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) Describe, how box plot can give information about whether the value of an attribute is symmetrically distributed.

b) What is meant by concept hierarchy? Discuss various types of hierarchies by providing one example for each type.

c) Give an example to show that items in a strong association rule may actually be negatively correlated.

d) State differences between predictive modeling and descriptive modeling.

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

f) Explain from salient differences between OLAP and OLTP systems.

g) Regarding the coupling of data mining system with a data base or data warehouse system. Distinguish between no coupling, loose coupling, semi tight coupling and tight coupling.

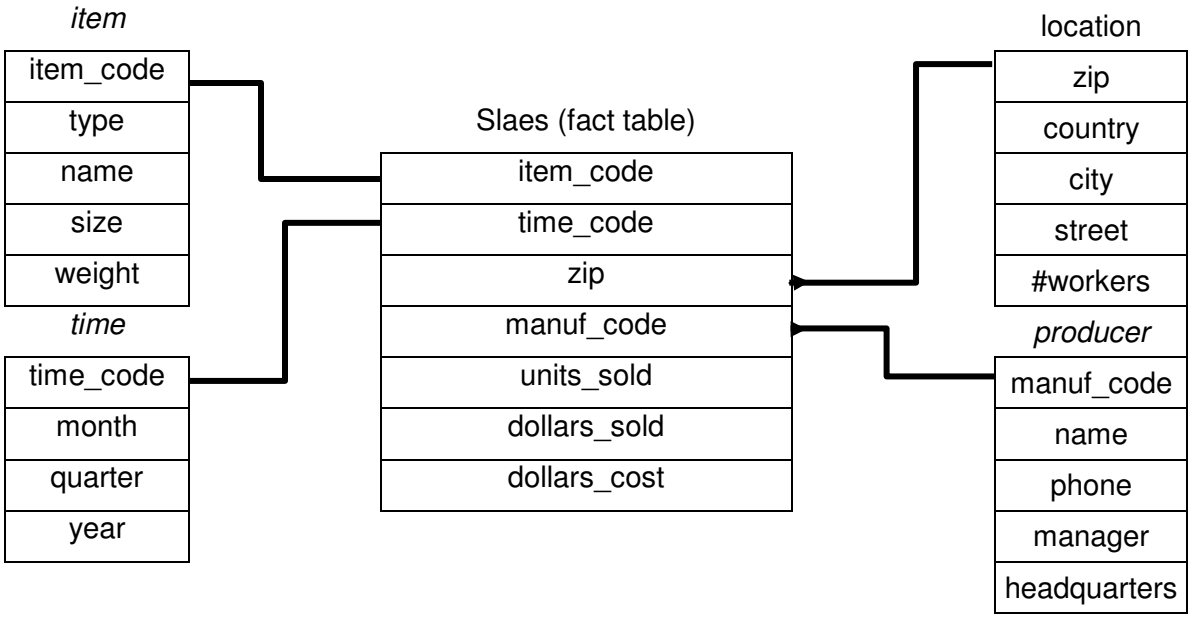
(7x4)

2.

a) Why is tree pruning required in decision tree induction? What are the two approaches for tree pruning?

b) How does snow flake schema overcome the disadvantages of star schema?

c) Given the following star schema, convert it into snowflake schema.



d) Explain three preprocessing steps required before feeding data into a data warehouse.

(6+2+4+6)

3.

a) An analyst wants to use association rule mining approach to analyze test results. The test consists of 100 questions with ten possible answers each. Each question can have more than one correct answer.
   i) Convert this data into a form suitable for association analysis?
   ii) What type of attributes are there in data? How many of them are there?

b) What are the characteristics of Bayesian Belief networks?

c) List and explain three techniques used for supervised learning.

(6+6+6)

4.

a) Describe the K-means clustering algorithm.

b) Discuss the Hyperlink induced topic search (HITS) algorithm.

c) Cluster the following data into three clusters using agglomerative clustering method.
Data: { 2, 4, 10, 12, 3}.

(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) What are the four properties that a distance function must satisfy? If $O_1=\{a_{11}, a_{12}, a_{13}, a_{14}\}$ and $O_2=\{a_{21}, a_{22}, a_{23}, a_{24}\}$ are two objects, how are Euclidian and Manhattan distances computed between O_1 and O_2.

c) State three topological relationships between two spatial objects each with the help of an example.

(**6+6+6)**

6.

a) List four salient distinctions between data warehouse and operational database.

b) Discuss the architecture of data warehouse with a neat diagram. Explain in detail the functionality of each component.

c) Suppose that the data for analysis includes attribute Age. The age values for the data tuples are 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) Compute the five number summary for the above data.
   ii) Draw a box-plot of the data.

(6+6+6)

7.

a) Discuss the features of Data Mining query language with examples.

b) State the difference between web content mining and web structure mining.

c) Write down the algorithm for generating association rules from the set $\mathcal{L}$, (. set of frequent/large itemsets). Use mc as the minimum confidence level. Support of an itemset $x \in \mathcal{L}$, is given by Sup(x).

(6+6+6)