ROLL NO. _____

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

AMIETE – IT (Current Scheme)

Time: 3 Hours **JUNE 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. Which data mining technique is used to find correlation among given data?
(A) Association rule mining **(B)** Classification
(C) Clustering **(D)** Prediction

b. Which of the following statements is true?
(A) Data mining is the process of carry out multidimensional analysis
(B) Data mining is used for reporting
(C) Data mining is used for uncovering patterns that are present in data
(D) None of these

c. Which of the following statements are not true?
(A) Decision trees perform classification without requiring much computation.
(B) Decision trees can at best work on small samples of data and cannot easily approach large data sets
(C) Decision trees are more appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
(D) None of these

d. The process of standardizing the data that is extracted from the operational systems is known as:
(A) Transformation **(B)** Cleansing
(C) Population **(D)** Data Staging

e. This technique uses mean and standard deviation scores to transform real-valued attributes.
(A) Decimal scaling **(B)** Min-Max normalization
(C) Z-Score normalization **(D)** Logarithmic normalization

f. In which algorithm downward closure property must be satisfied?
(A) Apriori algorithm **(B)** K-nearest neighbor
(C) Decision tree induction **(D)** None of these

g. Which classification method does not involve sharp cut off for continuous attribute?
(A) decision tree induction **(B)** rule based classification
(C) fuzzy set approach **(D)** Bayesian classification

h. Which method is classified as a lazy learning?
(A) fuzzy set approach **(B)** genetic algorithm
(C) case base reasoning **(D)** rough set approach

i. Attributes may be eliminated from the target dataset during this step of the KDD process.
(A) Creating a target dataset **(B)** Data Pre-processing
(C) Data Transformation **(D)** Data Mining

j. The sequence of knowledge Discovery in data mining is
(A) Information -> Data -->Facts -->Knowledge
(B) Data ->Information -->Facts -->Knowledge
(C) Data -> Facts --> Information -->Knowledge
(D) None of these

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

Q.2 a. Discover the query examples in each one:
(i) Database processing
(ii) Data mining processing
Your answer must clearly highlight the queries to make clear difference between above two kind of processing. **(8)**

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. Explore the following types of datasets: <give example in each one> **(6)**
(i) Record
(ii) Graph
(iii) Ordered

b. What do you mean by descriptive data summarization techniques? What are the major roles of these kinds of techniques in data pre-processing? What set of measures are used in both, central tendency and dispersion of the data? Highlight each measure with the help of examples. **(10)**

Q.4 a. Why a data warehouse is separated from operational databases? Clearly indicate the reasons. **(8)**

b. What are the differences between the three main types of data warehouse usage: *information processing*, *analytical processing*, and *data mining?* Discuss the motivation behind *OLAP mining*. **(8)**

Q.5 a. Given a generalized relation, *R*, derived from a database, *DB*, suppose that a set, S-*DB*, of tuples needs to be deleted from *DB*. Outline an *incremental* updating procedure for applying the necessary deletions to *R*. **(8)**

b. For *class characterization*, what are the major differences between a data cube-based implementation and a relational implementation such as attribute-oriented induction? Discuss which method is most efficient and under what conditions this is so. **(8)**

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. **(6)**

b. What do you mean by classification and predication? In data mining, forecasting value may be discrete or continuous, clearly indicate where to use classification as well as prediction? **(4)**

c. Classification is a 3-step process; clearly discuss each step with its involved activities. **(6)**

Q.7 a. Define Bayesian Classification. Given a ***data tuple*** having the values *"sunny", "cool", "high" and "true"* for the attributes outlook, temp., humidity and windy respectively. What would be a naive Bayesian classification of the *Play* for the given tuple? Consider below dataset. **(10)**

Outlook	Temp	Humidity	Windy	Play?
sunny	hot	high	FALSE	**No**
sunny	hot	high	TRUE	**No**
overcast	hot	high	FALSE	**Yes**
rainy	mild	high	FALSE	**Yes**
rainy	cool	normal	FALSE	**Yes**
rainy	cool	Normal	TRUE	**No**
overcast	cool	Normal	TRUE	**Yes**
sunny	mild	High	FALSE	**No**
sunny	cool	Normal	FALSE	**Yes**
rainy	mild	Normal	FALSE	**Yes**
sunny	mild	normal	TRUE	**Yes**
overcast	mild	High	TRUE	**Yes**
overcast	hot	Normal	FALSE	**Yes**
rainy	mild	high	TRUE	**No**

b. Consider the following two-class confusion matrix. The matrix shows the classification results of a supervised model. The task of this model is to determine the action (Malicious or Benign).

		True Class	
		Malicious	Benign
Predicated Class	Malicious	70	10
	Benign	40	380

(i) Compute the precision of this model with respect to the malicious class.
(ii) Compute the recall of this model with respect to the malicious class. **(6)**

Q.8 a. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): **(6)**
(i) Compute the *Euclidean distance* between the two objects.
(ii) Compute the *Manhattan distance* between the two objects.
(iii) Compute the *Minkowski distance* between the two objects, using $p = 3$.

b. Both *k-means* and *k-medoids* algorithms can perform effective clustering. Illustrate the strength and weakness of *k-means* in comparison with the *k- medoids* algorithm. Also, illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme (such as AGNES). **(10)**

Q.9 a. There is a strong linkage between *statistical data analysis* and data mining. Some people think of data mining as automated and scalable methods for statistical data analysis. Do you agree or disagree with this perception? Present one statistical analysis method that can be automated and/or scaled up nicely by integration with the current data mining methodology. **(8)**

b. General-purpose computers and domain-independent relational database systems have become a large market in the last several decades. However, many people feel that generic data mining systems will not prevail in the data mining market. What do you think? For data mining, should we focus our efforts on developing *domain-independent* data mining tools or on developing *domain-specific* data mining solutions? Present your reasoning. **(8)**