SHRIYA SHAH
60004210226

UNSUPERVISED LEARNING

Pre-Processing :
It is done to increase accuracy.
Firstly, follow the basic steps like getting the dataset, importing libraries, importing datasets etc.

We need to firstly encode the given data into a numerical representation to apply the machine learning models.

The methods used to encode the data include:

- One-Hot Encoding: Best suited for categorical features. It creates columns with binary values c for each category (1 for present and 0 for absent)
- Response Coding : Represent the probability of the datapoints of a particular class. rP(class=X | category=A) = P(category=A ∩ class=X) / P(category=A)

Handle Null values :
We can do this using the following methods-

- We can eliminate the rows having null values, if the number of missing values were less compared to the data given.
- We can take the mean/median/mode value of the feature and replace the null values with it.
- Use regression to predict the missing values.
- We can also replace the null values with 'none' or empty strings.

Clustering Method Chosen:

K Means :

Reason for Choosing:

- This has been chosen since the date is large and using this would be better preferred.
- It is easier to use.
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

Understanding:

How it works-

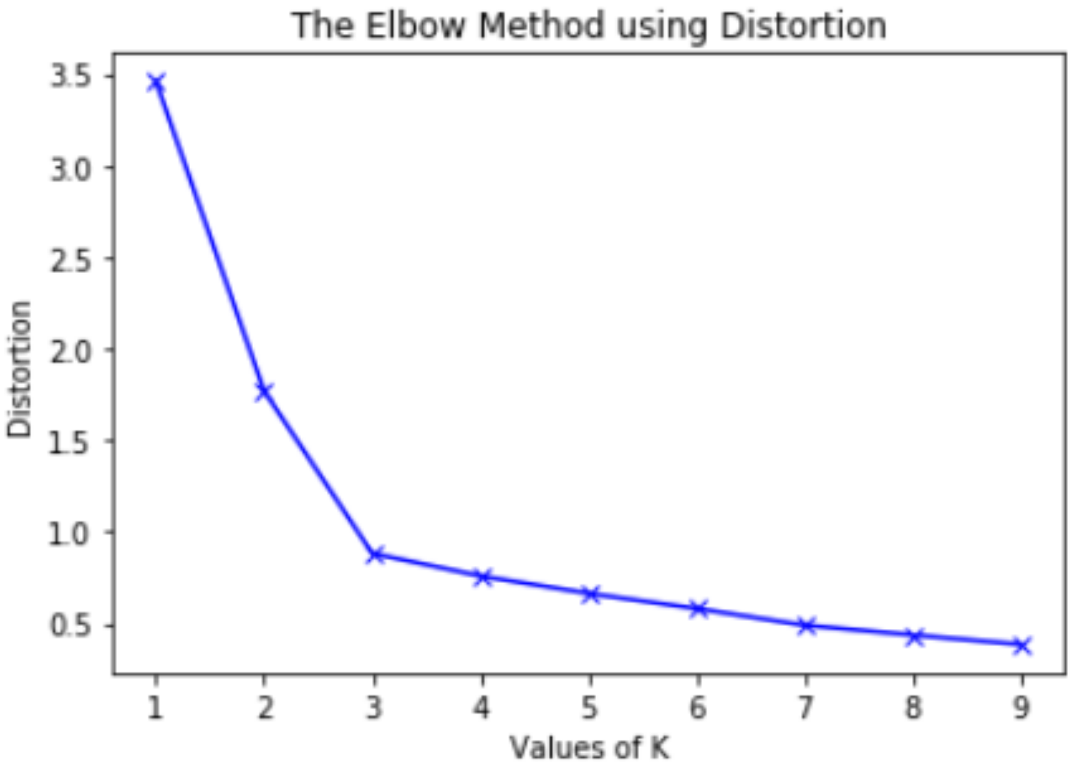
- Iterations are taken using different k(centroids) values.
- Initialise k number of centroids.
- Find out which points are near the centroid using Euclidian distance.
- Compute the average to update the centroid.
- The data points nearer to one centroid form the same cluster.

To decide number of clusters:
We need to determine the k-value

To decide k-value-
Elbow method:

Clustering Method Chosen:

K Means :



- For every iteration draw graph with k-value on the x-axis and WSCC on the y-axis (WSCC is short for Within Cluster, Sum of Square)
- The graph thus formed will have an abrupt change at one point and will have an almost constant value after this point.
- This point gives the k-value.
- Still, you need to confirm this by finding out the Silhouette coefficient.

Validating Clustering Method-
Silhouette:

- Find out the silhouette coefficient
- This variable varies from -1 to 1.
- The closer this value is to 1, the more accurate the iteration is.

$$s(\boldsymbol{o}) = \frac{b(\boldsymbol{o}) - a(\boldsymbol{o})}{\max\{a(\boldsymbol{o}), b(\boldsymbol{o})\}}$$

- Here, b(o) is the average of distance between the datapoints from 2 clusters
- a(o) is the average of the distance between 2 datapoints within a cluster.
- O is the datapoint.
- b(o) value should be larger than a(o) value for better formed clusters.

Points to Remember :

- If the s(i) value is negative for any cluster label, do not consider it for the k-value.
- Always prefer a larger number as the value of k, for creating a generalised mode.