SET- I

1 Consider the group of 12 sales price records that has been sorted as follows: 5, 10, 11, 13,15,35,50,55,72,92,204, and 215. Partition them into three bins by each of the following methods.
(a) equal-frequency (equi-depth) partitioning
(b) equal-width partitioning
(c) Clustering.
Implement the same using R.

2 A gadget factory has been quite successful for the past 10 years and Ms.Marry, the manager of the company wondering whether to expand the factory this year or not. The cost to expand factory is $2M. With no expansion, expected revenue is $4M if the economy stays good; while only $1.5M if the economy is bad. If manager expands the factory, expected to receive $7M. if economy is good and $3M if economy is bad. Assume that there is a 45% chance of a good economy and a 55% chance of a bad economy. Draw a Decision Tree showing these choices.

3 Apply Apriori Algorithm for given database below by assuming Minimum support = 2. Implement using WEKA for the given data.

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

4 Use following group of data: 200,300,400,600,1000
(a) min-max normalization by setting min = 0 and max = 1
(b) z-score normalization using the mean absolute deviation instead of standard deviation
(c) normalization by decimal scaling

SET 2

1 Consider a group of people who are affected by blood pressure based on the diabetes dataset. Display it using scatterplot and bar chart (that is Blood Pressure vs Age employing dataset "diabetes.csv") using R.

2 Analyze the dataset "diabetes.csv" how the diabetes trend is for different age people, using Linear Regression and Multiple Regression.

3 Suppose a database has five transactions. Let minimum support = 50% (2) and minimum confidence = 80%.
Transactions Items

T1	(M, O, N, K, E, Y)
T2	(D, O, N, K, E, Y)
T3	(M, A, K, E)
T4	(M, U, C, K, Y)
T5	(C, O, O, K, I, E)

i) Implement using WEKA and find all frequent item sets using Apriori algorithm
ii) Also draw FP-Growth Tree

4 Prediction of Categorical Data using Decision Tree Algorithm through WEKA using any datasets. a) Tree b) Preprocess c) Logistic

SET 3

1. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 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. Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
2. Download the Dataset "water" From R dataset Link. Find out whether there is a linear relation between attributes "mortality" and "hardness" by plot function. Fit the Data into the Linear Regression model. Predict the mortality for the hardness = 88.
3. Create the dataset using ARFF file format:

Transaction ID	Items
T1	Hot Dogs, Buns, Ketchup
T2	Hot Dogs, Buns
T3	Hot Dogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	Hot Dogs, Coke, Chips

a. Find the frequent item-sets and generate association rules on this. Assume that minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%).
b. List the various rule generated by apriori and FP tree algorithm, mention whether it is accepted or rejected.
4. Prediction of Categorical Data using Rule base classification and decision tree classification through WEKA using any datasets. Compare the accuracy using two algorithm and plot the graph

SET 4

1 Imagine that you have selected data from the All-Electronics data warehouse for analysis. The data set will be huge! The following data are a list of All Electronics prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted: 1,1,5,5,5,5,5,8,8,10, 10,10,10,12,14,14,14,15,15,15,15,15,15,18,18,18,18,18, 18,18,18,20,20,20,20,20,20,20,21,21,21,21,25,25,25,25,25,28,28,30, 30,30 .
(i) Partition the dataset using an equal-frequency partitioning method with bin equal to 3
(ii) Apply data smoothing using bin means and bin boundary.
(iii) Plot Histogram for the above frequency division

2 Two Maths teachers are comparing how their Year 9 classes performed in the end of year exams. Their results are as follows:
Class A: 76,35,47,64,95,66,89,36,8476,35,47,64,95,66,89,36,84
Class B: 51,56,84,60,59,70,63,66,5051,56,84,60,59,70,63,66,50
(i) Find which class had scored higher mean, median and range.
(ii) Plot above in boxplot and give the inferences

3 Consider a Binary classification model that can be used to predict whether one or more ads on the website will be clicked or not. The models are used to optimize the ad inventory on websites by selecting which ads will have a better chance of being clicked.

4 Consider that Many businesses use cluster analysis to identify consumers who are similar to each other so they can tailor their emails sent to consumers in such a way that maximizes their revenue. Consider a business may collect the following information about consumers:
Percentage of emails opened

i) Number of clicks per email

ii) Time spent viewing email
Using these metrics, a business can perform various cluster analyses to identify consumers who use email in similar ways and tailor the types of emails and frequency of emails they send to different clusters of customers. Compare the performance of the applied clustering algorithm.