[This question paper contains 4 printed pages.]

Your Roll No.....

Sr. No. of Question Paper : 13135 **L**

Unique Paper Code : 32347611

Name of the Paper : Data Mining (CBCS)

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)* Determine the attribute type for each of the following : (2)
   (i) Movie ratings on a scale of 1 to 5 stars
   (ii) Temperature measured in degrees Celsius
   (iii) Postal/ZIP codes
   (iv) Number of children in a family.
   (b) Find the Euclidean distance between the data points X(1, 2, 0, −1) and Y(0, −1, 2, 1). (2)
   (c) If the recall and precision of a classifier are 0.6 and 0.8 respectively, compute the value of F_1 measure. (2)
   (d) State the *Apriori principle*. If itemset {p, q} is found to be infrequent, what can be said about itemset {p, q, r} ? Explain. (2)
   (e) Briefly describe *two* methods used for handling missing values in a dataset. (3)
   (f) How is an eager learner classifier different from a lazy learner classifier ? Support your answer with one example of each. (3)

P.T.O.

(g) Consider a categorical attribute Quality with values {very poor, poor, average, good, excellent}. Show the representation of each value when this attribute is converted to binary form using : (3)

(i) 3 bits

(ii) 5 asymmetric binary attributes.

(h) Explain the following terms with reference to the DBSCAN clustering algorithm : (2)

(i) Core point

(ii) Border point

(i) Normalize the following set of values using min-max normalization : {15, 19, 22, 28}. (3)

(j) The confusion matrix for a 2-class problem is given below. Calculate the *Accuracy*, *Sensitivity*, *Specificity* and *Error rate* : (5)

		Predicted Class	
		Class = 1	**Class = 0**
Actual Class	**Class = 1**	350	150
	Class = 0	100	400

(k) What is the difference between noise and outliers ? Are noise objects always outliers? Are outliers always noise objects ? Justify with examples. (4)

(l) What is k-fold cross-validation ? How does it differ from the holdout method ? List *two* advantages of k-fold cross-validation over the holdout method. (4)

Section B

2. Consider the following dataset for a 2-class classification problem : (10)

A	B	Class
T	F	+
T	T	+
T	T	−
T	F	−
T	T	+
F	F	−
F	F	−
F	T	+
F	F	−
T	T	+

(i) Calculate the Gini index when splitting on attributes A and B.

(ii) Which attribute would the decision tree induction algorithm choose to split on?

(iii) Draw the resulting decision tree showing the number of instances of each class at the leaf nodes.

(iv) How many instances are misclassified by the resulting decision tree?

3. Consider the following transaction dataset. Assume the minimum support is 40% and the minimum confidence is 60%. (10)

TID	Items
T1	{a, b, c}
T2	{a, b, d}
T3	{b, c, d}
T4	{a, c, d}
T5	{a, b, c, d}
T6	{b, c}
T7	{a, b, d}
T8	{b, c, d}
T9	{a, b, c, d}
T10	{a, c}

(i) Show clearly the steps to find frequent itemsets of all sizes using the Apriori algorithm.
(ii) Generate all the strong association rules from the frequent 3-itemsets.

4. Given the following data points : 3, 6, 8, 14, 16, 22, 28, 32. Assume $k = 2$, and the initial cluster centroids $\mu_1 = 3$ and $\mu_2 = 14$. Show clearly the clusters obtained and the new cluster centres after each iteration of the K-means algorithm for two iterations. (10)

5. The DM Cafe sells cookies with optional toppings : chocolate-chips, walnuts, and oats. Suppose, you have tried five cookies (C1 to C5) and kept a record of which ones you liked : (10)

	Chocolate-chips	Walnuts	Oats	Liked
C1	True	True	False	True
C2	False	True	True	False
C3	True	False	True	True
C4	True	True	True	False
C5	False	False	True	True

Show binarization of the above data and use it to calculate Euclidean distances, to demonstrate how the k-Nearest-Neighbor (k-NN) classifier with majority voting would classify a tuple <True, False, True>, for $k = 1$ and $k = 3$ respectively.

6. Consider the dataset shown below : (10)

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

(i) Estimate the conditional probabilities P(Outlook|Yes), P(Temperature|Yes), P(Humidity|Yes), P(Windy|Yes), P(Outlook|No), P(Temperature|No), P(Humidity|No), and P(Windy|No).

(ii) Use these estimates of conditional probabilities to predict the class label (Play Golf) for a test sample (Outlook = Rainy, Temperature = Cool, Humidity = High, Windy = True) using the naive Bayes approach.

7. Use the distance matrix given below to perform hierarchical clustering using the *single link* approach. Show the dendrogram clearly indicating the merging height at each step. (10)

	P1	P2	P3	P4	P5
P1	0				
P2	0.30	0			
P3	0.25	0.18	0		
P4	0.40	0.32	0.20	0	
P5	0.35	0.15	0.30	0.28	0

8. Consider the following rule set for animal classification : (10)

R1 : (Habitat = land) ∧ (Has_Fur = yes) ⟶ Mammals
R2 : (Has_Wings = yes) ∧ (Has_Fur = no) ⟶ Birds
R3 : (Has_Scales = yes) ∧ (Habitat = water) ⟶ Fish
R4 : (Habitat = land) ∧ (Has_Fur = no) ∧ (Has_Wings = no) ⟶ Reptiles
R5 : (Habitat = water) ∧ (Has_Scales = no) ⟶ Amphibians

Consider the following test instances :

Name	Habitat	Has_Fur	Has_Wings	Has_Scales	Class
eagle	land	no	yes	no	?
dog	land	yes	no	no	?
trout	water	no	no	yes	?
frog	water	no	no	no	?

(i) Are these rules mutually exclusive ? Justify your answer.

(ii) Are these rules exhaustive ? Justify with an example if not.

(iii)Identify which rule(s) cover each test instance and predict the class for each.

(200)