INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)
Dundigal, Hyderabad -500 043

Computer Science and Engineering

COURSE INFORMATION SHEET

Course Title	**FOUNDATIONS OF DATA SCIENCE**			
Course Code	**BCS001**			
Programme	**M.Tech**			
Semester	**I**			
Course Type	**Core**			
Regulation	**R18**			
Course Structure	**Lectures**	**Tutorials**	**Practicals**	**Credits**
	4	**-**	**-**	**3**
Course Coordinator	**Mr. C Raghavendra, Assistant Professor**			
Course Faculty	**Mr. C Raghavendra, Assistant Professor**			

I. COURSE OVERVIEW:

To meet the challenge of ensuring excellence in engineering education, the issue of quality needs to be addressed, debated and taken forward in a systematic manner. Accreditation is the principal means of quality assurance in higher education. The major emphasis of accreditation process is to measure the outcomes of the program that is being accredited.

In line with this, faculty of Institute of Aeronautical Engineering, Hyderabad has taken a lead in incorporating philosophy of outcome based education in the process of problem solving and career development. So, all students of the institute should understand the depth and approach of course to be taught through this question bank, which will enhance learner's learning process.

II. COURSE PRE-REQUISITES:

Level	Course Code	Semester	Prerequisites	Credits
UG	AHS010	II	Probability and Statistics	4
UG	ACS005	IV	Database Management Systems	4
UG	ACS012	VII	Big Data and Business Analytics	4
UG	ACS014	VIII	Machine Learning	4

III. MARKS DISTRIBUTION

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

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 unit 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 unit. Each question carries 14 marks.

Continuous Internal Assessment (CIA):
CIA is conducted for a total of 30 marks, with 25 marks for Continuous Internal Examination (CIE) and 05 marks for Quiz / Alternative Assessment Tool (AAT).

Continuous Internal Examination (CIE):
The CIE exam is conducted for 25 marks of 2 hours duration consisting of two parts. Part–A shall have five compulsory questions of one mark each. In part–B, four out of five questions have to be answered where, each question carries 5 marks. Marks are awarded by taking average of marks scored in two CIE exams.

Quiz / Alternative Assessment Tool (AAT):
Two Quiz exams shall be online examination consisting of 20 multiple choice questions and are be answered by choosing the correct answer from a given set of choices (commonly four). Marks shall be awarded considering the average of two quizzes for every course. The AAT may include seminars, assignments, term paper, open ended experiments, micro projects, five minutes video and MOOCs.

IV. DELIVERY / INSTRUCTIONAL METHODOLOGIES:

√	CHALK & TALK	√	QUIZ	√	ASSIGNMENTS	√	MOOCs
√	LCD / PPT	√	SEMINARS	√	MINI PROJECT	√	VIDEOS
√	OPEN ENDED EXPERIMENTS						

V. ASSESSMENT METHODOLOGIES – DIRECT

√	CIE EXAMS	√	SEE EXAMS	√	ASSIGNEMNTS	√	SEMINARS
√	LABORATORY PRACTICES	√	STUDENT VIVA	√	MINI PROJECT	√	CERTIFICATION
√	TERM PAPER						

VI. ASSESSMENT METHODOLOGIES – INDIRECT

√	ASSESSMENT OF COURSE OUTCOMES (BY FEEDBACK, ONCE)	√	STUDENT FEEDBACK ON FACULTY (TWICE)
√	ASSESSMENT OF MINI PROJECTS BY EXPERTS		

VII. 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.

VIII . COURSE LEARNING OUTCOMES:

Students, who complete the course, will have demonstrated the ability to do the following:

CBCS212.01	Understand and intuition of the whole process line of extracting knowledge from data
CBCS212.02	Deep insight in one of the specialisations within the program, depending on the study and the choice of the master project
CBCS212.03	Solid knowledge in a broad range of methods based on statistics and informatics and can use these for data management, analysis and problem solving

CBCS212.04	Experience in deriving theoretical properties of methods involved in Data Science
CBCS212.05	Experience in implementation/modification of methods involved in Data Science
CBCS212.06	Describe what Data Science is and the skill sets needed to be a data scientist.
CBCS212.07	Use R to carry out basic statistical modelling and analysis
CBCS212.08	Motivate and explain trade-offs in big data processing technique design and analysis in written and oral form.
CBCS212.09	Understand significance of exploratory data analysis (EDA) in data science. Apply basic tools (plots, graphs, summary statistics) to carry out EDA.
CBCS212.10	Describe the Data Science Process and how its components interact.
CBCS212.11	Apply basic machine learning algorithms (Linear Regression, k-Nearest Neighbors (k-NN), k-means, Naive Bayes) for predictive modelling.
CBCS212.12	Understand basic terms what Statistical Inference means. Identify probability distributions commonly used as foundations for statistical modeling
CBCS212.13	Identify common approaches used for Feature Generation.
CBCS212.14	Create effective visualization of given data (to communicate or persuade)
CBCS212.15	Work effectively in teams on data science projects.

IX. HOW PROGRAM OUTCOMES ARE ASSESSED:

Program Outcomes		Level	Proficiency assessed by
PO 1	Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems (Engineering Knowledge).	H	Assignments
PO 2	Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences (Problem Analysis).	H	Assignments
PO 3	Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations (Design/Development of Solutions).	S	Mini Project
PO 4	Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions (Conduct Investigations of Complex Problems).	S	Open ended experiments /
PO 5	Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations (Modern Tool Usage).	S	Mini Project
PO 6	Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice (The Engineer and Society).	N	---
PO 7	Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development (Environment and Sustainability).	N	---.
PO 8	Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice (Ethics).	N	---
PO 9	Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings (Individual and Team Work).	N	---
PO 10	Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions (Communication).	S	Seminars / Term Paper / 5 minutes video
PO 11	Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.	N	---

Program Outcomes		Level	Proficiency assessed by
PO 12	Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change (Life-long learning).	S	Certifications

N= None S= Supportive H = Highly Related

X. HOW PROGRAM SPECIFIC OUTCOMES ARE ASSESSED:

Program Specific Outcomes		Level	Proficiency assessed by
PSO 1	**Professional Skills:** The ability to research, understand and implement computer programs in the areas related to algorithms, system software, multimedia, web design, big data analytics, and networking for efficient analysis and design of computer-based systems of varying complexity.	H	Lectures, Assignments
PSO 2	**Problem-Solving Skills:** The ability to apply standard practices and strategies in software project development using open-ended programming environments to deliver a quality product for business success.	H	Projects
PSO 3	**Successful Career and Entrepreneurship:** The ability to employ modern computer languages, environments, and platforms in creating innovative career paths, to be an entrepreneur, and a zest for higher studies.	S	Guest Lectures

N - None S - Supportive H - Highly Related

XI. 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 II
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 III
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 IV
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 V
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 Press, 2nd Edition, 2011.

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.
4	Peter Flach, "Machine Learning", Cambridge University Press, 1st Edition, 2012.

XII. COURSE PLAN:

The course plan is meant as a guideline. There may probably be changes.

Lecture No	Topic Outcomes	Topic/s to be covered	Reference
1	**Understand** Data scientist and data engineer	Introduction to Data Science, roles and projects	T1:2.1
2	**Discuss and Work** about R	Importance of R and R programming	T1:2.3
3- 5	**Explain** concepts of statistics	Summary statistics, probability distribution	T1:2.3.1
4	**Discuss** data science	Data scientist, terminologies	T1:7.2,7.3
6-9	**Understanding** analytics and big data	Reporting and analysis, types	T1:10.3.1
10-11	**Discuss** data Science Technology	NoSQL, SQL, R, ANOVA	T1:11.2, 12.1.1,
12-13	**Discuss** database connectivity	XML, JSON	T1:13.3.2, 13.4.1
14	**Discuss** Data Analysis	Correlation analysis, regression analysis	T2:17.1.1, 17.1.3
15-17	**Define** data models	Data Models	T1:18.1, 18.2.1
18-25	**Understand** Data models	Evaluating and validating	T2:18.3.4, 18.3.4.1
26-30	**Discuss** various methods	Cluster analysis	T1:22.12, 19.1.2
31-33	**Define** neural network and its representation	Introduction to Artificial neural Networks	T3:18.4, 18.4.3
34-35	**Discuss** back propagation	Problems and algorithms	T3:23.1.1, 23.1.3
36	**Define** motivation and sampling theory	Evaluation hypothesis	T3:18.3.4, 18.3, 4.1
37-49	**Define** errors and different approaches	Learning algorithms	T2:24.2,28.4
50-55	**Understanding** Neural Networks	Documentation and deployment	T3:24.3.1, 24.3.3,24.3.4
56-60	**Discuss** Graphical analysis	Plots, matrix plots, case studies	T3:24.3.6, 24.3.9

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

S NO	DESCRIPTION	PROPOSED ACTIONS	RELEVANCE WITH POs	RELEVANCE WITH PSOs
1	Machine Learning and Statistics	Seminars / Guest Lectures / NPTEL	PO 1, PO 2, PO 3	PSO 1
2	NoSQL, SQL, XML, ANOVA	Work Shops/ Guest Lectures / NPTEL/ Laboratory Practices	PO 2, PO 3	PSO 1,PSO 2
3	Laboratory Practice on R	Work Shops/ Laboratory Practices	PO 1, PO 3, PO 5	PSO 2,PSO 2

XIV. MAPPING COURSE OBJECTIVES LEADING TO THE ACHIEVEMENT OF PROGRAM OUTCOMES AND PROGRAM SPECIFIC OUTCOMES:

Course Objectives	Program Outcomes												Program Specific Outcomes		
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
I	H	H			S							S	H	S	
II			S		S		H						S	H	
III	S	H							S				H		S

S= Supportive **H = Highly Related**

XV. MAPPING COURSE OUTCOMES LEADING TO THE ACHIEVEMENT OF PROGRAM OUTCOMES AND PROGRAM SPECIFIC OUTCOMES:

Course learning Outcomes	Program Outcomes												Program Specific Outcomes		
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
CBCS212.01	H	S											S		
CBCS212.02	H	S											H		
CBCS212.03		H			H									S	
CBCS212.04			S										H		
CBCS212.05		H	H										S		
CBCS212.06		S			H							S	S		
CBCS212.07			H	S										S	
CBCS212.08		H		S									S		
CBCS212.09		H													
CBCS212.10				S										H	
CBCS212.11	S	S		H										S	
CBCS212.12					H						S		S		
CBCS212.13	S													S	
CBCS212.14					H				S		S			S	
CBCS212.15		S	H						S				S	H	

S= Supportive **H = Highly Related**

XVI. DESIGN BASED PROBLEMS (DP) / OPEN ENDED PROBLEM:

1. Assume you need to generate a predictive model of a quantitative outcome variable using multiple regression. Explain how you intend to validate this model.

2. How to very efficiently cluster 100 billion web pages, for instance with a tagging or indexing algorithm?

3. You have data on the duration of calls to a call center. Generate a plan for how you would code and analyze these data. Explain a plausible scenario for what the distribution of these durations might look like. How could you test (even graphically) whether your expectations are borne out?

Prepared By:
Mr. C Raghavendra, Assistant Professor

HOD, CSE