| R•I•T | Rochester Institute of Technology<br>Golisano College of Computing and Information Sciences<br>School of Information |
|---|---|

**Name: ___________________________________________**

# ISTE-600 Lab #2: Classifications using Weka

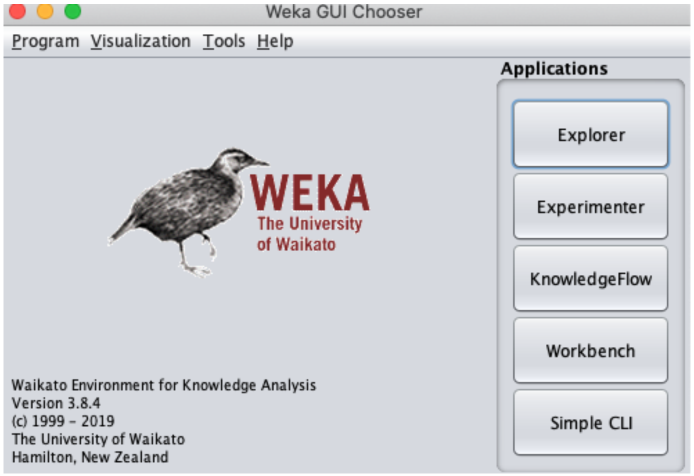
Weka is an open source software for data mining. Weka offers a graphical workbench as well as a command line, and a programming interface (API). It contains dozens of powerful machine learning algorithms and useful visualization tools. Before you continue this lab, please get familiar first with Weka, and especially read the documentation about the Weka Explorer and about Weka's Attribute-Relation File Format. In this lab, you will learn how to read a dataset in ARFF file with class annotations into Weka and apply different classification procedures to the data. Submit 1) ***Lab02_YourName***.model, 2) **Lab02_unknown_YourName.arff and 3) a scan copy of answered this Lab #2** to the Lab #2 Drop Box by Friday, 9/18 @11:59PM.

## Part 1: Data Exploration

**Environment setup:**

Download latest stable version of Weka. Follow the steps below:

1) Browse to URL: http://www.cs.waikato.ac.nz/ml/weka/downloading.html
2) Select a download suitable for your machine.
3) Run and install setup.
4) Now click Start menu → All programs → Weka
5) GUI Chooser appears, click the Explorer button to use the Weka Explorer.



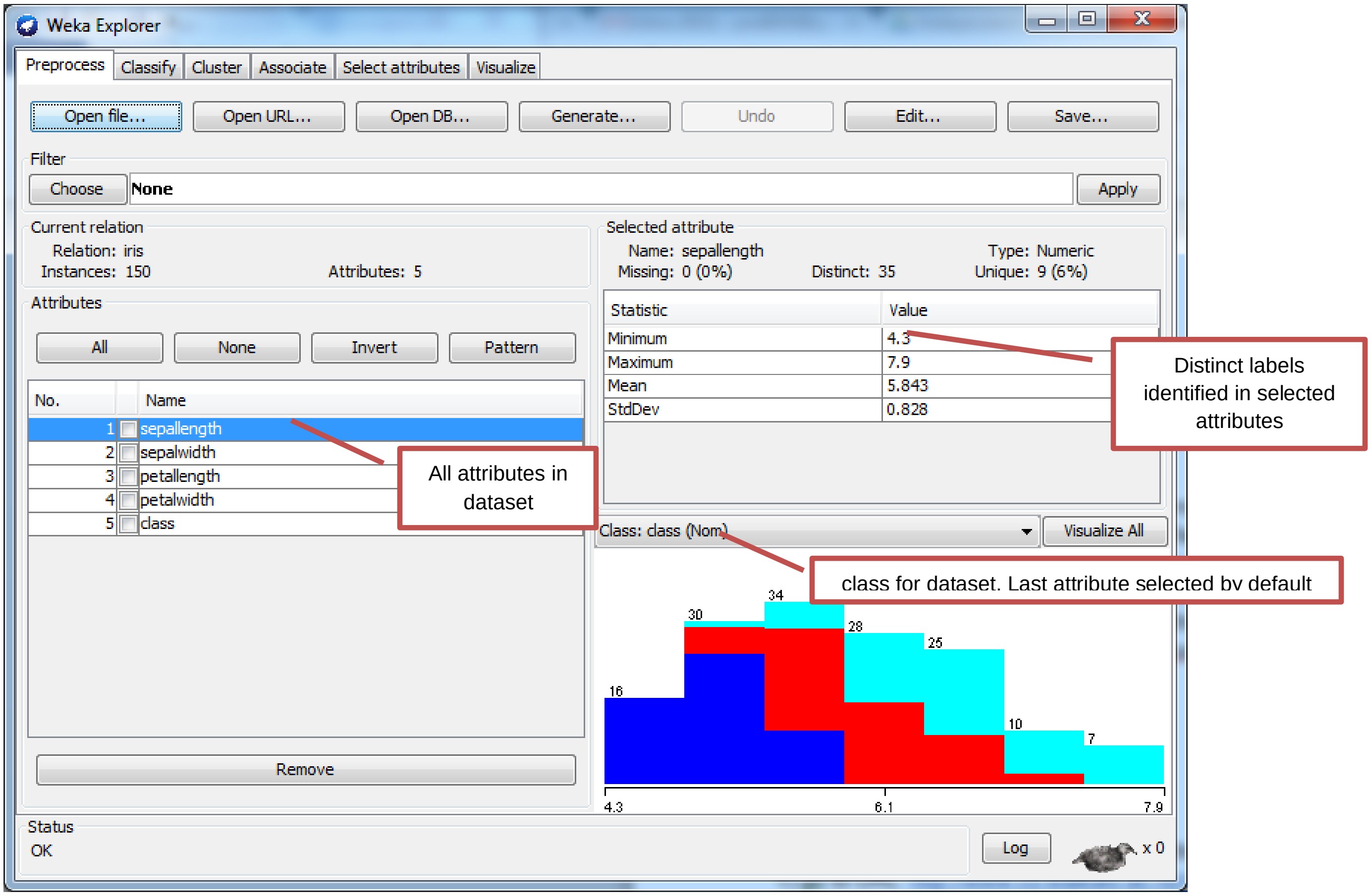
Waikato Environment for Knowledge Analysis
Version 3.8.4
(c) 1999 - 2019
The University of Waikato
Hamilton, New Zealand

## Data preparation and loading:

Weka provides sample datasets for learning purposes. These files can be found under

<Weka installation directory>/data/*iris*.arff.

You can also try to open the file using notepad and check for format and annotations used. In the Preprocess tab, click on open file and select the iris dataset.



Filter: None

Current relation
Relation: iris
Instances: 150
Attributes: 5

| No. | Name |
|---|---|
| 1 | □ sepallength |
| 2 | □ sepalwidth |
| 3 | □ petallength |
| 4 | □ petalwidth |
| 5 | □ class |

Selected attribute
Name: sepallength
Type: Numeric
Missing: 0 (0%)
Distinct: 35
Unique: 9 (6%)

| Statistic | Value |
|---|---|
| Minimum | 4.3 |
| Maximum | 7.9 |
| Mean | 5.843 |
| StdDev | 0.828 |

Class: class (Nom)

Status
OK

Distinct labels identified in selected attributes

All attributes in dataset

class for dataset. Last attribute selected bv default

### Questions (5 points each)

Q1.1 How many instances are there? _________

Q1.2 How many attributes are there? _________

Q1.3 How many possible values does the *class* attribute have? __________

Now open the **contact-lenses** dataset.

Q1.4 How many instances are there?  _________

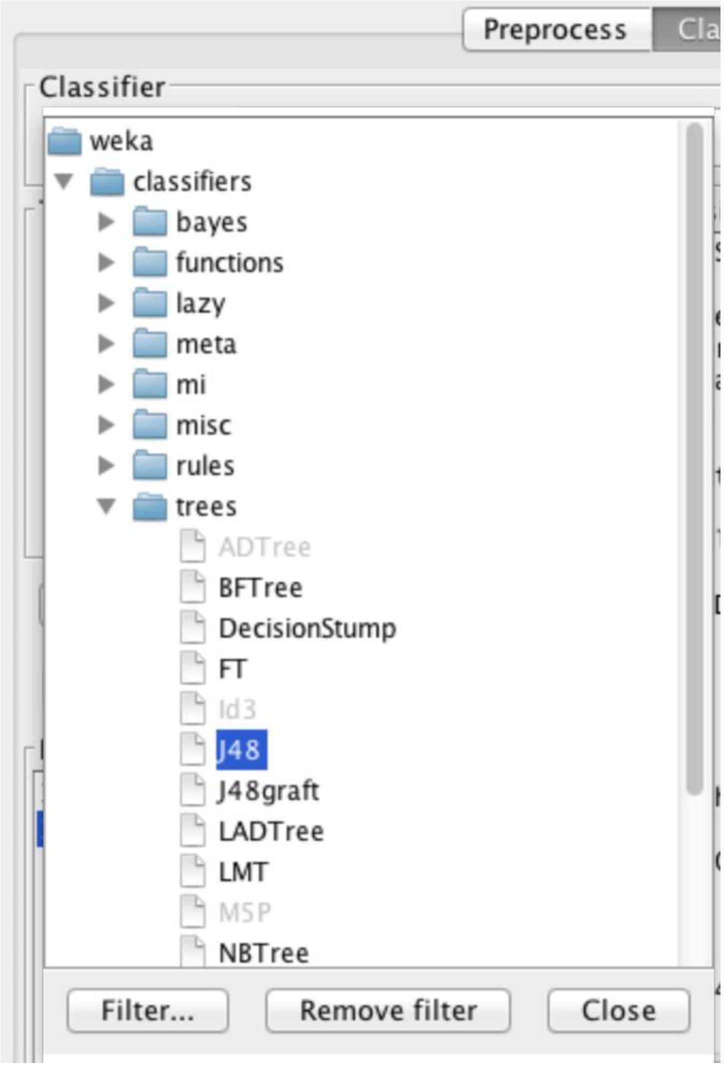
Q1.5 How many attributes are there?  _________

Q1.6 How many possible values are there for the *age* attribute? __________

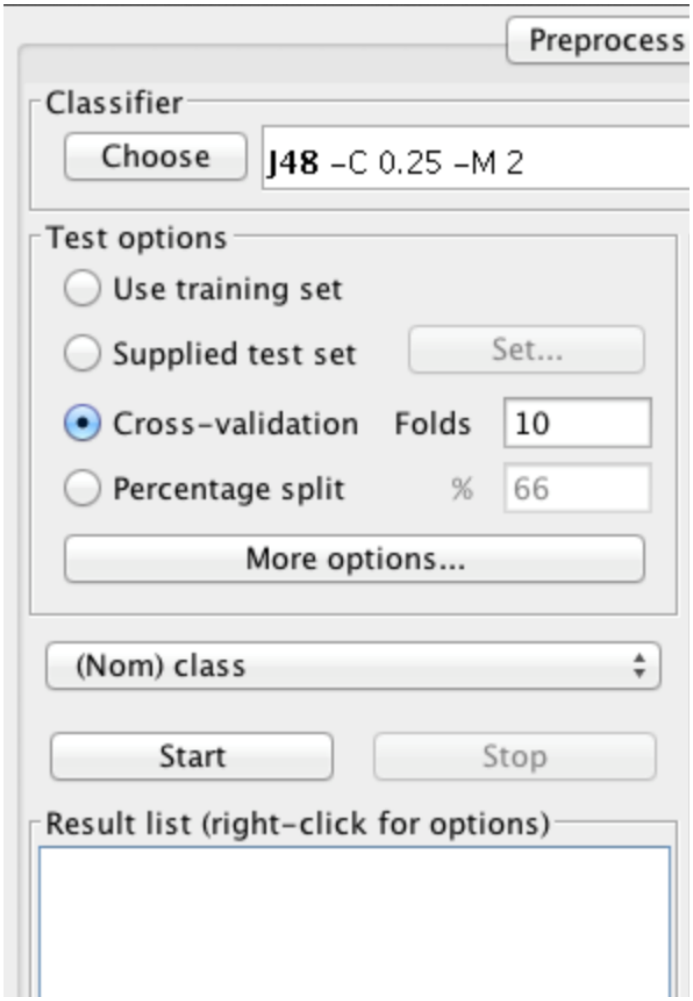
Q1.7 Which of these attributes has *myope* as a possible value? ___________________

## Part 2. Classification

Open the **iris** dataset. Navigate to the Classify tab in Weka explorer.
Click on ‘Choose’ and select “J48” as show in figure below:



Now in test options, select the default test option “Cross-validation Folds 10” and click start.



Classifier: J48 -C 0.25 -M 2

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

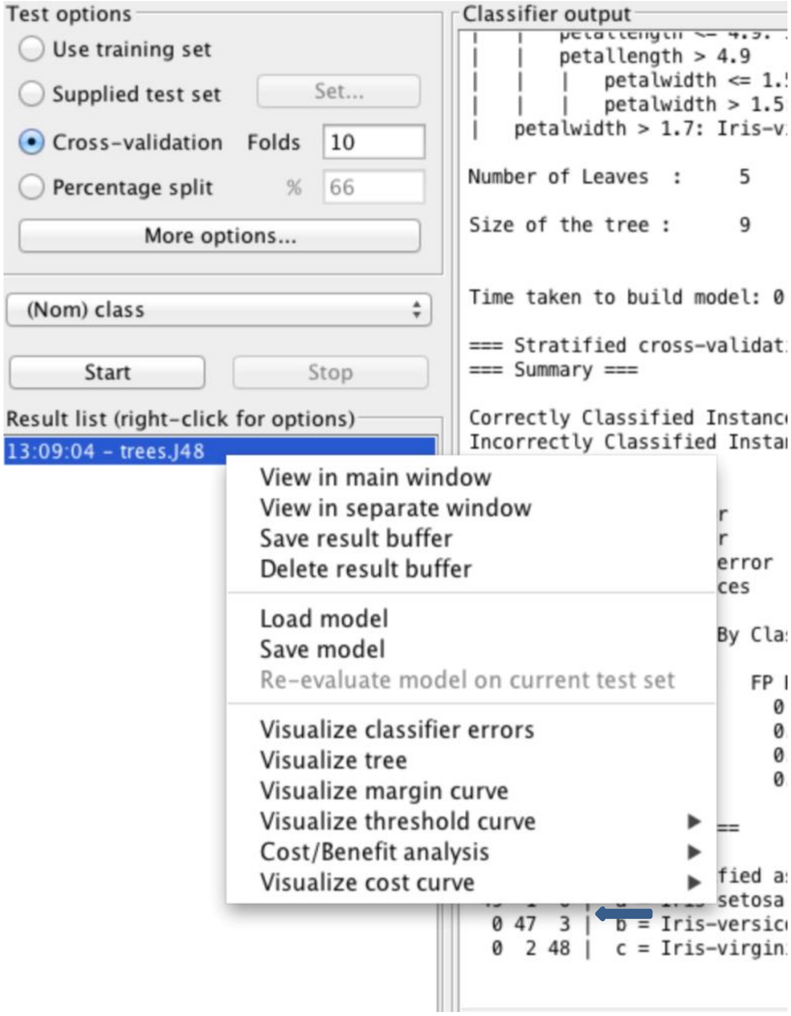
(Nom) class

## Output representation:

Analysis reports can be seen on the right, which includes the following:

- Classification Algorithm Used
- Attributes used in analysis
- Rules used to create the tree
- Classification accuracy
- Error metrics obtained in analysis
- Confusion Matrix

Now right click on the model and click on “Visualize tree”



Cross-validation Folds 10

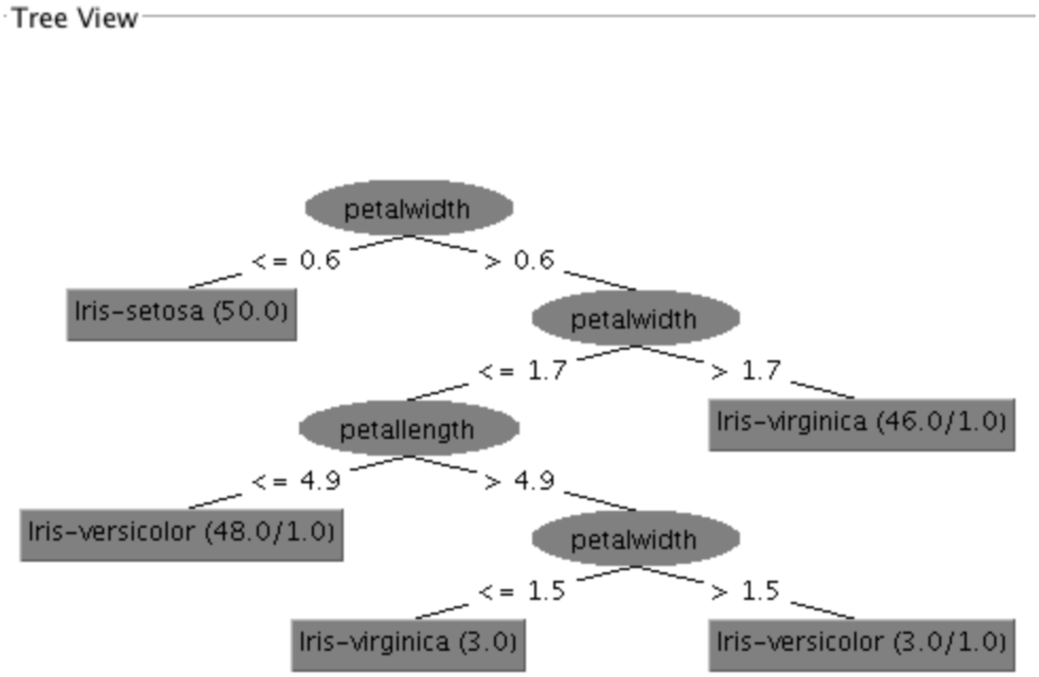
Percentage split % 66

(Nom) class

13:09:04 - trees.J48

```
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5
|   |   |   petalwidth > 1.5
|   petalwidth > 1.7: Iris-v

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0

=== Stratified cross-validat
=== Summary ===

Correctly Classified Instanc
Incorrectly Classified Insta


                                    r
                                    r
                                    error
                                    ces

                                    By Cla

                                       FP
                                        0
                                        0
                                        0
                                        0

                                    ==

                                    fied a
                                    setosa
  0 47  3 |   b = Iris-versic
  0  2 48 |   c = Iris-virgin
```

You can see a decision tree created for the algorithm:



**Questions**

Q2.1 What is the percentage of correctly classified instances? (5 pts.) _________________

Q2.2 How many instances are misclassified? (5 pts.)______________

Q2.3 “Visualize classifier errors” by right-clicking on the Result list, and use the visualization to determine the instance numbers of the misclassified instances. Remember to use the Jitter slider. Sometimes points sit right on top of each other. List instance numbers. (15 pts.)

______________________________________________________________

Q2.4 Which instance(s) of type *Iris-virginica* is misclassified as *Iris-versicolor*? (5 pts.)

______________________________________________________________

Save your results as ***Lab02_YourName***.model and submit it to MyCourses Drop Box before the due date.

Q2.5 Using the decision tree model (***Lab02_YourName***.model ) you have built, classify new unseen data given below. The ARFF file (**Lab02_unknown_YourName.arff**) with cases to predict needs to have the same structure as the file used to learn the model. The difference is that the value attribute is "?" (w/o "") for all instances.

1) Load the saved model with the right click on the "Result list" panel.
2) In the "Test Options", select "Supplied test" to load the test dataset & select "(Nom) class" from the list of attributes.
3) Click "More options" and choose PlainText from 'Output predictions'
4) Finally, right click in the model and run "Re-evaluate model on current test set".

| Sepal-length | Sepal-width | Petal-length | Petal-width | Class |
|---|---|---|---|---|
| 5.0 | 3.4 | 1.4 | 0.3 | __________________ |
| 6.7 | 3.1 | 5.3 | 1.3 | __________________ |
| 6.0 | 4.4 | 2.4 | 1.5 | __________________ |

Submit your **Lab02_unknown_YourName.arff** to Dropbox.

## Part 3. Confusion Matrix

A classification was performed on another dataset with class label as humidity. The confusion matrix obtained is shown below:

**Classification Model: M Confusion Matrix**

| | Humidity | Predicted class | | |
|---|---|---|---|---|
| | | High | Medium | Low |
| Actual class | High | 23 | 4 | 0 |
| | Medium | 6 | 13 | 3 |
| | Low | 9 | 2 | 20 |

Q3.1 How many actual instances of each class label exist? (5 points)

Q3.2 Calculate the number of correctly classified instances? Where did you find it? (10 points)

Q3.3 Calculate the classification accuracy. Show all calculations. (10points)

Q3.4 The following is a **Cost Matrix**. Calculate the cost of the **Classification Model M** shown in the above. (15 points)

| | Humidity | Predicted class | | |
|---|---|---|---|---|
| | | High | Medium | Low |
| Actual class | High | -1 | 5 | 10 |
| | Medium | 5 | 0 | 10 |
| | Low | 100 | 10 | 0 |

## Part 4. Decision Tree (20 points)

Supposed the weather data shown below is to use for building a Decision Tree. The first step is to find the root node choosing an attribute from four attributes: outlook, temperature, humidity and windy. Find the first optimal splitting node using the GINI measure. Choose one and show your work.

Answer: outlook___; temperature___; humidity____; or windy____

| outlook | temperature | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | FALSE | no |
| sunny | hot | high | TRUE | no |
| overcast | hot | high | FALSE | yes |
| rainy | mild | high | FALSE | yes |
| rainy | cool | normal | FALSE | yes |
| rainy | cool | normal | TRUE | no |
| overcast | cool | normal | TRUE | yes |
| sunny | mild | high | FALSE | no |
| sunny | cool | normal | FALSE | yes |
| rainy | mild | normal | FALSE | yes |
| sunny | mild | normal | TRUE | yes |
| overcast | mild | high | TRUE | yes |
| overcast | hot | normal | FALSE | yes |
| rainy | mild | high | TRUE | no |

**Show your work on pages 11/12.**

**Part 4. Show your work**

**Part 4. Show your work (cont.)**

| | |
|---|---|
| R•I•T | **Rochester Institute of Technology**<br>**Golisano College of Computing and Information Sciences**<br>**School of Information** |

ISTE-600 LAB 2 Gradesheet Name: ______________________________

| Questions | Max. Points | Points Earned | Comments |
|---|---|---|---|
| Part 1 | | | |
| Q1.1 | 5 | | |
| Q1.2 | 5 | | |
| Q1.3 | 5 | | |
| Q1.4 | 5 | | |
| Q1.5 | 5 | | |
| Q1.6 | 5 | | |
| Q1.7 | 5 | | |
| Part 2 | | | |
| Q2.1 | 5 | | |
| Q2.2 | 5 | | |
| Q2.3 | 5 | | |
| Q2.4 | 5 | | |
| Q2.5 | 15 | | |
| No model file | -10 | | |
| No unknown arff file | -10 | | |
| Part 3 | | | |
| Q3.1 | 5 | | |
| Q3.2 | 5 | | |
| Q3.3 | 5 | | |
| Q3.4 | 15 | | |
| Part 4 | | | |
| | 20 | | |
| TOTAL | 120 | | |