Data Mining Tutorial

Mark A. Austin

University of Maryland

austin@umd.edu
ENCE 688P, Fall Semester 2021

October 16, 2021

Overview

Part 05

Working with

Weka

Introduction

WEKA

WEKA (Waikato Environment for Knowledge Acquisition) is a workbench for data mining and machine learning.

Software Download and Installation

- WEKA is written in Java, so it will run on both PCs and Macs.
- Download from: https://www.cs.waikato.ac.nz/weka/

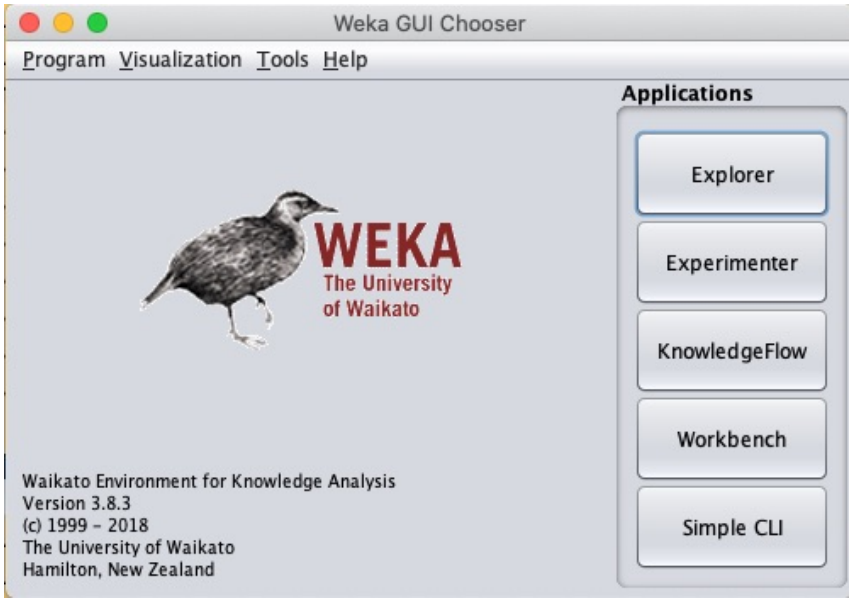
Online Resources

- See class web page for evolving list of links to WEKA resources ...
- Videos learning machine learning with WEKA are available on YouTube.

Getting Started

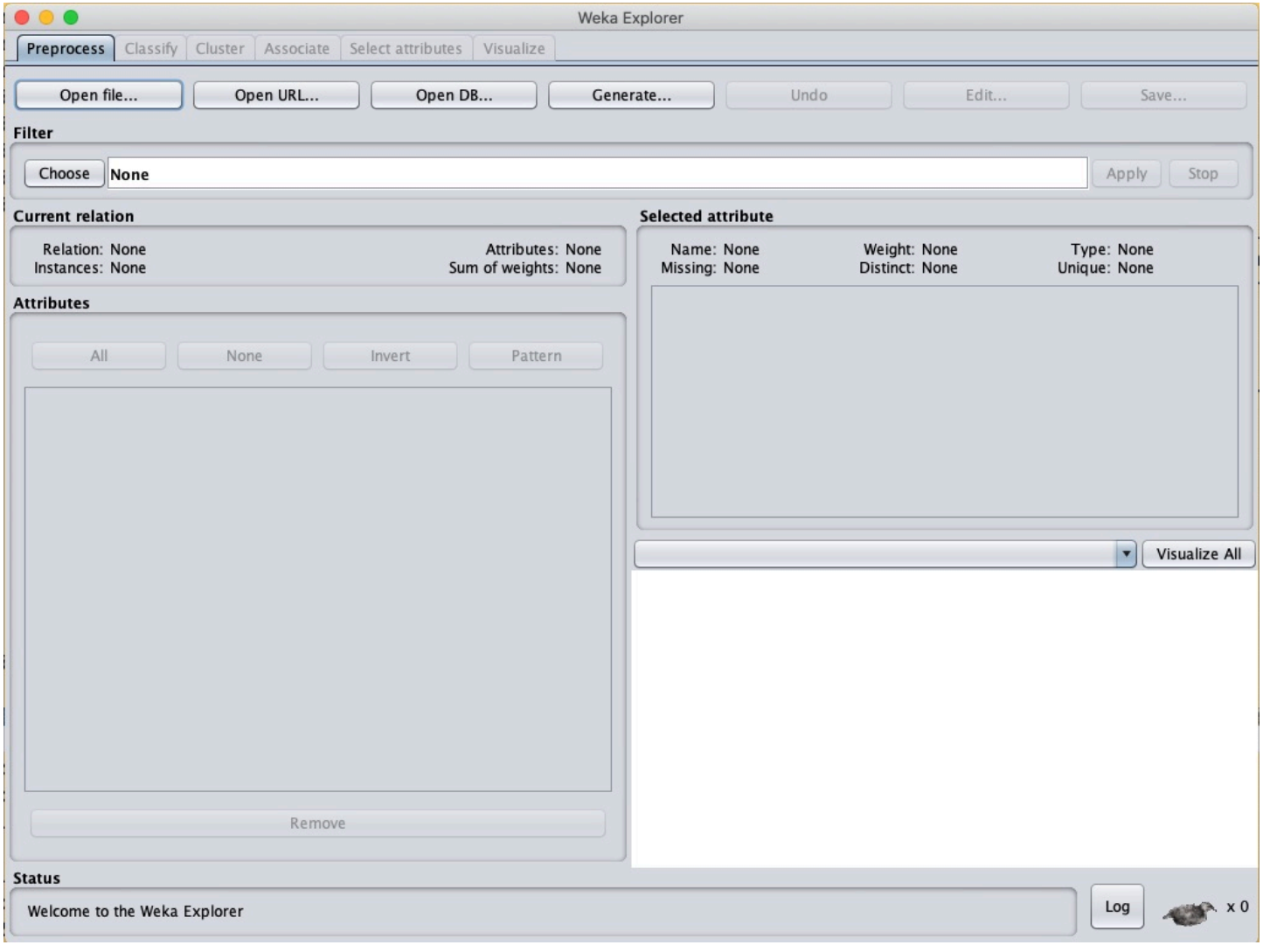
From the Terminal Window

```
prompt >> java -jar weka.jar
```

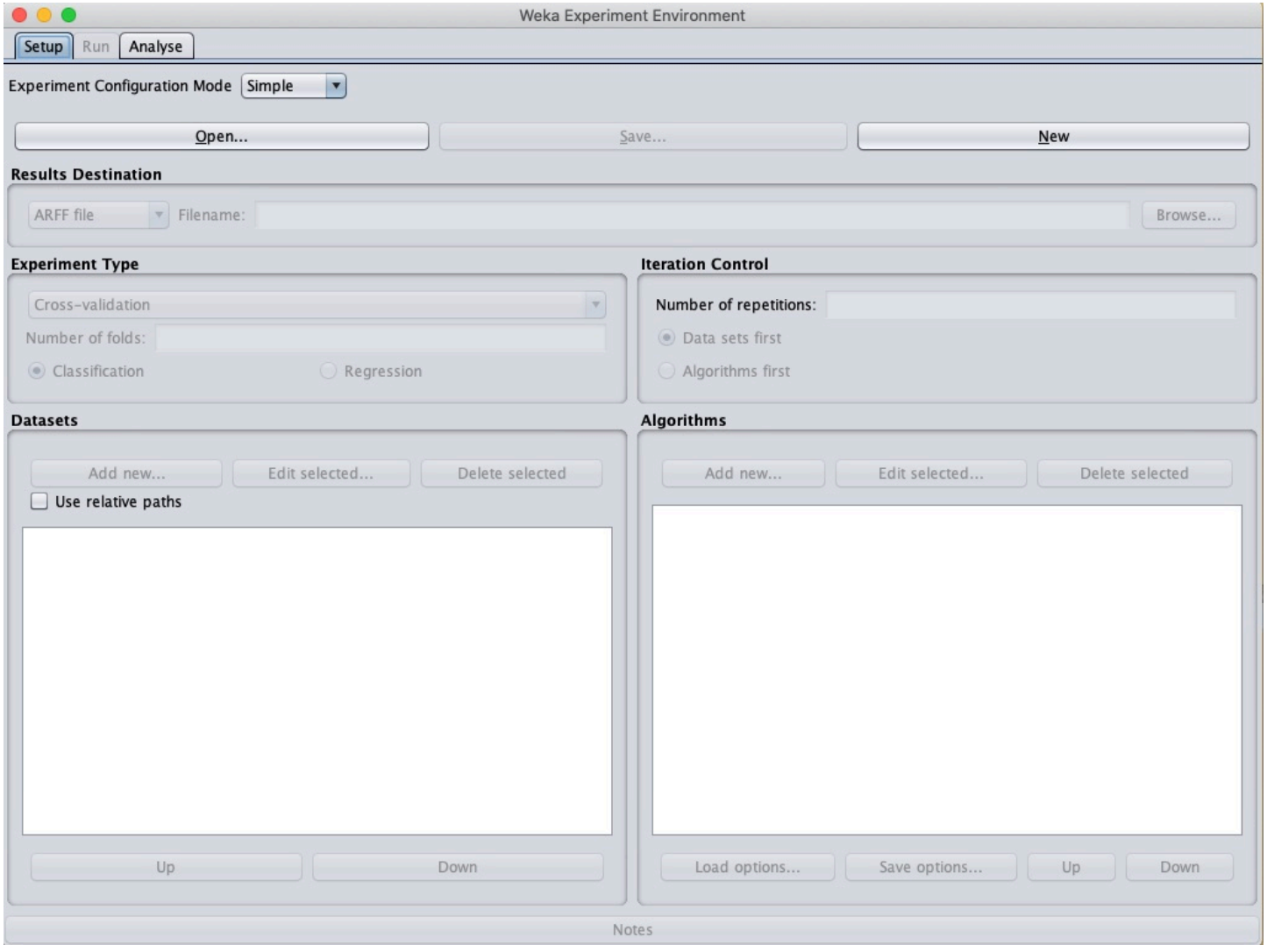


You can also write and run custom applications through the WEKA API.

Weka GUI Explorer



Weka GUI Experimenter



Data Mining

Examples

Example 1. Will Customer Buy Computer?

Input datafile (arff format)

```
% ============================================================
% ENCE 688P: Classification for Buy Computer?
% ============================================================

@relation 'computer'
@attribute id real
@attribute age { young, middle, senior}
@attribute income { low, medium, high}
@attribute student {yes, no}
@attribute credit { fair, excellent}
@attribute purchase { no, yes}

@data
1,young,high,no,fair,no
2,young,high,no,excellent,no
3,middle,high,no,fair,yes
4,senior,medium,no,fair,yes
5,senior,low,yes,fair,yes
6,senior,low,yes,excellent,no
7,middle,low,yes,excellent,yes
8,young,medium,no,fair,no
9,young,low,yes,fair,yes
10,senior,medium,yes,fair,yes
11,young,medium,yes,excellent,yes
12,middle,medium,no,excellent,yes
13,middle,high,yes,fair,yes
14,senior,medium,no,excellent,no
```

Example 1. Will Customer Buy Computer?

Java Program Source Code

See: java-code-ml-weka2018/src/ence688p/ClassificationTask.java

Abbreviated Program Output (J48 unpruned tree)

```
age = young
|   student = yes: yes (2.0)
|   student = no: no (3.0)
age = middle: yes (4.0)
age = senior
|   credit = fair: yes (3.0)
|   credit = excellent: no (2.0)

Number of Leaves  :  5
Size of the tree :  8
```

Example 1. Will Customer Buy Computer?

Classification Accuracy wrt Training Dataset

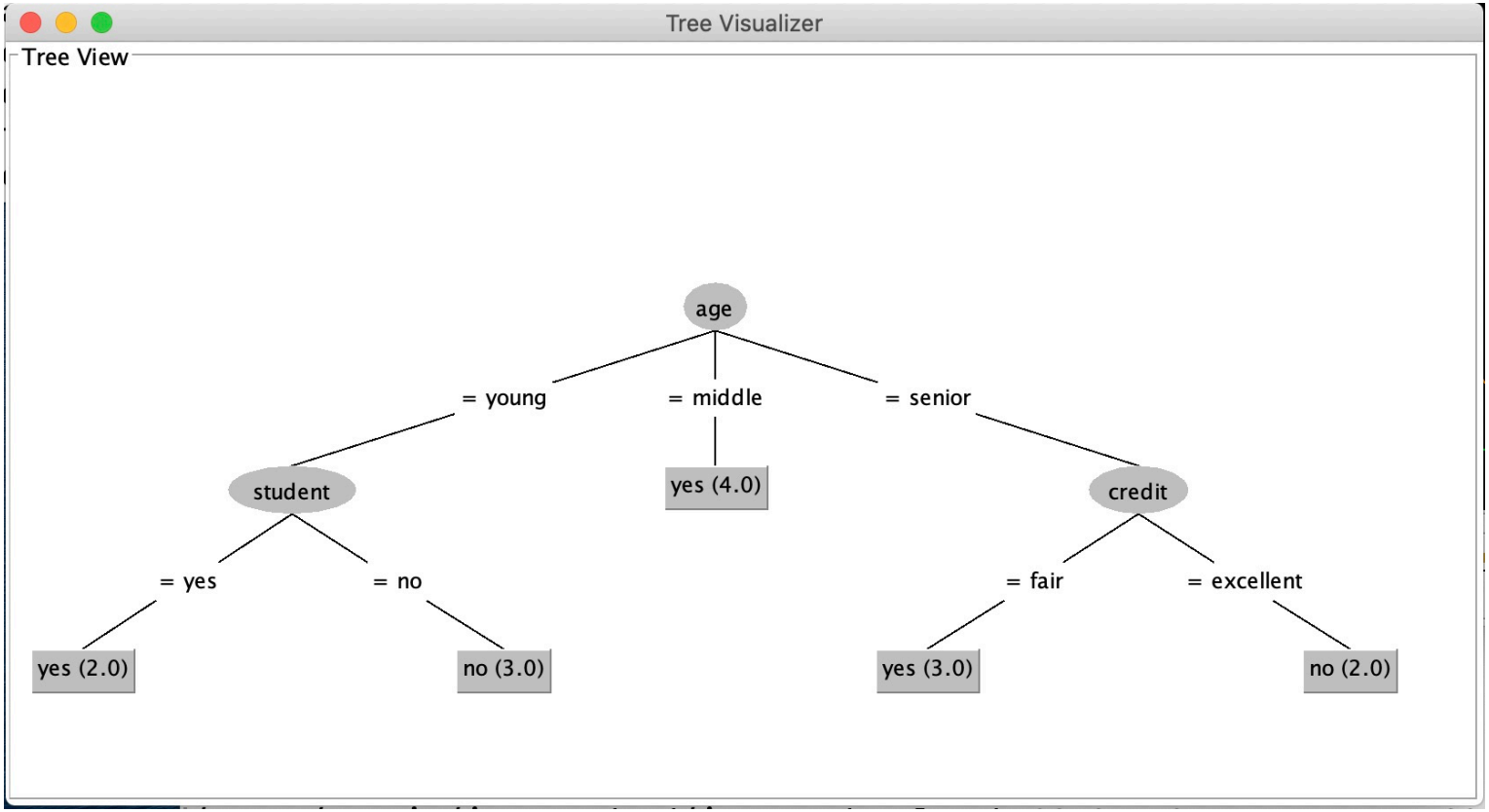
```
Correctly Classified Instances          14     100 %
Incorrectly Classified Instances         0       0 %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0 %
Root relative squared error              0 %
Total Number of Instances               14

=== Confusion Matrix ===

    a b   <-- classified as
    5 0 | a = no
    0 9 | b = yes
```

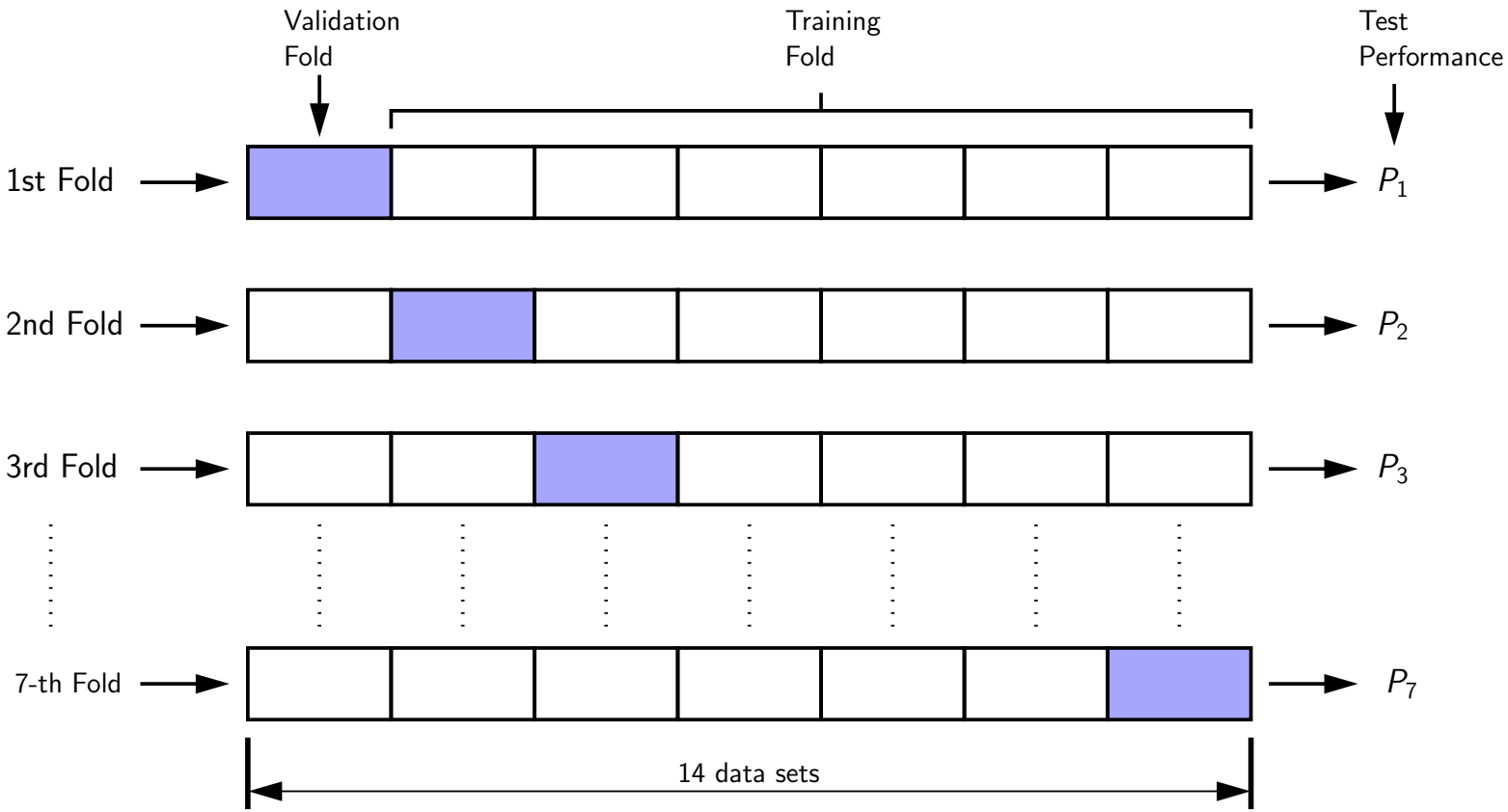
Example 1. Will Customer Buy Computer?

Classification Accuracy wrt Training Dataset



Example 1. Will Customer Buy Computer?

Cross Validation Model (nofolds = 7)



Example 1. Will Customer Buy Computer?

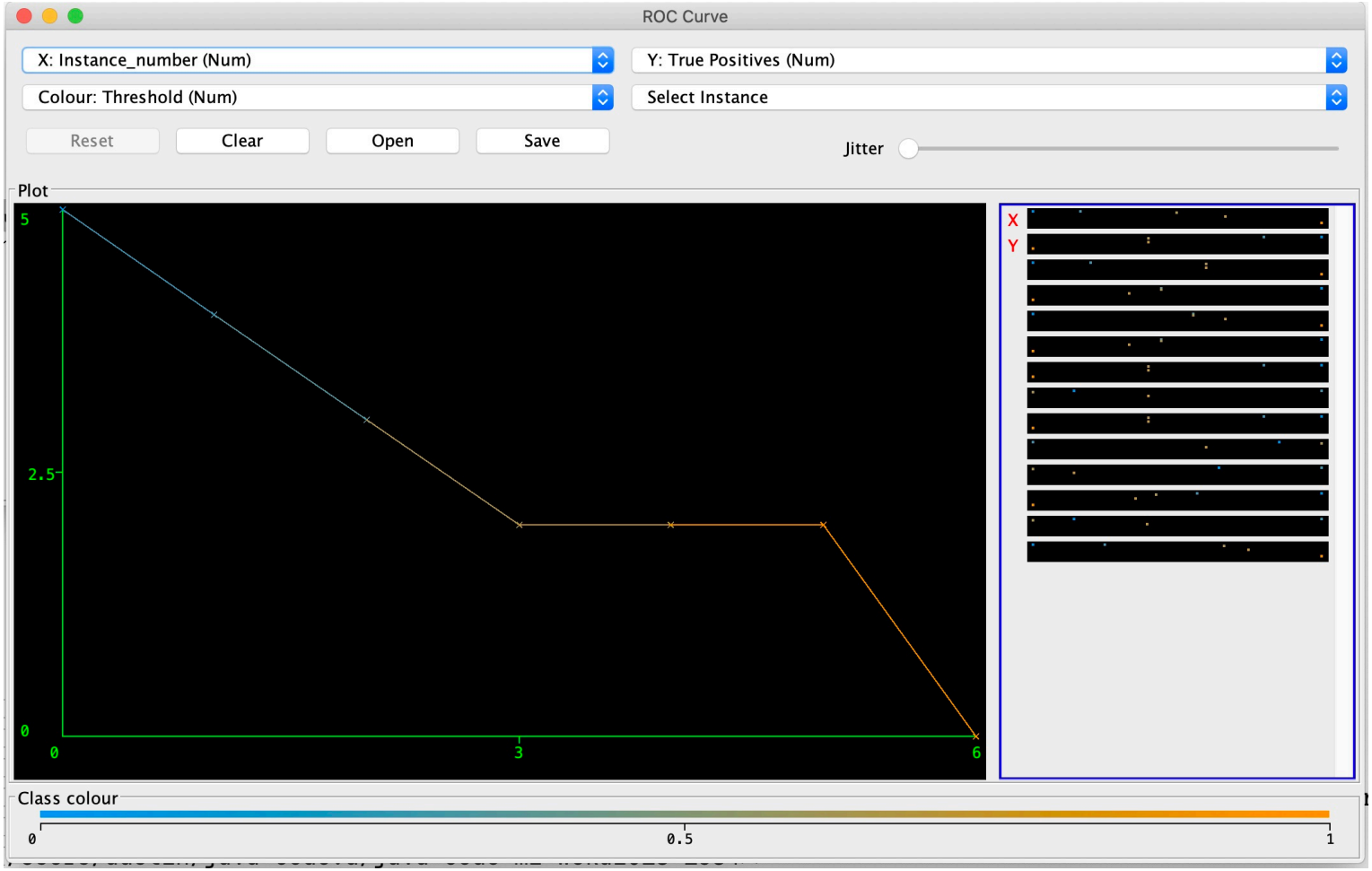
Cross Validation Model (after classification) (nofolds = 7)

```
Correctly Classified Instances          10               71.4286 %
Incorrectly Classified Instances         4               28.5714 %
Kappa statistic                          0.3778
Mean absolute error                      0.2798
Root mean squared error                  0.4393
Relative absolute error                 58.3333 %
Root relative squared error             88.6322 %
Total Number of Instances               14

=== Confusion Matrix ===

   a b   <-- classified as
   3 2 | a = no
   2 7 | b = yes
```

Example 1. Will Customer Buy Computer?



Example 2. Milk, Diapers and Beer

Input datafile (arff format)

```
1  % ================================================================
2  % ENCE 688P: Customer purchases at supermarket ..
3  %
4  % Mark Austin                                              March, 2021
5  % ================================================================
6
7  @relation 'supermarket'
8  @attribute id real
9  @attribute beer {t}
10 @attribute bread {t}
11 @attribute coke {t}
12 @attribute diapers {t}
13 @attribute eggs {t}
14 @attribute milk {t}
15
16 @data
17 1,?,t,?,?,?,t
18 2,t,t,?,t,t,?
19 3,t,?,t,t,?,t
20 4,t,t,?,t,?,t
21 5,?,t,t,t,?,t
```

Example 2. Milk, Diapers and Beer

Java Program Source Code (Weka Code)

See: java-code-ml-weka2018/src/ence688p/Supermarket.java

Abbreviated Program Output (Print modified input file)

```
@relation supermarket-weka.filters.unsupervised.attribute.Remove-R1

@attribute beer {t}
... attributes removed ...
@attribute milk {t}

@data
?,t,?,?,?,t
t,t,?,t,t,?
t,?,t,t,?,t
t,t,?,t,?,t
?,t,t,t,?,t
```

Example 2. Milk, Diapers and Beer

Abbreviated Program Output (Apriori Model)

```
Size of set of large itemsets L(1): 6
Size of set of large itemsets L(2): 13
Size of set of large itemsets L(3): 12
Size of set of large itemsets L(4): 4

Best rules found:

 1. beer=t 3 ==> diapers=t 3 <conf:(1)> lift:(1.25) lev:(0.12) [0] conv
 2. coke=t 2 ==> diapers=t 2 <conf:(1)> lift:(1.25) lev:(0.08) [0] conv
 3. coke=t 2 ==> milk=t 2 <conf:(1)> lift:(1.25) lev:(0.08) [0] conv:(0
 4. beer=t bread=t 2 ==> diapers=t 2 <conf:(1)> lift:(1.25) lev:(0.08)
 5. beer=t milk=t 2 ==> diapers=t 2 <conf:(1)> lift:(1.25) lev:(0.08) [
 6. coke=t milk=t 2 ==> diapers=t 2 <conf:(1)> lift:(1.25) lev:(0.08) [
 7. coke=t diapers=t 2 ==> milk=t 2 <conf:(1)> lift:(1.25) lev:(0.08) [
 8. coke=t 2 ==> diapers=t milk=t 2 <conf:(1)> lift:(1.67) lev:(0.16) [
 9. eggs=t 1 ==> beer=t 1 <conf:(1)> lift:(1.67) lev:(0.08) [0] conv:(0
10. eggs=t 1 ==> bread=t 1 <conf:(1)> lift:(1.25) lev:(0.04) [0] conv:(
```

Example 2. Milk, Diapers and Beer

Abbreviated Program Output (FPGrowth Model)

```
FPGrowth found 38 rules (displaying top 10)

 1. [coke=t]: 2 ==> [milk=t]: 2 <conf:(1)> lift:(1.25) lev:(0.08) conv:
 2. [beer=t]: 3 ==> [diapers=t]: 3 <conf:(1)> lift:(1.25) lev:(0.12) co
 3. [coke=t]: 2 ==> [diapers=t]: 2 <conf:(1)> lift:(1.25) lev:(0.08) co
 4. [eggs=t]: 1 ==> [diapers=t]: 1 <conf:(1)> lift:(1.25) lev:(0.04) co
 5. [eggs=t]: 1 ==> [bread=t]: 1 <conf:(1)> lift:(1.25) lev:(0.04) conv
 6. [eggs=t]: 1 ==> [beer=t]: 1 <conf:(1)> lift:(1.67) lev:(0.08) conv:
 7. [milk=t, beer=t]: 2 ==> [diapers=t]: 2 <conf:(1)> lift:(1.25) lev:(
 8. [coke=t]: 2 ==> [milk=t, diapers=t]: 2 <conf:(1)> lift:(1.67) lev:(
 9. [milk=t, coke=t]: 2 ==> [diapers=t]: 2 <conf:(1)> lift:(1.25) lev:(
10. [diapers=t, coke=t]: 2 ==> [milk=t]: 2 <conf:(1)> lift:(1.25) lev:(

--- ===================================== ...
--- Finished !! ...
```

References

- Jaynes E.T., Information Theory and Statistical Mechanics. II, Phys. Rev. 108, 171, October 1957.
- Kapur J.N., Maximum-Entropy Models in Science and Engineering, John Wiley and Sons, 1989.
- Mitchell T.M., Machine Learning and Data Mining, Communications of the ACM, Vol. 42., No. 11, November 1999.
- Russell S., and Norvig P., Artificial Intelligence: A Modern Approach (Third Edition), Prentice-Hall, 2010.
- Shanon C.E., and Weaver W., The Mathematical Theory of Communication, University of Illinois, Urbana, Chicago, 1949.
- Witten I.H., Frank E., Hall M.A., and Pal C.J., Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2017.