IDM Project 1
Classification Report

Aniketh Sukhtankar (UF ID 7819 9584)

Introduction

The goal of the project is to increase familiarity with the classification packages, available in R to do data mining analysis on real-world problems. Several different classification methods were used on the given Life Expectancy dataset. The dataset was obtained from the Wikipedia website. The continent column was added as per the requirements to be used as class label. kNN, Support Vector Machine, C4.5 and RIPPER were the classification methods used on the data set. The following steps were run to get the desired results.

Dataset Preparation

Before you can start working on the classification algorithms, it is important to prepare your data. The following section will outline two ways in which you can do this: by normalizing your data (if necessary) and by splitting your data in training and testing sets. Before Normalization, Dimensionality reduction was performed on the data set in which irrelevant features like 'Country'(Entity) that contains no information that is useful in the identification of 'Continent' were excluded from dataset.

Normalization

As part of data preparation, we need to normalize our data so that its consistent. Normalization makes it easier for the different classification algorithms to learn. There are two types of normalization:

- Example normalization is the adjustment of each example individually, while
- Feature normalization indicates that you adjust each feature in the same way across all examples.

When we execute the summary () function and study the minimum and maximum values of all the (numerical) attributes, we see that the attributes have a wide range of values, which creates the need to normalize the Life Expectancy dataset, so that the classification is not dominated by any feature. Normalization adjusts the range of all features, so that distances between variables

with larger ranges are not over-emphasized. The normalization was performed by first making a normalize () function. This function was then used as an argument in another command, where I put the results of the normalization in a data frame through as.data.frame() after the function lapply() returned a list of the same length as the data set that was passed. Each element of that list was the result of the application of the normalize argument to the data set that served as input. For the Life Expectancy dataset, the normalize argument was applied on the four numerical attributes of the data set (Rank, Overall Life, Male Life, Female Life) and the results were placed in a data frame.

Training and Test Sets

To assess the model's performance, the data set was divided into two parts: a training set and a test set. The first was used to train the system, while the second is used to evaluate the learned or trained system. According to the requirements the data set was split into two disjoint sets where 80% of the original data set was used as the training set, while the 20% that remains composed the test set. The "runif" command is used to randomly assign either a 1 or a 0 and then use the assigned random numbers to divide the dataset into training and testing dataset.

Classification Methods Used

K-Nearest Neighbor(KNN)

Packages: caret

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor. Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN. The best k value in my case was obtained using a 10-fold cross-validation based search of k, repeated 3 times to obtain the trainControl. Accuracy was used to select the optimal model using the largest value. The final values used for the model was the k value with the largest accuracy among the ones tested. Prediction was done using the obtained fit and test data. Finally, the confusion matrix was obtained using the predicted model and test data class label values.

Support Vector Machine

Packages: e1071

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well. Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line). The e1071 package in R is used to create Support Vector Machines with ease. Tuning parameters value for machine learning algorithms effectively improves the model performance. Some important parameters having higher impact on model performance are "kernel", "gamma" and "C".

- Linear kernel is chosen if you have substantial number of features (>1000) because it is more likely that the data is linearly separable in high dimensional space. In our model we use the radial/RBF kernel and perform cross validation for its parameters as to avoid over-fitting.
- Gamma is the Kernel coefficient for 'rbf', 'poly' and 'sigmoid'. A higher value of gamma will try to exact fit the as per training data set i.e. generalization error and cause over-fitting problem.
- The Penalty parameter C of the error term controls the tradeoff between smooth decision boundary and classifying the training points correctly.

In my script I use tune.svm to get an effective combination of these parameters and avoid over-fitting. The kernel parameter is tuned to Radial. The gamma value is tuned by setting the "Gamma" parameter to best.parameters$gamma as returned by tune.svm. The C value is tuned by the "Cost" parameter in R to best.parameters$cost as returned by tune.svm. Prediction was done using the obtained fit and test data. Finally, the confusion matrix was obtained using the predicted model and test data class label values.

C4.5 Decision Tree

Packages: RWeka, caret

C4.5 algorithm is a classification algorithm producing decision tree based on information theory. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest

normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sublists. This algorithm has a few base cases. All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class. None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class. Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value. In the given script I used caret to create a 10-fold training set. Then, applied the J48 method which is an open source Java implementation of the C4.5 algorithm available in the RWeka data mining package using the 10-fold trainControl to obtain a model fit. Accuracy was used to select the optimal model using the largest value. The final values used for the model were the C and M values that returned the highest accuracy. Prediction was done using the obtained fit and test data. Finally, the confusion matrix was obtained using the predicted model and test data class label values.

Ripper Decision Tree

Packages: RWeka, caret

Repeated Incremental Pruning to Produce Error Reduction (RIPPER) is an optimized version of IREP. It 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 ts 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. Used a 10-fold cross-validation, repeated 3 times to obtain the trainControl. Then used JRip that implements a propositional rule learner, “Repeated Incremental Pruning to Produce Error Reduction” using the trainControl to obtain a model fit. Accuracy was used to select the optimal model using the largest value. The final values used for the model were the NumOpt, NumFolds and MinWeights values that returned the highest accuracy. Prediction was done using the obtained fit and test data. Finally, the confusion matrix was obtained using the predicted model and test data class label values.

Classification Results and Analysis

Support Vector Machine (SVM)

1. SAMPLE 1 (Seed: 1707)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model was cost = 16, gamma = 0.25.

```
 Accuracy
0.5405405
> precision
[1] 1.0000000 0.4117647 0.5384615 0.0000000       NaN       NaN
> recall
[1] 0.6000000 0.7777778 1.0000000 0.0000000 0.0000000 0.0000000
> fMeasure
[1] 0.7500000 0.5384615 0.7000000       NaN       NaN       NaN
```

Confusion Matrix and Statistics

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

Overall Statistics

Accuracy	0.5405
95% CI	(0.3692, 0.7051)
No Information Rate	0.2703
P-Value [Acc > NIR]	0.0004561
Kappa	0.4066
Mcnemar's Test P-Value	NA

Statistics by Class:

	Class: Africa	Class: Asia	Class: Europe	Class: North America	Class: Oceania	Class: South America
Sensitivity	0.6000	0.7778	1.0000	0.00000	0.0000	0.00000
Specificity	1.0000	0.6429	0.8000	0.96875	1.0000	1.00000
Pos Pred Value	1.0000	0.4118	0.5385	0.00000	NaN	NaN
Neg Pred Value	0.8710	0.9000	1.0000	0.86111	0.8919	0.94595
Prevalence	0.2703	0.2432	0.1892	0.13514	0.1081	0.05405
Detection Rate	0.1622	0.1892	0.1892	0.00000	0.0000	0.00000
Detection Prevalence	0.1622	0.4595	0.3514	0.02703	0.0000	0.00000
Balanced Accuracy	0.8000	0.7103	0.9000	0.48438	0.5000	0.50000

2. SAMPLE 2 (Seed: 1234)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model was cost = 16, gamma = 0.25.

```
> accuracy
 Accuracy
0.6428571
> precision
[1] 1.0000000 0.4705882 0.5833333 0.5000000       NaN       NaN
> recall
[1] 0.7857143 0.8888889 0.8750000 0.2500000 0.0000000 0.0000000
> fMeasure
[1] 0.8800000 0.6153846 0.7000000 0.3333333       NaN       NaN
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            11    0      0             0       0             0
  Asia               2    8      1             3       3             0
  Europe             1    1      7             0       2             1
  North America      0    0      0             1       1             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.6429
                 95% CI : (0.4803, 0.7845)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 3.989e-05

                  Kappa : 0.5344
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.7857      0.8889        0.8750              0.25000         0.0000              0.00000
Specificity                 1.0000      0.7273        0.8529              0.97368         1.0000              1.00000
Pos Pred Value              1.0000      0.4706        0.5833              0.50000            NaN                  NaN
Neg Pred Value              0.9032      0.9600        0.9667              0.92500         0.8571              0.97619
Prevalence                  0.3333      0.2143        0.1905              0.09524         0.1429              0.02381
Detection Rate              0.2619      0.1905        0.1667              0.02381         0.0000              0.00000
Detection Prevalence        0.2619      0.4048        0.2857              0.04762         0.0000              0.00000
Balanced Accuracy           0.8929      0.8081        0.8640              0.61184         0.5000              0.50000
```

3. SAMPLE 3 (Seed: 1111)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model was cost = 16, gamma = 0.25.

```
> accuracy
Accuracy
   0.625
> precision
[1] 0.9166667 0.5000000 0.5238095 0.6666667       NaN       NaN
> recall
[1] 1.0000000 0.5454545 0.8461538 0.1818182 0.0000000 0.0000000
> fMeasure
[1] 0.9565217 0.5217391 0.6470588 0.2857143       NaN       NaN
> cf
Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 0.625
                 95% CI : (0.4735, 0.7605)
    No Information Rate : 0.2708
    P-Value [Acc > NIR] : 3.011e-07

                  Kappa : 0.5017
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 1.0000      0.5455        0.8462              0.18182        0.00000              0.00000
Specificity                 0.9730      0.8378        0.7143              0.97297        1.00000              1.00000
Pos Pred Value              0.9167      0.5000        0.5238              0.66667            NaN                  NaN
Neg Pred Value              1.0000      0.8611        0.9259              0.80000        0.97917              0.97917
Prevalence                  0.2292      0.2292        0.2708              0.22917        0.02083              0.02083
Detection Rate              0.2292      0.1250        0.2292              0.04167        0.00000              0.00000
Detection Prevalence        0.2500      0.2500        0.4375              0.06250        0.00000              0.00000
Balanced Accuracy           0.9865      0.6916        0.7802              0.57740        0.50000              0.50000
```

4. SAMPLE 4 (Seed: 2222)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model was cost = 16, gamma = 0.25.

```
> accuracy
Accuracy
0.547619
> precision
[1] 1.0000000 0.2307692 0.6250000 0.0000000       NaN       NaN
> recall
[1] 0.9090909 0.5000000 0.9090909 0.0000000 0.0000000 0.0000000
> fMeasure
[1] 0.9523810 0.3157895 0.7407407       NaN       NaN       NaN
> cf
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            10    0      0             0       0             0
  Asia               1    3      1             2       4             2
  Europe             0    1     10             2       2             1
  North America      0    2      0             0       0             1
  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 : 0.4251
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.9091     0.50000        0.9091              0.00000         0.0000              0.00000
Specificity                 1.0000     0.72222        0.8065              0.92105         1.0000              1.00000
Pos Pred Value              1.0000     0.23077        0.6250              0.00000            NaN                  NaN
Neg Pred Value              0.9688     0.89655        0.9615              0.89744         0.8571              0.90476
Prevalence                  0.2619     0.14286        0.2619              0.09524         0.1429              0.09524
Detection Rate              0.2381     0.07143        0.2381              0.00000         0.0000              0.00000
Detection Prevalence        0.2381     0.30952        0.3810              0.07143         0.0000              0.00000
Balanced Accuracy           0.9545     0.61111        0.8578              0.46053         0.5000              0.50000
```

5. SAMPLE 5 (Seed: 3333)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model was cost = 4, gamma = 0.25.

```
> accuracy
 Accuracy
0.5357143
> precision
[1] 0.8666667 0.4285714 0.4705882 0.0000000       NaN       NaN
> recall
[1] 0.9285714 0.6923077 0.7272727 0.0000000 0.0000000 0.0000000
> fMeasure
[1] 0.8965517 0.5294118 0.5714286       NaN       NaN       NaN
> cf
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            13    2      0             0       0             0
  Asia               0    9      2             5       3             2
  Europe             0    1      8             6       1             1
  North America      1    1      1             0       0             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5357
                 95% CI : (0.3974, 0.6701)
    No Information Rate : 0.25
    P-Value [Acc > NIR] : 4.474e-06

                  Kappa : 0.4016
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North America Class: Oceania Class: South America
Sensitivity                 0.9286      0.6923        0.7273                0.00000       0.00000                0.00000
Specificity                 0.9524      0.7209        0.8000                0.93333       1.00000                1.00000
Pos Pred Value              0.8667      0.4286        0.4706                0.00000           NaN                    NaN
Neg Pred Value              0.9756      0.8857        0.9231                0.79245       0.92857                0.94643
Prevalence                  0.2500      0.2321        0.1964                0.19643       0.07143                0.05357
Detection Rate              0.2321      0.1607        0.1429                0.00000       0.00000                0.00000
Detection Prevalence        0.2679      0.3750        0.3036                0.05357       0.00000                0.00000
Balanced Accuracy           0.9405      0.7066        0.7636                0.46667       0.50000                0.50000
```

K Nearest Neighbor (KNN)

1. SAMPLE 1 (Seed: 1707)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model was k = 23.

```
> accuracy
 Accuracy
0.5135135
> precision
[1] 1.0000000 0.3888889 0.6250000 0.5000000       NaN 0.0000000
> recall
[1] 0.6000000 0.7777778 0.7142857 0.2000000 0.0000000 0.0000000
> fMeasure
[1] 0.7500000 0.5185185 0.6666667 0.2857143       NaN       NaN
```

Confusion Matrix and Statistics

	Reference					
Prediction	Africa	Asia	Europe	North.America	Oceania	South.America
Africa	6	0	0	0	0	0
Asia	2	7	1	4	2	2
Europe	1	0	5	0	2	0
North.America	0	1	0	1	0	0
Oceania	0	0	0	0	0	0
South.America	1	1	1	0	0	0

Overall Statistics

Accuracy	0.5135
95% CI	(0.344, 0.6808)
No Information Rate	0.2703
P-Value [Acc > NIR]	0.00143
Kappa	0.3805
Mcnemar's Test P-Value	NA

Statistics by Class:

	Class: Africa	Class: Asia	Class: Europe	Class: North.America	Class: Oceania	Class: South.America
Sensitivity	0.6000	0.7778	0.7143	0.20000	0.0000	0.00000
Specificity	1.0000	0.6071	0.9000	0.96875	1.0000	0.91429
Pos Pred Value	1.0000	0.3889	0.6250	0.50000	NaN	0.00000
Neg Pred Value	0.8710	0.8947	0.9310	0.88571	0.8919	0.94118
Prevalence	0.2703	0.2432	0.1892	0.13514	0.1081	0.05405
Detection Rate	0.1622	0.1892	0.1351	0.02703	0.0000	0.00000
Detection Prevalence	0.1622	0.4865	0.2162	0.05405	0.0000	0.08108
Balanced Accuracy	0.8000	0.6925	0.8071	0.58437	0.5000	0.45714

2. SAMPLE 2 (Seed: 1234)

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 9.

```
> accuracy
 Accuracy
0.6190476
> precision
[1] 1.0000000 0.6250000 0.6666667 0.2000000 0.5000000 0.5000000
> recall
[1] 0.7857143 0.5555556 0.7500000 0.5000000 0.1666667 1.0000000
> fMeasure
[1] 0.8800000 0.5882353 0.7058824 0.2857143 0.2500000 0.6666667
```

Confusion Matrix and Statistics

Prediction	Reference: Africa	Asia	Europe	North.America	Oceania	South.America
Africa	11	0	0	0	0	0
Asia	0	5	0	2	1	0
Europe	1	0	6	0	2	0
North.America	2	3	2	2	1	0
Oceania	0	1	0	0	1	0
South.America	0	0	0	0	1	1

Overall Statistics

Accuracy	0.619
95% CI	(0.4564, 0.7643)
No Information Rate	0.3333
P-Value [Acc > NIR]	0.0001396
Kappa	0.5241
Mcnemar's Test P-Value	NA

Statistics by Class:

	Class: Africa	Class: Asia	Class: Europe	Class: North.America	Class: Oceania	Class: South.America
Sensitivity	0.7857	0.5556	0.7500	0.50000	0.16667	1.00000
Specificity	1.0000	0.9091	0.9118	0.78947	0.97222	0.97561
Pos Pred Value	1.0000	0.6250	0.6667	0.20000	0.50000	0.50000
Neg Pred Value	0.9032	0.8824	0.9394	0.93750	0.87500	1.00000
Prevalence	0.3333	0.2143	0.1905	0.09524	0.14286	0.02381
Detection Rate	0.2619	0.1190	0.1429	0.04762	0.02381	0.02381
Detection Prevalence	0.2619	0.1905	0.2143	0.23810	0.04762	0.04762
Balanced Accuracy	0.8929	0.7323	0.8309	0.64474	0.56944	0.98780

3. SAMPLE 3 (Seed: 1111)

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 15.

```
> accuracy
 Accuracy
0.5416667
> precision
[1] 0.8461538 0.5000000 0.5833333 0.3750000 0.0000000 0.3333333
> recall
[1] 1.0000000 0.3636364 0.5384615 0.2727273 0.0000000 1.0000000
> fMeasure
[1] 0.9166667 0.4210526 0.5600000 0.3157895       NaN 0.5000000
> cf1
Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 0.5417
                 95% CI : (0.3917, 0.6863)
    No Information Rate : 0.2708
    P-Value [Acc > NIR] : 6.618e-05

                  Kappa : 0.4204
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North.America Class: Oceania Class: South.America
Sensitivity                 1.0000     0.36364        0.5385               0.2727        0.00000              1.00000
Specificity                 0.9459     0.89189        0.8571               0.8649        0.91489              0.95745
Pos Pred Value              0.8462     0.50000        0.5833               0.3750        0.00000              0.33333
Neg Pred Value              1.0000     0.82500        0.8333               0.8000        0.97727              1.00000
Prevalence                  0.2292     0.22917        0.2708               0.2292        0.02083              0.02083
Detection Rate              0.2292     0.08333        0.1458               0.0625        0.00000              0.02083
Detection Prevalence        0.2708     0.16667        0.2500               0.1667        0.08333              0.06250
Balanced Accuracy           0.9730     0.62776        0.6978               0.5688        0.45745              0.97872
```

4. SAMPLE 4 (Seed: 2222)

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 23.

```
> accuracy
 Accuracy
0.5714286
> precision
[1] 0.8333333 0.3750000 0.6250000 0.1666667       NaN       NaN
> recall
[1] 0.9090909 0.5000000 0.9090909 0.2500000 0.0000000 0.0000000
> fMeasure
[1] 0.8695652 0.4285714 0.7407407 0.2000000       NaN       NaN
> cf1
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North.America Oceania South.America
  Africa            10    0      0             0       2             0
  Asia               0    3      1             2       1             1
  Europe             0    1     10             1       2             2
  North.America      1    2      0             1       1             1
  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 : 0.4538
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North.America Class: Oceania Class: South.America
Sensitivity                 0.9091     0.50000        0.9091              0.25000         0.0000              0.00000
Specificity                 0.9355     0.86111        0.8065              0.86842         1.0000              1.00000
Pos Pred Value              0.8333     0.37500        0.6250              0.16667            NaN                  NaN
Neg Pred Value              0.9667     0.91176        0.9615              0.91667         0.8571              0.90476
Prevalence                  0.2619     0.14286        0.2619              0.09524         0.1429              0.09524
Detection Rate              0.2381     0.07143        0.2381              0.02381         0.0000              0.00000
Detection Prevalence        0.2857     0.19048        0.3810              0.14286         0.0000              0.00000
Balanced Accuracy           0.9223     0.68056        0.8578              0.55921         0.5000              0.50000
```

5. SAMPLE 5 (Seed: 3333)

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.

```
> accuracy
 Accuracy
0.5178571
> precision
[1] 0.8571429 0.5333333 0.4666667 0.1666667 0.0000000 0.3333333
> recall
[1] 0.85714286 0.61538462 0.63636364 0.09090909 0.00000000 0.33333333
> fMeasure
[1] 0.8571429 0.5714286 0.5384615 0.1176471       NaN 0.3333333
> cf1
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North.America Oceania South.America
  Africa            12    2      0             0       0             0
  Asia               0    8      0             3       3             1
  Europe             0    2      7             5       1             0
  North.America      0    1      3             1       0             1
  Oceania            2    0      0             1       0             0
  South.America      0    0      1             1       0             1

Overall Statistics

               Accuracy : 0.5179
                 95% CI : (0.3803, 0.6534)
    No Information Rate : 0.25
    P-Value [Acc > NIR] : 1.532e-05

                  Kappa : 0.3935
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North.America Class: Oceania Class: South.America
Sensitivity                 0.8571      0.6154        0.6364               0.09091        0.00000               0.33333
Specificity                 0.9524      0.8372        0.8222               0.88889        0.94231               0.96226
Pos Pred Value              0.8571      0.5333        0.4667               0.16667        0.00000               0.33333
Neg Pred Value              0.9524      0.8780        0.9024               0.80000        0.92453               0.96226
Prevalence                  0.2500      0.2321        0.1964               0.19643        0.07143               0.05357
Detection Rate              0.2143      0.1429        0.1250               0.01786        0.00000               0.01786
Detection Prevalence        0.2500      0.2679        0.2679               0.10714        0.05357               0.05357
Balanced Accuracy           0.9048      0.7263        0.7293               0.48990        0.47115               0.64780
```

Ripper Decision Tree

1. SAMPLE 1 (Seed: 1707)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were NumOpt = 5, NumFolds = 6 and MinWeights = 4.

```
> accuracy
 Accuracy
0.4864865
> precision
[1] 0.6666667 0.4285714 0.4545455 0.3333333       NaN       NaN
> recall
[1] 0.6000000 0.6666667 0.7142857 0.2000000 0.0000000 0.0000000
> fMeasure
[1] 0.6315789 0.5217391 0.5555556 0.2500000       NaN       NaN
```

```
Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 0.4865
                 95% CI : (0.3192, 0.656)
    No Information Rate : 0.2703
    P-Value [Acc > NIR] : 0.004058

                  Kappa : 0.3374
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North.America Class: Oceania Class: South.America
Sensitivity                 0.6000      0.6667        0.7143              0.20000         0.0000              0.00000
Specificity                 0.8889      0.7143        0.8000              0.93750         1.0000              1.00000
Pos Pred Value              0.6667      0.4286        0.4545              0.33333            NaN                  NaN
Neg Pred Value              0.8571      0.8696        0.9231              0.88235         0.8919              0.94595
Prevalence                  0.2703      0.2432        0.1892              0.13514         0.1081              0.05405
Detection Rate              0.1622      0.1622        0.1351              0.02703         0.0000              0.00000
Detection Prevalence        0.2432      0.3784        0.2973              0.08108         0.0000              0.00000
Balanced Accuracy           0.7444      0.6905        0.7571              0.56875         0.5000              0.50000
```

2. SAMPLE 2 (Seed: 1234)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were NumOpt = 3, NumFolds = 3 and MinWeights = 2.

```
> accuracy
 Accuracy
0.4761905
> precision
[1] 1.0000000 0.6666667 0.2592593       NaN 0.0000000       NaN
> recall
[1] 0.7857143 0.2222222 0.8750000 0.0000000 0.0000000 0.0000000
> fMeasure
[1] 0.8800000 0.3333333 0.4000000       NaN       NaN       NaN
```

Confusion Matrix and Statistics

Prediction	Reference: Africa	Asia	Europe	North.America	Oceania	South.America
Africa	11	0	0	0	0	0
Asia	0	2	1	0	0	0
Europe	3	6	7	4	6	1
North.America	0	0	0	0	0	0
Oceania	0	1	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 : 0.3333
    P-Value [Acc > NIR] : 0.03837

                  Kappa : 0.3211
 Mcnemar's Test P-Value : NA
```

Statistics by Class:

	Class: Africa	Class: Asia	Class: Europe	Class: North.America	Class: Oceania	Class: South.America
Sensitivity	0.7857	0.22222	0.8750	0.00000	0.00000	0.00000
Specificity	1.0000	0.96970	0.4118	1.00000	0.97222	1.00000
Pos Pred Value	1.0000	0.66667	0.2593	NaN	0.00000	NaN
Neg Pred Value	0.9032	0.82051	0.9333	0.90476	0.85366	0.97619
Prevalence	0.3333	0.21429	0.1905	0.09524	0.14286	0.02381
Detection Rate	0.2619	0.04762	0.1667	0.00000	0.00000	0.00000
Detection Prevalence	0.2619	0.07143	0.6429	0.00000	0.02381	0.00000
Balanced Accuracy	0.8929	0.59596	0.6434	0.50000	0.48611	0.50000

3. SAMPLE 3 (Seed: 1111)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were NumOpt = 4, NumFolds = 3 and MinWeights = 5.

```
> accuracy
Accuracy
     0.5
> precision
[1] 0.4782609 0.5714286 0.6000000       NaN 0.0000000       NaN
> recall
[1] 1.0000000 0.3636364 0.6923077 0.0000000 0.0000000 0.0000000
> fMeasure
[1] 0.6470588 0.4444444 0.6428571       NaN       NaN       NaN
> cf2
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North.America Oceania South.America
  Africa            11    3      3             4       1             1
  Asia               0    4      1             2       0             0
  Europe             0    2      9             4       0             0
  North.America      0    0      0             0       0             0
  Oceania            0    2      0             1       0             0
  South.America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.5
                 95% CI : (0.3523, 0.6477)
    No Information Rate : 0.2708
    P-Value [Acc > NIR] : 0.0006077

                  Kappa : 0.3514
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North.America Class: Oceania Class: South.America
Sensitivity                 1.0000     0.36364        0.6923               0.0000        0.00000               0.00000
Specificity                 0.6757     0.91892        0.8286               1.0000        0.93617               1.00000
Pos Pred Value              0.4783     0.57143        0.6000                  NaN        0.00000                   NaN
Neg Pred Value              1.0000     0.82927        0.8788               0.7708        0.97778               0.97917
Prevalence                  0.2292     0.22917        0.2708               0.2292        0.02083               0.02083
Detection Rate              0.2292     0.08333        0.1875               0.0000        0.00000               0.00000
Detection Prevalence        0.4792     0.14583        0.3125               0.0000        0.06250               0.00000
Balanced Accuracy           0.8378     0.64128        0.7604               0.5000        0.46809               0.50000
```

4. SAMPLE 4 (Seed: 2222)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were NumOpt = 4, NumFolds = 2 and MinWeights = 4.

```
> accuracy
 Accuracy
0.5714286
> precision
[1] 0.5789474 0.4285714 0.6923077 0.3333333       NaN       NaN
> recall
[1] 1.0000000 0.5000000 0.8181818 0.2500000 0.0000000 0.0000000
> fMeasure
[1] 0.7333333 0.4615385 0.7500000 0.2857143       NaN       NaN
> cf2
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North.America Oceania South.America
  Africa            11    0      2             0       3             3
  Asia               0    3      0             2       1             1
  Europe             0    1      9             1       2             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 : 0.5714
                 95% CI : (0.4096, 0.7228)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 2.158e-05

                  Kappa : 0.4433
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North.America Class: Oceania Class: South.America
Sensitivity                 1.0000     0.50000        0.8182              0.25000         0.0000              0.00000
Specificity                 0.7419     0.88889        0.8710              0.94737         1.0000              1.00000
Pos Pred Value              0.5789     0.42857        0.6923              0.33333            NaN                  NaN
Neg Pred Value              1.0000     0.91429        0.9310              0.92308         0.8571              0.90476
Prevalence                  0.2619     0.14286        0.2619              0.09524         0.1429              0.09524
Detection Rate              0.2619     0.07143        0.2143              0.02381         0.0000              0.00000
Detection Prevalence        0.4524     0.16667        0.3095              0.07143         0.0000              0.00000
Balanced Accuracy           0.8710     0.69444        0.8446              0.59868         0.5000              0.50000
```

5. SAMPLE 5 (Seed: 3333)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were NumOpt = 4, NumFolds = 5 and MinWeights = 4.

```
> accuracy
Accuracy
   0.375
> precision
[1] 0.3333333       NaN 0.5000000       NaN       NaN       NaN
> recall
[1] 1.0000000 0.0000000 0.6363636 0.0000000 0.0000000 0.0000000
> fMeasure
[1] 0.50  NaN 0.56  NaN  NaN  NaN
> cf2
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North.America Oceania South.America
  Africa            14   11      4             7       3             3
  Asia               0    0      0             0       0             0
  Europe             0    2      7             4       1             0
  North.America      0    0      0             0       0             0
  Oceania            0    0      0             0       0             0
  South.America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.375
                 95% CI : (0.2492, 0.5145)
    No Information Rate : 0.25
    P-Value [Acc > NIR] : 0.02588

                  Kappa : 0.1813
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North.America Class: Oceania Class: South.America
Sensitivity                 1.0000      0.0000        0.6364               0.0000        0.00000              0.00000
Specificity                 0.3333      1.0000        0.8444               1.0000        1.00000              1.00000
Pos Pred Value              0.3333         NaN        0.5000                  NaN            NaN                  NaN
Neg Pred Value              1.0000      0.7679        0.9048               0.8036        0.92857              0.94643
Prevalence                  0.2500      0.2321        0.1964               0.1964        0.07143              0.05357
Detection Rate              0.2500      0.0000        0.1250               0.0000        0.00000              0.00000
Detection Prevalence        0.7500      0.0000        0.2500               0.0000        0.00000              0.00000
Balanced Accuracy           0.6667      0.5000        0.7404               0.5000        0.50000              0.50000
```

C4.5 Decision Tree

1. SAMPLE 1 (Seed: 1707)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were C = 0.01 and M = 3.

```
> accuracy
 Accuracy
0.4864865
> precision
[1] 0.8571429 0.4375000 0.4166667       NaN       NaN 0.0000000
> recall
[1] 0.6000000 0.7777778 0.7142857 0.0000000 0.0000000 0.0000000
> fMeasure
[1] 0.7058824 0.5600000 0.5263158       NaN       NaN       NaN
```

Confusion Matrix and Statistics

	Reference					
Prediction	Africa	Asia	Europe	North.America	Oceania	South.America
Africa	6	0	0	1	0	0
Asia	2	7	1	3	1	2
Europe	2	1	5	1	3	0
North.America	0	0	0	0	0	0
Oceania	0	0	0	0	0	0
South.America	0	1	1	0	0	0

Overall Statistics

```
               Accuracy : 0.4865
                 95% CI : (0.3192, 0.656)
    No Information Rate : 0.2703
    P-Value [Acc > NIR] : 0.004058

                  Kappa : 0.3411
 Mcnemar's Test P-Value : NA
```

Statistics by Class:

	Class: Africa	Class: Asia	Class: Europe	Class: North.America	Class: Oceania	Class: South.America
Sensitivity	0.6000	0.7778	0.7143	0.0000	0.0000	0.00000
Specificity	0.9630	0.6786	0.7667	1.0000	1.0000	0.94286
Pos Pred Value	0.8571	0.4375	0.4167	NaN	NaN	0.00000
Neg Pred Value	0.8667	0.9048	0.9200	0.8649	0.8919	0.94286
Prevalence	0.2703	0.2432	0.1892	0.1351	0.1081	0.05405
Detection Rate	0.1622	0.1892	0.1351	0.0000	0.0000	0.00000
Detection Prevalence	0.1892	0.4324	0.3243	0.0000	0.0000	0.05405
Balanced Accuracy	0.7815	0.7282	0.7405	0.5000	0.5000	0.47143

2. SAMPLE 2 (Seed: 1234)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were C = 0.01 and M = 1.

```
> accuracy
Accuracy
0.547619
> precision
[1] 1.0000000 1.0000000 0.4117647 0.0000000 0.5000000 0.0000000
> recall
[1] 0.7857143 0.3333333 0.8750000 0.0000000 0.3333333 0.0000000
> fMeasure
[1] 0.88 0.50 0.56  NaN 0.40  NaN
> cf3
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North.America Oceania South.America
  Africa            11    0      0             0       0             0
  Asia               0    3      0             0       0             0
  Europe             3    2      7             1       3             1
  North.America      0    2      1             0       1             0
  Oceania            0    1      0             1       2             0
  South.America      0    1      0             2       0             0

Overall Statistics

               Accuracy : 0.5476
                 95% CI : (0.3867, 0.7015)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 0.003433

                  Kappa : 0.4316
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North.America Class: Oceania Class: South.America
Sensitivity                 0.7857     0.33333        0.8750              0.00000        0.33333              0.00000
Specificity                 1.0000     1.00000        0.7059              0.89474        0.94444              0.92683
Pos Pred Value              1.0000     1.00000        0.4118              0.00000        0.50000              0.00000
Neg Pred Value              0.9032     0.84615        0.9600              0.89474        0.89474              0.97436
Prevalence                  0.3333     0.21429        0.1905              0.09524        0.14286              0.02381
Detection Rate              0.2619     0.07143        0.1667              0.00000        0.04762              0.00000
Detection Prevalence        0.2619     0.07143        0.4048              0.09524        0.09524              0.07143
Balanced Accuracy           0.8929     0.66667        0.7904              0.44737        0.63889              0.46341
```

3. SAMPLE 3 (Seed: 1111)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were C = 0.1325 and M = 5.

```
> accuracy
 Accuracy
0.5416667
> precision
[1] 0.8461538 0.6666667 0.5714286 0.5000000 0.0000000 0.1428571
> recall
[1] 1.0000000 0.3636364 0.6153846 0.1818182 0.0000000 1.0000000
> fMeasure
[1] 0.9166667 0.4705882 0.5925926 0.2666667       NaN 0.2500000
> cf3
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North.America Oceania South.America
  Africa            11    1      1             0       0             0
  Asia               0    4      1             1       0             0
  Europe             0    2      8             4       0             0
  North.America      0    1      1             2       0             0
  Oceania            0    2      0             2       0             0
  South.America      0    1      2             2       1             1

Overall Statistics

               Accuracy : 0.5417
                 95% CI : (0.3917, 0.6863)
    No Information Rate : 0.2708
    P-Value [Acc > NIR] : 6.618e-05

                  Kappa : 0.4316
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North.America Class: Oceania Class: South.America
Sensitivity                 1.0000     0.36364        0.6154              0.18182        0.00000              1.00000
Specificity                 0.9459     0.94595        0.8286              0.94595        0.91489              0.87234
Pos Pred Value              0.8462     0.66667        0.5714              0.50000        0.00000              0.14286
Neg Pred Value              1.0000     0.83333        0.8529              0.79545        0.97727              1.00000
Prevalence                  0.2292     0.22917        0.2708              0.22917        0.02083              0.02083
Detection Rate              0.2292     0.08333        0.1667              0.04167        0.00000              0.02083
Detection Prevalence        0.2708     0.12500        0.2917              0.08333        0.08333              0.14583
Balanced Accuracy           0.9730     0.65479        0.7220              0.56388        0.45745              0.93617
```

4. SAMPLE 4 (Seed: 2222)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were C = 0.01 and M = 5.

```
> accuracy
 Accuracy
0.5714286
> precision
[1] 0.8333333 0.3750000 0.5000000       NaN       NaN       NaN
> recall
[1] 0.9090909 0.5000000 1.0000000 0.0000000 0.0000000 0.0000000
> fMeasure
[1] 0.8695652 0.4285714 0.6666667       NaN       NaN       NaN
> cf3
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North.America Oceania South.America
  Africa            10    0      0             0       2             0
  Asia               0    3      0             2       2             1
  Europe             1    3     11             2       2             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 : 0.5714
                 95% CI : (0.4096, 0.7228)
    No Information Rate : 0.2619
    P-Value [Acc > NIR] : 2.158e-05

                  Kappa : 0.4367
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North.America Class: Oceania Class: South.America
Sensitivity                 0.9091     0.50000        1.0000              0.00000         0.0000              0.00000
Specificity                 0.9355     0.86111        0.6452              1.00000         1.0000              1.00000
Pos Pred Value              0.8333     0.37500        0.5000                  NaN            NaN                  NaN
Neg Pred Value              0.9667     0.91176        1.0000              0.90476         0.8571              0.90476
Prevalence                  0.2619     0.14286        0.2619              0.09524         0.1429              0.09524
Detection Rate              0.2381     0.07143        0.2619              0.00000         0.0000              0.00000
Detection Prevalence        0.2857     0.19048        0.5238              0.00000         0.0000              0.00000
Balanced Accuracy           0.9223     0.68056        0.8226              0.50000         0.5000              0.50000
```

5. SAMPLE 5 (Seed: 3333)

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were C = 0.255 and M = 5.

```
> accuracy
 Accuracy
0.4464286
> precision
[1] 0.85714286 0.33333333 0.50000000 0.08333333 0.00000000 0.25000000
> recall
[1] 0.85714286 0.23076923 0.72727273 0.09090909 0.00000000 0.33333333
> fMeasure
[1] 0.85714286 0.27272727 0.59259259 0.08695652        NaN 0.28571429
> cf3
Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 0.4464
                 95% CI : (0.3134, 0.5853)
    No Information Rate : 0.25
    P-Value [Acc > NIR] : 0.001073

                  Kappa : 0.3053
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Africa Class: Asia Class: Europe Class: North.America Class: Oceania Class: South.America
Sensitivity                 0.8571     0.23077        0.7273              0.09091        0.00000              0.33333
Specificity                 0.9524     0.86047        0.8222              0.75556        0.98077              0.94340
Pos Pred Value              0.8571     0.33333        0.5000              0.08333        0.00000              0.25000
Neg Pred Value              0.9524     0.78723        0.9250              0.77273        0.92727              0.96154
Prevalence                  0.2500     0.23214        0.1964              0.19643        0.07143              0.05357
Detection Rate              0.2143     0.05357        0.1429              0.01786        0.00000              0.01786
Detection Prevalence        0.2500     0.16071        0.2857              0.21429        0.01786              0.07143
Balanced Accuracy           0.9048     0.54562        0.7747              0.42323        0.49038              0.63836
```

Conclusion

This project was mainly concerned with performing basic classification algorithms such as SVM, KNN, C4.5 and RIPPER with the help of R. The Life Expectancy data set that was used was small and over viewable.

The splitting of dataset and selection of trainset and testset has a significant impact on the overall accuracies from various classification algorithms. The method for cross-validating and tuning the parameters of various algorithms determines the predicted model accuracies.

The following are the cumulated results obtained for each of the algorithms for all the five seed values in the form of average and standard deviation. As we can see below, KNN was the most accurate among all the classification methods used with an accuracy of approximately 58%. It was also the faster among all the algorithms.

	Mean	Standard Deviation
K Nearest Neighbors	0.57834	0.051333
Support Vector Machine	0.5635	0.039333
RIPPER Decision Tree	0.48182	0.07039
C4.5 Decision Tree	0.51872	0.051004

References

- https://www.datacamp.com/community/tutorials/machine-learning-in-r
- http://www.saedsayad.com/k_nearest_neighbors.htm
- https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
- https://en.wikipedia.org/wiki/C4.5_algorithm
- https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification/JRip