Name: Muhammad Faizan Majid

Roll no: Su92-BSDSM-F23-012

Project Report

**Topic: Predicting Used Car Prices**

**Git hub link**

https://github.com/Faizan-select/lab-task/blob/main/AI%20Lab/Predicting%20used%20car%20prices.ipynb

**Objective**

The objective of this project is to predict the prices of used cars based on various features such as mileage, engine capacity, kilometers driven, and year of manufacture. This involves preprocessing data, applying feature engineering, and using machine learning models to predict prices accurately.

**Data Preprocessing**

1. **Handling Missing Values**:
   * Columns with missing values (e.g., mileage, engine specifications) were filled using appropriate strategies such as imputing with the mode or extracting numerical values.
2. **Feature Engineering:**
   * **Kilometers Driven**: Numerical scaling was applied to standardize this column due to its wide range of values.
   * **Year**: The age of the car was calculated as the difference between the current year and the year of manufacture.
   * **Manufacturer Extraction**: The manufacturer was extracted from the Name column.
   * **Engine**: Engine capacity (in CC) was extracted and converted to a numeric format.
3. **Data Encoding:**
   * Categorical variables like Manufacturer were encoded using label encoding or one-hot encoding as needed.
4. **Splitting the Data:**
   * The dataset was split into training and testing sets using an 80-20 split to ensure proper validation.

**Exploratory Data Analysis**

Key insights derived from the data:

* Mileage and kilometers driven showed significant variability, impacting the car's price.
* Older cars (higher age values) tend to have lower prices.
* Engine capacity also correlated positively with the price.

**Modeling**

1. **Linear Regression:**
   * A baseline model was trained and evaluated.
   * R² score was used to measure performance.
2. **Random Forest Regressor**:
   * Achieved better performance than linear regression due to its ability to handle nonlinear relationships.
   * Confusion matrix and heatmaps were used to evaluate predictions for categorical price brackets.

**Evaluation**

* **Metrics:**
  + R² Score: Measures how well the model explains the variance in the price.
  + Confusion Matrix: Used to evaluate classification tasks if prices were categorized.
* **Random Forest Performance**:
  + Significantly outperformed other models with higher accuracy and robustness.

**Conclusion**

The project demonstrates a comprehensive approach to predicting used car prices. Random Forest was the most effective model, benefiting from its flexibility and ability to capture nonlinear relationships in the data. Future improvements could include:

* Hyperparameter tuning for models.
* Experimenting with advanced algorithms like Gradient Boosting or XGBoost.
* Using additional external data sources for features like market trends or fuel prices.

