Computer Science (CO-SCIENCE)
Vol. 5 No. 2 Juli 2025
Akreditasi Sinta 4 Nomor SK : 230/E/KPT/2022

Model Prediksi Risiko Kesehatan Perkotaan Berbasis Lingkungan dengan XGBoost

Muhammad Kahfi Aulia[1*], Eka Utaminingsih[2], Nanang Prihatin[3]

[1,2]Universitas Bumi Persada
Alue Awe, Muara Dua, Lhokseumawe, Aceh, Indonesia

[3]Politeknik Negeri Lhokseumawe
Buketrata, Blang Mangat, Lhokseumawe, Aceh, Indonesia

e-mail: [1]auliamuhammadkahfi@gmail.com, [2]ekautami921@gmail.com, [3]nanang@pnl.ac.id

(*) Corresponding Author

Artikel Info : Diterima : 11-06-2025 | Direvisi : 29-06-2025 | Disetujui : 10-07-2025

Abstrak - Kualitas udara perkotaan yang buruk merupakan isu kesehatan masyarakat yang serius, terutama di wilayah dengan urbanisasi tinggi. Penelitian ini bertujuan memprediksi risiko kesehatan akibat polusi udara menggunakan metode pembelajaran mesin berbasis variabel lingkungan. Dataset yang digunakan adalah *Urban Air Quality and Health Impact*, terdiri atas 1.000 baris dan 46 kolom, mencakup suhu, kelembapan, kecepatan angin, titik embun, indeks ultraviolet (UV), dan skor risiko kesehatan dari kota-kota besar di Amerika Serikat. Sebagai peningkatan dari studi sebelumnya yang menggunakan regresi linier dan *Random Forest* (R-squared 0,89; Mean Squared Error/MSE 0,65), penelitian ini menerapkan model *Extreme Gradient Boosting* (XGBoost) yang dioptimasi menggunakan teknik *Randomized Search* terhadap beberapa hiperparameter utama. Model ini dilatih dan diuji dengan pembagian data 80:20, menghasilkan *R-squared* sebesar 0,9692 dan MSE sebesar 0,0122. Titik embun dan kecepatan angin merupakan fitur paling berpengaruh. Dataset yang digunakan bersifat sintetis namun menyerupai pola lingkungan di kota-kota Indonesia. Penelitian ini tidak menggunakan kerangka *text mining*, melainkan pendekatan regresi terawasi berbasis data lingkungan. Kebaruan utama terletak pada penerapan pertama XGBoost yang dioptimasi dengan fitur kompleks seperti suhu terasa untuk estimasi risiko kesehatan perkotaan. Keterbatasan penelitian ini meliputi belum adanya validasi dengan data lokal serta belum dianalisisnya hubungan antarvariabel secara mendalam.

Kata Kunci : kualitas udara, risiko kesehatan, XGBoost, pembelajaran mesin, regresi lingkungan

Abstracts *- Poor urban air quality is a major public health concern, especially in highly urbanized areas. This study aims to predict health risks associated with air pollution using machine learning techniques based on environmental variables. The dataset used, Urban Air Quality and Health Impact, contains 1,000 rows and 46 columns, including temperature, humidity, wind speed, dew point, ultraviolet (UV) index, and health risk scores from major U.S. cities. As an improvement over previous studies using linear regression and Random Forest (R-squared 0.89; Mean Squared Error/MSE 0.65), this research implements an optimized Extreme Gradient Boosting (XGBoost) model. The model was fine-tuned using Randomized Search on key hyperparameters and evaluated with an 80:20 data split. It achieved an R-squared of 0.9692 and MSE of 0.0122. Dew point and wind speed were identified as the most influential features. Although synthetic, the dataset reflects environmental patterns similar to Indonesian urban areas. This study does not adopt a text mining framework but instead uses a supervised regression approach based on environmental features. Its main novelty lies in the first application of an optimized XGBoost model using complex variables such as feels-like temperature to estimate urban health risk. Limitations include the absence of real-world validation with Indonesian data and the lack of analysis on interactions between variables.*

Keywords : air quality, health risk, XGBoost, fine-tuning, machine learning, environmental variables

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
Copyright (c) 2025 The Autour(s)

PENDAHULUAN

Pertumbuhan pesat populasi di wilayah perkotaan telah menyebabkan meningkatnya paparan masyarakat terhadap risiko lingkungan yang kompleks, seperti pencemaran udara dan kondisi cuaca ekstrem. Menurut *World Health Organization* (2021), polusi udara menjadi penyebab lebih dari tujuh juta kematian dini setiap tahunnya, setara dengan risiko kesehatan global lainnya seperti malnutrisi dan penyakit menular. Konsentrasi polutan seperti $PM_{2.5}$, NO_2, dan O_3 telah terbukti secara signifikan meningkatkan insiden penyakit pernapasan dan kardiovaskular, terutama di wilayah urban dengan tingkat emisi tinggi (Rosatul Umah & Eva Gusmira, 2024). Di Indonesia, dampak polusi udara telah diamati secara luas. Garmini dan Purwana (2020) menunjukkan bahwa paparan SO_2 dan polutan dalam ruang tertutup meningkatkan risiko ISPA pada balita. Studi serupa oleh Inayah (2025) juga menegaskan bahwa kejadian ISPA pada balita berhubungan erat dengan konsentrasi PM dan NO_2 di lingkungan ambien. Di DKI Jakarta, peningkatan tingkat polusi udara berkorelasi langsung dengan meningkatnya kasus pneumonia balita (Munggaran *et al.*, 2024), sebuah temuan yang konsisten dengan studi Nova *et al.* (2023) di sekitar kawasan industri baja.

Selain pencemaran udara, kondisi meteorologis seperti suhu tinggi, kelembaban, dan variasi angin juga mempengaruhi risiko kesehatan. Titik embun yang tinggi, misalnya, telah dikaitkan dengan eksaserbasi penyakit paru obstruktif kronik (Márovics *et al.*, 2024) dan peningkatan penyakit *Mycobacterium tuberculosis* (Krishnan *et al.*, 2022). Brimicombe *et al.* (2024) menyatakan bahwa indeks panas seperti *heat index* secara signifikan berdampak pada morbiditas ibu dan bayi, menjadikannya indikator penting dalam peringatan dini terhadap gelombang panas. Dampak gabungan antara polusi udara dan kondisi suhu ekstrem juga ditemukan dalam studi lintas negara oleh Castro *et al.* (2025), yang mengungkap bahwa interaksi antara $PM_{2.5}$ dan suhu tinggi meningkatkan risiko kematian secara signifikan. Di konteks nasional, studi Rahayuningtyas *et al.* (2025) di Kabupaten Bantul menunjukkan bahwa kombinasi faktor iklim seperti kelembaban, *dew point*, dan suhu bola basah berkontribusi terhadap prediksi kejadian demam berdarah dengue (DBD). Hal ini memperkuat argumen bahwa kondisi atmosfer memainkan peran penting dalam mempengaruhi beban penyakit masyarakat.

Untuk menjawab kompleksitas ini, pendekatan berbasis *machine learning* menjadi pilihan yang menjanjikan dalam mengembangkan sistem prediksi risiko kesehatan lingkungan. Beberapa pendekatan sebelumnya telah mencoba memprediksi dampak polusi udara terhadap kesehatan, salah satunya adalah studi oleh Pathak *et al.* (2024) yang menggunakan algoritma *Random Forest* dengan data lingkungan perkotaan. Meskipun menghasilkan nilai koefisien determinasi (R^2) sebesar 0,89, pendekatan tersebut masih memiliki keterbatasan, seperti akurasi yang kurang optimal dan ketidakmampuan menangkap relasi non-linear yang kompleks antar fitur lingkungan. *Random Forest* juga cenderung menghasilkan model yang sulit diinterpretasikan secara kebijakan karena tidak memprioritaskan variabel secara eksplisit.

Salah satu algoritma yang menunjukkan performa unggul adalah *Extreme Gradient Boosting* (XGBoost). Sebagai respons terhadap masalah yang diangkat pada penelitian ini, peneliti mengusulkan pemanfaatan XGBoost, sebuah algoritma pembelajaran *ensemble* yang menggabungkan kekuatan *decision tree* dalam bentuk boosting berurutan. XGBoost tidak hanya memiliki performa prediktif yang unggul, tetapi juga menawarkan kontrol regularisasi untuk mencegah *overfitting* serta kemampuan untuk menangani data tidak seimbang (Asnawi *et al.*, 2025). Zhou *et al.* (2025) menunjukkan bahwa XGBoost dapat secara akurat memprediksi volume kunjungan rumah sakit dengan memperhitungkan parameter lingkungan seperti $PM_{2.5}$ dan suhu. Kemampuan XGBoost dalam menangani relasi non-linear dan interaksi antar fitur menjadikannya sangat cocok untuk data lingkungan yang kompleks.

Selain akurasi, aspek interpretabilitas menjadi penting untuk keperluan kebijakan publik. XGBoost mendukung analisis *feature importance* yang membantu mengidentifikasi variabel lingkungan paling berpengaruh, seperti titik embun atau suhu terasa (Lev, 2022). Dengan pendekatan ini, di masa depan, sistem peringatan dini dapat difokuskan pada variabel utama seperti titik embun dan kecepatan angin kencang, sebagaimana yang juga disarankan dalam penelitian Sapna *et al.* (2024) dan Fauzianto & Ali (2024) yang menyoroti pentingnya pemantauan polutan dan kondisi atmosfer pasca kebakaran lahan.

Penelitian ini bertujuan untuk mengembangkan model prediksi risiko kesehatan perkotaan berbasis regresi XGBoost yang telah dioptimasi, menggunakan data kualitas udara dan parameter meteorologi. Selain menghasilkan model prediktif yang akurat, studi ini juga mengevaluasi kontribusi relatif dari setiap variabel lingkungan terhadap risiko kesehatan. Kontribusi utama penelitian ini terletak pada penerapan pertama model XGBoost yang telah di-*finetuning* dalam konteks analisis kesehatan urban di Indonesia, serta penyajian basis data empiris yang mendekati kondisi lokal melalui pemanfaatan data sintetis yang representatif.

.

METODE PENELITIAN

Penelitian ini menerapkan pendekatan kuantitatif analitik untuk membangun model prediksi risiko kesehatan perkotaan menggunakan algoritma XGBoost. Dataset berisi 1000 baris dan 46 kolom. Semua baris yang memiliki nilai kosong pada variabel target *Health Risk Score* dihapus, dan fitur numerik yang memiliki nilai hilang diimputasi menggunakan nilai median, karena metode ini relatif tahan terhadap outlier. Jumlah data setelah pra-pemrosesan serta distribusi variabel target dicatat untuk memastikan transparansi metodologis. Selanjutnya, data dibagi menjadi subset pelatihan sebesar 80% dan pengujian sebesar 20% dengan pembagian stratifikasi berdasarkan nilai *Health Risk Score*, dan penguncian parameter random_state dilakukan agar hasil dapat direproduksi secara konsisten.

Model dikembangkan menggunakan pustaka xgboost.XGBRegressor dari Python, dan dilakukan penyetelan (*fine-tuning*) hyperparameter dengan menggunakan *RandomizedSearchCV* serta validasi silang (*5-fold cross-validation*). Ruang pencarian mencakup parameter *n_estimators* (100–1.000), *max_depth* (3–9), *learning_rate* (0,01–0,2), dan *subsample* (0,5–1,0). Metode ini dipilih berdasarkan temuan bahwa *Randomized Search* lebih efisien dalam menjelajahi ruang parameter dibandingkan *Grid Search* (Pramudhyta & Rohman, 2024; Subaşı, 2024). XGBoost juga menyediakan regularisasi L2 internal yang secara empiris terbukti efektif dalam menekan risiko *overfitting* (Bentéjac *et al.*, 2021).

Evaluasi model dilakukan dengan dua metrik utama: *Mean Squared Error* (MSE) dan koefisien determinasi (*R-squared*). MSE mengukur rata-rata kuadrat selisih antara nilai aktual dan prediksi, dihitung menggunakan persamaan (1):

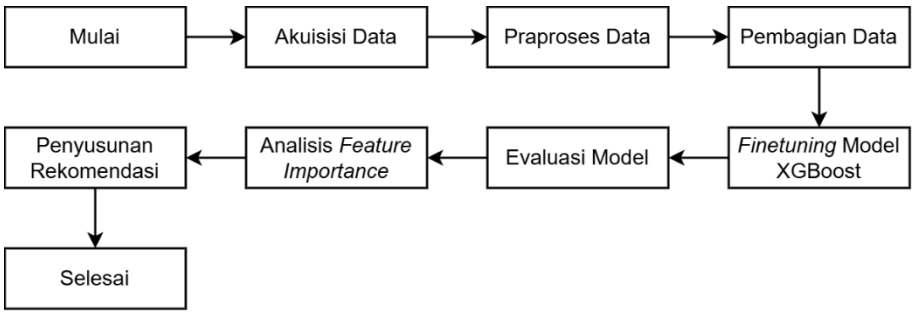
$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

dengan n adalah jumlah sampel, y_i adalah nilai aktual, dan $\hat{y}_i$ adalah nilai prediksi. MSE merupakan metrik yang umum digunakan dalam regresi karena mempertahankan satuan variabel target (Ozili, 2023). Sementara itu, *R-squared* mengukur proporsi variasi data yang dapat dijelaskan oleh model prediktif, dihitung dengan persamaan (2):

$$R^2 = 1 - \left(\frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}\right) \quad (2)$$

di mana $\bar{y}$ adalah rata-rata nilai aktual. Nilai R^2 mendekati 1 menunjukkan bahwa model memiliki kemampuan penjelasan yang tinggi terhadap variabilitas data (Chicco *et al.*, 2021). Kedua metrik ini dihitung pada data pengujian serta pada setiap lipatan *cross-validation* untuk menilai akurasi dan kemampuan generalisasi model.

Tahap selanjutnya adalah analisis *feature importance* menggunakan metrik *gain* dari XGBoost untuk mengidentifikasi fitur-fitur lingkungan yang paling berkontribusi terhadap prediksi skor risiko kesehatan. Kemudian tahap akhir dari penelitian ini adalah penyusunan rekomendasi berbasis hasil analisis *feature importance* dari model XGBoost. Setelah model dievaluasi, fitur-fitur lingkungan yang memiliki kontribusi tertinggi terhadap prediksi skor risiko kesehatan diidentifikasi menggunakan metrik *gain*. Berdasarkan fitur-fitur tersebut, rekomendasi awal dirumuskan dengan mengusulkan ambang batas atau kombinasi nilai parameter lingkungan yang dapat diasosiasikan dengan peningkatan risiko kesehatan, sehingga dapat digunakan sebagai indikator peringatan dini atau masukan kebijakan mitigasi. Rekomendasi ini bersifat eksploratif dan bertujuan menyediakan dasar awal bagi pengembangan sistem pemantauan kesehatan masyarakat yang berbasis data lingkungan secara lebih responsif dan adaptif. Rangkaian proses penelitian secara keseluruhan disajikan dalam bentuk blok diagram pada Gambar 1.



Sumber : Hasil Penelitian (2025)
Gambar 1. Rangkaian proses penelitian

HASIL DAN PEMBAHASAN

1. Dataset

Dataset *Urban Air Quality and Health Impact* (Abdullah & Yaqoob, 2024) adalah dataset publik dari Kaggle yang berisi 1.000 observasi dan 46 variabel yang merekam parameter lingkungan serta dampaknya terhadap kesehatan masyarakat di kota-kota Amerika Serikat. Fitur-fitur dalam dataset ini mencakup tujuh kategori

utama, yaitu: (1) identifikasi spasial-temporal (misalnya tanggal, kota, musim, dan hari), (2) parameter termal seperti suhu maksimum, suhu persepsi, titik embun, dan *heat index*, (3) kondisi atmosfer seperti kelembapan, tekanan, angin, dan jarak pandang, (4) presipitasi dan salju, (5) radiasi matahari dan indeks cuaca ekstrem, (6) variabel kesehatan berupa *Health Risk Score* dan deskripsi kondisi atmosfer, serta (7) metadata tambahan seperti fase bulan dan sumber data.

Setiap fitur telah melalui proses validasi kualitas dan dapat dimanfaatkan dalam studi epidemiologi lingkungan, prediksi risiko kesehatan berbasis cuaca, serta analisis perubahan iklim mikro perkotaan. Variabel-variabel termal dan kelembapan khususnya memberikan wawasan penting terkait fenomena *heat stress* dan implikasinya terhadap kesehatan populasi. Contoh sampel data dan nama fiturnya disajikan pada Tabel 1.

Tabel 1. Sampel dataset dan fiturnya

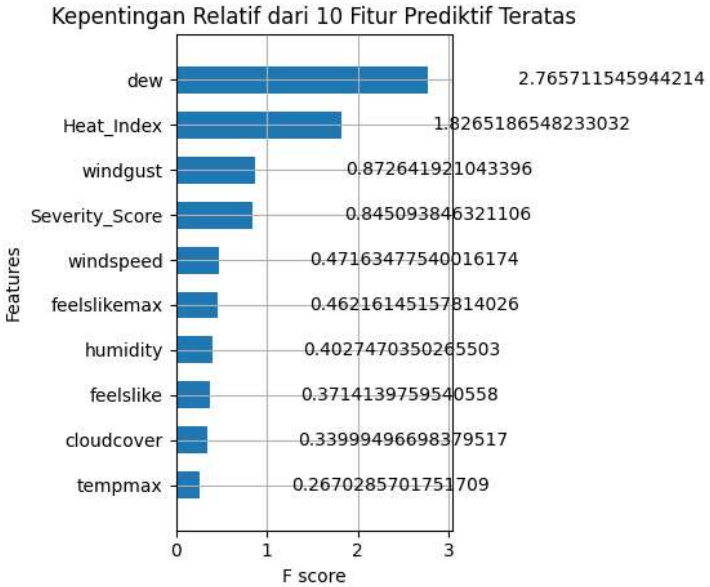
tempmax	tempmin	temp	feelslikemax	feelslikemin	Health_Risk_Score
106,1	91,0	98,5	104,0	88,1	10,52217031099
103,9	87,0	95,4	100,5	84,7	10,06233162414
105,0	83,9	94,7	99,9	81,6	9,673386510582
106,1	81,2	93,9	100,6	79,5	9,411519469002
106,1	82,1	94,0	101,0	80,0	9,515178582170

Sumber: Dataset *Urban Air Quality and Health Impact* (2024)

2. Hasil Penelitian

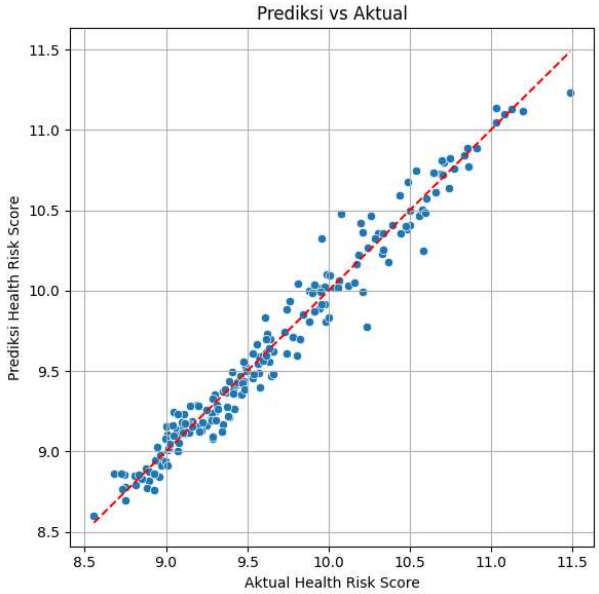
Model *XGBoost Regression* yang telah dioptimasi melalui proses *tuning hyperparameter* dengan *RandomizedSearchCV* pada 5-fold cross-validation menghasilkan performa prediktif yang sangat tinggi. Parameter optimal yang diperoleh adalah n_estimators=700, max_depth=3, learning_rate=0.05, subsample=0.6, colsample_bytree=0.6, reg_alpha=0.5, reg_lambda=1.5, dan gamma=0. Dengan konfigurasi ini, model mencapai nilai Mean Squared Error (MSE) sebesar 0,0122 dan koefisien determinasi (R^2) sebesar 0,9692 pada data uji. Artinya, model mampu menjelaskan hampir 97% variasi dalam skor risiko kesehatan berdasarkan parameter lingkungan, dengan rata-rata kesalahan prediksi sebesar $\sqrt{0,0122} \approx 0,11$ unit pada skala risiko yang diasumsikan berkisar antara 0–12.

Gambar 2 menyajikan *feature importance* dari 10 fitur utama berdasarkan metrik *gain*, yang mengukur kontribusi fitur terhadap penurunan fungsi loss dalam pemodelan. Fitur dew point menjadi yang paling signifikan dengan skor 2,76, menunjukkan pengaruh besar dari kelembapan absolut terhadap prediksi risiko. Fitur berikutnya adalah Heat Index (1,83), yang mewakili kombinasi suhu dan kelembapan dalam mencerminkan beban panas terhadap tubuh manusia. Wind Gust, atau kecepatan hembusan angin, menempati posisi ketiga (0,87) dan mengindikasikan bahwa variabilitas atmosfer ekstrem turut meningkatkan risiko kesehatan. Fitur-fitur lainnya seperti *Severity_Score*, *feelslikemax*, *humidity*, dan *cloudcover* juga memberikan kontribusi bermakna.



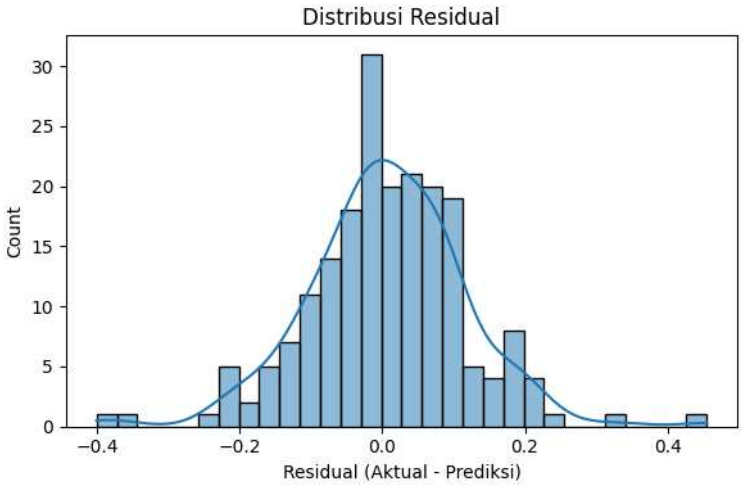
Sumber : Hasil Penelitian (2025)
Gambar 2. *Feature importance* dari 10 fitur utama.

Gambar 3 menunjukkan hubungan antara nilai prediksi dan nilai aktual dari *Health Risk Score* pada data uji. Titik-titik data tersebar sangat dekat dengan garis referensi (y = x) yang digambarkan dalam garis putus-putus merah, menandakan tingkat kecocokan yang tinggi antara hasil prediksi dan observasi sebenarnya. Konsistensi ini mengindikasikan bahwa model mampu memberikan estimasi yang akurat di seluruh rentang skor, tidak hanya pada nilai tengah. Pola ini juga menunjukkan bahwa tidak terdapat pola sistematis terhadap overestimasi atau underestimasi pada rentang skor tertentu.



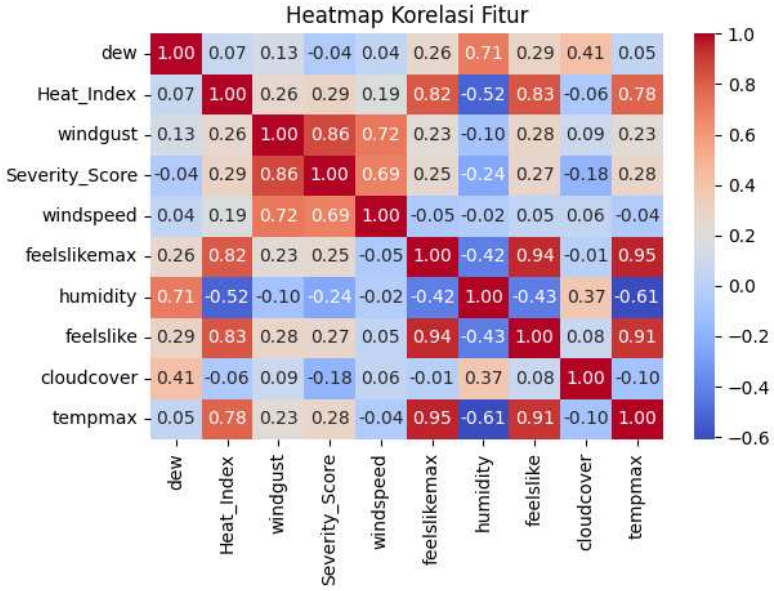
Sumber : Hasil Penelitian (2025)
Gambar 3. Hubungan antara nilai prediksi dan nilai aktual dari *Health Risk Score*

Distribusi error (residual) ditampilkan dalam Gambar 4, yang memperlihatkan histogram dari selisih antara nilai aktual dan prediksi. Distribusi ini berbentuk simetris dan mendekati distribusi normal, dengan puncak yang jelas di sekitar nol. Artinya, kesalahan model tidak condong ke arah positif maupun negatif, dan prediksi yang dihasilkan tidak memiliki bias sistematis. Ini penting sebagai indikator bahwa model tidak hanya akurat secara statistik, tetapi juga seimbang secara prediktif dalam berbagai kondisi atmosfer.



Sumber : Hasil Penelitian (2025)
Gambar 4. Distribusi residual

Gambar 5 adalah *heatmap* korelasi antar 10 fitur terpenting yang digunakan dalam model. Korelasi sangat tinggi ditemukan antara *feelslikemax*, *feelslike*, dan *tempmax* ($r > 0{,}9$), yang menunjukkan adanya redudansi informasi suhu di antara fitur-fitur ini. Korelasi yang kuat ini dapat dijadikan pertimbangan dalam penyederhanaan model di masa depan melalui reduksi dimensi atau seleksi fitur. Di sisi lain, korelasi rendah antara *windgust*, *Severity_Score*, dan fitur lainnya menunjukkan bahwa fitur-fitur tersebut menyumbang informasi yang unik dan independen terhadap output model. Korelasi antara *dew* dan *humidity* ($r \approx 0{,}71$) mengindikasikan bahwa titik embun berhasil menangkap dimensi kelembapan lingkungan secara lebih efektif daripada kelembapan relatif semata.



Sumber : Hasil Penelitian (2025)
Gambar 5. *Heatmap* korelasi antar 10 fitur terpenting

Jika dibandingkan dengan studi Pathak *et al.* (2024), model XGBoost dalam penelitian ini menunjukkan peningkatan kinerja yang nyata. Pathak menggunakan Random Forest untuk prediksi dampak kualitas udara dan hanya mencapai R^2 sebesar 0,89 dan RMSE 0,65. Model yang dikembangkan dalam studi ini tidak hanya mencapai $R^2 = 0{,}9692$ (peningkatan lebih dari 8%), tetapi juga memberikan interpretasi fitur yang jauh lebih rinci dan akurat berkat metrik gain dalam XGBoost. Perbedaan signifikan ini menunjukkan bahwa dengan tuning dan analisis fitur mendalam, XGBoost mampu memberikan prediksi yang lebih presisi dan bermanfaat secara aplikatif.

Secara keseluruhan, hasil penelitian ini memperlihatkan bahwa model XGBoost Regression yang dikembangkan tidak hanya akurat secara kuantitatif, tetapi juga kuat secara diagnostik—dengan distribusi error yang seimbang, dominasi fitur yang konsisten dengan teori medis, dan struktur internal antar fitur yang mendukung interpretabilitas lanjutan.

3. Pembahasan

Dengan hasil evaluasi MSE sebesar 0,0122 dan R^2 sebesar 0,9692, model ini menunjukkan kinerja yang sangat superior dibandingkan metode regresi tradisional. Dalam studi Pathak *et al.* (2024), *Random Forest* dilaporkan hanya mencapai R^2 sebesar 0,89 dan RMSE sekitar 0,65 dalam konteks prediksi dampak lingkungan terhadap kesehatan, tanpa integrasi eksplisit analisis fitur atau proses tuning parameter secara sistematis. Dibandingkan dengan itu, pendekatan XGBoost dalam studi ini tidak hanya menunjukkan akurasi lebih tinggi, tetapi juga menghasilkan model yang lebih stabil dan dapat dijelaskan secara interpretatif melalui analisis *feature importance*.

Pentingnya fitur *dew point* dalam model selaras dengan literatur medis yang menyebutkan bahwa kelembapan absolut memainkan peran lebih signifikan dalam memicu eksaserbasi pernapasan dibanding kelembapan relatif. Kombinasi dengan *heat index* dan *wind gust* sebagai prediktor utama memperkuat bukti bahwa variabilitas termal dan atmosferik memiliki dampak besar terhadap beban penyakit. Keunikan *wind gust* sebagai fitur dengan korelasi rendah namun *gain* tinggi juga menunjukkan bahwa elemen-elemen cuaca ekstrem yang tidak selalu muncul sebagai tren jangka panjang tetap dapat meningkatkan risiko secara tajam dalam jangka pendek.

Korelasi antar fitur dalam heatmap memperlihatkan pola hubungan yang dapat dimanfaatkan dalam rekayasa fitur selanjutnya. Penghapusan fitur dengan korelasi sangat tinggi dapat mengurangi multikolinearitas dan mempercepat waktu komputasi tanpa kehilangan informasi signifikan. Sebaliknya, fitur dengan korelasi rendah namun skor gain tinggi sebaiknya dipertahankan sebagai bagian dari fitur esensial yang menyumbang informasi independen terhadap variabel target.

Dari sisi implementasi, hasil ini sangat aplikatif dalam konteks perkotaan Indonesia. Kota-kota besar seperti Jakarta, Surabaya, Bandung, dan Makassar sering mengalami kelembapan tinggi, fluktuasi suhu ekstrem, dan paparan polusi kronis. Model ini dapat diintegrasikan ke dalam sistem pemantauan kualitas udara dan kesehatan lingkungan oleh Dinas Kesehatan atau instansi mitigasi bencana seperti BPBD. Misalnya, ketika nilai *dew point* dan *wind gust* melebihi ambang tertentu, sistem dapat mengeluarkan peringatan dini bagi kelompok rentan seperti lansia dan penderita penyakit kronis. Selain itu, perencanaan kota juga dapat mengadopsi temuan ini dalam perancangan ruang terbuka hijau dan ventilasi alami.

Namun, penelitian ini masih memiliki batasan. Dataset yang digunakan adalah dataset sekunder dari luar negeri dan belum dikalibrasi dengan kondisi atmosfer dan profil kesehatan lokal Indonesia. Variabel sosial-

ekonomi dan kepadatan populasi yang juga berperan dalam menentukan dampak lingkungan terhadap kesehatan belum dimasukkan dalam model. Oleh karena itu, validasi eksternal dan pengembangan lanjutan dengan data lokal sangat direkomendasikan untuk memperluas jangkauan dan efektivitas implementasi model di Indonesia.

Sebagai simpulan bagian ini, pendekatan XGBoost Regression yang telah dioptimasi dan divalidasi dalam studi ini bukan hanya unggul dibanding metode lain dalam hal akurasi, tetapi juga memberikan informasi bernilai tinggi tentang faktor risiko lingkungan yang dapat ditindaklanjuti secara operasional. Keunggulan ini menjadikan model ini layak untuk diadopsi dalam sistem pemantauan kesehatan lingkungan berbasis data di wilayah urban Indonesia.

KESIMPULAN

Penelitian ini menunjukkan bahwa *XGBoost Regression* yang dioptimasi melalui penyetelan hiperparameter mampu memprediksi risiko kesehatan masyarakat perkotaan dengan akurasi tinggi (MSE = 0,0122; R^2 = 0,9692). Model secara konsisten mengidentifikasi *dew point* dan *wind gust* sebagai fitur dominan, yang mencerminkan pentingnya kombinasi tekanan termal dan variabilitas atmosfer dalam mempengaruhi beban kesehatan. Temuan ini berkontribusi secara teoretis pada literatur pemodelan prediktif berbasis lingkungan, sekaligus memberikan implikasi praktis berupa peluang penerapan sistem peringatan dini dan pemantauan risiko berbasis data, terutama di wilayah urban tropis seperti kota-kota besar di Indonesia.

Namun demikian, penggunaan dataset dari Amerika Serikat menghadirkan potensi bias jika diterapkan langsung di Indonesia tanpa kalibrasi lokal. Model ini juga belum mempertimbangkan faktor sosial-ekonomi, status kesehatan dasar, dan distribusi kerentanan populasi. Oleh karena itu, penelitian selanjutnya perlu memanfaatkan data lokal yang mencerminkan kondisi iklim, demografi, dan perilaku masyarakat Indonesia. Integrasi dengan data real-time (sensor udara atau catatan medis), serta penerapan pendekatan *explainable AI* seperti SHAP, akan memperkuat relevansi kebijakan dan kepercayaan pengguna. Dengan pendekatan yang adaptif dan berbasis bukti, sistem ini dapat menjadi fondasi bagi perencanaan kota yang lebih tanggap terhadap risiko lingkungan.

ACKNOWLEDGEMENTS

Peneliti mengucapkan terima kasih kepada pengembang dataset *Urban Air Quality and Health Impact* (Abdullah & Yaqoob, 2024) yang telah membagikan data secara terbuka melalui Kaggle di bawah lisensi MIT. Dataset ini menjadi dasar utama dalam proses pengembangan dan evaluasi model prediktif pada studi ini. Penelitian ini tidak menerima pendanaan eksternal dan seluruh analisis dilakukan secara independen.

CONFLICT OF INTEREST

Peneliti menyatakan tidak memiliki konflik kepentingan, baik secara finansial, institusional, maupun pribadi, yang dapat memengaruhi hasil atau interpretasi dari penelitian ini.

REFERENSI

Abdullah, M., & Yaqoob, S. (2024). *Urban Air Quality and Health Impact Analysis*. Kaggle. https://doi.org/https://doi.org/10.34740/kaggle/dsv/9341077

Asnawi, M. F., Fitriyanto, N., & Pamoengkas, M. A. (2025). The Application Of XGBoost Classification For Fraud Detection In Credit Card. *Clean Energy and Smart Technology*, *03*(02), 41–48.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. In *Artificial Intelligence Review* (Vol. 54, Issue 3). Springer Netherlands. https://doi.org/10.1007/s10462-020-09896-5

Brimicombe, C., Conway, F., Portela, A., Lakhoo, D., Roos, N., Gao, C., Solarin, I., & Jackson, D. (2024). A scoping review on heat indices used to measure the effects of heat on maternal and perinatal health. *BMJ Public Health*, *2*(1), e000308. https://doi.org/10.1136/bmjph-2023-000308

Castro, E., Healy, J., Liu, A., Wei, Y., Kosheleva, A., & Schwartz, J. (2025). Interactive effects between extreme temperatures and PM2.5 on cause-specific mortality in thirteen U.S. states. *Environmental Research Letters* , *20*(1), 14011. https://doi.org/10.1088/1748-9326/ad97d1

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623

Fauzianto, F. S., & Ali, M. (2024). Analisis Komparatif Pemantauan Kualitas Udara Ambien di Surabaya Pada

Tahun 2023. *Ocean Engineering : Jurnal Ilmu Teknik Dan Teknologi Maritim*, *3*(2), 01–13. https://doi.org/10.58192/ocean.v3i2.2085

Garmini, R., & Purwana, R. (2020). Polusi Udara Dalam Rumah Terhadap Infeksi Saluran Pernafasan Akut pada Balita di TPA Sukawinatan Palembang. *Jurnal Kesehatan Lingkungan Indonesia*, *19*(1), 1. https://doi.org/10.14710/jkli.19.1.1-6

Inayah, N. (2025). Hubungan polutan udara ambien dengan kejadian ISPA pada balita. *Jurnal Kesehatan Tambusai*, *6*(2), 59–64. https://doi.org/10.31004/jkt.v6i2.44236

Krishnan, R., Thiruvengadam, K., Jayabal, L., Selvaraju, S., Watson, B., Malaisamy, M., Nagarajan, K., Tripathy, S. P., Chinnaiyan, P., & Chandrasekaran, P. (2022). An influence of dew point temperature on the occurrence of Mycobacterium tuberculosis disease in Chennai, India. *Scientific Reports*, *12*(1), 1–10. https://doi.org/10.1038/s41598-022-10111-4

Lev, A. (2022). *XGBoost versus Random Forest*. Qwak's Blog. https://www.qwak.com/post/xgboost-versus-random-forest

Márovics, G., Pozsgai, É., Németh, B., Czigány, S., Soós, S., Németh-Simon, S., & Girán, J. (2024). Weather Variability and COPD: A Risk Estimation Identified a Vulnerable Sub-population in Hungary. *In Vivo*, *38*(4), 1690–1697. https://doi.org/10.21873/invivo.13619

Munggaran, G. A., Kusnoputranto, H., & Ariyanto, J. (2024). Korelasi Polusi Udara dengan Insiden Pneumonia Balita di DKI Jakarta pada Tahun 2017-2020. *Jurnal Promotif Preventif*, *7*(1), 123–135. https://doi.org/10.47650/jpp.v7i1.1071

Nova, L. S., Siahainenia, H. E., & Novianti, P. (2023). Gambaran Kejadian ISPA Pada Anak Balita di Sekitar Industri Baja Menurut Jarak dan Kondisi Lingkungan. *Jurnal Bidang Ilmu Kesehatan*, *13*(1), 24–33. https://doi.org/10.52643/jbik.v13i1.2521

Ozili, P. K. (2023). The acceptable R-square in empirical modelling for social science research. *Social Research Methodology and Publishing Results: A Guide to Non-Native English Speakers*, *115769*, 134–143. https://doi.org/10.4018/978-1-6684-6859-3.ch009

Pathak, S., Kuchkorov, T., Yusupov, I., Makhkamov, B., Zaynidinov, K., & Ather, D. (2024). Urban Air Quality and Health Impact Analysis Based on Machine Learning Models. *Proceedings - International Conference on Information Science and Communications Technologies - Applications, Trends and Opportunities, ICISCT 2024*, 357–362. https://doi.org/10.1109/ICISCT64202.2024.10957146

Pramudhyta, N. A., & Rohman, M. S. (2024). Perbandingan Optimasi Metode Grid Search dan Random Search dalam Algoritma XGBoost untuk Klasifikasi Stunting. *Jurnal Media Informatika Budidarma*, *8*(1), 19. https://doi.org/10.30865/mib.v8i1.6965

Rahayuningtyas, D., Pascawati, N., Alfanan, A., & Dharmawan, R. (2025). Model prediksi kasus DBD berdasarkan perubahan iklim: Studi kohort dengan data NASA di Kabupaten Bantul. *Jurnal Kesehatan Lingkungan Indonesia*, *24*(1), 84–94. https://doi.org/10.14710/jkli.24.1.84-94

Rosatul Umah, & Eva Gusmira. (2024). Dampak Pencemaran Udara terhadap Kesehatan Masyarakat di Perkotaan. *Profit: Jurnal Manajemen, Bisnis Dan Akuntansi*, *3*(3), 103–112. https://doi.org/10.58192/profit.v3i3.2246

Sapna, A., Qolbi, L., Salsabilla, R., Salsilla, R., Anggriyani, R., & Ardi. (2024). Dampak dan pencegahan polusi udara akibat kebakaran lahan terhadap kesehatan masyarakat di Padang (Sumatera Barat). *Prosiding Seminar Nasional Biologi*, *3*(2), 912–923. https://doi.org/10.24036/prosemnasbio/vol3/797

Subaşı, N. (2024). Comprehensive Analysis of Grid and Randomized Search on Dataset Performance. *European Journal of Engineering and Applied Sciences*, *7*(2), 77–83. https://doi.org/10.55581/ejeas.1581494

World Health Organization. (2021). WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. In *World Health Organization*. WHO. https://www.who.int/publications/i/item/9789240034228

Zhou, L., Zhu, Q., Chen, Q., Wang, P., & Huang, H. (2025). Predicting hospital outpatient volume using XGBoost: a machine learning approach. *Scientific Reports*, *15*(1), 17028. https://doi.org/10.1038/s41598-025-01265-y