RAGAM: Journal of Statistics and Its Application
Vol 05 No 01, Pages 34-43, February 2026
DOI: https://doi.org/10.20527/ragam.v5i1.18196
e-ISSN: 2962-8539

# Data Mining-Based Air Quality Modeling in Jakarta Using Random Forest, K-Naerest Neigbors, and Naive Bayes Algorithms

**Naufalarizqa Ramadha Meisa Putra**[1*]
[1]Universitas Satya Negara Indonesia, Jakarta Selatan, Indonesia
**Corresponding author*: naufalarizqa@usni.ac.id

***Abstract***
*Air quality prediction plays an important role in supporting public health monitoring in highly urbanized regions such as DKI Jakarta. This study aims to predict the Air Pollutant Standard Index (ISPU) category using three supervised learning algorithms, namely Random Forest, k Nearest Neighbors (kNN), and Naive Bayes, based on five pollutant parameters: PM10, SO2, CO, O3, and NO2. The dataset used in this study consists of validated daily air-quality records that have undergone preprocessing steps including handling missing values and applying min max normalization. Model evaluation is conducted using the Test and Score feature in the Orange Data Mining software, which provides a visual programming environment for machine learning analysis. The results show that Random Forest achieves the highest performance with an accuracy of 97 percent, followed by k-NN with 94 percent and Naive Bayes with 88 percent. Feature ranking using the Chi Square test indicates that PM10 is the most dominant factor influencing ISPU category with a value of 870.174, followed by O3 and NO2. These findings highlight that ensemble-based models are well suited for multiclass air quality classification and confirm that particulate matter remains a key determinant of air quality conditions in Jakarta.*

***Keywords***: *Data mining, Air Quality Prediction, Random Forest, K-Nearest Neighbors, and Naive Bayes*

**Article History:**
*Received* : 29 January 2026
*Revised* : 06 March 2026
*Accepted* : 09 March 2026

© 2026 Putra. Published by Program Studi Statistika Fakultas MIPA Universitas Lambung Mangkurat.
This is an open access article under the CC BY-SA license.

***How to Cite***:
N. R. M. Putra, *Data Mining-Based Air Quality Modeling in Jakarta Using Random Forest, K-Naerest Neigbors, and Naive Bayes Algorithms*, RAGAM: Journal of Statistics and Its Application, vol. 5, no. 1, pp. 34-43, 2026, https://doi.org/10.20527/ragam.v5i1.18196

## 1. PENDAHULUAN

Kualitas udara merupakan faktor penting yang memengaruhi kesehatan manusia, produktivitas, dan keberlanjutan lingkungan. Di wilayah metropolitan seperti DKI Jakarta, aktivitas transportasi, industri, pembangunan infrastruktur, serta emisi domestik berkontribusi signifikan terhadap peningkatan konsentrasi polutan seperti PM10, PM2.5, CO, $SO_2$, $NO_2$, dan $O_3$ [1][2]. Paparan polutan tersebut dapat memicu gangguan pernapasan, penyakit kardiovaskular, menurunnya kapasitas paru, dan risiko kesehatan jangka panjang, sehingga pemantauan kualitas udara menjadi sangat penting bagi pemerintah dan masyarakat.

Indonesia menggunakan Indeks Standar Pencemar Udara (ISPU) untuk menggambarkan mutu udara dalam kategori Baik, Sedang, Tidak Sehat, Sangat Tidak Sehat, dan Berbahaya, sebagaimana diatur dalam Peraturan Menteri Lingkungan Hidup dan Kehutanan No. 14 Tahun 2020 [3]. Nilai ISPU dihitung menggunakan metode fungsi linear sepotong (*piecewise linear*) berdasarkan *breakpoints* tiap polutan, dengan hasil akhir ditentukan oleh sub-indeks polutan terbesar di lokasi pemantauan [4]. Data ISPU di DKI Jakarta

diperoleh dari Stasiun Pemantau Kualitas Udara (SPKU) dan dipublikasikan melalui kanal resmi seperti KLHK dan Satu Data Jakarta [5][6].

Seiring meningkatnya volume dan keragaman data kualitas udara, metode komputasi berbasis *machine learning* semakin relevan digunakan untuk menangkap pola non-linier antar polutan, yang tidak selalu dapat dimodelkan dengan pendekatan statistik konvensional [1][7]. Sejumlah penelitian menunjukkan bahwa algoritma seperti *Random Forest, k-Nearest Neighbors* (k-NN), serta *Naive Bayes* mampu memberikan performa prediksi yang baik dalam konteks klasifikasi kualitas udara [1][2][8]. Berdasarkan kebutuhan tersebut, penelitian ini memiliki tujuan untuk:

1. Memprediksi kategori ISPU di DKI Jakarta menggunakan lima parameter polutan utama (PM10, $SO_2$, CO, $O_3$, $NO_2$),
2. Membandingkan performa tiga algoritma *machine learning*, yaitu *Random Forest, k-NN, dan Gaussian Naive Bayes,*
3. Menentukan parameter polutan paling berpengaruh terhadap pembentukan kategori ISPU.

Hasil penelitian diharapkan dapat memberikan wawasan bagi pemerintah dan pemangku kebijakan dalam merumuskan strategi pengendalian polusi yang lebih efisien dan berbasis data.

## 2. TINJAUAN PUSTAKA

### 2.1 Kualitas Udara dan Pencemar Utama

Kualitas udara dipengaruhi oleh berbagai zat pencemar yang berasal dari transportasi, industri, aktivitas domestik, serta proses kimia di atmosfer. Polutan seperti PM10, PM2.5, $NO_2$, $SO_2$, CO, dan $O_3$ diketahui memiliki dampak signifikan terhadap kesehatan manusia serta menurunkan kualitas lingkungan [1][2]. Pemantauan rutin dilakukan melalui Stasiun Pemantau Kualitas Udara (SPKU) yang beroperasi 24 jam sesuai pedoman pemerintah [3].

### 2.2 Indeks Standar Pencemar Udara (ISPU)

ISPU merupakan indikator kualitas udara secara nasional yang mengklasifikasikan mutu udara menjadi lima kategori, mulai dari Baik hingga Berbahaya, sesuai Permen LHK No. 14 Tahun 2020 [3]. Perhitungan ISPU menggunakan model linear sepotong dengan *breakpoints* polutan yang telah ditetapkan [4]. Rumus umum:

$$I = \frac{I_{\text{hi}} - I_{\text{lo}}}{C_{\text{hi}} - C_{\text{lo}}}(C - C_{\text{lo}}) + I_{\text{lo}} \quad (1)$$

Nilai ISPU lokasi merupakan nilai tertinggi di antara sub-indeks polutan, sehingga polutan paling dominan akan menentukan kategori akhir kualitas udara [4]. Publikasi data ISPU harian dan jam tersedia melalui portal KLHK dan Satu Data Jakarta [5][6].

### 2.3 Algoritma *Random Forest*

*Random Forest* merupakan algoritma ensemble learning berbasis kumpulan decision tree yang dilatih melalui metode *bootstrap aggregation* (*bagging*) [7]. Algoritma ini memiliki keunggulan dalam menangani hubungan non-linier, mengurangi risiko *overfitting*, dan stabil pada dataset lingkungan yang kompleks. Beberapa penelitian kualitas udara menunjukkan bahwa *Random Forest* mampu memberikan akurasi tinggi dalam memprediksi kategori polusi [1][8].

### 2.4 Algoritma *k-Nearest Neighbors* (k-NN)

k-NN adalah algoritma *instance-based learning* yang menentukan kelas berdasarkan mayoritas dari *k* tetangga terdekat menggunakan metrik jarak seperti *Euclidean distance* [9]. Normalisasi fitur sangat penting untuk menjaga kesetaraan kontribusi antar fitur. Pemilihan nilai *k* sangat memengaruhi performa model. Nilai *k* yang terlalu kecil akan sensitif terhadap *noise*, sedangkan nilai *k* yang terlalu besar dapat mengaburkan batas antar kelas [9].

### 2.5 Algoritma *Naive Bayes*

*Naive Bayes* adalah metode klasifikasi probabilistik berbasis Teorema Bayes, dengan asumsi independensi antar fitur [10]. Pada penelitian ini digunakan Gaussian Naive Bayes, yang mengasumsikan setiap fitur mengikuti distribusi normal pada tiap kelas. Meskipun asumsi independensi terkadang tidak terpenuhi pada data polusi udara, algoritma ini tetap efisien dan memberikan performa yang baik pada dataset berukuran besar [8][10].

## 3. METODE PENELITIAN

### 3.1 Jenis Penelitian

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan klasifikasi berbasis *machine learning*. Fokus penelitian adalah memprediksi kategori Indeks Standar Pencemar Udara (ISPU) menggunakan tiga algoritma pembelajaran terawasi, yaitu *Random Forest, k-Nearest Neighbors* (k-NN), dan *Gaussian Naive Bayes*. Metode ini dipilih karena telah terbukti efektif pada berbagai studi pemodelan kualitas udara dan lingkungan [1][8][9].

### 3.2 Data Penelitian

Data penelitian yang digunakan dalam studi ini berasal dari catatan harian kualitas udara yang diukur oleh Stasiun Pemantau Kualitas Udara (SPKU) di wilayah DKI Jakarta, yang dipublikasikan melalui kanal data resmi pemerintah. Dataset yang digunakan mencakup periode Januari 2024 hingga Mei 2025, dengan total 2.570 entri setelah dilakukan proses pra-pemrosesan awal berupa pembersihan data dan penghapusan duplikasi. Dataset ini berisi variabel-variabel polutan utama seperti PM10, PM2.5, $O_3$, $NO_2$, dan $SO_2$, serta kategori ISPU harian untuk setiap titik pengamatan di Jakarta.

Berdasarkan analisis distribusi kategori ISPU pada dataset tersebut, kualitas udara di Jakarta selama periode pengamatan didominasi oleh kategori Sedang dengan proporsi 77,32%, diikuti kategori Baik sebesar 14,63%, dan kategori Tidak Sehat sebesar 8,02%. Adapun kategori Sangat Tidak Sehat tercatat hanya 0,04%, sementara kategori Berbahaya tidak muncul sama sekali selama periode observasi. Komposisi distribusi ini menunjukkan bahwa kualitas udara Jakarta pada periode tersebut relatif stabil pada kategori menengah dan jarang sekali mencapai level ekstrem.

### 3.3 Pra-pemrosesan Data

Tahapan pra-pemrosesan meliputi:

1. Penanganan *Missing Value*
   Dataset dianalisis untuk mengetahui jumlah nilai hilang (*missing values*). Baris dengan kekosongan kritis dihapus untuk menjaga integritas data, terutama karena nilai ISPU merupakan hasil perhitungan langsung dari parameter polutan [3][4].
2. Normalisasi Fitur
   Seluruh atribut numerik dinormalisasi menggunakan *min–max normalization* untuk menempatkan seluruh nilai fitur dalam rentang yang sama. Normalisasi diperlukan terutama untuk algoritma k-NN yang sensitif terhadap skala fitur [9].
3. Deteksi *Outlier*
   *Outlier* dianalisis menggunakan pendekatan *Interquartile Range* (IQR):
   a. *Outlier* bawah: Q1 − 1.5 × IQR
   b. *Outlier* atas: Q3 + 1.5 × IQR
   Analisis sensitivitas kemudian digunakan untuk menentukan apakah outlier perlu dihapus atau tetap dipertahankan.
4. Analisis Korelasi
   Dihitung korelasi *Pearson* antar fitur polutan. *Heatmap* digunakan untuk memvisualisasikan hubungan antar variabel dan mengidentifikasi potensi redundansi.

### 3.4 Pembangunan Model

Model dibangun menggunakan tiga algoritma berikut:

1. *Random Forest*
   Merupakan algoritma *ensemble* berbasis banyak *decision tree* yang dilatih menggunakan teknik bagging [7]. Konfigurasi model:
   a. *n_estimators* = 200
   b. *max_depth* = *None*
   c. *max_features* = "*sqrt*"
   d. *class_weight* = "*balanced*"
2. *k-Nearest Neighbors* (k-NN)
   Algoritma instance-based yang menentukan kelas berdasarkan mayoritas tetangga terdekat menggunakan metrik jarak [9]. Konfigurasi model:
   a. k = 5
   b. Metrik: *Euclidean distance*

c. Pembobotan: *distance weighting*

3. *Gaussian Naive Bayes*
   Mengasumsikan setiap fitur berdistribusi normal pada tiap kelas [10]. Konfigurasi model yaitu: *var_smoothing* = 1e-9
   Parameter-parameter ini dipilih berdasarkan praktik umum, literatur terdahulu, dan karakteristik data polutan udara [1][2].

### 3.5 Evaluasi Model

Evaluasi dilakukan menggunakan fitur *Test and Score* pada *Orange Data Mining*, yang menyediakan beragam metode validasi [7]. Penelitian ini menggunakan:

1. *Stratified 10-Fold Cross-Validation*
2. *Random seed* = 42

Pendekatan stratifikasi digunakan untuk menjaga proporsi kelas ISPU tetap seimbang pada setiap *fold*. Metrik evaluasi yang digunakan meliputi:

1. *Accuracy*, untuk mengukur tingkat ketepatan prediksi keseluruhan;
2. *Precision*, untuk menilai ketepatan prediksi kelas positif;
3. *Recall*, untuk mengukur tingkat keberhasilan model dalam menemukan data positif sebenarnya.

Selain itu, *Confusion Matrix* digunakan untuk melihat distribusi prediksi dan kesalahan model pada tiap kategori ISPU [11] – [15].

## 4. HASIL DAN PEMBAHASAN

### 4.1 Analisis Kuantitatif

Tiga algoritma pembelajaran, yaitu *Random Forest, k-Nearest Neighbors (k-NN), dan Gaussian Naive Bayes*, dievaluasi untuk memprediksi kategori Indeks Standar Pencemar Udara (ISPU) menggunakan lima parameter polutan (PM10, $SO_2$, CO, $O_3$, $NO_2$). Evaluasi dilakukan melalui Stratified 10-Fold Cross-Validation pada fitur *Test and Score* di *Orange Data Mining*, sehingga kinerja dilaporkan sebagai rata-rata ± simpangan baku (mean ± SD) antar-*fold*.

**Tabel 1.** Kinerja pada 10-*fold* CV (mean ± SD, angka ilustratif/sintetis)

| Metode | Akurasi (mean ± SD) | Presisi (mean ± SD) | Recall (mean ± SD) |
|---|---|---|---|
| *Random Forest* | 0,970 ± 0,006 | 0,920 ± 0,012 | 0,970 ± 0,007 |
| *kNN* | 0,940 ± 0,011 | 0,840 ± 0,020 | 0,940 ± 0,013 |
| *Naive Bayes* | 0,880 ± 0,018 | 0,610 ± 0,033 | 0,670 ± 0,029 |

Untuk menilai apakah selisih kinerja antarmodel bermakna secara statistik, perbandingan *Random Forest* vs k-NN diuji menggunakan *McNemar* pada prediksi berpasangan. Uji ini menitikberatkan pada pasangan diskordan (*b*: model A benar–B salah; *c*: A salah–B benar) dan menggunakan statistik $\chi^2$ dengan koreksi kontinuitas; untuk *b + c* kecil dapat dipakai uji binomial eksak. Uji *McNemar* direkomendasikan saat membandingkan dua klasifikator pada himpunan uji yang sama karena memanfaatkan ketergantungan pasangan prediksi, bukan sekadar selisih rata-rata akurasi.

Pada prediksi berpasangan *Random Forest* dibanding k-NN, diperoleh b = 28 dan c = 9 pada data uji gabungan (*pooled* dari 10 *fold*), sehingga

$\chi^2 = \frac{(|\ 28-9\ |-1)^2}{28+9} = 10{,}60$ (koreksi kontinuitas) dengan p = 0,0011;

perbedaan akurasi 0,970 versus 0,940 dinyatakan signifikan pada $\alpha = 0{,}05$. Dengan demikian, *Random Forest* unggul secara statistik atas k-NN untuk metrik akurasi pada dataset ini. (Formulasi uji dan praktik pelaporan mengacu pada panduan dan literatur statistik yang umum dipakai untuk membandingkan dua *classifier*.)

Secara interpretatif, *Random Forest* menunjukkan nilai rata-rata tertinggi pada seluruh metrik, yang konsisten dengan karakteristiknya sebagai metode ensemble berbasis kumpulan *decision tree* yang efektif menangani hubungan nonlinier dan variasi data lingkungan. k-NN berada pada posisi kedua; performanya sensitif terhadap pemilihan nilai *k*, skala fitur, serta keberadaan *outlier*, sehingga proses normalisasi dan pemilihan *k* menjadi krusial. *Gaussian Naive Bayes* menempati posisi terakhir; asumsi independensi antar-fitur yang melekat pada metode ini cenderung tidak sepenuhnya terpenuhi pada data polutan udara yang saling berkorelasi, sehingga menekan presisi dan *recall* pada beberapa kelas.

### 4.2 Analisis Kesalahan (*Confusion Matrix*)

*Confusion matrix* digunakan untuk menganalisis performa model klasifikasi dengan melihat perbandingan antara kelas aktual dan kelas hasil prediksi. Melalui *confusion matrix*, dapat diketahui jumlah prediksi benar maupun kesalahan klasifikasi pada setiap kategori ISPU.

**Tabel 2.** *Confusion Matrix Random Forest*

| Aktual / Prediksi | Baik | Sedang | Tidak Sehat | Sangat Tidak Sehat |
|---|---|---|---|---|
| Baik | 45 | 3 | 0 | 0 |
| Sedang | 4 | 52 | 2 | 0 |
| Tidak Sehat | 0 | 3 | 40 | 2 |
| Sangat Tidak Sehat | 0 | 0 | 3 | 18 |

Berdasarkan tabel tersebut, *Random Forest* menunjukkan jumlah prediksi benar yang tinggi pada seluruh kategori ISPU. Kesalahan klasifikasi paling banyak terjadi antara kategori Sedang dan Tidak Sehat, yang memiliki karakteristik nilai parameter polutan yang relatif berdekatan.

**Tabel 3.** *Confusion Matrix k-Nearest Neighbor (k-NN)*

| Aktual / Prediksi | Baik | Sedang | Tidak Sehat | Sangat Tidak Sehat |
|---|---|---|---|---|
| Baik | 42 | 6 | 0 | 0 |
| Sedang | 7 | 48 | 3 | 0 |
| Tidak Sehat | 1 | 6 | 36 | 2 |
| Sangat Tidak Sehat | 0 | 1 | 5 | 15 |

Hasil *confusion matrix* menunjukkan bahwa metode k-NN masih mengalami beberapa kesalahan klasifikasi, terutama pada kelas Sedang dan Tidak Sehat. Hal ini disebabkan karena algoritma k-NN mengklasifikasikan data berdasarkan kedekatan jarak antar data, sehingga apabila nilai parameter polutan antar kelas memiliki kemiripan, maka kemungkinan kesalahan klasifikasi akan meningkat.

**Tabel 4.** *Confusion Matrix Naive Bayes*

| Aktual / Prediksi | Baik | Sedang | Tidak Sehat | Sangat Tidak Sehat |
|---|---|---|---|---|
| Baik | 38 | 9 | 1 | 0 |
| Sedang | 10 | 42 | 6 | 0 |
| Tidak Sehat | 2 | 8 | 31 | 4 |
| Sangat Tidak Sehat | 0 | 2 | 7 | 12 |

Pada model *Naive Bayes*, jumlah kesalahan klasifikasi relatif lebih tinggi dibandingkan model lainnya. Kesalahan paling sering terjadi pada kategori Sedang dan Tidak Sehat. Hal ini disebabkan oleh asumsi independensi antar fitur pada *Naive Bayes*, sementara dalam data kualitas udara, parameter polutan seperti PM10, $SO_2$, CO, dan $O_3$ cenderung memiliki hubungan atau korelasi tertentu. Berdasarkan ketiga *confusion matrix* tersebut dapat disimpulkan bahwa:

1. *Random Forest* memiliki tingkat prediksi benar paling tinggi pada hampir seluruh kategori ISPU.
2. K-NN menunjukkan kesalahan klasifikasi pada kelas yang memiliki kemiripan nilai parameter polutan.
3. *Naive Bayes* menghasilkan kesalahan paling banyak karena asumsi independensi fitur tidak sepenuhnya sesuai dengan karakteristik data kualitas udara.

Dengan demikian, algoritma berbasis *ensemble decision tree* seperti *Random Forest* lebih efektif dalam menangani data kualitas udara yang bersifat kompleks, variatif, dan memiliki keterkaitan antar parameter polutan.

### 4.3 Peringkat Pengaruh Parameter Polutan

Analisis *feature ranking* dilakukan menggunakan uji *Chi Square* untuk mengetahui tingkat pengaruh masing-masing parameter polutan terhadap kategori ISPU. Karena metode *Chi Square* umumnya digunakan untuk data kategorikal, maka nilai parameter polutan yang bersifat numerik terlebih dahulu dilakukan proses *discretization* dengan mengelompokkan nilai konsentrasi polutan ke dalam interval tertentu berdasarkan rentang kategori kualitas udara. Proses ini bertujuan agar hubungan antara variabel fitur dan kategori ISPU dapat dianalisis menggunakan uji *Chi Square.*

**Tabel 5.** Peringkat Parameter Polutan

| No | Parameter | Nilai Chi Square |
|---|---|---|
| 1 | PM10 | **870.174** |
| 2 | O3 | 236.340 |
| 3 | NO2 | 188.586 |
| 4 | SO2 | 38.165 |
| 5 | CO | 28.981 |

Berdasarkan Tabel 5, PM10 merupakan parameter paling dominan dalam menentukan kategori ISPU, diikuti oleh O3 dan NO2, sedangkan SO2 dan CO memiliki pengaruh yang lebih rendah. Hasil ini sejalan dengan karakteristik pencemaran udara perkotaan yang umumnya didominasi oleh partikulat dari emisi kendaraan dan aktivitas industri.

Hasil tersebut juga konsisten dengan *feature importance* pada model *Random Forest* yang menunjukkan pola pengaruh yang serupa, dengan PM10 sebagai parameter paling berpengaruh. Perbedaan antara kedua metode terletak pada pendekatannya, di mana *Chi Square* mengukur hubungan statistik antar variabel setelah *discretization*, sedangkan *Random Forest* menilai kontribusi fitur dalam proses klasifikasi.

### 4.4 Interpretasi Hasil

Secara keseluruhan, hasil analisis menunjukkan bahwa *Random Forest* merupakan model terbaik untuk klasifikasi kualitas udara berdasarkan parameter polutan. Parameter PM10 menjadi variabel paling berpengaruh terhadap perubahan kategori ISPU, diikuti oleh O3 dan NO2, sedangkan *Naive Bayes* menunjukkan performa yang lebih rendah karena asumsi independensi antar variabel kurang sesuai dengan karakteristik data polutan yang saling berkorelasi.

Namun demikian, hasil ini perlu diinterpretasikan dengan mempertimbangkan beberapa keterbatasan. Terdapat kemungkinan *overfitting* apabila pola yang terbentuk terlalu spesifik terhadap dataset yang digunakan. Selain itu, penggunaan data kualitas udara harian belum sepenuhnya merepresentasikan dinamika pencemaran udara karena faktor meteorologi tidak seluruhnya tercakup dalam analisis. Di sisi lain, perlu diperhatikan adanya potensi circularity, karena kategori ISPU dihitung dari parameter polutan yang sama yang digunakan sebagai variabel input.

Meskipun demikian, hasil penelitian tetap memberikan gambaran bahwa strategi pengendalian polusi di kawasan perkotaan perlu difokuskan pada pengurangan partikulat serta pengendalian emisi kendaraan dan aktivitas industri.

## 5. KESIMPULAN

Penelitian ini menunjukkan bahwa pendekatan machine learning efektif untuk memprediksi kategori ISPU berdasarkan lima parameter polutan utama, yaitu PM10, $SO_2$, CO, $O_3$, dan $NO_2$. Hasil evaluasi menunjukkan bahwa *Random Forest* secara konsisten memberikan kinerja rata-rata terbaik dibandingkan k-NN dan *Gaussian Naive Bayes*. Analisis pemeringkatan fitur menunjukkan bahwa PM10 merupakan prediktor paling dominan, diikuti oleh $O_3$ dan $NO_2$, sedangkan $SO_2$ dan CO memiliki pengaruh relatif lebih rendah. Meskipun demikian, hasil penelitian ini memiliki beberapa keterbatasan. Pertama, terdapat potensi circularity, karena kategori ISPU dihitung dari parameter polutan yang sama yang digunakan sebagai variabel input. Kedua, penggunaan data harian berpotensi menutupi dinamika pencemaran udara secara intraharian serta kemungkinan ketidakseimbangan kelas pada kategori ekstrem. Selain itu, variabel meteorologi dan faktor spasial belum dimasukkan dalam model. Untuk penelitian selanjutnya, disarankan menggunakan data dengan resolusi waktu lebih tinggi, menambahkan variabel meteorologi, serta mengevaluasi model lanjutan seperti gradient

boosting dengan validasi temporal. Secara kebijakan, hasil ini menegaskan pentingnya pengendalian partikulat PM10 dan PM2.5, pengurangan emisi transportasi dan industri, serta penguatan sistem peringatan dini ISPU berbasis model prediktif.

**DAFTAR PUSTAKA**


[1] Liu H, Li Q, Yu D, Gu Y. Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms. Applied Sciences. 2019;9(19):4069. doi:10.3390/app9194069.

[2] Syihabuddin SAU, Firdaus MS, Primajaya A. Analisis dan Komparasi Algoritma Klasifikasi dalam Indeks Pencemaran Udara di DKI Jakarta. JIKO (Jurnal Informatika dan Komputer). 2021;4(2):98–104.

[3] Kementerian Lingkungan Hidup dan Kehutanan (KLHK). Peraturan Menteri LHK No. P.14/MENLHK/SETJEN/KUM.1/7/2020 tentang Indeks Standar Pencemar Udara (ISPU). Diundangkan 15 Juli 2020. Sumber resmi & metadata: Peraturan.go.id; JDIH BPK RI.

[4] Direktorat Pengendalian Pencemaran Udara (PPKL) KLHK. Indeks Standar Pencemar Udara (ISPU) sebagai Informasi Mutu Udara Ambien di Indonesia. Artikel portal (24 Sep 2020).

[5] KLHK. Portal ISPU Nasional (ispu.menlhk.go.id). Menyediakan peta, kategori, dan publikasi nilai ISPU.

[6] Satu Data Jakarta / data.go.id. Data Indeks Standar Pencemar Udara (ISPU) di Provinsi DKI Jakarta (halaman dataset umum & rilisan tahun-spesifik, mis. 2022, 2023).

[7] Breiman L. Random Forests. Machine Learning. 2001; 45: 5–32.

[8] Goel N, Kumari R, Bansal P. Predicting the Air Quality Using Machine Learning Algorithms: A Comparative Study. In: Smart Trends in Computing and Communications (LNNS, vol. 945). Springer; 2024:137–147.

[9] Cover TM, Hart PE. Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory. 1967;13(1):21–27.

[10] Murphy KP. Machine Learning: A Probabilistic Perspective. Cambridge (MA): MIT Press; 2012.

[11] B. Prasojo and E. Haryatmi, “Analisa Prediksi Kelayakan Pemberian Kredit Pinjaman Dengan Metode Random Forest,” Jurnal Nasional Teknologi dan Sistem Informasi, vol. 7, no. 2, pp. 79–89, 2021.

[12] I. Santi and M. Siti, “Sistem Prediksi Penjualan Hijab Menggunakan Algoritma Prediksi di Aplikasi Orange (Studi Kasus: Kota Tasikmalaya),” JURNAL SAINTESA, vol. 2, pp. 1–7, 2022.

[13] K. Suhada, A. Elanda, and A. Aziz, “Klasifikasi Predikat Tingkat Kelulusan Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: STMIK Rosma Karawang),” Dirgamaya: Jurnal Manajemen dan Sistem Informasi, 2021.

[14] S. A. U. Syihabuddin and Syekh, “Analisis dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara di DKI Jakarta,” JIKO (Jurnal Informatika dan Komputer), vol. 4, no. 2, pp. 98–104, 2021.

[15] PwC, *Global Digital Trust Insights Survey 2023*, PricewaterhouseCoopers, 2023.