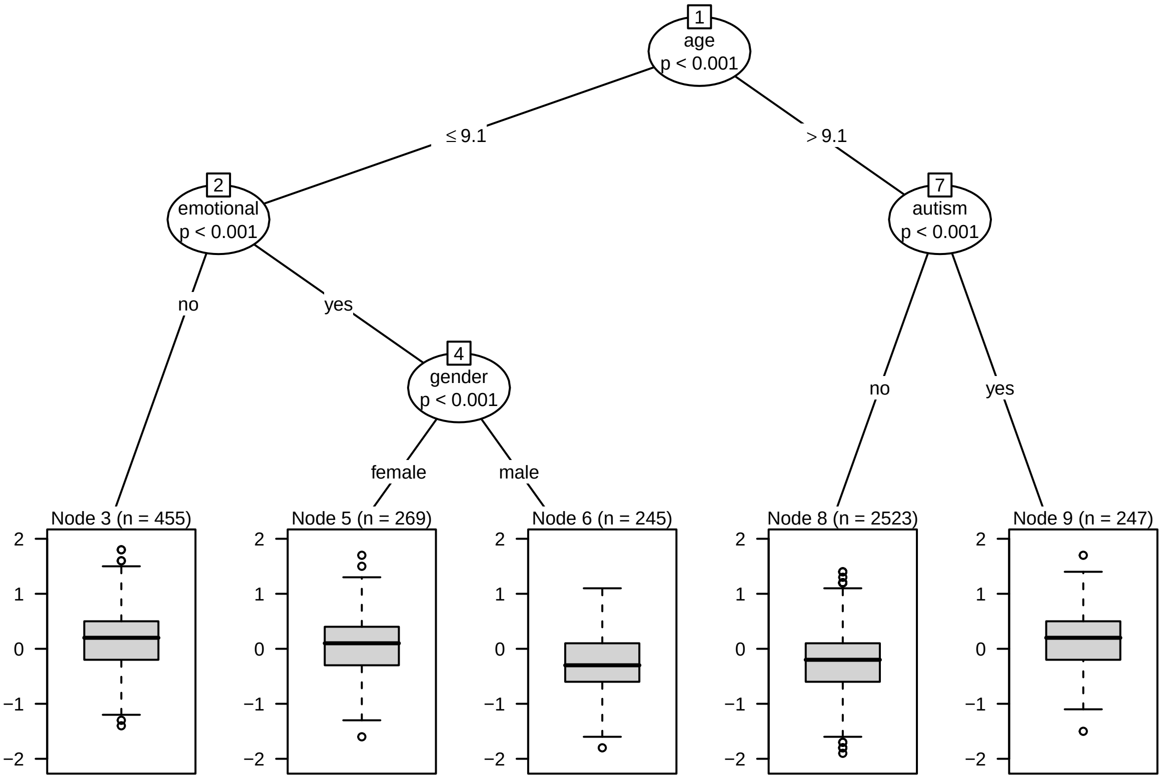
Assess stability of splitting variables and values

Load package and data:

```
library("glmertree")
data <- read.table("UK_MH mimic data.txt")
```

Fit tree on full data:

```
full_tree <- lmertree(outcome ~ 1 | cluster_id | age + gender + emotional +
                         autism + impact + conduct, data = data)
plot(full_tree$tree, gp = gpar(cex = .6))
```



Assess stability of splitting variables and values

We first install the latest version of the `stablelearner` package (note: it is not available on CRAN yet and you only have to do this once):

```
install.packages("stablelearner", repos="http://R-Forge.R-project.org")
```

Then we load the package:

```
library("stablelearner")
```

Using the `stabletree()` function, we can assess stability of splits. The `stabletree()` function takes the fitted tree as input. It then takes `B` random subsamples (500, by default) from the data that was used to construct the tree. On each of those samples, it fits a tree and evaluates which splitting variables and values appear in the tree. The random subsamples can be bootstrap samples (that is, observations are sampled with replacement to generate a sample of the same size as the original data), subsamples (that is, observations are smapled without replacement, the size of the subsample will be smaller than that of the original sample), or an alternative approach (type `?bootstrap` for more info). By default, `stabletree()` employs bootstrap sampling, but this can yield higher inclusion frequencies for noise variables. I therefore prefer to use subsampling with `v = .8` or `v = .9`, yielding subsamples with 80 or 90 percent of the observations in the original sample, respectively.

```
set.seed(42)
stab <- stabletree(full_tree, sampler = subsampling(B = 500, v = .8))
```

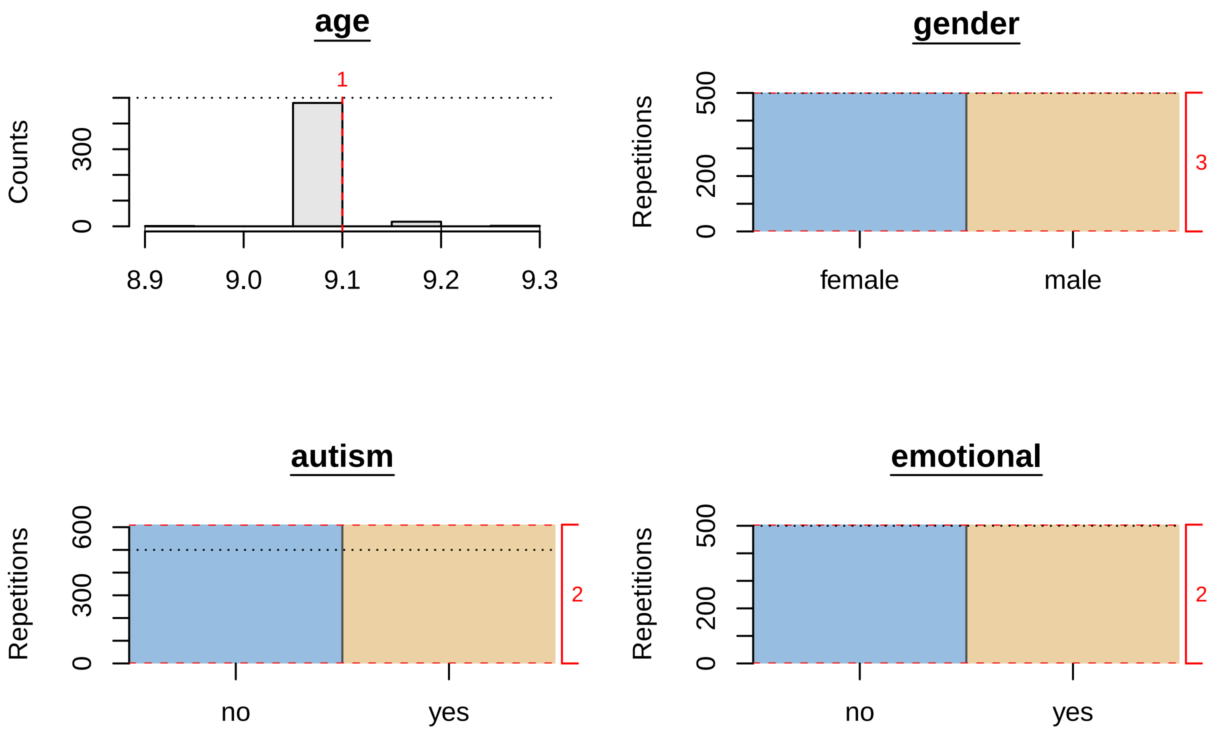
```
summary(stab)
```

```
##
## Call:
## lmertree(formula = outcome ~ 1 | cluster_id | age + gender +
##     emotional + autism + impact + conduct, data = data)
##
## Sampler:
## B = 500
## Method = Subsampling with 80.0% data
##
## Variable selection overview:
##
##            freq *  mean *
## age       1.000 1 1.002 1
## gender    1.000 1 1.002 1
## emotional 1.000 1 1.008 1
## autism    1.000 1 1.222 1
## impact    0.012 0 0.012 0
## conduct   0.002 0 0.002 0
## (* = original tree)
```

Using the `summary()` function, we see that `age`, `gender`, `emotional` and `autism` were always selected as splitting variables, and `impact` and `conduct` only very rarely, indicating that the same variables are selected for splitting in almost all subsamples. The `mean` column indicates the average number of times each variables was used for splitting in a tree. These values can be > 1, as a variable can appear multiple times in a single tree.

Next, we assess the stability of the splitting values. We only select variables that appeared in the original tree and were often selected for splitting in the subsamples (otherwise we get too many plots):

```
plot(stab, select = c("age", "gender", "autism", "emotional"))
```



The plots reveal that the same splitting value for `age` was used in almost al trees. For for `gender`, `autism` and `emotional`, the exact same splitting value was used in all trees, but this is not a very surprising or interesting result for binary factors, as those only have a single possible splitting value. Note that the red lines in the plot reveal the splitting values in the original trees, and the number of the split in the original tree in which the variable and value was used.