Lib. copy. 20/05/22
(Morning.)

[This question paper contains 7 printed pages.]

Your Roll No..............

Sr. No. of Question Paper	: **1174**	**A**
Unique Paper Code	: 32347611	
Name of the Paper	: Data Mining	
Name of the Course	: **B.Sc. (Hons.) 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. Question No. **1 (Section A)** is compulsory.
3. Attempt any **4** Questions from Nos. **2** to **8 (Section B)**.
4. Parts of a question must be answered together.

Section A

1. (a) How are accuracy rate and error calculated for evaluation of a classification model? (2)

(b) Briefly describe the aggregation technique in data-preprocessing? (2)

(c) Normalize the age of four students, given by the values {18, 21, 22, 25}. (2)

(d) Explain briefly the significance of dimensionality reduction. (2)

(e) What is an outlier in context of a dataset? (2)

(f) What kind of Association Rules do you think would be stronger and more interesting – the rules with high support and low confidence or the rules with low support and high confidence? Why? (3)

(g) Define the use of sampling in data mining? Name two sampling methods. (3)

(h) What are the three factors that affect the computational complexity of Apriori algorithm? (3)

(i) Distinguish between the following type of clustering schemes :

(i) Exclusive vs. Fuzzy Clustering

(ii) Complete vs. Partial Clustering (4)

(j) What do you understand by the term missing data in data mining? Briefly describe two methods for dealing with missing data. (4)

(k) Define the terms scalability and heterogeneity? What challenges do they pose while mining the data? (4)

(l) Define precision and recall metrics used for classification. (4)

Section B

2. (a) Explain discretization and binarization in context of data pre-processing. (4)

(b) Consider a categorical attribute Customer satisfaction {unsatisfactory, poor, neutral, good, very good}

(i) Convert the above categorical attribute to three binary attributes. (2)

(ii) Convert the same attribute to five asymmetric binary attributes. (2)

(c) State the Apriori Principle. (2)

3. For the given employee table, identify the type of each attribute (nominal, ordinal, interval- scaled, ratio-scaled), giving justification for your choice. For each attribute that has missing values, briefly state how will you handle missing values therein. (10)

Emp_id	Gender	Age	Home _pin _code	Date_ of_ joining	Desig.	Contact_No	Email_id
1001	M	32	232322	16/4/10	Captain	981828706	b@gma.com
1002	F	31	222321	21/3/11	Captain	981121072	f@gma.com
1003	F	34	243431	23/4/08	Major	992665007	??
1004	M	??	232432	21/5/09	Captain	987654390	r@gma.com
1005	M	35	454656	13/4/07	Colonel	981123456	d@gma.com
1006	??	36	465645	04/5/05	Colonel	786789564	a@gma.com
1007	F	30	234123	09/7/12	Captain	885678909	??
1008	M	32	676878	18/7/10	Major	??	x@gma.com
1009	M	33	565768	24/6/11	Colonel	989967890	e@gma.com
1010	M	30	498976	05/9/12	Major	??	d@gma.com

4. (a) Consider the following dataset where each data object has a class label along with five features associated with it.

Class	Cap Shape	Bruises	Odour	Stalk Shape	Habitat
Edible	Flat	Yes	anise	Tapering	grasses
poisonous	Convex	Yes	pungent	enlargening	grasses
Edible	Convex	Yes	almond	enlargening	grasses
Edible	Convex	Yes	almond	Tapering	meadows
Edible	Flat	Yes	anise	enlargening	woods
Edible	flat	No	none	enlargening	urban
poisonous	conical	Yes	pungent	enlargening	urban
Edible	flat	Yes	anise	enlargening	meadows
poisonous	convex	Yes	pungent	enlargening	urban

Consider the following pair of rules :

- *(Odour = pungent) and (habitat = urban)* → (Class = *poisonous)*
- *(Bruises = yes)* → *(Class = edible)*

(i) Are the two rules mutually exclusive? Justify your answer. (2)

(ii) Calculate coverage and accuracy for each of the rules. (4)

(b) Consider the one-dimensional labeled data set given below :

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

Classify the data point x = 4.0 according to the 5-nearest neighbours, using the majority voting scheme. (4)

5. (a) What are the three conditions needed to be satisfied by a distance measure, so that it can be established as a distance metric? (3)

(b) Show whether Euclidean Distance, used for finding distance between two data objects $o_1(x_1, y_1)$ and $o_2(x_2, y_2)$, can be treated as a distance metric. (6)

(c) With the help of a diagram, explain the usage of a dendrogram. (1)

6. Consider a transaction database D, consisting of nine transactions, as shown in the following table. Suppose the minimum support is set at 45% and the minimum confidence is set at 70%, show clearly the steps for finding out frequent itemsets of all sizes using the Apriori algorithm. Also generate the strong association rules from the frequent itemsets of size 3. (10)

TID	List of Items
T1	A,B,C,F
T2	B,D
T3	B,C
T4	A,B,C
T5	A,C,F
T6	B,C,F
T7	A,D
T8	A,B,C,E,F
T9	A,B,C

7. Consider a dataset of images of dogs and cats. Suppose there are 500 images of dogs and cats each. The classification model predicts 340 correct images of dogs and 410 correct images of cat. Perform the operations that follow :

(a) Draw the confusion matrix for this problem.

(b) Compute the classifier accuracy, error and sensitivity. (4+6)

8. Given the following data points: 4, 9, 18, 13, 11, 2, 6, 25, k = 3 and initial centroids $\mu_1 = 5$, $\mu_2 = 10$ and $\mu_3 = 15$. Show clearly the clusters and new cluster centres obtained after each iteration of K-means algorithm for two iterations of the algorithm. (10)