Gain expertise from experienced DATA SCIENCE professionals.

DIGI SAMURAI

# Data Science with Python

## Why Data Science?

Harnessing the power of data is no longer a luxury but a necessity in today's data-driven world, where insights are sought after and challenges abound.

A career in data science offers numerous advantages, including high demand, the opportunity to address the skills gap, diverse roles, challenging work, the ability to derive actionable insights from data, and the ability to contribute toward a data-driven world. Choosing a data science career can provide both professional satisfaction and the opportunity to make a significant impact in today's increasingly data-centric society.

Explore the Fascinating World of Data Science and Transform Your Career

Prepare yourself with critical knowledge to analyze data effectively, derive actionable insights, and navigate the data-driven world confidently and skillfully.

# DATA SCIENCE COURSE OVERVIEW

## Module 1: Python Programming

` **1.1 Introduction to Python**

- Overview and History of Python
- Installation and Setup (Anaconda, Jupyter Notebook, IDEs)
- Python Syntax and Interactive Shell

**1.2 Python Basics**

- Data Types (Integers, Floats, Strings, Booleans)
- Variables and Constants
- Basic Operators (Arithmetic, Comparison, Logical)
- Control Structures
- Conditional Statements (if, elif, else)
- Looping (for, while)
- Functions
- Defining Functions
- Function Arguments and Return Values
- Lambda Functions

**1.3 Data Structures**

- **Lists**
  - Indexing, slicing, and common list methods for manipulating and accessing elements.
  - Introduction to list comprehensions for creating lists efficiently.
- **Tuples**
  - Exploring the immutability of tuples and their use cases.
  - Working with tuple operations such as accessing elements and unpacking.
- **Sets**
  - Discussing the properties of sets, including uniqueness and unordered nature.
  - Performing set operations such as union, intersection, difference, and symmetric difference.
- **Dictionaries**
  - Understanding key-value pairs and dictionary characteristics.
  - Accessing, modifying, and managing data using dictionaries.
  - Exploring the concept of dictionaries containing other dictionaries and their applications in representing complex data structures.

**1.4 File Handling**

- Reading and Writing Text Files
- File Modes and File Operations
- Working with CSV Files

**1.5 Error Handling**

- Exception Handling (try, except, else, finally)
- Raising Exceptions
- Custom Exceptions

## Module 2: Python libraries for Data Science

**2.1 Pandas**

- Introduction to Pandas Data Structures
- Creating and Manipulating DataFrames and Series
- Indexing, Selecting, and Filtering Data
- Handling Missing Data
- DataFrame Operations (Merging, Joining, Concatenating)
- Grouping and Aggregating Data
- Pivot Tables and Cross-tabulations
- Time Series Data Manipulation

**2.2 Numpy**

- Numpy Arrays and Array Operations
- Array Creation Techniques (arange, linspace, zeros, ones)
- Indexing, Slicing, and Reshaping Arrays
- Broadcasting and Vectorized Operations
- Mathematical Functions and Linear Algebra Operations
- Random Number Generation and Statistical Operations

**2.3 Matplotlib**

- Basic Plotting (Line, Bar, Scatter, Histogram)
- Plot Customization (Labels, Titles, Legends, Styles)
- Subplots and Layout Adjustments
- Saving and Exporting Figures
- Plotting Multiple Series and Customizing Colors

**2.4 Seaborn**

- Introduction to Seaborn and Comparison with Matplotlib
- Statistical Plots (Distribution, Box, Violin, Pair Plots)
- Categorical Plots (Bar, Count, Boxen, Strip, Swarm)
- Matrix Plots (Heatmaps, Clustermaps)
- Plot Aesthetics and Themes
- Advanced Customizations (Annotations, Plotting on Facets)

## Module 3: Database Management System (DBMS) and SQL

### 3.1 Introduction to DBMS

- Basic Concepts and Terminology
- Types of Databases (Relational, NoSQL)
- Database Design and ER Diagrams

### 3.2 SQL Basics

- SQL Syntax and Query Structure
- Data Definition Language (DDL)
- Creating, Altering, and Dropping Tables
- Data Manipulation Language (DML)
- Inserting, Updating, and Deleting Data
- Data Querying (SELECT, WHERE, ORDER BY)
- Aggregate Functions (COUNT, SUM, AVG, MAX, MIN)
- Grouping Data (GROUP BY, HAVING)

### 3.3 Advanced SQL

- Joins (INNER, LEFT, RIGHT, FULL, SELF)
- Subqueries and Nested Queries
- Set Operations (UNION, INTERSECT, EXCEPT)
- Window Functions
- Indexing and Optimization Techniques
- Transactions and ACID Properties

## Module 4: Excel for Data Science

### 4.1 Excel Basics

- Data Entry and Formatting
- Basic Formulas and Functions (SUM, AVERAGE, IF)
- Cell Referencing (Relative, Absolute, Mixed)

### 4.2 Data Analysis Tools

- Pivot Tables and Pivot Charts
- Data Visualization (Charts and Graphs)
- Data Cleaning and Transformation (Text to Columns, Remove Duplicates)

### 4.3 Advanced Excel Features

- Lookup Functions (VLOOKUP, HLOOKUP, INDEX-MATCH)
- Data Validation and Conditional Formatting
- What-If Analysis (Goal Seek, Data Tables, Scenarios)
- Introduction to Macros and VBA Basics

## Module 5: Statistics

### 5.1 Descriptive Statistics

- Measures of Central Tendency (Mean, Median, Mode)
- Measures of Dispersion (Range, Variance, Standard Deviation)
- Data Visualization (Histograms, Box Plots)
- Skewness and Kurtosis

**5.2 Probability Basics**

- Basic Probability Concepts and Rules
- Conditional Probability and Bayes' Theorem
- Probability Distributions (Uniform, Normal, Binomial, Poisson)
- Central Limit Theorem

**5.3 Inferential Statistics**

- Sampling and Sampling Distributions
- Hypothesis Testing (Null and Alternative Hypotheses)
- Type I and Type II Errors
- t-tests (One-sample, Two-sample, Paired)
- ANOVA (Analysis of Variance)
- Chi-square Test for Independence
- Correlation and Causation
- Simple and Multiple Linear Regression Analysis

## Module 6: Feature Engineering

**6.1 Data Preprocessing**

- Handling Missing Data (Imputation, Dropping)
- Encoding Categorical Variables (One-hot Encoding, Label Encoding)
- Scaling and Normalization (Min-Max, Standardization)
- Outlier Detection and Treatment

**6.2 Feature Selection**

- Filter Methods (Correlation, Chi-square, ANOVA)
- Wrapper Methods (Forward, Backward, Recursive Feature Elimination)
- Embedded Methods (Lasso, Ridge, Decision Trees)

**6.3 Feature Creation**

- Polynomial Features and Interactions
- Binning and Discretization
- Feature Extraction (PCA, LDA)
- Time-based Features

## Module 7: Machine Learning

**7.1 Supervised Learning**

- **Linear Regression:**
  - Model Assumptions and Diagnostics

- Training and Evaluation (MSE, R^2 score)
- Regularization Techniques (Ridge, Lasso)

- **Logistic Regression:**
  - Binary Classification and Sigmoid Function
  - Model Evaluation (Confusion Matrix, Precision, Recall, F1 Score)
  - ROC Curve and AUC
- **K-Nearest Neighbours (KNN):**
  - Distance Metrics (Euclidean, Manhattan)
  - Choosing the Value of K (Cross-Validation)
  - Pros and Cons of KNN
- **Decision Trees (DT):**
  - Tree Construction and Splitting Criteria (Gini, Entropy)
  - Overfitting and Pruning
  - Decision Tree Visualization
  - Feature Importance

**7.2 Unsupervised Learning**

- **Clustering:**
  - K-Means:
    - Elbow Method and Silhouette Score
    - Choosing the Number of Clusters
  - Hierarchical Clustering:
    - Dendrograms and Linkage Methods (Single, Complete, Average)
    - Agglomerative vs. Divisive Clustering
  - Other Clustering Techniques (DBSCAN, Mean Shift)
- **Dimensionality Reduction:**
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - t-SNE for Visualization

## Module 8: Data Analytics using BI Tools

**8.1 Introduction to BI Tools**

- Importance of Business Intelligence (BI)
- Overview of Popular BI Tools (Power BI, Tableau)

**8.2 Power BI/Tableau Basics**

- Data Connection and Import
- Data Cleaning and Transformation
- Creating and Customizing Visualizations (Bar, Line, Pie, Map)
- Calculated Fields and Measures
- Filters and Slicers

- Interactive Dashboards and Storytelling
- Sharing and Publishing Reports

## Module 9: Deep Learning

**9.1. Introduction to Deep Learning**
- Overview and applications
- Key differences between deep learning and traditional machine learning

**9.2 Concepts of Neural Networks**
- Neurons, layers, activation functions
- Loss functions and optimization

**9.3. Artificial Neural Networks (ANN)**
- Architecture and training
- Backpropagation and gradient descent

**9.4. TensorFlow 2.0**
- Introduction to TensorFlow and its ecosystem
- Building and training models with TensorFlow

**9.5. Convolutional Neural Networks (CNN)**
- Convolutional layers, pooling layers
- Applications in image recognition

**9.6. Recurrent Neural Networks (RNN)**
- Sequence modeling, LSTM, and GRU networks
- Applications in time series and text data

**9.7. Transfer Learning**
- Utilizing pre-trained models for new tasks
- Fine-tuning and feature extraction

**9.8. Generative Adversarial Networks (GANs)**
- Concept and architecture
- Applications in image generation

**9.9. Introduction to Time Series**
- Key concepts and components
- Applications in forecasting

**9.10. End-to-End Project using Deep Learning**
- Practical project encompassing data preprocessing, model building, evaluation, and deployment

## Module 10: Natural Language Processing (NLP)

**10.1. Introduction to NLP**
- Overview and applications
- Key challenges in NLP

**10.2. Basic Text Processing**
- Tokenization and text normalization
- Stemming and lemmatization

- Stop words removal

**10.3. Smoothing and Sequential Tagging**

- Smoothing techniques (e.g., Laplace smoothing)
- Part-of-Speech (POS) tagging
- Named Entity Recognition (NER)

**10.4. Parsing**

- Syntactic parsing (constituency and dependency parsing)
- Parsing algorithms (e.g., CYK algorithm, shift-reduce parsing)

**10.5. Advanced NLP**

- Transformers and attention mechanisms
- Pre-trained language models (e.g., BERT, GPT)
- Applications in sentiment analysis and machine translation

## Module 11: Big Data

**11.1. Introduction and Overview**

- Definition, characteristics, and importance of Big Data
- Key technologies: Hadoop, Spark, NoSQL databases, Apache Kafka Apache Flink, Apache Cassandra, Apache HBase, Apache Storm

**11.2. Data Storage and Processing**

- Distributed file systems (HDFS, GFS, Amazon S3)
- Data processing frameworks: Batch, real-time, and stream processing

**11.3. Data Acquisition, Analysis, and Visualization**

- Methods and tools for data collection and ingestion
- Data analysis techniques, machine learning, and visualization tools

**11.4. Data Management and Governance**

- Governance, security, privacy, and compliance in Big Data
- Scalability, performance optimization, and cluster management

**11.5. Applications and Future Trends**

- Practical applications in various industries (IoT, healthcare, finance, etc.)
- Current challenges and emerging trends (edge computing, AI integration)

## Module 12: Project Work

**12.1 Project Planning**

- Defining the Problem Statement and Objectives
- Data Collection and Source Identification
- Project Workflow and Timeline

**12.2 Data Collection and Preparation**

- Data Acquisition (APIs, Web Scraping, Public Datasets)
- Data Cleaning and Preprocessing
- Exploratory Data Analysis (EDA)

**12.3 Model Building and Evaluation**

- Model Selection and Implementation
- Hyperparameter Tuning and Optimization
- Model Evaluation and Validation (Cross-Validation, Confusion Matrix)

**12.4 Presentation and Reporting**

- Visualizations and Insights Communication
- Report Writing and Documentation
- Presentation and Q&A

# Course Pre-requisite

**No prior knowledge of data science is required, but a basic understanding of statistics and the ability to wield a wand, or rather a programming language, will be advantageous.**

# Course Duration

**60 Hours to Explore Data Science**

## CONTACT US

www.digisamurai.co.in

+91 8910632224

+91 7595887833