Prediksi Harga Mobil Menggunakan Linear Regression

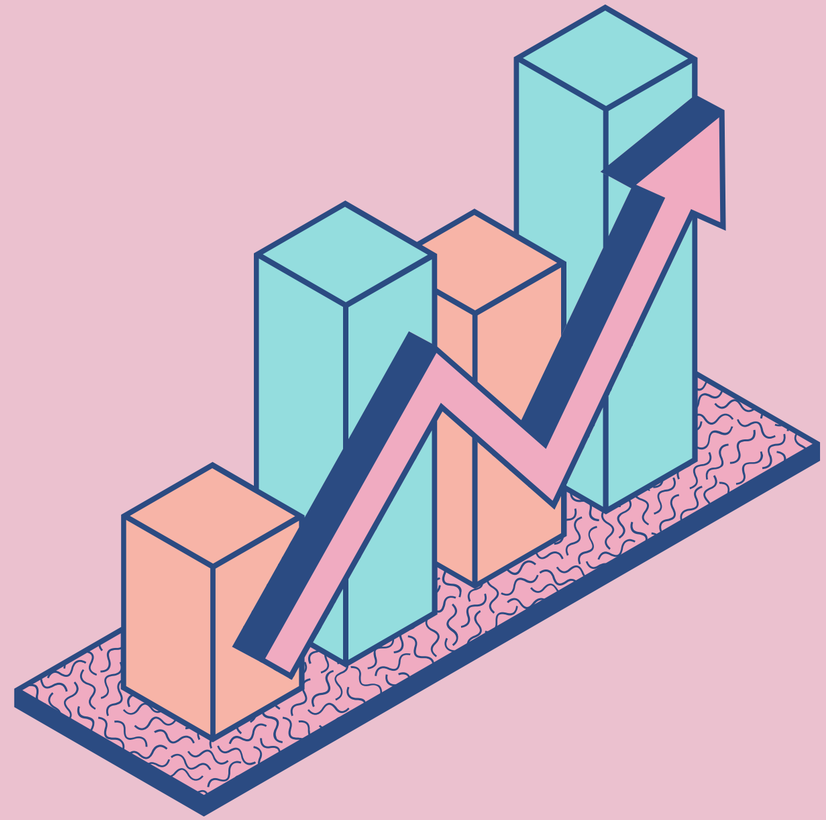
Fadhlurrahman Faiz
2310631170079

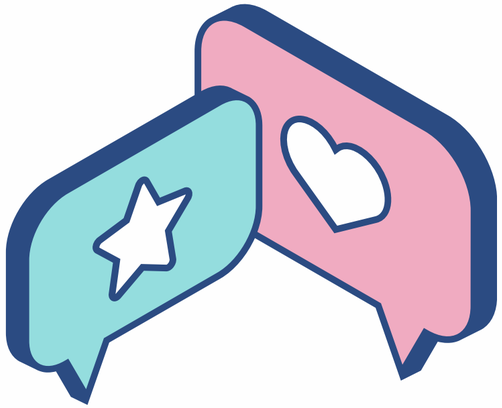
Pendahuluan

Regresi Linear adalah metode statistik yang digunakan untuk mencari hubungan antara satu atau lebih variabel bebas dengan satu variabel target.

Dalam proyek ini, model digunakan untuk melihat bagaimana faktor seperti tahun mobil, jarak tempuh, dan tenaga mesin memengaruhi harga mobil bekas.

Sederhananya, model mencoba menggambarkan pola hubungan tersebut dalam bentuk garis lurus.





Preprocessing

Sebelum membuat model, data perlu dipersiapkan dulu agar bersih dan rapi.

Dataset yang digunakan diambil dari Kaggle dengan judul: “Used Cars Price Prediction”, berisi sekitar 6000 data mobil bekas di India.

Langkah awal yaitu memuat data, melihat jumlah baris dan kolom, serta memahami tipe datanya.

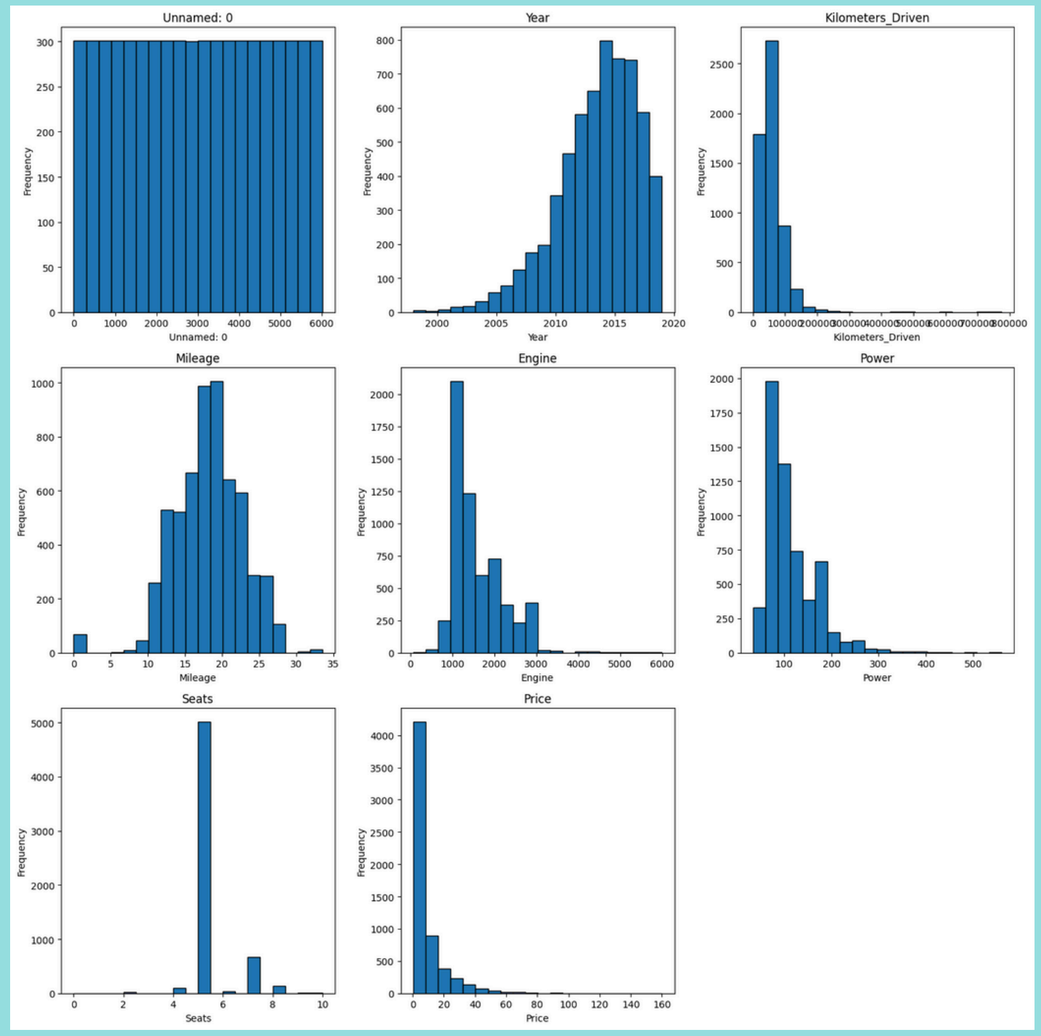
Tujuannya supaya kita tahu isi data dulu sebelum lanjut ke tahap analisis.

Exploratory Data Analysis (EDA)

Histogram

Grafik histogram memperlihatkan distribusi setiap kolom numerik.

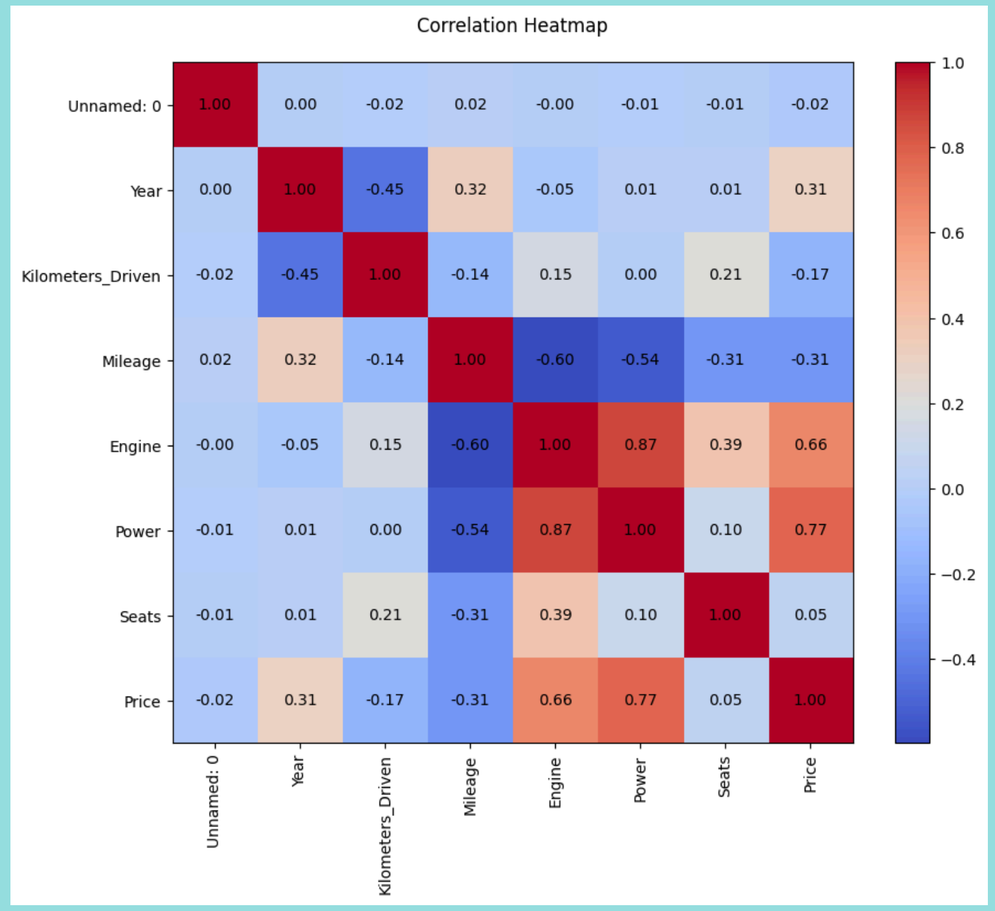
- Kolom Year paling banyak di sekitar tahun 2015.
- Kilometers_Driven, Engine, Power, dan Price memiliki sebaran yang condong ke kanan karena ada nilai ekstrem.
- Mileage hampir berdistribusi normal.
- Seats paling banyak bernilai 5, artinya mobil 5 kursi paling umum.



Exploratory Data Analysis (EDA)

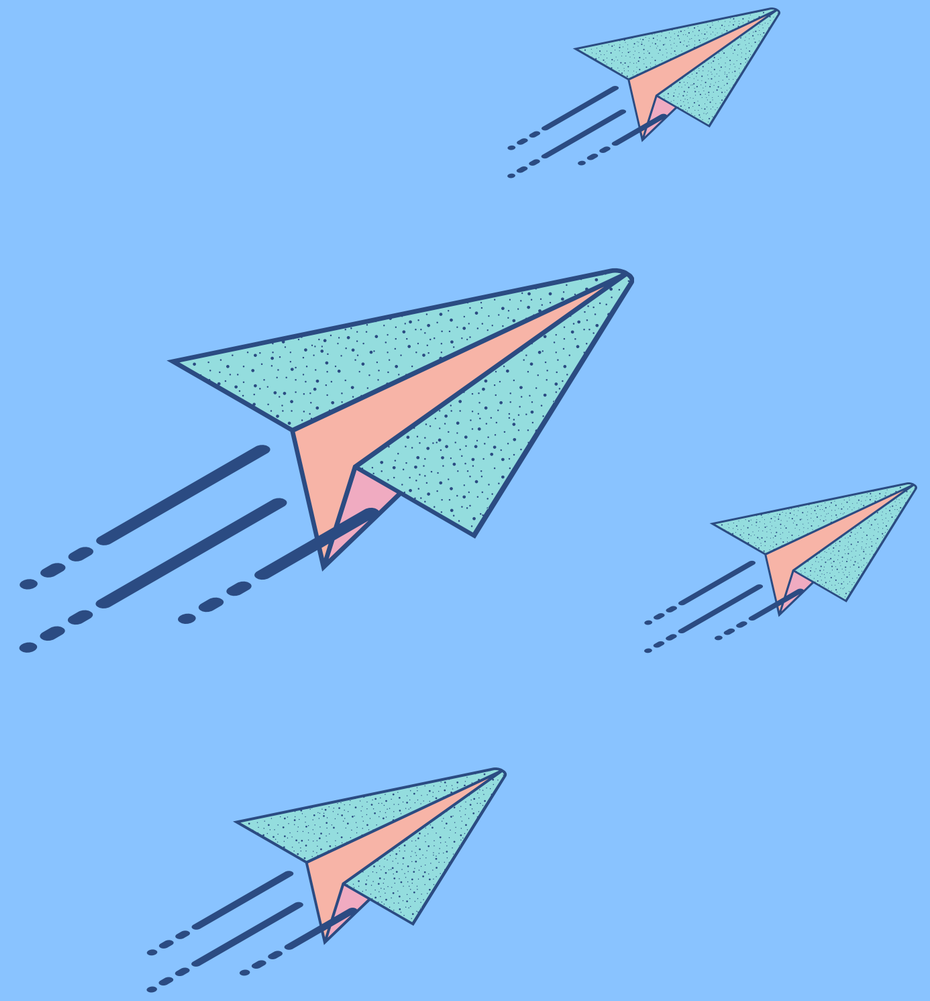
Heatmap

- Engine dan Power punya korelasi positif kuat terhadap Price — mesin besar dan tenaga tinggi bikin harga naik.
- Mileage berkorelasi negatif dengan Price, berarti mobil irit biasanya tenaganya lebih kecil.
- Year juga punya korelasi positif, artinya mobil baru harganya lebih mahal.



Data Cleaning

- Tahap ini dilakukan untuk membersihkan data dari hal-hal yang bisa mengganggu model.
- Nilai kosong (missing value) diisi dengan median untuk data numerik dan modus untuk data kategorikal.
- Kolom seperti Mileage, Engine, dan Power diubah dari string (misalnya “20 km/kg”) menjadi angka murni.
- Data duplikat dan outlier dibuang menggunakan metode IQR (Interquartile Range).
- Setelah dibersihkan, dataset jadi lebih konsisten dan siap dipakai untuk modelling.



Modelling

Pada tahap modelling, data yang telah dibersihkan dibagi menjadi 80% data latih dan 20% data uji menggunakan train_test_split() dengan random_state=64 agar hasilnya konsisten.

Kolom kategorikal seperti Fuel_Type dan Transmission diubah menjadi angka melalui One Hot Encoding, sementara data numerik dinormalisasi menggunakan StandardScaler supaya skalanya seragam.

Setelah itu, model Linear Regression dilatih menggunakan data latih untuk mempelajari hubungan antara berbagai fitur mobil dan harga jualnya.

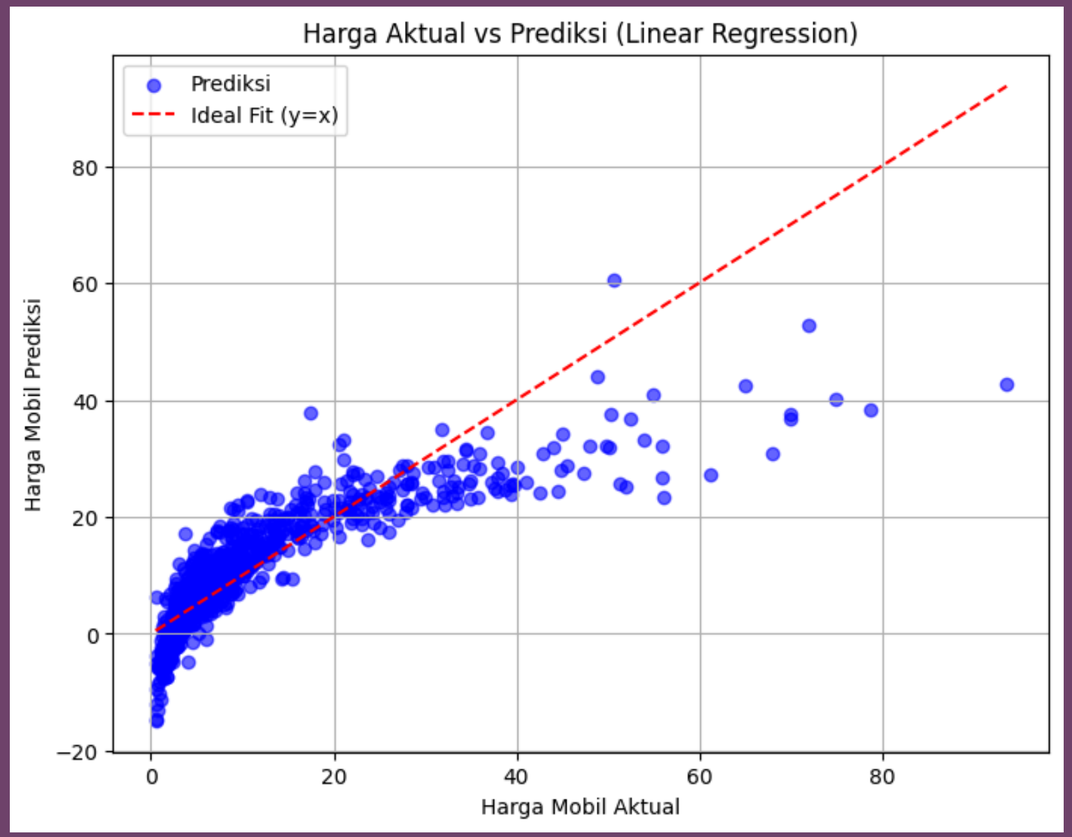
Model kemudian diuji dengan data uji untuk menilai seberapa baik kemampuannya dalam memprediksi harga mobil bekas.

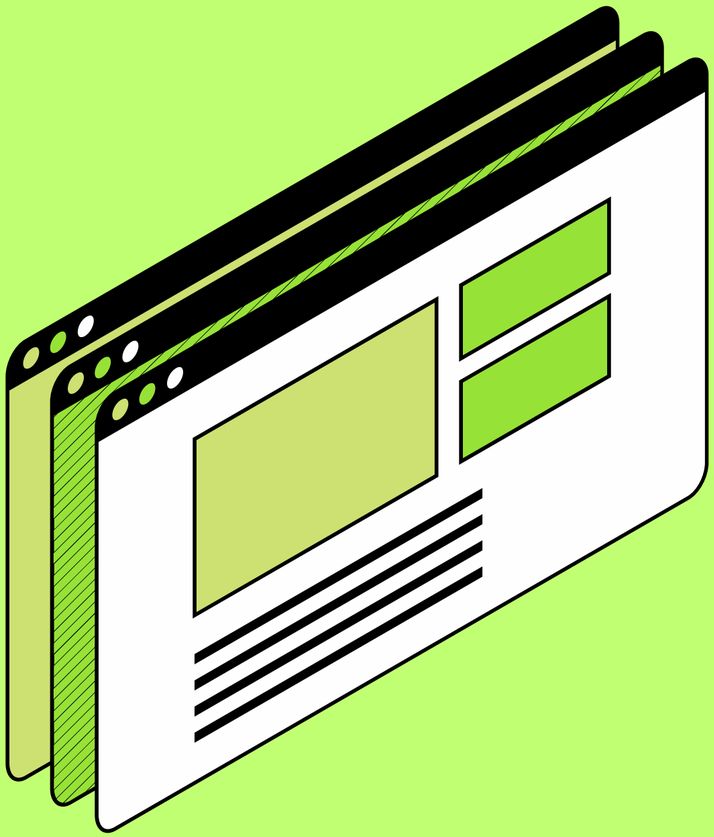
Hasil dan Evaluasi

Model diuji menggunakan data uji (X_test_scaled), dan hasil evaluasinya sebagai berikut:

- R^2 Score: 0.72
- MAE: 1.26
- MSE: 3.30
- RMSE: 1.81

Artinya, model mampu menjelaskan sekitar 72% variasi harga mobil bekas dari data yang ada.
Visualisasi scatter plot menunjukkan sebagian besar prediksi sudah mendekati nilai aktual, meskipun masih ada sedikit penyimpangan di beberapa titik.





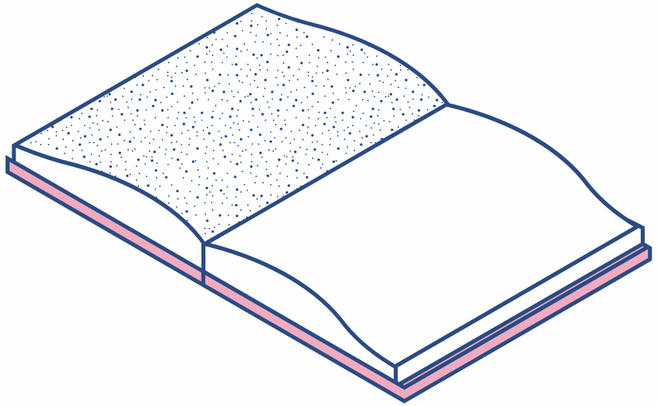
Kesimpulan

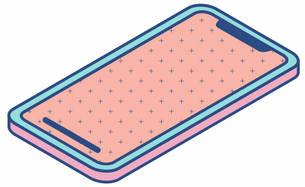
Model Linear Regression berhasil digunakan untuk memprediksi harga mobil bekas dengan hasil yang cukup baik, ditunjukkan oleh nilai R^2 sebesar 0.72.

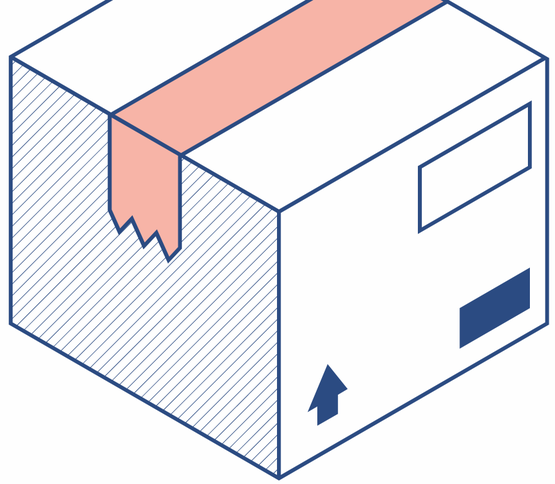
Fitur yang paling berpengaruh terhadap harga adalah tahun pembuatan, diikuti oleh kapasitas dan tenaga mesin.

Tahapan data cleaning dan scaling terbukti penting dalam meningkatkan performa model.

Meskipun hasilnya sudah memadai, masih ada beberapa prediksi yang kurang akurat — sehingga model bisa dikembangkan lebih lanjut menggunakan algoritma lain seperti Random Forest atau XGBoost untuk hasil yang lebih optimal.







TERIMA KASIH