Abdul Saboor (20L-1113 | BDS-6A1)

Created	@March 2, 2023
Course	Data Mining

Cancer Data

Preprocessing

- There were no null values in the dataset
- remove the column of **ID** as it is of no use
- normalize the dataset for easier visualization

Classification

Decision Tree

```
|   |   texture_mean <= 0.3324315184308842: B (9.0/1.0)
|   |   texture_mean > 0.3324315184308842: M (10.0)
|   concavity_mean > 0.1677600749765698: M (164.0)

Number of Leaves  :     15

Size of the tree :      29


Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         531               93.3216 %
Incorrectly Classified Instances        38                6.6784 %
Kappa statistic                          0.8585
Mean absolute error                      0.0709
Root mean squared error                  0.2554
Relative absolute error                 15.1651 %
Root relative squared error             52.8146 %
Total Number of Instances              569

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.934    0.067    0.892      0.934    0.912      0.859    0.930     0.874     M
                 0.933    0.066    0.960      0.933    0.946      0.859    0.930     0.924     B
Weighted Avg.    0.933    0.066    0.934      0.933    0.934      0.859    0.930     0.905

=== Confusion Matrix ===

   a   b   <-- classified as
 198  14 |   a = M
  24 333 |   b = B
```

Random Forest

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities -batch-size 1000

Time taken to build model: 0.14 seconds

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

Correctly Classified Instances         551               96.8366 %
Incorrectly Classified Instances        18                3.1634 %
Kappa statistic                          0.9319
Mean absolute error                      0.0751
Root mean squared error                  0.1708
Relative absolute error                 16.0652 %
Root relative squared error             35.3334 %
Total Number of Instances              569

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.943    0.017    0.971      0.943    0.957      0.932    0.989     0.988     M
                 0.983    0.057    0.967      0.983    0.975      0.932    0.989     0.989     B
Weighted Avg.    0.968    0.042    0.968      0.968    0.968      0.932    0.989     0.989

=== Confusion Matrix ===

   a   b   <-- classified as
 200  12 |   a = M
   6 351 |   b = B
```

Naive Bayes

```
  [total]                  215.0  360.0

fractal_dimension_worst
  '(-inf-0.09279]'         126.0  313.0
  '(0.09279-inf)'           88.0   46.0
  [total]                  214.0  359.0



Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         540               94.9033 %
Incorrectly Classified Instances        29                5.0967 %
Kappa statistic                          0.8909
Mean absolute error                      0.0539
Root mean squared error                  0.2167
Relative absolute error                 11.525  %
Root relative squared error             44.8193 %
Total Number of Instances              569

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.929    0.039    0.934      0.929    0.931      0.891    0.985     0.981     M
                 0.961    0.071    0.958      0.961    0.959      0.891    0.985     0.987     B
Weighted Avg.    0.949    0.059    0.949      0.949    0.949      0.891    0.985     0.985

=== Confusion Matrix ===

   a   b   <-- classified as
 197  15 |   a = M
  14 343 |   b = B
```

Analysis

Considering the following dataset, **Random Forest** performed best as it gave **96.8%** accuracy. The highest among the selection of algorithms.

Diabetes Data

Preprocessing

- There are no null values to be removed in dataset
- normalize the dataset for easier visualization

Classification

Decision Tree

```
|   |   |   |   Alopecia = No: Positive (5.0)
|   |   |   delayed healing = No: Positive (11.0)
|   |   Itching = No: Positive (30.0)

Number of Leaves  :     22

Size of the tree :      43


Time taken to build model: 0.04 seconds

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

Correctly Classified Instances         499               95.9615 %
Incorrectly Classified Instances        21                4.0385 %
Kappa statistic                          0.9156
Mean absolute error                      0.0549
Root mean squared error                  0.1975
Relative absolute error                 11.5905 %
Root relative squared error             40.5926 %
Total Number of Instances              520

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.950    0.025    0.984      0.950    0.967      0.916    0.966     0.975     Positive
                 0.975    0.050    0.924      0.975    0.949      0.916    0.966     0.910     Negative
Weighted Avg.    0.960    0.035    0.961      0.960    0.960      0.916    0.966     0.950

=== Confusion Matrix ===

   a   b   <-- classified as
 304  16 |   a = Positive
   5 195 |   b = Negative
```

Random Forest

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.15 seconds

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

Correctly Classified Instances         505               97.1154 %
Incorrectly Classified Instances        15                2.8846 %
Kappa statistic                          0.9392
Mean absolute error                      0.0563
Root mean squared error                  0.1395
Relative absolute error                 11.9003 %
Root relative squared error             28.6698 %
Total Number of Instances              520

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.972    0.030    0.981      0.972    0.976      0.939    0.998     0.999     Positive
                 0.970    0.028    0.956      0.970    0.963      0.939    0.998     0.997     Negative
Weighted Avg.    0.971    0.029    0.971      0.971    0.971      0.939    0.998     0.998

=== Confusion Matrix ===

   a   b   <-- classified as
 311   9 |   a = Positive
   6 194 |   b = Negative
```

Naive Bays

```
  [total]                322.0    202.0

Obesity
  Yes                     62.0     28.0
  No                     260.0    174.0
  [total]                322.0    202.0



Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         453               87.1154 %
Incorrectly Classified Instances        67               12.8846 %
Kappa statistic                          0.734
Mean absolute error                      0.149
Root mean squared error                  0.3184
Relative absolute error                 31.4632 %
Root relative squared error             65.4511 %
Total Number of Instances              520

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.856    0.105    0.929      0.856    0.891      0.738    0.945     0.969     Positive
                 0.895    0.144    0.796      0.895    0.842      0.738    0.945     0.909     Negative
Weighted Avg.    0.871    0.120    0.878      0.871    0.872      0.738    0.945     0.946

=== Confusion Matrix ===

   a   b   <-- classified as
 274  46 |   a = Positive
  21 179 |   b = Negative
```

Analysis

Considering the following dataset, **Random Forest** performed best as it gave **97.1%** accuracy. The highest among the selection of algorithms.