Used Car Price Prediction

Avneesh Jarangal, COPC, ajarangal_be22@thapar.edu
Prerit Bhagat, COPC, pbhagat_be22@thapar.edu

1. **Introduction:**
   A. Technical background of project:

   With the rise of e-commerce platforms for used cars, accurate pricing mechanisms are vital. This project explores the application of machine learning to predict car prices by analyzing data attributes like mileage, age, and brand. This shift from manual valuation to data-driven predictions improves accuracy and transparency.

   B. Technical Concepts used:
   - **Regression Models:** Linear Regression and advanced tree-based algorithms.
   - **Feature Engineering:** Transforming categorical and numerical data.
   - **Evaluation Metrics:** Mean Absolute Error (MAE), R2-score, etc.
   - **Model Optimization:** Hyperparameter tuning via grid search and cross-validation.

   C. Motivation:

   The ambiguity in used car pricing motivates the creation of a tool for accurate price predictions, empowering buyers and sellers with reliable insights.

   D. Problem Statement:

   To design and implement a machine learning system capable of predicting a car's price based on historical data, addressing variables like brand, model, fuel type, and mileage.

   E. Area of application:

   Marketplaces: Enhancing customer trust with dynamic pricing.
   Dealerships: Aiding in trade-ins and inventory management.
   Finance and Insurance: Valuations for loans and premium calculations.

   F. Dataset and input format:

   The dataset includes detailed records of used cars with attributes such as:

   - **Manufacturing Year:** The year the car was produced.
   - **Brand/Company:** The manufacturer or make of the car.
   - **Mileage:** Total distance driven by the vehicle.
   - **Fuel Type:** The type of fuel used (e.g., petrol, diesel).

2. **Objective**
   A. Main Objective:

      The primary objective of this project is to develop a machine learning-based predictive model that accurately estimates the price of used cars. The model aims to analyze key factors such as manufacturing year, brand, mileage, fuel type, and transmission. This ensures a reliable and data-driven pricing mechanism for buyers, sellers, and stakeholders in the automobile market. By leveraging historical data and advanced algorithms, the project seeks to address the ambiguity and inefficiency often associated with manual car price evaluations.

3. **Methodology**
   A. Steps:

1. Problem Definition & Objective

- Define the objective: Predict the price of used cars based on features like year, make, model, mileage, etc.

2. Data Collection & Acquisition

- Acquire the dataset that contains car listings with features such as `make`, `model`, `year`, `mileage`, and `price`.

3. Data Exploration & Understanding (EDA)

- Perform exploratory data analysis:
  - Check for missing values and inspect the data types.
  - Visualize the distribution of features like car price, mileage, and year.
  - Analyze the correlations between features and the target variable (price).

4. Data Preprocessing

- Clean and preprocess the data:
  - Handle missing values (e.g., impute or remove).
  - Encode categorical variables like car brand and model.
  - Scale or normalize numerical features as needed.
  - Split the data into training and testing sets.

5. Feature Engineering

- Create new features or transform existing ones to enhance the model:
  - Calculate the car's age from the year of manufacture.
  - Transform mileage data or other variables for better model performance.

6. Model Selection & Training

- Choose machine learning algorithms to train the model:
  - Train baseline models such as Linear Regression.
  - Experiment with more complex models like Random Forest or Gradient Boosting.

7. Model Evaluation

- Evaluate the performance of the models using appropriate metrics:
  - Calculate MAE, RMSE, and R^2 to measure model accuracy.
  - Visualize the residuals and compare predicted vs actual values.

8. Model Optimization & Hyperparameter Tuning

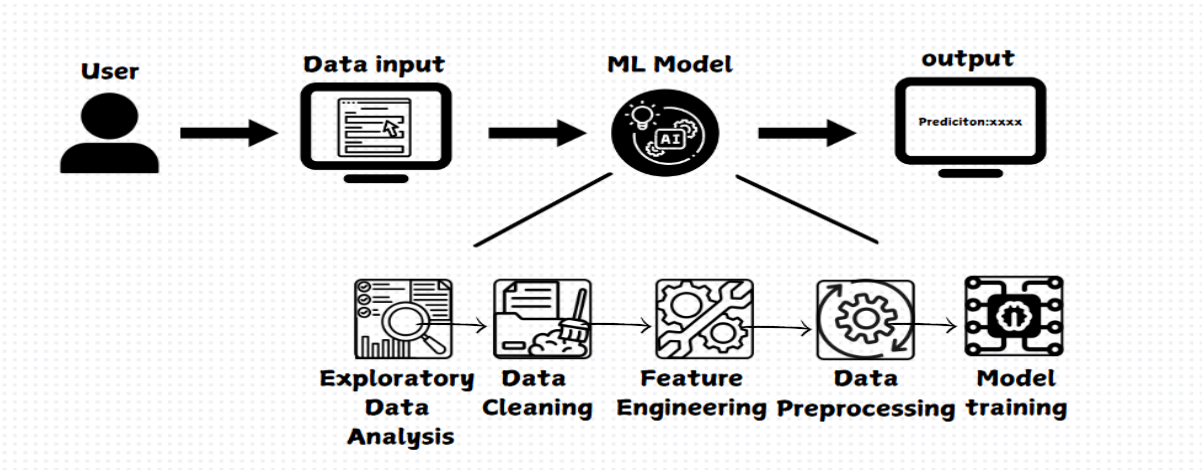
- Fine-tune the model by optimizing hyperparameters:
  - Use techniques like Grid Search or Random Search to find the best parameters.
  - Apply cross-validation to improve model generalization.

B. Deliverable of each steps or phase:

- **Problem Definition & Objective**: Predict car prices using features such as year, make, model, and mileage.
- **Data Collection & Acquisition**: Gather the dataset with car listings containing relevant features.
- **Data Exploration & Understanding (EDA)**: Analyze missing values, visualize feature distributions, and explore correlations.
- **Data Preprocessing**: Handle missing data, encode categorical features, scale numerical values, and split into training and testing sets.
- **Feature Engineering**: Create additional features, like car age, to improve model performance.
- **Model Selection & Training**: Train baseline and complex models (e.g., Linear Regression, Random Forest).
- **Model Evaluation**: Evaluate models using MAE, RMSE, and R^2.
- **Model Optimization & Hyperparameter Tuning**: Use Grid or Random Search and cross-validation to fine-tune the model for better generalization.

4. **Working Model**

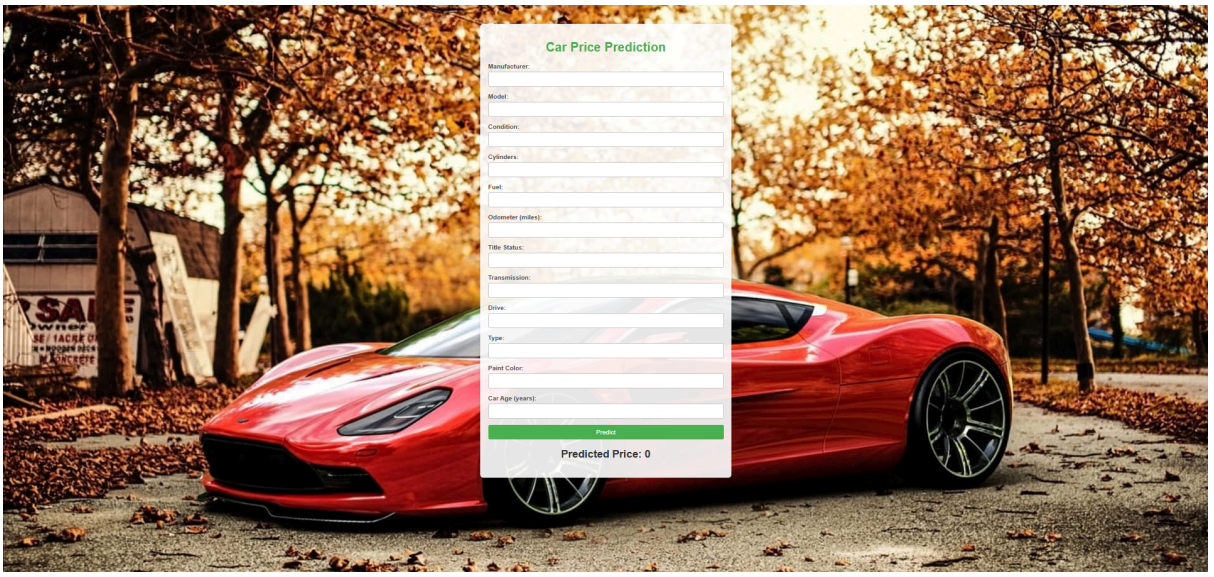
A. Technical Diagram



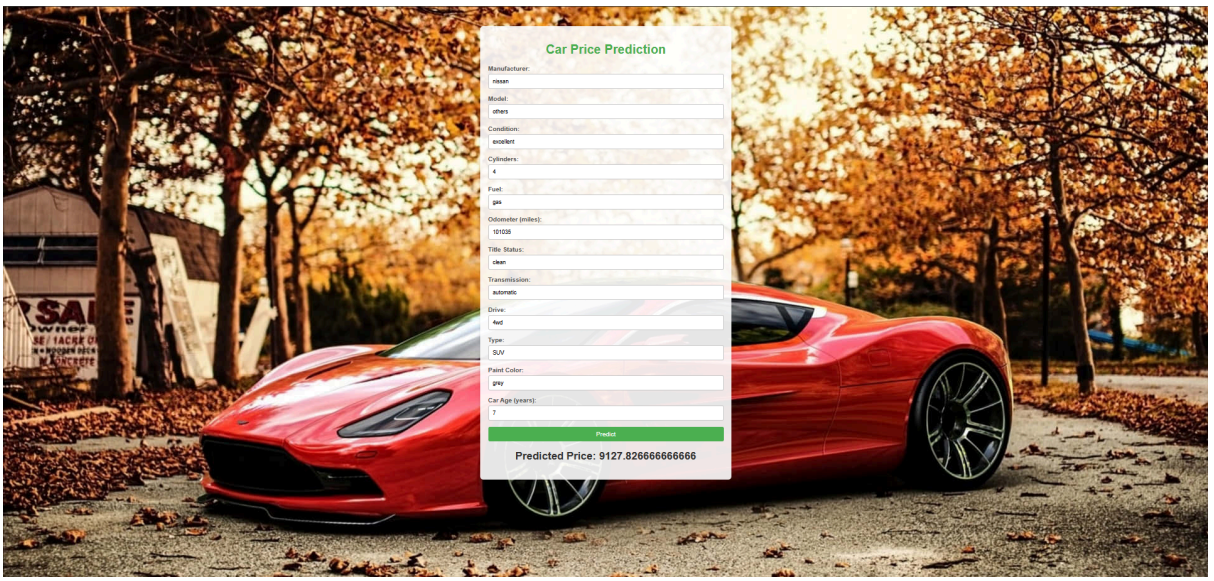
B. Working Module

This web-based car price prediction module allows users to input car details like manufacturer, model, condition, and mileage. Once the user fills in the form and clicks **Predict**, the backend uses a trained regression model (e.g., Random Forest) to estimate the car's price. The model was selected based on performance metrics like accuracy and error rates. The predicted price is displayed immediately on the interface. This application uses machine learning to help users estimate car values, making it useful for resale, buying, and selling decisions.

a. Before Prediction



b. After Prediction



C. Attained Deliverable

1. **Accurate Price Prediction**: Developed a functional car price prediction tool that accurately estimates vehicle prices based on user inputs, with a user-friendly interface.
2. **Regression Model Implementation**: Successfully implemented a machine learning regression model (such as Random Forest) for price prediction, using various features such as car make, model, mileage, and condition.

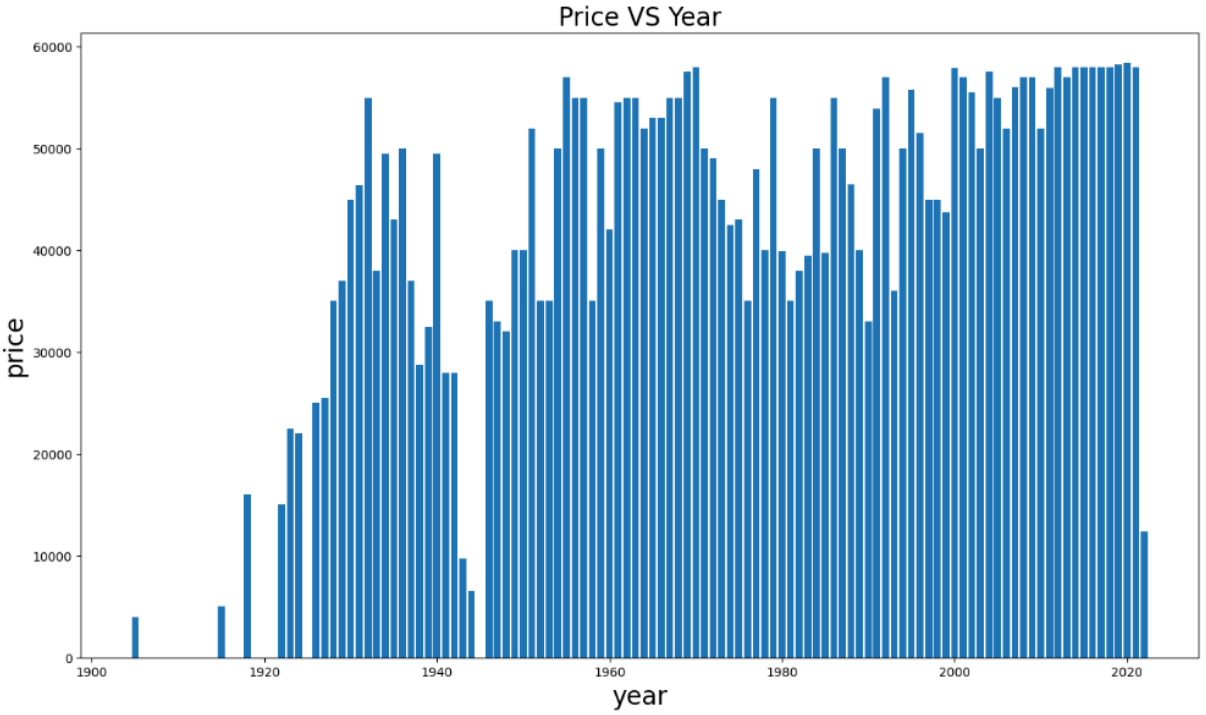
3. **Real-time Prediction**: Integrated real-time price prediction capabilities, allowing users to receive an immediate estimate after filling in the car details.
4. **Improved User Experience**: Designed a responsive and intuitive UI that makes it easy for users to interact with the system and get results quickly.
5. **Deployment**: Successfully deployed the web application, making it accessible to users for car price estimation.

5. Results

A. Outcome Graphs

a. Vehicle Price Trends Across Manufacturing Years

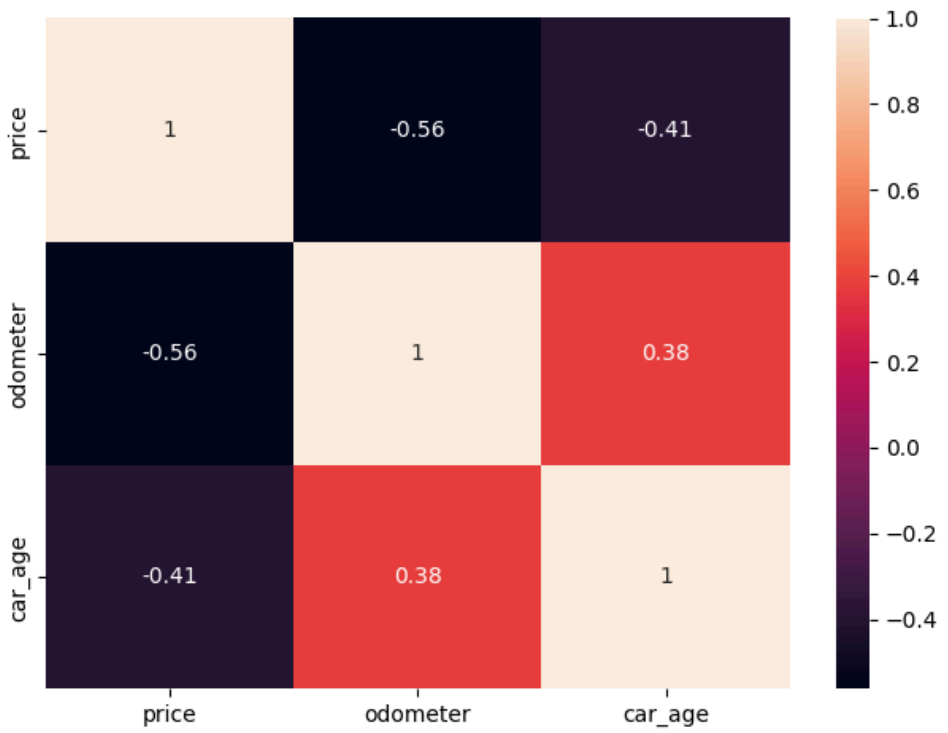
This bar chart shows the relationship between the manufacturing year of vehicles and their prices. Older cars (pre-1940s) have relatively low prices, while prices increase for cars manufactured after 1980. Vehicles from recent years (2000 and later) maintain consistently higher prices. The graph illustrates a positive trend in vehicle prices over time, possibly reflecting advancements in technology, features, or demand for modern vehicles.



b. Correlation Analysis of Key Features

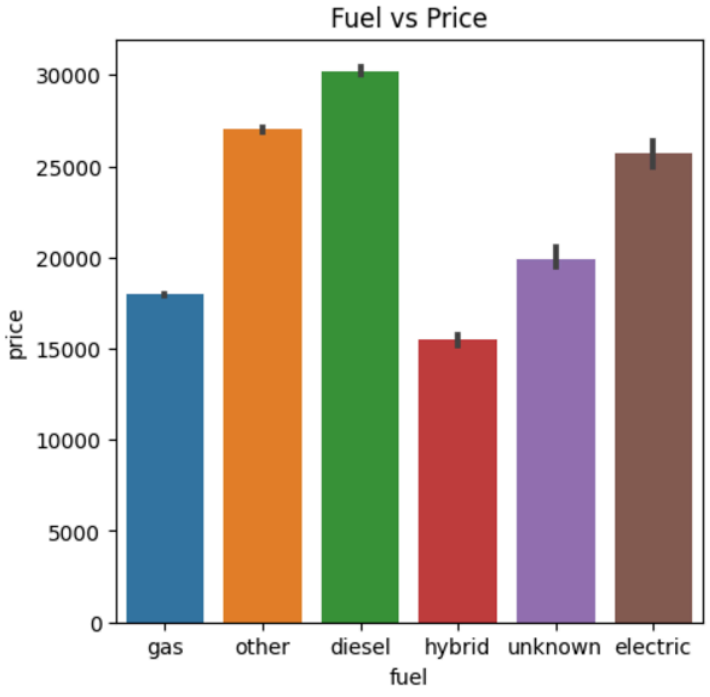
This heatmap visualizes the correlation coefficients between variables: `price`, `odometer`, and `car_age`.

- **Key observations:**
  - `Price` has a moderate negative correlation with `Odometer` (-0.56), indicating that cars with higher mileage tend to have lower prices.
  - `Price` also has a weaker negative correlation with `Car Age` (-0.41), suggesting that older cars are generally less expensive.
  - `Odometer` and `Car Age` show a moderate positive correlation (0.38), indicating that older cars often have higher mileage.



c. Impact of Fuel Type on Vehicle Prices

This bar plot compares the average price of vehicles across different fuel types. Diesel vehicles have the highest average price, followed by electric cars, while gas-powered vehicles have the lowest prices. The differences in prices could be due to the varying efficiency, environmental impact, and adoption rates of different fuel types.



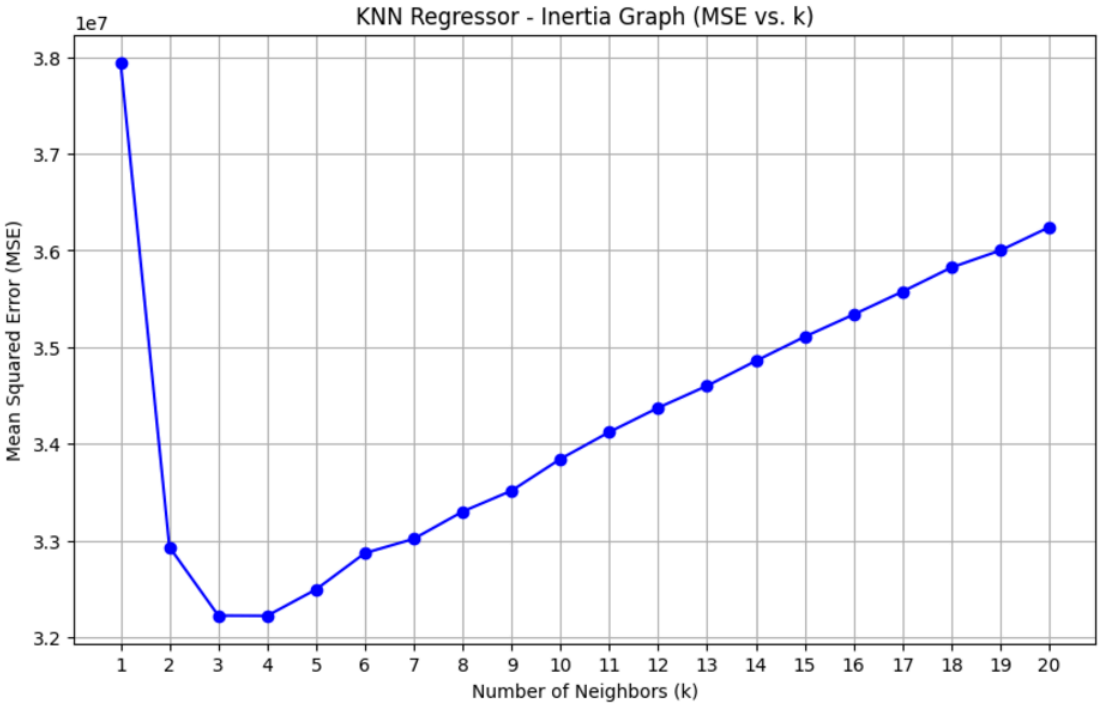
B. Comparative Studies

a. **Evaluation of Different Regression Algorithms Across Metrics**

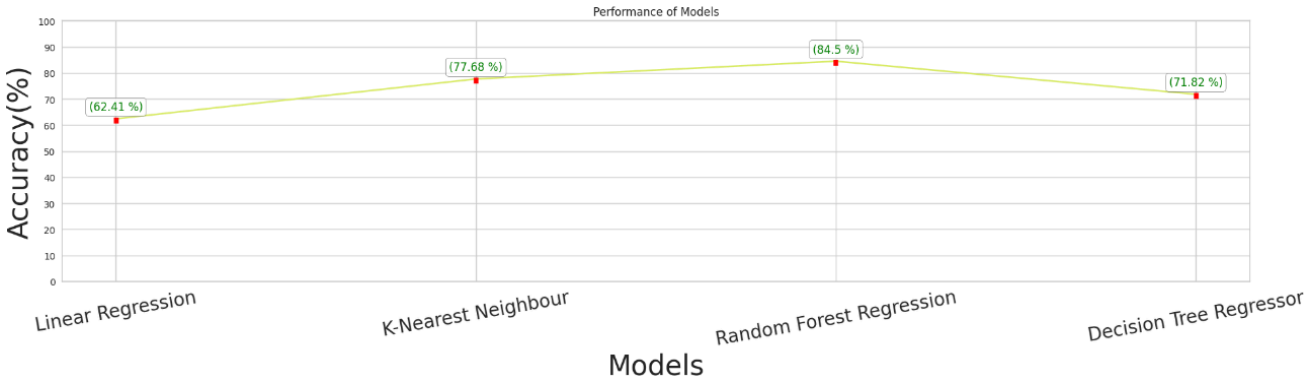
	Linear Regression	**KNN Regressor**	**Random Forest Regressor**	**Decision Trees Regressor**
R2	0.6264	0.7781	0.8460	0.7232
Accuracy	62.6378	77.8073	84.5975	72.3157
MSE	54250663.20	32224245.03	22364692.71	40198080.26
MAE	5560.76	3244.18	2638.73	3223.86
Root MSE	7365.50	5676.64	4729.13	6340.20

b. **Impact of K on Mean Squared Error (MSE) in KNN Regression**

This line graph demonstrates how the mean squared error (MSE) of the K-Nearest Neighbors (KNN) regressor changes with the number of neighbors (k). The lowest error is observed at k=3, suggesting that this is the optimal value for k in this dataset. Beyond this, the MSE gradually increases, indicating diminishing accuracy with higher k values.



c. Comparison of Model Accuracy Across Regression Algorithms



6. Conclusion

A. Justification of Objectives:

The primary objective of developing a machine learning model for predicting car prices was achieved. The dataset provided a diverse set of features that allowed for comprehensive analysis and accurate predictions. The implementation of regression models, feature selection, and data preprocessing directly aligned with the project's goal of creating a practical and efficient pricing tool.

B. Future Scope:

Model Deployment: Develop a user-friendly web or mobile application for real-time car price predictions.
Multi-Regional Analysis: Extend the model to accommodate data from different regions, addressing variations in market demand and supply.
Dynamic Pricing Models: Include real-time market trends and seasonal variations to adapt the predictions dynamically.
Explainable AI: Implement techniques to make the predictions more interpretable, ensuring trust and transparency for end-users.

7. References

https://proceedings.mlr.press/v7/niculescu09/niculescu09.pdf

Kaggle: https://www.kaggle.com/

Scikit-learn Documentation: https://scikit-learn.org/

Python Documentation for Data Analysis Libraries: Pandas, NumPy, Matplotlib, and Seaborn, https://docs.python.org/

Submitted by:

Avneesh Jarangal
Roll No. 102217029
Prerit Bhagat
Roll No.102217030

Submitted to:

Dr. Nitin Arora
Assistant Prof.
Dept. of CSE
TIET, Patiala