IV Year B.Tech. ECM II - Semester	L	T	P	To	C
	4	-	-	4	4

CS443 BIG DATA ANALYTICS (ELECTIVE VI)

Course Description and Objectives:

The main objectives of this course is to enable the students with basic data analytic skills like regression analysis, classification techniques, clustering techniques, association rule mining. Further, this course also enables the students how to scale the above algorithms with different data environments like massive amount of data, streaming data, distributed data and provides hands on experience on real world problems using above theoretical background.

Course Outcomes

- *Necessary theory background for processing analytics.*
- *Processing analytics on small scale data.*
- *Mining from massive datasets.*
- *Mining from distributed datasets.*

UNIT I - Introduction To Big Data

Introduction to BigData Platform – Traits of Big data -Challenges of Conventional Systems - Web Data – Evolution Of Analytic Scalability - Analytic Processes and Tools - Analysis vs Reporting - - Statistical Concepts: Sampling Distributions - Re-Sampling - Statistical Inference - Prediction Error.

UNIT II - Data Analysis

Regression Modeling - Multivariate Analysis - Bayesian Modeling - Inference - Support Vector and Kernel Methods - Analysis of Time Series: Linear Systems Analysis - Nonlinear Dynamics - Rule Induction - Neural Networks: Learning

UNIT III - Advanced Learning And Introduction To Streaming

Generalization - Competitive Learning - Principal Component Analysis and Neural Networks - Fuzzy Logic: Extracting Fuzzy Models from Data – Fuzzy c-Means- Stochastic Search Methods. Introduction to Streams Concepts – Stream Data Model and Architecture - Stream Computing - Sampling Data in a Stream – Filtering Streams

UNIT IV - Frequent Itemsets And Clustering

Mining Frequent Itemsets - Market Based Model – Apriori Algorithm, FP-Growth, Dynamic Item set Algorithm – Clustering Techniques – Hierarchical – K-Means, K-medoid, CURE- Clustering High Dimensional Data – CLIQUE– Clustering in Non-Euclidean Space – Clustering for Streams and Parallelism.

UNIT V - Frameworks And Visualization

MapReduce – Hadoop, Hive, MapR – Sharding – NoSQL Databases - S3 - Hadoop Distributed

File Systems – Visualizations - Visual Data Analysis Techniques - Interaction Techniques; Systems and Analytics Applications.

TEXT BOOKS:

1. Michael Berthold, David J. Hand, “Intelligent Data Analysis”, Springer, 2007.
2. AnandRajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, Cambridge University Press, 2012.

REFERENCE BOOKS:

1. Bill Franks, “Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics”, John Wiley & sons, 2012.
2. Glenn J. Myatt, “Making Sense of Data”, John Wiley & Sons, 2007
3. Pete Warden, “Big Data Glossary”, O’Reilly, 2011.
4. Jiawei Han, MichelineKamber “Data Mining Concepts and Techniques”, Second Edition, Elsevier, Reprinted 2008.