**CO2 Emission by vehicles**

*A report submitted in partial fulfilment of the requirements for the Award of Degree of*

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

**By**

**Inti Sindhu**

**Regd. No: 20B91A05B3**

**Under Supervision of Mr. Gundala Nagaraju**

**Henotic Technology Pvt Ltd, Hyderabad**

**(Duration: 7th July, 2022 to 6th September, 2022)**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING SAGI RAMA KRISHNAM RAJU ENGINEERING COLLEGE**

(An Autonomous Institution)

Approved by AICTE, NEW DELHI and Affiliated to JNTUK, Kakinada

CHINNA AMIRAM, BHIMAVARAM,

ANDHRA PRADESH

SAGI RAMA KRISHNAM RAJU ENGINEERING COLLEGE

(Autonomous)

Chinna Amiram, Bhimavaram

DEPARTMENT OF

Computer Science and Engineering



**CERTIFICATE**

This is to certify that the “**Summer** **Internship Report”** submitted by **Inti Sindhu, 20B91A05B3** is work done by him/her and submitted during 2021 - 2022 academic year, in partial fulfilment of the requirements for the award of the Summer Internship Program for **Bachelor of Technology in Computer Science and Engineering,** at **Henotic Technology Pvt Ltd, Hyderabad** from *07.07.2022 to 06.09.2022*

|  |  |  |
| --- | --- | --- |
| **Department Internship Coordinator** | **Dean -T & P Cell** | **Head of the Department** |

**Table of Contents**

[**1.0** **Introduction** 7](#_Toc111803085)

[**1.1.** **What are the different types of Machine Learning?** 8](#_Toc111803086)

[**i.** **Supervised learning** 8](#_Toc111803087)

[**ii.** **Unsupervised learning** 8](#_Toc111803088)

[**iii.** **Reinforcement learning** 8](#_Toc111803089)

[**1.2.** **Benefits of Using Machine Learning in Transportation** 9](#_Toc111803090)

[**1.3.** **About Industry (Transportation)** 10](#_Toc111803091)

[**1.3.1** **AI / ML Role in Transportation** 10](#_Toc111803092)

[**2.0** **CO2 Emission by vehicles** 11](#_Toc111803093)

[**2.1.** **Main Drivers for AI CO2 prediction** 11](#_Toc111803094)

[**2.2.** **Internship Project - Data Link** 11](#_Toc111803095)

[**3.0** **AI / ML Modelling and Results** 12](#_Toc111803096)

[**3.1.** **Problem Statement** 12](#_Toc111803097)

[**3.2.** **Data Science Project Life Cycle** 12](#_Toc111803098)

[**3.2.1** **Data Exploratory Analysis** 13](#_Toc111803099)

[**3.2.2** **Data Pre-processing** 13](#_Toc111803100)

[**3.2.2.1.** **Check the Duplicate and low variation data** 13](#_Toc111803101)

[**3.2.2.2.** **Identify and address the missing variables** 13](#_Toc111803102)

[**3.2.2.3.** **Handling of Outliers** 13](#_Toc111803103)

[**3.2.2.4.** **Categorical data and Encoding Techniques** 13](#_Toc111803104)

[**3.2.2.5.** **Feature Scaling** 14](#_Toc111803105)

[**3.2.3** **Selection of Dependent and Independent variables** 14](#_Toc111803106)

[**3.2.4** **Data Sampling Methods** 14](#_Toc111803107)

[**3.2.4.1.** **Stratified sampling** 14](#_Toc111803108)

[**3.2.4.2.** **Simple random sampling** 14](#_Toc111803109)

[**3.2.5** **Models Used for Development** 14](#_Toc111803110)

[**3.2.5.1.** **Model 01** 14](#_Toc111803111)

[**3.2.5.2.** **Model 02** 15](#_Toc111803112)

[**3.2.5.3.** **Model 03** 15](#_Toc111803113)

[**3.2.5.4.** **Model 04** 15](#_Toc111803115)

[**3.2.5.5.** **Model 05** 15](#_Toc111803117)

[**3.2.5.6.** **Model 06** 15](#_Toc111803118)

[**3.2.5.7.** **Model 07** 15](#_Toc111803119)

[**3.2.5.8.** **Model 08** 15](#_Toc111803120)

[**3.2.5.9.** **Model 09** 15](#_Toc111803121)

[**3.2.5.10.** **Model 10** 16](#_Toc111803122)

[**3.3.** **AI / ML Models Analysis and Final Results** 16](#_Toc111803123)

[**3.3.1** **Decision Tree Rgressor Model codes** 16](#_Toc111803124)

[**3.3.2** **Extra Tree Regressor Python Code** 16](#_Toc111803125)

[**3.3.3** **Extra Trees Python code** 16](#_Toc111803126)

[**4.0** **Conclusions and Future work** 18](#_Toc111803127)

[**5.0** **References** 19](#_Toc111803128)

[**6.0** **Appendices** 20](#_Toc111803129)

[**6.1.** **Python code Results** 20](#_Toc111803130)

[**6.2.** **List of Charts** 20](#_Toc111803131)

[**6.2.1** **Chart 01: Count of different fuels** 20](#_Toc111803132)

[**6.2.2** **Chart 02: Analysis of Engine size** 21](#_Toc111803134)

[**6.2.3** **Chart 03: Correlation between Features** 21](#_Toc111803135)

[**6.2.4** **Chart 04: Engine Size vs CO2 Emission** 22](#_Toc111803136)

[**6.2.5** **Chart 06: True values Vs Predicted values** 22](#_Toc111803137)

**Abstract**

Traffic emissions are considered one of the leading causes of environmental impact in megacities and their dangerous effects on human health. This paper presents a hybrid model based on data mining and GIS models designed to predict vehicular Carbon Dioxide (CO2) emitted from traffic on the New Klang Valley Expressway, Malaysia. The hybrid model was developed based on the integration of GIS and the optimized Artificial Neural Network algorithm that combined with the Correlation based Feature Selection (CFS) algorithm to predict the daily vehicular CO2 emissions and generate prediction maps at a microscale level in a small urban area by using a field survey and open source data, which are the main contributions to this paper. The other contribution is related to the case study, which represents the spatial and quantitative variations in the vehicular CO2 emissions between toll plaza areas and road networks.

The proposed hybrid model consists of three steps: the first step is the implementation of the correlation-based Feature Selection model to select the best model’s predictors; the second step is the prediction of vehicular CO2 by using a multilayer perceptron neural network model; and the third step is the creation of micro scale prediction maps. The model was developed using six traffic CO2 predictors: number of vehicles, number of heavy vehicles, number of motorbikes, temperature, wind speed and a digital surface model. The network architecture and its hyperparameters were optimized through a grid search approach. The traffic CO2 concentrations were observed at 15-min intervals on weekends and weekdays, four times per day. The results showed that the developed model had achieved validation accuracy of 80.6 %. Overall, the developed models are found to be promising tools for vehicular CO2 simulations in highly congested areas.

**Keywords:** traffic CO2; traffic CO2 prediction; CO2 emission; GIS; land use/land cover (LULC)

# **Introduction**

With the increasing power of computer technology, companies and institutions can nowadays store large amounts of data at reduced cost. The amount of available data is increasing exponentially and cheap disk storage makes it easy to store data that previously was thrown away. There is a huge amount of information locked up in databases that is potentially important but has not yet been explored. The growing size and complexity of the databases makes it hard to analyses the data manually, so it is important to have automated systems to support the process. Hence there is the need of computational tools able to treat these large amounts of data and extract valuable information.

In this context, Data Mining provides automated systems capable of processing large amounts of data that are already present in databases. Data Mining is used to automatically extract important patterns and trends from databases seeking regularities or patterns that can reveal the structure of the data and answer business problems. DataMining includes learning techniques that fall into the field of Machine learning. The growth of databases in recent years brings data mining at the forefront of new business technologies.

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 tougher all over the world, so improving the fuel economy of cars is strongly desired. From now onwards, it is considered that the 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.

Different types of traffic CO prediction approaches are mentioned in the literature Early methods of traffic CO modelling were based on traditional techniques using data sampling and global position system (GPS) techniques. Several thematic maps and vehicle emission equations are combined to model the traffic emission distribution in a region and produce informative maps that could help in effective decision making Recent methods are mostly based on land-use regression analysis using statistical and soft computing algorithms These statistical and computing techniques allow the input of various traffic and road geometry factors. Almost all these models are designed by using experimental samples; consequently, these models are highly influenced by the traffic flow condition and the measurement style and the geographic locations The main drawback of these models is that they can not be generalized because of the local environment like vehicle model and type and the weather presented an approach of recognizing the road geometric features from positioning information surveyed by collecting vehicle data.

## **What are the different types of Machine Learning?**

There are four types of machine learning algorithms: supervised, unsupervised and reinforcement.

### **Supervised learning**

In supervised learning, the machine is taught by example. The operator provides the machine learning algorithm with a known dataset that includes desired inputs and outputs, and the algorithm must find a method to determine how to arrive at those inputs and outputs. While the operator knows the correct answers to the problem, the algorithm identifies patterns in data, learns from observations and makes predictions

1. **Classification**: In classification tasks, the machine learning program must draw a conclusion from observed values and determine to  
   what category new observations belong.
2. **Regression**: In regression tasks, the machine learning program must estimate – and understand – the relationships among variables. Regression analysis focuses on one dependent variable and a series of other changing variables – making it particularly useful for prediction and forecasting.
3. **Forecasting**: Forecasting is the process of making predictions about the future based on the past and present data, and is commonly used to analyze trends.

### **Unsupervised learning**

Here, the machine learning algorithm studies data to identify patterns. There is no answer key or human operator to provide instruction. Instead, the machine determines the correlations and relationships by analyzing available data. In an unsupervised learning process, the machine learning algorithm is left to interpret large data sets and address that data accordingly.

1. **Clustering**: Clustering involves grouping sets of similar data (based on defined criteria). It’s useful for segmenting data into several groups and performing analysis on each data set to find patterns.
2. **Dimension reduction**: Dimension reduction reduces the number of variables being considered to find the exact information required.

### **Reinforcement learning**

Reinforcement learning focuses on regimented learning processes, where a machine learning algorithm is provided with a set of actions, parameters and end values. By defining the rules, the machine learning algorithm then tries to explore different options and possibilities, monitoring and evaluating each result to determine which one is optimal. Reinforcement learning teaches the machine trial and error. It learns from past experiences and begins to adapt its approach in response to the situation to achieve the best possible result.

## **Benefits of Using Machine Learning in Transportation**

The transport infrastructure like expressways and roads has a significant importance in the development of any country’s economy by linking cities. These infrastructures are rapidly developing due to the changing in the traffic modes, leading to congested roads. Hence, road traffic emissions are increasing, creating many negative impacts on air quality on roadways, intersections and toll roads. Traffic emissions, such as carbon dioxide (CO2), are the primary contributor to overall air pollution from this infrastructure, and the primary source of traffic emissions is vehicular exhausts.

Spatial prediction models are effectively used as a decision-making support tool for prediction and simulation of traffic emissions on road networks There are various negative impacts that can result from inappropriate traffic levels, including high levels of noise and high concentrations of gaseous pollutants Several diseases e.g., cancers, heart diseases, respiratory problems and preterm births, can occur when human beings are exposed to high concentrations of CO2 . The measurement of vehicular emissions on roadways and toll gates may be costly, risky and requires a lot of time and effort. Moreover, the designers do not have the opportunity to determine the vehicular emissions through the design process. In the most recent planning techniques for design of highways and road networks, traffic emission models are often required to support sustainable transportation planning and the reduction of traffic emissions from sources such as congestion and tollgate areas.

Thus, the GIS-based modeling of traffic emission and intra-urban air pollution exposure can be an effective tool in the environmental assessment for sustainable road planning. This tool can distinguish the areas affected by different types of pollutants and the related ecological and social factors. This would be able to determine the best strategy to support the decision makers . On the other hand, GIS can save costs and time in the traffic emission modeling and can therefore be used in sustainable planning. Different types of traffic CO2 prediction approaches are mentioned in the literature Early methods of traffic CO2 modeling were based on traditional techniques using data sampling and global position system (GPS) techniques. Several thematic maps and vehicle emission equations are combined to model the traffic emission distribution in a region and produce informative maps that could help in effective decision making . Recent methods are mostly based on land-use regression analysis using statistical and soft computing algorithms . These statistical and computing techniques allow the input of various traffic and road geometry factors.

Almost all these models are designed by using experimental samples; consequently, these models are highly influenced by the traffic flow condition and the measurement style and the geographic locations . The main drawback of these models is that they can not be generalized because of the local environment like vehicle model and type and the weather presented an approach of recognizing the road geometric features from positioning information surveyed by collecting vehicle data.

## **About Industry (Transportation)**

The transport infrastructure like expressways and roads has a significant importance in the development of any country’s economy by linking cities. These infrastructures are rapidly developing due to the changing in the traffic modes, leading to congested roads. Hence, road traffic emissions are increasing, creating many negative impacts on air quality on roadways, intersections and toll roads. Traffic emissions, such as carbon dioxide (CO2), are the primary contributor to overall air pollution from this infrastructure, and the primary source of traffic emissions is vehicular exhausts.

### **AI / ML Role in Transportation**

Spatial prediction models are effectively used as a decision-making support tool for prediction and simulation of traffic emissions on road networks There are various negative impacts that can result from inappropriate traffic levels, including high levels of noise and high concentrations of gaseous pollutants Several diseases e.g., cancers, heart diseases, respiratory problems and preterm births, can occur when human beings are exposed to high concentrations of CO2 . The measurement of vehicular emissions on roadways and toll gates may be costly, risky and requires a lot of time and effort. Moreover, the designers do not have the opportunity to determine the vehicular emissions through the design process. In the most recent planning techniques for design of highways and road networks, traffic emission models are often required to support sustainable transportation planning and the reduction of traffic emissions from sources such as congestion and tollgate areas.

Thus, the GIS-based modeling of traffic emission and intra-urban air pollution exposure can be an effective tool in the environmental assessment for sustainable road planning. This tool can distinguish the areas affected by different types of pollutants and the related ecological and social factors. This would be able to determine the best strategy to support the decision makers . On the other hand, GIS can save costs and time in the traffic emission modeling and can therefore be used in sustainable planning. Different types of traffic CO2 prediction approaches are mentioned in the literature Early methods of traffic CO2 modeling were based on traditional techniques using data sampling and global position system (GPS) techniques. Several thematic maps and vehicle emission equations are combined to model the traffic emission distribution in a region and produce informative maps that could help in effective decision making . Recent methods are mostly based on land-use regression analysis using statistical and soft computing algorithms . These statistical and computing techniques allow the input of various traffic and road geometry factors.

# **CO2 Emission by vehicles**

With the increasing power of computer technology, companies and institutions can nowadays store large amounts of data at reduced cost. The amount of available data is increasing exponentially and cheap disk storage makes it easy to store data that previously was thrown away. There is a huge amount of information locked up in databases that is potentially important but has not yet been explored. The growing size and complexity of the databases makes it hard to analyses the data manually, so it is important to have automated systems to support the process. Hence there is the need of computational tools able to treat these large amounts of data and extract valuable information.

## **Main Drivers for AI CO2 prediction**

Predictive modelling allows for simultaneous consideration of many variables and quantification of their overall effect. When a large amount of co2 gas emission are analysed, patterns regarding the characteristics of the vehicles that drive begin to emerge.

The following are the main drivers which influencing the gas emission

|  |  |
| --- | --- |
| * Engine size * Exposures * Limits * Leakages * Fuel Type * Mixture info * Prior loss experience * Quality check * Geography based * Land covered * Hill places * Roads | * Other Factors * Weather condition * Traffic time * Fuel quality * Engine manufacture * Life time of vehicle * Vehicle size * Expenses of vehicle * Run time * Quality of fuel * Details from Traffic |

## **Internship Project - Data Link**

The internship project data has taken from Kaggle, and the link is www.kaggle.com/datasets/sindhuinti/co2-emission

# **AI / ML Modelling and Results**

## **Problem Statement**

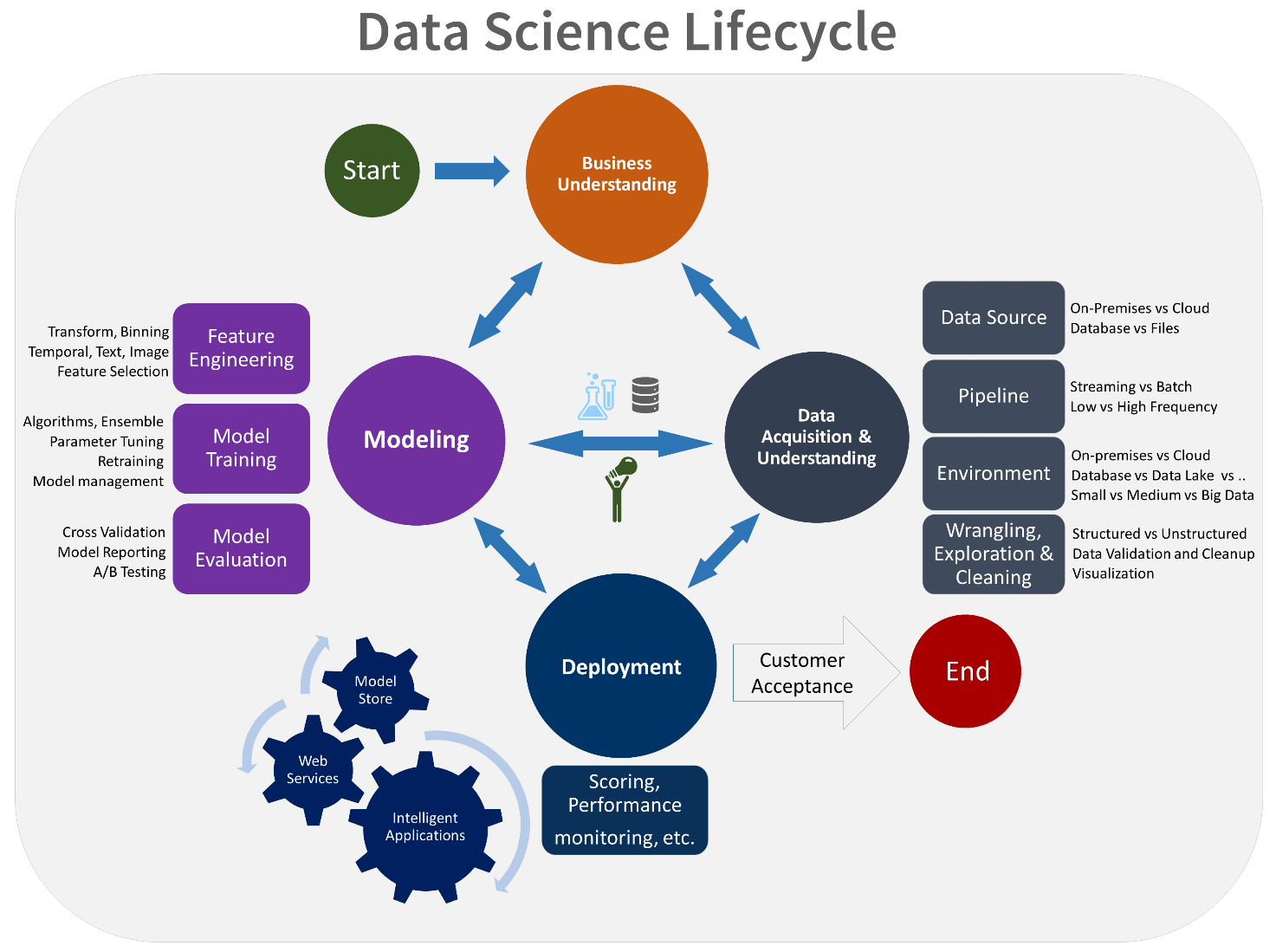
Predictive models are most effective when they are constructed using a well-defined data and research data since this allows the model to recognize the specific nature of a data exposure as well as its patterns. The construction of the model also involves input from the collected data the process, as well as consideration of industry leading research and benchmarks.

Predictive modelling can be used to quantify the impact to the Co2 emission resulting various environmental pollutants which leads to various health diseases.

* Early identification of the amount of co2 emission from vehicles prevents from global warming due to harmful gases.
* Recognition of different vehicles emits more amount of polluting gases into the atmosphere.

## **Data Science Project Life Cycle**

Data Science is a multidisciplinary field of study that combines programming skills, domain expertise and knowledge of statistics and mathematics to extract useful insights and knowledge from data.



### **Data Exploratory Analysis**

Exploratory data analysis has been done on the data to look for relationship and correlation between different variables and to understand how they impact or target variable.

The exploratory analysis is done for Auto Quote / Policy Conversion with different parameters and all the charts are presented in **Appendices 6.2 - List of charts (6.2.1 to 6.2.9)**

### **Data Pre-processing**

We removed variables which does not affect our target variable (CO2 Emissions(g/km)) as they may add noise and also increase our computation time ,we checked the data for anomalous data points and outliers. We did principal component analysis on the data set to filter out unnecessary variables and to select only the important variables which have greater correlation with our target variable.

### **Check the Duplicate and low variation data**

Duplicate values in your data can be a big problem! It can lead to substantial errors and over estimate your results. But finding and removing them from your data is actually a necessary part of data preprocessing. An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The standard deviation is a number that measures how far data values are from their mean

### **Identify and address the missing variables**

**Missing data**, or missing values, occur when you don’t have data stored for certain [variables](https://www.scribbr.com/methodology/types-of-variables/) or participants. Data can go missing due to incomplete data entry, equipment malfunctions, lost files, and many other reasons. In any dataset, there are usually some missing data. In [quantitative research](https://www.scribbr.com/methodology/quantitative-research/), missing values appear as blank cells in your spreadsheet. We can remove them using different imputation techniques like mean, mode and median.

### **Handling of Outliers**

[An outlier is an](https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm)**observation that lies an abnormal distance from other values in a random sample from a population.** There is, of course, a degree of ambiguity. Qualifying a data point as an anomaly leaves it up to the analyst or model to determine what is abnormal—and what to do with such data points.

### **Categorical data and Encoding Techniques**

The performance of a machine learning model not only depends on the model and the hyperparameters but also on how we process and feed different types of variables to the model. Since most machine learning models only accept numerical variables, pre-processing the categorical variables becomes a necessary step. We need to convert these categorical variables to numbers such that the model is able to understand and extract valuable information.

### **Feature Scaling**

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing. It is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

### **Selection of Dependent and Independent variables**

The dependent or target variable here CO2 emission. The target variable tells the amount of carbon dioxide emitted by specific vehicle. The independent variables are selected after doing exploratory data analysis and we used Boruta to select which variables are most affecting our target variable.

### **Data Sampling Methods**

The data we have is highly unbalanced data so we used some sampling methods which are used to balance the target variable so we our model will be developed with good accuracy and precision. We used three Sampling methods

### **Stratified sampling**

Stratified sampling randomly selects data points from majority class so they will be equal to the data points in the minority class. So, after the sampling both the class will have same no of observations.

It can be performed using strata function from the library sampling.

### **Simple random sampling**

Simple random sampling is a sampling technique where a set percentage of the data is selected randomly. It is generally done to reduce bias in the dataset which can occur if data is selected manually without randomizing the dataset.

We used this method to split the dataset into train dataset which contains 80% of the total data and test dataset with the remaining 20% of the data.

### **Models Used for Development**

We built our predictive models by using the following ten algorithms

### **Model 01**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

### **Model 02**

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

### **Model 03**

### Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

### **Model 04**

### An extra-trees regressor. This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

### **Model 05**

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

### **Model 06**

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

### **Model 07**

Gaussian process regression is nonparametric (i.e. not limited by a functional form), so rather than calculating the probability distribution of parameters of a specific function, GPR calculates the probability distribution over all admissible functions that fit the data.

### **Model 08**

The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero.

### **Model 09**

Gradient boosting Regression calculates the difference between the current prediction and the known correct target value. This difference is called residual. After that Gradient boosting Regression trains a weak model that maps features to that residual.

### **Model 10**

SVR uses the same basic idea as Support Vector Machine (SVM), a classification algorithm, but applies it to predict real values rather than a class. SVR acknowledges the presence of non-linearity in the data and provides a proficient prediction model.  It is one among the popular Machine Learning models that can be used in classification problems or assigning classes when the data is not linearly separable

## **AI / ML Models Analysis and Final Results**

We used our train dataset to build the above models and used our test data to check the accuracy and performance of our models.

We used metrices for calculating accuracy score, mean square error, R2 score and adjusted r2 score of our models and compare and select the best model for given dataset of size ~ 10310 vehicle info.

### **Decision Tree Rgressor Model codes**

* The Python code for models with stratified sampling technique as follows:

from sklearn.tree import DecisionTreeRegressor

ModelDT = DecisionTreeRegressor()

# Fit the model and predict

ModelDT.fit(X\_train, y\_train)

y\_pred = ModelDT.predict(X\_test)

### **Extra Tree Regressor Python Code**

* The Python code for models with stratified sampling technique as follows:

from sklearn.ensemble import ExtraTreesRegressor

ModelET = ExtraTreesRegressor()

# Fit the model

ModelET.fit(X\_train, y\_train)

# Prediction

y\_pred = ModelET.predict(X\_test)

### **Extra Trees Python code**

* The Python code for models with stratified sampling technique as follows:

from sklearn.ensemble import RandomForestRegressor

ModelRF = RandomForestRegressor()

# Fit the model

ModelRF.fit(X\_train, y\_train)

# Prediction

y\_pred = ModelRF.predict(X\_test)

**Stratified Sampling**: Stratified random sampling is a method of sampling that involves the division of a population into smaller subgroups known as strata. In stratified random sampling, or stratification, the strata are formed based on members' shared attributes or characteristics, such as income or educational attainment.

**Simple Random Sampling**: Simple random sampling is a type of probability sampling in which the researcher randomly selects a subset of participants from a population. Each member of the population has an equal chance of being selected. Data is then collected from as large a percentage as possible of this random subset.

# **Conclusions and Future work**

The model results in the following order by considering the model accuracy, R2 score and adjusted r2 score.

1. **Extra Trees Regressor** with Stratified and Random Sampling
2. **Decision Trees Regressor** with Simple Random Sampling
3. **Random** with Simple Random Sampling

We recommend model – **Extra Tree Regressor** with Stratified and Random Sampling technique as a best fit for the give n CO2 emission dataset. We considered Extra Tree because it uses bootstrap aggregation which can reduce bias and variance in the data and can leads to good predictions with the dataset.

Table

Description automatically generated

**The future work to calculate the “Other vehicle’s emission” in Transport industry by using regression methods.**

• **Test machine learning on real data**: When ABB has produced a couple of thousands experimental data points it would be a good idea to redo this project to determine the suitability of Machine Learning. Alternatively, another system could be examined, one that already has a lot of data to it.

• **Try out other machine learning platforms:** Because of the current popularity of Machine Learning there are many platforms available. Python, MATLAB and Google’s TensorFlow could all be interesting to explore.

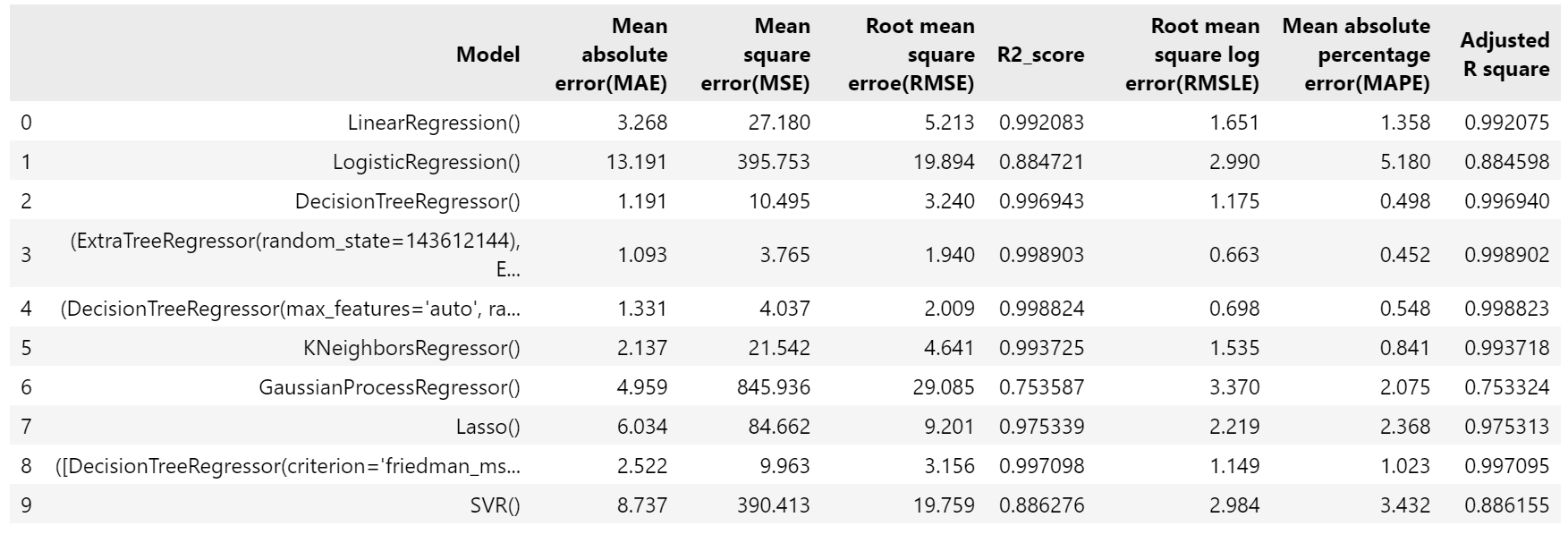
• **Set up sophisticated data-driven experimental structures**: to collect clean and useful data is a challenge, even more if the results have to be very accurate. To use big chunks of data, processes must be set up, to ensure high quality and quantity of data, access to it and the ability to draw conclusions from it. To implement real-time analytics there are software solutions such as Spark, Hadoop or similar software offered by Amazon Web Services and Microsoft Azure

# **References**

* <https://docs.microsoft.com/en-us/azure/architecture/data-science-process/lifecycle>
* [www.kaggle.com/datasets/sindhuinti/co2-emission](http://www.kaggle.com/datasets/sindhuinti/co2-emission)
* [www.geeksforgeeks.com](http://www.geeksforgeeks.com)
* [www.wikipedia.com/co2emission](http://www.wikipedia.com/co2emission)
* [www.scikit-learn.org](http://www.scikit-learn.org)
* [www.seaborn.pydata.org](http://www.seaborn.pydata.org)
* [www.matplotlib.org](http://www.matplotlib.org)
* [www.pandas.pydata.org](http://www.pandas.pydata.org)

# **Appendices**

## **Python code Results**



## **List of Charts**

### **Chart 01: Count of different fuels**

### 

### Chart Description automatically generated**Chart 02: Analysis of Engine size**

### **Chart 03: Correlation between Features**

A picture containing text, clipart

Description automatically generated

### **Chart 04: Engine Size vs CO2 Emission**

### **Chart 06: True values Vs Predicted values**

