Car Price Prediction

Linear Regression

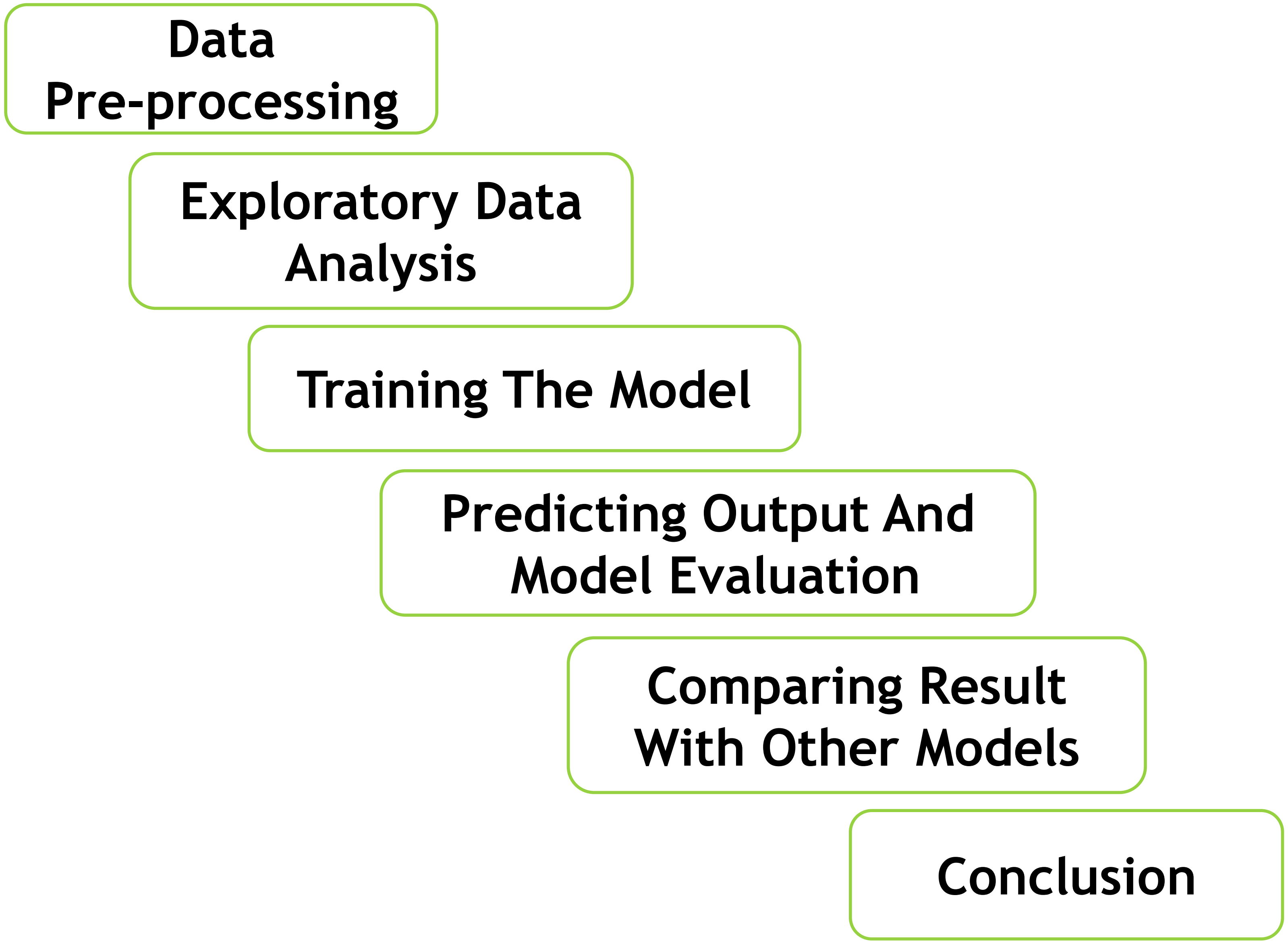
Group 2

R. Srushtitha

D.Srujana

G.Shylaja

Machine Learning Timeline Analysis



Data Pre-processing

Exploratory Data Analysis

Training The Model

Predicting Output And Model Evaluation

Comparing Result With Other Models

Conclusion

About the Data:-

Technical analysis looks at historical prices, economic growth rates, and other related factors, formulating an approximate price. Then, to get a more accurate picture of the market, the process turns to fundamental analysis.

Objective:-

The main aim of this project is to predict the price of used cars using the various Machine Learning (ML) models. This can enable the customers to make decisions based on different factors like Brand, model of car, type of fuel, Budget, mileage etc..

Path:-

Implementing multiple machine learning models to fit best model for the dataset.

Data And Data Quality Check

Data Introduction:

The data consists of 18 columns and 19237 observations

Columns:

‘ID’,’Price’,’Levy’,’Manufacture’,’Model’,’Prod Year’, ‘Category’,’Leather interiror’,’Fuel Type’, Engine Volume’,’Mileage’,’Cylinders’,’Gear box type’,’Drive wheels’,’Doors’,’Wheels’,’Color’,’Airbags’.

Missing Values:

The data is complete, with no missing values.

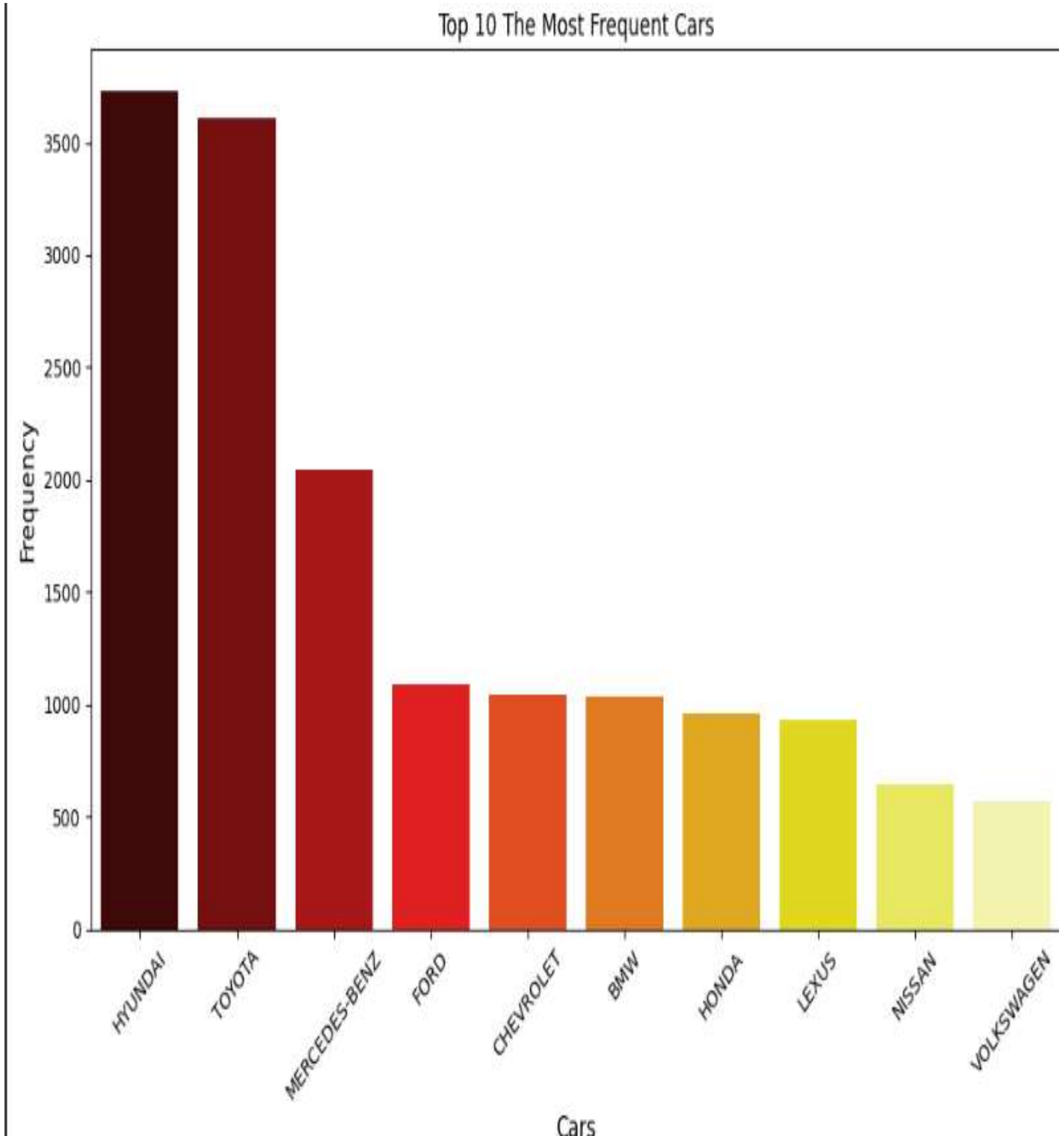
Duplicate Values:

There are 313 duplicate values removed out of 19237 observations.

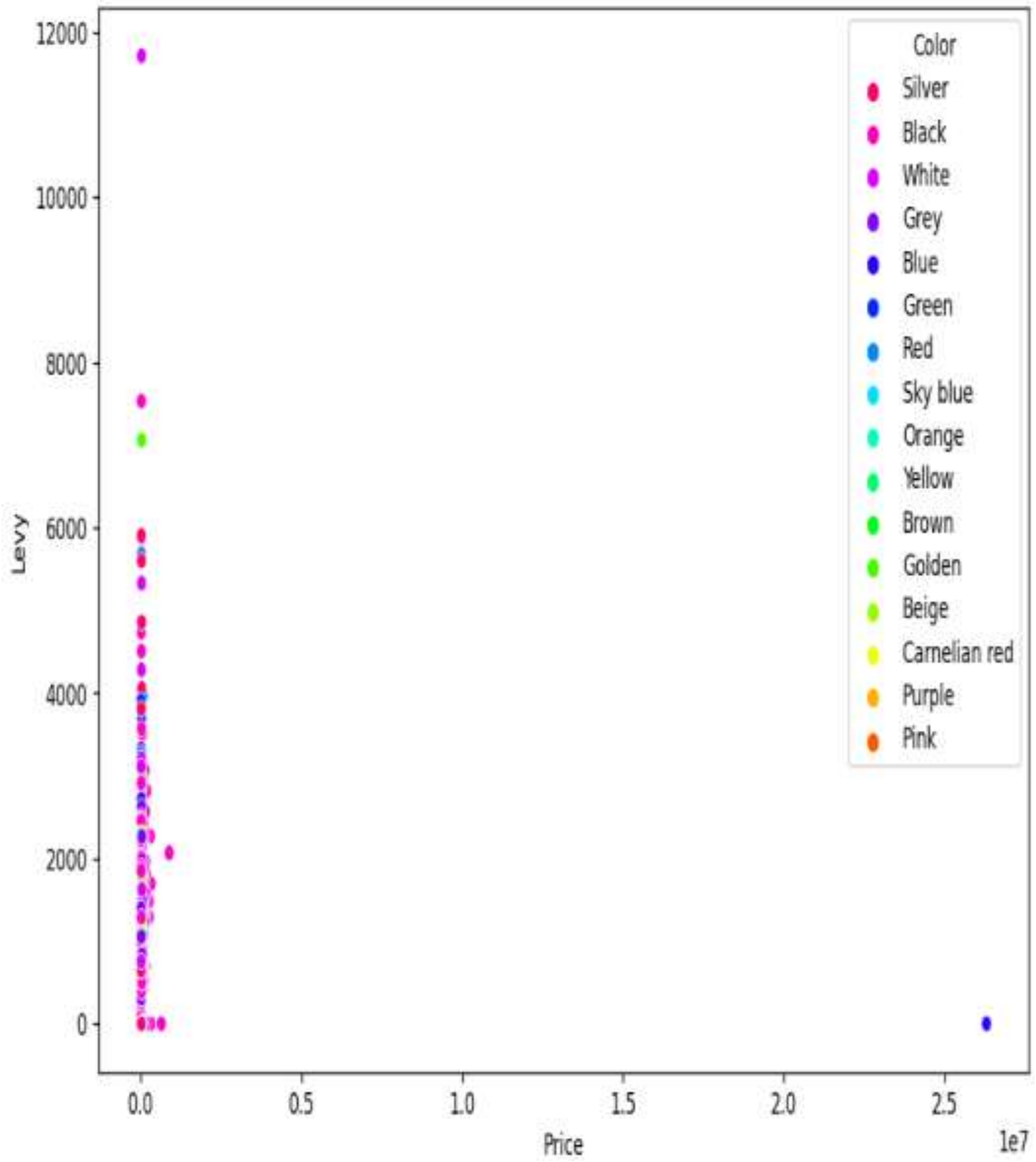
Dropped Columns:

- ID
- Doors

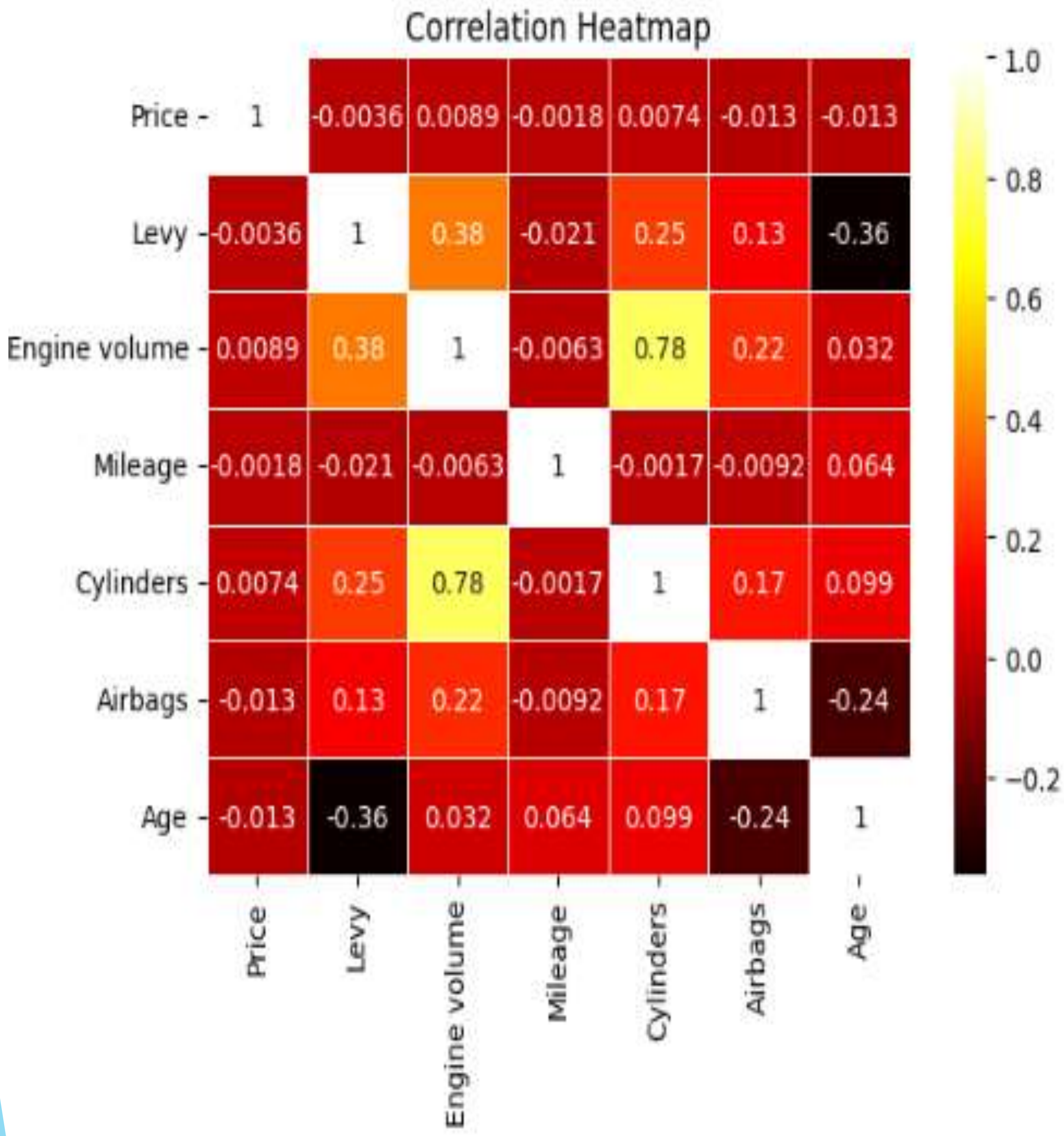
EXPLORATORY DATA ANALYSIS(EDA)



- Count plot using Seaborn to visualize the frequency of cars from the 'Manufacturer' column.
- The bars represent the number of occurrences of each manufacturer, and they are ordered based on their frequency.
- This type of plot is useful for understanding the distribution of categorical data and identifying the most common categories within a specific column.



- A scatter plot using Seaborn to visualize the relationship between the 'Price' and 'Levy' columns in your dataset.
- The points on the scatter plot are colored based on the 'Color' column. It allows to visually identify if there are specific color trends or patterns in the distribution of 'Price' and 'Levy'.

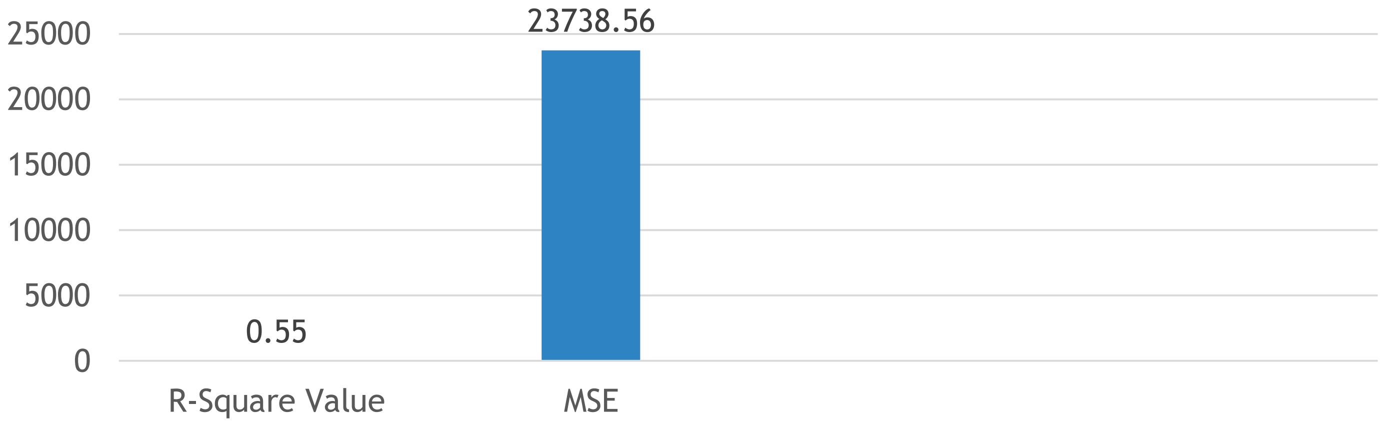


- The numerical values within each square of the heatmap represent the correlation coefficient between pairs of variables. Values closer to 1 indicate a strong positive correlation, values closer to -1 indicate a strong negative correlation, and values near 0 indicate a weak correlation.

Algorithms:

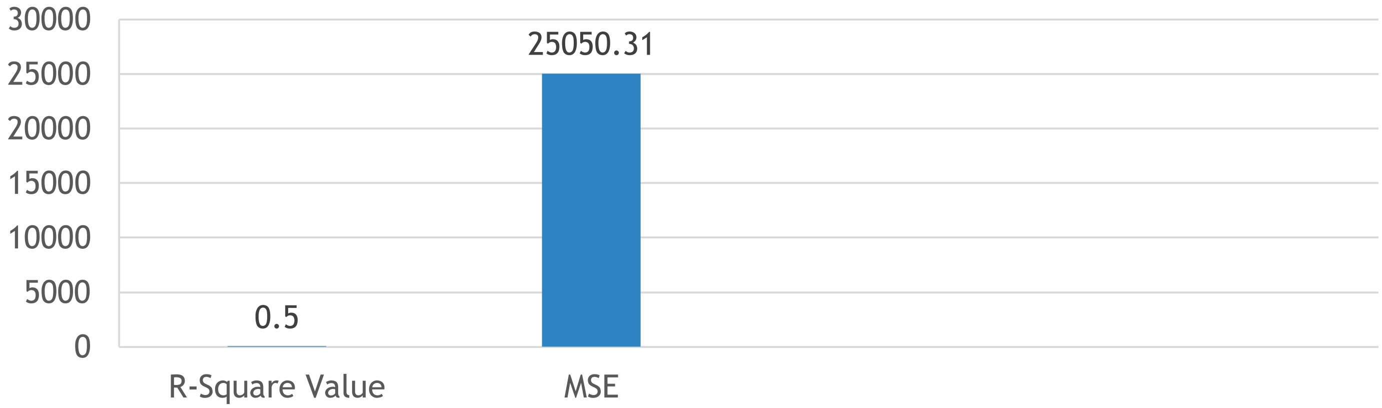
Linear Regression:

Linear regression goal is to predict a continuous numerical output based on one or more input features. Linear regression models establish a linear relationship between the input features and the output variable. It's a simple yet effective algorithm for understanding the relationship between variables.



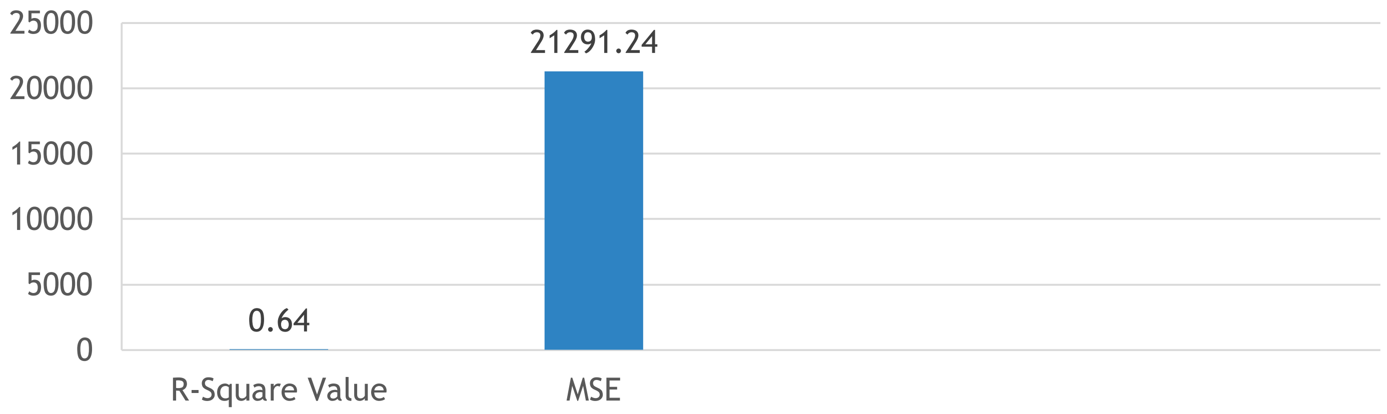
Decision Tree Regression:

Decision tree is a hierarchical structure that make decisions by splitting data based on variables. It can handle both regression and classification tasks. It is easy to interpret and visualize.



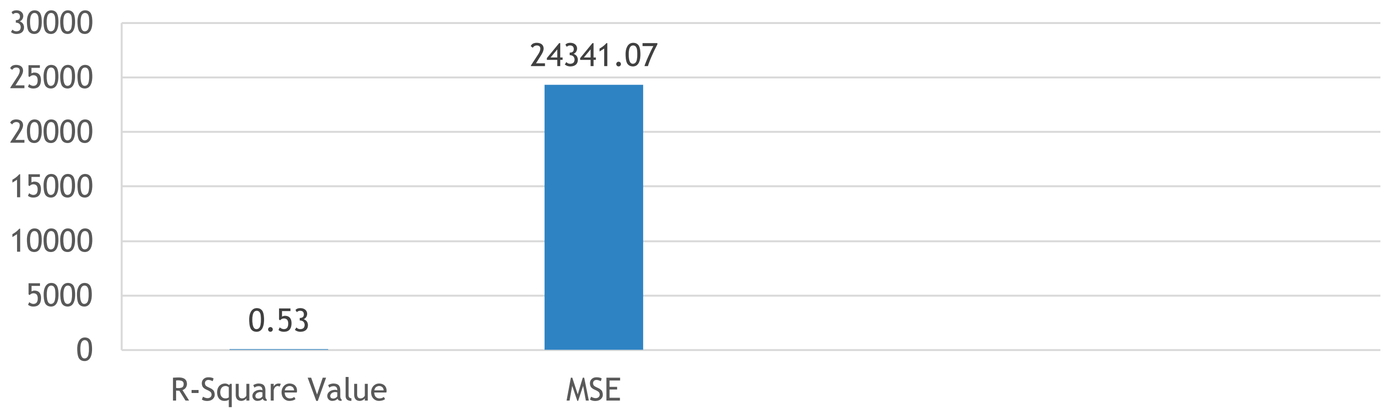
Random Forest Regression:

- Random Forest Regressor is used for solving regression problems. It is an extension of the decision tree algorithm that combines multiple decision trees to make more accurate predictions. Random forests are known for their high predictive accuracy and ability to handle complex datasets.



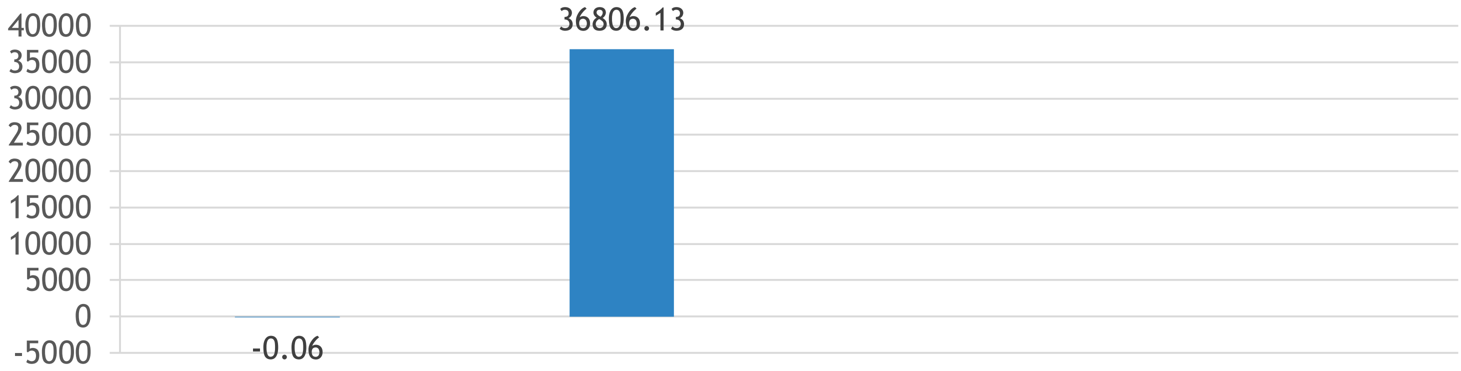
K - Nearest Neighbor Regression:

KNN Regression is a useful algorithm when you need a simple and interpretable model for regression tasks. It's suitable for situations where you want to capture non-linear relationships. It is used to predict continuous numerical values based on the values of neighboring data points.



Support Vector Machine:

SVM works by finding a hyperplane in a high-dimensional space that best separates data into different classes. It aims to maximize the margin while minimizing classification errors. It is used for regression and classification tasks.



Comparision of Implemented Models :

Algorithms	Mean Square Error
Linear Regression	23738.56
Decision Tree Classifier	25050.31
Random Forest Classifier	21291.24
Gradient Boosting Regression	24341.07
Support Vector Machine	36806.13

Summary And Recommendations

- For Car Price Prediction dataset , we performed five types of machine learning algorithms: Linear Regression, Decision Tree Regression, Random Forest Regression, SVM and K- Nearest Neighbours Regression.

- The visualization of data performed in exploratory data analysis showed clearly which of the features are more and less correlated with the target variable which is Price.

- From all the analysis and implementations of the algorithms, we found the best result in Random Forest algorithm with Mean Square Error (MSE) of **21291.241317** which is the least error value among all the other errors.

- So, here we can conclude that random Forest Algorithm can be considered the best model for the given Student Grade dataset.

Future Insights :

- Analyze the impact of transmission types (gearbox) and drive wheel configurations on car prices. For example, automatic transmissions or specific drive wheel types may influence prices differently.
- Understand the relationship between mileage and car prices. Typically, lower mileage is associated with higher prices, but there may be nuances based on other factors.
- Investigate how the number of doors and the presence of airbags contribute to the pricing of used cars. Safety features may be significant determinants.
- Explore whether the color of the car plays a role in pricing. Certain colors may be more popular and affect resale values.
- Investigate customer preferences by analyzing the popularity of specific features or combinations of features. This can help in understanding market demand and influencing pricing strategies.

Presented by:

- R.Srushtitha
- D.Srujana
- G.Shylaja

Click here to Refer Code

https://colab.research.google.com/drive/1yDtt2ul7aqLxafdD1C1ewglfru30yDT3?usp=sharing