IMLA MINI PROJECT

Prediction of Miles Per Gallon

Group 5

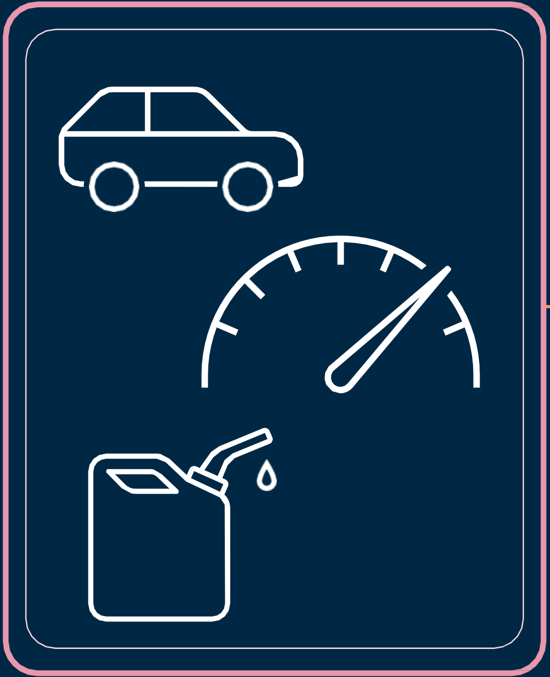
Presented By:
PA15 Vikram Deshmukh
PA36 Palak Mallawat
PA38 Rashi Madnani
PA41 Dhanashree Lodhe

TABLE OF CONTENTS

TABLE OF CONTENTS

PROBLEM STATEMENT

Prediction of Miles per Gallon (mpg) on the basis of different features like displacement, horsepower, origin of a car, weight etc. We need to find which factors mostly affect the fuel consumption of a car in order to improve the mpg value. Hence build a model to predict the mpg value of each car.



DATASET

COLUMN NAME	DATATYPE	DESCRIPTION
CYLINDERS	Int64	contains the number of cylinders present in the car
DISPLACEMENT	Float64	contains the Displacement of the car
HORSEPOWER	Float64	contains the Horsepower of the car
WEIGHT	Float64	contains the weight of the car
ACCLERATION	Float64	contains the Acceleration of the car
MODEL YEAR	Int64	contains the model year of the car
ORIGIN	Int64	contains the origin country which car belong to
CAR NAME	Object	contains the name of the car(Brand-Model-Variant)
MPG	Float64	contains the fuel consumption value(in Miles per Gallon) for car

EXPLORATORY DATA ANALYSIS

CHECK NULL AND DUPLICATE VALUES

1

2

CHECK FOR CATEGORICAL VALUES

CHECK FOR OUTLIERS

3

4

DIMENSION REDUCTION

PREPROCESSING

Assigning X and y values

X: Features
Y: Target (mpg)

Encoding categorical values

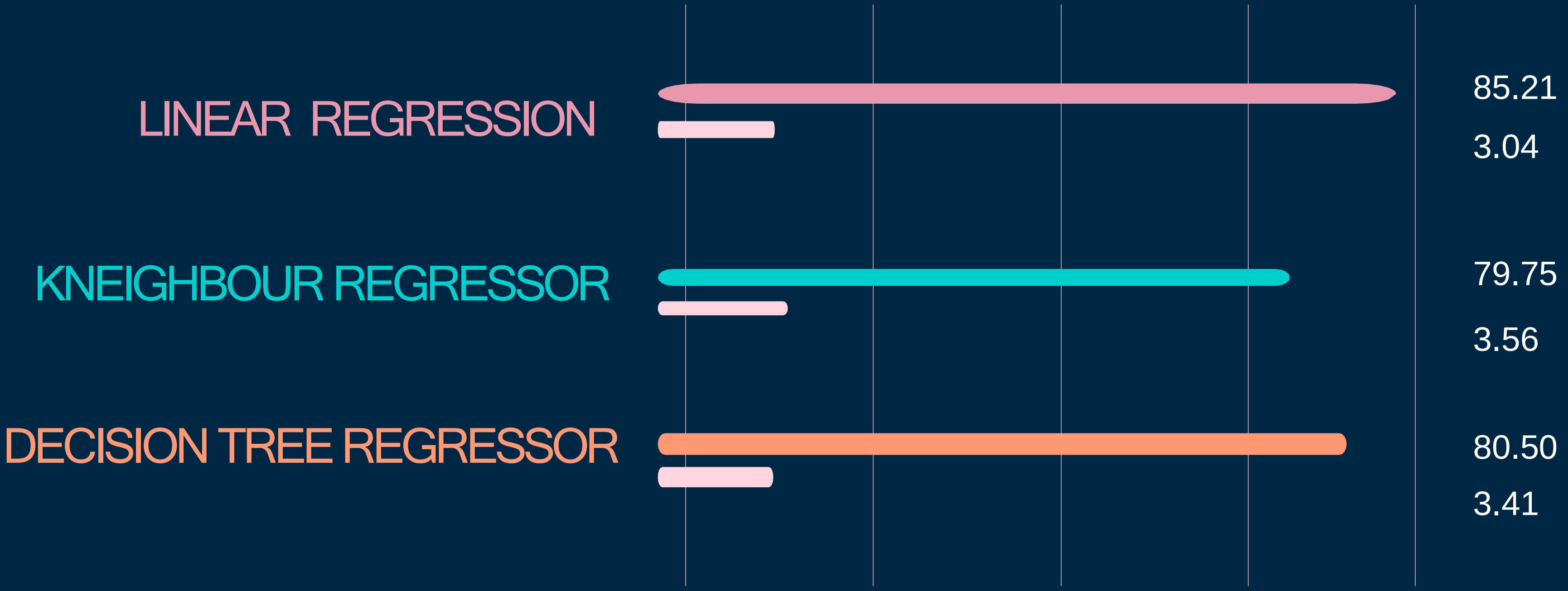
OneHotEncoder is used on cylinder and origin column.
LabelEncoder is used on model_year and brand_name.

Train – Test Split and Feature Scaling

Test-size: 0.3
Train-size: 0.7
StandardScaler is used for Normalisation

MODELS

ML MODELS IMPLEMENTED



LINEAR REGRESSION

Linear Regression is a machine learning algorithm based on supervised learning. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output)

KNEIGHBOUR REGRESSOR

K nearest neighbors is a simple algorithm that stores all available cases and predict the numerical target based on a similarity measure (e.g., distance functions). A simple implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbors. Another approach uses an inverse distance weighted average of the K nearest neighbors. KNN regression uses the same distance functions as KNN classification. We can use the Euclidean or Manhattan distance.

DECISION TREE REGRESSOR

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. It employs a top-down, greedy search and recursive divide and conquer method. The ID3 algorithm can be used to construct a decision tree for regression by replacing Information Gain with Standard Deviation Reduction.

ENSEMBLE

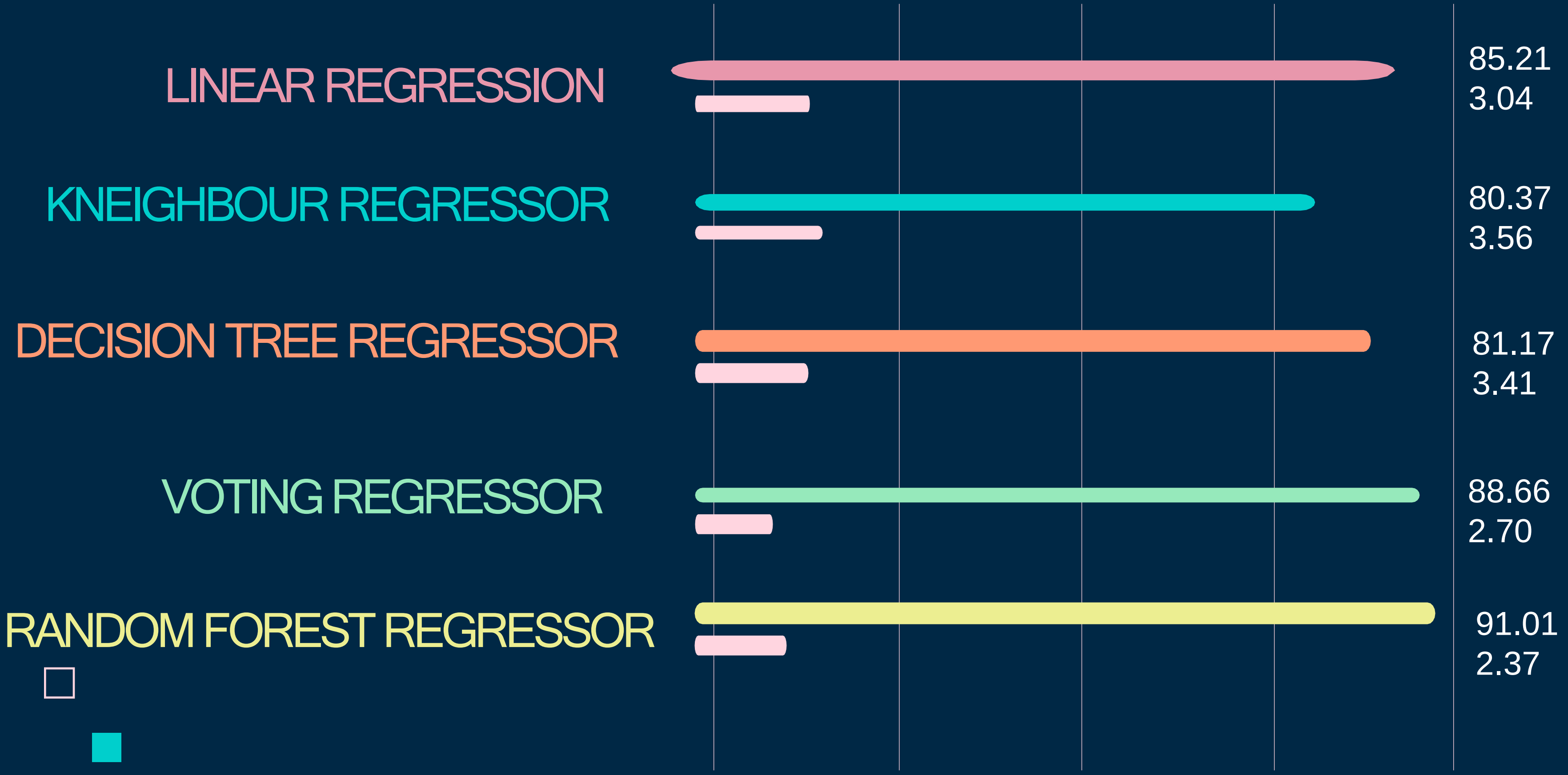
VOTING REGRESSOR

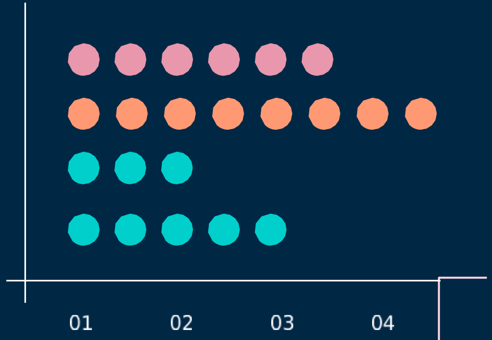
A voting regressor (or a “majority voting ensemble“) is an ensemble machine learning model that combines the predictions from multiple other models. It averages the individual predictions to form a final prediction. It is a technique that may be used to improve model performance, ideally achieving better performance than any single model used in the ensemble. It follows the averaging method of ensemble and is a prime example of bagging.

RANDOM FOREST REGRESSOR

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. In the case of a regression problem, the final output is the mean of all the outputs. This part is Aggregation.

MODEL SCORE/ACCURACY COMPARISON





CLUSTERING

K-MEANS CLUSTERING

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

THANK YOU