Machine Learning Assignment- 3

A1) option d- All of the above

A2) option d- None

A3) option c- Reinforcement learning and Unsupervised learning

A4) option b- The tree representing how close the data points are to each other

A5) option d- None

A6) option c- k-nearest neighbour is same as k-means

A7) option d- 1,2 and 3

A8) option a- i) only

A9) option a- 2

A10) option b- Given a database of information about your users, automatically group them into different market segments.

A11) option a

A12) option b

A13) Importance of Clustering:

Clustering is important because of following reasons-

1) it determines the intrinsic grouping among the unlabelled data present
2) it is an essential component and makes life so much easier in creating new machine learning methods
3) 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.

4) This clustering analysis has been used for model analysis, vector region of attraction.
5) 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.
6) Clustering quality depends on the methods and the identification of hidden patterns.
7) They play a wide role in applications like marketing economic research and weblogs to identify similarity measures, Image processing, and spatial research.
8) They are used in outlier detections to detect credit card fraudulence.

A14) How to improve clustering performance-

We can opt for better and more powerful algorithms in order to improve clustering performance such as k-means++. We can also improve the performance using feature weights learning.

SQL Worksheet- 3

```
A1) CREATE TABLE Customers (
  customerNumber int NOT NULL,
customerName varchar,
contactLastName varchar,
contactFirstName varchar,
phone int,
addressline1 varchar,
addressline2 varchar,
city varchar,state varchar,
postalCode int,
Country varchar,
salesRepEmployeeNumber int,
creditLimit int,
PRIMARY KEY (customerNumber)
FOREIGN KEY (salesRepEmployeeNumber) REFERENCE
employees(employeeNumber)


);
```

```
A2) CREATE TABLE Orders(
orderNumber int NOT NULL,
orderDate DATE,
requiredDate DATE,
shippedDate DATE,

status varchar,

comment varchar,

customerNumber int,

PRIMARY KEY (orderNumber),

FOREIGN KEY (customerNumber) REFERENCE Customers(customerNumber)

);
```

```
A3) select * from Orders;
```

```
A4) select comment from Orders;
```

```
A5) SELECT orderDate, COUNT(orderNumber) FROM Orders GROUP BY
orderDate;
```

```
A6) SELECT employeNumber, lastName, firstName FROM employees;
```

```
A7) SELECT o.orderNumber, c.customerName FROM orders o
LEFT JOIN customers c ON c.customerNumber = o.customerNumber;
```

```
A8) SELECT c.customerName , e.firstName FROM customers c
LEFT JOIN employees e ON c. salesRepEmployeeNumber =
e.employeeNumber;
```

```
A9) SELECT paymentDate, COUNT(amount) FROM payments GROUP BY
paymentDate;
```

```
A10) SELECT productName, MSRP, productDescription FROM products;
```

```
A11) SELECT p.productName, p.productDescription FROM orderdetails AS o
   INNER JOIN products AS p ON o.productCode = p.productCode
  GROUP BY o.productCode
  ORDER BY SUM(o.quantityOrdered) DESC LIMIT 3
```

```
A12)  SELECT c.city

FROM Customers AS c LEFT JOIN orders AS o ON
c.customerNumber=o.customerNumber GROUP BY c.city ORDER BY COUNT(*);
```

```
A13) SELECT state FROM Customers GROUP BY state ORDER BY 2 DESC;
```

```
A14) SELECT employeeNumber, CONCAT(firstName, lastName) AS fullName
FROM employees;
```

```
A15)   SELECT o.orderNumber, c.customerName,
SUM(ord.quantityOrdered*ord.priceEach) AS
    Total_Amount FROM Customers c inner join orders o
    on c.CustomerNumber = o.CustomerNumber join
    orderdetails ord on o.orderNumber = ord.orderNumber
    GROUP BY o.customerNumber;
```

STATISTICS WORKSHEET 3

A1) option b- Total Variation = Residual Variation + Regression Variation
A2) option c- binomial
A3) option a- 2
A4) option a- Type I error
A5) option b- Size of the test
A6) option a- decreases
A7) option b-Hypothesis
A8) option d- All of the mentioned
A9) option a-0

A10) Bayes Theorem- It is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome having occurred in similar circumstances. Bayes' theorem provides a way to revise existing predictions or theories or update probabilities given new or additional evidence.

A11) z-score- A z-score is also known as standard score. It gives an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is.

A12) t-test- A t-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually influences the population of interest, or whether two groups are different from one another.
t-test makes some assumptions about data and these assumptions are:

1. Data is independent
2. Data is normally distributed approximately.
3. Data has a similar amount of variance within each group being compared

A13) Percentile- A percentile is a comparison score between a particular score and the scores of the rest of a group. It shows the percentage of scores that a particular score surpassed.

A14) ANOVA- 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. If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

A15) ANOVA is helpful for testing three or more variables. If we have to compare variance between more than two groups than we use ANOVA. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.