Penerapan Algoritma *Machine Learning* untuk Mengklasifikasikan Polusi Udara di Wilayah DKI Jakarta

1st Muhammad Arya Suhendi
Universitas Buana Perjuangan
Karawang, Indonesia
if21.muhammdsuhendi@mhs.ubpkarawang.ac.id

2rd Tatang Rohana
Universitas Buana Perjuangan
Karawang, Indonesia
tatang.rohana@ubpkarawang.ac.id

3rd Jamaludin Indra
Universitas Buana Perjuangan
Karawang, Indonesia
jamaludin.indra@ubpkarawang.ac.id

4th Ayu Ratna Juwita
Universitas Buana Perjuangan
Karawang, Indonesia
ayurj@ubpkarawang.ac.id

Abstract— Polusi udara di DKI Jakarta merupakan masalah serius, dengan tingkat polusi tertinggi di Asia Tenggara. Sumber utamanya berasal dari transportasi, industri, dan pembakaran sampah. Keterbatasan sistem pemantauan konvensional mendorong pemanfaatan kecerdasan buatan, khususnya algoritma machine learning, untuk meningkatkan akurasi klasifikasi kualitas udara. Penelitian ini membandingkan performa empat algoritma Support Vector Machine (SVM), Gradient Boosting, Random Forest, dan Decision Tree dalam mengklasifikasikan tingkat polusi udara di Jakarta. Dataset yang digunakan terdiri dari 1.675 data Indeks Standar Pencemar Udara (ISPU) yang diperoleh dari Dinas Lingkungan Hidup Jakarta selama periode Januari hingga November 2024, dengan parameter meliputi PM10, PM2.5, SO2, CO, O3, dan NO2. Proses penelitian mencakup tahapan pembersihan data, normalisasi, reduksi dimensi menggunakan Principal Component Analysis, pembangunan model melalui pembagian data latih dan uji (80:20), serta evaluasi performa menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil evaluasi menunjukkan bahwa seluruh algoritma memberikan tingkat akurasi yang tinggi, dengan Random Forest mencapai performa terbaik 93,71%, diikuti Decision Tree 93,41%, Gradient Boosting 92,81% serta SVM 92,51%. Temuan ini mendukung penerapan machine learning sebagai solusi pemantauan polusi udara yang lebih efektif di Jakarta.

Kata kunci — Polusi Udara, Jakarta, *Machine learning*, ISPU, Klasifikasi

I. PENDAHULUAN

Polusi udara merupakan penurunan kualitas udara akibat zat atau partikel berbahaya yang berdampak negatif terhadap kesehatan dan lingkungan. DKI Jakarta termasuk kota dengan tingkat polusi tertinggi di Asia Tenggara, dengan konsentrasi rata-rata PM2.5 mencapai 43,8 µg/m³ pada 2023, jauh melebihi ambang batas *WHO* sebesar 5 µg/m³ [1]. Sumber utamanya mencakup emisi kendaraan, aktivitas industri, pembakaran sampah, serta faktor alam seperti kebakaran hutan [2].

Transportasi menyumbang sekitar 75% total emisi polutan, diikuti industri (20%) dan aktivitas domestik (5%) (Listyarini et al., 2023). Jumlah kendaraan bermotor yang mencapai 20,5 juta unit pada 2023 memperparah kondisi tersebut (Rachmayani, 2023). Selain dampak terhadap kesehatan masyarakat, polusi udara juga menimbulkan kerugian ekonomi hingga Rp46,8 triliun per tahun [4].

Sistem pemantauan kualitas udara yang ada dinilai belum mampu memberikan hasil yang akurat dan *real-time*, sehingga memerlukan pendekatan baru yang lebih adaptif. Teknologi kecerdasan buatan, khususnya algoritma *machine learning*, telah terbukti efektif dalam menganalisis data lingkungan secara cepat dan akurat. Berbagai algoritma seperti *Support Vector Machine (SVM), Gradient Boosting, Random Forest*, dan *Decision Tree* menunjukkan kinerja menjanjikan dalam klasifikasi kualitas udara [5].

Data Indeks Standar Pencemaran Udara (ISPU), yang berlokasi di Jalan Mandala V No.67, Cililitan, Kramat Jati, Jakarta Timur, Daerah Khusus Ibukota Jakarta 13640, menjadi sumber penelitian ini. Di kota Jakarta, dinas lingkungan hidup bertanggung jawab untuk menjaga kesehatan lingkungan masyarakat. Penelitian tersebut dilakukan di Laboratorium Riset UBP Karawang. Tujuan dari penelitian ini adalah untuk menganalisis dan membandingkan kinerja empat algoritma pembelajaran mesin (*SVM, Gradient Boosting, Random Forest*, dan *Decision Tree*) dalam klasifikasi kualitas udara di Jakarta. Ada evaluasi yang dilakukan untuk menemukan algoritma terbaik berdasarkan akurasi, presisi, recall, dan skor F1. Penelitian ini diharapkan dapat berfungsi sebagai dasar untuk pengembangan sistem pemantauan udara yang lebih efisien dan membantu proses membuat kebijakan yang tepat untuk mengurangi polusi.

II. TINJAUAN PUSTAKA

A. *Polusi Udara*

Transportasi, bisnis, dan pembakaran sampah menyebabkan polusi udara yang signifikan di kota Jakarta.[6]. PM2.5, NO_2, dan O_3 terbukti memicu penyakit paru dan jantung [7]. Kualitas udara diklasifikasi dari “sehat” hingga “berbahaya” menurut standar WHO.

B. *Machine Learning*

Komputer dapat belajar dari data untuk klasifikasi melalui metode *machine learning,* yang mencakup *supervised, unsupervised,* dan lainnya [8]. Untuk hasil optimal, algoritma harus sesuai dengan karakteristik dataset.

C. *Algoritma Klasifikasi*

- *SVM* merupakan algoritma *supervised learning* yang mencari *hyperplane* optimal untuk memisahkan kelas data [9] . *SVM* efektif dalam menangani data berdimensi tinggi dan *non-linear*. Dalam konteks kualitas udara, *SVM* mampu mengklasifikasikan data pencemar seperti PM10, NO_2, dan CO secara akurat [10].

- *Gradient Boosting* membangun model secara bertahap untuk meminimalkan kesalahan prediksi menggunakan metode *gradient descent* [11]. Algoritma ini unggul dalam menangani data *non-linear* dan kompleks, serta mampu mencapai akurasi tinggi dalam klasifikasi polusi udara [12].

- *Random Forest (RF)* adalah metode *ensemble* yang menggunakan teknik *bagging* untuk menggabungkan berbagai pilihan pohon untuk meningkatkan akurasi dan mengurangi *overfitting*. [13]. *RF* terbukti akurat dalam klasifikasi polusi udara, dengan studi menunjukkan akurasi lebih dari 99% dalam klasifikasi ISPU Jakarta [14].

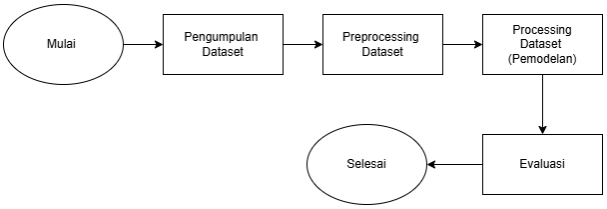
- Decision Tree menghasilkan model dalam bentuk struktur pohon yang mudah diinterpretasi dan cocok untuk data kategorikal dan numerik [15]. Algoritma ini telah berhasil digunakan dalam klasifikasi kualitas udara dengan tingkat akurasi mendekati 100% di berbagai studi, termasuk di Yogyakarta dan Klang [16].

D. *Confusion Matrix*

Metode evaluasi performa klasifikasi yang dikenal sebagai *confusion matrix* berbentuk tabel yang mencatat hasil prediksi model terhadap data aktual. Matriks ini memuat nilai *True Positive, False Positive, True Negative*, dan *False Negative* yang digunakan untuk menghitung metrik evaluasi seperti akurasi, presisi, *recall*, dan *F1-score* [17].

III. METODE PENELITIAN

Tahapan dalam penelitian ini dijalankan secara tersruktur guna mencapai hasil dan tujuan yang telah ditetapkan. Berikut merupakan tahapan proses penelitian ini:



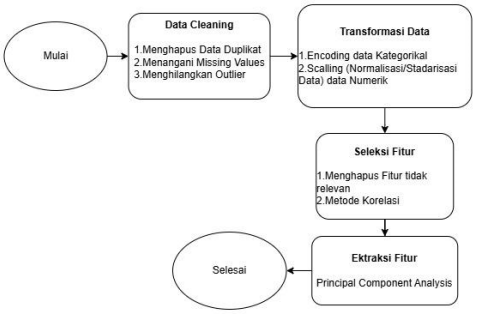
Gambar 1. Prosedur Penelitian

A. Pengumpulan *Dataset*

Data yang digunakan adalah data pengamatan Indeks standar pencemaran udara (ISPU) pada bulan Januari sampai November tahun 2024 yang diperoleh dari website Dinas Lingkungan Hidup kota Jakarta. Secara umum, dataset ini berisi 1675 data yang dicatat terdiri dari 11 atribut dan 1 kelas. Data ISPU yang diperoleh dari berbagai stasiun pemantau kualitas udara. Parameter yang digunakan dalam pengumpulan data mencakup konsentrasi *PM10*,*PM2.5* SO2, CO, O3, dan *NO2* yang merupakan komponen fundamental dalam penentuan nilai ISPU Pengumpulan *dataset.*

B. *Preprocessing Dataset*

Tahapan Tahap preprocessing dan analisis data merupakan bagian penting dalam menyiapkan data sebelum dilakukan pemodelan. Proses ini diawali dengan analisis data, setelah itu dilakukan pembersihan data *(data cleaning)*, selanjutnya data ditransformasikan, kemudian dinormalisasi menggunakan *MinMaxScaler* pada data numerik. Selain itu, dilakukan juga ekstraksi atau seleksi fitur, salah satunya menggunakan *PCA (Principal Component Analysis),* guna mereduksi dimensi dan meningkatkan efisiensi komputasi. Tahap ini bertujuan agar data bersih, terstruktur, dan optimal untuk menghasilkan model klasifikasi yang akurat dan andal.



Gambar 2. *Preprocessing Dataset*

C. Pemodelan

Dataset ISPU dibagi menjadi data latihan (*training data)* dan data uji (*testing data*) pada tahap pemrosesan data. Ada 1340 data latihan dan 335 data uji, masing-masing dengan rasio 80:20. Data uji digunakan untuk mengevaluasi akurasi prediksi, sedangkan data latih digunakan untuk melatih model. Selanjutnya, pelatihan model dilakukan dengan menggunakan empat algoritma pembelajaran mesin: *Support Vector Machine (SVM), Random Forest, Gradient Boosting*, dan *Decision Tree*. Setiap algoritma dilatih untuk mengidentifikasi pola hubungan antara parameter pencemar udara dan kategori ISPU menggunakan data latihan. Setelah model dibuat, data uji dimasukkan ke dalamnya untuk melakukan prediksi; setelah itu, hasil prediksi dicatat dan disiapkan untuk dievaluasi.Tahap ini sangat penting karena menentukan seberapa baik model yang dibangun mampu menggeneralisasi data baru dan memberikan hasil klasifikasi yang akurat.

D. Evaluasi Model

Evaluasi performa model dilakukan menggunakan confusion matrix yang menghasilkan metrix berupa *accuracy, precision, recall,* dan *F1-score.* Hasil evaluasi menunjukkan efektivitas masing-masing algoritma dalam mengklasifikasikan tingkat pencemaran udara. Analisis komprehensif terhadap hasil evaluasi memberikan *insight* mengenai kelebihan dan limitasi setiap model yang diimplementasikan. Evaluasi dan validasi hasil hitung menggunakan rumus akurasi, *precision*, *recall* dan *f-measure*.

IV. HASIL PEMBAHASAN

A. Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari Dinas Lingkungan Hidup DKI Jakarta berupa Indeks Standar Pencemar Udara (ISPU) periode Januari hingga November 2024 dengan total 1675 data dan 12 *attribute*, mencakup parameter utama seperti *PM10, PM2.5, CO, SO2, NO2*, dan *O3* yang menjadi indikator tingkat pencemaran udara. Proses pengumpulan dilakukan secara daring melalui portal resmi pemerintah dan disimpan dalam format *CSV* https://satudata.jakarta.go.id/open-data/data-indeks-standar-pencemar-udara-ispu-di-provinsi-dki-jakarta. Data dibaca menggunakan perintah *pd.read_csv* ('nama_file.csv'). Perintah ini akan memuat seluruh isi file *CSV* ke dalam bentuk *Data Frame*, yaitu struktur data dua dimensi yang memudahkan manipulasi dan analisis data. Setelah data dimuat, ditampilkan beberapa baris pertama menggunakan *df*.

	periode_data	tanggal	stasiun	pm_sepuluh	pm_duakomalima	sulfur_dioksida	karbon_monoksida	ozon	nitrogen_dioksida	max	parameter_pencemar_kritis	kategori
0	202401	2024-01-21	DKI3 Jagakarsa	51.0	65.0	45.0	9.0	8.0	79.0	79.0	NaN	SEDANG
1	202401	2024-01-22	DKI3 Jagakarsa	27.0	34.0	45.0	5.0	8.0	56.0	56.0	NaN	SEDANG
2	202401	2024-01-23	DKI3 Jagakarsa	NaN	52.0	46.0	6.0	9.0	51.0	52.0	PM25	SEDANG
3	202401	2024-01-24	DKI3 Jagakarsa	46.0	65.0	46.0	8.0	9.0	38.0	65.0	PM25	SEDANG
4	202401	2024-01-25	DKI3 Jagakarsa	37.0	55.0	47.0	7.0	11.0	28.0	55.0	PM25	SEDANG
...	...	...	...	...	...	...	...	...	...	...	...	...
1670	202410	2024-10-10	DKI4 Lubang Buaya	58.0	82.0	53.0	20.0	41.0	20.0	82.0	PM25	SEDANG
1671	202410	2024-10-11	DKI4 Lubang Buaya	64.0	66.0	53.0	23.0	34.0	26.0	66.0	PM25	SEDANG
1672	202410	2024-10-12	DKI4 Lubang Buaya	58.0	64.0	54.0	21.0	33.0	25.0	64.0	PM25	SEDANG
1673	202410	2024-10-13	DKI4 Lubang Buaya	59.0	60.0	53.0	21.0	36.0	20.0	60.0	PM25	SEDANG
1674	202410	2024-10-14	DKI4 Lubang Buaya	78.0	76.0	53.0	29.0	43.0	30.0	78.0	PM10	SEDANG

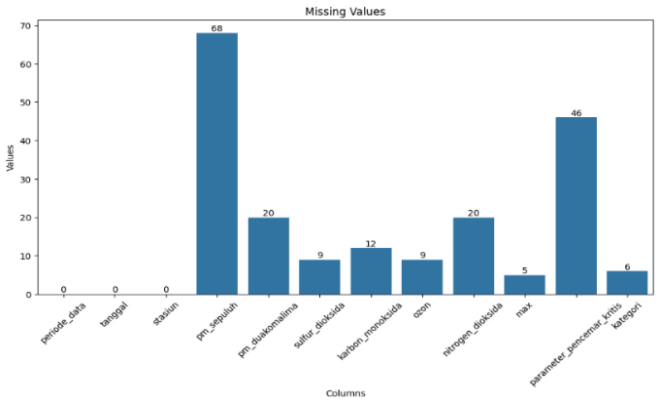
1675 rows × 12 columns

Gambar 3. Pengumpulan Data

B. *Preprocessing*

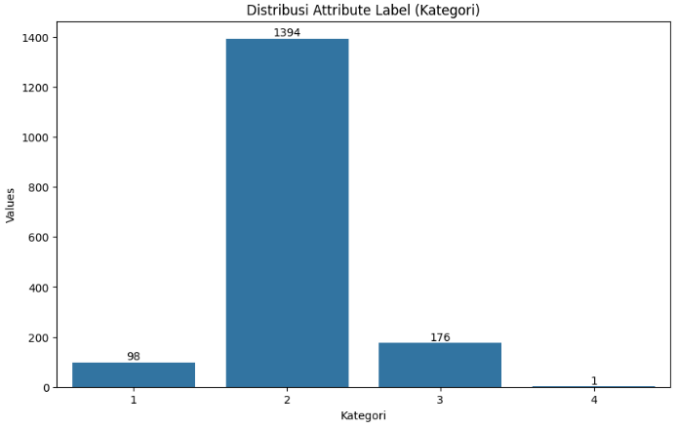
Selanjutnya, dilakukan pemeriksaan terhadap keberadaan *missing value*. Untuk mengetahui sejauh mana kelengkapan data, dilakukan visualisasi terhadap jumlah nilai hilang di setiap kolom. Berdasarkan hasil eksplorasi data pada Gambar 4 ditemukan bahwa beberapa kolom numerik *seperti pm_sepuluh, pm_duakomalima, sulfur_dioksida, karbon_monoksida, ozon, nitrogen_dioksida,* dan *max* memiliki nilai yang tidak lengkap. Oleh karena itu, dilakukan imputasi menggunakan strategi *mean* (rata-rata) untuk mengisi nilai yang hilang pada numerik. Untuk menangani nilai hilang dalam data, langkah pertama adalah mengidentifikasi kolom numerik yang memiliki *missing value*, seperti *pm_sepuluh* dan *sulfur_dioksida*. Digunakan *SimpleImputer* dari *sklearn.impute* dengan strategi *mean* untuk mengganti nilai kosong dengan rata-rata kolom. Proses ini

dilakukan menggunakan *it_transform* dan hasilnya disimpan kembali ke *DataFrame*. Sementara itu, untuk kolom kategorikal seperti kategori dan stasiun, digunakan strategi *most_frequent* untuk mengganti nilai kosong dengan nilai yang paling sering muncul.



Gambar 4. Grafik Missing Value

Langkah selanjutnya transformasi data yaitu mengubah data bertipe *object* menjadi data dalam bentuk angka. Gambar 5 menunjukkan visualisasi mengubah nilai 'BAIK' menjadi 1, 'SEDANG' 2, 'TIDAK SEHAT' 3, 'SANGAT TIDAK SEHAT' 4 dengan jumlah data perlabel.



Gambar 5. Tranformasi Data Label

Setelah memahami distribusi data dan dinamika kondisi udara sepanjang tahun, proses korelasi *heatmap* dihitung menggunakan metode *Pearson* preprocessing. Metode *Pearson correlation coefficient*, digunakan untuk mengukur kekuatan dan arah hubungan linier antara dua variabel numerik. Pada Gambar 6, perhitungan ini dilakukan secara otomatis oleh fungsi *.corr()* dari pustaka *pandas*. Secara matematis, koefisien korelasi *Pearson* r antara dua variabel X dan Y dihitung pada persamaan (1) :

$$r = \frac{\sum(X_i - \underline{X})(Y_i - \underline{Y})}{\sqrt{\sum (X_i - \underline{X})^2 \sum(Y_i - Y_i)^2}} \quad (1)$$

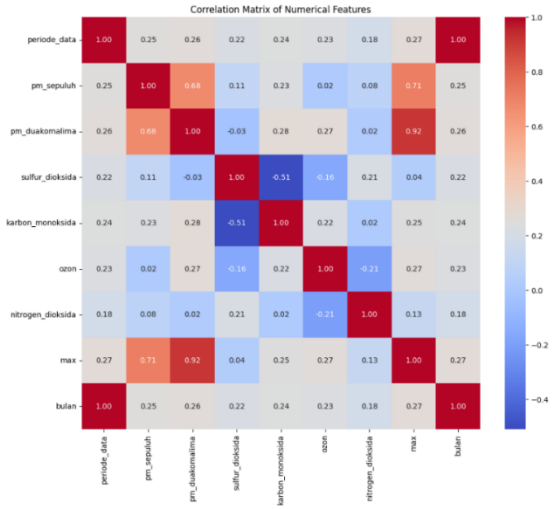
Di mana :

- X_i dan Y_i adalah nilai dari masing-masing observasi
- $\underline{X}$ dan Y adalah rata-rata dari masing_masing variabel

Nilai r yang dihasilkan berada dalam rentang -1 dan 1 :

- r = 1: hubungan positif sempurna,
- r = -1: hubungan negartif sempurna,
- r = 0: tidak ada hubungan linier

ISSN: 2715-2766
Vol. 7 No. 1, Januari 2026



Gambar 6 Heatmap Korelasi

Data preprocessing dilakukan untuk mempersiapkan data agar siap digunakan dalam pemodelan. Selanjutnya dilakukan normalisasi data menggunakan metode *MinMax Scaler*. Tujuannya adalah menyetarakan skala seluruh fitur numerik dalam rentang 0 hingga 1. Proses ini diterapkan pada semua kolom bertipe numerik, kecuali kolom target klasifikasi (*numeric_kategori*). Transformasi dilakukan dengan fungsi *fit_transform()* dari pustaka *scikit-learn*, sehingga semua fitur numerik memiliki skala yang seragam.

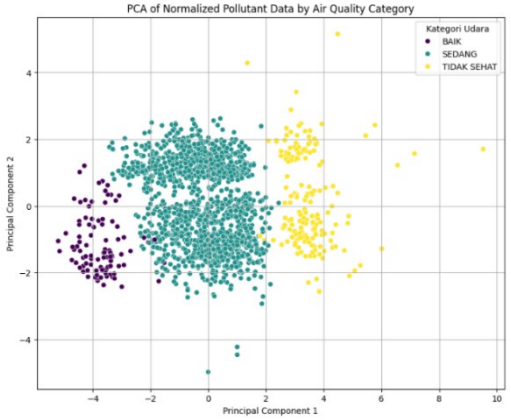
```
   periode_data    tanggal        stasiun  pm_sepuluh  pm_duakomalima  \
0           0.0 2024-01-21  DKI3 Jagakarsa    0.213873        0.374150
1           0.0 2024-01-22  DKI3 Jagakarsa    0.075145        0.163265
2           0.0 2024-01-23  DKI3 Jagakarsa    0.236994        0.285714
3           0.0 2024-01-24  DKI3 Jagakarsa    0.184971        0.374150
4           0.0 2024-01-25  DKI3 Jagakarsa    0.132948        0.306122

   sulfur_dioksida  karbon_monoksida      ozon  nitrogen_dioksida       max  \
0         0.361905          0.089552  0.044643              0.385  0.297143
1         0.361905          0.029851  0.044643              0.270  0.165714
2         0.371429          0.044776  0.053571              0.245  0.142857
3         0.371429          0.074627  0.053571              0.180  0.217143
4         0.380952          0.059701  0.071429              0.130  0.160000

  parameter_pencemar_kritis kategori  numeric_kategori  bulan
0                      PM25   SEDANG                 2      1
1                      PM25   SEDANG                 2      1
2                      PM25   SEDANG                 2      1
3                      PM25   SEDANG                 2      1
4                      PM25   SEDANG                 2      1
```

Gambar 7. Normalisasi Data

Selanjutnya, dilakukan *ekstraksi fitur menggunakan Principal Component Analysis (PCA)* untuk mengurangi dimensi data. *PCA* bertujuan untuk menyederhanakan data berdimensi tinggi tanpa mengurangi informasi signifikan. Proses dimulai dengan mengambil kolom numerik dari data yang telah dibersihkan dan menyimpannya dalam variabel X, sementara kolom target disimpan dalam variabel y. Data kemudian dinormalisasi menggunakan *MinMaxScaler* untuk memperoleh distribusi dengan rata-rata 0 dan standar deviasi 1. *PCA* kemudian diterapkan dengan parameter *n_components=2*, menghasilkan dua komponen utama (*PC1* dan *PC2*). Persentase variasi data yang dijelaskan oleh masing-masing komponen ditampilkan untuk mengukur kontribusi informasi yang dipertahankan. Hasil dari *PCA* divisualisasikan dalam bentuk *scatter plot*, dengan pewarnaan berdasarkan kategori kualitas udara, yang membantu untuk mengidentifikasi pola keterpisahan antar kategori kualitas udara.



Gambar 8. *Scatter Plot PCA*

C. Hasil Pemodelan

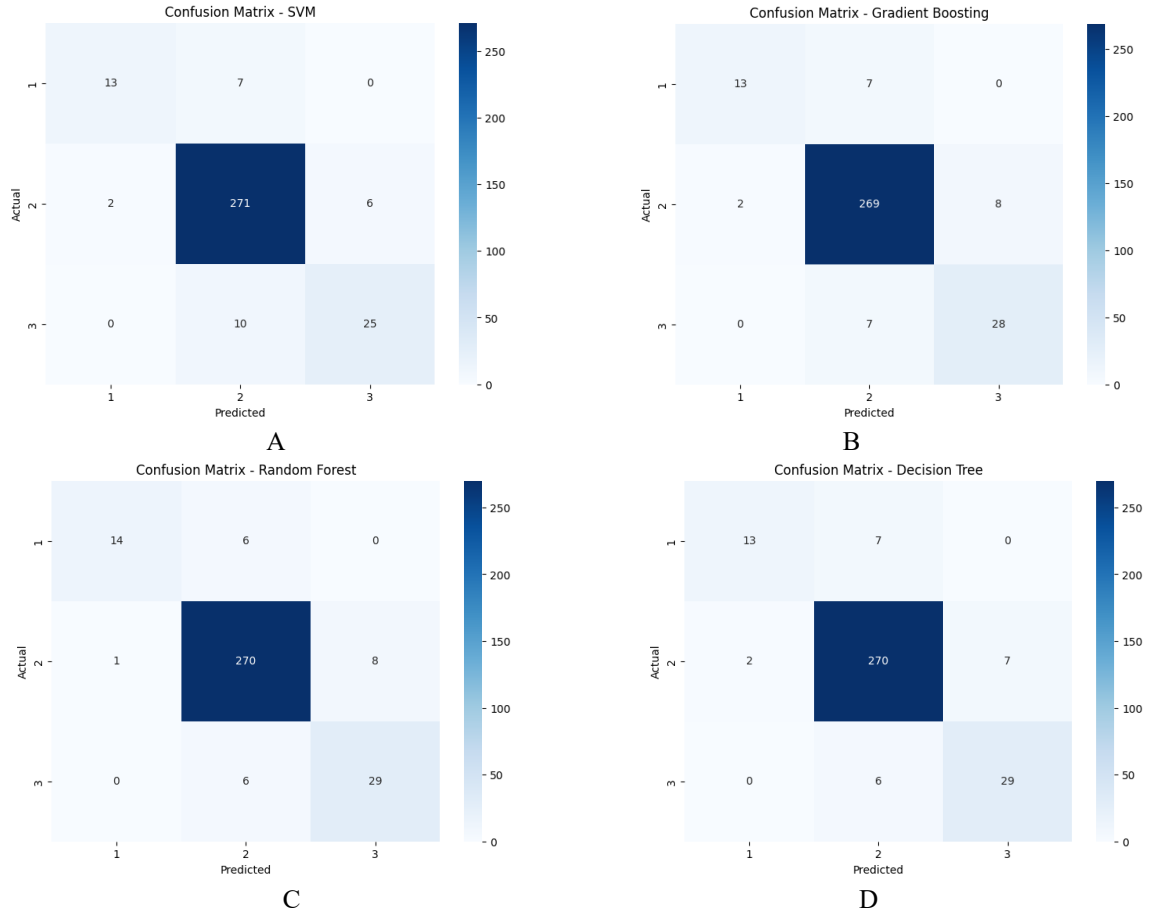
Pemodelan Pada penelitian ini, empat algoritma *machine learning*, yaitu *Support Vector Machine (SVM), Gradient Boosting, Random Forest*, dan *Decision Tree*, diuji untuk mengklasifikasikan kualitas udara di Jakarta setelah dilakukan reduksi dimensi menggunakan *Principal Component Analysis (PCA)*.

- ***Support Vector Machine (SVM)***: Setelah pembagian dataset menjadi 80% data latih dan 20% data uji, model *SVM* diinisialisasi dengan kernel linear dan parameter regulasi C=1. Model ini kemudian dilatih pada data latih dan diuji dengan data uji. Hasil evaluasi menunjukkan akurasi 92,5%, mengindikasikan klasifikasi dengan ketepatan yang tinggi
- ***Gradient Boosting***: Model *Gradient Boosting* diinisialisasi dengan *n_estimators=100* dan *learning_rate=0.1*. Setelah pelatihan, model diuji dan dievaluasi dengan metrik akurasi, *precision, recall*, dan *F1-score*. Hasilnya menunjukkan akurasi 92,81%, *precision* 0,86, *recall* 0,80, dan *F1-score* 0,83. *Confusion Matrix* menunjukkan performa klasifikasi yang baik.
- ***Random Forest***: Model *Random Forest* diinisialisasi dengan parameter *n_estimators=100* dan pembagian data latih dan uji dengan proporsi 80:20. Setelah pelatihan, model menghasilkan akurasi 93,71%, dengan *F1-score* yang sangat baik di setiap kelas. Model ini menunjukkan kestabilan dan keseimbangan yang sangat baik dalam klasifikasi
- ***Decision Tree***:Model *Decision Tree* diinisialisasi dengan *random_state=42* dan dilatih pada data latih. Hasil evaluasi menunjukkan akurasi 93,41%, menandakan bahwa penggunaan *PCA* dapat menyederhanakan data tanpa mengurangi akurasi. Model ini efektif dalam mengklasifikasikan data setelah reduksi dimensi

Secara keseluruhan, semua algoritma yang diuji menunjukkan performa yang baik, dengan *Random Forest* mencapai akurasi tertinggi (93,71%) diikuti oleh *Decision Tree* (93,41%), *Gradient Boosting* (92,81%), dan *SVM* (92,5%). Hasil ini menunjukkan bahwa model yang telah dilatih menggunakan *PCA* dapat mengklasifikasikan kualitas udara di Jakarta secara akurat.

D. *Evaluasi* Model dengan *Confusion Matrix*

Langkah selanjutnya adalah mengevaluasi performa model terhadap data uji (test set) menggunakan *confusion matrix*. Evaluasi ini bertujuan untuk mengukur sejauh mana model dapat melakukan generalisasi terhadap data baru yang tidak pernah dilihat sebelumnya selama proses pelatihan. *Confusion matrix* menghasilkan gambaran visual mengenai bagaimana model mengklasifikasikan data, dengan menunjukkan hasil dari jumlah prediksi yang benar dan salah dalam masing-masing kelas.



Gambar 9. Hasil *Confusion Matrix*

Berdasarkan Gambar A, berikut keterangannya:

Tabel 1. Evaluasi Manual per kelas pada *SVM*

Kelas	TP	FP	FN	Precision	Recall	F1-Score
1(Rendah)	13	2	7	13 / (13 + 2) = 0.87	13 / (13 + 7) = 0.65	2 × (0.87 × 0.65) / (0.87 + 0.65) = 0.74
2(Sedang)	271	17	8	271 / (271 + 17) = 0.94	271 / (271 + 8) = 0.97	2 × (0.94 × 0.97) / (0.94 + 0.97) = 0.96
3(Tinggi)	25	6	10	25 / (25 + 6) = 0.81	25 / (25 + 10) = 0.71	2 × (0.81 × 0.71) / (0.81 + 0.71) = 0.76

Secara umum, *confusion matrix* ini menunjukkan bahwa model memiliki performa yang baik, dengan kesalahan klasifikasi yang relatif kecil *confusion matrix* diatas, berikut hasil perhitungan hasil akurasi :

$$Accuracy = \frac{TP_{total}}{Jumlah\ total\ sampel} = \frac{13+271+25}{334} = \frac{309}{334} = 0.9251\ atau\ 92.51\% \quad (2)$$

Berdasarkan Gambar B, berikut keterangannya:

Tabel 2. Evaluasi Manual per kelas pada *Gradient Boosting*

Kelas	TP	FP	FN	Precision	Recall	F1-Score
1(Rendah)	13	2	7	13 / (13 + 2) = 13/15 = 0.87	13 / (13 + 7) = 13/20 = 0.65	2 × (0.87 × 0.65) / (0.87 + 0.65) = 0.74
2 (Sedang)	269	14	10	269 / (269 + 14) = 269/283 = 0.95	269 / (269 + 10) = 269/279 = 0.96	2 × (0.95 × 0.96) / (0.95 + 0.96) = 0.96
3 (Tinggi)	28	8	7	28 / (28 + 8) = 28/36 = 0.78	28 / (28 + 7) = 28/35 = 0.80	2 × (0.78 × 0.80) / (0.78 + 0.80) = 0.79

Hasil Gambar B Gradient Boosting mampu mengklasifikasikan sebagian besar data dengan akurat, terutama pada kelas 3.Kesalahan klasifikasi terjadi dalam jumlah kecil, yang menunjukkan performa model tergolong baik. Berikut perhitungan total akurasi dimulai pada persamaan (3) :

$$Accuracy = \frac{TP_{total}}{Jumlah\ total\ sampel} = \frac{13+269+28}{334} = \frac{310}{334} = 0.9281\ atau\ 92.81\% \quad (3)z$$

Berdasarkan Gambar C, berikut keterangannya:

Tabel 3. Evaluasi Manual per kelas pada *Random Forest*

Kelas	TP	FP	FN	Precision	Recall	F1-Score
1(Rendah)	14	1	6	14 / (14 + 1) = 14/15 = 0.93	14 / (14 + 6) = 14/20 = 0.70	2 × (0.93 × 0.70) / (0.93 + 0.70) = 0.80
2(Sedang)	270	12	9	270 / (270 + 12) = 270/282 = 0.96	270 / (270 + 9) = 270/279 = 0.97	2 × (0.96 × 0.97) / (0.96 + 0.97) = 0.96
3(Tinggi)	29	8	6	29 / (29 + 8) = 29/37 = 0.78	29 / (29 + 6) = 29/35 = 0.83	2 × (0.78 × 0.83) / (0.78 + 0.83) = 0.81

Pada Gambar C model *Random Forest* cukup akurat, dengan sebagian besar prediksi tepat. Hasil ini menunjukkan bahwa model *Random Forest* memiliki performa yang baik dengan kesalahan klasifikasi yang sangat minim. Berikut perhitungan total akurasi dimulai pada persamaan (4):

$$Accuracy = \frac{TP_{total}}{Jumlah\ total\ sampel} = \frac{14+270+29}{334} = \frac{313}{334} = 0.9371\ atau\ 93.71\% \quad (4)$$

Berdasarkan Gambar D, berikut keterangannya:

Kelas	TP	FP	FN	Precision	Recall	F1-Score
1(Rendah)	13	2	7	13 / (13 + 2) = 13/15 = 0.87	13 / (13 + 7) = 13/20 = 0.65	(2 × 0.87 × 0.65) / (0.87 + 0.65) = 0.74
2(Sedang)	270	14	9	270 / (270 + 14) = 270/284 = 0.95	270 / (270 + 9) = 270/279 = 0.97	(2 × 0.95 × 0.97) / (0.95 + 0.97) = 0.96
3(Tinggi)	29	7	6	29 / (29 + 7) = 29/36 = 0.81	29 / (29 + 6) = 29/35 = 0.83	(2 × 0.81 × 0.83) / (0.81 + 0.83) = 82

Secara keseluruhan, *Decision Tree* mampu mengenali mayoritas data dengan tepat, namun sedikit lebih banyak kesalahan klasifikasi dibandingkan model *Random Forest*. performa klasifikasi. Berikut perhitungan total akurasi dimulai pada persamaan (5):

$$Accuracy = \frac{TP_{total}}{Jumlah\ total\ sampel} = \frac{13+270+29}{334} = \frac{312}{334} = 0.9341\ atau\ 93.41\% \quad (5)$$

V. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengimplementasikan empat algoritma *machine learning SVM, Gradient Boosting, Random Forest*, dan *Decision Tree* untuk mengklasifikasikan tingkat polusi udara di DKI Jakarta berdasarkan data ISPU tahun 2024. Proses dilakukan secara sistematis, mulai dari pengumpulan dan praproses data hingga pelatihan dan evaluasi model. Hasil evaluasi menunjukkan bahwa seluruh algoritma memberikan performa klasifikasi yang sangat baik, dengan *Random Forest* mencatat akurasi tertinggi sebesar 93,71%, diikuti oleh *Decision Tree* 93,41%*, Gradient Boosting* 92,81%, dan *SVM* 92,51%.

Untuk pengembangan selanjutnya, disarankan agar penelitian menggunakan dataset yang lebih lengkap dengan cakupan waktu yang lebih luas serta mempertimbangkan variabel eksternal seperti suhu, kelembaban, dan curah hujan. Selain itu, penerapan teknik lanjutan seperti hyperparameter tuning dan cross-validation dapat meningkatkan akurasi dan generalisasi model. Pengembangan sistem pemantauan atau dashboard visual juga direkomendasikan agar hasil model dapat digunakan secara praktis dalam mendukung pengambilan keputusan oleh pihak terkait.

PENGAKUAN

Naskah ilmiah ini merupakan bagian dari penelitian Tugas Akhir yang dilakukan oleh Muhammad Arya Suhendi yang berjudul "Penerapan Algoritma *Machine Learning* Untuk Mengklasifikasikan Polusi Udara Di Wilayah DKI Jakarta*"* Penelitian ini dibimbing oleh Bapak Tatang Rohana, S.T.,M.M., M.Kom dan Bapak Jamaludin Indra, S.Kom.,M.Kom.

DAFTAR PUSTAKA

[1] IQAir, "World Air Quality Report 2023," *IQAir*, pp. 1–45, 2023, [Online]. Available: https://www.iqair.com/world-most-polluted-countries

[2] Geneva: World Health Organization, "WHO global air quality guidelines," *Part. matter (PM2.5 PM10), ozone, nitrogen dioxide, sulfur dioxide carbon monoxide*, pp. 1–360, 2021.

[3] Asiva Noor Rachmayani, "Statistik Transportasi Provinsi DKI Jakarta 2022/2023," p. 6, 2023.

[4] G. Syuhada *et al.*, "Impacts of Air Pollution on Health and Cost of Illness in Jakarta, Indonesia," *Int. J. Environ. Res. Public Health*, vol. 20, no. 4, 2023, doi: 10.3390/ijerph20042916.

[5] 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.

[6] A. Hidayat, "Dampak Perubahan Iklim Terhadap Pertanian Dan Strategi Adaptasi Yang Diterapkan Oleh Petani (2)," *Univ. Medan Area*, 2023.

[7] 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.

[8] M. Pandia, "Kajian Literatur Multimedia Retrieval : Machine Learning Untuk Pengenalan Wajah," *J. Ilmu Komput. dan Sist. Inf.*, vol. 7, no. 1, pp. 161–166, 2024, doi: 10.55338/jikomsi.v7i1.2758.

[9] I. I. Ridho and G. Mahalisa, "Analisis Klasifikasi Dataset Indeks Standar Pencemaran Udara (Ispu) Di Masa Pandemi Menggunakan Algoritma Support Vector Machine (Svm)," *Technol. J. Ilm.*, vol. 14, no. 1, p. 38, 2023, doi: 10.31602/tji.v14i1.8005.

[10] R. E. Putra, M. Kalista, and ..., "Klasifikasi prediksi kualitas udara Menggunakan metode Support Vector Machine (SVM)," *eProceedings …*, vol. 10, no. 4, pp. 3790–3796, 2023, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20808%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/download/20808/20307

[11] J. Nasional, S. Informasi, N. Christina, and T. Linda, "Komparasi Algoritma Naïve Bayes dan Gradient Boosting untuk Prediksi Pasien Diabetes," vol. 02, pp. 118–125, 2024.

[12] A. T. Pratiwi, A. Barizi, M. I. Maulana, and P. Rosyani, "Systematic Literature Review Penerapan Gradient Boosting Untuk Klasifikasi Penyakit Diabetes Tipe 2," vol. 2, no. 3, pp. 454–458, 2024.

[13] M. Asgari, W. Yang, and M. Farnaghi, "Spatiotemporal data partitioning for distributed random forest algorithm: Air quality prediction using imbalanced big spatiotemporal data on spark distributed framework," *Environ. Technol. Innov.*, vol. 27, p. 102776, 2022, doi: 10.1016/j.eti.2022.102776.

[14] H. Hasna, Nonong Amalita, Dony Permana, and Admi Salma, "Random Forest Implementation for Air Pollution Standard Index Classification in DKI Jakarta 2022," *UNP J. Stat. Data Sci.*, vol. 2, no. 2, pp. 226–233, 2024, doi: 10.24036/ujsds/vol2-iss2/173.

[15] A. Ilyasa *et al.*, "Diagnosa Penyakit Kulit Wajah Dengan Metode Decision Tree dan Algoritma C4 . 5," vol. 6, pp. 88–98, 2025.

[16] A. Z. D. Nur Adiya, A. F. Desvita, A. Fidela, D. Amelia, and T. Astuti, "Penerapan Data Mining Untuk Klasifikasi

Kualitas Udara di Daerah Istimewa Yogyakarta Menggunakan Algoritma C4.5," *JDMIS J. Data Min. Inf. Syst.*, vol. 2, no. 2, pp. 59–65, 2024, doi: 10.54259/jdmis.v2i2.2800.

[17] A. W. Mucholladin, F. A. Bachtiar, and M. T. Furqon, "Klasifikasi Penyakit Diabetes menggunakan Metode Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 2, pp. 622–633, 2024, [Online]. Available: http://j-ptiik.ub.ac.id