Code No: R1632052

$\boxed{\text{R16}}$

$\left(\text{SET - 1}\right)$

### III B. Tech II Semester Regular Examinations, April/May - 2019
## DATA WAREHOUSING AND MINING
(Computer Science and Engineering)

Time: 3 hours                                                                                     Max. Marks: 70

Note: 1. Question Paper consists of two parts (**Part-A** and **Part-B**)
2. Answer **ALL** the question in **Part-A**
3. Answer any **FOUR** Questions from **Part-B**
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

### PART –A

| | | | |
|---|---|---|---|
| 1. | a) | What are the steps involved in KDD process. | [2M] |
| | b) | State why data preprocessing is an important issue for data warehousing and data mining. | [2M] |
| | c) | What is decision tree classifier? | [2M] |
| | d) | What is Bayesian Belief Networks? | [3M] |
| | e) | How association rules mined from large databases? | [3M] |
| | f) | Define density based method. | [2M] |

### PART -B

| | | | |
|---|---|---|---|
| 2. | a) | What is data Mining? Explain the differences between Knowledge discovery and data mining. | [7M] |
| | b) | Define Data Visualization & data transformation? Explain with examples. | [7M] |
| | | | |
| 3. | a) | Write short notes on the following:<br>(i) Data Preprocessing    (ii) Data Discretization    (iii) Concept Hierchy | [6M] |
| | b) | Given the following measurement for the variable age:<br>18, 22, 25, 42, 28, 43, 33, 35, 56, 28<br>Standardize the variables by the following:<br>(i) Compute the mean absolute deviation for age.<br>(ii) Compute the Z-score for the first four measurements. | [8M] |
| | | | |
| 4. | a) | Explain different classification Techniques. | [7M] |
| | b) | (i) What are over fitted models? Explain their effects on performance.<br>(ii) What are the advantages and disadvantages of decision trees over other classification methods? | [7M] |
| | | | |
| 5. | a) | Explain Naive Baye's Classification. | [7M] |
| | b) | Explain Baye's theorem. Develop an algorithm for classification using Bayesian classification. | [7M] |
| | | | |
| 6. | a) | Discuss Apriori Algorithm with a suitable example and explain how its efficiency can be improved? | [7M] |
| | b) | Write the algorithm to discover frequent item sets without candidate generation and explain it with an example. | [7M] |
| | | | |
| 7. | a) | Describe K means clustering with an example. | [7M] |
| | b) | (i) What are the requirements for cluster analysis? Explain briefly.<br>(ii) What is an outlier? Explain the types of outliers. | [7M] |

\*\*\*\*\*\*

**R16**

**SET - 2**

### III B. Tech II Semester Regular Examinations, April/May - 2019
## DATA WAREHOUSING AND MINING
(Computer Science and Engineering)

Time: 3 hours                                                                                     Max. Marks: 70

Note: 1. Question Paper consists of two parts (**Part-A** and **Part-B**)
2. Answer **ALL** the question in **Part-A**
3. Answer any **FOUR** Questions from **Part-B**

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

## PART –A

| | | | |
|---|---|---|---|
| 1. | a) | List the five primitives for specifying a data mining task. | [2M] |
| | b) | Write the strategies for data reduction. | [2M] |
| | c) | List the approaches for filling in the missing values. | [2M] |
| | d) | What is pattern evaluation & correlation analysis? | [3M] |
| | e) | Define support and confidence in Association rule mining. | [3M] |
| | f) | What is an outlier? Mention its applications. | [2M] |

## PART -B

| | | | |
|---|---|---|---|
| 2. | a) | What is data mining? Briefly explain the Knowledge discovery process. | [7M] |
| | b) | Describe the various descriptive statistical measures for data mining. | [7M] |
| | | | |
| 3. | a) | Explain in detail about data pre-processing. | [7M] |
| | b) | What is the need of dimensionality reduction? Explain any two techniques for dimensionality reduction. | [7M] |
| | | | |
| 4. | a) | Discuss K- Nearest neighbor classification algorithm and its characteristics. | [7M] |
| | b) | What is association and correlation? With an example describe classification and prediction. | [7M] |
| | | | |
| 5. | a) | State Bayes theorem and discuss how Bayesian classifiers work? | [7M] |
| | b) | What are Bayesian classifiers? With an example, describe how to predict a class label using Naive Bayesian classification. | [7M] |

6.     A database has four transactions. Let min_sup=60% and min_conf=80%                [14M]

| TID | date | items_bought |
|---|---|---|
| 100 | 10/15/2018 | {K, A, B, D} |
| 200 | 10/15/2018 | {D, A, C, E, B} |
| 300 | 10/19/2018 | {C, A, B, E} |
| 400 | 10/22/2018 | {B, A, D} |

i) Find all frequent items using Apriori & FP-growth, respectively. Compare the efficiency of the two meaning process.

ii) List all of the strong association rules (with support '*s*' and confidence '*c*') matching the following meta-rule where X is a variable representing customers, and item i denotes variables representing items (e.g., "A", "B",etc.):
Vx $\in$ transactions, buys(X,item1) ^ buys(X,item2) =>buys(X,item3)[s,c].

| | | | |
|---|---|---|---|
| 7. | a) | What is Density based clustering? Describe DBSCAN clustering algorithm. | [7M] |
| | b) | Describe how categorization of major clustering methods is being done? | [7M] |

\*\*\*\*\*

|''|'''||''||'''||||

**R16**

**SET - 3**

### III B. Tech II Semester Regular Examinations, April/May - 2019
# DATA WAREHOUSING AND MINING
(Computer Science and Engineering)

Time: 3 hours                                                                                          Max. Marks: 70

Note: 1. Question Paper consists of two parts (**Part-A** and **Part-B)**
2. Answer **ALL** the question in **Part-A**
3. Answer any **FOUR** Questions from **Part-B**

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

## PART –A

| | | | |
|---|---|---|---|
| 1. | a) | What is data mining? | [2M] |
| | b) | How concept hierarchies are useful in data mining? | [2M] |
| | c) | List similarity measures. | [2M] |
| | d) | What is rule classification? | [3M] |
| | e) | List the techniques to improve the efficiency of Apriori algorithm. | [2M] |
| | f) | What is the objective function of the K-means algorithm? | [3M] |

## PART -B

| | | | |
|---|---|---|---|
| 2. | a) | Explain data mining as a step-by-step process of knowledge discovery. Mention the Functionalities of Data mining. | [7M] |
| | b) | What is data cleaning? Describe the approaches to fill missing values. | [7M] |
| 3. | a) | Write a note on subset selection in attributes for data reduction. | [7M] |
| | b) | Discuss briefly about data cleaning techniques. | [7M] |
| 4. | a) | What is Decision tree? With an example, briefly describe the algorithm for generating decision tree. | [7M] |
| | b) | What is prediction? Explain the various prediction techniques. Explain about Decision tree Induction classification technique. | [7M] |
| 5. | a) | Describe the data classification process with a neat diagram. How does the Naive Bayesian classification works? Explain. | [7M] |
| | b) | What is misclassification rate of a classifier? Describe sensitivity and specificity measures of a classifier. | [7M] |

6. Make a comparison of Apriori and FP-Growth algorithms for frequent item set mining in transactional databases. Apply these algorithms to the following data: [14M]

| 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 |

7. 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 ) [14M]

Illustrate the K-means partitioning algorithms using the above data set.

\*\*\*\*\*

|"|'''||"||'''||||

**III B. Tech II Semester Regular Examinations, April/May - 2019**
**DATA WAREHOUSING AND MINING**
(Computer Science and Engineering)

Time: 3 hours                                                                 Max. Marks: 70

Note: 1. Question Paper consists of two parts (**Part-A** and **Part-B**)
2. Answer **ALL** the question in **Part-A**
3. Answer any **FOUR** Questions from **Part-B**

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

## PART –A

| | | | |
|---|---|---|---|
| 1. | a) | Define Discretization. | [2M] |
| | b) | List the three important issues that have to be addressed during data integration. | [2M] |
| | c) | Define Pre-pruning and post-pruning. | [2M] |
| | d) | Mention any three measures of Similarity. | [3M] |
| | e) | Define Association rule mining two step processes. | [2M] |
| | f) | Define outliers. List various outlier detection approaches. | [3M] |

## PART -B

| | | | |
|---|---|---|---|
| 2. | a) | Discuss in detail about the steps of knowledge discovery? | [7M] |
| | b) | What is noisy data? Explain the binning methods for data smoothening. | [7M] |
| | | | |
| 3. | a) | What is data normalization? Explain any two normalization methods. | [7M] |
| | b) | Briefly describe various forms of data pre-processing. | [7M] |
| | | | |
| 4. | a) | What is attribute selection measure? Briefly describe the attribute selection measures for decision tree induction. | [7M] |
| | b) | Describe the criteria used to evaluate classification and prediction methods. | [7M] |
| | | | |
| 5. | a) | What are Bayesian classifiers? With an example, describe how to predict a class label using Naive Bayesian classification. | [7M] |
| | b) | What is misclassification rate of a classifier? Describe sensitivity and specificity measures of a classifier. | [7M] |
| | | | |
| 6. | a) | What is Association rule mining? Briefly describe the criteria for classifying association rules. | [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 |

| | | | |
|---|---|---|---|
| 7. | a) | Describe any one Hierarchical clustering algorithm. | [7M] |
| | b) | What is cluster analysis? Describe the dissimilarity measures for interval-scaled variables and binary variables. | [7M] |

**\*\*\*\*\***

|''|'''||''||'''|||