HR Attrition Analysis

Veerasak Kritsanapraphan

4/25/2018

HR Attrition Analysis

```
hremployee <- read.csv("WA_Fn-UseC_-HR-Employee-Attrition.csv",
                       header=TRUE)
```

Sampling

```
library(fifer)
```

```
## Loading required package: MASS
```

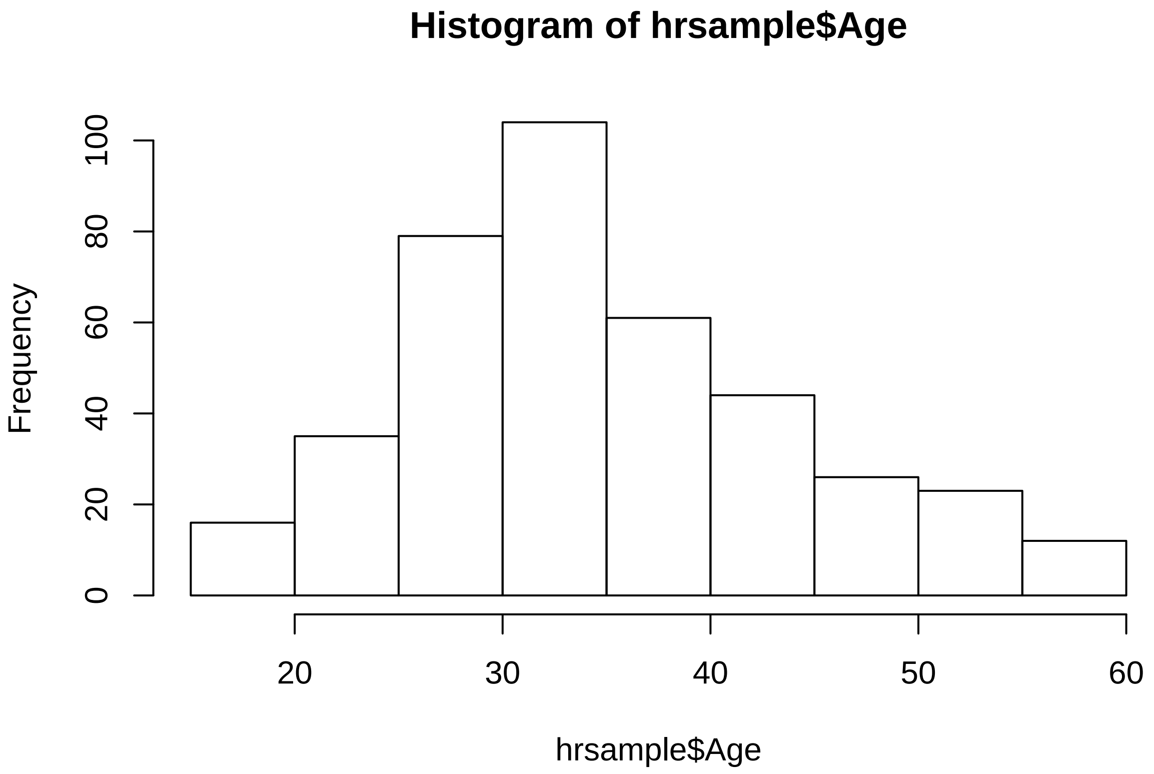
```
yeshr <- stratified(hremployee, "Attrition", 200,
           select = list(Attrition = c("Yes")))
nohr <- stratified(hremployee, "Attrition", 200,
                    select = list(Attrition = c("No")))
hrsample <- rbind(yeshr,nohr)

set.seed(1234)
ind <- sample(2, nrow(hrsample), replace=TRUE,
              prob=c(0.6,0.4))
trainData <- hrsample[ind==1,]
testData <- hrsample[ind==2,]
str(hrsample)
```

```
## 'data.frame':    400 obs. of  35 variables:
##  $ Age                     : int  28 34 49 20 24 28 34 29 21 19 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 2 3 3 3 2 3
##  $ DailyRate               : int  1434 699 1475 1097 693 654 296 341 337 303 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 2 2 2 2 3 2 3 3 3 2 ...
##  $ DistanceFromHome        : int  5 6 28 11 3 1 6 1 7 2 ...
##  $ Education               : int  4 1 2 3 2 2 2 3 1 3 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 6 4 2 4 2 2 3 4 3 2 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  65 31 1420 1016 720 741 555 896 1780 243 ...
##  $ EnvironmentSatisfaction : int  3 2 1 4 1 1 4 2 2 2 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 2 2 2 1 1 1 1 1 2 2 ...
##  $ HourlyRate              : int  50 83 97 98 65 67 33 48 31 47 ...
##  $ JobInvolvement          : int  3 3 2 2 3 1 1 2 3 2 ...
##  $ JobLevel                : int  1 1 2 1 2 1 1 1 1 1 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 3 7 3 7 8 7 9 9 9 3
##  $ JobSatisfaction         : int  3 1 1 1 3 2 3 3 2 4 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 3 3 3 3 3 1 1 3 3 ...
##  $ MonthlyIncome           : int  3441 2960 4284 2600 4577 2216 2351 2800 2679 1102 ...
##  $ MonthlyRate             : int  11179 17102 22710 18275 24785 3872 12253 23522 4567 9241 ...
##  $ NumCompaniesWorked      : int  1 2 3 1 9 7 0 6 1 1 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 2 1 2 1 1 ...
##  $ PercentSalaryHike       : int  13 11 20 15 14 13 16 19 13 22 ...
##  $ PerformanceRating       : int  3 3 4 3 3 3 3 3 3 4 ...
##  $ RelationshipSatisfaction: int  3 3 1 1 1 4 4 3 2 3 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  0 0 0 0 0 0 1 3 0 0 ...
##  $ TotalWorkingYears       : int  2 8 20 1 4 10 3 5 1 1 ...
##  $ TrainingTimesLastYear   : int  3 2 2 2 3 4 3 3 3 3 ...
##  $ WorkLifeBalance         : int  2 3 3 3 3 3 2 3 3 2 ...
##  $ YearsAtCompany          : int  2 4 4 1 2 7 2 3 1 1 ...
##  $ YearsInCurrentRole      : int  2 2 3 0 2 7 2 2 0 0 ...
##  $ YearsSinceLastPromotion : int  2 1 1 0 2 3 1 0 1 1 ...
##  $ YearsWithCurrManager    : int  2 3 3 0 0 7 0 2 0 0 ...
```

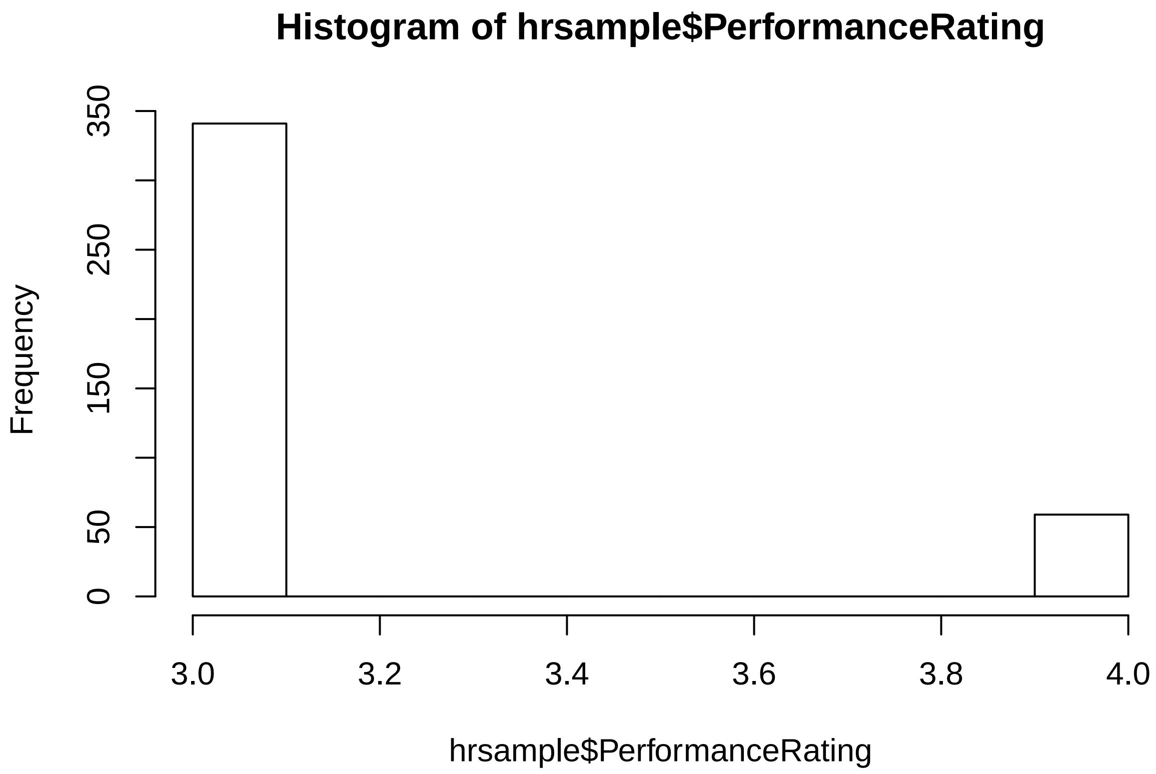
```
table(hrsample$JobSatisfaction)
```

```
## 
##   1   2   3   4 
##  91  77 121 111
```

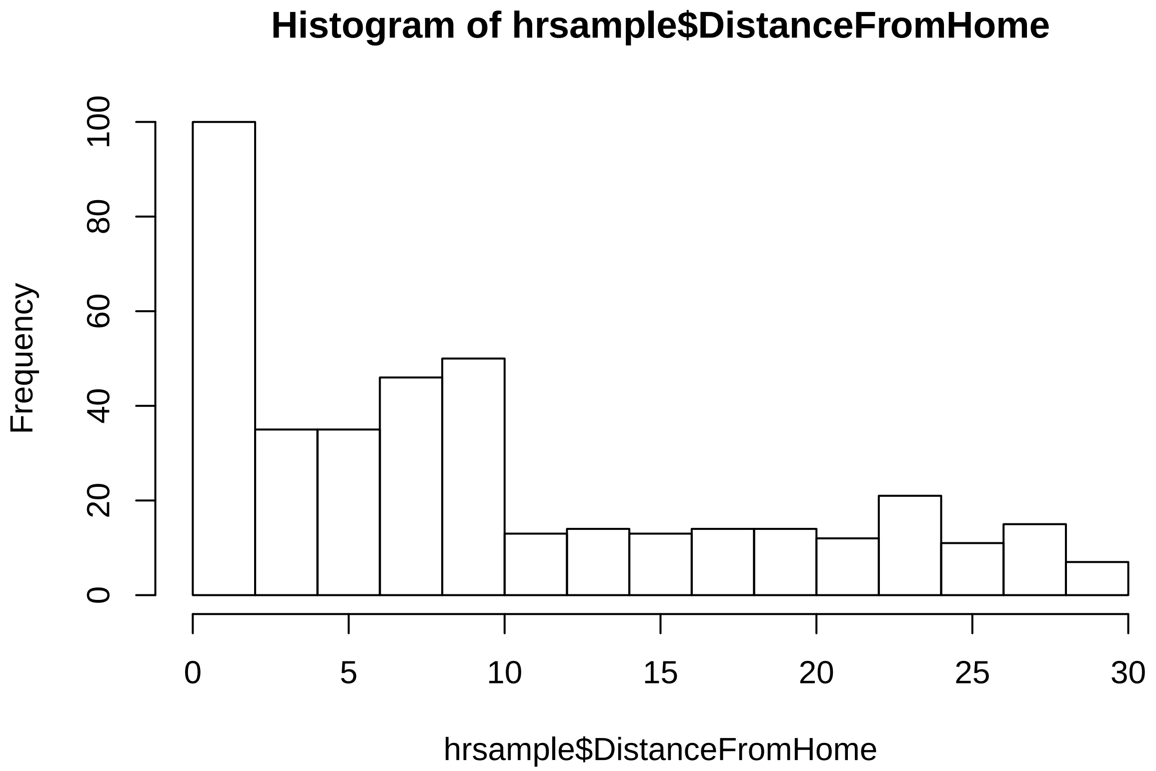
```
hist(hrsample$Age)
```



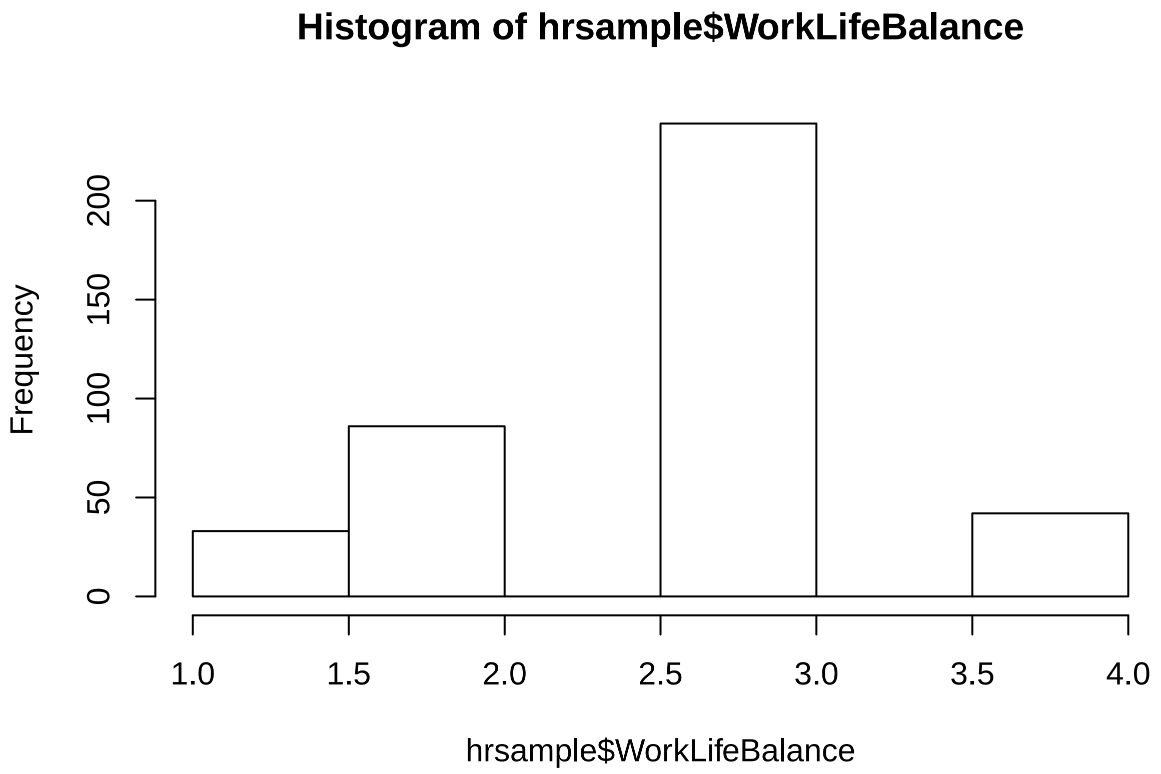
```
hist(hrsample$PerformanceRating)
```



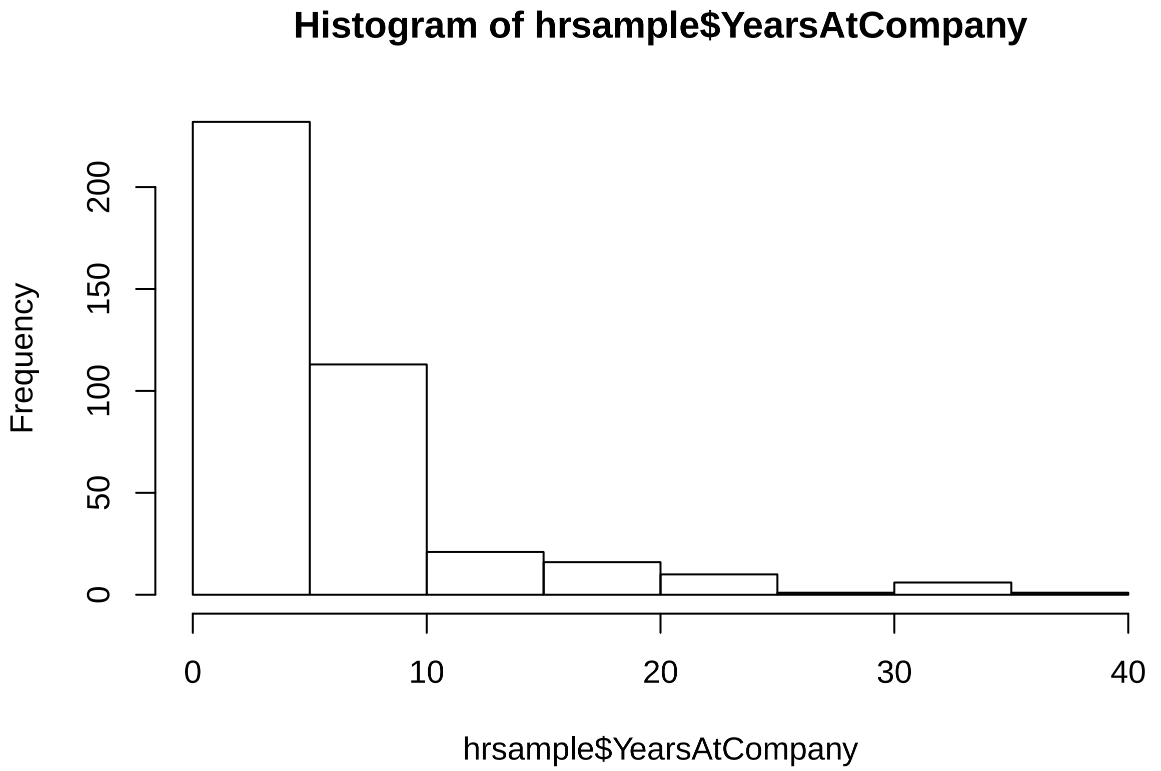
```r
hist(hrsample$DistanceFromHome)
```



```r
hist(hrsample$WorkLifeBalance)
```



```
hist(hrsample$YearsAtCompany)
```



```
#myformula <- Attrition ~ .
myformula <- Attrition ~ JobSatisfaction +
                         Age + PerformanceRating +
                         DistanceFromHome +
                         WorkLifeBalance +
                         YearsAtCompany
table(trainData$Attrition)
```

```
##
##  No Yes
## 123 126
```

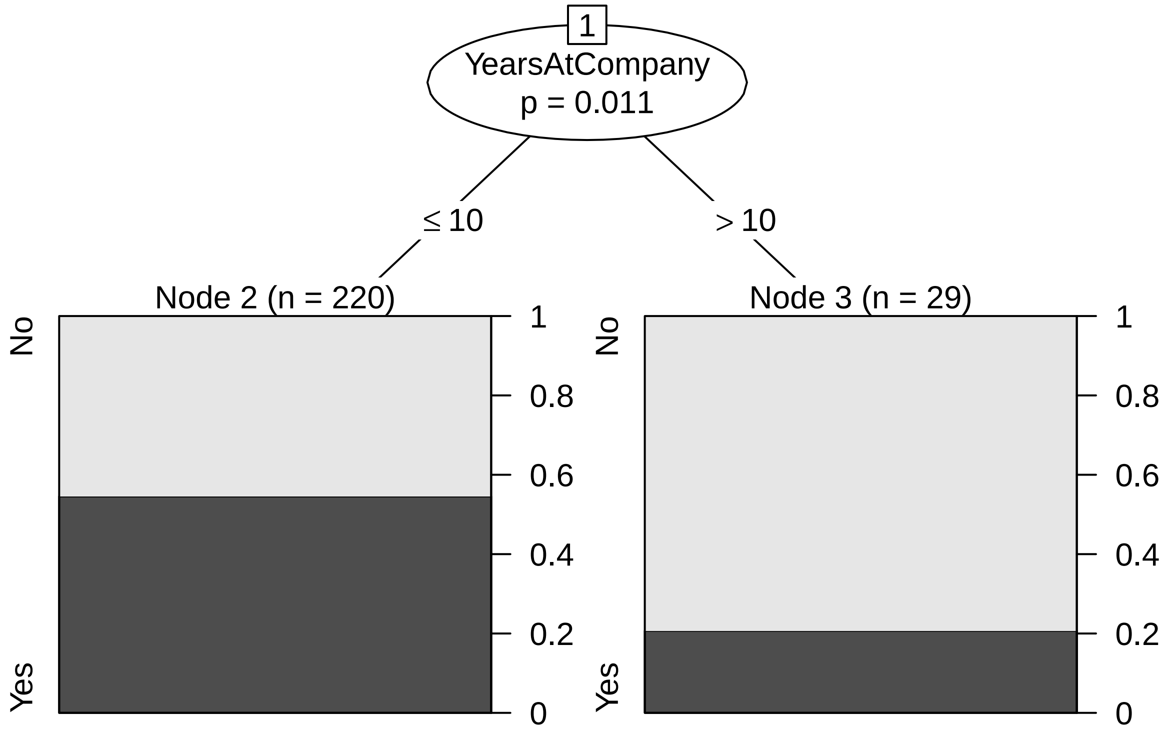
```
table(testData$Attrition)
```

```
##
##  No Yes
##  77  74
```

```
library(party)
```

```
## Warning: package 'party' was built under R version 3.4.3

## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich
```

```
ctree_model <- ctree(myformula, data=trainData)
plot(ctree_model)
```



```r
testpred <- predict(ctree_model,newdata=testData)
table(testpred,
      testData$Attrition)
```

```
##
## testpred No Yes
##      No  16  10
##      Yes 61  64
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.3

## Loading required package: lattice

## Loading required package: ggplot2

## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2018c.
## 1.0/zoneinfo/Asia/Bangkok'
```

```r
confusionMatrix(testpred, testData$Attrition)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  16  10
##        Yes 61  64
##
##                Accuracy : 0.5298
##                  95% CI : (0.447, 0.6114)
##     No Information Rate : 0.5099
##     P-Value [Acc > NIR] : 0.3422
##
```

```
##                    Kappa : 0.0717
##  Mcnemar's Test P-Value : 2.958e-09
##
##              Sensitivity : 0.2078
##              Specificity : 0.8649
##           Pos Pred Value : 0.6154
##           Neg Pred Value : 0.5120
##               Prevalence : 0.5099
##           Detection Rate : 0.1060
##     Detection Prevalence : 0.1722
##        Balanced Accuracy : 0.5363
##
##         'Positive' Class : No
##
```

```r
myformula <- Attrition ~ .
table(trainData$Attrition)
```

```
##
##  No Yes
## 123 126
```

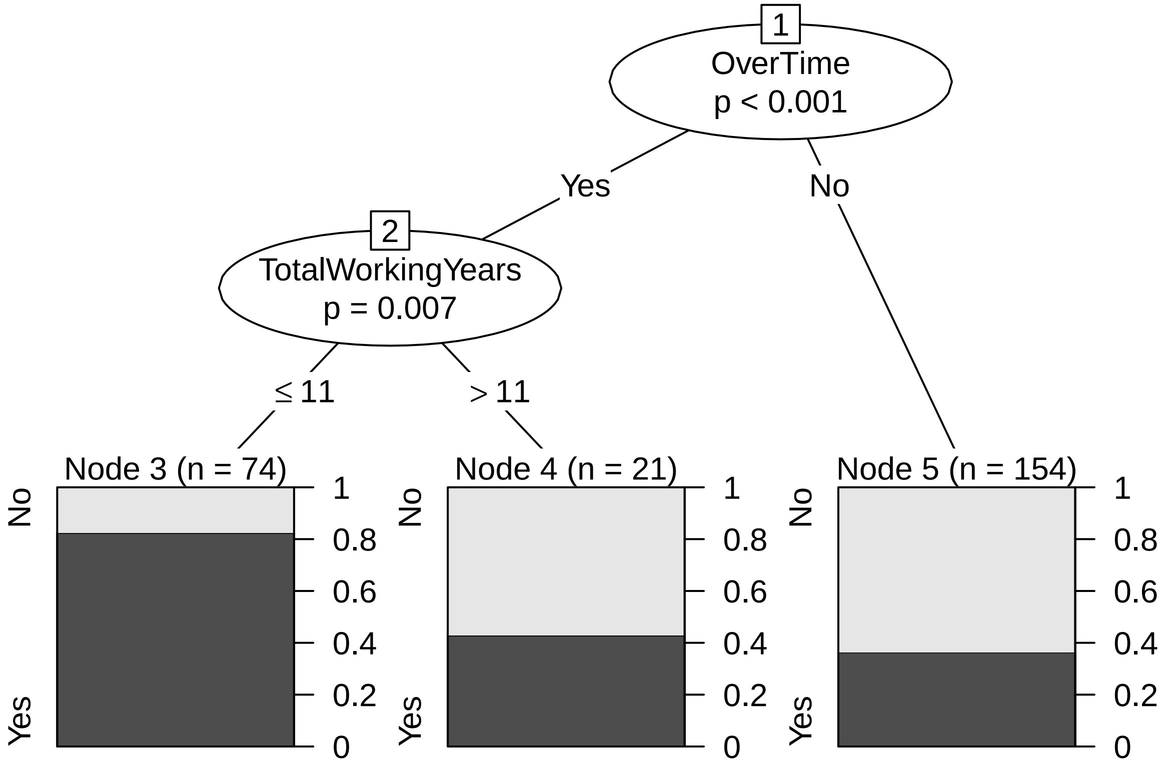
```r
table(testData$Attrition)
```

```
##
##  No Yes
##  77  74
```

```r
library(party)
ctree_model <- ctree(myformula, data=trainData)
```

```
## Warning in factor_trafo(x): factors at only one level may lead to problems
```

```r
plot(ctree_model)
```



```
testpred <- predict(ctree_model,newdata=testData)
```

```
## Warning in factor_trafo(x): factors at only one level may lead to problems
```

```
table(testpred,
      testData$Attrition)
```

```
##
## testpred No Yes
##      No  71  42
##      Yes  6  32
```

```
library(caret)

confusionMatrix(testpred, testData$Attrition)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  71  42
##        Yes  6  32
##
##                Accuracy : 0.6821
##                  95% CI : (0.6015, 0.7554)
##     No Information Rate : 0.5099
##     P-Value [Acc > NIR] : 1.338e-05
##
##                   Kappa : 0.3579
##  Mcnemar's Test P-Value : 4.376e-07
##
```

```
##             Sensitivity : 0.9221
##             Specificity : 0.4324
##          Pos Pred Value : 0.6283
##          Neg Pred Value : 0.8421
##              Prevalence : 0.5099
##          Detection Rate : 0.4702
##    Detection Prevalence : 0.7483
##       Balanced Accuracy : 0.6773
##
##        'Positive' Class : No
##
```