PROJECT NAME:-

WORKSHEET SET 3

SUBMMITED BY:-

ATUL DAHIMA

MACHINE LEARNING ASSIGNMENT 3

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

Ans:- D) All of the above

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

Ans:- D) None

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

Ans:- C) Reinforcement learning and unsupervised learning

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

Ans:- B) The tree representing how close the data points are to each other

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

Ans:- D) None

6. Which is the following is wrong?

e. k-means clustering is a vector quantization method
f. k-means clustering tries to group n observations into k clusters
g. k-nearest neighbour is same as k-means
h. None

Ans:- 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

Ans:- D) 1,2 and 3

8. Which of the following are true?

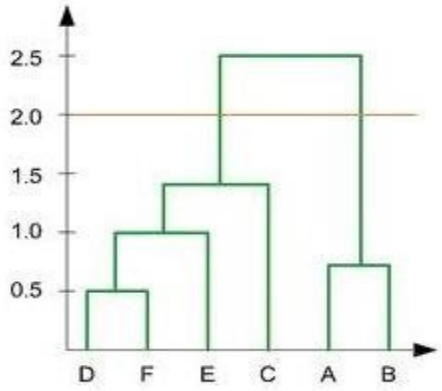
i. Clustering analysis is negatively affected by multicollinearity of features
ii. Clustering analysis is negatively affected by heteroscedasticity

Options:

a. 1 only
b. 2 only
c. 1 and 2
d. None of them

Ans:- 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 clustersformed?



i. 2
j. 4
k. 3
l. 5

Ans:- i) 2

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

m. Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.
n. Given a database of information about your users, automatically group them into different market segments.
o. 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

Ans:- B) given a database of information about your users , automatically group them into different market segments

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

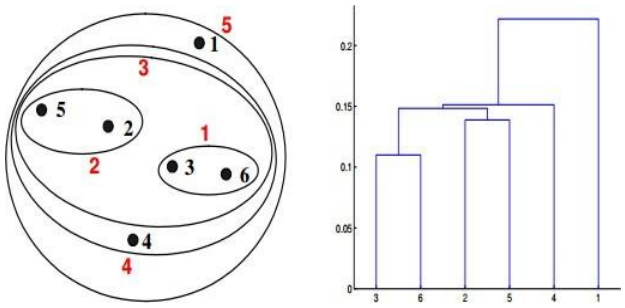
Table : X-Y coordinates of six points.

	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

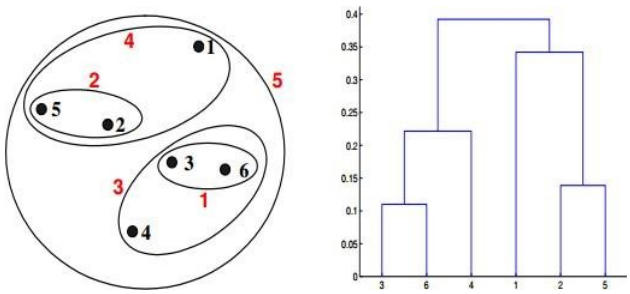
Table : Distance Matrix for Six Points

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

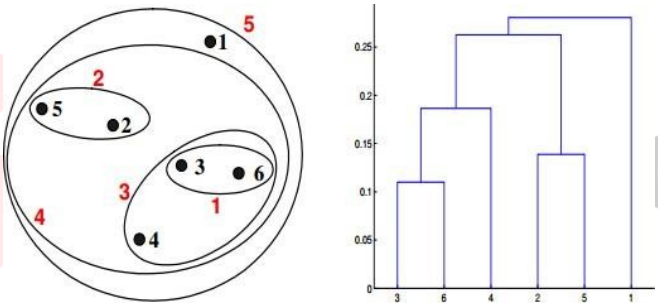
a.



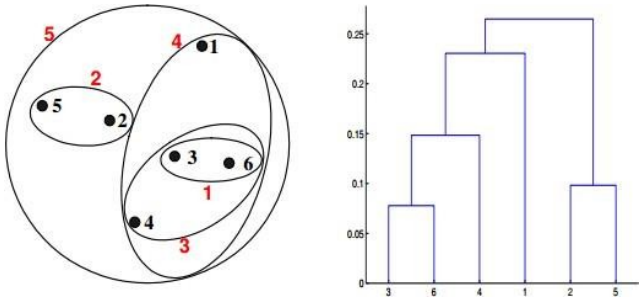
b.



c.



d.



Ans:- B

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

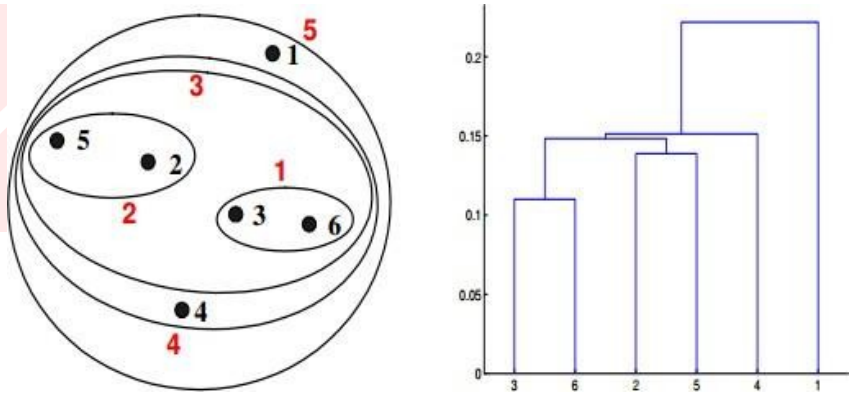
Table : X-Y coordinates of six points.

	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

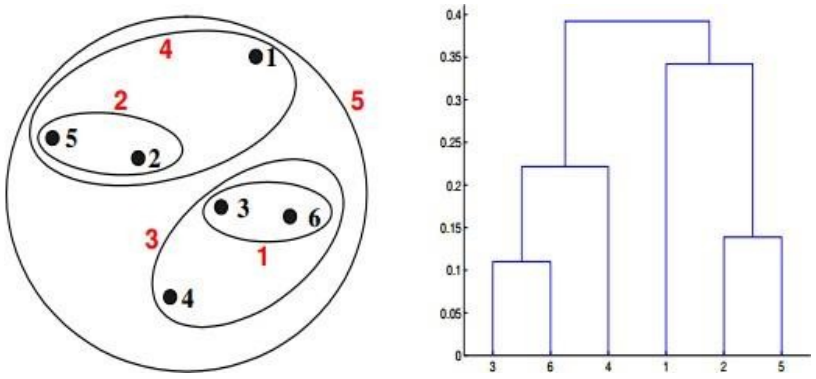
Table : Distance Matrix for Six Points

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

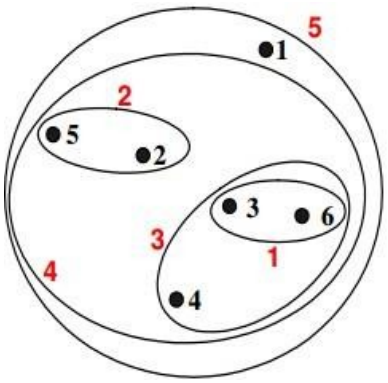
a.

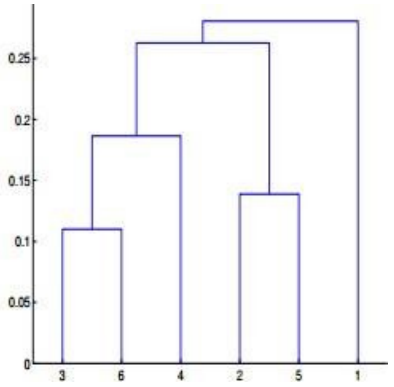


b.

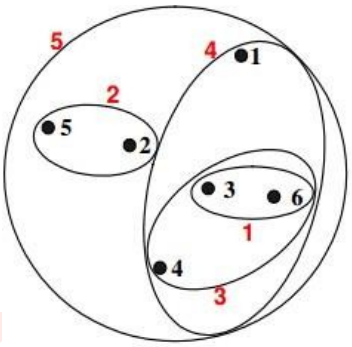


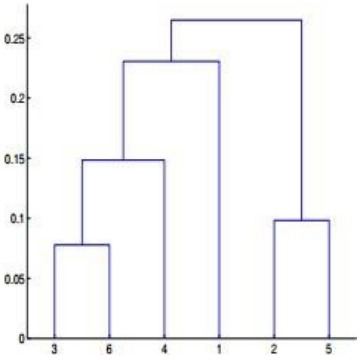
c.





d.





Ans:- D

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

13.What is the importance of clustering?

Ans:- clustering is used to gain important insights from data by observing what groups (or clusters) the data points fall into when they apply a clustering algorithm to the data. By definition, unsupervised learning is a type of machine learning that searches for patterns in a data set with no pre-existing labels and a minimum of human intervention. Clustering can also be used for anomaly detection to find data points that are not part of any cluster, or outliers.

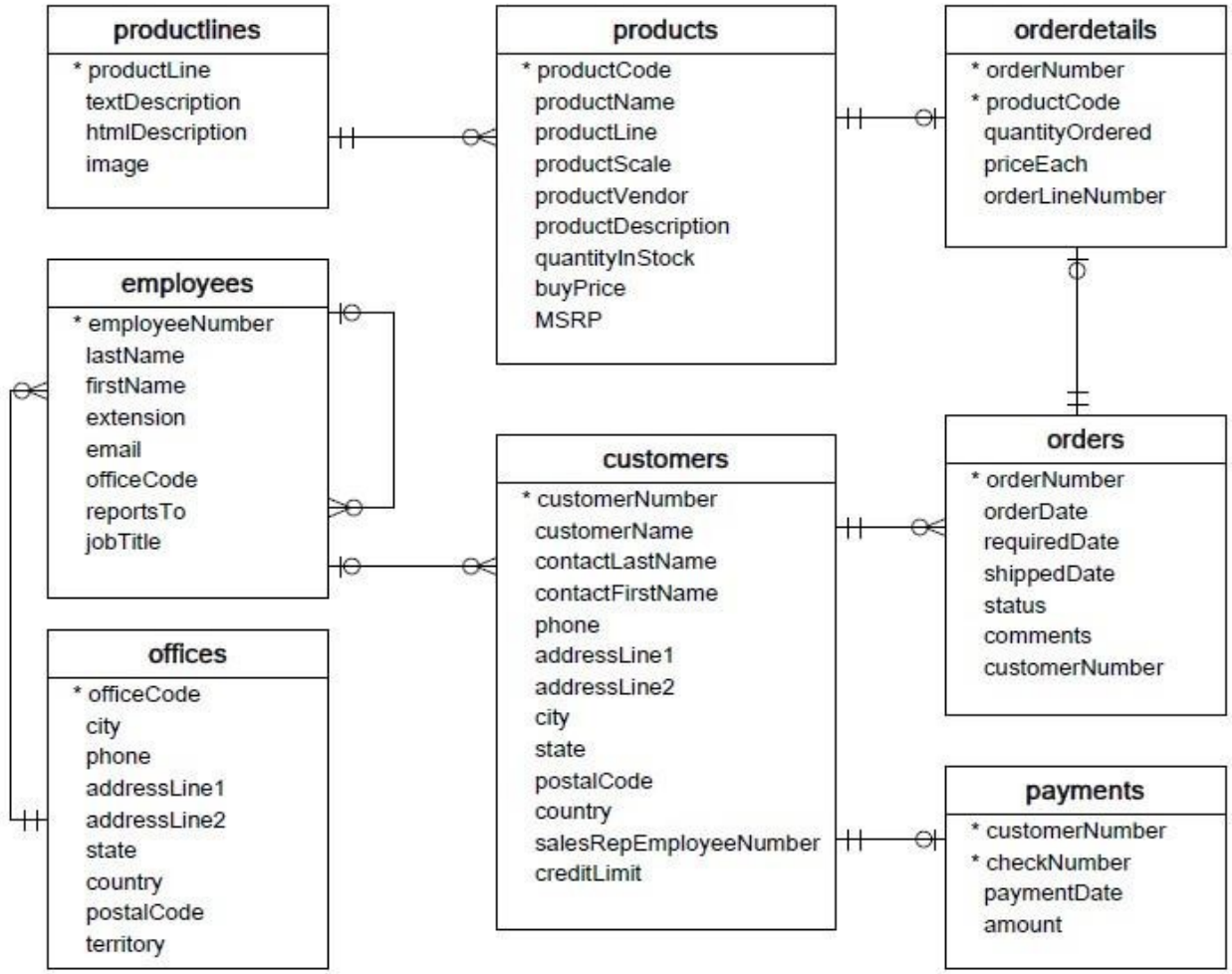
Clustering is used to identify groups of similar objects in datasets with two or more variable quantities. In practice, this data may be collected from marketing, biomedical, or geospatial databases, among many other places.

14. How can I improve my clustering performance?

Ans:-

WORKSHEET 3
SQL

Refer the following ERD and answer all the questions in this worksheet. You have to write the queries using mysql for the required Operation.



- **Customers**: stores customer's data.
- **Products**: stores a list of scale model cars.
- **Product Lines**: stores a list of product line categories.
- **Orders**: stores sales orders placed by customers.
- **Order Details**: stores sales order line items for each sales order.
- **Payments**: stores payments made by customers based on their accounts.
- **Employees**: stores all employee information as well as the organization structure such as who reports to whom.
- **Offices**: stores sales office data.

1. Write SQL query to create table **Customers.**

Ans:- CRETAE TABLE customers

(

customerNumber int NOT NULL,

customerName char(20) NOT NULL,

contactLastName char(20) NOT NULL,

….

)

2. Write SQL query to create table **Orders.**

Ans:- CRETAE TABLE orders

```
(
   orderNumber int NOT NULL,
   orderDate int NOT NULL,
    requiredDate int NOT NULL,
    ….
)
```

3. Write SQL query to show all the columns data from the **Orders** Table.

Ans:-

4. Write SQL query to show all the comments from the **Orders** Table.

5. Write a SQL query to show orderDate and Total number of orders placed on that date, from **Orders** table.

6. Write a SQL query to show employeNumber, lastName, firstName of all the employees from **employees** table.

7. Write a SQL query to show all orderNumber, customerName of the person who placed the respective order.

8. Write a SQL query to show name of all the customers in one column and salerepemployee name inanothercolumn.

9. Write a SQL query to show Date in one column and total payment amount of the payments made on that datefrom the **payments** table.

10. Write a SQL query to show all the products productName, MSRP, productDescription from the **products** table.

Write a SQL query to print the productName, productDescription of the most ordered product.

11. Write a SQL query to print the city name where maximum number of orders were placed.

12. Write a SQL query to get the name of the state having maximum number of customers.

13. Write a SQL query to print the employee number in one column and Full name of the employee in thesecond column for all the employees.

14. Write a SQL query to print the orderNumber, customer Name and total amount paid by the customer for thatorder (quantityOrdered × priceEach).

STATISTICS WORKSHEET-3

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is the correct formula for total variation?

a) Total Variation = Residual Variation – Regression Variation

b) Total Variation = Residual Variation + Regression Variation

c) Total Variation = Residual Variation * Regression Variation

d) All of the mentioned

Ans:- B) total variation= residual variation+regression variation

2. Collection of exchangeable binary outcomes for the same covariate data are called outcomes.

a) random

b) direct

c) binomial

d) none of the mentioned

Ans:- C) binomial

3. How many outcomes are possible with Bernoulli trial?

a) 2

b) 3

c) 4

d) None of the mentioned

Ans:- A) 2

4. If Ho is true and we reject it is called

a) Type-I error

b) Type-II error

c) Standard error

d) Sampling error

Ans:- A) Type 1 error

5. Level of significance is also called:

a) Power of the test

b) Size of the test

c) Level of confidence

d) Confidence coefficient

Ans:- C) Level of confidence

6. The chance of rejecting a true hypothesis decreases when sample size is:

a) Decrease

b) Increase

c) Both of them

d) None

Ans:- B) increase

7. Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Ans:- B) hypothesis

8. What is the purpose of multiple testing in statistical inference?

a) Minimize errors

b) Minimize false positives

c) Minimize false negatives

d) All of the mentioned WORKSHEET

Ans:- D) all of the mentioned WORKSHEET

9. Normalized data are centred at and have units equal to standard deviations of the original data

a) 0

b) 5

c) 1

d) 10

Ans:- A) 0

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What Is Bayes' Theorem?

Ans:- In statistics and probability theory, the Bayes' theorem (also known as the Bayes' rule) is a mathematical formula used to determine the conditional probability of events. Essentially, the Bayes' theorem describes the probability of an event based on prior knowledge of the conditions that might be relevant to the event.

11. 11. What is z-score?

Ans:- A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean.

It is also known as a standard score, because it allows comparison of scores on different kinds of variables by standardizing the distribution. A standard normal distribution (SND) is a normally shaped distribution with a mean of 0 and a standard deviation (SD) of 1

12. What is t-test?

Ans:- A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances. A t-test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population.

13. What is percentile?

Ans:- In statistics, percentiles are used to understand and interpret data. The *n*th percentile of a set of data is the value at which *n* percent of the data is below it. In everyday life, percentiles are used to understand values such as test scores, health indicators, and other measurements. For example, an 18-year-old male who is six and a half feet tall is in the 99th percentile for his height. This means that of all the 18-year-old males, 99 percent have a height that is equal to or less than six and a half feet. An 18-year-old male who is only five and a half feet tall, on the other hand, is in the 16th percentile for his height, meaning only 16 percent of males his age are the same height or shorter.

14. What is ANOVA?

Ans:- Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

15. How can ANOVA help?

Ans:- ANOVA is helpful for **testing three or more variables**. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources