HW_DecisionTree

경제학과 2020110210 공소연

2022-11-06

Q1-1

```
library(rpart)
```

```
## Warning: 패키지 'rpart'는 R 버전 4.2.2에서 작성되었습니다
```

```
data(stagec)
data <- stagec
str(data)
```

```
## 'data.frame':    146 obs. of  8 variables:
##  $ pgtime : num  6.1 9.4 5.2 3.2 1.9 4.8 5.8 7.3 3.7 15.9 ...
##  $ pgstat : int  0 0 1 1 1 0 0 0 1 0 ...
##  $ age    : int  64 62 59 62 64 69 75 71 73 64 ...
##  $ eet    : int  2 1 2 2 2 1 2 2 2 2 ...
##  $ g2     : num  10.26 NA 9.99 3.57 22.56 ...
##  $ grade  : int  2 3 3 2 4 3 2 3 3 3 ...
##  $ gleason: int  4 8 7 4 8 7 NA 7 6 7 ...
##  $ ploidy : Factor w/ 3 levels "diploid","tetraploid",..: 1 3 1 1 2
1 2 3 1 2 ...
```

```
test.df <- data[145:146,-8]
train.df <- data[1:144,]
```

Q1-2

```
my_control <- rpart.control(xval = 10, cp = 0,
                            minsplit = 4)
tree_model <- rpart(ploidy~.,method = "class",
                    control = my_control, data = train.df)
tree_model
```

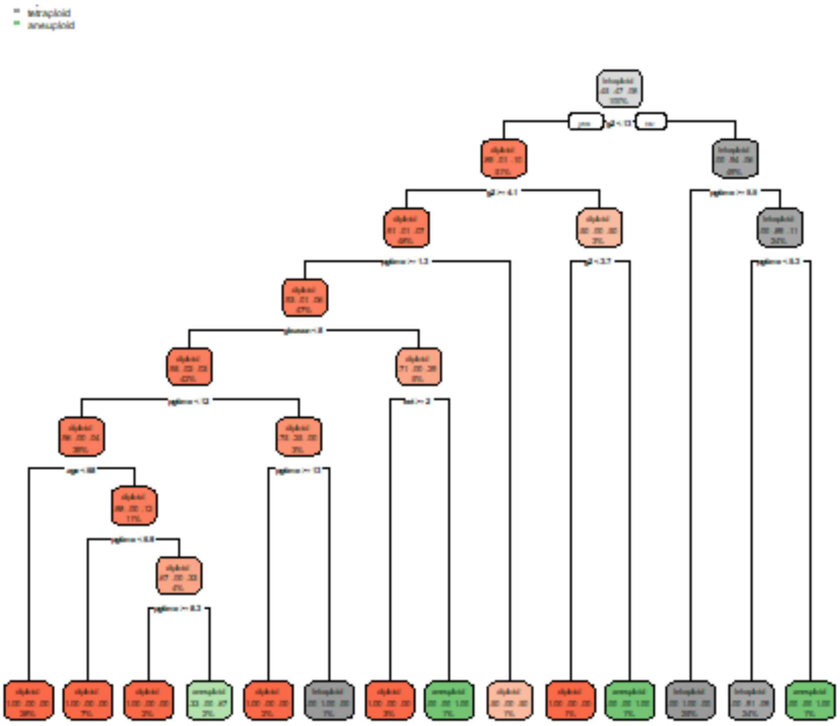
```
## n= 144
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##   1) root 144 76 tetraploid (0.45138889 0.47222222 0.07638889)
##     2) g2< 13.055 73  8 diploid (0.89041096 0.01369863 0.09589041)
##       4) g2>=4.14 69  6 diploid (0.91304348 0.01449275 0.07246377)
##         8) pgtime>=1.25 67  5 diploid (0.92537313 0.01492537
0.05970149)
##          16) gleason< 7.5 60  3 diploid (0.95000000 0.01666667
0.03333333)
##            32) pgtime< 12.45 56  2 diploid (0.96428571 0.00000000
```

```
0.03571429)
##              64) age< 67.5 40  0 diploid (1.00000000 0.00000000
0.00000000) *
##              65) age>=67.5 16  2 diploid (0.87500000 0.00000000
0.12500000)
##               130) pgtime< 6.9 10  0 diploid (1.00000000 0.00000000
0.00000000) *
##               131) pgtime>=6.9 6  2 diploid (0.66666667 0.00000000
0.33333333)
##                 262) pgtime>=8.3 3  0 diploid (1.00000000
0.00000000 0.00000000) *
##                 263) pgtime< 8.3 3  1 aneuploid (0.33333333
0.00000000 0.66666667) *
##            33) pgtime>=12.45 4  1 diploid (0.75000000 0.25000000
0.00000000)
##              66) pgtime>=13.3 3  0 diploid (1.00000000 0.00000000
0.00000000) *
##              67) pgtime< 13.3 1  0 tetraploid (0.00000000
1.00000000 0.00000000) *
##          17) gleason>=7.5 7  2 diploid (0.71428571 0.00000000
0.28571429)
##            34) eet>=1.5 5  0 diploid (1.00000000 0.00000000
0.00000000) *
##            35) eet< 1.5 2  0 aneuploid (0.00000000 0.00000000
1.00000000) *
##         9) pgtime< 1.25 2  1 diploid (0.50000000 0.00000000
0.50000000) *
##       5) g2< 4.14 4  2 diploid (0.50000000 0.00000000 0.50000000)
##        10) g2< 3.67 2  0 diploid (1.00000000 0.00000000 0.00000000)
*
##        11) g2>=3.67 2  0 aneuploid (0.00000000 0.00000000
1.00000000) *
##     3) g2>=13.055 71  4 tetraploid (0.00000000 0.94366197
0.05633803)
##       6) pgtime>=5.5 36  0 tetraploid (0.00000000 1.00000000
0.00000000) *
##       7) pgtime< 5.5 35  4 tetraploid (0.00000000 0.88571429
0.11428571)
##        14) pgtime< 5.3 34  3 tetraploid (0.00000000 0.91176471
0.08823529) *
##        15) pgtime>=5.3 1  0 aneuploid (0.00000000 0.00000000
1.00000000) *
```

```
library(rpart.plot)
```

```
## Warning: 패키지 'rpart.plot'는 R 버전 4.2.2에서 작성되었습니다
```

```
rpart.plot(tree_model)
```

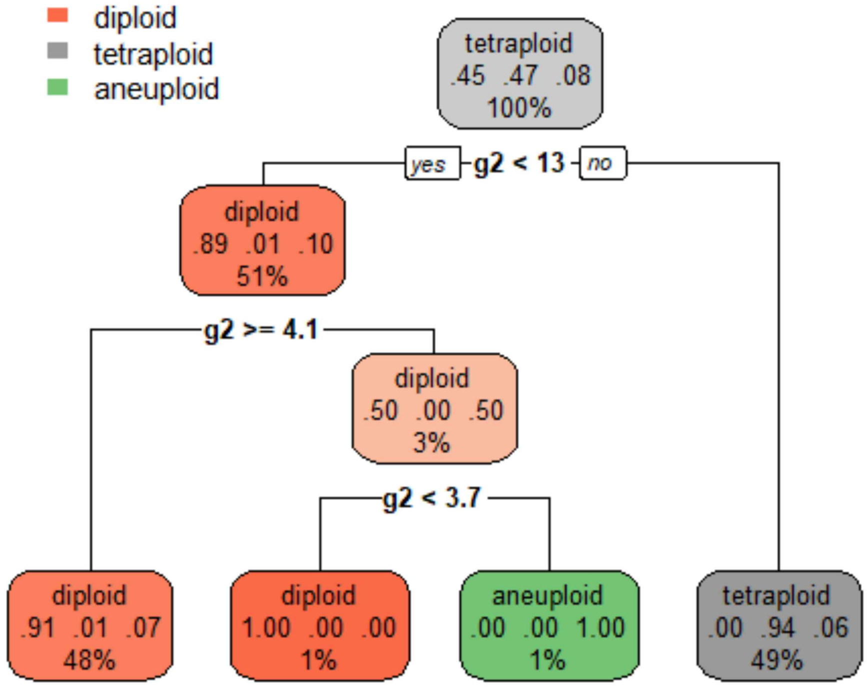


Q1-3

```
printcp(tree_model)
```

```
##
## Classification tree:
## rpart(formula = ploidy ~ ., data = train.df, method = "class",
##     control = my_control)
##
## Variables actually used in tree construction:
## [1] age     eet     g2      gleason pgtime
##
## Root node error: 76/144 = 0.52778
##
## n= 144
##
##          CP nsplit rel error  xerror     xstd
## 1 0.8421053      0  1.000000 1.18421 0.076440
## 2 0.0131579      1  0.157895 0.15789 0.043640
## 3 0.0087719      3  0.131579 0.18421 0.046778
## 4 0.0065789      6  0.105263 0.19737 0.048233
## 5 0.0043860     10  0.078947 0.23684 0.052219
## 6 0.0000000     13  0.065789 0.26316 0.054605
```

```
pruned_model <- prune.rpart(tree_model, cp = 0.013)
rpart.plot(pruned_model)
```



Q1-4

```
predict(pruned_model, newdata = test.df, type = "class")
```

```
##     145     146
## diploid diploid
## Levels: diploid tetraploid aneuploid
```

Q2-1

```
library("TH.data")
```

```
## Warning: 패키지 'TH.data'는 R 버전 4.2.2에서 작성되었습니다
```

```
## 필요한 패키지를 로딩중입니다: survival
```

```
## 필요한 패키지를 로딩중입니다: MASS
```

```
##
## 다음의 패키지를 부착합니다: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
data(bodyfat)
data2 <- bodyfat
str(data2)
```

```
## 'data.frame':    71 obs. of  10 variables:
##  $ age          : num  57 65 59 58 60 61 56 60 58 62 ...
##  $ DEXfat      : num  41.7 43.3 35.4 22.8 36.4 ...
##  $ waistcirc   : num  100 99.5 96 72 89.5 83.5 81 89 80 79 ...
##  $ hipcirc     : num  112 116.5 108.5 96.5 100.5 ...
##  $ elbowbreadth: num  7.1 6.5 6.2 6.1 7.1 6.5 6.9 6.2 6.4 7 ...
##  $ kneebreadth : num  9.4 8.9 8.9 9.2 10 8.8 8.9 8.5 8.8 8.8 ...
##  $ anthro3a    : num  4.42 4.63 4.12 4.03 4.24 3.55 4.14 4.04 3.91 
3.66 ...
##  $ anthro3b    : num  4.95 5.01 4.74 4.48 4.68 4.06 4.52 4.7 4.32 
4.21 ...
##  $ anthro3c    : num  4.5 4.48 4.6 3.91 4.15 3.64 4.31 4.47 3.47 
3.6 ...
##  $ anthro4     : num  6.13 6.37 5.82 5.66 5.91 5.14 5.69 5.7 5.49 
5.25 ...
```

```
?bodyfat
```

```
## httpd 도움말 서버를 시작합니다 ...
```

```
##  완료
```

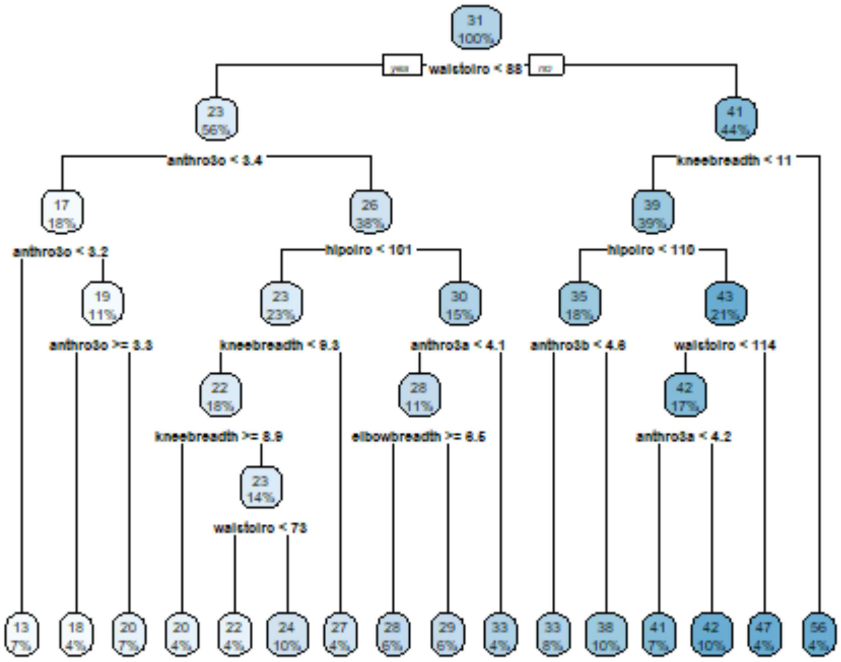
```
# DEXfat is the dependent variable.
```

Q2-2

```
set.seed(123)
my_control2 <- rpart.control(xval = 10, cp = 0,
                             minsplit = 8)
tree_model2 <- rpart(DEXfat~., method = "anova",
                     control = my_control2, data = data2)
tree_model2
```

```
## n= 71 
## 
## node), split, n, deviance, yval
##       * denotes terminal node
## 
##  1) root 71 8535.984000 30.78282  
##    2) waistcirc< 88.4 40 1315.358000 22.92375  
##      4) anthro3c< 3.42 13  145.993100 16.83692  
##        8) anthro3c< 3.165 5   11.818520 13.03600 *
##        9) anthro3c>=3.165 8   16.792550 19.21250  
##         18) anthro3c>=3.31 3    3.139467 17.90667 *
##         19) anthro3c< 3.31 5    5.468120 19.99600 *
##      5) anthro3c>=3.42 27  455.819300 25.85444  
##       10) hipcirc< 101.35 16  120.142900 23.31937  
##         20) kneebreadth< 9.25 13   66.003320 22.49538  
##           40) kneebreadth>=8.85 3   13.596800 19.79000 *
##           41) kneebreadth< 8.85 10   23.862010 23.30700  
##             82) waistcirc< 72.5 3    2.886667 22.00667 *
##             83) waistcirc>=72.5 7   13.728770 23.86429 *
##         21) kneebreadth>=9.25 3    7.065000 26.89000 *
##       11) hipcirc>=101.35 11   83.287160 29.54182  
##         22) anthro3a< 4.09 8   12.389350 28.31250  
##            44) elbowbreadth>=6.45 4    3.575600 27.51000 *
##            45) elbowbreadth< 6.45 4    3.661700 29.11500 *
##          23) anthro3a>=4.09 3   26.568600 32.82000 *
##    3) waistcirc>=88.4 31 1562.162000 40.92355
##      6) kneebreadth< 11.15 28  615.525900 39.26036
##       12) hipcirc< 109.9 13  136.296000 35.27846
##         24) anthro3b< 4.605 6   19.088200 32.61000 *
##         25) anthro3b>=4.605 7   37.862970 37.56571 *
##       13) hipcirc>=109.9 15   94.469970 42.71133
##         26) waistcirc< 113.5 12   30.609800 41.69000
##           52) anthro3a< 4.155 5   16.973800 40.66000 *
##           53) anthro3a>=4.155 7    4.542571 42.42571 *
##         27) waistcirc>=113.5 3    1.272867 46.79667 *
##      7) kneebreadth>=11.15 3  146.280300 56.44667 *
```

```
rpart.plot(tree_model2)
```



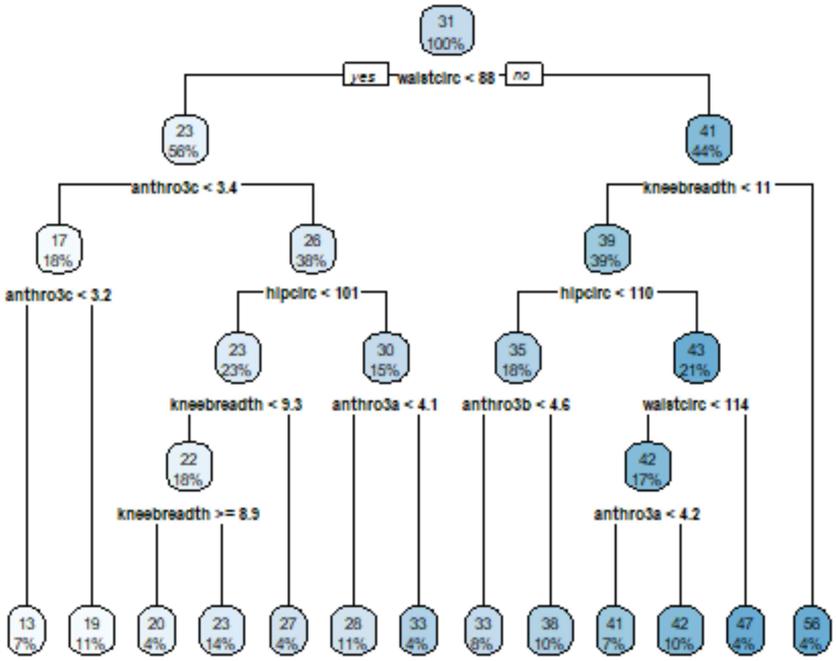
Q2-3

```
printcp(tree_model2)
```

```
##
## Regression tree:
## rpart(formula = DEXfat ~ ., data = data2, method = "anova", control
= my_control2)
##
## Variables actually used in tree construction:
## [1] anthro3a     anthro3b     anthro3c     elbowbreadth hipcirc
## [6] kneebreadth  waistcirc
```

```
##
## Root node error: 8536/71 = 120.23
##
## n= 71
##
##            CP nsplit rel error  xerror     xstd
## 1  0.66289544      0  1.000000 1.03108 0.169866
## 2  0.09376252      1  0.337105 0.39741 0.091802
## 3  0.08359261      2  0.243342 0.37083 0.070027
## 4  0.04507506      3  0.159749 0.33000 0.066175
## 5  0.02956768      4  0.114674 0.28050 0.057943
## 6  0.01375143      5  0.085107 0.24686 0.066795
## 7  0.00929533      6  0.071355 0.24186 0.066630
## 8  0.00733217      7  0.062060 0.22636 0.063950
## 9  0.00551484      8  0.054728 0.22152 0.063786
## 10 0.00519322      9  0.049213 0.22133 0.063795
## 11 0.00334402     10  0.044020 0.21419 0.063774
## 12 0.00106531     11  0.040676 0.19240 0.053999
## 13 0.00095888     12  0.039610 0.19004 0.053767
## 14 0.00084894     13  0.038651 0.19072 0.053745
## 15 0.00060357     14  0.037803 0.19072 0.053745
## 16 0.00000000     15  0.037199 0.19122 0.053845
```

```
pruned_model2 <- prune.rpart(tree_model2, cp = 0.00096)
rpart.plot(pruned_model2)
```


31
100%
yes
waistcirc < 88
no
23
56%
41
44%
anthro3c < 3.4
kneebreadth < 11
17
18%
26
38%
39
39%
anthro3c < 3.2
hipcirc < 101
hipcirc < 110
23
23%
30
15%
35
18%
43
21%
kneebreadth < 9.3
anthro3a < 4.1
anthro3b < 4.6
waistcirc < 114
22
18%
42
17%
kneebreadth >= 8.9
anthro3a < 4.2
13
7%
19
11%
20
4%
23
14%
27
4%
28
11%
33
4%
33
8%
38
10%
41
7%
42
10%
47
4%
56
4%