

BECOME AN EXPERT IN

# **DATA** **SCIENCE**

*STATISTIC / MACHINE LEARNING /  
NLP / R / PYTHON/SPARK*

## **About the Course**

Data Science is the study of the generalize extraction of knowledge from data. Being a data Scientist requires an integrated skill set spanning mathematics, statistics, machine learning, databases and programming languages along with a good understanding of the craft of problem formulation to engineer effective solutions.

This course will introduce students to this rapidly growing field and equip them with some of its basic principles and tools as well as its general mindset.

- Students will learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration, exploratory data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication.
- The focus in the treatment of these topics will be a balanced approach on breadth and depth, and emphasis will be placed on integration and synthesis of concepts and their application to real time problems.
- To make the learning contextual, real datasets from a variety of disciplines will be used.

## **Program Highlights**

- ✓ Most Comprehensive Curriculum
- ✓ Trained by passionate and Industry experts
- ✓ Each concept will be explained by golden rule

Theory → Example → Software Implementation (R/Python) → Real-Time applicability

- ✓ Designed for the Industry

- ✓ Live Project
- ✓ Placement Assistance

## **Audience**


Any degree. No programming and Statistics knowledge required.

## **Duration & Mode of Training**




- ✓ 6 months, Online Training and classroom.

## **Course Content**

### **INTRODUCTION**

-  Introduction to Data Science –
  - What is Data Science?
  - Why now?
  - Where Data Science is applicable?

### **DATA EXPLORATION USING STATISTICAL METHODS – DESCRIPTIVE AND INFERENCE STATISTICS**

-  Introduction to statistics
-  Summarizing Data
  - Central Tendency measures – Mean, Median and Mode
  - Measures of Variability – Range, Interquartile Range, Standard Deviation and Variance
  - Measures of Shape – Skewness and Kurtosis
  - Covariance, Correlation
-  Data Visualization
  - Histograms

- Pie charts
- Bar Graphs
- Box Plot

#### Probability basics

#### Parametric and Non parametric Statistical Tests

- 'f' Test
- 'z' Test
- 't' Test
- Chi-Square test

#### Probability Distributions

- Expected value and variance
- Discrete and Continuous
- Bernoulli Distribution
- Binomial Distribution
- Poisson Distribution
- Normal Distribution
- Exponential Distribution

#### Sampling methods and Central Limit Theorem

- Overview
- Random sampling
- Stratified sampling
- Cluster sampling
- Central Limit Theorem

#### Hypothesis Testing

- Type I error
- Type II error
- Null and Alternate Hypothesis
- Reject or Acceptance criterion
- P-value

#### Confidence Intervals

#### ANOVA

- Assumptions

- One way
- Two way

## MACHINE LEARNING – INTRODUCTION

- Introduction to Machine Learning
  - What is Machine Learning?
  - Statistics (vs) Machine Learning
  - Types of Machine Learning
    - Supervised Learning
    - Un-Supervised Learning
    - Reinforcement Learning

## SUPERVISED MACHINE LEARNING

- Classification
  - Nearest Neighbor Methods (knn)
  - Logistic
- Tree based Models – Decision Tree
  - Basics
  - Classification Trees
  - Regression Trees
- Probabilistic methods
  - Bayes Rule
  - Naïve Bayes
- Regression Analysis
  - Simple Linear Regression
  - Assumptions
  - Model development and interpretation
  - Sum of Least Squares
  - Model validation
  - Multiple Linear Regression

## Regression Shrinkage Methods

- Lasso
- Ridge

## Advanced Models – Black Box

- Support Vector Machine
- Neural Networks

## Ensemble Models

- Bagging
- Boosting
- Random Forests

## Optimization

- Gradient Descent (Batch and Stochastic)

## Recommendation Systems

- Collaborative filtering
  - User based filtering
  - Item based filtering

# UNSUPERVISED MACHINE LEARNING

## Association Rules (Market Basket Analysis)

- Apriori







## Cluster Analysis

- Hierarchical clustering
- K-Means clustering

## Dimensionality Reduction

- Principal Component Analysis
- Discriminant Analysis (LDA/GDA)






## MODEL VALIDATION

-  Confusion Matrix
-  ROC Curve (AUC)
-  Gain and Lift Chart
-  Kolmogorov-Smirnov Chart
-  Root Mean Square Error (RMSE)
-  Cross Validation
  - Leave one out cross validation (LOOCV)
  - K-fold cross validation

## NATURAL LANGUAGE PROCESSING

-  Introduction to Natural Language Processing
-  Sentiment Analysis
-  Text Similarity

## R Programming Language

-  Introduction
  - R Overview
  - Installation of R and RStudio software
  - Important R Packages
  - Datatypes in R – Vectors, Lists, Matrices, Arrays, Data Frames
-  Decision making & Loops
  - If-else, while, for
  - Next, break, try-catch
-  Functions
  - Writing functions
  - Nested functions
-  Built-in functions
  - Vapply, Sapply, Tapply, Lapply etc.
-  Data Preparation/Manipulation
  - Reading and Writing Data
  - Summarize and structure of data

- Exploring different datasets in R
- Sub Setting Data Frames
- String manipulation in Data Frames
- Handling Missing Values, Changing Data types, Data Binning Techniques, Dummy Variables

#### Data Visualization using ggplot2

- Basic charts – Histograms, Bar plots, Line graphs, Scatter plots etc.

## Python Programming Language

#### Introduction

- How is Python different from R
- Installing Anaconda- Python
- Setting up with spyder

#### Datatypes in Python

#### Importing modules

#### Introduction to Strings

#### String manipulation

#### Control loops:

- For
- While
- If else

#### Functions

- Lambda
- apply

#### Numpy

#### Pandas

- Introduction to Dataframes
- Conversion of written R codes into python

#### Scipy-Machine Learning in Python

#### Beautiful Soup

#### Matplotlib

# Spark and Scala Programming Language

## Module 1

Introduction to Big Data and Spark

Overview of BigData and Spark

Hadoop limitations

Spark History

Spark Architecture

Spark and Hadoop Advantages

Benefits of Spark + Hadoop

Introduction to Spark Eco-system

Spark Installation

## Module 2

Introduction to Scala

Scala foundation

Features of Scala

Setup Spark and Scala on Unbuntu and Windows OS

Install IDE's for Scala

Run Scala Codes on Scala Shell

Understanding Data types in Scala

Implementing Lazy Values

Control Structures

Looping Structures

Functions

Procedures

Collections



Arrays and Array Buffers

Map's, Tuples and Lists

### Module 3

Object Oriented Programming in Scala

Implementing Classes

Implementing Getter & Setter

Object & Object Private Fields

Implementing Nested Classes

Using Auxilary Constructor

Primary Constructor

Companion Object

Apply Method

Understanding Packages

Override Methods

Type Checking

Casting

Abstract Classes

### Module 4

Functional Programming in Scala

Understanding Functional programming in Scala

Implementing Traits

Layered Traits

Rich Traits

Anonymous Functions

Higher Order Functions

Closures and Currying

Performing File Processing

## Module 5

Foundation to Spark

Spark Shell and PySpark

Basic operations on Shell

Spark Java projects

Spark Context and Spark Properties

Persistence in Spark

HDFS data from Spark

Implementing Server Log Analysis using Spark

## Module 6

Working with Resilient Distributed DataSets (RDD)

Understanding RDD

Loading data into RDD

Scala RDD, Paired RDD, Double RDD & General RDD Functions

Implementing HadoopRDD, Filtered RDD, Joined RDD

Transformations, Actions and Shared Variables

## Module 7

Spark Eco-system - Spark Streaming & Spark SQL

Introduction to Spark Streaming

Architecture of Spark Streaming

Discretized streams RDD.

Applying Transformations and Actions on Streaming Data

Integration with Flume and Kafka.

Integration with Cassandra

Monitoring streaming jobs.

Introduction to Spark SQL

Querying Files as Tables

Text file Format

JSON file Format

Spark SQL performance optimization

Implementing Data visualization in Spark

### ***Project details:***

#### ***1) Prediction of Rainfall:***

- . Read the data sets.*
- . Exploratory Analysis.*
- . Prepossessing and cleaning the data sets .*
- . Build the model.*
- . Prediction using Linear Regression.*

#### ***2) Prediction of credit card loan :***

- . Read the data sets.*

- . Exploratory Analysis.
- . Prepossessing and cleaning the data sets .
- . Build the model.
- . Prediction using decision tree.

### 3) Prediction of Air Quality :

- . Read the data sets.
- . Exploratory Analysis.
- . Prepossessing and cleaning the data sets .
- . Build the model.
- . Prediction using random forest.

*Note: like we will provide more prediction by using different models on projects.*

*Mock Test :*

- . I'll assign mock test for each module.*
- . Provide the some programs of that particular module.*
- . Cover the interview question of module.*