Exercise Sheet 6

(until 24.05.2019)

What factors explain excessive alcohol consumption among students? The record for the task sheet comes from a survey of students who attended mathematics and Portuguese courses and contains many interesting details about their sociodemographics, life circumstances and learning success.

The ordinal scaled variables `Dalc` and `Walc` give information about the alcohol consumption of the students on weekdays and weekends. Create a binary target variable `alc_prob` as follows:

```
library(stringr)
library(readr)
library(dplyr)
# (Adapt Path)
student <- read_csv(str_c(dirname(getwd()), "/Data/student_alc.csv"))
student <- student %>%
mutate(alc_prob = ifelse(Dalc + Walc >= 6, "alc_p", "no_alc_p"))
```

1. Calculate the Gini index for the target variable `alc_prob` and the _Gini index_ for each variable with respect to `alc_prob`. Determine the 5 variables with the highest _Gini Gain_.
2. Learn 2 different decision trees with `alc_prob` as target variable. For the first tree, nodes should be further partitioned until the class distribution of all resulting leaf nodes is pure. For the second tree, nodes with a cardinality of less than 20 instances should not be further partitioned. Determine the quality of the trees by calculating sensitivity (_True Positive Rate_) and specificity (_True Negative Rate_) for a 70%:30% split in training and test sets. Display the decision trees graphically and discuss the differences in quality measures.
3. Use `randomForest::randomForest()` to create a random forest with 200 trees. As candidates for a split within a tree a random sample of 5 variables should be drawn. Calculate Accuracy, Sensitivity and Specificity for the Out-of-the-Bag instances and show the most important variables (`?importance`).

Dataset: http://isgwww.cs.uni-magdeburg.de/cv/lehre/VisAnalytics/material/exercise/datasets/student_alc.csv

(**Source**: https://www.kaggle.com/uciml/student-alcohol-consumption)