Vol. 10, No. 1, April 2025, Page. 175-184
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Analisis Implementasi Support Vector Machine dan Random Forest untuk Prediksi Kategori Indeks Kualitas Udara Jakarta

Evander Banjarnahor[*1], Ronald Belferik[2], Wiputra Cendana[3], Yohanes Adi Saputra Abraham[4]

[1]Matematika, Universitas Pelita Harapan, Tangerang, Indonesia
[2]Informatika, Universitas Pelita Harapan, Tangerang, Indonesia
[3]Pendidikan Guru Sekolah Dasar, Universitas Pelita Harapan, Tangerang, Indonesia
[4]Center for Teaching and Learning, Universitas Pelita Harapan, Tangerang, Indonesia

Email: [1]evander.banjarnahor@uph.edu, [2]ronald.belferik@uph.edu, [3]wiputra.cendana@uph.edu, [4]yohanes.abraham@uph.edu

Abstrak

Kualitas udara yang buruk di Jakarta berdampak signifikan terhadap kesehatan masyarakat dan lingkungan. Oleh karena itu, diperlukan metode prediksi untuk membantu pengambilan kebijakan mitigasi polusi udara. Penelitian ini memprediksi kategori indeks kualitas udara dengan metode Support Vector Machine (SVM) dan Random Forest menggunakan data polutan (PM10, PM2.5, SO_2, CO, O_3, NO_2) dari Kaggle tahun 2021, meliputi PM10, PM2.5, SO2, CO, O3, dan NO2. Analisis korelasi menunjukkan bahwa PM10 dan PM2.5 memiliki hubungan yang sangat kuat (r = 0.96), menandakan keterkaitan erat dalam menentukan tingkat polusi udara. SVM dan Random Forest disimulasikan dengan berbagai rasio pembagian data latih dan uji (10:90, 15:85, 20:80, 25:75, dan 30:70), serta menggunakan *stratified k-fold cross-validation* untuk meningkatkan validitas hasil dan mengurangi potensi overfitting. Hasil evaluasi menunjukkan bahwa kedua model memberikan performa yang sangat baik dengan akurasi lebih dari 97% pada seluruh skenario pembagian data. Random Forest mencapai akurasi maksimum 100% pada rasio 15:85, sementara SVM mencatatkan akurasi tertinggi 98,9% pada rasio 25:75. Hasil *cross-validation* menunjukkan bahwa Random Forest mencapai akurasi 100% pada simulasi menggunakan *5-folds*, dengan nilai presisi, recall, dan F1-score yang juga 100%. Di sisi lain SVM menunjukkan akurasi sedikit lebih rendah yaitu 97,30% namun lebih konsisten dengan standar deviasi 2,50%.

Kata kunci: analisis korelasi, cross-validation, indeks kualitas udara, SVM, Random Forest

Abstract

Poor air quality in Jakarta significantly impacts public health and the environment. Therefore, accurate predictive methods are necessary to support policy-making for air pollution mitigation. This study predicts the Air Quality Index Categories using Support Vector Machine (SVM) and Random Forest methods with pollutant data from Kaggle (2021), including PM10, PM2.5, SO2, CO, O3, and NO2. Correlation analysis shows that PM10 and PM2.5 have a powerful relationship (r = 0.96), indicating a close relationship in determining air pollution levels. SVM and Random Forest were simulated with various ratios of training and test data splits (10:90, 15:85, 20:80, 25:75, and 30:70), and stratified k-fold cross-validation was used to enhance the validity of results and reduce the potential for overfitting. The evaluation results show that both models performed excellently with accuracies above 97% across all data split scenarios. Random Forest achieved a maximum accuracy of 100% at the 15:85 ratio, while SVM recorded a highest accuracy of 98.9% at the 25:75 ratio. Cross-validation results show that Random Forest achieved 100% accuracy in simulations using 5-folds, with precision, recall, and F1-score also reaching 100%. On the other hand, SVM showed a slightly lower accuracy of 97.30%, but was more consistent with a standard deviation of 2.50%.

Keywords: *correlation analysis, cross-validation, air quality index, Support Vector Machine, Random Forest*

Vol. 10, No. 1, April 2025, Page. 175-184
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

1. PENDAHULUAN

Kualitas udara merupakan isu krusial di kota-kota besar dunia, tak terkecuali Jakarta. Sebagai ibu kota Indonesia, Jakarta menghadapi tantangan kompleks terkait pencemaran udara yang berdampak signifikan terhadap kesehatan masyarakat, lingkungan, dan perekonomian. Kualitas udara yang baik sangat krusial bagi kelangsungan hidup manusia dan makhluk hidup lainnya. Udara yang bersih dan sehat akan melindungi kita dari berbagai risiko penyakit, seperti gangguan pernapasan, penyakit jantung, kanker, dan masalah kesehatan lainnya [1]. Menurut BPS Provinsi DKI Jakarta pada tahun 2022 data Indeks Kualitas Udara (IKU) Jakarta menunjukkan fluktuasi dan tren yang mengkhawatirkan dari tahun ke tahun. Hal ini menuntut adanya solusi yang efektif dan akurat untuk memprediksi kualitas udara sebagai dasar pengambilan keputusan yang tepat [2].

Penelitian oleh Silva dkk [3], mengatakan bahwa pengurangan aktivitas industri dan logistik selama pandemi membuat penurunan emisi polutan udara seperti nitrogen dioksida, sulfur dioksida, karbon monoksida, dan partikel debu, yang pada akhirnya berkontribusi pada dampak positif terhadap kualitas udara. Konsentrasi polutan di DKI Jakarta mencapai 57 mikrogram per meter kubik dan kualitas udara di wilayah tersebut sangat buruk [4]. Selama tiga tahun, dari 2017 hingga 2019, Jakarta mengalami penurunan kualitas udara yang signifikan. *Particulate Matter* (PM) merupakan komponen utama polusi udara yang berfungsi sebagai indikator kualitas udara serta berpotensi menimbulkan dampak negatif terhadap lingkungan dan kesehatan makhluk hidup [5]. Harjono dkk [6] berpendapat bahwa polusi udara di Jakarta disebabkan oleh sumber-sumber alami dan buatan manusia. Peristiwa alam seperti letusan gunung berapi, kebakaran hutan, dan aktivitas mikroba melepaskan polutan seperti asap, gas, dan debu. Penyebab utama buruknya kualitas udara di DKI Jakarta adalah emisi dari kendaraan bermotor yang menggunakan bahan bakar berbasis minyak bumi, yang merupakan salah satu kontributor terbesar terhadap polusi udara. Secara khusus, emisi karbon dioksida (CO2) dari kendaraan-kendaraan ini rata-rata 40-50% lebih tinggi dibandingkan dengan emisi hidrokarbon (HC), nitrogen oksida (NOx), dan sulfur dioksida (SO2) dari kendaraan-kendaraan yang sama [7]. Pertumbuhan penduduk, urbanisasi, pembangunan yang tidak seimbang, dan kurangnya kesadaran masyarakat merupakan faktor lainnya yang mempengaruhi pencemaran udara [8].

Prediksi kualitas udara menjadi krusial untuk memberikan informasi dini kepada masyarakat, membantu pemerintah dalam merumuskan kebijakan pengendalian pencemaran, serta memungkinkan industri untuk mengambil langkah-langkah preventif [9]. Berbagai metode telah dikembangkan untuk memprediksi kualitas udara, mulai dari metode statistik konvensional hingga pendekatan machine learning yang lebih canggih. Dalam beberapa tahun terakhir, machine learning telah muncul sebagai pendekatan yang menjanjikan dalam prediksi kualitas udara. Algoritma machine learning mampu menangani data yang kompleks dan berdimensi tinggi, serta mengekstrak pola-pola non-linear yang sulit ditangkap oleh metode konvensional. Gupta dkk [10] melakukan analisis regresi dengan support vector regression (SVR), random forest regression (RFR), dan CatBoost regression (CR) untuk memprediksi indek kualitas udara. Sama halnya dengan Kumar & Pande [11], melakukan prediksi polusi udara di kota-kota yang ada di India dengan model machine learning. Hasil dari penelitiannya menunjukkan bahwa XGboost memperoleh akurasi tertinggi hingga 91% dibandingkan dengan model lainnya seperti SVM dan Gaussian Naive Bayes. Berbeda dengan penelitian-penelitian tersebut, Purwono dan Gata menggunakan teknik optimasi hiperparameter grid search CV pada model klasifikasi SVM. Dari penelitian tersebut akurasi model sebelum optimasi adalah 73,31%, sedangkan setelah optimasi dengan grid search CV menjadi 94,8% [12].

Dua algoritma yang populer dan sering digunakan dalam prediksi kualitas udara adalah Support Vector Machine (SVM) dan Random Forest [13]. SVM adalah algoritma supervised learning yang efektif untuk klasifikasi dan regresi [14]. SVM bekerja dengan mencari hyperplane optimal yang memisahkan data ke dalam kelas-kelas yang berbeda dengan margin terbesar [15]. Keunggulan SVM terletak pada kemampuannya menangani data berdimensi tinggi dan mengatasi masalah overfitting. SVM juga memiliki kemampuan untuk menggunakan kernel trick untuk memetakan data ke dalam ruang berdimensi lebih tinggi, sehingga memungkinkan pemisahan data yang tidak linear [16]. Random Forest, di sisi lain, adalah ensemble learning method yang membangun banyak decision tree secara acak dan menggabungkan prediksi dari masing-masing pohon untuk menghasilkan prediksi akhir [17]. Random Forest memiliki keunggulan dalam menangani data yang kompleks, serta memberikan estimasi kepentingan fitur. Random Forest juga kurang rentan terhadap overfitting dibandingkan dengan decision

tree. Penelitian ini bertujuan untuk mengisi kesenjangan tersebut dengan membandingkan kinerja algoritma SVM dan Random Forest dalam memprediksi kualitas udara di DKI Jakarta. Hasil penelitian sebelumnya menunjukkan bahwa kedua algoritma, SVM dan Random Forest, menunjukkan kinerja yang baik dalam memprediksi kualitas udara di DKI Jakarta. SVM, dengan kemampuannya menangani data berdimensi tinggi dan mengatasi overfitting, mencapai akurasi yang kompetitif [18]. Random Forest, sebagai ensemble learning method, juga menunjukkan kinerja yang baik dengan kemampuannya menangani data yang kompleks dan memberikan estimasi kepentingan fitur. Dengan demikian, penelitian ini akan membandingkan 2 machine learning model, khususnya Support Vector Machine (SVM) dan dan Random Forest untuk memprediksi indeks standar pencemar udara (ISPU) di provinsi DKI Jakarta pada tahun 2021. Hasil penelitian ini dapat digunakan sebagai dasar untuk pengembangan sistem aplikasi prediksi kualitas udara yang lebih akurat dan efektif, serta untuk pengambilan kebijakan yang lebih baik dalam upaya perbaikkan kualitas udara di DKI Jakarta.

Penelitian sebelumnya telah banyak mengkaji penerapan metode machine learning untuk prediksi kualitas udara di berbagai wilayah dunia, termasuk penggunaan algoritma seperti Support Vector Machine (SVM) dan Random Forest. Namun demikian, masih terdapat gap yang perlu diperdalam khususnya dalam konteks data kualitas udara Jakarta yang memiliki fitur yang lebih lengkap dai tahun 2021. Berbeda dengan penelitian sebelumnya yang menggunakan data dalam rentang tahun yang luas dan melakukan imputasi untuk menangani data yang hilang, penelitian ini secara khusus hanya menggunakan data tahun 2021 tanpa imputasi data. Kondisi ini menghasilkan jumlah data yang relatif terbatas, yang memungkinkan eksplorasi lebih mendalam mengenai pengaruh ukuran dataset terhadap kinerja model.

Selain itu, penelitian ini menghadirkan kebaruan dengan membandingkan kinerja algoritma pada beberapa rasio pembagian data latih dan uji (10:90, 15:85, 20:80, 25:75, dan 30:70), serta mengaplikasikan stratified k-fold *cross-validation* khususnya dengan nilai k = 5 dan k = 10 untuk meningkatkan validitas hasil. Penelitian ini diharapkan mampu memberikan kontribusi signifikan dalam meningkatkan akurasi prediksi kualitas udara, sehingga membantu pengambilan keputusan kebijakan mitigasi polusi udara dan lingkungan yang lebih tepat dan efektif di DKI Jakarta.

2. DATA DAN METODE PENELITIAN

2.1. Data Penelitian

Dalam penelitian ini, dilakukan analisis performa model Support Vector Machine (SVM) dan Random Forest untuk memprediksi kualitas udara di Jakarta. Data yang digunakan bersifat open source yang diperoleh dari Kaggle dengan judul dataset "Air Quality Index in Jakarta" yang dapat diakses pada link berikut: https://www.kaggle.com/datasets/senadu34/air-quality-index-in-jakarta-2010-2021.

Dataset awal terdiri dari 4.383 record yang mencakup berbagai parameter pencemar udara, yaitu PM10, SO2, CO, O3, NO2, max, dan PM2.5. Sebelum melakukan pemrosesan data, dilakukan cleaning data dimana bagian-bagian yang hilang (*missing value*) di hapus. Keputusan untuk menghapus data dibandingkan melakukan imputasi didasarkan pada tingginya proporsi missing value pada fitur pm pada pm25, yaitu sekitar 4.018 record (91,6%). Penggunaan teknik imputasi dalam kondisi ini dinilai berisiko menimbulkan bias dan mengurangi representasi data aktual. Setelah preprocessing selesai dilakukan, data yang tersisa dan siap digunakan dalam penelitian ini menjadi sebanyak 365 *record*, atau kurang dari 10% dari dataset awal. Distribusi akhir record per kategori kualitas udara adalah sebagai berikut: Baik: 57 *record* (15,61%), Sedang: 294 *record* (80,54%) dan Tidak Sehat: 14 *record* (3,85%).
Tabel 1 berikut adalah ilustrasi penggalan dataset Indeks Pencemaran Kualitas Udara di Jakarta yang di simulasikan

Tabel 1. Ilustrasi Data Indeks Pencemaran Kualitas Udara di Jakarta

Date	Station	pm10	so2	co	o3	no2	max	critical	pm25	Category
01/01/2021	DKI1 (Bunderan HI)	38.0	29.0	6.0	31.0	13.0	53.0	PM25	53.0	SEDANG
02/01/2021	DKI1 (Bunderan HI)	27.0	27.0	7.0	47.0	7.0	47.0	O3	46.0	BAIK
03/01/2021	DKI1 (Bunderan HI)	44.0	25.0	7.0	40.0	13.0	58.0	PM25	58.0	SEDANG
04/01/2021	DKI1 (Bunderan HI)	30.0	24.0	4.0	32.0	7.0	48.0	PM25	48.0	BAIK
…	…	…	…	….	…	…	…	…	…	…

										TIDAK
30/08/2021	DKI1 (Bunderan HI)	79.0	27.0	17.0	39.0	41.0	112	PM25	112.0	SEHAT
28/12/2021	DKI1 (Bunderan HI)	51.0	53.0	15.0	18.0	13.0	65	PM25	65.0	SEDANG
29/12/2021	DKI1 (Bunderan HI)	31.0	54.0	10.0	24.0	11.0	54	SO2	49.0	SEDANG
30/12/2021	DKI1 (Bunderan HI)	55.0	53.0	16.0	23.0	14.0	71	PM25	71.0	SEDANG
31/12/2021	DKI1 (Bunderan HI)	62.0	52.0	23.0	20.0	14.0	85	PM25	85.0	SEDANG

2.2. Metode Penelitian

Dalam penelitian ini membandingan dua metode Machine Learning yang biasa digunakan yaitu Support Vector Machine dan R Random Forest.

2.2. 1 Support Vector Machine

Support Vector Machine (SVM) adalah algoritma machine learning yang populer digunakan untuk masalah klasifikasi dan regresi [19]. Algoritme ini berfungsi dengan menentukan hyperplane terbaik yang membagi data ke dalam berbagai kategori. Hyperplane yang dipilih bertujuan untuk memaksimalkan margin, yang mengacu pada jarak antara hyperplane dan titik data terdekat dari setiap kategori. SVM sangat efektif dan kuat ketika menangani data berdimensi tinggi [20]. Algoritma SVM dapat dijelaskan dengan rumus:

$$f(x) = \omega\varphi(x) + b \quad (1)$$

di mana ω dan b masing-masing mewakili vektor normal dan bias term. $\varphi(x)$ adalah fitur dalam ruang dimensi tinggi yang dapat diekspresikan sebagai fungsi pemetaan nonlinier [21].

2.2.2 Random Forest

Random Forest adalah metode pembelajaran ensemble yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi. Setiap pohon keputusan dibangun dengan menggunakan subset yang dipilih secara acak dari kumpulan data. Salah satu keuntungan signifikan dari hutan acak adalah batasan inherennya pada overfitting, yang meningkatkan kapasitas model untuk menggeneralisasi secara efektif pada data baru yang belum pernah ada sebelumnya [22]. Prediksi keseluruhan dari Random Forest dicapai dengan rata-rata (untuk regresi) atau melalui pemungutan suara (untuk klasifikasi) di antara semua pohon keputusan. Salah satu keuntungan dari Random Forest adalah kemampuannya untuk mengelola data berdimensi tinggi dan fitur-fitur yang saling berkorelasi. Salah satu Algoritma dari Random Forest yaitu [23].

$$G = 1 - \sum_{i=1}^{C} p_i^2 \quad (2)$$

Dimana:

- pi adalah probabilitas memilih titik data dari kelas i dalam dataset.

2.2.3 Evaluasi Model

Dalam mengevaluasi kinerja model machine learning, *accuracy*, *precision*, *recall*, dan *F1-score* digunakan dalam penelitian ini dengan persamaan berikut ini [24].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

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

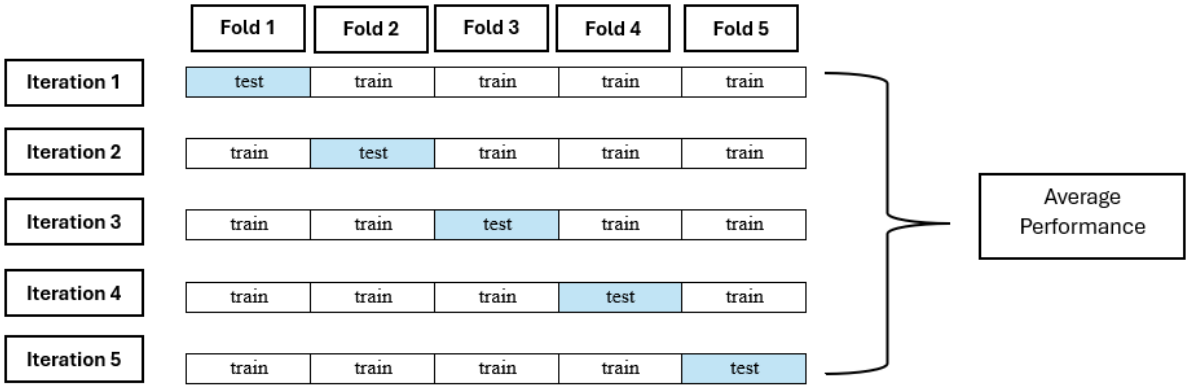
$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 - score = \frac{Precision \times Recall}{Precision+Recall} \quad (6)$$

di mana TP adalah *true positive rate*, TN adalah *true negative*, FP adalah *false positive rate*, dan FN adalah *false negative rate*[24].

Selain itu model juga di evaluasi dengan *k-fold cross validation*. *K-fold cross validation* adalah salah satu teknik dari validasi silang, yang mana memecah data menjadi *k* bagian set data dengan ukuran yang sama. Penggunaan *k-fold cross validation* untuk menghilangkan bias pada data [25]. Gambar 1 adalah penerapan *k-fold cross validation*

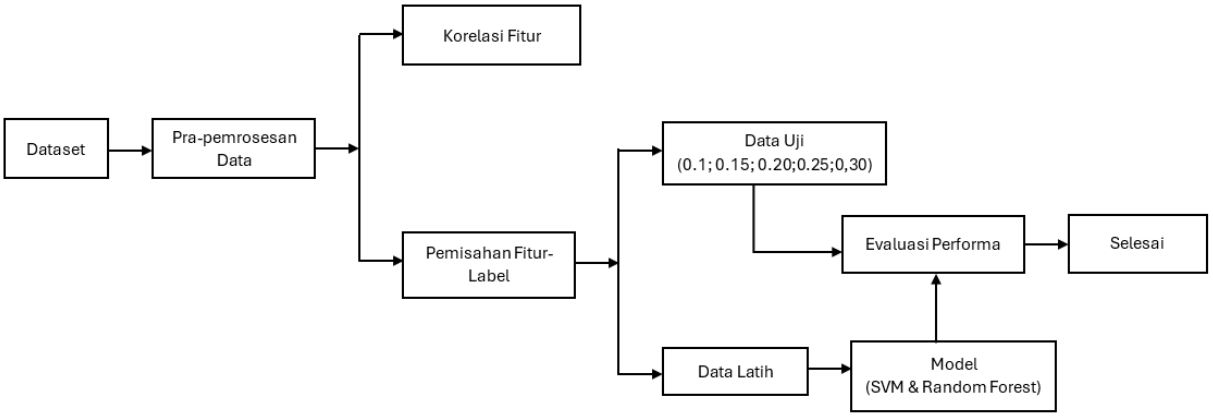
Vol. 10, No. 1, April 2025, Page. 175-184
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI



Gambar 1. Penerapan *k-fold cross validation*

2.2.4 Alur Penelitian

Adapun alur penelitian ini dapat dilihat pada Gambar 2 di bawah ini

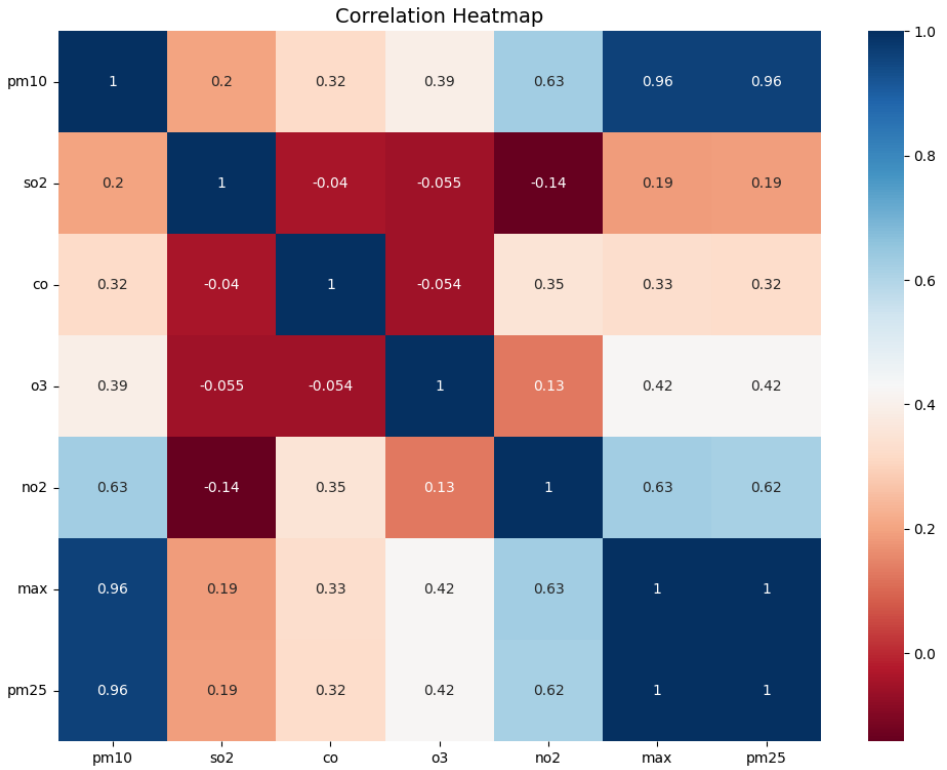


Gambar 2. Alur Penelitian

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, dilakukan analisis performa 2 machine learning model yaitu, Support Vector Machine (SVM) dan Random Forest untuk memprediksi kualitas udara di Jakarta. Sebelum pembagian data menjadi data latih dan data uji, dilakukan analisis korelasi untuk memahami hubungan antar variabel, sehingga dapat membantu dalam pemilihan fitur yang paling relevan untuk model prediksi. Gambar 3 menunjukkan peta korelasi (*correlation heatmap*) dari fitur-fitur pada indeks standar pencemaran udara (ISPU).

Vol. 10, No. 1, April 2025, Page. 175-184
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI



Gambar 3. Peta Korelasi fitur-fitur Indeks Standar Pencemaran Udara (ISPU)

Gambar 3 menunjukkan peta korelasi (correlation heatmap) yang menggambarkan hubungan linear antara berbagai parameter kualitas udara di Jakarta, termasuk PM10, SO2, CO, O3, NO2, nilai maksimum polusi udara (max) dan nilai PM2.5. Dari analisis korelasi, PM10 dan PM2.5 memiliki korelasi yang sangat tinggi sebesar 0.96, yang menunjukkan bahwa kedua parameter ini cenderung meningkat atau menurun bersamaan, kemungkinan karena sumber emisi yang serupa, seperti kendaraan bermotor dan aktivitas industry.

Selain itu, NO2 juga memiliki korelasi positif yang signifikan dengan PM10 dan PM2.5, masing-masing sebesar 0.63 dan 0.62, yang menunjukkan bahwa peningkatan polutan gas ini berkaitan erat dengan peningkatan partikulat di udara. Sementara itu, korelasi yang lebih rendah atau negatif terlihat antara SO2 dengan CO (-0.04) dan O3 (-0.055), yang menunjukkan bahwa faktor-faktor atmosferik atau reaksi kimia di udara mempengaruhi hubungan antar-polutan ini. O3 memiliki korelasi yang relatif rendah dengan polutan lainnya, dengan nilai tertinggi hanya sebesar 0.42 terhadap PM2.5 dan max, mengindikasikan bahwa faktor pembentukan ozon lebih dipengaruhi oleh kondisi lingkungan dibandingkan dengan sumber emisi langsung.

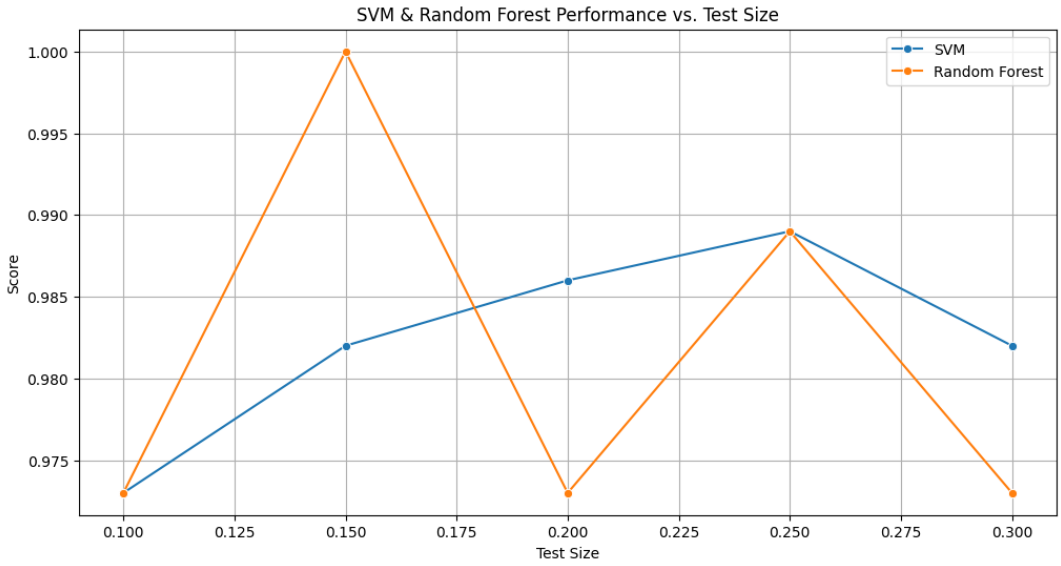
Setelah melakukan analisis korelasi, data dibagi menjadi data latih dan data uji dengan proporsi data uji sebesar 10%, 15%, 20%, 25%, dan 30%. Model Support Vector Machine (SVM) dan Random Forest digunakan untuk melakukan prediksi terhadap kualitas udara berdasarkan parameter yang telah dipilih. Tabel 2 menunjukkan hasil evaluasi dari metode SVM dan Random Forest untuk setiap rasio data uji dan data latih. Gambar 4 menunjukkan visualisasi dari perbadingan metrik akurasi, presisi, recall, dan F1-score dari model SVM dan random forest.

Tabel 2. Perbandingan Evaluasi Performa Model Random Forest dan Support Vector Machine

Test/Train Ratio	Model	Accuracy	Precision	Recall	F1 Score
10:90	Random Forest	97,30%	97,30%	97,30%	97,30%
15:85	Random Forest	100,00%	100,00%	100,00%	100,00%
20:80	Random Forest	97,30%	97,30%	97,30%	97,30%
25:75	Random Forest	98,90%	98,90%	98,90%	98,90%
30:70	Random Forest	97,30%	97,30%	97,30%	97,30%
10:90	SVM	97,30%	97,30%	97,30%	97,30%

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

15:85	SVM	98,20%	98,20%	98,20%	98,20%
20:80	SVM	98,60%	98,60%	98,60%	98,60%
25:75	SVM	98,90%	98,90%	98,90%	98,90%
30:70	SVM	98,20%	98,20%	98,20%	98,20%



Gambar 4. Perbandingan Evaluasi Performa Random Forest dan Support Vector Machine

Hasil evaluasi yang disajikan dalam Tabel 2 menunjukkan bahwa kedua model memiliki performa yang tinggi dalam memprediksi kualitas udara berdasarkan data yang tersedia. Akurasi model berkisar antara 97,30% hingga 100%, yang mengindikasikan bahwa metode yang digunakan mampu menghasilkan prediksi yang sangat baik. Random Forest menunjukkan akurasi sempurna (100%) pada rasio data uji sebesar 15%, yang menandakan bahwa model ini mampu mengklasifikasikan seluruh sampel uji dengan benar dalam skenario tersebut. Namun, pada rasio data uji lainnya, akurasi Random Forest sedikit menurun tetapi tetap berada di atas 97%, yang menunjukkan bahwa model memiliki kestabilan performa dalam berbagai skenario pembagian data. Metrik presisi, recall, dan F1-score juga menunjukkan konsistensi yang tinggi dengan nilai yang identik dengan akurasi, mencerminkan keseimbangan dalam mendeteksi berbagai kategori kualitas udara.

Di sisi lain, SVM juga menunjukkan performa yang sangat baik dengan akurasi yang berkisar antara 97,30% hingga 98,90%. Model ini menunjukkan peningkatan performa ketika rasio data latih lebih besar, dengan akurasi tertinggi (98,90%) pada rasio 25% data uji. Hal ini menandakan bahwa SVM memiliki ketahanan dalam menangani variasi data, tetapi masih sedikit lebih sensitif terhadap perubahan proporsi data dibandingkan Random Forest.

Namun, setelah menemukan hasil akurasi 100% pada Random Forest pada rasio data uji 15%, yang dapat mengindikasikan overfitting, dilakukan evaluasi ulang dengan Stratified K-Fold Cross-Validation untuk meningkatkan validitas hasil. *Cross-validation* ini memberikan gambaran yang lebih baik mengenai performa model dengan membagi data menjadi beberapa lipatan/*folds* dan menjaga distribusi kelas yang seimbang di setiap *folds*. Pada penelitian ini digunakan $k = 5$ dan $k = 10$ di mana penggunaan 5 *folds* dan 10 *folds* yang memungkinkan untuk menggunakan sebagian besar data sebagai data pelatihan dan menghasilkan evaluasi yang lebih akurat dan mengurangi potensi overfitting. Tabel 3 adalah perbandingan hasil pada model SVM dan Random Forest dengan menggunakan *5-fold* dan *10-fold cross validation*

Tabel 3. Perbandingan Hasil pada Model SVM dan Random Forest dengan Menggunakan *stratified k-fold cross validation*

Model	**Metric**	**5-folds**		**10-folds**	
		Mean Score	**Std Dev**	**Mean Score**	**Std Dev**
Random Forest	Accuracy	100,00%	00,00%	99,70%	0,80%

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Random Forest	Precision	100,00%	00,00%	99,70%	0,80%
Random Forest	Recall	100,00%	00,00%	99,70%	0,80%
Random Forest	F1-Score	100,00%	00,00%	99,70%	0,80%
SVM	Accuracy	97,30%	2,50%	97,00%	3,70%
SVM	Precision	97,30%	2,50%	97,00%	3,70%
SVM	Recall	97,30%	2,50%	97,00%	3,70%
SVM	F1-Score	97,30%	2,50%	97,00%	3,70%

Hasil evaluasi menggunakan cross-validation menunjukkan bahwa Random Forest mencapai akurasi 99,7% pada 10-folds, dengan nilai presisi, recall, dan F1-score juga masing-masing 99,7%, dan standar deviasi sebesar 0,80%. Pada *5-folds*, Random Forest menunjukkan performa sempurna dengan akurasi 100% pada semua metrik, yang menunjukkan konsistensi model yang sangat baik. Sebaliknya, SVM memiliki mean score yang sedikit lebih rendah, yaitu 97,00% pada *10-folds*, dengan standar deviasi yang lebih tinggi (3,70%), menunjukkan fluktuasi yang lebih moderat dibandingkan Random Forest. Pada 5-folds, SVM mencatatkan akurasi 97,30%, dengan standar deviasi yang lebih kecil (2,50%), menunjukkan bahwa model ini lebih konsisten dibandingkan dengan Random Forest.

Secara keseluruhan, kedua model menunjukkan performa yang sangat baik dalam memprediksi kualitas udara di Jakarta, dengan Random Forest lebih stabil dalam berbagai skenario dan SVM menunjukkan sedikit fluktuasi namun tetap kompetitif. Berdasarkan hasil ini, pemilihan model dapat disesuaikan dengan kompleksitas data dan kebutuhan analisis yang lebih spesifik. Penelitian ini memberikan wawasan mengenai efektivitas machine learning dalam memprediksi kualitas udara, yang dapat digunakan untuk mendukung kebijakan mitigasi polusi udara di Jakarta.

4. KESIMPULAN

Penelitian ini mengevaluasi prediksi kualitas udara (ISPU) di Jakarta menggunakan metode Support Vector Machine (SVM) dan Random Forest dengan berbagai rasio pembagian data latih dan uji. Analisis korelasi menunjukkan bahwa variabel PM10 dan PM2.5 memiliki korelasi yang sangat tinggi (0.96), yang mengindikasikan keterkaitan erat antara kedua partikel tersebut dalam menentukan tingkat polusi udara. Selain itu, variabel NO2 juga menunjukkan korelasi yang cukup kuat dengan PM10 dan PM2.5, yang menunjukkan bahwa polutan gas tersebut memiliki kontribusi signifikan terhadap kualitas udara. Sebaliknya, variabel SO2 dan O3 memiliki korelasi yang relatif lemah terhadap partikel-partikel polutan utama, yang menunjukkan pola distribusi dan dampak yang berbeda.

Hasil evaluasi performa model menunjukkan bahwa Random Forest dan SVM memberikan akurasi lebih dari 97% pada seluruh skenario pembagian data. Random Forest mencapai akurasi 100% pada rasio 15:85, sementara SVM mencapai 98,9% pada rasio 25:75. Namun, Random Forest menunjukkan akurasi sempurna pada rasio tertentu, yang mengindikasikan potensi overfitting. Untuk mengatasi hal ini, dilakukan evaluasi menggunakan stratified *k-fold cross-validation*. Pada hasil simulasi *5-folds* menunjukkan bahwa Random Forest tetap mempertahankan akurasi 100%, sementara SVM memiliki akurasi 97,30% dengan standar deviasi 2,50%, menunjukkan kestabilan yang lebih baik terhadap data baru. Selain itu, hasil evaluasi dengan *10-fold cross-validation* mengkonfirmasi hasil ini. Random Forest mencapai akurasi 99,7% dengan standar deviasi 0,80% pada semua lipatan, sementara SVM menunjukkan akurasi 97,00% dengan standar deviasi 3,70%, yang menunjukkan fluktuasi lebih besar dan sedikit lebih tinggi dibandingkan dengan Random Forest. Secara keseluruhan, kedua model menunjukkan performa yang sangat baik dan dapat diandalkan untuk memprediksi kualitas udara di Jakarta. Keda model ini juga dapat memberikan sumbangsi dalam kebijakan mitigasi polusi udara dan lingkungan di Jakarta.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

UCAPAN TERIMA KASIH

Penelitian ini didukung oleh Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) Universitas Pelita Harapan. Kami mengucapkan terima kasih atas dukungan dan fasilitas yang diberikan untuk kelancaran penelitian ini.

DAFTAR PUSTAKA

[1] T. Xayasouk and H. Lee, "AIR POLLUTION PREDICTION SYSTEM USING DEEP LEARNING," Jun. 2018, pp. 71–79. doi: 10.2495/AIR180071.

[2] S. Khan *et al.*, "Modelling the Impact of Road Dust on Air Pollution: A Sustainable System Dynamics Approach," *E3S Web of Conferences*, vol. 430, p. 01176, Oct. 2023, doi: 10.1051/e3sconf/202343001176.

[3] A. C. T. Silva, P. T. B. S. Branco, and S. I. V. Sousa, "Impact of COVID-19 Pandemic on Air Quality: A Systematic Review," *Int J Environ Res Public Health*, vol. 19, no. 4, p. 1950, Feb. 2022, doi: 10.3390/ijerph19041950.

[4] A. Amalia, A. Zaidiah, and I. N. Isnainiyah, "Prediksi Kualitas Udara Menggunakan Algoritma K-Nearest Neighbor," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 7, no. 2, pp. 496–507, May 2022, doi: 10.29100/jipi.v7i2.2843.

[5] N. F. Prih Waryatno, N. P. Kinanti, and Taryono, "Kondisi Pencemaran Udara pada Saat Periode Lebaran 2022 di Wilayah Jakarta," *Buletin GAW Bariri*, vol. 3, no. 2, Dec. 2022, doi: 10.31172/bgb.v3i2.68.

[6] C. Harjono, L. Gianto, R. Sidik, and D. L. Widaningrum, "FORECASTING AIR POLLUTION DRIVEN BY VEHICLE GROWTH, PUBLIC TRANSPORT, INDUSTRY, AND HOUSEHOLD WASTE," *Journal of Environmental Science and Sustainable Development*, vol. 7, no. 2, Dec. 2024, doi: 10.7454/jessd.v7i2.1272.

[7] P. Lestari, M. K. Arrohman, S. Damayanti, and Z. Klimont, "Emissions Inventory of Air Pollutants from Anthropogenic Sources in Jakarta," May 15, 2023. doi: 10.5194/egusphere-egu23-6686.

[8] A. H. R. Inaku and C. Novianus, "Pengaruh Pencemaran Udara PM 2,5 dan PM 10 Terhadap Keluhan Pernapasan Anak di Ruang Terbuka Anak di DKI Jakarta," *ARKESMAS (Arsip Kesehatan Masyarakat)*, vol. 5, no. 2, pp. 9–16, Dec. 2020, doi: 10.22236/arkesmas.v5i2.4990.

[9] S. Jia, "Effect of combined strategy on mitigating air pollution in China," *Clean Technol Environ Policy*, vol. 23, no. 3, pp. 1027–1043, Apr. 2021, doi: 10.1007/s10098-020-02013-8.

[10] A. Gupta, H. K. Mall, and S. Janarthanan., "Rainfall Prediction Using Machine Learning," in *2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR)*, IEEE, Mar. 2022, pp. 1–5. doi: 10.1109/ICAITPR51569.2022.9844203.

[11] K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities," *International Journal of Environmental Science and Technology*, vol. 20, no. 5, pp. 5333–5348, May 2023, doi: 10.1007/s13762-022-04241-5.

[12] A. Toha, P. Purwono, and W. Gata, "Model Prediksi Kualitas Udara dengan Support Vector Machines dengan Optimasi Hyperparameter GridSearch CV," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 4, no. 1, pp. 12–21, May 2022, doi: 10.12928/biste.v4i1.6079.

[13] Anisa Ma'u Luthfi and Fatkhurokhman Fauzi, "Perbandingan Klasifikasi Random Forest, Support Vector Machines, dan LGBM Pada Klasifikasi Kualitas Udara di Jakarta," *JUSTINDO (Jurnal Sistem dan Teknologi Informasi Indonesia)*, vol. 9, no. 2, pp. 99–108, Aug. 2024, doi: 10.32528/justindo.v9i2.1912.

[14] A. S. Handayani, S. Soim, T. E. Agusdi, and N. L. Husni, "Air Quality Classification Using Support Vector Machine," *Computer Engineering and Applications Journal*, vol. 10, no. 1, pp. 55–69, Feb. 2021, doi: 10.18495/comengapp.v10i1.350.

[15] S. Ketu and P. K. Mishra, "Scalable kernel-based SVM classification algorithm on imbalance air quality data for proficient healthcare," *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2597–2615, Oct. 2021, doi: 10.1007/s40747-021-00435-5.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

[16] Z. Almheiri, M. Meguid, and T. Zayed, "Failure modeling of water distribution pipelines using meta-learning algorithms," *Water Res*, vol. 205, p. 117680, Oct. 2021, doi: 10.1016/j.watres.2021.117680.

[17] S. Gocheva-Ilieva, A. Ivanov, and M. Stoimenova-Minova, "Prediction of Daily Mean PM10 Concentrations Using Random Forest, CART Ensemble and Bagging Stacked by MARS," *Sustainability*, vol. 14, no. 2, p. 798, Jan. 2022, doi: 10.3390/su14020798.

[18] H. Sunaryanto, M. A. Hasan, and G. Guntoro, "Classification Analysis of Unilak Informatics Engineering Students Using Support Vector Machine (SVM), Iterative Dichotomiser 3 (ID3), Random Forest and K-Nearest Neighbors (KNN)," *IT Journal Research and Development*, vol. 7, no. 1, pp. 36–42, Aug. 2022, doi: 10.25299/itjrd.2022.8912.

[19] D. Mustafa Abdullah and A. Mohsin Abdulazeez, "Machine Learning Applications based on SVM Classification A Review," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 81–90, Apr. 2021, doi: 10.48161/qaj.v1n2a50.

[20] A. Louati, H. Louati, E. Kariri, F. Alaskar, and A. Alotaibi, "Sentiment Analysis of Arabic Course Reviews of a Saudi University Using Support Vector Machine," *Applied Sciences*, vol. 13, no. 23, p. 12539, Nov. 2023, doi: 10.3390/app132312539.

[21] J. Wang and J. Hu, "A robust combination approach for short-term wind speed forecasting and analysis – Combination of the ARIMA (Autoregressive Integrated Moving Average), ELM (Extreme Learning Machine), SVM (Support Vector Machine) and LSSVM (Least Square SVM) forecasts using a GPR (Gaussian Process Regression) model," *Energy*, vol. 93, pp. 41–56, Dec. 2015, doi: 10.1016/j.energy.2015.08.045.

[22] S. M. Simon, P. Glaum, and F. S. Valdovinos, "Interpreting random forest analysis of ecological models to move from prediction to explanation," *Sci Rep*, vol. 13, no. 1, p. 3881, Mar. 2023, doi: 10.1038/s41598-023-30313-8.

[23] X. Xie, M.-J. Yuan, X. Bai, W. Gao, and Z.-H. Zhou, "On the Gini-impurity Preservation For Privacy Random Forests," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Curran Associates, Inc., 2023, pp. 45055–45082. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/8d6b1d775014eff18256abeb207202ad-Paper-Conference.pdf

[24] N. Syahira and D. B. Arianto, "PREDIKSI TINGKAT KUALITAS UDARA DENGAN PENDEKATAN ALGORITMA K-NEAREST NEIGHBOR," *Jurnal Ilmiah Informatika Komputer*, vol. 29, no. 1, pp. 45–59, Apr. 2024, doi: 10.35760/ik.2024.v29i1.10069.

[25] D. Fitrianah, S. Dwiasnati, H. H. H, and K. A. Baihaqi, "Penerapan Metode Machine Learning untuk Prediksi Nasabah Potensial menggunakan Algoritma Klasifikasi Naïve Bayes," *Faktor Exacta*, vol. 14, no. 2, p. 92, Aug. 2021, doi: 10.30998/faktorexacta.v14i2.9297.