BE6-R4: DATA WAREHOUSE 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) Briefly explain process of Knowledge Discovery from data.

b) Write down basic difference in single and complete linkage of hierarchical clustering. Show difference between two with an example.

c) Define outlier. Does K-Means clustering algorithm detect outlier or not? Explain briefly.

d) What is the main drawback of using Information Gain? How is it overcome in Gain Ratio?

e) What is Bagging? How does it improve performance?

f) Name tree structure used to represent process of hierarchical clustering. Explain it by giving one example.

g) What are major challenges of mining a huge amount of data compared to data with hundreds of records only?

(7x4)

2. Differentiate between (give examples for each case):

a) i) Data and knowledge

ii) Supervised learning and unsupervised learning

iii) Data characterization and Data Discrimination

b) In which of mining techniques attribute selection method is used? How are discrete-valued and continuous valued attributes handled in attribute selection method? Explain by giving example for each case.

([3x4]+[2+2+2])

3.

a) How is a data warehouse differ from an operational database? Briefly explain star-schema and snowflake-schema used for data modeling in a data warehouse.

b) Draw 3-tier architecture of a data warehouse and briefly explain each stage.

(10+8)

4.

a) Suppose data for analysis include attribute age whose 15 values in increasing order are 16,16, 20, 20,20,20,22,23,25,25,25,25,25,30,33. Compute the following:

i) Mean and median of the data.

ii) What is mode of data? Comment on data modality.

b) What is Naive Bayes Classifies? What is the weakness of the assumption in the method?

c) Which type of clusters are generated by K-means and why? What is the pre-requisite for using K-means for clustering?

(6+8+4)

5.

a) Define precision and accuracy of a classifier. For the given confusion matrix, compute precision and sensitivity of a classifier.

		Predicted class		
	Classes	**Yes**	**No**	**Total**
Actual class	**Yes**	190	10	200
	No	210	90	300
	Total	400	100	-

b) Using the following transactional data, compute all frequent 2-itemsets using Apriori (minimum support is 50%).

TID	I1	I2	I3	I4	I5
T_1	1	1	1	0	0
T_2	1	1	1	1	1
T_3	1	0	1	1	0
T_4	1	0	1	1	1
T_5	1	1	1	1	0

c) Write any two ways to improve efficiency of Apriori algorithm.

(8+6+4)

6.

a) What is the need of data normalization? Briefly explain min-max normalization and z-score normalization. Use min-max normalization to transform the value 35 for age onto range [0.0,1.0] where min-max values of age are [20,80].

b) Differentiate between nominal and ordinal variables. Give one example for each. Also write down the mechanism used for computing dissimilarity between two objects having nominal attributes.

(10+8)

7.

a) Explain the following terms:
   i) Spatial data mining
   ii) Temporal Databases
   iii) DBSCAN

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

([3x4]+6)