Data mining Term project (option 1)

Comparison of Decision tress and Support Vector Machine

Johanna Del Pino (jmd53), Kijit Desai (kd227)

Data Description 3

Decision Tree (C4.5) 4

- Data Preprocessing 4
- Running the algorithm 10

Decision Tree 12

Classifier Output 13

Results 15

Support Vector Machines 16

Running the Algorithm 18

Classifier Output 21

Results 27

Cross Validation Results and Comparisons 28

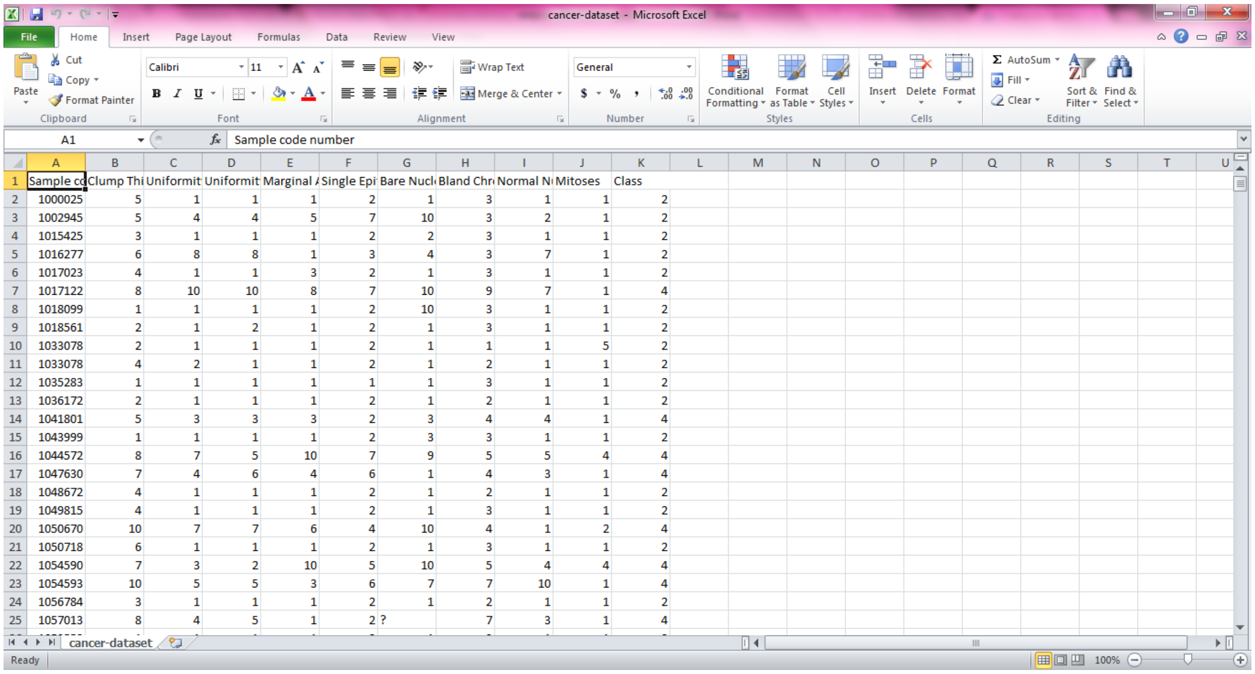
Data Description

Title: Breast Cancer Wisconsin (Prognostic) Data Set

Url: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29

This training dataset contains data from 699 breast cancer cases. Each case could be diagnosed as benign or malignant. For the diagnosis, each case has ten attributes. These attributes are listed in the table below. The last attribute in the table represents the class, or label with which the case was diagnosed. The dataset was storage as a CSV file in order to use it in WEKA (Figure 1).

Attribute		Domain
1	Sample code number	id number
2	Clump Thickness	1 - 10
3	Uniformity of Cell Size	1 - 10
4	Uniformity of Cell Shape	1 - 10
5	Marginal Adhesion	1 - 10
6	Single Epithelial Cell Size	1 - 10
7	Bare Nuclei	1 - 10
8	Bland Chromatin	1 - 10
9	Normal Nucleoli	1 - 10
10	Mitoses	1 - 10
11	Class	2 for benign, 4 for malignant



A1 — Sample code number

	A	B	C	D	E	F	G	H	I	J	K
1	Sample cd	Clump Thi	Uniformit	Uniformit	Marginal /	Single Epi	Bare Nucl	Bland Chr	Normal N	Mitoses	Class
2	1000025	5	1	1	1	2	1	3	1	1	2
3	1002945	5	4	4	5	7	10	3	2	1	2
4	1015425	3	1	1	1	2	2	3	1	1	2
5	1016277	6	8	8	1	3	4	3	7	1	2
6	1017023	4	1	1	3	2	1	3	1	1	2
7	1017122	8	10	10	8	7	10	9	7	1	4
8	1018099	1	1	1	1	2	10	3	1	1	2
9	1018561	2	1	2	1	2	1	3	1	1	2
10	1033078	2	1	1	1	2	1	1	1	5	2
11	1033078	4	2	1	1	2	1	2	1	1	2
12	1035283	1	1	1	1	1	1	3	1	1	2
13	1036172	2	1	1	1	2	1	2	1	1	2
14	1041801	5	3	3	3	2	3	4	4	1	4
15	1043999	1	1	1	1	2	3	3	1	1	2
16	1044572	8	7	5	10	7	9	5	5	4	4
17	1047630	7	4	6	4	6	1	4	3	1	4
18	1048672	4	1	1	1	2	1	2	1	1	2
19	1049815	4	1	1	1	2	1	3	1	1	2
20	1050670	10	7	7	6	4	10	4	1	2	4
21	1050718	6	1	1	1	2	1	3	1	1	2
22	1054590	7	3	2	10	5	10	5	4	4	4
23	1054593	10	5	5	3	6	7	7	10	1	4
24	1056784	3	1	1	1	2	1	2	1	1	2
25	1057013	8	4	5	1	2	?	7	3	1	4

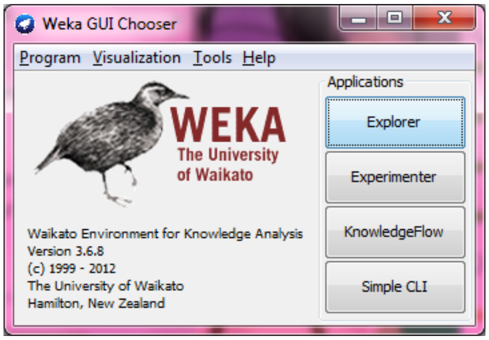
Figure 1 - CSV file of Training dataset

Decision Tree (C4.5)

The ID3 algorithm classifies data with discrete values; in order for WEKA to recognize the dataset values as discrete a preprocessing phase must be done in the data before running the algorithm. In this phase the Discretization of the dataset is done, along with the filtering of the "Sample code number" attribute that is not needed for the classification.

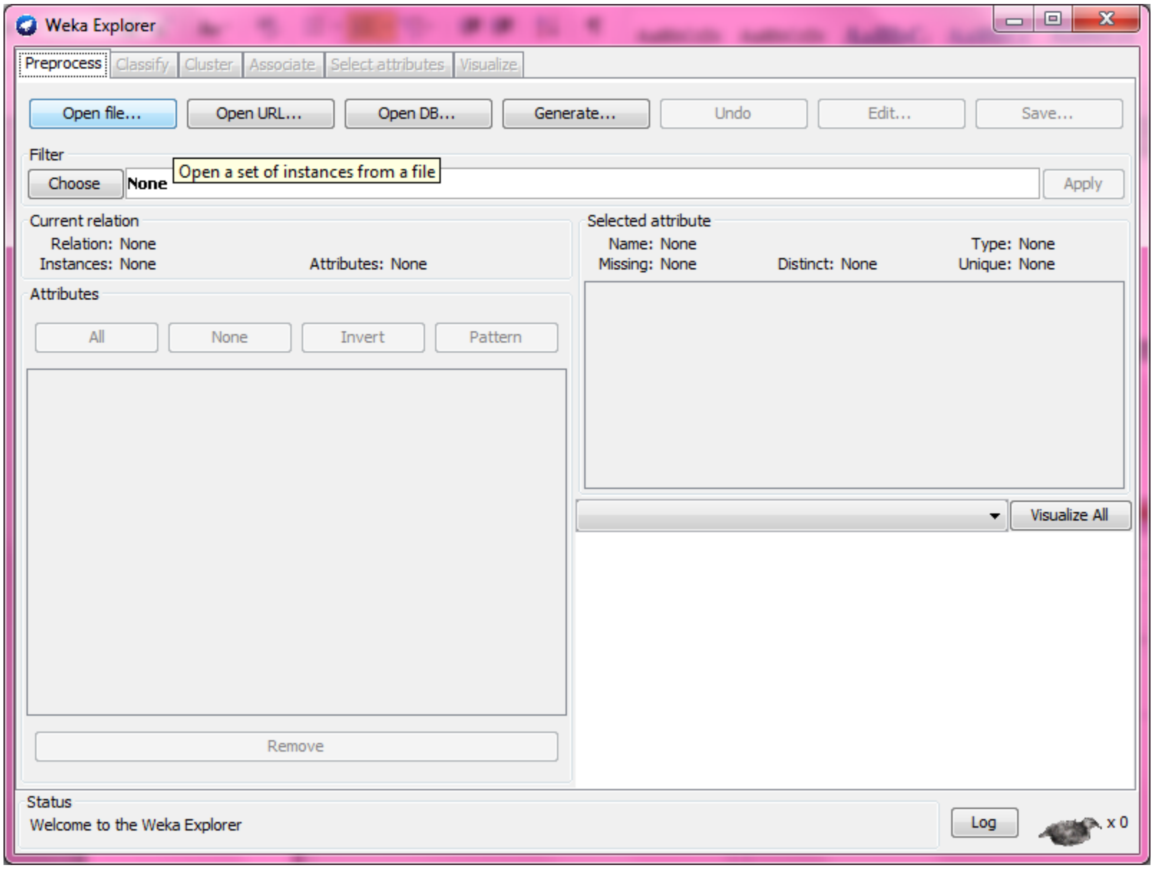
Data Preprocessing

1. **Opening WEKA:** We open the WEKA tool and click in the "Explorer" button



Waikato Environment for Knowledge Analysis
Version 3.6.8
(c) 1999 - 2012
The University of Waikato
Hamilton, New Zealand

2. **Loading the dataset:** In the new window we choose the "Open file" option to upload the CSV file. Then we choose the CSV file and WEKA will display the data and its attributes.



Filter: None

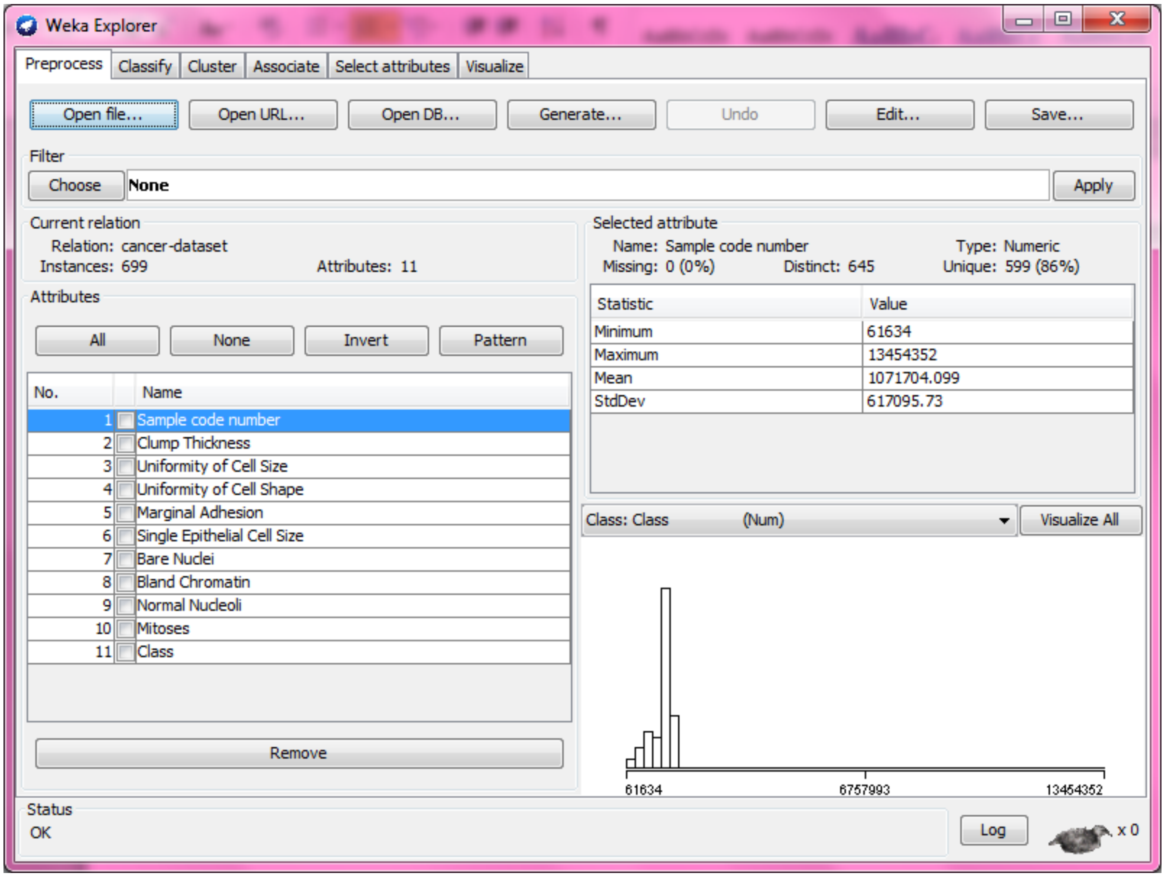
Open a set of instances from a file

Current relation
Relation: None
Instances: None
Attributes: None

Attributes

Selected attribute
Name: None
Missing: None
Distinct: None
Type: None
Unique: None

Status
Welcome to the Weka Explorer



Filter: None

Current relation
Relation: cancer-dataset
Instances: 699 Attributes: 11

Selected attribute
Name: Sample code number Type: Numeric
Missing: 0 (0%) Distinct: 645 Unique: 599 (86%)

Statistic	Value
Minimum	61634
Maximum	13454352
Mean	1071704.099
StdDev	617095.73

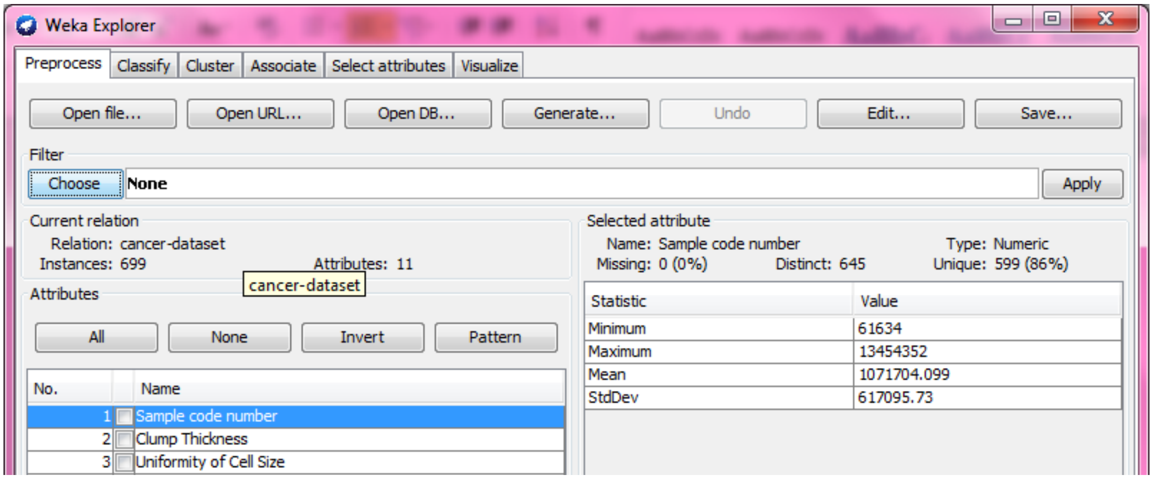
Attributes

No.	Name
1	□ Sample code number
2	□ Clump Thickness
3	□ Uniformity of Cell Size
4	□ Uniformity of Cell Shape
5	□ Marginal Adhesion
6	□ Single Epithelial Cell Size
7	□ Bare Nuclei
8	□ Bland Chromatin
9	□ Normal Nucleoli
10	□ Mitoses
11	□ Class

Class: Class (Num)

Status
OK

3. **Removing the first attribute:** We proceed to create a filter to remove the "Sample code number" attribute by clicking in the *Choose* button located under the word *Filter*, and then we choose the option filters -> unsupervised -> attribute -> remove.



Filter: None

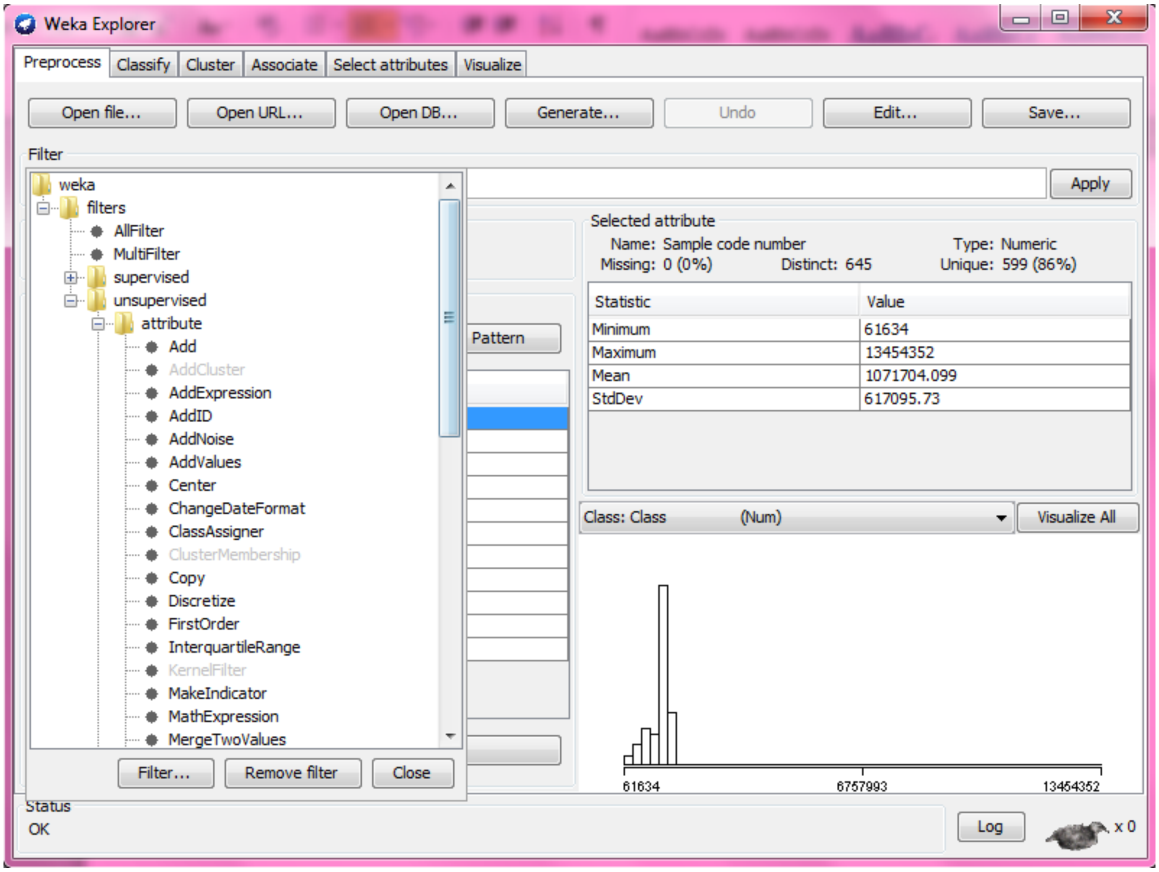
Current relation
Relation: cancer-dataset
Instances: 699 Attributes: 11

Selected attribute
Name: Sample code number Type: Numeric
Missing: 0 (0%) Distinct: 645 Unique: 599 (86%)

Statistic	Value
Minimum	61634
Maximum	13454352
Mean	1071704.099
StdDev	617095.73

Attributes

No.	Name
1	□ Sample code number
2	□ Clump Thickness
3	□ Uniformity of Cell Size



Selected attribute

Name: Sample code number Type: Numeric

Missing: 0 (0%) Distinct: 645 Unique: 599 (86%)

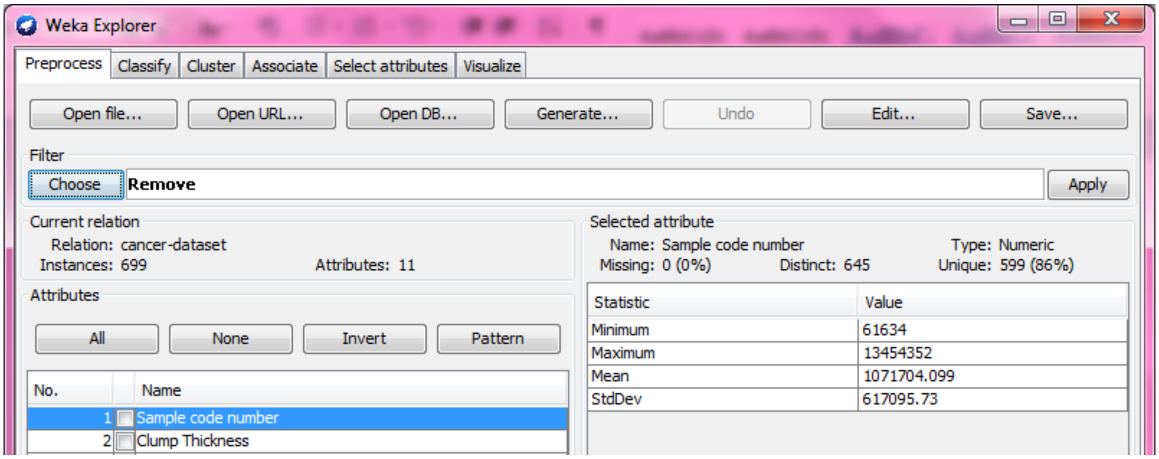
Statistic	Value
Minimum	61634
Maximum	13454352
Mean	1071704.099
StdDev	617095.73

Class: Class (Num)

Status

OK

We double click in the text field with the word “Remove” in it. A small window will appeared in which we must specify the number of the attribute we want to remove, and then press the *OK* button. Note that the *invertSelection* option must be set as false.



Remove

Current relation

Relation: cancer-dataset

Instances: 699 Attributes: 11

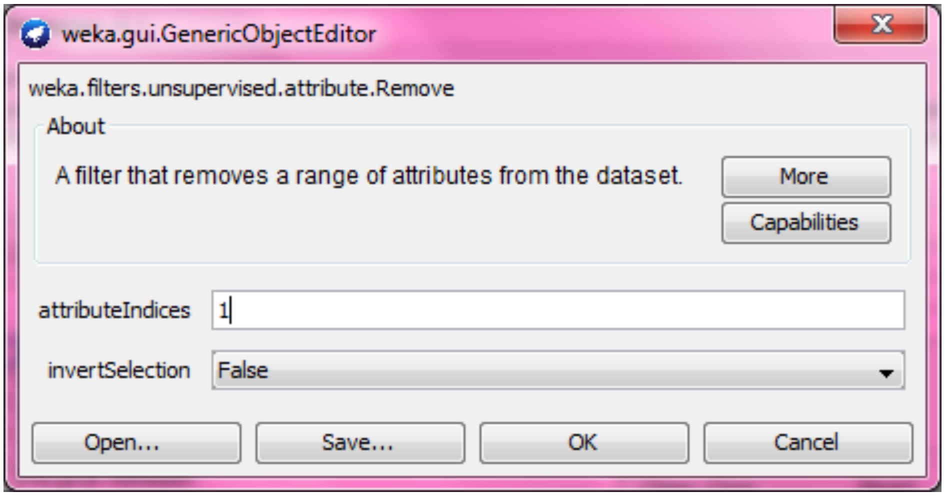
No.	Name
1	□ Sample code number
2	□ Clump Thickness

Selected attribute

Name: Sample code number Type: Numeric

Missing: 0 (0%) Distinct: 645 Unique: 599 (86%)

Statistic	Value
Minimum	61634
Maximum	13454352
Mean	1071704.099
StdDev	617095.73



weka.filters.unsupervised.attribute.Remove

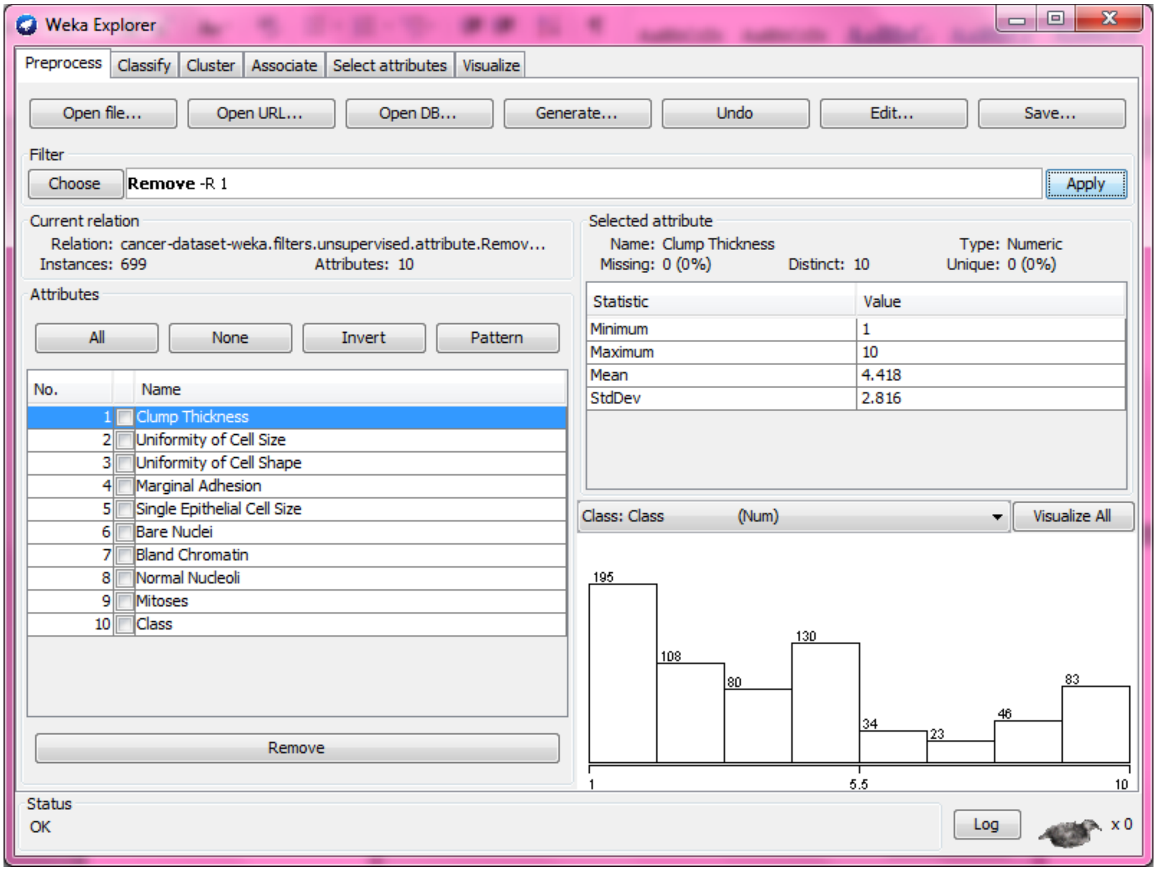
About

A filter that removes a range of attributes from the dataset.

attributeIndices: 1

invertSelection: False

Then we click on the *Apply* button next to the text field and the first attribute will be removed. After this we save this dataset as an "arff" file by clicking the *Save* button on the upper left corner of the window.



Filter: **Remove** -R 1

Current relation
Relation: cancer-dataset-weka.filters.unsupervised.attribute.Remov...
Instances: 699   Attributes: 10

Selected attribute
Name: Clump Thickness   Type: Numeric
Missing: 0 (0%)   Distinct: 10   Unique: 0 (0%)

Attributes

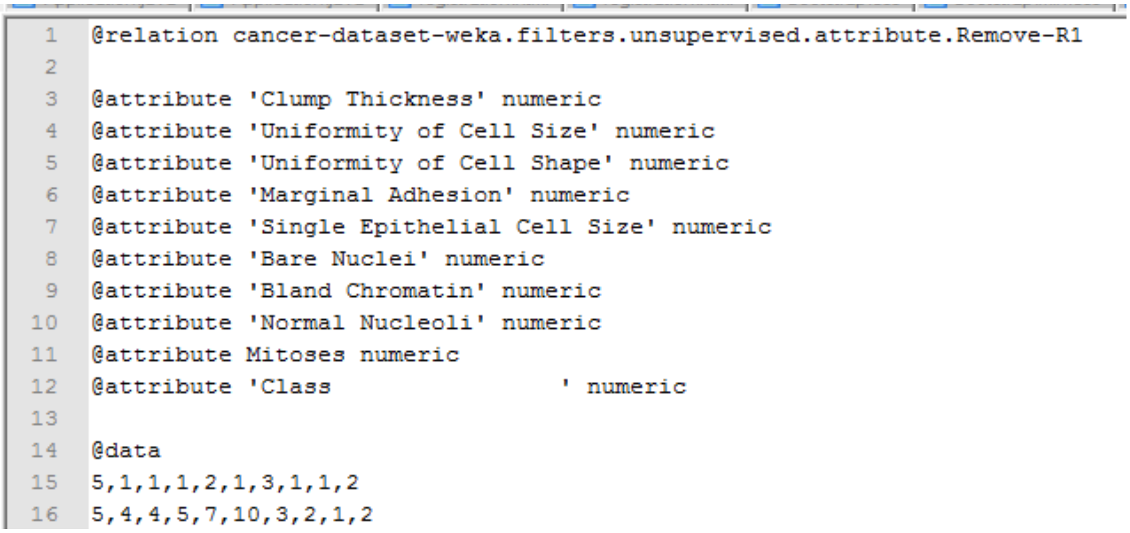
No.	Name
1	□ Clump Thickness
2	□ Uniformity of Cell Size
3	□ Uniformity of Cell Shape
4	□ Marginal Adhesion
5	□ Single Epithelial Cell Size
6	□ Bare Nuclei
7	□ Bland Chromatin
8	□ Normal Nucleoli
9	□ Mitoses
10	□ Class

Statistic	Value
Minimum	1
Maximum	10
Mean	4.418
StdDev	2.816

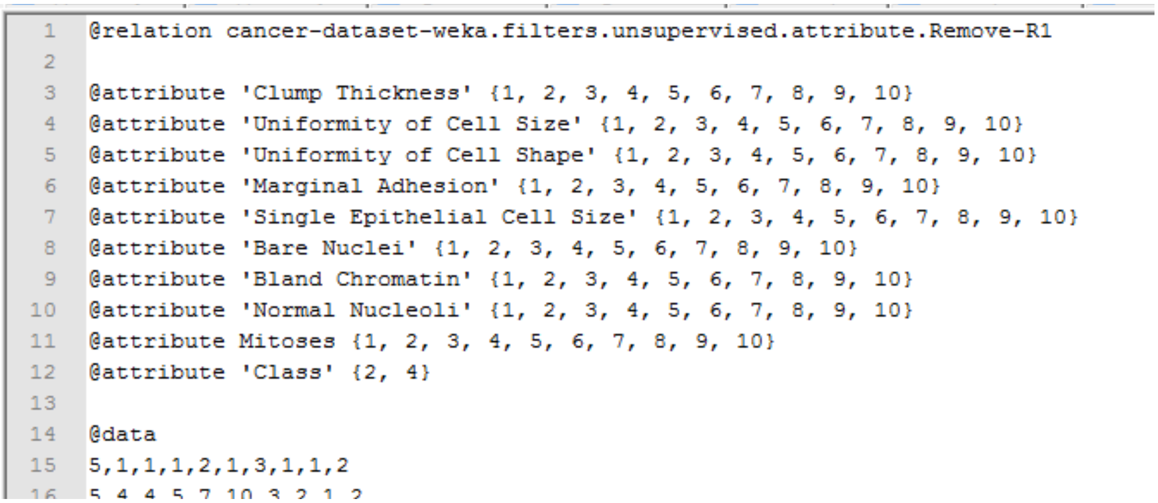
Class: Class (Num)

Status
OK

4. **Discretizing the dataset:** For this we need to go to the *arrf* file and open it as text. The *numeric* type must be replaced by the group of discrete numbers that the attribute can have. In our dataset case, all the attributes can have a value between 1-10, so the *numeric* word will be replaced by {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. The Class attribute can only have 2 and 4 as values so its discrete numbers will be {2, 4}

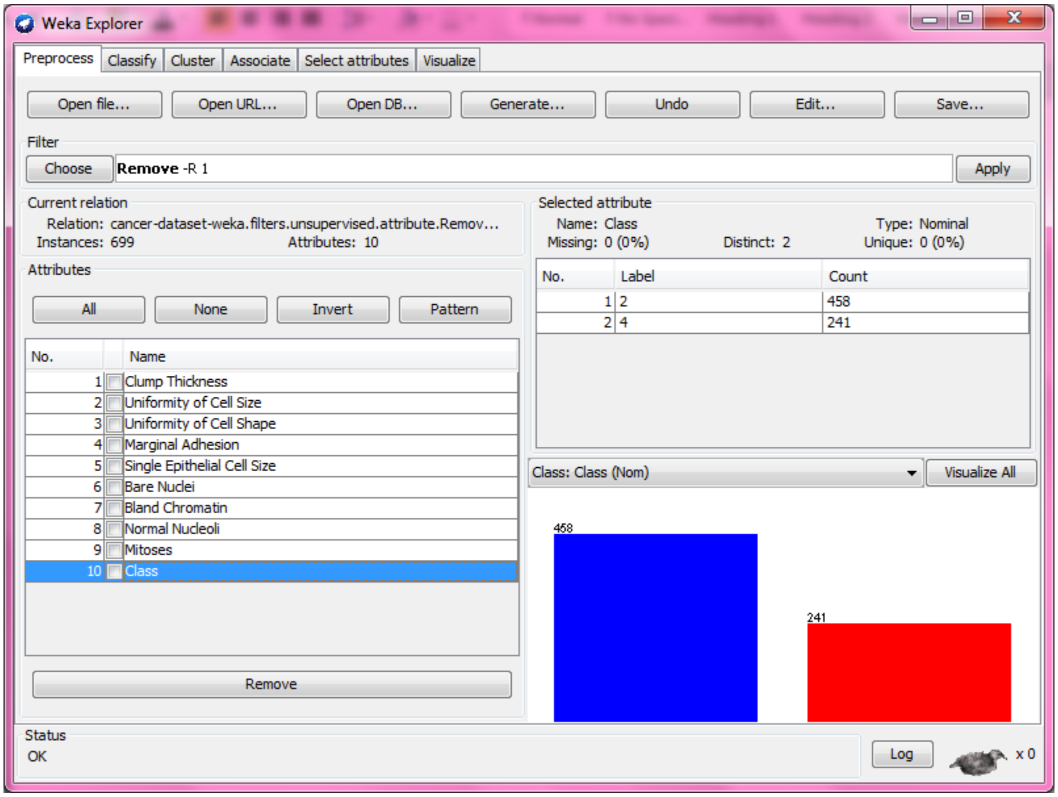


```
@relation cancer-dataset-weka.filters.unsupervised.attribute.Remove-R1

@attribute 'Clump Thickness' numeric
@attribute 'Uniformity of Cell Size' numeric
@attribute 'Uniformity of Cell Shape' numeric
@attribute 'Marginal Adhesion' numeric
@attribute 'Single Epithelial Cell Size' numeric
@attribute 'Bare Nuclei' numeric
@attribute 'Bland Chromatin' numeric
@attribute 'Normal Nucleoli' numeric
@attribute Mitoses numeric
@attribute 'Class                ' numeric

@data
5,1,1,1,2,1,3,1,1,2
5,4,4,5,7,10,3,2,1,2
```



```
@relation cancer-dataset-weka.filters.unsupervised.attribute.Remove-R1

@attribute 'Clump Thickness' {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
@attribute 'Uniformity of Cell Size' {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
@attribute 'Uniformity of Cell Shape' {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
@attribute 'Marginal Adhesion' {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
@attribute 'Single Epithelial Cell Size' {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
@attribute 'Bare Nuclei' {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
@attribute 'Bland Chromatin' {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
@attribute 'Normal Nucleoli' {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
@attribute Mitoses {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
@attribute 'Class' {2, 4}

@data
5,1,1,1,2,1,3,1,1,2
5,4,4,5,7,10,3,2,1,2
```

After this step we reload our *arff* file, and now the data will be displayed with different colors depending on the class which the case belongs. *Blue* for benign cases and *Red* for malign cases.



Filter

Remove -R 1

Current relation

Relation: cancer-dataset-weka.filters.unsupervised.attribute.Remov...

Instances: 699 Attributes: 10

Selected attribute

Name: Class Type: Nominal

Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Attributes

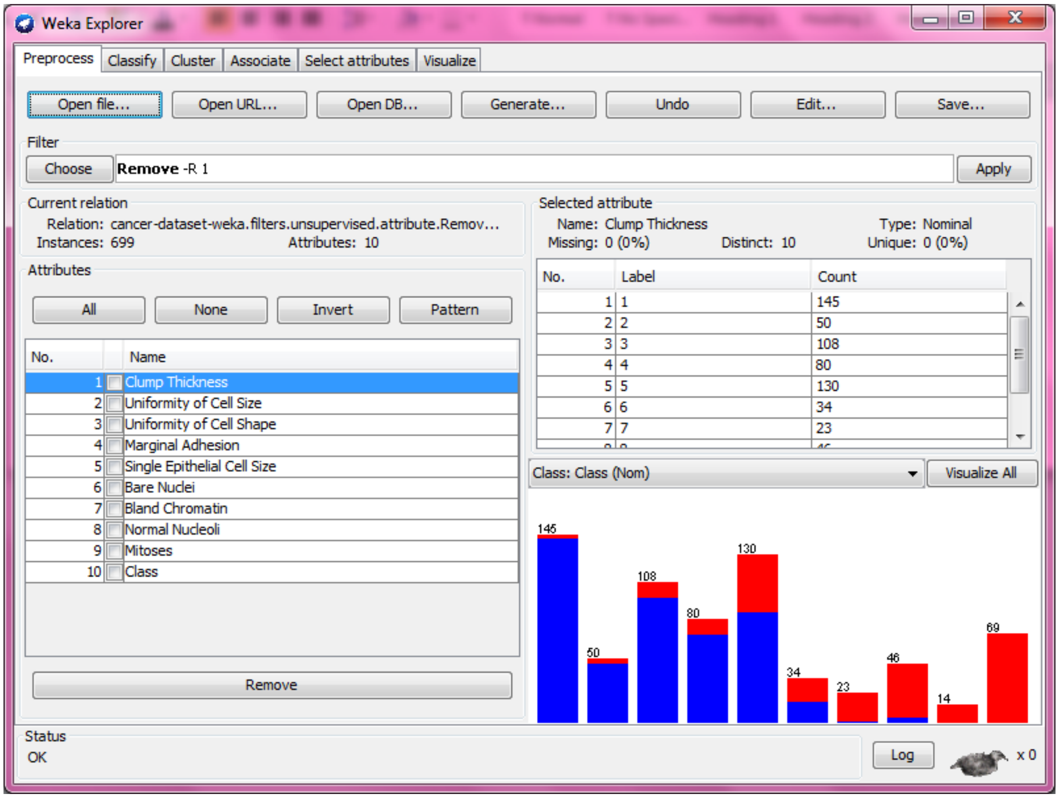
No.		Name
1	□	Clump Thickness
2	□	Uniformity of Cell Size
3	□	Uniformity of Cell Shape
4	□	Marginal Adhesion
5	□	Single Epithelial Cell Size
6	□	Bare Nuclei
7	□	Bland Chromatin
8	□	Normal Nucleoli
9	□	Mitoses
10	□	Class

No.	Label	Count
1	2	458
2	4	241

Class: Class (Nom)

Status

OK



Filter

Remove -R 1

Current relation

Relation: cancer-dataset-weka.filters.unsupervised.attribute.Remov...

Instances: 699 Attributes: 10

Selected attribute

Name: Clump Thickness Type: Nominal

Missing: 0 (0%) Distinct: 10 Unique: 0 (0%)

Attributes

No.		Name
1	□	Clump Thickness
2	□	Uniformity of Cell Size
3	□	Uniformity of Cell Shape
4	□	Marginal Adhesion
5	□	Single Epithelial Cell Size
6	□	Bare Nuclei
7	□	Bland Chromatin
8	□	Normal Nucleoli
9	□	Mitoses
10	□	Class

No.	Label	Count
1	1	145
2	2	50
3	3	108
4	4	80
5	5	130
6	6	34
7	7	23

Class: Class (Nom)

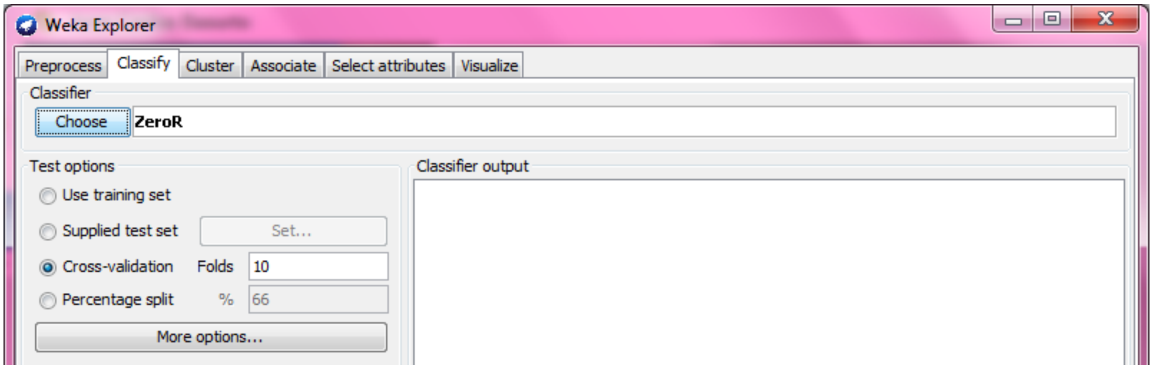
Status

OK

Running the algorithm

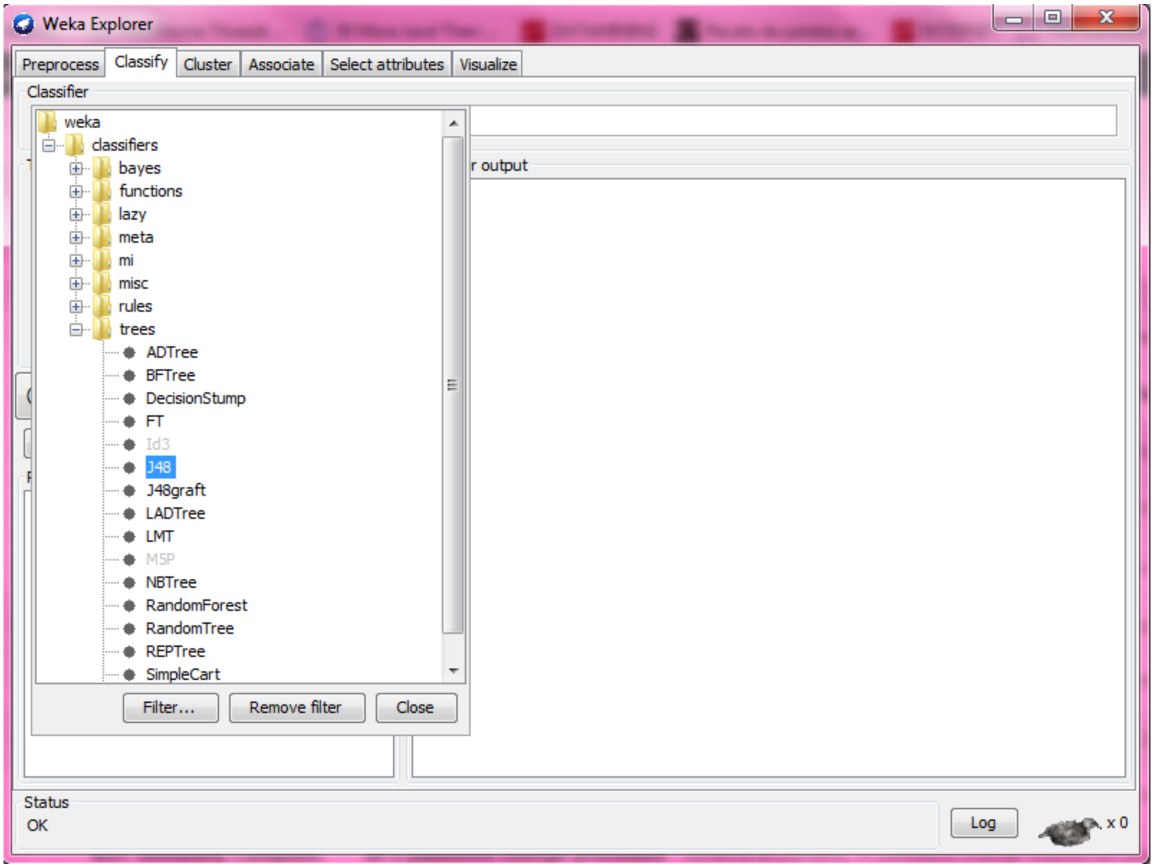
Note: in WEKA the ID3 algorithm only operates on nominal attributes; and because the attributes of our dataset are numeric the J48 algorithm will be used instead. It must be clear that the ID3 algorithm is the precursor of the J48 algorithm so won't be any alteration in the results because of this change.

1. **Setting the Algorithm:** Select the *Classify* tab*.* Then click on the *Choose* button below the text "Classifier" and next to the text field. Then, in the list of algorithm displayed we proceed to choose the option classifiers-> trees -> J48.



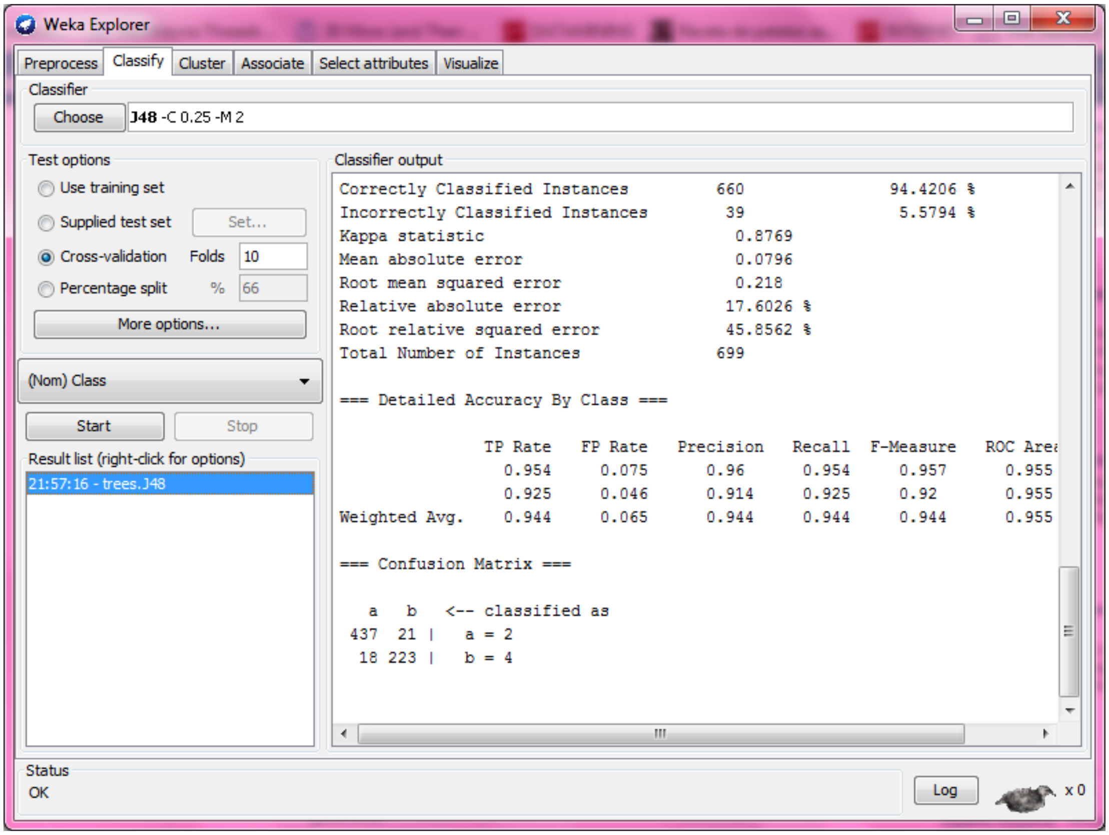
Choose ZeroR

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



Status
OK

To do the evaluation using the 10-Fold Cross Validation, we need to set this method in the *Test options* area by choosing the *Cross-validation* option. Also we need to set the number of folds as ten. To run the algorithm we click on the *Start* button and WEKA will proceed to make the evaluations in the dataset and present the results in the *Classifier output* area.



J48 -C 0.25 -M 2

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

More options...

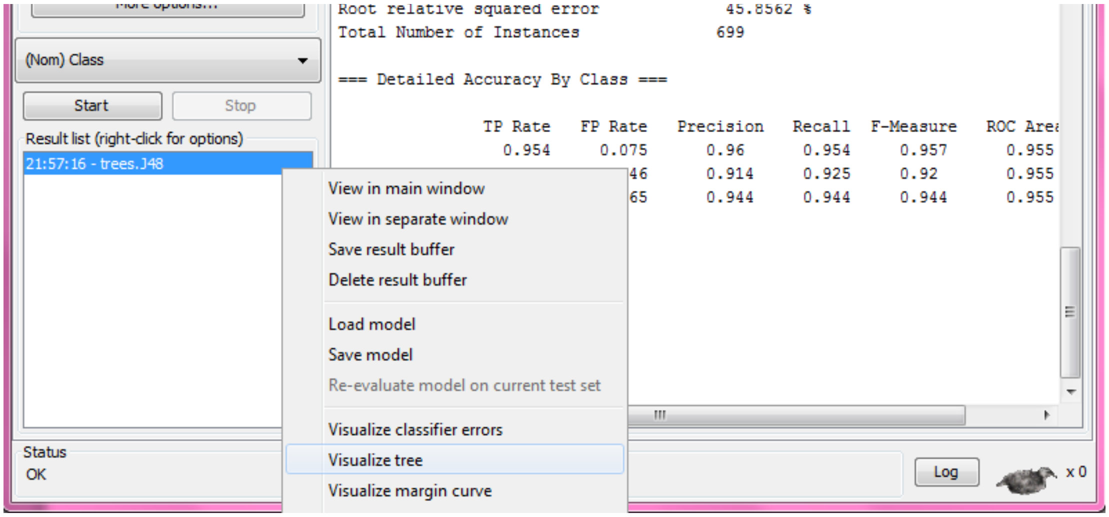
(Nom) Class

21:57:16 - trees.J48

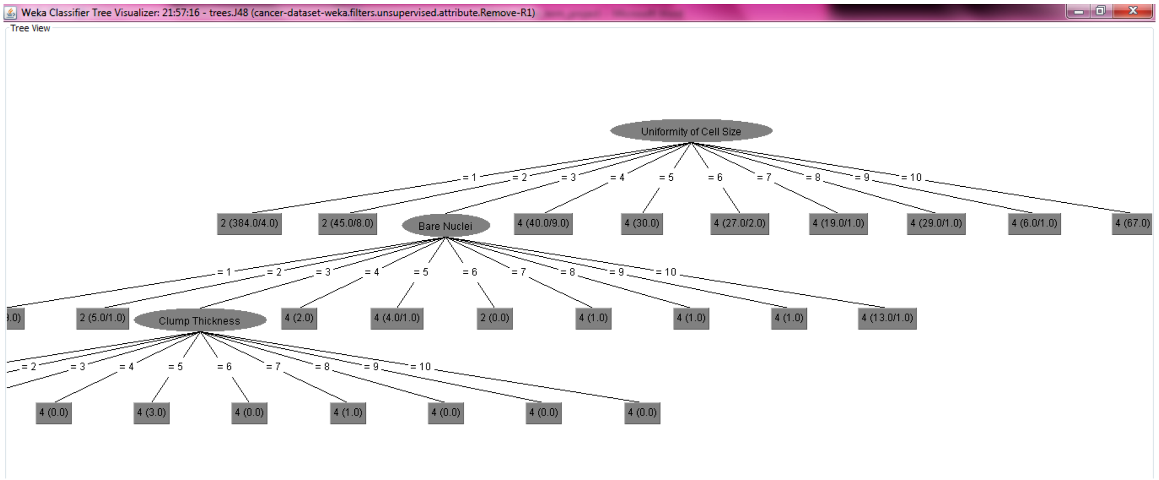
Status: OK

```
Correctly Classified Instances         660               94.4206 %
Incorrectly Classified Instances        39                5.5794 %
Kappa statistic                          0.8769
Mean absolute error                      0.0796
Root mean squared error                  0.218
Relative absolute error                 17.6026 %
Root relative squared error             45.8562 %
Total Number of Instances              699

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.954     0.075     0.96      0.954     0.957      0.955
                 0.925     0.046     0.914     0.925     0.92       0.955
Weighted Avg.    0.944     0.065     0.944     0.944     0.944      0.955

=== Confusion Matrix ===

   a   b   <-- classified as
 437  21 |   a = 2
  18 223 |   b = 4
```

To view the decision tree of the dataset make a right click on the result description displayed in the *Result list* area. Then select the option *Visualize tree*, a new window will appear with the decision tree display on it.



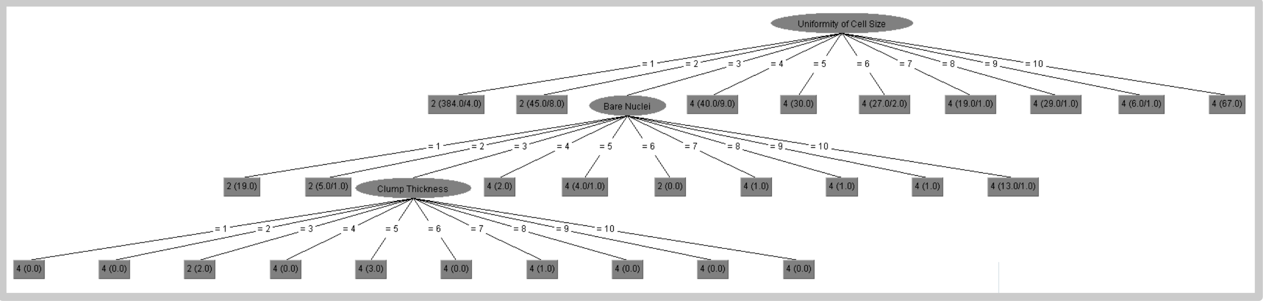
```
Root relative squared error             45.8562 %
Total Number of Instances              699

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.954     0.075     0.96      0.954     0.957      0.955
                             46      0.914     0.925     0.92       0.955
                             65      0.944     0.944     0.944      0.955
```



Weka Classifier Tree Visualizer: 21:57:16 - trees.J48 (cancer-dataset-weka.filters.unsupervised.attribute.Remove-R1)

Decision Tree

- Number of Leaves : 28
- Size of the tree: 31



Classifier Output

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     cancer-dataset-weka.filters.unsupervised.attribute.Remove-R1
Instances:    699
Attributes:   10
              Clump Thickness
              Uniformity of Cell Size
              Uniformity of Cell Shape
              Marginal Adhesion
              Single Epithelial Cell Size
              Bare Nuclei
              Bland Chromatin
              Normal Nucleoli
              Mitoses
              Class
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Uniformity of Cell Size = 1: 2 (384.0/4.0)
Uniformity of Cell Size = 2: 2 (45.0/8.0)
Uniformity of Cell Size = 3
|   Bare Nuclei = 1: 2 (19.0)
|   Bare Nuclei = 2: 2 (5.0/1.0)
|   Bare Nuclei = 3
|   |   Clump Thickness = 1: 4 (0.0)
|   |   Clump Thickness = 2: 4 (0.0)
|   |   Clump Thickness = 3: 2 (2.0)
|   |   Clump Thickness = 4: 4 (0.0)
|   |   Clump Thickness = 5: 4 (3.0)
|   |   Clump Thickness = 6: 4 (0.0)
|   |   Clump Thickness = 7: 4 (1.0)
|   |   Clump Thickness = 8: 4 (0.0)
|   |   Clump Thickness = 9: 4 (0.0)
|   |   Clump Thickness = 10: 4 (0.0)
|   Bare Nuclei = 4: 4 (2.0)
|   Bare Nuclei = 5: 4 (4.0/1.0)
|   Bare Nuclei = 6: 2 (0.0)
|   Bare Nuclei = 7: 4 (1.0)
|   Bare Nuclei = 8: 4 (1.0)
|   Bare Nuclei = 9: 4 (1.0)
|   Bare Nuclei = 10: 4 (13.0/1.0)
```

```
Uniformity of Cell Size = 4: 4 (40.0/9.0)
Uniformity of Cell Size = 5: 4 (30.0)
Uniformity of Cell Size = 6: 4 (27.0/2.0)
Uniformity of Cell Size = 7: 4 (19.0/1.0)
Uniformity of Cell Size = 8: 4 (29.0/1.0)
Uniformity of Cell Size = 9: 4 (6.0/1.0)
Uniformity of Cell Size = 10: 4 (67.0)

Number of Leaves  :     28

Size of the tree :      31


Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         660               94.4206 %
Incorrectly Classified Instances        39                5.5794 %
Kappa statistic                          0.8769
Mean absolute error                      0.0796
Root mean squared error                  0.218
Relative absolute error                 17.6026 %
Root relative squared error             45.8562 %
Total Number of Instances              699

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.954     0.075      0.96      0.954     0.957      0.955    2
                 0.925     0.046      0.914     0.925     0.92       0.955    4
Weighted Avg.    0.944     0.065      0.944     0.944     0.944      0.955

=== Confusion Matrix ===

   a   b   <-- classified as
 437  21 |   a = 2
  18 223 |   b = 4
```

Results

- True Benign: 437
- False Benign: 21
- True Malign: 223
- False Malign: 18

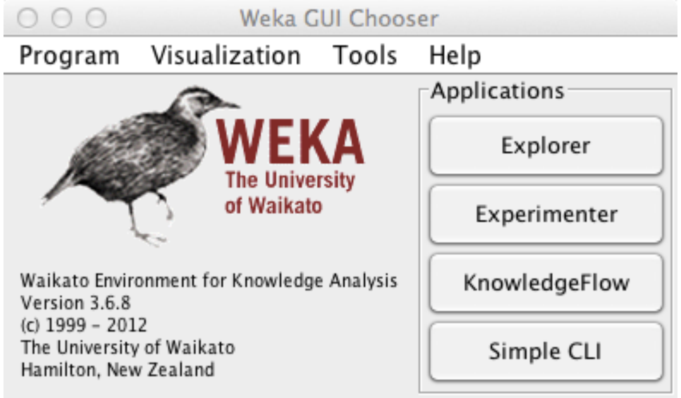
Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
2	0.954	0.075	0.96	0.954	0.957	0.955
4	0.925	0.046	0.914	0.925	0.92	0.955
Weighted Avg.	**0.944**	**0.065**	**0.94**	**0.944**	**0.944**	**0.955**

Support Vector Machines

Support Vector Machines are supervised models in data mining that analyze data, recognize patterns and classify the data into classes. In the dataset we will be using for this tool, we have to apply pre-processing to obtain discrete data before running the algorithm.

1. Starting Weka tool

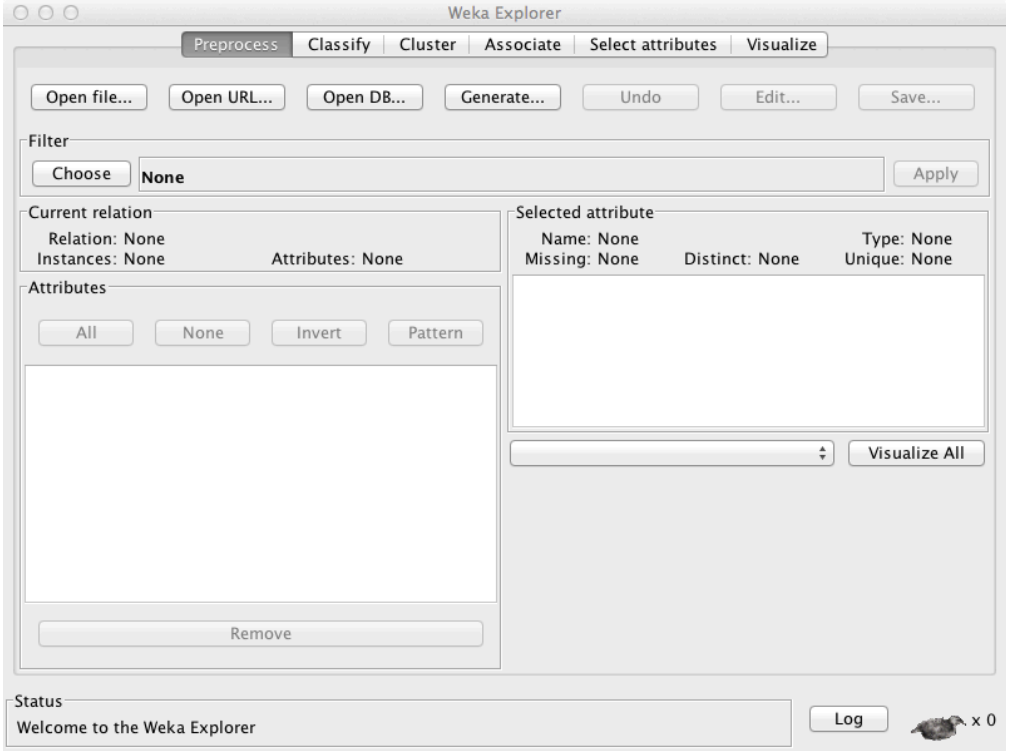
Open the Weka toolkit and click on the Explorer button.



Waikato Environment for Knowledge Analysis
Version 3.6.8
(c) 1999 - 2012
The University of Waikato
Hamilton, New Zealand

2. Loading Dataset

After clicking on the Explorer button the following window will open.



Filter: None

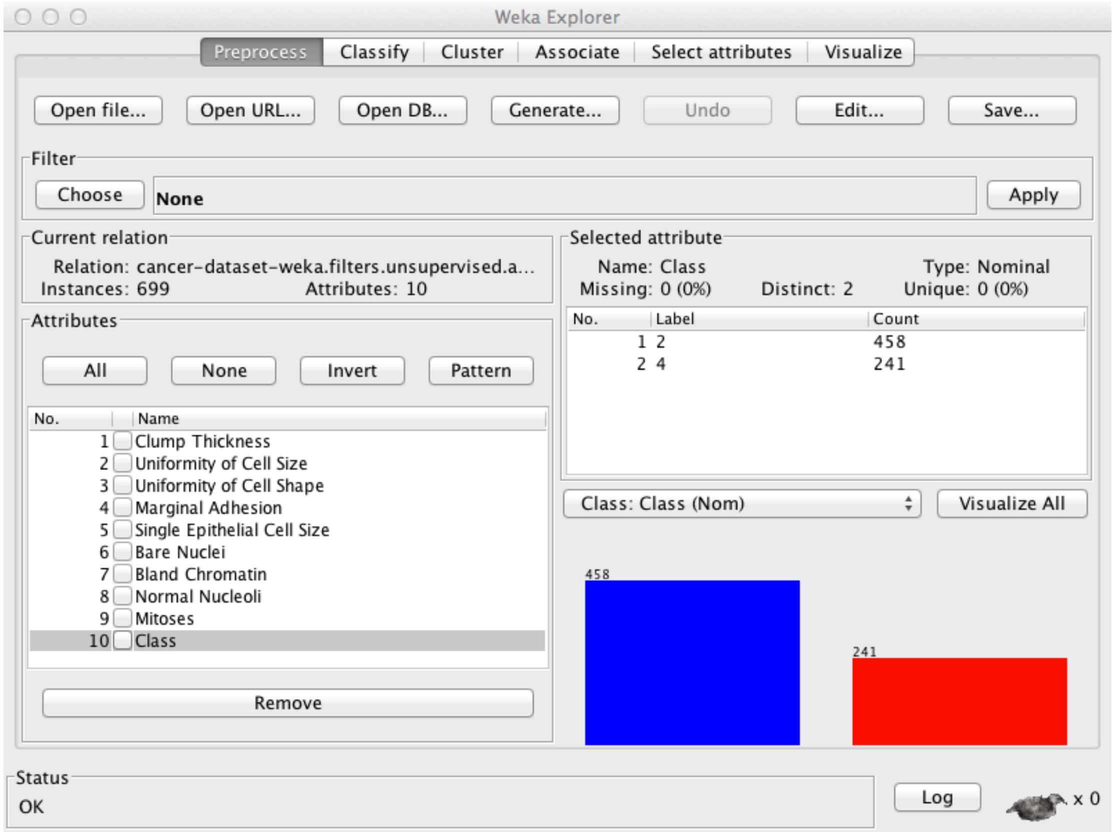
Relation: None
Instances: None
Attributes: None

Name: None
Type: None
Missing: None
Distinct: None
Unique: None

Welcome to the Weka Explorer

Now, Click on the “Open file…” button and select you dataset(it can be a .csv or .arff file). If it is a .csv file, you will have to discretize the dataset following steps for data-preprocessing as elaborated in the Decision Tree(C4.5) tutorial. After than you can continue with the remaining steps of this tutorial.

Once your .arff file has been loaded, Weka will display different colors depending on the class which the case belongs. *Blue* for benign cases and *Red* for malign cases.



Filter: None

Current relation
Relation: cancer-dataset-weka.filters.unsupervised.a...
Instances: 699 Attributes: 10

Selected attribute
Name: Class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	2	458
2	4	241

Attributes

No.	Name
1	☐ Clump Thickness
2	☐ Uniformity of Cell Size
3	☐ Uniformity of Cell Shape
4	☐ Marginal Adhesion
5	☐ Single Epithelial Cell Size
6	☐ Bare Nuclei
7	☐ Bland Chromatin
8	☐ Normal Nucleoli
9	☐ Mitoses
10	☐ Class

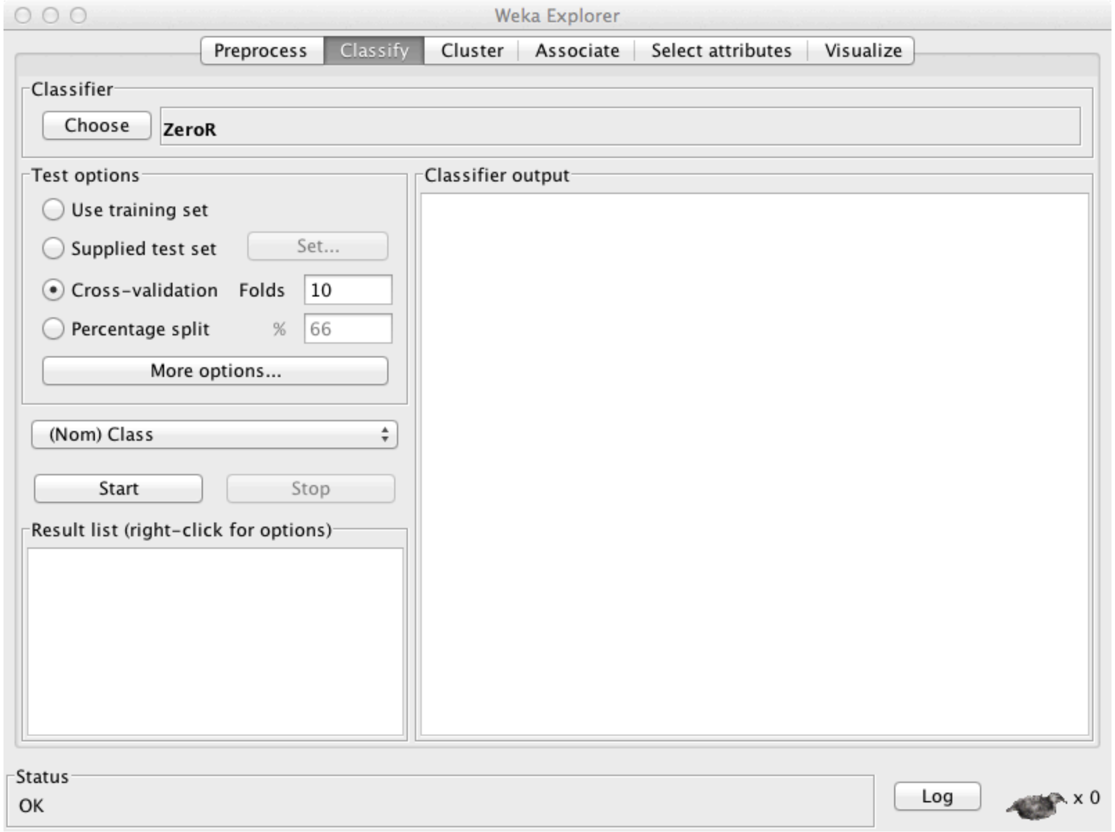
Class: Class (Nom)

Status
OK

Running the Algorithm

3. Classifying

Now, Click on the “Classify” tab at the top of the window and the window should look like this,



Classifier: ZeroR

● Cross-validation Folds 10

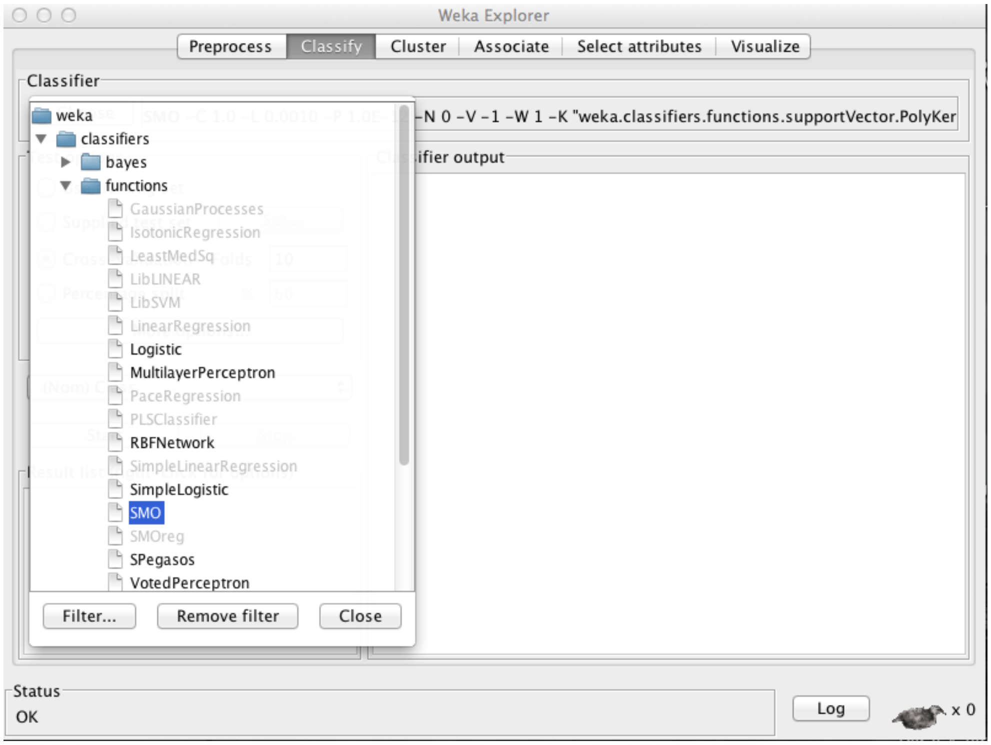
○ Percentage split % 66

(Nom) Class

Status
OK

4.Selecting Classifier Filter

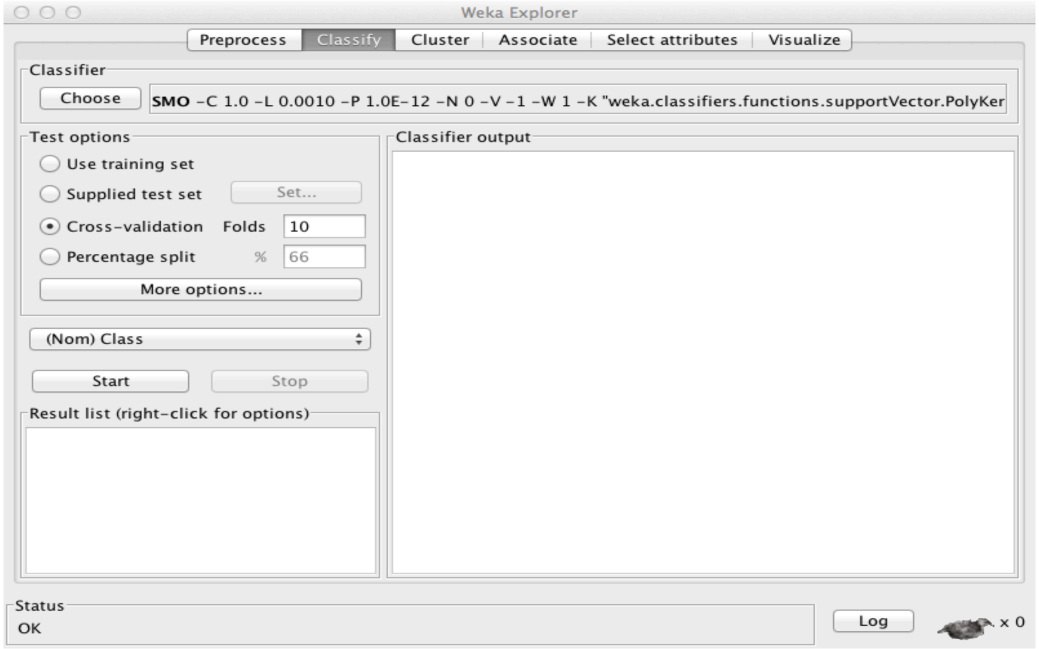
Now, We select the Classifier, click on the “Choose” button below Classifier and select the “SMO” filter from the given option . SMO is a SVM filter.



```
-N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKer
```

5.Running the SVM classifier

Now, under Test options, select “Cross-Validation” with 10 folds(you can specify a different number of folds as per your requirement).



Classifier

SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKer

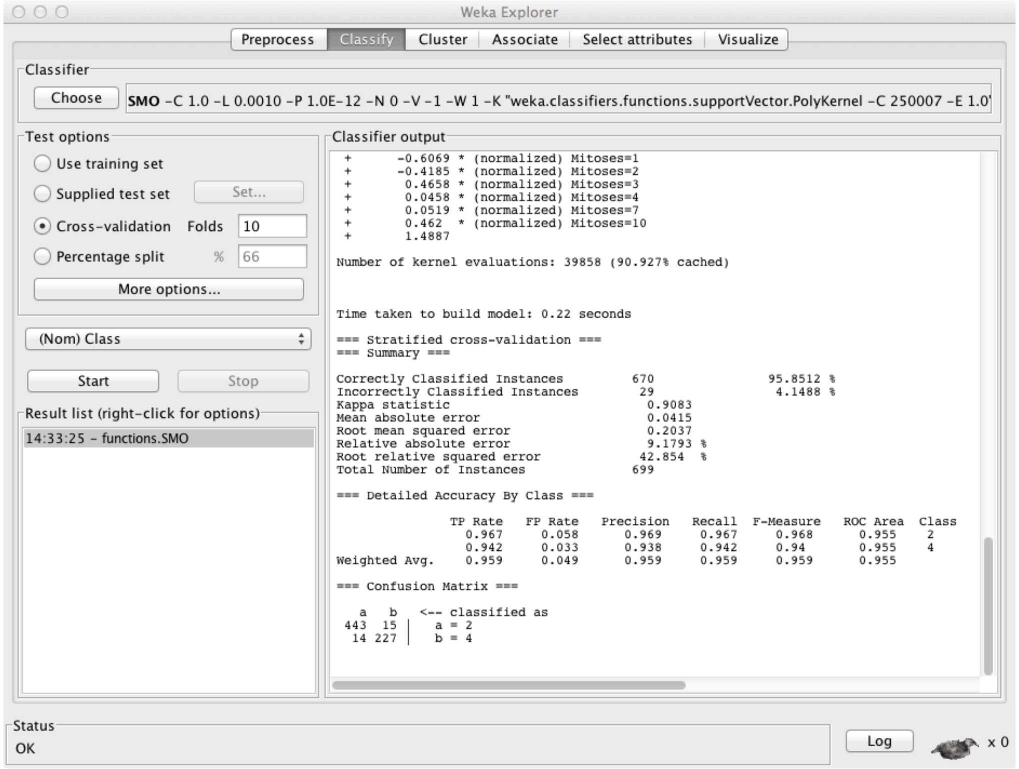
Test options

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

(Nom) Class

Status
OK

After selecting the Test option, Click start and Weka will run the SVM and the result will be displayed as per the next window.



Classifier

SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0'

Test options

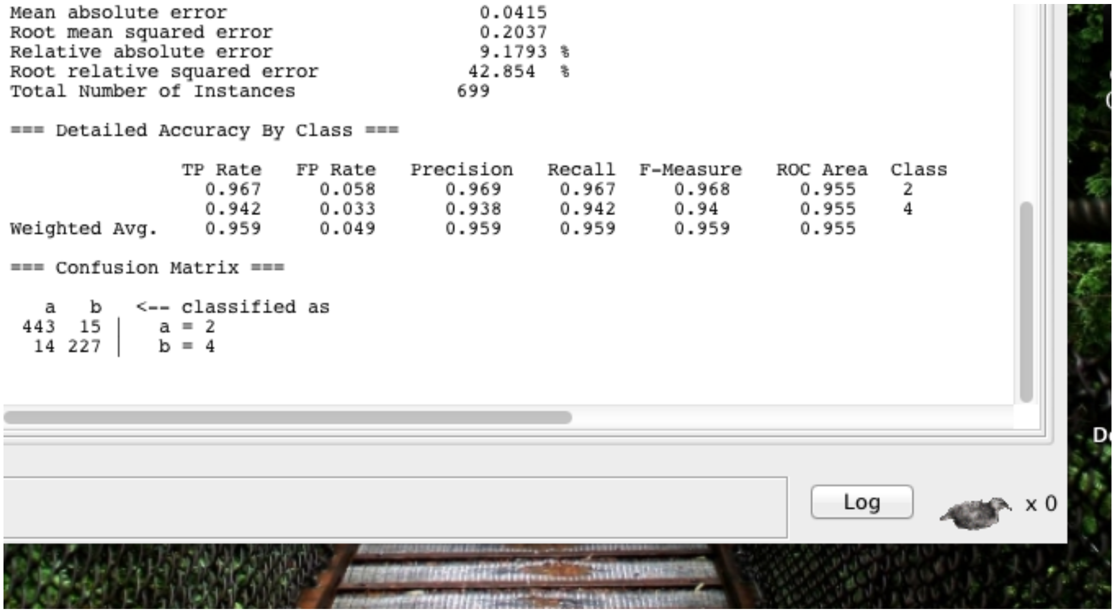
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Class

14:33:25 - functions.SMO

```
 +      -0.6069 * (normalized) Mitoses=1
 +      -0.4185 * (normalized) Mitoses=2
 +       0.4658 * (normalized) Mitoses=3
 +       0.0458 * (normalized) Mitoses=4
 +       0.0519 * (normalized) Mitoses=7
 +       0.462  * (normalized) Mitoses=10
 +       1.4887

Number of kernel evaluations: 39858 (90.927% cached)



Time taken to build model: 0.22 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         670               95.8512 %
Incorrectly Classified Instances        29                4.1488 %
Kappa statistic                          0.9083
Mean absolute error                      0.0415
Root mean squared error                  0.2037
Relative absolute error                  9.1793 %
Root relative squared error             42.854  %
Total Number of Instances              699

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.967     0.058      0.969     0.967     0.968      0.955    2
                 0.942     0.033      0.938     0.942     0.94       0.955    4
Weighted Avg.    0.959     0.049      0.959     0.959     0.959      0.955

=== Confusion Matrix ===

   a   b   <-- classified as
 443  15 |   a = 2
  14 227 |   b = 4
```

Status
OK



```
Mean absolute error                      0.0415
Root mean squared error                  0.2037
Relative absolute error                  9.1793 %
Root relative squared error             42.854  %
Total Number of Instances              699

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.967     0.058      0.969     0.967     0.968      0.955    2
                 0.942     0.033      0.938     0.942     0.94       0.955    4
Weighted Avg.    0.959     0.049      0.959     0.959     0.959      0.955

=== Confusion Matrix ===

   a   b   <-- classified as
 443  15 |   a = 2
  14 227 |   b = 4
```

The output obtained gives us the true positive, false positive , precision , recall ROC for the test sample. This allows us to grade the accuracy for the classifier.

Classifier Output

=== Run information ===

Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"

Relation: cancer-dataset-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-R1,2,3,4,5,6,7,8,9

Instances: 699

Attributes: 10

Clump Thickness

Uniformity of Cell Size

Uniformity of Cell Shape

```
              Marginal Adhesion
              Single Epithelial Cell Size
              Bare Nuclei
              Bland Chromatin
              Normal Nucleoli
              Mitoses
              Class
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

Classifier for classes: 2, 4

BinarySMO

Machine linear: showing attribute weights, not support vectors.

        -0.7999 * (normalized) Clump Thickness=1
 +       0.1609 * (normalized) Clump Thickness=2
 +      -0.0973 * (normalized) Clump Thickness=3
 +       -0.4661 * (normalized) Clump Thickness=4
 +       -0.4235 * (normalized) Clump Thickness=5
 +       -0.757  * (normalized) Clump Thickness=6
 +        0.7853 * (normalized) Clump Thickness=7
 +        0.0781 * (normalized) Clump Thickness=8
 +        0.377  * (normalized) Clump Thickness=9
 +        1.1424 * (normalized) Clump Thickness=10
 +       -0.5701 * (normalized) Uniformity of Cell Size=1
 +        0.0147 * (normalized) Uniformity of Cell Size=2
 +       -0.054  * (normalized) Uniformity of Cell Size=3
 +       -0.3618 * (normalized) Uniformity of Cell Size=4
 +        0.1242 * (normalized) Uniformity of Cell Size=5
 +        0.1607 * (normalized) Uniformity of Cell Size=6
 +       -0.2776 * (normalized) Uniformity of Cell Size=7
 +        0.0331 * (normalized) Uniformity of Cell Size=8
 +        0.0828 * (normalized) Uniformity of Cell Size=9
 +        0.848  * (normalized) Uniformity of Cell Size=10
 +       -0.6141 * (normalized) Uniformity of Cell Shape=1
 +       -0.8623 * (normalized) Uniformity of Cell Shape=2
 +        0.3367 * (normalized) Uniformity of Cell Shape=3
 +        0.0722 * (normalized) Uniformity of Cell Shape=4
 +        0.5204 * (normalized) Uniformity of Cell Shape=5
 +       -0.023  * (normalized) Uniformity of Cell Shape=6
 +       -0.3138 * (normalized) Uniformity of Cell Shape=7
 +        0.2503 * (normalized) Uniformity of Cell Shape=8
```

```
+       0.4934 * (normalized) Uniformity of Cell Shape=9
+       0.14   * (normalized) Uniformity of Cell Shape=10
+      -0.2767 * (normalized) Marginal Adhesion=1
+      -0.2161 * (normalized) Marginal Adhesion=2
+      -0.4446 * (normalized) Marginal Adhesion=3
+       0.5876 * (normalized) Marginal Adhesion=4
+      -0.215  * (normalized) Marginal Adhesion=5
+       0.0432 * (normalized) Marginal Adhesion=6
+       0      * (normalized) Marginal Adhesion=7
+       0.1526 * (normalized) Marginal Adhesion=8
+      -0.0749 * (normalized) Marginal Adhesion=9
+       0.444  * (normalized) Marginal Adhesion=10
+       0.0491 * (normalized) Single Epithelial Cell Size=1
+      -0.093  * (normalized) Single Epithelial Cell Size=2
+      -0.1685 * (normalized) Single Epithelial Cell Size=3
+       0.1564 * (normalized) Single Epithelial Cell Size=4
+       0.3348 * (normalized) Single Epithelial Cell Size=5
+       0.2325 * (normalized) Single Epithelial Cell Size=6
+      -1.2362 * (normalized) Single Epithelial Cell Size=7
+      -0.2752 * (normalized) Single Epithelial Cell Size=8
+       1      * (normalized) Single Epithelial Cell Size=10
+      -1.0405 * (normalized) Bare Nuclei=1
+      -0.5017 * (normalized) Bare Nuclei=2
+      -0.3412 * (normalized) Bare Nuclei=3
+      -0.2828 * (normalized) Bare Nuclei=4
```

```
 +      -0.15   * (normalized) Bare Nuclei=5
 +       0.8277 * (normalized) Bare Nuclei=6
 +      -0.2789 * (normalized) Bare Nuclei=7
 +       0.2527 * (normalized) Bare Nuclei=8
 +       0.7264 * (normalized) Bare Nuclei=9
 +       0.7882 * (normalized) Bare Nuclei=10
 +      -0.6998 * (normalized) Bland Chromatin=1
 +      -0.3522 * (normalized) Bland Chromatin=2
 +      -0.1926 * (normalized) Bland Chromatin=3
 +      -0.0886 * (normalized) Bland Chromatin=4
 +       0.0426 * (normalized) Bland Chromatin=5
 +      -0.1697 * (normalized) Bland Chromatin=6
 +       0.1967 * (normalized) Bland Chromatin=7
 +       0.3212 * (normalized) Bland Chromatin=8
 +       0.2344 * (normalized) Bland Chromatin=9
 +       0.708  * (normalized) Bland Chromatin=10
 +      -0.1236 * (normalized) Normal Nucleoli=1
 +      -1.1778 * (normalized) Normal Nucleoli=2
 +       0.1662 * (normalized) Normal Nucleoli=3
 +       0.2401 * (normalized) Normal Nucleoli=4
 +      -0.2096 * (normalized) Normal Nucleoli=5
 +       0.1674 * (normalized) Normal Nucleoli=6
 +      -0.1995 * (normalized) Normal Nucleoli=7
 +      -0.5755 * (normalized) Normal Nucleoli=8
 +       0.3762 * (normalized) Normal Nucleoli=9
 +       1.3362 * (normalized) Normal Nucleoli=10
 +      -0.6069 * (normalized) Mitoses=1
 +      -0.4185 * (normalized) Mitoses=2
 +       0.4658 * (normalized) Mitoses=3
 +       0.0458 * (normalized) Mitoses=4
 +       0.0519 * (normalized) Mitoses=7
 +       0.462  * (normalized) Mitoses=10
 +       1.4887

Number of kernel evaluations: 39858 (90.927% cached)



Time taken to build model: 0.24 seconds


=== Stratified cross-validation ===
=== Summary ===


Correctly Classified Instances         670               95.8512 %
Incorrectly Classified Instances        29                4.1488 %
Kappa statistic                          0.9083
Mean absolute error                      0.0415
Root mean squared error                  0.2037
Relative absolute error                  9.1793 %
Root relative squared error             42.854  %
```

```
Total Number of Instances             699


=== Detailed Accuracy By Class ===


             TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class

               0.967     0.058     0.969     0.967     0.968      0.955    2

               0.942     0.033     0.938     0.942     0.94       0.955    4

Weighted Avg.    0.959     0.049     0.959     0.959     0.959      0.955



=== Confusion Matrix ===


   a   b   <-- classified as

 443  15 |   a = 2

  14 227 |   b = 4
```

Results

- True Benign: 443
- False Benign: 15
- True Malign: 227
- False Malign: 14

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
2	0.967	0.058	0.969	0.967	0.968	0.955
4	0.942	0.033	0.938	0.942	0.94	0.955
Weighted Avg.	**0.959**	**0.049**	**0.959**	**0.959**	**0.959**	**0.955**

Cross Validation Results and Comparisons

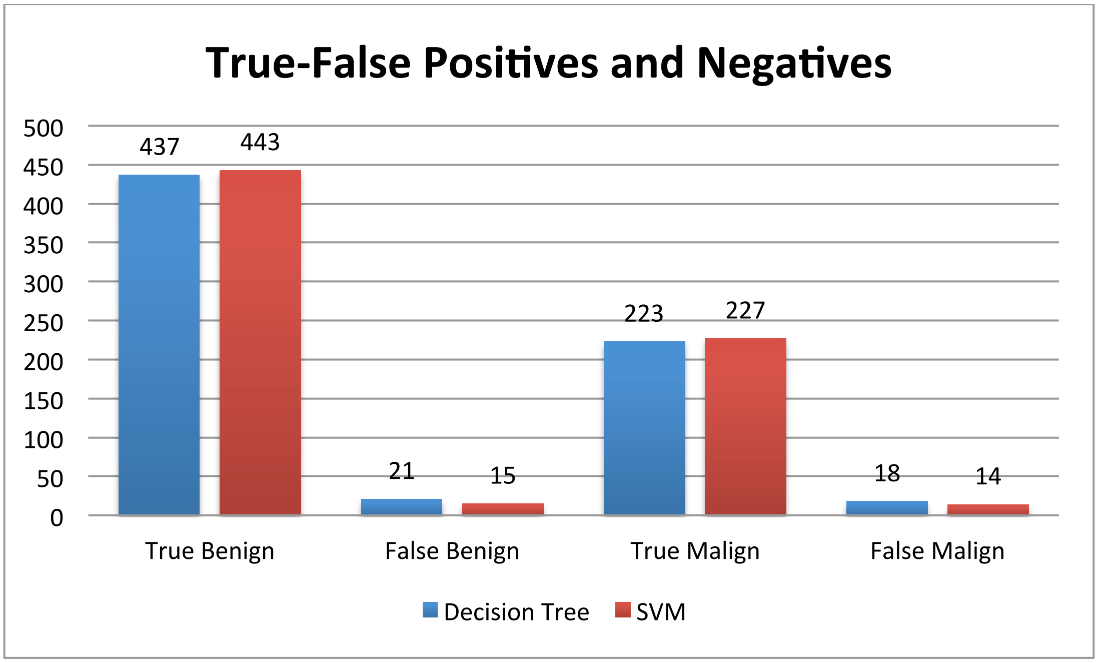


Figure 2 - True - False Positives and Negatives chart

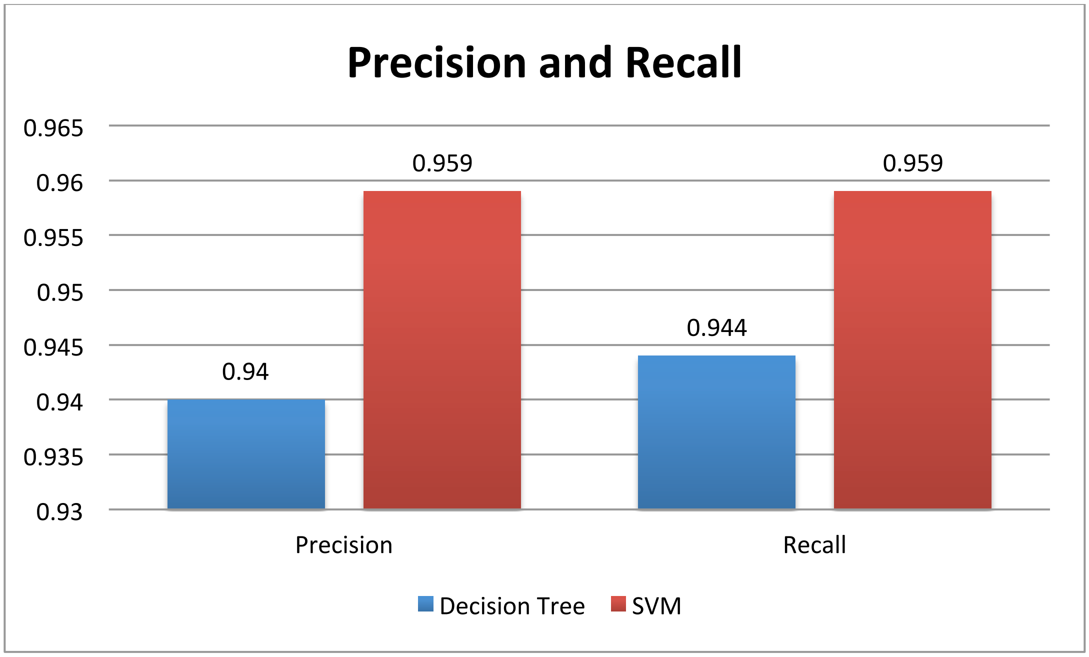


Figure 3 - Precision and Recall

With Respect to the charts above, we can infer from the results that the SVM algorithm had better performance compared to the Decision tree algorithm. As we can observe from Figure 2, that both algorithms gave us close results. However the SVM algorithm predicted more cases correctly than Decision tree algorithm as well as the number of cases predicted wrong were less compared to decision tree. Figure 3 shows that SVM classifies the cases more accurately and the results are more relevant. So, we can conclude that SVM is the more optimal algorithm that can be used for this dataset.