# Mining Massive Datasets
## (Professional Elective –III)

| | | | | | |
|---|---|---|---|---|---|
| **Course Code** | 20IT4701E | **Year** | IV | **Semester** | I |
| **Course Category** | **PE3** | **Branch** | IT | **Course Type** | Theory |
| **Credits** | 3 | **L-T-P** | 3-0-0 | **Prerequisites** | Data mining |
| **Continuous Internal Evaluation:** | 30 | **Semester End Evaluation :** | 70 | **Total Marks:** | 100 |

**<u>Correlation between CO – PO, CO- PSO</u>** (Use √ symbol for representing correlation)

| | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 | PSO1 | PSO2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CO1 | √ | | | | | | | | | | | | | |
| CO2 | | √ | | | | | | | | | | | | |
| CO3 | | √ | | | | | | | | | | | | |
| CO4 | | | | √ | | | | | | | | | √ | √ |

**<u>Strength of Correlation between CO – PO, CO- PSO in scale of 1-3</u>**

1: Slight (low), 2: Moderate (medium) 3: Substantial (High)

| | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 | PSO1 | PSO2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CO1 | 3 | | | | | | | | | | | | | |
| CO2 | | 3 | | | | | | | | | | | | |
| CO3 | | 3 | | | | | | | | | | | | |
| CO4 | | | | 3 | | | | | | | | | 3 | 3 |
| Overall course | 3 | 3 | | 3 | | | | | | | | | 3 | 3 |

| **Course Outcomes** | | **Blooms Taxonomy Level** |
|---|---|---|
| **Upon Successful completion of course, the student will be able to** | | |
| CO1 | Recollecting fundamentals of data mining. | L2 |
| CO2 | Apply the concept of Map reduce and data streams for storing and processing of massive data sets | L3 |
| CO3 | Analyze the issues underlying the effective applications of massive data sets | L4 |
| CO4 | Evaluate different clustering algorithms and analyze various decomposition techniques | L4 |

| **Syllabus** | | |
|---|---|---|
| **Unit No** | **Contents** | **Mapped CO** |
| **I** | **Data Mining: Introduction,** Statistical Modeling, Machine Learning, Computational Approaches to Modeling, Feature Extraction, Statistical Limits on Data Mining, Hash Functions, Indexes, Natural Logarithms, Power Laws. | **CO1** |

| | | |
|---|---|---|
| **II** | **Map Reduce and the New Software Stack:** Distributed File Systems, Map Reduce, Algorithms Using MapReduce, Extensions to MapReduce, Complexity Theory for MapReduce. | **CO2** |
| **III** | **Mining Data Streams:** The Stream Data Model, Sampling Data in a Stream, Filtering Streams, Counting Distinct Elements in a Stream, Counting Ones in a Window, Decaying Windows. | **CO1,CO2** |
| **IV** | **Frequent Item sets:** The Market-Basket Model, Market Baskets and the A-Priori Algorithm, Handling Larger Datasets in Main Memory, Limited-Pass Algorithms, Counting Frequent Items in a Stream. | **CO1,CO3** |
| **V** | **Clustering:** Introduction to Clustering Techniques, Hierarchical Clustering, K-means Algorithms, The CURE Algorithm, Clustering in Non-Euclidean Spaces, and Clustering for Streams and Parallelism.<br>**Dimensionality Reduction:** Eigen values and Eigenvectors of Symmetric Matrices, Principal-Component Analysis, Singular-Value Decomposition, CUR Decomposition | **CO1,CO4** |
| **Learning Resources** | | |
| **Text Books** | | |
| 1.Mining of Massive Datasets - Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman" (LaTeX) | | |