MACHINE LEARNING WORKSHEET

1 – D

2 – D

3 – C

4 – B

5 – D

6 – C

7 – D

8 – A

9 – A

10 – B

11 – A

12 – B

13 – Importance of Clustering are as follows:

- Having clustering methods helps in restarting the local search procedure and remove the inefficiency. In addition, clustering helps to determine the internal structure of the data.
- This clustering analysis has been used for model analysis, vector region of attraction.
- Clustering helps in understanding the natural grouping in a dataset. Their purpose is to make sense to partition the data into some group of logical groupings.
- Clustering quality depends on the methods and the identification of hidden patterns.
- They play a wide role in applications like marketing economic research and weblogs to identify similarity measures, Image processing, and spatial research.
- They are used in outlier detections to detect credit card fraudulence.

14 - K-means clustering algorithm can be significantly improved by using a better initialization technique, and by repeating (re-starting) the algorithm or by avoiding unbalanced cluster size.

STATISTICS WORKSHEET

1. B
2. C
3. A
4. A
5. B
6. A
7. B
8. D
9. A
10. 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". It is given by the formula

    P(A|B) = P(A) P(B|A)/P(B) where P(B) != 0
11. Z-score indicates how much a given value differs from the standard deviation. The Z-score, or standard score, is the number of standard deviations a given data point lies above or below mean. Standard deviation is essentially a reflection of the amount of variability within a given data set. It is given by the formula:

    $z = (x-\mu)/\sigma$ where x is the raw score, μ is the population mean, and σ is the population standard deviation
12. A t-test is **a statistical test that compares the means of two samples**. It is used in hypothesis testing, to determine whether a process actually has an effect on population of interest or whether two groups are different from one another.
13. It is a value on a scale of 100 that indicates the percent of a distribution that is equal to or below it.
14. 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. The one-way ANOVA can help you know whether or not there are significant differences between the means of your independent variables. When we understand how each independent variable's mean is different from the others, we can begin to understand which of them has a connection to your dependent variable, and begin to learn what is driving that behavior.

WORKSHEET 3 SQL

1. import sqlite3

   db = sqlite3.connect('Customers.db')

   cur = db.cursor()

   cur.execute('create table Customers(customerNumber int primary key,customerName varchar (30),contactLastName varchar (20),contactFirstName varchar (20), phone int, addressLine1 varchar (40), addressLine2 varchar (40), city varchar (20) ,state varchar (20),postalCode int,country varchar (20),salesRepEmployeeNumber varchar (20),creditLimit float)')

2. db = sqlite3.connect('orders.db')

   cur = db.cursor()

   cur.execute('create table orders (orderNumber varchar (20) primary key,orderDate int, requiredDate int, shippedDate int, status varchar (30), comments varchar (30), customerNumber varchar (20))')

3. results = cur.execute("select * from orders")

   results.fetchall()

4. result = cur.execute("Select comments from orders")

   for i in result:

   print(i)

5. order = cur.execute(" SELECT orderDate, COUNT(orderNumber) from orders GROUP BY orderDate '")

   for i in order:

   print(i)

6. results = cur.execute("Select employeNumber,lastName, firstName from employees")

   for row in results:

      print(row)

7. cur.execute(" SELECT orderNumber,customerName FROM orders o JOIN customers c ON o.customerNumber = c.customerNumber ")

8. cur.execute("SELECT c.customerName, e.firstName FROM customers c JOIN employees e ON c.salesRepEmployeeNumber=e.employeeNumber")

9. cur.execute("SELECT paymentDate,SUM(amount) FROM payments GROUP BY paymentDate")

10. cur.execute("SELECT productName,MSRP, productDescription FROM products")

11. cur.execute("SELECT p.productName, p.productDescription, COUNT(od.productCode)*od.quantityOrdered FROM products p LEFT JOIN orderdetail od ON p.productCode = od.productCode GROUP BY od.productCode ORDER BY COUNT(od.productCode)*od.quantityOrdered DESC LIMIT 1")

12. cur.execute("SELECT c.city FROM customers c JOIN orders o ON c.customerNumber = o.customerNumber GROUP by c.customerNumber ORDER BY COUNT(o.customerNumber) DESC LIMIT 1")

13. cur.execute("SELECT state,COUNT(state) FROM customers GROUP BY state ORDER BY COUNT(state) DESC LIMIT 1")

14. cur.execute("SELECT employeeNumber,concat(firstName,' ',lastName) FROM employees")

15. cur.execute("SELECT o.orderNumber, c.customerName,SUM(od.quantityOrdered*od.priceEach) FROM customers c JOIN orders o ON c.customerNumber = o.customerNumber JOIN orderdetail od ON od.orderNumber = o.orderNumber GROUP BY o.orderNumber")