JOMLAI: Journal of Machine Learning and Artificial Intelligence
Vol. 4, No. 2, Juni 2025, pp. 116~126
ISSN: 2828-9102 (Print) | 2828-9099 (Online)
DOI: 10.55123/jomlai.v4i2.5369

Evaluasi Perbandingan Algoritma MLR dan SVM untuk Prediksi Kualitas Udara DKI Jakarta

Comparative Evaluation of MLR and SVM Algorithms for DKI Jakarta Air Quality Prediction

Arfany Dhimas Muftareza
Sekolah Tinggi Meteorologi Klimatologi dan Geofisika, Indonesia

Article Info

Genesis Artikel:
Diterima, 24 April 2025
Direvisi, 17 Mei 2025
Disetujui, 20 Juni 2025

Kata Kunci:
Kualitas Udara
Machine Learning
Prediksi
MLR
SVM

ABSTRAK

Penelitian ini mengeksplorasi penerapan Machine Learning menggunakan algoritma Multiple Linear Regression (MLR) dan Support Vector Machine (SVM) untuk memprediksi kategori kualitas udara di Jakarta berdasarkan parameter polutan utama, seperti PM10, PM2.5, NO2, CO, SO2, dan O3. Dataset yang digunakan berasal dari data ISPU yang diukur dari lima stasiun pemantauan kualitas udara di Provinsi DKI Jakarta pada tahun 2021. Proses penelitian meliputi pengumpulan data, pembersihan data, implementasi model menggunakan pustaka scikit-learn, dan evaluasi kinerja model menggunakan metrik Accuracy, R-Squared, RMSE, dan MAE. Hasil evaluasi kinerja model menunjukkan bahwa SVM memiliki kinerja yang lebih baik daripada MLR, sebagaimana dibuktikan oleh nilai akurasi yang lebih tinggi (91,78% vs. 90,41%), nilai R-kuadrat yang lebih tinggi (69,63% vs. 64,56%), nilai RMSE yang lebih rendah (0,2867 vs. 0,3097), dan nilai MAE yang lebih rendah (0,0822 vs. 0,0959), yang menunjukkan bahwa kesalahan dalam model SVM lebih kecil daripada MLR. Studi ini membuktikan efektivitas model berbasis pembelajaran mesin dalam memberikan prediksi kategori kualitas udara yang akurat, meskipun masih ada tantangan dalam memprediksi kategori "Baik" yang memerlukan pengembangan lebih lanjut, seperti penyeimbangan data dan rekayasa fitur tingkat lanjut untuk meningkatkan akurasi prediksi semua kategori.

Keywords:
Air Quality
Machine Learning
Prediction
MLR
SVM

ABSTRACT

This research explores the application of Machine Learning using Multiple Linear Regression (MLR) and Support Vector Machine (SVM) algorithms to predict air quality categories in Jakarta based on key pollutant parameters, such as PM10, PM2.5, NO2, CO, SO2, and O3. The dataset used comes from ISPU data measured from five Air quality monitoring stations in DKI Jakarta Province in 2021. The research process includes data collection, data cleaning, model implementation using the scikit-learn library, and model performance evaluation using Accuracy, R-Squared, RMSE, and MAE metrics. The results of model performance evaluation show that SVM performs better than MLR, as evidenced by higher accuracy value (91.78% vs. 90.41%), higher R-squared value (69.63% vs. 64.56%), lower RMSE value (0.2867 vs. 0.3097), and lower MAE value (0.0822 vs. 0.0959), indicating that the error in SVM model is smaller than MLR. This study proves the effectiveness of machine learning-based models in providing accurate air quality category predictions, although there are still challenges in predicting the "Good" category that require further development, such as balancing data and advanced feature engineering to improve the prediction accuracy of all categories.

Penulis Korespondensi:
Arfany Dhimas Muftareza,
Program Studi Instrumentasi-MKG,
Sekolah Tinggi Meteorologi Klimatologi dan Geofisika,
Email: arfanydhimuf@gmail.com

1. PENDAHULUAN

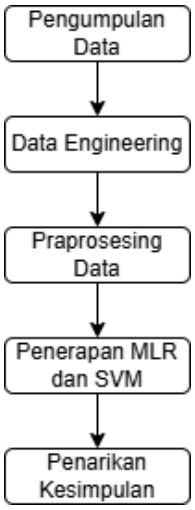
Dunia telah berkembang pesat dalam hal industri dan teknologi di era pertumbuhan populasi dan permintaan yang maju ini. Namun, kelemahan dari kemajuan ini sering kali diabaikan. Contohnya termasuk pembakaran bahan bakar fosil, emisi gas berbahaya dari mobil yang tidak terkendali, dan pembuangan limbah industri. Aktivitas-aktivitas ini berkontribusi terhadap

masalah polusi udara global, yang memperburuk kualitas udara. Polusi udara terjadi ketika zat-zat berbahaya dilepaskan ke atmosfer, membahayakan kesehatan manusia dan organisme hidup lainnya[1]. Menurut laporan Dinas Lingkungan Hidup DKI Jakarta, indeks kualitas udara Jakarta sering kali berada dalam kategori "Tidak Sehat" terutama pada musim kemarau. Selain itu, studi dari IQAir tahun 2022 menunjukkan bahwa Jakarta termasuk dalam 10 kota besar dengan kualitas udara terburuk di dunia selama beberapa bulan dalam setahun. Dalam beberapa dekade terakhir, Jakarta telah mengalami urbanisasi yang substansial, perluasan, dan peningkatan jumlah kendaraan di jalan, yang semuanya telah menyebabkan konsumsi energi yang lebih tinggi[2].

Karena dampak kualitas udara yang buruk, risiko kematian akibat berbagai penyakit dapat meningkat, yang berkontribusi terhadap sebanyak 7 juta kematian secara global setiap tahun[3]. Untuk mengatasi masalah ini, prediksi kualitas udara menjadi penting sehingga tindakan mitigasi dapat dilaksanakan. Dalam beberapa tahun terakhir, para peneliti semakin memanfaatkan pembelajaran mesin untuk mengatasi tantangan polusi udara perkotaan. Pembelajaran mesin adalah bagian dari kecerdasan buatan yang memungkinkan sistem untuk belajar dari data daripada pemrograman eksplisit[4]. Penerapan pembelajaran mesin sebagai metode prediktif menjadi semakin populer karena dapat menangani sejumlah besar data dan mengidentifikasi pola yang rumit[5]. Dengan memanfaatkan algoritma pembelajaran mesin untuk menganalisis kumpulan data yang luas dari sumber-sumber seperti stasiun pemantauan, metode ini meningkatkan sistem pemantauan kualitas udara[6]. Pendekatan konvensional untuk memprediksi kualitas udara memiliki kekurangan dalam mengatasi faktor-faktor rumit yang memengaruhi kualitas udara.

Penelitian ini bertujuan untuk menerapkan algoritma Regresi Linier Berganda (MLR) dan Support Vector Machine (SVM) untuk memperkirakan kategori kualitas udara. MLR berfungsi sebagai alat yang efektif untuk memodelkan hubungan antara beberapa variabel independen, dalam hal ini PM10, PM2.5, NO2, CO, SO2, dan O3, dengan kategori kualitas udara sebagai variabel dependen[7][8]. Selain itu, SVM digunakan karena kemampuannya yang tinggi dalam menangani data non-linier dan memberikan prediksi yang baik. Support Vector Machine (SVM) adalah teknik pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Ini beroperasi dengan menemukan hyperplane optimal yang memisahkan kelas data dengan margin maksimum, meningkatkan kemampuan generalisasi dalam prediksi[9]. Penelitian oleh Weizhen Lu dan rekan-rekannya (2002) menunjukkan bahwa Support Vector Machines (SVM) lebih unggul dibandingkan Radial Basis Function (RBF) dalam memprediksi konsentrasi polutan udara di Hong Kong. Meskipun disebutkan bahwa parameter bebas SVM berpengaruh terhadap performa model, penelitian ini tidak mendalami teknik optimasi parameter secara sistematis[10].

2. METODE PENELITIAN



Gambar 1. Diagram Alir Penelitian

Untuk memprediksi kualitas udara menggunakan metode regresi linier berganda (MLR) dan Support Vector Machine (SVM), terdapat beberapa tahapan penting yang dilakukan. Pengumpulan data memanfaatkan sumber sekunder yang diperoleh dari pemerintah atau organisasi swasta, yang memiliki keunggulan dalam ketersediaan dan efisiensi biaya, meskipun tetap diperlukan evaluasi terhadap kualitas dan relevansinya[21]. Data kualitas udara di DKI Jakarta diperoleh dari Kaggle, mencakup parameter utama seperti PM10, PM2.5, NO2, CO, SO2, O3, dan kategori kualitas udara. Dalam penelitian ini, variabel konsentrasi polutan yang berpengaruh terhadap kualitas udara ditetapkan sebagai variabel independen, sedangkan kategori kualitas udara dijadikan sebagai variabel dependen[22]. Transformasi data dilakukan dengan mengonversi tipe data ke format yang sesuai, seperti datetime untuk analisis berbasis waktu serta format numerik untuk kategori kualitas udara. Proses pembersihan data mencakup penanganan data hilang, duplikasi, dan outlier menggunakan metode imputasi dengan nilai rata-rata dari parameter serta deteksi outlier melalui boxplot[23][24]25][26]. Setelah itu, data dibagi menjadi set pelatihan dan pengujian untuk membangun serta mengevaluasi model[27]. Model MLR dan SVM diterapkan menggunakan pustaka Scikit-Learn, dengan evaluasi kinerja berdasarkan nilai akurasi, R-squared, Root Mean Squared Error, dan Mean Absolute Error untuk menilai efektivitas kedua pendekatan[16][28].

2.1 Indeks Kualitas Udara

Statistik yang dapat dipahami untuk menilai kualitas udara dan dampaknya terhadap kesehatan manusia adalah indeks kualitas udara. Dengan aplikasi di ponsel mereka seperti AirVisual, warga dapat dengan cepat memperoleh AQI waktu nyata. Indeks kualitas udara Amerika Serikat (AS) berfungsi sebagai dasar untuk indeks kualitas udara yang umum digunakan di Indonesia. Indeks ini berasal dari studi komprehensif yang meneliti hubungan antara polusi udara di AS dan status kesehatan penduduknya. Namun karena AS dan Indonesia memiliki kondisi iklim dan komposisi genetik yang berbeda, AQI tidak secara akurat mewakili kualitas udara di Indonesia[14]. Oleh karena itu, setiap negara harus menentukan ambang batas indeks kualitas udara untuk wilayahnya, disesuaikan dengan faktor lingkungan dan komposisi genetik yang dimilikinya. Indonesia akan memodifikasi nilai ambang batas dan kategori dampak kesehatan berdasarkan studi regional yang mempertimbangkan unsur-unsur lingkungan untuk memberikan AQI yang secara akurat mewakili kondisi lokal.

2.2 Indeks Standar Pencemaran Udara (ISPU)

Peraturan Nomor 14 Tahun 2020 tentang Indeks Baku Pencemar Udara diterbitkan oleh Kementerian Lingkungan Hidup dan Kehutanan pada tahun 2020. Peraturan ini menggantikan Keputusan Menteri Lingkungan Hidup Nomor 45 Tahun 1997 tentang Perhitungan, Pelaporan, dan Informasi Indeks Baku Pencemar Udara. Dalam peraturan baru tersebut, perhitungan Indeks Baku Mutu Udara (ISPU) didasarkan pada tujuh parameter, yaitu PM10, PM2.5, NO2, CO, SO2, O3, dan HC. Dua parameter telah ditambahkan dari peraturan sebelumnya karena HC dan PM2.5 memiliki risiko kesehatan yang signifikan. Kualitas udara ambien suatu wilayah dapat dijelaskan menggunakan ISPU, yaitu angka tanpa satuan. ISPU didasarkan pada pengaruh terhadap makhluk hidup, daya tarik estetika, dan kesehatan masyarakat. ISPU dibuat untuk membantu masyarakat umum dalam melakukan penyeragaman data tentang keadaan udara ambien pada waktu dan tempat tertentu. Hal ini juga diharapkan dapat diperhitungkan oleh pemerintah federal dan pemerintah daerah ketika mereka bersiap untuk mengendalikan polusi udara[15]. AQI Amerika menggunakan ambang batas berdasarkan penelitian di wilayah beriklim subtropis, sementara ISPU Indonesia memperhitungkan konteks lokal yang berkontribusi terhadap perbedaan sensitivitas indeks terhadap polutan.

Tabel 1. Konversi Nilai Konsentrasi Parameter Kualitas Udara

ISPU	PM_{10} ($\mu g/m^3$)	$PM_{2.5}$ ($\mu g/m^3$)	SO_2 ($\mu g/m^3$)	CO ($\mu g/m^3$)	O_3 ($\mu g/m^3$)	NO_2 ($\mu g/m^3$)	HC ($\mu g/m^3$)	Informasi
0 – 50	50	15.5	52	4000	120	80	45	Baik
51 – 100	150	55.4	180	8000	235	200	100	Sedang
101 – 200	350	150.4	400	15000	400	1130	215	Tidak Sehat
201 – 300	420	250.4	800	30000	800	2260	432	Sangat Tidak Sehat
>300	500	500	1200	45000	1000	3000	648	Berbahaya

2.3 Regresi Linear Berganda

Salah satu metode untuk analisis dan prediksi data adalah regresi linier. Regresi linier sederhana melibatkan pembuatan model bivariat untuk memprediksi variabel respons menggunakan satu variabel penjelas. Di sisi lain, regresi linier berganda membentuk model multivariat dengan memasukkan beberapa variabel penjelas. Dengan memperkenalkan variabel penjelas tambahan, regresi linier berganda memperluas prinsip regresi linier sederhana. Istilah "linier" berlaku untuk kedua metode karena diasumsikan bahwa variabel respons terkait erat dengan kombinasi linier dari variabel penjelas[16]. Mirip dengan persamaan regresi linier dasar, persamaan regresi linier berganda mengandung istilah tambahan:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Gambar 2. Persamaan Regresi Linear Berganda

Dalam konteks ini, Y menunjukkan variabel dependen yang dimaksudkan untuk prediksi. X1, X2, …, Xn mewakili variabel independen yang digunakan untuk meramalkan Y. Istilah β0 mengacu pada intersep atau konstanta, yang menunjukkan nilai Y ketika semua variabel independen ditetapkan menjadi nol. Koefisien β1, β2, …, βn adalah koefisien regresi yang menunjukkan perubahan rata-rata dalam Y untuk setiap peningkatan satu unit dalam variabel independen masing-masing, dengan asumsi variabel independen lainnya tetap tidak berubah. Akhirnya, istilah kesalahan atau residual menandakan variasi dalam Y yang tidak dapat dijelaskan oleh variabel independen[16]. Meskipun MLR umumnya digunakan untuk prediksi variabel kontinu, dalam studi ini digunakan untuk memprediksi kategori ISPU yang telah diubah menjadi variabel numerik diskret.

2.4 Support Vector Machine (SVM)

Ruang hipotetis dalam bentuk fungsi linier di semua ruang fitur berdimensi tinggi digunakan untuk klasifikasi oleh Support Vector Machine (SVM), sistem pembelajaran terbimbing yang dilatih menggunakan semua algoritma pembelajaran berdasarkan teori optimasi dan semua bias pembelajaran. Karena SVM lebih mampu menggeneralisasi data daripada metode sebelumnya, sistem ini dibuat untuk mengatasi masalah klasifikasi[17]. Menemukan hiperbidang terbaik untuk membagi data ke dalam kelas-kelas yang berbeda adalah cara kerja SVM. Garis dalam dua dimensi atau bidang dalam dimensi yang lebih tinggi yang

memisahkan titik data dari dua kelas disebut hiperbidang. Margin, atau jarak antara hiperbidang dan titik data terdekat dari setiap kelas—dikenal sebagai vektor pendukung—adalah apa yang ingin dimaksimalkan oleh SVM. Dalam dunia pembelajaran mesin, SVM saat ini menjadi isu hangat, yang menghasilkan kegembiraan sebanyak yang pernah dilakukan oleh jaringan saraf buatan. Dengan representasi model yang mudah dipahami, SVM adalah metode yang ampuh untuk klasifikasi generik (nonlinier), regresi, dan deteksi outlier[18].

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0, \quad i = 1, 2, 3, \ldots, n$$

Gambar 3. Persamaan Support Vector Machine

Pada pengenalan pola, fungsi diskriminatif linier pada ruang n-dimensi adalah g(x) = $\omega.x$ + b, persamaan hyperplane klasifikasi dapat dituliskan g(x) = ($\omega.x$) + b = 0. Diasumsikan bahwa kelas -1 dan +1 (kelas 2) dapat dipisahkan secara sempurna oleh hyperplane. Untuk kelas -1 sebagai sampel negatif, diformulasikan sebagai pola yang memenuhi pertidaksamaan $\omega.x$ + b ≤ -1, sedangkan untuk kelas +1 memenuhi $\omega.x$ + b ≥ 1. Margin terbesar ditentukan dengan memaksimalkan jarak antara hyperplane dan titik terdekat, yaitu 1/‖w‖. Hal ini diformulasikan sebagai masalah Quadratic Programming (QP), yaitu mencari margin maksimum dengan mempertimbangkan persamaan di atas[17]. SVM sebagai metode klasifikasi sangat cocok karena mampu menangani data yang tidak linear dan multi-kelas, serta efektif dalam memisahkan kategori polusi udara yang memiliki overlap fitur cukup tinggi.

2.5 Evaluasi Model

Efisiensi model Segmented Multiple Linear Regression dievaluasi oleh peneliti menggunakan indikator penilaian kinerja seperti Mean Absolute Percentage Error (MAPE) dan Root Mean Square Error (RMSE)[19]. Mean Absolute Error (MAE) dan Root Mean Squared Error (RMSE) adalah dua metrik yang sering digunakan dalam evaluasi model. Sampel observasi dan prediksi model yang cocok digunakan untuk menghitung MAE dan RMSE. Akar kuadrat dari mean squared error menghasilkan root mean square error, atau RMSE. Akar kuadrat tidak mengubah peringkat relatif model, tetapi menghasilkan ukuran yang direpresentasikan dalam satuan yang sama dengan y. Jarak Manhattan dan norma Euclidean, yang masing-masing merupakan norma L2 dan L1, adalah varian rata-rata dari MSE dan MAE. Keputusan antara RMSE dan MAE memiliki solusi rasional menurut teori probabilitas. Gagasan bahwa MAE hanya relevan untuk kesalahan yang didistribusikan secara seragam adalah salah, meskipun RMSE bekerja dengan baik untuk kesalahan yang didistribusikan secara normal. Meskipun MAE lebih andal, ada pilihan yang lebih baik[20]. Mereka mengukur tingkat kesalahan prediksi model dengan cara yang berbeda dan memiliki kegunaan khusus dalam analisis regresi.

3. HASIL DAN PEMBAHASAN

Pada bagian ini akan dijelaskan hasil temuan yang diperoleh dari analisis data dan interpretasi hasil yang dihasilkan oleh model MLR dan SVM. Proses ini bertujuan untuk mengevaluasi kinerja model dalam memprediksi kategori kualitas udara berdasarkan variabel polutan udara yang telah diolah sebelumnya. Setiap tahapan, mulai dari pembersihan data hingga evaluasi model, akan dijelaskan secara rinci untuk memberikan gambaran yang jelas.

3.1. Pengumpulan Data

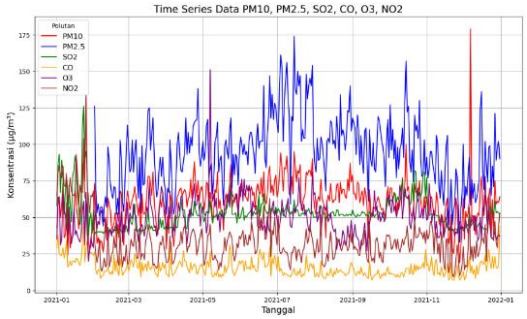
	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	categori	location
0	1/1/2021	43	NaN	58	29	35	65	65	O3	SEDANG	DKI2
1	1/2/2021	58	NaN	86	38	64	80	86	PM25	SEDANG	DKI3
2	1/3/2021	64	NaN	93	25	62	86	93	PM25	SEDANG	DKI3
3	1/4/2021	50	NaN	67	24	31	77	77	O3	SEDANG	DKI2
4	1/5/2021	59	NaN	89	24	35	77	89	PM25	SEDANG	DKI3
...	...	...	...	...	...	...	...	...	...	...	...
360	12/27/2021	75	121.0	61	23	40	47	121	PM25	TIDAK SEHAT	DKI4
361	12/28/2021	59	89.0	53	16	34	33	89	PM25	SEDANG	DKI4
362	12/29/2021	61	98.0	54	15	37	29	98	PM25	SEDANG	DKI4
363	12/30/2021	60	102.0	53	17	38	44	102	PM25	TIDAK SEHAT	DKI4
364	12/31/2021	64	90.0	52	44	37	53	90	PM25	SEDANG	DKI4

Gambar 4. Dataset Kualitas Udara DKI Jakarta Tahun 2021

Data yang digunakan diperoleh dari dataset publik yang tersedia di platform Kaggle. Dataset ini berisi data Indeks Standar Pencemaran Udara (ISPU) yang diukur dari lima Stasiun Pemantauan Kualitas Udara (SPKU) yang tersebar di Provinsi DKI

Jakarta sepanjang tahun 2021, dengan total data sebanyak 365 baris. Dataset tersebut terdiri dari 11 fitur utama yang mewakili berbagai parameter, yaitu tanggal, pm10, pm25, so2, co, o3, no2, max, critical, category, dan location.
URL: https://www.kaggle.com/datasets/derryderajat/indeks-pencemaran-udara-dki?resource=download

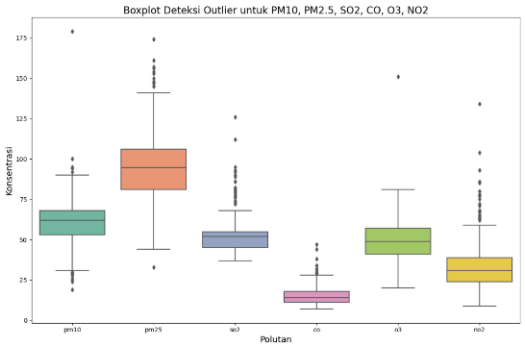
3.2. Data Deret Waktu Kualitas Udara DKI Jakarta 2021



Gambar 5. Rangkaian Waktu Dataset

Gambar di atas merupakan visualisasi data time series yang menunjukkan konsentrasi enam polutan utama, yaitu PM10, PM2.5, NO2, CO, SO2, dan O3 yang diukur dari lima Stasiun Pemantauan Kualitas Udara (SPKU) di DKI Jakarta sepanjang tahun 2021. Grafik ini menunjukkan perubahan konsentrasi polutan harian sepanjang periode pengamatan. PM10 dan PM2.5 merupakan polutan partikulat yang menunjukkan fluktuasi yang cukup signifikan sepanjang tahun. PM2.5 memiliki nilai yang lebih tinggi dibandingkan PM10. Peningkatan yang signifikan terlihat pada pertengahan hingga akhir tahun pada parameter PM10.

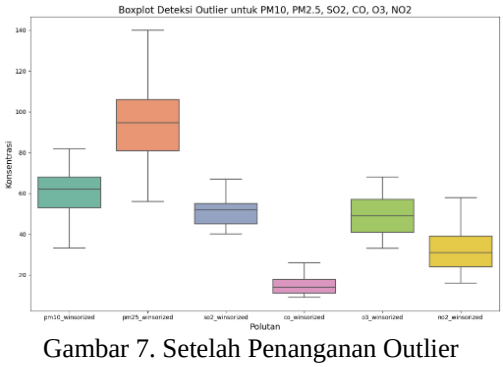
3.3 Deteksi Outlier untuk Setiap Variabel



Gambar 6. Outlier Pada Dataset

Boxplot ini memberikan gambaran umum tentang distribusi data dan membantu mengidentifikasi nilai outlier pada setiap variabel. Semua parameter memiliki nilai outlier yang ditunjukkan oleh data yang berada di luar batas atas dan bawah. Semua nilai outlier pada setiap variabel perlu ditangani agar data terdistribusi secara normal.

3.4 Penanganan Data Outlier

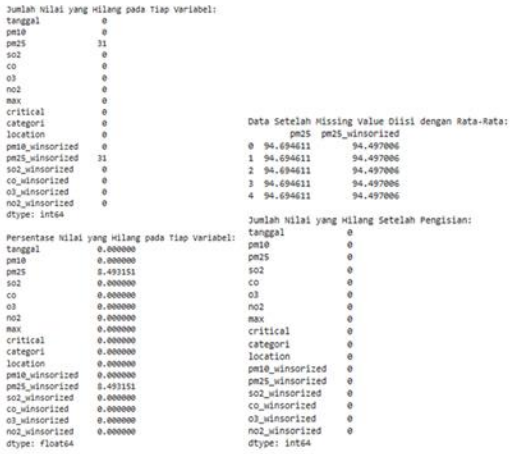


Gambar 7. Setelah Penanganan Outlier

Winsorizing adalah teknik yang menghasilkan estimator lokasi dan variabilitas yang lebih andal dengan mengurangi dampak outlier pada mean dan varians. Winsorizing menggunakan nilai tertinggi berikutnya di ekor bawah dan terendah di ekor atas untuk menetapkan kembali nilai ke proporsi kasus di kedua ekor distribusi[29]. Untuk menangani outlier, persentil ke-5 digunakan untuk menangani nilai yang lebih rendah dari batas bawah boxplot dan persentil ke-95 untuk nilai yang lebih tinggi

dari batas atas boxplot. Hasilnya menunjukkan bahwa tidak ada lagi nilai outlier yang terdeteksi dalam data. Winsorization dipilih karena lebih stabil daripada trimming yang menghapus data secara langsung. Metode ini mempertahankan jumlah total observasi sambil mengurangi efek ekstrem yang bisa mengganggu model linier dan kernel-based.

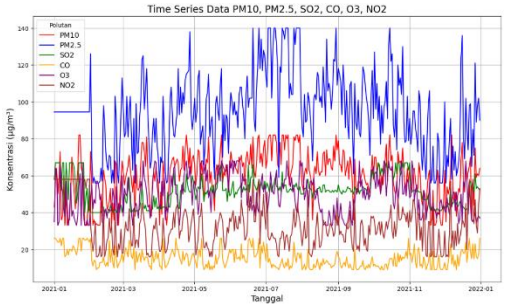
3.5 Imputasi Nilai pada Data Kosong



Gambar 8. Cek dan Input Data Kosong

Dari hasil pemeriksaan dapat diketahui bahwa variabel PM2.5 memiliki 31 nilai kosong. Variabel lainnya seperti pm10, so2, co, o3, no2, max, critical, category, dan location tidak memiliki nilai kosong. Persentase nilai kosong untuk variabel PM2.5 adalah 8,49%. Nilai rata-rata dihitung berdasarkan data yang ada, dan nilai-nilai kosong diisi dengan nilai rata-rata tersebut. Setelah nilai-nilai kosong diisi dengan nilai rata-rata, langkah selanjutnya adalah melakukan verifikasi apakah semua nilai kosong telah terisi. Pemilihan imputasi dengan rata-rata dilakukan karena nilai yang hilang hanya terdapat pada satu variabel (PM2.5) dengan persentase yang masih tergolong rendah (<10%). Selain itu, rata-rata cocok untuk distribusi yang mendekati normal dan menghindari penghapusan data yang bisa mengurangi representativitas sampel

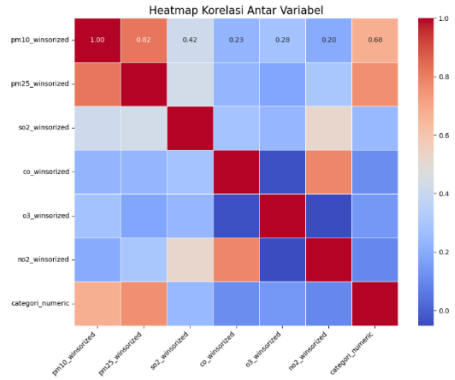
3.6 Data Deret Waktu Setelah Pembersihan Data



Gambar 9. Grafik Deret Waktu Setelah Pembersihan

Grafik ini memiliki hasil pembersihan data dengan mengganti nilai outlier dengan metode winsorization dan memasukkan nilai kosong ke dalam rata-rata. Grafik ini menunjukkan fluktuasi yang lebih stabil dan berdistribusi normal. Misalnya, nilai PM10 sekitar bulan November 2021, terjadi fluktuasi yang tidak biasa dan berbeda dengan data lainnya, yaitu mencapai lebih dari 180 µg/m³. Setelah diproses, grafik ini menunjukkan distribusi yang lebih baik, tanpa lonjakan nilai yang ekstrem. Nilai outlier dan data kosong pada tiap variabel juga sudah tergantikan dengan nilai pada metode yang dipilih.

3.7 Nilai Korelasi Setiap Variabel



Gambar 10. Nilai Korelasi Parameter

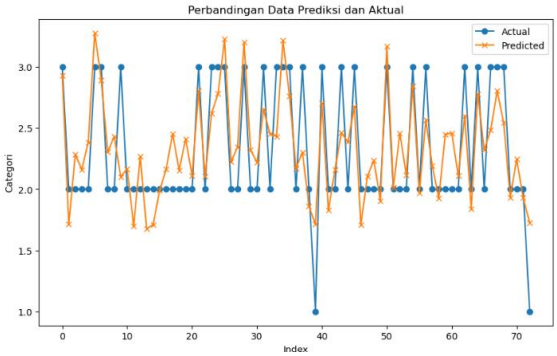
Matriks korelasi yang ditunjukkan di bawah ini menunjukkan tingkat korelasi antar masing-masing variabel yang telah melalui proses pembersihan data. Kategori kualitas udara sangat dipengaruhi oleh variabel PM2.5 dengan nilai sebesar 0,763673 dan diikuti oleh PM10 dengan nilai sebesar 0,678556. NO2 dan Kategori Kualitas Udara memiliki korelasi yang sangat lemah dengan nilai sebesar 0,093402 yang berarti kedua variabel ini hampir tidak memiliki hubungan. PM10 dan PM2.5 memiliki korelasi yang sangat tinggi dengan nilai sebesar 0,8195 yang menunjukkan bahwa kedua variabel ini sering menunjukkan pola perubahan yang serupa.

3.8 Pemisahan Data Training dan Data Testing

Proses pemisahan data dilakukan untuk menyiapkan data yang akan digunakan dalam model training dan testing. Variabel X berisi kumpulan fitur pm10, pm25, so2, co, o3, dan no2 yang akan dijadikan input dalam model. Sedangkan variabel y merupakan variabel target yang menunjukkan kategori kualitas udara yang telah dikonversi menjadi nilai numerik. Sebanyak 20% dari total data akan digunakan sebagai data testing, sedangkan 80% sisanya akan digunakan untuk training. Proporsi 80:20 dipilih karena merupakan praktik umum dalam pembelajaran mesin yang menawarkan trade-off optimal antara cukupnya data latih dan validasi kinerja model pada data uji yang representatif [27].

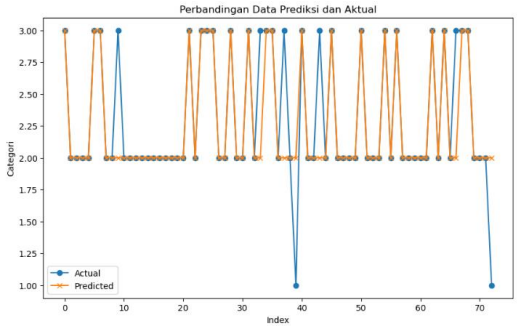
3.9 Penerapan Regresi Linear Berganda

Regresi Linier Berganda untuk memodelkan hubungan antara variabel polutan udara dengan kategori kualitas udara. Setelah model dilatih menggunakan data latih, diperoleh hasil berupa koefisien intersep dan slope untuk masing-masing variabel. Nilai Intersep yang diperoleh sebesar 0,8916751857397582. Koefisien Slope masing-masing variabel adalah PM10 (0,00578014), PM2,5 (0,01633731), SO2 (-0,00168198), CO (0,00171387), O3 (-0,00276095), dan NO2 (-0,00638805). Berdasarkan hasil koefisien yang diperoleh, diperoleh persamaan regresi linier berganda, yang memungkinkan untuk memprediksi kategori kualitas udara (nilai y) berdasarkan nilai variabel x. Koefisien tertinggi dimiliki oleh PM2.5 (0.0163), menunjukkan bahwa peningkatan 1 satuan PM2.5 cenderung menaikkan skor kategori ISPU secara signifikan. Sebaliknya, nilai negatif pada NO2 (-0.00638) menunjukkan hubungan terbalik, meskipun pengaruhnya relatif kecil.



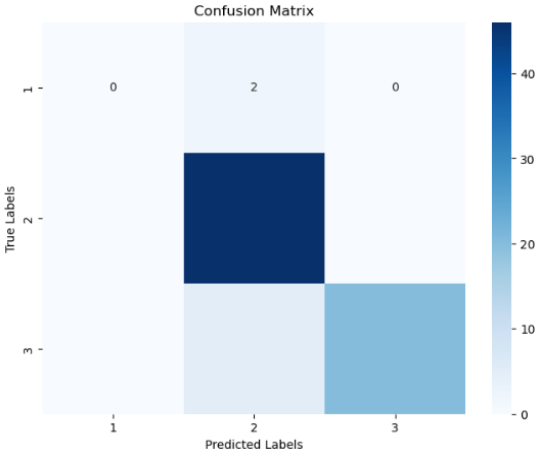
Gambar 11. Perbandingan Data Prediksi dan Aktual Regresi Linear Berganda

Setelah model regresi linier berganda dilatih, langkah selanjutnya adalah membandingkan nilai aktual dan nilai prediksi pada data uji. Sebagai sampel, pada indeks data ke-193, data aktual bernilai 3, sedangkan data prediksi bernilai 2,929882. Nilai Aktual merupakan kategori kualitas udara aktual (Baik = 1, Sedang = 2, Tidak Sehat = 3). Nilai Prediksi merupakan hasil prediksi model regresi linier berganda yang menunjukkan kategori kualitas udara prediksi berdasarkan data polutan udara.



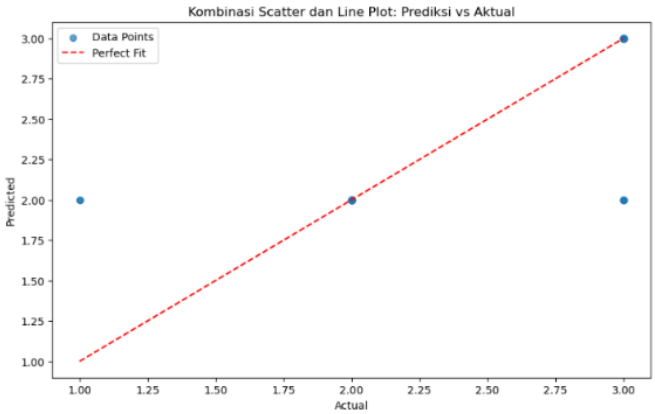
Gambar 12. Perbandingan Data Prediksi dan Aktual Regresi Linear Berganda Setelah Konversi

Namun, nilai prediksi yang dihasilkan bukanlah bilangan bulat karena model regresi linier menghasilkan nilai kontinu. Untuk menyesuaikan hasil prediksi ini agar sesuai dengan kategori kualitas udara yang ditentukan (1, 2, atau 3), diperlukan proses penskalaan. Untuk mengelompokkan hasil prediksi ke dalam kategori yang sesuai, aturan yang digunakan adalah Nilai kurang dari 1,5 dikategorikan sebagai 1 (Baik); Nilai antara 1,5 dan 2,5 dikategorikan sebagai 2 (Sedang); dan Nilai di atas 2,5 dikategorikan sebagai 3 (Tidak Sehat). Setelah proses penskalaan, hasil prediksi model menjadi lebih jelas dan dapat dikelompokkan ke dalam kategori kualitas udara yang sesuai.



Gambar 13. Confusion Matrix Regresi Linear Berganda

Confusion Matrix ini menunjukkan kinerja model Regresi Linier Berganda (MLR) dalam mengklasifikasikan kualitas udara ke dalam tiga kategori: Baik, Sedang, dan Tidak Sehat. Model tersebut memprediksi dengan tepat 46 sampel “Sedang” dan 20 sampel “Tidak Sehat”. Akan tetapi, tidak ada sampel “Baik” yang diprediksi dengan tepat, dengan 2 sampel “Baik” yang salah diklasifikasikan sebagai “Sedang”. Selain itu, 5 sampel “Tidak Sehat” salah diprediksi sebagai “Sedang”. Hasil ini menunjukkan bahwa model tersebut sangat baik untuk kategori “Sedang” tetapi perlu ditingkatkan dalam mendeteksi kategori “Baik” dan mengurangi kesalahan dalam kategori “Tidak Sehat”.

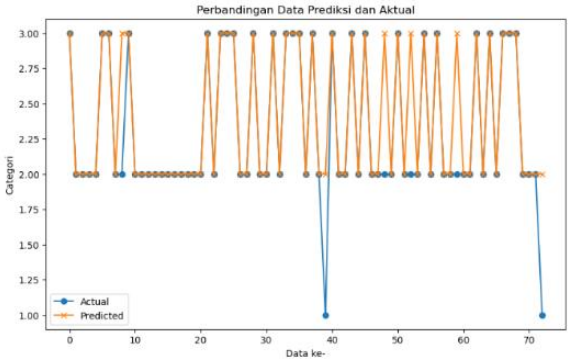


Gambar 14. Kombinasi Scatter dan Line Plot

Gambar di atas menunjukkan kombinasi diagram sebar dan diagram garis yang membandingkan nilai prediksi dengan nilai aktual dari model MLR. Titik Biru Mewakili pasangan data aktual dan prediksi dan Garis Putus-putus Merah secara diagonal mewakili kondisi ideal di mana prediksi model sama persis dengan nilai aktual. Beberapa titik biru berada di garis merah, yang menunjukkan bahwa model mampu memberikan prediksi yang hampir akurat. Namun, ada beberapa nilai prediksi yang tidak sesuai dengan nilai aktual dalam data uji.

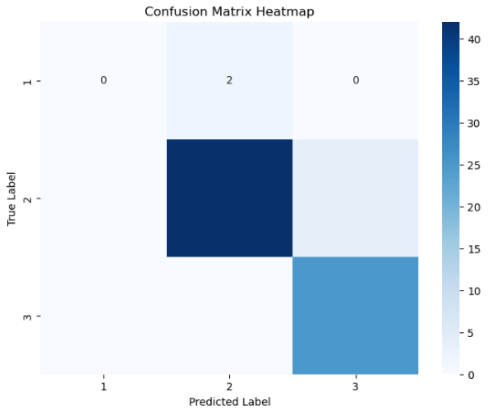
3.10 Penerapan Support Vector Machine

Implementasi model Support Vector Machine (SVM) diawali dengan penggunaan fungsi Standard Scaler untuk melakukan standarisasi data. Langkah ini memastikan setiap fitur memiliki skala yang sama, sehingga SVM dapat bekerja secara optimal. Selanjutnya, Kernel RBF dipilih karena mampu menangani hubungan non-linier antar fitur, dan nilai C=100 dipilih untuk meminimalkan error pada data latih, berdasarkan eksperimen grid search sederhana yang menunjukkan kombinasi ini memberikan akurasi tertinggi tanpa overfitting signifikan. Parameter C yang besar bertujuan untuk mengurangi kesalahan klasifikasi pada data training dengan memaksimalkan margin keputusan. Hasil aplikasi SVM ini nantinya akan dibandingkan dengan model Multiple Linear Regression (MLR) untuk mengevaluasi akurasi prediksi dan kemampuan masing-masing model dalam mengklasifikasi kategori kualitas udara. Data diacak dengan menggunakan nilai *random state* yang sama dengan yang diterapkan pada MLR untuk mengukur performa.



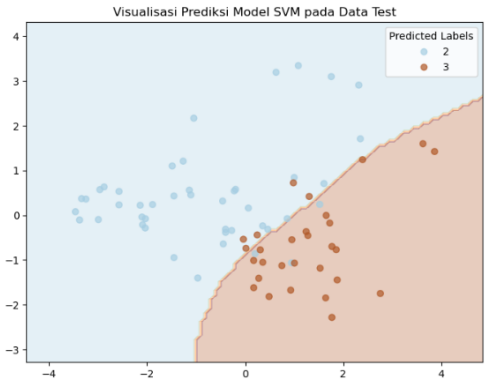
Gambar 15. Perbandingan Data Prediksi dan Aktual Support Vector Machine

Grafik di atas menunjukkan perbandingan antara nilai aktual dan nilai prediksi yang dihasilkan oleh model SVM untuk 73 data uji. Mayoritas prediksi sesuai dengan nilai aktual, yang menunjukkan kinerja model yang baik. Namun, terdapat 6 kesalahan klasifikasi, di mana beberapa data yang seharusnya diklasifikasikan sebagai Baik (1) diprediksi sebagai Sedang (2) dan data yang seharusnya Sedang (2) diprediksi sebagai Tidak Sehat (3). Namun, kesalahan ini relatif kecil, yang menunjukkan bahwa model SVM memiliki akurasi tinggi dalam memprediksi kategori dengan tingkat generalisasi yang baik.



Gambar 16. Confusion Matrix Support Vector Machine

Confusion Matrix menunjukkan bahwa model Support Vector Machine (SVM) tidak mampu memprediksi kategori Baik (1) dengan tepat, karena semua data aktual pada kategori ini diprediksi secara tidak tepat sebagai Sedang (2). Sebanyak 42 dari 46 data aktual pada kategori Sedang (2) diprediksi secara tepat, sedangkan 4 data diprediksi secara tidak tepat sebagai Tidak Sehat (3). Untuk kategori Tidak Sehat (3), model SVM mampu memprediksi dengan sempurna, yaitu 25 data sesuai dengan nilai aktualnya. Secara keseluruhan, model ini akurat dalam memprediksi kategori Tidak Sehat, tetapi memiliki kekurangan dalam membedakan kategori Baik dan Sedang.



Gambar 17. Pemisahan Data Menggunakan PCA

Gambar tersebut menunjukkan visualisasi prediksi model SVM pada data uji yang telah direduksi dimensinya menggunakan PCA. Principal Component Analysis (PCA) adalah metode analisis data yang digunakan untuk mengurangi dimensionalitas suatu dataset dengan mengubah variabel asli menjadi variabel baru yang disebut komponen utama (PC)[30]. PCA digunakan di sini untuk memfasilitasi visualisasi dan analisis hanya menggunakan dua komponen utama yang memuat sebagian besar informasi dari data asli. Area biru dan coklat mewakili batas keputusan antara kategori Moderat (2) dan Tidak Sehat (3). Model SVM berhasil memisahkan kedua kategori tersebut dengan baik, meskipun beberapa data berada di dekat batas keputusan, yang menunjukkan potensi kesalahan prediksi di area antara.

3.11 Evaluasi Model

Pada bagian Evaluasi Model, perbandingan kinerja dilakukan antara dua model, yaitu Multiple Linear Regression (MLR) dan Support Vector Machine (SVM), dengan mengukur empat metrik evaluasi yang berbeda: Akurasi, R-Squared, Root Mean Squared Error (RMSE), dan Mean Absolute Error (MAE).

Tabel 2. Evaluasi Model

Metrik	MLR	SVM
Akurasi	90.41%	91.78%
R-Squared	64.56%	69.63%
RMSE	0.3097	0.2867
MAE	0.0959	0.0822

Berdasarkan hasil Evaluasi Model yang dilakukan pada kedua model, dapat dilihat perbandingan yang menunjukkan kinerja yang sedikit lebih unggul pada model SVM. Model MLR memiliki Akurasi sebesar 90,41% dengan R-kuadrat sebesar 64,56%, yang menunjukkan bahwa model ini mampu menjelaskan sekitar 64,56% variabilitas data. Akan tetapi, model SVM berhasil mencapai Akurasi sebesar 91,78% yang sedikit lebih tinggi dibandingkan dengan MLR, dan memiliki R-kuadrat sebesar 69,63%, yang menunjukkan kemampuan model SVM yang lebih baik dalam menjelaskan variabilitas data. Selain itu, SVM juga memiliki nilai RMSE yang lebih rendah, yaitu sebesar 0,2867 dibandingkan dengan MLR yang memiliki nilai RMSE sebesar 0,3097, yang menunjukkan bahwa SVM memiliki galat prediksi yang lebih kecil. Begitu pula pada MAE, SVM memperoleh nilai sebesar 0,0822, lebih rendah dibandingkan dengan MLR yang memiliki nilai MAE sebesar 0,0959, yang berarti SVM memberikan prediksi yang lebih akurat dan stabil. Secara keseluruhan, meskipun kedua model memberikan hasil yang baik, SVM menunjukkan kinerja yang lebih unggul dalam hal akurasi dan nilai R-squared serta memiliki nilai error yang kecil.

4. KESIMPULAN

Penelitian ini berhasil menerapkan Machine Learning menggunakan algoritma Multiple Linear Regression (MLR) dan Support Vector Machine (SVM), untuk memprediksi kategori kualitas udara di Jakarta. Variabel X berisi kumpulan fitur pm10, pm25, so2, co, o3, dan no2 yang akan digunakan sebagai input dalam model. Variabel y merupakan variabel target yang menunjukkan kategori kualitas udara yang telah diubah menjadi nilai numerik. Hasil evaluasi menunjukkan bahwa kinerja model SVM mengungguli MLR, sebagaimana dibuktikan oleh nilai akurasi yang lebih tinggi (91,78% vs 90,41%), nilai R-kuadrat yang lebih tinggi (69,63% vs 64,56%), nilai RMSE yang lebih rendah (0,2867 vs 0,3097), dan nilai MAE yang lebih rendah (0,0822 vs 0,0959). SVM terbukti lebih efektif dalam menangani data non-linier dan menawarkan generalisasi yang lebih baik, menjadikannya alat yang lebih baik untuk prediksi kualitas udara. Namun, kedua model tersebut menunjukkan tantangan dalam memprediksi kategori kualitas udara “Baik” secara akurat, kemungkinan karena representasinya yang terbatas dalam kumpulan data, dalam hal ini disebut sebagai *imbalance data*. Hal ini menyoroti perlunya penyeimbangan data yang lebih baik atau rekayasa fitur tingkat lanjut untuk meningkatkan akurasi prediksi semua kategori.

UCAPAN TERIMA KASIH

Saya ingin mengucapkan terima kasih kepada dosen metode penelitian saya, atas saran, dorongan, dan dukungannya yang penting selama pengembangan studi ini. Pengetahuannya tentang metode penelitian telah meningkatkan pemahaman saya terhadap topik ini dan memungkinkan saya untuk berhasil menerapkan ide-ide ini dalam tugas ini. Umpan balik yang tulus dan rekomendasi terperinci sangat penting dalam menyempurnakan studi ini, dan saya sangat berterima kasih atas hal itu.

REFERENSI

[1] S. Halsana, "Air Quality Prediction Model using Supervised Machine Learning Algorithms," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 3307, pp. 190–201, 2020, doi: 10.32628/cseit206435.
[2] S. D. A. Kusumaningtyas, E. Aldrian, T. Wati, D. Atmoko, and Sunaryo, "The recent state of ambient air quality in Jakarta," *Aerosol Air Qual. Res.*, vol. 18, no. 9, pp. 2343–2354, 2018, doi: 10.4209/aaqr.2017.10.0391.
[3] D. A. Kristiyanti, E. Purwaningsih, E. Nurelasari, A. Al Kaafi, and A. H. Umam, "Implementation of Neural Network Method for Air Quality Forecasting in Jakarta Region," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012037.
[4] A. Maulana *et al.*, "Optimizing University Admissions: A Machine Learning Perspective," *J. Educ. Manag. Learn.*, vol. 1, no. 1, pp. 1–7, 2023, doi: 10.60084/jeml.v1i1.46.
[5] B. Mahesh, "Machine Learning Algorithms - A Review," *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 381–386, 2020, doi: 10.21275/art20203995.
[6] G. M. Idroes *et al.*, "Urban Air Quality Classification Using Machine Learning Approach to Enhance Environmental Monitoring," *Leuser J. Environ. Stud.*, vol. 1, no. 2, pp. 62–68, 2023, doi: 10.60084/ljes.v1i2.99.
[7] S. Rath, A. Tripathy, and A. R. Tripathy, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 5, pp. 1467–1474, 2020, doi: 10.1016/j.dsx.2020.07.045.
[8] S. Fashoto, E. Mbunge, G. Ogunleye, and J. Van den Burg, "Implementation of Machine Learning for Predicting Maize Crop Yields Using Multiple Linear Regression and Backward Elimination," *Malaysian J. Comput.*, vol. 6, no. 1, p. 679, 2021, doi: 10.24191/mjoc.v6i1.8822.
[9] C. Cortes and V. Vapnik, "Support-Vector Networks," *Kluwer Acad. Publ.*, vol. 20, no. 2, pp. 273–297, 1995, doi: 10.1111/j.1747-0285.2009.00840.x.
[10] W. Lu *et al.*, "Air pollutant parameter forecasting using support vector machines," *Proc. Int. Jt. Conf. Neural Networks*, vol. 1, no. February, pp. 630–635, 2002, doi: 10.1109/ijcnn.2002.1005545.
[11] D. Iskandaryan, F. Ramos, and S. Trilles, "Air quality prediction in smart cities using machine learning technologies based on sensor data: A review," *Appl. Sci.*, vol. 10, no. 7, 2020, doi: 10.3390/app10072401.
[12] H. Liu, Q. Li, D. Yu, and Y. Gu, "Air quality index and air pollutant concentration prediction based on machine learning algorithms," *Appl. Sci.*, vol. 9, no. 19, 2019, doi: 10.3390/app9194069.
[13] Y. C. Liang, Y. Maimury, A. H. L. Chen, and J. R. C. Juarez, "Machine learning-based prediction of air quality," *Appl. Sci.*, vol. 10, no. 24, pp. 1–17, 2020, doi: 10.3390/app10249151.
[14] D. Adryanti Felicia Sampe *et al.*, "Pilot study of air quality index assessment of nitrogen pollutant using lichen as bioindicators in Jakarta and Depok, Indonesia," *E3S Web Conf.*, vol. 211, pp. 1–13, 2020, doi: 10.1051/e3sconf/202021102014.
[15] Menteri Lingkungan Hidup dan Kehutanan Republik Indonesia, "Peraturan Menteri Lingkungan Hidup dan Kehutanan Republik Indonesia Nomor P.14/MENLHK/SETJEN/KUM.1/7/2020 Tentang Indeks Standar Pencemaran Udara," pp. 1–16, 2020, [Online]. Available: https://peraturan.bpk.go.id/Details/163466/permen-lhk-no-14-tahun-2020
[16] M. Tranmer, J. Murphy, M. Elliot, and M. Pampaka, "Multiple Linear Regression (2nd Edition)," *Cathie Marsh Cent. Census Surv. Res.*, vol. 5, no. 5, pp. 1–5, 2020.
[17] N. H. Ovirianti, M. Zarlis, and H. Mawengkang, "Support Vector Machine Using A Classification Algorithm," *SinkrOn*, vol. 7, no. 3, pp. 2103–2107, 2022, doi: 10.33395/sinkron.v7i3.11597.
[18] D. Meyer, "Support Vector Machines," *R-News*, vol. 1, pp. 3–9, 2009.
[19] M. Zidan and A. Kamil, "Forecasting the Air Quality Index (AQI) in Jakarta, Indonesia by Using a Linear Regression Model," *ResearchGate*, no. September, 2024, doi: 10.13140/RG.2.2.34971.89122.
[20] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, 2022, doi: 10.5194/gmd-15-5481-2022.
[21] R. Heiss, "Data, Types of," *Int. Encycl. Commun. Res. Methods*, pp. 1–6, 2017, doi: 10.1002/9781118901731.iecrm0062.
[22] C. Andrade, "A Student's Guide to the Classification and Operationalization of Variables in the Conceptualization and Design of a Clinical Study: Part 1," *Indian J. Psychol. Med.*, vol. 43, no. 2, pp. 177–179, 2021, doi: 10.1177/0253717621994334.
[23] H. Müller and J. Freytag, "Problems, Methods, and Challenges in Comprehensive Data Cleansing," *Informatics reports // Inst. Comput. Sci. Humboldt Univ. Berlin*, no. HUB-IB-164, Humboldt University Berlin, pp. 1–23, 2003, [Online]. Available: http://www.dbis.informatik.hu-berlin.de/fileadmin/research/papers/techreports/2003-hub_ib_164-mueller.pdf
[24] F. Ridzuan and W. M. N. Wan Zainon, "A review on data cleansing methods for big data," *Procedia Comput. Sci.*, vol. 161, pp. 731–738, 2019, doi: 10.1016/j.procs.2019.11.177.
[25] E. Hartini, "Classification of Missing Values Handling Method During Data Mining: Review," *Sigma Epsil.*, vol. 21, no. 2, pp. 49–60, 2017.
[26] A. ur Rehman and S. B. Belhaouari, "Unsupervised outlier detection in multidimensional data," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00469-z.
[27] A. Nurhopipah and U. Hasanah, "Dataset Splitting Techniques Comparison For Face Classification on CCTV Images," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 14, no. 4, p. 341, 2020, doi: 10.22146/ijccs.58092.
[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, doi: 10.1289/EHP4713.
[29] B. E. Blaine, "Winsorizing," *Fish. Digit. Publ.*, pp. 1817–1818, 2018.
[30] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, 2016, doi: 10.1098/rsta.2015.0202.