Syllabus

Online Post Graduate Diploma in Data Science

Two Semesters ***Duration: One Year***

Department of Statistics

Assam University, Silchar

2024

Course Titles and Credits of Online Post Graduate Diploma in Data Science offered by

Department of Statistics: Assam University, Silchar

Duration of the Program: 1 year **Number of Semesters: 2**

Total Credits: 20 + 20 = 40 **Hours of Teaching: 240 hours**

First Semester					
Course Number	**Course Title**	**Total Credit**	**Hours of Teaching**	**Marks Distribution**	
				End Semester Exam	**Continuous Assessment**
101	Fundamentals of Data Science	5	30 hours	70	30
102	Statistics Essentials	5	30 hours	70	30
103	R Programming	5	30 hours	70	30
104	Machine Learning	5	30 hours	70	30
Second Semester					
201	Multivariate Data Analysis	5	30 hours	70	30
202	Python Programming	5	30 hours	70	30
203	Analysis of Big Data using Hadoop	5	30 hours	70	30
204	Project Work	5	30 hours	70	30

Department of Statistics: Assam University, Silchar

Program Name: Online Post Graduate Diploma in Data Science

Paper No. 101

Course Title: Fundamentals of Data Science

Credit: 5 **Hours of Teaching: 30**

Full marks: 100 End Semester (70) Internal Assessment (30)
Pass Marks: 40 End Semester (28) Internal Assessment (12)

Learning Objectives: This paper shall give the students, knowledge about the basic concepts of Data Science. Also this course aims to provide students with introductory knowledge of several data science techniques that can be used for data analytics. The topics covered in this paper can be utilized for further studies in Data Science of for the purpose of research. At the end of this course, students should be able to

i) Understand the basic concepts and philosophy of Data Science
ii) Develop the ability to build statistical analysis of data
iii) Learn about different domains of knowledge where Data Science can be applied
iv)

Unit I : Data Science Basics

Introduction, Key Components of Data Science, Main Application Areas of Data Science, Data Science process, Defining Big Data, Sources of Big Data, Statistical Thinking in the age of big data, Exploring Solutions for big data problems, Understanding Hadoop, Job Prospects with Data Science

Unit II: Extracting Meaning from Data

Exploratory Data Analysis (EDA), Concept of Population and Sample, Population and sample for Big Data, Using models to explore data, Comparing model expectations to reality, Use of numerical methods in Data Science, Application of Data Science in Business Decision Process

Unit III: Modelling Spatial Data and Data from Social Networks and Financial Modelling

Spatial data and its characteristics, predicting using statistical models, kriging algorithms, surface analysis on spatial data.
Social Network Analysis, Centrality analysis, Data Journalism, Preparing Financial Data, log returns, volatility measurement, exponential weighting.

Unit IV: Data Visualization

Exploratory Data Analysis through Graphs, Basic principles of Data Visualization design, Standard Charts (Univariate, bivariate and multivariate), Topological structures, spatial plots and maps, Overview of D3js, Online data visualization platforms, online geographic tools.

Unit V: Causality and Epidemiology

Causation, Causal effect, confounders, Observational studies, Rubin Causal Model, Visualizing Causality
Introduction to Epidemiology, Overview of clinical study designs and its types, Randomized Clinical Trials, Combatting Confounding,

Reference

1. Lillian Pierson (2015) *Data Science for Dummies*, John Wiley & Sons, Inc., New Jersey, USA
2. Rachel Schutt and Cathy O'Neil (2014) *Doing Data Science*, O' Really, California, USA
3. Peng R. D and Matsui E. (2015) *The Art of Data Science: A Guide for Anyone Who Works with Data*, Skybrude Consulting, LLC
4. Tukey, J. (1977) *Exploratory Data Analysis*, Addison Wesley Publishing Company

Department of Statistics: Assam University, Silchar

Program Name: Online Post Graduate Diploma in Data Science

Paper No. 102

Course Title: Statistics Essentials

Credit: 5 **Hours of Teaching: 30**

Full marks: 100 End Semester (70) Internal Assessment (30)
Pass Marks: 40 End Semester (28) Internal Assessment (12)

Learning Objectives: Knowledge of Basic Statistics is considered as the first step to understand analytics. As the participants of the program shall be from diverse background, so the course is designed to train the takers from fundamental knowledge of Statistics and computing. Along with the most essential statistical techniques necessary to explore the world of analytics the procedure of their computation in Excel and R will also be displayed. At the end of this course, students should be able to

i) Perform computation of different measures of descriptive statistics
ii) Draw basic statistical graphs with all relevant accessories using software
iii) Understand the concept of probability and probability distributions
iv) Simulation from standard probability distributions- both discrete and continuous
v) Acquire knowledge on interval estimation and their computation
vi) Perform parametric and non-parametric tests in R and Excel and for drawing inferences from them

UNIT-1: Basic Descriptive Statistics

Tabular and graphical presentation of data, measures of central tendency, dispersion, skewness, kurtosis, simple correlation, regression, measures of agreement- along with their computational aspects in R and Excel.

UNIT-II: Probability and Probability Distributions

Basic concepts of events, sample space and probability, conditional probability and independence, Bayes theorem, Probability mass function and probability density function of a random variable and mathematical expectation. Standard probability distributions including Binomial, Poisson, normal and exponential – computation of probabilities and simulation from the standard distributions in R and Excel.

UNIT-III: Estimation and Confidence Interval

Sampling distribution of a statistic, point estimation using maximum likelihood estimation, confidence interval estimation for mean, standard deviation and proportion, concept of prediction interval

UNIT-IV: Parametric Tests

Large and small sample tests concerning single mean, variance; difference of two means, paired sample mean, Analysis of variance (ANOVA)- one-way, two-way with and without replication, Multiple Regression, Logistic regression - and their computation in R and Excel (wherever applicable).

UNIT-V: Non-Parametric Tests

Introduction for Non-parametric tests, when non-parametric test can be used, Wilcoxon rank-sum test and Mann–Whitney test, Wilcoxon signed-rank test, the Kruskal–Wallis test, Chi-square tests for goodness of fit, independence of attributes, run test for randomness - and their computation in R.

Reference:

1. Anderson D.R., Sweeney D.J. and Williams T.A. (2012) *Essentials of Modern Business Statistics with Microsoft Excel* (10^{th} edition), Thomson South Western
2. Sharpe N. R,, Deveaux R. D. and Velleman P. P. (2010) *Business Statistics (2^{nd} edition)*, Addison Wesley
3. Field, A., Miles, J., and Field, Z. (2012). *Discovering statistics using R*. SAGE Publications.
4. Moore, D.S. (2010) *Basic Practice of Statistics*, Palgrave Macmillan
5. Paul, T. (2012) *R Cookbook*, O' Really, California, USA

Department of Statistics: Assam University, Silchar

Program Name: Online Post Graduate Diploma in Data Science

Paper No. 103

Course Title: R Programming

Credit: 5 **Hours of Teaching: 30**

Full marks: 100 End Semester (70) Internal Assessment (30)
Pass Marks: 40 End Semester (28) Internal Assessment (12)

Learning Objective: R has become a very powerful tool for statistical computation, graphics and programming. It is a free, open source system. The purpose of this course is to develop a basic understanding of the R working environment. We will introduce the necessary arithmetic and logical operators, salient functions for manipulating data, and getting help using R. Though R these days are used in areas beyond the purpose for which it was basically built, but the course shall focus on applications of R in Data Science including data analytics. The course shall focus on the construction of different statistical models using R. The paper provides an in-depth coverage on the computational aspects of various statistical techniques and goodness of fit tests used for data analytics. By the end of the course students shall be confident and equipped with:

i) Creating Vectors, handle variables and perform several basic functions
ii) Getting control over inputs and withdraw the desired outputs from R environment
iii) Perform Statistical computations and graphics in R
iv) Handling complex computation involving Matrices in R
v) Using of R for some advanced statistical modelling and applications

Unit I : Basics of R

Introduction to R, installing R, R studio, R package, R Data structures, Vectors, Factors, Managing Data with R, Exploring and understanding Data, Reading from and writing to CSV files, Reading data from the web, Arithmetic operators and Logical operations, Dealing with R packages, Working with R script.

Unit II : Data Structures

Working with vectors, dealing with Categorical variable, Combining multiple vectors, Creating list, working with individual elements of a list, Matrix initialization, Matrix Algebra in R, editing Data frames, combining Data frames.

Unit III : Data Transformation and Working with Strings

Splitting Vector into groups, application of functions to different arrangements of data like element, row, column, sub-groups etc. Working with strings, extracting substrings, string manipulation functions, working with dates,

Unit IV : Descriptive Statistics and Graphics in R

Basic descriptive statistics for grouped data, Correlation: Karl Pearson's, Spearman's, Kendall's tau, Point bi-serial, Phi, Regression models- linear and non-linear (simple and multiple), Ridge regression, Generalized Regression models, Change point analysis, Statistical graphs like- scatter plots (with all variants), box plots (simple and compound), bubble plots, histograms and its variants, mosaic plot, scatter-plot matrix.

Unit V : Statistical Models in R

Mathematical Models, Models with random components, simulation from discrete and continuous distributions, sampling from populations, checking model assumptions, resampling methods and its applications. Time Series analysis- Auto-regressive models and Auto-regressive Moving Average models.

Reference:

1. Schmuller, J. (2017). *Statistical Analysis with R For Dummies*. John Wiley & Sons.
2. Mailund, T. (2017). *Beginning Data Science in R: Data Analysis, Visualization, and Modelling for the Data Scientist*. Apress.
3. Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R*, JAGS, and Stan.
4. Teetor, P. (2011) *R Cookbook*, O'Reilly Media Inc. CA
5. Maindonald J. and Braun J. (2003) *Data Analysis and Graphics using R- an Example based Approach*, Cambridge University Press
6. Adler, J. (2010) *R in a Nutshell*, O'Reilly Media Inc. CA

Department of Statistics: Assam University, Silchar

Program Name: Online Post Graduate Diploma in Data Science

Paper No. 103

Course Title: Machine Learning

Credit: 5 **Hours of Teaching: 30**

Full marks: 100 End Semester (70) Internal Assessment (30)
Pass Marks: 40 End Semester (28) Internal Assessment (12)

Learning Objectives: The purpose of this course is to provide an accessible, yet comprehensive, basic concepts and applications of machine learning. It is intended to make the students understand the basic concepts of artificial intelligence, information theory and the other disciplines that incorporate machine learning, with balanced coverage of theory and practice. The course does not assume any background in artificial intelligence in different techniques that a data scientist needs to know for analyzing big data. The viewpoint is to provide the underlying key ideas about the machine learning solution and create a basis for developments and scopes in many application domains. On completing the course students shall be equipped about:

(i) Knowledge about the most important concepts of machine learnings
(ii) They shall understand the different applications of machine learning and why it is an important tool for data scientists
(iii) Handling the computational complexity of Machine Learning applications using different packages
(iv) Gather knowledge about the core areas of machine learning and apply them to analyze live datasets

Unit I : Basics of Machine Learning

Definition of learning systems, Goals and applications of machine learning, Aspects of developing a learning system: training data, concept representation, applications to other fields. Boolean Algebra, Class of Boolean functions, version spaces for learning.

Unit II : Classification

Classification using *k*-Nearest neighborhood and its application, Classification via Bayes' Rule, Classification using Expectation Maximization Algorithm and its application. Probabilistic Classification.

Unit III : Supervised and Unsupervised Learning

Parameter smoothing. Generative vs. discriminative training. Support Vector Machines, Random Forests, and Decision Tree.
Learning from unclassified data, Clustering: Hierarchical Agglomerative Clustering, k-means partitioned clustering, Expectation maximization (EM) for soft clustering. Semi-supervised learning with EM using labelled and unlabelled data.

Unit IV : Neural Networks and Forecasting Methods

Understanding Neural Networks, Activation Functions, Network Topology, backpropagation algorithm and its applications.
Understanding multiple regression, logistic regression and its applications, adding regression to trees.

Unit V : Temporal Difference Learning and Reinforcement Learning

Temporal pattern and prediction methods, supervised and temporal difference methods. Introduction to reinforcement learning, Q learning- functions and algorithms, Non-deterministic rewards and actions.

References:

1. Mitchell TM. (1997) *Machine Learning*, McGraw Hill Science.
2. Lantz B. (2013) *Machine Learning with R*, Packt Publishing Limited, UK
3. Smola, A., & Vishwanathan, S. V. N. (2008). *Introduction to Machine Learning*. Cambridge University, UK, 32(34), 2008.
4. Kodratoff, Y. (2014). *Introduction to Machine Learning*. Elsevier.
5. Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists.* O' Reilly Media, Inc, CA.
6. Jannach, D., Zanker, M. and Felfernig, A. (2010) *Recommender Systems: An Introduction*, Cambridge University Press

Department of Statistics: Assam University, Silchar

Program Name: Online Post Graduate Diploma in Data Science

Paper No. 201

Course Title: Multivariate Data Analysis

Credit: 5 **Hours of Teaching: 30**

Full marks: 100 End Semester (70) Internal Assessment (30)
Pass Marks: 40 End Semester (28) Internal Assessment (12)

Learning Objectives: When one needs to simultaneously study several characteristics observed from many subjects and is concerned with the inter-relationship between these characteristic then one resorts to multivariate analysis. Basically, all practical data analytics converges to the analysis of multivariate data. This paper shall provide the students knowledge about the philosophy of multivariate techniques and teaches them some most common tools of multivariate analysis. Analysis of multivariate data is computationally difficult and needs the application of statistical packages. The curriculum describes in details the purpose of different multivariate tools of data analysis, their computation using statistical packages and interpretation of the outcome of the analysis. The analysis of real life data using multivariate tools like- factor analysis, discriminant analysis, cluster analysis, principal component analysis etc. using appropriate statistical package shall also be covered. At the end of this course, students should be able to

i) Clean multivariate data and prepare it for the purpose of analysis
ii) Choose appropriate tool for analysing multivariate analysis
iii) Perform common multivariate analysis using statistical packages
iv) Interpret the outcome of the analytical results
v) Develop a comprehensive roadmap of data analysis by converting real life problems into quantitative models

Unit I : Multivariate Data Basics

Introduction of multivariate data sets, Application of Multivariate analysis in different domains of knowledge, Common multivariate distributions like: Multivariate Normal Distribution, Wishart Distribution, Distribution of sample mean vector and variance co-variance matrix.

Unit II: Multivariate Inference

Estimation of mean vector and the variance-covariance matrix, Different tests for mean vectors, Test of equality of two mean vectors, Test of equality of several mean vectors, Tests concerning covariance and correlation matrices.

Unit III: Principal Component Analysis and Factor Analysis

Introduction to Principal Components, Methods of extracting Principal Components from Correlation matrix, decision regarding number of principal components.
Introduction to Factor Analysis, estimation of factor loadings, methods of factor extraction, rotation of factors.

Unit IV: Cluster Analysis and Canonical Correlation

Introduction to Cluster Analysis, Distance and matching measures, Formation of clusters, *k*-mean clustering and hierarchical clustering.
Canonical correlation analysis and its interpretation, applications of Canonical correlation and relevant statistical tests.

Unit V: Discrimination, Classification and MANOVA

Introduction to classification methods, Bayes Classification Rules, Discriminant Analysis, Different methods of Discriminant Analysis, Fishers Discriminating Function, Inference concerning Discriminating Function.
Multivariate Analysis of variance: Assumptions, one-way and two-way classification and their applications.

Reference:

1. Bhuyan K. C. (2005) *Multivariate Analysis and its Applications*, New Central Book Agency, Kolkata, India.
2. Johnson, R.A. and Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis*, Sixth Edition, Pearson & Prentice- Hall.
3. Izenman, A. J. (2009), *Modern Multivariate Statistical Techniques*, Springer, Singapore
4. Alboukadel Kassambara (2017) *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*, STHDA
5. Alboukadel Kassambara (2017) *Practical Guide to Principal Component Methods in R*, STHDA
6. Everitt B. and Hothorn T. (2011) *An Introduction to Applied Multivariate Analysis with R*, Springer

Department of Statistics: Assam University, Silchar

Program Name: Online Post Graduate Diploma in Data Science

Paper No. 202

Course Title: Python Programming

Credit: 5 **Hours of Teaching: 30**

Full marks: 100 End Semester (70) Internal Assessment (30)
Pass Marks: 40 End Semester (28) Internal Assessment (12)

Learning Objectives: The course is designed to provide knowledge of Python to participants having little or no prior programming experience. The knowledge of Python language is essential for participants to develop their computational approaches to problem solving. Python programming is intended for software engineers, system analysts, program managers and user support personnel for solving problems from basic to advanced level. The main objective of the paper is to teach the participants programming skills in core Python and to acquire Object Oriented Skills. The course shall also enable the participants to write database applications in Python. On successful completion of this course the student will be able to:

i) Learn Python language for expressing computations
ii) Learn a systematic approach to organizing, writing and debugging Python programs
iii) Defining an ambiguous problem statement to computation formulation leading to solution of the problem
iv) Application of random techniques and simulation for solving problems that does not succumb to closed-form simulations
v) Learn how to utilize Python to apply statistical and visualization tools to model datasets
vi) To apply Python for web scrapping.

Unit I : Introduction to Python language, Python Keywords, Identifiers and Python Comments

What is Python? Basic elements of Python, Python Applications, Features of Python Programming Language, Python keywords or reserved words, Python Identifiers, class names, variable names, function names, method names, and Identifier naming rules. Implementations of Python, Modes of Programming in Python.
Comments for Understanding Python code, Python Comment Syntax, Python Single line comment, Multiline comment in Python, and writing Python comments.

Unit II : Python Variables, Data Types and Operators

Define Variables, Declaration of Variables, Assigning Values to Variables, Initialization, Reading, Variable naming restrictions, and Types of Python Variables.
Define Data Type, Python Numbers, Python Strings, Python Boolean data type.
Python Arithmetic, Comparison/Relational Operators, Increment Operators, Logical operators, Branching Programs, Python Identity Operators and Python Operators Precedence.

Unit III : Python Control Flow: Decision making, Looping and Branching

Decision Making / Conditional Statements in Python, Simple If Structure, if-else structure, if else if structure, and nested Structure.
Python Control Flow Statements, Python Loop Statements. Python while loop, Python for loop, Python range, Python Nested Loop Structures.
Python Flow Control: Branching Statements, Python Branching Statements.

Unit IV : Function, Objects, Lists, Testing and Debugging

Introduction on functions, functions and modules, built-in functions, writing functions (function basics, parameter passing, and documenting functions), higher order functions,
Using objects, string objects, file objects, using lists, building lists, lists and functions.
Testing, Black box and Glass box testing, Debugging, Abstract data types and classes, inheritance.

Unit V : Algorithm quality, Stochastic Programs, Plotting and Web Scrapping

Introduction to algorithm complexity, some important complexity classes, search and sort algorithms. Plotting with PyLab.
Stochastic programs, Statistical Inference and simulations, Monte Carlo simulation, Web Scrapping, different approaches of scrapping a webpage.
Building interfaces and developing applications in Python

Reference:

1. McGrath M., (2013) *Python in Easy Steps*, McGraw Hill Education (India) Pvt. Ltd.
2. Guttag, J.V., (2016) *Introduction to Computation and Programming using Python*, PHI Learning Pvt. Ltd.
3. Madhavan S. (2015). *Mastering Python for Data Science*, Packt Publishing Limited.
4. Lutz M. (2013). *Learning Python*, O'Reilly, 5th Edition.
5. Urban M. and Murach J. (2016). *Murach's Python Programming*, Shroff Publishers & Distributors Pvt. Ltd.
6. Baezly D M. (2009). *Python Essential Reference*, Fourth Edition. Addison-Wesley Professional.
7. Baezly D M. (2013). *Python Cookbook*, Third Edition, O'Reilly,.CA

Department of Statistics: Assam University, Silchar

Program Name: Online Post Graduate Diploma in Data Science

Paper No. 203

Course Title: Analysis of Big Data using Hadoop

Credit: 5 **Hours of Teaching: 30**

Full marks: 100 End Semester (70) Internal Assessment (30)
Pass Marks: 40 End Semester (28) Internal Assessment (12)

Learning Objective: Storing, processing and drawing inferences by analysing Big data is one of the main reasons for which one needs to learn Data Science. Any program on Data Science remains incomplete without proper knowledge of handling big data sets. Hadoop is an open source, Java based framework used for storing and processing of big data. The data is stored on inexpensive commodity servers that run as clusters. Its distributed file system enables concurrent processing and fault tolerance. Hadoop uses the MapReduce programming model for faster storage and retrieval of data from its nodes. This course try to provide the participants a comprehensive knowledge in working with Hadoop file system, concept of MapReduce and Spark programming and apply these to different data management strategies of big data. On completion of this course, the participants will be able to:

i) Understand the preliminary concept of Big Data
ii) Understand Hadoop file system and concepts of MapReduce
iii) Apply parallel programming using Spark.
iv) Apply different data management strategies in big data scenarios

Unit 1: Basics of Big Data

Introduction to big data: Definition, Vs defining Big data, Different structures of data, Life cycle of analytics project, Big data enabling technologies: Apache Hadoop, Map reduce, Hadoop Ecosystem, Hive, Spark, Zoo keeper, No SQL, Cassandra, Hbase, Spark, Kafka, Hadoop Stack for Big Data: Basics of Hadoop, Scalability, Reliability, Basic Modules of Hadoop, Architecture of Hadoop, Hadoop Distributed file system (HDFS), Map Reduce Engine, versions of Hadoop, Yarn, Apache Scoop, PIG, Hive, Oozie, Zookeeper, Impala.

Unit 2: Hadoop Distributed File System (HDFS) and MapReduce

Introduction to HDFS, HDFS design concepts, Basic architecture of HDFS, Key components of HDFS, Various HDFS designs, federation block pools, HDFS performance Measure, HDFS

block size, No of blocks in a file, no of files and their impact, read write process in HDFS, HDFS Tuning parameters, HDFS block size, Replication in HDFS, robustness in HDFS, Map Reduce 1.0: The task tracker, execution steps in Mapreduce job, distributed file system, motivation for Map Reduce, Map and reduce steps in details, MapReduce function with wordcount example. Map Reduce 2.0: Application of Mapreduce, the YARN scheduler, Map Reduce Examples.

Unit 3: Parallel programming and Spark

Introduction to spark, Scala vs Java API, Scala and functional programming, Spark concepts, main primitives, resilient distributed database (RDD) and fault tolerance, creating RDD, transformations and actions, working with key value pairs, controlling the level of parallelism, page rank algorithm, page rank performance.

Need of spark, operations in spark, application of spark, execution of spark, distributed programming, flume in java, Available APIs, Built in libraries in Spark: standard libraries for big data, machine learning libraries, GraphX, Spark streaming, Design of Key Value Stores: Key value abstraction, relational database and it's limitation, key-value/No-SQL model, Design of Apache Casandra, internal operations in Casandra.

Unit 04: Data Management in Hadoop

Data Placement strategies, bloom filter, compaction, reads, deletes, membership, Casandra vs RDBMS, CAP Theorem: Availability, consistency, partition tolerance, CAP theorem fallout, CAP Tradeoff, eventual consistency, RDBMS vs Key Value stores, consistency in Casandra, Quorums, Casandra Consistency levels. Consistency Solutions: eventual consistency, newer consistency models, comparisons of consistency models, Zookeeper: why do we need it, class distribution system, fault tolerant distribution system, race condition, deadlock, coordination, Zookeeper, design goals of Zookeeper, Architecture, election in Zookeeper, sessions, states, guarantees, operations, Access control lists, Zookeeper applications, Cassandra Query Language(CQL): Features of CQL, C* Model, introduction to CQL, CQL/Casandra mapping,

Unit 05: HBase, Spark and Kafka

Introduction to HBase, HBase architecture, components of HBase, HBase Data Model, HBase Storage Hierarchy, Cross Datacentre replication, Auto sharing and distribution, Bloom filter, Fold, Store and Shift.
Big streaming data processing, Fault tolerant stream processing, Spark Ecosystem, Spark Streaming, Spark Streaming features, working of spark streaming, Spark application, Spark streaming workflow, smart window based countByValue, Smart window based reduce, arbitrary stateful computation, Spark Streaming – Dstreams, Batches and RDD, Dstream classes, Spark Streaming Operation, fault tolerance, Sliding window analytics, Sliding window function, moving average example.
Batch vs Streaming, Kafka History, Use cases, Kafka Data Model, Partition distribution, producer, consumers, Kafka Architecture, publish-Subscribe system, brokers, Kafka Guarantees, Persistence in Kafka.

References:

1. White T. (2015) *Hadoop: The Definitive Guide*, 4th Edition, O'Reilly Publication, CA
2. Sammer E. (2012) *Hadoop Operations*, O'Reilly Publication, CA
3. Lam C. (2010) *Hadoop in Action*, First Edition, Manning Publications
4. Liu, H.H. (2012) *Hadoop Essentials: A Quantitative Approach*, 1st edition, Createspace Independent Pub
5. Prajapati V. (2015) *Big Data Analytics with R and Hadoop*, 2015, authored by Packt Publishing
6. EMC Education Services (2015) *Data Science and Big Data Analytics Discovering, Analyzing, Visualizing and Presenting Data*, Willey Publication

Department of Statistics: Assam University, Silchar

Program Name: Online Post Graduate Diploma in Data Science

Paper No. 204

Course Title: Project Work

Credit: 5 **Contact Hours : 30**

Full marks: 100 End Semester (70) Internal Assessment (30)
Pass Marks: 40 End Semester (28) Internal Assessment (12)

Learning Objective: This paper enables to train the participants of the program to undertake projects individually and also aims at providing opportunity to the participants to apply the various machine learning methods and applications in R and Python in a live project. The projects shall enable the students to take up their own study and to understand the application of data science methods that they learned during the course. The project requires the students to synthesize the topics from the course into some theme of practical use. They are expected to design computer programs or use statistical models in the project. For the project each student shall work under the supervision of a faculty member allotted by the department. The topic of the project shall be decided by the participant in consultation with the supervisor. The project shall be individual and no group project shall be allowed in this paper. Participants are to attain at least one objective in their project. They are required to apply packages like R and/or Excel and/or Python and/or Hadoop etc. and perform advanced analytical tools of machine learning and multivariate analysis. *Thus, students shall get some exposure to practical analysis of Big Data, one of the demanding qualification for modern day data analyst.*

Implementation of the Project Work

In consultation of the supervisor, students shall decide on a researchable topic for their project. The proposal of the project shall contain the detailed objective(s) and methodology. Once approved by the supervisor the student shall start working on the project. There shall be two mid-term evaluations of the project to check the continuous progress of the student. Before the start of the end-semester examination students are required to submit the project report/dissertation in hard copy in duplicate. During the end semester examination students shall present the same, whereby they shall be evaluated by an external examiner.

***** *****