ADVANCED DATA SCIENCE

Lecture 1: Introduction to Data Science and Data Handling in Python

By

Dr. Sumit Kumar Singh

Associate Professor

SR University

COURSE MOTIVATION

- Data drives discovery, innovation, and decision-making.
- Data Science blends mathematics, statistics, and computing.
- Advanced Data Science focuses on the theory and algorithms behind modern AI.

WHAT IS ADVANCED DATA SCIENCE?

- Data Science: Extracting insights and knowledge from data.
- Machine Learning: Building algorithms that learn from data.
- Advanced DS: Focuses on the mathematical, statistical, and algorithmic foundations.

STRUCTURE OF THE COURSE

1. Importing, Summarizing, and Visualizing Data
2. Statistical Learning
3. Monte Carlo Methods
4. Unsupervised Learning
5. Regression and Regularization
6. Classification
7. Decision Trees and Ensembles
8. Deep Learning

DATA SCIENCE WORKFLOW

1. Data Acquisition
2. Data Cleaning & Structuring
3. Data Summarization & Visualization
4. Modeling (Statistical or ML-based)
5. Evaluation & Deployment

PYTHON ECOSYSTEM FOR DATA SCIENCE

- Data Handling: Pandas, NumPy
- Visualization: Matplotlib, Seaborn
- Machine Learning: Scikit-learn
- Deep Learning: TensorFlow, PyTorch
- Development Environment: Jupyter/Colab

IMPORTING AND STRUCTURING DATA

- Data originates from random experiments.
- Features (columns) and observations (rows).
- Quantitative features: continuous or discrete.
- Qualitative features: categorical (nominal or ordinal).

SUMMARY STATISTICS AND TABLES

- Descriptive statistics: mean, median, variance, quantiles.
- Frequency tables and cross-tabulation.
- Pandas methods: describe(), value_counts(), crosstab().

DATA VISUALIZATION

- Qualitative → Bar Charts
- Quantitative → Histograms, Boxplots, ECDF
- Bivariate → Scatter Plots, Grouped Boxplots

DISCUSSION / LAB PREVIEW

- Practice loading a dataset.
- Identify variable types.
- Visualize distributions with Matplotlib or Seaborn.