Otto Friedrich University Bamberg

Data Streams and Complex Event Processing

Assignment - 05

Supervisor:
Prof. Dr. Daniela Nicklas
Mobile Software Systems

Submitted By :
Manasa Reddy Bandari - 1942692
Akshay Sharma - 1943984
Kummari Daniel Raj - 1973448

Feb 4, 2019

Contents

1 Exercise 1:

To investigate the influence of streamed continuous data on the efficiency of the learning results, we use two different datasets: one is from cars sales data and the second one is from the health sector (both data sets can be found on the VC under the folder "Datasets"). The goal of the task is to observe how a classification algorithm creates classification models based on varying sizes of data. By using different sizes for the training and testing, we try to imitate the behavior of streamed data. The class to be observed in the cars dataset is the price and in the second dataset, the final attribute is the class. (increased binding protein, decreased binding protein and negative)

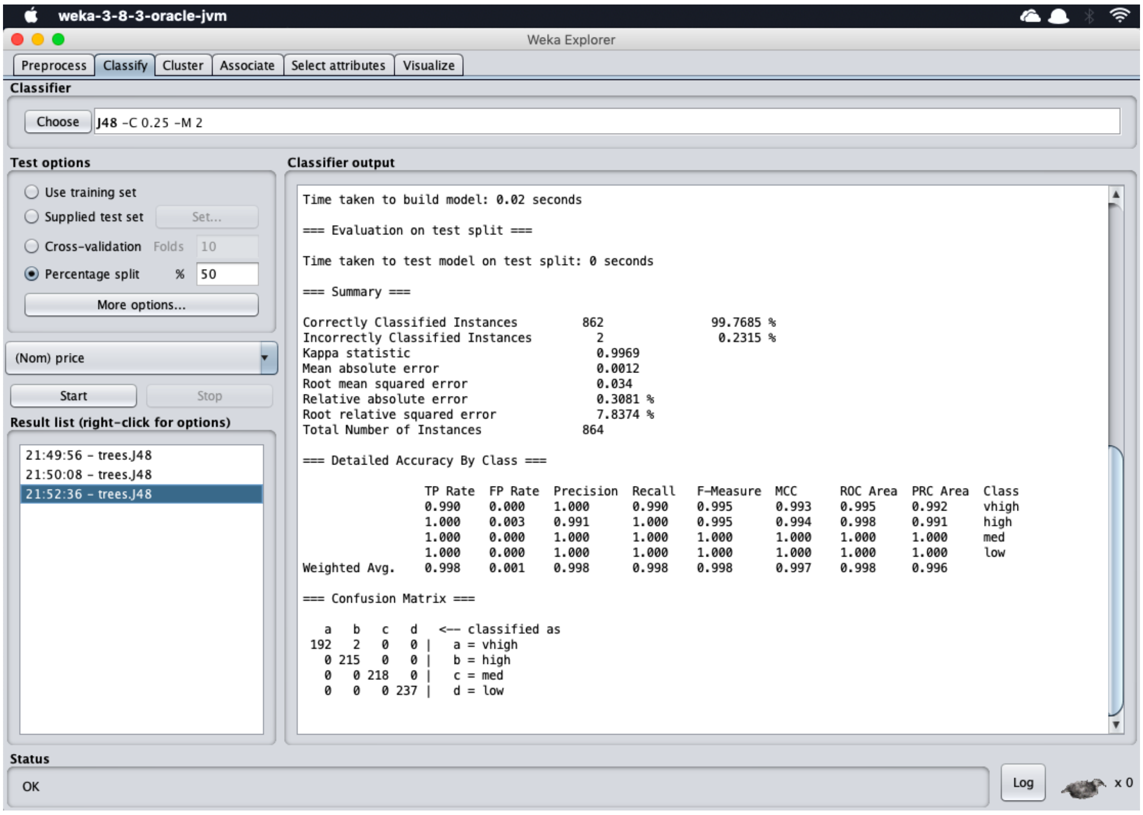
1.1 Have a look at both datasets and give a brief explanation on both datasets and what do they describe?

Car Dataset includes characteristics like number of people, number of doors, overall cost, safety values and much more. There are dependencies in between many characteristics like capacity and boot space. It shows an entire domain to check and decide the best option for the desired input.

Health Dataset includes information about the person's age and gender and will determine on the number of factors whether they have thyroid or not.

1.2 To learn the effect of decreasing training set on the data, we use the first dataset (cars) with the classification algorithm J48 and record the precision results in three different settings:

1.2.1 Use 50% of the dataset for training and the other 50% for testing



J48 -C 0.25 -M 2

● Percentage split % 50

(Nom) price

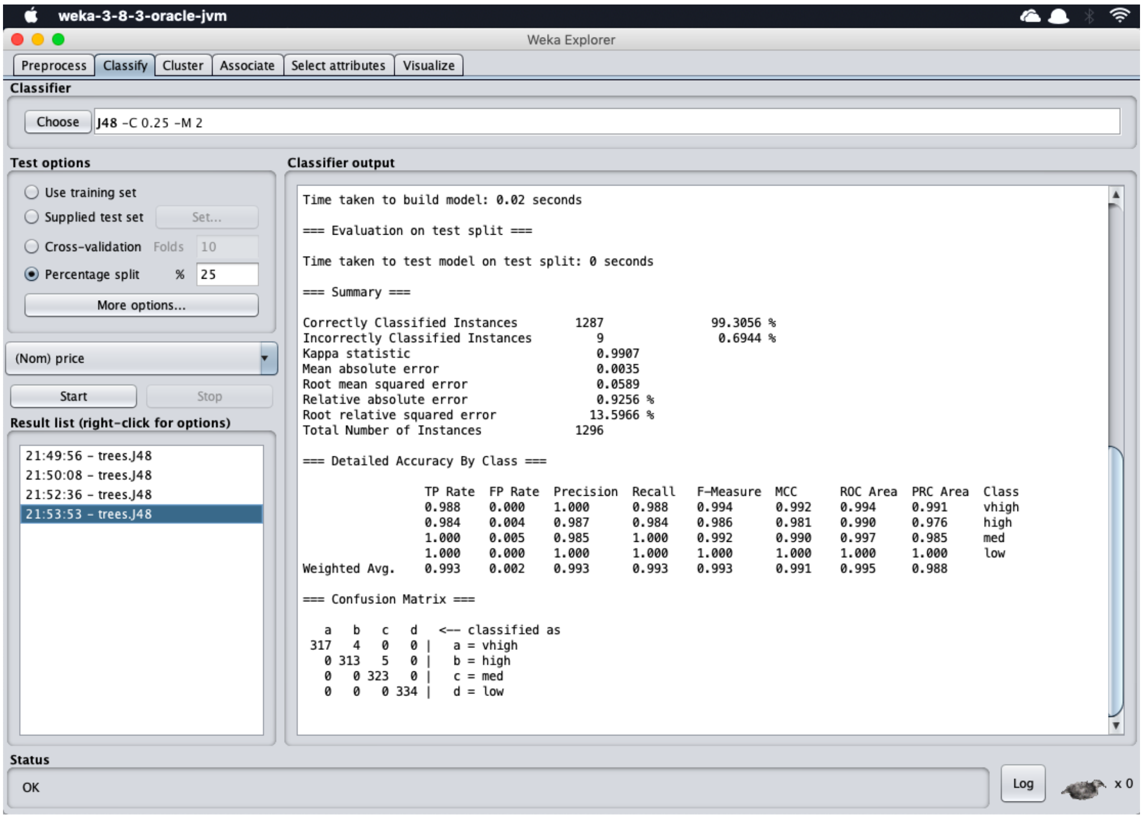
21:49:56 - trees.J48
21:50:08 - trees.J48
21:52:36 - trees.J48

```
Time taken to build model: 0.02 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         862               99.7685 %
Incorrectly Classified Instances         2                0.2315 %
Kappa statistic                          0.9969
Mean absolute error                      0.0012
Root mean squared error                  0.034
Relative absolute error                  0.3081 %
Root relative squared error              7.8374 %
Total Number of Instances              864

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.990    0.000    1.000      0.990    0.995      0.993    0.995     0.992     vhigh
                 1.000    0.003    0.991      1.000    0.995      0.994    0.998     0.991     high
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     med
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     low
Weighted Avg.    0.998    0.001    0.998      0.998    0.998      0.997    0.998     0.996

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 192   2   0   0 |   a = vhigh
   0 215   0   0 |   b = high
   0   0 218   0 |   c = med
   0   0   0 237 |   d = low
```

Status: OK

Figure 1: 50% Training Dataset

1.2.2 Use 25% of the dataset for training and the other 75% for testing



J48 -C 0.25 -M 2

● Percentage split % 25

(Nom) price

```
21:49:56 - trees.J48
21:50:08 - trees.J48
21:52:36 - trees.J48
21:53:53 - trees.J48
```

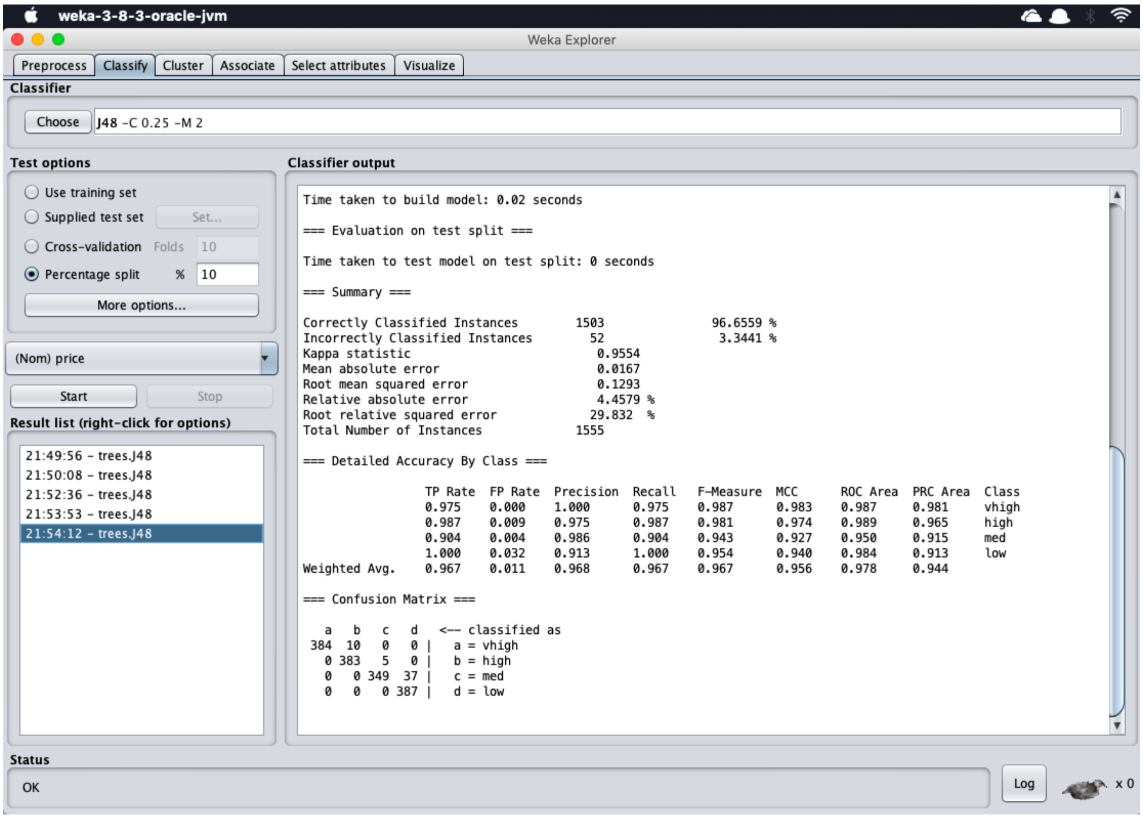
```
Time taken to build model: 0.02 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances        1287               99.3056 %
Incorrectly Classified Instances         9                0.6944 %
Kappa statistic                          0.9907
Mean absolute error                      0.0035
Root mean squared error                  0.0589
Relative absolute error                  0.9256 %
Root relative squared error             13.5966 %
Total Number of Instances             1296

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.988    0.000    1.000      0.988    0.994      0.992    0.994     0.991     vhigh
                 0.984    0.004    0.987      0.984    0.986      0.981    0.990     0.976     high
                 1.000    0.005    0.985      1.000    0.992      0.990    0.997     0.985     med
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     low
Weighted Avg.    0.993    0.002    0.993      0.993    0.993      0.991    0.995     0.988

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 317   4   0   0 |   a = vhigh
   0 313   5   0 |   b = high
   0   0 323   0 |   c = med
   0   0   0 334 |   d = low
```

Status

OK

Figure 2: 25% Training Dataset

1.2.3 Use 10% of the dataset for training and the other 90% for testing



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 10

(Nom) price

Result list (right-click for options)

21:49:56 - trees.J48
21:50:08 - trees.J48
21:52:36 - trees.J48
21:53:53 - trees.J48
21:54:12 - trees.J48

Classifier output

```
Time taken to build model: 0.02 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances        1503               96.6559 %
Incorrectly Classified Instances        52                3.3441 %
Kappa statistic                          0.9554
Mean absolute error                      0.0167
Root mean squared error                  0.1293
Relative absolute error                  4.4579 %
Root relative squared error             29.832  %
Total Number of Instances             1555

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.975    0.000    1.000      0.975    0.987      0.983    0.987     0.981     vhigh
                 0.987    0.009    0.975      0.987    0.981      0.974    0.989     0.965     high
                 0.904    0.004    0.986      0.904    0.943      0.927    0.950     0.915     med
                 1.000    0.032    0.913      1.000    0.954      0.940    0.984     0.913     low
Weighted Avg.    0.967    0.011    0.968      0.967    0.967      0.956    0.978     0.944

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 384  10   0   0 |   a = vhigh
   0 383   5   0 |   b = high
   0   0 349  37 |   c = med
   0   0   0 387 |   d = low
```

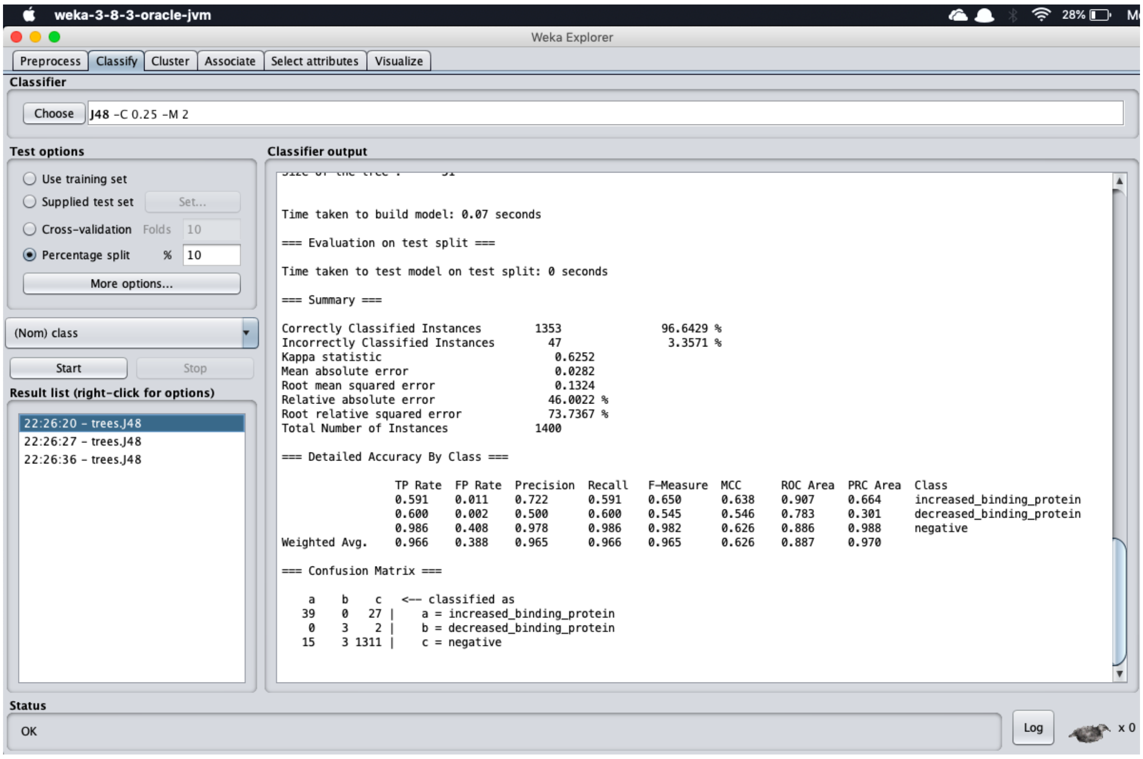
Status

OK

Figure 3: 10% Training Dataset

1.3 Use the second dataset (thyroid) with the same algorithm and record the precision results in three different settings:

1.3.1 Use 50% of the dataset for training and the other 50% for testing



J48 -C 0.25 -M 2

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

(Nom) class

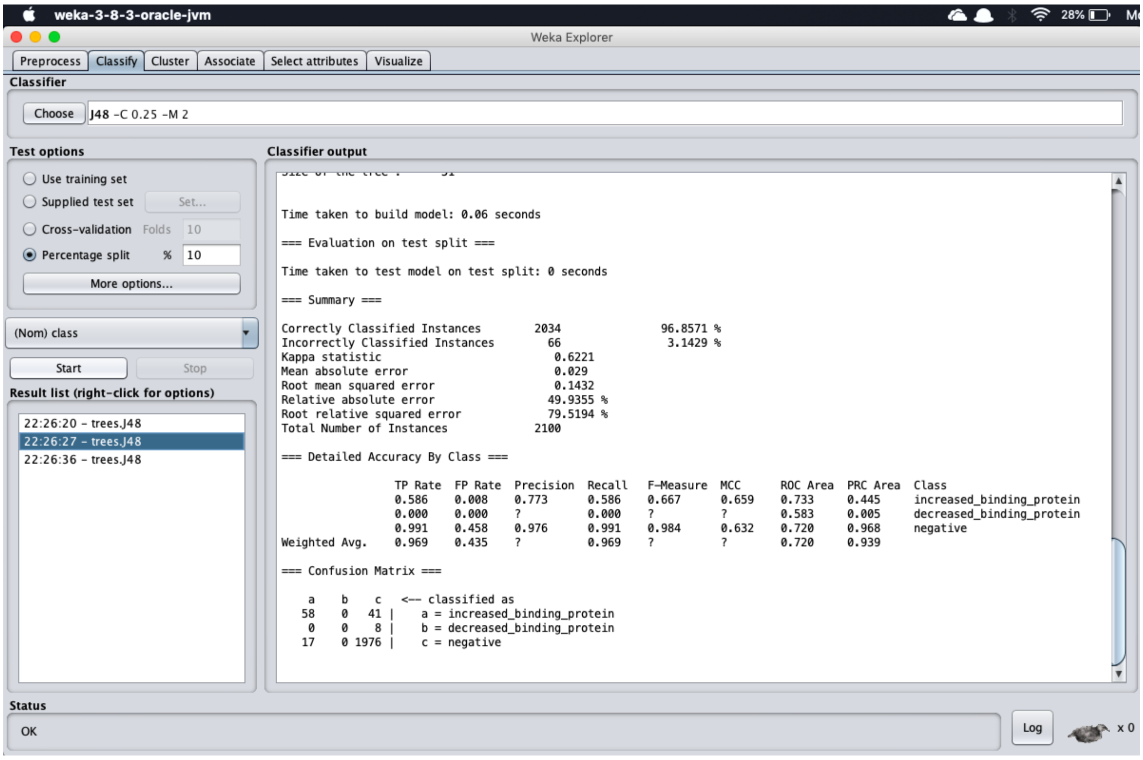
22:26:20 - trees.J48
22:26:27 - trees.J48
22:26:36 - trees.J48

```
Time taken to build model: 0.07 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances        1353               96.6429 %
Incorrectly Classified Instances        47                3.3571 %
Kappa statistic                          0.6252
Mean absolute error                      0.0282
Root mean squared error                  0.1324
Relative absolute error                 46.0022 %
Root relative squared error             73.7367 %
Total Number of Instances             1400

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.591    0.011    0.722      0.591    0.650      0.638    0.907     0.664     increased_binding_protein
                 0.600    0.002    0.500      0.600    0.545      0.546    0.783     0.301     decreased_binding_protein
                 0.986    0.408    0.978      0.986    0.982      0.626    0.886     0.988     negative
Weighted Avg.    0.966    0.388    0.965      0.966    0.965      0.626    0.887     0.970

=== Confusion Matrix ===

    a    b    c   <-- classified as
   39    0   27 |    a = increased_binding_protein
    0    3    2 |    b = decreased_binding_protein
   15    3 1311 |    c = negative
```

OK

Figure 4: 50% Training Dataset

1.3.2 Use 25% of the dataset for training and the other 75% for testing



Classifier: J48 -C 0.25 -M 2

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

(Nom) class

Result list (right-click for options)

```
22:26:20 - trees.J48
22:26:27 - trees.J48
22:26:36 - trees.J48
```

Classifier output

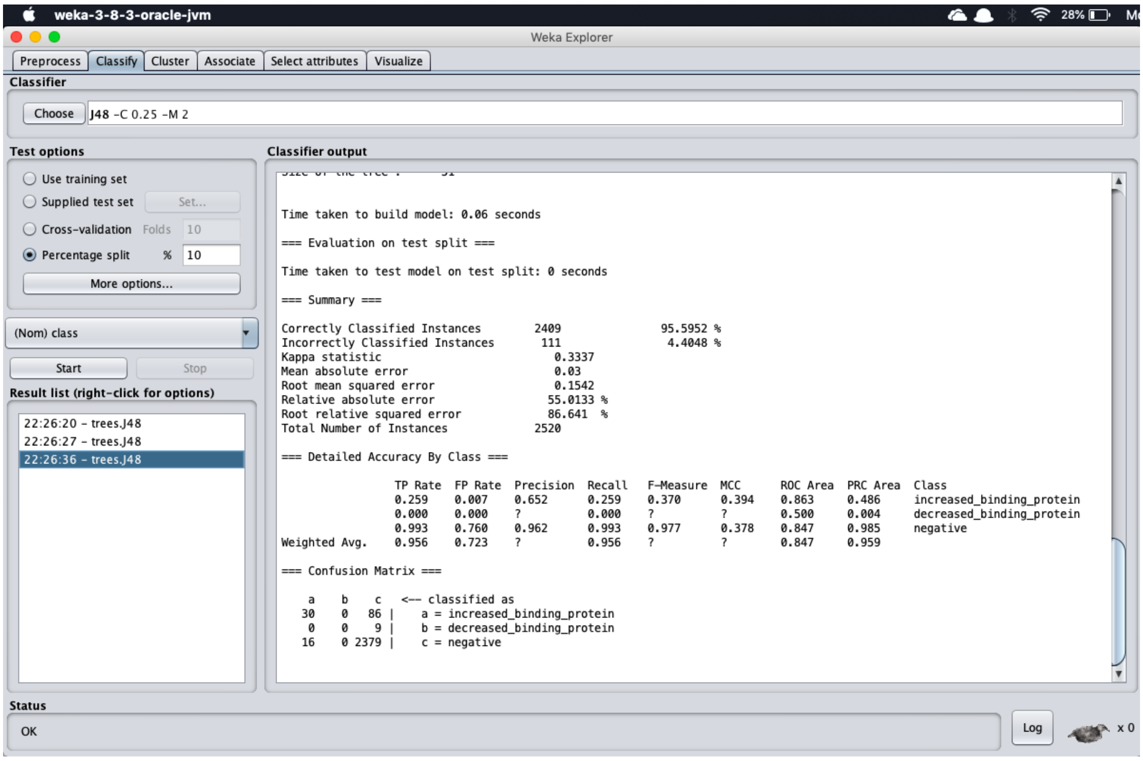
```
Time taken to build model: 0.06 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances        2034               96.8571 %
Incorrectly Classified Instances        66                3.1429 %
Kappa statistic                          0.6221
Mean absolute error                      0.029
Root mean squared error                  0.1432
Relative absolute error                 49.9355 %
Root relative squared error             79.5194 %
Total Number of Instances             2100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.586    0.008    0.773      0.586    0.667      0.659    0.733     0.445     increased_binding_protein
                 0.000    0.000    ?          0.000    ?          ?        0.583     0.005     decreased_binding_protein
                 0.991    0.458    0.976      0.991    0.984      0.632    0.720     0.968     negative
Weighted Avg.    0.969    0.435    ?          0.969    ?          ?        0.720     0.939

=== Confusion Matrix ===

    a    b    c   <-- classified as
   58    0   41 |    a = increased_binding_protein
    0    0    8 |    b = decreased_binding_protein
   17    0 1976 |    c = negative
```

Status

OK

Figure 5: 25% Training Dataset

1.3.3 Use 10% of the dataset for training and the other 90% for testing



Classifier: J48 -C 0.25 -M 2

● Percentage split % 10

(Nom) class

```
22:26:20 - trees.J48
22:26:27 - trees.J48
22:26:36 - trees.J48
```

```
Time taken to build model: 0.06 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances        2409               95.5952 %
Incorrectly Classified Instances       111                4.4048 %
Kappa statistic                          0.3337
Mean absolute error                      0.03
Root mean squared error                  0.1542
Relative absolute error                 55.0133 %
Root relative squared error             86.641  %
Total Number of Instances             2520

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.259    0.007    0.652      0.259    0.370      0.394    0.863     0.486     increased_binding_protein
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.004     decreased_binding_protein
                 0.993    0.760    0.962      0.993    0.977      0.378    0.847     0.985     negative
Weighted Avg.    0.956    0.723    ?          0.956    ?          ?        0.847     0.959

=== Confusion Matrix ===

    a    b    c   <-- classified as
   30    0   86 |    a = increased_binding_protein
    0    0    9 |    b = decreased_binding_protein
   16    0 2379 |    c = negative
```

Status: OK

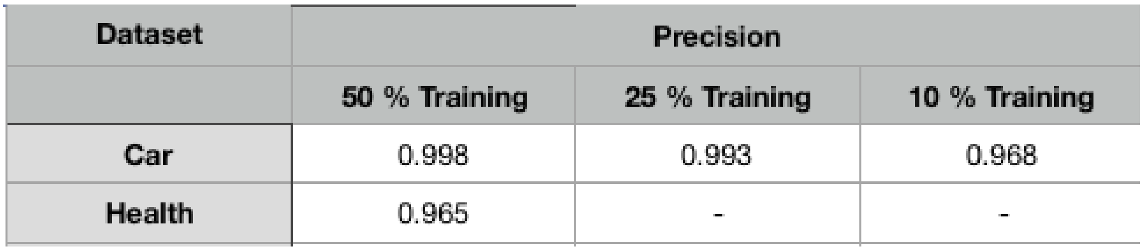
Figure 6: 10% Training Dataset

1.4 What are the effects of the different sizes of training and testing datasets on the learned models and accuracy?

Data is necessary for training to predict the accurate results so we reduce the training data percentage then we saw a gradual decline of the precision. In order to produce near accurate results, more training data is required.
Also, we noticed that the correctly classified instances were getting increased as the training percentage was getting increased.

1.5 Compare the different precision results between the two datasets and explain the change of accuracy.

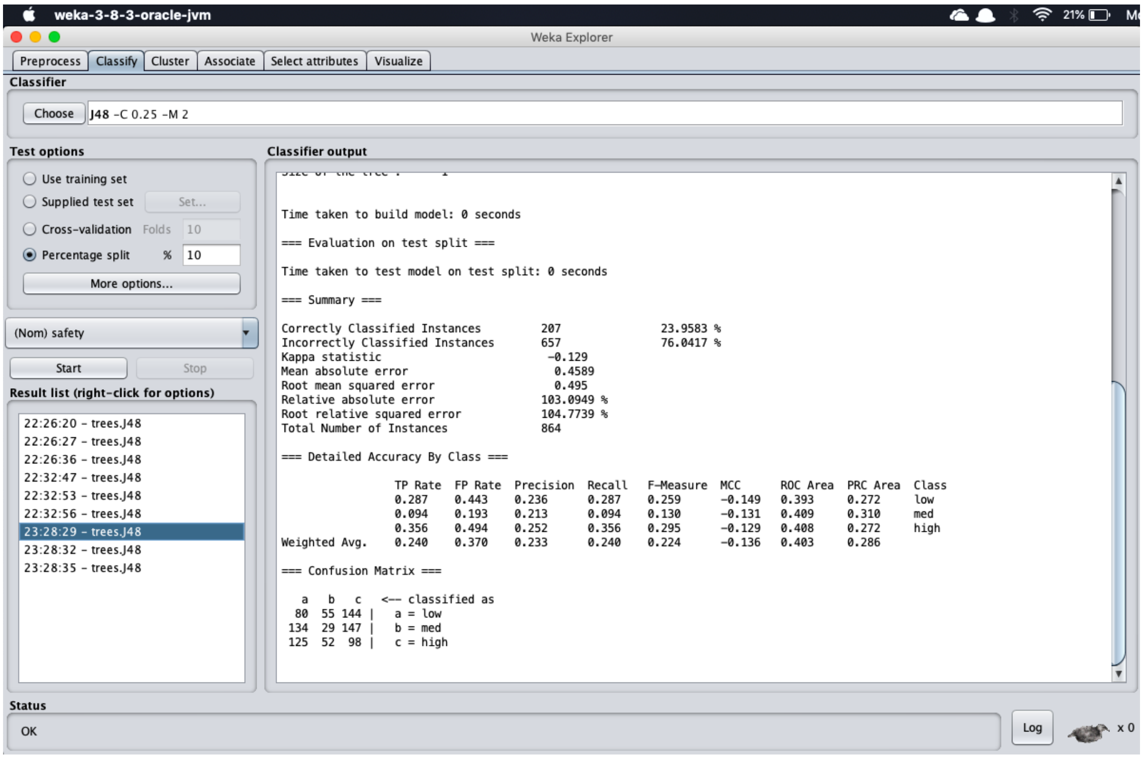
With change in training dataset, there will always be change in precision and accuracy altogether. We see the decline of precision and accuracy when the training dataset was reduced.



Dataset	Precision		
	50 % Training	25 % Training	10 % Training
Car	0.998	0.993	0.968
Health	0.965	-	-

Figure 7: Comparison

1.6 For the first dataset, use the safety attribute as a class, how accurate is the model with the same settings (50-50%, 25-75%. 10-90%)



J48 -C 0.25 -M 2

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

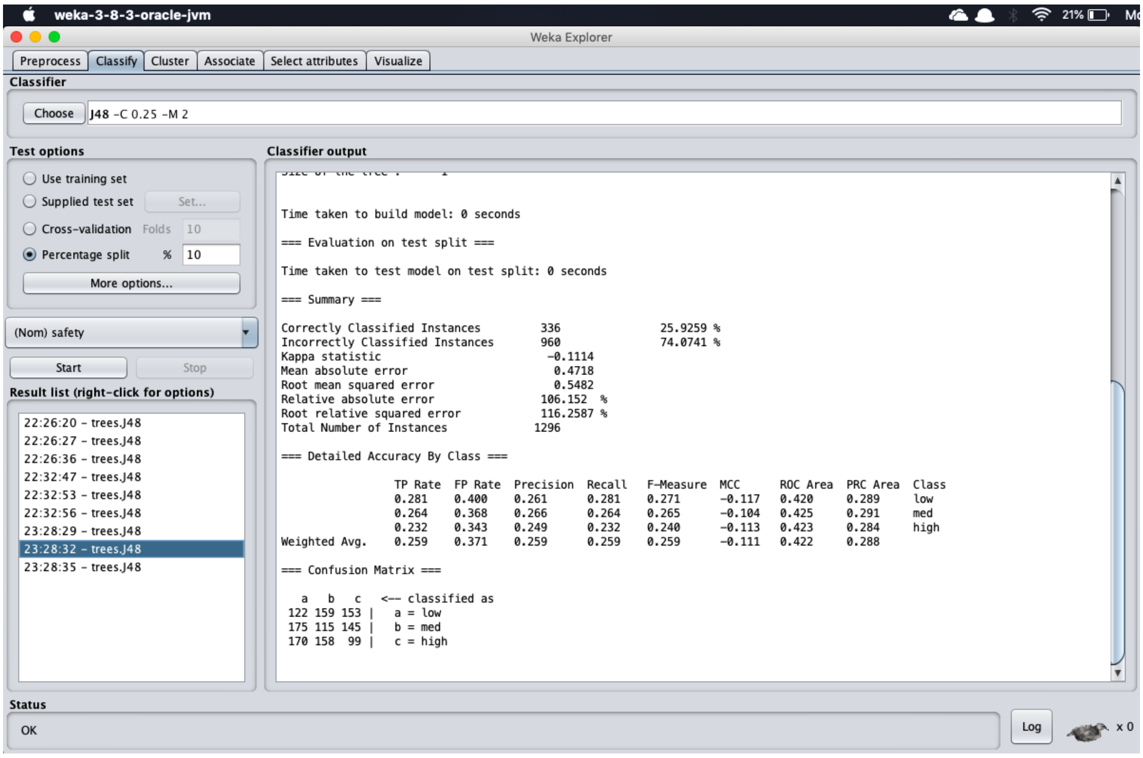
(Nom) safety

22:26:20 - trees.J48
22:26:27 - trees.J48
22:26:36 - trees.J48
22:32:47 - trees.J48
22:32:53 - trees.J48
22:32:56 - trees.J48
23:28:29 - trees.J48
23:28:32 - trees.J48
23:28:35 - trees.J48

```
Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         207               23.9583 %
Incorrectly Classified Instances       657               76.0417 %
Kappa statistic                         -0.129
Mean absolute error                      0.4589
Root mean squared error                  0.495
Relative absolute error                103.0949 %
Root relative squared error            104.7739 %
Total Number of Instances              864

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.287    0.443    0.236      0.287    0.259      -0.149   0.393     0.272     low
                 0.094    0.193    0.213      0.094    0.130      -0.131   0.409     0.310     med
                 0.356    0.494    0.252      0.356    0.295      -0.129   0.408     0.272     high
Weighted Avg.    0.240    0.370    0.233      0.240    0.224      -0.136   0.403     0.286

=== Confusion Matrix ===

   a   b   c   <-- classified as
  80  55 144 |   a = low
 134  29 147 |   b = med
 125  52  98 |   c = high
```

OK

Figure 8: 50% Training Dataset



Classifier

J48 -C 0.25 -M 2

Test options

● Percentage split % 10

(Nom) safety

Result list (right-click for options)

```
22:26:20 - trees.J48
22:26:27 - trees.J48
22:26:36 - trees.J48
22:32:47 - trees.J48
22:32:53 - trees.J48
22:32:56 - trees.J48
23:28:29 - trees.J48
23:28:32 - trees.J48
23:28:35 - trees.J48
```

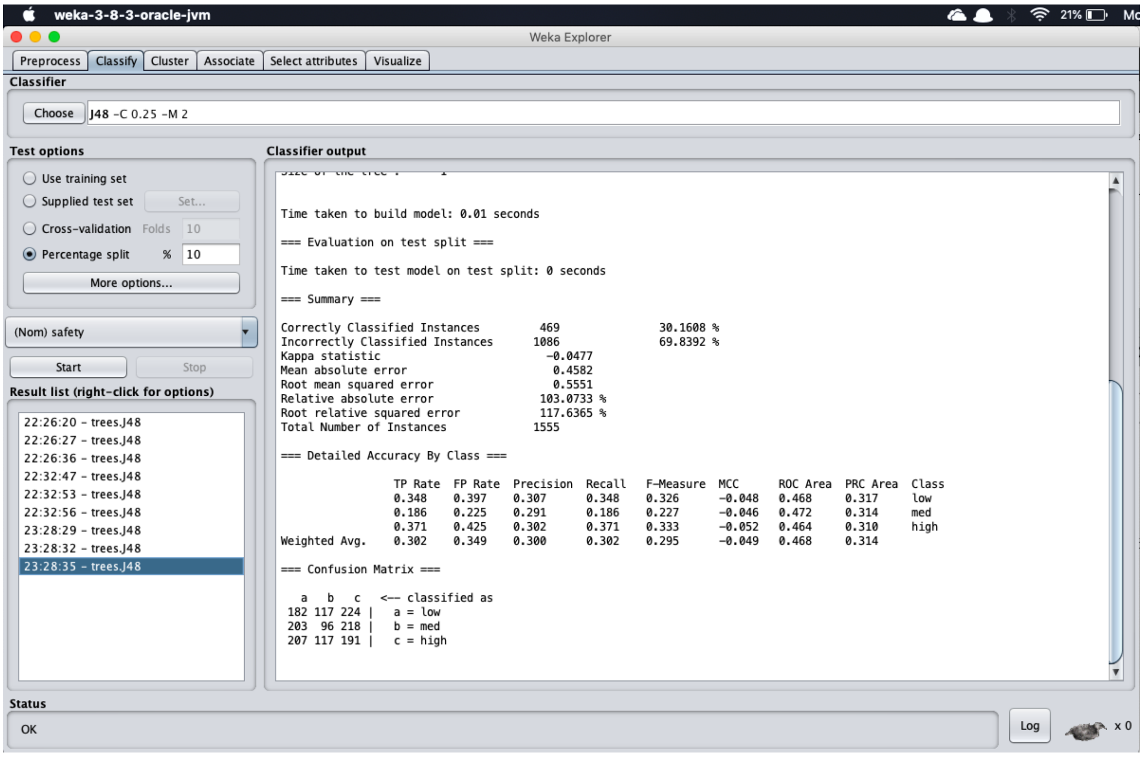
Classifier output

```
Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         336               25.9259 %
Incorrectly Classified Instances       960               74.0741 %
Kappa statistic                         -0.1114
Mean absolute error                      0.4718
Root mean squared error                  0.5482
Relative absolute error                106.152  %
Root relative squared error            116.2587 %
Total Number of Instances             1296

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.281    0.400    0.261      0.281    0.271      -0.117   0.420     0.289     low
                 0.264    0.368    0.266      0.264    0.265      -0.104   0.425     0.291     med
                 0.232    0.343    0.249      0.232    0.240      -0.113   0.423     0.284     high
Weighted Avg.    0.259    0.371    0.259      0.259    0.259      -0.111   0.422     0.288

=== Confusion Matrix ===

   a   b   c   <-- classified as
 122 159 153 |   a = low
 175 115 145 |   b = med
 170 158  99 |   c = high
```

Status

OK

Figure 9: 25% Training Dataset



Classifier

J48 -C 0.25 -M 2

Test options

● Percentage split % 10

(Nom) safety

Result list (right-click for options)

```
22:26:20 - trees.J48
22:26:27 - trees.J48
22:26:36 - trees.J48
22:32:47 - trees.J48
22:32:53 - trees.J48
22:32:56 - trees.J48
23:28:29 - trees.J48
23:28:32 - trees.J48
23:28:35 - trees.J48
```

Classifier output

```
Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         469               30.1608 %
Incorrectly Classified Instances      1086               69.8392 %
Kappa statistic                         -0.0477
Mean absolute error                      0.4582
Root mean squared error                  0.5551
Relative absolute error                103.0733 %
Root relative squared error            117.6365 %
Total Number of Instances             1555

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.348    0.397    0.307      0.348    0.326      -0.048   0.468     0.317     low
                 0.186    0.225    0.291      0.186    0.227      -0.046   0.472     0.314     med
                 0.371    0.425    0.302      0.371    0.333      -0.052   0.464     0.310     high
Weighted Avg.    0.302    0.349    0.300      0.302    0.295      -0.049   0.468     0.314

=== Confusion Matrix ===

   a   b   c   <-- classified as
 182 117 224 |   a = low
 203  96 218 |   b = med
 207 117 191 |   c = high
```

Status

OK

Figure 10: 10% Training Dataset