DATA SCIENCE

I Semester: CSE

Course Code	Category	Hours / Week			Credits	Maximum Marks		
		L	T	P	C	CIA	SEE	Total
BCSB06	Elective	3	-	-	3	30	70	100
Contact Classes:45	Tutorial Classes: Nil	Practical Classes: Nil				Total Classes: 45		

COURSE OBJECTIVES:

The course should enable the students to:

I. Summarize the fundamental knowledge on basics of data science and R programming.
II. Develop programs in R language for understanding and visualization of data using statistical functions and plots.
III. Learn to apply hypotheses and data into actionable predictions.
IV. Understand a range of machine learning algorithms along with their strengths and weaknesses.
V. Able to document and transfer the results and effectively communicate the findings using visualization techniques.

COURSE OUTCOMES (COs):

CO 1: Understand the process and different stages of data science and relevant data descriptions in R language.
CO 2: Illustrate various SQL, NOSQL databases connecting with R and perform correlation and regression analysis.
CO 3: Evaluate different data models and perform clustering analysis.
CO 4: Solve various real time problems using artificial neural networks techniques and comparing different learning algorithms.
CO 5: Explore on various ways to deliver results through documentation and plots of multivariate data and matrix data.

COURSE LEARNING OUTCOMES (CLOs):

1. Understand and develop relevant programming abilities.
2. Understand and intuition of the whole process line of extracting knowledge from data.
3. Equip with the fundamental knowledge on basics of data science and R programming
4. Critically analyze and evaluate variety of NoSQL databases.
5. Develop the ability to build and assess Data-based models.
6. Analyze data analysis and make models using regression analysis.
7. Familiarize with variety of machine learning tasks: clustering, dimensionality reduction, regression and classification.
8. Understand how to formalize practical problems using methods of machine learning.
9. Understand neural networks techniques solve real time problems.
10. Understand the different learning algorithms.
11. Chose an appropriate learning Algorisms to solve particular problems.
12. Based on delivering results make documentation for various results sets.
13. Understand how to plot graphs for multivariate and matrix data.

UNIT-I:	INTRODUCTION	Classes: 10
Data science process, roles, stages in data science project, working with data from files, working with relational databases, exploring data, managing data, cleaning and sampling for modeling; Introduction to R: Introduction to various data types, numeric, character, date, data frame, array, matrix etc., reading and writing datasets, working with different file types .txt, .csv, outliers, R functions and loops; Summary statistics: Summary, str, aggregate, subset, head, tail; Probability distribution.		
UNIT-II	**SQL, NOSQL AND DATA ANALYSIS**	**Classes: 10**
SQL using R, excel and R, introduction to NoSQL, connecting R to NoSQL databases, R with XML, JSON; Correlation analysis; Covariance analysis, ANOVA, forecasting, heteroscedasticity, autocorrelation; Regression analysis: Regression modeling, multiple regression.		
UNIT-III	**DATA MODELS**	**Classes: 08**
Choosing and evaluating models, mapping problems to machine learning, evaluating clustering models, validating models Cluster analysis: K-means algorithm, Naive Bayes memorization methods, unsupervised methods		
UNIT-IV	**ARTIFICIAL NEURAL NETWORKS**	**Classes: 09**
Artificial neural networks: Introduction, neural network representation, appropriate problems for neural network learning, perceptions, multilayer networks and the back propagation algorithm, remarks on the back propagation algorithm; Evaluation hypotheses: Motivation, estimation hypothesis accuracy, basics of sampling theory, a general approach for deriving confidence intervals, difference in error of two hypotheses, comparing learning algorithms.		
UNIT-V	**DELIVERING RESULTS**	**Classes: 08**
Documentation and deployment, producing effective presentations, introduction to graphical analysis, plot() function, displaying multivariate data, matrix plots, multiple plots in one window, exporting graph, using graphics parameters, case studies.		

Text Books:

1. Nina Zumel, John Mount, "Practical Data Science with R", Manning Publications, 1st Edition, 2014.
2. William N. Venables, David M. Smith, "An Introduction to R", Network Theory Limited, 2nd Edition, 2009.
3. Stephen Marsland, "Machine Learning: An Algorithmic Perspective", Taylor & Francis CRC.

Web References:

1. G. Jay Kerns, "Introduction to Probability and Statistics Using R", Youngstown State University, USA, 1st Edition, 2011.
2. William W Hsieh, "Machine Learning Methods in the Environmental Sciences", Neural Networks, Cambridge University Press, 1st Edition, 2009.
3. Chris Bishop, "Neural Networks for Pattern Recognition", Oxford University Press, 1st Edition, 1995.