Code No: RT32052 **R13** SET - 1

III B. Tech II Semester Supplementary Examinations, November/December-2016

DATA WARE HOUSING AND MINING

(Common to CSE and IT)

Time: 3 hours Maximum Marks: 70

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

*****

PART –A

1	a)	What is outlier mining? Define Data characterization.	[3M]
	b)	Define data mining? Mention the steps in the data mining process?	[4M]
	c)	What is clustering? What are the requirements of clustering?	[4M]
	d)	Define Dimensional Modeling? List out its advantages.	[4M]
	e)	Merits of Data Warehouse. What are the characteristics of Data Warehouse?	[4M]
	f)	What is support and confidence? What is its purpose in association mining?	[3M]

PART -B

2	a)	How is a *data warehouse* different from a database? How are they similar?	[5M]
	b)	Explain the OLAP operations in multidimensional model?	[7M]
	c)	Discuss the components of Data warehouse?	[4M]
3	a)	How might you determine *outliers* in the data? What other methods are there for *data smoothing*?	[5M]
	b)	List out and describe the primitives for specifying a data mining task.	[6M]
	c)	i) What are the value ranges of the following *normalization methods*? (a) min-max normalization (b) z-score normalization (c) normalization by decimal scaling ii) Use the two methods below to *normalize* the following group of data: 200*;* 300*;* 400*;* 600*;* 1000 (a) min-max normalization by setting *min* = 0 and *max* = 1 (b) z-score normalization	[5M]
4	a)	Briefly discuss about data mining task premitives.	[8M]
	b)	What is data mining? Draw and explain the architecture of a typical data mining system?	[8M]
5	a)	Compare the Advantages and Disadvantages of ***Eager Classification*** (e.g., decision tree, Bayesian, neural network) versus ***Lazy Classification*** (e.g., k-nearest neighbor, case-based reasoning).	[9M]
	b)	Explain the issues regarding Classification and Prediction?	[7M]

6a) Consider the following data set for a binary class problem. [9M]

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

i) Calculate the information gain when splitting on *A* and *B*. Which attribute would the decision tree induction algorithm choose?
ii) Calculate the gain in the Gini index when splitting on *A* and *B*. Which attribute would the decision tree induction algorithm choose?

b) What is decision tree? Explain the algorithm for generating a decision tree with a suitable example? [7M]

7a) What is cluster Analysis? Briefly explain K-means also with an example? Write its advantages and disadvantages. [9M]

b) Explain the Model-based method of clustering? [7M]

*****