Data Science Fundamentals

Duration: 16 hours (2 days)

Pre-requisites

- Basic Python programming skills
- Familiarity with Excel or CSV data formats
- Fundamental understanding of statistics (mean, median, variance)
- Optional: Prior exposure to pandas or NumPy

Course Outcomes

By the end of this 2-day workshop, participants will:

- Understand the data science lifecycle from data to insights
- Perform data cleaning, transformation, and analysis using Python
- Visualize data using matplotlib and seaborn
- Build and evaluate simple machine learning models
- Apply end-to-end data science workflows on real datasets

Day 1: Data Exploration & Analysis

1. **Introduction to Data Science**
   - What is Data Science?
   - Data Science lifecycle: Define → Acquire → Prepare → Analyze → Act
   - Tools of the trade (Python, Jupyter, pandas, scikit-learn)
2. **Python for Data Science Refresher**
   - Lists, dictionaries, functions, loops
   - Working with Jupyter Notebooks
3. **Data Manipulation with pandas**
   - Loading CSV, Excel files
   - DataFrames: filter, sort, group, merge
   - Handling missing values, duplicates

4. **Exploratory Data Analysis (EDA)**
    - Descriptive statistics (mean, median, std)
    - Value counts, correlation, outliers
    - Hands-on EDA with a real dataset
5. **Data Visualization**
    - Line, bar, pie, scatter, histogram
    - Using matplotlib and seaborn
    - Customizing plots for reports
6. **Hands-on Lab: Titanic Dataset Analysis**
    - Clean, explore, and visualize insights from Titanic dataset

---

Day 2: Data Science in Action (Model Building)

7. **Introduction to Machine Learning**
    - Supervised vs. Unsupervised learning
    - Classification vs. Regression
8. **Feature Engineering**
    - Encoding categorical variables
    - Scaling numerical values
    - Splitting data into train/test sets
9. **Model Building with scikit-learn**
    - Train a simple Logistic Regression model
    - Evaluate with accuracy, precision, recall
    - Use Confusion Matrix and ROC curve
10. **Model Improvement Techniques**
    - Cross-validation
    - Hyperparameter tuning (GridSearchCV)
    - Avoiding overfitting
11. **Real-World Use Case: Predict Customer Churn**

- Load and analyze a churn dataset
- Build and evaluate a classification model

12. **Deployment & Next Steps**
    - Saving model with joblib
    - Basic intro to deployment using Streamlit