RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

New Scheme Based On AICTE Flexible Curricula

Computer Science & Information Technology, VIII-Semester

CSIT-801 Data Science

Objective:
The objective of this course is to familiarize students with the roles of a data scientist and enable them to analyze data to derive meaningful information from it.

Course Outcomes: After the completion of this course, the students will be able to:

1. Demonstrate proficiency with statistical analysis of data.
2. Build and assess data-based models.
3. Execute statistical analyses with professional statistical software.
4. Demonstrate skill in data management.
5. Apply data science concepts and methods to solve problems in real-world contexts and will co mmunicate these solutions effectively

Unit I
Data Science and Big Data Overview: Types of data, Sources of data, Data collection, Data storage and management, Big Data Overview, Characterization of Big data, Drivers of Big Data, Challenges, Big Data Use Cases, Defining Big Data Analytics and examples of its use cases, Data Analytics Lifecycle: Discovery, Data Preparation, Model Planning, Model Building, Communicate Results, Operationalize.

Unit II
Advanced Analytical Theory and Methods: Clustering, K-means, Additional Clustering Algorithms, Association Rules, Apriori Algorithm, Applications of Association Rules, Regression, Linear Regression, Logistic Regression, Classification, Decision Trees, Naive Bayes, Additional Classification Methods, Text Analysis, Text Analysis Steps, Determining Sentiments.

Unit III
Advanced Analytics-Technology and Tools: Analytics for Unstructured Data Use Cases, MapReduce, Apache Hadoop, Traditional database vs. Hadoop, Hadoop Core Components, HDFS, Design of HDFS, HDFS Components, HDFS Architecture, Hadoop 2.0 Architecture, Hadoop-2.0 Resource Management, YARN.

Unit IV
The Hadoop Ecosystem: Introduction to Hive, Hbase, Hive Use Cases: Face book, Healthcare; Hive Architecture, Hive Components. Integrating Data Sources, Dealing with Real-Time Data Streams, Complex Event Processing, Overview of Pig, Difference between Hive and Pig, Use

Cases of Pig, Pig program structure, Pig Components, Pig Execution, Pig data models, Overview of Mahout, Mahout working.

Unit V

Introduction to R, Basic Data Analytics Methods Using R, Communicating and Operationalizing an Analytics Project, Creating the Final Deliverables, Data Visualization Basics.

Recommonded Books:

1. EMC Education Services, “Data Science and Big Data Analytics”, Wiley, 2015.
2. Judith Hurwitz, Alan Nugent, Fern Halper, and Marcia Kaufman, “Big Data for Dummies”,Wiley & Sons,2013.
3. VigneshPrajapati, “Big Data Analytics with R and Hadoop” ,Packt Publishing, 2013.
4. David Dietrich, Barry Heller, and Beibei Yang“Data Science and Big Data Analytics:Discovering, Analyzing, Visualizing and Presenting Data”, John Wiley & Sons, Inc.

List of Experiments:

1. Introduction to R tool for data analytics science
2. Basic Statistics and Visualization in R
3. K-means Clustering
4. Association Rules
5. Linear Regression
6. Logistic Regression
7. Naive Bayesian Classifier
8. Decision Trees
9. Simulate Principal component analysis
10. Simulate Singular Value Decomposition