SEMESTER-III

COURSE 7: DATA MINING TECHNIQUES USING R

Theory Credits: 3 3 hrs/week

Aim and objectives of Course:

- To understand Data mining techniques and algorithms.
- Comprehend the data mining environments and application.

Learning outcomes of Course:

Students who complete this course will be able to

- Compare various conceptions of data mining as evidenced in both research andapplication.
- Evaluate mathematical methods underlying the effective application of data mining.
- Should be able to apply the type of techniques based on the problems considered.
- Can find out the market patterns and association amongst different products.

UNIT I:

An idea on Data Warehouse, Data mining-KDD versus data mining, Stages of the Data MiningProcess-Task primitives., Data Mining Techniques – Data mining knowledge representation.

UNIT II

Data mining query languages- Integration of Data Mining System with a Data Warehouse-Issues, Data pre-processing – Data Cleaning, Data transformation – Feature selection – Dimensionality reduction

UNIT III

Concept Description: Characterization and comparison What is Concept Description,Data Generalization by Attribute-Oriented Induction(AOI), AOI for Data Characterization, Efficient Implementation of AOI.

Mining Frequent Patterns, Associations and Correlations: Basic Concepts, FrequentItemset Mining Methods: Apriori method, generating Association Rules, Improvingthe Efficiency of Apriori,Pattern-Growth Approach for mining Frequent Item sets.

UNIT-IV

Classification Basic Concepts: Basic Concepts, Decision Tree Induction: Decision TreeInduction Algorithm, Attribute Selection Measures, Tree Pruning. Bayes Classification Methods.

UNIT-V

Association rule mining: Antecedent, consequent , muti-relational association rules,

ECLAT.Case study on Market Basket Analysis.

Cluster Analysis: Cluster Analysis, Partitioning Methods, Hierarchal methods, Density basedmethods-DBSCAN.

TEXT BOOKS:

1. Jiawei Han, MichelineKamber, Jian Pei."Data Mining: Concepts and Techniques", 3rd Edition,Morgan Kaufmann Publishers, 2011.
2. AdelchiAzzalini, Bruno Scapa, "Data Analysis and Data mining" , 2ndEdiiton, Oxford Univeristy Press Inc., 2012.
3. Data Mining, The Textbook (2015) by Charu Aggarwal.

REFERENCES BOOKS:

1. Alex Berson and Stephen J. Smith, "Data Warehousing, Data Mining & OLAP", 10th Edition, TataMcGraw Hill Edition , 2007.
2. G.K. Gupta, "Introduction to Data Mining with Case Studies", 1st Edition, EasterEconomy Edition, PHI, 2006.

Student Activities:

1. Students should be able to implement Data Mining algorithms provided the relevantdata
2. Given the data, students can visualize all statistical measures
3. Differentiate the types of mining problems and identify what type of algorithms are tobe implemented.

Continuous assessment:

Let the students be tested in the following questions from each unit

1. What is Data Mining and KDD? Where Data Mining fits in KDD Process
2. Describe all Preprocessing methods
3. Explain Data Description and AOI Algorithm
4. Explain Classification and Write any Decision tree induction algorithm
5. Explain the concept of clustering and write any algorithm to form clusters.

SEMESTER-III

COURSE 7: DATA MINING TECHNIQUES USING R

Practical | Credits: 1 | 3 hrs/week

1. Get and Clean data using dplyr exercises.
2. Visualize all Statistical measures(Mean ,Mode, Median, Range, InterQuartile Range etc.,using Histograms, Boxplots and Scatter Plots).
3. Create a data frame with atleast 10 entries of columns EMPID,EMPNAME,SALARY,STARTDATE

   a. Extract two column names using column name.
   b. Extract the first two rows and then all columns.
   c. Extract 3^{rd} and 5^{th} row with 2^{nd} and 4^{th} column.

4. Create a data frame with 10 observations and 3 variables and add new rows and columns to itusing 'rbind' and 'cbind' function.
5. Create a function to discretize a numeric variable into 3 quantiles and label them as low, medium, and high. Apply it on each attribute of any dataset to create a new data frame. 'discrete' with Categorical variables and the class label.
6. Create a simple scatter plot using any dataset using 'dplyr' library. Use the samedata to indicate distribution densities using box whiskers.
7. Write R Programs to implement k-means clustering, k-medoids clustering and density based clustering on any datasets.
8. Write a R Program to implement decision trees using 'reading Skills' dataset.
9. Implement decision trees using any dataset using package party and 'rpart'.
10. Generate top 5 association rules using apriori.
11. Generate top 5 association rules using ECLAT.
12. Write an R program to implement Naïve bayes Classification.