Regression Analysis using Machine Learning Algorithms to Predict CO 2 Emissions

Conference Paper · February 2024

DOI: 10.23919/INDIACom61295.2024.10499094

4 authors, including:

Lida Anna Joshy
Christ University
1 PUBLICATION **2** CITATIONS
SEE PROFILE

Rakoth Kandan S.
Christ University
25 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Jenefa Kerson
Christ University
25 PUBLICATIONS **100** CITATIONS
SEE PROFILE

Regression Analysis using Machine Learning Algorithms to Predict CO_2 Emissions

Lida Anna Joshy
Dept. of CSE
CHRIST (Deemed to be University)
Kengeri Campus, Bengaluru

Rakoth Kandan Sambandam
Dept. of CSE
CHRIST (Deemed to be University)
Kengeri Campus, Bengaluru
rakothsen@gmail.com

Divya Vetriveeran
Dept. of CSE
CHRIST (Deemed to be University)
Kengeri Campus, Bengaluru

J. Jenefa
Dept. of CSE
CHRIST (Deemed to be University)
Kengeri Campus, Bengaluru

***Abstract* - Precise measurement of fuel consumption and emissions plays an important role in evaluating the environmental effects of materials and stringent emission control methods, especially within the transportation sector. This sector represents a substantial contributor to both global greenhouse gas emissions and the release of hazardous pollutants, making accurate assessment imperative for addressing climate change. The primary objective is to construct accurate predictive models that estimate CO_2 emissions based on vehicle attributes, fostering a deeper understanding of the environmental impact of vehicular activities. Leveraging the "CO_2 Emissions_Canada.csv" dataset, the paper embarks on an extensive journey of data preprocessing, exploratory data analysis, and model training. These algorithms are meticulously fine-tuned and evaluated through metrics such as R-squared and mean absolute percentage error, rendering insights into their predictive accuracies. In essence, this paper pioneers a pathway towards environmentally responsible mobility solutions, capitalizing on the fusion of data science and environmental conservation.**

Keywords - Machine Learning, Emission Control, CO_2, k-NN, Random Forest.

I. Introduction

The rapid expansion of the automotive industry has been a major contributor to the global surge in carbon dioxide (CO_2) emissions, a leading catalyst of climate change. As societies become increasingly conscious of the environmental repercussions tied to greenhouse gas emissions, the imperative to curtail CO_2 emissions from vehicles has taken on paramount importance. Emissions from vehicles prominently rank among the most significant contributors to urban air pollution and the release of greenhouse gases. Increased CO_2 concentrations originating from motor vehicles represent the primary source of atmospheric pollutants, with the potential to result in adverse health consequences, such as headaches and diminished cognitive sharpness. Considering the influence of CO_2 and CO emissions on both air quality and climate change, it becomes essential to incorporate these aspects into emission measurement studies [1].

In this era of technological advancement and sustainability awareness, understanding the relationship between vehicle characteristics and CO_2 emissions is pivotal for regulatory bodies, manufacturers, and consumers alike. In general, vehicle emissions are contingent on their dynamic attributes, encompassing factors like speed, power, acceleration, braking, and other relevant features, in addition to the choice of fuel and engine type. Furthermore, the dynamic characteristics of a vehicle can vary in response to different traffic conditions [1].

This paper is focused to analyze and predict CO_2 emissions from vehicles by employing various machine learning algorithms, providing insights into the factors influencing emission levels and offering potential strategies for reducing environmental impacts. By delving into the dataset containing information about different vehicle attributes and their corresponding CO_2 emission levels, this paper aims to unravel patterns, correlations, and trends that can help to make informed decisions toward greener transportation options.

Deep Bidirectional LSTMS(Bi-LSTM) and Gated Recurrent Unit (GRU) have been used for the Air quality measurement [2] the dataset used in this proposed work is a real-time data taken from the Vishakhapatnam region. Principal Component Analysis (PCA) was used to reduce the complexity of the system.

IoT can also deploy into this air quality prediction with wearable devices to identify the air quality around them. In paper [3] authors proposed a model with Geographic Information System (GIS). Sensor networks have been used to connect with the focused area. Sensors can able to communicate in this area with external devices i.e. wireless, and wired. Hence sensors can able to predict the quality of the air, biological quantities, etc. [4].

II. Related Works

In recent years, the application of regression analysis using diverse machine learning algorithms to predict CO_2 emissions based on various vehicle features has gained substantial attention due to its potential to address pressing environmental concerns and optimize transportation systems. Many works have been done to exemplify the advancements and insights achieved in this domain, contributing to the understanding of vehicular emissions and their implications for sustainable mobility [5].

Over the past few decades, researchers have developed various models to estimate car emissions. An exemplary model known as "MEASURE," crafted by the Georgia Institute of Technology, leverages data-intensive parameters to compute emissions of carbon oxide, nitro oxide, and VOC across various vehicle operating modes. These modes encompass deceleration, cruise control, acceleration, and idle [6].

Another well-known framework for recording road vehicle emissions within European Environment Agency (EEA) member nations is "COPART" [6]. Exploring the links between a nation's natural energy consumption and the patterns of emissions stands as a crucial research endeavor within this context. In a recent investigation involving 70 diesel automobiles subjected to real-world driving conditions, a team of researchers harnessed a machine learning model to forecast both emissions and vehicle performance. To make immediate predictions regarding Nitrogen Oxide (NOx) emissions, they employed a combination of lookup tables, non-linear regression (NLR), and neural network multilayer perceptron (MLP) models [6].

However, it's worth nothing that this model focuses solely on estimating NOx emissions and does not consider CO_2 emissions. A different approach was suggested to evaluate the precision of predictions made by different machine learning models in the context of CO_2 emissions. This approach integrated Gaussian process regression (GPR) and produced positive outcomes. The research study utilized actual driving emission (RDE) data gathered from hybrid electric vehicles [6]. Notably, the study revealed that CO_2 emissions from hybrid electric vehicles do not demonstrate a straightforward, linear correlation with their acceleration and speed.

In paper [7] authors proposed a novel methodology for air quality monitoring using the Firefly algorithm and Support vector machine. In this approach EHR (Electronic Health Record) has been used for the result analysis and H-APIs used for the predictive analysis. Comparably with the existing methods the proposed method achieved a better air quality predictions [8].

III. Proposed Work

The proposed work flow contains two different parts, First workflow of the model begins with pre-processing, Data Analysis, etc. The second part of the model experiments with different algorithms and compares the result analysis for a better understanding of the proposed model.

A. Proposed Workflow

The primary aim of this research work is to develop accurate predictive models using regression analysis and machine learning algorithms to forecast CO_2 emissions from vehicles. By leveraging a dataset containing vehicle attributes and corresponding emission levels, the paper seeks to find the relationships between these factors. In this work, it is focused on the construction of various regression models. The aim is to predict carbon dioxide emissions for light-duty vehicle designs by employing ensemble learning models that leverage vehicle specifications. This approach significantly improves our capability to make accurate emissions forecasts. Through this exploration, the paper intends to facilitate informed decision-making for promoting sustainable mobility solutions, reducing carbon footprints, and advancing environmentally responsible practices within the automotive industry. The proposed work of this paper encompasses a comprehensive set of steps to achieve the main aim. Beginning with data preprocessing, the paper involves loading, cleaning, and structuring the "CO_2 Emissions_Canada.csv" dataset.

Exploratory data analysis follows, where patterns and correlations between vehicle attributes and emission levels are uncovered. Following the dataset division into training and testing sets for the purpose of developing and evaluating predictive models, a diverse set of regression algorithms, comprising Linear Regression, Decision Trees, k-Nearest Neighbors, Random Forests, and XGBoost, is applied in the construction of these models. Model performance is evaluated using metrics like R-squared and mean absolute percentage error, providing insights into their predictive accuracy. The paper further delves into the interpretation of model coefficients and feature importance's, shedding light on the relative influence of different attributes on CO_2 emissions. Insights derived from the models play a pivotal role in suggesting eco-friendly driving practices.

Ultimately, the proposed work seeks to contribute to the ongoing global effort to combat climate change by utilizing data-driven approaches to understand, predict, and mitigate the vehicular CO_2 emissions. Figure 1 represents the overall workflow of the paper.

B. Basic Model

The paper mainly deals with employing a variety of regression algorithms to predict CO_2 emissions from vehicles based on their distinct attributes. After importing and preprocessing the dataset, it finds the correlation between the emission levels and the features. The dataset is initially partitioned into training and testing sets, facilitating both model development and evaluation. Employing algorithms like Linear Regression, Random Forests, Decision Trees, k-Nearest Neighbors, and XGBoost, the paper constructs predictive models aimed at capturing the intricate relationships between features and CO_2 emissions. By quantifying the accuracy of these models using metrics like R-squared and mean absolute percentage error, the paper provides insights into the optimal algorithm for predicting emissions and facilitating informed decision-making in the context of vehicular environmental impact. The prediction outcomes derived from the machine

learning models presented in this paper can serve as a valuable index and point of reference for various stakeholders in different contexts. [9]

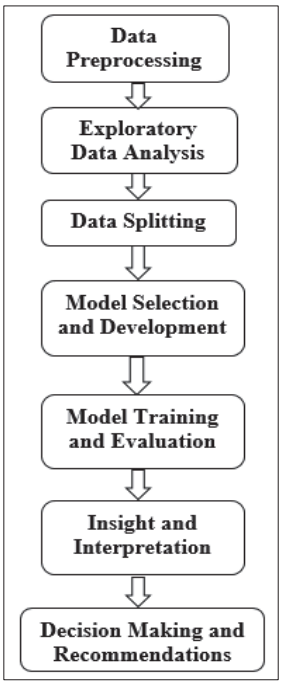

Fig. 1. Proposed Workflow

IV. Experimental and Analytical Work

In this proposed model, Experimental analysis done with various models for predicting the emission prediction among XGB gives method gives better results. The dataset used for this proposed work is real-time data set which well proven dataset that helps to predict emission levels accurately.

A. Dataset

The dataset used in this paper is CO_2 Emissions_Canada.csv which serves as the foundation for understanding and predicting CO_2 emissions from vehicles in the Canadian market. This dataset comprises comprehensive data pertaining to 7,384 light-duty vehicles spanning the years 2017 to 2021. This dataset likely comprises various attributes related to vehicles, such as engine size, cylinder count, fuel consumption, and other relevant characteristics. The dataset likely includes a target variable, "CO_2 Emissions (g/km)," representing the emissions in grams per kilometre for each vehicle. The dataset's richness enables exploratory data analysis to unveil relationships between these attributes and emission levels, helping to identify significant factors contributing to CO_2 emissions. The size of the dataset is (7384, 12).

B. Modeling, Analysis and Design

This paper encompasses several key processes aimed at predicting CO_2 emissions from vehicles using regression analysis and machine learning techniques. To create a prediction model, the initial step involves data collection from the internet. Subsequently, the raw data is transformed into an appropriate format. Following this data preprocessing, the model undergoes validation through a data cleaning process. To assess its accuracy and reliability, cross-validation is executed by partitioning the dataset into training data and testing data. Here it begins with data preprocessing, involving the loading and cleaning of the "CO_2 Emissions_Canada.csv" dataset. Exploratory data analysis follows, revealing patterns and relationships among vehicle attributes and CO_2 emissions.

Following this step, the dataset is partitioned into training and testing sets, which serves to streamline the process of model development and assessment [10]. A variety of regression algorithms, including Linear Regression, Random Forests, Decision Trees, k-Nearest Neighbors, and XGBoost, are utilized for the development of predictive models. These models undergo a fine-tuning process, are trained on the training set, and subsequently evaluated using the testing set. Performance metrics such as R-squared and mean absolute percentage error quantify their accuracy. Finally, insights from the models shed light on factors influencing emissions, contributing to informed decision-making for reducing environmental impact and fostering sustainable transportation practices.

V. Materials and Methods

In this section, we provide comprehensive details regarding the data source, the specific machine learning algorithms employed, and the mathematical models applied in our analysis. Machine learning algorithms have the capability to forecast future responses by analyzing past reactions and the dynamic conversion derived from correlated predictors.

The materials and methods employed in this paper constitute a structured framework for predicting CO_2 emissions from vehicles using regression analysis and machine learning. The primary material encompasses the "CO_2 Emissions_Canada.csv" dataset, containing information on vehicle attributes and CO_2 emissions. Data preprocessing techniques are applied to ensure data quality, including data loading, handling of missing values, and potential data transformations. The methods involve exploratory data analysis to uncover underlying patterns and correlations among the dataset's variables. Multiple regression algorithms, such as Linear Regression, Decision Trees, Random Forests, k-Nearest Neighbors, and XGBoost, are implemented to construct predictive models. Insights derived from the models provide a comprehensive understanding of how various vehicle attributes influence CO_2 emissions, contributing to informed decision-making for sustainable transportation strategies. This approach, amalgamating materials and methods, empowers the paper to analyze the intricate relationship between vehicle features and emissions, guiding towards more environmentally responsible mobility solutions. Fig. 2 represents the accuracy of the various algorithm in the work.

The paper employed a range of regression algorithms, including Linear Regression (LR), Decision Tree Regressor (DTR), Random Forest Regressor (RFR), k-Nearest Neighbors

Regressor (KNR), and XGBoost Regressor (XGB), to predict CO_2 emissions from vehicle attributes.

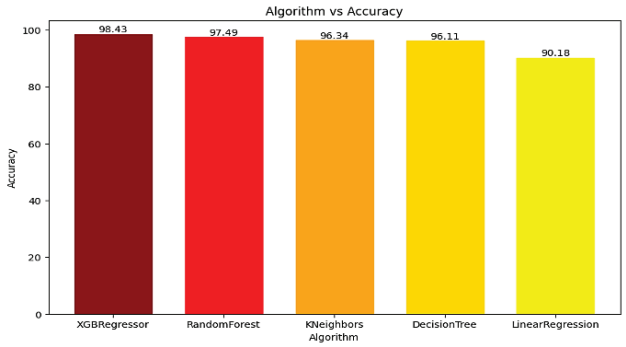

Fig. 2. Algorithm vs Accuracy Graph

The accuracy comparison of these algorithms based on R-squared scores (%) on the test data is as follows:

XGBoost Regressor achieved the highest accuracy, with an R-squared score of approximately 98.43%. Random Forest Regressor attained an accuracy of about 97.35% k-Nearest Neighbors Regressor achieved an accuracy of approximately 96.34%. Decision Tree Regressor achieved an accuracy of around 96.27%. Linear Regression achieved a comparatively lower accuracy of approximately 90.18%.

This comparison provides insights into the performance of each algorithm in predicting CO_2 emissions. Notably, XGBoost Regressor outperformed other algorithms by yielding the highest accuracy, suggesting its proficiency in capturing complex relationships within the dataset. Its ensemble nature and gradient boosting technique likely enabled it to provide the most accurate predictions. The substantial accuracy achieved by Random Forest Regressor, k-Nearest Neighbors Regressor, and Decision Tree Regressor underscores their effectiveness in modeling the dataset's patterns. However, Linear Regression yielded comparatively lower accuracy, possibly due to its simplicity in handling only linear relationships, which might not fully capture the complexity of the CO_2 emissions prediction task. Overall, the XGBoost Regressor emerged as the algorithm with the highest predictive accuracy, making it a promising choice for accurate CO_2 emission predictions in this paper.

VI. Results and Analysis

The results and analysis of this paper yield valuable insights into predicting CO_2 emissions from vehicles using a variety of regression algorithms and machine learning techniques. Through comprehensive data preprocessing, exploratory data analysis, and model training, the paper successfully developed predictive models capable of accurately estimating CO_2 emissions based on vehicle attributes. This study explored the capabilities of different machine learning methodologies. The performance comparison of different algorithms revealed that the XGBoost Regressor exhibited the highest accuracy with an R-squared score of approximately 98.43%, followed by Random Forest Regressor (97.35%), k-Nearest Neighbors Regressor (96.34%), Decision Tree Regressor (96.27%), and Linear Regression (90.18%). These findings highlight the significance of employing advanced techniques like XGBoost and ensemble methods for intricate emission prediction tasks. The analysis underscores the importance of factors such as engine size, fuel consumption, and vehicle class in influencing CO_2 emissions.

Additionally, the paper contributes to sustainable transportation strategies by providing insights into eco-friendly driving practices and informing decisions for reducing environmental impact. Overall, this paper serves as a testament to the effectiveness of machine learning in tackling complex environmental challenges and offers a pathway towards greener and more responsible mobility solutions. The implementations of the diverse models were compared and assessed utilizing a range of evaluation metrics, which encompassed mean square error (MSE), R-squared (R2), and root mean square error (RMSE). The calculated mean absolute percentage error (MAPE) value of approximately 0.013 in the paper holds significance as a quantifiable measure of the predictive accuracy of the model developed for CO_2 emission prediction. The MAPE value represents the average percentage difference between the actual CO_2 emission values and the corresponding predicted values. A lower MAPE value indicates that the model's predictions are very close to the actual values, with an error of around 1.3%.

The graph as shown in Fig. 3. depicted by the code snippet visually presents a comparison between the actual and predicted CO_2 emission values from the model developed in the paper. The x-axis represents the number of data points or instances, while the y-axis represents the CO_2 emission values. The graph's dual lines, one in dark blue representing the actual values and the other in pale green representing the predicted values, illustrate how closely the model's predictions align with the ground truth data.

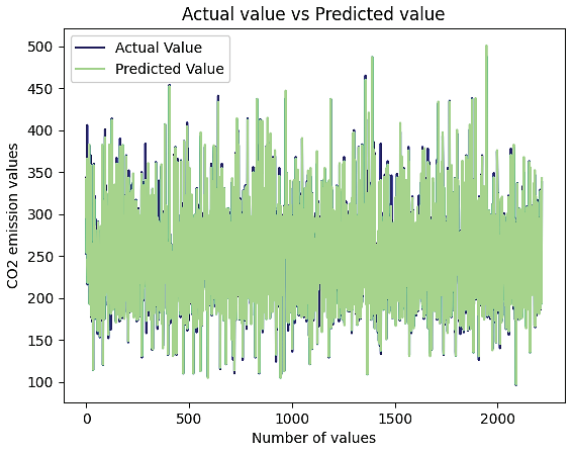

Fig. 3. Accuracy- Graphical representation of actual value and predicted value of CO_2 emission.

When the two lines closely overlap, it indicates that the model's predictions are accurate and in good agreement with

the actual emissions. In such cases, the pale green line aligns almost perfectly with the dark blue line. Deviations between the lines may occur due to variations in the dataset, complexities in the relationship between attributes and emissions, or inherent noise in the data [11].

This graph serves as a visual confirmation of the model's performance. The greater the alignment between the two lines, the more reliable the model's predictions are. Conversely, significant disparities between the lines may indicate areas where the model struggles to accurately predict emissions [12]. In summary, this visual representation offers an intuitive way to assess the model's precision and its capacity to generalize to unfamiliar data points. It contributes to a clearer understanding of how well the predictive models capture the nuances of CO_2 emissions based on vehicle attributes [13].

VII. Conclusion

This paper presents an observational and predictive study aimed at offering a comparative analysis of various brands and vehicle types concerning their fuel consumption and CO_2 emissions. This paper successfully demonstrates the power of regression analysis and machine learning in predicting CO_2 emissions from vehicles. By following a comprehensive pipeline encompassing data preprocessing, exploratory data analysis, model training, and evaluation, this paper has successfully crafted precise predictive models capable of estimating emissions based on vehicle attributes. The comparison of various regression algorithms highlighted the superiority of the XGBoost Regressor in achieving the highest accuracy, emphasizing the significance of employing advanced techniques for intricate prediction tasks. The paper's findings underscore the importance of vehicle attributes like engine size, fuel consumption, and class in influencing emission levels.

These insights can inform environmentally conscious driving practices and guide decisions for reducing carbon footprints within the automotive sector. The applications of this paper extend to real-world scenarios. The predictive models can be integrated into vehicle emissions estimation tools, aiding regulatory bodies, manufacturers, and policymakers in making informed decisions to mitigate environmental impacts. This research delves into a comprehensive examination of different vehicle types and brands, leveraging vehicle measurements to gain deeper insights into the automotive market and its environmental implications. The recommended vehicle attributes and predictive models put forth in this study can serve as valuable guides for both consumers and vehicle manufacturers, aiding them in making informed choices and undertaking measures to reduce their environmental footprint.

By encouraging the adoption of vehicles with lower emission levels and fostering eco-friendly driving habits, the paper contributes to a more sustainable future. Ultimately, this paper serves as a testament to the potential of data-driven approaches in addressing pressing environmental challenges and driving positive change within the automotive industry and beyond.

Acknowledgement

The authors are grateful for the facilities provided by the CHRIST (Deemed to be University), Kengeri Campus, Bengaluru for the facilities offered to carry out this work

References

[1] Chandrashekar, Pritha Chatterjee, Digvijay S. Pawar, “Estimation of CO2 and CO emissions from auto-rickshaws in Indian heterogeneous traffic”. Journal 2(5), pp. 99–110, 2016.

[2] Nisha Thakur, Sanjeev Karmakar, Ravi Shrivastava, “Hybrid deep learning algorithms for forecasting air quality index using dimension reduction technique in search of precise results”, International Journal of Information Technology, Vol.15(6), 2023.

[3] Azeddine Khiat et.al, “New approach based internet of things for a clean atmosphere”, International Journal of Information Technology, Vol.11 (1), 2019.

[4] Cui-lin Wu, Hong-di He, Rui-feng Song, Xing-hang Zhu, Zhong-ren Peng, Qing-yan Fu, Jun Pan, “A hybrid deep learning model for regional O3 and NO2 concentrations prediction based on spatiotemporal dependencies in air quality monitoring network”, Environmental Pollution, Vol.320, 2023.

[5] Deepak Narayan Paithankar, Abhijeet Rajendra Pabale, Rushikesh Vilas Kolhe, P. William, Madhukar Yawalkar, “Framework for implementing air quality monitoring system using LPWA-based IoT technique”, Measurement: Sensors, vol.26, 2023.

[6] Yuvaraj Natarajan, Gitanjali Wadhwa, K. R. Sri Preetha and Anand Paul, “Forecasting Carbon Dioxide Emissions of Light-Duty Vehicles with Different Machine Learning Algorithms”, Electronics, Vol.12 (10), 2023.

[7] Pankaj Rahi, Sanjay P. Sood, Rohit Bajaj, Yogesh Kumar, “Air quality monitoring for Smart eHealth system using firefly optimization and support vector machine”, International Journal of Information Technology, Vol.13(5), 2021.

[8] Manuel Mendez, Mercedes G.Merayo, Manuel Nunez, “Machine Learning Algorithms to Forecast Air Quality: A Survey”, Artificial Intelligence Review, vol. 56, 2023.

[9] Ümit Ağbulut, “Forecasting of transportation-related energy demand and CO2 emissions in Turkey with different machine learning algorithms, Sustainable Production and Consumption”, vol. 29, pp. 141-157, 2022.

[10] Abdul Motin Howlader, Dilip Patel, and Robert Gammariello, “Data-Driven Approach for Instantaneous Vehicle Emission Predicting Using Integrated Deep Neural Network”, Transportation Research Part D: Transport and Environment, vol. 116, 2023.

[11] Chowdaiah Chandrashekar, Rohan Singh Rawat, Pritha Chatterjee and Digvijay Sampatrao Pawar, “Evaluating the real-world emissions of diesel passenger Car in Indian heterogeneous traffic”, Environmental Monitoring and Assessment, Vol. 195, 2023.

[12] Yang Meng and Hossain Noman, “Predicting CO2 Emission Footprint Using AI through Machine Learning”, Atmosphere, vol. 13(11), 2022.

[13] Ngo Le Huy Hien and Ah-Lian Kor, “Analysis and Prediction Model of Fuel Consumption and Carbon Dioxide Emissions of Light-Duty Vehicles”, Applied Sciences, vol. 12(2), 2022.