UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2 Nomor 2 Tahun 2024 Halaman 226-233.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Random Forest Implementation for Air Pollution Standard Index Classification in DKI Jakarta 2022

Hanifa Hasna, Nonong Amalita*, Dony Permana, Admi Salma
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: nongmat@fmipa,unp,ac,id

Submitted : 24 Mei 2024
Revised : 31 Mei 2024
Accepted : 31 Mei 2024

ABSTRACT

*Air pollution is a serious challenge in various cities, including DKI Jakarta. Based on measurements of the Air Pollution Standard Index carried out by the DKI Jakarta Environmental Service, the air quality in DKI Jakarta is considered moderate to unhealthy. Deteriorating air quality in the Jakarta metropolitan area is very dangerous for humans and living things. Therefore, to prevent the problem, the classification of air quality based on pollutant content is carried out using Random Forest (RF). The application of RF will form several trees that can provide better predictions and are able to produce low errors. The result of this study obtained optimal tree formation, namely tree formation using a combination of mtry (*many input variables randomly selected in one sorting node)*=2 and ntree (number of trees in the forest) as many as 5000 trees. The resulting accuracy was 99.17% with an OOB error rate of 0.83%. This research identifies that particulate pollutants are the main factor causing air pollution in DKI Jakarta. Based on these results, it shows that RF is able to provide accurate predictions about the level of air pollution in DKI Jakarta and can be identify important factors that affect air pollution.*

Keywords: *Air Pollution, Air Pollution Standard Index, Classification, DKI Jakarta, Random Forest*

I. PENDAHULUAN

Udara merupakan sumber daya yang umum dan tidak terbatas yang mempengaruhi kehidupan manusia dan organisme hidup (Agista, dkk, 2020). Udara berperan penting dalam segala hal, sehingga kualitas udara yang baik memberikan manfaat penting. Namun polusi udara dapat mempengaruhi kualitas udara di suatu wilayah dan menimbulkan bahaya. Menurut Amalia dkk (2022), polusi udara dapat menimbulkan berbagai penyakit pada manusia seperti kanker paru-paru, infeksi saluran pernafasan, sesak nafas, dan penyakit lainnya. Selain itu, polusi udara juga berdampak pada perubahan ekosistem, pemanasan global, dan penipisan lapisan ozon. Banyak faktor yang dapat menyebabkan pencemaran udara, antara lain asap tembakau, aktivitas industri, pembakaran hutan dan lahan, transportasi, dan aktivitas lainnya.

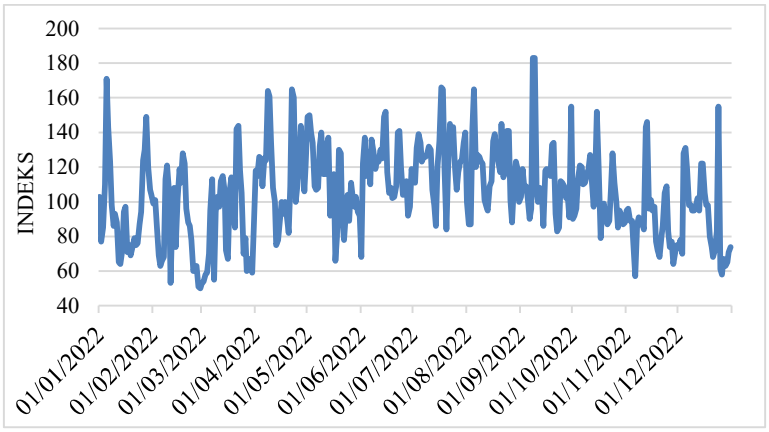
Berdasarkan pemantauan kualitas udara yang dilakukan *Air Quality Index* (AQI) Amerika Serikat pada Juni 2022, DKI Jakarta menduduki peringkat pertama kota dengan kualitas udara terburuk di Indonesia. DKI Jakarta tergolong kota dengan kualitas udara sedang hingga tidak sehat. DKI Jakarta adalah ibu kota Indonesia dan pusat pemerintahan, bisnis, dan budaya. Tren memburuknya kualitas udara di wilayah metropolitan Jakarta mungkin disebabkan oleh faktor alam dan manusia. Salah satu faktor alam yang dapat mencemari atmosfer adalah aktivitas gunung berapi. Sementara itu, aktivitas transportasi dan industri merupakan faktor manusia yang paling penting dalam pertumbuhan penduduk (Agista, dkk, 2020).

Dinas Lingkungan Hidup (DLH) Provinsi DKI Jakarta melakukan pengukuran udara yang dapat digunakan untuk memperkirakan kualitas udara di DKI Jakarta. Indeks Standar Pencemaran Udara (ISPU) adalah indeks yang mencatat pengukuran tersebut. Informasi mengenai kualitas udara bersih atau kotor serta dampaknya terhadap kesehatan dapat dilihat pada ISPU yang merupakan hasil pemantauan dan pelaporan kualitas udara. ISPU tidak

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/173

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

memiliki satuan dan direpresentasikan secara numerik (Amalia, dkk, 2022). Berdasarkan ISPU Dinas Lingkungan Hidup DKI Jakarta, Gambar 1 menggambarkan keadaan kualitas udara di DKI Jakarta (DLH).



Gambar 1. Kondisi Udara DKI Jakarta Berdasarkan ISPU harian Tahun 2022

Dari Gambar 1 terlihat bahwa ISPU DKI Jakarta cenderung berada pada kisaran 101 hingga 200 yang menempatkan kualitas udara DKI Jakarta pada kategori "tidak sehat". Sebaliknya, kualitas udara di wilayah metropolitan Jakarta yang tergolong kualitas udara sehat hanya terjadi pada satu hari dalam setahun, yakni pada tanggal 28 Februari 2022. Hal ini menunjukkan bahwa kondisi udara di wilayah metropolitan Jakarta didominasi oleh kualitas udara yang tidak sehat. Penentuan tingkat ISPU dapat dengan mudah dilakukan melalui proses klasifikasi untuk mencegah terjadinya ISPU yang lebih tinggi, sehingga membantu pemerintah dan DLH DKI Jakarta untuk mengambil tindakan yang dapat mengurangi pencemaran udara. Metode klasifikasi yang umum digunakan adalah hutan acak (*Random Forest)*.

Random Forest (RF) menggunakan sampel *bootstrap* untuk membentuk sebuah pohon, membentuk banyak pohon yang dapat mencapai hasil prediksi yang lebih baik. RF dapat diterapkan untuk memproses data pelatihan dalam jumlah yang sangat besar secara efisien. Selain itu, RF juga efektif dalam menangani data yang hilang dan dapat memberikan hasil dengan kesalahan yang lebih sedikit (Breman, 2001). Shofa (2023) melakukan klasifikasi ISPU di DKI Jakarta dengan membandingkan metode RF dan metode multilayer perceptron (MLP). Berdasarkan hasil penelitian menunjukkan bahwa metode RF mempunyai akurasi yang lebih tinggi dibandingkan dengan metode MLP, karena metode RF memperoleh nilai akurasi yang lebih tinggi (99%) dibandingkan dengan metode MLP.

Selanjutnya, Syukron (2018) mengembangkan sistem klasifikasi untuk evaluasi kredit menggunakan dataset German Credit, sedangkan Primajaya (2018) menerapkan pendekatan RF untuk mengkategorikan curah hujan. Temuan penelitian ini menunjukkan bahwa hasil akurasinya melebihi 90%, menunjukkan bahwa pendekatan RF dapat menghasilkan klasifikasi berkualitas tinggi karena nilai akurasinya yang tinggi. Oleh karena itu, tujuan penelitian ini adalah untuk mengklasifikasikan ISPU di DKI Jakarta pada tahun 2022 berdasarkan kekhawatiran yang telah dibahas.

II. METODE PENELITIAN

A. Jenis Penelitian dan Sumber Data

Jenis penelitian yang dilakukan adalah penelitian terapan. Penelitian ini menggunakan data sekunder yaitu data Indeks Standar Pencemaran Udara (ISPU) DKI Jakarta tahun 2022 yang disediakan oleh Dinas Lingkungan Hidup DKI Jakarta. Variabel yang digunakan dalam penelitian ini adalah satu variabel respon (kategori udara) dan enam variabel prediktor: materi Partikulat (PM_{10}), materi Partikulat ($PM_{2,5}$), Nitrogen Dioksida (NO_2), Sulfur Dioksida (SO_2), Karbon Monoksida (CO) dan Ozon (O_3).

B. Langkah-Langkah Analisis

Analisis data pada penelitian ini menggunakan RF dengan bantuan *software Python*. Adapun langkah-langkah analisis RF yang dilakukan dalam penelitian ini adalah sebagai berikut.

1. Mengumpulkan data Indeks Standar Pencemaran Udara Tahun 2022 yang diperoleh dari Dinas Lingkungan Hidup DKI Jakarta. Data yang digunakan adalah data harian yang terdiri dari 365 observasi dengan satu variabel respon dan enam variabel prediktor.
2. Melakukan eksplorasi data.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/173

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

3. Melakukan pengambilan sampel *bootstrap* ke-*i* sebanyak 2/3 dari *dataset original* dengan jumlah yang sama dengan *dataset original* menggunakan sistem *resampling*. Sedangkan sisanya 1/3 dari *dataset original* akan dijadikan sampel OOB ke-*i* (Breiman, 2001). Pengambilan sampel ini dilakukan untuk setiap pembentukan satu *tree*.
4. Melakukan pemilihan peubah penjelas X sebanyak parameter *mtry* √p=2 secara *random* dengan m < p (jumlah variabel). Terdapat tiga parameter utama yang digunakan metode *random forest* yaitu *mtry* (banyak input variabel secara acak terpilih dalam satu *node* pemilah), *ntree* (jumlah banyaknya *tree* dalam *forest*) dan *node size* (jumlah minimum amatan dalam sebuah *terminal node*) (Genuer R, 2008:5). Untuk perhitungan nilai mtry menggunakan $mtry_I=\sqrt{p}$. Pembentukan jumlah pohon *(ntree)* yang digunakan adalah 5000, namun terdapat enam nilai parameter *ntree* yang bisa digunakan yaitu 10, 50. 100, 500, 1000, 5000. Pembentukan pohon keputusan dalam *random forest* dimulai dengan menghitung nilai *entropy* sebagai penentu tingkat ketidakmurnian atribut dari nilai *information gain*. Pencarian nilai *entropy* menggunakan rumus:

$$Entropy\ (\mathrm{Y}) = -\sum_{i=1}^{n} p_i * log_2\ p_i \qquad (1)$$

Keterangan:
Y : himpunan kasus
n : jumlah partisi atribut *i*
p_i : proporsi nilai Y_i terhadap kelas Y
Sedangkan untuk mencari *information gain* yang digunakan untuk mengukur efektivitas suatu atribut dalam pengklasifikasian data dapat dihitung dengan menggunakan Persamaan (2).

$$Information\ Gain\ (\mathrm{Y},\alpha) = Entropy\ (\mathrm{Y}) - \sum_{i=1}^{n} \frac{Y_i}{Y} * Entropy\ (Y_i) \qquad (2)$$

Keterangan:
Y : himpunan kasus
α : atribut
n : jumlah partisi atribut *i*
Y_i : jumlah kasus pada partisi ke-*i*
5. Setelah *tree* dan *forest* terbentuk, selanjutnya menghitung besar nilai misklasifikasi menggunakan sampel *Out Of Bag* (OOB) berdasarkan *ntree* menggunakan persamaan laju galat OOB_i untuk satu *tree* seperti pada Persamaan (3).

$$\text{Laju Galat } OOB_i = \frac{1}{n}\sum_{i=1}^{n} 1_{Y_i \neq \hat{Y}_i} \qquad (3)$$

Keterangan:
$\sum_{i=1}^{n} 1_{Y_i \neq \hat{Y}_i}$: jumlah data hasil prediksi yang salah (misklasifikasi)
Y_i : hasil amatan sebenarnya ke-*i*
$\hat{Y}_i$: hasil amatan yang diprediksi ke-*i*
n : jumlah OOB ke-*i* menjadi sampel OOB
Setelah mendapatkan nilai besar nilai misklasifikasi dari sampel OOB ke-*i*, selanjutnya akan dihitung rataan tingkat misklasifikasi OOB dengan menggunakan Persamaan (4).

$$\text{Laju Galat OOB} = \frac{\sum OOB_i\ error\ rate}{k} \text{ x } 100\% \qquad (4)$$

dimana *k* yaitu banyak pohon yang terbentuk. Semakin akurat dan dapat diandalkan prediksi hutan, semakin rendah perkiraan tingkat kesalahan OOB yang dihasilkan.
6. Mengidentifikasi variabel *importance*
Salah satu *output* yang terdapat dalam penerapan RF adalah mengidentifikasi variabel *importance.* Variabel *importance* dapat dihitung menggunakan *Mean Decrease Accuracy* (MDA). Menurut Strobl, dkk. (2008), MDA merupakan teknik komputasi yang dapat menentukan tingkat kepentingan variabel prediktor dengan mempartisi data sampel menggunakan pengurutan dan OOB sampling. OOB memperkirakan nilai prediksi yang lebih akurat dengan menghitung nilai presisi OOB variabel X sebelum dan sesudah substitusi, serta dengan menghitung selisih kedua nilai tersebut. Perhitungan MDA dilakukan dengan menggunakan rumus seperti Persamaan (5).

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/173

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

$$\text{MDA}(X_h) = \frac{1}{k}\sum_{t=1}^{k}\frac{\sum_{i\in OOB(t)} I(y_i=\hat{y}_i^{(t)}) - \sum_{i\in OOB(t)} I(y_i=\hat{y}_{i,h}^{(t)})}{|OOB^{(t)}|} \quad (5)$$

Dengan asumsi bahwa $t \in \{1,2,,3,\dots,k\}$, sampel OOB untuk pohon ke-t direpresentasikan sebagai $OOB^{(t)}$. Nilai rata-rata selisih kelas prediksi sebelum permutasi X adalah $[\![\hat{y}_i^{(t)} = f^{(t)}(x_i)$, dan kelas prediksi setelah permutasi X yaitu $\hat{y}_{i,h}^{(t)} = f^{(t)}(x_{i,h})$ pada pengamatan ke-i, menunjukkan tingkat relevansi variabel X pada pohon ke-t.

7. Menghitung tingkat *accuracy, specificity, sensitivity* menggunakan *confusion matrix*.
Menurut Indriyanti (2017), matriks konfusi adalah perhitungan yang membandingkan hasil klasifikasi dan data berdasarkan data aktual dan total dataset. Hasil akhir dari matriks ini adalah tingkat presisi yang dinyatakan dalam persentase (%). Menurut Nurdalia (2023), tabel matriks konfusi untuk mengklasifikasikan dua kelas ditunjukkan pada Tabel 1.

Tabel 1. *Confusion Matrix* Kelasifikasi Dua Kelas

f_{ij}		Kelas hasil prediksi (j)	
		Kelas = 1	**Kelas = 0**
Kelas asli (i)	Kelas = 1	TP	FN
	Kelas = 0	FP	TN

Keterangan:
TP (*True Positif*) : Kelas yang diamati benar dengan hasil Kelas yang diprediksi benar.
TN (*True Negatif*) : Kelas yang diamati salah dengan hasil Kelas yang diprediksi salah.
FP (*False Positif*) : Kelas yang diamati salah dengan hasil Kelas yang diprediksi benar.
FP (*False Negatif*) : Kelas yang diamati benar dengan hasil Kelas yang diprediksi salah.
Menurut Mustika dkk (2021: 201-204) terdapat tiga *performance metrics* yang digunakan yaitu:

1. Akurasi (*Accuracy*)
Rasio prediksi akurat baik positif maupun negatif terhadap keseluruhan jumlah data yang digunakan untuk pengujian dan pelatihan dikenal sebagai akurasi. Sejauh mana suatu model dapat mengklasifikasikan dengan benar disebut sebagai keakuratannya dalam klasifikasi. Sejauh mana nilai yang diantisipasi menyerupai nilai sebenarnya dapat disebut sebagai akurasi. Persamaan (6) dapat dimanfaatkan untuk mencari nilai akurasi.

$$Accuracy = \frac{\text{TP+T}}{\text{TP+TN+FP+F}} \quad (6)$$

2. *Specificity*
Keakuratan prediksi negatif relatif terhadap semua data negatif dikenal sebagai spesifisitas (*Specificity)*. Persamaan (7) dapat digunakan untuk menghitung nilai spesifisitas.

$$Specificity = \frac{\text{TN}}{\text{TN+F}} \quad (7)$$

3. *Recall (Sensitivity)*
Rasio perkiraan yang akurat terhadap semua hasil proyeksi yang positif dan asli disebut *recall*. Persamaan (8) dapat digunakan untuk menghitung *recall.*

$$Sensitivity = \frac{\text{TP}}{\text{TP+F}} \quad (8)$$

8. Menarik kesimpulan.

III. HASIL DAN PEMBAHASAN

A. Deskripsi Data

Sebelum melakukan analisis RF, dilakukan analisis data terlebih dahulu. Ini akan membantu dalam mendapatkan gambaran lengkap tentang data Standar Pencemaran Udara (ISPU) DKI Jakarta 2022. Data ISPU DKI Jakarta terdiri dari 365 observasi dengan 6 variabel prediktor dan 1 variabel respon. Rangkuman masing-masing variabel prediktor pada data ISPU DKI Jakarta 2022 disajikan pada Tabel 2.

Tabel 2. Analisis Deskriptif Variabel Prediktor

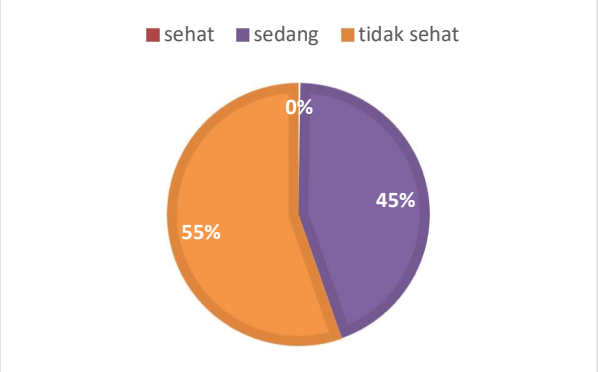
Variabel	Mean	Median	Min	Max	Std. Deviasi
Partikulat (PM_{10})	65,770	65	27	97	12,843

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/173

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Variabel	Mean	Median	Min	Max	Std. Deviasi
Partikulat (PM_{25})	103,595	103	50	171	24,610
Sulfida (SO_2)	48,592	50	38	59	4,549
Karbon Monoksida (CO)	20,266	19	10	60	6,834
Ozon (O_3)	66,869	63	15	183	27,876
Nitrogen Dioksida (NO_2)	32,280	32	8	54	8,840

Tabel 2 merupakan gambaran masing-masing variabel prediktor yang menunjukkan jenis kandungan polutan pemicu terjadinya pencemaran udara. Berdasarkan tabel tersebut terlihat bahwa kandungan udara yang memiliki nilai rata-rata tertinggi terdapat pada Partikulat (PM_{25}) sebesar 103,595. Hal ini menunjukkan bahwa rata-rata kondisi udara di DKI Jakarta dapat dikategorikan tidak sehat. Selain itu, terdapat kandungan udara yang memiliki nilai maksimum diatas 100 yang artinya terdapat hari-hari tertentu di DKI Jakarta yang memiliki kondisi udara yang tidak sehat. Berdasarkan keseluruhan kandungan udara di DKI Jakarta tahun 2022, Sulfida (SO_2) memiliki nilai keragaman yang kecil dengan angka yang masih tergolong baik. Dilihat dari rata-rata, kandungan udara Sulfida (SO_2) cenderung rendah yang artinya kondisi udara tidak terlalu berbeda antar hari. Selanjutnya dilakukan eksplorasi data untuk variabel respon. Eksplorasi data variabel respon disajikan seperti pada Gambar 2.



Gambar 2. Diagram Lingkaran ISPU DKI Jakarta 2022

Gambar 2 menunjukkan kondisi udara di DKI Jakarta tahun 2022 memiliki kualitas udara yang dapat dikategorikan sehat, sedang, dan tidak sehat. Berdasarkan gambar tersebut, kondisi udara yang memiliki kategori sehat hanya 1 amatan sebesar 0,274%. Sedangkan, kondisi udara yang memiliki kategori sedang 162 amatan, dan untuk kondisi udara dengan kategori tidak sehat sebanyak 202 amatan. Hal ini menunjukkan bahwa kondisi udara yang tidak sehat lebih banyak terjadi dibandingkan dengan kondisi udara yang sehat.

B. Analisis *Random Forest*

Klasifikasi kualitas udara DKI Jakarta menggunakan RF memerlukan pembentukan pohon yang optimal. Hal pertama yang perlu dilakukan dalam melakukan pembentukan pohon adalah membagi data menjadi dua bagian, yaitu 2/3 data asli (244 amatan) dijadikan sebagai sampel *boostrap* ke-*i*, dan 1/3 data lainnya (121 amatan) akan dijadikan sampel OBB ke-*i*. Selanjutnya, pembentukan pohon dilakukan dengan menggunakan kombinasi parameter *mtry* (banyak input variabel secara acak terpilih dalam satu *node* pemilah) dan *ntree* (jumlah banyaknya *tree* dalam *forest*). Pada penelitian ini, *ntree* yang digunakan sebanyak 10, 50, 100, 500, 1000, dan 5000 pohon dengan *mtry*=2. Kombinasi *ntree* dan *mtry* tersebut yang akan dilakukan analisis terhadap rataan besar nilai misklasifikasi *Out of Bag Error* (OOB). Hasil akurasi masing-masing kombinasi *ntree* dan *mtry*=2 dengan menggunakan bantuan *library randomforestclassifier* pada *Python* dapat dilihat seperti pada Tabel 3.

Tabel 3. Akurasi Setiap Kombinasi *Ntree* dan *Mtry=2*

Ntree	Akurasi (%)
10	97,52
50	97,52
100	97,52
500	98,35
1000	98,35
***5000**	**99,17**

*Pembentukan pohon optimal

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/173

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

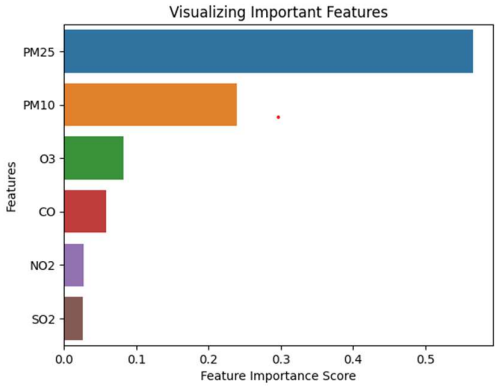
Pembentukan pohon yang optimal dipilih berdasarkan dari kombinasi *ntree* dan *mtry*=2 yang memperoleh nilai akurasi tertinggi dan laju galat OOB terendah. Berdasarkan Tabel 3 terlihat bahwa kombinasi *ntree* sebanyak 5000 pohon dengan *mtry*=2 memperoleh nilai akurasi tertinggi sebesar 99,17%. Sehingga, laju galat OOB yang diperoleh kombinasi tersebut sebesar 0,83%. Dengan demikian, pembentukan pohon yang optimal terdapat pada hasil pohon dengan kombinasi *mtry*=2 dengan *ntree* sebanyak 5000 pohon.

Setelah diperoleh pembentukan pohon yang optimal, selanjutnya mengidentifikasi variabel *importance* atau variabel prediktor penting yang menjadi pemicu terjadinya pencemaran udara di DKI Jakarta. Variabel *importance* didapatkan dari perhitungan *Mean Decrease Accuracy* (MDA) yang diperoleh dengan menggunakan *feature importance scores* pada *Python*. Hasil *feature importance scores* dapat dilihat seperti pada Tabel 4.

Tabel 4. *Feature Importance Scores*

Fitur	Skor
Partikulat (PM_{25})	0,596062
Partikulat (PM_{10})	0,212128
Ozon (O_3)	0,082539
Karbon Monoksida (CO)	0,054435
Nitrogen Dioksida (NO_2)	0,032979
Sulfida (SO_2)	0,021857

Tabel 4 menunjukkan urutan kandungan polutan yang menjadi pemicu terjadinya pencemaran udara di DKI Jakarta. Berdasarkan tabel tersebut terlihat bahwa jenis kandungan Polutan Partikulat (PM_{25}) memperoleh skor tertinggi yang diikuti dengan Partikulat (PM_{10}), Karbon Monoksida (CO), Ozon (O_3), Sulfida (SO_2), Nitrogen Dioksida (NO_2). Hal ini menjelaskan bahwa jenis kandungan polutan yang mendominasi terjadinya pencemaran udara di DKI Jakarta tahun 2022 terdapat pada kandungan Partikulat (PM_{25}). Sedangkan, pencemaran udara di DKI Jakarta tidak terlalu disebabkan oleh kandungan polutan Sulfida (SO_2). Untuk lebih jelas dapat dilihat visualisasi fitur *importance* seperti pada Gambar 3.



Gambar 3. Visualisasi Variabel *Importance*

Berdasarkan Gambar 3 terlihat bahwa polutan Partikulat mendominasi terjadinya pencemaran udara di DKI Jakarta tahun 2022. Sumber polutan partikulat dapat berasal dari pembangkit tenaga listrik, pembakaran bahan bakar fosil dari kendaraan, pembakaran hutan, debu vulkanik, dan dari proses-proses industri lainnya. Adanya pencemaran udara yang disebabkan oleh polutan partikulat dapat menyebabkan terjadinya penyakit pernafasan seperti asma, kanker paru-paru, atau bahkan dapat menyebabkan kematian.

Tahapan selanjutnya dalam melakukan analisis RF adalah menguji hasil performa algoritma RF dengan menggunakan *confusion matrix* (matriks konfusi). Matriks konfusi berisikan matriks dari hasil prediksi klasifikasi yang dibandingkan dengan data asli. Matriks konfusi dapat dilihat seperti pada Tabel 5.

Tabel 5. Matriks konfusi

Prediksi	Aktual	
	Sedang	Tidak Sehat
Sedang	72	0
Tidak Sehat	1	48

Tabel 5 merupakan hasil *confusion matrix* menggunakan pembentukan pohon dengan kombinasi *mtry*=2 dengan *ntree* sebanyak 5000 pohon. Berdasarkan tabel tersebut terlihat bahwa dari 72 kondisi udara yang memiliki kualitas udara sedang, terdapat 72 kondisi udara yang diprediksi memiliki kualitas udara sedang dan tidak ada

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/173

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

kondisi udara yang diprediksi memiliki kualitas udara tidak sehat. Selain itu, dari 49 kondisi udara yang memiliki kualitas tidak sehat, terdapat 48 kondisi udara yang diprediksi memiliki kualitas udara yang tidak sehat dan terdapat 1 kondisi udara yang diprediksi memiliki kualitas udara sedang. Berdasarkan hasil *confusion matrix* tersebut selanjutnya dilakukan perhitungan untuk melihat perbandingan *sensitivity*, *specificity*, dan *accuracy* yang dapat dilihat seperti pada Tabel 6.

Tabel 6. *Sensitivity*, *Specificity*, dan *Accuracy*

Sensitivity	***Specificity***	***Accuracy***
100%	97,96%	99,17%

Berdasarkan Tabel 6 terlihat bahwa nilai *sensitivity* yang diperoleh sebesar 100%. Hal ini menjelaskan bahwa kemampuan RF dalam mengklasifikasi data ISPU DKI Jakarta dengan benar untuk kondisi udara yang memiliki kualitas udara sedang adalah 100%. Sedangkan, untuk *specificity* diperoleh sebesar 97,96% yang berarti bahwa kemampuan RF dalam mengklasifikasi data ISPU DKI Jakarta dengan benar untuk kondisi udara dengan kualitas udara tidak sehat sebesar 97,96%. Sementara itu, untuk nilai *accuracy* yang diperoleh secara keseluruhan sebesar 99,17%. Hal ini menunjukan bahwa RF mampu mengklasifikasi data ISPU DKI Jakarta tahun 2022 dengan sangat baik, karena hasil *accuracy* yang didapatkan memperoleh nilai lebih dari 90%.

IV. KESIMPULAN

Penerapan RF dalam mengklasifikasi ISPU DKI Jakarta tahun 2022 menghasilkan kombinasi *mtry*=2 dengan *ntree* sebanyak 5000 pohon sebagai pembentukan pohon yang optimal. Pembentukan pohon tersebut memperoleh laju galat OOB yang rendah sebesar 0,83% dan akurasi sebesar 99,17%. Hal ini menunjukkan bahwa metode RF mampu mengklasifikasi dengan sangat baik karena tingkat akurasi yang diperoleh sangat tinggi. Selain itu, berdasarkan hasil penelitian terlihat bahwa urutan jenis polutan yang menjadi pemicu terjadinya pencemaran udara adalah Partikulat (PM_{25}) sebesar 0,596062 yang diikuti dengan Partikulat (PM_{10}), Karbon Monoksida (CO), Ozon (O_3), Sulfida (SO_2), Nitrogen Dioksida (NO_2). Berdasarkan hasil tersebut diharapkan dapat membantu pemerintah dan Dinas Lingkungan Hidup DKI Jakarta dalam membuat langkah-langkah penanggulangan pencemaran udara. Serta, untuk penelitian selanjutnya dapat menambahkan variabel jenis kandungan polutan lainnya dan dapat mengganti atau menambahkan kombinasi jumlah parameter *ntree* dan *mtry* yang digunakan.

DAFTAR PUSTAKA

Agista, P. I., Gusdini, N., & Maharani, M. D. (2020). Analisis Kualitas Udara dengan Indeks Standar Pencemar Udara (ISPU) dan Sebaran Kadar Polutannya di Provinsi DKI Jakarta. *SEOI, 2*(II), 39-57.

Amalia, A., Zaidiah, A., & Isnainiyah, I. N. (2022, Juni). Prediksi Kualitas Udara Menggunakan Algoritma K-Nearest Neighbor. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 07(02)*, 496-507.

Breiman, L. (2001). Random Forest. Berkeley: Statistics Department University of California.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1993). Classification And Regression Tree. New York: Chapman And Hall.

Dinas Lingkunngan Hidup DKI Jakarta

Fadilah, L. (2018). "Klasifikasi Random Forest Pada Data Imbalanced", *Skripsi,* 51 Hal., Universitas Islam Negeri Syarif Hidayatullah, Jakarta, Indonesia, Juni 2018.

Indriyanti, D., & Khoeroh, H. (2017). Evaluasi Penatalaksanaan Gizi Balita Stunting di Wilayah Kerja Puskesmas Sirampog. Unnes Journal of Public Health, 6(3), 189-195.

Keputusan Menteri Negara Lingkungan Hidup No.45/MENLH/10/1997 Tentang Indeks Standar Pencemar Udara

Mustika, Ardilla, Y., Manuhutu, A., Ahmad, N., Hasbi, I., Guntoro, . . . Ernawati, I. (2021). Data Mining Dan Aplikasinya. Bandung: Widina Bhakti Persada.

Nurdalia, Zilrahmi, Permana, D., & Salma, A. (2023). Comparison Between Naïve Bayesand K-Nearest Neighborfor DKI Jakarta Air Pollution Standard Index Classification. *UNP JOURNAL OF STATISTICS AND DATA SCIENCE, I (2)*, 67-73.

Peraturan Pemerintah No.41 tahun 1999 Tentang Pengendalian Pencemaran Udara.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/173

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), 1(1)*, 27-31.

Shofa, S. H. (2023). "Klasifikasi Kategori Indeks Standar Pencemar Udara (ISPU) DKI Jakarta Menggunakan Multilayer Perceptron dan Random Forest", *Tugas Akhir*, 67 Hal., Universitas Siliwangi, Tasikmalaya, Indonesia, Juni 2023.

Strobl, C. dkk. (2008). Conditional Variable Importance for Random Forests. Ludwig Maximilians Universitas Munchen.

Syukron, A., & Subekti, A. (2018). Penerapan Metode Random Over-Under Sampling dan Random Forest untuk Klasifikasi Penilaian Kredit. *JURNAL INFORMATIKA, 5(2)*, 175-185.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/173