

# Data Science and Big Data Analytics

**Duration:** 5 Days · **Training format:** Virtual Classroom / Online

## Description

---

"Data Science and Big Data Analytics" course is designed to provide participants with comprehensive knowledge and practical skills in the field of data science and big data analytics. The course covers a wide range of topics to equip participants with the necessary tools and techniques for data analysis, machine learning, and working with big data technologies.

## Outline

---

### PYTHON

- Introduction to Python
- Overview of Python Programming Language
- Python Development Environments and Configurations
  - Anaconda and Python Development Environment (Spyder, Jupyter Notebook)
  - Google Colaboratory (Environment for Practical Applications during the Training)

### DATA SCIENCE

- Foundations of Data Science
  - What is Data Analysis? What Can Be Done with Data Analysis?
  - What is Data Science?
  - Elements of Data Science
  - Stages of Extracting Useful Information from Data (Data Analytics)
    - Descriptive Analytics
    - Diagnostic Analytics
    - Predictive Analytics
    - Prescriptive Analytics
  - Applications of Data Science
- Implementation in the Data Science Process Cycle
  - CRISP-DM Methodology
    - Business Understanding
    - Data Understanding
    - Data Preparation
    - Modeling

- Evaluation
- Deployment
- Basic Stages of Application Development in Data Science
- Tools Used for Application Development in Data Science
  - Numpy
    - Array and Matrix Creation
    - Formatting Operations (Reshaping, Merging, Splitting)
    - Index Operations
    - Mathematical Operations (Random Number Operations, etc.)
    - Statistical Operations (min, max, mean, std, etc.)
  - Pandas
    - Series Operations (Creating Series, Features, etc.)
    - DataFrame Operations (Creation, Features, Element Operations, Merging, Grouping, Filtering, Apply, Pivot Tables)
    - Reading Data from Excel and CSV Files into DataFrame and Performing Operations on Data
  - Matplotlib/Seaborn
    - 2D Graph Usage (line, bar, scatter, histogram, pie, etc.)
    - 3D Graph Usage
    - Performing Operations on Graphs (Labeling Titles and Axes, Defining Colors, Legends, etc.)

## **Exploratory Data Analysis (EDA)**

- Data Literacy
  - What is Data Literacy?
  - Basic Concepts of Data Literacy
    - Population and Sample
    - Observation Unit
    - What is a Variable? What are the Types of Variables?
    - What is Scale? What are the Types of Scales?
    - Measures of Central Tendency
      - Mean, Median, Mode, Quartiles
    - Measures of Dispersion
      - Range, Standard Deviation, Variance, Skewness, Kurtosis
  - Data Definition
  - Organizing and Reducing Data
  - Data Representation
  - Data Analysis and Evaluation
  - Loading/Reading Data Sets from Files

- Learning About Data Size
- Sampling in Data
- Data Types in Data

## **Data Preprocessing and Cleaning**

- Data Identification
- Feature Viewing and Selection
- Sorting and Grouping
- Operations on Features
  - Adding Features
  - Renaming Feature Names
  - Deriving New Features from Existing Features
  - Using the re (regular expression) Module in Data Analysis
  - Deleting Features
- Operations on Observations
  - Display of Observations (From the Beginning, From the End, Randomly)
  - Adding Observations
  - Deleting Observations
- Data Filtering
  - Filtering Using Dictionaries and Lists
  - Filtering with Query
- Missing Data
  - Detecting Missing Data
  - Approaches to Deleting Missing Data
  - Approaches to Completing Missing Data
    - Complete with a Constant Value
    - Complete with the Mean
    - Complete with Data from the Previous and Next Observation
  - Proportional Operations on Missing Data
- Duplicate Data
  - Detecting Duplicate Data
  - Cleaning Duplicate Data
- Data Transformation Operations
  - Scaling and Normalization of Data
  - Merging-Aggregation
  - Categorical Data
- Detection of Outliers/Extreme Values

## **Statistical Operations on Numerical Data**

- Distribution
- Variance Analysis
- Correlation Analysis
- Data Visualization
  - Plotting (line, bar, pie, heat map, etc.)
  - Operations on Graphs

## **Practical Data Analysis on Ready Data Sets**

- Tools for Creating Data Analysis Reports

## **MACHINE LEARNING**

- Fundamentals of Machine Learning
  - What is Machine Learning?
  - Differences Between Machine Learning and Deep Learning
  - Real-Life Examples
  - Basic Concepts and Terminology
    - Types of Problems (Regression, Classification)
    - Model
    - Splitting the Data Set into Training and Test Sets
    - Overfitting
    - Model Validation
  - Types of Learning
    - Supervised Learning
    - Unsupervised Learning
    - Reinforcement Learning
  - Feature Engineering
    - Outlier Detection
    - Data Cleaning
    - Data Transformation (encoding scaling)
    - Data Reduction
    - Feature Extraction
  - Methods for Evaluating Machine Learning Model Performance
    - Confusion Matrix
      - Accuracy, Recall, Precision
    - R2 Score
    - F1 Score
    - AUC-ROC Curve
    - Mean Absolute Error (MAE)
    - Mean Squared Error (MSE)

- Tools for Implementing Machine Learning with Python
  - Scikit-Learn Module and Machine Learning

## **Machine Learning Algorithms/Models and Application Development**

- Regression Models (Theory, Model, Prediction)
  - Simple Linear
  - Multiple Linear
- Classification Models (Theory, Model, Prediction)
  - Classification with Logistic Regression
  - Classification with K-Nearest Neighbours (KNN)
  - Classification with Decision Trees (CART)
  - Classification with Support Vector Machines (SVM)
- Clustering Models (Theory, Model, Prediction) (Unsupervised Learning Application)
  - Clustering with K-Means Algorithm
  - Determining the Optimal Number of Clusters with Elbow Method
  - Hierarchical Clustering
- Boosting (Prediction) Models (Theory, Model, Prediction)
  - Gradient Boosting
  - XBoost
  - LightGBM
- Real-Time (On Streaming Data) Machine Learning Application
  - Developing an application for predicting data pulled from an online database (firebase) using an ML model

## **Saving Machine Learning Models and Transferring to Other Applications**

- Methods for Using Data Science Models in Different Applications
- Model Transfer with Pickle
- Model Transfer with Joblib
- Converting to Native Code with m2cgen

## **BIG DATA ANALYTICS**

### **BIG DATA**

- What is Big Data?
- Components and Characteristic Features of Big Data
- Use Cases of Big Data - Real-Life Examples
- Expected Skills for Employees in Big Data
- Skills Required for Big Data Expertise
- Big Data Technologies and Tools
  - Technologies Used in Big Data Architecture

- Apache Hadoop Ecosystem
- Apache Spark Technologies
- Application Development Model in Distributed Architectures for Big Data

## **BIG DATA ANALYTICS**

- Development Processes of Big Data Applications with PySpark
  - PySpark Installation
  - Basic Data Frame Operations with PySpark
  - SQL Operations with PySpark
  - Big Data Visualization with PySpark
- Machine Learning Applications on Big Data with PySpark
  - Customer Churn Analysis (Theory, Model, Prediction)
    - Data Preprocessing
    - Machine Learning with Gradient Boosting Machine (GBM) Algorithm
  - Individual Activity Analysis (Theory, Model, Prediction)
    - Data Preprocessing
    - Machine Learning on Big Data with Logistic Regression Algorithm