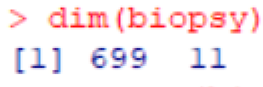
20 BME 7082/26 BE 7082/26 PH 7028
Autumn 2020
MB Rao

Homework Sheet No. 2 Due Date: September 10, 2020 Maximum Points: 30

Wisconsin Breast Cancer Data: A gold standard procedure for detecting breast cancer is biopsy. This is a definitive procedure. Women, fifty years or older, are advised get checked once every year for the presence or absence of cancer. Biopsy is painful, time-consuming, and expensive. It cannot be done every year. An alternative procedure is mammogram. This diagnostic test is not accurate. Its sensitivity is about 85% (True positives) and specificity 80% (True negatives). Wisconsin Medical Research Center proposed another diagnostic procedure (breast aspiration), which they touted more accurate than the mammogram. A needle is inserted into the breast and cells are extracted. Various properties (nine in all) of the cells are noted for each case (malignant) and control (benign). The determination of whether it is malignant or benign comes from the gold standard procedure. Download the data (biopsy) from the package (MASS). The response variable is 'class,' which is binary. We have nine predictors in all. They are cryptically labeled V1 through V9.

One needs to prepare the data before building a classification tree.

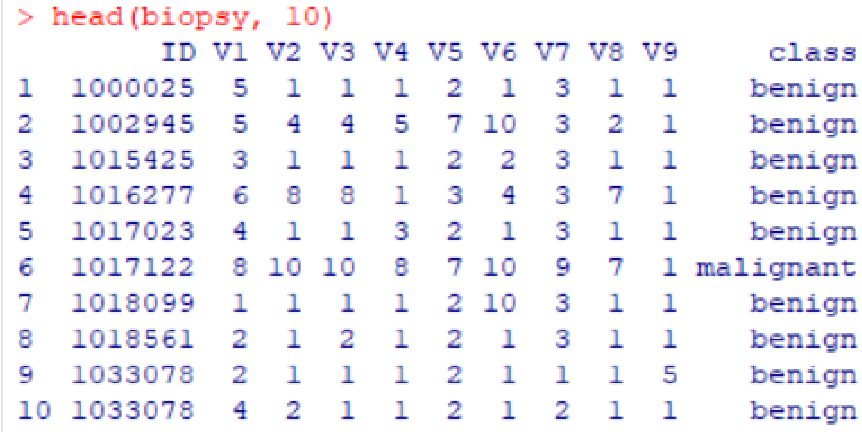
1. What is the dimension of the data? 1 point



```
> dim(biopsy)
[1] 699  11
```

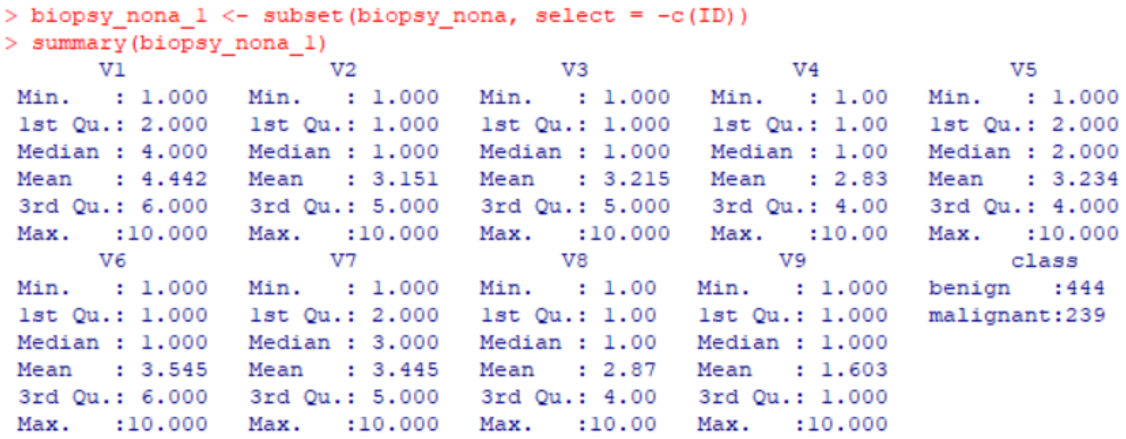
The data has 699 rows and 11 columns.

2. Show the top ten rows of the data. 1 point



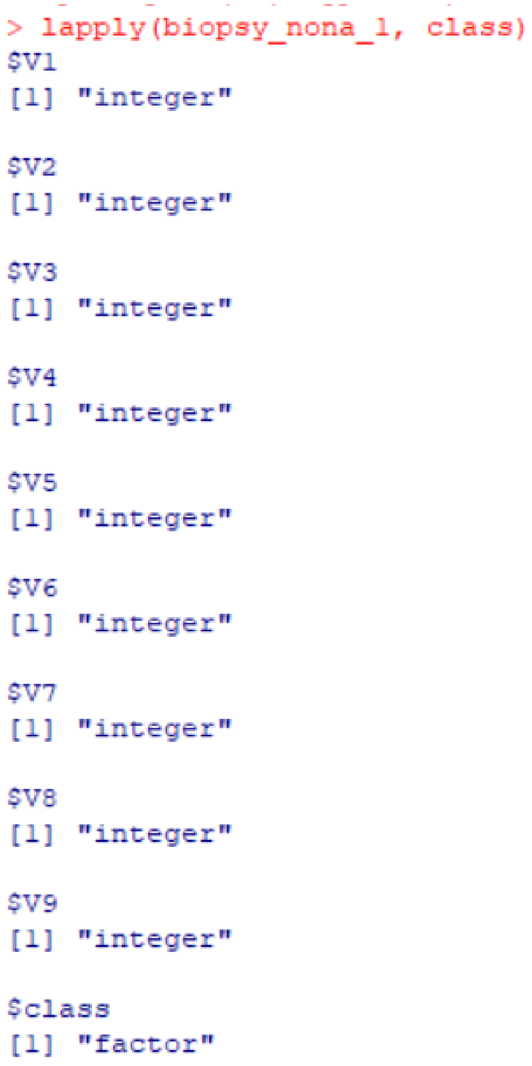
```
> head(biopsy, 10)
        ID V1 V2 V3 V4 V5 V6 V7 V8 V9     class
1  1000025  5  1  1  1  2  1  3  1  1    benign
2  1002945  5  4  4  5  7 10  3  2  1    benign
3  1015425  3  1  1  1  2  2  3  1  1    benign
4  1016277  6  8  8  1  3  4  3  7  1    benign
5  1017023  4  1  1  3  2  1  3  1  1    benign
6  1017122  8 10 10  8  7 10  9  7  1 malignant
7  1018099  1  1  1  1  2 10  3  1  1    benign
8  1018561  2  1  2  1  2  1  3  1  1    benign
9  1033078  2  1  1  1  2  1  1  1  5    benign
10 1033078  4  2  1  1  2  1  2  1  1    benign
```

3. The first column is id. This is a variable. It is useless. Create a new folder eliminating the first column. 1 point



```
> biopsy_nona_1 <- subset(biopsy_nona, select = -c(ID))
> summary(biopsy_nona_1)
       V1               V2               V3               V4              V5        
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.00   Min.   : 1.000  
 1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.00   1st Qu.: 2.000  
 Median : 4.000   Median : 1.000   Median : 1.000   Median : 1.00   Median : 2.000  
 Mean   : 4.442   Mean   : 3.151   Mean   : 3.215   Mean   : 2.83   Mean   : 3.234  
 3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.00   3rd Qu.: 4.000  
 Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.00   Max.   :10.000  
       V6               V7               V8              V9               class    
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.00   Min.   : 1.000   benign   :444  
 1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.00   1st Qu.: 1.000   malignant:239  
 Median : 1.000   Median : 3.000   Median : 1.00   Median : 1.000                  
 Mean   : 3.545   Mean   : 3.445   Mean   : 2.87   Mean   : 1.603                  
 3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 4.00   3rd Qu.: 1.000                  
 Max.   :10.000   Max.   :10.000   Max.   :10.00   Max.   :10.000                  
```

4. What is the class of each of the variables now? 1 point

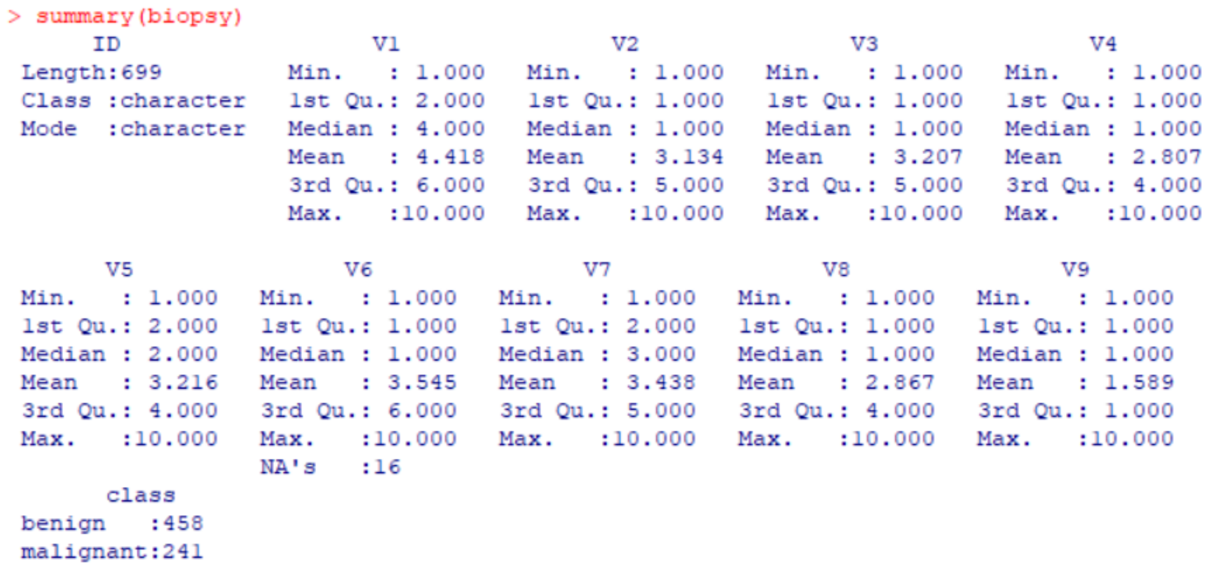


```
> lapply(biopsy_nona_1, class)
$V1
[1] "integer"

$V2
[1] "integer"

$V3
[1] "integer"

$V4
[1] "integer"

$V5
[1] "integer"

$V6
[1] "integer"

$V7
[1] "integer"

$V8
[1] "integer"

$V9
[1] "integer"

$class
[1] "factor"
```

5. Explain each variable. 4 points

Variables V1 through V9 are integer type meaning their numerical values will be used by R to classify (or predict) the result in the column entitled “Class”. Aforementioned command also enlists the last class column type as factor meaning the column will be used to categorize and store data in vectors or as in our case words “benign” or “malignant”.

6. Do summary statistics. Are there any missing observations? 1+1 points



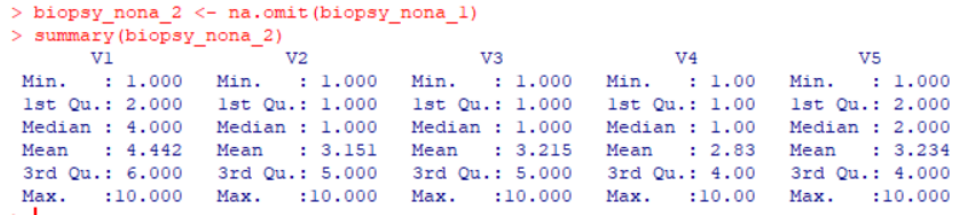
```
> summary(biopsy)
      ID                  V1               V2               V3               V4
 Length:699         Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
 Class :character   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000
 Mode  :character   Median : 4.000   Median : 1.000   Median : 1.000   Median : 1.000
                    Mean   : 4.418   Mean   : 3.134   Mean   : 3.207   Mean   : 2.807
                    3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.000
                    Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000

       V5               V6               V7               V8               V9
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
 1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
 Median : 2.000   Median : 1.000   Median : 3.000   Median : 1.000   Median : 1.000
 Mean   : 3.216   Mean   : 3.545   Mean   : 3.438   Mean   : 2.867   Mean   : 1.589
 3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 4.000   3rd Qu.: 1.000
 Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
                  NA's   :16
       class
 benign   :458
 malignant:241
```

Yes, there are 16 missing values spread in different rows.

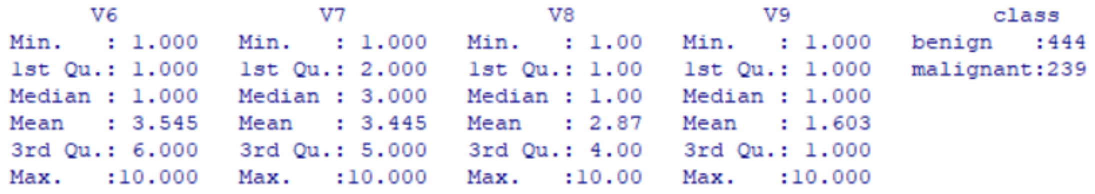
7. Eliminate the missing observations. (If any observation is missing in a row, the entire row is deemed missing.) (The R function complete.cases(biopsy) should help.) Or, find your own way to eliminate the missing observations. 4 points

I did this using na.omit.

> biopsy_nona_2 <- na.omit(biopsy_nona_1)



```
> biopsy_nona_2 <- na.omit(biopsy_nona_1)
> summary(biopsy_nona_2)
       V1               V2               V3               V4              V5
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.00   Min.   : 1.000
 1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.00   1st Qu.: 2.000
 Median : 4.000   Median : 1.000   Median : 1.000   Median : 1.00   Median : 2.000
 Mean   : 4.442   Mean   : 3.151   Mean   : 3.215   Mean   : 2.83   Mean   : 3.234
 3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.00   3rd Qu.: 4.000
 Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.00   Max.   :10.000
```



```
      V6              V7              V8             V9              class
Min.   : 1.000   Min.   : 1.000   Min.   : 1.00   Min.   : 1.000   benign   :444
1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.00   1st Qu.: 1.000   malignant:239
Median : 1.000   Median : 3.000   Median : 1.00   Median : 1.000
Mean   : 3.545   Mean   : 3.445   Mean   : 2.87   Mean   : 1.603
3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 4.00   3rd Qu.: 1.000
Max.   :10.000   Max.   :10.000   Max.   :10.00   Max.   :10.000
```

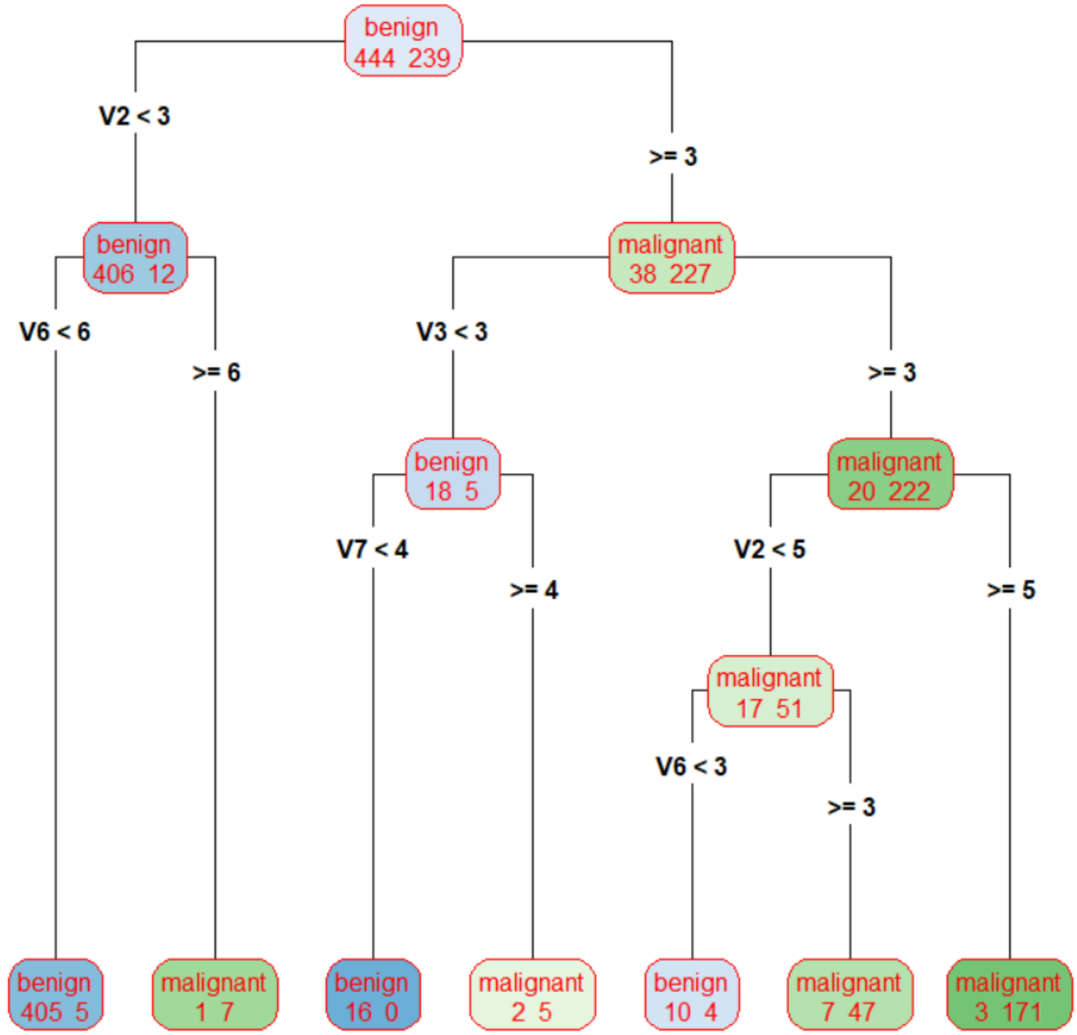
8. Build a classification tree. Show the tree. 3 points

I built the classification tree using the following lines of code:

MB <- rpart(class~., data=biopsy_nona_1)

> rpart.plot(MB, type = 4, extra =1, digits=3, col="red")

Output



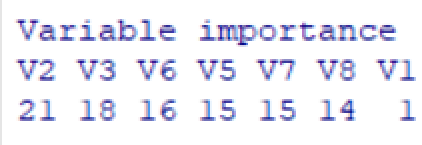
9. Calculate the misclassification rate. 2 points
Miscalculation Rate = 5+1+0+2+4+7+3/683 = 0.0322 or 3.22%

10. Provide a verbal description of the classification protocol of the tree. 5 points

- If V2 >=5 the cancer is malignant.
- If V2 >=3 and V2<5 and V6 >=3 the cancer is malignant.
- If V2 >=3 and V2<5 and V6 <3 the cancer is benign.
- If V2 >=3 and V3 <3 and V7 >= the cancer is malignant
- If V2 >=3 and V3 <3 and V7 < 4 the cancer is benign.
- If V2<3 and V6>=6 the cancer is malignant.
- If V2<3 and V6<6 the cancer is benign.

11. Identify the predictors that made a mark in the tree. 2 points

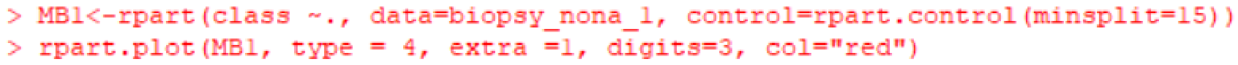
The most important predictor turned out to be V2 while the least important predictor was V1. As evident from the classification tree and the >summary(MB) command. The order of variable importance is shown below:



```
Variable importance
V2 V3 V6 V5 V7 V8 V1
21 18 16 15 15 14  1
```

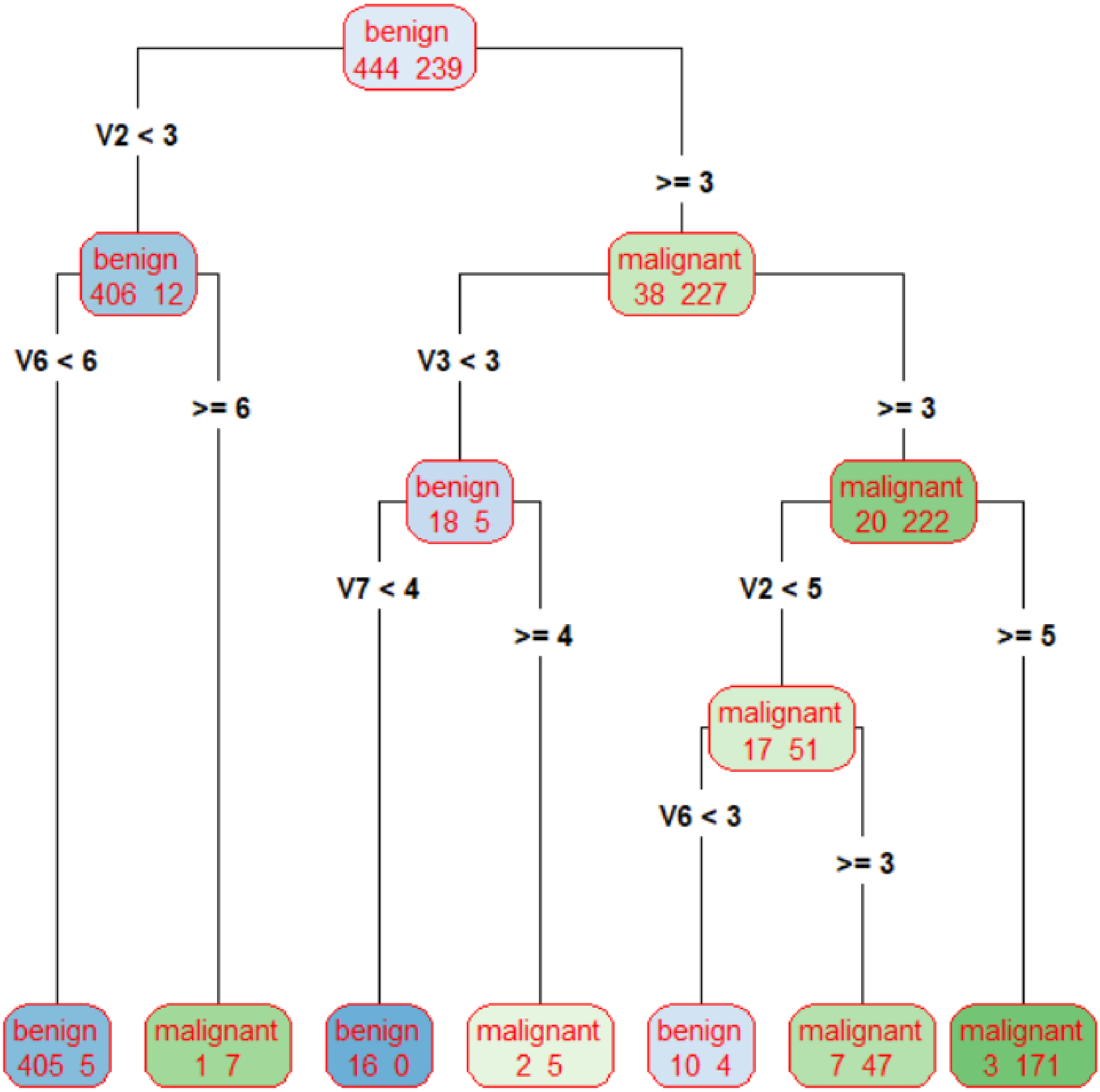
12. The default pruning principle stipulates that if the size of a node is 20 or less stop splitting the node. Suppose we change the size from 20 to 15. Explain how the tree changes and examine its impact on misclassification rate. 4 points

I used rpart.control to change the split size to 15 as shown below:



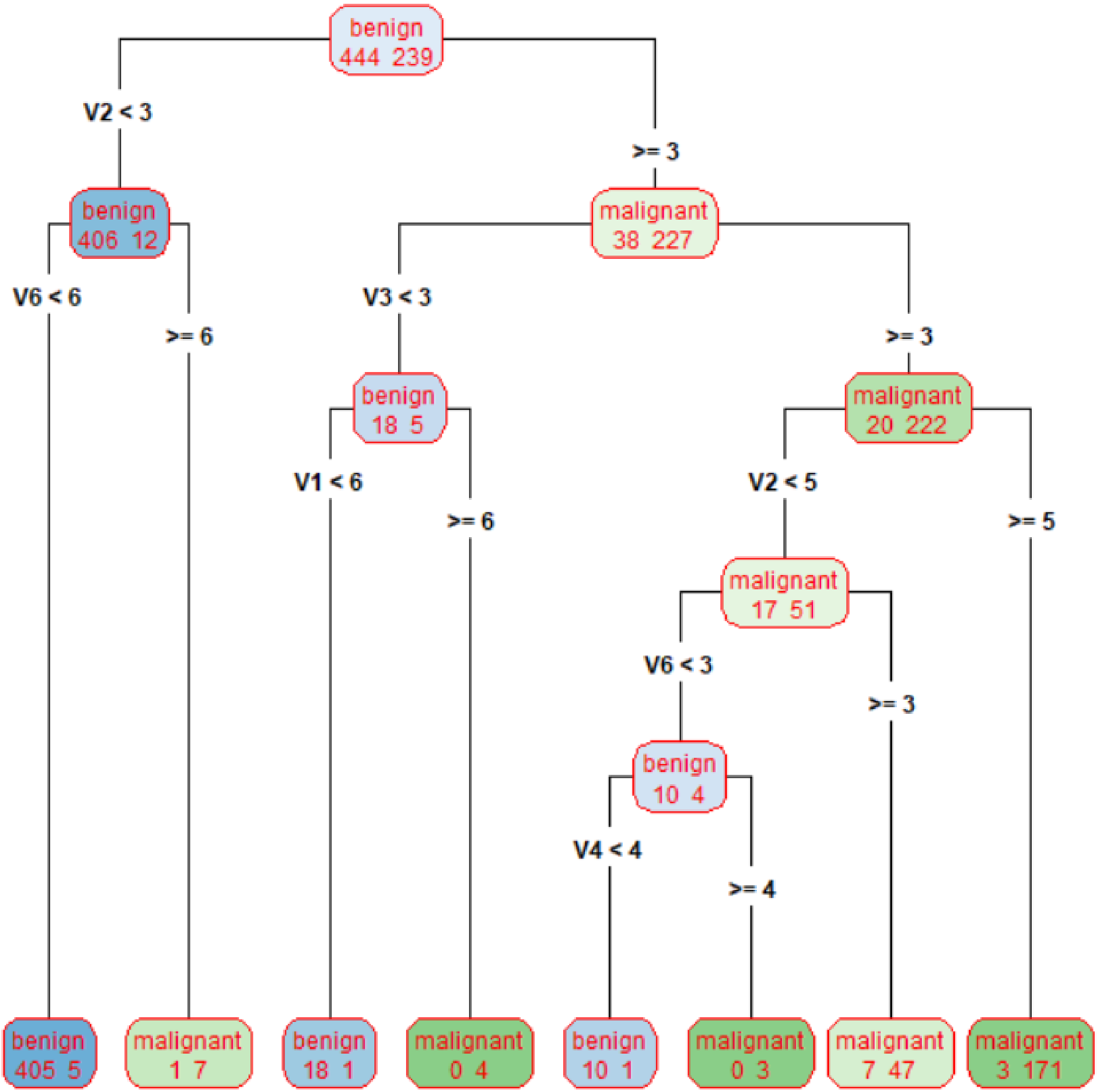
```
> MB1<-rpart(class ~., data=biopsy_nona_1, control=rpart.control(minsplit=15))
> rpart.plot(MB1, type = 4, extra =1, digits=3, col="red")
```

The classification tree obtained is shown below:



The misclassification rate remains the same as 22/683 = 3.22%

However, changing the minsplit node size to 10 results in the following classification tree that yields a Misclassification rate of 18/683 = 0.0263 or 2.63%



Hence a conclusion can be drawn from the above results that reducing the minsplit size reduces the misclassification rate. This leads to an improved prediction of the target with the same data and same variables.