Total No. of Questions: 6 *Total No. of Printed Pages:3*

Enrollment No.....

Faculty of Engineering

End Sem (Odd) Examination Dec-2022

IT3ED02 Data Mining & Warehousing

Programme: B.Tech. Branch/Specialisation: IT

Duration: 3 Hrs. **Maximum Marks: 60**

Note: All questions are compulsory. Internal choices, if any, are indicated. Answers of Q.1 (MCQs) should be written in full instead of only a, b, c or d.

Q.1 i. Snowflake schema has _____ table(s). **1**

(a) Sub-division (b) Sub-dimension

(c) Sub-Fact (d) None of these

ii. According to Inmon, a data warehouse is a subject-oriented, integrated, time-variant, and _____ collection of data. **1**

(a) Non-volatile (b) Volatile

(c) Summarized (d) Combined

iii. The KDD is abbreviation for- **1**

(a) Knowledge Database Definition

(b) Knowledge Discovery in Databases

(c) Knowledge Discovery Definition

(d) Knowledge Data Definition

iv. The various aspects of data mining methodologies is/are _____. **1**

I. Mining various and new kinds of knowledge.

II. Mining knowledge in multidimensional space.

III. Pattern evaluation and pattern or constraint-guided mining.

IV. Handling uncertainty, noise, or incompleteness of data.

(a) I, II and IV only (b) II, III and IV only

(c) I, II and III only (d) I, II, III and IV

v. What do you mean back propagation? **1**

(a) It is the transmission of error back through the network to adjust the inputs.

(b) It is the transmission of error back through the network to allow weights to be adjusted so that the network can learn.

(c) It is another name given to the curvy function in the perceptron

(d) None of these

P.T.O.

vi. Confidence (A-->B) = Support (A U B) / _____. **1**
(a) Support (A) (b) Support (B)
(c) Support (C) (d) None of these

vii. K-means is an example of- **1**
(a) Classification
(b) Association
(c) Clustering
(d) Prediction

viii. _____ is a clustering procedure characterized by the development of a tree-like structure. **1**
(a) Non-hierarchical clustering
(b) Hierarchical clustering
(c) Divisive clustering
(d) Agglomerative clustering

ix. Which of the following operations is not an OLAP operation? **1**
(a) Roll-up (b) Slice
(c) Drill-down (d) Zoom-in

x. MOLAP stands for- **1**
(a) Multidimensional Operational Analytic Processing
(b) Multidimensional Online Analytical Processing
(c) Mining Online Analytical Program
(d) Mining Operational Analytical Processing

Q.2 i. What is Data Mart? **2**

ii. Define data warehouse. Explain the data warehouse architecture with diagram. **3**

iii. Briefly explain different types of sources used in data warehouse from where data can be extracted. **5**

OR iv. Explain the star and snowflake schema of data warehouse. **5**

Q.3 i. What is data cleaning? Describe the approaches to fill missing values and noisy data. **4**

ii. Describe challenges to data mining regarding data mining methodology and user interaction issues. **6**

OR iii. Explain KDD process with the help of a diagram. **6**

Q.4 Attempt any two:

i. With the help of decision tree find means of predicting which company profiles will lead to a increase or decrease in profits based on the following data: **5**

Age	Competition	Type	Profit
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

Profit is class attribute.

ii. A database has five transactions. Let minimum support=60% and minimum confidence=80%. **5**
TID ITEMS_BOUGHT
T100 {M, O, N, K, E, Y}
T200 {D, O, N, K, E, Y}
T300 {M, A, K, E}
T400 {M, U, C, K, Y}
T500 {C, O, R, K, I, E}
Find all frequent itemsets using Apriori algorithm.

iii. Define FP-Growth algorithm with suitable example. **5**

Q.5 i. Define Clustering. What are the requirements for cluster analysis? **4**

ii. Explain DBSCAN algorithm with suitable example. **6**

OR iii. Suppose we have the following points: (1,1), (2,4), (3,4), (5,8), (6,2), (7,8). Use k - means algorithm (k = 2) to find two cluster. The distance function is Euclidean distance. **6**

Q.6 Attempt any two:

i. Describe typical OLAP operations with diagram. **5**

ii. Differentiate between OLTP and OLAP. **5**

iii. Explain types of OLAP. **5**

******

Marking Scheme
IT3ED02 Data Mining and Warehousing

Q.1	i)	Snow flake schema has _____ table(s). (b) Sub-dimension	**1**
	ii)	According to Inmon, a data warehouse is a subject-oriented, integrated, time-variant, and _____ collection of data. a) Non-volatile	**1**
	iii)	The KDD is abbreviation for (b) Knowledge Discovery in Databases	**1**
	iv)	The various aspects of data mining methodologies is/are _____ I. Mining various and new kinds of knowledge. II. Mining knowledge in multidimensional space. III. Pattern evaluation and pattern or constraint-guided mining. IV. Handling uncertainty, noise, or incompleteness of data. (d) All I, II, III and IV	**1**
	v)	Confidence (A-->B) = Support (A U B) / _____ (a) Support (A)	**1**
	vi)	What do you mean back propagation? (b) It is the transmission of error back through the network to allow weights to be adjusted so that the network can learn	**1**
	vii)	K-means is an example of (c) Clustering	**1**
	viii)	_____ is a clustering procedure characterized by the development of a tree-like structure. (b) Hierarchical clustering	**1**
	ix)	Which of the following operations is not an OLAP operation? (d) Zoom-in	**1**
	x)	MOLAP stands for: (b) Multidimensional Online Analytical Processing	**1**
Q.2	i.	What is Data Mart? 2 marks	**2**
	ii.	Define Data Warehouse. Explain the data warehouse architecture with diagram. Data Warehouse Definition: 1 mark Data warehouse architecture: 2 marks	**3**
	iii.	Briefly explain different types of sources used in Data Warehouse from where data can be extracted. 5 types: 5 marks	**5**
OR	iv.	Explain the Star and Snowflake schema of Data Warehouse. Star schema: 2.5 marks	**5**
		Snowflake schema: 2.5 marks	
Q.3	i.	What is Data Cleaning? Describe the approaches to fill missing values and noisy data. Data Cleaning: 1.5 Marks Approaches: 2.5 Marks	**4**
	ii.	Describe challenges to Data Mining regarding data mining methodology and user interaction issues. Challenges: 4 Marks User interaction issues: 2 Marks	**6**
OR	iii.	Explain KDD process with the help of a diagram. Diagram: 2 Marks KDD Explanation: 4 Marks	**6**
Q.4	i.	With the help of decision tree find means of predicting which company profiles will lead to a increase or decrease in profits based on the following data:	**5**

Age	**Competition**	**Type**	**Profit**
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

Profit is class attribute.
2 Marks for calculating information gain for 3 attributes
2 Marks for calculating 2nd level splitting attribute
1 Mark for drawing the tree

	ii.	A database has five transactions. Let minimum support=60% and minimum confidence=80%. TID ITEMS_BOUGHT T100 {M, O, N, K, E, Y} T200 {D, O, N, K, E, Y } T300 {M, A, K, E} T400 {M, U, C, K, Y} T500 {C, O, R, K, I, E}	**5**

		Find all frequent itemsets using Apriori algorithm. Support Calculation: 3 marks Confidence: 2 marks	
OR	iii.	Define FP-Growth Algorithm with suitable example. FP-Growth Algorithm: 2 Marks Example: 3 marks	**5**
Q.5	i.	Define Clustering. What are the requirements for cluster analysis? Clustering: 2 marks Requirements: 2 marks	**4**
	ii.	Explain DBSCAN Algorithm with suitable example. DBSCAN: 2 marks Algorithm: 4 marks	**6**
OR	iii.	Suppose we have the following points: (1,1), (2,4), (3,4), (5,8), (6,2), (7,8). Use k - means algorithm (k = 2) to find two cluster. The distance function is Euclidean distance. Stepwise marking: 6 marks	**6**
Q.6		Attempt any two:	
	i.	Describe typical OLAP operations with diagram. 5 Operations: 1 mark for each	**5**
	ii.	Differentiate between OLTP and OLAP: 1 mark for each difference.	**5**
	iii.	Explain types of OLAP. ROLAP: 2 marks MOLAP: 2 marks HOLAP: 1 mark	**5**

******