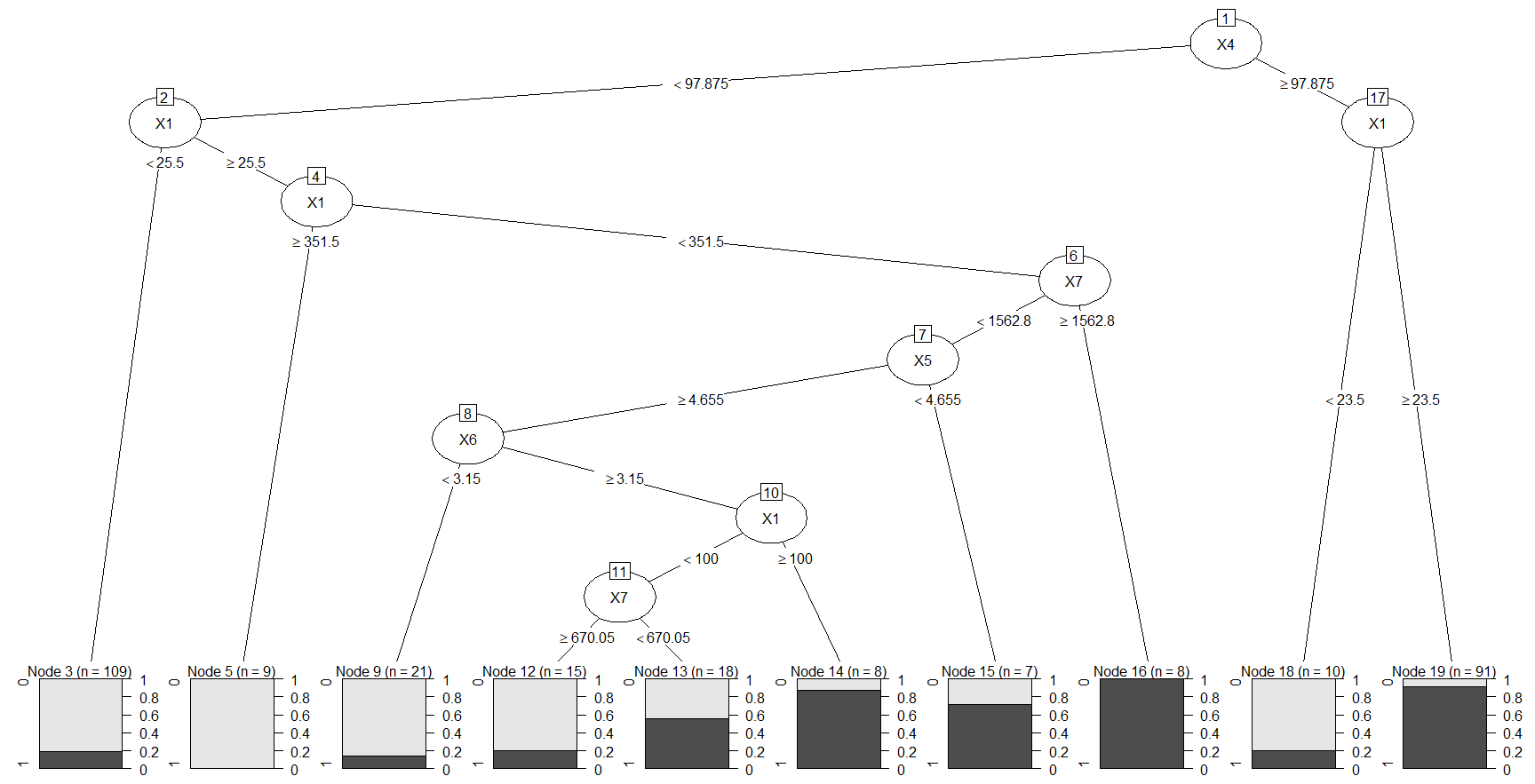
# Assessment brief

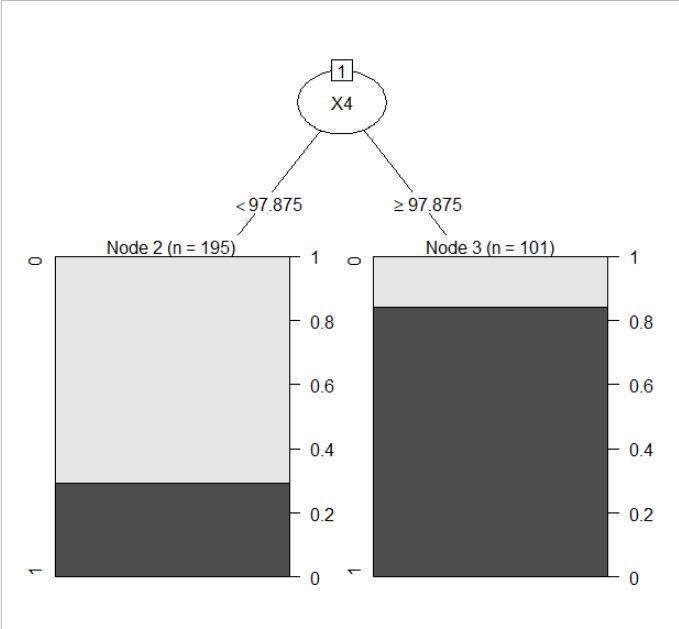
|  |  |
| --- | --- |
| Module Title: | Data Mining |
| Code: | B9DA103 |
| Module Leader: | Bahman Honari |
| Student code: | 10521647 |
| Student name: | Juliana Salvadori |
| Instruction for CA 1: | The provided Dataset contains 16 variables.  Response is the target variable to be predicted using the other variable, i.e. X1-X8 and Y1-Y8.  X variables are the actual measurements. Y variables are the categorical form of the X variables based of the recommended cut-off points by the previous studies. Group is the variable that might moderate the predictability of response by Xs or Ys.  The purpose of this study is to use “Decision Trees” to assess predictability of Response by X, Y or combination of them. As mentioned in above, this predictability might be affected by the Group variable, and therefore the analysis should be performed for all cases, as well as for each group separately to study the effect of variable Group. You would therefore need to check 9 models as  follow:  Using all data, run the models Response over Xs, Response over Ys, Response over Xs and Ys.  Repeat above for data with Group=0 and Group=1, separately. |

## Model 1 - DT model over Xs for all data

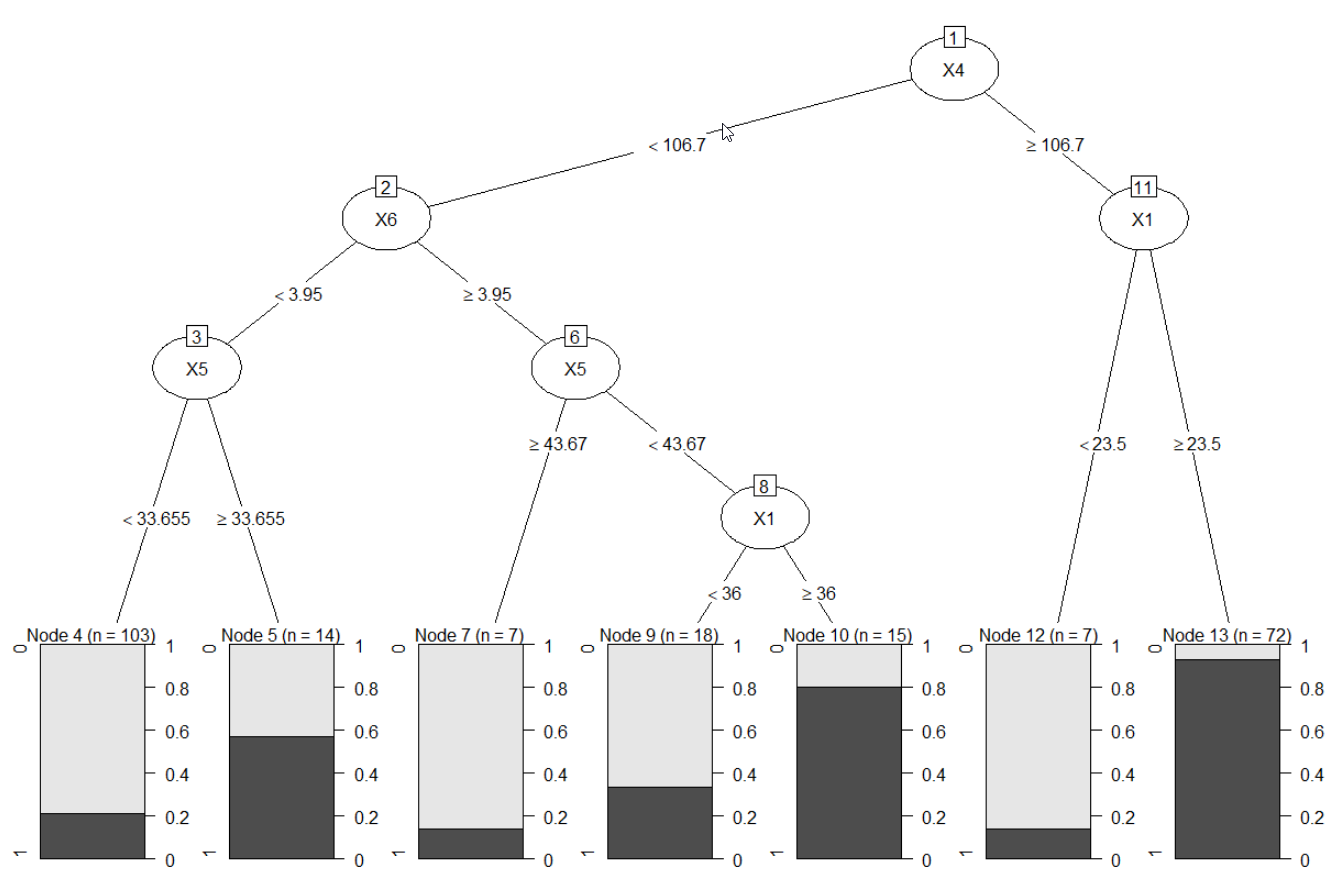
### Full tree



### After pruning



### Training model



Confusion matrix

> table\_matrix

pred

0 1

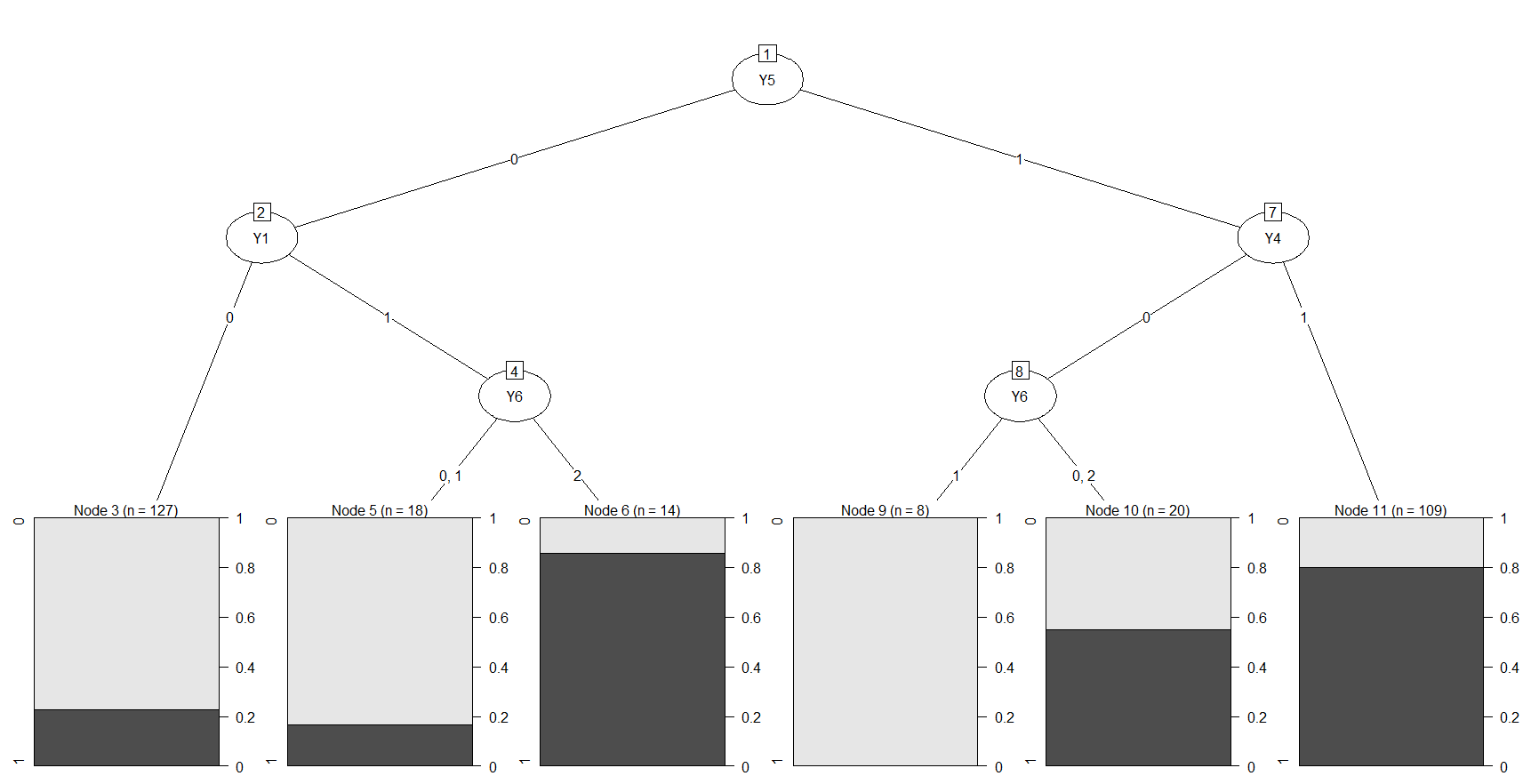
0 33 2

1 11 14

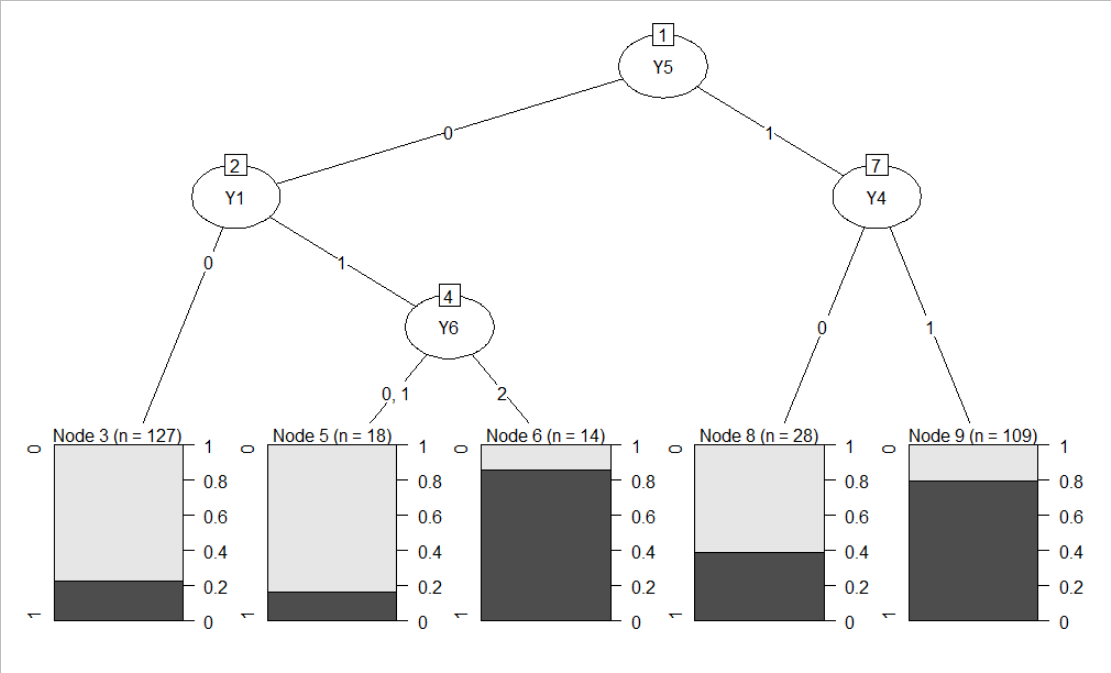
### Accuracy for model 1 = 0.783333333333333

## Model 2 - DT model over Ys for all data

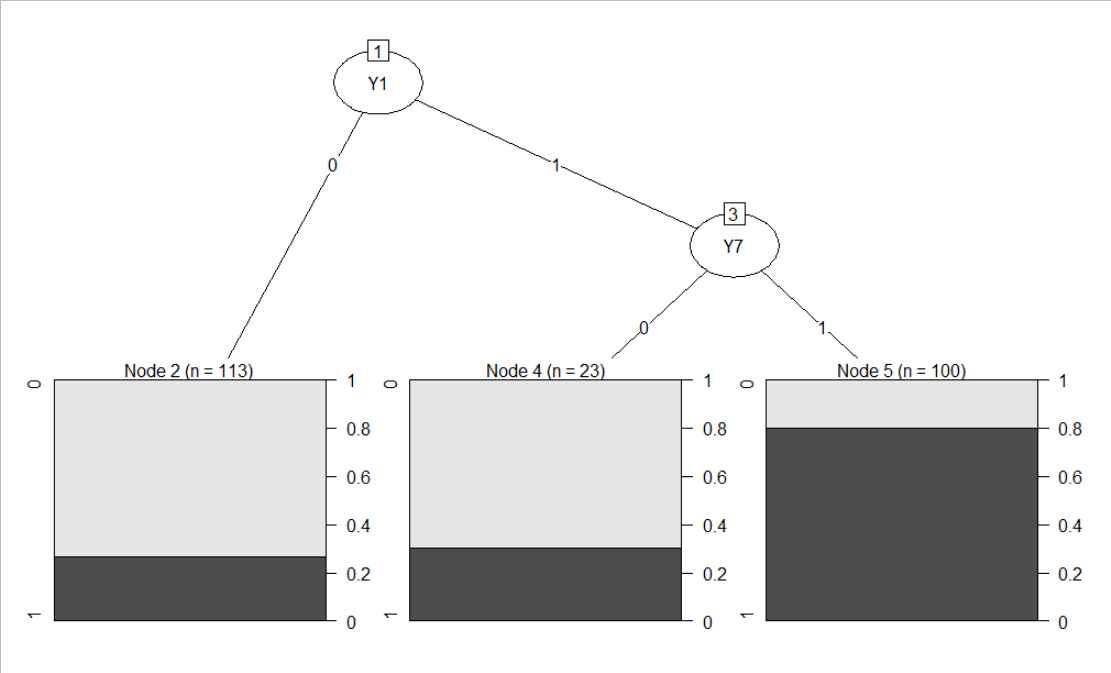
### Full tree



### After pruning



### Training model



Confusion matrix

> table\_matrix

pred

0 1

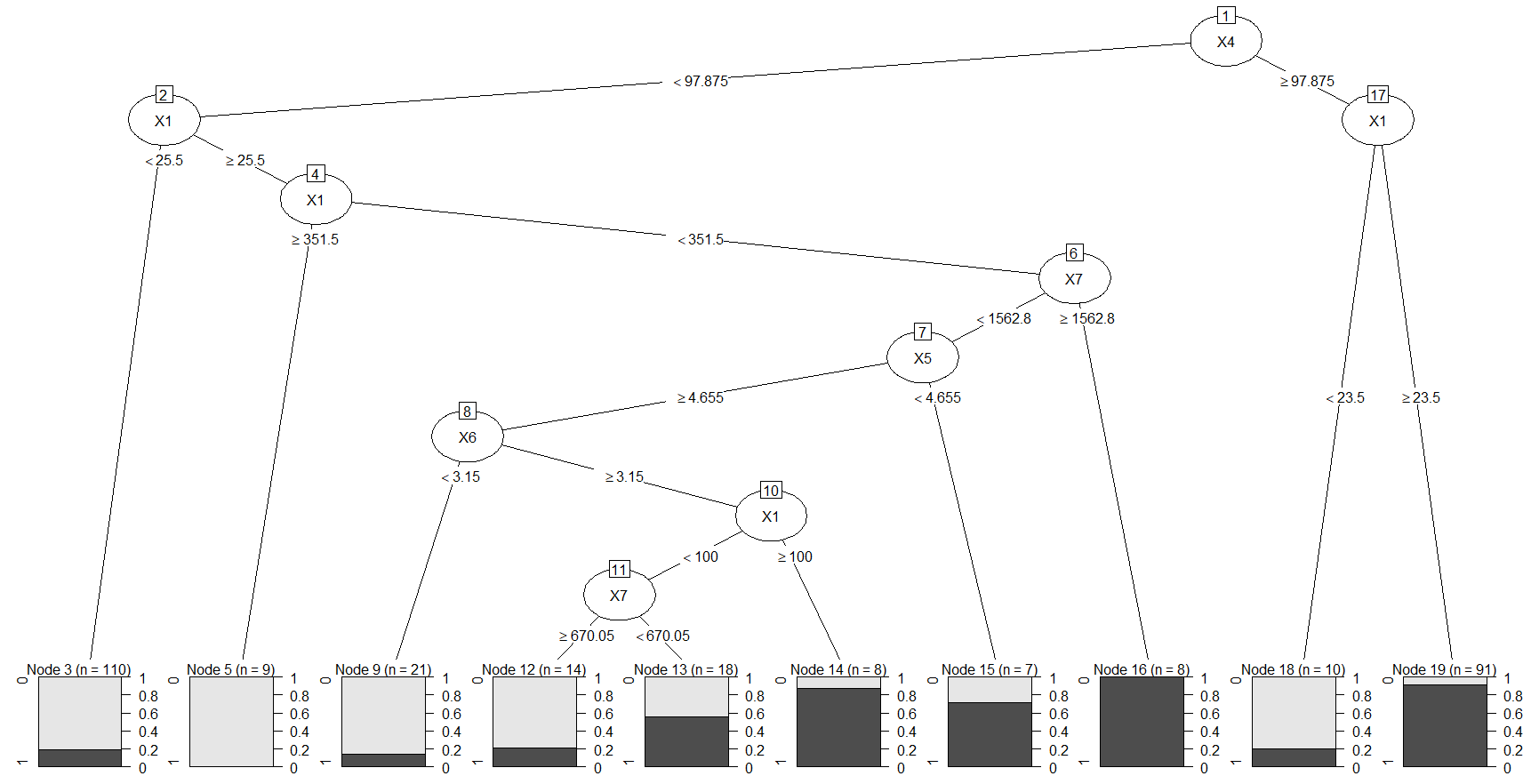
0 27 8

1 11 14

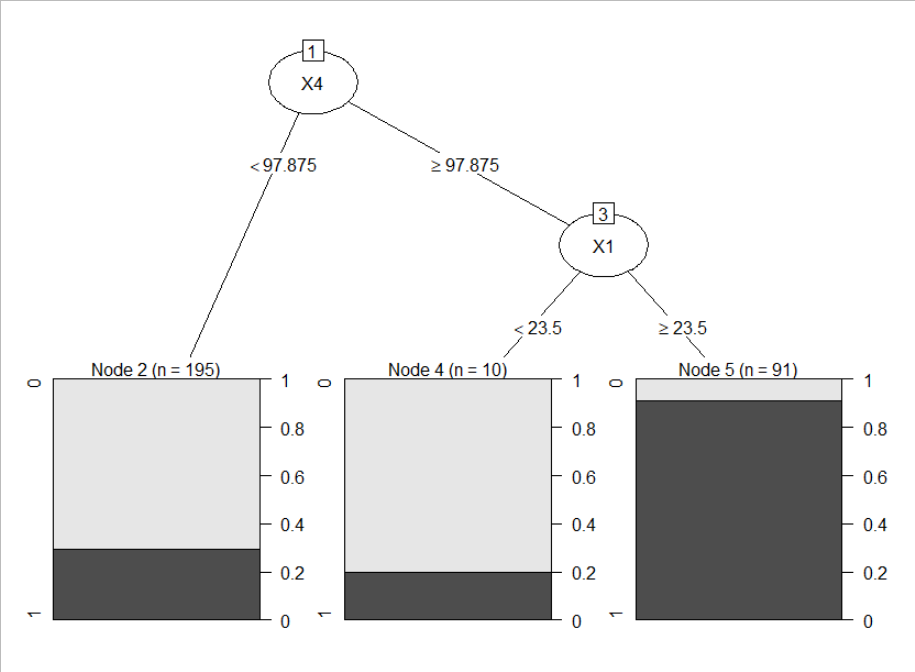
### Accuracy for model 2 = 0.683333333333333

## Model 3 - DT model over Xs and Ys for all data

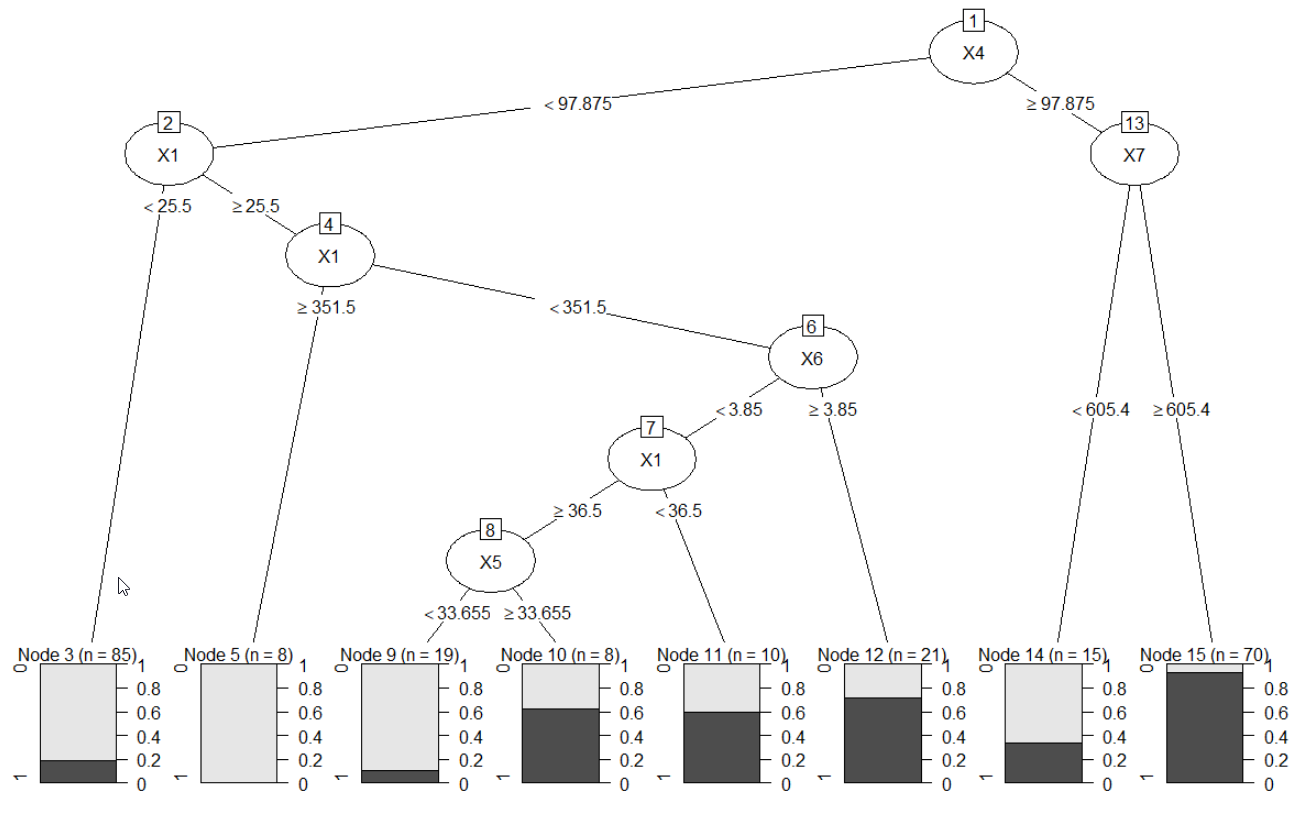
### Full tree



### After pruning



### Training model



Confusion matrix

> table\_matrix

pred

0 1

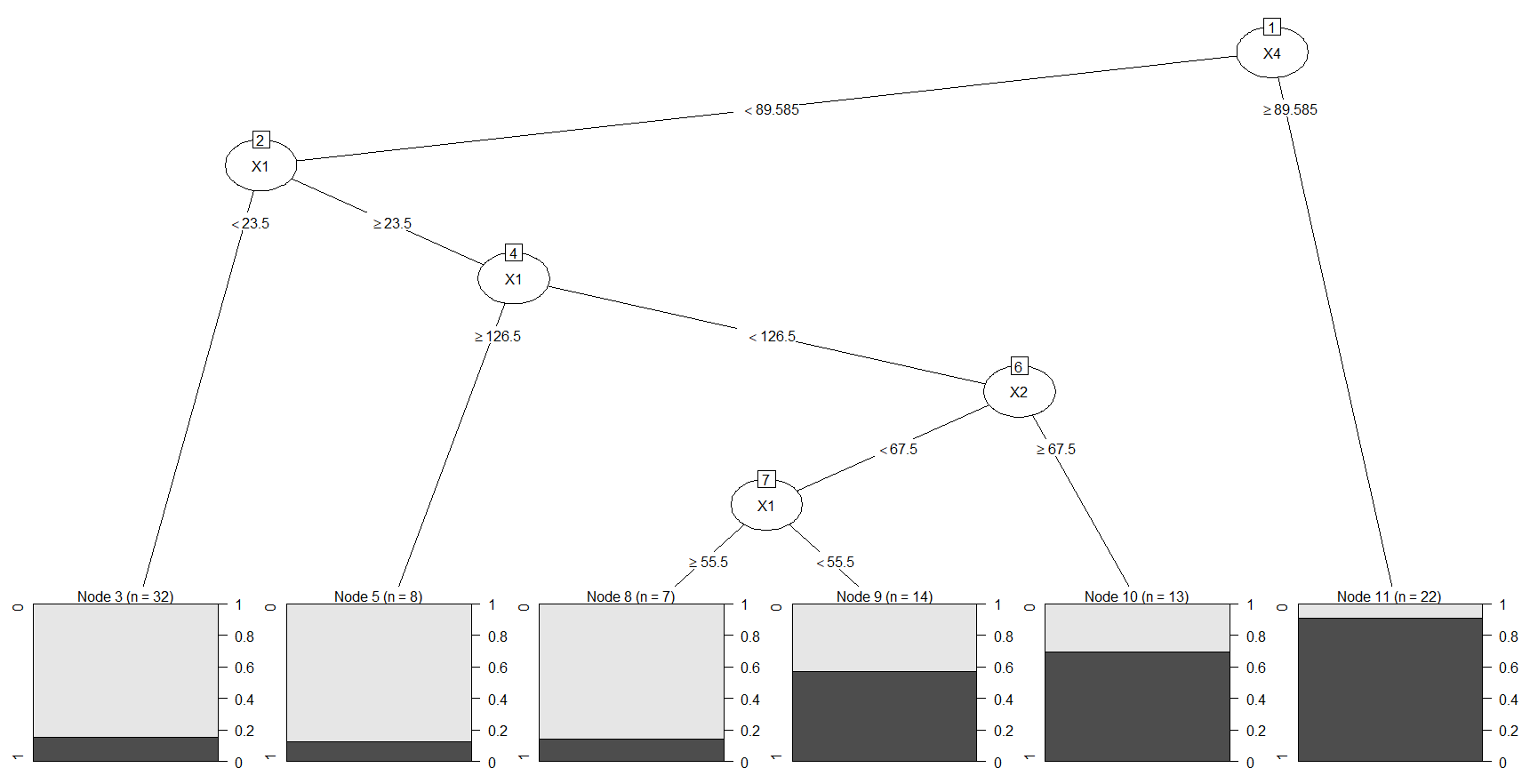
0 25 7

1 8 20

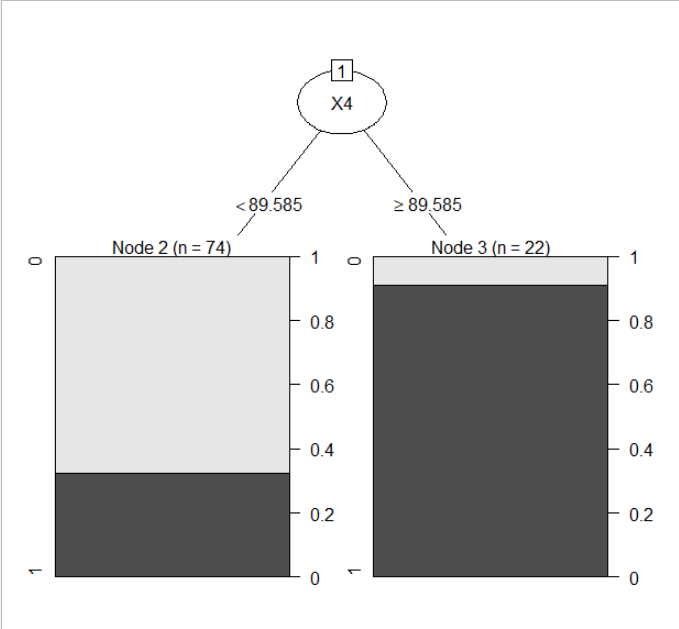
### Accuracy for model 3 = 0.75

## Model 4 - DT model over Xs for Group=0

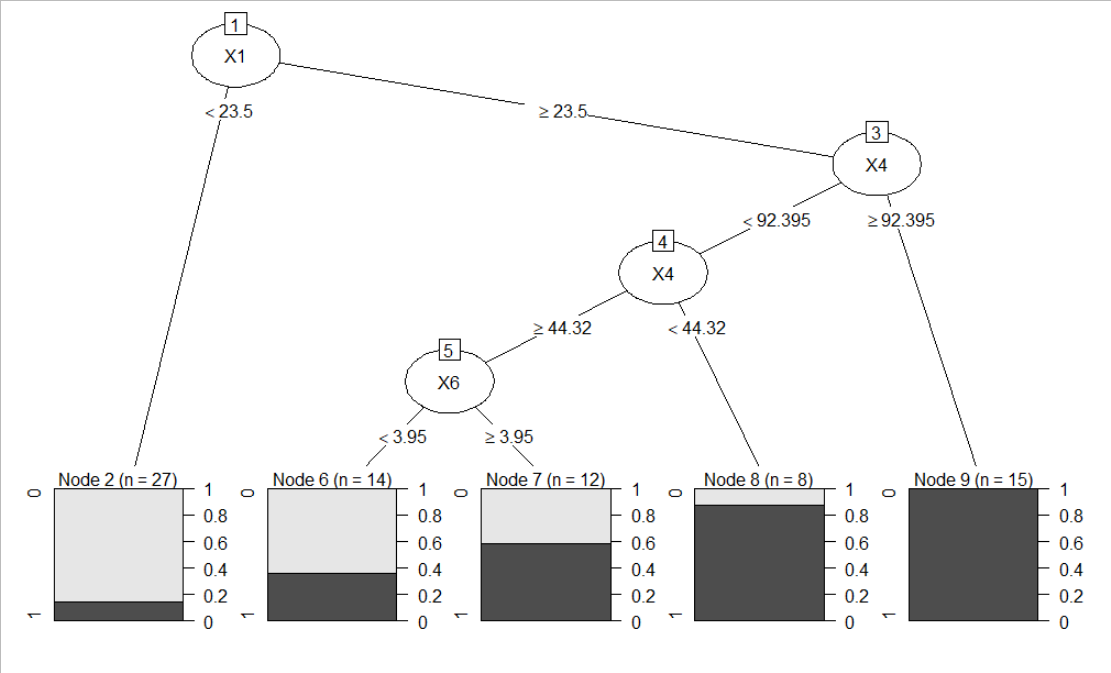
### Full tree



### After pruning



### Training model



Confusion matrix

> table\_matrix

pred

0 1

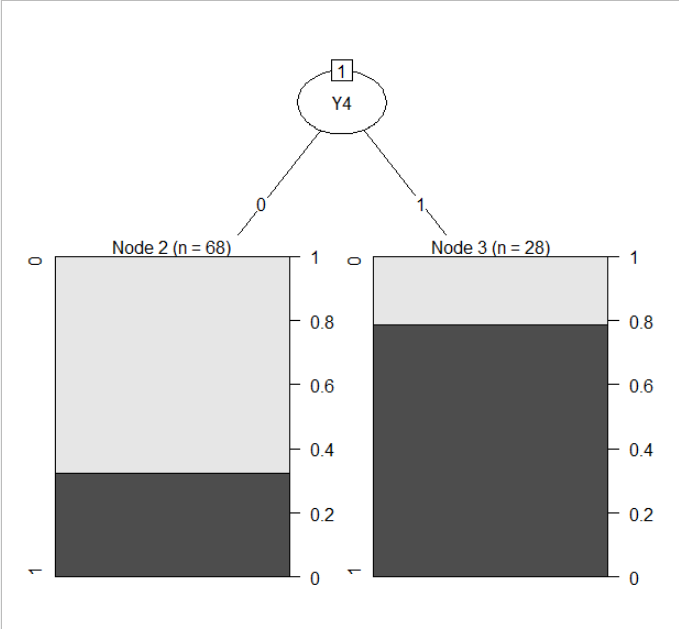
0 10 4

1 1 5

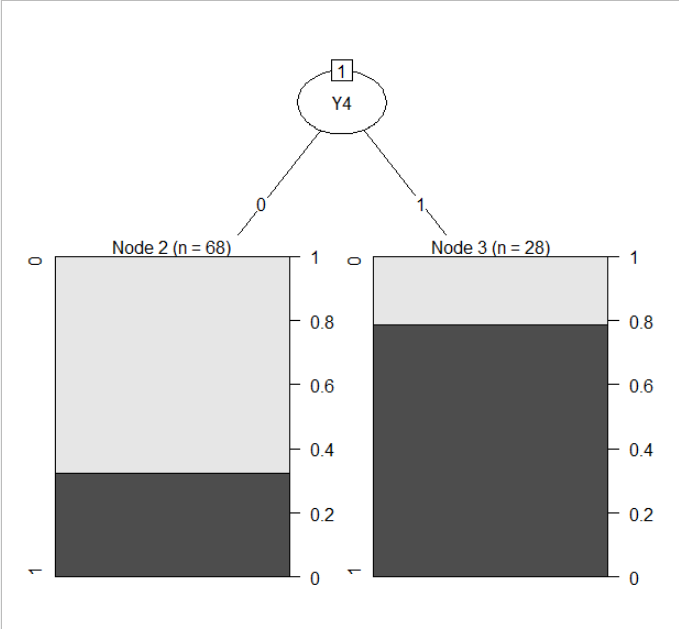
### Accuracy for model 4 = 0.75

## Model 5 - DT model over Ys for Group=0

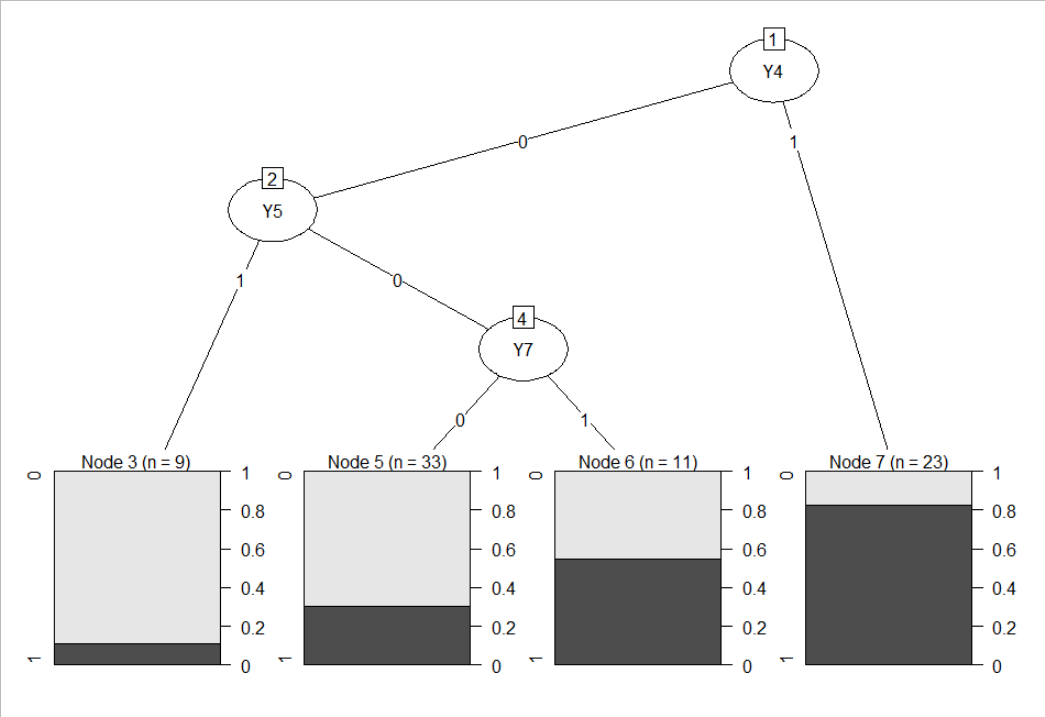
### Full tree



### After pruning



### Training model



Confusion matrix

> table\_matrix

pred

0 1

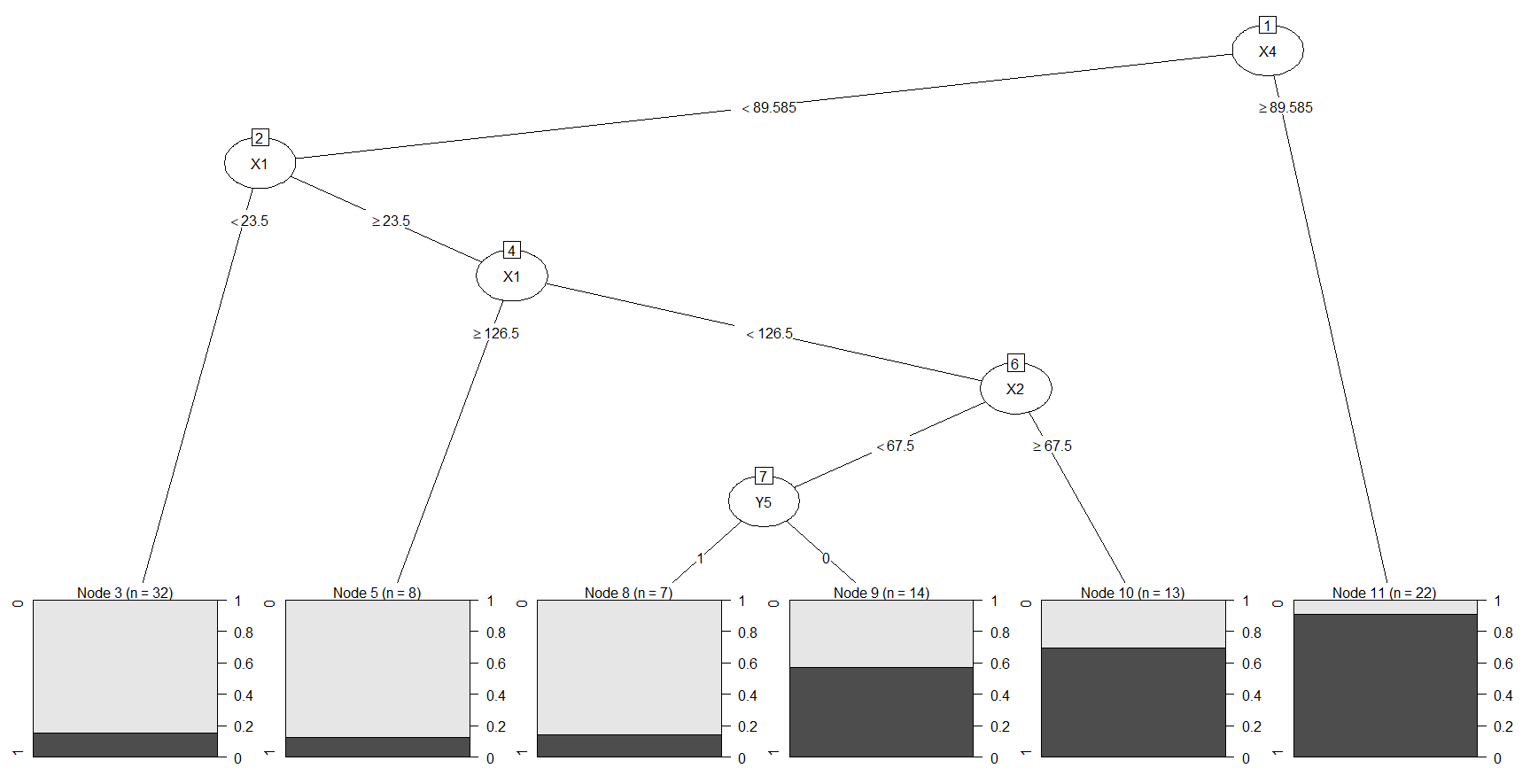
0 6 6

1 5 3

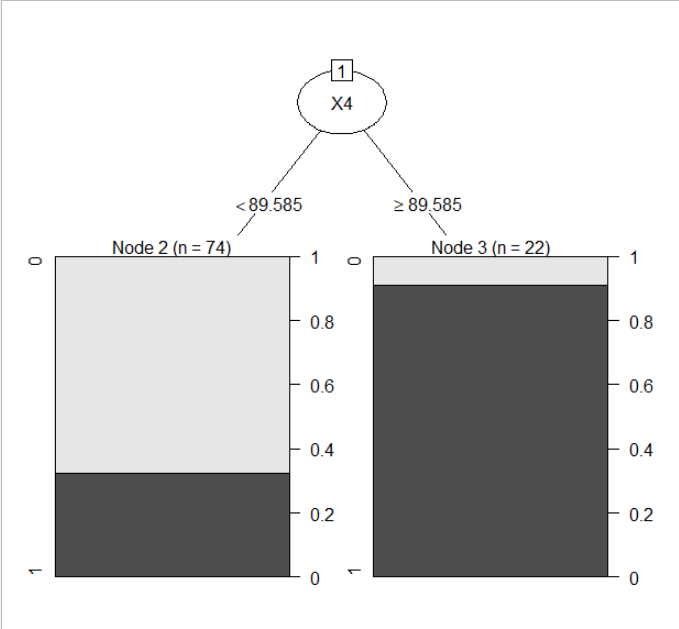
### Accuracy for model 5 = 0.45

## Model 6 - DT model over Xs and Ys for Group=0

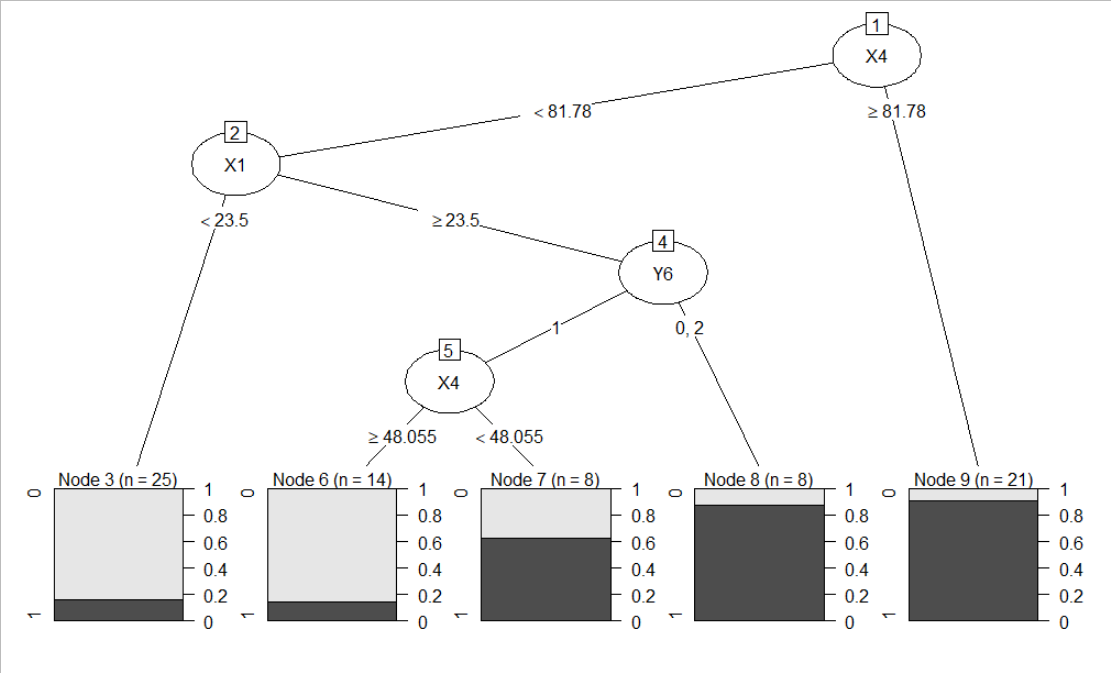
### Full tree



### After pruning



### Training model



Confusion matrix

> table\_matrix

pred

0 1

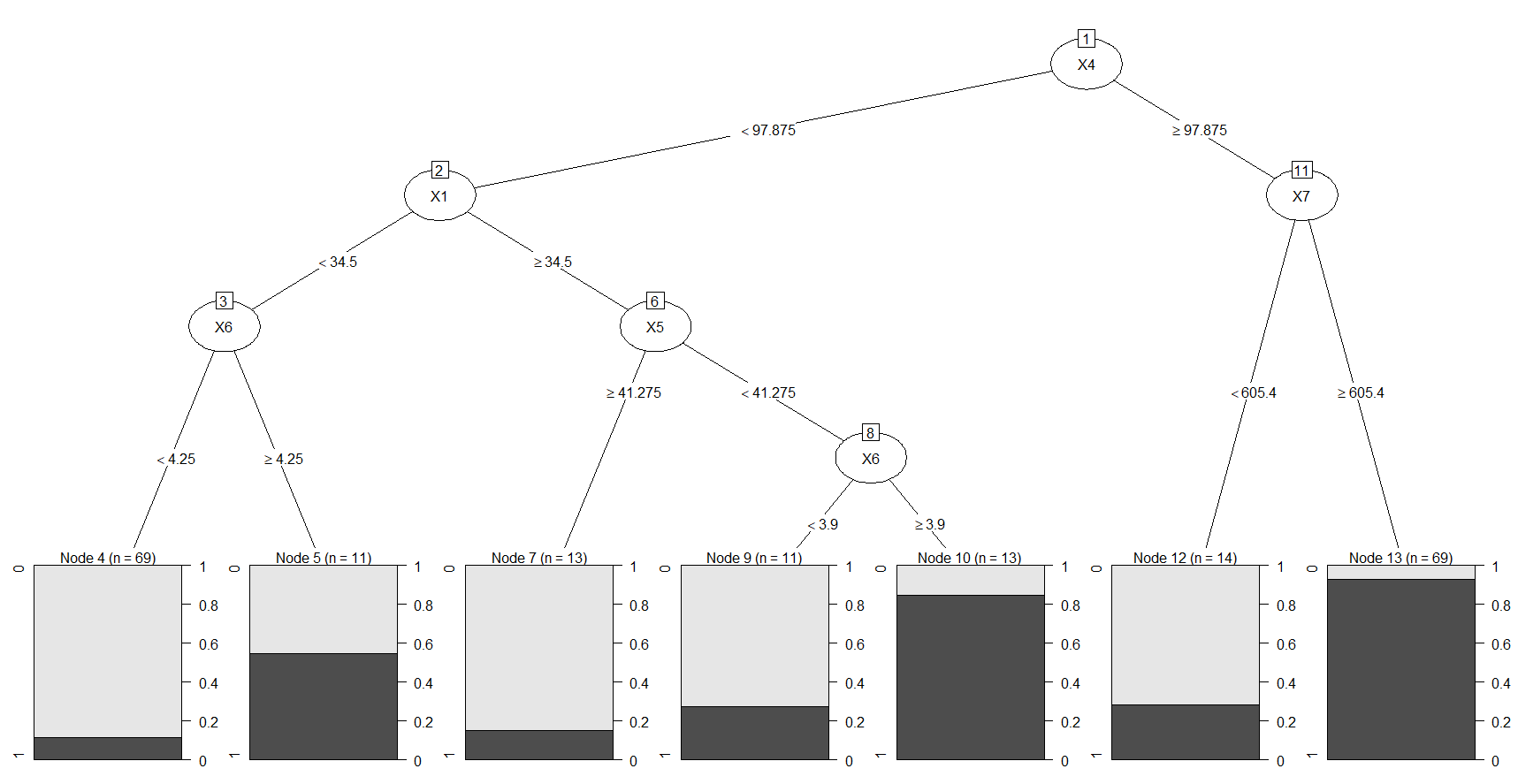
0 7 6

1 1 6

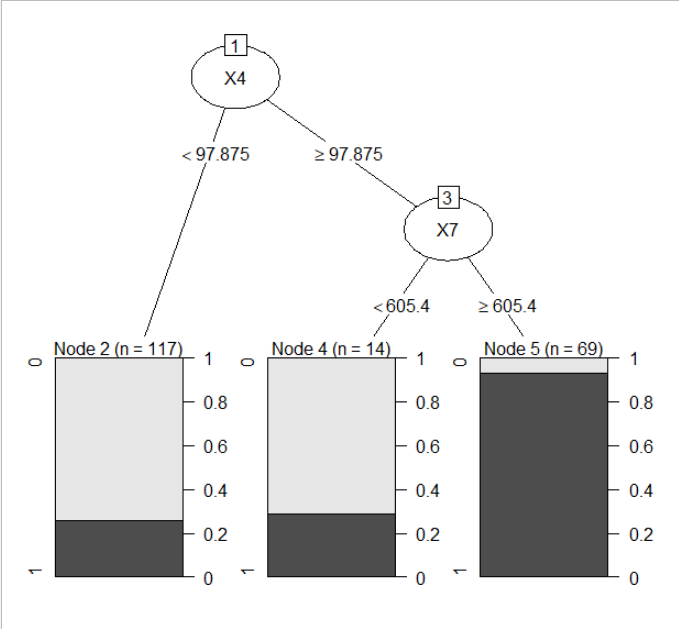
### Accuracy for model 6 = 0.65

## Model 7 - DT model over Xs for Group=1

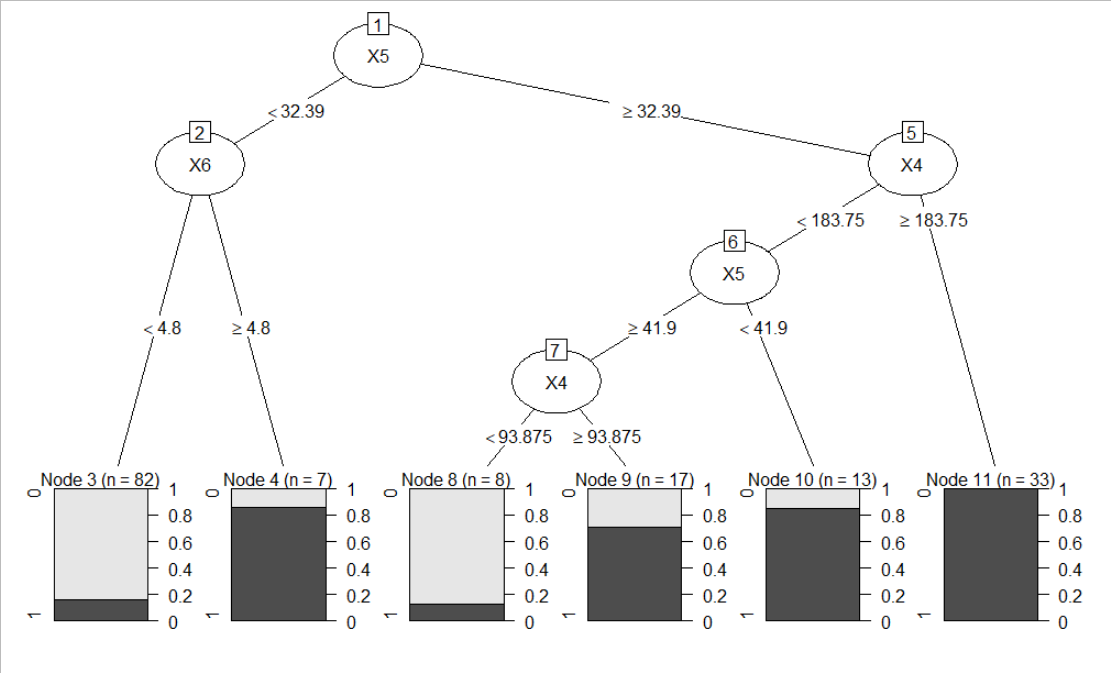
### Full tree



### After pruning



### Training model



Confusion matrix

> table\_matrix

pred

0 1

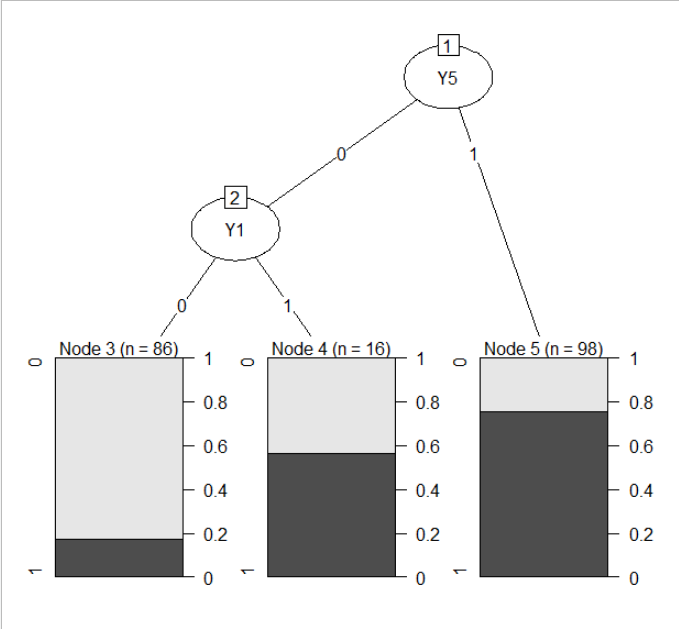
0 17 1

1 7 15

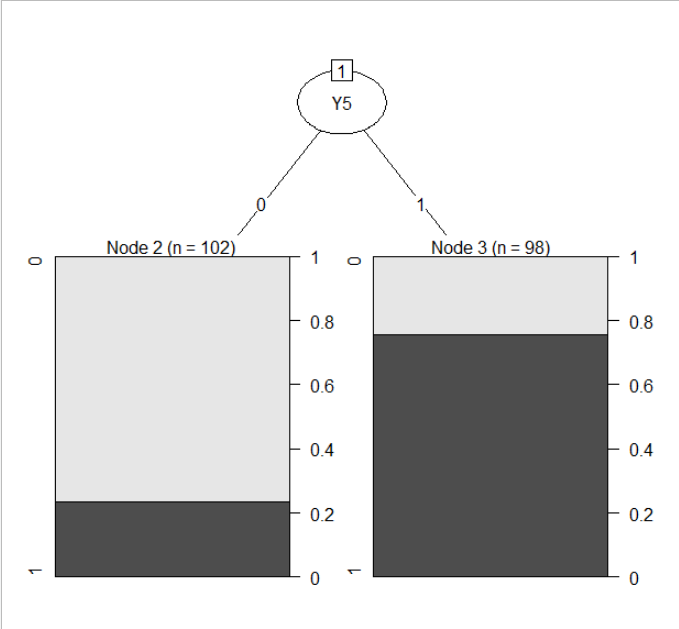
### Accuracy for model 7 = 0.8

## Model 8 - DT model over Ys for Group=1

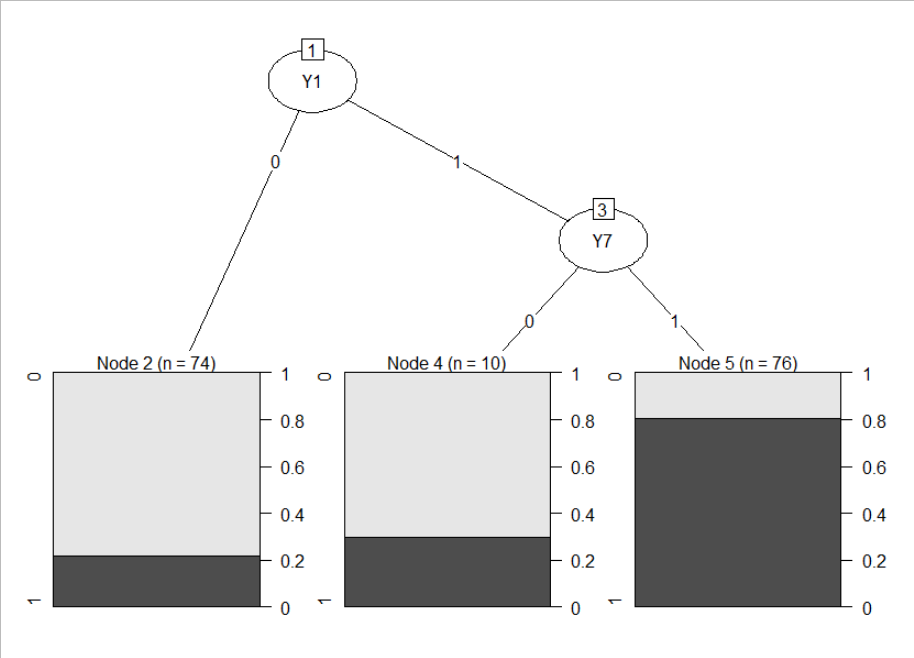
### Full tree



### After pruning



### Training model



Confusion matrix

> table\_matrix

pred

0 1

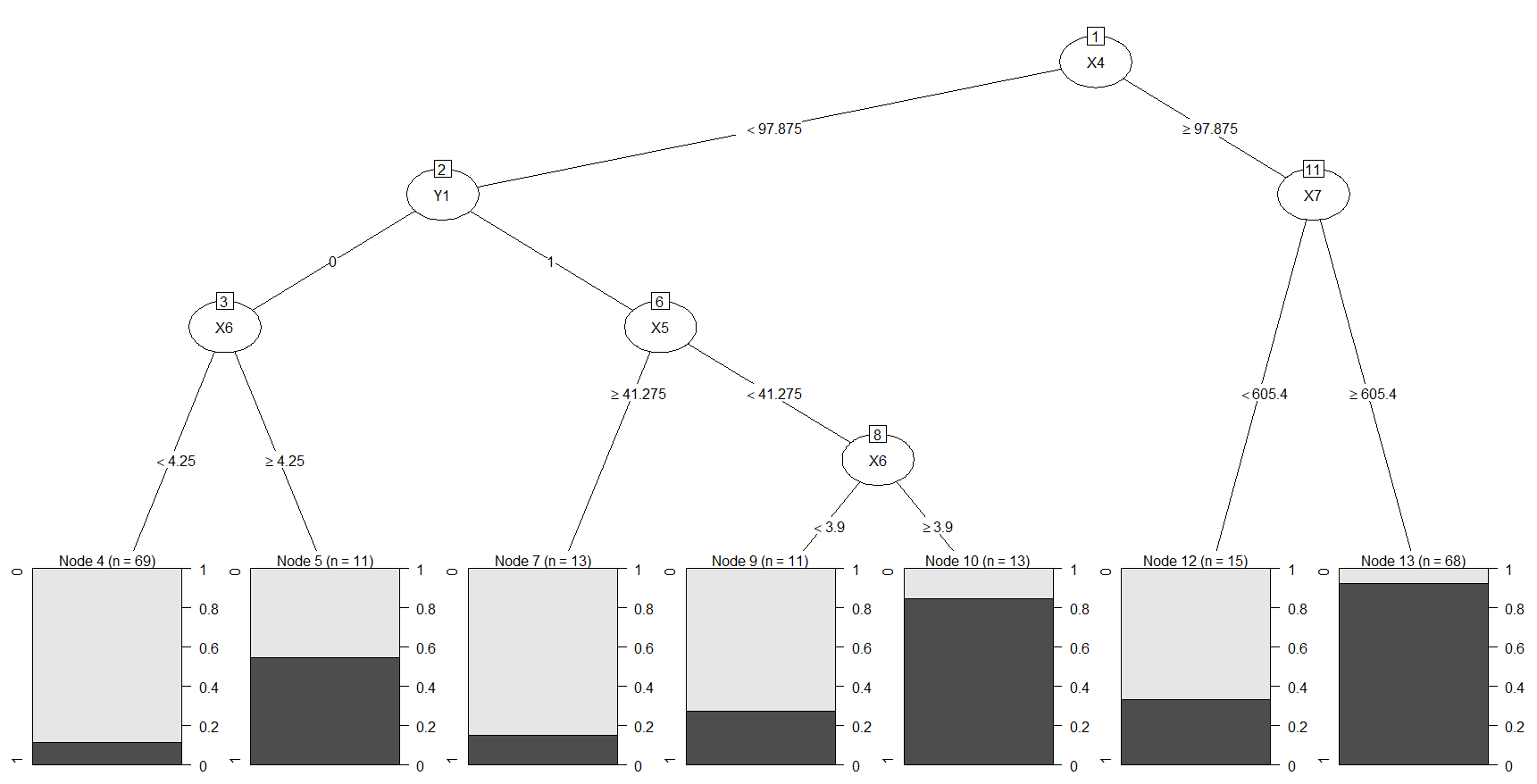
0 18 4

1 9 9

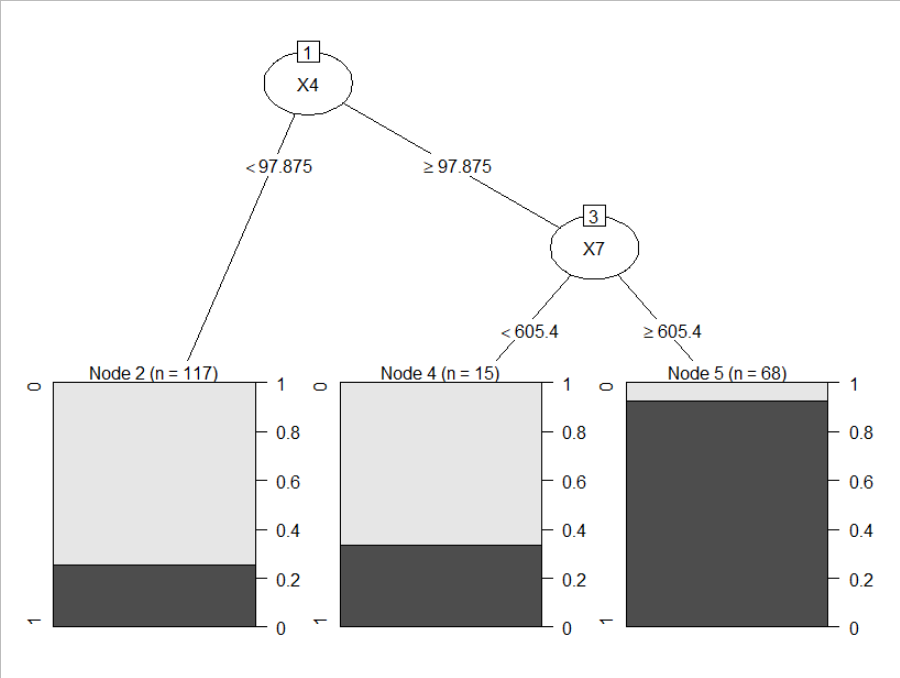
### Accuracy for model 8 = 0.675

## Model 9 - DT model over Xs and Ys for Group=1

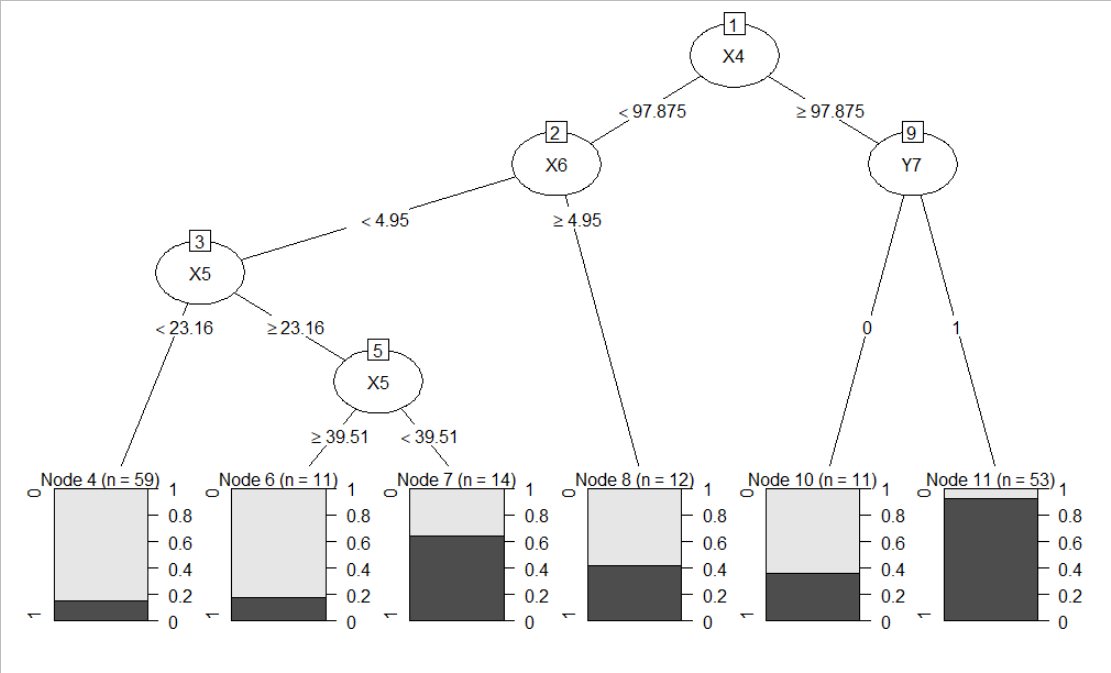
### Full tree



### After pruning



### Training model



Confusion matrix

> table\_matrix

pred

0 1

0 13 7

1 4 16

### Accuracy for model 9 = 0.725

# Summary

Based on the results (accuracy) obtained on the models described above it’s possible to conclude that the Xs variables are better predictors than the Ys variables or the combination of Xs and Ys variables. In all datasets analysed, the accuracy found on the models where only the X variables were used was always greater than the other models. Also, it’s possible to notice that the use of the Group variable can affect the predictability as the greatest accuracy was found when using Xs variables and Group = 1 (model 7).