This question paper contains 8 printed pages]

Roll No. □□□□□□□□□□□

S. No. of Question Paper : **9426A**

Unique Paper Code : **32347611** **HC**

Name of the Paper : **Data Mining**

Name of the Course : **B.Sc. (H) Computer Science : DSE-4**

Semester : **VI**

Duration : **3 Hours** Maximum Marks : **75**

(Write your Roll No. on the top immediately on receipt of this question paper.)

All questions are compulsory from Section A.

Attempt any *four* questions from Section B.

Section A

1. (*a*) What is the difference between Data mining and KDD ? 2

(*b*) Identify attribute types for the following : 2

(*i*) eye color

(*ii*) grades

(*iii*) dates in a calendar

(*iv*) age.

P.T.O.

(*c*) What are the maximum and minimum values of Gini Index ? Find Gini index for the following node : 2

Node N	**Count**
Class = 0	1
Class : 1	5

(*d*) Give *two* applications where graph data structure is used to model data. 2

(*e*) Given four points $p_1(0,2)$, $p_2(2,0)$, $p_3(3,1)$ and $p_4(5,1)$. Calculate Euclidian distance between the points p_1 and p_2, and p_3 and p_4. 2

(*f*) Let X denote the categorical attribute having possible values {poor, good, better, best}. What is the representation of each value when X is converted to binary form ? 2

(*g*) How are interval scaled attributes different from ratio scaled attributes ? Give an example of each. 3

(*h*) How is a eager learner different from lazy learner ? Support your answer with an example from both categories of classifiers. 3

(*i*) State the *Apriori principle*. Comment on the following statement :

"If an item set {*x*, *y*, *z*} is frequent, then its subset {*y*, *z*} will be frequent." 4

(*j*) Given the age of four students, normalize the values {18, 21, 22, 25}. 4

(*k*) Explain the following terms with reference to the DBSCAN algorithm : 2+2

(*i*) Core point

(*ii*) Noise point

(*l*) What are *mutually exclusive* rules in a rule based classifier ? What problem may arise if rules are not mutually exclusive ? How can such problem be resolved ? 5

Section B

2. (*a*) Consider the following transaction dataset : 6

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

(*i*) Compute the support of itemsets $\{e\}$, $\{b, d\}$, $\{b, d, e\}$, $\{a, b, d, e\}$

(*ii*) Compute the confidence of rules $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$.

(*iii*) Is confidence a symmetric measure ?

(*b*) Let A and B be two sets of integers. A distance measure ‘*d*’ is defined as *d*(A – B) = size(A – B), where ‘–’ denotes the set difference. Prove that ‘*d*’ is not a metric. 4

3. (*a*) Explain the concept of aggregation with the help of an example. List three uses of aggregation. 2+3

(*b*) What is the difference between noise and outliers ? Answer the following questions : 5

(*i*) Is noise ever interesting or desirable ?

(*ii*) Are outliers ever interesting or desirable ?

(*iii*) Are noise obejcts always outliers ?

(*iv*) Are outliers always noise objects ?

4. (*a*) For the following two-class problem, draw a *Confusion Matrix* and compute the *Accuracy* and *Error* from it : 6

Instance id	A	B	Predicted Class	Actual Class
1	T	F	+	+
2	T	T	+	+
3	T	T	+	–
4	T	F	–	–
5	T	T	+	+
6	F	F	–	+
7	F	F	–	–
8	F	F	–	–
9	T	T	–	+
10	T	F	–	+

(*b*) What is *k* - fold cross validation ? How is it different from the hold-out method ? 4

5. (*a*) What is the difference between hierarchical and partition based clustering ? Enumerate *two* advantages and disadvantages of hierarchical clustering. 4

(*b*) What is simple random sampling ? 2

(*c*) Consider the following rule set : 4

R_1 : (Give Birth = no) ^ (Can Fly = yes) → Birds

R_2 : (Give Birth = no) ^ (Live in Water = yes) → Fishes

R_3 : (Give Birth = yes) ^ (Blood Type = warm) → Mammals

R_4 : (Give Birth = no) ^ (Can Fly = no) → Reptiles

R_5 : (Live in Water = sometimes) → Amphibians.

Which rules cover the following tuples :

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	Warm	No	Yes	No	?
bear	Warm	Yes	No	No	?

6. (*a*) List the rules that can be generated from tlre 3- itemset ABE using the following transactional data set. Compute the confidence. 8

T_id	Itemset
t1	ACD
t2	BCE
t3	ABCE
t4	BDE
t5	ABCE
t6	ABCD

(*b*) Enumerate strong association rules if *minConf* = 0.5 2

7. (*a*) Given the following points : 2, 4, 10, 12, 3, 20, 30, 11, 25. Given $k = 3$, and the initial means, $\mu_1 = 2$, $\mu_2 = 4$ and $\mu_3 = 6$. Show the clusters obtained and the new means after each iteration using the K-means algorithm. 8

(*b*) What is the differences between Partial and Complete clustering scheme ? 2