**Component 2**

**Fuel consumption rating**

# **Introduction**

Every year, new vehicle models with unique features are created and added to the existing ones (Placek, 2022). It has become critical to measure the fuel consumption rates of these vehicles because this contributes to the world's current CO2 emissions.

According to a recent study, there is evidence of a link between increased CO2 emissions and poor health (Candanosa, 2021), global warming (Lindsey, 2022), as well as sea-level rise (Knox-Hayes & Gribkoff, 2021).

The project's goal is to identify factors that influence CO2 emissions in passenger vehicles.

# **Project Rationale**

The benefits of this project includes but not limited to:

* Improved understanding of the sources and patterns of CO2 emissions, which can help inform decision-making and policy-making related to reducing emissions.
* Improved ability to identify and address sources of high emissions, such as power plants or industrial facilities, and to target emissions reduction efforts where they will have the greatest impact.

# **Methodologies**

The steps I used to train this model includes:

1. I imported the necessary libraries
2. I loaded and cleaned the data
3. Split the data into training and testing set
4. Select the algorithm of choice and create the model
5. Evaluate the model performance
6. Use the model: Once the model has been trained, it can be used to make predictions

**The Dataset**

The data consist of 13 columns and 5359 rows, all the null entries were dropped from the data set and the columns were casted to the correct data types. The dataset contained five categorical features and the remainder were numerical continuous features.

**Data Analysis**

some exploratory data analysis was performed for better understanding of the features and to determine those with high predictive power. CO2 emissions decreased dramatically over a five-year period, from 2010 to 2014. This could be due to an economic factor, similar to what we saw in 2020 as a result of the COVID outbreak (Stanford Earth Matters Magazine, 2020).

Chart, line chart

Description automatically generated

Figure 1 - Change in CO2 emission by year

Among the manufacturers represented in the dataset, Chevrolet produced approximately 500 vehicles, followed by Ford by a factor of 100.

Chart, bar chart, histogram

Description automatically generated

Figure 2 - Count of vehicle production by manufacturers

Transmission rate corresponds to cylinder size for each recorded transmission type. Chart, bar chart

Description automatically generated

Figure 3 - Transmission vs cylinders

Chart

Description automatically generated with medium confidenceChart, bar chart

Description automatically generated

Figures 4 and 5 - Fuel type by CO2 emission and engine size

The violin plot above shows that, while diesel (D) is less commonly used based on the height of the violin, the width shows that it emits the most CO2, followed by ethanol (E). However, vehicles with larger fuel size tend to use ethanol, while those with smaller engine sizes use diesel.

Model year has a very low correlation, -0.13 with the CO2 emission. Also, from the regression table below, model year has the lowest weight, -2.007 on CO2 emission. Every other feature has a very high correlation and weight on CO2 emission.

A picture containing chart

Description automatically generated

Figure 6 – Correlation Map

Graphical user interface, text, application

Description automatically generated

Regression table

**Model Development**

The model was developed in three stages. The first step is to predict CO2 emissions using only numerical continuous data, followed by predicting every instance of categorical features using all data, and finally identifying groups in the data using only numerical continuous data.

For the first aspect, three regression models were created solely for the purpose of comparing the best fit. This was accomplished by first standardizing the data and then fitting selected features and label. A linear regression model was done using all the numerical features. From the exploratory data analysis carried out and the weight of all the features on the target, it was evident that the “model year” has little or no predictive power so it was dropped from the features and a second linear regression model was carried out. The third model is a decision tree regressor models. The R2 score was used to evaluate all models.

The phase of categorical data classification involved looping through a list of feature names and fitting the models to a random forest classifier, which was done after the features were encoded. On both training and testing sets, the developed model was evaluated using accuracy and f1 score.

The KMeans algorithm was used to define the groups, but first the optimum number of clusters were identified using the elbow method, and the model was fit to the entire dataset using the identified cluster size.

# **Results**

The first linear regression had an R2 score of 0.875, while the second one done after removing “model year” from the features had as R2 score of 0.874, this confirmed the fact the model year has no predictive power on the target. The decision tree had an R2 score of 0.985 which is the best fit.

Chart, bar chart

Description automatically generated

Figure 8 - Regression model performance

Chart, line chart

Description automatically generatedFuel type and make are significant predictors of CO2 emissions, according to the categorical features classification, while the model has little to no impact.

Figure 9 - Performance of the categorical features

Using the elbow method, the optimum number of clusters were identified as 3. However, a comparison was carried out using 4 clusters as well.

Chart, line chart

Description automatically generated Figure 10 - Elbow method result

The figure below shows the difference with cluster size of 4 and cluster size of 3 respectively

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 11 and 12 - Cluster 4 and 3 Comparison

**Evaluation Metrics**

I chose the R2 score for the regression analysis because it can provide insight into the types of errors that the model is making. For example, if the R2 score is close to 1, this indicates that the model is making very few errors and can accurately predict the true values. On the other hand, if the R2 score is close to 0, this indicates that the model is making a lot of errors and is not able to accurately predict the true values. This reason made the decision tree model better than the linear regression. For classification stage, comparing f1 score to accuracy can provide a more comprehensive evaluation of a model's performance and can help identify specific areas for improvement. To avoid overfitting, the features were normalized so the datapoints can be within the same range and scale

**Conclusion**

The project's goal was met satisfactorily with excellent model performance, and the model can now be used in production to help monitor CO2 emissions and identify features that contribute to that effect.

**References**

Candanosa, R. M. (2021). Health Benefits of Reducing Emissions to Mitigate Climate Change. NASA.

Available online:<https://www.nasa.gov/feature/esnt/2021/reducing-emissions-to-mitigate-climate-change-could-yield-dramatic-health-benefits-by-2030>

[Assessed 07/12/2022]

Knox-Hayes, J., & Gribkoff, E. (2021). *Sea Level Rise | MIT Climate Portal*. MIT Climate Portal. Available online: <https://climate.mit.edu/explainers/sea-level-rise>

[Assessed 07/12/2022]

Lindsey, R. (2022). Climate Change: Atmospheric Carbon Dioxide | NOAA Climate.gov. Climate.gov.,

Available online: <https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide>

[Assessed 07/12/2022]

Placek, M. (2022). *Motor vehicle production - Statistics & Facts*. Statista.

Available online: <https://www.statista.com/topics/975/motor-vehicle-production/#topicOverview>

[Assessed 01/12/2022]

Stanford Earth Matters Magazine. (2020). *COVID lockdown causes record drop in carbon emissions for 2020*. Stanford Doerr School of Sustainability.

Available online: <https://earth.stanford.edu/news/covid-lockdown-causes-record-drop-carbon-emissions-2020>

[Assessed 05/12/2022]

United Nations Climate Change. (n.d.). *The Paris Agreement*. UNFCCC.

Available online:<https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement>

[Assessed 07/12/2022]