**LAPORAN PEMODELAN REGRESI LINIER BERGANDA UNTUK MEMPREDIKSI KONSENTRASI CO BERDASARKAN DATA KUALITAS UDARA**

Oleh : Farid Al Farizi

NIM : 202410370110017

Kelas : J



**UNIVERSITAS MUHAMMADIYAH MALANG**

**FAKULTAS TEKNIK**

**PRODI INFORMATIKA**

**2025**

###### DAFTAR ISI

Daftar Isi

[BAB I AKUISISI DATA DAN PEMAHAMAN 1](#_Toc216549013)

[1.1 Perintah Akuisisi Data 1](#_Toc216549014)

[1.1.1 Perolehan Data 1](#_Toc216549015)

[1.1.2 Pemuatan Data 2](#_Toc216549016)

[1.1.3 Identifikasi Variabel 3](#_Toc216549017)

[1.2 Tugas Analisis Awal 5](#_Toc216549018)

[1.2.1 Tampilkan 10 Baris Pertama 5](#_Toc216549019)

[1.2.2 Laporkan persentase missing values untuk setiap kolom 6](#_Toc216549020)

[BAB II ANALISIS DATA EKSPLORATIF (EDA) 8](#_Toc216549021)

[2.1 Perintah Distribusi 8](#_Toc216549022)

[2.1.1 Distribusi Target: Histogram 8](#_Toc216549023)

[2.1.2 Deteksi Outliers 10](#_Toc216549024)

[2.2 Perintah Korelasi 12](#_Toc216549025)

[2.2.1 Korelasi Numerik: Hitung matriks korelasi 13](#_Toc216549026)

[2.2.2 Visualisasi: Heatmap Matriks Korelasi dan Identifikasi Prediktor Terkuat 14](#_Toc216549027)

[BAB III PREPROCESSING & FEATURE ENGINEERING 19](#_Toc216549028)

[3.1 Pembersihan Data 19](#_Toc216549029)

[3.1.1 Penanganan Missing Values 19](#_Toc216549030)

[3.2 Encoding dan Pembagian Data 20](#_Toc216549031)

[3.2.1 One-Hot Encoding 20](#_Toc216549032)

[BAB IV ANALISIS REGRESI & INTERPRETASI 22](#_Toc216549033)

[4.1 Perintah Pelatihan Model 22](#_Toc216549034)

[4.1.1 Implementasi: Inisialisasi dan Pelatihan Model Regresi Linier Berganda. 22](#_Toc216549035)

[4.1.2 Prediksi 22](#_Toc216549036)

[4.2 Perintah Evaluasi Model 23](#_Toc216549037)

[4.2.1 Hitung 4 Metrik Evaluasi pada Data Testing 23](#_Toc216549038)

[4.3 Perintah Interpretasi Koefisien 25](#_Toc216549039)

[4.3.1 Tampilkan semua koefisien 25](#_Toc216549040)

[4.3.2 Interpretasi 26](#_Toc216549041)

[4.4 Perintah Visualisasi Hasil 28](#_Toc216549042)

[4.4.1 Scatter Plot 28](#_Toc216549043)

[BAB V LAPORAN AKHIR 31](#_Toc216549044)

[5.1 Pendahuluan 31](#_Toc216549045)

[5.1.1 Latar Belakang 31](#_Toc216549046)

[5.1.2 Tujuan 31](#_Toc216549047)

[5.1.3 Identitas Dataset 31](#_Toc216549048)

[5.2 Metodologi 32](#_Toc216549049)

[5.2.1 Detail Langkah EDA dan Preprocessing 32](#_Toc216549050)

[5.3 Hasil dan Pembahasan Model 34](#_Toc216549051)

[5.3.1 Tabel ringkasan Metrik Evaluasi. 34](#_Toc216549052)

[5.3.2 Koefisien Determinasi 34](#_Toc216549053)

[5.3.3 Interpretasi Koefisien 35](#_Toc216549054)

[5.4 Kesimpulan dan Saran 36](#_Toc216549055)

# AKUISISI DATA DAN PEMAHAMAN

## Perintah Akuisisi Data

Akuisisi data pada penelitian ini dilakukan melalui beberapa tahapan, yaitu perolehan data dari sumber sekunder, proses perapian dan prapengolahan data, serta pemuatan data ke dalam perangkat lunak pengolah data. Data yang digunakan berupa berkas AirQuality.xlsx, yaitu versi data yang telah dirapikan dari dataset awal AirQuality. Dataset ini memuat pengukuran kualitas udara dan variabel meteorologi yang diambil secara berkala.

Setelah data mentah diperoleh dari sumber daring (Kaggle), peneliti terlebih dahulu melakukan proses pembersihan dan penyesuaian format di Microsoft Excel, seperti menghapus baris dengan nilai tidak lengkap, menstandarkan tipe data numerik, serta memilih variabel yang relevan untuk analisis. Data yang telah rapi kemudian disimpan dalam format Excel (AirQuality.xlsx) dan selanjutnya dimuat ke dalam lingkungan kerja Python menggunakan pustaka pandas untuk keperluan eksplorasi, pembersihan lanjutan, serta pemodelan regresi pada tahap berikutnya.

### Perolehan Data

Data penelitian diperoleh dari dataset kualitas udara AirQuality yang tersedia secara daring sebagai data sekunder. Dataset awal disediakan dalam format CSV dan berisi deretan pengamatan kualitas udara yang diukur secara berkala, meliputi konsentrasi beberapa gas polutan serta kondisi lingkungan seperti suhu dan kelembapan udara.

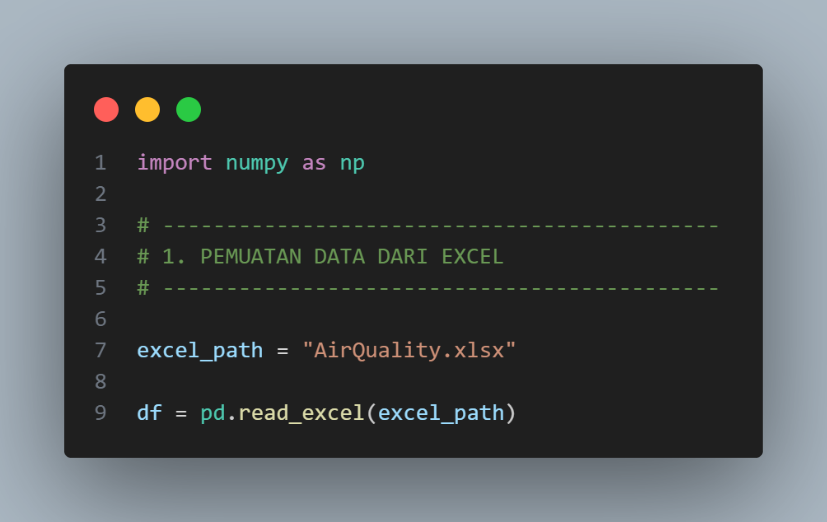
Proses perolehan dan perapian data dilakukan melalui langkah-langkah berikut:

1. Mengunduh dataset AirQuality dari sumber dataset daring dan menyimpannya dalam format awal (AirQuality.csv).
2. Membuka dataset tersebut di Microsoft Excel untuk melakukan proses kerapian data, antara lain:
   1. Menghapus atau menandai baris dengan nilai tidak wajar atau nilai kode hilang (misalnya nilai tertentu yang merepresentasikan data tidak tersedia).
   2. Memastikan setiap kolom numerik tersimpan sebagai tipe data numerik yang konsisten.
   3. Memilih dan mempertahankan hanya variabel-variabel yang relevan dengan tujuan penelitian.
3. Menyimpan hasil data yang telah dirapikan dalam format Excel dengan nama berkas: AirQuality.xlsx.

Dengan demikian, seluruh proses perolehan data bersifat sekunder, karena peneliti tidak melakukan pengukuran langsung di lapangan, tetapi memanfaatkan data yang sudah tersedia, kemudian melakukan proses kerapian (cleaning) dan penyesuaian format sebelum digunakan pada analisis lebih lanjut.

### Pemuatan Data

Setelah berkas AirQuality.xlsx diperoleh dan dirapikan, langkah selanjutnya adalah melakukan pemuatan data ke dalam lingkungan kerja Python agar dapat diolah lebih lanjut. Proses pemuatan data dilakukan menggunakan pustaka pandas. Perintah akuisisi data yang digunakan dapat dituliskan sebagai berikut:



Pada perintah di atas, fungsi read\_excel() digunakan karena data telah disimpan dalam format Excel (.xlsx) setelah melalui proses kerapian di Microsoft Excel. Setelah data berhasil dimuat, peneliti dapat melakukan pengecekan awal terhadap struktur data, seperti nama kolom, jumlah baris, serta keberadaan nilai hilang (missing value), sebelum masuk ke tahap pembersihan lanjutan dan analisis regresi.

### Identifikasi Variabel

Identifikasi variabel dilakukan untuk menentukan variabel target yang akan diprediksi serta variabel prediktor yang diduga berpengaruh terhadap variabel target. Tahap ini dilakukan setelah data selesai dirapikan sehingga hanya variabel yang relevan yang digunakan dalam pemodelan. Identifikasi variabel yang jelas penting agar model regresi yang dibangun memiliki dasar yang kuat dan sesuai dengan tujuan penelitian.

**Variabel Target**

Dalam penelitian ini, variabel target (variabel dependen) adalah:

CO(GT)  
Merupakan konsentrasi gas carbon monoxide (CO) yang terukur di udara ambien. Nilai ini digunakan sebagai indikator utama tingkat pencemaran udara oleh gas CO. Secara notasi, variabel target dapat dinyatakan sebagai:

Variabel CO(GT) dipilih sebagai target karena CO merupakan salah satu gas polutan yang berdampak signifikan terhadap kesehatan manusia dan kualitas lingkungan. Oleh karena itu, memprediksi konsentrasi CO berdasarkan kondisi lingkungan dan pembacaan sensor lainnya menjadi penting untuk mendukung pemantauan kualitas udara.

**Variabel Prediktor**

Variabel prediktor (variabel independen) adalah variabel-variabel yang diduga memengaruhi besarnya konsentrasi CO di udara. Berdasarkan hasil seleksi pada data yang telah dirapikan, beberapa variabel prediktor yang digunakan dalam model antara lain:

1. PT08.S1(CO)  
   Pembacaan sensor 1 yang sensitif terhadap gas CO. Sensor ini merepresentasikan respon instrumen terhadap keberadaan gas CO di udara.
2. C6H6(GT)  
   Konsentrasi benzena (C₆H₆) di udara. Benzena merupakan salah satu komponen polutan yang sering berkorelasi dengan emisi kendaraan dan aktivitas industri sehingga berpotensi berkaitan dengan pola konsentrasi CO.
3. PT08.S2(NMHC)  
   Pembacaan sensor 2 yang sensitif terhadap non-methane hydrocarbons (NMHC). Keberadaan hidrokarbon ini biasanya berkaitan dengan aktivitas pembakaran dan emisi kendaraan.
4. T (Temperature)  
   Suhu udara (°C). Suhu dapat memengaruhi dispersi polutan dan reaksi kimia di atmosfer, sehingga berpotensi memengaruhi konsentrasi CO yang terukur.
5. RH (Relative Humidity)  
   Kelembapan relatif (%). Kondisi kelembapan berperan dalam proses fisik dan kimia di atmosfer yang dapat memengaruhi akumulasi, pengenceran, maupun transformasi gas polutan.
6. AH (Absolute Humidity)  
   Kelembapan absolut. Menggambarkan jumlah uap air aktual di udara dan dapat terkait dengan karakteristik dispersi polutan di atmosfer.

Secara umum, hubungan antara variabel target dan variabel prediktor dalam model regresi linier berganda yang dibangun dalam penelitian ini dapat dituliskan sebagai berikut:

dengan keterangan:

: konsentrasi CO (CO(GT)),

: intercept,

: koefisien regresi masing-masing variabel prediktor,

: komponen error yang merepresentasikan faktor-faktor lain di luar model.

Model ini selanjutnya akan digunakan untuk mengkaji seberapa besar kontribusi masing-masing variabel prediktor terhadap konsentrasi CO, sekaligus mengevaluasi kemampuan model dalam memprediksi kualitas udara berdasarkan variabel-variabel yang telah dipilih.

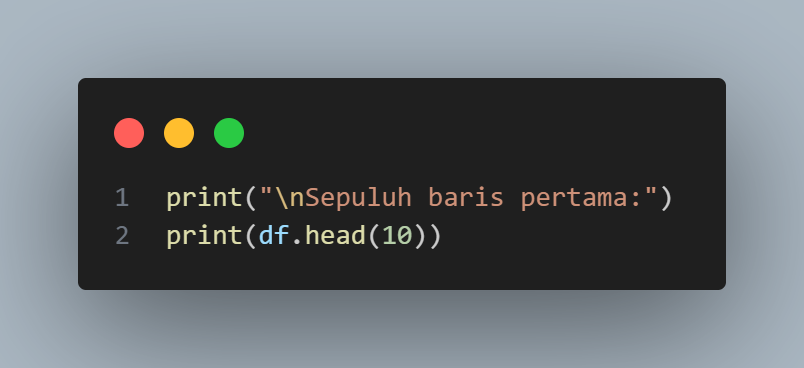
## Tugas Analisis Awal

Analisis awal dilakukan untuk memperoleh gambaran umum mengenai struktur dan kualitas data sebelum dilakukan pemodelan lebih lanjut. Pada tahap ini, peneliti menampilkan beberapa baris pertama dari dataset untuk memastikan bahwa proses pemuatan data telah berjalan dengan benar, serta menghitung persentase nilai hilang (missing values) pada setiap kolom. Informasi ini penting untuk menilai kelayakan data dan menentukan strategi penanganan data hilang pada tahap prapengolahan (preprocessing).

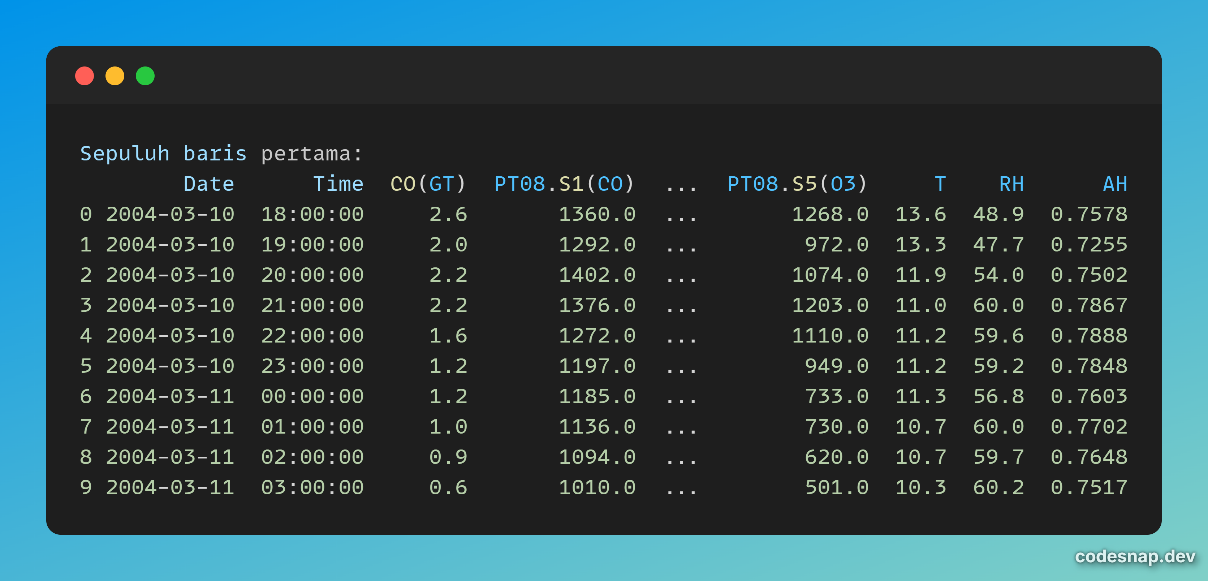
### Tampilkan 10 Baris Pertama

Langkah pertama dalam analisis awal adalah menampilkan sejumlah baris teratas dari dataset untuk melakukan inspeksi visual terhadap isi data. Dengan melihat sepuluh baris pertama, peneliti dapat memverifikasi bahwa setiap kolom telah terbaca dengan benar, tipe data sudah sesuai, dan tidak terjadi pergeseran kolom akibat kesalahan pemisah (delimiter) atau format saat pemuatan data dari berkas Excel.

Perintah Python yang digunakan untuk menampilkan 10 baris pertama adalah sebagai berikut:



Output:

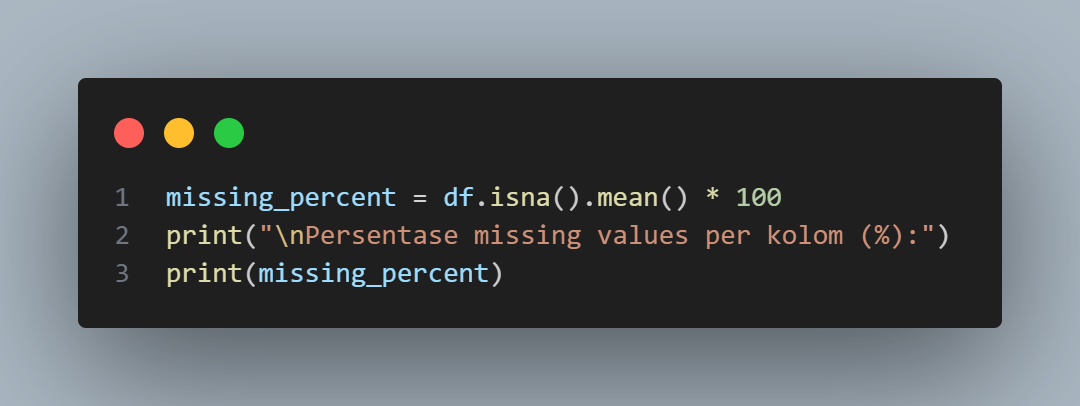


Perintah df.head(10) akan menghasilkan tampilan sepuluh observasi pertama dari dataset AirQuality.xlsx. Dari keluaran tersebut, peneliti dapat mengamati nama kolom, rentang nilai awal, serta memastikan bahwa variabel-variabel seperti CO(GT), PT08.S1(CO), C6H6(GT), PT08.S2(NMHC), T, RH, dan AH telah terbaca dengan benar dan siap digunakan pada analisis lanjutan.

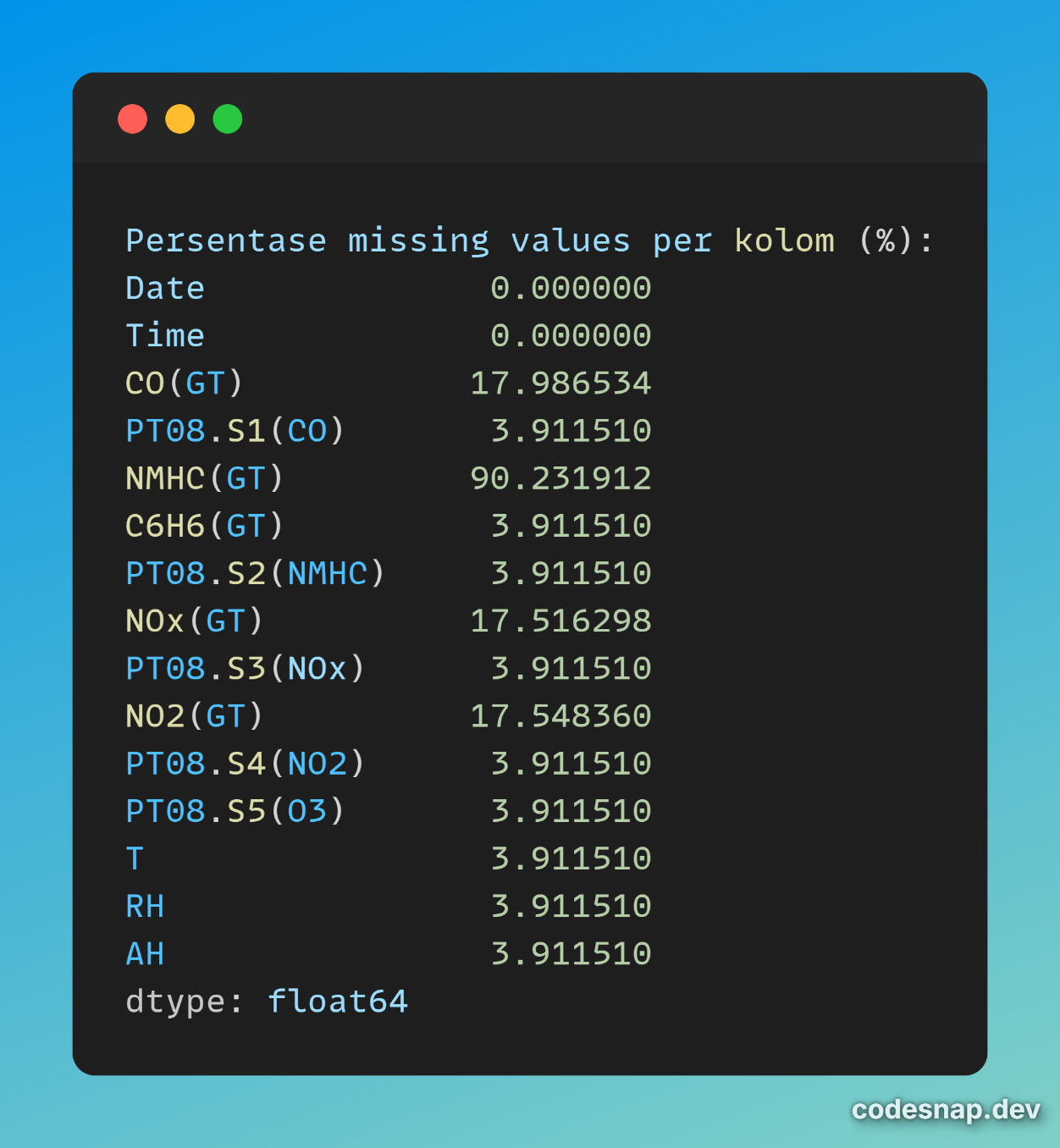
### Laporkan persentase missing values untuk setiap kolom

Tahap berikutnya adalah mengidentifikasi keberadaan nilai hilang (missing values) pada setiap kolom. Nilai hilang yang tidak ditangani dengan baik dapat memengaruhi hasil estimasi parameter regresi dan menurunkan kualitas model. Oleh karena itu, diperlukan perhitungan persentase missing values untuk setiap variabel agar peneliti dapat memutuskan apakah suatu kolom masih layak digunakan, perlu dilakukan imputasi, atau harus dikeluarkan dari analisis.

Perhitungan persentase missing values untuk setiap kolom dilakukan dengan menggunakan kombinasi fungsi isna(), mean(), dan dikalikan dengan 100 untuk memperoleh nilai dalam bentuk persen. Perintah Python yang digunakan adalah:



Output:



Perintah df.isna().mean() akan menghasilkan proporsi nilai hilang pada setiap kolom (antara 0 dan 1), kemudian dikalikan 100 untuk mengubahnya menjadi persen. Hasil yang diperoleh dapat disajikan dalam bentuk tabel, sehingga terlihat dengan jelas kolom mana yang memiliki persentase missing value rendah (misalnya di bawah 5%) dan kolom mana yang memiliki persentase missing value tinggi.

# ANALISIS DATA EKSPLORATIF (EDA)

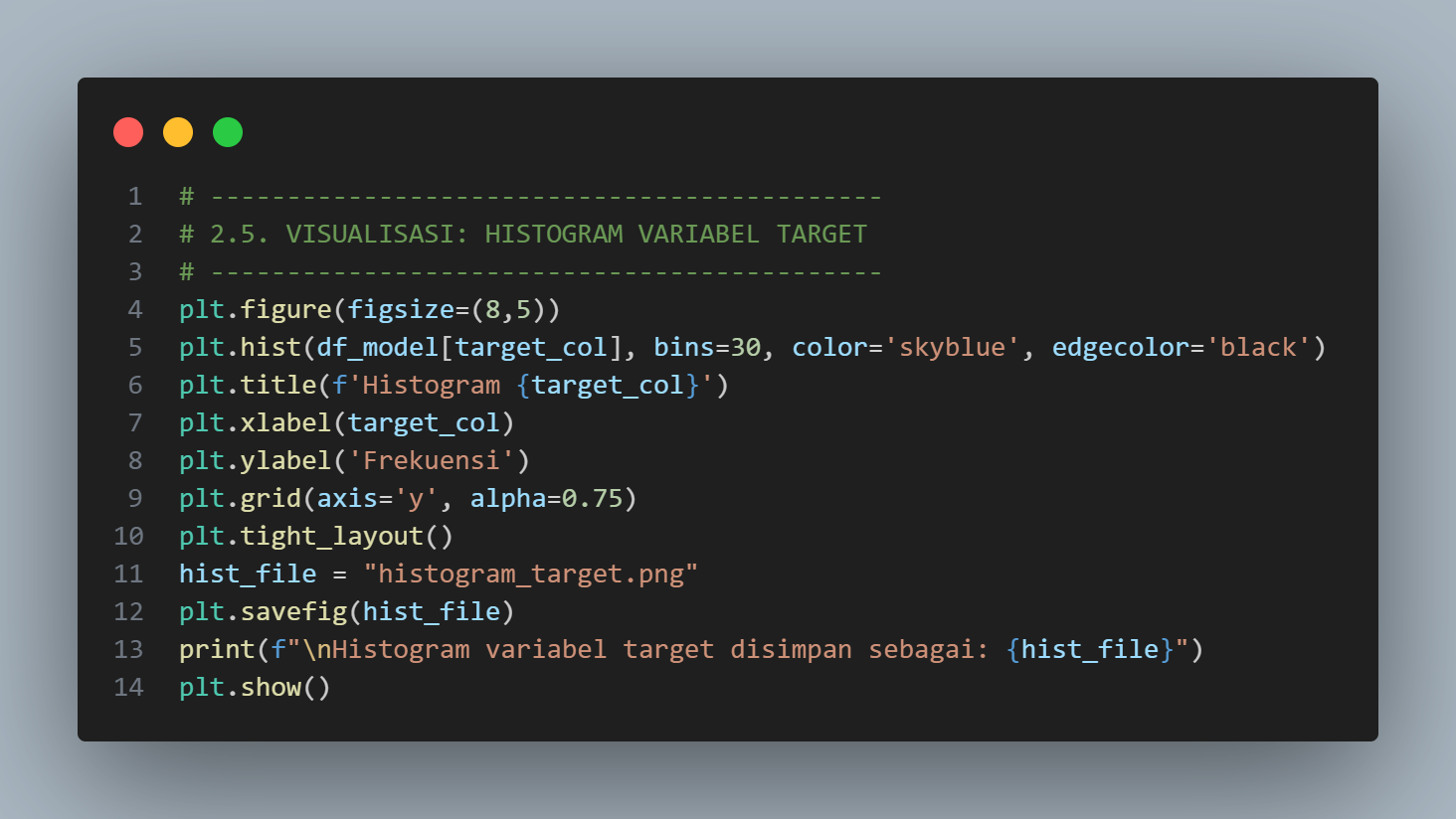
## Perintah Distribusi

Analisis distribusi dilakukan untuk memahami pola sebaran data, khususnya pada variabel target yang akan diprediksi. Tahap ini mencakup pembuatan histogram untuk melihat bentuk distribusi, pengujian apakah distribusi tersebut condong (skewed) ke kiri atau ke kanan, serta deteksi adanya nilai pencilan (outlier) menggunakan box plot. Informasi ini penting untuk menilai kelayakan penerapan regresi linier serta menentukan apakah diperlukan transformasi data atau penanganan khusus terhadap outlier.

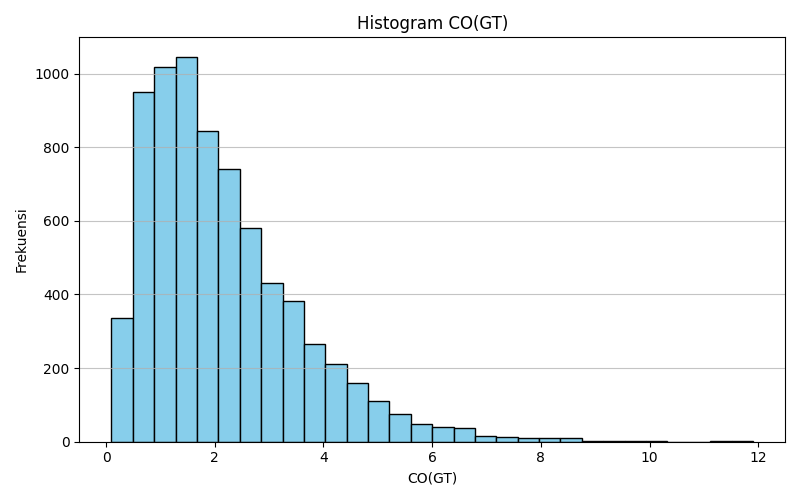
### Distribusi Target: Histogram

Pada subbagian ini, distribusi variabel target CO(GT) dianalisis dengan menggunakan histogram. Histogram memberikan gambaran visual mengenai frekuensi kemunculan nilai-nilai CO pada rentang tertentu, sehingga dapat dilihat apakah data terkonsentrasi pada nilai rendah, sedang, atau tinggi.

Perintah Python yang digunakan untuk membuat histogram variabel target adalah sebagai berikut (menggunakan df\_model yang sudah dibersihkan pada tahap sebelumnya):



Output:



Dari histogram tersebut, peneliti dapat mengamati apakah data CO(GT) cenderung menumpuk pada kisaran nilai tertentu dan apakah terdapat ekor distribusi yang memanjang ke arah nilai rendah atau nilai tinggi.

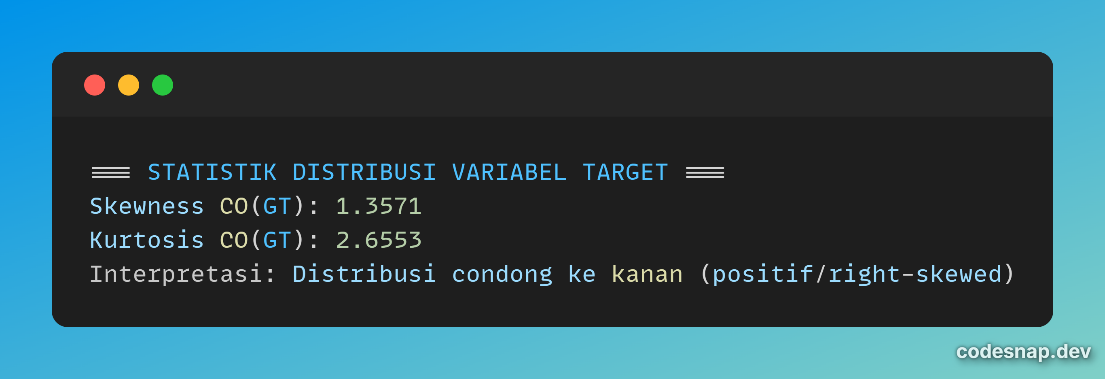
#### Distribusi Skewed

Untuk menilai apakah distribusi variabel CO(GT) bersifat simetris atau condong (skewed), selain melalui pengamatan visual histogram, dilakukan juga perhitungan nilai skewness secara kuantitatif. Nilai skewness dapat dihitung menggunakan fungsi skew() dari pandas.

Perintah Python yang digunakan adalah sebagai berikut:



Output:



Secara umum, interpretasi nilai skewness adalah sebagai berikut:

1. Skewness mendekati 0 → distribusi relatif simetris.
2. Skewness positif ( > 0 ) → distribusi right-skewed (ekor memanjang ke kanan, banyak nilai rendah, sedikit nilai sangat tinggi).
3. Skewness negatif ( < 0 ) → distribusi left-skewed (ekor memanjang ke kiri, banyak nilai tinggi, sedikit nilai sangat rendah).
4. Secara praktis:
   1. → skewness lemah
   2. → skewness sedang
   3. → skewness kuat

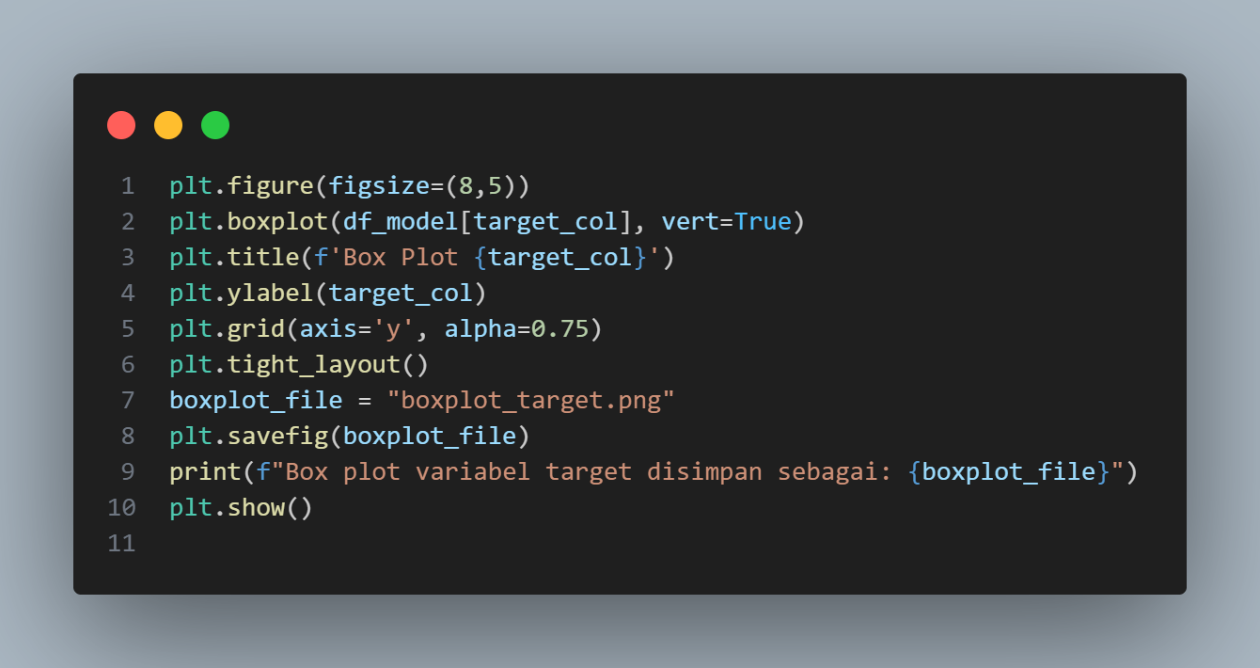
Berdasarkan hasil perhitungan, nilai skewness untuk variabel CO(GT) sebesar 1,3571. Nilai tersebut menunjukkan bahwa distribusi CO(GT) bersifat positif dan termasuk kategori kuat, sehingga sebaran data cenderung right skewed.

### Deteksi Outliers

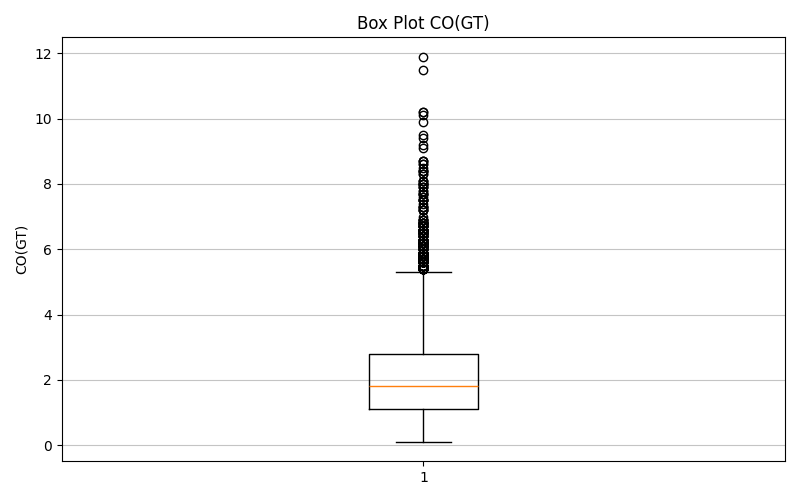
Selain bentuk distribusi secara umum, penting juga untuk mengidentifikasi apakah terdapat outlier, yaitu nilai-nilai yang berada jauh di luar pola sebaran mayoritas data. Kehadiran outlier dapat mempengaruhi estimasi koefisien regresi dan menurunkan kinerja model, sehingga perlu dideteksi sejak awal.

Deteksi awal outlier dapat dilakukan secara visual melalui boxplot. Box plot menampilkan median, kuartil pertama (Q1), kuartil ketiga (Q3), serta nilai-nilai yang berada di luar rentang dan , di mana .

Perintah Python untuk membuat box plot variabel target CO(GT) adalah sebagai berikut:



Output:



#### Box Plot

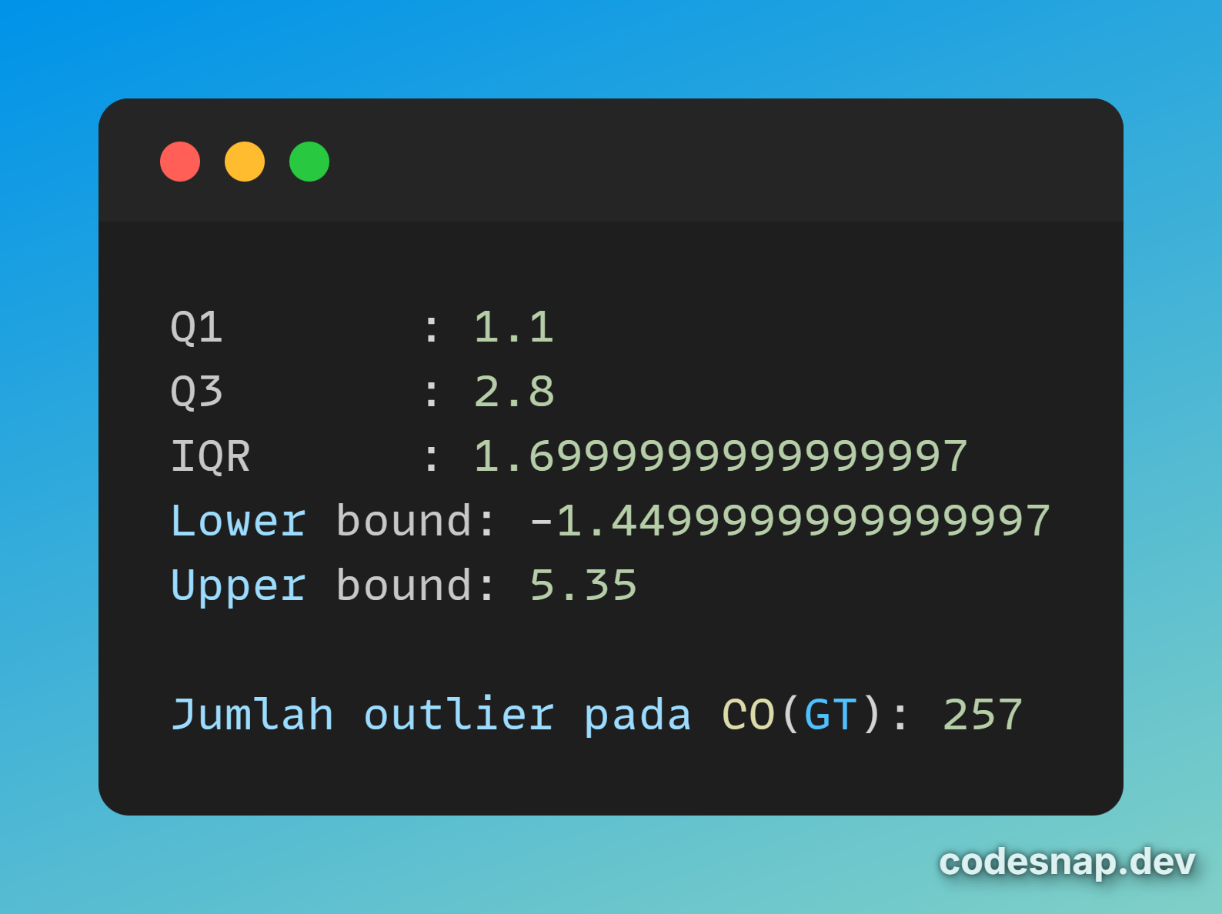
Agar deteksi outlier tidak hanya bersifat visual, peneliti juga dapat menggunakan metode Interquartile Range (IQR) untuk menentukan secara kuantitatif data mana yang tergolong outlier. Rumusnya:

1. Batas bawah (lower bound) :
2. Batas atas (upper bound) :

Observasi di bawah batas bawah atau di atas batas atas dikategorikan sebagai outlier. Perintah Python untuk menghitung batas IQR dan menghitung jumlah outlier pada CO(GT):



Output:



Berdasarkan perhitungan IQR, diperoleh nilai Q1 sebesar 1,1 dan Q3 sebesar 2,8. Dengan demikian, IQR yang dihasilkan adalah 1.6999999999999997, sehingga batas bawah dan batas atas untuk mendeteksi outlier masing-masing adalah -1.4499999999999997 dan 5,35. Hasil perhitungan menunjukkan bahwa terdapat 257 observasi yang berada di luar rentang tersebut, sehingga dikategorikan sebagai outlier pada variabel CO(GT).

## Perintah Korelasi

Analisis korelasi dilakukan untuk melihat hubungan linier antara variabel target dan variabel-variabel prediktor yang digunakan dalam model. Informasi ini penting untuk:

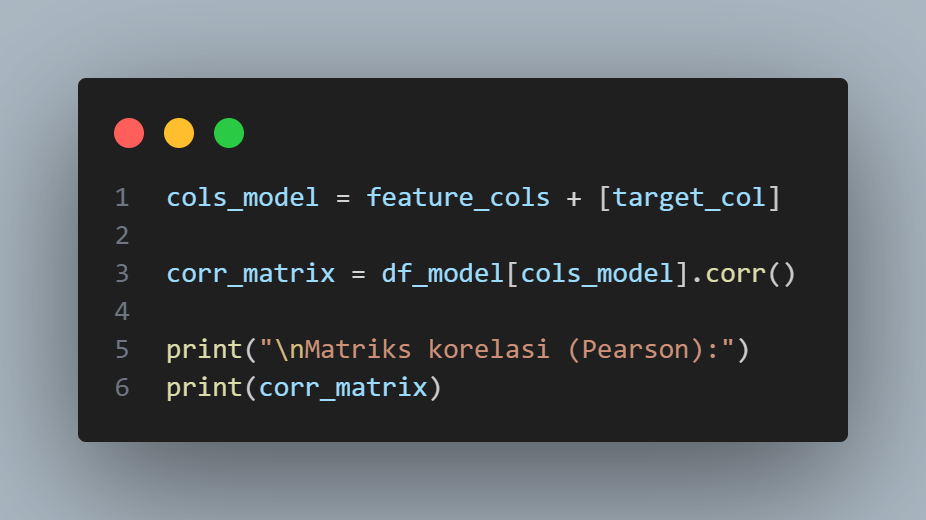
1. Mengidentifikasi prediktor mana yang memiliki hubungan paling kuat dengan variabel target.
2. Menilai potensi adanya multikolinearitas antar prediktor, yang dapat memengaruhi stabilitas estimasi koefisien regresi.

Pada tahap ini, terlebih dahulu dihitung matriks korelasi numerik, kemudian hasilnya divisualisasikan dalam bentuk heatmap untuk memudahkan interpretasi pola hubungan antar variabel.

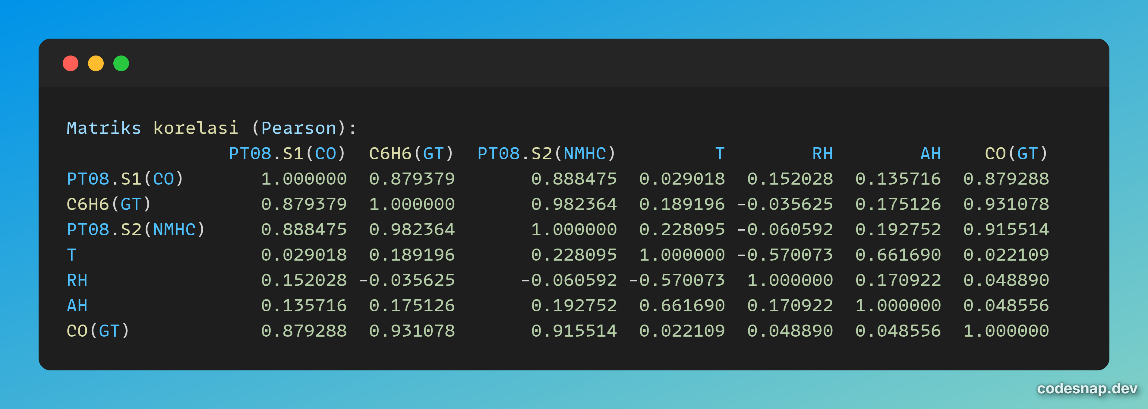
### Korelasi Numerik: Hitung matriks korelasi

Matriks korelasi dihitung menggunakan koefisien korelasi Pearson antar seluruh variabel numerik yang terlibat dalam model, yaitu variabel target CO(GT) dan variabel-variabel prediktor: PT08.S1(CO), C6H6(GT), PT08.S2(NMHC), T, RH, dan AH.

Perintah Python yang digunakan:



Output:



Matriks korelasi menunjukkan besaran dan arah hubungan linier antara variabel target CO(GT) dan variabel-variabel prediktor. Nilai korelasi yang mendekati +1 mengindikasikan hubungan positif yang kuat, sedangkan nilai yang mendekati −1 mengindikasikan hubungan negatif yang kuat. Nilai mendekati 0 menunjukkan hubungan linier yang lemah atau tidak ada.

Matriks korelasi Pearson dihitung untuk variabel target CO(GT) dan seluruh variabel prediktor yang digunakan dalam model, yaitu PT08.S1(CO), C6H6(GT), PT08.S2(NMHC), T, RH, dan AH. Hasil perhitungan menunjukkan bahwa tiga variabel sensor, yaitu C6H6(GT), PT08.S2(NMHC), dan PT08.S1(CO) memiliki korelasi positif yang sangat kuat terhadap CO(GT), masing-masing dengan nilai korelasi sekitar 0,931, 0,916, dan 0,879. Nilai korelasi yang tinggi ini mengindikasikan bahwa ketika pembacaan sensor-sensor tersebut meningkat, konsentrasi CO yang terukur di udara juga cenderung meningkat.

Sebaliknya, variabel-variabel meteorologi seperti T (Temperature), RH (Relative Humidity), dan AH (Absolute Humidity) menunjukkan nilai korelasi yang sangat rendah terhadap CO(GT), yaitu sekitar 0,022, 0,049, dan 0,049. Nilai tersebut berada sangat dekat dengan nol, sehingga hubungan linier antara kondisi suhu/kelembapan dengan konsentrasi CO dapat dikatakan lemah atau hampir tidak ada secara linier. Dengan demikian, secara numerik dapat disimpulkan bahwa variasi CO(GT) lebih banyak dijelaskan oleh perubahan pembacaan sensor gas dibandingkan oleh variasi faktor meteorologi.

Selain melihat hubungan antara variabel prediktor dan variabel target, matriks korelasi juga menunjukkan adanya hubungan yang sangat kuat di antara beberapa variabel prediktor itu sendiri. Korelasi antara C6H6(GT) dan PT08.S2(NMHC) mencapai sekitar 0,98, sedangkan korelasi antara PT08.S1(CO) dengan C6H6(GT) dan PT08.S2(NMHC) masing-masing sebesar kurang lebih 0,88 dan 0,89. Nilai korelasi yang sangat tinggi antar prediktor ini mengindikasikan adanya potensi multikolinearitas, yaitu kondisi di mana beberapa variabel penjelas saling berkorelasi kuat satu sama lain. Dalam konteks pemodelan regresi linier, multikolinearitas dapat menyebabkan koefisien regresi menjadi kurang stabil dan sulit diinterpretasikan secara terpisah. Meskipun demikian, pada penelitian ini variabel-variabel tersebut tetap digunakan karena secara substantif mewakili sensor-sensor gas yang saling berkaitan dalam sistem pengukuran kualitas udara, dan fokus utama analisis adalah pada kemampuan model dalam memprediksi konsentrasi CO secara keseluruhan.

### Visualisasi: Heatmap Matriks Korelasi dan Identifikasi Prediktor Terkuat

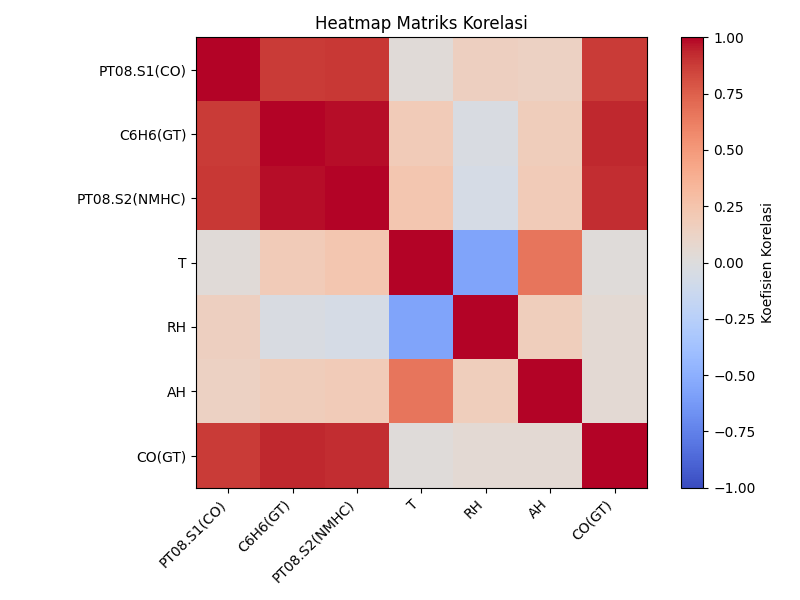
Untuk memudahkan interpretasi hubungan antar variabel, matriks korelasi divisualisasikan dalam bentuk heatmap. Pada heatmap, setiap sel merepresentasikan nilai korelasi antara sepasang variabel, dengan intensitas warna yang menggambarkan kekuatan hubungan.

Visualisasi Heatmap

Perintah Python (menggunakan matplotlib):



Output:

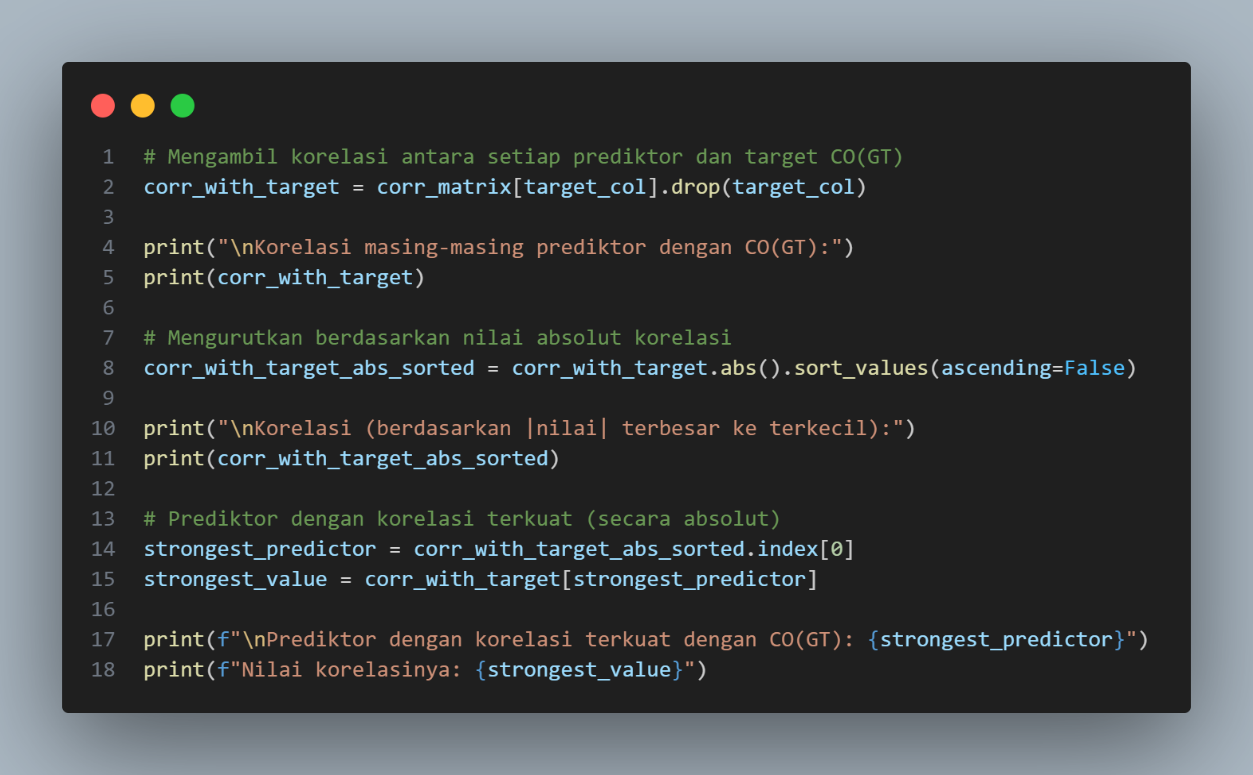


Heatmap matriks korelasi memberikan gambaran visual mengenai kekuatan dan arah hubungan antar variabel. Warna yang lebih “terang/gelap” (sesuai skala warna) menunjukkan korelasi yang lebih kuat, baik positif maupun negatif, sedangkan warna yang mendekati tengah menunjukkan korelasi yang lemah.

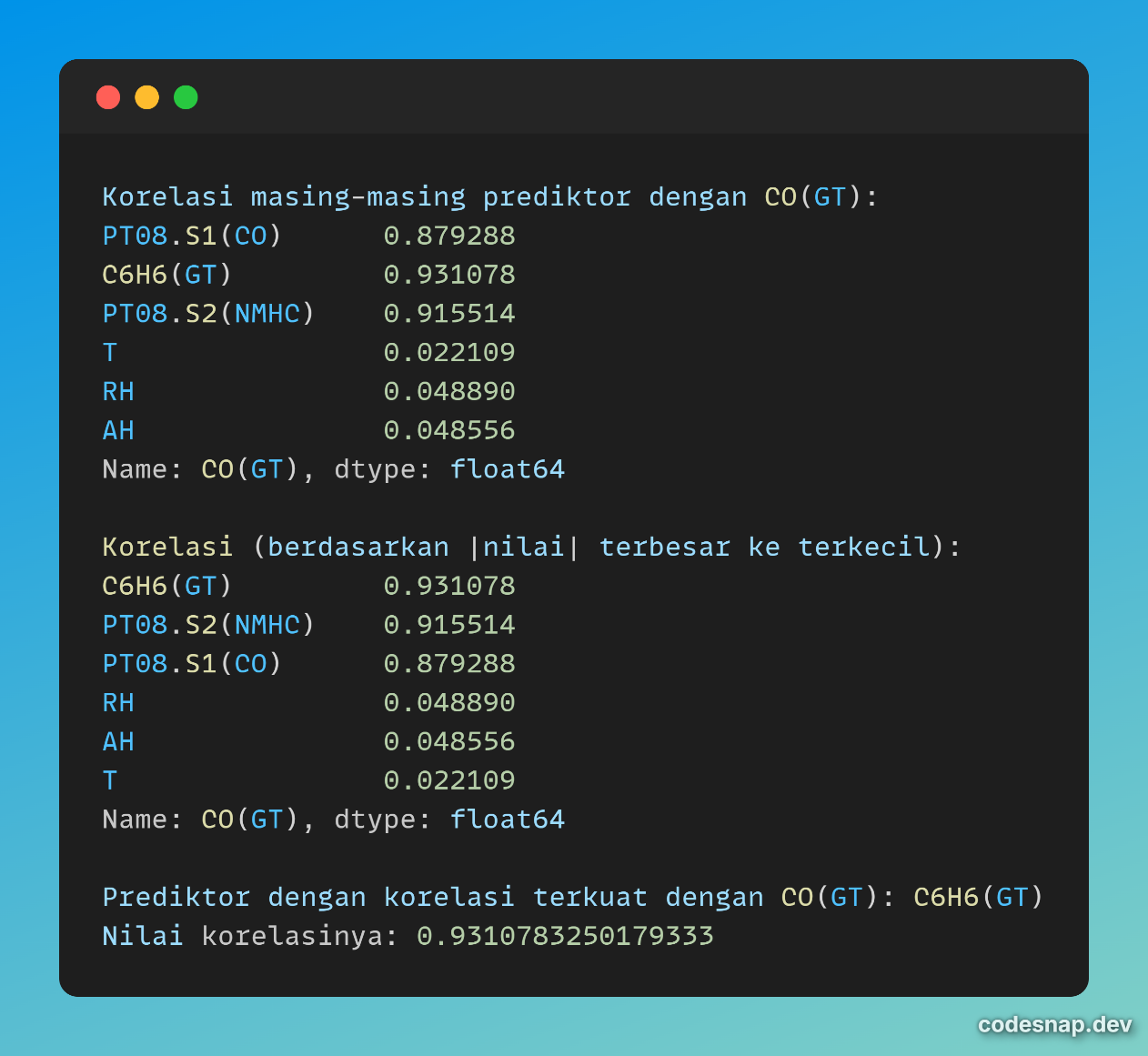
Identifikasi Prediktor yang Paling Kuat Korelasinya dengan Variabel Target

Untuk menentukan prediktor mana yang paling kuat korelasinya dengan variabel target CO(GT), dilakukan perhitungan korelasi Pearson antara CO(GT) dan setiap variabel prediktor, kemudian diurutkan berdasarkan nilai absolut korelasinya (supaya korelasi negatif yang kuat tetap terdeteksi).

Perintah Python:



Output:



Matriks korelasi yang telah dihitung kemudian divisualisasikan dalam bentuk heatmap untuk memudahkan interpretasi pola hubungan antar variabel. Pada heatmap tersebut terlihat bahwa sel-sel yang merepresentasikan pasangan CO(GT) dengan C6H6(GT), PT08.S2(NMHC), dan PT08.S1(CO) memiliki intensitas warna yang paling kuat dibandingkan pasangan variabel lainnya, sejalan dengan nilai korelasi numerik yang mendekati +1. Sementara itu, sel yang merepresentasikan hubungan antara CO(GT) dengan variabel T, RH, dan AH tampak dengan warna yang mendekati tengah skala, menandakan korelasi yang lemah.

Berdasarkan pengurutan nilai korelasi absolut antara setiap prediktor dan variabel target CO(GT), diperoleh bahwa prediktor dengan korelasi terkuat adalah C6H6(GT) dengan nilai korelasi sekitar 0,9311. Nilai ini menunjukkan adanya hubungan linier positif yang sangat kuat: semakin tinggi konsentrasi benzena (C6H6) yang terukur, semakin tinggi pula konsentrasi CO di udara ambien. Di bawahnya, PT08.S2(NMHC) dan PT08.S1(CO) juga memberikan kontribusi korelasi yang sangat kuat terhadap CO(GT). Hasil ini mengindikasikan bahwa variabel-variabel sensor gas merupakan kandidat prediktor utama dalam pemodelan regresi untuk memprediksi konsentrasi CO, sedangkan variabel meteorologi berperan relatif lebih kecil dalam menjelaskan variasi CO(GT) pada dataset yang digunakan.

# PREPROCESSING & FEATURE ENGINEERING

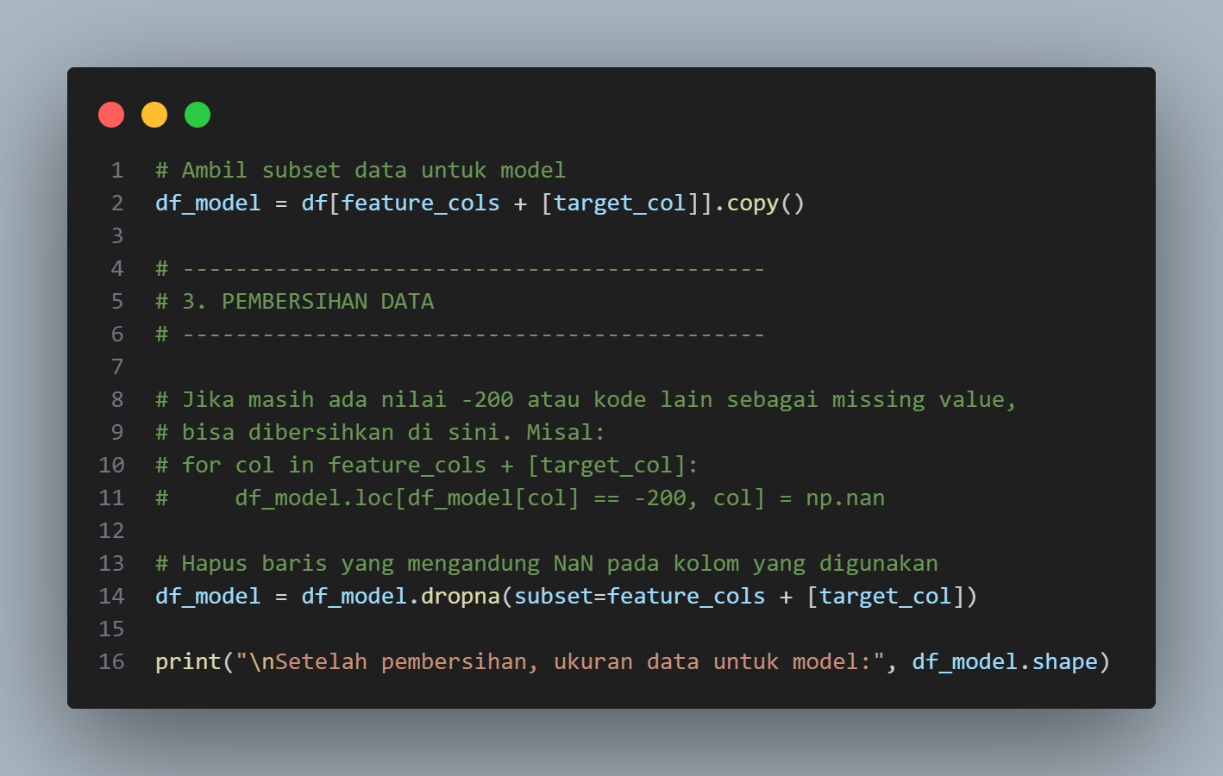
## Pembersihan Data

Pembersihan data dilakukan untuk memastikan bahwa data yang masuk ke dalam model regresi bebas dari nilai hilang dan inkonsistensi yang dapat mengganggu proses estimasi parameter. Pada tahap ini, fokus utama adalah penanganan missing values pada variabel-variabel yang digunakan dalam pemodelan, yaitu CO(GT) sebagai variabel target serta PT08.S1(CO), C6H6(GT), PT08.S2(NMHC), T, RH, dan AH sebagai variabel prediktor.

### Penanganan Missing Values

Berdasarkan hasil analisis awal, masih dimungkinkan terdapat nilai hilang (missing values) pada sebagian observasi, baik karena data tidak tercatat maupun karena sebelumnya direpresentasikan oleh kode khusus. Untuk mencegah bias dan error pada saat pelatihan model, dilakukan pembersihan data dengan cara menghilangkan baris yang mengandung nilai hilang pada kolom-kolom yang terlibat langsung dalam pemodelan.

Secara teknis, langkah penanganan missing values yang digunakan dalam skrip Python adalah:



Melalui langkah ini, hanya observasi yang memiliki informasi lengkap pada seluruh variabel model yang dipertahankan. Pendekatan listwise deletion ini dipilih karena proporsi data hilang relatif kecil pada variabel-variabel utama, sehingga penghapusan baris dengan nilai hilang tidak mengurangi jumlah sampel secara drastis dan tetap mempertahankan kualitas data untuk analisis regresi.

## Encoding dan Pembagian Data

Pada template tugas, subbab ini mencakup dua hal:

1. proses encoding variabel kategorikal menggunakan One-Hot Encoding, dan
2. pembagian data menjadi data latih (training) dan data uji (testing).

Namun, pada penelitian ini seluruh variabel yang digunakan dalam model (CO(GT), PT08.S1(CO), C6H6(GT), PT08.S2(NMHC), T, RH, dan AH) merupakan variabel numerik kontinu. Dengan demikian, tidak terdapat variabel kategorikal seperti Tipe\_Fitur di dalam dataset AirQuality, sehingga proses encoding tidak perlu diterapkan secara nyata.

Meskipun demikian, mekanisme one-hot encoding tetap dijelaskan secara konseptual sebagai bagian dari dokumentasi metodologi.

### One-Hot Encoding

Secara umum, apabila dalam suatu penelitian terdapat variabel kategorikal (misalnya Tipe\_Fitur dengan kategori seperti Sensor, Cuaca, dan sebagainya), variabel tersebut tidak dapat langsung digunakan dalam model regresi linier. Variabel kategorikal perlu diubah terlebih dahulu menjadi sejumlah variabel biner (0/1) menggunakan teknik One-Hot Encoding, sehingga setiap kategori direpresentasikan sebagai kolom tersendiri.

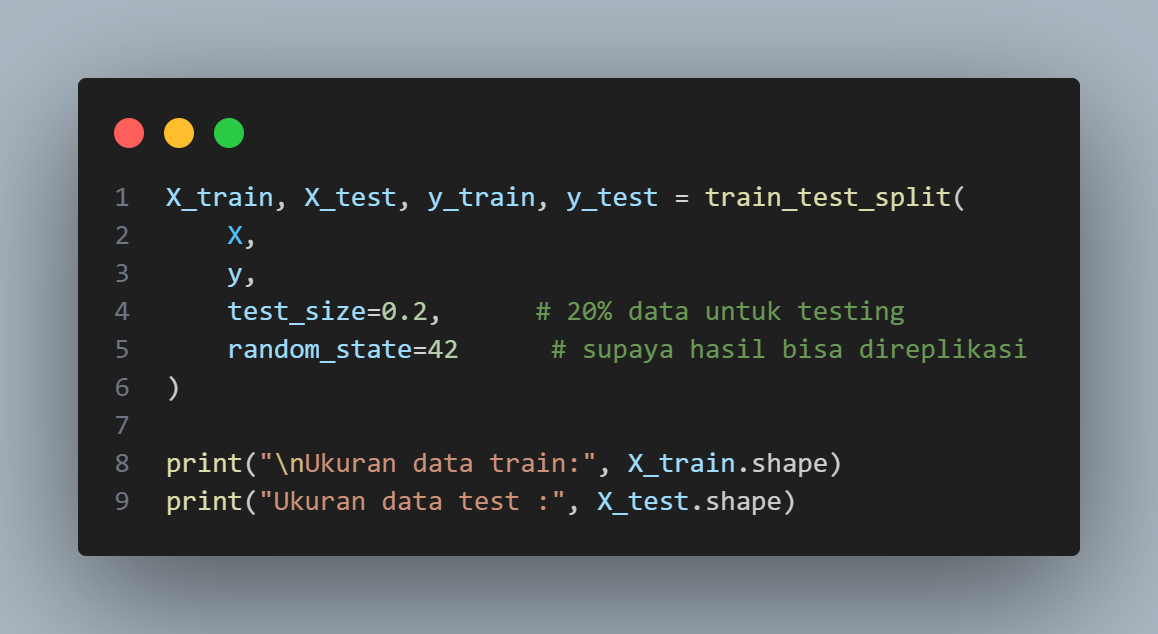
Contoh perintah Python untuk melakukan one-hot encoding terhadap variabel kategorikal Tipe\_Fitur adalah sebagai berikut:



Argumen drop\_first=True digunakan untuk menghindari jebakan dummy variable trap dengan membuang satu kategori sebagai baseline.

Namun, pada dataset AirQuality.xlsx yang digunakan dalam penelitian ini, seluruh variabel yang masuk ke dalam model sudah berbentuk numerik dan tidak memiliki kolom Tipe\_Fitur. Oleh karena itu, langkah one-hot encoding tidak diterapkan secara nyata dalam implementasi kode, dan proses pemodelan dapat langsung menggunakan df\_model tanpa tahap encoding tambahan.

Sebaliknya, setelah proses pembersihan data, tahapan yang benar-benar diterapkan adalah pembagian data menjadi data latih dan data uji, seperti berikut:



Pembagian ini memastikan bahwa kinerja model dapat dievaluasi secara objektif menggunakan data yang tidak pernah “dilihat” model selama proses pelatihan, sehingga hasil evaluasi (R², MAE, RMSE) lebih mencerminkan kemampuan generalisasi model terhadap data baru.

# ANALISIS REGRESI & INTERPRETASI

## Perintah Pelatihan Model

### Implementasi: Inisialisasi dan Pelatihan Model Regresi Linier Berganda

Pada tahap ini, model yang digunakan adalah Regresi Linier Berganda dengan variabel target CO(GT) dan enam variabel prediktor (PT08.S1(CO), C6H6(GT), PT08.S2(NMHC), T, RH, dan AH). Data terlebih dahulu dibagi menjadi data latih (training) dan data uji (testing) dengan proporsi 80% : 20%. Model kemudian diinisialisasi menggunakan kelas LinearRegression dari pustaka scikit-learn dan dilatih menggunakan data latih.

Perintah Python yang digunakan:

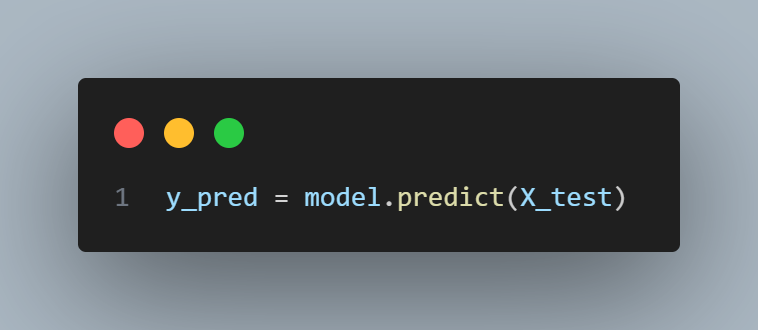


Dengan perintah ini, model akan mencari nilai koefisien regresi dan intercept yang meminimalkan jumlah kuadrat galat (Ordinary Least Squares) antara nilai aktual CO(GT) dan nilai prediksi model pada data latih.

### Prediksi

Setelah model selesai dilatih, langkah berikutnya adalah melakukan prediksi pada data uji (X\_test) yang tidak pernah digunakan pada saat pelatihan. Prediksi ini digunakan untuk menilai kemampuan model dalam melakukan generalisasi terhadap data baru.

Perintah Python yang digunakan:



Vektor y\_pred berisi nilai CO(GT) hasil prediksi model, sedangkan y\_test berisi nilai CO(GT) aktual pada data uji. Kedua vektor ini kemudian dibandingkan pada tahap evaluasi model.

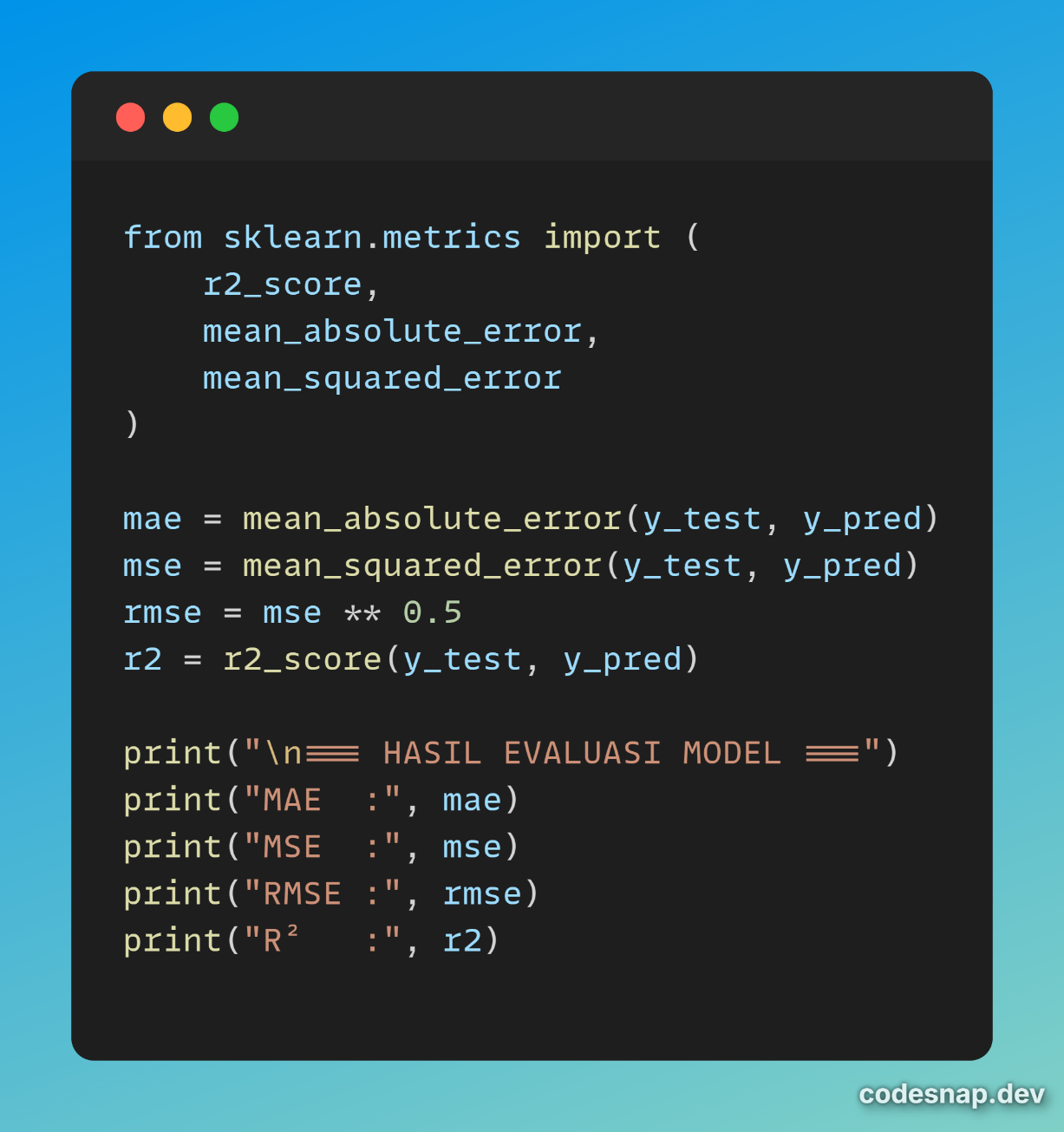
## Perintah Evaluasi Model

### Hitung 4 Metrik Evaluasi pada Data Testing

Untuk menilai kinerja model, digunakan empat metrik evaluasi utama, yaitu:

1. Mean Absolute Error (MAE)
2. Mean Squared Error (MSE)
3. Root Mean Squared Error (RMSE)
4. Koefisien Determinasi (R²)

Perintah Python yang digunakan :



Output:



#### Mean Absolute Error (MAE)

Nilai MAE yang diperoleh sebesar kurang lebih 0,2857. Artinya, secara rata-rata selisih absolut antara nilai aktual dan nilai prediksi konsentrasi CO pada data uji adalah sekitar 0,29 satuan CO(GT). Nilai MAE yang relatif kecil ini menunjukkan bahwa model mampu menghasilkan prediksi yang cukup dekat dengan nilai observasinya.

#### Mean Squared Error (MSE)

Nilai MSE yang dihasilkan adalah sekitar 0,1767. Metrik ini merepresentasikan rata-rata kuadrat kesalahan antara nilai aktual dan nilai prediksi. Karena kesalahan dikuadratkan, MSE sensitif terhadap error yang besar; nilai yang tidak terlalu besar mengindikasikan bahwa tidak banyak prediksi dengan kesalahan yang ekstrem.

#### Root Mean Squared Error (RMSE)

Nilai RMSE, yaitu akar kuadrat dari MSE, diperoleh sebesar kurang lebih 0,4204. RMSE memiliki satuan yang sama dengan variabel target, sehingga dapat diinterpretasikan bahwa secara rata-rata akar kuadrat kesalahan prediksi konsentrasi CO berada di kisaran 0,42 satuan CO(GT). Semakin kecil nilai RMSE, semakin baik kemampuan model dalam memprediksi konsentrasi CO, sehingga nilai ini memperkuat temuan bahwa model memiliki kinerja yang cukup baik.

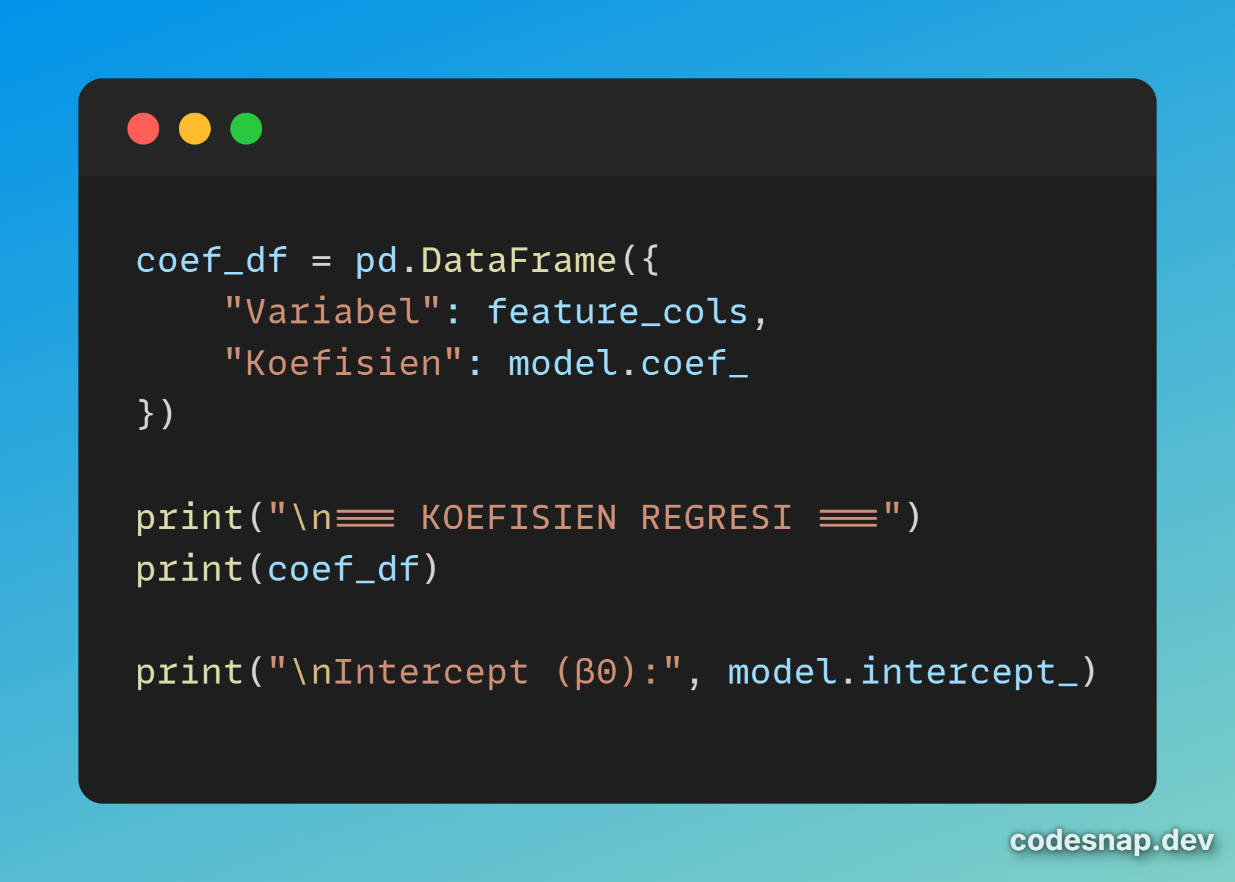
#### Koefisien Determinasi

Koefisien determinasi yang diperoleh adalah R² ≈ 0,9095. Hal ini berarti sekitar 90,95% variasi konsentrasi CO(GT) pada data uji dapat dijelaskan oleh kombinasi variabel PT08.S1(CO), C6H6(GT), PT08.S2(NMHC), T, RH, dan AH dalam model regresi linier berganda yang dibangun. Nilai R² yang mendekati 1 menunjukkan bahwa model memiliki kemampuan penjelasan yang sangat baik terhadap variasi data, sehingga secara keseluruhan model dapat dianggap cukup akurat untuk memprediksi kualitas udara dalam konteks dataset yang digunakan.

## Perintah Interpretasi Koefisien

### Tampilkan semua koefisien

Koefisien regresi menggambarkan besar dan arah pengaruh masing-masing variabel prediktor terhadap variabel target CO(GT) ketika variabel lain dianggap konstan. Koefisien dan intercept model dapat ditampilkan dengan perintah:



Output:



### Interpretasi

Berikut interpretasi masing-masing koefisien (dengan asumsi variabel lain konstan):

Intercept (β₀ = −0,5147)

Jika seluruh variabel prediktor bernilai nol, maka model memprediksi konsentrasi CO(GT) sebesar −0,5147. Secara fisik kondisi semua variabel bernilai nol mungkin tidak terjadi, sehingga intercept lebih berfungsi sebagai konstanta penyesuai model dan tidak diinterpretasikan secara substantif.

PT08.S1(CO) (β₁ = 0,001296)

Setiap kenaikan 1 satuan pembacaan sensor PT08.S1(CO) diperkirakan akan meningkatkan CO(GT) rata-rata sebesar 0,0013 satuan, dengan asumsi variabel lain tetap. Tanda positif sesuai dengan ekspektasi: semakin besar respon sensor CO, semakin tinggi konsentrasi CO yang terukur.

C6H6(GT) (β₂ = 0,142619)

Koefisien ini bernilai positif dan paling besar di antara seluruh prediktor. Artinya, setiap kenaikan 1 satuan konsentrasi benzena C6H6(GT) akan meningkatkan konsentrasi CO yang diprediksi sekitar 0,1426 satuan. Hal ini menunjukkan bahwa benzena berasosiasi sangat kuat dengan emisi sumber yang sama (misalnya lalu lintas atau aktivitas pembakaran), sehingga menjadi prediktor utama dalam menjelaskan variasi CO, konsisten dengan korelasi tinggi (≈ 0,93) yang diperoleh sebelumnya.

PT08.S2(NMHC) (β₃ = 0,000251)

Setiap kenaikan 1 satuan pembacaan sensor PT08.S2(NMHC) akan meningkatkan CO(GT) yang diprediksi sekitar 0,00025 satuan. Tanda positif menunjukkan bahwa peningkatan konsentrasi hidrokarbon non-metana cenderung diikuti peningkatan konsentrasi CO, meskipun besaran kontribusinya per satuan relatif kecil dibanding C6H6(GT). Namun secara keseluruhan, peran sensor ini tetap penting karena rentang nilainya cukup besar dan korelasinya dengan CO(GT) juga tinggi (≈ 0,92).

T – Temperature (β₄ = −0,017808)

Koefisien suhu bernilai negatif, yang berarti setiap kenaikan 1°C suhu udara akan menurunkan prediksi CO(GT) sekitar 0,0178 satuan, dengan variabel lain tetap. Secara fisik, hal ini dapat diinterpretasikan bahwa pada suhu yang lebih tinggi, kondisi atmosfer cenderung lebih mendukung proses pencampuran dan dispersi polutan, sehingga konsentrasi CO di permukaan menurun.

RH – Relative Humidity (β₅ = −0,000634)

Setiap kenaikan 1% kelembapan relatif diperkirakan akan menurunkan CO(GT) sekitar 0,00063 satuan, ceteris paribus. Nilainya sangat kecil, tetapi tanda negatif menunjukkan kecenderungan bahwa kelembapan yang lebih tinggi berasosiasi dengan konsentrasi CO yang sedikit lebih rendah.

AH – Absolute Humidity (β₆ = −0,149438)

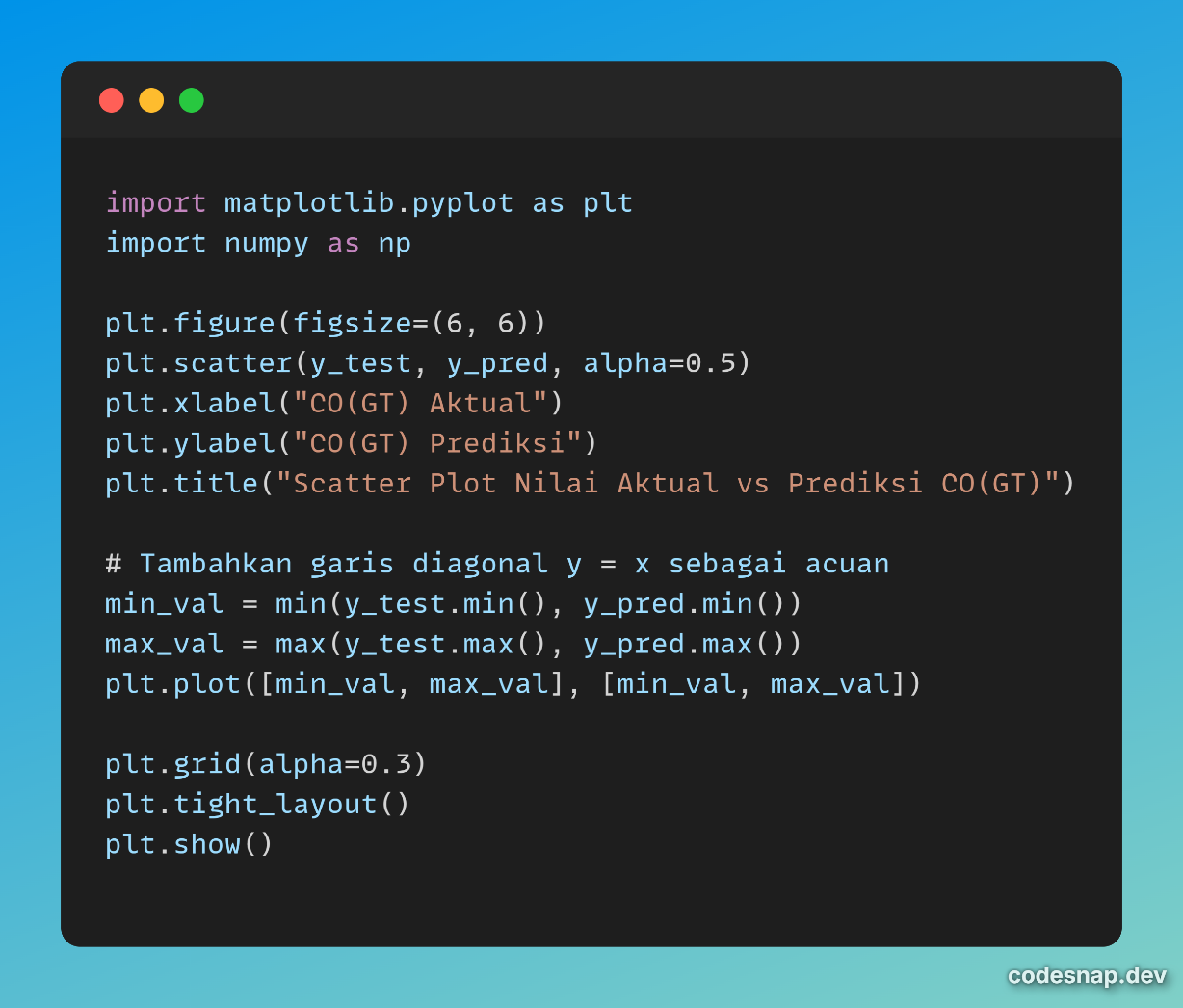
Koefisien AH juga bernilai negatif dan besarnya cukup signifikan. Setiap kenaikan 1 satuan kelembapan absolut diperkirakan akan menurunkan CO(GT) sekitar 0,1494 satuan. Hal ini mengindikasikan bahwa ketika kandungan uap air di udara meningkat, konsentrasi CO yang terukur cenderung menurun, kemungkinan karena perubahan sifat dispersi dan dinamika atmosfer.

## Perintah Visualisasi Hasil

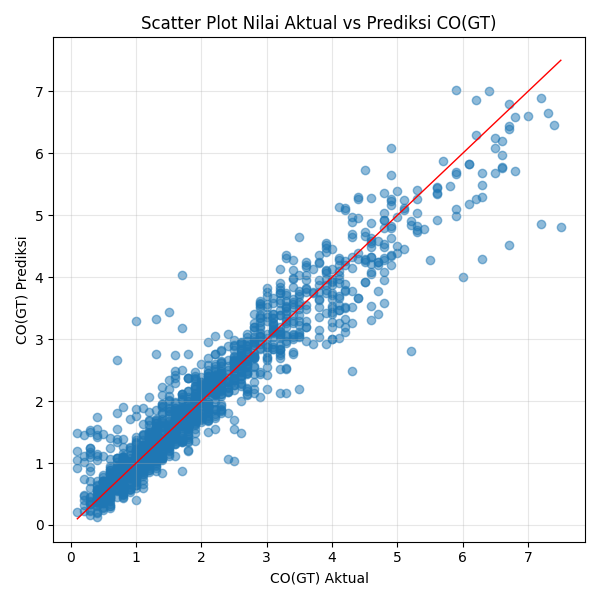
### Scatter Plot

Untuk mengevaluasi kesesuaian antara nilai prediksi model dan nilai aktual pada data uji, dibuat scatter plot antara y\_test (nilai aktual) dan y\_pred (nilai prediksi). Jika model bekerja dengan baik, titik-titik data akan banyak berkumpul di sekitar garis diagonal .

Perintah Python:



Output:



Untuk melihat kesesuaian antara nilai prediksi model dan nilai observasi, dibuat scatter plot yang membandingkan CO(GT) aktual pada sumbu X dan CO(GT) prediksi pada sumbu Y, dilengkapi dengan garis referensi diagonal . Titik-titik biru merepresentasikan pasangan , sedangkan garis merah menunjukkan posisi ideal ketika nilai prediksi sama persis dengan nilai aktual.

Dari Gambar scatter plot terlihat bahwa sebagian besar titik data berkelompok rapat di sekitar garis diagonal, terutama pada rentang konsentrasi CO yang rendah hingga sedang. Hal ini menunjukkan bahwa model regresi linier berganda mampu menghasilkan prediksi yang sangat dekat dengan nilai aktual pada sebagian besar observasi, sejalan dengan nilai R² ≈ 0,91 dan RMSE ≈ 0,42 yang telah diperoleh sebelumnya. Beberapa titik menyimpang lebih jauh dari garis diagonal, khususnya pada nilai CO yang lebih tinggi, yang mengindikasikan adanya sejumlah observasi dengan kesalahan prediksi yang lebih besar. Namun secara umum, pola sebaran titik yang mengikuti garis mengonfirmasi bahwa model memiliki kinerja prediktif yang baik dalam memperkirakan konsentrasi CO berdasarkan variabel-variabel prediktor yang digunakan.

# LAPORAN AKHIR

## Pendahuluan

### Latar Belakang

Kualitas udara merupakan salah satu faktor penting yang berpengaruh terhadap kesehatan manusia dan kelayakan lingkungan hidup. Salah satu gas polutan yang sering digunakan sebagai indikator kualitas udara adalah carbon monoxide (CO). Paparan CO dalam konsentrasi tinggi dapat mengganggu sistem pernapasan dan kardiovaskular, sehingga pemantauan serta pemodelan konsentrasi CO di udara menjadi hal yang krusial.

Perkembangan sensor lingkungan dan teknik pengolahan data memungkinkan dibangunnya model statistika untuk memprediksi konsentrasi CO berdasarkan pembacaan sensor gas dan kondisi meteorologis. Dengan model yang baik, proses pemantauan kualitas udara dapat dilakukan secara lebih efisien, serta memberikan informasi prediktif yang berguna bagi pengambil kebijakan maupun pengguna sistem pemantauan. Dalam konteks tersebut, analisis regresi linier berganda menjadi salah satu pendekatan yang relevan untuk mengkaji hubungan linier antara CO dan beberapa variabel penjelas yang terkait.

### Tujuan

Secara umum, tujuan dari proyek ini adalah:

1. Membangun model regresi linier berganda untuk memprediksi konsentrasi CO (CO(GT)) berdasarkan pembacaan beberapa sensor gas dan variabel meteorologis.
2. Menganalisis faktor-faktor yang berpengaruh terhadap variasi konsentrasi CO dengan melihat kekuatan hubungan dan tanda koefisien regresi.
3. Mengevaluasi kinerja model menggunakan beberapa metrik evaluasi (MAE, MSE, RMSE, dan R²) pada data uji.

Tujuan tambahan yang diharapkan adalah memberikan gambaran awal mengenai kualitas data AirQuality serta potensi pengembangan model yang lebih kompleks pada penelitian lanjutan.

### Identitas Dataset

Dataset yang digunakan dalam penelitian ini adalah dataset AirQuality yang memuat pengukuran kualitas udara dan variabel meteorologi dalam bentuk deret waktu. Data diperoleh secara sekunder dari sumber daring (kaggle), kemudian diunduh dan disimpan dalam berkas AirQuality.csv, lalu dirapikan menjadi AirQuality.xlsx.

Secara umum, karakteristik dataset adalah sebagai berikut:

Sumber data : Data sekunder dari repository daring (Kaggle) (AirQuality dataset).

Format : Awalnya CSV, kemudian diolah dan disimpan dalam format Excel (AirQuality.xlsx).

Jumlah observasi awal : 9.357 baris dengan 15 kolom.

Jumlah observasi yang digunakan dalam model : 7.344 baris setelah proses pembersihan dan penghapusan data tidak lengkap.

Variabel target : CO(GT) – konsentrasi carbon monoxide di udara ambien.

Variabel prediktor utama :

1. PT08.S1(CO) – pembacaan sensor CO,
2. C6H6(GT) – konsentrasi benzena,
3. PT08.S2(NMHC) – sensor non-methane hydrocarbons,
4. T – suhu udara (°C),
5. RH – kelembapan relatif (%),
6. AH – kelembapan absolut.

## Metodologi

### Detail Langkah EDA dan Preprocessing

Metodologi yang digunakan dalam penelitian ini dapat diringkas dalam beberapa tahap berikut:

1. Akuisisi dan pemuatan data  
   Data diunduh dalam format CSV kemudian dibuka di Microsoft Excel untuk proses kerapian awal. Data yang telah dirapikan disimpan sebagai AirQuality.xlsx dan dimuat ke dalam Python menggunakan pandas melalui fungsi read\_excel().
2. Analisis data eksploratif (EDA)
   1. Menampilkan 10 baris pertama dan menghitung persentase missing values pada setiap kolom untuk mendapatkan gambaran awal kualitas data.
   2. Membuat histogram untuk variabel target CO(GT) dan menghitung skewness untuk melihat bentuk distribusi.
   3. Membuat box plot dan menghitung batas IQR (Q1, Q3, IQR, lower bound, upper bound) untuk mengidentifikasi kemungkinan outlier pada CO(GT).
   4. Menghitung matriks korelasi Pearson dan memvisualisasikannya dalam bentuk heatmap untuk melihat hubungan linier antar variabel, baik antara prediktor dengan target maupun antarprediktor, sekaligus menilai potensi multikolinearitas.
3. Pembersihan data (preprocessing)
   1. Dibentuk subset df\_model yang hanya berisi variabel target (CO(GT)) dan variabel prediktor yang dipilih.
   2. Jika terdapat kode khusus untuk data hilang, kode tersebut dapat diubah menjadi NaN. Dalam implementasi saat ini, data hilang ditangani dengan cara menghapus baris yang memiliki nilai NaN pada salah satu kolom model (listwise deletion).
   3. Pendekatan ini dipilih karena persentase missing value pada variabel-variabel utama relatif kecil, sehingga penghapusan tidak mengurangi jumlah data secara drastis dan lebih sederhana dibandingkan melakukan imputasi yang berpotensi menambah asumsi baru.
4. Encoding variabel kategorikal  
   Template tugas mensyaratkan adanya penjelasan mengenai One-Hot Encoding. Namun, pada dataset AirQuality seluruh variabel yang digunakan dalam model merupakan variabel numerik kontinu dan tidak terdapat variabel kategorikal (seperti Tipe\_Fitur). Oleh karena itu, encoding tidak diterapkan secara nyata. One-hot encoding hanya dijelaskan secara konseptual sebagai teknik yang digunakan apabila terdapat variabel kategorikal dalam penelitian lain.
5. Pembagian data train–test
   1. Data yang sudah bersih dibagi menjadi data latih (80%) dan data uji (20%) menggunakan train\_test\_split dari scikit-learn dengan random\_state = 42 agar hasil dapat direplikasi.
   2. Variabel prediktor disimpan dalam matriks X, sedangkan variabel target disimpan dalam vektor y.
6. Pelatihan dan evaluasi model
   1. Model Regresi Linier Berganda diinisialisasi dengan LinearRegression() dan dilatih menggunakan X\_train dan y\_train.
   2. Prediksi dilakukan pada X\_test untuk memperoleh y\_pred.
   3. Kinerja model dinilai menggunakan empat metrik: MAE, MSE, RMSE, dan R², serta divisualisasikan dengan scatter plot antara y\_test dan y\_pred.

## Hasil dan Pembahasan Model

### Tabel ringkasan Metrik Evaluasi

Ringkasan hasil evaluasi model pada data uji ditunjukkan pada Tabel berikut (dibulatkan empat desimal):

|  |  |
| --- | --- |
| Metrik | Nilai |
| MAE | 0,2857 |
| MSE | 0,1767 |
| RMSE | 0,4204 |
| R² | 0,9095 |

Nilai-nilai ini menunjukkan bahwa model memiliki kesalahan prediksi yang relatif kecil dengan kemampuan penjelasan yang tinggi terhadap variasi data.

### Koefisien Determinasi

Nilai koefisien determinasi R² ≈ 0,9095 mengindikasikan bahwa sekitar 90,95% variasi konsentrasi CO(GT) pada data uji dapat dijelaskan oleh kombinasi variabel PT08.S1(CO), C6H6(GT), PT08.S2(NMHC), T, RH, dan AH dalam model regresi linier berganda.

Dalam konteks pemodelan kualitas udara, nilai R² yang mendekati 1 ini dapat dikategorikan sebagai sangat baik, karena hanya sekitar 9% variasi yang tidak dapat dijelaskan oleh model dan terakumulasi dalam komponen error. Hal ini juga selaras dengan hasil scatter plot nilai aktual vs prediksi, di mana titik-titik data terkonsentrasi di sekitar garis diagonal , menunjukkan bahwa prediksi model sangat mendekati nilai observasi pada sebagian besar pengukuran.

Meskipun demikian, masih terdapat beberapa titik yang menyimpang lebih jauh dari garis diagonal, terutama pada nilai CO yang lebih tinggi, yang berkontribusi terhadap nilai MSE dan RMSE yang tidak nol. Hal ini menunjukkan bahwa walaupun model linier cukup efektif secara umum, masih ada pola non-linier atau faktor lain yang belum sepenuhnya tertangkap oleh model.

### Interpretasi Koefisien

Model regresi linier berganda yang diperoleh dapat dituliskan (dengan pembulatan koefisien) sebagai berikut:

Interpretasi utama dari koefisien-koefisien tersebut adalah:

1. C6H6(GT) memiliki koefisien positif paling besar (≈ 0,1426). Artinya, setiap kenaikan 1 satuan konsentrasi benzena diperkirakan akan meningkatkan konsentrasi CO sekitar 0,14 satuan, dengan asumsi variabel lain tetap. Hal ini menunjukkan bahwa benzena merupakan prediktor paling kuat bagi CO, konsisten dengan nilai korelasinya yang sangat tinggi (≈ 0,93).
2. PT08.S1(CO) dan PT08.S2(NMHC) juga memiliki koefisien positif, masing-masing sekitar 0,0013 dan 0,0003. Peningkatan pembacaan sensor-sensor ini berasosiasi dengan peningkatan CO, menggambarkan bahwa sumber emisi yang meningkatkan hidrokarbon dan NMHC cenderung juga meningkatkan CO di udara.
3. Variabel meteorologis T, RH, dan AH memiliki koefisien negatif. Secara khusus, AH memiliki koefisien sekitar −0,1494, yang berarti peningkatan kelembapan absolut cenderung menurunkan konsentrasi CO yang diprediksi. Demikian pula, kenaikan suhu (T) sebesar 1°C diperkirakan menurunkan CO sekitar 0,0178 satuan. Hal ini dapat diinterpretasikan bahwa kondisi atmosfer dengan suhu dan kelembapan tertentu lebih mendukung proses dispersi dan pengenceran polutan, sehingga konsentrasi CO di permukaan menurun.

Secara keseluruhan, interpretasi koefisien menunjukkan bahwa variabel sensor gas (khususnya C6H6(GT), PT08.S1(CO), dan PT08.S2(NMHC)) merupakan faktor utama yang mendorong meningkatnya konsentrasi CO, sedangkan variabel meteorologi cenderung berperan sebagai faktor yang sedikit menekan konsentrasi CO.

## Kesimpulan dan Saran

Berdasarkan seluruh rangkaian analisis yang telah dilakukan, beberapa kesimpulan utama dapat dirangkum sebagai berikut:

1. Model regresi linier berganda yang dibangun memiliki kinerja yang sangat baik, dengan nilai R² ≈ 0,91, MAE ≈ 0,29, dan RMSE ≈ 0,42 pada data uji. Hal ini menunjukkan bahwa model mampu menjelaskan lebih dari 90% variasi konsentrasi CO(GT) dan memberikan prediksi yang cukup akurat.
2. Faktor yang paling berpengaruh terhadap konsentrasi CO adalah variabel sensor gas, khususnya:
   1. C6H6(GT) sebagai prediktor dengan korelasi dan koefisien terbesar,
   2. diikuti oleh PT08.S2(NMHC) dan PT08.S1(CO).  
      Variabel ini menggambarkan emisi polutan dari sumber-sumber pembakaran (misalnya kendaraan dan industri) yang secara langsung berkaitan dengan peningkatan CO di udara.
3. Variabel meteorologis (T, RH, dan AH) menunjukkan pengaruh yang relatif lebih kecil dan cenderung negatif, yang mengindikasikan bahwa kondisi suhu dan kelembapan tertentu dapat membantu mengurangi konsentrasi CO melalui proses pencampuran dan dispersi di atmosfer.
4. Penerapan teknik pembersihan data sederhana (drop baris dengan missing values) dan penggunaan model linier sudah cukup untuk menghasilkan performa yang baik pada dataset ini, namun masih terdapat beberapa observasi dengan kesalahan prediksi yang cukup besar, terutama pada konsentrasi CO yang tinggi.

Sebagai saran pengembangan model, beberapa langkah berikut dapat dipertimbangkan:

1. Mencoba model regresi lain seperti Ridge Regression dan Lasso Regression untuk menangani potensi multikolinearitas yang cukup tinggi di antara variabel sensor (C6H6(GT), PT08.S2(NMHC), dan PT08.S1(CO)), sekaligus melakukan seleksi variabel secara lebih sistematis.
2. Mengeksplorasi model non-linier (misalnya Random Forest Regression atau Gradient Boosting) untuk menangkap pola hubungan yang tidak sepenuhnya linier antara variabel prediktor dan konsentrasi CO.
3. Menggunakan validasi silang (cross-validation) untuk memperoleh estimasi kinerja model yang lebih stabil dan mengurangi risiko overfitting.
4. Menambahkan variabel lain yang relevan, seperti data lalu lintas, jenis sumber emisi, atau informasi spasial, apabila tersedia, guna meningkatkan kemampuan model dalam menjelaskan variasi konsentrasi CO secara lebih komprehensif.

Dengan pengembangan lebih lanjut, model yang dibangun berpotensi menjadi komponen pendukung yang kuat dalam sistem pemantauan dan peringatan dini kualitas udara, khususnya untuk mengantisipasi risiko paparan CO di lingkungan perkotaan.