EMISSION OF CO2 FROM CARS

Using Random Forest Regression

Developed by: Saloni MP, Chandana M, Aparna Rai,
Neha UK, Banushree DJ Smart Bridge-Remote
Summer Internship Program

1. INTRODUCTION

As we all know, Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The basic process of machine learning is to give training data to a learning algorithm. The learning algorithm then generates a new set of rules, based on inferences from the data. This is in essence generating a new algorithm, formally referred to as the machine learning model. By using different training data, the same learning algorithm could be used to generate different models.

Growth comes in tandem with industrialization, which involves the use of energy, leading to increased carbon emissions and environmental degradation. Therefore, as a country industrializes, its pollution levels will increase significantly. Less developed countries pollute less because they generally have low incomes, so people are unable to afford products which contribute to increased pollution, such as cars, planes, televisions, etc. Carbon dioxide emissions or CO2 emissions are emissions stemming from the burning of fossil fuels and the manufacture of cement; they include carbon dioxide produced during consumption of solid, liquid, and gas fuels as well as gas flaring.

Transport is responsible for nearly 30% of the EU's total CO2 emissions, of which 72% comes from road transportation. Significantly reducing CO2 emissions from transport will not be easy, as the rate of emission reductions has slowed. Other sectors have cut emissions since 1990, but as more people become more mobile, CO2 emissions from transport are increasing. If we focus on cars, Newer cars, registered after 2017, are taxed under an entirely different system. The first year of tax - when a car is brand new - is based on CO2 emissions. Choosing a vehicle with low CO2 emissions can save hundreds - or even thousands - of pounds in company car tax. Cars with low g/km CO2 ratings are placed in lower company car tax bands, reducing the level of tax that drivers pay.

Here we use Machine Learning to predict CO2 emission of a car using its data. The model uses a handful of variables to predict the CO2 emissions of a car. The purpose of this model is to make accurate CO2 emission predictions given a few variables, using the random forest regression as the foundation.
The accuracy of the models can be improved by including data from more years or including other datasets which relate to CO2 emissions. Adding more years of data gives the model more training data, which will improve accuracy. Finding more datasets, such as model, fuel type, car ownership percentage, oil and coal mining industry, and manufacturing can also improve model accuracy since these variables are likely correlated with the CO2 emissions of a car and can therefore improve the correctness of the model's predictions.
One application of these results is an increased understanding of the pattern of pollution in the world: Low income countries do not have enough money to afford the lifestyle to create pollution. Middle income countries are industrializing rapidly so they produce the highest amount of pollutants. High income countries pollute less because they can afford more clean energy technologies.

Another application of these results is new approaches and policies for sustainable development. From the variable importance from the Distributed Random Forest model, we can see the most important variables in determining carbon dioxide emissions are energy usage (in oil), renewable energy usage, urban population, and electricity usage. What this means is in order to reduce carbon emissions, people need to focus on reducing fossil fuel energy usage, increasing renewable energy, concentrate on urban development, and reducing electricity usage. As countries are able to predict the amount of pollutants entering the atmosphere based on their economic growth projection, they will be able to develop the necessary policies to keep pollutants under control in the future. Countries can fund research, development, and deployment of restricted oil usage, greater renewable energy usage, clean energy urban areas, and cutting down on electricity usage.

1.1 Overview

Nowadays, there are many factors affecting nature. Vehicle exhaust plays a major source of anthropogenic carbon dioxide (CO2) in metropolitan cities. Popular community modes (buses, trucks, taxis) and about 2.4 million private cars are the emission source of air pollution in India. Carbon dioxide is a greenhouse gas that traps the earth's heat and contributes to climate change. The relationship between transportation and air pollutants, such as CO2, CO has been well documented in a wide range of case studies. In addition, the

number of vehicles emitting CO2 beyond a particular threshold should be seized by concerning RTA region head. In transportation sector as the vehicles are rapidly growing and huge data volumes are available which represent the behavior and the risks around the sector are increased. Data Mining is one of the most motivating and vital area of research with the aim of extracting information from tremendous amount of accumulated data sets. Here a new model for classifying pollutant level in automobile sector by using Machine Learning concepts. The model has been built using data form automobile sector to predict the status of CO2 emission. One algorithms have been used to build the proposed model:-Random Forest. By using the algorithm a Flask model has been implemented and tested. Random Forest was selected as best algorithm based on accuracy.

1.2 Purpose

Our aim from the project is to make use of pandas, matplotlib, & seaborn libraries from python to extract the libraries for machine learning for the CO2 emission prediction.
Secondly, to learn how and check whether the vehicle is producing Co2 beyond the limit of threshold or not and to test and train the parameters accordingly using Random Forest regression algorithm.
And in the end, to predict whether the vehicle needs to be seized or not using ensemble techniques of combining the predictions from machine learning algorithm and withdrawing the conclusions.

2. LITERATURE SURVEY

Machine Learning is the process of analyzing data from different perspectives and extracting useful knowledge from it. It is the core of knowledge discovery process. The various steps involved in extracting knowledge from raw data as depicted in figure-1. Different data mining techniques include classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression etc. Regression is the most commonly applied data mining technique, which employs a set of pre classified examples to develop a model that can classify the records of emission at large scale. Fraud detection and getting seized risk applications are particularly well suited to regression technique. This approach frequently employs Random Forest Regression Algorithm. In classification, a training set is used to build the model as the classifier which can classify the data items into its appropriate classes. A test set is used to validate the model.

2.1 Existing Problem

The amount of Co2 emission from the transport sector (including cars) accounts for about 20% of total CO2 emissions. Accordingly, from the viewpoint of preventing global warming, reducing that proportion is a key issue. In regard to CO2 emissions from cars, fuel economy standards are getting together all over the world, so improving fuel economy of cars is strongly desired. From now onwards, it is considered that fuel economy of engines will be further improved by boosting engine efficiency and by hybridization (electrification) of cars. What's more, improving fuel economy by improving "driving operation" (i.e. the operation in which a car is driven) and by smoothing traffic flows will come into the picture in the near future. Under these circumstances, with concern for the environment from the viewpoint of reducing CO2 and other exhaust emissions, the Hitachi Group is comprehensively promoting a broad range of technical developments for reducing CO2 emissions from cars.

2.2 Proposed Solution

Machine Learning (Random Forest Regression):

Here in this project we are going to use random forest regression. By using this algorithm depends upon our input data the output of the vehicle are going to predict the CO2 emission of that particular car. So that if the emission of CO2 of that particular car is more than the threshold value then that car details should be sent to the particular RTA region head to seize the car .

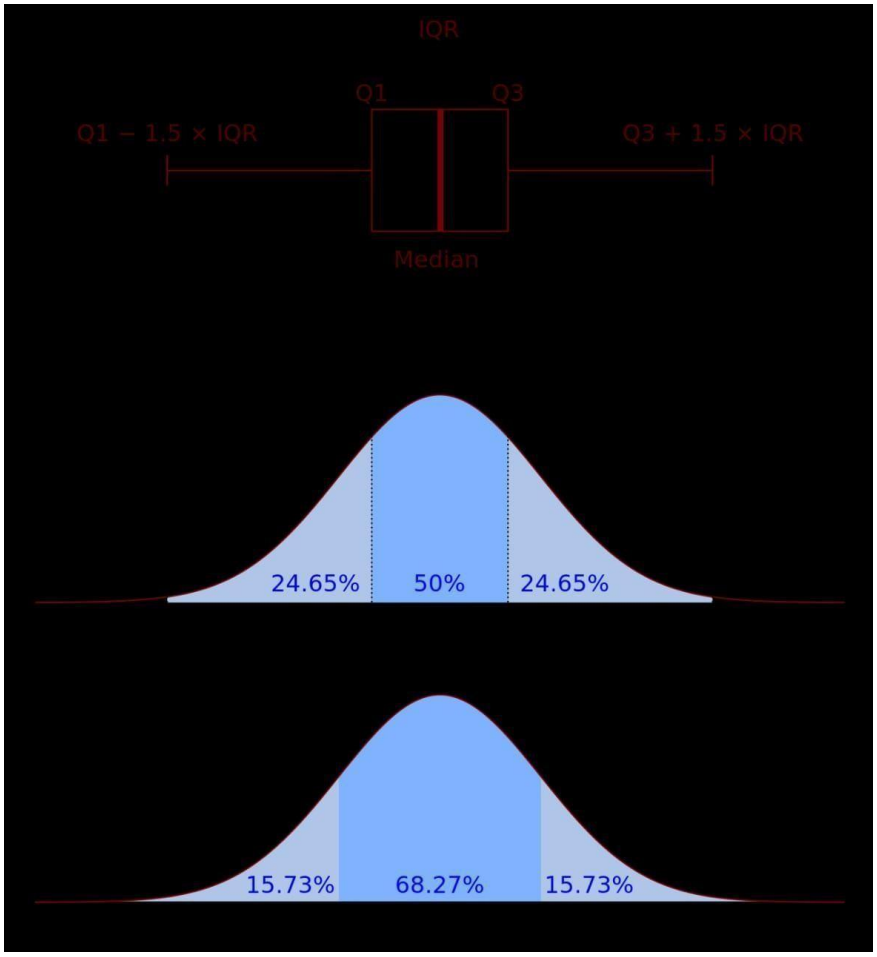
3. THEORETICAL ANALYSIS

While selecting the algorithm that gives an accurate prediction we gone through lot of algorithms which gives the results abruptly accurate and from them we selected only one algorithm for the prediction problem that is Random Forest Regression algorithm, it assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. That's how the prediction work happens with the Random Forest Algorithm.

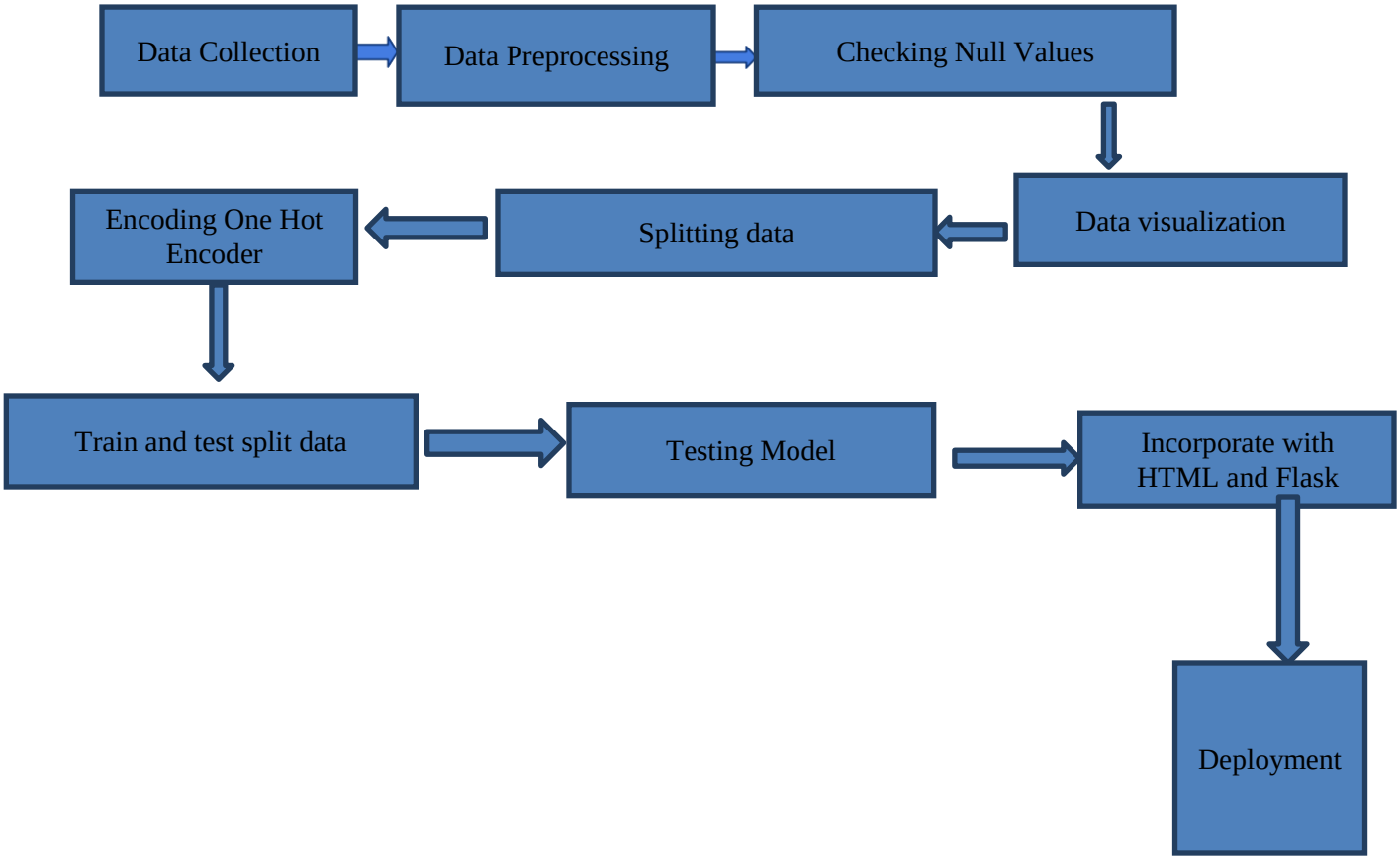
The peculiarity of this problem is collecting the vehicles details in real time and working with the prediction at the same time, so we developed a user interface for the people who'll be accessing the emission of pollutants (CO2) status prediction. Accuracy is defined as the ratio of the number of Samples correctly classified by the classifier to the total number of samples for a given test data set. The formula is as follows

$$\text{Accuracy} = [\text{IQR} = Q3\text{-}Q1]$$

At first we calculated the Q1 and Q3 values respectively to calculate IQR (inter quartile range) value. This is also called as mid-spread, H-spread, is a measure of statistical dispersion being equal tothe difference between 75^{th} and 25^{th} percentiles, or between upper and lower quartiles



3.1 Block Diagram



3.2. Software Designing

- Jupyter Notebook Environment
- Spyder Ide
- Machine Learning Algorithms
- Python (pandas, numpy, matplotlib, seaborn, sklearn)
- HTML
- Flask

We developed this emission of CO2 status prediction by using the Python language which is a interpreted and high level programming language and using the Machine Learning algorithms. For coding we used the Jupyter Notebook environment of the Anaconda distributions and the Spyder, it is an integrated scientific programming in the python language.

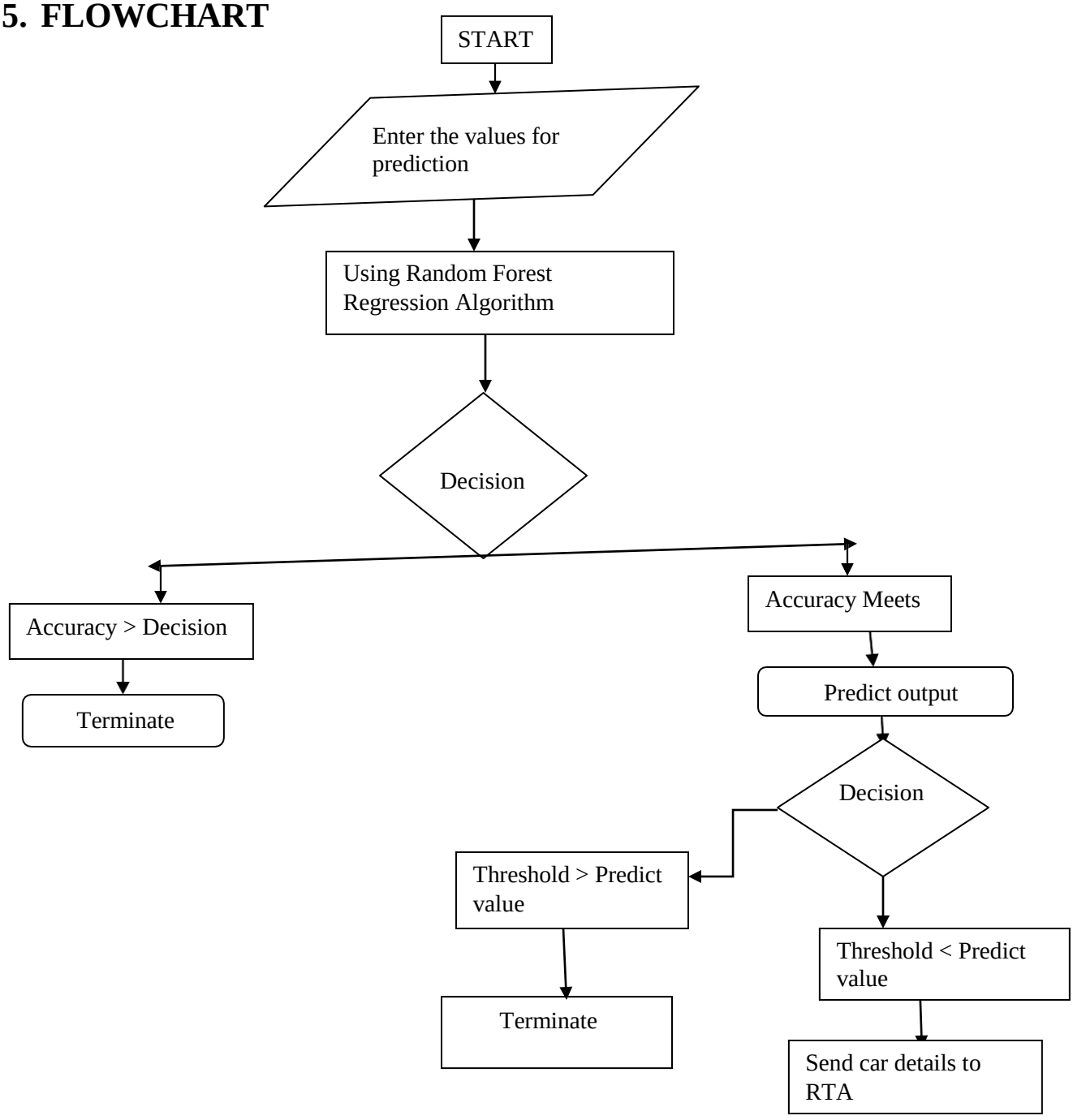
For creating a user interface for the prediction we used the Flask. It is a micro web framework written in Python. It is classified as a micro framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions, and a scripting language to create a webpage is HTML by creating the templates to use in the functions of the Flask and HTML.

4. EXPERIMENTAL INVESTIGATION

In this paper, the dataset we used is derived from https://www.kaggle.com/behrag/co2-emission-forecast-with-python-seasonal-arima/data.com.It contains more than 7385 original data of users with 12 attributes. After that, the missing values are filled in by means of mode interpolation, and the duplicate or meaningless attributes are deleted, finally we have retained to 12 attributes. Those attributes were shown below in the screenshot of the data set we used.

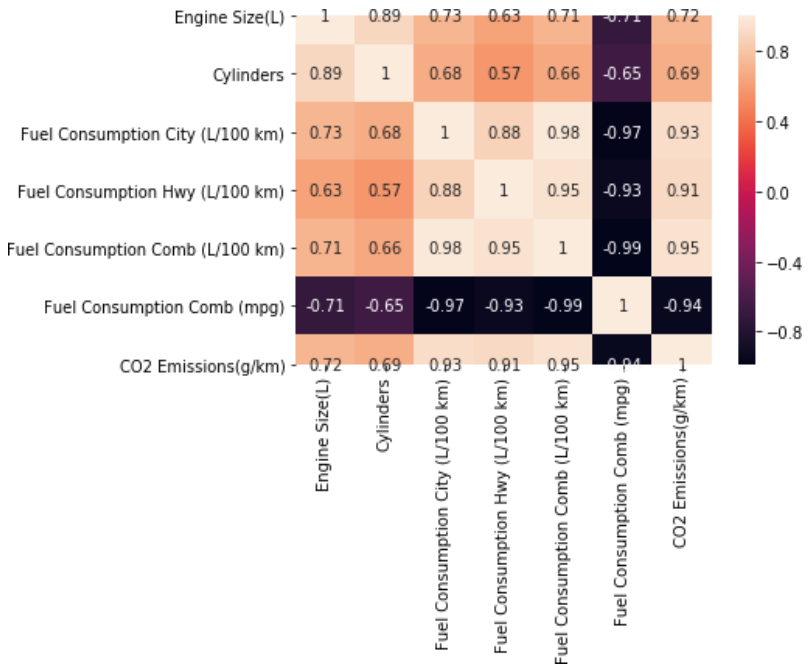
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Make	Model	Vehicle Cl	Engine Siz	Cylinders	Transmiss	Fuel Type	Fuel Cons	Fuel Cons	Fuel Cons	Fuel Cons	CO2 Emissions(g/km)	
2	ACURA	ILX	COMPACT	2	4	AS5	Z	9.9	6.7	8.5	33	196	
3	ACURA	ILX	COMPACT	2.4	4	M6	Z	11.2	7.7	9.6	29	221	
4	ACURA	ILX HYBRII	COMPACT	1.5	4	AV7	Z	6	5.8	5.9	48	136	
5	ACURA	MDX 4WD	SUV - SMA	3.5	6	AS6	Z	12.7	9.1	11.1	25	255	
6	ACURA	RDX AWD	SUV - SMA	3.5	6	AS6	Z	12.1	8.7	10.6	27	244	
7	ACURA	RLX	MID-SIZE	3.5	6	AS6	Z	11.9	7.7	10	28	230	
8	ACURA	TL	MID-SIZE	3.5	6	AS6	Z	11.8	8.1	10.1	28	232	
9	ACURA	TL AWD	MID-SIZE	3.7	6	AS6	Z	12.8	9	11.1	25	255	
10	ACURA	TL AWD	MID-SIZE	3.7	6	M6	Z	13.4	9.5	11.6	24	267	
11	ACURA	TSX	COMPACT	2.4	4	AS5	Z	10.6	7.5	9.2	31	212	
12	ACURA	TSX	COMPACT	2.4	4	M6	Z	11.2	8.1	9.8	29	225	
13	ACURA	TSX	COMPACT	3.5	6	AS5	Z	12.1	8.3	10.4	27	239	
14	ALFA ROM	4C	TWO-SEA	1.8	4	AM6	Z	9.7	6.9	8.4	34	193	
15	ASTON M	DB9	MINICOM	5.9	12	A6	Z	18	12.6	15.6	18	359	
16	ASTON M	RAPIDE	SUBCOMP	5.9	12	A6	Z	18	12.6	15.6	18	359	
17	ASTON M	V8 VANTA	TWO-SEA	4.7	8	AM7	Z	17.4	11.3	14.7	19	338	
18	ASTON M	V8 VANTA	TWO-SEA	4.7	8	M6	Z	18.1	12.2	15.4	18	354	
19	ASTON M	V8 VANTA	TWO-SEA	4.7	8	AM7	Z	17.4	11.3	14.7	19	338	
20	ASTON M	V8 VANTA	TWO-SEA	4.7	8	M6	Z	18.1	12.2	15.4	18	354	
21	ASTON M	VANQUISI	MINICOM	5.9	12	A6	Z	18	12.6	15.6	18	359	
22	AUDI	A4	COMPACT	2	4	AV8	Z	9.9	7.4	8.8	32	202	
23	AUDI	A4 QUATT	COMPACT	2	4	AS8	Z	11.5	8.1	10	28	230	

5. FLOWCHART

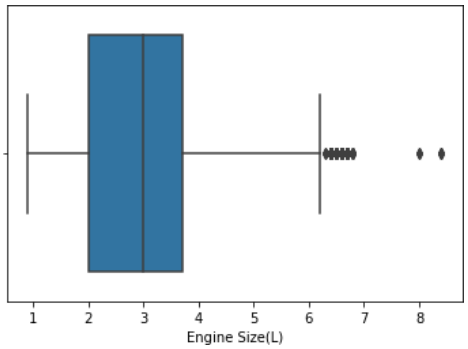

START
Enter the values for prediction
Using Random Forest Regression Algorithm
Decision
Accuracy > Decision
Terminate
Accuracy Meets
Predict output
Decision
Threshold > Predict value
Terminate
Threshold < Predict value
Send car details to RTA

6. RESULT

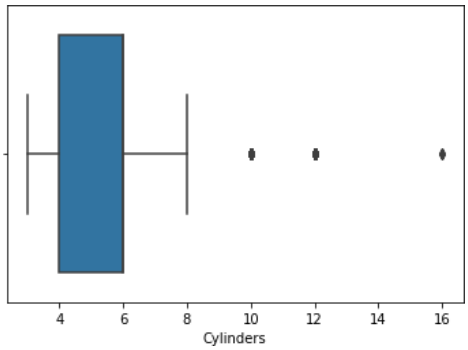
In this paper, the Random Forest algorithm is used to predict the vehicle performance with respect to CO2 emission from cars. The obtained results are displayed in Table below. The results shows that, the performance of vehicles on the basis of emission of CO2 is displayed as well as the accuracy on predicting the defaulters are mentioned in it respectively.

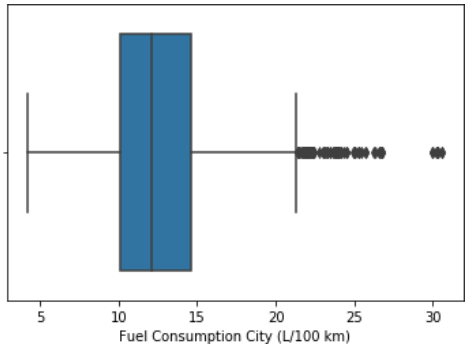


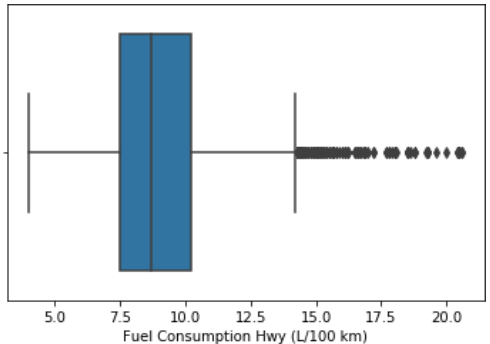
Engine Size(L)



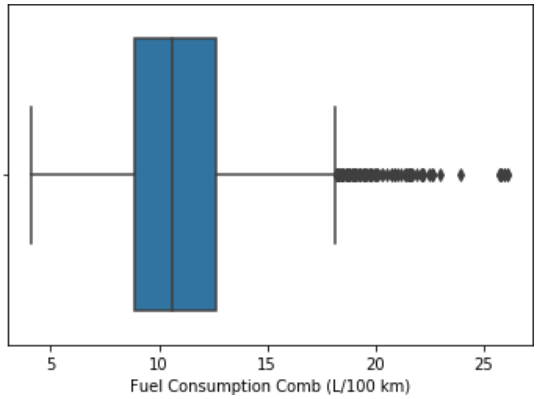
Cylinders



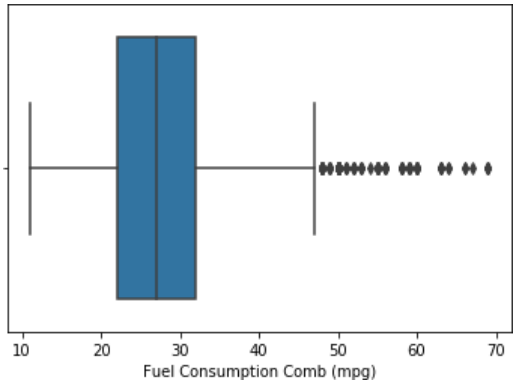




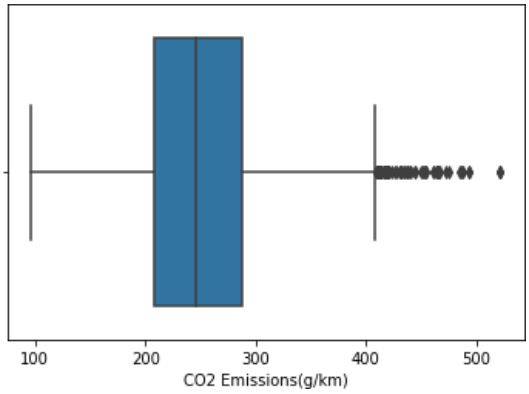
Fuel Consumption Comb (L/100 km)



Fuel Consumption Comb (mpg)



CO2 Emissions(g/km)



Algorithm used	**Accuracy**
Random Forest Regression	**98%**

7. ADVANTAGES AND DISADVANTAGES

Advantages:

- Easy and simple User Interface for the people who is going to evaluate the vehicle's pollution emission's status.
- Random Forest Regression give the accurate result of the prediction up to 98% which is the algorithm we used for prediction.
- It is widely used for managing risks in the transportation sector.
- It is composed using the HTML and Python for the web usage in real time.
- It can work in real time and predict as soon as the necessary details for prediction are given to the model.

Disadvantages:

- Gives 98% accuracy for the prediction status leading to over fitting of data.
- It could not work anywhere like a web-application, if one is using other should be static.
- Needs multiple value for the prediction.

8. APPLICATIONS

- It is widely used for managing risks in the transportation and ecological sector. Government as well as RTA officials need to know the credibility of customers they are dealing with in real time.
- To have an idea of customer relationship cycle such as customer acquisition, increasing value of the customer and customer retention and standards of vehicle hybridization.
- It is one of the most widely used areas of data mining in this sector. The consumer behavior with reference to product, credibility and quality channel can be analyzed by the department.
- Due to tremendous growth in data this industry deals with, analysis and transformation of the data into useful knowledge has become a task beyond human ability.
- So we use Machine Learning Algorithms to analyze the data and propose what transportation sector companies need to achieve their needs.

9. CONCLUSION

In this paper, the Random Forest algorithm is adopted to build a UI model for predicting pollutants default in the lending sector and the results are stored in the database. The experiment shows that the algorithm performs outstanding than the data mining algorithms in the prediction of defaulters and has strong ability of generalization. There is no definitive guide of which algorithms to use given any situation. What may work on some data sets may not necessarily work on others. Therefore, always evaluate methods using cross validation to get a reliable estimates.

10. FUTURE SCOPE

The scope of Machine Learning is not limited to the transport sector. Rather, it is expanding across all fields such as banking and finance, information technology, media & entertainment, gaming, and the automotive industry.

Furthermost, Random forest regression algorithm can be applied on other datasets available for co2 emission to further investigate its accuracy. An analysis of other machine learning algorithms can also be done to know the power of machine learning for predictions. In further studies, we will try to conduct experiments on larger datasets or try to tune the model so as to achieve the state of-art performance of the model and a great UI support system making it complete web application model.

.

11. BIBLIOGRAPHY

- ADAC, "ADAC-Test: Emissionsminderung durch Netzsteuerung", Munich, January 2013
- CE Delft / ECN / TNO, "CO2-reductie door gedragsverandering in de verkeerssector: Een quickscan van het CO2-reductie-potentieel en kosteneffectiviteit van een selectie van maatregelen", 2015
- Huber, T., "Driver Influence & advance Driving Strategy improves CO2 Saving in Real Driving", 2015 (includes description of Continental project GERICO)
- iMobility Forum, Working Group for Clean and Efficient Mobility (WG4CEM), "Identifying the most promising ITS solutions for clean and efficient mobility", November 2013
- Klunder, G.A et al, "Impact of Information and Communication Technologies on Energy Efficiency in Road Transport", TNO report, 2009
- KonSULT, "the Knowledgebase on Sustainable Urban Land use and Transport", 2014, www.konsult.leeds.ac.uk
- Navteq, "Green Streets" (white paper), 2010
- Niu, D. & Sun, J., "Eco-Driving versus Green Wave Speed Guidance for Signalized Highway Traffic: A multi-vehicle driving simulator study". 13th COTA International Conference of Transportation Professionals (CICTP 2013), www.sciencedirect.com/science/article/pii/S1877042813022507
- Pandazis, J-Ch., "ITS for Energy Efficiency", ERTICO Thematic Paper, Brussels, November 2014
- RAC Foundation / Wengraf, Ivo, "Easy on the Gas – the effectiveness of eco-driving", 2012. www.racfoundation.org/research/environment

APPENDIX

HTML:

```
<!DOCTYPE html>
<html >
<!--From https://codepen.io/frytyler/pen/EGdtg-->
<head>
  <meta charset="UTF-8">
  <title>ML API</title>
  <link href='https://fonts.googleapis.com/css?family=Pacifico' rel='stylesheet'
type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Arimo' rel='stylesheet' type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Hind:300' rel='stylesheet'
type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Open+Sans+Condensed:300'
rel='stylesheet' type='text/css'>
<link rel="stylesheet" href="{{ url_for('static', filename='css/stylelb.css') }}">

<style>
.login{
top: 20%;
}
</style>
</head>
<body>
    <div class="login">
    <center>
        <h1 style="color:black; font-size: 200%;">CO2 Emission Prediction</h1>

    <!-- Main Input For Receiving Query to our ML -->
    <form action="{{ url_for('y_predict')}}"method="post">
     <fieldset style="background-color: #008080;width:110%;line-height:150%">

<select name="Make" required="required">
     <option value="">Select Make </option>
     <option value="FORD">FORD</option>
     <option value="ROLLS-ROYCE">ROLLS-ROYCE</option>
     <option value="MERCEDES-BENZ">MERCEDES-BENZ</option>
     <option value="NISSAN">NISSAN</option>
     <option value="AUDI">AUDI</option>
     <option value="KIA">KIA</option>
     <option value="JAGUAR">JAGUAR</option>
```

```
<option value="BUGATTI">BUGATTI</option>
<option value="BMW">BMW</option>
<option value="PORSCHE">PORSCHE</option>
<option value="ACURA">ACURA</option>
</select>

<select name="Model" required="required">
<option value="">Select Model </option>
<option value="FOCUS FFV">FOCUS FFV</option>
<option value="COMPASS">COMPASS</option>
<option value="ATS">ATS</option>
<option value="ILX">ILX</option>
</select>

<select name="Vehicle Class" required="required">
<option value="">Select Vehicle Class </option>
<option value="SUV - SMALL">SUV - SMALL</option>
<option value="MID-SIZE">MID-SIZE</option>
<option value="COMPACT">COMPACT</option>
<option value="SUBCOMPACT">SUBCOMPACT</option>
</select>

<input type="number" name="Engine Size(L)" placeholder="Engine Size(L)"
required="required" />
<input type="number" name="Cylinders" placeholder="Cylinders" required="required"
/>

<select name="Transmission" required="required">
    <option value="">Select Transmission </option>
    <option value="AS6">AS6</option>
    <option value="AS8">AS8</option>
    <option value="M6">M6</option>
  <option value="A6">A6</option>
  <option value="AM7">AM7</option>
  <option value="A9">A9</option>
  <option value="A8">A8</option>
  <option value="AS7">AS7</option>
  <option value="M5">M5</option>
  <option value="AV">AV</option>
  <option value="AS10">AS10</option>
  <option value="AV6">AV6</option>
  <option value="AM6">AM6</option>
```

```
                    <option value="M7">M7</option>
                    <option value="AV7">AV7</option>
                    <option value="AS9">AS9</option>
                    <option value="A5">A5</option>
                    <option value="AM8">AM8</option>
                    <option value="AV8">AV8</option>
                    <option value="A4">A4</option>
                    <option value="A7">A7</option>
                    <option value="AS5">AS5</option>
                    <option value="AV10">AV10</option>
                    <option value="A10">A10</option>
                    <option value="AM9">AM9</option>
                    <option value="AS4">AS4</option>
                </select>

                <select name="Fuel Type" required="required">
                <option value="">Select Fuel Type </option>
                <option value="X">X</option>
                <option value="Z">Z</option>
                <option value="D">D</option>
                <option value="E">E</option>
                <option value="N">N</option>
                </select>

                <input type="number" name="Fuel Consumption City (L/100 km)" placeholder="Fuel Consumption City (L/100 km)" required="required" />
                <input type="number" name="Fuel Consumption Hwy (L/100 km)" placeholder="Fuel Consumption Hwy (L/100 km)" required="required" />
                <input type="number" name="Fuel Consumption Comb (L/100 km)" placeholder="Fuel Consumption Comb (L/100 km)" required="required" />
                <input type="number" name="Fuel Consumption Comb (mpg)" placeholder="Fuel Consumption Comb (mpg)" required="required" />

                    <button type="submit" class="btn btn-primary btn-block btn-large">Predict</button>

                </form>
                <br>
                <br>
                {{ prediction_text }}
            </div>
        </body>
</html>
```

APP.PY:

```
import numpy as np
from flask import Flask, request, jsonify,
render_template from joblib import load
app = Flask(_name_)
model= load('randomforestregressor.save')
trans1=load('transform1')
trans2=load('transform2')
trans3=load('transform3')
trans4=load('transform4')
trans5=load('transform5')
threshold=200


@app.route('/')
def home():
    return render_template('index.html')

@app.route('/y_predict',methods=['POST'])
def y_predict():
    '''
    For rendering results on HTML GUI
    '''
    x_test = [[x for x in request.form.values()]]
    print(x_test)

    test=trans1.transform(x_test)
    test=test[:,1:]

    test=trans2.transform(test)
    test=test[:,1:]

    test=trans3.transform(test)
    test=test[:,1:]

    test=trans4.transform(test)
    test=test[:,1:]

    test=trans5.transform(test)
    test=test[:,1:]

    print(test)
    #test = scaler.transform(test)
    prediction = model.predict(test)
    print(prediction)
    if prediction[0] > threshold:
        output = 'True'
    else:
        output = 'False'
```

```
    return render_template('index.html', prediction_text='CO2 emission for given vehicle = {} threshold = {}\n Vehicle seized- {}'.format(prediction[0],threshold,output))

'"@app.route('/predict_api',methods=['POST'])
def predict_api():

    #For direct API calls trought request

    data = request.get_json(force=True)
    prediction = model.y_predict([np.array(list(data.values()))])

    output = prediction[0]
    return jsonify(output)'"

if__name__== "_main_":
    app.run(debug=True)
```