ROLL NO. _____

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

AMIETE – IT (Current Scheme)

Time: 3 Hours **June 2018** **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. Bayesian classifiers is

(A) A class of learning algorithm that tries to find optimum classification of a set of examples using the probabilistic theory.

(B) Any mechanism employed by a learning system to constrain the search space of a hypothesis.

(C) An approach to the design of learning algorithm that is inspired by the fact that when people encounter new situations , they often explain them by reference to familiar experience, adapting the explanations to fit the new situation.

(D) None of these

b. KDD (Knowledge Discovery in Databases) is referred to

(A) Non-trivial extraction of implicit previously unknown and potentially useful information from data.

(B) Set of columns in a database table that can be used to identify each record within this uniquely.

(C) Collection of interesting and useful patterns in a database

(D) None of these

c. Online transaction processing is used because

(A) It is efficient **(B)** Disk is used for storing files

(C) It can handle random queries **(D)** Transactions occur in batches

d. _____ is used to load the information from operational database

(A) Replication technique **(B)** Reengineering technique

(C) Engineering technique **(D)** Transformation engineering

e. K-nearest neighbour is one of the _____.

(A) Learning technique **(B)** Clustering technique

(C) Purest search teanique **(D)** Data warehousing tool

f. The next stage to the data selection in KDD process is____.

(A) Enrichment **(B)** Coding

(C) Cleaning **(D)** Reporting

g. _____is the technique which is used for discovering patterns in dataset at the beginning of data mining process.
(A) Kohenon map **(B)** Visualization
(C) OLAP **(D)** SQL

h. A _____ acts as a bridge between data warehouse and database application.
(A) Data Mart **(B)** Operational data
(C) Meta data **(D)** Data Cube

i. Which one of the following is not true about OLAP?
(A) They create no new knowledge
(B) OLAP is powerful than data mining tool
(C) They cannot search for new solution
(D) OLAP tool store their data in special multidimensional format

j. The intermediate layers in a back –propagation network consists of _____.
(A) Photo receptors **(B)** Responders
(C) Hidden nodes **(D)** Associates

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

Q.2 Attempt all parts **(4x4=16)**

(i) What do you mean by data Mining? What are the techniques used in data cleaning process?
(ii) Describe why concept hierarchies are useful in data mining.
(iii) Explain all step in the process of knowledge Discovery form data (KDD).
(iv) What are Normalization techniques? Normalize the following group of data : 200,300,400,600,1000, for
a. Min-max normalization by setting min=0 and max=1
b. Z-score normalization
c. Decimal Normalization

Q.3 a. Briefly explain with examples OLAP operation on multidimensional data (Rollup, Drill down, Slice and dice and pivot). **(8)**

b. Give minimum six differences between OLAP and OLPT. **(8)**

Q.4 a. Briefly explain the following concepts. You may use an example to explain your point. **(8)**
(i) Snowflake schema
(ii) Fact constellation
(iii) Star Schema

b. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result:

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

Calculate
(i) The man, median, and standard deviation of age and % fat **(4)**
(ii) Draw the box plots for age and % fat **(4)**

Q.5 a. Find out the Information gain of age tuple in following table. **(8)**

RID	Age	Income	student	Credit raring	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

b. Explain the architecture of a data warehouse? Also explain the single tier & three tier architecture of a data warehouse. **(8)**

Q.6 a. Why tree pruning is useful in decision tree induction? What is the drawback of using a separate set of tuples to evaluate pruning? **(8)**

b. How does back propagation algorithm work? How can we design the topology of the neural network? Explain the input and output function in Hidden layer of Multilayer feed –Forward neural network. **(8)**

Q.7 a. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): **(8)**
a. Compute the Euclidean distance between the two objects.
b. Compute the Manhattan distance between the two objcets.
c. Compute the minkowski distance between the two objects, using q=3.

b. Explain following techniques used in cluster analysis **(8)**
(i) Chameleon
(ii) Clique

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

b. A data base has five transactions. Let min sup=60%and min conf=80% **(8)**

TID	Items bought
T100	M,O,N,K,E,Y
T200	D,O,N,K,E,Y
T300	M,A,K,E
T400	M,U,C,K,Y
T500	C,O,O,K,I,E

(i) Find all frequent item sets using Apriori
(ii) Find all the frequent rules.

Q.9 a. List out the application that the organization uses to build a query and reporting environment for the data warehouse?. **(8)**

b. What is spatial database? Explain the methods of mining spatial database. **(8)**