[This question paper contains 5 printed pages.]

Sr. No. of Question Paper : 5714 F Your Roll No.....

Unique Paper Code : 234611

Name of the Paper : Data Mining (CSHT-616 (iv)

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

Semester : VI

Duration : 3 Hours Maximum Marks : 75

Instructions for Candidates

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. **All** parts of Question 1(Part A) are compulsory.
3. Parts of a question must be answered together.
4. Attempt any **four** questions from Part B.
5. **All** questions in Part B carry equal marks.

Section A

Q1.

a) Distinguish between supervised and unsupervised learning. Give examples of data mining tasks that fall under the purview of each. 2+2

b) Is searching using database queries same as using any data mining technique for knowledge discovery. Comment. 2

c) Let student_enrolment_number be a primary key having integer values. Why is it considered a nominal attribute rather than a numeric attribute during data mining. 2

d) List two limitations of *k-means* clustering algorithm. 2

P.T.O.

e) Define support and confidence of an association rule with formal notations. 1+2

f) Consider the following set of frequent 2-itemsets: 3

L2={ {1,2}{1,3} {1,5}{2,3}{2,4} {2,5} }

List all possible candidate 3-itemsets that will be generated.

g) What is the difference between noise and outliers? Are noise objects *always* outliers? Are outliers *always* noise objects? 3

h) What is decision tree pruning? Why is it required? 3

i) Give any one application each of Classification, Clustering and Association rule mining. 3

j) Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). 5

(1) Time in terms of AM or PM.

(2) Angle as measured in degrees between 0° and 360°.

(3) Bronze, Silver, and Gold medals as awarded at the Olympics.

(4) Height above sea level.

(5) ISBN Number of a Book

k) What is a confusion matrix? How would you compute the accuracy, false positive and false negative rates of a classifier from it? 5

Section B

Q2. a) Consider the market basket database shown below:

TID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Bread, Butter, Diapers}
7	{ Bread, Butter, Diapers}
8	{ Beer, Diapers}
9	{Milk, Bread, Butter, Diapers}
10	{ Beer, Cookies}

i. Find an itemset of size 2 that has the largest support. 2
ii. List all itemsets of size 3 and their support counts 2
iii. Generate all the association rules that can be generated from the itemsets {Bread, Butter, Milk } 3

b) How does partitioning improve the efficiency of Apriori algorithm ? 3

3.
a) Explain with an example the difference between maximal and closed frequent itemsets. 2

b) How does FP-Growth algorithm improve upon the drawbacks of Apriori algorithm. 3

P.T.O.

c) Construct the FP Tree for the following dataset. 5

TID	*Items bought*
100	{*f, a, c, d, g, i, m, p*}
200	{*a, b, c, f, l, m, o*}
300	{*b, f, h, j, o, w*}
400	{*b, c, k, s, p*}
500	{*a, f, c, e, l, p, m, n*}

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

Customer ID	Gender	Car Type	Shirt Size	Class
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) Describe two stopping conditions to terminate the construction of a decision tree. 4

b) Enumerate the strengths and weaknesses of following classifiers. 6

i) k-nearest neighbour

ii) Naïve Bayes classifier

6.

a) Why is clustering also called database segmentation? 2

b) Distinguish between the Agglomerative and Divisive clustering methods 3

c) Cluster the following data into three clusters using agglomerative clustering method
Data: {2, 4, 10, 12, 3} 5

7.

a) Consider the following data: 6

{25, 25, 25, 25, 30, 33, 33, 35, 35, 36, 40, 45, 46, 52, 70}

i) Compute the five number summary for the above data
ii) Draw a boxplot of the data

b) What is the curse of dimensionality? Mention two techniques for dimensionality reduction. 4