Streamlit

Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration

Project CO2 emissions by vehicles

Executive summary

This project addresses the critical challenge of reducing CO_2 emissions from vehicles, a major contributor to global greenhouse gas emissions, particularly in the transportation sector. By analyzing a large dataset of vehicle registrations and emissions tests from the European Union, this study aimed to understand the key factors influencing CO_2 emissions and develop machine learning models to predict emissions based on vehicle characteristics.

Objectives:

- Analyze the relationship between vehicle characteristics (mass, engine capacity, fuel type) and CO_2 emissions.
- Develop predictive models to accurately estimate emissions based on these features.
- Identify the most significant factors that impact emissions, providing insights for vehicle manufacturers to design more sustainable cars.

Several machine learning algorithms were tested for regression and classification tasks. **Linear regression** was selected for predicting absolute CO_2 values. When fuel consumption was included as a predictor, the model achieved an **R^2 of 0.99** and an **RMSE of 5.6 g/km**. Without fuel consumption, the model still performed well, achieving an **R^2 of 0.87** and an **RMSE of 14.38 g/km**, indicating that fuel consumption plays a significant role but the model remains robust without it.

For classification tasks, **Random Forest** was chosen as the best model. When fuel consumption was included, the model achieved **an accuracy of 98.3%** and an **F1-score of 0.98**. Without fuel consumption, the model still performed well, with **an accuracy of 94.2%** and an **F1-score of 0.94**, demonstrating robustness even when key variables were excluded. The prediction time remained **0.1 seconds**, making it suitable for real-time applications. These models provide actionable insights for vehicle manufacturers and regulators, helping to design greener vehicles and meet sustainability goals.

Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration

Project CO2 emissions by vehicles

Methodology

Data Collection and Preprocessing

- **Choosing the dataset**: The project utilized a dataset from the European Environment Agency (EEA) containing CO_2 emission records for over 10 million vehicles registered in the EU and associated countries.
- **Data Cleaning**: Missing values were handled by removing columns with more than 50% missing data and imputing where necessary. Redundant features with high correlation, such as vehicle mass and fuel consumption, were removed to avoid multicollinearity.
- **Outlier Detection**: Box plots and Z-scores were used to detect outliers, which were retained if they represented legitimate vehicle data, such as high-performance cars.
- **Normalization**: Features like engine size and power were normalized using Min-Max scaling to ensure uniformity for machine learning algorithms.

Feature Engineering

Key features selected for the analysis included vehicle mass, engine capacity, engine power, fuel type, and CO_2 emissions (target variable). Innovative technologies like LED lighting were also encoded for analysis. The analysis was based on the work of Zubair et al., and Al-Nefaie et al.

Machine Learning Models

- **Regression**: Linear regression was chosen due to its balance between simplicity and performance. The model achieved an **R^2 of 0.99** and an **RMSE of 5.6 g/km** when fuel consumption was included, and an **R^2 of 0.87** and **RMSE of 14.38 g/km** without it.
- **Classification**: Random Forest was selected as the best model for classifying vehicles into seven CO_2 emission categories. When fuel consumption was included, it achieved **an accuracy of 98.3%** and an **F1-score of 0.98**. Without fuel consumption, it still performed well with **an accuracy of 94.2%** and **F1-score of 0.94**, demonstrating robustness even when key variables were excluded.

These models provide actionable insights into the relationship between vehicle characteristics and CO_2 emissions, offering potential solutions to reduce the environmental impact of the automotive industry.

Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration

Project CO2 emissions by vehicles

Data Overview

Raw Data

	ID	Country	Vehicle family identification number	Pool	Manufacturer name (EU stan
0	123,100,000	AT	IP-0401362-U5Y-1	KIA	KIA SLOVAKIA
1	123,100,001	AT	IP-0000035-WBS-1	BMW	BMW GMBH
2	123,100,002	AT	IP-0401368-U5Y-1	KIA	KIA SLOVAKIA
3	123,100,003	AT	IP-MQB37AS_A1_1033-WAU-1	VOLKSWAGEN	AUDI AG
4	123,100,004	AT	IP-0401361-U5Y-1	KIA	KIA SLOVAKIA

Dataset contains 1297738 rows and 40 columns.

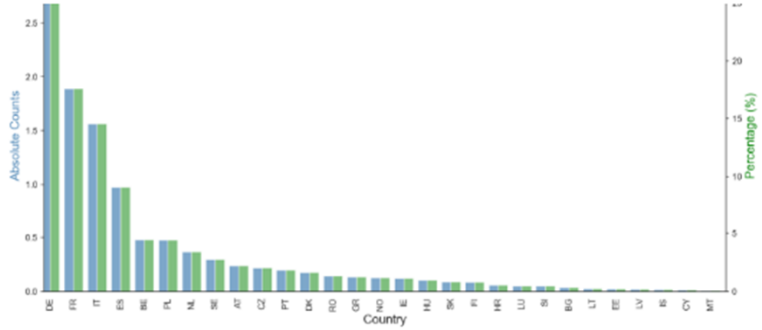
Data Type Inspection

	0
ID	int64
Country	object
Vehicle family identification number	object
Pool	object
Manufacturer name (EU standard)	object
Manufacturer name (OEM declaration)	object
Manufacturer name (MS registry denom	float64
Type approval number	object
Type	object
Variant	object

Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration



Missing Value Handling

Missing values per column:

	0
Pool	133,045
Manufacturer name (MS registry denomination)	1,889,599
Version	1
Specific CO2 Emissions in g/km (NEDC)	1,889,599
Wheel base in mm	1,889,599
Axle width steering axle in mm	1,889,599
Axle width other axle in mm	1,889,599
Engine capacity in cm3	309,733
Electric energy consumption in Wh/km	1,415,326
Innovative technology	577,943

- [ ] Fill Missing Values
- [x] Drop Missing Values

Rows with missing values dropped.

Cleaned Data

	Country	Pool	Manufacturer name (EU standard)	Type	Commercial name
0	AT	KIA	KIA SLOVAKIA	NQ5E	SPORTAGE/2F/GT-LINE/UVO/16 TG
1	AT	BMW	BMW GMBH	G234M	M3 COMPETITION M XDRIVE TOURIN
2	AT	KIA	KIA SLOVAKIA	NQ5E	SPORTAGE/PHEV/GOLD/UVO/16 TGD
3	AT	VOLKSWAGEN	AUDI AG	F3	Q3
4	AT	KIA	KIA SLOVAKIA	NQ5E	SPORTAGE/ANNIVERSARY/16 TGDI

Cleaned dataset contains 962684 rows and 15 columns.

Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration

	Country	Pool	Manufacturer name (EU standard)	Type	Commercial name
0	AT	KIA	KIA SLOVAKIA	NQ5E	SPORTAGE/2F/GT-LINE/UVO/16 TG
1	AT	BMW	BMW GMBH	G234M	M3 COMPETITION M XDRIVE TOURIN
2	AT	KIA	KIA SLOVAKIA	NQ5E	SPORTAGE/PHEV/GOLD/UVO/16 TGI
3	AT	VOLKSWAGEN	AUDI AG	F3	Q3
4	AT	KIA	KIA SLOVAKIA	NQ5E	SPORTAGE/ANNIVERSARY/16 TGDI

Cleaned dataset contains 962684 rows and 15 columns.

☐ Show Correlation Matrix

Pairplot of selected numerical columns

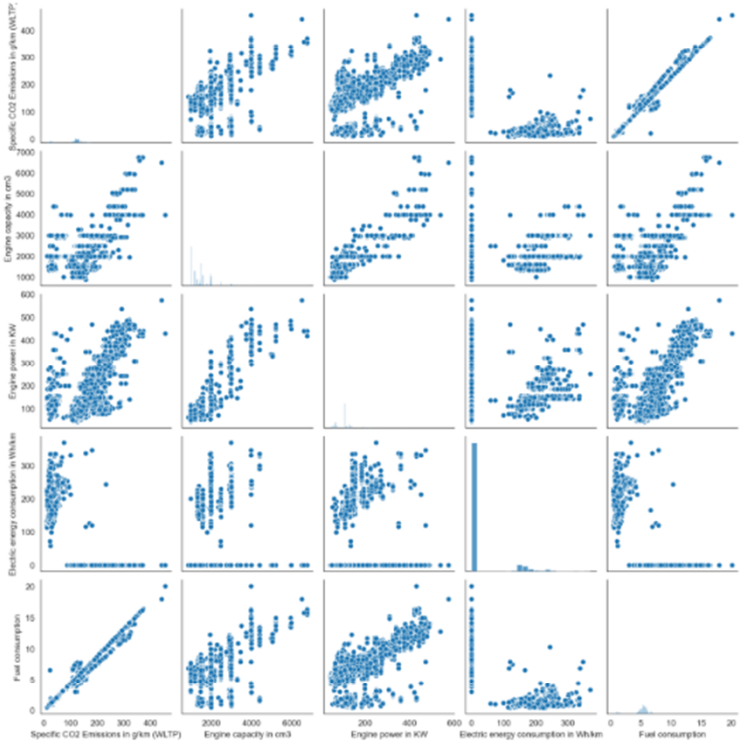


Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration

Project CO2 emissions by vehicles

Modelling

Data Preview

	Mass in running order (kg)	WLTP test mass	Specific CO2 Emissions in g/km (WLTP)	Engine capacity in cm
0	1,337	1,446	126	99
1	1,670	1,782	125	2,48
2	1,493	1,576	135	1,19
3	1,649	1,814	131	1,59
4	1,560	1,640	118	1,98

Dataset contains 7047745 rows and 9 columns.

Correlation Matrix

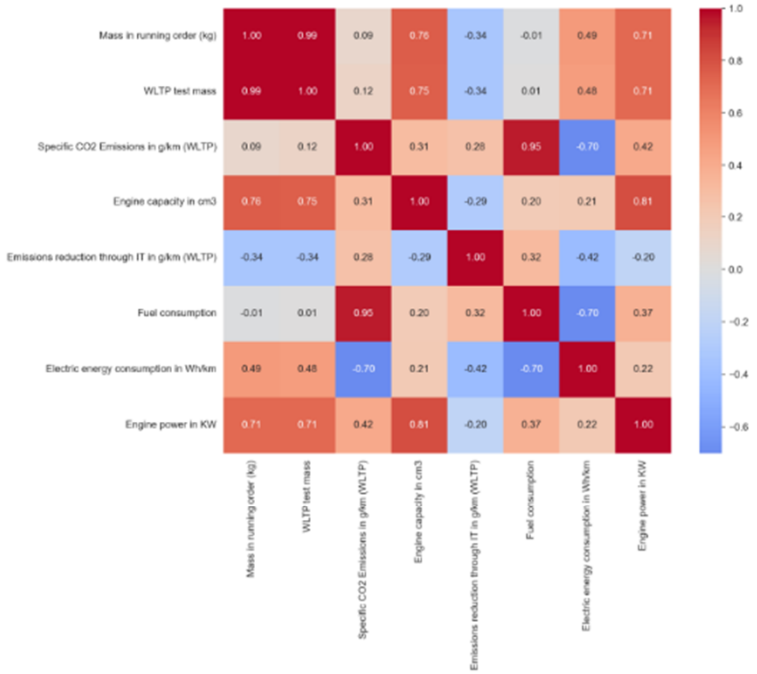


Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration

Choose the type of model:

With Fuel Consumption

Choose the type of task:

Classification

Data Preview

	Mass in running order (kg)	WLTP test mass	Specific CO2 Emissions in g/km (WLTP)	Engine capacity in cm
0	1,337	1,446	126	99
1	1,670	1,782	125	2,48
2	1,493	1,576	135	1,19
3	1,649	1,814	131	1,59
4	1,560	1,640	118	1,98

Dataset contains 7047745 rows and 9 columns.

Handle the 'Fuel type' Column

What would you like to do with the 'Fuel type' column?

- Encode using pd.get_dummies
- Remove from dataset

Data after encoding 'Fuel type' using pd.get_dummies:

	WLTP test mass	Specific CO2 Emissions in g/km (WLTP)	Engine capacity in cm3	Emissions reduction thro
0	1,446	126	999	
1	1,782	125	2,487	
2	1,576	135	1,199	
3	1,814	131	1,598	
4	1,640	118	1,987	

Choose the classification model:

Random Forest

Select Features for Classification

WLTP test mass × | Engine capacity i… × | Emissions reduc… ×
Electric energy c… × | Engine power in … × | Fuel type_diesel ×
Fuel type_diesel… × | Fuel type_e85 × | Fuel type_petrol × | Fuel type_lpg ×
Fuel type_petrol… × | Fuel consumption ×

Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration

Specific CO2 Emissions in g/km (WLTP)

Bins for Target Segmentation

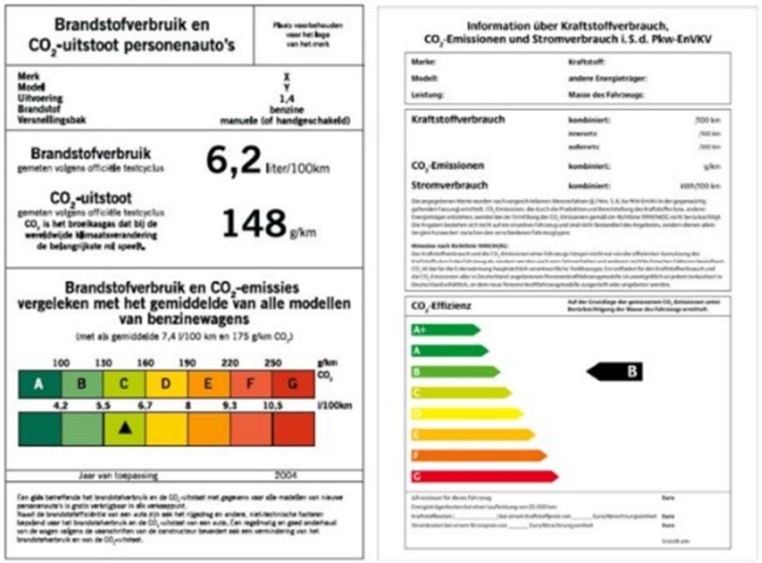
Enter bin ranges (comma-separated):

0, 100, 130, 160, 190, 220, 250, inf

Enter labels for the bins (comma-separated):

0, 1, 2, 3, 4, 5, 6

The `use_column_width` parameter has been deprecated and will be removed in a future release. Please utilize the `use_container_width` parameter instead.



Car Fuel Consumption and CO2 Emission Labels

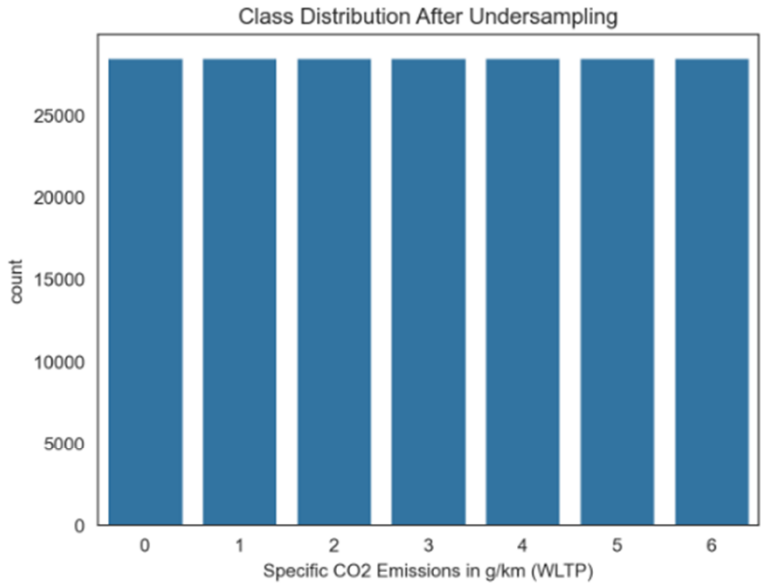
Using the following bins: [0.0, 100.0, 130.0, 160.0, 190.0, 220.0, 250.0, inf] with labels: [0, 1, 2, 3, 4, 5, 6]

Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration

Class Distribution After Undersampling



Test Size

0.20

0.10 0.50

Training set: 159527 samples, Test set: 39882 samples

Scale Features

Features have been scaled using StandardScaler.

Time taken for prediction: 0.3431 seconds

Confusion Matrix:

	0	1	2	3	4	5	6
0	1	1,346	3,975	255	0	0	17
1	1	5,053	665	0	0	0	0
2	0	3,862	1,802	9	0	0	0
3	4	3,766	1,680	215	1	0	0
4	188	3,975	1,542	11	0	0	0
5	34	4,540	1,181	40	0	0	0
6	2,811	2,721	3	184	0	0	0

Classification Report:

Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	5594
1	0.20	0.88	0.33	5719
2	0.17	0.32	0.22	5673
3	0.30	0.04	0.07	5666
4	0.00	0.00	0.00	5716
5	0.00	0.00	0.00	5795
6	0.00	0.00	0.00	5719
accuracy			0.18	39882
macro avg	0.10	0.18	0.09	39882
weighted avg	0.10	0.18	0.09	39882

Generate PCA Plot

SHAP Analysis for Multiclass Classification

Shape of shap_values: `(100, 12, 7)`

Shape of X_test: `(100, 12)`

SHAP Feature Importance for Class 0

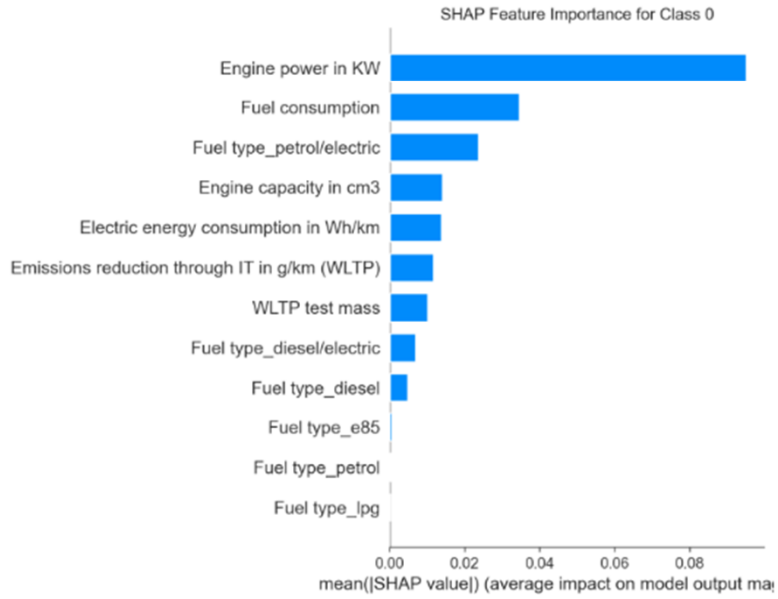


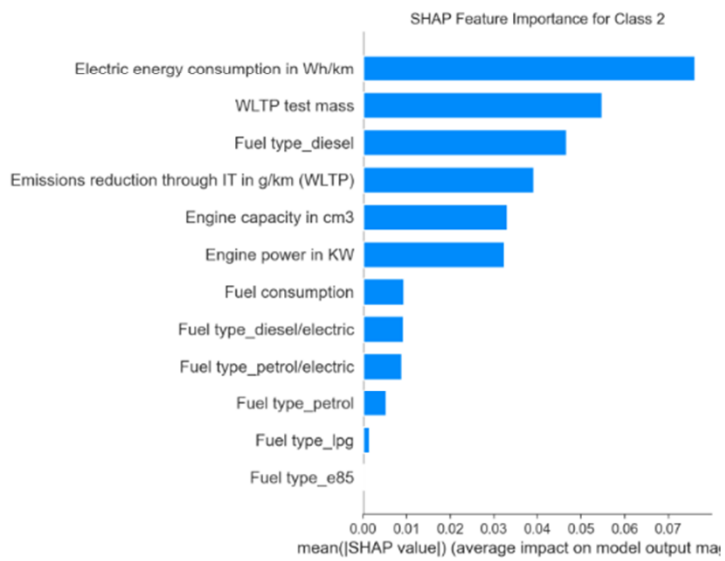
Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration

mean(|SHAP value|) (average impact on model output ma

SHAP Feature Importance for Class 2



SHAP Feature Importance for Class 3

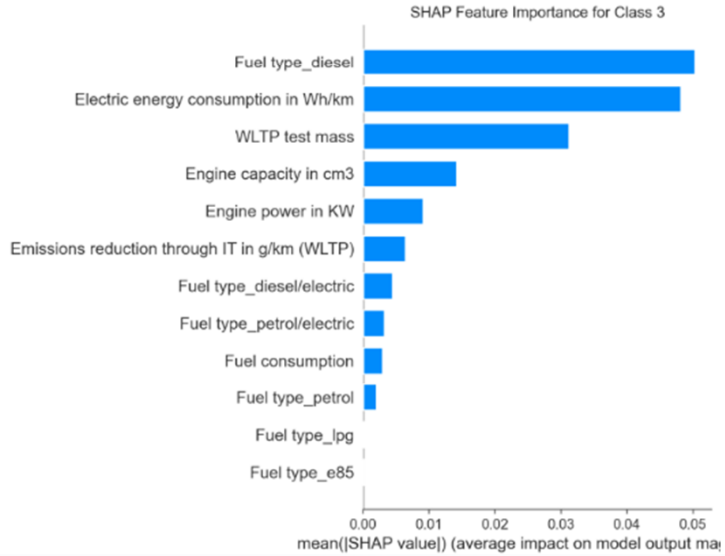
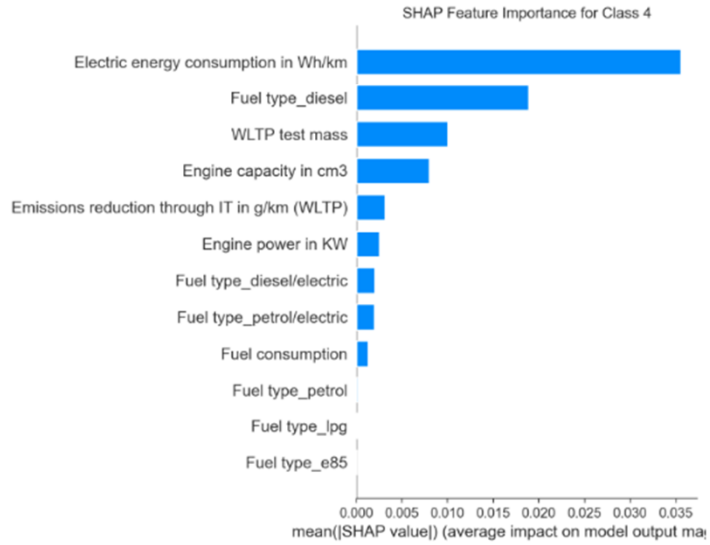


Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration

SHAP Feature Importance for Class 4



SHAP Feature Importance for Class 5

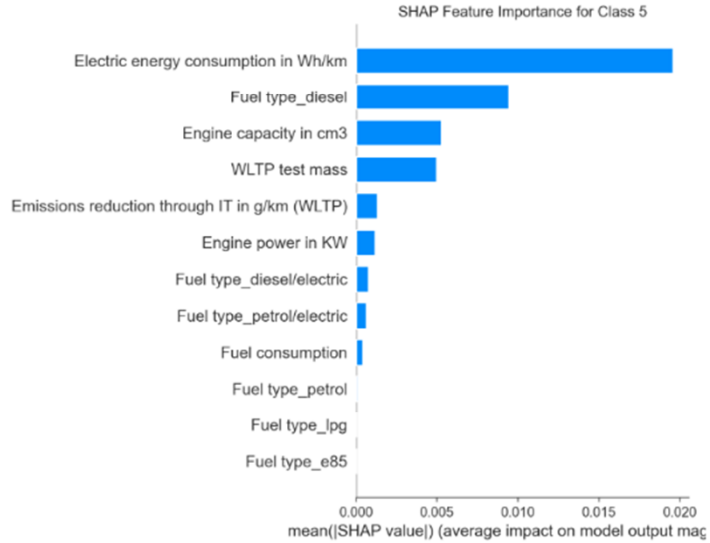
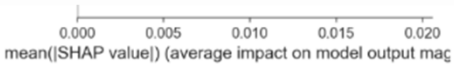


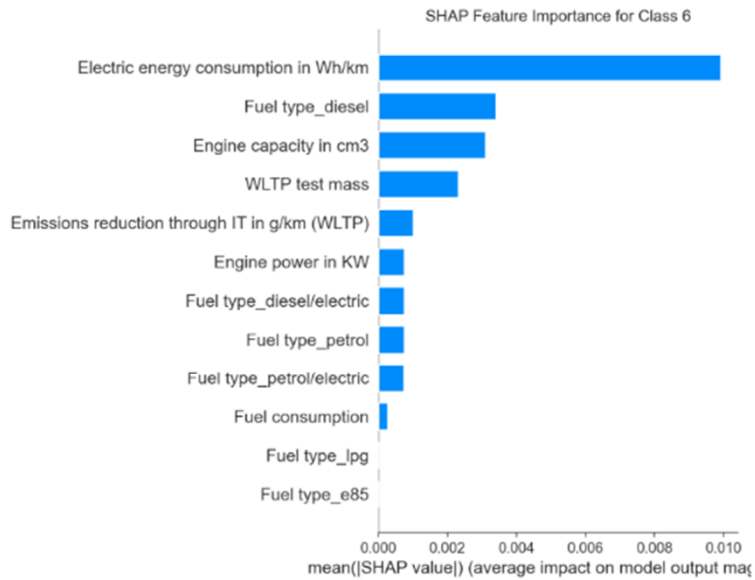
Table of contents

Go to

- Executive summary
- Methodology
- Data exploration
- Modelling
- Additional Exploration



SHAP Feature Importance for Class 6



Overall SHAP Feature Importance