

9. (a) What is data frame ? How do we load data frame ?  
Explore the R functions for data in data frame.
- (b) What are issues in decision tree learning in R ?  
Write R code for generating decision rules and  
plotting a decision tree. **8+7=15**

Roll No. ....

Total Pages : 04

**MCAQ/D-23**

**24027**

**DATA MINING AND INTEGRATION  
USING R  
MCA-20-32**

Time : Three Hours]

[Maximum Marks : 75

**Note :** Attempt *Five* questions in all, selecting *one* question each from Unit I to Unit IV. Q. No. **1** is compulsory. All questions carry equal marks.

**Compulsory Question**

1. (a) Write the important characteristics of data warehouse.
- (b) What are the different types of outlier ?
- (c) What is tree pruning ? Discuss its types.
- (d) Define neurode. How do you compute weight in a neural network ?
- (e) What is Global-as view and Local-as-view ?
- (f) Write a note on Phonetic Similarity Measures.
- (g) How do we convert data types in R ?
- (h) What is vector ? How do we declare vectors in R ?

**15**

### Unit I

2. (a) Outline and explain each and every component of data warehouse architecture.  
(b) What are the motivations behind data mining ? Discuss knowledge discovery life-cycle. **8+7=15**
3. (a) Briefly explain data mining functionalities. Why is it important to understand ?  
(b) Discuss the strategy for data cleaning and data transformation. **8+7=15**

### Unit II

4. (a) What are the different types of hierarchical clustering ? Frame the cluster for data points A(1, 1), B(2, 3), C(3, 5), D(4, 5), E(6, 6), and F(7, 5) using hierarchical clustering. Design dendrogram.  
(b) Which distance function is more appropriate for implementing decision tree ? How information gain is computed ? **8+7=15**
5. (a) Use the Apriori algorithm to find association rule for the following dataset :

Transaction ID	Items
T1	11,12,13
T2	12,13,15,16
T3	11,12,13,15

T4

12,15

T5

11,13,15

When minimum support count is of 2 and minimum confidence is of 75 per cent.

- (b) What is nearest neighborhood ? Write the challenges for implementing nearest neighborhood. **8+7=15**

### Unit III

6. (a) Discuss discrepancies between the source schemata and the mediated schema needs to be handle by schema mapping.  
(b) Illustrate Needleman-Wunch Measure for sequence based similarity measure. **8+7=15**
7. (a) Elaborate the challenges of schema matching and mapping. Explore the components of a typical schema matching system.  
(b) Write a detailed note on learning based matching. **8+7=15**

### Unit IV

8. (a) Discuss the advantages of R over other programming languages. How R deal with directories ?  
(b) State an illustration in R, showing the use of control statements and loops in R. **8+7=15**