FLIP ROBO

PROJECT NAME

WORKSHEET ON
MACHINE LEARNING
SQL AND
STATISTICS

SUBMITTED BY

VARSHA.VASANT.SHINDE

INDEX

WORKSHEET-3 ML

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

Q1. Which of the following is an application of clustering?

a. Biological network analysis
b. Market trend prediction
c. Topic modelling
d. All of the above
ANSWER: - d. All of the above

Q2. On which data type, we cannot perform cluster analysis?

a. Time series data
b. Text data
c. Multimedia data
d. None
ANSWER: - d. None

Q3. Netflix's movie recommendation system uses-

a. Supervised learning
b. Unsupervised learning
c. Reinforcement learning and Unsupervised learning
d. All of the above
ANSWER: - Reinforcement learning

Q4. 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
ANSWER: - b. The tree representing how close the data points are to each other.

Q5. 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
ANSWER: - d. None

Q6. 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
ANSWER: - c. k-nearest neighbour is same as k-means

Q7. 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

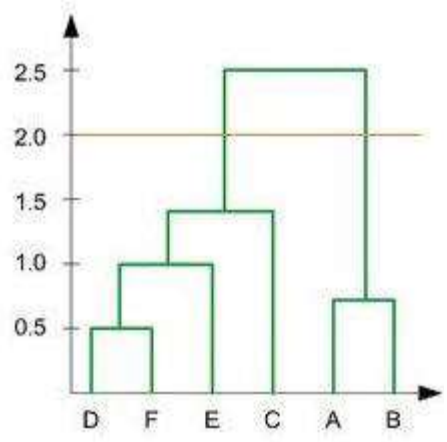
Options:
a. 1 and 2
b. 1 and 3
c. 2 and 3
d. 1, 2 and 3
ANSWER: - d. 1, 2 and 3

Q8. 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

Q9. 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

Q10. For which of the following tasks might clustering be a suitable approach?

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: - b and c

Q11. 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

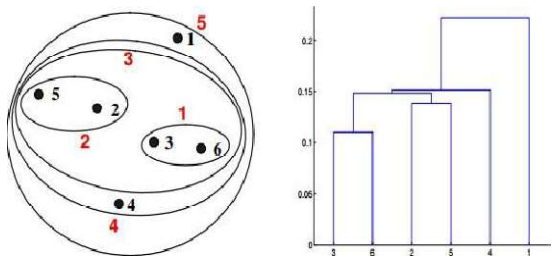
Table : X-Y coordinates of six points.

	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

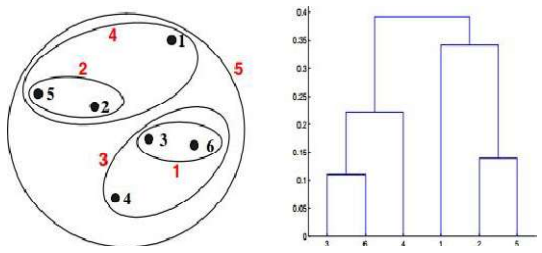
Table : Distance Matrix for Six Points

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

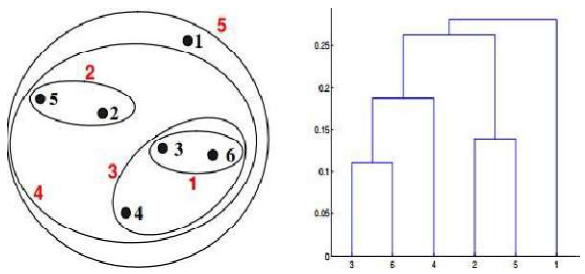
A.



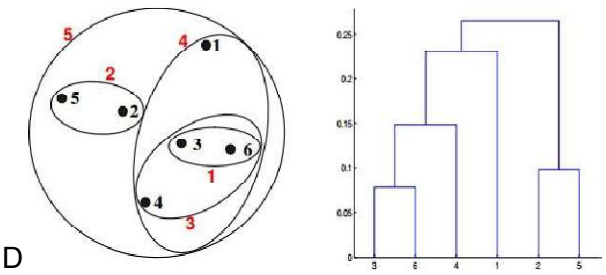
B.



C.



D



ANSWER :- A

Q12. 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

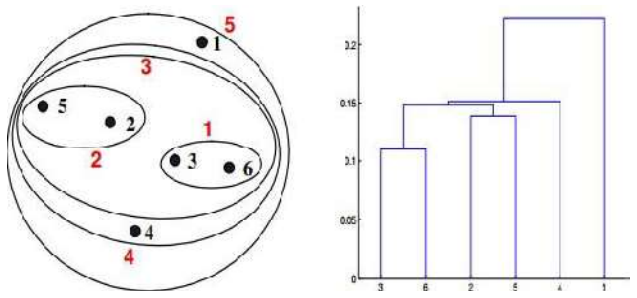
Table : X-Y coordinates of six points.

	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

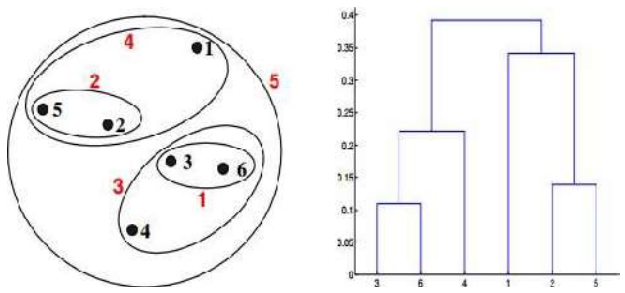
Table : Distance Matrix for Six Points

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

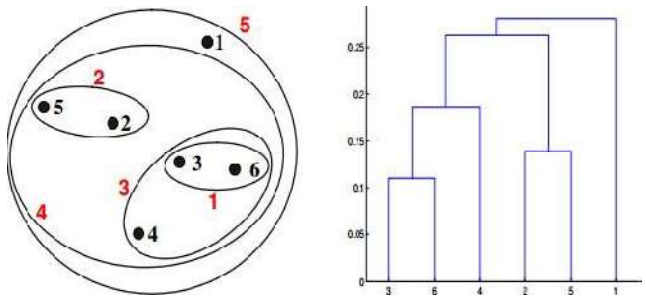
A.



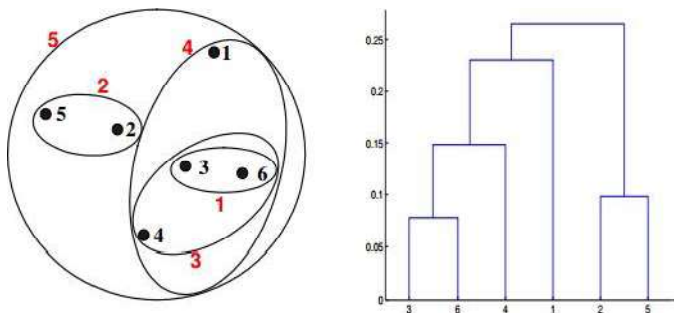
B.



C.



D.



ANSWER: - B

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

13. What is the importance of clustering?

ANSWER: -

- This clustering method helps grouping valuable data into clusters and picks appropriate results based on different techniques.
- For example, in information retrieval, the results of the query are grouped into small clusters, and each cluster has irrelevant results. By Clustering techniques, they are grouped into similar categories, and each category is subdivided into sub-categories to assist in the exploration of queries output.

- There are various types of clustering methods; they are

- Hierarchical methods
- Partitioning methods
- Density-based
- Model-based clustering
- Grid-based model

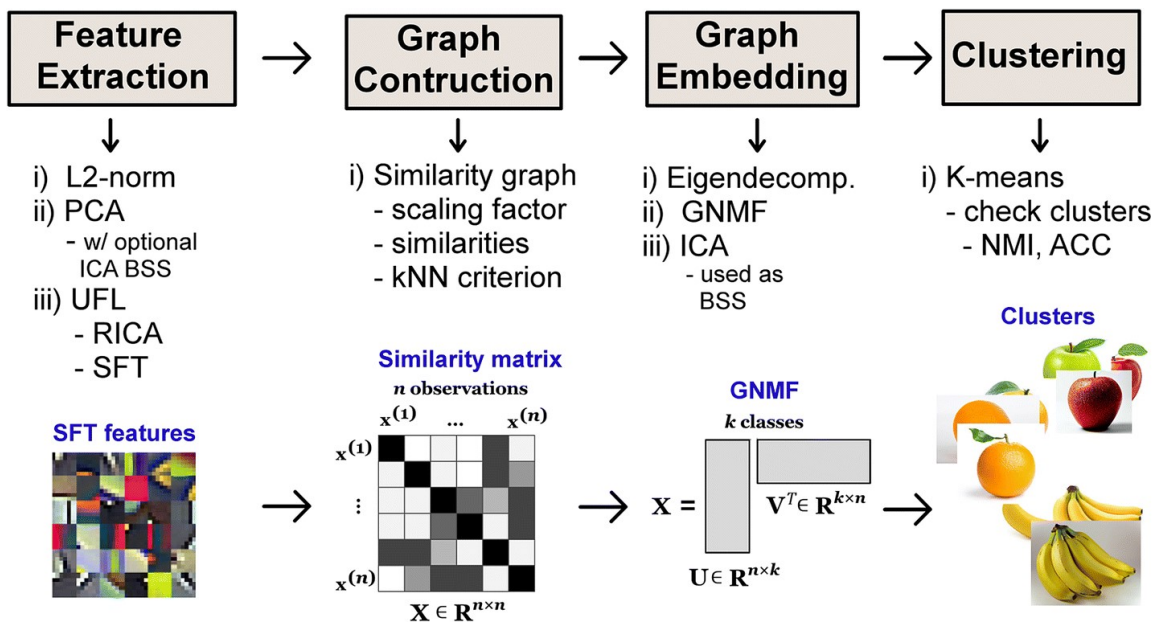
14. How can I improve my clustering performance?
ANSWER: -

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.

The clustering pipeline consists of four key components:

- feature extraction,
- graconstruction,
- graph embedding, and
- K-means clustering.

In the following, the datasets are first described and then the four components are introduced.



WORKSHEET-3 STATISTICS

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

Q1. 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
ANSWER:- b) Total Variation = Residual Variation + Regression Variation

Q2. Collection of exchangeable binary outcomes for the same covariate data are called outcomes.

a) Random
b) Direct
c) Binomial
d) None of the mentioned
ANSWER:- c) Binomial

Q3. How many outcomes are possible with Bernoulli trial?

a) 2
b) 3
c) 4
d) None of the mentioned
ANSWER:- d) None of the mentioned

Q4. If Ho is true and we reject it is called

a) Type-I error
b) Type-II error
c) Standard error
d) Sampling error
ANSWER:- a) Type-I error

Q5. Level of significance is also called:

a) Power of the test
b) Size of the test
c) Level of confidence
d) Confidence coefficient
ANSWER:- c) Level of confidence

Q6. The chance of rejecting a true hypothesis decreases when sample size is:

a) Decrease
b) Increase
c) Both of them
d) None
ANSWER:- b) Increase

Q7. Which of the following testing is concerned with making decisions using data?

a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned
ANSWER:- b) Hypothesis

Q8. 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
ANSWER:- d) All of the mentioned

Q9. Normalized data are centred at and have units equal to standard deviations of the original data

a) 0
b) 5
c) 1
d) 10
ANSWER:- a) 0

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

Q10. What Is Bayes' Theorem?
ANSWER:-

Bayes' theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of “causes”.

Q11. What is z-score?
ANSWER:-

- A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of **standard deviations** from the mean.
- If a Z-score is 0, it indicates that the data point's score is identical to the mean score.
- A Z-score of 1.0 would indicate a value that is one standard deviation from the mean.
- Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.
- A Z-Score is a statistical measurement of a score's relationship to the mean in a group of scores.
- A Z-score can reveal to a trader if a value is typical for a specified data set or if it is atypical.
- In general, a Z-score below 1.8 suggests a company might be headed for bankruptcy, while a score closer to 3 suggests a company is in solid financial positioning.

Q12. What is t-test?
ANSWER:-
A t-test is an inferential statistic used to determine if there is a significant difference between the means of two groups and how they are related. T-tests are used when the data sets follow a normal distribution and have unknown variances, like the data set recorded from flipping a coin 100 times.

The t-test is a test used for hypothesis testing in statistics and uses the t-statistic, the t-distribution values, and the degrees of freedom to determine statistical significance.

- A t-test is an inferential statistic used to determine if there is a statistically significant difference between the means of two variables.
- The t-test is a test used for hypothesis testing in statistics.
- Calculating a t-test requires three fundamental data values including the difference between the mean values from each data set, the standard deviation of each group, and the number of data values.
- T-tests can be dependent or independent.

13. What is percentile?
ANSWER:-

- “Percentile” is in everyday use, but there is no universal definition for it.
- The most common definition of a percentile is a number where **a certain percentage of scores fall below that number.**
- You might know that you scored 67 out of 90 on a test.
- But that figure has no real meaning unless you know what percentile you fall into.
- If you know that your score is in the 90th percentile, that means you scored better than 90% of people who took the test.

Q14. What is ANOVA?
ANSWER:-

- 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.

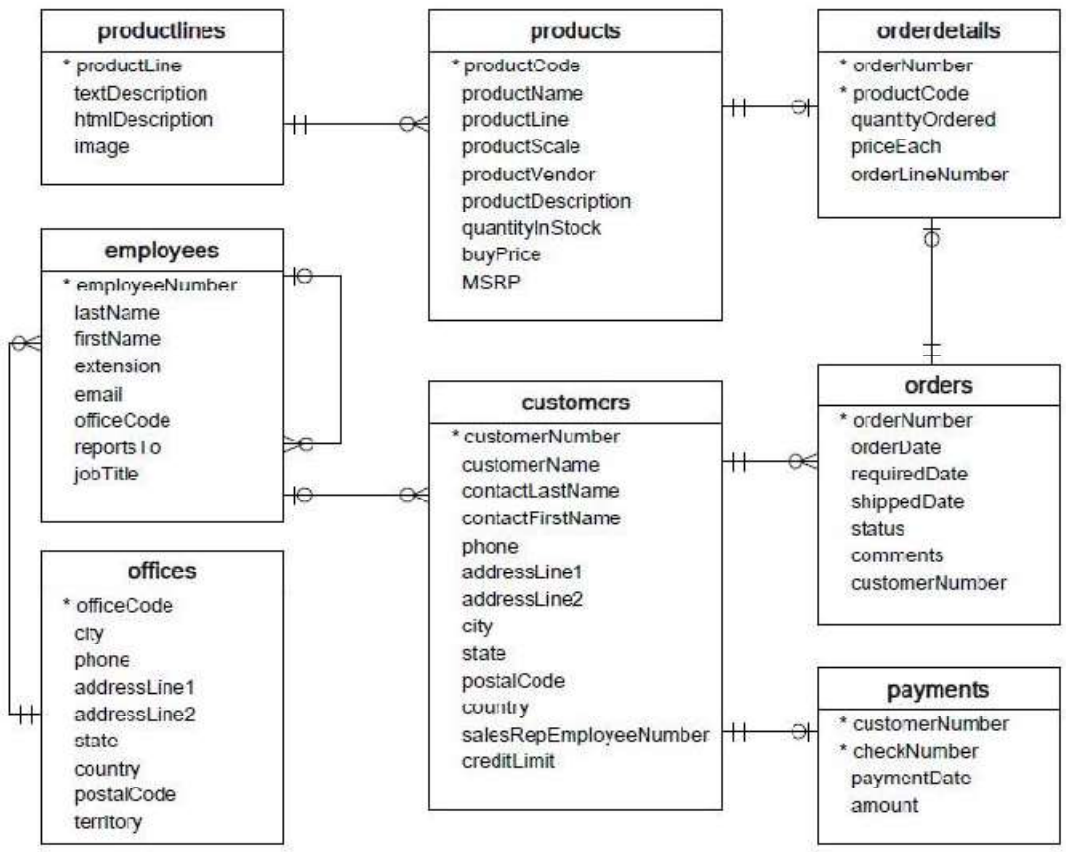
- The systematic factors have a statistical influence on the given data set, while the random factors do not.
- Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.
- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.
- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.

Q15. How can ANOVA help?
ANSWER:-

- 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.

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.

Q1. Write SQL query to create table **Customers.**
ANSWER:-

```
Create Table CUSTOMER1 (
    customerNumber int (20),
    CustomerName varchar(25),
    contactLastName varchar(50),
    ContactFirstName varchar (50),
    Phone int (20),
    AddressLine1 varchar (50),
    AddressLine2 varchar (50),
    City varchar (50),
    State varchar (50),
    Postal Code int (50),
    Country varchar (50),
    SalesRepEmployeeNumber int (50),
    Credit Limit varchar (50)
    );
```

Q2. Write SQL query to create table **Orders.**
ANSWER:-

```
CREATE TABLE ORDER  (
      orderNumber int ,
      orderDate int ,
      requiredDate int,
      shippedDate int,
      status varchar (20),
     comments varchar(50),
     customerNumber int
     );
```

Q3. Write SQL query to show all the columns data from the **Orders** Table.
ANSWER:-

```
SELECT * FROM ORDER;
```

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

```
SELECT comments FROM ORDER1;
```

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

ANSWER:-

SELECT orderDate, shippedDate FROM ORDER;

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

ANSWER:-

SELECT employeNumber, lastName, firstName FROM EMPLOYEE;

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

ANSWER:-

SELECT orderNumber, customerName FROM ORDER;

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

ANSWER:-

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

ANSWER:-

SELECT sum Date, total amount FROM PAYMENT;

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

ANSWER:-

SELECT productName, MSRP, productDescription FROM PRODUCT;

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

ANSWER:-

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

ANSWER:-

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

ANSWER:-

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

ANSWER:-

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

ANSWER:-