American International University-Bangladesh

Data Warehouse and Data Mining
Section: B

Supervised Learning: Flags Dataset.

Submitted By,
Name: Zaman, Wasif
ID: 16-32027-2

Submitted To,
Rahman Mohammod Hafizur
Assistant Professor,
Department of Computer Science
American International University-Bangladesh

Datasets: The dataset is used for supervised learning. Dataset is about Flags.

Supervised Learning: Flags Dataset

Data Set Name: Flags Data Set

Data Set Information:

This data file contains details of various nations and their flags. In this file the fields are separated by spaces (not commas). With this data you can try things like predicting the religion of a country from its size and the colors in its flag.
10 attributes are numeric-valued. The remainder are either Boolean- or nominal-valued
Attribute Information:

1. **Name**: Name of the country concerned
2. **Landmass**: 1=N. America, 2=S. America, 3=Europe, 4=Africa, 4=Asia, 6=Oceania
3. **Zone**: Geographic quadrant, based on Greenwich and the Equator; 1=NE, 2=SE, 3=SW, 4=NW
4. **Area**: in thousands of square km
5. **Population**: in round millions
6. **Language**: 1=English, 2=Spanish, 3=French, 4=German, 5=Slavic, 6=Other Indo-European, 7=Chinese, 8=Arabic, 9=Japanese/Turkish/Finnish/Magyar, 10=Others
7. **Religion**: 0=Catholic, 1=Other Christian, 2=Muslim, 3=Buddhist, 4=Hindu, 5=Ethnic, 6=Marxist, 7=Others
8. **Bars**: Number of vertical bars in the flag
9. **Stripes**: Number of horizontal stripes in the flag
10. **Colors**: Number of different colours in the flag
11. **Red**: 0 if red absent, 1 if red present in the flag
12. **Green**: same for green
13. **Blue**: same for blue
14. **Gold**: same for gold (also yellow)
15. White: same for white
16. **Black**: same for black
17. **Orange**: same for orange (also brown)
18. **Mainhue**: predominant color in the flag (tie-breaks decided by taking the top most hue, if that fails then the most central hue, and if that fails the leftmost hue)
19. **Circles**: Number of circles in the flag
20. **Crosses**: Number of (upright) crosses
21. **Saltires**: Number of diagonal crosses
22. **Quarters**: Number of quartered sections
23. **Sunstars**: Number of sun or star symbols
24. **Crescent**: 1 if a crescent moon symbol present, else 0
25. **Triangle**: 1 if any triangles present, 0 otherwise
26. **Icon**: 1 if an inanimate image present (e.g., a boat), otherwise 0
27. **Animate**: 1 if an animate image (e.g., an eagle, a tree, a human hand) present, 0 otherwise
28. **Text**: 1 if any letters or writing on the flag (e.g., a motto or slogan), 0 otherwise
29. **Topleft**: Colour in the top-left corner (moving right to decide tie-breaks)
30. **Botright**: Colour in the bottom-left corner (moving left to decide tie-breaks)
31. Religion

Solution:

For the solution, five classifiers have been used. These are:

Naive Bayes
J48
IBK
KSTAR
JRIP

1. Naive Bayes:

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         168               86.5979 %
Incorrectly Classified Instances        26               13.4021 %
Kappa statistic                          0.8319
Mean absolute error                      0.0339
Root mean squared error                  0.1701
Relative absolute error                 16.8777 %
Root relative squared error             53.791  %
Total Number of Instances              194

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.972    0.032    0.875      0.972    0.921      0.904    0.970     0.840     Muslim
                 0.667    0.028    0.667      0.667    0.667      0.639    0.960     0.551     Marxist
                 0.900    0.037    0.915      0.900    0.908      0.867    0.968     0.874     Cristian
                 0.875    0.013    0.946      0.875    0.909      0.888    0.961     0.935     Catholic
                 0.963    0.036    0.813      0.963    0.881      0.865    0.972     0.777     Ethnic
                 0.750    0.011    0.750      0.750    0.750      0.739    0.907     0.625     Buddhist
                 0.000    0.000    ?          0.000    ?          ?        0.705     0.080     Hindu
                 0.500    0.005    0.667      0.500    0.571      0.570    0.954     0.592     Others
Weighted Avg.    0.866    0.028    ?          0.866    ?          ?        0.959     0.809
```

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h   <-- classified as
 35  0  0  0  0  1  0  0 |  a = Muslim
  0 10  0  0  4  0  0  1 |  b = Marxist
  3  1 54  2  0  0  0  0 |  c = Cristian
  0  0  5 35  0  0  0  0 |  d = Catholic
  0  1  0  0 26  0  0  0 |  e = Ethnic
  2  0  0  0  0  6  0  0 |  f = Buddhist
  0  1  0  0  2  1  0  0 |  g = Hindu
  0  2  0  0  0  0  0  2 |  h = Others
```

By considering a,b,c,d as positive and e,f,g,h negative 2*2 confusion matrix

	+	-
+	141	6
-	6	37

Correctly Classified Instances	168	86.5979 %
Incorrectly Classified Instances	26	13.4021 %

TruePositiveRate(TPR)=0.96

False Positive Rate = 0.14

2. J48:

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

Correctly Classified Instances         194               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances              194

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Muslim
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Marxist
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Cristian
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Catholic
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Ethnic
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Buddhist
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Hindu
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Others
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===
```

```
=== Confusion Matrix ===

  a   b   c   d   e   f   g   h   <-- classified as
 36   0   0   0   0   0   0   0 |  a = Muslim
  0  15   0   0   0   0   0   0 |  b = Marxist
  0   0  60   0   0   0   0   0 |  c = Cristian
  0   0   0  40   0   0   0   0 |  d = Catholic
  0   0   0   0  27   0   0   0 |  e = Ethnic
  0   0   0   0   0   8   0   0 |  f = Buddhist
  0   0   0   0   0   0   4   0 |  g = Hindu
  0   0   0   0   0   0   0   4 |  h = Others
```

By considering a,b,c,d as positive and e,f,g,h negative 2*2 confusion matrix=

	+	-
+	151	0
-	0	43

Correctly Classified Instances 194 100%
Incorrectly Classified Instances 0 0%

TruePositiveRate(TPR)=1.000

False Positive Rate = 0.000

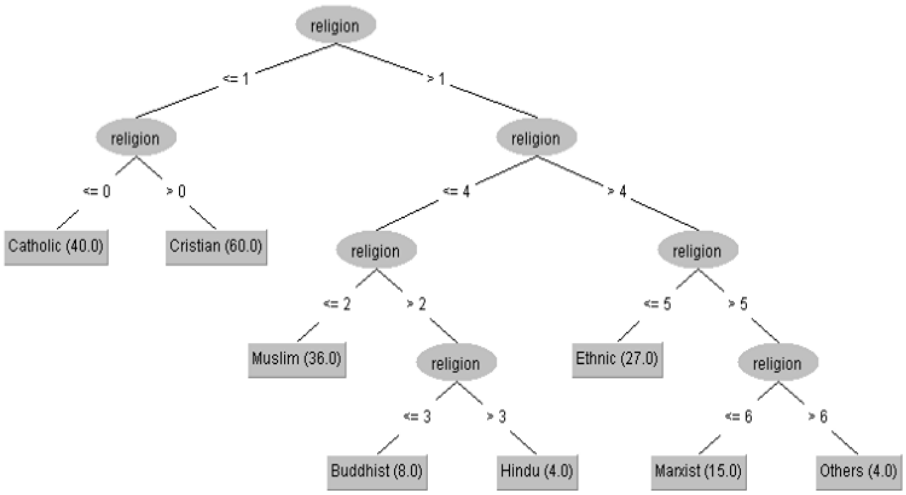

FIG: Decision Tree

3. IBK:

```
Correctly Classified Instances          93               47.9381 %
Incorrectly Classified Instances       101               52.0619 %
Kappa statistic                          0.3521
Mean absolute error                      0.134
Root mean squared error                  0.3531
Relative absolute error                 66.753  %
Root relative squared error            111.6218 %
Total Number of Instances              194

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.611    0.108    0.564      0.611    0.587      0.488    0.768     0.436     Muslim
                 0.400    0.067    0.333      0.400    0.364      0.307    0.641     0.184     Marxist
                 0.500    0.172    0.566      0.500    0.531      0.341    0.658     0.456     Cristian
                 0.525    0.149    0.477      0.525    0.500      0.363    0.701     0.373     Catholic
                 0.444    0.078    0.480      0.444    0.462      0.379    0.663     0.295     Ethnic
                 0.125    0.054    0.091      0.125    0.105      0.061    0.601     0.059     Buddhist
                 0.000    0.011    0.000      0.000    0.000      -0.015   0.539     0.023     Hindu
                 0.250    0.005    0.500      0.250    0.333      0.344    0.595     0.142     Others
Weighted Avg.    0.479    0.122    0.485      0.479    0.480      0.357    0.681     0.360
```

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h   <-- classified as
 22  2  2  2  7  0  1  0 |  a = Muslim
  1  6  2  1  0  4  0  1 |  b = Marxist
  6  4 30 17  0  2  1  0 |  c = Cristian
  2  0 13 21  3  1  0  0 |  d = Catholic
  4  2  3  3 12  3  0  0 |  e = Ethnic
  1  2  2  0  2  1  0  0 |  f = Buddhist
  2  0  1  0  1  0  0  0 |  g = Hindu
  1  2  0  0  0  0  0  1 |  h = Others
```

By considering a,b,c,d as positive and e,f,g,h negative 2*2 confusion matrix

	+	-
+	131	20
-	23	20

Correctly Classified Instances 93 47.9381 %
Incorrectly Classified Instances 101 52.0619 %

True Positive Rate (TPR) = 0.87

False Positive Rate = 0.53

4. Kstar:

```
=== Summary ===

Correctly Classified Instances         128               65.9794 %
Incorrectly Classified Instances        66               34.0206 %
Kappa statistic                          0.5671
Mean absolute error                      0.0844
Root mean squared error                  0.2613
Relative absolute error                 42.0338 %
Root relative squared error             82.5962 %
Total Number of Instances              194

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.722    0.127    0.565      0.722    0.634      0.544    0.925     0.769     Muslim
                 0.467    0.045    0.467      0.467    0.467      0.422    0.874     0.528     Marxist
                 0.800    0.127    0.738      0.800    0.768      0.659    0.905     0.800     Cristian
                 0.800    0.065    0.762      0.800    0.780      0.722    0.948     0.867     Catholic
                 0.519    0.042    0.667      0.519    0.583      0.531    0.916     0.686     Ethnic
                 0.125    0.011    0.333      0.125    0.182      0.184    0.700     0.241     Buddhist
                 0.000    0.011    0.000      0.000    0.000      -0.015   0.382     0.029     Hindu
                 0.000    0.000    ?          0.000    ?          ?        0.854     0.178     Others
Weighted Avg.    0.660    0.086    ?          0.660    ?          ?        0.896     0.719
```


```
=== Summary ===

Correctly Classified Instances         128               65.9794 %
Incorrectly Classified Instances        66               34.0206 %
Kappa statistic                          0.5671
Mean absolute error                      0.0844
Root mean squared error                  0.2613
Relative absolute error                 42.0338 %
Root relative squared error             82.5962 %
Total Number of Instances              194

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.722    0.127    0.565      0.722    0.634      0.544    0.925     0.769     Muslim
                 0.467    0.045    0.467      0.467    0.467      0.422    0.874     0.528     Marxist
                 0.800    0.127    0.738      0.800    0.768      0.659    0.905     0.800     Cristian
                 0.800    0.065    0.762      0.800    0.780      0.722    0.948     0.867     Catholic
                 0.519    0.042    0.667      0.519    0.583      0.531    0.916     0.686     Ethnic
                 0.125    0.011    0.333      0.125    0.182      0.184    0.700     0.241     Buddhist
                 0.000    0.011    0.000      0.000    0.000      -0.015   0.382     0.029     Hindu
                 0.000    0.000    ?          0.000    ?          ?        0.854     0.178     Others
Weighted Avg.    0.660    0.086    ?          0.660    ?          ?        0.896     0.719
```


By considering a,b,c,d as positive and e,f,g,h negative 2*2 confusion matrix

	+	-
+	144	7
-	24	19

Correctly Classified Instances	128	65.9794 %
Incorrectly Classified Instances	66	34.0206 %
True Positive Rate (TPR)	0.95	
False Positive Rate	0.56	

5. JRip:

```
Time taken to build model: 0.02 seconds

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

Correctly Classified Instances         190               97.9381 %
Incorrectly Classified Instances         4                2.0619 %
Kappa statistic                          0.974
Mean absolute error                      0.0065
Root mean squared error                  0.0686
Relative absolute error                  3.2427 %
Root relative squared error             21.7012 %
Total Number of Instances              194

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Muslim
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Marxist
                 1.000    0.030    0.938      1.000    0.968      0.954    0.994     0.975     Cristian
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Catholic
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Ethnic
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Buddhist
                 0.000    0.000    ?          0.000    ?          ?        0.937     0.143     Hindu
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Others
Weighted Avg.    0.979    0.009    ?          0.979    ?          ?        0.997     0.975
```

v

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h   <-- classified as
 36  0  0  0  0  0  0  0 |  a = Muslim
  0 15  0  0  0  0  0  0 |  b = Marxist
  0  0 60  0  0  0  0  0 |  c = Cristian
  0  0  0 40  0  0  0  0 |  d = Catholic
  0  0  0  0 27  0  0  0 |  e = Ethnic
  0  0  0  0  0  8  0  0 |  f = Buddhist
  0  0  4  0  0  0  0  0 |  g = Hindu
  0  0  0  0  0  0  0  4 |  h = Others
```

By considering a,b,c,d as positive and e,f,g,h negative 2*2 confusion matrix=

	+	-
+	151	0
-	4	39

Correctly Classified Instances 190 97.9381%
Incorrectly Classified Instances 4 2.0619 %

TruePositiveRate(TPR)=1

False Positive Rate = 0.10

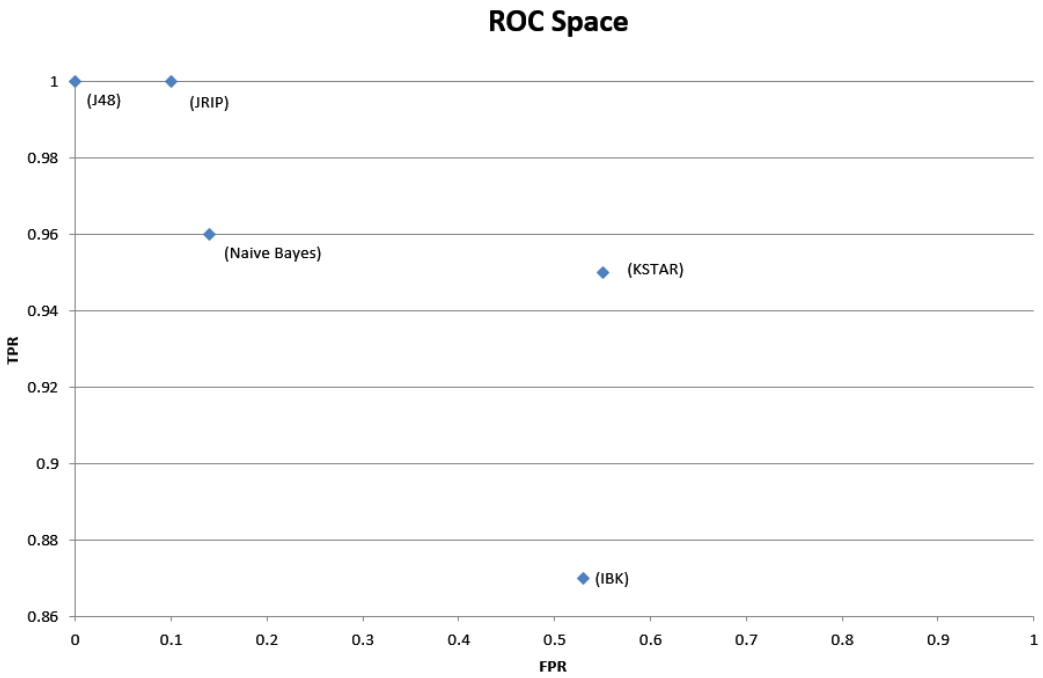

Fig: ROC Graph

ROC Graph:

I have applied 5 algorithms: **Naive Bayes, J48, IBk, Kstar, JRip**. With FPR and TPR values of these algorithms, I have plotted this ROC graph. From this ROC graph, we can see that J48 (1.00 , 0.00) point is closest among all other points to the best point (0,1). J48 generates the lowest False positive rate (FPR) which is 0.00 and the highest True positive rate (TPR) which is 1.00. It has also highest correctly classified instances which is 194. So, I will consider **J48** as the best classifier algorithm.