Duration: 3hrs **[Max Marks:80]**

N.B. : (1) Question No 1 is Compulsory.
(2) Attempt any three questions out of the remaining five.
(3) All questions carry equal marks.
(4) Assume suitable data, if required and state it clearly.

1 Attempt any FOUR **[20]**

a Differentiate between Agglomerative and Divisive clustering method **[5]**

b A dimension table is wide, the fact table deep. Explain **[5]**

c What data mining functionalities are required for disease detection system in healthcare domain. Think of kinds of patterns that can be mined. Can such pattern be generated alternatively by data query processing? **[5]**

d In real-world data, tuples with *missing values* for some attributes are common occurrences. Describe various methods for handling this problem. **[5]**

e Differentiate between OLTP and OLAP. **[5]**

2 a A database has five transactions as given below . Let min sup count =3 and min conf =70% find all frequent itemsets and strong association rules. **[10]**

TID	Items
10	1,3,4
20	2,3,5
30	1,2,3,5
40	2,5
50	1,3,5

b Suppose that a data warehouse consists of the three dimensions time, doctor and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit **[10]**
(i)Draw a star schema diagram for the above data warehouse.
(ii) Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2010?
(iii)To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).

X126Y56563CX126Y56563CX126Y56563CX126Y56563C

3 a Explain Data mining as a step in KDD. Give the architecture of typical data mining. **[10]**

b Why is entity – relationship modeling technique not suitable for data warehouse? How is dimensional modeling different? **[10]**

4 a Develop a model to predict the salary of college graduates with 10 years of work experience using linear regression. **[10]**

Year of experience (x)	Salary in $100 (y)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

b Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. **[10]**

(a) What is the *mean* of the data? What is the *median*?
(b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
(c) What is the *midrange* of the data?
(d) Can you find (roughly) the first quartile ($Q1$) and the third quartile ($Q3$) of the data?
(e) Give the *five-number summary* of the data.
(f) Show a *boxplot* of the data.

5 a The college wants to record the marks for the courses completed by students using the dimensions a) course b) student c)Time and measure of aggregate marks. **[10]**
Create a cube and describe following operations
i) Roll up ii) Drill Down iii)Slice iv) Dice

b Explain characteristics of data warehousing assuming AllElectronics store warehouse. **[10]**

X126Y56563CX126Y56563CX126Y56563CX126Y56563C

6 a Demonstrate Multidimensional and Multilevel association Rule mining with suitable example. **[10]**

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

X126Y56563CX126Y56563CX126Y56563CX126Y56563C