Savitribai Phule Pune University
Final Year of Computer Engineering (2012 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**	**06 Hours**

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?, Applications 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.		

Books:

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

Group A

1. Installation of Hadoop & R
2. Building Hadoop MapReduce Application for counting frequency of words/phrase in simple text file.

Group B

1. Study of R: Declaring Variable, Expression, Function and Executing R script.
2. Creating List in R – merging two lists, adding matrices in lists, adding vectors in list.
3. Manipulating & Processing Data in R – merging data sets, sorting data, plotting data, managing data using matrices & data frames
4. Implementation of K-Means Clustering with R
5. Text Analysis using R: analyzing minimum three different data sets

Group C

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