SEMESTER-V

COURSE 13 B: PYTHON FOR DATA SCIENCE

Theory **Credits: 3** **3 hrs/week**

Course Objectives:

1. Introduce foundational concepts of NumPy arrays and array operations for efficient numerical computing.
2. Teach key data structures and manipulation techniques using Pandas.
3. Enable students to perform data input/output operations and implement basic data cleaning workflows.
4. Explore string processing methods and feature engineering strategies in Pandas.
5. Guide learners in advanced data wrangling tasks including merging, reshaping, and hierarchical indexing.

Course Outcomes:

1. Demonstrate proficiency in creating and manipulating NumPy arrays for mathematical operations and simulations.
2. Apply Pandas Series and DataFrame operations for structured data handling and analysis.
3. Read, write, and clean diverse data formats using Python tools, addressing missing values and outliers.
4. Implement vectorized string operations and create derived features for enhanced model readiness.
5. Perform complex data wrangling tasks such as merging datasets, reshaping data structures, and generating group-level statistics.

Unit 1. NumPy Essentials:
NumPy ndarray: A Multidimensional Array Object, Creating ndarrays, Data Types for ndarrays, Arithmetic with Arrays, Basic Indexing and Slicing, Boolean Indexing, Fancy Indexing, Transposing Arrays, Swapping Axes, Universal Functions: Element-wise Operations, Basic Mathematical and Statistical Functions, Random Number Generation (basic use)

Unit 2. Pandas Basics and Data Structures:
Series, DataFrame, Index objects, Indexing and Selection, Filtering and Boolean Indexing, Arithmetic and Data Alignment, Sorting and Ranking, Dropping Entries, Handling Duplicate Indexes

Unit 3. Data Input, Output, and Cleaning:
Reading and Writing Data in Text Format (CSV, TXT), Working with JSON, Reading Microsoft Excel Files, Handling Missing Data, Dropping and Filling Missing Values, Replacing Values, Renaming Axis Indexes, Removing Duplicates, Filtering Outliers, Transforming Data Using Mapping or Functions

Unit 4. String Operations and Feature Engineering:
String Methods in pandas, Basic Regular Expressions, Vectorized String Functions, Creating Dummy/Indicator Variables, Permutation and Random Sampling.

Unit 5. Data Wrangling and Reshaping:
Merging and Joining Datasets, Concatenating Along an Axis, Combining Data with Overlap, Reshaping with Pivot, Stack, and Unstack, Basic Hierarchical Indexing, Summary Statistics by Group or Level

Textbooks

1. Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter, Wes McKinney, 3rd Edition, O'Reilly Media, 2022.
2. Python for Data Science For Dummies, Yuli Vasiliev, 2nd Edition, Wiley, 2022.

Reference Books

1. Python Data Science Handbook: Essential Tools for Working with Data, Jake VanderPlas, 2nd Edition, O'Reilly, 2022.

2. Introduction to Machine Learning with Python, Andreas Müller & Sarah Guido, O'Reilly Media, Reprint Edition, 2023.

3. Foundations for Analytics with Python: From Non-programmer to Hacker, Clinton Brownley, 2nd Edition, Pearson, 2020.

Activities:

Outcome: Demonstrate proficiency in creating and manipulating NumPy arrays for mathematical operations and simulations.

Activity: Create a NumPy-based simulation:

- Generate a 2D array representing temperature data over time
- Apply mathematical operations (mean, std, element-wise addition)
- Simulate random noise and visualize the effect

Evaluation Method: Code-based assessment (10-point scale):

- Correct use of np.array, np.random, and math functions
- Accuracy of simulation logic
- Output clarity and reproducibility

Outcome: Apply Pandas Series and DataFrame operations for structured data handling and analysis.

Activity: Analyze a CSV dataset (e.g., sales or COVID data):

- Load into a DataFrame
- Perform Series operations (filter, map, value_counts)

- Apply DataFrame methods (groupby, sort, describe)

Evaluation Method: 10-point scale checklist and peer review to verify:

- Proper use of Series vs DataFrame
- Logical data manipulation
- Insightful summary statistics

Outcome: Read, write, and clean diverse data formats using Python tools, addressing missing values and outliers.

Activity: Work with multiple formats:

- Read CSV, Excel, and JSON files
- Identify and handle missing values (dropna, fillna)
- Detect and treat outliers using IQR or Z-score

Evaluation Method: Rubric-based evaluation to check (10-point scale):

- File handling accuracy
- Cleaning completeness
- Outlier detection logic

Outcome: Implement vectorized string operations and create derived features for enhanced model readiness.

Activity: Prepare text data for modelling to:

- Use str methods to clean and standardize strings
- Extract features (e.g., domain from email, length of name)
- Encode categorical variables (e.g., get_dummies, LabelEncoder)

Evaluation Method: Feature report to check (10-point scale):

- Efficiency of vectorized operations
- Relevance of derived features
- Readiness for ML input

Outcome: Perform complex data wrangling tasks such as merging datasets, reshaping data structures, and generating group-level statistics.

Activity: Integrate and reshape datasets:

- Merge two datasets on a common key
- Reshape using pivot, melt, stack, unstack
- Generate group-level stats (e.g., mean sales per region)

Evaluation Method: Before-and-after comparison to validate:

- Accuracy of merge and reshape
- Correct use of aggregation
- Final structure suitability for analysis

SEMESTER-V

COURSE 13 B: PYTHON FOR DATA SCIENCE

Practical **Credits: 1** **2 hrs/week**

List of Practicals:

1. Create and Manipulate NumPy ndarrays; Explore Data Types
2. Perform Arithmetic Operations and Element-wise Calculations on Arrays
3. Practice Indexing, Slicing, Boolean, and Fancy Indexing on ndarrays
4. Use Universal Functions and Compute Basic Mathematical/Statistical Functions with NumPy
5. Create and Manipulate Pandas Series and DataFrames
6. Perform Indexing, Selection, Filtering, and Boolean Indexing in Pandas
7. Conduct Arithmetic Operations and Data Alignment in DataFrames
8. Sort, Rank, Drop Entries and Handle Duplicate Indexes in Pandas
9. Read and Write Data in CSV, TXT, JSON, and Excel Formats
10. Handle Missing Data: Detect, Drop, Fill, and Replace Missing Values
11. Rename Axis Indexes, Remove Duplicates, and Filter Outliers
12. Transform Data Using Mapping Functions and Apply String Operations
13. Perform String Operations and Use Regular Expressions on DataFrames
14. Create Dummy Variables and Perform Permutations and Random Sampling
15. Merge, Join, and Concatenate Datasets Using Pandas
16. Reshape Data Using Pivot, Stack, Unstack, and Perform Hierarchical Indexing
17. Compute Summary Statistics Grouped by Levels or Categories