Q.P. Code : 595002

(3 Hours) [Total Marks : 80

N.B. : (1) **Q1** is **compulsory**.
     (2) Attempt any **three** from the remaining.
     (3) Assume suitable data.

1. (a) Define "Data Mining". Enumerate five example applications that can benefit by using Data Mining.   **5**

   (b) What is Data Preprocessing? Explain the different methods for the Data Cleansing phase.   **5**

   (c) What is hierarchical clustering? Explain any two techniques for finding distance between the clusters in hierarchical clustering.   **5**

   (d) Explain the concept of a decision support system with the help of an example application.   **5**

2. (a) Partition the given data into 4 bins using Equi-depth binning method and perform smoothing according to the following methods.   **10**
Smoothing by bin mean
Smoothing by bin median
Smoothing by bin boundaries

    Data: 11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45, 45, 71, 72, 73, 75

   (b) For the same set of data points in question 2. (a)   **10**
     (a) Find Mean, Median and Mode.

     (b) Show a boxplot of the data. Clearly indicating the five- number summary.

[TURN OVER

3. (a) The table below shows a sample dataset of whether a customer **10**
responds to a survey or not. **"Outcome"** is the class label.
Construct a Decision Tree Classifier for the dataset. For a new example
(Rural, semidetached, low, No), what will be the predicted class label?

District	House Type	Income	Previous Customer	Outcome
Suburban	Detached	High	No	Nothing
Suburban	Detached	High	Yes	Nothing
Rural	Detached	High	No	Responded
Urban	Semi-detached	High	No	Responded
Urban	Semi-detached	Low	No	Responded
Urban	Semi-detached	Low	Yes	Nothing
Rural	Semi-detached	Low	Yes	Responded
Suburban	Terrace	High	No	Nothing
Suburban	Semi-detached	Low	No	Responded
Urban	Terrace	Low	No	Responded
Suburban	Terrace	Low	Yes	Responded
Rural	Terrace	High	Yes	Responded
Rural	Detached	Low	No	Responded
Urban	Terrace	High	Yes	Nothing

(b) Briefly explain Bagging and Boosting of Classifiers **10**

4. (a) Use the Apriori algorithm to identify the frequent item-sets in the following database. Then extract the strong association rules from these sets.   **10**

Min. Support = 30% Min. Confidence = 75%

TID	Items
01	A, B, D, E, F
02	B, C, E
04	A, B, D, E
04	A, B, C, E,
05	A, B, C, D, E, F
06	B, C, D
07	A, B, D, E

(b) Explain multidimensional and multi level Association rules with examples.   **10**

5. (a) Use any hierarchical clustering algorithm to cluster the following 8 examples into 3 clusters :   **10**

A1= (2, 10),   A2= (2, 5),   A3= (8,4),   A4 = (5, 8),
A5= (7, 5),   A6= (6, 4),   A7= (1,2),   A8= (4,9)

(b) What is an outlier? Describe methods that can be used for outlier analysis.   **10**

6. (a) Consider the following case study: An International chain of hotels wants to analyze and improve its performance using several performance indicators -quality of room, service facilities, check in, breakfast, popular time of visits, duration of stay etc.   **10**
For this case study design a BI system, clearly explaining all steps from data collection to decision making.

(b) Clearly explain the working of the DB_SCAN algorithm using appropriate diagrams.   **10**