SEMESTER-V

COURSE 12 B: DATA SCIENCE WITH R

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

Course Objectives

1. Introduce key mathematical and statistical tools essential for data analysis.
2. Explain the Data Science process, lifecycle, and its applications in diverse domains.
3. Develop proficiency in R programming for data manipulation and basic analytics.
4. Teach effective data handling, transformation, and visualization techniques using R.
5. Explore practical use cases in Data Science with modeling, clustering, and ethical awareness.

Course Outcomes

At the end of the course, students will be able to:

1. Apply statistical methods and probability distributions to analyze and interpret data.
2. Describe the Data Science workflow and perform exploratory data analysis (EDA).
3. Use R programming constructs and libraries for data input, control flows, and functions.
4. Handle and clean datasets using R tools like dplyr, tidyr, and manage missing/time-based data.
5. Implement basic machine learning models and evaluate performance using appropriate metrics and visual tools.

Unit 1. Mathematical & Statistical Foundations:
Sets, Functions, Probability, Random Variables, Descriptive Statistics: Mean, Median, Mode, Variance, Standard Deviation, Probability Distributions: Binomial, Normal, Poisson, Hypothesis Testing: t-test, Chi-square test, Correlation & Regression

Unit 2. Introduction to Data Science Process:
Introduction- Definition - Data Science in various fields - Examples - Impact of Data Science - Data Analytics Life Cycle - Data Science Toolkit - Data Scientist - Data Science Team, Exploratory Data Analysis (EDA), Feature Engineering & Data Transformation

Unit 3. Basics of R Programming:
Introduction to R and RStudio, Data Types, Variables, Operators, Control Structures (if, loops), Functions and Packages, Data Input/Output (CSV, Excel, XML, JSON).

Unit 4. Data Handling & Visualization in R:
Data Frames, Lists, Matrices, dplyr and tidyr for Data Wrangling, Handling Missing Data, Working with Date/Time in R.

Unit 5. Applications & Case Studies in Data Science:
Simple Linear Regression, Model Evaluation: Accuracy, Confusion Matrix, ROC.

K-Means Clustering, Text Mining & Word Clouds, Recommender Systems Basics, Ethical Issues in Data Science

Textbooks

1. An Introduction to Statistical Learning with Applications in R, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer, 2nd Edition, 2021
2. R for Data Science, Hadley Wickham and Garrett Grolemund, O'Reilly Media, 2017.

Reference Books

1. The Art of R Programming, Norman Matloff,, No Starch Press, 2011.
2. Modern Applied Statistics with S, W.N. Venables & B.D. Ripley, Springer, 2002.
3. Introduction to Data Science: Data Analysis and Prediction Algorithms with R, Rafael A. Irizarry, CRC Press, 2020.
4. Data Science from Scratch: First Principles with Python (for conceptual clarity only), Joel Grus,

Activities:

Outcome: Apply statistical methods and probability distributions to analyze and interpret data.

Activity: Analyze a dataset (e.g., student scores or sales data) to:

- Calculate mean, median, mode, variance, and standard deviation
- Fit and visualize a normal distribution
- Apply probability rules to answer questions (e.g., likelihood of scoring above 80)

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

- Accuracy of statistical calculations
- Correct use of probability formulas
- Interpretation of results and distribution plots

Outcome: Describe the Data Science workflow and perform exploratory data analysis (EDA).

Activity: Use a real-world dataset (e.g., Titanic or COVID data) to:

- Outline the steps of the Data Science workflow
- Perform EDA using summary statistics and visualizations (histograms, boxplots, scatterplots)

Evaluation Method: Presentation and checklist (10-point scale):

- Clear explanation of workflow stages
- Quality of EDA insights
- Use of appropriate plots and summaries

Outcome: Use R programming constructs and libraries for data input, control flows, and functions.

Activity: Write an R script that:

- Reads a CSV file
- Uses if, for, and while loops
- Defines and calls custom functions with arguments and return values

Evaluation Method: Code review and execution test to verify (10-point scale):

- Correctness of the syntax and logic
- Functionality of control structures
- Output accuracy and modularity

Outcome: Handle and clean datasets using R tools like dplyr, tidyr, and manage missing/time-based data.

Activity: Clean a messy dataset using:

- dplyr for filtering, selecting, and mutating
- tidyr for reshaping and handling missing values
- Time-based operations (e.g., filling gaps, formatting dates)

Evaluation Method: Before-and-after comparison (10 point score):

- Completeness of cleaning steps
- Use of appropriate functions
- Handling of missing/time data

Outcome: Implement basic machine learning models and evaluate performance using appropriate metrics and visual tools.

Activity: Build a simple classification model (e.g., logistic regression or decision tree) using R:

- Train/test split
- Predict outcomes
- Evaluate using confusion matrix, accuracy, precision, recall

Evaluation Method: Model report and demo (10 point scale):

- Correct implementation of model
- Use of evaluation metrics

SEMESTER-V

COURSE 12 B: DATA SCIENCE WITH R

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

List of Practicals:

1. Compute Mean, Median, Mode, Variance, and Standard Deviation
2. Visualize Binomial, Normal, and Poisson Distributions
3. Perform t-test and Chi-Square Test in R
4. Calculate Correlation and Build a Simple Linear Regression Model
5. Conduct Exploratory Data Analysis (EDA) on a Real-World Dataset
6. Apply Feature Engineering: Scaling, Normalization, and Encoding
7. Practice R Programming: Variables, Control Structures, and Functions
8. Read and Write Data from CSV, Excel, JSON, and XML Files
9. Use dplyr and tidyr for Data Wrangling Tasks
10. Handle Missing Data and Detect Outliers
11. Work with Dates and Times in R
12. Visualize Data Using ggplot2 (Bar, Scatter, Histogram, Boxplot)
13. Perform K-Means Clustering and Visualize Clusters
14. Evaluate Models Using Confusion Matrix, Accuracy, and ROC Curve
15. Perform Text Mining and Create a Word Cloud