1. **Introduction**

**Purpose of the Project**:

Vehicle emissions are a major source of air pollution, which has significant health and environmental consequences. In the UK, the government has established strict regulations to manage and reduce these emissions to protect public health and the environment. For instance, the UK is committed to reducing greenhouse gas emissions by 68% by 2030 under the Climate Change Act. The purpose of this project is to see if one can infer a vehicles emissions given the vehicles properties.

**Overview of the Dataset**:

The dataset used in this project contains vehicle specifications and their associated emissions. It consists of detailed records for 639 vehicles, capturing several key characteristics:

* **Year**: All vehicles in this dataset are from the year 2000.
* **Make**: This refers to the brand or manufacturer of the vehicle.
* **Model**: Specific model of the vehicle.
* **Vehicle Class**: The category of the vehicle (e.g., SUV, sedan, hatchback).
* **Engine Size**: The volume of the engine, typically measured in litres.
* **Cylinders**: The number of cylinders in the vehicle’s engine.
* **Transmission**: Type of gearbox used in the vehicle.
* **Fuel Type**: What kind of fuel the vehicle uses (e.g., gasoline, diesel).
* **Fuel Consumption**: Average fuel consumption, measured in litres per 100 kilometres.
* **CO2 Emissions**: Amount of carbon dioxide (CO2) the vehicle emits, typically measured in grams per kilometre.

The data is sourced from publicly available records(\*See Annex) that track vehicle specifications and emissions, possibly collected by a governmental or regulatory body overseeing transportation and environmental standards. Such datasets are often used to monitor compliance with environmental regulations and to study the impact of various vehicle types on air quality.

This dataset offers a comprehensive view of how different factors, like engine size or fuel type, contribute to the environmental impact of vehicles, making it a valuable resource for understanding and predicting vehicle emissions.

1. **Methodology**

**Data Preparation**:

The dataset was pre-cleaned with only 1 duplicate row that was removed in preprocessing.

**Model Selection**:

In the development of our predictive model, we carefully selected features that significantly influence vehicle emissions. The initial analysis included “Fuel Consumption” as a feature, which naturally has a strong correlation with emissions. However, to focus on underlying factors that affect fuel consumption and provide a broader perspective, we chose to omit this feature from the final model. This approach ensures that our model can identify other significant contributors to emissions without being overly dependent on direct consumption figures.

1. **Results**

**Model Performance**:

Model testing included several machine learning techniques, with Gradient Boosting providing the most accurate predictions:

**Initial Model Performance**:

* Linear Regression MSE: 1014.47
* Random Forest MSE: 375.49
* Gradient Boosting MSE: 378.58

After tuning the hyperparameters of our Gradient Boosting model, we achieved even better performance:

* **Optimized Gradient Boosting MSE**: 262.29

The optimised model, using parameters {'learning\_rate': 0.2, 'max\_depth': 4, 'n\_estimators': 150}, improved its accuracy significantly, highlighting the effectiveness of the model in predicting vehicle emissions.

This model's high R-squared score of 99% initially with fuel consumption and the excellent performance in subsequent validations (95% with fuel consumption omitted) indicate its strong predictive power and reliability, providing valuable tools for monitoring and planning in compliance with UK emission standards.

1. **Discussion**

**What the Results Mean**:

The performance metrics from our predictive model offer significant insights into its accuracy and usefulness in practical scenarios. The model achieved a low Mean Squared Error (MSE) and a high R-squared value, which are both indicators of the model's predictive capabilities:

* **Mean Squared Error (MSE)**: This metric measures the average squared difference between the actual and predicted emissions. Lower values indicate higher accuracy. For instance, the optimized Gradient Boosting model's MSE of 262.29 means the model's predictions are close to the actual emission values, which suggests a high level of precision in forecasts.
* **R-squared (R²)**: This value tells us how well the variations in emissions are being explained by the model. An R-squared score close to 95% (without considering fuel consumption) indicates that nearly all the variability in emissions across different vehicles can be explained by the model. This high score means that our model is extremely effective at forecasting emissions based on factors like engine size and vehicle class.

In practical terms, these metrics imply that the model can reliably predict the amount of CO2 emissions a vehicle will produce based on its specifications. This is particularly useful for:

* **Regulatory Compliance**: Ensuring vehicles meet environmental standards.
* **Environmental Impact Studies**: Helping researchers and policymakers understand the potential impact of various vehicle types on air quality.
* **Automotive Industry**: Assisting manufacturers in designing more environmentally friendly vehicles.

**Model Reliability**:

The reliability of our model is further underscored by the results of cross-validation, a robust statistical technique used to evaluate how the predictions of a model will generalise to an independent data set. The average MSE across multiple folds of data was consistently low, and the average R-squared remained high, which indicates consistent performance across different subsets of the dataset.

This consistency is important because it means the model does not just perform well on a single set of data it was initially tested on but also on new, unseen data. This robustness enhances confidence in the model's utility in real-world applications, assuring stakeholders of its reliability and accuracy in diverse conditions.

Overall, the strong performance metrics combined with the robust validation process confirm that our predictive model is not only accurate but also dependable for practical use in monitoring and planning vehicle emissions. These findings hold promise for contributing significantly to efforts aimed at reducing air pollution and meeting stringent environmental regulations.

1. **Conclusion and Next Steps**

**Summary of Findings**:

This project demonstrates the capability of machine learning models to predict vehicle emissions effectively, assisting in regulatory compliance and environmental management. By identifying key factors that influence emissions, stakeholders can make informed decisions that contribute to the UK's environmental goals.

**Future Work**:

Future work may include exploring additional data sources or features that could further enhance the model's accuracy. Continuous updates and validations of the model as new data becomes available will ensure its relevance and effectiveness in changing regulatory and environmental contexts.

1. **Annex**

**Data Source**: <https://www.kaggle.com/datasets/krupadharamshi/fuelconsumption>