

# DATA MINING

DIBUAT OLEH MIRZA RAMADHAN

MULAI

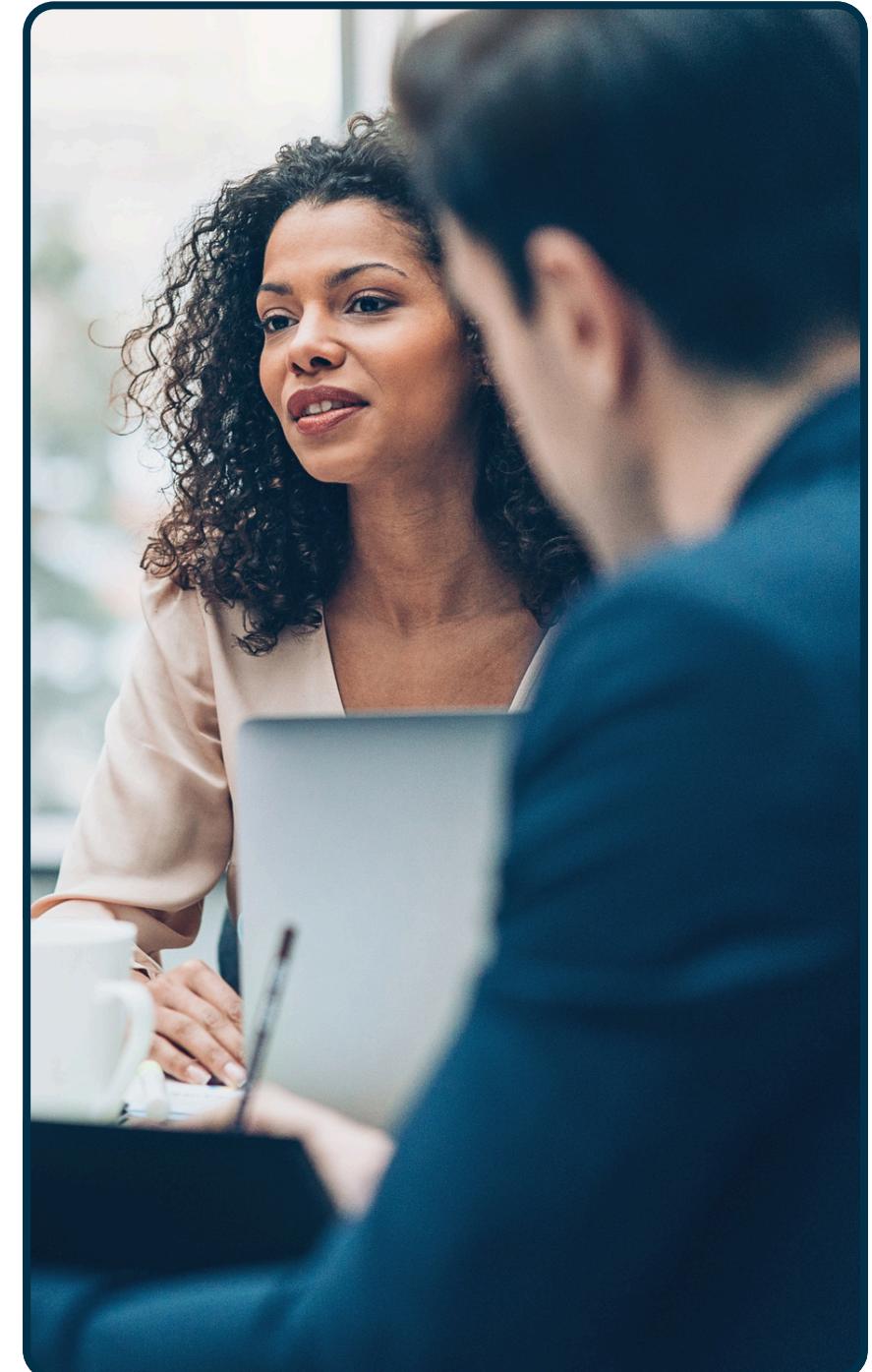
01



# PENDAHULUAN

Pasar mobil bekas terus berkembang pesat, menjadikannya pilihan populer bagi banyak konsumen. Namun, menentukan harga yang wajar untuk sebuah mobil bekas merupakan tantangan tersendiri, baik bagi penjual maupun pembeli. Harga dipengaruhi oleh berbagai faktor kompleks seperti merek, usia mobil, jarak tempuh, jenis bahan bakar, dan kondisi lainnya. Ketidakpastian ini seringkali menyebabkan transaksi yang kurang optimal.

Untuk mengatasi masalah tersebut, penelitian ini bertujuan untuk membangun sebuah model machine learning yang dapat memprediksi harga mobil bekas secara akurat berdasarkan atribut-atribut yang relevan. Dengan memanfaatkan data historis penjualan mobil, kita akan menerapkan model Regresi Linear untuk mengidentifikasi pola dan hubungan antara fitur-fitur mobil dengan harga jualnya.



02

Proses analisis akan melalui tiga tahap utama:

1. Preprocessing Data: Tahap ini mencakup pembersihan data dari nilai yang hilang dan tidak konsisten, analisis data eksplorasi (EDA) untuk memahami karakteristik data, serta rekayasa fitur (feature engineering) untuk menciptakan variabel baru yang lebih informatif.
2. Pemodelan (Modelling): Data yang telah bersih akan dibagi menjadi data latih dan data uji. Selanjutnya, model Regresi Linear akan dilatih menggunakan data latih.
3. Evaluasi Model: Kinerja model akan diukur menggunakan data uji dengan metrik evaluasi standar seperti R-squared (R<sup>2</sup>), Mean Absolute Error (MAE), dan Root Mean Squared Error (RMSE) untuk memastikan akurasi dan keandalannya.

Melalui pendekatan ini, diharapkan dapat dihasilkan sebuah model prediktif yang tidak hanya akurat tetapi juga dapat memberikan wawasan berharga mengenai faktor-faktor apa saja yang paling signifikan dalam menentukan harga mobil bekas.

# TUJUAN PENELITIAN

## Definisi Tujuan Penelitian

Tujuan utamanya adalah: Membangun model machine learning yang akurat dan andal untuk memprediksi harga jual mobil bekas berdasarkan fitur-fitur yang relevan.

### Tujuan Pertama

01

Mengidentifikasi faktor kunci yang paling mempengaruhi harga (seperti merek, usia, dan kilometer).



02

### Tujuan Kedua

Membuat alat praktis untuk membantu penjual dan pembeli dalam menentukan harga yang wajar.

●  
**04**  
●



# TAHAP 1 REPROCESSING DATA

## A. PROFILING DATA / PREPARATION

```
# import library yang diperlukan
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn import metrics

# memuat dataset dari file CSV
df = pd.read_csv('train-data.csv')

# menampilkan 5 baris pertama dari dataset
print("Data Awal (5 Baris Pertama):")
print(df.head())

# menampilkan informasi ringkas tentang tipe data dan nilai non-null
print("\nInformasi Dataset:")
df.info()

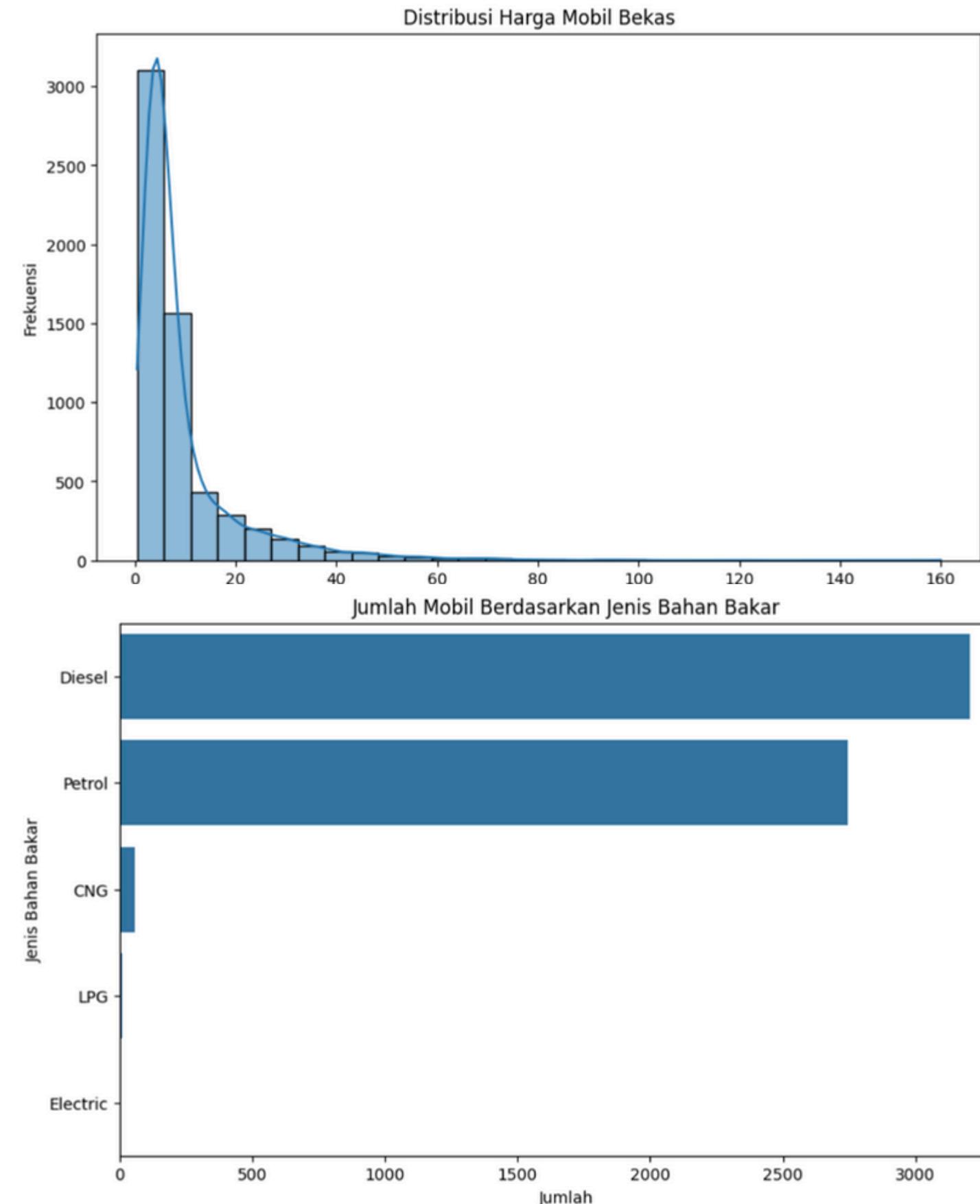
# menampilkan statistik deskriptif untuk kolom numerik
print("\nStatistik Deskriptif:")
print(df.describe())

# mengecek jumlah nilai yang hilang (missing values) di setiap kolom
print("\nJumlah Missing Values per Kolom:")
print(df.isnull().sum())
```

## B. EXPLORATORY DATA ANALYSIS (EDA)

```
# visualisasi distribusi harga mobil
plt.figure(figsize=(10, 6))
sns.histplot(df['Price'], kde=True, bins=30)
plt.title('Distribusi Harga Mobil Bekas')
plt.xlabel('Harga (dalam Lakh)')
plt.ylabel('Frekuensi')
plt.show()

# visualisasi jumlah mobil berdasarkan Jenis Bahan Bakar (Fuel_Type)
plt.figure(figsize=(10, 6))
sns.countplot(y='Fuel_Type', data=df, order = df['Fuel_Type'].value_counts().index)
plt.title('Jumlah Mobil Berdasarkan Jenis Bahan Bakar')
plt.xlabel('Jumlah')
plt.ylabel('Jenis Bahan Bakar')
plt.show()
```



## C. DATA CLEANING

```
# 1. menghapus kolom yang tidak relevan atau terlalu banyak missing values
# kolom Unnamed: 0 adalah duplikat dari index, dan New_Price memiliki >86% data hilang.
df.drop(columns=['Unnamed: 0', 'New_Price'], inplace=True)

# 2. membersihkan dan mengubah tipe data kolom 'Mileage', 'Engine', dan 'Power'
# fungsi untuk membersihkan string dan mengubahnya menjadi float
def clean_numeric_col(col):
    if isinstance(col, str):
        return float(col.split()[0])
    return col

df['Mileage'] = df['Mileage'].apply(clean_numeric_col)
df['Engine'] = df['Engine'].apply(clean_numeric_col)

# untuk kolom 'Power', ada string 'null bhp' yang perlu diubah menjadi NaN terlebih dahulu
df['Power'] = df['Power'].replace('null bhp', np.nan)
df['Power'] = df['Power'].apply(clean_numeric_col)

# 3. menangani missing values
# untuk kolom numerik, kita isi dengan nilai median
df['Mileage'] = df['Mileage'].fillna(df['Mileage'].median())
df['Engine'] = df['Engine'].fillna(df['Engine'].median())
df['Power'] = df['Power'].fillna(df['Power'].median())

# untuk kolom 'Seats', kita isi dengan modus
df['Seats'] = df['Seats'].fillna(df['Seats'].mode()[0])

# mengecek kembali missing values setelah cleaning
print("\nMissing Values Setelah Cleaning:")
print(df.isnull().sum())
```

## D. DATA TRANSFORMATIONS

```
# 1. feature Engineering: Membuat fitur 'Car_Age' dari 'Year'  
current_year = 2024  
df['Car_Age'] = current_year - df['Year']  
df.drop(columns=['Year'], inplace=True) # Hapus kolom Year asli  
  
# 2. menangani Variabel Kategorikal  
# ekstrak merek mobil dari kolom 'Name'  
df['Brand'] = df['Name'].apply(lambda x: x.split()[0])  
df.drop(columns=['Name'], inplace=True)  
  
# mengubah kolom kategorikal menjadi numerik menggunakan One-Hot Encoding  
categorical_cols = ['Location', 'Fuel_Type', 'Transmission', 'Owner_Type', 'Brand']  
df = pd.get_dummies(df, columns=categorical_cols, drop_first=True)
```

## E. DATA FINAL (PENGECEKAN ULANG)

```
print("\nData Final Setelah Preprocessing (5 Baris Pertama):")
print(df.head())

print("\nInformasi Dataset Final:")
df.info()
```

## TAHAP 2 MODELLING

```
# 1. Pisahkan fitur (X) dan target (y)
X = df.drop('Price', axis=1)
y = df['Price']

# 2. Split data menjadi data latih dan data uji (80% latih, 20% uji)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# 3. Data Scaling (Standard Scaler)
# Scaling PENTING untuk model linear agar fitur dengan skala besar tidak mendominasi
scaler = StandardScaler()

# Fit scaler HANYA pada data latih untuk menghindari data leakage
X_train = scaler.fit_transform(X_train)

# Transform data uji menggunakan scaler yang sudah di-fit
X_test = scaler.transform(X_test)

# 4. Melatih model Regresi Linear
model = LinearRegression()
model.fit(X_train, y_train)

print("\nModel Regresi Linear berhasil dilatih.")
```

Model Regresi Linear berhasil dilatih.

# TAHAP 3 EVALUASI MODEL

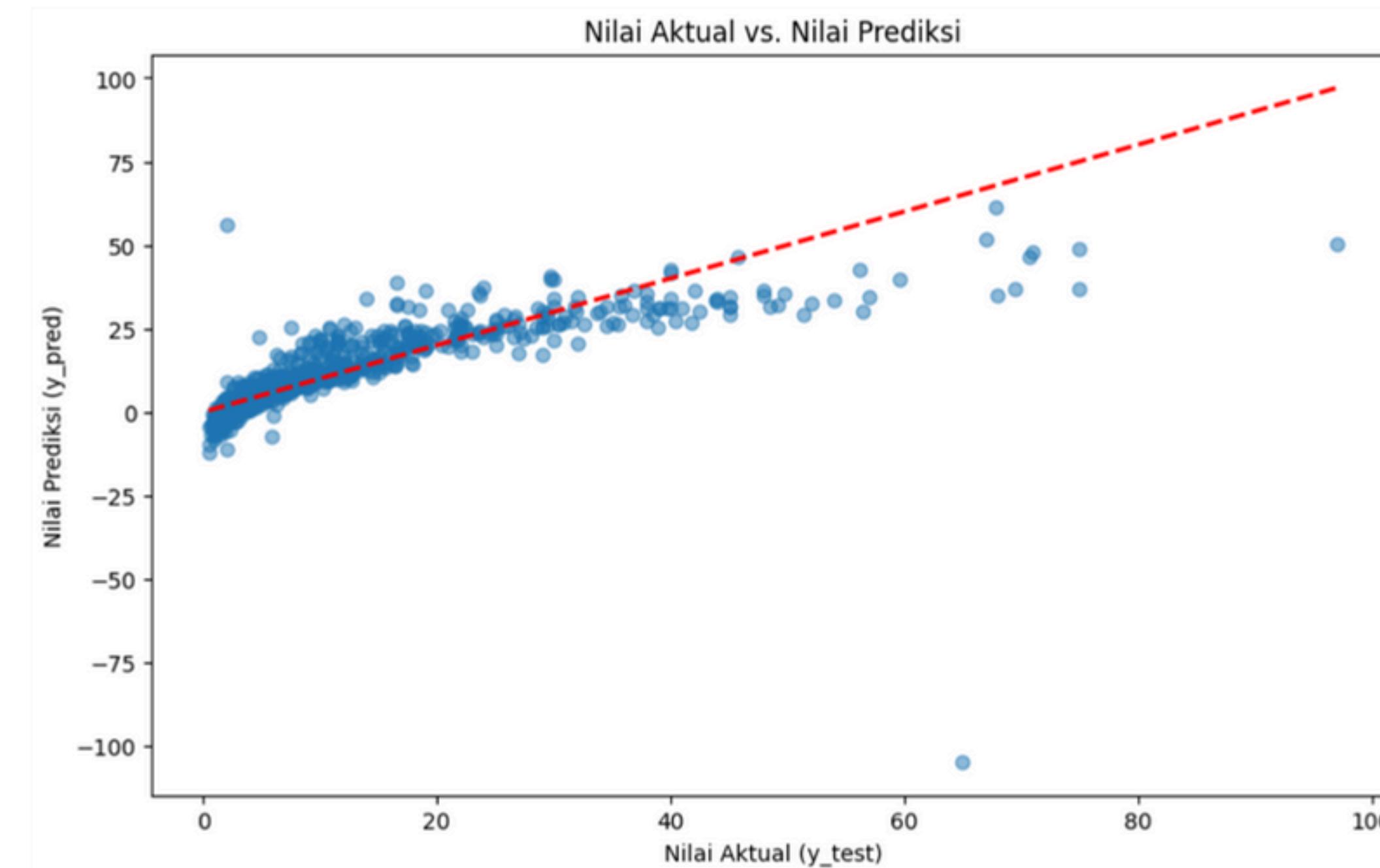
```
# 1. Membuat prediksi pada data uji
y_pred = model.predict(X_test)

# 2. Menghitung metrik evaluasi
r2 = metrics.r2_score(y_test, y_pred)
mae = metrics.mean_absolute_error(y_test, y_pred)
mse = metrics.mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)

print("\n--- Hasil Evaluasi Model ---")
print(f"R-squared (R2) Score: {r2:.4f}")
print(f"Mean Absolute Error (MAE): {mae:.4f}")
print(f"Mean Squared Error (MSE): {mse:.4f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.4f}")

# 3. Opsional: Visualisasi Hasil Prediksi vs Aktual
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred, alpha=0.5)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'r--', lw=2)
plt.title('Nilai Aktual vs. Nilai Prediksi')
plt.xlabel('Nilai Aktual (y_test)')
plt.ylabel('Nilai Prediksi (y_pred)')
plt.show()
```

# HASIL EVALUASI MODEL



# KESIMPULAN

**PROYEK INI BERHASIL MEMBANGUN SEBUAH MODEL REGRESI LINEAR UNTUK MEMPREDIKSI HARGA MOBIL BEKAS, YANG DIAWALI DENGAN TAHAP PREPROCESSING DATA YANG KRUSIAL UNTUK MEMBERSIHKAN DAN MENGUBAH DATA MENTAH MENJADI FORMAT YANG SIAP DIANALISIS. MODEL YANG DILATIH MENUNJUKKAN PERFORMA YANG KUAT DENGAN NILAI R-SQUARED (R<sup>2</sup>) SEBESAR 0.8732, YANG BERARTI MAMPU MENJELASKAN 87.3% VARIASI HARGA MOBIL. MESKIPUN RATA-RATA KESALAHAN PREDIKSI (MAE) SEKITAR 2.61 LAKH, MODEL INI TERBUKTI EFEKTIF SEBAGAI ALAT BANTU PENENTUAN HARGA YANG OBJEKTIF. UNTUK PENGEMBANGAN DI MASA DEPAN, DISARANKAN MENGGUNAKAN MODEL YANG LEBIH KOMPLEKS SEPERTI RANDOM FOREST UNTUK MENINGKATKAN AKURASI, TERUTAMA PADA SEGMENT MOBIL MEWAH.**

MIRZA RAMADHAN

# TERIMA KASIH ATAS PERHATIANNYA

Tidak Ada Kata Terima Kasih Yang Terlalu Kecil. Setiap Ungkapan Terima  
Kasih Adalah Bentuk Penghargaan Yang Besar

SELESAI