Project Documentation

# 1. Overview

This project is focused on predicting car prices using data from multiple Indian cities. It involves the following steps:  
1. Data cleaning and transformation  
2. Feature extraction  
3. Model training and hyperparameter tuning  
4. Model evaluation  
5. Saving the best model for future use.

# 2. Methodology

## 2.1 Data Cleaning and Preprocessing

- Files: Data from six Indian cities: Bangalore, Delhi, Kolkata, Hyderabad, Jaipur, and Chennai were used.  
- Columns Cleaned: 'new\_car\_overview', 'new\_car\_feature', and 'new\_car\_specs' were cleaned using the function clean\_column\_data.  
- Missing Values: Numerical columns were filled with the mean, and categorical columns were filled with the mode.  
- Encoding: One-hot encoding was used for nominal categorical columns, and label encoding was applied to ordinal columns.  
- Outlier Handling: Capping of extreme values was performed using the 5th and 95th percentiles.

## 2.2 Feature Engineering

Key features like kilometers driven, model year, seats, and engine displacement were extracted.  
Numerical columns were normalized using MinMax scaling.

## 2.3 Model Building

Six different regression models were tested:  
1. Linear Regression  
2. Lasso (L1)  
3. Ridge (L2)  
4. Decision Tree  
5. Random Forest  
6. Gradient Boosting

## 2.4 Hyperparameter Tuning

GridSearchCV was used for hyperparameter tuning for the Random Forest, Gradient Boosting, Lasso, and Ridge models.

# 3. Evaluation Metrics

The following metrics were used to evaluate the models:  
- Mean Squared Error (MSE)  
- Mean Absolute Error (MAE)  
- R² Score  
  
The best model was selected based on the lowest MSE, MAE, and highest R² score.

# 4. Results

The best model identified was Gradient Boosting. The following metrics were obtained for the test data:  
- Test MSE: <value>  
- Test MAE: <value>  
- Test R²: <value>

# 5. Model Saving

The best model was saved using joblib as 'best\_model.pkl' and the columns used for training were saved in 'model\_columns.pkl'.

# 6. Visualization

Several visualizations were created to explore the data:  
1. Scatter plot: Kilometers driven vs. price.  
2. Histogram: Distribution of model year.  
3. Box Plot: Price distribution by the number of owners.  
4. Correlation Heatmap: Numerical features.