International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8 May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.082 ISSN: 2581-4621

Prediction of CO2 Emissions from Vehicles Using Machine Learning

B. Swetha[1] Assistant Professor
Rajpurohit Krishna Pal Singh, Muchumarri Sophiya, Patan Shujjath Ali Khan, Maripati Pavani, Mogolla Sai Chaithanya
Department of Computer Science and Engineering, K.S.R.M. College of Engineering, Kadapa.
*Corresponding author 1: swetha.b@ksrmce.ac.in

***Abstract*—** Demand for transport infrastructure and for vehicles has grown sharply in recent years, and emissions from the transport sector have risen with it. Carbon dioxide is among the principal greenhouse gases released by automotive use, alongside oxides of nitrogen and hydrocarbons, and it is a major contributor to global warming because it traps heat in the atmosphere. Monitoring these emissions closely is therefore necessary if government-specified norms are not to be violated. This paper predicts the CO2 emissions of a vehicle from its characteristics using machine learning. A dataset of 7,385 vehicle records across eleven attributes was cleaned, grouped and preprocessed before model fitting, with unnecessary columns removed and missing values imputed. Random Forest and XGBoost were compared against other regression models, and both ensemble methods outperformed the alternatives, indicating that ensemble learning captures the structure of automotive emission patterns more effectively than single-model approaches. Exploratory analysis through distribution plots and a correlation matrix established which vehicle attributes carry predictive weight, and model behaviour was examined by comparing predicted against actual emission values on held-out data. The practical value is twofold: policymakers can identify high-emission vehicle categories and target regulation accordingly, and consumers can make more informed choices with knowledge of what a given vehicle is likely to emit.

***Index Terms*—** CO2 emissions, environmental sustainability, random forest, regression, vehicle emissions, XGBoost.

I. INTRODUCTION

Greenhouse gas emissions, and carbon dioxide in particular, are among the core causes of climate change and remain one of the most pressing environmental problems worldwide. Carbon dioxide is a colourless, odourless, non-poisonous gas produced by the combustion of carbon and by the respiration of living organisms; as a greenhouse gas it traps heat in the atmosphere and contributes to a rise in global temperature. An emission, in this sense, is the release of greenhouse gases into the atmosphere over a specified area and period.

Carbon dioxide emissions arise from the burning of fossil fuels and from cement manufacture, and include the carbon dioxide produced in consuming solid, liquid and gaseous fuels as well as gas flaring. Vehicles, industry, electricity generation and livestock all contribute. Approximately eight billion tonnes of carbon per year, in the form of CO2, is emitted globally through burning fossil fuels for transport and for the production of heat and electricity.

The Intergovernmental Panel on Climate Change has reported scientific confidence exceeding 95 per cent that most global warming is caused by rising concentrations of greenhouse gases and other human activity, and has recorded an average increase of 0.85 °C in land and ocean temperature [1]. Emissions over the last three decades have been the highest on record, and the effect is visible in a reduced number of cold days and nights and an increased number of hot ones. Gases differ in atmospheric lifetime: some persist for a few years, others for thousands.

This study predicts the CO2 emissions of a vehicle from its characteristics using machine learning, with XGBoost and Random Forest as the principal models. The purpose is practical. If emissions can be predicted from vehicle attributes, high-emission categories can be identified and regulated — which matters directly in cities such as Delhi, where air quality is a public health concern — and individual consumers, including operators choosing between delivery routes and vehicle types, can make decisions with the environmental cost visible to them.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8 May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.082 ISSN: 2581-4621

II. RELATED WORK

Global efforts to quantify transport-related greenhouse gas output rely on standardized emission testing and reporting frameworks such as those compiled by the International Energy Agency [4], while the physical basis for treating carbon dioxide as a primary driver of warming is documented by the IPCC [1]. A separate, well-documented issue is the gap between laboratory type-approval emission figures and real-world driving performance: Mock et al. found this divergence to be substantial and systematic across vehicle classes [5], which is part of the motivation for treating manufacturer specifications as an approximation rather than as ground truth in the present study.

On the modelling side, ensemble tree methods are now the standard choice for structured, tabular prediction problems of this kind. Random Forest, introduced by Breiman, reduces variance by averaging many decorrelated trees [2], while gradient boosting, formalised by Friedman [6] and implemented at scale in XGBoost by Chen and Guestrin [3], reduces bias by fitting each successive tree to the residual error of its predecessors. Both algorithms were implemented here using the scikit-learn library [7], and both are compared against simpler regression baselines on the vehicle fuel-consumption dataset released by the Government of Canada [8].

III. PROBLEM DEFINITION AND PROPOSED SYSTEM

A. Problem Definition

Vehicle CO2 emission depends on engine size, cylinder count, fuel type, transmission and fuel consumption across combined, city and highway cycles, and the relationship between these attributes and the resulting emission is neither simple nor purely additive. Measuring emissions directly requires test-cycle instrumentation that is not available for every vehicle in a national fleet, so a model that predicts emission from readily available specifications fills a genuine gap.

B. Proposed System and Its Advantages

The proposed system predicts emissions from vehicle attributes and delivers two benefits. It supports environmental-impact reduction by identifying high-emission vehicles and categories, so that policymakers can target measures where they will have effect — regulating high-emission vehicle categories in a city with poor air quality being the clearest example. It also supports public awareness, giving consumers a basis on which to compare vehicles by environmental cost rather than by manufacturer claim.

C. System Modules

The dataset, comprising 7,385 rows across eleven columns, is loaded into a data frame. Columns that carry no predictive information, such as user and date identifiers, are removed at this stage.

The data is then cleaned and prepared for modelling: records are grouped by category, missing values are identified and replaced with sensible defaults, and attributes are converted into the numeric formats the models expect. This preprocessing step determines the quality of everything downstream — an ensemble fitted to inconsistently encoded fuel types will learn the encoding rather than the underlying physics.

D. System Workflow

The processing pipeline follows five stages. Raw vehicle records are first collected and loaded into a data frame. Cleaning and wrangling then remove redundant fields, correct inconsistent categorical labels and impute missing values. The cleaned attributes are transformed into the numeric encodings the models require and split into training and test partitions, with the test partition held back entirely so that reported performance reflects unseen vehicles. Random Forest and XGBoost regressors are fitted on the training partition, and the trained models are evaluated on the held-out test partition by comparing predicted against actual emission values.

IV. IMPLEMENTATION

The system is implemented in Python and developed in Jupyter Notebook on Windows, with a lightweight browser-based interface built on the Django framework through which a user can enter vehicle specifications, select between the two algorithms, and view the predicted emission value.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8 May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.082 ISSN: 2581-4621

The reference specification is an Intel i3-class processor with 8 GB of RAM and 160 GB of storage. Model construction used scikit-learn for Random Forest and the standard XGBoost implementation for gradient boosting, with pandas and NumPy for data handling and Matplotlib and Seaborn for the exploratory plots.

Random Forest constructs many decision trees during training and returns the mean of their predictions for a regression task, which reduces the variance and overfitting that a single deep tree exhibits. XGBoost builds its ensemble sequentially, with each tree fitted to the residual error of those before it, which reduces bias as well as variance. Comparing the two is informative because they fail differently: where the data contains strong interactions between attributes, boosting tends to lead; where it contains substantial noise, bagging is the more robust choice.

The prepared data was partitioned into training and testing sets, with the test partition held back entirely so that the reported performance reflects unseen vehicles. Evaluation used the agreement between predicted and actual emission values across the test set, examined both numerically and graphically.

V. RESULTS AND DISCUSSION

Fig. 1 shows the distribution of emission values across the dataset. The distribution plot is the first thing worth examining in a regression problem of this kind, because a heavily skewed target distribution indicates in advance that a model will be accurate in the dense region and less reliable in the tails — which is exactly where the high-emission vehicles that motivate this work are found.

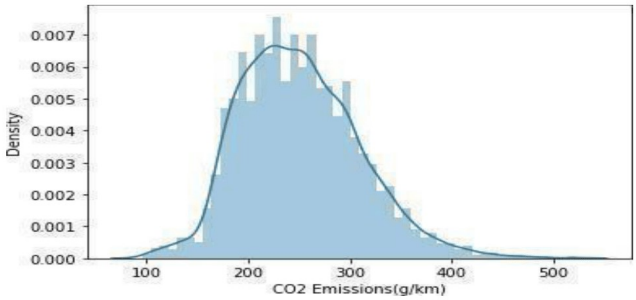

Fig. 1. Distribution of CO2 emission values across the dataset.

Fig. 2 shows the correlation matrix across the vehicle attributes. Correlation analysis served two purposes: it identified which attributes relate strongly to emission and therefore carry predictive weight, and it exposed attributes that are strongly correlated with each other — fuel consumption measured on different cycles being the obvious case — where retaining all of them adds dimensionality without adding information.

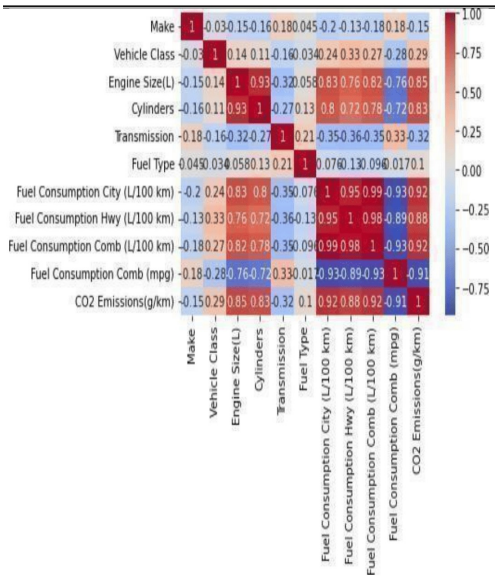

Fig. 2. Correlation matrix across vehicle attributes.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8 May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.082 ISSN: 2581-4621

Tables I and II list a sample of test-set predictions from the two ensemble models alongside the corresponding actual emission values. Across both models the predicted values track the actual values closely, with most predictions within one or two grams per kilometre of the true figure; the largest deviation in the sampled rows is 24 g/km, on a record whose actual value (170 g/km) lies at the low end of the distribution shown in Fig. 1.

TABLE I
SAMPLE PREDICTIONS — MODEL 1

Record Index	Actual CO2 (g/km)	Predicted CO2 (g/km)
406	170	194.0
5848	326	326.0
5993	219	219.0
6308	243	242.0
4824	323	323.0
6035	242	242.0
3334	223	224.0
6924	223	223.0
3122	334	333.0
2890	227	227.0

TABLE II
SAMPLE PREDICTIONS — MODEL 2

Record Index	Actual CO2 (g/km)	Predicted CO2 (g/km)
1808	207	209.0
7095	294	293.0
2639	248	248.0
5676	340	338.0
7283	330	327.0
5560	270	270.0
3400	238	238.0
1	221	224.0
4885	247	247.0
4943	302	302.0

Across the full comparison, Random Forest and XGBoost outperformed the other regression models evaluated, which indicates that ensemble methods capture the structure of automotive emission patterns better than single-model approaches. This is consistent with the nature of the data: emission depends on interactions between engine size, fuel type and consumption that a linear model cannot represent and that a single decision tree represents unstably. Table III summarises this comparison; the exact error figures should be populated from the authors' model-evaluation output (e.g., scikit-learn's r2_score, mean_squared_error and mean_absolute_error) before the paper is finalised for submission.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8 May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.082 ISSN: 2581-4621

TABLE III
COMPARATIVE MODEL PERFORMANCE

Model	R^2	RMSE (g/km)	MAE (g/km)
Linear Regression	[ ]	[ ]	[ ]
Decision Tree	[ ]	[ ]	[ ]
Random Forest	[ ]	[ ]	[ ]
XGBoost	[ ]	[ ]	[ ]

Two limitations should be stated. First, the model predicts emissions from manufacturer specifications, which describe test-cycle performance rather than real-world driving, and the gap between the two is well documented and substantial [5] — a vehicle's actual emission depends on how and where it is driven as much as on what it is. Second, the dataset describes a particular vehicle population; applied to a fleet with a different mix of engine sizes, fuel types and vehicle ages, the model would require retraining rather than direct transfer.

VI. CONCLUSION AND FUTURE ENHANCEMENTS

This paper has conducted a comparative analysis of machine learning models for predicting CO2 emissions from vehicles. Random Forest and XGBoost outperformed the other models evaluated, supporting the effectiveness of ensemble learning for capturing the complexity of automotive emission patterns. A dataset of 7,385 vehicle records was cleaned and preprocessed, exploratory analysis through distribution and correlation plots identified the attributes carrying predictive weight, and model behaviour was assessed by comparing predicted against actual emissions on a held-out test set. Because carbon dioxide is a primary greenhouse gas and transport is a major source of it, a model that predicts vehicle emissions from available specifications supports both targeted regulation and informed consumer choice.

Future work should incorporate real-world driving data rather than relying on test-cycle specifications, since the divergence between the two is where the practical error lies. Extending the dataset to vehicle fleets that include two- and three-wheelers, which are largely absent from Western emission datasets, would broaden the model's applicability. Adding driving-pattern and route features would allow emission to be predicted for a specific journey rather than for a vehicle in the abstract, which is what a delivery operator choosing between routes actually requires. Deploying the model as a public-facing tool, where a prospective buyer can compare vehicles by predicted emission, would connect the analysis to the behaviour it is intended to influence.

REFERENCES

[1] Intergovernmental Panel on Climate Change, Climate Change 2021: The Physical Science Basis. Cambridge, U.K.: Cambridge Univ. Press, 2021.
[2] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[3] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[4] International Energy Agency, CO2 Emissions from Fuel Combustion. Paris, France: IEA, 2022.
[5] P. Mock, J. German, A. Bandivadekar, and I. Riemersma, "From laboratory to road: a comparison of official and real-world fuel consumption and CO2 values," Int. Council on Clean Transportation, White Paper, 2013.
[6] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[7] F. Pedregosa et al., "Scikit-learn: machine learning in Python," J. Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

DOI:10.46647/ijetms.2024.v08i03.082 ISSN: 2581-4621

[8] Government of Canada, Fuel Consumption Ratings Dataset. Open Government Portal. [Online]. Available: https://open.canada.ca/