DATAWAREHOUSING AND DATAMINING LABORATORY

VI Semester: IT | CSE

Course Code	Category	Hours / Week			Credits	Maximum Marks		
		L	T	P	C	CIA	SEE	Total
AIT102	Core	-	-	3	2	30	70	100
Contact Classes: Nil	Tutorial Classes: Nil	Practical Classes: 36				Total Classes: 36		

OBJECTIVES:

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.

COURSE OUTCOMES(COs):

1. Understand Data Mining concepts and knowledge discovery process
2. Explore on data insights and preprocessing techniques
3. Extract association rules on frequent items in transaction data
4. Build and analyze the classification model using various algorithms.
5. Perform clustering using partition algorithms

COURSE LEARNING OUTCOMES (CLOs):

The students should enable to:

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

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_tut

orial_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