23CD21T1- INTRODUCTION TO DATA SCIENCE

(AI&DS)

Course Category:	Professional Core	Credits:	3
Course Type:	Theory	Lecture-Tutorial-Practical:	3-0-0
Prerequisite:	Mathematics and Statistics Programming Skills	**Sessional Evaluation:** **Univ. Exam Evaluation:** **Total Marks:**	30 70 100
Objectives:	**Students undergoing this course are expected:** - Knowledge and expertise to become a data scientist. - Essential concepts of statistics and machine learning that are vital for data science - Significance of exploratory data analysis (EDA) in data science. - Critically evaluate data visualizations presented on the dashboards - Suitability and limitations of tools and techniques related to data science process		

	Upon successful completion of the course, the students will be able to:		
Course Outcomes	CO1	Understand significance of Data Science. (L2)	
	CO2	Analyze large data(L4)	
	CO3	Understand the NoSQL movement for handling big data, including distributing data storage and processing with the Hadoop framework, the ACID principles. (L2)	
	CO4	Proficient in utilizing Neo4j and the Cypher query language for managing and querying graph databases, applying Python libraries such as NLTK and SQLite for text mining. (L3)	
	CO5	Perform Data reduction and apply visualization techniques. Apply the Data Science process for real world problems(L3)	
Course Content	### UNIT-I **Introduction:** Introduction to Data Science, benefits, and uses, facets of data, data science process in brief, big data ecosystem, and data science. **Data Science process:** Overview, defining goals and creating project charter, retrieving data, cleansing, integrating and transforming data, exploratory analysis, model building, presenting findings, and building applications on top of them. ### UNIT-II Applications of machine learning in Data science, the role of ML in DS, Python tools like sklearn, modeling process for feature engineering, model selection, validation and prediction, types of ML, semi-supervised learning. **Handling large data:** problems and general techniques for handling large data, programming tips for dealing large data, case studies on DS projects for predicting malicious URLs, for building recommender systems.		

UNIT-III

NoSQL movement for handling Bigdata: Distributing data storage and processing with Hadoop framework, case study on risk assessment for loan sanctioning, ACID principle of relational databases, CAP theorem, base principle of NoSQL databases, types of NoSQL databases, case study on disease diagnosis and profiling.

UNIT-IV

Tools and Applications of Data Science: Introducing Neo4jfor dealing with graph databases, graph query language Cypher, Applications graph databases, Python libraries like nltk and SQLite for handling Text mining and analytics, case study on classifying Reddit posts

UNIT-V

Data Visualization and Prototype Application Development: Data Visualization options, Crossfilter, the JavaScript MapReduce library, Creating an interactive dashboard with dc.js, Dashboard development tools.

Applying the Data Science process for real world problem solving scenarios as a detailed
case study.

Text Books & References Books	**TEXT BOOKS:** 1. Davy Cielen, Arno D.B.Meysman, and Mohamed Ali, "Introducing to Data Science using Python tools", Manning Publications Co, Dreamtech press, 2016 2. Prateek Gupta, "Data Science with Jupyter" BPB publishers, 2019 for basics. **REFERENCE BOOKS:** 1. Joel Grus, "Data Science From Scratch", OReilly, 2019 2. Doing Data Science: Straight Talk from the Frontline, 1 st Edition, Cathy O'Neil and Rachel Schutt, O'Reilly, 2013.
E-Resources	1. https://nptel.ac.in/courses/