**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/**

**MODEL QUESTION PAPER**

**INTRODUCTION TO DATA SCIENCE AND R PROGRAMMING**

**Time: Three Hours Maximum Marks: 60**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

PART - A

**Answer any Five of the following 5x4=20 marks**

1. Explain about R attributes?
2. What are the 4 V’s of Big Data?
3. Explain about using readr package
4. Explain about data cleaning
5. Explain about regularization.
6. What does "training a model" and "validating a model" mean
7. Differentiate between full batch learning, mini-batch learning, and online learning
8. Explain the concept of MapReduce.
9. Identify two key functions from the dplyr package
10. What are Vectorised Operations?

PART - B

**Answer all of the following 5 x 8 = 40 marks**

1. What are the different facets of data

OR

1. What are the different data structures in R? Explain
2. Explain about the data integration

OR

1. Explain about data transformation
2. Explain in detail about supervised machine learning

OR

1. Explain in detail about unsupervised machine learning
2. How choosing a right data structure improve efficiency when handling large datasets

OR

1. Explain perceptron in detail? How does it handle large data?
2. Explain in detail about control structures in R

OR

1. Explain about subsetting R objects

**QUESTION BANK**

**Unit-1** Defining Data Science and Big data, Benefits and Uses, facets of Data, Data Science Process.

History And Overview of R, Getting Started with R, R Nuts and Bolts

**short answer questions**

1. Which values are used to represent missing values and which functions are used to test for presence of missing values.
2. What are the 4 V’s of Big Data?
3. Explain about explicit coercion
4. Explain about naming objects in R

**Long answer questions**

1. What are the different facets of data
2. What are the different data structures in R? Explain
3. Explain about built in datatypes in R and R attributes

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 readr package, Interfaces to the outside world.

Unit-2

**short answer questions**

1. Explain about using readr package
2. Explain about interfaces to outside world
3. What is the importance of data preparation?
4. Explain about data cleaning

**Long answer questions**

1. Explain about the data science process
2. Explain in detail about data integration
3. Explain about exploratory data analysis
4. Explain in detail about data transformation

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 it is used in data science, The modeling process, Types of Machine Learning-Supervised and Unsupervised.

**Short answer questions**

1. What does "training a model" and "validating a model" mean
2. What is root cause analysis?
3. Explain about ensemble learning.
4. What is availability bias?
5. Explain about regularization.

**Long Answer Questions**

1. Explain in detail about supervised machine learning
2. Explain in detail about unsupervised machine learning
3. Explain Naive Bayes classification technique
4. Explain linear and multiple linear regression
5. i) Give applications of machine learning  
   ii) Explain about k-means clustering with an example
6. Explain about validation strategies with examples.

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 for dealing with large datasets.

**Short answer questions**

1. Differentiate between full batch learning, mini-batch learning, and online learning
2. Differentiate between online and Streaming Algorithms
3. Explain the concept of MapReduce

**Long Answer Questions**

1. Explain perceptron in detail? How does it handle large data?
2. Explain about common problems encountered with large amounts of data
3. How choosing a right data structure improve efficiency when handling 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.**

**Short answer questions**

1. Identify two key functions from the dplyr package
2. What are Vectorised Operations?
3. Explain briefly about debugging in R.

**Long Answer Questions**

1. Explain in detail about control structures in R with examples
2. What are Loop Functions? Explain with examples.
3. Explain about subsetting R objects