

LINEAR DAN POLYNOMIAL REGRESSION

disusun untuk memenuhi tugas
mata kuliah Pembelajaran mesin

Oleh:

Kelompok 10

Anggota :

Muhammad Habil Aswad	(2208107010013)
Rafli Afriza Nugraha	(2208107010028)
Muhammad Khalid Al Ghifari	(2208107010044)
Muhammad Ridho	(2208107010064)
Muhammad Ilzam	(2208107010087)



JURUSAN INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SYIAH KUALA

DARUSSALAM, BANDA ACEH

2025

BAB I

PENDAHULUAN

1.1 Latar Belakang

Emisi karbon dioksida (CO₂ emissions) dari kendaraan bermotor menjadi salah satu penyebab utama pencemaran udara dan perubahan iklim. Tingginya konsumsi bahan bakar berkontribusi langsung terhadap jumlah emisi CO₂ yang dilepaskan ke atmosfer. Oleh karena itu, pemahaman mengenai pola konsumsi bahan bakar serta faktor-faktor yang mempengaruhi emisi CO₂ sangat penting dalam upaya mengurangi dampak lingkungan. Dengan adanya data yang mencakup berbagai spesifikasi kendaraan, analisis lebih lanjut dapat membantu dalam menentukan faktor utama yang berkontribusi terhadap emisi yang dihasilkan.

Dataset "CO₂ Emissions Canada", yang berasal dari Environment Canada, menyediakan informasi mengenai konsumsi bahan bakar dan emisi CO₂ dari berbagai model kendaraan yang beredar di Kanada. Dataset ini mencakup variabel seperti ukuran mesin, jumlah silinder, jenis bahan bakar, serta konsumsi bahan bakar dalam berbagai kondisi. Dengan menganalisis dataset ini, kita dapat memahami hubungan antara karakteristik kendaraan dan emisi yang dihasilkan, serta membangun model prediksi menggunakan regresi linier dan regresi polinomial untuk memperkirakan jumlah emisi CO₂ berdasarkan spesifikasi kendaraan tertentu.

1.2 Tujuan

- Menganalisis hubungan antara konsumsi bahan bakar dan emisi CO₂ menggunakan teknik eksplorasi data dan visualisasi.
- Membangun dan mengevaluasi model prediksi emisi CO₂ menggunakan regresi linier dan regresi polinomial untuk menentukan model yang paling akurat.

BAB II

PEMBAHASAN

2.1 Deskripsi Dataset

Dataset yang digunakan dalam analisis ini diperoleh dari Kaggle dengan judul "CO2 Emissions by Vehicles". Dataset ini mencakup informasi mengenai emisi karbon dioksida (CO2) yang dihasilkan oleh kendaraan serta berbagai fitur yang dapat mempengaruhi emisi tersebut. Data ini dikumpulkan dari situs resmi pemerintah Kanada dan mencakup rentang waktu selama tujuh tahun. Secara keseluruhan, terdapat 7385 data kendaraan dengan 12 atribut yang mencakup informasi teknis dan konsumsi bahan bakar.

2.2 Sampel Data

Beberapa sampel data dari dataset ini ditampilkan sebagai berikut:

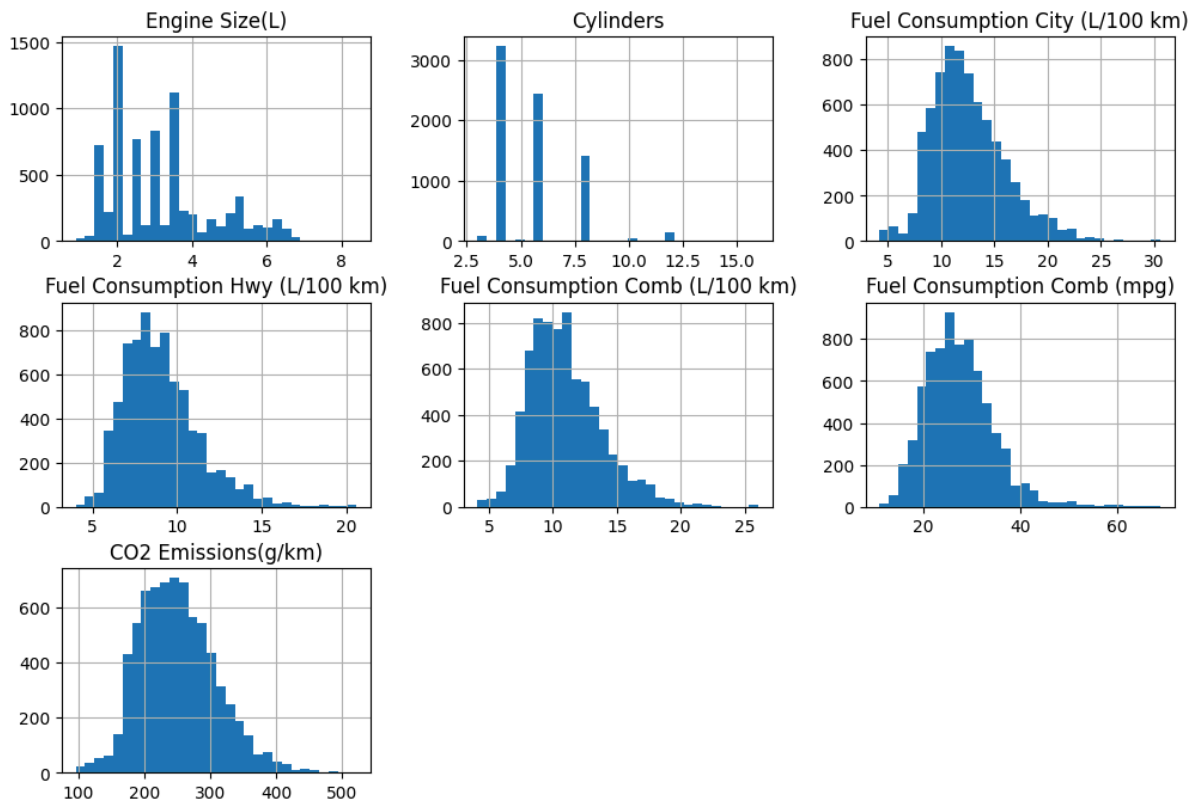
	Make	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmission	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)
0	ACURA	ILX	COMPACT	2.0	4	AS5	Z	9.9	6.7	8.5	33	196
1	ACURA	ILX	COMPACT	2.4	4	M6	Z	11.2	7.7	9.6	29	221
2	ACURA	ILX HYBRID	COMPACT	1.5	4	AV7	Z	6.0	5.8	5.9	48	136
3	ACURA	MDX 4WD	SUV - SMALL	3.5	6	AS6	Z	12.7	9.1	11.1	25	255
4	ACURA	RDX AWD	SUV - SMALL	3.5	6	AS6	Z	12.1	8.7	10.6	27	244

2.3 Data Loading

Proses pemuatan data dilakukan dengan membaca file CSV menggunakan library pandas. Data yang dimuat kemudian diperiksa untuk memastikan tidak ada nilai yang hilang dan bahwa semua atribut memiliki tipe data yang sesuai.

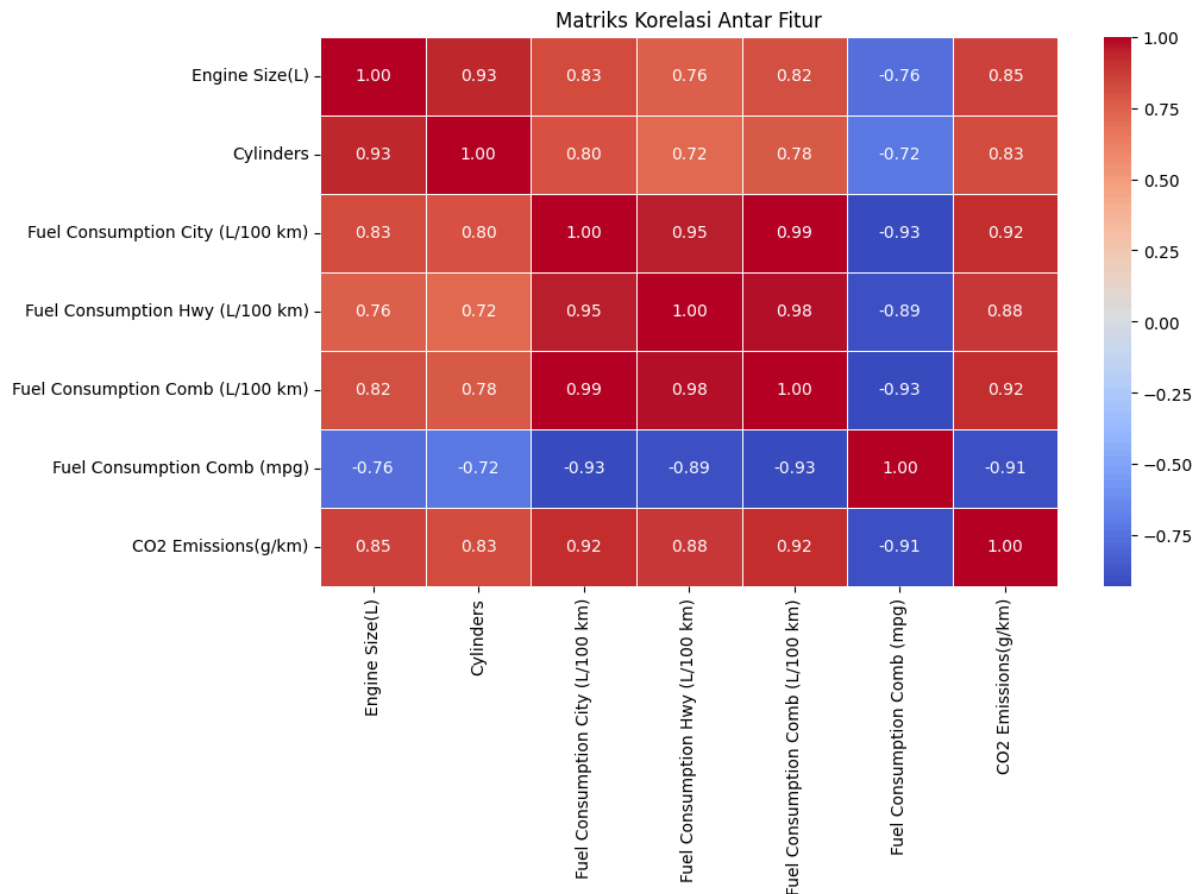
2.4 Distribusi Data

Distribusi Data



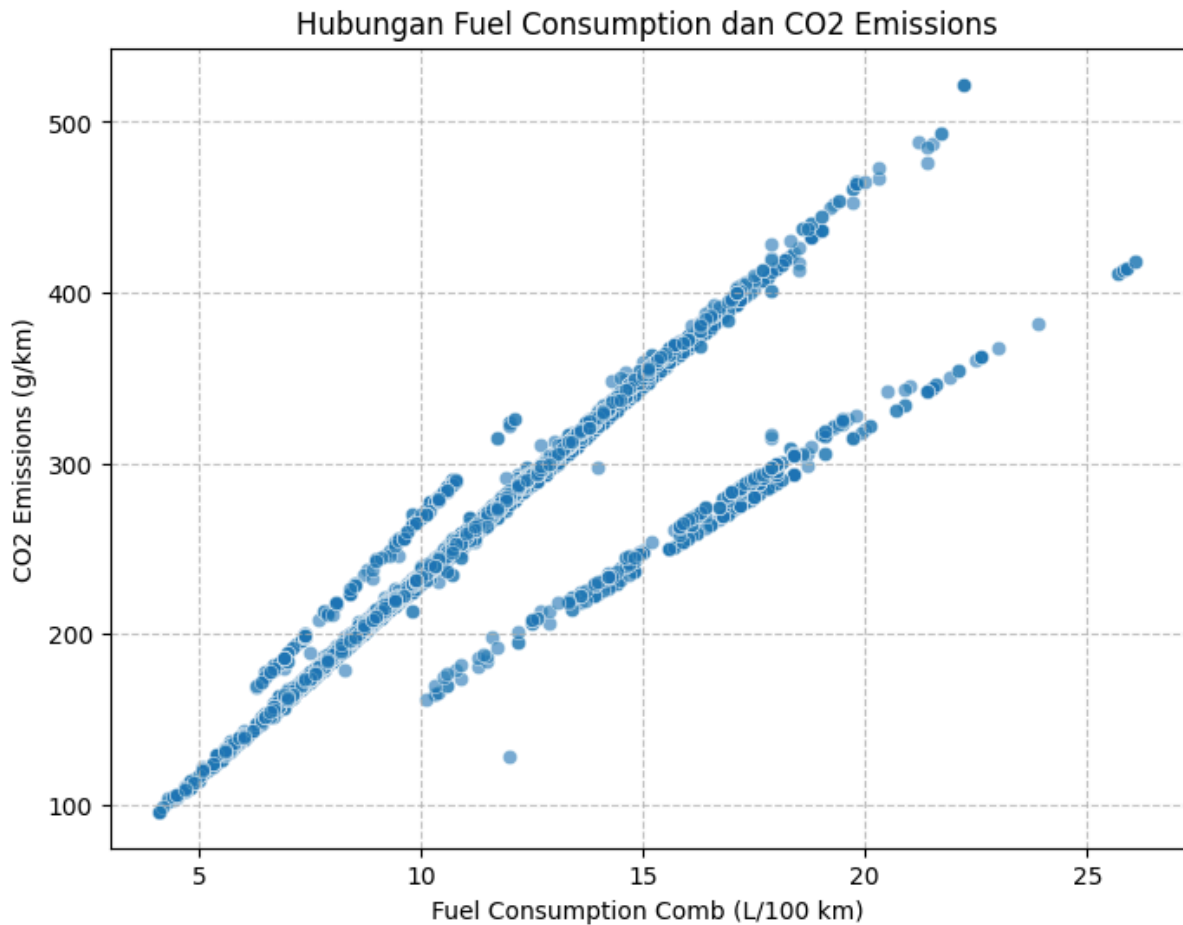
Distribusi data dianalisis untuk melihat bagaimana variabel-variabel utama, seperti konsumsi bahan bakar dan emisi CO₂, tersebar dalam dataset. Dari hasil analisis distribusi, ditemukan bahwa sebagian besar kendaraan memiliki konsumsi bahan bakar kombinasi antara 5 hingga 15 L/100 km dan emisi CO₂ berkisar antara 100 hingga 400 g/km.

2.5 Matriks Korelasi Antar Fitur



Matriks korelasi dihitung untuk menganalisis hubungan antara fitur-fitur dalam dataset. Hasilnya menunjukkan bahwa terdapat korelasi positif yang cukup kuat antara ukuran mesin (Engine Size) dan jumlah silinder (Cylinders) terhadap konsumsi bahan bakar dan emisi CO2. Artinya, semakin besar ukuran mesin dan jumlah silinder, semakin tinggi konsumsi bahan bakar dan emisi CO2 yang dihasilkan.

2.6 Scatter Plot Hubungan Fuel Consumption dan CO2 Emissions



Hubungan antara konsumsi bahan bakar dan emisi CO₂ divisualisasikan menggunakan scatter plot. Dari hasil visualisasi, terlihat bahwa konsumsi bahan bakar yang lebih tinggi berkorelasi dengan peningkatan emisi CO₂. Hal ini sejalan dengan ekspektasi bahwa kendaraan dengan konsumsi bahan bakar lebih boros akan menghasilkan lebih banyak emisi karbon dioksida.

2.7 Membuat Model

Dalam tahap ini, dilakukan pembangunan model untuk memprediksi emisi CO₂ berdasarkan fitur kendaraan. Dua metode yang digunakan adalah **Regresi Linier** dan **Regresi Polinomial**.

1. Regresi Linier

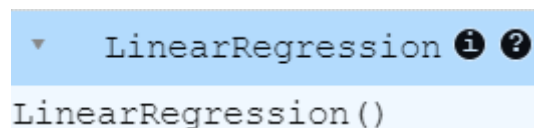
Model regresi linier digunakan untuk mencari hubungan linear antara fitur kendaraan dan emisi CO₂. Proses dimulai dengan membagi dataset menjadi data latih dan data uji, kemudian model diinisialisasi menggunakan **LinearRegression()** dan dilatih dengan data latih menggunakan metode **fit()**. Setelah model dilatih, bobot koefisien regresi diperoleh dan siap digunakan untuk prediksi.

2. Regresi Polinomial

Selain hubungan linear, regresi polinomial digunakan untuk menangkap pola hubungan non-linear antara fitur kendaraan dan emisi CO₂. Dalam metode ini, fitur awal dikonversi menjadi fitur polinomial menggunakan **PolynomialFeatures(degree=2)**, yang menghasilkan kombinasi kuadratik dari fitur asli. Setelah transformasi, model **LinearRegression()** dilatih kembali menggunakan fitur polinomial untuk meningkatkan akurasi prediksi.

Model yang telah dibangun ini akan dievaluasi menggunakan metrik seperti **Mean Squared Error (MSE)** dan **R-squared (R²)** untuk mengetahui performanya dalam memprediksi emisi CO₂ berdasarkan karakteristik kendaraan.

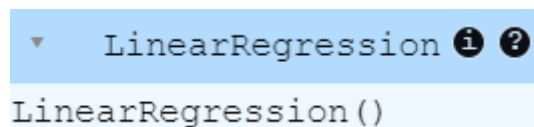
2.7.1 Regresi Linier



```
▼ LinearRegression ⓘ ?  
LinearRegression()
```

Menginisialisasi model **Linear Regression** dan melatihnya menggunakan data **X_train** (fitur) dan **y_train** (target). Setelah eksekusi, model akan mempelajari hubungan antara fitur dan target untuk digunakan dalam prediksi.

2.7.2 Regresi Polinomial



```
▼ LinearRegression ⓘ ?  
LinearRegression()
```

Model Linear Regression yang telah dilatih dengan fitur yang telah ditransformasikan ke dalam bentuk polinomial derajat 2. Matriks fitur X_train diubah menjadi bentuk polinomial menggunakan PolynomialFeatures, yang menambahkan kombinasi kuadratik dari fitur asli. Setelah transformasi, model LinearRegression() dilatih menggunakan data X_train_poly dan y_train.

2.8 Evaluasi Model

Setelah membangun model, langkah selanjutnya adalah mengevaluasi performanya dalam memprediksi emisi CO₂. Evaluasi dilakukan menggunakan beberapa metrik, yaitu Mean Squared Error (MSE), R-squared (R²), Mean Absolute Error (MAE), dan Mean Absolute Percentage Error (MAPE).

2.8.1 Evaluasi Model Regresi Linier

Evaluasi Regresi Linear:

MSE : 377.2062

R² : 0.8903

MAE : 12.9623

MAPE : 5.36%

Output hasil evaluasi **Regresi Linier**:

- **MSE (377.2062)** → Rata-rata kesalahan kuadrat antara prediksi dan nilai aktual.
- **R² (0.8903)** → Model menjelaskan 89.03% variasi data.
- **MAE (12.9623)** → Rata-rata selisih absolut antara prediksi dan nilai aktual (dalam g/km).
- **MAPE (5.36%)** → Kesalahan prediksi relatif kecil, menunjukkan model cukup akurat.

2.8.2 Evaluasi Model Regresi Polinomial

Evaluasi Regresi Polinomial:

MSE : 91.8307

R² : 0.9733

MAE : 5.6609

MAPE : 2.33%

Output hasil evaluasi **Regresi Polinomial**:

- **MSE (91.8307)** → Kesalahan kuadrat rata-rata lebih kecil dibanding regresi linier.
- **R² (0.9733)** → Model menjelaskan 97.33% variasi data, lebih baik dari regresi linier.
- **MAE (5.6609)** → Rata-rata selisih absolut lebih rendah, menunjukkan prediksi lebih akurat.
- **MAPE (2.33%)** → Kesalahan prediksi relatif kecil, menandakan model lebih presisi.

2.9 Perbandingan Evaluasi Model Linier dan Polinomial

	Metric	Regresi Linear	Regresi Polinomial
0	MSE	377.2062	91.8307
1	R ²	0.8903	0.9733
2	MAE	12.9623	5.6609
3	MAPE (%)	5.3619%	2.3281%

Perbandingan evaluasi antara **Regresi Linier** dan **Regresi Polinomial**:

- **MSE (377.2062 vs. 91.8307)** → Regresi Polinomial memiliki kesalahan lebih kecil.
- **R² (0.8903 vs. 0.9733)** → Regresi Polinomial lebih baik dalam menjelaskan variasi data.
- **MAE (12.9623 vs. 5.6609)** → Rata-rata kesalahan prediksi lebih rendah pada Regresi Polinomial.
- **MAPE (5.3619% vs. 2.3281%)** → Regresi Polinomial lebih presisi dibanding Regresi Linier.

Kesimpulannya, Regresi Polinomial memberikan hasil prediksi yang lebih akurat dibandingkan Regresi Linier.

2.10 Analisis Hasil

Berdasarkan hasil evaluasi, Regresi Polinomial menunjukkan performa yang lebih baik dibandingkan Regresi Linier dalam memprediksi emisi CO₂ kendaraan. Hal ini terlihat dari MSE yang lebih rendah (91.8307 vs. 377.2062), R² yang lebih tinggi (0.9733 vs. 0.8903), serta MAE dan MAPE yang lebih kecil.

2.10.1 Koefisien Regresi Linier

	Fitur	Koefisien
0	Intercept	18.932011
1	Fuel Type	10.737114
2	Fuel Consumption Comb (L/100 km)	14.999313
3	Engine Size(L)	10.140795

Output menunjukkan koefisien Regresi Linier, yang menginterpretasikan pengaruh masing-masing fitur terhadap emisi CO₂:

- **Intercept (18.9320)** → Nilai awal emisi saat semua fitur bernilai nol.

- **Fuel Type (10.7371)** → Jenis bahan bakar mempengaruhi emisi CO₂, dengan peningkatan tertentu menambah emisi sebesar 10.7371 satuan.
- **Fuel Consumption Comb (14.9993)** → Setiap kenaikan 1 L/100 km konsumsi bahan bakar meningkatkan emisi sebesar 14.9993 satuan.
- **Engine Size (10.1408)** → Mesin yang lebih besar cenderung menghasilkan lebih banyak emisi CO₂.

Koefisien positif menunjukkan bahwa peningkatan fitur tersebut akan meningkatkan emisi CO₂.

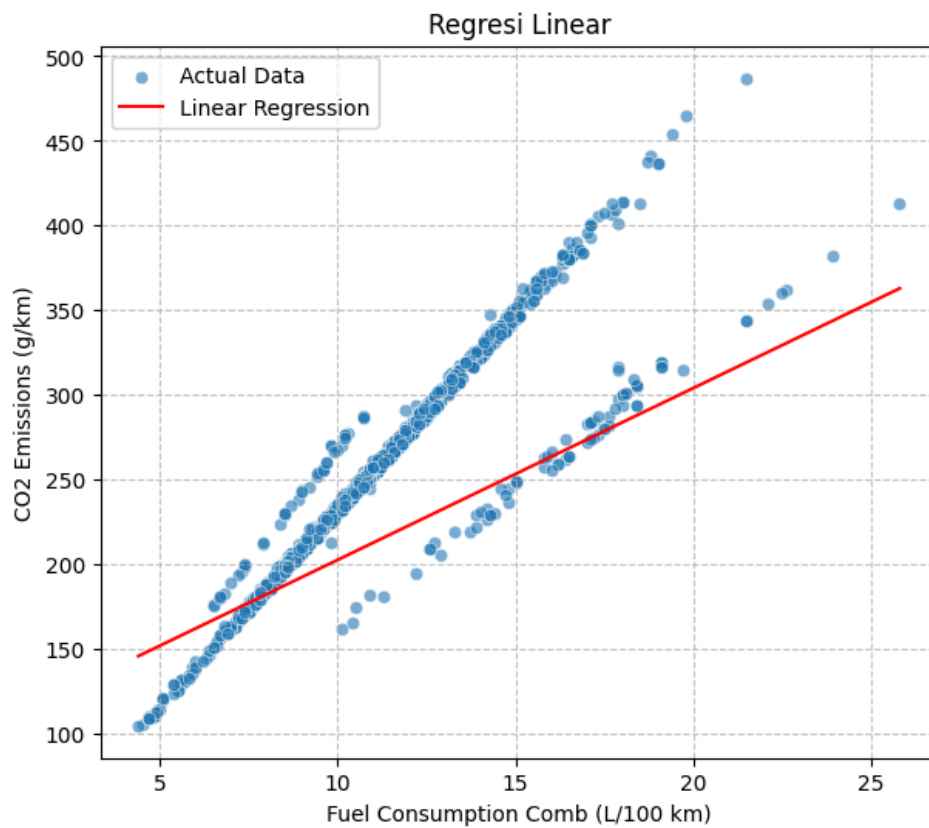
2.10.2 Koefisien Regresi Polinomial

	Fitur	Koefisien
0	Intercept	144.431428
1	1	0.000000
2	Fuel Type	-44.457472
3	Fuel Consumption Comb (L/100 km)	-0.034179
4	Engine Size(L)	35.260627
5	Fuel Type^2	-0.608633
6	Fuel Type Fuel Consumption Comb (L/100 km)	7.244580
7	Fuel Type Engine Size(L)	-8.734772
8	Fuel Consumption Comb (L/100 km)^2	0.043506
9	Fuel Consumption Comb (L/100 km) Engine Size(L)	-1.001662
10	Engine Size(L)^2	1.057230

Output menunjukkan koefisien Regresi Polinomial, yang mencerminkan hubungan non-linear antara fitur kendaraan dan emisi CO₂:

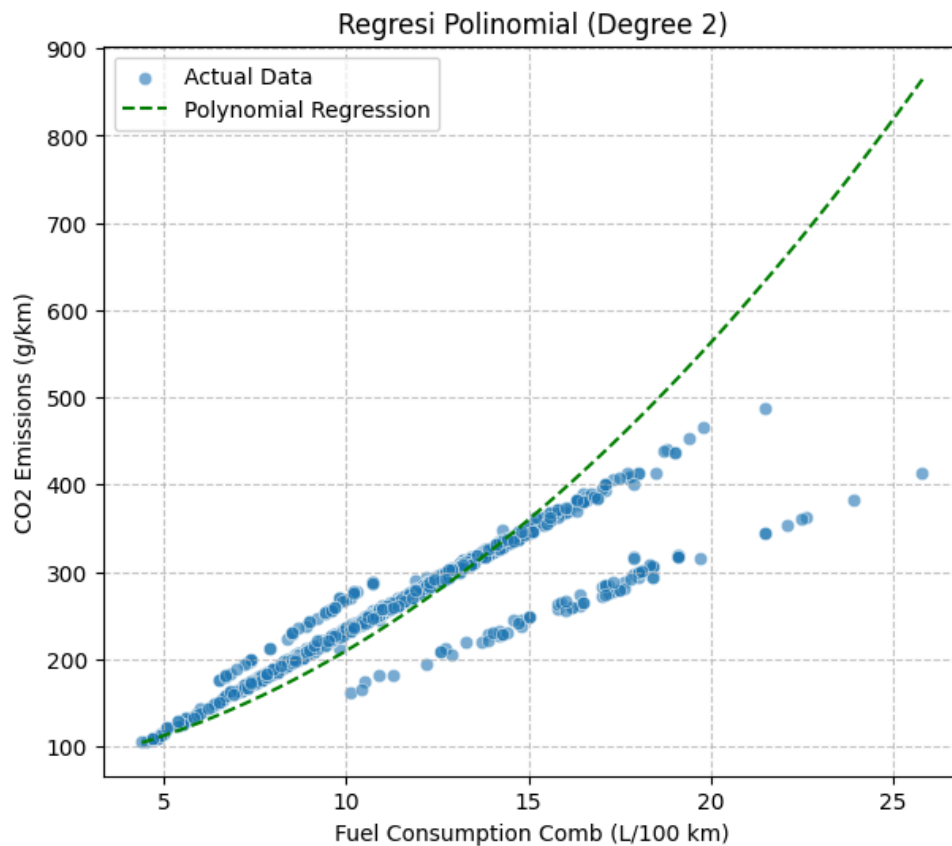
- **Intercept (144.4314)** → Nilai awal emisi saat semua fitur bernilai nol.
- **Fuel Type (-44.4575)** → Jenis bahan bakar memiliki pengaruh negatif terhadap emisi CO₂ dalam model ini.
- **Fuel Consumption Comb (-0.0342)** → Konsumsi bahan bakar berpengaruh kecil dalam bentuk linear, tetapi lebih signifikan dalam interaksi polinomial.
- **Engine Size (35.2606)** → Mesin yang lebih besar cenderung meningkatkan emisi CO₂.
- Koefisien polinomial (seperti Fuel Type^2, Engine Size^2, dan interaksi antar fitur) menunjukkan bahwa hubungan antara fitur dan emisi tidak hanya linear, tetapi juga dipengaruhi oleh interaksi antar fitur, menjelaskan mengapa regresi polinomial lebih akurat dalam menangkap pola data.

2.11 Plot Regresi Linier



Plot regresi linier ini menunjukkan hubungan antara konsumsi bahan bakar (Fuel Consumption Comb) dan emisi CO₂. Titik-titik biru mewakili data aktual, sementara garis merah adalah model regresi linier. Terlihat bahwa regresi linier tidak menangkap pola data dengan sempurna karena ada distribusi non-linear, yang menunjukkan bahwa hubungan antara variabel tidak sepenuhnya linier.

2.12 Plot Regresi Polinomial



Regresi polinomial pada gambar menunjukkan hubungan antara konsumsi bahan bakar dalam liter per 100 kilometer dan emisi CO₂ dalam gram per kilometer. Titik-titik biru merepresentasikan data pengukuran aktual, sedangkan garis hijau putus-putus menggambarkan model regresi polinomial berderajat dua yang mencoba memprediksi emisi CO₂ berdasarkan konsumsi bahan bakar. Ketika konsumsi bahan bakar meningkat, emisi CO₂ juga meningkat namun dengan laju yang tidak konstan, melainkan cenderung meningkat lebih cepat pada nilai konsumsi bahan bakar yang lebih tinggi, membentuk kurva melengkung ke atas. Data aktual tampak tersebar dalam beberapa kelompok berbeda yang mungkin mencerminkan kategori kendaraan yang berbeda, dan model regresi terlihat lebih akurat pada rentang konsumsi bahan bakar rendah hingga menengah dibandingkan pada nilai konsumsi yang lebih tinggi.