ANALISIS PERFORMA METODE RANDOM FOREST DAN CATBOOST DALAM PEMODELAN KUALITAS UDARA KOTA PALEMBANG

THESIS

Oleh :
NURCHAERANI KADIR
NIM. 220605210018

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2024

ANALISIS PERFORMA METODE RANDOM FOREST DAN CATBOOST DALAM PEMODELAN KUALITAS UDARA KOTA PALEMBANG

THESIS

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh :
NURCHAERANI KADIR
NIM. 220605210018

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2024

ANALISIS PERFORMA METODE RANDOM FOREST DAN CATBOOST DALAM PEMODELAN KUALITAS UDARA KOTA PALEMBANG

THESIS

Oleh :
NURCHAERANI KADIR
NIM. 220605210018

Telah diperiksa dan disetujui untuk diuji
Tanggal : 18 Mei 2024

Pembimbing I	Pembimbing II
Dr. M. Faisal, M.T	Dr. Fachrul Kurniawan, M. MT., IPM
NIP. 19740510 200501 1 007	NIP. 19771020 200912 1 001

Mengetahui,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian, M.Cs
NIP. 19740424 200901 1 008

ANALISIS PERFORMA METODE RANDOM FOREST DAN CATBOOST DALAM PEMODELAN KUALITAS UDARA KOTA PALEMBANG

THESIS

Oleh :
NURCHAERANI KADIR
NIM. 220605210018

Telah dipertahankan di Depan Dewan Penguji
Dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Magister Komputer (M. Kom)
Pada Tanggal 18 Mei 2024

Susunan Dewan Penguji		**Tanda Tangan**
Penguji I	: Dr. Usman Pagalay, M.Si NIP. 19650414 200312 1 001	( )
Penguji II	: Dr. Totok Chamidy, M. Kom NIP. 19691222 200604 1 001	( )
Pembimbing I	: Dr. Muhammad Faisal, M.T NIP. 19740510 200501 1 007	( )
Pembimbing II	: Dr. Fachrul Kurniawan, M. MT., IPM NIP. 19771020 200912 1 001	( )

Mengetahui,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian
NIP. 19740424 200901 1 008

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini :

Nama	: NURCHAERANI KADIR
Nim	: 220605210018
Program Studi	: Magister Informatika
Fakultas	: Sains dan Teknologi

Menyatakan dengan sebenarnya bahwa Thesis yang saya tulis ini benar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui sebagai tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan Thesis ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 18 Juni 2024

Yang Membuat Pernyataan

NURCHAERANI KADIR

NIM. 220605210018

PERSEMBAHAN

Puji syukur kehadirat Allah, shalawat serta salam bagi Rasul-Nya
Penulis persembahkan sebuah karya ini kepada:

Kedua orang tua penulis tercinta, Bapak Drs. ABD. Kadir dan Ibu Nurlia, S. Pd, Kakak laki-laki Muh. Ichsan beserta istri Yurnaningsih, Kakak perempuan Nur Ifah, serta Adik laki-laki Muh. Fauzan dan Husnudzan yang senantiasa memberikan do'a dan semangat.

Seluruh Civitas Akademik Magister Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang yang telah membimbing dan memberikan ilmunya yang sangat bermanfaat.

Seluruh rekan-rekan mahasiswa Magister Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang yang telah membantu dan memberikan dukungan, semangat, dan motivasi selama ini.

Serta rekan-rekan yang tak bisa penulis sebutkan satu per satu yang selalu memberikan semangat dan motivasinya kepada penulis untuk menyelesaikan Thesis ini.

KATA PENGANTAR

Bismillahirrahmanirrahim

Segala puji dan syukur penulis panjatkan ke hadirat Allah subhanahu wa ta'ala yang telah melimpahkan rahmat dan hidayahNya kepada penulis, sehingga penulis bisa menyelesaikan Thesis dengan judul "**Analisis Performa Metode *Random Forest* Dan *Catboost* Dalam Pemodelan Kualitas Udara Kota Palembang**".

Shalawat serta salam semoga tercurahkan kepada baginda Rasulullah Muhammad SAW, nabi yang telah menuntun umatnya dari alam yang gelap gulita menuju alam yang terang benderang. Tujuan dari penyusunan Thesis ini guna memenuhi salah satu syarat untuk menyelesaikan studi di Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Didalam pengerjaan Thesis ini telah melibatkan banyak pihak yang sangat membantu dalam banyak hal. Oleh sebab itu, dengan segala kerendahan hati, penulis sampaikan rasa terima kasih sedalam-dalamnya kepada:

1. Bapak Dr. Cahyo Crysdian, selaku Ketua Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Bapak Dr. Muhammad Faisal, M.T dan Bapak Dr. Fachrul Kurniawan, M. MT., IPM, selaku dosen pembimbing yang telah membimbing dalam penyusunan Thesis ini hingga selesai.
3. Seluruh Civitas Akademik Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim

Malang yang telah memberikan bimbingan, arahan, ilmu, serta wawasan bagi penulis.

4. Keluarga tercinta yang telah banyak memberikan doa dan dukungan kepada penulis secara moril maupun materil hingga Thesis ini dapat terselesaikan.
5. Semua pihak yang telah banyak membantu dalam penyusunan Thesis ini yang tidak bisa penulis sebutkan semuanya.

Penulis menyadari bahwa dalam penyusunan Thesis ini masih terdapat kekurangan dan penulis berharap semoga Thesis ini bisa memberikan manfaat kepada para pembaca khususnya bagi penulis secara pribadi.

Malang, 18 Juni 2024

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Kadir, Nurchaerani. 2024. **Analisis Performa Metode Random Forest Dan Catboost Dalam Pemodelan Kualitas Udara Kota Palembang.** Thesis Program Studi Magister Informatika Fakultas Sains Dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing : (I) Dr. Muhammad Faisal, MT., (II) Dr. Fachrul Kurniawan, M. MT., IPM

Kata Kunci : Kualitas udara, Random Forest, CatBoost

Polusi udara telah menjadi salah satu tantangan lingkungan terbesar yang dihadapi oleh banyak kota di seluruh dunia, mengancam kesehatan warga dan kelestarian lingkungan. Pemodelan kualitas udara yang akurat menjadi sangat penting dalam upaya mitigasi polusi udara di wilayah perkotaan, khususnya Kota Palembang. Penelitian ini berfokus pada membandingkan performa dua metode machine learning populer, yaitu Random Forest dan CatBoost dalam pemodelan kualitas udara di Kota Palembang. Data yang digunakan meliputi konsentrasi polutan udara seperti PM2.5, PM10, SO2, NO2, CO, dan O3, serta data meteorologi seperti temperature, humadity, wind speed, wind gust, dan wind direction. Kedua metode diimplementasikan dengan optimasi parameter untuk membangun model prediksi. Evaluasi model dilakukan menggunakan metrik akurasi dan Root Mean Squared Error (RMSE). Hasil menunjukkan bahwa CatBoost memiliki performa sedikit lebih baik dibandingkan Random Forest, dengan akurasi yang lebih tinggi mencapai 96.28% berbanding 94,88% untuk Random Forest, serta RMSE yang lebih rendah yaitu 0.56348 berbanding 0.66199 untuk Random Forest. Meskipun demikian, perbedaan performa tidak terlalu signifikan, dan keduanya menunjukkan performa yang cukup baik secara keseluruhan.

ABSTRACT

Kadir, Nurchaerani. 2024. **Performance Analysis of Random Forest and CatBoost Methods in Air Quality Modeling for Palembang City**.Thesis of Master's Degree Program in Informatics, Faculty of Science and Technology, Islamic State University Maulana Malik Ibrahim Malang. Supervisor: (I) Dr. Muhammad Faisal, MT., (II) Dr. Fachrul Kurniawan, M. MT., IPM

Kata Kunci : Air Quality, Random Forest, CatBoost

Air pollution has become one of the biggest environmental challenges faced by many cities around the world, threatening citizens' health and environmental sustainability. Accurate modeling of air quality predictions has become crucial in efforts to mitigate air pollution in urban areas, especially in Palembang City. The study focuses on comparing the performance of two popular machine learning methods, Random Forest and CatBoost, in modeling air quality forecasts in Palembang City. The data used included concentrations of air pollutants such as PM2.5, PM10, SO2, NO2, CO, and O3, as well as meteorological data such as temperature, humidity, wind speed, wind gust, and wind direction. Both methods are implemented with parameter optimization to build predictive models. The evaluation of the model was done using accuracy metrics and root mean squared error (RMSE). The results showed that CatBoost performed slightly better than Random Forest, with a higher precision of 96.28% versus 94,88% for Random forest, and a lower RMSE of 0.56348 compared to 0.66199 for random forest. Nevertheless, the performance differences are not very significant, and both indicate fairly good overall performance.

الملخص

قدير، نور خير اني. ٢٠٢٤ **.تحليل أداء طريقة Random Forest و CatBoost في نمذجة تنبؤ جودة الهواء في مدينة باليمبانج .**رسالة ماجستير في برنامج دراسة علوم الحاسوب، كلية العلوم والتكنولوجيا، جامعة الإسلامية الحكومية مولانا مالك إبراهيم مالانج . المشرف : (١) الدكتور محمد فيصل ، الماجستير. **(٢)** الدكتور فخر الكرنيوان، ماجستير في التكنولوجيا(.M. MT) ، مهندس محترف معتمد.(IPM)

كلمات البحث : جودة الهواء,CatBoost، Random Forest

تعتبر تلوث الهواء أصبح أحد أكبر التحديات البيئية التي تواجهها العديد من المدن في جميع أنحاء العالم، مهدداً صحة السكان واستدامة البيئة. يصبح نمذجة التنبؤ بجودة الهواء بدقة أمرًا بالغ الأهمية في جهود التخفيف من تلوث الهواء في المناطق الحضرية، خاصة في مدينة باليمبانغ. يركز هذا البحث على مقارنة في نمذجة التنبؤ بجودة الهواء CatBoost أداء اثنتين من طرق تعلم الآلة الشائعة، وهما الغابة العشوائية و ، SO2، PM10، PM2.5 في مدينة باليمبانغ. البيانات المستخدمة تشمل تركيز الملوثات الهوائية مثل ، بالإضافة إلى البيانات الأرصادية مثل درجة الحرارة، الرطوبة، سرعة الرياح، O3 ، وCO، NO2 عاصفة الرياح، واتجاه الرياح. تم تنفيذ الطريقتين مع تحسين المعايير لبناء نموذج التنبؤ. تم تقييم النموذج تمتلك أداءً CatBoost أظهرت النتائج أن .(RMSE) باستخدام مقياس الدقة وجذر متوسط مربع الخطأ أفضل قليلاً مقارنة بالغابة العشوائية، مع دقة أعلى تصل إلى ٩٦,٢٨% مقارنةً بـ ٩٤,٨٨% للغابة العشوائية، وجذر متوسط مربع الخطأ أقل وهو ٠,٥٦٣٤٨ مقارنةً بـ ٠,٦٦١٩٩ للغابة العشوائية. ومع ..ذلك، فإن الفرق في الأداء ليس كبيرًا جدًا، وكلاهما يظهر أداءً جيدًا بشكل عام

BAB I
PENDAHULUAN

1.1 Latar Belakang

Udara merupakan salah satu sumber daya alam yang tidak dapat diperbarui. Udara yang bersih dan sehat menjadi syarat pokok keberlangsungan setiap mahluk di muka bumi. Kualitas udara adalah persepsi kualitas udara yang kita hirup dan berfungsi sebagai ukuran untuk menunjukkan tingkat keparahan polusi udara terhadap lingkungan dan kesehatan manusia (W. W. Li, 2020).

Polusi udara kini menjadi salah satu masalah lingkungan yang paling signifikan di banyak kota di seluruh dunia (Calo et al., 2024). Kota Palembang merupakan salah satu kota dengan tingkat polusi udara terparah pada tahun 2023 (Annur, 2024). Kualitas udara yang buruk kini menjadi ancaman utama bagi masyarakat global dengan menyebabkan efek berjenjang pada individu, sistem medis, kesehatan ekosistem, dan ekonomi baik di negara berkembang maupun negara maju (Liang & Gong, 2020).

Meningkatnya aktivitas transportasi, pertumbuhan industri, dan urbanisasi yang pesat telah menyebabkan peningkatan emisi polutan udara, yang dapat membahayakan kesehatan masyarakat dan lingkungan (Lee, 2019; Gunasekar et al., 2022). Polusi udara adalah kontaminasi lingkungan dalam atau luar ruangan oleh zat kimia, fisik, atau biologis apa pun yang mengubah sifat alami atmosfer. Menurut Menteri Lingkungan Hidup dan Kehutanan (LHK) bahwa penyebab polusi udara antara lain karena meningkatnya jumlah kendaraan, kegiatan industri, dan pembangkit listrik tenaga uap (PLTU). Selain itu pencemaran udara yang paling umum, terjadi di area pemukiman yang disebabkan oleh tiga sumber utama

diantaranya asap dari aktivitas rumah tangga, pembakaran limbah, dan operasi fasilitas industri. Ketiga sumber ini melepaskan gas-gas buangan hasil pembakaran yang mengkontaminasi udara di sekitar dan menurunkan kualitas udara yang dihirup.

Udara yang bersih dan sehat adalah salah satu nikmat besar dari Allah. Namun, seiring dengan aktivitas manusia yang semakin masif, kualitas udara sering terancam oleh pencemaran. Sehubungan dengan perintah Allah dan Rasul-Nya untuk menjaga kelestarian alam, umat Islam wajib memprediksi dan memantau kualitas udara. Seperti halnya dalam Alqur'an surat Ar-Rum ayat 41 bahwa Allah telah menekankan larangan untuk merusak dan mengeksploitasi alam tanpa memperhatikan pemeliharaannya karena kerusakannya akan berdampak kepada manusia sendiri.

ظَهَرَ الْفَسَادُ فِي الْبَرِّ وَالْبَحْرِ بِمَا كَسَبَتْ أَيْدِي النَّاسِ لِيُذِيقَهُمْ بَعْضَ الَّذِي عَمِلُوا لَعَلَّهُمْ يَرْجِعُونَ (٤١)

Artinya: *"Telah nampak kerusakan di darat dan di laut disebabkan karena perbuatan tangan manusia, (Melalui hal itu) Allah memebuat mereka merasakan sebagian dar (akibat) perbuatan mereka agar mereka kembali (ke jalan yang benar). (QS. Ar Rum: 41)*

Ayat ini menerangkan bahwa segala bentuk kerusakan yang terjadi di permukaan bumi disebabkan oleh perbuatan manusia, dan akibatnya juga berdampak pada manusia itu sendiri. Menurut Tafsir Ibnu Katsir, ayat diatas menjelaskan bahwa kerusakan di muka bumi disebabkan oleh perbuatan manusia yang melampaui batas, seperti pencemaran lingkungan, penggundulan hutan, dan eksploitasi sumber daya alam secara berlebihan.

Berdasarkan data IQAir, indeks kualitas udara (IQA) dan polusi udara PM2.5 didunia, Indonesia menempati peringkat ke-17 negara yang memiliki kualitas udara

terburuk didunia dan Jakarta menempati peringkat ke-5 kota yang memiliki kualitas udara terburuk (IQAir, 2023).

Polusi udara terdiri dari gas berbahaya dan partikel halus (PM2.5) yang mempengaruhi kualitas udara (Kothandaraman et al., 2022). Polutan yang menjadi perhatian Menteri Lingkungan Hidup dan Kehutanan meliputi parameter PM_{10}, $PM_{2.5}$, karbon monoksida, ozon, nitrogen dioksida, sulfur dioksida, dan hidrokarbon (HC). Tingginya nilai konsentrasi HC dan $PM_{2.5}$ memiliki resiko yang besar terhadap kesehatan manusia (Kusnandar, 2020). Selain itu, beberapa faktor yang mempengaruhi kualitas udara diantaranya suhu, intensitas angin, dan kelembapan. Pada musim panas, indeks kualitas udara berkorelasi positif secara signifikan dengan suhu udara, karena ketika massa udara hangat, suhu akan meningkat dan sejumlah besar polutan akan terakumulasi. Adapun pada kondisi kelembapan rendah, pertumbuhan inti kondensasi di atmosfer memperburuk polusi, dan pada kondisi kelembapan tinggi, hal ini akan berdampak pada pembuangan polutan akibat pengendapan (Y. Liu et al., 2022).

Polusi udara dapat berdampak negatif terhadap kesehatan manusia, hewan, dan tumbuhan. Dampak kesehatan tersebut seperti iritasi mata, penyakit pernapasan, penyakit jantung, stroke, dan kanker paru-paru. Sehingga untuk mengatasi masalah pencemaran udara, diperlukan upaya-upaya pencegahan dan pengendalian. Salah satu upaya tersebut adalah melakukan pemodelan terhadap kualitas udara.

Dalam beberapa dekade terakhir, pemodelan kualitas udara melalui teknik machine learning merupakan pendekatan yang menjanjikan untuk memprediksi dan memantau kondisi udara secara efektif (H. Liu et al., 2019). Dua algoritma yang

sering digunakan dalam pemodelan ini adalah *Random Forest* dan *CatBoost* yang telah terbukti mampu memberikan performa terbaik dibandingkan teknik lainnya dalam berbagai tugas regresi dan klasifikasi (Akanksha et al., 2023; Matović & Nataša, 2021).

Metode *Random Forest* adalah metode ansambel learning yang dapat diterapkan dalam beragam skenario, yang mencakup tugas regresi dan klasifikasi. Selama tahap pelatihan, *Random Forest* membangun banyak *Decision Tree* dan menghasilkan kelas yang terkait dengan mode kelas untuk klasifikasi atau prediksi rata-rata untuk tugas regresi, yang berasal dari masing-masing pohon (Mihirani et al., 2023).

Sementara itu, metode *CatBoost* merupakan salah satu algoritma Boosting, yang dikenal dengan keandalan dan kemampuan prediksi yang sangat efisien (Guo et al., 2023). Algoritma *CatBoost* merupakan pengembangan lebih lanjut dari metode Gradient Boosting dan Decision Tree. Menurut (Ding et al., 2021), *CatBoost* menyempurnakan konsep pembelajaran ensemble (gabungan) yang efisien dengan menggunakan teknik pengurutan lifting dan Decision Tree simetris sebagai pengklasifikasi lemah. Selanjutnya (Guo et al., 2023) menegaskan bahwa *CatBoost* meningkatkan kinerja dengan mengombinasikan kerangka kerja Gradient Boosting dan Decision Tree secara inovatif, di mana Decision Tree simetris berperan sebagai pengklasifikasi lemah yang dikombinasikan secara efisien melalui proses pengurutan lifting dalam kerangka pembelajaran ensemble.

Random Forest telah terbukti menjadi metode yang efektif dalam memprediksi kualitas udara. Studi yang dilakukan oleh (Altinçöp & Oktay, 2019) dengan menerapkan metode *Random Forest* dalam memprediksi kualitas udara

menunjukkan hasil yang sangat akurat dan memiliki kinerja lebih baik dibandingkan dengan jaringan syaraf tiruan. Selain itu, (S. Li et al., 2021) dalam penelitiannya menerapkan beberapa metode machine learning untuk memprediksi kualitas udara diantaranya *Random forest, Decision Tree, dan Deep BackPropagation Neural Network*. Didapatkan bahwa metode *Random Forest* memiliki kinerja terbaik dalam memprediksi kualitas udara.

Di sisi lain, *CatBoost* juga telah banyak digunakan dalam prediksi kualitas udara. Penelitian yang dilakukan oleh (Ravindiran et al., 2023) menunjukkan bahwa metode *CatBoost* memiliki performa yang lebih baik dibandingkan dengan metode lain seperti *Random Forest*, *Adaboost*, dan *XGBoost* dalam memprediksi kualitas udara.

Random Forest dan *CatBoost* memiliki kemampuan yang baik dalam melakukan prediksi kualitas udara. Namun, belum diketahui dengan pasti metode mana yang lebih unggul dalam menangani kompleksitas data kualitas udara di Indonesia, khususnya di Kota Palembang. Selain itu, optimasi parameter juga merupakan faktor penting yang dapat mempengaruhi performa kedua metode ini. *Random Forest* dan *CatBoost* memiliki beberapa parameter yang dapat dioptimasi, seperti jumlah pohon, kedalaman pohon, dan learning rate. Salah satu teknik optimasi yang populer adalah RandomSearch, yang telah terbukti efektif dalam mengoptimalkan hyperparameter pada berbagai kasus (Probst et al., 2019).

Berdasarkan uraian latar belakang diatas, maka perlu dilakukan sebuah penelitian tentang analisis perbandingan antara *Random Forest* dan *CatBoost* dalam memprediksi kualitas udara karena kualitas udara yang baik memainkan peran kunci dalam menjaga kesehatan manusia, lingkungan, dan ekosistem alamiah.

Dengan pemahaman yang lebih baik tentang kualitas udara dan dampak polusi, masyarakat dan pemerintah dapat mengambil langkah-langkah yang lebih efektif untuk menjaga udara yang bersih dan sehat. Model ini diharapkan dapat memberikan prediksi akurat tentang kualitas udara di masa depan dan memungkinkan masyarakat untuk mengambil tindakan preventif dan membantu pemerintah dalam perencanaan kebijakan yang lebih efektif.

1.2 Pernyataan Masalah

Berdasarkan uraian latar belakang di atas, maka rumusan masalah penelitian ini adalah bagaimana menganalisa perbandingan performa antara metode *Random Forest* dan *CatBoost* dalam pemodelan kualitas udara di Kota Palembang?

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah untuk mengetahui hasil analisa performa antara metode *Random Forest* dan CatBoost dalam pemodelan kualitas udara di Kota Palembang.

1.4 Manfaat Penelitian

Manfaat dari penelitian ini diharapkan dapat membantu pemerintah dan lembaga terkait dalam upaya merencanakan kebijakan lingkungan yang lebih efektif, memberikan kontribusi pada upaya global untuk mengatasi isu pencemaran udara dan perubahan iklim, serta meningkatkan kesadaran masyarakat tentang pentingnya kualitas udara yang baik bagi kesehatan dan lingkungan.

1.5 Batasan Masalah

Adapun batasan masalah dalam penelitian ini adalah:

1. Daerah perkotaan yang menjadi fokus penelitian adalah Kota Palembang.

2. Data yang digunakan merupakan data yang dipantau mulai 16 Januari 2022 hingga 16 Mei 2022 berupa data polusi (PM 10, PM 25, SO2, CO, O3, dan NO2) dan data meteorologi (*Temperature*, *Humidity*, *Wind Speed*, *Wind Gust*, dan *Wind Direction*).

BAB II
LITERATUR REVIEW

2.1 *Pemodelan Kualitas Udara*

Polusi udara adalah masalah global di masa sekarang yang implikasinya semakin mengerikan setiap hari (Nair et al., 2023). Pemodelan kualitas udara dapat dicapai dengan menerapkan model transportasi kimia atmosfer, model statistik, dan algoritma pembelajaran mesin (Gao et al., 2022).

Pemodelan kualitas udara telah dibahas dalam beberapa jurnal diantaranya dalam penelitian yang dilakukan Ravindiran et al., (2023), salah satu upaya yang dilakukan untuk memodelkan kualitas udara yaitu menerapkan beberapa model *Machine learning* seperti *LightGBM, Random Forest, CatBoost*, *Adaboost*, dan *XGBoost.* Ditemukan parameter uji $PM_{2.5}$ dan PM_{10} merupakan faktor penting dalam menentukan nilai Indeks Kualitas Udara, sementara karakteristik metrologi mempunyai dampak yang minimal. Dari penelitian ini diperoleh hasil bahwa *CatBoost* mengungguli model lainnya dengan R^2 sebesar 0.9998, MAE sebesar 0.60, MSE sebesar 0.58, dan RMSE sebesar 0.76. *Adaboost* memiliki hasil pemodelan paling efektif dengan R^2 sebesar 0.9753. Pada penelitian yang akan dilakukan akan menerapkan beberapa parameter polusi udara dan faktor meteorologi, tetapi pemodelan kualitas udara berfokus di beberapa kota Indonesia yang memiliki kualitas udara terburuk dalam beberapa bulan terakhir.

Kumar & Pande, (2023) melakukan penelitian yang lebih menekankan efisiensi teknik *machine learning* dalam memodelkan kualitas udara dibandingkan dengan metode-metode terdahulu. Langkah pertama yang dilakukan dalam pemodelan kualitas udara adalah melakukan *preprocessing* data, kemudian seleksi

fitur analisis yang korelasi dan analisis data eksploratori. Setelah itu dilakukan teknik *resampling* data dan terakhir menerapkan algoritma *machine learning* seperti *Gaussian Naive Bayes, Support Vector Machine, XGBoost*, dan lainnya. Dari kelima algoritma tersebut didapatkan bahwa algoritma *Naïve Bayes* memiliki akurasi tertinggi. Sedangkan algoritma *XGBoost* menunjukkan kinerja terbaik dengan RMSE 1.465, RMSLE 0.045, MAE 0.298 dan R^2 0.612. Peneliti menyarankan penggunaan model *XGBoost* untuk memodelkan dan menganalisis kualitas udara karena model tersebut dinilai menunjukkan kinerja terbaik dan memiliki hubungan kuat antara hasil prediksi dan data aktual. Dalam penelitian ini, data yang digunakan merupakan data polutan, sementara penelitian yang akan dilakukan akan menggunakan data polusi udara dan data faktor meteorologi.

Mihirani et al., (2023), dalam penelitiannya yang bertujuan memberikan peringatan dini tentang tingkat polusi udara melalui prakiraan kualitas udara. Penelitian ini berfokus pada indeks kualitas udara sebagai matrik utama untuk mengukur tingkat polusi dan menerapkan beberapa metode *machine learning* untuk menganalisis data polusi udara diantaranya *Linear Regression, Lasso Regression, Random Forest Regression,* dan *K-Nearest Neighbor Regression*. Adapun polutan yang dijadikan pertimbangan dalam penelitian ini yaitu Partikel Halus ($PM_{2.5}$), Sulfur Dioksida (SO_2), Nitrogen Dioksida (NO_2), dan Karbon Monoksida (CO). Kinerja dari model *machine learning* dievaluasi berdasarkan *Mean Absolute Error (MAE), Mean-Squared Error (MSE), Root Mean Squared Error (RMSE),* dan akurasi. Hasil dari beberapa model *machine learning* yang diterapkan, kemudian dibandingkan satu sama lain, sehingga ditemukan bahwa *Random Forest* memiliki performa terbaik dengan akurasi yang tinggi sebesar 99.87, MAE 0.09 dan RMSE

paling rendah dibandingkan model lainnya yaitu sebasar 0.422. Pada penelitian ini polutan yang dijadikan pertimbangan untuk diuji yaitu $PM_{2.5}$, SO_2, NO_2, CO, sedangkan penelitian saat ini mencakup $PM_{10,}$ $PM_{2.5}$, SO_2, NO_2, O_3, CO, HC, dan beberapa faktor meteorologi.

Selain itu, Kothandaraman et al., (2022) menerapkan sistem cerdas untuk pemodelan kualitas udara. $PM_{2.5}$ merupakan konsentrasi polutan yang utama diamati dalam penelitian ini. Mereka membandingkan 5 algoritma *machine learning* untuk pemodelan kualitas udara berdasarkan polutan $PM_{2.5}$ diantaranya *Linier Regression, Random Forest, KNN, Ridge and Lasso, XGBoost,* dan *AdaBoost.* Dari kelima perbandingan algoritma tersebut didapatkan algoritma dengan kinerja yang baik dengan tingkat kesalahan yang lebih rendah dari algoritma yang lain yaitu 10.59 RMSE dan 9.23 MAE menggunakan algoritma *AdaBoost.* Konsentrasi data yang digunakan dalam penelitian ini hanya berfokus pada data polusi udara dengan parameter $PM_{2.5,}$ sedangkan penelitian saat ini berfokus pada beberapa parameter polusi udara. Seperti yang diketahui bahwa kualitas data yang lebih baik akan menghasilkan model yang lebih akurat.

Pemodelan kualitas udara juga dilakukan oleh (Gladkova & Saychenko, 2022) Dalam penelitiannya dibahas bahwa peningkatan polusi udara menjadi masalah kritis bagi manusia. Tujuan utama dalam penelitian ini adalah memprediksi perubahan konsentrasi $PM_{2.5}$ untuk pemantauan kualitas udara dan pencegahan risiko. Penelitian yang dilakukan ini mencakup pemeriksaan komprehensif terhadap data mentah, serta analisis komparatif penerapan berbagai metode pembelajaran mesin. Untuk mengatasi tantangan dalam memprediksi data deret waktu terkait konsentrasi materi partikulat, beberapa model pembelajaran mesin digunakan,

diantaranya *ARIMA, Facebook Prophet, dan LSTM*. Dari beberapa model yang digunakan, ditemukan bahwa metode LSTM yang memberikan hasil lebih baik yaitu dengan value 7.865484 RMSE dan 61.865833 MSE. Hasil penelitian menunjukkan bahwa saat ini dimungkinkan untuk memperkirakan perubahan nilai rata-rata konsentrasi polutan beberapa bulan sebelumnya. Namun, perlu ditekankan bahwa penerapan teknologi ini secara luas memerlukan data yang lebih akurat dan andal untuk meningkatkan presisi dan efektivitas prediksi tersebut. Sehingga penelitian saat ini tidak hanya berfokus pada satu parameter data uji yaitu $PM_{2.5}$ tetapi mencakup beberapa parameter polutan udara dan faktor meteorologi.

Sebuah studi juga dilakukan oleh (Y. Liu et al., 2022), penelitian dilakukan untuk mengembangkan pemodelan yang dapat memproyeksikan kualitas udara dengan mempertimbangkan faktor-faktor meteorologi dan *data real-time* mengenai gas buangan industri. Data yang digunakan berasal dari berbagai sumber untuk pengembangan model bertujuan untuk memprediksi tingkat polutan udara dengan tingkat akurasi yang tinggi. Selain itu, model-model tersebut kemudian divalidasi dengan data independen untuk mengukur tingkat akurasi mereka dalam pemodelan kualitas udara. Adapun model-model yang dibandingkan yaitu *Random Forest, Backpropagation (BP) Neural Network, Decision Tree,* dan *Least Squares Support Vector Machine (LSSVM)*. Dari beberapa model tersebut didapatkan kinerja yang lebih baik dengan RMSE 22.91 dan MAE 15.80 adalah model *Random Forest*. Hasil dari penelitian ini memiliki implikasi praktis dalam mengelola kualitas udara, terutama dalam kasus di mana pengaruh gas buangan industri memiliki dampak signifikan pada kualitas udara lokal. Model-model ini dapat digunakan sebagai alat penting dalam pemantauan dan pengendalian polusi udara, membantu masyarakat

dalam memahami dan mengatasi fluktuasi kualitas udara yang mungkin terjadi. Pemodelan kualitas udara berdasarkan faktor pencemar udara dan parameter meteorologi dengan menerapkan metode *Random Forest* juga dilakukan pada penelitian saat ini tetapi berfokus pada kota-kota besar di Indonesia.

Adapun studi yang dilakukan (S. Li et al., 2021) dalam pemodelan kualitas udara didaerah pegunungan Wuling, daerah yang terjal dan licin saat musim hujan. Penelitian ini menerapkan metode *machine learning* untuk pemodelan kualitas udara dengan tujuan mengembangkan model prediksi kualitas udara yang lebih baik. Adapun metode *machine learning* tersebut yaitu yaitu *Random Forest, Decision Tree, dan Deep Back Propagation Neural Network*. Ketiga algoritma tersebut dievaluasi dengan cara membandingkan *Mean Squared Error* (MSE) dan *Mean Absolute Error* (MAE) antara nilai kualitas udara yang diprediksi dan diamati. Sehingga diperoleh hasil bahwa *Random Forest* memiliki kinerja terbaik dalam pemodelan kualitas udara di Wilayah Pegunungan Wuling dengan nilai MSE (MAE) untuk *Random Forest*: 3.54 (1.28), *Decision Tree*: 8.35 (1.92), dan *Deep Back Propagation Neural Network*: 5.01 (1.61). Berdasarkan penelitian sebelumnya, metode *Random Forest* memiliki kinerja yang lebih baik sehingga pada penelitian saat ini akan mengeksplorisasi metode yang sama tetapi dengan data terbaru pada kota-kota besar di Indonesia.

Madhuri et al., (2020) melakukan penelitian dengan mengeksplorasi penggunaan beberapa teknik *Machine learning Supervised Learning* seperti *Regresi Linier (LR), Support Vector Machine (SVM) , Decision Tree (DT)*, dan *Random Forest* (RF) untuk pemodelan kualitas udara. Adapun beberapa pertimbangan parameter kebutuhan untuk pemodelan kualitas udara seperti

kelembaban relatif udara, CO, Timah oksida, hidrokarbon nonlogam, Benzena, Titanium, NO, Tungsten, Indium oksida, Suhu, dan lain-lain. Temuan dari penelitian ini menunjukkan bahwa metode *Random Forest* memiliki performa lebih baik dalam pemodelan kualitas udara dengan nilai RMSE sebesar 0.84. Seperti yang telah diketahui bahwa penerapan *machine learning* dapat memberikan hasil pemodelan kualitas udara yang baik dengan menggunakan parameter pencemar udara dan meteorologi, sehingga pada penelitian selanjutnya akan diterapkan metode yang sama, tetapi mempertimbangkan beberapa parameter uji yang memengaruhi kualitas udara seperti Benzena dan Titanium.

Pada penelitian yang dilakukan Ambika et al., (2019), peneliti menggunakan metode Regresi Linier untuk membangun model kualitas udara berdasarkan *Air Quality Index* (AQI). Studi ini mengidentifikasi hubungan antara berbagai parameter yang memengaruhi kualitas udara, seperti konsentrasi polutan udara ($PM_{2.5}$, PM_{10}, CO, NO2, dan lain-lain) dan faktor-faktor lingkungan (suhu, kelembaban, dan kecepatan angin). Model yang diusulkan kemudian diuji dan divalidasi menggunakan data pengamatan sebelumnya. Sehingga diperoleh hasil RMSE 3.44 dan MAE 2.43. Fokus penelitian saat ini adalah pemodelan kualitas udara dibeberapa wilayah di Indonesia dengan menerapkan konsentrasi polutan udara dan faktor meteorologi.

Adapun penelitian lain yang dilakukan oleh (Anurag et al., 2019) untuk pemodelan kualitas udara yang efisien sehingga memungkinkan untuk meramalkan perubahan-perubahan yang tidak diinginkan yang terjadi pada lingkungan dan menjaga agar emisi polutan tetap terkendali. Penelitian ini menerapkan algoritma *XGBoost* untuk pemodelan kualitas udara dan menghasilkan nilai RMSE yang

rendah yaitu 15.97 dibandingkan dengan metode lainnya seperti *Neural Network, Decision Tree, dan Multiple Linier Regression*. Data yang digunakan merupakan data isolasi faktor-faktor yang berkontribusi lebih besar terhadap Indeks Kualitas Udara yaitu data meteorologi dan data faktor pencemar udara.

2.2 *Theoretical Framework*

Theoretical framework pada penelitian ini membahas tentang teori dari beberapa penelitian terdahulu yang relevan atau terkait dengan pemodelan kualitas udara. Tahap *theoretical framework* ini untuk mengidentifikasi metode yang relevan atau sesuai dengan topik yang akan dibahas pada penelitian ini. Berikut beberapa penelitian terdahulu tentang pemodelan kualitas udara.

Tabel 2. 1 Penelitian terdahulu

No	Identitas	Input	Metode	Performance
1.	Anurag *et al.*, (2019)	**Parameter polutan :** Karbon Monoksida, Benzena, Et benzena, Nitrous oksida, MP Xylene, Nitrogen Oksida (Nox), Ozon, PM2.5, SO2 dan Toluene. **Parameter meteorologi:** tekanan, kelembapan relatif, kecepatan angin, suhu, dan arah angin	XGBoost, Neural Network, Decision Tree, dan Multiple Linier Regression	XGBoost : 15.39
2.	Ambika *et al.* (2019)	**Parameter Polutan :** PM10, PM2.5, CO, NH3, SO2, NO2, PB **Parameter Kendaraan :** Jumlah kendaraan	Linier Regression	Linier Regression : 3.44

No	Identitas	Input	Metode	Performance
3.	Madhuri *et al.* (2020)	**Parameter Polutan :** CO, Tin oksida, Hidrokarbon non-logam, Benzena, Titanium, NO, Tungsten, Oksida Indium **Parameter Meteorologi:** kecepatan angin di atmosfer, arah angin, kelembaban relatif, dan suhu	Regresi Linier, SVM, Decision Tree, dan Random Forest	Random Forest : 0.84
4.	S. Li *et al. (2021)*	**Parameter Polutan**	Random Forest, Decision Tree, & Deep Back Propagation Neural Network	Random Forest : 3.54
5.	Y. Liu *et al.* (2022)	**Gas Limbah Industri Parameter Meteorologi:** curah hujan, suhu udara, kelembaban relatif, skala angin, tekanan udara, intensitas sinar matahari total dan curah hujan	Random Forest, Backpropagation Neural Network, Decision Tree, & Least Squares Support Vector Machine (LSSVM).	Random Forest : 22.91
6.	Gladkova & Saychenko (2022)	**Parameter Polutan:** PM2.5	ARIMA, Facebook Prophet, dan LSTM	LSTM : 7.86
7.	Kothandaraman *et al*. (2022)	**Parameter Polutan:** PM2.5	Linier Regression, Random Forest, KNN, Ridge and Lasso, XGBoost, dan AdaBoost	AdaBoost : 10.59

No	Identitas	Input	Metode	Performance
8.	Mihirani *et al*. (2023)	**Parameter Polutan**: PM2.5, SO2, NO2, CO	Linear Regression, Lasso Regression, Random Forest Regression, dan K-Nearest Neighbor Regression	Random Forest : 0.422
9.	Kumar & Pande, (2023)	**Parameter Polutan**	Gaussian Naive Bayes, SVM, XGBoost	XGBoost : 1.465
10.	Ravindiran *et al.* (2023)	**Parameter Polutan :** PM2.5, PM10, NO, NO2, NOx, NH3, SO2, CO, O3, Benzene, Toluene, dan Xylene **Parameter meteorologi:** Suhu, Kelembaban Relatif, Kecepatan Angin, Arah Angin, Radiasi Matahari, Tekanan Udara, Suhu, Curah Hujan dan Curah Hujan Total	LightGBM, Random Forest, *CatBoost*, Adaboost, dan XGBoost	*CatBoost :* 0.76

Pada tabel diatas menjelaskan beberapa metode dari penelitian terdahulu yang memiliki *performance* yang baik dalam menentukan kualitas udara. Dataset yang digunakan juga berpengaruh pada hasil kinerja dari masing-masing metode. Berikut rangkuman dari beberapa metode dengan hasil kinerja berdasarkan *error* yang paling kecil dimuat dalam tabel berikut.

Tabel 2. 2 Hasil Kinerja Metode

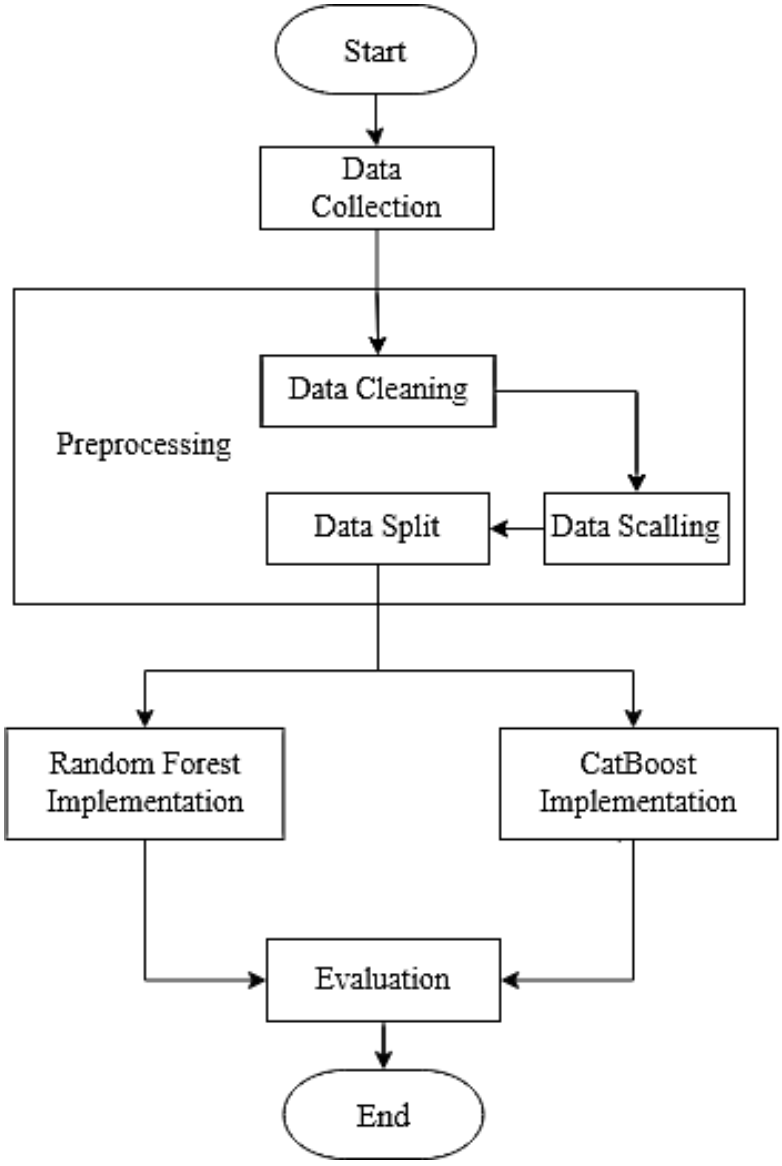
Identitas	Metode	Error
Mihirani *et al*. (2023)	Random Forest	0.422
Ravindiran *et al.* (2023)	*CatBoost*	0.76
Madhuri *et al.* (2020)	Random Forest	0.84
Anurag *et al.,* (2019)	XGBoost	15.39
Kumar & Pande, (2023)	XGBoost	1.465
Ambika *et al.* (2019)	Linier Regression	3.44
S. Li *et al. (2021)*	Random Forest	3.54
Gladkova & Saychenko (2022)	LSTM	7.86
Kothandaraman *et al*. (2022)	AdaBoost	10.59
Y. Liu *et al.* (2022)	Random Forest	22.91

Dari beberapa penelitian diatas dapat dilihat performa masing-masing metode. Metode yang memiliki hasil performa terbaik dibanding metode lainnya untuk pemodelan kualitas udara adalah *Random Forest* dan *CatBoost*. Oleh sebab itu, penelitian ini akan ditindaklanjuti untuk membandingkan metode *Random Forest* dan *CatBoost* dalam pemodelan kualitas udara pada Kota Palembang.

BAB III
METODOLOGI PENELITIAN

Tahap metodologi penelitian ini membahas beberapa hal mengenai prosedur penelitian yang akan dilakukan untuk pemodelan kualitas udara, seperti pengumpulan data, dimana input data yang digunakan meliputi konsentrasi polutan dan data meteorologi. Tahap berikutnya melibatkan *preprocessing* data untuk meningkatkan struktur data, termasuk pembersihan dan penskalaan data. Setelah itu, dataset dibagi menjadi bagian pengujian dan pelatihan untuk keperluan prediksi.

Pada tahap selanjutnya dilakukan implementasi metode. Metode yang diusulkan dalam penelitian ini adalah metode *Random Forest dan CatBoost*. Kemudian output yang dihasilkan berupa pemodelan kualitas udara berdasarkan faktor polusi udara dan faktor meteorologi dari masing-masing metode. Selanjutnya dilakukan evaluasi terhadap metode yang diusulkan guna melihat performa dari masing masing-masing metode. Berikut ilustrasi urutan langkah-langkah yang ditempuh dalam pelaksanaan penelitian ini ditunjukkan pada Gambar 3.1.



Gambar 3. 1 Desain Sistem

3. 1 *Data Collection*

Penelitian ini memanfaatkan data sekunder yang bersifat kuantitatif sebagai sumber informasinya. Jenis data kuantitatif merupakan jenis data yang berisi informasi berupa nilai numerik atau nilai angka yang dapat diukur. Dataset yang digunakan memuat informasi tentang konsentrasi polusi udara dan faktor meteorologi Kota Palembang**.**

3.1.1 Data Polutan

Sumber data polutan yang digunakan dalam penelitian ini dikumpulkan dari situs website (https://aqicn.org/). Situs yang memantau kualitas udara didunia yang berasal dari Badan Perlindungan Lingkungan Hidup atau *Environmental Protection Agencies* (EPA). Dataset yang digunakan berupa data kualitas udara Kota Palembang. Data yang digunakan merupakan data yang dipantau mulai 16 Januari pukul 17.00 hingga 16 Mei 2022 17.00. Dataset ini mencakup variabel partikulat matter 2.5 (PM2.5), partikulat matter 10 (PM 10), nitrogen dioksida (NO2), karbon monoksida (CO), sulfur dioksida (SO2), dan ozon (O3) . Variabel polutan udara yang digunakan dapat dilihat pada tabel 3.1 berikut ini.

Tabel 3. 1 Variabel Faktor Polusi Udara

Atribut	Unit	Deskripsi
PM 10	$\mu g/m^3$	Partikel udara dengan diameter kurang dari 2.5 mikrometer
PM 2.5	$\mu g/m^3$	Partikel udara dengan diameter kurang dari 10 mikrometer
SO2	$\mu g/m^3$	Sulfida
CO	$\mu g/m^3$	Karbon Monoksida
O3	$\mu g/m^3$	Ozon
NO2	$\mu g/m^3$	Nitrogen dioksida

3.1.2 Data Meteorologi

Beberapa faktor yang mempengaruhi kualitas udara adalah suhu, intensitas angin, dan kelembapan (Y. Liu et al., 2022). Dalam penelitian ini diterapkan juga data meteorologi untuk mengukur tingkat kualitas udara. Dataset diperoleh dari laman website resmi https://www.visualcrossing.com/ yang memuat tentang data historis dan prakiraan cuaca selama 24 jam perhari. Variabel data meteorologi yang

digunakan diantaranya *Temperature, wind speed, wind gust, wind direction,* dan *humadity*. Berikut variabel meteorologi ditujukan pada Tabel 3.2.

Tabel 3. 2 Variabel Faktor Meteorologi

Atribut	**Unit**	**Deskripsi**
Temp	*C*	*Temperature* / suhu
Hum	*%*	*Humidity* / Kelembapan
WS	*km/jam*	*Wind speed* / Kecepatan angin (Pergerakan udara secara umum)
WG	*km/jam*	*Wind Gust* / Hembusan Angin (Peningkatan udara secara signifikan dan mendadak)
WD	*0°- 360°*	*Wind direction / Arah angin*

Variabel target output kualitas udara berdasarkan Badan Perlindungan Lingkungan Hidup atau *Environmental Protection Agencies* (EPA). Sebuah platform yang dibawah dukungan PBB. Variabel target output disebut sebagai AQI (Air Quality Index), indeks yang dihitung berdasarkan konsentrasi berbagai polutan, termasuk yang diukur dalam µg/m³. AQI mengkonversi konsentrasi polutan yang diukur (dalam µg/m³ atau satuan lain) menjadi skala standar 0-500. Berikut variabel output ditujukan pada Tabel 3.3.

Tabel 3. 3 Variabel output

Nilai	Rentang Angka	Kategori
1	0 - 50	Baik
2	51 - 100	Sedang
3	101 - 200	Tidak sehat untuk kelompok sensitif
4	201 - 300	Tidak Sehat
5	301 - 400	Sangat Tidak Sehat
6	401 ≥	Berbahaya

Berikut contoh dataset kualitas udara sebagai berikut.

Tabel 3. 4 Contoh Dataset

name	datetime	tem	hur	w	w	w	pm1	pm2	o3	so	no	co	ac
Palembang	2024-01-16T17:00:00	30	70.29	14.8	13	350	2.25	1.41	96.92	0.91	2.49	507.63	45
Palembang	2024-01-16T18:00:00	25	88.68	11.5	13	300	2.42	1.43	96.92	0.83	2.11	511.53	45
Palembang	2024-01-16T19:00:00	25.1	92.81	6.1	2.7	300	2.8	1.95	96.08	0.77	1.77	502.63	44
Palembang	2024-01-16T20:00:00	25	94.19	5.4	3.6	10	3.18	2.47	95.25	0.7	1.44	493.73	44
Palembang	2024-01-16T21:00:00	25	94.19	6.8	3.6	280	3.56	3	94.41	0.63	1.1	484.82	44
Palembang	2024-01-16T22:00:00	25	94.75	16.2	3.6	50	3.36	2.6	98.47	0.64	1.17	484.55	46
Palembang	2024-01-16T23:00:00	25	94.19	20.9	3.6	10	3.15	2.2	102.52	0.64	1.23	484.27	47
Palembang	2024-01-17T00:00:00	25	100	18	3.6	360	2.95	1.8	106.57	0.65	1.29	483.99	49
Palembang	2024-01-17T01:00:00	25	97.67	7.9	5.4	315	3.58	2.19	116.05	0.82	1.27	491.22	54
Palembang	2024-01-17T02:00:00	25	94.19	7.9	7.6	360	4.21	2.58	125.53	0.99	1.24	498.45	57
Palembang	2024-01-17T03:00:00	25	94.19	10.8	7.6	300	4.84	2.96	135	1.16	1.22	505.69	69
Palembang	2024-01-17T04:00:00	24.9	95.03	10.1	0	0	4.37	2.77	137.45	1.17	0.97	500.4	72
Palembang	2024-01-17T05:00:00	23.3	96.43	7.2	5	310.1	3.89	2.58	139.89	1.18	0.72	495.12	75
Palembang	2024-01-17T06:00:00	23.1	96.43	4.3	4.3	295.2	3.42	2.39	142.34	1.18	0.47	489.83	78
Palembang	2024-01-17T07:00:00	25.1	97.96	9	3.6	205	2.85	2.02	143.11	1.22	0.55	496.23	79
Palembang	2024-01-17T08:00:00	26	94.23	12.2	9.4	300	2.27	1.65	143.89	1.26	0.63	502.63	80
Palembang	2024-01-17T09:00:00	27.6	78.38	19.4	9.4	318.8	1.7	1.28	144.66	1.3	0.71	509.02	81
Palembang	2024-01-17T10:00:00	29.2	79.08	17.6	9.4	340	2.34	1.69	141.2	1.17	2	532.39	77
Palembang	2024-01-17T11:00:00	29.9	65.75	16.9	12.2	330.6	2.98	2.1	137.75	1.04	3.29	555.75	72
Palembang	2024-01-17T12:00:00	31	66.38	16.2	14.8	280	3.62	2.52	134.29	0.91	4.58	579.12	68
Palembang	2024-01-17T13:00:00	30.6	71.87	15.8	18.4	350	4.51	2.88	131.31	0.89	4.08	589.41	64
Palembang	2024-01-17T14:00:00	29.9	66.96	15.1	11.2	352.9	5.41	3.24	128.33	0.87	3.59	599.7	60
Palembang	2024-01-17T15:00:00	24	100	44.6	22.3	300	6.3	3.6	125.35	0.85	3.09	609.99	57
Palembang	2024-01-17T16:00:00	24.8	95.59	16.9	14.8	320	5.7	3.24	125.11	0.82	2.66	607.49	56
Palembang	2024-01-17T17:00:00	24	100	18.7	9.4	20	5.11	2.87	124.87	0.8	2.23	604.99	56

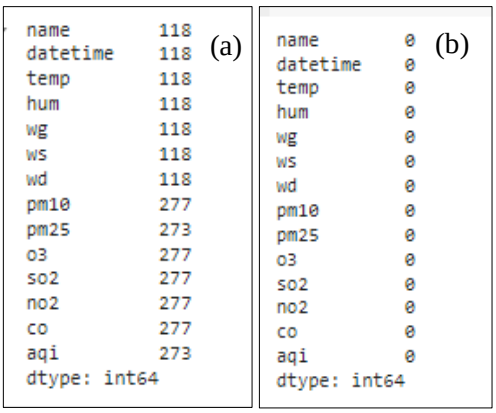
3. 2 *Data Pre-processing*

Tahap *preprocessing* merupakan tahap penting dalam proses pengolahan data. Terkadang data yang dimiliki tidak memenuhi kualitas standar untuk dilakukan proses. Oleh karena itu perlu dilakukan *preprocessing* untuk memperbaiki data yang tidak berkualitas menjadi format data yang berkualitas sehingga dapat dijadikan sebagai input untuk pengujian metode yang diusulkan dan menghasilkan akurasi yang tinggi. Berikut langkah-langkah yang dilakukan dalam *preprocessing* data.

3.2.1 Data Cleaning

Proses pembersihan data merupakan tahap awal *preprocessing yang* dilakukan untuk mengidentifikasi adanya missing value atau data yang hilang/tidak

lengkap. Tujuannya adalah untuk mengatasi masalah yang muncul ketika terdapat sel-sel kosong pada satu atau beberapa variabel data. Proses pembersihan data dilakukan agar dapat meningkatkan keakuratan hasil. Langkah ini penting sebagai upaya untuk memastikan kualitas dan kelengkapan data sebelum diproses lebih lanjut. Berikut contoh dataset yang telah dilakukan *cleaning.*



(a) Sebelum *cleaning*, (b) Sesudah *cleaning*

Gambar 3. 2 Contoh hasil data *cleaning*

Dari Gambar 3.2 diatas dapat diketahui bahwa pada bagian (a) dideteksi terdapat missing value pada masing-masing variabel. Sehingga langkah yang dilakukan untuk menangani hal tersebut dengan cara melakukan cleaning atau drop data pada variabel-variabel yang memiliki nilai *missing value,* seperti yang terlihat pada bagian (b).

3.2.2 Data Scalling

Dataset yang digunakan merupakan dataset yang memiliki satuan yang berbeda-beda, sehingga sebelum melakukan tahap training model yang diusulkan, terlebih dahulu dilakukan tahap pengskalaan ulang dataset menggunakan teknik normalisasi. Normalisasi data bertujuan untuk menyamakan skala antar variabel,

sehingga tidak ada variabel yang lebih dominan hanya karena memiliki rentang nilai yang lebih besar dibandingkan variabel lainnya. Dengan melakukan normalisasi, setiap variabel akan berkontribusi dengan besaran yang seimbang dalam proses analisis atau pemodelan, tanpa terpengaruh oleh perbedaan skala aslinya. Normalisasi ini dilakukan untuk mengubah skala data menjadi rentang data yang normal dengan nilai berada antara 0 dan 1. Berikut persamaan yang akan digunakan untuk normalisasi (Raschka, 2018).

$$\boldsymbol{x_{norm}} = \frac{\boldsymbol{x} - \boldsymbol{x_{min}}}{\boldsymbol{x_{max}} - \boldsymbol{x_{min}}}$$

dimana $\boldsymbol{x}$ merupakan nilai asli dari fitur, sedangkan $\boldsymbol{x_{min}}$ dan $\boldsymbol{x_{max}}$ merupakan nilai minimum dan maksimum dari fitur tersebut. Data berskala yang diperoleh berada pada rentang [0, 1]. Berikut contoh hasil normalisasi data dapat dilihat pada Tabel 3.5.

Tabel 3. 5 Contoh Hasil Normalisasi Data

temp	hum	wg	ws	wd	pm10	pm25	o3	so2	no2	co
0.5000	0.6700	0.2186	0.1746	0.8972	0.0065	0.0079	0.1139	0.0732	0.0854	0.0085
0.4167	0.7513	0.2127	0.1422	0.9250	0.0014	0.0018	0.0812	0.0565	0.0844	0.0073
0.4250	0.7273	0.1595	0.0043	0.8194	0.0016	0.0019	0.0732	0.0521	0.0818	0.0073
0.3333	0.8381	0.1492	0.0345	0.0389	0.0017	0.0020	0.0653	0.0478	0.0792	0.0074
0.3333	0.8381	0.1492	0.0237	0.0639	0.0018	0.0022	0.0574	0.0436	0.0766	0.0074
0.3833	0.8494	0.1433	0.0000	0.3056	0.0028	0.0035	0.0502	0.0559	0.0805	0.0078
0.3583	0.8758	0.1492	0.0237	0.0000	0.0039	0.0048	0.0431	0.0681	0.0844	0.0082
0.3333	0.9032	0.1551	0.1530	0.7833	0.0050	0.0060	0.0359	0.0804	0.0884	0.0086
0.2500	0.8197	0.1123	0.4763	0.8833	0.0089	0.0109	0.0926	0.0744	0.0734	0.0100
0.1667	0.9018	0.1226	0.2026	0.8889	0.0129	0.0157	0.1492	0.0682	0.0584	0.0113
0.1667	1.0000	0.1595	0.1164	0.7231	0.0169	0.0206	0.2059	0.0622	0.0434	0.0126
0.1833	0.8998	0.1329	0.0022	0.5278	0.0191	0.0233	0.2517	0.0578	0.0322	0.0126
0.1667	0.9018	0.1654	0.1164	0.7222	0.0214	0.0312	0.2975	0.0533	0.0210	0.0125
0.1667	1.0000	0.2230	0.2802	0.7500	0.0236	0.0255	0.3432	0.0487	0.0099	0.0125
0.2750	0.8312	0.2083	0.0086	0.7222	0.0248	0.0332	0.3528	0.0592	0.0131	0.0129
0.3083	0.8378	0.2186	0.2134	0.7500	0.0260	0.0427	0.3623	0.0695	0.0164	0.0132
0.4167	0.8128	0.1920	0.2414	0.8056	0.0272	0.0370	0.3719	0.0800	0.0197	0.0136
0.5000	0.6867	0.1861	0.1638	0.8056	0.0246	0.0301	0.3502	0.0960	0.0367	0.0141
0.5833	0.6485	0.2083	0.3966	0.8333	0.0221	0.0270	0.3286	0.1120	0.0537	0.0146
0.5583	0.5929	0.2230	0.3384	0.9472	0.0195	0.0238	0.3069	0.1280	0.0707	0.0152
0.5750	0.5553	0.2127	0.4741	0.9139	0.0213	0.0261	0.2474	0.1231	0.0802	0.0139

3.2.3 Data Splitting

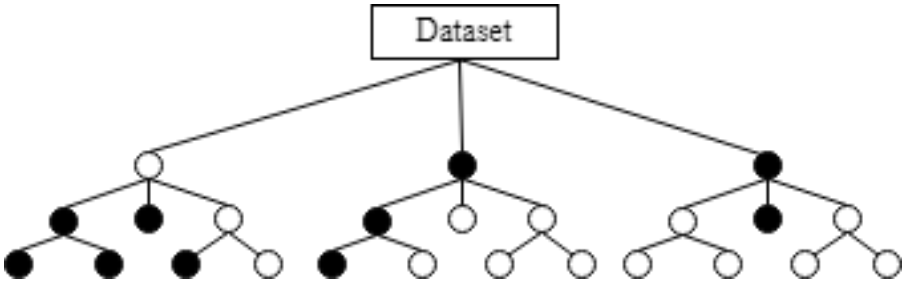
Setelah tahap *preprocessing* data selesai, dataset dibagi menjadi dua bagian yaitu data *training* dan data *testing*. Data *training* berfungsi untuk mengembangkan model yang diajukan, sementara data *testing* dimanfaatkan untuk menilai efektivitas model yang sudah dilatih tersebut.

Dalam penelitian ini, terdapat empat rasio pembagian data yang digunakan untuk memisahkan dataset menjadi dua bagian - data pelatihan (data training) dan data pengujian (data testing). Empat variasi rasio data yang diaplikasikan adalah 80:20, 70:30, 60:40, dan 50:50.

3. 3 *Implementation Random Forest*

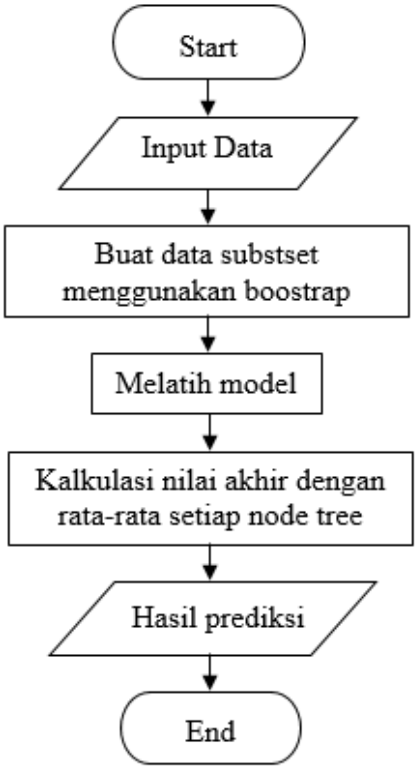
Random Forest adalah algoritma *machine learning* yang digunakan untuk regresi dan klasifikasi. Secara teoretis, algoritma ini terbentuk oleh sejumlah besar *decision tree* yang berkontribusi pada hasil prediksi berdasarkan kontribusi masing-masing *tree*. Salah satu keunggulan algoritma ini adalah kemampuan untuk mengatasi nilai yang hilang dan mencegah masalah *overfitting* melalui penggunaan banyak pohon, dan kinerja yang efisien dalam penanganan dataset yang besar. Beberapa parameter yang dapat diubah untuk meningkatkan kinerja algoritma diantaranya ukuran kumpulan prediksi, kedalaman maksimum pohon, jumlah atribut yang dipilih secara acak, jumlah iterasi, dan jumlah benih acak.

Prinsip kerja *Random Forest* adalah menyimpan seluruh dataset pelatihan dan mencari pola pelatihan yang paling mirip saat membuat prediksi atau menemukan tetangga terdekat. Pada dasarnya algoritma ini menggunakan pendekatan pencarian tetangga terdekat dan fungsi jarak.



Gambar 3. 3 Ilustrasi *Random Forest*

Algoritma *Random Forest* dibagi menjadi dua bagian. Yang pertama adalah membentuk pohon "k" untuk membuat hutan acak. Yang kedua adalah melakukan prediksi berdasarkan struktur acak yang terbentuk. Langkah-langkah penerapan teknik *Random Forest* adalah sebagai berikut:



Gambar 3. 4 Alur Metode *Random Forest*

Pada Gambar 3.4 diatas dimulai dari input data, dimana data yang digunakan sebagai input dalam perhitungan metode *Random Forest* adalah data yang telah

melalui tahap *preprocessing* dan *split* data yang menjadi data training dan data testing.

Langkah selanjutnya setelah input data adalah membuat boostrap dataset atau mengambil data secara acak. Selanjutnya, kita membuat subset acak dengan membangun pohon keputusan berdasarkan data yang dihasilkan dari proses bootstrap. Pohon keputusan ini dibuat menggunakan ukuran informasi atau metode CART (Classification and Regression Tree). Dalam CART, kriteria untuk menentukan pohon keputusan terbaik disebut Indeks Gini. Menghitung nilai informasi menggunakan Gini Impurity atau Indeks Gini (GI) untuk setiap variabel. Informasi ini digunakan untuk menentukan variabel mana yang akan menjadi node dalam pohon (T. H. Lee et al., 2020). Namun karena kasus ini memprediksi nilai "aqi" yang berupa nilai numerik (regresi), maka menggunakan metrics lain seperti seperti Variance atau Mean Squared Error (MSE) (Ramosaj & Pauly, 2019). Berikut persamaan untuk menghitung *Variance* adalah:

$$Variance = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

di mana y_i adalah nilai target dan $\bar{y}$ adalah rata-rata nilai target dalam node.

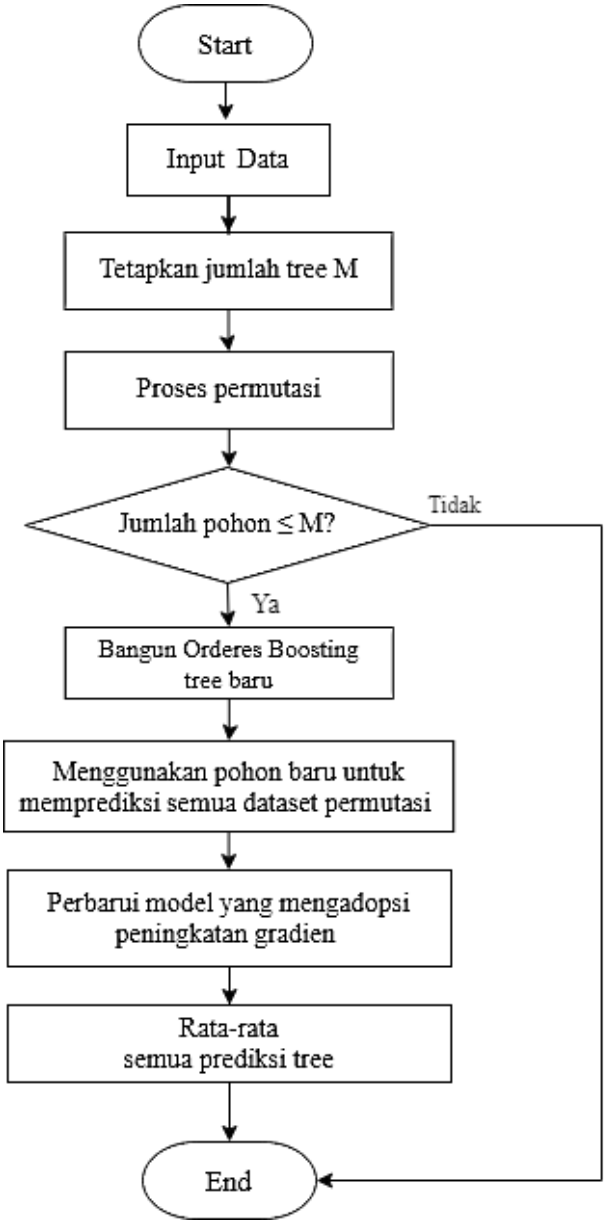
Setelah membangun sejumlah *Decision tree*, langkah terakhir adalah menggabungkan atau mengagregasi semua hasil prediksi dari *tree* tersebut untuk menghasilkan prediksi akhir. Dalam kasus regresi, metode yang umum digunakan adalah dengan menghitung rata-rata dari semua prediksi pohon.

3. 4 *Implementation CatBoost*

Selain Random Forest, algoritma *CatBoost* juga diimplementasikan dalam penelitian ini untuk pemodelan kualitas udara. *CatBoost* adalah algoritma *machine*

learning yang tergabung dalam keluarga *Gradient Boosted Decision Trees* (GBDT) yang sangat efektif dan memiliki kinerja yang tinggi untuk tugas klasifikasi, regresi (Jabeur et al., 2021). *CatBoost* menggunakan kombinasi peningkatan terurut, permutasi acak, dan pengoptimalan berbasis gradien untuk mencapai kinerja tinggi pada kumpulan data besar dan kompleks dengan fitur kategoris.

Model *CatBoost* merupakan kumpulan dari beberapa pohon keputusan dan prediksi akhir dibuat dengan cara menggabungkan atau menjumlahkan prediksi yang dihasilkan oleh setiap pohon tersebut (Yalçin et al., 2023). Adapun alur metode CatBoost untuk prediksi disajikan pada Gambar 3.5 berikut.



Gambar 3. 5 Alur Metode *CatBoost*
(Yalçin et al., 2023)

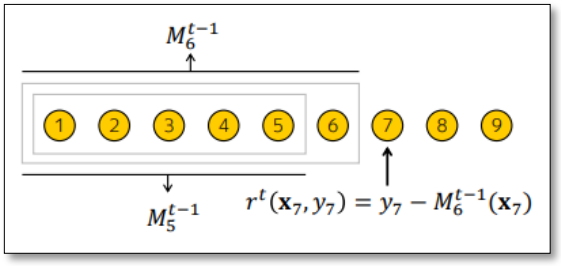
Gambar 3.5 diatas menunjukkan alur metode CatBoost, dimana proses dimulai dari input data. Setelah itu melakukan permutasi acak meminimalkan overfitting dan memaksimalkan penggunaan dataset secara keseluruhan dalam proses pelatihan. Adapun persamaan untuk proses permutasi acak menurut

(Dorogush et al., 2018) dapat dinyatakan melalui formula berikut. Misalkan σ = ($\sigma_1, \ldots, \sigma_n$) adalah permutasinya.

$$\frac{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k} = x_{\sigma_p,k}\right] Y_{\sigma_j} + \alpha \cdot P}{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k} = x_{\sigma_p,k}\right] + \alpha}$$

P mewakili *Prior* yang diperoleh dari nilai rata-rata label dalam dataset, sedangkan α adalah parameter positif atau parameter > 0 yang berfungsi sebagai bobot *prior*.

Langkah selanjutnya adalah membangun pohon *Ordered Boosting* baru yang akan digunakan untuk estimasi gradien atau residual. Proses *Ordered Boosting* melibatkan pengacakan urutan sampel pelatihan σ dan pemeliharaan serangkaian n model pendukung (M_1,…, M_n). Setiap model M_i dilatih menggunakan hanya i sampel pertama dalam urutan acak tersebut. Pada setiap tahap, residual untuk sampel j dihitung menggunakan model M_{j-1}. Konsep *Ordered Boosting* ini diilustrasikan dalam gambar berikut (Prokhorenkova et al., 2018) .



Gambar 3. 6 Ilustrasi *Ordered Boosting*

(Prokhorenkova et al., 2018)

Dengan menggunakan *tree* Ordered Boosting tersebut, melakukan prediksi untuk semua set permutasi, dan terus memperbarui matriks M untuk setiap set permutasi. Berdasarkan iterasi, prediksi dilakukan dengan menghitung rata-rata dari semua prediksi iterasi *tree* tersebut.

3. 5 Evaluation

Evaluasi adalah tahap yang dilakukan setelah tahap implementasi model. Tahap evaluasi merupakan tahap yang berfungsi untuk menganalisis kinerja model yang telah diterapkan. Untuk mengukur kinerja suatu model diperlukan metrik pengukur seperti RMSE dan akurasi. Seberapa akurat hasil prediksi yang diperoleh dapat dianalisis dengan mempertimbangkan nilai Root Mean Square Error (RMSE), dan koefisien determinasi (R2) (Gladkova & Saychenko, 2022)

Root Mean Squared Error adalah akar kuadrat dari rata-rata selisih kuadrat antara nilai target dan nilai yang diprediksi oleh model. Dalam matematika, RMSE adalah akar kuadrat dari *mean squared error* (MSE), yang merupakan perbedaan absolut minimum antara nilai keluaran aktual yang diamati dan nilai prediksi model (Kothandaraman et al., 2022).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{x_i} - x)^2}$$

dimana:

$n = jumlah\ total\ dataset$

$\widehat{x_i}$ = Nilai prediksi

x = Nilai sebenarnya

R-square mengindikasikan persentase variasi dalam hasil yang dapat dijelaskan oleh variabel terikat (R2). Indeks kinerja R-square mencerminkan sejauh mana kesesuaian antara nilai prediksi dan nilai sebenarnya (Madhuri et al., 2020). Untuk menghitung R-square, kita dapat memanfaatkan fungsi r2_score yang terdapat dalam modul sklearn.metrics (Y. Liu et al., 2022).

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{x_i} - \bar{x})^2}{\sum_{i=1}^{n}(x - \bar{x})^2}$$

Dimana n adalah jumlah total dataset, $\widehat{x_i}$ adalah nilai prediksi, x adalah nilai sebenarnya, dan $\bar{x}$ menunjukkan rata-rata nilai sebenarnya.

Performa model pembelajaran mesin akan semakin optimal ketika *Root Mean Squared Error* (RMSE) mendekati nilai minimum, sementara nilai R-kuadrat semakin mendekati nilai maksimum.

BAB IV

HASIL DAN PEMBAHASAN

Penelitian ini berisi tentang analisis perbandingan metode *Random Forest* dan *CatBoost* dalam pemodelan kualitas udara di Kota Palembang. Hasil dari masing-masing model kemudian dievaluasi untuk melihat performa yang baik dari masing-masing model.

4. 1 Pengujian Metode *Random Forest*

Pada penelitian ini, metode *Random Forest* dari perpustakaan Scikit-learn python digunakan untuk membangun model prediksi kualitas udara di Kota Palembang. Dalam mengimplementasikan *Random Forest* menggunakan Python, langkah awal yang dilakukan adalah membagi dataset yang berisi informasi kualitas udara menjadi data pelatihan dan data pengujian dengan rasio yang berbeda-beda, yaitu 80:20, 70:30, 60:40, dan 50:50. Pembagian dataset ini bertujuan untuk mengevaluasi performa model pada berbagai skenario rasio data yang berbeda.

Selanjutnya, pada setiap skenario rasio data, model *Random Forest Regressor* diinisialisasi dengan menggunakan *random_state* yang berbeda, yaitu 45, 60, dan 75. Parameter *random_state* ini digunakan untuk mengendalikan proses pengacakan dalam pembangunan pohon keputusan (*decision tree*) di dalam *Random Forest*. Setelah diinisialisasi, model dilatih dengan data pelatihan menggunakan *metode fit()*.

Untuk mengevaluasi performa model, metrik *Root Mean Squared Error* (RMSE) dan akurasi digunakan. RMSE dihitung dengan menghitung akar dari rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi. Sementara akurasi dihitung

dengan membandingkan nilai prediksi dengan nilai aktual pada data pengujian. Berikut potongan *source code* yang digunakan dalam pemodelan kualitas udara Kota Palembang menggunakan metode *Random Forest*.

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state =42)
model = RandomForestRegressor(n_estimators =1000,
                              max_depth=60,
                              min_samples_split=5,
                              min_samples_leaf=2,
                              max_features=5,
                              oob_score=True,
                              random_state =75)
model.fit(X_train, y_train)
predict=model.predict(X_test)
```

Potongan kode di atas menunjukkan penggunaan model *Random Forest* Regressor dari *scikit-learn* untuk memprediksi kualitas udara. Bagian kode "*test_size=0.2*" diubah sesuai dengan rasio data yang berbeda-beda untuk setiap pengujian. Sementara itu, tulisan "*random_state =75"* yang ditandai dengan warna merah merujuk pada parameter *random_state* dalam inisialisasi model *RandomForestRegressor*. Parameter ini berfungsi untuk mengendalikan proses pengacakan dalam pembangunan pohon keputusan *(decision tree)* di dalam *Random Forest.*

4.1.1 Skenario Pengujian 1

Skenario pengujian pertama kualitas udara di Kota Palembang, dilakukan uji coba dengan rasio pembagian data data training dan data testing yaitu 80:20. Setiap pengujian, dilakukan penerapan tiga variasi jumlah *random_state* , yaitu 45, 60, dan 75. Dibawah ini (Tabel 4.1) dapat dilihat hasil uji coba pemodelan kualitas udara di Kota Palembang dengan jumlah *random_state* = 45.

Tabel 4. 1 Hasil pemodelan RF Skenario 1 *random_state =45*

Index	Actual_Value	Predicted_Value	Deviation
23-4-2022 4:00	34	33.78	-0.2
23-4-2022 5:00	30	29.85	-0.2
23-4-2022 6:00	26	25.93	-0.1
23-4-2022 7:00	28	28.31	0.3
23-4-2022 8:00	31	31.57	1
...	...	...	...
16-5-2022 13:00	31	31.02	0.0
16-5-2022 14:00	23	23.90	1
16-5-2022 15:00	24	24.41	0.4
16-5-2022 16:00	27.2	27.47	0.3
16-5-2022 17:00	25	24.80	-0.2

Hasil pengujian pada Tabel 4.1 diatas terlihat bahwa perbandingan hasil prediksi dengan hasil actual dengan rasio perbandingan data 80:20 (23 April 2022 sampai 16 Mei 2022) dan *random_state =45* menunjukkan hasil yang cukup akurat. Ini terbukti dari beberapa hasil yang memiliki nilai deviasi yang cukup kecil yaitu bernilai -1 dan -2, meskipun terdapat beberapa kasus yang memiliki selisih antara nilai prediksi dengan nilai aktual bernilai 1. Adapun pengujian dengan *random_state 60.*

Tabel 4. 2 Hasil pemodelan RF Skenario 1 *random_state =60*

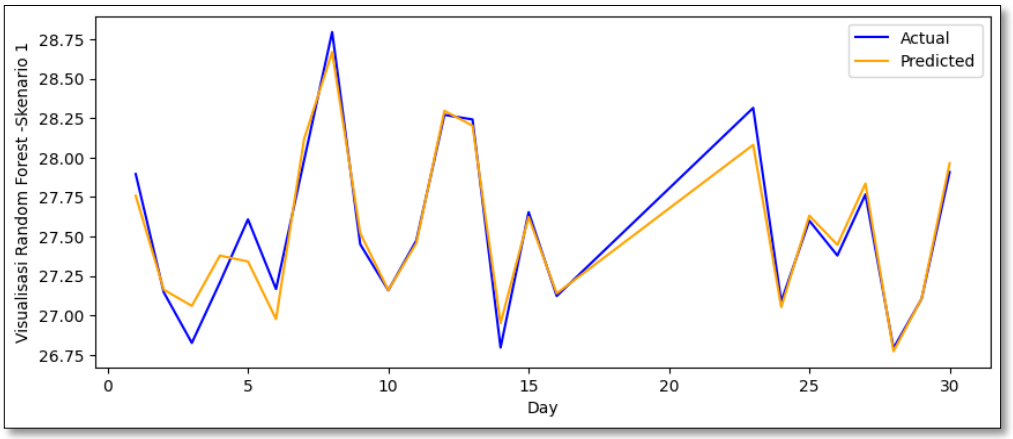
Index	Actual_Value	Predicted_Value	Deviation
23-4-2022 4:00	34	33.78	-0.2
23-4-2022 5:00	30	29.85	-0.1
23-4-2022 6:00	26	25.92	-0.1
23-4-2022 7:00	28	28.32	0.3
23-4-2022 8:00	31	31.63	1
...	...	...	...
16-5-2022 13:00	31	31.01	0.0
16-5-2022 14:00	23	23.88	1
16-5-2022 15:00	24	24.39	0.4
16-5-2022 16:00	27.2	27.52	0.3
16-5-2022 17:00	25	24.74	-0.3

Berdasarkan model prediksi pada Tabel 4.2 dengan menggunakan nilai acak (*random_state=60*) menunjukkan hasil prediksi yang stabil. Hal ini dibuktikan dengan selisih atau deviasi antara nilai prediksi dengan nilai aktual relatif kecil dengan nilai selisih -0 dan 0 yang mengindikasikan bahwa model mampu memprediksi nilai dengan baik dan mendekati nilai sebenarnya. Selain itu, dilakukan pengujian dengan *random_state = 75*, sebagaimana ditunjukkan pada Tabel 4.3.

Tabel 4. 3 Hasil pemodelan RF Skenario 1 *random_state =75*

Index	Actual_Value	Predicted_Value	Deviation
23-4-2022 4:00	34	33.79	-0.2
23-4-2022 5:00	30	29.84	-0.2
23-4-2022 6:00	26	25.89	-0.1
23-4-2022 7:00	28	28.34	0.3
23-4-2022 8:00	31	31.54	1
...	...	...	...
16-5-2022 13:00	31	31.02	0.0
16-5-2022 14:00	23	23.94	1
16-5-2022 15:00	24	24.40	0.4
16-5-2022 16:00	27.2	27.46	0.3
16-5-2022 17:00	25	24.76	-0.2

Hasil evaluasi yang ditampilkan pada Tabel 4.3 diperoleh dengan menerapkan *random_state* 75. Pengujian ini menghasilkan outcome yang serupa dengan uji coba sebelumnya. Beberapa prediksi menunjukkan deviasi minimal dari nilai sebenarnya, dengan selisih kurang dari 1. Gambar 4.1 menyajikan representasi visual dari model kualitas udara Kota Palembang, menggunakan pembagian data latih dan uji dengan proporsi 80:20.



Gambar 4. 1 Visualisasi hasil pemodelan Skenario 1 *Random Forest*

Gambar 4.1 memperlihatkan komparasi nilai sebenarnya (ditunjukkan dengan garis biru) dan hasil prediksi (ditandai dengan garis orange) untuk kualitas udara. Data yang digunakan mencakup periode 23 April sampai 16 Mei 2022, dengan rasio pembagian 80:20. Secara umum, kedua garis menunjukkan kecenderungan yang mirip, mengindikasikan bahwa model cukup mampu menangkap pola keseluruhan data. Meski di beberapa titik prediksi sangat mendekati nilai aktual (terlihat dari garis yang hampir berhimpit), terdapat pula perbedaan yang cukup besar di beberapa bagian, terutama pada interval hari ke-3 hingga ke-6 dan hari ke-18 hingga ke-23.

Hasil kinerja berbagai uji coba metode *Random Forest* dalam melakukan pemodelan kualitas udara dengan rasio pembagian data 80:20 telah dirangkum dalam tabel di bawah ini**.**

Tabel 4. 4 Performance metode *Random Forest* Skenario 1

Random state	RMSE	Accuracy
45	**0.63811**	**95.13%**
60	0.64315	95.05%
75	0.64158	95.08%

Hasil analisis pada Tabel 4.4 mengungkapkan bahwa untuk pemodelan kualitas udara Kota Palembang dalam Skenario 1, dengan pembagian data 80:20, kinerja optimal dicapai melalui pengaturan khusus metode *Random Forest*. Penggunaan *random_state 45* menghasilkan performa terbaik, ditandai dengan nilai RMSE terendah sebesar 0,63811 dan tingkat akurasi tertinggi mencapai 95,13%.

4.1.2 Skenario Pengujian 2

Dalam Skenario pengujian kedua untuk kualitas udara Kota Palembang, dilakukan eksperimen dengan proporsi data latih dan uji sebesar 70:30. Berikut disajikan hasil prediksi kualitas udara kota tersebut menggunakan *random_state 45*.

Tabel 4. 5 Hasil pemodelan RF Skenario 2 *random_state =45*

Index	Actual_Value	Predicted_Value	Deviation
11-4-2022 9:00	34	33.67	-0.3
11-4-2022 10:00	30	29.80	-0.2
11-4-2022 11:00	26	25.90	-0.1
11-4-2022 12:00	28	28.32	0.3
11-4-2022 13:00	31	31.53	1
...	...	...	...
16-5-2022 13:00	29.9	30.04	0.1
16-5-2022 14:00	24	23.95	0.0
16-5-2022 15:00	31	30.84	-0.2
16-5-2022 16:00	27	25.67	-1.3
16-5-2022 17:00	32.1	31.82	-0.3

Pada Tabel 4.5 diatas terlihat bahwa perbandingan hasil prediksi dengan hasil actual dengan rasio perbandingan data 70:30 dan *random_state =45* menunjukkan hasil yang cukup akurat dan memuaskan. Ini dibuktikan selisih antara hasil prediksi lebih rendah dari nilai aktual yaitu selisih bernilai 0, meskipun pada *index* hari 2022-04-11 13:00:00 bernilai 1. Adapun pengujian dengan *random_state* 60.

Tabel 4. 6 Hasil pemodelan RF Skenario 2 *random_state =60*

Index	Actual_Value	Predicted_Value	Deviation
11-4-2022 9:00	34	33.68	-0.3
11-4-2022 10:00	30	29.83	-0.2
11-4-2022 11:00	26	25.91	-0.1
11-4-2022 12:00	28	28.30	0.3
11-4-2022 13:00	31	31.60	1
...	...	...	...
16-5-2022 13:00	29.9	30.03	0.1
16-5-2022 14:00	24	23.94	-0.1
16-5-2022 15:00	31	30.88	-0.1
16-5-2022 16:00	27	25.66	-1.3
16-5-2022 17:00	32.1	31.89	-0.2

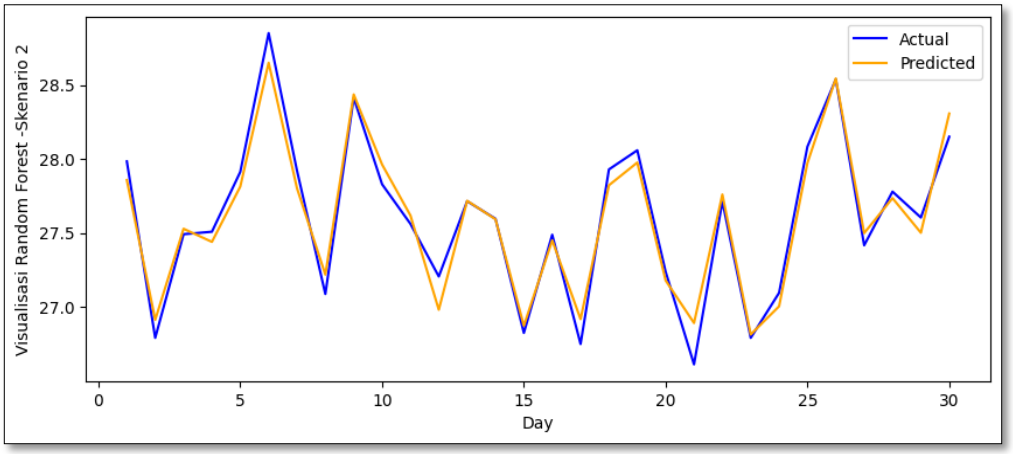
Pengujian dengan *random_state =60* memiliki hasil prediksi yang cukup stabil. Terlihat pada Tabel 4.6 menunjukkan selisih hasil prediksi dan nilai aktual memiliki hasil yang cukup stabil dari pengujian sebelumnya. Hal ini terbukti pada indeks ke 11-4-2022 13:00 yang memiliki nilai deviasi sebesar 1. Adapun hasil pengujian dengan *random_state =75*, ditunjukkan pada Tabel 4.7.

Tabel 4. 7 Hasil pemodelan RF Skenario 2 *random_state =75*

Index	Actual_Value	Predicted_Value	Deviation
11-4-2022 9:00	34	33.70	-0.3
11-4-2022 10:00	30	29.83	-0.2
11-4-2022 11:00	26	25.90	-0.1
11-4-2022 12:00	28	28.29	0.3
11-4-2022 13:00	31	31.50	0.5
...	...	...	...
16-5-2022 13:00	29.9	29.99	0.1
16-5-2022 14:00	24	23.96	0.0
16-5-2022 15:00	31	30.80	-0.2
16-5-2022 16:00	27	25.68	-1.3
16-5-2022 17:00	32.1	31.83	-0.3

Dari Tabel 4.7 diketahui bahwa pengujian dengan *random_state =75* dan rasio 70:30 memiliki hasil prediksi yang cukup stabil jika dilihat dari selisih nilai

antara nilai nyata dan nilai prediksi. Adapun visualisasi hasil pemodelan kualitas udara dengan rasio pembagian data 70:30 yaitu terlihat pada gambar 4.2 berikut.



Gambar 4. 2 Visualisasi hasil pemodelan Skenario 2 *Random Forest*

Visualisasi kualitas udara pada Gambar 4.2 diatas dengan rasio perbandingan data 70:30 atau dari tanggal 11 April hingga 16 Mei 2022 menampilkan perbandingan antara nilai aktual (garis biru) dan nilai prediksi (garis orange) dengan model prediksi cukup baik. Hal ini terbukti dari kedua garis yang mengikuti pola yang sama, meskipun ada beberapa titik yang memiliki perbedaan signifikan, seperti pada hari ke-13 dan rentang hari ke-17 hingga hari ke-21. Adapun hasil *performance* pengujian metode *Random Forest* dengan rasio pembagian data 70:30 dalam melakukan pemodelan kualitas udara telah dirangkum dalam tabel berikut.

Tabel 4. 8 Performance metode *Random Forest* Skenario 2

Random state	RMSE	Accuracy
45	0.65936	94.93%
60	**0.65877**	**94.94%**
75	0.65912	94.94%

Tabel 4.8 memaparkan kinerja *Random Forest* dalam memodelkan kualitas udara Kota Palembang pada Skenario 2. Analisis tabel tersebut menunjukkan bahwa hasil optimal diperoleh melalui konfigurasi khusus metode *Random Forest* dengan *random_state 60*. Konfigurasi ini menghasilkan nilai RMSE terendah sebesar 0,65877 dan tingkat akurasi mencapai 94,94%.

4.1.3 Skenario Pengujian 3

Skenario pengujian ke-3 kualitas udara di Kota Palembang, dilakukan dengan rasio pembagian data training dan data testing yaitu 60:40. Berikut ini hasil uji coba prediksi kualitas udara di Kota Palembang dengan jumlah *random_state* =45.

Tabel 4. 9 Hasil pemodelan RF Skenario 3 *random_state =45*

Index	Actual_Value	Predicted_Value	Deviation
30-3-2022 14:00	34	33.69	-0.3
30-3-2022 15:00	30	29.80	-0.2
30-3-2022 16:00	26	25.89	-0.1
30-3-2022 17:00	28	28.35	0.3
30-3-2022 18:00	31	31.91	1
...	...	...	...
16-5-2022 13:00	34	33.19	-0.8
16-5-2022 14:00	27	26.90	-0.1
16-5-2022 15:00	31	31.35	0.4
16-5-2022 16:00	33	32.19	-0.8
16-5-2022 17:00	29	28.75	-0.3

Berdasarkan Tabel 4.9 diketahui bahwa perbandingan antara data aktual dan hasil prediksi dengan random state=45 menunjukkan tingkat akurasi yang cukup baik. Pada index hari 30-3-2022 16:00 dan 16-5-2022 14:00, hasil prediksi sedikit lebih rendah dari kondisi nyata dengan selisih -0.1. Sebaliknya, pada hari 30-3-2022 18:00, hasil prediksi sedikit lebih tinggi dari kondisi nyata dengan selisih 1. Selain itu, terdapat pengujian dengan *random_state=60*, yaitu sebagai berikut**.**

Tabel 4. 10 Hasil pemodelan RF Skenario 3 *random_state =60*

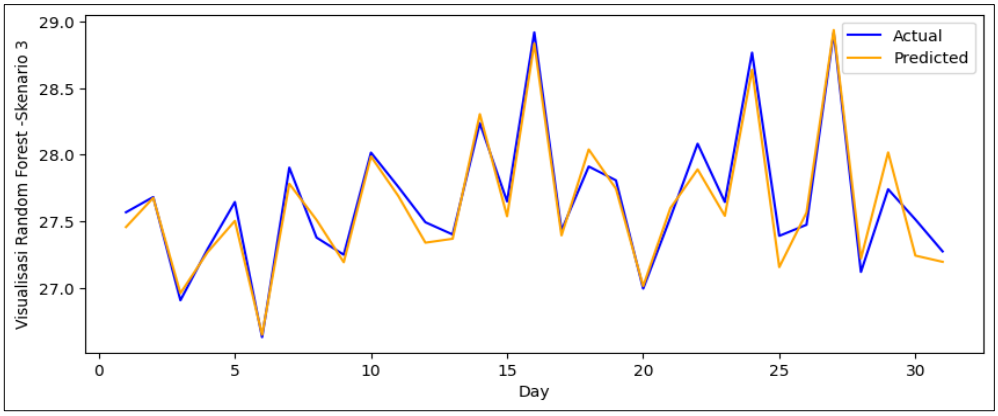
Index	Actual_Value	Predicted_Value	Deviation
30-3-2022 14:00	34	33.67	-0.3
30-3-2022 15:00	30	29.81	-0.2
30-3-2022 16:00	26	25.92	-0.1
30-3-2022 17:00	28	28.27	0.3
30-3-2022 18:00	31	31.93	1
...	...	...	...
16-5-2022 13:00	34	33.21	-0.8
16-5-2022 14:00	27	26.92	-0.1
16-5-2022 15:00	31	31.28	0.3
16-5-2022 16:00	33	32.12	-0.9
16-5-2022 17:00	29	28.76	-0.2

Evaluasi menggunakan *random_state 60*, seperti ditampilkan pada Tabel 4.10, menghasilkan prediksi yang konsisten dan akurat. Hal ini dibuktikan dengan hasil pada tanggal 30-3-2022 pukul 16:00 dan 16-5-2022 pukul 14:00, di mana perbedaan antara nilai prediksi dan aktual tetap sama dengan pengujian sebelumnya, yaitu -1. Sementara itu, hasil pengujian dengan *random_state 75* disajikan dalam Tabel 4.11.

Tabel 4. 11 Hasil pemodelan RF Skenario 3 *random_state =75*

Index	Actual_Value	Predicted_Value	Deviation
30-3-2022 14:00	34	33.67	-0.3
30-3-2022 15:00	30	29.78	-0.2
30-3-2022 16:00	26	25.91	-0.1
30-3-2022 17:00	28	28.29	0.3
30-3-2022 18:00	31	31.85	1
...	...	...	...
16-5-2022 13:00	34	33.19	-0.8
16-5-2022 14:00	27	26.94	-0.1
16-5-2022 15:00	31	31.25	0.3
16-5-2022 16:00	33	32.11	-0.9
16-5-2022 17:00	29	28.70	-0.3

Hasil pengujian pada Tabel 4.11 menunjukkan hasil prediksi yang hampir sama dengan pengujian sebelumnya. Dimana selisih nilai hasil pengujian data aktual dan hasil prediksi memiliki selisih yang cukup rendah seperti pada indeks hari 30-3-2022 16:00 dan 16-5-2022 14:00 yaitu -1. Adapun gambaran pola kualitas udara dengan rasio pembagian data 60:40 yaitu terlihat pada gambar 4.3 berikut.



Gambar 4. 3 Visualisasi hasil pemodelan Skenario 3 *Random Forest*

Pola kualitas udara dari tanggal 30 Maret hingga 16 Mei 2022 pada Gambar 4.3 diatas dengan rasio perbandingan data 60:40 menampilkan perbandingan antara nilai aktual (garis biru) dan nilai prediksi (garis orange) dengan pemodelan cukup baik. Hal ini terbukti dari kedua garis yang mengikuti pola yang sama, meskipun pada hari ke-13, 18, 21, 25, 29 dan 30 terdapat perbedaan yang signifikan.

Adapun performance dari masing-masing pengujian *Random Forest* dalam melakukan pemodelan kualitas udara dengan rasio pembagian data 60:40 ditunjukkan pada tabel berikut ini.

Tabel 4. 12 Performance metode *Random Forest* Skenario 3

Random state	RMSE	Accuracy
45	0.66559	94.84%
60	0.66716	94.82%
75	**0.66376**	**94.87%**

Tabel 4.12 memperlihatkan efektivitas *Random Forest* dalam memodelkan kualitas udara Kota Palembang untuk Skenario 3, menggunakan proporsi data 60:40. Analisis tabel ini mengungkapkan bahwa kinerja optimal *Random Forest* dicapai melalui penyesuaian khusus dengan *random_state 75*. Konfigurasi ini menghasilkan nilai RMSE terendah 0,66376 dan tingkat akurasi tertinggi 94,87%.

4.1.4 Skenario Pengujian 4

Pada skenario pengujian keempat untuk pemodelan kualitas udara Kota Palembang, eksperimen dilakukan dengan membagi data latih dan uji dalam proporsi 50:50. Setiap uji coba menerapkan tiga variasi random_state: 45, 60, dan 75. Berikut ini disajikan hasil evaluasi untuk masing-masing variasi state tersebut.

Tabel 4. 13 Hasil pemodelan RF Skenario 4 *random_state =45*

Index	Actual_Value	Predicted_Value	Deviation
18-3-2022 20:00	34	33.67	-0.3
18-3-2022 21:00	30	29.82	-0.2
18-3-2022 22:00	26	26.01	0.0
18-3-2022 23:00	28	28.36	0.4
19-3-2022 0:00	31	31.97	1
...	...	...	...
16-5-2022 13:00	34	33.42	-0.6
16-5-2022 14:00	28	27.08	-0.9
16-5-2022 15:00	25	25.52	1
16-5-2022 16:00	24	24.11	0.1
16-5-2022 17:00	22.9	24.78	2

Hasil pengujian dengan variasi *random state=45* dan rasio data 50:50 terlihat pada Tabel 4.13 diatas menunjukkan bahwa hasil prediksi kurang baik dibanding pengujian sebelumnya. Hal ini terbukti pada indeks hari 16-5-2022 17:00 selisih hasil prediksi lebih tinggi 2 dari hasil aktual. Tetapi pada indeks yang lain selisih nilai prediksi lebih rendah dari nilai aktual yaitu berkisar antara nilai -1 hingga 0. Adapun pengujian dengan *random_state =60.*

Tabel 4. 14 Hasil pemodelan RF Skenario 4 *random_state =60*

Index	Actual_Value	Predicted_Value	Deviation
18-3-2022 20:00	34	33.62	-0.4
18-3-2022 21:00	30	29.79	-0.2
18-3-2022 22:00	26	25.94	-0.1
18-3-2022 23:00	28	28.34	0.3
19-3-2022 0:00	31	31.83	1
...	...	...	...
16-5-2022 13:00	34	33.41	-0.6
16-5-2022 14:00	28	27.05	-0.9
16-5-2022 15:00	25	25.45	0.5
16-5-2022 16:00	24	24.11	0.1
16-5-2022 17:00	22.9	24.78	2

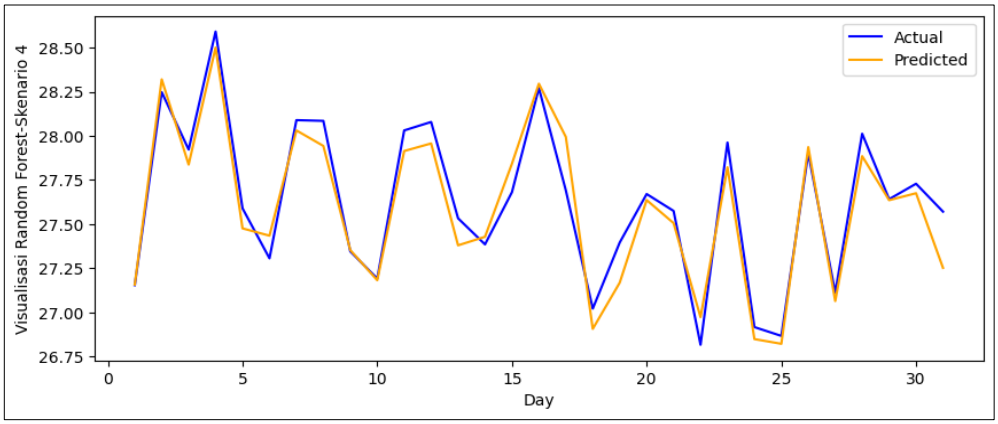
Pengujian dengan *random_state =60* seperti terlihat pada Tabel 4.14 memiliki hasil yang tidak jauh berbeda dengan pengujian *random_state =45* (Tabel 4.13). Ini dibuktikan pada indeks hari 16-5-2022 17:00 selisih nilai prediksi lebih tinggi yaitu 2 dari nilai aktual. Adapun hasil pengujian dengan *random_state =75*, ditunjukkan pada Tabel berikut.

Tabel 4. 15 Hasil pemodelan RF Skenario 4 *random_state =75*

Index	Actual_Value	Predicted_Value	Deviation
18-3-2022 20:00	34	33.64	-0.4
18-3-2022 21:00	30	29.81	-0.2
18-3-2022 22:00	26	25.95	-0.1
18-3-2022 23:00	28	28.34	0.3
19-3-2022 0:00	31	31.96	1
...	...	...	...
16-5-2022 13:00	34	33.44	-0.6
16-5-2022 14:00	28	27.03	-1.0
16-5-2022 15:00	25	25.43	0.4
16-5-2022 16:00	24	24.13	0.1
16-5-2022 17:00	22.9	24.76	2

Berdasarkan hasil pengujian pada Tabel 4.15 diketahui bahwa nilai prediksi dan nilai aktual kualitas udara dengan rasio pembagian data 50:50 dengan

random_state =75 memiliki hasil prediksi yang stabil, dimana selisih antara nilai aktual dan hasil prediksi tidak jauh berbeda pada pengujian dengan *random_state =60* dan *75*. Adapun visualisasi kualitas udara dengan rasio pembagian data 50:50 yaitu terlihat pada gambar 4.4 berikut.



Gambar 4. 4 Visualisasi hasil pemodelan Skenario 4 *Random Forest*

Visualisasi hasil kualitas udara Skenario 4 pada Gambar 4.4 diatas dengan rasio perbandingan data 60:40 atau dari indeks tanggal 18 Maret hingga 16 Mei 2022 menampilkan perbandingan antara nilai aktual (garis biru) dan nilai prediksi (garis orange) dengan model prediksi kurang baik dari sebelumnya. Hal ini terbukti dari kedua garis yang mengikuti pola memiliki perbedaan yang signifikan.

Adapun performance dari masing-masing pengujian *Random Forest* dalam melakukan prediksi kualitas udara dengan rasio pembagian data 50:50 ditunjukkan pada tabel berikut ini.

Tabel 4. 16 Performance metode *Random Forest* Skenario 4

Random state	RMSE	Accuracy
45	0.68834	94.54%
60	0.69051	94.50%
75	**0.68733**	**94.56%**

Hasil analisis Tabel 4.16 menunjukkan bahwa dalam pemodelan kualitas udara Kota Palembang Skenario 4, kinerja optimal dicapai melalui konfigurasi khusus metode *Random Forest* dengan random_state 75. Konfigurasi ini menghasilkan nilai RMSE terendah sebesar 0,68733 dan tingkat akurasi tertinggi 94,56%. Namun, perlu dicatat bahwa pengujian dengan konfigurasi lain juga menunjukkan performa yang sangat baik.

4. 2 Evaluasi Metode Random Forest

Pada fase evaluasi metode *Random Forest*, hasil terbaik dari setiap skenario pemodelan Kota Palembang dipilih untuk dianalisis lebih lanjut dan dibandingkan dengan metode lainnya. Proses pemodelan menggunakan variasi rasio pembagian antara data latih (*training*) dan data uji (*testing*), meliputi proporsi 80:20, 70:30, 60:40, dan 50:50. Selain mempertimbangkan rasio pembagian data, terdapat parameter lain yang dapat memengaruhi performa model, yaitu nilai *random state* (Fatriansyah et al., 2023). Mengubah nilai *random state* dapat mengubah akurasi model. Dalam penelitian ini, variasi jumlah random state yang digunakan yaitu 45, 60, dan 75. Evaluasi terhadap performa terbaik dari masing-masing skenario disajikan dalam Tabel 4.17 berikut ini.

Tabel 4. 17 Performance Metode *Random Forest*

Skenario	Rasio Data	Random state	RMSE	Accuracy
1	**80 : 20**	**45**	**0.63811**	**95.13%**
2	70 : 30	60	0.65877	94.94%
3	60 : 40	75	0.66376	94.87%
4	50 : 50	75	0.68733	94.56%

Berdasarkan Tabel 4.17, pemodelan kualitas udara di Kota Palembang menggunakan metode *Random Forest* memberikan hasil terbaik dengan rasio data pelatihan dan pengujian 80:20 serta menggunakan *random state = 45* dengan hasil RMSE terendah sebesar 0.63811 dan akurasi tertinggi 95.13%. Nilai error yang semakin rendah dan akurasi yang semakin tinggi sangat mempengaruhi efisiensi suatu metode (Gladkova & Saychenko, 2022). Hal ini sejalan dengan pendapat yang dikemukakan (Madhuri et al., 2020), suatu metode dianggap memiliki performa yang lebih baik jika menghasilkan nilai RMSE (Root Mean Squared Error) yang rendah dan akurasi yang tinggi.

4. 3 Pengujian Metode *CatBoost*

Pengujian model *CatBoost* dievaluasi menggunakan beberapa proporsi pembagian data latih dan uji: 80:20, 70:30, 60:40, dan 50:50. Variasi rasio ini diterapkan untuk menilai kinerja model dalam berbagai skenario komposisi data.

Selain itu, juga digunakan tiga nilai *random state* yang berbeda, yaitu 45, 60, dan 75 untuk mengendalikan proses pengacakan dalam pembangunan pohon keputusan (*decision tree*) di dalam *CatBoost*. Proses pemodelan dan evaluasi dilakukan menggunakan bahasa pemrograman Python dan library terkait seperti *CatBoost*, scikit-learn, dan numpy. Berikut potongan *source code* yang digunakan dalam pemodelan kualitas udara Kota Palembang menggunakan metode *CatBoost*.

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.5,random_state =42)
        model = CatBoostRegressor(n_estimators =1000,
                                  learning_rate=0.1,
                                  random_state =75)
model.fit(X_train, y_train)
predict=model.predict(X_test)
```

Potongan *source code* diatas menunjukkan penggunaan metode *CatBoost Regressor* untuk pemodelan kualitas udara. Bagian kode "*test_size=0.5*" diubah sesuai dengan rasio data yang berbeda-beda untuk setiap pengujian. Sementara itu, tulisan merah digunakan untuk membuat instance model *CatBoostRegressor* dengan *hyperparameter n_estimators*=1000 (jumlah *decision tree* yang akan dibangun) dan *random_state =75* (*seed* untuk proses pengacakan dalam membangun pohon keputusan). Parameter "*random_state* " ini berfungsi untuk mengendalikan proses pengacakan dalam pembangunan pohon keputusan (*decision tree*) di dalam *CatBoost*.

4.3.1 Skenario Pengujian 1

Pada skenario pengujian pertama untuk menilai kualitas udara di Kota Palembang menggunakan metode *CatBoost*, dilakukan percobaan dengan membagi data menjadi data latih dan data uji dengan rasio 80:20. Setiap pengujian melibatkan tiga variasi jumlah *random_state* , yaitu 45, 60, dan 75. Tabel 4.1 di bawah ini menunjukkan hasil pengujian pemodelan kualitas udara di Kota Palembang dengan jumlah *random_state = 45*.

Tabel 4. 18 Hasil pemodelan *CatBoost* Skenario 1 *random_state =45*

Index	Actual_Value	Predicted_Value	Deviation
23-4-2022 4:00	34	34.06	0.1
23-4-2022 5:00	30	29.90	-0.1
23-4-2022 6:00	26	25.60	-0.4
23-4-2022 7:00	28	27.91	-0.1
23-4-2022 8:00	31	31.31	0.3
...	...	...	...
16-5-2022 13:00	31	30.98	0.0
16-5-2022 14:00	23	22.93	-0.1
16-5-2022 15:00	24	24.20	0.2
16-5-2022 16:00	27.2	27.53	0.3
16-5-2022 17:00	25	24.98	0.0

Hasil pengujian yang tertera pada Tabel 4.18 menunjukkan perbandingan antara prediksi dan hasil aktual menggunakan metode *CatBoost* dengan rasio pembagian data 80:20 dan *random_state = 45*, yang memberikan hasil yang baik. Hal ini ditunjukkan oleh nilai deviasi antara prediksi dan nilai aktual yang bernilai 0. Pengujian berikutnya dilakukan dengan *random_state=60*.

Tabel 4. 19 Hasil pemodelan *CatBoost* Skenario 1 *random_state =60*

Index	Actual_Value	Predicted_Value	Deviation
23-4-2022 4:00	34	34.28	0.3
23-4-2022 5:00	30	29.93	-0.1
23-4-2022 6:00	26	26.00	0.0
23-4-2022 7:00	28	27.96	0.0
23-4-2022 8:00	31	31.30	0.3
...	...	...	...
16-5-2022 13:00	31	30.95	-0.1
16-5-2022 14:00	23	22.86	-0.1
16-5-2022 15:00	24	24.22	0.2
16-5-2022 16:00	27.2	27.65	0.5
16-5-2022 17:00	25	25.00	0.0

Hasil prediksi yang ditampilkan pada Tabel 4.19 dengan menggunakan rasio data 80:20 dan *random_state 60* menunjukkan hasil yang stabil. Hal ini dibuktikan oleh selisih atau deviasi antara nilai prediksi dan nilai aktual memiliki selisih yang rendah tidak mencapai nilai 1. Selain itu, terdapat juga pengujian dengan *random_state 75* yang ditampilkan pada Tabel 4.20.

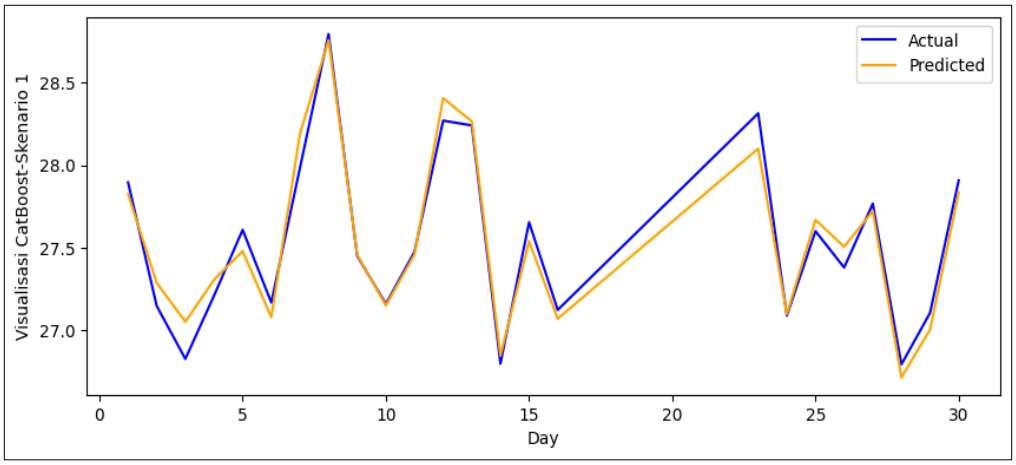
Tabel 4. 20 Hasil pemodelan *CatBoost* Skenario 1 *random_state =75*

Index	Actual_Value	Predicted_Value	Deviation
23-4-2022 4:00	34	34.03	0.0
23-4-2022 5:00	30	29.72	-0.3
23-4-2022 6:00	26	25.83	-0.2
23-4-2022 7:00	28	27.83	-0.2
23-4-2022 8:00	31	31.34	0.3
...	...	...	...

Index	Actual_Value	Predicted_Value	Deviation
16-5-2022 13:00	31	31.06	0.1
16-5-2022 14:00	23	23.04	0.0
16-5-2022 15:00	24	24.19	0.2
16-5-2022 16:00	27.2	27.26	0.1
16-5-2022 17:00	25	24.87	-0.1

Tabel 4.20 menunjukkan hasil pengujian yang dilakukan dengan menggunakan *random_state =75*. Pengujian menunjukkan hasil yang hampir sama seperti pengujian sebelumnya. Hal ini terlihat dari nilai deviasi antara nilai nyata dan hasil prediksi. Meskipun terdapat beberapa selisih hasil prediksi, tetapi kurang dari 1.

Adapun visualisasi hasil pemodelan kualitas udara menggunakan metode *CatBoost* dengan rasio pembagian data 80:20 yaitu terlihat pada gambar 4.2 berikut.



Gambar 4. 5 Visualisasi hasil pemodelan Skenario 1 *CatBoost*

Visualisasi kualitas udara pada Gambar 4.5 diatas menunjukkan perbandingan antara nilai aktual (garis biru) dan nilai prediksi (garis orange) untuk kualitas udara dengan rasio perbandingan data 80:20 atau dari indeks tanggal 23 April hingga 16 Mei 2022. Terlihat perbedaan yang sangat signifikan antara nilai aktual dan nilai prediksi pada hari ke-3, 13, dan rentang hari ke-16 hingga hari ke-

23, terdapat pula beberapa titik yang memiliki perbedaan, namun tidak terlalu signifikan dibanding beberapa hari diatas. Performa pengujian metode *CatBoost* dalam melakukan pemodelan kualitas udara dengan rasio pembagian data 80:20 dapat dilihat pada tabel 4.21 berikut ini.

Tabel 4. 21 Performance metode *CatBoost* Skenario 1

Random state	RMSE	Accuracy
45	0.52934	96.65%
60	0.52393	96.72%
75	**0.52019**	**96.76%**

Berdasarkan Tabel 4.21, pada pemodelan kualitas udara di Kota Palembang menggunakan metode *CatBoost* pada Skenario 1 dengan rasio pembagian data pelatihan dan data pengujian 80:20 menunjukkan performa terbaik dicapai pada modifikasi *random_state = 75*, yang menghasilkan nilai RMSE (*Root Mean Squared Error*) terendah sebesar 0.52019 dan akurasi tertinggi mencapai 96.76%.

4.3.2 Skenario Pengujian 2

Skenario pengujian kedua untuk pemodelan kualitas udara di Kota Palembang menggunakan metode *CatBoost* dilakukan dengan rasio pembagian data latih dan data uji 70:30. Berikut adalah hasil uji coba pemodelan kualitas udara di Kota Palembang dengan *random_state* = 45.

Tabel 4. 22 Hasil pemodelan *CatBoost* Skenario 2 *random_state = 45*

Index	Actual_Value	Predicted_Value	Deviation
11-4-2022 9:00	34	33.82	-0.2
11-4-2022 10:00	30	29.77	-0.2
11-4-2022 11:00	26	25.68	-0.3
11-4-2022 12:00	28	27.74	-0.3
11-4-2022 13:00	31	31.16	0.2
...	...	...	...

Index	Actual_Value	Predicted_Value	Deviation
16-5-2022 13:00	29.9	30.04	0.1
16-5-2022 14:00	24	23.63	-0.4
16-5-2022 15:00	31	31.21	0.2
16-5-2022 16:00	27	25.67	-1.3
16-5-2022 17:00	32.1	31.92	-0.2

Pengujian dengan *random_state =45* terlihat pada Tabel 4.22 menghasilkan nilai prediksi yang cukup akurat dan memuaskan. Ini dibuktikan selisih antara hasil prediksi lebih rendah dari nilai aktual bernilai -1 pada indeks hari 16-5-2022 16:00 dan bernilai 0 pada indeks lainnya. Adapun pengujian dengan *random_state =60*, ditunjukkan pada Tabel 4.23.

Tabel 4. 23 Hasil pemodelan *CatBoost* Skenario 2 *random_state=60*

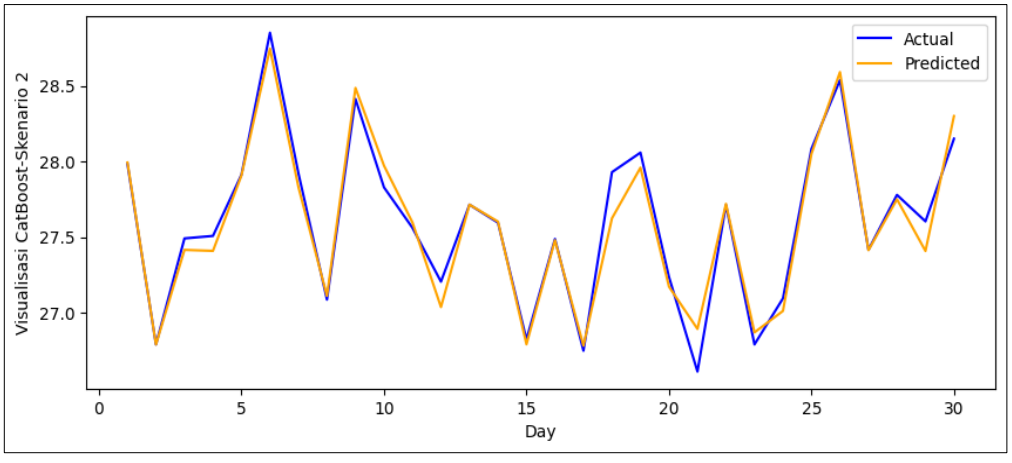
Index	Actual_Value	Predicted_Value	Deviation
11-4-2022 9:00	34	34.05	0.0
11-4-2022 10:00	30	29.84	-0.2
11-4-2022 11:00	26	26.25	0.3
11-4-2022 12:00	28	27.60	-0.4
11-4-2022 13:00	31	31.24	0.2
...	...	...	...
16-5-2022 13:00	29.9	30.30	0.4
16-5-2022 14:00	24	23.82	-0.2
16-5-2022 15:00	31	31.16	0.2
16-5-2022 16:00	27	26.00	-1.0
16-5-2022 17:00	32.1	31.55	-0.5

Adapun pengujian dengan random state=60 seperti yang ditampilkan pada Tabel 4.23 diketahui bahwa perbandingan antara data aktual dan hasil prediksi menunjukkan tingkat akurasi yang baik dan stabil. Hal ini terlihat bahwa terdapat beberapa indeks data yang memiliki selisih hasil prediksi lebih rendah dari hasil nyata dengan nilai -1 seperti pada indeks hari 16-5-2022 16:00. Selain itu terdapat pengujian dengan random state=75 seperti ditunjukkan pada Tabel 4.24 berikut ini**.**

Tabel 4. 24 Hasil pemodelan *CatBoost* Skenario 2 *random_state = 75*

Index	Actual_Value	Predicted_Value	Deviation
11-4-2022 9:00	34	33.95	-0.1
11-4-2022 10:00	30	29.72	-0.3
11-4-2022 11:00	26	25.77	-0.2
11-4-2022 12:00	28	27.57	-0.4
11-4-2022 13:00	31	31.20	0.2
...	...	...	...
16-5-2022 13:00	29.9	30.55	1
16-5-2022 14:00	24	23.61	-0.4
16-5-2022 15:00	31	31.17	0.2
16-5-2022 16:00	27	25.74	-1.3
16-5-2022 17:00	32.1	31.84	-0.3

Berdasarkan Tabel 4.24, pada pengujian dengan menggunakan *random_state =75*, hasil prediksi menunjukkan stabilitas yang cukup baik jika dilihat dari selisih antara nilai prediksi dengan nilai aktual. Meskipun terdapat sedikit peningkatan nilai deviasi pada indeks hari 16-5-2022 13:00 yaitu bernilai 1. Namun secara keseluruhan selisih nilai antara aktual dan prediksi relatif kecil dan stabil. Adapun visualisasi hasil pemodelan kualitas udara dengan rasio pembagian data 70:30 yaitu terlihat pada gambar 4.2 berikut.



Gambar 4. 6 Visualisasi hasil pemodelan Skenario 2 *CatBoost*

Berdasarkan Gambar 4.6 visualisasi hasil kualitas udara Skenario 2 diatas dengan rasio perbandingan data 70:30 menampilkan perbandingan antara nilai aktual (garis biru) dan nilai prediksi (garis orange) dengan model prediksi cukup baik. Hal ini terbukti dari kedua garis antara nilai aktual dan nilai prediksi yang mengikuti pola memiliki persamaan yang cukup baik. Meskipun terdapat perbedaan yang besar pada rentang indeks hari ke-18 hingga hari ke-21. Terdapat juga perbedaan yang disignifikan pada hari ke-28 hingga ke-29. Adapun performance dari masing-masing pengujian metode *CatBoost* dalam melakukan pemodelan kualitas udara dengan rasio pembagian data 70:30 ditampilkan pada tabel 4.25

Tabel 4. 25 Performance metode *CatBoost* Skenario 2

Random state	RMSE	Accuracy
45	**0.55634**	**96.39%**
60	0.56098	96.33%
75	0.56692	96.25%

Berdasarkan Tabel 4.25, pada pemodelan kualitas udara di Kota Palembang menggunakan metode *CatBoost* pada Skenario 2 dengan rasio pembagian data 70:30 menunjukkan performa terbaik diperoleh ketika menggunakan *random_state = 45*. Pada kondisi ini, model menghasilkan nilai RMSE *(Root Mean Squared Error)* terendah sebesar 0.55634 dan akurasi tertinggi mencapai 96.39%.

4.3.3 Skenario Pengujian 3

Pada skenario pengujian ketiga, model kualitas udara di Kota Palembang menggunakan metode *CatBoost* dengan rasio pembagian data latih dan data uji 60:40. Berikut adalah hasil uji coba prediksi kualitas udara di Kota Palembang dengan *random_state = 45*.

Tabel 4. 26 Hasil pemodelan *CatBoost* Skenario 3 *random_state = 45*

Index	Actual_Value	Predicted_Value	Deviation
30-3-2022 14:00	34	33.83	-0.2
30-3-2022 15:00	30	29.65	-0.3
30-3-2022 16:00	26	26.08	0.1
30-3-2022 17:00	28	27.66	-0.3
30-3-2022 18:00	31	31.66	1
...	...	...	...
16-5-2022 13:00	34	33.88	-0.1
16-5-2022 14:00	27	26.79	-0.2
16-5-2022 15:00	31	31.73	1
16-5-2022 16:00	33	32.26	-0.7
16-5-2022 17:00	29	28.85	-0.2

Berdasarkan hasil pengujian pada Tabel 4.26 dengan menggunakan random_state 45, terlihat hasil prediksinya cukup memuaskan. Ini terbukti dari deviasi nilai prediksi yang relatif kecil dibandingkan dengan nilai aktual, meskipun pada indeks 30-3-2022 18:00 dan 16-5-2022 15:00 tampak selisih lebih besar yaitu deviasi bernilai 1. Selain itu, terdapat pengujian dengan random_state =60 seperti terlihat pada Tabel 4.27**.**

Tabel 4. 27 Hasil pemodelan *CatBoost* Skenario 3 *random_state=60*

Index	Actual_Value	Predicted_Value	Deviation
30-3-2022 14:00	34	34.04	0.0
30-3-2022 15:00	30	29.67	-0.3
30-3-2022 16:00	26	26.27	0.3
30-3-2022 17:00	28	27.93	-0.1
30-3-2022 18:00	31	31.87	1
...	...	...	...
16-5-2022 13:00	34	33.98	0.0
16-5-2022 14:00	27	26.75	-0.3
16-5-2022 15:00	31	31.71	1
16-5-2022 16:00	33	32.00	-1.0
16-5-2022 17:00	29	29.06	0.1

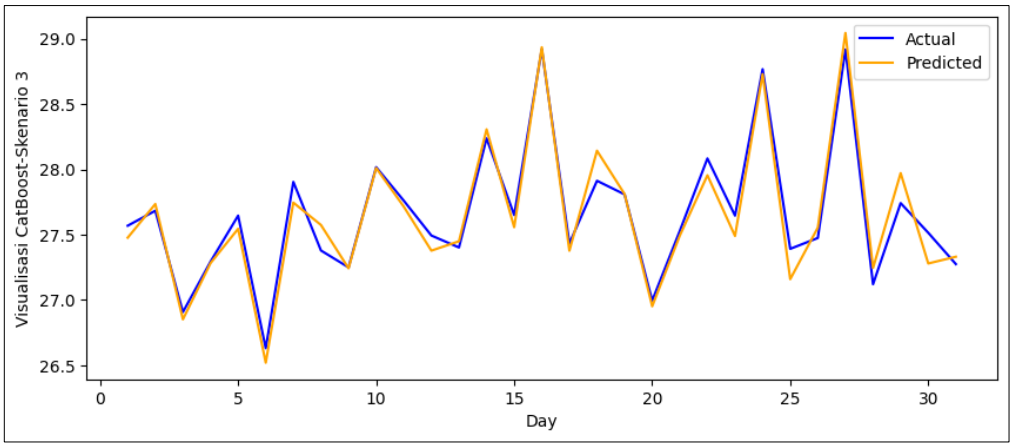
Pengujian dengan *random_state =60* yang tercatat dalam Tabel 4.27 menghasilkan nilai prediksi yang cukup stabil. Hal ini terbukti pada indeks 30-3-

2022 18:00 dan 16-5-2022 15:00 menunjukkan deviasi hampir sama dari pengujian sebelumnya yaitu sebesar 1 antara hasil prediksi dengan kondisi nyata. Adapun hasil pengujian dengan *random_state =75*, ditunjukkan pada Tabel 4.28.

Tabel 4. 28 Hasil pemodelan *CatBoost* Skenario 3 *random_state = 75*

Index	Actual_Value	Predicted_Value	Deviation
30-3-2022 14:00	34	34.17	0.2
30-3-2022 15:00	30	29.56	-0.4
30-3-2022 16:00	26	25.82	-0.2
30-3-2022 17:00	28	27.88	-0.1
30-3-2022 18:00	31	31.54	1
...	...	...	...
16-5-2022 13:00	34	34.02	0.0
16-5-2022 14:00	27	26.71	-0.3
16-5-2022 15:00	31	31.62	1
16-5-2022 16:00	33	31.90	-1.1
16-5-2022 17:00	29	29.04	0.0

Hasil pengujian dengan *random_state =75* pada Tabel 4.28 diatas terlihat menunjukkan hasil prediksi yang cukup baik. Ini terbukti dari selisih hasil prediksi dengan nilai nyata menunjukkan nilai deviasi yang stabil. Adapun visualisasi hasil pemodelan kualitas udara dengan rasio pembagian data 60:40 yaitu terlihat pada gambar 4.7 berikut.



Gambar 4. 7 Visualisasi hasil pemodelan Skenario 3 *CatBoost*

Gambar 4.7 menunjukkan perbandingan visual kualitas udara antara 30 Maret hingga 16 Mei 2022, menggunakan rasio data 60:40. Plot ini membandingkan nilai aktual (ditunjukkan oleh garis biru) dengan nilai prediksi model (ditampilkan oleh garis orange). Model prediksi menunjukkan kinerja yang cukup baik, terlihat dari kecenderungan kedua garis yang mengikuti pola serupa. Namun, terdapat perbedaan yang cukup signifikan dibeberapa titik waktu tertentu, terutama pada hari ke-8, 13, 18, 19, 25, 29, dan 30. Perbedaan-perbedaan ini menunjukkan area di mana akurasi prediksi model masih bisa ditingkatkan.

Adapun performa dari masing-masing pengujian metode *CatBoost* dalam melakukan pemodelan kualitas udara dengan rasio pembagian data 60:40 ditampilkan pada tabel berikut ini.

Tabel 4. 29 Performance metode *CatBoost* Skenario 3

Random state	RMSE	Accuracy
45	0.58135	96.06%
60	**0.57505**	**96.15%**
75	0.58626	96.00%

Tabel 4.29 menunjukkan performa pada pemodelan kualitas udara di Kota Palembang menggunakan metode *CatBoost* pada Skenario 3 dengan rasio pembagian data 60:40. Pada skenario ini kinerja terbaik diperoleh ketika menggunakan *random_state=60*, dimana model menghasilkan nilai RMSE terendah sebesar 0.57505 dan akurasi tertinggi mencapai 96.15%.

4.3.4 Skenario Pengujian 4

Pada skenario pengujian keempat, model kualitas udara di Kota Palembang menggunakan metode *CatBoost* dengan rasio pembagian data latih dan data uji

50:50. Berikut adalah hasil uji coba prediksi kualitas udara di Kota Palembang dengan *random_state = 45*.

Tabel 4. 30 Hasil pemodelan *CatBoost* Skenario 4 *random_state = 45*

Index	Actual_Value	Predicted_Value	Deviation
18-3-2022 20:00	34	33.98	0.0
18-3-2022 21:00	30	29.85	-0.1
18-3-2022 22:00	26	25.70	-0.3
18-3-2022 23:00	28	28.06	0.1
19-3-2022 0:00	31	31.39	0.4
...	...	...	...
16-5-2022 13:00	34	33.61	-0.4
16-5-2022 14:00	28	27.43	-0.6
16-5-2022 15:00	25	25.28	0.3
16-5-2022 16:00	24	24.14	0.1
16-5-2022 17:00	22.9	23.88	1.0

Hasil pengujian dengan variasi *random state=45* dan rasio data 50:50 terlihat pada Tabel 4.30 diatas menunjukkan bahwa hasil prediksi cukup baik. Hal ini terbukti pada indeks hari 16-5-2022 17:00, selisih nilai prediksi lebih tinggi yaitu 1 dari nilai aktual. Adapun pengujian dengan *random_state =60*.

Tabel 4. 31 Hasil pemodelan *CatBoost* Skenario 4 *random_state=60*

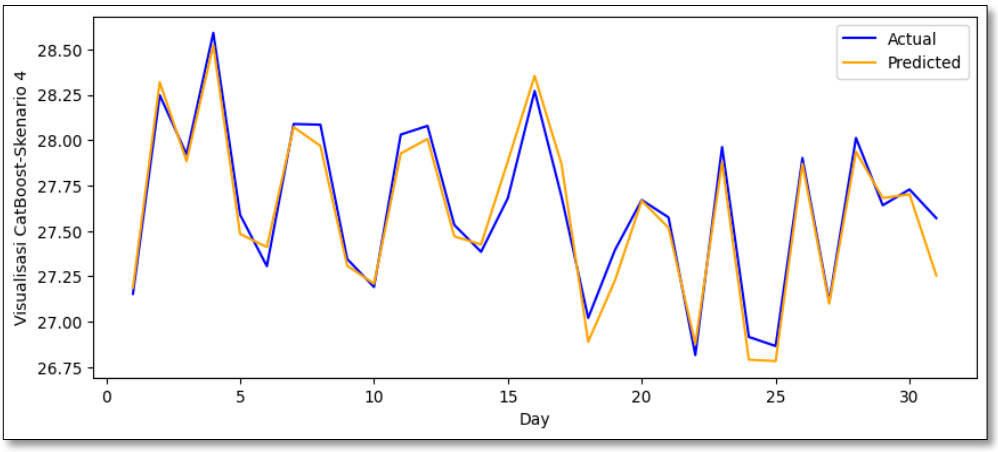
Index	Actual_Value	Predicted_Value	Deviation
18-3-2022 20:00	34	33.94	-0.1
18-3-2022 21:00	30	29.90	-0.1
18-3-2022 22:00	26	25.86	-0.1
18-3-2022 23:00	28	27.79	-0.2
19-3-2022 0:00	31	31.70	1
...	...	...	...
16-5-2022 13:00	34	34.06	0.1
16-5-2022 14:00	28	27.58	-0.4
16-5-2022 15:00	25	25.09	0.1
16-5-2022 16:00	24	24.15	0.1
16-5-2022 17:00	22.9	24.18	1.3

Pengujian dengan *random_state =60* seperti terlihat pada Tabel 4.31 memiliki hasil yang tidak jauh berbeda dengan pengujian *random_state =45* (Tabel 4.30). Meskipun terdapat sedikit peningkatan nilai deviasi antara hasil prediksi dengan nilai aktual pada indeks hari 19-3-2022 0:00 dan 16-5-2022 17:00 dengan nilai sebesar 1. Namun secara keseluruhan selisih nilai antara aktual dan prediksi relatif kecil dan stabil. Adapun hasil pengujian dengan *random_state =75*, ditunjukkan pada Tabel berikut.

Tabel 4. 32 Hasil pemodelan *CatBoost* Skenario 3 *random_state =75*

Index	Actual_Value	Predicted_Value	Deviation
18-3-2022 20:00	34	33.90	-0.1
18-3-2022 21:00	30	29.78	-0.2
18-3-2022 22:00	26	25.73	-0.3
18-3-2022 23:00	28	27.98	0.0
19-3-2022 0:00	31	31.73	1
...	...	...	...
16-5-2022 13:00	34	33.74	-0.3
16-5-2022 14:00	28	27.29	-0.7
16-5-2022 15:00	25	25.05	0.1
16-5-2022 16:00	24	24.06	0.1
16-5-2022 17:00	22.9	23.94	1.0

Dari hasil pengujian pada Tabel 4.32 diketahui bahwa nilai prediksi dan nilai aktual kualitas udara dengan rasio pembagian data 50:50 dengan *random_state =75* memiliki hasil prediksi yang stabil, dimana selisih antara nilai aktual dan hasil prediksi tidak jauh berbeda pada pengujian dengan *random_state =60 dan 75*. Adapun visualisasi hasil pemodelan kualitas udara dengan rasio pembagian data 50:50 yaitu terlihat pada gambar 4.8 berikut.



Gambar 4. 8 Visualisasi hasil pemodelan Skenario 4 *CatBoost*

Gambar 4.8 menampilkan visualisasi hasil pemodelan kualitas udara untuk Skenario 4, menggunakan metode *Catboost* dengan pembagian data sama rata (50:50). Grafik ini mencakup periode dari 18 Maret hingga 16 Mei 2022, membandingkan nilai aktual (ditunjukkan oleh garis biru) dengan nilai prediksi model (garis orange). Berbeda dengan hasil sebelumnya, model prediksi pada skenario ini menunjukkan performa yang kurang memuaskan. Hal ini terlihat jelas dari adanya perbedaan yang cukup besar antara garis aktual dan prediksi di sepanjang periode, mengindikasikan bahwa model ini kurang akurat dalam memprediksi kualitas udara dibandingkan dengan skenario-skenario sebelumnya. Adapun performance dari masing-masing pengujian metode *CatBoost* dalam melakukan pemodelan kualitas udara dengan rasio pembagian data 50:50 ditunjukkan pada tabel berikut ini.

Tabel 4. 33 Performance metode *CatBoost* Skenario 4

Random state	RMSE	Accuracy
45	0.61118	95.69%
60	**0.60233**	**95.82%**
75	0.60590	95.77%

Tabel 4.33 menunjukkan performa pada pemodelan kualitas udara di Kota Palembang menggunakan metode *CatBoost* pada Skenario 4 dengan rasio pembagian data 50:50. Pada skenario ini kinerja terbaik diperoleh ketika menggunakan *random_state=60*, dimana model menghasilkan nilai RMSE terendah sebesar 0.60233 dan akurasi tertinggi mencapai 95.82%.

4. 4 Evaluasi Metode *CatBoost*

Dalam tahap evaluasi pengujian metode *CatBoost*, hasil pemodelan dengan performa terbaik dari masing-masing skenario pemodelan Kota Palembang diambil sebagai bahan analisis perbandingan dengan metode *Random Forest*. Dalam proses pemodelan, rasio pembagian data antara data pelatihan (*training*) dan data pengujian (*testing*) divariasikan dengan menggunakan beberapa rasio, yaitu 80:20, 70:30, 60:40, dan 50:50. Selain mempertimbangkan rasio pembagian data, terdapat parameter lain yang dapat memengaruhi performa model, yaitu nilai *random state* (Fatriansyah et al., 2023). Mengubah nilai *random state* dapat mengubah akurasi model. Dalam penelitian ini, variasi jumlah random state yang digunakan yaitu 45, 60, dan 75. Evaluasi terhadap performa terbaik dari masing-masing skenario disajikan dalam Tabel 4.34 berikut ini.

Tabel 4. 34 Performance Metode *CatBoost*

Skenario	Rasio Data	Random state	RMSE	Accuracy
1	**80 : 20**	**75**	**0.52019**	**96.76%**
2	70 : 30	45	0.55634	96.39%
3	60 : 40	60	0.57505	96.15%
4	50 : 50	60	0.60233	95.82%

Berdasarkan Tabel 4.34, prediksi kualitas udara di Kota Palembang menggunakan metode *CatBoost* memberikan hasil terbaik dengan rasio data pelatihan dan pengujian 80:20 serta menggunakan *random state = 75*. Dalam kasus ini, skenario dengan RMSE terendah sebesar 0.52019 dan akurasi tertinggi 96.76% dianggap sebagai hasil terbaik. Nilai error yang semakin rendah dan akurasi yang semakin tinggi sangat mempengaruhi efisiensi suatu metode (Gladkova & Saychenko, 2022). Sejalan dengan pendapat yang dikemukakan (Madhuri et al., 2020), suatu metode dianggap memiliki performa yang lebih baik jika menghasilkan nilai RMSE (Root Mean Squared Error) yang rendah dan akurasi yang tinggi.

4. 5 Pembahasan

Penelitian ini menerapkan algoritma *Random Forest* dan *CatBoost* dalam pemodelan kualitas udara Kota Palembang. Kedua metode *machine learning* ini telah dikenal luas dan menunjukkan efektivitas dalam berbagai kasus klasifikasi dan regresi. Meski demikian, setiap metode memiliki ciri khas dan pendekatan unik dalam pembentukan model. Oleh karena itu, diperlukan analisis evaluasi yang komprehensif terhadap hasil eksperimen dari kedua metode ini, dengan mempertimbangkan berbagai skenario dan modifikasi yang telah diimplementasikan. Evaluasi performa kedua metode tersebut dilakukan dengan merujuk pada metrik utama, yaitu akurasi dan *Root Mean Squared Error* (RMSE). Berikut hasil pengujian dari metode *Random Forest* dan *CatBoost*.

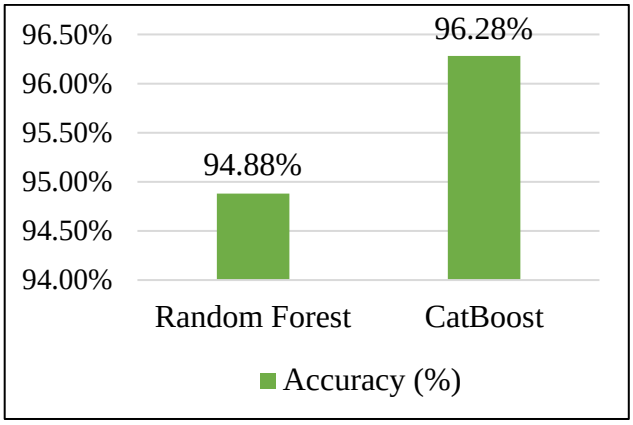
Tabel 4. 35 Hasil Pengujian Metode *Random Forest* dan *CatBoost*

Skenario (Rasio Data)	***Random Forest***			***CatBoost***		
	Random state	RMSE	Accuracy	Random state	RMSE	Accuracy
1 (80:20)	45	0.63811	95.13%	75	0.52019	96.76%
2 (70:30)	60	0.65877	94.94%	45	0.55634	96.39%
3 (60:40)	75	0.66376	94.87%	60	0.57505	96.15%
4 (50:50)	75	0.68733	94.56%	60	0.60233	95.82%
	Average	0.66199	94.88%	Average	0.56348	96.28%

Hasil pengujian yang tercantum pada Tabel 4.35 merupakan hasil terbaik dari setiap skenario pengujian. Dimana dalam pengujian metode *Random Forest* diterapkan 4 jenis skenario data yaitu 80:20, 70:30, 60:40,50:50, serta 3 jenis modifikasi parameter *random state* yang digunakan diataranya 45, 60, dan 75. Dalam pengujian tersebut diperoleh akurasi tertinggi adalah 95.13% dan RMSE 0.63811 dengan rasio perbandingan data 80:20 dan random state 45.

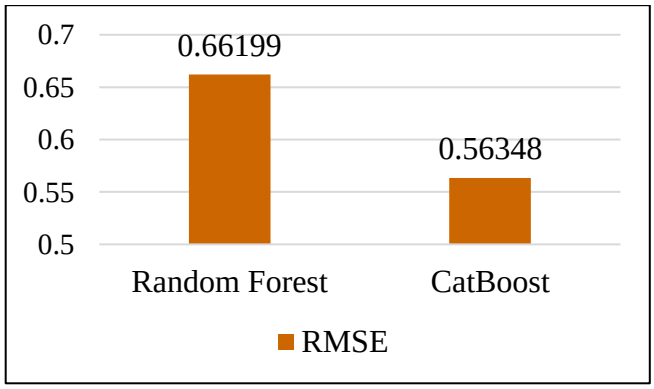
Sedangkan dalam pengujian metode *CatBoost* juga menggunakan 4 jenis skenario data yaitu 80:20, 70:30, 60:40,50:50, serta 3 jenis modifikasi parameter *random state* yang digunakan diataranya 45, 60, dan 75. Dari hasil pengujian diperoleh akurasi terbaik sebesar 96.76% dan RMSE terendah 0.52019 dengan rasio perbandingan data 80:20 dan *random_state =75*..

Berdasarkan hasil evaluasi komparatif antara *Random Forest* dan *CatBoost*, seperti terlihat pada Tabel 4.35, menghasilkan nilai rata-rata akurasi dan RMSE untuk setiap skenario optimal dari kedua metode. Berikut visualisasi yang menggambarkan perbandingan tingkat akurasi kedua metode tersebut.



Gambar 4. 9 Perbandingan hasil akurasi *Random Forest* dan *CatBoost*

Perbandingan hasil akurasi metode *Random Forest* dan *CatBoost* pada Gambar 4.9 terlihat bahwa metode *CatBoost* mengungguli metode *Random Forest* dalam hal akurasi pada pemodelan kualitas udara di Kota Palembang. Metode *CatBoost* menunjukkan performa yang lebih baik dengan tingkat akurasi 96,28%, sementara *Random Forest* mencapai akurasi 94,88%. Selisih akurasi antara kedua metode tidak terlalu signifikan, hanya sekitar 1,4%. Meskipun demikian, kedua metode tetap menampilkan akurasi yang sangat tinggi, melebihi 90%, mengindikasikan bahwa kedua metode tersebut mampu memberikan performa yang memuaskan dalam pemodelan kualitas udara di Kota Palembang. Adapun visualisasi perbandingan antara metode *Random Forest* dan *CatBoost* berdasarkan nilai RMSE.



Gambar 4. 10 Perbandingan hasil RMSE metode *Random Forest* dan *CatBoost*

Dari Gambar 4.10 diatas menunjukkan bahwa metode *CatBoost* memiliki nilai RMSE yang lebih rendah (0.56348) dibandingkan dengan metode *Random Forest* (0.66199), yang mengindikasikan bahwa metode *CatBoost* lebih baik dalam melakukan pemodelan kualitas udara di Kota Palembang karena menghasilkan kesalahan prediksi yang lebih kecil.

4. 6 Pemodelan Kualitas Udara menurut Pandangan Islam

Udara yang bersih dan sehat adalah salah satu nikmat besar dari Allah yang sering kali diabaikan oleh manusia. Namun, seiring dengan perkembangan aktivitas manusia yang semakin masif, kualitas udara sering terancam oleh pencemaran. Dalam pandangan Islam, menjaga kelestarian lingkungan dan meminimalisir kerusakan di muka bumi merupakan salah satu kewajiban bagi setiap muslim. Pencemaran udara yang disebabkan oleh aktivitas manusia telah menjadi masalah global yang serius dan memerlukan upaya penanganan yang tepat. Salah satu upaya yang dapat dilakukan adalah dengan melakukan pemodelan kualitas udara yang bertujuan untuk mengantisipasi dan mengurangi dampak buruk dari pencemaran udara. Dalam Al-Qur'an Surah Al-A'raf ayat 56, Allah berfirman:

وَلَا تُفْسِدُوْا فِى الْاَرْضِ بَعْدَ اِصْلَاحِهَا وَادْعُوْهُ خَوْفًا وَّطَمَعًاۗ اِنَّ رَحْمَتَ اللّٰهِ قَرِيْبٌ مِّنَ الْمُحْسِنِيْنَ

Artinya: *"Dan janganlah kamu membuat kerusakan di muka bumi, sesudah (Allah) memperbaikinya dan berdoalah kepada-Nya dengan rasa takut (tidak akan diterima) dan harapan (akan dikabulkan). Sesungguhnya rahmat Allah amat dekat kepada orang-orang yang berbuat baik." (QS. Al-A'raf: 56).*

Ayat ini menegaskan larangan untuk melakukan kerusakan di muka bumi, termasuk pencemaran udara yang dapat mengganggu keseimbangan ekosistem dan merugikan makhluk hidup lainnya. Sebagaimana dijelaskan dalam Tafsir

Departemen Agama Jilid IV halaman 218, “Larangan berbuat kerusakan di muka bumi adalah mencakup semua bidang kehidupan”.

Selain itu, dalam pemodelan kualitas udara, kerjasama dan kolaborasi antara berbagai pihak, seperti pemerintah, akademisi, dan masyarakat, sangat penting. Hal ini sejalan dengan prinsip persatuan dan kerjasama dalam Islam, seperti dijelaskan dalam Al-Qur’an,

وَاعْتَصِمُوْا بِحَبْلِ اللّٰهِ جَمِيْعًا وَّلَا تَفَرَّقُوْاۚ

Artinya: *“Dan berpeganglah kamu semuanya kepada tali (agama) Allah, dan janganlah kamu bercerai berai.” (QS. Ali Imran: 103).*

Dalam tafsir Jalalain, ayat diatas menganjurkan untuk bersatu dan saling bekerjasama dalam semua hal kebaikan (Al-Mahali & Jalaludin, 2003), khususnya dalam mengatasi polusi udara. Melalui pemodelan kualitas udara yang akurat dan efektif, kita dapat mengidentifikasi area-area yang membutuhkan perhatian khusus dan mengambil tindakan yang tepat untuk mengurangi polusi udara. Hal ini sejalan dengan prinsip Islam yang menganjurkan manusia untuk selalu berusaha memperbaiki diri dan lingkungannya. Allah berfirman,

اِنَّ اللّٰهَ لَا يُغَيِّرُ مَا بِقَوْمٍ حَتّٰى يُغَيِّرُوْا مَا بِاَنْفُسِهِمْۗ

Artinya: “Sesungguhnya Allah tidak akan mengubah keadaan suatu kaum sebelum mereka mengubah keadaan diri mereka sendiri. (QS. Ar-Ra’d: 11).

Tafsir Departemen Agama Republik Indonesia Jilid 4 menjelaskan bahwa ayat ini menganjurkan manusia untuk berusaha memperbaiki diri dan lingkungannya sebelum Allah memberikan perubahan. Dengan pemodelan kualitas udara yang baik, kita dapat memastikan bahwa udara yang kita hirup sehat dan

bersih. Hal ini sangat penting dalam Islam, karena kesehatan dan kebersihan merupakan aspek penting dalam ibadah. Rasulullah bersabda,

اِنَّ اللهَ طَيِّبٌ يُحِبُّ الطَّيِّبَ نَظِيْفٌ يُحِبُّ النَّظَافَةَ

Artinya: *"Sesungguhnya Allah itu baik dan menyukai kebaikan, bersih dan menyukai kebersihan." (HR. Tirmidzi).*

Dengan udara yang bersih dan sehat, kita dapat melaksanakan ibadah dengan lebih khusyuk dan nyaman. Dengan demikian, pemodelan kualitas udara tidak hanya bermanfaat secara lingkungan, tetapi juga memiliki dampak positif pada kehidupan spiritual dan ibadah kita. Dengan menjaga lingkungan dan udara yang bersih, kita telah menjalankan perintah Allah untuk menjaga kelestarian bumi dan menjaga kesehatan diri sendiri. Selain itu, pemodelan kualitas udara juga membantu kita untuk menjaga keseimbangan alam dan menghindari perilaku yang merusak. Melalui upaya ini, kita turut berperan aktif dalam melestarikan kehidupan di bumi sebagaimana yang diperintahkan oleh agama Islam.

BAB V

KESIMPULAN

5. 1 Kesimpulan

Berdasarkan hasil analisis performa metode *Random Forest* dan *CatBoost* dalam pemodelan kualitas udara di Kota Palembang, dapat ditarik kesimpulan bahwa metode *CatBoost* memiliki performa yang sedikit lebih baik dibandingkan dengan *Random Forest* dalam pemodelan kualitas udara di Kota Palembang. Hal ini dapat dilihat dari dua metrik evaluasi utama, yaitu akurasi dan *Root Mean Squared Error* (RMSE).

Dalam hal akurasi, *CatBoost* memiliki akurasi yang lebih tinggi, mencapai 96.28%, sedangkan *Random Forest* hanya mencapai 94,88%. Meskipun perbedaannya tidak terlalu signifikan, yaitu 1,4%, namun *CatBoost* tetap unggul dalam menghasilkan model yang lebih akurat. Selanjutnya, dari segi RMSE yang mengukur kesalahan model, *CatBoost* juga mengungguli *Random Forest* dengan nilai RMSE yang lebih rendah, yaitu 0.56348 berbanding 0.66199. Nilai RMSE yang lebih rendah mengindikasikan bahwa *CatBoost* mampu meminimalkan kesalahan model dengan lebih baik dibandingkan *Random Forest*.

Meskipun demikian, perbedaan performa antara kedua metode tidak terlalu signifikan, dan keduanya menunjukkan performa yang cukup baik secara keseluruhan dalam pemodelan kualitas udara di Kota Palembang.

5. 2 Saran

Berdasarkan hasil dan kesimpulan di atas, diharapkan pada penelitian selanjutnya untuk melakukan riset dengan membagi data menjadi beberapa subset

berdasarkan kondisi spesifik seperti musim, waktu, atau lokasi, untuk melihat apakah terdapat perbedaan signifikan dalam performa kedua metode pada kondisi yang berbeda. Hal ini dapat memberikan wawasan lebih lanjut tentang kesesuaian masing-masing metode dalam situasi tertentu.

DAFTAR PUSTAKA

Akanksha, A., Maurya, N., Jain, M., & Arya, S. (2023). Prediction and Analysis of Air Pollution Using Machine Learning Algorithms. *2023 3rd International Conference on Intelligent Technologies, CONIT 2023*, 1–6. https://doi.org/10.1109/CONIT59222.2023.10205615

Al-Mahali, I. J., & Jalaludin, A.-S. I. (2003). Kitab Tafsir Al Jalalain (Asbabun Nuzul Ayat Surah alfatiah s.d Al-Isra). *1*, 1–1121.

Altinçöp, H., & Oktay, A. B. (2019). Air Pollution Forecasting with Random Forest Time Series Analysis. *2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018*, 8–12. https://doi.org/10.1109/IDAP.2018.8620768

Ambika, G. N., Singh, B. P., Sah, B., & Tiwari, D. (2019). Air quality index prediction using linear regression. *International Journal of Recent Technology and Engineering*, *8*(2), 4247–4252. https://doi.org/10.35940/ijrte.B2437.078219

Annur, C. M. (2024). *10 Cities with the Worst Air Pollution in Indonesia 2023*. Databoks. https://databoks.katadata.co.id/datapublish/2024/01/17/10-kota-dengan-polusi-udara-terburuk-2023-tangsel-teratas

Anurag, N. V., Burra, Y., Sharanya, S., & Gireeshan, M. G. (2019). Air quality index prediction using meteorological data using featured based weighted xgboost. *International Journal of Innovative Technology and Exploring Engineering*, *8*(11 Special Issue), 1026–1029. https://doi.org/10.35940/ijitee.K1211.09811S19

Calo, S., Bistaffa, F., Jonsson, A., Gómez, V., & Viana, M. (2024). Spatial air quality prediction in urban areas via message passing. *Engineering Applications of Artificial Intelligence*, *133*. https://doi.org/10.1016/j.engappai.2024.108191

Ding, Y., Chen, Z., Lu, W., & Wang, X. (2021). A CatBoost approach with wavelet decomposition to improve satellite-derived high-resolution PM2.5 estimates in Beijing-Tianjin-Hebei. *Atmospheric Environment*, *249*(August 2020), 118212. https://doi.org/10.1016/j.atmosenv.2021.118212

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *CatBoost: gradient boosting with categorical features support*. 1–7.

Fatriansyah, J. F., Dhaneswara, D., Hanifa, M., Hartoyo, F., Pradana, A. F., Anis, M., & Fauzi, A. (2023). Perancangan Program Pengestimasi Probabilitas Kegagalan Peralatan Penukar Panas Akibat Korosi Seragam Berbasis Deep Neural Network. *Syntax Literate; Jurnal Ilmiah Indonesia*, *8*(3), 1827–1839.

Gao, L., Cai, C., & Hu, X.-M. (2022). Air Quality Prediction Using Machine Learning. In *Machine Learning in Chemical Safety and Health* (pp. 267–288). https://doi.org/https://doi.org/10.1002/9781119817512.ch11

Gladkova, E., & Saychenko, L. (2022). Applying machine learning techniques in air quality prediction. *Transportation Research Procedia*, *63*, 1999–2006. https://doi.org/10.1016/j.trpro.2022.06.222

Gunasekar, S., Joselin Retna Kumar, G., & Pius Agbulu, G. (2022). Air Quality Predictions in Urban Areas Using Hybrid ARIMA and Metaheuristic LSTM. *Computer Systems Science and Engineering*, *43*(3), 1271–1284. https://doi.org/10.32604/csse.2022.024303

Guo, Z., Wang, X., & Ge, L. (2023). Classification prediction model of indoor PM2.5 concentration using CatBoost algorithm. *Frontiers in Built Environment*, *9*(July), 1–10. https://doi.org/10.3389/fbuil.2023.1207193

IQAir. (2023). *World Air Quality*. IQAir. https://www.iqair.com/id/world-air-quality

Jabeur, S. Ben, Gharib, C., Mefteh-Wali, S., & Arfi, W. Ben. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, *166*(January), 120658. https://doi.org/10.1016/j.techfore.2021.120658

Kothandaraman, D., Praveena, N., Varadarajkumar, K., Madhav Rao, B., Dhabliya, D., Satla, S., & Abera, W. (2022). Intelligent Forecasting of Air Quality and Pollution Prediction Using Machine Learning. *Adsorption Science and Technology*. https://doi.org/10.1155/2022/5086622

Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: a case study of Indian cities. *International Journal of Environmental Science and Technology*, *20*(5), 5333–5348. https://doi.org/10.1007/s13762-022-04241-5

Kusnandar, M. (2020). Republic of Indonesia Government Regulations. *Regulation of the Minister of Environment and Forestry of the Republic of Indonesia Number 14 of 2020 Concerning the Air Pollution Standard Index*, 1–16.

Lee, C. (2019). Impacts of urban form on air quality in metropolitan areas in the United States. *Computers, Environment and Urban Systems*, *77*. https://doi.org/10.1016/j.compenvurbsys.2019.101362

Lee, T. H., Ullah, A., & Wang, R. (2020). Bootstrap Aggregating and Random Forest. *Advanced Studies in Theoretical and Applied Econometrics*, *52*, 389–429. https://doi.org/10.1007/978-3-030-31150-6_13

Li, S., Deng, X., & Tang, B. (2021). Using Machine Learning Methods for Prediction of Air Quality in Wuling Mountain Area in China. *2021 International Conference on Electronic Information Technology and Smart*

Agriculture (ICEITSA), 426–430. https://doi.org/10.1109/ICEITSA54226.2021.00087

Li, W. W. (2020). *Air pollution, air quality, vehicle emissions, and environmental regulations* (H. Khreis, M. Nieuwenhuijsen, J. Zietsman, & T. B. T.-T.-R. A. P. Ramani (eds.); pp. 23–49). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-818122-5.00002-8

Liang, L., & Gong, P. (2020). Urban and air pollution: a multi-city study of long-term effects of urban landscape patterns on air quality trends. *Scientific Reports*, *10*(1), 1–13. https://doi.org/10.1038/s41598-020-74524-9

Liu, H., Li, Q., Yu, D., & Gu, Y. (2019). Air quality index and air pollutant concentration prediction based on machine learning algorithms. *Applied Sciences (Switzerland)*, *9*(19). https://doi.org/10.3390/app9194069

Liu, Y., Wang, P., Li, Y., Wen, L., & Deng, X. (2022). Air quality prediction models based on meteorological factors and real-time data of industrial waste gas. *Scientific Reports*, *12*(1), 1–15. https://doi.org/10.1038/s41598-022-13579-2

Madhuri, M., Samyama Gunjal, G. H., & Kamalapurkar, S. (2020). Air pollution prediction using machine learning supervised learning approach. *International Journal of Scientific and Technology Research*, *9*(4), 118–123.

Matović, K., & Nataša, V. (2021). Air Quality Prediction in Smart City. *2021 15th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS)*, 287–290. https://doi.org/10.1109/TELSIKS52058.2021.9606405

Mihirani, M., Yasakethu, L., & Balasooriya, S. (2023). Machine Learning-based Air Pollution Prediction Model. *2023 IEEE IAS Global Conference on Emerging Technologies, GlobConET 2023*, *2*, 1–6. https://doi.org/10.1109/GlobConET56651.2023.10150203

Nair, A. S., Khare, S., & Thakur, A. (2023). *Air Quality Index Prediction of Bangalore City Using Various Machine Learning Methods - Information and Communication Technology for Competitive Strategies (ICTCS 2022): Intelligent Strategies for ICT* (M. S. Kaiser, J. Xie, & V. S. Rathore (eds.); pp. 391–406). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-9304-6_37

Probst, P., Boulesteix, A. L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, *20*, 1–22.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *2018-Decem*(Section 4), 6638–6648.

Ramosaj, B., & Pauly, M. (2019). Consistent estimation of residual variance with random forest Out-Of-Bag errors. *Statistics and Probability Letters*, *151*, 49–57. https://doi.org/10.1016/j.spl.2019.03.017

Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. *Chemosphere*, *338*(May). https://doi.org/10.1016/j.chemosphere.2023.139518

Yalçin, T., Paradell Solà, P., Stefanidou-Voziki, P., Domínguez-García, J. L., & Demirdelen, T. (2023). Exploiting Digitalization of Solar PV Plants Using Machine Learning: Digital Twin Concept for Operation. *Energies*, *16*(13), 1–17. https://doi.org/10.3390/en16135044

Zalakeviciute, R., Rybarczyk, Y., Alexandrino, K., Bonilla-Bedoya, S., Mejia, D., Bastidas, M., & Diaz, V. (2021). Gradient boosting machine to assess the public protest impact on urban air quality. *Applied Sciences (Switzerland)*, *11*(24). https://doi.org/10.3390/app112412083

LAMPIRAN

Cara Kerja Metode Random Forest

Contoh data input :

ws	pm25	aqi
0	3.78	16
7.6	3.68	16
9.4	8.57	36
11.2	28.47	86
16.6	13.47	53
13	18.37	64
13	24.11	76
10.3	32.04	93

Boostrap Pertama

- **Langkah 1.** Sampling dengan Pengembalian (Bootstrapping)

  Setelah input data, membuat boostrap dataset atau mengambil data secara acak.

ws	pm25	aqi
7.6	3.68	16
16.6	13.47	53
0	3.78	16
13	18.37	64
9.4	8.57	36
0	3.78	16
10.3	32.04	93
13	24.11	76

- **Langkah 2:** Feature Selection

  Misalnya variabel yang digunakan adalah *WS* dan PM25

- **Langkah 3: Menghitung Gini Impurity/ Menghitung Variance**

  Menghitung nilai informasi Gini Impurity atau *Gini Index* (GI) variabel *Temp* dan PM25, dimana informasi tersebut digunakan untuk menentukan variabel yang menjadi node dalam tree. Namun karena kasus ini memprediksi nilai “aqi”

yang berupa nilai numerik (regresi), maka menggunakan metrics lain seperti Mean Squared Error (MSE).

Total variance: $\bar{y} = \frac{16+53+16+64+36+16+93+76}{8}$ = 46.75

ws	pm25	aqi
Subset 1/ Left node → pm25 ≤ 10		
7.6	3.68	16
0	3.78	16
0	3.78	16
9.4	8.57	36
Subset 2/Right Node → pm25 > 10		
16.6	13.47	53
13	18.37	64
10.3	32.04	93
13	24.11	76

Left Node :

$\bar{y} = \frac{16+16+16+36}{4}$ = 21

$Variance_{left} = \frac{1}{4}[(16-21)^2 + (16-21)^2 + (16-21)^2 + (36-21)^2] = 75$

Right Node :

$\bar{y} = \frac{53+64+93+76}{4}$ = 71.5

$$Variance_{right} = \frac{1}{4}[(53-71.5)^2 + (64-71.5)^2 + (93-71.5)^2 + (76-71.5)^2]$$
$$= 220.75$$

Setelah itu, dilakukan perhitungan *Variance Split* (VS) pada masing-masing subset, terlihat bahwa subset 2 memiliki nilai aqi yang cukup beragama (53, 64, 93, 76). Hal ini menunjukkan bahwa node tersebut masih memiliki varians yang signifikan, yang berarti ada potensi untuk memperbaiki prediksi dengan melakukan split lebih lanjut. Sehingga variabel yang digunakan sebagai internal node adalah VS pada subset 2. Oleh karena itu, subset yang dihasilkan adalah sebagai berikut.

ws	pm25	aqi
Subset 3/Left Node → pm25 ≤ 20		
16.6	13.47	53
13	18.37	64
Subset 4/Right Node → pm25 > 20		
10.3	32.04	93
13	24.11	76

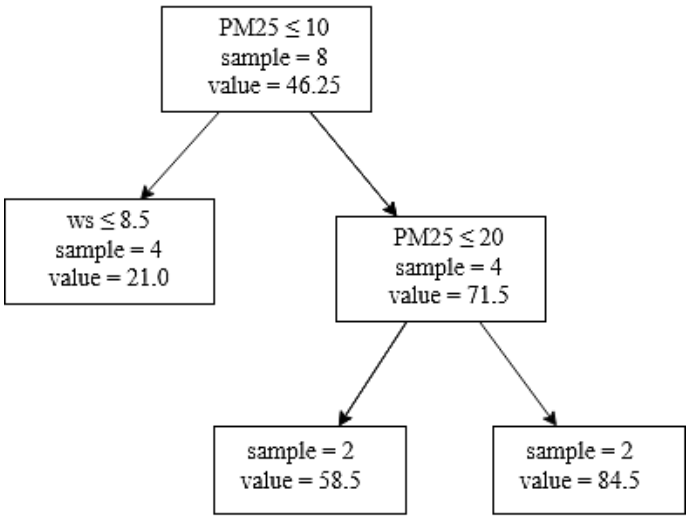
$\bar{y}_{left} = \frac{53+64}{2} = 58.5$

$Variance_{left} = \frac{1}{2}[(53 - 58.5)^2 + (64 - 58.5)^2] = 30.25$

$\bar{y}_{right} = \frac{93+76}{2} = 84.5$

$Variance_{right} = \frac{1}{2}[(93 - 84.5)^2 + (76 - 84.5)^2] = 72.25$

Untuk memberikan pemahaman yang lebih rinci, berikut adalah struktur *decision tree* yang terbentuk.



Boostrap Kedua

- **Langkah 1**. Sampling dengan Pengembalian (Bootstrapping)

ws	pm25	aqi
7.6	3.68	16
16.6	13.47	53
0	3.78	16
13	18.37	64
9.4	8.57	36
0	3.78	16
10.3	32.04	93
13	24.11	76

- **Langkah 2:** Menghitung Gini Impurity/ Menghitung Variance

ws	pm25	aqi
Subset 1/ Left node → ws <= 10		
7.6	3.68	16
0	3.78	16
9.4	8.57	36
0	3.78	16
		21
Subset 2/ Right node → ws > 10		
16.6	13.47	53
13	18.37	64
10.3	32.04	93
13	24.11	76

ws	pm25	aqi
Subset 3/ Left node → pm25 <= 15		
16.6	13.47	53
Subset 4/ Right node → ws pm25 > 15		
13	18.37	64
10.3	32.04	93
13	24.11	76

Setelah pembentukan pohon-pohon, maka dapat dilakukan prediksi kualitas udara berdasarkan data testing. Setelah itu, hasil akhir dari pengujian data akan ditentukan berdasarkan hasil voting dari setiap pohon keputusan yang telah dibangun.

Hasil Prediksi

ws	pm25	aqi	Prediksi 1	Prediksi 2	Rata-rata
7.6	3.68	16	21	21	21
16.6	13.47	53	59.25	53	56.125
0	3.78	16	21	21	21
13	18.37	64	59.25	77.67	68.46
9.4	8.57	36	21	21	21
0	3.78	16	21	21	21
10.3	32.04	93	52.75	77.67	65.21
13	24.11	76	84.75	77.67	81.21

Cara Kerja Metode Catboost

- **Langkah 1:** Menentukan Target dan Fitur

Ws	pm25	aqi
0	3.78	16
7.6	3.68	16
9.4	8.57	36
11.2	28.47	86
16.6	13.47	53
13	18.37	64
13	24.11	76
10.3	32.04	93

Fitur: ws, pm25

Target: aqi

- **Langkah 2:** Inisialisasi Model

Dimulai dengan nilai prediksi awal, misalnya rata-rata dari nilai target aqi.

$$Prediksi\ Awal = rata - rata\ dari\ aqi$$

$$= \frac{16 + 16 + 36 + 86 + 53 + 64 + 76 + 93}{8} = 55$$

- **Langkah 3:** Hitung Residual (Kesalahan)

Residual pada iterasi pertama dihitung sebagai selisih antara nilai aktual dan prediksi awal.

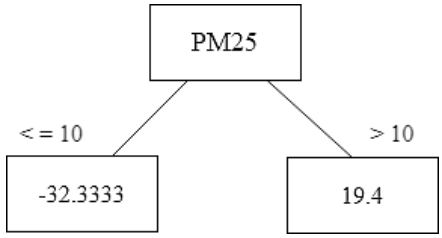
$$Residual = aqi - Prediksi\ Awal$$

ws	pm25	aqi	Prediksi awal	Residual
0	3.78	16	55	-39
7.6	3.68	16	55	-39
9.4	8.57	36	55	-19
11.2	28.47	86	55	31
16.6	13.47	53	55	-2
13	18.37	64	55	9
13	24.11	76	55	21
10.3	32.04	93	55	38

- **Langkah 4:** Bangun Pohon Keputusan (Tree) Berdasarkan Residual

  Bangun pohon keputusan menggunakan residual sebagai target. Misalnya menggunakan PM25 untuk membangun *Decision Tree* awal.

ws	pm25	aqi	Prediksi awal	Residual	Average
PM25 < = 10					
0	3.78	16	55	-39	-32.3333
7.6	3.68	16	55	-39	
9.4	8.57	36	55	-19	
PM25 > 10					
11.2	28.47	86	55	31	19.4
16.6	13.47	53	55	-2	
13	18.37	64	55	9	
13	24.11	76	55	21	
10.3	32.04	93	55	38	



- **Langkah 5:** Melakukan Prediksi

  Dengan menggunakan learning rate (misalnya 0.1)

  Untuk data dengan pm25 <= 10:

  Prediksi baru = 55 + 0.1 * (-32.33) = 55 - 3.233 = 51.767

  Untuk data dengan pm25 > 10:

  Prediksi baru = 55 + 0.1 * (19.4) = 55 + 1.94 = 56.94

ws	pm25	aqi	Prediksi Awal	Residual	Prediksi Baru
0	3.78	16	55	-39	51.767
7.6	3.68	16	55	-39	51.767
9.4	8.57	36	55	-19	51.767
11.2	28.47	86	55	31	56.94
16.6	13.47	53	55	-2	56.94
13	18.37	64	55	9	56.94
13	24.11	76	55	21	56.94
10.3	32.04	93	55	38	56.94