ROLL NO. _____

Code: AT78 Subject: DATA MINING & WAREHOUSING

AMIETE – IT

Time: 3 Hours **DECEMBER 2012** **Max. Marks: 100**

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. Bayes Theorem is:

(A) P(H|X)=P(X|H)P(H)/P(X) **(B)** P(H|X)=P(X|H)P(X)/P(H)
(C) P(X|H)=P(X|H)P(H)/P(X) **(D)** P(X|H)=P(X|H)P(X)/P(H)

b. The generic two-level data warehouse architecture includes which of the following?

(A) At least one data mart
(B) Data that can be extracted from numerous internal and external sources
(C) Near real-time updates
(D) All of these

c. Which of the following is not a Data Mining Functionality?

(A) Association analysis **(B)** Cluster Analysis
(C) Classification & Prediction **(D)** Randomization

d. Which of the key words does not distinguish warehouses from other repositories?

(A) Subject-oriented **(B)** Time –variant
(C) Volatile **(D)** Integrated

e. Which of the following is not an OLAP server?

(A) ROLAP **(B)** DOLAP
(C) MOLAP **(D)** HOLAP

f. Which of the following is a data smoothing technique?

(A) Histogram **(B)** Regression
(C) Correlation **(D)** Induction

g. Which of the following is not a basis for classification of an association rules?

(A) types of values **(B)** dimensions of data involved
(C) volume of data involved **(D)** levels of abstractions involved

h. A decision tree is a flow chart like structure where:

(A) each internal node denotes a test on an attribute
(B) each internal node is of type "yes" or "no"
(C) each branch represents an outcome of a test
(D) the leaf nodes represents the classes or class distribution

i. Which of the following is not a category of clustering methods?

(A) Partitioning method **(B)** Relational method
(C) Hierarchical method **(D)** Grid-based method

j. Which of the following is not a commercial Data Mining System.

(A) Mineset **(B)** Clementine
(C) D2kMiner **(D)** Enterprise Miner

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

Q.2 a. List and describe the five primitives for specifying a data mining task. **(8)**

b. How data mining (DM) is different from knowledge discovery in databases (KDD)? Explain. **(8)**

Q.3 a. Use the following two methods below to normalize the given group of data:
200, 300, 400, 600, 1000
(i) min-max normalization by setting min = 0 and max = 1
(ii) Z-score normalization **(8)**

b. Explain the Three-tier data warehouse architecture. What are the three data warehouse models from architecture point of view. Explain **(8)**

Q.4 a. Explain different types of problems in data, which the data-cleaning methods can deal. What are the different methods to deal with "missing values"? **(9)**

b. State why, for the integration of multiple heterogeneous information source, many companies in industry prefer the update-driven approach (which

constructs and uses data warehouse), rather than the query-driven approach (which applies wrappers and integrators). Describe situation where the query-driven approach is preferable over the update-driven approach. **(7)**

Q.5 a. How does data mining relate to information processing and OLAP? Discuss in detail. **(9)**

b. For class characterization, what are the major differences between a data cube-based implementation and a relational implementation such as attribute-oriented induction? Discuss which method is most efficient and under what conditions this is so. **(7)**

Q.6 a. Discuss the criterion used for comparison and evaluation of the classification and prediction method. **(8)**

b. Association rule mining often generates a large number of rules. Discuss effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules. **(8)**

Q.7 a. What is boosting? State why it may improve the accuracy of decision tree induction. **(8)**

b. How classification is done by back-propagation? Give an example of a general multilayered feed-forward neural network. **(8)**

Q.8 a. Discuss the various types of typical requirements for clustering in data-mining. **(9)**

b. What are model based clustering methods? Discuss two major approaches viz. Statistical approach and neural network based approach used here. **(7)**

Q.9 a. Discuss how the data mining can be applied to Biomedical & DNA Data analysis. **(8)**

b. What are the major challenges faced in bringing data mining research to market? Illustrate one data mining research issue that, in your view, may have a strong impact on the market and on society. Discuss how to approach such a research issue. **(8)**