Course Content – Applied Data Analytics using R

APPLIED DATA ANALYTICS USING R **Duration: 40 Hours + Case Study**

Minimum hours per session	:	*7.5 – 8 hours*
Optimum number of participants	:	*15*
Location	:	*IIT-Madras Research Park, Chennai*
Total duration	:	*40 hours/ (5 days) + Case Study offline*
Total hours of theory	:	*18 hours (~ 36 %)*
Total hours of hands on training	:	*22 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 1: Data science – Class of problems
Module 2: R Programming
Module 3: Statistical modelling
Module 4: Data Preparation
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 1: Data science – Class of problems (2 hours)

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 2: R programming (6 hours)

Module learning outcomes:

1. Participants will be introduced to basics of 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:

- RStudio and its GUI
- Data types
- Importing and exporting data in R
- Built-in functions
- Data visualization

Module 3: Statistical modelling (7 hours – 4 hours' theory + 3 hours R training)

Module contents:

- Probability
- Principle of counting
- Conditional probability, Bayes' theorem
- Random variables, expectation
- 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 4: Data preparation (3 hours – 1-hour 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:

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

Module 5: Predictive modelling (7 hours – 4 hours' theory + 3 hours R training)

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 (12 hours – 6 hours' theory + 6 hours R training)

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.

8-Week Schedule:

WEEK 1:

Modules	*: Data Science – Introduction and Class of Problems, R programming*
Classroom Session	*: 8 hours (including theory and hands on training)*
Participant efforts	*: 19 hours (includes assignments, pre-read topics and self-study topics)*
Offline support	*: 1-2 hours (from GDPL staff)*

WEEK 2:

Modules	*: Statistical Modelling*
Classroom Session	*: 7 hours (including theory and hands on training)*
Participant efforts	*: 13 hours (includes assignments, pre-read topics and self-study topics)*
Offline support	*: 1-2 hours (from GDPL staff)*

WEEK 3:

Modules	*: Data Preparation, Predictive modelling*
Classroom Session	*: 7 hours (including theory and hands on training)*
Participant efforts	*: 13 hours (includes assignments, pre-read topics and self-study topics)*
Offline support	*: 1-2 hours (from GDPL staff)*

WEEK 4:

Modules	*:Predictive modelling, Machine Learning*
Classroom Session	*: 8 hours (including theory and hands on training)*
Participant efforts	*: 14 hours (includes assignments, pre-read topics and self-study topics)*
Offline support	*: 2-3 hours (from GDPL staff)*

WEEK 5:

Modules	*: Machine Learning, Case Study*
Classroom Session	*: 8 hours (including theory and hands on training)*
Participant efforts	*: 20 hours (includes assignments, pre-read topics and self-study topics)*
Offline support	*: 3-6 hour (from GDPL staff)*

WEEK 6:

Modules	***: Case Study – problem conceptualization and solution strategy***
Participant efforts	***: 5 or 10 hours (1 or 2 case studies depending on participant being a professional or fresh graduate)***
Offline support	***: 2-3 hours (from GDPL staff)***

Week 7:

Modules	***: Case Study – solution strategy implementation***
Participant efforts	***: 5 or 10 hours (1 or 2 case studies depending on participant being a professional or fresh graduate)***
Offline support	***: 2-3 hour (from GDPL staff)***

Week 8:

Modules	***: Case Study – final queries and presentation***
Participant efforts	***: 5 or 10 hours (1 or 2 case studies depending on participant being a professional or fresh graduate)***
Offline support	***: 2-3 hour (from GDPL staff)***