BE6-R4: DATA WAREHOUSING AND DATA MINING

NOTE:

1. **Answer question 1 and any FOUR from questions 2 to 7.**
2. **Parts of the same question should be answered together and in the same sequence.**

Time: 3 Hours **Total Marks: 100**

1.

a) Differentiate between Classification and Regression with the help of an example.
b) Differentiate between ordinal and nominal data attributes with the help of an example.
c) What is the motivation for building a data warehouse?
d) Give example of two business problems that can be solved by clustering.
e) What is data cleaning? List two methods of data cleaning.
f) Give two applications of text mining explaining the objective and the data mining task employed to achieve it.
g) What is a meant by Online analytical processing?

(7x4)

2.

a) What is the sequence of tasks to load data in warehouse? Explain in detail including the tools used.
b) What is meant spatio-temporal mining? Describe with the help of one application.
c) Where supervised and unsupervised learning are used and what purpose do they serve?

(6+6+6)

3.

a) In a star schema for tracking the shipments for a distribution company, the following dimension tables are present:
   i) Time
   ii) Customer ship-to
   iii) Ship-from
   iv) Product

   List three possible attributes for each of the dimension tables and designate a primary key.
b) Which data mining task can accomplish the following? Give reasons for your answer in not more than 30 words.
   i) partitioning customer database into three groups
   ii) finding whether a telephone call is going to fail or succeed based on the past experience of the incoming and outgoing traffic from exchange.

(12+6)

4.

a) Write an algorithm for K-Nearest neighbor classification?
b) Explain in detail the FASMI characteristics of OLAP system.
c) Explain through an example as how and why information gain is used to construct a decision tree?

(6+6+6)

5. A telecom company intends to mine the data of call detail records containing fields:
   i) calling no.
   ii) called no.
   iii) call duration
   iv) time of day
   v) type of calling phone (cdma/gps/gprs)
   vi) type of called phone (cdma/gps/gprs)
   vii) type of connection (prepaid/ postpaid)

viii) number of intermediate exchanges to route the call
ix) call status (success/failure)
x) facilities used (call diversion, call waiting).

a) List two fields which can be removed to reduce dimensionality. Justify.
b) Which two fields can be discretized? Justify.
c) List two interesting associations that can be examined. The antecedent must have at-least two items.

(6+6+6)

6.

a) What is Naive Bayes classifies? What is the weakness of the assumption in the method?
b) Generate all possible association rules with confidence values, from the following set of item-sets, such that the antecedent has exactly one item. The support is given against each.

{(a:10),(b:12),(d:8),(f:16),(a d: 6),(a b:8),(a f:10),(a b d:6),(a b f:4),(a b d f:3)}

(9+9)

7. Write short notes on **three** of the following:

a) Neural networks
b) Multimedia databases
c) K-means clustering algorithm
d) Apriori Algorithm for mining association rules

(3x6)