Certificate Course
in
Data Science using R and Python

Total Hours: **120**

Aim:

To equip students with the advanced knowledge and skills on R and Python programming as Statistical programming languages used for data analysis.

Objectives:

- To provide hands on experience on R and Python programming.
- To impart theoretical and practical knowledge on R and Python programming.
- To study advanced aspects of Data Science.
- To familiarize the underlying Statistical concepts in Artificial intelligence and Machine learning.

Outcomes:

Students get an exposure on advanced aspects of Data Science and are equipped to do Statistical Data Analysis using R and Python independently.

Unit I

Revisiting the basics of R(for new comers)(5hrs)-Installing the base R system and R-Studio, Installing and loading packages, Review of the elementary data types in R, Familiarizing with popular packages and functions in R, Writing functions in R, Control statements in R.

Importing data in R(5 hrs)- Importing data from flat files with utils, readr& data.table, importing Excel data, Reproducible Excel work with XLConnect, Importing data from database files, Importing data from web, Importing data from statistical software packages.

Basics of Data Manipulation in R(5hrs)- Cleaning data in R: Introduction and exploring raw data, tidying data, preparing data for analysis, case studies.

Data Manipulation with dplyr package(6hrs)- select, mutate, filter, arrange and summarise verbs, the pipe operator, joining data sets with dplyr: mutating joins, filtering joins and set operations, assembling data, advanced joining, case studies.

Data visualization with ggplot2(6hrs)-Introduction, data, aesthetics, geometries.

Exploratory data analysis in R(6hrs)-Exploring categorical data, Exploring numerical data, Numerical summaries, case studies.

Unit II

Introduction to machine learning(6hrs)-Terminology in machine learning: supervised, unsupervised and reinforcement learning, statistical principles used in machine learning.

Supervised learning methods(11hrs)-Linear regression, decision trees, random forests, classification problem: knn method, logistic regression, support vector machine, Fisher’s linear discriminant.

Unsupervised learning(10hrs)-Clustering: Hierarchical, k-means, Principal component analysis.

Unit III

Basics of Python programming(10hrs)-Installing basic Python IDE, ANACONDA system, installing packages, basic data types in Python, popular packages in Python, Basic syntax of Python.

Intermediate Python(10hrs)- Matlibplot, Dictionaries and Pandas, Logic, Control flow and filtering.

Importing data in Python(5hrs)- Importing from flat files such as .txts and .csvs, from files native to other software such as Excel spreadsheets, Stata, SAS and MATLAB files, from relational databases such as SQLite & PostgreSQL, from the web and from Application Programming Interfaces, also known as APIs.

Manipulating dataframes with pandas(6hrs)-Extracting and transforming data, advanced indexing, rearranging and reshaping data, grouping data.

Cleaning data in Python(5hrs)- Exploring data, tidying data for analysis, combining data for analysis, cleaning data for analysis, case studies.

Unit IV

Exploratory data analysis in Python(6hrs)-Graphical exploratory analysis, numerical exploratory analysis.

Supervised learning with Python(9hrs)- Linear regression, decision trees, random forests, classification problem: knn method, logistic regression, support vector machine, Fisher’s linear discriminant.

Unsupervised learning with python(9hrs)- Clustering: Hierarchical, k-means, Principal component analysis.

Project Work

References

1. Crawley, Michael J. The R book. John Wiley & Sons, 2012.
2. Wickham, Hadley, and Garrett Grolemund. R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.", 2016.
3. Peng, Roger D. R programming for data science. Leanpub, 2016.
4. McKinney, Wes. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. "O'Reilly Media, Inc.", 2012.
5. VanderPlas, Jake. Python data science handbook: essential tools for working with data. "O'Reilly Media, Inc.", 2016.
6. Grus, Joel. Data science from scratch: first principles with python. O'Reilly Media, 2019.
7. Zinoviev, Dmitry. Data Science Essentials in Python: Collect-Organize-Explore-Predict-Value. Pragmatic Bookshelf, 2016.
8. Madhavan, Samir. Mastering Python for Data Science. Packt Publishing Ltd, 2015.