Classification

Preventing excessive drug consumption among young adults is a social responsibility. Publicly sponsored Drug Prevention Programs (DPPs) are one possible preventive measure against drug addiction. However, DPPs are relatively expensive. Budget limits of public institutions restrict the number of people who can participate in DPPs. Many DPPs are targeted at youths with a high risk to become drug addicted. The objective of this exercise is to predict the risk of young adults to become drug addicted based on socio-economic characteristics. These predictions could be used to assign youths to DPPs.

The file `drugs.RData` contains information about drug consumption and socio-economic characteristics for 10'001 young adults in the United States. Table 1 describes the observable variables.

Table 1: Variable descriptions.

Variable	Description
Gender	String for female or male
Age	Age categories: 16-17 years, 18-19 years, and 20-24 years
Ethnicity	Ethnicity categories: Afro-American, Hispanic, White, and Other
Employment	String for at least once employed or unemployed during the previous year
Earning	Earnings categories (measured in preceding year): <1k USD, 1-5k USD, 5-10k USD, and >10k USD
Partner_Status	String for single or partner
Delinquency	String for arrested or not arrested in the previous year
Soft_Drug	Logical for soft drug consumption
Hard_Drug	Logical for hard drug consumption

Group Home Assignment (max. 4 points)

The mandatory group home assignment has to be submitted before 12:00 o'clock prior PC-session 5. It is obligatory to solve the assignment in R Markdown with `echo = TRUE` option for every code chunk and generate a PDF file with the solution. Generating an HTML file from the R Markdown and converting it into the PDF is also possible. Make sure that the final PDF file has no readability issues. The PDF with the answers to the six questions below as well as the file with the R Markdown code has to be submitted via Canvas.
Download the data set `drugs.RData` from Canvas. Load the data into R. Install and load the packages `rpart` and `rpart.plot`.

1. How large is the share of males who consume soft drugs (in percent)? (0.5 points)

2. How large is the difference between the share of male and female hard drug consumers (in percentage points)? (0.5 points)

3. Report the shares of young adults who consume soft drugs for each age group (16-17 years, 18-19 years, and 20-24 years). Is soft drug consumption increasing or decreasing with age? (0.5 points)

4. Tabulate the observations by earnings category and soft drug consumption. Perform a chi-squared test to evaluate whether soft drug consumption is independent of the earnings. Can you reject the independence hypothesis at a 5% significance level? (0.5 points)
   Hint: You can learn how to perform a chi-squared test in the help section of the `chisq.test()` *function.*

5. Draw 500 times a random subsample of 500 observations in your dataset and record the average soft drug consumption in each subsample. Draw a histogram of your results. Are the recorded subsample means close to the average drug consumption in the full sample? (1 point)

6. Write a function that allows to specify multiple subsample size and multiple number of draws. Your function should perform the same procedure as in the previous exercise, but for all combinations of specified subsample sizes and number of draws, and return the average drug consumption information of every single draw. Run your function for `N_runs=c(100, 500, 2500)` and `sample_sizes=c(100, 500, 2500)`. Use the `geom_density()` function to draw two kernel density estimates of your results. In the first, fix the number or draws at 500 and visualize the density estimate for the three different subsample sizes. In the second, fix the subsample size at 500 and visualize the density estimate for the three different number of draws. What asymptotic behavior do you observe in each plot? (1 point)

The exercises 1 and 2 will be solved during the PC session. They are not part of the group home assignment.

Exercise 1: Soft Drug Consumption

1. Partition randomly the data into a 75% training and 25% test sample.
2. Fit a classification tree in the training sample using gender as only predictor for soft drug consumption. Plot and interpret the tree model.
3. Predict soft drug consumption in the test sample. How good is the prediction accuracy?
4. Fit a classification tree in the training sample with gender, age, ethnicity, employment, earnings, partnership status, and delinquency as predictors for soft drug consumption. Grow a fully stratified deep tree with the complexity parameter `cp` = -1.
5. Prune the tree to its optimal level. Plot and interpret the tree model.
6. Predict soft drug consumption in the test sample. Compare the prediction accuracy of the shallow, deep, and pruned trees.

Exercise 2: Hard Drug Consumption

1. Fit a classification tree in the training sample with gender, age, ethnicity, employment, earnings, partnership status, and delinquency as predictors for hard drug consumption. Grow a fully stratified deep tree with the complexity parameter `cp` = -1.
2. Prune the tree to its optimal level. Plot and interpret the tree model.
3. Predict hard drug consumption in the test sample. Compare the prediction accuracy of the deep and pruned trees.

Useful links:

- A description of the `rpart` package can be found under: `https://cran.r-project.org/web/packages/rpart/rpart.pdf`.