

Pemanfaatan Algoritma CatBoost dalam Analisis Kualitas Udara di Jakarta Berdasarkan PM2.5 dan NO

Fahmi Hammam Taqiyuddin¹, Aviolla Terza Damaliana², Andri Fauzan Adziima³

¹Fahmi Hammam Taqiyuddin (Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur)

¹22083010026@student.upnjatim.ac.id

³andri.fauzan.fasilkom@upnjatim.ac.id

²Aviolla Terza Damaliana, S.Si., M.Stat. (Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur)

Corresponding author email: ²aviolla.terza.sada@upnjatim.ac.id

ABSTRAK

Pencemaran udara di wilayah perkotaan menjadi isu serius karena peningkatan aktivitas transportasi dan industrialisasi menyebabkan tingginya konsentrasi polutan, khususnya PM2.5 dan SO₂. Kondisi ini menimbulkan masalah kesehatan dan menuntut adanya sistem pemantauan kualitas udara yang akurat. Penelitian sebelumnya menunjukkan bahwa metode berbasis machine learning, termasuk CatBoost, mampu memberikan performa prediksi yang baik, namun sebagian studi masih bergantung pada banyak variabel atau menghadapi kendala ketidakseimbangan kelas. Penelitian ini menawarkan solusi dengan membangun model klasifikasi kualitas udara menggunakan algoritma CatBoost hanya dengan dua fitur utama, yaitu PM2.5 dan SO₂, sehingga lebih sederhana namun tetap efektif. Data kualitas udara harian tahun 2021–2025 digunakan sebagai dasar pelatihan, dengan pembagian 80% data latih dan 20% data uji serta penerapan class weighting otomatis. Hasil evaluasi menunjukkan bahwa model mampu mengklasifikasikan sebagian besar kategori dengan baik, ditandai akurasi 0,95 dan weighted f1-score 0,95, meskipun kategori SANGAT TIDAK SEHAT tidak terbaca akibat minimnya data. Temuan ini mengindikasikan bahwa CatBoost merupakan pendekatan yang layak dan efisien untuk mendukung sistem pemantauan kualitas udara berbasis data di perkotaan.

Kata Kunci: kualitas udara, CatBoost, PM2.5, nitrogen monoksida, Jakarta

I. PENDAHULUAN

Pencemaran udara di kawasan perkotaan menjadi isu penting karena urbanisasi yang pesat meningkatkan tekanan terhadap lingkungan. Berbagai aktivitas seperti transportasi, industri, konstruksi, dan pemanasan domestik menghasilkan beragam polutan yang memperburuk kualitas udara. Kondisi ini menimbulkan konsekuensi serius bagi kesehatan masyarakat karena paparan jangka panjang terhadap polutan seperti PM, NO₂, dan SO₂ telah dikaitkan dengan penyakit pernapasan dan kardiovaskular. Kompleksitas sumber emisi di wilayah padat penduduk menunjukkan bahwa penurunan kualitas udara muncul dari interaksi berbagai faktor yang saling berkaitan. Oleh karena itu, pemahaman ilmiah mengenai dinamika polusi udara di kota menjadi sangat penting sebagai dasar pengembangan strategi mitigasi dan kebijakan lingkungan yang efektif [1].

Permasalahan kualitas udara di kota besar semakin kompleks karena urbanisasi yang cepat meningkatkan jumlah emisi dari aktivitas manusia. Berbagai sumber seperti transportasi, industri, dan aktivitas domestik memberikan kontribusi signifikan terhadap meningkatnya konsentrasi polutan. Polusi ini semakin parah ketika kondisi meteorologis tertentu menyebabkan polutan terperangkap di permukaan kota dan mengurangi tingkat dispersi alami. Dampaknya dapat terlihat dari risiko kesehatan yang meningkat, terutama di wilayah dengan kepadatan penduduk tinggi dan aktivitas ekonomi intensif. Oleh sebab

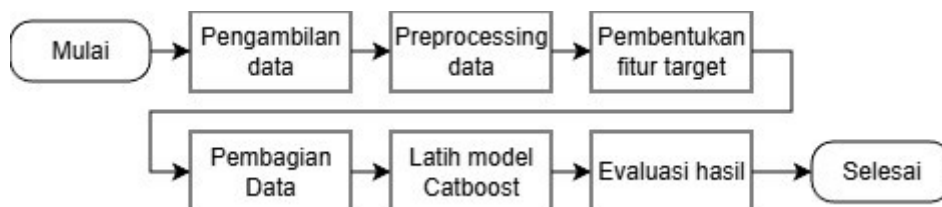
itu, masalah kualitas udara memerlukan pendekatan terpadu yang mempertimbangkan berbagai sumber pencemar, faktor lingkungan, serta dinamika sosial yang memengaruhi pembentukan polusi[2].

Penelitian sebelumnya telah menggunakan algoritma CatBoost untuk memprediksi indeks kualitas udara berdasarkan data SPKU Jakarta. Studi tersebut membangun empat model dengan variasi rasio pelatihan dan pengujian untuk melihat pengaruh komposisi data terhadap performa prediksi. Proses optimasi parameter dilakukan menggunakan GridSearchCV sehingga model dapat mencapai kombinasi parameter terbaik. Hasil pengujian menunjukkan bahwa model dengan proporsi data latih lebih besar mampu menghasilkan akurasi tertinggi karena pola data dapat dipelajari dengan lebih efektif. Temuan ini menegaskan bahwa CatBoost merupakan metode yang akurat dan andal untuk memprediksi kualitas udara di wilayah Jakarta[3].

Pemanfaatan algoritma CatBoost dalam pemantauan kualitas udara memberikan potensi peningkatan akurasi dalam mengidentifikasi kategori udara perkotaan. Model ini mampu menangani data polutan yang kompleks dan tidak lengkap sehingga lebih stabil dalam memproses variasi konsentrasi PM10, SO₂, CO, O₃, dan NO₂. Performa tinggi yang ditunjukkan oleh akurasi, presisi, dan recall pada hasil pengujian memperlihatkan bahwa metode ini layak diterapkan dalam sistem pemantauan udara berbasis data. Integrasi model CatBoost dengan visualisasi dan analisis fitur dominan seperti O₃ dan PM10 membantu perencana kota dalam memahami faktor utama yang memengaruhi kualitas udara. Dengan demikian, pendekatan ini dapat menjadi dasar pengembangan strategi pemantauan lingkungan yang lebih efektif dan adaptif di wilayah perkotaan[4].

II. METODOLOGI PENELITIAN

Sumber data dalam penelitian ini adalah data kualitas udara harian Jakarta periode 22 Januari 2021 hingga 31 Desember 2025 yang diunduh dari portal resmi Satu Data Jakarta pada laman <https://satudata.jakarta.go.id/home>. Dataset berisi 6.139 baris data dan 12 variabel, mencakup informasi stasiun pemantauan (SPKUA), tanggal pencatatan, serta parameter polutan utama seperti PM10, PM2.5, SO₂, CO, O₃, dan NO₂. Data tersebut diperlakukan sebagai data sekunder dan digunakan untuk membangun model prediksi kualitas udara dengan algoritma CatBoost. Penelitian ini bertujuan membangun model klasifikasi berbasis CatBoost untuk memprediksi kategori kualitas udara harian, sekaligus melakukan prediksi kondisi udara untuk 7 hari ke depan. Tahapan penelitian disusun mengikuti alur sistematis yang ditunjukkan pada diagram alir berikut.



Gambar 1. Diagram Alir Penelitian

Informasi mengenai tahapan penelitian pada gambar 1 diatas akan disajikan dalam penjelasan berikut:

1. Pengumpulan Data

Dataset kualitas udara yang memuat parameter PM2.5, SO₂, tanggal pencatatan, dan kategori kualitas udara diunduh dan dibaca menggunakan pandas. Data ini menjadi dasar untuk membangun model klasifikasi CatBoost.

2. Pembersihan dan Persiapan Data
Beberapa langkah pembersihan dan persiapan data dilakukan, antara lain:
 - Mengubah variabel tanggal menjadi tipe datetime.
 - Mengurutkan data berdasarkan waktu untuk keperluan peramalan.
 - Menangani nilai hilang pada data dengan diisi nilai 0
 - Memastikan format kolom kategori sudah konsisten.
3. Pembentukan Fitur
Pemilihan fitur mengacu pada dua polutan utama yang paling dominan memengaruhi kualitas udara yang dimana PM2.5 dan SO₂ Sebagai Fitur dan kategori kualitas udara sebagai Target.
4. Pembagian Data
Dataset dibagi menjadi 80% data latih dan 20% data uji, mengikuti praktik pemodelan prediksi deret waktu agar model dapat belajar pola jangka panjang.
5. Pelatihan Model
CatBoost merupakan algoritma Gradient Boosting Decision Tree (GBDT) yang dikembangkan oleh Yandex. Berbeda dengan metode boosting konvensional, CatBoost menggunakan pendekatan ordered boosting untuk mengurangi overfitting dan mampu menangani fitur kategorikal tanpa perlu dikonversi menggunakan one-hot encoding. Mekanisme ini menjadikan CatBoost lebih stabil pada dataset berukuran kecil hingga menengah serta lebih tangguh terhadap distribusi kelas yang tidak seimbang. Pendekatan ini secara luas juga digunakan dalam penelitian kualitas udara sebelumnya karena kemampuannya memberikan akurasi tinggi dalam memodelkan hubungan non-linear pada parameter polutan udara[5]. Secara konseptual, CatBoost membangun model secara bertahap dengan menambahkan pohon keputusan baru yang berfungsi memperbaiki kesalahan prediksi pada iterasi sebelumnya. Proses ini dapat dijelaskan melalui rumus pembaruan berikut:

$$F_1(x) = F_0(x) + \eta \cdot h_1(x) \quad (1)$$

yang menunjukkan bahwa model akhir terbentuk dari akumulasi perbaikan bertahap terhadap kesalahan prediksi. Nilai η mengatur seberapa besar kontribusi koreksi yang diberikan oleh pohon baru pada setiap iterasi.

Parameter Model yang Digunakan Model CatBoostClassifier dibangun menggunakan konfigurasi parameter berikut:

- iterations: 1500
 - learning_rate: 0.1
 - depth: 8
 - l2_leaf_reg: 8
 - loss_function: MultiClass
 - eval_metric: TotalF1
 - auto_class_weights: Balanced
 - random_seed: 42
6. Evaluasi Model
Setelah pelatihan, model dievaluasi menggunakan: classification report dari scikit-learn yang memuat precision, recall, f1-score, dan accuracy untuk setiap kelas serta nilai rata-ratanya, confusion matrix untuk melihat sebaran prediksi benar dan salah pada masing-masing kelas. Penggunaan kombinasi metrik ini membantu menilai performa model tidak hanya dari satu sisi, tetapi dari keseimbangan antara ketepatan prediksi (precision), kemampuan menemukan kelas yang benar (recall), serta ringkasan keseluruhan (f1-score dan accuracy), terutama pada kondisi kelas yang tidak seimbang.

III. HASIL DAN ANALISIS

Evaluasi dilakukan untuk menilai kemampuan model CatBoost dalam mengklasifikasikan kategori kualitas udara berdasarkan dua parameter polutan, yaitu

	precision	recall	f1-score	support
BAIK	0.73	0.99	0.84	144
SANGAT TIDAK SEHAT	0.00	0.00	0.00	1
SEDANG	0.99	0.94	0.97	915
TIDAK ADA DATA	0.90	1.00	0.95	9
TIDAK SEHAT	0.99	0.97	0.98	159
accuracy			0.95	1228
macro avg	0.72	0.78	0.75	1228
weighted avg	0.96	0.95	0.95	1228

Gambar 2. Classification Report Model CatBoost

PM2.5 dan SO₂. Pengujian menggunakan 20% data uji menghasilkan metrik precision, recall, f1-score, dan akurasi keseluruhan sebagaimana ditampilkan pada Gambar 2. Secara umum, model mampu mengenali pola utama kualitas udara dengan cukup baik meskipun distribusi kelas tidak seimbang. Evaluasi ini juga memberikan gambaran awal mengenai konsistensi model dalam menangani variasi nilai polutan pada setiap kategori. Selain itu, hasil perbandingan antar kelas menunjukkan bahwa model lebih responsif terhadap kategori dengan jumlah sampel yang besar, sehingga stabilitas prediksi pada kelas mayoritas dapat terjaga.

Hasil evaluasi pada gambar 2 menunjukkan bahwa kategori SEDANG, TIDAK ADA DATA, dan TIDAK SEHAT memiliki performa tinggi dengan f1-score masing-masing 0.97, 0.95, dan 0.98. Kategori BAIK juga teridentifikasi dengan baik melalui nilai recall 0.99, meskipun precision lebih rendah (0.73) karena sebagian prediksi dari kelas lain masuk ke kategori ini. Sebaliknya, kategori SANGAT TIDAK SEHAT tidak dapat diprediksi karena hanya memiliki satu sampel pada data uji, sehingga model tidak memperoleh pola pembelajaran yang cukup. Secara keseluruhan, akurasi model mencapai 0.95 dengan weighted f1-score 0.95, menandakan bahwa CatBoost cukup efektif dalam klasifikasi kualitas udara, meskipun kelas yang sangat jarang tetap menjadi tantangan utama.

IV. KESIMPULAN

Penelitian ini berhasil membangun model klasifikasi kualitas udara menggunakan algoritma CatBoost dengan memanfaatkan dua parameter polutan utama, yaitu PM2.5 dan SO₂. Proses pemodelan dilakukan melalui tahapan pengolahan data, pembentukan fitur, penanganan ketidakseimbangan kelas, serta pelatihan model menggunakan konfigurasi parameter yang telah dioptimalkan. Berdasarkan evaluasi, model menunjukkan kemampuan yang kuat dalam mengidentifikasi sebagian besar kategori kualitas udara dengan akurasi keseluruhan sebesar 0.95 dan nilai weighted f1-score sebesar 0.95. Hasil ini menegaskan bahwa CatBoost mampu mempelajari hubungan non-linear antar polutan dengan cukup baik dan memberikan prediksi yang stabil pada kategori dengan representasi data yang memadai.

Meski demikian, model masih menghadapi keterbatasan dalam mengenali kategori SANGAT TIDAK SEHAT karena jumlah sampelnya sangat sedikit. Kondisi ini

menyebabkan model tidak memiliki pola pembelajaran yang cukup untuk membentuk batas keputusan yang akurat pada kelas tersebut. Oleh karena itu, penelitian selanjutnya dapat mempertimbangkan penambahan data, teknik penyeimbangan kelas, atau pendekatan berbasis aturan untuk meningkatkan akurasi pada kategori yang jarang muncul. Secara keseluruhan, hasil penelitian ini menunjukkan bahwa CatBoost merupakan metode yang layak dan efektif untuk mendukung analisis kualitas udara berbasis data dan dapat menjadi dasar pengembangan sistem pemantauan lingkungan yang lebih adaptif di masa mendatang.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada seluruh pihak yang telah memberikan dukungan selama proses penelitian ini berlangsung. Apresiasi khusus diberikan kepada rekan-rekan yang turut membantu dalam pengumpulan data dan diskusi teknis, serta kepada pihak institusi yang menyediakan fasilitas sehingga penelitian ini dapat diselesaikan dengan baik. Tanpa dukungan tersebut, penyusunan penelitian ini tidak akan berjalan dengan optimal.

REFERENSI

- [1] T. Vyas and H. R. Varia, "Air quality analysis and modeling in urban area : A review study," 2023.
- [2] X. Li, "Visualization Study on Trends and Hotspots in the Field of Urban Air Pollution in Metropolitan Areas and Megacities : A Bibliometric Analysis via Science Mapping," 2025.
- [3] M. Arif, A. Syukur, and T. Chamidy, "Prediksi Kualitas Udara Menggunakan Metode CatBoost," vol. 10, no. 2, pp. 249–258, 2025.
- [4] G. M. Idroes, T. R. Noviandy, A. Maulana, and Z. Zahriah, "Urban Air Quality Classification Using Machine Learning Approach to Enhance Environmental Monitoring," vol. 1, no. 2, 2023, doi: 10.60084/ljes.v1i2.99.
- [5] Y. D. Prasetyo, F. N. Rahmadani, and M. Idhom, "Prediction of the Air Quality Index in DKI Jakarta Province Using the CatBoost Method," vol. 4, no. 3, pp. 1–5, 2025.