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) Comment and Discuss: "*Snowflake schema is a variant of star schema*".
b) Explain differences between Rollup and Drill down with an example.
c) Describe various methods for handling missing values in data sets?
d) How is Web Content Mining different from Web Usage Mining?
e) Discuss issues to consider during data integration.
f) What is the importance of correlation Analysis? Give an example to show that "Correlation does not imply causality".
g) What is an outlier? Explain outlier mining and its applications.

(7x4)

2.

a) What is a data cube? Name three categories of measures based on the kind of aggregate functions used in computing data cube.
b) How multilevel association rules can be mined efficiently using concept hierarchy.
c) Write an algorithm for K-Nearest neighbor classification.

(8+5+5)

3.

a) 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 of the data? What is the median?
   ii) What is the mode of the data? Comment on the data's modality.
   iii) What is the midrange of the data?
   iv) What is the first quartile (Q1) and the third quartile (Q3) of the data?
   v) Give the five-number summary of the data.
b) Suppose that the data for analysis includes the attribute age. The age values are: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70
   i) Plot an equi-width histogram of width 10.
   ii) Use smoothing by bin method to smooth above data, using bin depth of 3.
   iii) Use min-max normalization to transform the value 35 for age on to the range [0.0, 1.0].
c) Explain the various basic heuristics methods for attribute subset selection.

(6+6+6)

4.

a) Suppose that a data warehouse for Big University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination.
   i) Draw a snowflake schema diagram for the data warehouse.
   ii) Starting with the base cuboid [student; course; semester; instructor], what specific OLAP operations (e.g., roll-up from semester to year) are to be performed in order to list the average grade of CS courses for each Big University student.
   iii) If each dimension has five levels (including all), such as "student < major < status < university < all", how many cuboids will this cube contain (including the base and apex cuboids)?
b) Explain with an example how information gain is used to construct a decision tree?

(10+8)

5.

a) What is Apriori Algorithm? With the help of it generate the frequent item sets for the following transaction table. Take minimum support count=2.

TID	Items
T1	{1,3,4}
T2	{2,3,5}
T3	{1,2,3,5}
T4	{2,5}
T5	{2,3,5}

b) Explain the following terms with respect to classification activity of data mining:
   i) Supervised learning
   ii) Unsupervised Learning
   iii) Training Set and testing Set
   iv) Classification accuracy

(10+8)

6.

a) The support and confidence measures are insufficient at filtering out uninteresting association rules. Explain with an example how correlation can be used to overcome weakness of the support and confidence?

b) Explain Attribute Oriented Induction (AOI).

c) How rule based classifier technique is different from tree based classification?

(8+6+4)

7.

a) What are the requirements of clustering in data mining? What are the types of data that often occur in cluster analysis? How data is pre-processed in cluster analysis?

b) How one can choose an efficient training dataset in case of Artificial Neural Networks?

c) Explain the three tier architecture of data warehouse.

(6+6+6)