JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 10, No. 2, MEI, 2025, Pp. 249 – 258
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Prediksi Kualitas Udara Menggunakan Metode CatBoost

Mohamad Arif Abdul Syukur [(1)*], Suhartono [(2)], Totok Chamidy [(3)]
Departemen Teknik Informatika, UIN Maulana Malik Ibrahim, Malang, Indonesia
e-mail : 200605110044@student.uin-malang.ac.id, {suhartono,to2k2013}@ti.uin-malang.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 26 Juni 2024, direvisi 2 Januari 2025, diterima 8 Januari 2025, dan dipublikasikan 31 Mei 2025.

Abstract

Air is essential for life, but industrial activities, forest fires, cigarette smoke, and transportation contribute to air pollution. AirVisual AQI 2024 data ranks Jakarta in 11th place globally, with the highest level of pollution, reaching 127, which is unhealthy for sensitive groups and poses a risk of causing serious illnesses, including skin and respiratory diseases. This research uses the CatBoost method to predict the air quality index using Jakarta SPKU data taken from Kaggle. The data is processed through pre-processing and divided into four models with different comparisons of training and testing data. Each model was tested with the parameters iteration, depth, learning rate, and l2_leaf_reg, using GridSearchCV to find the optimal combination. The results show that the model with 90% training data and 10% testing data provides the best accuracy of 97%, due to the larger proportion of training data. This research demonstrates that the CatBoost method can yield accurate air quality predictions, which is crucial in supporting efforts to mitigate the impact of pollution and enhance public health.

Keywords: Prediction, Air Quality, Gradient Boosting, CatBoost, GridSearchCV

Abstrak

Udara penting bagi kehidupan, namun aktivitas industri, pembakaran hutan, asap rokok dan transportasi meningkatkan polusi udara. Data AirVisual AQI 2024 menempatkan Jakarta pada peringkat 11 dunia dengan tingkat polusi tertinggi, mencapai 127 yang tidak sehat bagi kelompok sensitif, dan berisiko menimbulkan penyakit serius seperti penyakit kulit dan pernapasan. Penelitian ini menggunakan metode CatBoost untuk memprediksi indeks kualitas udara dengan data SPKU Jakarta yang diambil dari Kaggle. Data tersebut diolah melalui pra-pemrosesan dan dibagi menjadi empat model dengan perbandingan data pelatihan dan pengujian yang berbeda. Setiap model diuji dengan parameter *iteration*, *depth*, *learning_rate*, dan *l2_leaf_reg*, menggunakan GridSearchCV untuk menemukan kombinasi terbaik. Hasilnya menunjukkan bahwa model dengan 90% data pelatihan dan 10% data pengujian memberikan akurasi terbaik sebesar 97%, karena proporsi data pelatihan yang lebih besar. Penelitian ini menunjukkan bahwa metode CatBoost dapat memberikan prediksi kualitas udara yang akurat, yang penting untuk mendukung upaya mengurangi dampak polusi dan meningkatkan kesehatan masyarakat.

Kata Kunci: Prediksi, Kualitas Udara, Gradient Boosting, CatBoost, GridSearchCV

1. PENDAHULUAN

Prediksi kualitas udara di kota Jakarta sangat penting karena dampaknya yang signifikan terhadap kesehatan masyarakat dan keadilan lingkungan (Apte et al., 2017). Variasi tingkat kualitas udara dapat mempunyai implikasi yang besar, sehingga penting untuk mengukur dan mengelola polusi udara secara efektif. Tingginya variasi tingkat polusi udara di Jakarta, khususnya polutan seperti PM10, SO2, NO2, O3, dan CO, mempunyai implikasi langsung terhadap kesehatan masyarakat dan kesejahteraan lingkungan (Chandra et al., 2022; Iqbal et al., 2025). Terbatasnya jumlah stasiun pemantauan kualitas udara di Jakarta mengharuskan pengembangan model statistik untuk mengkarakterisasi secara komprehensif distribusi polusi udara secara spasial dan temporal (Lestari et al., 2022; Syuhada et al., 2023).

Pengembangan model prediksi berdasarkan data untuk memperkirakan tingkat polusi udara di Jakarta pada tahun-tahun mendatang menandakan diperlukannya pendekatan proaktif untuk

mengelola dan meningkatkan kualitas udara di kota tersebut (Ramadhani et al., 2022). Analisis terpadu terhadap polusi udara dan kondisi meteorologi di Jakarta menunjukkan adanya variasi musiman dalam konsentrasi polutan, sehingga menunjukkan perlunya model prediktif untuk mengantisipasi perubahan kualitas udara sepanjang tahun (Handhayani, 2023). Penggunaan AI dan sistem data besar untuk prediksi kualitas udara semakin menekankan kemajuan teknologi yang diperlukan untuk memperkirakan kualitas udara secara akurat dan efisien (Jufriansah et al., 2023).

Penelitian sebelumnya telah menggunakan berbagai metode pembelajaraan secara efektif dalam memprediksi tingkat kualitas udara seperti Adaptive Boosting (AdaBoost), jaringan saraf tiruan (ANN), Random Forest, *stacking ensemble*, dan Support Vector Machine (SVM) dalam memperkirakan tingkat indeks kualitas udara (AQI) (Lei et al., 2023; Liang et al., 2020; Ravindiran et al., 2025). Akan tetapi sebagian besar penelitian menghadapi tantangan dalam menangani data yang tidak seimbang, fitur kategori yang kompleks, serta efisiensi komputasi yang rendah ketika mengelola *dataset* yang besar. Selain itu eksplorasi terhadap optimasi parameter yang kurang pada model yang digunakan dapat mengurangi potensi akurasi prediksi.

Berbagai model pembelajaran mesin telah digunakan untuk memprediksi kualitas udara dengan memasukkan parameter meteorologi. Support Vector Machine (SVM) terkenal karena fleksibilitas dan skalabilitasnya dalam prakiraan kualitas udara (Castelli et al., 2020). Selain itu, jaringan memori jangka pendek yang dioptimalkan dengan *hyperparameter* telah digunakan untuk prediksi tingkat kualitas udara, yang menunjukkan semakin besarnya adopsi metode berbasis pembelajaran mesin dalam mengatasi tantangan prediksi kualitas udara (Kim & Kim, 2015).

Pada penelitian ini, peneliti mencoba menggunakan metode CatBoost untuk mengatasi permasalahan tersebut dalam memprediksi indeks kualitas udara di wilayah Jakarta. Metode CatBoost unggul dalam menangani data kategori tanpa memerlukan pra-pemrosesan yang ekstensif dan memiliki performa yang baik pada data yang tidak seimbang serta dibandingkan dengan metode lain, CatBoost memiliki kemampuan mengurangi *overfitting* melalui algoritma *boosting* berbasis *gradient* dan pemrosesan yang lebih cepat pada *dataset* yang besar. Metode ini juga telah terbukti memberikan akurasi tinggi pada *dataset* polusi udara di India (Ramesh, 2023).

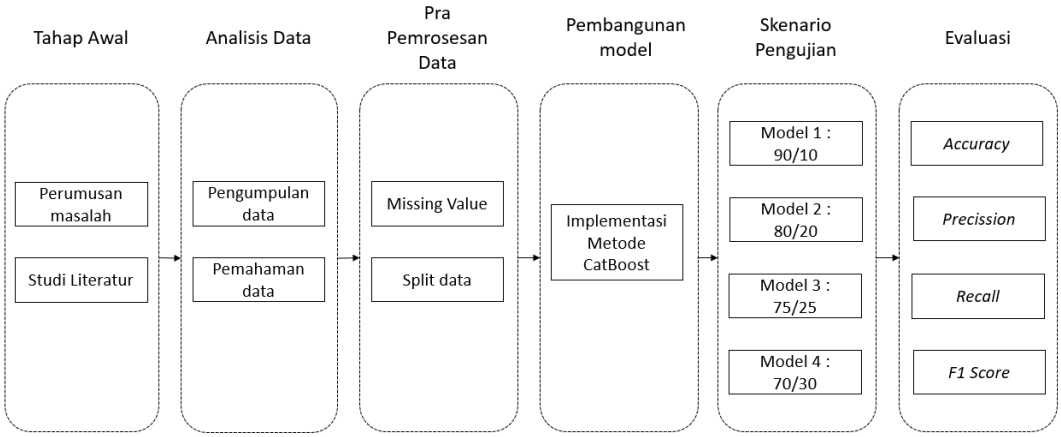
Penelitian ini menguji algoritma CatBoost dan menggabungkannya dengan teknik GridSearchCV untuk mencari parameter terbaik dan optimal. CatBoost digunakan dengan data polusi di wilayah Jakarta. Hasilnya terdapat empat model dengan pembagian data yang berbeda, model 1 dengan pembagian data latih 90% dan pengujian 10%, model 2 dengan pembagian data latih 80% dan pengujian 20%, model 3 dengan pembagian data latih 75% dan pengujian 25%, dan model 4 dengan pelatihan data latih 70% dan pengujian 30%. Data yang digunakan dalam penelitian ini adalah data indeks kualitas udara di wilayah Jakarta tahun 2020 yang diperoleh dari website penyedia data yaitu Kaggle.

Oleh karena itu, penelitian ini bertujuan untuk mengetahui kinerja metode CatBoost dalam memprediksi indeks kualitas udara berdasarkan data indeks standar polutan udara di *dataset* wilayah Jakarta yang bersumber dari Kaggle untuk meningkatkan akurasi dalam klasifikasi prediksi. Harapanya penelitian ini dapat memberikan kontribusi signifikan terhadap penggunaan metode Catboost dan teknik GridSearchCV untuk memprediksi indeks kualitas udara di wilayah Jakarta yang lebih akurat, mendukung pengambilan keputusan untuk pengelolaan lingkungan yang lebih baik dan dapat memberikan solusi yang lebih komprehensif, efektif dan efisien.

2. METODE PENELITIAN

Metode Penelitian merupakan alur proses dari awal sampai akhir yang dilakukan untuk penelitian. Agar penelitian dapat efektif, diperlukan alur sistem yang terstruktur. Dalam penelitian ini sistem dimulai dari pendekatan penelitian perumusan masalah dan dilanjutkan dengan analisis data, kemudian setelah itu pengolahan data dengan pra-pemrosesan dan pengolahan data menggunakan Jupyter Notebook dilanjutkan dengan evaluasi model dan pembahasan skenario

pengujian hingga penarikan kesimpulan. Penelitian ini memiliki kerangka diagram alir proses untuk mencapai tujuan tersebut, seperti pada Gambar 1.



Gambar 1 Desain Penelitian

2.1 Pengumpulan Data

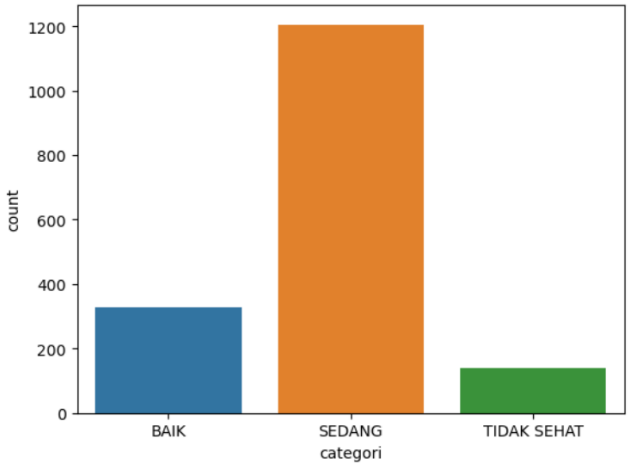
Pada tahap pengumpulan data yang dilakukan adalah mengumpulkan informasi mengenai topik kualitas udara. Data yang dikumpulkan merupakan data standar indeks polusi udara di wilayah Jakarta yang diambil dari Kaggle. Data ini dapat digunakan karena bersifat publik dan dapat diunduh dengan mudah melalui *website* Kaggle. Namun data tersebut merupakan data yang valid dan dapat dijadikan bahan pengolahan data yang kemudian dapat membantu dalam pengambilan keputusan. Sehingga data indeks standar pencemar udara di Jakarta dapat digunakan untuk penelitian tugas akhir. Datanya berjumlah 1830 dan terdiri dari 10 fitur yaitu tanggal, stasiun, pm10, so2, co, 03, no2, max, kritis, dan kategori. Tipe data setiap fitur berbeda-beda, ada yang bertipe objek, ada pula yang bertipe numerik. Data yang dikumpulkan akan digunakan sebagai masukan dalam proses pembangunan model dalam penelitian ini.

2.2 Pemahaman Data

Terdapat fitur tipe numerik yaitu fitur pm10, so2, co, o3, no2, dan max. Dan terdapat fitur tipe objek atau kategori yaitu tanggal, stasiun, kritis, dan kategori. Pada data ini terdapat tiga kategori indeks kualitas udara pada *dataset*, yaitu udara baik, sedang, dan tidak sehat. Hal ini dikarenakan pada *dataset* yang digunakan nilai tertinggi untuk indeks kualitas udara adalah 191 sehingga tidak ada kategori berbahaya di atas 300. Tabel 1 dan Gambar 2 menunjukkan deskripsi fitur dan gambar histogram analisis sebaran kategori *dataset*.

Tabel 1 Fitur *Dataset*

No.	Fitur	Keterangan
1	Tanggal	Tanggal Pengambilan Data
2	Stasiun	Lokasi Pengambilan Data
3	Pm10	Partikulat salah satu yang diukur
4	So2	Sulfida
5	Co	Karbon Monoksida
6	O3	Ozon
7	No2	Nitrogen Dioksida
8	Max	Nilai tertinggi dari semua parameter yang diukur
9	Critical	Parameter tertinggi yang diukur
10	Categori	Kategori yang dihasilkan dari perhitungan indeks standar pencemaran udara



Gambar 2 Data Fitur Categori

2.3 *Missing Values*

Pada pengolahan data *pre-processing* biasanya terdapat data yang tidak seimbang atau data yang berada di luar jangkauan. Sehingga perlu adanya perbaikan data dengan pengolahan untuk menemukan dan memperbaiki data yang tidak sesuai keakuratannya atau terdapat nilai yang hilang pada *dataset*. Dalam penelitian ini data yang tidak sesuai dan data yang tidak digunakan akan dihilangkan. Data yang akan dihapus adalah fitur “tanggal”, “stasiun”, dan “kritis”. Hal ini dilakukan karena fitur tersebut merupakan tipe objek atau tipe kategori dan pengaruhnya terhadap indeks kualitas udara juga kecil, bahkan hampir tidak ada.

2.4 *Split Data*

Pemisahan data dalam pengolahan data sangat penting dalam penelitian ini dengan tujuan untuk membagi data menjadi dua bagian yaitu data latih dan data uji. Data latih digunakan sebagai data untuk membangun suatu model, sedangkan data uji digunakan untuk menguji model yang telah dibuat dan juga untuk mengevaluasi kinerja model. Untuk pembagian data split pada penelitian ini dibagi menjadi 4 model yaitu model 1 dengan split 90:10 artinya 90% data latih dan 10% data uji. Model 2 dengan pembagian 80:20, artinya 80% data pelatihan dan 20% data pengujian. Model 3 dengan pembagian 75:25, artinya 75% data latih dan 25% data uji. Dan model 4 adalah pembagian 70:30, artinya 70% data pelatihan dan 30% data pengujian. Pembagian ini dapat memberikan performa yang baik untuk evaluasi keempat model dan juga dapat mencegah terjadinya *overfitting*. Selain itu juga dapat memberikan hasil kinerja yang akurat.

Dalam proses pembentukan data latih, penelitian ini menggunakan setiap komposisi model yang telah ditentukan dari keseluruhan jumlah data yang digunakan. Berdasarkan data pelatihan yang diperoleh dapat digunakan untuk melatih model CatBoost. Sedangkan data uji dengan komposisi masing-masing model yang ada merupakan data sisa dari keseluruhan data yang telah digunakan untuk data latih. Jadi data yang digunakan untuk data latih tidak digunakan untuk data uji. Dengan cara ini, data pengujian dapat digunakan untuk menguji model CatBoost dengan tingkat akurasi terbaik. Dengan keempat model yang telah ditentukan maka akan diambil akurasi terbaik untuk proses pengujian dan model tersebut akan digunakan dalam memprediksi indeks kualitas udara.

2.5 Implementasi CatBoost

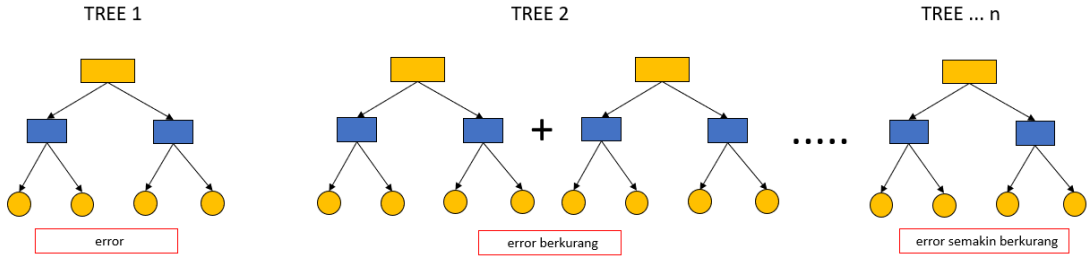
Dalam pengolahan data ada langkah yang sangat penting yang harus dilakukan yaitu pengembangan model. Pengembangan model pada penelitian ini di mana algoritma CatBoost mempelajari data latih yang telah ditentukan pada proses sebelumnya. Model pelatihan kerja

adalah menyesuaikan bobot dan bias terbaik pada algoritma dengan tujuan mengurangi fungsi kerugian dari fungsi kerugian yang ada dalam rentang nilai prediksi.

Pembuatan model menerapkan gradien yang setiap contoh dihitung berdasarkan contoh sebelumnya. Gradien yang sesuai dengan prediksi pertama dihitung menggunakan persamaan pada Pers. (1) (Dewi, 2021). Ketika pohon baru ditambahkan ke ansambel, setiap skor dihitung dari jumlah kandidat yang terpisah. Dalam hal ini gradien diperkirakan menggunakan kesamaan kosinus (*cosine similarity*). Dari fungsi skor tersebut perlu dilakukan pemilihan pohon di antara calon pohon yang ada. Misalnya, diberikan kandidat T_C, maka nilai fungsi skor kesamaan kosinus dapat dihitung menggunakan rumus pada Pers. (2) (Dewi, 2021). Visualisasi dari proses pembentukan model menggunakan algoritma CatBoost ditunjukkan pada Gambar 3.

$$grad_{r,j}(i) = \frac{\partial L(y_i, s)}{\partial s}|_{s=Mr,j\ (i)} \quad (1)$$

$$Cosine = \frac{\sum_{i=1}^{n} w_i \,.\, \Delta_i \,.\, g_i}{\sqrt{\sum_{i=1}^{n} w_i \Delta_i^2} \,.\, \sqrt{\sum_{i=1}^{n} w_i g_i^2}} \quad (2)$$



Gambar 3 Visualisasi CatBoost

2.6 Skenario Pengujian

Dalam penelitian ini, peneliti membagi data menjadi empat model seperti yang ditunjukkan pada Tabel 2. Model 1 merupakan model dengan rasio 90:10 yaitu 90% untuk data latih dan 10% untuk data uji. Model 2 merupakan model dengan rasio 80:20 yaitu 80% untuk data latih dan 20% untuk data uji. Model 3 merupakan model perbandingan 75:25 yaitu 75% untuk data latih dan 25% untuk data uji. Model 4 perbandingannya 70:30 yaitu 70% untuk data latih dan 30% untuk data uji. Dengan perbandingan yang berbeda maka akan ditemukan prediksi yang lebih akurat.

Tabel 2 Skenario Pengujian

No.	Model	*Training*	*Testing*
1	Model 1	90%	10%
2	Model 2	80%	20%
3	Model 3	75%	25%
4	Model 4	70%	30%

2.7 Evaluasi

Pada tahap evaluasi model dalam penelitian ini digunakan matriks konfusi (*confusion matrix*) sebagai dasar perhitungan metrik evaluasi seperti akurasi, presisi, *recall*, dan skor F1. Akurasi adalah sejauh mana nilai prediksi yang dihasilkan model mendekati nilai sebenarnya pada data dengan mengetahui jumlah data klasifikasi yang benar, dengan rumus perhitungannya disajikan pada Pers (3). Presisi adalah perbandingan nilai-nilai relevan berdasarkan seluruh nilai yang dipilih dengan membandingkan jumlah informasi relevan dengan jumlah total informasi yang dipilih dengan rumus perhitungannya pada Pers. (4).

Recall merupakan pemilihan rasio nilai relevan berdasarkan jumlah nilai relevan yang tersedia dengan cara membandingkan jumlah informasi relevan dengan jumlah total informasi relevan dalam informasi tersebut dengan rumus perhitungannya pada Pers. (5). Nilai presisi dan *recall* dihitung untuk mengevaluasi keakuratan dan cakupan model dalam memprediksi kategori tertentu. Selanjutnya, skor F1 sebagai rata-rata harmonis dari nilai presisi dan *recall* untuk mendapatkan gambaran seimbang antara kedua metrik tersebut, yang dijelankan pada Pers. (6) (Amalia et al., 2022).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}\ x\ 100\% \quad (3)$$

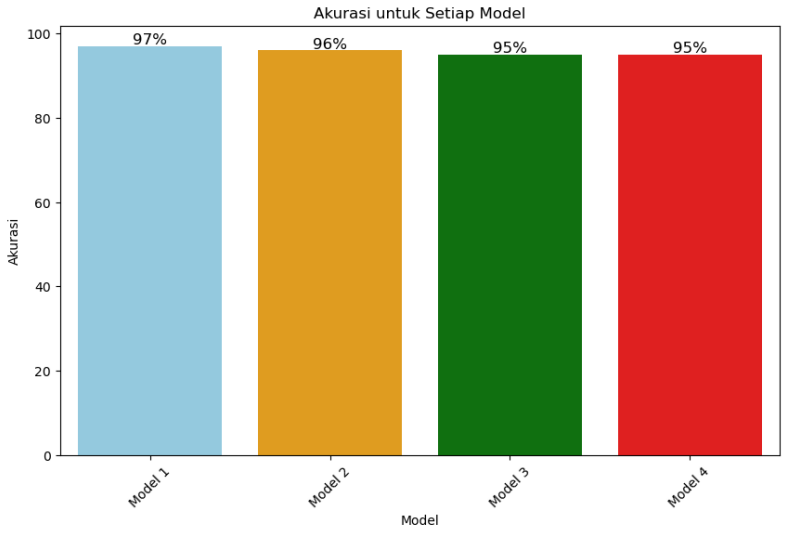
$$Precission = \frac{TP}{TP + FP}\ x\ 100\% \quad (4)$$

$$Recall = \frac{TP}{TP + FN}\ x\ 100\% \quad (5)$$

$$F1\ score = 2\ \frac{precission\ x\ recall}{precission + recall}\ x\ 100\% \quad (6)$$

3. HASIL DAN PEMBAHASAN

Setiap model diuji menggunakan beberapa kombinasi parameter *depth*, *learning_rate*, *iteration*, dan *l2_leaf_reg*. Sehingga dalam pengujian semua model menggunakan nilai parameter yang sama. Nilai parameter *depth* yang digunakan adalah 6, 8, dan 10. Nilai parameter *learning_rate* yang digunakan adalah 0,1 dan 0,01. Nilai parameter iteration yang digunakan adalah 500, 1000, dan 1500. Sedangkan nilai parameter *l2_leaf_reg* yang digunakan adalah 1, 2, dan 3. Pada proses pengujian ini, pencarian parameter optimal dilakukan dengan menggunakan GridSearchCV. Sehingga mencari kombinasi parameter terbaik yang akan digunakan untuk model klasifikasi. Selanjutnya akan dievaluasi hasil masing-masing model.

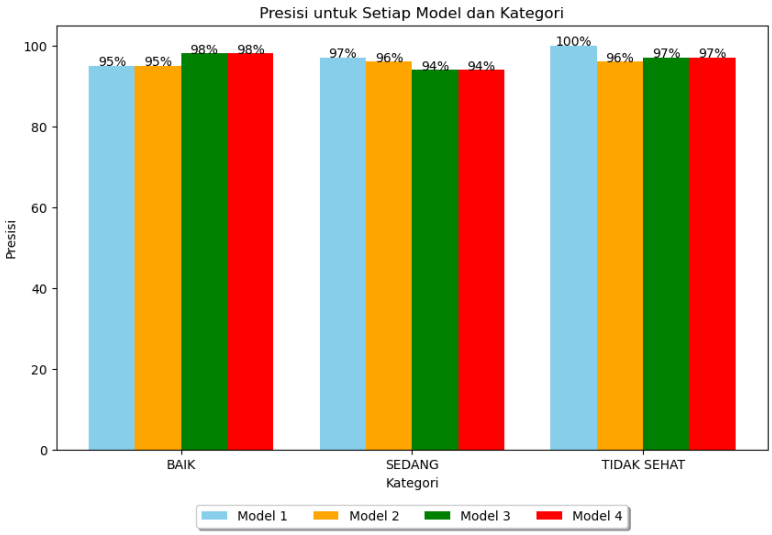


Gambar 4 Hasil Akurasi

Pada Gambar 4, Model 1 mendapatkan nilai akurasi yang lebih tinggi dibandingkan ketiga model lainnya. Hal ini dikarenakan oleh pembagian data antara data latih dan data uji. Model 1 menggunakan proporsi data latih sebesar 90% dan data uji 10%, sehingga model memiliki lebih banyak data untuk proses pelatihannya dan mampu mempelajari pola dengan lebih baik. Sementara, jumlah data uji yang lebih sedikit memungkinkan sistem melakukan pengujian secara

optimal karena model telah cukup terlatih. Model 1 terbukti menjadi model terbaik dibandingkan ketiga model lainnya berdasarkan hasil evaluasi akurasi. Pembagian data latih dan data uji sangat penting karena sangat mempengaruhi seberapa baik model dalam mengenali pola dari data baru (Okprana & Winanjaya, 2022). Dengan pembagian ini, dapat dilihat seberapa baik model dalam memprediksi data yang sebelumnya belum pernah terlihat.

Evaluasi yang dilakukan secara presisi memudahkan keakuratan model dalam mendeteksi data positif. Dengan demikian, jika nilai presisi meningkat maka jumlah kesalahan dalam memprediksi data positif akan berkurang karena presisi memberikan pemahaman mengenai keakuratan model dalam memprediksi data positif (Nainggolan & Sinaga, 2023). Hasil nilai presisi masing-masing model ditunjukkan pada Gambar 5. Dari kombinasi parameter optimal yang ditemukan yaitu pada parameter depth=6, iteration=1500, l2_leaf_reg=1, dan learning_rate=0.1 dengan hasil mean_test_score *cosine similirity* sebesar 0,958870. Seperti pada Gambar 5, model 1 memiliki nilai presisi yang paling tinggi dibandingkan dengan model lainnya. Walaupun nilai presisi pada kategori “baik” hanya mencapai 95%, namun pada kategori “sedang” mencapai 97% dan “tidak sehat” memperoleh presisi sempurna 100%.

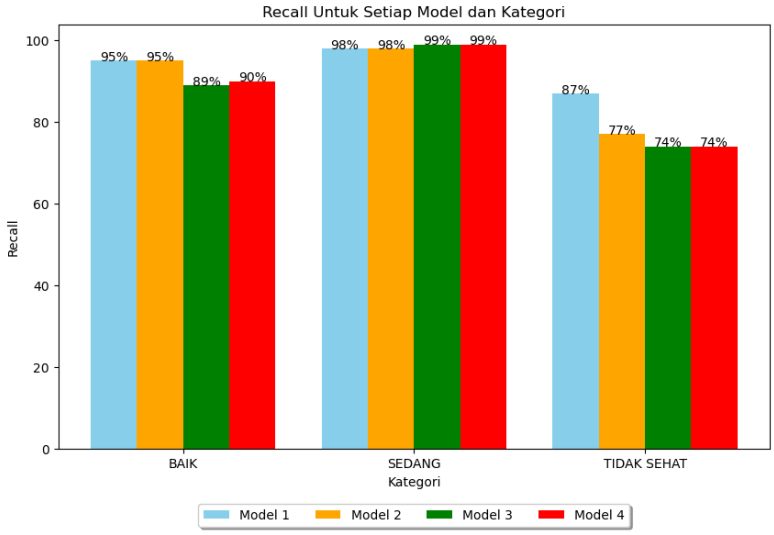


Gambar 5 Hasil Presisi

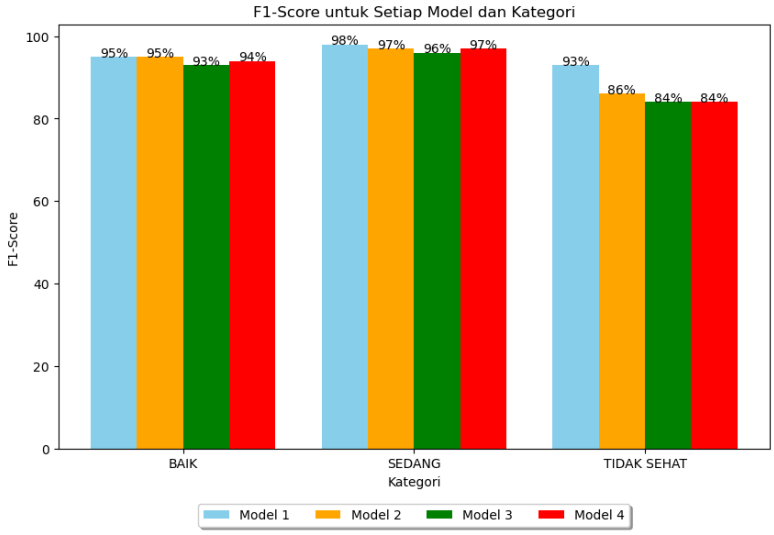
Berikutnya adalah *recall* yang merupakan perbandingan antara prediksi *True Positives* (TP) dengan seluruh data *true* positif. Sehingga evaluasi yang dilakukan dengan *recall* memberikan pemahaman seberapa sering model memprediksi positif padahal data sebenarnya positif (Saputro & Sari, 2020). Hasil nilai *recall* masing-masing model ditunjukkan pada Gambar 6. Semua model mendapatkan nilai *recall* yang tinggi sehingga dapat dikatakan banyak data positif yang diprediksi benar oleh model. Model 1 memperoleh nilai *recall* tertinggi pada kategori “tidak sehat” sebesar 87%, meskipun pada kategori “baik”, nilai *recall* model 1 dan model 2 bernilai sama yaitu 95%, dan pada kategori “sedang”, nilai *recall* model 1 sedikit lebih rendah dibandingkan dengan model 3 dan 4, yaitu sebesar 98%.

Selanjutnya evaluasi dilakukan menggunakan metrik *F1-score* yang merupakan nilai rata-rata harmonik tingkat presisi dan *recall* (Baharuddin et al., 2019). *F1-score* merupakan integrasi antara nilai presisi dan recall yang menjadi satu nilai sebagai kualitas model prediksi secara keseluruhan. Sehingga model dapat dinilai apakah akurat atau tidak dalam memprediksi data positif dengan benar. Hasil nilai *F1-score* masing-masing model ditunjukkan pada Gambar 7. Nilai *F1-score* pada semua model mencapai hasil yang baik. Sehingga dapat dikatakan bahwa semua model memiliki nilai keseimbangan untuk memprediksi sebagian besar kategori yang tepat (*recall)* dan menemukan prediksi yang akurat (*precision*). Nilai *F1-score* model 1 pada kategori “baik” hasilnya sama dengan model 2 yaitu 95%. Nilai *F1-score* tertinggi dihasilkan oleh model 1

pada kategori “sedang” sebesar 98%, nilai tersebut tertinggi dibandingkan ketiga model lainnya. Begitupun pada kategori “tidak sehat”, nilai pada model 1 juga menghasilkan nilai *F1-score* tertinggi dibandingkan dengan ketiga model lainnya, yaitu 93%.



Gambar 6 Hasil *Recall*



Gambar 7 Hasil *F1-Score*

4. KESIMPULAN

Berdasarkan skenario pengujian yang dilakukan pada penelitian ini, model dibagi menjadi 4 bagian, dengan perbandingan data latih dan data uji sebesar 90:10 pada model 1, 80:20 pada model 2, 75:25 pada model 3, serta 70:30 pada model 4. Model 1 menghasilkan nilai akurasi paling tinggi sehingga dapat dikatakan model 1 merupakan model terbaik dibandingkan model lainnya dalam memprediksi indeks kualitas udara di wilayah Jakarta. Data indeks standar pencemar udara di wilayah Jakarta diambil dari Kaggle dengan pemodelan menggunakan metode CatBoost.

Fungsi GridSearchCV diterapkan untuk mendapatkan kombinasi parameter optimal dari beberapa parameter yang ditentukan. Kombinasi parameter optimal yang didapat adalah pada

depth=6, iteration=1500, l2_leaf_reg=1, dan learning_rate=0.1 dengan hasil mean_test_score *cosine similirity* sebesar 0,958870. Model 1 menunjukkan performa yang sangat baik dengan nilai akurasi sebesar 97%. Pada evaluasi berdasarkan kategori, nilai presisi pada kategori "baik" mencapai 95%, kategori "sedang" 97%, dan kategori "tidak sehat" 100%. Sementara itu, nilai *recall* pada kategori "baik" mencapai 95%, kategori "sedang" 98%, dan kategori "tidak sehat" 87%. Selain itu, nilai *F1-score* yang merupakan kombinasi dari presisi dan *recall* menunjukkan hasil pada kategori "baik" sebesar 95%, kategori "sedang" 98%, dan kategori "tidak sehat" 93%.

Dapat disimpulkan bahwa pada penelitian ini, metode CatBoost dengan penerapan GridSearchCV dapat meningkatkan nilai akurasi pada proses prediksi indeks kualitas udara. Model terbaik diperoleh pada scenario pembagian data sebesar 90% untuk data latih dan 10% untuk data uji, dengan hasil prediksi masuk dalam kategori sangat baik. Pembagian data tersebut memberikan keuntungan karena jumlah data latih yang besar memungkinkan model untuk belajar secara optimal, sedangkan jumlah data uji yang lebih sedikit memudahkan sistem dalam melakukan evaluasi terhadap data yang belum pernah dilihat sebelumnya.

Penelitian ini menunjukkan kontribusi dalam pengembangan metode prediksi indeks kualitas udara dengan mengkombinasikan metode CatBoost dan GridSearchCV serta variasi parameter yang berbeda. Tentunya perlu adanya pengembangan untuk penelitian lebih lanjut mengenai penggunaan metode CatBoost atau penggunaan GridSearchCV dalam model prediksi atau klasifikasi sehingga dapat meningkatkan nilai akurasi yang ada. Selain itu, pengujian terhadap *dataset* yang lebih besar dan kompleks, serta perbandingan dengan algoritma pembelajaran mesin lainnya seperti XGBoost atau LightGBM, juga dapat memberikan wawasan tambahan mengenai keunggulan relatif dari pendekatan ini.

DAFTAR PUSTAKA

Amalia, A., Zaidiah, A., & Isnainiyah, I. N. (2022). Prediksi Kualitas Udara Menggunakan Algoritma K-Nearest Neighbor. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, *7*(2), 496–507. https://doi.org/10.29100/jipi.v7i2.2843

Apte, J. S., Messier, K. P., Gani, S., Brauer, M., Kirchstetter, T. W., Lunden, M. M., Marshall, J. D., Portier, C. J., Vermeulen, R. C. H., & Hamburg, S. P. (2017). High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data. *Environmental Science & Technology*, *51*(12), 6999–7008. https://doi.org/10.1021/acs.est.7b00891

Baharuddin, M. M., Azis, H., & Hasanuddin, T. (2019). Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca. *ILKOM Jurnal Ilmiah*, *11*(3), 269–274. https://doi.org/10.33096/ilkom.v11i3.489.269-274

Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A Machine Learning Approach to Predict Air Quality in California. *Complexity*, *2020*, 1–23. https://doi.org/10.1155/2020/8049504

Chandra, W., Resti, Y., & Suprihatin, B. (2022). Implementation of a Breakpoint Halfway Discretization to Predict Jakarta's Air Quality. *Inovasi Matematika (Inomatika)*, *4*(1), 1–10. https://doi.org/10.35438/inomatika.v4i1.310

Dewi, N. K. (2021). *Deteksi Fake Follower Instagram Menggunakan Catboost Classifer* [UIN Syarif Hidayatullah]. https://repository.uinjkt.ac.id/dspace/handle/123456789/56737

Handhayani, T. (2023). An Integrated Analysis of Air Pollution and Meteorological Conditions in Jakarta. *Scientific Reports*, *13*(1), Article ID: 5798. https://doi.org/10.1038/s41598-023-32817-9

Iqbal, M., Susilo, B., & Hizbaron, D. R. (2025). How Local Pollution and Transboundary Air Pollution Impact Air Quality in Jakarta? *Papers in Applied Geography*, *11*(1), 49–62. https://doi.org/10.1080/23754931.2024.2399626

Jufriansah, A., Khusnani, A., Pramudya, Y., Sya'bania, N., Leto, K. T., Hikmatiar, H., & Saputra, S. (2023). AI Big Data System to Predict Air Quality for Environmental Toxicology Monitoring. *Journal of Novel Engineering Science and Technology*, *2*(01), 21–25. https://doi.org/10.56741/jnest.v2i01.314

Kim, D. J., & Kim, J. Y. (2015). Generation Technique of Dynamic Monster's Behavior Pattern Based on User's Behavior Pattern Using FuSM. *Journal of Next-Generation Convergence*

Information Services Technology, *1*(1), 9–18. https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002141142

Lei, T. M. T., Ng, S. C. W., & Siu, S. W. I. (2023). Application of ANN, XGBoost, and Other ML Methods to Forecast Air Quality in Macau. *Sustainability*, *15*(6), Article ID: 5341. https://doi.org/10.3390/su15065341

Lestari, P., Arrohman, M. K., Damayanti, S., & Klimont, Z. (2022). Emissions and Spatial Distribution of Air Pollutants from Anthropogenic Sources in Jakarta. *Atmospheric Pollution Research*, *13*(9), Article ID: 101521. https://doi.org/10.1016/j.apr.2022.101521

Liang, Y. C., Maimury, Y., Chen, A. H. L., & Juarez, J. R. C. (2020). Machine Learning-Based Prediction of Air Quality. *Applied Sciences*, *10*(24), Article ID: 9151. https://doi.org/10.3390/app10249151

Nainggolan, S. P., & Sinaga, A. (2023). Comparative Analysis of Accuracy of Random Forest and Gradient Boosting Classifier Algorithm for Diabetes Classification. *Sebatik*, *27*(1), 97–102. https://doi.org/10.46984/sebatik.v27i1.2157

Okprana, H., & Winanjaya, R. (2022). Analisis Pengaruh Komposisi Data Training dan Testing Terhadap Akurasi Algoritma Resilient Backpropagation (RProp). *BRAHMANA: Jurnal Penerapan Kecerdasan Buatan*, *4*(1), 89–95. https://doi.org/10.30645/brahmana.v4i1.138

Ramadhani, R. F., Prasetiyowati, S. S., & Sibaroni, Y. (2022). Performance Analysis of Air Pollution Classification Prediction Map with Decision Tree and ANN. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 536–543. https://doi.org/10.47065/josyc.v3i4.2117

Ramesh, L. (2023). Prediction of Air Pollution and an Air Quality Index Using Machine Learning Techniques. *International Journal of Advanced Research in Computer Science*, *14*(02), 51–55. https://doi.org/10.26483/ijarcs.v14i2.6972

Ravindiran, G., Karthick, K., Rajamanickam, S., Datta, D., Das, B., Shyamala, G., Hayder, G., & Maria, A. (2025). Ensemble Stacking of Machine Learning Models for Air Quality Prediction for Hyderabad City in India. *IScience*, *28*(2), Article ID: 111894. https://doi.org/10.1016/j.isci.2025.111894

Saputro, I. W., & Sari, B. W. (2020). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Creative Information Technology Journal*, *6*(1), 1–11. https://doi.org/10.24076/citec.2019v6i1.178

Syuhada, G., Akbar, A., Hardiawan, D., Pun, V., Darmawan, A., Heryati, S. H. A., Siregar, A. Y. M., Kusuma, R. R., Driejana, R., Ingole, V., Kass, D., & Mehta, S. (2023). Impacts of Air Pollution on Health and Cost of Illness in Jakarta, Indonesia. *International Journal of Environmental Research and Public Health*, *20*(4), Article ID: 2916. https://doi.org/10.3390/ijerph20042916