Name of the Programme: MCA
Course Code: CSA-604
Title of the Course: Data science
Number of Credits: 4 (2L-2T-0P)
Effective from AY: 2022-23

Prerequisites for the course	Statistics and probability theory and python programming. Python programming and Data science theory fundamentals.	
Objectives	To get started with basics of Data Science and learn all aspects of data science in its entirety. Main objectives are as under - ● to understand basic process of data science ● Python and Jupyter notebooks ● An applied understanding of how to manipulate and analyze uncurated datasets ● Basic statistical analysis and basic machine learning methods like linear regression . ● How to effectively visualize results using python APIs or tools.	
Content	**Unit -1:** Basics of Data Science: Introduction; Typology of problems-**Data Science in a big data world:** Benefits and uses of data science and big data-Facets of data-The data science process-The big data ecosystem and data science-**The data science process:** Overview of the data science process- Defining research goals and creating a project charter- Retrieving data-Cleansing, integrating, and transforming data-Exploratory data analysis-Build the models-Presenting findings and building applications on top of them. **Unit -2** Mathematics for Data science ● Importance of linear algebra, statistics and optimization from a data science perspective; Structured thinking for solving data science problems. ● Linear Algebra: Matrices and their properties (determinants, traces, rank, nullity, etc.); Eigenvalues and eigenvectors; Matrix factorizations; Inner products; Distance measures; Projections; Notion of hyperplanes; half-planes. ● Probability, Statistics and Random Processes: Probability theory and axioms; Random variables; Probability distributions and density functions (univariate and multivariate); Expectations and moments; Covariance and correlation; Statistics and sampling distributions; Hypothesis testing of means, proportions, variances and correlations; Confidence (statistical) intervals; Correlation functions; White-noise process. **Unit -3** Introduction to Data Science Methods: Linear regression as an exemplar function approximation problem; Linear classification problems. **Unit -4** Handling large data on a single computer ● The problems you face when handling large data-General techniques for handling large volumes of data-General programming tips for dealing with large data sets-Case study 1: Predicting malicious URLs-**First steps in big data-** Distributing data storage and processing with frameworks **Unit 5:** Join the NoSQL movement-Introduction to NoSQL **Unit 6: The rise of graph databases** ● Introducing connected data and graph databases ● Introducing Neo4j: a graph database **Unit 7: Data visualization to the end user**	4 hours 2 hours 2 hours 2 hours 4 hours 4 hours 4 hours

	● Data visualization options ● Crossfilter, the JavaScript MapReduce library ● Creating an interactive dashboard with dc.js ● Dashboard development tools ● Data science Story telling.	4 hours 4 hours
	Assignments to be discussed during the Tutorial slots -	30 hours
	1. Python libraries – Numpy, Matplotlib, seaborn, pandas. 2. Write program to do Exploratory data analysis using the libraries above.- Data collection(Kaggle, github and Machine learning repository),data cleaning (removing missing values, reformatting data etc. 3. Write program to do univariate analysis using tools like Box plot, histogram etc. 4. Write program to do bivariate analysis using tools like scatter plots, box plots. 5. Demo on business intelligence tools -Business intelligence tools help an organization analyze huge chunks of data; they provide insights with actionable recommendations - Tableau, Qlik,splunk,Trillium,Logi analytics, powerBI 6. Write program to implement PCA. 7. Write program to implement SVD 8. Use tools like tableau/Power BI to do Visualizatiation for large data set and create dashboard 9. Mini Project: With the tools of Jupyter notebooks, numpy, pandas, and Visualization, you're ready to do sophisticated analysis on your own. You'll pick a dataset we've worked with already and perform an analysis for this first project 10. Machine Learning: To take your data analysis skills one step further, write program to do basics of machine learning and how to use sci-kit learn - a powerful library for machine learning. 11. Working with Text and Databases: You'll find yourself often working with text data or data from databases. This week will give you the skills to access that data. For text data, we'll also give you a preview of how to analyze text data using ideas from the field of Natural Language Processing and how to apply those ideas using the Natural Language Processing Toolkit (NLTK) library. 12. Final Project: These weeks let you showcase all your new skills in an end-to-end data analysis project. You'll pick the dataset, do the data munging, ask the research questions, visualize the data, draw conclusions, and present your results.	
Pedagogy	Lectures/ Tutorials/Hands-on assignments/Self-study. Lab assignments/ research paper reading/ discussion/ tools demonstration/ mini project.	
References/ Readings	1. Practical statistics for data science by peter bruce and andrew bruce 2. Naked statistics by charles wheelon 3. Business data science by matt taddy 4. Elements of statistical learning by Trevor Hastie, Robert and jerome 5. Python for data analysis 6. Data science and big data analytics -EMC2 7. Hands-On Data Structures and Algorithms with Python — By Dr. Basant Agarwal. 8. 3. The Art of Data Science — by Roger D. Peng and Elizabeth	

	Matsui. 9. . Automate the Boring Stuff With Python: Practical Programming— by Al Sweigart.	
Course Outcomes	At the end of the course, the students will – 1. Enrich one's knowledge with overall basics of data science 2. appreciate Data Science to be able to get started in the direction. 3. Students should be able to carry out mini Data Science projects using python libraries.	

(Back to top)