MACHINE LEARNING

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is an application of clustering?
    a. Biological network analysis
    b. Market trend prediction
    c. Topic modeling
    d. **All of the above**

Clustering technique is used in various applications such as market research and customer segmentation, biological data and medical imaging, search result clustering, recommendation engine, pattern recognition, social network analysis, image processing, etc.

2. On which data type, we cannot perform cluster analysis?
    a. Time series data
    b. Text data
    c. Multimedia data
    d. **None**

Cluster analysis is a statistical method for processing data. It works by organizing items into groups, or clusters, on the basis of how closely associated they are.

3. Netflix's movie recommendation system uses-
    a. **Supervised learning**
    b. Unsupervised learning
    c. Reinforcement learning and Unsupervised learning
    d. All of the above

4. The final output of Hierarchical clustering is-
    a. The number of cluster centroids
    b. The tree representing how close the data points are to each other
    c. A map defining the similar data points into individual groups
    d. **All of the above**

Hierarchical clustering results in a clustering structure consisting of nested partitions

5. Which of the step is not required for K-means clustering?
    a. A distance metric
    b. Initial number of clusters
    c. Initial guess as to cluster centroids
    d. **None**

K-means clustering is **a type of unsupervised learning, which is used when you have unlabelled data (i.e., data without defined categories or groups)**.

6. Which is the following is wrong?
    a. k-means clustering is a vector quantization method
    b. k-means clustering tries to group n observations into k clusters
    c. **k-nearest neighbour is same as k-means**
    d. None

MACHINE LEARNING

Answer – c. k-nearest neighbour is same as k-means

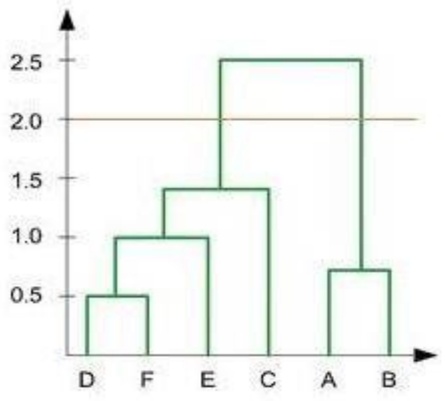
7. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?
   i. Single-link
   ii. Complete-link
   iii. Average-link Options:
   a. 1 and 2
   b. 1 and 3
   c. 2 and 3
   d. **1, 2 and 3**

Answer – d. 1, 2 and 3

8. Which of the following are true?
   i. Clustering analysis is negatively affected by multicollinearity of features
   ii. Clustering analysis is negatively affected by heteroscedasticity Options:
   a. **1 only**
   b. 2 only
   c. 1 and 2
   d. None of them

Answer – a. 1 only

9. In the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?



a. **2**
b. 4
c. 3
d. 5

Answer – a. 2

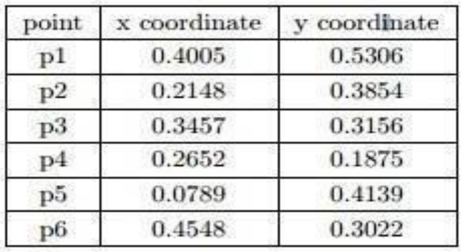
10. For which of the following tasks might clustering be a suitable approach?

MACHINE LEARNING

a. **Given sales data from a large number of products in a supermarket, estimate future sales for each**
of these products.
b. Given a database of information about your users, automatically group them into different market segments.
c. Predicting whether stock price of a company will increase tomorrow.
d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

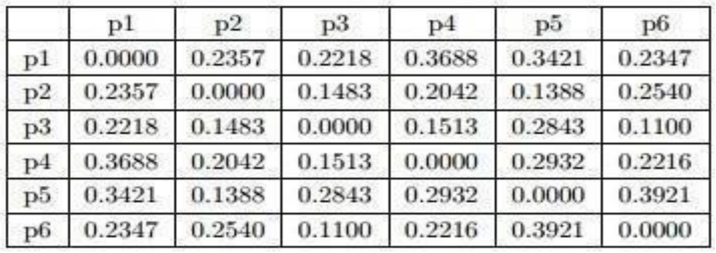
Answer – a. Given sales data from a large number of products in a supermarket, estimate future sales for each

11. Given, six points with the following attributes:



point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

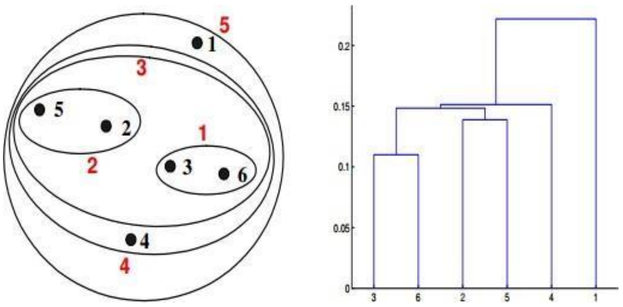
Table : X-Y coordinates of six points.



	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

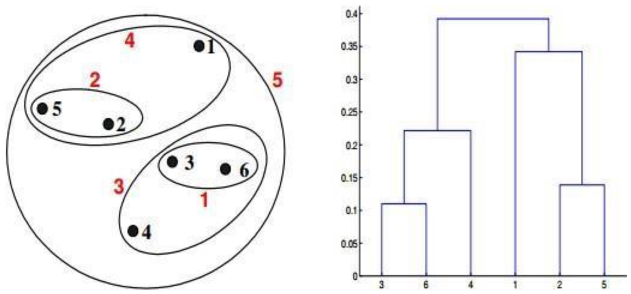


a.

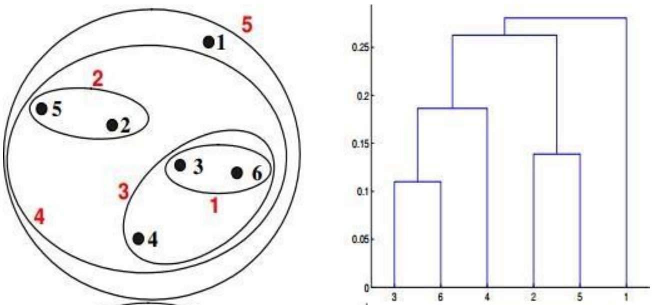
MACHINE LEARNING

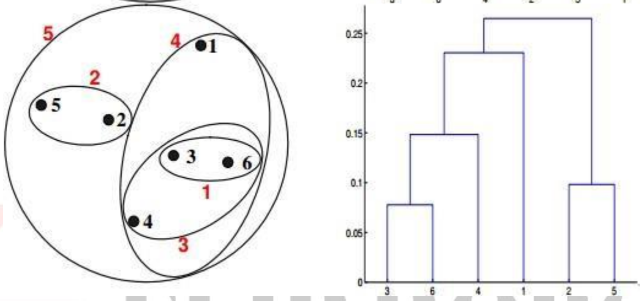
b.

c.



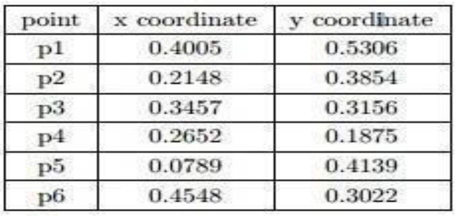
d.





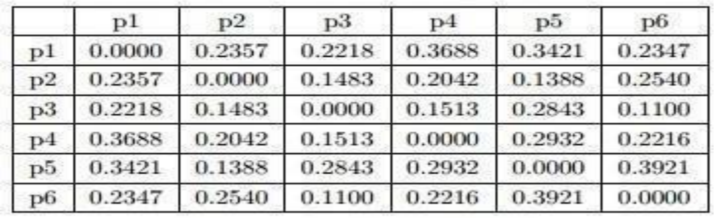
Answer – C

12. Given, six points with the following attributes:



point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table : X-Y coordinates of six points.

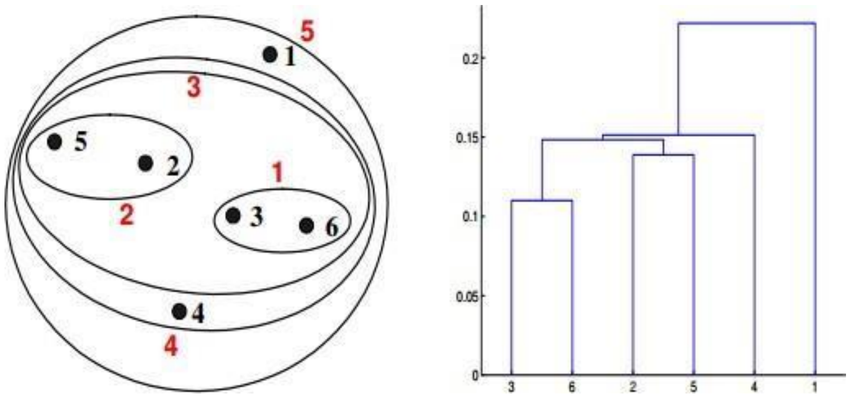


	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

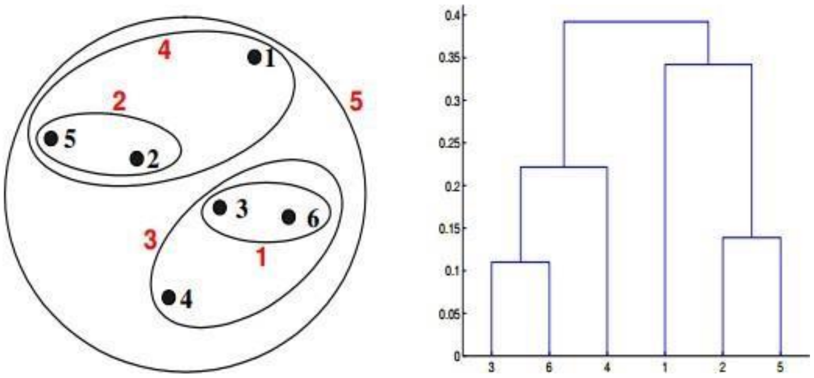
Table : Distance Matrix for Six Points

Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering.

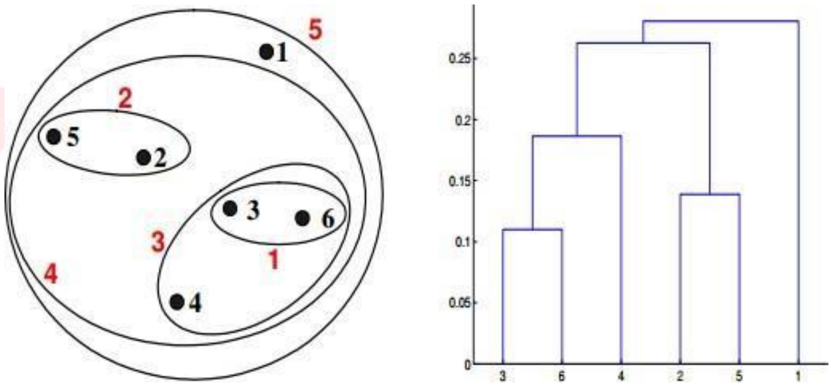
a.



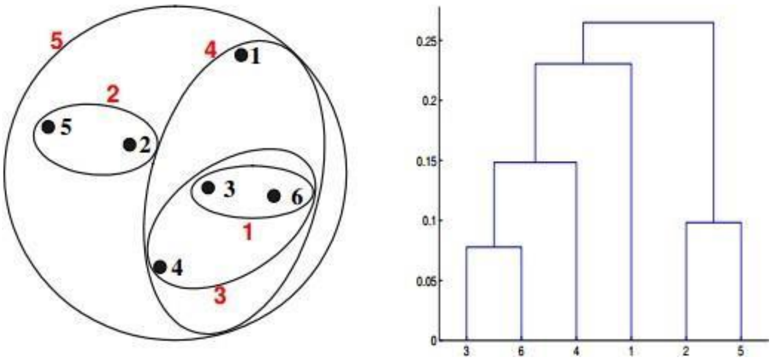
b.



c.



d.



Answer – D

MACHINE LEARNING

Q13 to Q14 are subjective answers type questions, Answers them in their own words briefly

13. What is the importance of clustering?

- **Increased resource availability: If one Intelligence Server in a cluster fails, the other Intelligence Servers in the cluster can pick up the workload. This prevents the loss of valuable time and information if a server fails.**
- **Strategic resource usage: You can distribute projects across nodes in whatever configuration you prefer. This reduces overhead because not all machines need to be running all projects, and allows you to use your resources flexibly.**
- **Increased performance: Multiple machines provide greater processing power.**
- **Greater scalability: As your user base grows and report complexity increases, your resources can grow.**
- **Simplified management: Clustering simplifies the management of large or rapidly growing systems.**

14. How can I improve my clustering performance?

- **Graph-based clustering performance can easily be improved by applying ICA blind source separation during the graph Laplacian embedding step.**
- **Applying unsupervised feature learning to input data using either RICA or SFT, improves clustering performance.**