Data Mining Practice - Decision Tree

1. Install packages - rpart

```
install.packages("rpart")
```

```
library(rpart)
```

2. Growing Trees

2.1. Example : kyphosis

```
data(kyphosis, package = "rpart")
# help(kyphosis)

str(kyphosis)
```

```
## 'data.frame':    81 obs. of  4 variables:
##  $ Kyphosis: Factor w/ 2 levels "absent","present": 1 1 2 1 1 1 1 1 1 2
 ...
##  $ Age     : int  71 158 128 2 1 1 61 37 113 59 ...
##  $ Number  : int  3 3 4 5 4 2 2 3 2 6 ...
##  $ Start   : int  5 14 5 1 15 16 17 16 16 12 ...
```

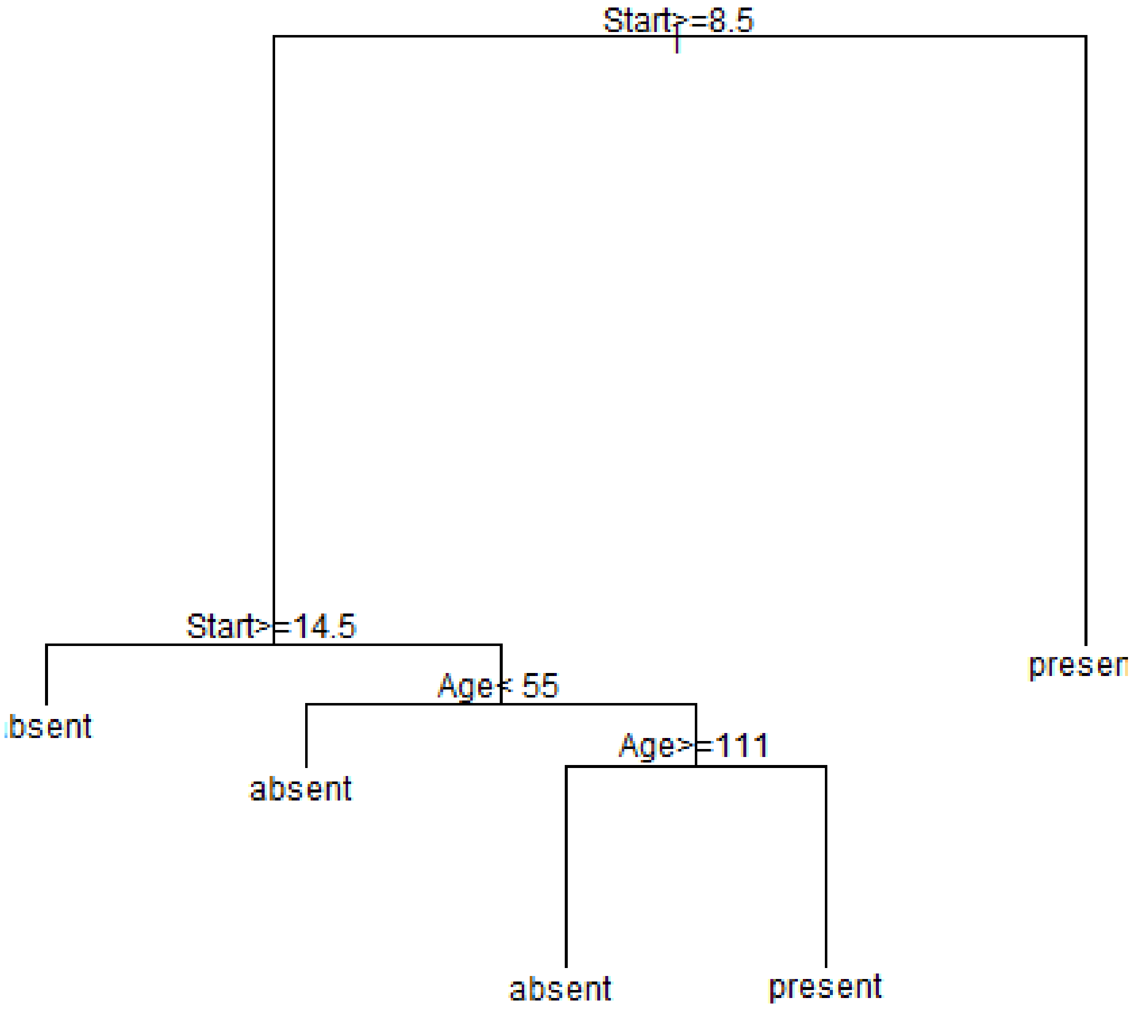
```
# 나무 모형 생성
kyphosis.tr <- rpart(Kyphosis ~ ., data = kyphosis)

print(kyphosis.tr)
```

```
## n= 81
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 81 17 absent (0.79012 0.20988)
##    2) Start>=8.5 62  6 absent (0.90323 0.09677)
##      4) Start>=14.5 29  0 absent (1.00000 0.00000) *
##      5) Start< 14.5 33  6 absent (0.81818 0.18182)
##       10) Age< 55 12  0 absent (1.00000 0.00000) *
##       11) Age>=55 21  6 absent (0.71429 0.28571)
##         22) Age>=111 14  2 absent (0.85714 0.14286) *
##         23) Age< 111 7  3 present (0.42857 0.57143) *
##    3) Start< 8.5 19  8 present (0.42105 0.57895) *
```

```
attributes(kyphosis.tr)
```

```
## $names
##  [1] "frame"               "where"               "call"
##  [4] "terms"               "cptable"             "method"
##  [7] "parms"               "control"             "functions"
## [10] "numresp"             "splits"              "variable.importance"
## [13] "y"                   "ordered"
##
## $xlevels
## named list()
##
## $ylevels
## [1] "absent"  "present"
##
## $class
## [1] "rpart"
```

```
# 나무 모형 그림 그리기
plot(kyphosis.tr)
text(kyphosis.tr)
```



plot of chunk unnamed-chunk-3

2.2. Example : iris

```
data(iris)
# help(stagec)
```

```
str(iris)
```

```
## 'data.frame':	150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1
 1 1 1 1 1 ...
```

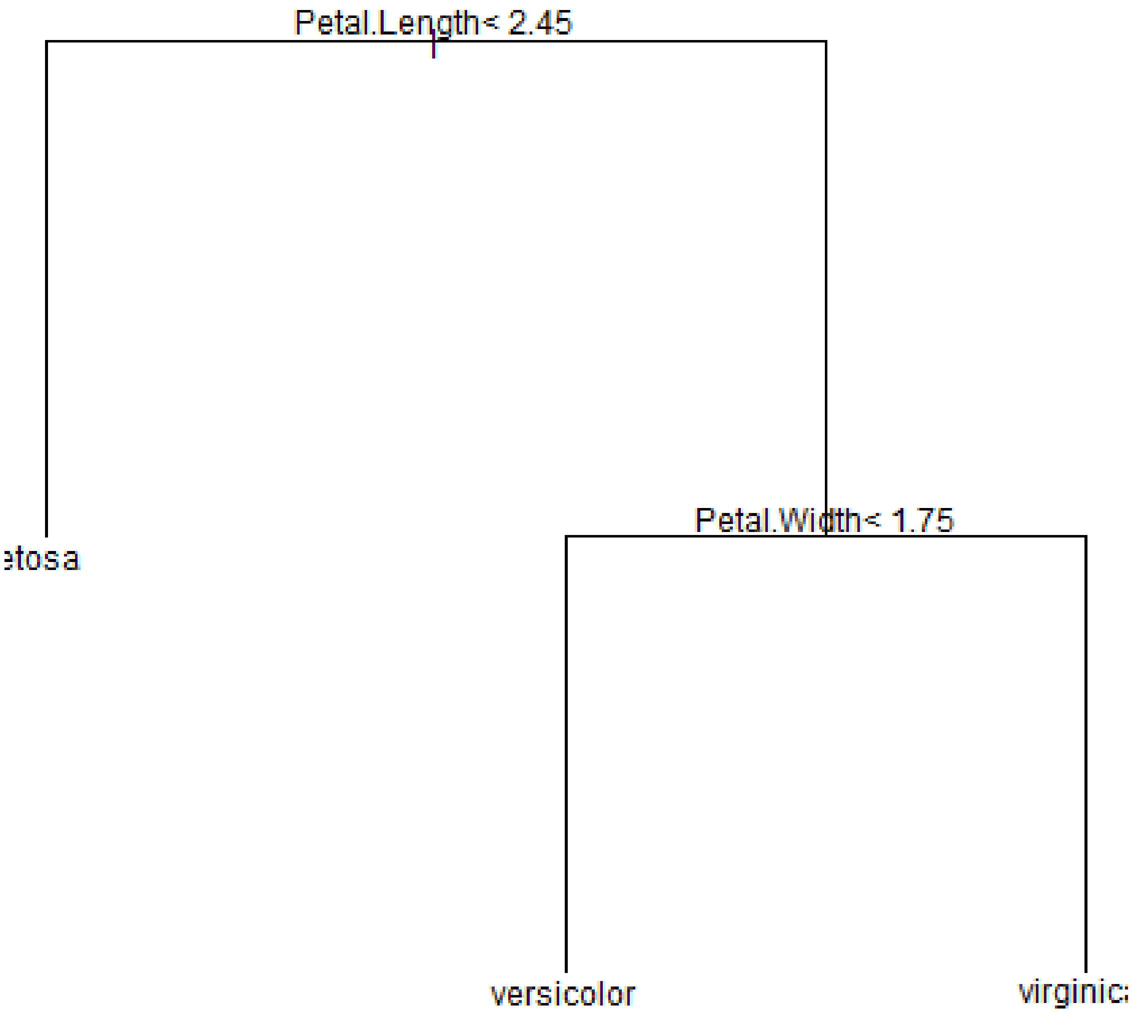
```
# 나무 모형 생성
iris.tr <- rpart(Species ~ ., data = iris)

print(iris.tr)
```

```
## n= 150
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 150 100 setosa (0.33333 0.33333 0.33333)
##   2) Petal.Length< 2.45 50   0 setosa (1.00000 0.00000 0.00000) *
##   3) Petal.Length>=2.45 100  50 versicolor (0.00000 0.50000 0.50000)
##     6) Petal.Width< 1.75 54   5 versicolor (0.00000 0.90741 0.09259) *
##     7) Petal.Width>=1.75 46   1 virginica (0.00000 0.02174 0.97826) *
```

```
attributes(iris.tr)
```

```
## $names
##  [1] "frame"               "where"               "call"
##  [4] "terms"               "cptable"             "method"
##  [7] "parms"               "control"             "functions"
## [10] "numresp"             "splits"              "variable.importance"
## [13] "y"                   "ordered"
##
## $xlevels
## named list()
##
## $ylevels
## [1] "setosa"     "versicolor" "virginica"
##
## $class
## [1] "rpart"
```

```
# 나무 모형 그림 그리기
plot(iris.tr)
text(iris.tr)
```



plot of chunk unnamed-chunk-4

2.3. Example : adult

```
adult <- read.table("cleanadult.data", col.names = c("age", "workcls", "fn
lwgt",
    "edu", "edu_num", "martial-status", "occupation", "relationship", "rac
e",
    "sex", "capital_gain", "capital_loss", "hours_per_week", "native_count
ry",
    "target"))
```

```
str(adult)
```

```
## 'data.frame':    30162 obs. of  15 variables:
##  $ age           : Factor w/ 72 levels "17,","18,","19,",..: 23 34 22 3
7 12 21 33 36 15 26 ...
##  $ workcls       : Factor w/ 7 levels "Federal-gov,",..: 6 5 3 3 3 3 3
5 3 3 ...
##  $ fnlwgt        : Factor w/ 20263 levels "100009,","100029,",..: 19136
 19378 9637 10837 15160 13251 4393 9189 17424 4319 ...
##  $ edu           : Factor w/ 16 levels "10th,","11th,",..: 10 10 12 2 1
0 13 7 12 13 10 ...
##  $ edu_num       : Factor w/ 16 levels "1,","10,","11,",..: 5 5 16 14 5
 6 12 16 6 5 ...
##  $ martial.status: Factor w/ 7 levels "Divorced,","Married-AF-spouse,
",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation    : Factor w/ 14 levels "Adm-clerical,",..: 1 4 6 6 10 4
 8 4 10 4 ...
##  $ relationship  : Factor w/ 6 levels "Husband,","Not-in-family,",..: 2
 1 2 1 6 6 2 1 2 1 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo,",..: 5 5 5 3
 3 5 3 5 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female,","Male,": 2 2 2 2 1 1 1
2 1 2 ...
##  $ capital_gain  : Factor w/ 118 levels "0,","10520,",..: 33 1 1 1 1 1
1 1 12 94 ...
##  $ capital_loss  : Factor w/ 90 levels "0,","1092,","1138,",..: 1 1 1 1
 1 1 1 1 1 1 ...
##  $ hours_per_week: Factor w/ 94 levels "1,","10,","11,",..: 35 5 35 35
35 35 8 40 46 35 ...
##  $ native_country: Factor w/ 41 levels "Cambodia,","Canada,",..: 39 39
39 39 5 39 23 39 39 39 ...
##  $ target        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2
 2 ...
```

```
# 변수 속성 변경
adult$age <- as.numeric(adult$age)
adult$fnlwgt <- as.numeric(adult$fnlwgt)
adult$capital_gain <- as.numeric(adult$capital_gain)
adult$capital_loss <- as.numeric(adult$capital_loss)
adult$hours_per_week <- as.numeric(adult$hours_per_week)

str(adult)
```

```
## 'data.frame':    30162 obs. of  15 variables:
##  $ age           : num  23 34 22 37 12 21 33 36 15 26 ...
##  $ workcls       : Factor w/ 7 levels "Federal-gov,",..: 6 5 3 3 3 3 3
5 3 3 ...
##  $ fnlwgt        : num  19136 19378 9637 10837 15160 ...
##  $ edu           : Factor w/ 16 levels "10th,","11th,",..: 10 10 12 2 1
0 13 7 12 13 10 ...
##  $ edu_num       : Factor w/ 16 levels "1,","10,","11,",..: 5 5 16 14 5
```

```
 6 12 16 6 5 ...
##  $ martial.status: Factor w/ 7 levels "Divorced,","Married-AF-spouse,
",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation    : Factor w/ 14 levels "Adm-clerical,",..: 1 4 6 6 10 4
 8 4 10 4 ...
##  $ relationship  : Factor w/ 6 levels "Husband,","Not-in-family,",..: 2
 1 2 1 6 6 2 1 2 1 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo,",..: 5 5 5 3
 3 5 3 5 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female,","Male,": 2 2 2 2 1 1 1
2 1 2 ...
##  $ capital_gain  : num  33 1 1 1 1 1 1 1 12 94 ...
##  $ capital_loss  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ hours_per_week: num  35 5 35 35 35 35 8 40 46 35 ...
##  $ native_country: Factor w/ 41 levels "Cambodia,","Canada,",..: 39 39
39 39 5 39 23 39 39 39 ...
##  $ target        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2
 2 ...
```

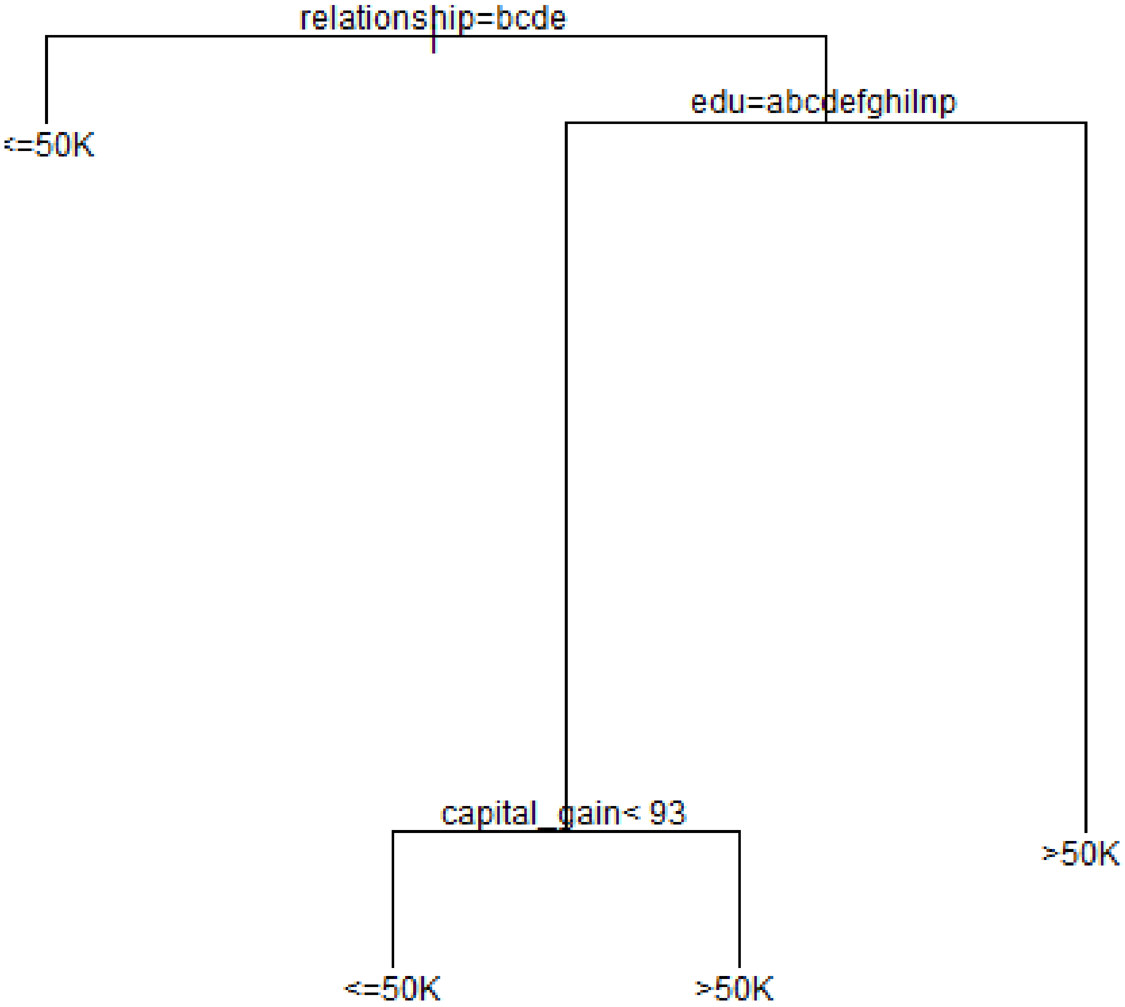
```
# 나무 모형 생성
adult.tr <- rpart(target ~ ., data = adult)

print(adult.tr)
```

```
## n= 30162
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 30162 7508 <=50K (0.751078 0.248922)
##    2) relationship=Not-in-family,,Other-relative,,Own-child,,Unmarried,
 16293 1135 <=50K (0.930338 0.069662) *
##    3) relationship=Husband,,Wife, 13869 6373 <=50K (0.540486 0.459514)

##      6) edu=10th,,11th,,12th,,1st-4th,,5th-6th,,7th-8th,,9th,,Assoc-acd
m,,Assoc-voc,,HS-grad,,Preschool,,Some-college, 9719 3322 <=50K (0.658195
0.341805)
##       12) capital_gain< 93 9338 2944 <=50K (0.684729 0.315271) *
##       13) capital_gain>=93 381    3 >50K (0.007874 0.992126) *
##      7) edu=Bachelors,,Doctorate,,Masters,,Prof-school, 4150 1099 >50K
(0.264819 0.735181) *
```

```
attributes(adult.tr)
```

```
## $names
##  [1] "frame"               "where"               "call"
##  [4] "terms"               "cptable"             "method"
##  [7] "parms"               "control"             "functions"
## [10] "numresp"             "splits"              "csplit"
## [13] "variable.importance" "y"                   "ordered"
##
## $xlevels
## $xlevels$workcls
## [1] "Federal-gov,"      "Local-gov,"        "Private,"
## [4] "Self-emp-inc,"     "Self-emp-not-inc," "State-gov,"
## [7] "Without-pay,"
##
## $xlevels$edu
##  [1] "10th,"         "11th,"         "12th,"         "1st-4th,"
##  [5] "5th-6th,"      "7th-8th,"      "9th,"          "Assoc-acdm,"
##  [9] "Assoc-voc,"    "Bachelors,"    "Doctorate,"    "HS-grad,"
## [13] "Masters,"      "Preschool,"    "Prof-school,"  "Some-college,"
##
## $xlevels$edu_num
##  [1] "1,"  "10," "11," "12," "13," "14," "15," "16," "2,"  "3,"  "4,"
## [12] "5,"  "6,"  "7,"  "8,"  "9,"
##
## $xlevels$martial.status
## [1] "Divorced,"              "Married-AF-spouse,"
## [3] "Married-civ-spouse,"    "Married-spouse-absent,"
## [5] "Never-married,"         "Separated,"
## [7] "Widowed,"
##
## $xlevels$occupation
##  [1] "Adm-clerical,"      "Armed-Forces,"      "Craft-repair,"
##  [4] "Exec-managerial,"   "Farming-fishing,"   "Handlers-cleaners,"
##  [7] "Machine-op-inspct," "Other-service,"     "Priv-house-serv,"
## [10] "Prof-specialty,"    "Protective-serv,"   "Sales,"
## [13] "Tech-support,"      "Transport-moving,"
##
## $xlevels$relationship
## [1] "Husband,"        "Not-in-family,"  "Other-relative," "Own-child,"

## [5] "Unmarried,"      "Wife,"
##
## $xlevels$race
## [1] "Amer-Indian-Eskimo," "Asian-Pac-Islander," "Black,"
## [4] "Other,"              "White,"
##
## $xlevels$sex
## [1] "Female," "Male,"
##
## $xlevels$native_country
##  [1] "Cambodia,"                   "Canada,"
##  [3] "China,"                      "Columbia,"
##  [5] "Cuba,"                       "Dominican-Republic,"
##  [7] "Ecuador,"                    "El-Salvador,"
##  [9] "England,"                    "France,"
## [11] "Germany,"                    "Greece,"
## [13] "Guatemala,"                  "Haiti,"
## [15] "Holand-Netherlands,"         "Honduras,"
## [17] "Hong,"                       "Hungary,"
## [19] "India,"                      "Iran,"
## [21] "Ireland,"                    "Italy,"
## [23] "Jamaica,"                    "Japan,"
## [25] "Laos,"                       "Mexico,"
## [27] "Nicaragua,"                  "Outlying-US(Guam-USVI-etc),"
## [29] "Peru,"                       "Philippines,"
## [31] "Poland,"                     "Portugal,"
## [33] "Puerto-Rico,"                "Scotland,"
## [35] "South,"                      "Taiwan,"
## [37] "Thailand,"                   "Trinadad&Tobago,"
## [39] "United-States,"              "Vietnam,"
## [41] "Yugoslavia,"
##
##
## $ylevels
## [1] "<=50K" ">50K"
##
## $class
## [1] "rpart"
```

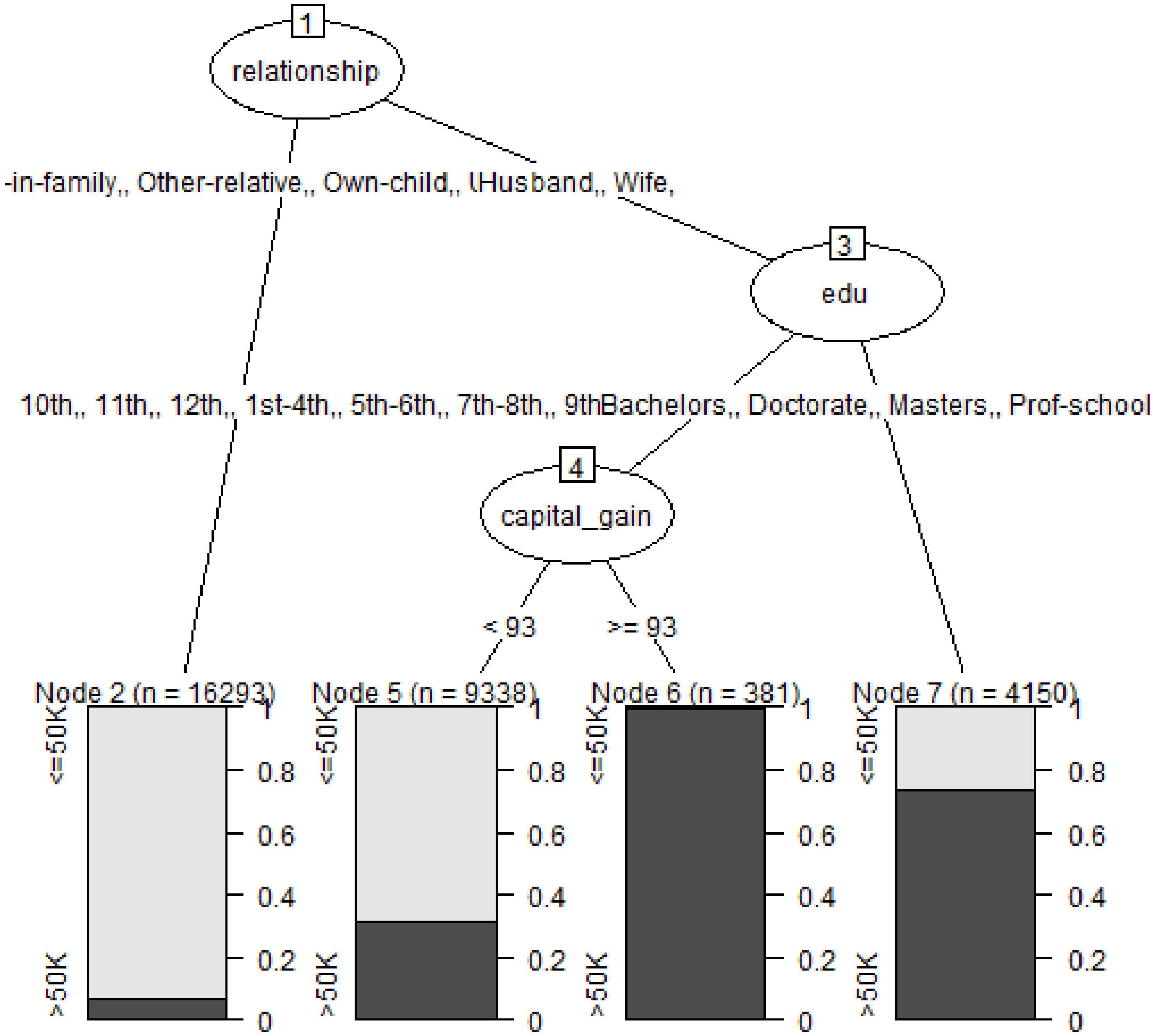
```
# 나무 모형 그림 그리기
plot(adult.tr)
text(adult.tr)
```



plot of chunk unnamed-chunk-5

```
library(partykit)
```

```
# 더 예쁜 그림
plot(as.party(adult.tr))
```



plot of chunk unnamed-chunk-7