Code No: R2031053 **R20** SET - 1

III B. Tech I Semester Regular Examinations, Dec/Jan -2022-23
DATA WAREHOUSING AND DATA MINING
(Computer Science and Engineering)

Time: 3 hours Max. Marks: 70

Answer any **FIVE** Questions **ONE** Question from **Each unit**
All Questions Carry Equal Marks
*****

UNIT-I

1. a) Explain kinds of data can be mined? Give examples. [7M]
   b) Differentiate Operational database systems and data warehousing. Explain the star schema and fact constellation schemas. [7M]

(OR)

2. a) List and describe the five primitives for specifying a data mining task. [7M]
   b) What are the differences between the three main types of data warehouse usage: information processing, analytical processing, and data mining? Discuss the motivation behind OLAP mining (OLAM). [7M]

UNIT-II

3. a) Suppose a group of 12 sales price records has been sorted as follows: 5,10,11,13,15,35,50,55,72,92,204,215. Partition them into three bins by each of the following methods: (i) equal-frequency (equal-depth) partitioning (ii) equal-width partitioning (iii) clustering [7M]
   b) Write short notes on the following: (i) Data Preprocessing (ii) Data Discretization (iii) Concept Hierarchy [7M]

(OR)

4. a) What are the value ranges of the following normalization methods? (i) min-max normalization (ii) z-score normalization (iii) z-score normalization using the mean absolute deviation instead of standard deviation (iv) normalization by decimal scaling [7M]
   b) Explain in detail about data pre-processing. [7M]

UNIT-III

5. a) Use the C4.5 algorithm to build a decision tree for classifying the followingobjects: [7M]

Class	Size	Color	Shape
A	Small	Yellow	Round
A	Big	Yellow	Round
A	Big	Red	Round
A	Small	Red	Round
B	Small	Black	Round
B	Big	Black	Cube
B	Big	Yellow	Cube
B	Big	Black	Round
B	Small	Yellow	Cube

   b) Why information gain is considered as attribute selection measure? Illustrate with an example. [7M]

(OR)

6. a) Explain the decision tree induction algorithm with appropriate examples. Discuss the disadvantages of this approach? What is over fitting, and how can it be prevented for decision trees? [7M]
   b) What is visual mining? Explain the application of decision tree induction algorithm in it. [7M]

UNIT-IV

7. a) Why is the process of discovering association rules relatively simple compared to generating large itemsets in transactional databases? [7M]
   b) Can we design a method that mines the complete set of frequent item sets without candidate generation? If yes, explain it with the following table: [7M]

TID List of items
001 milk, dal, sugar, bread
002 Dal, sugar, wheat,jam
003 Milk, bread, curd, paneer
004 Wheat, paneer, dal, sugar
005 Milk, paneer, bread
006 Wheat, dal, paneer, bread

(OR)

8. a) Discuss Apriori Algorithm with a suitable example and explain how its efficiency can be improved? [7M]
   b) Consider the transaction data-set: **TransIDItems** [7M]

T1 {a,b}
T2 {b,c,d}
T3 {a,c,d,e}
T4 {a,d,e}
T5 {a,b,c}
T6 {a,b,c,d}
T7 {a}
T8 {a,b,c}
T9 {a,b,d}
T10 {b,c,e}

Construct the FP tree by showing the trees separately after reading each transaction.

UNIT-V

9. a) Consider five points {X1 , X2 , X3 , X4 , X5 } with the following co ordinates as a two dimensional sample for clustering : X1 = (0.5, 2.5); X2 = (0,0); X3 = (1.5,1); X4 = (5,1); X5 = (6,2) Illustrate the K-means partitioning algorithms using the above data set. [7M]
   b) What is cluster analysis? Describe the dissimilarity measures for interval-scaled variables and binary variables. [7M]

(OR)

10. a) Describe k-means clustering algorithms in terms of the following criteria: (i) shapes of clusters that can be determined; (ii) input parameters that must be specified; and (iii) limitations. [7M]
    b) What is Cluster Analysis? What are some typical applications of clustering? What are some typical requirements of clustering in data mining? [7M]

Code No: R2031053 **R20** SET - 2

III B. Tech I Semester Regular Examinations, Dec/Jan -2022-23
DATA WAREHOUSING AND DATA MINING
(Computer Science and Engineering)

Time: 3 hours Max. Marks: 70

Answer any **FIVE** Questions **ONE** Question from **Each unit**
All Questions Carry Equal Marks
*****

UNIT-I

1. a) Explain what kinds of pattern scan be mined? Give examples. [7M]
   b) State why, for the integration of multiple heterogeneous information sources, many companies in industry prefer the update-driven approach, rather than the query-driven approach. [7M]

(OR)

2. a) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. Enumerate three classes of schemas that are popularly used for modeling data warehouses and explain. [7M]
   b) Briefly compare the following concepts using examples [7M]
   Discovery-driven cube, multi feature cube, virtual warehouse

UNIT-II

3. a) What is data integration and discuss issues to consider during data integration. [7M]
   b) Given the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. [7M]
   (i) Use min-max normalization to transform the value 35 for age onto the range [0.0,1.0].
   (ii) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.
   (iii) Use normalization by decimal scaling to transform the value 35 for age.

(OR)

4. a) What is the need of dimensionality reduction? Explain any two techniques for dimensionality reduction [7M]
   b) Suppose a group of 12 sales price records has been sorted as follows: 5,10,11,13,15,35,50,55,72,92,204,215. [7M]
   Partition them into three bins by each of the following methods:
   (i) equal-frequency (equal-depth) partitioning
   (ii) equal-width partitioning
   (iii) clustering

UNIT-III

5. a) What are the new features of C4.5 algorithm comparing with original Quinlan's ID3 algorithm for decision-tree generation? [7M]
   b) What is attribute selection measure? Briefly describe the attribute selection measures for decision tree induction. [7M]

(OR)

6. a) Given a decision tree, you have the option of (i) converting the decision tree to rules and then pruning the resulting rules, or (ii) pruning the decision tree and then converting the pruned tree to rules. What advantage does (i) have over (ii)? [7M]

   b) The following table consists of training data from an employee database. The data have been generalized. For example, "31 … 35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row. [7M]

department status age salary count
sales senior 31…35 46K…50K 30
sales junior 26…30 26K…30K 40
sales junior 31…35 31K…35K 40
systems junior 21…25 46K…50K 20
systems senior 31…35 66K…70K 5
systems junior 26…30 46K…50K 3
systems senior 41…45 66K…70K 3
marketing senior 36…40 46K…50K 10
marketing junior 31…35 41K…45K 4
secretary senior 46…50 36K…40K 4
secretary junior 26…30 26K…30K 6

Let status be the class label attribute.Construct a decision tree from the given data.

UNIT-IV

7. a) Given a simple transactional database X: [7M]

TID Items
T01 A, B, C, D
T02 A, C, D, F
T03 C, D, E, G, A
T04 A, D, F, B
T05 B, C, G
T06 D, F, G
T07 A, B, G
T08 C, D, F, G

Using the threshold values support = 25% and confidence = 60%, find all large item sets in database X

   b) How association rules mined from large databases? Explain. [7M]

(OR)

8. a) What is a frequent item set? How to find frequent item sets for a transactional database? Explain any one approach with illustrations. [7M]

   b) Find frequent item sets for the following table using FP-Growth algorithm. Assume relevant thresholds. [7M]

Tid List of item ids
T1I1, I3, I5
T2I2, I4, I1
T3I1, I2, I3, I4
T4I5, I3, I2
T5I1, I2, I5
T6 I3, I4, I5

UNIT-V

9. a) Given the samples X1 = {1, 0}, X2 = {0, 1}, X3 = {2, 1}, and X4 = {3, 3}, suppose that the samples are randomly clustered into two clusters C1 = {X1, X3} and C2 = {X2, X4}. Apply oneiteration of the K-means partitional-clustering algorithm, and find a new distribution of samples in clusters. [7M]

   b) Describe how categorization of major clustering methods is being done? [7M]

(OR)

10. a) Suppose that the data-mining task is to cluster the following eight points (representing location) into three clusters: A1 (2;10) ; A2 (2;5) ; A3 (8;4) ; B1 (5;8) ; B2 (7;5) ; B3 (6;4) ; C1 (1;2) ; C2 (4;9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to determine: the three cluster centers after the first round of execution. [7M]

    b) What are the advantages and disadvantages of k-means clustering against model-based clustering? [7M]

Code No: R2031053 **R20** SET - 3

III B. Tech I Semester Regular Examinations, Dec/Jan -2022-23
DATA WAREHOUSING AND DATA MINING
(Computer Science and Engineering)

Time: 3 hours Max. Marks: 70

Answer any **FIVE** Questions **ONE** Question from **Each unit**
All Questions Carry Equal Marks
*****

UNIT-I

1. a) Briefly compare the following concepts with example Snowflake schema, fact constellation, starnet query model. [7M]
   b) Explain technologies used for data mining? [7M]

(OR)

2. a) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.
   Draw a schema diagram for the above data warehouse. [7M]
   b) Present an example where data mining is crucial to the success of a business. What data mining functionalities does this business need? Can such patterns be generated alternatively by data query processingor simple statistical analysis? [7M]

UNIT-II

3. a) Use these methods to normalize the following group of data: 200,300,400,600,1000
   (i) min-max normalization by setting min = 0 and max = 1
   (ii) z-score normalization
   (iii) z-score normalization using the mean absolute deviation instead of standard deviation. [7M]
   b) What is data consolidation? In detail discuss various techniques used to consolidate data. [7M]

(OR)

4. a) Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results: [7M]

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) Normalize the two attributes based on z-score normalization.
   (ii) Calculate the correlation coefficient (Pearson's product moment coefficient). Are these two attributes positively or negatively correlated? Compute their covariance.
   b) Describe the problem of data quality with some examples. Explain the usage of feature subset selection in data preprocessing. [7M]

UNIT-III

5. a) Given a training data set Y: [7M]

A B C Class
15 1 A C1
20 3 B C2
25 2 A C1
30 4 A C1
35 2 B C2
25 4 A C1
15 2 B C2
20 3 B C2

Find the best split point for decision tree for attribute A.

b) Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning? [7M]

(OR)

6. a) Make a decision tree for the following database using Gini Index. Indicate all intermediate steps. [7M]

Example	***Colour***	***Shape***	***Size***	***Class***
1	Red	Square	Big	+
2	Blue	Square	Big	+
3	Red	Circle	Big	+
4	Red	Circle	Small	–
5	Green	Square	Small	–
6	Green	Square	Big	–

b) Given data set, D, the number of attributes, n, and the number of training tuples, |D|, show that the computational cost of growing a tree is at most n × |D| × log(|D|). [7M]

UNIT-IV

7. a) What are the frequent item sets with a minimum support of 3 for the given set of transactions? [7M]

TID Items
101 A,B,C,D,E
102 A,C,D
103 D,E
104 B,C,E
105 A,B,D,E
106 A,B
107 B,D,E
108 A,B,D
109 A,D
110 D,E

b) Write the algorithm to discover frequent item sets without candidate generation and explain it with an example. [7M]

(OR)

8. a) Consider the following table to find frequent item sets using vertical data format. Support threshold 30% [7M]

TidList of items

T01 Milk, biscuits, surf powder, teabags
T02 Teabags, sugar, soap
T03 Milk, sugar, bread, soap
T04 Bread, teabags, biscuits
T05 Chocolates, milk, biscuits
T06 Milk, teabags, bread
T07 Bread, biscuits, chocolate
T08 Milk, surf powder, bread

b) Discuss Apriori Algorithm with a suitable example and explain how its efficiency can be improved? [7M]

UNIT-V

9. a) Cluster the following data into three clusters, using the *k*-means method. [7M]

x	***y***
10.9	12.6
2.3	8.4
8.4	12.6
12.1	16.2
7.3	8.9
23.4	11.3
19.7	18.5
17.1	17.2
3.2	3.4
1.3	22.8
2.4	6.9
2.4	7.1
3.1	8.3
2.9	6.9
11.2	4.4
8.3	8.7

b) Describe K means clustering with an example. [7M]

(OR)

10. a) Consider five points {X1 , X2 , X3 , X4 , X5 } with the following coordinates as a two dimensional sample for clustering : X1 = (0.5,2.5); X2 = (0,0); X3 = (1.5,1); X4 = (5,1); X5 = (6,2) Illustrate the K-means partitioning algorithms using the above data set. [7M]

b) What is cluster analysis? Describe the dissimilarity measures for interval-scaled variables and binary variables. [7M]

Code No: R2031053 **R20** SET - 4

III B. Tech I Semester Regular Examinations, Dec/Jan -2022-23
DATA WAREHOUSING AND DATA MINING
(Computer Science and Engineering)

Time: 3 hours Max. Marks: 70

Answer any **FIVE** Questions **ONE** Question from **Each unit**
All Questions Carry Equal Marks
*****

UNIT-I

1. a) Explain which kinds of applications are targeted for data mining? [7M]
   b) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. [7M]
   Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2010? Explain the concepts involved.

(OR)

2. a) A data warehouse can be modeled by either a star schema or a snowflake schema. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. [7M]
   b) Explain the difference and similarity between discrimination and classification, between characterization and clustering, and between classification and regression. [7M]

UNIT-II

3. a) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. [7M]
   b) Discuss in detail about data transformation with suitable examples. [7M]

(OR)

4. a) Explain various data pre-processing methods with appropriate examples. [7M]
   b) Give the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. [7M]
   How might you determine outliers in the data? Relate it with data cleaning.

UNIT-III

5. a) Describe the classification task in induction and deduction phases. Explain with example classification tasks. [7M]

   b) Calculate the gain in the Gini Index when splitting on A and B. Which attribute would the decision tree induction algorithm choose? [7M]

A	***B***	***Class Label***
T	F	+
T	T	+
T	T	+
T	F	–
T	T	+
F	F	–
F	F	–
F	F	–
T	T	–
T	F	–

(OR)

6. a) Why information gain is considered as attribute selection measure? Illustrate with an example. [7M]

   b) Identify the attribute that will act as the root node of a decision tree to predict golf play for following database with Gini Index. Indicate all the intermediate steps. [7M]

Outlook	**Wind**	**PlayGolf**
rain	strong	no
sunny	weak	yes
overcast	weak	yes
rain	weak	yes
sunny	strong	yes
rain	strong	no
overcast	strong	no

UNIT-IV

7. a) Assume 5 transactions and explain the two-step approach to generate frequent item sets and to mine association rules using Apriori algorithm. [7M]

   b) A database has four transactions. Let min_sup=60% and min_conf=80% [7M]

TID date items_bought
100 10/15/2022{K, A, B, D}
200 10/15/2022{D, A, C, E, B}
300 10/19/2022{C, A, B, E}
400 10/22/2022{B, A, D}
Find all frequent items using Apriori& FP-growth, respectively. Compare the efficiency of the two-meaning process.

(OR)

8. a) Write the algorithm to discover frequent item sets without candidate generation and explain it with an example. [7M]

   b) Make a comparison of Apriori and FP-Growth algorithms for frequent item set mining in transactional databases. Apply these algorithms to the following data: [7M]

TID LIST OF ITEMS

1 Bread, Milk, Sugar, TeaPowder, Cheese, Tomato
2 Onion, Tomato, Chillies, Sugar, Milk
3 Milk, Cake, Biscuits, Cheese, Onion
4 Chillies, Potato, Milk, Cake, Sugar, Bread
5 Bread, Jam, Mik, Butter, Chilles
6 Butter, Cheese, Paneer, Curd, Milk, Biscuits
7 Onion, Paneer, Chilies, Garlic, Milk
8 Bread, Jam, Cake, Biscuits, Tomato

UNIT-V

9. a) Suppose that the data mining task is to cluster points into three clusters, where the points are A1(2,10),A2(2,5),A3(8,4),B1(5,8),B2(7,5),B3(6,4),C1(1,2),C2(4,9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only the three cluster centers after the first round of execution [7M]

   b) What are the requirements for cluster analysis? Explain briefly. [7M]

(OR)

10. a) Given the points x1 = {1, 0}, x2 = {0,1}, x3={2, 1}, and x4 = {3, 3}. Suppose that these points are randomly clustered into two clusters: C1 = {x1, x3} and C2 = {x2, x4}. Apply one iteration of K-means partitional-clustering algorithm and find new distribution of elements in clusters. What is the change in a total square error? [7M]

    b) Use an example to show why the k-means algorithm may not find the global optimum, that is, optimizing the within-cluster variation [7M]