Seat No.: _____ Enrolment No._____

GUJARAT TECHNOLOGICAL UNIVERSITY

BE - SEMESTER– VII • EXAMINATION – WINTER 2014

Subject Code: 171601 Date: 25/11/2014

Subject Name: Data Warehousing and Data Mining

Time: 10:30am TO 01:00 pm Total Marks: 70

Instructions:

1. **Attempt all questions.**
2. **Make suitable assumptions wherever necessary.**
3. **Figures to the right indicate full marks.**

Q.1 **(a)** Define KDD. How data mining techniques applied over multimedia database, temporal database and spatial database to extract useful knowledge. **07**

(b) What is concept hierarchy? List and explain types of concept hierarchy in detail. **07**

Q.2 **(a)** What is data cleaning? Discuss various ways of handling missing values during data cleaning. **07**

(b)

1. Explain Star and Fact Galaxy schemas used in data warehouse for multidimensional database. **05**
2. Differentiate OLAP vs. OLTP. **02**

OR

(b)

1. What is Cuboid? Explain various OLAP operations on data cube with suitable example. **05**
2. Differentiate Fact table vs. Dimension table. **02**

Q.3 **(a)** Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order): **07**

13, 15, 16, 16, 19, 20, 23, 29, 35, 41, 44, 53, 62, 69, 72

i) Use min-max normalization to transform the value 45 for age onto the range [0:0, 1:0]

ii) Use z-score normalization to transform the value 45 for age, where the standard deviation of age is 20.64 years.

(b) State the Apriori Property. Generate large itemsets and association rules using Apriori algorithm on the following data set with minimum support value and minimum confidence value set as 50% and 75% respectively. **07**

TID	*Items Purchased*
T101	Cheese, Milk, Cookies
T102	Butter, Milk, Bread
T103	Cheese, Butter, Milk, Bread
T104	Butter, Bread

OR

Q.3 **(a)** What is noise? Explain data smoothing methods as noise removal technique to divide given data into bins of size 3 by bin partition (equal frequency), by bin means, by bin medians and by bin boundaries. Consider the data: 10, 2, 19, 18, 20, 18, 25, 28, 22 **07**

(b) List two shortcomings of the algorithms which helped in improving the efficiency of Apriori algorithm. Discuss any TWO variations of the Apriori algorithm to improve the efficiency. **07**

Q.4 **(a)** How K-Mean clustering method differs from K-Medoid clustering method? Discuss the process of K-Mean clustering. Also outline major drawbacks of K-Mean clustering technique. **07**

(b) Explain how the accuracy of a classifier can be measured. How *Bagging* strategy helps improving the classifier accuracy? **07**

OR

Q.4 **(a)** What is supervised learning? Using the given table, show how the ROOT splitting attribute is selected using *InfoGain* measure in the overall process of decision tree induction. **07**

No.	Attributes				Class
	Outlook	Temperature	Humidity	Windy	
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	True	N
3	Overcast	Hot	High	False	P
4	Rain	Mild	High	False	P
5	Rain	Cool	Normal	False	P
6	Rain	Cool	Normal	True	N
7	Overcast	Cool	Normal	True	P
8	Sunny	Mild	High	False	N
9	Sunny	Cool	Normal	False	P
10	Rain	Mild	Normal	False	P
11	Sunny	Mild	Normal	True	P
12	Overcast	Mild	High	True	P
13	Overcast	Hot	Normal	False	P
14	Rain	Mild	High	True	N

(b) Explain Linear Regression and Non-linear Regression techniques of prediction. **07**

Q.5 **(a)** What is web log? Explain web structure mining and web usage mining in detail. **07**

(b) Discuss the application of data warehousing and data mining in government sector. **07**

OR

Q.5 **(a)** Explain the information retrieval methods used in text mining. **07**

(b) What are neural networks? Describe the various factors which make them useful for classification and prediction in data mining. Explain how the topology of neural network is designed. **07**

****