Introduction to classification and regression trees, random forests and model-based recursive partitioning in R

Exercise 6: Logistic Regression Based Recursive Partitioning

For this exercise, we'll use the Pima Indians Diabetes dataset. We will partition the dataset based on a logistic regression model, predicting the probability of having diabetes based on the outcome of a glucose tolerance test.

```
# install.packages("mlbench")  ## use only if package mlbench is not yet installed
library(partykit)
```

```
## Loading required package: grid
```

```
data("PimaIndiansDiabetes", package = "mlbench")
?mlbench::PimaIndiansDiabetes
```

```
## starting httpd help server ...
```

```
##  done
```

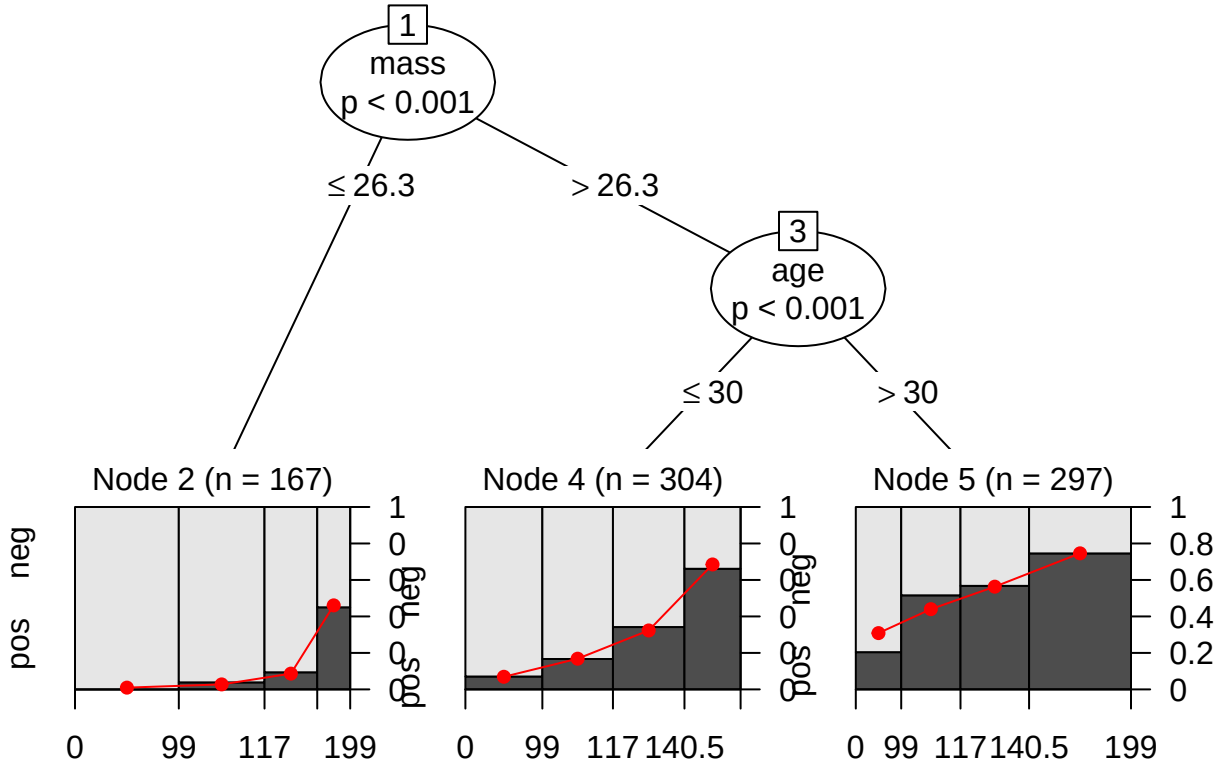
a) Use the function `glmtree()` from package `partykit` to fit a logistic regression based recursive partition. Regress the diabetes variable on the glucose variable and specify all other variables as potential partitioning variables. Add the argument `family = binomial` to specify a logistic regression model.

```
pid_tree <- glmtree(diabetes ~ glucose | pregnant + pressure + triceps + insulin +
mass + pedigree + age, data = PimaIndiansDiabetes, family = binomial)
```

b) Plot and print the tree and interpret the results. Do the subgroups differ in terms of the association between the outcome of the glucose intolerance test and the probability of having diabetes? Or do the subgroups mostly differ in their overall probablity of having diabetes?

```
plot(pid_tree)
```

```
## Loading required namespace: vcd
```




```
print(pid_tree)
```

```
## Generalized linear model tree (family: binomial)
##
## Model formula:
## diabetes ~ glucose | pregnant + pressure + triceps + insulin +
##     mass + pedigree + age
##
## Fitted party:
## [1] root
## |   [2] mass <= 26.3: n = 167
## |       (Intercept)     glucose
## |       -9.95150963  0.05870786
## |   [3] mass > 26.3
## |   |   [4] age <= 30: n = 304
## |   |       (Intercept)     glucose
## |   |       -6.70558554  0.04683748
## |   |   [5] age > 30: n = 297
## |   |       (Intercept)     glucose
## |   |       -2.77095386  0.02353582
##
## Number of inner nodes:    2
## Number of terminal nodes: 3
## Number of parameters per node: 2
## Objective function (negative log-likelihood): 355.4578
```

```
# The latter. Note that logistic regression coefficients are a bit more tricky
# to interpret than linear regression coefficients. The intercept is lowest in
# the left node, highest in the right node. The probability of having diabetes
# varies accordingly.
#
# The slope is highest in the left node, and lowest in the right node, but
# these are log odds that have to be interpreted with respect to the intercept.
#
# I didn't do the calculations, but the effect of a unit increase of the glucose
# intolerance test result on the probability of having diabetes may well be equal
# in each of the terminal nodes (subgroups).
```