IBM – NAAN MUDHALVAN APPLIED DATA SCIENCE ANALYSIS

Project 2 : Covid-19 Vaccines Analysis

Phase – 2 – INNOVATION

INTRODUCTION:

Nowadays, machine learning (ML) used in every area of computational work where algorithms are designed, and performance is increased .In the last years, learning from unbalanced data sets has become a critical problem in machine learning and is frequently found in several applications such as computer security.

Machine learning, Like some other technologies, played a crucial role in Determining the virus's triggers and conditions . It was an Attempt to clean up the noisy data that had scattered across The world in order to educate biological areas where research Was attempting to understand how the virus resides beyond The human body and the effects of various factors such as Climate, population, and on COVID-19 spread.

Clustering is often a valuable function for Learning data. Uncontrolled clustering is known as the Segmentation of data into clusters that contain the same data, Mainly to make homogeneous groups.

METHODOLOGY:

Here we use two machine learning techniques they are;

- ❖ Clustering.
- ❖ Time series forecasting.

Clustering Algorithms:

A clustering algorithm divides a data set into many classes, With the similarity inside each group being greater than the Similarity within groups.

Clustering algorithms are commonly used for data Structure and categorization, as well as data compression and Model creation.

Now,we going to use two cluster they are;

- K-means clustering algorithm.
- Hierarchical clustering algorithm.

K-means clustering algorithm:

k-means algorithm is a clear partition process. It separated (N) data objects into (K) cluster sets to obtain low cross similarity and high intracluster specificity.

Steps:

Step 1: Get started: As initial centers, select k data items at

random from Data collection D. K is the number of clusters.

Step 2: Repetition:

a) Assume that each cluster is a centroid.

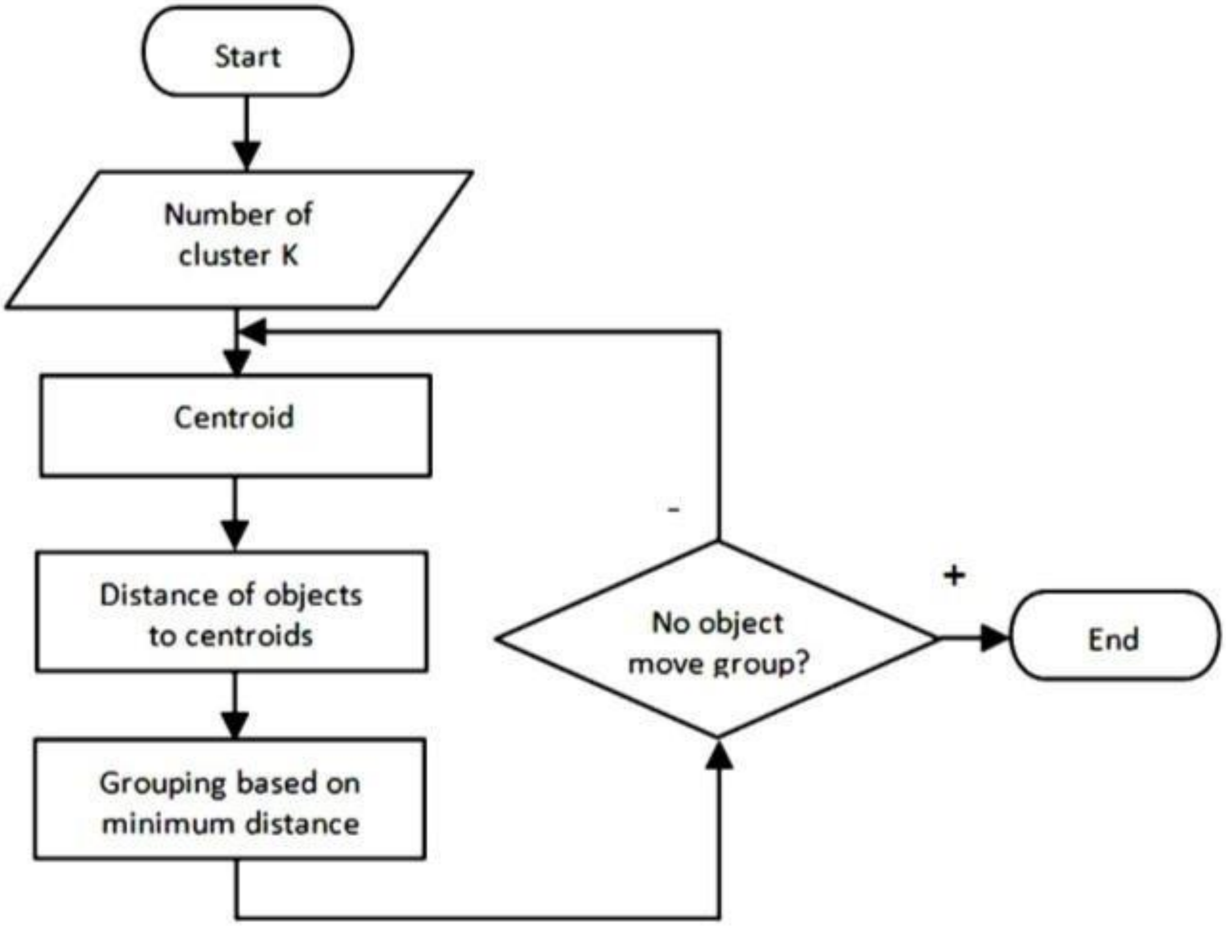
b) Evaluate the difference between all of the data points, as well as the centroids.

c) Allocate di to the cluster that is nearest to you as a data object.

Step 3: By each cluster j (1=j=k), create an update. Calculate the cluster center once more.

Step 4: Repeat until there is no difference in the cluster's core.

Step 5: Finish.


Start
Number of cluster K
Centroid
Distance of objects to centroids
Grouping based on minimum distance
No object move group?
-
+
End

K-means Clustering Process.

Hierarchical Clustering Algorithm:

By creating a hierarchy of clusters, also known as a dendrogram, the hierarchical Clustering approach combined or separates identical data items . The hierarchical Clustering approach creates clusters in a step-by-step manner.

Agglomerative and Divisive algorithms are the two Kinds of hierarchical clustering algorithms. Agglomerative Hierarchical clustering algorithm: Agglomerative hierarchical Clustering is a bottom-up approach that starts from each Person within a cluster.

Steps:

Step 1: Get started: Assign a cluster number equivalent to the Number of objects.

Step 2: Repetition: When the number of clusters is set to 1 or

when the consumer specifies the number of clusters.

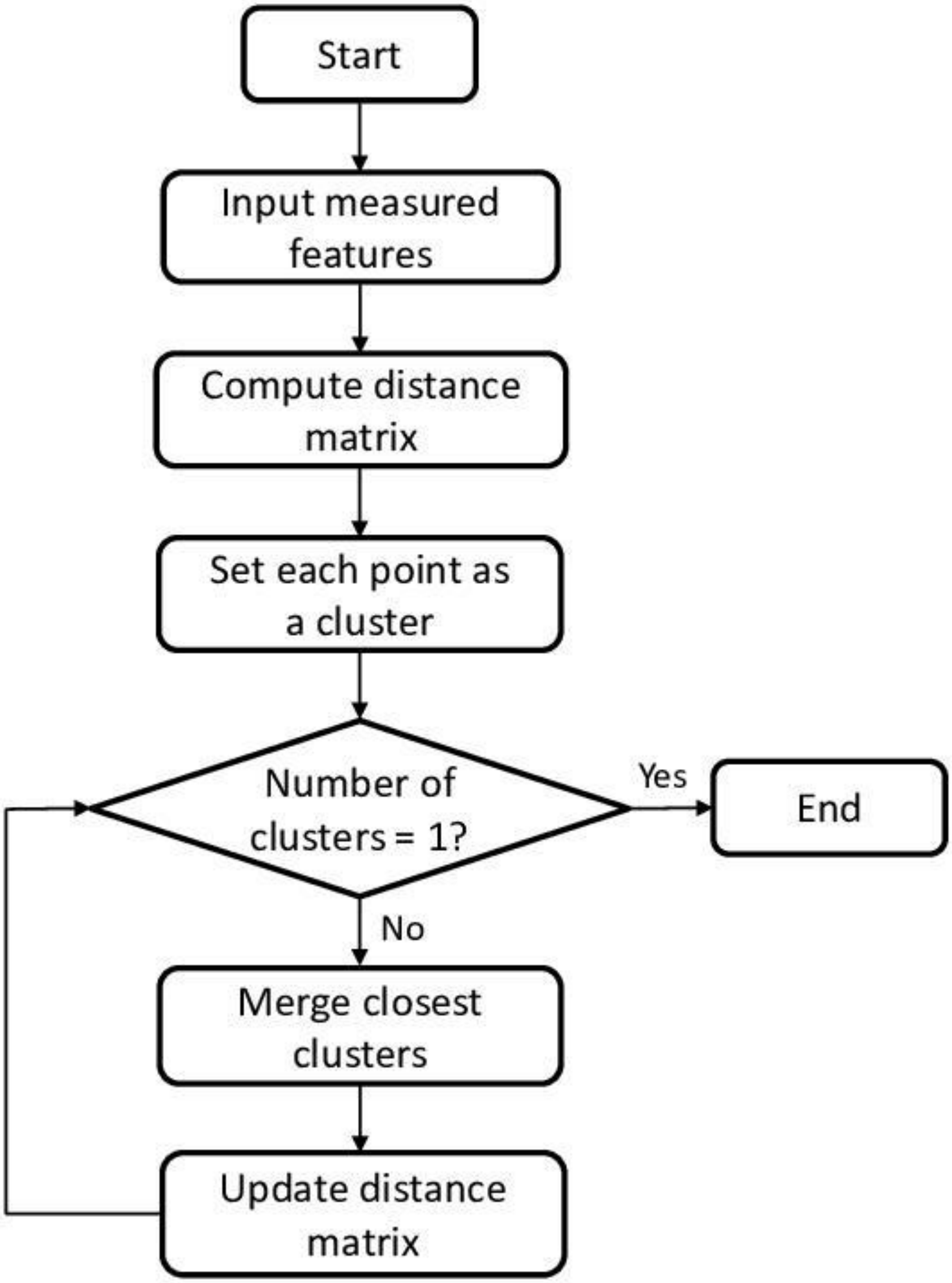
a) Calculate the shortest inters cluster size.

b) Combine the clusters with the smallest inter-cluster

distance.

Step 3: Finish.

Divisive hierarchical clustering algorithm: This is a hierarchical clustering algorithm, as opposed to a bottom-up approach.



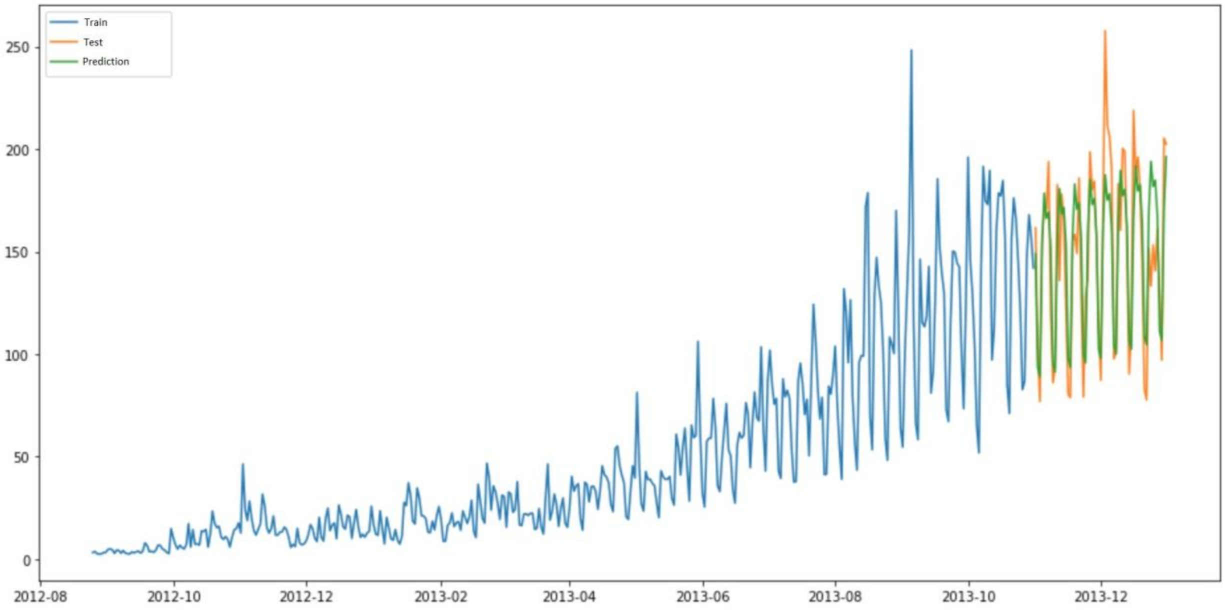
Flowchart Agglomerative Hierarchical Clustering.

Time series forecasting:

Time series forecasting is a technique used to make predictions about future data points in a time-ordered sequence. It’s widely used in various fields, such as finance, economics, and weather forecasting.

Machine Learning:

Methods like linear regression, decision trees, and neural networks can be used for time series forecasting. LSTM (Long Short-Term Memory) networks are particularly effective for sequential data.


Train
Test
Prediction