COURSE CODE	**DS-401**	
COURSE NAME	**INTRODUCTION TO DATA SCIENCE**	
CREDIT HOURS	Theory:	02
	Practical:	01
	Total:	03
CONTACT HOURS	Theory:	32
	Practical:	48
	Total:	80
PREREQUISITE	Nil	

MODE OF TEACHING:

Instruction:	Two hours of lecture per week	67%
Practical:	Three hours of Lab work per week	33%

COURSE DESCRIPTION:

Data Science is the study of the generalizable extraction of knowledge from data. Being a data scientist requires an integrated skill set spanning mathematics, statistics, machine learning, databases and other branches of computer science along with a good understanding of the craft of problem formulation to engineer effective solutions. This course will introduce students to this rapidly growing field and equip them with some of its basic principles and tools as well as its general mindset. Students will learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration, exploratory data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication. The focus in the treatment of these topics will be on breadth, rather than depth, and emphasis will be placed on integration and synthesis of concepts and their application to solving problems. To make the learning contextual, real datasets from a variety of disciplines will be used.

COURSE OBJECTIVES:

The students are expected to achieve understand.

a. Describe what Data Science is and the skill sets needed to be a data scientist.
b. Explain in basic terms what Statistical Inference means. Identify probability distributions commonly used as foundations for statistical modeling. Fit a model to data.
c. R/Python to carry out basic statistical modeling and analysis.
d. Explain the significance of exploratory data analysis (EDA) in data science. Apply basic tools (plots, graphs, summary statistics) to carry out EDA.
e. Describe the Data Science Process and how its components interact.
f. Use APIs and other tools to scrap the Web and collect data.
g. Apply EDA and the Data Science process in a case study.

RELEVANT PROGRAM LEARNING OUTCOMES (PLOs):

The course is designed so that students will achieve the PLOs:

1	Engineering Knowledge:	☐	7	Ethics:	☐
2	Problem Analysis:	☑	8	Individual and Collaborative Team Work:	☐
3	Design/Development of Solutions:	☑	9	Communication:	☐
4	Investigation:	☐	10	Project Management:	☐
5	Tool Usage:	☑	11	Lifelong Learning:	☐
6	The Engineer and Society:	☐			☐

COURSE LEARNING OUTCOMES:

Upon successful completion of the course, students will be able to:

Sr. No	CLO	Domain	Taxonomy Level	PLO
1	Understand and explain the key concepts, processes, and methodologies used in data science, including data collection, preprocessing, analysis, and interpretation.	Cognitive	2	2
2	Apply Python programming skills and data science libraries such as Pandas, NumPy, and Matplotlib to clean, manipulate, and analyze datasets.	Cognitive	3	5
3	Develop and implement basic machine learning models (e.g., regression, classification) to make predictions from data and evaluate model performance using appropriate metrics.	Cognitive	4	3
4	Display advanced proficiency in executing the design and implementation of Data Science modern tools usage.	Psychomotor	4	5

TOPICS COVERED:

Theory:

No.	Topic
1	Introduction: What is Data Science? (1) Big Data and Data Science hype - and getting past the hype (2) Why now? (3) Current landscape of perspectives (4) Skill sets needed
2	Statistical Inference: (1) Populations and samples (2) Statistical modeling, probability distributions, fitting a model (3) Intro to R
3	Exploratory Data Analysis and the Data Science Process (1) Basic tools (plots, graphs and summary statistics) of EDA (2) Philosophy of EDA (3) The Data Science Process (4) Case Study
4	Three Basic Machine Learning Algorithms (1) Linear Regression (2) k-Nearest Neighbors (k-NN) (3) k-means
5	One More Machine Learning Algorithm and Usage in Applications (1) Motivating application: Filtering Spam (2) Why Linear Regression and k-NN are poor choices for Filtering Spam

	(3) Naive Bayes and why it works for Filtering Spam (4) Data Wrangling: APIs and other tools for scrapping the Web
6	Feature Generation and Feature Selection (Extracting Meaning From Data) (1) Motivating application: user (customer) retention (2) Feature Generation (brainstorming, role of domain expertise, and place for imagination) (3) Feature Selection algorithm - Filters; Wrappers; Decision Trees; Random Forests
7	Recommendation Systems: Building a User-Facing Data Product (1) Algorithmic ingredients of a Recommendation Engine (2) Dimensionality Reduction (3) Singular Value Decomposition
8	WP No.20-50th ACM 28th Dec 2017 Principal Component Analysis Exercise: build your own recommendation system
9	Mining Social-Network Graphs Social networks as graphs
10	Clustering of graphs Direct discovery of communities in graphs
11	Partitioning of graphs Neighborhood properties in graphs
12	Partitioning of graphs Neighborhood properties in graphs
13	Data Visualization Basic principles, ideas and tools for data visualization
14	Examples of inspiring (industry) projects, Exercise: create your own visualization of a complex dataset
15	Data Science and Ethical Issues Discussions on privacy, security, ethics
16	A look back at Data Science
17	Next-generation data scientists
18	**ESE**

Practicals:

Sr. No.	Labs
1.	Python programming language basics, focusing on libraries commonly used in data science (e.g., NumPy, Pandas)
2.	Data Wrangling and Cleaning
3.	Exploratory Data Analysis (EDA)
4.	Loading and processing large datasets using Pandas and Dask, sampling techniques, and performance optimization
5.	Create advanced visualizations (e.g., heatmaps, pair plots, bar charts) using Seaborn, Matplotlib, and Plotly. Customize plots with labels, legends, and formatting.
6.	Train simple machine learning models (e.g., linear regression, classification) using the Scikit-learn library.
7.	Implement linear regression, evaluate models using metrics such as R^2 and RMSE, and visualize regression lines

TEXT AND MATERIAL:

a. Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk From The Frontline. O'Reilly. 2014.
b. Foster Provost and Tom Fawcett. Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking. ISBN 1449361323. 2013.

ASSESMENT SYSTEM:

1. **CLOs Assessment**

Cognitive	**Psychomotor**	**Affective**
Spreadsheet	Rubrics	-

2. **Relative Grading**

Theoretical/Instruction			67%
	Assignments 10-15%		
	Quizzes 10-15%		
	Mid Semester Exam 25-30%		
	End Semester Exam 40-50%		
Practical Work			33%
Laboratory Work		*80%*	
	Laboratory Report / Rubrics 60%		
	Laboratory Quiz20%		
Viva/Quiz		*20%*	
Total			**100%**