ROLL NO. _____

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

AMIETE – IT

Time: 3 Hours **DECEMBER 2016** **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. 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

b. Fact tables are _____.

(A) Completely not normalized **(B)** Completely normalized
(C) Partially not normalized **(D)** Partially normalized

c. Extreme values that occur infrequently are called as _____.

(A) Rare Values **(B)** Outliers
(C) Dimensionality reduction **(D)** Both **(A)** and (B)

d. All set of items whose support is greater than the user-specified minimum support are called as_____.

(A) Maximal frequent set **(B)** Frequent set
(C) Border sector lattice **(D)** Both **(A)** and (B)

e. The number of iterations in a priori is _____.

(A) Increases with the size of the maximum frequent set.
(B) Decreases with increase in size of the maximum frequent set.
(C) Increases with the size of the data.
(D) Both **(A)** and **(C)**

f. In which algorithm each cluster is represented by the center of gravity of the cluster.

(A) K- Medoid **(B)** ROCK
(C) K- Mean **(D)** Both **(A)** and **(C)**

g. The transformed prefix paths of a node 'a' form a truncated database of pattern which co-occur with 'a' is called _____.

(A) Suffix path **(B)** Conditional pattern base
(C) Conditional FP-Tree **(D)** All of these

h. In web mining, _____ is used to know which URLs tend to be requested together.

(A) Clustering **(B)** Association Rule Mining
(C) Classification **(D)** Both **(A)** and **(C)**

i. _____ describes the discovery of useful information from the web contents.

(A) Web usage mining **(B)** Web Structure mining
(C) Web content mining **(D)** All of these

j. Which one manages both current and historic transactions?

(A) OLTP **(B)** Both **(A)** and **(D)**
(C) RDBMS **(D)** OLAP

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

Q.2 a. Describe the steps involved in data mining when viewed as a process of knowledge discovery. **(4)**

b. What is *data mining*? Is it a simple transformation of technology developed from databases, statistics, and machine learning? **(4)**

c. Suppose your task as a software engineer at *Big-University* is to design a data mining system to examine their university course database, which contains the following information: the name, address, and status (e.g., undergraduate or graduate) of each student, the courses taken, and their cumulative grade point average (GPA). Describe the *architecture* you would choose. What is the purpose of each component of this architecture? **(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. Discuss issues to consider during *data integration*. **(8)**

Q.4 a. What is a data warehouse? How is a data warehousing different from a database? Enlist the characteristics of data warehouse. **(8)**

b. Explain stars and snowflakes schema using example. **(8)**

Q.5 a. There are several typical cube computation methods, such as *multiway array computation* (MultiWay) , *BUC* (Bottom-Up Computation), and *Star-Cubing*. Briefly describe these three methods (i.e., use one or two lines to outline the key points), and compare their feasibility and performance under the following conditions:
(i) Computing a dense full cube of low dimensionality (e.g., less than 8 dimensions)
(ii) Computing an iceberg cube of around 10 dimensions with a highly skewed data distribution
(iii) Computing a sparse iceberg cube of high dimensionality (e.g., over 100 dimensions) **(12)**

b. What is Generalization? How it is useful in data transformation? Give an example. **(4)**

Q.6 a. What is frequent pattern analysis? Give a typical example of frequent itemset mining. **(4)**

b. Comments on the following: *constrained frequent itemsets, approximate frequent itemsets, near-match frequent itemsets, and top-k frequent itemsets*. **(6)**

c. What is the goal of data classification? Enlist the similarities and dissimilarities between classification and prediction methods. **(6)**

Q.7 a. Why are decision tree classifiers so popular? Support your answer with relevant points. **(6)**

b. What is overfitting in decision tree? Discuss the approaches to avoid overfitting. **(6)**

c. Discuss IF-THEN rules for classification. How do you measure coverage and accuracy of a rule? **(4)**

Q.8 a. What is the cluster analysis? Give examples of clustering applications. **(8)**

b. Briefly outline how to compute the dissimilarity between objects described by asymmetric binary variables. **(8)**

Q.9 a. 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. **(10)**

b. Explain how data mining techniques are useful for implementation of Intrusion Detection System? **(6)**