문제1번

카페에 올라와 있는 Heart.csv 데이터를 다운 받으세요

caret 패키지에 있는 createDataPartition 함수를 이용하여 train:test = 7:3 이 되도록 데이터를 나눠주세요 >> **set.seed(4)로 설정, 종속변수는 AHD**

< 데이터 설명 >
Heart 데이터는 흉부외과 환자 303명을 관찰한 데이터로, AHD 칼럼에 각 환자들이 심장병이 있는지 여부가 기록되어 있습니다. heart.csv에 담긴 데이터들의 칼럼은 아래와 같이 14개가 있습니다.

1.age : 나이 (int)
2.sex : 성별 (1, 0 / int)
3.chest pain type (4 values) : 가슴 통증 타입 (0 ~ 3 / int)
4.resting blood pressure : 혈압
5.serum cholestoral in mg/dl : 혈청 콜레스테롤
6.fasting blood sugar > 120 mg/dl : 공복 혈당
7.resting electrocardiographic results : 심전도
8.maximum heart rate achieved : 최대 심장박동 수
9.exercise induced angina : 운동 유도 협심증
10.oldpeak = ST depression induced by exercise relative to rest : 노약 = 운동에 의해 유발되는 St 우울증
11.the slope of the peak exercise ST segment ST : 세그먼트의 기울기
12.number of major vessels (0-3) colored by flourosopy : 혈관의수
13.thal : 3 = normal; 6 = fixed defect; 7 = reversible defect: thalassemia이라고 불리우는 혈관질병여부

```
## 'data.frame':    303 obs. of  15 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age      : int  63 67 67 37 41 56 62 57 63 53 ...
##  $ Sex      : int  1 1 1 1 0 1 0 0 1 1 ...
##  $ ChestPain: Factor w/ 4 levels "asymptomatic",..: 4 1 1 2 3 3 1 1 1 1 ...
##  $ RestBP   : int  145 160 120 130 130 120 140 120 130 140 ...
##  $ Chol     : int  233 286 229 250 204 236 268 354 254 203 ...
##  $ Fbs      : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ RestECG  : int  2 2 2 0 2 0 2 0 2 2 ...
##  $ MaxHR    : int  150 108 129 187 172 178 160 163 147 155 ...
##  $ ExAng    : int  0 1 1 0 0 0 0 1 0 1 ...
##  $ Oldpeak  : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ Slope    : int  3 2 2 3 1 1 3 1 2 3 ...
##  $ Ca       : int  0 3 2 0 0 0 2 0 1 0 ...
##  $ Thal     : Factor w/ 3 levels "fixed","normal",..: 1 2 3 2 2 2 2 2 3 3 ...
##  $ AHD      : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 1 2 1 2 2 ...
```

문제2번

rpart 알고리즘을 이용하여 tree를 만들기

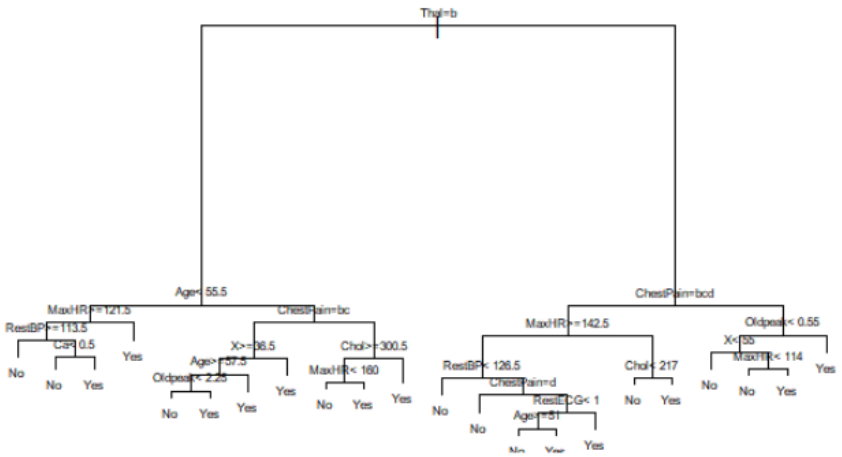
2-1 먼저 사전가지치기를 위해, 가장 accurary를 높게하는 minsplit를 '그리드서치'를 통해서 찾으세요.

이때 i는 1부터 20까지로 해주세요. 이때 max accuracy를 갖는 minsplit이 여러 개 나왔다면 가장 큰 i값

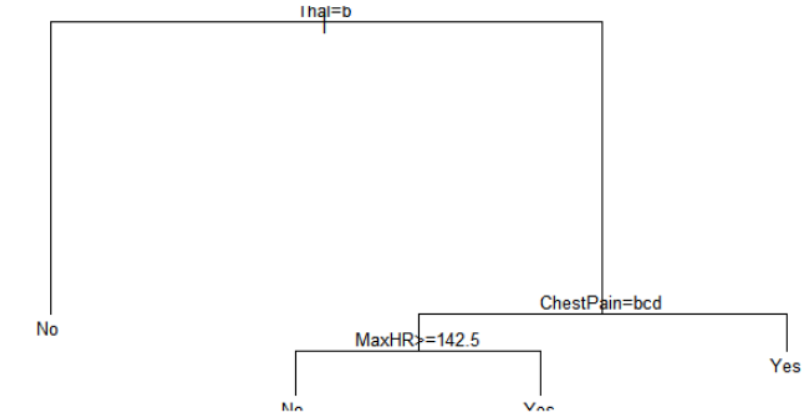
=minsplit값을 확인해 줍니다. 힌트 : for loop이용하기, loop안의 seed는 234

2-2 2-1에서 구한 minsplit값을 이용하여 세운 rpart 트리를 plotting해주세요. (model fit전에 seed 234써주기)

2-3 2-2에서 세운 rpart모델의 cp table과 cp ploting을 참고하여 과적합 방지를 위한 사후가지치기를 하고 다시 plotting 하세요.



사후 가지치기 후



문제3번

2번에서 만든 모델 중 가지치기를 하기 전 모델과 한 후의 모델의 성능을 test data에서 Confusion matrix 를 통해 비교해보세요.

```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  30  10
       Yes 19  31

               Accuracy : 0.6778
                 95% CI : (0.571, 0.7725)
    No Information Rate : 0.5444
    P-Value [Acc > NIR] : 0.006912

                  Kappa : 0.3619

 Mcnemar's Test P-Value : 0.137395

            Sensitivity : 0.6122
            Specificity : 0.7561
         Pos Pred Value : 0.7500
         Neg Pred Value : 0.6200
             Prevalence : 0.5444
         Detection Rate : 0.3333
   Detection Prevalence : 0.4444
      Balanced Accuracy : 0.6842

       'Positive' Class : No
```

사후 가지치기 후

```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  35  13
       Yes 14  28

               Accuracy : 0.7
                 95% CI : (0.5943, 0.7921)
    No Information Rate : 0.5444
    P-Value [Acc > NIR] : 0.001859

                  Kappa : 0.3964

 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.7143
            Specificity : 0.6829
         Pos Pred Value : 0.7292
         Neg Pred Value : 0.6667
             Prevalence : 0.5444
         Detection Rate : 0.3889
   Detection Prevalence : 0.5333
      Balanced Accuracy : 0.6986

       'Positive' Class : No
```