[This question paper contains 8 printed pages.]

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

Sr. No. of Question Paper : 3032 **H**

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 **7 (Section B)**.
4. Parts of a question must be answered together.
5. Use of Scientific Calculator is allowed.

Section A

1. (a) Determine the attribute type for the following :

(i) Price of a Book

(ii) Number of students in a class

P.T.O.

(iii) Eye color

(iv) Dates in a Calendar (2)

(b) How many association rules can be generated from a transactional dataset that contains five items? (2)

(c) Consider an association rule between items from market basket domain, which has high support and low confidence. What does it signify? (2)

(d) Consider a categorical attribute with three values {Grade A, Grade B, Grade C}. Covert this attribute to asymmetric binary attributes. (2)

(e) Give an example to show how the Apriori principle uses the anti-monotone property of support to prune the number of candidate itemsets. (2)

(f) Explain the problem of class imbalance with the help of an example. (2)

(g) Explain the following terms with respect to a density-based clustering algorithm: Core point, Border point, and Noise point. (3)

(h) List two ways to handle missing values and outliers in a dataset. (4)

(i) Enumerate four factors that affect the computational complexity of Apriori Algorithm. (4)

(j) List two points of difference between the following clustering schemes : (4)

(i) Complete vs. Partial clustering

(ii) Partitional vs. Hierarchical clustering

(k) Consider the following values for two attributes corresponding to four data points: $P_1(1, 3)$, $P_2(3, 1)$, $P_3(4, 2)$, and $P_4(6, 3)$. Compute the proximity matrix using Euclidean Distance. (4)

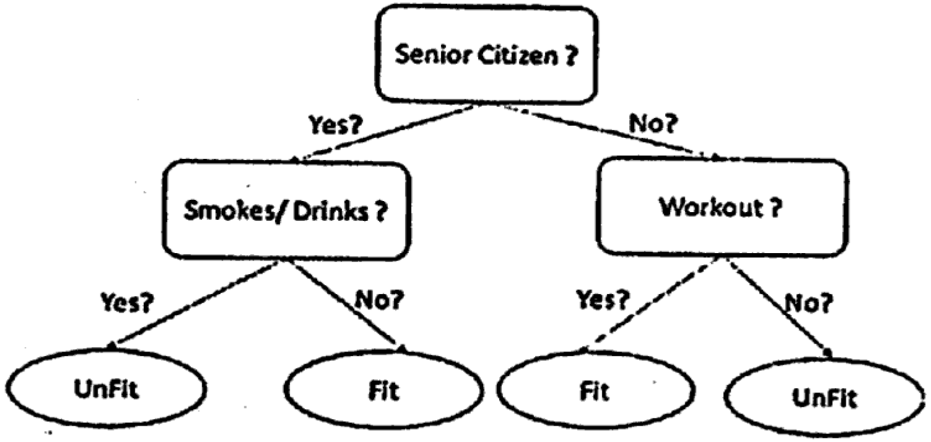
(l) Consider the given dataset with two attributes: Age and Loan Amount measured on different scales. What problem might arise if the dataset is directly used for K-means clustering with Euclidean distance as proximity measure? Show how can this problem be resolved on the data given in following table. (4)

	Age (in years)	Loan Amount (in rupees)
1	42	690000
2	25	510000
3	32	560000
4	37	590000
5	49	850000
6	36	700000

P.T.O.

Section B

2. Consider the following decision tree to classify health of a senior citizen and perform the listed tasks.



(a) Extract all the classification rules from the tree. Arrange the rules according to class-based ordering. (5)

(b) Are the extracted set of rules exhaustive? Justify. (2)

(c) Classify the following records : (3)

- Senior Citizen = Yes, Smokes/Drinks = No, Health = ?
- Senior Citizen = No, Workout = Yes, Health = ?

3. Suppose we have a dataset of 10 patients, where each patient is classified as either 'having a heart disease' or "not having a heart disease'. The dataset contains three categorical features: the patient's age group (x1). blood pressure level (x2), and cholesterol level (x3). A Naive Bayes classifier is to be used to predict whether a new patient is likely to have a heart disease based on the age group, blood pressure level, and cholesterol level. The dataset and their class labels are given below :

Age Group (x1)	Blood Pressure (x2)	Cholesterol Level (x3)	Heart Disease (y)
< 40	Normal	Normal	No
< 40	Normal	High	No
< 40	High	Normal	No
40-49	High	Normal	Yes
40-49	High	High	Yes
50-59	Normal	Normal	No
50-59	Normal	High	Yes
50-59	High	High	Yes
60+	High	Normal	Yes
60+	High	High	Yes

P.T.O.

(a) Estimate the conditional probabilities for *P(Age Group=60+\Yes), P(Blood Pressure-High\Yes), P(Cholesterol Level=Normal\Yes), P(Age Group=40-49\No), P(Blood Pressure=High\No),* and *P(Cholesterol Level-High(No).* (6)

(b) Use the estimate of conditional probabilities to predict the class label *(Heart Disease)* for a test sample *(Age Group = 50-59, Blood Pressure = High, Cholesterol Level=Normal)* using the Naive Bayes approach. (4)

4. (a) Given a 2×2 confusion matrix for a binary classifier, where n_{ij} denotes number of instances of class j predicted as class i. Further, $\Sigma_i \Sigma_j n_{ij} = N$ is the total number of instances. Write expressions for accuracy, precision, recall, and FI-score and compute these metrics for the following confusion matrix : (5)

	Predicted Class	
Original Class	Positive	Negative
Positive	513	2
Negative	4	101

(b) What is k-fold cross-validation? Suggest one advantage over the holdout method. (2)

(c) What is 'curse of dimensionality'? What is the difference between feature subset selection and dimensionality reduction techniques. (3)

5. Consider the following dataset, which consists of 10 tuples with two features, X1 and X2, and a binary class label Y.

Tuple No.	X1	X2	Y
T1	0.5	500	A
T2	1	750	A
T3	1.5	250	A
T4	2	1000	A
T5	2.5	600	B
T6	3	200	B
T7	3.5	800	B
T8	4	900	B
T9	4.5	400	A
T10	5	700	A
T11	3	600	?

(a) Scale the features X1 and X2 to range [0 – 1] using min-max normalization. (4)

(b) Use the k-Nearest Neighbor (k-NN) classifier with k = 5 and proximity measure as Euclidean distance to classify the tuple T11 (3, 600). Which class would the k-NN classifier assign to the tuple T11? Justify. (6)

6. Consider the following data set with nine transactions. Use Apriori algorithm to compute all frequent itemsets of size one and two, considering 1/3 as the minimum support. Also, generate strong association rules using frequent 2-itemsets, considering 0.65 as the minimum confidence. (10)

TID	Items
T1	Pen, Pencil, Paper
T2	Pencil, Sharpener
T3	Pencil, Eraser
T4	Pen, Pencil, Sharpener
T5	Pen, Eraser
T6	Pencil, Eraser
T7	Pen, Eraser
T8	Pen, Pencil, Eraser, Paper
T9	Pen, Pencil, Eraser

7. Use complete link agglomerative clustering to group the data described by the following distance matrix. Show the dendrogram. Use Euclidean distance as proximity measure. (10)

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0