FUEL EFFICIENCY PREDICTION

M. Aditya Vamsi, B. Raja Rishita, M. Amin Qurishi, N. Siva Sandeep, K. Raghu Ram

Department of Computer Science Engineering, GITAM, deemed to be University, Visakhapatnam, India-530046

Abstract: - Car makers are continually streamlining their cycles to increment eco-friendliness, in light of the fact that the vehicle business has been developing for more than two centuries, and fuel costs are persistently expanding. Clients are turning out to be more specific about highlights, so car creators are continually changing their cycles to increment eco-friendliness. In any case, imagine a scenario where there was a precise assessor for a vehicle's MPG (Miles per Gallon) or the fuel utilization indicator in light of a few known details. By having a more positive vehicle that is likewise more proficient, you could beat rivals on the lookout, increment interest for the item and put more creation into the market. Utilizing Machine Learning, we are currently planning the expectation models and lessening the Error values for cars that have been made in the beyond couple of year.

Here we will use the datasets available to machine learning practitioners to create a model to predict fuel efficiency of various kinds of vehicles across various periods. As part of the model, we will include descriptions of many different cars from different periods. These descriptions will include things such as cylinders, displacement, horsepower, and weight. The ML method is appropriate for this type of analysis since it can be applied to learn patterns in data and construct models from them. Besides that, deep learning concepts will be implemented to create other models. Based on the analysis, it will be seen whether which model will produce less error with better efficiency.

Keywords: Machine Learning, Fuel Efficiency Prediction, Random Forest, Decision tree, KNN, Linear Regression.

1 Introduction

Having a decent comprehension of what influences fuel utilization, and afterward having the option to foresee it, is vital to upgrading eco-friendliness. In the transportation business, the Miles per Gallon, or MPG, is utilized to work out a vehicle's proficiency as a component of the energy it consumes. MPG fluctuates by beginning. To check the MPG content in vehicles, we have made chart models. The diagram models relate with this MPG in the vehicles in view of chambers, removal, drive, and weight. Motors are estimated by relocation, or their volume, which is normally communicated in liters or cubic centimeters. The beginning is a discrete number going from 1 to 3. In light of this dataset, we accepted that 1 addresses a vehicle from America, 2 addresses a vehicle from Europe, and 3 addresses a vehicle from Asia or different spots. A few of the qualities in this dataset may be erroneous, so we will address those qualities during the pre-handling of the information. Demonstrating fuel utilization in expressways is more straightforward, since outer factors, for example, traffic and street conditions don't fundamentally impact fuel utilization. Besides, by having the option to anticipate the fuel utilization the proprietors may likewise distinguish potential fuel extortion if any.

Makers, controllers, and clients are totally intrigued by vehicle fuel utilization models. They are expected during all phases of the vehicle's life cycle. The objective of this work is to display average fuel utilization for weighty vehicles all through activity and upkeep. As a general rule, there are three sorts of techniques for creating fuel utilization models:

- Material science-based models, these are the models that are framed from an exhaustive handle of the actual framework. These models utilize exhaustive numerical conditions to depict the elements of the vehicle's parts at each time step.
- Factual models, which are additionally information driven and lay out a planning between the likelihood circulation of a chose set of indicators and the objective result.
- AI models, which are information driven and address a theoretical planning from an info space comprising of a chose set of indicators to a result space that addresses the objective result, for this situation normal fuel utilization.

Compromises between the above techniques not entirely settled by cost and precision, contingent upon the necessities of the planned application.

Without exact information on the vehicle's actual properties and estimations, the technique should apply and adjust to a wide scope of vehicle advancements (counting future ones) and arrangements for every vehicle. While gauging the required exactness versus the expense of creating and adjusting an individualized model for every vehicle, AI arises as the strategy for decision. There have been a few past models created for both immediate and normal fuel use. Since they can address the elements of the framework's movement at different time steps, physical science-based models are the most appropriate for assessing momentary fuel use. Since figuring out designs in genuine opportunity information is troublesome, AI models can't foresee prompt fuel utilization with an elevated degree of exactness. These calculations, then again, are prepared to do precisely recognizing and learning patterns in normal fuel use. Recently proposed AI strategies for normal fuel utilization utilize a bunch of indicators accumulated after some time to gauge fuel utilization in gallons per mile or liters per kilometer. While our recommended method is as yet centered around normal fuel utilization, it changes from prior models in that the indicators' feedback space is quantized with respect to a proper distance instead of a decent time span. All indicators in the proposed models are collected regarding a proper window that addresses the vehicle's distance voyaged, bringing about a superior planning from the information space to the model's result space. Past AI models, then again, needed to not just gain proficiency with the examples in the info information, yet additionally convert from the info area's time sensitive scale to the result space's distance-based scale (i.e., normal fuel utilization). Involving similar scale for the model's feedback and result regions enjoys different benefits.

2 Related Work

In [1], creators assessed the prescient capacity of three AI expectation models to anticipate the fuel utilization of a significant distance public transport. A few fundamental qualities, including as burden, motor RPM, and traffic, are excluded from the chose dataset, regardless of the way that they straightforwardly impact fuel utilization. Indeed, even without such basic elements, they showed that the RF model could all the more precisely foresee fuel utilization while gathering information patterns. An illustration of such a model is the discovery of fuel extortion by contrasting the real utilization of the vehicle with the anticipated worth in view of different boundaries like distance, area, rise, speed, and day of the week. They wanted to fuse more factors affecting fuel use, like traffic, climate, and transport load, as a component of future work to further develop the forecast capacities much more. They're additionally dealing with a module that will walk you through the most common way of reengineering strategies to decrease fuel use through better armada planning and driving propensities. The informational index that they have used to assemble their model is the data about a particular significant distance public transport in Sri Lanka. Around 4:00 p.m., the transport leaves from Depot and goes to Colombo (i.e., business capital). The transport then, at that point, leaves Colombo at 7:00 p.m. furthermore, drives along the A2, A4, and AB10 interstates, showing up at the objective around 7:00 a.m. the following day.

Altogether, the transport covers 365 kilometers in a single heading. The return trip follows similar way and happens between 4:00 p.m. what's more, 7:00 a.m. the following morning. A rugged locale makes up about 33% of the course. A GPS-based global positioning framework and a capacitive, high-accuracy fuel sensor are introduced on the transport. Over a 3G association, gathered information is sent in close to continuous to a cloud server. Between May 13 and August 31, 2015, the dataset contains outbound and inbound ventures. In light of the characteristics and the information present in the dataset, they have carried out different AI calculations and systems to anticipate the fuel utilization of those armada vehicles in that specific course confronting the different latitudinal and longitudinal variables.

They have utilized Random Forest, Gradient Boosting and Artificial Neural Network for executions, in the wake of testing the different mistakes including Bias, MAE and RMSE

In [2], creators fostered a model that could precisely appraise a vehicle's MPG given a few data about the vehicle and concluded the code by getting an altered RMSE score of 1.97 from a more noteworthy worth of 3.26 as the underlying RMSE score. This model could be prepared with fresher vehicle information and used to anticipate even the R2 score, as we had the option to code and found a consistent expansion in the worth from 0.82 to 0.91, demonstrating that the model is substantially more solid being used and furthermore helpful for our rival's future MPG evaluations for impending vehicles, permitting.

Organizations to possibly zero in assets as of now spent on making more productive, more well-known vehicles that eclipses contenders. Despite the fact that their model might be off base sometimes, they examined how their dataset may contain mistaken values for MPG, yet by and large, the expectations are more exact than the qualities in the dataset. The information gathered from fresher vehicles is significantly more solid, so their model will actually want to work all the more proficiently with various, more precise datasets.

They have used the information from the earlier years vehicles as a piece of their information base to anticipate the eco-friendliness or the miles per gallon utilizing the best AI model which they found as Linear Regression in view of testing different calculations.

The calculations have been carried out and displayed so that the RMSE blunder is focused on to be just about as productive as conceivable i.e., in the middle of 1 and 2 for the better precision of the forecasts to be made on the vehicle’s information.

3 METHODOLOGIES

In this project, we will be implementing Linear regression, Random Forest, Decision tree, Neural Network etc. algorithms. At first, we will perform the data preprocessing on our data along with its data cleaning.

All the visualizations related to the data can be seen here, various graphs and plots will be used. Cleaning of the data will include filling the null units in the dataset if any or rectifying any other defects.

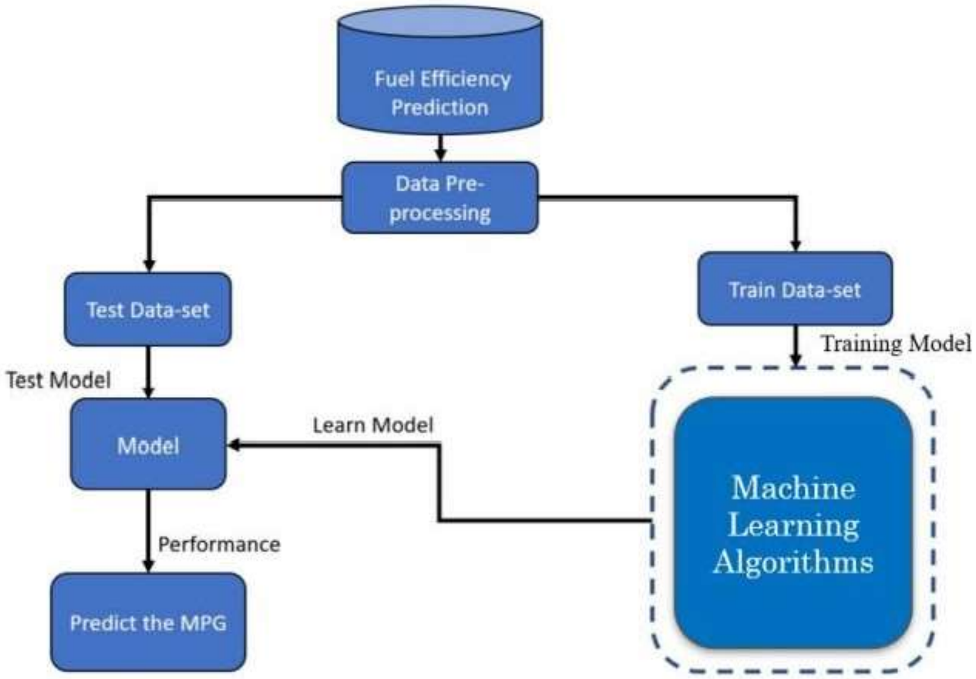
Then, we will be constructing different models with the various algorithms and start performing the training process on the data. Each different model will be constructed using the different algorithms. The predictions values will be dependent on the proper training of data with the ones from the dataset. After all this, testing will be done on the datasets with the datasets for MPG or the fuel efficiency predicted value.

The RMSE values of all the models will be compared in order to find the better suited and efficient model. Once the models are finalized, the deployment of it can be done.

Deployment process includes the creation of a web page with the suitable models and deploys them in that web page. The user either from the automobile industry side or the customer will visit the page or give the new inputs as per their models to be checked.

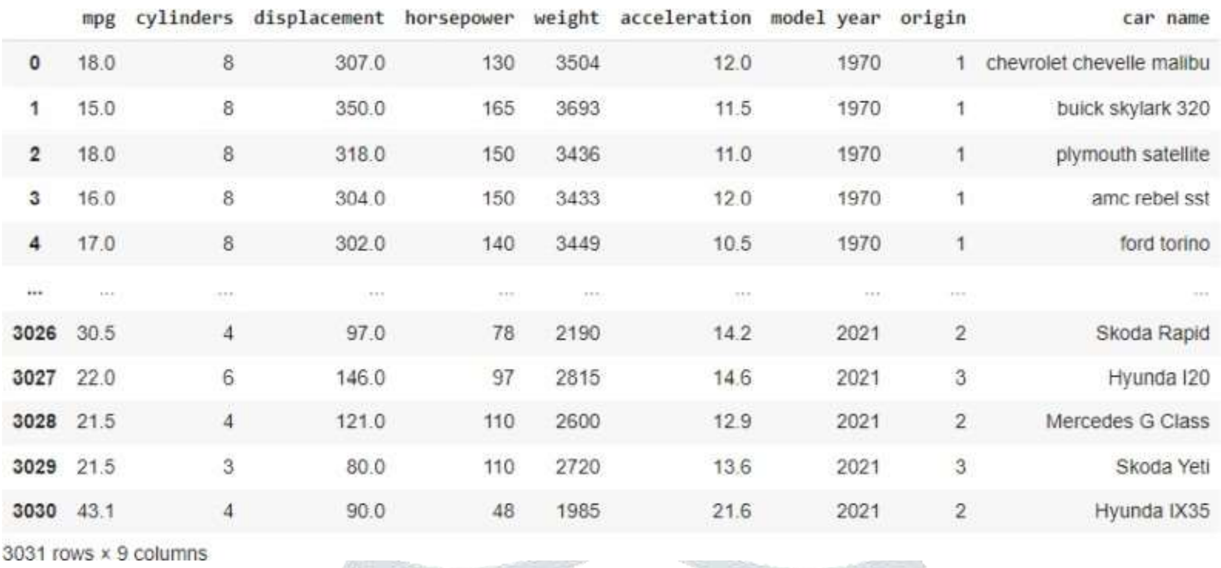
The given input will be fed to the model and predictions of the fuel efficiency will be calculated accordingly. Finally, the predicted output which will be the MPG or fuel efficiency will be displayed on to the screen.

FLOWCHART



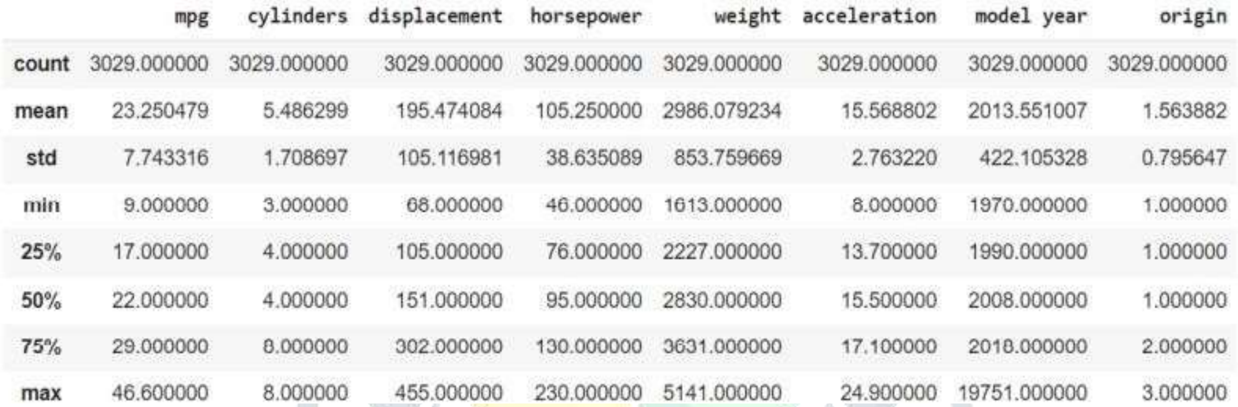
4 DATA PREPROCESSING

As a data set (also known as a dataset) is a collection of information. A data set correlates to one or more database tables in the case of tabular data, where each column of a table represents a specific variable and each row corresponds to a specific record of the data set in the requirements. The dataset that we have used in our project consists of 3032 rows and 8 columns. This dataset is taken from Kaggle and combined it with 100,000 used cars dataset. We have chosen our dataset with respect to the requirement of the project. The following consists of sample data from our dataset.



	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
0	18.0	8	307.0	130	3504	12.0	1970	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	1970	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	1970	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	1970	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	1970	1	ford torino
...	...	...	...	...	...	...	...	...	...
3026	30.5	4	97.0	78	2190	14.2	2021	2	Skoda Rapid
3027	22.0	6	146.0	97	2815	14.6	2021	3	Hyunda I20
3028	21.5	4	121.0	110	2600	12.9	2021	2	Mercedes G Class
3029	21.5	3	80.0	110	2720	13.6	2021	3	Skoda Yeti
3030	43.1	4	90.0	48	1985	21.6	2021	2	Hyunda IX35

3031 rows × 9 columns

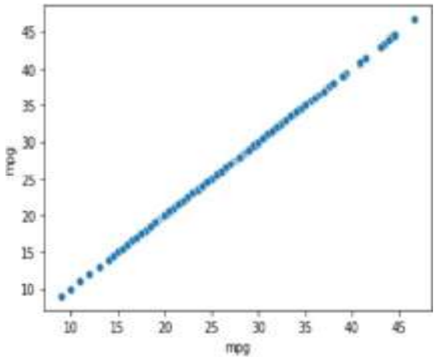


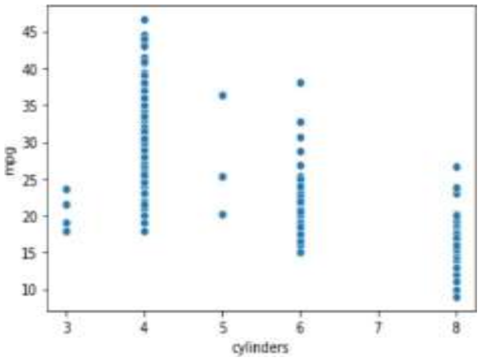
	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
count	3029.000000	3029.000000	3029.000000	3029.000000	3029.000000	3029.000000	3029.000000	3029.000000
mean	23.250479	5.486299	195.474084	105.250000	2986.079234	15.568802	2013.551007	1.563882
std	7.743316	1.708697	105.116981	38.635089	853.759669	2.763220	422.105328	0.795647
min	9.000000	3.000000	68.000000	46.000000	1613.000000	8.000000	1970.000000	1.000000
25%	17.000000	4.000000	105.000000	76.000000	2227.000000	13.700000	1990.000000	1.000000
50%	22.000000	4.000000	151.000000	95.000000	2830.000000	15.500000	2008.000000	1.000000
75%	29.000000	8.000000	302.000000	130.000000	3631.000000	17.100000	2018.000000	2.000000
max	46.600000	8.000000	455.000000	230.000000	5141.000000	24.900000	19751.000000	3.000000

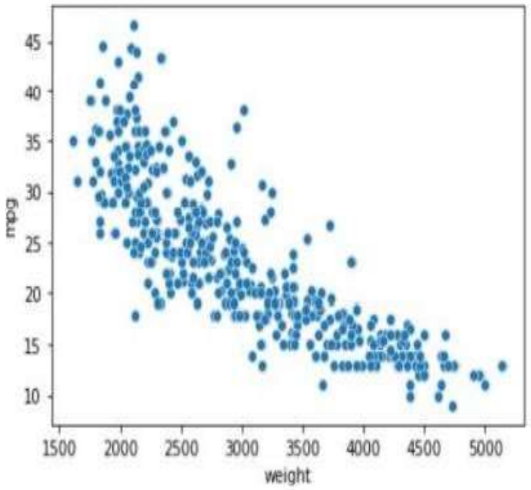
Data has been preprocessed accordingly and made the information and representations in the format requires and easy for making the predictions more effective and precise.

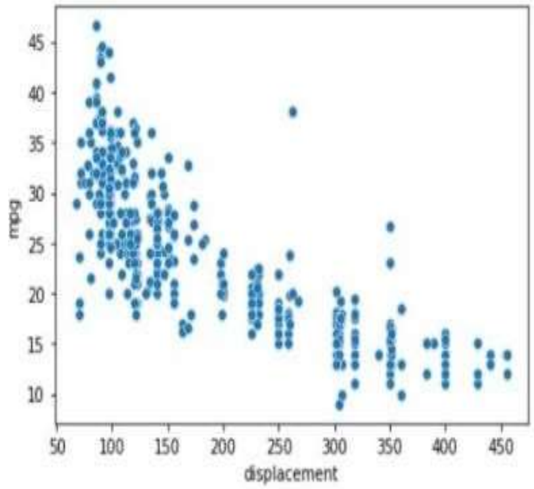
DATA VISUALIZATION:

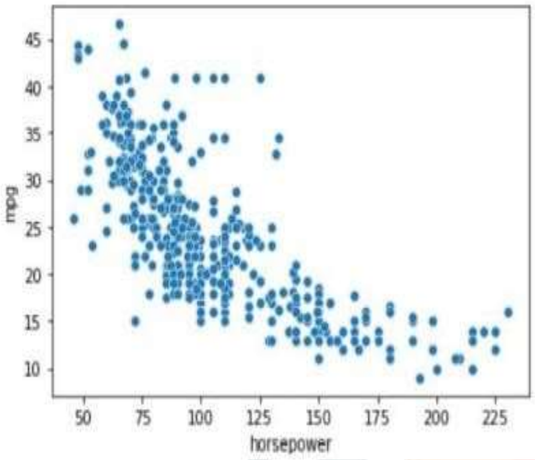
Various attributes in the dataset have been visualized individually to use its representations for the analysis and predictions to be made on the data in the machine learning algorithms.

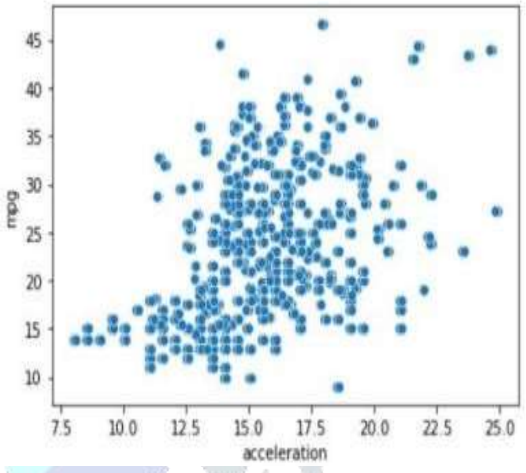


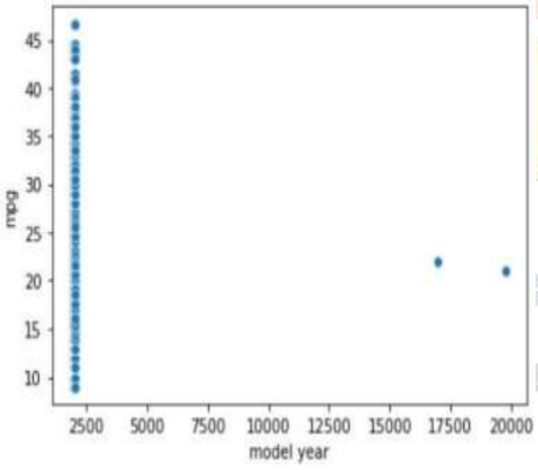


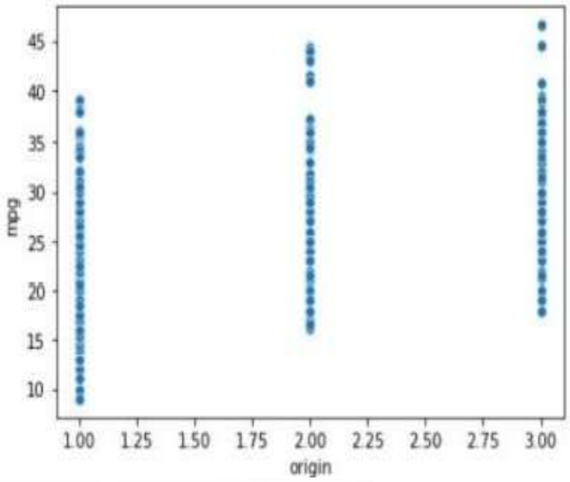

mpg
10
15
20
25
30
35
40
45
1500
2000
2500
3000
3500
4000
4500
5000
weight

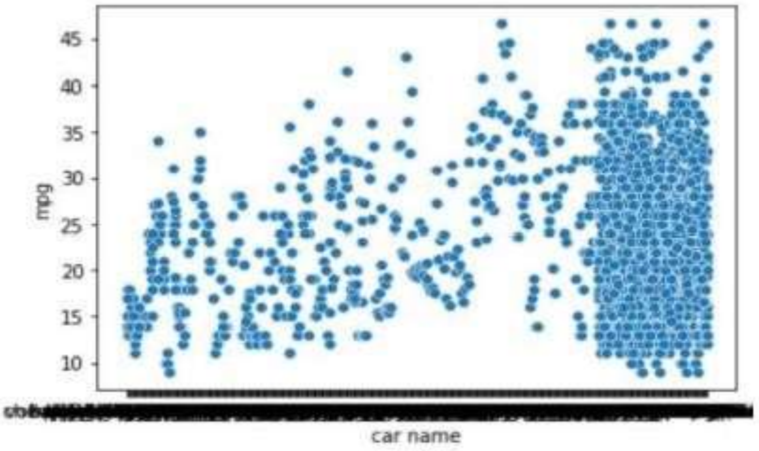

mpg
10
15
20
25
30
35
40
45
50
100
150
200
250
300
350
400
450
displacement

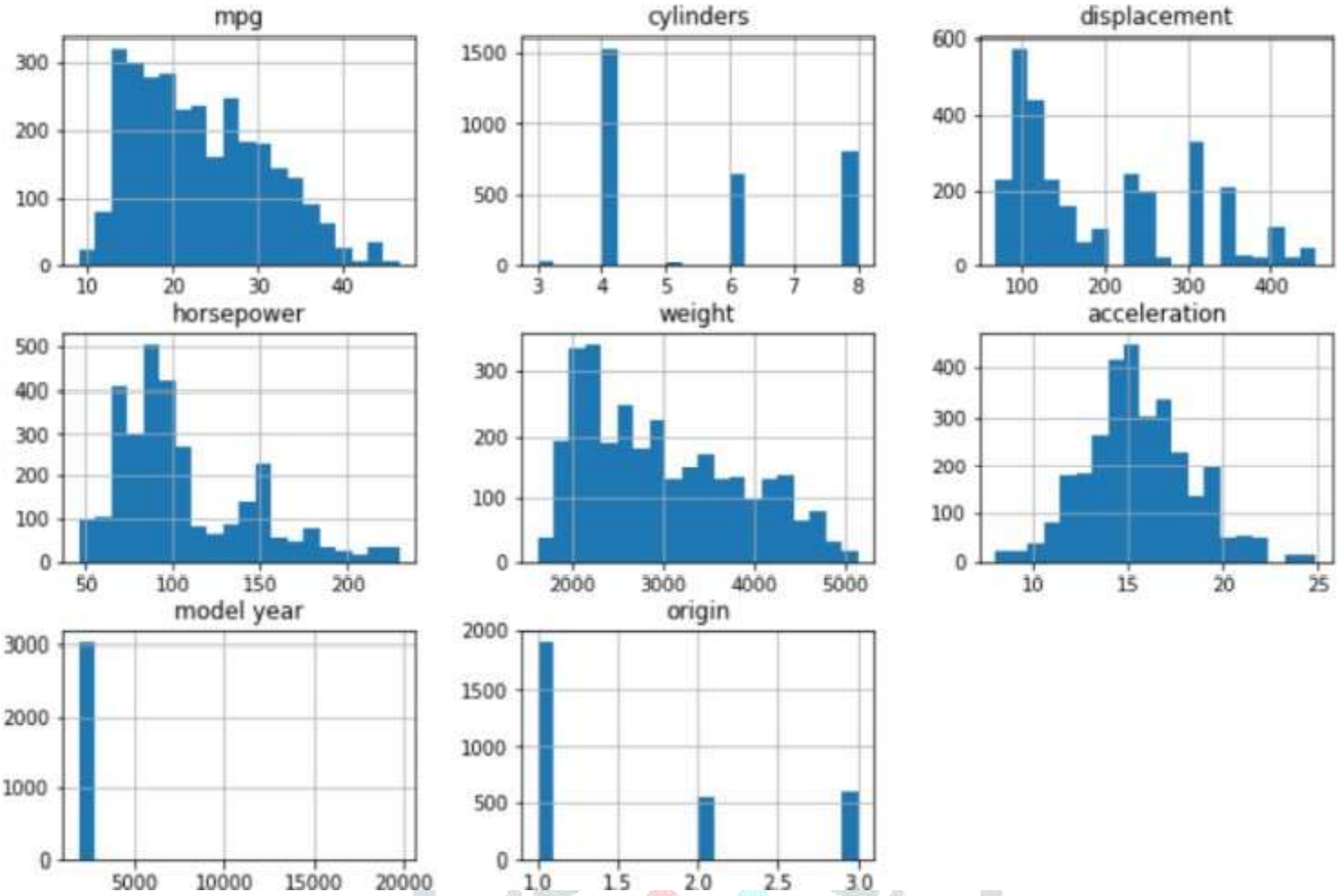

mpg
10
15
20
25
30
35
40
45
50
75
100
125
150
175
200
225
horsepower


mpg
10
15
20
25
30
35
40
45
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
acceleration


mpg
10
15
20
25
30
35
40
45
2500
5000
7500
10000
12500
15000
17500
20000
model year

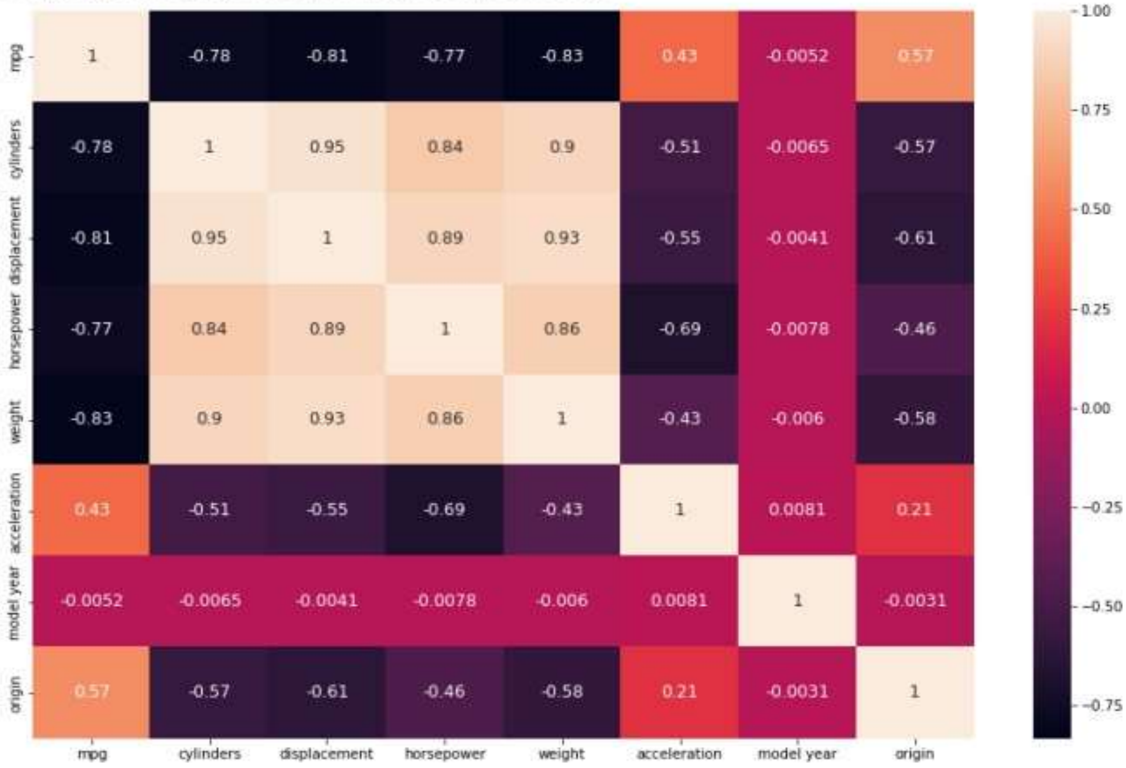

mpg
10
15
20
25
30
35
40
45
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
origin


mpg
10
15
20
25
30
35
40
45
car name



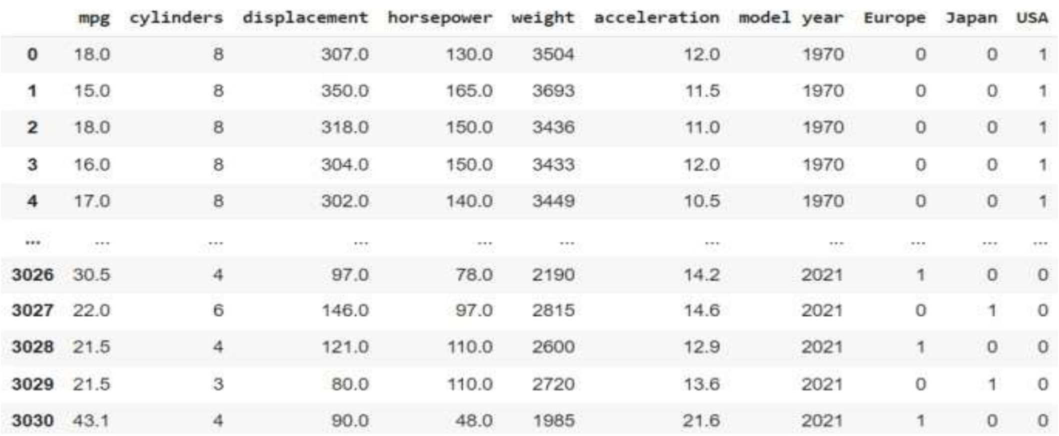
ONE-HOT ENCODING:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fdf6d208110>
```



	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
mpg	1	-0.78	-0.81	-0.77	-0.83	0.43	-0.0052	0.57
cylinders	-0.78	1	0.95	0.84	0.9	-0.51	-0.0065	-0.57
displacement	-0.81	0.95	1	0.89	0.93	-0.55	-0.0041	-0.61
horsepower	-0.77	0.84	0.89	1	0.86	-0.69	-0.0078	-0.46
weight	-0.83	0.9	0.93	0.86	1	-0.43	-0.006	-0.58
acceleration	0.43	-0.51	-0.55	-0.69	-0.43	1	0.0081	0.21
model year	-0.0052	-0.0065	-0.0041	-0.0078	-0.006	0.0081	1	-0.0031
origin	0.57	-0.57	-0.61	-0.46	-0.58	0.21	-0.0031	1

The importance and the values to be considered among all the attributes are known by this. Accordingly, the assumptions and analysis are done to build the model.

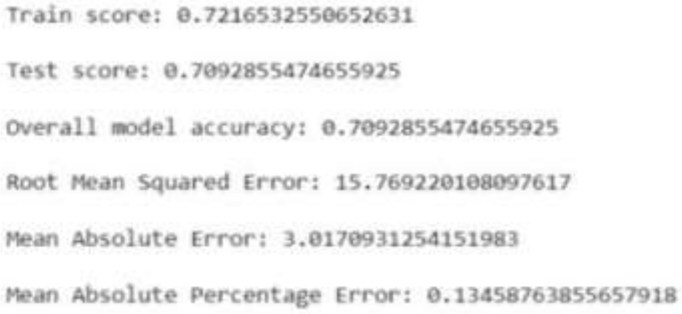


	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	Europe	Japan	USA
0	18.0	8	307.0	130.0	3504	12.0	1970	0	0	1
1	15.0	8	350.0	165.0	3693	11.5	1970	0	0	1
2	18.0	8	318.0	150.0	3436	11.0	1970	0	0	1
3	16.0	8	304.0	150.0	3433	12.0	1970	0	0	1
4	17.0	8	302.0	140.0	3449	10.5	1970	0	0	1
...	...	...	...	...	...	...	...	...	...	...
3026	30.5	4	97.0	78.0	2190	14.2	2021	1	0	0
3027	22.0	6	146.0	97.0	2815	14.6	2021	0	1	0
3028	21.5	4	121.0	110.0	2600	12.9	2021	1	0	0
3029	21.5	3	80.0	110.0	2720	13.6	2021	0	1	0
3030	43.1	4	90.0	48.0	1985	21.6	2021	1	0	0

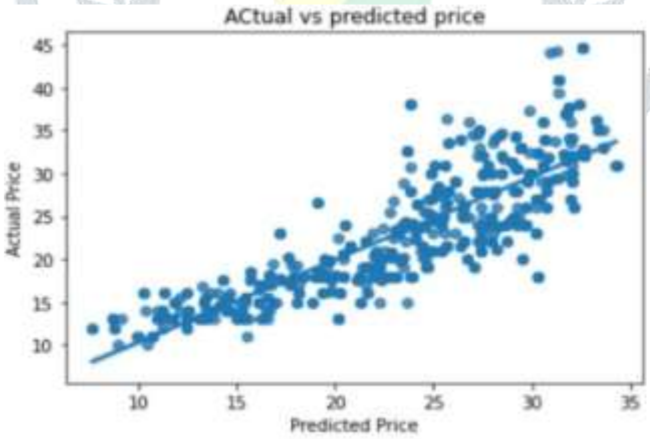
5 Training

Training the models is an essential process in making it ready for the testing phase and able to make the required functions for doing the required predictions. The following are the various models that are used for our predictions.

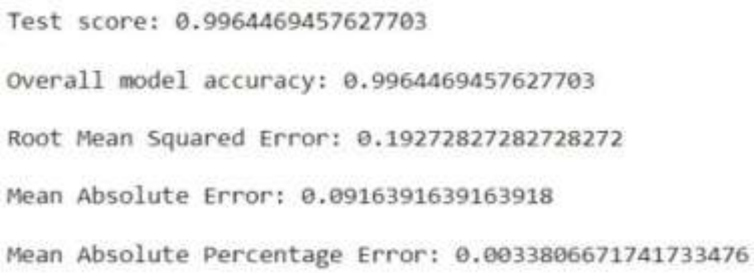
a) Linear Regression,
Below are the parameters been checked along with the predicted price versus the actual price that has been received upon using this model for predictions.



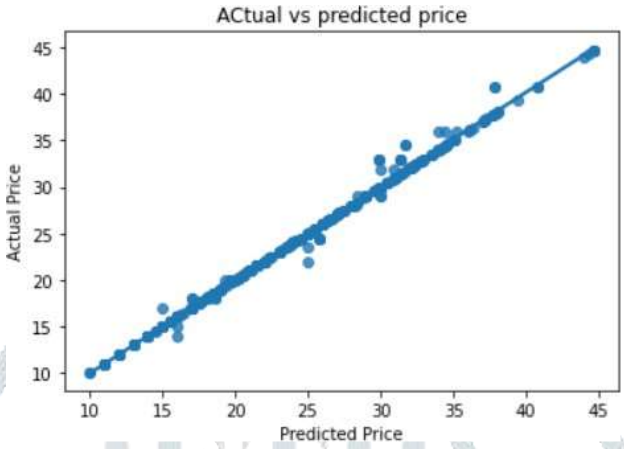
```
Train score: 0.7216532550652631

Test score: 0.7092855474655925

Overall model accuracy: 0.7092855474655925

Root Mean Squared Error: 15.769220108097617

Mean Absolute Error: 3.0170931254151983

Mean Absolute Percentage Error: 0.13458763855657918
```



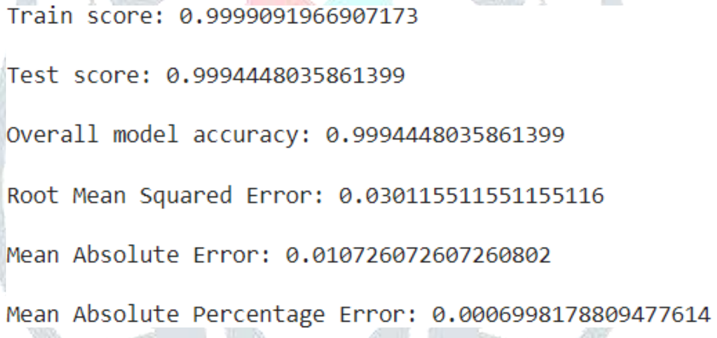
b) KNN
Below are the parameters been checked along with the predicted price versus the actual price that has been received upon using this model for predictions.



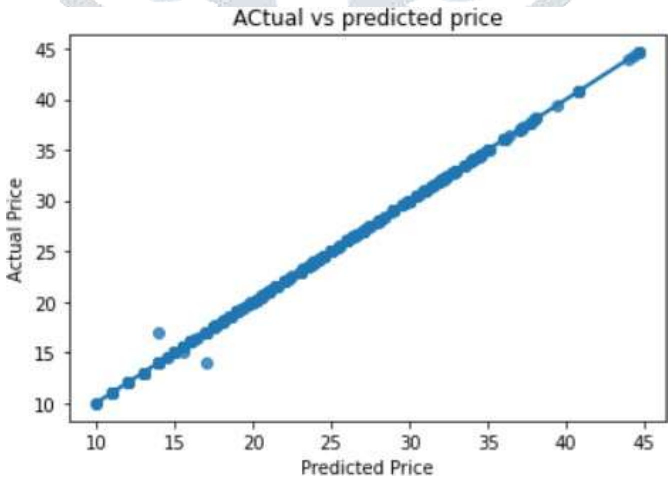
```
Test score: 0.9964469457627703

Overall model accuracy: 0.9964469457627703

Root Mean Squared Error: 0.19272827282728272

Mean Absolute Error: 0.0916391639163918

Mean Absolute Percentage Error: 0.0033806671741733476
```



c) Decision Tree
Below are the parameters been checked along with the predicted price versus the actual price that has been received upon using this model for predictions.



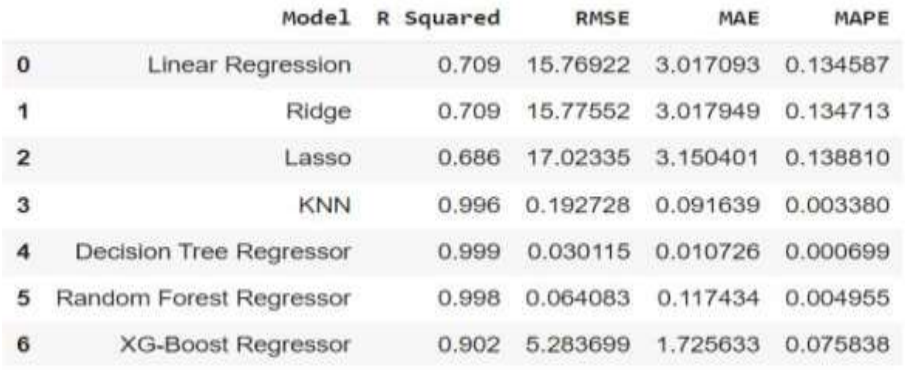
```
Train score: 0.9999091966907173

Test score: 0.9994448035861399

Overall model accuracy: 0.9994448035861399

Root Mean Squared Error: 0.030115511551155116

Mean Absolute Error: 0.010726072607260802

Mean Absolute Percentage Error: 0.0006998178809477614
```



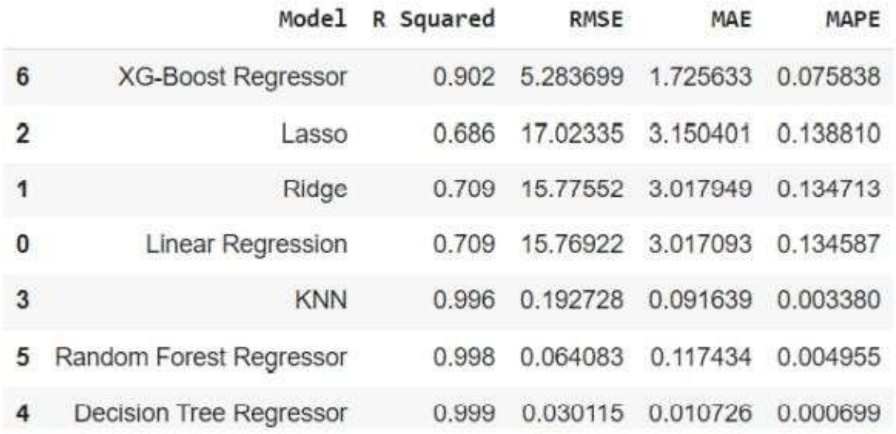
6 TESTING

We have trained the model with various machine learning algorithms and conducted testing on them to see the various comparisons by comparing the various errors and results. We have chosen our best model based on these results.

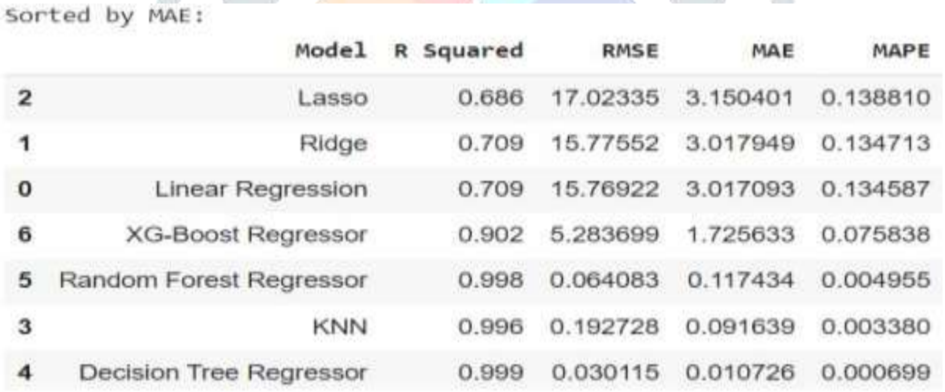
Model Performances



	Model	R Squared	RMSE	MAE	MAPE
0	Linear Regression	0.709	15.76922	3.017093	0.134587
1	Ridge	0.709	15.77552	3.017949	0.134713
2	Lasso	0.686	17.02335	3.150401	0.138810
3	KNN	0.996	0.192728	0.091639	0.003380
4	Decision Tree Regressor	0.999	0.030115	0.010726	0.000699
5	Random Forest Regressor	0.998	0.064083	0.117434	0.004955
6	XG-Boost Regressor	0.902	5.283699	1.725633	0.075838

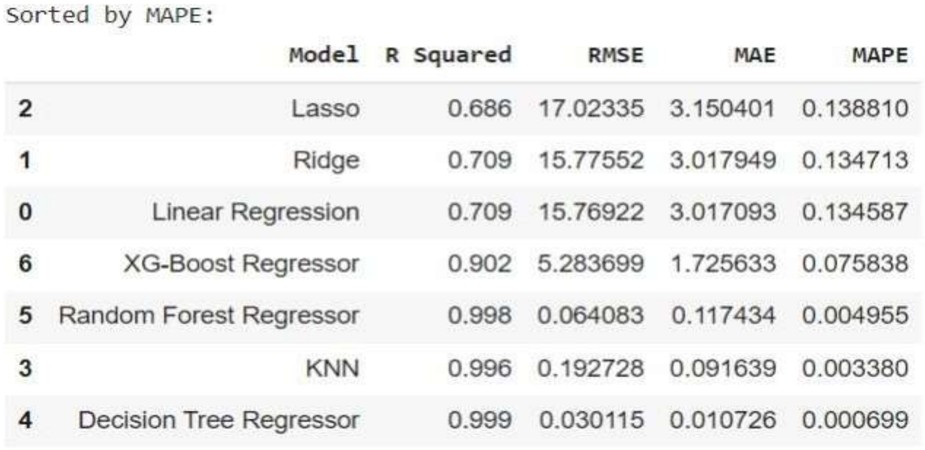


	Model	R Squared	RMSE	MAE	MAPE
6	XG-Boost Regressor	0.902	5.283699	1.725633	0.075838
2	Lasso	0.686	17.02335	3.150401	0.138810
1	Ridge	0.709	15.77552	3.017949	0.134713
0	Linear Regression	0.709	15.76922	3.017093	0.134587
3	KNN	0.996	0.192728	0.091639	0.003380
5	Random Forest Regressor	0.998	0.064083	0.117434	0.004955
4	Decision Tree Regressor	0.999	0.030115	0.010726	0.000699



Sorted by MAE:

	Model	R Squared	RMSE	MAE	MAPE
2	Lasso	0.686	17.02335	3.150401	0.138810
1	Ridge	0.709	15.77552	3.017949	0.134713
0	Linear Regression	0.709	15.76922	3.017093	0.134587
6	XG-Boost Regressor	0.902	5.283699	1.725633	0.075838
5	Random Forest Regressor	0.998	0.064083	0.117434	0.004955
3	KNN	0.996	0.192728	0.091639	0.003380
4	Decision Tree Regressor	0.999	0.030115	0.010726	0.000699



Sorted by MAPE:

	Model	R Squared	RMSE	MAE	MAPE
2	Lasso	0.686	17.02335	3.150401	0.138810
1	Ridge	0.709	15.77552	3.017949	0.134713
0	Linear Regression	0.709	15.76922	3.017093	0.134587
6	XG-Boost Regressor	0.902	5.283699	1.725633	0.075838
5	Random Forest Regressor	0.998	0.064083	0.117434	0.004955
3	KNN	0.996	0.192728	0.091639	0.003380
4	Decision Tree Regressor	0.999	0.030115	0.010726	0.000699

TEST RESULT:

Upon comparing and testing all the models, we found that Decision tree is the most efficient one among all the models with an accuracy of 99.9%.

7 CONCLUSION

Fuel prices are increasing rapidly each day, and the demand of vehicles with better fuel efficiency or Miles per gallon is growing tremendously. This situation leads consumers to choose vehicles wisely, on the other hand the vehicle manufacturers also have a tight competition and close margins to deal with to have a better vehicle in the market. In this kind of situations our model to predict the fuel efficiency of vehicles will come into action for making effective vehicles by knowing its specifications beforehand and make more popular vehicles that outshine competitors.

During this project, our main objective was to predict the vehicle’s fuel efficiency or the MPG (Miles per gallon). We have done the data preprocessing to make the dataset free from null values and other disturbances, then we performed data visualization of the data represent and know well about the attributes in the dataset. We have implemented various machine learning models and checked for their errors and accuracies until we get the best effective model for the data taken. Upon attaining the best fit model, we have deployed it. In the deployment page, it calculates the result value and gives that answer based on the probability and calculations been made by that model. Here the user can give the dataset taken or their own values as an input to the model and it gives the output based on the comparison of the probability of that MPG.

REFERENCES

[1] Sandareka Wickramanayake and H.M.N. Dilum Bandara, “Fuel Consumption Prediction of Fleet Vehicles Using Machine Leaning: A Comparative Study”.

[2] Varun Shirbhayye, Deepesh Kurmi, Siddharth Dyavanapalli, Agraharam Sai Hari Prasad, Nidhi Lal, “An Accurate Prediction of MPG (Miles Per Gallon) using Linear Regression Model of Machine Learning”.

[3] J. Lindberg, “Fuel consumption prediction for he’avy vehicles using machine learning on log data,” Master’s thesis, KTH, School of Computer Science and Communications (CSC), 2014.

[4] J. Gondar, M. Earleywine, and W. Sparks “Analyzing vehicle fuel saving opportunities through intelligent driver feedback,” SAE World Congr., Detroit, Michigan, 2012.

[5] J. S. Stichter, “Investigation of vehicle and driver aggressivity and relation to fuel economy testing,” Master’s thesis, University of Iowa, Iowa, 2012.

[6] I. M. Berry, “The effects of driving style and vehicle performance on the real-world fuel consumption of U.S. light-duty vehicles,” Master’s thesis, Massachusetts Institute of Technology, Cambridge, 2010.

[7] L.Rokach, (2009, Nov, 19), Ensemble-based classifiers, [Online], Available: http://www.ise.bgu.ac.il/faculty/liorr/AI.pdf

[8] A. Liaw and M. Wiener, “Classification and regression by random forest,” R News, vol. 2, no. 3, Dec. 2002.