QPC: RD20BTECH327 AR 20 Reg. No □□□□□□□□□□

GIET UNIVERSITY, GUNUPUR – 765022

B. Tech (Fifth Semester – Regular) Examinations, December – 2022

BPCCS5010 / BPCCT5010 - Data Mining & Data Warehousing

(CSE & CST)

Time: 3 hrs Maximum: 70 Marks

Answer ALL Questions

The figures in the right hand margin indicate marks.

PART – A: (Multiple Choice Questions) **(1 x 10 = 10 Marks)**

Q.1. Answer ***ALL*** questions CO # PO #

a. What does Apriori algorithm do? CO-3 PO-1

i. It mines all frequent patterns through pruning rules with lesser support
ii. It mines all frequent patterns through pruning rules with higher support
iii. Both 1 and 2
iv. None of the above

b. What is not true about FP growth algorithms? CO-2 PO-2

i. It mines frequent itemsets without candidate generation.
ii. There are chances that FP trees may not fit in the memory
iii. FP trees are very expensive to build
iv. It expands the original database to build FP trees.

c. What is Gini index? CO-3 PO-1

i. It is a type of index structure
ii. It is a measure of purity
iii. Both options except none
iv. None of the options

d. Which one of these is not a tree based learner? CO-2 PO-2

i. CART
ii. ID3
iii. Bayesian classifier
iv. Random Forest

e. The following technology is not well-suited for data mining: CO-3 PO-1

i. Expert system technology
ii. Data visualization
iii. Technology limited to specific data types such as numeric data types
iv. Parallel architecture

f. Which of the following features usually applies to data in a data warehouse? CO-3 PO-1

i. Data are often deleted
ii. Most applications consist of transactions
iii. Data are rarely deleted
iv. Relatively few records are processed by applications

g. In the relational database terminology, a table is synonymous with: CO-1 PO-1

i. A column
ii. A row
iii. An attribute
iv. A relation

h. A null value indicates: CO-1 PO-1

i. A numeric value with value 0
ii. The absence of a value
iii. A very small value
iv. An erroneous value

i. The following is a major disadvantage while using a neural network CO-2 PO-2

i. It is very difficult to find optimal or near optimal parameters for the network
ii. Interpretation of the model becomes very difficult
iii. It becomes difficult to model non-linear relation between input and output variables
iv. The number of inputs it can handle are limited

j. In training a neural network using back propagation algorithm CO-2 PO-2

i. Chain rule of differentiation is used in computing gradient of the error surface
ii. Activation functions are chosen so that they are differentiable in nature
iii. The connecting weights can be generated initially at random in the range of (0.0, 1.0)
iv. All of the above

PART – B: (Short Answer Questions) (2 x 10 = 20 Marks)

Q2. Answer ALL questions		CO #	PO #
a.	What is Knowledge Discovery?	CO-1	PO-1
b.	What is the need of data warehouses?	CO-2	PO-2
c.	Define fact table.	CO-4	PO-1
d.	Define metadata and explain the types of metadata	CO-3	PO-1
e.	Define support and confidence.	CO-3	PO-1

f.	Find the cosine similarity between the given two term frequency vectors:
	X=[3,2,0,5,0,0,0,2,0,0]
	Y=[1,0,0,0,0,0,0,1,0,2]
		CO-2	PO-1

g.	What is attribute selection measure?	CO-3	PO-1
h.	Briefly describe the k-NN classification algorithm.	CO-3	PO-3
i.	Give two examples of activation function used in neural networks.	CO-3	PO-2
j.	Explain the principle of hierarchical clustering.	CO-3	PO-1

PART – C: (Long Answer Questions) (10 x 4 = 40 Marks)

Answer ***ALL*** questions	Marks	CO #	PO #
3.a. Briefly outline how to compute the *dissimilarity* between objects described by the following types of variables: i. Numerical (interval-scaled) variables ii. Categorical variables iii. Ratio-scaled variables iv. Nonmetric vector objects	5	CO-1	PO-2
b. Explain the steps of KDD, with the help of a diagram.	5	CO-1	PO-1
(OR)			
c. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:	10	CO-2	PO-2

Age	23	23	27	27	39	41	47	49	50
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
Age	52	54	54	56	57	58	58	60	61
% fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

i. Calculate the mean, median, and standard deviation of age and %fat.

ii. Find out the covariance and correlation among these two attributes.

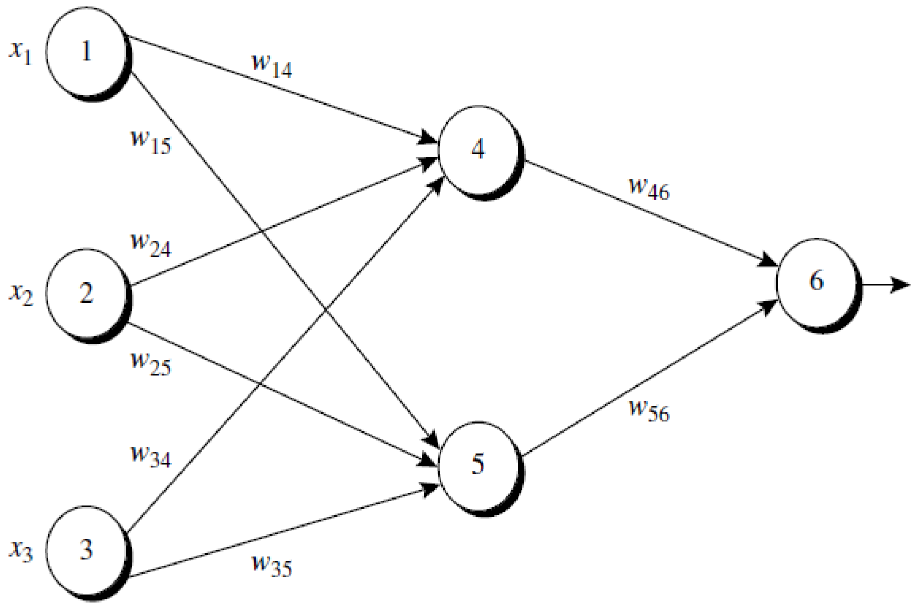
	Marks	CO #	PO #
4.a. Explain how Apriori Algorithm is used for mining frequent item sets.	5	CO-2	PO-1
b. What are the measures of interestingness for an association rule? Define a strong association rule.	5	CO-2	PO-2
(OR)			
c. There are five transactions (T1,T2,T3,T4,T5) with items (A,B,C,D) purchased as T1(B,C),T2(A,C,D),T3(B,C), T4(A,B,C,D), T5(B,D). The min_sup=2. Show how Apriori Rule Mining Algorithm can generate the association rules for the above dataset.	10	CO-3	PO-2

5.a. What is decision trees algorithm? List down the attribute selection measures used by the ID3 algorithm to construct a Decision Tree. 5 CO-2 PO-2

b. Write short answer on Naïve Bayes classifier. 5 CO-2 PO-1

(OR)

c. A multilayer feed-forward neural network is shown in below Figure. Let the learning rate be 0.9. The initial weight and bias values of the network are given in Table below, along with the first training tuple, ***X* = (1, 0, 1)**, with a class label of 1. Compute Net input, output and error at each node and update weight and bias values just once. Use logistic activation function at nodes 4, 5 and 6. 10 CO-3 PO-2



Initial Input, weight and Bias values:

x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{24}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	-0.3	0.4	0.1	-0.5	0.2	-0.3	-0.2	-0.4	0.2	0.1

6.a. Why is outlier mining important? Briefly describe the different approaches behind distanced-based outlier detection and density based local outlier detection. 5 CO-2 PO-2

b. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): Compute the *Minkowski distance* between the two objects, using *q* = 3. 5 CO-2 PO-1

(OR)

c. Both *k-means* and *k-medoids* algorithms can perform effective Clustering. Illustrate the strength and weakness of *k-means* in comparison with the *k-medoids* algorithm. 5 CO-3 PO-2

d. Suppose that the data mining task is to cluster the following eight points (with (*x*, *y*) representing location) into three clusters: 5 CO-3 PO-2

*A*1(2, 10), *A*2(2, 5), *A*3(8, 4), *B*1(5, 8), *B*2(7, 5), *B*3(6, 4), *C*1(1, 2), *C*2(4, 9): The distance function is Euclidean distance. Suppose initially we assign *A*1, *B*1, and *C*1 as the center of each cluster, respectively.

Use the *k-means* algorithm to show *only*

i. The three cluster centers after the first round execution
ii. The final three clusters

--- End of Paper ---