30/09/2024

# Project CO2 emission by vehicles

Final report

Lenovo
DATA SCIENTEST

# Table des matières

1 Introduction....3

1.1 Context....3

1.2 Objectives....3

2 Understanding and Manipulation of Data....4

2.1 Framework....4

2.2 Pre-processing and Feature Engineering....5

2.2.1 *Handling Missing Data*....6

2.2.2 *Data Normalization*....6

2.2.3 *Handling Redundant Features*....6

2.2.4 *Feature Transformation: Innovative Technologies*....6

2.2.5 *Feature Selection*....7

2.2.6 *Outlier Detection and Treatment*....7

2.2.7 *Train-Test Split and Sampling*....7

2.2.8 *Final Dataset*....8

2.3 Visualization and Statistics....8

2.3.1 *Distribution of $CO_2$ Emissions*....8

2.3.2 *Vehicle Mass (kg), Engine Capacity ($cm^3$) and Engine Power (KW) vs. $CO_2$ Emissions*....9

2.3.3 *Emissions by Fuel Type*....13

2.3.4 *Correlation Matrix*....14

2.3.5 *Innovative Technologies*....16

3 Modeling and problem solving....17

3.1 Classification of the problem....17

3.2 Model recommended for regression....17

3.3 Model recommended for classification....18

3.3.1 *Approach with 7 classes*....18

3.3.2 *Approach with 10 classes*....21

4 Challenges faced....29

5 Conclusion and Perspectives....29

6 Bibliography....31

# 1 Introduction

## 1.1 Context

The pressing challenge of reducing $CO_2$ emissions is at the forefront of global efforts to combat climate change. The transportation sector, particularly the automotive industry, is one of the largest contributors to greenhouse gas emissions, especially in Europe. The European Union has implemented stringent emissions standards to curb this trend, creating a significant impact on car manufacturers and prompting the development of more eco-friendly technologies. This project, situated within the intersection of **economic, technical, scientific, and business** considerations, seeks to analyze vehicle data to understand and predict $CO_2$ emissions.

From an **economic perspective**, car manufacturers face increasing pressure to reduce emissions due to both regulatory requirements and shifting consumer preferences towards greener options. Compliance with emissions standards directly affects market competitiveness, with penalties for non-compliance and incentives for low-emission vehicles.

On a **technical level**, vehicle manufacturers are constantly innovating to reduce emissions. The integration of electric vehicles (EVs), hybrid technologies, and fuel-efficient designs has made it increasingly important to understand which vehicle characteristics—such as weight, engine size, and fuel type—most significantly influence emissions. The technical aspect of this project involves using machine learning to identify these key predictors from large datasets.

From a **scientific perspective**, understanding vehicle $CO_2$ emissions is a complex problem that involves multiple fields, including environmental science, mechanical engineering, and data science. Our project employs statistical modeling and machine learning techniques to analyze the vast amounts of data generated by vehicle registrations and emissions tests. This scientific approach will help illuminate the relationships between vehicle characteristics and their environmental impact.

In the **business context**, the findings of this project could assist manufacturers in designing vehicles that are not only compliant with emissions regulations but also optimized for lower emissions, thus offering a competitive edge. Companies are looking to innovate by adopting greener technologies, and the insights from this project can contribute to more informed decision-making processes in product development and regulatory compliance strategies.

In an academic setting, this project fits within a larger discourse on sustainability and environmental engineering. It leverages data science and machine learning to generate actionable insights, offering a multidisciplinary approach to a global challenge.

## 1.2 Objectives

The main objectives of this project are to:

1. **Analyze the relationship between vehicle characteristics and $CO_2$ emissions**: We aim to understand how different factors, such as vehicle mass, engine capacity, and fuel type, contribute to $CO_2$ emissions.

2. **Develop a predictive model for $CO_2$ emissions**: Using machine learning algorithms, we will create a model that can accurately predict $CO_2$ emissions based on vehicle specifications.
3. **Identify the most important factors affecting emissions**: By quantifying the impact of each variable, we aim to inform vehicle design decisions that could lead to lower emissions.
4. **Provide insights to help manufacturers design more sustainable vehicles**: The ultimate goal is to contribute to the ongoing efforts to reduce the automotive industry's carbon footprint, potentially influencing future vehicle designs.

While the team has not engaged with external experts, we have drawn from a variety of academic sources and prior research, such as the works by **Zubair et al.** and **Al-Nefaie et al.**, to shape our approach to both data analysis and model development.

# 2 Understanding and Manipulation of Data

## 2.1 Framework

The dataset used in this project was obtained from the European Environment Agency (EEA), which systematically monitors and publishes data on $CO_2$ emissions from vehicles registered across the European Union (EU) and associated countries. The dataset specifically focuses on the emissions performance of vehicles based on the World Harmonized Light Vehicles Test Procedure (WLTP), a global standard for measuring fuel consumption and emissions. For our analysis, we selected the most recent dataset from 2022, which provides a comprehensive view of the emissions landscape for new vehicles sold within that year.

This dataset is both extensive and highly granular, containing 10.7 million individual vehicle entries and 40 variables detailing a variety of vehicle characteristics. These entries span across the 27 member states of the EU, along with additional data from Norway and Iceland. This coverage makes the dataset particularly valuable for understanding not only the trends in individual countries but also the broader EU context, including variations in vehicle emissions due to regional policies and consumer preferences.

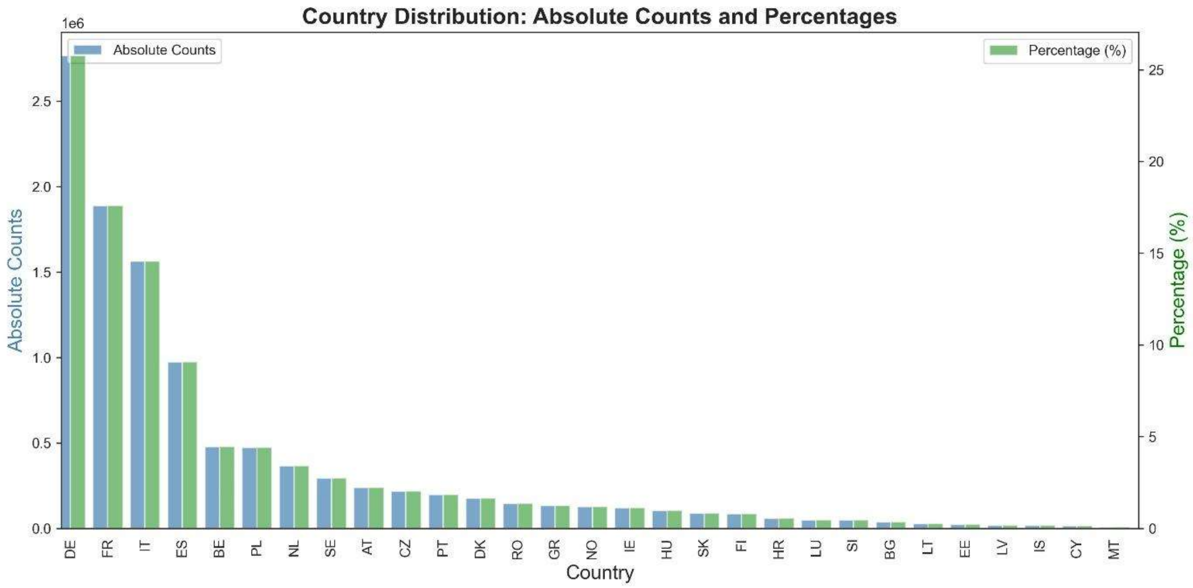


*__Figure 1__: Distribution of the data across all the European countries.*

Key variables in the dataset include:

**Vehicle Mass**: The dataset provides two types of mass measurements: curb weight (the weight of the vehicle without passengers or cargo) and WLTP test mass (the weight used during testing under the WLTP). Both of these measures, expressed in kilograms, are crucial for understanding the relationship between a vehicle's weight and its emissions output.

**Engine Specifications**: Detailed engine-related variables include engine displacement (measured in cubic centimeters), engine power (measured in kilowatts), the number of cylinders, and fuel type (such as petrol, diesel, electric, or hybrid). For hybrid and fully electric vehicles, additional features like battery capacity and electric range (in kilometers) are also recorded. These specifications are pivotal in determining the vehicle's overall performance, efficiency, and environmental impact.

**Innovative Technologies**: The dataset tracks the adoption of eco-innovations, or technologies aimed at reducing emissions and improving fuel efficiency. Examples include LED lighting systems, start-stop technology, and other emissions-reducing mechanisms approved under the WLTP standards. These technologies are often denoted by specific codes, which we preprocessed to simplify their analysis.

**$CO_2$ Emissions**: One of the most critical variables is the vehicle's specific $CO_2$ emissions, measured in grams per kilometer (g/km). This value is calculated during standardized testing using the WLTP, which ensures a consistent and reliable measure of how much $CO_2$ each vehicle emits per kilometer driven. This is the primary target variable for our analysis, as it reflects the environmental impact of each vehicle.

Given the size of the dataset, which totals approximately 2.4 GB, significant preprocessing was required to prepare it for machine learning and statistical analysis. The dataset initially contained several variables with a high percentage of missing values, and many fields required normalization to ensure comparability across different vehicles. Additionally, the large volume of data meant that computational efficiency was a critical consideration, necessitating the use of optimized tools and techniques for both data cleaning and analysis.

## 2.2 Pre-processing and Feature Engineering

Pre-processing and feature engineering were essential steps in transforming the raw dataset into a format suitable for machine learning and statistical analysis. Given the size and complexity of the dataset—10.7 million rows and 40 columns—our goal was to clean the data, handle missing values, and optimize it for predictive modeling without losing critical information. This section outlines the steps taken to prepare the data.

### 2.2.1 Handling Missing Data

The dataset included several columns with missing or incomplete data, which required careful treatment. We adopted a tiered approach to managing these missing values:

1. **Dropping Columns with High Missing Data**: Columns where more than 50% of the entries were missing were excluded from the analysis. These columns were deemed too incomplete to provide meaningful insights or contribute effectively to the model's performance.

2. **Selective Removal of Missing Rows**: For columns with a smaller percentage of missing data, rows containing null values were dropped, provided that doing so would not significantly reduce the dataset's size or diversity.

3. **Imputation**: For specific columns where missing data carried inherent meaning, such as "Emissions Reduction through Innovative Technology (g/km, WLTP)," missing values were replaced with zeros. In this case, a missing entry likely indicated that the vehicle did not possess any emissions-reducing technology, and thus a value of zero would more accurately reflect this condition.

### 2.2.2 Data Normalization

Given the wide range of values across different features (e.g., fuel consumption ranging from 0.1 to 29 liters per 100 km, and engine capacity between 658 cm$^3$ and 7997 cm$^3$), normalization was essential to ensure that variables with large numeric ranges did not dominate the analysis. We applied Min-Max Normalization to scale all numeric features to a range between 0 and 1. This method preserves the relationships between values while making them more suitable for machine learning algorithms that assume similar scales for inputs.

### 2.2.3 Handling Redundant Features

Several features in the dataset displayed high collinearity, which could introduce redundancy and multicollinearity into our predictive models. To address this, we reviewed the correlations between numerical variables:

**Vehicle Mass**: Both "Curb Mass" and "WLTP Test Mass" were provided in the dataset. A correlation analysis revealed that these variables had a 99% correlation, meaning they captured essentially the same information. As a result, we retained only the "WLTP Test Mass" variable to eliminate redundancy while maintaining the integrity of the data.

**Fuel Consumption and $CO_2$ Emissions**: There was a 95% correlation between fuel consumption and $CO_2$ emissions, which is expected given that higher fuel consumption generally leads to higher emissions. While this relationship is crucial for understanding emissions, including both variables in predictive modeling could lead to overfitting. Therefore, fuel consumption was excluded from the final model to avoid multicollinearity.

### 2.2.4 Feature Transformation: Innovative Technologies

The column "Innovative Technologies" represented various emissions-reducing features approved under the WLTP standards, such as LED lights and energy-saving technologies. However, these technologies were encoded using space-separated numbers (e.g., "e13 33 37"), making them challenging to interpret directly.

To simplify this feature for analysis, we made the following transformations:

**"Has_LED" Boolean Variable**: LED lighting (code 37) is a common innovative technology associated with $CO_2$ reduction. We created a binary feature called has_LED, where a value of 1 indicates that the vehicle is equipped with LED lighting, and 0 means it is not.

**"Additional_IT" Categorical Variable**: In addition to LEDs, other eco-innovations (codes 32, 33, 37, etc.) were considered. We created a categorical feature called additional_IT, which could take one of four values: [32, 33, 37, NONE]. This transformation simplified the multi-value encoding while preserving the information about which technologies were present.

### *2.2.5 Feature Selection*

To ensure that the predictive model remained interpretable and effective, we carefully selected features based on domain knowledge from previous studies on $CO_2$ emissions. The following features were retained as key drivers of emissions:

- **Vehicle Mass (WLTP Test Mass)**
- **Engine Capacity ($cm^3$)**
- **Engine Power (KW)**
- **Fuel Type**
- **Transmission Type**
- **Innovative Technologies (Has_LED, Additional_IT)**
- **$CO_2$ Emissions (g/km, WLTP) (Target Variable)**

These features were chosen because they had a clear relationship with vehicle emissions based on the literature and our own exploratory analysis.

### *2.2.6 Outlier Detection and Treatment*

Given the large number of vehicles in the dataset, outliers could skew the results. Outliers were identified using box plots and z-scores for key variables like engine capacity, vehicle mass, and $CO_2$ emissions. We observed that most outliers corresponded to either high-performance or low-efficiency vehicles (e.g., luxury SUVs, supercars), which were retained in the dataset as they represent legitimate use cases in the real world. However, extreme outliers with implausible values (e.g., negative emissions or zero engine capacity) were removed.

### *2.2.7 Train-Test Split and Sampling*

Given the large size of the dataset (over 10 million rows), we opted for stratified sampling to ensure that our training and test sets reflected the distribution of key variables, particularly fuel type and country of registration. A typical 80/20 split was used, with 80% of the data allocated for model training and 20% for testing and validation. Stratification ensured that rare categories, such as electric vehicles and hybrid models, were adequately represented in both sets.

### *2.2.8 Final Dataset*

After pre-processing, the dataset was significantly reduced in size, both in terms of the number of rows (after handling missing data) and the number of features (after removing redundant or highly correlated variables). This clean, structured dataset served as the foundation for our exploratory data analysis and machine learning model development.

## 2.3 Visualization and Statistics

To better understand the dataset and identify key patterns before applying models, we conducted exploratory data analysis (EDA). This involved visualizing the distributions of key features, examining relationships between variables, and calculating summary statistics. These visualizations and statistics provided insights into the structure of the data and helped guide feature engineering and model selection.

We did some basic statistical analysis on the target variable to decide if we would work on all data or only on the data of particular countries. Based on the analysis we can drop Italy for the rest of the analysis because it’s not adding insight:

- The Mean of Italy is close to that of the overall data while the mean of Germany is higher and that of France is lower
- In the case of Italy, the data is less dispersed, so it does not bring any additional specificity to the analysis that is worth exploring.

*Table 1: Box plot parameters*

| **Type of data** | **Size of the data set** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **max** |
|---|---|---|---|---|---|---|---|---|
| **Overall** | 7063310.0 | 126.761808 | 39.731316 | 3.0 | 116.0 | 128.0 | 142.0 | 461.0 |
| **Germany** | 2140118.0 | 138.999046 | 45.440096 | 10.0 | 124.0 | 136.0 | 155.0 | 456.0 |
| **France** | 1543750.0 | 116.502216 | 32.389164 | 10.0 | 111.0 | 122.0 | 133.0 | 456.0 |
| **Italy** | 1360690.0 | 125.401174 | 25.366009 | 10.0 | 112.0 | 124.0 | 135.0 | 456.0 |

### *2.3.1 Distribution of $CO_2$ Emissions*

A key focus of this project is to analyze $CO_2$ emissions. This analysis explores $CO_2$ emissions (g/km) for vehicles in Germany, France, and Italy. Using box plots, we can visualize the distribution of emissions in each country and compare them to the overall distribution.

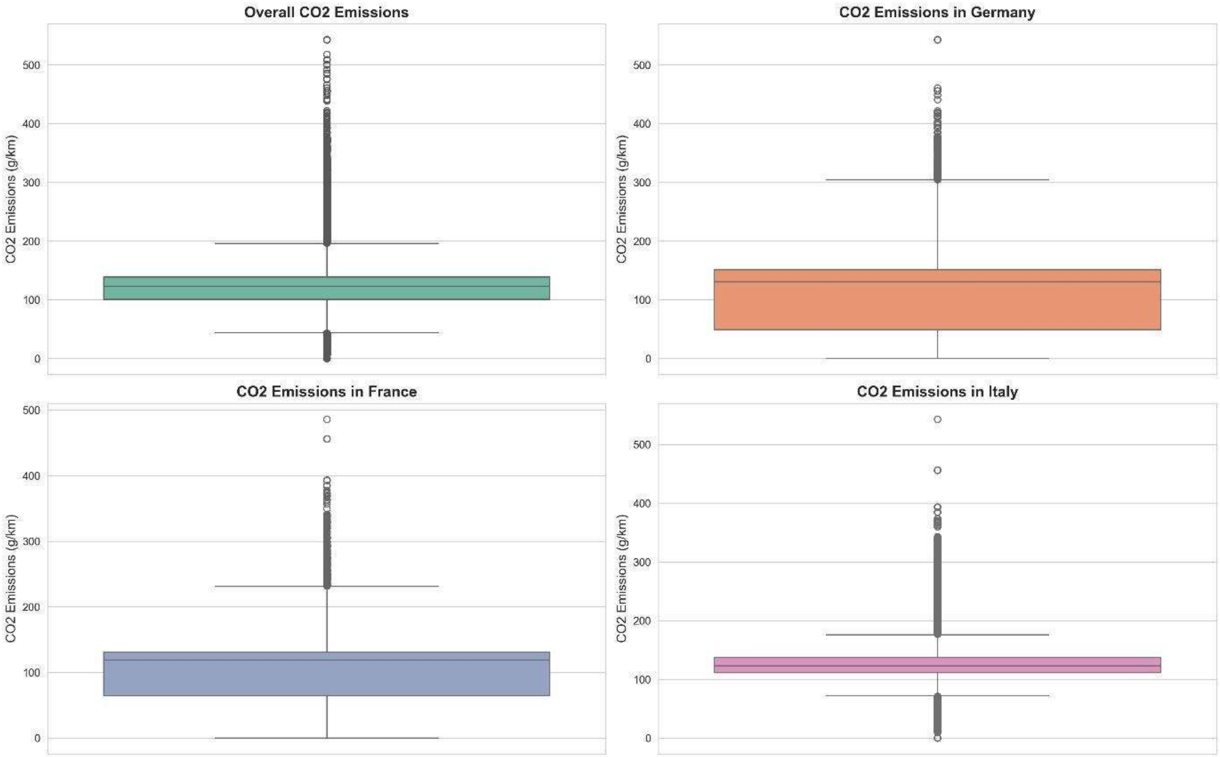


***Figure 2**: CO2 emissions distributions: All, Germany, France, Italy*

**Variability:** The box plots reveal significant variability in $CO_2$ emissions across all three countries, suggesting that factors such as vehicle type, driving habits, and fuel efficiency standards play a crucial role in determining emissions.

**Outliers:** The presence of outliers, especially high-emission vehicles, indicates a need to address tail-end emissions to reduce the overall environmental impact.

**Country Differences:** While there are some variations between the countries, the overall distributions are relatively similar, suggesting that common factors such as vehicle technology and fuel availability influence emissions across the region.

This analysis provides insights into $CO_2$ emissions for vehicles in Germany, France, and Italy. While there are some differences between the countries, the overall picture highlights the need for continued efforts to improve vehicle efficiency and reduce emissions to mitigate climate change. Further research could delve into specific factors contributing to the observed variations and explore potential policy interventions to promote cleaner transportation.

### *2.3.2 Vehicle Mass (kg), Engine Capacity ($cm^3$) and Engine Power (KW) vs. $CO_2$ Emissions*

To explore the relationship between Vehicle Mass (kg), Engine Capacity ($cm^3$) and Engine Power (KW) with $CO_2$ emissions, we generated a scatter plot showing how these two variables interact for different fuel types. Due the size of the dataset,  we have decided to use the first 2 countries with the largest dataset: **Germany** and **France**.

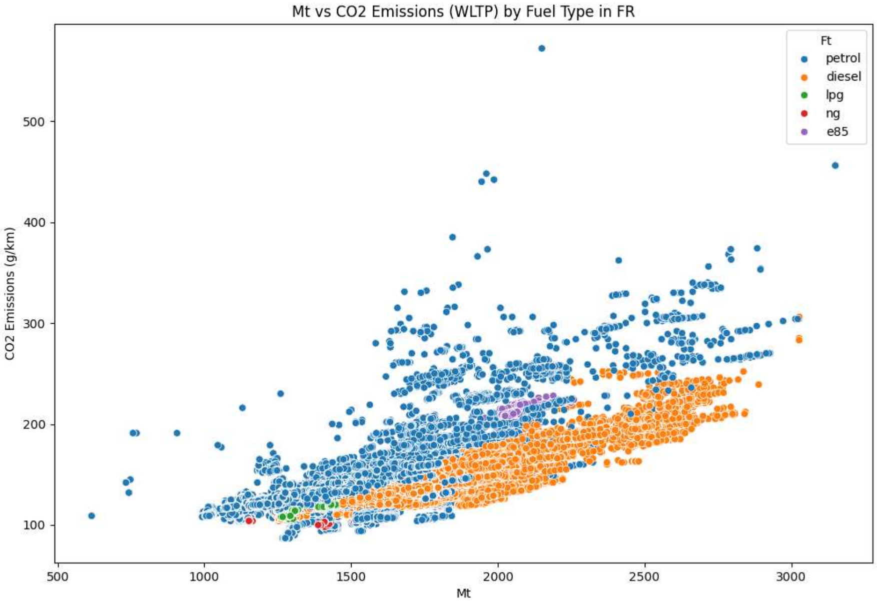


***Figure 3**: Mt (WLTP test mass) vs CO2 Emissions (WLTP) by Fuel Type in France (2022).*

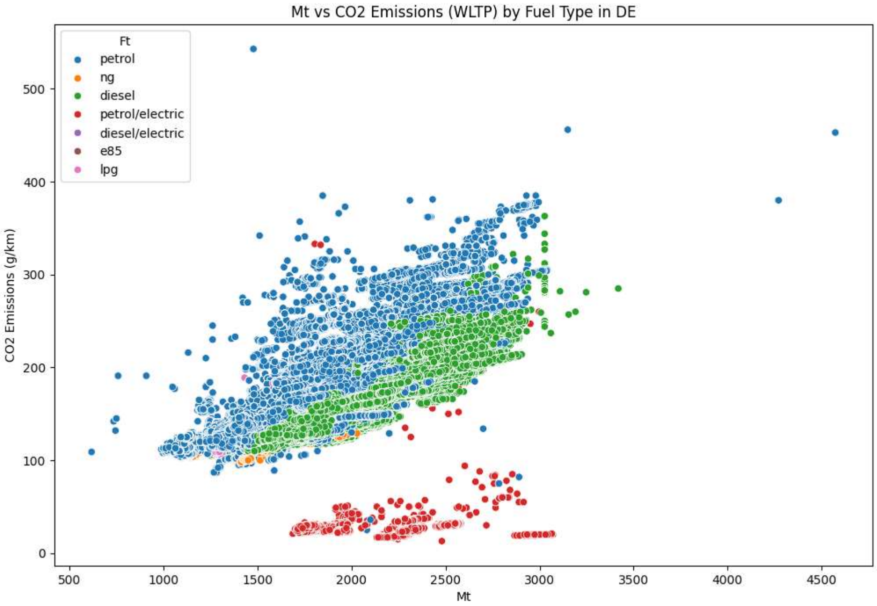


***Figure 4**: Mt (WLTP test mass) vs CO2 Emissions (WLTP) by Fuel Type in Germany (2022).*

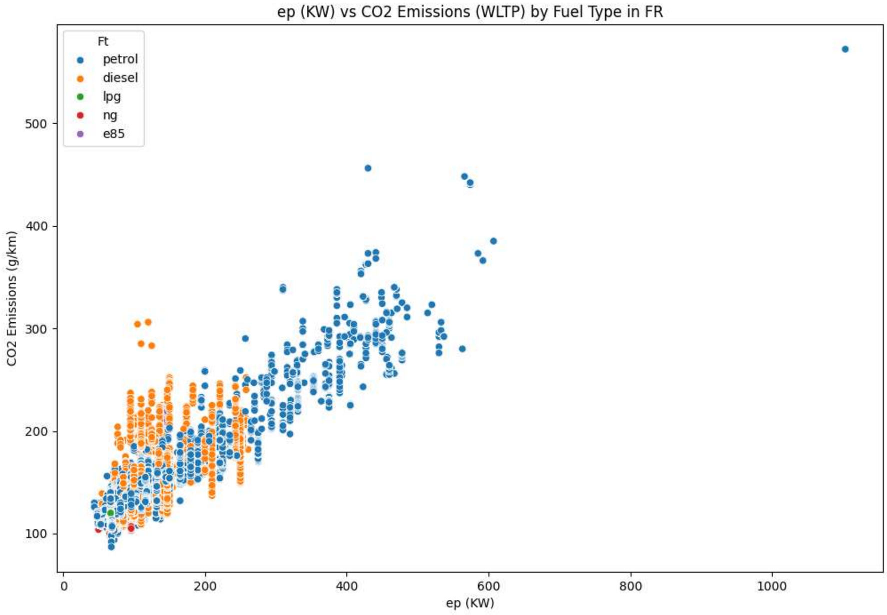


***Figure 5**: ep (KW) vs CO2 Emissions (WLTP) by Fuel Type in France (2022).*

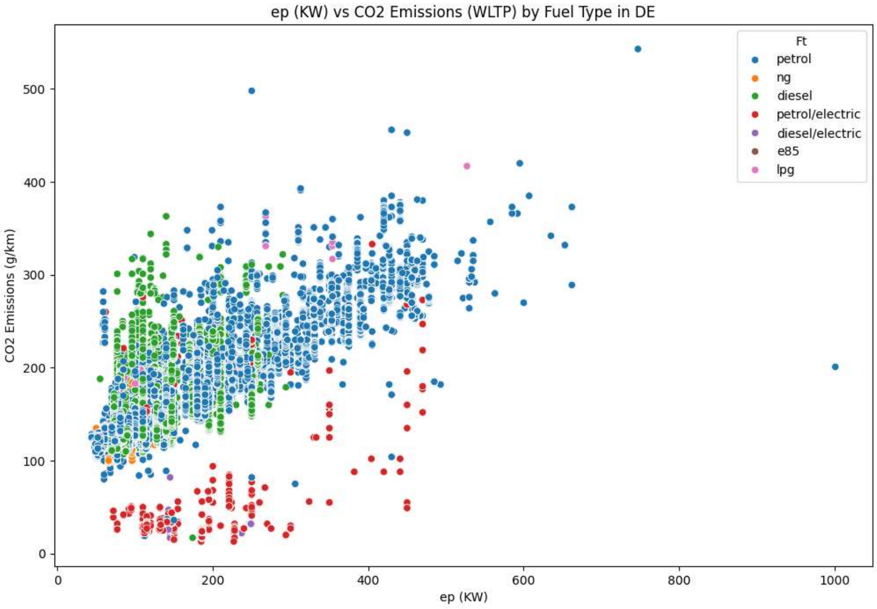


***Figure 6**: ep (KW) vs CO2 Emissions (WLTP) by Fuel Type in Germany (2022).*

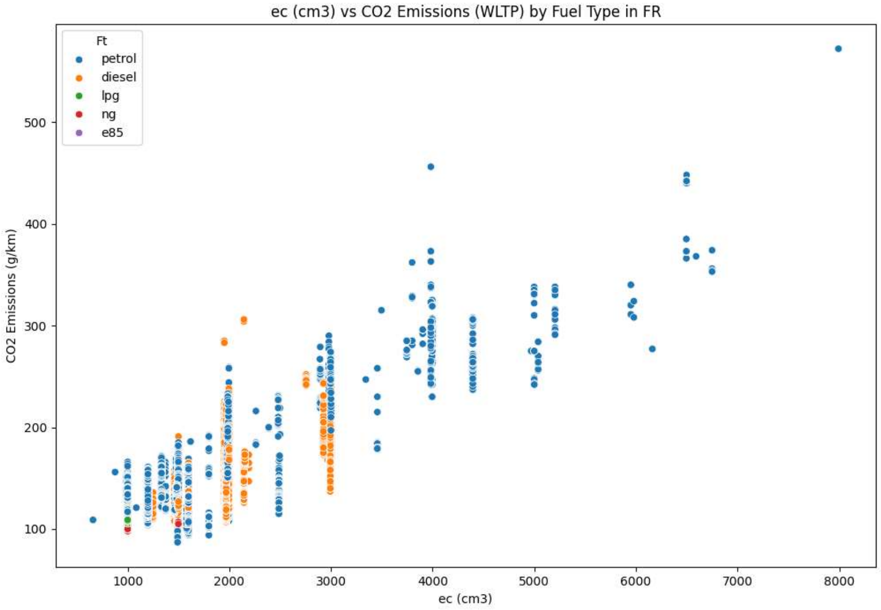


***Figure 7**: ec (cm3) vs CO2 Emissions (WLTP) by Fuel Type in France (2022).*

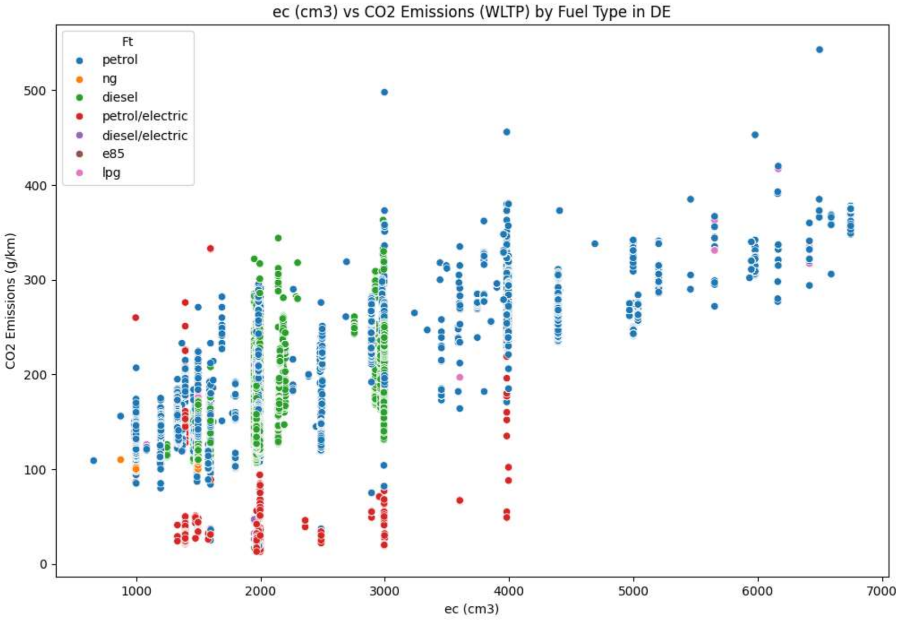


***Figure 8**: ec (cm3) vs CO2 Emissions (WLTP) by Fuel Type in Germany (2022).*

**Consistent Positive Correlation:** Larger, heavier vehicles with more powerful engines tend to emit more CO2 regardless of fuel type.

- **Fuel Type Influence:** Hybrid vehicles consistently have lower CO2 emissions than traditional petrol and diesel options. LPG and e85 performance varies between countries.
- **Specific Relationships:** Engine capacity, vehicle mass, and engine power are all positively correlated with CO2 emissions, but the impact can vary slightly between fuel types.

- **Clustering and Outliers:** Data points cluster in certain regions within each fuel type, and outliers exist, especially for petrol and diesel vehicles.
- **Potential Insights:** Vehicle design and efficiency, fuel type impact, technological advancements, and country-specific factors influence CO2 emissions.

Overall, the analysis highlights the need for continued efforts to promote vehicle efficiency, reduce emissions, and consider the environmental impact of different fuel options.

### 2.3.3 *Emissions by Fuel Type*

We also analyzed the differences in CO₂ emissions across various fuel types (e.g., petrol, diesel, hybrid) by creating a box plot.

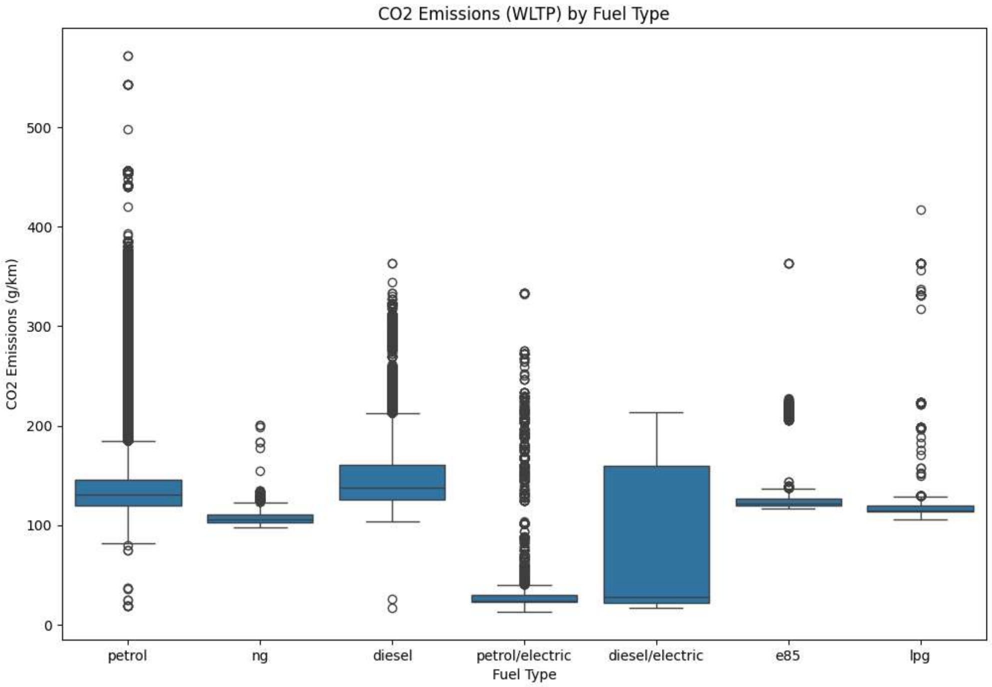


***Figure 9**: CO2 Emissions (WLTP) by Fuel Type in Germany and France (2022).*

**Hybrid Vehicles (petrol/electric, diesel/electric):** Hybrid vehicles consistently exhibit significantly lower CO₂ emissions compared to traditional petrol and diesel options. This suggests that hybrid technology is more efficient in reducing emissions.

**Petrol and Diesel Vehicles:** Both petrol and diesel vehicles show a wide range of emissions. While there might be some overlap, the overall trend indicates that diesel vehicles tend to have slightly lower emissions on average compared to petrol vehicles.

**Alternative Fuels (e85, LPG):** The performance of e85 and LPG vehicles varies. Some e85 vehicles have emissions comparable to diesel, while others exhibit higher levels. LPG vehicles generally have emissions similar to petrol vehicles.

Interpretation

**Hybrid Dominance:** Hybrid vehicles clearly outperform traditional fuel types in terms of $CO_2$ emissions, emphasizing the effectiveness of hybrid technology in reducing environmental impact.

**Fuel Type Variations:** Within petrol and diesel categories, there are variations in emissions, likely influenced by factors such as engine size, driving habits, and vehicle age.

**Alternative Fuel Potential:** While e85 and LPG show some promise, their performance is less consistent compared to hybrid vehicles. Further research and technological advancements are needed to optimize their environmental benefits.

Conclusion

This analysis demonstrates the significant potential of hybrid technology in reducing $CO_2$ emissions from vehicles. While petrol and diesel vehicles continue to be widely used, there is a clear trend towards cleaner alternatives. As technology advances, it is expected that alternative fuel options will become more competitive and contribute to a more sustainable transportation sector.

### 2.3.4 *Correlation Matrix*

To identify relationships between multiple features, we calculated a correlation matrix for the numerical variables in the dataset. This was visualized using a heatmap to highlight both positive and negative correlations.

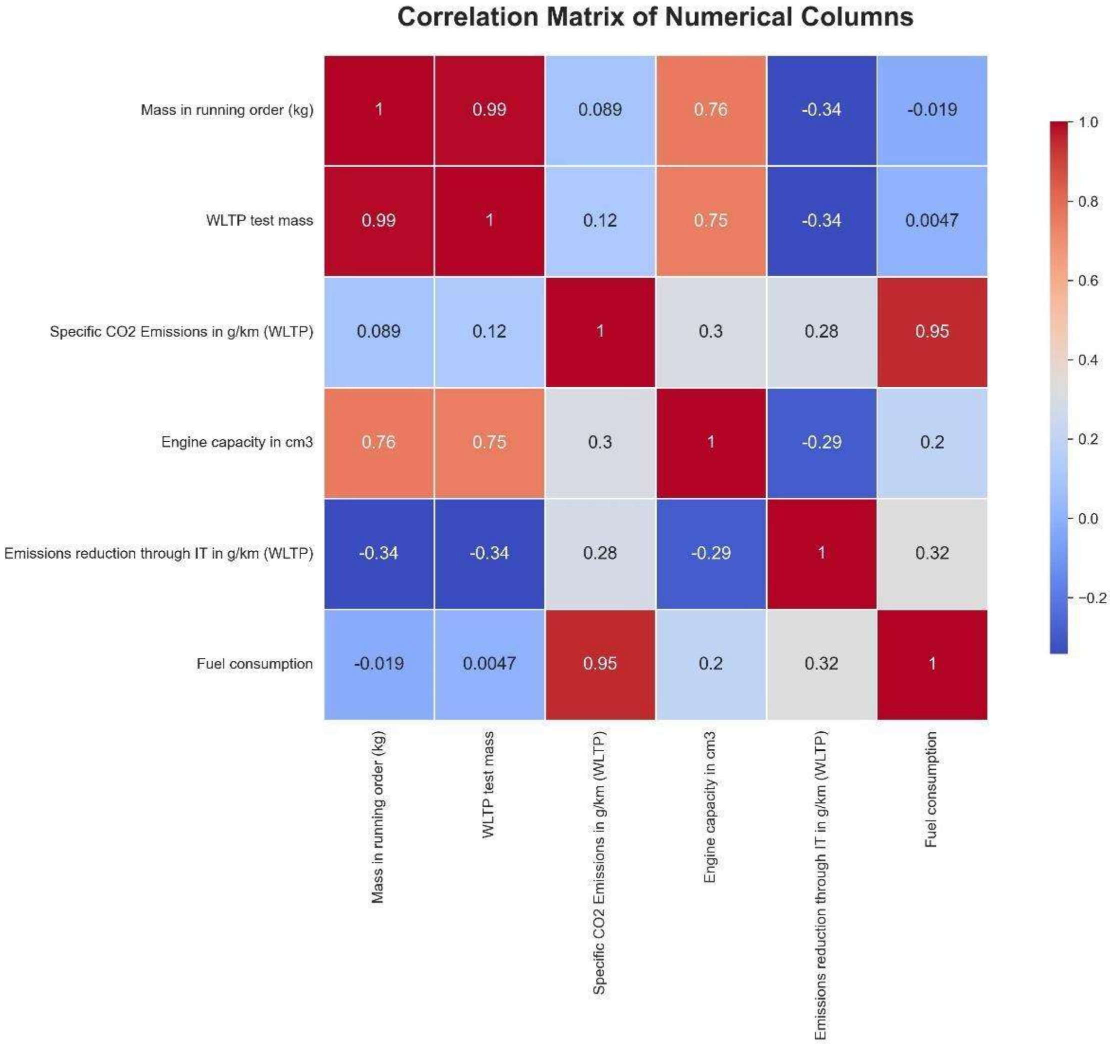


***Figure 10***: *Correlation between the numerical variables*

**Vehicle Weight and Emissions:** The strong positive correlations between vehicle weight (measured in different ways) and CO₂ emissions highlight the importance of reducing vehicle mass to improve fuel efficiency and reduce environmental impact.

**Engine Size and Emissions:** Larger engines are generally associated with higher CO₂ emissions, emphasizing the need for more efficient engine designs and technologies.

**Technology and Emissions Reduction:** The negative correlation between emissions reduction through IT and CO₂ emissions indicates that technological advancements can play a crucial role in mitigating climate change.

This correlation matrix provides valuable insights into the relationships between various factors influencing CO₂ emissions. By understanding these relationships, we can identify potential areas for improvement in vehicle design, technology, and policies to promote more sustainable transportation.

### 2.3.5 *Innovative Technologies*

The following Figures show the combination of the innovative technologies and the effect on the total emissions, per Fuel type.

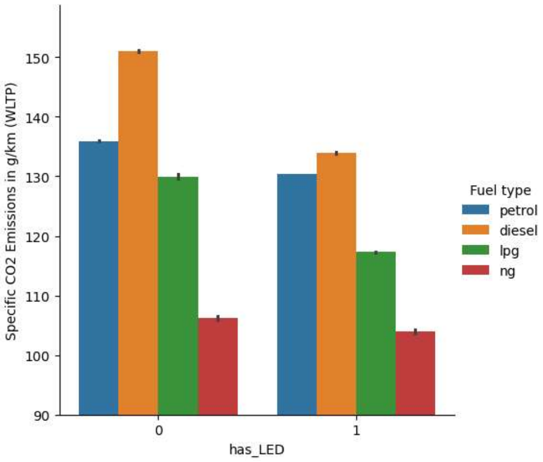


***Figure 11**: Influence on CO2 Emissions by Fuel type divided by has_LED.*

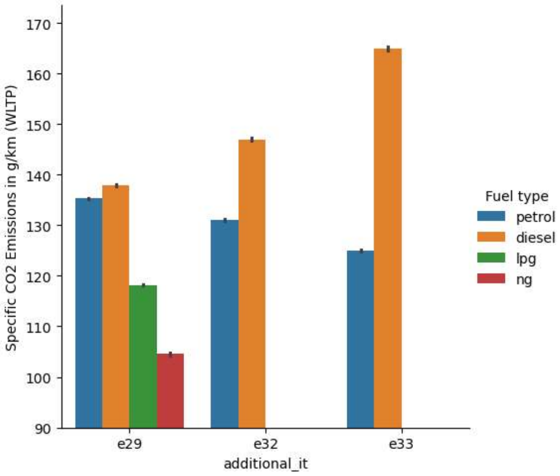


***Figure 12**: Influence on CO2 Emissions by Fuel type divided by technologies.*

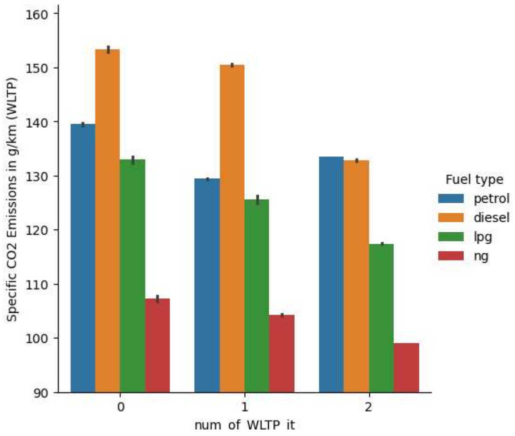


***Figure 13**: Influence on CO2 Emissions by Fuel type divided by WLTP it.*

This analysis suggests that while LED technology can contribute to energy efficiency in various applications, its impact on $CO_2$ emissions from vehicles is limited. Specific technologies can influence $CO_2$ emissions, their impact varies depending on the fuel type. And while WLTP it might have some influence on $CO_2$ emissions for certain fuel types, the overall impact is limited. Hybrid vehicles continue to be the most efficient option in terms of reducing emissions. Further research and data collection are needed to better understand the potential benefits and drawbacks of different technologies in improving fuel efficiency and reducing $CO_2$ emissions.

# 3 Modeling and problem solving

## 3.1 Classification of the problem

This problem can be solved using two different types of machine learning algorithms:

- **Regression**: The absolute value of the CO2 emission can be predicted using the technical features of the vehicle
- **Classification**: The CO2 emission category can be predicted. A first approach consists of using, the car labels (Haq and Weiss 2016.) used in some European countries (e.g., Belgium, Germany, Spain, UK) can be used for our modeling. These labels are presented hereafter:
    - A: CO2 emission < 100 g/km
    - B: CO2 emission between 100 g/km and 130 g/km
    - C: CO2 emission between 130 g/km and 160 g/km
    - D: CO2 emission between 160 g/km and 190 g/km
    - E: CO2 emission between 190 g/km and 220 g/km
    - F: CO2 emission between 220 g/km and 250 g/km
    - G: CO2 emission greater 250 g/km

Another approach for classification consisting of 10 classes instead of 7 was also explored. This approach was completed to make the overall problem harder since the first approach gave excellent results even without the fuel consumption.

## 3.2 Model recommended for regression

We tested 3 different models for regression: linear regression, lasso regression, and the ridge regression. The ridge regression and the lasso regression after optimization were giving the same results as the linear regression in the case of the model without fuel consumption: an R-square around 0.87 and an RMSE of 14.38. We recommend the linear regression model for three reasons: (i) linear regression is giving the same performance results as Ridge and Lasso (ii) it is simpler than Ridge and Lasso regression that require hyperparameter tuning (iii) It easy to interpret. The regression coefficients (Fig. 1) allow us to easily interpret the importance of each feature on the final results. The

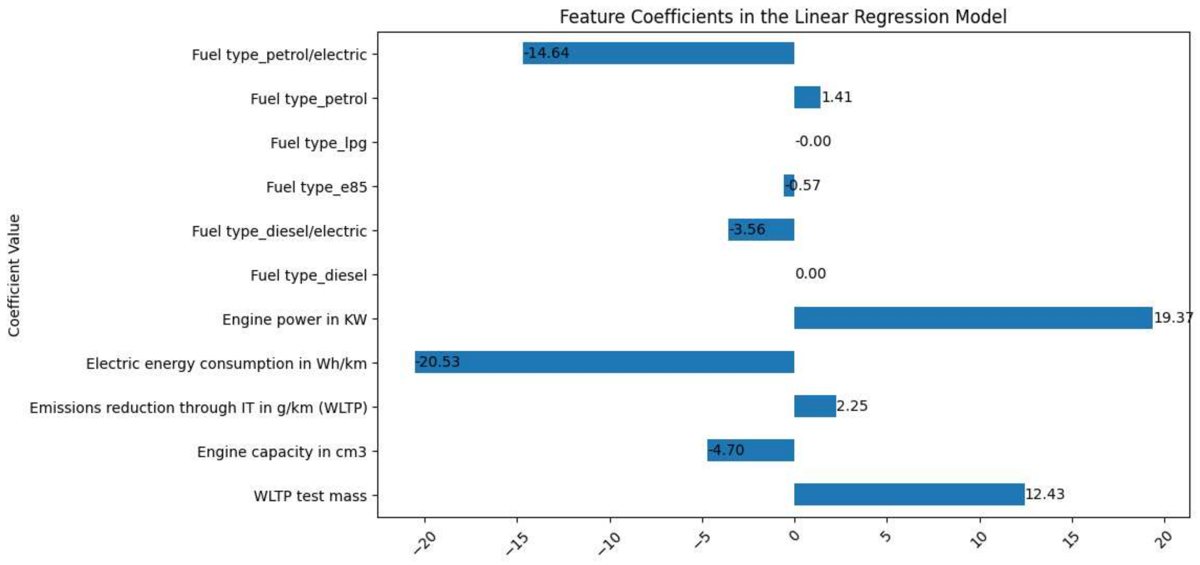


***Figure 14**: Model without Fuel consumption – Features coefficients in the Linear Regression model*

## 3.3 Model recommended for classification

### *3.3.1 Approach with 7 classes*

For the classification with 7 classes, Random Forest is the recommended model for our problem. Random Forest gives the best result in terms of performance with an average score of 0.95 with all classes having an F1 score greater than 0.9 which is a very good results. Moreover, Random Forest prediction time is relatively ~0.1 seconds which allows its deployment easily in production. The interpretability analysis give us good confidence in the choice of Random Forest because features that features highlighted make physical sense.

*Table 2: Random Forest Confusion Matrix for Dataset from 2022.*

| | **Predicted values** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Model with fuel consumption** | | | | | | | | **Model without fuel consumption** | | | | | |
| **Real values** | **Class** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | | **0** | **1** | **2** | **3** | **4** | **5** | **6** |
| | **0** | 5617 | 1 | 1 | 1 | 0 | 0 | 3 | | 5615 | 7 | 0 | 1 | 0 | 0 | 0 |
| | **1** | 2 | 5577 | 41 | 2 | 0 | 0 | 1 | | 5 | 5345 | 261 | 9 | 1 | 2 | 0 |
| | **2** | 0 | 41 | 5538 | 43 | 1 | 0 | 0 | | 1 | 195 | 5241 | 186 | 0 | 0 | 0 |
| | **3** | 13 | 5 | 22 | 5535 | 47 | 1 | 0 | | 3 | 4 | 161 | 5253 | 191 | 11 | 0 |
| | **4** | 1 | 0 | 1 | 31 | 5535 | 55 | 0 | | 1 | 1 | 1 | 183 | 5125 | 305 | 7 |
| | **5** | 2 | 1 | 0 | 1 | 60 | 5542 | 17 | | 3 | 0 | 0 | 16 | 252 | 5209 | 143 |
| | **6** | 0 | 0 | 1 | 4 | 2 | 15 | 5601 | | 0 | 0 | 2 | 4 | 6 | 147 | 5464 |

*Table 3: Random Forest Classification Report for Dataset from 2022.*

| | **Model with fuel consumption** | | | | **Model without fuel consumption** | | |
|---|---|---|---|---|---|---|---|
| Class | Precision | Recall | F1- Score | | Precision | Recall | F1- Score |
| 0 | 1 | 1 | 1 | | 1 | 1 | 1 |
| 1 | 0.99 | 0.99 | 0.99 | | 0.96 | 0.95 | 0.96 |
| 2 | 0.99 | 0.98 | 0.99 | | 0.92 | 0.93 | 0.93 |
| 3 | 0.99 | 0.98 | 0.98 | | 0.93 | 0.93 | 0.93 |
| 4 | 0.98 | 0.98 | 0.98 | | 0.92 | 0.91 | 0.92 |
| 5 | 0.99 | 0.99 | 0.99 | | 0.92 | 0.93 | 0.92 |
| 6 | 1 | 1 | 1 | | 0.97 | 0.97 | 0.97 |
| | | | | | | | |
| Accuracy | | | 0.99 | | | | 0.95 |
| Macro average | 0.99 | 0.99 | 0.99 | | 0.95 | 0.95 | 0.95 |
| Weighted average | 0.99 | 0.99 | 0.99 | | 0.95 | 0.95 | 0.95 |

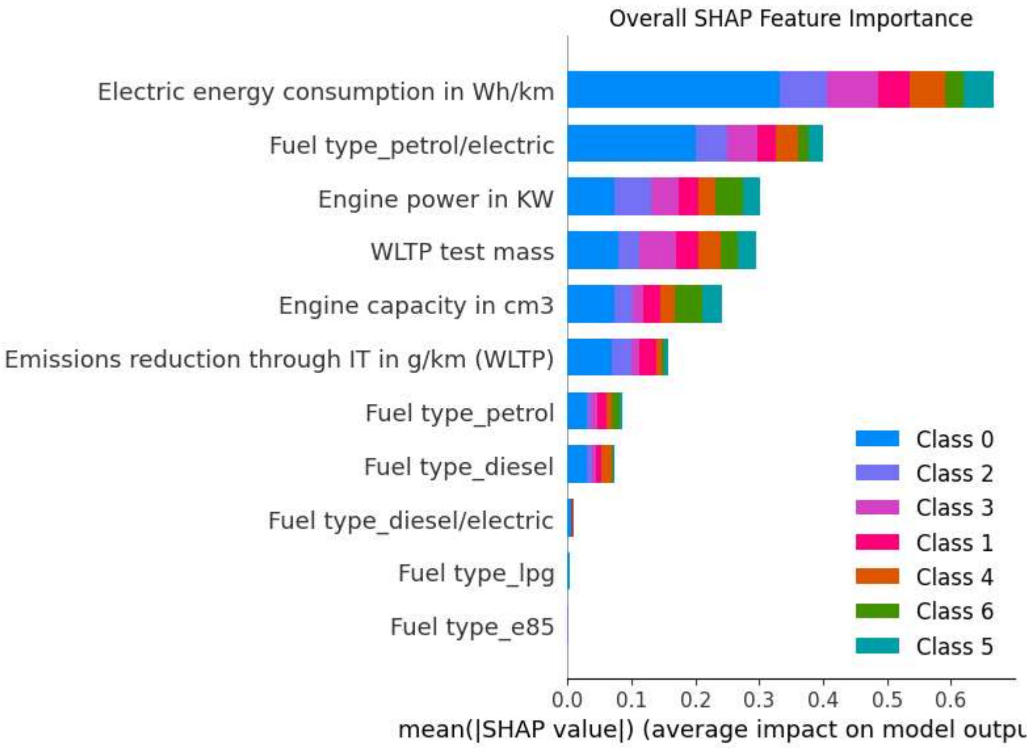


***Figure 15**: Model without Fuel consumption – SHAP analysis on the Random Forest model (best model)*

For example, the interpretability of the Random Forest help us identify 4 parameters that are the most critical to CO2 emission:

- Factors that increase CO2 emission:
  - Engine power in KW
  - Mass of the vehicle
- Factors that decrease CO2 emission
  - Petrol/electric as fuel type
  - Electric energy consumption in KW

To meet the CO2 emission target, car manufacturers can follow two different strategies

**Strategy 1:** Car manufacturers can design light vehicles thus reducing the weight and the engine power.

**Strategy 2**: They can do R&D to have better hybrid vehicles, vehicles with high electric engine power.

Not only we explain why we selected Random Forest, but we also did the extra mile to explain why the other models tested. The other models tested can be separated in two groups:

Models that are close Random forest could be selected

- Bagging: It gave the accuracy as Random Forest (an average F1 score of 0.95), but is computationally intensive: the prediction time of bagging is 1434 ms, 15 times the prediction of the Random Forest. Moreover, Bagging can sometimes require significant memory and processing power
- XgBoost: It had an average F1-score of 0.93, 0.02 point lower than Random Forest. Also, hyperparameter tuning can be complex
- KNN: It gave an average F1-score of 0.93, 0.02 point lower than Random Forest. However, the prediction was 18 times higher than Random Forest

Models that could simply not be selected

- Logistic regression: Low average F1-score: 0.68
- SVM: Low average F1-score: 0.54
- Decision Tree: Low average F1-score: 0.77
- Decision Tree Boosting: Low average F1-score: 0.83
- Deep Learning: Low average F1-score: 0.88

In general, an average score of greater than 0.9 is preferred.

*Table 4: Comparative Performance of Machine Learning Models for Predicting CO2 with and without Fuel Consumption Data (2022 Dataset).*

| | **Model with Fuel Consumption** | **Model without Fuel Consumption** | **Model with Fuel Consumption** | **Model without Fuel Consumption** |
|---|---|---|---|---|
| | Performance metrics | Prediction speed (ms) | Performance metrics | Prediction speed (ms) |
| Logistic regression | Average F1-score: 0.88 | 4 | Average F1-score: 0.68 | 6 |
| Logistic regression with Grid Search | Average F1-score: 0.91 | 8 | Average F1-score: 0.69 | 5 |
| SVM | Average F1-score: 0.64 | 4 | Average F1-score: 0.54 | 4 |
| SVM with Grid Search optimization | Average F1-score: 0.64 | 4 | Average F1-score: 0.54 | 5 |
| KNN | Average F1-score: 0.98 | 16 705 | Average F1-score: 0.93 | 17 300 |
| Decision Tree | Average F1-score: 0.96 | 4 | Average F1-score: 0.77 | 4 |
| Decision Tree Boosting | Average F1-score: 0.96 | 647 | Average F1-score: 0.83 | 651 |
| Bagging | Average F1-score: 0.99 | 1079 | Average F1-score: 0.95 | 1434 |
| Random Forest | Average F1-score: 0.99 | 89 | Average F1-score: 0.95 | 99 |
| XgBoost | Average F1-score: | 70 | Average F1-score: | 77 |

| | | | | |
|---|---|---|---|---|
| | 0.98 | | 0.91 | |
| Deep Learning | Average F1-score: 0.97 | 337 730 | Average F1-score: | 331 154 |

This table presents a comparison of various classification models applied to a dataset. The models are evaluated based on their average F1-score and prediction speed. The table includes models with and without fuel consumption as a predictor variable, as well as models optimized using grid search.

### *3.3.2 Approach with 10 classes*

The classification problem using EU labels gave too good results. Therefore, we have decided to execute another classification that would be slightly harder, which we enforced through a different discretization method as well as a slightly different preprocessing step. As we've done for the first classification, we will show the results for two scenarios: with and without considering the Fuel Consumption explanatory variable.

**Pre-processing**

- First of all, we've decided to remove all hybrid-electric vehicles (diesel/electric & petrol/electric), since these are quite easy to predict, as they don't have a reflection of their entire emission in the dataset.
- Removal of outliers on the target variable (the emissions).
- While adding some slightly-correlated preprocessing variables didn't have much impact on the results when the Fuel consumption variable was considered, adding them helped improve the results when it was disregarded.
- Including 'dummies' for Innovative technologies extraction, fuel types, countries, pools, etc.

**Discretization**

Instead of using the A-G standards, we've decided to make the problem harder by discretizing the target variable into more classes. For that purpose, we wrote a function called "discretize', which takes an integer **n** and splits the data into n equally populated classes. This "equally populated" choice also helped us solve the issue of the target variable being dispersed mainly around 2 classes (as shown in a graph on section 3 above), which absolved the need for over/under-sampling. We chose **n=10** as the number of classes we'd like to discretize our target variable into, but it'd be extremely easy to explore other discretizations.

**Training Method**

In order to find the ideal hyperparameters, all of the algorithms were first trained on a smaller dataset (under 100k records). Then, in order to be able to perform a fair comparison, all of the algorithms were trained on a 2 million records dataset.

We've tried multiple classifiers and combined them with hyperparameters, to eventually come up with the following winners:

### Random Forest Classifier:

*Table 5: Random Forest Classifier*

| | With Fuel Consumption | Without Fuel Consumption |
|---|---|---|
| Training Accuracy | 0.9881 | 0.9423 |
| Accuracy | 0.9830 | 0.9421 |
| F1-score | 0.9829 | 0.9418 |
| Precision | 0.98 | 0.9419 |
| Recall | 0.98 | 0.9417 |

**HyperParameters:**

- With Fuel Consumption:
  *criterion*='entropy', *max_depth*=None, *max_features*='log2', *n_estimators*=350
- Without Fuel Consumption:
  *criterion*='entropy', *max_depth*=None, *max_features*='sqrt', n_estimators=200

**Advantages: High accuracy, robust to noise, provides feature importance.**

**Constraints:** Computationally expensive, less interpretable than individual trees.

**Key Takeaways:**

- Fuel consumption is crucial for CO2 prediction.
- Random Forest performs well, but requires careful hyperparameter tuning.
- Consider computational costs for large datasets

**DTC with Adaptive Boosting:**

**DTC alone:**

*Table 5: DTC*

| | With Fuel Consumption | Without Fuel Consumption |
|---|---|---|
| Training Accuracy | 0.9788 | 0.9310 |
| Accuracy | 0.9772 | 0.9245 |
| F1-score | 0.9772 | 0.9241 |
| Precision | 0.9772 | 0.9242 |
| Recall | 0.9771 | 0.9240 |

**HyperParameters:**

- With Fuel Consumption:
  *criterion*='entropy', *max_depth*=20, *min_samples_leaf*=15, *min_samples_split*=15
- Without Fuel Consumption:
  *criterion*='gini', *max_depth*=25, *min_samples_leaf*=15, *min_samples_split*=15

**Advantages:** Interpretable, efficient, handles non-linear relationships.

**Constraints:** Prone to overfitting, sensitive to features, limited flexibility.

**Key Takeaways:**

- Fuel consumption is a key predictor.
- Hyperparameter tuning is crucial.
- Interpretable, but watch for overfitting.

DTC is a good choice for CO2 prediction, but requires careful tuning.

**DTC with Adaptive Boosting:**

*Table 6: DTC with Adptive Boosting*

| | With Fuel Consumption | Without Fuel Consumption |
|---|---|---|

| | | |
|---|---|---|
| Training Accuracy | 0.9861 | 0.9432 |
| Accuracy | 0.9810 | 0.9310 |
| F1-score | 0.9810 | 0.9307 |
| Precision | 0.9810 | 0.9309 |
| Recall | 0.9811 | 0.9307 |

**HyperParameters:**

- *algorithm*='SAMME', *n_estimators*=100, *learning_rate*=1.01

**Advantages:** Improves accuracy, robust to noise, provides insights.

**Constraints:** Computationally expensive, depends on base classifier.

**Key Takeaways:**

- Fuel consumption is crucial.
- AdaBoost enhances performance.
- Hyperparameter tuning is essential.

Promising approach for CO2 prediction, but consider computational costs and base classifier quality.

**XGBoost:**

*Table 7: XGBoost*

| | With Fuel Consumption | Without Fuel Consumption |
|---|---|---|
| Training Accuracy | 0.9822 | 0.9377 |
| Accuracy | 0.9791 | 0.9265 |
| F1-score | 0.9790 | 0.9263 |
| Precision | 0.9790 | 0.9265 |
| Recall | 0.9790 | 0.9262 |

|  |  |  |
|---|---|---|

**HyperParameters:**

- With Fuel Consumption:
  *max_depth*=20, *learning_rate*=0.2, *subsample*=0.7
- Without Fuel Consumption:
  *max_depth*=25, *learning_rate*=0.2, *subsample*=1

**Advantages:** Ensemble approach, gradient boosting, regularization.

**Constraints:** Computationally expensive, less interpretable.

**Key Takeaways:**

- Fuel consumption is crucial.
- XGBoost offers high accuracy and handles complexity.
- Hyperparameter tuning is essential.

XGBoost is a promising choice for CO2 prediction, but consider computational costs.

**Bagging:**

*Table 8: Bagging*

|  | With Fuel Consumption | Without Fuel Consumption |
|---|---|---|
| Training Accuracy | 0.9879 | 0.9573 |
| Accuracy | 0.9823 | 0.9385 |
| F1-score | 0.9823 | 0.9381 |
| Precision | 0.9823 | 0.9382 |
| Recall | 0.9823 | 0.9381 |

**HyperParameters:**

- With Fuel Consumption:
  *max_features*=0.92, *max_samples*=0.95, *bootstrap_features*=False, *oob_score*=True, *n_estimators*=200

- Without Fuel Consumption:
  *max_features*=0.92, *max_samples*=0.9, *bootstrap_features*=False, *oob_score*=True, *n_estimator*s=200

**Advantages:** Ensemble approach, parallelizable, out-of-bag estimation.

**Constraints:** Computationally expensive, less interpretable.

**Key Takeaways:**

- Fuel consumption is crucial.
- Bagging offers strong performance and robustness.
- Hyperparameter tuning is essential.

Bagging is a promising choice for CO2 prediction, but consider computational costs.

**K-Nearest Neighbors (KNN):**

*Table 9: KNN*

| | With Fuel Consumption | Without Fuel Consumption |
|---|---|---|
| Training Accuracy | 0.9840 | - |
| Accuracy | 0.9798 | - |
| F1-score | 0.9798 | - |
| Precision | 0.9797 | - |
| Recall | 0.9798 | - |

**HyperParameters:**

- *n_neighbors*=3, *metric*='minkowski'

**Advantages:** Simple, non-parametric, lazy learning.

**Constraints:** Computationally expensive, sensitive to distance metric, suffers from curse of dimensionality.

**Key Takeaways:**

- Fuel consumption is crucial.
- KNN performs well, especially for smaller datasets.

- Consider computational costs and distance metric choice.

KNN is a good option for CO2 prediction, especially when interpretability and smaller datasets are priorities.

**Voting Classifier:**

*Table 10: Voting Classifier*

| | With Fuel Consumption | Without Fuel Consumption |
|---|---|---|
| Training Accuracy | 0.9873 | 0.9557 |
| Accuracy | 0.9828 | 0.9388 |
| F1-score | 0.9827 | 0.9384 |
| Precision | 0.9827 | 0.9385 |
| Recall | 0.9828 | 9.9384 |

**HyperParameters:**

- With Fuel Consumption:
  *estimators*=[ ('ab_dt', ab_dt_clf), ('xgb',  xgb_clf), ('rf', rf_clf), ('bagging', bag_clf), ('knn', knn_clf) ] , *voting*='soft'
- Without Fuel Consumption:
  *estimators*=[ ('ab_dt', ab_dt_clf), ('xgb',  xgb_clf), ('rf', rf_clf), ('bagging', bag_clf) ] , *voting*='soft'

**Advantages:** Combines multiple models, reduces overfitting, can be interpretable.

**Constraints:** Computationally expensive, depends on base classifier quality.

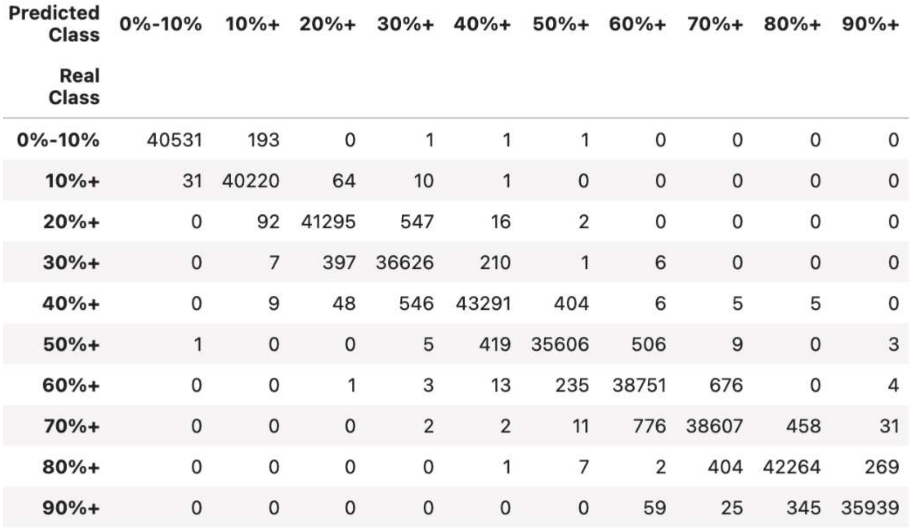
**Key Takeaways:**

- Fuel consumption is crucial.
- Voting classifiers offer strong performance.
- Tune hyperparameters carefully.

Voting classifiers are promising for CO2 prediction, but consider computational costs and base classifier selection.

**Crosstab (Voting - With Fuel Consumption):**

*Table 55: Crosstab Voting Classifier*

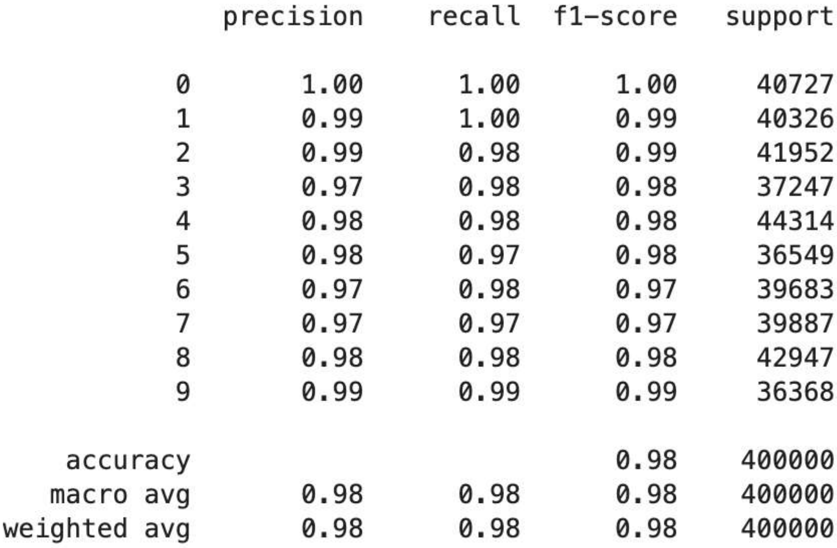


| Predicted Class | 0%-10% | 10%+ | 20%+ | 30%+ | 40%+ | 50%+ | 60%+ | 70%+ | 80%+ | 90%+ |
|---|---|---|---|---|---|---|---|---|---|---|
| **Real Class** | | | | | | | | | | |
| **0%-10%** | 40531 | 193 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| **10%+** | 31 | 40220 | 64 | 10 | 1 | 0 | 0 | 0 | 0 | 0 |
| **20%+** | 0 | 92 | 41295 | 547 | 16 | 2 | 0 | 0 | 0 | 0 |
| **30%+** | 0 | 7 | 397 | 36626 | 210 | 1 | 6 | 0 | 0 | 0 |
| **40%+** | 0 | 9 | 48 | 546 | 43291 | 404 | 6 | 5 | 5 | 0 |
| **50%+** | 1 | 0 | 0 | 5 | 419 | 35606 | 506 | 9 | 0 | 3 |
| **60%+** | 0 | 0 | 1 | 3 | 13 | 235 | 38751 | 676 | 0 | 4 |
| **70%+** | 0 | 0 | 0 | 2 | 2 | 11 | 776 | 38607 | 458 | 31 |
| **80%+** | 0 | 0 | 0 | 0 | 1 | 7 | 2 | 404 | 42264 | 269 |
| **90%+** | 0 | 0 | 0 | 0 | 0 | 0 | 59 | 25 | 345 | 35939 |

The provided confusion matrix evaluates the performance of a classification model, likely a Voting Classifier, on a dataset with CO2 emission predictions. The rows represent the true classes (real CO2 emission levels), while the columns represent the predicted classes.

**Classification Report (Voting - With Fuel Consumption):**

*Table 56: Classification Report Voting Classifier*



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     40727
           1       0.99      1.00      0.99     40326
           2       0.99      0.98      0.99     41952
           3       0.97      0.98      0.98     37247
           4       0.98      0.98      0.98     44314
           5       0.98      0.97      0.98     36549
           6       0.97      0.98      0.97     39683
           7       0.97      0.97      0.97     39887
           8       0.98      0.98      0.98     42947
           9       0.99      0.99      0.99     36368

    accuracy                           0.98    400000
   macro avg       0.98      0.98      0.98    400000
weighted avg       0.98      0.98      0.98    400000
```

**Overall Performance:**

- **High Accuracy:** The overall accuracy of 0.98 indicates that the model is performing well in general.
- **Balanced Metrics:** The macro and weighted averages of precision, recall, and F1-score are all close to 0.98, suggesting balanced performance across classes.

**Class-Specific Performance:**

- **Class 0:** Perfect precision, recall, and F1-score, indicating excellent performance for this class.
- **Classes 1-9:** All classes exhibit high precision, recall, and F1-scores, with minor variations.
- **Class 1:** The model has perfect recall for Class 1, indicating it correctly identifies all instances of this class.

**Key Takeaways:**

- The Voting Classifier with fuel consumption as a predictor demonstrates strong overall performance.
- The model excels in classifying most instances accurately.
- Minor variations in performance across classes might be due to class imbalance or inherent difficulty in distinguishing certain classes.

The classification report indicates that the Voting Classifier is an effective model for CO2 emission prediction. Its high accuracy and balanced performance across classes suggest that it can reliably classify instances into different CO2 emission categories. However, further analysis of specific classes and error cases can provide valuable insights for potential improvements.

# 4 Challenges faced

Throughout the project, we encountered two primary challenges:

1. **Iterative Data Processing and Model Tuning**: A significant amount of time was spent going back and forth between data preprocessing and modeling to achieve acceptable performance metrics. The complexity of feature engineering, along with refining the target variable ($CO_2$ emissions) into appropriate categories, required multiple adjustments. Identifying the most influential variables and optimizing hyperparameters for different models was particularly challenging, especially when trying to balance accuracy with interpretability. This iterative process, while essential, consumed considerable time and effort.
2. **Computational Limitations**: Handling a massive dataset posed a significant challenge in terms of computational resources. Despite migrating to Google Colab to leverage more processing power, certain models, particularly ensemble techniques like Random Forest and XGBoost, took upwards of 10 minutes to train. This not only restricted the number of models we could test but also limited our ability to perform more extensive hyperparameter tuning or deeper exploration of the data. The lengthy training times hindered real-time experimentation and required us to carefully prioritize model choices and configurations.

# 5 Conclusion and Perspectives

In this study, we aimed to predict $CO_2$ emissions based on vehicle technical features using both regression and classification models. After extensive testing and analysis, we identified the best-performing models for each problem type: Linear Regression for regression and Random Forest for classification. These models were selected not only for their performance but also for their interpretability and computational efficiency, making them ideal candidates for deployment in real-world applications.

**Regression Conclusion**

For predicting the absolute value of $CO_2$ emissions, we recommend using Linear Regression. This model achieved an $R^2$ of approximately 0.87 and an RMSE of 14.38, performing comparably to Lasso and Ridge regression. However, Linear Regression stands out due to its simplicity, as it does not require hyperparameter tuning, making it easier to implement and interpret. The regression coefficients provide clear insights into the influence of different technical features on $CO_2$ emissions, allowing for actionable insights for vehicle manufacturers. For instance, vehicle mass and engine power were identified as key factors that increase emissions, while electric energy consumption and hybrid technologies help reduce emissions.

**Classification Conclusion**

For the classification problem, particularly with 7 $CO_2$ emission categories, Random Forest emerged as the best model. It demonstrated high accuracy, with an average F1-score of 0.95 and a prediction time of just 0.1 seconds, making it suitable for real-time applications. The model's interpretability was enhanced through feature importance analysis, which confirmed the relevance of key variables such as engine power and vehicle mass. Additionally, the performance of Random Forest remained robust even when fuel consumption data was excluded, further validating its generalizability.

In more complex scenarios with 10 $CO_2$ classes, we also tested other advanced models like Bagging, XGBoost, and Voting Classifiers, which yielded strong results but at higher computational costs. For instance, XGBoost and Voting Classifiers showed high accuracy and F1-scores but required careful hyperparameter tuning and significant computational resources. While these models are promising, especially for large-scale datasets, they are more computationally intensive and less interpretable compared to Random Forest.

**Why Choose These Models?**

Linear Regression is recommended for its balance between performance and simplicity. It provides clear and interpretable results, making it ideal for use cases where understanding the relationship between variables is essential.

Random Forest is highly accurate, robust to noise, and relatively fast, making it the optimal choice for classification tasks. Its interpretability through feature importance and SHAP analysis offers valuable insights into the factors driving $CO_2$ emissions.

Bagging, XGBoost, and Voting Classifiers are recommended for more complex tasks where higher accuracy is needed, but computational costs should be considered.

**Perspectives**

Moving forward, the insights gained from this study can be applied in various ways. Vehicle manufacturers can use these models to design lighter, more energy-efficient cars, thus reducing $CO_2$ emissions. Additionally, regulatory bodies can leverage these models to set more accurate and informed emission standards. Future work could focus on enhancing the models by incorporating more advanced features, such as real-time driving behavior data for larger datasets.

In conclusion, the chosen models provide a reliable and interpretable framework for predicting and classifying $CO_2$ emissions, offering practical solutions to help achieve sustainability goals in the automotive industry.

# 6 Bibliography

1. Al-Nefaie, A.H.; Aldhyani, T.H.H. Predicting $CO_2$ Emissions from Traffic Vehicles for Sustainable and Smart Environment Using a Deep Learning Model. *Sustainability* **2023**, *15*, 7615. https://doi.org/10.3390/su15097615
2. Zubair, M., Chen, S., Ma, Y. *et al.* Impact of Features on $CO_2$ Emission from Fueling Vehicles. *Iran J Sci Technol Trans Civ Eng* (2024). https://doi.org/10.1007/s40996-024-01439-0
3. Gary Haq and Martin Weiss, *CO2 labelling of passenger cars in Europe: Status, challenges, and future prospects,* Energy Policy, Volume 95, August 2016, Pages 324-335
4. Datascientest training