PROJECT TITLE:

Predicting Vehicle Fuel Efficiency Using Machine Learning

INTRODUCTION

In the transportation and automotive domain, accurately predicting a vehicle's fuel efficiency is essential for developing energy-efficient designs, complying with environmental regulations, and guiding consumer choices.

OBJECTIVE

Traditional fuel efficiency predictions often rely on static measurements or simple correlations, which may fail to capture complex interactions between vehicle characteristics.

This project aims to improve the prediction of fuel efficiency (miles per gallon - MPG) using machine learning models that analyze vehicle-specific attributes.

INFORMATION RELATED TO DATASET

The dataset used in this project contains various attributes related to vehicles and their corresponding MPG (miles per gallon), which is the target variable.

There are a total of **8 features**, broadly categorized into

Feature Name	Description
fuel efficiency	Distance covered per gallon of fuel consumed
Engine config	Total number of combustion chambers (cylinders) in the engine
engine volume	Engine capacity measured in cubic inches
power output	Maximum engine power output measured in horsepower
vehicle mass	Overall weight of the vehicle in pounds

Accel capability	Time taken to accelerate from 0 to 60 mph (in seconds)
release year	Model release year (two-digit format)
manufacture region	Region where the car was manufactured (1 = USA, 2 = Europe, 3 = Japan)

METHODOLOGY

The task is to **predict MPG**, a continuous numerical variable, based on various vehicle characteristics. Hence, it is a **Regression** problem.

Algorithms/Models Applied

- **Linear Regression:** Used as a baseline model to capture linear relationships between features and MPG.
- **Decision Tree Regressor:** To handle non-linear relationships and feature interactions.
- **XGBoost Regressor:** Applied for its robustness and superior predictive accuracy in complex regression tasks.

Tools & Libraries Used

- **Python**
- **NumPy, Pandas:** Data loading and preprocessing
- **Matplotlib, Seaborn:** Data visualization
- **Scikit-learn:** Model implementation and evaluation
- **XGBoost:** Gradient boosting model for regression

IMPLEMENTATION

- **Data Cleaning:**
    - Handled missing values (especially in horsepower)

    - Converted categorical columns (e.g., Origin) using one-hot encoding
    - Checked for data consistency and feature correlations
- **Feature Scaling:**
    - Standardized numerical features to bring them to a common scale (especially for models sensitive to scale like linear regression)
- **Train-Test Split:**
    - 80% of the data used for training
    - 20% reserved for testing

Evaluation Metrics Used

As this is a regression task, the following metrics were used to evaluate model performance:

- **R^2 Score:** Proportion of variance explained by the model
- **MSE (Mean Squared Error):** Average of squared differences between predicted and actual values
- **RMSE (Root Mean Squared Error):** Square root of MSE
- **MAE (Mean Absolute Error):** Average absolute error in predictions