

What is “R”? R is both a programming language and a free software for analytics and graphics. It runs on pretty much any operating system, and is widely used among statisticians and data miners for data analysis and presentations.

R is a powerful language used widely for data analysis and statistical computing. ... Inclusion of powerful packages in R has made it more and more powerful with time. Packages such as dplyr, tidyr, readr, data.table, SparkR, ggplot2 have made data manipulation, visualization and computation much faster

R is free for everyone to use because it is an open source programming language. Programming codes of R can be used across all platforms like Linux, Windows, and Mac. There are no limits with respect to subscription costs or license management, which makes it easily available to data geeks. Also, you can have free access to the R programming libraries. Nevertheless, there are some commercial libraries meant for enterprises dealing with data in terabytes. Hadoop is a good example.



R Programming

1. Introduction to R :

- What is R, What is S
- R language for statistical programming
- History of R,
- Features of R
- Comparing Tools, Installing R
- R interfaces, R Library

2. Data types and Data Structures:

- Data Types, Attributes Entering Input
- Explicit Coercion
- Data Frame, List
- Factors
- Vectors Objects
- Matrix Dimensions
- Matrix Design
- cbind and rbind

3. Data Management:

- Reading Data
- Writing data
- Reading data files with tables
- Files connection
- Reading lines of Text files

4. Data Manipulation:

- Data sorting
- Find and remove duplicates record
- Cleaning data
- Recoding data
- Merging data
- Slicing of Data
- Merging Data
- Aggregating Data

5. Control Flow structures:

- If, if-else
- for, while
- nested loops
- repeat and break
- next, return
- Apply family (Apply, Lapply, Sapply, Tapply, Mapply)

6. Functions:

- Build in functions
- User Defined functions
- Argument and parameters in a function

7. TidyR (Data Cleaning):

- Clean datasets according to standards
- Date and Times
- Pattern Matching
- Regular Expression

8. Plotting using R:

- Basic plotting and Advance plotting (ggplot2)
- Graphical Parameters
- Combining Plot
- Density Plot, Histogram
- Dot Plot, Bar Plots, Line charts, Pie charts
- Box plot, Scatter Plot
- Saving plots

9. Project based on R:

- Twitter Sentiment analysis

Data Science & Analytics

1. What is Machine Learning?

- Supervised Learning
- Unsupervised Learning
- Introduction to Predictive modelling
- Introduction to Different Phase of Machine Learning
- Types of Analytics
- Descriptive Analysis
- Predictive Analysis
- Prescriptive Analysis

2. Data Exploration for modeling:

- Identifying any problems with the data
- Identify missing data
- Identify outlier data
- Univariate analysis
- Bivariate analysis

3. Measure of Dispersion:

- Standard Deviation
- Range
- Interquartile range

4. Measure of central tendency:

- Mean
- Median
- Mode

5. Introduction to Probability:

- Concept of Probability
- Probability mass function
- Random Variables-Discrete and Continuous
- Binomial Distribution
- Poisson Distribution
- Normal Distribution
- Kurtosis
- Shapes (Skewness)
- Central limit theorem

6. Introduction to Parametric and Non Parametric tests:

- Z test
- T test
- Chi square Test

7. Linear Regression: Solving regression problems:

- Introduction/Applications
- Assumptions of linear Regression
- Building Linear Regression Model
- Understanding standard metrics (Variable signification, R-square/adjusted R-square)
- Assess the overall effectiveness of the model
Validation of models (Rerunning Vs. Scoring)

8. Logistic Regression: Solving classification problems:

- Linear Regression vs. Logistic Regression vs. Generalized Linear Regression
- Assumptions of Logistic Regression
- Binary Logistic Model
- Multinomial Logistic Regression
Checking and improving accuracy using ROC curve, Precision, Recall, Confusion Matrix.

9.KNN:

- What is KNN & Applications?
- KNN for missing treatment
- KNN for solving regression problems
- KNN for solving classification problems
- Validating KNN model
- Model fine tuning with hyper parameters

10. Decision Trees

- Decision Trees - Introduction
- Types of Decision Tree
- Growing and Pruning tree
- Entropy, Information Gain, Gini Index
- Overfitting - Best Practices to avoid

11. Support Vector Machines

- Support Vector Regression
- Support Vector classifier
- Hard Margin
- Soft Margin
- Kernel functions
- Interpretation of Outputs and Fine tune the models with hyper parameters

12. Ensemble Learning

- Concept of Ensembles
- Methods of Ensembling
- Stacking
- Bagging
- Boosting
- Bagging Algorithms
- Random Forest
- Boosting Algorithms
- Gradient Boosting
- Adaboost
- XGBoost

13. Unsupervised Learning

- Dimension Reduction using PCA
- Clustering Algorithms
- K- Mean clustering
- Hierarchical clustering
- Project/ Case Studies(Data Science):
Will be based on the above Techniques.

Fee: 17,500 RS/-

Duration: 3 Months