FUEL CONSUMPTION PREDICTION

*A report submitted in partial fulfillment of the requirements for the Award of Degree of*

**BACHELOR OF TECHNOLOGY**

in

**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

By

**SATHI VENKATA GOWTHAM REDDY**

**Reg. No: 23B91A61G3**

Submitted to

**Sri Hemalatha Dendukuri**

(Assistant Professor) &

**Dr. Chigurupati Ravi Swaroop**

(Assistant Professor)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**S.R.K.R. ENGINEERING COLLEGE(A)**

SRKR MARG, CHINNA AMIRAM, BHIMAVARAM-534204, A.P

(Recognized by A.I.C.T.E New Delhi) (Accredited by NBA & NAAC) (Affiliated to JNTU, KAKINADA)

SAGI RAMA KRISHNAM RAJU ENGINEERING COLLEGE

(Autonomous)

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**



**CERTIFICATE**

This is to certify that the **ML-CONNECT** Report titled **“FUEL CONSUMPTION PREDICTION”** is the bonafide work done by **Mr. SATHI VENKATA GOWTHAM REDDY** bearing **Register Number: 23B91A61G3** in the second year first semester at **SRKR Engineering College, Bhimavaram** from 8th January 2025 to 9th January 2025 in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering.

| **Dr. Bh.V.S.R.K.Raju** |  | **Dr. Chigurupati Ravi Swaroop** |  | **Hemalatha Dendukuri** |
| --- | --- | --- | --- | --- |
| **(Professor and Head)** |  | **(Assistant Professor)** |  | **(Assistant Professor)** |

**TABLE OF CONTENTS**

1.0 Introduction

1.1 What are the different types of Machine Learning?

1.2 Benefits of using Machine Learning in Employee Salary Prediction

2.0 Employee Salary Prediction

3.0 AI / ML Modelling and Results

3.1 Problem Statement – Employee Salary Prediction

3.2 Data Science Project Life Cycle

3.2.1 Data Exploratory Analysis

3.2.2 Data Pre-processing

3.2.2.1 Managing Duplicate and Low Variation Data

3.2.2.2 Identifying and Addressing Missing Variables

3.2.2.3 Handling Outliers

3.2.2.4 Categorical data and Encoding Techniques

3.2.2.5 Feature Scaling

3.2.3 Selection of Dependent and Independent Variables

3.2.4 Data Balancing

3.2.5 Models used for Development

3.2.5.1 Logistic Regression

3.2.5.2 Decision Tree Model

3.2.5.3 Random Forest Model

3.2.5.4 Extra Trees Classification Model

3.2.5.5 K- Nearest Neighbors Model

3.2.5.6 Light Gradient Boosting Model

3.2.5.7 Gaussian Naive Bayes Model

3.2.5.8 Bagging classifier

3.2.5.9 Gradient Boosting classifier

3.3 AI / ML Models Analysis and Final Results

3.3.1 Combined Code for all Models

3.3.2 Extra Tree Classifier using GridSearchCV– Code

3.3.3 Random Forest using GridSearchCV– Code

3.3.4 Light Gradient Boosting Classifier using GridSearchCV– Code

4.0 Conclusions and Future Work

5.0 References

6.0 Appendices

**ABSTRACT**

This project focuses on predicting fuel consumption for vehicles based on several key factors such as vehicle brand, model, engine size, weight, fuel type, and age. The prediction of fuel consumption is vital for optimizing vehicle usage, reducing operational costs, and enhancing fuel efficiency, benefiting both consumers and businesses. By accurately forecasting fuel consumption, consumers can make more informed purchasing decisions, choosing vehicles that offer better fuel economy. For businesses operating fleets, this model can assist in minimizing fuel costs, optimizing vehicle deployment, and planning preventive maintenance to extend vehicle life. Additionally, it can contribute to sustainability efforts by reducing fuel consumption and lowering emissions.

The project employs machine learning techniques to develop a predictive model, leveraging historical data to generate insights that lead to more effective decision-making and resource management in both individual and business contexts The project begins with a comprehensive data exploration, addressing missing values and ensuring data quality. Exploratory Data Analysis (EDA) uncovers patterns and relationships, setting the stage for insightful feature engineering. Key variables are extracted, and categorical features are encoded to prepare the data for model training.

Multiple regression models, including linear regression, decision trees and random forest, are implemented and fine-tuned to predict fuel consumption. Evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared provide a comprehensive assessment of model performance.

1. **Introduction**

Fuel consumption prediction plays a crucial role in modern transportation, with widespread applications in both consumer and business contexts. As the global demand for fuel-efficient vehicles increases, predicting fuel consumption accurately has become an essential tool for making informed decisions about vehicle selection, usage, and maintenance. This project aims to develop a predictive model for fuel consumption based on a range of factors, including vehicle brand, model, engine size, weight, fuel type, and the age of the vehicle. These variables, individually and in combination, significantly impact the fuel efficiency of a vehicle.Accurately predicting fuel consumption is vital for optimizing vehicle operations and minimizing fuel costs. For consumers, such predictions can guide purchasing decisions, helping them choose vehicles that align with their fuel efficiency needs and reduce long-term operational costs. For businesses, particularly those managing fleets of vehicles, predicting fuel consumption is integral to reducing overall operational expenses, scheduling maintenance more effectively, and improving sustainability by minimizing the environmental impact of fuel usage. By incorporating machine learning techniques, this project seeks to provide a reliable tool for understanding and forecasting fuel consumption, ultimately benefiting both individual consumers and fleet operators.

The potential impact of this model extends beyond cost savings. It offers the opportunity to foster sustainable practices by encouraging the use of more fuel-efficient vehicles, thereby reducing carbon emissions and contributing to a cleaner environment. By empowering consumers and businesses with accurate, data-driven insights into fuel consumption, this project aims to drive better decision-making and enhance the overall efficiency of vehicle fleets across diverse industries.

**1.1 What are the different types of Machine Learning?**

Machine Learning is a subset of Artificial Intelligence (AI). Basically, machine learning automates the prediction of data by learning the trends of data and improving performance.

Machine Learning comprises of set of algorithms which has their own unique way of learning the data and predicting the outcomes. This Machine Learning is then classified into four types.

* + 1. **Supervised Machine Learning**

Supervised machine learning is based on supervision in which we are subjected to train labelled dataset. The main goal of the supervised learning it to map the input variables with the output variables and provide an optimal prediction.

The following are the two types of problems in supervised learning:

1. Regression
2. Classification
   * 1. **Unsupervised Machine Learning**

In unsupervised machine learning the machine interprets the unlabeled dataset and trains and predicts the output without supervision.

In this type of machine learning the model is trained with the data which is neither classified nor labelled and groups the unsorted dataset based on their patterns and similarities with differences under consideration.

Unsupervised learning is again classified into the following types

1. Clustering
2. Association
   * 1. **Semi-Supervised Machine Learning**

Semi-Supervised machine learning is one of the machine learning algorithms which lies between Supervised and Unsupervised machine learning. To overcome the complications and drawbacks of Supervised and unsupervised machine learning can be overcome by using Semi-supervised machine learning technique.

**1.1.4 Reinforcement Machine Learning**

Reinforcement learning is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation.

Reinforcement learning is all about making decisions sequentially. In simple words, we can say that the output depends on the state of the current input and the next input depends on the output of the previous input

* 1. **Benefits of using Machine Learning in Fuel Consumption Prediction.**

Machine learning (ML) offers significant advantages when applied to the task of predicting fuel consumption for vehicles. Leveraging advanced algorithms and data analysis techniques, ML enables the creation of accurate, efficient, and adaptive models that can provide valuable insights and improve decision-making in both individual and business contexts

Machine learning models can learn from large datasets with complex relationships between various factors that influence fuel consumption, such as vehicle brand, engine size, weight, fuel type, and age. By analyzing historical data, ML algorithms can detect patterns and make precise predictions, leading to more accurate fuel consumption forecasts compared to traditional methods or simple statistical approaches.

Fuel consumption is influenced by a variety of complex and non-linear factors. Machine learning algorithms, especially advanced models like neural networks, decision trees, and random forests, are well-equipped to capture these intricate relationships. For example, the interaction between engine size and vehicle weight may not be linear, but machine learning models can effectively learn and account for such non-linear relationships.ML applications in salary prediction also extend to identifying discrepancies and biases in pay structures that might otherwise go unnoticed. This can include anything from identifying gender pay gaps to analyzing regional differences. IBM Watson Analytics is an excellent example of how integrating machine learning with organizational data can aid in creating transparent and equitable salary frameworks.

The use of machine learning in fuel consumption prediction enables more accurate, scalable, and adaptive solutions compared to traditional methods. By leveraging data-driven insights, consumers and businesses can optimize fuel efficiency, reduce costs, enhance sustainability, and make informed decisions that improve both economic and environmental outcomes. Machine learning offers a powerful tool for tackling the growing demand for fuel-efficient solutions in today’s world.

**2.0 Fuel Consumption Prediction**

**1. Data Collection:**

Gather a dataset containing information about vehicles and their fuel consumption. This dataset should ideally have features like Brand, model, engine size, weight, fuel type and age etc.

**2. Data Preprocessing:**

Clean the data by handling missing values, encoding categorical variables (like vechicle id or vechicle name), and scaling numerical features.

**3. Feature Selection/Engineering:**

Identify the most relevant features that affect fuel consumption. This could include features such as vehicle brand, model, engine size, weight, fuel type, age, and driving conditions. You may also want to engineer new features by creating dummy variables for categorical data or deriving new features from existing ones.

**4. Splitting the Data:**

Divide the dataset into training and testing sets. The training set will be used to train the model, while the testing set will help evaluate the model's performance. This ensures that the model generalizes well to new, unseen data and prevents overfitting to the training data.

**5. Model Selection and Training:**

Choose a suitable machine learning algorithm for regression tasks, as fuel consumption is a continuous variable. Algorithms like Linear Regression, Random Forest Regressor, Gradient Boosting, or Neural Networks are commonly used for such tasks. Train the model using the training dataset, where the algorithm learns patterns and relationships between input features (e.g., vehicle characteristics, driving conditions) and fuel consumption.

**6. Model Evaluation:**

Evaluate the model's performance on the testing dataset using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or R-squared score. These metrics help you understand how well the model predicts fuel consumption and its ability to generalize to new data.

**7. Hyperparameter Tuning (Optional):**

Optimize the model's performance by tuning its hyperparameters. This involves adjusting parameters that affect the learning process but are not directly learned from the data. Hyperparameter tuning can significantly improve the accuracy of the model.

**8. Prediction:**

Once the model is trained and evaluated, use it to predict fuel consumption for new vehicle data. The model will take input features (e.g., vehicle type, engine size, age, and driving conditions) and provide a fuel consumption prediction, helping consumers and businesses make informed decisions about vehicle usage and fuel efficiency

**3.0 AI / ML Modelling and Results**

**3.1 Problem Statement – Fuel Consumption Prediction**

Predicting fuel consumption using machine learning involves understanding the various factors that affect fuel usage, such as vehicle type, engine specifications, driving patterns, terrain, and environmental conditions. Here's a breakdown of steps to address this problem:

1. **Data Collection:**Gather a dataset containing information about vehicles, fuel usage, and related factors. Some of the features to collect may include:

* Brand Model
* Engine\_Size
* Fuel\_Type
* Weight
* Mileage
* Fuel\_Consumption
* Age
* Transmission

1. **Data Preprocessing:**Clean the data by handling missing values, identifying and addressing outliers, encoding categorical variables, and scaling numerical features if needed (e.g., normalizing speed, distance, etc.).
2. **Feature Engineering:**Create new features if necessary. For example:

* Converting categorical variables like vehicle type or fuel type into numerical form using one-hot encoding or label encoding.
* Deriving additional features such as fuel consumption rate per distance traveled or average speed during a trip.

1. **Splitting Data:**Divide the dataset into training and testing sets. Typically, a 70-30 or 80-20 split is used to ensure that the model is well-trained and can generalize well on unseen data.
2. **Choosing a Model:**Select appropriate machine learning models for regression, considering algorithms such as:

* Linear Regression
* Decision Trees
* Random Forest Regressors

1. **Training the Model:**Train the chosen models on the training dataset, ensuring that the model learns the relationship between the input features and fuel consumption.
2. **Model Evaluation:**Evaluate the models using appropriate metrics such as:

* Mean Absolute Error (MAE)
* Mean Squared Error (MSE)
* Root Mean Squared Error (RMSE)  
  These metrics will help assess the model’s prediction accuracy and overall performance.

1. **Hyperparameter Tuning:**Optimize the model’s performance by tuning its hyperparameters. This can be done using techniques like GridSearchCV or RandomizedSearchCV to find the best model configuration.
2. **Predictions:**Use the trained model to predict fuel consumption for new or unseen data, such as predicting fuel usage for a future trip based on relevant vehicle and environmental factors.

**10. Model Deployment:** Deploy the model into a production environment, ensuring that it can handle real- time predictions if needed. For example, integrating it into a vehicle fleet management system to estimate fuel usage during daily operations.

**11. Predictions:**

Use the trained model to predict fuel consumption for new or unseen data.

**12. Model Deployment:**

Deploy the fuel consumption prediction model into a production environment,making sure it can handle real-time predictions.

**3.2 Data Science Project Life Cycle**

Data Science is a multidisciplinary field that combines programming skills, domain expertise, and knowledge of statistics and mathematics to extract valuable insights from data. The process of developing a fuel consumption prediction model involves several steps, from data exploration to model evaluation.

**3.2.1 Data Exploratory Analysis**

Exploratory Data Analysis (EDA) is performed to identify relationships and correlations between various variables and their impact on fuel consumption. By analyzing the data, we can understand how features like engine size, vehicle type, driving conditions, and weather influence fuel consumption.

The exploratory analysis for the **fuel consumption prediction** model includes visualizations such as scatter plots, correlation matrices, and distribution graphs to identify significant patterns. The results of this analysis are presented in Appendices 6.1 and 6.2.

**3.2.2 Data Pre-processing**

During data pre-processing, we focus on removing irrelevant variables, handling missing data, and addressing other issues such as outliers. Only the variables that have a direct impact on fuel consumption are retained for analysis and prediction.

**3.2.2.1 Managing Duplicate and Low Variation Data**

In this dataset, duplicates are not present since each record represents individual vehicle or driving data. This has been verified using appropriate checks and coding.

**3.2.2.2 Identifying and Addressing Missing Variables**

Missing data is handled in the following way:

* **Missing fuel consumption values** are filled using the mean value of the existing data.
* For other variables, such as vehicle type, engine size, and driving conditions, missing categorical values are filled with the most frequent value (mode) within each column.

**3.2.2.3 Handling Outliers**

Outliers in fuel consumption data could represent incorrect measurements or extreme cases that don't match the rest of the dataset. After applying boxplots for each relevant variable (e.g., engine size, distance traveled), we observe and remove outliers that could distort the analysis and predictions.

**3.2.2.4 Categorical Data and Encoding Techniques**

Categorical data, such as vehicle type (sedan, SUV, etc.), fuel type (gasoline, diesel, electric), and driving conditions (city, highway, mixed) are converted into numerical values through encoding.

* **Label Encoding** is used to assign a unique numerical value to each category in the dataset.
* **One-Hot Encoding** may also be used if needed, particularly for variables with multiple categories (e.g., different fuel types).

This ensures that categorical data can be processed by machine learning models that require numerical inputs.

**3.2.2.5 Feature Scaling**

Feature scaling normalizes the continuous variables, such as **engine size** and **mileage**, so that they are on the same scale. **MinMaxScaler** is applied to these continuous features to transform their values into the range [0, 1].

**3.2.3 Selection of Dependent and Independent Variables**

* **Dependent Variable**: The target variable for this project is **fuel consumption**, measured in liters per 100 kilometers (L/100 km) or miles per gallon (MPG).
* **Independent Variables**: The independent variables include features such as:
  + **Vehicle Type**
  + **Engine Size**
  + **Fuel Type**
  + **Vehicle Weight**

These variables are selected based on insights from exploratory data analysis and their expected impact on fuel consumption.

**3.2.4 Data Balancing**

Fuel consumption data in this case is well-balanced, with no significant skewness between the different types of vehicles or fuel types. The target variable (fuel consumption) is continuous, so balancing techniques like oversampling or undersampling are not required. However, it's important to ensure the dataset covers a wide range of vehicle types and driving conditions to improve model generalization.

**3.2.5 Models Used for Development**

Various machine learning models are used to predict fuel consumption based on the available features. Below are the models applied to this dataset:

**3.2.5.1 Linear Regression**

Linear regression is used to model the relationship between fuel consumption (dependent variable) and the independent variables (e.g., engine size, driving conditions). It assumes a linear relationship and predicts a continuous value.

**3.2.5.2 Decision Tree Model**

Decision Trees are used to predict fuel consumption by recursively splitting the data based on feature values. Each split in the tree represents a decision based on an important feature, and the leaf nodes represent fuel consumption predictions. This model is easy to interpret but can be prone to overfitting.

**3.2.5.3 Random Forest Model**

Random Forest is an ensemble method that combines multiple decision trees to improve prediction accuracy. It reduces overfitting by averaging predictions from various trees and introducing randomness in the training process. This method is ideal for predicting fuel consumption across different driving scenarios.

**3.2.5.4 Extra Trees Classification Model**

Extra Trees are a variation of Random Forest, where trees are constructed with high randomness in choosing splits. This technique is computationally efficient and can help in making faster predictions for fuel consumption in real-time applications.

**3.2.5.5 K-Nearest Neighbors Model (KNN)**

The K-Nearest Neighbors (KNN) algorithm uses the proximity of similar data points to predict the target variable (fuel consumption). For each prediction, KNN identifies the nearest neighbors in the feature space and takes the average of their fuel consumption values.

**3.2.5.6 Light Gradient Boosting Model (LightGBM)**

LightGBM is a gradient boosting framework that enhances model performance and reduces memory consumption. It works well for large datasets with high dimensionality and can efficiently predict fuel consumption even when dealing with complex relationships between variables.

**3.2.5.7 Gaussian Naive Bayes Model**

Gaussian Naive Bayes is based on Bayes' theorem and assumes that the features are normally distributed. This model is useful for predicting fuel consumption when dealing with a large number of categorical or continuous variables.

**3.2.5.8 Bagging Classifier**

Bagging (Bootstrap Aggregating) uses multiple models (such as decision trees) to reduce variance and improve the accuracy of predictions. This method combines predictions from several models to get a more reliable estimate of fuel consumption.

**3.2.5.9 Gradient Boosting Classifier**

Gradient Boosting builds a strong predictive model by combining several weak learners (decision trees) in a sequential manner. This method is highly effective for predicting fuel consumption based on complex interactions between features.

**3.3 AI / ML Models Analysis and Final Results**

The dataset is split into a training set (80%) and a testing set (20%) to evaluate the performance of each model. After training the models on the training set, the predictions are tested on the testing set.

The **Confusion Matrix** is used to evaluate the accuracy, precision, recall, and F1-score for each model. Additionally, regression metrics like **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **R-squared** are used to measure the model’s predictive performance for fuel consumption.

The model with the highest performance metrics is selected as the best model for predicting fuel consumption and is then ready for deployment.

This process ensures that we can predict **fuel consumption** accurately based on vehicle and driving-related features. The final model can be deployed to provide real-time fuel consumption predictions for users, vehicle fleet managers, or environmental monitoring systems.

**3.3.1 Combined Code for all Models**

#importing the libraries

import pandas as pd

import numpy as np

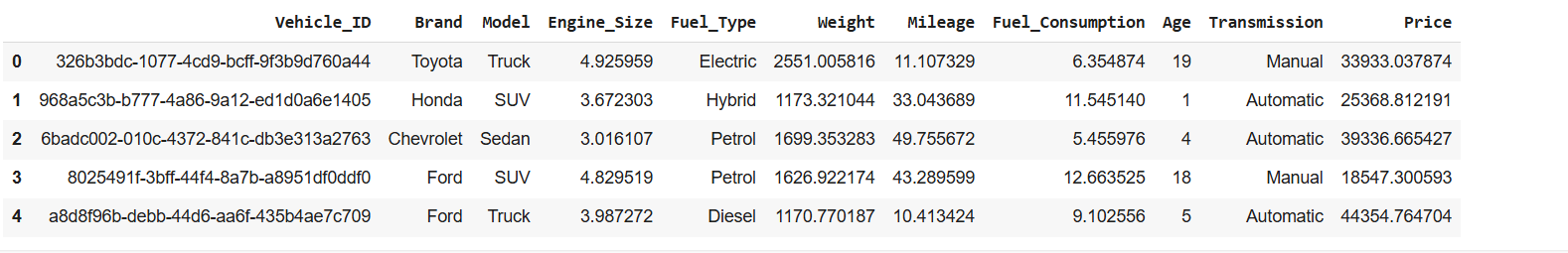
from google.colab import drive

drive.mount('/content/drive')

# Reading dataset

df = pd.read\_csv('/content/drive/MyDrive/dataset/Fuel\_Consumption\_Prediction.csv')

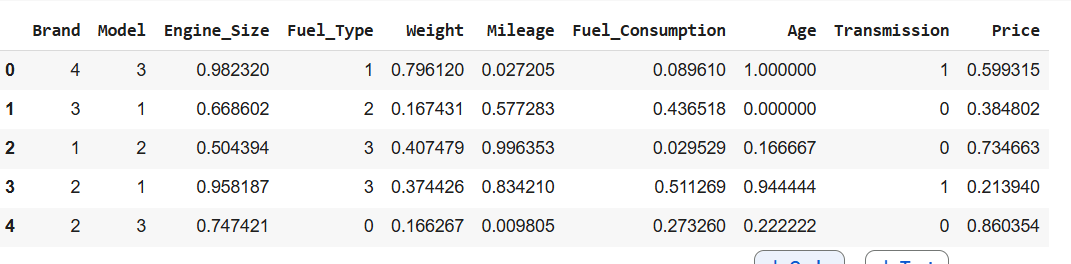
df.head()



# Remove unwanted columns

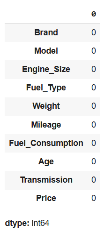
df.drop(columns=['Vehicle\_ID'], inplace=True)

# specify the columns to remove



# find the null values

df.isnull().sum()

****

from sklearn.preprocessing import LabelEncoder

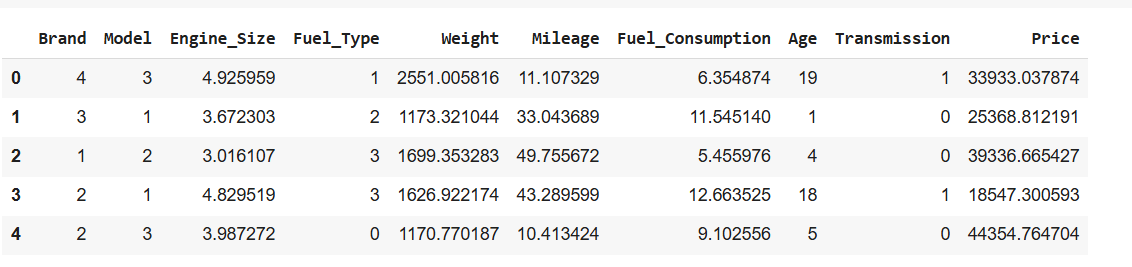
from sklearn.preprocessing import LabelEncoder

lebel\_encoder = LabelEncoder()

for column in cat\_columns:

df[column]=label\_encoder.fit\_transform(df[column])

df.head()



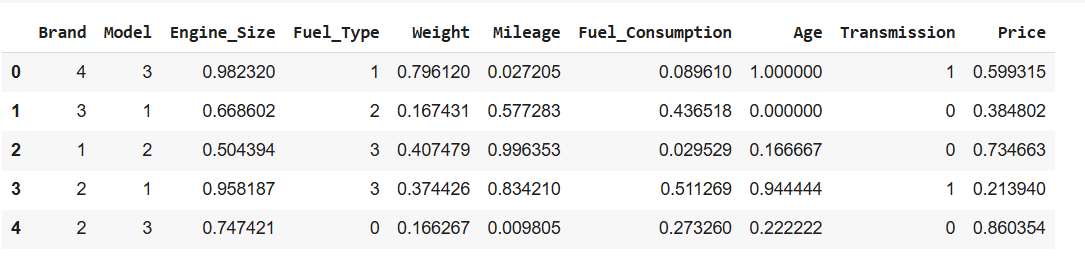
from sklearn.preprocessing import MinMaxScaler

fro, sklearn.preprocessing import MinMaxScaler

scaler=MinMaxScaler()

df[num\_columns]=scaler.fit\_transform(df[num\_columns])

maindata.head()



X=df.drop(‘fuel\_consumption’,axis=1) # --independent variable

y=df[‘fuel\_consumption’] # --dependent variable

#importing ML processes

from sklearn.linear\_model import LinearRegression

from sklearn.tree import DecisionTreeRegressor

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import classification\_report

# Splitting the dataset into train and test set

from sklearn.model\_selection import train\_test\_split as tts

x\_train,x\_test,y\_train,y\_test=tts(X,y,test\_size=0.2,random\_state=42)

# Train Logistic Regression model

from sklearn.linear\_model import LinearRegression

linreg\_model = LinearRegression()

linreg\_model.fit(X\_train, y\_train)

# Evaluate Logistic Regression model

from sklearn.metrics import mean\_squared\_error, mean\_absolute\_error, r2\_score

linreg\_pred = linreg\_model.predict(X\_test)

# Evaluation metrics

linreg\_mse = mean\_squared\_error(y\_test, linreg\_pred)

linreg\_mae = mean\_absolute\_error(y\_test, linreg\_pred)

linreg\_r2 = r2\_score(y\_test, linreg\_pred)

print(f'Mean Squared Error: {linreg\_mse}')

print(f'Mean Absolute Error: {linreg\_mae}')

print(f'R-squared: {linreg\_r2}')

Mean Squared Error: 0.06861866113105708

Mean Absolute Error: 0.21068495202990742

R-squared: -0.04719394401728927

#Train Random Forest Classifier model

from sklearn.ensemble import RandomForestRegressor

rf\_model = RandomForestRegressor()

rf\_model.fit(X\_train, y\_train)

# Evaluate Random Forest Classifier model

rf\_pred = rf\_model.predict(X\_test)

rf\_mse = mean\_squared\_error(y\_test, rf\_pred)

rf\_mae = mean\_absolute\_error(y\_test, rf\_pred)

rf\_r2 = r2\_score(y\_test, rf\_pred)

print(f'Mean Squared Error: {rf\_mse}')

print(f'Mean Absolute Error: {rf\_mae}')

print(f'R-squared: {rf\_r2}')

Mean Squared Error: 0.07785062027268348

Mean Absolute Error: 0.22623333920381214

R-squared: -0.18808348550894904

from sklearn.tree import plot\_tree

import matplotlib.pyplot as plt

# Plotting the first tree in the forest

plot\_tree(rf\_model.estimators\_[0], filled=True)

plt.show

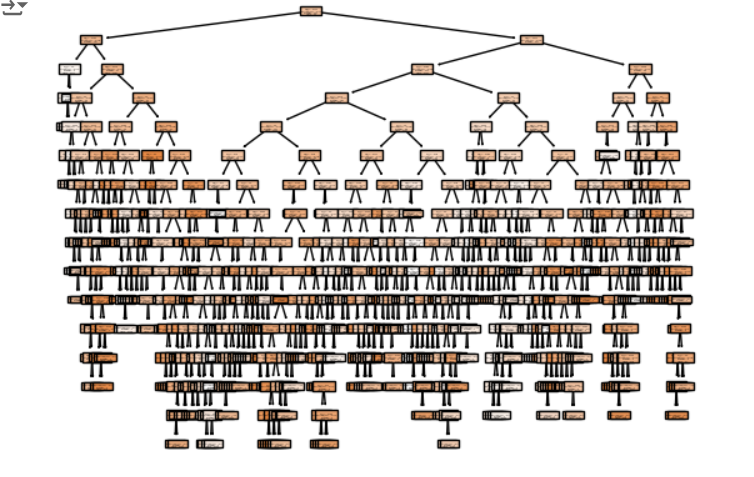
from sklearn.tree import plot\_tree

import matplotlib.pyplot as plt

# Plotting the first tree in the forest

plot\_tree(rf\_model.estimators\_[0], filled=True)

plt.show()



# Save the best model as a .pkl file

import pickle

with open('best\_model.pkl', 'wb') as file:

    pickle.dump(best\_model, file)

print("Best model saved as 'best\_model.pkl'")

Best model saved as 'best\_model.pkl'

**5.0 References**

* Association Of Computer Engineers SRKR

**6.0 Appendices**

**6.1 List of charts**