

**American International University-Bangladesh**

**Data Warehouse and Data Mining**

Section: B

Assignment-1 (Supervised Learning; Flags Dataset)

Assignment-2 (Unsupervised Learning; Cereal Data Set)

Submitted By,

**Name: Zahid Hossain**

**ID: 16-32597-2**

**Assignment-1 (Supeervised)**

**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. colours: 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 topmost 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

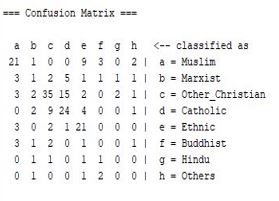
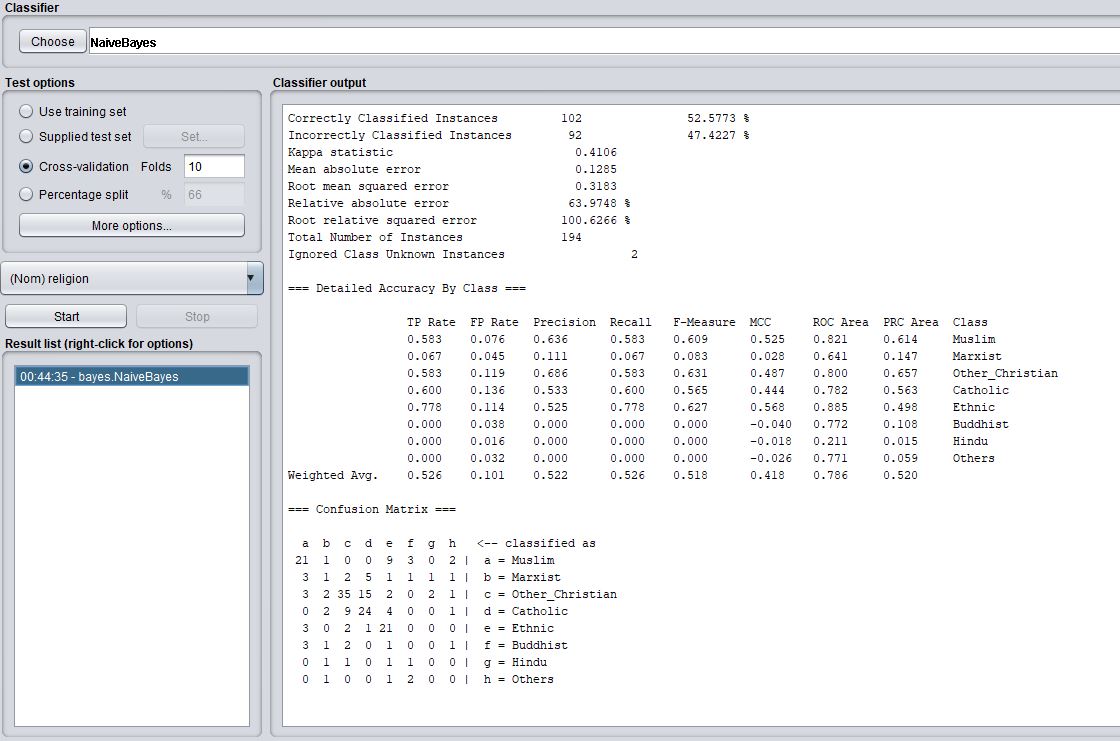
29. topleft: colour in the top-left corner (moving right to decide tie-breaks)

30. botright: Colour in the bottom-left corner (moving left to decide tie-breaks)

**Solution:**

For the solution, 5 classifiers have been used. These are:

1. Naive Bayes
2. J48 (Decision Tree)
3. IBk (Nearest Neighbor)
4. Random Tree
5. JRip
6. **Naive Bayes:**



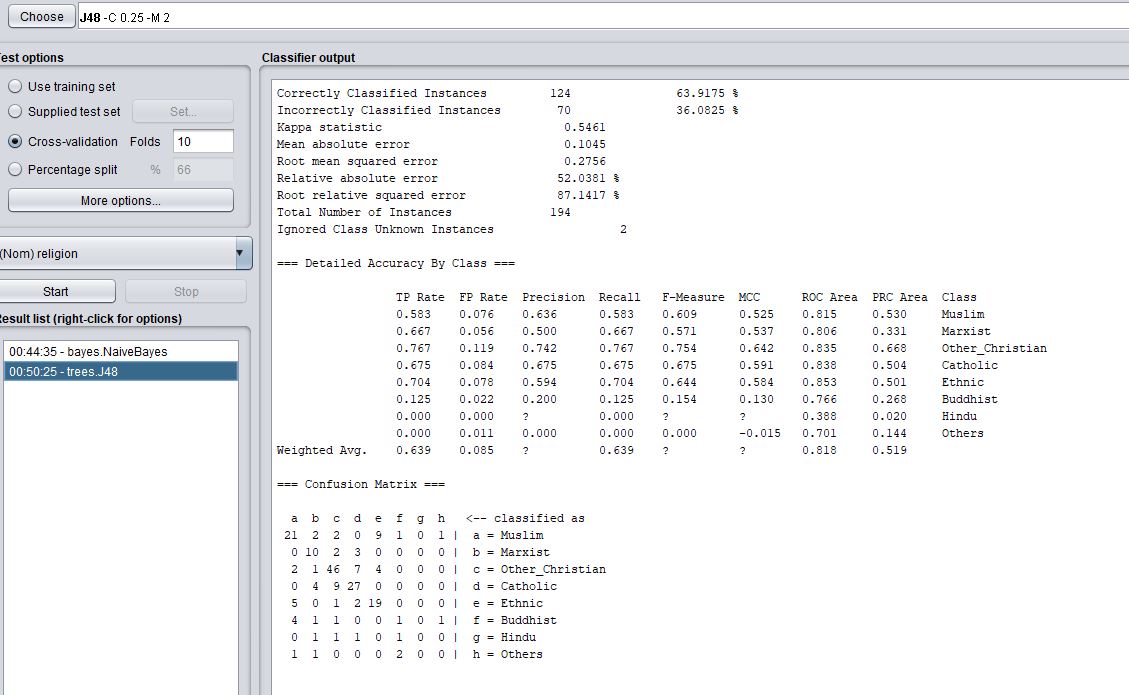
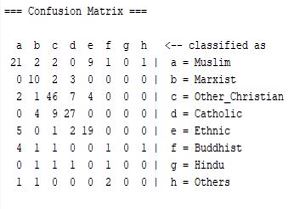
Correctly classified Instances = 102 [52.5773%]

Incorrectly classified Instances = 92 [47.4227%]

True Positive Rate (TPR) = 0.526

False Positive Rate = 0.101

1. **J48:**



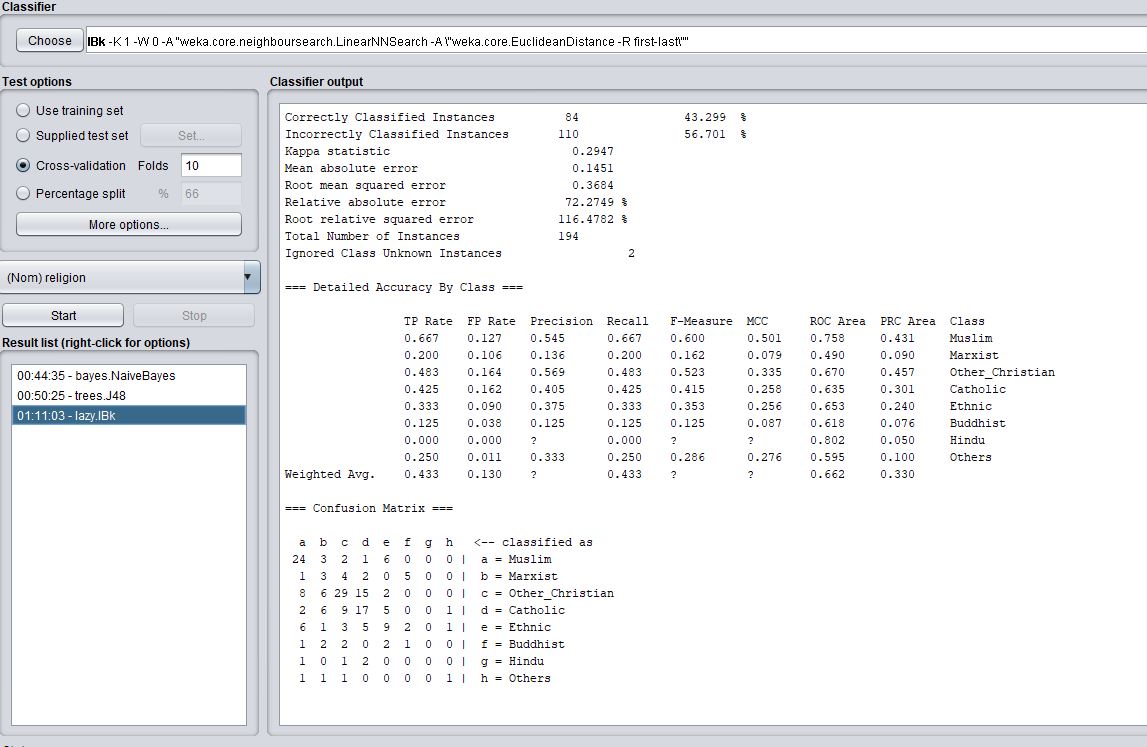
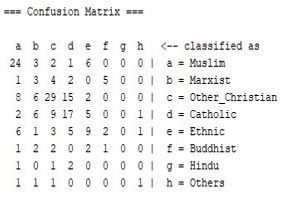
Correctly classified Instances = 124 [63.9175%]

Incorrectly classified Instances = 70 [36.0825%]

True Positive Rate (TPR) = 0.639

False Positive Rate = 0.085

1. **IBk:**



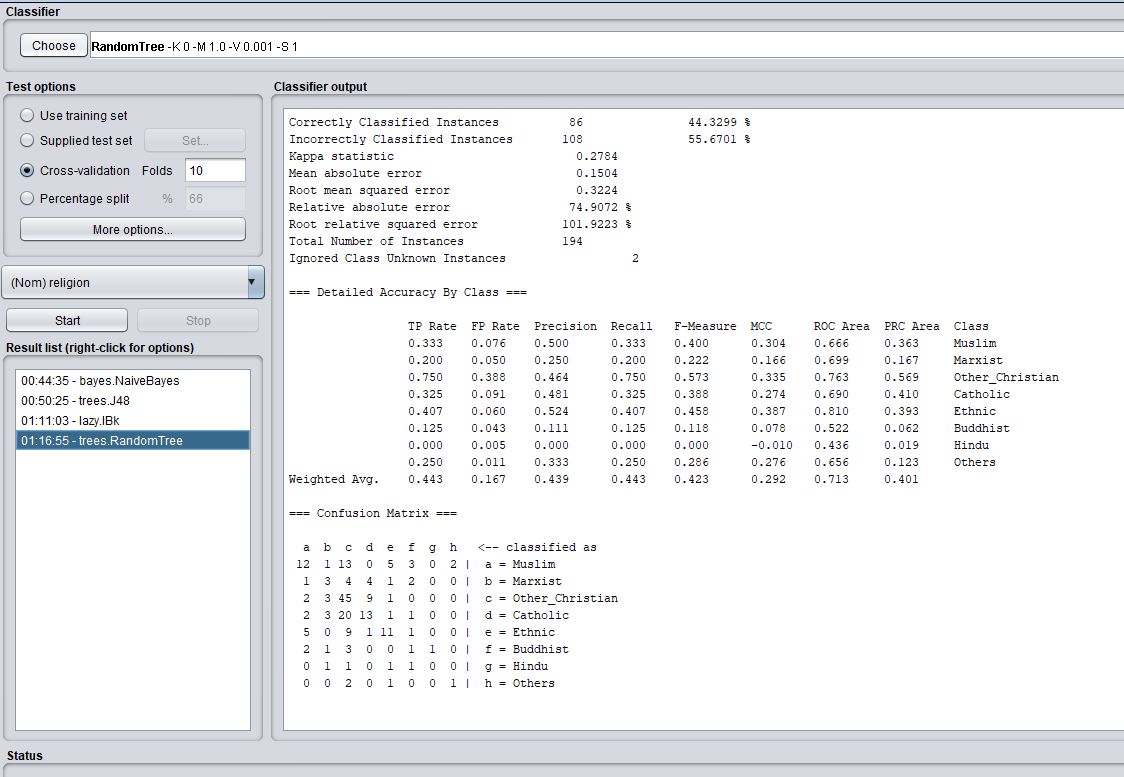
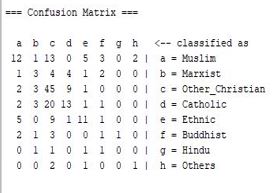
Correctly classified Instances = 84 [43.299%]

Incorrectly classified Instances = 110 [56.701%]

True Positive Rate (TPR) = 0.433

False Positive Rate = 0.130

1. **Random Tree:**



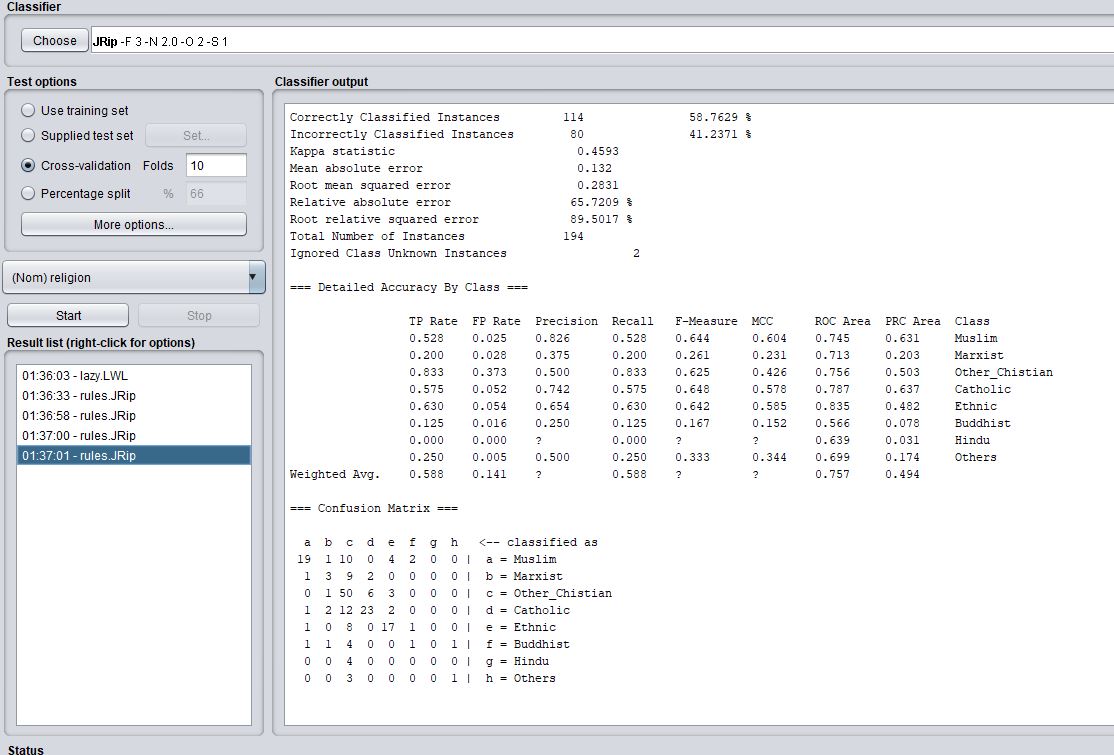
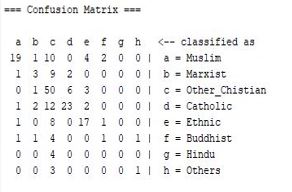
Correctly classified Instances = 86 [44.3299%]

Incorrectly classified Instances = 108 [55.6701%]

True Positive Rate (TPR) = 0.443

False Positive Rate = 0.167

1. **JRip:**



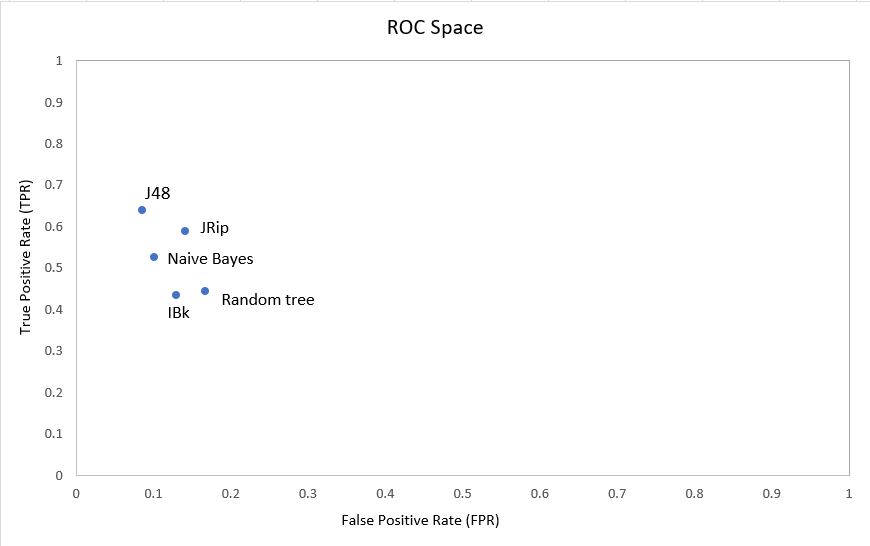
Correctly classified Instances = 114 [58.7629%]

Incorrectly classified Instances = 80 [41.2371%]

True Positive Rate (TPR) = 0.588

False Positive Rate = 0.141

**ROC Graph:**



I have applied 5 algorithms; Naïve Bayes, J48, IBk, Random Tree, JRip. With FPR and TPR values of these algorithms, I have plotted this ROC graph. From this ROC graph, we ca see that J48 (0.085,0.639) point is closest among all other points to the best point (0,1). J48 generates the lowest False positive rate (FPR) which is 0.085 and the highest True positive rate (TPR) which is 0.639. It has also highest correctly classified instances which is 124. So, I will consider J48 as the best classifier algorithm.