Introduction to Data Mining (CIS 6930)

Fall 2017

Project 1: Classification Report

Shikha Dharmendra Mehta

UFID 48519256

shikha.mehta@ufl.edu

DATASET PREPARATION

Life Expectancy Table is the main dataset used for this project. This dataset (obtained from Reference [1]) requires an additional Continent column, mapping every entity to its continent. I added this column and saved the complete dataset in an Excel format for importing into R.

Training and Test Set Creation

- The dataset is divided into training set (80%) and test set (20%) using data partitioning function *createDataPartition* in ***caret*** package.
- This partitioning is done based on our categorical variable Continent, to ensure a balanced partitioning between the training and test sets.
- *createDataPartition* randomly selects observations from the dataset to add to the training set, with a probability of 0.8.
- All the remaining observations are transferred to test set.

Training and Test Set Groups

- I independently repeat the above steps for creating 5 different sets of training and test sets.
- For each classification method, the standard deviation and average value of predictions over all the sets has been computed and presented in the Results section.

Data Preprocessing

- We can observe from the dataset that the range of numeric values for each of the feature variables (Rank, Overall life expectancy at birth, Male life expectancy at birth, Female life expectancy at birth) differ. To avoid a bias in prediction due to this, we need to normalize these values.
- In the R script, we are doing this by setting the preProcess parameter in the *train* function (from ***caret***). This parameter is set with values “center” and “scale”, ensuring that the training set is normalized before a classification model is trained on it.

CLASSIFICATION METHODS & PARAMETERS

1. Support Vector Machine Classification

- Algorithm:
  Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.
- Implementation: package used – ***caret***
  svmLinear3 method was used as a parameter to *train* function of ***caret*** package. To train this classifier, data was centered and scaled for normalization. Model obtained from this is then used to predict Continent label for the test data set and confusion matrix is generated again to get prediction metrics for the algorithm.
- Parameters involved: **cost**, **Loss** function
  We specify the tuneLength parameter as 3 in the *train* method to tell the algorithm how many different values of each parameter it should try. We have used the default control parameters (simple bootstrap resampling) on the training set to determine accuracies for the different combinations of parameter values it tries. Based on our specifications, the *train* method automatically selects the model that yields the best accuracy, and hence the best values for cost and Loss function.

2. k - Nearest Neighbour (kNN) Classification

- Algorithm:
  k - nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance). k-NN has

been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

- Implementation: package used - ***caret***
  caret package function *train* was used to train the KNN classifier, and a model was generated. Since there is significant difference between ranges of 'Rank' and other three attributes, preprocessing steps of centering the data and scaling was done. The model was used to predict Continent label for the test data set and a confusion matrix was obtained. We get Accuracy, Precision, Recall and F1 measures from the confusion matrix.
- Parameters involved: **k**
  We specify the tuneLength parameter as 9 in the *train* method to tell the algorithm how many different values of k it should try. We have used the default control parameters (simple bootstrap resampling) on the training set to determine accuracies for the different values of k it tries. Based on our specifications, the *train* method automatically selects the model that yields the best accuracy, and hence the best value of k.

3. RIPPER Classification

- Algorithm:
  Repeated Incremental Pruning to Produce Error Reduction (RIPPER), is based in association rules with reduced error pruning (REP), a very common and effective technique found in decision tree algorithms. In REP for rules algorithms, the training data is split into a growing set and a pruning set. First, an initial rule set is formed that over the growing set, using some heuristic method. This overlarge rule set is then repeatedly simplified by applying one of a set of pruning operators typical pruning operators would be to delete any single condition or any single rule. At each stage of simplification, the pruning operator chosen is the one that yields the greatest reduction of error on the pruning set. Simplification ends when applying any pruning operator would increase error on the pruning set.
- Implementation: package used - ***rJava, RWeka, caret***
  JRip method was used as a parameter to *train* function of ***caret*** package. To train this classifier, data was centered and scaled for normalization. Model obtained

from this is then used to predict Continent label for the test data set and confusion matrix is generated again to get prediction metrics for the algorithm.

- Parameters involved: **NumOpt, NumFolds, MinWeights**
  We specify the tuneLength parameter as 3 in the *train* method to tell the algorithm how many different values of each parameter it should try. We have used the default control parameters (simple bootstrap resampling) on the training set to determine accuracies for the different combinations of parameter values it tries. Based on our specifications, the *train* method automatically selects the model that yields the best accuracy, and hence the best values for cost and Loss function.

4. C4.5 Classification

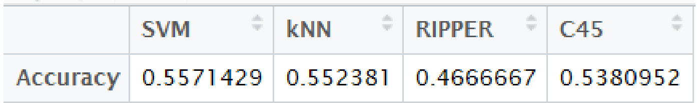
- Algorithm:
  C4.5 is an algorithm used to generate a decision tree is an extension of earlier ID3 Algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.
- Implementation: package used - ***rJava, RWeka, caret***
  J48 method was used as a parameter to *train* function of ***caret*** package. To train this classifier, data was centered and scaled for normalization. Model obtained from this is then used to predict Continent label for the test data set and confusion matrix is generated again to get prediction metrics for the algorithm.
- Parameters involved: **C, M**
  We specify the tuneLength parameter as 3 in the *train* method to tell the algorithm how many different values of each parameter it should try. We have used the default control parameters (simple bootstrap resampling) on the training set to determine accuracies for the different combinations of parameter values it tries. Based on our specifications, the *train* method automatically selects the model that yields the best accuracy, and hence the best values for cost and Loss function.

CLASSIFICATION RESULTS & ANALYSIS

Summary of Entire Dataset

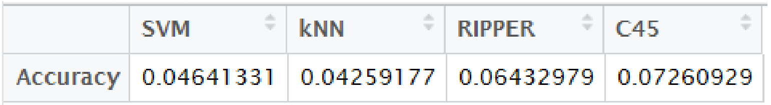
Total Number of Observations	223
Total Number of Feature Variables	4
Total Number of Class Labels	6

Accuracies (Average) for all Classifiers over the 5 iterations



	SVM	kNN	RIPPER	C45
Accuracy	0.5571429	0.552381	0.4666667	0.5380952

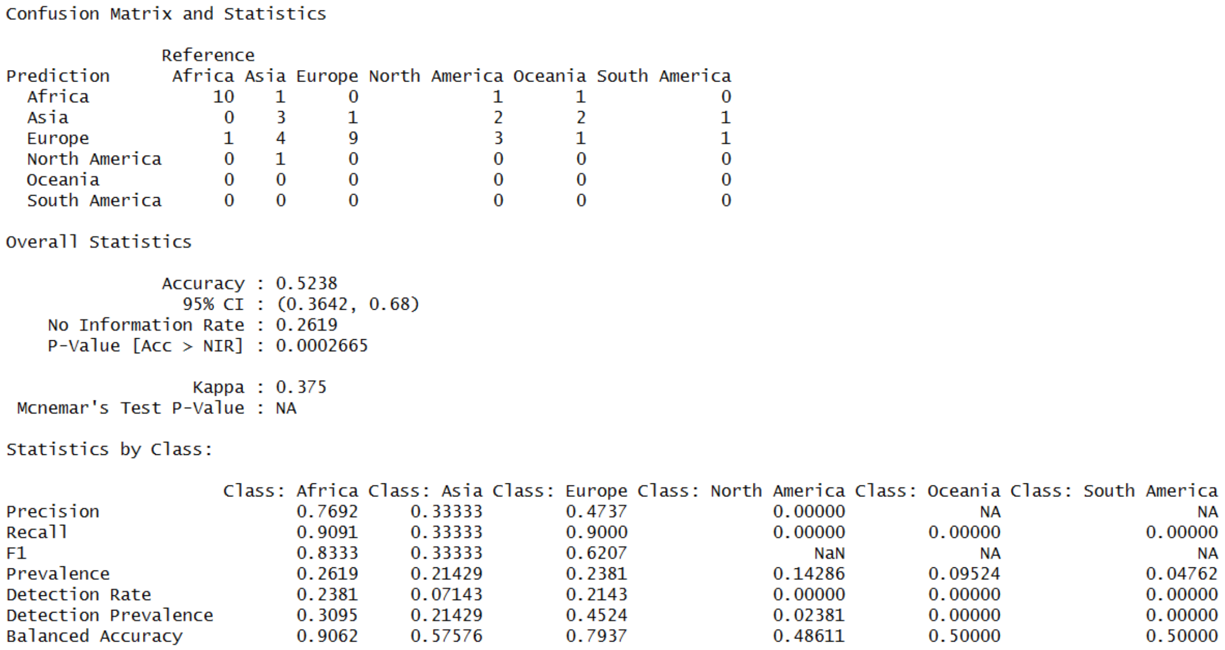
Accuracies (Standard Deviation) for all Classifiers over the 5 iterations



	SVM	kNN	RIPPER	C45
Accuracy	0.04641331	0.04259177	0.06432979	0.07260929

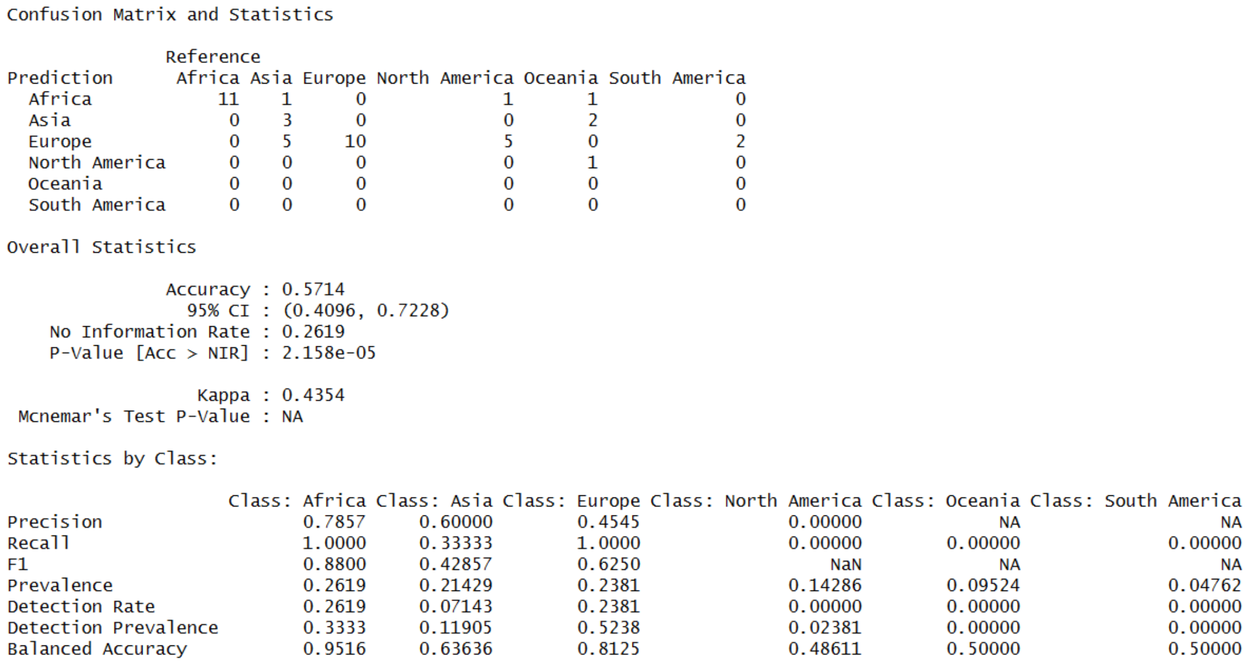
SVM Classification - confusion matrices and metrics

Iteration 1:



```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            10    1      0             1       1             0
  Asia               0    3      1             2       2             1
  Europe             1    4      9             3       1             1
  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 : 0.5238
                 95% CI : (0.3642, 0.68)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 0.0002665

                  Kappa : 0.375
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.7692     0.33333        0.4737              0.00000             NA                   NA
Recall                      0.9091     0.33333        0.9000              0.00000        0.00000              0.00000
F1                          0.8333     0.33333        0.6207                  NaN             NA                   NA
Prevalence                  0.2619     0.21429        0.2381              0.14286        0.09524              0.04762
Detection Rate              0.2381     0.07143        0.2143              0.00000        0.00000              0.00000
Detection Prevalence        0.3095     0.21429        0.4524              0.02381        0.00000              0.00000
Balanced Accuracy           0.9062     0.57576        0.7937              0.48611        0.50000              0.50000
```

Iteration 2:



```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            11    1      0             1       1             0
  Asia               0    3      0             0       2             0
  Europe             0    5     10             5       0             2
  North America      0    0      0             0       1             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5714
                 95% CI : (0.4096, 0.7228)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 2.158e-05

                  Kappa : 0.4354
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.7857     0.60000        0.4545              0.00000             NA                   NA
Recall                      1.0000     0.33333        1.0000              0.00000        0.00000              0.00000
F1                          0.8800     0.42857        0.6250                  NaN             NA                   NA
Prevalence                  0.2619     0.21429        0.2381              0.14286        0.09524              0.04762
Detection Rate              0.2619     0.07143        0.2381              0.00000        0.00000              0.00000
Detection Prevalence        0.3333     0.11905        0.5238              0.02381        0.00000              0.00000
Balanced Accuracy           0.9516     0.63636        0.8125              0.48611        0.50000              0.50000
```

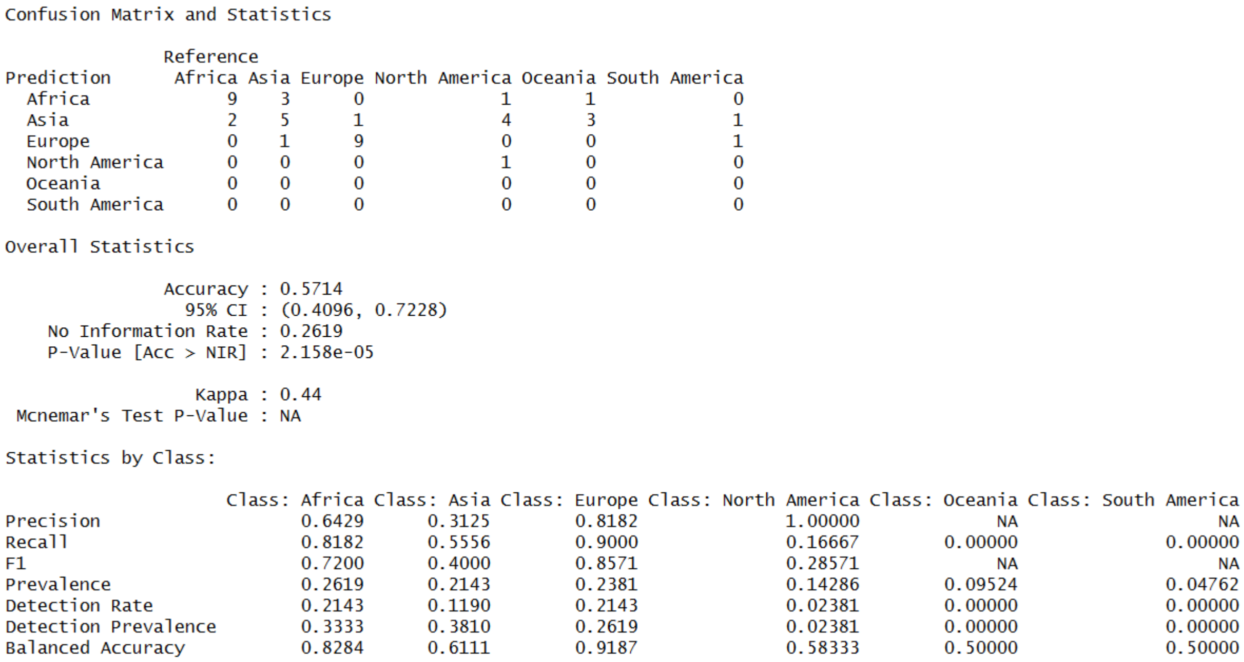
Iteration 3:

```
Confusion Matrix and Statistics

               Reference
Prediction       Africa Asia Europe North America Oceania South America
  Africa              9    1      0             0       2             0
  Asia                2    3      1             4       1             1
  Europe              0    4      9             2       1             1
  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 : 0.5             
                 95% CI : (0.3419, 0.6581)
    No Information Rate : 0.2619          
    P-Value [Acc > NIR] : 0.000819        
                                          
                  Kappa : 0.3457          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.7500     0.25000        0.5294              0.00000             NA                   NA
Recall                      0.8182     0.33333        0.9000              0.00000        0.00000              0.00000
F1                          0.7826     0.28571        0.6667                  NaN             NA                   NA
Prevalence                  0.2619     0.21429        0.2381              0.14286        0.09524              0.04762
Detection Rate              0.2143     0.07143        0.2143              0.00000        0.00000              0.00000
Detection Prevalence        0.2857     0.28571        0.4048              0.02381        0.00000              0.00000
Balanced Accuracy           0.8607     0.53030        0.8250              0.48611        0.50000              0.50000
```

Iteration 4:

```
Confusion Matrix and Statistics

               Reference
Prediction       Africa Asia Europe North America Oceania South America
  Africa             10    2      0             0       0             1
  Asia                0    6      1             4       2             0
  Europe              1    1      9             1       2             1
  North America       0    0      0             1       0             0
  Oceania             0    0      0             0       0             0
  South America       0    0      0             0       0             0

Overall Statistics
                                          
               Accuracy : 0.619           
                 95% CI : (0.4564, 0.7643)
    No Information Rate : 0.2619          
    P-Value [Acc > NIR] : 1.21e-06        
                                          
                  Kappa : 0.5015          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.7692      0.4615        0.6000              1.00000             NA                   NA
Recall                      0.9091      0.6667        0.9000              0.16667        0.00000              0.00000
F1                          0.8333      0.5455        0.7200              0.28571             NA                   NA
Prevalence                  0.2619      0.2143        0.2381              0.14286        0.09524              0.04762
Detection Rate              0.2381      0.1429        0.2143              0.02381        0.00000              0.00000
Detection Prevalence        0.3095      0.3095        0.3571              0.02381        0.00000              0.00000
Balanced Accuracy           0.9062      0.7273        0.8562              0.58333        0.50000              0.50000
```

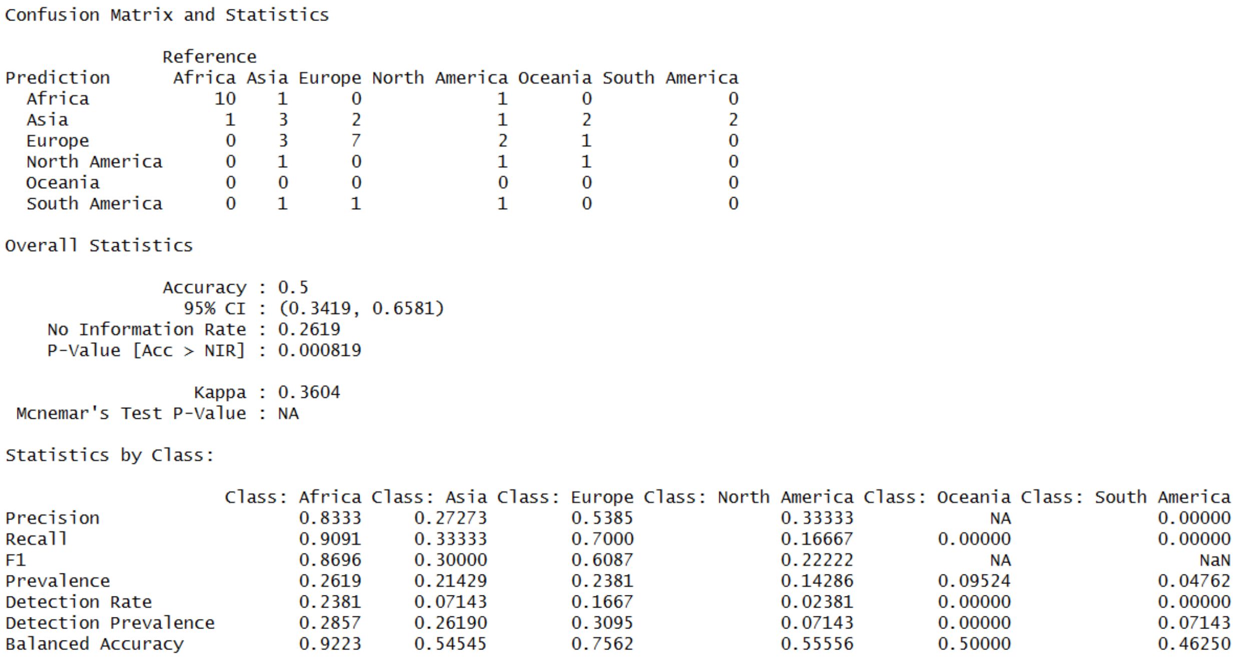
Iteration 5:



```
Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 0.5714
                 95% CI : (0.4096, 0.7228)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 2.158e-05

                  Kappa : 0.44
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.6429      0.3125        0.8182              1.00000             NA                   NA
Recall                      0.8182      0.5556        0.9000              0.16667        0.00000              0.00000
F1                          0.7200      0.4000        0.8571              0.28571             NA                   NA
Prevalence                  0.2619      0.2143        0.2381              0.14286        0.09524              0.04762
Detection Rate              0.2143      0.1190        0.2143              0.02381        0.00000              0.00000
Detection Prevalence        0.3333      0.3810        0.2619              0.02381        0.00000              0.00000
Balanced Accuracy           0.8284      0.6111        0.9187              0.58333        0.50000              0.50000
```

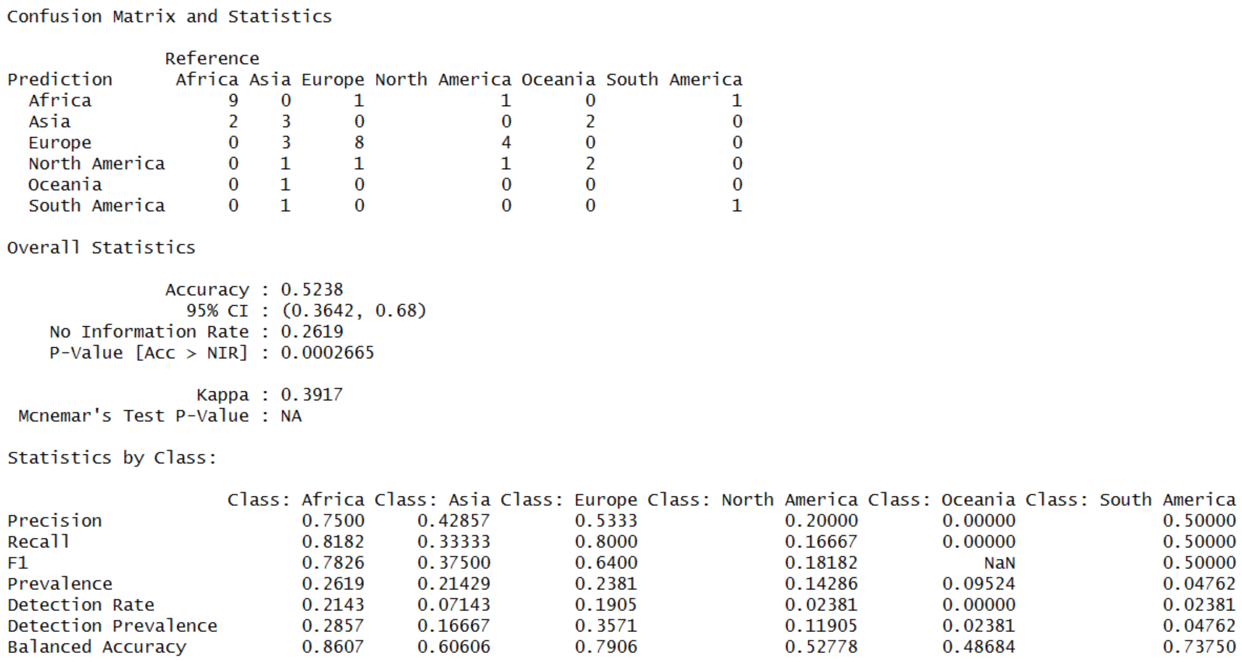
k-NN Classification - confusion matrices and metrics

Iteration 1:



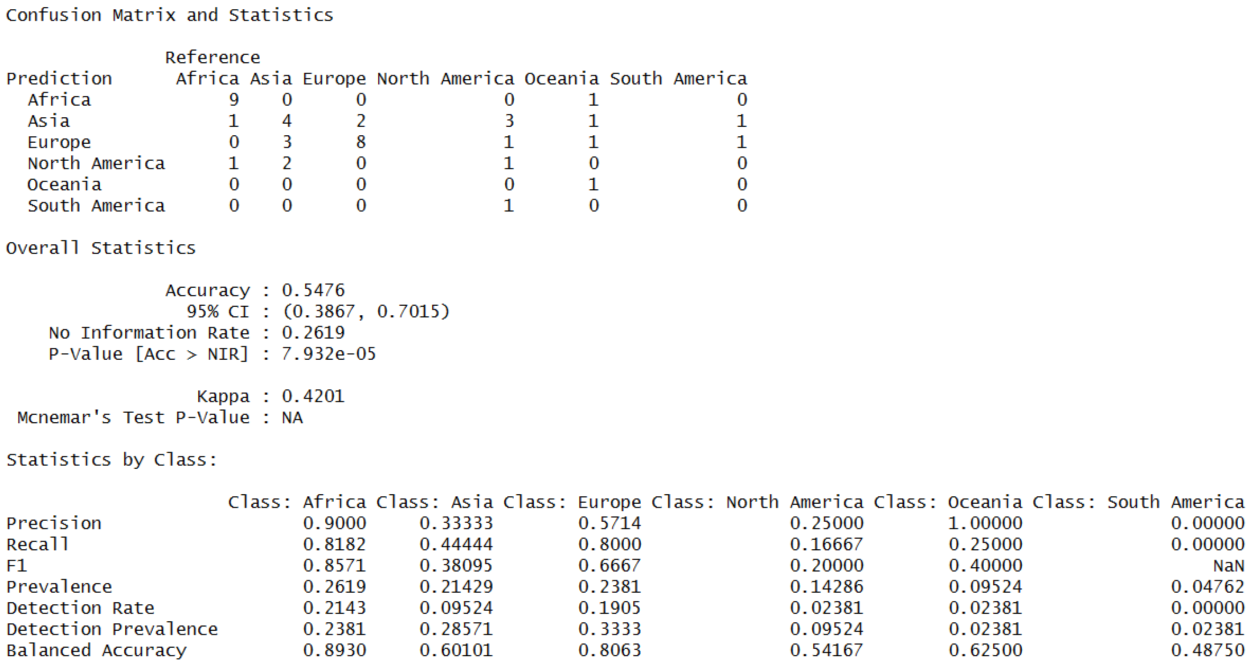
```
Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 0.5
                 95% CI : (0.3419, 0.6581)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 0.000819

                  Kappa : 0.3604
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.8333     0.27273        0.5385              0.33333             NA              0.00000
Recall                      0.9091     0.33333        0.7000              0.16667        0.00000              0.00000
F1                          0.8696     0.30000        0.6087              0.22222             NA                  NaN
Prevalence                  0.2619     0.21429        0.2381              0.14286        0.09524              0.04762
Detection Rate              0.2381     0.07143        0.1667              0.02381        0.00000              0.00000
Detection Prevalence        0.2857     0.26190        0.3095              0.07143        0.00000              0.07143
Balanced Accuracy           0.9223     0.54545        0.7562              0.55556        0.50000              0.46250
```

Iteration 2:



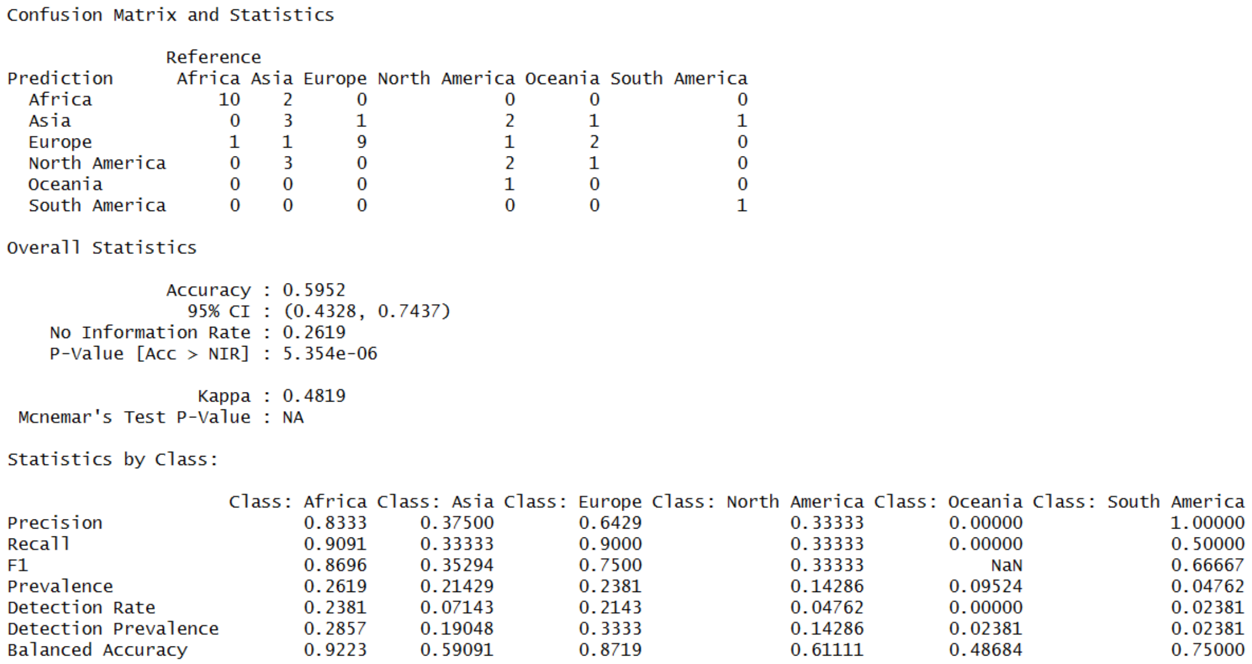
```
Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 0.5238
                 95% CI : (0.3642, 0.68)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 0.0002665

                  Kappa : 0.3917
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.7500     0.42857        0.5333              0.20000        0.00000              0.50000
Recall                      0.8182     0.33333        0.8000              0.16667        0.00000              0.50000
F1                          0.7826     0.37500        0.6400              0.18182            NaN              0.50000
Prevalence                  0.2619     0.21429        0.2381              0.14286        0.09524              0.04762
Detection Rate              0.2143     0.07143        0.1905              0.02381        0.00000              0.02381
Detection Prevalence        0.2857     0.16667        0.3571              0.11905        0.02381              0.04762
Balanced Accuracy           0.8607     0.60606        0.7906              0.52778        0.48684              0.73750
```

Iteration 3:



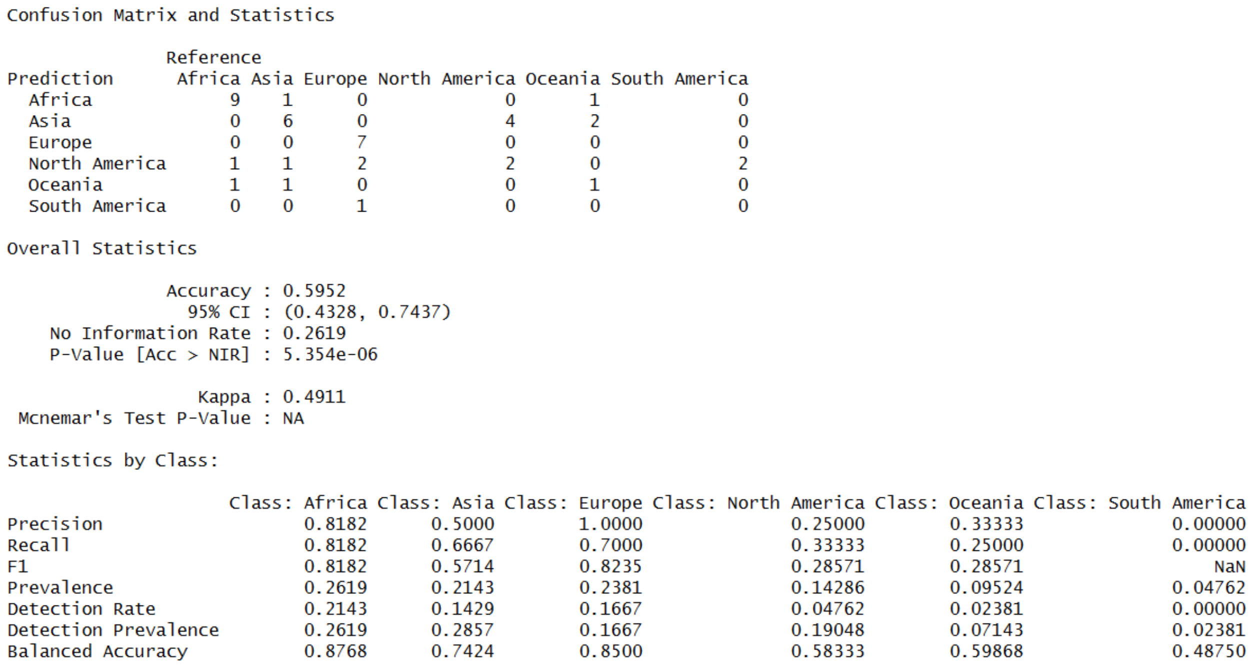
```
Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 0.5476
                 95% CI : (0.3867, 0.7015)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 7.932e-05

                  Kappa : 0.4201
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.9000     0.33333        0.5714              0.25000        1.00000              0.00000
Recall                      0.8182     0.44444        0.8000              0.16667        0.25000              0.00000
F1                          0.8571     0.38095        0.6667              0.20000        0.40000                  NaN
Prevalence                  0.2619     0.21429        0.2381              0.14286        0.09524              0.04762
Detection Rate              0.2143     0.09524        0.1905              0.02381        0.02381              0.00000
Detection Prevalence        0.2381     0.28571        0.3333              0.09524        0.02381              0.02381
Balanced Accuracy           0.8930     0.60101        0.8063              0.54167        0.62500              0.48750
```

Iteration 4:



```
Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 0.5952
                 95% CI : (0.4328, 0.7437)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 5.354e-06

                  Kappa : 0.4819
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.8333     0.37500        0.6429              0.33333        0.00000              1.00000
Recall                      0.9091     0.33333        0.9000              0.33333        0.00000              0.50000
F1                          0.8696     0.35294        0.7500              0.33333            NaN              0.66667
Prevalence                  0.2619     0.21429        0.2381              0.14286        0.09524              0.04762
Detection Rate              0.2381     0.07143        0.2143              0.04762        0.00000              0.02381
Detection Prevalence        0.2857     0.19048        0.3333              0.14286        0.02381              0.02381
Balanced Accuracy           0.9223     0.59091        0.8719              0.61111        0.48684              0.75000
```

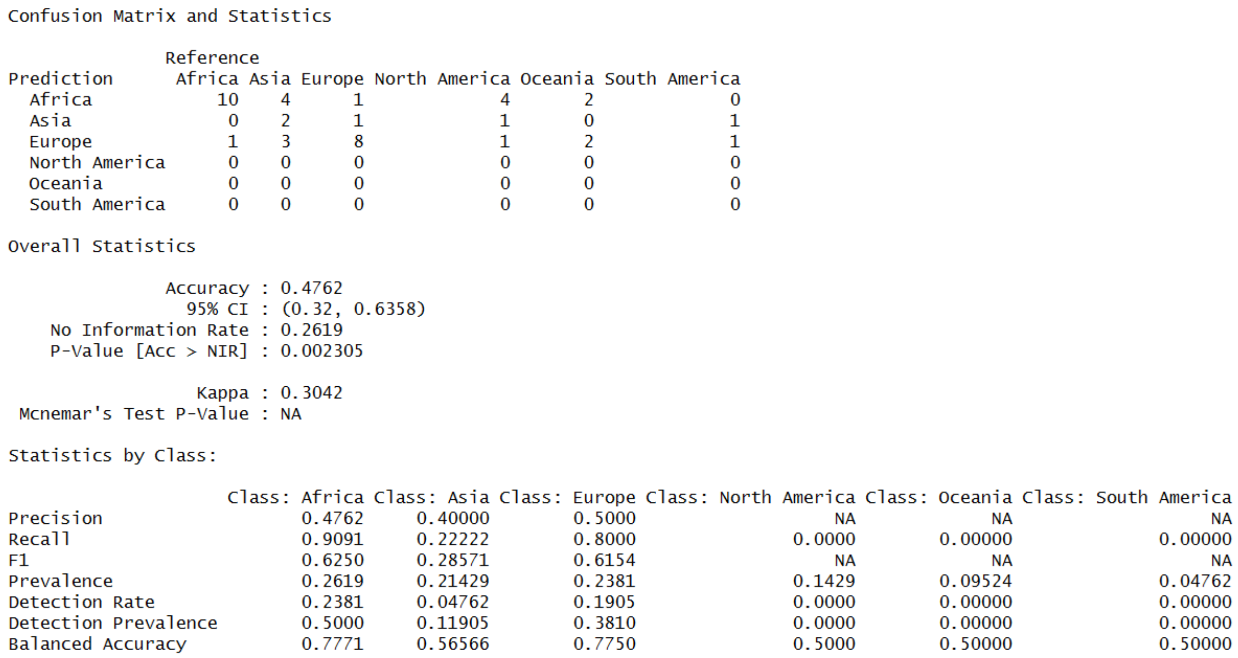
Iteration 5:



```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             9    1      0             0       1             0
  Asia               0    6      0             4       2             0
  Europe             0    0      7             0       0             0
  North America      1    1      2             2       0             2
  Oceania            1    1      0             0       1             0
  South America      0    0      1             0       0             0

Overall Statistics

               Accuracy : 0.5952
                 95% CI : (0.4328, 0.7437)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 5.354e-06

                  Kappa : 0.4911
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.8182      0.5000        1.0000              0.25000        0.33333              0.00000
Recall                      0.8182      0.6667        0.7000              0.33333        0.25000              0.00000
F1                          0.8182      0.5714        0.8235              0.28571        0.28571                  NaN
Prevalence                  0.2619      0.2143        0.2381              0.14286        0.09524              0.04762
Detection Rate              0.2143      0.1429        0.1667              0.04762        0.02381              0.00000
Detection Prevalence        0.2619      0.2857        0.1667              0.19048        0.07143              0.02381
Balanced Accuracy           0.8768      0.7424        0.8500              0.58333        0.59868              0.48750
```

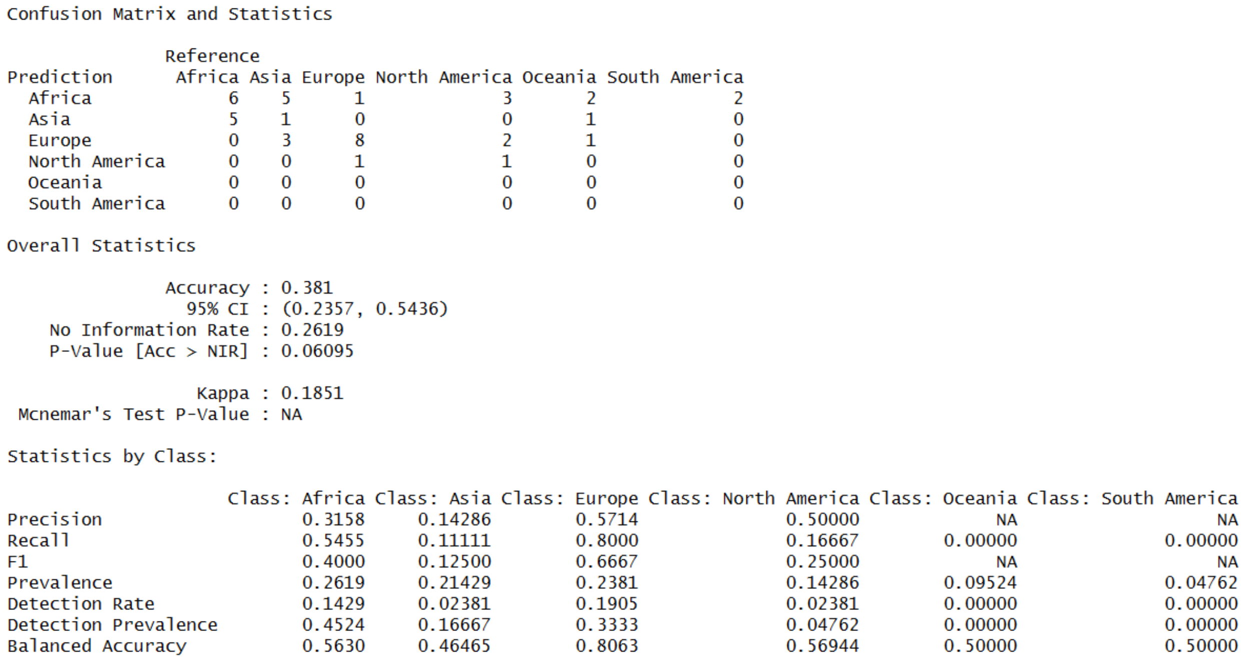
RIPPER Classification - confusion matrices and metrics

Iteration 1:



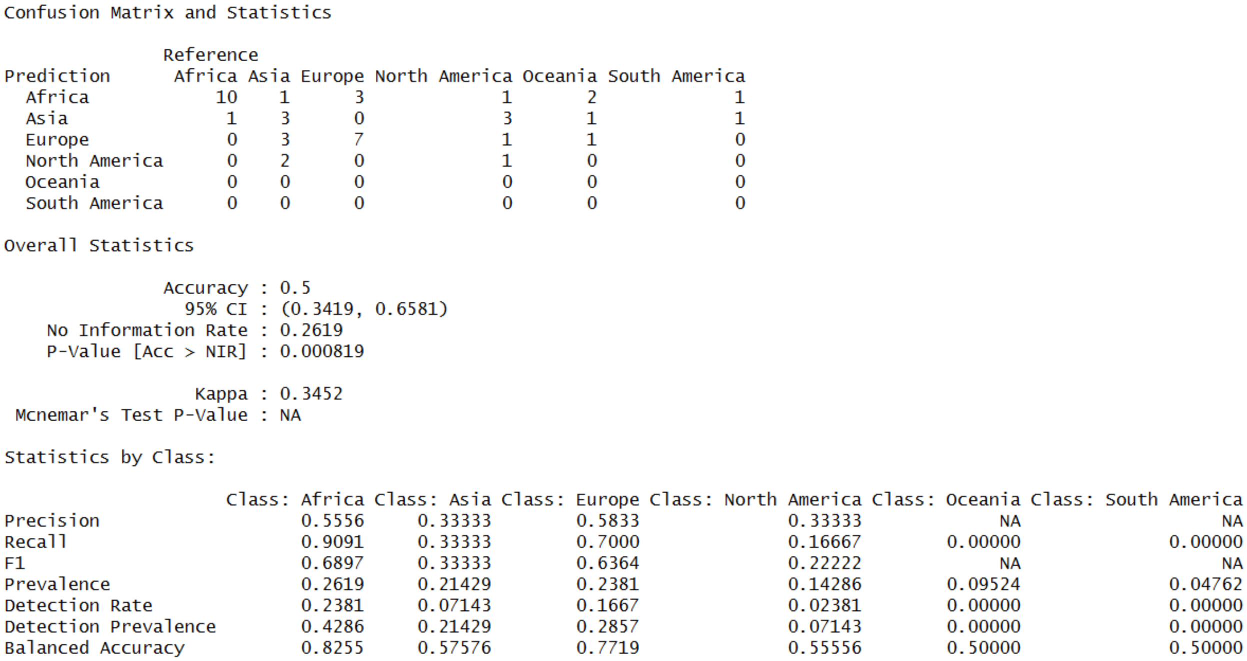
```
Confusion Matrix and Statistics

               Reference
Prediction       Africa Asia Europe North America Oceania South America
  Africa            10    4      1             4       2             0
  Asia               0    2      1             1       0             1
  Europe             1    3      8             1       2             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 : 0.4762
                 95% CI : (0.32, 0.6358)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 0.002305

                  Kappa : 0.3042
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.4762     0.40000        0.5000                   NA             NA                   NA
Recall                      0.9091     0.22222        0.8000               0.0000        0.00000              0.00000
F1                          0.6250     0.28571        0.6154                   NA             NA                   NA
Prevalence                  0.2619     0.21429        0.2381               0.1429        0.09524              0.04762
Detection Rate              0.2381     0.04762        0.1905               0.0000        0.00000              0.00000
Detection Prevalence        0.5000     0.11905        0.3810               0.0000        0.00000              0.00000
Balanced Accuracy           0.7771     0.56566        0.7750               0.5000        0.50000              0.50000
```

Iteration 2:



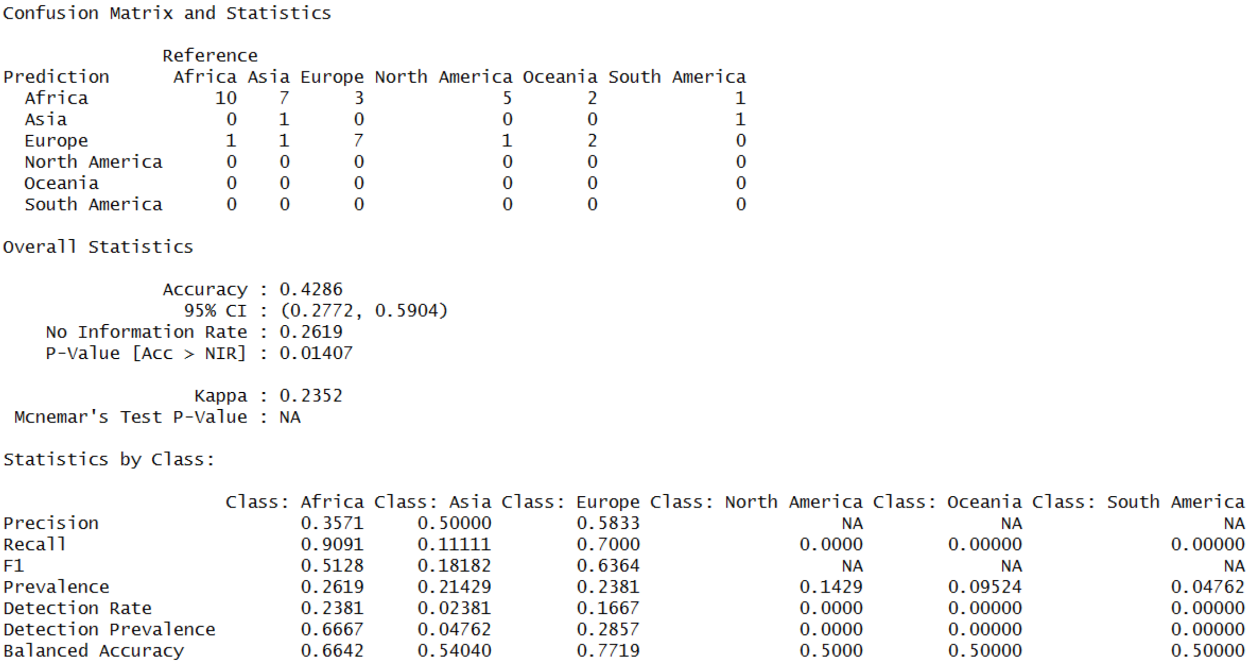
```
Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 0.381
                 95% CI : (0.2357, 0.5436)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 0.06095

                  Kappa : 0.1851
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.3158     0.14286        0.5714              0.50000             NA                   NA
Recall                      0.5455     0.11111        0.8000              0.16667        0.00000              0.00000
F1                          0.4000     0.12500        0.6667              0.25000             NA                   NA
Prevalence                  0.2619     0.21429        0.2381              0.14286        0.09524              0.04762
Detection Rate              0.1429     0.02381        0.1905              0.02381        0.00000              0.00000
Detection Prevalence        0.4524     0.16667        0.3333              0.04762        0.00000              0.00000
Balanced Accuracy           0.5630     0.46465        0.8063              0.56944        0.50000              0.50000
```

Iteration 3:



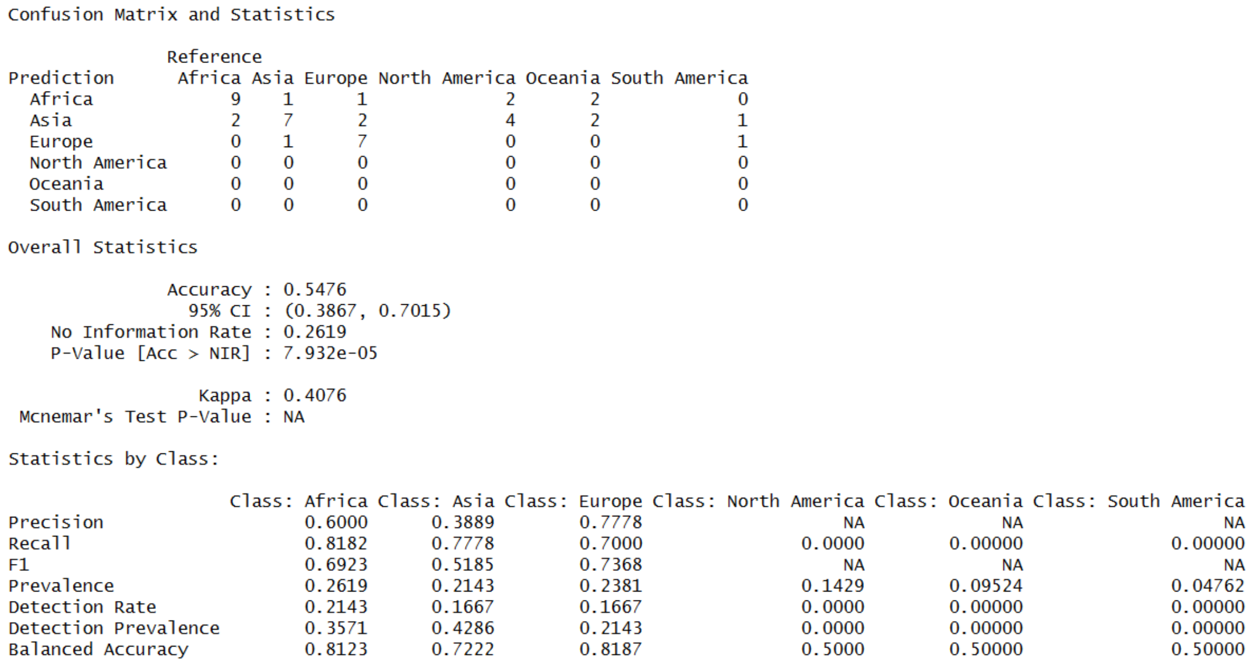
```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            10    1      3             1       2             1
  Asia               1    3      0             3       1             1
  Europe             0    3      7             1       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 : 0.5
                 95% CI : (0.3419, 0.6581)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 0.000819

                  Kappa : 0.3452
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.5556     0.33333        0.5833              0.33333             NA                   NA
Recall                      0.9091     0.33333        0.7000              0.16667        0.00000              0.00000
F1                          0.6897     0.33333        0.6364              0.22222             NA                   NA
Prevalence                  0.2619     0.21429        0.2381              0.14286        0.09524              0.04762
Detection Rate              0.2381     0.07143        0.1667              0.02381        0.00000              0.00000
Detection Prevalence        0.4286     0.21429        0.2857              0.07143        0.00000              0.00000
Balanced Accuracy           0.8255     0.57576        0.7719              0.55556        0.50000              0.50000
```

Iteration 4:



```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            10    7      3             5       2             1
  Asia               0    1      0             0       0             1
  Europe             1    1      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 : 0.4286
                 95% CI : (0.2772, 0.5904)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 0.01407

                  Kappa : 0.2352
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.3571     0.50000        0.5833                   NA             NA                   NA
Recall                      0.9091     0.11111        0.7000               0.0000        0.00000              0.00000
F1                          0.5128     0.18182        0.6364                   NA             NA                   NA
Prevalence                  0.2619     0.21429        0.2381               0.1429        0.09524              0.04762
Detection Rate              0.2381     0.02381        0.1667               0.0000        0.00000              0.00000
Detection Prevalence        0.6667     0.04762        0.2857               0.0000        0.00000              0.00000
Balanced Accuracy           0.6642     0.54040        0.7719               0.5000        0.50000              0.50000
```

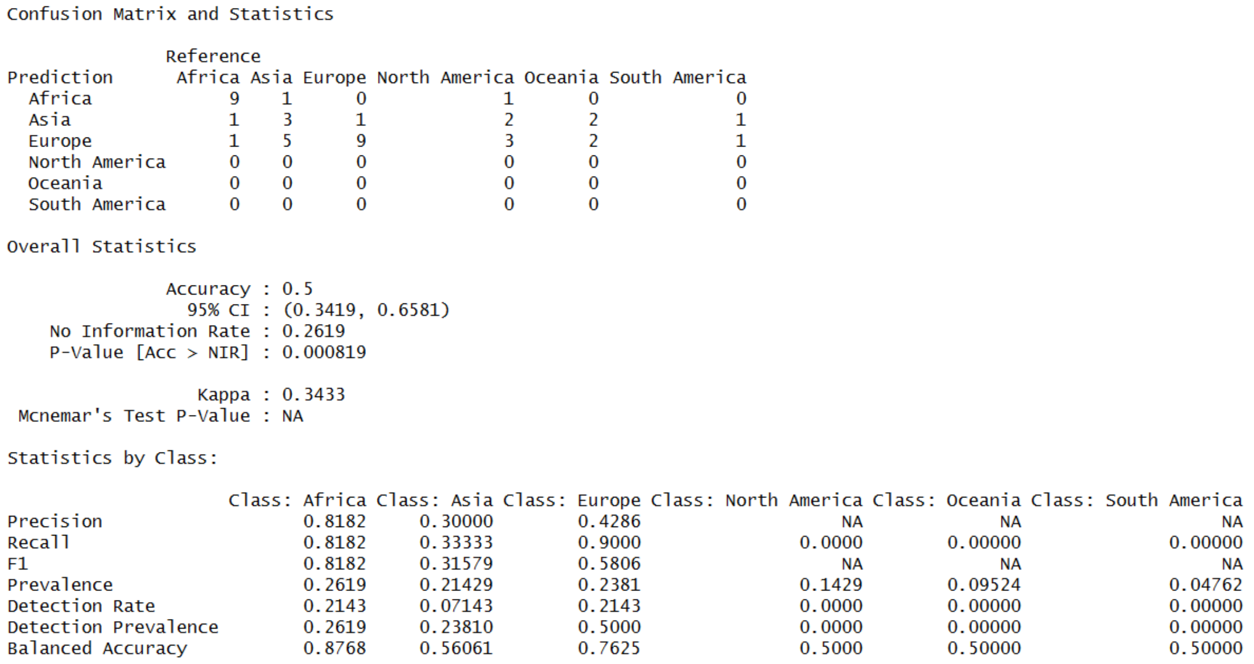
Iteration 5:



```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             9    1      1             2       2             0
  Asia               2    7      2             4       2             1
  Europe             0    1      7             0       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 : 0.5476
                 95% CI : (0.3867, 0.7015)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 7.932e-05

                  Kappa : 0.4076
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.6000      0.3889        0.7778                   NA             NA                   NA
Recall                      0.8182      0.7778        0.7000               0.0000        0.00000              0.00000
F1                          0.6923      0.5185        0.7368                   NA             NA                   NA
Prevalence                  0.2619      0.2143        0.2381               0.1429        0.09524              0.04762
Detection Rate              0.2143      0.1667        0.1667               0.0000        0.00000              0.00000
Detection Prevalence        0.3571      0.4286        0.2143               0.0000        0.00000              0.00000
Balanced Accuracy           0.8123      0.7222        0.8187               0.5000        0.50000              0.50000
```

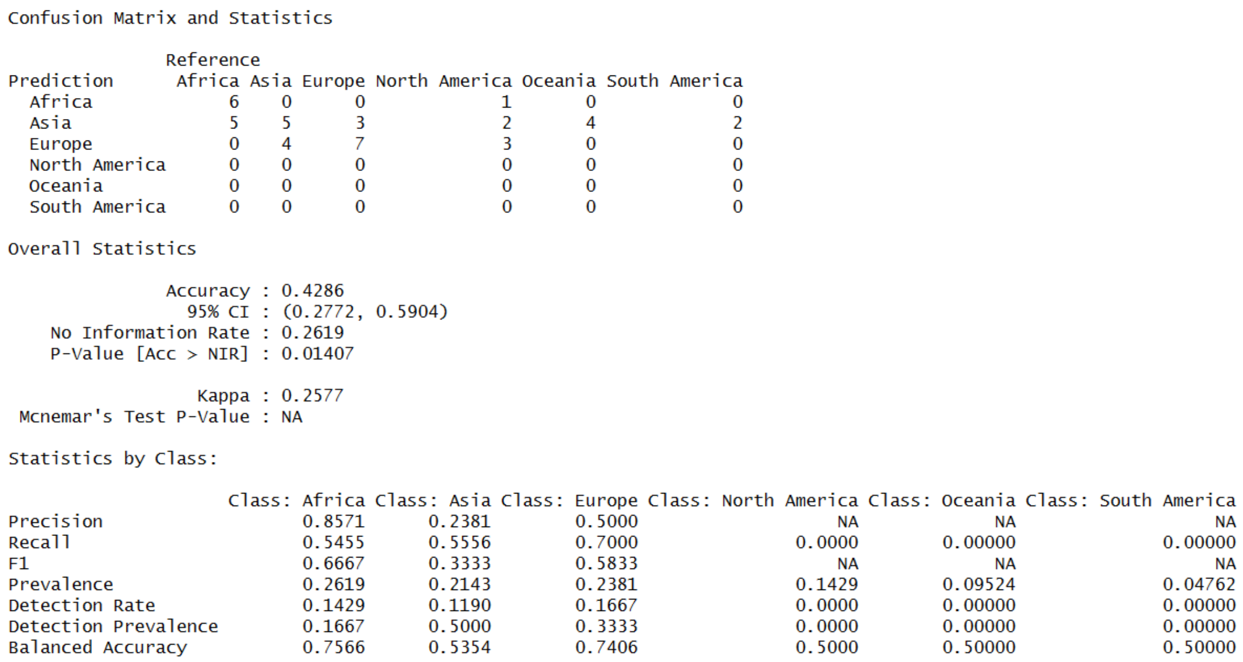
C4.5 Classification - confusion matrices and metrics

Iteration 1:



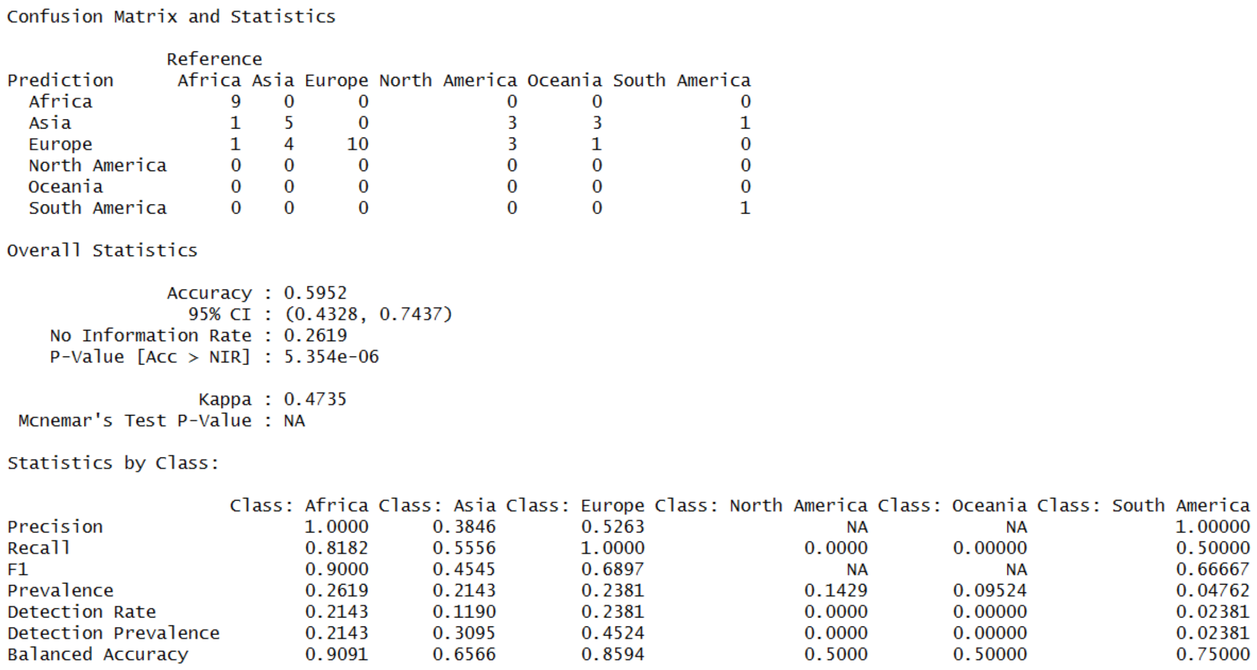
```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             9    1      0             1       0             0
  Asia               1    3      1             2       2             1
  Europe             1    5      9             3       2             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 : 0.5
                 95% CI : (0.3419, 0.6581)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 0.000819

                  Kappa : 0.3433
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.8182     0.30000        0.4286                   NA             NA                   NA
Recall                      0.8182     0.33333        0.9000               0.0000        0.00000              0.00000
F1                          0.8182     0.31579        0.5806                   NA             NA                   NA
Prevalence                  0.2619     0.21429        0.2381               0.1429        0.09524              0.04762
Detection Rate              0.2143     0.07143        0.2143               0.0000        0.00000              0.00000
Detection Prevalence        0.2619     0.23810        0.5000               0.0000        0.00000              0.00000
Balanced Accuracy           0.8768     0.56061        0.7625               0.5000        0.50000              0.50000
```

Iteration 2:



```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             6    0      0             1       0             0
  Asia               5    5      3             2       4             2
  Europe             0    4      7             3       0             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 : 0.4286
                 95% CI : (0.2772, 0.5904)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 0.01407

                  Kappa : 0.2577
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.8571      0.2381        0.5000                   NA             NA                   NA
Recall                      0.5455      0.5556        0.7000               0.0000        0.00000              0.00000
F1                          0.6667      0.3333        0.5833                   NA             NA                   NA
Prevalence                  0.2619      0.2143        0.2381               0.1429        0.09524              0.04762
Detection Rate              0.1429      0.1190        0.1667               0.0000        0.00000              0.00000
Detection Prevalence        0.1667      0.5000        0.3333               0.0000        0.00000              0.00000
Balanced Accuracy           0.7566      0.5354        0.7406               0.5000        0.50000              0.50000
```

Iteration 3:



```
Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 0.5952
                 95% CI : (0.4328, 0.7437)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 5.354e-06

                  Kappa : 0.4735
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   1.0000      0.3846        0.5263                   NA             NA              1.00000
Recall                      0.8182      0.5556        1.0000               0.0000        0.00000              0.50000
F1                          0.9000      0.4545        0.6897                   NA             NA              0.66667
Prevalence                  0.2619      0.2143        0.2381               0.1429        0.09524              0.04762
Detection Rate              0.2143      0.1190        0.2381               0.0000        0.00000              0.02381
Detection Prevalence        0.2143      0.3095        0.4524               0.0000        0.00000              0.02381
Balanced Accuracy           0.9091      0.6566        0.8594               0.5000        0.50000              0.75000
```

Iteration 4:

```
Confusion Matrix and Statistics

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

Overall Statistics
                                          
               Accuracy : 0.5952          
                 95% CI : (0.4328, 0.7437)
    No Information Rate : 0.2619          
    P-Value [Acc > NIR] : 5.354e-06       
                                          
                  Kappa : 0.4742          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.7692     0.57143        0.5263                   NA        0.00000               1.00000
Recall                      0.9091     0.44444        1.0000               0.0000        0.00000               0.50000
F1                          0.8333     0.50000        0.6897                   NA            NaN               0.66667
Prevalence                  0.2619     0.21429        0.2381               0.1429        0.09524               0.04762
Detection Rate              0.2381     0.09524        0.2381               0.0000        0.00000               0.02381
Detection Prevalence        0.3095     0.16667        0.4524               0.0000        0.04762               0.02381
Balanced Accuracy           0.9062     0.67677        0.8594               0.5000        0.47368               0.75000
```

Iteration 5:

```
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             9    1      0             0       0             0
  Asia               2    7      2             5       3             1
  Europe             0    1      8             1       1             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 : 0.5714          
                 95% CI : (0.4096, 0.7228)
    No Information Rate : 0.2619          
    P-Value [Acc > NIR] : 2.158e-05       
                                          
                  Kappa : 0.4417          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Precision                   0.9000      0.3500        0.6667                   NA             NA                    NA
Recall                      0.8182      0.7778        0.8000               0.0000        0.00000               0.00000
F1                          0.8571      0.4828        0.7273                   NA             NA                    NA
Prevalence                  0.2619      0.2143        0.2381               0.1429        0.09524               0.04762
Detection Rate              0.2143      0.1667        0.1905               0.0000        0.00000               0.00000
Detection Prevalence        0.2381      0.4762        0.2857               0.0000        0.00000               0.00000
Balanced Accuracy           0.8930      0.6919        0.8375               0.5000        0.50000               0.50000
```

CONCLUSION

The objectives of this project have been met. For our dataset, Support Vector Machines gave the best average accuracy among the 4 chosen classification algorithms in predicting Continent class labels. A couple of interesting observations made were:

- The division of the main dataset and selection of training data and test data has varying impact on the overall accuracies from various classification models.
- The control parameters for determining training model accuracies help tune the algorithm parameters, which ultimately determine the prediction model accuracies.
- For most classification algorithms in R’s caret package, greater values of tuneLength parameter lead to better model accuracy.

REFERENCES

[1] Wikipedia Website (section: List by the CIA (2016))

[2] http://www.investopedia.com/terms/d/datamining.asp

[3] https://en.wikipedia.org/wiki/Data_mining

[4] http://dataaspirant.com/2017/01/09/knn-implementation-r-using-caret-package/

[5] http://dataaspirant.com/2014/09/16/data-mining/

[6] http://www.statisticshowto.com/normalized/

[7] http://www.saedsayad.com/k_nearest_neighbors.htm

[8] https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification/JRip

[9] https://en.wikipedia.org/wiki/Support_vector_machine

[10] http://www.milanor.net/blog/read-excel-files-from-r/