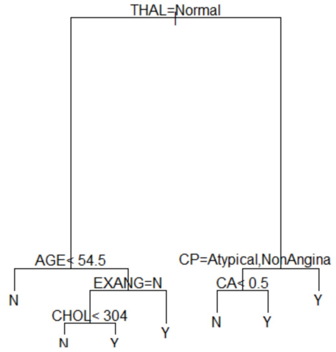
Question 1

1. The best tree used the following variables:
   THAL, THALACH, CA, EXANG, OLDPEAK, CP, AGE, SEX, CHOL, TRESTBPS

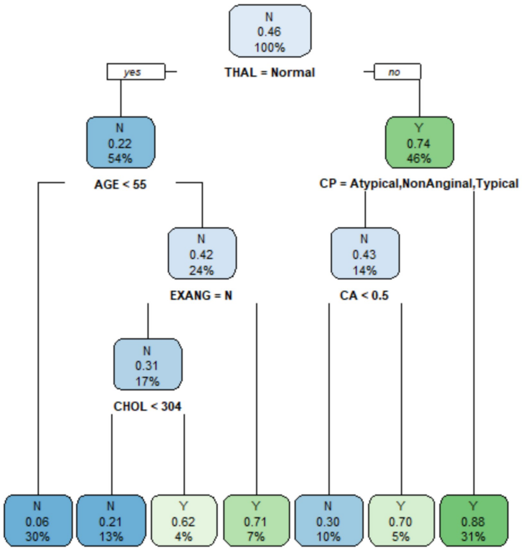
   The tree have 7 leaves.

2. The plot of tree:



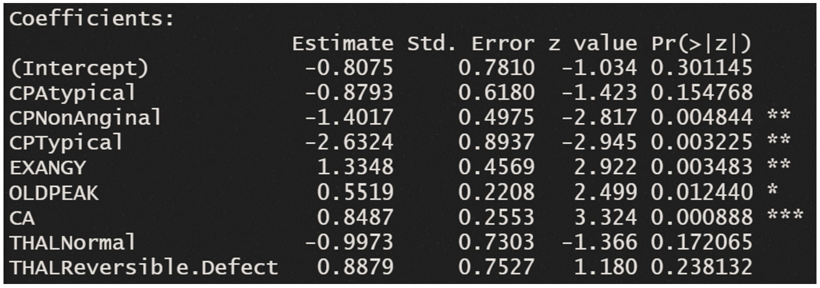
   People whos THAL are normal, age < 55 do not have heart disease.
   People whos THAL are normal, age >= 55, do EXANG and CHOL < 304 do not have heart disease
   People whos THAL are normal, age >= 55, do EXANG and CHOL > 304 have heart disease.
   People whos THAL are normal, age >= 55 and do not EXANG have heart disease.
   People whos THAL are not normal, CP = Atypical/NonAnginal/Typical and CA < 0.5 do not have heart disease.
   People whos THAL are not normal, CP = Atypical/NonAnginal/Typical and CA >= 0.5 do have heart disease.
   People whos THAL are not normal and CP is not Atypical/NonAnginal/Typical have heart disease.

3. The graph:



4. People whos THAL are normal, age >= 55 and do not EXANG will have 71% of chance to have heart disease.

5. According to the output of the code:



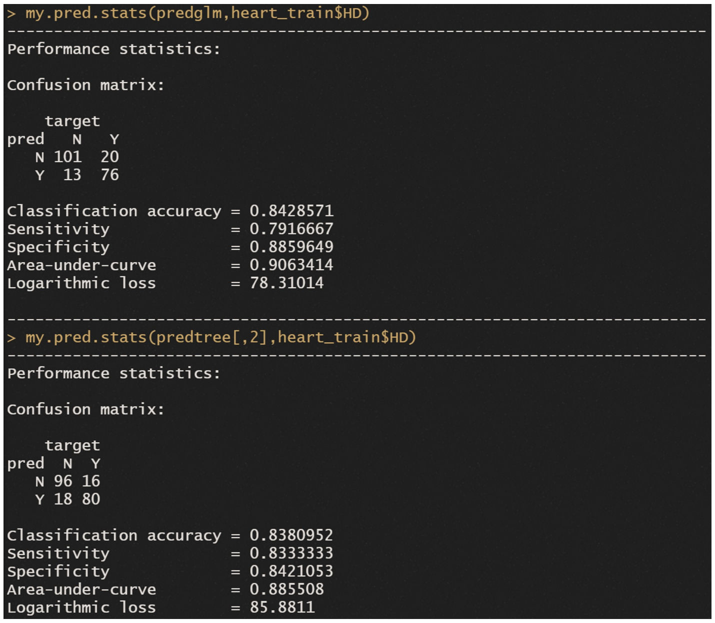
```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.8075     0.7810  -1.034 0.301145
CPAtypical             -0.8793     0.6180  -1.423 0.154768
CPNonAnginal           -1.4017     0.4975  -2.817 0.004844 **
CPTypical              -2.6324     0.8937  -2.945 0.003225 **
EXANGY                  1.3348     0.4569   2.922 0.003483 **
OLDPEAK                 0.5519     0.2208   2.499 0.012440 *
CA                      0.8487     0.2553   3.324 0.000888 ***
THALNormal             -0.9973     0.7303  -1.366 0.172065
THALReversible.Defect   0.8879     0.7527   1.180 0.238132
```

CA is the most important predictor in the logistic regression.

6. The regression equation of the logistic regression model:

$$P(\text{HD}' = Y \mid \text{CPAtypical}, \text{CPNonAnginal}, \text{CPTypical}. \text{EXANGY}, \text{OLDPEAK}, \text{CA}, \text{THALNormal}, \text{THALReversible.Defect} = \frac{\quad}{1 + \exp(-(0.8075 + 0.8793 \times \text{CPAtypical} + 1.4017 \times \text{CPNonAnginal} + 2.6324 \times \text{CPTypical} + 1.3348 \times \text{EXANGY} + 0.5519 \times \text{OLDPEAK} + 0.8487 \times \text{CA} + 0.9973 \times \text{THALNormal} + 0.8879 \times \text{THALReversible.Defect}.}$$

7. According to the output:



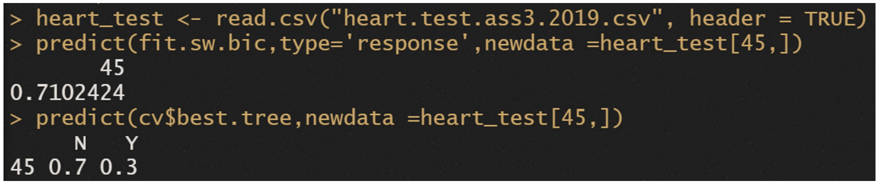
```
> my.pred.stats(predglm,heart_train$HD)
---------------------------------------------------------------------------
Performance statistics:

Confusion matrix:

    target
pred   N   Y
   N 101  20
   Y  13  76

Classification accuracy = 0.8428571
Sensitivity             = 0.7916667
Specificity             = 0.8859649
Area-under-curve        = 0.9063414
Logarithmic loss        = 78.31014

---------------------------------------------------------------------------
> my.pred.stats(predtree[,2],heart_train$HD)
---------------------------------------------------------------------------
Performance statistics:

Confusion matrix:

    target
pred  N  Y
   N 96 16
   Y 18 80

Classification accuracy = 0.8380952
Sensitivity             = 0.8333333
Specificity             = 0.8421053
Area-under-curve        = 0.885508
Logarithmic loss        = 85.8811
```

Logistic Regression model has better accuracy, specificity and AUC. And tree model got better sensitivity.
In high sensitivity model, the person who got heart disease is most likely to be diagnosed as a patient. In high specificity model, the person who do not have heart disease is most likely to be diagnosed as not a patient. In another word, High sensitivity model focus on lower rate of missed diagnosis, where high specificity model focus on lower misdiagnosis rate.

Heart disease is a very serious disease, missed diagnosis is way better than misdiagnosis if a person really have a heart disease.

So the tree model would be more preferable.

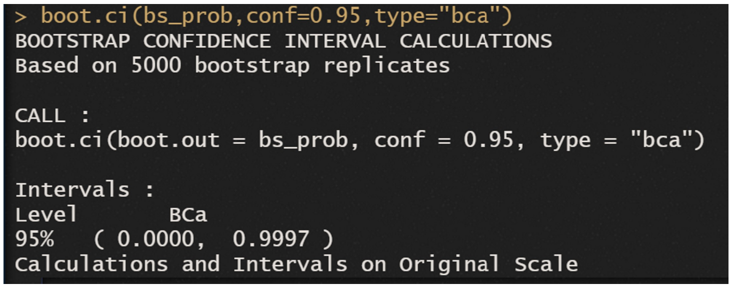
8. The probability of two different model:



```
> heart_test <- read.csv("heart.test.ass3.2019.csv", header = TRUE)
> predict(fit.sw.bic,type='response',newdata =heart_test[45,])
       45
0.7102424
> predict(cv$best.tree,newdata =heart_test[45,])
     N   Y
45 0.7 0.3
```

Logistic Regression model suggests the person has 71% chance to have heart disease.
Tree model suggests that the person only has 30% chance to have heart disease.

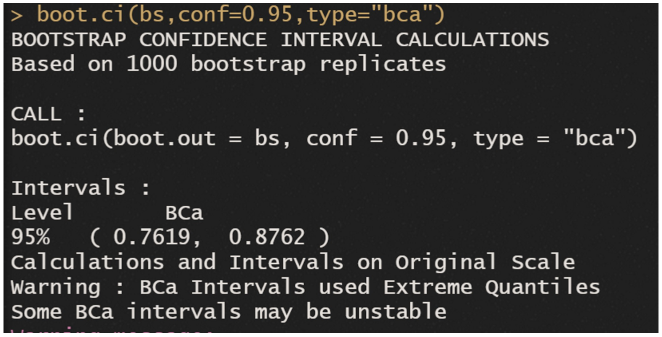
9. The output of confidence interval:



```
> boot.ci(bs_prob,conf=0.95,type="bca")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

CALL :
boot.ci(boot.out = bs_prob, conf = 0.95, type = "bca")

Intervals :
Level       BCa
95%   ( 0.0000,  0.9997 )
Calculations and Intervals on Original Scale
```

We are 95% confident that the probability of the patient at 45th row have heart disease is between 0.00 and 0.99.

10. The output of confidence interval:



```
> boot.ci(bs,conf=0.95,type="bca")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = bs, conf = 0.95, type = "bca")

Intervals :
Level       BCa
95%   ( 0.7619,  0.8762 )
Calculations and Intervals on Original Scale
Warning : BCa Intervals used Extreme Quantiles
Some BCa intervals may be unstable
```

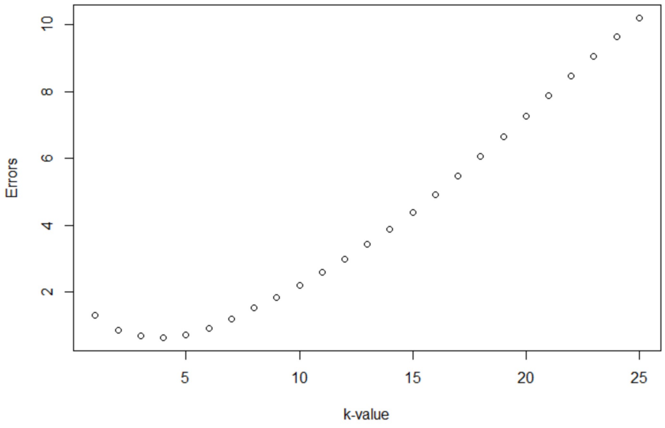
According to the output, we are 95% confidence that the classification accuracy is between 0.7619 and 0.8762

Question 2

1. a.

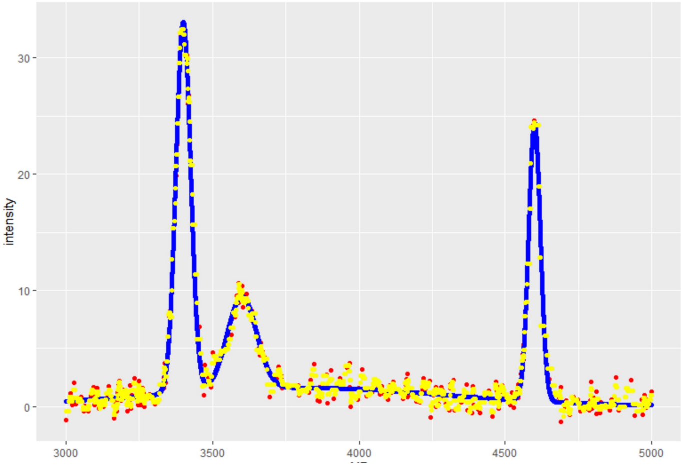
Mean square error = 0.6946

Plot of Error vs k-value:

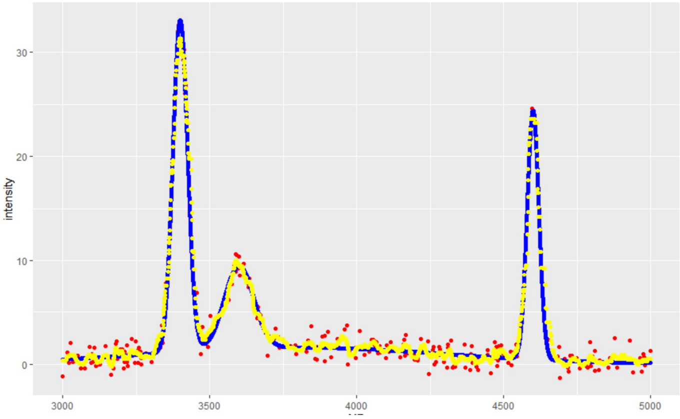


b.

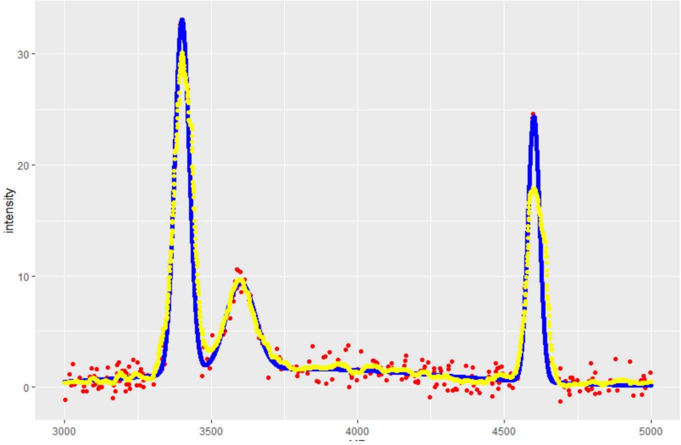
Plot of k = 2:



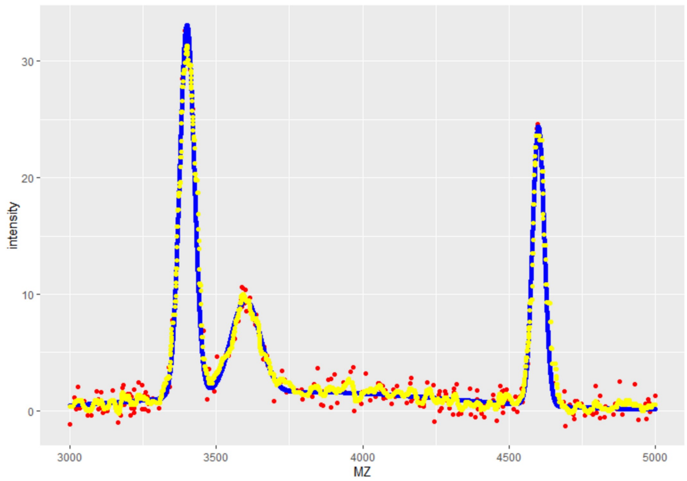
Plot of k = 5:



Plot of k = 10:



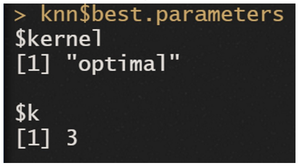
Plot of k = 25:



c.
According to the plot, when k = 10, the data fits the best due to its smoother. When k = 10 or 25, the fit is less discrete, so there's less noise.

2. According to the output:

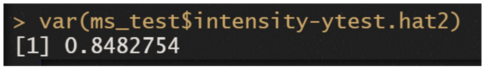
3.



```
> knn$best.parameters
$kernel
[1] "optimal"

$k
[1] 3
```

The method select 3 as k value.

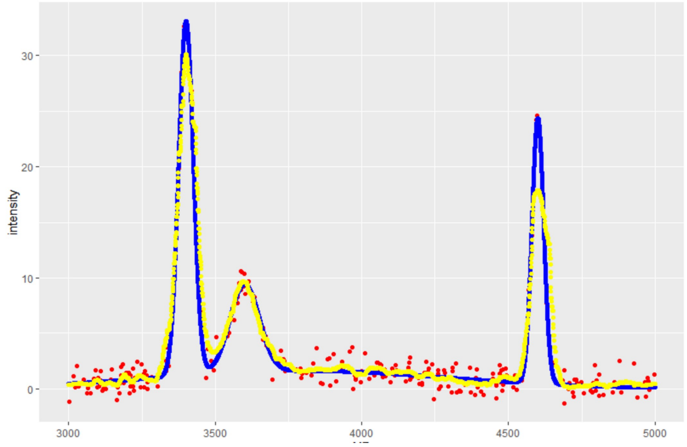
3. According to the output:



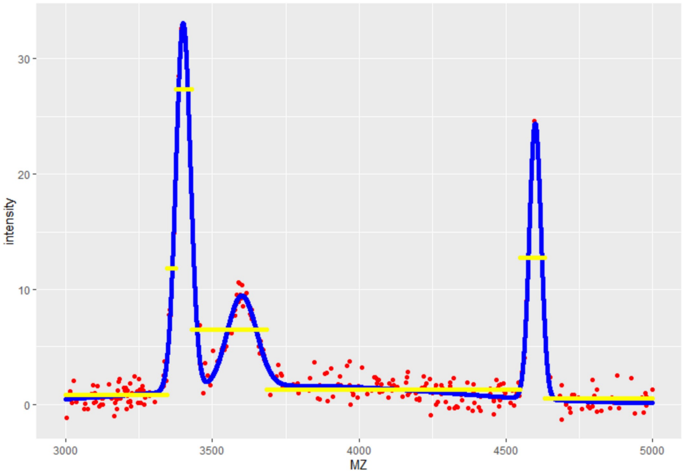
```
> var(ms_test$intensity-ytest.hat2)
[1] 0.8482754
```

The variance is 0.848

4. The estimated spectra of k value = 10 provides a respectively smooth, low-noise estimate. Due to the data are mostly well fitted on the curve as the following graph shows



5. The plot produced:



The knn method fits better due to the tree spectra is completely discrete where knn fits a relatively smooth trend with lower noise.

6. Knn method is better than tree in this scenario.