

Data Science

Code:

Nature of the Course: TH+PR

Year: Fourth

Semester: VII

Credit: 3

Contact Hours:45

Course Objectives

The course provides a solid foundation in the principles and techniques used in data science, as well as hands-on experience using popular data science tools. Students will learn how to clean and prepare data for analysis, explore data using visualizations and summary statistics, build statistical models and use machine learning algorithms for classification and regression tasks, visualize data and communicate insights to stakeholders.

Course Descriptions

Data Science is an interdisciplinary field that involves using statistical and computational methods to extract insights from data. This course introduces the field of data science, including the tools and techniques used to collect, clean, analyze, and visualize data. The course covers topics such as data exploration, statistical analysis, machine learning, and data visualization.

The course will be delivered through a combination of lectures, hands-on exercises, and projects. Students will have access to data science software tools to practice and apply the concepts learned. The course will also include guest lectures from industry experts to provide students with insights into real-world applications of data science.

Course Outcome: Upon completing the course on Data Science, students should be able to:

1. Understand the key concepts, principles, and techniques of Data Science.
2. Collect, clean, and preprocess data for analysis.
3. Perform statistical analysis and modeling on data using appropriate methods.
4. Apply machine learning algorithms and techniques to solve real-world problems.
5. Create effective data visualizations to communicate insights and findings.

CONTENT

1. Introduction to Data Science

[6 Hours]

- 1.1. Overview of Data Science
- 1.2. Importance of Data science
- 1.3. Concept of Big Data
- 1.4. Skills and Tools required for Data Science

- 1.5. **Ethical Considerations in Data Science:** Privacy and Security in Data Science, Bias and Fairness in Data Science, Transparency and Accountability in Data Science
- 1.6. **Applications of Data Science:** Business Analytics, Social Media Analysis, Healthcare Analytics, Environmental Analytics, and Local use cases

2. Statistical Analysis [10 Hours]

- 2.1. **Random Sampling and Sample Bias:** Bias, Random Selection, Size Versus Quality, Sample Mean Versus Population Mean, Mean versus Median
- 2.2. **Inferential Statistics:** Central Limit Theorem, and Standard Error, Confidence Intervals, Normal Distribution, Student's t-Distribution, and Binomial Distribution
- 2.3. Hypothesis Tests
- 2.4. Statistical Significance
- 2.5. Regression and multi-variate analysis
- 2.6. Time Series Analysis
- 2.7. Conditional Probability: Bayes Theorem

3. Data Collection and Preprocessing [10 hours]

- 3.1. **Data Sources:** Reading Files, Scraping the Web, and Using APIs
- 3.2. **Data Cleaning:** Handling Missing Data, Noisy Data, and Incomplete Data
- 3.3. **Data Wrangling:** Join, Combine, and Reshape
- 3.4. Data Integration
- 3.5. Data Transformation
- 3.6. Handling Categorical Data

4. Data Exploration and Visualization [10 hours]

- 4.1. **Exploratory Data Analysis (EDA):** Philosophy of Exploratory Data Analysis, Estimates of Location, Estimates of Variability, Exploring the Data Distribution, Exploring Binary and Categorical Data
- 4.2. **Data Visualization Techniques:** Percentiles and Boxplots, Frequency Tables, Histograms, Estimates Line Plots, Bar Plots, Density Plots, and Scatter or Point Plots

5. Introduction to Data Science Tools [9 hours]

- 5.1. Basic Concept of NumPy and Pandas
- 5.2. Basic Concept of Matplotlib and Seaborn
- 5.3. Machine Learning with **scikit-learn**

Practical: Using Python or R tools students should practice enough on real-world data intensive problems to analyze and extract insights from data using effective data visualizations to communicate insights and findings.

References

1. Wes McKinney - Python for Data Analysis - Data Wrangling with pandas, NumPy, and Jupyter - O'Reilly Media (2022)
2. Joel Grus - Data Science from Scratch First Principles with Python, Second Edition-O'Reilly Media (2019)

3. Fawcett Tom, Provost Foster - Data science for business-O'Reilly (2013)
4. Peter Bruce, Andrew Bruce and Peter Gedeck - Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python- O'Reilly Media, 2020
5. O'Neil, Cathy, Schutt, Rachel - Doing data science-O'Reilly Media (2014)