Dosen Pengampu: Ratna Mufidah, S.Kom., M.Kom.

Data Mining

Memprediksi Harga Mobil Bekas Menggunakan

LINEAR REGRESSION

Present By :
Abdul Burhanudin
NPM: 2310631170001
Kelas: 5A Informatika

Fakultas Ilmu Komputer - UNSIKA

Pendahuluan

A Tujuan

Mengembangkan model prediksi berbasis machine learning untuk memperkirakan harga jual mobil bekas di India berdasarkan data spesifikasi dan kondisi kendaraan.

B Data set

Menggunakan dataset “Used Cars Price Prediction” yang diambil dari Kaggle, berisi sekitar 6.000+ data mobil dengan 12 atribut seperti: tahun, kilometer, bahan bakar, tenaga, dan harga.

C Model

Menerapkan algoritma Linear Regression dengan bantuan library scikit-learn (Python) untuk menganalisis hubungan antar fitur dan memprediksi harga.

Analisis ini dilakukan menggunakan Google Colab dengan bahasa pemrograman Python.

Tahap Proses

Pre-Processing

Profiling Data

EDA

Data Cleaning

Data Transformations

Data Final

Evaluasi Model

R^2

MAE

MSE

RMSE

Modelling

Memisahkan fitur & target

Membagi data training & testing

Data scalling

Model Linear Regression

Profilling data

Sumber Dataset (Github)

```
# load data Used Cars Price Prediction
!wget https://raw.githubusercontent.com/FarrelllAdityaaa/dataset-uts-datamining/refs/heads/main/used_cars_price_fiks.csv
```

Target

Price (Harga mobil bekas)

Kolom Numerik

8 kolom

Kolom Kategori

5 kolom

```
# Menampilkan informasi dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6019 entries, 0 to 6018
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Unnamed: 0         6019 non-null   int64
 1   Name               6019 non-null   object
 2   Location           6019 non-null   object
 3   Year               6019 non-null   int64
 4   Kilometers_Driven  5719 non-null   float64
 5   Fuel_Type          6019 non-null   object
 6   Transmission       6019 non-null   object
 7   Owner_Type         6019 non-null   object
 8   Mileage            6017 non-null   float64
 9   Engine             5983 non-null   float64
 10  Power              5876 non-null   float64
 11  Seats              5977 non-null   float64
 12  Price              6019 non-null   float64
dtypes: float64(6), int64(2), object(5)
memory usage: 611.4+ KB
```

Exploratory Data Analysis (EDA)

1. Analisis Distribusi Harga Mobil

Menggunakan histogram untuk melihat sebaran harga mobil.

2. Korelasi Antar Variabel Numerik

Menggunakan heatmap untuk melihat hubungan antara Engine, Power, Mileage, dan Price

3. Hubungan Tahun Produksi dengan Harga

Scatter plot menunjukkan mobil keluaran baru cenderung lebih mahal.

4. Perbandingan Berdasarkan Kategori

menggunakan boxplot untuk melihat perbedaan rata-rata harga di setiap kategori.

DATA CLEANING

1. Menangani Missing Value

memisahkan kolom numerik dan kategorikal, lalu mengisi nilai kosong dengan metode yang sesuai

2. Menghapus Duplikat dan Memastikan Tipe Data

Setelah nilai kosong diisi, saya memeriksa kembali data yang duplikat dan menghapusnya jika ada.

3. Cek Ulang Kondisi Data

Setelah dilakukan pembersihan, saya memastikan kembali bahwa tidak ada lagi missing value di seluruh kolom.

DATA TRANSFORMATION

1. Menghapus Kolom yang Tidak Diperlukan

Kolom Name saya hapus karena hanya berisi nama mobil, yang tidak berpengaruh secara langsung terhadap harga mobil.

2. Mengubah Data Kategorikal Menjadi Numerik

machine learning tidak dapat membaca data berupa teks,
jadi saya menggunakan metode One-Hot Encoding untuk mengubah kolom kategori menjadi variabel numerik (dummy variable).

3. Mengecek Hasil Transformasi

Kolom Name saya hapus karena hanya berisi nama mobil, yang tidak berpengaruh secara langsung terhadap harga mobil.

DATA FINAL

1. Mengecek Ukuran Dataset dan Nilai Kosong

memastikan kembali jumlah kolom dan baris, serta memastikan semua nilai kosong sudah terisi.

2. Mengecek Tipe Data

memeriksa kembali tipe data pada beberapa kolom untuk memastikan semuanya sudah dalam format numerik.

3. Melihat Data Siap Pakai

menampilkan beberapa baris awal data hasil preprocessing untuk memastikan data sudah seragam dan tidak ada kolom teks tersisa.

MODELLING

1. Memisahkan Fitur dan Target

2. Membagi Data Training dan Testing

3. Melakukan Data Scaling

4. Melatih Model Linear Regression

EVALUASI MODEL

Hasil Evaluasi Linear Regression

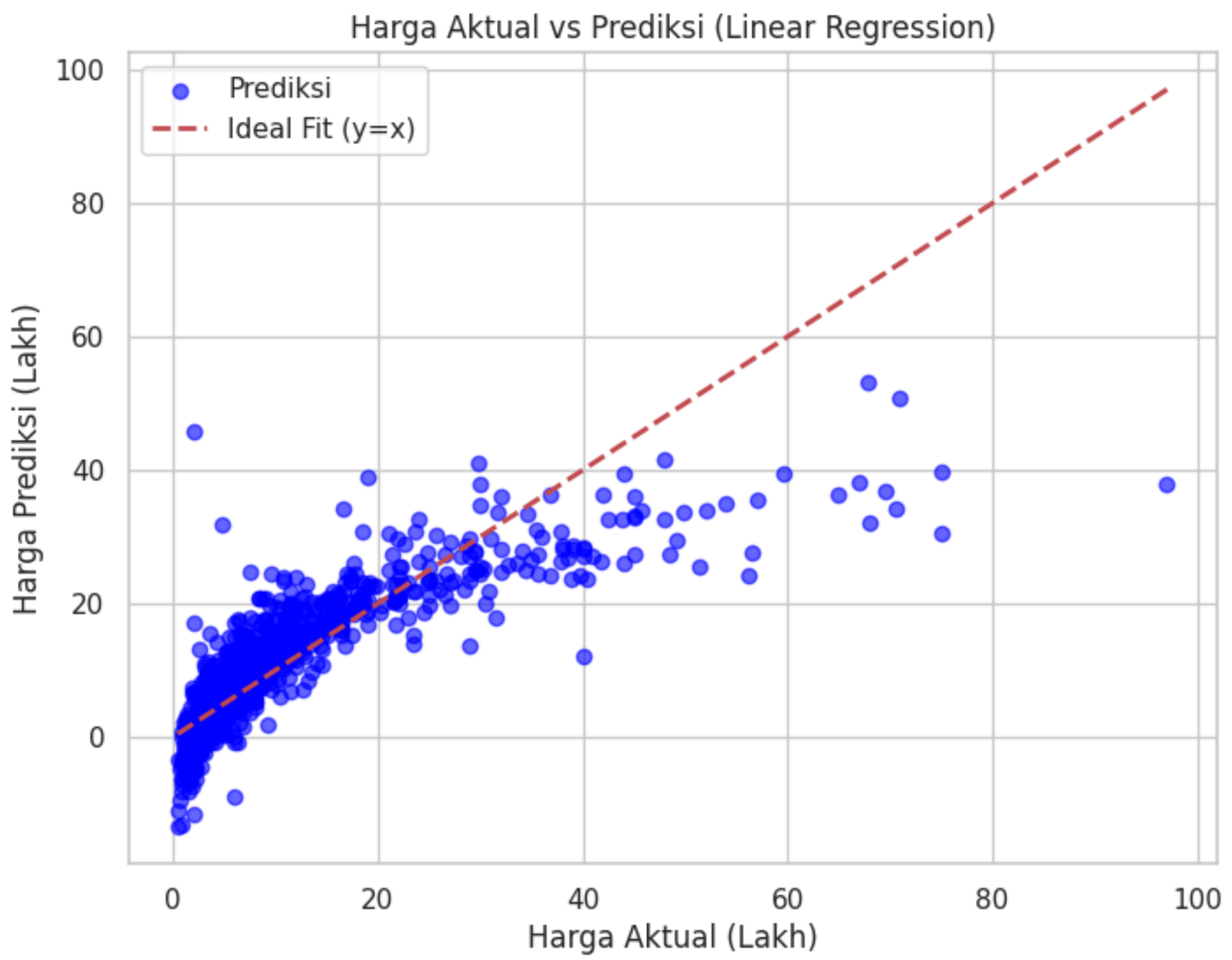
R^2 = 0.69 → Model mampu menjelaskan sekitar 69% variasi harga mobil,
menunjukkan performa yang cukup baik namun masih ada ruang untuk perbaikan.
MAE = 3.79 → Rata-rata selisih antara hasil prediksi dengan harga aktual sekitar 3,8 satuan harga.
MSE = 37.77 → Menunjukkan besarnya rata-rata error kuadrat, yang relatif moderat untuk dataset ini.
RMSE = 6.14 → Secara rata-rata, model salah memprediksi sekitar 6 satuan harga, menandakan hasil prediksi cukup dekat dengan nilai sebenarnya.

```
#Evaluasi
r2 = r2_score(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)

print("R² Score  :", r2)
print("MAE       :", mae)
print("MSE       :", mse)
print("RMSE      :", rmse)
```

```
R² Score  : 0.6930355425383958
MAE       : 3.798146536749495
MSE       : 37.77489016450676
RMSE      : 6.146128062813755
```

visualisasi dengan Scatter plot



Kesimpulan

Model Linear Regression yang dibangun mampu memprediksi harga mobil bekas dengan hasil yang cukup baik. Nilai R^2 sebesar 0.69 menunjukkan bahwa model dapat menjelaskan sekitar 69% variasi harga mobil, menandakan hubungan yang cukup kuat antara variabel-variabel fitur dengan harga mobil. Sementara itu, nilai MAE, MSE, dan RMSE menunjukkan tingkat kesalahan prediksi yang masih tergolong wajar. Dengan demikian, model ini sudah cukup akurat dan stabil untuk digunakan sebagai dasar dalam analisis awal maupun estimasi harga mobil bekas di masa mendatang.

THANK YOU

Salam hangat : Burhan