CSCI4030U: Big Data Analytics
Lab08

Syed Naqvi
100590852

April 8, 2024

Ecoli Dataset

(a) **C4.5 (weka.classifier.trees.J48)**
Misclassification Rate: 15.7738%
Runtime: 0s

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         283               84.2262 %
Incorrectly Classified Instances        53               15.7738 %
Kappa statistic                          0.7824
Mean absolute error                      0.0486
Root mean squared error                  0.1851
Relative absolute error                 26.5877 %
Root relative squared error             61.3413 %
Total Number of Instances              336
```

(b) **RIPPER (weka.classifier.rules.JRip)**
Misclassification Rate: 19.3452%
Runtime: 0.03s

```
Time taken to build model: 0.03 seconds

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

Correctly Classified Instances         271               80.6548 %
Incorrectly Classified Instances        65               19.3452 %
Kappa statistic                          0.7311
Mean absolute error                      0.0608
Root mean squared error                  0.2013
Relative absolute error                 33.2586 %
Root relative squared error             66.7354 %
Total Number of Instances              336
```

(c) **Naive Bayesian Classification (weka.classifiers.bayes.NaiveBayes)**
Misclassification Rate: 14.5833%
Runtime: 0s

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         287               85.4167 %
Incorrectly Classified Instances        49               14.5833 %
Kappa statistic                          0.8002
Mean absolute error                      0.0429
Root mean squared error                  0.1639
Relative absolute error                 23.461  %
Root relative squared error             54.3314 %
Total Number of Instances              336
```

(d) **k-Nearest Neighbor (weka.classifiers.lazy.IBk)**
Misclassification Rate: 19.6429%
Runtime: 0s

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         270               80.3571 %
Incorrectly Classified Instances        66               19.6429 %
Kappa statistic                          0.7295
Mean absolute error                      0.0535
Root mean squared error                  0.2189
Relative absolute error                 29.238  %
Root relative squared error             72.5574 %
Total Number of Instances              336
```

(e) **Neural Networks (weka.classifiers.functions.MultilayerPerceptron)**
Misclassification Rate: 13.9881%
Runtime: 0.3s

```
Time taken to build model: 0.3 seconds

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

Correctly Classified Instances         289               86.0119 %
Incorrectly Classified Instances        47               13.9881 %
Kappa statistic                          0.8066
Mean absolute error                      0.0484
Root mean squared error                  0.1704
Relative absolute error                 26.479  %
Root relative squared error             56.4913 %
Total Number of Instances              336
```

The MultilayerPerceptron algorithm has the lowest misclassification rate but also had the highest runtime of 0.3 seconds. It is likely context dependent if the incresed runtime costs are worth the improved accuracy. The k-nearest neighbor algorithm had the highest misclassification rate with a runtime of 0s while the RIPPER algoirithm had the second highest misclassification rate with a runtime of 0.03s.

Glass Dataset

(a) **C4.5 (weka.classifier.trees.J48)**
Misclassification Rate: 34.1121%
Runtime: 0s

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         141               65.8879 %
Incorrectly Classified Instances        73               34.1121 %
Kappa statistic                          0.5412
Mean absolute error                      0.1059
Root mean squared error                  0.2928
Relative absolute error                 50.0098 %
Root relative squared error             90.2088 %
Total Number of Instances              214
```

(b) **RIPPER (weka.classifier.rules.JRip)**
Misclassification Rate: 30.3738%
Runtime: 0.01s

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         149               69.6262 %
Incorrectly Classified Instances        65               30.3738 %
Kappa statistic                          0.5741
Mean absolute error                      0.1139
Root mean squared error                  0.2657
Relative absolute error                 53.8052 %
Root relative squared error             81.8743 %
Total Number of Instances              214
```

(c) **Naive Bayesian Classification (weka.classifiers.bayes.NaiveBayes)**
Misclassification Rate: 50.4673%
Runtime: 0s

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         106               49.5327 %
Incorrectly Classified Instances       108               50.4673 %
Kappa statistic                          0.334
Mean absolute error                      0.1521
Root mean squared error                  0.3343
Relative absolute error                 71.8506 %
Root relative squared error            102.9939 %
Total Number of Instances              214
```

(d) **k-Nearest Neighbor (weka.classifiers.lazy.IBk)**
Misclassification Rate: 29.4393%
Runtime: 0s

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         151               70.5607 %
Incorrectly Classified Instances        63               29.4393 %
Kappa statistic                          0.6017
Mean absolute error                      0.0897
Root mean squared error                  0.2852
Relative absolute error                 42.3765 %
Root relative squared error             87.8768 %
Total Number of Instances              214
```

(e) **Neural Networks (weka.classifiers.functions.MultilayerPerceptron)**
Misclassification Rate: 30.8411%
Runtime: 0.3s

```
Time taken to build model: 0.24 seconds

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

Correctly Classified Instances         148               69.1589 %
Incorrectly Classified Instances        66               30.8411 %
Kappa statistic                          0.5677
Mean absolute error                      0.1067
Root mean squared error                  0.2471
Relative absolute error                 50.3806 %
Root relative squared error             76.124  %
Total Number of Instances              214
```

For the Glass dataset there was a clear loser with the Naive Bayes algorithm. This algorithm had a misclassification rate of 50.4673%. It seems that the classification rates for this dataset are lower in general with the k-nearest neighbors algorithm performing the best at 29.4393% misclassification rate.

Image Dataset

(a) **C4.5 (weka.classifier.trees.J48)**
Misclassification Rate: 10.9524%
Runtime: 0.01s

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         187               89.0476 %
Incorrectly Classified Instances        23               10.9524 %
Kappa statistic                          0.8722
Mean absolute error                      0.0333
Root mean squared error                  0.1652
Relative absolute error                 15.5312 %
Root relative squared error             50.4698 %
Total Number of Instances              210
```

(b) **RIPPER (weka.classifier.rules.JRip)**
Misclassification Rate: 20.4762%
Runtime: 0.01s

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         167               79.5238 %
Incorrectly Classified Instances        43               20.4762 %
Kappa statistic                          0.7611
Mean absolute error                      0.0625
Root mean squared error                  0.2068
Relative absolute error                 29.1466 %
Root relative squared error             63.1781 %
Total Number of Instances              210
```

(c) **Naive Bayesian Classification (weka.classifiers.bayes.NaiveBayes)**
Misclassification Rate: 22.381%
Runtime: 0s

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         163               77.619  %
Incorrectly Classified Instances        47               22.381  %
Kappa statistic                          0.7389
Mean absolute error                      0.0557
Root mean squared error                  0.2269
Relative absolute error                 25.9739 %
Root relative squared error             69.315  %
Total Number of Instances              210
```

(d) **k-Nearest Neighbor (weka.classifiers.lazy.IBk)**
Misclassification Rate: 12.8571%
Runtime: 0s

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         183               87.1429 %
Incorrectly Classified Instances        27               12.8571 %
Kappa statistic                          0.85
Mean absolute error                      0.0397
Root mean squared error                  0.1761
Relative absolute error                 18.5207 %
Root relative squared error             53.8043 %
Total Number of Instances              210
```

(e) **Neural Networks (weka.classifiers.functions.MultilayerPerceptron)**
Misclassification Rate: 11.4286%
Runtime: 0.45s

```
Time taken to build model: 0.45 seconds

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

Correctly Classified Instances         186               88.5714 %
Incorrectly Classified Instances        24               11.4286 %
Kappa statistic                          0.8667
Mean absolute error                      0.0377
Root mean squared error                  0.1499
Relative absolute error                 17.563  %
Root relative squared error             45.7901 %
Total Number of Instances              210
```

For the image dataset there is larger variance in the misclassification rates acorss the various algorithms. The C4.5 algorithm has the lowest misclassification rate while the Naive Bayesian algorithm has the highest.

Conclusions:
Generally, the misclassification rates do not vary too much. It seems that the Neural Network algorithm consistenlty has the longest runtime while the c4.5 algoirithm is usually among the better performers. The closest thing to a clear loser would likely be the Naive Bayesian algorithm is it had the highest misclassification rates for both the Image and Glass datasets.