III B.Tech I Semester

23A39501	**DATA WRANGLING AND PRE-PROCESSING (Professional Core)**	**L**	**T**	**P**	**C**
		3	**0**	**0**	**3**

Course Objectives (COs)

- To introduce students to data wrangling techniques using Python and other tools.
- To familiarize students with various data formats such as CSV, JSON, XML, and databases.
- To enable students to clean and preprocess data by handling missing values, duplicates, outliers, and normalization.
- To equip students with skills in data exploration, visualization, and transformation for analysis.
- To provide practical knowledge of web scraping techniques and data acquisition from various sources.

Course Outcomes

- Demonstrate proficiency in data wrangling techniques for structured and unstructured data.
- Apply data extraction and transformationtechniques on various file formats (CSV, JSON, XML, Excel, PDF).
- Perform data cleaning operations, including handling missing values, outlier detection, duplicates, and normalization.
- Analyze datasets by performing exploratory data analysis (EDA) using visualization tools.
- Develop web scraping scripts using Python libraries such as Scrapy, BeautifulSoup, Selenium to gather real-time data.

UNIT - I: Introduction to Data Wrangling

What Is Data Wrangling, Importance of Data Wrangling, Tasks of Data Wrangling, Data Wrangling Tools, Introduction to Python for Data Wrangling, Python Basics for Data Wrangling, Handling Structured Data: CSV, JSON, and XML Formats, Data Meant to Be Read by Machines

UNIT - II: Working with Excel Files, PDFs, and Databases

Installing Python Packages for Data Wrangling, Parsing Excel Files, Programmatic Approaches to PDF Parsing, Converting PDF to Text (pdfminer), Acquiring and Storing Data, Introduction to Databases for Data Wrangling, Relational Databases: MySQL and PostgreSQL, Non-Relational Databases: NoSQL and Alternative Data Storage

UNIT - III: Data Cleaning and Exploration

Why Clean Data? Basics of Data Cleanup, Identifying and Formatting Data for Clean-Up, Finding Outliers and Bad Data, Removing Duplicates and Fuzzy Matching, Using Regular Expressions (RegEx) for Data Cleaning, Normalization and Standardization of Data, Saving Cleaned Data and Testing with New Data, Data Exploration: Table Functions and Joining Datasets

UNIT - IV: Data Preprocessing and Reduction

Data Quality: Why Preprocess Data?, Major Tasks in Data Preprocessing, Handling Missing Values in Data, Identifying and Removing Noisy Data, Data Integration and Entity Identification Problem,

Redundancy and Correlation Analysis in Data, Detection and Resolution of Data Conflicts, Tuple Duplication and Its Impact

UNIT - V: Data Transformation and Web Scraping

Overview of Data Transformation Strategies, Normalization and Standardization, Discretization by Binning and Histogram Analysis, Clustering, Sampling, and Data Cube Aggregation, Web Scraping: What to Scrape and How, Analyzing and Parsing Web Pages with LXML and XPath, Advanced Web Scraping Using Selenium and Scrapy.

Textbooks (Core Learning Materials)

1. Data Wrangling with Python: Tips and Tools to Make Your Life Easier – Dr. Jacqueline Kazil and Katharine Jarmul, O'Reilly Media.
2. Data Preprocessing for Machine Learning in Python" – M.G. Sumithra, CRC Press.

Reference Books (Supplementary Learning)

1. Web Scraping with Python: Collecting More Data from the Modern Web" – Ryan Mitchell, O'Reilly Media.
2. Data Cleaning and Exploration with Machine Learning" – Michael Walker, Packt Publishing.