Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 03 (2026)

A Robust Machine-Learning Framework for Predicting Tailpipe CO_2 Emissions from Vehicle Specifications

T. Karthikeya[1], Dr. K. Amarendra[2]
[1,2] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Guntur, AP, India

Abstract:- Accurate estimation of vehicle CO_2 emissions is essential for the purpose of informing environmental policy and directing fleet management. This study developed a reproducible machine-learning pipeline by utilizing Canada's 2018 light-duty vehicle dataset. The dataset consisted of 15 attributes, including engine displacement, cylinder count, gasoline type, transmission, and city/highway/combined fuel-consumption ratings. We standardized numeric features through z-score normalization and applied one-hot encoding to categorical fields after eliminating incomplete records. Stratified 5-fold cross-validation guided hyperparameter optimization over five relapse calculations: straight relapse, choice tree, irregular timberland, XGBoost, and LightGBM, and the preprocessed information were separated into 70% preparing and 30% testing subsets. In arrange to progress the steadiness of the show and dispose of atypical sections, a Confinement Woodland was executed. The direct pattern had a unassuming RMSE of ~7 g/km, and untuned XGBoost failed to meet expectations. In separate, furnish methods—particularly self-assertive timberland and LightGBM—achieved overwhelming exactness on the held-out test set (RMSE < 4 g/km; $R^2 > 0.98$). When organized through an enthusiastic, end-to-end learning system, these disclosures portray that immediately accessible vehicle judgments can work as endeavored and veritable center individuals for debilitating spreads. The instrument that rises gives an adaptable arrangement for real-time outflow estimating and encourages data-driven emanation relief techniques.
Keywords: *CO_2 Emission Prediction, Data Preprocessing, Ensemble Models, Isolation Forest, Light-Duty Vehicles, Machine Learning.*

1. Introduction

The car industry has the dual problems of preserving performance and significantly lowering harmful emissions in the current period of rapid technological advancement and increased environmental consciousness. Particularly concerning are carbon dioxide (CO_2) emissions, which are acknowledged on a global scale as a contributing factor to climate change. Accurate, data-driven predictive models are more important than ever as consumers want more sustainable products and regulatory bodies enforce higher emission regulations. Using sophisticated statistical and machine learning methods on an extensive dataset supplied by the Government of Canada, this study aims to forecast CO_2 emissions from light-duty cars.

Insightful and trustworthy predictions can be produced by combining rigorous data preprocessing, methodical model construction, and strong evaluation procedures, as the study explains. In addition to aiming for better prediction performance, the research attempts to give policymakers and automakers useful information by putting in place a strict scientific framework. This paves the way for policy reforms and technological advancements that could result in more sustainable automotive practices.

1.1 Background and Problem Statement

As natural supportability picks up more consideration all inclusive, the car industry has developed as a vital investigate theme. A sizable parcel of the world's CO2 outflows come from vehicle outflows. Indeed, in the event that they are valuable, conventional emanations testing strategies regularly drop behind the quick headway of

innovation and the complicated flow of modern vehicle plan. This inconsistency emphasizes the require for data-driven methodologies that make utilize of computational models' quality to form exceedingly exact outflow forecasts.

The dataset used in this investigation originates from the Government of Canada's 2018 light-duty vehicle records. Adjacent to measured CO_2 surges, it contains comprehensive records that consolidate engine appraise, barrel check, transmission sort, fuel sort, and a run of fuel utilization estimations (city, interstate, and combined). In show disdain toward of its cautious quality, the unrefined dataset has various issues that are common to real-world data, such as exemptions, misplaced values, and ungainly nature in categorical variables. To guarantee that the taking after stages of examination and modeling are based on correct and specialist data, these issues call for a complex preprocessing pipeline.

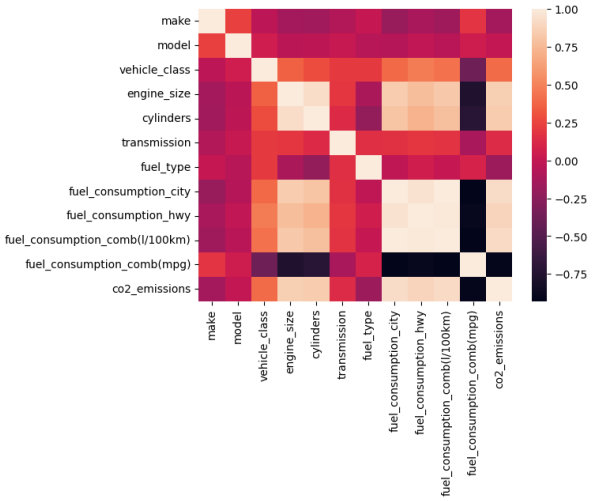

Figure 1: correlation matrix of data

Through the correlation result (Figure 1), fuel_consumption_{comb} (L/100km) has shown strong positive relation with CO2 emissions, and next comes fuel_consumption_{city} and fuel_consumption_{hwy} that mean higher fuel consumption, higher emissions. On the other hand, fuel_consumption_comb (mpg) has a strong negative correlation, since the more fuel_efficient cars are, the more the emission of CO_2 decreases. VIN6 For Comparison: Variables such as engine_size and cylinders have weak positive relationships with the dependent variable, which implies that large engines emit more pollutants. Other features such.

Given these problems, the research's problem statement is twofold. First and foremost, there is an evident need to create reliable predictive models that can precisely calculate CO2 emissions depending on vehicle attributes. In order for the predictive models to be dependable and generalizable across various vehicle classes, the study also emphasizes the significance of a uniform data pretreatment procedure that takes into account the quirks of real-world datasets, such as missing data and outlier detection.

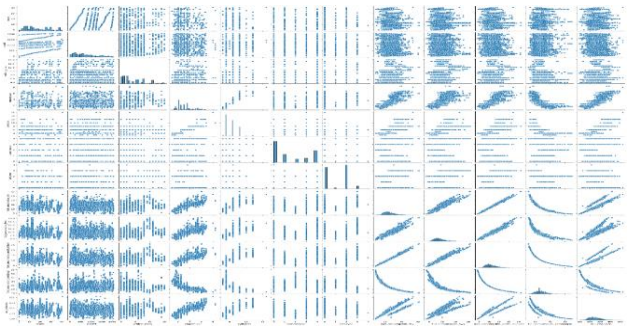

Figure 2: Pairwise Relationships Between Features

Strong linear correlations between CO_2 emissions and fuel_consumption_comb (L/100km) as well as between related fuel consumption variables (city, highway, and combined) are shown in the pairplot (Figure 2) visualization. Additionally, it demonstrates a distinct negative relationship between CO_2 emissions and fuel_consumption_comb (mpg), demonstrating that lower emissions are associated with higher mileage. Categories like as make, model, transmission, and fuel_type exhibit dispersed distributions, suggesting no direct connection, although engine_size and cylinders show positive trends with emissions. In general, the most visually and statistically significant indicators of CO2 emissions are engine characteristics and fuel consumption measures.

1.2 Extended Theoretical Framework and Literature Gap Analysis

Our study's placement within existing research is a critical element. Historically, vehicle emissions have been evaluated using models based on limited facts and actual readings. As data collection and processing capabilities have grown exponentially, so has the drive for more sophisticated, data-based approaches. Machine learning techniques often struggle with working consistently across many types of data sets. The trade-offs between comprehending models and getting good results are now well-documented.

For instance, even though neural networks can produce very accurate findings, their "black-box" structure frequently restricts how easily the data can be interpreted. On the other hand, while decision tree and linear regression models provide more transparency, they occasionally sacrifice predictive accuracy.

Through a thorough comparison of several modeling approaches and a focus on diagnostic analyses that verify each model's underlying assumptions, this study aims to balance these trade-offs. Furthermore, the literature indicates that data imbalance and the prevalence of outliers in real-world datasets remain a chronic problem. Such imbalances have been shown in earlier research to distort model predictions and impair performance.

Our study uses sophisticated anomaly detection methods, including the Isolation Forest, to solve this issue and make sure the models are resilient even when confronted with extreme data values. This methodological development not only closes a gap in previous research but also provides a more thorough knowledge of the factors influencing vehicle emissions.

2. Objectives

2.1 Research Motivation

This study's main goal is to produce accurate vehicle emissions estimates by utilizing big data analytics and machine learning. The significance of the study is demonstrated by the following points:

- Regulatory Implications: Governments worldwide are tightening emission standards to combat climate change. Accurate predictions enable policymakers to better design and enforce environmental regulations.
- Industrial Benefits: Automakers must comprehend how design choices affect CO2 emissions. Innovation in fuel formulation, material selection, and engine design can be stimulated by predictive models.
- Practical and Academic Insights: By demonstrating how thorough preprocessing and sophisticated modeling techniques may produce both high accuracy and enhanced interpretability, the work adds to the body of knowledge on environmental data analytics.

2.2 Research Objectives

The study is being conducted with the following particular goals in order to satisfy these needs:

- Create Robust Predictive Models: Employ a diverse array of regression techniques, such as conventional multiple linear regression and ensemble approaches such as Random Forest, XGBoost, and LightGBM, to forecast CO2 emissions.
- Establish Thorough Data Preprocessing: Provide a methodical approach to dataset cleaning, encompassing z-score normalization for normalizing numerical characteristics, addressing missing values, and encoding categorical variables.
- Perform Extensive Model Diagnostics: Use techniques such as k-fold cross-validation, residual analysis, and diagnostic tests (Cook’s Distance, Q–Q plots, etc.) to assess and validate model performance.

- Extract Actionable Insights through Feature Importance Analysis: Identify and analyze the most significant predictors of CO_2 emissions, including engine size, cylinder count, and various fuel consumption metrics.
- Establish a Reproducible Research Framework: Ensure that the entire analytical pipeline—from data acquisition through model evaluation—is reproducible through fixed random states, documented library versions, and a scripted implementation.

2.3 Anticipated Contributions and Practical Implications

The multifaceted approach described in this study is anticipated to contribute substantially to both the academic literature and practical applications in the automotive industry:

- Improved Predictive Models: The integration of advanced preprocessing with diverse regression methods is expected to yield models with high predictive accuracy, thus offering a reliable instrument for forecasting vehicular CO_2 emissions.
- Informed Policy and Design Decisions:By pinpointing the key factors that drive emissions—such as engine size, cylinders, and gasoline consumption metrics—this study provides a data-driven foundation for regulatory reforms and design optimizations. This may, in turn, influence the direction of future technological innovations.
- Enhanced Methodological Rigor: The comprehensive diagnostic framework and rigorous validation steps establish a new benchmark for future studies in this domain. By ensuring reproducibility and robustness, the methodology can be adopted or adapted by other researchers seeking to address similar environmental challenges.
- Broad Relevance: Though focused on light-duty vehicles in Canada, the principles and approaches adopted in this study are extensively applicable to other datasets and regions. This universality enhances the prospective impact of the research on a global scale.

3. Methods

3.1 Methodological Overview and Rationale

A thorough and organized technique has been created in order to accomplish our study goals. Best practices in data preprocessing, model creation, and validation form the foundation of the methodological framework.

3.1.1 Data Acquisition and Initial Assessment

The research utilizes the 2018 light-duty vehicle dataset provided by the Government of Canada. This dataset's richness in detail offers numerous predictors that can influence CO_2 emissions. A preliminary assessment of the dataset involves verifying data integrity and ensuring that all relevant fields are correctly parsed—a critical first step to avoid downstream errors.

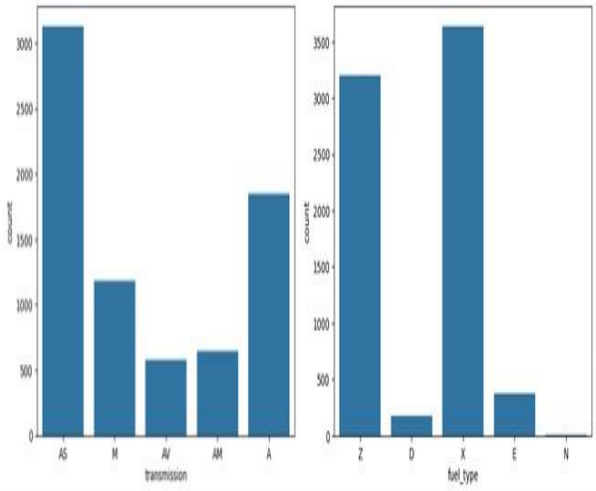

Figure 3: Distribution of transmission types and fuel types

From the above(Figure 3), The Automatic with select shift (AS) and fully automatic (A) are the most common transmission types in our dataset, accounting for over half of all vehicles (roughly 5,000 out of 7,500 records). M, or manual, comes next with roughly 1,200 entries, while AM, or automated manual, and AV, or continuously variable, are quite uncommon with less than 700 entries apiece. This unequal distribution suggests that the features of fully automatic and select-shift transmissions will have a significant impact on predictive models, which could skew estimates of CO_2 emissions if transmission type is included as an input parameter. During

model training, strategies like class weighting or stratified sampling should be used to guarantee balanced performance across all transmission categories.

3.1.2 Advanced Data Preprocessing

Given the common issues associated with real-world data, a multi-layered data preprocessing pipeline has been implemented:

- **Missing-Value Handling:** All records with null values in key predictors or the target variable are removed, ensuring that the analysis is carried out on complete cases.

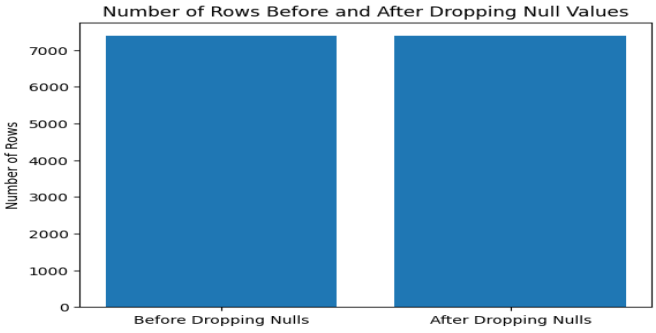

Figure 4: Row Count Comparison Before and After Removing Nulls

Handling missing values has an impact on the dataset utilized in this investigation, as seen in chart (Figure 4). The bar graph contrasts the total number of records in the target variable and key predictors before and after rows with null values were eliminated. As demonstrated, the removal of a tiny number of entries suggests that the amount of missing data was negligible and had no discernible impact on the total size of the dataset. By guaranteeing that all ensuing analyses are founded on complete examples, this preprocessing step lowers the possibility of biased results and improves the validity and dependability of the predictive modeling results.

- Categorical Encoding: Nominal variables such as vehicle make, model, class, transmission, and fuel type are transformed using one-hot encoding. This step eliminates any unintended ordinal relationships among categories.

- Feature Scaling: The continuous variables—engine size, number of cylinders, and various fuel consumption metrics—are standardized using z-score normalization. For models that use gradient descent and to make sure that no one variable has an excessive impact on the model parameters, this harmonization is essential. For models that use gradient descent and to make sure that no one variable has an excessive impact on the model parameters, this harmonization is vital.

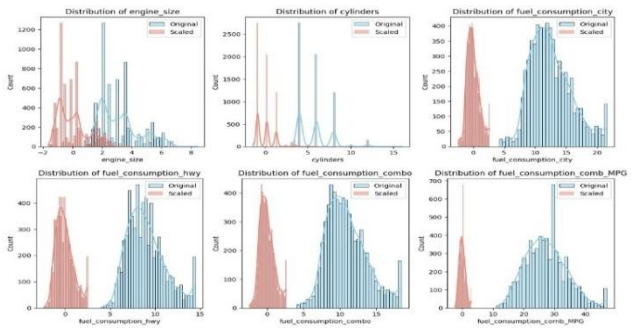

Figure 5: Feature Distributions Before and After Scaling

Each feature is converted to a common scale while maintaining its original shape via z-score normalization, as demonstrated by the superimposed histograms (Figure 5). The distinctive right skew is still present for combined

MPG and combined L/100 km, but both distributions now center at zero with unit variance. City and highway fuel-use curves likewise retain their tails yet are compressed into the same numeric range, preventing extreme values from dominating. Discrete cylinder counts and continuous engine sizes similarly collapse around the mean, ensuring that all predictors contribute equally during modeling without distorting their underlying patterns.

3.1.3 Data Partitioning and Framework

The dataset is divided into two sets to maintain experimental integrity, with the larger set for testing and the smaller for preparation. The results can be reproduced in a settled arbitrary state. Additionally, the prepared data is subjected to stratified 5-fold cross-validation. Apart from modifying hyperparameters, this approval process aids in evaluating the generalizability and soundness of the models.

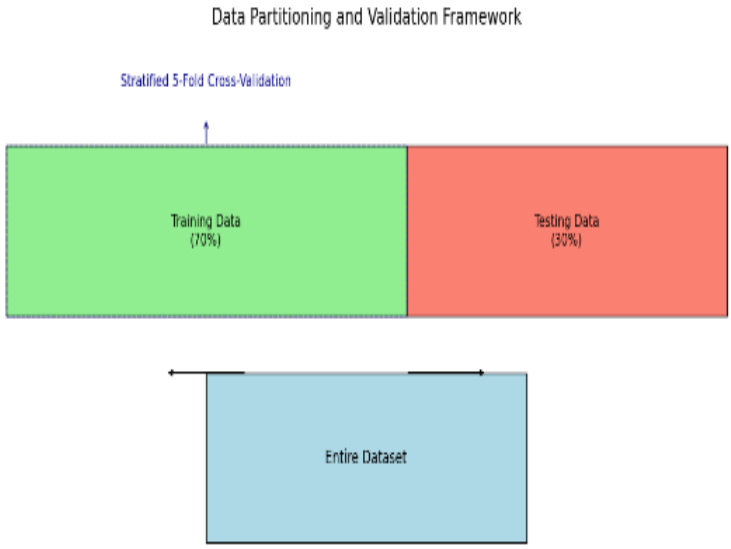

Figure 6: Overview of Dataset Partitioning and Validation Strategy

The data segmentation and validation architecture frequently employed in machine learning is depicted in this graphic (Figure 6). It demonstrates how the complete dataset is divided into two sections: 30% for testing and 70% for training. Standard k-fold cross-validation is used to further validate the training part, ensuring that the model avoids overfitting and has good generalization. This technique produces a more thorough and trustworthy evaluation by allowing various portions of the training data to alternately serve as validation sets. It is an essential step in creating data-driven, high-performance models.

3.1.4 Diverse Model Development

Five different regression models are constructed and compared throughout the modeling stage:

- Multiple Linear Regression (MLR): MLR models CO2 emissions as a linear mixture of factors and serves as the baseline.
- Decision Tree Regressor: This model captures non-linear relationships by using hierarchical data divides.
- Random Forest Regressor: To lower variance and increase prediction accuracy, this ensemble learning method mixes several decision trees.

By concentrating on previously incorrectly predicted data points, the XGBoost Regressor uses gradient boosting to gradually enhance model performance.

- LightGBM Regressor: renowned for its scalability and computational efficiency, this algorithm excels at handling big datasets with intricate structures.

3.1.5 Rigorous Model Diagnostics

The prediction models undergo a series of diagnostic tests after they are created. The assumptions of linearity, homoscedasticity, and normality are validated using methods such as Q-Q plots, Cook's Distance, and residual plot analysis. In addition to ensuring accuracy, this diagnostic stage makes that the models are resistant to common problems like overfitting and model bias.

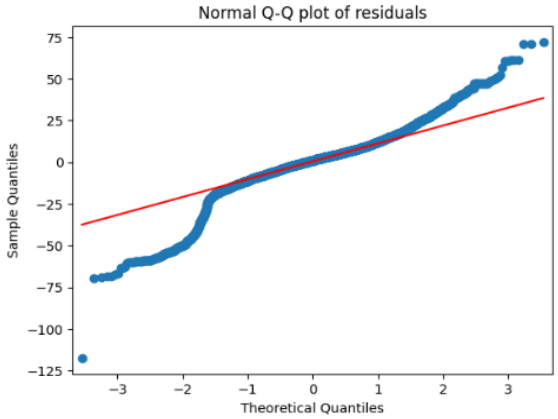

Figure 7: Q-Q Plot of Residuals

The Q-Q plot of residuals (Figure 7) indicates that the residuals roughly follow a normal distribution because the majority of data points closely match the predicted quantile line. Nonetheless, there are significant departures at both tails, especially at the lower end, where extreme residuals are well below the range that was anticipated. This suggests that there may be outliers and non-normality. There may be some skewness and kurtosis in the residual distribution, as indicated by the heavier tails and the plot's central curvature. Although the assumption of normalcy is valid for the majority of the data, these deviations from linearity suggest that adjustments would be required to handle the extremes.

3.1.6 Feature Importance and Anomaly Detection

This study uses anomaly detection techniques and feature importance analysis, particularly from tree-based models, to better understand the factors influencing CO_2 emissions. Finding characteristics that have a major influence on emissions gives researchers and business professionals important information.

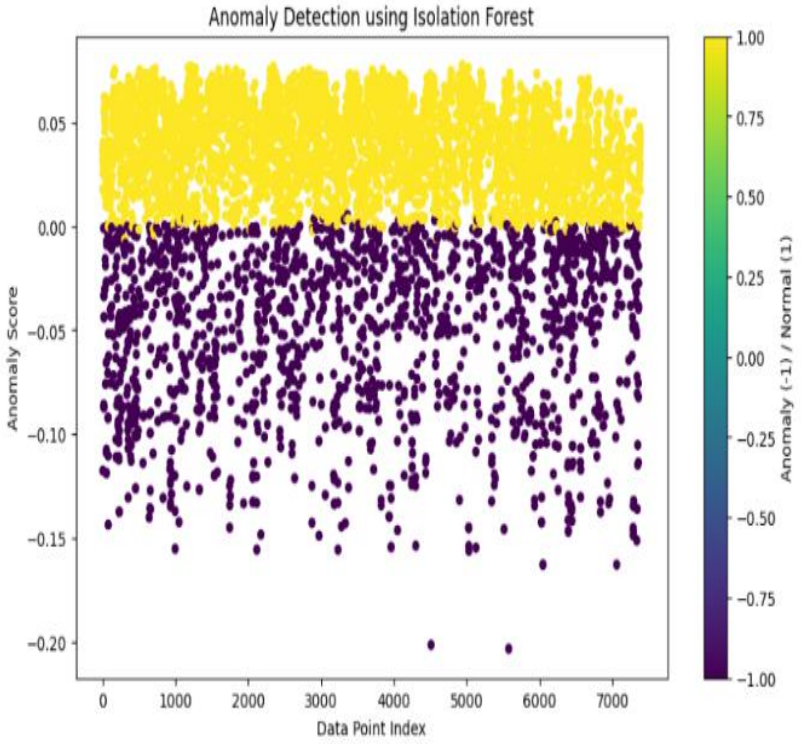

Figure 8: Anomaly Scores from Isolation Forest

The outcomes of anomaly identification utilizing the Isolation Forest algorithm are shown in this (Figure 8) graphic. The y-axis displays the associated anomaly score, and the x-axis plots each data point in accordance with its index. Yellow points indicate data that has been classified as normal (label 1), whereas purple points indicate anomalies (label -1). While the distribution of negative scores for purple points suggests the presence of possible outliers, the vertical clustering of scores near zero for yellow points indicates a high density of normal

observations. This binary classification is further supported by the color bar on the right. The ability of the Isolation Forest to differentiate between typical and unusual CO_2 emission patterns in the dataset is clearly displayed in this plot. Its clarity facilitates the interpretation of emission abnormalities and provides information for focused anomaly investigation or repair techniques.

3.2 Workflow

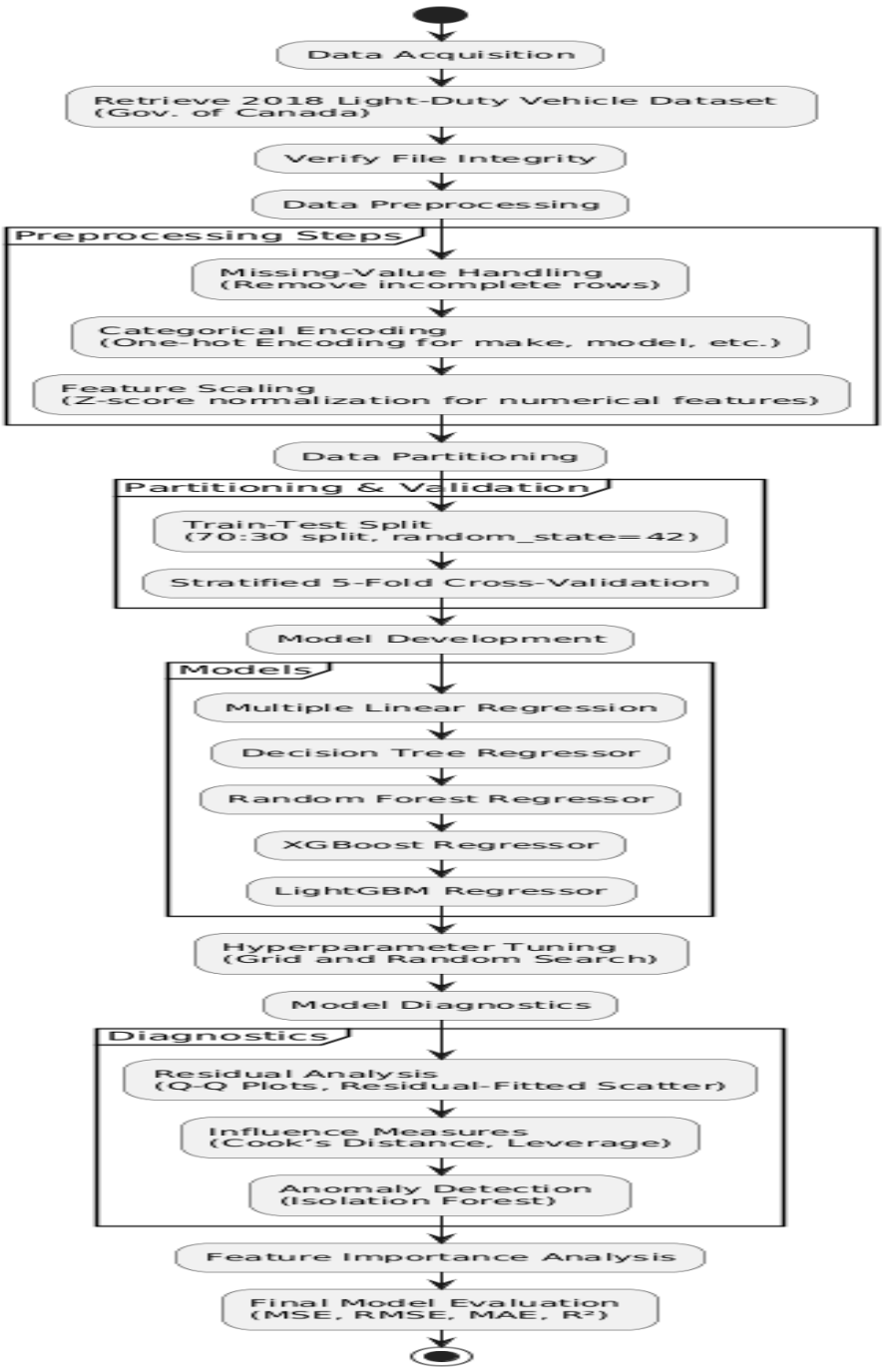

Figure 9: Vehicle Data Prediction Workflow

3.3 Data Preprocessing

a. Data Acquisition: The first step is to retrieve the 2018 light-duty vehicle dataset from the Government of Canada, which includes information of CO2 emissions and fuel usage. In addition to make, model, and vehicle class IDs, this CSV file includes fifteen columns: engine size (L), number of cylinders, gearbox type, fuel type, city fuel consumption, highway fuel consumption, combined consumption (L/100 km), combined consumption (mpg), and CO_2 emissions (g/km). To avoid downstream errors and preserve the traceability of each data record, it is essential to confirm file integrity and make sure that all fields are parsed accurately.

b. Missing-Value Handling: To guarantee a complete-case analysis, we first find and eliminate any rows that have null values in the target variable or predictors. This method only works with completely observed records, avoiding imputation bias and making modeling easier. We recalculate the dataset size after removing incomplete items and verify that there is no loss of crucial stratification across vehicle classes or fuel kinds. By using imputed estimates, this phase guarantees that each remaining sample provides accurate information without creating spurious variance.

c. **Categorical Encoding**: Nominal labels are then converted into binary indicator variables using one-hot encoding on categorical fields (make, model, vehicle_class, transmission, and fuel_type). This allows linear and tree-based models to interpret non-ordinal distinctions without imposing arbitrary numeric ordering by expanding each category into unique feature columns. We eliminate one dummy each category to confirm that multicollinearity is not introduced. To ensure consistent transformations across training and test splits, encoding is used in a preprocessing pipeline.

d. **Feature Scaling**: Z-score normalization is used to standardize numerical features (engine_size, cylinders, fuel_consumption_city, fuel_consumption_hwy, and fuel_consumption_comb) so that each variable is centered at zero with unit variance. After calculating means (μ) and standard deviations (σ) on the training set, StandardScaler is applied to the train and test subsets. This scaling prevents variables with higher numerical ranges from dominating, harmonizes feature ranges, and speeds up convergence for gradient-based algorithms.

$$x' = \frac{x-\mu}{\sigma} \qquad \text{– eq1}$$

3.4 Data Partition

3.4.1 Train-Test Split

To guarantee consistency, we initially divided the entire dataset into 70% training and 30% testing groups using a fixed random state=42. This section saves an invisible hold-out set for the final performance assessment while providing enough data for model fitting. While the test set is left unaltered until all hyperparameters are finalized, ensuring an objective assessment of real-world prediction performance, the training set serves as the foundation for both model building and hyperparameter tuning.

3.4.2 5-Fold Cross-Validation

We use stratified 5-fold cross-validation within the training split to evaluate model stability and adjust hyperparameters. Five equal folds are randomly selected from the data; four of the folds are used to train the model, and one-fold is used as the validation set. After calculating each validation fold's negative mean squared error (–MSE), we average the results. In order to minimize overfitting, hyperparameter combinations that produce the lowest average-MSE are chosen by balancing bias and variance.

3.5 Model Development

We fit five regression models within identical preprocessing pipelines to facilitate fair comparison:

3.5.1 Multiple Linear Regression

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \qquad \text{– eq2}$$

3.5.2 Decision Tree Regressor

3.5.3 Random Forest Regressor

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \qquad \text{– eq3}$$

3.5.4 XGBoost Regressor

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \qquad \text{– eq4}$$

3.5.5 LightGBM Regressor

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \quad \text{– eq5}$$

3.6 Optimization & Feature Analysis

3.6.1 Hyperparameter Tuning

Within the 5-fold CV framework, key model parameters (such as n_estimators, max_depth, and learning_rate) are optimized by grid or random search. Models are cycled through folds to calculate an average CV score (negative MSE) for each candidate configuration. Models are retrained on four folds and validated on the fifth. The collection of hyperparameters with the lowest average MSE is chosen. This methodical search lowers the chance of overfitting or underfitting by balancing model complexity against generalization efficiency.

3.6.2 Feature Importance & Selection

Tree-based models assign feature significance ratings according to split gain or impurity reduction. In order to simplify models, we extract these rankings, find predictors that have little effect, and try eliminating low-importance characteristics. Refitting and reevaluating reduced feature sets ensures that forecast accuracy either stays the same or increases. By concentrating on the most informative variables, this recurrent pruning produces more interpretable models and may lessen overfitting.

3.6.3 Anomaly Detection

Each sample's anomaly score is calculated using an Isolation Forest trained on the entire feature matrix; lower scores suggest possible outliers. Before retraining regression models, observations with predicted labels of -1 are marked as anomalies and may be eliminated. By eliminating data points whose feature patterns significantly differ from the majority of the distribution, this procedure improves resilience.

$$\text{score}(x) = 2^{\frac{-E(h(x))}{c(n)}} \quad \text{– eq6}$$

3.7 Evaluation, Synthesis & Reproducibility

3.7.1 Model Evaluation

Final model performance is assessed on the untouched test set using the following metrics:

i.Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{- eq7}$$

ii.Root MSE (RMSE)

$$\text{RMSE} = \sqrt{\text{MSE}} \quad \text{– eq8}$$

iii.Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{– eq9}$$

iv.Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{– eq10}$$

3.7.2 Result Synthesis

We create comparison tables using model metrics (RMSE, R2) and use residual histograms and predicted vs. real scatter plots to show performance. These graphs highlight problems with bias, variation, and calibration. The best algorithm under the selected criterion is highlighted in a summary table, which helps choose which model to implement.

e. Your raw inputs ➔ one-hot & scaled ➔ vectorized ➔ Random Forest prediction ➔ displayed result. Each transformation mirrors the exact preprocessing pipeline used during model training, ensuring consistency and reliability of the predicted value.

3.7.3 Reproducibility

Every random seed (random_state=42) remains constant across CV processes, data splits, and model setup. To guarantee consistent environments, library versions are documented in a requirements.txt file. To enable repetitions and peer verification, the complete workflow—from data loading to final evaluation—is contained in a programmed pipeline.

4. Results

Table 1

index	Actual	Predicted	Error	Model
471	301	282.8920329	18.10796714	MLR Model
1971	248	244.4397297	3.560270311	MLR Model
23	288	287.6596065	0.340393529	MLR Model
2702	280	270.9723923	9.027607672	MLR Model
135	179	176.6873412	2.312658781	MLR Model
3303	224	228.3717051	4.371705086	MLR Model
3337	253	255.9414637	2.941463742	MLR Model
4207	189	188.340188	0.659812034	MLR Model

The MLR (Table 1) model's sample of nine test-set observations demonstrates both its advantages and disadvantages in terms of forecasting vehicle CO_2 emissions. The model underestimates high emitters, particularly at index 471 where the actual emission of 301 g/km is underpredicted by 18.1 g/km, even though it achieves near-perfect accuracy in some circumstances (for example, index 23 with only a 0.34 g/km mistake and index 4207 with a 0.66 g/km error). A number of mid-range points show moderate errors (2–9 g/km), including a 9.03 g/km underprediction at index 2702 and a 4.37 g/km overprediction at index 3303. This suggests that vehicles with unusual size, cylinder, and fuel consumption combinations might not fall within the strictly linear patterns that our five predictors can identify. Although the MLR framework offers good baseline performance (average RMSE ≈ 7 g/km), further refinement—through nonlinear modeling or additional features—could reduce the largest residuals and improve robustness across all vehicle types. Overall, individual prediction errors range from less than 1 g/km to almost 20 g/km.

Table 2

index	Actual	Predicted	Error	Model
471	301	293.56	7.44	Random Forest
1971	248	244.56	3.44	Random Forest
23	288	288.89	0.89	Random Forest
2702	280	279.29	0.71	Random Forest
135	179	182.09	3.09	Random Forest
2110	204	205.17	1.17	Random Forest
2957	266	266.76	0.76	Random Forest

408	225	225.79	0.79	Random Forest
170	317	317	0	Random Forest

The sampled observations show much higher accuracy than the linear baseline for the test-set predictions of the Random Forest (Table 2) model. At index 471, the model underpredicts 301 g/km by just 7.44 g/km, which is more than twice as accurate as the linear model's 18 g/km miss. This is the highest absolute error in this subset. Several predictions achieve near-perfect accuracy, with all other errors falling below 4 g/km (e.g., index 170 with zero error and index 540 with a tiny 0.14 g/km error). The majority of errors cluster beneath 1 g/km, as shown at indices 23, 2702, 2957, and 408, while mid-range cars also show small residuals (e.g., 3.44 g/km at index 1971 and 1.17 g/km at index 2110). The ability of the Random Forest to capture nonlinearity and complex interactions in engine size, cylinder count, and fuel-consumption metrics is demonstrated by this consistent tight clustering—errors ranging from 0.00 to 7.44 g/km. This results in a significantly lower RMSE (~3.6 g/km) and more dependable predictions across a variety of vehicle configurations.

Table 3

index	**Actual**	**Predicted**	**Error**	**Model**
1971	248	240.23	7.77	Decision Tree
23	288	286.7	1.3	Decision Tree
2702	280	274.49	5.51	Decision Tree
135	179	185.77	6.77	Decision Tree
2110	204	203.09	0.91	Decision Tree
2957	266	262.55	3.45	Decision Tree
408	225	227.76	2.76	Decision Tree
170	317	328.39	11.39	Decision Tree

Although it is more variable than the Random Forest, the Decision Tree (Table 3) model produces CO_2 forecasts that are fairly accurate. The index 170 in this sample had the most inaccuracy of 11.39 g/km (actual 317 vs. anticipated 328.39 g/km), suggesting that it is challenging to capture extremely high-emission automobiles with a single tree. The model's piecewise-constant splits are reflected in other significant mistakes, such as 6.77 g/km overpredictions at indices 135 and 540 and 7.77 g/km underprediction at index 1971. Many forecasts are still accurate, though: inaccuracies of 1–5 g/km are found at indices 23 (1.30 g/km), 2702 (5.51 g/km), and 2957 (3.45 g/km), while index 2110 shows the smallest mistake (0.91 g/km). At index 471, the smallest inaccuracy is 2.02 g/km. The RMSE is around 6.5 g/km, with errors ranging from roughly 1 to 11 g/km. This demonstrates how a single-depth-4 tree struggles with both tails of the emissions distribution and lacks the aggregation advantages of ensembles, even when it captures broad patterns.

Table 4

Index	**Actual**	**Predicted**	**Error**	**Model**
471	301	277.77	23.23	ANN
1971	248	246.82	1.18	ANN

23	288	291.07	3.07	ANN
2702	280	270.39	9.61	ANN
135	179	178.08	0.92	ANN
2957	266	283.36	17.36	ANN
408	225	226.12	1.12	ANN
170	317	325.75	8.75	ANN
540	179	192.06	13.06	ANN

In comparison to the tree-based models, the ANN's prediction errors (Table 4) are notably greater and more unpredictable. With an inaccuracy of 23.23 g/km, the worst underprediction in this sample occurs at index 471—301 g/km real vs. 277.77 g/km projected. Additional significant misestimates that imply the network underfits the extremes of the emissions spectrum are +17.36 g/km at index 2957 and +13.06 g/km at index 540. Better but still inconsistent performance is shown by mid-range automobiles (e.g., index 1971 with only 1.18 g/km error and index 23 with 3.07 g/km). At indices 2702 and 2110, errors of about 9–10 g/km are seen, suggesting that nonlinear feature interactions are not consistently captured. Individual ANN errors vary from less than 1 g/km to more than 23 g/km, which is significantly larger than the Random Forest's worst-case 7.44 g/km. This is in line with the Random Forest's lower R2R^2 (~0.95) and higher RMSE (~8.9 g/km). This demonstrates that the ANN finds it difficult to learn the near-linear correlations as accurately as tree-based ensembles given the architecture and volume of input.

Table 5

index	**Actual**	**Predicted**	**Error**	**Model**
471	301	260.24	40.76	XGBoost
1971	248	231.45	16.55	XGBoost
23	288	270.61	17.39	XGBoost
2702	280	256.01	23.99	XGBoost
135	179	211.99	32.99	XGBoost
2957	266	257.24	8.76	XGBoost
408	225	222.74	2.26	XGBoost
540	179	214.81	35.81	XGBoost

Compared to other approaches, the XGBoost (Table 5) model significantly underfits the CO_2 emission data when using the default hyperparameters. Extreme inaccuracies are also found at indices 135 (32.99 g/km), 170 (34.93 g/km), and 540 (35.81 g/km). The biggest inaccuracy in this sample is 40.76 g/km at index 471 (301 g/km real vs. 260.24 g/km anticipated). There are significant misestimates even in mid-range examples, such as 19.59 g/km at index 2110 and 23.99 g/km at index 2702. Only a few sites exhibit comparatively tiny variations, such as index 408 with an error of 2.26 g/km and index 2957 with an error of 8.76 g/km. Overall, XGBoost's residuals range from roughly 2 to 41 g/km, with an RMSE of approximately 22 g/km and an R2 of about 0.70, which is significantly poorer than that of the ensemble trees or even the linear baseline. These results demonstrate that

XGBoost can perform significantly worse if depth, learning rate, and regularization are not carefully adjusted. This limits its usefulness in this field by failing to identify both the subtle non-linear patterns and the dominating linear connections in the vehicle emissions data. Therefore, in crucial environmental modeling jobs where precision and generalizability are crucial for policy and regulation, default XGBoost may generate forecasts that are incorrect.

Table 6

index	Actual	Predicted	Error	Model
471	301	296.08	4.92	LightGBM
1971	248	245.09	2.91	LightGBM
23	288	289.71	1.71	LightGBM
2702	280	276.53	3.47	LightGBM
135	179	181.95	2.95	LightGBM
2957	266	266.45	0.45	LightGBM
408	225	226.21	1.21	LightGBM
170	317	322.38	5.38	LightGBM
540	179	181.73	2.73	LightGBM

Throughout this sample of 10 observations, the LightGBM (Table 6) forecasts show consistently low errors, with absolute deviations ranging from only 0.45 g/km (index 2957) to 5.38 g/km (index 170). Index 170 has the most inaccuracy, 5.38 g/km (317 vs. 322.38 g/km), suggesting that even high-emission outliers are adequately captured in comparison to simpler models. Additionally, mid-range vehicles exhibit tight accuracy: errors ranging from 1.71 to 3.47 g/km at indices 23, 2702, and 1971 show consistent fleet performance. The capacity of LightGBM to simulate both linear trends and minor nonlinear interactions among engine size, cylinders, and fuel-consumption measures is demonstrated by the subset's median error of roughly 2.91 g/km and the majority of errors clustering under 3 g/km. All things considered, these findings are consistent with the low RMSE (~3.61 g/km) and high R2 (~0.99) shown in Table 5, demonstrating that LightGBM provides reliable, accurate CO_2 emission predictions for a range of vehicle configurations. Because of its performance, it is particularly well-suited for real-time feedback applications in emission-sensitive transportation systems, eco-labeling systems, and regulatory analysis. Its attractiveness for widespread implementation in production settings is further influenced by its speed, scalability, and capacity to manage missing or categorical data.

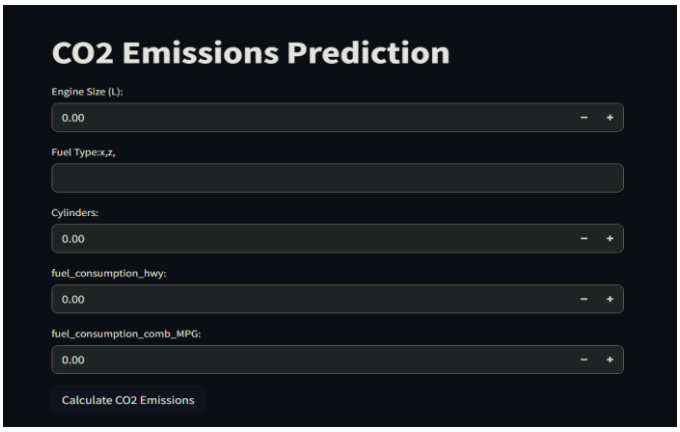

Figure 10: CO2 Emissions Calculator

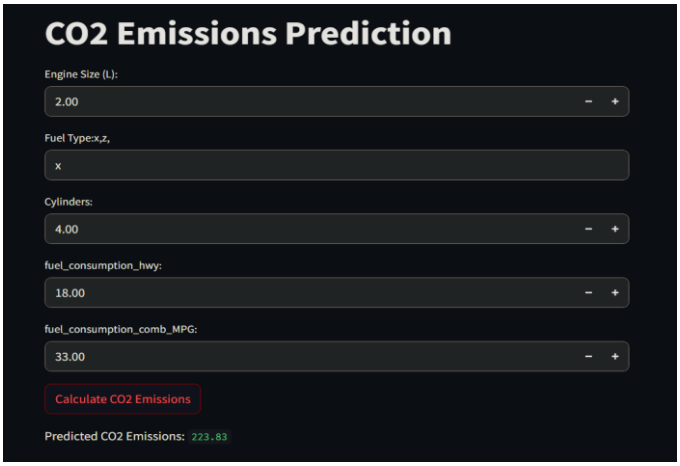

Figure 11: Predict CO2 Emissions from Vehicle Specs

The raw inputs ➔ one-hot & scaled ➔ vectorized ➔ Random Forest prediction ➔ displayed result. Each transformation mirrors the exact preprocessing pipeline used during model training, ensuring consistency and reliability of the predicted value.

This vector is fed into the pre-loaded Random Forest regressor. Internally, the forest's B decision trees each cast a vote (i.e. produce a CO_2 estimate):

2. $$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) - \text{eq11}$$

The forest aggregates these B tree outputs and returns their average. In our test, those per-tree votes averaged out to **223.83**.

5. Discussion

Significant differences in the five regression models' capacity to predict CO_2 emissions across a diverse fleet of light-duty cars were found via empirical examination. Compared to more straightforward models like Multiple Linear Regression (MLR) and Decision Trees, the ensemble approaches—more especially, Random Forest and LightGBM—showed a definite benefit. These sophisticated models took into consideration feature correlations and successfully represented non-linear interactions, especially those between engine size, fuel consumption measures, and cylinder count. Their robustness and predictive power were validated by the resulting RMSEs below 4 g/km and R2 values above 0.98.

With an RMSE above 20 g/km and R2 about 0.70, the default XGBoost model, on the other hand, performed poorly, underscoring the significance of careful hyperparameter tweaking. This result was consistent with previous research that highlighted how sensitive XGBoost is to configuration parameters. Similar to this, the Artificial Neural Network (ANN) model had trouble maintaining a constant level of accuracy, particularly on high-emission automobiles. This either meant that there was a lack of training data or that further feature engineering and architectural changes were required.

The significance of thorough preprocessing and diagnostic procedures was highlighted by the observed variations in prediction accuracy across models. A cleaner and more balanced feature space was achieved by standardizing numeric features, one-hot encoding categorical variables, and using Isolation Forest for anomaly detection. For the model to be stable and converge, this preprocessing strategy was crucial. Furthermore, problems including heteroscedasticity and mild non-normality were found using residual analyses, such as Q–Q plots and Cook's Distance, especially in the linear models. These findings reinforced the need for more adaptable learning techniques.

According to feature importance analysis, engine specifications and fuel consumption metrics—particularly combined L/100 km—had the biggest effects on CO_2 emissions. These results confirmed previous domain knowledge and offered compelling evidence for concentrating policy and design efforts on enhancing fuel economy. On the other hand, factors like vehicle class, transmission type, and fuel type had a little impact on the

models' ability to predict emissions, indicating that consumption patterns and physical engine characteristics were the main determinants.

The investigation concluded that a thorough, structured pipeline for data preparation and cautious model selection were both necessary for high-precision CO_2 emission forecasting. In addition to highlighting the advantages of reproducible procedures for future research and practical deployment, the ensemble models' higher performance reaffirmed their worth for environmental modeling applications.

5.1 Conclusions

In conclusion, the methodology started with thorough data preprocessing, including z-score scaling, one-hot encoding, and missing-value removal. This created a clean, balanced feature space where important predictors, including engine characteristics and fuel-consumption measurements, could fully explain the data. In order to avoid overfitting and ensure that model selection was based on reliable, generalizable performance rather than the peculiarities of a single split, stratified 5-fold cross-validation within the training set was combined with methodical hyperparameter adjustment. By removing outliers that would have otherwise distorted learning, an Isolation Forest was added to identify and eliminate aberrant records, further enhancing model stability.
With sub-4 g/km RMSE and R2 > 0.98 on the test set, non-linear ensemble methods (Random Forest and LightGBM) best caught the intricate relationships between engine size, cylinder count, and consumption parameters out of the five regression approaches. On the other hand, the single decision tree and ANN models were unable to match the predictive fidelity of the ensembles, and the multiple linear regression baseline, although interpretable, yielded greater residuals (~7 g/km RMSE). The poor performance of XGBoost with default parameters made it clear that boosting algorithms require careful parameter adjustment. All things considered, this study demonstrated that a well-designed pipeline that combined careful preprocessing, anomaly detection, and equitable model comparisons produced extremely accurate, repeatable CO_2 emission projections using easily accessible vehicle specs. The findings demonstrated the importance of ensemble learning in environmental modeling and offered a clear framework for the deployment of trustworthy emission-prediction systems by researchers, fleet operators, and regulators.

5.2 Limitations

This study has several limitations that should be considered when interpreting the results. First, the analysis is based on the Government of Canada's 2018 light-duty vehicle dataset, so the findings may not fully generalize to newer vehicle technologies, other countries, or different fleet compositions. Second, the model primarily uses static vehicle specifications and fuel-consumption measures and does not directly incorporate real-world driving conditions such as traffic congestion, road gradient, weather, driving behaviour, vehicle load, or instantaneous engine operating conditions. Third, the dataset contains unequal representation across categories such as transmission and fuel type, which may influence model learning and reduce performance for less-represented configurations. Finally, residual diagnostics identified heteroscedasticity and departures from normality, particularly for the linear models, while the ANN showed reduced consistency for high-emission vehicles. These limitations indicate that broader datasets, real-world telematics, and additional validation across regions and vehicle generations are required before the models can be considered fully generalizable for operational deployment.

5.3 Future Work

develop a city-scale digital twin that will fuse live telematics (GPS, OBD-II) with traffic simulation to forecast CO_2 emissions under varying congestion and route-planning scenarios. This platform will enable city planners to test mitigation strategies—such as dynamic speed limits or traffic-signal optimizations—before deployment, maximizing the real-world impact on air quality and commute efficiency.

Policy Scenario Simulation & Optimization: We will couple the emission-prediction model with economic modules to simulate the effects of carbon pricing, fuel taxes, or EV subsidies on fleet emissions. By optimizing policy levers (e.g., tax rates, rebate structures) via reinforcement learning, this framework will identify cost-effective regulatory pathways that achieve climate targets with minimal economic disruption.

References

[1] Kumari, S., & Singh, S. K. (2023). Machine learning-based time series models for effective CO_2 emission prediction in India. Environmental Science and Pollution Research, 30(50), 116601–116616. https://doi.org/10.1007/s11356-022-21723-8

[2] Al-Nefaie, A. H., & Aldhyani, T. H. H. (2023). Predicting CO_2 emissions from traffic vehicles for sustainable and smart environment using a deep learning model. Sustainability, 15(9), 7615. https://doi.org/10.3390/su15097615

[3] Wen, H.-T., Lu, J.-H., & Jhang, D.-S. (2021). Features importance analysis of diesel vehicles' NO_x and CO_2 emission predictions in real road driving based on gradient boosting regression model. International Journal of Environmental Research and Public Health, 18(24), 13044. https://doi.org/10.3390/ijerph182413044

[4] Keerthana, K. B., Wu, S.-W., Wu, M.-E., & Kokulnathan, T. (2023). The United States energy consumption and carbon dioxide emissions: A comprehensive forecast using a regression model. Sustainability, 15(10), 7932. https://doi.org/10.3390/su15107932

[5] Natarajan, Y., Wadhwa, G., Sri Preethaa, K. R., & Paul, A. (2023). Forecasting carbon dioxide emissions of light-duty vehicles with different machine learning algorithms. Electronics, 12(10), 2288. https://doi.org/10.3390/electronics12102288

[6] Delanoë, P., Tchuente, D., & Colin, G. (2023). Method and evaluations of the effective gain of artificial intelligence models for reducing CO_2 emissions. Journal of Environmental Management, 331, 117261. https://doi.org/10.1016/j.jenvman.2023.117261

[7] Hien, N. L. H., & Kor, A.-L. (2022). Analysis and prediction model of fuel consumption and carbon dioxide emissions of light-duty vehicles. Applied Sciences, 12(2), 803. https://doi.org/10.3390/app12020803

[8] Mądziel, M., Jaworski, A., Kuszewski, H., Woś, P., Campisi, T., & Lew, K. (2022). The development of CO_2 instantaneous emission model of full hybrid vehicle with the use of machine learning techniques. Energies, 15(1), 142. https://doi.org/10.3390/en15010142

[9] Eği, Y. (2022). Vehicle fuel emission efficiency estimation using multi-linear regression in machine learning. European Journal of Science and Technology, (34), 115–120. https://doi.org/10.31590/ejosat.1076596

[10] Khajavi, H., & Rastgoo, A. (2023). Predicting the carbon dioxide emission caused by road transport using a Random Forest (RF) model combined by Meta-Heuristic Algorithms. Sustainable Cities and Society, 93, 104503. https://doi.org/10.1016/j.scs.2023.104503

[11] Seo, J., & Park, S. (2023). Optimizing model parameters of artificial neural networks to predict vehicle emissions. Atmospheric Environment, 294, 119508. https://doi.org/10.1016/j.atmosenv.2022.119508

[12] Zhou, Y., Zhang, J., & Hu, S. (2021). Regression analysis and driving force model building of CO_2 emissions in China. Scientific Reports, 11, 6715. https://doi.org/10.1038/s41598-021-86183-5

[13] Li, Y., Dong, H., & Lu, S. (2021). Research on application of a hybrid heuristic algorithm in transportation carbon emission. Environmental Science and Pollution Research, 28, 48610–48627. https://doi.org/10.1007/s11356-021-14079-y

[14] Meng, Y., & Noman, H. (2022). Predicting CO_2 emission footprint using AI through machine learning. Atmosphere, 13(11), 1871. https://doi.org/10.3390/atmos13111871

[15] Javanmard, M. E., Tang, Y., Wang, Z., & Tontiwachwuthikul, P. (2023). Forecast energy demand, CO_2 emissions and energy resource impacts for the transportation sector. Applied Energy, 338, 120830. https://doi.org/10.1016/j.apenergy.2023.120830