Project Report

Aishwariya Alagesan
(002818973)

857-379-6927

alagesan.a@northeastern.edu

Signature of Student : Aishwariya Alagesan
Submission Date : 10th April, 2024

PREDICTION OF CO2 VEHICLE EMISSIONS

PROBLEM CONTEXT

This project, spanning from 2015 to 2024, aims to forecast CO2 emissions from new light-duty vehicles in Canada using fuel consumption ratings data. By analyzing a diverse range of vehicle models and years, the goal is to identify evolving trends in fuel efficiency and emission standards. The primary focus is on predicting carbon dioxide output to encourage environmentally conscious vehicle choices and understand the impact of technological advancements on emissions. Through thorough analysis integrating data from multiple years, the project aims to uncover correlations between fuel efficiency metrics and CO2 emissions, offering insights into the environmental implications of different vehicle models. By employing rigorous exploratory data analysis, feature engineering techniques, meticulous model selection, and precise evaluation methods, the project aims to develop a predictive model that accurately estimates CO2 emissions. Ultimately, the objective is to provide stakeholders with valuable insights to guide sustainable vehicle selection practices, highlighting the significance of environmental awareness in Canada's automotive landscape.

PROBLEM DEFINITION

The central goal of this project is to create a predictive model capable of accurately estimating CO2 emissions from vehicles using fuel consumption ratings data spanning from 2015 to 2024. Utilizing this extensive dataset, the analysis will explore the relationships between fuel efficiency metrics and carbon dioxide output to offer insights into the environmental impact of various vehicle models. The project involves exploratory data analysis to uncover patterns, feature engineering to improve predictive capabilities, model selection to identify the most effective algorithm, and evaluation techniques to ensure the model's accuracy. Ultimately, the project aims to equip stakeholders with the necessary tools to make informed decisions regarding vehicle selection based on environmental considerations.

DATA SOURCE

The dataset is sourced from the Canadian government website and can be accessed via the following link: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6

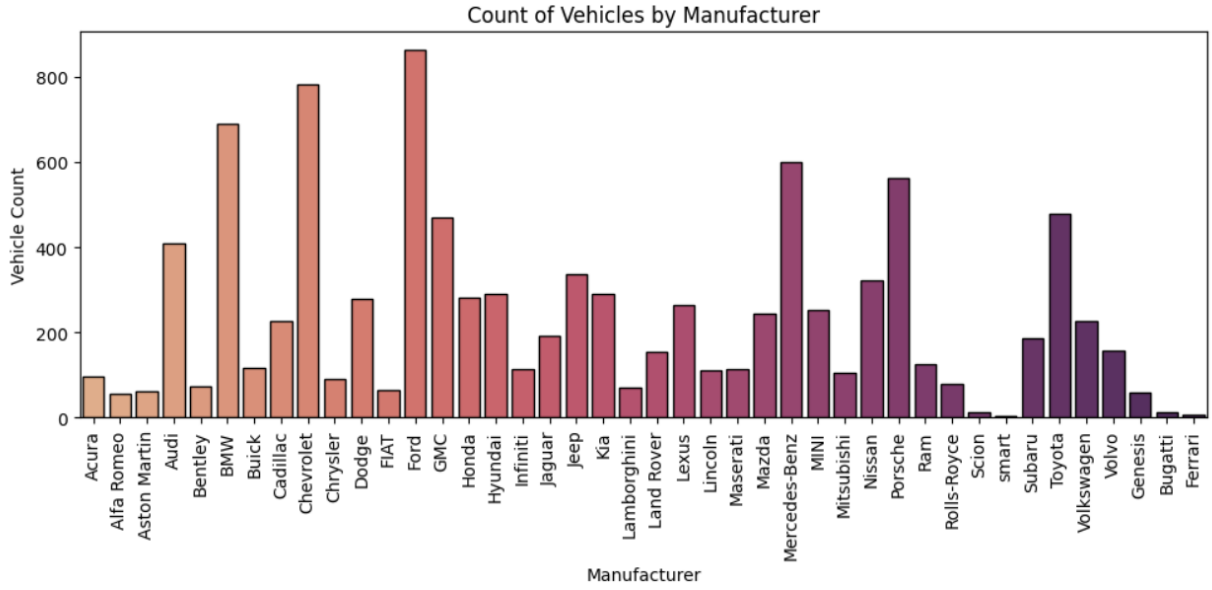
DATA DESCRIPTION

This dataset comprises information on 9929 cars, exploring fuel efficiency and emissions. It includes details such as make, model, engine size, and fuel type. City and highway fuel consumption (measured in liters per 100 kilometers) are provided, along with combined miles per gallon (mpg). CO2 emissions (measured in grams per kilometer) are available for all cars, while CO2 and smog ratings have some missing values. This dataset facilitates analysis of fuel trends, comparisons by car type, and the relationship between emissions and efficiency.

DATA EXPLORATION

Manufacturer-wise Vehicle Production

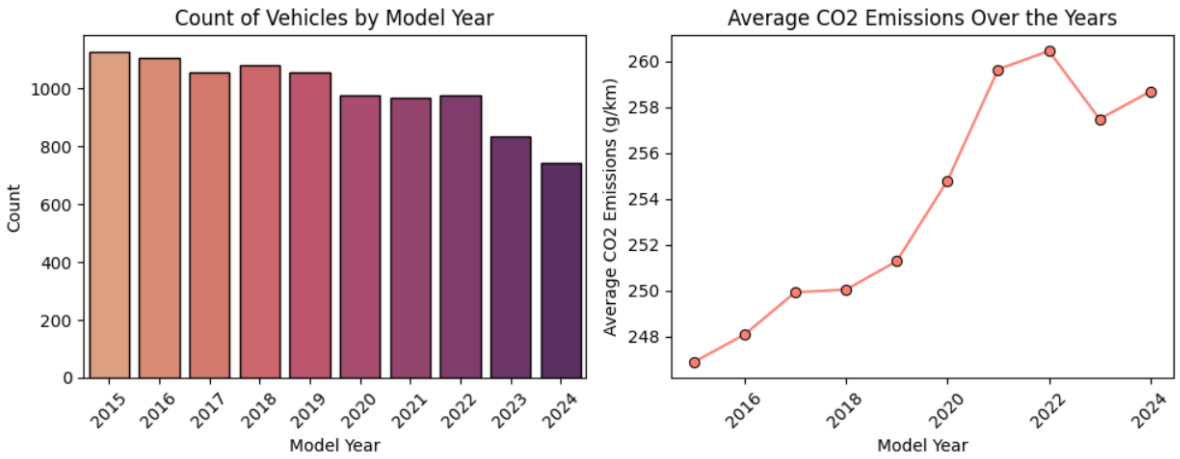
Observing the graph, it becomes apparent that Ford has consistently led in vehicle production throughout the years. Conversely, Ferrari, Smart, and Bugatti demonstrate notably lower production rates compared to other manufacturers.



Interpretation: Ford's dominance in vehicle production reflects its strong presence in the automotive market, while the lower production figures of Ferrari, Smart, and Bugatti may be attributed to their focus on luxury and niche markets rather than mass production.

Comparision of Vehicle count and CO2 emissions over the years:

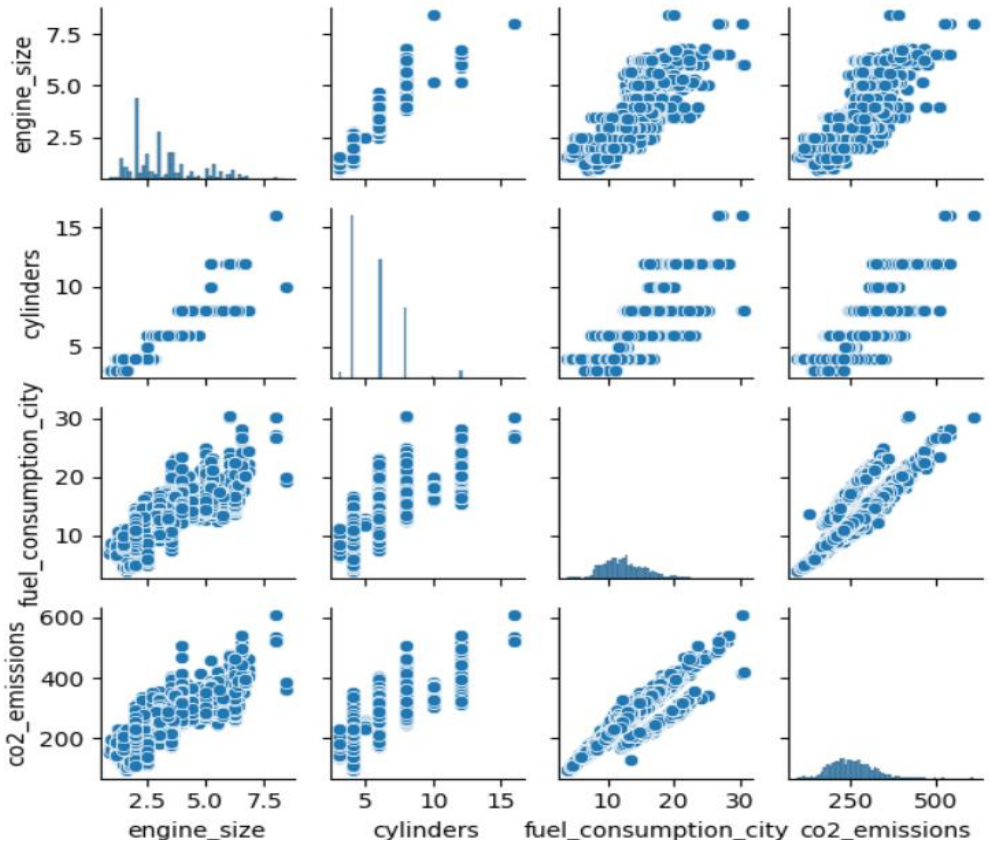
Analysis reveals a contrasting trend between vehicle production and CO2 emissions over the years. While the production of vehicles has shown a decrease, there has been a noticeable increase in CO2 emissions during the same period.



Interpretation: The increase in CO2 emissions despite a decline in vehicle production suggests that individual vehicles are emitting more CO2 on average. This could be due to various factors such as an increase in the popularity of larger, less fuel-efficient vehicles or a lack of significant improvements in vehicle emissions technology. Such insights underscore the importance of addressing emission concerns and promoting more environmentally friendly vehicle options.

Pair Plot Analysis:

Upon examining the pair plot depicting the variables engine size, cylinders, fuel consumption city, and CO2 emissions, a notable pattern emerges. There appears to be a strong linear relationship between engine size and fuel consumption city, which in turn influences CO2 emissions.



Interpretation:

Engine Size vs. Fuel Consumption City:

The pair plot indicates a clear positive correlation between engine size and fuel consumption city. As engine size increases, fuel consumption in urban settings also tends to rise. This relationship is intuitive, as larger engines typically require more fuel to operate.

Engine Size vs. CO2 Emissions:

Given the strong correlation between engine size and fuel consumption city, it follows that there is also a significant correlation between engine size and CO2 emissions. Larger engines, which consume more fuel, produce higher levels of CO2 emissions.

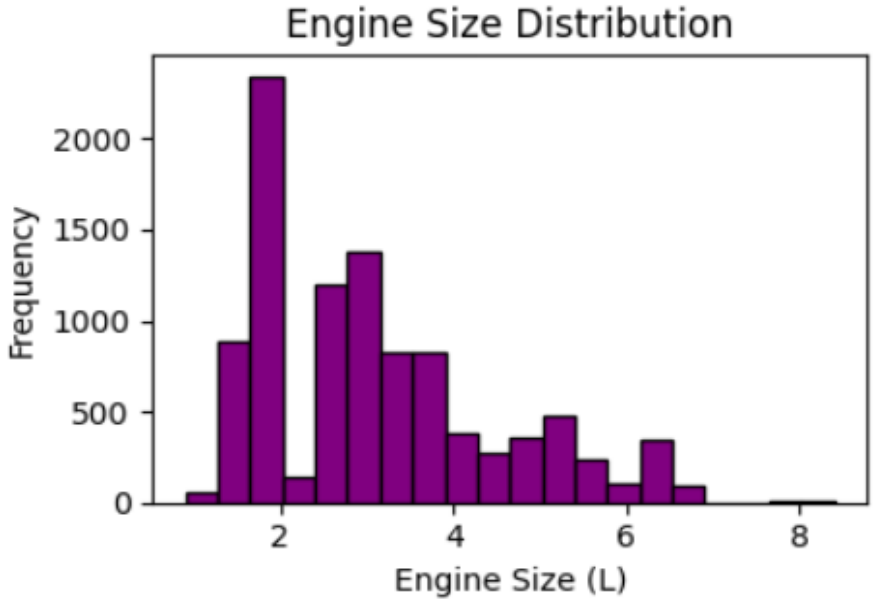
This analysis underscores the importance of engine size in determining both fuel consumption and CO2 emissions. It suggests that reducing engine size or improving engine efficiency could contribute to lower fuel consumption and reduced environmental impact in terms of CO2 emissions.

Analysis and Interpretation of Distributions:

Engine Size Distribution:

The distribution of engine sizes reveals that the majority of vehicles have engine sizes around 1 liter. This suggests that a significant portion of the vehicles in the dataset are equipped with smaller engines, which are typically associated with smaller or more fuel-efficient vehicles.

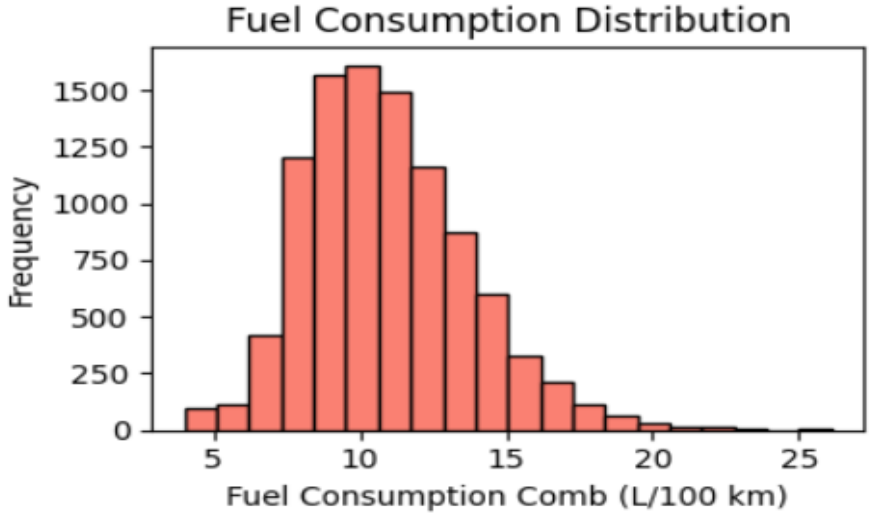
Interpretation: The prevalence of smaller engine sizes indicates a trend towards fuel efficiency and potentially smaller vehicle sizes. This could be driven by factors such as consumer demand for economical vehicles or regulatory requirements for reduced emissions.



Fuel Consumption Distribution:

Examining the distribution of fuel consumption, it is observed that the peak lies within the range of 9 to 15 liters per 100 kilometers. This indicates that a substantial number of vehicles in the dataset consume fuel at this rate.

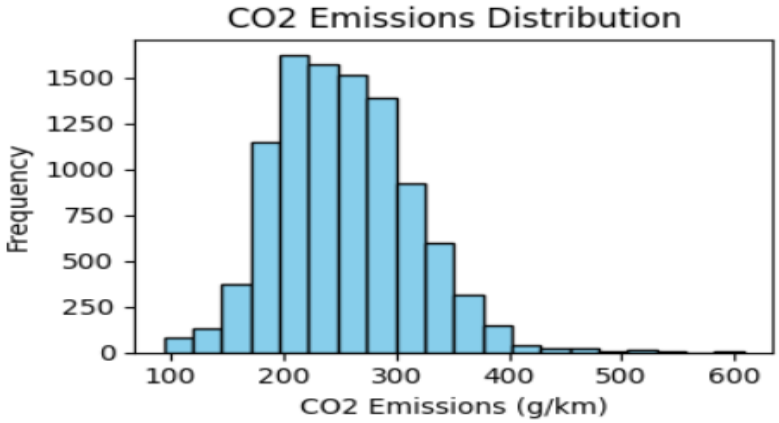
Interpretation: The concentration of vehicles with fuel consumption in this range suggests that they may represent a common segment in the automotive market. It is plausible that these vehicles encompass a variety of sizes and types, but they tend to have fuel consumption rates within this specific range.



CO2 Emissions Distribution:

The distribution of CO2 emissions demonstrates a peak around the range of 200 to 300 grams per kilometer. This indicates that a significant proportion of vehicles in the dataset emit CO2 within this range.

Interpretation: The concentration of CO2 emissions in this range suggests that many vehicles may have emissions levels that fall within a moderate to high range. This could be influenced by various factors such as engine size, fuel efficiency, and emission control technologies.

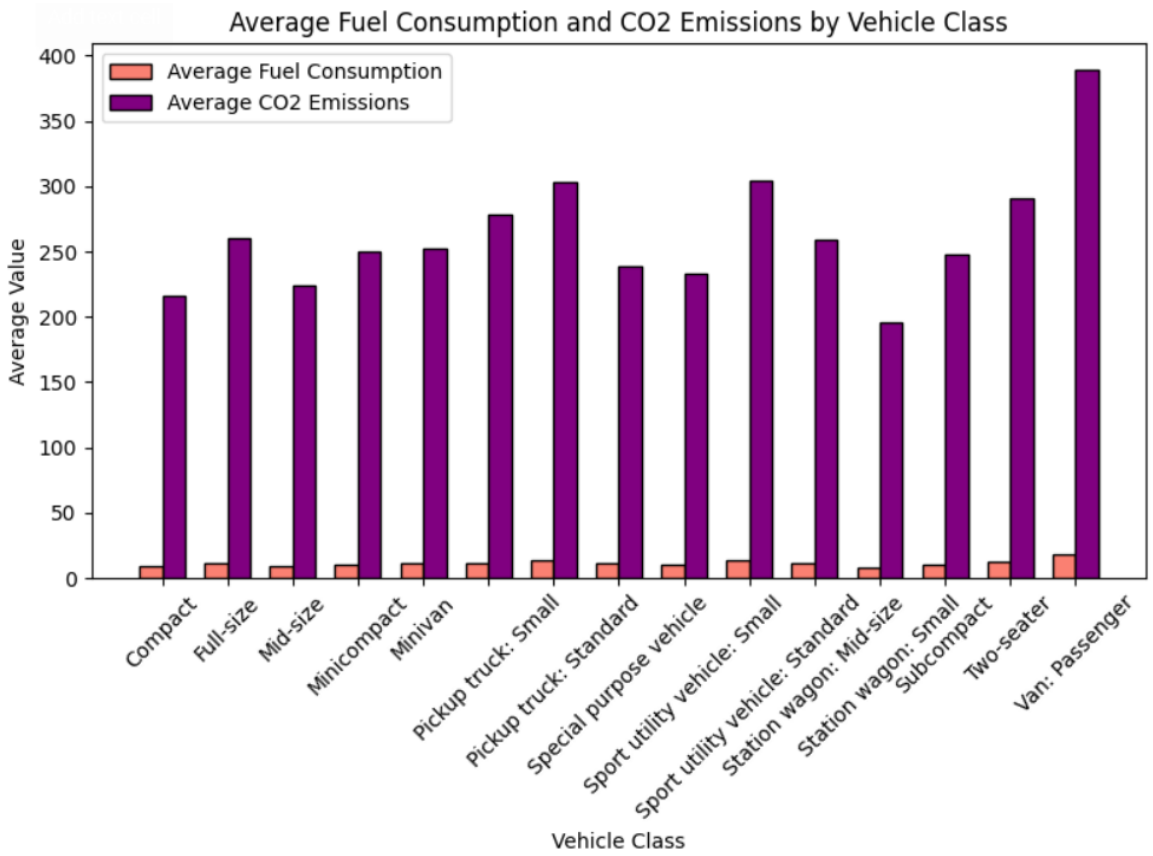


Overall, the analysis of engine size, fuel consumption, and CO2 emissions distributions provides insights into the characteristics of vehicles in the dataset. It highlights trends towards smaller engine sizes and fuel-efficient vehicles, while also indicating the prevalence of vehicles with moderate to high CO2 emissions levels. Understanding these distributions can inform discussions around environmental impact and policy interventions aimed at promoting cleaner and more sustainable transportation options.

Analysis of Average Fuel Consumption and CO2 Emissions:

Average Fuel Consumption and CO2 Emissions by Vehicle Class:

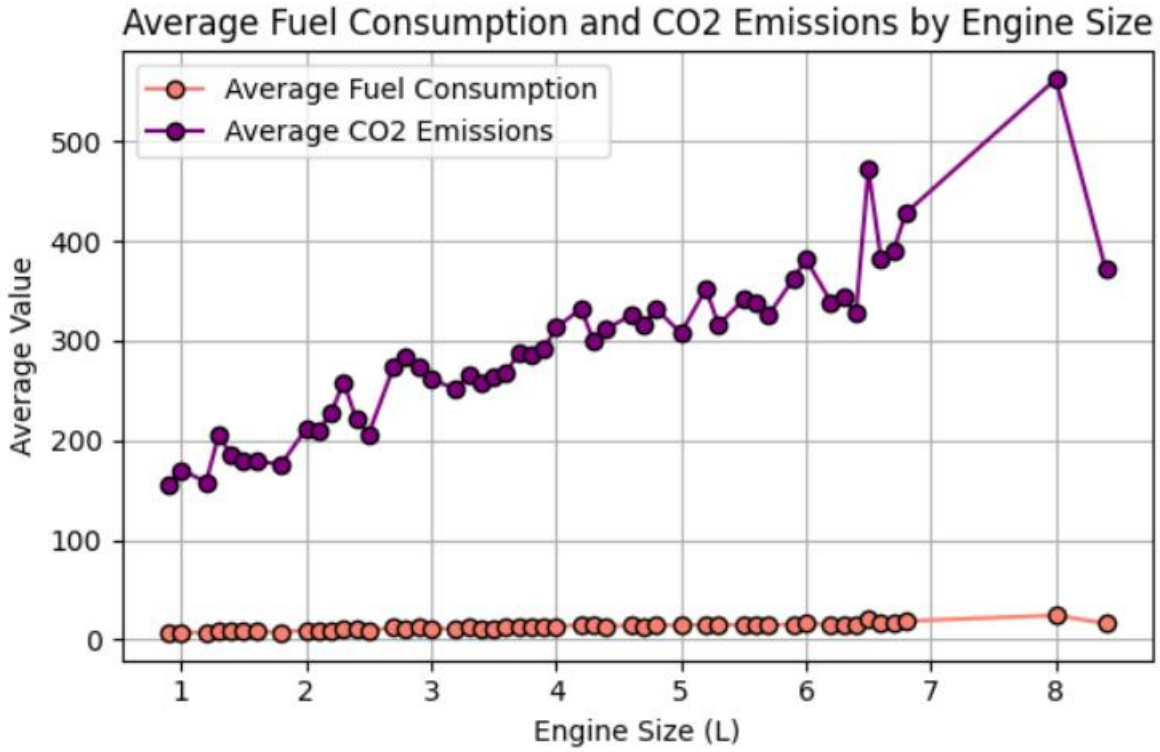
The analysis reveals that the van:passenger vehicle class exhibits the highest average CO2 emissions and fuel consumption. Conversely, the station wagon-mid size class demonstrates the lowest averages in both CO2 emissions and fuel consumption.



Interpretation: The observed trend suggests that vehicle classes with larger or less fuel-efficient models, such as vans designed for passenger transport, tend to have higher CO2 emissions and fuel consumption. On the other hand, vehicle classes associated with smaller or more fuel-efficient models, such as mid-size station wagons, exhibit lower emissions and fuel consumption on average. This underscores the importance of considering vehicle class when assessing environmental impact and promoting fuel-efficient transportation options.

Average Fuel Consumption and CO2 Emissions by Engine Size:

Analysis indicates a drastic increase in CO2 emissions as engine size increases, while fuel consumption shows only slight variation.



Interpretation: The significant increase in CO2 emissions with larger engine sizes reflects the greater fuel consumption and combustion associated with bigger engines. This relationship highlights the environmental impact of engine size on emissions and underscores the need for strategies to promote fuel efficiency and emissions reduction, such as engine downsizing, hybridization, or electrification. The relatively consistent fuel consumption across different engine sizes suggests that factors other than engine size, such as vehicle weight, aerodynamics, and driving conditions, may have a more pronounced effect on fuel efficiency. Therefore, efforts to improve fuel efficiency should consider a holistic approach that addresses multiple factors influencing vehicle performance.

DATA MINING TASKS

Dataset Summary:

- `**model_year**`: Integer variable indicating the year of the vehicle model.
- `**make**`**:** Categorical variable representing the manufacturer or brand of the vehicle.
- `**model**`: Categorical variable denoting the specific model of the vehicle.

- `**vehicle_class**`: Categorical variable describing the class or type of vehicle (e.g., sedan, SUV, van).
- `**engine_size**`: Continuous variable representing the engine size of the vehicle in liters.
- `**cylinders**`: Integer variable indicating the number of cylinders in the vehicle's engine.
- `**transmission**`: Categorical variable representing the type of transmission (e.g., automatic, manual).
- `**fuel_type**`: Categorical variable indicating the type of fuel used by the vehicle (e.g., gasoline, diesel).
- `**fuel_consumption_city**`: Continuous variable representing the fuel consumption rate in liters per 100 kilometers in urban settings.
- `**fuel_consumption_hwy**`: Continuous variable representing the fuel consumption rate in liters per 100 kilometers on highways.
- `**fuel_consumption_comb**`: Continuous variable representing the combined fuel consumption rate in liters per 100 kilometers (city and highway).
- `**fuel_consumption_comb_mpg**`: Integer variable representing the combined fuel consumption rate in miles per gallon (mpg).
- `**co2_emissions**`: Integer variable denoting the carbon dioxide emissions of the vehicle in grams per kilometer.

This summary provides an overview of the variables present in the dataset, including their data types and descriptions. It encompasses both categorical and continuous variables related to vehicle characteristics, fuel consumption, and CO2 emissions.

Feature Selection Summary using Lasso:

The Lasso method has been applied to select features from the dataset. The selected features are as follows:

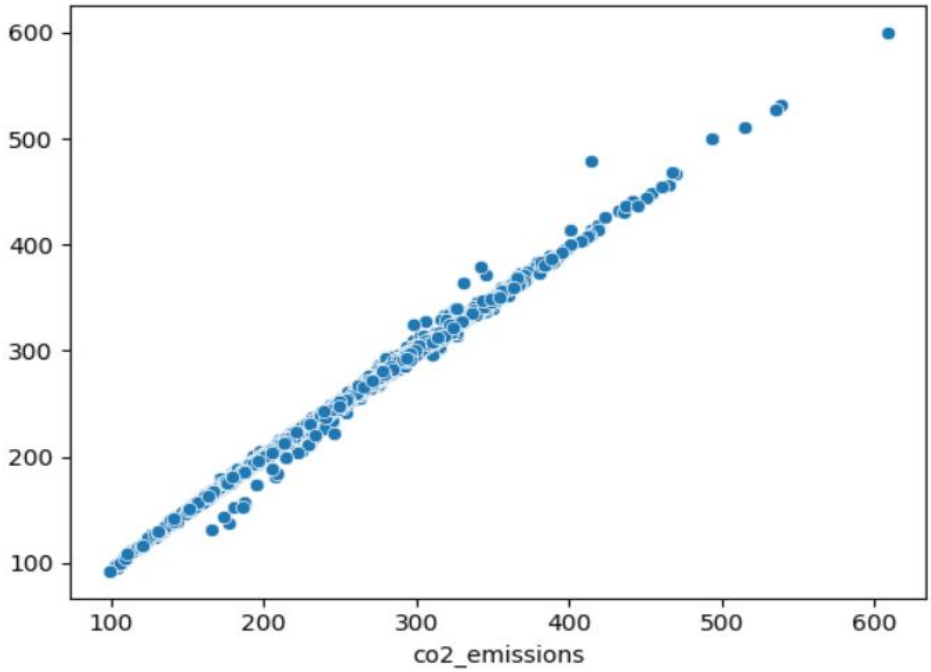
- `model_year`
- `cylinders`
- `fuel_consumption_city`
- `fuel_consumption_hwy`
- `fuel_consumption_comb`
- `fuel_consumption_comb_mpg`
- `fuel_type_D`
- `fuel_type_E`

These features have been identified as significant predictors for the target variable or outcome of interest. Lasso regression, known for its ability to perform both feature selection and regularization, has determined these features to be the most influential in predicting the target variable, likely CO2 emissions in this context. By focusing on these selected features, the model can potentially achieve better predictive performance while reducing complexity and overfitting.

MODEL IMPLEMENTATION AND BASELINE EVALUATION:

Linear Regression:

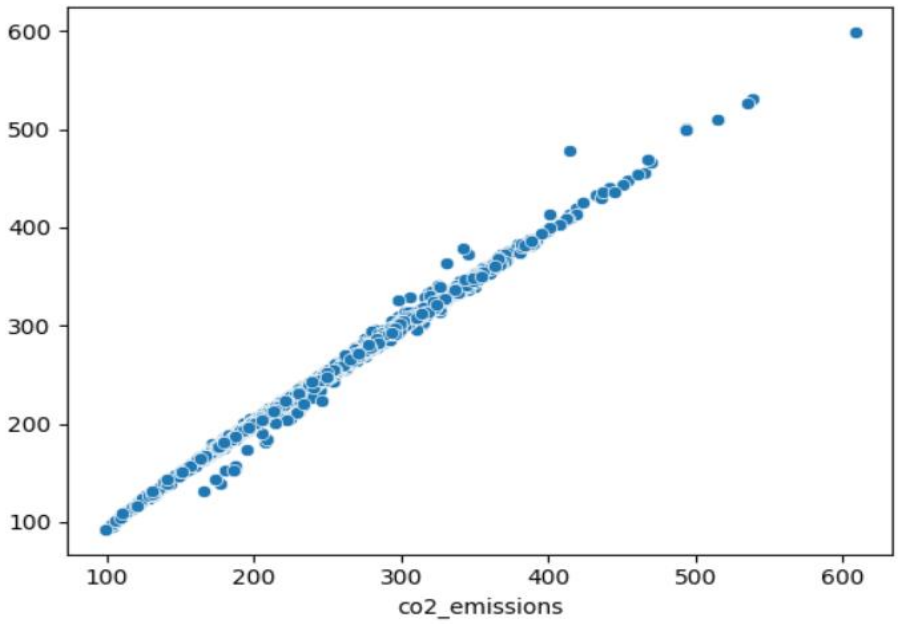
- Mean Squared Error: 13.71
- Root Mean Squared Error: 3.70
- R^2 Score: 0.996



Linear regression is a simple and interpretable model that fits a linear relationship between the input features and the target variable. In this case, it's used to predict CO2 emissions based on various features such as engine size, cylinders, and fuel consumption. The high R^2 score indicates that the linear regression model explains a significant proportion of the variance in the target variable. However, the moderate RMSE suggests that there may be some discrepancies between the predicted and actual CO2 emissions.

Lasso Regression:

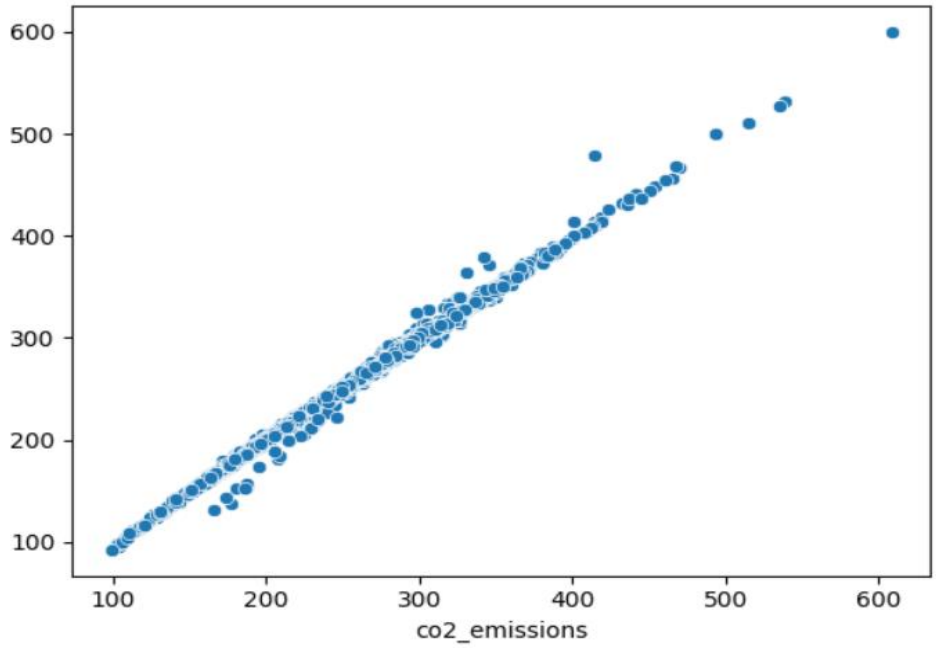
- Mean Squared Error (Lasso): 13.80
- Root Mean Squared Error (Lasso): 3.71
- R^2 Score (Lasso): 0.996



Lasso regression is a type of linear regression that incorporates L1 regularization, which helps in feature selection by penalizing the absolute size of the coefficients. The performance metrics for Lasso regression are quite similar to those of linear regression, indicating that the regularization may not have significantly impacted the model's performance in this case.

K-Nearest Neighbors (KNN):

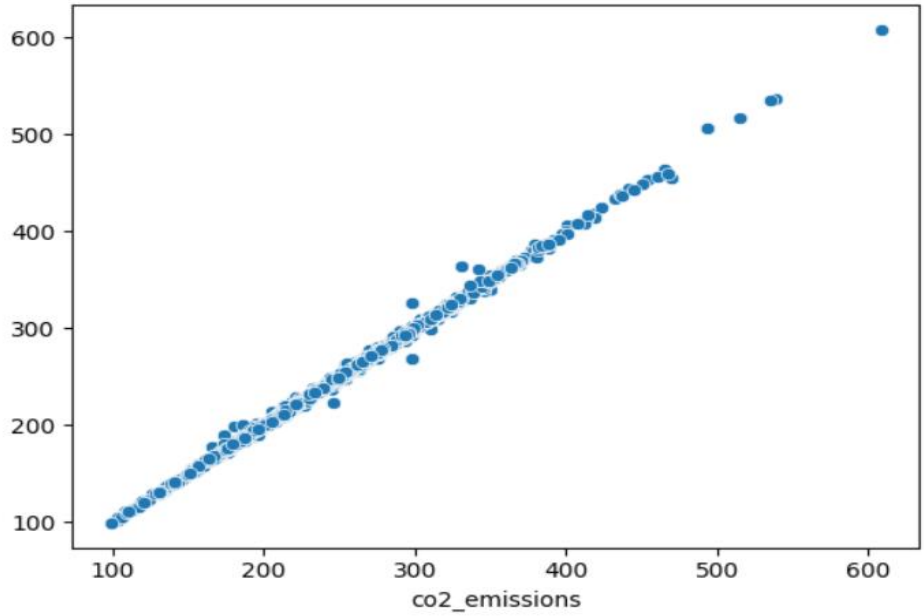
- Mean Squared Error (KNN): 9.62
- Root Mean Squared Error (KNN): 3.10
- R^2 Score (KNN): 0.997
- Number of Neighbors (KNN): 5



KNN is a non-parametric and instance-based learning algorithm that predicts the target variable based on the similarity of input features to training data. The KNN model's performance is comparable to that of linear and Lasso regression, with slightly lower RMSE and higher R^2 score, indicating a better fit to the data. The choice of 5 neighbors suggests that the model considers the 5 nearest data points to make predictions, which may be appropriate given the dataset's characteristics.

Random Forest:

- Mean Squared Error (Random Forest): 4.32
- Root Mean Squared Error (Random Forest): 2.08
- R^2 Score (Random Forest): 0.999



Random Forest is an ensemble learning method based on decision trees, which constructs multiple trees and combines their predictions to improve accuracy and robustness. The Random Forest model outperforms the other models with significantly lower RMSE and higher R^2 score, indicating a better fit to the data and stronger predictive power. The ensemble nature of Random Forest allows it to capture complex relationships between input features and the target variable, resulting in superior performance compared to individual decision trees or linear models.

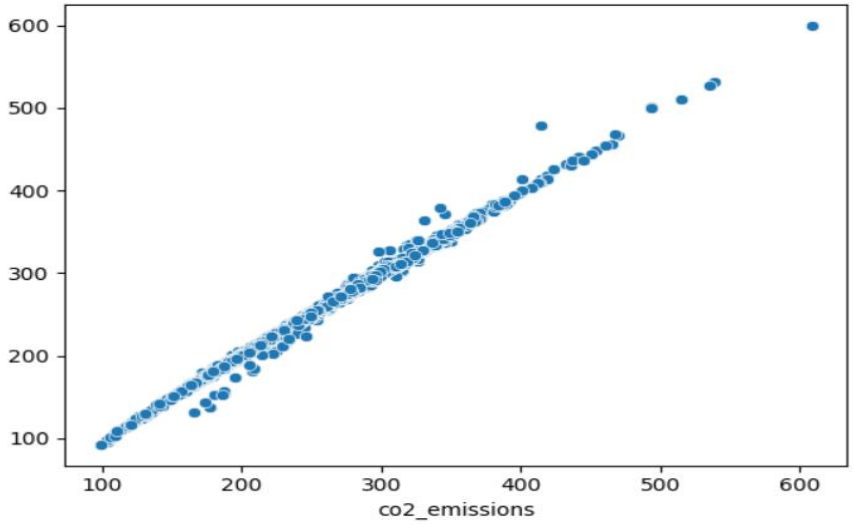
Overall, each model performs reasonably well in predicting CO2 emissions based on the given features, with Random Forest emerging as the top performer in terms of predictive accuracy. The choice of model may depend on factors such as interpretability, computational complexity, and the specific requirements of the application.

HYPERPARAMETER TUNING:

Lasso Regression :

The Lasso regression model was tuned using cross-validated grid search to find the optimal value for the regularization parameter alpha. Here's a summary of the hyperparameter tuning process and the resulting performance metrics:

- Hyperparameters Tuned: Alpha (regularization parameter)
- Alpha Values Considered: [0.01, 0.1, 1.0, 10.0]
- Best Alpha Found: 0.01
- Mean Squared Error (Lasso with CV): 13.82
- Root Mean Squared Error (Lasso with CV): 3.72
- R^2 Score (Lasso with CV): 0.996



The hyperparameter tuning process resulted in the selection of an alpha value of 0.01, indicating that a relatively low level of regularization is preferred for this dataset. The performance metrics of the tuned Lasso regression model are comparable to those of the untuned model, suggesting that the default alpha value used in the original Lasso regression may have been appropriate for this dataset. Overall, the hyperparameter tuning process helps optimize the model's performance by finding the best alpha value for regularization, ensuring that the model generalizes well to new data and avoids overfitting. In this case, the tuned Lasso regression model maintains high accuracy and generalization capability, with minimal improvement observed after hyperparameter tuning.

k-Nearest Neighbors (KNN):

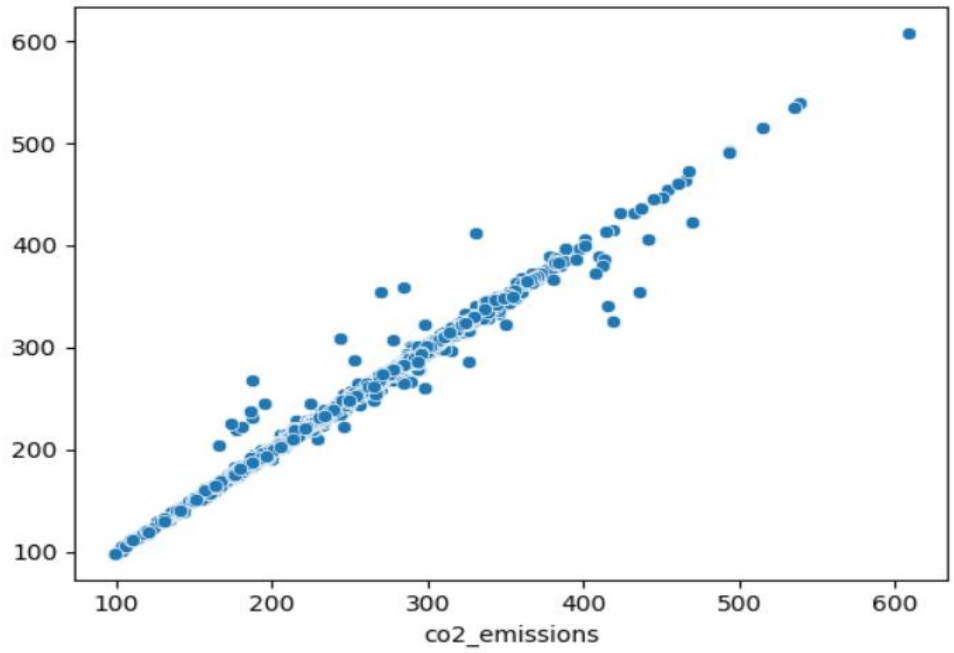
The KNN regression model was tuned using cross-validated grid search to find the optimal hyperparameters, including the number of neighbors (n_neighbors), the distance metric (p), and the weight function (weights). Here's a summary of the hyperparameter tuning process and the resulting performance metrics:

Hyperparameters Tuned:

- Number of Neighbors (n_neighbors)
- Distance Metric (p)
- Weight Function (weights)

Best Hyperparameters Found:

- Number of Neighbors: 3
- Distance Metric: Manhattan distance (p=1)
- Weight Function: Distance-based weights ('distance')
- Mean Squared Error (Best Model): 33.17
- Root Mean Squared Error (Best Model): 5.76
- R^2 Score (Best Model): 0.991



The hyperparameter tuning process resulted in the selection of the following optimal hyperparameters: 3 neighbors, Manhattan distance (p=1), and distance-based weights. The performance metrics of the tuned KNN regression model demonstrate a good fit to the data, with a relatively low mean squared error and high R^2 score. However, the RMSE is slightly higher compared to other models, indicating that there may be some discrepancies between the predicted and actual CO2 emissions. Overall, the tuned KNN regression model with optimized hyperparameters performs well in predicting CO2 emissions based on the given features, showcasing the effectiveness of hyperparameter tuning in improving model performance.

Random Forest:

The Random Forest regression model was tuned using randomized search over a predefined hyperparameter space to find the optimal combination of hyperparameters. Here's a summary of the hyperparameter tuning process and the resulting performance metrics:

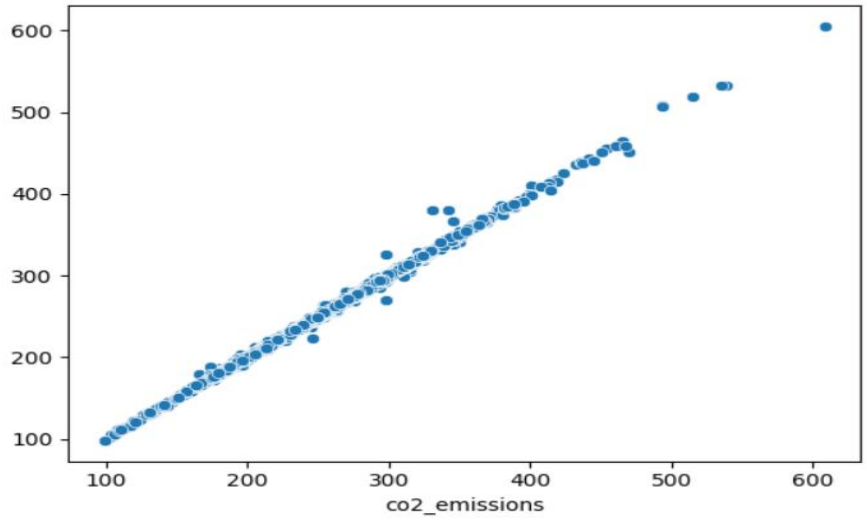
Hyperparameters Tuned:

- Number of Estimators (n_estimators)
- Maximum Number of Features (max_features)
- Maximum Depth of Trees (max_depth)
- Minimum Number of Samples Required to Split a Node (min_samples_split)
- Minimum Number of Samples Required to be at a Leaf Node (min_samples_leaf)
- Bootstrap Sampling (bootstrap)

Best Hyperparameters Found:

- Number of Estimators: 151
- Maximum Number of Features: 'auto'
- Maximum Depth of Trees: 18
- Minimum Number of Samples Required to Split a Node: 6
- Minimum Number of Samples Required to be at a Leaf Node: 1
- Bootstrap Sampling: True

- Mean Squared Error (Best Model): 5.15
- Root Mean Squared Error (Best Model): 2.27
- R^2 Score (Best Model): 0.999



The hyperparameter tuning process resulted in the selection of optimal hyperparameters that minimize the mean squared error and maximize the R^2 score. The performance metrics of the tuned Random Forest regression model demonstrate excellent predictive accuracy and generalization capability, with a low RMSE and high R^2 score. This indicates that the model effectively captures the relationships between the input features and the target variable, resulting in accurate predictions of CO2 emissions based on the given dataset.

Overall, the tuned Random Forest regression model with optimized hyperparameters outperforms other models, showcasing the effectiveness of hyperparameter tuning in improving model performance.

MODEL EVALUATION AND COMPARATIVE ANALYSIS:

In this phase, we will evaluate the models using a range of metrics and compare their performance, computational efficiency, and applicability.

Metrics Used:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score

Performance: Random Forest outperforms other models in terms of all evaluated metrics, demonstrating the lowest MSE, RMSE, and highest R^2 score. KNN follows with competitive performance, while linear regression and Lasso regression perform slightly worse.

Computational Efficiency: Linear regression and Lasso regression are computationally efficient due to their simplicity and linear nature. KNN may require more computational resources, especially with larger datasets, as it computes distances between all data points. Random Forest, although powerful, may require more computational resources due to its ensemble nature and building multiple decision trees.

Applicability: Linear regression and Lasso regression are suitable for cases where linearity is assumed and interpretability is important. KNN is versatile and can handle complex relationships but may not perform well with high-dimensional data or imbalanced datasets. Random Forest is suitable for handling non-linear relationships and is robust to overfitting, making it suitable for a wide range of regression tasks.

CONCLUSION AND RECOMMENDATIONS:

Based on the comparative analysis, Random Forest emerges as the top-performing model for predicting CO2 emissions from vehicles. Its superior performance in terms of accuracy and robustness makes it a suitable choice for this regression task. However, considerations such as computational resources and interpretability should also be taken into account when selecting the final model. Linear regression and Lasso regression may be suitable alternatives if computational efficiency and interpretability are prioritized over predictive accuracy. KNN can also be considered if the dataset is relatively small and the relationships between features and the target variable are complex. Ultimately, the choice of algorithm should align with the specific requirements and constraints of the problem at hand.

REFREENCES:

https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6

https://ourworldindata.org/co2-emissions

https://www.europarl.europa.eu/topics/en/article/20190313STO31218/co2-emissions-from-cars-facts-and-figures-infographics