|  |  |
| --- | --- |
| MONO | **DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**  **NATIONAL INSTITUTE OF TECHNOLOGY PATNA**  Ashok Raj Path, PATNA 800 005 (Bihar), India |
| Phone No.: 0612 – 2372715, 2370419, 2370843, 2371929, 2371930, 2371715 Fax – 0612- 2670631 Website: [www.nitp.ac.in](http://www.nitp.ac.in/) |

***CSX449 Data Science***

**L-T-P-Cr: 3-0-0**

**Pre-requisites:** Machine Learning

**Objectives/Overview:**

* To visualize, manipulate and predict useful results from different data sources.
* To learn unsupervised, supervised and machine learning algorithms on the real-world data using Python.
* To introduce big data architecture.

**Course Outcomes:**

At the end of the course, a student should be able to:

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Outcome** | **Mapping to POs** |
|  | Handle and summarize data from a statistical perspective. | PO1, PO4 |
|  | Learn fundamentals of python programming necessary for handling massive data. | PO3 |
|  | Perform data manipulation and visualization programmatically. | PO3 |
|  | Perform classification and machine learning techniques using various APIs. | PO3, PO6 |
|  | Perform map-reduce and unsupervised operations on real world data. | PO3, PO6 |

**UNIT I: Lectures: 1**

Data Science applications, recommendation system, image recognition, digital advertising, big data.

**UNIT II: Lectures: 9**

Basics of NumPy Arrays, Computation on NumPy Arrays: Universal Functions, Aggregations: Min, Max, and Everything in Between, Computation on Arrays, Comparisons, Masks, and Boolean Logic, Sorting Arrays, Structured Data: NumPy’s Structured Arrays.

**UNIT III: Lectures: 8**

Data Indexing and Selection, Operating on Data in Pandas, Handling Missing Data, Hierarchical Indexing, Combining Datasets, Aggregation and Grouping, Pivot Tables, Vectorized String Operations, Working with Time Series.

**UNIT IV: Lectures: 3**

Line Plots, Scatter Plots, Visualizing Errors, Density and Contour Plots, Histograms, Binnings, Customizing Colorbars.

**UNIT V: Lectures: 7**

Naive Bayes Classification, Gaussian Naive Bayes, Multinomial Naive Bayes, Linear Regression, Multiple Regression, logistic Regression, Introduction to Scikit-Learn, Data Representation in Scikit-Learn, Scikit-Learn’s Estimator API.

**UNIT VI: Lectures: 7**

Support Vector Machines, Maximizing the Margin, Decision Trees, Entropy, and Creation of decision tree and Random Forests, Perceptrons, feed-forward neural network, backpropagation.

**UNIT VII: Lectures: 7**

Clustering: k-Means Algorithm, Expectation–maximization, Network Analysis, betweenness entrality, eigenvector centrality, directed graphs and page ranks, MapReduce, mapper and reducer functions with examples.

**Text/Reference Books:**

1. Data Science from Scratch by Joel Grus, O’Reilly.
2. Python Data Science Handbook by Jake VanderPlas, O’Reilly.
3. Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools by [Davy Cielen](https://www.amazon.in/s/ref=dp_byline_sr_book_1?ie=UTF8&field-author=Davy+Cielen&search-alias=stripbooks), [Arno D.B. Meysman](https://www.amazon.in/s/ref=dp_byline_sr_book_2?ie=UTF8&field-author=Arno+D.B.+Meysman&search-alias=stripbooks), [Mohamed Ali](https://www.amazon.in/s/ref=dp_byline_sr_book_3?ie=UTF8&field-author=Mohamed+Ali&search-alias=stripbooks) , Manning.