# CERTIFICATE COURSE IN DATA SCIENCE

## (6 Months)


**STATE BOARD OF TECHNICAL EDUCATION AND TRAINING**

**SANKETHIKA VIDYA BHAVAN, MASAB TANK, TELANGANA:HYDERABAD**

# Certificate Course in Data Science

**Duration of the Course** : **6 Months**

**Eligibility** : **Intermediate or its equivalent**

**Total Teaching Hrs** : **250 Hrs**

## Scheme of Instruction and Examination

| Sub Code | Subject Name | Instruction Period/Week | | Total Periods | Scheme of Examination | | | |
|---|---|---|---|---|---|---|---|---|
| | | Theory | Practical | | Duration | Internal Marks | End Exam Marks | Total Marks |
| **THEORY** | | | | | | | | |
| DS-101 | Data Science-I | 03 | - | 50 | 3Hrs | 0 | 100 | 100 |
| DS-102 | Data Science-II | 03 | - | 50 | 3Hrs | 0 | 100 | 100 |
| **PRACTICALS** | | | | | | | | |
| DS-103 | Data Science-I Lab | - | 04 | 75 | 3Hrs | 40 | 60 | 100 |
| DS-104 | Data Science-II Lab | - | 04 | 75 | 3Hrs | 40 | 60 | 100 |
| | TOTAL | 06 | 08 | 250 | | 80 | 320 | 400 |

# Certificate Course in Data Science

| | | |
|---|---|---|
| **Subject Code** | **:** | DS-101 |
| **Subject Name** | : | Data Science-I |
| **Periods/Week** | : | 03 Hrs |
| **Total Periods** | : | 50 Hrs |

## About Course

Data Science Certification Training using R Programming helps you gain expertise in various concepts of Statistics, Time Series and Data Management, Pre Processing and preparation, data transformation. In this Course, you'll be solving real-life case studies on Media, Healthcare, Social Media, and HR.

**Course Objectives:** This course enables the students to

- Acquire Data Gathering, Preprocessing and Data Management Skill
- Apply Various Supervised and Un-Supervised Techniques.
- Create Solutions for various Business Problems.

**Course Outcomes:** On completion of this course the students are able to

- Automate data analysis using R
- Work with real-time data
- Learn tools and techniques for predictive modeling
- Validate data by applying preprocessing techniques
- Explain Time Series and its related concepts

### UNIT-1: Introduction to Data Science

What is Data Science, What does Data Science involve, Era of Data Science , Business Intelligence vs. Data Science , Life cycle of Data Science ,Tools of Data Science.

### UNIT-2: R Programming Essentials

Commands and Syntax, Packages and Libraries, Introduction to Data Types, Data Structures in R – Vectors, Matrices, Arrays, Lists, Factors, Data Frames, Importing and Exporting Data.
Control structures and Functions

## UNIT-3: Descriptive Statistics

Data exploration (histograms, bar chart, box plot, line graph, scatter plot), Qualitative and Quantitative Data, Measure of Central Tendency (Mean, Median and Mode), Measure of Positions (Quartiles, Deciles, Percentiles and Quantiles), Measure of Dispersion (Range, Median, Absolute deviation about median, Variance and Standard deviation), Other Measures: Quartile and Percentile, Interquartile Range

## UNIT-4: Data Acquisition

Gather information from different sources, Internal systems and External systems, Web APIs, Open Data Sources, Data APIs, Web Scrapping, Relational Database access (queries) to process/access data.

## UNIT-5: Data Pre-Processing and Preparation

Data Munging, Wrangling, Plyr packages, Cast/Melt

## UNIT-6: Data Quality and Transformation

Data imputation, Data Transformation (minmax, log transform, z-score transform etc.,), Binning, Classing and Standardization, Outlier/Noise& Anomalies

# Certificate Course in Data Science

| | | |
|---|---|---|
| **Subject Code** | **:** | DS-102 |
| **Subject Name** | : | Data Science-II |
| **Periods/Week** | : | 03 Hrs |
| **Total Periods** | : | 50 Hrs |

## About Course

This course helps you gain expertise in various machine learning algorithms such as regression, clustering, decision trees, random forest and Naïve Bayes. And different classes of machine learning algorithms like supervised, unsupervised and reinforcement algorithms. In this Course, you'll be solving real-life case studies on Media, Healthcare, Social Media, and HR.

**Course Objectives:** This course enables the students to

- Acquire Data Gathering, Preprocessing and Data Management Skill
- Apply Various Supervised and Un-Supervised Techniques.
- Create Solutions for various Business Problems.

**Course Outcomes:** On completion of this course the students are able to

- Automate data analysis using R
- Describe Machine Learning
- Work with real-time data
- Learn tools and techniques for predictive modeling
- Discuss Machine Learning algorithms and their implementation
- Validate Machine Learning algorithms

## UNIT-1: Handling of Text Data

Bag-of-words, Regular Expressions, Sentence Splitting and Tokenization, Punctuations and Stop words, incorrect spellings, Properties of words and Word cloud, Lemmatization and Term-Document TxD computation, Sentiment Analysis (Case Study)

## UNIT-2: Predictive Analytics: Regression

Simple Linear Regression, Multiple Linear Regression, Industry Relevance of Linear Regression and Logistic Regression.

## UNIT-3: Unsupervised Learning: Clustering

Distance measures, Different clustering methods (Distance, Density, Hierarchical) , Iterative distance-based clustering, Dealing with continuous, categorical values in K-Means, Constructing a hierarchical cluster , K-Medoids, k-Mode and density-based clustering , Measures of quality of clustering

## UNIT-4: Supervised Learning: Classification

Naïve Bayes Classifier, K-Nearest Neighbors, Support Vector Machines, Decision Trees, Random Forest.

## UNIT-5: Association Rules Mining and Recommendation Systems

What are Association Rules, Association Rule Parameters, Calculating Association Rule Parameters, Recommendation Engines, Collaborative Filtering, Content Based Filtering.

## UNIT-6: Business Use Cases:

Case Study1: Recommendation Systems and Collaborative filtering
Case Study2: Customer Segmentation and Value

# Certificate Course in Data Science

| | | |
|---|---|---|
| **Subject Code** | **:** | DS-103 |
| **Subject Name** | : | Data Science-I LAB |
| **Periods/Week** | : | 04 Hrs |
| **Total Periods** | : | 75 Hrs |

**Course Objectives:** This course will enable students to

- Acquire Data Gathering, Preprocessing and Data Management Skill
- Apply Various Supervised and Un Supervised Techniques.
- Create Solutions for various Business Problems.

## List of Experiments

1. a) Write an R script, to create R objects for calculator application and save in a specified location in disk

   b) Write an R script to find basic descriptive statistics using summary, str, quartile function on mtcars & cars datasets.

2. Write an R script to find subset of dataset by using subset (), aggregate () functions on iris dataset.
3. Reading different types of data sets (.txt, .csv) from web and disk and writing in file in specific disk location
4. a) Find the data distributions using box and scatter plot.
   b) Plot the histogram, bar chart and pie chart on sample data.

5. Plot the correlation plot on dataset and visualize giving an overview of relationships among data on iris data
6. Analysis of covariance: variance (ANOVA), if data have categorical variables on iris data.

# Certificate Course in Data Science

| | | |
|---|---|---|
| **Subject Code** | **:** | DS-104 |
| **Subject Name** | : | Data Science-II LAB |
| **Periods/Week** | : | 04 Hrs |
| **Total Periods** | : | 75 Hrs |

**Course Objectives:** This course will enable students to

- Acquire Data Gathering, Preprocessing and Data Management Skill
- Apply Various Supervised and Un Supervised Techniques.
- Create Solutions for various Business Problems.

## List of Experiments

1. a) Find out relation between variables that are affecting the admission of a student in a institute based on his or her GRE score, GPA obtained and rank of the student using Logistic Regression with R Programming.
   b) Apply multiple regressions, if data have a continuous independent variable. Apply on above dataset.
2. Apply regression Model techniques to predict the data on above dataset.
3. Install relevant package for classification, Choose classifier for classification problem and evaluate the performance of classifier.
4. Clustering algorithms for unsupervised classification, Plot the cluster data using R visualizations.
5. Write a program to implement k-Nearest Neighbor algorithm to classify the iris data set. Print both correct and wrong predictions.
6. a) Assuming a set of documents that need to be classified, use the naïve Bayesian Classifier model to perform this task. Calculate the accuracy, precision, and recall for your data set

b) Write a program to construct a Bayesian network considering medical data. Use this model to demonstrate the diagnosis of heart patients using standard Heart Disease Data Set.