Data Analytics Assignment 7

Dan Sun(das225) Yue Su(yus55)

1.Classification Models for Home -vote

a.BFTree

The accuracy will be 95.1724%

Classifier output

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

Best-First Decision Tree

physician_fee_freeze=(n)|(w)
|  adoption_of_the_budget_resolution=(y)|(n)
|  |  adoption_of_the_budget_resolution=(y)|(w): democrat(224.0/0.0)
|  |  adoption_of_the_budget_resolution!=(y)|(w): democrat(23.0/2.0)
|  adoption_of_the_budget_resolution!=(y)|(n)
|  |  mx_missile=(n)|(y)
|  |  |  education_spending=(n)|(w): democrat(5.0/0.0)
|  |  |  education_spending!=(n)|(w): republican(1.0/1.0)
|  |  mx_missile!=(n)|(y): republican(2.0/0.0)
physician_fee_freeze!=(n)|(w)
|  synfuels_corporation_cutback=(y)
|  |  adoption_of_the_budget_resolution=(w)|(y)
|  |  |  nti_satellite_test_ban=(n)|(w): democrat(6.0/0.0)
|  |  |  nti_satellite_test_ban!=(n)|(w): republican(3.0/0.0)
|  |  adoption_of_the_budget_resolution!=(w)|(y)
|  |  |  el_salvador_aid=(n): democrat(2.0/0.0)
|  |  |  el_salvador_aid!=(n)
|  |  |  |  export_administration_act_sa=(w)|(n): republican(8.0/3.0)
|  |  |  |  export_administration_act_sa!=(w)|(n): republican(10.0/0.0)
|  synfuels_corporation_cutback!=(y)
|  |  duty_free_exports=(y)
|  |  |  immigration=(n): republican(2.0/2.0)
|  |  |  immigration!=(n): republican(9.0/0.0)
|  |  duty_free_exports!=(y)
|  |  |  adoption_of_the_budget_resolution=(y)
|  |  |  |  export_administration_act_sa=(w): republican(1.0/1.0)
|  |  |  |  export_administration_act_sa!=(w): republican(12.0/0.0)
|  |  |  adoption_of_the_budget_resolution!=(y): republican(118.0/0.0)

Size of the Tree: 29

Number of Leaf Nodes: 15

Time taken to build model: 0.12 seconds
```

When we change the seed value from 1 to 3 and minNumObj from 2 to 10 , the accuracy will increased to 95.6322%

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

Best-First Decision Tree

physician_fee_freeze=(n)|(w): democrat(253.0/5.0)
physician_fee_freeze!=(n)|(w)
|  synfuels_corporation_cutback=(y): republican(21.0/11.0)
|  synfuels_corporation_cutback!=(y)
|  |  duty_free_exports=(y): republican(11.0/2.0)
|  |  duty_free_exports!=(y)
|  |  |  adoption_of_the_budget_resolution=(y): republican(13.0/1.0)
|  |  |  adoption_of_the_budget_resolution!=(y): republican(118.0/0.0)

Size of the Tree: 9

Number of Leaf Nodes: 5

Time taken to build model: 0.1 seconds

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

Correctly Classified Instances         416               95.6322 %
Incorrectly Classified Instances        19                4.3678 %
Kappa statistic                          0.9088
```

b.Decision Stump

The accuracy will be 95.6322%. This method doesn't provide parameters.We cannot improve the accuracy by changing parameter.

```
Decision Stump

Classifications

physician_fee_freeze = y : republican
physician_fee_freeze != y : democrat
physician_fee_freeze is missing : democrat

Class distributions

physician_fee_freeze = y
republican      democrat
0.9209039548022598      0.07909604519774012
physician_fee_freeze != y
republican      democrat
0.01937984496124031     0.9806201550387597
physician_fee_freeze is missing
republican      democrat
0.38620689655172413     0.6137931034482759


Time taken to build model: 0 seconds

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

Correctly Classified Instances         416               95.6322 %
Incorrectly Classified Instances        19                4.3678 %
```

c.FT

The accuracy will be 96.7816%. By changing the parameters, we cannot improve the accuracy.

```
Time taken to build model: 0.21 seconds

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

Correctly Classified Instances         421               96.7816 %
Incorrectly Classified Instances        14                3.2184 %
Kappa statistic                          0.9323
Mean absolute error                      0.0395
Root mean squared error                  0.1731
Relative absolute error                  8.3238 %
Root relative squared error             35.5574 %
Total Number of Instances              435
```

d.Id3

The accuracy will be94.2529%. No way to change the parameters.

```
Time taken to build model: 0.02 seconds

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

Correctly Classified Instances         410               94.2529 %
Incorrectly Classified Instances        22                5.0575 %
Kappa statistic                          0.8933
Mean absolute error                      0.0509
Root mean squared error                  0.2257
Relative absolute error                 10.7989 %
Root relative squared error             46.4492 %
UnClassified Instances                   3                0.6897 %
Total Number of Instances              435
```

e.J48

The accuracy will be 94.9425%.

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         413               94.9425 %
Incorrectly Classified Instances        22                5.0575 %
Kappa statistic                          0.894
Mean absolute error                      0.068
Root mean squared error                  0.2051
Relative absolute error                 14.3367 %
Root relative squared error             42.1278 %
Total Number of Instances              435
```

When we change the minNumObj to 5 as well as numFolds to 5, the accuracy will increase to 95.4023%

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         415               95.4023 %
Incorrectly Classified Instances        20                4.5977 %
Kappa statistic                          0.9041
Mean absolute error                      0.0728
Root mean squared error                  0.2013
Relative absolute error                 15.3501 %
Root relative squared error             41.3398 %
Total Number of Instances              435
```

f.LMT

The accuracy will be 96.7816%. Cannot improve anymore.

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

Correctly Classified Instances         421               96.7816 %
Incorrectly Classified Instances        14                3.2184 %
Kappa statistic                          0.9324
Mean absolute error                      0.0556
Root mean squared error                  0.1698
Relative absolute error                 11.7247 %
Root relative squared error             34.8824 %
Total Number of Instances              435
```

2.Classification Models for Iris

a.BFTree

The accuracy will be 94.6667%.

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

Correctly Classified Instances         142               94.6667 %
Incorrectly Classified Instances         8                5.3333 %
Kappa statistic                          0.92
Mean absolute error                      0.041
Root mean squared error                  0.1754
Relative absolute error                  9.2315 %
Root relative squared error             37.2061 %
Total Number of Instances              150
```

By changing the parameters,we can improve the accuracy to 96%

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

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.044
Root mean squared error                  0.1734
Relative absolute error                  9.9004 %
Root relative squared error             36.7788 %
Total Number of Instances              150
```

b.Decision Stump

The accuracy will be 66.6667%. No way to improve.

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

Correctly Classified Instances         100               66.6667 %
Incorrectly Classified Instances        50               33.3333 %
Kappa statistic                          0.5
Mean absolute error                      0.2222
Root mean squared error                  0.3333
Relative absolute error                 50      %
Root relative squared error             70.7107 %
Total Number of Instances              150
```

c.FT

The accuracy will be96.6667%.It's the highest accuracy.

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

Correctly Classified Instances         145               96.6667 %
Incorrectly Classified Instances         5                3.3333 %
Kappa statistic                          0.95
Mean absolute error                      0.0316
Root mean squared error                  0.1343
Relative absolute error                  7.1172 %
Root relative squared error             28.4908 %
Total Number of Instances              150
```

d.Id3

We find out that we cannot use Id3 model for this dataset.

e.J48

The accuracy will be 96%.Its the highest accuracy.

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Total Number of Instances              150
```

f.LMT

The accuracy will be 94%.Cannot improve.

```
Time taken to build model: 0.14 seconds

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

Correctly Classified Instances         141               94      %
Incorrectly Classified Instances         9                6      %
Kappa statistic                          0.91
Mean absolute error                      0.0439
Root mean squared error                  0.1542
Relative absolute error                  9.8675 %
Root relative squared error             32.7159 %
Total Number of Instances              150
```

Conclusions:

1. For House -vote , the highest accuracy is 96.7816%, and we use FT and LMT tree model to come out with this result.
2. For Iris, the highest accuracy is 96.6667%, and we use Ft tree model to come out with this result.
3. We can draw the conclusion that FT is the best method for dataset classification in this project.It gets highest accuracy in both datasets.