SEMESTER I

DISCIPLINE SPECIFIC CORE (DSC) COURSES

Name of the Programme : MSc. in Data **Science**
Course Code : **CSD-500**
Title of the **Course** : Fundamentals of Data Science (Theory)
Number of Credits : 2(2L-0T- 0P)
Contact hours : **30 hours (30L-0T-0P)**
Effective from AY : **2023-24**

Pre-requisites for the course	Statistics and probability theory and python programming	
Objectives	The objective is to gain a comprehensive understanding of data science, covering fundamental concepts, tools, and techniques.	
Content	Unit I: Introduction: Typology of problems - Data science in a big data world: Benefits and uses of data science and big data-Facets of data-The data science process-The big data ecosystem and data science-The data science process: Overview of the data science process- Defining research goals and creating a project charter-Retrieving data-Cleansing, integrating, and transforming data-Exploratory data analysis-Build the models- Presenting findings and building applications on top of them. Mathematics for Data Science — A quick Review: Importance of linear algebra, statistics and optimization from a data science perspective; Structured thinking for solving data science problems. Linear Algebra: Matrices and their properties (determinants, traces, rank, nullity, etc.); Eigenvalues and eigenvectors; Matrix factorizations; Inner products; Distance measures; Projections; Notion of hyperplanes; half-planes. Probability, Statistics and Random Processes: Probability theory and axioms; Random variables; Probability distributions and density functions (univariate and multivariate); Expectations and moments; Covariance and correlation; Statistics and sampling distributions; Hypothesis testing of means, proportions, variances and correlations; Confidence (statistical) intervals; Correlation functions; White-noise process. Data clearing (EDA) Introduction to Data Science Methods: Linear regression as an exemplar function approximation problem; Linear classification problems-PCA	IS **hours**

	Unit II Handling large data on a single computer - The problems you face when handling large data-General techniques for handling large volumes of data-General programming tips for dealing with large data sets - Case study 1: Predicting malicious URLs - First steps in big data-Distributing data storage and processing with frameworks Introduction to NoSQL The rise of graph databases Introducing connected data and graph databases Introducing Neo4j: a graph database Data visualization to the end user Data visualization options Cross filter, the JavaScript MapReduce library Creating an interactive dashboard with dc.js Dashboard development tools	15 **hours**
Pedagogy	Lectures/ Tutorials/Hands-on assignments/Self-study/Flipped classroom	
References / Readings	1. Baesens, B. (2014). Analytics in a big data world: The essential guide to data science and its applications. John Wiley & Sons. 2. Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical statistics for data scientists: 50+ essential concepts using R and Python. O'Reilly Media. 3. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer. 4. McKinney, W. (2022). Python for data analysis. " O'Reilly Media, Inc.". 5. Taddy, M. (2019). Business data science. 6. Wheelan, C. Naked Statistics: Stripping the Dread from the Data.	
Course Outcomes	1. Understanding of fundamental concepts and techniques in data science. 2. Proficiency in data manipulation, analysis, and visualization using tools like Python or R. 3. Introduction to machine learning algorithms and evaluation methods. 4. Awareness of ethical considerations and responsible practices in data science.	

(Back to Index)