Pandemic Information Mining

Sunny Joseph - FIT18CS120
Melvyn Joseph - FIT18CS082
Kevin Poly - FIT18CS073
Rakenth Pc - FIT18CS105

FISAT

September 19, 2022

Contents

Problem Statement

Here our objective is to find 4 possible methods for the implementation of Pandemic Information Mining.

Methodology

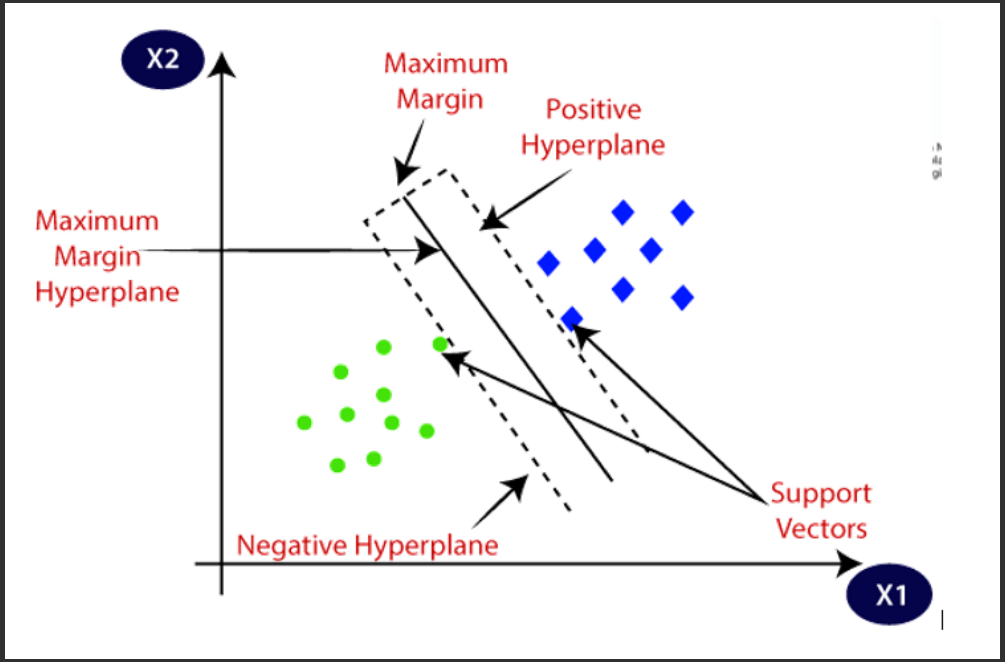
- Our aim is to mine information based upon the recent pandemic outbreak.
- We have also implement it with 4 different data mining algorithm and compare them to come with a conclusion of which algorithm is the best.

1. SUPPORT VECTOR MACHINE
2. K-NEAREST NEIGHBOR
3. NAIVE BAYES
4. K-MEANS ALGORITHM

Support Vector Machine(SVM)

- SVM is a supervised machine learning algorithm. It is commonly used for classification and regression challenges.
- In the SVM algorithm, each point is represented as a data item within the n-dimensional space where the value of each feature is the value of a specific coordinate.
- Its main task is to find the best hyper plane that can separate data perfectly into its two classes.
- Its main task is to find the best hyper plane that can separate data perfectly into its two classes.

Support Vectors:



- **Hyper plane: As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.**
- **Margin: It may be defined as the gap between two lines on the closet data points of different classes.**

Algorithm

- Step-1: START.
- Step-2: Load all the Data sets required(Depends on the pandemic the user wants to predict).
- Step-3: Let future forecast be used for predicting the outbreak for n number of days .
- Step-4: Select the parameters for SVM such as kernel ,c, gamma, epsilon and shrinking.
- Step-5: We can assign svmsearch as Randomized Search CV to build a model of svm by using the necessary parameters .
- Step-6: Fitting the data by using svm search.fit(x train confirmed,y train confirmed).
- Step-7: Let svm confirmed be assigned with the estimate values from the Randomized Search CV.
- Step-8: Future forecast can be predicted by using svm confirmed.predict(future forecast).
- Step-9: STOP

- By using SVM we will be able to approximately predict a pandemic outbreak or future pandemic data by comparing existing data sets with svm algorithm.
- The objective of applying SVMs is to find the best hyperplane in order to help us separate our space into classes. The hyperplane (line) is found through the maximum margin, i.e., the maximum distance between data points of both classes.

Advantages

- SVM works relatively well when there is a clear margin of separation between classes.
- SVM is more effective in high dimensional spaces.
- SVM is effective in cases where the number of dimensions is greater than the number of samples.
- SVM is relatively memory efficient.

Disadvantages

- SVM algorithm is not suitable for large data sets.
- SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.
- In cases where the number of features for each data point exceeds the number of training data samples, the SVM will under perform .
- As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification..

KNN Algorithm

- **K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.**

What is Supervised Learning?

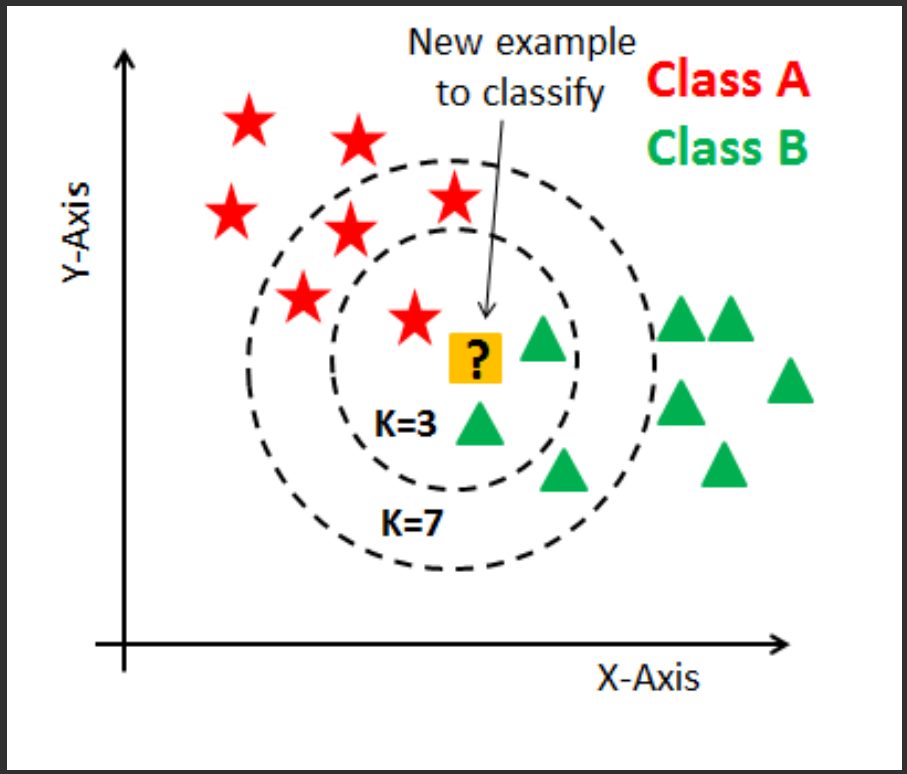
- **Supervised learning algorithm is a kind of algorithm where it relies on labelled input to learn and predicts the output.**

KNN Algorithm

- KNN Algorithm is a simple algorithm which stores all the available data and classifies a new data point based on the similarity.
- K-NN algorithm uses distance metrics in order to find similarities or dissimilarities.
- KNN is a non-parametric algorithm. [A non-parametric algorithm is computationally slower, but makes fewer assumptions about the data.]
- It is also called a lazy learner algorithm. [It only memorizes and does nothing to learn by itself.]

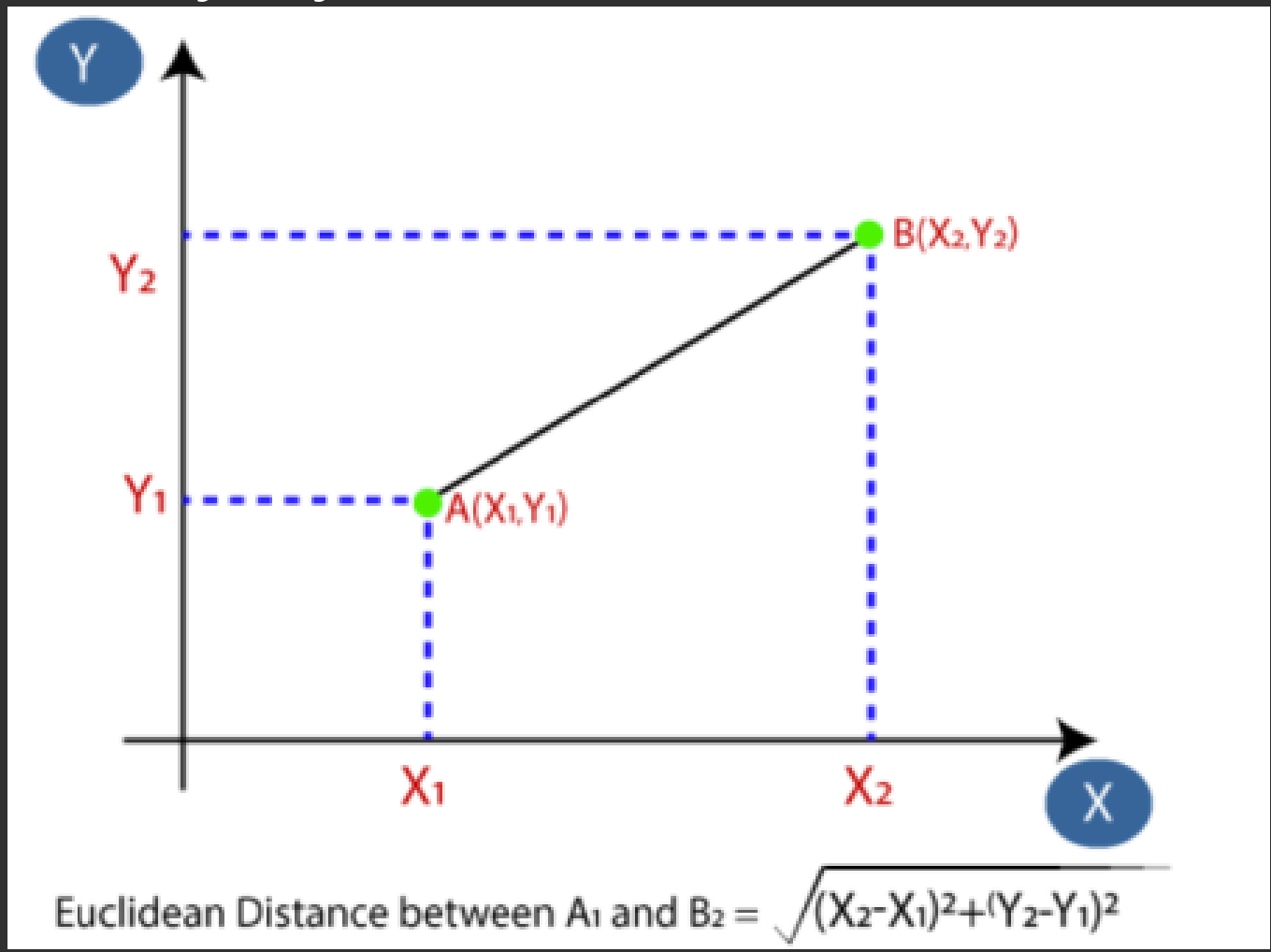
How does Knn algorithm work?

- **KNN Algorithm is based on feature similarity.**
- **K in KNN algorithm represents the number of nearest neighbours.**



How distance in calculated in KNN?

- **There are many ways to calculate distance.**



Small Example how it works!

Perform KNN - classification Algorithm on following dataset and Predict the class for $x(P_1 = 3$ and $P_2 = 7)$. $k = 3 \rightarrow 3$ nearest neighbours.

	P_1	P_2	Class
(i)	7	7	False
(ii)	7	4	False
(iii)	3	4	True
(iv)	1	4	True

Euclidean Distance $= \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$

X_2 → Observed Value, X_1 → Actual Value

D(i) $= \sqrt{(3-7)^2 + (7-7)^2} = 4 \rightarrow N3 \rightarrow$ False

D(ii) $= \sqrt{(3-7)^2 + (7-4)^2} = \sqrt{16+9} = 5$

D(iii) $= \sqrt{(3-3)^2 + (7-4)^2} = 3 \rightarrow N_1 \rightarrow$ TRUE

D(iv) $= \sqrt{(3-1)^2 + (7-4)^2} = \sqrt{4+9} = 3.6 \rightarrow N_2 \rightarrow$ TRUE

2 TRUE > 1 FALSE

Therefore $(P_1 = 3$ and $P_2 = 7)$ will belong to class TRUE

Algorithm

- Step-1: START.
- Step-2: Load the training dataset [.csv or .xls file]
- Step-3: Initialize the value of K.
- Step-4: Calculate the Euclidean distance of K number of neighbours .
- Step-5: Take the K nearest neighbours as per the calculated Euclidean distance.
- Step-6: Among these K neighbours, count the number of the data points.
- Step-7: Assign the new data points to that category for which the number of the neighbour is maximum.
- Step-8: STOP

Advantages

- KNN is veryeasy to implement.
- No Training Period is required for KNN.Thus it is called Lazy Learner (Instance based learning). It does not learn anything in the training period.
- Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm.

Disadvantages

- **The main disadvantage of the KNN is it does not work well with large dataset.**
- **Sensitive to noisy data, missing values and outliers.**
- **K value must be decided carefully.**

NAÏVE BAYES Algorithm

- **Naive Bayes is a machine learning model that is used for large volumes of data.**
- **It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors.**
- **It uses Bayes' Theorem.**
- **Bayes' Theorem predicts the occurrence of any event.**

Bayes' Theorem

- It is a theorem that works on conditional probability.
- Conditional probability is the probability that something will happen, given that something else has already occurred
- The conditional probability can give us the probability of an event using its prior knowledge.

Conditional Probability:

- P(H/E)=(P(E/H)*P(H))/P(E)
- Where: P(H): The probability of hypothesis H being true. This is known as prior probability.
- P(E): The probability of the evidence.
- P(E|H): The probability of the evidence given that hypothesis is true.
- P(H|E): The probability of the hypothesis given that the evidence is true

Working

- **Consider the following data set,**

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

- The data set are ‘Weather’ and its corresponding target ‘Play’.
- In the first step the data set is converted to Frequency table.
- Then the likelihood table is created by finding the probabilities like overcast probability and probability of playing.
- Now Bayes’ Theorem is used to calculate the posterior probability for each class .The class with the highest posterior probability is the outcome of prediction.

Problem: "Players will play if the weather is sunny".

Bayesian equation: P(H/E)=(P(E/H)*P(H))/P(E)

P(Yes|Sunny)=(P(Sunny|Yes)*P(Yes))/P(Sunny)

P(Sunny|Yes)=3/9=0.33

P(Yes)=9/14-0.64

P(Sunny)=5/14=0.36

Thus, P(Yes|Sunny)=0.33*0.64/0.36=0.60 , Has the highest probability.So the players will play.

Algorithm

- Step-1: START.
- Step-2: : Handling the data. Import the data from the file using the import function and then test it.
- Step-3: Summarise the data. Summarising the data will help in calculating the probability and make prediction.
- Step-4: : Make a particular prediction. For making a particular prediction we could use the summarised data.
- Step-5: Generate the predictions from the given training data set and the summarised data set.
- Step-6: Evaluating the accuracy. The accuracy of the predictions are evaluated as the percentage correct of the data set.
- Step-7: Finally all the data are tied together and a Naive Bayes classifier model is build.
- Step-8: STOP

Advantages

- It is a highly extensible algorithm which is very fast.
- It can be used for both binary as well as multi class classification.
- It can be easily trained on small datasets and can be used for large volumes of data as well.

Disadvantages

- **The main disadvantage of the NB is considering all the variables independent that contributes to the probability.**

K-MEANS ALGORITHM

- K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.
- It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

- It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum.
- The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

WORKING

- In this approach, the data objects ('n') are classified into 'k' number of clusters in which each observation belongs to the cluster with nearest mean.
- It defines 'k' sets (the point may be considered as the center of a one or two dimensional figure), one for each cluster k n. The clusters are placed far away from each other.
- Then , it organizes the data in appropriate data set and associates to the nearest set. If there is no data pending, first step is complicated to perform, in this case an early grouping is done. It is necessary to re-calculate 'k' new set as barycenters of the clusters from previous step.

- After having these 'k' new sets, the same data set points and the nearest new sets are bound together.
- Finally, a loop is generated. As a result of this loop, the 'k' sets change their location step by step until no more changes are made.

Finally, this algorithm aims at minimizing an objective function as:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where,

$x_i^{(j)}$ = data point

c_j = cluster centre

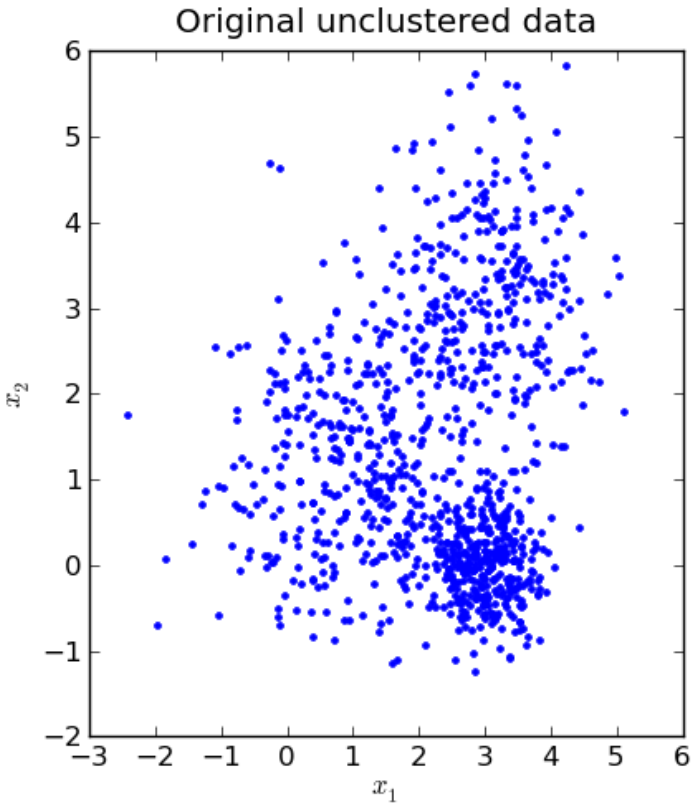
n = Number of data points

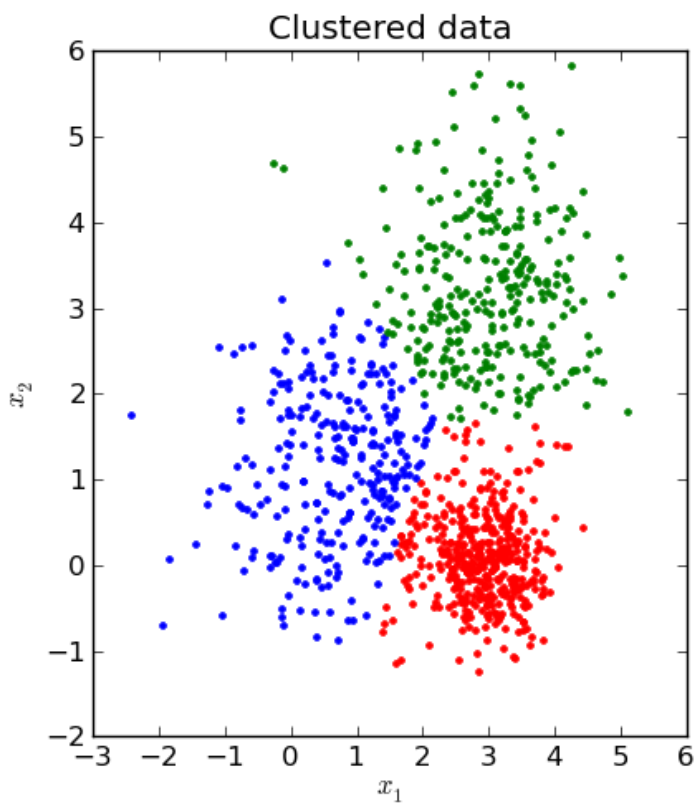
k = Number of cluster

$||x_i^{(j)} - c_j||^2$ = distance between a data point $x_i^{(j)}$ and cluster centre c_j.

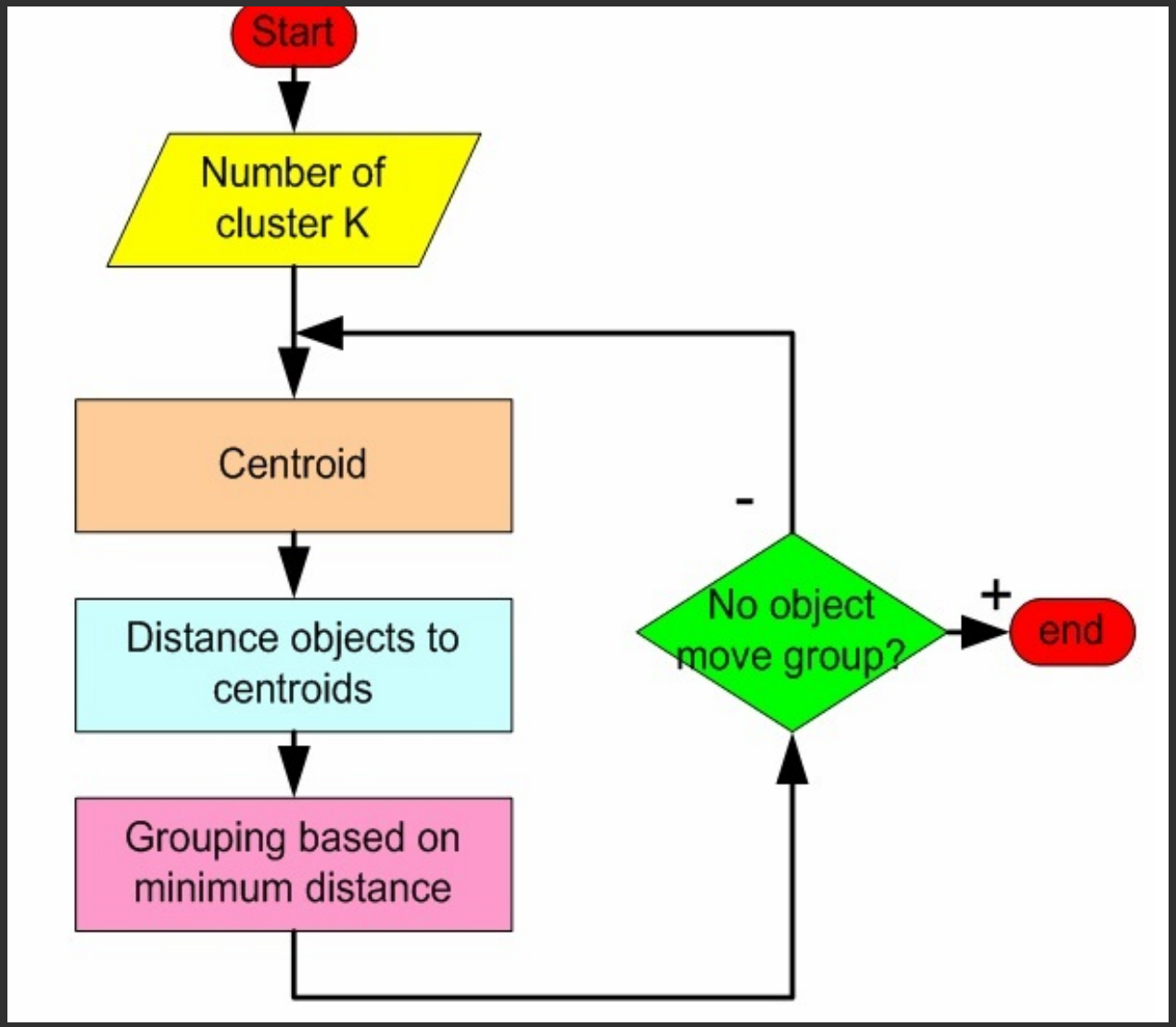
WORKING

- **A FORM OF RAW UNSORTED(UNCLUSTERED) DATA IS FED TO THE PROGRAMAND THE DATA IS CLUSTERED INTO K SIMILAR GROUPS(in this case the data is sorted into 3 groups)**


Original unclustered data
x_1
x_2


Clustered data
x_1
x_2

Flowchart


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

Algorithm

- Step-1: START.
- Step-2: Import the data set.
- Step-3: Specify number of clusters K.
- Step-4: Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
- Step-5: Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
- Step-6:STOP

Advantages

- **Easy to implement.**
- **With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small).**

Disadvantages

- **Difficult to predict the number of clusters (K-Value).**

Conclusion

- **As we have seen at all the 4 different algorithm that can be used in Pandemic Information Mining.**

References

1. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
2. https://www.educba.com/knn-algorithm/
3. https://hackerbits.com/data/k-nearest-neighbor-knn-data-mining-algorithm/
4. https://www.youtube.com/watch?v=6kZ-OPLNcgEab-channel=edureka
5. https://intellipaat.com/blog/tutorial/machine-learning-tutorial/svm-algorithm-in-python
6. https://www.youtube.com/watch?v=sHWKN5dakPw
7. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
8. https://www.analyticssteps.com/blogs/what-naive-bayes-algorithm-machine-learning
9. https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

Thank You!!