IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

# Prediction of Greenhouse Gas Emission in Cars using Machine Learning

**[1]Mr. Amit A Bhalerao, [2]Dr. Shantakumar B Patil**
[1]Student, [2]Head of Department
[1]Computer Science and Engineering,
[1]Sai Vidya Institute of Technology, Bangalore, India

***Abstract:*** Automobile industry is one of the biggest sources of emission of a major Greenhouse Gas i.e., CO2 (Carbon-Dioxide). Unless transport emissions are monitored and brought under control, national and international climate goals will be missed. To meet commitments, we need to track emissions from automobiles and build technologies that would help us to decarbonize them effectively. We need every tool to tackle CO2 emissions from automobiles and early prediction of such emissions using statistical data can help people across the globe in aiding transformative changes that might end up delivering requisite huge cuts in emission. The project aims at predicting CO2 emission levels by analyzing dataset containing official record of statistical data from various car makers. The concept of Regression under Machine Learning is implemented to predict the emission rate and a final study of overall analysis is carried out to determine the best means of predicting rate(s) of emission.

***IndexTerms*** **- Greenhouse Gas Emissions, Automobile Industry, Machine Learning, Regression**

## I. INTRODUCTION

Greenhouse gas emissions results in the phenomenon of Global Warming, which is the rise in the average of Earth's surface temperature. This has happened due to some physical factors like emission of carbon-dioxide, nitrous oxide, methane, etc and is increasing simultaneously on a daily basis. It was observed over the past century that the surface temperature of Earth has been increasing due to increase in the greenhouse gas emission. It is estimated that the average temperature will be increasing by six degree Celsius in the next 200 years and this will be slowly leading to devastating effects across the globe soon. Global warming occurs when the greenhouse gases absorb sunlight and prevents the radiation from escaping back into the atmosphere, thus leading to the problem of average increase in the Earth's surface temperature. The outcomes of these happenings is lethal and has resulted in melting of glaciers on a daily basis, increasing sea levels across globe, flooding along coastal regions and many more to add. The likely impact of climate change on the monthly mean maximum and minimum temperature in Chaliyar river basin using Artificial Neural Networks has been described by NR Chithra [10]. Umair Shahzad describes effects of global warming in his paper titled 'Global Warming: Causes, Effects and Solutions' [16]. H Hassani explains climatic change through his paper titled 'Big Data and Climate Change' [17]. M. W. Gardner explains ANN through his paper 'Artificial Neural Networks (The Multilayer Perception) - A review of applications in the atmospheric sciences' [21]. The prediction of air quality has been explained by A Russo through his paper titled 'Air quality prediction using optimal neural networks with stochastic variables' [23]. Basic information pertaining to temperature, global warming, etc by Y. Song is found in 'The forecasting research of early warning systems for atmospheric pollutants: A case in Yangtze River Delta region' [24]. To get relief from the destructive results of global warming, many people are trying their best to reduce the emission of greenhouse gases. In addition, people are also trying to create awareness regarding harmful effects due to excess emission of greenhouse gases. Complexities have arisen due to several issues that have highlighted over the past few years. Mitigating the risks evolving due to increasing levels of emission will play a great role in shaping our future and defining the society in which we are living. Any single step taken towards reducing emissions can be of great help in turning this world to become emission free and a beautiful place to live.

## II. LITERATURE REVIEW

The issues pertaining to global warming have been referred to by several authors over the past few years. In this project, we are trying to predict the emission levels through means of Regression. 'A new approach for simulating and forecasting the rainfall run-off process within the next two month' by M.J Alizadeh, M.R. Kavianpour, Ozgur Kisi and Vahid Nourani [1] deals with prediction of rainfall. In this, ANN and SVR techniques are used. However, it provides very limited vital information related to emission of greenhouse gases. 'Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms' by S. Salcedo-Sanz, R. C. Deo, L. Carro-Calvo and B. Saavedra Moreno [2] is also a prediction-based idea. In this project, SVR and multilayer perceptron methods are used. Even in this, only the temperature is focused and other physical factors are not taken care of which are responsible for the process of Global Warming. Another prediction based idea is provided by A.

Elmahdi and M. A. Imteaz in the paper 'Multiple Regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes' [3]. In this project, regression and artificial neural network are used to predict the rainfall. This idea is also focusing on rainfall but not upon emission of greenhouse gases. 'Development and Analysis of ANN Models for Rainfall Prediction by Using Time-Series Data' by Neelam Mishra, Hemant Kumar Soni, Sanjiv Sharma, AK Upadhyay [4] is also used as a reference. In this project, the concept of regression, MSE and MRE are used. This idea also focused only on rainfall and not on temperature or greenhouse gas emission. The use of Artificial Neural Networks to observe and forecast rainfall was mentioned in the paper titled 'Application of Artificial Neural Networks to Rainfall Forecasting in Queensland' by Jennifer Merchasy [5]. However, the idea doesn't explain about how rainfall patterns are affected by emission of CO2. The methodology of rainfall prediction through neural networks by Pallavi Gupta is given in her paper titled 'Comparison of neural network configuration in the long-range forecast of south-west monsoon rainfall over India' [6]. Again, there's no relation explained between rainfall and its pattern affected due to evolving problem of Global Warming. Prediction of increased air temperature as a result of global warming is explained by Robert F.C. in his paper titled 'Support Vector Regression with reduced training sets for air temperature prediction with artificial neural network'. 'Analysis of Global Warming Using Machine Learning' by Harvey Zheng [8] is also used as a reference. In this project, concept of SVM, lasso and random forest regression is used. The idea has focused on global warming but the explanation of this idea is more complex and not so clear. 'A hybrid Double Feed forward Neural Network for Suspended Sediment Load Estimation' [9] by Kowk Wing Chau is also a prediction-based paper focused on factors affecting temperature. 'Research on weather forecast based on neural networks' by Yuan Quan and Lu Yuchang [11] papers gives the concept of weather prediction using neural network but doesn't focus on global warming. 'An interactive predictive system for weather forecasting' by Ayham Omary, Ahmad Wedyan, Ahmed Zghoul, Ahmad Banihai and Izzat Alsmadi [12] is also based on the overall weather prediction, not to a specific field. 'Weather Prediction Using Data mining' by Prashant Biradar [13] is also a weather prediction based paper but does not focus on concerns related to global warming. 'An Integrated Approach for Weather Forecasting based on Data Mining and Forecasting Analysis' by G. Vamsi Krishna [14] has good explanation focused the overall weather, not a particular thing. 'Machine Learning Applied to Weather Forecasting' [15], 'Atmospheric Temperature Prediction using Support Vector Machines' [18], Localized Precipitation Forecasts from a Numeric Weather Prediction Model Using Artificial Neural Networks' [19], 'Neural Network Local Forecasting with Weather Ensemble Predictions' [20], 'Multistage Artificial Neural Network Short-Term Load Forecasting Engine with Front-End Weather Forecast' [22], 'Neural network based short-term load forecasting using weather compensation' [25] were used as reference paper too. These all don't focus upon issue of global warming due to emissions of CO2 gas.

## III. OBJECTIVES

Automobile industry is one of the biggest source of emission of a major Greenhouse Gas - CO2. Unless transport emissions are monitored and brought under control, national and international climate goals will be missed. To meet commitments, we need to track emissions from automobiles and build technologies that would help us to decarbonise them effectively. The project aims at predicting CO2 emission levels by analysing dataset containing official record of statistical data from various car makers. The concept of Regression under Machine Learning is implemented to predict the CO2 emission rate and a final study of overall analysis is carried out to determine the best means of predicting rate(s) of emission.

## IV. PROPOSED SYSTEM

Automobile industry is one of the biggest sources of emission of a major Greenhouse Gas i.e., CO2 (Carbon-Dioxide). Unless transport emissions are monitored and brought under control, national and international climate goals will be missed. To meet commitments, we need to track emissions from automobiles and build technologies that would help us to decarbonize them effectively. We need every tool to tackle CO2 emissions from automobiles and early prediction of such emissions using statistical data can help people across the globe in aiding transformative changes that might end up delivering requisite huge cuts in emission.

The project aims at achieving the following:

1. Predicting CO2 emission levels by analyzing dataset containing official record of statistical data from various car makers as per data provided by the official sources of Canadian Government.
2. The concept of Regression under Machine Learning is implemented to predict the emission rate and a final study of overall analysis is carried out to determine the best means of predicting rate(s) of emission.

There are many technologies which can be used for the prediction of emission data. We have tried to test each of these to get the highest accuracy. All of them have different working models which are described below:

1. Linear Regression: It's a simple, yet most powerful supervised machine learning technique which is used to determine the relationship between dependent variable and one or more independent variables. The line of best-fit, also known as the regression line, depicts the relationship between the variables. It is given by the equation $\mathbf{y = mx + c + \epsilon}$, where **y** is the dependent variable, **x** is the independent variable, **c** denotes the intercept, **m** denotes the slope and $\epsilon$ denotes the error term. Linear Regression involving one independent variable is known as simple linear regression whereas the one involving multiple independent variables is known as multi-variable linear regression.
2. Lasso Regression: It's a great regularization technique which is used for feature selection and prevention of over-fitting of the training data to the model. It works by penalizing the sum of absolute value of weights found through the method of regression.
3. Support Vector Regression: In order to solve problems related to classification or regression, we make use of a supervised machine learning technique known as support vector regression. Here, we plot each of the data item as a point in the n dimensional space with the value of each feature being the particular value of the graph co-ordinate. Hyperplane is the straight line that is required for fitting the data and the data points present on either side of the line are referred to as support vectors, which influence the hyperplane orientation, thereby helping us to build the SVR Model.

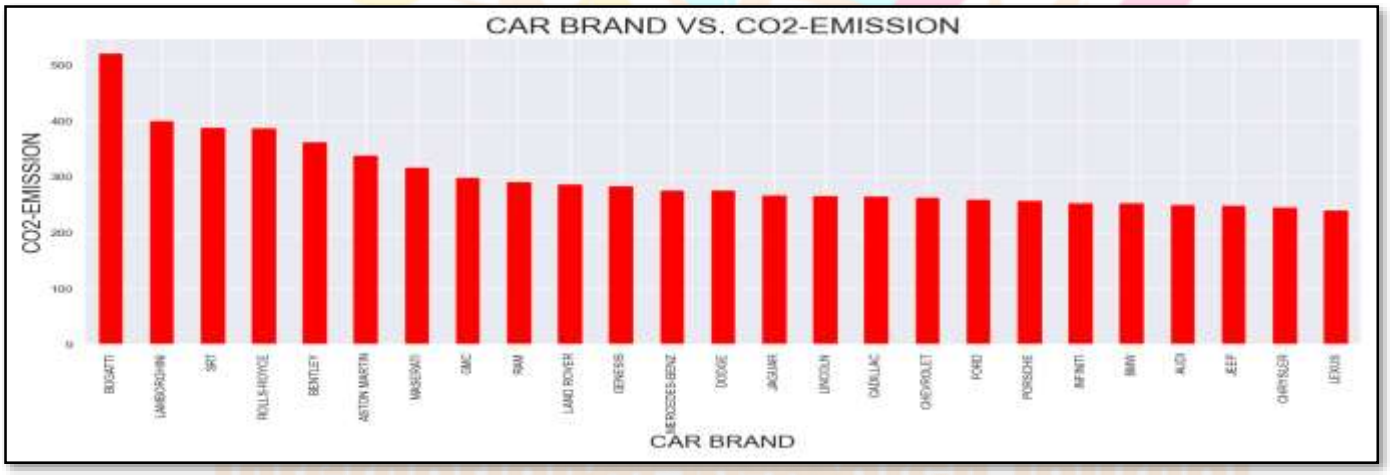
The prediction of CO2 emission rate is possible through three different regression models i.e., Linear Regression Model, Lasso Regression Model and Support Vector Regression Model implemented under this project. The methodology is described through following modules:

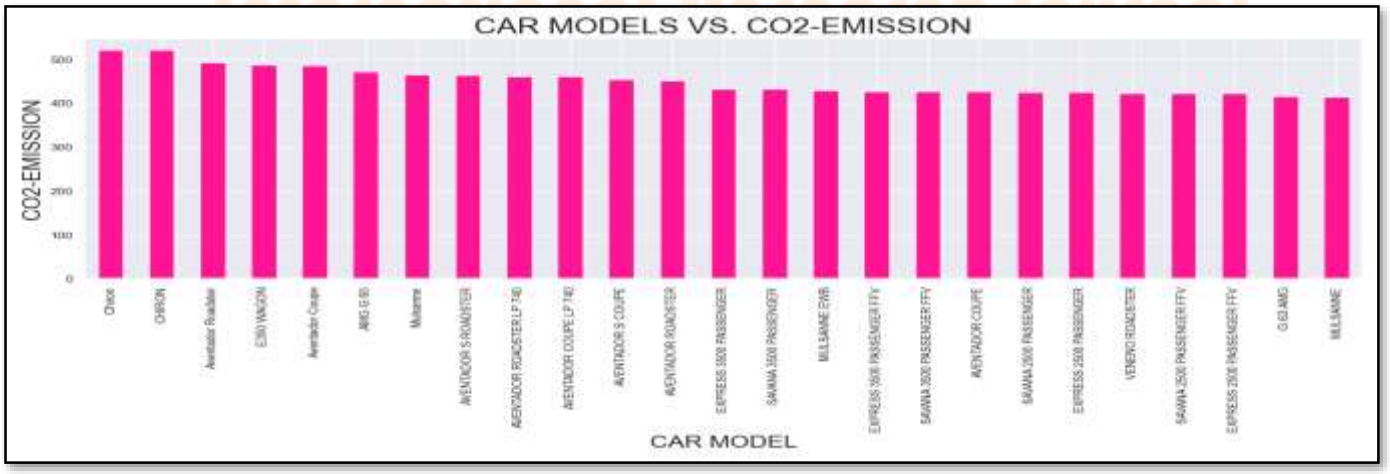
1. Data Collection: The raw data is collected from the data set provided by the official sources of Canadian Government.
2. Data Pre-Processing: The data is initially cleaned and grouped as per requirements. In addition, we check if there are any missing values in the data set or not. If there are some missing values, then change it by any default value. After that, if any data needs to change its format, it is done. Post this step, processed data is utilized for the data prediction.
3. Data Prediction: The processed data prediction can be done by any regression processes which are mentioned above. For that, the processed data is splitted into two groups for the purpose of training and testing. A predictive object is created to predict the test value which is trained by the trained value. Then the object is used to forecast emission data for set of cars from various automobile makers.
4. Data Visualization: Comparison of predicted values vs. actual values is made by making use of separate graphical interface. In addition, we predict the values of intercept, coefficient, mean difference, mean squared error and R2 score for each of the regression models.

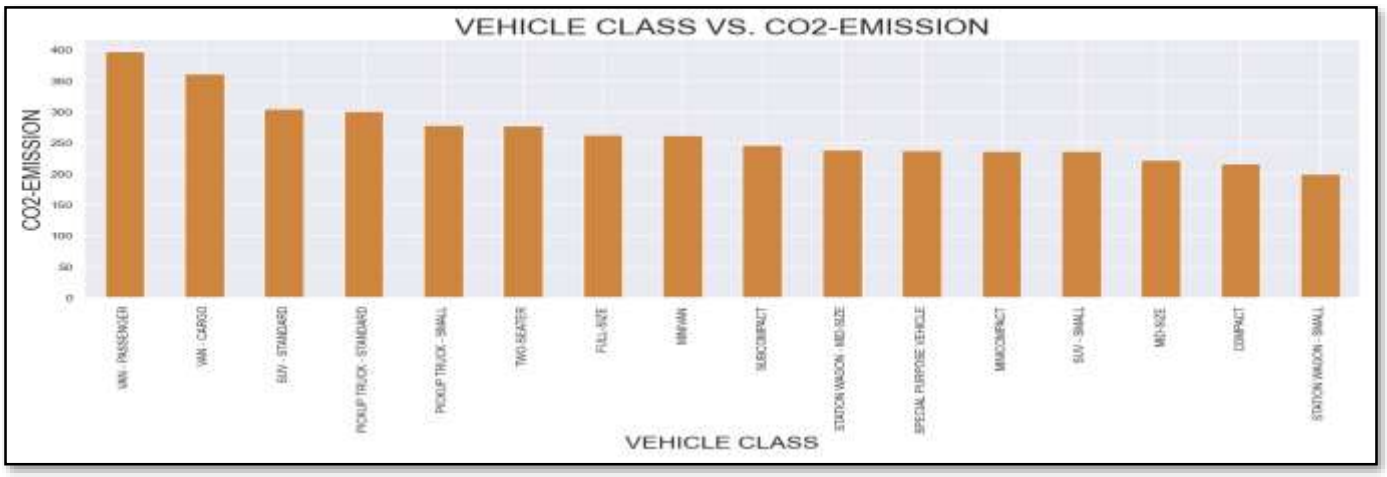
## V. DATATSET

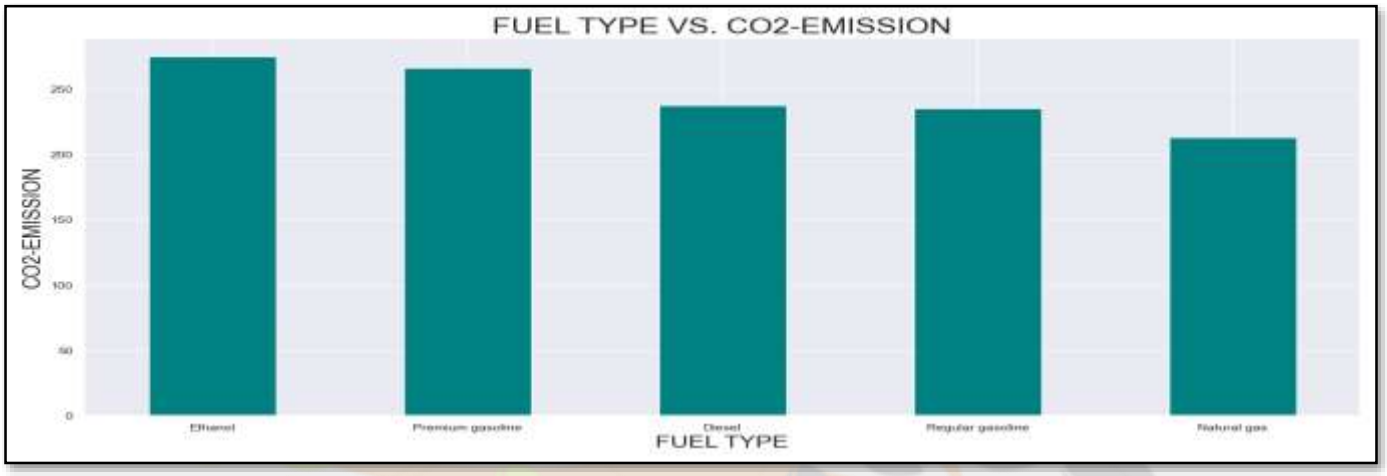
The dataset is provided by the official agencies of Canadian Government consisting of emission data provided by several automobile makers and is available over the Kaggle platform. A total of 7,385 records have enabled us to have statistical soundness of the results provided by statistical analysis done through models of regression.

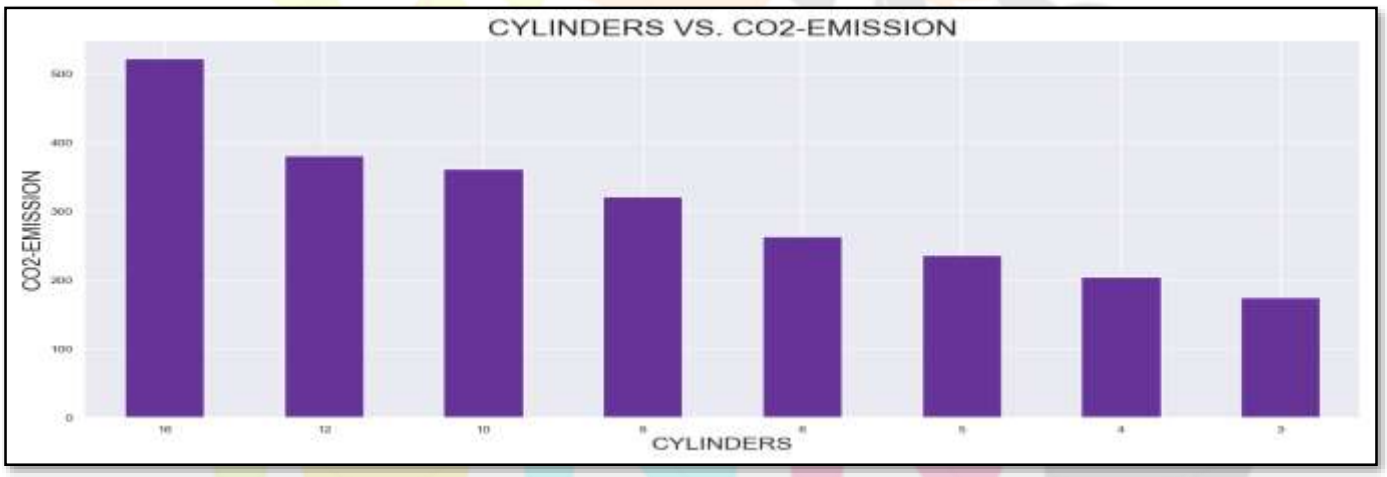
## VI. SNAPSHOTS

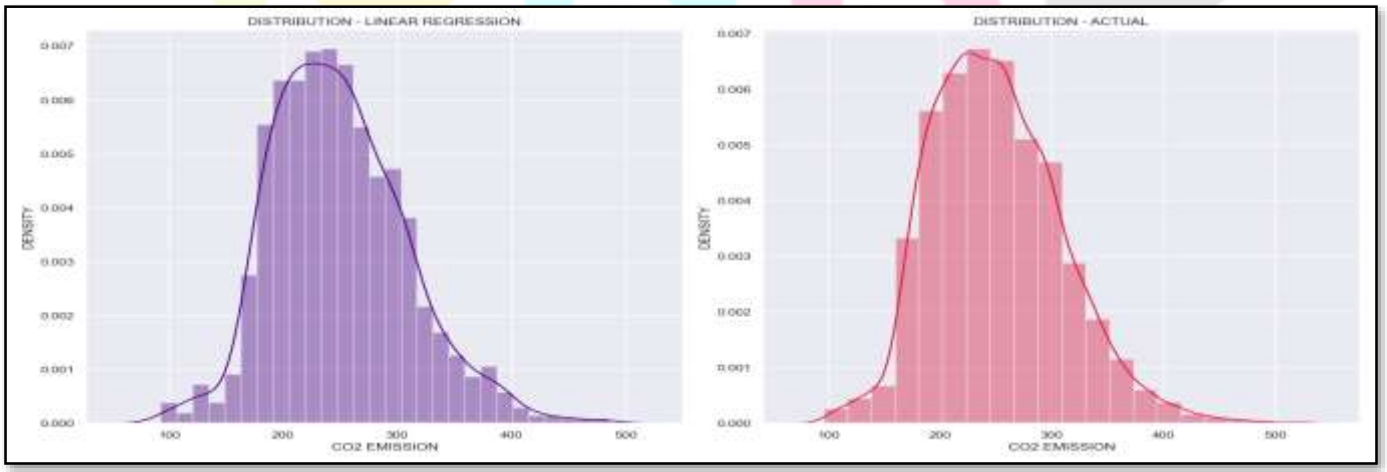



VEHICLE CLASS VS. CO2-EMISSION
CO2-EMISSION
VEHICLE CLASS


FUEL TYPE VS. CO2-EMISSION
CO2-EMISSION
Ethanol
Premium gasoline
Diesel
Regular gasoline
Natural gas
FUEL TYPE


CYLINDERS VS. CO2-EMISSION
CO2-EMISSION
CYLINDERS

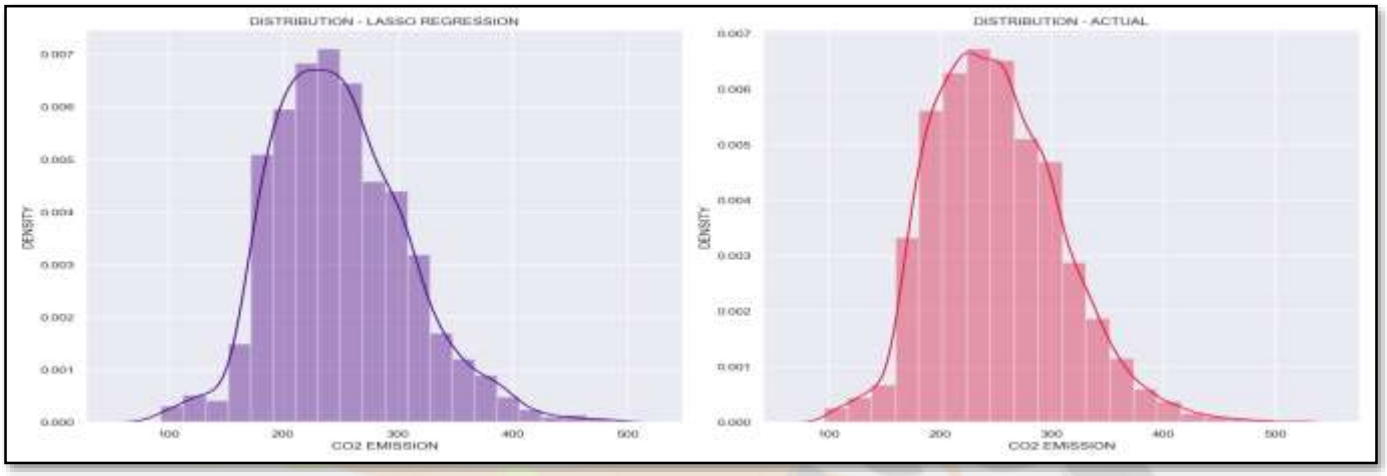

DISTRIBUTION - LINEAR REGRESSION
DISTRIBUTION - ACTUAL
DENSITY
CO2 EMISSION

```
CO2_Emissions Mean:   250.58978873239437

*****LINEAR REGRESSION*****

      Pred_Values  Real_Values  Diff
0          359.06        368.0  8.94
1          292.97        290.0  2.97
2          377.60        382.0  4.40
3          210.80        211.0  0.20
4          192.94        193.0  0.06
...           ...          ...   ...
1472       233.49        235.0  1.51
1473       262.28        263.0  0.72
1474       341.78        346.0  4.22
1475       193.05        193.0  0.05
1476       177.97        177.0  0.97

[1477 rows x 3 columns]
Intercept:  250.98357880480785
Coefficient:  [  0.09846243  -0.09126632  -0.03196586   0.25410999   2.17469937  24.28719842  13.07132844
  20.53745833  -6.35439345 -30.19620392 -15.35456651 -15.09009955  -0.35425518  -0.32284677
  -0.17696755  -0.3645601 ]
Mean Difference:  2.9790589031821257
Mean Squared Error:  4.918260935039378
R2 Score:  0.993041824997087
```

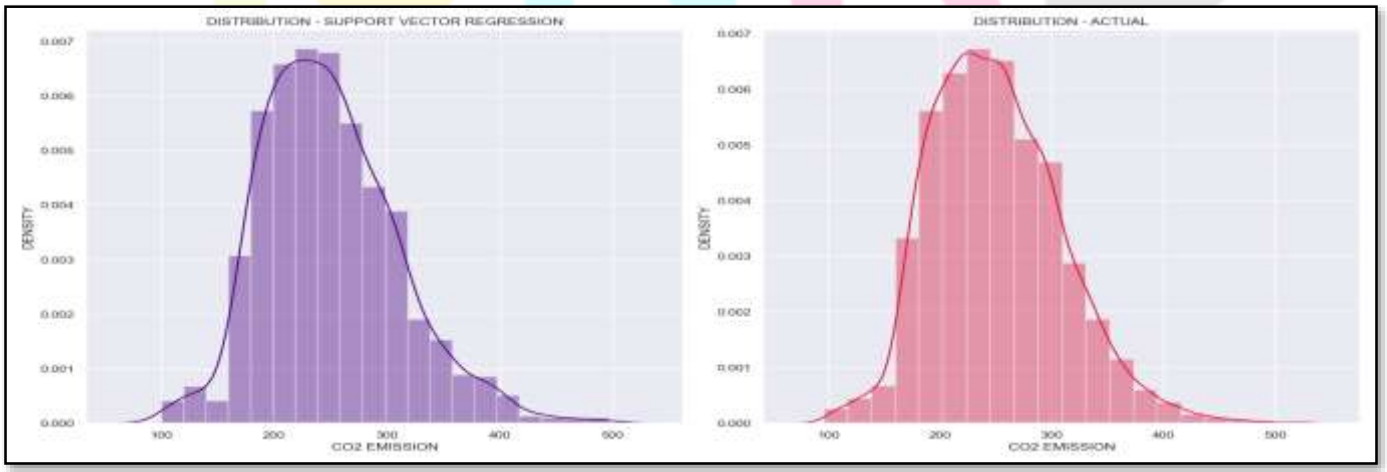



```
CO2_Emissions Mean:   250.58978873239437

*****LASSO REGRESSION*****

      Pred_Values  Real_Values  Diff
0          358.80        368.0  9.20
1          292.87        290.0  2.87
2          376.81        382.0  5.19
3          211.39        211.0  0.39
4          193.37        193.0  0.37
...           ...          ...   ...
1472       233.45        235.0  1.55
1473       262.26        263.0  0.74
1474       341.31        346.0  4.69
1475       193.12        193.0  0.12
1476       178.02        177.0  1.02

[1477 rows x 3 columns]
Intercept:  250.98357880480785
Coefficient:  [ 4.10688699e-02 -3.23723948e-02 -3.10957933e-02  2.68895563e-01  2.36357321e+00  3.30421624e+01
  1.77015362e+01  6.80685134e+00 -6.32948570e+00 -2.86969837e+01 -1.24701105e+01 -1.23090508e+01
  0.00000000e+00  0.00000000e+00 -3.51753527e-02 -1.06025946e-01]
Mean Difference:  3.0695125253893027
Mean Squared Error:  5.010845670292353
R2 Score:  0.9927773882674095
```



```
CO2_Emissions Mean:  250.58978873239437

******SUPPORT VECTOR REGRESSION******

      Pred_Values  Real_Values  Diff
0          363.53         368.0  4.47
1          294.12         298.0  4.12
2          383.39         382.0  1.39
3          210.69         211.0  0.31
4          192.50         193.0  0.50
...           ...           ...   ...
1472       233.26         235.0  1.74
1473       263.99         263.0  0.99
1474       344.80         346.0  1.20
1475       191.77         193.0  1.23
1476       177.00         177.0  0.00

[1477 rows x 3 columns]
Intercept:  [251.38738858]
Coefficient:  [ 9.65857614e-02 -1.40512914e-01 -7.64048072e-02  2.48774128e-01  7.23862185e-01  2.74644624e+01
  1.49695713e+01  2.21484610e+01 -2.21602915e+00 -3.14421271e+01 -1.41138753e+01 -1.39874529e+01
  1.98250128e-01  1.86550282e-01  2.90031891e-02  8.55908589e-03]
Mean Difference:  2.8031753554502368
Mean Squared Error:  5.434780633123261
R2 Score:  0.9915035745194277
```

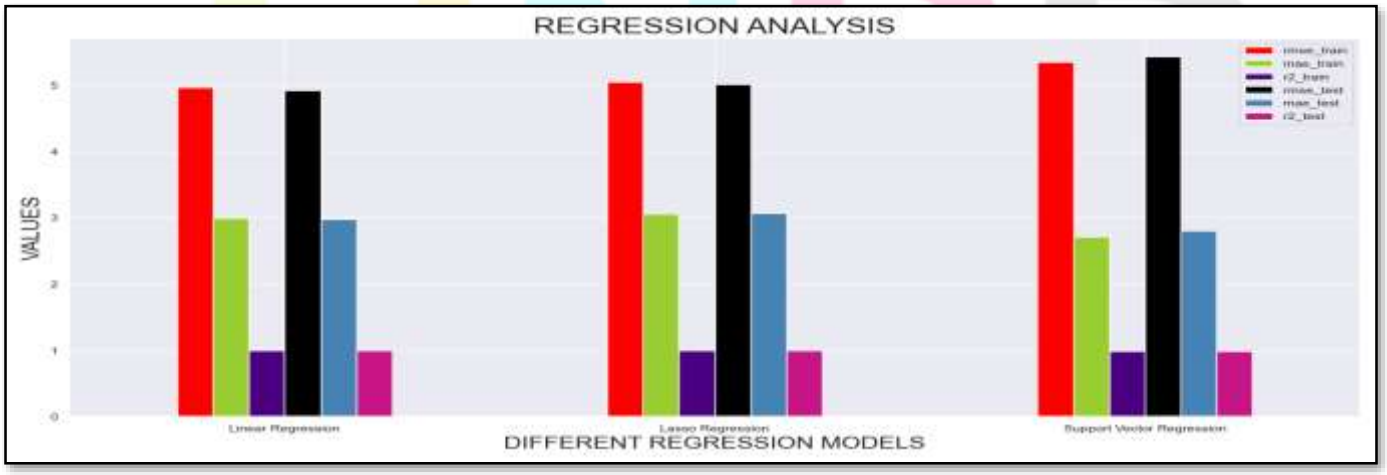
## VII. RESULTS AND CONCLUSION

A total of **7385** records have enabled us to have statistical soundness of the results provided by statistical analysis and meaningful insights have been derived from data under study. A model is considered to be a **BEST FIT** if it has **LOW-RMSE** and **HIGH R2** value(s) whereas it is considered to be a **WORST FIT** if it has **HIGH-RMSE** and **LOW R2** value(s). From the analysis of outcomes of various regression models, we derive the following information:

a. In case of **TRAINING MODEL:**
   - **BEST FIT –** For **LINEAR REGRESSION,** the **RMSE** is the lowest (**4.962163**) and the **R2** is the highest **(0.992778).**
   - **WORST FIT –** For **SUPPORT VECTOR REGRESSION**, the **RMSE** is the highest (**5.345007)** and the **R2** is the lowest (**0.991621**).

b. In case of **TESTING MODEL:**
   - **BEST FIT –** For **LINEAR REGRESSION,** the **RMSE** is the lowest (**4.918261**) and the **R2** is the highest **(0.993042).**
   - **WORST FIT –** For **SUPPORT VECTOR REGRESSION**, the **RMSE** is the highest (**5.429471**) and the **R2** is the lowest (**0.991520**).

| | models | rmse_train | mae_train | r2_train | rmse_test | mae_test | r2_test |
|---|---|---|---|---|---|---|---|
| 0 | Linear Regression | 4.962163 | 2.996013 | 0.992778 | 4.918261 | 2.979052 | 0.993042 |
| 1 | Lasso Regression | 5.044863 | 3.063387 | 0.992535 | 5.010846 | 3.069379 | 0.992777 |
| 2 | Support Vector Regression | 5.345007 | 2.713701 | 0.991621 | 5.429471 | 2.802910 | 0.991520 |




It can finally be concluded from the above results that **LINEAR REGRESSION** is the model of best fit and **SUPPORT VECTOR REGRESSION** is the model of worst fit for the given case under study. The progress achieved in the study of prediction of emission data from this project can help the people across the globe to reach net-zero emissions.

## VIII. SCOPE FOR FUTURE WORK

The project is in its preliminary phase to be able to contribute on a larger scale to the society. The results obtained are significant in understanding the key role played by several factors in contributing to the growth of greenhouse gas emission from

automobiles. In further stages of development, we can expect this project to be implemented as a key feature in vehicle infotainment system that shall generate live emission data, thus providing the government and its associated agencies a chance to understand the patterns of pollution levels across different This can be a game changer and can play a major role in controlling the levels of pollution!

**ACKNOWLEDGMENT**

I express special, in-depth, sincere gratitude to my project guide Dr. Shantakumar B Patil, Head of Department, Computer Science & Engineer, Sai Vidya Institute of Technology, Bengaluru for their constant motivation, guidance and unconditional support.

**REFERENCES**

[1] "A new approach for simulating and forecasting the rainfall-runoff process within the next two months" by M.J Alizadeh, M.R. Kavianpour, Ozgur Kisi; Vahid Nourani, Journal of Hydrology (Vol – 548); May-2017
[2] "Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms" by S. Salcedo-Sanz, R. C. Deo, L. Carro-Calvo, B. SaavedraMoreno; Theoretical and Applied Climatology (Vol – 125); July-2016
[3] "Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes" by F. Mekanik, M.A. Imteaz, S. Gato-Trinidad, A. Elmahdi; Journal of Hydrology (Vol – 503); October-2013
[4] "Development and Analysis of ANN Models for Rainfall Prediction by Using Time-Series Data" by Neelam Mishra, Hemant Kumar Soni, Sanjiv Sharma, AK Upadhyay; International Journal of Intelligent Systems and Applications; January-2018
[5] "Application of Artificial Neural Networks to Rainfall Forecasting in Queensland, Australia" by John Abbot and Jennifer Marohasy, Advances in Atmospheric Sciences; July-2012
[6] "Comparison of neural network configuration in the long-range forecast of southwest monsoon rainfall over India" by Snehasish Chakraverty and Pallavi Gupta; Neural Computing and Applications (Vol – 17); March2008
[7] "Support vector regression with reduced training sets for air temperature prediction a comparison with artificial neural networks" by Robert F. C, Gerrit Hoogenboom, Ronald W. McClendon, J. A. Paz; Neural Computing and Applications (Vol – 20); February-2011
[8] "Analysis of Global Warming Using Machine Learning" by HarveyZheng, Computational Water, Energy; and Environmental Engineering (Vol - 7); July2018
[9] "A hybrid Double Feed forward Neural Network for Suspended Sediment Load Estimation" by Xiao Yun Chen and Kwok Wing Chau; Water Resources Management (Vol - 30); May-2016
[10] "Prediction of the likely impact of climate change on monthly mean maximum and minimum temperature in the Chaliyar river basin, India, using ANN- based models" by N.R. Chithra, S.G. Thampi, S. Surapaneni, R. Nannapaneni, A.A.K. Reddy, J.D. Kumar Theoretical and Applied Climatology (Vol – 121); August-2014
[11] "Research on weather forecast based on neural networks" by Yuan Quan and Lu Yuchang; August-2002
[12] "An interactive predictive system for weather forecasting" by Ayham Omary, Ahmad Wedyan, Ahmed Zghoul, Ahmad Banihai, Izzat Alsmadi; IEEE; June-2012
[13] "Weather Prediction Using Data Mining" by Prashant Biradar, Sarfraz Ansari, Yashavant Paradkar, Savita Lohiya; IJERD (Vol – 5); 2017
[14] "An Integrated Approach for Weather Forecasting based on Data Mining and Forecasting Analysis" by G. Vamsi Krishna; International Journal of Computer Application (Vol – 120); June-2015
[15] "Machine Learning Applied to Weather Forecasting" by Mark Holmstrom, Dylan Liu, Christopher Vo; 2016
[16] "Global Warming: Causes, Effects and Solutions" by Umair Shahzad; Durreesamin Journal (Vol – 1), August2015
[17] "Big Data and Climate Change" by H. Hassani, Xu Huang and E. Silva; February-2019
[18] "Atmospheric Temperature Prediction using Support Vector Machines" by Y.Radhika and M.Shashi; International Journal of Computer Theory and Engineering (Vol – 1); April-2009
[19] "Localized Precipitation Forecasts from a Numeric Weather Prediction Model Using Artificial Neural Networks" by R. J. Kuligowski and A. P. Barros; American Meterological Society (Vol – 13)
[20] "Neural Network Local Forecasting with Weather Ensemble Predictions" by J. W. Taylor and R. Buizza; IEEE Transactions and Power Systems (Vol – 17); August-2002
[21] "Artificial Neural Networks (The Multilayer Perception) – A review of applications in the atmospheric sciences" by M. W. Gardner and S. R. Dorling; Atmospheric Enviroment (Vol – 32), June-1998
[22] "Multistage Artificial Neural Network Short-Term Load Forecasting Engine with Front-End Weather Forecast" by K. Methaprayoon; IEEE Transactions on Industry Application (Vol – 43); December-2007
[23] "Air quality prediction using optimal neural networks with stochastic variables" by A. Russo, F. Raischel, P. G. Lind, Atmospheric Environment 2013
[24] "The forecasting research of early warning systems for atmospheric pollutants: A case in Yangtze River Delta region" by Y. Song, S. Qin, J. Qu, Feng Liu; Atmospheric Environment; 2015
[25] "Neural network based short-term load forecasting using weather compensation" by T.W.S. Chow and C.T. Leung; IEEE Transactions on Power Systems (Vol – 11), Nov – 1996