[This question paper contains 6 printed pages.]

Sr. No. of Question Paper : 6096 **D** **Your Roll No.....**

Unique Paper Code : 234611

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

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

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. Question No. **1** is compulsory.
3. Parts of a questions must be answered together.
4. Attempt any **four** questions from Part **B.**
5. **All** questions in Part **B** carry equal marks.

PART A

Question No. **1** (Compulsory).

1. (a) Mention any two ways of outlier detection. Give an application of outlier detection. (4)

   (b) Under what situation is stratified sampling preferred over random sampling. (2)

   (c) What is a frequent itemset ? When is an association rule said to be strong ? (3)

   (d) Which of the two is more informative in data mining, (a) Low support and high confidence (b) High support and low confidence ? Support your answer with appropriate example. (4)

P.T.O.

(e) Consider the following set of frequent 3-itemsets :

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

List all the possible candidate 4-itemsets using the $F_{k-1} \times F_{k-1}$ merging strategy. (3)

(f) What is a confusion matrix ? With a suitable example show how it is used to evaluate the accuracy of a classifier. (3)

(g) Give the formula for calculating the Gini Index and Entropy to measure the impurity of a class distribution. For a binary class problem, derive the possible maximum and minimum value for both these measures. (4)

(h) Why is the k-Nearest Neighbor (kNN) classifier known as a lazy classifier ? (3)

(i) Why is clustering also known as unsupervised learning ? (3)

(j) Mention one strength and one weakness each of *k-means clustering* in comparison with the *h-medoids* algorithm. (3)

(k) Present conditions under which *density-based* clustering is more suitable than *partitioning-based* clustering and hierarchical clustering. (3)

PART B

Attempt any **four** questions from this part.
All questions carry equal marks.

2. (a) What is the five number summary of a distribution. Draw a boxplot to explain the same. (6)

(b) For a standard laboratory weight of 1 g, the following five values were observed experimentally : {1.015, 0.990, 1.013, 1.001, 0.986}. Calculate (a) precision (b) bias. (4)

3. (a) What is the *curse of dimensionality*. Mention two techniques commonly used for dimension reduction. (4)

(b)

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

For the transaction dataset given above,

(i) Generate the complete FP-tree.

(ii) Generate the list of frequent itemsets ordered by the suffix {c}. (3+3)

4.

TID	Items
1	A,B,C
2	B,C
3	C,D,E
4	C,D,E,F
5	A,B
6	B
7	A,D

P.T.O.

For the given dataset,

(i) Apply the Apriori algorithm with support count 2. Show the candidate and frequent itemsets for each database scan.

(ii) Indicate the association rules that are generated, sort them by confidence.

(6+4)

5. Consider the following data for a binary class problem :

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	–
T	T	+
F	F	–
F	F	–
F	F	–
T	T	–
T	F	–

(a) Calculate the overall Gini before the split.

(b) Calculate the gain in the Gini Index when split on A.

(c) Calculate the gain in the Gini Index when split on B.

(d) Which attribute would the decision tree induction algorithm choose ? Why ?

(2+3+3+2)

6. (a) Consider the one dimensional dataset :

X	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
Y	–	–	+	+	+	–	–	+	–	–

(i) Classify the data point x = 5.0 according to its 3 and 5 nearest neighbors (using majority vote). Use Euclidean distance as the distance metric.

(ii) Repeat part (i) using a distance-weighted voting approach. (2+4)

(b) Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. One-fifth of the college students are graduate students and the rest are undergraduates.

(i) What is the probability that a student who smokes is a graduate student ? Use the Bayes' theorem.

(ii) Suppose 30% of the graduate students live in a dorm but only 10% of the undergraduate students live in a dorm. If a student smokes and lives in the dorm, is he or she more likely to be a graduate or undergraduate student ? Assume independence between students who live in a dorm and those who smoke. (2+2)

7. (a) Suppose that the data mining task is to cluster the following eight points (with *(x; y)* representing location) into three clusters.

A1(2; 10); A2(2; 5); A3(8; 4); A4(5; 8); A5(7; 5); A6(6; 4); A7(1; 2}; A8(4; 9).

The distance function is Euclidean distance. Suppose initially we assign A1, A4, and A7 as the center of each cluster, respectively. Use the k-means algorithm to show the three clusters and mention the cluster centers after the first round of execution. (5)

(b) Following is the Euclidean distance matrix for 6 points. Perform the single link hierarchical clustering. Show the results by drawing a dendogram. The dendogram should clearly show the order in which the points are merged.

P.T.O.

	P1	P2	P3	P4	P5	P6
P1	0.0	0.24	0.22	0.37	0.34	0.23
P2		0.0	0.15	0.20	0.14	0.25
P3			0.0	0.15	0.28	0.11
P4				0.0	0.29	0.22
P5					0.0	0.39
P6						0.0

(5)