

PREDIKSI INDEKS KUALITAS UDARA OZON (ISPU O₃) MENGGUNAKAN RANDOM FOREST DAN REGRESI VEKTOR DUKUNGAN: STUDI PERBANDINGAN PADA DATA TAHUN 2022

Syaeful Bachri Barelang, Ase Ananda Putra, Jundana Abdullah Hanif,
Muhammad Rayyan Ardiano, Muhammad Athallah Izma, Vitri Tundjungsari

Teknik Informatika, Universitas Esa Unggul

Jl. Harapan Indah Boulevard No.2, Pusaka Rakyat, Kec. Tarumajaya, Kabupaten Bekasi, Jawa Barat

bachri.barelang@student.esaunggul.ac.id

ABSTRAK

Penurunan kualitas udara merupakan permasalahan lingkungan yang semakin serius, khususnya di wilayah perkotaan dengan aktivitas industri dan transportasi yang tinggi. Ozon (O₃) sebagai salah satu parameter utama dalam Indeks Standar Pencemar Udara (ISPU) memiliki dampak signifikan terhadap kesehatan manusia dan lingkungan, sehingga diperlukan metode prediksi yang akurat untuk mendukung sistem peringatan dini. Permasalahan utama dalam prediksi O₃ terletak pada kompleksitas hubungan *nonlinier* antar polutan yang sulit dimodelkan menggunakan pendekatan konvensional. Oleh karena itu, penelitian ini bertujuan untuk memprediksi nilai ISPU O₃ serta membandingkan kinerja metode *supervised learning Random Forest* dan *Support Vector Regression* (SVR) menggunakan data kualitas udara tahun 2022. Metode penelitian meliputi tahap praproses data, pemilihan fitur PM₁₀, SO₂, CO, dan NO₂, pelatihan model, serta evaluasi menggunakan metrik MAE, RMSE, MSE, dan koefisien determinasi (R²). Hasil penelitian menunjukkan bahwa model Random Forest memberikan performa terbaik dengan nilai R² sebesar 0,49, sedangkan model pembanding menghasilkan nilai R² sebesar 0,38. Temuan ini menunjukkan bahwa pendekatan *supervised learning* efektif untuk prediksi ISPU O₃, meskipun masih diperlukan penambahan variabel meteorologi untuk meningkatkan akurasi model.

Kata kunci : Kualitas udara, Ozon (O₃), ISPU, Random Forest, Support Vector Regression, Supervised Learning

1. PENDAHULUAN

Kualitas udara merupakan salah satu indikator penting dalam menentukan tingkat kesehatan lingkungan dan risiko terhadap kesehatan manusia, terutama di wilayah perkotaan dengan tingkat aktivitas industri dan transportasi yang tinggi. Salah satu parameter utama dalam pengukuran kualitas udara di Indonesia adalah Indeks Standar Pencemar Udara (ISPU), di mana Ozon (O₃) menjadi salah satu polutan kritis yang berpengaruh terhadap kesehatan pernapasan dan lingkungan [1]. Prediksi konsentrasi O₃ menjadi semakin penting guna mendukung sistem peringatan dini dan pengambilan kebijakan mitigasi polusi udara.

Seiring berkembangnya teknologi dan ketersediaan data lingkungan, pendekatan berbasis *machine learning* semakin banyak digunakan dalam pemodelan dan prediksi kualitas udara karena mampu menangkap pola nonlinier dan kompleks yang sulit dianalisis menggunakan pendekatan statistik konvensional [2], [3]. Di antara berbagai metode *supervised learning*, Random Forest dan Support Vector Regression (SVR) merupakan dua algoritma yang terbukti efektif dalam memodelkan hubungan antara variabel meteorologi dan konsentrasi polutan udara [1], [4], [5]. Beberapa penelitian menunjukkan bahwa kedua metode tersebut mampu memberikan performa prediksi yang baik dalam konteks pemodelan O₃ dan indeks kualitas udara di berbagai wilayah [1], [6].

Random Forest dikenal memiliki kemampuan generalisasi yang baik, ketahanan terhadap *overfitting*,

serta mampu menangani data berdimensi tinggi dan kompleks [3], [6]. Sementara itu, SVR menawarkan keunggulan dalam pemodelan regresi berbasis margin yang mampu memberikan estimasi yang stabil pada data nonlinier dengan dukungan fungsi kernel [2], [3], [4], [7]. Sejumlah studi sebelumnya telah melakukan analisis komparatif antara metode *machine learning* untuk prediksi kualitas udara, termasuk prediksi O₃, dan menunjukkan adanya variasi performa antara model tergantung karakteristik data dan parameter model yang digunakan [1], [8], [9].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk memprediksi nilai ISPU O₃ menggunakan metode Random Forest dan Support Vector Regression serta melakukan analisis komparasi kinerja kedua model berdasarkan data kualitas udara tahun 2022. Hasil penelitian diharapkan dapat memberikan kontribusi terhadap pengembangan model prediksi kualitas udara berbasis *machine learning* yang lebih akurat serta mendukung pengambilan keputusan dalam pengelolaan lingkungan dan kesehatan masyarakat.

2. TINJAUAN PUSTAKA

Penelitian mengenai prediksi kualitas udara berbasis *machine learning* menunjukkan bahwa metode Random Forest dan Support Vector Regression (SVR) memiliki performa yang efektif dalam memodelkan konsentrasi Ozon (O₃) dan indeks kualitas udara [1], [4]. Random Forest dikenal stabil, akurat, serta mampu menangani data kompleks dan berdimensi tinggi [3], [9], sedangkan SVR unggul

dalam memodelkan hubungan nonlinier melalui fungsi kernel [2], [4]. Sejumlah studi komparatif menyatakan bahwa kedua metode tersebut sering menunjukkan performa prediksi yang baik pada berbagai kondisi data [5], [7], [9]. Oleh karena itu, penelitian lebih lanjut masih diperlukan untuk mengevaluasi perbandingan kinerjanya pada dataset spesifik.

Berdasarkan tinjauan pustaka tersebut, terlihat bahwa masih diperlukan penelitian lanjutan yang membandingkan performa Random Forest dan SVR dalam konteks data spesifik, termasuk data kualitas udara tahunan dengan fokus pada parameter O_3 . Oleh karena itu, penelitian ini berupaya melanjutkan kontribusi ilmiah dengan melakukan analisis komparatif kedua metode tersebut pada data kualitas udara tahun 2022 untuk memperoleh gambaran kinerja model yang lebih komprehensif.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode *supervised learning* untuk memprediksi nilai Indeks Standar Pencemar Udara (ISPU) Ozon (O_3). Pendekatan *machine learning* dipilih karena kemampuannya dalam memodelkan hubungan nonlinier dan kompleks antar variabel pencemar udara yang sulit ditangkap oleh metode statistik konvensional [1]. Data yang digunakan dalam penelitian ini merupakan data kualitas udara tahun 2022 yang mencakup parameter PM_{10} , SO_2 , CO , NO_2 , serta nilai maksimum harian (MAX). Variabel O_3 ditetapkan sebagai variabel target, sedangkan parameter lainnya digunakan sebagai variabel prediktor. Secara umum, tahapan penelitian meliputi pengumpulan data, praproses data, pemodelan menggunakan *Random Forest* dan *Support Vector Regression* (SVR), serta evaluasi dan perbandingan kinerja model.

3.1. Pengumpulan dan Eksplorasi Data

Data kualitas udara diperoleh dari hasil pemantauan rutin selama tahun 2022. Tahap eksplorasi data dilakukan untuk memahami karakteristik data, distribusi nilai polutan, serta hubungan awal antar variabel. Pemeriksaan ini penting untuk memastikan kesiapan data sebelum dilakukan pemodelan [2], [3].

3.2. Praproses Data

Tahap praproses mencakup pembersihan data, penghapusan atribut yang tidak relevan, serta pemeriksaan nilai hilang dan anomali. Normalisasi data dilakukan untuk menyamakan skala antar fitur numerik agar proses pelatihan model menjadi lebih stabil. Selain itu, analisis korelasi digunakan untuk memastikan bahwa fitur yang digunakan memiliki relevansi terhadap variabel target O_3 [1].

3.3. Pemodelan Random Forest Regressor

Random Forest Regressor digunakan sebagai metode utama karena kemampuannya dalam menangani data berdimensi tinggi dan hubungan

nonlinier melalui pendekatan *ensemble learning* [4], [5]. Model ini membangun sejumlah pohon keputusan dengan teknik *bootstrap sampling* dan *random feature selection*, sehingga mampu mengurangi risiko *overfitting* serta meningkatkan stabilitas prediksi. Selain itu, Random Forest memungkinkan analisis *feature importance* yang memberikan wawasan mengenai kontribusi masing-masing variabel dalam memprediksi O_3 .

3.4. Pemodelan Support Vector Regression (SVR)

Support Vector Regression digunakan sebagai metode pembanding karena memiliki kemampuan yang baik dalam pemodelan regresi nonlinier berbasis margin [6], [7]. SVR memanfaatkan fungsi kernel untuk memetakan data ke ruang berdimensi lebih tinggi sehingga hubungan kompleks antar variabel dapat dimodelkan secara optimal. Metode ini banyak digunakan dalam penelitian prediksi kualitas udara karena stabilitasnya terhadap data dengan ukuran sampel terbatas [6].

3.5. Pelatihan dan Validasi Model

Dataset dibagi menjadi data latih dan data uji menggunakan skema *train-test split*. Untuk meningkatkan keandalan hasil, proses pelatihan juga divalidasi menggunakan *k-fold cross-validation*. Pendekatan ini bertujuan untuk mengurangi bias pemodelan dan memastikan bahwa model memiliki kemampuan generalisasi yang baik [3], [8].

3.6. Evaluasi Kinerja Model

Kinerja kedua model dievaluasi menggunakan beberapa metrik regresi, yaitu Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), dan koefisien determinasi (R^2). Metrik-metrik tersebut umum digunakan dalam penelitian prediksi kualitas udara untuk menilai tingkat akurasi dan kesalahan model [1], [4]. Hasil evaluasi selanjutnya digunakan untuk membandingkan performa Random Forest dan SVR dalam memprediksi nilai ISPU O_3 .

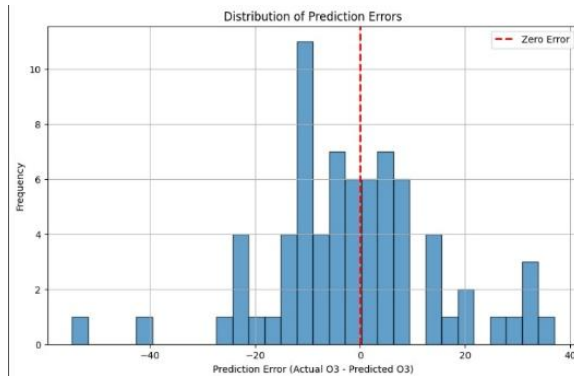
Melalui metodologi ini, penelitian diharapkan mampu memberikan gambaran yang komprehensif mengenai efektivitas metode *supervised learning* dalam prediksi kualitas udara, khususnya parameter O_3 , serta menjadi dasar pengembangan model prediksi yang lebih akurat di masa mendatang.

4. HASIL DAN PEMBAHASAN

4.1. Random Forest

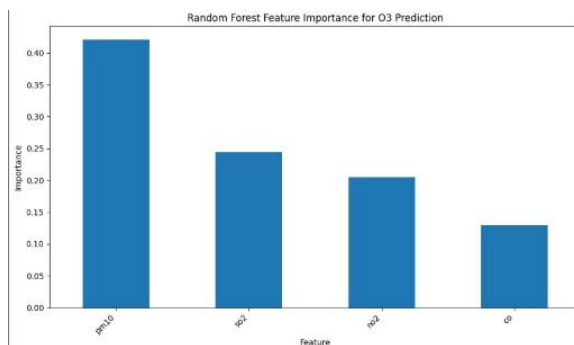
Hasil analisis menunjukkan bahwa model Random Forest Regressor memiliki kemampuan prediksi yang cukup baik terhadap nilai ISPU O_3 , ditunjukkan dengan nilai R^2 sebesar 0,49 yang berarti sekitar 49% variasi konsentrasi O_3 dapat dijelaskan oleh variabel input yang digunakan. Nilai MAE sebesar 12,15 mengindikasikan bahwa rata-rata selisih antara nilai prediksi dan nilai aktual berada pada kisaran 12 satuan, sedangkan nilai RMSE sebesar

16,15 yang lebih tinggi dari MAE menunjukkan adanya beberapa kesalahan prediksi yang relatif besar pada sebagian data.



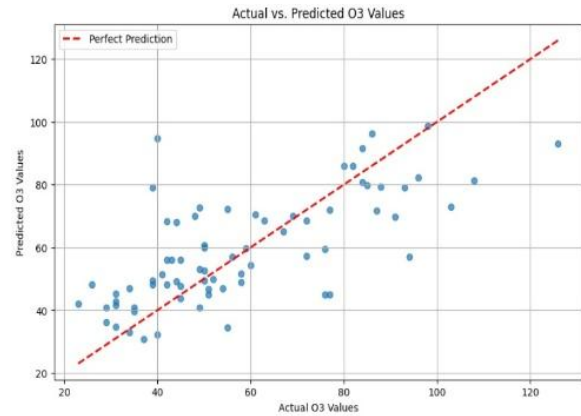
Gambar 1. Prediction Error

Berdasarkan analisis *feature importance*, PM10 menjadi prediktor paling dominan dengan kontribusi sekitar 42%, diikuti oleh SO₂ sekitar 24% dan NO₂ sekitar 20,5%. Sementara itu, CO memiliki pengaruh paling rendah, yaitu sekitar 13%. Temuan ini sejalan dengan pemahaman ilmiah bahwa partikulat dan gas pencemar lain memiliki keterkaitan dengan pembentukan dan dinamika O₃ di atmosfer.



Gambar 2. Feature Importance

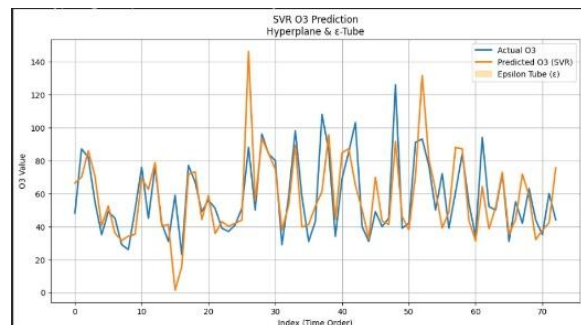
Visualisasi *actual vs predicted* menunjukkan korelasi positif di mana sebagian besar titik mendekati garis prediksi sempurna pada nilai O₃ rendah hingga menengah. Namun, penyebaran titik semakin melebar pada konsentrasi O₃ tinggi, menunjukkan kecenderungan model untuk mengestimasi lebih rendah pada kondisi puncak O₃. Hal ini juga diperkuat oleh distribusi *prediction error* yang sedikit condong ke kanan, menandakan lebih banyak kasus *underprediction*. Secara umum, model mampu memberikan prediksi yang cukup stabil, namun masih menunjukkan keterbatasan dalam memodelkan nilai O₃ yang tinggi dan variabilitas yang lebih kompleks.



Gambar 3. Actual vs Predicted

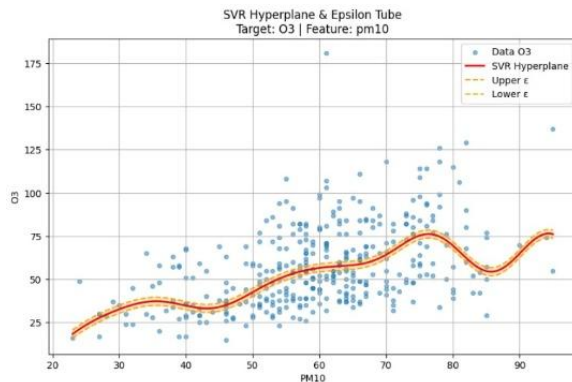
4.2. Support Vector Regression

Pada pengujian tahap kedua, pemodelan prediksi O₃ dilakukan menggunakan variabel fitur periode_data, PM10, SO₂, CO, NO₂, dan MAX sebagai prediktor. Data yang digunakan berasal dari tahun 2022 dengan variasi nilai polutan yang merepresentasikan kondisi kualitas udara pada beberapa periode waktu. Berdasarkan hasil evaluasi, model menghasilkan nilai Mean Squared Error (MSE) sebesar 317,43 dan nilai koefisien determinasi (R²) sebesar 0,3845. Nilai R² tersebut menunjukkan bahwa sekitar 38,45% variasi konsentrasi O₃ berhasil dijelaskan oleh model, sedangkan sisanya masih dipengaruhi oleh faktor lain yang tidak tercakup dalam fitur yang digunakan atau ketidakteraturan pola data.



Gambar 4. SVR Actual vs Prediction

Nilai MSE yang relatif besar mengindikasikan bahwa terdapat selisih kesalahan yang cukup signifikan antara nilai prediksi dan nilai aktual pada sebagian data. Kondisi ini menunjukkan bahwa meskipun model mampu menangkap kecenderungan umum variasi O₃, akurasi prediksi masih terbatas, terutama pada kondisi nilai O₃ tertentu yang mungkin bersifat ekstrem atau fluktuatif. Secara umum, model menunjukkan kemampuan prediktif pada tingkat moderat, namun performanya belum optimal. Hal ini membuka indikasi bahwa penambahan variabel lain, khususnya parameter meteorologi seperti suhu, kelembaban, dan kecepatan angin, berpotensi meningkatkan kualitas prediksi pada tahap pengembangan model selanjutnya.

Gambar 5. *Hyperplane & Epsilon Tube*

Bagian pembahasan ini menganalisis hasil kinerja kedua model supervised learning dalam memprediksi nilai ISPU O_3 serta membandingkan efektivitasnya berdasarkan parameter evaluasi dan karakteristik prediksi. Pada model pertama yang menggunakan Random Forest Regressor, diperoleh nilai R^2 sebesar 0,49 yang menunjukkan bahwa hampir separuh variasi konsentrasi O_3 dapat dijelaskan oleh variabel input. Nilai MAE sebesar 12,15 dan RMSE sebesar 16,15 mengindikasikan adanya tingkat kesalahan yang masih dapat diterima namun menunjukkan keberadaan error yang lebih besar pada sebagian sampel data. Sementara itu, model kedua menghasilkan R^2 sebesar 0,3845 dengan MSE sebesar 317,43, yang merefleksikan akurasi prediksi yang relatif lebih rendah dan tingkat kesalahan yang lebih tinggi dibandingkan model pertama. Secara umum, perbedaan nilai R^2 pada kedua model menunjukkan bahwa Random Forest memiliki kemampuan generalisasi yang lebih baik dalam menangkap pola hubungan antar variabel polutan terhadap pembentukan O_3 , sejalan dengan temuan penelitian sebelumnya yang menyatakan bahwa metode ensemble cenderung lebih stabil dalam memodelkan data lingkungan yang bersifat non-linear dan fluktuatif [1], [2].

Dari sisi analisis fitur, model pertama menunjukkan bahwa PM10, SO_2 , dan NO_2 memiliki kontribusi dominan dalam mempengaruhi prediksi O_3 , sementara CO memiliki pengaruh yang paling rendah. Hal ini sejalan dengan studi atmosfer yang menyebutkan bahwa interaksi antar polutan memainkan peran penting dalam pembentukan ozon troposferik [8]. Pada model kedua, meskipun menggunakan variasi fitur yang lebih beragam, performa yang dihasilkan tidak lebih baik. Hal ini menunjukkan bahwa penambahan variabel tidak selalu menjamin peningkatan akurasi, terutama jika variabel tersebut tidak memiliki korelasi kuat atau data memiliki tingkat variasi temporal yang tinggi [1], [8].

Dari aspek visualisasi, model pertama menunjukkan hubungan yang lebih baik antara nilai aktual dan nilai prediksi, terutama pada kisaran konsentrasi rendah hingga menengah. Namun, terdapat kecenderungan underprediction pada nilai O_3 tinggi, yang mengindikasikan bahwa model masih

menghadapi tantangan dalam memodelkan kondisi ekstrem, suatu fenomena yang umum terjadi pada pemodelan kualitas udara [8]. Model kedua menunjukkan pola yang lebih menyebar dengan error yang lebih bervariasi, sehingga tingkat keandalan prediksinya relatif lebih rendah.

Secara keseluruhan, dapat disimpulkan bahwa model Random Forest pada pengujian pertama memiliki performa yang lebih baik dibandingkan model kedua, baik dari segi nilai R^2 , tingkat error, maupun konsistensi prediksi. Namun, kedua model sama-sama menunjukkan bahwa variabilitas O_3 tidak hanya dipengaruhi oleh polutan lain, tetapi juga sangat dipengaruhi oleh faktor meteorologi dan dinamika atmosfer yang belum dimasukkan dalam pemodelan. Oleh karena itu, pengembangan model selanjutnya direkomendasikan untuk menambahkan variabel suhu, radiasi matahari, kelembaban, dan kecepatan angin guna meningkatkan akurasi prediksi serta kemampuan model dalam menangkap kondisi ekstrem konsentrasi O_3 .

5. KESIMPULAN DAN SARAN

Berdasarkan hasil pemodelan dan evaluasi yang telah dilakukan, dapat disimpulkan bahwa model supervised learning mampu memprediksi konsentrasi ISPU O_3 dengan tingkat akurasi yang moderat. Model Random Forest Regressor menunjukkan performa terbaik dengan nilai R^2 sebesar 0,49, MAE 12,15, dan RMSE 16,15, yang menandakan kemampuan model dalam menjelaskan sekitar 49% variasi data O_3 . Sementara itu, model kedua hanya mampu mencapai R^2 sebesar 0,3845 dengan nilai error yang lebih besar, sehingga memiliki akurasi prediksi yang lebih rendah. Analisis feature importance menunjukkan bahwa PM10, SO_2 , dan NO_2 merupakan variabel yang paling berpengaruh terhadap pembentukan O_3 , sedangkan CO memiliki kontribusi yang lebih kecil. Hasil ini memperkuat temuan studi sebelumnya bahwa pembentukan ozon troposferik sangat dipengaruhi oleh interaksi polutan serta dinamika atmosfer yang kompleks [1], [2]. Selain itu, kedua model masih menunjukkan keterbatasan dalam memprediksi nilai O_3 ekstrem, terutama pada konsentrasi tinggi, sebagaimana juga dilaporkan dalam penelitian kualitas udara lainnya [1], [8]. Berdasarkan temuan penelitian ini, terdapat beberapa rekomendasi untuk pengembangan penelitian selanjutnya. Pertama, perlu dilakukan penambahan variabel pendukung seperti suhu, radiasi matahari, kelembaban, dan kecepatan angin guna meningkatkan kemampuan model dalam menangkap dinamika atmosfer dan fenomena ekstrem konsentrasi O_3 [7], [10]. Kedua, penggunaan algoritma yang lebih kompleks seperti Gradient Boosting, XGBoost, atau metode deep learning dapat dipertimbangkan untuk memperoleh performa prediksi yang lebih baik [1], [2]. Ketiga, diperlukan pengujian model pada periode data yang lebih panjang dan beragam untuk meningkatkan generalisasi model. Terakhir, diperlukan validasi lapangan dan integrasi

dengan sistem pemantauan kualitas udara agar hasil prediksi dapat dimanfaatkan secara praktis dalam mendukung kebijakan pengendalian polusi udara dan perencanaan lingkungan.

DAFTAR PUSTAKA

- [1] Q. Pan, F. Harrou, and Y. Sun, "A comparison of machine learning methods for ozone pollution prediction," *J Big Data*, vol. 10, no. 1, p. 63, 2023, doi: 10.1186/s40537-023-00748-x.
- [2] L. Widyarini and H. Purnomo, "Air Quality Prediction Using the Support Vector Machine Algorithm," *Journal of Information Systems and Informatics*, vol. 6, no. 2, Jun. 2024, doi: 10.51519/journalisi.v6i2.705.
- [3] V. Tundjungsari, *Dasar Machine Learning*, 2nd ed., vol. 1. Jakarta: DeePublish, 2024.
- [4] E. Avşar, S. Dümen, and U. Çevik, "Performance Evaluation of Various Regression Models and Features for Prediction of Ozone Concentration," *Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, vol. 35, no. 3, pp. 567–574, 2020, doi: 10.21605/cukurovaummfd.845985.
- [5] V. Tundjungsari, "Open Health Data Development for Machine Learning-based Resource Sharing: Indonesia Case Study," *International Journal of Artificial Intelligence Research*, vol. 6, no. 2, pp. 2579–7298, 2022, doi: 10.29099/ijair.v6i2.581.
- [6] A. Setiawan, "Random Forest Algorithm to Measure the Air Pollution Standard Index," *Knowledge Engineering and Data Science*, vol. 7, Nov. 2024, doi: 10.17977/um018v7i12024p86-100.
- [7] S. G. Gunjal and S. Kamalapurkar, "Air Pollution Prediction Using Machine Learning Supervised Learning Approach", [Online]. Available: www.ijstr.org
- [8] A. Yafouz *et al.*, "Comprehensive comparison of various machine learning algorithms for short-term ozone concentration prediction," *Alexandria Engineering Journal*, vol. 61, no. 6, pp. 4607–4622, 2022, doi: <https://doi.org/10.1016/j.aej.2021.10.021>.
- [9] V. Kumar, S. Singh, Z. Ahmed, and N. Verma, "Air Pollution Prediction using Machine Learning Algorithms: A Systematic Review." [Online]. Available: www.ijert.org
- [10] A.-L. Balogun and A. Tella, "Modelling and investigating the impacts of climatic variables on ozone concentration in Malaysia using correlation analysis with random forest, decision tree regression, linear regression, and support vector regression," *Chemosphere*, vol. 299, p. 134250, 2022, doi: <https://doi.org/10.1016/j.chemosphere.2022.134250>