ROLL NO. _____

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

AMIETE – IT (Current Scheme)

Time: 3 Hours **JUNE 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. _____ is an example for case based-learning.
(A) Decision trees **(B)** Neural networks
(C) Genetic algorithm **(D)** K-nearest neighbor

b. The next stage to data selection in KDD process _____.
(A) Enrichment **(B)** coding
(C) cleaning **(D)** reporting

c. Metadata describes _____
(A) A contents of the database
(B) Structure of the contents of database
(C) Structure of database
(D) Database itself

d. K-nearest neighbor is one of the _____.
(A) Learning Technique **(B)** OLAP Tool
(C) Search Technique **(D)** Datawarehousing Tool

e. In K-nearest neighbor algorithm, K denotes _____.
(A) Number of neighbors that are investigated
(B) Number of iterations
(C) Number of total records
(D) Random number

f. Mode of the data set {13, 3, 11, 24, 5, 3, 9, 2} is _____.
(A) 24 **(B)** 5
(C) 3 **(D)** None of these

g. OLAP is used to explore the _____ knowledge.
(A) Shallow **(B)** Deep
(C) Multidimensional **(D)** Hidden

h. Which of the following is the data mining tool?
(A) C **(B)** Weka
(C) C++ **(D)** VB

i. Removing duplicate records is a process called_____.
(A) recovery **(B)** data cleaning
(C) data cleansing **(D)** data pruning

j. MDDB stands for _____.
(A) Multiple Data Doubling **(B)** Multidimensional Databases
(C) Multiple Double dimension **(D)** Multi-Dimension Doubling

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

Q.2 a. Write the steps for discovering knowledge from database using data mining technique. What are the applications of data mining? **(8)**

b. Describe major challenges regarding data mining & society along with efficiency & scalability issue. **(8)**

Q.3 a. Suppose that the data for analysis includes the attribute total obtain marks in a particular subject. The marks 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, 35, 40, 45, 46, 52. **(10)**
(i) What is the mean of the data?
(ii) What is the median of the data?
(iii) What is the mode of the data?
(iv) What is the variance of the data?
(v) What is the standard deviation of the data?

b. Discuss briefly the different steps in preprocessing of data. **(6)**

Q.4 a. Explain the process of designing a data warehouse. **(8)**

b. Explain OLAP operations on multidimensional data. **(8)**

Q.5 a. What are the differences among the three main types of data warehouse usage: information processing, analytical processing, and data mining? **(8)**

b. Explain multidimensional data mining in cube space. **(8)**

Q.6 a. Explain Enterprise Warehouse, Data Mart, Virtual Warehouse and Metadata. **(8)**

b. Explain Bayesian Belief Networks. **(8)**

Q.7 a. Explain basic k-means algorithm. List the limitation of k-means algorithm. **(8)**

b. Explain the features one should look while selecting an appropriate data mining product available in the market. **(8)**

Q.8 Consider the following training data

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

(i) Draw the decision tree of these data. **(6)**

(ii) Calculate information gain of age, income, student, credit_rating relative to these training examples. **(5)**

(iii) What is the best split and gain ratio among ID, age, income, student, credit_rating according to information gain? **(5)**

Q.9 a. What are the requirements of clustering in data mining? Explain briefly. **(8)**

b. Explain backpropogation algorithm. **(8)**