CE3-R3: DATA WAREHOUSING AND MINING

NOTE:

1. **Answer question 1 and any FOUR questions from 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) List and explain four salient differences between OLAP and OLTP systems.

b) Construct a data cube from Table given below. Is this a dense or sparse data cube? If it is sparse, identify the empty cells.

Product ID	Location ID	Number Sold
1	1	29
1	3	8
2	1	5
2	2	23

c) Differentiate between descriptive and predictive data mining tasks. Give two examples for each.

d) How many possible association rules can be constructed from a data set with 4 items?

e) What is meant by hierarchical clustering? Explain, how agglomerative method is different from divisive clustering method.

f) What is meant by concept hierarchy? Explain with the help of a suitable example.

g) Explain the difference between web usage mining and web content mining with the help of a suitable example.

(7x4)

2.

a) Why is tree pruning useful in decision tree induction? Describe two commonly used strategies for tree pruning.

b) Name any three advantages of the Star schema? List one disadvantage of star schema?

c) Explain Iceberg queries with the help of a suitable example.

(6+6+6)

3.

a) Find frequent item sets from the following database using Aprori Algorithm. Use ms=0.5 and mc=0.75 and show the candidate and frequent item set at each level.

bid	transaction
01	AB DE
02	B D F
03	A B C
04	C D
05	C E
06	A C E
07	A B D E
08	A B D
09	A B C D
10	A B C E

b) What are Bayesian belief networks?
c) Give one application each of clustering and association rule mining in text data.
d) Give mathematical formulation of classification problem.

(8+2+4+4)

4.

a) Write the algorithm for K-means clustering. What is the difference between centroid and medoid of a cluster?
b) Describe the characteristics, architecture and issues associated with ROLAP tool.
c) Give formula and explain the following attribute selection measures:
   i) Information gain
   ii) Gain Ratio
   iii) Gini Index

(6+6+6)

5.

a) What is meant by Multi level association rule? Discuss any two approaches for mining multi level association rules with examples.
b) State the different topological relationships between two spatial objects.
c) How is the distance computed between:
   i) interval scaled data
   ii) binary data
   iii) categorical data

(6+6+6)

6.

a) List and explain four OLAP operators.
b) What do you understand by the classification accuracy of a classifier? How is it computed when the classes are mutually exclusive?
c) Define an outlier? Explain any one technique for eliminating outliers with an example.

(6+6+6)

7.

a) Discuss 3- tier architecture of data warehouse with a neat diagram. Explain in detail the functionality of each component.
b) A data cube C, has n dimensions, and each dimension has exactly p distinct values in the base cuboid. Assume that there are no concept hierarchies associated with the dimensions.
   i) What are the maximum and minimum numbers of cells possible in the base cuboid?
   ii) What are the maximum and minimum numbers of cells possible in data cube C?
c) With the help of a neat diagram, explain the process of knowledge discovery. Write a brief note for each of the steps.

(6+6+6)