FIT2086 Assignment 3

Name: Haw Xiao Ying

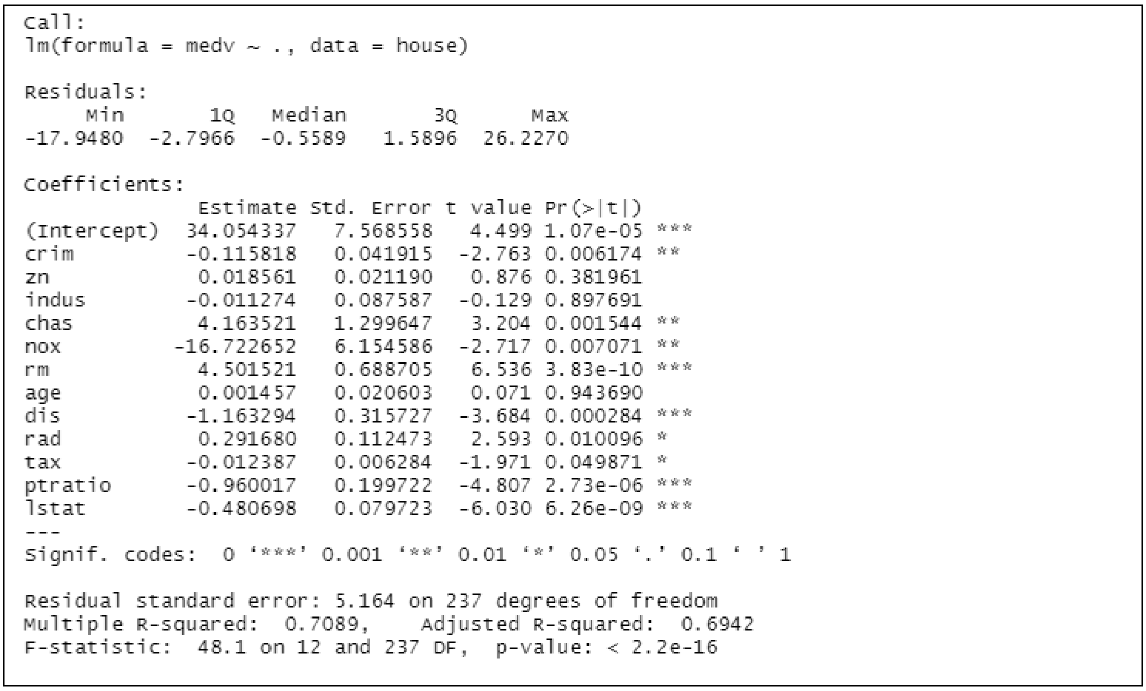
Student ID: 29797918

Question 1

1) R code that fits a multiple linear model to the housing data:

```
house = read.csv('housing.ass3.2020.csv')
fit = lm(medv~.,house)
summary(fit)
```

R output from fitting a multiple linear model to medv:



```
Call:
lm(formula = medv ~ ., data = house)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.9480  -2.7966  -0.5589   1.5896  26.2270 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  34.054337   7.568558   4.499 1.07e-05 ***
crim         -0.115818   0.041915  -2.763 0.006174 ** 
zn            0.018561   0.021190   0.876 0.381961    
indus        -0.011274   0.087587  -0.129 0.897691    
chas          4.163521   1.299647   3.204 0.001544 ** 
nox         -16.722652   6.154586  -2.717 0.007071 ** 
rm            4.501521   0.688705   6.536 3.83e-10 ***
age           0.001457   0.020603   0.071 0.943690    
dis          -1.163294   0.315727  -3.684 0.000284 ***
rad           0.291680   0.112473   2.593 0.010096 *  
tax          -0.012387   0.006284  -1.971 0.049871 *  
ptratio      -0.960017   0.199722  -4.807 2.73e-06 ***
lstat        -0.480698   0.079723  -6.030 6.26e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.164 on 237 degrees of freedom
Multiple R-squared:  0.7089,	Adjusted R-squared:  0.6942 
F-statistic:  48.1 on 12 and 237 DF,  p-value: < 2.2e-16
```

The predictors that are possibly associated with median house value (medv) included crim, chas, nox, rm, dis, rad, tax, ptratio and lstat since their p-values are small. From this output, it appears that the predictors mentioned above are associated with median value of owner-occupied homes in $1,000s (medv). We can say that if those predictors were not associated with median house value (medv) at the population level, the chance of seeing an association as strong as the one we have observed, or stronger, just by chance is so unlikely. This data therefore offers a strong evidence against the null hypothesis that those predictors and median house value (medv) are not associated. Therefore, we can conclude that crim, chas, nox, rm, dis, rad, tax, ptratio and lstat are possibly associated with median house value (medv) due to the small p-value (< 0.05) they have.

We see that the three predictors: average number of rooms per dwelling (rm), percentage of “lower status” of the population (lstat) and pupil-teacher ratio (ptratio) have the three smallest p-value among all of others and hence we can say that these three variables have strong association with the median house value (medv). In conclusion, these three variables appear to be the strongest predictors of housing price.

2) Let α be our significance level where $\alpha = 0.05$. Then, if we do p different tests (where p=12 in this case since there are 12 predictors) then the Bonferroni procedure says we should only reject the null hypothesis for each of the tests if p-value $< \alpha/\text{p}$.

$$\alpha/\text{p} = 0.05/12 = 0.0042$$

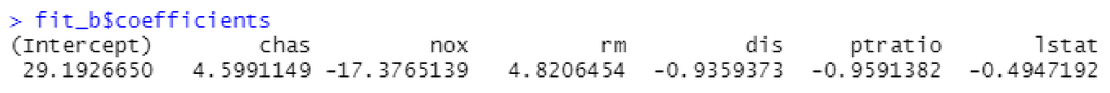
Since the p-value needs to be less than 0.0042, the predictors left are: does the suburb front the Charles River (chas), average number of rooms per dwelling (rm), weighted distances to five Boston employment centres (dis), percentage of “lower status” of the population (lstat) and pupil-teacher ratio (ptratio).

3) The per-capita crime rate (crim) seems to affect the median house value to become less valuable if the crime rate increases. The effect a suburb having frontage on the Charles River (chas) has on the median house price (medv) for that suburb will be the median house value (medv) is higher.

4) Code to prune out potentially unimportant variables with BIC:

```
fit_b = step(fit, k=log(length(house$medv)))
fit_b$coefficients
```

Results:



```
> fit_b$coefficients
(Intercept)        chas         nox          rm         dis     ptratio       lstat 
 29.1926650   4.5991149 -17.3765139   4.8206454  -0.9359373  -0.9591382  -0.4947192 
```

The final regression equation obtained after pruning:

E[medv] = 29.193 + 4.599*chas – 17.377*nox + 4.821*rm – 0.936*dis – 0.959*ptratio – 0.495*lstat

5) To improve the median house value in their suburb, the suburb needs to have frontage on the Charles River, decrease the nitric oxides concentration, increase the average number of rooms per dwelling, decrease the weighted distances to five Boston employment centres, decrease the pupil-teacher ratio and decrease the percentage of “lower status” of the population.

6) $Median\ house\ value = 29.193 + 4.599(0) - 17.377(0.573) + 4.821(6.03) - 0.936(2.505) - 0.959(21) - 0.495(7.88)$

$Median\ house\ value = 29.193 - 9.957021 + 29.07063 - 2.34468 - 20.139 - 3.9006$

$Median\ house\ value = 21.922329$

Predicted median house price is \$21,922.

Code to find the variance of the dataset:

```
var(house$medv)
```

Result:

```
[1] 87.19116
```

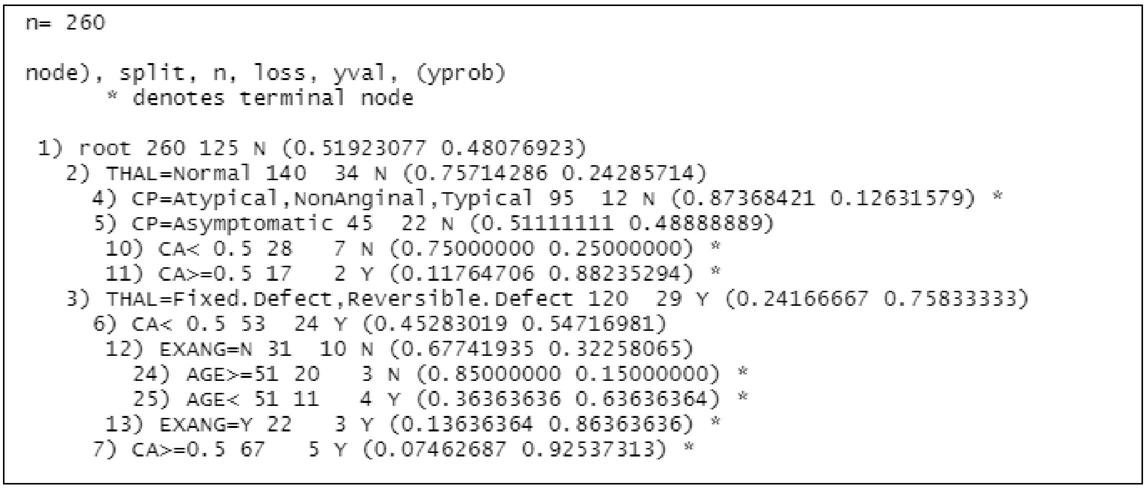
The 95% confidence interval for the predicted median house price can be calculated as below:

$$\left(21.922329 - 1.96\left(\sqrt{\frac{87.19116}{250}}\right), 21.922329 + 1.96\left(\sqrt{\frac{87.19116}{250}}\right)\right)$$

which is (20.7648, 23.0798).

Question 2

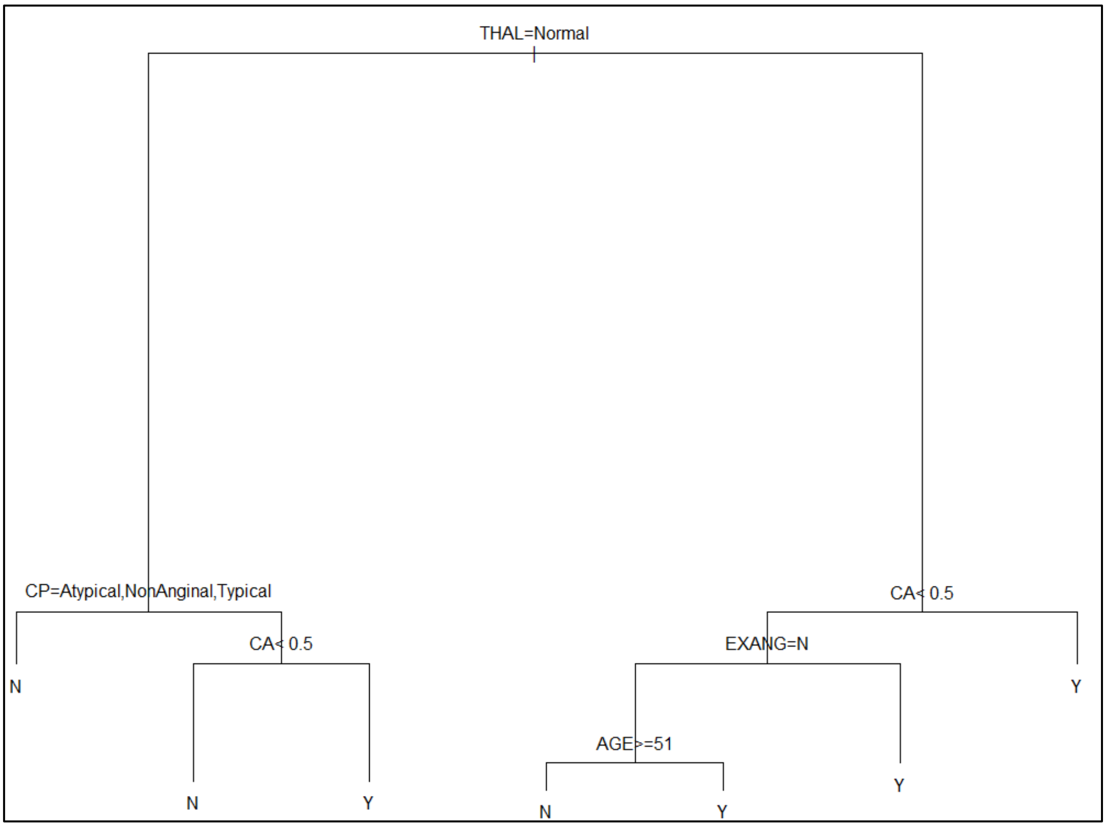
1) Result:



```
n= 260

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 260 125 N (0.51923077 0.48076923)
   2) THAL=Normal 140  34 N (0.75714286 0.24285714)
     4) CP=Atypical,NonAnginal,Typical 95  12 N (0.87368421 0.12631579) *
     5) CP=Asymptomatic 45  22 N (0.51111111 0.48888889)
      10) CA< 0.5 28   7 N (0.75000000 0.25000000) *
      11) CA>=0.5 17   2 Y (0.11764706 0.88235294) *
   3) THAL=Fixed.Defect,Reversible.Defect 120  29 Y (0.24166667 0.75833333)
     6) CA< 0.5 53  24 Y (0.45283019 0.54716981)
      12) EXANG=N 31  10 N (0.67741935 0.32258065)
        24) AGE>=51 20   3 N (0.85000000 0.15000000) *
        25) AGE< 51 11   4 Y (0.36363636 0.63636364) *
      13) EXANG=Y 22   3 Y (0.13636364 0.86363636) *
     7) CA>=0.5 67   5 Y (0.07462687 0.92537313) *
```

According to the result produced above, variables that have been used in the best tree include THAL, CP, CA, EXANG and AGE. The best tree has 7 leaves (terminal nodes).

2) The tree below shows the relationship between the predictors and heart disease:



From the tree above, we can interpret the relationship between the predictors and heart disease as below:

- ✓ If the Thallium scanning results (THAL) by a patient is **normal**, chest pain type (CP) is **atypical, nonanginal, typical**, the patient will probably have **no heart disease**.

- ✓ If the Thallium scanning results (THAL) by a patient is **normal**, chest pain type (CP) is **not atypical, nonanginal, typical** and their number of major vessels coloured by fluoroscopy (CA) is **less than 0.5**, they probably have **no heart disease**.

- ✓ If the Thallium scanning results (THAL) by a patient is **normal**, chest pain type (CP) is **not atypical, nonanginal, typical** and their number of major vessels coloured by fluoroscopy (CA) is **not less than 0.5**, they probably **have a heart disease**.

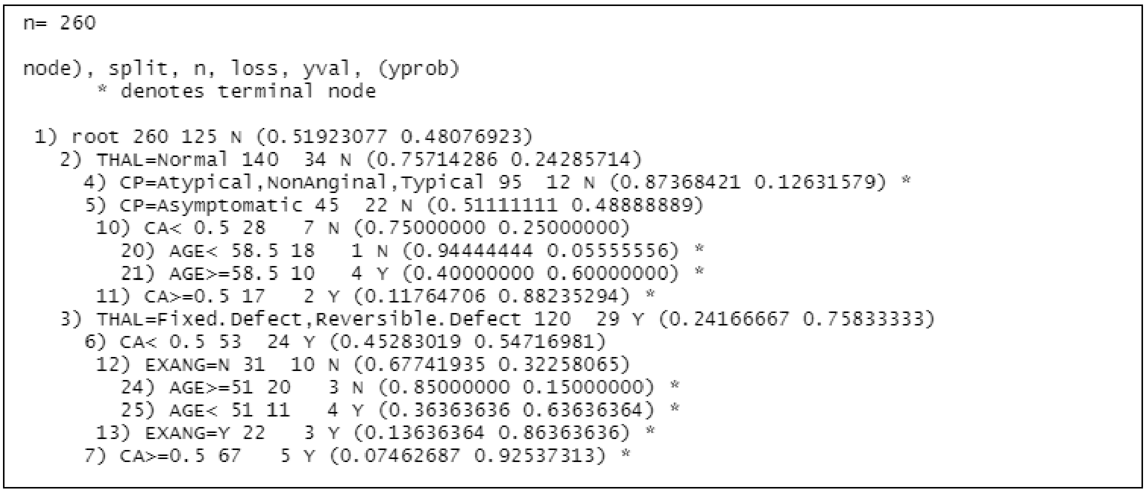
- ✓ If the Thallium scanning results (THAL) by a patient is **not normal** and their number of major vessels coloured by fluoroscopy (CA) is **not less than 0.5**, they will probably **have a heart disease**.

- ✓ If the Thallium scanning results (THAL) by a patient is **not normal**, their number of major vessels coloured by fluoroscopy (CA) is **less than 0.5** and **exercise included angina** (EXANG=Y), they will **have a heart disease**.

- ✓ If the Thallium scanning results (THAL) by a patient is **not normal**, their number of major vessels coloured by fluoroscopy (CA) is **less than 0.5**, **exercise did not include angina** (EXANG=N) and their age (AGE) is **not more or equals to 51**, they will probably **have a heart disease**.

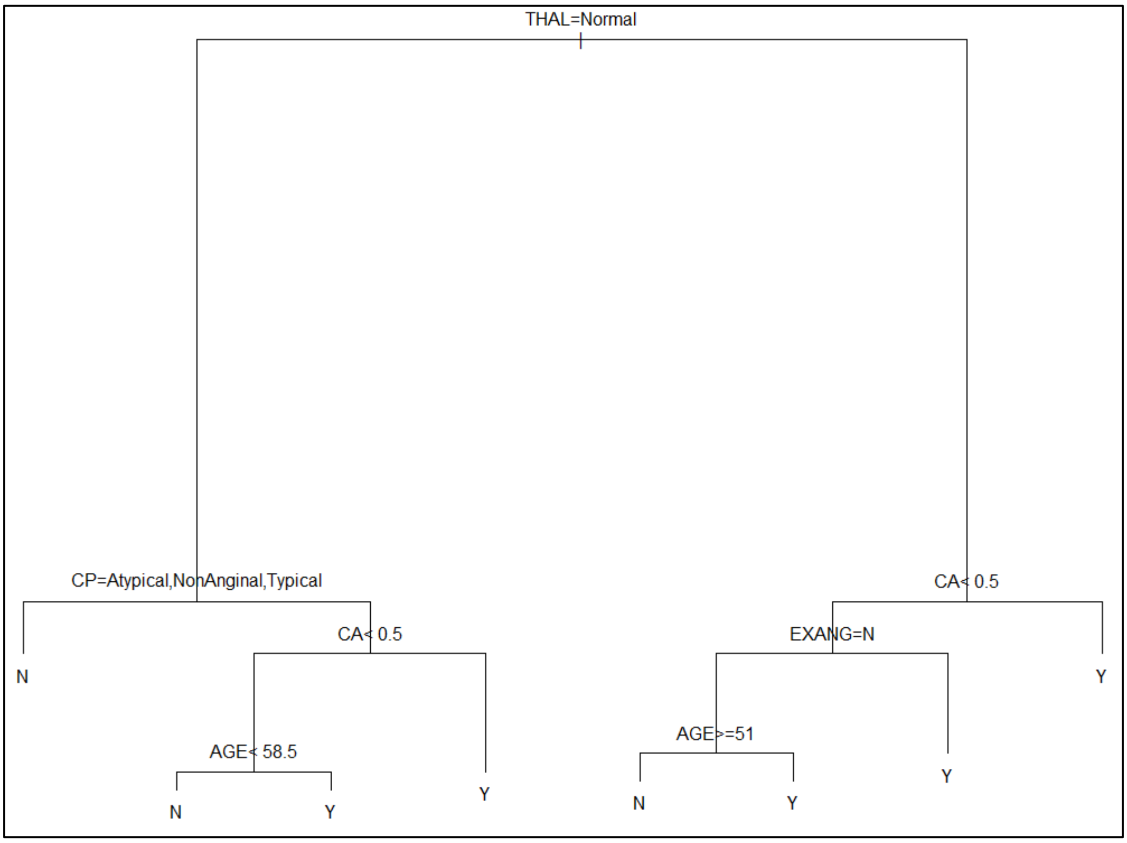
- ✓ If the Thallium scanning results (THAL) by a patient is **not normal**, their number of major vessels coloured by fluoroscopy (CA) is **less than 0.5**, **exercise does not include angina** (EXANG=N) and their age (AGE) is **more or equals to 51**, they will probably **not be having heart disease**.

3) Textual representation of the tree (the tree is not pruned since the question does not mention it):

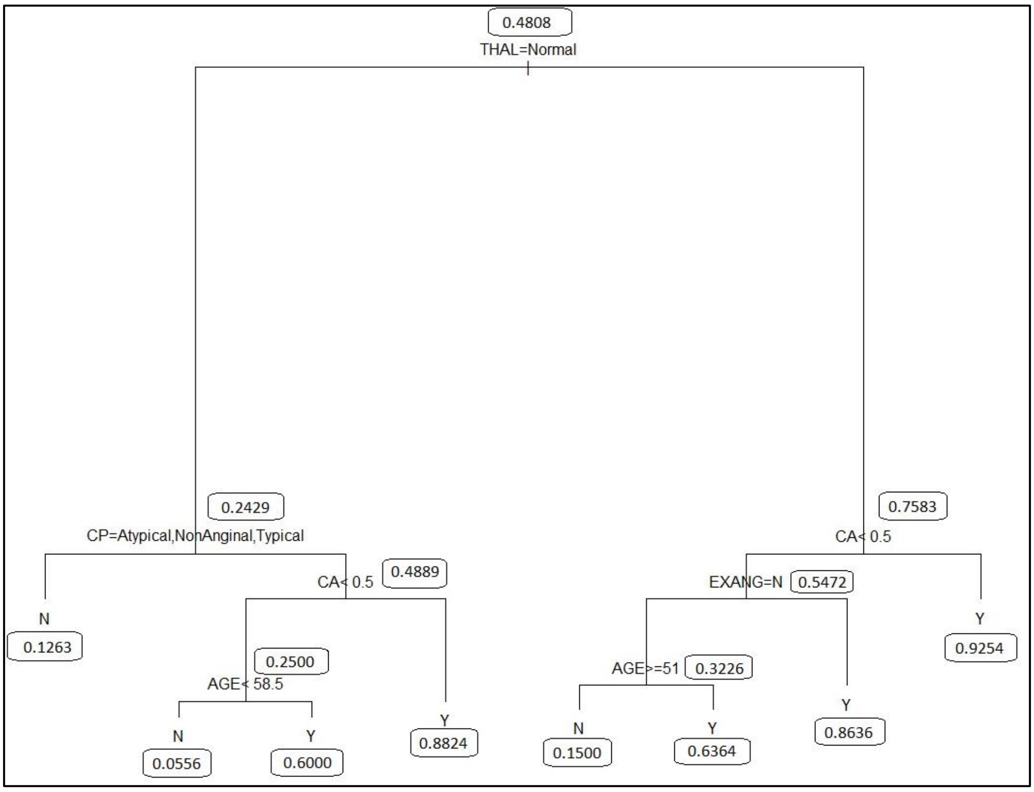


```
n= 260

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 260 125 N (0.51923077 0.48076923)
   2) THAL=Normal 140  34 N (0.75714286 0.24285714)
     4) CP=Atypical,NonAnginal,Typical 95  12 N (0.87368421 0.12631579) *
     5) CP=Asymptomatic 45  22 N (0.51111111 0.48888889)
      10) CA< 0.5 28   7 N (0.75000000 0.25000000)
        20) AGE< 58.5 18   1 N (0.94444444 0.05555556) *
        21) AGE>=58.5 10   4 Y (0.40000000 0.60000000) *
      11) CA>=0.5 17   2 Y (0.11764706 0.88235294) *
   3) THAL=Fixed.Defect,Reversible.Defect 120  29 Y (0.24166667 0.75833333)
     6) CA< 0.5 53  24 Y (0.45283019 0.54716981)
      12) EXANG=N 31  10 N (0.67741935 0.32258065)
        24) AGE>=51 20   3 N (0.85000000 0.15000000) *
        25) AGE< 51 11   4 Y (0.36363636 0.63636364) *
      13) EXANG=Y 22   3 Y (0.13636364 0.86363636) *
     7) CA>=0.5 67   5 Y (0.07462687 0.92537313) *
```

The plot of the tree:

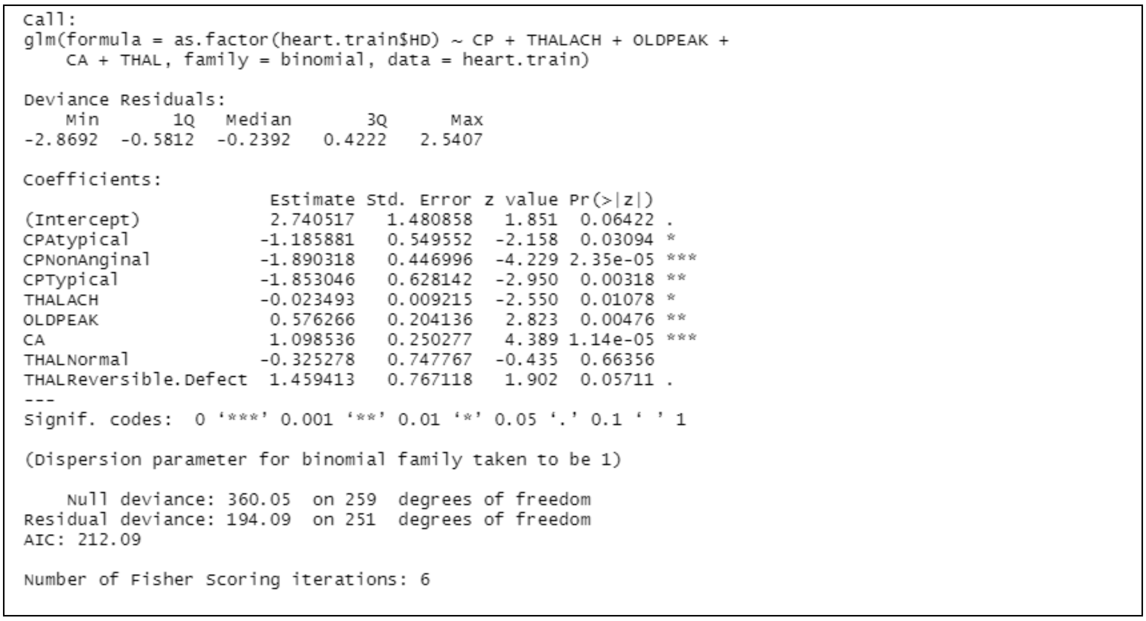


Annotated plot of tree:



4) The combination when the Thallium scanning results (THAL) by a patient is not normal and their number of major vessels coloured by fluoroscopy (CA) is not less than 0.5 gives the highest probability (0.9254) of having a heart disease.

5) Result:

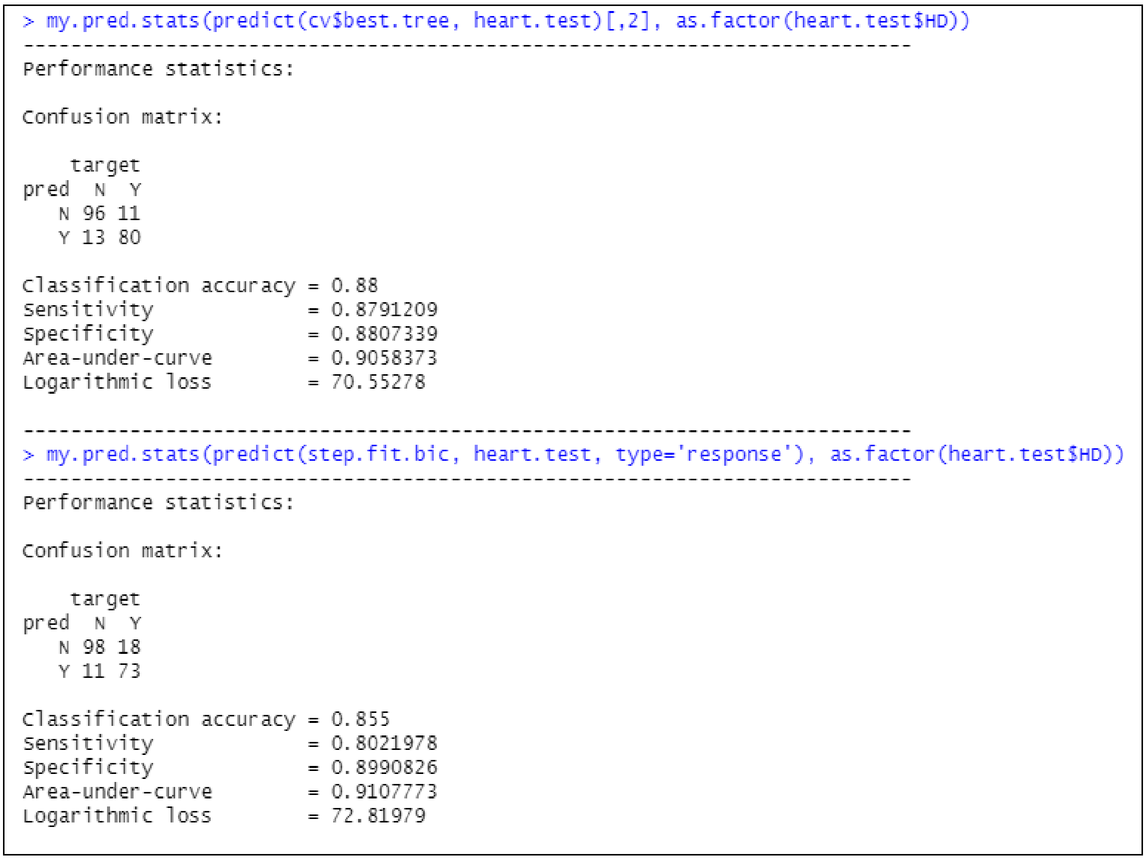


```
Call:
glm(formula = as.factor(heart.train$HD) ~ CP + THALACH + OLDPEAK +
    CA + THAL, family = binomial, data = heart.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8692  -0.5812  -0.2392   0.4222   2.5407

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.740517   1.480858   1.851  0.06422 .
CPAtypical            -1.185881   0.549552  -2.158  0.03094 *
CPNonAnginal          -1.890318   0.446996  -4.229 2.35e-05 ***
CPTypical             -1.853046   0.628142  -2.950  0.00318 **
THALACH               -0.023493   0.009215  -2.550  0.01078 *
OLDPEAK                0.576266   0.204136   2.823  0.00476 **
CA                     1.098536   0.250277   4.389 1.14e-05 ***
THALNormal            -0.325278   0.747767  -0.435  0.66356
THALReversible.Defect  1.459413   0.767118   1.902  0.05711 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 360.05  on 259  degrees of freedom
Residual deviance: 194.09  on 251  degrees of freedom
AIC: 212.09

Number of Fisher Scoring iterations: 6
```

The variables that include in the final model are CP, THALACH, OLDPEAK, CA and THAL. Compare to the variables used by the tree estimated by CV, it has extra variables such as THALACH and OLDPEAK. Besides the extra variables, it does not have variables such as EXANG and AGE. CA is the most important variable in the logistic regression. We can determine this by looking at the value of their Pr(>|z|), the smaller the value, the more important it would be.

6) $Regression\ equation = 2.7405 - CPAtypical \times (1.1859) - CPNonAnginal \times (1.8903) - CPTypical \times (1.8530) - THALACH \times (0.0235) + OLDPEAK \times (0.5763) + CA \times (1.0985) - THALNormal \times (0.3253) + THALReversible.Defect \times (1.4594)$

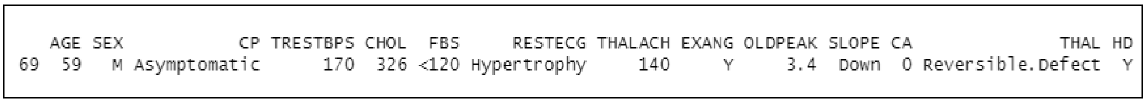
7) Result



```
> my.pred.stats(predict(cv$best.tree, heart.test)[,2], as.factor(heart.test$HD))
---------------------------------------------------------------------------
Performance statistics:

Confusion matrix:

    target
pred  N  Y
   N 96 11
   Y 13 80

Classification accuracy = 0.88
Sensitivity             = 0.8791209
Specificity             = 0.8807339
Area-under-curve        = 0.9058373
Logarithmic loss        = 70.55278

---------------------------------------------------------------------------
> my.pred.stats(predict(step.fit.bic, heart.test, type='response'), as.factor(heart.test$HD))
---------------------------------------------------------------------------
Performance statistics:

Confusion matrix:

    target
pred  N  Y
   N 98 18
   Y 11 73

Classification accuracy = 0.855
Sensitivity             = 0.8021978
Specificity             = 0.8990826
Area-under-curve        = 0.9107773
Logarithmic loss        = 72.81979
```

The classification accuracy of the tree model found using cross-validation is 0.88 which is higher (better) than the step-wise logistic regression model's classification accuracy, 0.855. Next, the tree model also has a better sensitivity since it has a greater value compared to the logistic regression model (0.8791>0.8022). For the specificity, the logistic regression model is now slightly better than the tree model as it has a specificity of 0.8991 while the tree model has a specificity of 0.8807. The logistic regression model also has a slightly better value of area-under-curve which is 0.9107 whereas the tree model has a value of 0.9058. Lastly, the logarithmic loss for the logistic regression model is more than the tree model's, which is bad in this case because a model with smaller logarithmic loss does a better job of estimating the probabilities of an individual being in one of the two target classes than a model with a larger logarithmic loss. In conclusion, the tree model found using cross validation does a better work than the logistic regression model since it has a better value of classification accuracy, sensitivity and logarithmic loss. Moreover, the specificity and are-under-curve are just very slightly worse than the step-wise logistic regression model. Therefore, I would say that the tree model found using cross validation is more preferred as a diagnostic test.

8) Data of row 69th:



```
   AGE SEX            CP TRESTBPS CHOL  FBS     RESTECG THALACH EXANG OLDPEAK SLOPE CA              THAL HD
69  59   M Asymptomatic      170  326 <120 Hypertrophy     140     Y     3.4  Down  0 Reversible.Defect  Y
```

(a) Tree model found using cross validation

Probability having a heart disease of the combination above = 0.8636

Probability not having a heart disease of the combination above = 1-0.8636 = 0.1364

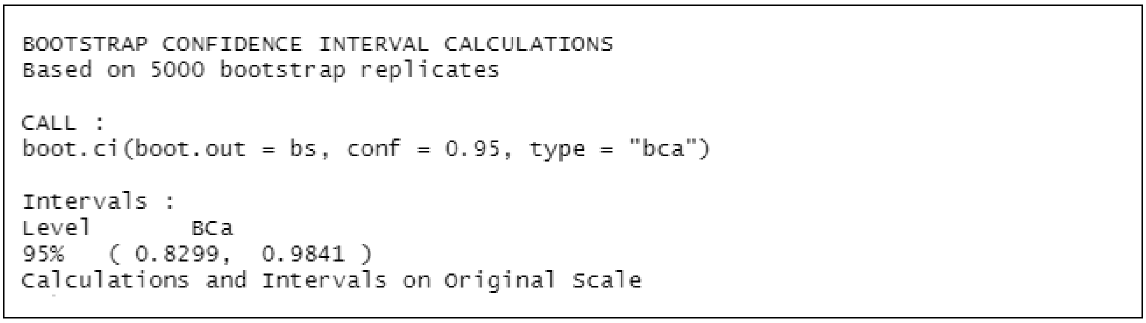
Odds of having heart disease for the patient in row 69th $= \frac{0.8636}{0.1364} = 6.3314$

(b) Step-wise logistic regression model

Odds of having heart disease for the patient in row 69th = 17.6397

The odds of having heart disease for the patient in the 69th row of the test dataset for the tree model found using cross validation is 6.3314, which is lesser than the odds of having heart disease for step-wise logistic regression model that has a value of 17.6397. Since the patient of row 69th is having a heart disease, the predicted odds of having heart disease for the patient in the 69th row of the test dataset for the logistic regression model seems to have a better result than the tree model because it has a higher odds value.

9) Result:



```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

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

Intervals :
Level       BCa
95%   ( 0.8299,  0.9841 )
Calculations and Intervals on Original Scale
```

The 95% confidence interval for the probability of having heart disease for patient in the 69th row in the test data is (0.8299, 0.9841). The predicted probability of having heart disease using the tree model is 0.8636, which is within the 95% confidence interval. For the logistic regression model, the probability can be calculated using the odds:

$$p/(1-p) = 17.6397$$

$$p = 17.6397 - 17.6397p$$

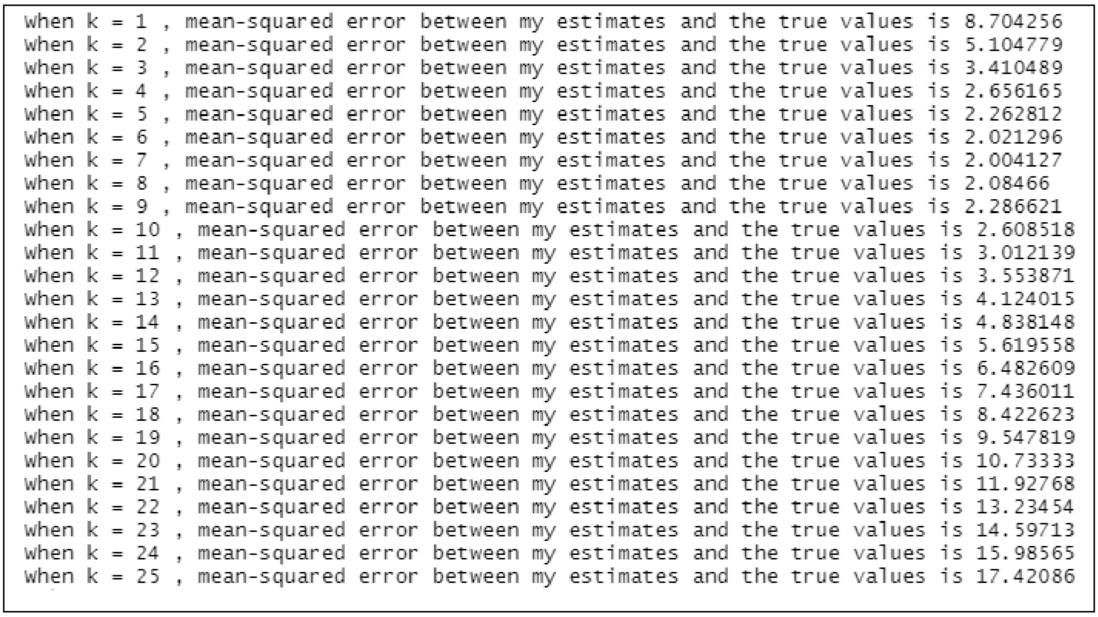
$$p = 0.9463511$$

As it is shown, the predicted probability of having heart disease for logistic regression model is also within the 95% confidence interval. The predicted probability using the tree model seems to be more accurate as it is more to the centre of the range 0.7850 to 0.9878.

Question 3

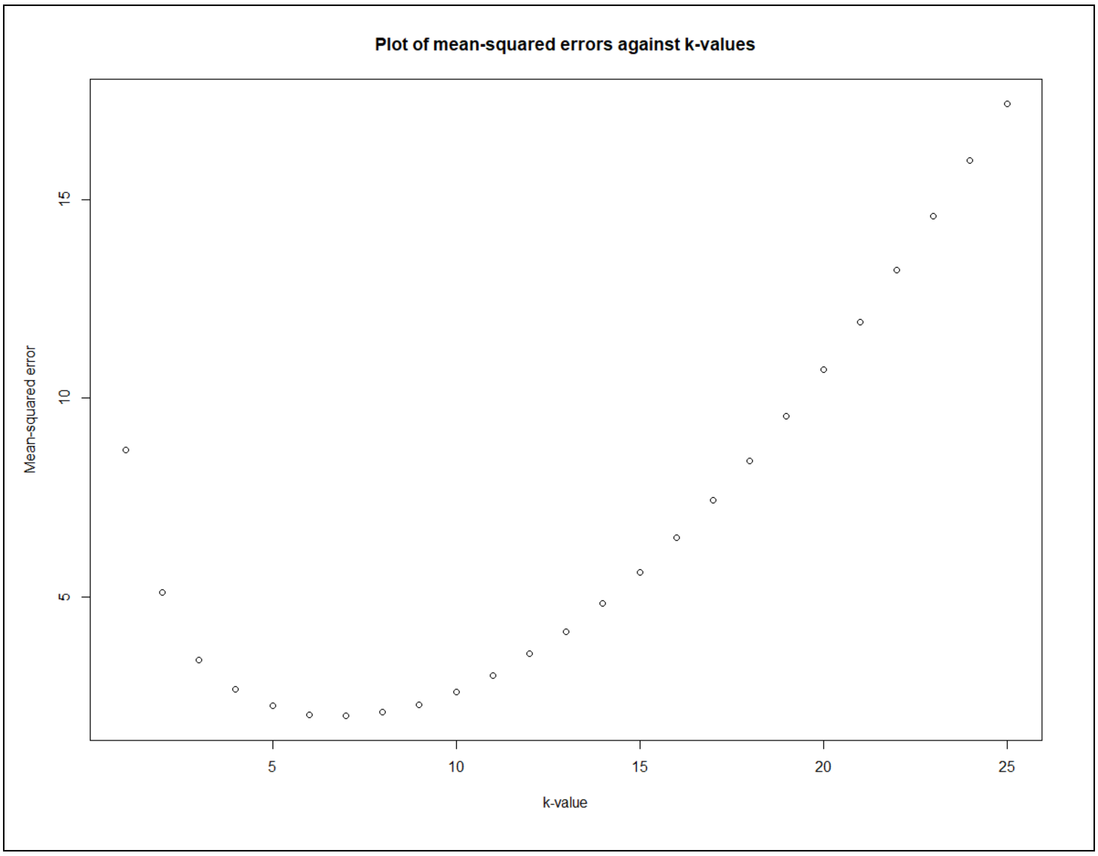
1)

(a) For each value of k = 1, 2, 3 … 25, the mean-squared errors between my estimates of the spectrum and the true values in ms.test$intensity:

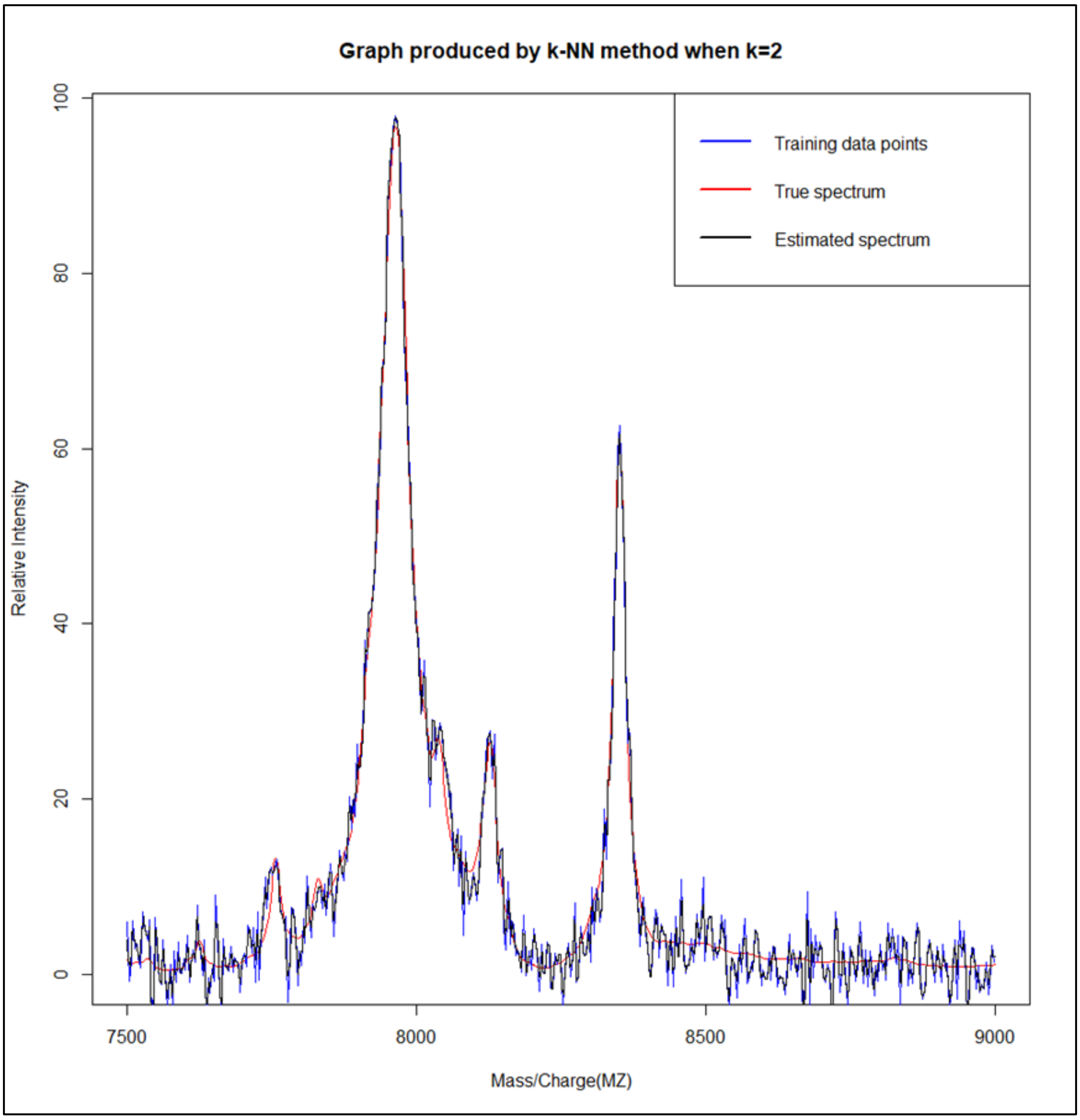


```
When k = 1 , mean-squared error between my estimates and the true values is 8.704256
When k = 2 , mean-squared error between my estimates and the true values is 5.104779
When k = 3 , mean-squared error between my estimates and the true values is 3.410489
When k = 4 , mean-squared error between my estimates and the true values is 2.656165
When k = 5 , mean-squared error between my estimates and the true values is 2.262812
When k = 6 , mean-squared error between my estimates and the true values is 2.021296
When k = 7 , mean-squared error between my estimates and the true values is 2.004127
When k = 8 , mean-squared error between my estimates and the true values is 2.08466
When k = 9 , mean-squared error between my estimates and the true values is 2.286621
When k = 10 , mean-squared error between my estimates and the true values is 2.608518
When k = 11 , mean-squared error between my estimates and the true values is 3.012139
When k = 12 , mean-squared error between my estimates and the true values is 3.553871
When k = 13 , mean-squared error between my estimates and the true values is 4.124015
When k = 14 , mean-squared error between my estimates and the true values is 4.838148
When k = 15 , mean-squared error between my estimates and the true values is 5.619558
When k = 16 , mean-squared error between my estimates and the true values is 6.482609
When k = 17 , mean-squared error between my estimates and the true values is 7.436011
When k = 18 , mean-squared error between my estimates and the true values is 8.422623
When k = 19 , mean-squared error between my estimates and the true values is 9.547819
When k = 20 , mean-squared error between my estimates and the true values is 10.73333
When k = 21 , mean-squared error between my estimates and the true values is 11.92768
When k = 22 , mean-squared error between my estimates and the true values is 13.23454
When k = 23 , mean-squared error between my estimates and the true values is 14.59713
When k = 24 , mean-squared error between my estimates and the true values is 15.98565
When k = 25 , mean-squared error between my estimates and the true values is 17.42086
```

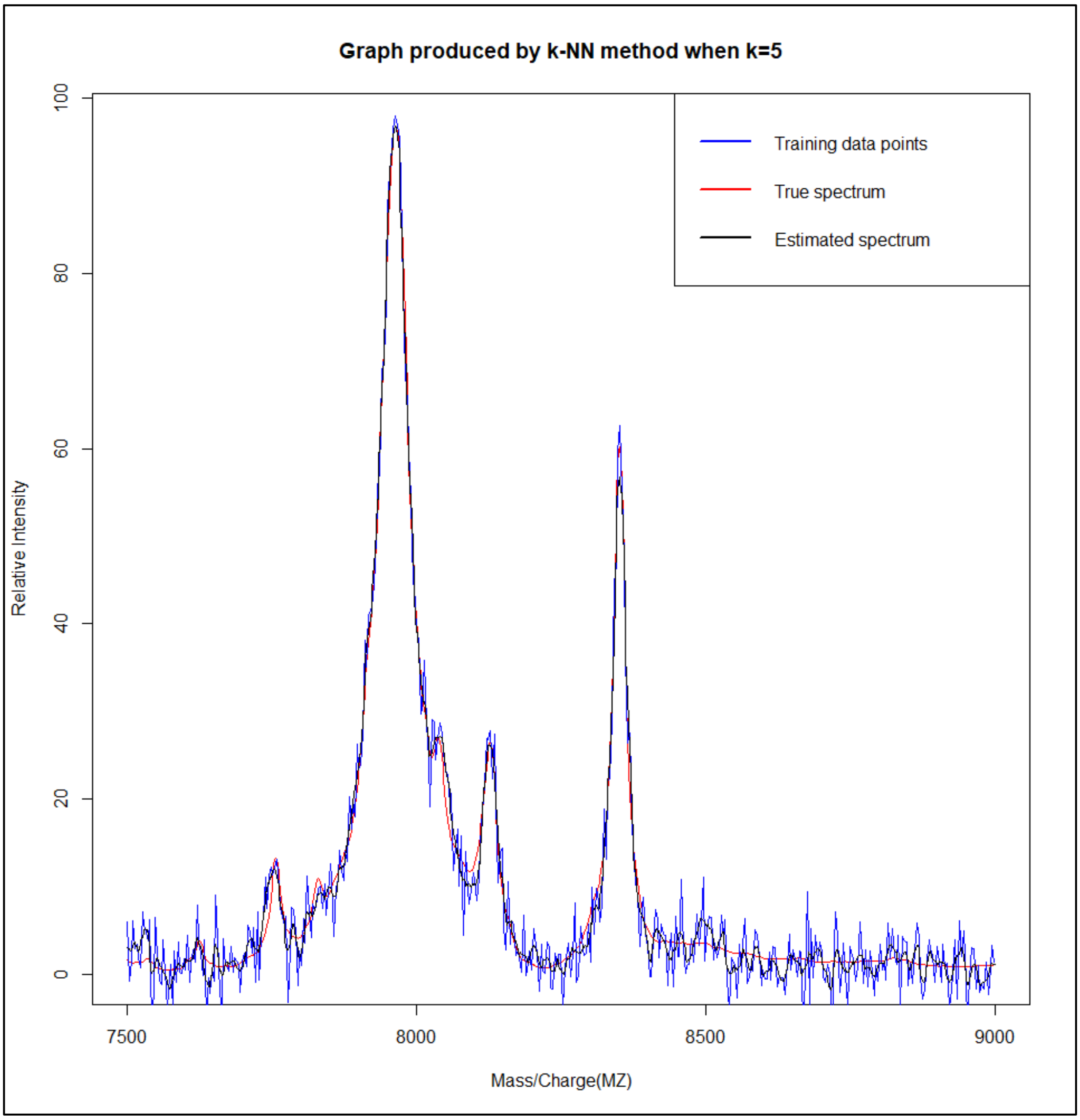
Plot of these errors against the various values of k:



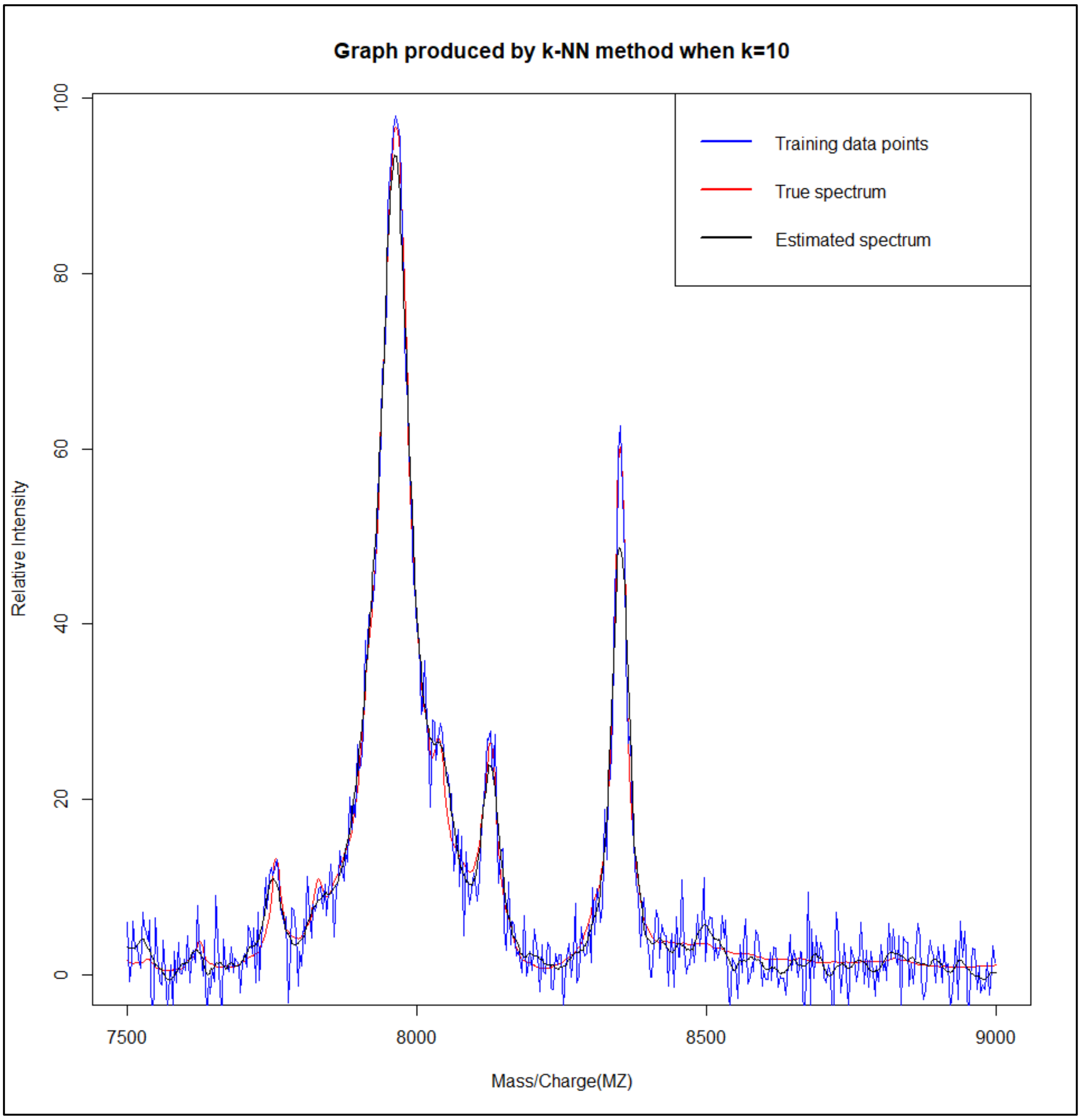
(b) Graph showing: the training data points (ms.train$intensity), the true spectrum (ms.test$intensity) and the estimated spectrum (predicted intensity values for the MZ values in ms.test.csv) produced by the k-NN method for k=2:



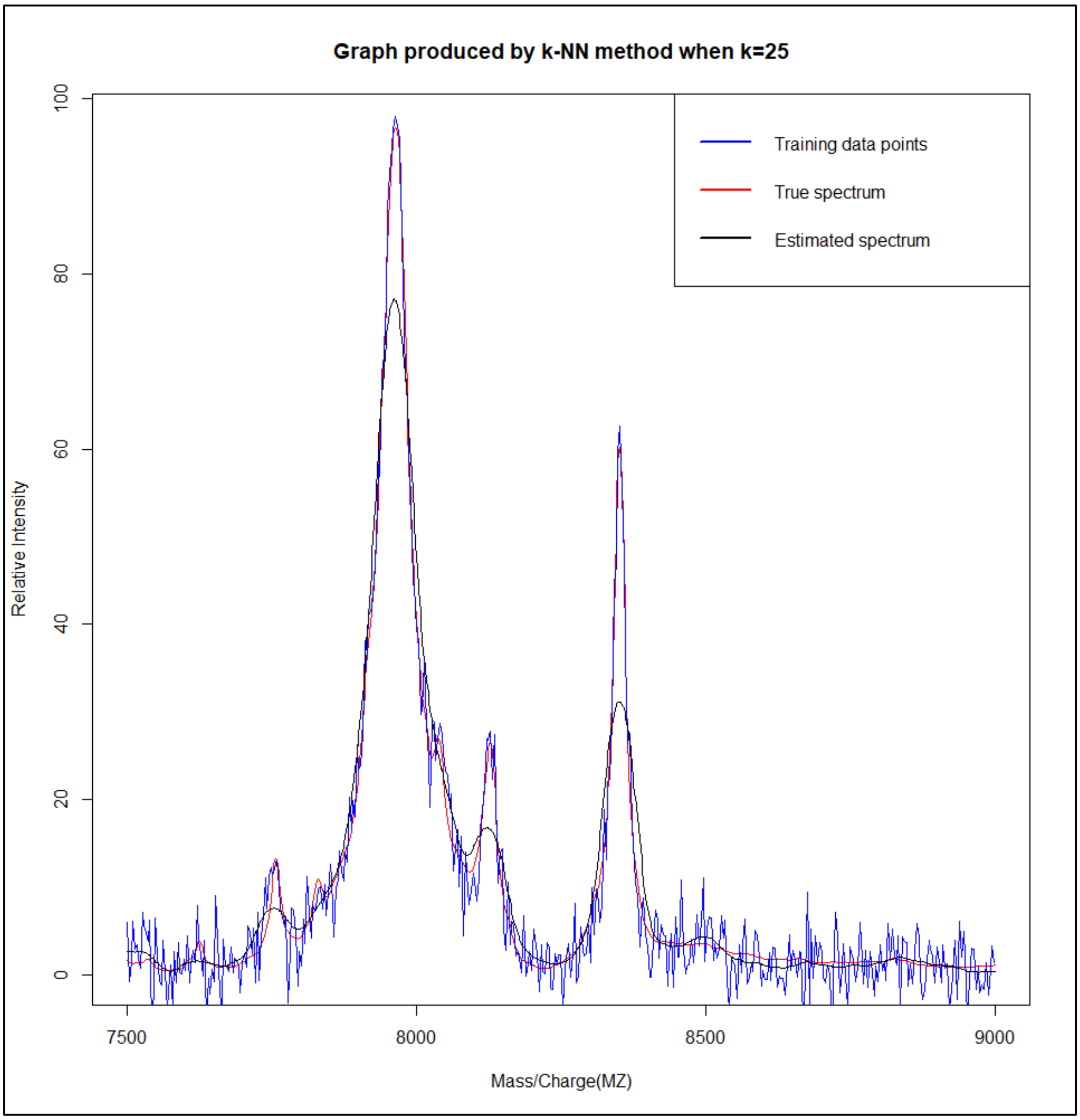
Graph showing: the training data points (ms.train$intensity), the true spectrum (ms.test$intensity) and the estimated spectrum (predicted intensity values for the MZ values in ms.test.csv) produced by the k-NN method for k=5:



Graph showing: the training data points (ms.train$intensity), the true spectrum (ms.test$intensity) and the estimated spectrum (predicted intensity values for the MZ values in ms.test.csv) produced by the k-NN method for k=10:



Graph showing: the training data points, the true spectrum and the estimated spectrum produced by the k-NN method for k=25:

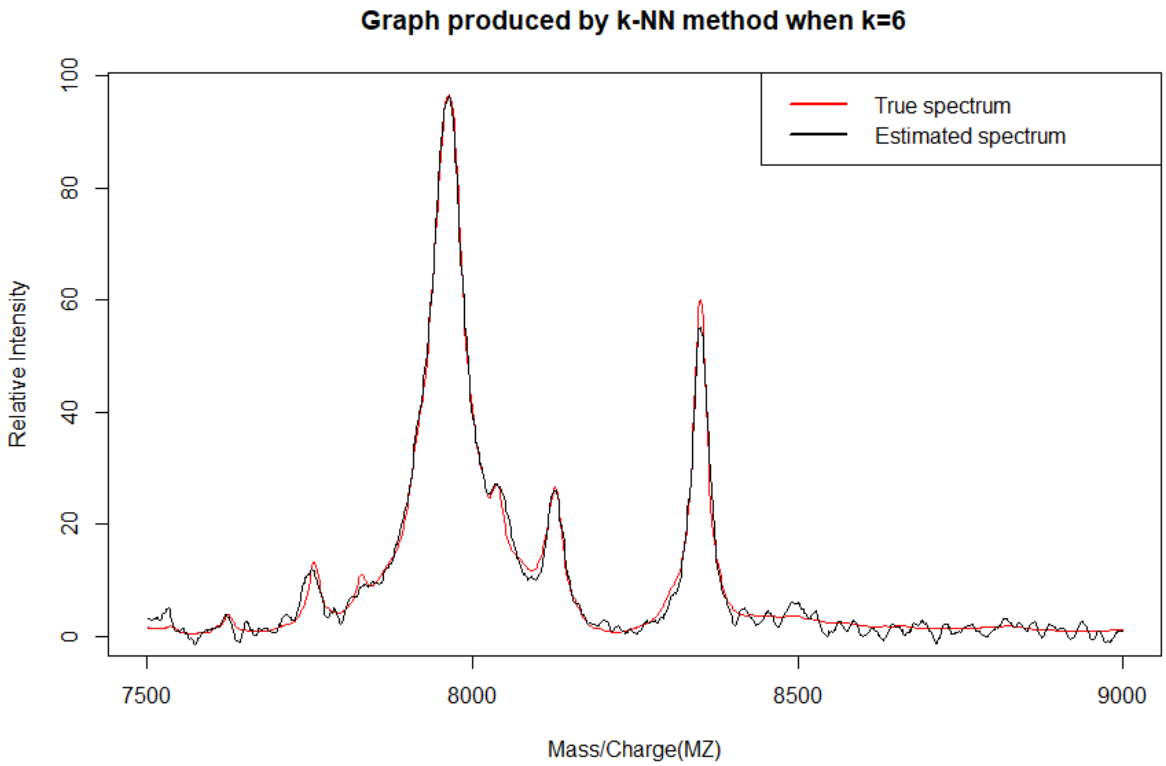


(c) By taking a look at Question 3.1a, we can see that the mean-squared errors on the true spectrum found when k=2, 5, 10 and 25. When k=5, the mean-squared error is at the least. Then followed by k=10, k=2 and k=25. Besides that, the plot of errors against the various values of k shows a shape of the curve where the minimal mean-squared errors occurred when k is around 5, 6, 7. When k=2, the coordinate of the point in the plot is higher than k=10 but lower than k=25. Therefore, we can conclude that among k=2, 5, 10 and 25, when k=5, the estimate is most accurate. When k=10, the estimate is better than k=2 and k=25. When k=2, the estimate is better than k=25. When k=25, the estimate is at worst.

2) The model selected k=6. But for Question3.1a, we have found out that the lowest mean-squared error occurred when k=7. The difference of the mean-squared error when k=6 and k=7 found in Question3.1a is: 2.021296 (k = 6) − 2.004127 (k = 7) = 0.017169.

3) Using the estimates of the curve produced in the previous question (Question 3.2), the estimate of the variance of the sensor/measurement noise that has corrupted our intensity measurements is 2.0196.

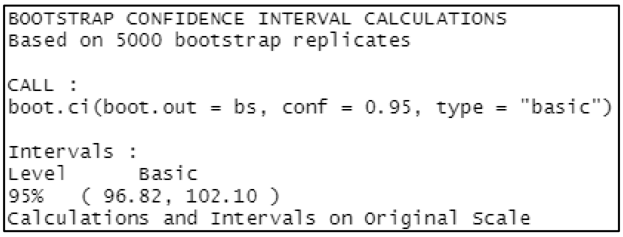
4) Yes. When k=6, the graph produced by k-NN method with true spectrum and estimated spectrum is shown as below. By comparing them, we can see that the black line provides a smooth, low-noise estimate of background level as well as accurate estimation of the peaks.



k-NN method is able to achieve this aim since we can use cross-validation to try and estimate the predictive performance of the k-NN algorithm with different choices for the parameters, and choose those values that lead to the best (estimated) prediction accuracy.

5) From the smoothed signal produced using the value of k found in Question 3.2 (k=6), the value of MZ corresponds to the maximum estimated abundance is 7963.3 mass/charge.

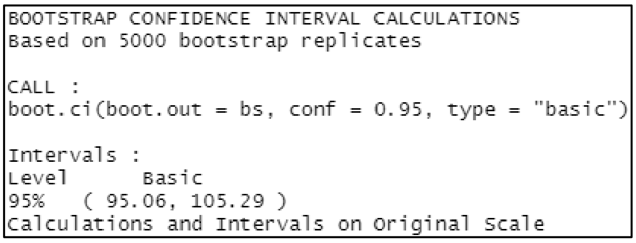
6) When k=3,



```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

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

Intervals :
Level      Basic
95%   ( 96.82, 102.10 )
Calculations and Intervals on Original Scale
```

95% confidence interval for the estimate of relative abundance at the MZ value is (96.82, 102.10)

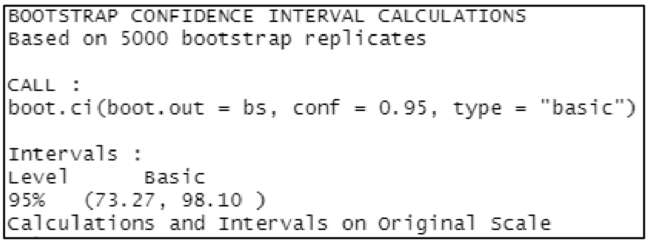
When k=6,



```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

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

Intervals :
Level      Basic
95%   ( 95.06, 105.29 )
Calculations and Intervals on Original Scale
```

95% confidence interval for the estimate of relative abundance at the MZ value is (95.06, 105.29)

When k=20,



```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

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

Intervals :
Level      Basic
95%   (73.27, 98.10 )
Calculations and Intervals on Original Scale
```

95% confidence interval for the estimate of relative abundance at the MZ value is (73.27, 98.10)

From the results above, we can see that the 95% confidence interval for the estimate of relative abundance at the MZ value is decreasing for increasing k-values. These confidence intervals vary in size for different values of k is because when the k-value is small, it has high-noise and is not smooth. Hence, the estimate is almost the same as the data. As the k-value increases, the noise is decreasing and the peak of MZ value (estimate of relative abundance at the MZ value) is also decreasing due to the smoothing.