PREDIKSI INDEKS KUALITAS UDARA MENGGUNAKAN METODE *CATBOOST*

SKRIPSI

Oleh :
MOHAMAD ARIF ABDUL SYUKUR
NIM. 200605110044

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

PREDIKSI INDEKS KUALITAS UDARA MENGGUNAKAN METODE *CATBOOST*

SKRIPSI

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

Oleh :
MOHAMAD ARIF ABDUL SYUKUR
NIM. 200605110044

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

HALAMAN PERSETUJUAN

PREDIKSI INDEKS KUALITAS UDARA MENGGUNAKAN METODE *CATBOOST*

SKRIPSI

Oleh :
MOHAMAD ARIF ABDUL SYUKUR
NIM. 200605110044

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 05 Juni 2024

Pembimbing I,	Pembimbing II,
Prof. Dr. Suhartono, M.Kom	Dr. Totok Chamidy, M.Kom
NIP. 19680519 200312 1 001	NIP. 19691222 200604 1 001

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

Dr. Fachrul Kurniawan, M.MT, IPM
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

PREDIKSI INDEKS KUALITAS UDARA MENGGUNAKAN METODE *CATBOOST*

SKRIPSI

Oleh :
MOHAMAD ARIF ABDUL SYUKUR
NIM. 200605110044

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 07 Juni 2024

Susunan Dewan Penguji

Ketua Penguji	: Okta Qomaruddin Aziz, M.Kom NIP. 19911019 201903 1 013	( )
Anggota Penguji I	: Ajib Hanani, M.T NIP. 19840731 202321 1 013	( )
Anggota Penguji II	: Prof. Dr. Suhartono, M.Kom NIP. 19680519 200312 1 001	( )
Anggota Penguji III	: Dr. Totok Chamidy, M.Kom NIP. 19691222 200604 1 001	( )

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

Dr. Fachrul Kurniawan, M.MT, IPM
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Mohamad Arif Abdul Syukur
NIM	: 200605110044
Fakultas / Prodi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Prediksi Indeks Kualitas Udara Menggunakan Metode *CatBoost.*

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

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

Malang, 08 Juni 2024
Yang membuat pernyataan,

Mohamad Arif Abdul Syukur
NIM.200605110044

MOTTO

... Jika tidak segera dilakukan maka kamu akan terus berfikir bagaimana melakukannya, jika memang kamu menginginkannya, Lakukanlah dengan DUIT Do'a, Usaha, Ikhtiyar, Tawakal…

HALAMAN PERSEMBAHAN

Alhamdulillahirabbil'alamin, Dengan tulus dan rasa bersyukur sebagai bentuk dedikasi dan bukti tanggung jawab dengan selesainya skripsi ini kepada kedua orang tua saya yaitu Cariwan dan Wasmi. Terima kasih sudah selalu mendoakan saya di setiap waktunya serta senantiasa memberikan dukungan dengan baik. Terima kasih atas kesabaran, kekuatan dan kepercayaan yang selalu kalian berikan sehingga menjadi sumber semangat untuk menuntut ilmu dan menyelesaikan skripsi ini.

KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh.

Tidak ada kata-kata yang pantas diucapkan selain rasa syukur kepada Allah *Subhanahu Wa Ta'ala* atas karunia dan rahmat-Nya sehingga skripsi ini dapat diselesaikan dengan baik. Shalawat dan salam juga tetap tercurahkan kepada baginda Nabi Muhammad *Shalallahu 'Alaihi Wasallam* yang telah membimbing dari zaman kegelapan kepada zaman yang terang benderang yaitu *Addiinul Islam.* Skripsi ini ditulis untuk memenuhi syarat kelulusan mahasiswa Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang dengan gelar sarjana komputer (S.Kom).

Melalui kesempatan ini, penulis menyampaikan rasa terimakasih kepada semua pihak yang telah mendukung terselesaikannya skripsi ini. Terutama yaitu kedua orang tua yang selelau memberikan dukungan moral dan material serta selalu mendoakan kapan pun dan diamana pun. Selain itu penulis juga mengucapkan terimaksih kepada:

1. Prof. Dr. H. M. Zainuddin, M.A., selaku rector Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Prof. Dr. Hj. Sri Harini, M.Si., selaku dekan Fakultas Sains dan Teknologi Universitas UIN Maulana Malik Ibrahim Malang.
3. Dr. Fachrul Kurniawan, M.MT., selaku ketua program studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Prof. Dr. Suhartono, M.Kom., selaku dosen pembimbing I dan Dr. Totok Chamidy, M.Kom selaku dosen pembimbing II yang telah bersedia

meluangkan waktu untuk membimbing dan memberikan dukungan serta dorongan sehingga dapat menyelesaikan penulisan skripsi ini dengan tepat waktu.

5. Okta Qomaruddin Aziz, M.Kom., selaku dosen penguji I dan Ajib Hanani, M.Kom selaku dosen penguji II yang telah menguji, menasehati, dan memberikan sara untuk menjadikan skripsi ini menjadi lebih baik.
6. Seluruh dosen dan segenap staff program studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang yang telah memberikan segala ilmu dan pengalaman sehingga dapat menjadi motivasi dan memudahkan dalam penyelesaian skripsi ini.
7. Teman-teman Musyrif/ah dan seluruh civitas Pusat Mahad Al-Jamiah UIN Maulana Malik Ibrahim Malang, khususnya teman-teman Mabna Ibnu Rusyd 2021-2022, Mabna Ibnu Khaldun 2022-2023, dan Mabna Al-Muhasibi 2023-2024, terimakasih atas kebersamaan, pengalaman dan motivasi kalian semua.
8. Teman-teman Jurusan Teknik Informatika Angkatan 2020 “INTEGER” yang telah memberikan pengalaman, menghabiskan waktu, dan berjuang bersama semasa kuliah di UIN Maulana Malik Ibrahim Malang.
9. Semua pihak dan tempat yang tidak dapat penulis sebutkan satu persatu yang telah membantu dalam penyusunan skripsi ini secara langsung ataupun tidak langsung.
10. Diri sendiri yang sudah sabar, kuat, dan semangat dalam menuntut ilmu dan menyelesaikan kuliahnya hingga selesai tepat waktu.

Akhir kata, penulis menyadari bahwa penulisan skripsi ini masih belum mencapai kata sempurna karena masih terdapat banyak kekurangan. Dengan demikian penulis berdoa semoga skripsi ini dapat menjadi amal ibadah yang baik di sisi Allah swt dan dapat menjadi manfaat bagi siapapun di dunia ataupun di akhirat. Semoga karya ini menjadi bentuk kontribusi dalam perkembangan ilmu pengetahuan serta bentuk tanggung jawab penulis sebagai hamba Allah yang menjalankan tugasnya.

Malang, 09 Juni 2024

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Syukur, Mohamad Arif Abdul. 2024. **Prediksi Indeks Kualitas Udara Menggunakan Metode *CatBoost*.** Skripsi. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M.Kom Pembimbing (II) Dr. Totok Chamidy, M.Kom.

.

Kata kunci: *CatBoost*, *GridSearchCV*, Indeks Kualitas Udara, Prediksi

Udara merupakan sumber kehidupan bagi makhluk hidup di bumi. Tanpa adanya udara semua makhluk di bumi tidak dapat hidup dengan baik. Dengan banyaknya aktifitas industri yang sudah maju pada masa sekarang dan banyaknya pembakaran hutan, asap rokok, transportasi menyebabkan adanya polusi udara. Berdasarkan data *AirVisual* dari AQI pada tahun 2024, Negara Indonesia tepatnya di kota Jakarta telah memasuki daftar ke-11 menjadi negara dengan tingkat polusi tertinggi di dunia hingga mencapai angka 127 dalam artian udara tidak sehat bagi kelompok yang sensitive. Dengan demikian polusi udara dapat menyebabkan banyaknya penyakit seperti penyakit kulit dan bahkan penyakit pernafasan yang dapat menimbulkan kematian. Salah satu cara untuk menekan kematian tersebut adalah dengan pemantauan prediksi indeks kualitas udara dengan cepat dan efisien. Sehingga dengan informasi yang cepat dan akurat dapat mendukung pemerintah atau masyarakat dalam kelangsungan hidup, mengurangi dampak penyakit, dan meningkatkan standar hidup bagi mereka yang sensitive terhadap udara. Penelitian ini memanfaatkan metode *CatBoost* untuk memprediksi indeks kualitas udara dengan cepat dan akurat. Tujuan penelitian ini adalah untuk mengetahui performa metode *CatBoost* dalam memprediksi indeks kualitas udara pada data indeks pencemar udara di spku daerah Jakarta yang diambil dari Kaggle. Data diproses melalui tahap pra pemrosesan data yang meliputi *missing value* dan *split* data. *Split* data yang dilakukan terbagi menjadi 4 model yaitu model 1 dengan perbandingan 90% data *training* dan 10% data *testing*, model 2 dengan perbandingan 80% data *training* dan 20% data *testing*, model 3 dengan perbandingan 75% data *training* dan 25% data *testing*, model 4 dengan perbandingan 70% data *training* dan 30% data *testing*. Pada masing-masing model akan dibandingkan dengan beberapa parameter yang sama yaitu *iterations* 500, 1000, 1500, *depth* 6, 8, 10, *learning_rate* 0,1 dan 0,01, *l2_leaf_reg* 1, 2, 3. Kemudian akan dicari kombinasi parameter terbaik dengan menggunakan *GridSearchCV* untuk dijadikan parameter sebagai model yang terbaik. Pada penelitian ini didapatkan nilai akurasi terbaik pada model 1 dengan akurasi mencapai 97%. Hal tersebut dipengaruhi oleh adanya pembagian data *training* sebesar 90% dan data *testing* 10%. Semakin besar data *training* dan semakin kecil data *testing*, maka akurasi semakin tinggi.

ABSTRACT

Syukur, Mohamad Arif Abdul. 2024. **Air Quality Index Prediction Using the CatBoost Method.** Thesis. Department of Informatics Engineering, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University, Malang. Supervisor: (I) Prof. Dr. Suhartono, M.Kom Supervisor (II) Dr. Totok Chamidy, M.Kom.

Air is the source of life for living creatures on earth. Without air, all creatures on earth cannot live well. With so many advanced industrial activities nowadays and lots of forest burning, cigarette smoke, transportation causes air pollution. Based on AirVisual data from AQI in 2024, Indonesia, specifically the city of Jakarta, has entered the 11th list as the country with the highest level of pollution in the world, reaching 127 in terms of unhealthy air for sensitive groups. Thus, air pollution can cause many diseases such as skin diseases and even respiratory diseases which can cause death. One way to reduce these deaths is by monitoring air quality index predictions quickly and efficiently. So that fast and accurate information can support the government or community in survival, reduce the impact of disease, and improve the standard of living for those who are sensitive to air. This research utilizes the CatBoost method to predict air quality indices quickly and accurately. The aim of this research is to determine the performance of the CatBoost method in predicting the air quality index on air pollution index data in the Jakarta area taken from Kaggle. Data is processed through the data pre-processing stage which includes missing values and split data. The split data carried out is divided into 4 models, namely model 1 with a comparison of 90% training data and 10% testing data, model 2 with a comparison of 80% training data and 20% testing data, model 3 with a comparison of 75% training data and 25% testing data , model 4 with a comparison of 70% training data and 30% testing data. Each model will be compared with the same parameters, namely iterations 500, 1000, 1500, depth 6, 8, 10, learning_rate 0.1 and 0.01, l2_leaf_reg 1, 2, 3. Then the best parameter combination will be searched for. by using GridSearchCV to parameterize the best model. In this study, the best accuracy value was obtained in model 1 with an accuracy of 97%. This is influenced by the division of training data by 90% and testing data by 10%. The larger the training data and the smaller the testing data, the higher the accuracy.

Key words: Air Quality Index, CatBoost. GridSearchCV, Prediction

مستخلص البحث

الشكر ,محمد عارف عبد. ٢٠٢٤. التنبؤ بمؤشر جودة الهواء باستخدام طريقة *CatBoost* . أُطرُوحَة. قسم الهندسة المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية، مالانج. المشرف: (١) الأستاذ. دكتور. سوهارتونو، الماجستير المشرف (٢) دكتور توتوك شميدي، الماجستير.

الكلمات المفتاحية: بمؤشر جودة الهواء، *CatBoost, GridSearchCV*، التنبؤ

الهواء هو مصدر الحياة للكائنات الحية على الأرض. بدون الهواء، لا يمكن لجميع الكائنات على الأرض أن تعيش بشكل جيد. مع وجود العديد من الأنشطة الصناعية المتقدمة في الوقت الحاضر والكثير من حرق الغابات، يتسبب دخان السجائر ووسائل النقل في تلوث الهواء. واستنادا إلى بيانات *AirVisual* الصادرة عن تنظيم القاعدة في العراق عام ٢٠٢٤، دخلت إندونيسيا، وتحديدا مدينة جاكرتا، في القائمة الـ١١ كدولة ذات أعلى مستوى من التلوث في العالم، حيث وصلت إلى المركز ١٢٧ من حيث الهواء غير الصحي للفئات الحساسة. وبالتالي فإن تلوث الهواء يمكن أن يسبب العديد من الأمراض مثل الأمراض الجلدية وحتى أمراض الجهاز التنفسي التي يمكن أن تسبب الوفاة. إحدى الطرق لتقليل هذه الوفيات هي مراقبة تنبؤات مؤشر جودة الهواء بسرعة وكفاءة. بحيث يمكن للمعلومات السريعة والدقيقة أن تدعم الحكومة أو المجتمع في البقاء على قيد الحياة، وتقليل تأثير المرض، وتحسين مستوى المعيشة لأولئك الذين لديهم حساسية للهواء. يستخدم هذا البحث طريقة *CatBoost* للتنبؤ بمؤشرات جودة الهواء بسرعة ودقة. الهدف من هذا البحث هو تحديد أداء طريقة *CatBoost* في التنبؤ بمؤشر جودة الهواء على بيانات مؤشر تلوث الهواء في منطقة جاكرتا المأخوذة من *Kaggle*. تتم معالجة البيانات من خلال مرحلة المعالجة المسبقة للبيانات والتي تتضمن القيم المفقودة والبيانات المقسمة. تم تقسيم البيانات المقسمة إلى ٤ نماذج، وهي النموذج ١ مع مقارنة ٩٠٪ من بيانات التدريب و١٠٪ من بيانات الاختبار، النموذج ٢ مع مقارنة ٨٠٪ من بيانات التدريب و٢٠٪ من بيانات الاختبار، النموذج ٣ مع مقارنة ٧٥٪ بيانات تدريب و ٢٥٪ بيانات اختبار نموذج ٤ مع مقارنة ٧٠٪ بيانات تدريب و ٣٠٪ بيانات اختبار. ستتم مقارنة كل نموذج بنفس المعلمات، وهي *iteration* ٥٠٠، ١٠٠٠، ١٥٠٠، *depth* ٦، ٨، ١٠، *learning_rate* ٠,١ و ٠,٠١، *l2_leaf_reg* ١,٢,٣ . ثم سيتم البحث عن أفضل مجموعة من المعلمات باستخدام *GridSearchCV* لتحديد أفضل نموذج. في هذه الدراسة تم الحصول على أفضل قيمة دقة في النموذج ١ بدقة ٩٧٪. ويتأثر ذلك بتقسيم بيانات التدريب بنسبة 90% وبيانات الاختبار بنسبة 10%. كلما كانت بيانات التدريب أكبر وبيانات الاختبار أصغر، زادت الدقة.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Udara adalah gas yang bercampur dan menempati lapisan permukaan bumi (Jusuf et al., 2023). Udara yang bersih merupakan anugerah pemberian dari Tuhan Yang Maha Kuasa untuk semua makhluk di Bumi yang dikehendaki, salah satunya yaitu manusia (Bardan et al., 2023). Manusia dapat bertahan hidup karena adanya udara sebagai sumber oksigen untuk bernafas, tanpa adanya oksigen, sel-sel tubuh manusia akan mati dalam beberapa waktu (Maha & Susilawati, 2023). Dengan demikian berarti udara berperan sangat penting dalam kehidupan manusia dan makhluk lainnya yang berada di Bumi. Sehingga udara di Bumi harus dijaga kualitasnya dan tidak diperbolehkan tercemar karena suatu zat yang kotor (Ridho & Mahalisa, 2023).

Indeks Kualitas Udara bagi manusia sangat penting untuk selalu dipantau keadaanya. Penurunan indeks kualitas udara akan menyebabkan dampak buruk pada kesehatan manusia dan menyebabkan terjadinya penyakit menular, keracunan dan gangguan syaraf serta penyakit lain pada organ pernapasan ataupun organ kardiovaskuler (Husen et al., 2023). Berdasarkan informasi yang dilansir dari Badan Meteorologi, Klimatologi dan Geofisika (BMKG) daerah Pontianak bahwa beberapa tahun yang lalu pada tanggal 20 September 2019 kualitas udara pada konsentrasi partikulat (PM_{10}) telah jauh melampaui nilai ambang batas (NAB) sekitar 495.05 µgram/m3. Hal ini dikarenakan adanya karhutla atau kebakaran

hutan dan lahan yang menyebabkan terjadinya kabut asap dan mengakibatkan penderitaan infeksi saluran pernapasan (ISPA) pada 6.025 warga (Putra & Rismawan, 2023). Untuk mengetahui indeks kualitas udara dapat ditemukan dengan menggunakan teknik data mining.

Datamining adalah metode yang banyak dipakai dan menjanjikan dalam menganalisis data, baik itu data yang sederhana ataupun data yang besar. Metode ini menggunakan berbagai teknik dalam proses analisa atau pengamatannnya untuk melihat hubungan yang tidak diketahui dalam jumlah besar agar dimengerti kegunaannya dalam pemilihan data dan memutuskan kesimpulan yang berguna (Hendra Di Kesuma et al., 2022). Metode ini juga merupakan proses yang baik dan akurat dalam menemukan informasi pada *datasheet* atau kumpulan data. Dalam prosesnya menggunakan teknik tertentu, metode ini berkaitan dengan algoritma atau rumus seperti statistika, matematika ataupun *machine learning* (Sholeh et al., 2023). *Dataminig* terbagi menjadi beberapa metode sesuai dengan karakteristik masing-masing. Jika digunakan untuk melihat kelakuan dari beberapa kejadian khusus yang memunculkan hubungan asosiasi maka yang paling cocok adalah menggunakan Asosiasi sedangkan jika digunakan untuk mengetahui nilai yang akan datang dari suatu hubungan, yang paling cocok adalah Prediksi (Fitri Boy, 2020).

Prediksi merupakan suatu proses untuk memperkirakan nilai yang akan datang berdasararkan data informasi yang terdahulu secara sistematis melaluli pendekatan-pendekatan yang akurat. Jawaban dari prediksi tidak harus nilai yang sangat pasti dalam keadaan yang akan terjadi pada masa depan, melainkan

pendekatan yang paling dekat. Dengan demikian nantinya akan menunjukan hasil nilai yang dirasa paling dekat sehingga akan dijadikan sebagai bahan masukan dalam pengambilan keputusan atau proses perencanaan di masa yang akan datang (Kriswantara & Sadikin, 2022). Dalam memprediksi sesuatu perlu adanya model yang dirancang dari algoritma untuk membantu lebih akuratnya hasil yang akan ditemukan. Dengan bantuan algoritma proses prediksi yang dilakukan akan lebih mudah dan tidak menguras waktu yang sangat banyak (Saadah & Salsabila, 2021). Terdapat banyak algoritma untuk memprediksi suatu data, seperti algoritma *Boosting* yang menggabungkan metode berurutan dan meningkatkan pengamatan secara iteratif. Dalam hal ini yang dapat digunakan untuk memprediksi data adalah algoritma *CatBoost (Categorikal Boosting)* dengan pendekatan *machine learning* ansambel yang dikuatkan dengan kerangka *Gradient Boosting* (Nababan et al., 2023). Terbukti dalam penelitian yang dilakukan oleh N. Srinivasa Gupta dan temannya yang berjudul "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis" yang di dipublikasikan oleh jurnal Hindawi dengan menbandingkan 3 metode yaitu Support Vector Regression, Random Forest Regression dan *CatBoost* Regressor. Dan hasil yang didapatkan dari perbandingan tersebut, metode *CatBoost* Regressor menghasilkan akurasi yang paling terbaik dibandingkan dengan dua metode lainnya. Sehingga metode *CatBoost* Regressor dapat digunakan untuk penelitian lain dengan data yang berbeda untuk membuktikan apakah dengan data yang berbeda metode *CatBoost* Regressor akan tetap menghasilkan akurasi yang baik.

CatBoost (Categorikal Boosting) merupakan algoritma yang termasuk kedalam gabungan keluarga *Gradient Boosted Decision Trees* (GBDT) dan masih dalam lingkup *ensemble learning.* Algoritma ini dikembangkan dan dibuka secara umum dalam *Supervised machine learning* dengan mengaitkan inovasi *Ordered Boosting* dan *Ordered Target Statistics* (Christian et al., 2022). Algoritma ini dikembangkan oleh para peneliti dari Yandex sebagai penerus dari algoritma MatrixNet yang dapat menyelesaikan beberapa masalah di berbagai bidang. Pada alagoritma ini diterapkan teknik *Gradient Boosting* untuk meningkatkan stabilitas dan keakuratan hasil prediksi yang baik dengan data yang heterogen. Algortima ini juga dapat mengurangi masalah *overfitting* karena menggunakan kode yang efisien dan juga dapat bekerja dengan baik dalam penentuan fitur yang bersifat kategori. Algoritma *CatBoost* juga memperhitungkan data objek atau data yang bukan numerik, tidak seperti metode-meteode yang sebelumnya dimana data yang bukan numerik harus diubah dulu kedalam bahasa angka supaya bisa diterjemahkan. Sehingga dapat mempengaruhi hasil keakuratan data yang ditemukan (Dewi, 2021). Maka dari itu penggunaan algoritma *CatBoost* diharapkan dapat memperbaiki kekurangan yang ada dan dapat meningkatkan hasil keakuratan data yang ditemukan.

Dalam Al Qur’an dijelaskan bahwa dalam hujan yang turun dari langit dan dalam pergantian siang dan malam dapat memberi kehidupan di Bumi dan dari pergeseran udara ke berbagai arah dengan kekuatan dan suhu yang berbeda serta bahaya dan manfaatnya terdapat kebesaran Allah bagi kaum yang mengerti karena akalnya. Sehingga kita sebagai manusia dianjurkan mengerti mengenai kualitas

udara supaya dapat berkontribusi dalam memperkirakan kualitas udara yang baik dengan harapan dapat dijadikan sebagai hal yang berguna dan bermanfaat bagi kesehatan manusia dan makhluk hidup lainnya. Sebagaimana Al-Qur'an Surah Al-Jatsiyah ayat 5 yang berbunyi sebagai berikut.

وَٱخۡتِلَٰفِ ٱلَّيۡلِ وَٱلنَّهَارِ وَمَآ أَنزَلَ ٱللَّهُ مِنَ ٱلسَّمَآءِ مِن رِّزۡقٖ فَأَحۡيَا بِهِ ٱلۡأَرۡضَ بَعۡدَ مَوۡتِهَا وَتَصۡرِيفِ ٱلرِّيَٰحِ ءَايَٰتٞ لِّقَوۡمٖ يَعۡقِلُونَ

"Dan pada pergantian malam dan siang dan hujan yang diturunkan Allah dari langit lalu dihidupkan-Nya dengan air hujan itu bumi sesudah matinya; dan pada perkisaran angin terdapat tanda-tanda (kekuasaan Allah) bagi kaum yang berakal" (QS. Al-Jatsiyah: 5).

Dijelaskan pada Tafsir Al-Mukhtashar "*Dan didalam pergantian siang dan malam, di dalam hujan yang diturunkan Allah dari langit lalu dengannya Allah menghidupkan bumi, dengan menghidupkan tumbuh-tumbuhan setelah sebelumnya berupa tanah mati, tidak ada tumbuh-tumbuhan padanya, dan didakam perkisaran angin dengan mendatangkannya sesekali dari satu arah dan sesekali dari arah yang lain agar engkau mendapat manfaat, terdapat bukti-bukti bagi orang-orang yang berakal, lalu dengannya mereka menjadikan dalil atas keesaan Allah dan kekuasaan-Nya untuk menbangkitkan serta kekuasaan-Nya atas segala sesuatu*" (TafsirWeb, n.d.-b).

Di surat Al-An'am ayat 98 juga dijelaskan bahwa telah dijelaskan kepada orang-orang yang mengetahui tentang tanda-tanda kebesaran Allah swt. Dan Allah telah menciptakan manusia seorang diri, sehingga manusia mendapatkan tempat yang tetap dan tempat untuk simpanan.

وَهُوَ ٱلَّذِىٓ أَنشَأَكُم مِّن نَّفْسٍ وَٰحِدَةٍ فَمُسْتَقَرٌّ وَمُسْتَوْدَعٌ ۗ قَدْ فَصَّلْنَا ٱلْءَايَٰتِ لِقَوْمٍ يَفْقَهُونَ

"*Dan Dialah yang menciptakan kamu dari seorang diri, maka bagimu terdapat tempat tetap dan tempat simpanan. Sesungguhnya telah Kami jelaskan tanda-tanda kebesaran Kami kepada orang-orang yang mengetahui*" *(QS. Al-An'am: 98).*

Tafsir Al-Madinah Al-Munawarah dibawah pengawasan Syaikh Prof. Dr. Imad Zuhair Hafidz, beliau adalah seorang professor fakultas al-qur'an Universitas Islam Madinah. Menjelaskan "*Dan termasuk tanda-tanda kekuasaannya, Dia menciptakan kalian dari Adam dan menjadikan permulaan kalian menetap dalam rahim dan tersimpan dalam tulang sulbi. Kemudian kehidupan dimulai dengan pertumbuhan dan penyebaran, menjadi berbagai jenis dan warna kulit, suku bangsa, dan kabilah*" (TafsirWeb, n.d.-a). Dalam tafsir tersebut dapat ditarik kesimpulan bahwa Allah menjelaskan ayat ini agar bukti keesaan Allah tetap jelas bagi orang-orang yang berusaha untuk melihat dan memahami hikmah dibalik itu semua.

Pada Al-Qur'an surat Al-An'am ayat 59 juga dijelaskan bahwa semua yang terjadi didunia sudah tertulis dalam kitab yang mana kitab itu diartikan sebagai Lauh Mahfudz. Sehingga udara yang tercemar atau tidak tercemar ataupun udara yang sehat dan tidak sehat yang terjadi di dunia sudah tercatat semuanya di lauh mahfudz. Maka dengan demikian kita sebagai manusia harus bersyukur dan mencari tahu bagaimana cara mengetahui udara dilingkungan kita ini sehat atau berbahaya. Dengan harapan berusaha untuk menyelamatkan diri sendiri dan orang lain serta makhluk hidup lainnya dari mara bahaya yang disebabkan oleh polusi udara.

وَعِندَهُۥ مَفَاتِحُ ٱلْغَيْبِ لَا يَعْلَمُهَآ إِلَّا هُوَ ۚ وَيَعْلَمُ مَا فِى ٱلْبَرِّ وَٱلْبَحْرِ ۚ وَمَا تَسْقُطُ مِن وَرَقَةٍ إِلَّا يَعْلَمُهَا وَلَا حَبَّةٍ فِى ظُلُمَٰتِ ٱلْأَرْضِ وَلَا رَطْبٍ وَلَا يَابِسٍ إِلَّا فِى كِتَٰبٍ مُّبِينٍ

"*Dan pada sisi Allah-lah kunci-kunci semua yang ghaib; tidak ada yang mengetahuinya kecuali Dia sendiri, dan Dia mengetahui apa yang di daratan dan di lautan, dan tiada sehelai daun pun yang gugur melainkan Dia mengetahuinya, dan tidak jatuh sebutir biji-pun dalam kegelapan bumi, dan tidak sesuatu yang basah atau kering, melainkan tertulis dalam kitab yang nyata (Lauh Mahfudz)" (QS. Al-An'am: 59).*

Berdasarkan latar belakang yang sudah dipaparkan diatas, dimana indeks kualitas udara yang sangat penting dalam kehidupan makhluk hidup, maka harus selalu dipantau kualitasnya. Dalam hal ini dengan mengacu pada penelitian yang dilakukan oleh N. Srinivasa Gupta pada tahun 2023 dalam jurnal Hindawi metode *CatBoost* dapat diterapakan sebagai algoritma untuk memprediksi indeks kualitas udara dengan data spku daerah Jakarta yang diperoleh dari kaggle, dengan harapan metode ini dapat menghasilkan akurasi yang baik. Dengan demikian penelitian ini akan mengimplementasikan metode *CatBoost* dalam memprediksi indeks kualitas udara di Jakarta pada data yang diperoleh dari kaggle. Sehingga penelitian ini berjudul **"Prediksi Indeks Kualitas Udara Menggunakan Metode *CatBoost"***.

1.2 Rumusan Masalah

Berdasarkan dengan adanya latar belakang diatas, maka permasalahannya adalah bagaimana performa metode *CatBoost* untuk memprediksi indeks kualitas udara berdasarkan data indeks standar pencemar udara di spku daerah jakarta dataset yang bersumber dari *kaggle*?

1.3 Batasan Masalah

Dari permasalahan yang ada, maka batasan masalah penelitian ini adalah data yang digunakan merupakan data publik bersumber dari *kaggle* pada tahun 2020.

1.4 Tujuan Penelitian

Tujuan penelitian adalah untuk mengetahui performa metode *CatBoost* dalam memprediksi indeks kualitas udara berdasarkan data indeks standar pencemar udara di spku daerah Jakarta dataset yang bersumber dari *kaggle*.

1.5 Manfaat Penelitian

Manfaat penelitian dari penelitian yang diharapkan adalah:

1. Dapat dijadikan sebagai referensi penelitian selanjutnya dalam penerapan metode *CatBoost* untuk prediksi kualitas udara.
2. Memberikan wawasan dan membantu peneliti dalam mengembangkan topik penelitian metode *CatBoost.*
3. Memberikan wawasan kepada pembaca mengenai indeks kualitas udara.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terkait

Penelitian oleh Yudiskara dan temannya memprediksi polusi udara di kota Jakarta menggunakan metode *Recurrent Neural Network-Gated Units* yang dilatih menggunakan data polusi udara dari tahun 2010 hingga 2021 di Jakarta. Perhitungan RMSE yang dihasilkan adalah 9,8044521 untuk pm10, 8,77282145 untuk so2, 7.24068196 untuk co, 118.02030243 untuk o3, dan 10.63599659 untuk no2. Sehingga dengan tingkat keakuratan prediksi model yang telah dibuat berhasil memprediksi partikulat udara di kota Jakarta dengan cukup baik. Dan terbukti bahwa metode *Recurrent Neural Network-Gated Units* mempunyai performa prediksi untuk polusi udara di Jakarta dengan cukup baik (Yudiskara et al., 2023).

Kualitas udara di Jakarta juga telah diteliti oleh Jayadi et.al pada tahun 2023 dengan mengimplementasikan metode yang berbeda yaitu perbandingan antara metode *K-Nearest Neighbor* dan *Support Vector Machine.* Dalam penelitiannya mereka menggunakan *dataset* indeks pencemaran udara yang diambil dari website Data Jakarta yang tersedia secara publik. Dalam pengujiannya dilakuakan beberapa klasifikasi kategori tingkat kualitas udara dengan menggunakan KNN dan SVM dengan data latih 3506 dan data uji sebanyak 877. Hasil menggunakan metode KNN yang terbaik adalah pada K = 6 dengan nilai akurasi mencapai 96%, presisi mencapai 96%, *recall* mencapai 93% dan *F1-Score* 94% setelah percobaan dari K = 2 hingga K = 10. Sedangkan untuk SVM dilakukan percobaan mengguanakan

kernel polynomial, rbf dan linear dengan parameter kernel 1, 10 dan 100 terdapat hasil terbaik yaitu pada percobaan menggunakan rbf dengan parameter 100 dengan nilai akurasi mencapai 98% presisi mencapai 97% *recall* mencapai 97% dan *F1-Score* mencapai 97%. Sehingga dapat disimpulkan bahwa kinerja algoritma SVM lebih baik dibandingkan dengan metode KNN dalam klasifikasi kategori indeks standar pencemaran udara di Jakarta (Jayadi et al., 2023).

Penelitian lain mengenai kualitas udara juga dilakukan oleh Putri pada tahun 2023 dengan menggunakan metode yang berbeda yaitu metode *Artificial Neural Network* (ANN) algoritma *Backpropagation* dalam mengklasifikasi kualitas udara di Provinsi DKI Jakarta yang mana dapat memberikan pemahaman yang mendalam. Dari penelitian ini data yang digunakan masih terdapat data kategorikal yang mengharuskan untuk diubah terlebih dahulu kedalam data numerik. Percobaan yang dilakukan dalam penelitian ini menggunakan model *Backpropagation* dengan nilai *learning rate* 0.001, 0.01, dan 0.1. sedangkan jumlah *hidden layer* yang akan dicoba adalah nilai 1 hingga 5 dengan *epoch* 50, 200, 500, 1000, dan 5000. Dengan masing-masing percobaan menggunakan parameter yang berbeda menghasilkan nilai yang optimal pada *learning rate* 0.001 dengan nilai *epoch* 5000 yaitu dengan hasil akurasi sebesar 94%, presisi 90% dan *recall* 100%. Sehingga metode ANN dengan algoritma *Backpropagation* terbukti dapat mengklasifikasi kualitas udara di Provinsi DKI Jakarta dengan sangat baik dan dapat dijadikan sebagai referensi untuk penelitian yang selanjutnya (Putri, 2023).

Penelitian lain tentang kualitas udara juga dilakukan oleh Ridho dan Mahalisa pada tahun 2023 dengan judul “Analisis Klasifikasi Dataset Indeks

Standar Pencemaran Udara (ISPU) di Masa Pandemi menggunakan Algoritma *Support Vector Mechine* (SVM)". Penelitian ini menggunakan dataset yang diambil dari website *kaggle* dan di idnetifikasi tipe data yang mengguanakan tipe integer. Setelah itu analisis data dengan *preprocessing* dan *splitting* data dengan membagi data *training* dan *testing*. Berdasarkan pengujian yang sudah dilakukan terdapat bahwa udara yang sedang lebih banyak dibandingkan dengan udara yang tidak sehat pada masa covid. Dengan menggunakan metode SVM untuk mengklasifikasi dataset dari ISPU berhasil menemukan akurasi yang tertinggi yaitu sebesar 97% sehingga terbukti bahwa pada masa covid udara menjadi lebih sehat (Ridho & Mahalisa, 2023).

Penelitian terkait tentang kualitas udara juga dilakukan oleh Nababan et,al pada tahun 2023 dengan menggunakan metode *Extreme Gradient Boosting* (XGBoost) dengan *Synthetic Minority Oversampling Technique* (SMOTE) pada prediksi kualitas udara berdasarkan indeks standar pencemaran udara. SMOTE digunakan untuk menangani data yang tidak seimbang dan data yang digunakan adalah data yang dikumpulkan oleh Badan Lingkungan Hidup Jakarta dalam 1 bulan dari jam 7 pagi sampai jam 8 pagi. Dalam pengujian prediksi kualitas udara menggunakan metode XGBoost dengan Repeated K-fold validation menunjukan tingkat performa yang sangat baik yaitu didapatkan hasil nilai akurasi mencapai 98%, presisi mencapai 79%, *recall* mencapai 79%, nilai f1-score sebesar 98% dan ROC AUC sebesar 99%. Dengan demikian terbukti bahwa model XGBoost dapat memprediksi kualitas udara dengan tingkat keakuratan yang tinggi. (Nababan et al., 2023).

Kualitas udara juga diteliti oleh N. Srinivasa Gupta dan temannya dengan penelitian yang berjudul "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis" dengan menbandingkan 3 metode yaitu Support Vector Regression, Random Forest Regression dan *CatBoost* Regressor. Dalam penelitiannya terbukti jika menggunakan metode SMOTE hasil akurasi akan lebih tinggi dibandingakan dengan hasil akurasi tanpa menggunkaan metode SMOTE. Dengan menggunakan SMOTE akurasi meningkat dari 76,6786% menjadi 93,5658% untuk metode Support Vector Regression dan dari 93.5658% menjadi 97,6080% untuk metode Random Forest Regressor. Sedangkan untuk metode *CatBoost* Regressor meningkat dari 77,8991% menjadi 96,7529%. Secara keseluruhan dalam menggunakan metode SMOTE akan meningkatkan akurasi model yang digunakan. Dari ketiga metode yang sudah diterapkan hasil yang paling tinggi adalah hasil dari metode *CatBoost* Regressor (Gupta et al., 2023).

Penelitian dengan mengimplementasikan metode *CatBoost* juga dilakukan oleh Christian et,al pada tahun 2022 untuk memprediksi pembatalan pesanan kamar hotel dengan data yang digunakan adalah data bersumber dari *kaggle* yaitu data hotel booking.csv. Dengan melakukan optimasi algoritma *CatBoost* dapat mengahsilkan parameter yang terbaik untuk mengklasifikasikan pesanan kamar hotel. Percobaan dengan menggunakan dept = 10, *iterations* = 2000, 12_leaf_reg = 1 dan *learning_rate* = 0.012 menghasilkan model klasifikasi dengan nilai akurasi sebesar 88% dan nilai presisi sebesar 86%. Sehingga dapat dikatakan metode *CatBoost* menghasilkan akurasi yang tinggi. Dengan demikian metode *CatBoost*

dapat dijadikan metode yang tepat untuk penelitian yang selanjutnya dengan sedikit perkembangan (Christian et al., 2022).

Penelitian lain yang menggunakan metode *CatBoost* juga dilakukan oleh Dewi pada tahun 2021 dengan menggunakan metode *CatBoost Classifier* untuk mendeteksi *Fake Follower* Instagram dan membuktikan hasil performa yang sangat tinggi dengan menggunakan SMOTE-NC dibandingkan tanpa menggunakan SMOTE-NC. Nilai akurasi meningkat hingga sampai 2% dan nilai presisi juga meningkat hingga 6%, nilai *recall* meningkat hingga 10%, F1-score meningkat hingga 9%. Akan tetapi pelatihan dengan menggunakan data latih mengalami keterlambatan selama 2 detik karena data latih yang terlalu banyak, hal ini dapat diselesaikan dengan menerapkan fitur seleksi agar mempercepat waktu. Maka dari itu model yang terbaik yang dihasilkan pada penelitian ini adalah model *CatBoost* dengan SMOTE-NC dan penerapan fitur seleksi yang mana menghasilkan nilai akurasi mencapai 98,04%, nilai presisi mencapai 97,71%, nilai *recall* mencapai 98,61%, dan nilai ROC AUC mencapai 99.71%. Waktu yang dibutuhkan untuk pelatihan adalah 11.1 detik yang termasuk kedalam kategori cepat (Dewi, 2021).

Tabel 2. 1 Penelitian Terkait

No	Referensi	Objek	Metode	Hasil	Perbedaan
1	Yudiskara et al., (2023)	Kualitas Udara di Jakarta tahun 2010 hingga 2021	*Recurrent Neural Network-Gated Units*	Perhitungan RMSE yang dihasilkan adalah 9,8044521 untuk pm10, 8,77282145 untuk so2, 7.24068196 untuk co, 118.02030243 untuk o3, dan 10.63599659 untuk no2	Metode dan data yang digunakan berbeda akan tetapi objeknya sama yaitu polusi udara di Jakarta.
2	Jayadi et al., (2023)	Indeks Standar Pencemaran Udara (ISPU) di	*K-Nearest Neighbor* dan *Support Vector Machine*	Hasil menggunakan metode KNN yang terbaik adalah pada K = 6 dengan nilai akurasi mencapai	Metode yang digunakan berbeda akan tetapi untuk objek

No	Referensi	Objek	Metode	Hasil	Perbedaan
		DKI Jakarta		96%, presisi mencapai 96%, *recall* mencapai 93% dan *F1-Score* 94% setelah percobaan dari K = 2 hingga K = 10. Sedangkan untuk SVM dilakukan percobaan mengguanakan kernel polynomial, rbf dan linear dengan parameter kernel 1, 10 dan 100 terdapat hasil terbaik yaitu pada percobaan menggunakan rbf dengan parameter 100 dengan nilai akurasi mencapai 98% presisi 97% *recall* 97% dan *F1-Score* sebesar 97%.	penelitiannya sama yaitu Indeks Standar Pencemaran Udara.
3	Putri, (2023)	Kualitas Udara di Provinsi DKI Jakarta tahun 2021	*Artificial Neural Network* (ANN) algoritma *Backpropagation*	Menghasilkan nilai yang optimal pada *learning rate* 0.001 dengan nilai *epoch* 5000 yaitu dengan hasil akurasi sebesar 94%, presisi 90% dan *recall* 100%.	Menggunakan metode yang berbeda namun objek yang diteliti sama, yaitu kualitas udara.
4	Ridho & Mahalisa, (2023)	Indeks Standar Pencemaran Udara pada masa covid	*Support Vector Machine*	Metode SVM untuk mengklasifikasi dataset dari ISPU berhasil menemukan akurasi yang tertinggi yaitu sebesar 97%	Metode yang berbeda, pada penelitian terdahulu ini menggunakan SVM sedangkan peneliti akan mengembangkan mengguanakn *CatBoost* dengan objek yang diteliti sama yaitu kualitas udara
5	Nababan et al., (2023)	Kualitas Udara berdasarkan Indeks Standar Pencemaran Udara	*Extreme Gradient Boosting* (XGBoost) dengan *Synthetic Minority Oversamplin*	Dalam pengujian prediksi kualitas udara menggunakan metode XGBoost dengan Repeated K-fold validation menunjukan tingkat performa yang sangat baik yaitu didapatkan	Sama sama meneliti objek kualitas udara, akan tetapi metodenya berbeda yaitu dengan XGBoost

No	Referensi	Objek	Metode	Hasil	Perbedaan
			g Technique (SMOTE)	hasil nilai akurasi mencapai 98%, nilai presisi mencapai 79%, *recall* mencapai 79%, f1-score mencapai 98% dan ROC AUC sebesar 99%.	menggunakan SMOTE.
6	Gupta et al., (2023)	*Air Quality Index*	*Support Vector Refression, Random Forest Regression, dan CatBoost Regression.*	Dengan menggunakan SMOTE akurasi meningkat dari 76,6786% menjadi 93,5658% untuk metode Support Vector Regression dan dari 93.5658% menjadi 97,6080% untuk metode Random Forest Regressor. Sedangkan untuk metode *CatBoost* Regressor meningkat dari 77,8991% menjadi 96,7529%.	Berbeda pada data yang digunakan. Penelitian yang dilakukan Gupta menggunakan data daerah luar negeri.
7	Christian et al., (2022)	Booking Pesananan Hotel menggunakan data hotel booking.csv	Metode *Categorical Boosting* (*CatBoost*)	Percobaan dengan menggunakan dept = 10, *iterations* = 2000, 12_leaf_reg = 1 dan *learning_rate* = 0.012 menghasilkan model klasifikasi dengan nilai kaurasi sebesar 88% dan nilai presisi sebesar 86%.	Sama sama menggunakan metode *CatBoost* akan tetapi data dan objeknya berbeda.
8	Dewi, (2021)	*Fake Follower* Instagram	*CatBoost* dengan SMOTE-NC dan penerapan fitur seleksi	Menghasilkan nilai akurasi mencapai 98,04%, nilai presisi mencapai 97,71%, nilai *recall* mencapai 98,61%, dan nilai ROC AUC mencapai 99.71%. Waktu yang dibutuhkan untuk pelatihan adalah 11.1 detik yang termasuk kedalam kategori cepat.	Metode yang sama digunakan yaitu *CatBoost.* Namun Objek dan data yang diteliti berbeda

2.2 Udara

Udara adalah zat paling penting setelah air karena udara berkontribusi memeberikan kehidupan di Bumi dengan oksigen. Fungsi lain udara adalah sebagai penghantar suara dan mendinginkan benda yang panas. Selain itu udara juga dapat menimbulkan nilai negatif seperti dapat menyebarkan penyakit pada makhluk hidup. Udara yang normal mengandung nilai nitrogen sebesar 78.1 %, oksigen 20,93 % dan karbon dioksida sebesar 0.03% (Shinta Enggar Maharan, 2021). Udara bersih adalah udara yang tidak tercampur dengan zat atau gas yang berbahaya ataupun merugikan seperti bercamupurnya dengan debu, nitrogen dioksida, karbon dioksida dan gas yang berbahaya lainnya(Amalia et al., 2022). Udara yang terdapat di alam pada kenyataanya tidak setiap hari bersih dan sehat karena adanya gas yang disebabkan oleh kemajuan teknologi pada masa sekarang ini yang menyebabkan kualitas udara mengalami penurunan. Dalam buku yang diterbitkan oleh Institut Pertanian Bogor yaitu tentang Pengendalian Kebakaran Hutan dan Lahan menjelaskan bahwa partikel yang kurang dari ukuran 10 mikrometer seperti partikel halus yag berukuran lebih rendah dari 2.5 mikrometer dapat masuk kedalam tubuh manusia salah satunya paru-paru. Sehingga polusi partikel yang masuk dapat menyebabkan kesulitan pernafasan, penyakit asma dan sesuatu yang berkaitan dengan kematian (Hariyadi et al., 2023).

2.3 Indeks Kualitas Udara

Kualitas udara merupakan sesuatu yang dijadikan ukuran untuk baik atau buruk mengenai udara yang dihasilkan oleh suatu campuran gas yang ada di Bumi yang mana nilai campurannya tidak setiap hari konstan. Jika udara tercampur

dengan masuknya suatu zat atau gas kedalam udara yang dilakukan karena adanya kegiatan manusia maka bisa dikatakan udara tersebut adalah udara yang tercemar sehingga akan menyebabkan penurunan kualitas udara ambien dan berdampak pada kehidupan makhluk hidup. Terdapat beberapa zat pencemar udara yang berdampak kepada penurunan kualitas udara yaitu Nitrogen Dioksida (NO_2), Karbon Monoksida (CO), Sulfur Dioksida (SO_2), Ozon (O_3) dan Partikulat Debu (PM_{10}) (Hermawan & Sela, 2019). Berdasarkan keputusan Nomor 45 tahun 1997 yang diputuskan oleh Menteri Negara Lingkungan Hidup tentang Indeks Standar Pencemar Udara (ISPU) dan keputusan Nomor 107 tahun 1997 melalui Keputusan Kepala Bapedal tentang dampak lingkungan dan pedoman teknis perhitungan, pelaporan serta informasi Indeks Standar Pencemar Udara, maka berdasar kepada dampak kesehatan manusia dan perlindungan untuk makhluk hidup di Bumi dapat digambarkan melalui kualitas udara ambien dengan Indeks Standar Kualitas Udara (Wangintan & Sofyan, 2019). Mulai dari tahun 2015 hingga tahun 2030 terdapat tujuan pembangunan yang berkelanjutan dengan 3 pilar yaitu sosial, ekonomi, dan lingkungan yang dirancang oleh Persekutuan Bangsa-Bangsa (PBB). Beberapa tujuannya adalah mengenai kesehatan, pola konsumsi dan produksi tanggung jawab, kota dan komunitas berkelanjutan, perubahan iklim dan mengenai ekosistem daratan. Semua tujuannya berkaitan dengan tingkat kualitas udara yang mana pencemaran udara sifatnya tidak dapat dibatasi dan disekat sehingga dampak kepada kehidupan makhluk hidup yang ada di Bumi sulit untuk dihindari (Bernadet et al., 2023). Indeks Standar Pencemaran Udara (ISPU) adalah kondisi sebagai dasar kualitas udara ambien yang terdapat disuatu wilayah untuk mengetahui

dampak pada kesehatan makhluk hidup. Kategori baik buruknya kualitas udara yang dihasilkan didasari dengan nilai ISPU sesuai dengan pencemaran zat utama sesuai dengan keputusan kepala Bapedal. Nilai ISPU dengan kategori baik adalah 0-50 yaitu kualitas udara yang sehat, kategori sedang adalah 51-100 dengan kualitas udara yang dapat mempengaruhi pandangan, kategori tidak sehat adalah 101-199 dengan kualitas udara yang menyebabkan kotor karena debu, kategori sangat tidak sehat adalah 200-299 menyebabkan kesensitifan pada penderita asma dan bronkitis, sedangkan nilai diatas 300 adalah kategori yang berbahaya (Khumaidi et al., 2020).

2.4 *Machine Learning*

Machine learning merupakan cabang kecerdasan buatan atau *Artificial Intelligence* yang biasa disebut dengan AI menurut IBM. *Machine learning* merupakan salah satu cabang ilmu koomputer yang berfokus kepada algoritma pengolahan data untuk memecahkan masalah. Sehingga dapat membuat keputusan yang bisa menirukan manusia untuk meningkatkan komputer dalam kemampuan belajarnya secara bertahap dan dapat meningkatkan akurasi. Dengan demikian yang difokuskan oleh *machine learning* adalah untuk pengembangan sistem yang membuat keputusan mandiri dan memiliki kemampuan belajar tanpa adanya program ulang oleh manusia dan bahkan mampu beradaptasi dengan perubahan yang baru karena tidak hanya menentukan tindakan yang optimal dalam keputusannya (Pratama et al., 2023). *Machine learning* akan berjalan secara otomatis dan dengan menyajikan *insight* yang bersifat prediktif akan menimbulkan lebih optimalnya pengambilan keputusan yang dilakukan. Untuk pengambilan secara otomatis berdasarkan kumpulan data dapat menggunakan algoritma

pembelajaran untuk memprediksi hasil yang ditemukan. Sehingga *machine learning* berkaitan dengan studi pembelajaran tentang konsep dan pola yang dikenali pada pembelajaran komputasi menggunakan algoritma (Wardhana et al., 2023).

Tujuan *Machine learning* adalah untuk mengetahui hasil ekspresi prediksi dalam pengolahan data yang sederhana dan dapat mudah dipahami oleh manusia (Nurkholifah et al., 2023). Sehingga untuk meniru dalam penalaran seperti manusia membutuhkan data yang cukup agar memberikan informasi yang akurat dalam pengambilan keputusan yang proses operasinya tanpa adanya bantuan dari manusia. Dengan demikian proses yang dilakukan dalam penelitian tidak bergantung kepada manusia melainkan memanfaatkan bantuan komputer untuk menganalisa dan memprediksi nilai indeks kualitas udara (Leni et al., 2023). Selain itu *machine learning* juga mengembangakan model statistika dan matematika untuk mengeksploitasi algoritma dalam pembelajaran pada data. Paradigma yang difokuskan adalah pada tujuan prediksi atau klasifikasi yang terdiri dari optimalnya pemetaan ataupun domain data yang menimbulkan perkembangan algoritma pembelajaran. Hal ini dikatakan sebagai *supervised learning*, yang mana pada prosesnya harus terdapat data latih yang berlabel, data validitas dan data uji. Data latih berguna untuk menentukan model yang parameternya optimal, dan data validitas digunakan untuk menghindari *overfitting* sedangkan data uji adalah data yang berguna untuk pengujian model dalam memprediksi suatu data (Urrochman et al., 2023).

Machine learning dibagi menjadi empat jenis teknik berdasarkan metode dan cara pembelajaranya. Yang pertama adalah *Supervised Machine Learning,* yaitu teknik yang mempunyai karakteristik harus terdapat label data yang akan dimasukan ke dalam algoritma dengan tujuan untuk menemukan solusi dan juga mempelajari data utama dalam kumpulan data untuk menemukan hasil yang diinginkan. Sehingga jika data dimasukan kedalam algoritma dengan data utama maka teknik ini harus menghasilkan prediksi yang sesuai atau mendekati dengan akurat. Yang kedua adalah *Unsupervised Machine Learning,* yaitu teknik yang mempunyai karakteristik tidak terdapat label data yang akan dimasukan ke dalam algoritma dengan tujuan untuk menemukan solusi dan yang mencoba untuk menemukan struktur tersembunyi dari kumpulan data yang digunakan. Bentuk dasar dari teknik ini adalah pengelompokan yang melibatkan kategori pada data sehingga dapat mengidentifikasi kelompok yang mirip dengan kelompok yang lain. Yang ketiga adalah *Semi-supervised Machine Learning,* yaitu teknik yang berada diantara teknik pertama dan teknik kedua. Hanya terdapat beberapa data yang berlabel dengan data yang banyak. Dalam prosesnya akan menggunakan pengelompokan dalam mengidentifikasi kelas dalam beberapa kumpulan data dengan menggunakan data berlabel yang sama. Teknik ini mempunyai kelebihan yaitu tidak perlu melabeli data satu per satu dan tidak menggunakan waktu yang lama. Yang keempat adalah *Reinforcement Learning,* yaitu teknik yang menggunakan paradigma pembelajaran dalam mengontrol sistem untuk memaksimalkan data type numerik yang menghasilkan tujuan jangka panjang. Tujuan dari teknik ini adalah untuk mengembangkan model algoritma pembelajaran

yang efektif dan efisien sehingga dapat menemukan perbedaan antara kelebihan dan kekurangan pada algoritma yang digunakan (Saputra et al., 2023).

2.5 *Boosting*

Boosting merupakan algoritma yang muncul pada tahun 1990 dikemukakan oleh Schapire pada tahun 1990, lalu pada tahun 1995 dikembangkan oleh Freund dan dikembangkan lagi oleh Schapire pada tahun 1999 yang dipengaruhi dengan adanya teori pembelajaran yang dikemukakan oleh Valiant pada tahun 1989 dimana pada proses klasifikasi masih terbilang lemah sehingga digabungkan untuk menghasilkan prediksi ansamble yang unggul. Pada intinya *boosting* merupakan jenis dari ensemble learning yang merubah model lemah menjadi model yang kuat. Algoritma ini dirancang untuk mengatasi permasalahan klasifikasi, akan tetapi dapat dikembangankan untuk masalah regresi (Fadlisyah & Muhathir, 2023).

Terdapat beberapa algoritma *Boosting* yang dapat digunakan untuk pembuatan model dalam *machine learning* yaitu AdaBoost (Adaptive Boosting), *Gradient Boosting,* XGBoost (*Extreme Gradient Boosting),* LightGBM (*Light Gradient Boosting Machine),* dan *CatBoost* (*Categorical Boosting).*

AdaBoost merupakan algoritma *boosting* yang dikemukakan dari kolaborasi antara Freund dan Schapire pada tahun 1995 dengan dasar konsep untuk peningkatan bobot yang salah dalam klasifikasi. Algoritma ini secara sekuensial membangun model gabungan pohon pada setiap iterasi yang bobotnya selalu di ubah dengan tujuan untuk mengoreksi data yang tidak sesuai dalam pengklasifikasian yang dilakukan sebelumnya. Data yang menerima jumlah bobot lebih besar akan dianggap sebagai data yang salah sedangkan data yang menerima

bobot lebih kecil akan dianggap sebagai data yang benar. Label pengelompokan pada algoritma ini ditentukan oleh pengamatan yang dilakukan oleh model yang dibangun setelah itu dipilih dan diprediksi adanya data baru yang didasari paa bobot yang mayoritas (Rahmi et al., 2023).

Gradient Boosting adalah algoritma ensemble keluarga *decission tree* yang pada proses peningkatan gradien dalam mengoreksi kesalahan sebelumnya dilakukan oleh setiap prediktor. Berbeda dengan AdaBoost karena dalam pelabelan setiap prediktor dilatih mengguanakan kesalahan yang sebelumnya sehingga bobot pelatihan tidak disesuaikan. Tahap pada algoritma ini diawali pada pembangunan pohon klasifikasi yang dilanjutkan melakukan iterasi berulang-ulang sehingga dapat memperbaiki pohon klasifikasi sebelumnya dengan adanya iterasi yang baru. Dalam pendekatan ini yang terpenting adalah penyusutan yang terjadi pada proses yang dilakukan karena setiap pohon dikalikan dengan cepatnya pembelajaran yang mana jumlahnya hanya berkisar 0 hingga 1. Dengan demikian pada setiap iterasi yang dilakukan jumlah prediksi dalam ansambel berkurang. Kelebihan algoritma ini adalah bisa digunakan pada semua jenis data, dan ketika menyelesaikan data pencilan dapat bertahan dengan baik sehingga memiliki hasil prediksi yang akurat (Diantika et al., 2023).

XGBoost *(Extreme Gradient Boosting)* merupakan teknik *machine learning* untuk mengatasi permasalahan regresi atupun klasifikasi dengan didasari dengan GBDT (*Gradient Boosting Decission Tree).* Algoritma ini adalah algoritma yang diusulkan pada tahun 2014 oleh Dr. Tianqi Chen yang berasal dari University of Washington yang merupakan pengembangan dari *Gradient Boosting.* Algoritma

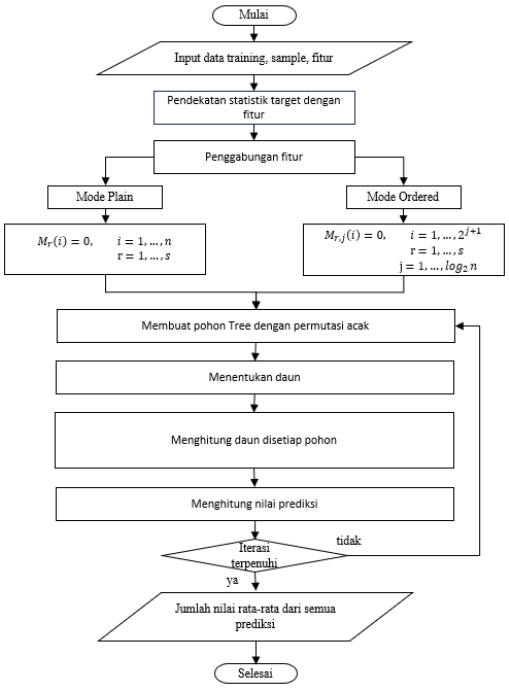
ini didasari dengan konsep penyesuaian parameter pembelajaran dengan cara berulang-ulang yang tujuannya untuk menurunkan *loss function* atau mekanisme evaluasi untuk model. Untuk membangun struktur pohon regresi algoritma ini menggunakan model yang teratur. Oleh karena itu mampu mengurangi model yang kompleks untuk menghindari overfitting sehingga dapat memberikan kinerja yang lebih baik (Agustin et al., 2023). Kemudian pada tahun 2016 C. Guestrin mengembangkan lagi dalam *library gradient boosting* yang terdistribusi menjadi optimal untuk desain yang fleksibel dan efisien. XGBoost digunakan untuk mengatasi masalah dengan mengguanakan beberapa fitur dari kumpulan data latih dalam memprediksi target pada *supervised learning*. Dengan adanya pendekatan ini memungkinkan adanya peningkatan residu yang akan bermanfaat pada peningkatan prediksi yang lebih akurat (Delima et al., 2023).

LightGBM (*Light Gradient Boosting Machine)* merupakan perkembangan baru dari GBDT yang dikembangkan untuk mengatasi tantangan dalam mempelajari pohon keputusan ketika dimensi fitur tinggi dan jumlah data yang besar. Algoritma ini menggunakan histogram untuk memproses data fitur dengan fokus utamanya yaitu mengidentifikasi nilai fitur kontinu kemudian data diiterasi untuk mengakumulasi nilai yang telah diidentifikasi sehingga menemukan nilai titik yang optimal. Algoritma ini juga mempunyai kelebihan mengurangi waktu dan konsumsi memori dalam prosesnya sehingga sangat membantu dalam penggunaan memori yang lebih sedikit dan waktu dalam pelatihan model yang cepat (Wang et al., 2023). LightGBM merupakan penggabungan dari algoritma pengambilan pengambilan sampel satu sisi (GOSS) berbasis gradien untuk mengurangi volume

data dengan berfokus kepada gradien yang signifikan dan mengabaikan gradient nilai yang kecil. Akan tetapi hal ini menimbulkan bias terhadap gradient yang lebih besar sehingga dapat merubah distribusi data asli. Sehingga untuk mengatasinya sampel diambil secara acak dengan data gradien yang kecil. Selain itu algoritma ini juga mengimplementasikan *Exclusive Featrue Bundling* (EFB) dalam mengatasi sebaran outlier dalam kumpulan data. Teknik ini adalah gabungan fitur tertentu yang menghasilkan pengurangan jumlah fitur namun tetap mempertahankan fitur yang krusial (Ivanoti et al., 2023).

CatBoost (Categorical Boosting) merupakan algoritma yang seperti *gradient boosting* dan *decission tree* yang memberikan nilai hasil prediksi akurat. Dengan mengimplementasikan pohon yang simetris maka algortima ini dapat digunakan untuk menyelesaikan permasalahan fitur kategorikal dan jika dibandingkan dengan algoritma lain yang kuat, algoritma *CatBoost* akan lebih cepat dalam prosesnya. Algoritma ini memanfaatkan teknik *boosting* untuk meningkatkan tingkat prediksi akurasi yang tinggi dengan menggabungkan model yang lemah supaya dapat menemukan model yang kuat (Syandika & Yustanti, 2023). Dengan perkembangan algoritma yang sudah meluas dalam pengolahan data, algoritma *CatBoost* adalah algoritma yang paling kompeten dan bekerja dengan baik pada data yang besar dan juga memiliki persyaratan latensi yang rendah. Algoritma ini digunakan untuk model ansambel yang iterasinya dilakukan berulang-ulang. Algoritma mempelajari pohon pertama pada iterasi pertama untuk mengurangi kesalahan karena biasanya terdapat kesalahan yang signifikan sehingga untuk meningkatkan nilai prediksi tidak baik dengan cara membangun pohon yang

besar. Sedangkan pada iterasi kedua algoritma mempelajari kesalahan pada iterasi pertama untuk mengurangi hasil kesalahanya (Barua et al., 2021).



Gambar 2. 1 Algoritma *CatBoost*

Pada gambar 2.1 memperlihatkan bahwa algoritma *CatBoost* termasuk menggunakan trik percepatan dalam membangun pohon yang kompleksitasnya dikurangi. Pada trik percepetan yang digunakan dilakukan menggunakan mode ordered. Jika tidak menggunakan trik percepatan maka yang dilakukan menggunakan mode plain. Dengan menggunakan mode ordered nilai prediksi sebelumnya akan dipertahankan yang diaproksimasi untuk contoh *iterasi* selanjutnya. Sehingga dengan demikian jumlah prediksi tidak akan lebih besar.

Pada algoritma *CatBoost* terdapat fungsi Buildtree yang menjelaskan secara rinci dalam langkah mebangun pohon. Fungsi tersebut digunakan untuk menggambarkan percocokan antara contoh pelatihan iterasi dengan daun yang

digunakan dengan permutasi acak. Misalnya jika terdapat contoh dengan fitur x yang dijadikan masukan untuk pelatihan maka akan dihitung dengan mode ordered berdasarkan permutasi acak yang didapatkan. Kemudian setelah itu akan memilih daun pada pohon dengan fitur x yang sesuai dengan pendekatan target ordered yang diperoleh. Jika sudah diperoleh maka algoritma dalam pencarian daun digunakan sebagai ganti dari permutasi acak. Sehingga pendekatan target ordered yang didapatkan akan digunakan dalam penerapan model yang sudah dilatih untuk contoh baru pada tahap pengujian model yang selanjutnya. Begitu juga seterusnya hingga mendapatakan model yang dianggap paling terbaik dengan performa prediksi yang akurat.

2.6 Metode Evaluasi

Hasil yang didapatkan dari pemodelan diidentifikasi dari data pelatihan yang sudah diproses untuk dievaluasi. Dalam evaluasi yang dilakukan pada model yang sudah dibuat pada proses pelatihan maka berdasarkan data pelatihan yang dimasukan akan memprediksi label kategori sebagai *output* yang hasilnya "1 atau 0" dan "ya atau tidak". Untuk mengevaluasi kinerja model terdapat beberapa teknik yang harus diterapkan untuk mengetahui kesalaha yang ada pada model seperti *Confusion Matrix* yaitu *Accuracy, Precision dan Recall, F1 score*, dan *Area under ROC curve (AUC)* (Fadlisyah & Muhathir, 2023).

Confusion Matrix digunakan untuk melihat performa pada model dari hasil analisis evaluasi. Teknik ini menggunakan perhitungan nilai *accuration, presicion, recall dan f1 score* dari model *machine learning.* Pada perhitungan *confusion*

matrix terdapat empat kasus dalam kondisi tertentu. Berikut adalah tabel *confusion matrix* (Delima et al., 2023).

Tabel 2. 2 Kondisi *Confusion Matrix*

		Predict Values	
		1	**0**
Aktual Values	**1**	*True Positif (TP)*	*False Negatif (FN)*
	0	*False Positif (FP)*	*True Negatif (TN)*

True positives (TP) adalah kondisi ketika hasil pada data aktual benar dan hasil prediksi juga benar. *True Negatif (TN)* adalah kondisi ketika hasil pada data aktual salah sedangkan hasil prediksi benar. *False Positif (FP)* adalah kondisi ketika hasil prediksi salah sedangkan hasil pada data aktual benar. *False Negatif (FN)* adalah kondisi ketika hasil prediksi salah dan hasil pada data aktual juga salah (Diantika et al., 2023).

Accuracy merupakan metrik yang digunakan dalam kondisi saat kategori variabel target yang ada dalam data dianggap kira-kira seimbang. Selain itu metrik ini adalah metrik klasifikasi yang paling sederhana untuk digunakan dan dianggap dapat menentukan prediksi benar dengan adanya jumlah total dari prediksi yang ditemukan (Agustin et al., 2023).

Precision dan *recall* merupakan sesuatu perhitungan yang sangat penting dalam pengambilan informasi dengan kategori kelas positif dan negatif yang mana kelas positif lebih penting dibandingkan dengan kelas negatif. Metrik presisi digunakan sebagai perhitunagn yang dapat mengatasi keterbatasannya akurasi dan juga dapat digunakan untuk menentukan kebenaran proporsi prediksi kelas positif. Sedangkan *recall* adalah nilai persen dari total positif yang diperkirakan positif dengan benar yang tujuannya untuk menghitung kebenaran proporsi prediksi

positif realita yang salah diprediksi. Hal ini dapat dihitung dengan prediksi yang sebenarnya benar terhadap total jumlah positif yang diprediksi dengan benar *true positif* ataupun *false negatif* (Nainggolan & Sinaga, 2023).

F1 score merupakan metrik untuk evaluasi model klasifikasi dengan type data biner untuk kelas positif berdasarkan prediksi yang dilakukan. Perhitungan ini dibantu dengan perhitungan *presicion* dan *recall* karena *f1 score* merupakan nilai tunggal yang mewakili keduanya sehingga perhitungan dapat dijadiakn sebagai rata-rata harmonik dari keduannya dengan masing-masing diberikan bobot yang sama. *F1 score* digunakan dengan baik karena dapat bekerja secara baik pada data yang besar dan tidak seimbang (Nababan et al., 2023).

Area Under Curve (AUC) dan *Receiver Operating Characteristic (ROC)* digunakan dalam pengolahan data untuk memvisualisasikan kinerja model klasifikasi. Sehingga AUC dan ROC sangat penting penggunaanya dalam evaluasi kinerja model klasifikasi. ROC mewakili grafik yang menujukan kinerja model pada tingkat yang berbeda. Terdapat dua kurva diplot parameter yaitu TPR *(True Positive Rate)* dan FPR *(False Positive Rate).* Untuk menghitung nilai disetiap titik kurva ROC dapat dilakukan dengan evaluasi model regresi berulang-ulang dengan data yang berbeda. Akan tetapi ROC tidak efisien dibandingkan dengan AUC karena AUC menghitung area dua dimensi dibawah seluruh kurva ROC. Nilai AUC adalah antara 0 hingga 1 sehingga jika prediksi 100% salah maka nilai AUC 0.0 sedangkan jika prediksi 100% benar makan nilai AUC 1.0. Untuk mengukur seberapa baik prediksi yang diurutkan berdasarkan nilai absolutnya harus menggunakan AUC dalam pengukurannya. Nilai diagnosa ROC dimulai dari 0.50-

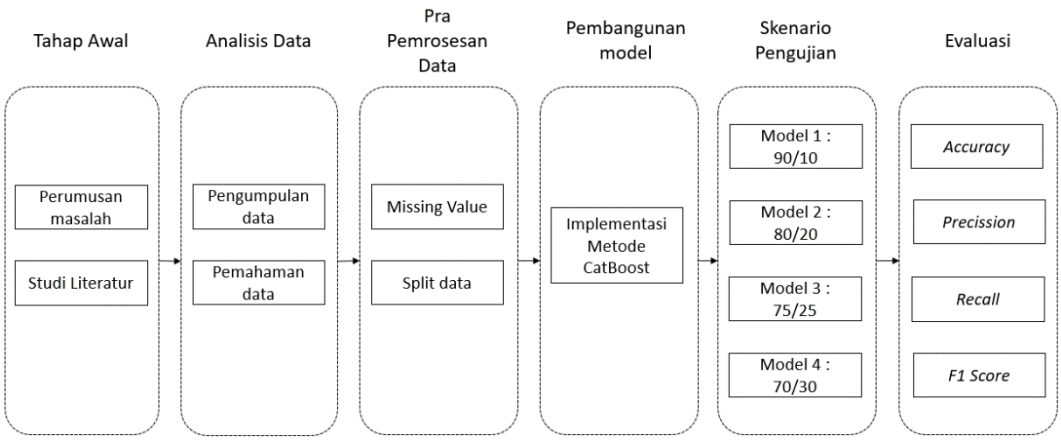
0.60 yang memiliki kriteria klasifikasi salah, 0.60-0.70 memiliki kriteria klasifikasi buruk, nilai 0.70-0.80 memiliki kriteria klasifikasi cukup, nilai 0.80-0.90 memiliki kriteria klasifikasi baik, dan nilai 0.90-1.00 adalah kriteria klasifikasi sangat baik (Nababan et al., 2023).

BAB III

DESAIN DAN IMPLEMENTASI

3.1 Desain Penelitian

Desain penelitian merupakan sebuah alur proses dari awal hingga akhir yang dilakukan untuk penelitian. Pada penelitian ini sistem dimulai dari pendekatan penelitian rumusan masalah dan diteruskan analisis data lalu setelah itu pengolahan data dengan pra-pemrosesan dan pemrosesan data dengan menggunakan Jupyter Notebook dilanjutkan dengan evaluasi model dan pembahasan skenario pengujian hingga penarikan kesimpulan.



Gambar 3. 1 Desain Penelitian

3.2 Tahap Awal

Pada tahap ini ada 2 tahap yaitu tahap perumusan masalah dan studi literatur. Tujuan perumusan masalah untuk mengetahui masalah yang akan diteliti pada penelitian ini. Sedangkan studi literatur dilakukan dengan tujuan untuk mengumpulkan informasi-informasi penting mengenai permasalahan yang

berkaitan dalam proses penelitian yang akan dilakukan serta mendalami pemahaman mengenai pembahasan yang akan diteliti.

3.2.1 Perumusan Masalah

Pada bagian perumusan masalah merumuskan permasalahan tentang topik kualitas udara di Jakarta dengan data publik yang diambil dari kaggle. Analisis dan pemahaman terhadap informasi-informasi dan data pada penelitain ini dilakukan dengan tujuan untuk menemukan masalah yang terkait dengan indeks kualitas udara. Sehingga dengan melakukan perumusan masalah peneliti dapat terbantu dalam membuat perencanaan yang akan dilakukan dalam penelitian dan juga penelitian yang akan dilakukan terarah dengan tujuan yang jelas.

3.2.2 Studi Literatur

Studi literatur mengkaji berbagai referensi yang diambil dari jurnal ataupun website yang berkaitan dengan topik permasalahan yang diangkat pada penelitain ini. Sehingga referensi mengenai indeks kualitas udara, algoritma *CatBoost*, metode evaluasi dan juga datashet dikaji terlebih dahulu dengan tujuan untuk mendapatkan informasi yang relevan dan dapat diterapkan untuk penelitian yang akan dilakukan.

3.3 Analisis Data Penelitian

3.3.1 Pengumpulan Data

Pengumpulan data akan mengumpulkan informasi mengenai topik kualitas udara. Data yang dikumpulkan adalah data indeks standar pencemar udara di spku wilayah Jakarta yang diambil dari kaggle. Data ini dapat digunakan oleh siapapun karena data ini bersifat publik dan dapat di unduh dengan mudah melalui website

kaggle. Akan tetapi data ini adalah data yang valid dan dapat dijadikan sebagai bahan untuk pengolahan data yang kemudian dapat membantu untuk pengambilan keputusan. Sehingga data indeks standar pencemar udara di spku Jakarta dapat digunakan untuk penelitian tugas akhir. Data tersebut berjumlah 1830 dan terdiri dari 10 fitur yaitu tanggal, stasiun, pm10, so2, co, 03, no2, max, critical dan juga categori. Tipe data di masing-masing fitur berbeda, ada yang bertipe object ada juga yang bertipe numerik. Data yang dukumpulkan akan digunakan untuk inputan dalam proses pembangunan model.

	tanggal	stasiun	pm10	so2	co	o3	no2	max	critical	categori
0	2020-01-01	DKI1 (Bunderan HI)	30	20	10	32	9	32.0	O3	BAIK
1	2020-01-02	DKI1 (Bunderan HI)	27	22	12	29	8	29.0	O3	BAIK
2	2020-01-03	DKI1 (Bunderan HI)	39	22	14	32	10	39.0	PM10	BAIK
3	2020-01-04	DKI1 (Bunderan HI)	34	22	14	38	10	38.0	O3	BAIK
4	2020-01-05	DKI1 (Bunderan HI)	35	22	12	31	9	35.0	PM10	BAIK
...	...	...	...	...	...	...	...	...	...	...
1825	2020-12-27	DKI5 (Kebon Jeruk) Jakarta Barat	18	32	4	41	---	41.0	CO	BAIK
1826	2020-12-28	DKI5 (Kebon Jeruk) Jakarta Barat	22	33	5	35	3	35.0	CO	BAIK
1827	2020-12-29	DKI5 (Kebon Jeruk) Jakarta Barat	15	28	4	27	---	28.0	PM25	BAIK
1828	2020-12-30	DKI5 (Kebon Jeruk) Jakarta Barat	16	7	3	21	2	21.0	CO	BAIK
1829	2020-12-31	DKI5 (Kebon Jeruk) Jakarta Barat	18	13	6	24	3	24.0	CO	BAIK

1830 rows × 10 columns

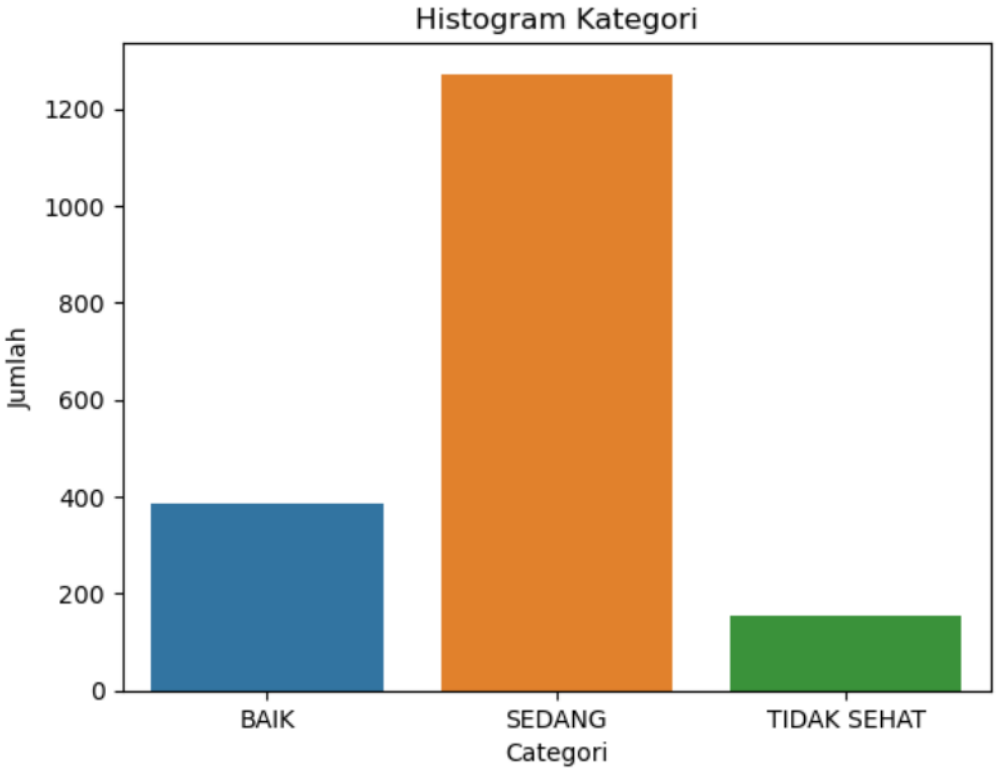
Gambar 3. 2 Dataset yang digunakan

3.3.2 Pemahaman Data

Pada dataset terdapat 10 fitur, yaitu tanggal, stasiun, pm10, so2, co, o3, no2, max, critical dan juga categori dimana fitur tanggal ini merupakan tanggal pengambilan data kualitas udara, fitur stasiun yaitu lokasi atau tempat pengambilan data kualitas udara, fitur pm10 yaitu parameter data nilai partikulat salah satu yang diukur, so2 adalah parameter data nilai sulfida pengaruh kualitas udara yang didapatkan, fitur co merupakan parameter data nilai carbon monoksida pengaruh kualitas udara yang didapatkan, fitur o3 merupakan parameter ozon yang berpengaruh kepada kualitas udara, fitur no2 merupakan parameter nilai nitrogen dioksida yang berpengaruh kepada kualitas udara, fitur max merupakan nilai ukur paling tinggi dari semua parameter dengan waktu pengambilan yang sama, fitur critical merupakan parameter yang hasil pengukurannya paling tinggi, dan yang terakhir adalah fitur categori yang merupakan hasil perhitungan indeks pencemaran udara. Terdapat fitur yang bertipe numerik yaitu fitur pm10, so2, co, o3, no2 dan juga max. Dan terdapat fitur yang bertipe object atau kategori yaitu, tanggal, stasiun, critical, dan juga categori. Pada data ini terdapat 3 kategori untuk indeks kualitas udara yang ada pada dataset yaitu udara baik, sedang dan tidak sehat. Hal ini dikarenakan pada dataset yang digunakan nilai paling tinggi pada indeks kualitas udara adalah 191 sehingga tidak ada kategori yang berbahaya diatas 300. Berikut adalah tabel keterangan fitur dan gambar histogram analisis distribusi dataset pada fitur kategori.

Tabel 3. 1 Fitur Dataset

No	Atribut	Keterangan
1	Tanggal	Tanggal pengambilan data
2	Stasiun	Lokasi pengambilan data
3	Pm10	Partikulat salah satu yang diukur
4	So2	Sulfida (dalam bentuk SO2)
5	Co	Carbon Monoksida
6	O3	Ozon
7	No2	Nitrogen Dioksida
8	*Max*	Nilai ukur paling tinggi dari seluruh parameter yang diukur dalam waktu yang sama
9	*Critical*	Parameter yang hasil pengukurannya paling tinggi
10	*Categori*	Kategori hasil perhitungan indeks standar pencemaran udara



Gambar 3. 3 Histogram Distribusi Data Kategori

Pada Gambar 3.3 terlihat bahwa pada atribut categori terdapat beberapa kelas yaitu kelas BAIK, SEDANG, dan TIDAK SEHAT. Kelas tersebut merupakan kelas kategori indeks kualitas udara yang ada pada dataset yang digunakan untuk prediksi pada penelitain ini. sehingga kelas tersebut adalah kelas yang dijadikan

sebagai target untuk prediksi. Jumlah data pada kelas BAIK adalah 385 data, jumlah data pada kelas SEDANG adalah 1272 data dan jumlah data pada kelas kategori TIDAK SEHAT adalah 154 data.

3.4 Pra Pemrosesan Data

Pada proses pengolahan data terdapat beberapa langkah yang akan dilakukan pada penelitian ini. Langkah pertama dalam pengolahan data pada penelitian ini adalah pra pemrosesan data yang mana digunakan untuk memperbaiki dan menganalisis data dari data yang tidak sesuai dan juga membuat data supaya cocok dengan pemodelan *machine learning* yang digunakan. Pada pra pemrosesan data terdapat tahapan penting yang harus dilakukan yaitu tahap *misssing value*, dan membagi data.

3.4.1 *Missing Value*

Dalam melakukan pra-pemrosesan pengolahan data biasanya terdapat data yang tidak seimbang atau data yang diluar jangkauan. Sehingga perlu adanya perbaikan data dengan pemrosesan untuk menemukan dan memperbaiki data yang tidak sesuai keakuratan ataupun terdapat nilai yang hilang dalam dataset. Pada penelitian ini data yang tidak sesuai dan data yang tidak digunakan untuk akan dihilangkan. Data yang akan dihilangkan adalah fitur “tanggal”, “stasiun” dan “critical”. Hal ini dilakukan karena fitur tersebut merupakan tipe object atau tipe kategori dan pengaruh kepada indeks kualitas udara juga sedikit, bahkan hampir tidak ada. Berikut merupakan data sebelum dan sesudah *missing value.*

	tanggal	stasiun	pm10	so2	co	o3	no2	max	critical	categori
0	2020-01-01	DKI1 (Bunderan HI)	30	20	10	32	9	32.0	O3	BAIK
1	2020-01-02	DKI1 (Bunderan HI)	27	22	12	29	8	29.0	O3	BAIK
2	2020-01-03	DKI1 (Bunderan HI)	39	22	14	32	10	39.0	PM10	BAIK
3	2020-01-04	DKI1 (Bunderan HI)	34	22	14	38	10	38.0	O3	BAIK
4	2020-01-05	DKI1 (Bunderan HI)	35	22	12	31	9	35.0	PM10	BAIK

Gambar 3. 4 Data Sebelum *Missing Value*

	pm10	so2	co	o3	no2	max	categori
0	30	20	10	32	9	32.0	BAIK
1	27	22	12	29	8	29.0	BAIK
2	39	22	14	32	10	39.0	BAIK
3	34	22	14	38	10	38.0	BAIK
4	35	22	12	31	9	35.0	BAIK

Gambar 3. 5 Data Sesudah *Missing Value*

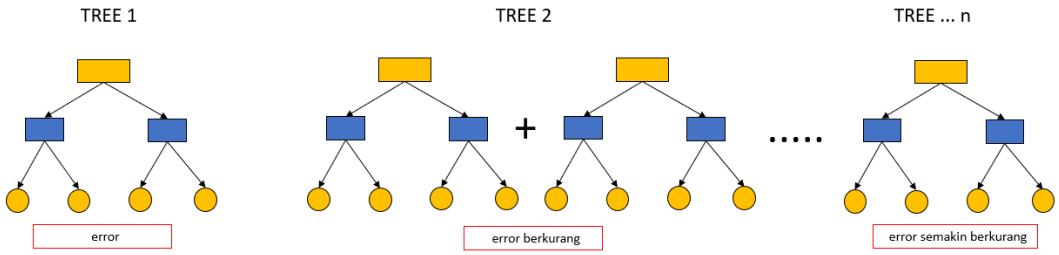
3.4.2 Split Data

Split data pada pemrosesan data sangat penting dilakukan pada penelitian ini guna untuk membagi data menjadi dua, yang pertama data untuk *training* dan yang kedua untuk *testing*. Data *training* digunakan sebagai data pelatihan pembangunan model, sedangkan data *testing* untuk menguji model yang sudah dibuat dan juga untuk mengevaluasi kinerja model. Untuk pembagian split data pada penelitian ini dibagi menjadi 4 model yaitu model 1 dengan pembagian 90/10 dalam artian untuk *training* 90% dan untuk *testing* 10%. Model 2 dengan pembagian 80/20 dengan artian untuk *training* 80% dan untuk *testing* 20%. Model 3 dengan pembagian 75/25 dengan artian untuk *training* 75% dan untuk *testing* 25%. Dan model 4 adalah pembagian 70/30 dengan artian untuk *training* 70% dan

untuk *testing* 30%. Pembagian ini dapat memberikan kinerja yang baik untuk evaluasi keempat model dan juga dapat mencegah overfitting. Selain itu juga dapat memberikan hasil performa yang akurat. Dalam proses pembentukan data *training* penelitian ini menggunakan masing-masing komposisi dari model yang telah ditentukan dari seluruh jumlah data yang digunakan. Berdasarkan data *training* yang diperoleh maka dapat digunakan untuk melatih model *CatBoost*. Sedangkan data *testing* dengan komposisi masing-masing model yang ada adalah sisa data dari data keseluruhan yang sudah digunakan untuk *training*. Sehingga data yang digunakan untuk *training* tidak digunakan untuk *testing*. Dengan demikian data *testing* dapat digunakan untuk menguji model *CatBoost* dengan tingkat akurasi yang terbaik. Dengan keempat model yang telah ditentukan akan diambil akurasi yang terbaik untuk proses pengujian dan model akan digunakan dalam prediksi indeks kualitas udara.

3.5 Pembangunan Model *CatBoost*

Pada pemrosesan data terdapat langkah yang sangat penting untuk dilakukan yaitu pembangunan model. Pembangunan model pada penelitian ini adalah dimana proses algoritma *CatBoost* mempelajari data pelatihan yang sudah ditentukan pada proses sebelumnya. Pekerjaan pelatihan model adalah untuk menyesuaikan bobot dan bias terbaik ke algoritma dengan tujuan untuk mengurangi fungsi kerugian *loss function* yang ada dalam rentang nilai prediksi.



Gambar 3. 6 Visualisasi *CatBoost*

3.5.1 Memilih Struktur Pohon

Pemilihan struktur pohon terbaik didasari dengan perhitungan split yang berbeda. Dengan demikian pohon akan dibangun menggunakan data split tersebut dan menetapkan nilai daun yang diperoleh serta menilai pohon yang terbaik. Pada pemilihan pohon *CatBoost* memiliki dua mode yaitu plain dan ordered. Mode plain adalah mode kombinasi algoritma GBDT standar dengan ordered target statistic. Selama proses pembelajaran dengan mengguanakan mode ordered boosting dipertahankan model pedukung dengan prediksi saat itu berdasarkan contoh pertama yang dilakukan. Dengan menerapkan gradient yang setiap contoh dihitung berdasarkan contoh sebelumnya. Gradient yang sesuai dengan prediksi pertama dihitung dengan menggunakan persamaan berikut (Dewi, 2021) :

$$grad_{r,j}(i) = \frac{\partial L(y_i,s)}{\partial s}|_{s=Mr,j\ (i)} \quad (3.1)$$

Keterangan :
L : fungsi kerugian
y_i : target
$M_{r,j}(i)$: prediksi saat ini

Pada saat pertambahan pohon baru ke ensemble masing-masing skor dihitung dari banyaknya kandidat yang terpisah. Pada hal ini gradient diestimasi dengan

menggunakan *cosine similarity cos*. Dari fungsi skor itulah diperlukan untuk memilih pohon diantara pohon-pohon kandidat yang ada. Misalkan diberikan kandidat T_C maka fungsi skor *cosine similarity cos* dapat dihitung dengan menggunakan rumus berikut (Dewi, 2021) :

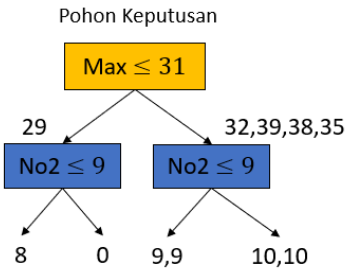
$$Cosine = \frac{\sum_{i=1}^{n} w_i \cdot \Delta_i \cdot g_i}{\sqrt{\sum_{i=1}^{n} w_i \Delta_i^2} \cdot \sqrt{\sum_{i=1}^{n} w_i g_i^2}} \quad (3.2)$$

Keterangan :
w_i : bobot objek ke-i
g_i : gradient pada objek ke-i yang sesuai dengan fungsi kerugian
Δ_i : nilai daun pada objek ke-i

Ketika struktur pohon T_C dibangun maka pohon tersebut akan digunakan sebagai peningkatan semua model yang selanjutnya. Dengan demikian berarti struktur pohon umum akan digunakan pada semua model akan tetapi pohon tersebut ditambahkan kepada nilai awal yang berbeda dengan himpunan nilai daun yang berbeda tergantung pada hasil yang ditemukan. Pohon yang dibentuk oleh *CatBoost* adalah pohon simetri atau biasa disebut dengan *Oblivious Decision Tree.* Sehingga pohon yang dibuat seimbang dan tidak terlalu *overfitting.* Pohon yang dibangun adalah pohon biner dengan setiap tingkatan menggunakan kriteria pemisah yang sama dan menggunakan fitur yang sama. Sehingga dapat membagi sampel kiri dan kanan dengan baik dan memiliki kecepatan prediksi 10 kali lebih cepat dibanding dengan pohon non simetris. Berikut adalah contoh pohon simetri dengan tabel keputusannya (Dewi, 2021).

Tabel Data

Pm10	So2	Co	O3	No2	Max	Categori
30	20	10	32	9	32	0
27	22	12	29	8	29	0
39	22	14	32	10	39	0
34	22	14	38	10	38	0
35	22	12	31	9	35	0



Gambar 3. 7 Pohon Simetri

Berdasarkan contoh pohon keputusan diatas perhitungan untuk memilih struktur pohon dengan *cosine* similarity adalah sebagai berikut :

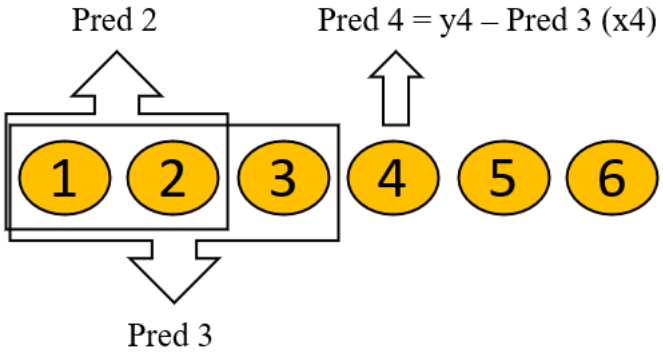
$$\textit{Cosine Similarity} = \frac{(32\ x\ 32)+(29\ x\ 0)+(39\ x\ 39)+(38+38)+(35+35)}{\sqrt{32^2+29^2+39^2+28^2+35^2}\ .\ \sqrt{32^2+0^2+39^2+38^2+35^2}}$$

$$= \frac{5234}{\sqrt{6055}\ .\ \sqrt{5214}}$$

$$= 0{,}958$$

3.5.2 *Ordered Boosting*

Pada tahap pelatihan terdapat residu yang harus dihitung dengan menggunakan model pelatihan tanpa adanya contoh pelatihan. Dengan perhitungan yang dilakukan residu akan tidak bias pada pengguanan semua contoh pelatihan. Hal ini dapat diimplementasikan dengan menggunakan prinsip ordering sehingga dalam mode ordered boosting dilakukan permutasi secara acak dari data pelatihan dengan mempertahankan nilai model yang berbeda.



Gambar 3. 8 Visualisasi Ordered Boosting

3.5.3 Menghitung Nilai Daun

Nilai daun dihitung dengan menggunakan estimasi gradien untuk semua pohon yang dibangun. Nilai daun yang dihasilkan dari model akhir dihitung dengan menggunakan prosedur gradient boosting standart yang sama pada kedua mode. Penerapan pada *CatBoost* adalah nilai pertama ditambah 1 permutasi acak dari data pelatihan. Permutasi kedua hingga akhir digunakan untuk pemisahan yang mendefinisikan struktur pohon sedangkan permutasi awal akan digunakan untuk memilih daun. Pada tahap evaluasi nilai daun dihasilkan secara individu dengan rata-rata gradient boosting dari contoh sebelumnya. sebuah list dibentuk dari kandidat pasangan fitur split untuk ditugaskan ke daun sebagai split. Kemudian sejumlah fungsi dihitung untuk setiap objek, dengan syarat semua kandidat yang diperoleh dari langkah pertama sudah ditugaskan ke daun. Sehingga split dengan nilai yang terkecil akan dipilih. Begitu juga seterusnya hingga selesai. Kandidat dipilih berdasarkan data *training* dan *testing* pada awal tahapan. Hasil akhir pohon yang dihasilkan adalah pohon yang simetri. Sehingga setiap indeks daun dapat

dikodekan sebagai vektor biner dengan panjang yang sama sesuai kedalaman pohon. Adapun langkah-langkah untuk menghitung daun sebagai berikut :

1. Membentuk list dari kandidat pasangan fitur pembagian untuk ditugaskan sebagai daun pembagian
2. Menghitung fungsi untuk setiap objek, dengan syarat kandidat yang didapatkan dari langkah awal sudah ditugaskan sebagai daun.
3. Pembagian dengan nilai terkecil dipilih.

Langkah-langkah proses pelatihan *CatBoost* dilakukan adalah sebagai berikut :

1. Inisialisasi model dan parameter

   Model :

   Model 1 = 90/10

   Model 2 = 80/20

   Model 3 = 75/25

   Model 4 = 70/30

   Parameter :

   depth = 6, 8, 10

   learning rate = 0,1 dan 0,01

   Iterations = 500, 1000, 1500

   l2_leaf_reg = 1, 2, 3

2. Hitung residu awal

   Nilai konstan target yang dijadikan sebagai nilai prediksi pertama = 0

   Nilai observed adalah nilai target pada data

   Residu awal = observed – prediksi pertama

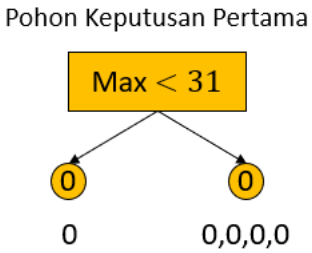
   = 0 – 0

   = 0

Pm10	So2	Co	O3	No2	Max	Categori	Prediksi	Residu
30	20	10	32	9	32	0	0	0
27	22	12	29	8	29	0	0	0
39	22	14	32	10	39	0	0	0
34	22	14	38	10	38	0	0	0
35	22	12	31	9	35	0	0	0

Gambar 3. 9 Nilai Residu

3. Buat pohon keputusan pertama

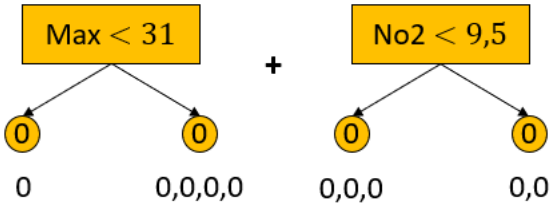


Gambar 3. 10 Pohon Keputusan Pertama

4. Buat pohon keputusan selanjutnya

   Prediksi baru = prediksi sebelumnya + learning rate x data terbaru pada residu

   = 0 + 0,1 x 0 = 0



Gambar 3. 11 Pohon Keputusan Selanjutnya

5. Menghitung prediksi akhir

$$\text{Prediksi akhir} = \frac{Jumlah\ Residual}{jumlah\ prediksi\ sebelumnya\ .\ (1 - prediksi\ sebelumnya)}$$

$$= \frac{0}{0\ .\ 1}$$

$$= 0$$

3.6 Skenario Pengujian

Pada penelitian ini peneliti membagi data menjadi 4 model. Model 1 adalah model dengan perbandingan 90:10 yaitu 90% *training* dan 10% *testing*. Model 2 adalah model dengan perbandingan 80:20 yaitu 80% *training* dan 20% *testing*. Model 3 adalah model perbandingan 75:25 yaitu 75% *training* dan 25% *testing*. Model 4 adalah perbandingan 70:30 yaitu 70% *training* dan 30% *testing*. Dengan perbedaan perbandingan maka akan ditemukan prediksi yang lebih akurat.

Tabel 3. 2 Skenario Pengujian

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

Pada masing-masing model akan diimplementasikan dengan algoritma *CatBoost* dengan nilai parameter yang sama. Parameter yang diinisialisasi adalah

parameter *iterations*, *depth*, *learning_rate* dan juga *l2_leaf_reg*. Untuk penentuan parameter itu sendiri peneliti mengacu kepada penelitan yang dilakukan oleh Johannes Christian dan temannnya dalam memprediksi pembatalan pesanan kamar hotel yang datanya berjumlah 16.494 pada tahun 2022 dengan hasil penemuan parameter terbaik pada *iterations* 2000, *depth* 10, learning rate 0,012 dan *l2_leaf_reg* sebanyak 1 (Christian et al., 2022). Sehingga dengan perbandingan ini akan menghasilkan beberapa prediksi yang berbeda. Dengan demikian hasil prediksi dengan nilai akurasi terbaik akan dipilih dan dijadikan sebagai model terbaik. Parameter *iterations* berpengaruh kepada jumlah pohon akhir yang memungkinkan kurang dari jumlah yang ditentukan dalam parameter ini dan dapat memecahkan masalah pada proses pelatihan. Parameter *depth* berpengaruh kepada kedalaman pohon pada sistem dengan kisaran nilai yang didukung bergantung kepada tipe unit pemrosesan dan tipe fungsi kerugian. Parameter *learning_rate* berpengaruh kepada kecepatan pembelajaran yang digunakan untuk mengurangi langkah gradien. Parameter *l2_leaf_reg* berpengaruh kepada koefisien regularisasi dengan nilai positif apapun diperbolehkan.

Tabel 3. 3 Parameter

No	Parameter	Nilai	Penjelasan
1	*iterations*	500, 1000, 1500	Jumlah maksimum pohon yang dibuat
2	*depth*	6, 8, 10	Jumlah kedalaman pohon
3	*learning_rate*	0,1 dan 0,01	Besaran tingkat pembelajaran
4	*l2_leaf_reg*	1, 2, 3	Koefisisen l2 sebagai koefisien regularisasi model

3.7 Evaluasi

Pada tahap evaluasi model pada penelitian ini akan menggunakan *confusion matrix* dan setelah berhasil mendapatkannya akan dihitung dengan perhitungan

terhadap nilai *accuracy, precission, recall,* dan *f1 score.* Nilai *accuracy* dihitung untuk mengevaluasi jumlah data yang diklasifikasikan dengana benar. Nilai *recall* dihitung untuk mengevaluasi cakupan sebuah model dalam melakukan prediksi kategori tertentu dan nilai *precission* dihitung untuk mengevaluasi ketepatan model dalam memprediksi kategori tertentu juga. Sedangkan nilai *f1 score* dihitung untuk mengetahui rata rata dari nilai *precission* dan *recall.*

1. *Accuracy*

   Accuracy merupakan tingkatan dekatnya nilai prediksi yang dihasilkan model serta nilai aktual yang ada pada data dengan mengetahui jumlah data klasifikasi yang benar. Berikut adalah rumus perhitungannya (Amalia et al., 2022).

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \quad (3.3)$$

2. *Precission*

   Precission merupakan rasio nilai relevan berdasarkan seluruh nilai terpilih dengan melakukan perbandingan jumlah informasi relevan terhadap jumlah seluruh informasi yang terpilih. Berikut adalah rumus perhitungannya (Amalia et al., 2022).

$$Precission = \frac{TP}{TP+FP}\ x\ 100\% \quad (3.4)$$

3. *Recall*

   Recall merupakan terpilihnya rasio nilai relevan berdasarkan jumlah nilai relevan yang ada dengan melakukan perbandingan jumlah informasi relevan terhadap jumlah seluruh informasi relevan yang ada dalam informasi. Berikut adalah rumus perhitungannya (Amalia et al., 2022).

$$Recall = \frac{TP}{TP+FN}\ x\ 100\% \quad (3.5)$$

4. *F1score*

   F1 score merupakan nilai rata-rata dari nilai *precission* dan *recall* . Berikut adalah rumus perhitungannya (Amalia et al., 2022).

$$F1\ score = 2\ \frac{precission\ x\ recall}{precission+recall}\ x\ 100\% \quad (3.6)$$

BAB IV

UJI COBA DAN PEMBAHASAN

4.1 Hasil Uji Coba

Pada hasil uji coba ini terdapat penganalisisan hasil uji coba pengujian model yang berdasarakan dengan skenario pengujian, dimana model terbagi menjadi 4 bagian. Masing-masing model akan diuji dengan menggunakan beberapa parameter *depth*, learning rate, *iterations* dan l2 leaf reg. Sehingga dalam pengujian pada semua model menggunakan nilai parameter yang sama. Nilai parameter *depth* yang digunakan adalah 6, 8 dan 10. Nilai parameter learning rate yang digunakan adalah 0,1 dan 0,01. Nilai parameter *iterations* yang digunakan adalah 500, 1000 dan 1500. Sedangkan nilai parameter l2 leaf reg yang digunakan adalah 1, 2 dan 3. Pada proses pengujian ini dilakukan pencarian parameter yang optimal dengan menggunakan *GridSearchCV*. *GridSearchCV* digunakan untuk mencari parameter terbaik dalam membentuk model yang optimal sehingga menghasilkan nilai akurasi yang terbaik dengan *cross validation.* Nilai yang dihasilkan adalah berupa skor terbaik yang merupakan skor dari rata-rata akurasi silang tertinggi (Priya, 2021). Tujuan pengujian ini adalah mengetahui performa metode *CatBoost* dalam prediksi indeks kualitas udara dengan menggunakan data sebanyak 1811. Setelah mendapatkan nilai akurasi menggunakan metode *CatBoost,* selanjutnya model yang telah dibangun perlu dilakukan evaluasi untuk mengetahui kinerja metode *CatBoost* dalam mengklasifikasi prediksi indeks kualitas udara menggunakan *confussion matriks* pada sub bab 3.7.

4.1.1 Pengujian Model 1

Pengujian menggunakan data *training* sebanyak 1629 dan data *testing* sebanyak 182. Pengujian yang dilakukan mendapatkan kombinasi parameter terbaik yang dihasilkan dengan menggunakan *GrideSearchCV*. Berikut adalah nilai hasil pengujian parameter untuk kombinasi parameter yang terbaik menggunakan *GridSearchCV*.

Tabel 4. 1 Hasil Rata-rata Skor Uji *Cosine similarity* Model 1

No	**params**				***mean_test_score (cosine similarity)***
	depth	***iterations***	***l2_leaf_reg***	***learning_rate***	
1	6	500	1	0,1	0.957643
2	6	500	1	0,01	0.945979
3	6	500	2	0,1	0.953959
4	6	500	2	0,01	0.945979
5	6	500	3	0,1	0.953346
6	6	500	3	0,01	0.945979
7	6	1000	1	0,1	0.957643
8	6	1000	1	0,01	0.947821
9	6	1000	2	0,1	0.956415
10	6	1000	2	0,01	0.947207
11	6	1000	3	0,1	0.956415
12	6	1000	3	0,01	0.946593
13	6	1500	1	0,1	0.958870
14	6	1500	1	0,01	0.950890
15	6	1500	2	0,1	0.958257
16	6	1500	2	0,01	0.947821
17	6	1500	3	0,1	0.957643
18	6	1500	3	0,01	0.947821
19	8	500	1	0,1	0.955801
20	8	500	1	0,01	0.945979
21	8	500	2	0,1	0.955187
22	8	500	2	0,01	0.944138
23	8	500	3	0,1	0.953346
24	8	500	3	0,01	0.944138
25	8	1000	1	0,1	0.956415
26	8	1000	1	0,01	0.949048
27	8	1000	2	0,1	0.955801
28	8	1000	2	0,01	0.946593
29	8	1000	3	0,1	0.954573
30	8	1000	3	0,01	0.945365
31	8	1500	1	0,1	0.957029
32	8	1500	1	0,01	0.950276
33	8	1500	2	0,1	0.955187
34	8	1500	2	0,01	0.949662
35	8	1500	3	0,1	0.954573
36	8	1500	3	0,01	0.948435

No	params				*mean_test_score (cosine similarity)*
	depth	*iterations*	*l2_leaf_reg*	*learning_rate*	
37	10	500	1	0,1	0.956415
38	10	500	1	0,01	0.945365
39	10	500	2	0,1	0.955187
40	10	500	2	0,01	0.945365
41	10	500	3	0,1	0.956415
42	10	500	3	0,01	0.944138
43	10	1000	1	0,1	0.957643
44	10	1000	1	0,01	0.949048
45	10	1000	2	0,1	0.957029
46	10	1000	2	0,01	0.947821
47	10	1000	3	0,1	0.955801
48	10	1000	3	0,01	0.945365
49	10	1500	1	0,1	0.958257
50	10	1500	1	0,01	0.950890
51	10	1500	2	0,1	0.957643
52	10	1500	2	0,01	0.950890
53	10	1500	3	0,1	0.956415
54	10	1500	3	0,01	0.949662

Pada tabel 4.1 menunjukan bahwa nilai parameter terbaik pada model 1 untuk setiap kombinasinya adalah pada *depth* 6, *iterations* 1500, *l2_leaf_reg* 1 dan *learning_rate* 0,1 dengan hasil *mean_test_score cosine similarity* sebesar 0.958870. Sehingga pada pengujian model 1 yang dilakukan telah mendapatkan kombinasi parameter yang optimal pada kombinasi tersebut.

Fokus pada parameter *depth* yang pertama adalah nilai 6 dengan *iterations* 500 *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0.957643. Nilai parameter *depth* kedua adalah 8 dengan *iterations* 500 *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0.955801. Dan nilai parameter *depth* ketiga adalah 10 dengan *iterations* 500 *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,956415. Terlihat bahwa bertambahnya nilai *depth* dari 6, 8, dan 10 skor paling baik adalah pada *depth* 6. Kemungkinan besar jika nilai *depth* sedikit nilai *mean_test_score cosine similarity* dapat maksimal dan jika berlebihan juga akan

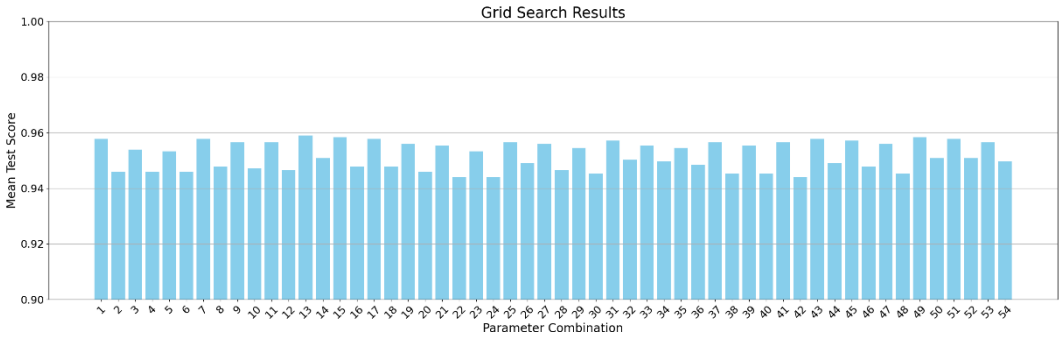
menyebabkan penurunan skor sesuai dengan data yang digunakan. Sehingga nilai parameter *depth* akan membantu meningkatkan *mean_test_score cosine similarity* jika pada nilai yang tepat tidak kurang dan tidak lebih.

Fokus pada parameter *iterations* yang pertama adalah 500 dengan nilai *depth* 6, *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,957643. Nilai parameter *iterations* kedua adalah 1000 dengan nilai *depth* 6, *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,957643. Dan nilai parameter ketiga yaitu 1500 dengan nilai *depth* 6, *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,958870. Terlihat bahwa bertambahnya nilai parameter *iterations* nilai *mean_test_score cosine similarity* meningkat akan tetapi pada nilai 500 dan 1000 nilainya sama. Sehingga nilai parameter *iterations* akan membantu meningkatkan nilai *mean_test_score cosine similarity* jika pada nilai yang banyak dan tepat.

Fokus pada parameter *l2_leaf_reg* yang pertama adalah 1 dengan nilai *depth* 6, *iterations* 500 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,957643. Nilai parameter *l2_leaf_reg* kedua adalah 2 nilai *depth* 6, *iterations* 500 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,953959. Dan nilai parameter *l2_leaf_reg* yang ketiga adalah 3 dengan nilai *depth* 6, *iterations* 500 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,953346. Terlihat bahwa bertambahnya nilai parameter *l2_leaf_reg* nilai *mean_test_score cosine similarity* juga semakin menurun.

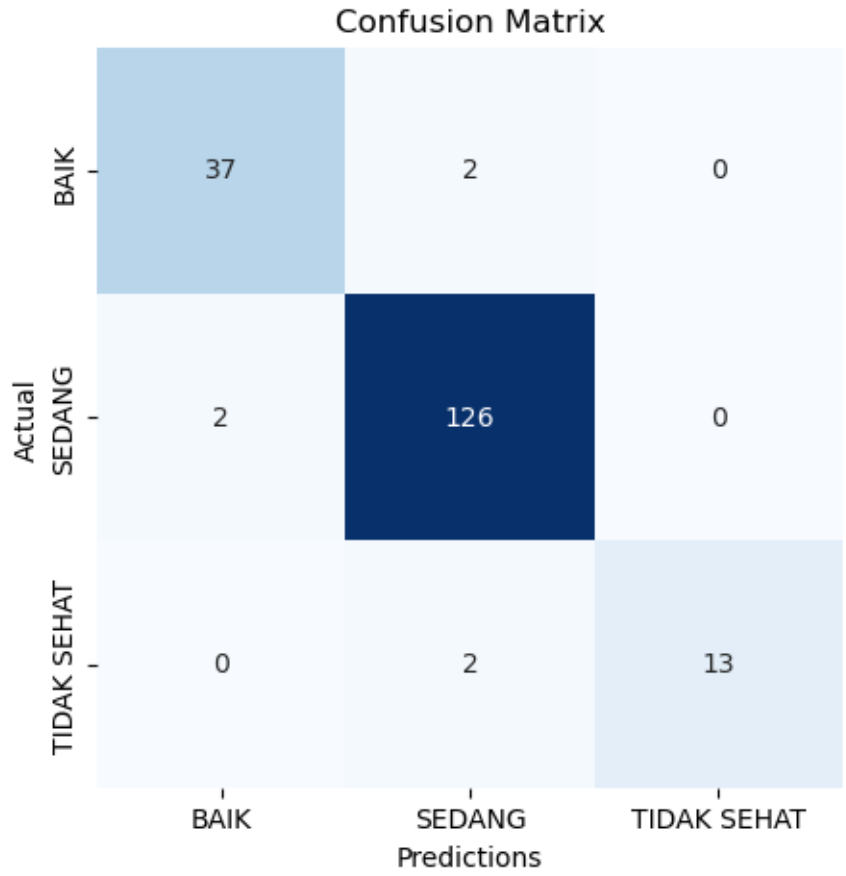
Sehingga parameter *l2_leaf_reg* membantu meningkatkan nilai *mean_test_score cosine similarity* jika pada nilai yang tepat dan tidak lebih atau kurang.

Fokus pada parameter *learning_rate* yang pertama adalah nilai 0,1 dengan nilai *depth* 6, iterations 500 dan *l2_leafIreg* 1 skor nilai uji *mean_test_score cosine similarity* adalah 0,957643 sedangkan pada nilai parameter *learning_rate* kedua yaitu 0,01 dengan nilai *depth* 6, iterations 500 dan *l2_leafIreg* 1 skor nilai uji *mean_test_score cosine similarity* adalah 0,945979. Terlihat bahwa parameter *learning_rate* semakin bertambah maka semakin baik nilai skor uji *mean_test_score cosine similarity*. Sehigga parameter *learning_rate* dapat membantu meningkatkan skor uji *mean_test_score cosine similarity* jika nilainya semakin tinggi. Untuk lebih jelasnya berikut adalah gambar kombinasi parameter pengujian pada model 1.



Gambar 4. 1 Rata-rata *Cosine Similarity* Model 1

Parameter optimal yang telah ditemukan akan digunakan untuk pengujian pada data *training* dan *testing*. Sehingga pengujian pada data *training* menghasilkan skor akurasi 1.0 atau sempurna dan pada data *testing* menghasilkan skor akurasi 97%. Berikut adalah gambar confusion matrix pengujian model 1.



Gambar 4. 2 Confusion Matrix Model 1

Dari gambar 4.2 menunjukan bahwa model 1 memprediksi 37 data kualitas udara baik dan pada hasil yang sebenarnya terdeteksi baik, memprediksi 2 data kualitas udara sedang akan tetapi pada data sebenarnya adalah baik, memprediksi 126 data kualitas udara sedang dan pada data yang sebenarnya adalah sedang, memprediksi 2 data kualitas udara baik akan tetapi pada data sebenarnya adalah sedang, memprediksi 13 data kualitas udara tidak sehat dan pada data yang sebenarnya adalah tidak sehat, memprediksi 2 data kualitas udara sedang akan tetaapi pada data yang sebenarnya adalah tidak sehat.

Dengan melihat gambar 4.2 dapat diketahui nilai dari *accuracy*, *precision*, *recall,* dan *f1-score* dari masing-masing kelas dengan menggunakan teknik evaluasi. Mengacu kepada gambar 4.2 untuk kelas kategori BAIK nilai TP adalah 37, FN adalah 2, FP adalah 2 dan TN adalah 141. Perhitungan dari nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{37 + 126 + 13}{182}\ x\ 100\% = 967 \text{ dibulatkan } 97\%$$

$$Precission \text{ kategori BAIK} = \frac{37}{37 + 2}\ x\ 100\% = 948 \text{ dibulatkan } 95\%$$

$$Recall \text{ kategori BAIK} = \frac{37}{37 + 2}\ x\ 100\% = 948 \text{ dibulatkan } 95\%$$

$$F1\ score \text{ kategori BAIK} = 2\ \frac{0{,}948\ x\ 0{,}948}{0{,}948 + 0{,}948}\ x\ 100\% = 948 \text{ dibulatkan } 95\%$$

Mengacu kepada gambar 4.2 untuk kelas kategori SEDANG nilai TP adalah 126, nilai FN adalah 2, nilai FP adalah 4, nilai TN adalah 50. Perhitungan dari nilai tersebut adalah sebagai berikut.

$$Precission \text{ kategori SEDANG} = \frac{126}{126 + 4}\ x\ 100\% = 969 \text{ dibulatkan } 97\%$$

$$Recall \text{ kategori SEDANG} = \frac{126}{126 + 2}\ x\ 100\% = 984 \text{ dibulatkan } 98\%$$

$$F1\ score \text{ kategori SEDANG} = 2\ \frac{0{,}984\ x\ 0{,}969}{0{,}984 + 0{,}969}\ x\ 100\%$$

$$= 976\ dibulatkan\ 98\%$$

Mengacu kepada gambar 4.2 untuk kelas kategori TIDAK SEHAT nilai TP adalah 13, nilai FN adalah 2, nilai FP adalah 0, nilai TN adalah 167. Perhitungan dari nilai tersebut adalah sebagai berikut.

$Precission$ kategori TIDAK SEHAT $= \frac{13}{13+0}\ x\ 100\% = 100\%$

$Recall$ kategori TIDAK SEHAT $= \frac{13}{13+2}\ x\ 100\% = 866$ dibulatkan 87%

$F1\ score$ kategori TIDAK SEHAT $= 2\ \frac{1\ x\ 0{,}866}{1+0{,}866}\ x\ 100\%$

$= 928$ dibulatkan 93%

4.1.2 Pengujian Model 2

Pengujian menggunakan data *training* sebanyak 1448 dan data *testing* sebanyak 363. Pengujian yang dilakukan mendapatkan kombinasi parameter terbaik yang dihasilkan dengan menggunakan *GrideSearchCV*. Berikut adalah nilai hasil pengujian parameter untuk kombinasi parameter yang terbaik menggunakan *GridSearchCV*.

Tabel 4. 2 Hasil Rata-rata Skor Uji *Cosine Similarity* Model 2

No	**params**				***mean_test_score (cosine similarity)***
	depth	***iterations***	***l2_leaf_reg***	***learning_rate***	
1	6	500	1	0,1	0.957882
2	6	500	1	0,01	0.949595
3	6	500	2	0,1	0.957882
4	6	500	2	0,01	0.947521
5	6	500	3	0,1	0.958572
6	6	500	3	0,01	0.947521
7	6	1000	1	0,1	0.958574
8	6	1000	1	0,01	0.950973
9	6	1000	2	0,1	0.958574
10	6	1000	2	0,01	0.951664
11	6	1000	3	0,1	0.957882
12	6	1000	3	0,01	0.950283
13	6	1500	1	0,1	0.959264
14	6	1500	1	0,01	0.953735
15	6	1500	2	0,1	0.959264
16	6	1500	2	0,01	0.950285
17	6	1500	3	0,1	0.958572
18	6	1500	3	0,01	0.950283
19	8	500	1	0,1	0.959262
20	8	500	1	0,01	0.949593
21	8	500	2	0,1	0.957882
22	8	500	2	0,01	0.948903

No	params				*mean_test_score (cosine similarity)*
	depth	*iterations*	*l2_leaf_reg*	*learning_rate*	
23	8	500	3	0,1	0.958572
24	8	500	3	0,01	0.947521
25	8	1000	1	0,1	0.957192
26	8	1000	1	0,01	0.950973
27	8	1000	2	0,1	0.957882
28	8	1000	2	0,01	0.950973
29	8	1000	3	0,1	0.958572
30	8	1000	3	0,01	0.950283
31	8	1500	1	0,1	0.957882
32	8	1500	1	0,01	0.956499
33	8	1500	2	0,1	0.957882
34	8	1500	2	0,01	0.954426
35	8	1500	3	0,1	0.958572
36	8	1500	3	0,01	0.952355
37	10	500	1	0,1	0.958572
38	10	500	1	0,01	0.950283
39	10	500	2	0,1	0.959262
40	10	500	2	0,01	0.948212
41	10	500	3	0,1	0.957882
42	10	500	3	0,01	0.948903
43	10	1000	1	0,1	0.958572
44	10	1000	1	0,01	0.954426
45	10	1000	2	0,1	0.957882
46	10	1000	2	0,01	0.950973
47	10	1000	3	0,1	0.957882
48	10	1000	3	0,01	0.950973
49	10	1500	1	0,1	0.958572
50	10	1500	1	0,01	0.957881
51	10	1500	2	0,1	0.957882
52	10	1500	2	0,01	0.953735
53	10	1500	3	0,1	0.957192
54	10	1500	3	0,01	0.953735

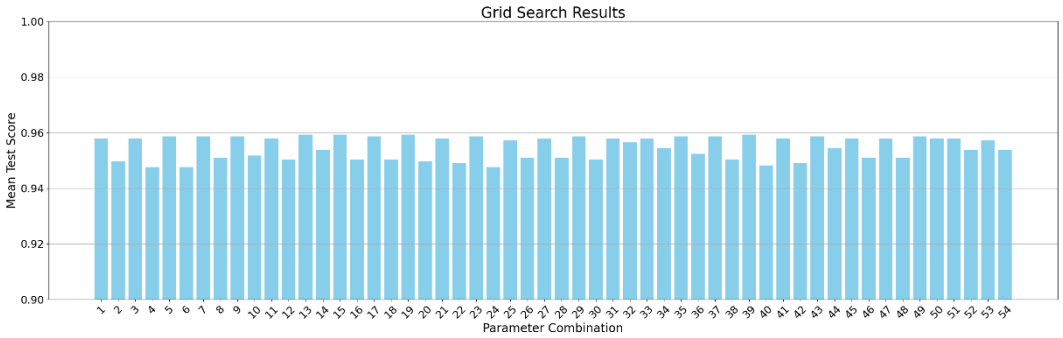
Pada tabel 4.2 menunjukan bahwa nilai parameter terbaik pada model 2 untuk setiap kombinasinya terdapat dua nilai tertinggi yang sama yaitu pada *depth* 6, *iterations* 1500, *l2_leaf_reg* 1 dan *learning_rate* 0,1 dan pada dept 6, *iterations* 1500, l2_leaaf_reg 2 dan *learning_rate* 0,1 dengan hasil *mean_test_score cosine similarity* sebesar 0.959264. Sehingga pada pengujian model 2 yang dilakukan telah mendapatkan kombinasi parameter yang optimal pada kombinasi tersebut.

Fokus pada parameter *depth* yang pertama adalah nilai 6 dengan *iterations* 500 *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0.957882. Nilai parameter *depth* kedua adalah 8 dengan *iterations* 500 *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0.959262. Dan nilai parameter *depth* ketiga adalah 10 dengan *iterations* 500 *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,958572. Terlihat bahwa bertambahnya nilai *depth* dari 6, 8, dan 10 nilai paling baik adalah pada *depth* 8. Kemungkinan besar jika nilai *depth* sedikit nilai *mean_test_score cosine similarity* kurang maksimal dan jika berlebihan juga akan menyebabkan penurunan skor. Sehingga nilai parameter *depth* akan membantu meningkatkan nilai *mean_test_score cosine similarity* jika pada nilai parameter yang tepat tidak kurang dan tidak lebih.

Fokus pada parameter *iterations* yang pertama adalah 500 dengan nilai *depth* 6, *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,957882. Nilai parameter *iterations* kedua adalah 1000 dengan nilai *depth* 6, *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,958574. Dan nilai parameter ketiga yaitu 1500 dengan nilai *depth* 6, *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,959264. Terlihat bahwa bertambahnya nilai parameter *iterations* nilai *mean_test_score cosine similarity* semakin meningkat. Sehingga nilai parameter *iterations* akan membantu meningkatkan nilai *mean_test_score cosine similarity* jika pada nilai parameter yang banyak.

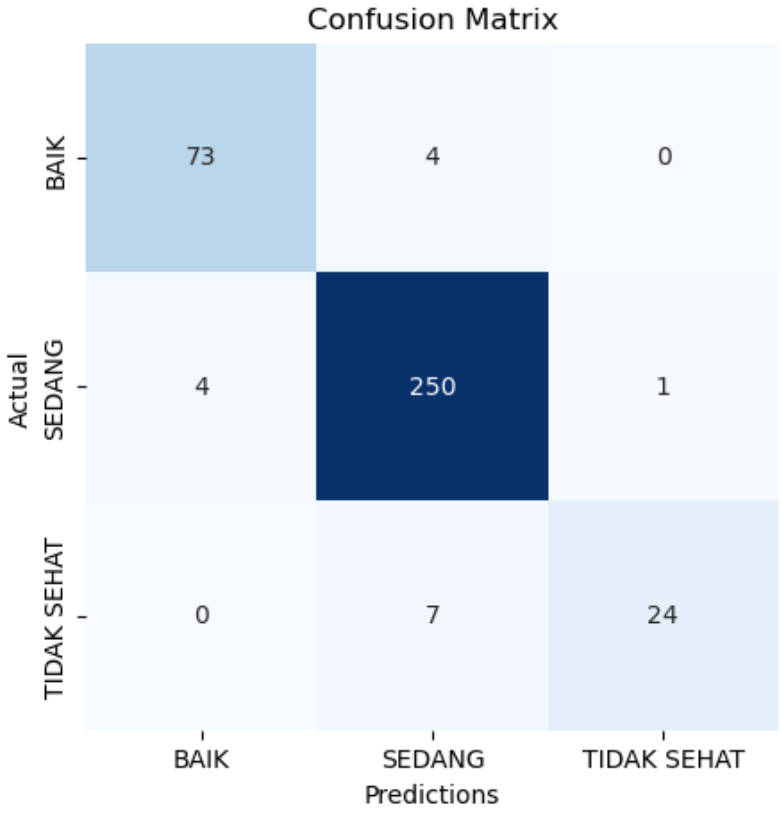
Fokus pada parameter *l2_leaf_reg* yang pertama adalah 1 dengan nilai *depth* 6, *iterations* 500 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,957882. Nilai parameter *l2_leaf_reg* kedua adalah 2 nilai *depth* 6, *iterations* 500 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,957882. Dan nilai parameter *l2_leaf_reg* yang ketiga adalah 3 dengan nilai *depth* 6, *iterations* 500 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,958572. Terlihat bahwa bertambahnya nilai parameter *l2_leaf_reg* nilai *mean_test_score cosine similarity* juga bertambah akan tetapi pada nilai 1 dan 2 bernilai sama. Sehingga parameter *l2_leaf_reg* membantu meningkatkan *mean_test_score cosine similarity* jika pada nilai parameter yang tepat dan banyak.

Fokus pada parameter *learning_rate* yang pertama adalah nilai 0,1 dengan nilai *depth* 6, iterations 500 dan *l2_leafIreg* 1 skor nilai uji *mean_test_score cosine similarity* adalah 0,957882 sedangkan pada nilai parameter *learning_rate* kedua yaitu 0,01 dengan nilai *depth* 6, iterations 500 dan *l2_leafIreg* 1 skor nilai uji *mean_test_score cosine similarity* adalah 0,949595. Terlihat bahwa parameter *learning_rate* semakin bertambah maka semakin baik nilai skor uji *mean_test_score cosine similarity*. Sehigga parameter *learning_rate* dapat membantu meningkatkan *mean_test_score cosine similarity* jika nilainya semakin tinggi. Untuk lebih jelasnya berikut adalah gambar kombinasi parameter pengujian pada model 2.



Gambar 4. 3 Rata-rata *Cosine Similarity* Model 2

Parameter optimal yang telah ditemukan akan digunakan untuk pengujian pada data *training* dan *testing*. Sehingga pengujian pada data *training* menghasilkan skor akurasi 1.0 atau sempurna dan pada data *testing* menghasilkan skor akurasi 96%. Berikut adalah gambar confusion matrix pengujian model 2.



Gambar 4. 4 Confusion Matrix Model 2

Dari gambar 4.4 menunjukan bahwa model 2 memprediksi 73 data kualitas udara baik dan pada hasil yang sebenarnya terdeteksi baik, memprediksi 4 data kualitas udara sedang akan tetapi pada data sebenarnya adalah baik, memprediksi 250 data kualitas udara sedang dan pada data yang sebenarnya adalah sedang, memprediksi 4 data kualitas udara baik akan tetapi pada data sebenarnya adalah sedang, memprediksi 24 data kualitas udara tidak sehat dan pada data yang sebenarnya adalah tidak sehat, memprediksi 7 data kualitas udara sedang akan tetapi pada data yang sebenarnya adalah tidak sehat dan memprediksi 1 data tidak sehat akan tetapi pada data yang sebenarnya adalah sedang.

Dengan melihat gambar 4.4 dapat diketahui nilai dari *accuracy*, *precision*, *recall*, dan *f1-score* dari masing-masing kelas dengan menggunakan teknik evaluasi. Mengacu kepada gambar 4.4 untuk kelas kategori BAIK nilai TP adalah 73, FN adalah 4, FP adalah 4 dan TN adalah 282. Perhitungan dari nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{73 + 250 + 54}{363}\ x\ 100\% = 9559 \text{ dibulatkan } 96\%$$

$$Precission \text{ kategori BAIK} = \frac{73}{73 + 4}\ x\ 100\% = 948 \text{ dibulatkan } 95\%$$

$$Recall \text{ kategori BAIK} = \frac{73}{73 + 4}\ x\ 100\% = 948 \text{ dibulatkan } 95\%$$

$$F1\ score \text{ kategori BAIK} = 2\ \frac{0{,}948\ x\ 0{,}948}{0{,}948 + 0{,}948}\ x\ 100\% = 948 \text{ dibulatkan } 95\%$$

Mengacu kepada gambar 4.4 untuk kelas kategori SEDANG nilai TP adalah 250, nilai FN adalah 5, nilai FP adalah 11, nilai TN adalah 97. Perhitungan dari nilai tersebut adalah sebagai berikut.

$Precission$ kategori SEDANG $= \frac{250}{250+11}\ x\ 100\% = 957$ dibulatkan 96%

$Recall$ kategori SEDANG $= \frac{250}{250+5}\ x\ 100\% = 980$ dibulatkan 98%

$F1\ score$kategori SEDANG $= 2\ \frac{0{,}957\ x\ 0{,}980}{0{,}957+0{,}980}\ x\ 100\%$

$= 968$ dibulatkan 97%

Mengacu kepada gambar 4.4 untuk kelas kategori TIDAK SEHAT nilai TP adalah 24, nilai FN adalah 7, nilai FP adalah 1, nilai TN adalah 331. Perhitungan dari nilai tersebut adalah sebagai berikut.

$Precission$ kategori TIDAK SEHAT $= \frac{24}{24+1}\ x\ 100\% = 96\%$

$Recall$ kategori TIDAK SEHAT $= \frac{24}{24+7}\ x\ 100\% = 774$ dibulatkan 77%

$F1\ score$ kategori TIDAK SEHAT $= 2\ \frac{0{,}96\ x\ 0{,}774}{0{,}96+0{,}774}\ x\ 100\%$

$= 857$ dibulatkan 86%

4.1.3 Pengujian Model 3

Pengujian menggunakan data *training* sebanyak 1358 dan data *testing* sebanyak 453. Pengujian yang dilakukan mendapatkan kombinasi parameter terbaik yang dihasilkan dengan menggunakan *GrideSearchCV*. Berikut adalah nilai hasil pengujian parameter untuk kombinasi parameter yang terbaik menggunakan *GridSearchCV*.

Tabel 4. 3 Hasil Rata-rata Skor Uji *Cosine Similarity* Model 3

No	**params**				***mean_test_score (cosine similarity)***
	depth	***iterations***	***l2_leaf_reg***	***learning_rate***	
1	6	500	1	0,1	0.956555
2	6	500	1	0,01	0.951402
3	6	500	2	0,1	0.955819
4	6	500	2	0,01	0.949192
5	6	500	3	0,1	0.956555
6	6	500	3	0,01	0.949929
7	6	1000	1	0,1	0.956555
8	6	1000	1	0,01	0.952141
9	6	1000	2	0,1	0.956555
10	6	1000	2	0,01	0.949931
11	6	1000	3	0,1	0.956555
12	6	1000	3	0,01	0.949195
13	6	1500	1	0,1	0.956555
14	6	1500	1	0,01	0.955083
15	6	1500	2	0,1	0.956555
16	6	1500	2	0,01	0.952141
17	6	1500	3	0,1	0.956555
18	6	1500	3	0,01	0.952877
19	8	500	1	0,1	0.958026
20	8	500	1	0,01	0.949929
21	8	500	2	0,1	0.958028
22	8	500	2	0,01	0.950666
23	8	500	3	0,1	0.958026
24	8	500	3	0,01	0.949929
25	8	1000	1	0,1	0.955817
26	8	1000	1	0,01	0.952141
27	8	1000	2	0,1	0.956555
28	8	1000	2	0,01	0.952877
29	8	1000	3	0,1	0.956555
30	8	1000	3	0,01	0.949193
31	8	1500	1	0,1	0.955817
32	8	1500	1	0,01	0.956555
33	8	1500	2	0,1	0.956555
34	8	1500	2	0,01	0.955085
35	8	1500	3	0,1	0.956555
36	8	1500	3	0,01	0.953613
37	10	500	1	0,1	0.956555
38	10	500	1	0,01	0.950666
39	10	500	2	0,1	0.958028
40	10	500	2	0,01	0.950666
41	10	500	3	0,1	0.957291
42	10	500	3	0,01	0.950666
43	10	1000	1	0,1	0.956555
44	10	1000	1	0,01	0.952877
45	10	1000	2	0,1	0.957291
46	10	1000	2	0,01	0.952879
47	10	1000	3	0,1	0.956555
48	10	1000	3	0,01	0.949932
49	10	1500	1	0,1	0.956555
50	10	1500	1	0,01	0.958028

No	params				*mean_test_score (cosine similarity)*
	depth	*iterations*	*l2_leaf_reg*	*learning_rate*	
51	10	1500	2	0,1	0.956555
52	10	1500	2	0,01	0.956556
53	10	1500	3	0,1	0.956555
54	10	1500	3	0,01	0.953613

Pada tabel 4.3 menunjukan bahwa nilai parameter terbaik pada model 3 untuk setiap kombinasinya terdapat dua nilai tertinggi yang sama yaitu pada *depth* 8, *iterations* 500, *l2_leaf_reg* 2 dan *learning_rate* 0,1 dan pada *depth* 10, *iterations* 500, l2_leaaf_reg 2 dan *learning_rate* 0,1 dengan hasil *mean_test_score cosine similarity* sebesar 0.958028. Sehingga pada pengujian model 3 yang dilakukan telah mendapatkan kombinasi parameter yang optimal pada kombinasi tersebut.

Fokus pada parameter *depth* yang pertama adalah nilai 6 dengan *iterations* 500 *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0.956555. Nilai parameter *depth* kedua adalah 8 dengan *iterations* 500 *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0.958026. Dan nilai parameter *depth* ketiga adalah 10 dengan *iterations* 500 *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,956555. Terlihat bahwa bertambahnya nilai *depth* dari 6, 8, dan 10 skor paling baik adalah pada *depth* 8. Kemungkinan besar jika nilai *depth* sedikit nilai *mean_test_score cosine similarity* kurang maksimal dan jika berlebihan juga akan menyebabkan penurunan. Sehingga nilai parameter *depth* akan membantu meningkatkan nilai *mean_test_score cosine similarity* jika pada nilai parameter yang tepat tidak kurang dan tidak lebih.

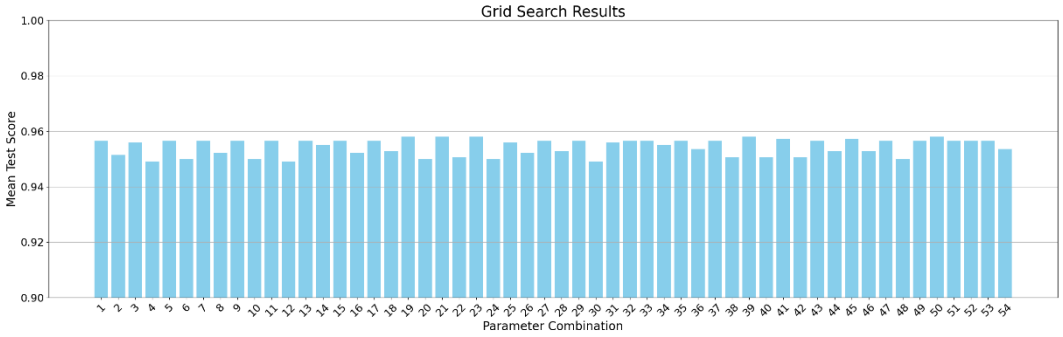
Fokus pada parameter *iterations* yang pertama adalah 500 dengan nilai *depth* 6, *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine*

similarity adalah 0,956555. Nilai parameter *iterations* kedua adalah 1000 dengan nilai *depth* 6, *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,956555. Dan nilai parameter ketiga yaitu 1500 dengan nilai *depth* 6, *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,956555. Terlihat bahwa bertambahnya nilai parameter *iterations* nilai *mean_test_score cosine similarity* tidak ada perubahan karena ketiga skor hasilnya adalah sama. Sehingga nilai parameter *iterations* akan membantu meningkatkan nilai *mean_test_score cosine similarity* jika pada nilai parameter yang tepat dan pada kombinasi yang cocok.

Fokus pada parameter *l2_leaf_reg* yang pertama adalah 1 dengan nilai *depth* 6, *iterations* 500 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,956555. Nilai parameter *l2_leaf_reg* kedua adalah 2 nilai *depth* 6, *iterations* 500 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,955819. Dan nilai parameter *l2_leaf_reg* yang ketiga adalah 3 dengan nilai *depth* 6, *iterations* 500 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,956555. Terlihat bahwa bertambahnya nilai parameter *l2_leaf_reg* nilai *mean_test_score cosine similarity* menurun akan tetapi pada nilai 1 dan dan 3 bernilai sama. Sehingga parameter *l2_leaf_reg* membantu meningkatkan *mean_test_score cosine similarity* jika pada nilai parameter yang tepat dan banyak.

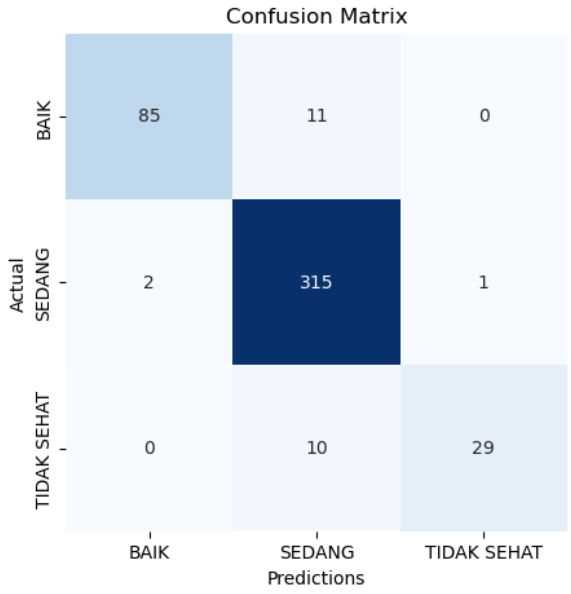
Fokus pada parameter *learning_rate* yang pertama adalah nilai 0,1 dengan nilai *depth* 6, iterations 500 dan *l2_leafIreg* 1 skor nilai uji *mean_test_score cosine similarity* adalah 0,956555 sedangkan pada nilai parameter *learning_rate* kedua

yaitu 0,01 dengan nilai *depth* 6, iterations 500 dan *l2_leafIreg* 1 skor nilai uji *mean_test_score cosine similarity* adalah 0,951402. Terlihat bahwa parameter *learning_rate* semakin bertambah maka semakin baik nilai *mean_test_score cosine similarity*. Sehigga parameter *learning_rate* dapat membantu meningkatkan *mean_test_score cosine similarity* jika nilai parameter semakin tinggi. Untuk lebih jelasnya berikut adalah gambar kombinasi parameter pengujian pada model 3.



Gambar 4. 5 Rata-rata *Cosine Similarity* Model 3

Parameter optimal yang telah ditemukan akan digunakan untuk pengujian pada data *training* dan *testing*. Sehingga pengujian pada data *training* menghasilkan skor akurasi 1.0 atau sempurna dan pada data *testing* menghasilkan skor akurasi 95%. Berikut adalah gambar confusion matrix pengujian model 3.



Gambar 4. 6 Confusion Matrix Model 3

Dari gambar 4.6 menunjukan bahwa model 3 memprediksi 85 data kualitas udara baik dan pada hasil yang sebenarnya terdeteksi baik, memprediksi 11 data kualitas udara sedang akan tetapi pada data sebenarnya adalah baik, memprediksi 315 data kualitas udara sedang dan pada data yang sebenarnya adalah sedang, memprediksi 2 data kualitas udara baik akan tetapi pada data sebenarnya adalah sedang, memprediksi 29 data kualitas udara tidak sehat dan pada data yang sebenarnya adalah tidak sehat, memprediksi 10 data kualitas udara sedang akan tetapi pada data yang sebenarnya adalah tidak sehat dan memprediksi 1 data tidak sehat akan tetapi pada data yang sebenarnya adalah sedang.

Dengan melihat gambar 4.6 dapat diketahui nilai dari *accuracy*, *precision*, *recall*, dan *f1-score* dari masing-masing kelas dengan menggunakan teknik evaluasi. Mengacu kepada gambar 4.6 untuk kelas kategori BAIK nilai TP adalah

85, FN adalah 11, FP adalah 2 dan TN adalah 355. Perhitungan dari nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{85 + 315 + 29}{453}\ x\ 100\% = 947 \text{ dibulatkan } 95\%$$

$$Precission \text{ kategori BAIK} = \frac{85}{85 + 2}\ x\ 100\% \ = 977 \text{ dibulatkan } 98\%$$

$$Recall \text{ kategori BAIK} = \frac{85}{85 + 11}\ x\ 100\% = 885 \text{ dibulatkan } 89\%$$

$$F1\ score \text{ kategori BAIK} = 2\ \frac{0{,}977\ x\ 0{,}885}{0{,}977 + 0{,}885}\ x\ 100\% = 928 \text{ dibulatkan } 93\%$$

Mengacu kepada gambar 4.6 untuk kelas kategori SEDANG nilai TP adalah 315, nilai FN adalah 3, nilai FP adalah 21, nilai TN adalah 114. Perhitungan dari nilai tersebut adalah sebagai berikut.

$$Precission \text{ kategori SEDANG} = \frac{315}{315 + 21}\ x\ 100\% = 937 \text{ dibulatkan } 94\%$$

$$Recall \text{ kategori SEDANG} = \frac{315}{315 + 3}\ x\ 100\% = 99\%$$

$$F1\ score \text{ kategori SEDANG} = 2\ \frac{0{,}937\ x\ 0{,}99}{0{,}937 + 0{,}99}\ x\ 100\%$$

$$= 0{,}962 \text{ dibulatkan } 96\%$$

Mengacu kepada gambar 4.2 untuk kelas kategori TIDAK SEHAT nilai TP adalah 29, nilai FN adalah 10, nilai FP adalah 1, nilai TN adalah 413. Perhitungan dari nilai tersebut adalah sebagai berikut.

$$Precission \text{ kategori TIDAK SEHAT} = \frac{29}{29 + 1}\ x\ 100\% = 966 \text{ dibulatkan } 97\%$$

$$Recall \text{ kategori TIDAK SEHAT} = \frac{29}{29 + 10}\ x\ 100\% = 743 \text{ dibulatkan } 74\%$$

$$F1\ score \text{ kategori TIDAK SEHAT} = 2\ \frac{0{,}96\ x\ 0{,}743}{0{,}96 + 0{,}743}\ x\ 100\%$$

$$= 839 \text{ dibulatkan } 84\%$$

4.1.4 Pengujian Model 4

Pengujian menggunakan data *training* sebanyak 1267 dan data *testing* sebanyak 544. Pengujian yang dilakukan mendapatkan kombinasi parameter terbaik yang dihasilkan dengan menggunakan *GrideSearchCV*. Berikut adalah nilai hasil pengujian parameter untuk kombinasi parameter yang terbaik menggunakan *GridSearchCV*.

Tabel 4. 4 Hasil Rata-rata Skor Uji *Cosine Similarity* Model 4

No	**params**				***mean_test_score (cosine similarity)***
	depth	***iterations***	***l2_leaf_reg***	***learning_rate***	
1	6	500	1	0,1	0.963687
2	6	500	1	0,01	0.951850
3	6	500	2	0,1	0.960532
4	6	500	2	0,01	0.950274
5	6	500	3	0,1	0.959742
6	6	500	3	0,01	0.948695
7	6	1000	1	0,1	0.963687
8	6	1000	1	0,01	0.955796
9	6	1000	2	0,1	0.963687
10	6	1000	2	0,01	0.954220
11	6	1000	3	0,1	0.962111
12	6	1000	3	0,01	0.952640
13	6	1500	1	0,1	0.963687
14	6	1500	1	0,01	0.958162
15	6	1500	2	0,1	0.963687
16	6	1500	2	0,01	0.955798
17	6	1500	3	0,1	0.963687
18	6	1500	3	0,01	0.955796
19	8	500	1	0,1	0.962108
20	8	500	1	0,01	0.951850
21	8	500	2	0,1	0.962898
22	8	500	2	0,01	0.951062
23	8	500	3	0,1	0.960532
24	8	500	3	0,01	0.951062
25	8	1000	1	0,1	0.962898
26	8	1000	1	0,01	0.958162
27	8	1000	2	0,1	0.964476
28	8	1000	2	0,01	0.955008
29	8	1000	3	0,1	0.962898

No	params				*mean_test_score (cosine similarity)*
	depth	*iterations*	*l2_leaf_reg*	*learning_rate*	
30	8	1000	3	0,01	0.953428
31	8	1500	1	0,1	0.958948
32	8	1500	1	0,01	0.959742
33	8	1500	2	0,1	0.963686
34	8	1500	2	0,01	0.958162
35	8	1500	3	0,1	0.964476
36	8	1500	3	0,01	0.956584
37	10	500	1	0,1	0.962898
38	10	500	1	0,01	0.951062
39	10	500	2	0,1	0.962898
40	10	500	2	0,01	0.951850
41	10	500	3	0,1	0.962108
42	10	500	3	0,01	0.950274
43	10	1000	1	0,1	0.962108
44	10	1000	1	0,01	0.955796
45	10	1000	2	0,1	0.964477
46	10	1000	2	0,01	0.955008
47	10	1000	3	0,1	0.962108
48	10	1000	3	0,01	0.953430
49	10	1500	1	0,1	0.961318
50	10	1500	1	0,01	0.958954
51	10	1500	2	0,1	0.964477
52	10	1500	2	0,01	0.957374
53	10	1500	3	0,1	0.962110
54	10	1500	3	0,01	0.956584

Pada tabel 4.4 menunjukan bahwa nilai parameter terbaik pada model 4 untuk setiap kombinasinya terdapat dua nilai tertinggi yang sama yaitu pada *depth* 10, *iterations* 1000, *l2_leaf_reg* 2 dan *learning_rate* 0,1 dan pada dept 10, *iterations* 1500, l2_leaaf_reg 2 dan *learning_rate* 0,1 dengan hasil *mean_test_score cosine similarity* sebesar 0.964477. Sehingga pada pengujian model 4 yang dilakukan telah mendapatkan kombinasi parameter yang optimal pada kombinasi tersebut.

Fokus pada parameter *depth* yang pertama adalah nilai 6 dengan *iterations* 500 *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0.963687. Nilai parameter *depth* kedua adalah 8 dengan *iterations* 500 *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0.962108. Dan nilai parameter *depth* ketiga adalah 10 dengan *iterations* 500

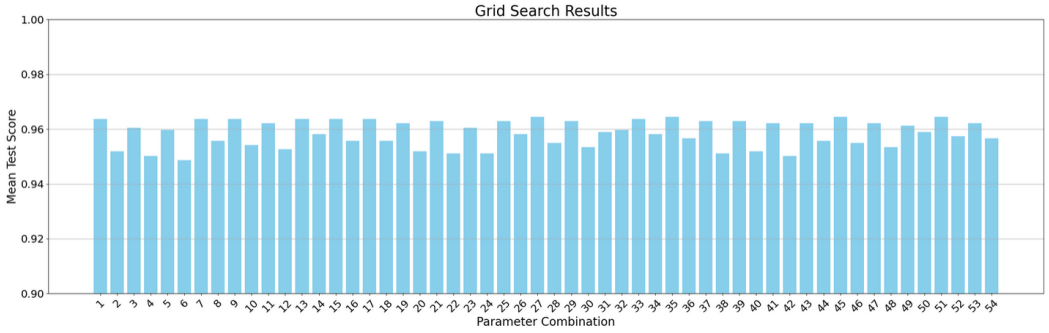
l2_leaf_reg 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,962898. Terlihat bahwa bertambahnya nilai *depth* dari 6, 8, dan 10 nilai paling baik adalah pada *depth* 6. Kemungkinan besar jika nilai *depth* sedikit skor uji *mean_test_score cosine similarity* akan maksimal dan jika berlebihan juga akan menyebabkan penurunan skor. Sehingga nilai parameter *depth* akan membantu meningkatkan *mean_test_score cosine similarity* jika pada nilai parameter yang tepat tidak kurang dan tidak lebih.

Fokus pada parameter *iterations* yang pertama adalah 500 dengan nilai *depth* 6, *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,963687. Nilai parameter *iterations* kedua adalah 1000 dengan nilai *depth* 6, *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,963687. Dan nilai parameter ketiga yaitu 1500 dengan nilai *depth* 6, *l2_leaf_reg* 1 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,963687. Terlihat bahwa bertambahnya nilai parameter *iterations* skor uji *mean_test_score cosine similarity* tetap sama. Sehingga nilai parameter *iterations* akan membantu meningkatkan *mean_test_score cosine similarity* jika pada nilai parameter yang tepat dan pada kombinasi yang cocok.

Fokus pada parameter *l2_leaf_reg* yang pertama adalah 1 dengan nilai *depth* 6, *iterations* 500 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,963687. Nilai parameter *l2_leaf_reg* kedua adalah 2 nilai *depth* 6, *iterations* 500 dan *learning_rate* 0,1 skor uji *mean_test_score cosine similarity* adalah 0,960532. Dan nilai parameter *l2_leaf_reg* yang ketiga adalah 3 dengan nilai *depth* 6, *iterations* 500 dan *learning_rate* 0,1 skor uji *mean_test_score cosine*

similarity adalah 0,959742. Terlihat bahwa bertambahnya nilai parameter *l2_leaf_reg* skor uji *mean_test_score cosine similarity* mengalami penurunan. Sehingga parameter *l2_leaf_reg* membantu meningkatkan *mean_test_score cosine similarity* jika pada nilai parameter yang tepat dan pada kombinasi yang cocok.

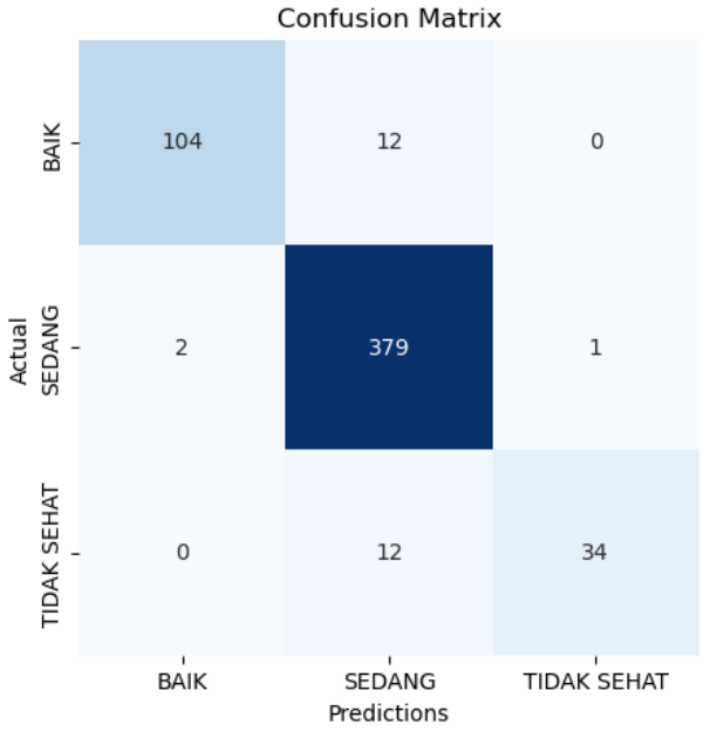
Fokus pada parameter *learning_rate* yang pertama adalah nilai 0,1 dengan nilai *depth* 6, iterations 500 dan *l2_leafIreg* 1 skor nilai uji *mean_test_score cosine similarity* adalah 0,963687 sedangkan pada nilai parameter *learning_rate* kedua yaitu 0,01 dengan nilai *depth* 6, iterations 500 dan *l2_leafIreg* 1 skor nilai uji *mean_test_score cosine similarity* adalah 0,951850. Terlihat bahwa parameter *learning_rate* semakin bertambah maka semakin baik nilai skor uji *mean_test_score cosine similarity*. Sehigga parameter *learning_rate* dapat membantu meningkatkan skor uji *mean_test_score cosine similarity* jika nilainya semakin tinggi. Untuk lebih jelasnya berikut adalah gambar kombinasi parameter pengujian pada model 4.



Gambar 4. 7 Rata-rata *Cosine Similarity* Model 4

Parameter optimal yang telah ditemukan akan digunakan untuk pengujianpada data *training* dan *testing*. Sehingga pengujian pada data *training*

menghasilkan skor akurasi 1.0 atau sempurna dan pada data *testing* menghasilkan skor akurasi 95%. Berikut adalah gambar confusion matrix pengujian model 4.



Gambar 4. 8 Confusion Matrix Model 4

Dari gambar 4.8 menunjukan bahwa model 4 memprediksi 104 data kualitas udara baik dan pada hasil yang sebenarnya terdeteksi baik, memprediksi 12 data kualitas udara sedang akan tetapi pada data sebenarnya adalah baik, memprediksi 379 data kualitas udara sedang dan pada data yang sebenarnya adalah sedang, memprediksi 2 data kualitas udara baik akan tetapi pada data sebenarnya adalah sedang, memprediksi 34 data kualitas udara tidak sehat dan pada data yang sebenarnya adalah tidak sehat, memprediksi 12 data kualitas udara sedang akan tetapi pada data yang sebenarnya adalah tidak sehat dan memprediksi 1 data tidak sehat akan tetapi pada data yang sebenarnya adalah sedang.

Dengan melihat gambar 4.8 dapat diketahui nilai dari *accuracy*, *precision*, *recall,* dan *f1-score* dari masing-masing kelas dengan menggunakan teknik evaluasi. Mengacu kepada gambar 4.8 untuk kelas kategori BAIK nilai TP adalah 104, FN adalah 12, FP adalah 2 dan TN adalah 426. Perhitungan dari nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{104 + 379 + 34}{544}\ x\ 100\% = 950 \text{ dibulatkan } 95\%$$

$$Precission \text{ kategori BAIK} = \frac{104}{104 + 2}\ x\ 100\% = 981 \text{ dibulatkan } 98\%$$

$$Recall \text{ kategori BAIK} = \frac{104}{104 + 12}\ x\ 100\% = 896 \text{ dibulatkan } 90\%$$

$$F1\ score \text{ kategori BAIK} = 2\ \frac{0{,}981\ x\ 0{,}896}{0{,}981 + 0{,}896}\ x\ 100\% = 936 \text{ dibulatkan } 94\%$$

Mengacu kepada gambar 4.8 untuk kelas kategori SEDANG nilai TP adalah 379, nilai FN adalah 3, nilai FP adalah 24, nilai TN adalah 138. Perhitungan dari nilai tersebut adalah sebagai berikut.

$$Precission \text{ kategori SEDANG} = \frac{379}{379 + 24}\ x\ 100\% = 940 \text{ dibulatkan } 94\%$$

$$Recall \text{ kategori SEDANG} = \frac{379}{379 + 3}\ x\ 100\% = 992 \text{ dibulatkan } 99\%$$

$$F1\ score \text{ kategori SEDANG} = 2\ \frac{0{,}940\ x\ 0{,}992}{0{,}940 + 0{,}992}\ x\ 100\%$$

$$= 9653 \text{ dibulatkan } 97\%$$

Mengacu kepada gambar 4.8 untuk kelas kategori TIDAK SEHAT nilai TP adalah 34, nilai FN adalah 12, nilai FP adalah 1, nilai TN adalah 497. Perhitungan dari nilai tersebut adalah sebagai berikut.

$Precission$ kategori TIDAK SEHAT $= \frac{34}{34+1}\ x\ 100\% = 971$ dibulatkan 97%

$Recall$ kategori TIDAK SEHAT $= \frac{34}{34+12}\ x\ 100\% = 739$ dibulatkan 74%

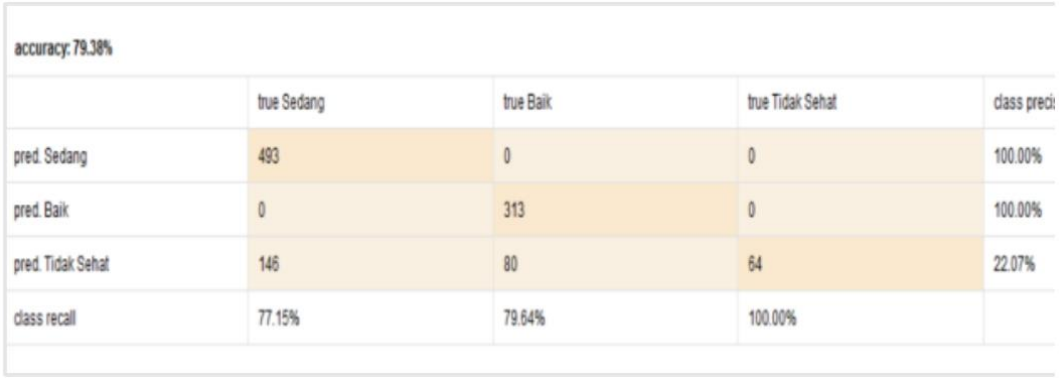
$F1\ score$ kategori TIDAK SEHAT $= 2\ \frac{0{,}97\ x\ 0{,}739}{0{,}97+0{,}739}\ x\ 100\%$

$= 839$ dibulatkan 84%

4.2 Pembahasan

Pada sub-bab ini merupakan penjelasan analisis dari hasil uji coba pada setiap model yang telah dilakukan. Penelitian ini menggunakan dataset yang diambil dari website kaggle.com yaitu data indeks standar pencemar udara di spku wilayah Jakarta dengan jumlah data sebanyak 1830 dan terdapat 10 fitur yaitu tanggal, stasiun, pm10, so2, co, o3, no2, max, critical, dan categori. Namun pada tahap pra pemrosesan data fitur tanggal, stasiun, dan critical tidak dipilih sehingga atribut tersebut dihilangkan karena fitur tersebut tidak dibutuhkan dan tidak relevan. Sehingga hanya 7 fitur yang akan digunakan yaitu 6 fitur sebagai independen dan 1 fitur sebagai dependen yaitu atribut categori. Fitur categori berisi data sedang sebanyak 1272, data baik sebanyak 385, data tidak sehat sebanyak 154, tidak ada data sebanyak 16 dan sangat tidak sehat sebanyak 3 data. Pada tahap pra pemrosesan data yang tidak ada dan data yang sangat tidak sehat akan dihilangkan karena data tersebut tidak mendukung dalam klasifikasi yang akan dilakukan. Sehingga pada atribut fitur categori hanya terdapat 3 kelas yaitu kelas sedang, baik, dan tidak sehat yang mana kelas tersebut akan menjadi kelas yang akan diprediksi sehingga menjadi kelas target dalam penelitian ini.

Penelitian sebelumnya yang dilakukan oleh Fitri dan yang lainnya pada tahun 2023 bertujuan untuk mengidentifikasi perbedaan kondisi lingkungan antara DKI Jakarta dan Kota Tangerang khususnya pada pengaruh kualitas udara. Metode yang digunakan adalah metode Naïve Bayes karena cocok dengan data yang digunakan yaitu data indeks kualitas udara yang bersifat kontinu. Data yang digunakan sebanyak 1096 data dengan 10 atribut. Terlihat pada gambar 4.9 bahwa model memiliki nilai akurasi total mencapai 79,38% sehingga menunjukan hasil kinerja model yang positif. Model yang dibuat juga mengidentifikasi nilai presisi untuk kelas Sedang dan Baik sebesar 100% yang artinya semua yang diprediksi sebagai kelas tersebut benar benar milik kelas tersebut, namun pada kelas Tidak Sehat nilai presisi sebesar 22% yang artinya sejumlah besar data diprediksi tidak sehat akan tetapi sebenarnya bukan merupakan data yang tidak sehat. Nilai recall yang didapatkan juga terlihat pada gambar 4.9 bahwa untuk kelas Sedang nilai recallnya adalah 77.15% , untuk kelas Baik adalah 79.64%, sedangkan untuk kelas Tidak Sehat nilai recallnya adalah 100% (Widiawati et al., 2023). Meskipun model memiliki nilai presisi yang lebih tinggi untuk kelas Sedang dan Baik akan tetapi pada kelas Tidak Sehat nilai presisinya rendah. Namun sebaliknya untuk nilai recall pada kelas Tidak Sehat sangat tinggi yang menunjukan bahwa kemampuan model dalam mengidentifikasi kelas dapat menunjukan sebagian besar dari kelas tersebut.

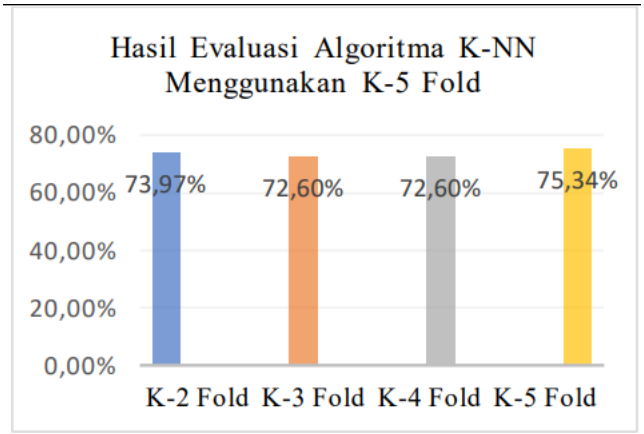


accuracy: 79.38%

	true Sedang	true Baik	true Tidak Sehat	class preci
pred. Sedang	493	0	0	100.00%
pred. Baik	0	313	0	100.00%
pred. Tidak Sehat	146	80	64	22.07%
class recall	77.15%	79.64%	100.00%	

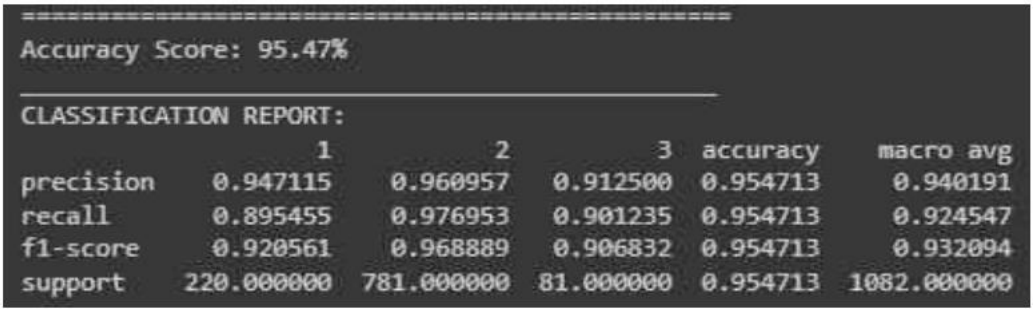
Gambar 4. 9 Hasil Akurasi Penelitian Fitri Widiawati dkk

Penelitian lain yang dilakukan oleh Wiranata dkk pada tahun 2023 dengan mengangkat permasalahan polusi udara di DKI Jakarta yang serius sehingga menyebabkan masalah kesehatan dan mempengaruhi kualitas lingkungan. Tujuan penelitian yang dilakukan oleh mereka adalah untuk mengetahui akurasi dari pengklasifikasian kualitas udara yang ada di provinsi DKI Jakarta dengan menggunakan metode *K-Nearest Neightbors* atau KNN. Dalam penelitiannya menggunakan K-5 fold diperoleh hasil akurasi pada K-2 fold mencapai 73,97%, K-3 fold mencapai 72,60%, K-4 fold mencapai 72,60%, dan pada K-5 fold mencapai 75,35%. Data yang digunakan untuk penelitian oleh Wiranata dkk berjumlah 244 data dengan pembagian 171 *training* dan 73 *testing*. Atribut pada dataset tersebut berjumlah 9 atribut yaitu pm10, so2, co, o3, no2, max, critical, lokasi, dan categori (Wiranata et al., 2023). Berikut adalah hasil akurasi penelitian dari Wiranata dkk.



Gambar 4. 10 Grafik Hasil Akurasi Penelitian Wiranata dkk

Penelitain yang dilakukan oleh Wahyudiyanta dan Supriyati pada analisis kualitas udara Jakarta dengan menggunakan metode *Support Vector Machine* (SVM) juga mencapai hasil yang baik. Tujaun penelitian mereka adalah untuk mengklasifikasikan kualitas udara di DKI Jakarta. Data yang digunakan adalah bersumber dari Jakarta Open Data pada tahun 2020 yang berjumlah 1663 data dengan 10 atribut. Atribut tersebut melibatkan parameter Indeks Standar Pencemar Udara (ISPU) yaitu tanggal, wilayah, pm10, so2, co, o3, no2, max, critical, dan categori. Dari hasil evaluasi pada penelitian yang dilakukan oleh Wahyudiyanta dan Supriyati mendapatkan akurasi mencapai 95,47%, presisi mencapai 96,09%, recall mencapai 97,70% dan f1-score mencapai 96,88% (Wahyudiyanta & Supriyati, 2024). Dengan demikian dapat disimpulkan bahwa performa model tersebut memiliki kualitas yang baik untuk membuat prediksi dan dapat digunakan untuk mengembangkan penelitian yang selanjutnya. Berikut adalah hasil akurasi dari penelitain Wahyudiyanta dan Supriyati.



```
================================================
Accuracy Score: 95.47%
------------------------------------------------
CLASSIFICATION REPORT:
                    1           2           3  accuracy    macro avg
precision    0.947115    0.960957    0.912500  0.954713     0.940191
recall       0.895455    0.976953    0.901235  0.954713     0.924547
f1-score     0.920561    0.968889    0.906832  0.954713     0.932094
support    220.000000  781.000000  81.000000  0.954713  1082.000000
```

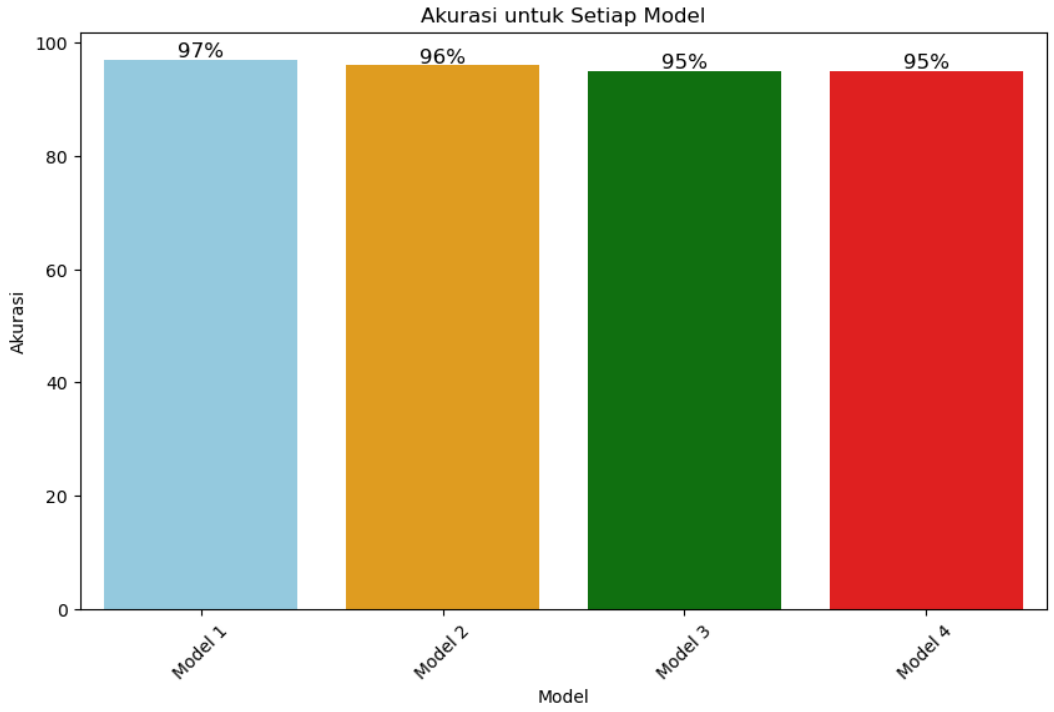
Gambar 4. 11 Hasil Akurasi Penelitian Wahyudiyanta dan Supriyati

Pada penelitian yang dilakukan oleh peneliti untuk memprediksi indeks kualitas udara di daerah Jakarta menggunakan metode *CatBoost* dengan data indeks pencemar udara di spku wilayah Jakarta yang diambil dari Kaggle. Peneliti membagi menjadi 4 model yaitu dengan perbandingan 90/10 untuk model 1, 80/20 untuk model 2, 75/25 untuk model 3, dan 70/30 untuk model 4. Pada tahap pemodelan peneliti menggunakan *GridSearchCV* untuk penyetelan beberapa parameter yang ditentukan dengan tujuan meningkatkan nilai akurasi. Parameter yang digunakan adalah *depth*, *iterations*, *learning_rate* dan *l2_leaf_reg*. Proses yang pertama dilakukan adalah preprocessing data yang merupakan analisis data dan merubah data mentah menjadi data yang siap digunakan untuk membuat model klasifikasi dengan melalui beberapa tahapan (Kohsasih & Situmorang, 2022). Pada presprocessing data yang pertama dilakukan adalah *missing value* yang bertujuan untuk membersihkan data yang tidak sesuai ataupun data yang isinya kosong. Sehingga data yang tidak sesuai dan kosong akan dihilangkan. Data yang dihilangkan adalah atribut tanggal, stasiun, dan critical. Setelah *missing value*, proses yang akan dilakukan adalah split data dengan membagi data sesuai skenario yang telah ditentukan. Proses berikutnya adalah Pembangunan model

menggunakan metode *CatBoost* dengan penyetelan parameter menggunakan *GridSearchCV* untuk menemukan kombinasi parameter terbaik. Parameter yang digunakan adalah *depth* dengan nilai 6,8 dan 10, *learning_rate* 0,1 dan 3, *iterations* 500, 1000, dan 1500, *l2_leaf_reg* 1,2, dan 3. Sehingga menemukan kombinasi parameter terbaik yang akan digunakan untuk model klasifikasi. Selanjutnya hasil dari setiap model akan dievaluasi. Berikut adalah hasil akurasi dari setiap model.

Tabel 4. 5 Hasil Akurasi Setiap Model

No	**Model**	**Banyaknya Data = 1811**				**Akurasi**
		Training		***Testing***		
		Presentase	**Jumlah**	**Presentase**	**Jumlah**	
1	Model 1	90%	1629	10%	182	97%
2	Model 2	80%	1448	20%	363	96%
3	Model 3	75%	1353	25%	453	95%
4	Model 4	70%	1267	30%	544	95%

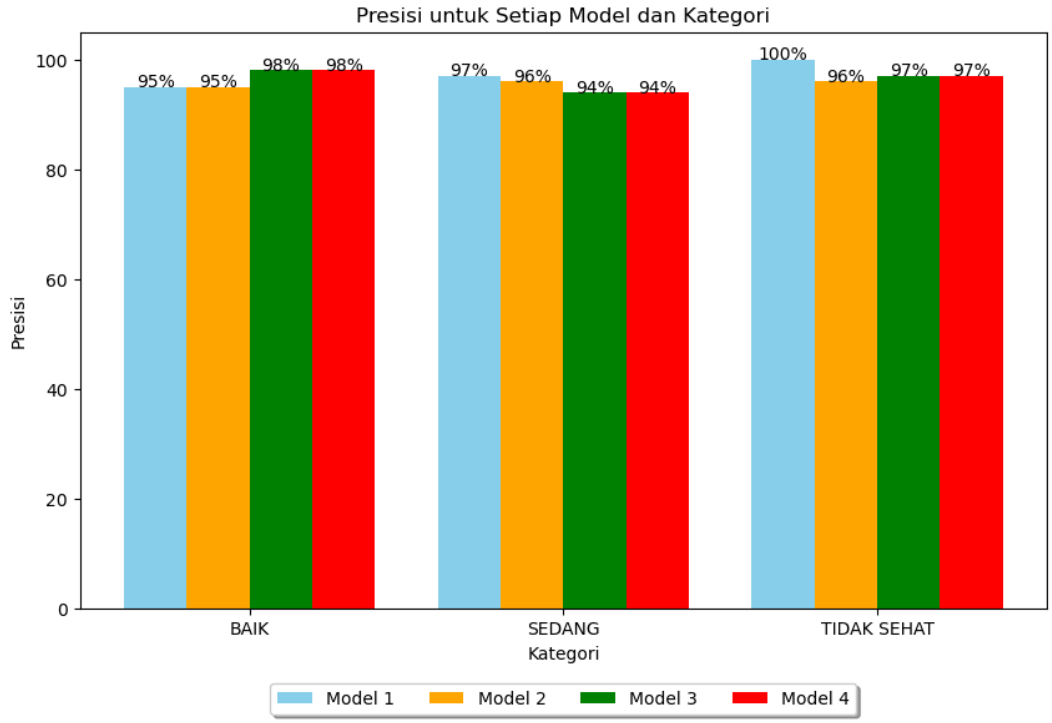


Gambar 4. 12 Hasil Akurasi Setiap Model

Pada tabel 4.5 menunjukan bahwa metode *CatBoost* memberikan performa yang baik dalam melakukan klasifikasi untuk memprediksi indeks kualitas udara pada data indeks standar pencemar udara di spku wilayah Jakarata. Hasil akurasi terbaik diperoleh dari pengujian model 1 dengan penyetelan parameter terbaik pada kombinasi *depth* 6, *iterations* 1500, *l2_leaf_reg* 1 dan *learning_rate* 0,1 dengan skor uji 0,958870. Model 1 menggunakan perbandingan data *training* 90% dengan jumlah data 1629 dan data *testing* 10% dengan jumlah data 182. Sehingga mendapatkan hasil akurasi keseluruhan mencapai 0,97. Terlihat pada gambar 4.12 Model 1 mendapatkan nilai akurasi yang lebih tinggi dibandingkan ketiga model yang lain karena pengaruh dari pembagian data *training* dan *testing*. Model 1 dengan adanya data *training* 90% akan lebih banyak data untuk pelatihan sehingga dengan banyaknya data *training* sistem akan melatih data dengan maksimal, dan pada data *testing* 10% data lebih sedikit sehingga sistem dapat memprediksi dan menguji dengan maksimal pada data yang sedikit karena sistem telah belajar melatih data yang banyak dan pada prediksi pengujian data semakin sedikit. Model 1 adalah model paling terbaik dibandingkan dengan model ketiga lainnya dengan perbandingan data yang telah dilakukan. Pembagian data *training* dan data *testing* ini sangat penting karena sangat berpengaruh kepada nilai akurasi yang didapatkan (Okprana & Winanjaya, 2022). Dengan demikian dapat terlihat seberapa bagus model dapat memperkirakan hasil berdasarkan data yang sebelumnya tidak terlihat dengan adanya pembagian data *training* dan *testing*.

Evaluasi model yang dilakukan menggunakan *confussion matrix* tidak hanya mendapatakan pengetahuan mengenai nilai akurasi saja, akan tetapi juga

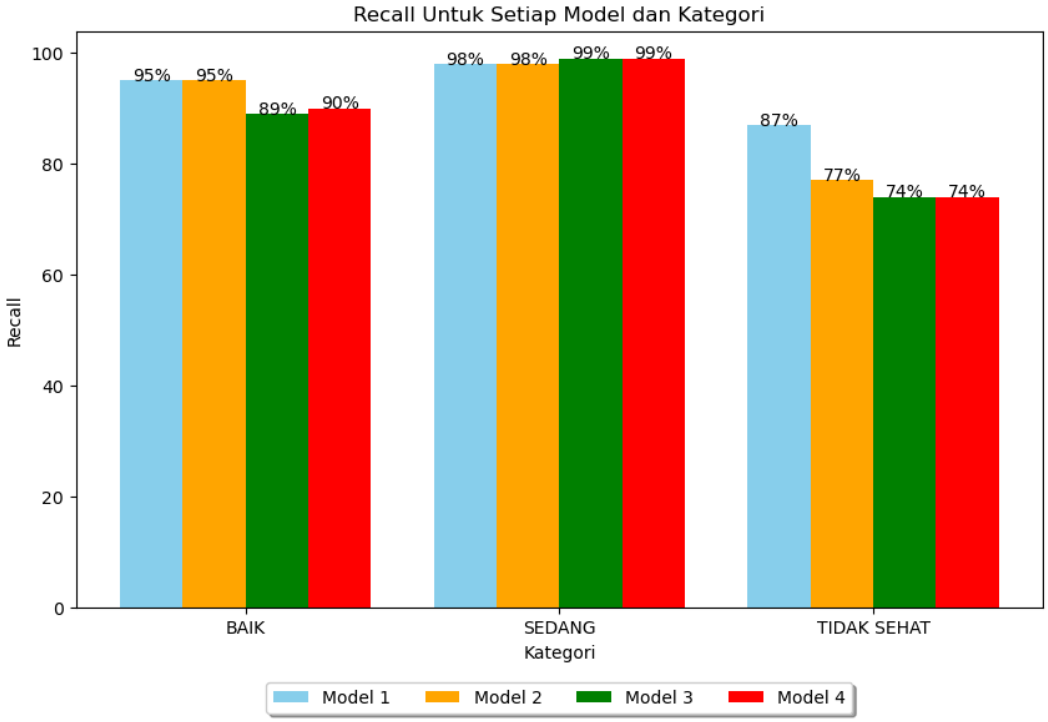
prescission, recall dan *f1-score*. *Pressicion* merupakan perbandingan antara prediksi *True Positives* (TP) dengan total nilai positif yang didapatkan. Sehingga pada evaluasi yang dilakukan *pressicion* memfasilitasi keakuratan model dalam mendeteksi data positif. Dengan demikian jika nilai presisi meningkat maka jumlah kesalahan dalam memprediksi data positif semakin berkurang karena *pressicion* memberikan pemahaman tentang akuratnya model dalam memprediksi data yang positif. Berikut merupakan hasil nilai *pressicion* dari setiap model.



Gambar 4. 13 Hasil Presisi Setiap Model

Pada gambar 4.13 terlihat bahwa nilai presisi pada model 1 adalah yang paling tinggi walaupun pada kategori baik nilai presisinya hanya mencapai 0.95 namun pada kategori sedang mencapai 0.97 dan kategori tidak sehat mencapai 1.00. Nilai presisi yang tinggi menunjukan bahwa prediksi pada data yang positif

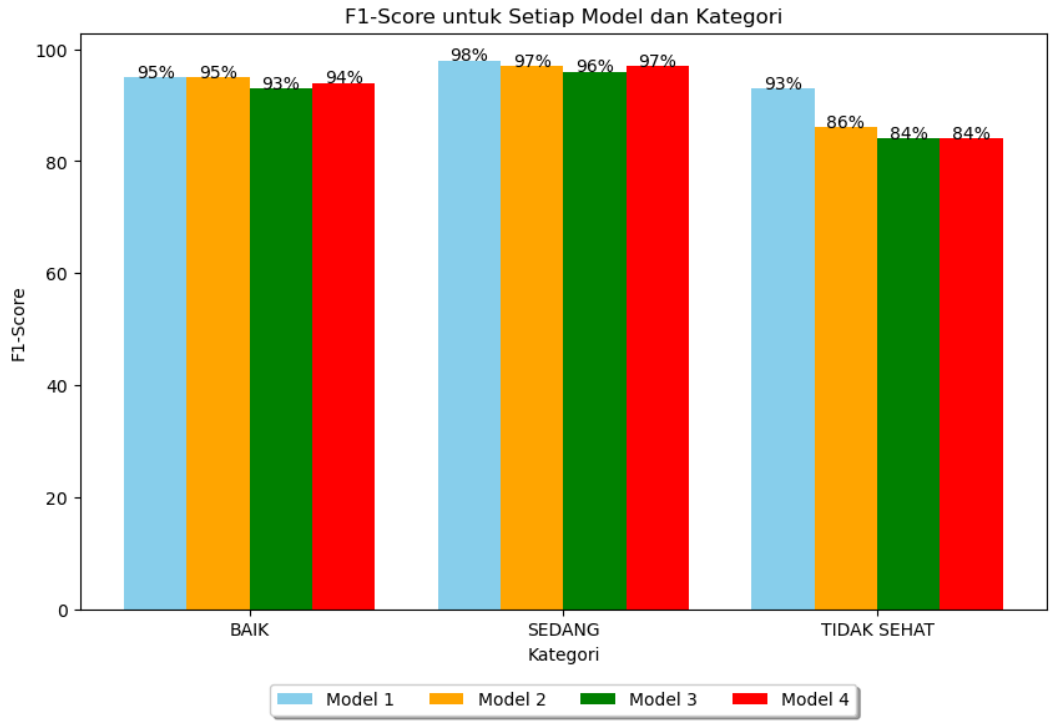
memang hasilnya yang sebenarnya adalah positif (Siregar et al., 2020). Evaluasi yang selanjutnya adalah *recall* yang merupakan perbandingan antara prediksi *True Positives* (TP) dengan seluruh data yang benar positif. Sehingga pada evaluasi yang dilakukan *recall* memberikan pemahaman seberapa sering model memprediksi positif ketika data sebenarnya adalah positif (Saputro & Sari, 2020). Berikut adalah hasil nilai *recall* pada setiap model.



Gambar 4. 14 Hasil Recall Setiap Model

Jika *pressicion* lebih mengarah kepada tingkat kesalahan model dalam memprediksi data sebagai positif namun seharusnya negatif, maka *recall* mengarah kepada tingkat kesalahan model yang rendah dengan tidak memprediksi sebagai posistif namun seharusnya positif. Pada gambar 4.14 terlihat bahwa pada semua model mendapatakn nilai *recall* yang tinggi sehingga dapat dikatakan banyak data

positif yang diprediksi benar oleh model. Model 1 pada kategori tidak sehat adalah nilai tertinggi walaupun pada kategori baik, model 1 dan model 2 bernilai sama dan pada kategori sedang model 1 nilainya lebih rendah dibandingkan dengan model 3 dan 4. Evaluasi yang selanjutnya adalah *f1-score* yang merupakan nilai rata-rata harmonic dari tingkat presisi dan perolehan (Baharuddin et al., 2019). *F1-score* adalah integrasi antara nilai *pressicion* dan *recall* yang menjadi satu nilai sebagai kualitas model prediksi secara keseluruhan. Sehingga dapat menilai akurat atau tidaknya model dalam memprediksi data positif dengan benar. Selain itu juga membantu menyeimbangkan antara presisi dan perolehan. Berikut adalah hasil nilai *f1-score* pada setiap model.



Gambar 4. 15 Hasil F1-Score Setiap Model

Terlihat pada gambar 4.15 nilai *f1-score* pada semua model mencapai hasil yang baik. Sehingga dapat dikatakan bahwa model memiliki nilai keseimbangan untuk memprediksi sebagian besar kategori yang tepat (*recall)* dan menemukan prediksi yang akurat (*pressicion*). Pada setiap model dari beberapa parameter yang ditentukan didapatkan kombinasi parameter terbaik yang berbeda. Model 1 kombinasi parameter terbaik yang didapatkan adalah pada *depth* 6, *iterations* 1500, *l2_leaf_reg* 1 dan *learning_rate* 0,1 dengan hasil rata-rata sebesar 0.958870. Model 2 kombinasi parameter terbaik yang didapatkan adalah pada *depth* 6, *iterations* 1500, *l2_leaf_reg* 1 dan *learning_rate* 0,1 dan pada *depth* 6, *iterations* 1500, l2_leaaf_reg 2 dan *learning_rate* 0,1 dengan hasil rata-rata sebesar 0.959264. Model 3 kombinasi terbaik yang didapatkan adalah pada *depth* 8, *iterations* 500, *l2_leaf_reg* 2 dan *learning_rate* 0,1 dan pada *depth* 10, *iterations* 500, l2_leaaf_reg 2 dan *learning_rate* 0,1 dengan hasil rata-rata sebesar 0.958028. Model 4 kombinasi terbaik yang didapatkan adalah pada *depth* 10, *iterations* 1000, *l2_leaf_reg* 2 dan *learning_rate* 0,1 dan pada dept 10, *iterations* 1500, l2_leaaf_reg 2 dan *learning_rate* 0,1 dengan hasil rata-rata sebesar 0.964477. Terdapat beberapa perbedaan kombinasi pada setiap model yang dibangun. Hal ini dikarenakan perbedaan pembagian pada data *training* dan *testing* ataupun karena beberapa eksperimen berbagai kombinasi parameter untuk menemukan performa yang terbaik pada pengujian. Berikut adalah perbandingan hasil akurasi peneliti dengan peneliti yang lainya.

Tabel 4. 6 Analisis Parameter *depth*

No	**Model 1**				***mean_test_score (cosine similarity)***
	depth	***iterations***	***l2_leaf_reg***	***learning_rate***	
1	6	500	1	0,1	0.957643
2	8	500	1	0,1	0.955801
3	10	500	1	0,1	0.956415
	Model 2				
4	6	500	1	0,1	0.957882
5	8	500	1	0,1	0.959262
6	10	500	1	0,1	0.958572
	Model 3				
7	6	500	1	0,1	0.956555
8	8	500	1	0,1	0.958026
9	10	500	1	0,1	0.956555
	Model 4				
10	6	500	1	0,1	0.963687
11	8	500	1	0,1	0.962108
12	10	500	1	0,1	0.962898

Terlihat pada tabel 4.6 parameter *depth* pada model 1 yang pertama adalah nilai 6 skor ujinya adalah 0.957643 dan pada parameter *depth* 8 skor ujinya adalah 0.955801 dan pada parameter *depth* 10 skor ujinya adalah 0,956415. Terlihat bahwa bertambahnya nilai *depth* dari 6, 8, dan 10 skor rata-rata pada masing-masing model terdapat skor yang meningkat, terdapat skor yang sama dan juga terdapat skor yang menurun. Sehingga dapat disimpulkan jika nilai *depth* lebih kecil, skor rata-rata dari *mean_test_score cosine similarity* dapat optimal dan jika lebih besar dapat menyebabkan penurunan skor *mean_test_score cosine similarity* pada kombinasi parameter yang tepat sesuai dengan data yang digunakan. Sehingga nilai parameter *depth* akan membantu meningkatkan skor *mean_test_score cosine similarity* jika pada nilai yang tepat, tidak kurang dan tidak lebih.

Tabel 4. 7 Analisis Parameter *iterations*

No	Model 1				*mean_test_score (cosine similarity)*
	depth	***iterations***	***l2_leaf_reg***	***learning_rate***	
1	6	500	1	0,1	0.957643
2	6	1000	1	0,1	0.957643
3	6	1500	1	0,1	0.958870
	Model 2				
4	6	500	1	0,1	0.957882
5	6	1000	1	0,1	0.958574
6	6	1500	1	0,1	0.959264
	Model 3				
7	6	500	1	0,1	0.956555
8	6	1000	1	0,1	0.956555
9	6	1500	1	0,1	0.956555
	Model 4				
10	6	500	1	0,1	0.963687
11	6	1000	1	0,1	0.963687
12	6	1500	1	0,1	0.963687

Terlihat pada tabel 4.7 bahwa bertambahnya nilai *iterations* dari 500, 1000, dan 1500 skor *mean_test_score cosine similarity* pada masing-masing model terdapat skor yang meningkat dan terdapat skor yang sama, akan tetapi tidak terdapat nilai skor yang menurun. Sehingga dapat simpulkan jika nilai *iterations* lebih kecil, skor rata-rata *mean_test_score cosine similarity* tidak maksimal dan jika lebih besar akan menyebabkan peningkatan skor *mean_test_score cosine similarity* pada kombinasi parameter yang tepat sesuai dengan data yang digunakan. Sehingga nilai parameter *iterations* akan membantu meningkatkan skor *mean_test_score cosine similarity* jika parameter *iterations* bertambah pada nilai yang tepat.

Tabel 4. 8 Analisis Parameter *l2_leaf_reg*

No	Model 1				*mean_test_score (cosine similarity)*
	depth	***iterations***	***l2_leaf_reg***	***learning_rate***	
1	6	500	1	0,1	0.957643
2	6	500	2	0,1	0.953959
3	6	500	3	0,1	0.953346
	Model 2				
4	6	500	1	0,1	0.957882
5	6	500	2	0,1	0.957882
6	6	500	3	0,1	0.958572

No	Model 1				*mean_test_score (cosine similarity)*
	depth	***iterations***	***l2_leaf_reg***	***learning_rate***	
	Model 3				
7	6	500	1	0,1	0.956555
8	6	500	2	0,1	0.955819
9	6	500	3	0,1	0.956555
	Model 4				
10	6	500	1	0,1	0.963687
11	6	500	2	0,1	0.960532
12	6	500	3	0,1	0.959742

Terlihat pada tabel 4.8 bahwa bertambahnya nilai *l2_leaf_reg* dari 1, 2, dan 3, nilai *mean_test_score cosine similarity* pada masing-masing model terdapat skor yang meningkat, terdapat skor yang sama dan juga terdapat nilai skor yang menurun. Sehingga dapat disimpulkan jika nilai *l2_leaf_reg* lebih kecil, skor *mean_test_score cosine similarity* tidak maksimal dan jika lebih besar akan menyebabkan peningkatan dan penurunan skor *mean_test_score cosine similarity* pada kombinasi parameter yang tepat sesuai dengan data yang digunakan. Sehingga nilai parameter *l2_leaf_reg* akan membantu meningkatkan skor *mean_test_score cosine similarity* jika pada kombinasi parameter yang tepat.

Tabel 4. 9 Analisis Parameter *learning_rate*

No	Model 1				*mean_test_score (cosine similarity)*
	depth	***iterations***	***l2_leaf_reg***	***learning_rate***	
1	6	500	1	0,1	0.957643
2	6	500	1	0,01	0.945979
	Model 2				
3	6	500	1	0,1	0.957882
4	6	500	1	0,01	0.949595
	Model 3				
5	6	500	1	0,1	0.956555
6	6	500	1	0,01	0.951402
	Model 4				
7	6	500	1	0,1	0.963687
8	6	500	1	0,01	0.951850

Terlihat pada tabel 4.9 bahwa bertambahnya nilai *learning_rate* dari 0,01 ke 0,1 skor *mean_test_score cosine similarity* pada masing-masing model

meningkat. Sehingga dapat disimpulkan jika nilai *learning_rate* lebih kecil, nilai *mean_test_score cosine similarity* tidak maksimal dan jika lebih besar akan menyebabkan peningkatan nilai *mean_test_score cosine similarity*. Dengan demikian nilai parameter *learning_rate* akan membantu meningkatkan *mean_test_score cosine similarity* dengan menambahkan nilai parameter.

Tabel 4. 10 Perbandingan Peneliti Dengan Penelitian Lain

No	Sitasi	Metode	Akurasi Terbaik
1	(Widiawati et al., 2023)	*Naive Bayes* menggunakan *RapidMiner*	79,38%
2	(Wiranata et al., 2023)	*K-Nearest Neighbors* dengan K-5 *fold* menggunakan *RapidMiner*	75,34% pada K-5 *fold*
3	(Wahyudiyanta & Supriyati, 2024)	*Support Vector Machine* (SVM)	95,47%
4	**Peneliti**	***CatBoost* + *GridSearchCV***	**97%**

Pada table 4.10 terlihat bahwa peneliti menggunakan *GridSearchCV* sehingga dapat mendapatakn hasil akurasi yang optimal. Hal ini dikarenakan cara kerja *GridSerachCV* yang membagi data kedalam beberapa lipatan melalui pendekatan *cross validation* kemudian dilatih dan diuji menggunakan metode *CatBoost* dengan parameter yang terbaik. Sehingga dapat menghasilkan kombinasi parameter terbaik untuk mendapatkan model yang optimal dengan hasil akurasi yang optimal dan lebih baik. Hasil akurasi terbaik yang didapatkan adalah 96,70% dapat dibulatkan menjadi 97%. Hasil penelitian yang dilakukan oleh Darmawan dan Dianta (2023) juga mendapatkan nilai yang optimal menggunakan *GridSearchCV* dengan pendekatan *cross validation* karena dapat menunjukan adanya peningkatan nilai akurasi setelah adanya proses dari *GridSearchCV* hingga mencapai nilai akurasi 86,0% yang sebelumnya adalah 83.51% (Darmawan &

Fauzan Dianta, 2023). Dalam penelitian yang dilakukan oleh Subarkah dkk menjelaskan bahwa nilai akurasi dapat dikelompokan kedalam nilai yang baik atau gagal. Pada penelitian subarkah menjelaskan menurut Gorunescu, nilai akurasi dari 90% - 100% masuk kedalam kategori yang sangat baik, nilai akurasi dari 80%-90% adalah kategori baik, nilai akurasi dari 70% - 80% adalah kategori cukup baik, nilai akurasi dari 60% - 70% adalah kategori kurang baik dan nilai akurasi dari 50% - 60% adalah kategori yang gagal (Subarkah et al., 2019). Sehingga berdasarkan kelompok tersebut maka dapat dikatakan penelitain prediksi indeks kualitas udara menggunakan metode *CatBoost* yang telah dilakukan dapat dikatakan kedalam kelompok nilai akurasi yang sangat baik.

4.3 Integrasi Penelitian dengan Al-Qur'an

Polusi udara merupakan permasalahan yang dampaknya sangat berpengaruh kepada kehidupan makhluk hidup di bumi baik itu tumbuhan, hewan ataupun manusia. Polusi udara dapat menyebabkan banyaknya kemunculan penyakit yang serius seperti penyakit pada pernafasan. Kualitas udara juga berpengaruh kepada ekosistem serta kesehatan manusia pada kualitas hidup secara keseluruhan (Maulana et al., 2024). Berdasarkan data *AirVisual* dari AQI pada tahun 2024, Negara Indonesia tepatnya Kota Jakarta telah memasuki daftar ke-11 menjadi negara dengan tingkat polusi tertinggi di dunia hingga mencapai angka 127 dalam artian tidak sehat bagi kelompok yang sensitive (IQAir, n.d.). Dengan demikian polusi udara di Jakarta telah menjadi masalah yang serius karena berdampak buruk pada kesehatan dan ekosistem. Hal ini disebabkan oleh beberapa aktifitas manusia yang mempengaruhi kualitas udara di muka bumi seperti asap

rokok, pembakaran hutan, transportasi, dan aktivitas industri lainya sesuai pada firman Allah SWT pada QS. Ar-Rum ayat 41 yang berbunyi :

ظَهَرَ ٱلْفَسَادُ فِى ٱلْبَرِّ وَٱلْبَحْرِ بِمَا كَسَبَتْ أَيْدِى ٱلنَّاسِ لِيُذِيقَهُم بَعْضَ ٱلَّذِى عَمِلُوا۟ لَعَلَّهُمْ يَرْجِعُونَ

"*Telah tampak kerusakan di darat dan di laut disebabkan perbuatan tangan manusia. (Melalui hal itu) Allah membuat mereka merasakan sebagian dari (akibat) perbuatan mereka agar mereka kembali (ke jalan yang benar)*" *(QS. Ar-Rum: 41).*

Tafsir Wajiz yang dipublikasikan oleh NU Online menjelaskan bahwa "*Bila pada ayat-ayat sebelumnya Allah menjelaskan sifat buruk orang musyrik Mekah yang menuhankan hawa nafsu, melalui ayat ini Allah menegaskan bahwa kerusakan di bumi adalah akibat mempertuhankan hawa nafsu. Telah tampak kerusakan di darat dan dilaut, baik itu kota maupun desa yang disebabkan karena perbuatan tangan manusia yang dikendalikan oleh hawa nafsu dan jauh dari tuntunan fitrah. Allah menghendaki agar mereka merasakan sebagian dari akibat perbuatan buruk mereka agar kembali ke jalan yang benar dengan menjaga kesesuaian perilakuknya dengan fitrahnya*" (NUOnline, n.d.). Dalam tafsir tersebut dapat ditarik kesimpulan bahwa Allah menjelaskan ayat ini agar manusia dapat menjaga kesesuain perilakunya dengan fitrahnya supaya ekosistem dan kehidupan dimuka bumi menjadi lebih baik sehingga polusi udara berkurang dan lingkungan serta hewan dan tumbuhan juga sehat.

Jika kehidupan di muka bumi masih tidak stabil dan masih banyak kerusakan seperti polusi udara yang disebabkan oleh tangan manusia sendiri. Maka akan banyak permasalahan-permasalahan yang muncul sehingga menyebabkan adanya penyakit pernafasan dan penyakait lainnya. Jika hal ini tidak segera

ditangani maka akan terus berlanjut dan semakin parah sehingga dapat menyebabkan tingkat kematian yang tinggi. Harapan besar dengan adanya sistem prediksi indeks kualitas udara dengan menggunakan metode *CatBoost* yang lebih cepat dan efisien untuk memprediksi indeks kualitas udara dapat membantu tenaga kesehatan atau pemerintahan dalam menangani polusi udara untuk ditinjak lanjuti. Sehingga dapat memberikan informasi sejak dini dan dapat mengurangi adanya penyakit yang disebabkan oleh tidak stabilnya kualitas udara. Sehingga sistem prediksi indeks kualitas udara ini dapat menjadi obat penyakit sejak dini dengan memberikan informasi kualitas udara yang ada. Seperti pada hadist nabi yang menjelaskan bahwa semua penyakit pasti ada obatnya (Hadits, n.d.).

Rasulullah Shallahu ‘Alaihi Wa Salam bersabda :

حَدَّثَنَا هَارُونُ بْنُ مَعْرُوفٍ وَأَبُو الطَّاهِرِ وَأَحْمَدُ بْنُ عِيسَى قَالُوا حَدَّثَنَا ابْنُ وَهْبٍ أَخْبَرَنِي عَمْرٌو وَهُوَ ابْنُ الْحَارِثِ عَنْ عَبْدِ رَبِّهِ بْنِ سَعِيدٍ عَنْ أَبِي الزُّبَيْرِ عَنْ جَابِرٍ عَنْ رَسُولِ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ أَنَّهُ قَالَ لِكُلِّ دَاءٍ دَوَاءٌ فَإِذَا أُصِيبَ دَوَاءُ الدَّاءِ بَرَأَ بِإِذْنِ اللَّهِ عَزَّ وَجَلَّ

“*Telah menceritakan kepada kami Harun bin Ma’ruf dan Abu Ath Thahir serta Ahmad bin Isa mereka berkata; Telah menceritakan kepada kami Ibnu Wahb; Telah mengabarkan kepadaku Amru yaitu Ibnu Al Harits dari Abdu Rabbih bin Sa’id dari Abu Zubair dari Jabir dari Rasulullah shallallahu ‘alaihi wa sallam, beliau bersabda : ‘Setiap penyakit ada obatnya. Apabila ditemukan obat yang tepat untuk suatu penyakit, maka akan sembuhlah penyakit itu dengan izin Allah ‘azza wajalla*’” (*HR Muslim*).

Prediksi kualitas udara merupakan bentuk pencegahan dini dalam menangani penyakit yang disebabkan oleh adanya polusi udara yang buruk. Adanya prediksi kualitas udara dapat membantu tenaga kesahatan atau pemerintahan untuk mendeteksi kualitas udara yang baik, sehingga dapat menjadi penanganan sejak dini

dan dapat mengurangi penyakit pada manusia ataupun kerusakan lingkungan. Dengan demikian sistem prediksi indeks kualitas udara ini dapat menjadi kebaikan yang bermanfaat untuk kehidupan dimuka bumi. Dengan kebaikan tersebut janganlah membuat kerusakan lain dan berdoalah dengan rasa takut serta berharaplah tidak akan terjadi kerusakan seperti polusi udara yang menyebabkan penyakit dan dapat menimbulkan tingkat kematian yang tinggi karena sesungguhnya orang baik dekat dengan rahmat Allah swt. Seperti yang dijelaskan pada firman Allah SWT pada Al quran surat Al-a'raf ayat 56 :

وَلَا تُفْسِدُوا۟ فِى ٱلْأَرْضِ بَعْدَ إِصْلَٰحِهَا وَٱدْعُوهُ خَوْفًا وَطَمَعًا ۚ إِنَّ رَحْمَتَ ٱللَّهِ قَرِيبٌ مِّنَ ٱلْمُحْسِنِينَ

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

Tafsir Ibnu Katsir (Ringkas)/Fatkhul Karim Mukhtashar Tafsir al-qur an al-adzhim, karya Syaikh Prof. Dr. Hikmat bin Basyir bin Yasin, professor fakultas al-qur an Universitas Islam Madinah mengatakan bahwa "*Firman Allah SWT: (Dan janganlah kalian membuat kerusakan di muka bumi, sesudah (Allah) memperbaikinya) Allah SWT melarang berbuat kerusakan di bumi dan hal-hal yang memberi kemudharatan setelah adanya perbaikan. Sesungguhnya jika segala sesuatu berjalan sesuai dengan kelestariannya, kemudian terjadi kerusakan setelah itu, maka itu memberi kemudharatan kepada semua hamba. Jadi Allah SWT melarang hal itu, dan memerintahkan untuk menyembah dan berdoa kepadaNya serta merendahkan diri dan tunduk kepadaNya. Lalu Allah SWT berfirman: (dan berdoalah kepada-Nya dengan rasa takut (tidak akan diterima) dan harapan (akan*

dikabulkan)) yaitu dengan takut terhadap siksaan yang ada di sisiNya dan berharap kepada pahala melimpah yang ada di sisiNya. Kemudian Allah SWT berfirman: (Sesungguhnya rahmat Allah amat dekat kepada orang-orang yang berbuat baik) yaitu sesungguhnya rahmat Allah mendatangi orang-orang yang berbuat baik yang mana mereka mengikuti perintah-perintahNya dan menjauhi larangan-laranganNya. Sebagaimana Allah SWT berfirman: (dan rahmat-Ku meliputi segala sesuatu. Maka akan Aku tetapkan rahmat-Ku untuk orang-orang yang bertakwa, yang menunaikan zakat dan orang-orang yang beriman kepada ayat-ayat Kami" (156) (Yaitu) orang-orang yang mengikut Rasul, Nabi yang ummi) (Surah Al-A'raf) dan Dia berfirman (dekat) tidak dikatakan "qariibatun" karena hal itu mengandung kata "rahmat" yang bermakna pahala, atau karena dimudhafkan kepada Allah, Oleh karena itu Allah berfirman (Qariibun minal muhsiniin) "amat dekat kepada orang-orang yang berbuat baik". (TafsirWeb, n.d.-c)

Dengan penjelasan tafsir diatas kita sebagai manusia yang hidup di bumi yang sudah diciptakan oleh Allah swt dengan baik harus menjaganya dan tidak diperbolehkan untuk merusak karena menimbulkan hal-hal yang tidak baik atau kemudharatan. Sehingga muncul larangan Allah swt untuk merusak muka bumi setelah diperbaiki. Allah swt memerintahkan untuk menyembah kepada-Nya dengan tunduk dan berdoa dengan rasa takut terhadap siksaan Allah swt serta mengaharapkan pahala yang banyak dan melimpah. Sesungguhnya orang-orang ynag berbuat baik sangat dekat dengan rahmat Allah swt. Maka dari itu untuk mengurangi banyaknya penyakit dan tingkat kematian yang disebabkan oleh

adanya polusi udara akibat tangan manusia sendiri kita harus menjaganya dengan berusaha dalam pemantauan prediksi indeks kualitas udara secara baik dan efisien. Dengan adanya sistem prediksi indek kualitas udara dapat bermanfaat dalam mengukur kualitas udara sehingga dapat dijadikan sebagai solusi atau obat supaya tidak terjadi penyakit di muka bumi sehingga hal ini juga merupakan kebaikan dengan harapan supaya mendapatkan rahmat Allah swt. Karena rahmat Allah swt dekat dengan orang-orang yang berbuat baik.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan skenario uji coba pada penelitian ini dengan beberapa pembagian model menjadi 4. Model 1 90% *training* dan 10% *testing*, model 2 80% *training* dan 20% *testing*, model 3 75% *training* dan 25% *testing*, dan model 4 70% *training* dan 30% *testing*. Model 1 menghasilkan nilai akurasi yang paling tertinggi sehingga dapat dikatakan model 1 adalah model yang paling terbaik dibandingkan dengan model yang lain dalam memprediski indeks kualitas udara di wilayah Jakarta pada data indeks standar pencemar udara di spku wilayah Jakarta yang diambil dari Kaggle menggunakan metode *CatBoost*. Dengan menerapkan *GridSearchcCV* untuk mendapatkan kombinasi parameter yang optimal dari beberapa parameter yang ditentukan. Kombinasi parameter optimal yang didapatkan adalah pada *depth* 6, *iterations* 1500, *l2_leaf_reg* 1 dan *learning_rate* 0,1 dengan hasil *mean_test_score cosine similarity* sebesar 0.958870. Nilai akurasi yang didapatkan dari model 1 mencapai 96.70% yang dibulatkan menjadi 97%. Nilai *precision* pada kategori baik mencapai 95%, kategori sedang 97%, dan kategori tidak baik 100%. Nilai *recall* pada kategori baik mencapai 95%, kategori sedang mencapai 98%, dan kategori tidak sehat mencapai 87%. Sedangkan nilai *f1-sscore* pada kategori baik mencapai 95%, kategori sedang mencapai 98%, dan kategori tidak baik mencapai 93%. Dapat disimpulkan bahwa pada penelitian ini metode *CatBoost* dengan menerapkan *GridSearchCV* dapat meningkatkan nilai

akurasi dalam proses prediksi indeks kualitas udara pada perbandingan 90% untuk data *training* dan 10% untuk data *testing* dengan kategori sangat baik. Sehingga pengaruh dari pembagian data pada Model 1 dengan data *training* 90% lebih banyak data untuk pelatihan dan dengan banyaknya data *training* sistem akan melatih data dengan maksimal, dan pada data *testing* 10% data lebih sedikit sehingga sistem dapat memprediksi dan menguji dengan maksimal pada data yang sedikit karena sistem telah belajar melatih data yang banyak dan pada prediksi pengujian data semakin sedikit. Tentu perlu adanya pengembangan untuk penelitian lebih lanjut tentang penggunaan metode *CatBoost* ataupun dalam penggunaan *GridSearchCV* pada model prediksi ataupun klasifikasi sehingga dapat meningkatkan nilai akurasi yang ada.

5.2 Saran

Peneliti sadar bahwa penelitain yang telah dilakukan masih jauh dari kata sempurna dan masih banyak kekurangan. Sehingga peneliti menyadari perlu adanya kritik dan saran untuk mengembangkan dan meningkatkan penelitian yang selanjutnya. Saran dari peneliti yang dapat diberikan untuk penelitian selanjutnya adalah:

1. Dapat menjelajahi metode klasifikasi untuk prediksi yang berbeda dengan penerapan *GridSearchCV* sehingga dapat membandingkan nilai akurasi yang didapatkan antara metode *CatBoost* dengan metode yang lain.
2. Dapat mencoba teknik lain untuk mencari kombinasi parameter yang optimal seperti *RandomSearchCV* untuk metode *CatBoost.*

3. Dapat mencoba kombinasi parameter lain selain *depth*, *iterations*, *learning_rate* dan *l2_leaf_reg* untuk meningkatkan akurasi pada metode *CatBoost*.
4. Dapat menjelajahi beberapa perbandingan dalam pembagian data *training* dan *testing*.

DAFTAR PUSTAKA

Agustin, E., Eviyanti, A., & Azizah, N. L. (2023). Deteksi Penyakit Epilepsi Melalui Sinyal EEG Menggunakan Metode DWT dan Extreme Gradient Boosting. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *7*(1), 117–127. https://doi.org/10.30865/mib.v7i1.5412

Amalia, A., Zaidiah, A., & Isnainiyah, I. N. (2022). Prediksi Kualitas Udara Menggunakan Algoritma K- Nearest Neighbor. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *7*(2), 496–507. https://doi.org/10.33387/jiko.v4i2.2871

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

Bardan, F., Razali, S., & Sari, C. M. (2023). Pelestarian Lingkungan Dalam Bentuk Penghijauan di IAI Al-Aziziyah Samalanga Bireuen Aceh. *KHADEM: Jurnal Pengabdian Kepada Masyarakat*, *2*(1), 55–64.

Barua, S., Gavandi, D., Sangle, P., Shinde, L., & Ramteke, J. (2021). Swindle : Predicting the Probability of Loan Defaults using *CatBoost* Algorithm. *Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021) IEEE, Iccmc*, 1710–1715.

Bernadet, Listyarini, S., & Warlina, L. (2023). Pengaruh Kebijakan Pencemaran Udara Sektor Transportasi Terhadap Nilai Indeks Kualitas Udara (Iku) Di Dki Jakarta. *Jurnal Ilmiah Pendidikan Lingkungan Dan Pembangunan*, *24*(01), 1–13. https://doi.org/10.21009/plpb.v24i01.30798

Christian, J., Ernawati, I., & ... (2022). Implementasi Penggunaan Algoritma Categorical Boosting (*CatBoost*) Dengan Optimisasi Hiperparameter Dalam Memprediksi Pembatalan Pesanan Kamar Hotel. *Seminar Nasional Mahasiswa Ilmu Komputer Dan Aplikasinya (SENAMIKA)*, *8*(1), 641–651. https://conference.upnvj.ac.id/index.php/senamika/article/view/2230

Darmawan, E. M. Z., & Fauzan Dianta, A. (2023). Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM. *Teknologi: Jurnal Ilmiah Sistem Informasi*, *13*(1), 8–15. https://doi.org/10.26594/teknologi.v13i1.3098Tersediaonlinediwww.journal.unipdu.ac.idHalamanjurnaldiwww.journal.unipdu.ac.id/index.php/teknologi

Delima, R., Hosianna, M., Pebrianty, D., & Amalia, J. (2023). Credit Risk Analysis dengan Algoritma Extreme Gradient Boosting dan Adaptive Boosting. *Journal of Information System, Graphics, Hospitality and Technology*, *05*(01), 1–7.

Dewi, N. K. (2021). Deteksi Fake Follower Instagram menggunakan *CatBoost*

Classifer. In *UIN Syarif Hidayatullah Jakarta Fakultas Sains dan Teknologi*.

Diantika, S., Nalatissifa, H., Supriyadi, R., Maulidah, N., & Fauzi, A. (2023). Implementasi Multi-Class Gradient Boosting Untuk Mengklasifikasikan Jenis Hewan Pada Kebun Binatang. *ANTIVIRUS: Jurnal Ilmiah Teknik Informatika*, *17*(1), 32–40.

Fadlisyah, & Muhathir. (2023). Performance Evaluation Of Variations Boosting Algorithms For Classifying Formalin Fish From Photos. *JITE (Journal of Informatics and Telecommunication Engineering)*, *6*(2), 621–629.

Fitri Boy, A. (2020). Implementasi Data Mining Dalam Memprediksi Harga Crude Palm Oil (CPO) Pasar Domestik Menggunakan Algoritma Regresi Linier Berganda (Studi Kasus Dinas Perkebunan Provinsi Sumatera Utara). *Journal of Science and Social Research*, *4307*(2), 78–85. http://jurnal.goretanpena.com/index.php/JSSR

Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. *Hindawi Journal OfEnvironmental and Public Health Supervised*, *14*(02), 51–55. https://doi.org/10.26483/ijarcs.v14i2.6972

Hadits, K. (n.d.). *Hadits Muslim Nomor 4084*. Ilmu Islam Portal Belajar Agama Islam. Retrieved April 24, 2024, from https://ilmuislam.id/hadits/27960/hadits-muslim-nomor-4084

Hariyadi, D., Kusumaningrum, E., Sumarsono, S., Fazlurrahman, F., & Setiyadi, B. (2023). Analisis Kualitas Udara Berbasis Dashboard Menggunakan ELK Stack. *JIKO (Jurnal Informatika Dan Komputer)*, *7*(1), 46. https://doi.org/10.26798/jiko.v7i1.685

Hendra Di Kesuma, Apriadi, D., Juliansa, H., & Etriyanti, E. (2022). Implementasi Data Mining Prediksi Mahasiswa Baru Menggunakan Algoritma Regresi Linear Berganda. *Jurnal Ilmiah Binary STMIK Bina Nusantara Jaya Lubuklinggau*, *4*(2), 62–66. https://doi.org/10.52303/jb.v4i2.74

Hermawan, A., & Sela, E. I. (2019). SPKU: Sistem Prediksi Kualitas Udara (Studi Kasus: Dki Jakarta). *Program Studi Teknik Informatika, Fakultas Bisnis Dan Teknologi Informasi Universitas Teknologi Yogykarta*. http://eprints.uty.ac.id/3552/

Husen, O. O., Mukaddas, J., & Ishak, A. (2023). Analisis Karbonmonoksida (CO), OksidaNitrogen (NOx) dan Sulfurdioksida (SO2) pada Kualitas Lingkungan Udara Ambien Jalan Raya Kota Kendari. *Sang Pencerah: Jurnal Ilmiah Universitas Muhammadiyah Buton*, *9*(2), 411–418. https://doi.org/10.35326/pencerah.v9i2.3021

IQAir. (n.d.). *Rangking Kota Besar Berpolusi Langsung*. IQAir. Retrieved April 24, 2024, from https://www.iqair.com/id/world-air-quality-ranking

Ivanoti, V. I., P, M. H., Triyono, G., & Utami, D. P. (2023). Decision Support System For Predicting Employee Leave Using The Light Gradient Boosting Machine (Lightgbm) And K-Means. *Jurnal Teknik Informatika (JUTIF)*, *4*(3), 657–667.

Jayadi, B. V., Handhayani, T., & Lauro, M. D. (2023). Perbandingan KNN Dan SVM Untuk Klasifikasi Kualitas Udara Di Jakarta. *Jurnal Ilmu Komputer Dan Sistem Informasi*, 1–7.

Jusuf, H., Prasetya, E., & Igirisa, N. (2023). Analisis Risiko Kesehatan Lingkungan Pajanan Particulate Matter (Pm10) Dan Karbon Monoksida (Co) Pada Masyarakat Di Desa Buata Kecamatan Botupingge. *Jurnal Sulolipu : Media Komunikasi Sivitas Akademika Dan Masyarakat*, *23*(1), 187–198. https://doi.org/10.1002/0471740039.vec1866

Khumaidi, A., Raafi'udin, R., & Solihin, I. P. (2020). Pengujian Algoritma Long Short-Term Memory untuk Prediksi Kualitas Udara dan Suhu Kota Bandung. *Jurnal Telematika*, *15*(1), 13–18.

Kohsasih, K. L., & Situmorang, Z. (2022). Analisis Perbandingan Algoritma C4.5 dan Naïve Bayes Dalam Memprediksi Penyakit Cerebrovascular. *Jurnal Informatika*, *9*(1), 13–17. https://doi.org/10.31294/inf.v9i1.11931

Kriswantara, B., & Sadikin, R. (2022). Used Car Price Prediction with Random Forest Regressor Model. *Journal of Information Systems, Informatics and Computing Issue Period*, *6*(1), 40–49. https://doi.org/10.52362/jisicom.v6i1.752

Leni, D., Sumiati, R., Haris, & Ardiansyah. (2023). Perancangan Metode Machine Learning Berbasis Web Untuk Prediksi Sifat Mekanik Aluminium. *Jurnal Rekayasa Mesin*, *14*(2), 611–626. https://doi.org/10.21776/jrm.v14i2.1370

Maha, I. K., & Susilawati. (2023). Dampak pencemaran lingkungan terhadap kesehatan masyarakat pesisir. *Journal Of Health And Medical Research*, *3*(4), 315–322.

Maulana, A., Purnamasari, A. I., & Ali, I. (2024). Analisis Klasifikasi Indeks Kualitas Udara Kota Di Indonesia Menggunakan Metode K-Nearest Neighbor Dan Naïve Bayes. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), 215–222.

Nababan, A. A., Jannah, M., Aulina, M., & Andrian, D. (2023). Prediksi Kualitas Udara Menggunakan Xgboost Dengan Synthetic Minority Oversampling Technique (Smote) Berdasarkan Indeks Standar Pencemaran Udara (Ispu). *JTIK (Jurnal Teknik Informatika Kaputama)*, *7*(1), 214–219. https://doi.org/10.59697/jtik.v7i1.66

Nainggolan, S. P., & Sinaga, A. (2023). Comparative Analysis of Accuracy of Random Forest and Gradient Boosting Classifier Algorithm for Diabetes Classification. *Sebatik*, *27*(1), 97–102.

https://doi.org/10.46984/sebatik.v27i1.2157

NUOnline. (n.d.). *Ar-Rum Ayat 41 Terjemah dan Tafsirnya*. NU ONLINE. Retrieved April 24, 2024, from https://quran.nu.or.id/ar-rum/41

Nurkholifah, M., Jasmarizal, Umar, Y., & Rahmaddeni. (2023). Analisa Performa Algoritma Machine Learning Dalam Prediksi Penyakit Liver. *Jurnal Indonesia : Manajemen Informatika Dan Komunikasi*, *4*(1), 164–172.

Okprana, H., & Winanjaya, R. (2022). Analisis Pengaruh Komposisi Data Training dan Testing Terhadap Akurasi Algoritma Resilient Backpropagation (RProp). *Brahmana : Jurnal Penerapan Kecerdasan Buatan*, *4*(1), 89–95. https://tunasbangsa.ac.id/pkm/index.php/brahmana/article/view/138

Pratama, R., Herdiansyah, M. I., Syamsuar, D., & Syazili, A. (2023). Prediksi Customer Retention Perusahaan Asuransi Menggunakan Machine Learning. *Jurnal SISFOKOM (Sistem Informasi Dan Komputer)*, *12*(1), 96–104.

Priya, B. C. (2021). *Cross-Validation and Hyperparameter Search in Scikit-Learn - A Complete Guide*. Dev Community. https://dev.to/balapriya/cross-validation-and-hyperparameter-search-in-scikit-learn-a-complete-guide-5ed8

Putra, A. E., & Rismawan, T. (2023). Klasifikasi Kualitas Udara Berdasarkan Indeks Standar Pencemaran Udara (ISPU) Menggunakan Metode Fuzzy Tsukamoto. *Coding : Jurnal Komputer Dan Aplikasi*, *11*(02), 190–196.

Putri, L. A. (2023). Implementasi Metode Artificial Neural Network (ANN) Algoritma Backpropagation untuk Klasifikasi Kualitas Udara di Provinsi DKI Jakarta Tahun 2021. *Statistics*, *3*(2), 184–191.

Rahmi, I. A., Afendi, F. M., & Kurnia, A. (2023). Metode AdaBoost dan Random Forest untuk Prediksi Peserta JKN-KIS yang Menunggak. *JAMBURA JOURNAL OF MATHEMATICS*, *5*(1), 83–94.

Ridho, I. I., & Mahalisa, G. (2023). Analisis Klasifikasi Dataset Indeks Standar Pencemaran Udara (ISPU) Di Masa Pandemi Menggunakan Algoritma Support Vector Machine (SVM). *Technologia*, *14*(1), 38–41.

Saadah, S., & Salsabila, H. (2021). Prediksi Harga Bitcoin Menggunakan Metode Random Forest. *Jurnal Politeknik Caltex Riau*, *7*(1), 24–32.

Saputra, E. P., Nurajizah, S., Maulidah, M., Hidayati, N., & Rachman, T. (2023). Komparasi Machine Learning Berbasis Pso Untuk Prediksi Tingkat Keberhasilan Belajar Berbasis E-Learning Comparation Of Pso-Based Learning Machine For E-Learning-Based. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *10*(2), 321–328. https://doi.org/10.25126/jtiik.2023106469

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

Shinta Enggar Maharan. (2021). Dampak Buruk Polusiudara Bagikesehatandancarameminimalkanrisikonya. *I Wayan Redi Aryanta, Shinta Enggar Maharani*, *3*, 1–12.

Sholeh, M., Nurnawati, E. K., & Lestari, U. (2023). Penerapan Data Mining dengan Metode Regresi Linear untuk Memprediksi Data Nilai Hasil Ujian Menggunakan RapidMiner. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *8*(1), 10–21. https://doi.org/10.14421/jiska.2023.8.1.10-21

Siregar, A. M., Faisal, S., Cahyana, Y., & Priyatna, B. (2020). Perbandingan Algoritma Klasifikasi Untuk Prediksi Cuaca. *Accounting Information System*, 15–24.

Subarkah, P., Marcos, H., Arsi, P., Prediksi, K., & Nasabah, A. (2019). Perbandingan Algoritme CART dan Naive Bayes Untuk Prediksi Kelayakan Nasabah BMT Khonsa Cilacap. *CITISEE*, *1*(3), 111–116.

Syandika, N. D., & Yustanti, W. (2023). Deteksi Anomali Terhadap Pembatalan Transaksi Pada Platform Tiktok Shop dengan Algoritma Categorical Boosting (*CatBoost*). *JINACS (Journal of Informatics and Computer Science)*, *05*(2), 149–156.

TafsirWeb. (n.d.-a). *Surat Al-An'am ayat 98 Arab, Latin, Terjemah Dan Tafsir*. TafsirWeb. Retrieved November 4, 2023, from https://tafsirweb.com/2222-surat-al-anam-ayat-98.html

TafsirWeb. (n.d.-b). *Surat Al-Jatsiyah ayat 5 Arab, Latin, Terjemah Dan Tafsir*. TafsirWeb. Retrieved November 4, 2023, from https://tafsirweb.com/9497-surat-al-jatsiyah-ayat-5.html

TafsirWeb. (n.d.-c). *Tafsir Berharga Terkait Dengan Surat Al-A'raf Ayat 56*. TafsirWeb. Retrieved June 5, 2024, from https://tafsirweb.com/2510-surat-al-araf-ayat-56.html

Urrochman, M. Y., Setyati, E., & Kristian, Y. (2023). Prediksi Timing Financial Distress Pada Bank Perkreditan Rakyat di Indonesia Menggunakan Machine Learning. *Jutisi: Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi*, *12*(2), 576–584.

Wahyudiyanta, S. A., & Supriyati. (2024). Analisis Kualitas Udara Jakarta dan Prediksi Tingkat Polusi dengan Metode Mesin Pembelajaran SVM Analysis of Jakarta ' s Air Quality and Prediction of Pollution Levels using Support Vector. *Prosiding SAINTEK*, *3*(1), 278–284.

Wang, J., Wang, Z., Li, J., & Peng, Y. (2023). An Interpretable Depression Prediction Model for the Elderly Based on ISSA Optimized LightGBM. *Journal of Beijing Institute of Technology*, *32*(2), 168–180. https://doi.org/10.15918/j.jbit1004-0579.2023.010

Wardhana, R. ., Wang, G., & Sibuea, F. (2023). Penerapan Machine Learning Dalam Prediksi Tingkat Kasus Penyakit Di Indonesia. *Journal of Information*

System Management (JOISM), *5*(1), 40–45.

Widiawati, F., Kurniawan, R., & Suprapti, T. (2023). Klasifikasi Data Tingkat Kualitas Udara Di Tangerang Selatan Menggunakan Algoritma Naive Bayes. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(6), 3739–3745. https://doi.org/10.36040/jati.v7i6.8261

Wiranata, A. D., Soleman, Irwansyah, Sudaryana, I. K., & Rizal. (2023). Klasifikasi Data Mining Untuk Menentukan Kualitas Udara Di Provinsi Dki Jakarta Menggunakan Algoritma K-Nearest Neighbors (K-NN). *Infotech: Journal of Technology Information*, *9*(1), 95–100.

Yudiskara, I. M. N., Dwidasmara, I. B. G., & Widiartha, I. M. (2023). Prediksi Polusi Udara Kota Jakarta Menggunakan Recurrent Neural Network-Gated Recurrent Units. *JURNAL PENGABDIAN INFORMATIKA*, *1*(3), 807–814.