

DWDM Mini Project

GUIDED BY : Prof. Vaishali Suryawanshi ma'am

PG 03 Omkar Najan

PRN NO. 1032180002

PG 09 Mayur Pingle

PRN NO. 10321800083

PG 13 Payal Patil

PRN NO. 1032180129

PG 14 Vaishnavi Salunke

PRN NO. 1032180146

Problem Definition

- Predicting a brand of a car using attributes like Miles/Gallon , Cylinders , cubic Inches , HorsePower ,weight(in lbs) , time to reach 60 miles , and year of release .
- Perform different data mining operations , and ML algorithm and find suitable algorithm for the predicting Machine Learning Algorithm.
- ML Algorithms to be used for this categorical data are -
 - Decision Tree Algorithm ,
 - K-Means Algorithm,
 - K-Nearest-Neighbors.

Objectives

1. Select Dataset
2. Preprocess this dataset
3. Perform Data Mining Operations
4. Visualize the results



Tools Used



- ❖ Jupyter Notebook
- ❖ IBM Skill Network Lab
- ❖ Python Libraries:
 - Numpy
 - Matplotlib
 - Pandas
 - Seaborn
 - Sklearn

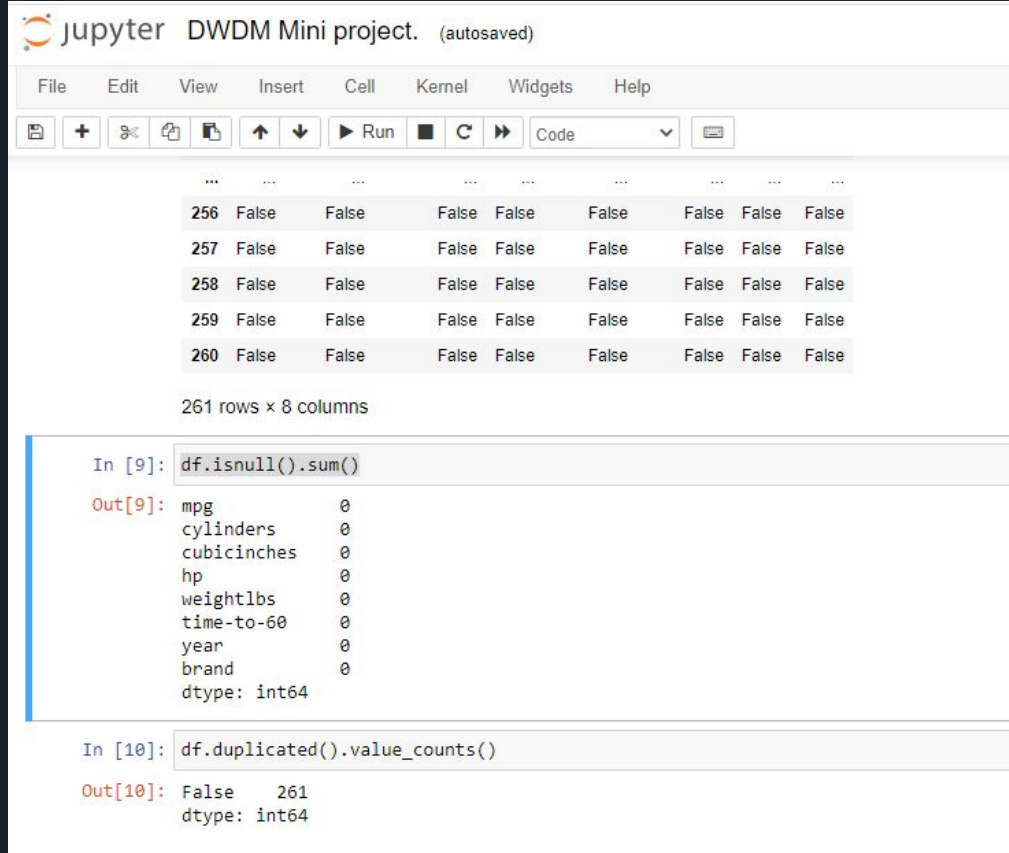


Dataset Description

- The "car.csv" file contains 258 Cars
- Data has Information about 3 brands/make of cars. Namely US, Japan, Europe.
- Target of the data set to find the brand of a car using the parameters such as horsepower, Cubic inches, Make year, etc.

Data Preprocessing

- Null values in the dataset were checked using isnull function. The null or empty values were replaced wherever needed.
- Duplicate Values are checked using ,duplicated() function , As there are no duplicate values so no values are dropped.
- Data types of all dataset are checked if all are int or float , str values are converted into int where possible.
- Categorical values are converted into dummy numerical values using preprocessing LabelEncoder function.



The screenshot shows a Jupyter Notebook titled "DWDM Mini project. (autosaved)". The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for saving, adding cells, undo, redo, and running code. A data preview table is displayed, showing rows 256 to 260 with 8 columns of boolean values (False). Below the table, it indicates "261 rows x 8 columns".

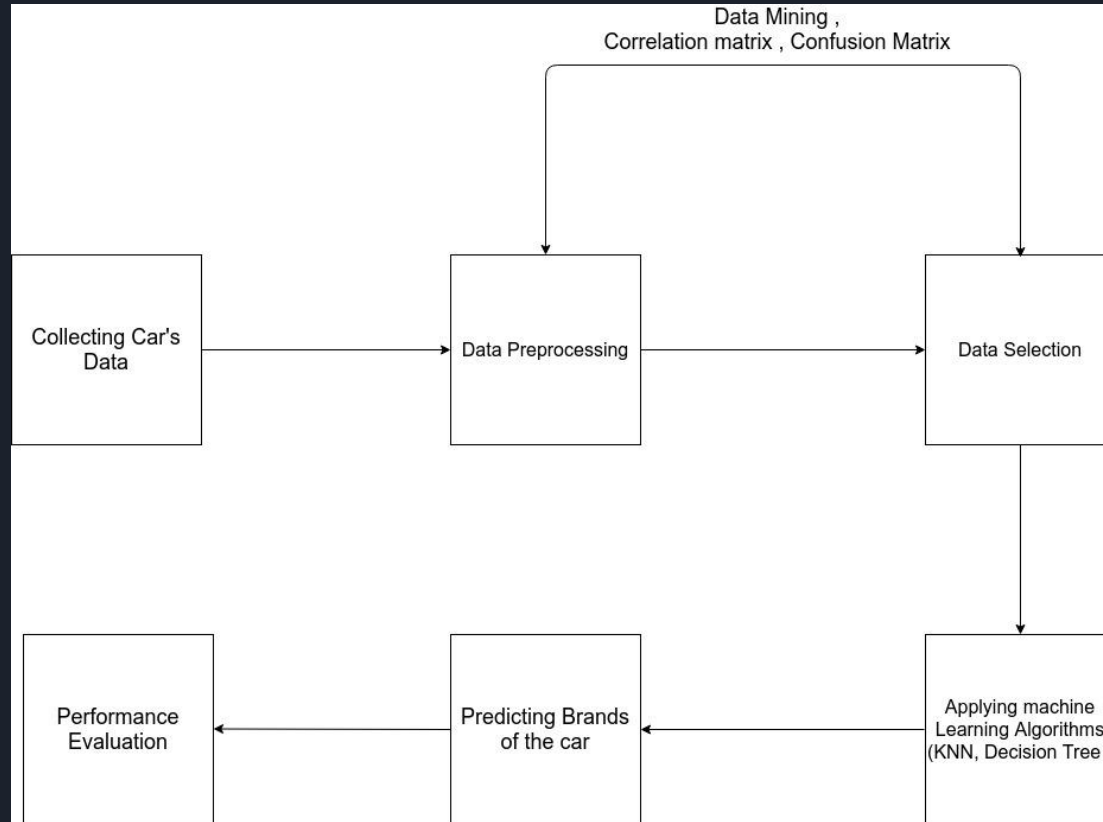
```
In [9]: df.isnull().sum()
Out[9]: mpg          0
cylinders          0
cubicinches        0
hp                 0
weightlbs          0
time-to-60         0
year               0
brand              0
dtype: int64
```

```
In [10]: df.duplicated().value_counts()
Out[10]: False      261
dtype: int64
```

System architecture

Input

output



Data Mining as a whole process

The whole process of Data Mining comprises of three main phases:

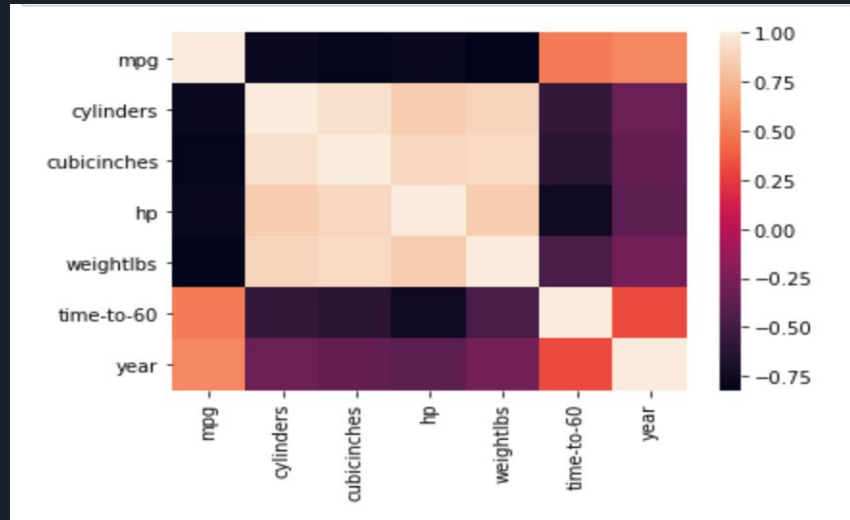
1. Data Pre-processing – Data cleaning, selection and transformation takes place
2. Data Extraction – Occurrence of exact data mining
3. Data Evaluation and Presentation – Analyzing and presenting results

Data Mining algorithms -used

- Decision Tree
- K-means
- The k-nearest neighbors (KNN)

Data Mining Task Performed

The data mining task performed on this project is clustering
The method to perform clustering is K-Means.

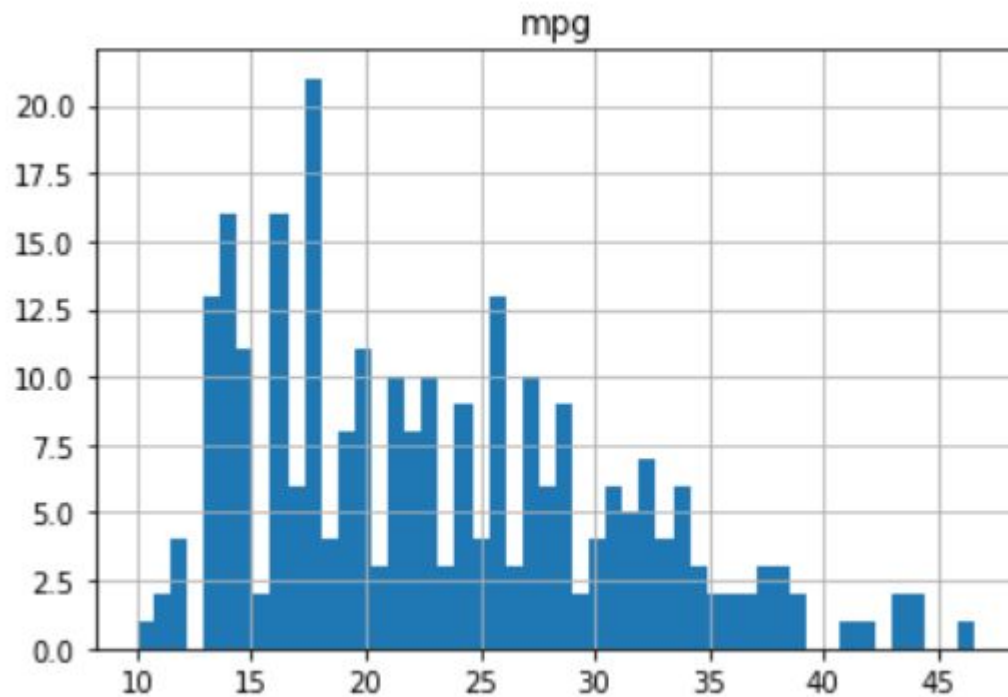


Correlation Heatmap

Decision trees

- It maps out all the decision paths in form of tree
- How to build Decision Tree :
 - Decision Tree are built by splitting training set in distinct nodes. Where one node contains All of / Most of One category of data.
 - Each Internal node corresponds to results of test.
 - Each Leaf node assigns a classification

KNN



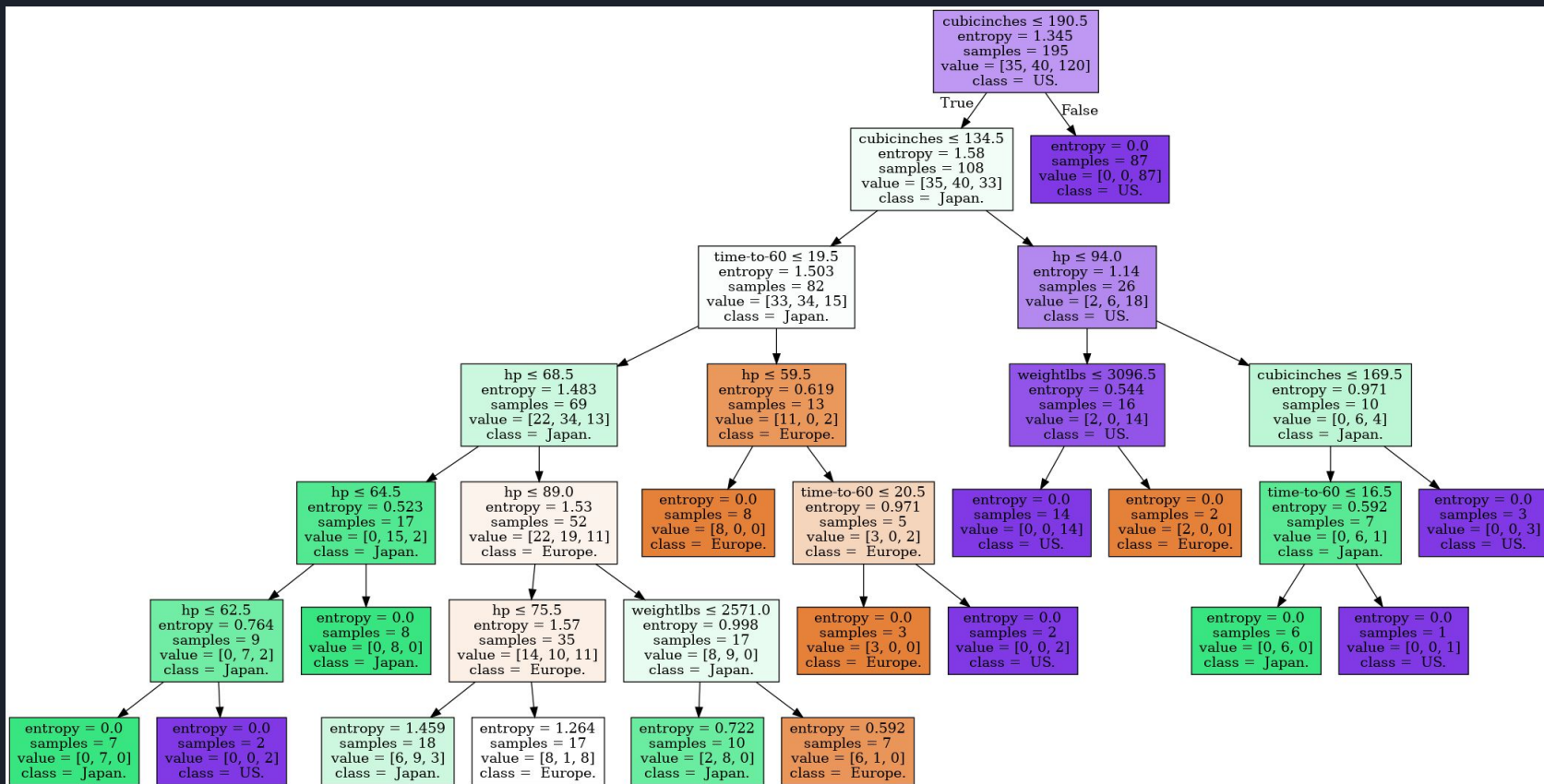
Histogram of mpg (miles per gallon)

The k-nearest neighbors (KNN) algorithm

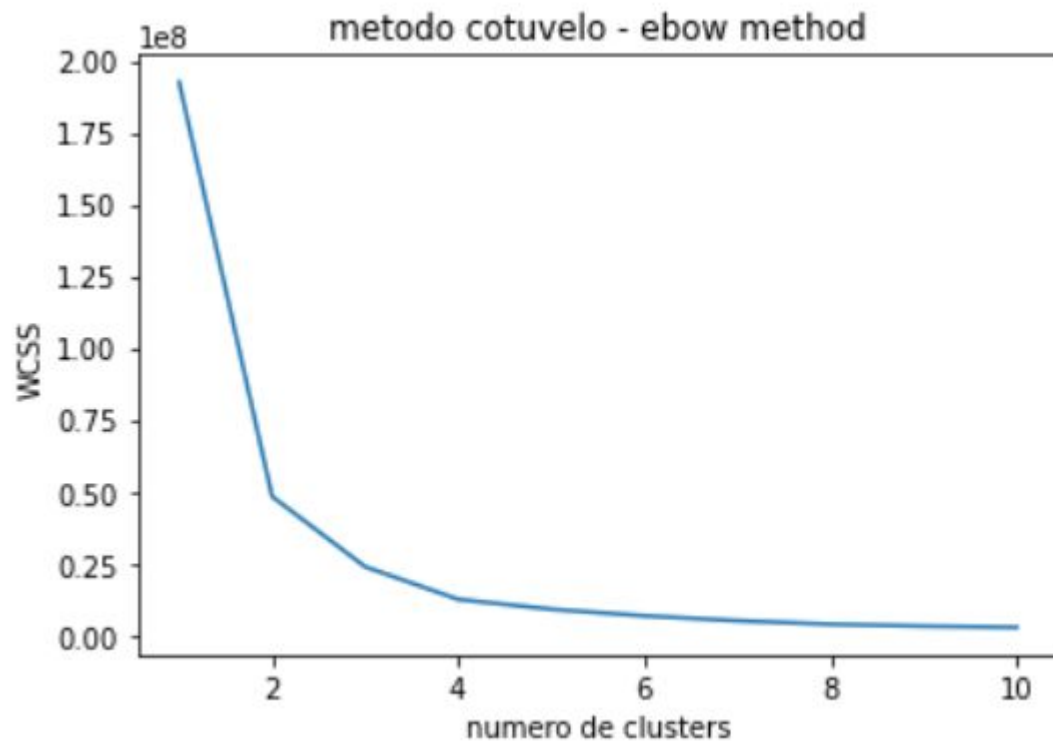
What is KNN:

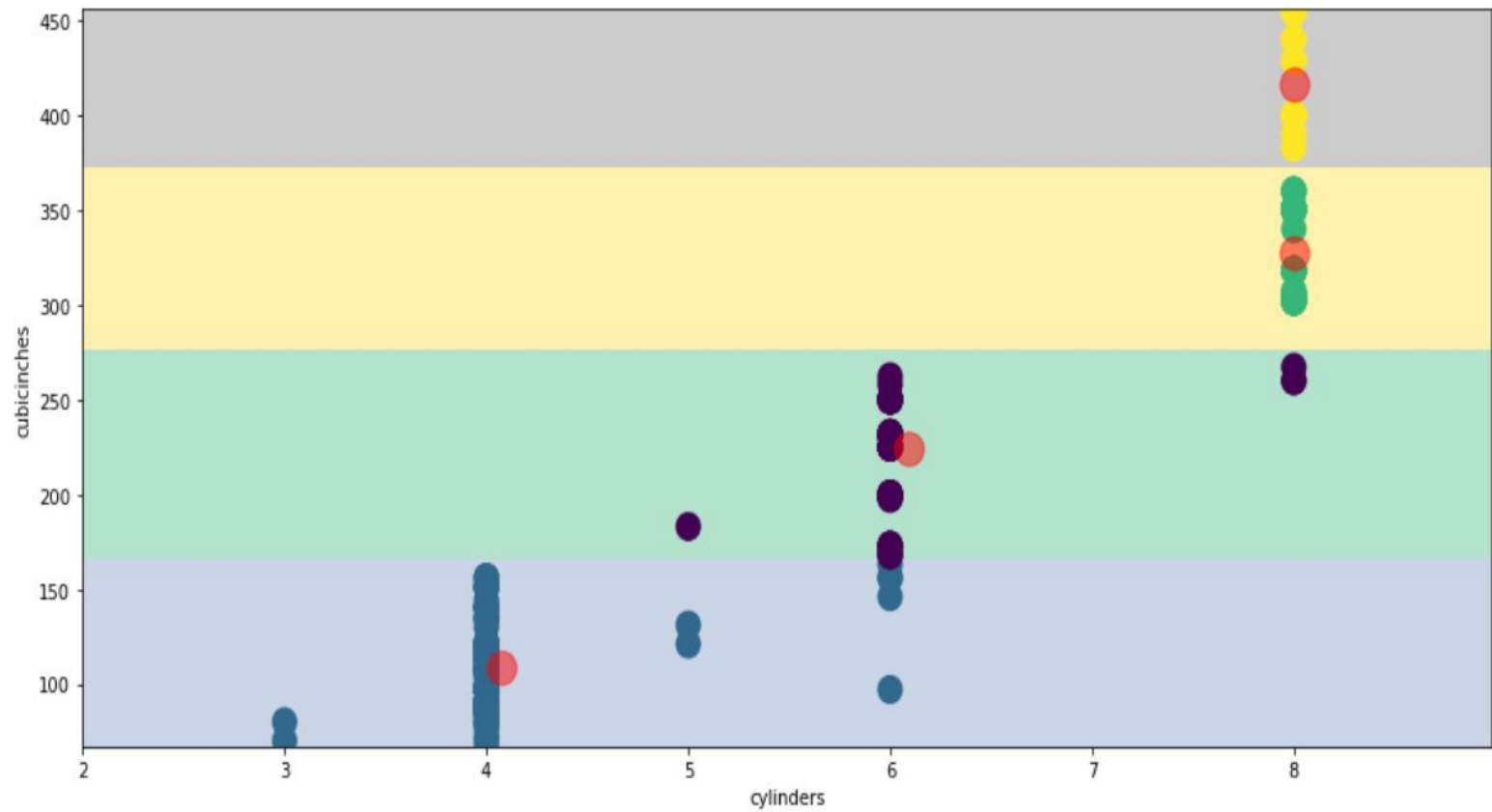
- It is a classification algorithm which takes bunch of labeled points and use them to label another points.
- This method of classification is based on similarity of other neighbors.
- Data points which are nearer to each other are neighbors.
- This is based on Idea that “Similar cases with same class label are neat to each other

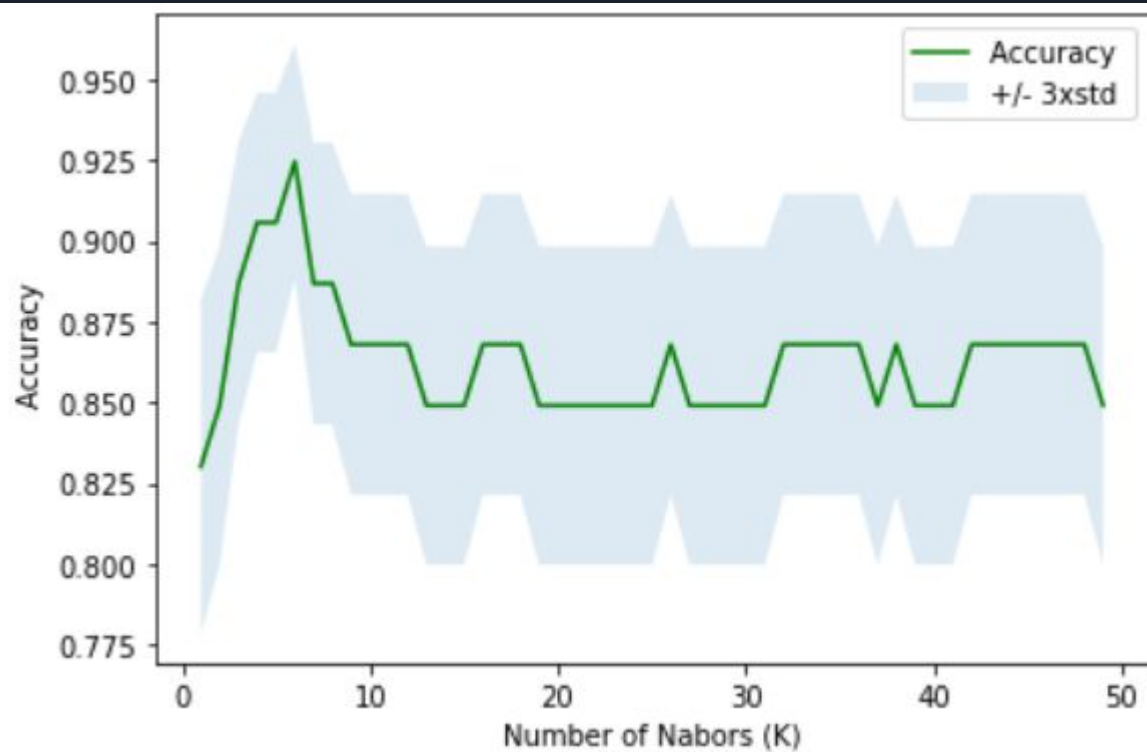
Output : Decision Tree



K Means







Conclusion:

	Model	Score_train	Score_test
0	Decision Tree Classifier	89.230769	77.272727
1	k-Nearest Neighbors	93.269231	92.452830

As we can see KNN is more efficient than Decision tree with training accuracy **93.26923** and testing accuracy **92.45830**

Conclusion

- Thus, a Data Mining task: Clustering was implemented on a dataset. The dataset was a kaggale sourced cars database.
- Preprocessing was performed on the dataset in order to prepare the dataset for the required operations of data mining over it.
- The output results were displayed as well as visualized.
- We are grateful to our mentor Prof. Vaishali Suryawanshi ma'am to provide us with mini project, it was great a learning opportunity for our group.