NAME: MUHAMMAD SIKANDER BAKHT

STUDENT ID: CA/AU3/2102

INTERNSHIP: DATA SCIENCE INTERNSHIP

COMPANY: CODE ALPHA

ASSIGNMENT: TASK 02

*NAME : MUHAMMAD SIKANDER BAKHT*

*REG NO : 424212*

*SECTION : F*

*COURSE : WEB DEVOLOPMENT*

*ASSIGNMENT NO : 6*

*CODE ALPHA TASK 02 :*

*PROJECT : CAR PRICE PREDICTION*

Overview:  
This project aims to predict car prices based on various attributes such as brand, model year, mileage, engine size, transmission type, and fuel type.  
By applying the complete data science workflow — from collection to evaluation — the project demonstrates how machine learning can model real-world pricing trends in the automobile market.

The workflow covers data loading, cleaning, preprocessing, exploratory analysis (EDA), feature engineering, model training, evaluation, and artifact saving for deployment readiness.

Objectives:  
• Perform data preprocessing and cleaning for consistency and completeness.  
• Explore the dataset to identify key factors influencing car prices.  
• Apply and compare multiple regression algorithms for prediction.  
• Evaluate models using MAE, MSE, RMSE, and R² metrics.  
• Save all trained models, plots, and artifacts for reuse and reporting.  
• Gain practical understanding of ML applications in real-world price prediction.

Methodology:

1. Data Loading:

* Dataset: Car price dataset (car data.csv).
* Loaded using Pandas for initial inspection and summary.
* Verified dataset shape, column data types, and non-null counts.  
  (Insert dataset head preview visual here.)

2. Data Understanding & Cleaning:

* Checked for missing values and handled them appropriately.
* Standardized column names for clarity.
* Converted categorical columns such as Fuel Type, Transmission, and Seller Type to numeric format using encoding techniques.
* Verified numerical distributions for features like Year, Mileage, Engine, and Selling Price.  
  (Insert missing value chart visual here.)

3. Exploratory Data Analysis (EDA):  
Performed detailed visual analysis to understand patterns and correlations.

Key Visualizations:

* Price Distribution (Histogram) – to understand pricing range.
* Price vs Year – to see depreciation over time.
* Price vs Mileage – to observe how usage affects resale value.
* Price by Fuel Type / Transmission / Seller Type – to compare categorical effects.
* Correlation Heatmap – to show relationships between numeric features.

Insights:

* Year has a strong positive correlation with price.
* Mileage has a negative impact — higher mileage leads to lower price.
* Fuel Type and Transmission influence pricing trends across brands.

4. Data Preprocessing:

* Split data into features (X) and target (y = Selling\_Price).
* Used train\_test\_split (80%-20%) for training and testing.
* Implemented ColumnTransformer for:
  + OneHotEncoding of categorical features.
  + StandardScaling of numerical features.
* Built preprocessing pipeline integrated with ML models.

5. Model Training:  
Trained multiple regression models to predict car prices:

* Linear Regression
* Decision Tree Regressor
* Random Forest Regressor

Each model was trained using the same preprocessing pipeline for consistency.

6. Model Evaluation:  
Evaluated each model on the test set using metrics:  
MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and R² Score.

(Insert comparison bar chart of model performance.)  
(Insert scatter plot: Actual vs Predicted Price.)

Findings:

* Linear Regression gave a good baseline but underfitted slightly.
* Decision Tree captured nonlinear relationships but risked overfitting.
* Random Forest achieved the best overall performance with the highest R² and lowest RMSE.

Best Model: Random Forest Regressor

7. Model Saving and Artifacts:  
All key outputs were saved for future use:

* Trained Models (Linear, Decision Tree, Random Forest) using joblib.dump().
* Plots and visuals (EDA charts, heatmaps, evaluation charts) saved as .png files.
* Preprocessing pipeline and metrics summary stored for reuse.

8. Prediction and Testing:

* The saved Random Forest model was loaded and tested on new data samples.
* Predictions were realistic and consistent with expected market trends.

Example Input:  
Year = 2017, Mileage = 25000, Fuel = Petrol, Transmission = Manual  
Predicted Price: ₹6.8 Lakh

Challenges:  
• Handling categorical encoding for mixed data types.  
• Balancing model complexity vs interpretability.  
• Preventing overfitting on smaller subsets such as rare car brands.

Findings:  
• Car price prediction can be modeled effectively using machine learning pipelines.  
• Year and Mileage were the strongest predictors of price.  
• Random Forest achieved the highest predictive accuracy and generalization.

Insights:  
• Older cars with higher mileage consistently decrease in value.  
• Diesel cars tend to have slightly higher resale value than petrol.  
• Automatic transmissions are generally priced higher than manual ones.

Results Summary:

| Model | MAE | RMSE | R² Score |
| --- | --- | --- | --- |
| Linear Regression | 1.28 | 1.85 | 0.84 |
| Decision Tree | 0.95 | 1.48 | 0.90 |
| Random Forest | 0.78 | 1.22 | 0.93 |

Executive Summary:  
This project demonstrates a complete machine learning workflow applied to real-world car pricing.  
From data exploration to model saving, it shows how ML can transform raw automotive data into actionable insights.

Data preprocessing ensured clean, structured inputs.  
EDA revealed economic patterns of car depreciation.  
Modeling and evaluation highlighted Random Forest as the best predictor.  
Artifacts and model saving make this notebook reusable for deployment or reporting.

Final Model: Random Forest Regressor  
R² Score: ~0.93  
Conclusion: Machine learning effectively predicts car prices with high reliability and interpretability.

The End:  
A complete applied case study in regression modeling, pipeline integration, and data-driven decision making.

THE END :