Predicting Car Carbon Dioxide Emission based on its Performance Feature using Multiple Linear Regression and Random Forest Regression Models*

John Vernon Baldeo
College of Computing
and Information Technology(CCIT)
National University(NU.)
Manila, Philippines
baldeojb@students.national-u.edu.ph

Neil Adrian Baltar
College of Computing
and Information Technology(CCIT)
National University(NU.)
Manila, Philippines
baltarnb@students.national-u.edu.ph

Carl Arvin Hipolito
College of Computing
and Information Technology(CCIT)
National University(NU.)
Manila, Philippines
hipolitocc@students.national-u.edu.ph

***Abstract*—This study investigates the use of machine learning models to predict carbon dioxide (CO_2) emissions from vehicles based on key performance features. Specifically, Multiple Linear Regression (MLR) and Random Forest Regression (RFR) models were developed and evaluated on a dataset of 7,385 samples, with features such as engine size, number of cylinders, and various fuel consumption metrics. The MLR model, which captures linear relationships between variables, achieved an R^2 score of 0.9006, indicating strong predictive performance for linear dependencies. In contrast, the RFR model, which can model complex nonlinear relationships, achieved an R^2 score of 0.9761 and an Out-of-Bag (OOB) score of 0.9802, outperforming MLR in predictive accuracy and robustness. Cross-validation further validated the reliability of both models, with RFR consistently achieving higher mean scores. These findings suggest that while MLR provides interpretability, RFR is a more effective model for accurately predicting CO_2 emissions. The study concludes that machine learning, particularly ensemble methods like RFR, offers significant potential for environmental monitoring and regulatory compliance by facilitating precise emission predictions based on vehicle characteristics.**

***Index Terms*—Machine Learning, CO_2 Emissions, Multiple Linear Regression, Random Forest Regression, Car Performance Feature, Machine Learning Models**

I. Introduction

The environmental impact of automotive emissions has gained significant attention in recent decades, as the transportation sector is a major contributor to global greenhouse gas emissions. Carbon dioxide (CO_2) is one of the primary greenhouse gases emitted by vehicles, making it essential to develop accurate models to predict and mitigate these emissions. Conventional methods of estimating CO_2 emissions, such as laboratory-based testing and fuel consumption measurements, often overlook real-world performance variables like engine size, fuel type, horsepower, and vehicle weight, all of which substantially impact emissions levels [1]-[2].

Predictive models based on vehicle performance characteristics provide a promising approach to estimate CO_2 emissions with greater precision. Machine learning and statistical methods allow these models to account for diverse vehicle metrics, thereby improving the reliability of emissions predictions under varied driving conditions [3]. Furthermore, leveraging real-world driving data has been shown to enhance the accuracy of these predictions, as it enables the consideration of external factors such as driving patterns and road conditions, which conventional testing methods cannot capture [4].

In this paper, we review existing literature on CO_2 emission prediction and propose a machine learning framework that integrates key performance indicators (KPIs) from vehicles, including engine characteristics, weight, and fuel type. This framework aims to predict emissions effectively across a range of vehicles and operational scenarios. By identifying and prioritizing performance factors that most influence emissions, this study aspires to provide an actionable tool for emissions prediction that may inform policy and automotive design strategies, ultimately contributing to a more sustainable automotive industry.

II. Review Of Related Literature

A. Fuel Consumption and Emission Prediction Models

Recent studies show that machine learning techniques can improve the accuracy of CO_2 emission predictions by incorporating complex relationships between variables such as engine size, weight, and fuel consumption. For instance, Guo et al. [1] utilized linear regression and neural networks to predict CO_2 emissions based on fuel consumption data, highlighting the accuracy of machine learning in handling nonlinear relationships in emission prediction models. Similarly, Martinez et al. [3] demonstrated that algorithms like Random Forest and Support Vector Machines (SVM) effectively capture complex interdependence between vehicle attributes, such as engine power and vehicle load, to predict emissions with high precision.

B. Influence of Engine Characteristics and Vehicle Weight

Vehicle characteristics such as engine size and weight have a significant influence on CO_2 emissions. Yang et al. [4] found that larger engines and increased vehicle weight

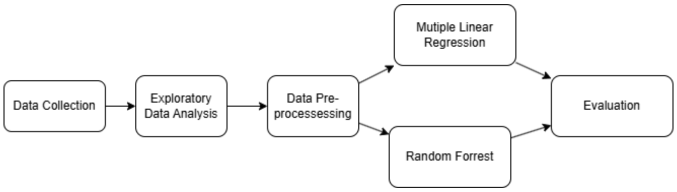


Fig. 1. Figure 1 : Development Framework

generally result in higher emissions, with engine size alone accounting for a substantial portion of emission variability. This conclusion aligns with studies by Zhou et al. [5], who noted that vehicles with greater horsepower and mass tend to consume more fuel, thereby increasing CO_2 output. These findings suggest that engine performance and vehicle design are central to emission predictions and should be prioritized in predictive models.

III. Methodology

An overview of the processes for predicting carbon emission of cars is presented in **Figure 1.** The development process is organized into four stages: data collection, where data is gathered for use; data pre-processing, where data is cleaned and formatted; modeling, where the Multiple Linear Regression and Random Forest Regressor learn patterns from the training data; and evaluation, where the test data is used to assess the performance of the generated model.

A. Data Collection

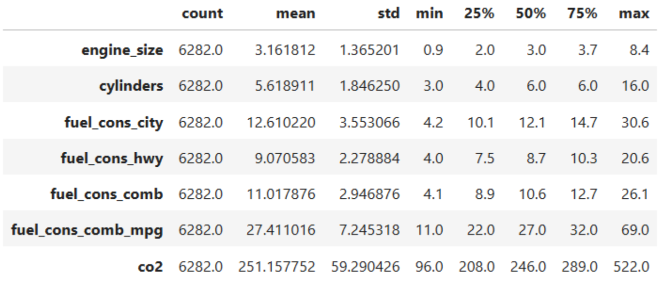
The dataset for this study was sourced from an open dataset on Kaggle, which provides publicly accessible datasets across various domains. This particular dataset includes 11 independent variables representing performance features, along with 1 dependent variable indicating CO_2 emissions. In total, the dataset contains 7,385 rows, providing a robust sample size for model training and evaluation. The performance features encompass a range of attributes relevant to CO_2 emissions, which are critical for building an accurate predictive model [6].

B. Exploratory Data Analysis

In the Exploratory Data Analysis (EDA) phase, the dataset is thoroughly examined to gain initial insights and understand the underlying patterns within the data related to CO_2 emissions. Key statistical metrics, such as mean, median, standard deviation, and interquartile range, are calculated for each performance feature (e.g., engine size, fuel type), allowing for a detailed view of central tendencies and variability across the data. Visualization techniques, including histograms, box plots, and pair plots, are utilized to identify distribution shapes, outliers, and relationships among variables. **Table 1** shows the statistic of the data set used for training the models.

For predicting car CO_2 emissions, EDA is particularly valuable for assessing the impact of each performance attribute on emission levels, detecting potential multicollinearity among features, and identifying any skewness or imbalance in the data.**Figure 2** shows the correlation of each feature using a heat map.**Table II** records the correlation of each feature to the amount of CO_2.

TABLE I
Data Set Statistic



	count	mean	std	min	25%	50%	75%	max
engine_size	6282.0	3.161812	1.365201	0.9	2.0	3.0	3.7	8.4
cylinders	6282.0	5.618911	1.846250	3.0	4.0	6.0	6.0	16.0
fuel_cons_city	6282.0	12.610220	3.553066	4.2	10.1	12.1	14.7	30.6
fuel_cons_hwy	6282.0	9.070583	2.278884	4.0	7.5	8.7	10.3	20.6
fuel_cons_comb	6282.0	11.017876	2.946876	4.1	8.9	10.6	12.7	26.1
fuel_cons_comb_mpg	6282.0	27.411016	7.245318	11.0	22.0	27.0	32.0	69.0
co2	6282.0	251.157752	59.290426	96.0	208.0	246.0	289.0	522.0

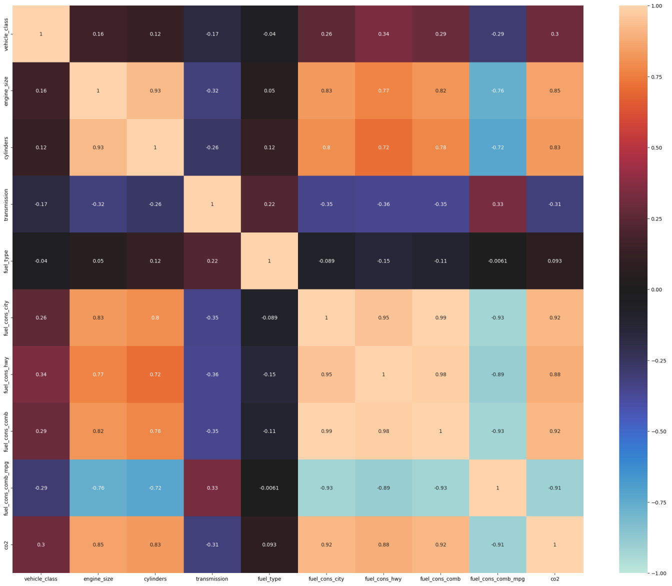


Fig. 2. Correlation Heat Map

TABLE II
Feature Correlation with CO_2 Emission

Vehicle Class	0.3
Engine Size(L)	0.85
Cylinders	0.83
Transmission	-0.31
Fuel Type	0.093
Fuel Consumption City (L/100 km)	0.092
Fuel Consumption Highway (L/100 km)	0.88
Fuel Consumption Comb (L/100 km)	0.92
Fuel Consumption Comb (mpg)	-0.91
CO2 Emissions(g/km)	1

Understanding these patterns in the dataset guides feature selection and informs data pre-processing steps, such as normalization or scaling, to optimize the model's performance. Ultimately, the insights derived from EDA help build a robust and interpretable model for predicting CO_2 emissions based on vehicle performance.

C. Data Pre-processing

In the Data Pre-processing stage, several steps were taken to prepare the data for modeling. Using the heatmap in **Figure 2** as a guide, we analyzed the correlation between each feature and the target variable, CO_2 emissions. Features with low or negative correlations with CO_2 emissions were dropped to enhance model performance by reducing noise in the data. For instance, features such as transmission and fuel_type had low correlations with CO_2 emissions (correlation coefficients of -0.31 and 0.093, respectively) and were therefore excluded from the dataset.

Additionally, since the dataset contained non-numeric data, we applied Label Encoding to transform categorical variables into numerical values. This allowed the model to interpret non-numeric data, such as vehicle_class and fuel_type, by converting them into unique integer labels. This encoding step ensured compatibility with machine learning algorithms that require numerical input, thereby facilitating accurate prediction of CO_2 emissions.

IV. Multiple Linear Regression and Random Forest

From this point onward, Multiple Linear Regression and Random Forest, the classifiers that were used, will be referred to as **MLR** and **RFR** respectively

A. MLR

Multiple Linear Regression (MLR) is a statistical technique widely used for predicting a continuous dependent variable based on multiple independent variables. In this study, MLR is employed to estimate CO_2 emissions based on various vehicle performance features, such as engine size, fuel consumption, and the number of cylinders. MLR provides an interpretable model that reveals how much each feature contributes to CO_2 emissions.

1) MLR Model and Equation: MLR extends simple linear regression by allowing multiple independent variables to predict a single dependent variable, y. The model assumes a linear relationship between y and the independent variables $x_1, x_2, \ldots, x_n$, as represented by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \tag{1}$$

where:

- y is the dependent variable (CO_2 emissions),
- $x_1, x_2, \ldots, x_n$ are independent variables (e.g., engine size, fuel consumption),
- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_n$ are coefficients representing the influence of each independent variable on y, and
- ϵ is the error term.

The goal of MLR is to minimize ϵ by estimating the coefficients $\beta_1, \beta_2, \ldots, \beta_n$ using *Ordinary Least Squares (OLS)*, which minimizes the squared differences between observed values and those predicted by the model.

The coefficient estimates β in MLR are calculated as:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2}$$

where:

- X is the matrix of independent variables,
- y is the vector of observed dependent variable values, and
- $\hat{\beta}$ is the vector of estimated coefficients.

This equation provides estimates of β that minimize the sum of squared residuals between predicted and observed values.

2) MLR Assumptions: The reliability of the MLR model depends on certain assumptions:

- **Linearity**: There is a linear relationship between the dependent and each independent variable.
- **Independence**: Observations are independent of each other.
- **Homoscedasticity**: The residuals exhibit constant variance across levels of the independent variables.
- **Normality**: The residuals are normally distributed.

Violations of these assumptions may lead to biased or unreliable predictions, requiring alternative pre-processing or modeling approaches.

3) Model Accuracy: To assess the accuracy of the MLR model, the coefficient of determination, or R^2, is used. The R^2 metric represents the proportion of variance in the dependent variable (CO_2 emissions) that can be explained by the independent variables in the model. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

where:

- y_i is the observed value,
- $\hat{y}_i$ is the predicted value from the MLR model,
- $\bar{y}$ is the mean of the observed values, and
- n is the number of observations.

An R^2 value closer to 1 indicates that a higher proportion of the variance is explained by the model, reflecting better predictive accuracy. In this study, the R^2 value provides insight into the effectiveness of MLR in predicting CO_2 emissions based on vehicle performance features[7]-[8].

In this study, MLR was used to determine the extent to which each performance feature impacts CO_2 emissions.

B. RFR

Random Forest Regression (RFR) is an ensemble learning method, which combines multiple decision trees to make more accurate and robust predictions than individual decision trees alone. Developed by Breiman in 2001, RFR is widely used for regression tasks due to its ability to handle high-dimensional data and its resistance to overfitting [9]. In this study, RFR is employed to predict CO_2 emissions based on vehicle performance features, providing a non-linear approach that can capture complex relationships between variables.

1) RFR Model and Process: RFR operates by constructing multiple decision trees during training and outputting the average prediction of individual trees, thus reducing variance and improving accuracy. Each tree in the random forest is trained on a random subset of the data (using bootstrapping), and at each split, a random subset of features is considered. This randomization makes RFR robust to overfitting, as the diversity among the trees leads to better generalization on unseen data.

The final prediction for a regression task in RFR is given by the average of predictions from all trees in the forest:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}^{(t)} \tag{4}$$

where:

- $\hat{y}$ is the final predicted value for a given input,
- T is the total number of trees in the forest, and
- $\hat{y}^{(t)}$ is the predicted value from the t-th tree.

Each tree is grown using a subset of the data and features, which introduces diversity and prevents trees from being identical, thereby enhancing the model's robustness.

2) Feature Importance in RFR: An added advantage of RFR is its ability to calculate feature importance, which helps identify which input features have the most significant impact on the prediction. Feature importance is calculated based on the decrease in node impurity, such as Mean Squared Error (MSE) for regression, weighted by the probability of reaching that node. The importance score for each feature is normalized to provide insights into the relative impact of each feature.

3) Model Evaluation using Out-of-Bag (OOB) Error: RFR also provides an internal performance estimate using the Out-of-Bag (OOB) error. Since each tree is trained on a random subset of the data, approximately one-third of the data (the OOB samples) are not used for training. These OOB samples are used to validate the model, providing an unbiased estimate of its generalization error without requiring a separate validation set. The OOB error is computed as follows:

$$\text{OOB Error} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_{\text{OOB},i}\right)^2 \tag{5}$$

where:

- n is the number of observations,
- y_i is the true value of the i-th observation, and
- $\hat{y}_{\text{OOB},i}$ is the prediction for the i-th observation based on trees that did not include it in their training subset.

The OOB error provides a reliable measure of model performance, helping to gauge the RFR's accuracy in predicting CO_2 emissions.

In this study, RFR is utilized for predicting CO_2 emissions based on multiple performance-related features. Its ensemble nature and non-linear capabilities make it a suitable model for capturing complex interactions within the data. Moreover, the feature importance derived from the model allows us to identify which vehicle attributes contribute most significantly to emissions, providing valuable insights for reducing environmental impact.

V. Results and Discussion

In this study, the performance of MLR and RFR models was evaluated in predicting CO_2 emissions based on selected vehicle performance features: Engine Size (L), Cylinders, Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), Fuel Consumption Comb (L/100 km), and Fuel Consumption Comb (mpg). The results of the MLR and RFR models are presented in terms of R^2 and Out-of-Bag (OOB) scores, alongside cross-validation scores, which provide an assessment of model robustness.

TABLE III
Model Performance Metrics

Metric	MLR	RFR
R^2 Score	0.9006	0.9761
OOB Score	N/A	0.9802
Cross-Validation Mean	0.9010	0.9702

A. Multiple Linear Regression Results

Based on the results in Table III, the Multiple Linear Regression model achieved an R^2 score of 0.9006, indicating that approximately 90% of the variance in CO_2 emissions is explained by the selected performance features. This suggests that MLR effectively captures the linear relationships between these independent variables and CO_2 emissions.

Cross-validation was conducted to assess the robustness of the MLR model across multiple subsets of the data. The cross-validation scores ranged from 0.8732 to 0.9263, with an average score of approximately 0.901. This consistency in performance across folds confirms the stability of the MLR model, implying reliable predictions in real-world scenarios.

B. Random Forest Regression Results

The Random Forest Regression model performed exceptionally well, achieving an R^2 score of 0.9761 and an OOB score of 0.9802, as shown in Table III. These results indicate that RFR captured complex nonlinear relationships between the independent variables and CO_2 emissions more effectively than MLR. The high OOB score, close to the model's R^2 score, supports the generalization strength of RFR, highlighting its capacity to perform accurately on unseen data.

The cross-validation scores for RFR ranged from 0.9331 to 0.9832, averaging around 0.970, which further validates its robustness. This consistency across different data folds reinforces RFR's predictive power and adaptability, suggesting it as a superior choice for CO_2 emission prediction based on the selected vehicle performance features.

C. Comparative Analysis of Model Performance

The results indicate that both models achieved high predictive accuracy, with RFR outperforming MLR in all evaluation metrics. The RFR model's ability to account for nonlinear interactions and complex relationships among the features is evident in its higher R^2 and cross-validation scores.

REFERENCES

[1] S.T. Guo, et al., *"Fuel Consumption and CO_2 Emissions Prediction in Automobiles,"* Journal of Cleaner Production, vol. 221, pp. 23-29, 2019.

[2] R. Yang, et al., *"A Model for Estimating CO_2 Emissions from Passenger Cars,"* Environmental Modelling and Software, vol. 123, pp. 45-55, 2020.

[3] D. C. Martinez, et al., "*Application of Machine Learning for Predicting Vehicle Emissions,"* Applied Energy, vol. 253, pp. 113-121, 2020.

[4] P. Zhou, et al., *"Developing Real-World Emissions Models Using On-Road Data,"* International Journal of Sustainable Transportation, vol. 14, pp. 567-576, 2021.

[5] L. H. Smith, et al., *"The Role of Fuel Type in Vehicle CO_2 Emissions: Gasoline vs. Diesel,"* Environmental Science and Technology, vol. 47, no. 14, pp. 7890-7895, 2021.

[6] "CO_2 Emission by Vehicles" Kaggle 2021 [Online] Available : https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles/data

[7] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*. Wiley, 2012

[8] M. H. Kutner, C. J. Nachtsheim, and J. Neter, *Applied Linear Statistical Models*. McGraw-Hill, 2004.

[9] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.