

Model Prediksi AQI Index Sebagai Solusi Kebijakan Smart City di Wilayah Asia Timur dan Asia Tenggara

Siti Nur Salamah

Departemen Matematika, Fakultas
Matematika dan Ilmu Pengetahuan
Alam, Universitas Indonesia
Depok, Jawa Barat
siti.nur26@ui.ac.id

Maryesta Apriliani Sihombing

Departemen Matematika, Fakultas
Matematika dan Ilmu Pengetahuan
Alam, Universitas Indonesia
Depok, Jawa Barat
maryesta.apriliani@ui.ac.id

Raissa Anggia Maharani

Departemen Matematika, Fakultas
Matematika dan Ilmu Pengetahuan
Alam, Universitas Indonesia
Depok, Jawa Barat
raissa.anggia@ui.ac.id

Abstrak — Polusi udara di Asia Tenggara dan Asia Timur terus menjadi perhatian serius karena dampaknya yang signifikan terhadap kesehatan masyarakat dan lingkungan. Kedua kawasan ini dipilih sebagai fokus penelitian karena memiliki karakteristik unik. Asia Tenggara menghadapi tantangan dari urbanisasi yang pesat dengan standar industri yang beragam, sedangkan Asia Timur mengalami dampak industrialisasi yang masif. Penelitian ini mengembangkan model prediksi Indeks Kualitas Udara (AQI) dengan memanfaatkan data energi, lingkungan, dan polutan untuk mendukung pengambilan keputusan berbasis data. Metode Ridge Regression, dengan parameter tuning solver 'auto' dan nilai α 0.01, dipilih sebagai model terbaik setelah evaluasi komprehensif. Hasil model Ridge Regression menunjukkan tingkat akurasi yang sangat tinggi, dengan koefisien determinasi (R^2) mencapai 1 untuk kedua wilayah yang mengindikasikan kemampuan model untuk menghasilkan prediksi yang akurat dan stabil serta menangkap hubungan linier secara sempurna. Hasil penelitian menunjukkan bahwa $PM_{2.5}$ dan PM_{10} merupakan faktor utama yang memengaruhi AQI di kedua kawasan, dengan pengaruh vegetasi perkotaan juga signifikan dalam mengurangi nilai AQI. Simulasi kebijakan berbasis data yang dilakukan menunjukkan potensi pengurangan polusi melalui strategi kebijakan lingkungan yang terarah. Hasil penelitian ini diharapkan dapat mendukung upaya peningkatan kualitas udara, pengurangan dampak polusi, dan kesehatan masyarakat secara berkelanjutan.

Kata Kunci—kesehatan masyarakat, *machine learning*, polusi udara, metrik *error*, strategi lingkungan.

I. PENDAHULUAN

A. Latar Belakang

Menurut World Health Organization (WHO), sebanyak 7 juta orang terancam kesehatannya akibat polusi udara [1]. Dengan pesatnya perkembangan dan pertumbuhan populasi di kota-kota metropolitan, berbagai masalah lingkungan semakin menjadi perhatian utama.

Polusi udara memiliki dampak buruk pada kesehatan manusia. Sistem pemantauan kualitas udara menjadi langkah penting untuk mengurangi dampak polusi udara terhadap masyarakat. Air Quality Index (AQI) adalah sistem yang digunakan untuk menilai dan mengkomunikasikan tingkat polusi udara berdasarkan beberapa parameter, termasuk $PM_{2.5}$, PM_{10} , karbon monoksida (CO), sulfur dioksida (SO_2), nitrogen dioksida (NO_2), dan ozon (O_3). Sistem ini

memberikan informasi kepada masyarakat mengenai kualitas udara di sekitar mereka sehingga dapat meningkatkan kesadaran akan potensi bahaya kesehatan, terutama bagi kelompok rentan seperti anak-anak, lansia, serta mereka yang memiliki penyakit kardiovaskular atau gangguan pernapasan [2].

Kondisi polusi udara di Asia Tenggara dan Asia Timur menunjukkan tantangan serius bagi kesehatan masyarakat dan lingkungan. Di Asia Tenggara, laporan IQAir mengungkapkan bahwa hanya 2,7% dari 296 kota yang memiliki kualitas udara yang memenuhi standar kesehatan internasional untuk $PM_{2.5}$. Indonesia mencatatkan tingkat polusi tertinggi di kawasan ini, dengan Jakarta dan Surabaya termasuk dalam daftar kota dengan konsentrasi $PM_{2.5}$ tertinggi di dunia, mencapai $43.8 \mu\text{g}/\text{m}^3$, yang jauh melebihi batas aman yang ditetapkan oleh World Health Organization (WHO) sebesar $5 \mu\text{g}/\text{m}^3$ [3][4]. Sementara itu, di Asia Timur, negara-negara seperti China menghadapi masalah serupa dengan tingkat polusi yang sangat tinggi, terutama di kota-kota besar seperti Beijing dan Shanghai, di mana emisi industri dan kendaraan bermotor menjadi penyebab utama pencemaran udara [5].

Asia Timur dan Asia Tenggara menjadi fokus penting untuk penelitian karena perbedaan dalam industrialisasi, pola konsumsi energi, dan kebijakan lingkungan. Asia Timur, khususnya China, telah mengalami industrialisasi yang cepat dan peningkatan urbanisasi yang signifikan, menyebabkan peningkatan emisi polutan [6]. Di sisi lain, Asia Tenggara, meskipun kurang terindustrialisasi dibandingkan Asia Timur, juga menghadapi tantangan serupa akibat pertumbuhan populasi dan urbanisasi yang pesat [7]. Dalam konteks ini, pentingnya model prediksi Air Quality Index (AQI) menjadi semakin jelas sebagai alat untuk membantu pengambil kebijakan merancang kebijakan berbasis data guna mengatasi polusi udara yang semakin meningkat.

Model prediksi Air Quality Index (AQI) mendukung pengambil kebijakan dalam merumuskan strategi efektif untuk mengatasi polusi udara. Dengan konsep *smart cities*, pendekatan *data-driven* memungkinkan analisis pola polusi untuk pembatasan kendaraan, pengaturan emisi, dan pemantauan kualitas udara. Model ini dapat membantu meningkatkan kesehatan masyarakat di Asia Timur dan Tenggara serta mengurangi dampak polusi udara.

B. Rumusan Masalah

Berdasarkan latar belakang tersebut, didapatkan rumusan masalah sebagai berikut.

1. Bagaimana hubungan faktor energi dan lingkungan terhadap Air Quality Index (AQI) di Asia Tenggara dan Asia Timur?
2. Bagaimana model prediksi Air Quality Index (AQI) terbaik berdasarkan data energi dan lingkungan di masing-masing wilayah?
3. Faktor mana yang paling signifikan dalam mempengaruhi Air Quality Index (AQI) di kedua wilayah?
4. Bagaimana hasil prediksi Air Quality Index (AQI) dapat digunakan untuk menyusun simulasi kebijakan lingkungan yang efektif di wilayah Asia Tenggara dan Asia Timur?

C. Tujuan Penelitian

Dengan rumusan masalah tersebut, maka tujuan dari penelitian ini yaitu sebagai berikut.

1. Menganalisis hubungan antara faktor energi dan lingkungan terhadap Air Quality Index (AQI) di Asia Tenggara dan Asia Timur.
2. Mengidentifikasi dan menganalisis model prediksi terbaik untuk memprediksi Air Quality Index (AQI) berdasarkan data energi dan lingkungan di masing-masing wilayah.
3. Mengidentifikasi faktor-faktor yang paling signifikan dalam memengaruhi Air Quality Index (AQI) di kedua wilayah.
4. Menggunakan hasil prediksi Air Quality Index (AQI) untuk menyusun simulasi kebijakan lingkungan yang efektif di wilayah Asia Tenggara dan Asia Timur.

D. Manfaat Penelitian

Dengan penelitian ini, penulis berharap dapat memberikan manfaat sebagai berikut.

1. Memberikan wawasan mengenai hubungan antara faktor energi dan lingkungan terhadap Air Quality Index (AQI) di Asia Tenggara dan Asia Timur sehingga dapat menjadi acuan dalam penelitian terkait kualitas udara.
2. Menyediakan referensi mengenai model prediksi terbaik untuk memprediksi Air Quality Index (AQI) berdasarkan data energi dan lingkungan di masing-masing wilayah, yang dapat dimanfaatkan oleh peneliti maupun praktisi.
3. Mengidentifikasi faktor-faktor signifikan yang memengaruhi Air Quality Index (AQI) di kedua wilayah yang dapat digunakan sebagai dasar dalam merancang kebijakan lingkungan yang lebih efektif.
4. Menghasilkan simulasi kebijakan lingkungan yang efektif berdasarkan hasil prediksi Air Quality Index (AQI) yang dapat digunakan untuk mendukung pengambilan keputusan dalam upaya meningkatkan kualitas udara di Asia Tenggara dan Asia Timur.

II. PEMBAHASAN

A. Landasan Teori

1. Pengertian Air Quality Index (AQI)

Air Quality Index (AQI) adalah indikator kualitas udara yang menggambarkan dan mengevaluasi status kualitas udara, dengan menyederhanakan konsentrasi beberapa polutan menjadi satu bentuk numerik. AQI dihitung berdasarkan standar kualitas udara baru (GB3095-2012), yang mencakup enam polutan, yaitu sulfur dioksida (SO_2), nitrogen dioksida (NO_2), $\text{PM}_{2.5}$, PM_{10} , ozon (O_3), dan karbon monoksida (CO) [8]. Bagi masyarakat umum, Air Quality Index (AQI) adalah indeks penting untuk memahami apakah kualitas udara buruk atau baik, serta membantu dalam interpretasi data untuk pengambilan keputusan terkait tindakan pengurangan polusi dan pengelolaan kualitas udara.

Nilai Air Quality Index (AQI) dihitung dari konsentrasi polutan dengan rumus berikut [9].

$$I = \frac{I_{tinggi} - I_{rendah}}{C_{tinggi} - C_{rendah}} (C - I_{rendah}) + I_{rendah} \quad (1)$$

Dengan,

I = Air Quality Index.

C = Konsentrasi polutan.

C_{rendah} = Titik pemutusan konsentrasi yang lebih kecil dari C .

C_{tinggi} = Titik pemutusan konsentrasi yang lebih besar atau sama dengan C .

I_{rendah} = Titik pemutusan indeks yang sesuai dengan C_{rendah} .

I_{tinggi} = Titik pemutusan indeks yang sesuai dengan C_{tinggi} .

Air Quality Index individu dihitung untuk setiap konsentrasi polutan terpisah, dan nilai tertinggi dari semua nilai tersebut menentukan klasifikasi AQI lokasi pada waktu tertentu. Partikel materi, sulfur dioksida, ozon permukaan tanah, nitrogen dioksida, dan karbon monoksida adalah kontributor penting dalam perhitungan AQI.

2. Smart cities

Smart cities didefinisikan sebagai kota inovatif yang menggunakan teknologi informasi dan komunikasi (TIK) untuk meningkatkan kualitas hidup, efisiensi operasi dan layanan perkotaan, serta daya saing, dengan memastikan bahwa kota tersebut memenuhi kebutuhan generasi sekarang dan masa depan dalam aspek ekonomi, sosial, dan lingkungan [10].

Dalam konteks ini, pentingnya Air Quality Index (AQI) menjadi sangat relevan karena AQI berfungsi sebagai indikator kualitas udara yang dapat mempengaruhi kesehatan masyarakat. Dengan memanfaatkan *Artificial Intelligence* untuk memantau AQI, *smart cities* dapat mengambil keputusan yang lebih baik dalam pengelolaan transportasi dan industri, sehingga meningkatkan efisiensi operasional dan merancang kebijakan lingkungan yang lebih baik. Selain itu, kota dengan kualitas udara yang baik lebih menarik bagi

investasi dan pengembangan bisnis sehingga meningkatkan daya saing ekonomi [11].

3. Data Science dan Machine Learning

Data science dan machine learning saling melengkapi dalam mendukung pengembangan *smart cities*. Data science berfokus pada studi data kota untuk mengekstraksi wawasan atau pengetahuan yang berguna untuk pengambilan keputusan cerdas di berbagai aplikasi dunia nyata [12]. Sementara itu, machine learning mengotomatisasi pembuatan model analitik dengan belajar dari data kota yang dikumpulkan yang memungkinkan analisis yang lebih efisien dan akurat [13]. Dengan menggabungkan kedua konsep ini, layanan berbasis data dapat dikembangkan untuk mendukung solusi inovatif pada penerapan *smart cities*.

4. Ridge Regression

Ridge Regression adalah salah satu metode regresi linier yang digunakan untuk mengatasi masalah multikolinearitas pada data [14]. Multikolinearitas terjadi ketika terdapat korelasi yang kuat antara dua atau lebih variabel independen dalam model regresi. Hal ini dapat menyebabkan hasil yang tidak akurat dan tidak stabil dalam memprediksi variabel dependen. Ridge Regression memperkenalkan penalti pada koefisien regresi untuk mengurangi efek multikolinearitas dan meningkatkan stabilitas model [15].

Pada Ridge Regression, digunakan regularisasi L2 untuk menambahkan penalti pada koefisien regresi. Regularisasi L2 mengurangi nilai koefisien regresi yang besar dan mempertahankan koefisien yang kecil sehingga mengurangi efek multikolinearitas dan meningkatkan stabilitas model. Ridge Regression juga dapat digunakan untuk mengatasi *overfitting* pada model regresi [16]. Berikut merupakan rumus dari Ridge Regression [16].

$$(X^T X + \alpha I) \beta = X^T y \quad (2)$$

Dengan,

X = Matriks fitur (*input*).

y = Vektor target (*output*).

$X^T X$ = Matriks kovarians dari data fitur.

α = Parameter regularisasi.

β = Vektor koefisien regresi yang akan diestimasi.

I = Matriks identitas dengan dimensi yang sama seperti $X^T X$.

5. Random Forest Regression

Random Forest Regression adalah sebuah metode pembelajaran ensemble yang menggunakan kombinasi dari beberapa pohon keputusan (*decision trees*) untuk melakukan tugas regresi. Dalam pendekatan ini, setiap pohon keputusan dilatih menggunakan subset acak dari data pelatihan dengan metode *bootstrapping*, di mana subset fitur acak dipertimbangkan pada setiap pemisahan node [17].

6. Gradient Boosting Regressor

Gradient Boosting Regressor adalah teknik *machine learning* yang efektif untuk menyelesaikan masalah regresi dengan menggabungkan beberapa model prediksi lemah menjadi satu model yang lebih kuat. Algoritma ini bekerja secara iteratif dengan menambahkan *weak learners*, umumnya berupa *decision trees*, yang bertujuan untuk

memperbaiki kesalahan dari model sebelumnya. Proses dimulai dengan inisialisasi model menggunakan rata-rata dari nilai target sebagai prediksi awal. Selanjutnya, *residual error* dihitung sebagai selisih antara nilai aktual dan prediksi, dan model baru dibangun untuk memprediksi residual ini. Setiap model yang ditambahkan berfokus pada kesalahan yang tersisa dari model sebelumnya, sehingga meningkatkan akurasi keseluruhan [18].

7. XGBoost Regressor

Boosting adalah metode yang mengubah beberapa classifier yang tidak efektif menjadi satu classifier yang sangat efektif. Teknik XGBoost dikembangkan berdasarkan dasar dari *gradient boosting* [19]. Varian XGBoost dari *gradient boosting* memiliki performa yang lebih baik dibandingkan dengan versi aslinya dalam hal kemampuan generalisasi, skalabilitas, dan efisiensi pemrosesan [20]. Saat menggunakan XGBoost, pengorganisasian data sangat penting. Karena XGBoost hanya menerima vektor numerik sebagai input, semua data kategori akan diubah menjadi nilai numerik yang sesuai.

8. Support Vector Machine (SVM)

Algoritma SVM adalah teknik *machine learning* yang bekerja dengan memanfaatkan garis, bidang, atau hyperplane untuk memisahkan data ke dalam dua atau lebih dimensi. Pendekatan ini bertujuan menemukan garis pemisah terbaik yang mampu memaksimalkan jarak antara kelompok data dari kelas yang berbeda, sehingga dapat memperjelas perbedaan antar kelas. Dalam prosesnya, SVM mempertimbangkan tiga parameter utama: gamma, C, dan jenis fungsi. Jenis fungsi yang digunakan dapat berupa linear, non-linear, polinomial, atau berbasis radial, tergantung pada sifat data *input*. Parameter gamma dan C berperan penting dalam menghindari model dari masalah *overfitting* maupun *underfitting* [21].

9. Linear Regression

Hubungan antara variabel respons dan variabel penjelas ditentukan melalui analisis regresi linear, yang banyak digunakan dalam aplikasi statistik dan memerlukan proses pelatihan untuk menghitung nilai dari banyak koefisien:

$$z_i = b_0 + b_{1x_1} + b_{2x_2} + \dots + b_{ix_i} \quad (3)$$

Di mana koefisien regresi b_i mewakili kekuatan dan arah hubungan antara variabel independen x_i dan variabel dependen z_i . Semakin besar nilai koefisien regresi, semakin besar pengaruh variabel independen terhadap prediksi nilai variabel dependen. Koefisien ini mencerminkan sejauh mana perubahan dalam variabel independen dapat memengaruhi nilai variabel dependen [22].

10. CatBoost

CatBoost adalah pengembangan dari framework *decision tree* dan *gradient boosting* [23]. Boosting didasarkan pada hipotesis bahwa beberapa model lemah yang relatif sederhana dapat digabungkan untuk menghasilkan model prediksi yang sangat kompetitif dan secara marginal melampaui peluang acak [24]. Dengan menyesuaikan rangkaian *decision tree*, di mana masing-masing pohon

belajar dari kesalahan iterasi sebelumnya, *gradient boosting* mampu mengurangi error. Proses ini terus diulang dengan menambahkan fungsi baru ke dalam kombinasi hingga fungsi *loss* yang dipilih tidak lagi diminimalkan.

Berbeda dari model *gradient boosting* konvensional, CatBoost menggunakan proses yang berbeda dalam membangun *decision tree*. Dengan mengharuskan semua node pada level yang sama untuk menguji prediktor yang sama di bawah kondisi yang sama, CatBoost menghasilkan "*oblivious trees*." Hal ini memungkinkan indeks sebuah *leaf* dapat dihitung hanya dengan operasi *bitwise* [25].

11. LightGBM

LightGBM adalah *framework* peningkatan gradien berbasis GBDT yang dirancang untuk menggabungkan M *decision tree* regresi yang lemah menjadi satu model regresi yang kuat secara bertahap [26]. Inovasi yang diterapkan dalam LightGBM membantu mengurangi penggunaan memori dan meningkatkan kecepatan prediksi. *Framework* ini juga menggunakan metode histogram dan algoritma *leaf-wise* dengan batasan yang dirancang secara cermat. Optimisasi ini meningkatkan efisiensi komputasi serta pengelolaan memori. Metode histogram bekerja dengan membagi nilai kontinu menjadi M interval, membangun histogram berdasarkan interval tersebut, lalu memproses data hingga terbentuk *decision tree* yang lengkap [27].

12. Lasso

Lasso menerapkan regularisasi di mana sebagian koefisien regresi dikurangi hingga menjadi nol. Dalam proses seleksi fitur, semua koefisien dengan nilai yang tidak nol dipilih, sementara kesalahan prediksi diminimalkan. Jika nilai parameter regularisasi ini sangat tinggi, maka koefisien variabel regresi dapat menjadi nol. Teknik regresi ini banyak digunakan karena mampu memberikan akurasi prediksi yang baik sekaligus mengurangi *overfitting* [28].

13. ElasticNet

ElasticNet adalah metode regularisasi dan seleksi variabel baru yang sering kali menunjukkan kinerja lebih baik daripada lasso dengan tingkat kepadatan representasi yang serupa. Selain itu, ElasticNet mendorong efek pengelompokan, di mana prediktor yang saling berkorelasi cenderung masuk atau keluar dari model secara bersamaan. ElasticNet sangat berguna ketika jumlah prediktor (p) jauh lebih besar daripada jumlah observasi (n), sedangkan lasso tidak cukup efektif sebagai metode seleksi variabel dalam kasus $p \gg n$ [29].

14. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) MAE adalah metrik yang mengukur rata-rata selisih absolut antara nilai peramalan dan nilai aktual, dirumuskan dengan:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

dengan n adalah jumlah observasi, y_i adalah nilai aktual pada observasi ke- i , dan \hat{y}_i adalah nilai prediksi pada observasi ke- i . Semakin kecil nilai MAE, maka semakin baik kinerja model prediksi *machine learning* [30].

15. Root Mean Squared Error (RMSE)

RMSE mengukur akar kuadrat dari rata-rata selisih kuadrat antara nilai peramalan dan nilai aktual, dirumuskan dengan:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

dengan n adalah jumlah observasi, y_i adalah nilai aktual pada observasi ke- i , dan \hat{y}_i adalah nilai prediksi pada observasi ke- i [30].

16. Mean Absolute Percentage Error (MAPE)

MAPE mengukur rata-rata persentase kesalahan peramalan dengan membandingkan selisih absolut antara nilai peramalan dan aktual dibagi dengan nilai aktual kemudian dikalikan 100 untuk mendapatkan persentase, dengan rumus:

$$MAPE = \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \quad (6)$$

dengan n adalah jumlah observasi, y_i adalah nilai aktual pada observasi ke- i , dan \hat{y}_i adalah nilai prediksi pada observasi ke- i . MAPE sering digunakan untuk mengevaluasi seberapa baik model dalam meramalkan nilai dalam konteks persentase. Nilai MAPE yang lebih rendah menunjukkan model yang lebih akurat [30].

17. Mean Squared Error (MSE)

MSE adalah rata-rata perbedaan yang dikuadratkan diantara nilai yang diramalkan dengan yang diamati. Rumusnya adalah sebagai berikut:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (7)$$

dengan n adalah jumlah observasi, y_i adalah nilai aktual pada observasi ke- i , dan \hat{y}_i adalah nilai prediksi pada observasi ke- i [30].

18. R-Squared

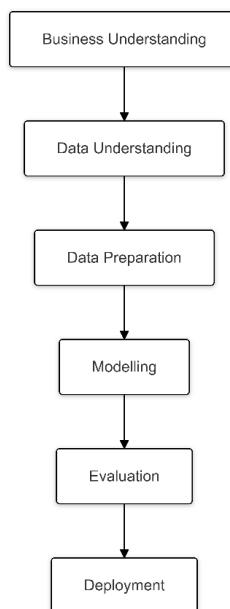
Metrik kinerja R-squared menunjukkan seberapa baik nilai prediksi cocok dengan nilai aktual. Rumusnya yaitu sebagai berikut.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

dengan y_i adalah nilai observasi aktual untuk data ke- i . \bar{y} merupakan nilai prediksi untuk data ke- i berdasarkan model yang digunakan. \bar{y} adalah rata-rata dari semua nilai y_i (nilai rata-rata dari data aktual) dan n merupakan jumlah total data yang digunakan dalam perhitungan [31].

B. Metode Penelitian

Metode penelitian yang diterapkan dalam penelitian ini adalah CRISP-DM (Cross-Industry Standard Process for Data Mining). Untuk mempermudah pemahaman dan memberikan gambaran yang jelas terkait proses yang dilalui, setiap tahapannya disajikan dalam bentuk diagram alur yang dapat dilihat pada Gambar 1.



Gambar 1. Diagram alur penelitian

1. Business Understanding

Business understanding pada penelitian ini berfokus pada pentingnya prediksi Air Quality Index (AQI) untuk mendukung pengambil kebijakan dalam merumuskan strategi berbasis data guna mengatasi polusi udara di Asia Tenggara dan Asia Timur. Penelitian ini bertujuan untuk menganalisis hubungan antara faktor energi dan lingkungan terhadap AQI, mengidentifikasi model prediksi terbaik, serta menentukan faktor yang paling signifikan memengaruhi AQI di kedua wilayah. Dengan pendekatan ini, penelitian ini diharapkan dapat memberikan dasar untuk simulasi kebijakan lingkungan yang efektif dalam meningkatkan kualitas udara dan kesehatan masyarakat di kawasan tersebut.

2. Data Understanding

Data penelitian diperoleh dari dataset lomba dengan nama “environmental_dataset” yang mencakup 9858 baris dan 21 kolom. Dataset ini berisi berbagai informasi lingkungan, energi, dan polutan yang relevan untuk mendukung prediksi Air Quality Index (AQI). Berikut adalah informasi mengenai dataset:

TABEL 1. INFORMASI DATASET

No	Kolom	Tipe Data	Keterangan
1	SensorID	Kategorik	Nomor ID
2	SensorLocation	Kategorik	Lokasi sensor
3	Pollutant_PM2.5_µg/m³	Numerik	Konsentrasi PM2.5
4	Pollutant_PM10_µg/m³	Numerik	Konsentrasi PM10
5	Pollutant_O3_ppb	Numerik	Konsentrasi Ozon
6	Pollutant_NO2_ppb	Numerik	Konsentrasi NO2
7	Pollutant_CO_ppm	Numerik	Konsentrasi CO
8	Pollutant_SO2_ppb	Numerik	Konsentrasi SO2
9	UrbanVegetationArea_m2	Numerik	Area vegetasi
10	Humidity_%	Numerik	Tingkat kelembaban
11	AirTemperature_C	Numerik	Suhu udara

12	EnergySavingTechnology	Kategorik	Teknologi hemat
13	AnnualEnergySavings_%	Numerik	Penghematan energi
14	PopulationDensity_people/km²	Numerik	Kepadatan populasi
15	RetrofitData	Kategorik	Data retrofit
16	RenewableEnergyPercentage_%	Numerik	Energi terbarukan
17	AnnualEnergyConsumption_kWh	Numerik	Konsumsi energi
18	GreenSpaceIndex_%	Numerik	Ruang hijau
19	HistoricPollutantLevels	Numerik	Polusi historis
20	Country	Numerik	Nama negara
21	AQI_Index	Numerik	Indeks AQI

Penelitian ini berfokus pada wilayah Asia Timur dan Asia Tenggara, sehingga data penelitian dibagi menjadi dua subset berdasarkan negara yang termasuk dalam masing-masing wilayah. Wilayah Asia Tenggara mencakup negara-negara berikut, yaitu Kamboja, Thailand, Filipina, Malaysia, Laos, Myanmar, Brunei, Singapura, Vietnam, Indonesia, dan Timor Leste. Sementara itu, wilayah Asia Timur meliputi negara-negara berikut, yaitu Jepang, Cina, Mongolia, dan Korea Selatan. Didapatkan datanya seperti berikut.

- Asia Tenggara: Dataset ini terdiri dari 3375 baris dan 21 kolom.
- Asia Timur: Dataset ini terdiri dari 1213 baris dan 21 kolom.

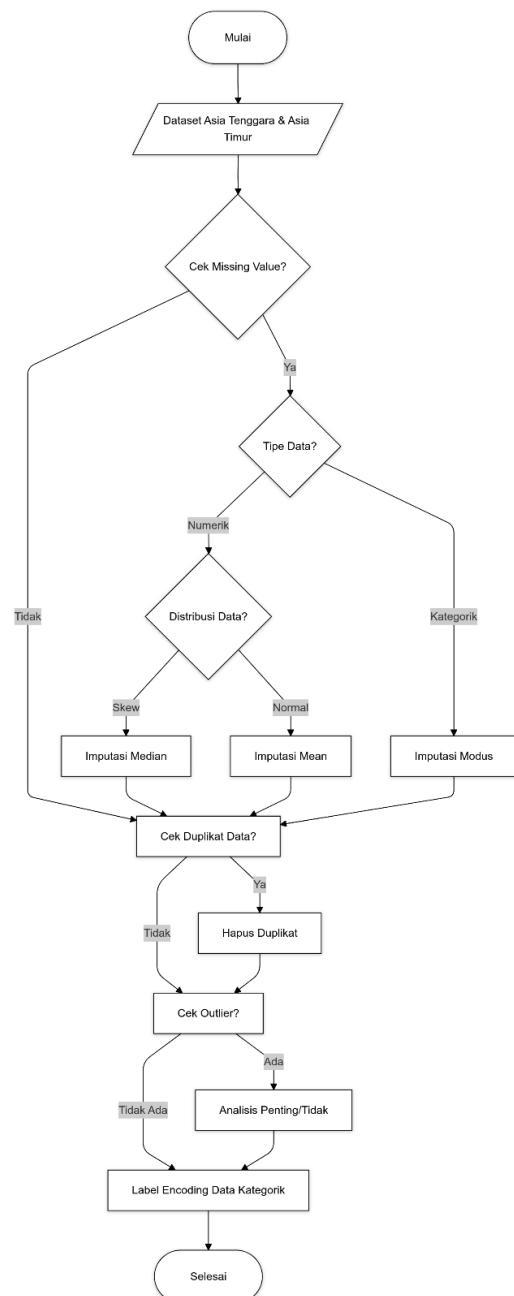
3. Data Preparation

Data preparation merupakan langkah penting dalam memastikan kualitas data yang akan digunakan untuk analisis. Pada tahap ini, data dari wilayah Asia Tenggara dan Asia Timur diperiksa secara terpisah untuk mengidentifikasi masalah seperti nilai yang hilang, data duplikat, dan *outlier*. Berikut merupakan diagram alur yang menggambarkan proses data *preparation* secara lebih terperinci.

TABEL 3. MISSING VALUE DATA ASIA TIMUR

Kolom	Tipe Data	Jumlah Missing Value
UrbanVegetationArea_m2	Numerik	303
EnergySavingTechnology	Kategorik	134
PopulationDensity_people/km ²	Numerik	133
RenewableEnergyPercentage %	Numerik	98

Sebelum menangani *missing value*, akan dicek terlebih dahulu distribusi kolom numerik yang memiliki *missing value*.



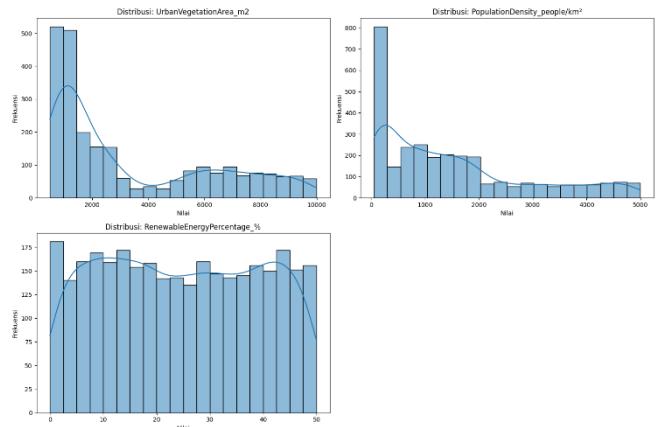
Gambar 2. Diagram alur tahapan Data Preparation

- *Missing Values*

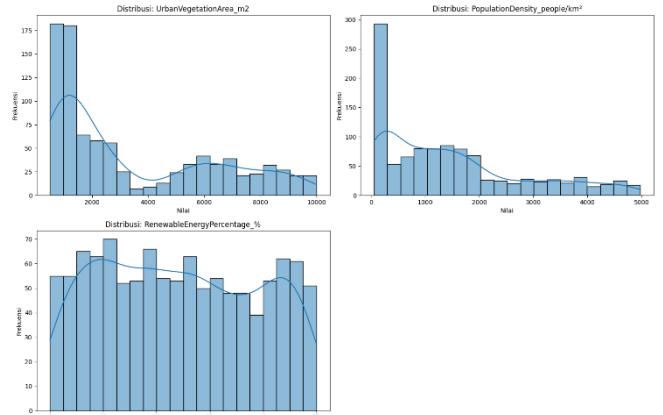
Untuk menangani nilai hilang (*missing values*) pada dataset Asia Tenggara dan Asia Timur, langkah awal yang dilakukan adalah mengidentifikasi fitur-fitur yang memiliki *missing values*. Tabel berikut menunjukkan jumlah nilai hilang pada masing-masing fitur berdasarkan tipe data di kedua wilayah tersebut.

TABEL 2. MISSING VALUE DATA ASIA TENGGARA

Kolom	Tipe Data	Jumlah Missing Value
UrbanVegetationArea_m2	Numerik	881
EnergySavingTechnology	Kategorik	367
PopulationDensity_people/km ²	Numerik	368
RenewableEnergyPercentage %	Numerik	282



Gambar 3. Distribusi missing value kolom numerik Asia Tenggara



Gambar 4. Distribusi missing value kolom numerik Asia Timur

Setelah memeriksa distribusi fitur numerik yang memiliki nilai hilang, diputuskan metode imputasi berdasarkan pola distribusinya. Kolom seperti "UrbanVegetationArea_m2" dan "PopulationDensity_people/km²" memiliki distribusi data yang tidak simetris (*skewed*), sehingga metode imputasi *median* digunakan. Hal ini dilakukan karena *median* lebih tahan terhadap pengaruh *outlier* dan mampu menjaga representasi data yang lebih stabil. Sementara itu, untuk kolom seperti "RenewableEnergyPercentage %" yang memiliki distribusi mendekati normal, metode imputasi *mean* dipilih agar tetap mencerminkan nilai rata-rata keseluruhan data.

Untuk fitur kategorik seperti "EnergySavingTechnology", imputasi dilakukan menggunakan *mode* berdasarkan konteks geografis, seperti kolom "Country" dan "SensorLocation". Jika *mode* tidak dapat ditentukan, nilai *default* "Unknown" digunakan. Pendekatan ini memastikan konsistensi nilai

kategorik dengan kondisi geografis yang relevan, sekaligus menghindari kehilangan informasi penting. Kombinasi metode ini diharapkan dapat meminimalkan distorsi data akibat imputasi.

- Duplikat Data

Pada tahap ini, data diperiksa untuk mendeteksi adanya duplikat yang dapat memengaruhi hasil analisis. Hasil pemeriksaan jumlah data duplikat dapat dilihat pada tabel berikut.

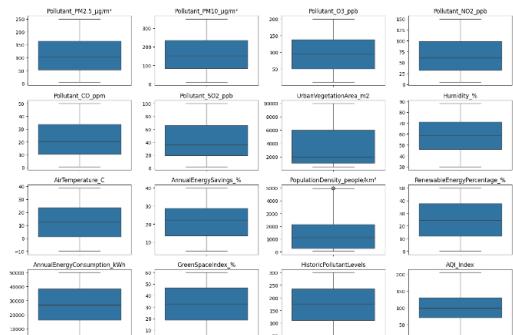
TABEL 4. JUMLAH DUPLIKAT DATA

Data	Jumlah Duplikat Data
Asia Tenggara	0
Asia Timur	0

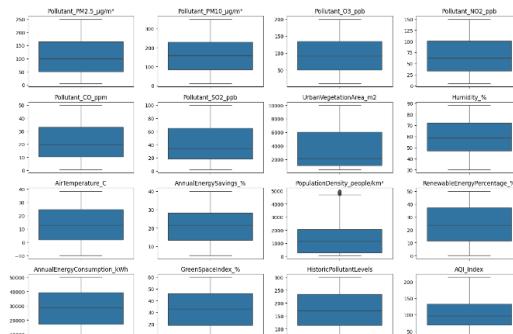
Berdasarkan tabel di atas, tidak ditemukan adanya data duplikat pada dataset Asia Tenggara maupun Asia Timur. Hal ini menunjukkan bahwa data dalam kedua wilayah tersebut bersih dari duplikasi, sehingga tidak memerlukan langkah penghapusan data duplikat.

- Outlier Data

Pada tahap ini, data akan diperiksa untuk mendeteksi adanya outlier. Berikut visualisasi *boxplot* yang digunakan untuk memberikan gambaran distribusi data dan keberadaan outlier pada masing-masing fitur numerik, seperti yang ditunjukkan pada gambar berikut.



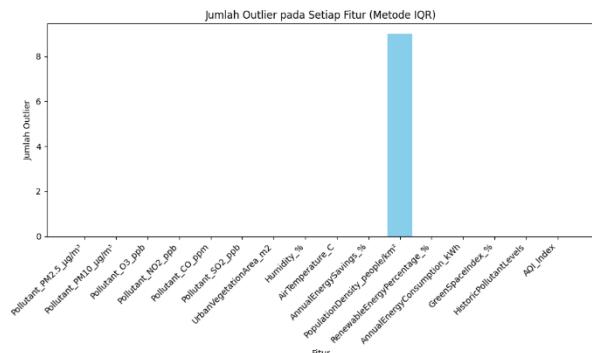
Gambar 5. Box plot data Asia Tenggara



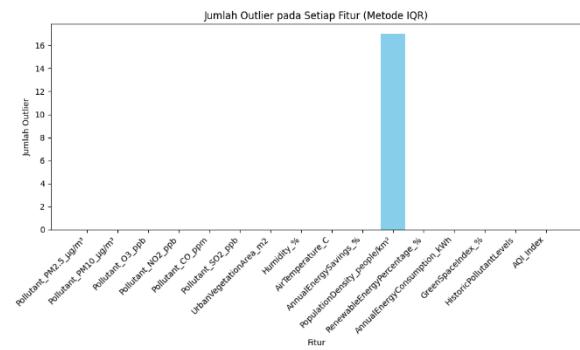
Gambar 6. Box plot data Asia Timur

Berdasarkan visualisasi *box plot* pada dataset Asia Tenggara dan Asia Timur, ditemukan adanya *outlier* pada fitur numerik, khususnya pada kolom "PopulationDensity_people/km²". Langkah selanjutnya adalah menghitung jumlah dan persentase *outlier* pada setiap

fitur numerik menggunakan metode Interquartile Range (IQR). Metode ini dipilih karena mampu mendeteksi nilai-nilai ekstrim di luar batas wajar berdasarkan distribusi data.



Gambar 7. Jumlah *outlier* data Asia Tenggara



Gambar 8. Jumlah *outlier* data Asia Timur

TABEL 5. PERSENTASE KOLOM OUTLIER

Data	Kolom Outlier	Persentase Outlier (%)
Asia Tenggara	PopulationDensity_people/km ²	0.266667
Asia Timur	PopulationDensity_people/km ²	1.401484

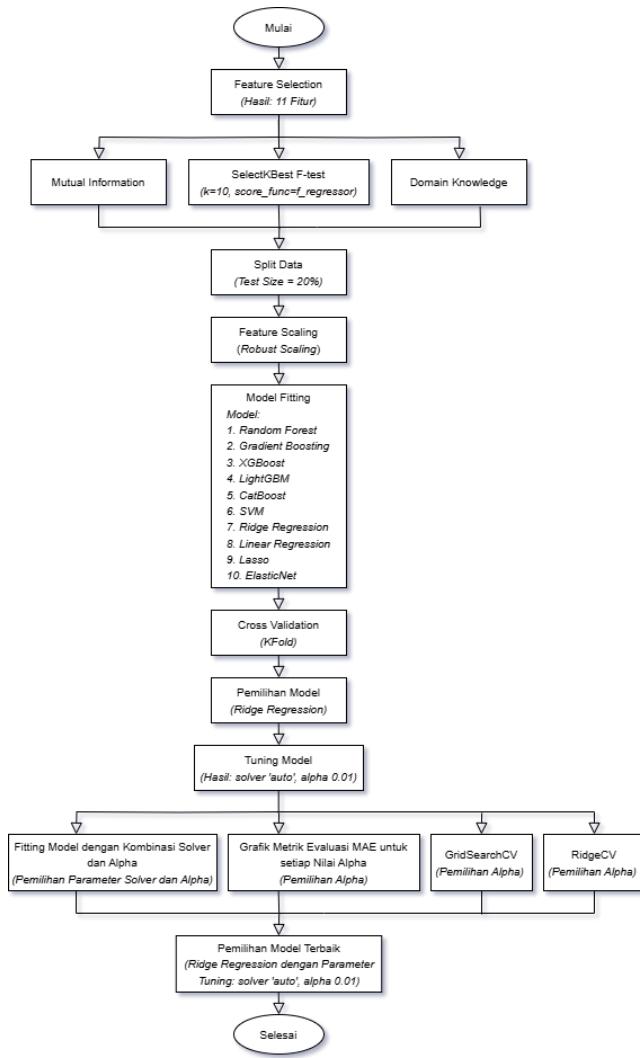
Berdasarkan hasil perhitungan, fitur "PopulationDensity_people/km²" pada dataset Asia Tenggara dan Asia Timur memiliki persentase *outlier* yang cukup rendah, yaitu 0.27% dan 1.40% berturut-turut. Mengingat jumlah *outlier* yang minimal, data tersebut diputuskan untuk tetap dipertahankan dalam analisis. Hal ini dilakukan karena *outlier* yang terdeteksi tidak menunjukkan kejanggalan signifikan dan tetap merepresentasikan karakteristik distribusi data.

- Label Encoding Data Kategorik

Pada tahap ini, fitur kategorik dalam dataset diubah menjadi bentuk numerik menggunakan dua teknik utama, yaitu label *encoding* dan *one-hot encoding*, agar dapat digunakan dalam analisis dan pemodelan. *One-hot encoding* diterapkan pada fitur seperti "SensorLocation" dan "EnergySavingTechnology" yang memiliki banyak kategori unik. Sementara itu, label *encoding* digunakan pada fitur seperti "RetrofitData" dan "Country", yang memiliki jumlah kategori lebih sedikit, dengan mengubah setiap kategori menjadi nilai *integer* unik.

4. Modelling & Evaluation

Pada langkah *Modelling & Evaluation*, diterapkan beberapa model regresi seperti untuk memprediksi Air Quality Index (AQI) di Asia Tenggara dan Asia Timur. Evaluasi dilakukan menggunakan metrik seperti MSE, RMSE, MAE, MAPE, dan R² untuk mengukur performa model. Selanjutnya, dilakukan Hyperparameter Tuning untuk meningkatkan akurasi dan performa model. Tujuan utama dari langkah ini adalah untuk menghasilkan model yang optimal dalam memprediksi AQI sehingga dapat mendukung kebijakan yang lebih baik mengenai kualitas udara. Berikut merupakan diagram alur yang menggambarkan proses *modelling & evaluation* secara lebih terperinci.



Gambar 9. Diagram alur tahapan *Modeling & Evaluation*

- *Feature Selection*

Digunakan tiga pertimbangan untuk feature selection, yakni dengan Mutual Information, SelectKBest ($k=10$, $score_func = f_regressor$), dan pertimbangan *domain knowledge*. Tabel di bawah menunjukkan ringkasan dari hasil Mutual Information dan SelectKBest ($score_func = f_regressor$).

TABEL 6. OUTPUT MUTUAL INFORMATION DAN SELECTKBEST (F-TEST)

Fitur	Mutual Information		Top 10 SelectKBest (F-Test)	
	Asia Tenggara	Asia Timur	Asia Tenggara	Asia Timur
Pollutant_PM2.5_μg/m³	0.5269	0.5998	✓	✓
Pollutant_PM10_μg/m³	0.2973	0.3209	✓	✓
SensorLocation	0.2457	0.2611	✓	✓
PopulationDensity_people/km²	0.1996	0.1808	✓	✓
UrbanVegetationArea_m²	0.1451	0.2168	✓	✓
Pollutant_NO2_ppb	0.0957	0.1045	✓	✓
Pollutant_SO2_ppb	0.0911	0.0397	✓	✓
Pollutant_O3_ppb	0.0825	0.0759	✓	✓
Humidity_%	0.0694	0.0677	✓	✓
Pollutant_CO_ppm	0.0543	0.0602	✓	✓
RenewableEnergyPercentage_%	0.0526	0.0302	✗	✗
AnnualEnergySavings_%	0.0384	0.0231	✗	✗
AirTemperature_C	0.0267	0.0168	✗	✗
Country	0.0241	0.0026	✗	✗
EnergySavingTechnology	0.0172	0.0000	✗	✗
GreenSpaceIndex_%	0.0086	0.0000	✗	✗
RetrofitData	0.0037	0.0000	✗	✗
AnnualEnergyConsumption_kWh	0.0000	0.0000	✗	✗
HistoricPollutantLevels	0.0000	0.0262	✗	✗

Berdasarkan hasil *Mutual Information* dan *F-Test* (melalui SelectKBest), 10 fitur teratas dipilih karena memiliki hubungan yang signifikan terhadap Air Quality Index (AQI). Perhatikan, pada dataset Asia Tenggara, fitur ‘RenewableEnergyPercentage_%’ menunjukkan nilai *Mutual Information* yang relevan (>0.05), meskipun tidak dipilih oleh *Top 10 Feature F-Test*. Oleh karena itu, fitur ini dimasukkan dalam model untuk memastikan bahwa aspek energi terbarukan tetap terwakili dalam analisis. Selain itu, fitur-fitur yang memiliki kontribusi rendah pada Mutual Information dikeluarkan dari model karena diduga tidak memberikan pengaruh signifikan terhadap prediksi AQI.

Berdasarkan hasil analisis di atas dan pertimbangan *domain knowledge*, 11 fitur berikut dipilih untuk digunakan dalam model prediksi AQI di Asia Tenggara dan Asia Timur: Pollutant_PM2.5_μg/m³, Pollutant_PM10_μg/m³, SensorLocation, PopulationDensity_people/km², UrbanVegetationArea_m², Pollutant_NO2_ppb, Pollutant_SO2_ppb, Pollutant_O3_ppb, Humidity %, Pollutant_CO_ppm, RenewableEnergyPercentage %

- *Model Fitting*

Data untuk prediksi Air Quality Index (AQI) di Asia Tenggara dan Asia Timur melalui beberapa tahapan penting sebelum dilakukan pemodelan. Tahap pertama adalah pembagian dataset menjadi data *train* dan data *test* dengan proporsi *test size* sebesar 20%, sehingga data pelatihan mencakup 80% dari keseluruhan dataset. Proses ini menggunakan *random state* (dengan nilai 42) untuk

memastikan *reproducibility*, yaitu agar hasil yang diperoleh tetap konsisten pada setiap eksekusi *code*. Variabel target yang digunakan adalah ‘AQI_Index’, sementara variabel prediktor dipilih berdasarkan proses feature selection yang telah dilakukan sebelumnya, yaitu sebanyak 11 fitur.

Sebelum *fitting model*, data dilakukan *scaling* menggunakan *Robust Scaler* untuk menormalkan nilai-nilai pada fitur sehingga lebih stabil terhadap pencilinan. Sepuluh model regresi digunakan dalam penelitian ini, yaitu Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost, Support Vector Machine, Ridge Regression, Linear Regression, Lasso, dan ElasticNet. Setiap model di-fit pada data *train* dan diuji pada data *test* dengan parameter *default* menggunakan *solver auto*. Evaluasi performa model dilakukan berdasarkan lima metrik: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), dan R². Metrik hasil *modeling* yang diukur pada data *test* untuk memprediksi AQI adalah sebagai berikut

TABEL 7. HASIL EVALUASI MODEL AWAL UNTUK ASIA TENGGARA

Model	MSE	RMSE	MAE	MAPE	R ²
Random Forest	165.525	40.685	31.941	0.0369	0.9890
Gradient Boosting	87.475	29.576	23.842	0.0286	0.9942
XGBoost	122.255	34.965	27.794	0.0316	0.9919
LightGBM	58.385	24.163	18.966	0.0216	0.9961
CatBoost	0.6378	0.7986	0.6114	0.0070	0.9996
SVM	785.638	88.636	58.919	0.0817	0.9480
Ridge Regression	0.0016	0.0397	0.0317	0.0004	1
Linear Regression	0.0000	0.0029	0.0025	0.0000	1
Lasso	100.664	31.728	26.473	0.0349	0.9933
ElasticNet	3.952.905	198.819	160.462	0.2106	0.7383

TABEL 8. HASIL EVALUASI MODEL AWAL UNTUK ASIA TIMUR

Model	MSE	RMSE	MAE	MAPE	R ²
Random Forest	335.768	57.945	45.530	0.0546	0.9774
Gradient Boosting	133.216	36.499	28.738	0.0339	0.9910
XGBoost	241.237	49.116	39.199	0.0455	0.9838
LightGBM	126.182	35.522	27.150	0.0316	0.9915
CatBoost	23.886	15.455	11.296	0.0132	0.9984
SVM	1.977.537	140.625	104.293	0.1306	0.8670
Ridge Regression	0.0096	0.0980	0.0796	0.0010	10.000
Linear Regression	0.0000	0.0028	0.0024	0.0000	10.000
Lasso	94.304	30.709	26.077	0.0337	0.9937
ElasticNet	3.277.572	181.041	150.354	0.1913	0.7796

Hasil evaluasi model prediksi Air Quality Index (AQI) di Asia Tenggara dan Asia Timur mengidentifikasi bahwa Linear Regression, Ridge Regression, dan CatBoost merupakan model dengan performa terbaik. Linear Regression menunjukkan hasil yang sangat unggul dengan nilai Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), dan Mean Absolute Percentage Error (MAPE) yang sangat rendah, serta nilai R² sebesar 1, yang menandakan

kemampuan model ini untuk menangkap hubungan linier antara fitur dan AQI secara sempurna. Ridge Regression, yang menerapkan regularisasi L2, memiliki performa yang hampir identik dengan Linear Regression, menjadikannya alternatif yang stabil dan andal, terutama dalam mengurangi risiko *overfitting* pada data dengan fitur yang saling berkorelasi.

Di sisi lain, CatBoost unggul dalam menangani hubungan non-linier dan fitur kategorikal dengan performa yang sangat baik, memberikan nilai error yang rendah di kedua wilayah Asia. Model berbasis *gradient boosting* lainnya, seperti LightGBM dan Gradient Boosting, juga menunjukkan hasil yang kompetitif meskipun sedikit di bawah CatBoost. Model seperti Support Vector Machine (SVM), ElasticNet, dan Lasso menunjukkan performa yang kurang optimal dengan nilai error yang lebih tinggi.

- Pemilihan Model Terbaik

Selain melihat metrik evaluasi pada data *test*, dilakukan juga Cross Validation menggunakan KFold splits sebanyak 10 untuk mendukung keputusan dalam pemilihan model terbaik. Pendekatan ini bertujuan untuk menguji stabilitas dan generalisasi model terhadap data baru. Dalam proses ini, data dipecah menjadi 10 lipatan (*folds*), di mana setiap lipatan menjadi data validasi secara bergantian, sementara sisanya digunakan sebagai data pelatihan (*train*). Skor evaluasi dihitung untuk setiap lipatan, lalu dirata-rata untuk mendapatkan performa keseluruhan model.

TABEL 9. HASIL EVALUASI MODEL AWAL UNTUK ASIA TENGGARA (DENGAN KFOLD CROSS-VALIDATION)

Model	MSE	RMSE	MAE	MAPE	R ²
Random Forest	195.998	44.272	34.540	0.0385	0.9868
Gradient Boosting	92.540	30.420	23.985	0.0271	0.9938
XGBoost	127.609	35.722	28.268	0.0318	0.9914
LightGBM	66.520	25.792	20.143	0.0226	0.9955
CatBoost	0.7633	0.8737	0.6626	0.0075	0.9995
SVM	728.951	85.379	55.896	0.0704	0.9513
Ridge Regression	0.0018	0.0425	0.0339	0.0004	1
Linear Regression	0.0000	0.0029	0.0025	0.0000	1
Lasso	98.364	31.363	25.894	0.0322	0.9934
ElasticNet	3.813.892	195.292	156.897	0.1928	0.7451

TABEL 10. HASIL EVALUASI MODEL AWAL UNTUK ASIA TIMUR (DENGAN KFOLD CROSS-VALIDATION)

Model	MSE	RMSE	MAE	MAPE	R ²
Random Forest	352.233	59.349	45.638	0.0527	0.9778
Gradient Boosting	158.583	39.823	31.787	0.0370	0.9900
XGBoost	256.802	50.676	39.230	0.0461	0.9837
LightGBM	143.662	37.903	29.316	0.0337	0.9909
CatBoost	33.935	18.421	13.200	0.0154	0.9979
SVM	2.652.427	162.863	120.409	0.1458	0.8336
Ridge Regression	0.0131	0.1143	0.0927	0.0011	0.9999
Linear Regression	0.0000	0.0030	0.0026	0.0000	1
Lasso	92.230	30.369	25.558	0.0333	0.9942
ElasticNet	3.803.285	195.020	159.758	0.2008	0.7606

Hasil evaluasi menggunakan KFold Cross-Validation menunjukkan bahwa tiga model terbaik untuk prediksi AQI di Asia Tenggara dan Asia Timur adalah Linear Regression, Ridge Regression, dan CatBoost. Model linear, seperti Linear Regression dan Ridge Regression, menunjukkan performa yang sangat baik, dengan nilai MSE, RMSE, MAE, dan MAPE yang paling rendah dibandingkan model lainnya. Secara spesifik, Linear Regression memiliki metrik performa yang sedikit lebih baik dibandingkan Ridge Regression. Namun, Ridge Regression tetap memberikan hasil yang hampir setara dengan Linear Regression dan memiliki keunggulan dalam menangani kompleksitas data.

Meskipun Linear Regression memiliki performa dengan metrik evaluasi yang sedikit lebih baik, Ridge Regression akan dipilih untuk memprediksi AQI karena memiliki kemampuan regularisasi (yakni L_2) yang lebih baik. Ridge Regression secara khusus dapat mengatasi potensi multikolinearitas di antara fitur-fitur yang digunakan, yang sering menjadi tantangan dalam dataset lingkungan dan energi seperti dataset pada kasus ini. Regularisasi pada Ridge Regression membantu mengurangi risiko *overfitting*, sehingga menghasilkan model yang lebih *robust* dan lebih stabil saat diterapkan pada data baru. Dengan kombinasi metrik evaluasi dan analisis Cross Validation, Ridge Regression dipilih sebagai model utama karena memberikan keseimbangan antara akurasi, stabilitas, dan kemampuan untuk menghadapi multikolinearitas pada data dengan banyak fitur. Jadi, Ridge Regression dipilih menjadi model terbaik dalam memprediksi AQI untuk Asia Tenggara dan Asia Timur, memastikan keseimbangan antara akurasi prediksi dan stabilitas model.

- *Tuning Model*

Ridge Regression telah dipilih sebagai model utama untuk memprediksi Air Quality Index (AQI). Untuk meningkatkan akurasi dan stabilitas prediksi, dilakukan tuning parameter pada model ini. Dua parameter utama yang akan dituning adalah *solver* dan *alpha*, yang memiliki peran penting dalam menentukan kinerja Ridge Regression. Proses tuning dilakukan melalui empat pendekatan yang dirancang untuk mengeksplorasi kombinasi parameter secara sistematis dan memastikan performa optimal model pada data.

Pada pendekatan pertama, *tuning* dilakukan dengan mengombinasikan parameter *solver* dan *alpha* pada Ridge Regression, diikuti evaluasi menggunakan K-Fold Cross-Validation dengan 10 folds. Hasil evaluasi menunjukkan bahwa *solver auto* memberikan performa yang konsisten dan optimal di seluruh metrik, termasuk MSE, RMSE, MAE, MAPE, dan R^2 , dengan hasil yang identik dengan *solver* lain seperti *svd*, *cholesky*, dan *lsqr*. *Solver auto* menunjukkan fleksibilitas dengan secara otomatis memilih algoritma terbaik berdasarkan sifat data, sehingga mempermudah pemodelan tanpa memerlukan intervensi manual. Selain itu, analisis terhadap parameter *alpha* menunjukkan bahwa semakin kecil nilai *alpha*, semakin baik performa model dalam hal akurasi prediksi, meskipun tetap mempertahankan manfaat regularisasi. Hal ini mengindikasikan bahwa *alpha* dengan nilai rendah, seperti 0.01, memberikan keseimbangan yang baik antara bias

dan varians, menghasilkan model yang stabil dan akurat. Tabel di bawah ini memberikan ringkasan dari kombinasi *solver* dan *alpha* untuk data Asia Tenggara dan Asia Timur

TABEL 11. HASIL EVALUASI MODEL DENGAN KOMBINASI SOLVER DAN ALPHA UNTUK ASIA TIMUR (DENGAN KFOLD CROSS-VALIDATION)(PENDEKATAN SATU)

<i>Solver</i>	<i>Alpha</i>	MSE	RMSE	MAE	MAPE	R²
<i>auto</i>	0.01	0.0000	0.0029	0.0025	0.0000	1
<i>svd</i>	0.01	0.0000	0.0029	0.0025	0.0000	1
<i>cholesky</i>	0.01	0.0000	0.0029	0.0025	0.0000	1
<i>lsqr</i>	0.01	0.0000	0.0029	0.0025	0.0000	1
<i>sparse_cg</i>	0.01	0.0000	0.0041	0.0034	0.0000	1
<i>sag</i>	0.01	0.0000	0.0043	0.0035	0.0000	1
<i>saga</i>	0.01	0.0000	0.0038	0.0031	0.0000	1
<i>auto</i>	0.10	0.0000	0.0051	0.0042	0.0000	1
<i>svd</i>	0.10	0.0000	0.0051	0.0042	0.0000	1
<i>cholesky</i>	0.10	0.0000	0.0051	0.0042	0.0000	1
<i>lsqr</i>	0.10	0.0000	0.0051	0.0042	0.0000	1
<i>sparse_cg</i>	0.10	0.0000	0.0060	0.0048	0.0001	1
<i>sag</i>	0.10	0.0000	0.0058	0.0047	0.0001	1
<i>saga</i>	0.10	0.0000	0.0059	0.0048	0.0001	1
<i>auto</i>	1.00	0.0018	0.0424	0.0339	0.0004	1
<i>svd</i>	1.00	0.0018	0.0424	0.0339	0.0004	1
<i>cholesky</i>	1.00	0.0018	0.0424	0.0339	0.0004	1
<i>lsqr</i>	1.00	0.0018	0.0424	0.0339	0.0004	1
<i>sparse_cg</i>	1.00	0.0018	0.0426	0.0340	0.0004	1
<i>sag</i>	1.00	0.0018	0.0425	0.0339	0.0004	1
<i>saga</i>	1.00	0.0018	0.0423	0.0338	0.0004	1
<i>auto</i>	10.00	0.1724	0.4148	0.3300	0.0039	0.9999
<i>svd</i>	10.00	0.1724	0.4148	0.3300	0.0039	0.9999
<i>cholesky</i>	10.00	0.1724	0.4148	0.3300	0.0039	0.9999
<i>lsqr</i>	10.00	0.1724	0.4148	0.3300	0.0039	0.9999
<i>sparse_cg</i>	10.00	0.1724	0.4149	0.3301	0.0039	0.9999
<i>sag</i>	10.00	0.1723	0.4146	0.3299	0.0039	0.9999
<i>saga</i>	10.00	0.1724	0.4148	0.3300	0.0039	0.9999
<i>auto</i>	100.00	127.825	35.720	28.419	0.0341	0.9915
<i>svd</i>	100.00	127.825	35.720	28.419	0.0341	0.9915
<i>cholesky</i>	100.00	127.825	35.720	28.419	0.0341	0.9915
<i>lsqr</i>	100.00	127.825	35.720	28.419	0.0341	0.9915
<i>sparse_cg</i>	100.00	127.830	35.721	28.418	0.0341	0.9915
<i>sag</i>	100.00	127.816	35.719	28.418	0.0341	0.9915
<i>saga</i>	100.00	127.837	35.722	28.420	0.0341	0.9914

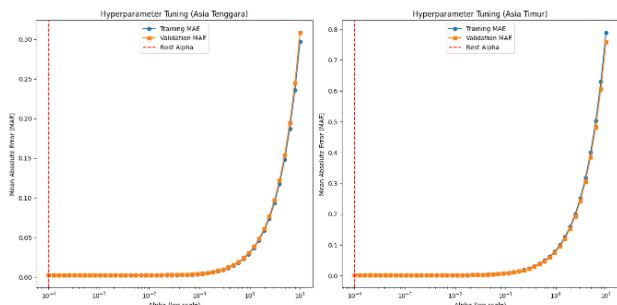
TABEL 12. HASIL EVALUASI MODEL DENGAN KOMBINASI SOLVER DAN ALPHA UNTUK ASIA TIMUR (DENGAN KFOLD CROSS-VALIDATION) (PENDEKATAN SATU)

<i>Solver</i>	<i>Alpha</i>	MSE	RMSE	MAE	MAPE	R²
<i>auto</i>	0.01	0.01	0.0000	0.0032	0.0027	0.0000
<i>svd</i>	0.01	0.01	0.0000	0.0032	0.0027	0.0000
<i>cholesky</i>	0.01	0.01	0.0000	0.0032	0.0027	0.0000
<i>lsqr</i>	0.01	0.01	0.0000	0.0032	0.0027	0.0000
<i>sparse_cg</i>	0.01	0.01	0.0000	0.0041	0.0033	0.0000
<i>sag</i>	0.01	0.01	0.0000	0.0044	0.0036	0.0000
<i>saga</i>	0.01	0.01	0.0000	0.0040	0.0033	0.0000
<i>auto</i>	0.10	0.10	0.0001	0.0119	0.0095	0.0001
<i>svd</i>	0.10	0.10	0.0001	0.0119	0.0095	0.0001
<i>cholesky</i>	0.10	0.10	0.0001	0.0119	0.0095	0.0001
<i>lsqr</i>	0.10	0.10	0.0001	0.0119	0.0095	0.0001
<i>sparse_cg</i>	0.10	0.10	0.0001	0.0121	0.0098	0.0001
<i>sag</i>	0.10	0.10	0.0001	0.0120	0.0097	0.0001
<i>saga</i>	0.10	0.10	0.0001	0.0121	0.0098	0.0001
<i>auto</i>	1.00	1.00	0.0131	0.1140	0.0927	0.0011
<i>svd</i>	1.00	1.00	0.0131	0.1140	0.0927	0.0011
<i>cholesky</i>	1.00	1.00	0.0131	0.1140	0.0927	0.0011
<i>lsqr</i>	1.00	1.00	0.0131	0.1140	0.0926	0.0011
<i>sparse_cg</i>	1.00	1.00	0.0130	0.1140	0.0926	0.0011
<i>sag</i>	1.00	1.00	0.0131	0.1142	0.0928	0.0011
<i>saga</i>	1.00	1.00	0.0130	0.1136	0.0923	0.0011
<i>auto</i>	10.00	10.00	11.705	10.798	0.8791	0.0106

<i>svd</i>	10.00	10.00	11.705	10.798	0.8791	0.0106
<i>cholesky</i>	10.00	10.00	11.705	10.798	0.8791	0.0106
<i>lsqr</i>	10.00	10.00	11.704	10.797	0.8791	0.0106
<i>sparse_cg</i>	10.00	10.00	11.708	10.799	0.8791	0.0106
<i>sag</i>	10.00	10.00	11.694	10.793	0.8786	0.0105
<i>saga</i>	10.00	10.00	11.703	10.797	0.8790	0.0106
<i>auto</i>	100.00	100.00	618.797	78.531	64.231	0.0783
<i>svd</i>	100.00	100.00	618.797	78.531	64.231	0.0783
<i>cholesky</i>	100.00	100.00	618.797	78.531	64.231	0.0783
<i>lsqr</i>	100.00	100.00	618.788	78.530	64.230	0.0783
<i>sparse_cg</i>	100.00	100.00	618.809	78.532	64.232	0.0783
<i>sag</i>	100.00	100.00	618.807	78.531	64.232	0.0783
<i>saga</i>	100.00	100.00	618.810	78.532	64.232	0.0783

Berdasarkan analisis di atas, akan dipilih *solver* ‘auto’ untuk tuning parameter alpha selanjutnya. *Solver* ‘auto’ dipilih karena fleksibilitasnya dalam secara otomatis menentukan algoritma terbaik berdasarkan karakteristik data, memastikan efisiensi tanpa intervensi manual. Evaluasi menunjukkan performa *solver* ‘auto’ identik dengan *solver* seperti *svd*, *cholesky*, dan *lsqr* pada metrik utama, menjadikannya pilihan optimal untuk pemodelan dengan karakteristik data yang kompleks.

Pada pendekatan kedua, digunakan grafik metrik evaluasi MAE untuk melihat parameter *alpha*. Dalam proses tuning parameter *alpha* dengan *solver* ‘auto’, terlihat jelas dari grafik bahwa semakin kecil nilai *alpha*, semakin rendah Mean Absolute Error (MAE), yang menunjukkan peningkatan kinerja model Ridge Regression dalam memprediksi AQI. Nilai *alpha* terkecil yang terpilih sebagai titik optimal menunjukkan performa terbaik, dengan MAE mencapai titik terendah sebelum naik kembali seiring bertambahnya nilai *alpha*.



Gambar 10. Grafik Metrik Evaluasi MAE untuk setiap Nilai Alpha (Pendekatan 2)

Pendekatan ketiga, menggunakan GridSearchCV yang dipilih berdasarkan MAE dan RMSE dengan *solver* ‘auto’. Dalam tuning parameter *alpha* Ridge Regression menggunakan GridSearchCV dengan *solver* ‘auto’, fokusnya adalah meminimalkan MAE dan RMSE. Algoritma yang diterapkan untuk data Asia Tenggara dan Timur menemukan bahwa *alpha* terbaik adalah 0.0, di mana menghasilkan nilai MAE dan RMSE yang paling rendah. Algoritma ini menggunakan RepeatedKFold untuk memastikan konsistensi dan keakuratan model, menunjukkan bahwa *alpha* yang lebih rendah cenderung menghasilkan prediksi yang lebih akurat.

TABEL 13. HASIL OPTIMAL PARAMETER *ALPHA* DARI GRIDSEARCHCV UNTUK RIDGE REGRESSION PADA PREDIKSI AQI (PENDEKATAN 3)

Wilayah	Best Alpha	Best MAE	Best RMSE
Asia Tenggara	0.0	0.002	0.003
Asia Timur	0.0	0.003	0.003

TABEL 14. HASIL OPTIMAL PARAMETER ALPHA DARI RIDGECV UNTUK RIDGE REGRESSION PADA PREDIKSI AQI (PENDEKATAN 4)

Wilayah	Best Alpha (MAE)	Best Alpha (RMSE)
Asia Tenggara	0.0	0.0
Asia Timur	0.0	0.0

Pada pendekatan keempat ini, pemilihan parameter *alpha* dilakukan menggunakan RidgeCV dengan *solver* ‘auto’, berdasarkan metrik MAE dan RMSE. Hasilnya, nilai *alpha* terbaik untuk kedua metrik pada wilayah Asia Tenggara dan Asia Timur adalah 0.0, yang berarti tidak ada regularisasi yang diterapkan, menjadikan Ridge Regression setara dengan Linear Regression. Pemilihan *alpha* ini dievaluasi menggunakan K-fold cross-validation untuk menilai kinerja model dalam memprediksi AQI.

Berdasarkan empat pendekatan di atas, akan digunakan parameter *solver* ‘auto’ dan *alpha* 0.01. Dipilih *solver* ‘auto’ karena hasil cross-validation menunjukkan bahwa *solver* ini menghasilkan kinerja yang optimal dengan nilai MSE, RMSE, MAE, dan MAPE yang identik dengan *solver* lain seperti *svd*, *cholesky*, dan *lsqr*, yang juga memberikan performa terbaik. Selain itu, *solver* ‘auto’ memiliki keunggulan utama dalam fleksibilitasnya, karena secara otomatis memilih algoritma terbaik berdasarkan sifat data. Ini membuat *solver* ‘auto’ lebih andal untuk digunakan dalam berbagai kondisi data tanpa memerlukan intervensi manual, sehingga memberikan efisiensi dalam proses pemodelan. Hasil dari GridSearchCV dan RidgeCV menunjukkan bahwa nilai parameter terbaik untuk *alpha* adalah 0. Namun, ketika *alpha* = 0, Ridge Regression menjadi identik dengan Linear Regression, karena tidak ada regularisasi yang diterapkan. Untuk tetap memanfaatkan manfaat regularisasi Ridge Regression, nilai *alpha* dipilih sedikit di atas 0, yaitu 0.01. Dengan *alpha* 0.01, Ridge Regression dapat mengurangi *overfitting* dan memberikan solusi yang lebih stabil, sekaligus tetap mempertahankan model yang optimal. Jadi, diperoleh model terbaik untuk dataset Asia Tenggara dan Asia Timur adalah Ridge Regression dengan parameter *solver* ‘auto’ dan *alpha* 0.01.

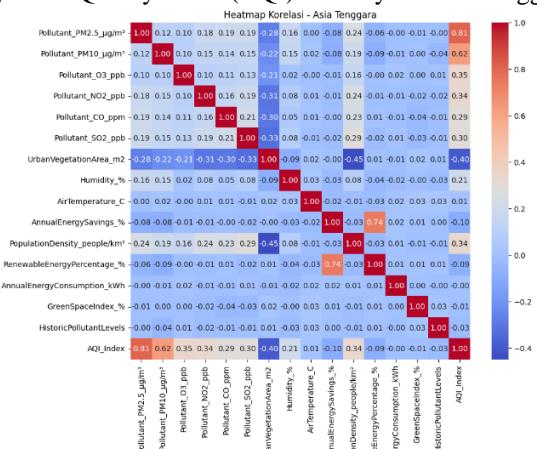
5. Deployment

Pada tahap *deployment*, model terbaik untuk Asia Tenggara dan Asia Timur diterapkan dalam simulasi kebijakan menggunakan *platform* Dash. Simulasi ini dirancang untuk memvisualisasikan hasil prediksi secara interaktif, dengan tujuan mengevaluasi dampak kebijakan lingkungan terhadap Air Quality (AQI). Fitur-fitur yang akan digunakan dalam simulasi kebijakan diidentifikasi berdasarkan analisis *feature importance* dan SHAP, sehingga memastikan kebijakan yang disimulasikan fokus pada variabel yang paling berpengaruh. Parameter kebijakan dapat diatur melalui kontrol interaktif pada *dashboard*, memungkinkan pengguna untuk menguji berbagai skenario secara fleksibel. Simulasi ini diharapkan mampu memberikan wawasan yang komprehensif mengenai efektivitas kebijakan yang diusulkan dalam memperbaiki kualitas udara di kedua kawasan tersebut.

C. Hasil dan Pembahasan

1. Hubungan faktor energi dan lingkungan terhadap Air Quality Index (AQI) di Asia Tenggara dan Asia Timur

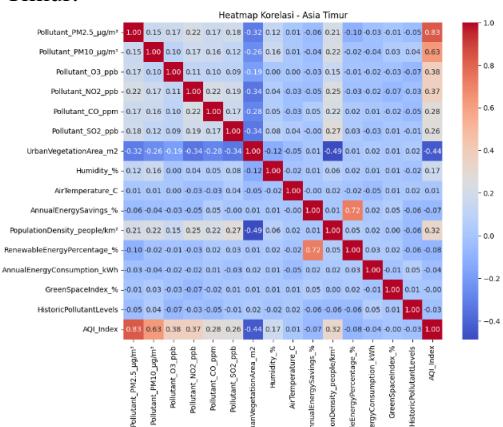
Melalui visualisasi menggunakan *heatmap*, diperoleh gambaran hubungan antara faktor energi dan lingkungan dengan Air Quality Index (AQI) di wilayah Asia Tenggara.



Gambar 11. *Heatmap* Korelasi Fitur Asia Tenggara

Faktor-faktor yang mempengaruhi AQI di Asia Tenggara menunjukkan adanya korelasi positif dan negatif. Secara positif, “Pollutant_PM2.5_µg/m³” dengan korelasi sebesar 0.81 menjadi faktor dominan dengan dampak besar terhadap kualitas udara, diikuti oleh “Pollutant_PM10_µg/m³” dengan korelasi sebesar 0.62 yang juga berkontribusi signifikan. Selain itu, gas polutan seperti “Pollutant_O3_ppb” dan “Pollutant_NO2_ppb”, serta “PopulationDensity_people/km²”, memiliki korelasi sedang yang turut meningkatkan AQI. Sebaliknya, terdapat korelasi negatif pada faktor-faktor yang dapat mengurangi AQI, seperti “UrbanVegetationArea_m²” dengan korelasi sebesar -0.40 yang menunjukkan peran signifikan ruang hijau di wilayah perkotaan terhadap penurunan AQI, sementara “AnnualEnergySavings_%” dengan korelasi sebesar -0.10 dan “RenewableEnergyPercentage_%” dengan korelasi sebesar -0.09 memiliki kontribusi kecil dalam menurunkan AQI.

Berikut merupakan *heatmap* korelasi fitur pada data di Asia Timur.



Gambar 12. Hasil analisis korelasi fitur Asia Tenggara

Faktor-faktor yang mempengaruhi AQI di Asia Timur menunjukkan pola korelasi positif dan negatif yang serupa dengan Asia Tenggara. “Pollutant_PM2.5_µg/m³” dengan korelasi sebesar 0.83 tetap menjadi faktor utama yang meningkatkan AQI, diikuti oleh “Pollutant_PM10_µg/m³” dengan korelasi sebesar 0.63 yang memberikan kontribusi signifikan. Faktor lain seperti “Pollutant_O3_ppb”, “Pollutant_NO2_ppb”, dan “PopulationDensity_people/km²” juga memiliki dampak sedang terhadap peningkatan AQI. Sebaliknya, faktor yang berkorelasi negatif seperti “UrbanVegetationArea_m²” dengan nilai korelasi -0.44 menunjukkan bahwa vegetasi di Asia Timur memiliki pengaruh lebih besar dalam mengurangi AQI dibandingkan Asia Tenggara. Sementara itu, “RenewableEnergyPercentage_%” dengan nilai korelasi sebesar -0.08 menunjukkan pengaruh energi terbarukan terhadap AQI tetap kecil dan sebanding dengan kondisi di Asia Tenggara.

Perbandingan antara Asia Tenggara dan Asia Timur menunjukkan bahwa PM_{2.5} dan PM₁₀ adalah faktor dominan yang mempengaruhi AQI di kedua wilayah, dengan Asia Timur memiliki korelasi yang sedikit lebih tinggi, yaitu pada Asia Timur untuk PM_{2.5} sebesar 0.83 dan PM₁₀ sebesar 0.63 dan pada Asia Tenggara untuk PM_{2.5} sebesar 0.81 dan PM₁₀ sebesar 0.62. Faktor "UrbanVegetationArea" menunjukkan dampak ruang hijau lebih besar di Asia Timur dibandingkan Asia Tenggara, mengindikasikan bahwa urbanisasi di Asia Timur lebih terkendali melalui peningkatan area vegetasi.

2. Model prediksi Air Quality Index (AQI) terbaik berdasarkan data energi dan lingkungan di Asia Tenggara dan Asia Timur.

Pada bagian *Metode Penelitian* Subbab *Modeling dan Evaluation* telah dijabarkan langkah-langkah hingga didapatkan model terbaik untuk memprediksi AQI di Asia Tenggara dan Asia Timur, yakni Ridge Regression dengan parameter tuning (*solver* ‘auto’ dan *alpha* 0.01).

a) Perbandingan Metrik Evaluasi Sebelum dan Setelah Parameter Tuning pada Ridge Regression

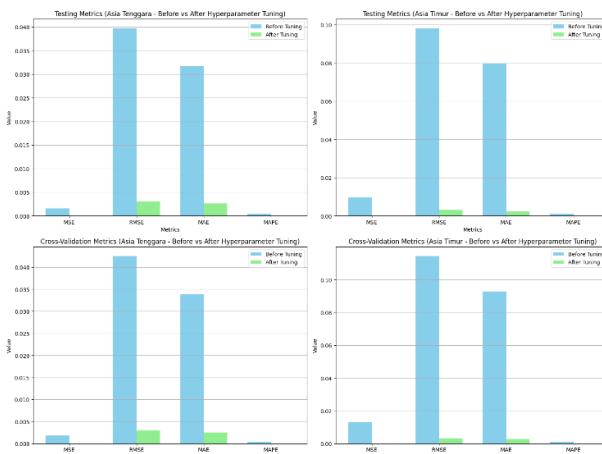
Berikut merupakan metriks evaluasi sebelum dan setelah *parameter tuning* dengan model Ridge Regression pada data Asia Tenggara dan Asia Timur.

TABEL 15. TABEL PERBANDINGAN SEBELUM DAN SESUDAH TUNING PADA MODEL RIDGE REGRESSION PADA ASIA TENGGARA

Metrik	Before Tuning	After Tuning
MSE	0.001576	0.000009
RMSE	0.039694	0.003000
MAE	0.031723	0.002589
MAPE	0.000406	0.000031
R^2	0.999999	1

TABEL 16. TABEL PERBANDINGAN SEBELUM DAN SESUDAH TUNING PADA MODEL RIDGE REGRESSION PADA ASIA TIMUR

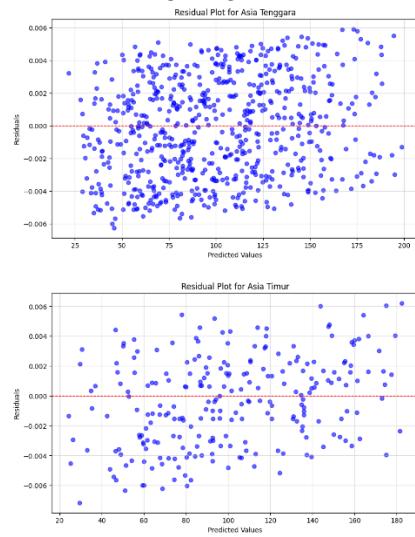
Metrik	Before Tuning	After Tuning
MSE	0.009604	0.000009
RMSE	0.097998	0.002988
MAE	0.079554	0.002520
MAPE	0.000953	0.000032
R^2	0.999994	1



Gambar 13. Perbandingan Metrik Evaluasi Sebelum dan Setelah Tuning pada Ridge Regression pada data Asia Tenggara (kiri) dan Asia Timur (kanan)

Gambar dan tabel di atas adalah hasil akhir perbandingan model sebelum dan sesudah *parameter tuning*. Hasil *tuning* model menunjukkan performa yang lebih baik dibandingkan dengan parameter *default*. Sebelum *tuning*, model menunjukkan nilai yang tinggi pada metrik evaluasi seperti MSE, RMSE, dan MAE. Setelah *tuning* dengan parameter *solver* ‘auto’ dan *alpha* 0.01, metrik evaluasi mengalami penurunan yang signifikan, terutama pada RMSE dan MAE, yang menunjukkan bahwa model menghasilkan prediksi yang lebih akurat dan lebih stabil. Grafik di atas menggambarkan peningkatan performa model, dengan penurunan nilai MSE dan RMSE setelah tuning pada data *testing* maupun *cross-validation*. Hal ini menunjukkan bahwa *tuning* parameter tidak hanya mengoptimalkan hasil prediksi, tetapi juga meningkatkan stabilitas model pada berbagai set data, menjadikannya lebih dapat diandalkan dalam memprediksi AQI Index.

b) Residual Ridge Regression



Gambar 14. Residual Plot model Ridge Regression Terbaik untuk Asia Tenggara (atas) dan Asia Timur (bawah)

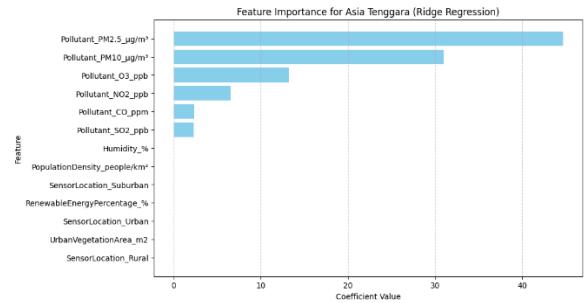
Residual plots untuk Ridge Regression di Asia Tenggara dan Asia Timur menunjukkan pola acak di sekitar garis nol, menandakan bahwa model berhasil menangkap hubungan antara fitur dan AQI dengan baik. Meskipun terdapat beberapa titik residual yang lebih besar, sebagian besar error model berada dekat garis nol yang mana menunjukkan prediksi yang relatif akurat. Tidak ada indikasi *overfitting* atau *underfitting*, dan model menunjukkan performa yang stabil dengan error yang kecil di kedua wilayah.

3. Faktor signifikan dalam mempengaruhi Air Quality Index (AQI) di kedua wilayah

Untuk mencari faktor signifikan yang mempengaruhi Air Quality Index (AQI) di Asia Tenggara dan Asia Timur, digunakan dua metode, yaitu dengan *feature importance* dan SHAP.

a) Feature importance

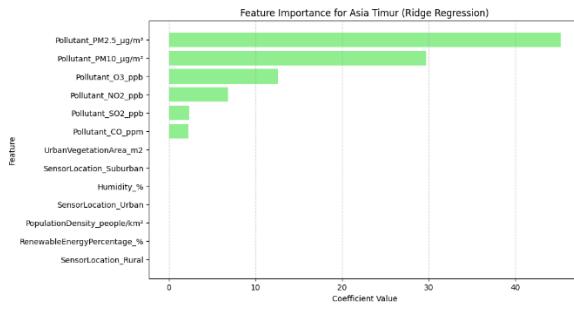
Analisis *feature importance* memberikan wawasan tentang seberapa besar pengaruh setiap faktor *input* terhadap hasil prediksi yang dihasilkan oleh model. Dengan model Ridge Regression yang digunakan untuk memprediksi Air Quality Index (AQI), *feature importance* digunakan untuk memahami faktor-faktor yang paling signifikan dalam memprediksi kualitas udara [32]. Berikut merupakan hasil dari *feature importance* di wilayah Asia Tenggara.



Gambar 15. Feature Importance Asia Tenggara

Di Asia Tenggara, faktor yang paling signifikan dalam mempengaruhi AQI adalah “Pollutant_P M2.5_\mu g/m^3” dengan nilai 44.667, yang menunjukkan bahwa PM_{2.5} memiliki pengaruh terbesar terhadap kualitas udara di kawasan ini. Diikuti oleh “Pollutant_P M10_\mu g/m^3” dengan nilai 30.947, yang juga menjadikan PM₁₀ faktor penting dalam menentukan nilai AQI. Meskipun kontribusinya lebih kecil dibandingkan dengan PM_{2.5} dan PM₁₀, “Pollutant_O3_ppb” (ozon) dengan nilai 13.193 tetap memegang peranan penting dalam kualitas udara. Di sisi lain, faktor-faktor dengan pengaruh rendah termasuk “UrbanVegetationArea_m2” (-0.000089), “RenewableEnergyPercentage_ %” (0.000050), dan “Humidity_ %” (0.000048) yang memiliki dampak sangat kecil terhadap AQI. Variabel kategorik seperti “SensorLocation” juga menunjukkan pengaruh yang hampir tidak signifikan dalam memprediksi nilai AQI.

Berikut merupakan hasil dari *feature importance* di wilayah Asia Timur.



Gambar 16. Feature Importance Asia Timur

Di Asia Timur, faktor yang paling dominan dalam mempengaruhi AQI adalah “Pollutant_PM2.5_µg/m³” dengan nilai 45.268, yang tetap menjadi faktor utama yang memengaruhi kualitas udara di kawasan ini. “Pollutant_PM10_µg/m³” dengan nilai 29.691 juga memberikan kontribusi besar terhadap AQI, menjadikannya faktor kedua yang signifikan. Selain itu, “Pollutant_O3_ppb (ozon)” dengan nilai 12.650 menjadi faktor ketiga yang berperan penting dalam kualitas udara. Di sisi lain, variabel-variabel seperti “UrbanVegetationArea_m²” (0.000309), “RenewableEnergyPercentage_%” (-0.000123), dan “Humidity_%” (0.000029) memiliki pengaruh kecil terhadap AQI. Variabel kategorik seperti SensorLocation juga menunjukkan pengaruh yang hampir tidak signifikan dalam memprediksi AQI.

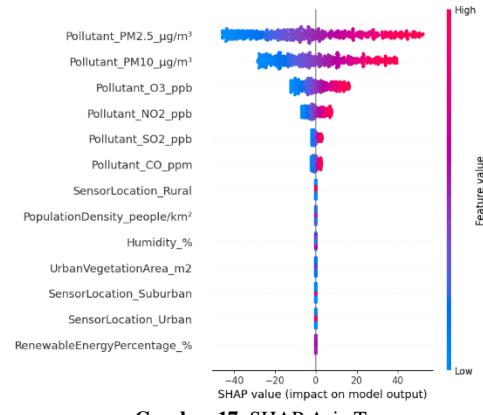
PM_{2.5} dan PM₁₀ merupakan faktor utama yang memengaruhi kualitas udara di Asia Tenggara dan Asia Timur, dengan konsentrasi keduanya menjadi penyebab utama penurunan AQI di kedua wilayah. Asia Timur menunjukkan sensitivitas yang sedikit lebih tinggi terhadap PM_{2.5} dibandingkan Asia Tenggara, sementara ozon (“Pollutant_O3_ppb”) memberikan kontribusi yang sedikit lebih besar di Asia Tenggara. Faktor-faktor lain seperti vegetasi urban (“UrbanVegetationArea_m²”) dan kebijakan energi (“RenewableEnergyPercentage_%”) memiliki pengaruh yang sangat rendah terhadap AQI, begitu pula variabel kategorik seperti “SensorLocation” yang menunjukkan dampak yang hampir tidak signifikan.

Kesimpulannya, PM_{2.5} menjadi faktor dominan yang paling signifikan di kedua wilayah, diikuti oleh PM₁₀. Asia Timur sedikit lebih sensitif terhadap PM_{2.5} dibandingkan Asia Tenggara, sementara Asia Tenggara menunjukkan pengaruh yang sedikit lebih besar dari ozon. Faktor minor lainnya, seperti vegetasi urban, kelembaban, dan kebijakan energi, hanya memberikan kontribusi kecil terhadap kualitas udara, menunjukkan bahwa faktor-faktor ini tidak terlalu mempengaruhi AQI secara signifikan.

b) SHAP

Nilai Shapley adalah metode yang digunakan untuk mengukur kontribusi setiap fitur dalam model prediksi. Hasil prediksi memberikan *knowledge* yang lebih mendalam melalui penggunaan nilai Shapley. Secara umum, distribusi nilai Shapley memberikan *knowledge* tentang dampak, pola, dan hubungan antar fitur yang berbeda. Pada level yang lebih spesifik, nilai Shapley

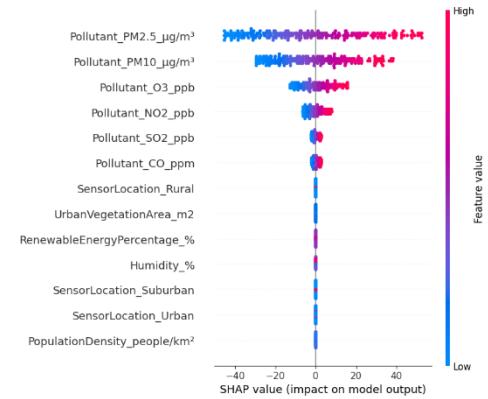
jugalah dapat digunakan untuk pengukuran yang akurat terhadap kontribusi masing-masing fitur terhadap prediksi untuk setiap sampel individu [32]. Berikut merupakan hasil dari SHAP pada wilayah Asia Tenggara.



Gambar 17. SHAP Asia Tenggara

Fitur utama yang memengaruhi model adalah “Pollutant_PM2.5_µg/m³”, dengan nilai SHAP rata-rata tertinggi, yaitu 22.77, menunjukkan pengaruh paling signifikan terhadap prediksi, diikuti oleh “Pollutant_PM10_µg/m³” dengan rata-rata 15.55 yang juga berdampak besar pada prediksi model. “Pollutant_O3_ppb” dengan rata-rata 6.69 memiliki pengaruh moderat, sementara “Pollutant_NO2_ppb”, “SO2_ppb”, dan “CO_ppm” menunjukkan kontribusi kecil namun signifikan terhadap hasil prediksi, terutama saat konsentrasi tinggi. Sebaliknya, fitur seperti “SensorLocation”, “PopulationDensity_people/km²”, dan “Humidity_%” memiliki kontribusi hampir nol terhadap prediksi, menunjukkan bahwa model tidak terlalu bergantung pada fitur ini.

Berikut merupakan hasil SHAP pada wilayah Asia Timur.



Gambar 18. SHAP Asia Timur

Hasil analisis SHAP menunjukkan bahwa “Pollutant_PM2.5_µg/m³” adalah fitur yang paling berpengaruh terhadap prediksi model, dengan rata-rata tertinggi, yaitu 23.65. Kemudian, diikuti oleh “Pollutant_PM10_µg/m³” dengan rata-rata 15.69 yang juga memberikan pengaruh besar. “Pollutant_O3_ppb” memiliki kontribusi moderat dengan rata-rata 6.76,

sementara “Pollutant_NO2_ppb”, “Pollutant_SO2_ppb”, dan “Pollutant_CO_ppm” menunjukkan pengaruh yang lebih kecil tetapi tetap signifikan, terutama pada nilai konsentrasi tinggi. Sebaliknya, fitur seperti “SensorLocation”, “UrbanVegetationArea_m²”, “RenewableEnergyPercentage_%”, “Humidity_%”, dan “PopulationDensity_people/km² memiliki kontribusi yang mendekati nol terhadap prediksi. Nilai SHAP yang tinggi untuk PM_{2.5} dan PM₁₀ menunjukkan bahwa peningkatan konsentrasi ini memberikan dampak positif yang signifikan terhadap prediksi model, sedangkan fitur lain memiliki pengaruh yang lebih kecil atau tidak signifikan.

Berdasarkan analisis dengan metode *feature importance* dan SHAP, “Pollutant_PM2.5_µg/m³” dan “Pollutant_PM10_µg/m³” konsisten menjadi faktor paling signifikan yang memengaruhi Air Quality Index (AQI) di Asia Tenggara dan Asia Timur dengan PM_{2.5} memberikan kontribusi terbesar. *Feature importance* menunjukkan bahwa Asia Timur memiliki sensitivitas sedikit lebih tinggi terhadap PM_{2.5} dibandingkan Asia Tenggara, yaitu nilai 45.268 untuk Asia Timur dan 44.667 untuk Asia Tenggara, sementara SHAP juga mendapatkan pola yang sama dengan rata-rata nilai SHAP 23.65 di Asia Timur dibandingkan 22.77 di Asia Tenggara. Sebaliknya, *feature importance* menunjukkan bahwa “Pollutant_O3_ppb” (ozon) memiliki pengaruh sedikit lebih besar di Asia Tenggara, yaitu dengan nilai 13.193 dibandingkan Asia Timur dengan nilai 12.650. Hal ini berbeda dengan SHAP bahwa ozon justru memiliki pengaruh lebih besar di Asia Timur, yaitu sebesar 6.76 dibandingkan Asia Tenggara dengan rata-rata 6.69. Faktor-faktor lain seperti “UrbanVegetationArea_m²”, “RenewableEnergyPercentage_%”, “Humidity_%”, dan variabel kategorik seperti “SensorLocation” menunjukkan pengaruh yang sangat kecil atau hampir tidak signifikan terhadap AQI pada kedua wilayah. Kedua metode ini menegaskan bahwa konsentrasi PM_{2.5} dan PM₁₀ adalah faktor kunci yang secara konsisten mempengaruhi penurunan kualitas udara di kedua wilayah.

4. Simulasi Kebijakan

Pada simulasi kebijakan dilakukan menggunakan model terbaik yang telah ditetapkan untuk Asia Tenggara dan Asia Timur, yaitu Ridge Regression. Parameter yang digunakan dalam simulasi kebijakan di kedua wilayah memanfaatkan fitur-fitur utama yang telah dipilih berdasarkan analisis *feature importance* dan SHAP, yaitu “Pollutant_PM2.5_µg/m³”, “Pollutant_PM10_µg/m³”, “Pollutant_O3_ppb”, dan “Pollutant_NO2_ppb” yang konsisten menunjukkan pengaruh signifikan terhadap kualitas udara di kedua wilayah. Selain itu, fitur tambahan seperti “UrbanVegetationArea_m²” dan “RenewableEnergyPercentage_%” juga digunakan meskipun kontribusinya rendah dalam model prediksi.

Polutan PM_{2.5} dan PM₁₀ diprioritaskan dalam simulasi kebijakan karena keduanya secara konsisten menunjukkan pengaruh terbesar terhadap nilai Air Quality Index (AQI) di Asia Tenggara dan Asia Timur. Konsentrasi PM_{2.5} memiliki dampak langsung pada kesehatan manusia, terutama pada

sistem pernapasan, karena partikel ini dapat masuk ke saluran pernapasan bagian bawah. Demikian pula, PM₁₀ memiliki dampak signifikan terhadap AQI dan kualitas udara, meskipun kontribusinya sedikit lebih kecil dibandingkan PM_{2.5}.

“Pollutant_O3_ppb” (ozon) dan “Pollutant_NO2_ppb” dipilih karena keduanya berkaitan erat dengan aktivitas transportasi dan industri. Ozon permukaan sering terbentuk sebagai hasil reaksi kimia dari polutan lain di atmosfer, sehingga pengurangannya memberikan efek tidak langsung pada perbaikan kualitas udara. Sementara itu, nitrogen dioksida (NO₂) merupakan indikator utama dari emisi kendaraan bermotor dan pabrik, sehingga pengurangannya menjadi langkah penting dalam pengendalian polusi.

Fitur “UrbanVegetationArea_m²” dan “RenewableEnergyPercentage_%” juga diterapkan dalam simulasi kebijakan, meskipun kontribusinya terhadap prediksi AQI relatif kecil. Hal ini disebabkan oleh perannya yang strategis dalam mendukung keberlanjutan lingkungan. Peningkatan area vegetasi urban memiliki manfaat penting, seperti menyerap polutan udara, termasuk karbon dioksida, mengurangi suhu perkotaan, meningkatkan kesehatan masyarakat, dan mendukung pengelolaan lingkungan yang lebih baik. Di sisi lain, peningkatan penggunaan energi terbarukan dapat secara signifikan mengurangi emisi polutan yang berasal dari energi fosil, seperti PM_{2.5}, PM₁₀, dan NO₂, yang merupakan kontributor utama terhadap penurunan kualitas udara. Meskipun dampak langsungnya terhadap model prediksi terbatas, langkah transisi menuju energi bersih dan peningkatan ruang hijau menjadi elemen kebijakan yang esensial untuk menciptakan lingkungan yang lebih sehat dan berkelanjutan dalam jangka panjang.

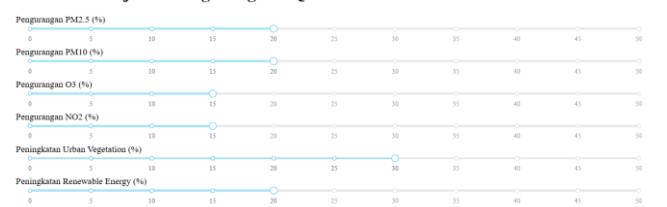
Dalam menentukan fitur kebijakan yang diterapkan dalam simulasi Dash untuk wilayah Asia Tenggara dan Asia Timur ditentukan serupa, karena hasil analisis *feature importance* dan SHAP menunjukkan pola kontribusi fitur utama yang konsisten di kedua wilayah, meskipun tingkat pengaruh masing-masing fitur berbeda. Berikut ini adalah simulasi kebijakan yang menggunakan Dash untuk menghadirkan visualisasi interaktif dan pengaturan parameter kebijakan secara fleksibel yang dirancang khusus untuk wilayah Asia Tenggara dan Asia Timur.

Simulasi Kebijakan Pengurangan AQI di Asia Tenggara



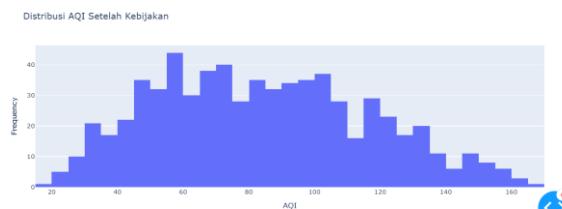
Gambar 19. Antarmuka Dash untuk Simulasi Kebijakan AQI di Asia Tenggara

Simulasi Kebijakan Pengurangan AQI di Asia Timur

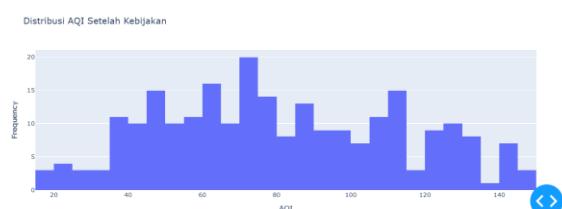


Gambar 20. Antarmuka Dash untuk Simulasi Kebijakan AQI di Asia Timur

Gambar di atas menampilkan antarmuka Dash dengan slider interaktif yang digunakan untuk mengatur persentase pengurangan polutan dan peningkatan fitur strategis. Slider ini memungkinkan pengguna untuk menentukan parameter kebijakan secara fleksibel dan langsung melihat hasil simulasi dalam bentuk visualisasi. Berikut hasil distribusi AQI setelah penerapan kebijakan yang ditunjukkan pada Gambar 21 dan Gambar 22



Gambar 21. Distribusi AQI Setelah Kebijakan di Asia Tenggara



Gambar 22. Distribusi AQI Setelah Kebijakan di Asia Timur

Simulasi kebijakan menggunakan Dash yang dikembangkan tersebut memberikan visualisasi interaktif untuk memprediksi dampak berbagai skenario kebijakan terhadap Air Quality Index (AQI) di Asia Tenggara dan Asia Timur. Dash ini telah dirancang berdasarkan model terbaik, yaitu Ridge Regression. Kemudian, antarmuka ini memungkinkan pengguna untuk secara fleksibel mengatur parameter kebijakan, seperti pengurangan polutan ($PM_{2.5}$, PM_{10} , O_3 , dan NO_2) serta peningkatan fitur strategis, yaitu area vegetasi urban dan persentase energi terbarukan.

Sebagai contoh, diterapkan skenario kebijakan di wilayah Asia Tenggara dan Asia Timur dengan parameter:

- Pengurangan $PM_{2.5}$ sebesar 20%
- Pengurangan PM_{10} , O_3 , dan NO_2 masing-masing sebesar 15%
- Peningkatan area vegetasi urban sebesar 30%
- Peningkatan persentase energi terbarukan sebesar 20%

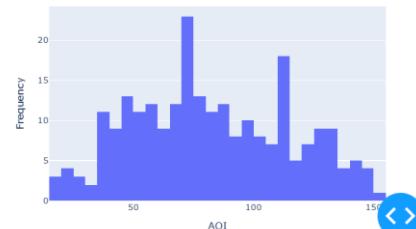


Gambar 23. Distribusi AQI Setelah Kebijakan di Asia Tenggara

Gambar 23 menampilkan distribusi AQI setelah skenario kebijakan di Asia Tenggara yang menunjukkan bagaimana penerapan parameter kebijakan seperti pengurangan konsentrasi polutan ($PM_{2.5}$, PM_{10} , O_3 , NO_2) dan peningkatan

fitur strategis (area vegetasi urban dan persentase energi terbarukan) memengaruhi distribusi nilai AQI. Hasil simulasi ini menunjukkan adanya pergeseran nilai AQI ke arah yang lebih rendah setelah kebijakan diterapkan yang menandakan peningkatan kualitas udara.

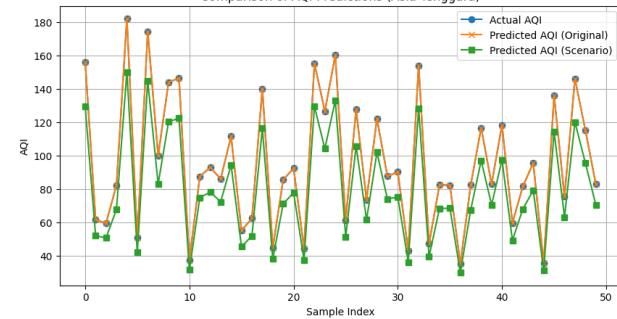
Distribusi AQI Setelah Kebijakan



Gambar 24. Distribusi AQI Setelah Kebijakan di Asia Timur

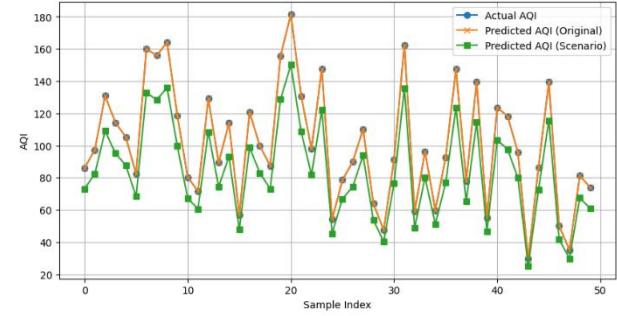
Gambar 24 menunjukkan distribusi AQI setelah skenario kebijakan di Asia Timur, di mana pola serupa terlihat dengan pergeseran nilai AQI ke arah yang lebih baik. Hasil ini mengindikasikan bahwa kebijakan yang diterapkan memberikan dampak positif terhadap pengurangan tingkat polusi udara di wilayah tersebut.

Comparison of AQI Predictions (Asia Tenggara)



Gambar 25. Perbandingan Prediksi AQI Sebelum dan Setelah Kebijakan di Asia Tenggara

Comparison of AQI Predictions (Asia Timur)



Gambar 26. Perbandingan Prediksi AQI Sebelum dan Setelah Kebijakan di Asia Timur

Gambar 25 dan 26 merupakan perbandingan prediksi AQI sebelum dan setelah Kebijakan di Asia Tenggara dan Asia Timur. Grafik ini memvisualisasikan nilai AQI aktual, prediksi original (sebelum kebijakan), dan prediksi skenario (setelah kebijakan). Terlihat bahwa prediksi skenario menunjukkan penurunan AQI yang signifikan dibandingkan dengan prediksi original yang mencerminkan keberhasilan simulasi kebijakan dalam mengurangi tingkat polusi udara.

TABEL 17. PENURUAN AQI SETELAH KEBIJAKAN

Wilayah	Rata-rata penuruan
Asia Tenggara	16.57
Asia Timur	16.72

Pada Tabel 17, ditampilkan rata-rata penurunan AQI di kedua wilayah, yaitu 16.57 untuk Asia Tenggara dan 16.72 untuk Asia Timur. Hal ini menunjukkan bahwa meskipun terdapat variasi kontribusi fitur di kedua wilayah, kebijakan yang diterapkan mampu memberikan dampak yang sebanding.

Simulasi kebijakan yang dilakukan menggunakan model terbaik berbasis Ridge Regression dan antarmuka Dash telah menunjukkan efektivitasnya dalam memvisualisasikan dampak berbagai skenario kebijakan terhadap kualitas udara di Asia Tenggara dan Asia Timur. Berdasarkan hasil simulasi, pengurangan konsentrasi polutan utama ($PM_{2.5}$, PM_{10} , O_3 , NO_2) secara konsisten memberikan dampak positif terhadap penurunan AQI. Selain itu, peningkatan area vegetasi urban dan penggunaan energi terbarukan juga berkontribusi dalam mendukung keberlanjutan lingkungan meskipun pengaruhnya kecil dalam model prediksi.

Hasil ini menegaskan pentingnya penerapan kebijakan berbasis data untuk meningkatkan kualitas udara di kawasan tersebut, serta menunjukkan bahwa Dash sebagai alat simulasi kebijakan dapat memberikan wawasan yang mendalam dan interaktif bagi pembuat kebijakan dalam mengambil langkah strategis untuk mencapai lingkungan yang lebih sehat dan berkelanjutan.

III. PENUTUP

A. Kesimpulan

Dari penelitian ini, didapatkan model prediksi Indeks Kualitas Udara (AQI) untuk mendukung kebijakan lingkungan berbasis data di kawasan Asia Tenggara dan Asia Timur. Dengan memanfaatkan metode Ridge Regression yang telah dioptimalkan (`solver 'auto'`, `alpha 0,01`), model ini memberikan performa prediksi yang sangat baik. Didapatkan metrik nilai akhir Mean Absolute Error (MAE) sebesar 0.002589, Mean Squared Error (MSE) sebesar 0.000009, Root Mean Squared Error (RMSE) sebesar 0.003, Mean Absolute Percentage Error (MAPE) sebesar 0.000031, dan koefisien determinasi (R^2) sebesar 1 untuk Asia Tenggara. Kemudian, Mean Absolute Error (MAE) sebesar 0.002520, Mean Squared Error (MSE) sebesar 0.000009, Root Mean Squared Error (RMSE) sebesar 0.002988, Mean Absolute Percentage Error (MAPE) sebesar 0.000032, dan koefisien determinasi (R^2) sebesar 1 untuk Asia Timur menunjukkan akurasi dan stabilitas yang tinggi dalam memprediksi kualitas udara.

Faktor utama yang mempengaruhi kenaikan AQI di kedua kawasan adalah konsentrasi $PM_{2.5}$ dan PM_{10} . Kemudian kontribusi positif ruang hijau perkotaan membantu menurunkan nilai AQI. Temuan ini memperlihatkan perbedaan dampak urbanisasi dan industrialisasi antara Asia Tenggara dan Asia Timur, di mana ruang hijau memiliki peran lebih signifikan di kawasan yang lebih padat seperti Asia Timur.

Hasil simulasi kebijakan menunjukkan bahwa pendekatan berbasis data dapat memberikan solusi strategis untuk

mengurangi polusi udara, seperti pengaturan emisi, pengelolaan vegetasi, dan perencanaan energi yang lebih efisien. Dengan hasil penelitian ini, diharapkan dapat menjadi referensi untuk mendukung formulasi kebijakan lingkungan yang lebih efektif dan berdampak luas dalam meningkatkan kualitas udara dan kesehatan masyarakat di wilayah Asia Tenggara dan Asia Timur.

DAFTAR PUSTAKA

- [1] World Health Organization. (2019). 7 million premature deaths annually linked to air pollution. Diakses pada https://www.who.int/phe/eNews_63.pdf
- [2] U.S. Environmental Protection Agency. (2024). *Research on health effects from air pollution*. Diakses pada <https://www.epa.gov/air-research/research-health-effects-air-pollution>
- [3] CNBC Indonesia. (2023). Alamak, cuma 8 kota di ASEAN yang udaranya bersih. Diakses pada <https://www.cnbcindonesia.com/news/20230316205623-4-422408/alamak-cuma-8-kota-di-asean-yang-udaranya-bersih-jakarta>
- [4] Open Parliament. (2023). 9 negara berpolusi di Asia Tenggara pada 2022. Diakses pada <https://openparliament.id/2023/09/06/9-negara-berpolusi-di-asia-tenggara-pada-2022/>
- [5] Sindonews. (2024). Meningkatnya Polusi Udara di China Perparah Angka Kematian Masyarakat. Diakses pada <https://international.sindonews.com/read/1403455/45/meningkatnya-polusi-udara-di-china-perparah-angka-kematian-masyarakat-1719364002>
- [6] Qu, Z., Henze, D. K., Worden, H. M., Jiang, Z., Gaubert, B., Theys, N., & Wang, W. (2022). Sector - based top - down estimates of NOx, SO2, and CO emissions in East Asia. *Geophysical research letters*, 49(2), e2021GL096009.
- [7] Hashmi, S. H., Fan, H., Habib, Y., & Riaz, A. (2021). Non-linear relationship between urbanization paths and CO2 emissions: A case of South, South-East and East Asian economies. *Urban Climate*, 37, 100814.
- [8] Sheng, N., Tang, U.W., (2015). The first official ranking city by air quality in China – A review and analysis. *Cities* 51, 139–149.
- [9] Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE access*, 7, 128325–128338.
- [10] Smart Sustainable Cities: An Analysis of Definitions – ITU Report (2014). Diakses pada <https://www.itu.int/en/ITU-T/ssc/Pages/info-ssc.aspx>
- [11] Hasibuan, A., & Sulaiman, O. K. (2019). Smart city, konsep kota cerdas sebagai alternatif penyelesaian masalah perkotaan kabupaten/kota, di kota-kota besar Provinsi Sumatera Utara. *Buletin Utama Teknik*, 14(2), 127–135.
- [12] Sarker, I. H. (2021). Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science*, 2(5), 377.
- [13] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3), 160.
- [14] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.
- [15] Rajan, M. P. (2022). An efficient Ridge regression algorithm with parameter estimation for data analysis in machine learning. *SN Computer Science*, 3(2), 171.
- [16] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: Springer.
- [17] H. Wu and W. Li, Downscaling Land Surface Temperatures Using a Random Forest Regression Model With Multitype Predictor Variables, *IEEE Access*, vol. 7, pp. 21904–21916, 2019. doi: 10.1109/ACCESS.2019.2896241.
- [18] Pawanekar, S. S., Kallimani, J. S., Udgirkar, G., Rashmi, M. R., Sanitha, M. C., & Pin, L. H. (2024). Efficient AQI Prediction: A Comparative Study of Artificial Neural Networks, LSTM, Random Forest, and Gradient Boosting Techniques. In *2024 8th International*

- Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 1597-1603). IEEE.
- [19] Mahesh, T. R., Vinoth Kumar, V., Muthukumaran, V., Shashikala, H. K., Swapna, B., & Guluwadi, S. (2022). Performance analysis of xgboost ensemble methods for survivability with the classification of breast cancer. *Journal of Sensors*, 2022(1), 4649510.
 - [20] Ali, Z. A., Abduljabbar, Z. H., Taher, H. A., Sallow, A. B., & Almufti, S. M. (2023). Exploring the power of eXtreme gradient boosting algorithm in machine learning: A review. Academic Journal of Nawroz University, 12(2), 320-334.
 - [21] Bozdağ, A., Dokuz, Y., & Gökçek, Ö. B. (2020). Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey. *Environmental pollution*, 263, 114635.
 - [22] Kumari, K., & Yadav, S. (2018). Linear regression analysis study. *Journal of the practice of Cardiovascular Sciences*, 4(1), 33-36.
 - [23] Zhang, Y., Zhao, Z., & Zheng, J. (2020). CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. *Journal of Hydrology*, 588, 125087.
 - [24] Prokhortenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.
 - [25] Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. *Chemosphere*, 338, 139518.
 - [26] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
 - [27] McCarty, D. A., Kim, H. W., & Lee, H. K. (2020). Evaluation of light gradient boosted machine learning technique in large scale land use and land cover classification. *Environments*, 7(10), 84.
 - [28] Melkumova, L. E., & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia engineering*, 201, 746-755.
 - [29] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2), 301-320.
 - [30] Amansyah, I., Indra, J., Nurlaelasari, E., & Juwita, A. R. (2024). Prediksi Penjualan Kendaraan Menggunakan Regresi Linear: Studi Kasus pada Industri Otomotif di Indonesia. *Innovative: Journal Of Social Science Research*, 4(4), 1199-1216.
 - [31] Murukonda, V. S. N. M., & Gogineni, A. C. (2022). Prediction of air quality index using supervised machine learning.
 - [32] Pande, C. B., Radhadevi, L., & Satyanarayana, M. B. (2024). Evaluation of machine learning and deep learning models for daily air quality index prediction in Delhi city, India. *Environmental Monitoring and Assessment*, 196(12), 1215.