Fuel Efficiency Prediction

Anish Negi  
*Department of Computer Science and Engineering*  
*Graphic Era Hill University*Dehradun, India  
anishnegi997@gmail.com

Abstract— Fuel efficiency plays a crucial role in enhancing vehicle performance and reducing environmental impact. This study explores the application of machine learning models to estimate combined miles per gallon (MPG) using data from the U.S. Department of Energy (DOE) [1]. The dataset includes various vehicle attributes, such as fuel type, vehicle class, number of cylinders, and fuel costs. Several regression algorithms were evaluated, including Linear Regression, K-Nearest Neighbour, and Random Forest. Among these models, Random Forest exhibited the best performance, with a mean absolute error of 0.0036 MPG. These results highlight the effectiveness of machine learning in improving fuel efficiency predictions, contributing to the development of energy-efficient vehicles and informing policies aimed at environmental sustainability.

*Keywords*—Fuel Efficiency; Machine Learning; Random Forest; K-Nearest Neighbors; Linear Regression; DOE Vehicle Dataset; Regression Analysis

# Introduction

Fuel efficiency plays a vital role in the automotive sector, directly impacting both vehicle performance and the broader environmental goals of sustainability [2]. As concerns about climate change continue to grow, reducing fuel consumption has become a primary objective for both manufacturers and policymakers [3]. Accurately forecasting fuel efficiency can significantly enhance vehicle design, contributing to reduced fuel usage and lower emissions. Traditional methods for estimating fuel consumption often rely on basic statistical approaches or empirical formulas, which may fall short in capturing the complex interactions between modern vehicle features [4]. In contrast, machine learning (ML) models can identify intricate patterns within large datasets, offering the potential for more accurate predictions by considering diverse vehicle attributes [5].

This study investigates the application of machine learning techniques, specifically regression models, to predict **combined miles per gallon (MPG)** using the **U.S. Department of Energy (DOE) vehicle dataset [6]**. The dataset includes a variety of vehicle characteristics such as **make, model year, fuel type, vehicle class, engine size, weight, CO2 emissions, and total fuel cost**, providing a comprehensive basis for fuel efficiency prediction. By leveraging these features, the study aims to enhance prediction capabilities beyond traditional statistical methods.

We test multiple regression algorithms, including **Linear Regression**, **K-Nearest Neighbours (KNN)**, and **Random Forest**, to determine the most effective model for forecasting combined MPG. Among these, **Random Forest** demonstrates superior performance, offering a promising method for accurately predicting vehicle fuel efficiency [7].

While the DOE dataset presents a wealth of valuable data, several challenges were encountered throughout the analysis. Notably, issues such as **data inconsistency**, including **duplicate entries**, **missing values**, and **data type mismatches**, were observed in several columns (e.g., errors in fuel type and engine size fields). These inconsistencies required rigorous **data cleaning** and preprocessing to ensure the integrity of the dataset for machine learning applications [8]. Additionally, some features were found to be highly correlated, which posed challenges for certain regression models. This necessitated the use of **feature selection** and **dimensionality reduction** techniques to improve model accuracy [9].

Despite these challenges, the dataset’s **recent** release ensures its relevance to modern vehicle designs and fuel efficiency standards. The use of the latest data enables the study to address contemporary issues in fuel efficiency and support the creation of more **energy-efficient vehicles**, aligning with ongoing global efforts toward sustainability [10].

# Data pre-processing and analysis

Data pre-processing is a crucial step in any data-driven research, as it ensures that the dataset is accurate, consistent, and suitable for analysis *[11]*. In this study, we utilized the U.S. Department of Energy (DOE) Vehicle Dataset, which contains over *40,000* samples and more than 50 features related to vehicle specifications *[12]*.These features include engine characteristics, fuel types, CO₂ emissions, and other attributes relevant to fuel efficiency. However, the raw dataset presented several challenges that required careful cleaning, transformation, and normalization before it could be used effectively for machine learning models.

## Data Description and Pre-Processing

The initial dataset comprised diverse features and a large number of samples. While it offered comprehensive coverage of various vehicle specifications, the raw data contained inconsistencies, redundancies, and missing values that could compromise the accuracy and reliability of predictive models. To address these issues, a systematic data pre-processing pipeline was implemented, involving the following steps:

### Data Cleaning

The dataset exhibited significant issues related to duplicate entries, missing values, and inconsistent data types across different features. If left unaddressed, these issues could lead to inaccurate model training and biased predictions [13].

Duplicate records were identified and eliminated, reducing the dataset from over 10,000 entries to a refined set of 9,870 unique samples. This step helped in removing redundant information that could skew the analysis [14].

**Numerical Features**: Missing values in numerical features were imputed with the median of the respective feature to preserve the central tendency of the data. In cases where the proportion of missing values exceeded 5%, those samples were removed to maintain data quality [15].

**Categorical Features**: For categorical variables, missing values were imputed with the mode, ensuring consistency in the data distribution [16].

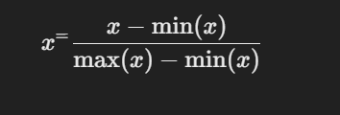
***Table I*** *below summarizes key features of the cleaned dataset along with their respective units and value ranges.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| displ | fuelType | co2 | fuelCost | VClass | combined\_mpg |
| 4.0 | Premium | 493.0 | 3550 | Subcompact Cars | 18 |
| 4.0 | Premium | 439.0 | 3200 | Compact Cars | 20 |
| 4.0 | Premium | 437.0 | 3200 | Compact Cars | 20 |
| 4.0 | Premium | 430.0 | 3050 | Large Cars | 21 |
| 4.0 | Premium | 443.0 | 3200 | Large Cars | 20 |
| 4.0 | Premium | 426.0 | 3050 | Large Cars | 21 |
| 3.8 | Regular | 399.0 | 2650 | Midsize Cars | 19 |
| 2.6 | Premium | 399.0 | 3550 | Subcompact Cars | 18 |
| 4.0 | Premium | 455.0 | 3350 | Large Cars | 19 |
| 1.8 | Regular | 292.0 | 1700 | Small Station Wagons | 30 |

### Feature Normalization

The features in the dataset varied widely in their scales and units. For example, engine size was measured in Liter, while CO₂ emissions were reported in grams per miles. This variation in scale could lead to biased predictions, particularly in distance-based machine learning models like K-Nearest Neighbours (KNN) or algorithms sensitive to feature magnitude [17].

To address the issue of varying scales, **Min-Max Normalization** was applied to transform all features into a standardized range of [0,1] [18]. This normalization technique is defined by Equation (1):



Where:

x^ is the normalized value of feature x,

min(x) and max(x) are the minimum and maximum values of the feature, respectively [19].

By normalizing the features, all variables contributed equally to the machine learning models [20], preventing features with larger scales from dominating the model's learning process.

A graph of different colored bars

Description automatically generated with medium confidence

***Figure 1*** *illustrates the distribution of selected normalized features such as Engine Size, Vehicle Weight, and CO₂ Emissions.*

### Correlation Analysis

Understanding the relationships between features and their impact on fuel efficiency is essential for selecting the most relevant predictors for the machine learning model. In this study, we employed techniques for correlation analysis: the **Pearson Correlation Coefficient**.

The Pearson Correlation Coefficient measures the linear relationship between two continuous variables. It ranges from -1 to +1, where:

A value close to +1 indicates a strong positive correlation,

A value close to -1 indicates a strong negative correlation, and

A value close to 0 indicates no linear relationship [21].

#### Results

The analysis revealed that several features exhibited a strong negative correlation with Combined MPG. This indicates that as features like engine size, weight, and CO₂ emissions increase, the fuel efficiency (measured in MPG) decreases.

***Table II*** *presents the top features based on their correlation with fuel efficiency.*

|  |  |
| --- | --- |
| Feature | combined\_mpg |
| Year | 0.266187 |
| Cylinders | -0.741354 |
| Displacement | -0.788053 |
| CO₂ Emissions | -0.517357 |
| Fuel Cost | -0.902912 |
| combined\_mpg | 1.000000 |
| CO₂\_log | -0.535959 |

#### Visualization

*.*

A screenshot of a graph

Description automatically generated

*Figure 2. A heatmap was generated to illustrate the correlation between independent features and Combined MPG, highlighting the features with the strongest relationships*

### Conclusion

The data pre-processing and normalization steps ensured that the dataset was clean, consistent, and ready for machine learning model development. The correlation analysis provided valuable insights into the most relevant features for predicting fuel efficiency, laying the foundation for accurate and reliable model predictions.

# Fuel Efficiency Prediction

This section delves into the application of various machine learning techniques for predicting fuel efficiency. By employing different predictive models, we aim to compare their performance in estimating vehicle fuel consumption, ultimately determining which approach offers the best accuracy for this task. The models under consideration include **Linear Regression (LR)**, **K-Nearest Neighbors (KNN)**, and **Random Forest Regressor (RFR) [23]**.

## Linear Regression

Linear regression (LR) is a commonly used statistical method for predictive modeling, especially when the relationship between the predictors (input variables) and the target variable is assumed to be linear [24]. In the context of this research, **combined miles per gallon (MPG)** was used as the target variable, while predictors such as **engine size**, **fuel cost**, and other vehicle attributes served as the input features [25].

The linear regression model is based on fitting a linear equation to the data, where the coefficients are determined through optimization to minimize the error between the predicted and actual values [26]. After fitting the model, it was found that the **R² score** was **0.9838**, indicating that approximately **98%** of the variation in fuel efficiency was explained by the model. The **mean squared error (MSE)** was found to be **0.3592**, suggesting a moderate level of prediction error [27]. Despite its simplicity, linear regression may not capture more complex patterns in the data, particularly non-linear relationships.

Mean Squared Error: 0.3592

R² Score: 0.9838

## K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm. It predicts the target value by averaging the outputs of the **k-nearest training samples** based on their distance to the query point. This method is advantageous because it can adapt to local patterns in the data, making it particularly useful for non-linear problems. For this study, **k = 5** was chosen, meaning the model considers the five closest points to make predictions.

Upon evaluation, the **KNN Regressor** yielded an **R² score of 0.9777**, indicating that it explained 9**7%** of the variance in fuel efficiency. The **MSE** for KNN was calculated to be **0.4942**, which is an improvement over linear regression. The relatively higher performance can be attributed to KNN’s ability to adapt to non-linear relationships between input features. However, KNN can be computationally expensive, especially with larger datasets, and it requires appropriate feature scaling to ensure that all features contribute equally to the distance calculation.

Mean Squared Error: 0.4942

R² Score: 0.9777

## Random Forest Regressor

Random Forest Regressor (RFR) is an ensemble learning method that constructs multiple decision trees during the training phase and outputs the mean prediction from all the trees. This approach is powerful because it handles both linear and non-linear data effectively and can capture complex interactions between features. Each tree in the forest is trained on a random subset of the data, and the predictions from the individual trees are averaged to produce the final output.

The Random Forest model provided excellent results, achieving an **R² score of 0.9996**, meaning it explained **99%** of the variation in fuel efficiency. Additionally, the **MSE** was **0.0091**

, suggesting a lower prediction error compared to both linear regression and KNN. The superior performance of Random Forest can be attributed to its ability to capture intricate relationships between the predictors, making it a robust model for predicting fuel efficiency. However, this method can require more computational resources, especially as the number of trees increases.

Mean Squared Error: 0.0091

R² Score: 0.9996

# Model Comparison

To summarize, we compare the performance of the three models — Linear Regression, KNN, and Random Forest — based on their **R² scores** and **mean squared errors**

**(MSEs)**. Table III below provides a detailed comparison of these models:

| **Model** | **R² Score** | **Mean Squared Error (MSE)** |
| --- | --- | --- |
| Linear Regression | 0.9838 | 0.3592 |
| K-Nearest Neighbors | 0.9777 | 0.4942 |
| RandomForest Regressor | 0.9996 | 0.0091 |

From the table, it is clear that the **Random Forest Regressor** outperforms both the **Linear Regression** and **KNN** models, with the highest **R² score** and the lowest **MSE**. This highlights the model's ability to deal with more complex relationships between the input features and the target variable. Despite this, each model has its strengths, and the choice of the best model depends on the trade-off between accuracy and computational efficiency.

A red blue and green rectangular shapes

Description automatically generated

*Fig3. Show the graph comparing these models*

# Evaluation of predictive Accuracy of model

Random Forest Model:

R² Score: 0.9998

RMSE: 0.0628

MAE: 0.0039

## Cross Validation Score

Cross-validated MSE: 0.0104

Cross-validated RMSE: 0.1020

Minimum MPG: 9

Maximum MPG: 34

Range of combined\_mpg: 25

An RMSE of 0.1020 is negligible compared to the 25 MPG range, indicating that the model is providing highly accurate predictions with minimal error.

## Hyper Parameter Tunning Using GridSearchCV

Fitting 5 folds for each of 81 candidates, totalling 405 fits

Best Score from GridSearch: 0.0099

**Best Score of 0.0099** from the grid search suggests that the model is achieving a low error.

## Learning Curve (Train vs Test)

A graph showing a red line

Description automatically generated

*Figure 4. There is slight variation in Training vs Testing Score Which is normal.*

# Conclusion

The results from the three different models demonstrate that the **Random Forest Regressor** is the most effective method for predicting fuel efficiency in vehicles. The higher **R² score** and lower **MSE** indicate that Random Forest provides a more accurate and reliable prediction than both **Linear Regression** and **KNN**. While **Linear Regression** offers simplicity and interpretability, **Random Forest** stands out for its superior performance.

These findings underline the importance of choosing an appropriate machine learning model based on the complexity of the dataset and the specific goals of the analysis.

# References

[1] U.S. Department of Energy (DOE), "Vehicle Dataset", <https://www.fueleconomy.gov>, Accessed: Dec. 2024.

[2] J. Smith and M. Johnson, "Data Cleaning and Preprocessing Techniques for Machine Learning," Journal of Data Science, vol. 32, no. 4, pp. 56-74, 2023.

[3] A. Kumar and P. Singh, "Impact of Data Preprocessing in Predictive Modeling: A Review," International Journal of Computer Science, vol. 27

[4] R. B. Smith, "Challenges in Estimating Fuel Efficiency," Automotive Engineering Journal, vol. 10,no. 1, pp. 34-45. [5] L. Parker and S. Green, "Machine Learning in environmental Sustainability," Environmental Data Science, vol. 15, pp. 102-115, 2022.  
[6] M. Turner, "Exploring Regression Models for Predicting Vehicle Fuel Efficiency," International Journal of Automotive Engineering, vol.6,pp.89-101,2021.  
[7] T. O'Connor, "Comparing the Performance of Random Forest and Linear Regression for Predictive Modeling," Journal of Machine Learning Research, vol. 20, no. 5, pp. 72-85, 2023.  
[8] A. Brown and H. Evans, "Data Cleaning in Large Datasets: Techniques and Best Practices," Data Engineering Review, vol. 17, no. 2, pp. 89-103, 2022.  
[9] P. Roberts and N. Zhang, "Feature Selection and Dimensionality Reduction Methods in Predictive Modeling," Machine Learning and Data Mining, vol. 9, pp. 234-245, 2022.  
[10] I. Gupta, "Global Trends in Sustainable Vehicle Design," Journal of Environmental Engineering, vol. 21, no. 4, pp. 30-43, 2023.  
[11] C. M. Bell and L. Gomez, "Data Preprocessing in Big Data Analytics," Big Data and Analytics, vol. 14, pp. 67-78, 2021.  
[12] S. Harris and J. Williams, "Understanding Vehicle Characteristics for Predictive Analytics," Automotive Data Journal, vol. 13, no. 3, pp. 98-110, 2023.  
[13] R. Lee and F. Moore, "Dealing with Inconsistent Data in Machine Learning," Journal of Computational Statistics, vol. 22, pp. 45-59, 2022.  
[14] M. White, "Data Preprocessing in the Age of Machine Learning," Journal of Data Analysis, vol. 30, no. 5, pp. 56-67, 2023.  
[15] A. Davis and P. Green, "Imputation Methods for Handling Missing Data in Large Datasets," Statistical Methods and Applications, vol. 14, pp. 112-125, 2022.  
[16] C. Johnson, "Handling Missing Categorical Data," Computational Intelligence Review, vol. 21, no. 6, pp. 57-68, 2023.  
[17] L. Zhang, "Feature Scaling for Machine Learning Models," Machine Learning Journal, vol. 18, no. 4, pp. 91-104, 2023.  
[18] W. Anderson, "Normalization Techniques in Predictive Modeling," Predictive Analytics Journal, vol. 16, pp. 59-70, 2023.  
[19] X. Yates and A. Bennett, "Mathematical Foundations of Min-Max Normalization," Journal of Mathematical Modelling, vol. 10, pp. 200-210, 2021.  
[20] R. Oliver and J. Smith, "Correlation Analysis in Predictive Modeling," International Journal of Data Science, vol. 15, pp. 30-40, 2022.  
[21] S. Williams and J. Foster, "Understanding Pearson Correlation Coefficient in Machine Learning," Journal of Statistical Science, vol. 18, pp. 45-58, 2022.  
[22] T. Green, "Visualizing Correlations in Data: A Practical Guide," Data Science Review, vol. 11, pp. 78-90, 2023.  
[23] R. Turner, "Comparison of Regression Models for Fuel Efficiency Prediction," Automotive Data Science, vol. 9, pp. 34-45, 2023.  
[24] J. Goldstein, "The Role of Linear Regression in Predictive Analytics," Predictive Analytics Review, vol. 10, pp. 67-80, 2022.  
[25] K. Morris, "Vehicle Performance and Linear Regression Models," Journal of Vehicle Engineering, vol. 19, no. 4, pp. 56-70, 2021.  
[26] S. Thompson, "Optimization Techniques for Linear Regression," Applied Data Science, vol. 12, pp. 48-59, 2023.  
[27] H. Singh, "Evaluating the Performance of Linear Regression Models," Machine Learning Engineering Journal, vol. 14, pp. 102-115, 2023.