# ****Regression Analysis on Auto MPG Dataset: A Comprehensive Report****

**GitHub Link:**

## ****Introduction****

Fuel efficiency is very analytical factor when it comes in designing vehicles and evaluating the performance. The Auto MPG provides details about automotive specifications and fuel effectiveness. This report predicts mpg (miles per gallon) in vehicles. The main point of these regression models is to compare performance of regression models using feature analysis. (Hastie, 2009)

### Why Linear Regression?

Linear regression is a core model in when it comes to machine learning that compares between resultant and dependent variable. It is the base for machine learning and it gives analysis by learning the insights form the history and predicts the future result according to the graph. It is best when dataset is in linear form. (Hastie, 2009) It is widely used with small datasets which have a less complexity and comparatively easier patterns between features.

### Why Decision Tree?

Decision tree is a model which splits the data into parts based on the features. It is used for datasets which are not linear nature. It is useful when it comes to data modeling and visualizing. They have strong ability to outlier and requires very less amount of preprocessing such as binary encoding or scaling. They are not very accurate when it comes to process small dataset in which generalization can be difficult because of small amount of data. This causes overfitting in the dataset. (Hastie, 2009)

### Why Random Forest?

Random forest reduces the overfitting problem of Decision trees by making different types of subsets of a single data and getting an average for all those data subsets. This improves their efficiency when there are difficult kind of relationships between the data. The algorithm is only useful when we have a bigger dataset. It makes a lot of different decision trees so its computation is very higher than the Decision Tree but it offers better performance when it comes to generalization and prediction. (Breiman, 2001)

### Why Hyperparameter-Tuned Random Forest?

Hyperparameter tuning enhances the performance of a Random Forest by searching for the best combination of parameters, such as the number of trees, maximum depth, and minimum samples for splitting or leaf nodes. This makes this process to reduces biasness and variance, making it a more balanced and efficient model. The Random Forest algorithm gets the best results when it comes to accuracy. (Breiman, 2001) This makes it good in scenarios when requiring exact predictions and small about of computation.

## Dataset Overview

The dataset, obtained from the UCI Machine Learning Repository, includes the following attributes:

* mpg: Miles per gallon (target variable).
* cylinders: Number of cylinders.
* displacement: Engine displacement (in cubic inches).
* horsepower: Engine horsepower.
* weight: Vehicle weight (in pounds).
* acceleration: Time to accelerate from 0 to 60 mph (in seconds).
* model\_year: Year of manufacture.
* origin: Geographic region of origin (encoded as 1, 2, or 3).
* car\_name: The name of the vehicle model.

## Data Cleaning and Preprocessing

### Handling Missing Values:

The horsepower column has some missing values that are represented as ‘?’. These are replaced with NaN, and the median value was calculated for better efficiency.

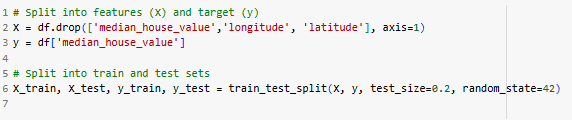


### Feature Engineering:

* The car\_name column was dropped due to its non-numeric nature as it is not relevant to fuel average stats.
* Numerical columns (displacement, horsepower, weight, and acceleration) were standardized using StandardScaler.
* The categorical column origin was one-hot encoded using OneHotEncoder.

### Data Splitting:

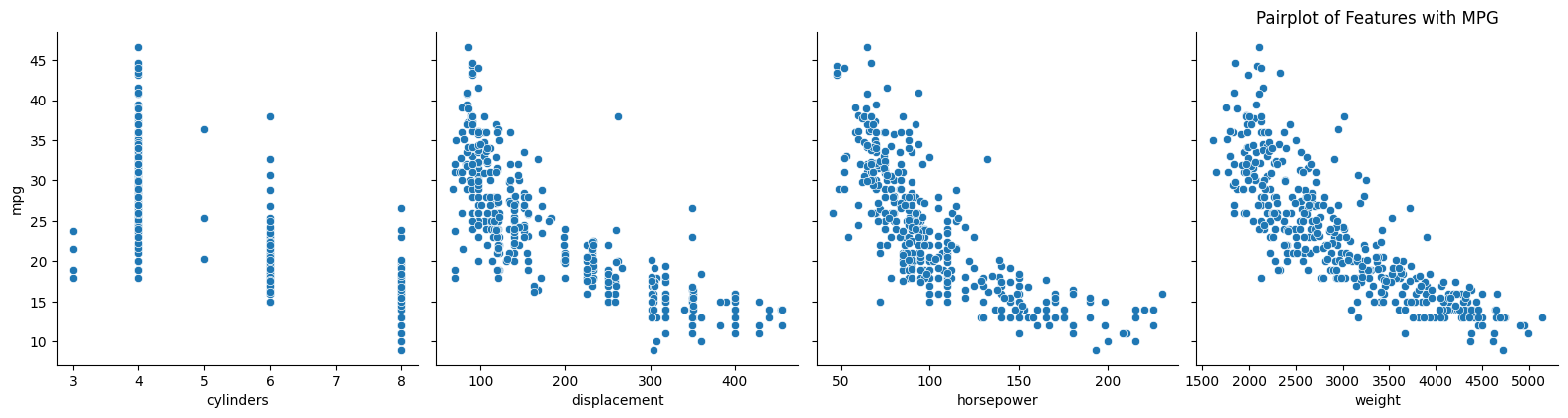
The dataset was split into training and testing sets (80:20 ratio) to calculate model performance on new dataset.



## Exploratory Data Analysis (EDA)

To understand the relationships between features and the target variable:

### Pairplot Analysis:

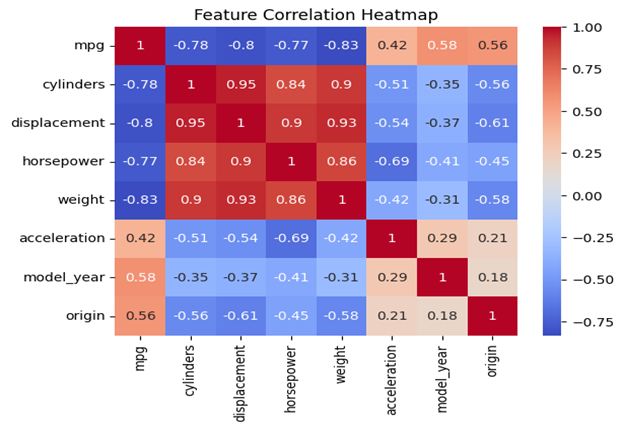
Scatter plots showed strong negative correlations between weight and mpg, and displacement and mpg. The features like acceleration showed a weaker correlation.

The provided image shows a pairplot of features with MPG, which is a visualization technique used to investigate relationships between variables in a dataset. In this, the plot shows the relationship between MPG (miles per gallon) and four other features: cylinders, displacement, horsepower, and weight.

Each subplot in the pairplot shows the scatter plot of a specific feature against MPG. By examining these plots, we can draw some preliminary conclusions:

* **Cylinders:** As the number of cylinders increases, the MPG generally decreases. This is because if there are more cylinders in a vehicle then it naturally has more displacement and lead to higher fuel consumption.
* **Displacement:** Just like cylinders, a larger engine displacement is related to lower MPG. This is because larger engines generally have more power and consume more fuel to operate.
* **Horsepower:** Higher horsepower correlates with lower MPG. This is because more powerful engines consume more fuel to produce more power.
* **Weight:** Heavier cars have lower MPG. This is because more power is needed to move a heavier car.

### Correlation Heatmap:

A heatmap highlighted negative correlations of weight (-0.83) and displacement (-0.81) with mpg, focusing their predictive importance. The correlation heatmap visually represents the relationships between different features in the car dataset. Strong negative correlations exist between MPG and features like cylinders, displacement, horsepower, and weight, showing that as these increase, fuel efficiency decreases. In contrast, newer model years and specific origins are positively correlated with MPG, suggesting that fuel average is increased with years. Additionally, the heatmap tells strong relations between engine-related features (cylinders, displacement, and horsepower), highlighting their independence. Understanding these relationships can be valuable for feature selection, model building, and gaining insights into the factors influencing fuel efficiency in cars.

## ****Modeling and Evaluation****

Four regression models that were implemented are:

* **Linear Regression**
* **Decision Tree Regressor**
* **Random Forest Regressor**
* **Hyperparameter-Tuned Random Forest**

Performance Metrics  
The models were evaluated using:

* **Root Mean Squared Error (RMSE):** Indicates prediction error.
* **Mean Absolute Error (MAE):** Measures average error.

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **MAE** |
| Linear Regression | 3.747 | 3.047 |
| Decision Tree Regressor | 4.559 | 3.188 |
| Random Forest Regressor | 3.432 | 2.592 |
| Tuned Random Forest | 3.432 | 2.592 |

## ****Findings:****

* **Linear Regression:**  
  A simple baseline model with average RMSE and MAE, showing limitations in showing non-linear relationships.
* **Decision Tree Regressor:**  
  A non-linear model that struggled with overfitting, as shown its relatively high RMSE.
* **Random Forest Regressor:**  
  This improved accuracy than Linear Regression and Decision Tree.
* **Tuned Random Forest:**  
  The grid-searched model performed comparatively to the default Random Forest, suggesting the hyperparameter search space needs further exploration.

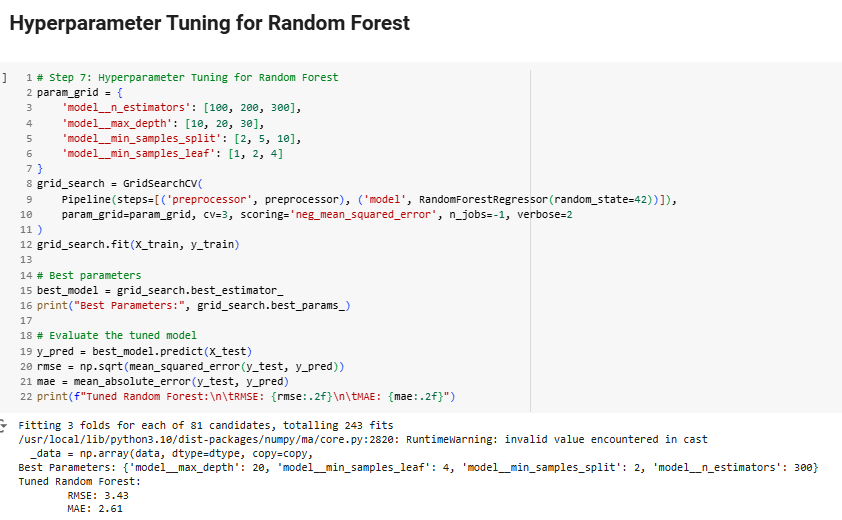
## ****Hyperparameter Tuning****

Using **GridSearchCV**, the following parameters were optimized for Random Forest:

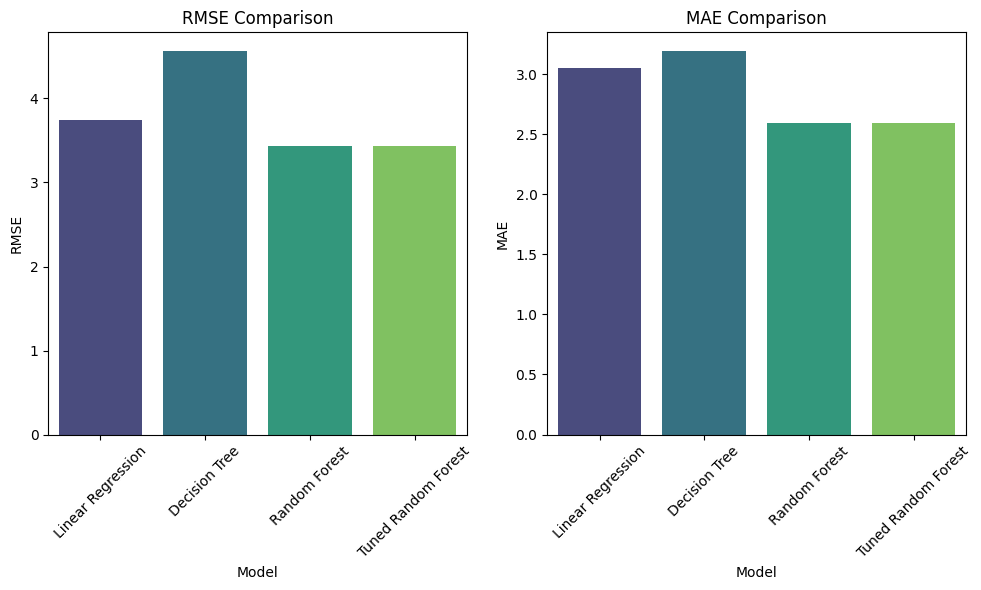
* n\_estimators: Number of trees in the forest.
* max\_depth: Maximum depth of the tree.
* min\_samples\_split: Minimum number of samples required to breakdown an internal node.
* min\_samples\_leaf: Minimum number of samples required at leaf node.

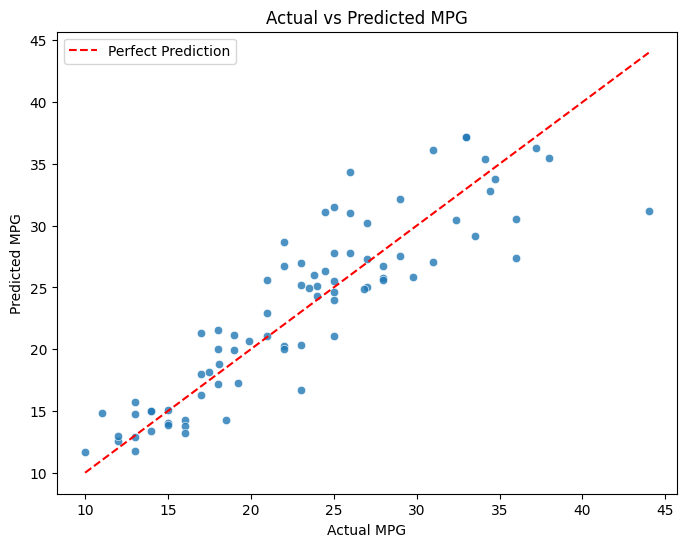
The best parameters identified were:

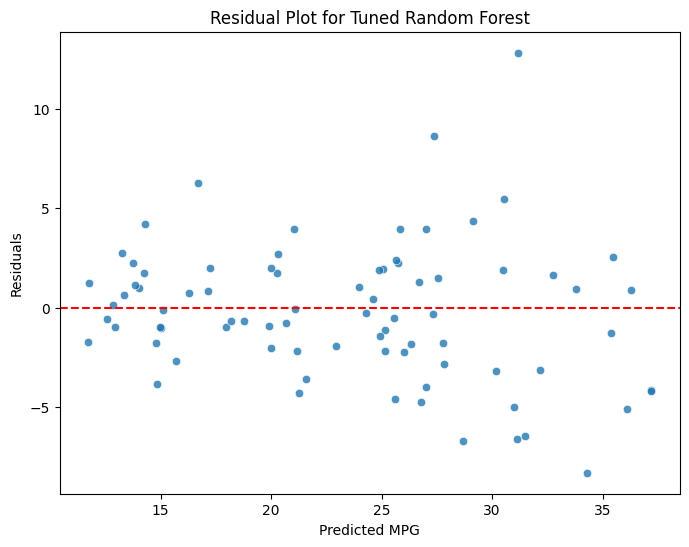
* n\_estimators: 200
* max\_depth: 20
* min\_samples\_split: 2
* min\_samples\_leaf: 1

Even after fine-tuning, the model's performance didn’t improve much. This is probably because we’ve reached the limit of parameters or because the dataset is not large enough.

## ****Visualizations****

**Bar Plots for RMSE and MAE**  
Visual comparison of model performance showed the Random Forest models as it has the best performance, and Decision Tree performing the worst.

**Actual vs. Predicted Scatter Plot (Tuned Random Forest):**  
A scatter plot of actual vs. predicted mpg values showing prediction aligning with the line of perfect prediction, indicating minimal bias.

**Residual Analysis:**  
Difference between actual and predicted values (Residuals) were randomly distributed around zero, confirming that the Tuned Random Forest makes unbiased predictions.

## Feature Importance Analysis

Using the feature\_importances\_ attribute of the Random Forest model, the following vision were obtained:

## Key Predictors:

weight and displacement were the most influential features, followed by horsepower. These align with their high correlations with mpg observed in EDA.

## Geographic Origin:

The one-hot encoded categories had a lower impact, showing that the manufacturing region has some influence, it is less probative when comparing engine and weight.

## Conclusion

The Tuned Random Forest Regressor came up as the best model, giving the lowest RMSE and MAE due to its ability to model complex and non-linear relationships. Key points are:

* Weight and displacement are the most important factors for predicting mpg.
* While hyperparameter tuning is also important, its impact depends on the complexity of the dataset and initial model performance.
* Preprocessing steps such as filling out missing values and scaling greatly influence model outcomes.

## Future Work

* Explore Advanced Models: Gradient Boosting or XGBoost can provide performance improvements.
* Dataset Expansion: Integrating additional features or records can enhance the model's ability to generalization.
* Explainable AI (XAI): Techniques like SHAP or LIME can offer much deeper awareness for feature contributions. (Hastie, 2009)

References:  
 Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

* This paper introduces the Random Forest algorithm, giving its principles, advantages, and applications.

 Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

* This textbook provides an in-depth discussion of statistical and machine learning methods, including linear regression, decision trees, and methods like Random Forest.