Final Year of Computer Engineering (2015 P Course)
Big Data & Data Analytics

Teaching Scheme: TH: 04 Hours/Week	**Credit**	**Examination Scheme:** In-Sem (Paper) : 30 Marks End-Sem (Paper) : 70 Marks

Prerequisite: Data Mining, Knowledge of probability theory, statistics, and programming

Course Objectives:

- To understand Data Analytics Life Cycle and Business Challenges
- To understand Analytical Techniques and Statically Models
- To understand Statically Modelling Language

Course Outcomes:

On completion of the course, student will be able to–

- Deploying the Data Analytics Lifecycle to address big data analytics projects
- Reframing a business challenge as an analytics challenge
- Applying appropriate analytic techniques and tools to analyze big data, create statistical models, and identify insights that can lead to actionable results
- Selecting appropriate data visualizations to clearly communicate analytic insights to business sponsors and analytic audiences
- Using tools such as: R and R Studio, MapReduce/Hadoop, in-database analytics,
- Explain how advanced analytics can be leveraged to create competitive advantage

Course Contents

Unit I	**Introduction to Big Data**	**06Hours**

Business Intelligence, Decision Support Systems, Data Warehousing; Definition of Big Data, Big data characteristics & considerations, Introduction to Hadoop

Unit II	**Big Data Analytics**	**06 Hours**

Big data analytics, Drivers of Big data analytics, Big Data Stack, Typical analytical architecture, Virtualization & Big Data, Virtualization Approaches, Business Intelligence Vs Data science, Applications of Big data analytics.

Unit III	**Data Analytics Lifecycle**	**06 Hours**

Need of Data analytic lifecycle, Key roles for successful analytic projects, **various phases of Data analytic lifecycle:** Discovery, Data Preparation, Model Planning, Model Building, Communicating Results, Operationalization.

Unit IV	**Machine Learning: Supervised Learning**	**08 Hours**

What is Machine Learning? Application of Machine Learning; Supervised Learning Structure of Regression Model, Linear Regression, Logistics Regression, Time series analysis, Support Vector Machine.

Unit V	**Classification &Unsupervised Learning**	**08 Hours**

Classification: Classification Problem, Classification Models, Classification Trees, Bayesian Method; Association Rule: Structure of Association Rule, Apriori Algorithm, General Association; Clustering: Clustering Methods, Partition Methods, Hierarchical Methods.

Unit VI	**Exploring Data in R**	**06 Hours**

Basic features of R, Exploring R GUI, Data Frames & Lists, Handling Data in R Workspace, Reading Data Sets & Exporting Data from R, Manipulating & Processing Data in R.

Book:

Text:

1. David Dietrich, Barry Hiller, “Data Science & Big Data Analytics”, EMC education services, Wiley publications, 2012
2. Trevor Hastie, Robert Tibshirani, Jerome Friedman, "The Elements of Statistical Learning", Springer, Second Edition, 2011.

Reference Books:

1. Business Intelligence – Data Mining and Optimization for Decision Making – Carlo Vercellis – Wiley Publications.
2. Big Data & Analytics – Seema Acharya & Subhashini Chellappan – Wiley Publications
3. Big Data (Black Book) – DT Editorial Services – Dreamtech Press.
4. Data Mining: Concepts and Techniques Second Edition – Jiawei Han and Micheline Kamber – Morgan KaufMan Publisher
5. Beginning R: The Statistical Programming Language – Mark Gardner – Wrox Publication

List of Experiments: 410255- Laboratory Practices -IV

1. Installation of R
2. Study of R: Declaring Variable, Expression, Function and Executing R script.
3. Creating List in R – merging two lists, adding matrices in lists, adding vectors in list.
4. Manipulating & Processing Data in R – merging data sets, sorting data, plotting data, managing data using matrices & data frames

Mini Project (Any One)

1. Twitter Data Analysis with R
2. Sentiment Analysis of WhatsApp data with R

UNIT I:
Introduction to Big Data

CONTENT

UNIT II:
Big Data Analytics

2.1 Big data analytics

2.2 Drivers of Big data analytics

2.3. Big Data Stack

2.4. Typical analytical architecture

2.5. Virtualization & Big Data

2.6. Virtualization Approaches

1. Big Data Server Virtualization:

2 Big Data Application Virtualization:

3 Big Data Network Virtualization:

4 Big Data and Storage Virtualization

5 Big Data processor and Memory Virtualization

2.7. Business Intelligence vs. Data science

2.8 Applications of Big data analytics.

UNIT III:
Data Analytics Lifecycle

3.1. Need of Data analytic lifecycle

3.2 Key roles for successful analytic projects

3.3. Various phases of Data analytic lifecycle

3.3.1 Discovery

3.3.2 Data Preparation

3.3.3 Model Planning

3.3.4 Model Building

3.3.5. Communicating Results

3.3.6 Operationalization.

UNIT IV
Machine Learning: Supervised Learning

4.1 What is Machine learning?

4.2Application of Machine learning

4.2.1 Supervised learning

4.2.2 Unsupervised learning

4.3 Structure of Regression Model

4.4 Linear Regression

4.5. Logistics Regression

4.6 Time series analysis

4.7 Support Vector Machine

UNIT V
Classification &Unsupervised Learning

5.1 Classification: Classification Problem

5.2 Classification Models

5.3 Classification Trees

5.3.1 Bayesian Method;

5.3.2. Association Rule: Structure of Association Rule

5.3.3. Apriori Algorithm

5.4. General Association

5.4.1. Clustering Methods

5.4.2. Partition Methods

5.4.3 Hierarchical Methods.

UNIT VI
Exploring Data in R

6.1. What is R? Its advantages

6.2 Basic features of R

6.3 Exploring R GUI

6.3.1 Managing graphics

6.3.1.1. Opening several graphics devices

6.3.1.2. Partitioning graphics

6.4. Data Frames & Lists

6.4.1 Data Frames

6.4.1.1. Making Data Frames

6.4.1.2. attach ( ) and detach ( )

6.4.1.3 Working with data frames

6.4.1.4 Attaching arbitrary list

6.4.1.5. Managing the search path

6.4.2 Lists

6.4.2.1 Constructing and modifying lists

6.4.2.2. Concatenating lists

6.5. Handling Data in R Workspace

6.6. Reading Data Sets & Exporting Data from R
