Data Science Course Content
By Dr. Vishwanath Rao

Introduction to Data Science
- Need for Data Scientists
- Foundation of Data Science
- What is Business Intelligence
- What is Data Analysis, Data Mining, and Machine Learning
- Analytics vs Data Science
- Value Chain
- Types of Analytics
- Lifecycle Probability
- Analytics Project Lifecycle

Data
- Basis of Data Categorization
- Types of Data
- Data Collection Types
- Forms of Data and Sources
- Data Quality, Changes and Data Quality Issues, Quality Story
- What is Data Architecture
- Components of Data Architecture
- OLTP vs OLAP
- How is Data Stored?

Big Data
- What is Big Data?
- 5 Vs of Big Data
- Big Data Architecture, Technologies, Challenge and Big Data Requirements
- Big Data Distributed Computing and Complexity
- Hadoop
- Map Reduce Framework
- Hadoop Ecosystem

Data Science Deep Dive

- What is Data Science?
- Why are Data Scientists 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 and Data Science Project Life Cycle & Stages
- Map-Reduce Framework
- Hadoop Ecosystem
- Data Acquisition
- Where to source data
- Techniques
- Evaluating input data
- Data formats, Quantity and Data Quality
- Resolution Techniques
- Data Transformation
- File Format Conversions
- Anonymization

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
- Usage of IDE R Studio and various GUI

R Programming Concepts
- The datatypes in R and its uses
- Built-in functions in R
- Subsetting 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

Data Import Techniques in R
- Import data from spreadsheets and text files into R
- Importing data from statistical formats
- Packages installation for database import
- Connecting to RDBMS from R using ODBC and basic SQL queries in R
- Web Scraping
- Other concepts on Data Import Techniques

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
- Some fancy plots
- Use-cases for EDA using R

Data Visualization in R
- Storytelling with Data
- Principle tenets
- Elements of Data Visualization
- Infographics vs Data Visualization
- Data Visualization & Graphical functions in R
- Plotting Graphs
- Customizing Graphical Parameters to improvise the plots
- Various GUIs
- Spatial Analysis
- Other Visualization concepts

Statistics + Machine Learning

Statistics
What is Statistics?

- Descriptive Statistics
- Central Tendency Measures
- The Story of Average
- Dispersion Measures
- Data Distributions
- Central Limit Theorem
- What is Sampling
- Why Sampling
- Sampling Methods
- Inferential Statistics
- What is Hypothesis testing
- Confidence Level
- Degrees of freedom
- what is pValue
- Chi-Square test
- What is ANOVA
- Correlation vs Regression
- Uses of Correlation and Regression

Machine Learning

Machine Learning Introduction

- ML Fundamentals
- ML Common Use Cases
- Understanding Supervised and Unsupervised Learning Techniques
- Clustering
- Similarity Metrics
- Distance Measure Types: Euclidean, Cosine Measures
- Creating predictive models
- Understanding K-Means Clustering
- Understanding TF-IDF, Cosine Similarity and their application to Vector Space Model
- Case study
- Implementing Association rule mining
- Case study
- Understanding Process flow of Supervised Learning Techniques
- Decision Tree Classifier
- How to build Decision trees

- Case study
- Random Forest Classifier
- What is Random Forests
- Features of Random Forest
- Out of Box Error Estimate and Variable Importance
- Case study
- Naive Bayes Classifier
- Case study
- Project Discussion
- Problem Statement and Analysis
- Various approaches to solving a Data Science Problem
- Pros and Cons of different approaches and algorithms
- Linear Regression
- Case study
- Logistic Regression
- Case study
- Text Mining
- Case study
- Sentimental Analysis
- Case study

Python
Getting Started with Python
- Python Overview
- About Interpreted Languages
- Advantages/Disadvantages of Python pydoc
- Starting Python
- Interpreter PATH
- Using the Interpreter
- Running a Python Script
- Python Scripts on UNIX/Windows, Editors and IDEs
- Using Variables
- Keywords
- Built-in Functions
- StringsDifferent Literals
- Math Operators and Expressions
- Writing to the Screen
- String Formatting
- Command Line Parameters and Flow Control

Sequences and File Operations

- Lists
- Tuples
- Indexing and Slicing
- Iterating through a Sequence
- Functions for all Sequences
- Using Enumerate()
- Operators and Keywords for Sequences
- The xrange() function
- List Comprehensions
- Generator Expressions
- Dictionaries and Sets

Deep Dive – Functions Sorting Errors and Exception Handling

- Functions
- Function Parameters
- Global Variables
- Variable Scope and Returning Values. Sorting
- Alternate Keys
- Lambda Functions
- Sorting Collections of Collections, Dictionaries and Lists in Place
- Errors and Exception Handling
- Handling Multiple Exceptions
- The Standard Exception Hierarchy
- Using Modules
- The Import Statement
- Module Search Path
- Package Installation Ways

Regular Expressionist's Packages and Object – Oriented Programming in Python

- The Sys Module
- Interpreter Information
- STDIO
- Launching External Programs
- path directories and Filenames
- Walking Directory Trees
- Math Function
- Random Numbers
- Dates and Times
- Zipped Archives

- Introduction to Python Classes
- Defining Classes
- Initializers
- Instance Methods
- Properties
- Class Methods and Data Static Methods
- Private Methods and Inheritance
- Module Aliases and Regular Expressions

Debugging, Databases and Project Skeletons

- Debugging
- Dealing with Errors
- Using Unit Tests
- Project Skeleton
- Required Packages
- Creating the Skeleton
- Project Directory
- Final Directory Structure
- Testing your Setup
- Using the Skeleton
- Creating a Database with SQLite 3
- CRUD Operations
- Creating a Database Object.

Machine Learning Using Python

- Introduction to Machine Learning
- Areas of Implementation of Machine Learning
- Why Python
- Major Classes of Learning Algorithms
- Supervised vs Unsupervised Learning
- Learning NumPy
- Learning Scipy
- Basic plotting using Matplotlib
- Machine Learning application

Supervised and Unsupervised learning

- Classification Problem
- Classifying with k-Nearest Neighbours (kNN)

Algorithm
- General Approach to kNN
- Building the Classifier from Scratch
- Testing the Classifier
- Measuring the Performance of the Classifier
- Clustering Problem
- What is K-Means Clustering
- Clustering with k-Means in Python and an

Application Example
- Introduction to Pandas
- Creating Data Frames
- GroupingSorting
- Plotting Data
- Creating Functions
- Converting Different Formats
- Combining Data from Various Formats
- Slicing/Dicing Operations.

Scikit and Introduction to Hadoop
- Introduction to Scikit-Learn
- Inbuilt Algorithms for Use
- What is Hadoop and why it is popular
- Distributed Computation and Functional Programming
- Understanding MapReduce Framework Sample MapReduce Job Run

Hadoop and Python
- PIG and HIVE Basics
- Streaming Feature in Hadoop
- Map Reduce Job Run using Python
- Writing a PIG UDF in Python
- Writing a HIVE UDF in Python
- Pydoop and MRjob Basics

IF STUDENTS WANTS TO UNDERSTAND DATA VERY WELL – HADOOP COURSE CAN BE ADDED TO THE CONTENTS.