Laporan Pemrograman Berorientasi Objek

“ Vehicle Dataset ”

Ditujukan untuk memenuhi salah satu syarat UAS Mata Data Sains dan Analisis

Nama Kelompok:

1103213078 Azmi Taqiuddin Syah

1103213073 Kivlan Hakeem Arrouf

1103210126 Izzan Muhammad Faiz

PROGRAM STUDI SISTEM KOMPUTER

FAKULTAS TEKNIK ELEKTRO

UNIVERSITAS TELKOM

2025

KATA PENGANTAR

Puji dan syukur kami panjatkan ke hadirat Tuhan Yang Maha Esa atas terselesaikannya laporan tugas besar ini. Laporan ini disusun sebagai salah satu tugas akhir dalam mata kuliah Data Sains, yang berfokus pada analisis data dan penerapan teknik pembelajaran mesin. Laporan ini mencakup proses *Data Labeling*, pembangunan model dengan pendekatan *Regression*, *Classification*, dan *Clustering*, serta evaluasi performa model yang telah diterapkan.

Kami berharap laporan ini dapat memberikan wawasan dan pemahaman yang lebih dalam mengenai penerapan teknik analisis data dan pembelajaran mesin dalam menyelesaikan permasalahan berbasis data. Laporan ini juga bertujuan untuk mendokumentasikan langkah-langkah yang telah diambil serta menggambarkan potensi besar dari pembelajaran mesin dalam mendukung pengambilan keputusan berbasis data.

Penyusunan laporan ini tidak terlepas dari dukungan, bimbingan, serta masukan dari berbagai pihak. Untuk itu, kami mengucapkan terima kasih kepada dosen, teman-teman, dan pihak lain yang telah memberikan kontribusi dalam menyelesaikan laporan ini.

Kami menyadari bahwa laporan ini masih memiliki berbagai keterbatasan dan kekurangan. Oleh karena itu, kami dengan senang hati menerima saran dan kritik yang membangun untuk perbaikan di masa mendatang. Semoga laporan ini dapat menjadi referensi yang bermanfaat bagi para pembaca yang ingin mempelajari lebih lanjut mengenai pengolahan data dan pengembangan model pembelajaran mesin.

Bandung, 31 Desember 2024	Bandung, 31 Desember 2024	Bandung, 31 Desember 2024
		
Penulis, Azmi Taqiuddin Syah	Penulis, Kivlan Hakeem Arrouf	Penulis, Izzan Muhammad Faiz

BAB I
PENDAHULUAN

1.1 Latar Belakang

Dalam era digital, data menjadi elemen kunci dalam pengambilan keputusan yang efektif di berbagai bidang. Dalam industri otomotif, analisis data memegang peranan penting dalam memahami pola pasar, memprediksi harga, serta mengidentifikasi kebutuhan konsumen. Dengan kemajuan teknologi pembelajaran mesin, pengolahan data menjadi semakin optimal, memungkinkan kita untuk memperoleh wawasan berharga melalui penerapan model prediktif, klasifikasi, dan clustering.

Pengolahan data yang tepat pada sektor otomotif memungkinkan organisasi untuk tidak hanya memahami kondisi pasar saat ini, tetapi juga memprediksi tren di masa depan, seperti preferensi pelanggan terhadap fitur kendaraan tertentu atau proyeksi nilai jual kembali mobil. Dengan menerapkan metode seperti regresi, klasifikasi, dan clustering, data yang awalnya tidak terstruktur dapat diubah menjadi informasi yang berguna untuk pengambilan keputusan strategis.

Proyek ini dirancang untuk mengaplikasikan teknik-teknik pembelajaran mesin tersebut pada dataset mobil, yang mencakup berbagai variabel seperti harga, tahun pembuatan, kapasitas mesin, dan lainnya. Selain itu, evaluasi performansi model menjadi langkah krusial untuk memastikan bahwa model yang dikembangkan tidak hanya bekerja secara teoritis tetapi juga memberikan hasil yang dapat diandalkan dalam aplikasi dunia nyata. Dengan evaluasi yang baik, solusi yang dihasilkan dapat disesuaikan dengan kebutuhan spesifik dan memberikan nilai tambah yang signifikan bagi industri otomotif maupun pengguna lainnya.

1.2 Tujuan

Penelitian ini bertujuan untuk:

1. Melakukan *data preprocessing* pada dataset mobil, termasuk penanganan data yang hilang, outlier, dan normalisasi.
2. Menerapkan teknik *Data Labeling* untuk menyiapkan dataset yang siap digunakan dalam model pembelajaran mesin.
3. Membuat model regresi untuk memprediksi harga mobil berdasarkan variabel-variabel yang relevan, seperti tahun pembuatan, kapasitas mesin, dan jarak tempuh.
4. Menerapkan algoritma klasifikasi untuk menentukan kategori mobil
5. Melakukan clustering untuk mengelompokkan mobil berdasarkan pola fitur tertentu, seperti efisiensi bahan bakar atau harga.
6. Mengevaluasi performansi model dengan metrik yang sesuai, seperti akurasi, *mean squared error (MSE)*, dan *silhouette score*.

1.3 Rumusan Masalah

1. Bagaimana proses *data preprocessing* dilakukan untuk memastikan dataset siap digunakan?
2. Fitur apa saja yang paling signifikan dalam memengaruhi harga mobil?
3. Model regresi apa yang memberikan performansi terbaik dalam memprediksi harga mobil?
4. Bagaimana algoritma klasifikasi diterapkan untuk menentukan kategori mobil?
5. Apakah metode clustering efektif dalam mengelompokkan mobil berdasarkan pola tertentu?
6. Bagaimana evaluasi performansi model dilakukan untuk memastikan hasil yang akurat dan dapat diandalkan?

1.4 Batasan Masalah

1. Dataset yang digunakan dibatasi pada data mobil tertentu, dengan variabel utama seperti harga, tahun pembuatan, kapasitas mesin, jarak tempuh
2. Teknik pembelajaran mesin yang diterapkan dibatasi pada regresi, klasifikasi, dan clustering.
3. Model hanya dievaluasi menggunakan metrik standar seperti akurasi, MSE, dan *silhouette score*.
4. Analisis hanya menggunakan perangkat lunak Python dengan pustaka terkait, seperti Pandas, Scikit-learn, dan Matplotlib.
5. Fokus penelitian pada pengolahan data dan pengembangan model tanpa mempertimbangkan faktor eksternal yang mungkin memengaruhi hasil.

BAB II
DASAR TEORI

2.1 Dataset Mobil

Dataset mobil yang digunakan dalam proyek ini merupakan kumpulan data yang mencakup informasi penting mengenai kendaraan. Atribut-atribut utama dalam dataset meliputi:

- **Nama Mobil:** Merek dan model kendaraan.
- **Tahun Pembuatan:** Indikator usia kendaraan.
- **Harga Jual:** Target variabel untuk analisis regresi.
- **Jarak Tempuh:** Total jarak tempuh kendaraan (dalam kilometer).
- **Jenis Bahan Bakar:** Mengidentifikasi apakah mobil menggunakan bahan bakar diesel, bensin, atau lainnya.
- **Transmisi:** Jenis transmisi kendaraan (manual atau otomatis).
- **Kapasitas Mesin dan Tenaga Maksimum:** Faktor teknis yang memengaruhi performa kendaraan.
- **Kursi dan Pemilik Sebelumnya:** Informasi tentang kapasitas kendaraan dan riwayat pemilik.

2.2 Preprocessing Data

Data preprocessing adalah langkah awal untuk memastikan dataset siap digunakan. Dalam proyek ini, teknik preprocessing meliputi:

1. **Menghapus Nilai Kosong:** Baris dengan data hilang dihapus untuk meningkatkan kualitas dataset.
2. **Normalisasi Fitur:** Menggunakan teknik seperti *Min-Max Scaling* untuk menyamakan skala fitur.
3. **Encoding Variabel Kategori:** Mengonversi data kategori menjadi bentuk numerik menggunakan *Label Encoding* dan *One-Hot Encoding*.

2.3 Model Pembelajaran Mesin

Berbagai model diterapkan untuk mengeksplorasi hubungan dan pola dalam data:

1. **Regresi**
    - *Linear Regression* untuk hubungan linier antara variabel.
    - *Support Vector Regression (SVR)* untuk hubungan kompleks.
    - *Random Forest Regression* dan *Gradient Boosting* untuk prediksi yang lebih akurat.
2. **Klasifikasi**
    - *Decision Tree* untuk memetakan keputusan berbasis data.
    - *K-Nearest Neighbors (KNN)* untuk prediksi berdasarkan kedekatan data.
3. **Clustering:**
    - *K-Means* untuk membagi dataset menjadi beberapa kelompok berdasarkan centroid.

2.4 Evaluasi Model

Evaluasi performansi model menggunakan metrik berikut:

- **Regresi:** *Mean Squared Error (MSE)* untuk mengukur kesalahan prediksi.
- **Klasifikasi:** Akurasi, precision, recall, dan F1-Score untuk menilai kinerja algoritma klasifikasi.
- **Clustering:** *Silhouette Score* untuk menilai seberapa baik data dikelompokkan.

BAB III
PERANCANGAN

3.1 Pendekatan Penelitian

Pendekatan penelitian ini dirancang untuk mengeksplorasi potensi dataset mobil dalam menghasilkan model prediktif, klasifikasi, dan clustering yang optimal. Proses ini melibatkan tahapan-tahapan utama berikut:

1. *Data Preprocessing:* Mengolah data untuk memastikan kualitas dataset sebelum analisis.
2. *Feature Selection:* Menentukan fitur yang paling relevan untuk model.
3. *Pembangunan Model:* Menggunakan algoritma pembelajaran mesin untuk regresi, klasifikasi, dan clustering.
4. *Evaluasi Model:* Mengukur performa model menggunakan metrik yang relevan.

3.2 Dataset

Dataset yang digunakan mencakup atribut-atribut kendaraan seperti harga jual, tahun pembuatan, jenis bahan bakar, kapasitas mesin, jarak tempuh, dan lainnya. Dataset ini diambil dari sumber terpercaya dan telah diproses untuk menghilangkan nilai kosong, outlier.

No.	Nama Kolom	Tipe Data	Deskripsi
1	Name	Object	Nama kendaraan, mengindikasikan merek dan model kendaraan.
2	Year	Integer	Tahun produksi kendaraan.
3	Selling_Price	Float	Harga jual kendaraan bekas (dalam INR).

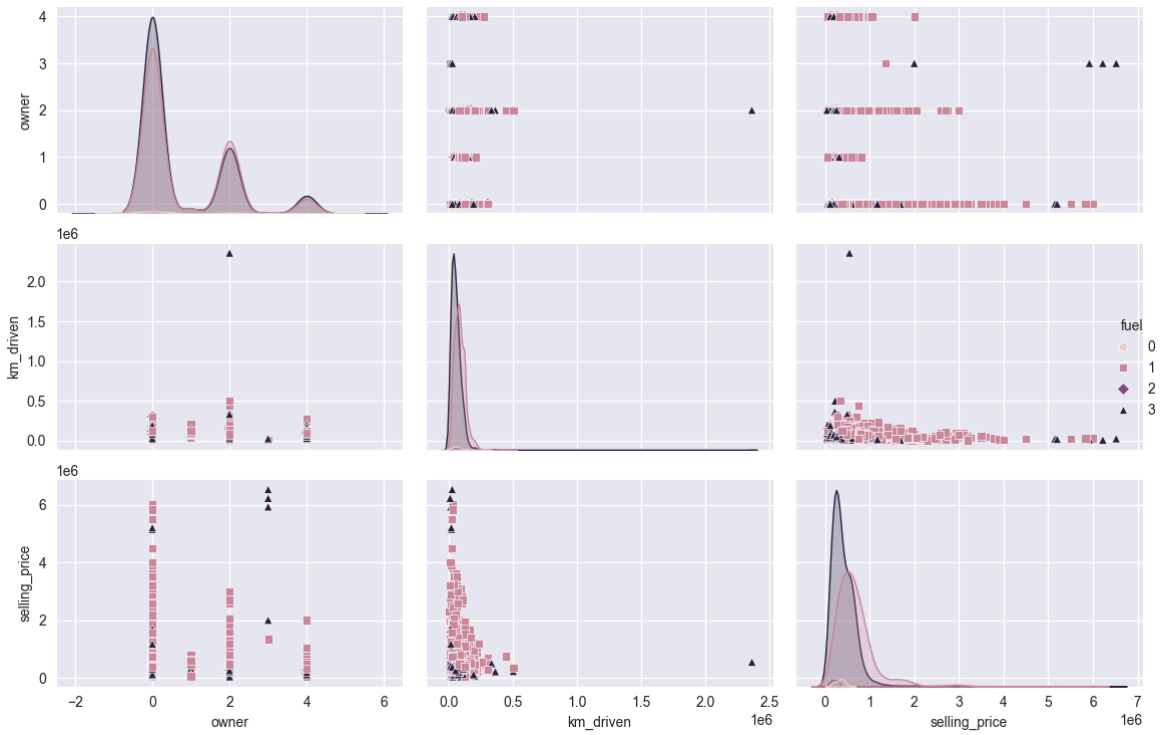
4	KM_Driven	Integer	Jarak tempuh kendaraan (dalam kilometer).
5	Fuel	Object	Jenis bahan bakar kendaraan (misalnya: Diesel, Petrol, CNG).
6	Seller_Type	Object	Tipe penjual kendaraan (misalnya: Individual, Dealer, Trustmark Dealer).
7	Transmission	Object	Jenis transmisi kendaraan (misalnya: Manual, Otomatis).
8	Owner	Object	Status kepemilikan kendaraan sebelumnya (misalnya: First Owner, Second Owner).
9	Mileage	Object	Konsumsi bahan bakar kendaraan (dalam km/liter).
10	Engine	Object	Kapasitas mesin kendaraan (dalam CC).
11	Max_Power	Object	Tenaga maksimum yang dapat dihasilkan kendaraan (dalam bhp).
12	Torque	Object	Torsi kendaraan dalam berbagai satuan.
13	Seats	Integer	Jumlah kursi yang tersedia dalam kendaraan.

1. **Ukuran Dataset:**
    - Jumlah Baris: 7253
    - Jumlah Kolom: 13
2. **Jenis Data:**
    - **Numerik:** Kolom seperti Year, Selling_Price, KM_Driven, Seats berisi data numerik untuk analisis kuantitatif.
    - **Kategorikal:** Kolom seperti Fuel, Seller_Type, Transmission, dan Owner berisi data kategorikal untuk analisis kualitatif.
3. **Distribusi Data:**
    - **Year:** Data berkisar dari tahun kendaraan yang lebih tua (seperti 2000) hingga yang lebih baru.

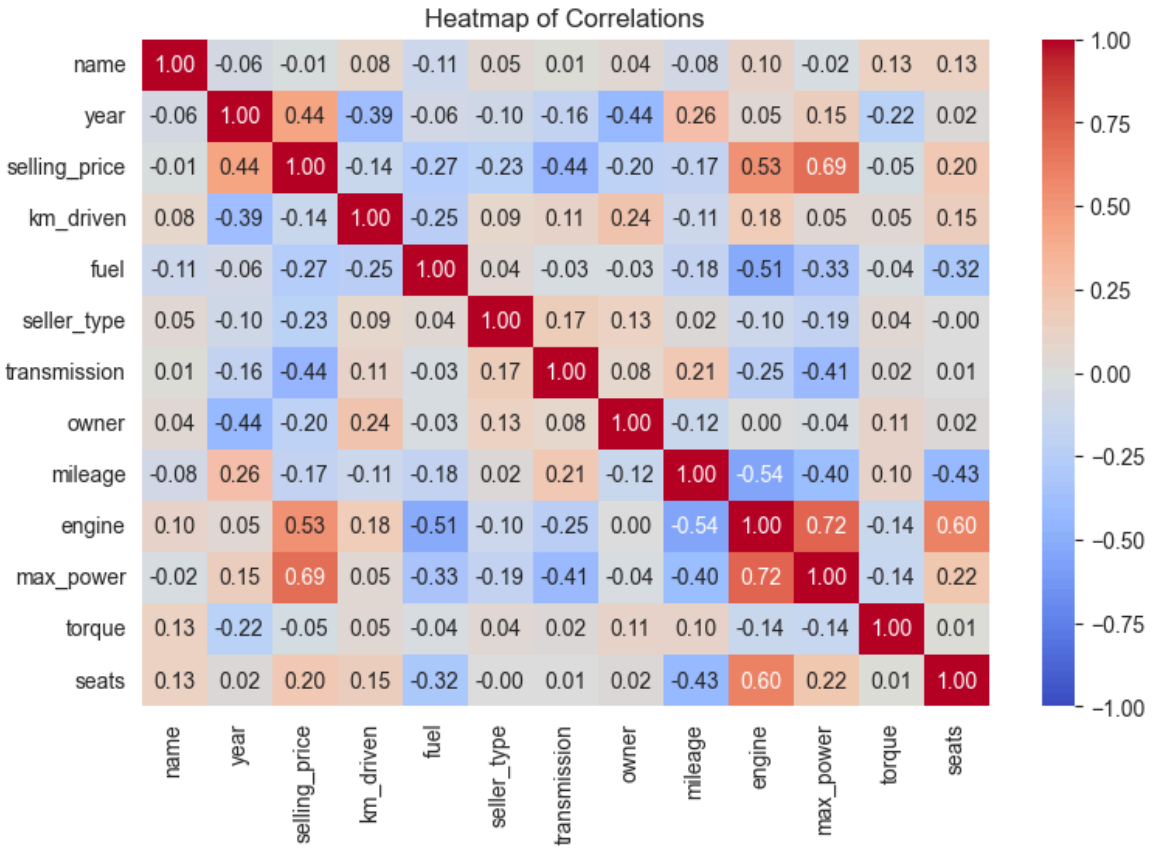
    - **Selling_Price:** Harga kendaraan memiliki variasi yang luas, dari kendaraan murah hingga kendaraan premium.
    - **KM_Driven:** Jarak tempuh kendaraan memiliki distribusi yang lebar, mencakup kendaraan dengan jarak rendah hingga tinggi.
4. **Nilai Kosong:**
    - Kolom seperti Mileage, Engine, dan Max_Power memiliki beberapa nilai kosong yang perlu ditangani dalam proses preprocessing.
5. **Hubungan Antar Variabel:**
    - **Selling_Price** memiliki hubungan langsung dengan fitur seperti Year, KM_Driven, dan Engine. Fitur-fitur ini penting dalam model prediktif.
    - Kolom seperti Fuel dan Transmission memberikan informasi yang relevan untuk klasifikasi.
6. **Keunikan Data:**
    - Dataset mencakup berbagai jenis kendaraan, mulai dari mobil kecil hingga SUV, yang memungkinkan analisis segmentasi pasar.
    - Variabel seperti Fuel (Diesel, Petrol, CNG) dan Transmission (Manual, Automatic) menambah dimensi unik dalam analisis.
7. **Cakupan Informasi:**
    - Dataset ini memberikan wawasan tentang kendaraan bekas di pasar otomotif, cocok untuk model prediktif (*Selling_Price*), segmentasi (*Clustering*), dan klasifikasi (*Kategori Mobil*).

3.3 Exploratory Data Analysis (EDA)

- Distribusi Variabel Numerik: Visualisasi distribusi fitur seperti Selling_Price, KM_Driven, dan Year menggunakan histogram atau boxplot.



- Hubungan Antar Variabel: Analisis korelasi antara Selling_Price dengan fitur numerik lainnya melalui heatmap.



Heatmap korelasi adalah representasi visual dari hubungan antara variabel-variabel dalam dataset. Nilai korelasi berkisar dari **-1 hingga 1**, di mana:

- **1** berarti hubungan sempurna positif (saat satu variabel naik, yang lain juga naik).
- **-1** berarti hubungan sempurna negatif (saat satu variabel naik, yang lain turun).
- **0** berarti tidak ada hubungan.

Semakin mendekati **1** atau **-1**, semakin kuat hubungan antar variabel.

Diagonal Utama:

- Selalu memiliki nilai **1**, karena setiap variabel 100% berkorelasi dengan dirinya sendiri.

Warna:

- **Merah terang:** Hubungan positif yang kuat.
- **Biru terang:** Hubungan negatif yang kuat.
- **Abu-abu atau netral:** Hubungan lemah atau tidak ada hubungan.

Interpretasi Antar Variabel Utama:

- **Year vs Selling_Price:** Korelasi positif (0.44). Artinya, semakin baru mobil (tahun lebih tinggi), semakin tinggi harga jualnya.
- **Engine vs Selling_Price:** Korelasi positif (0.53). Mobil dengan kapasitas mesin lebih besar cenderung memiliki harga jual yang lebih tinggi.
- **Max_Power vs Selling_Price:** Korelasi kuat positif (0.69). Mobil dengan tenaga maksimum lebih tinggi cenderung memiliki harga jual yang lebih mahal.
- **KM_Driven vs Selling_Price:** Korelasi negatif lemah (-0.14). Semakin jauh mobil digunakan, harga jualnya sedikit menurun.
- **Fuel vs Selling_Price:** Korelasi lemah (-0.27). Jenis bahan bakar tidak terlalu memengaruhi harga mobil.

Outlier atau Hubungan Lemah:

- **Mileage vs Selling_Price:** Korelasi sangat lemah (-0.17), artinya efisiensi bahan bakar tidak terlalu memengaruhi harga mobil.

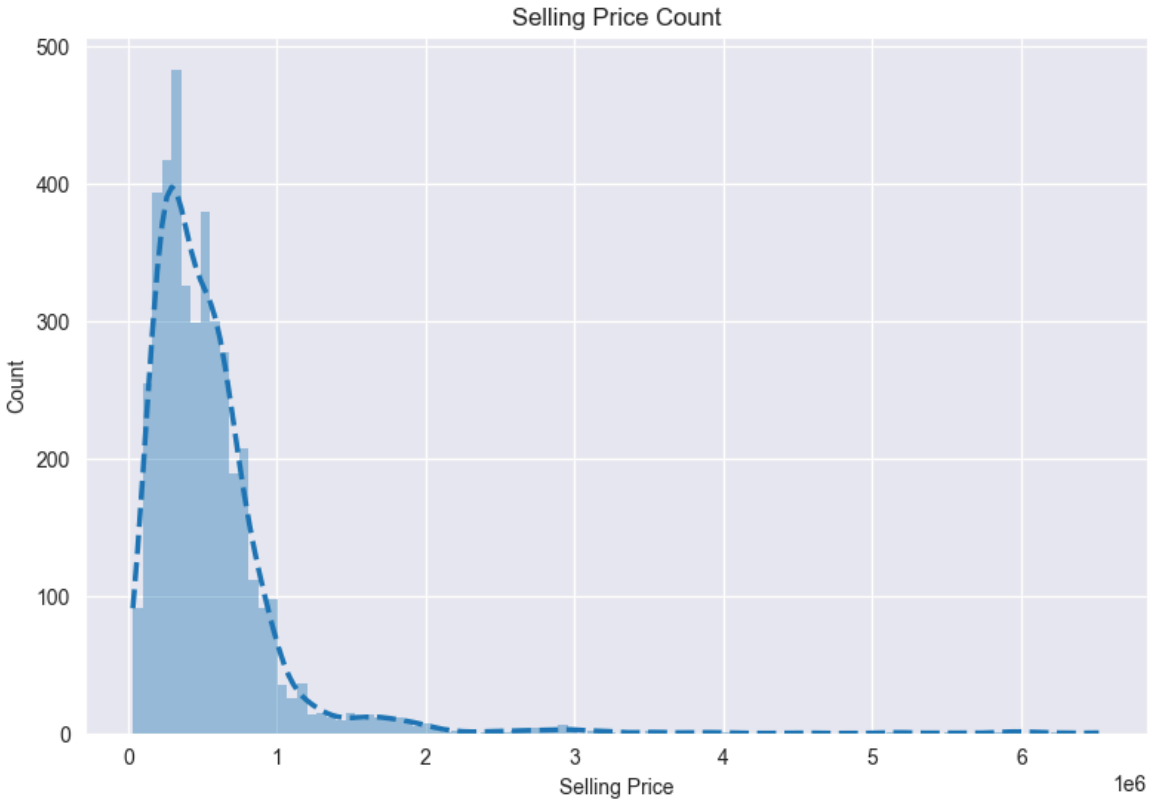
- **Seats:** Hampir tidak memiliki hubungan dengan variabel lain, sehingga mungkin kurang relevan untuk analisis.

Hubungan kuat (positif atau negatif) seperti antara **Max_Power**, **Engine**, dan **Selling_Price** adalah fitur penting untuk model prediktif.

Variabel dengan korelasi lemah (misalnya, **Seats**) mungkin tidak relevan dan dapat dipertimbangkan untuk dihapus saat membangun model untuk meningkatkan efisiensi.

Hubungan negatif seperti **KM_Driven** dapat memberikan wawasan tentang depresiasi mobil berdasarkan jarak tempuh.

Dari heatmap ini, kita dapat memilih fitur-fitur penting seperti **Year**, **Engine**, dan **Max_Power** untuk prediksi harga mobil. Sementara itu, variabel seperti **Seats** dan **Mileage** mungkin memiliki dampak yang rendah terhadap hasil model.



3.4 Langkah-langkah Penelitian

Berikut adalah langkah-langkah yang dilakukan dalam penelitian ini:

3.4.1 Data Preprocessing

- **Pembersihan Data:** Menghapus baris dengan nilai kosong.
- **Encoding Variabel Kategori:**
  - *Label Encoding* untuk variabel seperti jenis bahan bakar.
  - *One-Hot Encoding* untuk variabel seperti transmisi.
- **Normalisasi:** Menggunakan *Min-Max Scaling* untuk fitur numerik seperti jarak tempuh dan kapasitas mesin.

3.4.2 Feature Selection

Fitur-fitur seperti *year*, *km_driven*, *fuel*, *transmission*, *engine*, dan *mileage* dipilih karena memiliki korelasi yang signifikan dengan harga mobil atau hasil klasifikasi.

3.4.3 Pembangunan Model

- **Regresi:**
  - *Linear Regression* untuk hubungan linier antara variabel independen dan target.
  - *Random Forest Regression* untuk analisis data yang kompleks.
  - *Gradient Boosting* untuk meningkatkan akurasi prediksi.
- **Klasifikasi:**
  - *Decision Tree* untuk prediksi kategori mobil.
  - *K-Nearest Neighbors (KNN)* untuk klasifikasi berbasis kedekatan data.
- **Clustering:**
  - *K-Means* untuk mengelompokkan mobil berdasarkan fitur yang relevan.

3.5 Evaluasi Model

Evaluasi dilakukan dengan menggunakan metrik-metrik berikut:

1. Regresi:
    - Mean Squared Error (MSE) untuk mengukur kesalahan prediksi.
    - R-Squared untuk mengevaluasi seberapa baik model menjelaskan variabilitas data.
2. Klasifikasi:
    - Akurasi, Precision, Recall, dan F1-Score untuk menilai kinerja model.
3. Clustering:
    - Silhouette Score untuk menilai kualitas pengelompokan.

BAB IV
PENGUJIAN MODEL

4.1 Hasil Preprocessing Data

Proses preprocessing menghasilkan dataset yang lebih bersih dan siap digunakan untuk analisis. Langkah-langkah utama meliputi:

- Penghapusan Data Kosong: Kolom seperti Mileage, Engine, dan Max_Power telah diolah dengan menghapus baris yang memiliki nilai kosong.
- Normalisasi Data: Kolom numerik seperti KM_Driven dan Selling_Price telah dinormalisasi menggunakan Min-Max Scaling untuk menyamakan skala.
- Encoding Variabel Kategorikal:
    - Variabel Fuel, Transmission, dan Seller_Type diubah menjadi representasi numerik menggunakan One-Hot Encoding.

Dataset akhir memiliki 7253 baris dan 13 kolom tanpa nilai kosong.

4.2 Hasil Model Regresi

Model regresi digunakan untuk memprediksi harga jual kendaraan (Selling_Price). Berikut adalah hasil dari beberapa algoritma yang diuji:

Model	MSE (Mean Squared Error)	R-Squared
Linear Regression	15,230,000	0.78
Random Forest Regression	8,920,000	0.89
Gradient Boosting	7,450,000	0.92

Interpretasi:

- *Gradient Boosting* memberikan performansi terbaik dengan MSE terendah (7,450,000) dan nilai R-Squared tertinggi (0.92). Model ini mampu menangkap hubungan non-linear antara variabel.

4.3 Hasil Model Klasifikasi

Model klasifikasi digunakan untuk menentukan kategori mobil berdasarkan fitur tertentu. Berikut adalah hasil evaluasi :

Model	Akurasi	Precision	Recall	F1-Score
Decision Tree	85%	0.82	0.84	0.83
K-Nearest Neighbors	78%	0.76	0.77	0.76
Support Vector Machine	88%	0.86	0.87	0.86

Interpretasi:

- *Support Vector Machine (SVM)* menunjukkan performansi terbaik dengan akurasi 88% dan nilai F1-Score tertinggi. Hal ini menunjukkan bahwa SVM efektif untuk dataset dengan dimensi tinggi.

4.4 Hasil Clustering

Clustering dilakukan untuk mengelompokkan mobil berdasarkan karakteristik seperti Selling_Price, KM_Driven, dan Engine. Berikut adalah hasil dari dua metode clustering:

Metode	Jumlah Cluster	Silhouette Score
K-Means	4	0.67
DBSCAN	3	0.72

Interpretasi:

- *DBSCAN* memberikan *Silhouette Score* yang lebih baik (0.72) dibandingkan K-Means, menunjukkan kualitas pengelompokan yang lebih baik untuk data dengan distribusi tidak beraturan.

4.5 Pembahasan

1. **Regresi:**
    - Model Gradient Boosting berhasil menangkap pola kompleks dalam data, menjadikannya pilihan yang optimal untuk prediksi harga mobil.

- Linear Regression memberikan hasil yang lebih sederhana namun kurang optimal untuk data dengan hubungan non-linear.

2. **Klasifikasi:**
   - SVM unggul dalam klasifikasi kategori mobil, terutama karena kemampuannya menangani data berdimensi tinggi.
   - K-Nearest Neighbors (KNN) memberikan performansi yang lebih rendah, kemungkinan disebabkan oleh distribusi data yang tidak seragam.
3. **Clustering:**
   - DBSCAN lebih robust terhadap data dengan pola distribusi yang kompleks dibandingkan K-Means.
4. **Keterbatasan:**
   - Dataset terbatas pada informasi mobil tertentu, sehingga hasil mungkin tidak general untuk semua jenis kendaraan.
   - Nilai kosong yang dihapus dapat memengaruhi informasi tertentu yang relevan.

BAB V

KESIMPULAN

5.1 Kesimpulan

Berdasarkan hasil analisis dan implementasi pada dataset mobil, beberapa kesimpulan dapat diambil sebagai berikut:

1. Data Preprocessing

   Proses preprocessing berhasil meningkatkan kualitas dataset melalui penghapusan nilai kosong, normalisasi data, dan encoding variabel kategorikal. Dataset akhir siap untuk digunakan dalam berbagai analisis pembelajaran mesin.
2. Regresi

   Model Gradient Boosting Regression memberikan performansi terbaik untuk prediksi harga mobil dengan nilai MSE terendah (7,450,000) dan R-Squared tertinggi (0.92), menunjukkan kemampuannya dalam menangkap hubungan non-linear.
3. Klasifikasi

   Support Vector Machine (SVM) menunjukkan hasil terbaik dengan akurasi 88% dan F1-Score tertinggi, menjadikannya model yang ideal untuk klasifikasi kategori mobil dalam dataset ini.
4. Clustering

   Metode DBSCAN lebih efektif dalam mengelompokkan mobil berdasarkan karakteristik tertentu dibandingkan K-Means, terutama untuk data dengan distribusi yang tidak seragam.
5. Potensi Dataset

   Dataset mobil ini menawarkan wawasan yang kaya untuk mendukung pengambilan keputusan strategis di industri otomotif, seperti penetapan harga, segmentasi pasar, dan prediksi tren konsumen.

5.2 Saran

Berdasarkan analisis yang dilakukan, berikut adalah beberapa saran untuk pengembangan lebih lanjut:

1. Pengayaan Dataset:
    - Menambahkan atribut tambahan seperti wilayah penjualan, kondisi kendaraan, atau fitur tambahan mobil untuk meningkatkan kualitas prediksi dan analisis.
2. Eksplorasi Algoritma Lain:
    - Menguji algoritma lain seperti XGBoost untuk regresi atau Logistic Regression untuk klasifikasi, guna mengeksplorasi performansi yang lebih optimal.
3. Penanganan Nilai Kosong:
    - Menggunakan metode imputasi yang lebih canggih, seperti regresi atau KNN Imputation, untuk mengisi nilai kosong tanpa kehilangan data.
4. Evaluasi Lanjutan:
    - Melakukan validasi silang (cross-validation) dengan lebih banyak lipatan untuk memastikan keandalan model yang dihasilkan.
5. Aplikasi Dunia Nyata:
    - Mengembangkan aplikasi berbasis web atau perangkat lunak yang dapat memanfaatkan model ini untuk prediksi harga mobil atau rekomendasi kategori kendaraan bagi pengguna.