**Model Development and Selection – Documentation Report**

**1. Objective**

The goal of this phase was to build and evaluate multiple machine learning models to predict the prices of used cars accurately, and then select the best performing model for deployment based on measurable evaluation metrics.

**2. Dataset Used**

* **Input:** Preprocessed structured dataset (preprocessed\_dataset.csv) containing cleaned and normalized car data.
* **Target Variable:** prize(in lakhs) (price of used cars in Indian rupees, normalized to lakhs).

**3. Train-Test Split**

| **Aspect** | **Setting** |
| --- | --- |
| Train : Test Split Ratio | 80:20 |
| Random State | 42 (to ensure reproducibility) |

The data was randomly split, with 80% used for training and 20% held out for testing.

**4. Models Trained**

| **Model** | **Purpose** |
| --- | --- |
| **Linear Regression** | Baseline model for checking linear relationships. |
| **Ridge Regression** | Linear model with L2 regularization to handle small multicollinearity and overfitting. |
| **Decision Tree Regressor** | Non-linear model capturing decision rules from the data. |
| **Random Forest Regressor** | Ensemble model combining multiple decision trees to improve accuracy and reduce variance. |
| **Gradient Boosting Regressor** | Advanced boosting technique for handling bias and variance trade-offs. |

**5. Performance Metrics Evaluated**

| **Metric** | **Meaning** |
| --- | --- |
| MAE (Mean Absolute Error) | Average absolute difference between predicted and actual prices. |
| MSE (Mean Squared Error) | Square of errors, penalizing larger errors more. |
| R² Score (Coefficient of Determination) | Proportion of variance in target explained by the model; closer to 1 is better. |

**6. Model Performance Summary**

| **Model** | **MAE** | **MSE** | **R² Score** |
| --- | --- | --- | --- |
| Ridge Regression | 0.7955 | 1.8437 | 0.8685 |
| Linear Regression | 0.7962 | 1.8492 | 0.8681 |
| Random Forest Regressor | 0.8105 | 1.9877 | 0.8582 |
| Gradient Boosting Regressor | 1.2155 | 3.1523 | 0.7752 |
| Decision Tree Regressor | 1.2237 | 3.4721 | 0.7524 |

**7. Key Observations from Results**

| **Point** | **Inference** |
| --- | --- |
| **Ridge Regression performed best** | Achieved highest R² score (0.8685) with lowest MAE (0.7955) among all models. |
| Linear Regression was very close | However, Ridge slightly outperformed Linear due to regularization stabilizing predictions. |
| Tree-based models underperformed | Random Forest and Gradient Boosting had lower R² and higher MAE/MSE, indicating they were either overfitting or unable to model the problem better than Ridge. |
| Decision Tree was the weakest | Overfit-prone, less generalizable compared to ensemble methods or regularized linear models. |

**Ridge Regression emerged as the most suitable model for price prediction in this dataset.**

**8. Why Ridge Regression was Selected**

| **Reason** | **Explanation** |
| --- | --- |
| Stability | Ridge regression penalizes large coefficients, preventing overfitting. |
| Generalization | Provides a smoother, more generalizable model compared to Decision Trees or Random Forests. |
| Simplicity | Easier to deploy, explain, and maintain compared to complex ensemble models. |
| Best Performance | Achieved the best combination of MAE, MSE, and R² Score across all models tested. |
| Matches Data Structure | Since the car price behavior is relatively smooth and partially linear (newer cars, fewer kilometers → higher price), Ridge fits well. |

**Final Conclusion:**

**Ridge Regression was selected as the final model for deployment**  
 **It balances accuracy, stability, and practical deployment readiness.**