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) Is Data Mining a simple transformation of technology developed from databases, statistics, and machine learning? Justify your answer.

b) How is a data warehouse different from a database? How are they similar?

c) Using given confusion matrix, compute precision and sensitivity of a classifier:

Classes	True	False	Total
True	30	120	150
False	150	200	350
Total	180	320	

d) Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?

e) It is difficult to assess classification accuracy when individual data objects may belong to more than one class at a time. In such cases, what criteria would be used to compare different classifiers modelled using same data.

f) Using K Mean Clustering algorithm, divide following points (with (x, y) representing location) into three clusters, where 8 points are P1(2,10), P2(2,5), P3(8,4), P4(5,8), P5(7,5), P6(6,4), P7(1,2), P8(4,9). [Assume P1, P4, P7 as they initial cluster center for three clusters.]

g) What kinds of associations can be mined in multimedia data?

(7x4)

2.

a) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

b) Many companies in industry prefer the update-driven approach (which constructs and uses data warehouses), rather than the query-driven approach (which applies wrappers and integrators), for integrating multiple heterogeneous information sources. Give reasoning for the same. Describe situations where the query-driven approach is preferable over the update-driven approach.

(9x9)

3.

a) Consider following 6 Transactions. Let min support is 50% and confidence is 60%.

T1	Milk, Biscuits, Eggs
T2	Milk, Eggs, Bread, Butter
T3	Butter, Bread, Ice-Cream
T4	Ice-Cream, Milk, Eggs
T5	Milk, Bread, Butter
T6	Bread, Butter

List all frequent itemsets using Apriori's Algorithm.

b) Differentiate between associative classification and discriminative frequent pattern-based classification. Why are classification based on frequent patterns able to achieve higher classification accuracy in many cases than a classical decision-tree method?

c) How does 3-tier architecture of data warehouse different from 2-tier architecture? Write down their advantages of using 3-tier architecture of data warehouse.

(6+6+6)

4.

a) Write down any one real-life example where data mining is crucial to the success of a business. What data mining functionalities does this business need (e.g., think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis?

b) Match the following:

	Mining Type		**Problem Statement**
1.	Web Mining	**A.**	To identify patterns from natural language text
2.	Text Mining	**B.**	To examine gene and protein sequences
3.	Sequence Mining	**C.**	Prediction of weather
4.	Time-Series Analysis	**D.**	To find URL's that are accessed frequently

Give reasoning for each match.

(8+10)

5.

a) What is the outlier? What is need of identifying outliers? Which clustering also can identify outlier?

b) Briefly describe each of the following approaches used for clustering:
   i) Partioning-based
   ii) Hierarchical-based
   iii) Density-based

Name one clustering algorithm for each of the approach.

(9+9)

6.

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 should one perform 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?

b) Compare the advantages and disadvantages of eager classification (e.g., decision tree, Bayesian, neural network) versus lazy classification (e.g., k-nearest neighbour)

(10+8)

7.

a) Define OLAP. How does OLAP help in improving performance for queries retrieval from data warehouse? Write any three fundamental characteristics of OLAP.

b) Differentiate between Entropy and Gini index with reference to classification.

c) Briefly describe feed formal neural network. How are they used in Artificial Neural Network?

(7+5+6)