CSR302:DATA EXPLORATION AND PREPARATION

L:2 T:0 P:2 Credits:3

Course Outcomes: Through this course students should be able to

CO1 :: apply the fundamental concepts and importance of data science, including the data science lifecycle and key steps in data exploration and preparation.

CO2 :: identify skills to effectively collect and integrate data from various sources, including structured, unstructured, and semi-structured data, using techniques such as APIs and web scraping

CO3 :: identify common data quality issues by applying techniques for handling missing data, detecting and treating outliers, and performing data transformation and normalization.

CO4 :: explain transform and normalize data using techniques such as scaling, data encoding and transformation.

CO5 :: build comprehensive exploratory data analysis (EDA) to summarize the main characteristics of the data using descriptive statistics and data visualization techniques

CO6 :: illustrate how to implement feature engineering techniques to create, select, and transform features to improve the performance of predictive models.

Unit I

Introduction to Data Exploration and Preparation : Definition and Importance of Data Science, The Data Science Lifecycle, Roles in Data Science (Data Analyst, Data Scientist, Data Engineer), Data Exploration and Preparation Overview - Goals and Objectives, Key Steps, Importance of Data Quality, Tools and Technologies – Python, Pandas, NumPy, Matplotlib, Seaborn, IDE

Unit II

Data Collection and Integration : Data Sources and Collection Methods, Types of Data (Structured, Unstructured, Semi-structured), Data Collection Techniques (Surveys, Databases, APIs), Data Integration - Merging and Joining Data, Handling Different Data Formats (CSV, JSON, SQL), Web Scraping - Basics of Web Scraping with BeautifulSoup and Scrapy

Unit III

Data Cleaning : Introduction and Importance, Common Data Quality Issues (Missing Values, Outliers, Duplicates), Handling Missing Data - Identifying Missing Data Patterns, Techniques (Imputation, Deletion), Using Pandas for Missing Data Handling, Dealing with Outliers – Detection (IQR, Z-Score), Treatment Methods, Visualizing Outliers with Box Plots

Unit IV

Data Transformation and Normalization : Scaling and Normalization Techniques (Min-Max Scaling, Standardization), Data Encoding (Label Encoding, One-Hot Encoding), Log Transformation and Power Transformation

Unit V

Exploratory Data Analysis (EDA) : Introduction, Goals and Importance, Steps, Workflow and Best Practices, Descriptive Statistics - Measures of Central Tendency (Mean, Median, Mode), Measures of Dispersion (Variance, Standard Deviation, Range), Distribution Analysis (Skewness, Kurtosis), Plotting Techniques (Histograms, Box Plots, Scatter Plots), Using Matplotlib and Seaborn for Visualization, Advanced Visualization Techniques - Correlation Heatmaps, Pair Plots and Facet Grids, Time Series Plots and Geographical Plots

Unit VI

Feature Engineering : Introduction & Importance, Types of Features (Numerical, Categorical, Date-Time), Creating New Features - Deriving Features from Existing Data, Feature Extraction from Dates (Day, Month, Year), Feature Selection – Importance, Techniques, Dimensionality Reduction – PCA, LDA

List of Practicals / Experiments:

List of Practicals

- Data Loading and Inspection: Load the Airbnb dataset into Python using Pandas. Inspect the dataset to understand its structure, including the number of rows and columns, data types, and initial observations.

- Data Cleaning:Handle missing values: Identify columns with missing values and decide on appropriate strategies (e.g., imputation, deletion).Remove unnecessary columns that do not contribute to the analysis (e.g., URLs, irrelevant identifiers).
- Exploratory Data Analysis (EDA): Perform summary statistics: Compute descriptive statistics (e.g., mean, median, standard deviation) for numerical columns such as listing price and availability. Visualize distributions: Create histograms and box plots to visualize the distribution of numerical variables and identify outliers.Time series analysis (if applicable): Plot time series trends for variables like booking availability over time.
- Insights and Recommendations: Summarize key findings from the EDA: Identify trends, patterns, and insights discovered from the analysis. Provide actionable recommendations based on the insights: Suggest strategies for pricing optimization, targeting specific customer segments, or improving property listings based on the analysis.

Text Books:

1. FEATURE ENGINEERING FOR MACHINE LEARNING by ALICE ZHENG, AMANDA CASARI, SHROFF/O'REILLY

References:

1. PYTHON DATA SCIENCE HANDBOOK: ESSENTIAL TOOLS FOR WORKING WITH DATA by JAKE VANDERPLAS, SHROFF/O'REILLY
2. WEB SCRAPING WITH PYTHON by RYAN MITCHELL, SHROFF/O'REILLY