INFO523 Decicion Trees

Sebastian Deimen & Noah Giebink

21 März 2020

At first, we are going to make two sets of our spot-data: one only related to the music vaiables and one also including the socio- variables.

We splitted the spot_music_SOCIO data into training and test data, not using a validation set.

Our goal was to predict "country". We made a tree with the training data, used it to predict on our testdata and checked the results/error rates. The tree did surprisingly well with an error rate of 0 %.

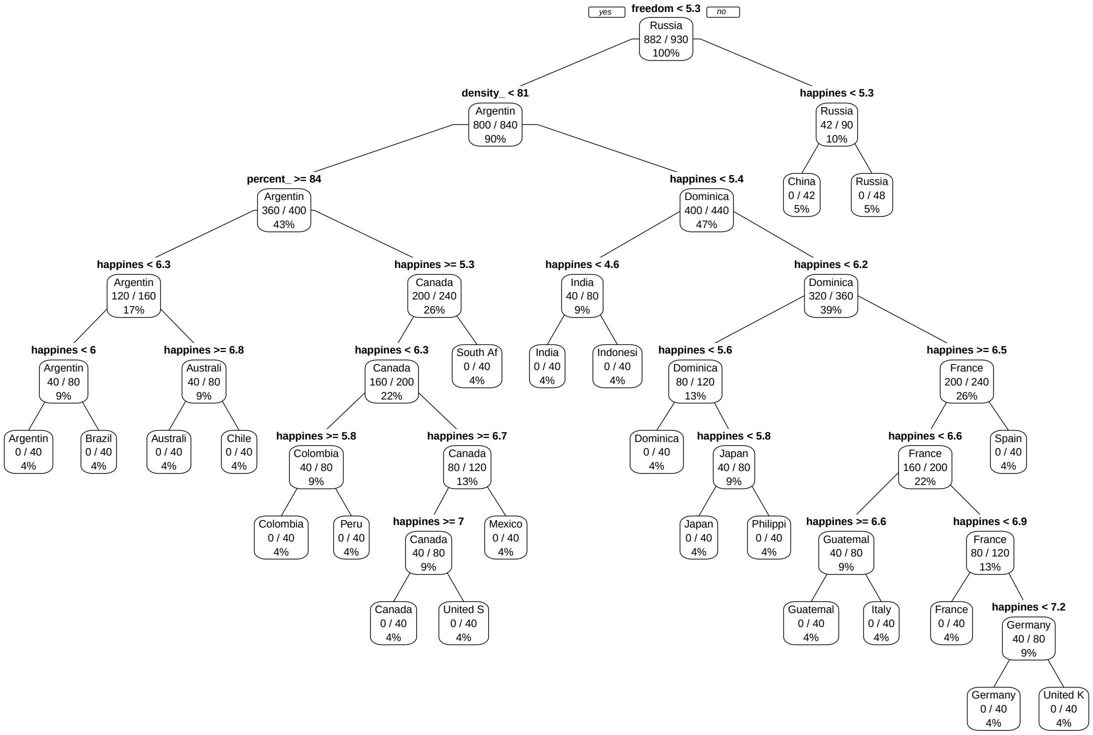
We also made a tree and prediction just for the music variables to predict "country", but the tree had a horrible error rate of around 82 %. So we decided to choose a different approach.

```
##             happiness          density_sqkm percent_internet_users
##             774.80000             681.46667              632.80000
##         percent_urban            median_age                freedom
##             618.13333             574.13333              418.03011
##                   gdp      track.popularity           danceability
##             330.99183              33.94272               16.00000
##           speechiness      instrumentalness          track.explicit
##              14.00000              11.00000               11.00000
##          acousticness              loudness               liveness
##               6.00000               3.00000                2.00000
```

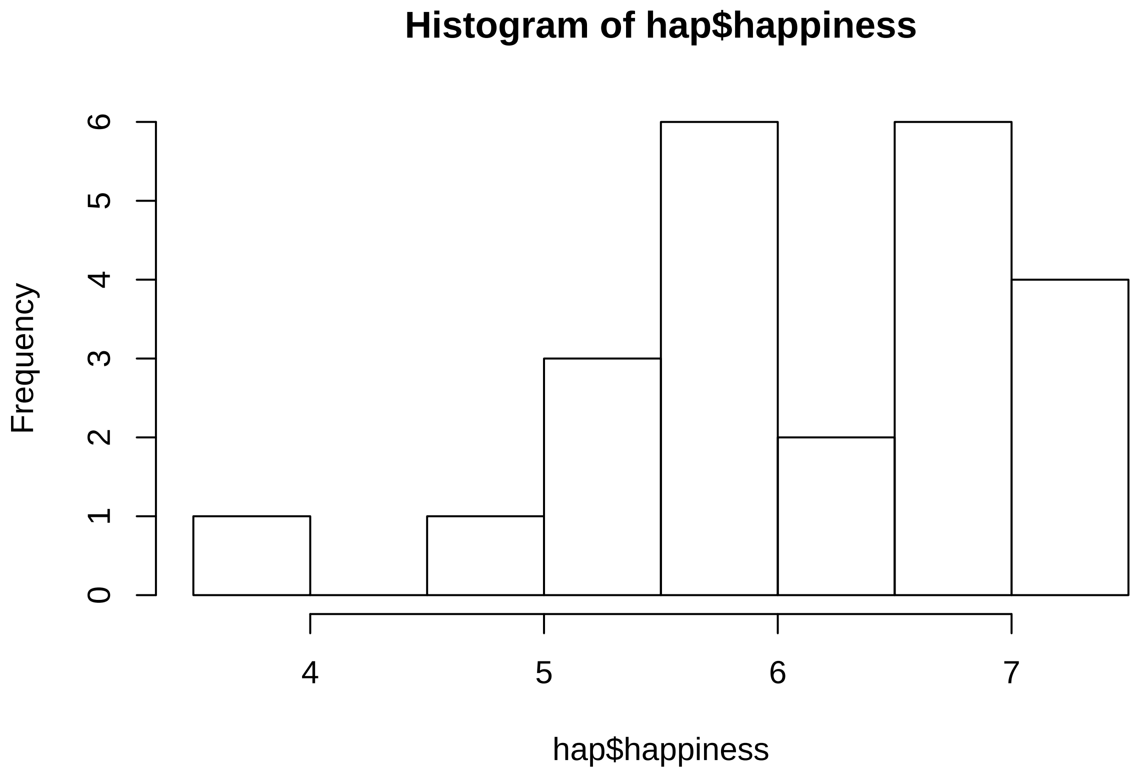
```
##            median_age             happiness percent_internet_users
##           591.8616306           569.1125660            546.9475318
##         percent_urban          density_sqkm                    gdp
##           445.1902581           426.2279907            402.4828312
##               freedom      track.popularity           acousticness
##           373.9048700            34.7183580             14.3690706
##              loudness              liveness           danceability
##            11.9792208             5.8925006              2.4552086
##      instrumentalness           speechiness                  tempo
##             1.5221542             0.9925037              0.4970497
```

```
##  error_create rate:  0
```

```
##  error_sample rate:  0
```



The different approach: We clustered countries by most important social feature (happiness) for classification. We decided to use two k = 2 to get “happy” and “unhappy” countries. We then bound the clusters to our solely music-varibale data and used this to grow the tree.



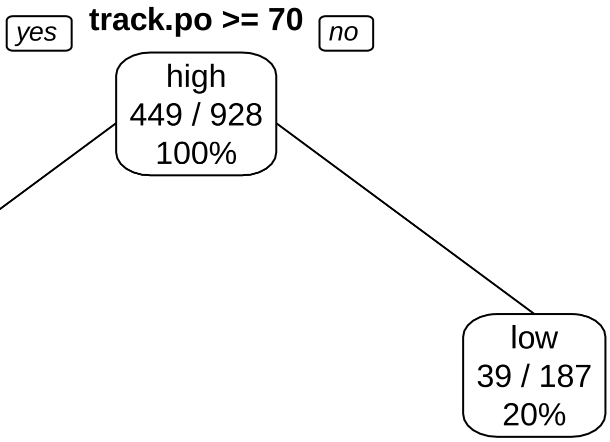
```
##  [1] 1 2 2 2 2 1 1 1 2 2 2 1 1 2 1 2 1 1 1 1 2 2 2

##               country happiness cluster
## 1           Argentina  5.792797       1
## 2               China  5.131434       1
## 3            Colombia  5.983512       1
## 4  Dominican Republic  5.433216       1
## 5               India  3.818069       1
## 6           Indonesia  5.340296       1
## 7               Japan  5.793575       1
## 8                Peru  5.679661       1
## 9         Philippines  5.869173       1
## 10             Russia  5.513500       1
## 11       South Africa  4.883922       1
## 12          Australia  7.176993       2
## 13             Brazil  6.190922       2
## 14             Canada  7.175497       2
## 15              Chile  6.436221       2
## 16             France  6.665904       2
## 17            Germany  7.118364       2
## 18          Guatemala  6.626592       2
## 19              Italy  6.516527       2
## 20             Mexico  6.549579       2
## 21              Spain  6.513371       2
## 22     United Kingdom  7.233445       2
## 23      United States  6.882685       2
```

Splitting, growing, predicting and plotting for the different approach:

The spot_music trees and error rates:

```
## track.popularity     danceability instrumentalness     acousticness
##       44.3195779        2.8440371        0.7110093        0.4740062
##      speechiness         loudness
##        0.4740062        0.2370031
```

```
## error_sample rate:  0.3991416
```



Plot the tree