MCAH201: DATA ANALYTICS WITH R/PYTHON

Hours/Week: 4 I.A. Marks: 30

Credits: 4 Exam. Marks: 70

Course Learning Objectives: Students will try to learn,

1. The probability distributions and density estimations to perform analysis of various kinds of data
2. The statistical analysis techniques using Python and R programming languages.
3. Expand the knowledge in R and Python to use it for further research.
4. The students will be able to carry out data analysis/statistical analysis effectively visualize the data.

Course Outcomes: After completing the course, the students will be able to,

CO1: Understand the fundamentals of data analytics and study the basic concepts of Excel spreadsheet Functions.

CO2: Realize the importance of filtering functions, charts and tables.

CO3: Identify the importance and usage of R package and its features

CO4: Learn the fundamentals of python programming

C05: Understand the various search methods and visualization techniques.

C06: Learn to use various techniques for mining data stream and applications using Map Reduce Concepts.

C07: Introduce programming tools PIG & HIVE in Hadoop echo system

---

UNIT-I **12Hrs.**

Basics of Data Analytics, Applications of Data Analytics, Phases in Data Analytics, Data Definitions and Analysis Techniques, Elements, Variables, and Data categorization, Levels of Measurement, Data Management and Indexing, Introduction to Statistical Learning and R-Programming.

UNIT-II **12Hrs.**

Introduction to R- Packages, Scientific Calculator- Inspecting Variables- Vectors, Matrices and Arrays- Lists and Data Frames- Functions- Strings and Factors- Flow Control and Loops- Advanced Looping- Date and Times. Introduction to Python Packages- Fundamentals of Python- Inserting and Exporting Data- Data Cleansing Checking and Filling Missing Data- Merging Data- Operations- Joins.

UNIT-III **12Hrs.**

Basic analysis techniques, Statistical hypothesis generation and testing, Chi-Square test, t-Test, Analysis of variance, Correlation analysis, Maximum likelihood test, Regression analysis, Classification techniques, Clustering, Association Rules Analysis, Practice and analysis with R/Python

UNIT-IV **12Hrs.**

Hadoop: History of Hadoop- the Hadoop Distributed File System – Components of Hadoop Analyzing the Data with Hadoop- Scaling Out- Hadoop Streaming- Design of HDFS-Java interfaces to HDFS Basics- Developing a Map Reduce Application-How Map Reduce Works- Anatomy of a Map Reduce Job run-Failures-Job Scheduling-Shuffle and Sort – Task execution - Map Reduce Types and Formats- Map Reduce Features Hadoop environment.

REFERENCE BOOKS:

1. Mukhopadhyay, Sayan. Advanced Data Analytics Using Python: With Machine Learning, Deep Learning and NLP Examples. Apress, 2018.
2. Jiawei Han, Micheline Kamber "Data Mining Concepts and Techniques", 2 nd Edition, Elsevier, Reprinted 2008.
3. Dalgaard, Peter, "Introductory statistics with R", Springer Science & Business Media, 2008.
4. McKinney, Wes. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. O'Reilly Media, Inc., 2012, 1st Edition.
5. E. Alpaydin, "Machine Learning", MIT Press, 2010.
6. Samir Madhavan, Mastering Python for Data Science, 2015.