INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)
Dundigal, Hyderabad -500 043

INFORMATION TECHNOLOGY

COURSE DESCRIPTOR

Course Title	DATA WAREHOUSING AND DATA MINING LABORATORY					
Course Code	AIT102					
Programme	B.Tech					
Semester	VI	CSE	IT			
Course Type	Core					
Regulation	IARE - R16					
Course Structure	**Theory**			**Practical**		
	Lectures	**Tutorials**	**Credits**	**Laboratory**	**Credits**	
	3	1	4	3	2	
Chief Coordinator	Dr. M Madhu Bala, Professor and Head, Dept. of CSE					
Course Faculty	Dr. D Kishore Babu, Associate Professor, Dept. of CSE Ms. M GeethabYadav, Assistant Professor, Dept. of CSE Mr. A Praveen, Assistant Professor, Dept. of IT Mr. Ch Suresh Kumar Raju, Assistant Professor, Dept. of CSE					

I. COURSE OVERVIEW:

This laboratory course provides an exposure on data mining features. It includes, creating data objects for new and existing datasets, explore on statistical and visual insights of data, perform preprocessing, and apply all data mining functionalities like association rule mining, clustering and classifications on various usecases.

II. COURSE PRE-REQUISITES:

Level	Course Code	Semester	Prerequisites
UG	ACS005	IV	Database Management Systems
UG	AHS010	II	Probability and Statistics

III. MARKSDISTRIBUTION:

Subject	SEE Examination	CIAExamination	Total Marks
Data Warehousing and Data Mining Laboratory	70 Marks	30 Marks	100

IV. DELIVERY / INSTRUCTIONAL METHODOLOGIES:

✘	Chalk & Talk	✘	Quiz	✘	Assignments	✘	MOOCs
✔	LCD / PPT	✘	Seminars	✘	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	**Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.	2	Laboratory Practices, viva
PO 2	**Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences	3	Laboratory Practices, viva
PO 3	**Design/development of solutions**: 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.	2	Laboratory Practices
PO 4	**Conduct investigations of complex problems**: 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.	2	Laboratory Practices

Program Outcomes (POs)		Strength	Proficiency assessed by
PO 5	**Modern tool usage**: 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.	3	Laboratory Practices

3 = High; 2 = Medium; 1 = Low

VII. HOW PROGRAM SPECIFIC OUTCOMES ARE ASSESSED:

Program Specific Outcomes (PSOs)		Strength	Proficiency assessed by
PSO 1	**Professional Skills:** The ability to understand, analyze and develop computer programs in the areas related to algorithms, system software, multimedia, web design, big data analytics, and networking for efficient design of computer-based systems of varying complexity.	2	Viva
PSO 2	**Software Engineering Practices**: The ability to apply standard practices and strategies in software service management using open-ended programming environments with agility to deliver a quality service for business success.	-	-
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.	1	Viva

3 = High; 2 = Medium; 1 = Low

VIII. COURSE OBJECTIVES (COs):

The course should enable the students to:	
I	Explore on data mining features.
II	Create and perform preprocessing on new and existing datasets.
III	Generate association rules on transactional data.
IV	Build the data models by using various classification and clustering algorithms.
V	Analyze the data models accuracy by varying the sample size.

IX. COURSE OUTCOMES (COs):

COs	Course Outcome	CLOs	Course Learning Outcome
CO 1	Understand Data Mining concepts and knowledge discovery process	CLO 1	Perform Matrix operations
CO 2	Explore on data insights and preprocessing techniques	CLO 2	Understand the given input dataset with the support of statistical and visual parameters
		CLO 3	Perform preprocessing on dataset to make it complete for analysis.
CO 3	Extract association rules on frequent items in transaction data	CLO 4	Perform association rule mining and generate frequent rules
CO 4	Build and analyze the classification model using various algorithms.	CLO 5	Create classification model using logistic regression
		CLO 6	Create and analyze classification model using K nearest neighbor (KNN).
		CLO 7	Create and analyze classification model using Decision trees
		CLO 8	Create and analyze classification model using support vector Machines(SVM)
CO 5	Perform clustering using partition algorithms	CLO 9	Evaluate the prediction models accuracy and visualize the results
		CLO 10	Perform clustering on the given data using k-means

X. 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
AIT102.01	CLO 1	Perform Matrix operations	PO 1; PSO 1	2
AIT102.02	CLO 2	Understand the given input dataset with the support of statistical and visual parameters	PO1;PO 2; PSO 1	2
AIT102.03	CLO 3	Perform preprocessing on dataset to make it complete for analysis.	PO 1;PO 2; PSO 1	2
AIT102.04	CLO 4	Perform association rule mining and generate frequent rules	PO1;PO2; PO 3; PO 4; PO 5; PSO 1	2
AIT102.05	CLO 5	Create classification model using logistic regression	PO1; PO2; PO 3; PO 4; PO 5	2
AIT102.06	CLO 6	Create and analyze classification model using K nearest neighbor (KNN).	PO1; PO2; PO 3; PO 4; PO 5	2
AIT102.07	CLO 7	Create and analyze classification model using Decision trees	PO1; PO2; PO 3; PO 4; PO 5; PSO 1	2
AIT102.08	CLO 8	Create and analyze classification model using support vector Machines(SVM)	PO1; PO2; PO 3; PO 4; PO 5	2
AIT102.09	CLO 9	Evaluate the prediction models accuracy and visualize the results	PO1; PO2; PO 3; PO 4; PO 5; PSO 1	2
AIT102.09	CLO 10	Perform clustering on the given data using k-means	PO 1; PO 2; PO 3; PO 4; PO 5; PSO 1	2

3 = High; 2 = Medium; 1 = Low

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

(CLOs)	Program Outcomes (POs)												Program Specific Outcomes (PSOs)		
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
CLO 1	3	-	-	-	1	-	-	-	-	-	-	-	3	-	-
CLO 2	3	2			1	-	-	-	-	-	-	-	1	-	-
CLO 3	3	2			1	-	-	-	-	-	-	-	1	-	-
CLO 4	1	3	2	2	3	-	-	-	-	-	-	-	2	-	-
CLO 5	1	3	2	2	3	-	-	-	-	-	-	-		-	1
CLO 6	1	3	2	2	3	-	-	-	-	-	-	-		-	1
CLO 7	1	3	2	2	3	-	-	-	-	-	-	-	2	-	1
CLO 8	1	3	2	2	3	-	-	-	-	-	-	-		-	1
CLO 9	1	3	2	2	3	-	-	-	-	-	-	-	2	-	1
CLO10	1	2	2	3	3	-	-	-	-	-	-	-	2	-	1

3 = High; 2 = Medium; 1 = Low

XII. ASSESSMENT METHODOLOGIES–DIRECT

CIE Exams	PO 1,PO 2, PO 3, PO 4, PO5	SEE Exams	PO 1,PO 2, PO 3, PO 4, PO5	Assignments	-	Seminars	-
Laboratory Practices	PO 1,PO5	Student Viva	PO 1,PO2, PSO1, PSO3	Mini Project	-	Certification	-
Term Paper	-						

XIII. ASSESSMENT METHODOLOGIES-INDIRECT

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

XIV. SYLLABUS

LIST OF EXPERIMENTS	
WEEK-1	**MATRIX OPERATIONS**
Introduction to Python libraries for Data Mining : NumPy, SciPy, Pandas, Matplotlib, Scikit-Learn Write a Python program to do the following operations: Library: NumPy a) Create multi-dimensional arrays and find its shape and dimension b) Create a matrix full of zeros and ones c) Reshape and flatten data in the array d) Append data vertically and horizontally e) Apply indexing and slicing on array f) Use statistical functions on array - Min, Max, Mean, Median and Standard Deviation	
WEEK-2	**LINEAR ALGEBRA ON MATRICES**
Write a Python program to do the following operations: Library: NumPy a) Dot and matrix product of two arrays b) Compute the Eigen values of a matrix c) Solve a linear matrix equation such as $3 * x^0 + x^1 = 9$, $x^0 + 2 * x^1 = 8$ d) Compute the multiplicative inverse of a matrix e) Compute the rank of a matrix f) Compute the determinant of an array	
WEEK-3	**UNDERSTANDING DATA**
Write a Python program to do the following operations: Data set: brain_size.csv Library: Pandas a) Loading data from CSV file b) Compute the basic statistics of given data - shape, no. of columns, mean c) Splitting a data frame on values of categorical variables d) Visualize data using Scatter plot	
WEEK-4	**CORRELATION MATRIX**
Write a python program to load the dataset and understand the input data Dataset : Pima Indians Diabetes Dataset Library : Scipy a) Load data, describe the given data and identify missing, outlier data items b) Find correlation among all attributes c) Visualize correlation matrix	
WEEK -5	**DATA PREPROCESSING – HANDLING MISSING VALUES**
Write a python program to impute missing values with various techniques on given dataset. a) Remove rows/ attributes b) Replace with mean or mode c) Write a python program to Perform transformation of data using Discretization (Binning) and normalization (MinMaxScaler or MaxAbsScaler) on given dataset.	
WEEK -6	**ASSOCIATION RULE MINING- APRIORI**
Write a python program to find rules that describe associations by using Apriori algorithm between different products given as 7500 transactions at a French retail store. Libraries: NumPy, SciPy, Matplotlib, Pandas Dataset: https://drive.google.com/file/d/1y5DYn0dGoSbC22xowBq2d4po6h1JxcTQ/view?usp=sharing a) Display top 5 rows of data b) Find the rules with min_confidence : .2, min_support= 0.0045, min_lift=3, min_length=2	
WEEK -7	**CLASSIFICATION – LOGISTIC REGRESSION**
Classification of Bank Marketing Data The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing	

campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The dataset provides the bank customers' information. It includes 41,188 records and 21 fields. The classification goal is to predict whether the client will subscribe (1/0) to a term deposit (variable y).
Libraries: Pandas, NumPy, Sklearn, Seaborn
Write a python program to

a) Explore data and visualize each attribute
b) Predict the test set results and find the accuracy of the model
c) Visualize the confusion matrix
d) Compute precision, recall, F-measure and support

WEEK-8	CLASSIFICATION - KNN

Dataset: The data set consists of 50 samples from each of three species of Iris: Iris setosa, Iris virginica and Iris versicolor. Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres.
Libraries:import numpy as np
Write a python program to
a) Calculate Euclidean Distance. b) Get Nearest Neighbors c) Make Predictions.

WEEK-9	CLASSIFICATION - DECISION TREES

Write a python program
a) to build a decision tree classifier to determine the kind of flower by using given dimensions.
b) training with various split measures(Gini index, Entropy and Information Gain)
c)Compare the accuracy

WEEK -10	CLUSTERING – K-MEANS

Predicting the titanic survive groups:
The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.
Libraries: Pandas, NumPy, Sklearn, Seaborn, Matplotlib
Write a python program

a) to perform preprocessing
b) to perform clustering using k-means algorithm to cluster the records into two i.e. the ones who survived and the ones who did not.

WEEK -11	CLASSIFICATION – BAYESIAN NETWORK

Predicting Loan Defaulters :
A bank is concerned about the potential for loans not to be repaid. If previous loan default data can be used to predict which potential customers are liable to have problems repaying loans, these "bad risk" customers can either be declined a loan or offered alternative products.
Dataset: The stream named bayes_bankloan.str, which references the data file named bankloan.sav. These files are available from the Demos directory of any IBM® SPSS® Modeler installation and can be accessed from the IBM SPSS Modeler program group on the Windows Start menu. The bayes_bankloan.str file is in the streams directory.

a) Build Bayesian network model using existing loan default data
b) Visualize Tree Augmented Naïve Bayes model
a) Predict potential future defaulters, and looks at three different Bayesian network model types (TAN, Markov, Markov-FS) to establish the better predicting model.

WEEK-12	CLASSIFICATION – SUPPORT VECTOR MACHINES (SVM)

A wide dataset is one with a large number of predictors, such as might be encountered in the field of bioinformatics (the application of information technology to biochemical and biological data). A medical researcher has obtained a dataset containing characteristics of a number of human cell samples extracted from patients who were believed to be at risk of developing cancer. Analysis of the original data showed that many of the characteristics differed significantly between benign and malignant samples.
Dataset: The stream named svm_cancer.str, available in the Demos folder under the streams subfolder. The data file is cell_samples.data. The dataset consists of several hundred human cell sample records, each of which contains the values of a set of cell characteristics.

a) Develop an SVM model that can use the values of these cell characteristics in samples from other patients to give an early indication of whether their samples might be benign or malignant. Hint: Refer UCI Machine Learning Repository for data set.
References:
1. https://www.dataquest.io/blog/sci-kit-learn-tutorial/ 2. https://www.ibm.com/support/knowledgecenter/en/SS3RA7_sub/modeler_tutorial_ddita/modeler_tutorial_ddita-gentopic1.html 3. https://archive.ics.uci.edu/ml/datasets.php
SOFTWARE AND HARDWARE REQUIREMENTS FOR A BATCH OF 24 STUDENTS: **HARDWARE:** Intel Desktop Systems: 24 Nos **SOFTWARE:** Application Software: Python, IBM SPSS Modeler - CLEMENTINE

XV. 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	Write a Python program to perform matrix operations	CLO 1	R1:1
2	Write a Python program to understand the given dataset	CLO 2	R1:1
3	Write a python program to load the dataset and understand the input data	CLO 1, CLO 2, CLO 3, CLO 4	R1:1
4	Write a python program to load the dataset and understand the input data. Find correlations	CLO 1, CLO 2, CLO 3, CLO 4	R1:1;R3:1
5	Write a python program to impute missing values with various techniques on given dataset.	CLO 1, CLO 2, CLO 3, CLO 4	R1:1;R3:1
6	Write a python program to find rules that describe associations by using Apriori algorithm between different products given as 7500 transactions at a French retail store.	CLO 2, CLO 3, CLO 4	R1:1;R3:1
7	Write a python program to a) Explore data and visualize each attribute b) Predict the test set results and find the accuracy of the model (Regression) c) Visualize the confusion matrix d) Compute precision, recall, F-measure and support	CLO 2, CLO 3, CLO 5	R1:1;R3:1
8	Write a python program to a) Calculate Euclidean Distance. b) Get Nearest Neighbors c) Make Predictions.	CLO 2, CLO 3, CLO 6	R1:1;R3:1
9	Write a python program a) to build a decision tree classifier to determine the kind of flower by using given dimensions. b) training with various split measures(Gini index, Entropy and Information Gain) c)Compare the accuracy	CLO 2, CLO 3, CLO 7	R1:1;R3:1
10	Write a python program a) to perform preprocessing b) to perform clustering using k-means algorithm to cluster the records into two i.e. the ones who survived and the ones who did not.	CLO 2, CLO 3, CLO 8, CLO10	R1:1;R3:1
11	a) Build Bayesian network model using existing loan default data b) Visualize Tree Augmented Naïve Bayes model a) Predict potential future defaulters, and looks at three different Bayesian network model types (TAN, Markov, Markov-FS) to establish the better predicting model.	CLO 2, CLO 3, CLO 9	R1:1;R2:1
12	Develop an SVM model that can use the values of these cell characteristics in samples from other patients to give an early	CLO 2, CLO 3, CLO 9	R1:1;R2:1

Week No.	Topics to be covered	Course Learning Outcomes (CLOs)	Reference
	indication of whether their samples might be benign or malignant.		

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

S NO	Description	Proposed actions	Relevance with POs	Relevance with PSOs
1	To improve standards and analyze the concepts.	Seminars	PO 1, PO 4	PSO 1;PSO 3
2	Conditional probability, Sampling distribution, correlation, regression analysis and testing of hypothesis	Seminars / NPTEL	PO 4, PO3	PSO 1; PSO 2
3	Encourage students to solve real time applications.	UCI / Kaggle data repositories	PO 2	PSO 1; PSO 2

Prepared by:

Dr. M MadhuBala, Professor, Dept. of CSE

HOD, IT