

GUJARAT TECHNOLOGICAL UNIVERSITY**MCA - SEMESTER-IV • EXAMINATION – SUMMER • 2014****Subject Code: 640005****Date: 03-06-2014****Subject Name: Data Warehousing and Data Mining (DWDM)****Time: 10:30 am - 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) What is Data Warehouse? **02**
 Why it is good to have a separate data warehouse? **02**
 Differentiate between Data warehouse and Operational Databases. **03**
- (b) i) What is Concept Hierarchy? **02**
 ii) Differentiate between classification and prediction **02**
 iii) What is Bitmap and Join Indexing? **03**
- Q.2** (a) What is Data Mining? **02**
 Explain in brief Knowledge discovery Process (KDD). **05**
- (b) Explain in brief Data mining major issues. **07**
- OR**
- (b) Why data in real world is dirty? **02**
 Discuss various ways to handle missing and noisy data. **05**
- Q.3** (a) Write and explain Apriori algorithm. **07**
- (b) i) Suppose a data warehouse of a hospital consists of four dimensions, namely time, doctor, location and patient with the concept hierarchy as follows:
 Time: day, month, quarter, year
 Doctor: doctor, specialization (e.g. pediatrician, etc.)
 Patient: patient, category (e.g. outdoor, indoor), address
 Location: location, address
- There are two measures, namely count and charge, where charge is doctor's consultation fees. **02**
- Draw a star and snowflake schema for the above data warehouse. **02**
- ii) Data: 50, 75, 125, 200, 150.
 Normalize the above data using following normalization methods:
 1) Min max normalization with min = 0 and max = 1. **02**
 2) Decimal scaling normalization. **01**
- OR**
- Q.3** (a) i) Differentiate between OLAP and OLTP. **04**
 ii) What is SelfExp, InExp and PathExp? **03**

- (b) A database has four transactions. Let min_sup = 60% and min_conf = 80%.

TID	Date	items_bought
1	18-Jan-2011	{A, B, D, K}
2	18-Jan-2011	{A, B, C, D, E}
3	19-Jan-2011	{A, B, C, E}
4	22-Jan-2011	{A, B, C, D}

- i) Find all frequent itemsets using Apriori algorithm. **04**
 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 itemi denotes variables representing items (e.g. "A", "B", etc.): **03**
 For all X ∈ transaction,

$$\text{buys}(X, \text{item1}) \wedge \text{buys}(X, \text{item2}) \Rightarrow \text{buys}(X, \text{item3}) [s, c]$$

- Q.4** (a) What is decision tree? **02**
 Discuss decision tree algorithm. **05**
 (b) i) Explain in brief attribute oriented induction. **04**
 ii) Explain Rollup, Drill down and "Slice & Dice" operations. **03**

OR

- Q.4** (a) Explain how to extract classification rules from decision trees. **07**
 (b) i) Please comment: "Strong rules are not necessarily interesting." **04**
 ii) What is Meta data? List contents of it. **03**

- Q.5** (a) Explain in brief outlier Analysis. **05**
 (b) Explain Data mining in Retail industry. **05**
 (c) Discuss Bagging and boosting methods for increasing the accuracy. **04**

OR

- Q.5** (a) Compare K-mean and K-medoids methods of clustering. **05**
 (b) How Data mining is useful in Telecom industry? **05**
 (c) Write Short Note: Bayes Theorem. **04**
