Course Outline: DATA SCIENCE MASTER (100 Hours)

1.Introduction

- What is Data Science
- Evolution of Analytics
- Data Science Components
- Data Scientist Skillset
- Types of Data Scientists
- Introduction to Machine Learning
- Data Science Process

2.Statistical Concepts and Application

2.1 Descriptive Statistics

- Data Basics
- Observations, variables, and data metrics
- Types of variables
- Relationships between variables
- Central Tendency
- Measures of Central Tendency
  - Arithmetic Mean / Average
  - Merits & Demerits of Arithmetic Mean
  - Mode
  - Merits & Demerits of Mode
  - Median
  - Merits & Demerits of Median
  - Variance

2.2 Data Visualization

- BAR Graph
- Pie Chart
- Box Plot
- Scatter Plot
- Histograms
- Bimodal & Multimodal Histograms
- Frequency Chart
- Line Charts

2.3 Probability Basics

- Notation and Terminology
- Unions and Intersections
- Conditional Probability and Independence

2.4 Probability Distribution

- Random Variable
- Probability Distributions
- Probability Mass Function
- Parameters vs. Statistics
- Binomial Distribution
- Poisson Distribution
- Normal Distribution
- Standard Normal Distribution
- Central Limit Theorem
- Cumulative Distribution function

2.5 Probability Distributions Sampling

- Random Sampling
- Systematic Random Sampling
- Stratified Random Sampling
- Cluster Random Sampling

2.6 Inferential Statistics

- Hypothesis Testing
  - Null Hypothesis
  - Alternate Hypothesis
  - Level of Significance
  - P-Value, Normality
  - Decision Criteria
- Tests of Hypothesis
  - Large Sample Test
  - Small Sample Test
  - One Sample: Testing Population Mean
  - Hypothesis in One Sample z-test
  - Two Sample: Testing Population Mean
  - One Sample t-test
  - Two Sample t-test
  - Paired t-test
  - Hypothesis in Paired Samples t-test
  - Chi-Square test
  - Hypothesis in Chi-Square test
  - F test, Hypothesis in F test

Course Outline: DATA SCIENCE MASTER (100 Hours)

2.7 ANOVA (Analysis of Variance)

- Hypothesis in Analysis of Variance
- General setup of ANOVA
- Non paremaetric Test and Parametric Test

3. Analytic Techniques using R

3.1 Introduction to R Programming

- When and Why to use R for Analytics
- Types of Objects in R
- Naming Conventions in R
- Creating Objects in R
- Data Structure in R
- Matrix, Data Frame, String, Vectors
- Understanding Vectors & Data input in R
- Lists, Data Elements
- Creating Data Files using R
- Importing Data Files from other sources.
- Know your Data

3.2 Data Manipulation & Exploration in R

- Sorting Data
- Sub-setting Data
- Selecting (Keeping) Variables
- Excluding (Dropping) Variables
- Selecting Observations and Selection using Subset Function
- Merging Data
- Adding Rows
- Data Type Conversion
- Built-In Numeric Functions
- Built-In Character Functions
- User Built Functions
- Control Structures
- Loop Functions
- Outlier & Missing Values

3.3 Hand on in R

- Basic Statistics & Data Visualization in R
- Probability Distributions in R
- Tests of Hypothesis using R
- Analysis of Variance Using R

4.Analytic Techniques using Python

4.1 Basics of Python Language

- When and Why to use Python for Analytics
- Introduction & Installation of Python
- Python Syntax
- Strings
- Lists and Dictionaries
- Loops
- Regular Expressions

4.2 Introduction to Pandas

- Selecting data from Pandas DataFrame
- Slicing and dicing using Pandas
- GroupBY / Aggregate
- Strings with Pandas
- Cleaning up messy data with Pandas
- Dropping Entries
- Selecting Entries

4.3 Data Manipulation using Pandas

- Data Alignment
- Sorting and Ranking
- Summary Statistics
- Missing values
- Merging data
- Concatenation
- Combining DataFrames
- Pivot
- Duplicates
- Binning

4.4 Scientific Libraries in Python

- Numpy
- Scikit-Learn

5.Machine Learning

5.1 Fundamentals of Machine Learning

- Overview & Terminologies
- What is Machine Learning?
- Why Learn?
- When is Learning required?
- Data Mining
- Application Areas and Roles

Course Outline: DATA SCIENCE MASTER (100 Hours)

- Types of Machine Learning
- Supervised Learning
- Unsupervised Learning
- Reinforcement learning

5.2 Machine Learning Concepts & Terminologies

- Steps in developing a Machine Learning application
- Key tasks of Machine Learning
- Modelling Terminologies
- Learning a Class from Examples
- Probability and Inference
- PAC Learning
- Noise
- Noise and Model Complexity
- Triple Trade-Off
- Association Rules
- Association Measures
- Sample Algorithms

6.Simple Linear Regression

- Correlation
- Regression
- Model Assumptions
- Estimation Process
- Least Squares Method
- The Coefficient of Determination
- Correlation and Regression Using R & Python
- Simple Linear Regression Assignments

7.Multiple Regression Analysis

- Introduction
- Design Requirements
- Assumptions
- Independence
- Normality, Homoscedasticity, Linearity
- Multiple Regression.
- Formal Statement of the Model
- Estimating parameters of the model
- F-test for the overall fit of the model
- Multiple regression model Building
- Selecting the best Regression equation
- Examples/Use Cases
- Interpreting the Final Model
- Multicollinearity and its Diagnostics
- Examples/Use Cases
- Qualitative Independent Variables
- Indicator variables
- Interpretation of Regression Coefficients
- Examples/Use Cases
- Regression Diagnostics and Residual Analysis
- Multiple Linear Regression Using R & Python
- Multiple Regression Assignment

8.Logistic Regression Analysis

- Theory Behind Logistic Regression
- Assessing the Model and Predictors
- When and Why do we Use Logistic Regression?
- Binary
- Multinomial
- Interpreting Logistic Regression
- Assumptions
- Sample size requirements
- The logistic function & Interpretation
- Methods for including variables
- Computational method
- Logistic Regression Model using R & Python
- Logistic Regression Assignment

9.Maximum Likelihood Estimation

- Bernoulli distribution
- Multinomial distribution
- Gaussian distribution
- Assessing the Model
- Assessing Changes in Models
- Assessing Predictors
- Methods of Regression
- Complete Separation
- Overdispersion
- MLE using Python

Course Outline: DATA SCIENCE MASTER (100 Hours)

10.Decision Trees

- Understanding the Concept
- Internal decision nodes
- Terminal leaves.
- Tree induction: Construction of the tree
- Classification Trees
- Entropy
- Selecting Attribute
- Information Gain
- Partially learned tree
- Overfitting
- Causes for over fitting
- Overfitting Prevention (Pruning) Methods
- Reduced Error Pruning
- Decision trees - Advantages & Drawbacks
- Ensemble Models
- Decision Trees using Python
- Decision Trees

11.Random Forests

- Introduction & Motivation
- Ensemble Methods - Bagging, Boosting & Random Forests
- Ensemble Classifiers
- Ensemble Models
- How random forests work?
- Gini Index
- Operation of Random Forest
- Random forest algorithm
- Common variables for random forests
- Random Forest – practical consideration
- Random Forest – Features, Advantages and Disadvantages
- Limitations of random forests
- Random Forest using Python

12.Support Vector Machine

- Problem Definition
- Separating Hyperplanes
- Linear separable case
- Formula for the Margin
- Finding the optimal hyperplane
- The optimization problem
- The Legrangin Dual Problem
- Importance of the Support Vectors
- VC dimension
- Non-linear SVM
- Mapping the data to higher dimension
- The Kernel Trick
- Important Kernel Issues
- Soft Margin
- The primal optimization problem
- The Dual Formulation
- The “C” Problem: Overfitting and Underfitting
- Model selection procedure
- SVM For Multi-class classification
- Applications of SVM
- Advantages & Drawbacks of SVM Model Building Exercises in Python

13.Bayesian Theory

- Axioms of Probability Theory
- Conditional Probability
- Independence
- Joint Distribution
- Baye’s Rule
- Bayesian Categorization
- Generative Probabilistic Models
- Naïve Bayes Generative Model
- Naïve Bayesian Categorization
- Example & Exercises
- Naïve Bayes Classifier using Python

14.K-Nearest Neighbor (K-NN)

- Non-parametric methods
- k-Nearest Neighbor Estimator
- How to Choose k or h
- Strengths and Weaknesses
- K-Nearest Neighbor using Python

Course Outline: DATA SCIENCE MASTER (100 Hours)

15.K Means Clustering

- Parametric Methods Recap
- Clustering
- Direct Clustering Method
- Mixture densities
- Classes v/s Clusters
- Non-Hierarchical Clustering
- K-Means
- Distance Metrics
- K-Means Algorithm
- K-Means Objective
- Color Quantization
- Vector Quantization
- Encoding/Decoding
- Soft Clustering
- Expectation Maximization (EM)
- EM Algorithm
- Feature Selection vs Extraction
- Seed Choice
- Uses of Clustering
- Clustering as Pre-processing
- K-Means Clustering using Python

16.Time Series

- The Art of Forecasting
- Forecasting Approaches
- Qualitative Forecasting Methods
- Quantitative Forecasting Methods
- Time Series & its Components
  - Trend
  - Cyclical
  - Seasonal
  - Irregular
- Smoothing Methods
  - Moving Average Method
  - Exponential Smoothing Method
- Forecast Effect of Smoothing Coefficient
- Linear Time-Series Forecasting Model
- Forecast using Trend Models
- The Linear Trend Model
- Time Series Plot
- Seasonality Plot
- Trend Analysis
- Quadratic Time-Series Forecasting Model
- Quadratic Time-Series Model Relationships
- Quadratic Trend Model
- Exponential Time-Series Forecasting Model
- Exponential Weight
- Exponential Trend Model
- Autoregressive Modeling
- Time Series Data Plot
- Auto-correlation Plot
- Evaluating Forecasts
- Quantitative Forecasting Steps
- Forecasting Guidelines
- Pattern of Forecast Error
- Residual Analysis
- Time Series Using Python

17.Power BI

- Overview of Power BI
- Data Integration with power BI
- Creation of analytics reports in power BI
- Publishing/sharing power BI dashboard on website

18.Natural Language Processing(NLP)

- What is Natural Language Processing?
- What Can Developers Use NLP Algorithms For?
- Open Source NLP Libraries

-NLTL
-Apache OpenNLP
-Stanford NLP

- Natural Language Processing Use Cases

19.Deep Learning

- What is deep learning?
- What is difference between ML, DL and AI?
- Why deep learning is important?
- Use cases of DL.

-Neural Networking
- Image Analytics