Fuel Consumption Analysis

Dr Rakesh Kumar M
Computer Science and Engineering
Rajalakshmi Engineering College,
Chennai, India
rakeshkumar.m@rajalakshmi.edu.in

Rohit M
Computer Science and Engineering
Rajalakshmi Engineering College,
Chennai, India
210701215@rajalakshmi.edu.in

Santhosh M
Computer Science and Engineering
Rajalakshmi Engineering College,
Chennai, India
210701233@rajalakshmi.edu.in

Abstract—This project analyzes fuel consumption and CO2 emissions data from various vehicle models using machine learning algorithms such as decision trees and random forests. Through exploratory data analysis and model building, insights into the factors influencing fuel consumption and emissions are gained. The study employs techniques like feature engineering, hyperparameter tuning, and evaluation metrics to optimize model performance. Results demonstrate the effectiveness of machine learning in predicting fuel consumption and CO2 emissions, contributing to environmental sustainability efforts in the automotive industry. Future research may focus on expanding the dataset and exploring additional predictive features for enhanced modeling accuracy.

Keywords— Fuel Consumption · Machine Learning · CO2 emissions · Exploratory Data Analysis

I. Introduction

In the modern automotive industry, understanding and managing fuel use is essential due to growing environmental concerns and rising gasoline costs. With the use of machine learning techniques, particularly Decision Tree and Random Forest, this study seeks to forecast fuel usage in automobiles. We build models to estimate fuel consumption in liters per 100 kilometers by assessing vehicle parameters such as vehicle class, engine size, number of cylinders, gearbox type, CO2 rating, and fuel type. By identifying the variables that affect fuel efficiency, this analysis helps manufacturers create more efficient automobiles, helps consumers make educated purchases, and promotes environmental sustainability.

The project begins with extensive data preprocessing, which includes handling missing values, encoding categorical variables, and scaling numerical features. Univariate and bivariate analyses help understand individual features and their relationships with fuel consumption, while chi-square tests assess the significance of associations between categorical variables. Outlier detection and handling ensure the accuracy and reliability of the dataset. Ordinal encoding and feature scaling prepare the data for model training.

Machine learning models are built using Decision Tree and Random Forest algorithms due to their ability to handle complex relationships and interactions between features. The models are trained on a split dataset, with hyperparameter tuning optimizing their performance. Metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) are used to evaluate model performance, ensuring robust and accurate predictions.

A key feature of this project is the development of an interactive user interface using Streamlit. This interface allows users to input vehicle specifications and receive real-time fuel consumption predictions. It also provides visualizations to explore the relationships between different vehicle attributes and fuel consumption. This tool is easy to use and does not require any prior knowledge of machine learning, making it available to a wide range of users.

This project integrates data analysis, machine learning, and user interface development to create a comprehensive tool for predicting and understanding vehicle fuel consumption. The knowledge acquired can help consumers make better decisions, advance environmental sustainability, and improve car design. The initiative makes a significant contribution to the development of more effective and sustainable transportation solutions by utilizing advanced algorithms and interactive visualizations.

II. Literature Survey

[1] The fuel of the automobile like cars mainly states the distance which is traveled by a car and particular fuel consumption is consumed by car or any vehicle. Fuel usage in cars and other vehicles contribute significantly to air pollution, and it differs greatly between countries. We are witnessing the fuel prices and customers being more particular about the features. To have a vehicle which is desirable and even more efficient, improvement of fuel has been carried out. Primarily, fuel prices differ throughout nations and also rely on the type of vehicle those nations utilize, as well as how frequently the customers use their cars.

[2] In the current situation, a vehicle's efficiency and performance analysis plays an important role. There are several situations in which the user is reluctant to give up the car. When this happens and the driver is unaware that the automobile needs to be disposed of, the relevant authorities need to step in and investigate whether the driver is really using the vehicle above its authorized mileage. To live a sustainable life, it is therefore becoming more and more

important to preserve the environment and nature. The car's engine type, cylinder count, fuel type, and other factors are taken into account while analyzing its performance.

[3] The fuel market affects the income distribution of countries in a direct or indirect way, which has an impact on a number of areas such as the stock market, cost of living, education, and vital commodities. In return, a variety of factors that also affect everyday life for the average person influence fuel prices. Therefore, it is evident that forecasting fuel price patterns is important for the benefit of drivers as well as for economists in every country who need to foresee the economic trends resulting from these price variations and be ready for everything.

[4] One major factor contributing to the problem of global warming is the use of personal vehicles. Gasoline used in cars emits an estimated 24 pounds of carbon dioxide and other greenhouse gases per gallon, which accounts for roughly 20% of all emissions. Over five pounds of heat-trapping pollutants are produced throughout the fuel's extraction, manufacture, and delivery, while it's important to note that cars release over 19 pounds per gallon directly from their tailpipes. At present, gas-powered vehicles typically achieve a fuel efficiency of around 22.0 miles per gallon, covering an annual distance of 11,500 miles. As a consequence, the combustion of one gallon of gasoline results in approximately 8,887 grams of CO2 being released into the atmosphere. In 1998, the automobile industry pledged to reduce new car emissions by 25% by 2008. At that time, CO2 emissions from new cars were approximately 203g/km. Currently, the average emissions of vehicles stand at approximately 170g/km and it is projected that they will not decrease to 140g/km until after 2025. The amount of CO2 emitted by a standard passenger car is typically around 4.6 metric tons per year, but this can fluctuate depending on factors such as fuel type, fuel efficiency, and mileage.

[5] The automotive sector is at the forefront of considerable change in an era when environmental sustainability meets with technical innovation. Accurately predicting vehicle fuel use has become a vital challenge and enormous opportunity as a result of the worsening consequences of climate change and the growing demand for energy efficiency worldwide. Enter the world of machine learning, a dynamic and powerful tool that is changing the way we think about automobile fuel efficiency.

[6] Driving behavior has a large impact on vehicle fuel consumption. Dedicated study on the relationship between the driving behavior and fuel consumption will decrease the energy cost of transportation and the development of the behavior assessment technology for the ADAS system. Therefore, it is important to evaluate this relationship in order to develop more ecological driving assistance systems and improve the vehicle fuel economy.

[7] The primary consumers of gasoline are automobiles. This has an adverse effect on the environment and significantly increases greenhouse gas emissions. Transportation was responsible for over 20% of global carbon dioxide emissions in 2000. Concerned with the long-term effects of carbon dioxide emissions are international environmental standards and specifications groups. The expected shortages in the production of petroleum products in the near future are contributing to the environmental effect and are driving forces behind the development of automobiles that are more fuel-efficient.

[8] A normal tree includes root, branches and leaves. The same structure is followed in Decision Tree. It contains root node, branches, and leaf nodes. Testing an attribute is on every internal node, the outcome of the test is on branch and class label as a result is on leaf node. A root node is parent of all nodes and as the name suggests it is the topmost node in Tree. A decision tree is a tree where each node shows a feature (attribute), each link (branch) shows a decision (rule) and each leaf shows an outcome (categorical or continuous value). As decision trees mimic the human level thinking so it's so simple to grab the data and make some good interpretations. The whole idea is to create a tree like this for the entire data and process a single outcome at every leaf.

[9] Random Forest is a tree-based ensemble with each tree depending on a collection of random variables. It can be used for either a categorical response variable, referred to in as "classification", or a continuous response, referred to as "regression". Similarly, the predictor variables can be either categorical or continuous.

[10] The random forest algorithm, proposed by L. Breiman in 2001, has been extremely successful as a general-purpose classification and regression method. The approach, which combines several randomized decision trees and aggregates their predictions by averaging, has shown excellent performance in settings where the number of variables is much larger than the number of observations.

III. Existing System

The existing system employs various regression techniques, including linear regression, ridge regression, and lasso regression, for predicting vehicle fuel consumption. These regression methods are commonly used in data analysis and provide insights into the relationships between independent variables (vehicle attributes) and the dependent variable (fuel consumption).The existing system employs various regression techniques, including linear regression, ridge regression, and lasso regression, for predicting vehicle fuel consumption. These regression methods are commonly used in data analysis and provide insights into the relationships between independent variables (vehicle attributes) and the dependent variable (fuel consumption).

Linear Regression: Linear regression is employed to establish a linear relationship between the independent variables and fuel consumption. This method assumes a linear combination of the independent variables to predict the dependent variable. The coefficients of the linear equation are estimated using the least squares method.
.

Ridge Regression: Ridge regression is a regularization technique that extends linear regression by

adding a penalty term to the least squares objective function. This penalty term helps mitigate overfitting by constraining the magnitudes of the regression coefficients. Ridge regression is particularly useful when dealing with multicollinearity among the independent variables.

Lasso Regression: Lasso regression, similar to ridge regression, introduces a penalty term to the least squares objective function. However, lasso regression uses the L1 norm penalty, which encourages sparsity in the coefficient estimates by shrinking some coefficients to zero. This property of lasso regression facilitates feature selection and can identify the most influential predictors of fuel consumption.

While linear regression, ridge regression, and lasso regression offer valuable insights into fuel consumption prediction, they have limitations. These include:

- Assumption of Linearity: Linear regression assumes a linear relationship between the independent and dependent variables, which may not always hold true in practice.
- Sensitivity to Outliers: Linear regression models are sensitive to outliers, which can distort the estimated coefficients and affect prediction accuracy.
- Model Complexity: Ridge and lasso regression introduce additional hyperparameters that need to be tuned, which adds complexity to model selection and interpretation.

IV. PROPOSED SYSTEM

A. *Objectives*

- To understand the relationship between various vehicle attributes and fuel consumption.
- To develop accurate predictive models for fuel consumption using machine learning techniques.
- To optimize the performance of these models through hyperparameter tuning.
- To create an interactive user interface using Streamlit that allows users to input vehicle specifications and obtain real-time predictions of fuel consumption.
- To provide insights and recommendations for improving fuel efficiency based on the analysis.

B. *Approach*

- Data Collection :

First, vehicle data such as vehicle class, engine size, number of cylinders, type of transmission, CO2 rating and fuel type were gathered by the system from the user.

Outlier Analysis: Detecting and handling outliers that could skew the results.

Ordinal Encoding: Converting categorical features such as transmission type and fuel type into numerical values.

The training dataset, is used to train the machine learning models by fitting them to the known outcomes. The test dataset, is used to evaluate the performance.

Feature Scaling: Normalizing the data to ensure that all features contribute equally to the model training process.

- Training The Model :

The core of the project involves building and evaluating machine learning models. Decision Tree and Random Forest algorithms are chosen for their ability to handle complex relationships and interactions between features. Splitting the dataset into training and test sets, and training the models on the training data.

- Model Evaluation :

Evaluating the models using metrics such as Mean Absolute Error(MAE), Mean Squared Error, to ensure robust and accurate predictions. Optimizing the models by fine-tuning hyperparameters to achieve the best performance.

Fig. Architecture diagram of BRIEFIFY

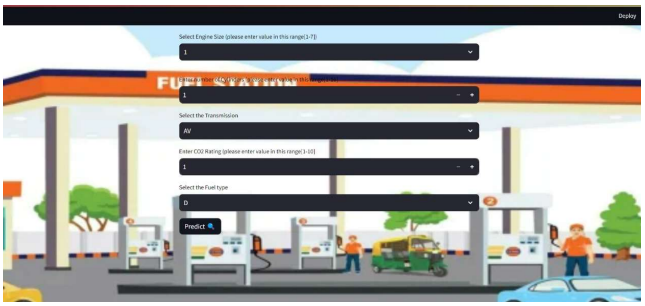
V . WORKING

- Data PreProcessing :

Data preprocessing involves cleaning and transforming the raw dataset into a suitable format for analysis. This includes handling missing values, encoding categorical variables, and scaling numerical features.

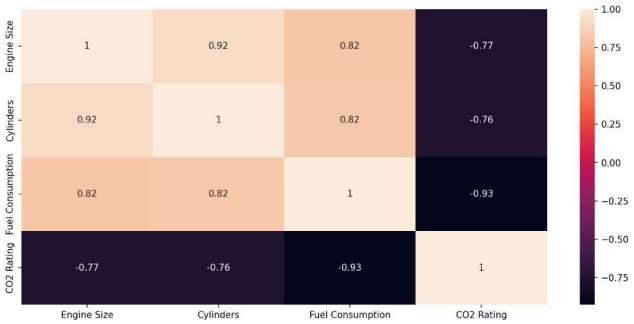
A. Data Collection

- The users need to specify the details of their vehicle, such as vehicle class, engine size, number of cylinders, transmission type, CO2 rating, and fuel type. These inputs were collected from the user through a simple user interface, ensuring that they can easily enter the necessary data.
- The entered data is verified by the system to make sure it is accurate and complete. This includes verifying that numerical values are within reasonable ranges and that categorical values are valid. Any missing or incorrect data is flagged for user correction, ensuring high-quality inputs for the predictive model.
- Once the user inputs and validates the vehicle specifications, the data is sent to the backend system. The collected data is then preprocessed to match the format required by the predictive models.



B. Data Analysis

- A correlation heatmap is generated to visualize the strength and direction of the relationships between all pairs of features. This heatmap uses color gradients to represent correlation coefficients, with values ranging from -1 to 1.
- Strong correlations (both positive and negative) can be easily identified, highlighting which features are most closely related to each other and to the target variable (fuel consumption). This information is crucial for feature selection and understanding multicollinearity, guiding the modeling process by focusing on the most impactful variables.



C. Decision Tree :

Decision trees are intuitive and powerful models that work by recursively splitting the dataset into subsets based on the value of input features. The goal is to create the most homogeneous subsets possible with respect to the target variable, which is the fuel consumption. The steps in building Decision Tree,

1. Root Node Selection: The process begins with selecting the feature that best splits the data. This is determined by calculating a metric such as Gini impurity or information gain for each feature and choosing the one that results in the highest purity (or lowest impurity).

2. Recursive Splitting: The chosen feature creates a root node, and the dataset is split into subsets. This process is repeated recursively for each subset, creating branches of the tree. At each step, the best feature for splitting the subset is chosen based on the same impurity criterion.

3. Stopping Criteria: The recursion continues until one of the stopping criteria is met, such as a maximum tree depth, a minimum number of samples in a node, or a node reaching a minimum impurity threshold.

D. Random Forest :

Random Forest is an ensemble learning method that improves the prediction accuracy of a single decision tree by aggregating the results of multiple trees. This algorithm is used to predict vehicle fuel consumption based on a range of vehicle attributes.

1. Bootstrap Sampling: Random Forest begins by creating multiple bootstrap samples from the training dataset. Each sample is created by randomly selecting data points with replacement, making sure that the samples are diverse.

2. Training Individual Trees: For each bootstrap sample, a decision tree is trained. During the training of each tree, a random subset of features is selected at each split. This randomness helps to reduce correlation among the trees and promotes the creation of diverse models.

3. Tree Depth and Splitting Criteria: Each tree in the forest grown to a specific depth, which is determined by factors such the maximum tree depth. Splits are created using common splitting criteria, such as information gain or Gini impurity.

VI . CONCLUSION

Our project has successfully developed and implemented machine learning models to predict vehicle fuel consumption with a high accuracy. By using advanced machine learning algorithms such as Decision Tree and Random Forest, we have been able to analyze and understand the complex relationships between various vehicle attributes and fuel efficiency. Through comprehensive data analysis, including univariate, bivariate, and correlation heatmap analysis, we have found the factors which influence fuel consumption.

Accurate predictions of fuel consumption have been made by the predictive models, which were trained and validated using a combination of training and test datasets. We increased efficiency and generalization to new data by optimizing the models using methods like feature scaling and hyperparameter tuning. The Streamlit user interface, which is implemented in Python, allows users to enter details of vehicles easily.

This project accurately predicts fuel consumption using advanced machine learning models, enhancing decision-making for consumers, manufacturers, and policymakers. By identifying key factors affecting fuel efficiency, it helps in designing more efficient vehicles and promotes environmental sustainability. Overall, it combines data analysis with practical application, supporting sustainable transportation solutions.

REFERENCES

[1] L. Shalini, S. Naveen and U. M. Ashwinkumar, "Prediction of Automobile MPG using Optimization Techniques," 2021 IEEE Madras Section Conference (MASCON), Chennai, India, 2021, pp. 1-6, doi: 10.1109/MASCON51689.2021.9563597.

[2] P. R, A. Choudhary, P. Jain and O. Kajave, "Vehicle Efficiency Prediction using Machine Learning Algorithms," 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), Trichy, India, 2023, pp. 392-399, doi: 10.1109/ICSMDI57622.2023.00076.

[3] M. Chaitanya Lahari, D. H. Ravi and R. Bharathi, "Fuel Price Prediction Using RNN," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India, 2018, pp. 1510-1514, doi: 10.1109/ICACCI.2018.8554642.

[4] S. Ramesh, S. S. I. M and J. J. Justus, "CO2 Emission Rating by Vehicles using Supervised Algorithms," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-8, doi: 10.1109/ACCAI58221.2023.10200707.

[5] Doruk, Alpay, and Muhammed Ali Bayram. "Predicting Vehicle Fuel Efficiency: A Comparative Analysis of Machine Learning Models on the Auto MPG Dataset." (2023).

[6] P. Ping, W. Qin, Y. Xu, C. Miyajima and K. Takeda, "Impact of Driver Behavior on Fuel Consumption: Classification, Evaluation and Prediction Using Machine Learning," in IEEE Access, vol. 7, pp. 78515-78532, 2019, doi: 10.1109/ACCESS.2019.2920489.

[7] Elalem, A., EL-Bourawi, M.S. Reduction of Automobile Carbon Dioxide Emissions. Int J Mater Form 3 (Suppl 1), 663–666 (2010). https://doi.org/10.1007/s12289-010-0857-2

[8] Harsh H. Patel, Purvi Prajapati, “Study and Analysis of Decision Tree Based Classification Algorithms,” International Journal of Computer Sciences and Engineering, Vol.6, Issue.10,pp.74-78,2018.https://doi.org/10.26438/ijcse/v6i10.7478

[9] Ensemble Machine Learning: Methods and Applications (pp.157-176)Chapter: 5 Publisher: SpringerEditors: Cha Zhang, Yunqian Ma.

[10]Biau, G., Scornet, E. A random forest guided tour. TEST 25, 197–227 (2016). https://doi.org/10.1007/s11749-016-0481-7