Project Report: Car Data Analysis

Introduction

This project analyzes a dataset of car attributes to uncover meaningful insights and patterns. The dataset includes information such as brand, model, manufacturing year, color, mileage, price, and condition. A range of data mining techniques were applied to explore trends, predict car condition, group similar vehicles, and estimate prices.

Key Findings

1. Exploratory Data Analysis (EDA):

- The dataset contains 292 entries, all complete with no missing values.
- Features are a mix of numerical (e.g., mileage, price) and categorical (e.g., brand, model, color, condition).
- A correlation matrix revealed:
  - A perfect negative correlation between Year and Age (-1), as expected.
  - Weak correlations among most other features.
- Visualizations highlighted patterns in distributions, particularly in mileage and price ranges.

2. Classification (Predicting Car Condition):

- A Decision Tree Classifier was trained to predict whether a car is New or Used based on variables like price, mileage, and age.
- Key patterns:
  - Lower mileage and higher price often indicate a New car.
  - Older vehicles with higher mileage are typically classified as Used.
- The model achieved reasonable classification accuracy, confirming its usefulness for basic condition prediction.

3. Clustering (Grouping Similar Cars):

- The Elbow Method was employed to determine the optimal number of clusters for K-Means clustering.
- Three distinct clusters were identified:
  - Cluster 1: High-priced, low-mileage New cars.
  - Cluster 2: Moderately priced cars with average mileage.
  - Cluster 3: Low-priced, high-mileage Used cars.

4. Association Rule Mining:

- Association rules were mined to find relationships among categorical features.
  - New cars typically show lower mileage and higher prices.
  - Used cars are generally associated with higher mileage and lower prices.
- Lift values were visualized, indicating weak to moderate associations between features.

5. Regression (Predicting Car Price):

- A Linear Regression model was built to predict car prices using variables such as mileage, year, and condition.
- The model yielded an R^2 value of ~0.45, indicating moderate predictive ability.
- Actual vs. Predicted Prices Visualization:

- The chart below compares real and estimated prices, revealing that while the model captures general trends, there is significant variability.

Visualization: Actual vs Predicted Prices

- Predictions are generally clustered within a limited range.
- The dashed diagonal line represents perfect predictions; deviations reflect prediction inaccuracies.
- The spread suggests the model may benefit from more features or a nonlinear approach.

Conclusion

This project demonstrates the value of data mining in analyzing and interpreting car-related data. Techniques such as classification, clustering, association mining, and regression provided actionable insights into car condition, pricing, and groupings. These findings can support better decision-making for both buyers and sellers.

Future Directions:

- Incorporate additional features (e.g., fuel type, transmission, engine specs) to enhance model accuracy.
- Explore advanced models (e.g., Random Forest, Gradient Boosting, or Neural Networks) to improve price prediction.
- Refine clustering and association analysis with dimensionality reduction and improved categorical encoding.