CIS6930 Project -1

Classification-Report

Name: Vagisha Tyagi

UFID: 0428-9808

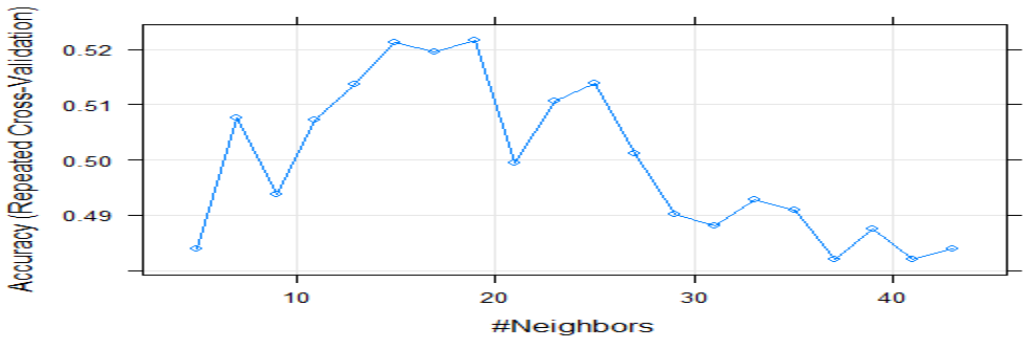
Dataset Preparation:

- Prepared the dataset by adding Continent column and separating columns into feature and class label role.
- Divided the dataset into training and test set in (80:20) ratio by creating own function which uses sample method to generate random permutation of training and test elements.
- Five groups of training and test set are created by passing 5 different seeds values so that same sample can be reproduced in future as well.
- As part of preprocessing, checked whether data has any NA with anyNA() method and dataset's attributes ranges were checked with summary().
- Converted the class label to a categorical variable by factoring it.
- For KNN classification, normalized the data by centering and scaling using preprocess method of caret package.

Classification methods:

1)KNN Classification

- Deployed KNN classification with help of preprocess and train method in caret package.
- Chose tunelength as 20 in train method and on basis of fit model results, it automatically selects best value.
- In our case as in the plot below, k=19 showed the maximum accuracy, hence k=19 was selected.



- KNN was called for 5 seeds and hence knn classification results for 5 groups were obtained.

  Average Accuracy across 5 groups :0.54

  Accuracy standard deviation across 5 groups:0.051
- To show results, confusion matrices and accuracy for each group is calculated.
- **KNN Results and Analysis**

```
myKnn(data,2018)
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            13    0      0             0       0             0
  Asia               1    3      3             2       2             1
  Europe             1    1      5             2       0             0
  North America      0    5      1             3       2             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5333
                 95% CI : (0.3787, 0.6834)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 0.004457

                  Kappa : 0.3958
 Mcnemar's Test P-Value : NA

> myKnn(data,2166)
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             9    4      0             2       2             0
  Asia               0    5      1             0       0             0
  Europe             2    2      7             1       2             0
  North America      0    1      0             3       1             0
  Oceania            0    0      0             1       0             0
  South America      0    0      1             0       0             1

Overall Statistics

               Accuracy : 0.5556
                 95% CI : (0.4, 0.7036)
    No Information Rate : 0.2667
    P-Value [Acc > NIR] : 3.924e-05

                  Kappa : 0.4368
 Mcnemar's Test P-Value : NA

> myKnn(data,2289)
Confusion Matrix and Statistics

               Reference
```

```
Prediction      Africa Asia Europe North America Oceania South America
  Africa             7    2      0             0       1             0
  Asia               0    3      1             2       2             2
  Europe             2    1     11             1       0             2
  North America      0    2      4             1       0             1
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.4889
                 95% CI : (0.337, 0.6423)
    No Information Rate : 0.3556
    P-Value [Acc > NIR] : 0.04539

                  Kappa : 0.3327
 Mcnemar's Test P-Value : NA

> myKnn(data,2322)
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             9    0      0             0       0             0
  Asia               1    6      1             1       2             1
  Europe             0    3      7             2       1             0
  North America      1    3      2             1       2             1
  Oceania            0    0      0             0       0             0
  South America      1    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5111
                 95% CI : (0.3577, 0.663)
    No Information Rate : 0.2667
    P-Value [Acc > NIR] : 0.0004137

                  Kappa : 0.3816
 Mcnemar's Test P-Value : NA

> myKnn(data,2408)
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            11    2      0             0       0             0
  Asia               1    6      1             2       2             0
  Europe             1    1      8             1       0             1
  North America      1    0      1             2       1             1
  Oceania            0    0      1             0       0             0
  South America      0    0      0             0       0             1

Overall Statistics

               Accuracy : 0.6222
                 95% CI : (0.4654, 0.7623)
    No Information Rate : 0.3111
```

```
   P-Value [Acc > NIR] : 1.666e-05

                  Kappa : 0.5118
Mcnemar's Test P-Value : NA
```

2) Ripper classification

- Deployed Ripper classification function with help of JRip method in RWeka package.
- Tuned the ripper classification by choosing Num of Folds as 50 in Weka Control function and on basis of fit model results, it will itself select the value that gives best accuracy results.
- Ripper function was called for 5 seeds and hence knn classification results for 5 groups were obtained.

  Average Accuracy across 5 groups :0.46

  Accuracy standard deviation across 5 groups:0.066
- To show results, confusion matrices and accuracy for each group is calculated.
- **Ripper Results and Analysis**

```
> myRipper(data,2018)
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            13    0      0             0       0             0
  Asia               0    4      3             2       1             1
  Europe             2    5      6             5       3             0
  North America      0    0      0             0       0             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5111
                 95% CI : (0.3577, 0.663)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 0.0103

                  Kappa : 0.358
 Mcnemar's Test P-Value : NA

> myRipper(data,2166)
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             9    9      2             6       3             1
  Asia               0    0      0             0       0             0
  Europe             2    3      7             1       2             0
  North America      0    0      0             0       0             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.3556
                 95% CI : (0.2187, 0.5122)
```

```
    No Information Rate : 0.2667
    P-Value [Acc > NIR] : 0.1206

                  Kappa : 0.1635
 Mcnemar's Test P-Value : NA

> myRipper(data,2289)
Confusion Matrix and Statistics

                Reference
Prediction       Africa Asia Europe North America Oceania South America
  Africa              8    4      3             1       0             2
  Asia                0    3      1             2       3             2
  Europe              1    1     12             1       0             1
  North America       0    0      0             0       0             0
  Oceania             0    0      0             0       0             0
  South America       0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5111
                 95% CI : (0.3577, 0.663)
    No Information Rate : 0.3556
    P-Value [Acc > NIR] : 0.02322

                  Kappa : 0.3483
 Mcnemar's Test P-Value : NA

> myRipper(data,2322)
Confusion Matrix and Statistics

                Reference
Prediction       Africa Asia Europe North America Oceania South America
  Africa             10    3      1             1       2             1
  Asia                1    6      3             1       1             1
  Europe              1    3      6             2       2             0
  North America       0    0      0             0       0             0
  Oceania             0    0      0             0       0             0
  South America       0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.4889
                 95% CI : (0.337, 0.6423)
    No Information Rate : 0.2667
    P-Value [Acc > NIR] : 0.001186

                  Kappa : 0.3159
 Mcnemar's Test P-Value : NA

> myRipper(data,2408)
Confusion Matrix and Statistics

                Reference
Prediction       Africa Asia Europe North America Oceania South America
  Africa             13    8      4             3       3             2
  Asia                0    0      0             0       0             0
  Europe              1    1      7             2       0             1
  North America       0    0      0             0       0             0
  Oceania             0    0      0             0       0             0
  South America       0    0      0             0       0             0

Overall Statistics
```

```
               Accuracy : 0.4444
                 95% CI : (0.2964, 0.6)
    No Information Rate : 0.3111
    P-Value [Acc > NIR] : 0.04112

                  Kappa : 0.2138
 Mcnemar's Test P-Value : NA
```

3) C4.5 classification

- Deployed C4.5 classification function with help of J48 method in RWeka package.
- Tuned the c4.5 classification by choosing reduced error pruning as true and minimum no. of instances as 9 in Weka Control function and on basis of fit model results, it will itself select the value that gives best accuracy results.
- C4.5 function was called for 5 seeds and hence c4.5 classification results for 5 groups were obtained.
  Average Accuracy across 5 groups :0.53
  Accuracy standard deviation across 5 groups:0.043
- To show results, confusion matrices and accuracy for each group is calculated.
- **C4.5 Results and Analysis**

```
> myc45(data,2018)
Confusion Matrix and Statistics

                Reference
Prediction       Africa Asia Europe North America Oceania South America
  Africa             13    0      0             0       0             0
  Asia                2    8      4             4       3             1
  Europe              0    1      5             3       1             0
  North America       0    0      0             0       0             0
  Oceania             0    0      0             0       0             0
  South America       0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5778
                 95% CI : (0.4215, 0.7234)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 0.0006533

                  Kappa : 0.4455
 Mcnemar's Test P-Value : NA

> myc45(data,2166)
Confusion Matrix and Statistics

                Reference
Prediction       Africa Asia Europe North America Oceania South America
  Africa              9    4      1             3       0             1
  Asia                0    5      1             2       3             0
  Europe              2    3      7             2       2             0
  North America       0    0      0             0       0             0
  Oceania             0    0      0             0       0             0
  South America       0    0      0             0       0             0

Overall Statistics
```

```
               Accuracy : 0.4667
                 95% CI : (0.3166, 0.6213)
    No Information Rate : 0.2667
    P-Value [Acc > NIR] : 0.003133

                  Kappa : 0.3037
 Mcnemar's Test P-Value : NA

> myc45(data,2289)
Confusion Matrix and Statistics

                Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             7    2      0             0       0             0
  Asia               0    4      2             3       3             2
  Europe             2    2     12             1       0             3
  North America      0    0      2             0       0             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5111
                 95% CI : (0.3577, 0.663)
    No Information Rate : 0.3556
    P-Value [Acc > NIR] : 0.02322

                  Kappa : 0.3418
 Mcnemar's Test P-Value : NA
> myc45(data,2322)
Confusion Matrix and Statistics

                Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             8    0      0             0       0             0
  Asia               2    8      2             2       4             2
  Europe             2    3      8             2       1             0
  North America      0    1      0             0       0             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5333
                 95% CI : (0.3787, 0.6834)
    No Information Rate : 0.2667
    P-Value [Acc > NIR] : 0.0001329

                  Kappa : 0.3803
 Mcnemar's Test P-Value : NA

> myc45(data,2408)
Confusion Matrix and Statistics

                Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            10    2      0             0       1             0
  Asia               3    6      2             3       2             1
  Europe             1    1      9             2       0             2
  North America      0    0      0             0       0             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0
```

```
Overall Statistics

               Accuracy : 0.5556
                 95% CI : (0.4, 0.7036)
    No Information Rate : 0.3111
    P-Value [Acc > NIR] : 0.0005798

                  Kappa : 0.4098
 Mcnemar's Test P-Value : NA
```

4) Support Vector Machine classification

- Deployed SVM classification function with help of tune method and using e1071 package.
- Tuned the svm fit classification by choosing kernel as linear and cost as 1 from tune method which gave best performance on cost.
- SVM function was called for 5 seeds and hence SVM classification results for 5 groups were obtained.
  Average Accuracy across 5 groups :0.55
  Accuracy standard deviation across 5 groups:0.040
- To show results, confusion matrices and accuracy for each group is calculated.
- **SVM Results and Analysis**

```
mySvm(data,2018)
Confusion Matrix and Statistics

                Reference
Prediction       Africa Asia Europe North America Oceania South America
  Africa             13    0      0             0       1             0
  Asia                1    5      3             3       2             1
  Europe              1    2      6             3       1             0
  North America       0    2      0             1       0             0
  Oceania             0    0      0             0       0             0
  South America       0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5556
                 95% CI : (0.4, 0.7036)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 0.001778

                  Kappa : 0.4163
 Mcnemar's Test P-Value : NA

> mySvm(data,2166)
Confusion Matrix and Statistics

                Reference
Prediction       Africa Asia Europe North America Oceania South America
  Africa              9    5      0             1       1             0
  Asia                0    4      1             3       3             0
  Europe              2    2      8             1       1             1
```

```
  North America      0    1      0             2       0             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5111
                 95% CI : (0.3577, 0.663)
    No Information Rate : 0.2667
    P-Value [Acc > NIR] : 0.0004137

                  Kappa : 0.3658
 Mcnemar's Test P-Value : NA

> mySvm(data,2289)
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             7    3      0             0       2             0
  Asia               0    3      0             2       1             3
  Europe             2    1     16             1       0             2
  North America      0    1      0             1       0             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.6
                 95% CI : (0.4433, 0.743)
    No Information Rate : 0.3556
    P-Value [Acc > NIR] : 0.0007157

                  Kappa : 0.4545
 Mcnemar's Test P-Value : NA

> mySvm(data,2322)
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            10    2      0             0       1             0
  Asia               0    4      2             1       1             2
  Europe             2    3      8             2       0             0
  North America      0    3      0             1       3             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5111
                 95% CI : (0.3577, 0.663)
    No Information Rate : 0.2667
    P-Value [Acc > NIR] : 0.0004137

                  Kappa : 0.3698
 Mcnemar's Test P-Value : NA
> mySvm(data,2408)
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            12    2      0             0       0             0
```

```
  Asia                1    5      2             3       3             0
  Europe              1    2      9             2       0             3
  North America       0    0      0             0       0             0
  Oceania             0    0      0             0       0             0
  South America       0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5778
                 95% CI : (0.4215, 0.7234)
    No Information Rate : 0.3111
    P-Value [Acc > NIR] : 0.0001935

                  Kappa : 0.436
 Mcnemar's Test P-Value : NA
```

Conclusion:

- On the given dataset, out of the 4 classification methods, maximum average accuracy over 5 groups was given by support vector machine as 56%.
- The classification can as well be tuned with other parameters to observe change in results.

Reference List:

[1]https://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy#List_by_the_CIA_.282016.29