

**Mini Project Report**

For the Subject

**Data Warehousing Data Mining**

**(Trimester VIII)**

Submitted by:

PG 03 Omkar Najan

PG 09 Mayur Pingale

PG 13 Payal Patil

PG 14 Vaishnavi Salunke

**Under the Guidance of**

**Prof.**

Vaishali Suryawanshi

**School of Computer Engineering and Technology**

**MIT World Peace University, Kothrud,**

**Pune 411 038, Maharashtra - India**

**2020-2021**

**Abstract**

Both the number and complexity of Data Mining projects has increased in late years. Unfortunately, nowadays there isn't a formal process model for this kind of projects, or existing approaches are not right or complete enough. In some sense, present situation is comparable to that in software that led to ”software crisis” in latest 60's. Software Engineering matured based on process models and methodologies. Data Mining's evolution is being parallel to that in Software Engineering. The research work described in this paper proposes a Process Model for Data Mining Projects based on the study of current Software Engineering Process Models (IEEE Std 1074 and ISO 12207) and the most used Data Mining Methodology CRISP-DM (considered as a "facto" standard) as basic references.

**Contents**

|  |  |  |
| --- | --- | --- |
| **Sr. no** | **Title** | **Page no.** |
| 1 | Introduction | 4 |
| 2 | Motivation | 5 |
| 3 | Problem Definition | 6 |
| 4 | Objectives | 7 |
| 5 | Tools Used | 8 |
| 6 | Dataset Description | 9 |
| 7 | Data Preprocessing | 10 |
| 8 | System Architecture | 11 |
| 9 | Data Mining Task Performed | 12 |
| 10 | Algorithm | 13 |
| 11 | Output | 14 |
| 12 | Visualization Screenshots | 15 |
| 13 | Conclusion | 19 |
| 14 | References | 20 |

**Introduction**

This project will take a car as ations to the input and target as output. The project uses a car Database as input to the model. The Model takes 'mpg', 'cylinders', 'cubicinches', 'hp', 'weightlbs', 'time-to-60', 'year', 'brand' as features of every car and forms clusters of car which have similar features using k-means clustering. When a car is given as input to model it will give the brand of a car using the parameters such as horsepower, Cubic inches, Make year

**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-Neighbours.

**Objectives**

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

**Tools Used**

* Python Libraries:

Numpy

Matplotlib

Pandas

Seaborn

sklearn

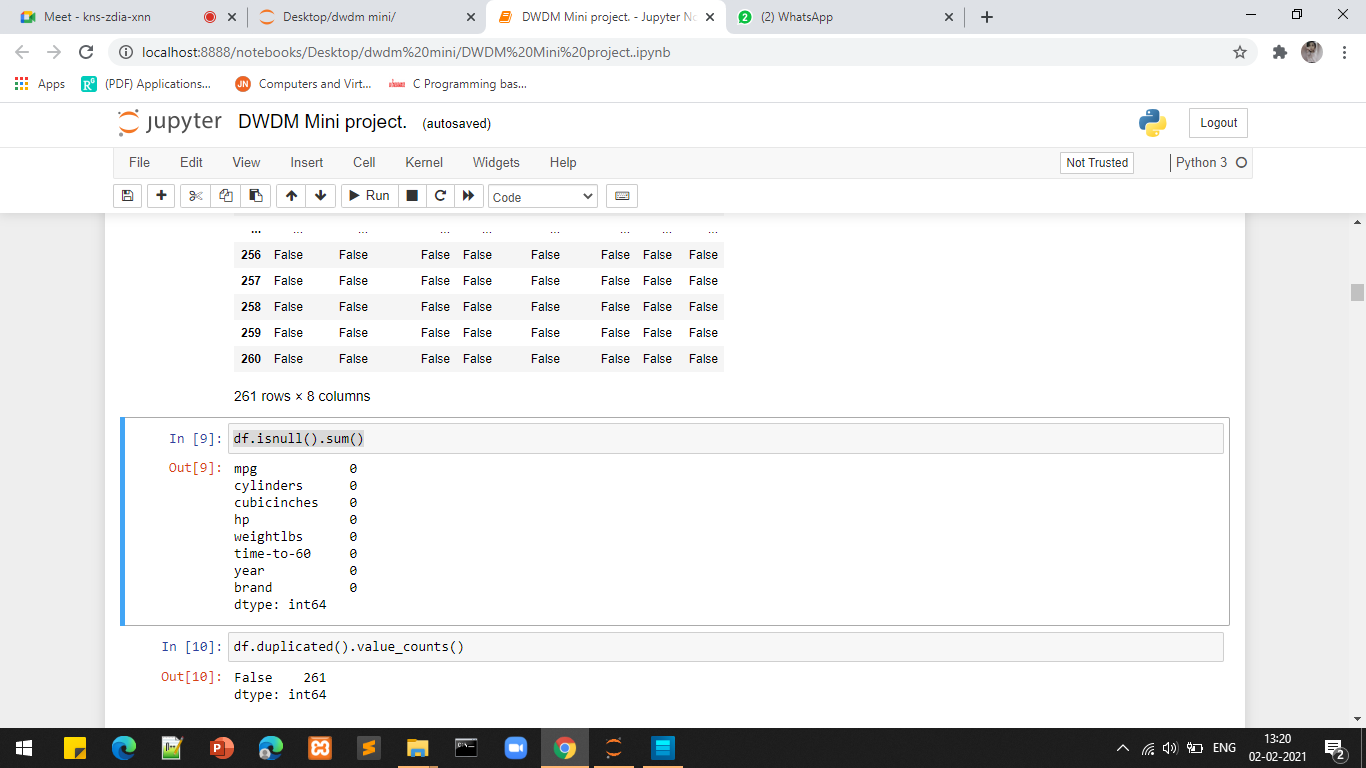
* Jupyter Notebook
* IBM Skill Network Lab

**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.



**Data Mining Task Performed**

The data mining task performed on this project is clustering

The method to perform clustering is K-Means.

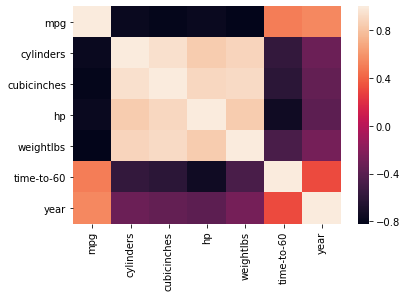
***k*-means clustering** is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. *k*-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances.

**Algorithm**

K-Means algorithm will categorize the items into k groups of similarity. To calculate that similarity, we will use the euclidean distance as measurement.

The algorithm works as follows:

1. Initialize k points, called means, randomly.
2. Categorize each item to its closest mean and we update the mean’s coordinates, which are the averages of the items categorized in that mean so far.
3. Repeat the process for a given number of iterations and at the end, we have our clusters.



**Decision trees:**

**Decision trees** are **used** to solve both **classification** and regression problems in the form of **trees** that can be incrementally updated by splitting the dataset into smaller datasets (numerical and categorical), where the results are represented in the leaf nodes.

**ALGORITHM**

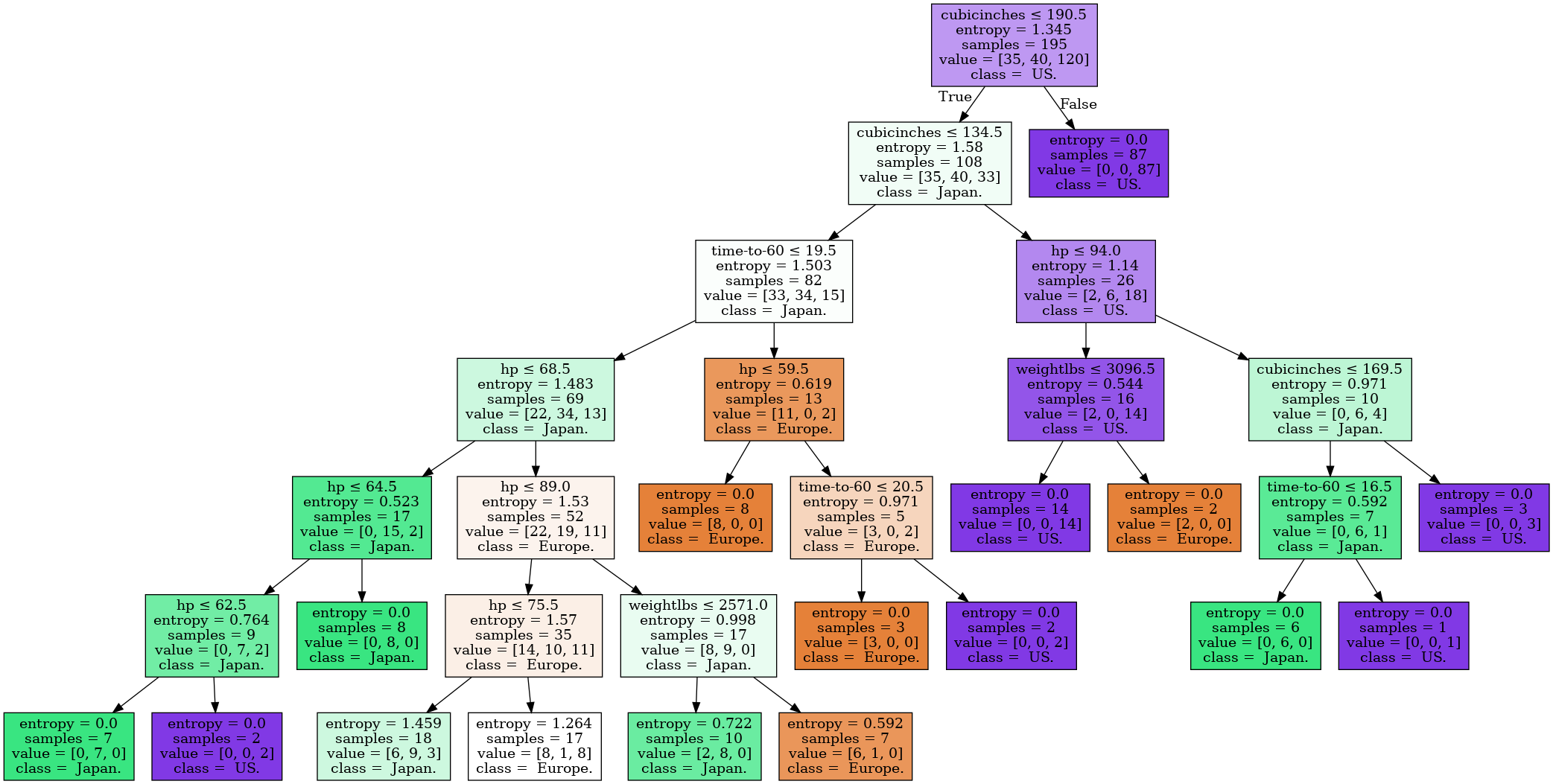
**Decision Tree:**

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data (training data).

In Decision Trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with the record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. The beginning, the whole training set is considered as the **root.**

feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.

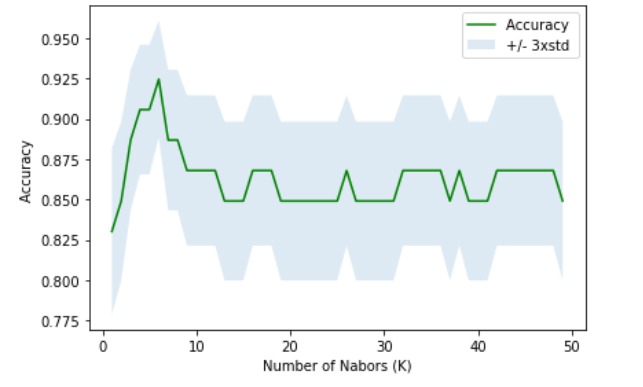
Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

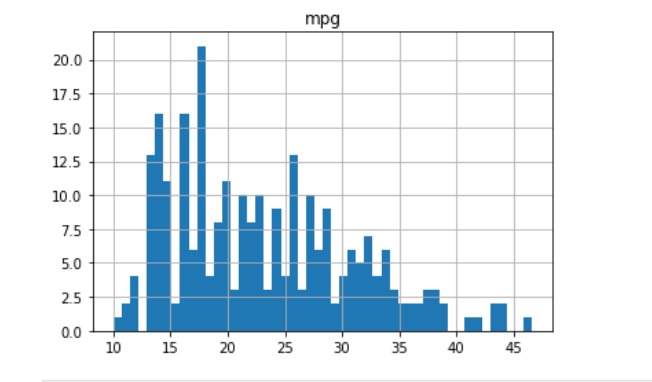


**Output**

The model will receive attributes of car as input, it will then compute all its features to identify which cluster does the belong to. Next the model will pick a few cars from test dataset preferably closest to the input car. This brand is displays as output.

**Visualization Screenshots**

****



**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.

For a given dataset, cars.csv **K-Nearest-Neighbors** is suitable algorithm for predicting brands of Cars with training accuracy 93% and testing accuracy 92%

Decision tree algorithm does not have good accuracy score in both training and testing dataset.

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.

**References**

[dataset](https://www.kaggle.com/abineshkumark/carsdata)

[Theory of KNN](https://www.geeksforgeeks.org/k-nearest-neighbours/)