ROLL NO. _____

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

AMIETE – IT (Current Scheme)

Time: 3 Hours **June 2019** **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. A data warehouse is a database, which is kept separate from the organization's _____database.

(A) Historical data **(B)** Operational
(C) Both A & B **(D)** None of these

b. Operational _____of data does not occur in the data warehouse environment.

(A) Refresh **(B)** Update & Refresh
(C) Update (**D)** None of these

c. Ordinal attribute:

(A) Has a fixed size of interval between data points.
(B) Represent rank order
(C) Has a true zero point
(D) None of these

d. Generalization of the data:

(A) Where summary or aggregation operations are applied to the data.
(B) Where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies.
(C) Where the attribute data are scaled so as to fall within a small specified range
(D) *Where new attributes are constructed* and added from the given set of attributes to help the mining process.

e. OLAP query needs only _____access of stored data.

(A) Write **(B)** Read and write both
(C) Read (**D)** None of these

f. A decision tree is a flow-chart-like tree structure where Internal node denotes a/an _____ on an attribute and branch represents a/an _____of the test.

(A) Test; outcome **(B)** Outcome; test

(C) Test; test **(D)** Outcome; outcome

g. decision tree classifiers are so popular because:

(i) The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for knowledge discovery.

(ii) The learning and classification steps of decision tree induction are simple and fast.

(iii) Decision trees can handle high dimensional data.

(iv) Representation of acquired knowledge in tree form is generally easy to humans.

(A) Only (i) & (ii) **(B)** Only (i), (ii), (iv)

(C) All of these (**D)** None of these

h. What is an essential process where intelligent methods are applied to extract data patterns?

(A) Data Selection **(B)** Data Warehousing

(C) Data Extraction **(D)** Data Mining

i. Extreme values that occur infrequently are called as

(A) Rare Values **(B)** Outliers

(C) Dimensionality reduction **(D)** Both A and B

j. Association rules are associated with:

(A) Population, confidence **(B)**Population, support, confidence

(C) Support, confidence **(D)** None of these

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

Q.2 a. What is *data mining*? In your answer, address the following: **(8)**

i. Explain how the evolution of database technology led to data mining.

ii. Describe the steps involved in data mining when viewed as a process of knowledge discovery.

b. Present an example where data mining is crucial to the success of a business. What *data mining functions* does this business need? Can they be performed alternatively by data query processing or simple statistical analysis? **(8)**

Q.3 a. In real-world data, tuples with *missing values* for some attributes are a common occurrence. Describe various methods for handling this problem. **(8)**

b. Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Use *smoothing by bin means* to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data. **(8)**

Q.4 a. State why, for the integration of multiple heterogeneous information sources, many companies in industry prefer the *update-driven approach* (which constructs and uses data warehouses), rather than the *query-driven approach* (which applies wrappers and integrators). Describe situations where the query-driven approach is preferable over the update-driven approach. **(10)** **(6)**

b. Briefly compare the concepts of Snowflake schema and Star schema. Use an example to explain your point(s).

Q.5

a. Often, the aggregate measure value of many cells in a large data cuboid is zero, resulting in a huge, yet sparse, multidimensional matrix. **(10)**

Design an implementation method that can elegantly overcome this sparse matrix problem. Note that you need to explain your data structures in detail and discuss the space needed, as well as how to retrieve data from your structures.

b. What is Generalization? How it is useful in data transformation? Give an example. **(6)**

Q.6 a. The Apriori algorithm uses *prior knowledge* of subset support properties. Prove that all nonempty subsets of a frequent itemset must also be frequent. **(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. Why is *naive Bayesian classification* called "naïve"? Briefly outline the major ideas of *naive* Bayesian classification. **(8)**

b. It is difficult to assess classification *accuracy* when individual data objects may belong to more than one class at a time. In such cases, comment on what criteria you would use to compare different classifiers modeled after the same data. **(8)**

Q.8 a. Briefly outline how to compute the *dissimilarity* between objects described by the following types of variables: Numerical (interval-scaled) variables; Asymmetric binary variables; Categorical variables **(8)**

b. Give an example of how specific clustering methods may be *integrated*, for example, where one clustering algorithm is used as a preprocessing step for another. In addition, provide reasoning on why the integration of two methods may sometimes lead to improved clustering quality and efficiency. **(8)**

Q.9 a. What is a collaborative recommender system? In what ways does it differ from a customer- or product- based clustering system? How does it differ from a typical classification or predictive modeling system? **(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)**