ROLL NO. ___________

Code: AT78 Subject: DATA MINING & WAREHOUSI

AMIETE – IT

Time: 3 Hours **JUNE 2013** **Max. Marks: 10**

PLEASE WRITE YOUR ROLL NO. AT THE SPACE PROVIDED ON EACH PAGE IMMEDIATELY AFTER RECEIVING THE QUESTION PAPER.

NOTE: There are 9 Questions in all.

- **Question 1 is compulsory and carries 20 marks. Answer to Q.1 must be written in the space provided for it in the answer book supplied and nowhere else.**
- **The answer sheet for the Q.1 will be collected by the invigilator after 45 minutes of the commencement of the examination.**
- **Out of the remaining EIGHT Questions answer any FIVE Questions. Each question carries 16 marks.**
- **Any required data not explicitly given, may be suitably assumed and stated.**

Q.1 Choose the correct or the best alternative in the following: (2×10)

a. Data mining can be classified as

(A) Database Technology **(B)** Machine learning
(C) Visualization **(D)** All of these

b. Data scrubbing is defined as

(A) a process to reject data from the data warehouse and to create the necessary indexes.
(B) a process to load the data in the data warehouse and to create the necessary indexes.
(C) a process to upgrade the quality of data after it is moved into a data warehouse.
(D) a process to upgrade the quality of data before it is moved into a data warehouse.

c. A star schema has what type of relationship between a dimension and fact table

(A) Many-to-many **(B)** One-to-one
(C) One-to-many **(D)** Many to one

d. Which of the following is the extract process?

(A) Capturing all of the data contained in various operational systems
(B) Capturing a subset of the data contained in various operational systems
(C) Capturing all of the data contained in various decision support systems
(D) Capturing a subset of the data contained in various decision support systems

e. Issues related to classification and prediction can be

(A) Data Cleaning
(B) Relevance Analysis
(C) Data Transformation and reduction
(D) All of these

f. An operational system is which of the following?

(A) A system that is used to run the business in real time and is based on historical data.
(B) A system that is used to run the business in real time and is based on current data.
(C) A system that is used to support decision making and is based on current data.
(D) A system that is used to support decision making and is based on historical data.

g. A multifield transformation

(A) Converts data from one field into multiple fields
(B) Converts data from multiple fields into one field
(C) Converts data from multiple fields into multiple fields
(D) All of these

h. Reconciled data is

(A) Data stored in the various operational systems throughout the organization
(B) Current data intended to be the single source for all decision support system
(C) Data stored in one operational system in the organization
(D) Data that has been selected and formatted for end-user support applications

i. ___________ stores multidimensional aggregate information.

(A) Data cube **(B)** Data Mart
(C) Both **(A)** & **(B)** **(D)** None of these

j. Which of the following process is the load and index?

(A) A process to reject data from the data warehouse and to create the necessary indexes.
(B) A process to load the data in the data warehouse and to create the necessary indexes.
(C) A process to upgrade the quality of data after it is moved into a data warehouse.
(D) A process to upgrade the quality of data before it is moved into a data warehouse.

ROLL NO. ___________

Code: AT78 Subject: DATA MINING & WAREHOUS

Answer any FIVE Questions out of EIGHT Questions.
Each question carries 16 marks.

Q.2 a. What is the difference between discrimination and classification? **(4)**

b. Describe three challenges to data mining regarding data mining methodology and user interaction issues. **(8)**

c. Describe why 'concept hierarchies' are useful in data mining. **(4)**

Q.3 a. Explain the following concepts:
(i) Data transformation
(ii) Data Reduction **(8)**

b. In real world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. **(8)**

Q.4 a. Suppose that a data warehouse consists of the four dimensions: date, spectator, location, and game, and the two measures: count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.
(i) Draw a star schema diagram for the data warehouse
(ii) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM Place in 2010? **(8)**

b. In data warehouse technology, a multiple dimensional view can be implemented by a relational database technique (ROLAP), or by a multidimensional database technique (MOLAP), or by a hybrid database technique (HOLAP).
(i) Briefly describe each implementation technique.
(ii) For each technique, explain how each of the following functions may be implemented:

- The generation of a data warehouse (including aggregation)
- Roll-up
- Drill-down
- Incremental updating **(8)**

Q.5 a. Explain Multiway Array Aggregation for full cube computation. **(8)**

b. Discuss how to support quality drill down although some low level cells may contain empty or too less data for reliable analysis. **(8)**

Q.6 a. Give a short example to show that items in a strong association rule may actually be negatively correlated. **(8)**

b. Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning? **(8)**

Q.7 a. Why is naive Bayesian classification called “naive”? Briefly outline the m ideas of naive Bayesian classification. **(8)**

b. What are ensemble methods? Explain in detail with the help of algorithm. **(8)**

Q.8 a. Briefly describe and give examples of each of the following approaches to clustering: partitioning methods, hierarchical methods, density-based methods, grid-based methods and model based methods. **(10)**

b. Describe any two of the following 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

- k-means
- k-medoids
- CLARA
- BIRCH **(6)**

Q.9 a. Why is the establishment of theoretical foundations important for data mining? Name and describe the main theoretical foundations that have been proposed for data mining. **(8)**

b. What are the differences between visual data mining and data visualization? **(4)**

c. Describe a situation in which you feel that data mining can infringe on your privacy. **(4)**