Faculty of Science

Course:	CSCI 4030U Big Data Analytics
Lab:	#6
Topic:	Data Mining with Weka
Due Date:	Apr. 2 at 11:59pm

Objective

The objective of this project is to gain hands-on experience in using data mining software to build models from real world data sets and to conduct evaluations of different data mining algorithms.

Software

The software to be used in this assignment is Weka [1]. You can download and install it on your own machine.

[1] http://www.cs.waikato.ac.nz/ml/weka/

You can use "java -Xmx256m -jar weka.jar" to modify the heap size when you invoke weka. You can increase the value of 256m if it is not enough.

Read attached documentation about WEKA

All Weka datasets can be found here:
https://drive.google.com/drive/folders/1UH-cuOnsepo_o1xkM3UyRr01rvvw8FIv?usp=sharing

A guide on how to use Weka's Explorer can be found here:
https://drive.google.com/file/d/1XTgnNRoR23AfuACGZ4iFIJ4LsVzOBgtO/view?usp=sharing

A page documenting the ARFF data format used by Weka can be found here:
https://drive.google.com/file/d/1sROl-8hdB1qBHjyjGj_FmN8hIc8NGLx1/view?usp=sharing

Task 1 (Lab 6)

Consider an attached lymphography data set (lymph.arff) that describes 148 patients with 19 attributes. The last attribute is the class attribute that labels a patient with one of the four categories (normal, metastases, malign_lymph, and fibrosis). Detailed information about the attributes is attached (lymph.txt). The data set is in the ARFF format used by Weka.

Use the following learning methods provided in Weka to learn a classification model from the data set with all the attributes:

- C4.5 (weka.classifier.trees.J48)
- RIPPER (weka.classifier.rules.JRip)

You will be submitting a pdf. For each learning method above, attach a screenshot of the output to the pdf, then research how both of the algorithms work and provide a one paragraph description for each of them.

Task 2 (Lab 6)

You are given a training data (monks-3.train.arff) set and a test data set (monks-3.test.arff) in which each training example is represented by seven **nominal** attributes. The last attribute is the class attribute that labels a training example with one of the two classes (0 and 1). The attribute information is given below:

Attribute	Values
a1	1, 2, 3
a2	1, 2, 3
a3	1, 2
a4	1, 2, 3
a5	1, 2, 3, 4
a6	1, 2
class	0, 1

Use the following learning methods provided in Weka to learn a classification model from the training data set and test the model on the test data set:

- C4.5 (weka.classifier.trees.J48)
- RIPPER (weka.classifier.rules.JRip)
- k-Nearest Neighbor (weka.classifiers.lazy.IBk)
- Naive Bayesian Classification (weka.classifiers.bayes.NaiveBayes)
- Neural Networks (weka.classifiers.functions.MultilayerPerceptron)

You will be submitting a pdf. For each learning method above, report the classification accuracy and confusion matrix of each algorithm on the test data set, then research how the algorithms work and provide a one paragraph description for each of them.