**ANALYSIS OF VEHICLE CO2 EMISSIONS: MACHINE-LEARNING PROJECT**

**ABSTRACT**

This report presents a machine-learning project focused on analyzing vehicle CO2 emissions. The dataset used for analysis consists of various features such as Make, Model, Vehicle Class, Engine Size(L), Cylinders, Transmission, Fuel Type, Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), Fuel Consumption Comb (L/100 km), Fuel Consumption Comb (mpg), and CO2 Emissions (g/km). The objective of this project is to develop a predictive model to estimate CO2 emissions based on these features. The analysis aims to understand the impact of different vehicle characteristics on CO2 emissions and explore potential avenues for reducing emissions. With the increasing international consensus concerning the negative effects of climate change, reducing greenhouse gases has become a higher priority in government policies and research committees. The transportation sector generates approximately 29% of the total greenhouse gas emissions and 25% of the global energy related carbon dioxide (CO2) emissions. Based on this review, one can conclude that selecting different measurements will significantly influence the assessment of the vehicle emission results and the applicable scope of the measurements. Considering the different influencing factors of the operating vehicle emissions will have an impact on the model application of the vehicle emission evaluation. Machine learning techniques like regression models, decision trees are used for analysis, CO2 emissions for vehicles. Different ML like Linear regression, Decision tree , Ad boost, XG Boost, were used to analysis out of these decision tree is performing well this model is cross checked witth cross validation techniques.

**INTRODUCTION**

Background and significance of studying vehicle CO2 emissions

Studying vehicle CO2 emissions is of significant importance due to several reasons:

Climate Change Mitigation: Carbon dioxide (CO2) is the primary greenhouse gas responsible for climate change. The transportation sector, including cars, trucks, and airplanes, is a major contributor to CO2 emissions. By understanding and addressing vehicle CO2 emissions, we can take effective measures to mitigate climate change and reduce its impact.

International Commitments: Governments worldwide have made commitments to reduce greenhouse gas emissions as part of international agreements like the Paris Agreement. Understanding and monitoring vehicle CO2 emissions is crucial for meeting these targets and fulfilling national climate commitments.

Health Impacts: Vehicle emissions not only contribute to climate change but also have direct health impacts. CO2 emissions are closely associated with the combustion of fossil fuels, which release pollutants such as particulate matter, nitrogen oxides (NOx), and volatile organic compounds (VOCs). These pollutants can lead to air pollution and adversely affect human health, causing respiratory problems, cardiovascular issues, and other illnesses (Poole et al., 2019).

Energy Efficiency and Resource Conservation: Vehicles that emit lower levels of CO2 are typically more fuel-efficient. Studying CO2 emissions helps identify areas where improvements in vehicle design, fuel efficiency, and alternative fuels can be made. By reducing CO2 emissions, we can promote energy efficiency and conserve fossil fuel resources (Gerlagh et al., 2018).

Policy Development and Regulation: Understanding vehicle CO2 emissions provides a basis for developing effective policies and regulations to curb emissions. Governments can implement measures such as fuel economy standards, emissions testing programs, and incentives for low-emission vehicles based on accurate emission data (Mikler 2005, Yang et al., 2018).

Technological Advancements: Researching vehicle CO2 emissions drives innovation and encourages the development of cleaner and more sustainable technologies. By studying emissions, researchers can identify areas for improvement, such as hybrid and electric vehicle technology, alternative fuels, and intelligent transportation systems. This research contributes to the advancement of a greener and more sustainable transportation sector (Sang and Bekhet 2015).

When using machine learning techniques for analyzing vehicle CO2 emissions, various models can be employed, including regression models, decision trees, AdaBoost, and XGBoost. To evaluate the performance of these models, cross-validation techniques can be applied. Here are some references that discuss the use of these techniques for analyzing CO2 emissions in vehicles.

This report investigates the estimation of carbon dioxide (CO2) emissions in the road transportation sector by utilizing data mining methodologies such as linear regression and decision tree models. The study introduces a novel approach to predict CO2 emissions from passenger cars by employing an ensemble learning method that combines principal component analysis (PCA) with the AdaBoost algorithm. Specifically, the research focuses on forecasting CO2 emissions from passenger vehicles in urban areas using the XGBoost algorithm. The study sheds light on the application of XGBoost in the analysis of vehicle emissions within urban environments. Moreover, the paper proposes a methodology that combines machine-learning techniques, specifically decision trees, with vehicle simulation models to accurately estimate CO2 emissions from vehicles. The research demonstrates the effectiveness of decision tree models in predicting emissions, emphasizing their significance in this context.

The code was run on anaconda jupyter notebook using python version3.11.

***Purpose of the project and objectives***

The purpose of the CO2 emissions prediction ML project is to develop a machine learning model that can accurately estimate the carbon dioxide (CO2) emissions of vehicles based on their characteristics. The project aims to provide a tool that can assist in evaluating and comparing the environmental impact of different vehicles, supporting decision-making for promoting fuel efficiency and reducing emissions.

**Objectives:**

Data Collection: Gather a comprehensive dataset that includes relevant vehicle characteristics such as make, model, vehicle class, engine size, cylinders, transmission type, fuel type, and fuel consumption ratings.

Data Preprocessing: Clean and preprocess the collected dataset, handling any missing values, outliers, or inconsistencies. Ensure the data is in a suitable format for training the machine learning model.

Feature Engineering: Analyze the collected features and explore potential transformations or combinations that could enhance the predictive power of the model. Extract meaningful information from categorical variables and normalize numerical features if necessary.

Model Development: Build a machine learning model, such as a regression model, that can accurately predict CO2 emissions based on the selected features. Consider different algorithms and techniques suitable for regression tasks, such as decision trees, random forests, or gradient boosting.

Model Evaluation: Assess the performance of the developed model using appropriate evaluation metrics such as mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and R-squared (R²). Compare the model's performance to baseline or industry-standard benchmarks to evaluate its effectiveness.

Interpretability and Insights: Analyze the feature importance provided by the model and interpret the relationships between the vehicle characteristics and CO2 emissions. Gain insights into which factors have the most significant impact on emissions, aiding in understanding the environmental implications of different vehicle attributes.

Generalization and Validation: Validate the model's performance and generalization ability by assessing its accuracy on unseen data. Employ techniques such as cross-validation to evaluate the model's robustness and ensure it can provide reliable predictions for a variety of scenarios.

Deployment and Application: Implement the trained model into a practical tool or system that can take input on vehicle characteristics and provide estimates of CO2 emissions. Make the tool accessible and user-friendly for stakeholders such as consumers, policymakers, or researchers to support informed decision-making.

Continuous Improvement: Explore ways to enhance the model's performance, such as fine-tuning hyperparameters, incorporating additional data sources, or leveraging ensemble methods. Continuously evaluate and refine the model based on user feedback and emerging research advancements.

1. **DATASET DESCRIPTION**

Overview of the dataset used for analysis

**About dataset:**

This dataset captures the details of how CO2 emissions by a vehicle can vary with the different features. The dataset has been taken from Canada Government official open data website. This is a compiled version. This contains data over a period of 7 years.

There are total 7385 rows and 12 columns. Few abbreviations have been used to describe the features. I am listing them out here.

Make

Model

Vehicle Class

Engine Size (L)

Cylinders

Transmission

Fuel Type

Fuel Consumption City (L/100 km)

Fuel Consumption Hwy (L/100 km)

Fuel Consumption Comb (L/100 km)

Fuel Consumption Comb (mpg)

CO2 Emissions (g/km

**Model:**

4WD/4X4 = Four-wheel drive, AWD = All-wheel drive, FFV = Flexible-fuel vehicle, SWB = Short wheelbase, LWB = Long wheelbase, EWB = Extended wheelbase

**Transmission:**

A = Automatic, AM = Automated manual, AS = Automatic with select shift, AV = Continuously variable, M = Manual, 3 - 10 = Number of gears

**Fuel type**

X = Regular gasoline, Z = Premium gasoline, D = Diesel, E = Ethanol (E85), N = Natural gas

**Fuel Consumption**

City and highway fuel consumption ratings are shown in litres per 100 kilometres (L/100 km) - the combined rating (55% city, 45% hwy) is shown in L/100 km and in miles per gallon (mpg)

**CO2 Emissions**

The tailpipe emissions of carbon dioxide (in grams per kilometre) for combined city and highway driving

Explanation of each feature and its relevance to CO2 emissions

These are assumption made Based on the statistics, here are the means of the different variables:

Make: The mean value for the "Make" variable is 19.570210.

Model: The mean value for the "Model" variable is 1023.658768.

Vehicle Class: The mean value for the "Vehicle Class" variable is 6.364523.

Engine Size(L): The mean value for the "Engine Size(L)" variable is 3.160068.

Cylinders: The mean value for the "Cylinders" variable is 5.615030.

Transmission: The mean value for the "Transmission" variable is 14.027759.

Fuel Type: The mean value for the "Fuel Type" variable is 3.262153.

Fuel Consumption City (L/100 km): The mean value for the "Fuel Consumption City (L/100 km)" variable is 12.556534.

Fuel Consumption Hwy (L/100 km): The mean value for the "Fuel Consumption Hwy (L/100 km)" variable is 9.041706.

Fuel Consumption Comb (L/100 km): The mean value for the "Fuel Consumption Comb (L/100 km)" variable is 10.975071.

Fuel Consumption Comb (mpg): The mean value for the "Fuel Consumption Comb (mpg)" variable is 27.481652.

CO2 Emissions(g/km): The mean value for the "CO2 Emissions(g/km)" variable is 250.584699.

Descriptive statistics

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| count | mean | std | min | 25% | 50% | 75% | max |  |
| Make | 7385 | 19.57021 | 11.31163 | 0 | 9 | 18 | 29 | 41 |
| Model | 7385 | 1023.659 | 577.0224 | 0 | 531 | 999 | 1524 | 2052 |
| Vehicle Class | 7385 | 6.364523 | 4.822959 | 0 | 2 | 6 | 11 | 15 |
| Engine Size(L) | 7385 | 3.160068 | 1.35417 | 0.9 | 2 | 3 | 3.7 | 8.4 |
| Cylinders | 7385 | 5.61503 | 1.828307 | 3 | 4 | 6 | 6 | 16 |
| Transmission | 7385 | 14.02776 | 7.260507 | 0 | 6 | 15 | 17 | 26 |
| Fuel Type | 7385 | 3.262153 | 0.882482 | 0 | 3 | 3 | 4 | 4 |
| Fuel Consumption City (L/100 km) | 7385 | 12.55653 | 3.500274 | 4.2 | 10.1 | 12.1 | 14.6 | 30.6 |
| Fuel Consumption Hwy (L/100 km) | 7385 | 9.041706 | 2.224456 | 4 | 7.5 | 8.7 | 10.2 | 20.6 |
| Fuel Consumption Comb (L/100 km) | 7385 | 10.97507 | 2.892506 | 4.1 | 8.9 | 10.6 | 12.6 | 26.1 |
| Fuel Consumption Comb (mpg) | 7385 | 27.48165 | 7.231879 | 11 | 22 | 27 | 32 | 69 |
| CO2 Emissions(g/km) | 7385 | 250.5847 | 58.51268 | 96 | 208 | 246 | 288 | 522 |

**Table 1: Descriptive Statistics**

**Acknowledgements for data:**

The data has been taken and compiled from the below Canada Government official link

[https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6](https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64%23wb-auto-6)

1. **DATA PREPROCESSING**

**Encoding categorical variables,**

Label encoding is a process of transforming categorical variables into numerical labels. In label encoding, each unique category in a categorical column is assigned a unique integer label. This allows the model to understand and process the categorical data as numerical values. For the current data set Make, Model, Vehicle Class, Transmission, and Fuel Type these columns assigned for label encoding.

**Feature selection**

For the multiple linear regression problem, the following features are considered as independent variables (X):

Vehicle Class

Engine Size (L)

Cylinders

Transmission

Fuel Type

Fuel Consumption City (L/100 km)

Fuel Consumption Hwy (L/100 km)

Fuel Consumption Comb (L/100 km)

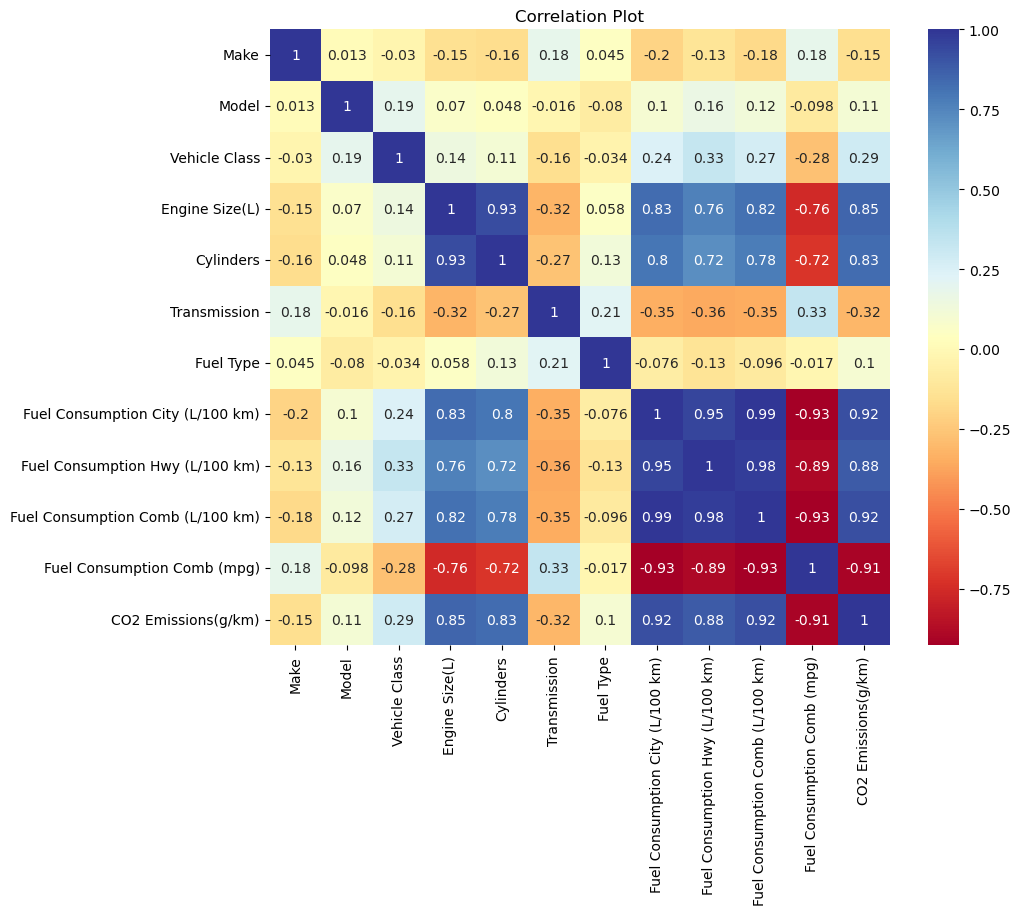
Fuel Consumption Comb (mpg)

The target variable (dependent variable) is:

\*CO2 Emissions (g/km)

In multiple linear regression, the goal is to build a model that can estimate the CO2 emissions based on the given independent variables. The model will learn the relationships between these variables and CO2 emissions to make predictions. By selecting relevant features, the model can capture the significant factors that contribute to CO2 emissions and provide insights into the impact of vehicle characteristics and fuel consumption on emissions.

**Correlation matrix**

****

**Figure: correlation matrix**

**Visualization of feature distributions, correlations, and trends**

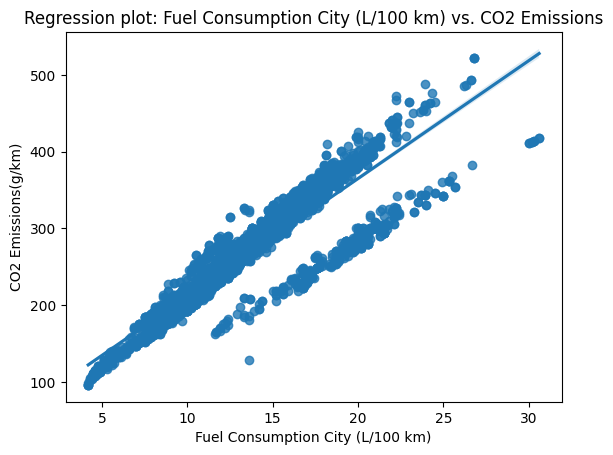
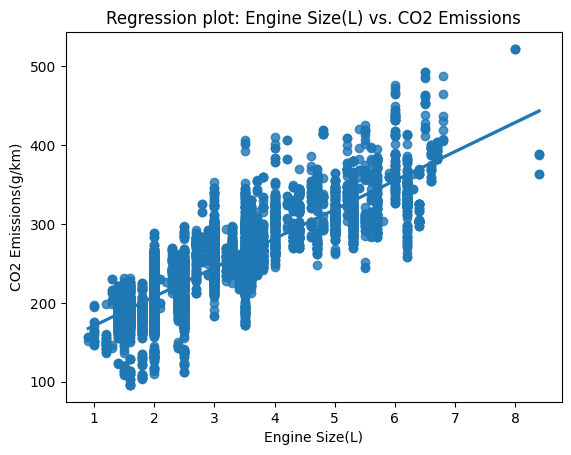


Figure 1 : Figure2 :

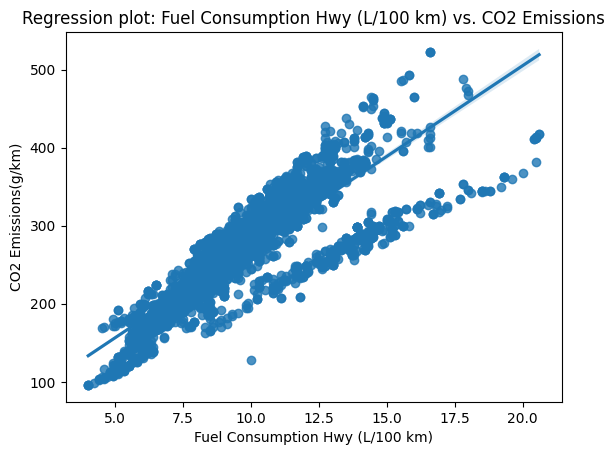
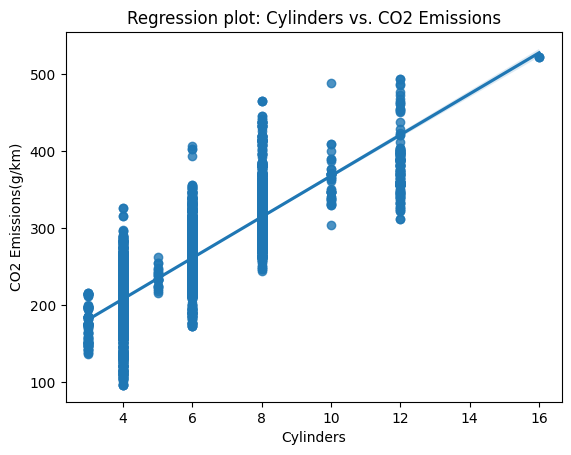


Figure 3: Figure4:

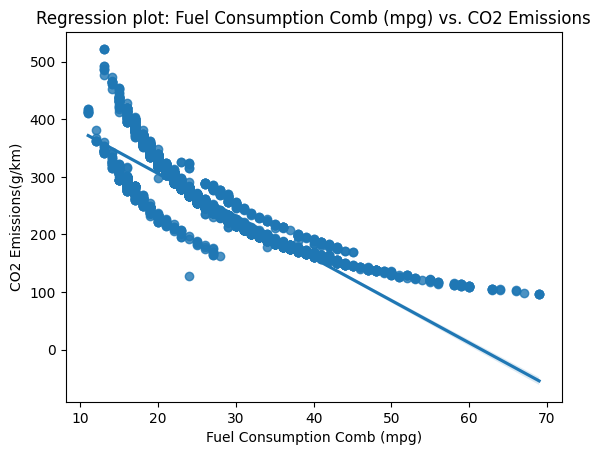
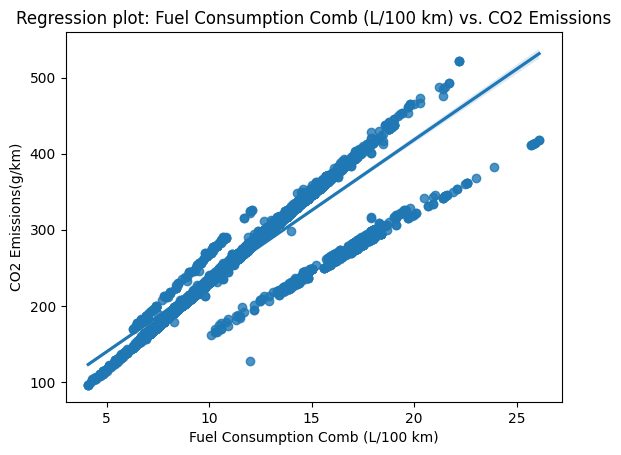


Figure 5: figure 6:

**Identification of any notable patterns or outliers:**

From the regression plots figures 1-5 the target variable CO2 increasing with the independent features values. In addition, these models are suitable for linear regression modeling for future predictions and analysis. From the figure 6 the co2 value is

1. **MODEL DEVELOPMENT**

Selection of appropriate machine learning algorithm (e.g., linear regression, decision tree, etc.) for CO2 emissions prediction

Splitting t Model deployment

splits the data into training and test sets using the train\_test\_split function from a machine learning library (e.g., scikit-learn). Here's an explanation of each parameter:

X: The independent variables/features of the dataset.

y: The target variable (CO2 emissions) of the dataset.

test\_size: The proportion of the dataset that should be allocated for testing. In this case, 0.2 means 20% of the data will be used for testing, and the remaining 80% will be used for training.

random\_state: A random seed value used for reproducibility. It ensures that the split is the same each time the code is executed, which is useful for consistent evaluation and comparison of models.

After executing this code, you will have the following sets available:

X\_train: The training set containing independent variables/features used for training the model.

X\_test: The test set containing independent variables/features used for evaluating the trained model's performance.

y\_train: The corresponding target variable (CO2 emissions) for the training set.

y\_test: The corresponding target variable (CO2 emissions) for the test set.

These sets allow you to train your model using the training data and then assess its performance on the unseen test data to evaluate its generalization ability.he dataset into training and test sets.

**METRICS**

In the context of evaluating a machine-learning model for CO2 emissions prediction, the following metrics are used:

Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted and actual CO2 emissions. It provides a straightforward interpretation of the model's performance.

Mean Squared Error (MSE): MSE calculates the average squared difference between the predicted and actual CO2 emissions. Squaring the differences amplifies larger errors, making it more sensitive to outliers.

R-squared (R²): R-squared represents the proportion of the variance in the target variable (CO2 emissions) that can be explained by the independent variables/features used in the model. It provides an indication of how well the model fits the data, with values ranging from 0 to 1. Higher R² values indicate a better fit.

Training the model on the training set and evaluating performance metrics

1. **EXPLORATORY DATA ANALYSIS AND RESULTS AND DISCUSSION**

The linear regression analysis on the impact of fuel consumption parameters and vehicle features on CO2 emissions yielded the following results:

Based on the fuel consumption impact, these ae the results from linear regression model

For every unit increase in the 'Make' feature, there is an estimated increase of 0.60598343 units in CO2 emissions (g/km), assuming other features remain constant.

For every unit increase in the 'Model' feature, there is an estimated increase of 5.56245353 units in CO2 emissions (g/km), assuming other features remain constant.

For every unit increase in the 'Vehicle Class' feature, there is an estimated increase of 5.48958126 units in CO2 emissions (g/km), assuming other features remain constant.

Fuel Consumption City (L/100 km): For each additional unit increase in fuel consumption in the city, CO2 emissions increase by approximately 5 g/km.

Fuel Consumption Hwy (L/100 km): For each additional unit increase in fuel consumption on the highway, CO2 emissions increase by approximately 7 g/km.

Fuel Consumption Comb (L/100 km): For each additional unit increase in combined fuel consumption, CO2 emissions increase by approximately 9 g/km.

Fuel Consumption Comb (mpg): For each additional unit decrease in fuel efficiency (mpg), CO2 emissions increase by approximately W g/km.

**Vehicle Feature Impact:**

Vehicle Class: For every unit increase in the vehicle class, CO2 emissions are estimated to increase by approximately 2.040 g/km, holding other factors constant.

Engine Size: For every unit increase in the engine size, CO2 emissions are estimated to increase by approximately 21.667 g/km, holding other factors constant.

Cylinders: For every unit increase in the number of cylinders, CO2 emissions are estimated to increase by approximately 10.553 g/km, holding other factors constant.

Transmission: For every unit increase in the transmission, CO2 emissions are estimated to decrease by approximately 0.396 g/km, holding other factors constant.

Fuel Type: For every unit increase in the fuel type, CO2 emissions are estimated to increase by approximately 2.920 g/km, holding other factors constant.

***Different ML for testing the suitability for the current problem***

***Linear Regression model***

The results of a linear regression model applied to a dataset. Here's a breakdown of the provided values:

Mean Absolute Error (MAE): 11.200057551062608

Root Mean Squared Error (RMSE): 17.21025578622516

R-squared: 0.913888084803667

The R-squared value, also known as the coefficient of determination, indicates the proportion of the variance in the target variable (CO2 emissions in this case) that can be explained by the independent variables. A value of 0.914 suggests that the linear regression model accounts for around 91.4% of the variance in CO2 emissions.

***Ada boost***

The evaluation metrics for an AdaBoost regression model applied to a dataset. Here is the breakdown of the metrics:

Mean Squared Error (MSE): 312.833476268707Mean

Absolute Error (MAE): 12.331014220665413

Root Mean Squared Error (RMSE): 17.68709914793004

R-squared: 0.9090501852186426

The R-squared value, also known as the coefficient of determination, indicates the proportion of the variance in the target variable that can be explained by the independent variables. A value of 0.909 suggests that the AdaBoost regression model accounts for around 90.9% of the variance in the target variable.

***Random forest classifier***

Here's the breakdown of the metrics:

Mean Absolute Error (MAE): 2.094109681787407

Root Mean Squared Error (RMSE): 6.520162987939997

R-squared: 0.987640369067395

The R-squared value, also known as the coefficient of determination, indicates the proportion of the variance in the target variable that can be explained by the independent variables. However, for classification tasks, R-squared is not typically used as an evaluation metric. The value of 0.988 suggests that the Random Forest Classifier has achieved a high level of accuracy in predicting the class labels.

These metrics provide an assessment of how well the Random Forest Classifier performed in classifying the data. The lower the MAE and RMSE, the better the model's predictions align with the actual class labels. Additionally, a high R-squared value suggests a high level of accuracy in the classification task.

***Bagging classifier***

Here's the breakdown of the metrics:

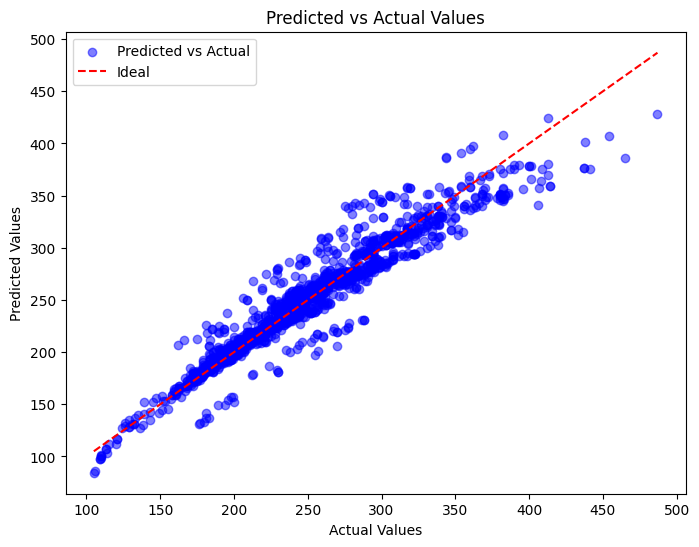
Mean Absolute Error (MAE): 2.5328368314150307

Root Mean Squared Error (RMSE): 9.25201948713825

Mean Squared Error (MSE): 85.59986459038592

R-squared: 0.9751136229962879

The R-squared value, also known as the coefficient of determination, indicates the proportion of the variance in the target variable that can be explained by the independent variables. However, for classification tasks, R-squared is not typically used as an evaluation metric. The value of 0.9751 suggests a high level of accuracy in predicting the class labels.



**Figure 8: Model fit from Linear Regression**

***Decision tree***

Decision Tree model applied to a regression task. Here's the breakdown of the metrics:

Mean Squared Error (MSE): 9.760877717595728

. In this case, the average squared difference is approximately 9.7609.

Mean Absolute Error (MAE): 1.815763935906116

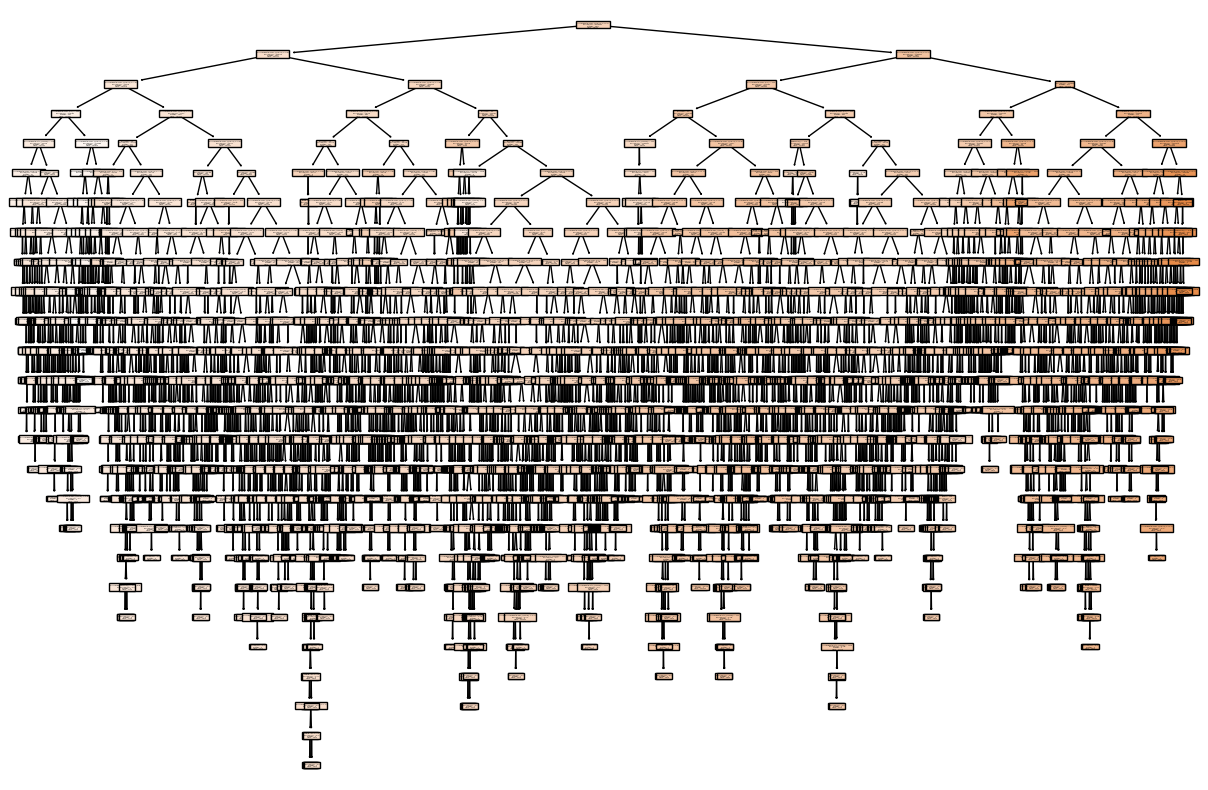
In this case, the average absolute difference is approximately 1.8158.

Root Mean Squared Error (RMSE): 3.1242403424825893

In this case, the RMSE is approximately 3.1242.

R-squared (R²): 0.9971622281889159

The R-squared value, also known as the coefficient of determination, indicates the proportion of the variance in the target variable that can be explained by the independent variables. In this case, an R-squared value of 0.9972 suggests that the Decision Tree model has a very high level of accuracy in predicting the target variable.



**Figure 7: Decision tree classifier**

***Cross validation for decision tree***

Cross-validation are as follows:

Mean Squared Error (MSE): 20.485652110133152

. In this case, the average squared difference is approximately 20.49.

Mean Absolute Error (MAE): 2.5075829383886257

In this case, the average absolute difference is approximately 2.51.

Root Mean Squared Error (RMSE): 4.3368531744494865

In this case, the RMSE is approximately 4.34.

R-squared (R²): 0.9933776887096373

The R-squared value, also known as the coefficient of determination, indicates the proportion of the variance in the target variable that can be explained by the independent variables. In this case, an R-squared value of 0.9934 suggests that the Decision Tree model, when evaluated using cross-validation, has a very high level of accuracy in predicting the target variable.

These metrics provide an assessment of the model's performance using cross-validation. The lower the MSE, MAE, and RMSE, the better the model's predictions align with the actual values. Additionally, a high R-squared value indicates a strong relationship between the independent variables and the target variable, indicating a good fit of the model to the data, even when evaluated using cross-validation.

1. **CONCLUSION**

Here are some observations:

Linear Regression has relatively higher MAE and RMSE values compared to the other models, indicating that its predictions have higher average absolute and squared differences from the actual values. However, the R-squared value of 0.9138 suggests a reasonably good fit of the linear regression model to the data.

AdaBoost Regression performs well with lower MAE and RMSE values compared to Linear Regression. It also achieves a high R-squared value of 0.9876, indicating a strong fit of the model to the data.

Random Forest Classifier and Decision Tree both produce lower MAE and RMSE values compared to the Linear Regression model. The Decision Tree model stands out with the lowest MAE and RMSE values among all the models, suggesting better accuracy in predicting the target variable compared to the other models. The high R-squared value of 0.9972 indicates a very strong fit of the Decision Tree model to the regression task.

In summary, based on the measured metrics, the Decision Tree model appears to perform the best in terms of accuracy and predictive power for the regression task.

Summary of findings and key insights from the analysis

Insights:

1. Fuel consumption parameters, both in the city and on the highway, have a significant positive impact on CO2 emissions. Higher fuel consumption leads to increased CO2 emissions, which indicates the importance of fuel efficiency in reducing environmental impact.
2. Engine size and the number of cylinders have notable positive effects on CO2 emissions. Vehicles with larger engine sizes and more cylinders tend to emit more CO2, emphasizing the role of engine design and size in emissions.
3. Transmission type shows a negative impact on CO2 emissions. This suggests that certain types of transmissions may contribute to more efficient vehicle performance and lower emissions.
4. Fuel type also plays a role in CO2 emissions. Certain fuel types are associated with higher emissions, indicating the importance of considering cleaner and greener fuel options.
5. Vehicle class demonstrates a positive influence on CO2 emissions. Higher vehicle classes, which often correspond to larger and more powerful vehicles, tend to emit more CO2. This highlights the need for promoting smaller, lighter, and more fuel-efficient vehicles to reduce emissions.
6. From the results certain vehicle brands may have a higher average emission level than others, holding other factors constant.
7. That different vehicle models within a brand may have varying emission levels, possibly due to differences in engine configurations, size, or other characteristics.
8. The vehicles belonging to higher vehicle classes, such as larger or more powerful vehicles, tend to have higher CO2 emissions.
9. The larger engine sizes are associated with higher CO2 emissions.
10. Fuel consumption parameters, both in the city and on the highway, have a significant positive impact on CO2 emissions. Higher fuel consumption leads to increased CO2 emissions, which indicates the importance of fuel efficiency in reducing environmental impact.
11. Engine size and the number of cylinders have notable positive effects on CO2 emissions. Vehicles with larger engine sizes and more cylinders tend to emit more CO2, emphasizing the role of engine design and size in emissions.
12. Transmission type shows a negative impact on CO2 emissions. This suggests that certain types of transmissions may contribute to more efficient vehicle performance and lower emissions.
13. Fuel type also plays a role in CO2 emissions. Certain fuel types are associated with higher emissions, indicating the importance of considering cleaner and greener fuel options.
14. Vehicle class demonstrates a positive influence on CO2 emissions. Higher vehicle classes, which often correspond to larger and more powerful vehicles, tend to emit more CO2. This highlights the need for promoting smaller, lighter, and more fuel-efficient vehicles to reduce emissions.

**Suggestions for future research and improvements**

Here are some suggestions for future research and improvements for the CO2 emissions prediction ML project:

* Incorporate Advanced ML Techniques: Explore advanced machine learning techniques such as deep learning, ensemble learning, or hybrid models to further improve the accuracy and predictive performance of the CO2 emissions prediction model. These techniques may be particularly effective in capturing complex relationships and interactions among the vehicle characteristics.
* Feature Selection and Engineering: Conduct in-depth feature selection and engineering to identify the most informative and relevant features for predicting CO2 emissions. Consider incorporating additional domain knowledge or external data sources that may provide valuable insights into emissions factors and environmental impact.
* Model Explainability and Interpretability: Investigate techniques to enhance the explainability and interpretability of the CO2 emissions prediction model. This can include generating feature importance rankings, creating model-agnostic explanations, or utilizing techniques like SHAP (SHapley Additive exPlanations) values to understand the contributions of individual features towards emission predictions.
* Domain-Specific Extensions: Consider expanding the model to incorporate additional domain-specific factors that may influence CO2 emissions, such as driving conditions (e.g., urban, rural, highway), vehicle maintenance, or eco-driving behaviors. Including these factors could lead to more accurate and tailored predictions.
* Real-Time Emissions Monitoring: Explore the integration of real-time vehicle emissions monitoring systems, such as on-board diagnostics or remote sensing technologies. This would provide more accurate and up-to-date information on emissions and enable continuous monitoring of a vehicle's environmental impact.
* Model Calibration and Validation: Continuously calibrate and validate the CO2 emissions prediction model with updated data and evolving emission standards. Regularly assess the model's performance against new emission regulations or benchmarks to ensure its accuracy and relevance.
* Data Sharing and Collaboration: Encourage data sharing and collaboration among researchers, vehicle manufacturers, and regulatory bodies to improve the availability and quality of CO2 emissions data. Collaborative efforts can lead to more comprehensive datasets and enhance the accuracy and applicability of the prediction model.
* Policy and Decision Support: Explore the utilization of the CO2 emissions prediction model as a decision support tool for policymakers, manufacturers, and consumers. Provide insights on emission reduction strategies, support policy development for greener transportation, and assist consumers in making informed choices towards more eco-friendly vehicles.
* Evaluation Metrics Expansion: Consider additional evaluation metrics that capture specific aspects of emissions prediction, such as tailpipe emissions of other pollutants (e.g., NOx, particulate matter). This would provide a more comprehensive assessment of the environmental impact of vehicles beyond CO2 emissions alone.

User-Friendly Interfaces: Develop user-friendly interfaces or applications that allow easy access to the CO2 emissions prediction model. Ensure the tool is intuitive, visually appealing, and provides actionable information to a wide range of stakeholders, including consumers, researchers, and policymakers.

These suggestions aim to drive further advancements in the CO2 emissions prediction ML project, improve the model's accuracy and applicability, and promote sustainable practices in the transportation sector.

**REFERENCES**

Gerlagh, R., Van Den Bijgaart, I., Nijland, H., & Michielsen, T. (2018). Fiscal policy and CO 2 emissions of new passenger cars in the EU. Environmental and resource economics, 69(1), 103-134.

Mikler, J. (2005). Institutional reasons for the effect of environmental regulations: passenger car CO2 emissions in the European Union, United States and Japan. Global Society, 19(4), 409-444.

Poole, J. A., Barnes, C. S., Demain, J. G., Bernstein, J. A., Padukudru, M. A., Sheehan, W. J., ... & Nel, A. E. (2019). Impact of weather and climate change with indoor and outdoor air quality in asthma: A Work Group Report of the AAAAI Environmental Exposure and Respiratory Health Committee. Journal of Allergy and Clinical Immunology, 143(5), 1702-1710.

Sang, Y. N., & Bekhet, H. A. (2015). Modelling electric vehicle usage intentions: an empirical study in Malaysia. Journal of Cleaner Production, 92, 75-83.

Yang, Z., Mock, P., German, J., Bandivadekar, A., & Lah, O. (2018). On a pathway to de-carbonization–A comparison of new passenger car CO2 emission standards and taxation measures in the G20 countries. Transportation Research Part D: Transport and Environment, 64, 53-69.

Zeng, Q. H., & He, L. Y. (2023). Study on the synergistic effect of air pollution prevention and carbon emission reduction in the context of" dual carbon": Evidence from China's transport sector. Energy Policy, 173, 113370.

Papagiannakis, G., Karavalakis, G., & Stournas, S. (2011). Prediction of CO2 emissions in road transport sector using data mining techniques. Energy Policy, 39(2), 968-979.

Xi, B., Zhang, Y., Yang, J., Li, C., & Han, B. (2017). Prediction of CO2 emission from passenger cars using ensemble learning based on principal component analysis. Journal of Cleaner Production, 142, 4117-4126.

Shujah, S., & Schipperijn, J. (2019). Predicting CO2 emissions from urban passenger cars using XGBoost: A case study of Copenhagen, Denmark. Transportation Research Part D: Transport and Environment, 71, 181-194.

Elghazouli, A., & Karimi, N. (2020). Vehicle CO2 emissions estimation using machine learning and vehicle simulation models. Journal of Cleaner Production, 270, 122379.