COURSE CONTENT

DATA SCIENCE LABORATORY								
I Semester: CSE								
Course Code	**Category**	**Hours / Week**			**Credits**	**Maximum Marks**		
BCSE12	**Core**	**L**	**T**	**P**	**C**	**CIA**	**SEE**	**Total**
		0	0	4	2	40	60	100
Contact Classes: Nil	**Total Tutorials: Nil**	**Total Practical Classes: 45**				**Total Classes: 45**		

I. COURSE OVERVIEW:

The course covers the basics of data analytics and R programming. Data analysis with statistical programming, analysis, and visualization of processed data is implemented using R programming. It includes the basics of mathematics, probability and statistical methods, and data communication. It also teaches students how to interpret large datasets and identify patterns to create predictive models.

II. COURSE OBJECTIVES:

The students will try to learn:

I. The concept of R objects.
II. Make use of different types of datasets for analysis using R.
III. Relations among variables using statistical analysis.
IV. The different data models for predictions using R

III. COURSE OUTCOMES:

After successful completion of the course, students will be able to:

CO1 Make Use of the container Data types for display the functional values.
CO2 Demonstrate the reading and writing operations from the web, and disk.
CO3 Illustrate the nature and relationships of data with effective visualizations for exploring the data.
CO4 Analyze the linear data models in data exploration.
CO5 Develop the classification and clustering models to label the data

IV. COURSE CONTENT:

Week 1: R AS CALCULATOR APPLICATION

a. Using with and without R objects on console
b. Using mathematical functions on console
c. Write an R script, to create R objects for the calculator application and save them in a specified location in the disk

Week 2: DESCRIPTIVE STATISTICS IN R

a. Write an R script to find basic descriptive statistics using summary, str, and quartile functions on mtcars and cars data sets.
b. Write an R script to find a subset of the dataset by using subset (), and aggregate () functions on the iris dataset.

Week 3: READING AND WRITING DIFFERENT TYPES OF DATASETS

a. Reading different types of data sets (.txt, .csv) from the web and disk and writing in files in specific disk locations.
b. Reading Excel data sheet in R.
   Reading XML dataset in R.

Week 4: VISUALIZATIONS

a. Find the data distributions using box and scatter plot.
b. Find the outliers using plot.
c. Plot the histogram, bar chart and pie chart on sample data.

Week 5: CORRELATION AND COVARIANCE

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

Week 6: REGRESSION MODEL

Import data from web storage. Name the dataset and now do Logistic Regression to find out the relation between variables that are affecting the admission of a student in an institute based on his or her GRE score, GPA obtained, and rank of the student. Also check whether the model is fit or not.
require (foreign), require(MASS).

Week 7: MULTIPLE REGRESSION MODEL

Apply multiple regressions, if data have a continuous independent variable. Apply on the above dataset.

Week 8: REGRESSION MODEL FOR PREDICTION

Apply regression Model techniques to predict the data on the above dataset.

Week 9: IMPLEMENT SVM CLASSIFICATION MODEL

a. Install relevant package for classification.
b. Choose a classifier for the classification problem.
c. Evaluate the performance of the classifier.

Week 10: IMPLEMENT NAÏVE BAYES CLASSIFICATION MODEL

a. Install relevant package for classification.
b. Choose a classifier for the classification problem.

c. Evaluate the performance of the classifier.

Week 11: IMPLEMENT DECISION TREE CLASSIFICATION MODEL

a. Install relevant package for classification.
b. Choose a classifier for the classification problem.
c. Evaluate the performance of the classifier.

Week 12: IMPLEMENT ENSEMBLE CLASSIFICATION MODEL

a. Install relevant package for classification.
b. Choose a classifier for the classification problem.
c. Evaluate the performance of the classifier.

Week 13: IMPLEMENT K-MEANS CLUSTERING MODEL

a. Clustering algorithms for unsupervised classification.
b. Plot the cluster data using R visualizations.

Week 14: IMPLEMENT DBSCAN CLUSTERING MODEL

a. Clustering algorithms for unsupervised classification.
b. Plot the cluster data using R visualizations

V. REFERENCE BOOKS:

1. Yanchang Zhao, "R and Data Mining: Examples and Case Studies", Elsevier, 1st edition, 2012.

VI. Web References:

1. http://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/
2. http://www.ats.ucla.edu/stat/r/dae/rreg.htm
3. http://www.coastal.edu/kingw/statistics/R-tutorials/logistic.html
4. http://www.ats.ucla.edu/stat/r/data/binary.csv

VII. MATERIALS ONLINE

1. Course template
2. Lab Manual