A Machine Learning Approach predicting Fuel Efficiency in Vehicles

Mr. ALLADA BALA DHARMA SASTHRA[1], Mr. S. RAJA[2]

1. DEPARTMENT OF COMPUTER SCIENCE AND SYSTEMS ENGINEERING, ANDHRA UNIVERSITY, COLLEGE OF ENGINEERING VISAKHAPATNAM-530 003.
2. Professor in DEPARTMENT OF COMPUTER SCIENCE AND SYSTEMS ENGINEERING, ANDHRA UNIVERSITY, COLLEGE OF ENGINEERING VISAKHAPATNAM-530 003.

ABSTRACT

Think about how the automobile industry has developed over the past 200 years. As fuel prices have increased and consumers have become more picky about features, automakers are continuously improving their production methods to improve fuel efficiency. But what if you could have a trustworthy estimator for cars given some known specifications about the vehicle improving vehicle fuel economy and combating fraud in fleet management require the ability to model and anticipate fuel usage. A vehicle's fuel consumption is influenced by both internal factors like distance, load, vehicle attributes, and driver behavior, as well as external elements like weather, traffic, and road conditions. However, not all of these variables may be quantifiable or accessible for the analysis of fuel efficiency. Hence, it is difficult to model and/or predict fuel economy using only the data that is now available while also indirectly accounting for as many influences from other internal and external sources. As the model can be constructed by learning the patterns in the data, machine learning is appropriate in this type of study. To circumvent these issues and to compare to the svm algorithm, we are utilizing two algorithms (random forest and svm). The closest result is produced by Random Forest, which is more accurate.

1 INTRODUCTION

In order to improve fuel economy of vehicles and stop fraudulent actions in fleet management, it is essential to be able to recognize the variables that affect fuel consumption and then be able to predict it. Take a long-distance bus, for instance, which connects a major metropolis with a remote area. The bus might run into traffic or different types of terrain along the trip, including crossing a mountainous location. Additionally, depending on the day of the week, the bus's load, weather, and traffic volume may change. Additionally, different drivers may operate the bus on other days. As a result, the bus's fuel usage may differ significantly across days. A wide range in gasoline usage creates room for fuel fraud. For instance, on days when the roads were clear and few people boarded the bus, many liters’ of fuel could be poured out of the tank covertly. These activities are common in nations where the cost of fuel is quite high. The bus owners, on the other hand, might wish to be aware of the elements that affect fuel use so they can implement appropriate process reengineering measures to cut fuel consumption. Additionally, owners may be able to spot potential fuel fraud by being able to estimate gasoline consumption. Vehicle owners can now record high resolution, multi-variate time series datasets linked to a vehicle's position, speed, engine conditions, and other factors thanks to the development of Global Positioning System (GPS) based tracking devices and precise fuel sensors.usage of fuel. Other Key Performance Indicators (KPIs) like as idling time, day of the week, and driver acceleration and breaking patterns can also be derived from such data. However, information about a number of other affecting elements, including load, traffic, weather, and driver, is still not gathered by a single tracking device and is typically not quantifiable for each vehicle. We analyze a situation where such information is not available (which is prevalent in rural areas and in underdeveloped nations), even though some of this information is accessible through other third-party systems/services. As a result, the problem is how to model and estimate fuel consumption when just a portion of the major fuel consumption-related elements are present. The development of large strong analytic tools is needed to turn the data into insightful knowledge. The two cultures of statistical modeling—the culture of data modelling and the culture of algorithmic modeling—are used to draw inferences from data. In the culture of data modelling, the black box, which independent variables enter and response variables exit, is assumed to have a stochastic data model. The culture of algorithmic modelling views this black box's interior as complicated and unknowable. They start with the result and locate a method, y = f(x), that uses x (predictor variables) to forecast replies, y. When interactions, high-dimensional data, linear and non-linear correlations, and other complex linkages between the predictors and responses are present, the machine ML performs better than conventional statistical models. As a result, ML is appropriate for predicting fuel usage because the model can be created by discovering patterns in the available data, which can then be applied to prediction. The majority of prior ML-based studies on fuel consumption have employed datasets from highway-bound vehicle traffic.

2. LITERATURESURVEYANDRELATEDWORK

[1] "Fuel Consumption Prediction of Fleet Vehicles Using Machine Leaning: A Comparative Study", Sandareka Wickramanayake and H.M.N. DilumBandara. Improving vehicle fuel economy and combating fraud in fleet management require the ability to model and anticipate fuel usage. A vehicle's fuel consumption is influenced by both internal factors like distance, load, vehicle attributes, and driver behavior, as well as external elements like weather, traffic, and road conditions. For the examination of fuel usage, some of these variables might not be measured or available. When only a portion of the aforementioned characteristics are accessible as a multi-variate time series from a long-distance public bus, that situation is what we are going to be looking at. In light of this, the task is to estimate and/or predict the fuel use using the information that is already available, but also inadvertently incorporating influences from additional internal and external elements. Machine learning (ML) is appropriate in this analysis since the model may be built by discovering patterns in the data. In this study, given all the relevant parameters as a time series, we assess the prediction performance of three ML approaches in estimating the fuel consumption of the bus. The investigation leads to the conclusion that, when compared to gradient boosting and neural networks, the random forest technique yields a more accurate forecast.

[2]Use of Artificial Neural Networks to Predict Fuel Consumption Based on Technical Parameters of Vehicles by JaroslawZiókowski, Mateusz Oszczypa, Jerzy Maachowski, and Joanna Szkutnik-Rogoz. This article provides a thorough examination. a discussion of the environmental impact of people and how much fuel motor vehicles use, with a focus on passenger cars. The study methodology chosen is based on the usage of artificial neural networks to develop a prediction model that can be used to estimate the fuel consumption of motor vehicles. The artificial neural networks were trained using a database of 1750 records, which contained data on vehicles produced in the previous ten years. From the generated neural networks, the MLP (Multi-Layer Perceptron) 22-10-3 network was chosen and then subjected to analysis.

[3] "Study on Prediction Method of Flight Fuel Consumption with Machine Learning", Wu Zixuan, Zhang Ning, Hong Weijun, and Yu Sheng. A flight fuel forecast model based on machine learning is proposed in order to decrease the additional fuel carrying of the flight and enhance fuel economy and commercial payload of the fleet. The feature vector of flight fuel consumption is built using the flight plan, aircraft operation data, risk control data, and aircraft performance data. The random forest approach is used to construct the regression relationship between the feature vector and the flight's fuel usage. The flight fuel prediction model is trained using historical flight data, and a comparative experiment is created to confirm the model's predictions. validity and the model's capacity for fitting. The model will be put into use soon to aid X airline's flight dispatchers in creating flight plans.

[4]"Fuel Efficiency Modelling and Prediction for Automotive Vehicles: A Data-Driven Approach", Xunyuan Yin, Zhaojian Li, Sirish L. Shah, and Lisong Zhang. The primary focus of this study is the modelling and prediction of fuel efficiency for typical automobiles using a useful vehicle database. The mutual information index (MII) is used in conjunction with the historical database processing to find a set of traits that have a substantial impact on fuel efficiency. To create fuel efficiency prediction models, five distinct machine learning algorithms are used. Quantile regression, a natural extension of traditional least square estimation, is proven to have better performance among these methods.

3 PROPOSED WORK AND ALGORITHM

A system's structure, behavior, and other aspects are defined by its system architecture, a conceptual model. A formal description and representation of a system that facilitates inferences about its structure and behavior is called an architecture description.

A system representation that incorporates the functional mapping of hardware and software components, the software architecture mapping to the hardware architecture, and the interaction of humans with these components. The suggested solution involves training the model so that it can more accurately estimate fuel usage as an output. Miles per gallon (MPG) prediction is presented utilizing Random forest and Support vector machines. In this study, UCI machine learning Auto MPG was used.

Benefits of Proposed System: • With the help of this project, we can quickly determine the MPG for each car based on the vehicle's pistons, journey distance, overall weight, and point of origin.
- Predicting fuel economy using a machine learning technique.
- Free of charge
- The dataset is compact and contains all the data needed to train and test the model.
- Simple to use

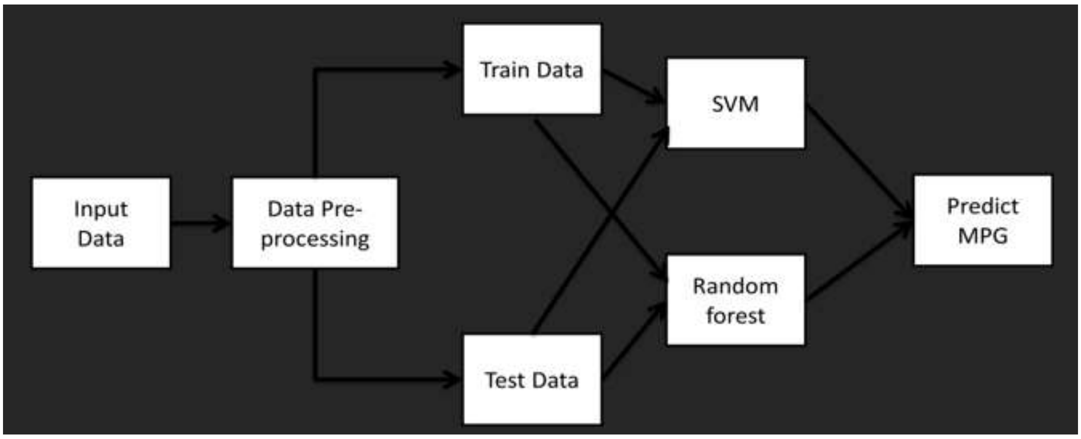


Fig 1: System architecture

4 METHODOLOGIES

Datasets are groups of data. The contents of a single database table or statistical data matrix, where each column of the table corresponds to a specific variable and each row to a specific member of the dataset under consideration, serve as the most typical examples of datasets. The data set includes values for each variable, such as an object's height or weight, for each dataset participant. A data structure of some kind is used to organise a data set. A data set in a database, for instance, can include a selection of company data (names, salary, contact details, sales numbers, etc.). As well as the data sets included within it, the database itself can be regarded as a data set. tied to a certain type of data, such sales statistics for a specific corporate division.

PREPROCESSING OF DATA

A data mining technique called data pre-processing is used to convert raw data into a format that is both practical and effective. Before implementing machine learning algorithms, this step is taken. It changes the original data into a format that a specific algorithm can utilise. The several steps involved in data pre-processing include data cleansing, feature selection, and data transformation.

Cleaning of Data

Data cleaning is the process of eliminating or changing data that is inaccurate, lacking and unnecessary, duplicated, or formatted incorrectly in order to prepare it for analysis. When it comes to data analysis, this information is typically neither necessary nor beneficial because it could incorrect outcomes. Depending on how the data is stored and the questions that need to be answered, there are many techniques for cleaning the data. Data cleaning is not just about deleting data to create room for new data; rather, it is about figuring out how to increase a data set's accuracy without necessarily deleting data. For starters, data cleaning is more than just deleting data; it also involves addressing spelling and grammatical flaws, standardising data sets, and repairing errors like empty fields, missing codes, and locating duplicate data points. Data cleaning is regarded as a key component of the fundamentals of data science since it is crucial to the discovery of valid conclusions and the analytical process.

Visualisation of data

A graphic depiction of information and data is known as data visualisation. Utilising visual Data visualisation tools offer a simple way to examine and comprehend trends, outliers, and patterns in data using features like charts, graphs, and maps. To analyse vast volumes of data and make data-driven decisions, data visualisation tools and technologies are crucial in the world of big data. Colours and patterns catch our attention. We can swiftly distinguish between red and blue, and a square from a circle. Everything in our culture is visual, from TV and movies to ads and art. Another sort of visual art that captures our attention and keeps it fixed on the message is data visualisation. We can immediately see trends and outliers when we look at a chart. We easily internalise something if we can see it. It's narrative with a goal. When you've If you've ever tried to see a trend in a huge spreadsheet of data, you know how much more impactful a visualisation can be.

Transformation of Data

The process of changing data from one format to another, usually from that of a source system into that needed by a destination system is known as data transformation. Most data integration and management operations, including data wrangling and data warehousing, include some type of data transformation. Data transformation, a step in the ELT/ETL process, can be categorized as "simple" or "complex" depending on the types of adjustments that must be made to the data before it is sent to its intended destination. The data transformation procedure can be carried out automatically, manually, or using a combination of both.

SPLITTING A DATASET

Data is typically divided into training and test sets in machine learning. The explanation for this is simple: it is impractical to try to evaluate your system using the data that you used to train it. The entire purpose of a machine learning system is to be able to operate with unobserved data; therefore, if you know that all potential values will be included in your training data, you might as well use a lookup method. A crucial step in reviewing data mining algorithms is dividing the data into training and testing sets.

Practice Set

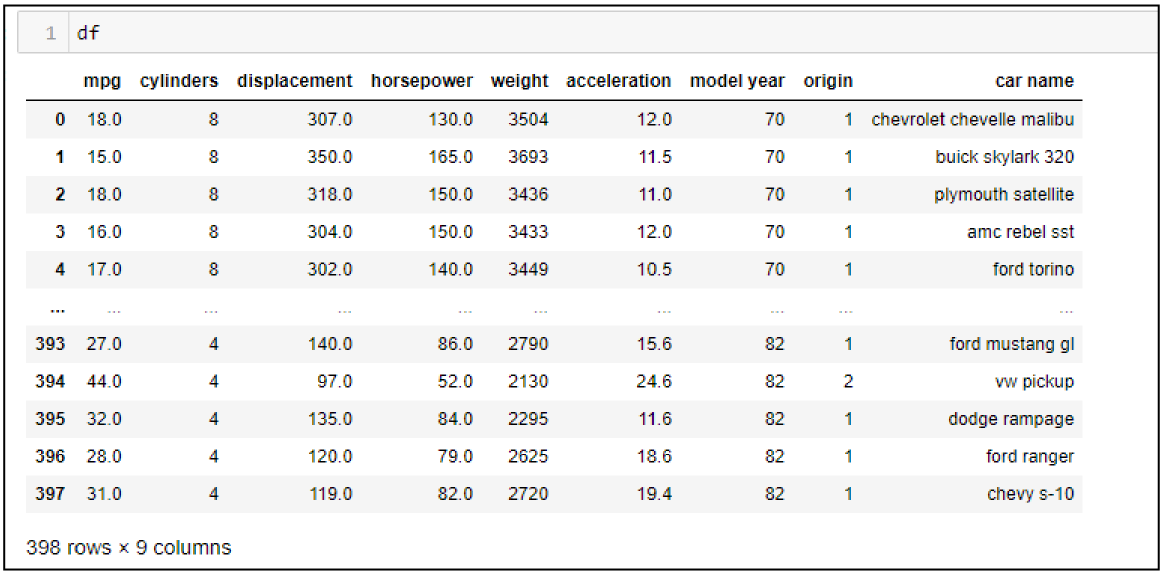
The training set's observations serve as the learning experience for the algorithm. Each observation in supervised learning problems consists of one or more observed input variables and an output variable that has been observed.

In this project, 398 records of training data for the xtrain and ytrain variables with a test size of 0.2 are provided. These 398 data are made up of 9 different car-related characteristics.

Exam Set

The test set is a collection of data that is used to assess the model's effectiveness using various performance indicators. The test set cannot contain any observations from the training set. It will be challenging to determine if the algorithm has learned to generalise from the training set or has merely memorised it if the test set does contain examples from the training set.

5. RESULTSANDDISCUSSIONSCREENSHOTS

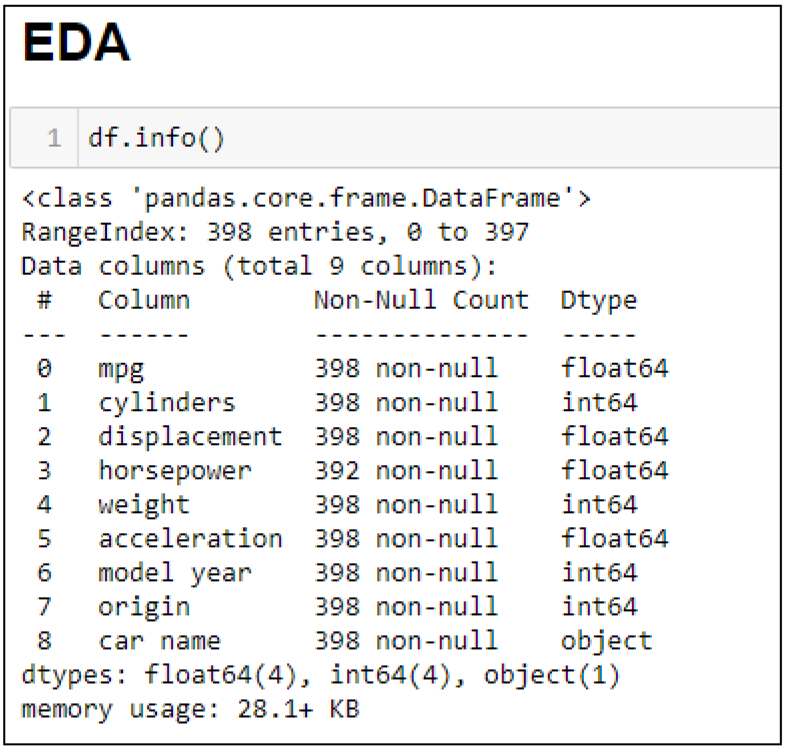


```
df
```

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
0	18.0	8	307.0	130.0	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165.0	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150.0	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150.0	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140.0	3449	10.5	70	1	ford torino
...	...	...	...	...	...	...	...	...	...
393	27.0	4	140.0	86.0	2790	15.6	82	1	ford mustang gl
394	44.0	4	97.0	52.0	2130	24.6	82	2	vw pickup
395	32.0	4	135.0	84.0	2295	11.6	82	1	dodge rampage
396	28.0	4	120.0	79.0	2625	18.6	82	1	ford ranger
397	31.0	4	119.0	82.0	2720	19.4	82	1	chevy s-10

398 rows × 9 columns

Fig 2: sample record in dataset

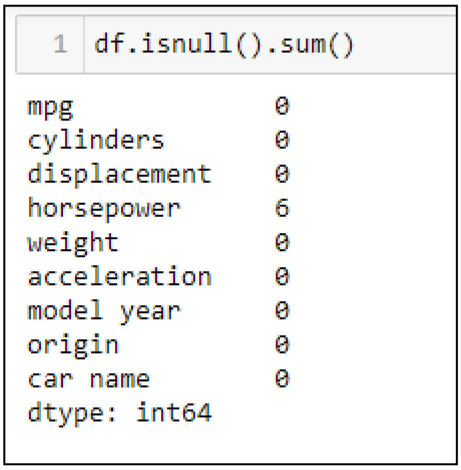


EDA

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 398 entries, 0 to 397
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   mpg           398 non-null    float64
 1   cylinders     398 non-null    int64
 2   displacement  398 non-null    float64
 3   horsepower    392 non-null    float64
 4   weight        398 non-null    int64
 5   acceleration  398 non-null    float64
 6   model year    398 non-null    int64
 7   origin        398 non-null    int64
 8   car name      398 non-null    object
dtypes: float64(4), int64(4), object(1)
memory usage: 28.1+ KB
```

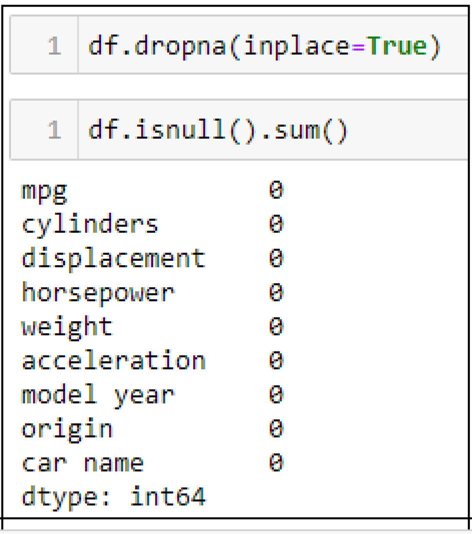
Fig 3: Eda analysis



```
df.isnull().sum()
```

```
mpg             0
cylinders       0
displacement    0
horsepower      6
weight          0
acceleration    0
model year      0
origin          0
car name        0
dtype: int64
```

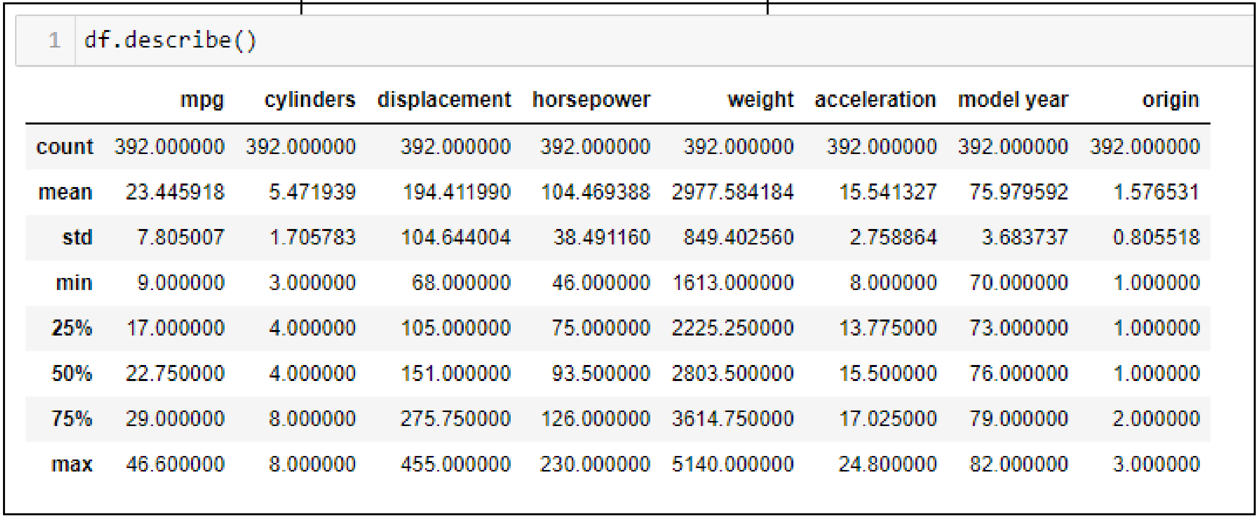
Fig 4: Checking the null values



```
df.dropna(inplace=True)
```

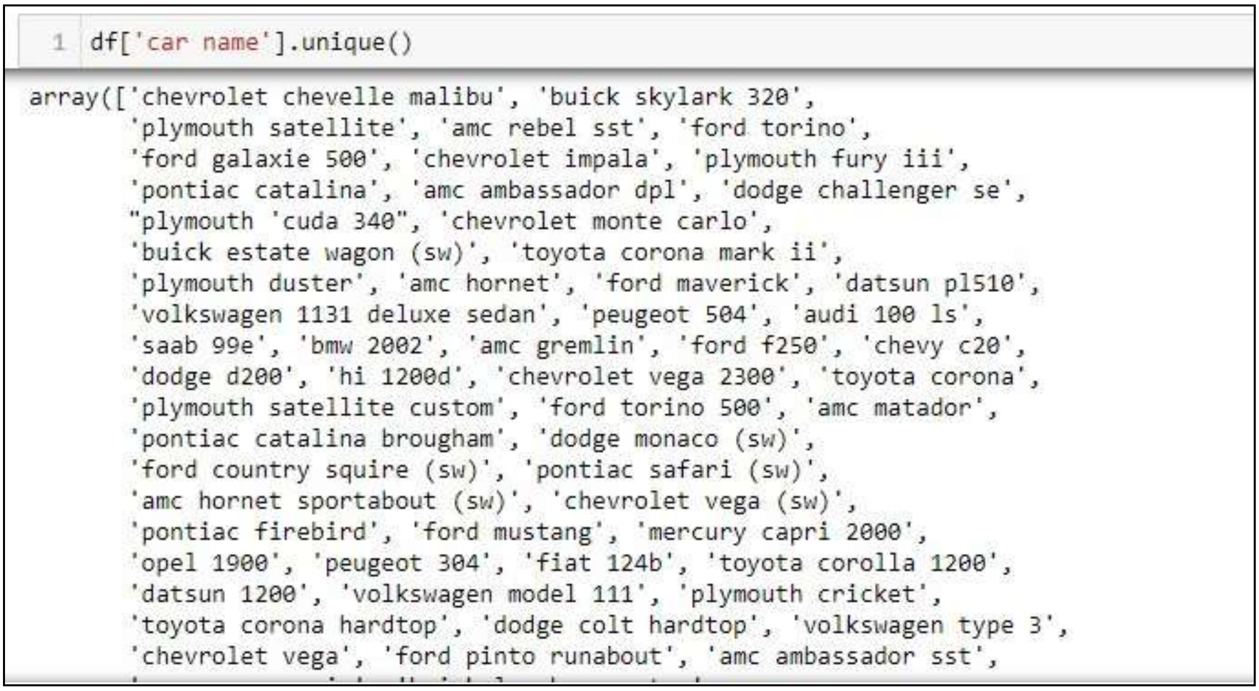
```
df.isnull().sum()
```

```
mpg             0
cylinders       0
displacement    0
horsepower      0
weight          0
acceleration    0
model year      0
origin          0
car name        0
dtype: int64
```



```
df.describe()
```

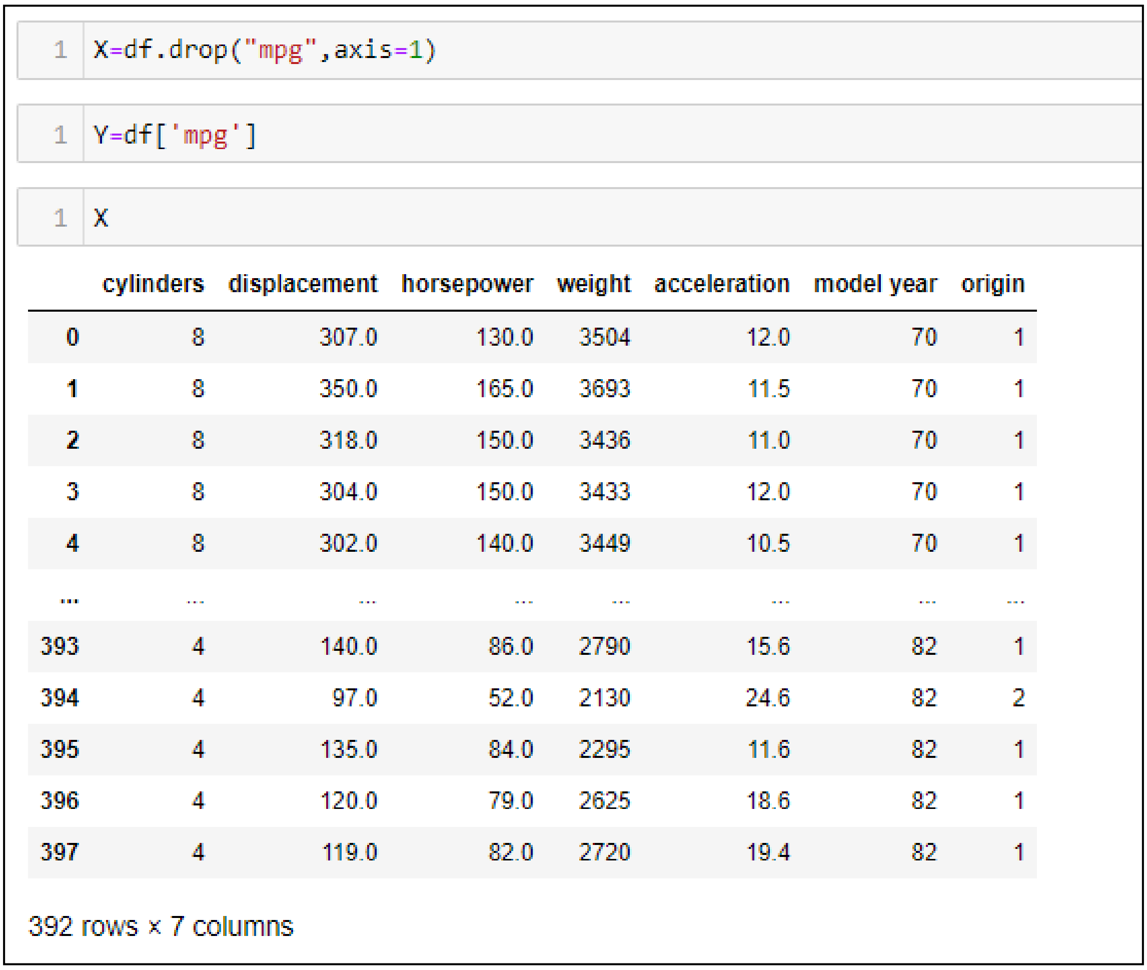
	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
count	392.000000	392.000000	392.000000	392.000000	392.000000	392.000000	392.000000	392.000000
mean	23.445918	5.471939	194.411990	104.469388	2977.584184	15.541327	75.979592	1.576531
std	7.805007	1.705783	104.644004	38.491160	849.402560	2.758864	3.683737	0.805518
min	9.000000	3.000000	68.000000	46.000000	1613.000000	8.000000	70.000000	1.000000
25%	17.000000	4.000000	105.000000	75.000000	2225.250000	13.775000	73.000000	1.000000
50%	22.750000	4.000000	151.000000	93.500000	2803.500000	15.500000	76.000000	1.000000
75%	29.000000	8.000000	275.750000	126.000000	3614.750000	17.025000	79.000000	2.000000
max	46.600000	8.000000	455.000000	230.000000	5140.000000	24.800000	82.000000	3.000000



```
df['car name'].unique()
```

```
array(['chevrolet chevelle malibu', 'buick skylark 320',
       'plymouth satellite', 'amc rebel sst', 'ford torino',
       'ford galaxie 500', 'chevrolet impala', 'plymouth fury iii',
       'pontiac catalina', 'amc ambassador dpl', 'dodge challenger se',
       "plymouth 'cuda 340", 'chevrolet monte carlo',
       'buick estate wagon (sw)', 'toyota corona mark ii',
       'plymouth duster', 'amc hornet', 'ford maverick', 'datsun pl510',
       'volkswagen 1131 deluxe sedan', 'peugeot 504', 'audi 100 ls',
       'saab 99e', 'bmw 2002', 'amc gremlin', 'ford f250', 'chevy c20',
       'dodge d200', 'hi 1200d', 'chevrolet vega 2300', 'toyota corona',
       'plymouth satellite custom', 'ford torino 500', 'amc matador',
       'pontiac catalina brougham', 'dodge monaco (sw)',
       'ford country squire (sw)', 'pontiac safari (sw)',
       'amc hornet sportabout (sw)', 'chevrolet vega (sw)',
       'pontiac firebird', 'ford mustang', 'mercury capri 2000',
       'opel 1900', 'peugeot 304', 'fiat 124b', 'toyota corolla 1200',
       'datsun 1200', 'volkswagen model 111', 'plymouth cricket',
       'toyota corona hardtop', 'dodge colt hardtop', 'volkswagen type 3',
       'chevrolet vega', 'ford pinto runabout', 'amc ambassador sst',
```

Fig 7: Features of data



```
X=df.drop("mpg",axis=1)
```

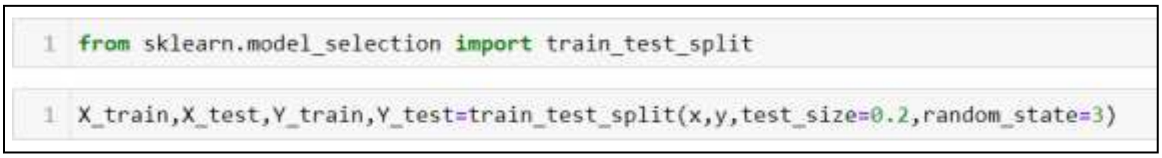
```
Y=df['mpg']
```

```
X
```

	cylinders	displacement	horsepower	weight	acceleration	model year	origin
0	8	307.0	130.0	3504	12.0	70	1
1	8	350.0	165.0	3693	11.5	70	1
2	8	318.0	150.0	3436	11.0	70	1
3	8	304.0	150.0	3433	12.0	70	1
4	8	302.0	140.0	3449	10.5	70	1
...	...	...	...	...	...	...	...
393	4	140.0	86.0	2790	15.6	82	1
394	4	97.0	52.0	2130	24.6	82	2
395	4	135.0	84.0	2295	11.6	82	1
396	4	120.0	79.0	2625	18.6	82	1
397	4	119.0	82.0	2720	19.4	82	1

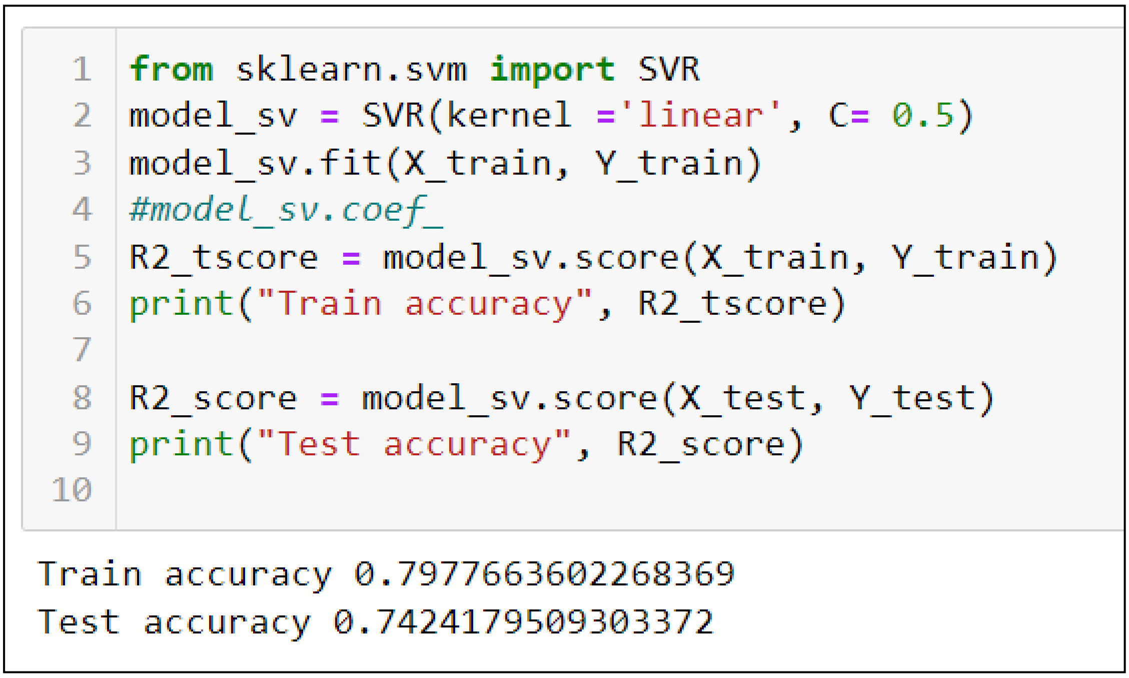
392 rows × 7 columns

Fig 8: **Features selected from data**



```
from sklearn.model_selection import train_test_split
```

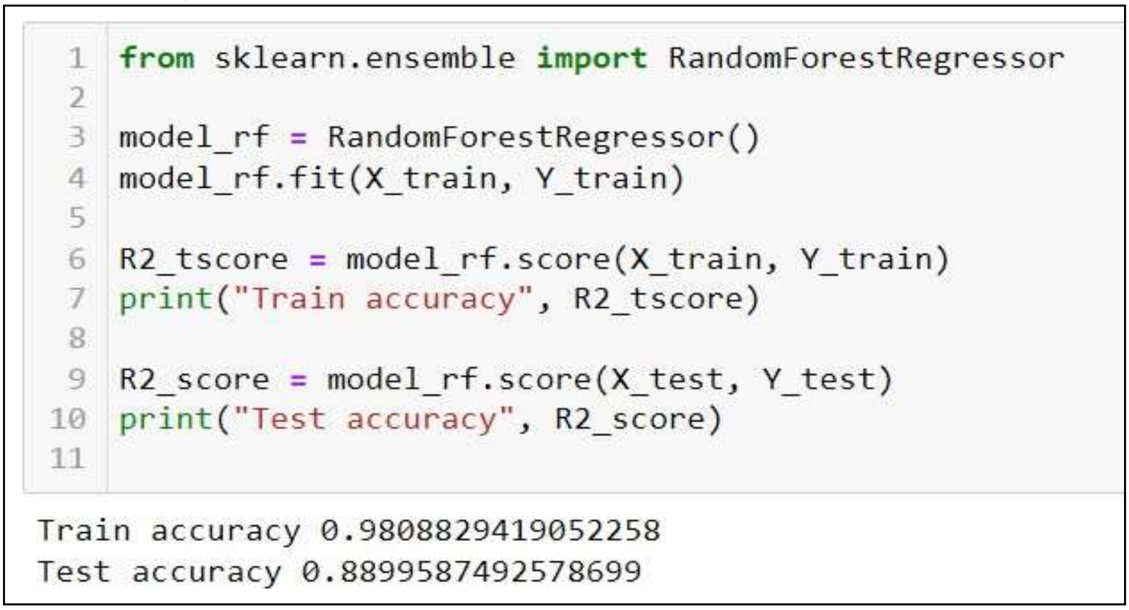
```
X_train,X_test,Y_train,Y_test=train_test_split(x,y,test_size=0.2,random_state=3)
```



```
from sklearn.svm import SVR
model_sv = SVR(kernel ='linear', C= 0.5)
model_sv.fit(X_train, Y_train)
#model_sv.coef_
R2_tscore = model_sv.score(X_train, Y_train)
print("Train accuracy", R2_tscore)

R2_score = model_sv.score(X_test, Y_test)
print("Test accuracy", R2_score)

```

```
Train accuracy 0.7977663602268369
Test accuracy 0.7424179509303372
```

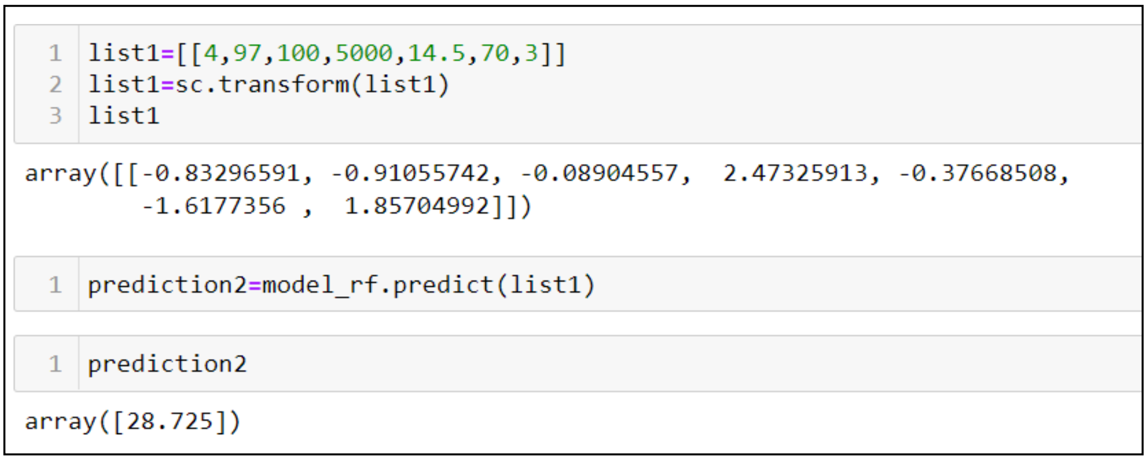
Fig 9: **Train and Test accuracy achieved using SVR Algorithm**



```
from sklearn.ensemble import RandomForestRegressor

model_rf = RandomForestRegressor()
model_rf.fit(X_train, Y_train)

R2_tscore = model_rf.score(X_train, Y_train)
print("Train accuracy", R2_tscore)

R2_score = model_rf.score(X_test, Y_test)
print("Test accuracy", R2_score)

```

```
Train accuracy 0.9808829419052258
Test accuracy 0.8899587492578699
```

Fig 10: **Train and Test accuracy using Random Forest Algorithm**



```
list1=[[4,97,100,5000,14.5,70,3]]
list1=sc.transform(list1)
list1
```

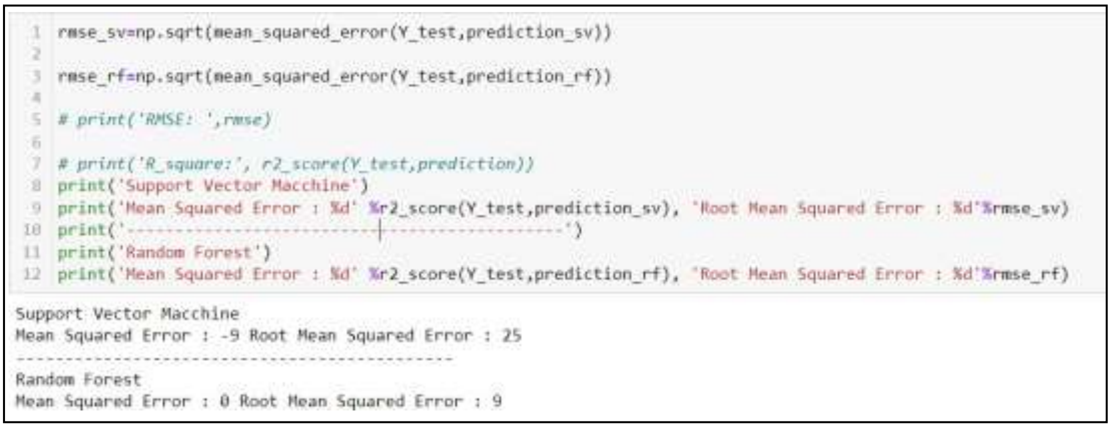
```
array([[-0.83296591, -0.91055742, -0.08904557,  2.47325913, -0.37668508,
        -1.6177356 ,  1.85704992]])
```

```
prediction2=model_rf.predict(list1)
```

```
prediction2
```

```
array([28.725])
```

Fig **9**: **predicted record using Random Forest Algorithm**



```
rmse_sv=np.sqrt(mean_squared_error(Y_test,prediction_sv))

rmse_rf=np.sqrt(mean_squared_error(Y_test,prediction_rf))

# print('RMSE: ',rmse)

# print('R_square:', r2_score(Y_test,prediction))
print('Support Vector Macchine')
print('Mean Squared Error : %d' %r2_score(Y_test,prediction_sv), 'Root Mean Squared Error : %d'%rmse_sv)
print('----------------------------------------------')
print('Random Forest')
print('Mean Squared Error : %d' %r2_score(Y_test,prediction_rf), 'Root Mean Squared Error : %d'%rmse_rf)
```

```
Support Vector Macchine
Mean Squared Error : -9 Root Mean Squared Error : 25
----------------------------------------------
Random Forest
Mean Squared Error : 0 Root Mean Squared Error : 9
```

Fig **10**: Mean square Error using SVR and Random Forest

6. CONCLUSION

On the basis of the technical parameters specified in the design assumptions, this project confirms the effectiveness of using Random forest (98% of train accuracy) to predict the fuel consumption of a passenger car powered by an internal combustion engine even at the stage of its design. The inclination of the automotive industry to reduce fuel usage in connection to the reduction of emissions of hazardous chemical compounds into the environment is one manifestation of the multifaceted nature of the problem. The goal of future research should be to increase the predictive models' accuracy by giving them more specifics and regularly updating the database.

7. REFERENCES

[1] L. Breiman, “Statistical Modeling: The Two Cultures,” Statistical Science, vol. 16, pp. 199-231, 2001.

[2] A. Viswanathan, “Data driven analysis of usage and driving parameters that affect fuel consumption of heavy vehicles,” Master’s thesis, Linköping University, Sweden, 2013.

[3] J. Lindberg, “Fuel consumption prediction for heavy 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.

Bicycle-based Public Transport System," Pervasive and Mobile Computing, vol. 6, no. 4, pp. 455–466, 2010.