yang tepat sasaran memerlukan data pengukuran kualitas udara dan analisis hasilnya yang akurat. Dalam pelaksanaannya diperlukan penentuan indeks kualitas udara yang terpantau terus menerus dengan mengolah data hasil pengukuran Stasiun Pemantauan Kualitas Udara (SPKU) tahun 2024 yang tersebar di DKI Jakarta. Penelitian ini diusulkan dengan tujuan menganalisis data kualitas udara di DKI Jakarta dan menentukan klasifikasi indeks kualitas udara dengan metode regresi logistik multinomial untuk kelas > 2 yang menjadi bagian dari pembelajaran mesin.

2. Tinjauan Pustaka

Metode *Autoregressive Integrated Moving Average* (ARIMA) dengan pendekatan Box-Jenkins dipakai Sophia (2020) untuk memprediksi partikulat PM_{10} menggunakan data Stasiun Pemantau Kualitas Udara (SPKU) Kelapa Gading tahun 2019 yang menghasilkan model ARIMA(1,1,1). Umri (2021) menganalisis dan membandingkan algoritma klasifikasi untuk menentukan tingkat indeks standar pencemaran udara (ISPU) DKI Jakarta. Dengan metode *Neural Network, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes*, dan *Decission Tree* untuk mengolah data

sulfur dioksida (SO_2), karbon monoksida (CO), ozon permukaan (O_3), nitrogen dioksida (NO_2), dan partikel debu (PM_{10}), diperoleh hasil terbaik nilai akurasi sebesar 99.80%, nilai RMSE terkecil dan di bawah 0,1 yakni 0,039, nilai kappa yang hampir sempurna yakni 0,996, serta waktu yang dibutuhkan hanya 0,8 detik dari metode *Decision Tree*.

Pengembangan model klasifikasi ISPU oleh Ayu (2023) dari data tahun 2022 menggunakan metode *Random Forest* dan SVM serta teknik oversampling *Synthetic Minority Over-sampling Technique (SMOTE)* menghasilkan akurasi 98% untuk *Random Forest* tanpa SMOTE, 99% untuk *Random Forest* dengan SMOTE, 91% untuk SVM tanpa SMOTE, dan 95% untuk SVM dengan SMOTE. Putri (2023) mengelompokkan ISPU dengan data tahun 2021 dan metode *Artificial Neural Network (ANN)* yang menghasilkan *accuracy* 94%, *precision* 90%, serta *recall* 100% dari arsitektur jaringan 5 *input layer*, 4 *hidden layer*, dan 2 *output layer*, 5000 *epoch* dan *learning rate* 0.001.

Pengelompokan ISPU untuk data tak seimbang dengan pembelajaran mesin dilakukan juga oleh Jayadi (2024). Data yang diolah (tahun 2010-2021) memakai metode *Logistic Regression, Naïve Bayes, Decision Tree*, SVM dan AdaBoost dan hasil terbaik didapatkan dari dengan *accuracy* 99%, *precision* 98%, *recall* 99%, dan *F1-score* 98%.

Luthfi (2024) membandingkan SVM, *Random Forest, Light Gradient-Boosting Machine (LGM)* dan *Random Forest* memberikan akurasi tertinggi 98%. Metode *eXtreme Gradient Boosting (XGBoost)* dengan bantuan SMOTE yang dipakai Sajiwo (2024) untuk mengolah data tahun 2022-2023 menghasilkan akurasi 99,63%. Data tahun 2016-2021 diklasifikasikan pula memakai metode *Random Forest* dengan hasil *accuracy* 99,94%, *precision* 99,98%, *recall* 99,9% dan *F1-score* 99,94%.

3. Metode

Pengelompokan IKU di DKI Jakarta ini dilakukan dengan urutan cara berikut: a) pengumpulan data, b) pemilahan data, c) trasformasi data, d) pemodelan, dan e) evaluasi model. Penelitian dimulai dengan mengumpulkan data berbentuk *dataset* dari https://satudata.jakarta.go.id/open-data. Data mentah dari file berformat JSON atau CSV yang diunduh berisi 1825 baris dan 12 kolom, dengan kepala kolom terdiri dari: periode_data, tanggal, stasiun, pm_sepuluh, pm_duakomalima, sulfur_dioksida, karbon_monoksida, ozon, nitrogen_dioksida, max, parameter_pencemar_kritis, dan kategori. Pemilahan data dilakukan dengan menghapus kolom yang tidak diperlukan, yaitu: periode_data, tanggal, max, stasiun, dan parameter_pencemar_kritis. Transformasi data dilakukan dengan: a) mengubah data pada kolom selain stasiun dan kategori menjadi numerik, b) mengganti sel yang kosong dengan nilai median dari kolom sel tersebut, c) mengubah kelas yang paling sedikit jumlahnya menjadi kelas lain yang (hampir) setara, dan d) menyeimbangkan kelas dengan metode *synthetic minority oversampling technique* (SMOTE). Pemodelan dan evaluasinya memakai metode regresi logistik multinomial dilakukan dengan bantuan aplikasi JASP. Secara teori, Matloff (2017) mengasumsikan model adalah log rasio peluang (*log-odds ratio*) dari kelas i relatif terhadap kelas 0 berbentuk linier,

$$\log \frac{P(Y=i \mid X=t)}{P(Y=0 \mid X=t)} = \log \frac{\gamma_i}{\gamma_0} = \beta_{0i} + \beta_{1i}t_1 + ... + \beta_{pi}, \quad i = 1, 2, ..., m-1 \tag{1}$$

dimana i mulai dari 1 (bukan dari 0) karena setiap kelas dibandingkan terhadap kelas 0. β_{ji} bisa dihitung dengan kemiripan maksimum (*maximum likelihood*), menghasilkan

$$\log \frac{\widehat{\gamma}_i}{\widehat{\gamma}_0} = \widehat{\beta}_{0i} + \widehat{\beta}_{1i}t_1 + ... + \widehat{\beta}_{pi} \tag{2}$$

Penerapan eksponensial selanjutnya akan menghasilkan rasio $\widehat{\gamma}_i/\widehat{\gamma}_0$ yang mana peluang $\widehat{\gamma}_i$ masing-masing diselesaikan dengan aljabar memakai kendala

$$\sum_{i=0}^{m-1} \widehat{\gamma}_i = 1 \tag{3}$$

4. Hasil dan Pembahasan

Statistik deskriptif meliputi rerata, simpangan baku, nilai Saphiro-Wilk, nilai p dari Saphiro-Wilk, minimum dan maksimum serta frekuensi variabel kategori disajikan Tabel 1 dan Tabel 2.

Tabel 1. Statistik Deskriptif

	PM_{10}	$PM_{2.5}$	SO_2	CO	O_3	NO_2
Mean	52.76	77.82	38.07	12.63	28.25	17.77
Std. Devition	14.96	22.02	13.33	6.594	13.03	8.621
Shapiro-Wilk	0.969	0.935	0.936	0.896	0.953	0.983
P-value of Shapiro-Wilk	< .001	< .001	< .001	< .001	< .001	< .001
Minimum	3.000	16.00	11.00	1.000	4.000	0.000
Maximum	163.0	287.0	89.00	55.00	81.00	53.00

Tabel 2. Frekuensi Kategori

Kategori	Frekuensi	Persen	Persen Kumulatif
BAIK	236	13.1	13.1
SANGAT TIDAK SEHAT	3	0.2	13.2
SEDANG	1,358	75.3	88.5
TIDAK SEHAT	207	11.5	100.0
Total	1,804	100.0	

Nilai p dari Saphiro-Wilk di Tabel 1 menunjukkan bahwa semua variabel kontinyu terdistribusi normal secara signifikan, sementara kategori yang merupakan variabel nominal punya 4 nilai (BAIK, SANGAT TIDAK SEHAT, SEDANG dan TIDAK SEHAT) dengan total 1804 data dari 1825 data awal atau berkurang 1,15% dari semula karena ada data kosong yang tidak dipakai atau dihilangkan. Kesenjangan kelas (*class imbalance*) terjadi akibat perbedaan jumlah data antar kelas yang mencolok, seperti halnya kategori SANGAT TIDAK SEHAT yang berjumlah hanya 3 data (0.2%) yang tidak sebanding dengan kategori SEDANG yang berjumlah 1358 data atau mewakili 75.3% seluruh data.

Data berkategori SANGAT TIDAK SEHAT diubah menjadi TIDAK SEHAT dan dilakukan *over-sampling* dengan SMOTE memakai bantuan *imbalanced-learn*, salah satu *library* Python, untuk menyeimbangkan kelas (*class balance*). Data hasil pengubahan dengan kategori BAIK (236 sampel), SEDANG (1358 sampel), dan TIDAK SEHAT (dari 207 menjadi 210 sampel) disampling ulang menjadi 1358 sampel untuk setiap kategori.

Regresi logistik multinomial dipakai untuk menganalisis hubungan antara dosis PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, dan NO_2 sebagai variabel prediktor dengan IKU atau kategori sebagai variabel respon yang modelnya dirangkum Tabel 3.

Tabel 3. Ringkasan Model

Family	Link	n(Train)	n(Test)	Test Accuracy
Multinomial	Logit	3.260	814	0,914

Tabel 4. Confusion Matrix

		Prediksi		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	2941	19	0
	SEDANG	39	247	12
	TIDAK SEHAT	0	0	256

Dari 4074 sampel, 3260 (80%) dipakai *training* dan 814 (20%) dipakai *testing* dengan kesesuaian antara nilai aktual dan prediksi (akurasi) 91,4% yang didukung *confusion matrix* pada Tabel 4. Karena kelas BAIK yang diinginkan, maka *true positive* berlaku jika nilai aktualnya BAIK diprediksi nilainya BAIK dan hasilnya 2941 sampel. Sedangkan kelas TIDAK SEHAT tidak dikehendaki, sehingga *true negative* berlaku jika nilai aktual sesuai dengan nilai prediksinya,yaitu TIDAK SEHAT dan hasilnya 256 sampel. Di samping itu ada *false positive* (nilai prediksinya BAIK dan nilai aktualnya selain BAIK) dengan hasil 39 sampel dan *false negative* (nilai aktualnya selain TIDAK SEHAT dan nilai prediksinya TIDAK SEHAT) dengan hasil 12 sampel.

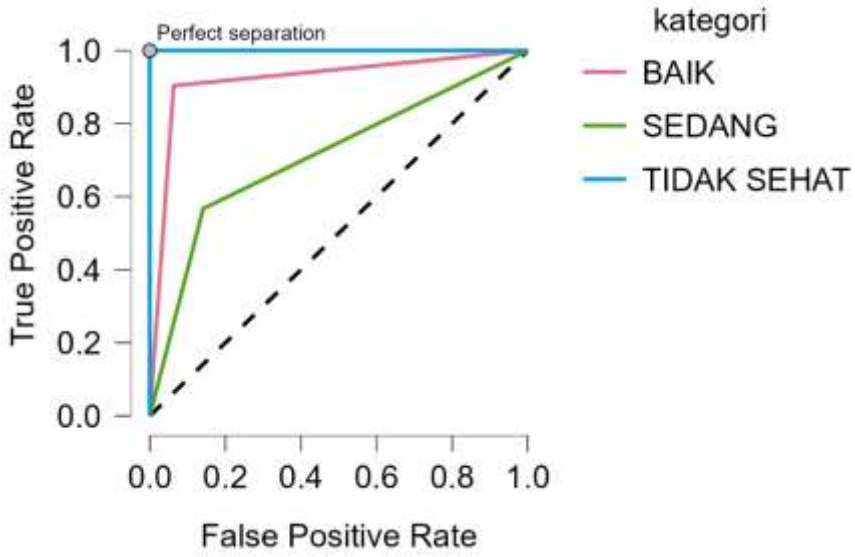
Tabel 5. Metrik Kinerja Model

	BAIK	SEDANG	TIDAK SEHAT	Average / Total
Support	260	298	256	814
Accuracy	0,929	0,914	0,985	0,943
Precision (Positive Predictive Value)	0,861	0,929	0,955	0,915
Recall (True Positive Rate)	0,927	0,829	1,000	0,914
F1 Score	0,893	0,876	0,977	0,913
Area Under Curve (AUC)	0,920	0,713	1,000	0,878
Threat Score	2,485	2,775	10,67	5,309
Statistical Parity	0,344	0,327	0,329	1,000

Tabel 5 merangkum metrik kinerja yang merupakan hasil *running* model dengan modul *Machine Learning* menggunakan aplikasi JASP. Pemilihan metrik kinerja yang tepat diperlukan pada data yang memiliki kesenjangan kelas karena pemakaian metrik yang keliru akan memberikan pemahaman yang keliru tentang bagaimana model bekerja dan selanjutnya tidak akan membantu dalam pengembangan atau pemilihan model yang sesuai untuk data.

Accuracy adalah proporsi semua klasifikasi, baik nilai positif ataupun negatif. Dalam hal ini *accuracy* hanya bisa digunakan untuk masing-masing kelas, tetapi tidak bagus untuk data yang punya kesenjangan kelas karena perlakuannya yang sama untuk semua kelas, yang berarti kinerja model pada kelas mayoritas akan mendominasi metrik ini (Huyen, 2022).

Precision adalah proporsi semua klasifikasi positif model yang aktualnya positif. *Recall (true positive rate)* adalah proporsi semua nilai aktual positif yang diklasifikasi benar sebagai nilai positif. Sedangkan *Area Under Curve* menunjukkan seberapa bagus model bisa memprediksi dengan tepat, seperti terlihat pada Gambar 1. *Statistical Parity* menunjukkan peluang yang hampir proporsional antara kategori BAIK (34,4%), SEDANG (32,7%) dan TIDAK SEHAT (32,9%).



Gambar 1. Kurva ROC

Matriks kepentingan fitur menampilkan variabel bebas yang berpengaruh terhadap hasil klasifikasi. Dimulai dari PM2,5 sebagai parameter IKU paling penting dengan *mean dropout loss* 2133,5 (terbesar), disusul PM10, SO2, O3, CO, dan terakhir NO2 sebagai parameter IKU tidak atau kurang penting dengan *mean dropout loss* 240,3 (terkecil). Hasil *mean dropout loss* yang didefinisikan sebagai *cross entropy* dari 50 permutasi setiap parameter dirangkum pada Tabel 6.

Tabel 6. Kepentingan Fitur

	Mean dropout loss
pmduakomalima	2133.5
pmsepuluh	803.6
sulfurdioksida	385.5
ozon	257.3
karbonmonoksida	242.3
nitrogendioksida	240.3

5. Kesimpulan

Data IKU DKI Jakarta tahun 2024 hasil pengukuran dosis PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, dan NO_2 bisa dipakai untuk menentukan klasifikasi kualitas udara dengan kategori SEHAT, SEDANG, dan TIDAK SEHAT. Kesenjangan kelas (*class imbalance*) diatasi dengan teknik SMOTE sehingga model regresi logistik multinomial masih mampu memprediksi kategori kualitas udara yang ada didukung metriks evaluasi yang

memadai dengan F1 score rata-rata 91,3%, dan AUC dalam ROC rata-rata 87,8%. PM2,5 menjadi parameter IKU paling penting, disusul PM10, SO2, O3, CO, dan terakhir NO2 sebagai parameter kurang atau tidak penting dalam klasifikasi IKU di DKI Jakarta ini.

Ucapan terima kasih
Penelitian ini dilakukan dengan dana DIPA PPNS tahun 2025.

Daftar Pustaka

Sophia, A. V. (2020). Model ARIMA untuk Prediksi Kualitas Udara PM10 DKI Jakarta dengan Metode Box-Jenkins (Studi Kasus SPKU Kelapa Gading). *Jurnal Teknologi Maritim*. 3, 1 (May. 2020), 10-13. DOI:https://doi.org/10.33863/jtm.v3i1.2910

Ayu, G., Lestari, N., Agus, K., & Aryanto, A. (2023). Peningkatan Akurasi Klasifikasi Kualitas Udara melalui Oversampling dengan Metode Support Vector Machine dan Random Forest. *Jurnal Sistem Dan Informatika (JSI)*. https://doi.org/10.30864/jsi.v18i1.596

Firdaus, R., Habibie, H., & Rizki, Y. (2024). Implementasi Algoritma Random Forest Untuk Klasifikasi Pencemaran Udara di Wilayah Jakarta Berdasarkan Jakarta Open Data. *JURNAL FASILKOM*. https://doi.org/10.37859/jf.v14i2.7669

Luthfi, A. M., & Fauzi, F. (2024). Perbandingan Klasifikasi Random Forest, Support Vector Machines, dan LGBM Pada Klasifikasi Kualitas Udara di Jakarta. *Justindo (Jurnal Sistem Dan Teknologi Informasi Indonesia)*. https://doi.org/10.32528/justindo.v9i2.1912

Putri, L. A., & Suwanda. (2023). Implementasi Metode Artificial Neural Network (ANN) Algoritma Backpropagation untuk Klasifikasi Kualitas Udara di Provinsi DKI Jakarta Tahun 2021. *Bandung Conference Series: Statistics*. https://doi.org/10.29313/bcss.v3i2.7826

Sajiwo, A. F. B., Rahmat, B., & Junaidi, A. (2024). Klasifikasi Indeks Standar Pencemaran Udaran (ISPU) Menggunakan Algoritma Xgboost dengan Teknik Imbalanced Data (SMOTE). *Jurnal Informatika Dan Teknik Elektro Terapan*. https://doi.org/10.23960/jitet.v12i3.4699

Umri, S. S. A., & Umri, S. S. A. (2021). Analisis dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara Di DKI JAKARTA. *JIKO (Jurnal Informatika Dan Komputer)*. https://doi.org/10.33387/jiko.v4i2.2871

Jayadi, B. V., Lauro, M. D., Rusdi, Z., & Handhayani, T. (2024). Air Quality Index Classification for Imbalanced Data using Machine Learning Approach. *Sistemasi: Jurnal Sistem Informasi*, *13*(3), 951-958.

The pandas development team. (2025). pandas-dev/pandas: Pandas (v2.3.0). Zenodo. https://doi.org/10.5281/zenodo.15597513

JASP Team (2025). JASP (Version 0.95)[Computer software]

Kluyver, Thomas, et al, (2016) Jupyter Notebooks – a publishing format for reproducible computational workflows. Loizides, Fernando and Scmidt, Birgit (eds.) In Positioning and Power in Academic Publishing: Players, Agents and Agendas. *IOS Press*. pp. 87-90 . (doi:10.3233/978-1-61499-649-1-87).

Polusi Jakarta Peringkat 1 di Dunia, Bagaimana Dampaknya pada Kesehatan? Retrieved September 29, 2025 from https://ohce.wg.ugm.ac.id/polusi-jakarta-peringkat-1-di-dunia-bagaimana-dampaknya-pada-kesehatan

Classification: Accuracy, recall, precision, and related metrics. Retrieved August 18, 2025, from https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

Huyen, Chip (2022) *Designing Machine Learning System*, O'Reilly
Matloff, Norman (2017) *Statistical Regression and Classification: From Linear Models to Machine Learning*, CRC Press