Course Content – Applied Data Analytics using R

APPLIED DATA ANALYTICS USING R **Duration: 48 Hours + Case Study**

Minimum hours per session	:	*7.5 – 8 hours*
Optimum number of participants	:	*15*
Location	:	*IIT-Madras Research Park, Chennai*
Total duration	:	*48 hours/ (6 days) + Case Study offline*
Total hours of theory	:	*20 hours (~ 36 %)*
Total hours of hands on training	:	*30 hours (~ 54 %)*

Objectives:

- Introduce the participants to the field of data analytics – background and key concepts
- Introduce the participants to problem types in data analytics – possible problem formulation framework
- Introduce and train participants to R programming.
- Introduce the participants to relevant topics with intensive work in R from an application viewpoint
- Introduce the participants to a real-life application of data analytics – a case study approach.

Modules:

Module 0: Pre-course work and assignment- Basics of R programming
Module 1: R Programming
Module 2: Data science – Class of problems
Module 3: Data Preparation
Module 4: Statistical modelling
Module 5: Predictive modelling
Module 6: Machine learning techniques
Module 7: Case study

Pre-requisite:

Bachelor's degree with understanding of basic statistics, matrix algebra and probability

Topics conceptual training structure:

CONCEPTUAL STRUCTURE – SINGLE DAY

Session split	Session split – details	Duration

Part 1	• Very brief explanation of technique/family of techniques in section using an illustrative example
	• Includes:
	  ○ Basic mathematical background
	  ○ Assumptions
	  ○ Interpretation and application
	• Topics similar to taught technique will be provided as self-study material
		2 - 3 hours

Part 2	• Case Study/example/illustration supporting 'part 1'.
	• Will include:
	  ○ Hands on training
	  ○ Introduction & exposure to important R packages
		5 - 6 hours

Conceptual Example for Part 1 explanation:

- *Suppose an example is taken to cover concepts in Module 3 on "Statistical Modelling".*
- *Said example is used to illustrate the "normal statistical distribution".*
- *Existence of other statistical distributions will be mentioned and provided as self-study material*
- *Assignments on other statistical distributions will also be provided requiring mandatory submission.*

Module 0: Pre-course work and assignment- Basics of R programming

Module learning outcomes:

1. Participants will be introduced to basics of R programming. Reading material for the basics will be shared with the them in advance.
2. Participants will have to submit an assignment testing their understanding of the basics of R programming.

Module contents:

- RStudio and its GUI
- Data types and operations
- Different objects in R

Module 1: R programming (6 hours)

Module learning outcomes:

1. Participants will be introduced to intermediate functions in R programming
2. Participants will be able to write their own programs and will learn to use the already existing data analytics modules in R
3. Participants will be able to import data from Excel, etc to the R platform

Module contents:

- Importing and exporting data in R
- Built-in functions
- Data visualization

Module 2: Data science – Class of problems (2 hours Theory)

Module learning outcomes:

1. Participants will learn to apply structured thinking to unstructured problems
2. Participants will be able to categorize and understand various data types
3. Participants will be able to convert imprecise business relevant problem statements to precise data analytic problems
4. Participants will learn the importance of visualization in the data analytics solution process

Module contents:

- Structured thinking and how it can help
- Conceptual understanding of data types
- Importance of quality of data
- Conceptual understanding of solution typology
- Introduction to problem formulation framework
- Impact of visualization
- Business relevant problem statements

Module 3: Data preparation (4 hours – 2 hours theory + 2 hours R training)

Module learning outcomes:

1. Participants will be able to setup appropriate sampling techniques
2. Participants will be able to apply techniques to address outliers and missing values

Module contents:

- Data preprocessing
- Treating outliers and missing values
- Data imputation
- Sampling techniques
- Stratified vs. Cluster sampling

Module 4: Statistical modelling (8 hours – 4 hours' theory + 4 hours R training)

Module learning outcomes:

1. Participants will become well versed in basic probability and statistics concepts
2. Participants will be able to setup hypothesis testing protocols
3. Participants will be able to interpret hypothesis test results

Module contents:

- Descriptive Statistics
- Continuous and discrete random variables and their distributions, statistical intervals
- Hypothesis testing
- One-sided and two-sided tests
- p-values, Type I and Type II errors
- Tests for hypothesis testing

Module 5: Predictive modelling (10 hours – 4 hours' theory + 6 hours R training)

Module learning outcomes:

1. Participants will be able to identify relationships between variables through correlation analysis
2. Participants will be able to develop predictive models between variables
3. Participants will be able to rationalize and assess the fidelity of models that are built

Module contents:

- Least Squares Regression
- Diagnostics and ANOVA
- Model assessment and validation
- Error analysis, iterative model building using domain experience
- Non-parametric testing

Module 6: Machine learning techniques (18 hours – 8 hours' theory + 10 hours R training)

Module learning outcomes:

1. Participants will be able to understand and develop algorithms for classification problems
2. Participants will be able to understand and develop algorithms for function approximation problems
3. Participants will be able to conceptualize novel algorithms

Module contents:

- Introduction to Linear Algebra
- Introduction to Optimisation
- Dimensionality reduction methods
- Classification methods
    - Linear discriminant analysis
    - Quadratic discriminant analysis

  - Logistic regression
  - K-neighborhood
  - Naïve Bayes classifier
  - Decision Trees
- Clustering methods
  - K means clustering
  - Fuzzy C-means clustering
  - Hierarchical clustering

Module 7: Discussion of a Gyan-data Case study 1 hour + (4 hours offline/7 hours offline)

Module learning outcomes:

1. Participants will learn to solve data analytics problems from conceptualization to the final solution and concomitant visualization of the solution

Module contents:

- Participants would be introduced an analytic problem for which data and problem statement would be provided. Participant to prepare a project work in groups and come with innovative solutions, which would be evaluated through a presentation.
- Each group would be allotted a minimum of two intervention sessions lasting two hours each.
