Savitribai Phule Pune University
T.Y.B.Sc. (Computer Science) – Sem - V
Course Type:DSEC – II **Course Code: CS - 354**
Paper Title : Foundations of Data Science

Teaching Scheme 03 lectures / week	No. of Credits 2	Examination Scheme IE : 15 marks UE: 35 marks

Prerequisites

- Problem solving using computers
- Basic mathematics and statistics
- Knowledge of Databases

Course Objectives

- Provide students with knowledge and skills for data-intensive problem solving and scientific discovery
- Be prepared with a varied range of expertise in different aspects of data science such as data collection, visualization, processing and modeling of large data sets.
- Acquire good understanding of both the theory and application of applied statistics and computer science based existing data science models to analyze huge data sets originating from diversified application areas.
- Be better trained professionals to cater the growing demand for data scientists in industry.

Course Outcomes

On completion of the course, student will be able to–

- Perform Exploratory Data Analysis
- Obtain, clean/process, and transform data.
- Detect and diagnose common data issues, such as missing values, special values, outliers, inconsistencies, and localization.
- Demonstrate proficiency with statistical analysis of data.
- Present results using data visualization techniques.
- Prepare data for use with a variety of statistical methods and models and recognize how the quality of the data and the means of data collection may affect conclusions.

Course Contents

Chapter 1	**Introduction to Data Science**	**6 lectures**

Introduction to data science, The 3 V’s: Volume, Velocity, Variety
Why learn Data Science?
Applications of Data Science
The Data Science Lifecycle
Data Scientist’s Toolbox
Types of Data

Structured, semi-structured, Unstructured Data, Problems with unstructured data
Data sources
Open Data, Social Media Data, Multimodal Data, standard datasets
Data Formats
Integers, Floats, Text Data, Text Files, Dense Numerical Arrays, Compressed or Archived Data, CSV Files, JSON Files, XML Files, HTML Files , Tar Files, GZip Files, Zip Files, Image Files: Rasterized, Vectorized, and/or Compressed

Chapter 2 **Statistical Data Analysis** **10 lectures**

2.1.Role of statistics in data science

2.2.Descriptive statistics

Measuring the Frequency

Measuring the Central Tendency: Mean, Median, and Mode

Measuring the Dispersion: Range, Standard deviation, Variance, Interquartile Range

2.3.Inferential statistics

Hypothesis testing, Multiple hypothesis testing, Parameter Estimation methods,

2.4.Measuring Data Similarity and Dissimilarity

Data Matrix versus Dissimilarity Matrix, Proximity Measures for Nominal Attributes, Proximity Measures for Binary Attributes, Dissimilarity of Numeric Data: Euclidean, Manhattan, and Minkowski distances, Proximity Measures for Ordinal Attributes

2.5.Concept of Outlier, types of outliers, outlier detection methods

Chapter 3 **Data Preprocessing** **10 lectures**

Data Objects and Attribute Types: What Is an Attribute?, Nominal , Binary, Ordinal Attributes, Numeric Attributes, Discrete versus Continuous Attributes

Data Quality: Why Preprocess the Data?

3.3.Data munging/wrangling operations

Cleaning Data - Missing Values, Noisy Data (Duplicate Entries, Multiple Entries for a Single Entity, Missing Entries, NULLs, Huge Outliers, Out-of-Date Data, Artificial Entries, Irregular Spacings, Formatting Issues - Irregular between Different Tables/Columns, Extra Whitespace, Irregular Capitalization, Inconsistent Delimiters, Irregular NULL Format, Invalid Characters, Incompatible Datetimes)

Data Transformation – Rescaling, Normalizing, Binarizing, Standardizing,Label and One Hot Encoding

Data reduction

Data discretization

Chapter 4 **Data Visualization** **10 lectures**

Introduction to Exploratory Data Analysis

Data visualization and visual encoding

Data visualization libraries

Basic data visualization tools

Histograms, Bar charts/graphs, Scatter plots, Line charts, Area plots, Pie charts, Donut charts

Specialized data visualization tools

Boxplots, Bubble plots, Heat map, Dendrogram, Venn diagram, Treemap, 3D scatter plots

Advanced data visualization tools- Wordclouds

Visualization of geospatial data

Data Visualization types

Reference Books:

1) Data Science Fundamentals and Practical Approaches, Gypsy Nandi, Rupam Sharma, BPB Publications, 2020.
2) The Data Science Handbook, Field Cady, John Wiley & Sons, Inc, 2017
3) Data Mining Concepts and Techniques, Third Edition, Jiawei Han, Micheline

Kamber, Jian Pei, Morgan Kaufmann, 2012.

4) A Hands-On Introduction to Data Science, Chirag Shah, University of Washington Cambridge University Press