Essentials of Data Science

Course Code: CS5113

Credits: **3-0-0**

Course Pre-requisites: **Probability, Statistics, Mathematics**

Data Science is a broad and interdisciplinary field that focuses on how we collect, manage, analyze, and visualize data to understand it better and make useful decisions. This course offers an introduction to Data Science. It is helpful for students who want to explore areas like Machine Learning, Data Management, and Statistics in more depth as well as for anyone who simply wants to learn how to work with data and use these skills in their own field of study or area of interest.

Course Outcomes:

After successful completion of course, students will be able to:

CO1: Explain the fundamental concepts, terminology, and applications of Data Science and identify appropriate data sources for solving real-world problems.

CO2: Acquire, clean, preprocess, and organize structured and unstructured data using appropriate data management techniques.

CO3: Apply statistical and exploratory data analysis techniques to identify patterns, relationships, and trends in datasets.

CO4: Develop and execute appropriate data queries and computational techniques to extract meaningful information from datasets.

CO5: Create suitable data visualizations and interpret analytical results to communicate insights effectively.

CO6: Apply the fundamental concepts of Data Science to formulate and analyze real-world problems and select appropriate analytical or machine-learning approaches.

Course Content:

Unit I: Introduction to data analytics

Introduction to Data Science, Representing Data, Data Models, collecting data, storing data, processing data, describing data, Introduction to data analytics, Types of Data Analytics. Data Visualization. Massive Data handling, Data streaming, Data sampling, Data warehousing.

Unit II: Basic Statistical learning and techniques

Statistical Foundations for Data Science, Sampling and Surveys, population, sampling methods. Introduction to statistical learning, statistical modelling, Algorithm modeling, Descriptive Statistics, Descriptive statistics with graphical approaches, Descriptive Statistics with measure of central tendency, Descriptive statistics with measure of location of dispersions

Unit III: **Probability and Inferential Statistics**

Probability Distribution and Inferential Statistics, Random variable and probability distributions, Inferential statistics – motivation, Inferential statistics- single sample tests. Bayesian inference. Basic analysis techniques, Statistical hypothesis generation and testing, Two sample tests, Type I and Type II errors, Confidence Intervals, chi-square test, t-test, Analysis of variance, correlation analysis, Maximum likelihood test, ANOVA and test of independence

Unit IV: Machine learning and causality

Introduction to machine learning, Supervised Learning, Unsupervised learning, Introduction to regression, ordinary least square regression, Simple and multiple regression, Regularization/ coefficient shrinkage. Singular Value Decomposition, PCA, random walks, Markov Chains. Causality, The Rubin causal model, potential outcomes, and counterfactuals. The basics of causal inference from observational data. From causal inference to data-driven decisions.

Grading criteria:

Quiz/In-class activity- 20%

Midsem- 20%

Endterm- 30%

Projects- 30%

Resources:

1. All content (slides, assignments, quiz, suggested readings, research papers, code, etc.) will be shared during the course.

Suggested readings:

1. Dennis Sun, Principles of Data Science, https://github.com/dlsun/pods

2. https://github.com/LearnDataSci/free-data-science-learning/blob/master/free-data-science-books.md

3. An Introduction to Statistical learning with application in Python, Gareth James et al.

https://www.statlearning.com/