Project Report on Predicting CO2 Emissions in Vehicles

Submitted by –
Rohan Sachidanand Chopade

Presented to:
Aethrone Aerospace

Index

1. Dataset Overview

The dataset provided for this analysis pertains to the fuel consumption and CO2 emissions of various vehicles. The data contains multiple features that describe the characteristics of the vehicles, such as engine size, fuel type, and the number of cylinders, among others. The goal is to predict CO2 emission based on these features using any 3 regression models. Below is a summary of the key features:

Column Name	Description
CO2EMISSIONS	The target variable representing the CO2 emissions in grams per kilometer.
ENGINESIZE	The size of the engine in liters (e.g., 2.0L, 3.5L).
CYLINDERS	The number of cylinders in the engine.
FUELCONSUMPTION_COMB	The combined fuel consumption (in liters per 100 km).
FUELTYPE	The type of fuel used by the vehicle (e.g., Gasoline, Diesel, etc.)

2. Data Cleaning Process

a. Missing Values:
The dataset was first checked for any missing values in important columns such as ENGINESIZE, CYLINDERS, FUELCONSUMPTION_COMB, and CO2EMISSIONS. Since no missing values were found, there was no need for imputation.

b. Outliers Detection:
Inspected the dataset for potential outliers in numerical features such as ENGINESIZE, CYLINDERS, and CO2EMISSIONS using boxplots and the IQR method. A few outliers were found but were deemed to be valid (e.g., high engine sizes in luxury vehicles). Therefore, no outlier removal was performed.

c. Categorical Variable Encoding:
Encoded categorical features such as VEHICLECLASS, TRANSMISSION, and FUELTYPE using one-hot encoding to convert them into numerical values for regression modelling. We ensured that the encoded variables were added without introducing multicollinearity.

d. Feature Selection
Certain columns such as MAKE, MODEL, and ID were dropped from the dataset because they were non-informative or identifiers that don't directly influence CO2 emissions. This reduced the dimensionality of the dataset and focused on relevant features.

e. Feature Scaling:

Standardized the feature set (ENGINESIZE, CYLINDERS, FUELCONSUMPTION_COMB) using **StandardScaler**. Feature scaling is necessary for algorithms sensitive to the scale of data, like Linear Regression, to improve model performance. However, tree-based models (like Random Forest and Decision Tree) are not sensitive to scaling, so this step was not needed for them.

3. Correlation Analysis

a. Pairplot:

A **pairplot** was generated to visualize the relationships between the features (ENGINESIZE, CYLINDERS, FUELCONSUMPTION_COMB) and the target variable CO2EMISSIONS.

The scatter plots and the kernel density estimate (KDE) distributions show:

- **ENGINESIZE** has a strong positive linear relationship with **CO2EMISSIONS**.
- **CYLINDERS** also shows a positive trend with **CO2EMISSIONS**.
- **FUELCONSUMPTION_COMB** has a direct correlation with **CO2EMISSIONS**.

b. Correlation Heatmap:

A **correlation heatmap** was generated to quantify these relationships numerically. We observed:

- **ENGINESIZE** and **CO2EMISSIONS** have a correlation of **0.87**, indicating a strong positive relationship.
- **CYLINDERS** has a strong positive correlation with **CO2EMISSIONS** (around **0.85**).
- **FUELCONSUMPTION_COMB** also has a strong positive correlation with **CO2EMISSIONS** (around **0.89**).

These high correlation values suggest that all three features are significant predictors of CO2 emissions.

4. Model Training and Evaluation

1. **Models Used**
   We trained three regression models to predict **CO2EMISSIONS**:
   a. **Linear Regression**
   b. **Random Forest Regressor**
   c. **Decision Tree Regressor**

2. **Model Performance Evaluation**
   Each model was evaluated using the following metrics:
   a. **Mean Absolute Error (MAE)**
   b. **Residual Sum of Squares (RSS)**
   c. **R^2 Score**

A. Linear Regression

- **MAE**: 3.46
- **RSS**: 6978.98
- **R^2 Score**: 0.99

Linear Regression showed excellent performance with a very high R^2 score, indicating that it explains almost all the variance in the target variable.

B. Random Forest Regressor

- **MAE**: 1.72
- **RSS**: 7635.51
- **R^2 Score**: 0.99

Random Forest performed very well, with the lowest MAE among the models. However, its RSS was slightly higher compared to Linear Regression, indicating more residual error.

C. Decision Tree Regressor

- **MAE**: 1.62
- **RSS**: 10360.00
- **R^2 Score**: 0.99

The Decision Tree model achieved the lowest MAE but had the highest RSS, suggesting that it might have overfitted to the training data, which can be typical for decision tree models.

5. Comparative Analysis

Model	MAE	RSS	R^2 Score
Linear Regression	3.46	6978.98	0.99
Random Forest Regressor	1.72	7635.51	0.99
Decision Tree Regressor	1.62	10360.00	0.99

- **Observations:**

1. **Linear Regression** provided very stable results and a solid R^2 score of 0.99. However, its MAE was higher than the other models, suggesting that it was not as precise in predicting individual values.
2. **Random Forest Regressor** achieved the lowest MAE and maintained a strong R^2 score, making it the most reliable model in terms of generalization.
3. **Decision Tree Regressor** had the lowest MAE but also the highest RSS, indicating overfitting. This model is more prone to capturing noise in the data rather than generalizing well.

6. Conclusion

Based on the evaluation metrics, the **Random Forest Regressor** is the best-performing model due to its balance between low MAE and relatively low RSS. It consistently performed well in predicting CO2 emissions while avoiding overfitting.