

Introduction to Data Science

Raj Kumar Biswokarma

What is Data Science ?

- Data Science is the process of extracting useful insights from data
- It converts raw data into useful information
- It Combines of:
 - Data
 - Statistics
 - Programming
 - Domain Knowledge
- Goal: Better Decision-making using data
- Data Science = Asking the right questions + using data to answer them

Why is Data Science Important?

- Data is growing very fast
- Manual analysis is no longer possible
- Organizations want:
 - Accuracy
 - Speed
 - Evidence-based decisions
- **Business:**
 - Better Way of Identifying who theirs customer
 - Information gained from data analysis used to communicate with customers
 - Allows business to see problems and respond them accordingly

Application of Data Science

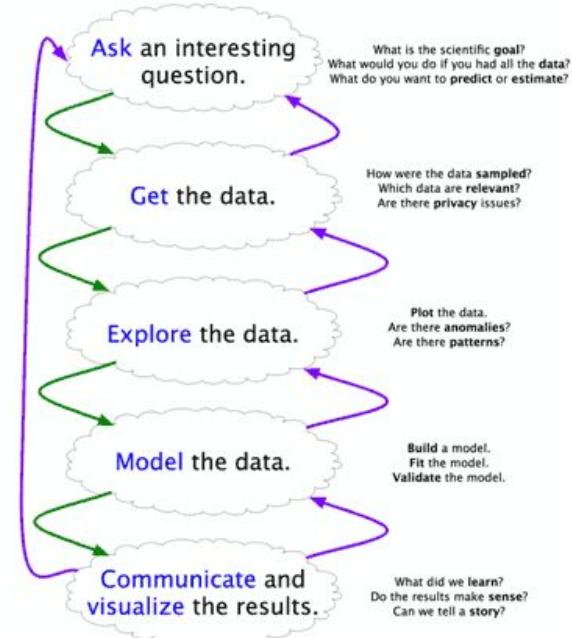
- Business → understanding customers
- Banking → loan and fraud analysis
- Government → census and planning
- Healthcare → disease trends
- Education → student performance

Various data science tools



Data Science Project Methodology

- **Problem Understanding**
 - What question are we trying to answer?
- **Data Collection**
 - Where is the data coming from?
- **Data Cleaning**
 - Removing errors, missing values, duplicates
- **Data Analysis and Visualization**
 - Finding patterns and showing them clearly
- **Insights & Decision**
 - Final goal: support action and decision-making



Tools of Choice-Python

- **Python**
 - Interpreted Programming Language
 - Developed by: Guido van Rossum
 - Versatile Programming language
 - Web Applications
 - Desktop GUI Applications
 - Software development
 - Scientific and Numeric
- **Why for Data Analysis**
 - Easy to Learn - Simplicity
 - Great Open-source community
 - Multiple platforms (Windows, Mac, Linux)
 - Large data science libraries
 - Widely used in Industry

Case study

- Sales Data Analysis
 - Monthly sales data is available
 - Management wants to know:
 - Which products are underperforming?
 - Which locations need attentions
 - Data Science helps to :
 - Compare Regions
 - Identify problem products
- Student Performance Analysis
 - Students performance data (exam / attendance)
 - Question:
 - Which Students need support?
 - Is attendance affecting performance
 - Data Science:
 - Identify weak students early
 - Improve overall results
- Loan & Customer Analysis
 - Customer and loan data is available
 - Questions:
 - Which customers are high risk?
 - Which factors affect loan repayment?
 - Data Science:
 - Analyze risk factors
 - Support loan approval decisions