ROLL NO. _____

Code: AT78 **Subject: DATA MINING & WAREHOUSING**

AMIETE – IT (Current Scheme)

Time: 3 Hours **December - 2017** **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. Strategic value of data mining is
(A) cost-sensitive **(B)** work-sensitive
(C) time-sensitive **(D)** technical-sensitive

b. Which view exposes the information being captured, stored, and managed by operational systems?
(A) top-down view **(B)** data warehouse view
(C) data source view **(D)** business query view

c. _____ stores multidimensional aggregate information.
(A) Data cube **(B)** Data Mart
(C) Both **(A)** & **(B)** **(D)** None of these

d. The load and index is
(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

e. 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)** $P(X|H)=P(X|H)P(X)/P(H)$

f. Which technique allows more data to be loaded into a single block?
(A) Compaction **(B)** Indexing
(C) Partitioning **(D)** Clustering

g. Which of the following schema supports the normalization in dimensional modeling?
(A) Star schema **(B)** Snow-Flake schema
(C) Fact-Constellation **(D)** None of these

h. What is created in association with metadata on inclusion of an external data in the data warehouse?
(A) Data Mart **(B)** Notification data
(C) External reference **(D)** Structure of data

i. 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

j. Data mining requires
(A) large quantities of operational data stored over a period of time
(B) lots of tactical data
(C) several tape drives to store archival data
(D) large mainframe computers

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

Q.2 a. Explain the different criteria based on which the data mining systems can be categorized. **(8)**

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

Q.3 a. What are various steps of data preprocessing? **(8)**

b. What do you mean by data reduction? Explain any two data reduction techniques. **(8)**

Q.4 a. Explain the architecture of a data warehouse. Also explain the single-tier and three tier architectures of a data warehouse. **(8)**

b. Explain different types of problems in data, which the data-cleaning methods can deal. What are the different methods to deal with “missing values”? **(8)**

Q.5 a. What is attribute oriented induction? What is its use in data characterization? Explain with examples. **(3+5)**

b. Explain Multiway Array Aggregation for full cube computation. **(8)**

Q.6 a. What is Data Classification Process? How it is differ than Predication? **(8)**

b. How does tree pruning work? Explain with examples two common approaches to tree pruning – pre pruning and post pruning. **(8)**

Q.7 a. Prediction is frequently referred to as the forecasting of missing numerical values. Justify the statement with an example. **(8)**

b. Why is naïve Bayesian classification called "naïve"? Briefly outline the major ideas of naïve Bayesian classification. **(8)**

Q.8 a. What is cluster analysis? How is it different from classification? Discuss the typical requirements of clustering in data mining. **(8)**

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. **(8)**

Q.9 a. What is collaborative recommender system? In what ways does it differ from customer or product based clustering system? Discuss the major challenges faced by collaborative recommender systems. **(8)**

b. 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)**