Prediction of CO2 Emission for vehicles using Machine Learning Algorithms

Nimisha Patel[1], Sakshi Shah[2], Astha Patel[3], Kareena Matwani[4]

Computer Science and Engineering Department, School of engineering and applied Sciences, Ahmedabad University
Ahmedabad, Gujarat , India - 380015

[1]nimisha.p1@ahduni.edu.in

[2]sakshi.s5@ahdni.edu.in

[3]astha.p@ahdni.edu.in

[4]kareena.m@ahdni.edu.in

Abstract*— *The massive contribution of Co2 in the environment is alarming [1], but there are not many actions taken on the fact of it. We aim to achieve a better psychology of the usage of the predominant factors through which Co2 emissions at large scale are affecting lives.The proper implementation of these highly impacting resources is must, The methods that are applied for the implementation purpose to accurately predict CO2 emissions) are multivariable linear regression, Decision Tree regression, Random forest regression and for classification purpose models like KNN, decision tree and random forest were applied through which a better understanding and approach can be generated for the reduction of the Co2 through many relevant ways to improve the AQI in the states of India as well. The results that we obtained on abundant amount of data (data from kaggle - canadian government data) were high accuracy scores on two algorithms of many implemented which were Random Forest Classifier and Decision Tree Regression, These may help in real world application as they generate higher accuracies with more precision which can help understand the trends of pollution in current time and take eco friendly step towards the same.

***Keywords*— CO2 emissions , transportation, influencing factors, machine Learning , vehicles, automobiles, regression, environment, pollution, classification.**

I. Introduction

The main goal is to predict CO2 emission with highest accuracy and to know which features mainly contribute to CO2 emission.The daunting and harsh exhaustion rates of Co2 increase thoroughly mostly which are reliable on the increase in usage of non renewable resources, the scrutiny and survey of the same is much more important and relevant in current times where the usage of fossil fuels in terms of burning fossil fuels , using it in vehicles or in manufacturing process is escalating. The investigations have been made on which some proposed results were policies in other countries , but mainly in India there are not many certified research papers to tell us about the operation in real terms , the study of the prediction of Co2 emissions thus becomes very important.The scope of machine learning algorithms is thus widely featured here to understand the fluctuating results of Co2 and other gases (CH4,No2 etc.) contributing to the atmosphere through which a more assured and certitude model can be implemented for the further use and reduction of Co2 in the environment through many different ways or application of policies.The CO2 emissions prediction will be helpful while designing and testing in green manufacturing, saves cost of trial to reduce CO2 emission by using different components, The features that mainly contribute to CO2 emissions can be regulated (Feature Importance).CO2 emissions prediction can help regulate laws and rules on the co2 emission like high taxation.

II. Literature survey

The algorithms taken into account in the current times for the prediction of Co2 emissions , The Scope of these type of methods applied will have better implication in our country through sensing of real time data and then supplemented it with these classifiers followed by Machine Learning & AI (Artificial Intelligence) based algorithms , the future scope of these methods is most important in the near future as India being a developing country is now getting bound to new generative environment friendly norms such as FAME policy [Faster adoption & Manufacturing of Hybrid and Electric Vehicle] that is leading to the change for better in the environment[2]. The usage of GPR(Gaussian Process Regression) has been used as a part of research under the prediction of Co2 emissions, with the smallest of RMSE , MSE and MAE the GPR is the best application to their model[3].

III. Implementation

The analyzing and modeling of the dataset was implemented in python using libraries like pandas, numpy, scikit learn, matplotlib, seaborn, pylab and more plotting libraries at the google collaboratory.

A. Data Reading & Handling:

The dataset used for analyzing and modeling the machine learning algorithms was obtained from the open platform kaggle. The dataset has been curated and developed with the help of the Canadian government and it includes the data for light duty vehicles of Canada during 2017-2021 with 7835 rows including 11 different features to predict CO2 emission observed. To get the reliable dataset, the dataset was cleaned by removing null values, duplicate values and outliers. After cleaning, the dataset has 5991 rows with unique and no null values. Features of the vehicle company, Model, Vehicle class, Engine Size, Cylinders, Transissions, Fuel type, Fuel consumption City (L/100 km), Fuel consumption Hwy (L/100 km), Fuel consumption Comb (L/100 km), Fuel consumption Comb (mpg).Outcome will be a real valued CO2 emission value in (g/km) or an class in which the vehicle belong to low emission, permissible emission, medium or high emission.

B. Feature Engineering :

The dataset includes numerical as well as non numerical features columns like Fuel type, transmission , model of vehicle, company and vehicle class which seems relevant in the CO2 emissions prediction, Thus feature engineering methods are used to handle non numeric feature columns with multi categorical values. Firstly, the count for the unique value of each of the non numerical columns is calculated, The column data is transformed to numeric data based on categorical data. The numeric values were assigned using the to_dict() function for transforming data to similar numeric values to similar category data.

C. Exploratory data analysis :

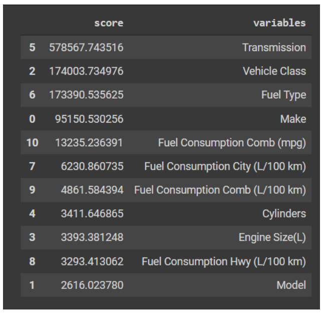
The univariate analysis like data distribution, outliers and frequency distributions were analyzed using bar graphs, scatter plots and histograms respectively. The univariate analysis helped find the trends and patterns that each feature column exhibits. Multivariate analysis helped analyze the variables and their dependency on each other in the dataset specially onto CO2 emissions. As a result, Most of the features have a strong correlation with CO2 emissions while Fuel consumption (mpg) which has a strong negative correlation with CO2 emissions. Model, Company, vehicle class and transmission are not strongly correlated to the CO2 emissions.EDA provided a context needed to develop an appropriate model to achieve the aim of predicting accurate value of CO2 emissions.

D. Feature selection :

As per given dataset, the CO2 emissions have 11 features in the dataset, but it is necessary to know which features affect the maximum CO2 emission rate. Also Along with this, the model complexity increases with the larger number of features. Thus, here are a few steps for feature analysis and feature selection performed to find most important/dependent features for CO2 emissions.

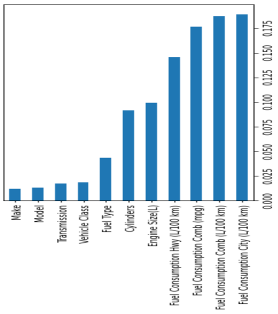
chi2 score : Higher the chi score higher the dependency of variable on to the CO2 emissions.

Feature Importance Score : That can be applied as per the model formulation, which is implemented further for decision tree and random forest models of classifier or regression. Higher the score, more relevant the feature is for the target variable. Implemented using the Tree Classifier model.



	score	variables
5	578567.743516	Transmission
2	174003.734976	Vehicle Class
6	173390.535625	Fuel Type
0	95150.530256	Make
10	13235.236391	Fuel Consumption Comb (mpg)
7	6230.860735	Fuel Consumption City (L/100 km)
9	4861.584394	Fuel Consumption Comb (L/100 km)
4	3411.646865	Cylinders
3	3393.381248	Engine Size(L)
8	3293.413062	Fuel Consumption Hwy (L/100 km)
1	2616.023780	Model

Chi2 score



Feature Importance (Tree classifier)

Correlation : Correlation among the feature variables to know the dependency and the cross validation method helped to recursively evaluate the accuracy and eliminate the features that are not important while prediction.The given data is splitted into 80% train dataset [because it leads to better model fitting] and 20% test dataset as it helps in better model training and in estimation of the generalization performance of the model.

E. Regression :

To better understand the data dependency among features for CO2 emission prediction, Single variable linear regression for each numerical as well as non numerical (converter to numerical data by categorical formulation).

Engine Size(L) : R2 : 0.63, MAE : 23.12, MSE : 918.86.

Cylinders : R2 : 0.58, MAE : 23.54, MSE : 957.99.

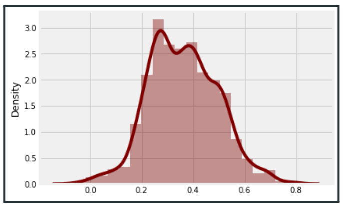
Fuel Type : R2 : -59.90, MAE : 45.13, MSE : 3242.76.

Fuel Consumption City (L/100 km) : R2 : 0.81, MAE : 14.24, MSE : 526.01.

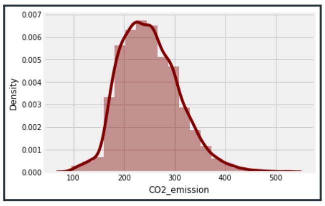
Fuel Consumption Comb (L/100 km) : R2 : 0.81, MAE : 13.86, MSE : 530.78.

Fuel Consumption Comb (mpg) : R2 : 0.77, MAE : 594.57, MSE : 594.57.

Multivariable Linear Regression To better fit the regression line onto the true values in the dataset. Firstly performed the multivariable linear regression on the overall important features variables 'Engine Size(L)','Fuel Consumption Comb (mpg)', 'Cylinders', 'Fuel Type', 'Fuel Consumption Comb (L/100 km)’,'Fuel Consumption Hwy (L/100 km)', 'Fuel Consumption City (L/100 km)', where the R2 score : 0.9256 , RMSE : 0.037. The multivariable linear regression model for all the feature variables, R2 score : 0.9275, RSME : 0.9292.



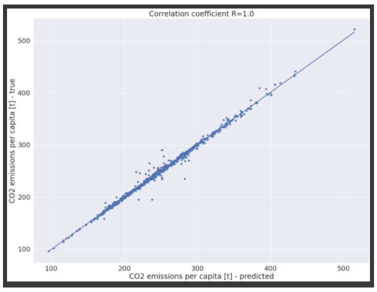
Train dataset



Test data set

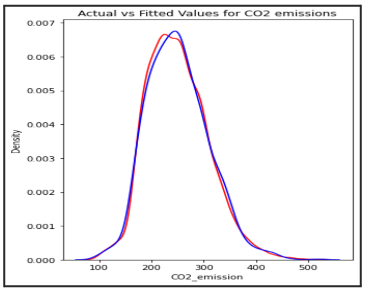
Polynomial regression : polynomial regression line for the important features like Engine Size, Fuel Type and Fuel consumption for degree 2 and degree 3 didn't fit the data and R2 score dropped to range 0.60 - 0.65.

Decision Tree Regression : The decision tree regressor works well to predict the answers with infinite possibility, which here is a real value of CO2 emission in g/km. The decision tree breaks down into smaller and smaller subsets by simple decision rules inferred from training data. Decision Tree Regression R2 score (Test Dataset) : 0.9961 (Including 11 feature variables). Feature selection can supplement the algorithm with lesser complexity and same accuracy by figuring out the most relevant feature. Most important features by feature importance in decision tree regression **Fuel Consumption Comb (L/100 km), Fuel Consumption Comb (mpg), Fuel Type, Fuel Consumption City (L/100 km), Engine Size(L).** Thus, here the model complexity can be decreased and yet maintain the accuracy score 0.9962. To see the effects of hyperparameters we tried tuning some hyperparameters and accuracy remained almost similar. Hyperparameters : The basic idea behind the decision tree is to find an independent point and split data into two parts, so mean squared error is minimized. Hyperparameter like n_estimators =1600, max_depth :20 and min sample split : 2 were tuned into the model.The hyperparameters on tuning exhibit an R2 score between 0.971 to 0.991 among 10 folds.So using hyperparameter, we get the value of **R2= 0.99, Mean Squared Error: MSE=26.58, Root Mean Squared Error: RMSE= 5.16.**



Regression line for decision tree

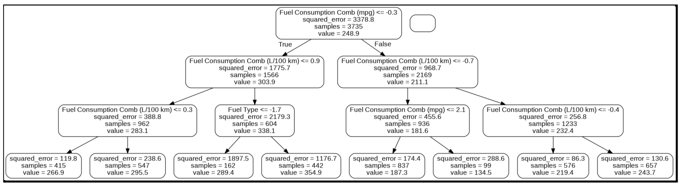
Random forest Regression method :The averaging results of each single decision tree makes a random forest better than a single decision tree and reduces overfitting, thus overall leads to better accuracy. Also Random forest method provides a feature importance method to find the relevant and important features for prediction, which in turn can reduce the model complexity and give the same accuracy.



RMSE: 4.5640 , MAE: 1.8745, R2_score: 0.9939, Accuracy = 99.21%

The random forest regressor importance score suggests the most important 5 features : **Fuel Consumption Comb (L/100 km), Fuel Consumption Comb (mpg), Fuel Type, Fuel Consumption City (L/100 km), Engine Size(L)** as a feature vector and CO2 emission as label vector. Model Analysis : **RMSE: 4.5461, MAE: 1.9547, R2_score: 0.9940**

Using available inbuilt algorithms in python, we found the hyperparameter for random tree forest. Hyperparameters further were tuned to increase the accuracy of the model. Here we are fining the features like: n_estimators - number of decision trees::**1600,** max_features - number of features to consider at every split : **sqrt,** max_depth - maximum number of levels in a tree : **20,** min_samples_split - minimum number of samples required to split a node : **2,** min_samples_leaf - minimum number of samples required at each leaf node : **1** The hyperparameters on tuning exhibit an R2 score between 0.992 to 0.998 among 10 folds. **R2 = 0.9947, Mean Squared Error: MSE = 18.0808, Root Mean Squared Error: RMSE = 4.2522.**



F. Classification:

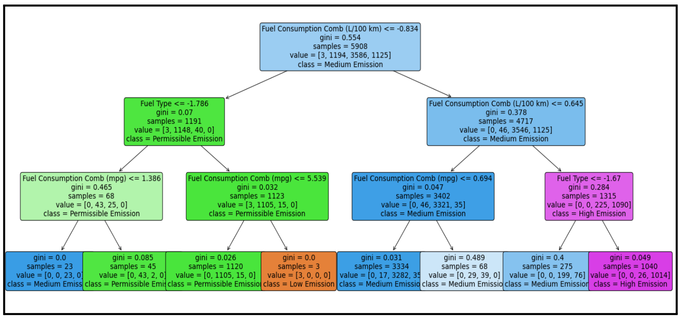
The main aim of classification is to classify the given light duty vehicle into the classes - low emission range(≤ 100), permissible emission(100-200 g/km), medium emission range (200 -300 g/km) and high emission - not permissible range(≥ 300 g/km).With the model set to predict CO2 Emissions from all the given features, we achieved the highest testing accuracy (0.99 percent) using the DecisionTreeClassifier.

Decision Tree Classifier Accuracy on dataset (Test Dataset) : 0.99

Decision Tree Regression R2 score (Test Dataset) : 0.94

Confusion matrix :

```
[[  2   0   0   0]
 [  0 291   3   0]
 [  0  12 873   4]
 [  0   0  12 280]]
```



Decision Tree classification

Random forest algorithm provides higher accuracy by cross validation for larger dataset with easy interpretability. random forest classification on the dataset including non categorical feature values,The accuracy of the model turned out to be as follows :

Accuracy of Random Forest classifier on training set: 1.00

Accuracy of Random classifier on test set: 0.98.

Confusion matrix :

```
[[  2   0   0   0]
 [  0 290   4   0]
 [  0   9 875   5]
 [  0   0   8 284]]
```

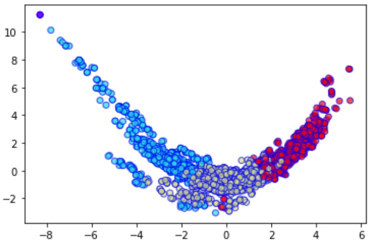
The class1 true predictions are less due to the smaller frequency of the particular class label. The **cross validation for feature selection** using random forest classifier does not eliminate any feature based on its importance and considers each feature equally, thus the number of features input cannot be reduced with feature importance by random classifier.

KNN : The accuracy declines and the misclassification rate in test set as well as train set increases (with the increase in K) . The data tends to underfit while the K value increases. Increasing the value of K, increases the training error (due to over smoothing).For the optimal value of K=1, the accuracy of the test dataset is the largest (0.97) and the generalization gap (train error - test error) is minimum. The true positive for class 1 has smaller value due less data points in the particular class. The confusion matrix helps us analyze how the prediction holds to the true values.

Confusion matrix :

```
[[  2   0   0   0]
 [  1 275  18   0]
 [  0  17 856  16]
 [  0   0  27 265]]
```

LDA & PCA : Linear Discriminant analysis is used for dimensionality reduction in supervised learning; The aim here is to reduce the 11 feature columns to 1 or 2 columns transformed into new space which can effectively segregate the 4 classes with no overlapping among them. Also tuned hyper parameters like solver : svd (inbuild) and the shrinkage value 0.1 which were not so efficient to increase model accuracy.



4 class - random forest classifier after Linear Discriminant analysis

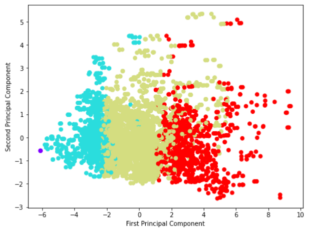
Accuracy of Random Forest classifier on training set: 0.91

Accuracy of Random classifier on test set: 0.90

Confusion matrix :

```
[[  0   2   0   0]
 [  0 266  22   6]
 [  0  20 814  55]
 [  0   0  41 251]]
```

Due to LDA dimensionality reduction the accuracy of random forest classifiers decreased. Similarly, performed PCA to transform data into new space with reduced dimensionality.



PCA components

Accuracy of Random Forest classifier on training set: 0.79. Accuracy of Random classifier on test set: 0.77. The reduced dimensions by PCA are not optimal for classification through random forest.

IV. Results

Regression : Accuracy achieved - 0.9961 for Decision Tree Regression, with 5 feature variable input the model predicts real value of CO2 Emission in g/km.

RMSE - train 0.95	RMSE - test 3.63	MAE - train 0.32	MAE - test 1.75	R2 - train 0.9997	R2 - test 0.9961

Classification : Accuracy achieved 0.98 for random forest classifier, input the feature variables and output will be the class label in which the vehicle belongs.

Accuracy - train set - 0.999	**Accuracy - test set - 0.98**

Confusion matrix :

```
[[  2   0   0   0]
 [  0 292   2   0]
 [  0  13 869   7]
 [  0   0  11 281]]
```

```
        precision    recall  f1-score   support
   0       1.00       1.00      1.00         2
   1       0.96       0.99      0.97       294
   2       0.99       0.98      0.98       889
   3       0.98       0.96      0.97       292

    accuracy                        0.98      1477
   macro avg       0.98      0.98   0.98      1477
weighted avg       0.98      0.98   0.98      1477
```

Features highly responsible for CO2 emissions (dependent) : After evaluating feature engineering methong, feature analysis, feature selection and feature importance scores for different models the top features that contribute to CO2 emissions largely and are strongly depend are :

Engine Size(L),Fuel Type,Fuel Consumption City (L/100 km), Fuel Consumption Comb (L/100 km), Fuel Consumption Comb (mpg).

While performing feature selection and cross validation method for feature importance, for random forest classifier , decision tree classifier, decision tree regression and knn the importance score of these 5 features remained higher and while modeling models with 5 features, accuracy was maintained.

V. CONCLUSION

The decision tree and random forest method works best in predicting the values among other algorithms applied and also for classification, and the higher accuracy model predictions will help in future implementation as follows :

Implementation approach that can be useful in the future

The main five features contributing to CO2 emissions can be regulated to lower the CO2 emissions due to transportation and vehicles. The manufacturers should aim for green manufacturing, where each new model developed would be responsible for least CO2 emissions. And the practical implementation like one point predictions can help them analyze the changes in designing and manufacturing. This implementation can be helpful to government and transportation authorities in order to impose taxation, duty fair, road tax and other policies on the basis of CO2 emissions of the vehicle to regulate the use of vehicles with lesser CO2 emissions

Acknowledgment

We **(Group 7- Discover Decipher)** express our special thanks to Professor **Mehul Raval** and **Teaching Assistants Vaishwi Patel & ArpitKumar Patel** for supporting us throughout the journey of the **CSE523 - Machine Learning** Course Project, which helped us gain practical knowledge that can be applied to the real time world.

Link to Github Repository

https://github.com/sakshi-shah14/CSE523-Machine-Learning-2022-Discover-Decipher.git

References

[1] F. Perera, “Pollution from fossil-fuel combustion is the leading environmental threat to Global Pediatric Health and Equity: Solutions Exist,” *International journal of environmental research and public health*, 23-Dec-2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5800116/ . [Accessed: 14-Apr-2022].

[2] “Faster adoption and manufacturing of (Hybrid &) Electric ...” [Online]. Available: https://heavyindustries.gov.in/UserView/index?mid=2418. [Accessed: 04-Mar-2022].

[3] N. Ma, W. Y. Shum, T. Han, and F. Lai, “Can machine learning be applied to carbon emissions analysis: An application to the CO2 emissions analysis using gaussian process regression,” *Frontiers*, 01-Jan-1AD. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fenrg.2021.756311/full. [Accessed:05 -Apr-2022].

[4] D. Podder, “CO2 emission by vehicles,” *Kaggle*, 05-Aug-2020. [Online]. Available: https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-veh

icles. [Accessed: 22-Apr-2022].

[5] J Mater Environ Sci.com, Journal of Materials and Environmental Science. [Online]. Available: https://www.jmaterenvironsci.com/Journal/vol11-2.html. [Accessed: 22-Apr-2022].

[6] Iea, "Tracking Transport 2020 – analysis," *IEA*, 01-Jun-2020. [Online]. Available: https://www.iea.org/reports/tracking-transport-2020. [Accessed: 18-Mar-2022].

[7] H. Ritchie and M. Roser, "Emissions by sector," *Our World in Data*, 11-May-2020. [Online]. Available: https://ourworldindata.org/emissions-by-sector#:~:text=The%20global%20breakdown%20for%20CO,transport%2C%20and%20manufacturing%20and%20construction. [Accessed: 15-Mar-2022].

[8] "Fuel consumption ratings," *Open Government Portal*. [Online]. Available: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6. [Accessed: 27-Mar-2022].

[9] R. Lee, "Are automotive suppliers ready for the move to electrification?," *LinkedIn*, 06-Apr-2021. [Online]. Available: https://www.linkedin.com/pulse/automotive-suppliers-ready-move-electrification-ron-lee. [Accessed: 16-Mar-2022].

[10] "Can machine learning be applied to carbon emissions ..." [Online]. Available: https://www.researchgate.net/publication/354838594_Can_Machine_Learning_be_Applied_to_Carbon_Emissions_Analysis_An_Application_to_the_CO2_Emissions_Analysis_Using_Gaussian_Process_Regression/. [Accessed: 02-Mar-2022].

[11] "Structural breaks in carbon emissions: A machine ... - imf.org." [Online]. Available: https://www.imf.org/-/media/Files/Publications/WP/2022/English/wpiea2022009-print-pdf.ashx. [Accessed: 10-Mar-2022].

[12] "Fuel consumption ratings," *Open Government Portal*. [Online]. Available: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6. [Accessed: 10-Mar-2022].