U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

June 2025 Semester End Main Examinations

Programme: B.E. **Semester: V**
Branch: Information Science and Engineering **Duration: 3 hrs.**
Course Code: 22IS5PCDMG **Max Marks: 100**
Course: Data Mining

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

READ FOR EXAMS 2024-25

			UNIT - I	CO	PO	Marks
1	a)		Provide definition of an attribute. Mention the different types of attributes. Highlight on examples and operations for each type.	CO1	PO1	05
	b)		Pre-processing involves transforming raw data into an understandable format suitable for mining. The following techniques provide data transformation and reduction. Elaborate on each of the techniques: (i) Feature creation (ii) Discretization and Binarization (iii) Aggregation	CO1	PO1	15
			OR			
2	a)		Elucidate the following data pre-processing techniques: (i) Sampling (ii) Dimensionality reduction (iii) Feature subset selection	CO1	PO1	15
	b)		For the following vectors, x and y, calculate the cosine, correlation, Euclidean and Jaccard measures. x = (0, 1, 0, 1), y = (1, 0, 1, 0)	CO2	PO1	05
			UNIT - II			
3	a)		Consider the following set of training samples:	CO3	PO2	15

Owns home	Married	Gender	Employed	Credit Rating	Risk Class
yes	yes	male	yes	A	B
no	no	female	yes	A	A
yes	yes	female	yes	B	C
yes	no	male	no	B	B
no	yes	female	yes	B	C
no	no	female	yes	B	A
no	no	male	no	B	B
yes	no	female	yes	A	A
no	yes	female	yes	A	C
yes	yes	female	yes	A	C

			The measures for computing impurities are (a) Gini and (b) Classification error. Find the best split by computing each of these impurity measures for the attributes in the above set of training samples.			
		b)	Rule ordering schemes are either rule-based or class-based. Explicate on these rule ordering schemes.	*CO1*	*PO1*	**05**
			OR			
	4	a)	Consider a training set that contains positive and negative examples. For each of the following candidate rules, determine which is the best candidate rule according to Rule coverage, Likelihood ratio, Laplace measure and m-estimate measure (with k: 2 and p: 0.6). Training set contains 36 positive and 25 negative examples (i) R1: covers 34 positive and 2 negative R2: covers 2 positive and 23 negative (ii) R3: covers 28 positive and 9 negative R4: covers 8 positive and 16 negative	*CO3*	*PO2*	**08**
		b)	Consider the training examples shown below for a binary classification problem. Apply Entropy measure to find the best split.	*CO3*	*PO1*	**12**

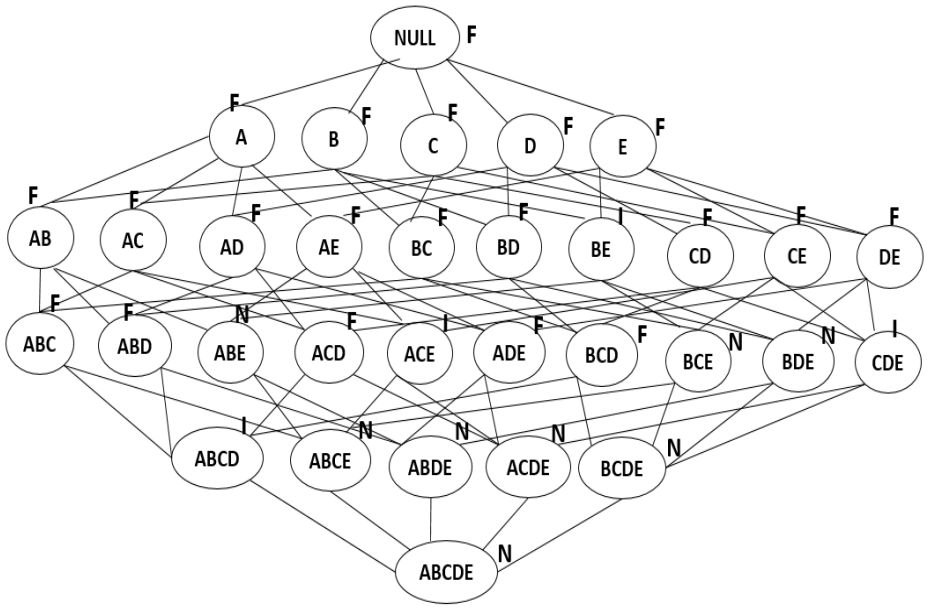
Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	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

			UNIT - III			
	5	a)	Explicate k-Nearest neighbor classification.	*CO1*	*PO1*	**08**
		b)	Apriori algorithm use support-based pruning to systematically control the exponential growth of candidate itemsets. Apply Apriori algorithm for the following set of transactions to find frequent item set, given minimum support as 30%.	*CO2*	*PO1*	**12**

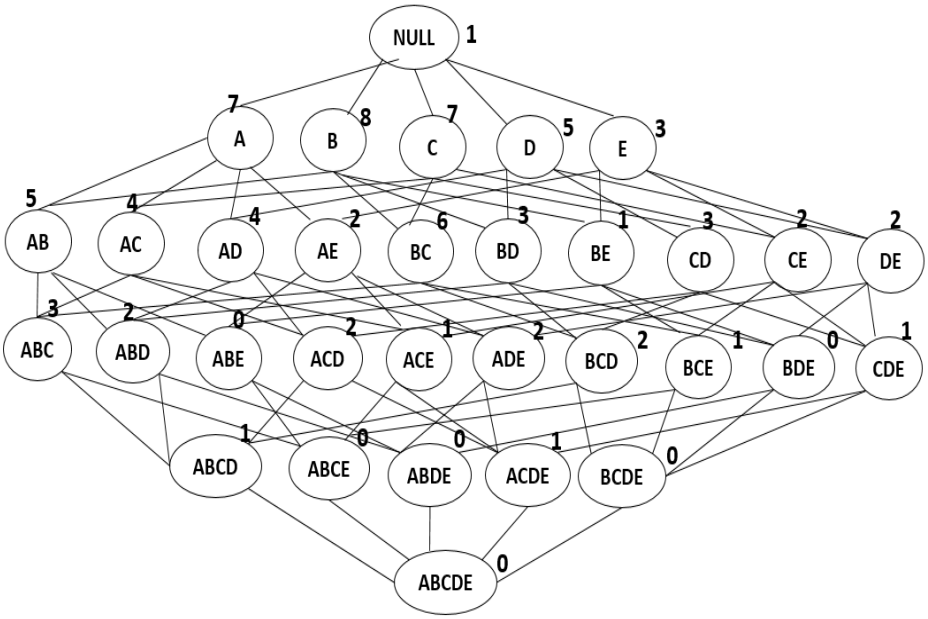
Transaction ID	Items Bought
1	{a,b,d,e}
2	{b,c,d}
3	{a,b,d,e}
4	{a,c,d,e}

			5 | {b,c,d,e} 6 | {b,d,e} 7 | {c,d} 8 | {a,b,c} 9 | {a,d,e} 10 | {b,d} Draw an itemset lattice representing the data set. Label each node in the lattice with the following letter(s): • N: If the itemset is not considered to be a candidate itemset by the Apriori algorithm. • F: If the candidate itemset is found to be frequent by the Apriori algorithm. • I: If the candidate itemset is found to be infrequent after support counting.			
			OR			
	6	a)	Explicate Naïve Bayes classification.	*CO1*	*PO1*	**08**
		b)	Apply Apriori algorithm for the following set of transactions to find frequent item set, with minimum support as 30%. TID | Items 1 | {Milk, Bread} 2 | {Bread, Butter, Jam} 3 | {Milk, Butter, Eggs} 4 | {Milk, Eggs} 5 | {Bread, Milk, Butter} 6 | {Bread, Butter} 7 | {Butter, Eggs} 8 | {Milk, Bread} 9 | {Bread, Eggs} Draw an itemset lattice representing the data set. Label each node in the lattice with the following letter(s): • N: If the itemset is not considered to be a candidate itemset by the Apriori algorithm. • F: If the candidate itemset is found to be frequent by the Apriori algorithm. I: If the candidate itemset is found to be infrequent after support counting.	*CO2*	*PO1*	**12**
			UNIT - IV			
	7	a)	Determine maximal frequent itemsets with minimum support as 20%. Lattice indicates frequent, infrequent and not considered itemsets.	*CO3*	*PO2*	**10**

REAPPEAR EXAMS 2024-25



Determine closed frequent itemsets with minimum support as 20%. Support count is provided in the lattice.



		b)	A compact data structure FP-tree extracts frequent itemsets. Apply FP-growth algorithm for the following transactions to find frequent item set with minimum support as 20%.	*CO2*	*PO1*	**10**

TID	Items
1	{a,b,e}
2	{b,d}
3	{b,c}
4	{a,b,d}
5	{a,c}
6	{b,c}
7	{a,c}
8	{a,b,c,e}
9	{a,b,c}

OR

8 For the following dataset, determine frequent itemsets using FP-Growth algorithm with minimum support 20%. *CO2* *PO1* **20**

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

Frequent Itemsets	
a	?
b	?
c	?
d	?
e	?

Generate the rule set with minimum confidence as 70%.

UNIT - V

9 a) A sample data consists of 6 two-dimensional points. *CO3* *PO2* **15**

	X	Y
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

Apply MIN, MAX and Group average techniques to define proximity between clusters and represent the resulting hierarchical clustering diagrams.

b) Clusters represent data distribution and differing densities. State and explain DBSCAN algorithm. *CO1* *PO1* **05**

OR

10 a) Elucidate Fuzzy c-means clustering. *CO1* *PO1* **05**

b) A sample data consists of 6 two-dimensional points. *CO3* *PO2* **15**

Point	**X**	**Y**
Q1	0.12	0.58
Q2	0.28	0.44
Q3	0.40	0.35
Q4	0.30	0.18
Q5	0.15	0.30
Q6	0.50	0.25

Perform hierarchical clustering using Single link, Complete link and Group average techniques that define proximity between clusters.

******

REAPPEARED MIS 2024-25