Name of the Programme : **M.Sc. in Data Science**
Course code : **CSD-501**
Title of the course : **Fundamentals of Data Science (Practical)**
Number of credits : **2(0L-0T-2P)**
Total contact hours : **60 hours (0L-0T-60P)**
Effective from AY : **2023-24**

Pre-requisites for the course	Basic programming skills, Statistics
Course Objectives	The course aims to provide an introduction to the fundamental processes of data science using Python and Jupyter notebooks, enabling participants to manipulate and analyze uncurated datasets, apply basic statistical analysis and machine learning methods, and effectively visualize the results.

Content	1. Create a Jupyter notebook and import the numpy library. Generate a 2D numpy array of size 5x5 filled with random integers between 1 and 100. Perform the following operations: a. Calculate the mean and standard deviation of the array. b. Find the sum of all elements in the array. c. Reshape the array into a 1D array and compute the median. 2. Download a dataset from an online source (e.g., Kaggle or UCI Machine Learning Repository) and load it into a pandas DataFrame. Perform the following tasks: a. Display the first five rows of the dataset. b. Check for missing values and handle them appropriately. c. Calculate summary statistics for numerical columns. d. Plot a histogram of one of the numerical variables. 3. Download a messy dataset containing missing values, duplicates, and inconsistent formatting. Use pandas to clean and prepare the data by: a. Handling missing values through imputation or removal. b. Identifying and removing duplicate entries. c. Standardizing formatting across columns (e.g., converting strings to lowercase). 4. Choose a dataset of your interest and create visualizations to explore its characteristics. Tasks include: a. Plotting a line chart to visualize the trend of a numerical variable over time (if applicable). b. Creating a scatter plot to examine the relationship between two numerical variables. c. Generating a bar chart or pie chart to display categorical data.	**60 hours**

	5. Apply machine learning techniques to analyze a dataset and make predictions. Tasks include: a. Preprocessing the data using numpy and pandas (e.g., handling missing values, encoding categorical variables). b. Splitting the dataset into training and testing sets.	

	c. Building and training a machine learning model using scikit-learn. d. Evaluating the model's performance using appropriate metrics (e.g., accuracy, precision, recall). 6. Select a dataset suitable for a classification or regression task. Apply machine learning techniques using scikit-learn to build and evaluate a predictive model. Requirements: a. Preprocess the data, including feature scaling and handling categorical variables. b. Split the dataset into training and testing sets. c. Choose an appropriate algorithm (e.g., decision tree, logistic regression) and train the model. d. Evaluate the model's performance using relevant metrics (e.g., accuracy, precision, recall). 7. Access a text dataset (e.g., movie reviews, news articles) and perform basic text analysis using NLTK. Requirements: a. Preprocess the text data by tokenizing, removing stopwords, b. and stemming or lemmatizing. c. Analyze the frequency of words and visualize the most common terms using word clouds or bar charts. d. Apply sentiment analysis to categorize the text into positive, negative, or neutral sentiments. 8. Connect to a database (e.g., SQLite, MySQL) using Python and perform basic operations. Requirements: a. Establish a connection to the database and retrieve data from one or more tables. b. Execute CRUD operations (Create, Read, Update, Delete) on the database using SQL queries or Python libraries (e.g., SQLAlchemy). c. Perform simple data analysis or visualization on the retrieved data. 9. Choose a dataset of interest and perform an end-to-end data analysis project, showcasing all your skills.	
Pedagogy	Tutorials/ Lab assignments/ Project work	

References/ Readings	1. Baesens, B. (2014). Analytics in a big data world: The essential guide to data science and its applications. John Wiley & Sons 2. Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical statistics for data scientists: 50+ essential concepts using R and Python. O'Reilly Media. 3. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer. 4. McKinney, W. (2022). Python for data analysis. " O'Reilly Media, Inc.". 5. Taddy, M. (2019). Business data science.
	6. Wheelan, C. Naked Statistics: Stripping the Dread from the Data.
Course Outcomes	1. Practical data analysis skills using data science tools. 2. Hands-on experience with real-world data projects. 3. Collaboration and teamwork in interdisciplinary settings. 4. Ethical considerations and responsible practices in data science 5. Experimentation and evaluation of data science techniques.

(Back to Index)