VASAVI COLLEGE OF ENGINEERING *(AUTONOMOUS)*, HYDERABAD
Accredited by NAAC with A++ Grade
B.E. (C.S.E.) VII-Semester Supplementary Examinations, July-2022
Data Mining (PE-II)

Time: **3 hours**   Max. Marks: **60**

*Note: Answer all questions from **Part-A** and any **FIVE** from **Part-B***

Part-A (10 × 2 = 20 Marks)

Q. No.	Stem of the question	M	L	CO	PO
1.	Draw a Boxplot for the data given as 85,92,78,88,90,88,89	2	1	1	1,2
2.	Suppose your midterm test score is 83 and your final exam score is 95. Using weights of 40% for the midterm and 60% for the final exam, compute the weighted average of your scores. If the minimum average for an A is 90, will you earn an A?	2	3	1	1,2
3.	Consider the following data. Fill the missing values of the attribute CGPA with mean, median, Linear interpolation.	2	2	2	1,2

HT.NO	Region	CGPA	Class
1	Urban	9.3	A
2	Rural	8.7	B
3	Rural	??	B
4	Rural	7.3	C
5	Urban	7.3	C

Q. No.	Stem of the question	M	L	CO	PO
4.	What is Concept hierarchy? Give concept hierarchy for Location dimension.	2	2	2	1,2
5.		2	3	3	1,2

	basketball	not basketball	sum(row)
cereal	2000	1750	3750
not cereal	1000	250	1250
sum(col.)	3000	2000	5000

Find the relation between basket ball playing and cereal eating habits by using Chi square test.

Q. No.	Stem of the question	M	L	CO	PO
6.	Draw a Contingency table for the rule B →D by using the transition database shown below. Use the contingency table and Compute the Confidence.	2	2	3	1,2

Transaction	Items
T1	ACDE
T2	BCE
T3	ABCE
T4	BCDE
T5	ABCE
T6	ABCD

Q.No						

7. Find Gini-index of the attribute 'A' in the table given below   2   3   4   1,2

Instance	A	B	C	Class
1	0	0	1	—
2	1	0	1	+
3	0	1	0	—
4	1	0	0	—
5	1	0	1	+
6	0	0	1	+
7	1	1	0	—
8	0	0	0	—
9	0	1	0	+
10	1	1	1	+

8. Compare Pre-pruning and Post-pruning of Decision tree.   2   2   4   1,2

9. Find the Cohesiveness of the cluster C1 given as C1={ P1(0,2) P2(3,2), P3 (3,5)}   2   2   5   1,2

10. Compute the SSE of the clusters as shown below   2   2   5   1,2
C1= { (0,1)(3,0)(2,2)} and C2= {(5,0)(7,8)}

Part-B (5 × 8 = 40 Marks)

11. a) Describe different kinds of patterns mined in data mining with appropriate examples.   3   2   1   1,2

    b) Compute the dissimilarity matrix for the student data given below   5   3   1   1,2

Roll No	Marks	Grades
1	96	A
2	87	B
3	83	B
4	96	A

12. a) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient and the two measures count & charge, where charge is the fee that a doctor charges a patient for a visit. Design a star schema used for modeling data warehouse   4   3   2   1,2

    b) Normalize the age attribute values 3,15,16,19,19,20,21,22,25,24,26,29,30,32,33,34,35,35,35,36,40,45,52,70 using Z-Score normalization   4   2   2   1,2

13. a) Explain different techniques to enhance Apriori algorithm.   3   2   3   1,2,3

    b) Consider the below tansaction in which B = Bread, J = Jelly, P = Peanut Butter, M = Milk and E = Eggs. Given that minimum threshold support = 40% , find frequent itemsets using FP-growth algorithm   5   3   3   1,2,3

TID	Items Bought
1	{B, J, P}
2	{B, P}
3	{B, M, P}
4	{E, B}
5	{E, M}

14. a)	Describe the following measures to find the best splitting attribute i) Gain ii) Gini Index iii) Gain ratio	3	2	4	1,2
b)	Find the predicted values by using Backpropogation after one iteration. Initial weights, bias factors, target outputs and network structure are as given below.	5	3	4	1,2,3

B1	B2	T1	T2
0.35	0.60	0.01	0.99

X1	X2	W1	W2	W3	W4	W5	W6	W7	W8
0.05	0.10	0.15	0.20	0.25	0.30	0.40	0.45	0.50	0.55
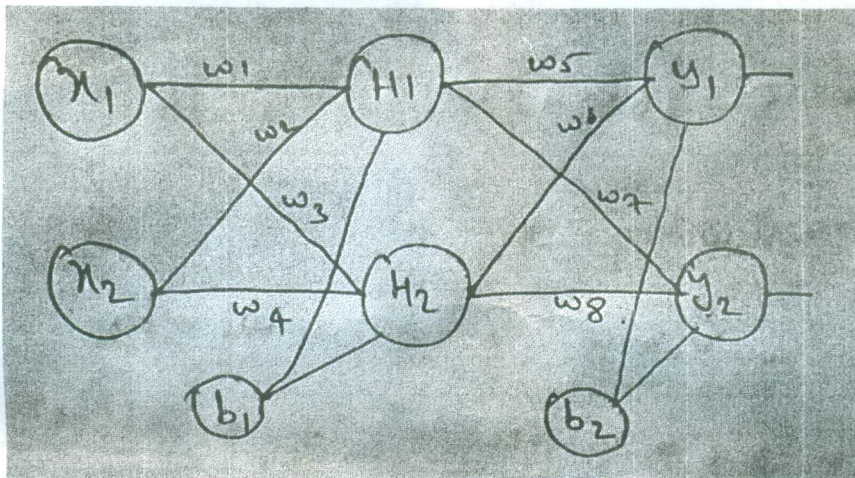
15. a)	Explain the following methods to find the clusters in Agglomerative clustering algorithm. a) Single-link b) Complete-link c) Group-average	3	2	5	1,2
b)	Given a set of seven 2-dimensional points p1=(0, 0), p2=(5, 0), p3=(5, 1), p4=(0, 1), p5=(0, 0.5), p6=(0, 9), and p7=(5.5, 1). Euclidean distance is the distance function. The DBSCAN algorithm is used to cluster the points. Epsilon = 1, and MinPts = 2 is used for DBSCAN. Find the clusters and outliers.	5	3	5	1,2,3
16. a)	(see table below)	3	3	1	1,2

Name	Gender	Fever	Cough	T1	T2	T3
Jack	M	Y	N	N	N	N
Mary	F	Y	N	N	P	N
Jim	M	Y	Y	N	N	N

Compute the S(Jack, Mary), S(Mary, Jim) and S(Jack, Jim) using Simple Matching Coefficient (SMC) and Jaccard Coefficient (JC) measures. Explain the significance of JC over SMC.

b)	Describe the 3-tier architecture of a Data Warehouse with a neat sketch.	5	1	2	1,2

17.						
		Answer any *two* of the following:				
	a)	Find frequent item sets and Strong Association rules using candidate generation for the following dataset by using Apriori algorithm with min_sup=50% Min_Conf=60%	4	3	3	1,2,3

TID	Items
T100	I1, I3, I4
T200	I2, I3, I5
T300	I1, I2, I3, I5
T400	I2, I5

	b)	Consider weather nominal data set given below. Find the class label for the test sample X= {outlook=overcast, temperature=mild, humidity = normal} by using Naïve bayes Classification algorithm.	4	3	4	1,2,3

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

	c)	Describe different type of outliers. Give example for each type of outlier.	4	2	5	1,2

M : Marks;   L: Bloom's Taxonomy Level;   CO; Course Outcome;   PO: Programme Outcome

i)	Blooms Taxonomy Level – 1	20%
ii)	Blooms Taxonomy Level – 2	35.52%
iii)	Blooms Taxonomy Level – 3 & 4	44.48%

*****

VII Th Sem All

ok