Subject: DATA MINING & WAREHOUSING

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

JUNE 2011

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. Which of the following represent information granularity?

(A) The extent of accuracy within the information can be updated by users.
(B) The extent of subjectivity within the information.
(C) The extent of detail within the information.
(D) The extent of strategy within the information.

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

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

c. In the data warehouse architecture, the _____ component interleaves with and connects other components.

(A) Metadata **(B)** Data cube
(C) Data mart **(D)** none of the above

d. 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)** All of the above

e. Data transformation includes which of the following?

(A) A process to change data from a detailed level to a summary level
(B) A process to change data from a summary level to a detailed level
(C) Joining data from one source into various sources of data.
(D) Separating data from one source into various source of data

f. 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)

g. Time variant nature of the data in the data warehouse includes

(A) Allows for the analysis of the past
(B) Relate information to the present
(C) Enables forecast the future
(D) All of the above

h. Which of the following is the extract process

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

i. _____ stores multidimensional aggregate information.

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

j. Data in _____ are typically fragmented and inconsistent.

(A) Data cube **(B)** Data Mart
(C) Operational systems **(D)** data warehouse

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

Q.2 a. Explain the architecture of a data mining system. **(8)**

b. Define each of the following data mining functionalities:
(i) Classification
(ii) Prediction
(iii) Evolution analysis
(iv) Clustering **(8)**

Q.3 a. Explain how to handle missing value in the data cleaning process. **(8)**

b. Discuss various steps of data preprocessing. **(8)**

Q.4 a. Discuss the problem related to use and storage of unstructured data in a data warehouse. What are the methods to capture and store external information? **(8)**

b. Explain the snowflake schema. Give suitable examples. **(8)**

Q.5 a. Describe two common approaches to control a generalization process. Illustrate with suitable example. **(8)**

b. How is class comparison performed? Explain the procedure. **(8)**

Q.6 a. Why are decision tree classifiers so popular? Explain. **(8)**

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

Q.7 a. Why is naïve Bayesian classification called "naive"? Briefly outline the major ideas of naïve Bayesian classification. **(8)**

b. Show that accuracy is a function of sensitivity and specificity. **(8)**

Q.8 a. What you mean by cluster analysis? What are the typical requirements of clustering in data mining? **(8)**

b. Briefly explain the model based clustering methods. **(8)**

Q.9 a. Describe few major applications of data mining in the business area. **(8)**

b. Write brief note on the following trends in data mining:

(i) Web mining.
(ii) Distributed data mining.
(iii) Real time data mining.
(iv) Biological data mining. **(8)**