Homework 4

Yanhao Li

Contents

```
library(tidyverse)
library(ISLR)
library(lasso2)
library(ISLR)
library(rpart)
library(rpart.plot)
library(randomForest)
library(ranger)
library(caret)
library(gbm)
```

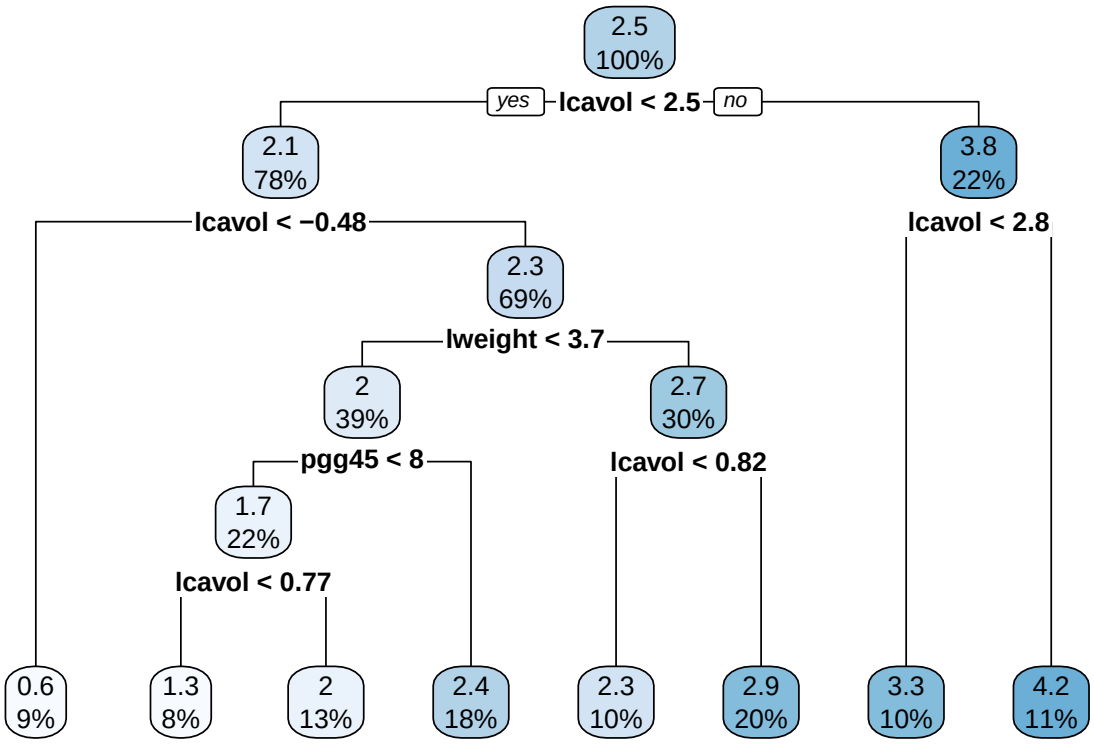
Question 1

Load, clean, and tidy data

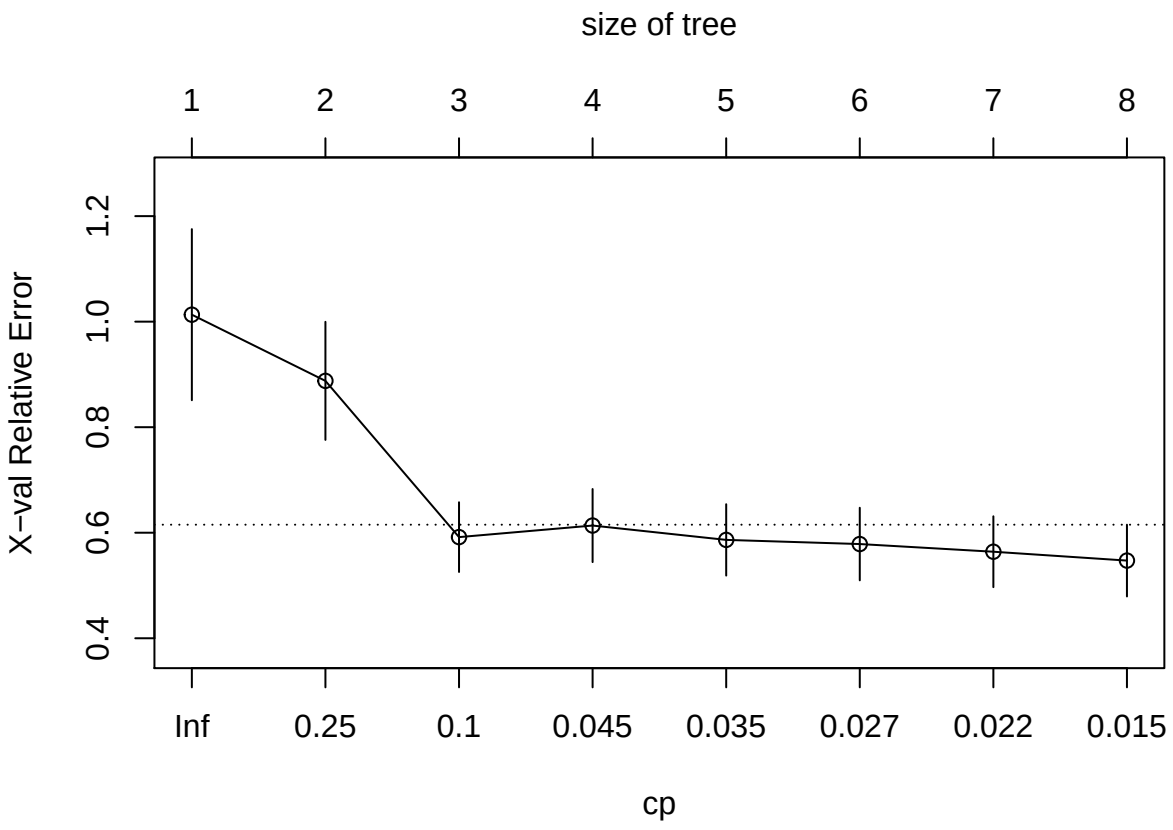
```
data("Prostate")

prostate = Prostate %>%
  janitor::clean_names()
```

Question a

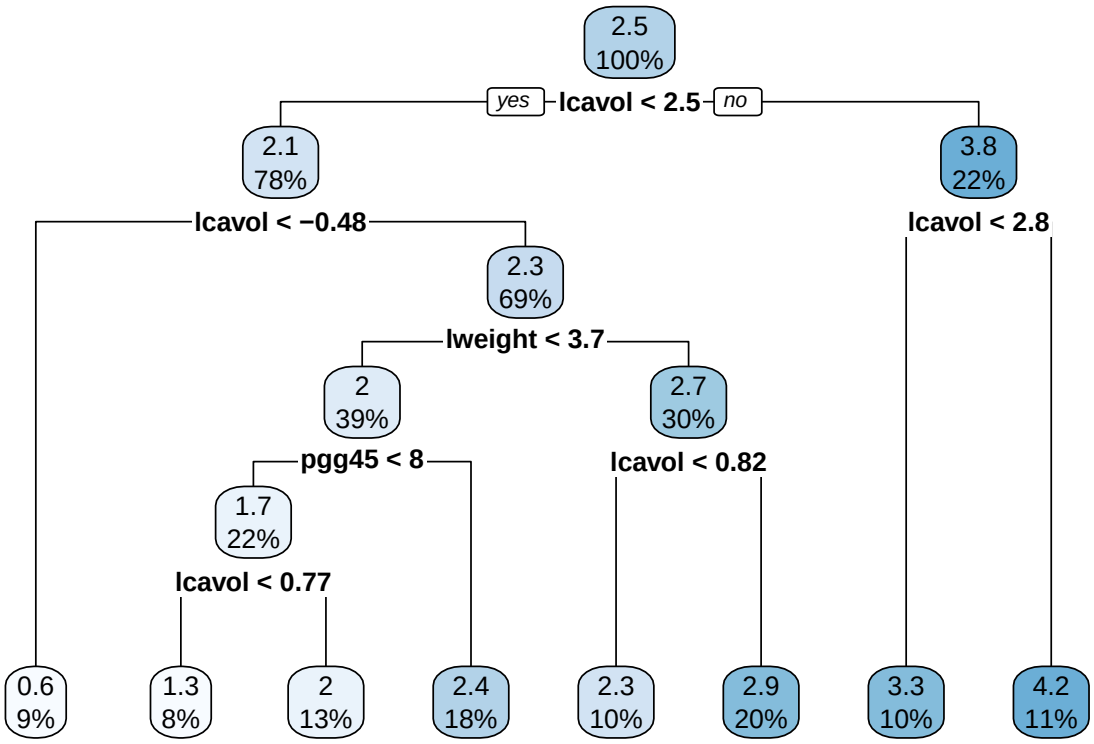
```
set.seed(1)

tree1 = rpart(formula = lpsa ~ .,
              data = prostate)

rpart.plot(tree1)
```



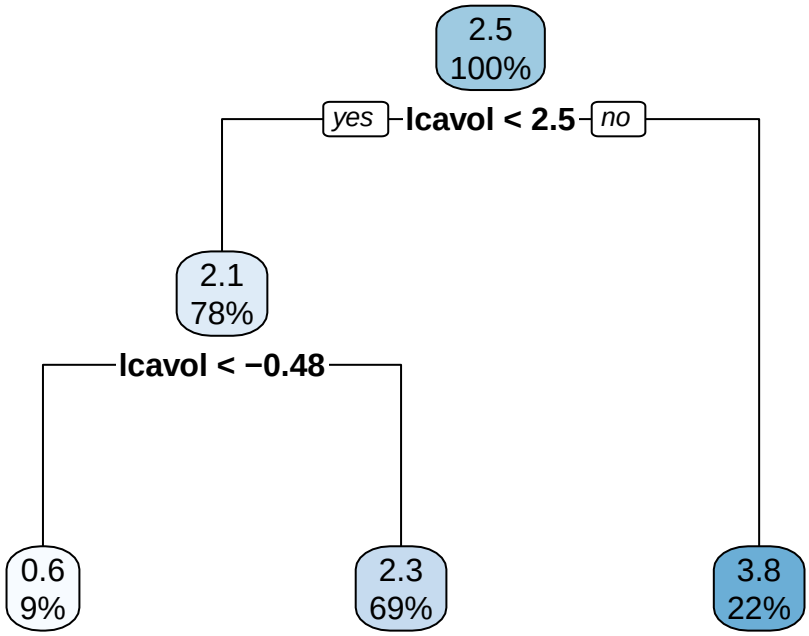

```
cpTable <- tree1$cptable

plotcp(tree1)
```




```
# minimum cross-validation error
minErr <- which.min(cpTable[,4])

tree2 <- prune(tree1, cp = cpTable[minErr,1])

rpart.plot(tree2)
```



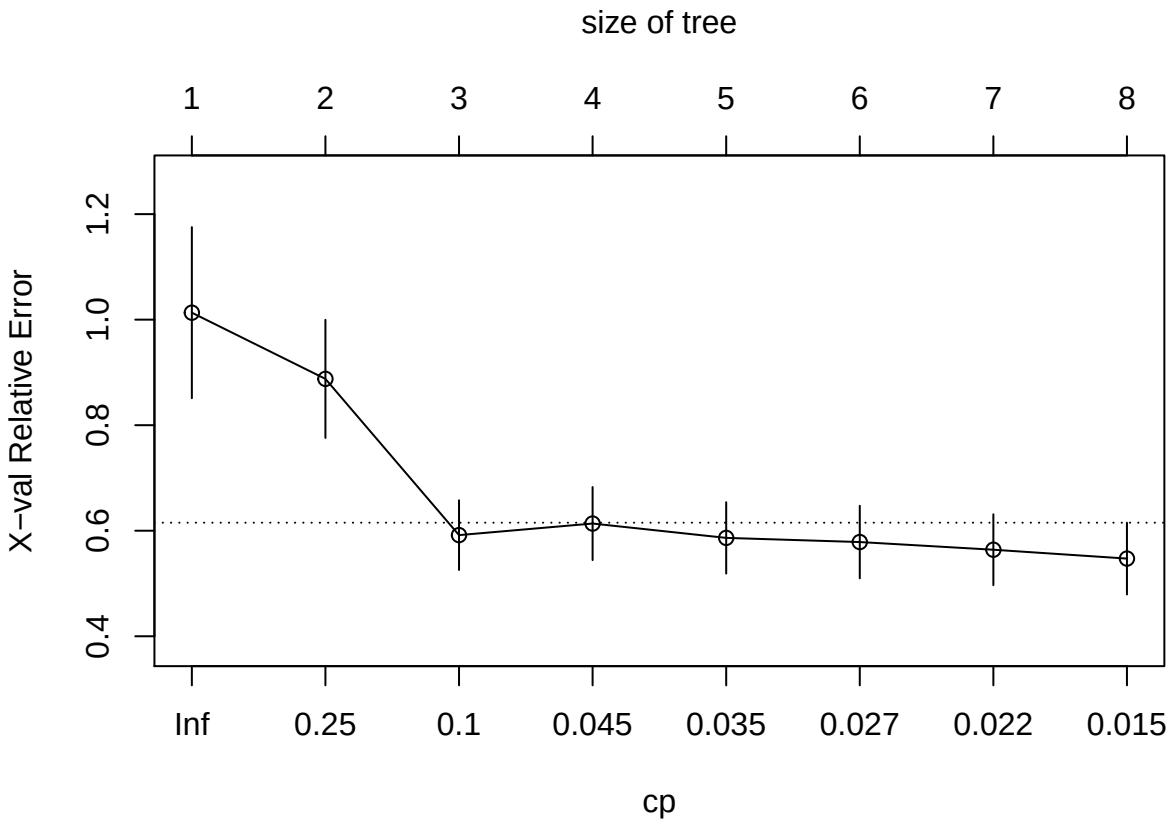

```
# 1SE rule
tree3 <- prune(tree1, cp = cpTable[cpTable[,4] < cpTable[minErr,4] + cpTable[minErr,5],1][1])

rpart.plot(tree3)
```



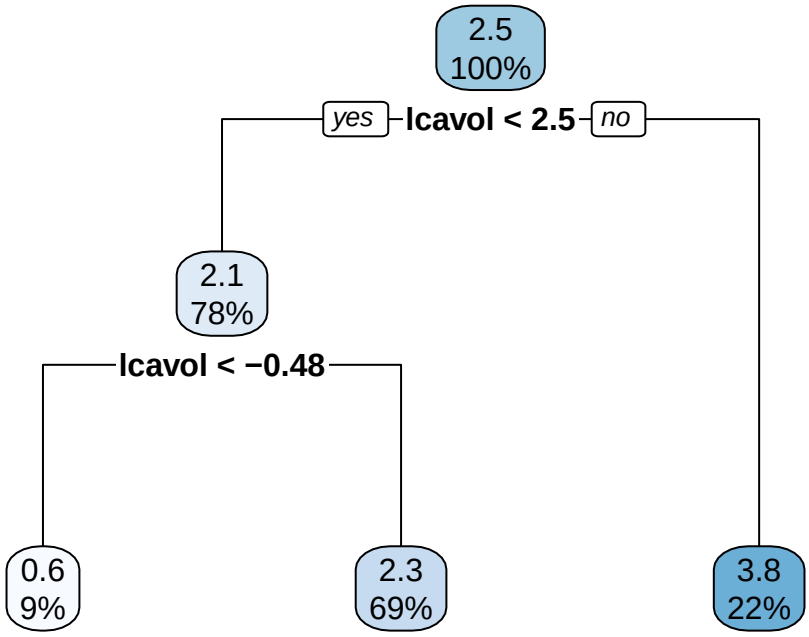
Tree size corresponds to the lowest cross-validation error is 8. It is different from the tree size obtained using the 1 SE rule, which is 3.

Question b

```
plotcp(tree1)
```




```
set.seed(1)

tree4 <- rpart(formula = lpsa ~ .,
               data = prostate,
               control = rpart.control(cp = 0.1))

rpart.plot(tree4)
```



A good choice of cp for pruning is often the leftmost value for which the mean lies below the horizontal line. According to the plot, I choose cp equals to 0.1 and size of tree equals to 3.

In terminal node where lcvol is less than -0.48, the mean lpsa is 0.6. This node contains 9% of the sample.

Question c

```
set.seed(1)

bagging <- randomForest(lpsa ~ . ,
                        prostate,
                        mtry = 8)

bagging$importance
```

```
##         IncNodePurity
## lcavol      76.557359
## lweight     16.761566
## age          5.875410
## lbph         5.123664
## svi          6.534788
## lcp          5.937665
## gleason      1.096503
## pgg45        5.776304
```

According to the table, variable importance from highest to lowest is lcavol, lweight, svi, lcp, age, pgg45, lbph, and gleason.

Question d

```
set.seed(1)

rf <- randomForest(lpsa ~ . ,
                   prostate,
                   mtry = 2)

rf$importance
```

```
##          IncNodePurity
## lcavol       34.596465
## lweight      18.743420
## age           8.892790
## lbph          7.317636
## svi          12.180308
## lcp          14.213581
## gleason       7.015983
## pgg45        11.246124
```

According to the table, variable importance from highest to lowest is lcavol, lweight, lcp, svi, pgg45, age, lbph, and gleason.

Question e

Question f

Question 2

Load, clean, and tidy data

Question a

Question b

Question c