Introduction to Data Science	Code & No:	CS 470
	Credits:	3(3+1+0)
	Pre-requisite:	STAT102
	Co-requisite:	
	Level:	9

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. 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.

Course Aims:

1) To cover the basics of data science
2) To give overview of statistical parameters used in data science
3) To demonstrate how to implement various machine learning algorithms for data analysis
4) To implement several machine learning algorithms in R

Course Learning Outcomes (CLOs):

1. Identify probability distributions commonly used as foundations for statistical modeling.
2. Apply basic tools (plots, graphs, summary statistics) to carry out Exploratory Data Analysis.
3. Apply basic machine learning algorithms (Linear Regression, k-Nearest Neighbors (k-NN), k-means, Naive Bayes) for predictive modeling.
4. Identify basic Feature Selection algorithms Decision Trees, Random Forests and use in applications
5. Use R language to carry out basic statistical modeling and analysis.

No.	Topics	Weeks	Teaching hours
1	Introduction: What is Data Science? - Big Data and Data Science hype - Current landscape of perspectives - Skill sets needed	1	3

2	Statistical Inference - Populations and samples - Statistical modeling, probability distributions, fitting a model - Intro to R	2	6
3	Exploratory Data Analysis and the Data Science Process - Basic tools (plots, graphs and summary statistics) of EDA	2	6
4	Three Basic Machine Learning Algorithms - Linear Regression - k-Nearest Neighbors (k-NN) - k-means	2	6
5	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	1	3
6	Feature Generation and Feature Selection (Extracting Meaning From Data) - Feature Selection algorithms-Decision Trees; Random Forests	2	6
7	Algorithmic ingredients of a Recommendation Engine - Dimensionality Reduction - Singular Value Decomposition - Principal Component Analysis	2	6
8	Data Visualization - Basic principles, ideas and tools for data visualization	1	3
9	Data Science and Ethical Issues - Discussions on privacy, security, ethics	1	3
	Total	**14**	**42**

Textbook:

- Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk From The Frontline. O'Reilly. 2014

Essential References:

- Jure Leskovek, Anand Rajaraman and Jeffrey Ullman. Mining of Massive Datasets. v2.1, Cambridge University Press. 2014.
- Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. ISBN 0262018020. 2013
- NINA ZUMEL, JOHN MOUNT, Practical Data Science with R, Manning Publications Co.,2014