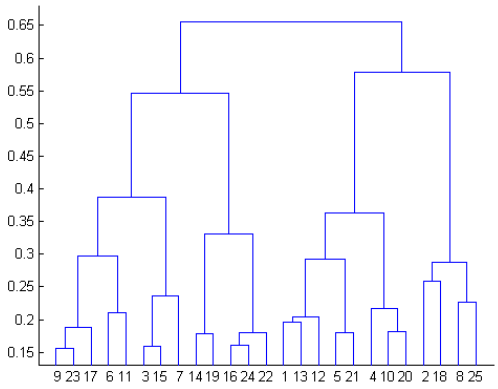
Name of the Assignment: Machine Learning (Worksheet-1)	
Submitted by	: Shalini Joshi
Designation	: Data Science Intern
Date of Submission	: 21st Dec,2022

Objective Type Questions:

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:



a) 2 b) 4 c) 6 d) 8

Ans: b) 4

2. In which of the following cases will K-Means clustering fail to give good results?
1. Data points with outliers 2. Data points with different densities 3. Data points with round shapes 4. Data points with non-convex shapes
Options: a) 1 and 2 b) 2 and 3 c) 2 and 4 d) 1, 2 and 4

Ans: d) 1,2 and 4

3. The most important part of is selecting the variables on which clustering is based.
a) interpreting and profiling clusters
b) selecting a clustering procedure
c) assessing the validity of clustering
d) formulating the clustering problem

Ans: d) Formulating the clustering problem

4. The most commonly used measure of similarity is the …….or its square.
a) Euclidean distance b) city-block distance c) Chebyshev's distance d) Manhattan distance
Ans: a) Euclidean distance

5. ……. is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.
a) Non-hierarchical clustering b) Divisive clustering c) Agglomerative clustering d) K-means clustering
Ans: b) Divisive clustering

6. Which of the following is required by K-means clustering?
a) Defined distance metric b) Number of clusters c) Initial guess as to cluster centroids d) All answers are correct
Ans: d) All answers are correct

7. The goal of clustering is to
a) Divide the data points into groups b) Classify the data point into different classes c) Predict the output values of input data points d) All of the above
Ans: a) Divide the data points into groups

8. Clustering is a
a) Supervised learning b) Unsupervised learning c) Reinforcement learning d) None
Ans: b) Unsupervised Learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?
a) K- Means clustering b) Hierarchical clustering c) Diverse clustering d) All of the above
Ans: d) All of the above

10. Which version of the clustering algorithm is most sensitive to outliers?
a) K-means clustering algorithm b) K-modes clustering algorithm c) K-medians clustering algorithm d) None
Ans: a) K-means clustering algorithm

11. Which of the following is a bad characteristic of a dataset for clustering analysis
a) Data points with outliers b) Data points with different densities c) Data points with non-convex shapes d) All of the above
Ans: d) All of the above

12. For clustering, we do not require

a) Labeled data b) Unlabeled data c) Numerical data d) Categorical data

Ans: a) Labeled data

Subjective Type Questions:

13.How is cluster analysis calculated?

Ans: Cluster analysis is an exploratory analysis that tries to identify structures within the data. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. Because it is exploratory, it does not make any distinction between dependent and independent variables.

It is calculated by measuring the distance between each data point and its centroid, squaring this distance, and summing these squares across one cluster.

14. How is cluster quality measured?

Ans: ***Measures for quality of clustering:***

- **Dissimilarity/Similarity metric**: The similarity between the clusters can be expressed in terms of a distance function, which is represented by d(i, j). Distance functions are different for various data types and data variables. Distance function measure is different for continuous-valued variables, categorical variables, and vector variables. Distance function can be expressed as Euclidean distance, Mahalanobis distance, and Cosine distance for different types of data.
- **Cluster completeness**: If any two data objects are having similar characteristics, then they are assigned to the same category of the cluster according to ground truth. Cluster completeness is high if the objects are of the same category.
- **Ragbag**: In some situations, there can be a few categories in which the objects of those categories cannot be merged with other objects. Then, the quality of those cluster categories is measured by the Rag Bag method. According to the rag bag method, we should put the heterogeneous object into a rag bag category.
- **Small cluster preservation**: If a small category of clustering is further split into small pieces, then those small pieces of cluster become noise to the entire clustering and thus it becomes difficult to identify that small category from the clustering.

15. What is cluster analysis and its types?

Ans: Cluster analysis is a multivariate data mining technique whose goal is to group objects (eg., products, respondents, or other entities) based on a set of user selected

characteristics or attributes. It is the basic and most important step of data mining and a common technique for statistical data analysis. Clusters should exhibit high internal homogeneity and high external heterogeneity.

Types of Cluster Analysis:

- **Hierarchical Cluster Analysis**

In this method, first, a cluster is made and then added to another cluster (the most similar and closest one) to form one single cluster. This process is repeated until all subjects are in one cluster. This particular method is known as Agglomerative method. Agglomerative clustering starts with single objects and starts grouping them into clusters.

- **The divisive method** is another kind of Hierarchical method in which clustering starts with the complete data set and then starts dividing into partitions.

- **Centroid-based Clustering**

In this type of clustering, clusters are represented by a central entity, which may or may not be a part of the given data set. K-Means method of clustering is used in this method, where k are the cluster centers and objects are assigned to the nearest cluster centers.

- **Distribution-based Clustering**

It is a type of clustering model closely related to statistics based on the modals of distribution. Objects that belong to the same distribution are put into a single cluster. This type of clustering can capture some complex properties of objects like correlation and dependence between attributes.

- **Density-based Clustering**

In this type of clustering, clusters are defined by the areas of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate clusters. The objects in these sparse points are usually noise and border points in the graph.

********** The End**********