INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal, Hyderabad -500 043

COMPUTER SCIENCE AND ENGINEERING

COURSE DESCRIPTOR

Course Title	**DATA SCIENCE**				
Course Code	**BCSB06**				
Programme	**M.Tech**				
Semester	**I**	**CSE**			
Course Type	**Elective**				
Regulation	**IARE-R18**				
Course Structure	**Theory**			**Practical**	
	Lectures	**Tutorials**	**Credits**	**Laboratory**	**Credits**
	3	**-**	**3**	**-**	**2**
Course Faculty	**Dr.Chukka Santhaiah, Associate Professor**				

I. COURSE OVERVIEW:

Data Science encompasses the use of mathematics, statistics, and computer science to study and evaluate data. This course is to extract valuable information for use in strategic decision making, product development, trend analysis and forecasting. It includes the processes derived from data engineering, statistics, programming, social engineering, data warehousing, machine learning and natural language processing.

II. COURSE PRE-REQUISITES:

Level	Course Code	Semester	Prerequisites	Credits
UG	**AHSB12**	II	Probability and Statistics	**3**
UG	**ACS005**	IV	Database Management Systems	**3**

III. MARKS DISTRIBUTION:

Subject	SEE Examination	CIA Examination	Total Marks
Foundations of Data Science	70 Marks	30 Marks	100

IV. DELIVERY / INSTRUCTIONAL METHODOLOGIES:

✔	LCD / PPT	✔	Seminars	✔	Videos	✔	MOOCs
✔	Mini Project			✘	Chalk & Talk	✔	Assignments
✘	Open Ended Experiments						

V. EVALUATION METHODOLOGY:

The course will be evaluated for a total of 100 marks, with 30 marks for Continuous Internal Assessment (CIA) and 70 marks for Semester End Examination (SEE).Out of 30 marks Allotted for CIA during the semester, marks are awarded by taking average of two CIA Examinations or the marks scored in the make-up examination.

Semester End Examination (SEE): The SEE is conducted for 70 marks of 3 hours duration. The syllabus for the theory courses is divided into five units and each module carries equal weight age in terms of marks distribution. The question paper pattern is as follows. Two full questions with "either" or "choice" will be drawn from each module. Each question carries 14 marks. There could be a maximum of two sub divisions in a question.

The emphasis on the questions is broadly based on the following criteria:

50 %	To test the objectiveness of the concept.
30 %	To test the analytical skill of the concept.
20 %	To test the application skill of the concept.

Continuous Internal Assessment (CIA):
CIA is conducted for a total of 30 marks (Table 1), with 20 marks for Continuous Internal Examination (CIE), 05 marks for Quiz and 05 marks for Technical Seminar and Term Paper.

Table 1: Assessment pattern for CIA

Component	**Theory**		**Total Marks**
Type of Assessment	**CIE Exam**	**Technical Seminar and Term Paper**	
CIA Marks	25	05	30

Continuous Internal Examination (CIE):
Two CIE exams shall be conducted at the end of the 9th and 17th week of the semester respectively. The CIE exam is conducted for 25 marks of 2 hours duration, consisting of 5 one mark compulsory questions in part-A and 4 questions in part-B. The student has to answer any 4 questions out of five questions, each carrying 5 marks. Marks are awarded by taking average of marks scored in two CIE exams.

Technical Seminar and Term Paper:

Two seminar presentations and the term paper with overview of topic are conducted during I Semester. The evaluation of Technical seminar and term paper is for maximum of 5 marks. Marks are awarded by taking average of marks scored in two Seminar Evaluations.

VI. HOW PROGRAM OUTCOMES ARE ASSESSED:

Program Outcomes (POs)		**Strength**	**Proficiency assessed by**
PO 1	Apply, Analyze a problem, identify and define computing requirements, design and implement appropriate solutions.	2	Seminar & Term paper
PO 2	Solve complex heterogeneous data intensive analytical based problems of real time scenario using state of the art hardware/software tools.	2	Seminar & Term paper

PO 3	Demonstrate a degree of mastery in emerging areas of CSE/IT like IoT, AI, Data Analytics, Machine Learning, cyber security, etc.	2	Seminar & Term Paper
PO 5	Independently carry out research/investigation and development work to solve practical problems	2	Seminar & Term Paper
PO 6	Function effectively on teams to establish goals, plan tasks, meet deadlines, manage risk and produce deliverables	1	Seminar & Term Paper

3 = High; 2 = Medium; 1 = Low

VII. HOW PROGRAM EDUCATIONAL OUTCOMES ARE ASSESSED:

Program Educational Outcomes (PEOs)		Strength	Proficiency assessed by
PEO 1	Independently design and develop computer software systems and products based on sound theoretical principles and appropriate software development skills..	2	Assignments
PEO 2	Demonstrate knowledge of technological advances through active participation in life-long learning.	2	Design Exercise
PEO 3	Accept to take up responsibilities upon employment in the areas of teaching, research, and software development.	2	Seminar
PEO 4	Exhibit technical communication, collaboration and mentoring skills and assume roles both as team members and as team leaders in an organization.	1	Term Paper

3 = High; 2 = Medium; 1 = Low

VIII. COURSE OBJECTIVES:

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.

IX. COURSE OUTCOMES (COs):

CO	Description	CLOs	Course Learning Outcome
CO 1	Understand the process and different stages of data science and relevant data descriptions in R language	CLO 1	Understand and develop relevant programming abilities
		CLO 2	Understand and intuition of the whole process line of extracting knowledge from data
		CLO 3	Equip with the fundamental knowledge on basics of data science and R programming
CO 2	Illustrate various SQL, NOSQL databases connecting with R and	CLO 4	Critically analyze and evaluate variety of NoSQL databases.

	perform correlation and regression analysis	CLO 5	Develop the ability to build and assess Data-based models.
		CLO 6	Analyze data analysis and make models using regression analysis
CO3	Evaluate different data models and perform clustering analysis.	CLO 7	Familiarize with variety of machine learning tasks: clustering, dimensionality reduction, regression and classification
		CLO8	Understand how to formalize practical problems using methods of machine learning
CO 4	Solve various real time problems using artificial neural networks techniques and comparing different learning algorithms.	CLO 9	Understand neural networks techniques solve real time problems
		CLO 10	Understand the different learning algorithms
		CLO 11	Chose a appropriate learning Algorisms to solve particular problems
CO5	Explore on various ways to deliver results through documentation and plots of multivariate data and matrix data	CLO 12	Based on delivering results make a documentation for various results sets
		CLO 13	Understand how to plot graphs for multivariate and matrix data

X. COURSE LEARNING OUTCOMES(CLOs):

CLO Code	CLOs	At the end of the course, the student will have the ability to	PO's Mapped	Strength of Mapping
BCSB06.01	CLO 1	Understand and develop relevant programming abilities	PO 1,PO2	2
BCSB06.02	CLO 2	Understand and intuition of the whole process line of extracting knowledge from data	PO 1, PO 2	2
BCSB06.03	CLO 3	Equip with the fundamental knowledge on basics of data science and R programming	PO 1, PO 2	2
BCSB06.04	CLO 4	Critically analyze and evaluate variety of NoSQL databases.	PO2, PO3	2
BCSB06.05	CLO 5	Develop the ability to build and assess Data-based models.	PO 2,PO3	2
BCSB06.06	CLO 6	Analyze data analysis and make models using regression analysis	PO 1,PO3	1
BCSB06.07	CLO 7	Familiarize with variety of machine learning tasks: clustering, dimensionality reduction, regression and classification	PO 1, PO 3 &PO5	2
BCSB06.08	CLO 8	Understand how to formalize practical problems using methods of machine learning	PO3, PO5	2
BCSB06.09	CLO 9	Understand neural networks techniques solve real time problems	PO 2, PO3	2
BCSB06.10	CLO 10	Understand the different learning algorithms	PO 3, PO5	2
BCSB06.11	CLO 11	Chose a appropriate learning Algorisms to solve particular problems	PO 3, PO5	1
BCSB06.12	CLO 12	Based on delivering results make a documentation for various results sets	PO 3, PO6	2
BCSB06.13	CLO 13	Understand how to plot graphs for multivariate and matrix data	PO 3, PO5	1

3= High; 2 = Medium; 1 = Low

XI. MAPING COURSE OUTCOMES LEADING TO THE ACHIEVEMENT OF PROGRAM OUTCOMES

Course Outcomes (COs)	Program Outcomes (PO)				
	PO 1	PO 2	PO 3	PO 5	PO 6
CO 1	2	2			
CO 2	2	2	1		
CO 3		1	2	1	
CO 4		1	1	1	1
CO 5			1	1	1

XII. MAPPING COURSE LEARNING OUTCOMES LEADING TO THE ACHIEVEMENT OF PROGRAM OUTCOMES

Course Learning Outcomes (CLOs)	Program Outcomes (PO)				
	PO 1	PO 2	PO 3	PO 5	PO 6
CLO 1	2	2			
CLO 2	2	2			
CLO 3	2	2			
CLO 4	1	2	2		
CLO 5		1	1		
CLO 6	2	2			
CLO 7	2		2		
CLO 8	2		1		
CLO 9	2			1	2
CLO 10	2				2
CLO 11		1			2
CLO 12				1	1
CLO 13				1	1

3 = High; 2 = Medium; 1 = Low

XIII. ASSESSMENT METHODOLOGIES –DIRECT

CIE Exams	PO1, PO3, PO5	SEE Exams	PO1, PO3, PO5	Seminar and Term Paper	PO1, PO2, PO3, PO5
Viva	-	Mini Project	-	Laboratory Practices	-

XIV. ASSESSMENT METHODOLOGIES -INDIRECT

✔	Early Semester Feedback	✔	End Semester OBE Feedback
✘	Assessment of Mini Projects by Experts		

XV. SYLLABUS

UNIT-I	INTRODUCTION
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-2	**SQL, NOSQL AND DATA ANALYSIS**
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-3	**DATA MODELS**
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-4	**ARTIFICIAL NEURAL NETWORKS**
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-5	**DELIVERING RESULTS**
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.	

XVI. COURSE PLAN:

The course plan is meant as a guideline. Probably there may be changes.

Lecture No	Topics to be covered	Course Outcomes (COs)	Reference
1	Introduction to Data Science, roles and projects, stages in data science project	CO 1	T1:1.1, 1.2
2	working with data from files, working with relational databases,	CO 1	T2:7.1, 7.2
3-4	exploring data, managing data, cleaning and sampling for modeling;	CO 1	T1:1.3, 1.4,1.6
5-6	Importance of R and R programming, Summary statistics, probability distribution	CO 2	T2:1.1, 1.3
7-8	Introduction to R, data types and functions	CO 2	T1:1.1, 1.8
9-11	Data scientist, terminologies, Reporting and analysis,	CO 2	
12-14	types NoSQL, SQL, R, ANOVA	CO 2	T2:11.2, 11.4
15-16	XML, JSON	CO 2	T2:11.6
17-18	Correlation analysis, regression analysis,	CO 2	T2:117
19-20	Regression modeling, multiple regression.	CO 2	T2:11.8
21-22	Data Models, Choosing and evaluating models, mapping problems to machine learning	CO 3	T1:6.1, 6.2
23-24	Evaluating and validating	CO 3	T2:18.3.4, 18.3.4.1
25-26	Cluster analysis, K-means algorithm, Naive Bayes memorization methods, unsupervised methods	CO 3	T3:14.1
27-28	Introduction to Artificial Neural Networks, neural network representation, appropriate problems for neural network learning, perceptions, multilayer networks	CO 4	T3:1.2, 1.3
29-30	Problems and algorithms, propagation algorithm, remarks on the back propagation algorithm	CO4	T3:4.2, 4.6
31-33	Motivation, estimation hypothesis accuracy, basics of sampling theory, a general approach for deriving confidence intervals	CO4	T3:18.3.4, 18.3, 4.1
34-35	Evaluation hypothesis, Learning algorithms	CO 4	T3:18.1
36-37	Documentation and deployment, producing effective presentations, introduction to graphical analysis	CO 5	T1: 8.1
38-39	Plots, matrix plots, multiple plots in one window,	CO 5	T2:12.1, 12.4
40-42	exporting graph, using graphics parameters, case studies	CO 5	T2:12.5, 12.8

XVII. GAPS IN THE SYLLABUS - TO MEET INDUSTRY / PROFESSION REQUIREMENTS:

S No	Description	Proposed actions	Relevance with POs
1	Problem reductions, Polynomial time and intractability	Seminars / Guest Lectures/ NPTEL	PO 1, PO 2& PO 4
2	String matching: Knuth-Morris-Pratt, Boyer-Moore, Edit distance, Longest increasing subsequence, Smith-Waterman algorithm	Seminars / Guest Lectures/ NPTEL	PO 2, PO 5
3	Encourage students to write programs based on the taught algorithms to solve problems	Laboratory Practices	PO 1, PO 3, PO 5

Prepared by:
Dr. Chukka Santhaiah, Associate Professor

HOD, CSE