

Course code	CC-A11			
Category	Core Subject			
Course title	Data science and Visualization			
Course ID	241/MCA/CC402			
Scheme and Credits	L	T	P	Credits
	3	-	2	4
Theory Internal	25			
Theory External	50			
Practical Internal	05			
Practical External	20			
Total	100			
Duration of Exam	3 hrs.			

Note: The examiner will set nine questions in total. Question one will have seven parts from all units and the marks of first question will be 20% of total marks of Question Paper and the remaining eight questions to be set by taking two questions from each unit and the marks of each question from Question no. 2 to 9 will be 20% of total marks of Question paper. The students have to attempt five questions in total, the first being compulsory and selecting one from each unit.

Course Outcomes:

CO1: Understand and implement the basics of programming in Python.

CO2: Apply the Numpy package for numerical calculations in Python.

CO3: Apply Pandas package for loading and preprocessing data in Python. Implement various data visualization tools of Python on real world data.

CO4: Understand and implement the Machine Learning Concepts in Python

Unit I

Overview of Python Programming Concepts: The concept of data types; variables, assignments; numerical types; operators and expressions; Control Structures; String manipulations; File Handling – creating, reading/writing text/number files; Dictionaries; Functions; OOPs Concepts

Unit II

Introduction to Numpy: Creation on Array, Array generation from Uniform distribution, Random array generation, reshaping, maximum and minimum, reshaping, Arithmetic operations, Mathematical functions, Bracket Indexing and Selection, Broadcasting, Indexing a 2D array (matrices); Data Manipulation with Pandas -Creating a Series - from lists, arrays and dictionaries, Storing data in series from intrinsic sources, Creating Data Frames, Imputation, Grouping and aggregation, Merging, Joining, Concatenation, Find Null Values or Check for Null Values, Reading data from csv, txt, excel, web.

Unit III

Introduction to Visualization - Installing and setting up visualization libraries, Canvas and Axes, Subplots, Common plots – scatter, histogram, boxplot, Logarithmic scale, Placement of ticks and custom tick labels, Pandas Viz, Style Sheets, Plot type, Area, Barplots, Histograms, Line Plots, Scatter Plots, BoxPlots, Hexagonal Bin Plot, Kernel Density Estimation plot (KDE), Distribution Plots, Categorical Data Plots, Combining Categorical Plots, Matrix Plots, Regression Plots, Grids; Python Visualizations toolkits/libraries.

Unit IV

Introduction to Machine Learning with SciKit-Learn & PyTorch– Data Representation and basic functions Estimator, parameters & model validation, Model Selection, Curve, Grid search, Feature engineering, Naive Bayes Classification, Linear regression, SVM etc; Overview of other Python ML/Deep Learning toolkits/Libraries. Introduction to NLP with NLTK and its functions, modules like speech tagging, tokenization, parsing, segmentation, recognition, cleaning & normalization of text etc; Overview of other Python NLP toolkits/Libraries.

Textbooks & References:

1. Charles Dierbach., Introduction to Python using Computer Science, Wiley Publications, Second Edition, 2015
2. Mark Lutz , Learning Python, O'Reilly publications , Fifth Edition, 2015
3. Jake VanderPlas, Python Data Science Handbook, O'Reilly , 2016
4. Paul Barry, Head First Python, Orielly Publications, Second Edition, 2010