Predicting Fuel Efficiency with Machine Learning

George Mason University
AIT-582-DL1 | Prof. Dr. Can Nguyen

Arika Bhattarai
Data Analytics and Engineering
George Mason University
Fairfax, VA.
abhatta@gmu.edu

Dhruval Dharmesh Bhau
Data Analytics and Engineering
George Mason University
Fairfax, VA.
dbhau@gmu.edu

Pragya Dhungana
Data Analytics and Engineering
George Mason University
Fairfax, VA.
pdhungana@gmu.edu

Rahul Guttapally
Data Analytics and Engineering
George Mason University
Fairfax, VA.
rguttapa@gmu.edu

Saesha Baniya
Data Analytics and Engineering
George Mason University
Fairfax, VA.
sbaniya@gmu.edu

Sai Saketh Cholleti
Data Analytics and Engineering
George Mason University
Fairfax, VA.
scholle4@gmu.edu

Siddarth Sajjansingh Sandu
Data Analytics and Engineering
George Mason University
Fairfax, VA.
ssandu@gmu.edu

Abstract- In this paper, we have attempted to investigate the role of fuel efficiency in the automotive industry, examining its impact on sustainability, economic viability, and technological progress. We have utilized the Auto MPG dataset and employed machine learning techniques and models to predict fuel economy, specifically concentrating on identifying significant vehicle attributes. To effectively make predictions, we utilized techniques such as data preprocessing, and model evaluation, which includes employing linear regression, decision tree, and random forest algorithms, we reveal key insights into the factors influencing miles per gallon (MPG). Our findings illuminate the significance of fuel-efficient vehicles in addressing environmental concerns, reducing operational costs, and fostering innovation. This research paper will delve deeper to understand fuel efficiency dynamics and the transformative potential of data-driven approaches in shaping the future of automotive industry and consumer preferences.

Keywords- Fuel efficiency, Automotive, Greenhouse, Miles per galloon (MPG), Machine learning, Engineering, Linear Regression, Random Forest, Decision Tree, Automobiles, Economic.

I. INTRODUCTION

In the automotive business, fuel economy is an essential indicator that reflects a vehicle's financial viability in addition to its environmental impact. The capacity to forecast and enhance fuel economy becomes more important as the emphasis on sustainability and energy conservation across the world increases. To find trends and predictions that can estimate a car's miles per gallon (MPG) based on its attributes, this study uses the Auto MPG dataset. Such data could help manufacturers design vehicles that are more fuel-efficient and help buyers make more educated selections.[1]

Car manufacturers are motivated to develop engines of low fuel consumption for the purpose of cutting their operations costs and becoming more competitive. Consumers, on the other hand, are attracted to cars with lower fuel expenses and lower emissions. In addition to this, as we contend with the climbing issues of climate change, governments put forward strict regulations that mandate adherence to stringent fuel efficiency standards. In such a situation, there is a significant inclination in the consumer's choice toward eco-friendly options, which in turn results in an increased longing for fuel-efficient vehicles. These opportunities are addressed by technological innovation as data analytics becomes a useful tool providing such MPG trends and further advancing vehicle design and engineering.[2]

The utilization of data sets like Auto MPG will favor the finding of interdependencies between the characteristics of vehicles and fuel efficiency and this in turn will help the industry in developing predictive models and estimating MPG. With an inclusive use of data-driven solutions, the automotive industry can speed up innovation, acquire a competitive advantage, and become the driving force of broader sustainability goals, thus leading to the evolution towards a more energy-efficient landscape of automobiles.

II. BACKGROUND

Fuel efficiency helps to know the capacity of the vehicle to go a specific distance based on the amount of fuel used. Fuel efficiency is also described as “miles per gallon (mpg)”. Gasoline is a non-renewable energy, and it takes millions of years to form non-renewable energy sources. Are we using gasoline in a proper manner? As we move to the 21st century, due to sustainability, and efficiency, the amount of transportation used has increased among the people. With the increased use of transportation, the use of fuel in a proper manner has become an extreme challenge. With the new innovative technology, gasoline transportation is getting replaced by electric cars, and hybrid cars which are environmentally friendly options to use. According to the EPA,

between 1975-2002, all automobiles' fuel efficiency grew by 101.5% from 13.1 to 26.4mpg [3]. This proves that as time passes by, people have been concerned about using fuel-efficient vehicles.

The use of fuel-efficient vehicles has advantages, such as low gasoline expenses, smaller carbon footprint, less reliance on foreign countries for oil, and many more. Whether someone is about to buy a vehicle or buys a new vehicle, the first question that will be asked is how the miles per gallon of the car is. This is a valid question to be asked as high miles per gallon cars means the vehicle has better fuel efficiency and people will choose the vehicle that gives high miles per gallon as it will reduce the gasoline expense and less emission of greenhouse gasses. With less emission of greenhouse gasses, it will also contribute to global climate change, reduction of carbon footprint, and decrease in pollution. Overall, the use of fuel-efficient vehicles is environmentally friendly, very helpful for energy conservation, as well as economically friendly for individuals as they can save money on gas expenses.

III. WHY THE PROBLEM IS IMPORTANT?

The search for fuel-efficient cars is important from both an environmental and financial viewpoint, as well as being an issue of personal economy. Higher fuel-efficient cars utilize less petroleum, a non-renewable resource, and emit fewer greenhouse gasses (GHGs), like carbon dioxide, which are the primary drivers of pollution in the air and global warming. Therefore, increasing fuel efficiency is essential to lowering the environmental impact of the automotive sector, battling climate change, and advancing energy independence by lowering the global oil demand. [4]

The pursuit of fuel efficiency stimulates technological innovation within the automotive sector. Manufacturers allocate many resources toward research and development endeavors aimed at enhancing engine efficiency, exploring alternative fuel technologies, and refining aerodynamics. These innovations aimed at fuel efficiency not only improve the performance of fuel-efficient vehicles but also accelerate progress in conventional automobiles, leading to a wider range of efficient and environmentally friendly transportation options for consumers.

Furthermore, from the perspective of the consumer, fuel efficiency has an immediate impact on operational expenses, making it a crucial consideration when buying a car. Understanding the factors that affect fuel efficiency can help automakers create more competitive, ecologically friendly, and efficient cars. Overall, the search for fuel-efficient cars is essential for addressing environmental, economic, and energy security challenges while fostering technological progress in the automotive industry.

IV. RESEARCH QUESTIONS AND HYPOTHESIS

A. Research Questions

- What vehicle characteristics are most indicative of fuel efficiency?
- Can a model be developed to predict a vehicle's MPG accurately using these characteristics?

B. Hypothesis:

We hypothesize that specific vehicle attributes, notably engine displacement, the number of cylinders, and overall weight, significantly influence fuel efficiency. By analyzing these features, we aim to develop a predictive model that accurately estimates MPG.

V. LITERATURE REVIEW

The use of Machine Learning (ML) techniques to predict fuel efficiency has significantly increased in recent years. The prediction by ML models for fuel efficiency has provided tangible benefits to the environmental sustainability and economic factors within the automotive sectors. Machine learning algorithms offer a powerful tool for developing predictive models that can accurately estimate MPG based on attributes such as engine specifications, vehicle weight, and aerodynamics.

Preliminary investigations into existing research reveal a burgeoning interest in applying machine learning techniques to predict automotive fuel efficiency. The foundational works in statistical learning by Hastie, Tibshirani, and Friedman (2009) provide a backdrop for our methodology. These sources underscore the potential of regression analysis and other predictive modeling techniques in understanding and forecasting fuel efficiency from vehicle attributes. [6]

Additionally, numerous studies have been conducted to explore the application of machine learning techniques to predict fuel efficiency in vehicles and the crucial need to reduce fuel consumption in transportation. One such study was conducted by a PhD student at West Virginia University which explores how machine learning methods can be applied to forecast fuel efficiency in vehicles, underscoring the importance of cutting fuel consumption in vehicles for both environmental and economic reasons. The research aims to develop a machine learning model that can be premised on fuel consumption history in the use of diesel heavy-duty trucks, using few engine parameters for easy implementation. It also focuses on creating advanced technologies to address sustainability issues in the transport sector mainly through improved fuel efficiency using machine learning techniques. [2]

Using predictive analysis and ML techniques for fuel consumption optimization to promote eco-friendly vehicle systems and telematic platforms is a prospective idea in the long run. The use of fuel-efficient is very promising in achieving environmental sustainability, economic considerations, and consumer decisions in the automotive industry.

VI. PROPOSED APPROACH

Our proposed approach for predicting fuel efficiency using the Auto MPG dataset encompasses a holistic and methodical strategy, starting with an in-depth exploration of the dataset to understand the intricacies and relationships between variables. This phase includes a thorough statistical analysis to identify patterns, outliers, and potential correlations. Following this, we will undertake meticulous data preprocessing, involving the handling of missing values, feature engineering to uncover new insights, and the normalization or standardization of numerical variables to ensure uniformity across the dataset. Categorical

variables will be encoded to facilitate their interpretation by machine learning algorithms. We'll establish a baseline model for performance benchmarking, then experiment with a range of more complex regression models, including decision trees, random forests, and gradient boosting machines, optimizing them through hyperparameter tuning to enhance predictive accuracy. The training and validation of models will be rigorously conducted, employing data splitting and k-fold cross-validation to assess their performance and generalizability on unseen data. We will explore a variety of models, including linear regression for its simplicity and interpretability, and more complex models like random forests and gradient boosting for their ability to handle nonlinear relationships and interactions between features. [4]

VII. DATASET

The dataset we chose covers a wide range of vehicle attributes organized by model and specifies each car by its unique features. It includes 399 rows and 9 columns. Parameters include mpg, cylinders, displacement of an engine, horsepower, weight, acceleration, model year, origin, and car name. Provided with it this data set is quite useful for many good purposes. The dataset provides essential insights for evaluating environmental impacts, assisting in the assessment of carbon emissions and fuel consumption linked to various car models. It also supplies predictive model purposes, thus making it possible to determine the fuel efficiency of a particular car using its peculiar attributes with the help of machine learning algorithms.

Moreover, this dataset is very valuable when we want to compare parallel analyses, identifying tendencies of automotive technology and the evolution of designs for different model years and origins. Fundamentally, this dataset will help us understand, forecast, and improve fuel efficiency in the automotive industry.

A. *Below is the dataset and their Data Type.*

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino

Figure 1: Dataset table

B. *The description of the dataset is given below:*

- mpg: The miles per gallon, which is the target variable for our prediction.
- cylinders: The number of cylinders in the engine.
- displacement: The engine displacement in cubic inches.
- horsepower: Engine horsepower as a string, which might need conversion to numeric and handling of any special values.
- weight: The weight of the car.
- acceleration: The car's acceleration.
- model year: The year of the car model.
- origin: A categorical feature indicating the origin of the car.
- car name: The name of the car.

C. *Data Types*

```
mpg             float64
cylinders         int64
displacement    float64
horsepower       object
weight            int64
acceleration    float64
model year        int64
origin            int64
car name         object
dtype: object
```

Figure 2: Data type of dataset element

VIII. PRELIMINARY RESULTS

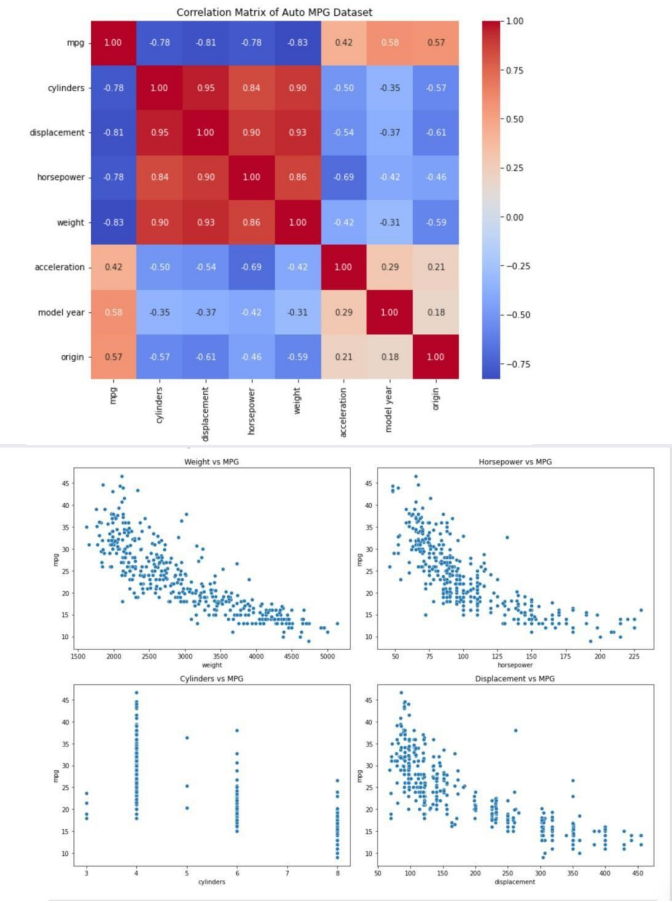

Figure 3: Correlation matrix and scatterplots between mpg and various features

The above correlation matrix and scatter plots provide insight into the relationships between mpg (miles per gallon) and various features: Weight, Horsepower, Cylinders, and Displacement show a negative correlation with mpg. This means that as these values increase, the fuel efficiency tends to decrease.

A. *Scatter plots Description*

1. Weight vs MPG: There's a clear negative trend, indicating that heavier cars tend to have lower MPG.
2. Horsepower vs MPG: Similarly, cars with more horsepower generally show lower MPG, indicating a trade-off between power and fuel efficiency.
3. Cylinders vs MPG: Vehicles with more cylinders tend to have lower MPG, likely due to larger, less efficient engines.
4. Displacement vs MPG: A similar negative trend is observed, where higher engine displacement is associated with lower MPG.

These relationships suggest that weight, horsepower, cylinders, and displacement are significant predictors of fuel efficiency.

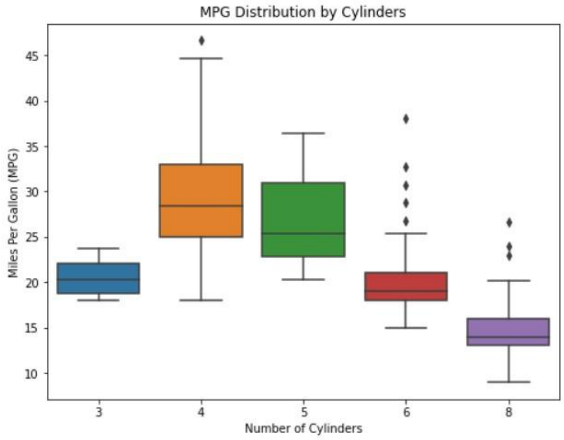

Figure 4: MPG distribution by Cylinders

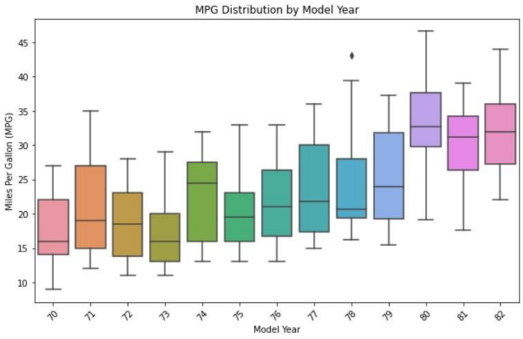

Figure 5: MPG Distribution by Model Year

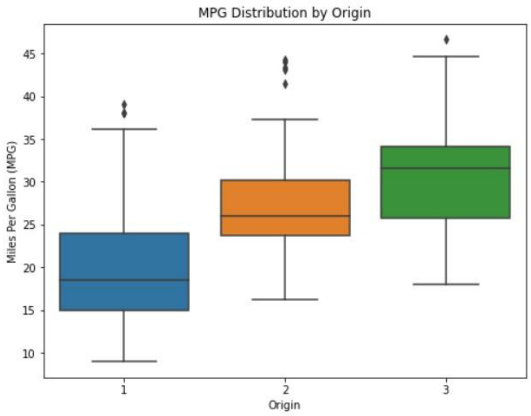

Figure 6: MPG distribution by Origin

B. *Box plots Description*

i. MPG Distribution by Cylinders: The boxplot in Figure 2 reveals how MPG varies with the number of cylinders. It's evident that vehicles with fewer cylinders tend to have higher MPG, reflecting a trend towards better fuel efficiency in engines with fewer cylinders.

ii. MPG Distribution by Model Year: This visualization in Figure 3 shows the progression of fuel efficiency over model years. There's a noticeable trend towards higher MPG in

newer models, indicating improvements in fuel efficiency over time.

iii. MPG Distribution by Origin: The boxplot in Figure 4 by origin shows differences in MPG based on the origin of the vehicle. This could reflect variations in automotive engineering priorities and regulations across different regions.

IX. EVALUATION METHODOLOGY

Our project will employ a diverse set of evaluation techniques to thoroughly assess the performance of various models, including linear regression, random forest, and decision tree algorithms. These evaluation methods will encompass both traditional statistical metrics and machine learning-specific metrics, ensuring a comprehensive understanding of model performance.

- Mean Absolute Error (MAE): MAE will be utilized to measure the average magnitude of errors between predicted and actual values for all models. It offers a straightforward indication of prediction accuracy.
- Root Mean Squared Error (RMSE): RMSE will complement MAE by providing a measure of the square root of the average of squared differences between predicted and actual values. RMSE emphasizes larger errors, offering insight into the overall predictive power of the models.
- Coefficient of Determination (R^2): R^2 will be employed to evaluate the proportion of variance in the dependent variable that can be explained by the independent variables. It serves as an indicator of the goodness of fit of the models.

These evaluation methods will be specifically tailored to each model:

1. Linear Regression: For linear regression models, we will analyze the coefficients of the predictor variables, as well as metrics such as adjusted R^2 to account for the number of predictors. This will help us understand the relationship between the independent and dependent variables and assess the model's overall fit.
2. Random Forest: For random forest models, in addition to the aforementioned metrics, we will examine feature importance scores generated by the algorithm. This will enable us to identify the most influential variables in predicting the target variable and evaluate the model's robustness.
3. Decision Tree: Decision tree models will be evaluated based on metrics such as Gini impurity or entropy to assess the quality of the splits in the tree. We will also visualize the decision tree structure to interpret how the model makes predictions and identify potential areas for improvement.

	Model	RMSE	R2 Score
0	Linear Regression	3.272746	0.790150
1	Decision Tree	3.380604	0.776090
2	Random Forest	2.383317	0.888712

Figure 7: RMSE and R2 square result of regression model

The evaluation of the three regression models on the test set yielded the following results:

Linear Regression - RMSE (Root Mean Square Error): 3.27, R^2 Score: 0.79
Decision Tree - RMSE (Root Mean Square Error): 3.38, R^2 Score: 0.78
Random Forest - RMSE (Root Mean Square Error): 2.38, R^2 Score: 0.89

The Random Forest model performs the best among the three, with the lowest RMSE and the highest R^2 score, indicating the highest prediction accuracy for fuel efficiency (MPG). This suggests that Random Forest is the most effective model for this dataset, capturing the complex relationships between various vehicle characteristics and fuel efficiency.

```
Linear Regression:
RMSE 95% Confidence Interval: [2.6005516  3.89525485]
R² Score 95% Confidence Interval: [0.68301222 0.86693912]

Decision Tree:
RMSE 95% Confidence Interval: [2.46418944 4.3777446 ]
R² Score 95% Confidence Interval: [0.58752011 0.87440684]

Random Forest:
RMSE 95% Confidence Interval: [1.85170907 2.93286154]
R² Score 95% Confidence Interval: [0.80596411 0.93630108]
```

Figure 8: Confidence Interval of Regression Model

This output provides the 95% confidence intervals for the RMSE and R^2 Score metrics obtained through bootstrapping for each regression model.

Random Forest has the narrowest confidence intervals for both RMSE and R^2 Score. This suggests that we're more confident about its performance compared to the other models.

Decision Tree has wider confidence intervals, indicating more uncertainty about its performance.

Linear Regression also has wider confidence intervals, suggesting less confidence in its performance compared to Random Forest.

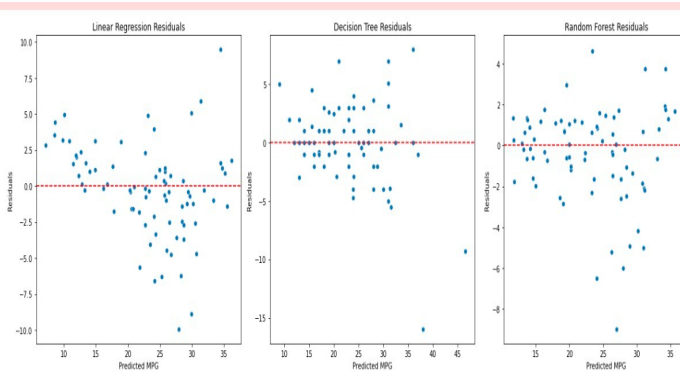

Figure 9: Scatterplot for Regression Residuals

Figure 9 represents the residual plots for the Linear Regression, Decision Tree, and Random Forest models:

1. Linear Regression Residuals: This plot shows the residuals (differences between actual and predicted values) for the Linear Regression model. Ideally, we want the residuals to be randomly dispersed around the horizontal line at 0, indicating that the model's errors are distributed evenly. While there's a general pattern of random dispersion, there might be some areas where the model consistently overpredicts or underpredicts MPG.

2. Decision Tree Residuals: The residuals for the Decision Tree model also show a random pattern, but with potentially more variability in errors compared to Linear Regression. This might indicate the model's higher sensitivity to specific data points, leading to larger errors for some predictions.

3. Random Forest Residuals: The Random Forest model's residuals appear to be more tightly clustered around the 0 line, suggesting that this model generally makes more accurate predictions with fewer large errors. The dispersion pattern indicates a good fit, with errors randomly distributed and no obvious pattern that would suggest systematic overfitting or underfitting.

These residual plots offer valuable insights into the prediction accuracy and reliability of each model. The more randomly the residuals are distributed, the more confidence we can have in the model's predictions across different values of MPG. The Random Forest model, in particular, shows a pattern that suggests it is the most reliable of the three models for predicting vehicle fuel efficiency, corroborating the earlier performance metrics (R^2 and RMSE).

X. CONCLUSION

This research paper focuses on importance of fuel efficiency within the automotive sector, underscoring its central role in promoting sustainability, economic robustness, and technological advancement. With the use of machine learning techniques on the Auto MPG dataset, we have identified influential vehicle attributes that impact fuel economy, providing valuable insights for manufacturers and consumers alike. Our findings highlight the necessity of prioritizing fuel-efficient vehicles to alleviate environmental effects, lower operational expenses, and comply with regulatory requirements. Furthermore, this paper also emphasizes the transformative capabilities of data-driven methodologies in shaping both the future of automotive engineering and consumer preferences. By utilizing advanced analytical tools and datasets, we can accelerate the development of eco-friendly vehicles and propel the industry towards a more sustainable and efficient future. In today's generation and the world of constant climate change, this research paper provides an effective way to promote fuel efficiency.

XI. REFERENCES

[1] Davis, Maggie. "Study: Fuel Efficiency Has Improved 35.4% | LendingTree." *Lending Tree*, 7 August 2023, https://www.lendingtree.com/auto/fuel-efficiency-study/ . (Accessed 25 February 2024).

[2] Hastie, T., Tibshirani, R., & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.)*. Springer Series in Statistics. Available at: https://hastie.su.domains/Papers/ESLII.pdf. (Accessed: 02 February 2024).

[3] Katreddi, S. (2023). Development of Machine Learning based approach to predict fuel edict fuel consumption and maintenance cost of Heavy-Duty Vehicles using diesel and alternative fuels. The Research Repository at WVU. Available at: https://www.researchgate.net/publication/333367045_A_Machine_Learning_Model_for_Average_Fuel_Consumption_in_Heavy_Vehicles (Accessed: 25 February 2024).

[4] Utah Department of Transportation Research & Innovation Division. (2021) *Utilizing machine learning to cross-check traffic data and Understand Urban Mobility*. Available at: https://rosap.ntl.bts.gov/view/dot/56869/dot_56869_DS1.pdf (Accessed: 02 February 2024).

[5] Venkataraman, R. (2020) *Predicting vehicle fuel efficiency*. Medium. Available at: https://towardsdatascience.com/predicting-vehicle-fuel-efficiency-c6065479a72f (Accessed: 01 February 2024).

[6] Xie, X., Sun, B., Li, X., Olsson, T., Maleki, N. and Ahlgren, F. (2023) Fuel consumption prediction models based on machine learning and *Mathematical Methods*. MDPI. Available at: https://www.mdpi.com/2077-1312/11/4/738 (Accessed: 01 February 2024).