**DATA PROCESSING**

1. import all libraries

read data

2. convert the Data into Dataframe for convenience

3. in case of any null values in AGE ,replace it by average value

4. in case of null values in the remaining column, replace by tab ' '

**FURTHER PROCESSING**

5. if we want to work on 1 column say job which has 3 types housemaid ,services,admin

we can use label encoding. this assigns numeric value to each category and they come in one row , thus further processing can be done

**CLUSTERING**

1. k means- exclusive

2. c means - overlapping

3. hierarchy - branched

Here the columns are independent and are information about the particular index .

since there seems to be no overlapping or hierarchy so the clustering method used would be K-means

In k means clustering method

1.same types of items are grouped together

2. item cannot share 2 clusters

3. say we want to divide in 3 clusters.

* random 3 points are chosen
* their distance is calculated from the 1s point
* nearest point is marked in cluster 1
* mean between these points are taken and the distance is calculated from this point and so onn
* distance calculated by Euclidean method- corresponds to Pythagoras in 2D

**NUMBER OF CLUSTERS**

1. number of clusters can be determined by the elbow graph
2. we make a loop starting with 1 cluster to say 10, in that
3. we make use of k means ++ which automatically chooses best points in the custer to calculate distance
4. we use kmeans\_inertia to find the wcss value and put it in array
5. we have an array of wcss for cluster 1 to 10
6. then plot a graph for cluster vs wcss
7. it will be an elbow graph
8. the point of sudden change or elbow point will be the no of clusters