(3 Hours) [Total Marks: 80]

Note: 1. Question no.1 is compulsory.
2. Attempt any three out of remaining five.
3. Assumptions made should be clearly indicated.
4. Figures to the right indicates full marks.
5.Assume suitable data whenever necessary.

Q. 1 Solve any four. **(20)**

A Every data structure in the data warehouse contains the time element. Why?

B In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

C What are the various methods for estimating a classifier's accuracy?

D Explain market basket analysis with an example.

E Describe K medoids algorithm.

F Explain CLARANS extension in web mining.

Q. 2 A Consider the quarterly sales of four companies C1, C2, C3.C4. The dimensions are
a) Time
b) Shopping category (Men's, Women's, Electronics, Home)
c) Company
Create a cube and describe all five OLAP operation. **(10)**

B Apply the Naïve Bayes classifier to classify the tuple <Red, SUV, Domestic> For the given dataset below. **(10)**

Instance no.	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	sports	Imported	Yes

Q.3 A Discuss the different types of attributes. **(10)**

B Suppose that the data mining task is to cluster the following points into 3 clusters .A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9).The distance function is Euclidean distance .Suppose we initially assign A1,B1,C1 as the center of each cluster respectively, Use the k means algorithm to show only a) the three cluster centers after the first round of execution b) The final three clusters. **(10)**

Q.4 A For a supermarket chain, consider the dimensions namely Product, Store, time,promotion. The schema contains the three facts namely units_sales, dollar_sales, and cost_ dollars.
Design a star schema and calculate the maximum number of base fact table records for the values given below:
Time period: 5 years
Stores: 300 reporting daily sales
Product: 40000 products in each store (about 4000 sell daily in each store
Promotion: a sold item may be in only one promotion in a store on a given day. **(10)**

B A database has five transactions. **(10)**

T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, K, I, E}

Let minimum support =3, Find all frequent itemsets using FP-growth algorithm.

Q.5 A What is web structure mining? Describe page ranking technique with the help of example. **(10)**

B Use agglomerative algorithm using the following data and plot a dendrogram using single link approach. The following figure contains sample data items indicating the distance between the elements. **(10)**

Item	E	A	C	B	D
E	0	1	2	2	3
A	1	0	2	5	3
C	2	2	0	1	6
B	2	5	1	0	3
D	3	3	6	3	0

Q. 6 A Apply apriori algorithm on the following dataset to find strong association rules. Minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%) **(10)**

Transaction ID	Items
T1	Hot dogs, Buns, Ketchup
T2	Hot dogs, Buns
T3	Hot dogs, Coke, Chips
T4	Coke, Chips
T5	Chips, Ketchup
T6	Hotdogs ,Coke, Chips

B Is Web mining different from classical data mining? Justify your answer. Describe types of web mining. **(10)**
