# CIS6930 Fall 2017: Introduction to Data Mining
# Project 1: Classification

Debarshi Mitra

UFID: 33813136

**Description of the dataset preparation:** Initially the dataset is taken from the Life Expectancy Table in the section "List by the CIA (2016)" from Wikipedia and data is read from project1.csv file, where the continent column is included along with the data from the initial dataset. The dataset1 data frame contains the 4 columns – Overall life expectancy at birth, Male life expectancy at birth, Female life expectancy at birth and continent. Using the sample method, the data in the dataset1 data frame is divided into 2 parts in the ratio 80:20, where 80% of the dataset is included as part of the training set and rest 20% is included as part of the test set. These data are written consecutively in Training.csv and Test.csv if they don't exist initially and they are saved so that the accuracy generated is same for the given files. A for loop is run over to generate 5 sets of training and test files randomly.

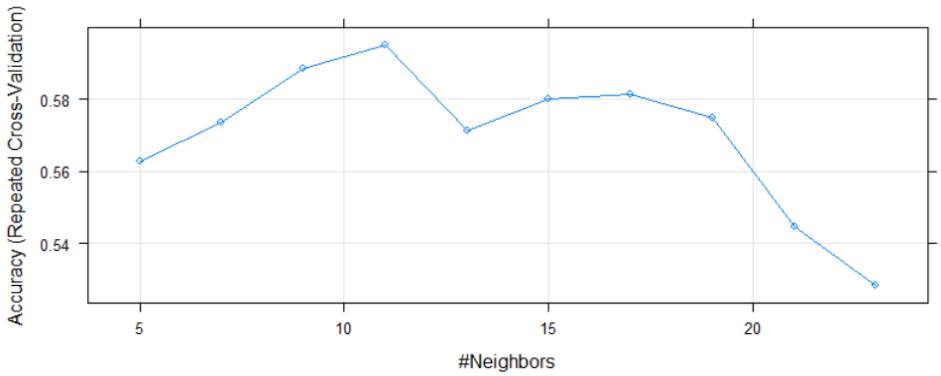
**Description of the classification methods:** The trainControl method is run with number equal to 10 and repeats equal to 3 (which gives the better results when different values are tried). The training set is trained using train method with classification models set to knn (K-Nearest Neighbor), followed by svmRadial (Support Vector Machine), J48(C4.5 of Decision Tree) and JRip (RIPPER of Decision Tree) with preprocess column set to center and scale and tuneLength set to value 10. All four of the classification methods are run inside a for loop for five iterations so for the five randomly selected training and test set files. Plots and confusion matrices are printed using the print statement for each of the classification model inside each of the iteration to generate the classification results and analyze the results.

**Classification results and analysis:**

Confusion Matrix and Statistics for Dataset1 (K-Nearest Neighbor):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            20    3      0             0       0             0
  Asia               1   10      6             3       0             0
  Europe             1    5     12             4       2             2
  North America      1    0      4             0       0             2
  Oceania            0    3      3             4       1             1
  South America      0    1      2             2       0             1

Overall Statistics

               Accuracy : 0.4681
                 95% CI : (0.3644, 0.5739)
    No Information Rate : 0.2872
    P-Value [Acc > NIR] : 0.0001535

                  Kappa : 0.3288
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8696      0.4545        0.4444                0.00000       0.33333              0.16667
Specificity                 0.9577      0.8611        0.7910                0.91358       0.87912              0.94318
Pos Pred Value              0.8696      0.5000        0.4615                0.00000       0.08333              0.16667
Neg Pred Value              0.9577      0.8378        0.7794                0.85057       0.97561              0.94318
Prevalence                  0.2447      0.2340        0.2872                0.13830       0.03191              0.06383
Detection Rate              0.2128      0.1064        0.1277                0.00000       0.01064              0.01064
Detection Prevalence        0.2447      0.2128        0.2766                0.07447       0.12766              0.06383
Balanced Accuracy           0.9137      0.6578        0.6177                0.45679       0.60623              0.55492
```

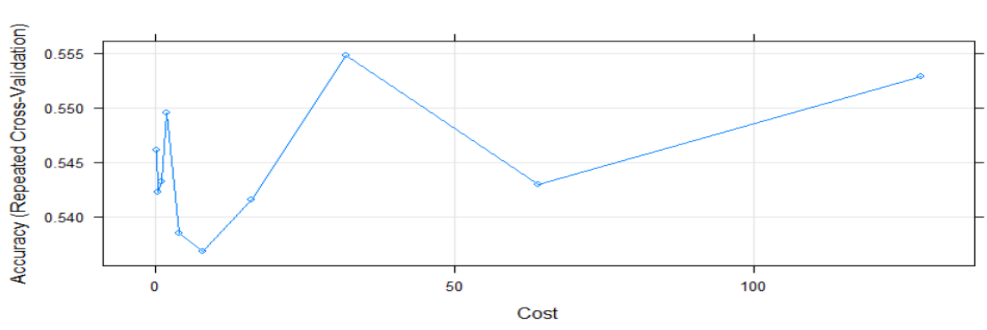
Tuning Hyperparameter:




Confusion Matrix and Statistics for Dataset1 (SVM):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            20    3      0             2       0             0
  Asia               0   12      5             3       0             1
  Europe             2    2     14             2       2             2
  North America      1    0      4             0       0             2
  Oceania            0    3      2             3       1             1
  South America      0    2      2             3       0             0

Overall Statistics

               Accuracy : 0.5
                 95% CI : (0.3951, 0.6049)
    No Information Rate : 0.2872
    P-Value [Acc > NIR] : 1.116e-05

                  Kappa : 0.3678
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8696      0.5455         0.5185                0.00000        0.33333                0.00000
Specificity                 0.9296      0.8750         0.8507                0.91358        0.90110                0.92045
Pos Pred Value              0.8000      0.5714         0.5833                0.00000        0.10000                0.00000
Neg Pred Value              0.9565      0.8630         0.8143                0.85057        0.97619                0.93103
Prevalence                  0.2447      0.2340         0.2872                0.13830        0.03191                0.06383
Detection Rate              0.2128      0.1277         0.1489                0.00000        0.01064                0.00000
Detection Prevalence        0.2660      0.2234         0.2553                0.07447        0.10638                0.07447
Balanced Accuracy           0.8996      0.7102         0.6846                0.45679        0.61722                0.46023
```

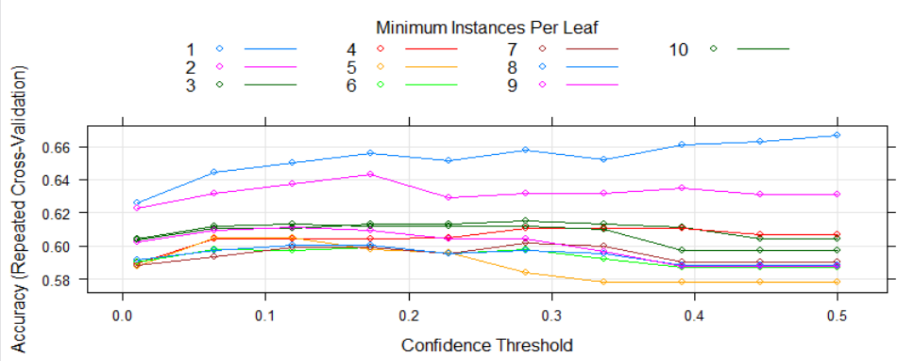
Tuning Hyperparameter:



## Confusion Matrix and Statistics for Dataset1 (Decision Tree – C4.5):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            20    4      0             1       0             0
  Asia               0    9      5             3       0             0
  Europe             1    3     10             4       2             3
  North America      0    2      3             0       0             0
  Oceania            2    2      5             3       1             3
  South America      0    2      4             2       0             0

Overall Statistics

               Accuracy : 0.4255
                 95% CI : (0.3241, 0.5318)
    No Information Rate : 0.2872
    P-Value [Acc > NIR] : 0.002883

                  Kappa : 0.2856
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8696     0.40909        0.3704              0.00000        0.33333              0.00000
Specificity                 0.9296     0.88889        0.8060              0.93827        0.83516              0.90909
Pos Pred Value              0.8000     0.52941        0.4348              0.00000        0.06250              0.00000
Neg Pred Value              0.9565     0.83117        0.7606              0.85393        0.97436              0.93023
Prevalence                  0.2447     0.23404        0.2872              0.13830        0.03191              0.06383
Detection Rate              0.2128     0.09574        0.1064              0.00000        0.01064              0.00000
Detection Prevalence        0.2660     0.18085        0.2447              0.05319        0.17021              0.08511
Balanced Accuracy           0.8996     0.64899        0.5882              0.46914        0.58425              0.45455
```

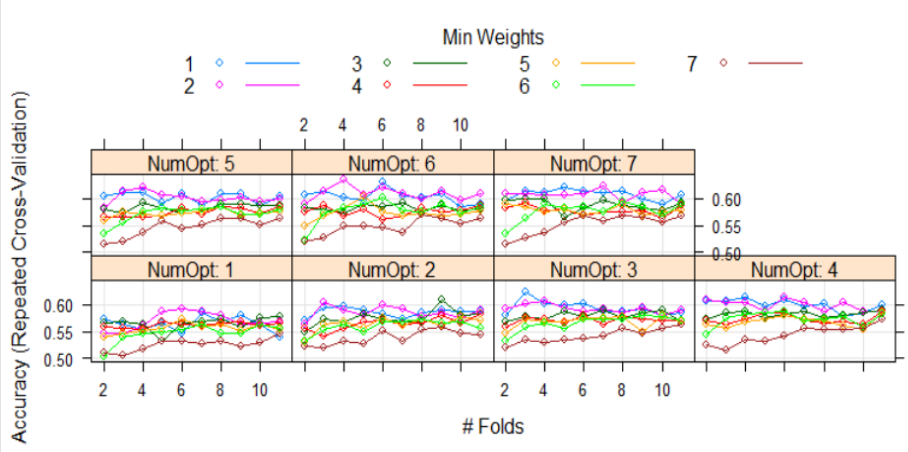
## Tuning Hyperparameter:




## Confusion Matrix and Statistics for Dataset1 (Decision Tree – Ripper):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            20    4      3             1       0             1
  Asia               0    8      5             5       0             1
  Europe             1    5     11             3       2             1
  North America      2    1      6             0       0             2
  Oceania            0    3      0             2       1             0
  South America      0    1      2             2       0             1

Overall Statistics

               Accuracy : 0.4362
                 95% CI : (0.3341, 0.5424)
    No Information Rate : 0.2872
    P-Value [Acc > NIR] : 0.00147

                  Kappa : 0.2814
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8696     0.36364        0.4074              0.0000        0.33333              0.16667
Specificity                 0.8732     0.84722        0.8209              0.8642        0.94505              0.94318
Pos Pred Value              0.6897     0.42105        0.4783              0.0000        0.16667              0.16667
Neg Pred Value              0.9538     0.81333        0.7746              0.8434        0.97727              0.94318
Prevalence                  0.2447     0.23404        0.2872              0.1383        0.03191              0.06383
Detection Rate              0.2128     0.08511        0.1170              0.0000        0.01064              0.01064
Detection Prevalence        0.3085     0.20213        0.2447              0.1170        0.06383              0.06383
Balanced Accuracy           0.8714     0.60543        0.6142              0.4321        0.63919              0.55492
```

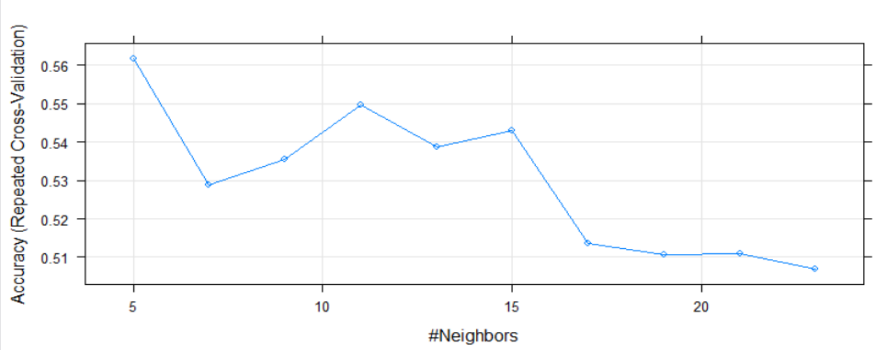
Tuning Hyperparameter:




Confusion Matrix and Statistics for Dataset2 (K-Nearest Neighbor):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            22    5      2             2       0             1
  Asia               2   14      0             1       2             4
  Europe             1    2     11             2       0             0
  North America      3    4      4             2       4             2
  Oceania            2    0      0             0       1             0
  South America      0    2      3             1       0             0

Overall Statistics

               Accuracy : 0.5051
                 95% CI : (0.4027, 0.6071)
    No Information Rate : 0.303
    P-Value [Acc > NIR] : 2.097e-05

                  Kappa : 0.3688
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.7333      0.5185        0.5500               0.25000        0.14286              0.00000
Specificity                 0.8551      0.8750        0.9367               0.81319        0.97826              0.93478
Pos Pred Value              0.6875      0.6087        0.6875               0.10526        0.33333              0.00000
Neg Pred Value              0.8806      0.8289        0.8916               0.92500        0.93750              0.92473
Prevalence                  0.3030      0.2727        0.2020               0.08081        0.07071              0.07071
Detection Rate              0.2222      0.1414        0.1111               0.02020        0.01010              0.00000
Detection Prevalence        0.3232      0.2323        0.1616               0.19192        0.03030              0.06061
Balanced Accuracy           0.7942      0.6968        0.7434               0.53159        0.56056              0.46739
```

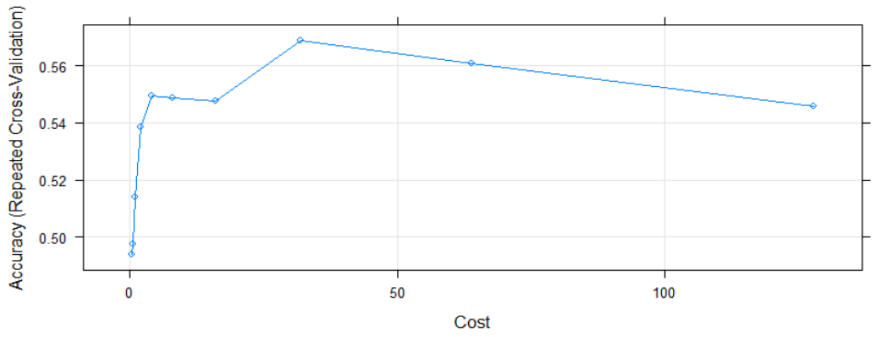
Tuning Hyperparameter:



## Confusion Matrix and Statistics for Dataset2 (SVM):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            22    3      0             0       0             0
  Asia               2    5      1             0       1             1
  Europe             1    5     15             3       0             3
  North America      2   12      2             4       5             3
  Oceania            2    1      0             1       1             0
  South America      1    1      2             0       0             0

Overall Statistics

               Accuracy : 0.4747
                 95% CI : (0.3734, 0.5776)
    No Information Rate : 0.303
    P-Value [Acc > NIR] : 0.0002469

                  Kappa : 0.3528
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.7333     0.18519        0.7500              0.50000        0.14286              0.00000
Specificity                 0.9565     0.93056        0.8481              0.73626        0.95652              0.95652
Pos Pred Value              0.8800     0.50000        0.5556              0.14286        0.20000              0.00000
Neg Pred Value              0.8919     0.75281        0.9306              0.94366        0.93617              0.92632
Prevalence                  0.3030     0.27273        0.2020              0.08081        0.07071              0.07071
Detection Rate              0.2222     0.05051        0.1515              0.04040        0.01010              0.00000
Detection Prevalence        0.2525     0.10101        0.2727              0.28283        0.05051              0.04040
Balanced Accuracy           0.8449     0.55787        0.7991              0.61813        0.54969              0.47826
```

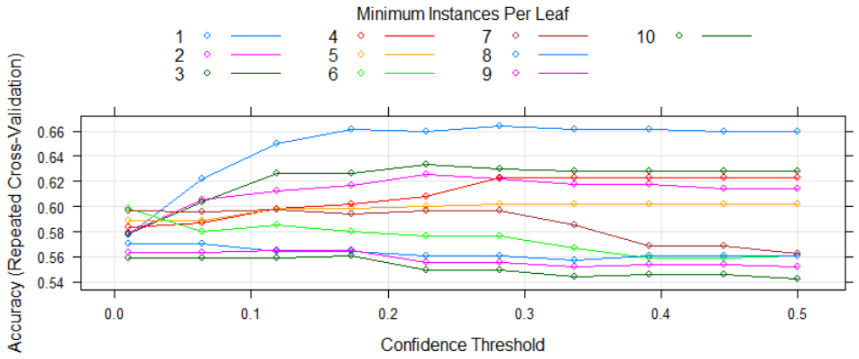
## Tuning Hyperparameter:




## Confusion Matrix and Statistics for Dataset2 (Decision Tree – C4.5):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            22    3      0             0       0             0
  Asia               2   13      0             1       2             2
  Europe             2    4     15             4       0             3
  North America      3    5      4             2       5             2
  Oceania            0    2      0             1       0             0
  South America      1    0      1             0       0             0

Overall Statistics

               Accuracy : 0.5253
                 95% CI : (0.4224, 0.6266)
    No Information Rate : 0.303
    P-Value [Acc > NIR] : 3.345e-06

                  Kappa : 0.3995
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.7333     0.4815        0.7500              0.25000        0.00000              0.00000
Specificity                 0.9565     0.9028        0.8354              0.79121        0.96739              0.97826
Pos Pred Value              0.8800     0.6500        0.5357              0.09524        0.00000              0.00000
Neg Pred Value              0.8919     0.8228        0.9296              0.92308        0.92708              0.92784
Prevalence                  0.3030     0.2727        0.2020              0.08081        0.07071              0.07071
Detection Rate              0.2222     0.1313        0.1515              0.02020        0.00000              0.00000
Detection Prevalence        0.2525     0.2020        0.2828              0.21212        0.03030              0.02020
Balanced Accuracy           0.8449     0.6921        0.7927              0.52060        0.48370              0.48913
```

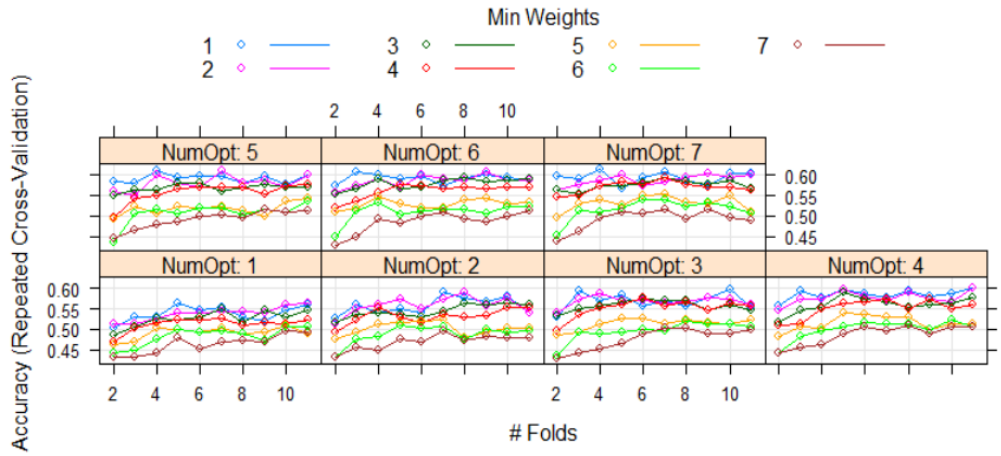
Tuning Hyperparameter:




Confusion Matrix and Statistics for Dataset2 (Decision Tree – Ripper):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            19    1      0             0       0             0
  Asia               3   17      2             3       5             3
  Europe             2    3     14             3       0             4
  North America      6    6      0             1       2             0
  Oceania            0    0      2             1       0             0
  South America      0    0      2             0       0             0

Overall Statistics

               Accuracy : 0.5152
                 95% CI : (0.4125, 0.6168)
    No Information Rate : 0.303
    P-Value [Acc > NIR] : 8.539e-06

                  Kappa : 0.3776
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.6333       0.6296        0.7000                0.12500        0.00000                0.00000
Specificity                 0.9855       0.7778        0.8481                0.84615        0.96739                0.97826
Pos Pred Value              0.9500       0.5152        0.5385                0.06667        0.00000                0.00000
Neg Pred Value              0.8608       0.8485        0.9178                0.91667        0.92708                0.92784
Prevalence                  0.3030       0.2727        0.2020                0.08081        0.07071                0.07071
Detection Rate              0.1919       0.1717        0.1414                0.01010        0.00000                0.00000
Detection Prevalence        0.2020       0.3333        0.2626                0.15152        0.03030                0.02020
Balanced Accuracy           0.8094       0.7037        0.7741                0.48558        0.48370                0.48913
```

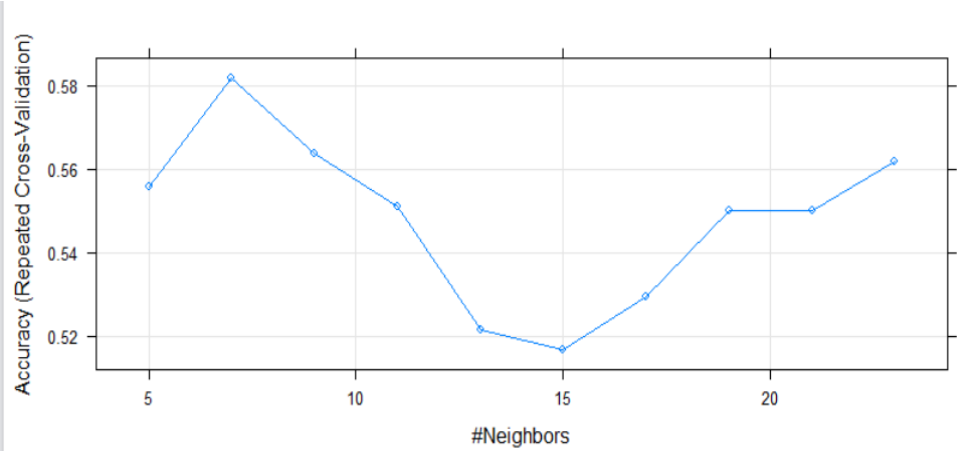
Tuning Hyperparameter:



## Confusion Matrix and Statistics for Dataset3 (K-Nearest Neighbor):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            20    1      0             1       1             0
  Asia               0    8      1             5       1             1
  Europe             1    9     15             3       1             5
  North America      0    1      8             0       1             0
  Oceania            2    3      0             2       2             1
  South America      0    2      1             0       1             0

Overall Statistics

               Accuracy : 0.4639
                 95% CI : (0.362, 0.5681)
    No Information Rate : 0.2577
    P-Value [Acc > NIR] : 9.557e-06

                  Kappa : 0.3219
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8696     0.33333        0.6000                0.0000        0.28571                0.00000
Specificity                 0.9595     0.89041        0.7361                0.8837        0.91111                0.95556
Pos Pred Value              0.8696     0.50000        0.4412                0.0000        0.20000                0.00000
Neg Pred Value              0.9595     0.80247        0.8413                0.8736        0.94253                0.92473
Prevalence                  0.2371     0.24742        0.2577                0.1134        0.07216                0.07216
Detection Rate              0.2062     0.08247        0.1546                0.0000        0.02062                0.00000
Detection Prevalence        0.2371     0.16495        0.3505                0.1031        0.10309                0.04124
Balanced Accuracy           0.9145     0.61187        0.6681                0.4419        0.59841                0.47778
```

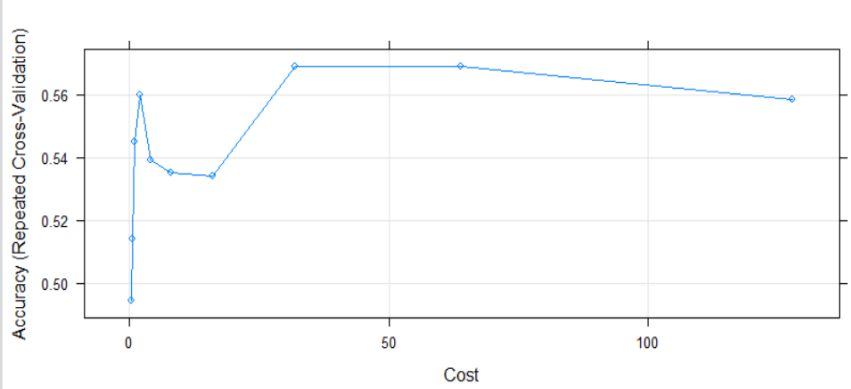
## Tuning Hyperparameter:




## Confusion Matrix and Statistics for Dataset3 (SVM):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            20    1      2             0       0             0
  Asia               0    9      2             5       2             1
  Europe             1    6     13             2       1             3
  North America      0    5      6             2       1             2
  Oceania            2    3      1             2       2             1
  South America      0    0      1             0       1             0

Overall Statistics

               Accuracy : 0.4742
                 95% CI : (0.3719, 0.5782)
    No Information Rate : 0.2577
    P-Value [Acc > NIR] : 3.663e-06

                  Kappa : 0.341
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8696     0.37500        0.5200                0.18182        0.28571                0.00000
Specificity                 0.9595     0.86301        0.8194                0.83721        0.90000                0.97778
Pos Pred Value              0.8696     0.47368        0.5000                0.12500        0.18182                0.00000
Neg Pred Value              0.9595     0.80769        0.8310                0.88889        0.94186                0.92632
Prevalence                  0.2371     0.24742        0.2577                0.11340        0.07216                0.07216
Detection Rate              0.2062     0.09278        0.1340                0.02062        0.02062                0.00000
Detection Prevalence        0.2371     0.19588        0.2680                0.16495        0.11340                0.02062
Balanced Accuracy           0.9145     0.61901        0.6697                0.50951        0.59286                0.48889
```

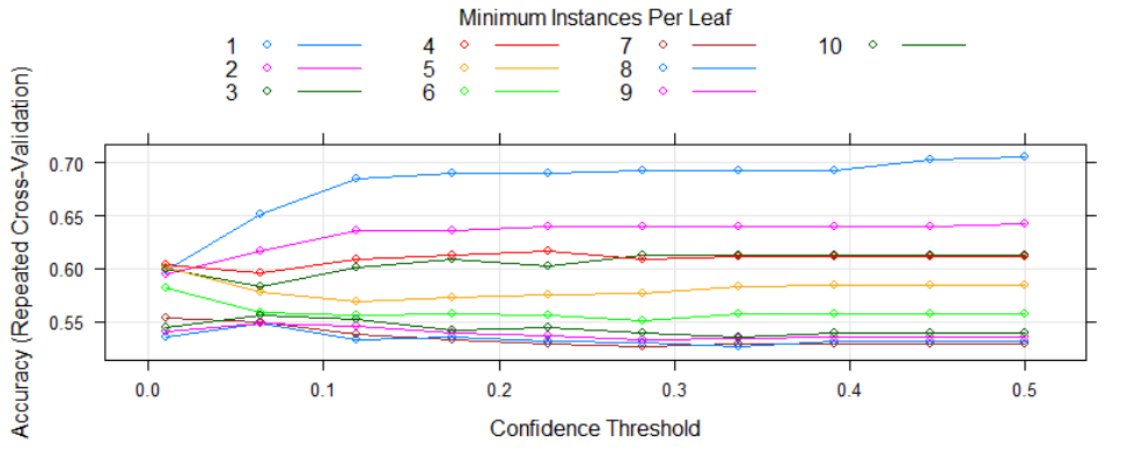
Tuning Hyperparameter:




Confusion Matrix and Statistics for Dataset3 (Decision Tree – C4.5):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            20    2      0             1       0             0
  Asia               0   10      1             4       2             2
  Europe             0    7     12             1       1             4
  North America      1    2     10             2       3             1
  Oceania            1    3      1             2       1             0
  South America      1    0      1             1       0             0

Overall Statistics

               Accuracy : 0.4639
                 95% CI : (0.362, 0.5681)
    No Information Rate : 0.2577
    P-Value [Acc > NIR] : 9.557e-06

                  Kappa : 0.3286
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8696      0.4167         0.4800               0.18182        0.14286              0.00000
Specificity                 0.9595      0.8767         0.8194               0.80233        0.92222              0.96667
Pos Pred Value              0.8696      0.5263         0.4800               0.10526        0.12500              0.00000
Neg Pred Value              0.9595      0.8205         0.8194               0.88462        0.93258              0.92553
Prevalence                  0.2371      0.2474         0.2577               0.11340        0.07216              0.07216
Detection Rate              0.2062      0.1031         0.1237               0.02062        0.01031              0.00000
Detection Prevalence        0.2371      0.1959         0.2577               0.19588        0.08247              0.03093
Balanced Accuracy           0.9145      0.6467         0.6497               0.49207        0.53254              0.48333
```

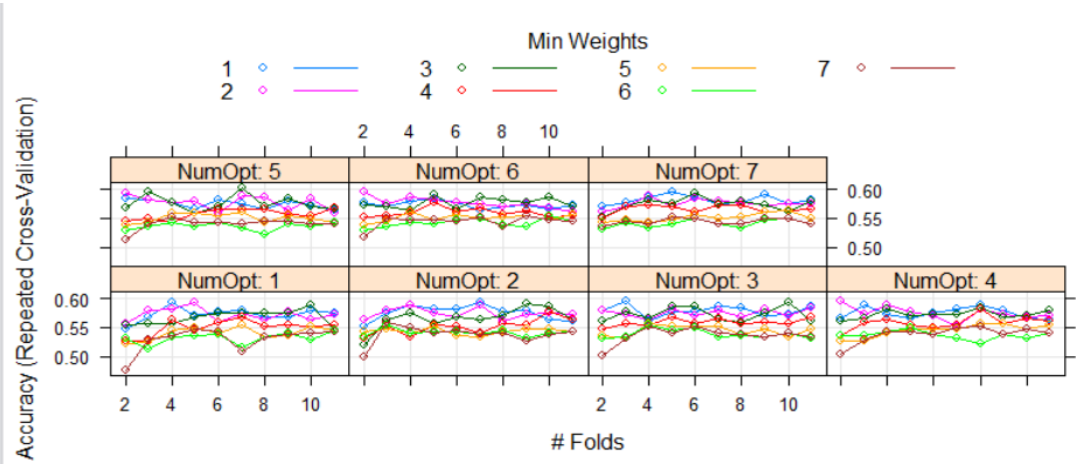
Tuning Hyperparameter:



Confusion Matrix and Statistics for Dataset3 (Decision Tree – Ripper):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            20    1      0             2       0             0
  Asia               1   14      2             4       4             2
  Europe             1    9     23             5       3             5
  North America      0    0      0             0       0             0
  Oceania            1    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5876
                 95% CI : (0.4831, 0.6867)
    No Information Rate : 0.2577
    P-Value [Acc > NIR] : 6.917e-12

                  Kappa : 0.4516
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8696      0.5833        0.9200                0.0000        0.00000              0.00000
Specificity                 0.9595      0.8219        0.6806                1.0000        0.98889              1.00000
Pos Pred Value              0.8696      0.5185        0.5000                   NaN        0.00000                  NaN
Neg Pred Value              0.9595      0.8571        0.9608                0.8866        0.92708              0.92784
Prevalence                  0.2371      0.2474        0.2577                0.1134        0.07216              0.07216
Detection Rate              0.2062      0.1443        0.2371                0.0000        0.00000              0.00000
Detection Prevalence        0.2371      0.2784        0.4742                0.0000        0.01031              0.00000
Balanced Accuracy           0.9145      0.7026        0.8003                0.5000        0.49444              0.50000
```

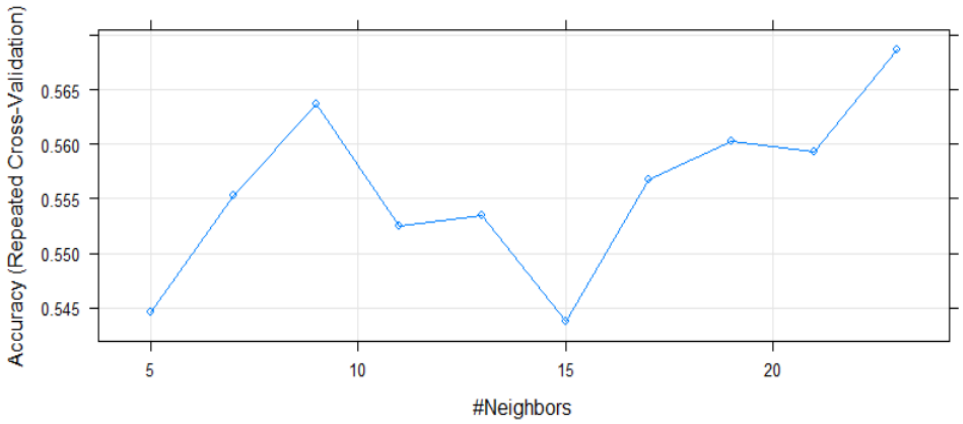
Tuning Hyperparameter:




Confusion Matrix and Statistics for Dataset4 (K-Nearest Neighbor):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            14    3      1             0       1             0
  Asia               3   15      9             6       4             5
  Europe             0    2     12             6       4             4
  North America      0    0      2             0       0             0
  Oceania            0    0      0             1       0             1
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.4409
                 95% CI : (0.338, 0.5476)
    No Information Rate : 0.2581
    P-Value [Acc > NIR] : 0.0001018

                  Kappa : 0.2857
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8235      0.7500        0.5000                0.00000       0.00000              0.0000
Specificity                 0.9342      0.6301        0.7681                0.97500       0.97619              1.0000
Pos Pred Value              0.7368      0.3571        0.4286                0.00000       0.00000                 NaN
Neg Pred Value              0.9595      0.9020        0.8154                0.85714       0.90110              0.8925
Prevalence                  0.1828      0.2151        0.2581                0.13978       0.09677              0.1075
Detection Rate              0.1505      0.1613        0.1290                0.00000       0.00000              0.0000
Detection Prevalence        0.2043      0.4516        0.3011                0.02151       0.02151              0.0000
Balanced Accuracy           0.8789      0.6901        0.6341                0.48750       0.48810              0.5000
```

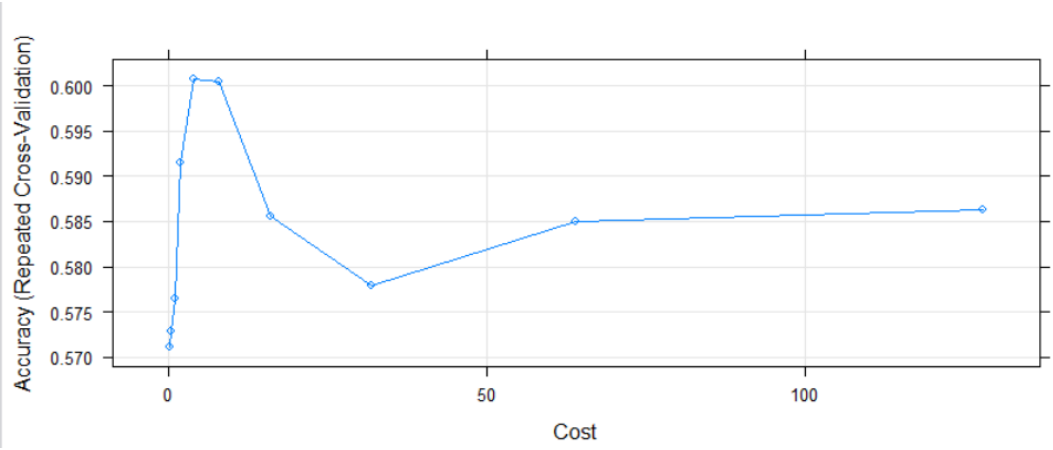
Tuning Hyperparameter:




Confusion Matrix and Statistics for Dataset4 (SVM):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            14    2      0             0       1             0
  Asia               3   16      5             8       5             7
  Europe             0    2     19             5       3             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 : 0.5269
                 95% CI : (0.4206, 0.6314)
    No Information Rate : 0.2581
    P-Value [Acc > NIR] : 3.025e-08

                  Kappa : 0.3903
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8235      0.8000        0.7917                0.0000        0.00000                0.0000
Specificity                 0.9605      0.6164        0.8116                1.0000        1.00000                1.0000
Pos Pred Value              0.8235      0.3636        0.5938                   NaN            NaN                   NaN
Neg Pred Value              0.9605      0.9184        0.9180                0.8602        0.90323                0.8925
Prevalence                  0.1828      0.2151        0.2581                0.1398        0.09677                0.1075
Detection Rate              0.1505      0.1720        0.2043                0.0000        0.00000                0.0000
Detection Prevalence        0.1828      0.4731        0.3441                0.0000        0.00000                0.0000
Balanced Accuracy           0.8920      0.7082        0.8016                0.5000        0.50000                0.5000
```

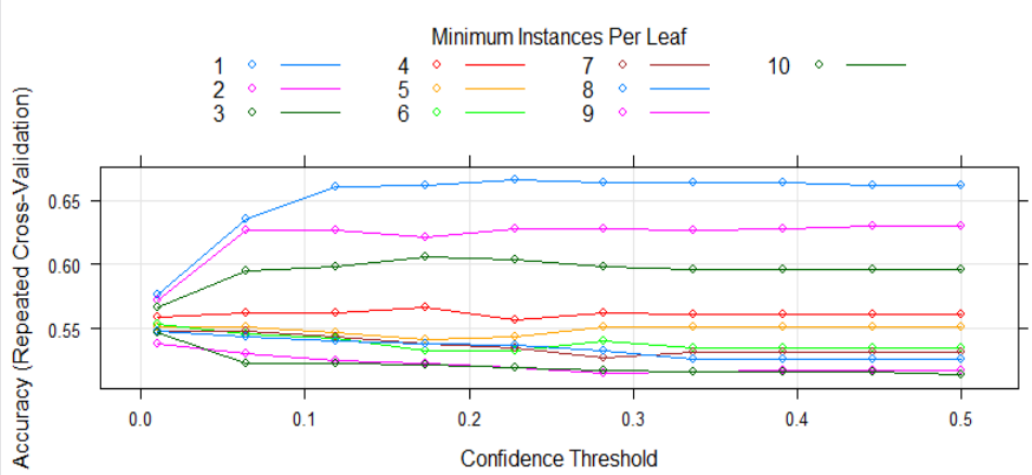
Tuning Hyperparameter:



## Confusion Matrix and Statistics for Dataset4 (Decision Tree – C4.5):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            15    3      2             1       1             1
  Asia               1    9      2             2       2             1
  Europe             0    5     14             6       4             4
  North America      0    0      3             2       1             2
  Oceania            0    1      0             0       1             1
  South America      1    2      3             2       0             1

Overall Statistics

               Accuracy : 0.4516
                 95% CI : (0.3481, 0.5583)
    No Information Rate : 0.2581
    P-Value [Acc > NIR] : 4.258e-05

                  Kappa : 0.3131
 Mcnemar's Test P-Value : 0.3758

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8824     0.45000        0.5833              0.15385        0.11111              0.10000
Specificity                 0.8947     0.89041        0.7246              0.92500        0.97619              0.90361
Pos Pred Value              0.6522     0.52941        0.4242              0.25000        0.33333              0.11111
Neg Pred Value              0.9714     0.85526        0.8333              0.87059        0.91111              0.89286
Prevalence                  0.1828     0.21505        0.2581              0.13978        0.09677              0.10753
Detection Rate              0.1613     0.09677        0.1505              0.02151        0.01075              0.01075
Detection Prevalence        0.2473     0.18280        0.3548              0.08602        0.03226              0.09677
Balanced Accuracy           0.8885     0.67021        0.6540              0.53942        0.54365              0.50181
```

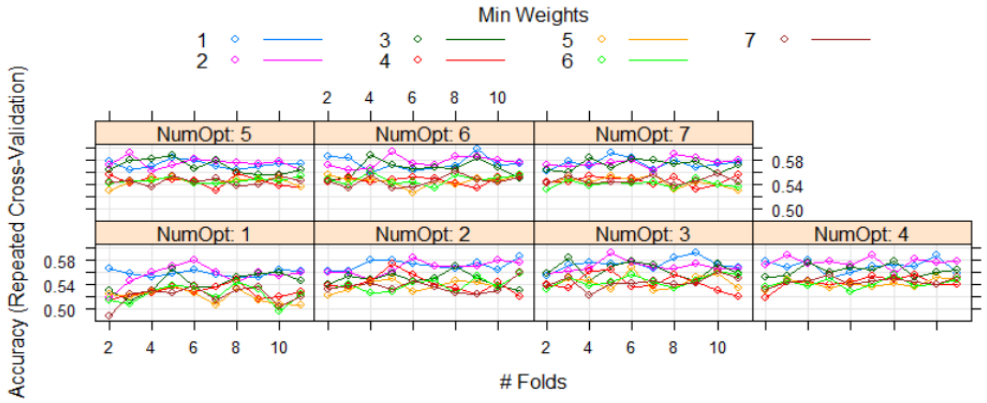
## Tuning Hyperparameter:




## Confusion Matrix and Statistics for Dataset4 (Decision Tree – Ripper):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            15    3      6             3       2             5
  Asia               2   12      4             4       2             3
  Europe             0    2     12             4       3             0
  North America      0    0      0             0       0             0
  Oceania            0    2      0             0       0             1
  South America      0    1      2             2       2             1

Overall Statistics

               Accuracy : 0.4301
                 95% CI : (0.3278, 0.5369)
    No Information Rate : 0.2581
    P-Value [Acc > NIR] : 0.0002333

                  Kappa : 0.2877
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8824     0.6000        0.5000              0.0000        0.00000              0.10000
Specificity                 0.7500     0.7945        0.8696              1.0000        0.96429              0.91566
Pos Pred Value              0.4412     0.4444        0.5714                 NaN        0.00000              0.12500
Neg Pred Value              0.9661     0.8788        0.8333              0.8602        0.90000              0.89412
Prevalence                  0.1828     0.2151        0.2581              0.1398        0.09677              0.10753
Detection Rate              0.1613     0.1290        0.1290              0.0000        0.00000              0.01075
Detection Prevalence        0.3656     0.2903        0.2258              0.0000        0.03226              0.08602
Balanced Accuracy           0.8162     0.6973        0.6848              0.5000        0.48214              0.50783
```

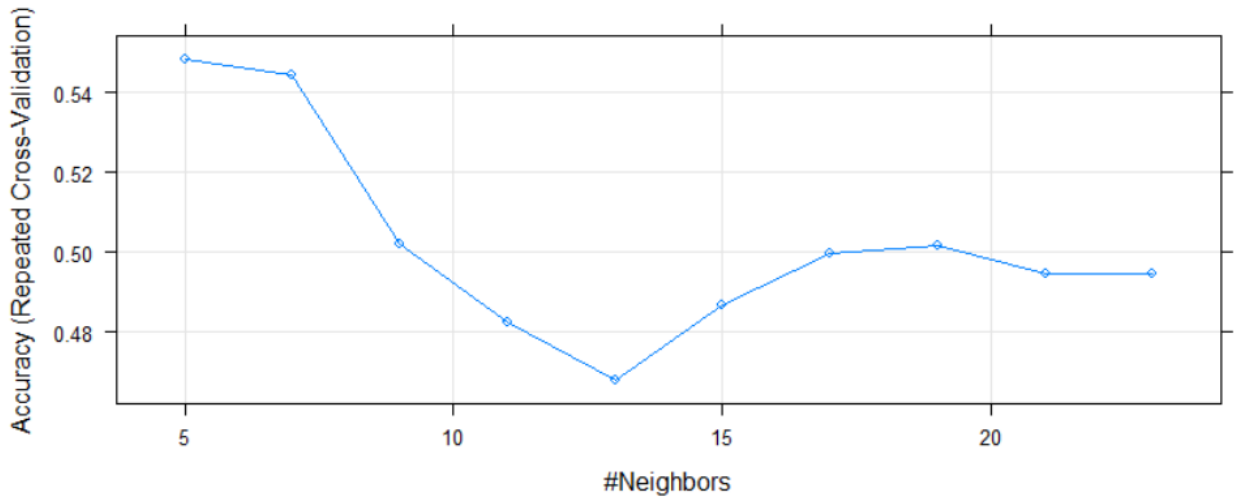
Tuning Hyperparameter:




Confusion Matrix and Statistics for Dataset5 (K-Nearest Neighbor):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            18    3      0             2       0             0
  Asia               1   10      3             3       0             1
  Europe             1    5     15             4       1             3
  North America      0    1      4             1       0             0
  Oceania            2    4      1             3       1             1
  South America      1    4      3             2       0             1

Overall Statistics

               Accuracy : 0.4646
                 95% CI : (0.3638, 0.5677)
    No Information Rate : 0.2727
    P-Value [Acc > NIR] : 3.488e-05

                  Kappa : 0.3318
 Mcnemar's Test P-Value : 0.1607

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.7826      0.3704         0.5769                0.06667        0.50000              0.16667
Specificity                 0.9342      0.8889         0.8082                0.94048        0.88660              0.89247
Pos Pred Value              0.7826      0.5556         0.5172                0.16667        0.08333              0.09091
Neg Pred Value              0.9342      0.7901         0.8429                0.84946        0.98851              0.94318
Prevalence                  0.2323      0.2727         0.2626                0.15152        0.02020              0.06061
Detection Rate              0.1818      0.1010         0.1515                0.01010        0.01010              0.01010
Detection Prevalence        0.2323      0.1818         0.2929                0.06061        0.12121              0.11111
Balanced Accuracy           0.8584      0.6296         0.6926                0.50357        0.69330              0.52957
```

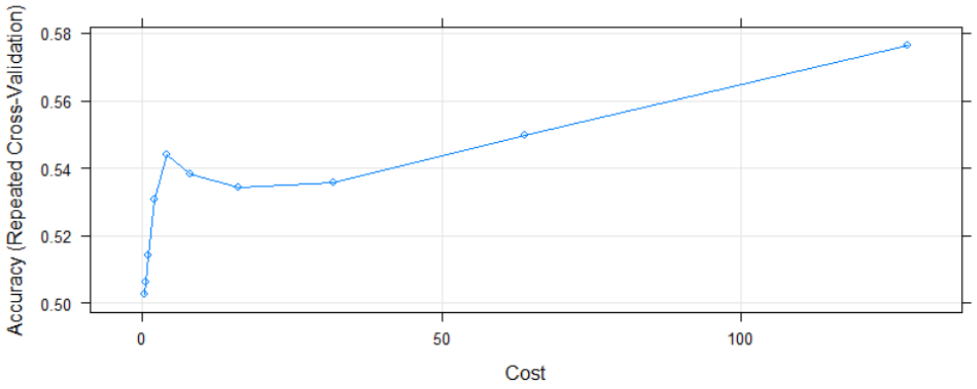
Tuning Hyperparameter:



## Confusion Matrix and Statistics for Dataset5 (SVM):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            20    3      1             2       0             1
  Asia               0   13      2             5       1             0
  Europe             1    5     18             4       1             3
  North America      0    1      3             1       0             0
  Oceania            1    2      0             1       0             1
  South America      1    3      2             2       0             1

Overall Statistics

               Accuracy : 0.5354
                 95% CI : (0.4323, 0.6362)
    No Information Rate : 0.2727
    P-Value [Acc > NIR] : 3.117e-08

                  Kappa : 0.4041
 Mcnemar's Test P-Value : 0.2311

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8696      0.4815        0.6923                0.06667        0.00000                0.16667
Specificity                 0.9079      0.8889        0.8082                0.95238        0.94845                0.91398
Pos Pred Value              0.7407      0.6190        0.5625                0.20000        0.00000                0.11111
Neg Pred Value              0.9583      0.8205        0.8806                0.85106        0.97872                0.94444
Prevalence                  0.2323      0.2727        0.2626                0.15152        0.02020                0.06061
Detection Rate              0.2020      0.1313        0.1818                0.01010        0.00000                0.01010
Detection Prevalence        0.2727      0.2121        0.3232                0.05051        0.05051                0.09091
Balanced Accuracy           0.8887      0.6852        0.7503                0.50952        0.47423                0.54032
```

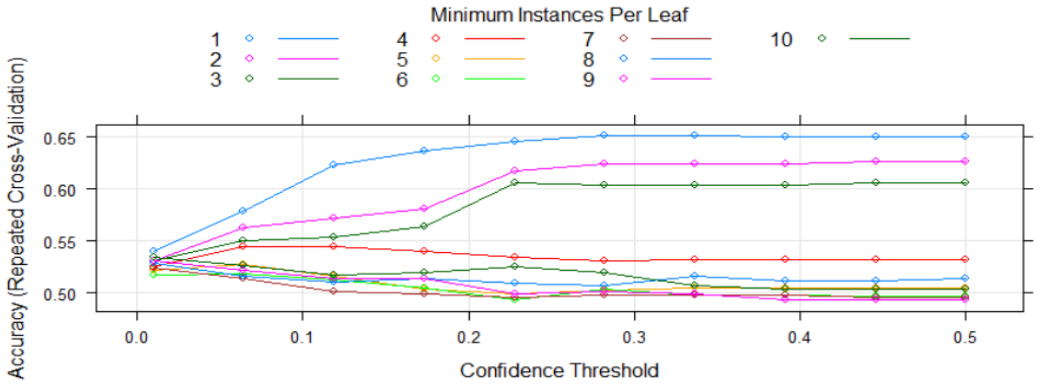
## Tuning Hyperparameter:




## Confusion Matrix and Statistics for Dataset5 (Decision Tree – C4.5):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            19    4      2             2       0             0
  Asia               1   12      3             5       1             1
  Europe             0    2     14             3       1             3
  North America      1    4      2             3       0             1
  Oceania            1    2      1             0       0             0
  South America      1    3      4             2       0             1

Overall Statistics

               Accuracy : 0.4949
                 95% CI : (0.3929, 0.5973)
    No Information Rate : 0.2727
    P-Value [Acc > NIR] : 2.172e-06

                  Kappa : 0.359
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.8261      0.4444        0.5385                0.2000         0.0000                0.16667
Specificity                 0.8947      0.8472        0.8767                0.9048         0.9588                0.89247
Pos Pred Value              0.7037      0.5217        0.6087                0.2727         0.0000                0.09091
Neg Pred Value              0.9444      0.8026        0.8421                0.8636         0.9789                0.94318
Prevalence                  0.2323      0.2727        0.2626                0.1515         0.0202                0.06061
Detection Rate              0.1919      0.1212        0.1414                0.0303         0.0000                0.01010
Detection Prevalence        0.2727      0.2323        0.2323                0.1111         0.0404                0.11111
Balanced Accuracy           0.8604      0.6458        0.7076                0.5524         0.4794                0.52957
```

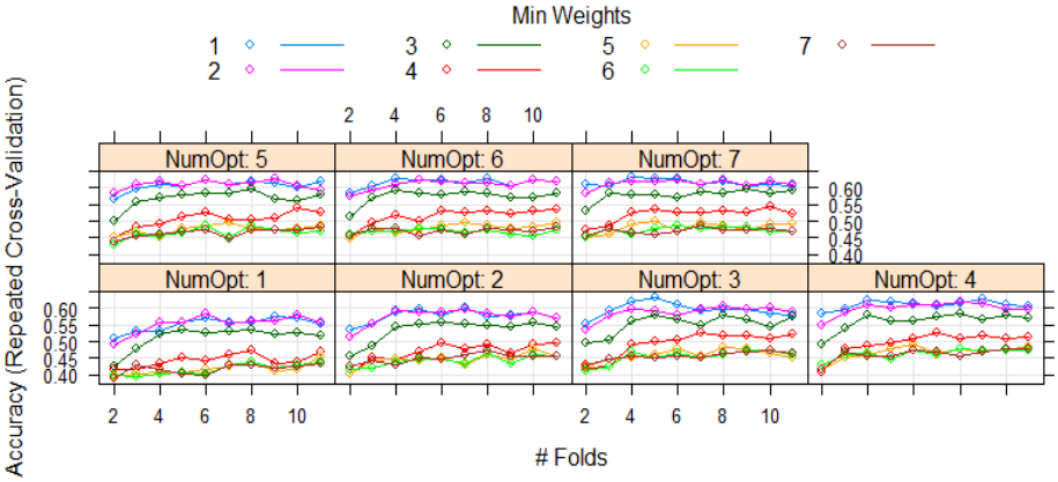
Tuning Hyperparameter:




Confusion Matrix and Statistics for Dataset5 (Decision Tree – Ripper):

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            18    8      1             4       1             2
  Asia               3   12      3             4       1             0
  Europe             0    2     17             5       0             3
  North America      0    0      2             0       0             0
  Oceania            0    1      0             1       0             0
  South America      2    4      3             1       0             1

Overall Statistics

               Accuracy : 0.4848
                 95% CI : (0.3832, 0.5875)
    No Information Rate : 0.2727
    P-Value [Acc > NIR] : 5.696e-06

                  Kappa : 0.3353
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.7826      0.4444        0.6538                0.0000        0.0000              0.16667
Specificity                 0.7895      0.8472        0.8630                0.9762        0.9794              0.89247
Pos Pred Value              0.5294      0.5217        0.6296                0.0000        0.0000              0.09091
Neg Pred Value              0.9231      0.8026        0.8750                0.8454        0.9794              0.94318
Prevalence                  0.2323      0.2727        0.2626                0.1515        0.0202              0.06061
Detection Rate              0.1818      0.1212        0.1717                0.0000        0.0000              0.01010
Detection Prevalence        0.3434      0.2323        0.2727                0.0202        0.0202              0.11111
Balanced Accuracy           0.7860      0.6458        0.7584                0.4881        0.4897              0.52957
```

Tuning Hyperparameter:



**Conclusion:** Overall, the SVM Classification gave more accuracy than the others though the highest accuracy percentage was found in Decision Tree – RIPPER followed by Decision Tree C4.5

**Reference:**


1. https://cran.r-project.org/web/packages/caret/caret.pdf
2. https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html
3. https://cran.r-project.org/web/packages/RWeka/index.html
4. https://www.youtube.com/watch?v=GtgJEVxl7DY&t=3s
5. http://dataaspirant.com/2017/01/09/knn-implementation-r-using-caret-package/
6. https://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy#List_by_the_CIA_.282016.29
7. https://www.rdocumentation.org/packages/graphics/versions/3.4.0/topics/plot
8. https://www.rdocumentation.org/packages/base/versions/3.4.1/topics/paste
9. https://www.rdocumentation.org/packages/base/versions/3.4.1/topics/sample
10. https://www.rdocumentation.org/packages/caret/versions/6.0-77/topics/trainControl
11. https://www.rdocumentation.org/packages/caret/versions/6.0-77/topics/train
12. https://www.rdocumentation.org/packages/raster/versions/2.5-8/topics/predict