DIPLOMA IN DATA SCIENCE

Module 1 (Introduction to Data Science)

- ✓ Need for Data Scientists
- ✓ Foundation of Data Science
- ✓ What is Data Analysis
- ✓ What is Data Mining
- ✓ Use Case of Data Science
- ✓ Analytics vs. Data Science
- ✓ Types of Analytics
- ✓ Analytics Project Lifecycle
- ✓ Business Intelligence vs. Data Science
- ✓ Life cycle of Data Science
- ✓ Tools of Data Science

Module 2 (DATA)

- ✓ Basis of Data Categorization
- ✓ Types of Data
- ✓ Data Collection Types
- ✓ Forms of Data & Sources

- ✓ Data Quality & Changes
- ✓ Data Quality Issues
- ✓ Data Quality Story
- ✓ What is Data Architecture
- ✓ Components of Data Architecture
- ✓ OLTP vs. OLAP
- ✓ How is Data Stored?

Module 3 (BIG DATA)

- ✓ What is Big Data?
- ✓ 5 Vs. of Big Data
- ✓ Big Data Architecture
- ✓ Big Data Technologies
- ✓ Big Data Challenge
- ✓ Big Data Requirements
- ✓ Big Data Distributed Computing & Complexity

Module 4 (Data Science Deep Dive)

- ✓ Get Inspired
- ✓ What Data Science is
- ✓ Why Data Scientists are in demand
- ✓ What is a Data Product
- ✓ The growing need for Data Science
- ✓ Large Scale Analysis Cost vs. Storage
- ✓ Data Science Skills
- ✓ Data Science Use Cases
- ✓ Data Science Project Life Cycle & Stages
- ✓ Map Reduce Framework
- ✓ Hadoop Ecosystem
- ✓ Data Acquisition
- ✓ Where to source data
- ✓ Techniques
- ✓ Evaluating input data
- ✓ Data formats
- ✓ Data Quantity
- ✓ Data Quality
- ✓ Resolution Techniques
- ✓ Data Transformation
- ✓ File format Conversions
- ✓ Annonymization

Module5 (Basics of Statistics & Probability)

- ✓ Statistics & Plotting
- ✓ Seabourn & Matplotlib - Introduction
- ✓ Univariate Analysis on a Data
- ✓ Plot the Data - Histogram plot
- ✓ Find the distribution
- ✓ Find mean, median and mode of the Data
- ✓ Take multiple data with same mean but different SD, same mean and SD but different kurtosis: find mean, SD, plot
- ✓ Multiple data with different distributions
- ✓ Making samples from the Data
- ✓ Making stratified samples - covered in bivariate analysis
- ✓ Find the mean of sample
- ✓ Central limit theorem
- ✓ Plotting

- ✓ Bivariate analysis
- ✓ Correlation
- ✓ Scatter plots
- ✓ Making stratified samples
- ✓ Categorical variables
- ✓ Class variable

Module 6 (Intro to R Programming)

- ✓ Introduction to R
- ✓ Business Analytics
- ✓ Analytics concepts
- ✓ The importance of R in analytics
- ✓ R Language community and eco-system
- ✓ Usage of R in industry
- ✓ Installing R and other packages
- ✓ Perform basic R operations using command line

R Programming Concepts

- ✓ The data types in R and its uses
- ✓ Built-in functions in R
- ✓ Sub-setting methods
- ✓ Summarize data using functions
- ✓ Use of functions like head(), tail(), for inspecting data
- ✓ Use-cases for problem solving using R

Data Manipulation in R

- ✓ Various phases of Data Cleaning
- ✓ Functions used in Inspection
- ✓ Data Cleaning Techniques
- ✓ Uses of functions involved
- ✓ Use-cases for Data Cleaning using R

Exploratory Data Analysis (EDA) using R

- ✓ What is EDA?
- ✓ Why do we need EDA?
- ✓ Goals of EDA
- ✓ Types of EDA
- ✓ Implementing of EDA
- ✓ Boxplots, cor() in R

- ✓ EDA functions
- ✓ Multiple packages in R for data analysis

Data Visualization in R

- ✓ Summary of Statics
- ✓ Data Distributions
- ✓ Data Transformations
- ✓ Out layer Detection and Management
- ✓ Charts, Histograms, Bar charts, Box plots
- ✓ Scatter Plots
- ✓ Inference and Variable Selection
- ✓ Fancy Charts, Bubble Charts
- ✓ Story telling with Data
- ✓ Principle tenets
- ✓ Elements of Data Visualization
- ✓ Info graphics vs. Data Visualization
- ✓ Data Visualization & Graphical functions in R
- ✓ Plotting Graphs
- ✓ Customizing Graphical Parameters to improvise the plots

Module 7 (Predictive Analytics)

Simple Linear Regression

- ✓ Simple Linear Regression Model
- ✓ Least-Square Estimation of the Parameters
- ✓ Hypothesis Testing on the Slope and Intercept
- ✓ Coefficient of Determination

Multiple Linear Regressions

- ✓ Multiple Regression Models
- ✓ Estimation of Model Parameters
- ✓ Hypothesis Testing in Multiple Linear Regression
- ✓ Multicollinearity

Model Adequacy Checking

- ✓ Residual Analysis
- ✓ The PRESS Statistic
- ✓ Detection and Treatment of Outliers
- ✓ Lack of Fit of the Regression Model

Polynomial Regression

- ✓ Polynomial Model in One/ Two /More Variable

Dummy Variables

- ✓ The General Concept of Indicator Variables

Variables Selection and Model Building

- ✓ Forward Selection/Backward Elimination
- ✓ Stepwise Regression

Generalized Linear Models

- ✓ Concept of GLM
- ✓ Logistic Regression
- ✓ Poisson Regression
- ✓ Negative Binomial Regression
- ✓ Exponential Regression

Autocorrelation

- ✓ Regression Models with Autocorrelation Errors

Applied Multivariate Analysis

- ✓ Measures of Central Tendency, Dispersion and Association
- ✓ Measures of Central Tendency/ Measures of Dispersion

Multivariate Normal Distribution

- ✓ Exponent of Multivariate Normal Distribution
- ✓ Positive Definite/Negative Definite/Semi Definite
- ✓ Eigenvalues and Eigenvectors
- ✓ Spectral Decomposition
- ✓ Single Value Decomposition

Sample Mean Vector and Sample Correlation

- ✓ Distribution of Sample Mean Vector
- ✓ Interval Estimate of Population Mean
- ✓ Inferences for Correlations

Discriminant Analysis

- ✓ Discriminant Analysis (Linear)
- ✓ Estimating Misclassification Probabilities

- ✓ Using Software-Real Time Problems

MANOVA

- ✓ MANOVA
- ✓ Test Statistics for MANOVA
- ✓ Hypothesis Tests
- ✓ MANOVA table
- ✓ Using Software-Real Time Problems

Module 8 (Machine Learning-Supervised Learning)

- ✓ Introduction
- ✓ Steps in Supervised Learning
- ✓ Supervised Learning
- ✓ Regression and Classification
- ✓ Training, Testing and Validation
- ✓ Measure of Performance

Linear Regression

- ✓ Simple Linear Regression
- ✓ Cost Functions
- ✓ Sum of Least Squares
- ✓ Variable Selection
- ✓ Model Development and Improvement

Classification Logistic Regression

- ✓ Variable Selection Methods
- ✓ Gradient Descent/Ascent Procedure
- ✓ Maximum Likelihood Method
- ✓ Measurements of Accuracy
- ✓ Interpretation and Implementation
- ✓ Bayes Law
- ✓ Naive Bayes
- ✓ Nearest-Neighbor Methods (K-NN Classifier)
- ✓ Using Software-Real Time Problems

Decision Trees

- ✓ Rule Based Learning
- ✓ Construction of rules
- ✓ The Basics of Decision Trees

- ✓ Regression Trees
- ✓ Classification Trees

Bagging and Random Forests

- ✓ Resampling Methods
- ✓ Resampling Methods with replacements
- ✓ Resampling Methods with-out replacements
- ✓ Random Forests

Boosting

- ✓ Adaboost
- ✓ Gradient Boosting-GBM
- ✓ Extreme Gradient Boosting –Xgboost

Cross Validation

- ✓ K-Fold Cross Validation
- ✓ Cross Validation Usage
- ✓ Bias and Variance

Module 9 (Machine Learning-Un-Supervised Learning)

Cluster Analysis (Segmentation)

- ✓ Hierarchical Clustering
- ✓ K-Means Procedure

Dimensionality Reduction Techniques

- ✓ Principal Component Analysis
- ✓ Using Software-Real Time Problems

Forecasting

- ✓ Time Series
- ✓ Time Series Analysis
- ✓ Components of time Series
- ✓ Arch and Garch
- ✓ Moving Averages
- ✓ Exponential Smoothing
- ✓ Arima and Arimax
- ✓ Additive and Multiplicative

Text Mining

- ✓ Cleaning Text Data
- ✓ Pre Processing
- ✓ Sentiment Analysis
- ✓ Text Classification
- ✓ Natural Language Processing(**NLP**)

Module 10 (Introduction to Python Programming)

- ✓ Introduction to Data Science
- ✓ Introduction to Python
- ✓ Basic Operations in Python
- ✓ Variable Assignment
- ✓ Functions: in-built functions, user defined functions
- ✓ Condition: if, if-else, nested if-else, else-if

Data Structure - Introduction

- ✓ List: Different Data Types in a List, List in a List
- ✓ Operations on a list: Slicing, Splicing, Sub-setting
- ✓ Condition(true/false) on a List
- ✓ Applying functions on a List
- ✓ Dictionary: Index, Value
- ✓ Operation on a Dictionary: Slicing, Splicing, Sub-setting
- ✓ Condition(true/false) on a Dictionary
- ✓ Applying functions on a Dictionary
- ✓ Numpy Array: Data Types in an Array, Dimensions of an Array
- ✓ Operations on Array: Slicing, Splicing, Sub-setting
- ✓ Conditional(T/F) on an Array
- ✓ Loops: For, While
- ✓ Shorthand for
- ✓ Conditions in shorthand for

Module 11 (PYTHON for Data Science)

- ✓ Use of Pandas
- ✓ File I/O
- ✓ Series: Data Types in series, Index
- ✓ Data Frame
- ✓ Series to Data Frame
- ✓ Re-indexing
- ✓ Operations on Data Frame: Slicing, Splicing (also Alternate), Sub-setting
- ✓ Pandas

- ✓ Stat operations on Data Frame
- ✓ Reading from different sources
- ✓ Missing data treatment
- ✓ Merge, join
- ✓ Options for look and feel of data frame
- ✓ Writing to file
- ✓ DB operations

Module 12(PYTHON for Data Science)

- ✓ Data Manipulation & Visualization
- ✓ Data Aggregation, Filtering and Transforming
- ✓ Lambda Functions
- ✓ Apply, Group-by
- ✓ Map, Filter and Reduce
- ✓ Visualization
- ✓ Matplotlib, pyplot
- ✓ Seaborn
- ✓ Scatter plot, histogram, density, heat-map, bar charts

Module 13 (Machine Learning Using Python)

- ✓ Linear Regression
- ✓ Regression - Introduction
- ✓ Linear Regression: Lasso, Ridge
- ✓ Variable Selection
- ✓ Forward & Backward Regression

Module 14 (Introduction to Deep Learning)

- ✓ Logistic Regression
- ✓ Logistic Regression: Lasso, Ridge
- ✓ Naive Bayes
- ✓ Neural Networks

Module 15 (Machine Learning Using Python)

- ✓ Introduction
- ✓ Distance Concepts
- ✓ Classification
- ✓ k nearest
- ✓ Clustering
- ✓ k means

- ✓ Multidimensional Scaling

Module 16 (Machine Learning Using Python)

- ✓ Random Forest
- ✓ Decision trees
- ✓ SVM

Module 17 (BIGDATA-HADOOP)

- ✓ Big Data and Hadoop Introduction
- ✓ Understand Hadoop Cluster Architecture
- ✓ Map Reduce Concepts
- ✓ Advanced Map Reduce Concepts
- ✓ Hadoop 2.0 & YARN

Module 18 (HADOOP Eco System)

- ✓ PIG
- ✓ HIVE
- ✓ HBASE
- ✓ SQOOP
- ✓ Flume & Oozie

Module 19 (Apache spark)

- ✓ What is Scala?
- ✓ Why Scala for Spark?
- ✓ Scala in other frameworks
- ✓ Introduction to Scala REPL
- ✓ Basic Scala operations
- ✓ Variable Types in Scala
- ✓ Control Structures in Scala
- ✓ Foreach loop, Functions and Procedures
- ✓ Collections in Scala- Array
- ✓ ArrayBuffer, Map, Tuples, Lists, and more
- ✓ OOPs and functional programming in scala

Classes in Scala

- ✓ Getters and Setters
- ✓ Custom Getters and Setters
- ✓ Properties with only Getters

- ✓ Auxiliary Constructor and Primary Constructor
- ✓ Singletons
- ✓ Extending a Class
- ✓ Overriding Methods
- ✓ Traits as Interfaces and Layered Traits
- ✓ Functional Programming
- ✓ Higher Order Functions
- ✓ Anonymous Functions, and more

Machine Learning with Spark

- ✓ Spark Context and Hive Context
- ✓ Data-frames on Spark
- ✓ PY-spark

Module 20
Conclude (Final Implementation in Real Time World)

- ✓ POCs and Real-Time Use Cases
- ✓ AWS and Windows Azure
- ✓ Python Project Work
- ✓ Real world project