

Paper Code – CSM402(01) Paper Name – Introduction to Data Science (Elective-II)		Full Marks: 100
Module	Topics	Lecture Hours
Module-1: Introduction	What is Data Science? ; Big Data and Data Science; Datafication; Current landscape of perspectives; - Skill sets needed	3
	Statistical Inference: Populations and samples; Statistical modelling, probability distributions; fitting a model; Introduction to R	3
	Exploratory Data Analysis and the Data Science Process; Basic tools (plots, graphs and summary statistics) of EDA; - Philosophy of EDA; The Data Science Process; Case Studies	2
Module-2: Three Basic Machine Learning Algorithms	Linear Regression; k-Nearest Neighbors (k-NN); k-means One More Machine Learning Algorithm and Usage in applications Motivating application: Filtering Spam Why Linear Regression and k-NN are poor choices for Filtering Spam Naive Bayes and why it works for Filtering Spam Data Wrangling: APIs and other tools for scrapping the Web	8
Module-4: Recommendation Systems: Building a User-Facing Data Product	Feature Generation and Feature Selection (Extracting Meaning from Data) Motivating application: user (customer) retention Feature Generation (brainstorming, role of domain expertise, and place for imagination) Feature Selection algorithms Filters; Wrappers; Decision Trees; Random Forests Algorithmic ingredients of a Recommendation Engine Dimensionality Reduction Singular Value Decomposition Principal Component Analysis Exercise: build your own recommendation system	8
Module-5: Mining Social-Network Graphs	Social networks as graphs Clustering of graphs Direct discovery of communities in graphs Partitioning of graphs Neighbourhood properties in graphs	6
Module-6: Data Visualization	Basic principles; ideas and tools for data visualization Examples of inspiring (industry) projects Exercise: create your own visualization of a complex dataset	6
Module-7: Data Science and Ethical Issues	Discussions on privacy, security, ethics A look back at Data Science Next-generation data scientists	4
Text Book: 1. Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk from the Frontline. O'Reilly. 2014. Additional references and books: 1. Jure Leskovek, Anand Rajaraman and Jeffery Ullman. Mining of Massive Datasets. v2.1, Cambridge University Press. 2014. (free online) 2. Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. ISBN 0262018020. 2013. 3. Foster Provost and Tom Fawcett. Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking. ISBN 1449361323. 2013. 4. Trevor Hastie, Robert Tibshirani and Jerome Friedman. Elements of Statistical Learning, Second Edition. ISBN 0387952845. 2009. (free online) 5. Avrim Blum, John Hopcroft and Ravindran Kannan. Foundations of Data Science. 6. Mohammed J. Zaki and Wagner Miera Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press. 2014. 7. Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, Third Edition. ISBN 0123814790. 2011		