This question paper contains 4+2 printed pages]

Roll No. | | | | | | | | | | | |

S. No. of Question Paper : 800

Unique Paper Code : 234611 G

Name of the Paper : Data Mining

Name of the Course : B.Sc. (H) Computer Science

Semester : VI

Duration : 3 Hours Maximum Marks : 75

(Write your Roll No. on the top immediately on receipt of this question paper.)

All parts of Question 1 (Section A) are compulsory.

Parts of a question must be answered together. Attempt any *four* questions from Part B. *All* questions in Section B carry equal marks.

Section A

All questions are compulsory.

Marks : 35

1. (*a*) Define : 2

(*i*) Closed frequent itemset

(*ii*) Maximal frequent itemset.

P.T.O.

(*b*) Indicate whether the following activities can be identified as a Data Mining Task or a Database Query : 4

(*i*) Computing the total sales of a company.

(*ii*) Sorting a student database based on student identification numbers.

(*iii*) Predicting the future stock price of a company using historical records.

(*iv*) Predicting whether a particular credit card transaction is likely to be fraudulent or not.

(*c*) What is data preprocessing ? Give *three* reasons for preprocessing the data. 1+3

(*d*) Describe the following terms briefly : 2+2

(*i*) Nominal attributes

(*ii*) Outlier analysis.

(*e*) State the Apriori property. Also state the *two* major drawbacks of the Apriori approach. 2+2

(*f*) Distinguish between supervised and unsupervised learning.Give *one* example of each. 2+2

(*g*) What is data transformation ? Describe *two* strategies used for a data transformation. 1+3

(*h*) Define support and confidence for Association Rules. 2+2

(*i*) Explain briefly the five number summary of a data distribution ? 5

Section B

2. Consider the market basket database shown below :

TID	Items Bought
1.	{Milk, bread, Diapers}
2.	{Bread, Butter, Milk}
3.	{Milk, Bread, Butter, Diapers}
4.	Bread, Butter, Diapers}
5.	{Milk, Bread, Butter, Diaper}

(1) List all itemsets of size 2 with their support counts. 5

(2) List all itemsets of size 3 with their support counts. 5

3. Give diagrammatical representation of the steps involved in the knowledge discovery from data. Explain each step in brief. 10

P.T.O.

4. Consider the training examples shown in the following table for a classification problem : 5×2=10

1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

(*a*) Compute the Gini Index for the customer ID attribute.

(*b*) Compute the Gini Index for the Car Type Attribute.

(*c*) Compute the Gini Index for the shirt size attribute.

(*d*) Which is a better attribute Car Type or Shirt Size ?

(*e*) Explain why Customer Id should not be used as the attribute test condition even though it has the lowest GINI value.

5. (*a*) What is a decision tree ? How is it used for classification ? 4

(*b*) Write short notes on the following classifiers : 6

(*i*) k-nearest neighbour

(*ii*) Naive Bayes classifier.

6. (*a*) Given the following data points : 2, 4, 10, 12, 3, 20, 30, 11, 25. Assume $k = 3$, and initial means $\mu 1 = 2$, $\mu 2 = 4$ and $\mu 3 = 6$. Show the clusters obtained using K-means_algorithm after one iteration, and show the new means for the next iteration. 6

(*b*) What is hierarchical clustering ? What is a dendogram ? 4

P.T.O.

7. (*a*) Let X be a random variable, denoting age of the students. Considering a random sample of size $n = 20$:

X = (69, 74, 68, 70, 72, 67, 66, 70, 76, 68, 72, 79, 74, 67, 66, 71, 74, 75, 75, 76)

Find the mean, median and mode for X. 6

(*b*) What is Noise in data ? How does binning help to smooth the noisy data ? 2+2