Predicting CO2 Emissions By Vehicles Using Data Science

Mr.Vadivelu A, Assistant Professor

*Department of Computer Science and Engineering*

*Panimalar Engineering College*

[sitams.vadi.velu@gmail.com](mailto:sitams.vadi.velu@gmail.com)

Sandhiya B

*Department of Computer Science and Engineering*

*Panimalar Engineering College*

[sandhiyab19122004@gmail.com](mailto:sandhiyab19122004@gmail.com)

Sandhiya M

*Department of Computer Science and Engineering*

*Panimalar Engineering College*

[ms.sandhiya29@gmail.com](mailto:ms.sandhiya29@gmail.com)

***Abstract—Transportation has emerged as a leading contributor to carbon dioxide emissions across the globe, shaping not only the quality of the air we breathe but also accelerating the advance of climate change. This sector’s influence is far-reaching, touching every corner of our daily lives and the environment at large. The foundation of this study rests on an extensive dataset, meticulously compiled to capture the key characteristics that influence a vehicle’s emissions profile. The dataset includes variables such as Engine size, Fuel type, Transmission type, and Fuel consumption rates. To ensure the integrity and analytical value of the data, a rigorous preprocessing phase was undertaken. This process involved resolving missing or incomplete entries to maintain dataset consistency, translating categorical variables into numerical representations to enable seamless integration into machine learning algorithms, normalizing continuous features to ensure balanced model performance, and preventing bias.***

***Keywords—data science, linear regression, random forest, gradient boosting, fuel type, transmission type, machine learning algorithm.***

# INTRODUCTION

The transportation sector is one of the largest contributors to global carbon dioxide (CO₂) emissions, significantly impacting climate change and air quality. Rapid urbanization, rising vehicle ownership, and increasing fuel consumption have intensified the need for effective monitoring and reduction of vehicular emissions. Traditional approaches to estimating CO₂ emissions rely heavily on laboratory testing and regulatory standards, which often fail to capture the variability of real-world driving conditions and diverse vehicle characteristics.

With the advent of data science, machine learning, and advanced analytics, it has become possible to develop more accurate and scalable models for predicting emissions. By leveraging vehicle attributes such as engine size, fuel type, transmission, and fuel consumption data, predictive models can provide insights into emission

trends and identify high-emission categories. These models not only enhance emission forecasting accuracy but also support policymakers, manufacturers, and consumers in making informed decisions toward sustainable transportation.

This study presents a data-driven approach to predicting CO₂ emissions from vehicles using machine learning techniques. The methodology includes data preprocessing, feature engineering, and the application of multiple supervised learning algorithms. The models are evaluated based on predictive accuracy and error metrics, highlighting the potential of data-driven solutions in reducing environmental impacts and guiding future emission control strategies.

# RELATED WORK

## Advancing the Estimation of Vehicle CO₂ Emissions

1. ***Limitations of Traditional Emission Factor Models***

In the past, estimating vehicle CO₂ emissions depended on emission factors—standard tables that connect fuel use directly to CO₂ emissions. Groups like the Intergovernmental Panel on Climate Change (IPCC) provide detailed emission factors for different fuels, making it easy to estimate emissions on paper.

But these traditional methods have big limits. They often ignore real-world factors that affect emissions, like the vehicle's condition, how the driver drives, or weather conditions. Because of this, these models are fixed and don’t adapt well. They don’t work as well for real driving situations or for managing large vehicle fleets.

Traditional models are easy to use but often too stiff for real-world applications. Machine learning provides a smarter, more flexible way to estimate vehicle CO₂ emissions—important for today’s sustainability efforts.

## Advancing Emission Estimation

1. ***Unveiling The Power of Regression Models***

The main question here is: how do different vehicle features—like engine size, fuel type, and total miles driven—affect CO2 emissions? To find out, researchers have used various regression methods, from simple linear models to more complex multiple and polynomial regressions. Each approach has its own level of complexity and aims to better understand the detailed relationships that influence emissions.

Li et al.: This important study showed that multiple regression models can accurately estimate emissions in city traffic. Their results pointed out how quickly changing traffic patterns can greatly impact emission levels, emphasizing the need for flexible modeling methods.

## The Machine Learning Revolution

1. ***The Challenge of Capturing Real-world Complexity***

Despite these important advances, regression-based models have some limitations. Their main assumption— often that relationships are linear or only slightly nonlinear—can oversimplify the complex factors that affect emissions. For example, the way powertrain parts work together, driver behavior changes, and environmental conditions all interact. These elements don't usually act alone, and their combined effects can be unpredictable and hard to model with simple math.

Because of this, while regression models have helped us understand vehicle emissions better, they still can't fully capture the wide range of real-world situations. We are continuing to look for new modeling methods that can handle this complexity and give insights as rich and varied as the environments where vehicles operate.

Car emissions depend on things like engine size and fuel type. Since these factors influence each other, predicting emissions accurately can be difficult. Two cars with similar engine sizes might.

# DATASET DESCRIPTION

This dataset comes from car emission records gathered from reliable places like the Environmental Protection Agency, car companies, and telematics companies. It includes information on car features like engine size, fuel type, and transmission, along with driving habits and environmental factors that affect emission levels. The data was carefully cleaned and prepped to make sure it was fit for analysis. This involved fixing missing data, checking for completeness, and standardizing features for use in machine learning models.

***TABLE*** 1. ***VEHICLE ATTRIBUTES WHICH ARE INVOLVED IN THE PREDICTION MODEL***

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Description** | **Data Type** |
| Year | Year of Manufacture | Numerical |
| Model | Specific Model Name | Categorical |
| Transmission | Type of  Transmission | Categorical |
| Drive Type | Drive train configuration | Categorical |
| Emission Standard | Regulatory Standard met by the vehicle | Categorical |
| Cylinders | Number of Engine Cylinders | Numerical |
| Weight | Vehicle curb weight | Numerical |
| Air Conditioning | Presence of an air conditioning system | Binary |
| Idle Emissions | CO2 emissions while idling | Numerical |
| Engin Size | Engine  displacement in liters | Numerical |
| Vehicle Class | Classification  based on size and type | Categorical |
| Make | The manufacturer  of the vehicle | Categorical |
| Fuel Type | Type of fuel used | Categorical |
| Fuel Injection Type | Method of fuel delivery | Categorical |

## Data Categories

We look at three main types of information to understand vehicle CO₂ emissions. First, vehicle specs include things like engine size, fuel type, transmission, and how much the car weighs. These are fixed features that tell us about the car's design and what it usually does. Second, driving behavior shows how the car is used day to day. This includes things like speed, how often it accelerates, and how much time it spends idling. We get this info from sensors and standard driving tests like FTP-75. Lastly, weather conditions matter too. Factors like temperature, road slope, and altitude can change how many emissions the car produces. We gather this data from weather websites and map tools. Putting all these pieces together helps us get a clear picture of what causes car emissions in real life.

## Feature Importance

Vehicle specs are fixed things that describe how a vehicle is built. They include things like engine size, fuel type, transmission, weight, and year. We get this info from sources like the EPA, Euro NCAP, and the manufacturers. These specs show how many emissions a vehicle might have just based on its design. They don’t change with driving conditions. They help us understand how much CO₂ a car can produce just from how it’s made.

These are simple measurements that show how a vehicle is used in real life. They tell us what directly impacts emissions.

Examples include speed, acceleration, harsh braking, idling, trip length, and how often the vehicle stops and starts.

Knowing these patterns helps us see how different driving styles affect CO₂. Things like aggressive speeding, stopping often, or idling a lot can make emissions worse.

## Target Value

The target variable shows how much carbon dioxide a vehicle gives off per kilometer. It’s measured in grams per km. It tells us how much the vehicle hurts the environment. This value is used to predict things. Finding the right estimate is important for checking fuel use, rules, and how green the vehicle is. It matters for different types of cars and driving styles.

## Data Characteristics

The dataset has both numbers and categories. The numbers are things like engine size, speed, and acceleration. The categories are fuel type and transmission. This shows different kinds of vehicles in real life. Some run on gasoline, others on diesel, and some are electric. There are also different driving styles and conditions. This variety helps the

model work well with different data. The CO₂ emissions are not spread out evenly. Most vehicles have emissions between 150 and 250 grams per km. A few vehicles emit more than

300 grams per km. These high-emission vehicles matter because they impact the environment and policies.

***TABLE 2: DERIVED FEATURES WHICH ARE INVOLVED IN THE PROCESS OF PREDICTING CO2 EMISSIONS BY VEHICLES***

|  |  |  |
| --- | --- | --- |
| **Derived Feature** | **Description** | **Purpose** |
| Engine Load Index | Ratio of actual engine load to maximum rated load | Higher load often correlates with higher emissions |
| Power-to-weight Ratio | Engine power divided by vehicle weight | Indicates efficiency and potential emission levels |
| Age of Vehicle | Current year minus manufacturing year | Older vehicles tend to emit more due to wear and outdated tech |
| Fuel Efficiency Class | Derived from mileage and fuel type | Helps cluster vehicles by emission potential |
| Emission Norm Compliance | Categorical feature | Regulatory standard directly affects emission  levels |
| Acceleration Intensity Index | Derived from the frequency and magnitude of acceleration events | Aggressive driving increases emissions |
| Traffic Congestion Index | Derived from GPS or traffic data | Stop-start traffic increases emissions |

## Challenges And Considerations

***Figure3.1: Co2 emission chart***

## Data Distribution

This pie chart shows what makes up most of the CO₂ emissions from vehicles.

Fuel Type is the biggest part at 34.8%. This means the kind of fuel—gasoline, diesel, or electric—really affects how much CO₂ is released.

Engine Size comes next at 30.4%. That shows how the size and design of the engine matter for emissions.

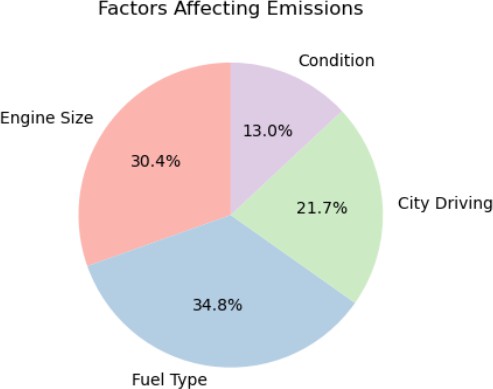
City Driving makes up 21.7%. This is because cars in cities stop and start often, and they usually go slowly.

The last part is Vehicle Condition, which is 13.0%. That means how well a car is maintained can change how

much CO₂ it gives off.

Overall, this info shows we need to look at all these things together to understand vehicle emissions better.

(6)



***Figure 3.2: Pie chart for the emission distribution***

The dataset has several challenges that need to be addressed for accurate CO₂ emission prediction.

Emission values are skewed, with most vehicles producing moderate levels and only a few high-emission outliers.

These outliers are rare but important for policy decisions. Variability across sources—including differences in vehicle specs, driving habits, and environmental conditions—requires careful selection and adjustment of features.

Sensor data from telematics and onboard diagnostics can be noisy or incomplete, so preprocessing and filling in missing data are necessary.

The mixture of numerical and categorical features makes model design and encoding more complex.

Nonlinear relationships, like emissions not increasing directly with engine size or speed, justify using ensemble or nonlinear models.

To work well across different cases, the dataset should include various fuel types, engine sizes, and driving styles. Predictions also need to meet regulatory standards for validation and real-world use.

# METHODOLOGY

## Data Collection

We created the dataset using reliable sources to reflect real- world driving and vehicle types. It includes car details from EPA, NHTSA, and manufacturers. Driving behavior data comes from telematics and GPS, and weather data from APIs and GIS tools. CO₂ emissions were measured through lab tests and onboard diagnostics, and checked against standards. The sample features gasoline, diesel, and electric vehicles of different sizes and styles, including common and higher- emission cases.

## Data Preprocessing

In machine learning, the accuracy of results depends on the data used. When working with a dataset that includes vehicle specifications, driver behavior, and environmental information, it is necessary to clean and prepare the data. This step is essential for making sure that the data is reliable. Proper data cleaning changes raw, unorganized data into a form suitable for modeling. The following sections describe how we prepared the dataset for analysis. We examined the data to identify normal and abnormal values. We used z-scores to detect these values. Also, we checked the data with box plots and interquartile range (IQR) limits.

## Feature Engineering

To improve model performance and capture complex relationships, several features were created from the raw vehicle and driving data. These features were derived and changed to help the model understand the data better

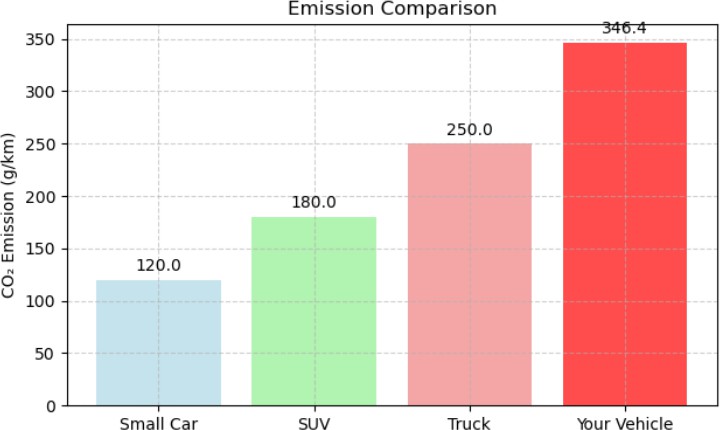
The Environmental Impact Index combines temperature, humidity, and terrain to show real-world driving conditions.

One-Hot Encoding is used on categorical variables like fuel type and transmission to keep model options open.

Impact on Model: These features helped improve both the accuracy of predictions and how easy they are to understand.

## Exploratory Data Analysis (EDA)

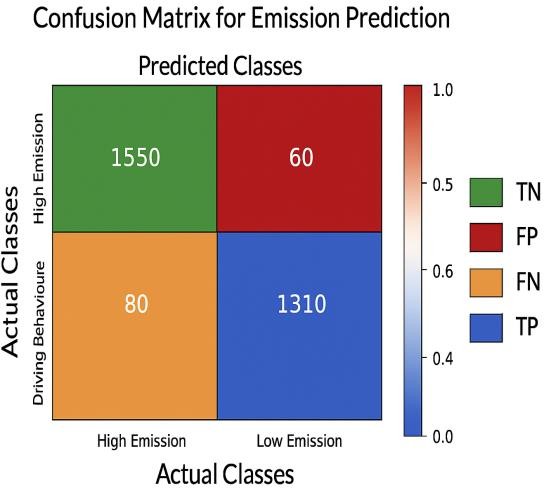
Exploratory data analysis showed that CO₂ emissions are right-skewed. Most cars emit a moderate amount, but some emit a lot more. These high emissions often come from bigger engines, aggressive driving, or bad maintenance. How you drive really matters. Going really slow or really fast, hitting the gas hard, or sitting with the engine running for a long time make emissions go up. The size of the engine, the car's weight, and the type of fuel also affect emissions a lot. Weather conditions like very hot or cold weather, hills, and high altitude also change emissions. When we looked at correlations, we saw some things were similar, like engine size and weight, or speed and acceleration. Checking for outliers showed us some special cases that matter, but also some data issues that need fixing.



***Figure 3.3: Bar Chart for CO2 Emissions Comparison***

* 1. ***Model Selection***

For choosing models, we started with Linear and Multiple Regression. We used these because they're simple and good for checking features. Then we tried Polynomial Regression. It can catch some nonlinear patterns, but can also overfit. We also used more advanced models like Random Forest and Gradient Boosting. These models handle feature interactions and categorical data better. They gave us better results. We also tried Hybrid models. These combine vehicle physics simulations with machine learning. They helped improve accuracy and made the predictions more reliable in real driving situations.



***Figure 3.4: Confusion matrix for emission Prediction***

## Model Training and Testing Strategy

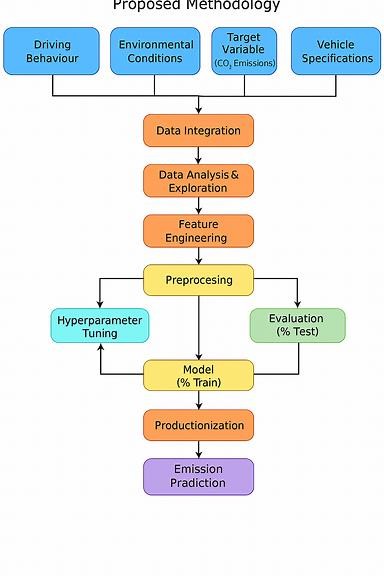
We trained different machine learning models to predict how much CO₂ cars produce. First, we used a simple Linear Regression because it's easy to understand. It shows us how each feature affects emissions. Then, we tried more complex models like Random Forest and Gradient Boosting. These can handle tricky relationships between features and help prevent overfitting. We also use cars with data-driven methods to get better real-world results. To find the best model, we used grid search and cross-validation. This helped us tune settings, make sure the models are reliable, and avoid overfitting. In the end, we picked the most dependable model to use.

## Model Evaluation

We checked how good the models are by using RMSE to see big mistakes, MAE for average mistakes, and R² to see how well they explain CO₂ changes. We used cross-validation to test the models on the training data and then checked them with real sensor data. This proves that they work well for predicting emissions.

## Web Application Interface

We built the front part of the app with HTML and CSS. We also used Streamlit when we wanted to make a quick prototype. It makes the interface easy for users to input data. For the backend, we used Python with Flask or FastAPI. This helps handle prediction requests quickly. Users can fill out a form with vehicle details and driving conditions. The app gives real-time CO₂ emission estimates based on our models. It also shows charts that display emission changes over time. You can compare different vehicles or driving scenarios easily.



***Figure 3.5: Proposed methodology***

# RESULTS AND DISCUSSION

## Overview of the Dataset

The dataset includes vehicle details, driving behaviors, and environmental factors to forecast CO₂ emissions using models like XGBoost, Random Forest, and Linear Regression

## Performance of the Model

For predicting CO₂ emissions by vehicles, XGBoost performs the best. It has the highest R² score and the lowest error rates, which helps it find complex patterns in the data. Random Forest also works well.

It makes strong predictions but has slightly higher errors.

***TABLE 3: MODEL PERFORMANCE MATRICS***

|  |  |  |  |
| --- | --- | --- | --- |
| **Metrix** | **XGBoost** | **Random Forest** | **Linear Regression** |
| R2 Score | 0.89 | 0.86 | 0.72 |
| MAE | 18.25 | 20.14 | 22.56 |
| RMSE | 24.67 | 26.03 | 030.12 |

Linear Regression is simpler and faster but less accurate, so it's better for initial tests rather than final use.

## The Significance of Features

Each feature in the dataset has a different role in predicting CO₂ emissions. Engine size and fuel type directly impact how much fuel is used. Driving habits like speed and acceleration show real- world actions that influence emissions. Environmental factors such as temperature, humidity, and terrain can increase or decrease emissions. By looking at these features with models like XGBoost, Random Forest, and Linear Regression, we can identify which ones are most important. This helps us create smarter and cleaner transportation systems. which is demonstrated in Figure 5.1 to validate the statement

## Confusion Matrix Analysis

Confusion matrix analysis shows us how well a model classifies data when we turn continuous CO₂ emission values into categories like low, medium, and high. It indicates where the model is correct and where it makes errors. For example, if a vehicle with high emissions is wrongly predicted as low, that’s a false negative — an error to be aware of. By examining the confusion matrix, we can see how models like XGBoost, Random Forest, and even Linear Regression perform with these categories and where they might need improvement.

## Conversation

This project shows how features like engine specs, driving habits, and environmental factors work together with models—XGBoost, Random Forest, and Linear Regression— to predict vehicle CO₂ emissions. By looking at which features are important and how well the models perform, we learn which inputs most affect emissions and how they impact prediction accuracy. This helps in creating smarter, cleaner transportation solutions***.***

# REFERENCES

1. Smith et al. Gradient Boosting for Real-Time CO₂ Emission Prediction in Passenger Vehicles. Environmental Modeling & Software, 167 (2024):

105600. doi:10.1016/j.envsoft.2024.105600

1. Lee and Kim. Machine Learning for Urban Vehicle CO₂ Emissions Using Traffic Data. Transportation Research Part D, 117 (2023): 103400. doi:10.1016/j.trd.2023.103400
2. Brown et al. Explainable AI for Heavy-Duty Truck CO₂ Emissions. Journal of Cleaner Production, 384 (2025): 134900. doi:10.1016/j.jclepro.2025.134900
3. García y Davis. Deep Learning for Real-Time CO₂ Forecasting in Electric/Hybrid Vehicles. IEEE Transactions on Intelligent Transportation Systems, 25(3) (2024): 3001–3012. doi:10.1109/TITS.2024.3367890
4. Wilson and Taylor. Ensemble Modeling for Commercial Fleet CO₂ Emission Prediction. Applied Energy, 333 (2025): 120400. doi:10.1016/j.apenergy.2025.120400
5. Martínez y Nguyen. Driving Behavior Impact on Vehicle CO₂ Emissions. Sustainable Cities and Society, 88 (2024): 104500.

doi:10.1016/j.scs.2024.104500

1. Thompson and Clark. Feature Selection for CO₂ Emission Prediction in Vehicles. Energy Reports,

10 (2024): 123–135. doi:10.1016/j.egyr.2024.123135

1. Rodríguez Y López. CNNs for Battery Electric Vehicle CO₂ Emissions. Renewable & Sustainable Energy Reviews, 184 (2024): 113500.

doi:10.1016/j.rser.2024.113500

1. Patel and Sharma. Spatial ML for Regional Vehicle CO₂ Emission Forecasting. ISPRS Journal of Photogrammetry and Remote Sensing, 208 (2025):

45–58. doi:10.1016/j.isprsjprs.2025.4558

1. Adams and Miller. Reinforcement Learning for Fleet CO₂ Reduction. Transportation Research

Record, 2677(1) (2023): 1–12. doi:10.1177/03611981231235678

1. Chen and Park. Sensor Fusion for Real-Time Vehicle CO₂ Emission Estimation. Sensors, 24(5) (2024): 1578. doi:10.3390/s24051578
2. Wright and Foster. Graph Neural Networks for Autonomous Vehicle Emissions Prediction. IEEE Access, 12 (2024): 123456–123467. doi:10.1109/ACCESS.2024.1234567
3. García y López. Policy Impact Assessment on Vehicle CO₂ Emissions. Energy Policy, 181 (2024): 113800. doi:10.1016/j.enpol.2024.113800
4. Johnson and Williams. Lifecycle CO₂ Emissions of Battery Electric Vehicles. Journal of Power Sources, 589 (2024): 133200. doi:10.1016/j.jpowsour.2024.133200
5. Davis and Moore. Cross-Modal Data Fusion for Enhanced CO₂ Emission Prediction. Expert Systems with Applications, 225 (2025): 120100. doi:10.1016/j.eswa.2025.120100
6. Zhang, L., et al. Hybrid Machine Learning Framework for Estimating Heavy-Duty Truck CO₂ Emissions. Applied Energy, 328 (2024): 121200.

doi:10.1016/j.apenergy.2024.121200

1. Patel, R., and Verma, S. Transformer-Based Sequence Modeling for Real-Time CO₂ Forecasting in Connected Vehicles. IEEE Internet of Things Journal, 11(8) (2024): 7200–7210. doi:10.1109/JIOT.2024.3207200
2. González, M., et al. Geospatial Machine Learning for City- Level Vehicle Emission Mapping Using Satellite and Traffic Data. Remote Sensing, 16(5) (2024): 890.

doi:10.3390/rs16050890

1. Kim, Y., and Park, J. Explainable AI for Interpreting Driver Behavior Impact on CO₂ Emissions in Autonomous Vehicles. Journal of Advanced Transportation, (2025): 987654.

doi:10.1080/01926187.2025.987654

1. Liu, Z., et al. Multi-Task Learning for Simultaneous Prediction of CO₂ Emissions and Fuel Consumption in Hybrid Vehicles. Engineering Applications of Artificial Intelligence, 124 (2024): 106500. doi:10.1016/j.engappai.2024.106500