DATA MINING AND WAREHOUSING LAB
(Common to CSE & IT)

			L	T	P	C
III B. Tech. - II Semester						
Course Code: A3CS28			-	-	3	2

COURSE OVERVIEW:

This course helps the students to practically understand a data warehouse, techniques and methods for data gathering and data pre-processing using OLAP tools. The different data mining models and techniques will be discussed in this course.

COURSE OBJECTIVES:

1. To teach principles, concepts and applications of data warehousing and data mining
2. To introduce the task of data mining as an important phase of knowledge recovery process
3. To inculcate Conceptual, Logical, and Physical design of Data Warehouses OLAP applications and OLAP deployment
4. To inculcate fundamental concepts that provide the foundation of data mining
5. Design a data warehouse or data mart to present information needed by management in a form that is usable for management client

COURSE OUTCOMES:

After undergoing the course, Students will be able to understand

1. Design a data mart or data warehouse for any organization
2. Extract knowledge using data mining techniques
3. Adapt to new data mining tools.
4. Explore recent trends in data mining such as web mining, spatial-temporal mining

SYLLABUS

DATA WAREHOUSE LAB
OBJECTIVES

Learn how to perform data mining tasks using a data mining toolkit (such as open source WEKA), Understand the data sets and data pre-processing, Demonstrate the working of algorithms for data mining tasks such as association rule mining, classification, clustering and regression, Exercise the data mining techniques with varied input values for different parameters.

UNIT - I
Explore machine learning tool "WEKA"

A. Explore WEKA Data Mining/Machine Learning Toolkit

- Downloading and/or installation of WEKA data mining toolkit,
- Understand the features of WEKA toolkit such as Explorer, Knowledge Flow interface, Experimenter, command-line interface.
- Navigate the options available in the WEKA
  (ex. Select attributes panel, Preprocess panel, Classify panel, Cluster panel, Associate panel and Visualize panel)
- Study the arff file format
- Explore the available data sets in WEKA.
- Load a data set (ex. Weather dataset, Iris dataset, etc.)
- Load each dataset and observe the following:
  1. List the attribute names and they types
  2. Number of records in each dataset
  3. Identify the class attribute (if any)
  4. Plot Histogram
  5. Determine the number of records for each class.
  6. Visualize the data in various dimensions

UNIT - 2
Perform data preprocessing tasks and Demonstrate performing association rule mining on data sets
- Explore various options available in Weka for preprocessing data and apply unsupervised filters like Discretization, Resample filter, etc. on each dataset
- Load weather. nominal, Iris, Glass datasets into Weka and run Apriori algorithm with different support and confidence values. Study the rules generated.
- Apply different discretization filters on numerical attributes and run the Apriori association rule algorithm. Study the rules generated. Derive interesting insights and observe the effect of discretization in the rule generation process.

UNIT - 3
Demonstrate performing classification on data sets
- Load each dataset into Weka and run 1d3, J48 classification algorithm. Study the classifier output. Compute entropy values, Kappa statistic.
- Extract if-then rules from the decision tree generated by the classifier, Observe the confusion matrix.
- Load each dataset into Weka and perform Naïve-bayes classification and k-Nearest Neighbour classification. Interpret the results obtained.
- Plot RoC Curves
- Compare classification results of ID3, J48, Naïve-Bayes and k-NN classifiers for each dataset, and deduce which classifier is performing best and poor for each dataset and justify.

UNIT - 4
Demonstrate performing clustering of data sets
- Load each dataset into Weka and run simple k-means clustering algorithm with different values of k (number of desired clusters). Study the clusters formed. Observe the sum of squared errors and centroids, and derive insights.
- Explore other clustering techniques available in Weka.
- Explore visualization features of Weka to visualize the clusters. Derive interesting insights and explain.

UNIT - 5
Demonstrate knowledge flow application on data sets
- A. Develop a knowledge flow layout for finding strong association rules by using Apriori,FP-Growth algorithms
- Set up the knowledge flow to load an ARFF file (batch mode) and perform a crossvalidation using J48 algorithm
- Demonstrate plotting multiple ROC curves in the same plot window by using j48 and Random forest tree

Resource Sites
- https://www.cs.waikato.ac.nz/ml/weka
- https://weka.wikispaces.com

Outcomes
- Ability to understand the various kinds of tools.
- Demonstrate the classification clusters and etc. in large data sets

DATA MINING LAB

OBJECTIVES
- To obtain practical experience using data mining techniques on real world data sets.
- Emphasize hands-on experience working with all real data sets.

Task 1: Credit Risk Assessment
Description:

The business of banks is making loans. Assessing the credit worthiness of an applicant is of crucial importance. You have to develop a system to help a loan officer decide whether the credit of a Customer is good, or bad. A bank's business rules regarding loans must consider two opposing factors. On the one hand, a bank wants to make as many loans as possible, Interest on these loans is the banks profit Source. On the other hand, a bank cannot afford to make too many bad loans. Too many bad loans could lead to the collapse of the bank. The bank's loan policy must involve a compromise: not too strict, and not too lenient.

To do the assignment, you first and foremost need some knowledge about the world of credit. You can acquire such knowledge in a number of ways.

1. Knowledge Engineering. Find a loan officer who is willing to talk. Interview her and try to represent her knowledge in the form of production rules.
2. Books. Find some training manuals for loan officers or perhaps a suitable textbook on finance. Translate this knowledge from text form to production rule form.
3. Common sense. Imagine yourself as a loan officer and make up reasonable rules which can be used to judge the credit worthiness of a loan applicant.
4. Case histories. Find records of actual cases where competent loan officers correctly judged when, and when not to, approve a loan application.

The German Credit Data
Actual historical credit data is not always easy to come by because of confidentiality rules. Here is one such dataset, consisting of 1000 actual cases collected in Germany. credit dataset (original) Excel spreadsheet version of the German credit data. In spite of the fact that the data is German, you should probably make use of it for this assignment. (Unless you really can consult a real loan officer!)

A few notes on the German dataset.
- DM stands for Deutsche Mark, the unit of currency, worth about 90 cents Canadian (but looks and acts like a quarter).
- Owns_telephone. German phone rates are much higher than in Canada so fewer people own telephones.
- Foreign_worker. There are millions of these in Germany (many from Turrkey). It is very hard to get German citizenship if you were not born of German parents.
- There are 20 attributes used in judging a loan applicant. The goal is the classify the applicant into one of two categories, good or bad.

Subtasks: (Turn in your answers to the following tasks)
List all the categorical (or nominal) attributes and the real-valued attributes separately by using German credit data set or any other built-in data sets available in weka. (5 marks)

1. What attributes do you think might be crucial in making the credit assessment? Come up with some simple rules in plain English using your selected attributes. (5 marks)
2. One type of model that you can create is a Decision Tree — train a Decision Tree using the complete dataset as the training data. Report the model obtained after training. (10 marks)
3. Suppose you use your above model trained on the complete dataset, and classify credit good/bad for each of the examples in the dataset. What % of examples can you classify correctly? (This is also called testing on the training set) Why do you think you cannot get 100 % training accuracy? (10 marks)
4. Is testing on the training set as you did above a good idea? Why or Why not? (10 marks)
5. One approach for solving the problem encountered in the previous question is using cross-validation? Describe what is cross-validation briefly. Train a Decision Tree again using cross-validation and report your results. Does your accuracy increase/decrease? Why? (10 marks)
6. Check to see if the data shows a bias against "foreign workers" (attribute 20), or "personal-status" (attribute 9). One way to do this (perhaps rather simple minded) is to remove these attributes from the dataset and see if the decision tree created in those cases is significantly different from the full dataset case which you have already done. To remove an attribute you can use the preprocess

tab in Weka's GUI Explorer. Did removing these attributes have any significant effect? Discuss. (10 marks)

7. Another question might be, do you really need to input so many attributes to get good results? Maybe only a few would do. For example, you could try just having attributes 2, 3, 5, 7, 10, 17 (andthe class attribute (naturally)). Try out some combinations. (You had removed two attributes in problem 7. Remember to reload the am data file to get all the attributes initially before you start selecting the ones you want.) (10 marks)

8. Sometimes, the cost of rejecting an applicant who actually has a good credit (case 1) might be higher than accepting an applicant who has bad credit (case 2). Instead of counting the misclassifcatjons equally in both cases, give a higher cost to the first case (say cost 5) and lower cost to the second case. You can do this by using a cost matrix in Weka. Train your Decision Tree again and report the Decision Tree and cross-validation results. Are they significantly different from results obtained in problem 6 (using equal cost)? (10 marks)

9. Do you think it is a good idea to prefer simple decision trees instead of having long complex decision trees? How does the complexity of a Decision Tree relate to the bias of the model? (10 marks)

10. You can make your Decision Trees simpler by pruning the nodes. One approach is to use Reduced Error Pruning – Explain this idea briefly. Try reduced error pruning for training your Decision Trees using cross-validation (you can do this in Weka) and report the Decision Tree you obtain? Also, report your accuracy using the pruned model. Does your accuracy increase? (10 marks)

11. (Extra Credit): How can you convert a Decision Trees into "if-thenelse rules". Make up your own small Decision Tree consisting of 2-3 levels and convert it into a set of rules. There also exist different classifiers that output the model in the form of rules – one such classifier in Weka is rules. PART, train this model and report the set of rules obtained. Sometimes just one attribute can be good enough in making the decision, yes, just one! Can you predict what attribute that might be in this dataset? OneR classifier uses a single attribute to make decisions (it chooses the attñbute based on minimum error). Report the rule obtained by training a one R classifier. Rank the performance of j48, PART and oneR. (10 marks)

Task Resources
1. Mentor lecture on Decision Trees
2. Andrew Moore's Data Mining Tutorials (See tutoals on Decision Trees and Cross Validation)
3. Decision Trees (Source: Tan, MSU) Tom Mitchell's book slides (See slides on Concept Learning and Decision Trees)

Weka resources:
1. Introduction to Weka (html version) (download ppt version)
2. Download Weka
3. Weka Tutorial
4. ARFF format
5. Using Weka from command line

Outcomes
1. Ability to add mining algorithms as a component to the exiting tools
2. Ability to apply mining techniques for realistic data.