iTrainAsia™

DATA SCIENCE

BIG DATA IS THE FOUNDATION OF ALL CURRENT MEGATRENDS, FROM SOCIAL TO MOBILE TO THE CLOUD TO GAMING

(!) The future belongs to people who can't transform data into strategic business decisions and value-driven products. Data Science is a field of Big Data which seeks to provide meaningful information from large amounts of complex data using various tools, algorithms and machine learning principles.

(i) Worldwide revenues for Big Data and business analytics will grow from $130.1 billion in 2016 to more than $203 billion in 2020, at a Compound Annual Growth Rate (CAGR) of 11.7%.
Source: IDC

Certified R Programmer for Data Science

Duration: Beginner to Intermediate (5 Days)

Course Overview

R is a programming language that is well-known for its power in statistical computing. The use of R in Data Science enables insights from data to be extracted, and these insights allow companies to get ahead of their competitors. This course provides an introduction to the fundamentals of R language, with a specific focus on how it can be used in Data Science.

You'll gain knowledge on how to gather data, and what you can do with it, starting from reading and cleansing, to manipulation and visualisation. You'll also be exposed to a wide range of topics including Big Data and data analytics lifecycle, exploratory data analysis and Shiny R package.

Prerequisites

All participants should have a basic statistical knowledge with some experience in programming but no specific language is required for this course.

Who Should Attend

This workshop is intended for individuals who are interested in learning Data Science, or who want to begin their career as a data scientist.

Exam Format

The Certified R Programming for Data Science certification exam duration is 2 hours, consisting of 50 Multiple-Choice Questions, with a passing score of 70%. You will receive a professional Certified R Programming for Data Science certification upon passing the exam.

Learning Outcomes

Upon completion of this course, you will be able to:

- Understand R language fundamentals, including basic syntax, variables, and types.
- Create functions and use control flow.
- Read and write data in R.
- Work with data in R.
- Create and customise visualisations using ggplot2.
- Perform predictive analytics using R.

Course Outline
Beginner to Intermediate (5 Days)

Day 1

Module 1: Introduction to Big Data Analytics

- What is Data?
- Why Data Collection is Important
- Types of Data
- What is Data Science?
- Characteristics of Big Data - The Three V's of Big Data
- Big Data Analytics and its Types

Module 2: Data Analytics Lifecycle

- Data Analytics Lifecycle Overview
- Detailed Explanation on Data Analytics Lifecycle

Module 3: Basic Programming Terminologies

- Variables
- Constants
- Keywords
- Comments
- Syntax

Module 4: Getting Started with R

- What is R?
- Install R and RStudio
- Explore RStudio Interface (With Lab Exercises)

Module 5: Data Types in R

- Numbers
- Strings
- Vectors
- Matrix
- Arrays

Day 2

Module 5: Data Types in R

- Data Frames
- Lists
- Factor (With Lab Exercises)

Module 6: Control Structures and Functions in R

- Conditional Statements
- Looping Statements
- Operators
- Functions Syntax
- Scoping Rules
- Subsetting
- Apply Functions (lapply, sapply, vapply)
- Debugging Tools
- Split Function

Module 7: Dealing with Date and Time in R

- Date Time Representation
- Date Time Arithmetic
- Date Time Comparison

Day 3

Module 8: Data Gathering

- Reading Data from CSV File
- Reading Data from JSON File
- Reading Data from XML File
- Reading Data from Web

Module 9: Data Cleansing and Exploration

- Extract, Transform and Load (ETL)
- Data Cleansing
- Aggregation, Filtering, Sorting, Joining
- Dealing with Missing Data
- Selecting Columns and Rows
- Data Wrangling
- Summarise and Group By

Module 10: Simulation and Profiling

- Random Sampling
- Generate Random Numbers
- R Profiler

Module 11: Data Visualisation

- What Is Visualisation?
- Need of Visualisation
- Types of Visualisation
- How to Handle the Properties for Chart Creation
- Activity

Day 4

Module 12: Getting Deeper into Data Visualisation

- Scatter Plots
- Boxplots
- Bar Charts
- Pie Charts
- Histograms

Module 13: Creating Graphs with ggplot

- Getting Started with ggplot
- Mapping Color, Shape and Size
- Creating Attractive Color Scheme
- Creating Bar Charts
- Creating Box Plots

Module 14: Advance Graphs in ggplot

- Correlation
- Deviation
- Ranking
- Distribution
- Composition
- Time Series Plots
- Groups
- Spatial

Day 5

Module 15: Shiny R Package

- Introduction
- How to Build a Simple Shiny Module?

Final Evaluation (Exam)

Certified R Programmer for Data Science
Duration: Advanced (4 Days)

Course Overview

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. In this course, we will first learn about various statistical functions used for data analysis and then explain about different types of Machine Learning algorithms using R programming language. We will also learn how to build different types of Machine Learning models based on different types of data. Most industries working with large amounts of data have recognised the value of Machine Learning technology.

By gleaning insights from this data (often in real time), organisations are able to work more efficiently or gain an advantage over competitors.

Prerequisites

All participants should have basic knowledge of artificial intelligence. Participants should be familiar with the concepts of data importing, data exploration and visualisation using R programming. If not, we recommend to complete the Basic to Intermediate course before starting this course.

Who Should Attend

This course is intended for individuals who have basic analytical knowledge on how to handle the large amount of data using R.

Exam Format

The Certified R Programming for Data Science certification exam duration is 2 hours, consisting of 50 Multiple-Choice Questions, with a passing score of 70%. You will receive a professional Certified R Programming for Data Science (Advanced) certification upon passing the exam.

Learning Outcomes

Upon completion of this course, you will be able to:

- Understand R language fundamentals, including basic syntax, variables, and types.
- Create functions and use control flow.
- Read and write data in R.

- Work with data in R.
- Create and customise visualisations using ggplot2.
- Perform predictive analytics using R.

Course Outline
Advanced (4 Days)

Day 1

Module 1: Statistical Data Analysis

- Analysing Mean, Median, Mode
- Variability
- Distributions
- Asymptotics
- Confidence Intervals
- Hypothesis Testing
- P-Values
- Bivariate Correlation
- Autocorrelation Test - Durbin-Watson Test, Newey-West Estimator
- Outlier Test - Cook's Distance, Studentised Residuals
- Normality Test - Kolmogorov-Smirnov, Cramer-von-Mises, Anderson Darling, Jarque-Bera Test

- T-Test
- Non Parametric Testing-Wilcoxon, Mann-Whitney, Kruskal-Wallis
- Chi-Squared Test
- Stationarity Test - Augmented Dickey-Fuller Test, Seasonal Augmented Dickey-Fuller Test
- Shapiro Wilk, Mobility Matrix
- Multicollinearity Test - Pearson's Correlation - Variance Inflation Factor
- Linearity Test
- Heteroscedasticity Test - White Test, Breusch-Pagan Test
- Regression Test - Out of Time Test

Module 4: Regression

- Linear Regression
- Multiple Regression
- Lab Exercise

Module 5: Logistic Regression

- Introduction to Logistic Regression
- Logistic Regression in R
- Lab Exercise

Module 6: Classification Algorithms

- Support Vector Machines
- Random Forest

Module 7: Naive Bayesian Classifiers

- Introduction to Naive Bayes Algorithm
- Naive Bayes Algorithms Working Principles
- Applications of Naive Bayes Algorithm
- Pros and Cons of Using Naive Bayes
- Steps to Build a Basic Naive Bayes Model

Day 2

Module 2: Machine Learning Algorithms and Its Types

- What is Machine Learning?
- Types of Machine Learning Algorithms

Module 3: Data Pre-Processing

- Understanding the Collected Data with Statistics
- Understanding Data with Visualisation
- Data Preparation

Day 3

Module 8: Decision Tree in R

- Introduction
- Types of Decision Tree
- Regression Trees vs Classification Trees
- Algorithm for Decision Tree
- Advantages and Disadvantages
- Tree-Based Model vs Linear Model

Module 9: Time Series Analysis

- Introduction
- Reading Time Series Data
- Plotting Time Series
- Decomposing Time Series
- Forecasts Using Exponential Smoothing
- ARIMA Models

Module 10: Text Mining

- Introduction to Text Analysis
- Reading the Data
- Create a Corpus and Term Document Matrix
- Sentiment Analysis
- Wordclouds
- N-gram Analysis
- Network Analysis

Module 11: KNN Algorithm

- Finding Nearest Neighbors
- Performance Metrics
- Automatic Workflows

Day 4

Module 12: Neural Network in R

- The Basics of Neural Network
- Fitting Neural Network in R
- Cross-Validation of a Neural Network

Final Evaluation (Exam)

Testimonials

Hear What Our Students Have To Say

"

The class is great and easily understood.
Raudhatul Musfirah Marzuki, Agrobank

Good trainer, Good Training.
Rose Adzreen, VADS

Very recommended.
Haizan bt Husin, Companies Commission of Malaysia

Great trainer as he explained the concept in detail and provided examples.
Sharifah Fazlinda, Assistant Manager, Sime Darby Holdings Berhad

Learned something new to apply to my job. Yes, it is fulfilling!
Ch'ng Ping Choon, Senior Engineering Specialist Manager, Measat Satellite Systems

"

Companies Who Learned From Us
Trusted by Public, Private and Education Sectors

PETRONAS

ISM Insurance Services MALAYSIA BERHAD

VADS
A TM Company

AGRO BANK
Sentiasa di sisi Anda

SSM
SURUHANJAYA SYARIKAT MALAYSIA
COMPANIES COMMISSION OF MALAYSIA