

24CSH202: Principles of Data Science

Hours/Week: 4
Credits: 4

I.A. Marks: 30
Exam. Marks: 70

Course Learning Objectives:

- The probability distributions and density estimations to perform analysis of various kinds of data
- The statistical analysis techniques using Python and R programming languages.
- Expand the knowledge in R and Python to use it for further research.
- The students will be able to carry out data analysis/statistical analysis and visualize the data.

Course Outcomes: After completing the course, the students will be able to,

- Understand the fundamentals of data analytics
- Clean and reshape messy datasets
- Realize the importance of filtering functions, charts and tables.
- Learn the usage of python programming for data analysis
- Understand the various search methods and visualization techniques.
- Use exploratory tools such as clustering and visualization tools to analyze data
- Perform linear regression analysis
- Use methods such as nearest neighbors, decision trees, and support vector machines to build a classifier

UNIT-I

12Hrs.

Basics of Data Analytics, Applications of Data Analytics, Phases in Data Analytics, Data Definitions and Analysis Techniques, Elements, Variables, and Data categorization, Levels of Measurement, Data Management and Indexing, Introduction to Statistical Learning.

UNIT-II

12Hrs.

Basic analysis techniques, Central Tendency and Dispersion, Probability Distributions, Sampling and Sampling Distribution, Distribution of sample: Mean, population and Variance, Confidence interval estimation, Statistical hypothesis generation and testing, Chi-Square test, t-Test, Analysis of variance, Correlation analysis, Maximum likelihood test, Regression analysis.

UNIT-III

12Hrs.

Data pre-processing techniques, Summarizing and Computing Descriptive Statistics, Handling Missing Data, Hierarchical Indexing, Data Loading, Storage and File formats, Data Wrangling: Clean, Transform, Merge, Plotting and Visualization, Data Aggregation and Group Operation.

UNIT-IV

12Hrs.

Classification techniques: CART, Attribute selection measures, Bayesian Classification, Rule based Classification, Classification by Backpropagation, SVM, Cluster Analysis: K-mean Clustering, Hierarchical Method of Clustering, Density based Methods, Grid Based

REFERENCE BOOKS:

1. McKinney, Wes. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. O'Reilly Media, Inc., 2012, 1st Edition.
2. Samir Madhavan, Mastering Python for Data Science, 2015.
3. Jiawei Han, Micheline Kamber "Data Mining Concepts and Techniques", 2 nd Edition, Elsevier, Reprinted 2008.
4. E. Alpaydin, "Machine Learning", MIT Press, 2010.

