INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal, Hyderabad -500 043

COMPUTER SCIENCE AND ENGINEERING

COURSE DESCRIPTOR

Course Title	**DATA PREPARATION AND ANALYSIS LABORATORY**				
Course Code	BCSB20				
Programme	M.Tech				
Semester	II	CSE			
Course Type	Core				
Regulation	IARE - R18				
Course Structure	**Theory**			**Practical**	
	Lectures	**Tutorials**	**Credits**	**Laboratory**	**Credits**
	-	-	-	4	2
Course Faculty	Ms. G Sulakshana, Assistant Professor, CSE				

I. COURSE OVERVIEW:

The course covers the basics of data preparation and data cleaning is an inevitable step in statistical analysis. In business environments, it is frequently required to transfer data from databases and perform statistical analysis. Establish a linkage between data marts and statistical packages is an important task which occurs in professional organizations. This course introduces you to the concepts and the techniques to prepare data located in business intelligent data marts for statistical analysis and covers reading, cleaning, pre-analyzing data and visualization.

II. COURSE PRE-REQUISITES:

Level	Course Code	Semester	Prerequisites	Credits
PG	BCSB10	I	Data Science Laboratory	2

III. MARKS DISTRIBUTION:

Subject	SEE Examination	CIA Examination	Total Marks
Data Preparation and Analysis Laboratory	70 Marks	30 Marks	100

IV. DELIVERY / INSTRUCTIONAL METHODOLOGIES:

✔	LCD / PPT	✔	Student viva	✔	Mini Project	✘	Videos
✔	Open Ended Experiments						

V. EVALUATION METHODOLOGY:

Each laboratory will be evaluated for a total of 100 marks consisting of 30 marks for internal assessment and 70 marks for semester end lab examination. Out of 30 marks of internal assessment, continuous lab assessment will be done for 20 marks for the day to day performance and 10 marks for the final internal lab assessment.

Semester End Examination (SEE): The semester end lab examination for 70 marks shall be conducted by two examiners, one of them being Internal Examiner and the other being External Examiner, both nominated by the Principal from the panel of experts recommended by Chairman, BOS.

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

20 %	To test the preparedness for the experiment.
20 %	To test the performance in the laboratory.
20 %	To test the calculations and graphs related to the concern experiment.
20 %	To test the results and the error analysis of the experiment.
20 %	To test the subject knowledge through viva – voce.

Continuous Internal Assessment (CIA):
CIA is conducted for a total of 30 marks (Table 1), with 20 marks for continuous lab assessment during day to day performance, 10 marks for final internal lab assessment.

Table 1: Assessment pattern for CIA

Component	**Laboratory**		**Total Marks**
Type of Assessment	Day to day performance	Final internal lab assessment	
CIA Marks	20	10	30

Continuous Internal Examination (CIE):
One CIE exams shall be conducted at the end of the 16th week of the semester. The CIE exam is conducted for 10 marks of 3 hours duration.

Preparation	**Performance**	**Calculations and Graph**	**Results and Error Analysis**	**Viva**	**Total**
2	2	2	2	2	10

VI. HOW PROGRAM OUTCOMES ARE ASSESSED:

Program Outcomes (POs)		Strength	Proficiency assessed by
PO 1	An ability to analyze a problem, and to identify and define the computing requirements appropriate to its solution.	3	Laboratory practices, student viva
PO 2	Solve complex heterogeneous data intensive analytical based problems of real time scenario using state of the art hardware/software tools	3	Laboratory practices, student viva
PO 7	To engage in life-long learning and professional development through self-study, continuing education, Professional and doctoral level studies.	3	Laboratory practices, Mini project

3 = High; 2 = Medium; 1 = Low

VII. COURSE OBJECTIVES (COs):

The course should enable the students to:	
I	Learn pre-processing method for multi-dimensional data
II	Practice on data cleaning mechanisms
III	Learn various data exploratory analysis
IV	Develop the visualizations for clusters or partitions

VIII. COURSE LEARNING OUTCOMES (CLOs):

CLO Code	CLO's	At the end of the course, the student will have the ability to:	PO's Mapped	Strength of Mapping
BCSB20.01	CLO 1	Analyze various data preprocessing methods on different data sets.		
BCSB20.02	CLO 2	Describe the fundamentals of data cleaning and implement various missing and noisy handling mechanisms.		
BCSB20.03	CLO 3	Gain knowledge to identify appropriate clustering techniques, and develop clusters for given dataset.		
BCSB20.04	CLO 4	Identify the association rule mining techniques, based on the requirements of the problem.		
BCSB20.05	CLO 5	Derive the hypothesis for association rules to discovery of strong association rules.		
BCSB20.06	CLO 6	Understand the concept of transformation techniques for numerical datasets.		
BCSB20.07	CLO 7	Learn various data visualization techniques and use them to solve statistical problems.		
BCSB20.08	CLO 8	Visualize the cluster datasets and convert the clusters into histograms.		
BCSB20.09	CLO 9	Understand hierarchical clustering and solve the problem for the given related datasets.		
BCSB20.10	CLO 10	Understand how scalability clustering done for apriori algorithm.		

3 = High; 2 = Medium; 1 = Low

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

Course Learning Outcomes (CLOs)	Program Outcomes					
	PO1	PO2	PO3	PO4	PO5	PO7
CLO 1	2					
CLO 2			3		2	
CLO 3				3		1
CLO 4	2				2	
CLO 5				3		
CLO 6				2		
CLO 7		3			3	2
CLO 8			3			
CLO 9	1					
CLO 10		3				

3 = High; 2 = Medium; 1 = Low

X. ASSESSMENT METHODOLOGIES – DIRECT

CIE Exams	PO2	SEE Exams	PO 2	Seminar and Term paper	PO 2, PO 3, PO 4	Laboratory Practices	PO 5
Student Viva	PO 5	Mini Project	PO 5	Certification	-		

XI. ASSESSMENT METHODOLOGIES - INDIRECT

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

XII. SYLLABUS

LIST OF EXPERIMENTS	
Week-1	**DATA PRE-PROCESSING AND DATA CUBE**
Data preprocessing methods on student and labor datasets Implement data cube for data warehouse on 3-dimensional data	
Week-2	**DATA CLEANING**
Implement various missing handling mechanisms ,Implement various noisy handling mechanisms	
Week-3	**EXPLORATORY ANALYSIS**
Develop k-means and MST based clustering techniques, Develop the methodology for assessment of clusters for given dataset	

Week-4	ASSOCIATION ANALYSIS
Design algorithms for association rule mining algorithms	
Week-5	**HYPTOTHYSIS GENERATION**
Derive the hypothesis for association rules to discovery of strong association rules; Use confidence and support thresholds.	
Week-6	**TRANSFORMATION TECHNIQUES**
Construct Haar wavelet transformation for numerical data, Construct principal component analysis (PCA) for 5-dimensional data.	
Week-7	**DATA VISUALIZATION**
Implement binning visualizations for any real time dataset, Implement linear regression techniques	
Week-8	**CLUSTERS ASSESSMENT**
Visualize the clusters for any synthetic dataset, Implement the program for converting the clusterinto histograms	
Week-9	**HIERARCHICAL CLUSTERING**
Write a program to implement agglomerative clustering technique ,Write a program to implement divisive hierarchical clustering technique	
Week-10	**SCALABILITY ALGORITHMS**
Develop scalable clustering algorithms ,Develop scalable a priori algorithm	
Reference Books:	
1. Sinan Ozdemir, "Principles of Data Science", Packt Publishers, 2016.	
Web References:	
1. https://paginas.fe.up.pt/~ec/files_1112/week_03_Data_Preparation.pdf 2. https://socialresearchmethods.net/kb/statprep.php 3. https://www.quest.com/solutions/data-preparation-and-analysis/	
SOFTWARE AND HARDWARE REQUIREMENTS FOR 18 STUDENTS: **SOFTWARE:** Open source Weka 3.8, Python **HARDWARE:** 18 numbers of Intel Desktop Computers with 4 GB RAM	

XIII. COURSE PLAN:

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

Week No.	Topics to be covered	Course Learning Outcomes (CLOs)	Reference
1	Data Gathering And Preparation	CLO 1, CLO 2, CLO 3, CLO 4	T2:1.4-1.5 T2:2.1-2.7
2	Parsing and Transformation	CLO 5, CLO 6	T2:3.1-3.5
3	Data Cleaning	CLO 5, CLO 6	T2: 5.2-5.3 T2: 6.1-6.6
4	Heterogeneous and Missing data	CLO 5, CLO 6, CLO 7	T2: 6.7 T2: 8.1-8.8 T2: 11.1- 11.5
5	Data Transformation	CLO 5, CLO 6, CLO 7, CLO 8	T2: 4.1-4.5
6	Exploratory Analysis	CLO 5, CLO 6, CLO 9	T1:7, 10 T2: 6.9 T2:10.1- 10.2

7	Clustering and Association	CLO 5, CLO 6, CLO 7, CLO 13	T2:10.3-10.5
8	Visualization	CLO 5, CLO 6, CLO 7, CLO 13, CLO 14	T2: 12.1- 12.4 T2:2.1-2.2
9	Correlations and connections	CLO 5, CLO 6, CLO 7	T2: 6.1-6.6
10	Hierarchies and networks	CLO 6, CLO 7, CLO 12	T1:8

Prepared by:
Ms. G Sulakshana, Assistant Professor **HOD, CSE**