FACULTY OF COMPUTER SCIENCE

Master of Computer Application (Sem-IV)

In Effect from Academic Year 2017-18

Subject Code : 1CS2010406	Subject Title: DATA ANALYTICS WITH R
Pre-requisite :	Any programming language

Course Objective:

The objectives of the course are to:

- Learn R Programming language,
- Learn data analytics, data visualization and statistical model for data analytics.

Teaching Scheme (Hours per week)				Evaluation Scheme (Marks)				
				Theory		Practical		
Lecture	Tutorial	Practical	Credit	University Assessment	Continuous Assessment	University Assessment	Continuous Assessment	Total
4	-	3	7	60	40	30	20	150

Subject			
Sr. No	Topic	Total Hours	Weight (%)
1	**Introduction to Data Analysis** Overview of Data Analytics, Need of Data Analytics, Nature of Data, Classification of Data: Structured, Semi-Structured, Unstructured, Characteristics of Data, Applications of Data Analytics.	06	15
2	**R Programming Basics** Overview of R programming, Environment setup with R Studio, R Commands, Variables and Data Types, Control Structures, Array, Matrix, Vectors, Factors, Functions, R packages.	12	25
3	**Data Visualization using R** **Reading and getting data into R (External Data):** Using CSV files, XML files, Web Data, JSON files, Databases, Excel files. **Working with R Charts and Graphs:** Histograms, Box plots, Bar Charts, Line Graphs, Scatter plots, Pie Charts	13	25
4	**Statistics with R** Random Forest, Decision Tree, Normal and Binomial distributions, Time Series Analysis, Linear and Multiple Regression, Logistic Regression, Survival Analysis	12	25
5	**Prescriptive Analytics** Creating data for analytics through designed experiments, Creating data for analytics though active learning, Creating data for analytics through reinforcement learning	05	10

Course Outcome:

At the end of this course, the student would be able to

- Apply statistical techniques using R Programming for data analytics and decision making.
- Become data analyst.

List of References:

1. An Introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics. W. N. Venables, D.M. Smith and the R Development Core Team. Version 3.0.1 (2013-05-16). URL: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
2. Jared P Lander, R for everyone: advanced analytics and graphics, Pearson Education, 2013
3. Dunlop, Dorothy D., and Ajit C. Tamhane. Statistics and data analysis: from elementary to intermediate. Prentice Hall, 2000.
4. G Casella and R.L. Berger, Statistical Inference, Thomson Learning 2002.
5. P. Dalgaard. Introductory Statistics with R, 2nd Edition. (Springer 2008)
6. Michael Berthold, David J. Hand, Intelligent Data Analysis, Springer
7. Hastie, Trevor, et al.The elements of statistical learning.Vol. 2. No. 1. New York: springer, 2009.
8. Montgomery, Douglas C., and George C. Runger.Appliedstatistics and probability for engineers. John Wiley &Sons, 2010

Indicative Practical List

1. Check the output of the following commands:
   help(), c(), length(), ls(), rm(), sum(), mean(), median(), var(), names(), data(), sqrt(), sd(), seq()
2. Write R script to perform arithmetic and logical operations.
3. Write R script to create 3 x 3 matrix to perform addition, subtraction, multiplication and division operations.
4. Write R script to create histogram and scatter plot for vector of x and y. Each vector contained 20 randomly selected elements from range between 0 to 9.
5. Create employee.csv file (emp_name, organization, mobile_no, email, salary, experience, and city) which contained 20 records. Write R script to read data from employee.csv and display it into R workspace.
6. Write R script to create data frame "student" with the fields of stud_id, stud_name, email_id and mobile_no. Perform following operations:
   a. Display data of data frame
   b. Display summary of data frame c. Display structure of data frame
   d. Extract and display only stud_name and mobile_no from data frame
7. The data below are the number of hours spent in social networking applications per week for a sample of 52 thousands.

18.5	29.1	22.5	20.7	22.0	26.6	5.9	9.0	14.5
11.8	26.0	20.4	16.8	17.0	16.9	16.0	11.8	23.0
9.8	19.7	13.0	11.9	29.0	17.7	25.5	24.3	27.0
10.6	19.9	21.5	15.1	16.2	19.4	26.9	30.0	12.5

   Perform following operations on above data set:
   a. Obtain histogram and box plot
   b. Find mean, median and standard deviation
   c. Generate smooth curve through histogram using "hist() and lines()" commands.
8. Consider a data set of 100 cricketers with their name, age, handed (left/right), average batting score etc. categorize these cricketers in various clusters on the basis of:
   a. Age
   b. Handed
   c. Average batting score
9. Consider some sample data and develop following models for Analysis Of Variance (ANOVA):
   a. Linear regression
   b. Multiple regression
10. Consider sample data and develop following models:
    a. Decision tree
    b. Random forest
    11. Consider the monthly average gold price in India starting from January 2005 to December 2016. Create time series object for all months of given period and plot it.