CAP787:DATA SCIENCE TOOLBOX

Course Outcomes: Through this course students should be able to

CO1 :: Understand data science and various Toolbox used for it

CO2 :: Use R and R studio as data science toolbox

CO3 :: Use Python for data analysis data science toolbox

CO4 :: Integrate Git and GitHub for version control which plays important role in data science

Unit I

Introduction to data and Data Science : • What is Data Science? • What is Data? • Getting Help • The Data Science Process • Types of Data Science Questions • Experimental Design • Download data from different sources and explore them.

Unit II

R and R studio : • Installing R • Installing R Studio • R Studio Tour • Loading data and basic analysis of data • R Packages • Projects in R • R Markdown

Unit III

Basic Data Analysis Using Python : • Introduction: Understanding the Dataset, Python package for data science, Importing and Exporting Data in Python, Basic Insights from Datasets. • Data Wrangling: Identify and Handle Missing Values, Data Formatting, Data Normalization, Sets, , Binning, Indicator variables, • Exploratory Data Analysis: Descriptive Statistics, Basic of Grouping, ANOVA, Correlation, Correlation 2.

Unit IV

Python Basics : • Python Data structure • Programming Fundamental • Working with data in python • Working with numpy Array

Unit V

Advance Data Analysis Using Python : • Model Development: Simple and Multiple Linear Regression, Model Evaluation using Visualization, Polynomial Regression and Pipelines, R-squared and MSE for In-Sample Evaluation, Prediction and Decision Making, • Working with Data in Python: Model Evaluation , Over Fitting, Under fitting and Model Selection , Ridge Regression, Grid Search , Model Refinement.

Unit VI

Version Control and Github : • Version Control • Github and Git • Linking Github and R Studio • Linking Github and python • Projects under Version Control

Text Books:

1. PYTHON FOR DATA ANALYSIS by WES MCKENNY, O'REILLY

References:

1. R FOR DATA SCIENCE: IMPORT, TIDY, TRANSFORM, VISUALIZE, AND MODEL DATA by HADLEY WICKHAM, O'REILLY
2. GITHUB ESSENTIALS by ACHILLEAS PIPINELLIS, PACKT PUBLISHING