JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -3 EXAMINATIONS-2022

B.Tech-VI Semester (CS&IT)

COURSE CODE (CREDITS): 18B1WCI635 (2) MAX. MARKS: 35

COURSE NAME: DATA MINING & DATA WAREHOUSING

COURSE INSTRUCTORS: Jagpreet MAX. TIME: 2 Hours

***Note:** All questions are compulsory. Marks are indicated against each question in square brackets.*

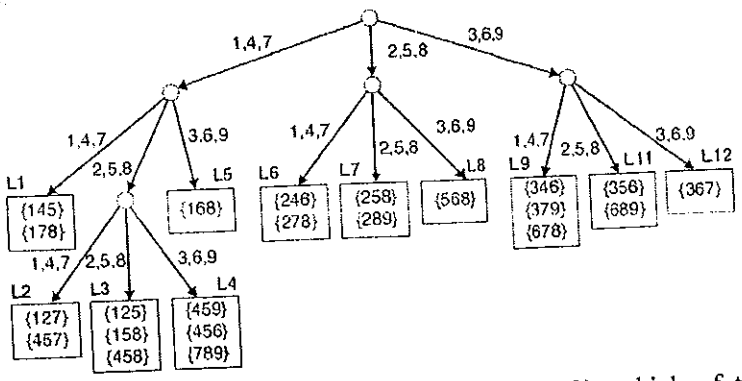
Q. No. 1 Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how following techniques can help: **[5*1 Marks] [CO-2]**

i. Data manipulation
ii. Clustering
iii. Classification
iv. Association rule mining
v. Anomaly detection

Q. No. 2 Use the distance measure in Table below, perform single link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. **[5 Marks] [CO-5]**

	P1	P2	P3	P4	P5	P6
P1	0.00	0.71	5.66	3.61	4.24	3.20
P2	0.71	0.00	4.95	2.92	3.54	2.50
P3	5.66	4.95	0.00	2.24	1.41	2.50
P4	3.61	2.92	2.24	0.00	1.00	0.50
P5	4.24	3.54	1.41	1.00	0.00	1.12
P6	3.20	2.50	2.50	0.50	1.12	0.00

Q. No. 3 The Apriori algorithm uses a hash tree data structure to efficiently count the support of candidate itemsets. Consider the hash tree for candidate 3-itemsets shown in figure below: **[2,3 Marks] [CO-4]**



i. Given a transaction that contains items {1,3,4,5,8}, which of the

hash tree leaf nodes will be visited when finding the candidates of the transaction?

ii. Use the visited leaf nodes in part (i) to determine the candidate item-sets that are contained in the transaction {1, 3, 4, 5, 8}.

Q. No. 4 How association rule mining is different from classification and clustering? Explain brute force method for association rule mining with example of shopping basket problem of five transactions and why it is computational expensive to perform? **[5 Marks] [CO-4]**

Q. No. 5 i. Traditional K-means has a number of limitations, such as sensitivity to outliers and difficulty in handling clusters of different sizes and densities, or with non-globular shapes. Comment on the ability of fuzzy c-means to handle these situations. **[3, 2 Marks] [CO-5]**

ii. Provide two examples where clustering is not a good technique to perform in data mining.

Q. No. 6 Consider the training examples shown in Table below for a binary classification problem. **[5 Marks] [CO- 4]**

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

i. Compute the Gini Index for the overall collection of training examples.

ii. Compute the Gini Index for the Customer ID attribute

iii. Compute the Gini Index for the Gender attribute.

iv. Compute the Gini Index for the Car Type attribute using multiway spilt.

v. Which attribute is better, Gender, Car Type or Shirt Size?

Q. No. 7 Write an algorithm for k-nearest neighbour classification given k and n, the number of attributes describing each tuple. **[5 Marks] [CO- 5]**