14210231 – Felipe Guth

Assignment 1 – COMP30120

The table 1 resumes all the tests performed in the Weka framework, where the correctly classified instances are depicted in percentage.

The better results were obtained using the method of training set as a test set, in most classifiers. This is expected since the classifier is overfitted in the data. In other words, it is optimized to classify the data it was trained on, but will have a bad performance in classifying new data. It will not be able to generalize in a proper way for new samples.

Excluding the overfitted classifiers, the best result was obtained with the 60%/40% training/testing method with the Naïve bayes classifier. However, the 10-fold cross validation method generated more consistent classifiers in general, compared to the 60-40 method.

	Training set as test set	60% training set 40% test set	10-fold cross validation
Naïve Bayes	67.58	**68.96**	63.44
KNN-1	**100**	55.17	60.68
KNN-3	77.24	53.44	65.51
SVM	73.79	63.79	**66.89**

Table 1 - Percentage of correctness of each classifier per method

Table 2 shows the round mean square error that is the distance, on average, of a data point from the fitted line, measured along a vertical line. The round mean square error once more reinforces the results depicted in Table 1.

Root mean squared error

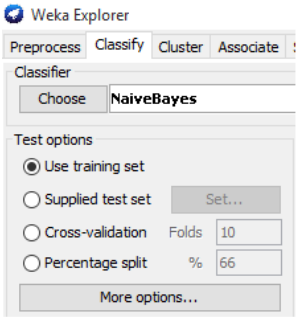
	Training set as test set	60% training set 40% test set	10-fold cross validation
Naïve Bayes	0.50	0.53	0.55
KNN-1	**0**	0.66	0.62
KNN-3	0.38	0.54	0.49
SVM	0.51	0.60	**0.57**

Table 2 – Root mean square error

Detailed Results

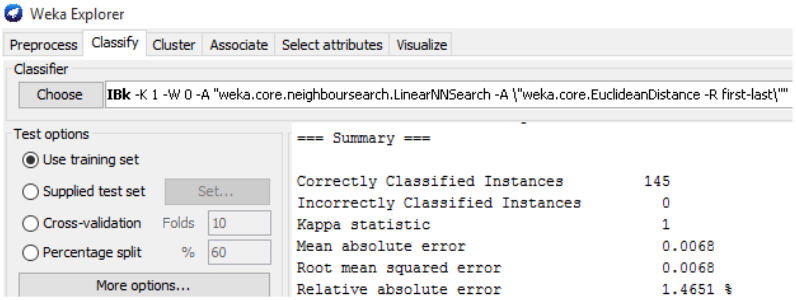
1 – Training set as test set

a – Naïve Bayes



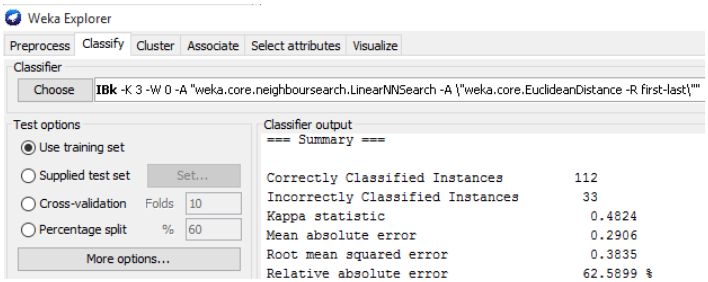

```
=== Summary ===

Correctly Classified Instances          98               67.5862 %
Incorrectly Classified Instances        47               32.4138 %
Kappa statistic                          0.3149
Mean absolute error                      0.3262
Root mean squared error                  0.5026
Relative absolute error                 70.2588 %
Root relative squared error            104.3608 %
Total Number of Instances              145

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.717     0.396      0.759     0.717      0.737      0.761    helpful
                 0.604     0.283      0.552     0.604      0.577      0.761    unhelpful
Weighted Avg.    0.676     0.355      0.683     0.676      0.679      0.761

=== Confusion Matrix ===

  a  b   <-- classified as
 66 26 |  a = helpful
 21 32 |  b = unhelpful
```

b – KNN-1



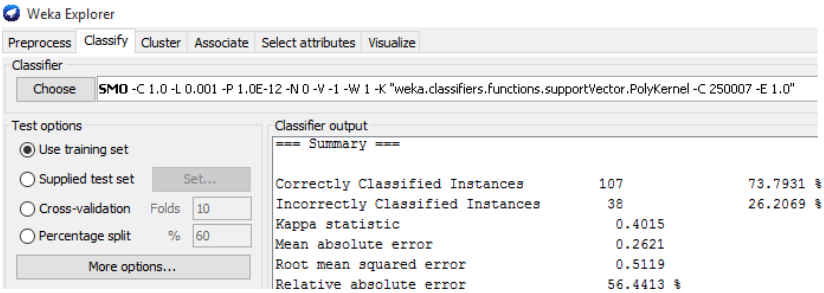

```
=== Summary ===

Correctly Classified Instances         145              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0068
Root mean squared error                  0.0068
Relative absolute error                  1.4651 %
Root relative squared error              1.4126 %
Total Number of Instances              145

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        helpful
                 1         0          1         1         1          1        unhelpful
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

  a  b   <-- classified as
 92  0 |  a = helpful
  0 53 |  b = unhelpful
```

c- KNN-3




```
=== Summary ===

Correctly Classified Instances         112               77.2414 %
Incorrectly Classified Instances        33               22.7586 %
Kappa statistic                          0.4824
Mean absolute error                      0.2906
Root mean squared error                  0.3835
Relative absolute error                 62.5899 %
Root relative squared error             79.6436 %
Total Number of Instances              145

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.891     0.434      0.781     0.891     0.832      0.838    helpful
                 0.566     0.109      0.75      0.566     0.645      0.838    unhelpful
Weighted Avg.    0.772     0.315      0.77      0.772     0.764      0.838

=== Confusion Matrix ===

  a  b   <-- classified as
 82 10 |  a = helpful
 23 30 |  b = unhelpful
```

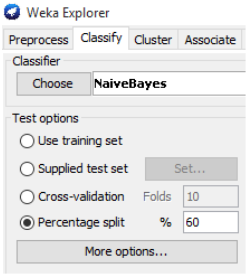
d) SVM




```
=== Summary ===

Correctly Classified Instances         107               73.7931 %
Incorrectly Classified Instances        38               26.2069 %
Kappa statistic                          0.4015
Mean absolute error                      0.2621
Root mean squared error                  0.5119
Relative absolute error                 56.4413 %
Root relative squared error            106.3018 %
Total Number of Instances              145

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.87      0.491      0.755     0.87      0.808      0.689    helpful
                 0.509     0.13       0.692     0.509     0.587      0.689    unhelpful
Weighted Avg.    0.738     0.359      0.732     0.738     0.727      0.689

=== Confusion Matrix ===

  a  b   <-- classified as
 80 12 |  a = helpful
 26 27 |  b = unhelpful
```

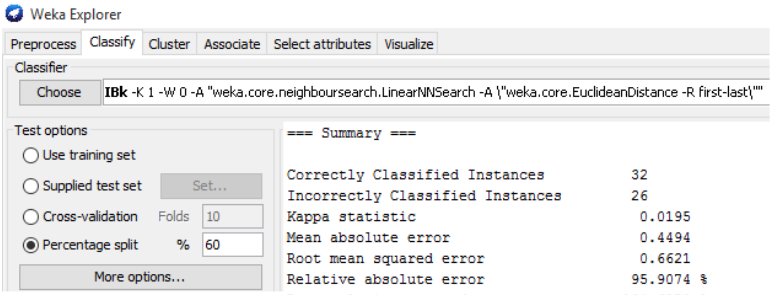
2 – 60% as training set 40% as test set

a- Naïve Bayes



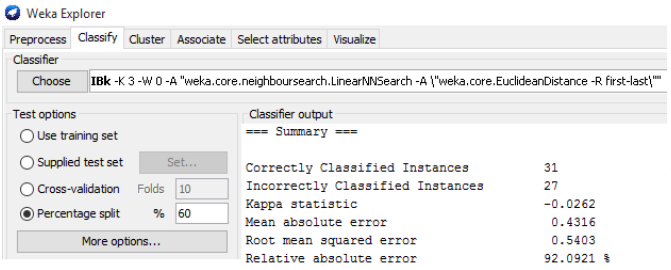

```
=== Summary ===

Correctly Classified Instances          40               68.9655 %
Incorrectly Classified Instances        18               31.0345 %
Kappa statistic                          0.3417
Mean absolute error                      0.3464
Root mean squared error                  0.5344
Relative absolute error                 73.9168 %
Root relative squared error            108.7217 %
Total Number of Instances               58

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.771     0.435      0.73      0.771     0.75       0.699    helpful
                 0.565     0.229      0.619     0.565     0.591      0.699    unhelpful
Weighted Avg.    0.69      0.353      0.686     0.69      0.687      0.699

=== Confusion Matrix ===

  a  b   <-- classified as
 27  8 |  a = helpful
 10 13 |  b = unhelpful
```

b- KNN-1



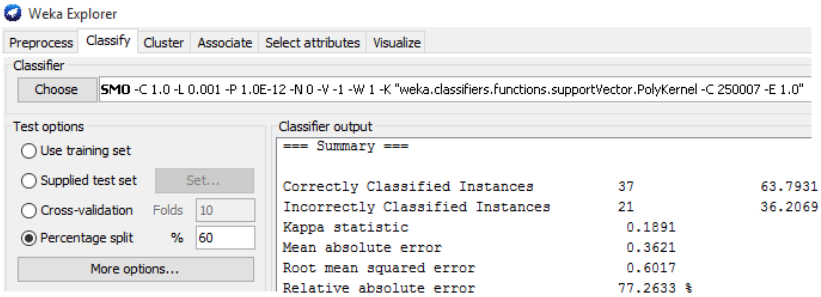

```
=== Summary ===

Correctly Classified Instances          32               55.1724 %
Incorrectly Classified Instances        26               44.8276 %
Kappa statistic                          0.0195
Mean absolute error                      0.4494
Root mean squared error                  0.6621
Relative absolute error                 95.9074 %
Root relative squared error            134.6879 %
Total Number of Instances               58

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.714     0.696      0.61      0.714     0.658      0.509    helpful
                 0.304     0.286      0.412     0.304     0.35       0.509    unhelpful
Weighted Avg.    0.552     0.533      0.531     0.552     0.536      0.509

=== Confusion Matrix ===

  a  b   <-- classified as
 25 10 |  a = helpful
 16  7 |  b = unhelpful
```

c- KNN-3




```
=== Summary ===

Correctly Classified Instances          31               53.4483 %
Incorrectly Classified Instances        27               46.5517 %
Kappa statistic                         -0.0262
Mean absolute error                      0.4316
Root mean squared error                  0.5403
Relative absolute error                 92.0921 %
Root relative squared error            109.9178 %
Total Number of Instances               58

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.714     0.739      0.595     0.714     0.649      0.563    helpful
                 0.261     0.286      0.375     0.261     0.308      0.563    unhelpful
Weighted Avg.    0.534     0.559      0.508     0.534     0.514      0.563

=== Confusion Matrix ===

  a  b   <-- classified as
 25 10 |  a = helpful
 17  6 |  b = unhelpful
```

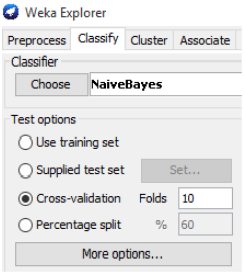
d- SVM




```
=== Summary ===

Correctly Classified Instances          37               63.7931 %
Incorrectly Classified Instances        21               36.2069 %
Kappa statistic                          0.1891
Mean absolute error                      0.3621
Root mean squared error                  0.6017
Relative absolute error                 77.2633 %
Root relative squared error            122.4121 %
Total Number of Instances               58

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.829     0.652      0.659     0.829     0.734      0.588    helpful
                   0.348     0.171      0.571     0.348     0.432      0.588    unhelpful
Weighted Avg.    0.638     0.462      0.624     0.638     0.615      0.588

=== Confusion Matrix ===

  a  b   <-- classified as
 29  6 |  a = helpful
 15  8 |  b = unhelpful
```

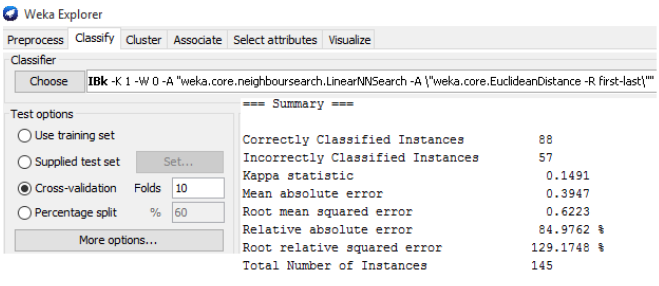
3 – 10-Fold Cross validation

a- Naïve Bayes



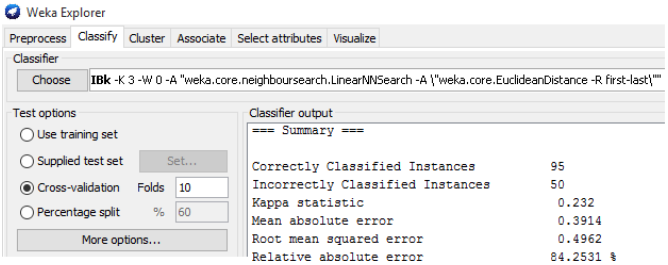

```
=== Summary ===

Correctly Classified Instances          92               63.4483 %
Incorrectly Classified Instances        53               36.5517 %
Kappa statistic                          0.2274
Mean absolute error                      0.375
Root mean squared error                  0.5534
Relative absolute error                 80.7326 %
Root relative squared error            114.8726 %
Total Number of Instances              145

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.685     0.453      0.724     0.685     0.704      0.676    helpful
                   0.547     0.315      0.5       0.547     0.523      0.676    unhelpful
Weighted Avg.    0.634     0.403      0.642     0.634     0.638      0.676

=== Confusion Matrix ===

  a  b   <-- classified as
 63 29 |  a = helpful
 24 29 |  b = unhelpful
```

b- KNN-1



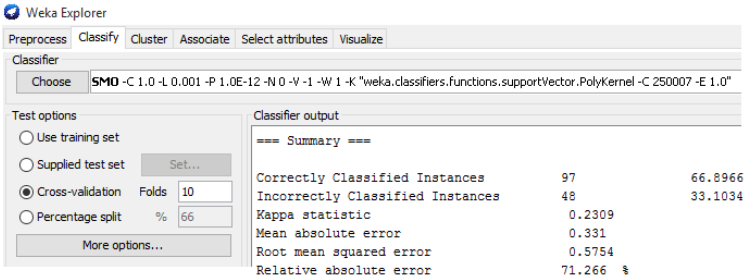

```
=== Summary ===

Correctly Classified Instances          88               60.6897 %
Incorrectly Classified Instances        57               39.3103 %
Kappa statistic                          0.1491
Mean absolute error                      0.3947
Root mean squared error                  0.6223
Relative absolute error                 84.9762 %
Root relative squared error            129.1748 %
Total Number of Instances              145

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.696     0.547      0.688     0.696     0.692      0.557    helpful
                   0.453     0.304      0.462     0.453     0.457      0.557    unhelpful
Weighted Avg.    0.607     0.458      0.605     0.607     0.606      0.557

=== Confusion Matrix ===

  a  b   <-- classified as
 64 28 |  a = helpful
 29 24 |  b = unhelpful
```

c- KNN-3




```
=== Summary ===

Correctly Classified Instances          95               65.5172 %
Incorrectly Classified Instances        50               34.4828 %
Kappa statistic                          0.232
Mean absolute error                      0.3914
Root mean squared error                  0.4962
Relative absolute error                 84.2531 %
Root relative squared error            103.0055 %
Total Number of Instances              145

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.772     0.547      0.71      0.772     0.74       0.607    helpful
                 0.453     0.228      0.533     0.453     0.49       0.607    unhelpful
Weighted Avg.    0.655     0.431      0.645     0.655     0.648      0.607

=== Confusion Matrix ===

  a  b   <-- classified as
 71 21 |  a = helpful
 29 24 |  b = unhelpful
```

d- SVM




```
=== Summary ===

Correctly Classified Instances          97               66.8966 %
Incorrectly Classified Instances        48               33.1034 %
Kappa statistic                          0.2309
Mean absolute error                      0.331
Root mean squared error                  0.5754
Relative absolute error                 71.266  %
Root relative squared error            119.4348 %
Total Number of Instances              145

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.837     0.623      0.7       0.837     0.762      0.607    helpful
                 0.377     0.163      0.571     0.377     0.455      0.607    unhelpful
Weighted Avg.    0.669     0.455      0.653     0.669     0.65       0.607

=== Confusion Matrix ===

  a  b   <-- classified as
 77 15 |  a = helpful
 33 20 |  b = unhelpful
```