Car prices predetor

This report summarizes the work carried out on the **Car Price Prediction** project using machine learning techniques.
The main goal of this project was to **build a predictive model** capable of estimating the price of a car based on multiple features such as the manufacturer, year of production, fuel type, transmission, and other specifications.

Dataset Description & Exploration

The dataset, provided in a CSV file, contained a large number of car listings with multiple attributes.

Key columns included:

- **Company Name: (e.g., Hyundai, Toyota, BMW)**
- **Model: Car model name or code**
- **Year of Manufacture: The production year of the car**
- **Fuel Type: Petrol, Diesel, Electric, Hybrid**
- **Transmission Type: Manual, Automatic, etc.**
- **Engine Capacity / Battery Size: Represented in CC or kWh**
- **Price: The target variable (car price in the market)**

Exploration steps performed:

- **Used df.head(), df.info(), and df.describe() to understand data structure and summary statistics.**
- **Inspected unique values in categorical fields like fuel type and transmission.**
- **Counted missing values with df.isnull().sum().**

Data Cleaning & Preprocessing

Data cleaning was crucial for ensuring the model received high-quality inputs.

✅ **Steps performed:**

- **Handling missing values:**
  - **Removed rows with critical missing values (like missing price or engine size).**
  - **Filled some non-critical missing entries with mean or mode values.**
- **Standardizing categorical data:**
  - **Normalized strings (e.g., "Petrol", "petrol", "PETROL" → "Petrol").**
- **Converting features into usable formats:**
  - **Converted "CC" or "kWh" values from text to numeric.**
  - **Encoded categorical variables (Fuel Type, Transmission) into numeric codes for ML algorithms.**
- **Outlier detection & treatment:**
  - **Removed extreme unrealistic entries (e.g., cars priced at $0 or absurdly high).**
- **Feature scaling:**
  - **Applied scaling to numerical values (using StandardScaler) to improve model performance.**

Exploratory Data Analysis (EDA)

Visual analysis was performed to gain insights into the dataset:

Techniques & graphs used:

- **Bar charts: Count of cars by manufacturer. (Showed which brands dominate the dataset – e.g., Hyundai, Toyota).**
- **Pie charts: Distribution of fuel types (Petrol was most common).**
- **Box plots: Price ranges per brand to identify expensive vs. affordable manufacturers.**
- **Heatmap (Correlation Matrix): Showed how strongly numerical features (Year, Engine Size) correlate with Price.**

Key insights from EDA:

- **Newer cars generally had higher prices.**
- **Engine capacity had a positive correlation with price.**
- **Petrol cars were more frequent in the dataset than diesel or hybrid cars.**

Model Building

The machine learning phase consisted of **training multiple regression algorithms** and comparing their performance.

Steps taken:

1. **Data Splitting:**
   - Dataset was split into **80% training** and **20% testing** using train_test_split().
2. **Models tested:**
   - **Linear Regression** – Baseline model for simple performance comparison.
   - **Random Forest Regressor** – Good for handling mixed feature types.
   - **Gradient Boosting Regressor (GBR)** – An ensemble model that builds strong predictions by combining multiple weak learners.
3. **Hyperparameter tuning (basic):**
   - For GBR, experimented with parameters like n_estimators, learning_rate, and max_depth to find the best combination.

Best Model:

- **Gradient Boosting Regressor** consistently produced **the highest accuracy**.

Model Evaluation

After training, models were evaluated using R^2 score and Mean Absolute Error (MAE).

Results for Gradient Boosting Regressor:

- **Training Accuracy: ~88.35%**
- **Test Accuracy: ~94.98%**
- **MAE: Low (indicating predictions were close to actual prices).**

This performance showed that the model generalized well — it didn't overfit on training data and performed well on unseen data.