Flag Dataset

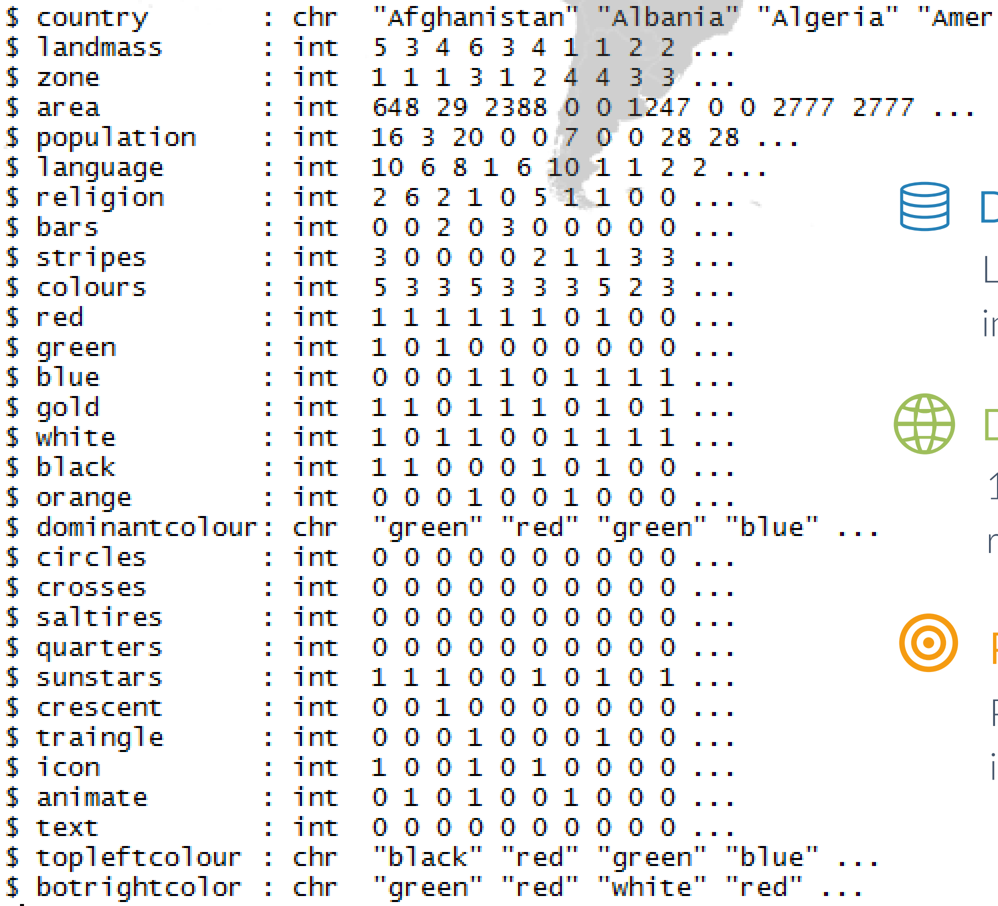
Classification & Clustering Techniques

Statistics for BA II

By Eva Giannatou

The Dataset

Data from the UCI Machine Learning Repository
containing details of various nations and their flags.



```
$ country        : chr  "Afghanistan" "Albania" "Algeria" "Amer
$ landmass       : int  5 3 4 6 3 4 1 1 2 2 ...
$ zone           : int  1 1 1 3 1 2 4 4 3 3 ...
$ area           : int  648 29 2388 0 0 1247 0 0 2777 2777 ...
$ population     : int  16 3 20 0 0 7 0 0 28 28 ...
$ language       : int  10 6 8 1 6 10 1 1 2 2 ...
$ religion       : int  2 6 2 1 0 5 1 1 0 0 ...
$ bars           : int  0 0 2 0 3 0 0 0 0 0 ...
$ stripes        : int  3 0 0 0 0 2 1 1 3 3 ...
$ colours        : int  5 3 3 5 3 3 3 5 2 3 ...
$ red            : int  1 1 1 1 1 1 0 1 0 0 ...
$ green          : int  1 0 1 0 0 0 0 0 0 0 ...
$ blue           : int  0 0 0 1 1 0 1 1 1 1 ...
$ gold           : int  1 1 0 1 1 1 0 1 0 1 ...
$ white          : int  1 0 1 1 0 0 1 1 1 1 ...
$ black          : int  1 1 0 0 0 1 0 1 0 0 ...
$ orange         : int  0 0 0 1 0 0 1 0 0 0 ...
$ dominantcolour: chr  "green" "red" "green" "blue" ...
$ circles        : int  0 0 0 0 0 0 0 0 0 0 ...
$ crosses        : int  0 0 0 0 0 0 0 0 0 0 ...
$ saltires       : int  0 0 0 0 0 0 0 0 0 0 ...
$ quarters       : int  0 0 0 0 0 0 0 0 0 0 ...
$ sunstars       : int  1 1 1 0 0 1 0 1 0 1 ...
$ crescent       : int  0 0 1 0 0 0 0 0 0 0 ...
$ traingle       : int  0 0 0 1 0 0 0 1 0 0 ...
$ icon           : int  1 0 0 1 0 1 0 0 0 0 ...
$ animate        : int  0 1 0 1 0 0 1 0 0 0 ...
$ text           : int  0 0 0 0 0 0 0 0 0 0 ...
$ topleftcolour : chr  "black" "red" "green" "blue" ...
$ botrightcolor : chr  "green" "red" "white" "red" ...
```

Dataset included

List of countries with demographics, geographic information and flag descriptions

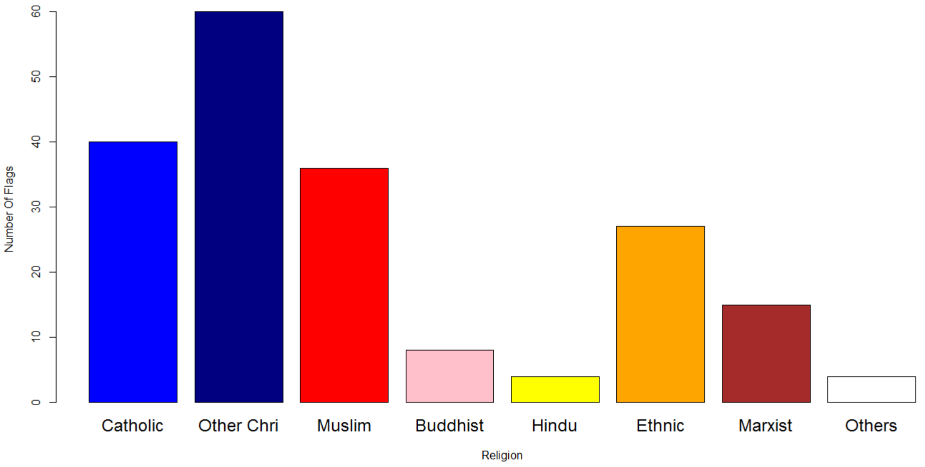
Dataset consists of

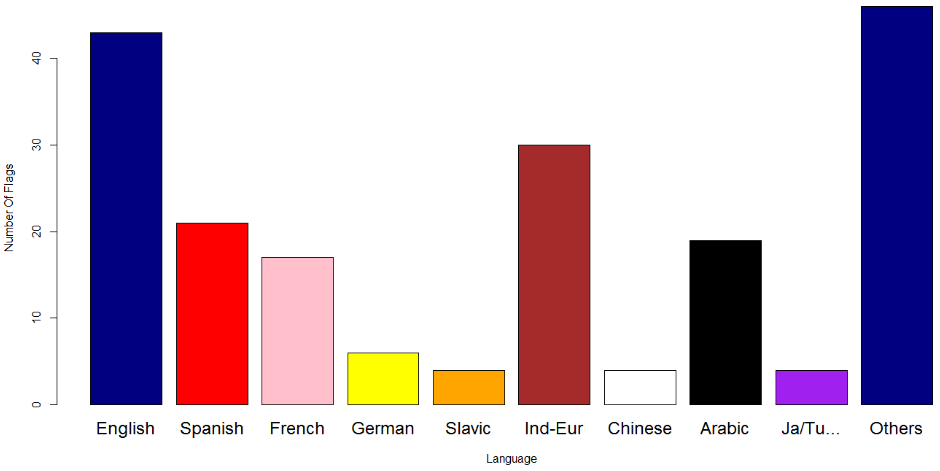
194 countries and 30 attributes (10 numeric and the rest of them were either Boolean or nominal)

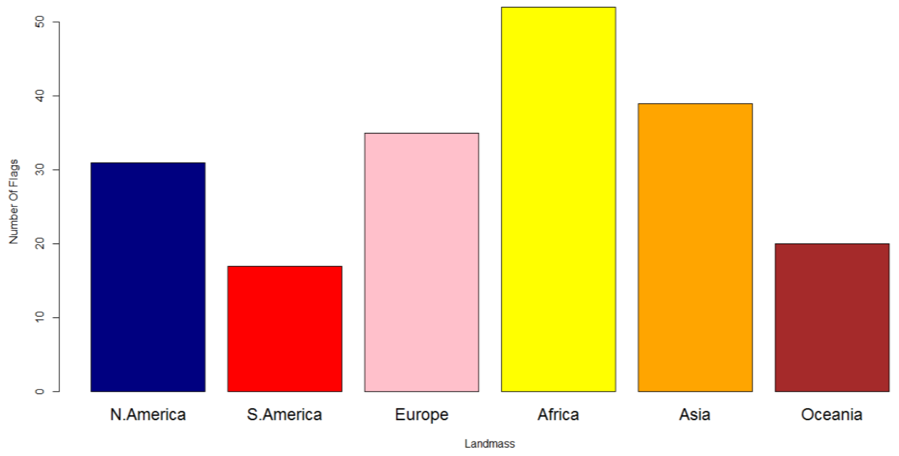
Project purpose

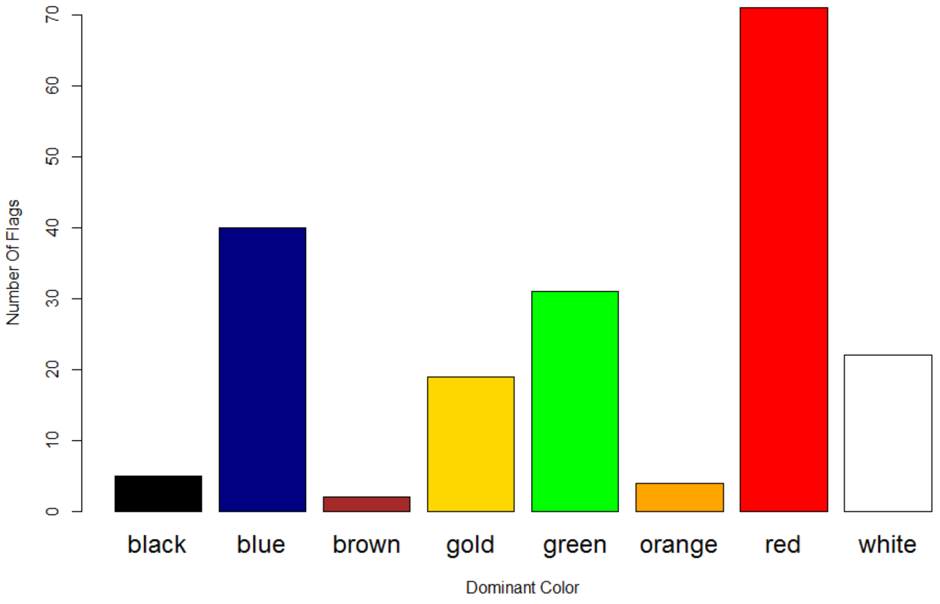
Predicting the religion of a country from its flag's characteristics

Data Visualization









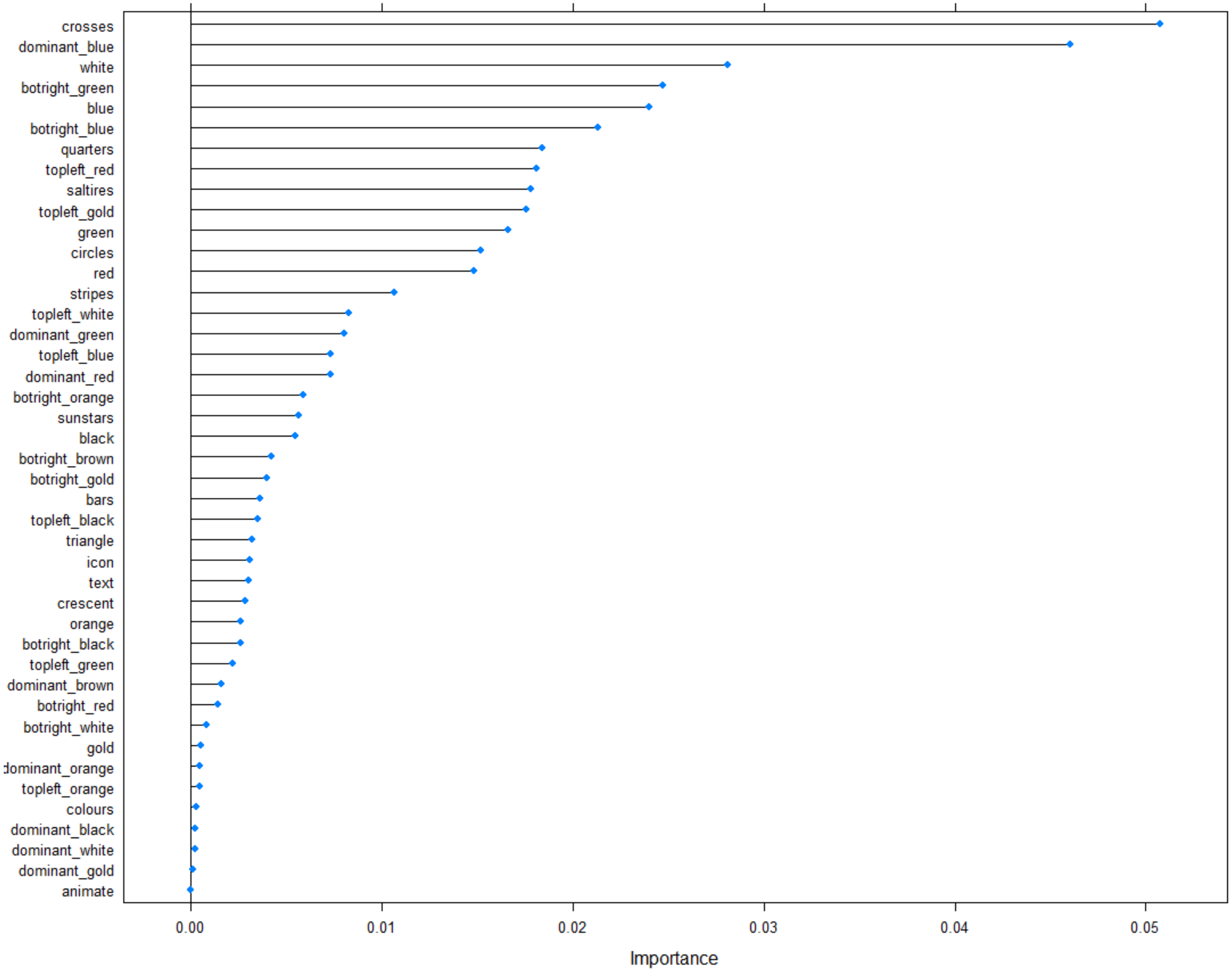
Important Variables

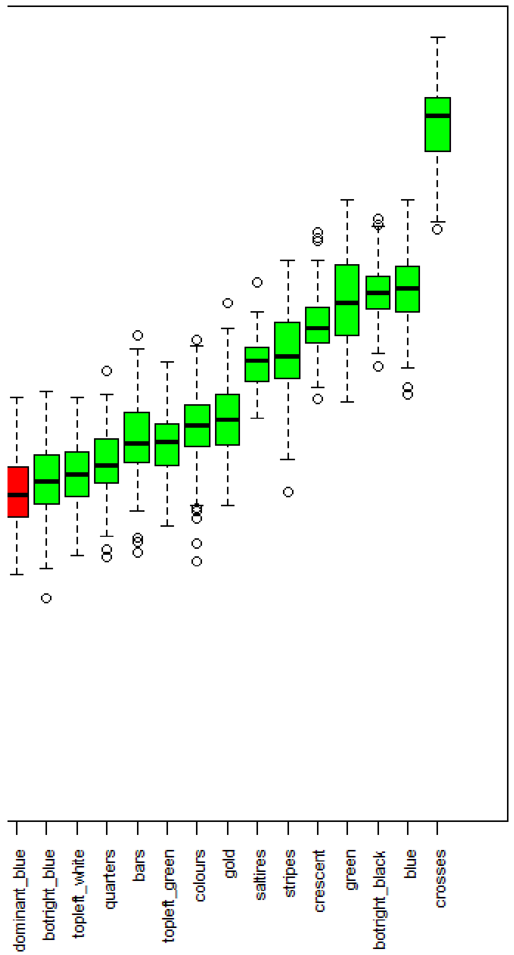
Which flag characteristics are correlated to the religion of a country?

crosses
white
botright_green
blue
botright_blue
dominant_blue

quarters
saltires
green
circles
bars
stripes

colours
gold
black
crescent
animate
topleft_white





Classification Models

Train models using training set
Test models using test set

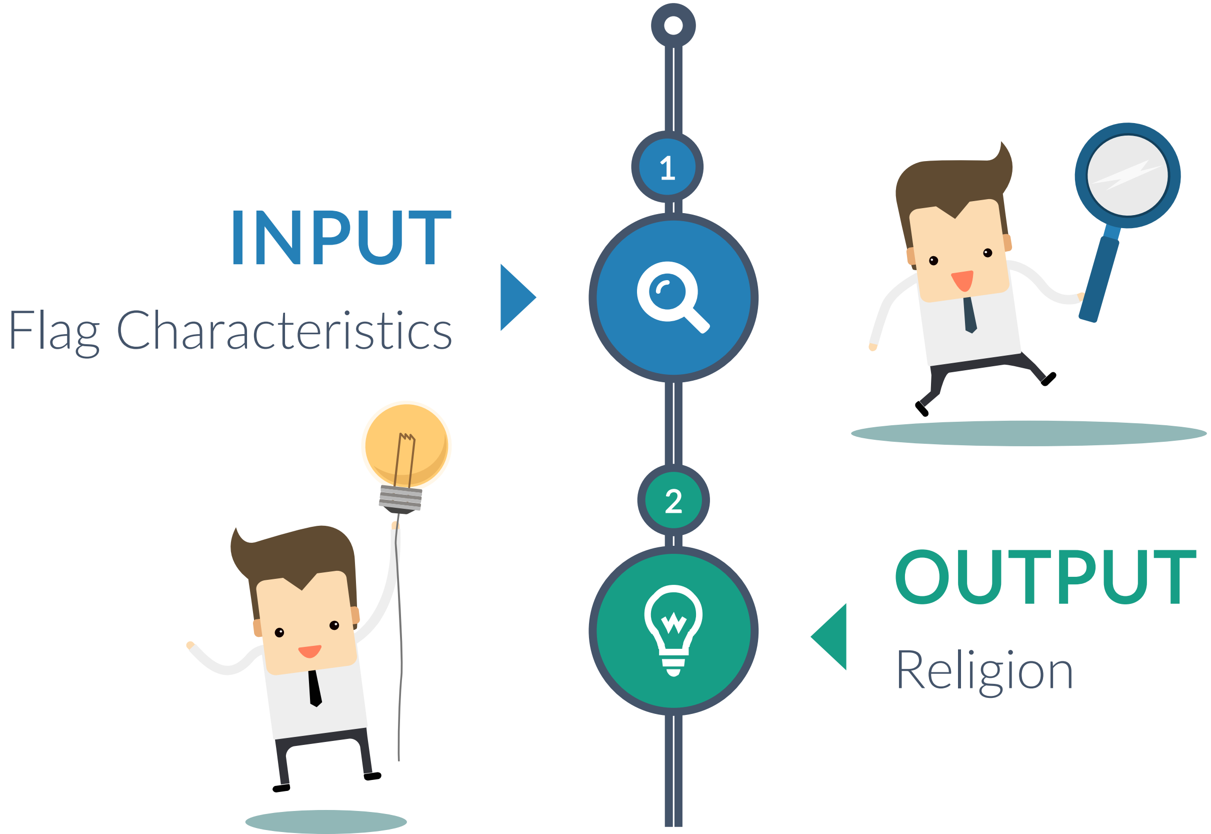
Top Performing Models

1 Rpart
Accuracy 84%
PCs Used 5

2 C5.0
Accuracy 61%
PCs Used 9

5 J48
Accuracy 58%
PCs Used 10

4 Bagging cart
Accuracy 55%
PCs Used 9



INPUT

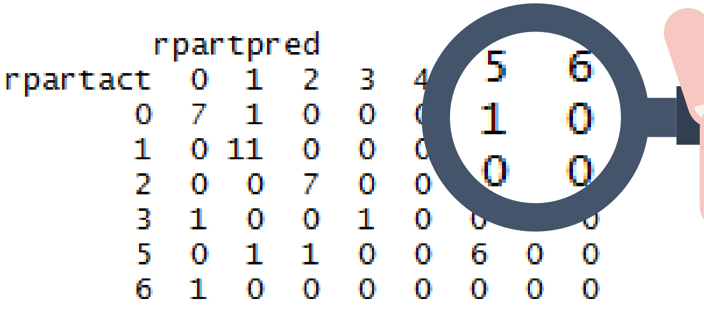
Flag Characteristics

OUTPUT

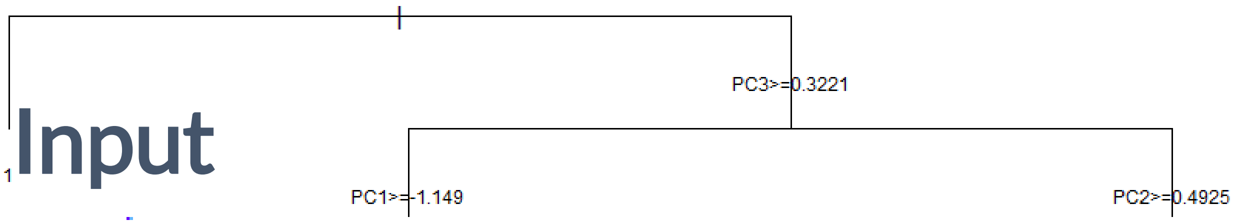
Religion

Better Performing Classification Model

Rpart Classification:
accuracy 84%



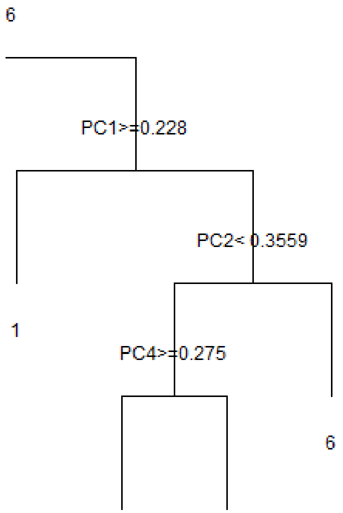
```
         rpartpred
rpartact  0  1  2  3  4
       0  7  1  0  0  0
       1  0 11  0  0  0
       2  0  0  7  0  0
       3  1  0  0  1  0
       5  0  1  1  0  0  6  0  0
       6  1  0  0  0  0  0  0  0
```



Input

```
'data.frame':   156 obs. of  11 variables:
 $ religion: int  6 1 0 1 0 0 1 2 2 1 ...
 $ PC1     : num  -1.264 1.273 -1.136 -0.104 1.172 ...
 $ PC2     : num  -1.523 -0.725 0.222 -0.565 2.635 ...
 $ PC3     : num  -0.756 -1.533 -0.355 -1.587 -1.65 ...
 $ PC4     : num  -1.019 -0.573 -2.758 0.438 -0.413 ...
 $ PC5     : num  -1.909 -0.895 -0.796 -1.205 0.662 ...
 $ PC6     : num  -1.311 1.038 -0.226 -0.483 0.725 ...
 $ PC7     : num  -0.2599 -0.0319 -0.5769 0.2665 0.2759 ...
 $ PC8     : num  0.544 -0.83 -1.43 -0.243 0.667 ...
 $ PC9     : num  -0.622 0.15 0.895 -0.703 0.157 ...
 $ PC10    : num  -0.902 -1.741 0.859 0.212 -0.281 ...
```

Output Decision Tree



Clustering Models

1. Find clusters of flags with respect to their characteristics
2. Discover the common characteristics of the countries within each flag cluster
3. Investigate possible connection between the clusters and the religions of the within countries

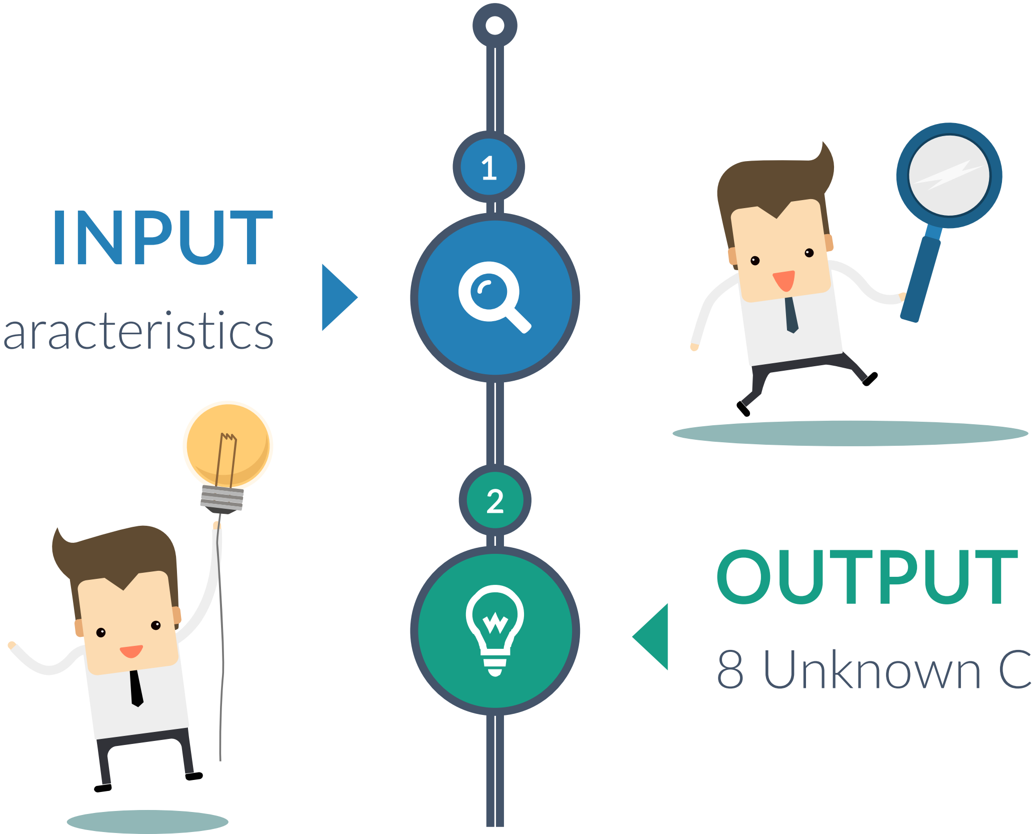
Models

1 Hierarchical

2 PAM

5 Mcluster

4 KMeans



INPUT

Flag Characteristics

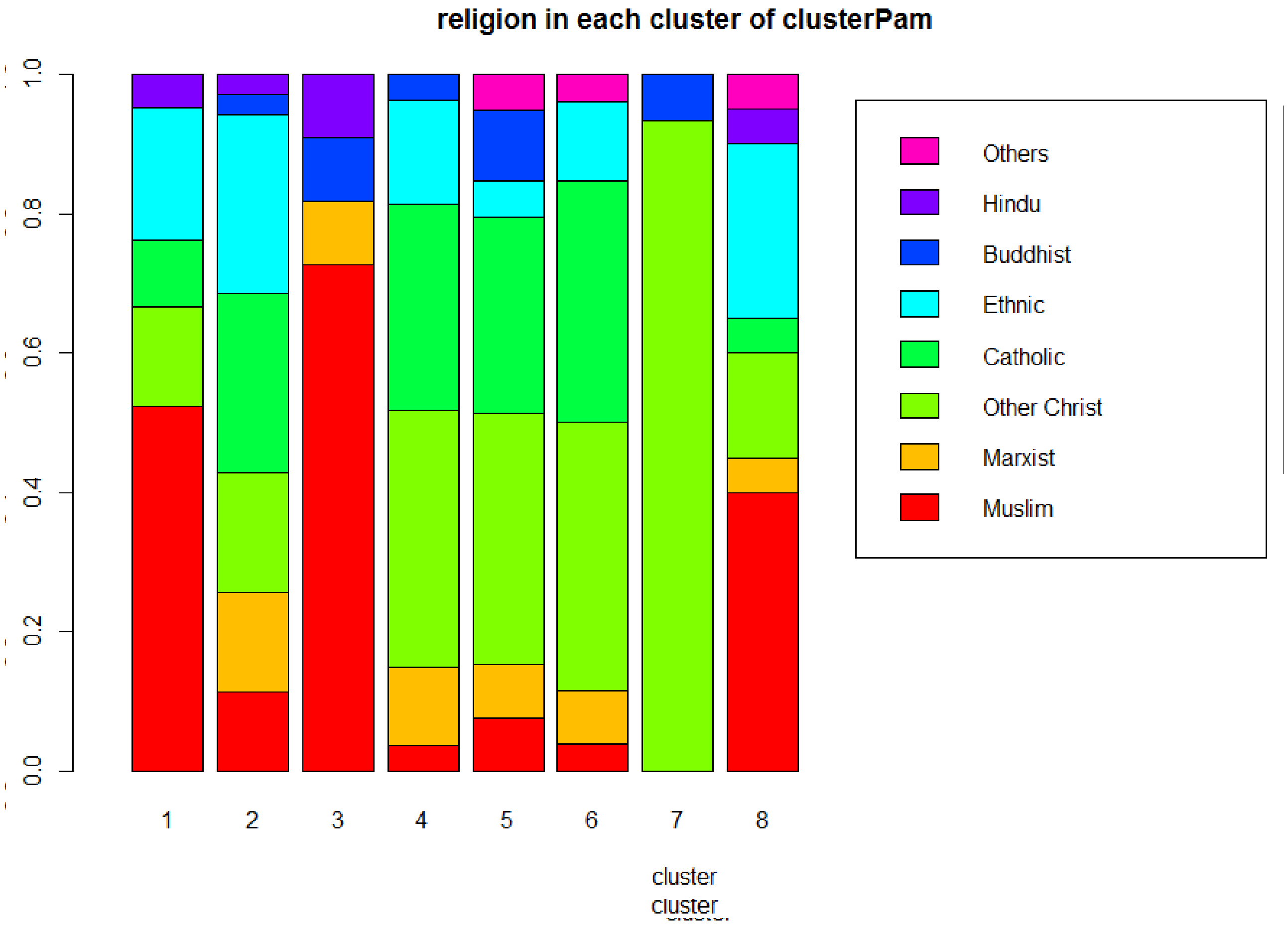
OUTPUT

8 Unknown Clusters

Cluster Analysis

Visualization of Cluster Characteristics

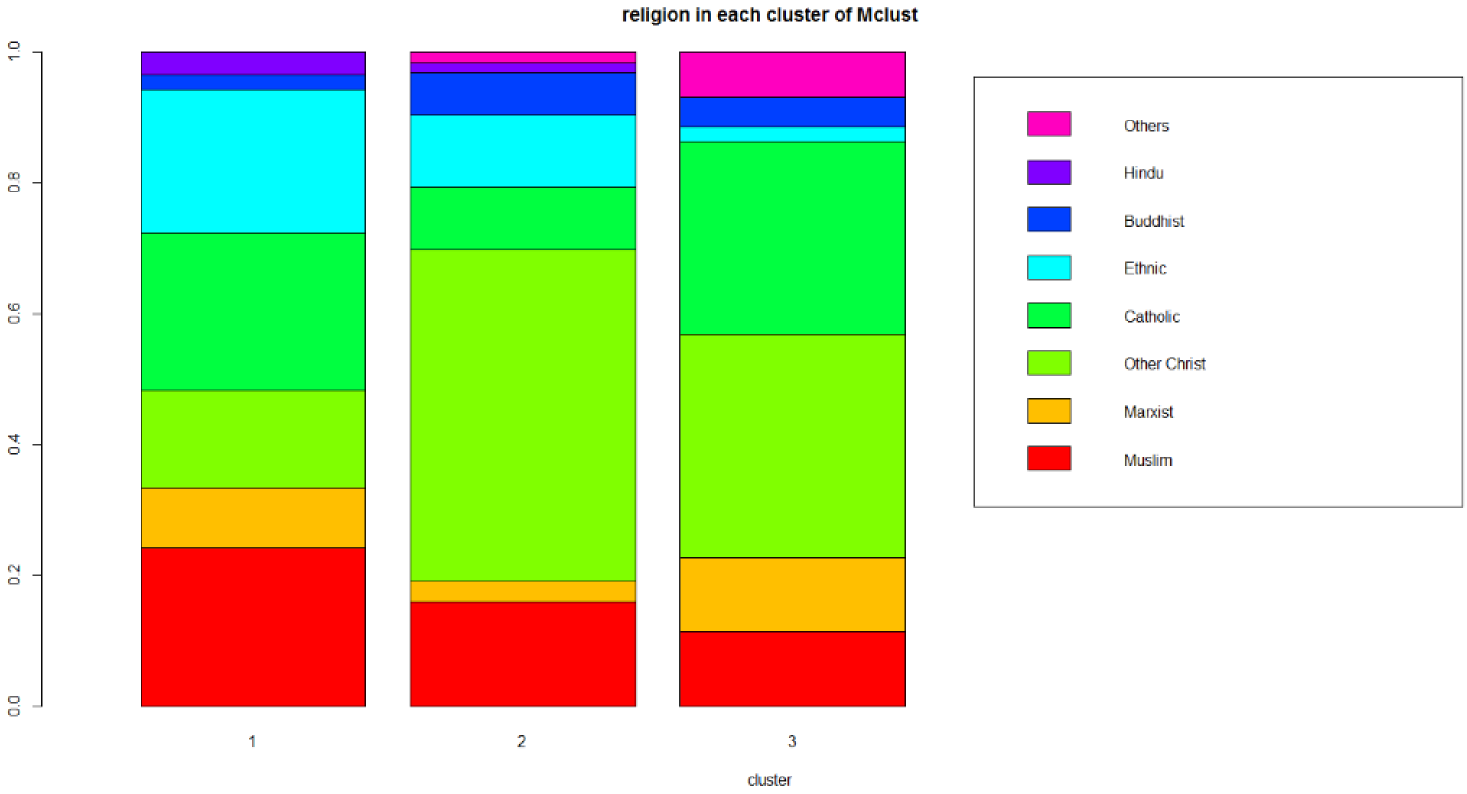
8 Clusters



Improved Cluster Analysis

Visualization of Cluster Characteristics

Recommended number of clusters: 3


religion in each cluster of Mclust
cluster
Others
Hindu
Buddhist
Ethnic
Catholic
Other Christ
Marxist
Muslim