Seminar Nasional Sains Data 2024 (SENADA 2024)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Perbandingan Klasifikasi PM2.5 di Daerah Khusus Jakarta Algoritma C5.0, Random Forest, dan SVM

Lourna Mariska Mauboy[1], Muhammad Raihan Abhirama[2], Syarifa Salsabila[3], Robert Kurniawan[4]

[1, 2, 3] *Program Studi Statistika, Politeknik Statistika STIS, Jakarta, Indonesia*
[4] *Program Studi Komputasi Statistik, Politeknik Statistika STIS, Jakarta, Indonesia*
[1]212112160@stis.ac.id
[2]21211221@stis.ac.id
[4]robertk@stis.ac.id
Corresponding author email: 212112389@stis.ac.id

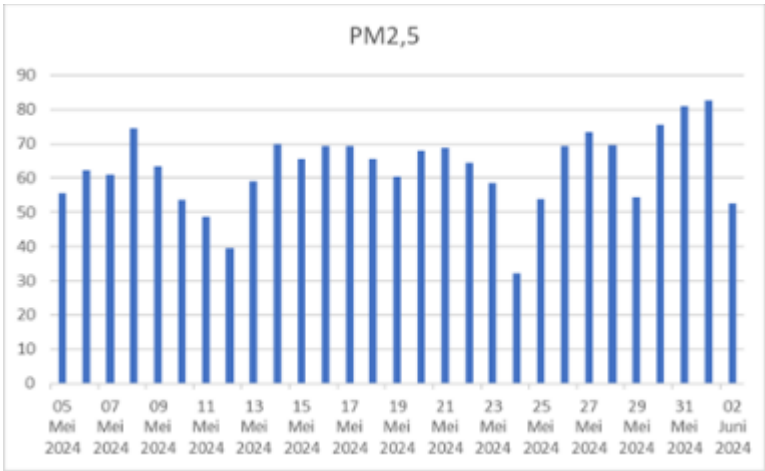
Abstract: *Air quality determines the health and quality of life of an environment. Good air quality will support quality life and better health. Poor air quality is caused by the increasing development of human activities to meet their needs. Poor air quality often occurs in developing countries with densely populated areas such as Jakarta. This research aims to classify Jakarta's air quality as indicated by the PM2.5 pollutant index value category. PM2.5 pollutants are influenced by meteorological factors. The data mining process is carried out to find information through certain patterns obtained from a set of data. The data mining method used in this research is classification which aims to classify Jakarta air quality from a PM2.5 perspective based on certain meteorological factors. Classification methods in the form of C5.0, Random Forest, and SVM were compared with certain evaluation criteria and the C5.0 method was selected as the best method with an accuracy value of 81,48%, precision 80%, recall 63,16%, and f1-score 70,59%.*
Keywords: *Classification, Air Quality, Random Forest, SVM, C5.0*

Abstrak: Kualitas udara menentukan kesehatan dan kualitas hidup dari sebuah lingkungan. Kualitas udara yang baik akan mendukung hidup yang berkualitas dan kesehatan yang lebih terjaga. Kualitas udara yang buruk diakibatkan semakin berkembangnya kegiatan manusia dalam memenuhi kebutuhannya. Kualitas udara yang buruk banyak terjadi di negara berkembang dengan wilayah berpenduduk padat seperti Jakarta. Penelitian ini bertujuan untuk mengklasifikasikan kualitas udara Jakarta yang diindikasikan oleh kategori nilai indeks polutan PM2.5. Polutan PM2.5 dipengaruhi oleh faktor-faktor meteorologi. Proses data mining dilakukan untuk menemukan informasi melalui pola tertentu yang didapatkan dari sekumpulan data. Metode data mining yang digunakan pada penelitian ini adalah klasifikasi yang bertujuan untuk mengklasifikasikan kualitas udara Jakarta dalam sudut pandang PM2.5 berdasarkan faktor meteorologi tertentu. Metode klasifikasi berupa C5.0, Random Forest, dan SVM dibandingkan dengan kriteria evaluasi tertentu dan metode C5.0 terpilih sebagai metode terbaik dengan nilai akurasi 81,48%, presisi 80%, recall 63,16%, dan f1-score 70,59%.
Kata kunci: Klasifikasi, Kualitas Udara, Random Forest, SVM, C5.0

I. PENDAHULUAN

Kualitas udara yang buruk berisiko terhadap kesehatan manusia [1]. Kualitas udara yang buruk dengan polusi menyebabkan kematian hampir 9 juta jiwa atau setara dengan 1 dari 6 kematian di dunia. Polusi udara adalah kontaminasi udara oleh *particular matter* (PM), ozon (O_3), nitrogen dioksida (NO_2), sulfur dioksida (SO_2), dan karbon monoksida (CO) [2]. Polusi udara menurunkan kualitas hidup manusia yang kemudian membebani sistem kesehatan serta menurunkan produktivitas dan pendapatan, terutama pada kelompok rentan [3]. Adanya polusi udara ini akibat dari kegiatan manusia yang semakin berkembang dan beragam berupa produksi energi, industri, rumah tangga, transportasi, dan pertanian. Kegiatan ini menghasilkan PM2.5 yang dapat menyebar di atmosfer dalam radius ratusan kilometer [4]. PM memiliki bahaya yang besar bagi manusia karena mampu menembus paru-paru, bahkan PM2.5 dapat menembus aliran darah dan berdampak pada kardiovaskuler dan pernapasan [5,6]. Pada tahun 2015, WHO melalui International Agency for Research on Cancer (IARC) mengklasifikasikan PM sebagai karsinogenik.

Polusi udara buruk banyak terjadi pada negara berkembang dengan penduduk yang padat [4]. Secara lebih spesifik, kota-kota pada wilayah yang memiliki kepadatan penduduk tinggi, memiliki kualitas udara yang tidak lebih baik. Hal seperti ini dapat terlihat pada kondisi udara Daerah Khusus Jakarta yang terlihat dari rata-rata PM2.5 diatas ambang batas aman yang ditetapkan WHO, $15\mu g/m^3$.



Gambar 1. Grafik konsentrasi PM2.5, sumber: iqair.com

Kondisi udara di DKJ pada bulan Mei 2024 secara rata-rata tergolong tidak sehat. Gambar 1. menunjukkan kontaminasi PM2.5 secara harian yang tinggi. Sumber dari polutan ini tidak hanya berasal dari kendaraan konvensional di jalan, namun berasal juga dari industri manufaktur [7,8].

Polutan PM2.5 sebagai partikel udara berukuran 2,5 mikrometer [9] yang menyerang saluran pernapasan menjadi ancaman serius daerah urban [10], terutama DKJ sebagai ibukota negara [11]. Selain dipengaruhi oleh polutan udara lainnya [12] dan kegiatan manusia, PM2.5 juga dipengaruhi oleh variabel meteorologi [13], diantaranya rata-rata temperatur atau suhu, total presipitasi, kecepatan angin, arah angin, dan tekanan udara. Secara khusus, indikator meteorologi merupakan faktor eksternal yang mempengaruhi kumpulan, persebaran hingga perkembangan polutan di suatu wilayah [14]

Polutan PM2.5 DKJ berkorelasi positif dengan temperatur [15]. Artinya peningkatan temperatur meningkatkan konsentrasi partikel PM2.5. Konsentrasi PM2.5 yang lebih tinggi cenderung terjadi pada hari-hari dengan suhu yang tinggi [16]. Kondisi ini cukup berisiko mendatangkan penyakit terutama disaat konsentrasi PM2.5 yang ekstrem dan suhu ekstrem terjadi bersamaan [17].

Banyaknya presipitasi hujan efektif menurunkan konsentrasi partikel di atmosfer, termasuk PM2.5 [18]. Hal ini disebabkan oleh di hari terjadinya hujan dan atau kabut terjadi penurunan sejumlah partikulat di atmosfer ke tanah [19]. Intensitas presipitasi hujan mampu memberikan perubahan pada kualitas udara [20].

Kecepatan angin yang rendah dapat menyebabkan partikel polutan PM2.5 terperangkap di atmosfer sehingga kedua variabel tersebut berkorelasi negative [21, 12],begitu pula efek variasi arah angin [11]. Pengaruh arah angin juga berkaitan dengan lokasi industri yang menghasilkan emisi dan aktivitas lainnya [22] seperti di Malaysia yakni polutan sebagian besar berasal dari arah timur laut karena dipengaruhi daerah sekitarnya seperti kawasan industri Johor Technology Park, Senai Technology Park, dsb, menjadikannya salah satu faktor yang menyebabkan polusi dari luar wilayah masuk menuju suatu wilayah [23].

Tekanan udara yang tinggi akan membuat atmosfer stabil sehingga menghambat pergerakan udara untuk menyebarkan atau mengencerkan PM2.5 sehingga konsentrasi polutan tersebut tetap tinggi [24]. Kombinasi setiap variabel mampu memberikan pengaruh yang lebih besar untuk semua lokasi [11]

Berdasarkan indikator meteorologi, dapat diidentifikasi pola tertentu hingga mengarah pada suatu informasi yang disebut dengan *data mining* [25]. Secara umum, ada tiga pendekatan utama prediksi PM2.5 :*statistical models*, metode numerik berupa *chemical transport and atmospheric dispersion modelling*, dan *machine learning*. Namun tidak seperti pendekatan lain yang bergantung pada keterbaruan data dan ketergantungan pada karakteristik geofisika yang kompleks, *machine learning* dapat mempertimbangkan beberapa parameter dalam satu model [26]. Proses identifikasi informasi melalui *machine learning* memungkinkan adanya model pembelajaran mesin untuk sistem perkiraan yang akurat namun dengan lebih sedikit pemrosesan menggunakan kumpulan data tertentu [27]. Salah satunya dengan klasifikasi. Metode ini memetakan variabel yang selanjutnya disebut atribut (input) ke dalam kategori atau kelas yang telah didefinisikan sebelumnya [28].

Kou et al., 2021 [29] menyatakan bahwa kondisi meteorologi suatu wilayah mampu berkontribusi positif terhadap peningkatan kualitas udara disertai implementasi kebijakan udara bersih. Sejalan dengan itu, Alpan & Sekeroglu, 2020 [30] menyatakan adanya prediksi konsentrasi polutan dengan akurasi yang tinggi menggunakan data meteorologi. Ameer, 2019 [31] melakukan klasifikasi polutan PM2.5 menggunakan variabel meteorologi dengan metode decision tree, random forest, MLP, Gradient Boosting dengan hasil random forest sebagai metode terbaik. Penelitian oleh Ejohwomu et al., 2022 [32] juga menggunakan variabel meteorologi seperti suhu dan kelembapan udara untuk memperoleh perkiraan konsentrasi PM2.5 yang andal dengan menggunakan ensemble models seperti XGBoost, ARIMA, Random Forest. Liu et al., 2019 menggunakan SVM [33] sebagai salah satu metode paling robust dan akurat dalam algoritma data mining untuk memprediksi kategori PM2.5 berdasarkan faktor meteorologi seperti tekanan atmosfer, kelembapan relatif, suhu udara, kecepatan angin, arah angin, dan total presipitasi. Didapatkan nilai akurasi yang tinggi sekaligus performa yang lebih baik dibandingkan ANN dan Adaboost menunjukkan SVM mampu memberikan ketepatan klasifikasi dan prediksi.

Beberapa penelitian dengan pendekatan klasifikasi menggunakan variabel PM2.5 dan parameter meteorologi telah dilakukan. Namun masih sedikit dilakukan di Indonesia. Terutama klasifikasi yang dilakukan berdasarkan pada polutan lainnya, sedikit yang membahas polutan berupa PM2.5 secara spesifik sekaligus kaitannya dengan parameter meteorologi dengan PM2.5. Padahal PM2.5 menimbulkan paling banyak permasalahan kesehatan. IQAir [34] menyebutkan di tahun 2019 dilaporkan 16 dari 44 kecamatan di Jakarta melaporkan adanya infeksi saluran pernapasan atas sebagai penyebab penyakit yang paling umum, sehingga penelitian penting dengan tujuan dilakukan pengklasifikasian menggunakan kondisi meteorologi yang dialami setiap hari. Penelitian ini diharapkan dapat berkontribusi dalam membantu pemangku kebijakan khususnya pemerintah DKJ dalam merumuskan kebijakan atau regulasi serta implikasi dalam kesehatan masyarakat berupa sistem peringatan dini dan pemantauan kualitas udara real-time menggunakan kondisi meteorologi.

II. METODE PENELITIAN

Metode klasifikasi yang digunakan diantaranya Decision tree C5.0, Random forest, SVM. Algoritma C5.0 menggunakan rasio *information gain* untuk memilih atribut split terbaik untuk pembentukan pohon keputusan dengan sebelumnya melakukan penghitungan *gain* dan *entropy* [35]. Rumus untuk mencari nilai *entropy* dan *gain* adalah sebagai berikut:

$$Entropy(S) = -\sum_{i=1}^{n} p_i log_2 p_i \quad (1)$$

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} x\ Entropy(S_i) \quad (2)$$

$$Gain\ Ratio = \frac{Gain(S,A)}{\sum_{i=1}^{m} Entropy(S_i)} \quad (3)$$

dengan:

S : himpunan kasus atau dataset, $S = \{S_1, S_2, S_3, \dots, S_v\}$

n : jumlah partisi S

p_i : proporsi dari S_i terhadap S, $0 \leq i \leq n$

S_i : himpunan kasus pada kategori ke-i, $S_i = \{s \in S | A(s) = a_i\}$

A : atribut (wspd, tavg, pres, wdir, prcp), $A = \{a_1, a_2, a_3, \dots, a_v\}$

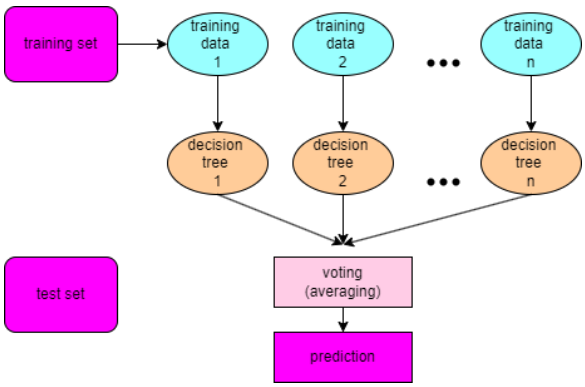
m : jumlah kelas pada atribut A

$|S_i|$: jumlah kasus pada partisi ke-i, $0 \leq i \leq n$

|S| : jumlah kasus dalam S

Proses dilakukan berulang pada setiap cabang hingga setiap kelas pada cabang memiliki kelas masing-masing [36]. Hal inilah yang membedakan algoritma C5.0 dengan algoritma decision tree sebelumnya seperti C4.5 yang berhenti sampai perhitungan gain [36]. Algoritma ini mampu memberikan keputusan yang bersifat akurat dan efisien namun kompleks. Akibatnya, terkadang bias pada data dengan nilai yang sangat besar [37]. Algoritma supervised ini cukup populer dan banyak digunakan untuk mengukur efek atribut dalam mengklasifikasikan dataset [38], termasuk menentukan faktor yang paling signifikan atau besar efeknya [39].

Random forest populer digunakan karena memberikan performa yang unggul dibandingkan metode klasifikasi tradisional, bergantung pada kualitas, kuantitas, dan distribusi data training dan data testing [40]. Random forest menggabungkan banyak decision tree untuk menghasilkan klasifikasi atau menangani persoalan regresi [41]. Algoritma ini dianggap lebih baik dibandingkan metode klasifikasi tunggal (decision tree) karena mengintegrasikan prediksi klasifikasi tunggal tersebut [42]. Akan tetapi, penentuan *hyperparameters* yang tidak tepat seperti jumlah pohon, kedalaman maksimum pohon, jumlah minimum sampel per leaf atau daun menyebabkan model menjadi kompleks atau rumit dan terjadi *overfitting* [43]. Ilustrasi cara kerja random forest dapat dilihat pada Gambar 2.

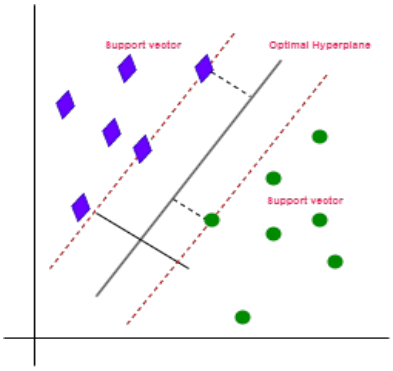


Gambar 2. Alur kerja random forest

SVM diperkenalkan sebagai model machine learning berbasis kernel untuk menangani klasifikasi dan regresi, terutama klasifikasi biner. SVM menggunakan hyperplane sebagai jarak antar kelas yang harus didapatkan secara maksimal [44]. Ilustrasi cara kerja SVM dapat dilihat pada Gambar 3.

Kemampuan generalisasi dan kekuatan diskriminatif (pemisahan kelas) menjadikan SVM sebagai salah satu metode yang paling sering digunakan beberapa tahun terakhir [45]. Di sisi lain, SVM terbatas pada dataset yang besar dan atau *noise* dan berakibat pada sensitivitas *outlier* sehingga mengarah pada

overfitting. Secara *default*, SVM tidak menyediakan estimasi probabilitas serta kebutuhan akan banyaknya *key parameters* yang mempengaruhi kinerja klasifikasi [46]. SVM diperkenalkan sebagai model *machine learning* berbasis kernel untuk menangani klasifikasi dan regresi, terutama klasifikasi biner. SVM menggunakan *hyperplane* sebagai jarak antar kelas yang harus didapatkan secara maksimal [44]. Ilustrasi cara kerja SVM dapat dilihat pada Gambar 3. Kemampuan generalisasi dan kekuatan diskriminatif (pemisahan kelas) menjadikan SVM sebagai salah satu metode yang paling sering digunakan beberapa tahun terakhir [45].



Gambar 3. SVM

Ketiga metode klasifikasi tersebut kemudian dibandingkan dengan metrik evaluasi berupa accuracy, precision, F1 score, dan recall [28, 47]. Hal ini dilakukan untuk melihat seberapa tepat kinerja algoritma dalam mengklasifikasikan data. Rincian confusion matrix untuk klasifikasi biner dapat dilihat pada tabel berikut[48]:

Tabel 1. Confusion Matrix

Class designation		*Actual class*	
		Yes	*No*
Predicted class	*Yes*	*True Positive*	*False Negative*
	No	*False Positive*	*True Negative*

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

$$Recall = \frac{TP}{TP + FP} \quad (5)$$

$$F1 - score = \frac{2\ x\ presisi\ x\ recall}{presisi\ x\ recall} \quad (6)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (7)$$

Data PM2.5 yang digunakan termasuk data harian Jakarta bulan januari 2021 hingga april 2024 yang bersumber dari Badan Meteorologi, Klimatologi dan Geofisika (BMKG) dan diakses lewat website https://aqicn.org/. Rincian klasifikasi data beserta penjelasannya dapat dilihat pada Tabel II. Makna setiap rentang nilai PM2.5 bersumber dari website https://www.iqair.com.

E-ISSN 2808-5841
P-ISSN 2808-7283

Tabel 2. Kategori Polutan PM2.5

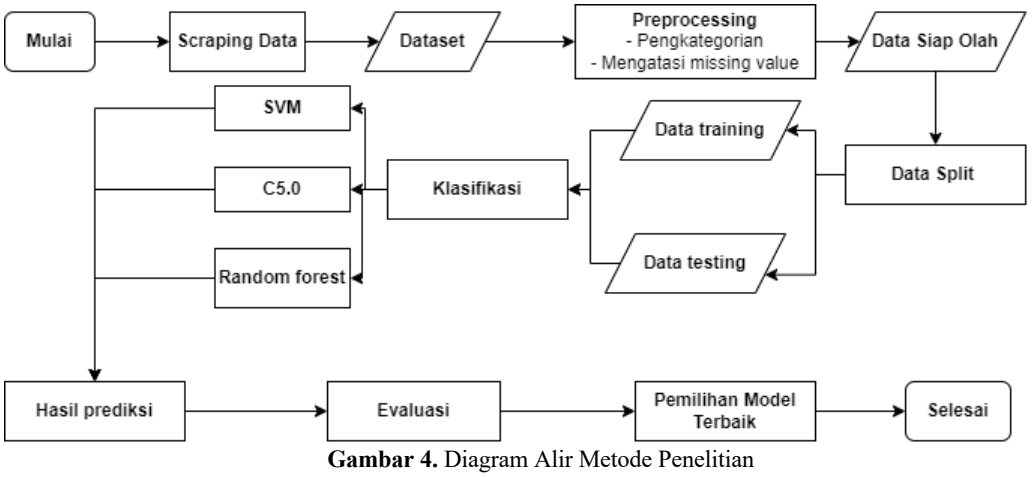
Level	PM2.5	Makna
Baik	0-15,5	Kualitas udara baik dan aman untuk melakukan aktivitas di luar ruangai
Sedang	15,6-55,4	Kualitas udara beresiko bagi kelompok tertentu (sensitif). Masih dapat melakukan aktivitas di luar ruangan namun kuantitasnya perlu diperhati
Tidak Sehat	55,5-150,4	Beresiko terjadi gangguan jantung dan paru-paru serta dampak kesehata lain pada masyarakat sehingga perlu menghindari aktivitas luar ruangan atau menggunakan masker.
Sangat Tidak Sehat	150,5-250,4	Kelompok sensitif akan mengalami penurunan daya tahan tubuh jika beraktivitas di luar ruangan. Sebaiknya tetap berada di dalam rumah.
Berbahaya	>250,4	Semua orang beresiko tinggi mengalami iritasi dan memicu penyakit kardiovaskular dan pernapasan.

Data meteorologi yang digunakan diambil dengan melakukan scraping pada website https://dev.meteostat.net/ menggunakan phyton. Deskripsi dan satuan data lebih rinci dapat dilihat pada Tabel III. Data yang dikumpulkan kemudian dianalisis dengan metode klasifikasi menggunakan Rstudio.

Tabel 3. Karakteristik Data

Variabel	Deskripsi	Satuan
PM2.5	partikel udara berukuran <= 2.5 mikrome	$\mu m/m^3$
Tavg (average temperature)	rata-rata suhu atau temperatur	°C
Prcp (total precipitation)	total presipitasi atau curah (air yang jatuh	mm
Wspd (wind Speed)	kecepatan angin	km/h
Wdir (wind direction)	arah angin	Degrees
Press(air pressure)	tekanan udara	hPa

Secara umum, tahapan metode penelitian dari fase pengambilan data hingga mendapatkan model terbaik dijabarkan secara runtun pada Gambar 2.



Gambar 4. Diagram Alir Metode Penelitian

Proses pekerjaan diawali dengan melakukan *preprocessing data. Preprocessing data* dimulai dengan melakukan *selection data*, yaitu memilih data dan variabel yang akan digunakan dan data mentah tersebut masih memiliki gangguan seperti *missing value* lalu diolah menjadi data yang siap digunakan (*data celan*) dalam proses analisis dengan cara penghapusan *missing value*. Langkah terakhir dalam *preprocessing data* adalah melakukan pengkategorian variabel respon. Pengkategorian variabel respon dilakukan agar data yang awalnya berupa data kontinu dapat diklasifikasikan secara tepat. Penentuan kategori didasarkan pada nilai PM2.5, di mana keseluruhan data variabel respon dibagi habis menjadi dua kategori, yaitu kategori dengan nilai PM2.5 yang aman untuk aktivitas di luar ruangan dan kategori dengan nilai PM2.5 yang tidak aman untuk aktivitas di luar ruangan. Pemilihan kategori baru didasarkan pada urgensi penelitian yang ingin membedakan kondisi udara yang baik bagi masyarakat untuk menghindari penyakit dari sisi polutan PM2.5.

Tabel 4. Hasil *Recode* PM2.5

Level	PM2.5	Kategori Baru
Baik	0 - 15,5	Safe
Sedang	15,6 - 55,4	
Tidak Sehat	55,5 - 150,4	Not Safe
Sangat Tidak Sehat	150,5 - 250,4	
Berbahaya	> 250,4	

Setelah *preprocessing data,* data secara habis dibagi menjadi dua, yaitu data *training* dan data *testing* sebelum melakukan klasifikasi. Data *training* merupakan data yang digunakan dalam penentuan dan pelatihan model klasifikasi, sedangkan data *testing* digunakan untuk mengevaluasi kinerja dari model yang telah terbentuk. Pada penelitian ini, terdapat beberapa set data *training* dan data *testing* yang nantinya akan dilakukan perbandingan untuk menentukan set data terbaik. Setelah dilakukan pembentukan data *training* dan data *testing*, maka data siap untuk diolah untuk pembentukan model klasifikasi. Pembentukan model klasifikasi dilakukan pada setiap set data *training* yang kemudian dievaluasi dengan set data *testing*.

Dalam melakukan klasifikasi dengan beberapa metode, perlu dilakukan evaluasi untuk menentukan metode dan perbandingan data set terbaik. Teknik evaluasi yang dapat digunakan adalah melihat nilai akurasi, presisi, *recall*, dan F-1 *score*. Proses evaluasi yang dilakukan adalah mencari metode dan perbandingan set data *training* serta *testing* dengan nilai akurasi tertinggi. Pada penelitian ini, akurasi tertinggi terdapat pada metode klasifikasi menggunakan algoritma C5.0 dengan proporsi data *training* adalah 95% dari total data. Proses yang dilakukan setelah melakukan evaluasi model klasifikasi adalah pengecekan *overfitting* untuk melihat adanya *overfit* atau kondisi di mana model terlalu cocok dengan data *training* sehingga tidak dapat mengolah data baru dengan baik. Model yang *overfit* memiliki kinerja yang sangat baik pada data training namun menghasilkan kinerja yang kurang baik pada data testing maupun data baru. Indikasi adanya *overfitting* adalah tingginya nilai akurasi pengecekan menggunakan data *training* dan rendahnya nilai akurasi pengecekan menggunakan data *testing*.

III. HASIL DAN PEMBAHASAN

Pembentukan model klasifikasi dilakukan pada setiap set data *training* yang kemudian dievaluasi dengan set data *testing*. Pada hasil penelitian yang tercantum pada Tabel 5, akurasi tertinggi, yaitu sebesar 81,48% pada metode klasifikasi menggunakan algoritma C5.0 dengan proporsi data *training*

adalah 95% dari total data. Metode dan perbandingan set data tersebut juga menghasilkan nilai presisi, *recall*, dan F1 *score* tertinggi dibandingkan metode dan perbandingan set data lainnya.

Tabel 5. Perbandingan Metode Klasifikasi

Perbandingan Data Training dan Data Testing	Evaluasi	Metode Klasifikasi		
		Random Forest	SVM	C5.0
70:30	Akurasi	70,81%	69,88%	70,19%
	Presisi	55,56%	60%	53,75%
	Recall	39,22%	14,71%	42,16%
	F1-Score	45,98%	23,62%	47,25%
80:20	Akurasi	69,77%	66,51%	68,37%
	Presisi	56,60%	50%	55,26%
	Recall	41,67%	12,5%	29,17%
	F1-Score	48%	20%	38,18%
90:10	Akurasi	67,59%	69,44%	75,93%
	Presisi	42,42%	36,36%	57,14%
	Recall	46,67%	13,33%	53,33%
	F1-Score	44,44%	19,51%	55,17%
95:5	Akurasi	70,37%	66,67%	81,48%
	Presisi	58,82%	60%	80%
	Recall	52,63%	15,79%	63,16%
	F1-Score	55,56%	25%	70,59%

Pada penelitian ini, nilai akurasi pengecekan menggunakan data *training* adalah 79,86% yang lebih kecil dari nilai akurasi pengecekan menggunakan data *testing*, yaitu 81,48%. Oleh karena itu, dapat dikatakan bahwa tidak adanya *overfitting* pada model klasifikasi algoritma C5.0 dengan data *training* sebesar 95% dari total data, sehingga model tersebut dapat mengklasifikan data baru dengan baik.

Metode klasifikasi dengan algoritma C5.0 menghasilkan *confusion matrix* yang menunjukkan kesesuaian pengkategorian prediksi dari data *testing*. Pada tabel 6 ditunjukkan bahwa 12 dari 19 observasi data kategori "Safe" diprediksikan secara benar ke dalam kategori "Safe", sedangkan sisanya diprediksi sebagai kategori "Not Safe". Untuk kategori "Not Safe" yang memiliki 35 observasi, 32 di antaranya dapat diklasifikasikan secara benar ke dalam kategori "Not Safe", sedangkan 3 observasi dikategorikan sebagai "Safe".

Tabel 6. Confusion Matrix C5.0

Referensi	Prediksi	
	Safe	Not Safe
Safe	12	7
Not Safe	3	32

Klasifikasi menggunakan algoritma C5.0 menghasilkan beberapa kriteria evaluasi yang tercantum pada Tabel 7. Nilai akurasi yang didapat sebesar 81,48%, sehingga dapat dikatakan bahwa 81,48% dari keseluruhan data dapat diprediksi ke dalam kategori yang benar atau sesuai dengan kategori referensi. Algoritma C5.0 juga menghasilkan presisi sebesar 80% yang mengukur seberapa baik model membuat prediksi secara benar untuk kategori "Safe" dari total prediksi ke dalam kategori "Safe" yang dilakukan. Nilai *recall* yang dihasilkan sebesar 63,16% yang menunjukkan persentase prediksi benar dari kategori "Safe" dari keseluruhan data referensi kategori "Safe". Nilai evaluasi yang terakhir adalah F1 *score* sebesar 70,59%. Nilai tersebut menunjukkan seberapa baik model dalam mengklasifikasikan ke dalam kategori "Safe" maupun "Not Safe".

Tabel 7. Nilai Evaluasi Klasifikasi C5.0

Akurasi	Presisi	Recall	F-1 Score
81,48%	80%	63,16%	70,59%

Untuk melihat pentingnya variabel terhadap pengklasifikasian dapat melihat *attribute usage* (penggunaan atribut) pada Tabel 8 yang menunjukkan seberapa banyak dan sering variabel digunakan dalam pembuatan model *decision tree*. Variabel yang sering digunakan dalam pembuatan model tentu merupakan variabel yang penting, di mana variabel yang penting mempunyai peranan yang signifikan dalam menentukan akurasi karena nilai akurasinya semakin baik ketika digunakan dalam algoritma *clustering* dan klasifikasi [49]. Hasil penelitian menunjukkan bahwa kecepatan angin merupakan variabel yang paling sering digunakan dalam penentuan klasifikasi PM2.5, di mana keseluruhan data kecepatan angin digunakan dalam pembentukan model. Variabel rata-rata suhu menyusul sebagai variabel yang paling banyak digunakan kedua dalam pemodelan klasifikasi. Dalam penelitian ini, variabel yang paling jarang digunakan dalam pembentukan *decision tree* pada algoritma C5.0 adalah total presipitasi.

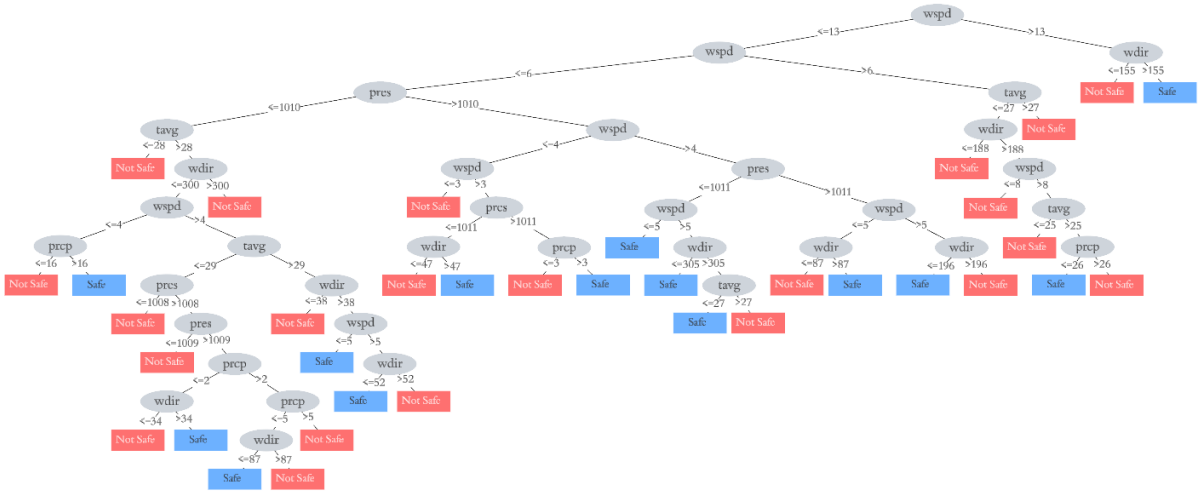
Tabel 8. Attribute Usage

wspd	tavg	pres	wdir	prcp
100%	84,28%	48,62%	43,52%	13,16%

Peningkatan konsentrasi polutan udara salah satu penyebabnya adalah faktor meteorologi [50]. Penelitian sebelumnya mengungkapkan bahwa suhu dan kecepatan angin umumnya dianggap sebagai faktor utama yang paling memengaruhi akumulasi PM dan gas polutan [51]. Hasil penelitian Chen et al, (2020) [52] juga menunjukkan bahwa angin dan suhu merupakan aspek yang paling sering menjadi faktor meteorologi dominan terhadap konsentrasi PM2.5 di di lokasi penelitiannya. Hal tersebut sesuai dengan penelitian ini, di mana kecepatan angin dan rata-rata suhu merupakan parameter yang paling sering digunakan dalam pemodelan klasifikasi PM2.5. Kecepatan angin erat kaitannya dengan kualitas udara karena berkontribusi terhadap pembersihan polutan [53] sedangkan suhu berpengaruh terhadap kualitas udara dalam menentukan kemampuan atmosfer untuk mengencerkan emisi [54].

Variabel yang menjadi penentu awal dalam pengklasifikasian kategori PM2.5 adalah kecepatan angin. Apabila kecepatan angin >13 km/h maka penentuan klasifikasi berikutnya didasarkan pada arah angin, yaitu ketika arah angin lebih dari 155 derajat maka PM2.5 akan diklasifikasikan sebagai "Safe" namun ketika arah angin kurang dari sama dengan 155 derajat maka PM2.5 akan diklasifikasikan sebagai "Not Safe". Akan tetapi, ketika kecepatan angin ≤13 km/h dan ≤6 km/h maka klasifikasi berikutnya berdasarkan tekanan udara sedangkan ketika kecepatan angin ≤13 km/h dan >6 km/h maka dasar klasifikasi selanjutnya adalah dengan rata-rata suhu. Proses tersebut akan terus berlanjut dengan dasar penentuan klasifikasi yang berbeda untuk setiap tahapnya. Dari Gambar 5 juga dapat dilihat bahwa kecepatan udara merupakan variabel yang sering menjadi penentu klasifikasi yang ditandai dengan sering munculnya variabel tersebut sebagai *node*.

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 5. Pohon Keputusan C5.0 Klasifikasi PM2.5

IV. KESIMPULAN

Pada penelitian ini dilakukan klasifikasi atas PM2.5 berdasarkan aspek-aspek meteorologi. Jumlah PM2.5 menjadi dasar pengkategorian kondisi kualitas udara apakah aman untuk berkegiatan di luar ruangan atau tidak. Terdapat beberapa metode yang digunakan dalam mengklasifikasikan PM2.5 yakni Random Forest, SVM, dan C5.0. Pada klasifikasi ini dilakukan beberapa skenario atas proporsi data training dan data testing yang digunakan yang dikombinasikan dengan ketiga metode yang kemudian metode terbaik. Pemilihan ini didasarkan pada tingkat akurasi yang dihasilkan metode tersebut dan didapat akurasi tertinggi adalah metode C5.0 dengan 95% data training. Terdapat 2 pola tingkat akurasi pada ketiga metode. Random forest dan C5.0 memiliki pola akurasi menurun hingga 90% data training dan meningkat pada 95% serta pola tidak beraturan pada SVM. Aspek wspd (*windspeed*) menjadi variabel yang paling sering digunakan pada decision tree hasil klasifikasi dengan metode C5.0.

REFERENSI

1. D. E. Schraufnagel *et al.*, "Health benefits of air pollution reduction," *Ann. Am. Thorac. Soc.*, vol. 16, no. 12, pp. 1478–1487, 2019, doi: 10.1513/AnnalsATS.201907-538CME.
2. R. Fuller *et al.*, "Pollution and health: a progress update," *Lancet Planet. Heal.*, vol. 6, no. 6, pp. e535–e547, 2022, doi: 10.1016/S2542-5196(22)00090-0.
3. F. M. R. da Silva Júnior, F. R. de Moura, R. de Lima Brum, and R. A. Tavella, "Air pollution—A look beyond big cities," *Integr. Environ. Assess. Manag.*, vol. 19, no. 2, pp. 295–297, 2023, doi: 10.1002/ieam.4720.
4. G. Shaddick, M. L. Thomas, P. Mudu, G. Ruggeri, and S. Gumy, "Half the world's population are exposed to increasing air pollution," *npj Clim. Atmos. Sci.*, vol. 3, no. 1, pp. 1–5, 2020, doi: 10.1038/s41612-020-0124-2.
5. Y. Ni, G. Shi, and J. Qu, "Indoor PM2.5, tobacco smoking and chronic lung diseases: A narrative review," *Environ. Res.*, vol. 181, no. September 2019, 2020, doi: 10.1016/j.envres.2019.108910.
6. P. Thangavel, D. Park, and Y. C. Lee, "Recent Insights into Particulate Matter (PM2.5)-Mediated Toxicity in Humans: An Overview," *Int. J. Environ. Res. Public Health*, vol. 19, no. 12, 2022, doi: 10.3390/ijerph19127511.
7. P. Ye, J. Li, W. Ma, and H. Zhang, "Impact of Collaborative Agglomeration of Manufacturing and Producer Services on Air Quality: Evidence from the Emission Reduction of PM2.5, NOx and SO2 in China," *Atmosphere (Basel).*, vol. 13, no. 6, 2022, doi: 10.3390/atmos13060966.
8. C. Li and S. Managi, "Contribution of on-road transportation to PM2.5," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, 2021, doi: 10.1038/s41598-021-00862-x.
9. W. Jung, J. S. Lee, S. Han, S. H. Ko, T. Kim, and Y. H. Kim, "An efficient reduced graphene-oxide filter

for PM2.5 removal," *J. Mater. Chem. A*, vol. 6, no. 35, pp. 16975–16982, 2018, doi: 10.1039/c8ta04587a.
10. R. Tong, J. Liu, W. Wang, and Y. Fang, "Health effects of PM2.5 emissions from on-road vehicles during weekdays and weekends in Beijing, China," *Atmos. Environ.*, vol. 223, no. July 2019, pp. 1–12, 2020, doi: 10.1016/j.atmosenv.2019.117258.
11. S. D. A. Kusumaningtyas, A. N. Khoir, E. Fibriantika, and E. Heriyanto, "Effect of meteorological parameter to variability of Particulate Matter (PM) concentration in urban Jakarta city, Indonesia," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 724, no. 1, pp. 1–7, 2021, doi: 10.1088/1755-1315/724/1/012050.
12. T. Istiana *et al.*, "Causality Analysis of Air Quality and Meteorological Parameters for PM2.5 Characteristics Determination: Evidence from Jakarta," *Aerosol Air Qual. Res.*, vol. 23, no. 9, pp. 1–18, 2023, doi: 10.4209/aaqr.230014.
13. N. Wardhani, H. Gani, S. Zuhriyah, H. Gani, and E. Vidyarini, "A Correlation Method for Meteorological Factors and Air pollution in association to covid-19 pandemic in the most affected city in Indonesia," *Ilk. J. Ilm.*, vol. 13, no. 3, pp. 195–205, 2021, doi: 10.33096/ilkom.v13i3.854.195-205.
14. M. Li *et al.*, "Exploring the regional pollution characteristics and meteorological formation mechanism of PM2.5 in North China during 2013–2017," *Environ. Int.*, vol. 134, no. July 2019, p. 105283, 2020, doi: 10.1016/j.envint.2019.105283.
15. W. L. Kusuma, W. Chih-Da, Z. Yu-Ting, H. H. Hapsari, and J. L. Muhamad, "Pm2.5 pollutant in asia—a comparison of metropolis cities in indonesia and taiwan," *Int. J. Environ. Res. Public Health*, vol. 16, no. 24, pp. 1–12, 2019, doi: 10.3390/ijerph16244924.
16. I. Gutiérrez-avila *et al.*, "Prediction of daily mean and one-hour maximum PM 2.5 concentrations and applications in Central Mexico using satellite-based machine-learning models," *J. Expo. Sci. Environ. Epidemiol.*, vol. 32, no. September 2022, pp. 917–925, 2022, doi: 10.1038/s41370-022-00471-4.
17. M. Yitshak-sade, J. F. Bobb, J. D. Schwartz, I. Kloog, and A. Zanobetti, "The association between short and long-term exposure to PM 2.5 and temperature and hospital admissions in New England and the synergistic effect of the short-term exposures," *Sci. Total Environ.*, vol. 639, pp. 868–875, 2018, doi: 10.1016/j.scitotenv.2018.05.181.
18. M. Yang *et al.*, "Polycyclic aromatic hydrocarbons (PAHs) associated with PM2.5 within boundary layer: Cloud/fog and regional transport," *Sci. Total Environ.*, vol. 627, pp. 613–621, 2018, doi: 10.1016/j.scitotenv.2018.01.014.
19. J. M. Yoo *et al.*, "New indices for wet scavenging of air pollutants (O3, CO, NO2, SO2, and PM10) by summertime rain," *Atmos. Environ.*, vol. 82, no. 2, pp. 226–237, 2014, doi: 10.1016/j.atmosenv.2013.10.022.
20. R. H. Virgianto, N. P. Kinanti, E. Ferdiansyah, and Q. A. Kartika, "The Effect of Precipitation on Scavenging of PM2.5 in Jakarta Based on Distributed Lag Non-Linear Models," *IPTEK J. Technol. Sci.*, vol. 32, no. 2, pp. 115–124, 2021, doi: 10.12962/j20882033.v32i2.7735.
21. J. Yang, B. Shi, Y. Shi, S. Marvin, Y. Zheng, and G. Xia, "Air pollution dispersal in high density urban areas: Research on the triadic relation of wind, air pollution, and urban form," *Sustain. Cities Soc.*, vol. 54, p. 101941, 2020, doi: 10.1016/j.scs.2019.101941.
22. N. Dahari, M. T. Latif, K. Muda, and N. Hussein, "Influence of meteorological variables on suburban atmospheric PM2.5 in the southern region of peninsular Malaysia," *Aerosol Air Qual. Res.*, vol. 20, no. 1, pp. 14–25, 2020, doi: 10.4209/aaqr.2019.06.0313.
23. B. Xu, W. Lin, and S. A. Taqi, "The impact of wind and non-wind factors on PM2.5 levels," *Technol. Forecast. Soc. Change*, vol. 154, no. September 2019, p. 119960, 2020, doi: 10.1016/j.techfore.2020.119960.
24. J. Xu *et al.*, "Grey correlation analysis of haze impact factor pm2.5," *Atmosphere (Basel).*, vol. 12, no. 11, pp. 1–15, 2021, doi: 10.3390/atmos12111513.
25. X. Shu and Y. Ye, "Knowledge Discovery : Methods from data mining and machine learning ☆," *Soc. Sci. Res.*, vol. 110, no. April 2022, p. 102817, 2023, doi: 10.1016/j.ssresearch.2022.102817.
26. J. Kleine Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, "Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters," *J. Electr. Comput. Eng.*, vol. 2017, 2017, doi: 10.1155/2017/5106045.
27. S. Saminathan and C. Malathy, "Ensemble-based classification approach for PM . concentration forecasting using meteorological data," *Front*, vol. Big Data, no. 6, p. 1175259, 2022.
28. I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," *IEEE Access*, vol. 6, pp. 33789–33795, 2018, doi: 10.1109/ACCESS.2018.2841987.

29. X. Kou, Z. Peng, M. Zhang, N. Zhang, and L. Lei, "Assessment of the Meteorological Impact on Improved PM 2 . 5 Air Quality Over North China During 2016 – 2019 Based on a Regional Joint Atmospheric Composition Reanalysis Data-Set," *J. Geophys. Res. Atmos.*, vol. 126, pp. 1–19, 2021, doi: 10.1029/2020JD034382.
30. K. Alpan, B. Sekeroglu, I. S. Engineering, P. Concentrations, M. Data, and S. City, "Prediction Of Pollutant Concentrations By Meteorological Data Using Machine Learning Algorithms," in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2020, vol. XLIV-4/W3-, no. October 2020, pp. 21–27, [Online]. Available: https://doi.org/10.5194/isprs-archives-XLIV-4-W3-2020-21-2020.
31. S. Ameer *et al.*, "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities," *IEEE Access*, vol. 7, pp. 128325–128338, 2019, doi: 10.1109/ACCESS.2019.2925082.
32. O. A. Ejohwomu *et al.*, "Modelling and Forecasting Temporal PM 2.5 Concentration Using Ensemble Machine Learning Methods," *Buildings*, vol. 12, no. 46, pp. 1–16, 2022.
33. W. Liu, G. Guo, F. Chen, and Y. Chen, "Meteorological pattern analysis assisted daily PM2.5 grades prediction using SVM optimized by PSO algorithm," *Atmos. Pollut. Res.*, vol. 10, no. 5, pp. 1482–1491, 2019, doi: 10.1016/j.apr.2019.04.005.
34. IQAir, "Air Quality Analysis and Statistics For Indonesia," 2019. .
35. K. V Uma, P. J. Padmaja, and D. Vinoodhini, "Stacked Feature Selection and C5.0 Classification Model with Tsallis Entropy for Medical Dataset," *J. Pharm. Negat. Results*, vol. 13, no. SO3, pp. 393–399, 2022, doi: 10.47750/pnr.2022.13.s03.065.
36. M. Kantardzic, "Data Mining Concept Models, Methods, and Algorithm," New Jersey, USA : A John Wiley & Sons, 2003.
37. C. Iorio, M. Aria, A. D'Ambrosio, and R. Siciliano, "Informative trees by visual pruning," *Expert Syst. Appl.*, vol. 127, pp. 228–240, 2019, doi: 10.1016/j.eswa.2019.03.018.
38. A. Z. Abdullah, B. Winarno, and D. R. S. Saputro, "The decision tree classification with C4.5 and C5.0 algorithm based on R to detect case fatality rate of dengue hemorrhagic fever in Indonesia," *J. Phys. Conf. Ser.*, vol. 1776, no. 1, pp. 1–10, 2021, doi: 10.1088/1742-6596/1776/1/012040.
39. J. H. Chen, H. H. Wei, C. L. Chen, H. Y. Wei, Y. P. Chen, and Z. Ye, "A practical approach to determining critical macroeconomic factors in air-traffic volume based on K-means clustering and decision-tree classification," *J. Air Transp. Manag.*, vol. 82, no. February 2019, p. 101743, 2020, doi: 10.1016/j.jairtraman.2019.101743.
40. C. Avci, M. Budak, N. Yagmur, and F. B. Balcik, "Comparison between random forest and support vector machine algorithms for LULC classification," *Int. J. Eng. Geosci.*, vol. 8, no. 1, pp. 1–10, 2023, doi: 10.26833/ijeg.987605.
41. Y. Zhang, S. Wei, L. Zhang, and C. Liu, "Comparing the Performance of Random Forest, SVM and Their Variants for ECG Quality Assessment Combined with Nonlinear Features," *J. Med. Biol. Eng.*, vol. 39, no. 3, pp. 381–392, 2019, doi: 10.1007/s40846-018-0411-0.
42. Q. Xu and J. Yin, "Application of Random Forest Algorithm in Physical Education," *Sci. Program.*, vol. 2021, pp. 1–10, 2021, doi: 10.1155/2021/1996904.
43. Y. Ao, H. Li, L. Zhu, S. Ali, and Z. Yang, "The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling," *J. Pet. Sci. Eng.*, vol. 174, no. August 2018, pp. 776–789, 2019, doi: 10.1016/j.petrol.2018.11.067.
44. F. Yesisca, D. E. Ratnawati, and B. Rahayudi, "Analisis Perbandingan Klasifikasi Topik Skripsi Mahasiswa menggunakan K-Nearest Neighbor dan Support Vector Machine (Studi Kasus: Jurusan Sistem Informasi, Fakultas Ilmu Komputer, Universitas Brawijaya)," vol. 6, no. 5, pp. 2328–2335, 2022, [Online]. Available: http://j-ptiik.ub.ac.id.
45. J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020, doi: 10.1016/j.neucom.2019.10.118.
46. A. F. A. H. Alnuaimi and T. H. K. Albaldawi, "An overview of machine learning classification techniques," *BIO Web Conf.*, vol. 97, pp. 1–24, 2024, doi: 10.1051/bioconf/20249700133.
47. B. Ayinla, "An Improved Collaborative Pruning Using Ant Colony Optimization and Pessimistic Technique of C5 . 0 Decision Tree Algorithm An Improved Collaborative Pruning Using Ant Colony Optimization and Pessimistic Technique of C5 . 0 Decision Tree Algorithm," *Int. J. Comput. Sci. Inf. Secur.*, vol. 18, no. December, pp. 111–123, 2020, doi: 10.5281/zenodo.4427699.
48. Ž. Vujović, "Classification Model Evaluation Metrics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp.

599–606, 2021, doi: 10.14569/IJACSA.2021.0120670.
49. U. Ojha, M. Jain, G. Jain, and R. K. Tiwari, "Significance of important attributes for decision making using C5.0," *8th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2017*, pp. 3–6, 2017, doi: 10.1109/ICCCNT.2017.8204031.
50. P. Chhabra, "The effect of different meteorological factors on the concentrations of air pollutants."
51. R. Li *et al.*, "Air pollution characteristics in China during 2015–2016: Spatiotemporal variations and key meteorological factors," *Sci. Total Environ.*, vol. 648, pp. 902–915, 2019, doi: 10.1016/j.scitotenv.2018.08.181.
52. Z. Chen *et al.*, "Influence of meteorological conditions on PM2.5 concentrations across China: A review of methodology and mechanism," *Environ. Int.*, vol. 139, no. July 2019, p. 105558, 2020, doi: 10.1016/j.envint.2020.105558.
53. N. Zhang, R. Ren, Q. Zhang, and T. Zhang, "Air pollution and tourism development: An interplay," *Ann. Tour. Res.*, vol. 85, no. March, 2020, doi: 10.1016/j.annals.2020.103032.
54. S. Oji and H. Adamu, "Correlation between air pollutants concentration and meteorological factors on seasonal air quality variation," *J. Air Pollut. Heal.*, vol. 5, no. 1, pp. 11–32, 2020, doi: 10.18502/japh.v5i1.2856.