SEMESTER-II

COURSE 3: INTRODUCTION TO DATA SCIENCE AND R PROGRAMMING

Theory Credits: 3 3 hrs/week

Aim and objectives of Course :

Data Science is a fast-growing interdisciplinary field, focusing on the analysis of data to extract knowledge and insight. This course will introduce students to the collection. Preparation, analysis, modelling and visualization of data, covering both conceptual and practical issues. Examples and casestudies from diverse fields will be presented, and hands- on use of statistical and data manipulation software will be included.

Learning outcomes of Course:

- Recognize the various discipline that contribute to a successful data science effort.
- Understand the processes of data science identifying the problem to be solved, datacollection, preparation, modeling, evaluation and visualization.
- Be aware of the challenges that arise in Data Sciences.
- Be able to identify the application of the type of algorithm based on the type of the problem.
- Be comfortable using commercial and open source tools such as the R/Python languageandits associated libraries for data analytics and Visualization.

UNIT I:
Defining Data Science and Big data, Benefits and Uses, facets of Data, Data Science Process. Historyand Overview of R, Getting Started with R, R Nuts and Bolts

UNIT II:
The Data Science Process: Overview of the Data Science Process-Setting the research goal, Retrieving Data, Data Preparation, Exploration, Modeling, data Presentation and Automation. GettingData in and out of R, Using reader package, Interfaces to the outside world.

UNIT III:
Machine Learning: Understanding why data scientists use machine learning-What is machine learning and why we should care about, Applications of machine learning in data science, Where itis used in data science, The modeling process, Types of Machine Learning-Supervised and Unsupervised.

UNIT IV:
Handling large Data on a Single Computer: The problems we face when handling large data, General Techniques for handling large volumes of data, Generating programming tips fordealing with large datasets.

UNIT V:

Sub setting R objects, Vectorised Operations, Managing Data Frames with the dplyr, Control structures, functions, Scoping rules of R, Coding Standards in R, Loop Functions, Debugging, Simulation. Case studies on preliminary data analysis.

TEXT BOOKS:

1. DavyCielen, Arno.D.B.Maysman, Mohamed Ali, “Introducing Data Science”ManningPublications, 2016.
2. Roger D. Peng, “R Programming for DataScience” Lean Publishing, 2015.

REFERENCE BOOKS:

1. Nina Zumel, John Mount, “Practical Data Science with R”, Manning Publications, 2014.
2. Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, AbhijitDasgupta, “PracticalData Science Cookbook”, Packt Publishing Ltd., 2014.

WebReferences for case studies:

1. https://www.kaggle.com/datasets

2. https://github.com/

SEMESTER-II

COURSE 3: INTRODUCTION TO DATA SCIENCE AND R PROGRAMMING

Practical Credits: 1 2 hrs/week

Lab/Practical/Experiments/Tutorials syllabus:

1. Installing R and R studio, with proper notes on version management, cosmetic settings and different
libraries.
2. Basic operations in r with arithmetic and statistics.
3. Getting data into R, Basic data manipulation, Loading Data into R
4. Basic plotting
5. Loops and functions
6. Create Vectors, Lists, Arrays, Matrices, Data frames and operations on them.
7. Demonstrate the visualization and graphics using visualization packages like ggplot2.
8. Implement Loop functions with lapply(), sapply(), tapply(), apply(), mapply().
9. Explore data using Single Variables: Unimodal, Bimodal, Histograms, Density Plots,Barcharts
10. Explore data using two Variables: Line plots, Scatter Plots, smoothing cures, Bar charts
11. Explore and implement commands using dplyr package
12. Download a dataset and work on basic data manipulation followed by inferential statistics.

RECOMMENDED TEXT BOOKS:

1. Mark Gardener, "Beginning R - The Statistical Programming Language", John Wiley &Sons, Inc., 2012.
2. W. N. Venables, D. M. Smith and the R Core Team, "An Introduction to R", 2013.
Recommended Reference books:
3. The art of R Programming: A tour of Statistical Software design. Norman Matloff.KindleEdition
4. The book of R : The first course in Programming and Statistics by Tilman M. Davies.

Recommended Co-curricular activities: (Co-curricular Activities should not promote copyingfromtext book or from others' work and shall encourage self/independent and grouplearning)

A. Measurable:
1. Assignments on:
2. Student seminars (Individual presentation of papers) on topics relating to:
3. Quiz Programmes on:
4. Individual Field Studies/projects:
5. Group discussion on:
6. Group/Team Projects on:

B. General

1. Collection of news reports and maintaining a record of paper-cuttings relatingtotopics covered in syllabus
2. Group Discussions on:
3. Watching TV discussions and preparing summary points recording personalobservations etc., under guidance from the Lecturers
4. Any similar activities with imaginative thinking.
5. Recommended Continuous Assessment methods: