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) What is the difference between gini index and entropy?
b) Describe three challenges to data mining regarding user interaction issues in data mining.
c) With the help of an example briefly compare data cleaning, data transformation and refresh.
d) Association rule mining often generates a large number of rules, how? Discuss effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules.
e) What is data reduction? Discuss the techniques used for reducing data in data mining.
f) Describe why concept hierarchies are useful in data mining.
g) What is boosting? State why it may improve the accuracy of decision tree induction.

(7x4)

2.

a) Differentiate between (give examples for each case):
   i) Data and knowledge
   ii) Supervised learning and unsupervised learning
   iii) Data characterization and Data Discrimination
b) Suppose that the data for analysis includes the attribute age. The age values 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, 36, 40, 45, 46, 52, 70.
   i) What is the mean and median of data?
   ii) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
   iii) What is the midrange of the data?
c) Prove that accuracy is a function of sensitivity and specificity,

(8+6+4)

3.

a) Mention the conditions under which density-based clustering is more suitable than i) partitioning-based clustering ii) hierarchical clustering. Give some sample data sets to support your argument for both.
b) Is it true to say that Data Mining and Knowledge Discovery in Databases(KDD) are same? Justify.
c) Discuss issues to consider during data integration.

(8+6+4)

4.

a) Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date and count is number of spectator. Spectators may be students, adults, or seniors, with each category having its own charge rate.
   i) Draw a star schema diagram for the data warehouse.
   ii) Starting with the base cuboid [date,spectator,location,game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM Place in 2004?
   iii) If each dimension has five levels (including all), such as "date < spectator < location < game < all", how many cuboids will this cube contain (including the base and apex cuboids)?
b) What is Web Mining? Elaborate the various techniques used for web mining.

(12+6)

5.

a) How Crossover and Mutation is performed in Genetic Algorithm? Explain with example.

b) Consider the training examples shown in Table below for a binary classification problem.

Sr. No	A1	A2	A3	Target Class
1	T	T	1	P
2	T	T	6	P
3	T	F	5	N
4	F	F	4	P
5	F	T	7	N
6	F	T	3	N
7	F	F	8	N
8	T	F	7	P
9	F	T	5	N

i) What is the entropy of this collection of training examples with respect to the positive class?

ii) What are the information gains of a1 and a2 relative to these training examples?

(8+10)

6.

a) A database has five transactions. Let min support = 60% and min confidence = 80%.
{M,O,N,K,E,Y},{D,O,N,K,E,Y}, {M,AK,E}, {M,U,C,K,Y}, {C,O,O,K,I,E}

i) Find all frequent itemsets using FP growth algorithm.

ii) List all of the strong association rules (with support s and confidence c) matching the following Meta rule, where X is a variable representing customers and item i denotes variables representing items (e.g., "A","B", etc.):
$\forall x \in$ transaction, buys(X, item1) $\wedge$ buys(X, item2) $\Rightarrow$ buys(X,item3) [s,c]

b) What is Back-Propagations? Which type of NN is used by Back-Propagation algorithm for learning? Explain.

(10+8)

7.

a) Recent applications pay special attention to spatiotemporal data streams. A spatiotemporal data stream contains spatial information that changes over time, and is in the form of stream data, i.e., the data flow in-and-out like possibly infinite streams.

i) Present three application examples of spatiotemporal data streams.

ii) Discuss what kind of interesting knowledge can be mined from such data streams, with limited time and resources.

iii) Identify and discuss the major challenges in spatiotemporal data mining.

b) Compare K-means clustering algorithm with K-medoids algorithm.

c) Discuss the importance of any two data mining packages?

(9+5+4)