|  |  |  |  |
| --- | --- | --- | --- |
| **BID – 703** | | **Advanced Data Science** |  |
| **LEARNING OBJECTIVES:** | | | |
| 1. Explain advance methods and algorithms of Data Science. 2. Discuss the concepts and various in-built functions associated with NumPy. 3. Familiarize with data science environment and data structures of NumPy (Python- Programming language). 4. Explore various packages and applications of Python required for the analysis and visualization of data. | | | |
| **LEARNING OUTCOMES:** | | | |
| **Students will be able to:**   1. Design models and scripts with interactive environment for solving problems using biological data. 2. Analyze different type of data using advanced functions of NumPy. 3. Understanding of Hadoop Architecture and its application in data science. 4. Construct predicting systems associated with computational biology, machine learning and artificial intelligence. | | | |
| **Unit I:** | **Advance Methods of Data Science and Algorithms** | |  |
| Central Limit Theorem; AB Testing; Recommended systems; Dimensionality Reduction; Singular Value Decomposition; Manifold Learning and Diffusion Maps; Spectral Clustering; Approximation Algorithms; Synchronization. Introduction to Scikit Learn; Bagging and Boosting Hyperparameters and Model Validation; Feature Engineering; Algorithms for Data Science: Linear Logistic Regression, Decision Trees, k-means, Principal Component Analysis. | | | |
| **Unit II:** | **Introduction to NumPy** | |  |
| Brief Python Data Types; NumPy arrays; Data Types; Array Functions; Universal Functions; Aggregations; Broadcasting; Fancy Functions; Sorting arrays: Partial Sort; Structures arrays; Data Manipulation using Pandas; Handling Missing Data; Hierarchical Indexing; Time Series Visualization Using Matplotlib. | | | |
| **Unit III:** | **Hadoop Architecture** | |  |
| Introduction to Hadoop; Framework; Modules: Hadoop Common, Hadoop YARN, Hadoop Distributed File Systems (HDFS), Hadoop MapReduce; Architecture; Environment Setup; Operation Modes; HDFS Overview: Features of HDFS, Architecture, Goals; HDFS Operations; Hadoop Command Reference; MapReduce: Algorithms, Terminologies, MapReduce Command and Jobs; Streaming: Mapper Phase Code, Reducer Phase Code; Multi-Node Cluster. | | | |
| **Unit IV:** | **Applications of NumPy in Data Analysis** | |  |
| Data Science case studies; TITANIC Data Science; Programming associated with Data Science (Online Notebooks); Applications in Machine Learning; Application in Python | | | |
| **Recommended Books:** | | | |
| 1. Doing Data Science, Straight Talk from the Frontline, Cathy O'Neil and Rachel Schutt, O'Reilly Media, Inc., (2013). 2. Joel Grus. Data Science from Scratch. O'Reilly Media, Inc., 2nd Edition (2019). 3. Hadoop: The Definitive Guide, Tom White, O'Reilly Media, Inc., 4th Edition (2015). 4. Python Data Science Handbook, Jake Vanderplas, O'Reilly Media, Inc., (2016). 5. Python for Data Analysis, Data Wrangling with Pandas, NumPy, and IPython, Wes McKinney, O'Reilly Media, Inc., 2nd Edition (2017). | | | |