

**PVP SIDDHARTHA INSTITUTE OF TECHNOLOGY, KANURU, VIJAYAWADA  
(AUTONOMOUS)  
INFORMATION TECHNOLOGY**

**MINING MASSIVE DATASETS  
(Professional Elective –III)**

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

**Strength of Correlation between CO – PO, CO- PSO in scale of 1-3**

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

<b>Course Outcomes</b>		<b>Blooms Taxonomy Level</b>
<b>Upon Successful completion of course, the student will be able to</b>		
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	<b>Data Mining: Introduction,</b> Statistical Modeling, Machine Learning, Computational Approaches to Modeling, Feature Extraction, Statistical Limits on Data Mining, Hash Functions, Indexes, Natural Logarithms, Power Laws.	CO1
II	<b>Map Reduce and the New Software Stack:</b> Distributed File Systems, Map Reduce, Algorithms Using MapReduce, Extensions to MapReduce, Complexity Theory for MapReduce.	CO2
III	<b>Mining Data Streams:</b> 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	<b>Frequent Item sets:</b> 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	<b>Clustering:</b> Introduction to Clustering Techniques, Hierarchical Clustering, K-means Algorithms, The CURE Algorithm, Clustering in Non-Euclidean Spaces, and Clustering for Streams and Parallelism. <b>Dimensionality Reduction:</b> 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)		