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 current global CO2 emissions are bizarre, and transportation largely contributes to the CO2 emissions. Our objective is to predict and monitor the CO2 emissions based on the fuel consumption, engine size, number of cylinders used, fuel type, model of vehicle, company of vehicle and vehicle class. In the following report, the regression methods (single variable linear regression, multivariable linear regression and polynomial regression) is applied to predict CO2 emission among the proper independent variables identified using feature selection and exploratory data analysis from a preprocessed and reliable dataset. The performance, efficiency and reliability of the regression models were evaluated using R2 score, Mean Squared error (MSE), Mean Absolute error (MAE) and root mean squared error (RMSE) and the regression methods showed utmost 0.9 R2 score for multivariable linear regression. The vehicles are also categorized on the basis of CO2 emission range, which is necessary for practical implementation like green manufacturing and deciding road tax for the vehicle which is done as per the range of CO2 emissions. Thus, the vehicles on the given specification were also classified into low, permissible, medium and high emission rate classes using different classification models. The results were analyzed using a confusion matrix which showed KNN to be the most accurate for categorizing with k=1 and accuracy = 0.977 ≃ 0.98.

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

I. Introduction

The daily CO2 exhaustion rates are increasing exponentially due to increasing population and increased reliance on fossil fuels resulting in the extreme overheating of the environment. Transportation has highest dependency on fossil fuels and is responsible for 24% global CO2 emissions yet due to less accuracies and developing phase of renewable energy fuels and electricity operated vehicles[1].Transportation is responsible for 24% of direct CO2 emissions due to fuel consumption[2] being the 2nd largest reason for global CO2 emissions.[3] Carbon footprint, a measure of carbon dioxide released in particular duration due to particular activity, is a widely used term and concept for regulating threats of global climate change. Along with this, recently many efforts have been put on the relevant solutions for CO2 emissions. In this effort, there is a need for certain effective regulating steps like green manufacturing and regulation of road taxation based on CO2 emissions that is sustainable, economically friendly and helps eliminate the risks associated with the environment. There are many traditional approaches being used and researched upon to predict the CO2 emissions. In this paper, the scope of machine learning algorithms in the prediction of CO2 emissions and categorizing particular vehicles into the class of low, permissible, medium and high emission rate class is discussed.

II. Literature Survey

There are many different processes implemented as a part of improvement in the nature and reduction in Co2 in different parts of the world , The usage of GPR(Gaussian Process Regression) has been used as a part of research under the prediction of Co2 emissions, in the same application different factors adding to Co2 are applied such as deforestation , pollution and their relation with the inputs and thus methods were applied, with the smallest of RMSE , MSE and MAE the GPR is the best application to their model[12]. The other methods that were used for estimating Co2 emissions from electrical generation, the linear models and SVM were applied for predicting the LCOE(levelized cost of energy) value , with these parameters different association algorithms with the neural networks were applied[13].The detection as well as sensing of such data through its affecting factors would help in detecting the pattern of the same and thus machine learning algorithms such as SVM , Decision tree and Random forest are most apt for the solving of the problem.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[14].

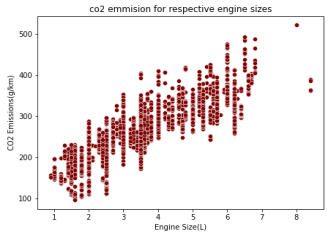
III. Implementation & Results

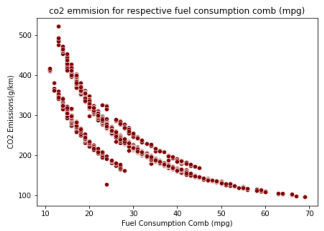
A. Data Processing :

The CO2 emissions dataset was gathered from kaggle (csv file) which was publicly available and formulated with the help of the Canadian Government official website. The dataset included 7385 rows [4]light duty vehicles and 12 columns [features] which was observed for 2017-21 in Canada. It included model specific fuel consumption, engine size, cylinders and estimated CO2 emissions for the respective models of the vehicles. To get a reliable dataset, we searched for the null values, typos, missing values, not applicable values and outliers which were not observed in our dataset. Further, there were 1103 duplicate values found in the dataset which were dropped and it led us to 5991 unique entries to model algorithms and determine the impacts of different features on CO2 emission.

B. Exploratory (Statistical) data analysis on dataset

In order to get a better understanding regarding the data distribution and dependency among the dataset, we performed statistical and exploratory data analysis on the obtained cleaned and reliable dataset like correlation coefficient, univariate and bivariate analysis by scatter plots using seaborn and matplotlib libraries. CO2 emission is strongly correlated (positive) to Fuel consumption city and comb(L/km).





Positive correlation (engine size) **Negative correlation (fuel consumption comb(mpg))**

Engine size, fuel consumption city, fuel consumption hwy, number of cylinders have strong (positive) correlation to CO2 emission and were positively and strongly correlated with each other which helped us determine the feature variables for multivariable linear regression.

C. Feature selection

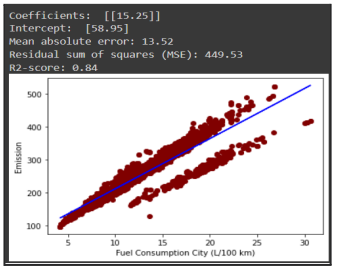
The chi square test can be used to select the dependent (more) features from the many in the dataset for better model performance. Higher chi square value indicates the dependence on the response variable.Observed from the bar graph and chi test score are ordered on the basis of CO2 emissions dependence on these variables as fuel consumption comb(mpg), fuel consumption city(L/100km), fuel consumption comb(L/100km) ,cylinders and engine size

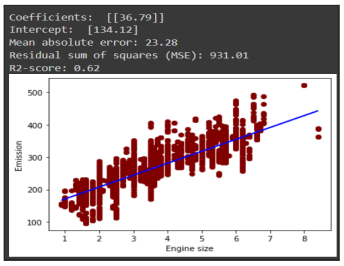
D. Data Splitting

The given data is splitted into 80% train dataset [because it leads to better model fitting] and 20% test dataset as it allows us to find hyperparameters and estimate the generalization performance.

E. Supervised Learning Techniques

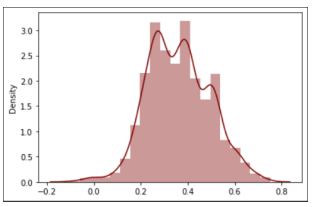
Single Variable Linear Regression: Considering the dataset, from 12 different feature variables we considered highly dependent and correlated variables obtained by feature selection for better determination of CO2 Emission so we get best fitting lines among the data points and accurate determination of CO2 emissions.



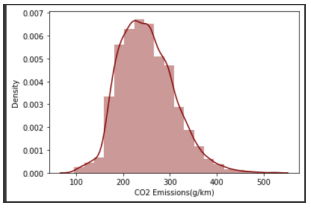


Over here, the value of A= 15.25 which describes that the model predicts the response 15.25 when the train parameter X is 0. The output (that is CO2 emission) increases by 58.95 (intercept - B) when X (engine size) is increased by 1. Linear regression line fits up to 84% .

Multiplevariable Linear Regression: Further the features with high R2 score from linear regression and high chi2 score are taken into consideration for the determination of accurate CO2 emissions. Linear regression for features : Engine Size(L) , Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), Fuel Consumption Comb (L/100 km), Fuel Consumption Comb (mpg) Cylinders to the output CO2 Emissions(g/km). The multiple regression fits the line 90% for three independent variables with RSME (test dataset) = 0.0424.



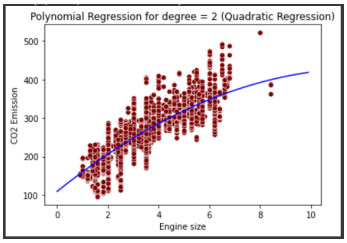
Difference in predicted and test values



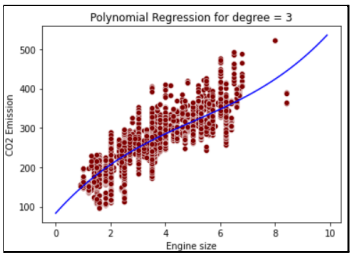
Difference in predicted and train values

Quadratic Regression: Quadratic regression can be achieved by plotting the graph for polynomial regression degree=2 for which we have considered the Engine Size (L) as independent variable and CO2 emissions as dependent variable. Over here, the value of θ1=53.33, the value of θ2 = -2.23 and the value of b=108.69.The Quadratic regression gives 63% model accuracy. The MSE is 902.34 and MAE is 23.03.

Polynomial Regression: To reach a better accuracy for a regression model we tried if a polynomial regression can work on the give datasets.For our dataset , we have performed the polynomial regression with degree = 3 for independent variable - engine size and dependent variable - CO2 emissions.The polynomial regression with degree =3 gives 62% accuracy with MAE = 22.51 and MSE = 925.76, having the value of θ1=77.93, θ2 = -9.34, t θ3 = 0.61 and b=83.22. It can be interpreted that quadratic regression gives better accuracy than polynomial regression with degree=3.The scatter plot for variable feature engine size (L) shows linear strong positive correlation with the dependent variable CO2 emissions and thus, with the increasing degree of polynomial, error rate seems to be increased and R2 score decreased.



Plot for Degree = 2



Plot for Degree = 3

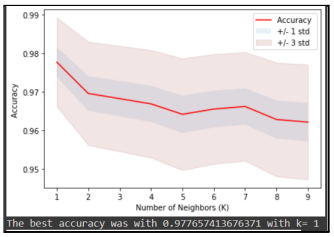
Decision Tree Regression: It is used to solve regression as well as classification models. It is used because it is simple to understand and non-linearity, high dimensionality does not affect the model performance. The decision tree regressor model gives good accuracy [95%] having MSE score 0.15 and will be implemented deeper further.

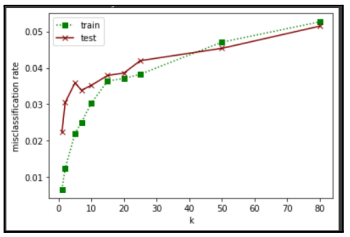
Classification based: The accuracy from linear regression, multivariate linear regression and polynomial regression was around 0.9 utmost for different combinations of dependent and correlated features, which shows non accurate prediction of CO2 emission value by regression. Thus, We would like to predict and analyze the likelihood that the data would fall into the predetermined categories. The data obtained from source have a real integer value not the categorical classification, thus, firstly the classes (lower-upper limit) were created in a way that is equal and balanced distribution among each category class where Co2 emission value ranged from 96g/Km to 522 g/km. The global standard also suggests upto 100 g/km emission as a low emission and above 300 as very high CO2 emission rate[4] . Thus the classes taken into consideration are :

Class 0 : <= 100 , frequency : 5	Low Emission Range
Class 1 : 100 - 200 frequency 1488	Permissible Emission Range
Class 2 : 200-300 frequency 4475	Medium Emission Range
Class 3 : >=300 frequency 1417	High Emission Range (Not permissible)

K-Nearest Neighbor classifier : Firsty, for implementation of KNN, we selected the correct value of the k analysing test- train accuracy and error rate.For k= 2 - Train set Accuracy: 0.98, Test set Accuracy: 0.976 = 0.98.For k= 5 - Train set Accuracy: 0.98, Test set Accuracy: 0. 967 = 0.97 . For k= 10 - Train set Accuracy: 0.97, Test set Accuracy: 0.96.

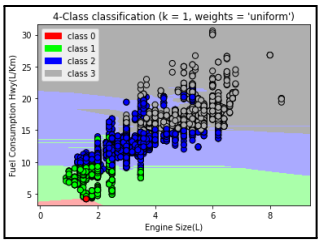
Accuracy vs K plot **misclassification rate vs K plot**



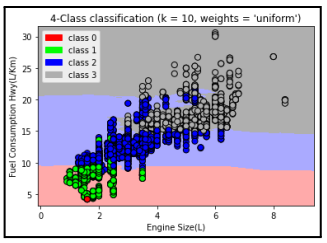


The best accuracy about 0.98 at k =1.The test error and train error is minimum at k =1, and the k value at which test error is minimum and the k value at which difference between test and train error is smaller, generalisation gap is smaller) leads to better accuracy while predicting on unseen data.

Boundary Plot for different value of K



k=1



k=10, underfitting

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). The KNN doesn't work well for high dimensions if the model is trained using more features due to the curse of dimensionality. Confusion Matrix is used to determine the

performance of the classifier function, the accuracy score, recall and f1 score that are class level and general performance measuring matrices are >0.95 which means that the model has good accuracy in 4 class classification.

```
k= 1
Train set Accuracy:  0.9933987813134733
Test set Accuracy:  0.977657413676371
[[287   0]
 [  1   0]]
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         1
           1       0.97      0.98      0.97       294
           2       0.99      0.98      0.98       881
           3       0.96      0.99      0.97       301

    accuracy                           0.98      1477
   macro avg       0.73      0.73      0.73      1477
weighted avg       0.98      0.98      0.98      1477
```

IV. Conclusions

At the culmination , we have seen that Multi variable Linear Regression gave the best accuracy among the various regression models which is 0.89 , thus after the decision of iterating the process with classification was made , the training set was classified by KNN classification model which gave accuracy of 0.98, The other classification models as well as regression model like decision tree are taken into consideration (at present) while we build our model to be a more precise model.

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.

IMPLEMENTATION USING GOOGLE COLLABORATORY

Group7_DiscoverDecipher_Mid_Sem_ProjectImplementation.ipynb

References

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

[2] 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].

[3] "(PDF) developing Machine Learning models to predict CO2 ..." [Online]. Available: https://www.researchgate.net/publication/352144108_Developing_machine_learning_models_to_predict_CO2_trapping_performance_in_deep_saline_aquifers. [Accessed: 10-Mar-2022].

[4] "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].

[5] 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].

[6] "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].

[7] "Towards the systematic reporting of the energy and carbon ..." [Online]. Available: https://jmlr.org/papers/volume21/20-312/20-312.pdf. [Accessed: 15-Mar-2022].

[8] "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].

[9] "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].

[10] C. Saleh1, N. R. Dzakiyullah2, and J. B. Nugroho1, "IOPscience," *IOP Conference Series: Materials Science and Engineering*, 01-Feb-2016. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/114/1/012148. [Accessed: 10-Mar-2022].

[11] C. Zhu and D. Gao, "A research on the factors influencing carbon emission of transportation industry in 'The belt and road initiative' countries based on panel data," *MDPI*, 22-Jun-2019. [Online]. Available: https://www.mdpi.com/1996-1073/12/12/2405. [Accessed: 18-Mar-2022].

[12] 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: 25-Mar-2022].

[13] M. Rao, "Machine learning in estimating CO2 Emissions from electricity generation," *IntechOpen*, 12-Apr-2021. [Online]. Available: https://www.intechopen.com/online-first/76238. [Accessed: 27-Mar-2022].

[14] "Faster adoption and manufacturing of (Hybrid &) Electric ..." [Online]. Available: https://heavyindustries.gov.in/UserView/index?mid=2418. [Accessed: 27-Mar-2022].