FIT2086 Assignment3

Name: Lianzheng Xie
Student ID: 32068611

Question1 :

1. First, the mean function was used to calculate the average number of daily reported cases, and then we needed to obtain the variance and sample size.

```
> fuel = read.csv('fuel.ass3.2022.csv')
> mlr = lm(formula = Comb.FE~. ,data = fuel)
> summary(mlr)
```

```
Call:
lm(formula = Comb.FE ~ ., data = fuel)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7617 -1.0503 -0.0885  0.7359 11.3772 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -2.167e+02  1.550e+02  -1.398 0.162706    
Model.Year                1.156e-01  7.679e-02   1.505 0.132856    
Eng.Displacement         -1.331e+00  1.861e-01  -7.151 3.22e-12 ***
No.Cylinders              5.730e-03  1.206e-01   0.048 0.962117    
AspirationOT             -1.034e-01  1.240e+00  -0.083 0.933569    
AspirationSC             -7.990e-01  4.064e-01  -1.966 0.049842 *  
AspirationTC             -1.217e+00  2.201e-01  -5.528 5.31e-08 ***
AspirationTS             -1.351e+00  6.720e-01  -2.010 0.044935 *  
No.Gears                 -1.940e-01  5.158e-02  -3.760 0.000191 ***
Lockup.Torque.ConverterY -5.621e-01  1.974e-01  -2.847 0.004602 ** 
Drive.SysA                6.138e-02  2.706e-01   0.227 0.820624    
Drive.SysF                1.535e+00  2.930e-01   5.239 2.41e-07 ***
Drive.SysP               -9.766e-01  5.639e-01  -1.732 0.083967 .  
Drive.SysR                2.081e-01  2.551e-01   0.816 0.415071    
Max.Ethanol              -8.956e-03  6.100e-03  -1.468 0.142704    
Fuel.TypeGM               8.096e-01  1.004e+00   0.806 0.420647    
Fuel.TypeGP               4.064e-01  2.425e-01   1.676 0.094372 .  
Fuel.TypeGPR              8.418e-02  2.458e-01   0.343 0.732106    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.688 on 482 degrees of freedom
Multiple R-squared:  0.6571,	Adjusted R-squared:  0.645 
F-statistic: 54.34 on 17 and 482 DF,  p-value: < 2.2e-16
```

Above is the result of fitting a linear model, where the factors that may be related to fuel efficiency are marked with * on the right, Eng.Displacement, No.Gears, Drive.SysF, AspirationSC, AspirationTC, AspirationTS and Lockup.Torque.ConverterY. Because their p-value is less than α =0.05.

And the three variables appear to be strongest predictors of fuel efficiency are Eng.Displacement, AspirationTC and Drive.SysF. Because their p-values are much lower than 0.05

2. Let α =0.05 , if we do p different tests where p = 17 in this case because there are 17 predictors. Then if we consider Bonferroni procedure, we need p-value < α/p.

$$\frac{\alpha}{p} = \frac{0.05}{17} = 0.00294$$

So now we need variables with p values less than 0.00294, which is just Eng.Displacement, AspirationTC, No.Gears and Drive.SysF.

3. Engine displacement(Eng.Displacement) refers to the exhaust volume. When the engine is working, the higher the power of the air input gas, the greater the engine energy released per unit time, and the higher the fuel conversion efficiency. Drive.sysF's drivetrain and powertrain are compact and have a short Drive route, resulting in high transmission efficiency. It can reduce the parts of the transmission system, reduce the cost and the weight of the vehicle, which is conducive to the fuel economy of the vehicle. According to the estimate in Q1.1, E(Comb.FE) requires -1.316 Eng.Displacement units for each additional unit. For each additional unit, E(Comb.FE) requires +1.5535e Drive.sysF units.

4.

```
> BIC = step(object = mlr, durectuin = 'both',k = log(length(fuel$Comb.FE)))
> summary(BIC)
```

```
Call:
lm(formula = Comb.FE ~ Eng.Displacement + Aspiration + No.Gears +
    Lockup.Torque.Converter + Drive.Sys, data = fuel)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8612 -0.9840 -0.1003  0.7283 11.4706

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              16.36119    0.46567  35.134  < 2e-16 ***
Eng.Displacement         -1.31647    0.07658 -17.192  < 2e-16 ***
AspirationOT              0.13369    1.22670   0.109 0.913262
AspirationSC             -0.57062    0.38796  -1.471 0.141985
AspirationTC             -1.07175    0.18662  -5.743 1.64e-08 ***
AspirationTS             -1.32489    0.64147  -2.065 0.039414 *
No.Gears                 -0.17477    0.05068  -3.448 0.000613 ***
Lockup.Torque.ConverterY -0.57320    0.19365  -2.960 0.003227 **
Drive.SysA                0.19340    0.25739   0.751 0.452771
Drive.SysF                1.54754    0.28015   5.524 5.40e-08 ***
Drive.SysP               -1.08018    0.55259  -1.955 0.051182 .
Drive.SysR                0.28424    0.25126   1.131 0.258518
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.693 on 488 degrees of freedom
Multiple R-squared:  0.6507,    Adjusted R-squared:  0.6428
F-statistic: 82.64 on 11 and 488 DF,  p-value: < 2.2e-16
```

E[Comb.FE] = 16.36119-1.316*Eng.Displacement+ 0.1336877*AspirationOT -0.5706151* AspirationSC -1.0717452* AspirationTC -1.3248883* AspirationTS -0.1747731* No.Gears -0.5731985* Lockup.Torque.ConverterY+ 0.1934040* Drive.SysA+ 1.5475411* Drive.SysF -1.0801801* Drive.SysP+ 0.2842354* Drive.SysR

5. a)Since the BIS model is based on a database with limited data, the population variance is unknown, so sample variance is needed. Use the formula below

$$\left(\hat{\mu} - t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}},\ \hat{\mu} + t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

```
result = fuel[33,]
predict(mlr,result,interval = 'confidence')
```

```
> predict(mlr,result,interval = 'confidence')
        fit      lwr     upr
33 13.37941 12.93791 13.8209
```

(12.93791, 13.8209)

b) The new car has a mean fuel efficiency is 13.37941,and the 95%CI range >11, so new car will have better fuel efficiency than current car. However, this data is based on the model prediction and cannot fully represent the real situation in reality. New cars are more fuel efficient.

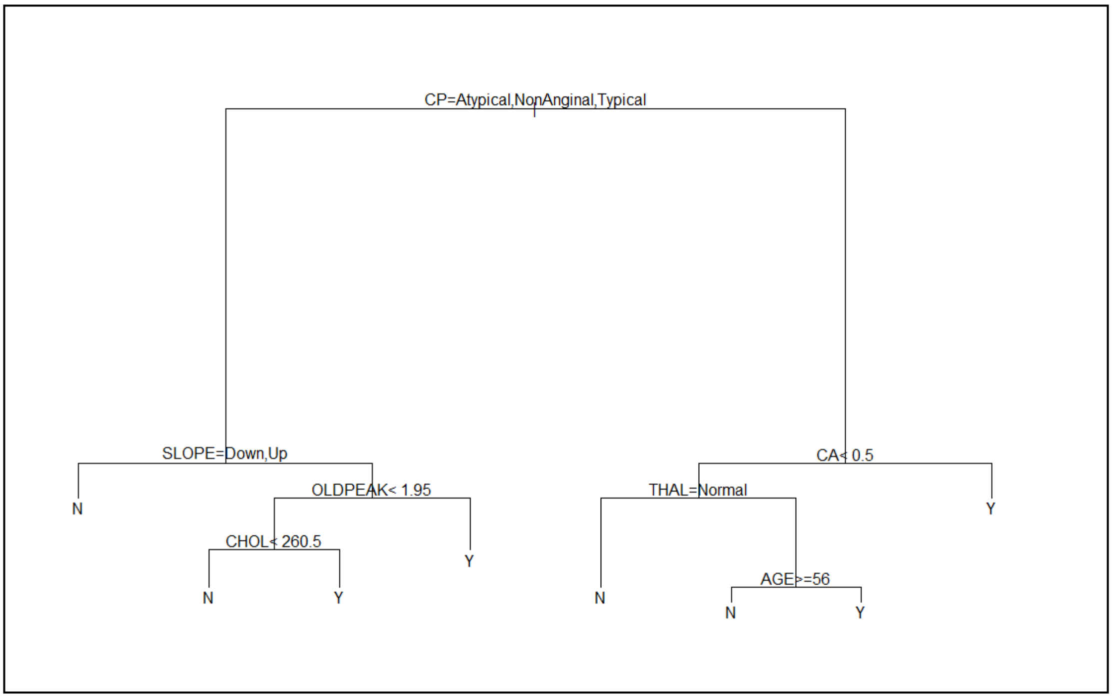
Question 2:

1. Using learn.tree.cv function get the result:

```
n= 210

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

 1) root 210 99 N (0.52857143 0.47142857)
   2) CP=Atypical,NonAnginal,Typical 106 23 N (0.78301887 0.21698113)
     4) SLOPE=Down,Up 69  6 N (0.91304348 0.08695652) *
     5) SLOPE=Flat 37 17 N (0.54054054 0.45945946)
      10) OLDPEAK< 1.95 30 10 N (0.66666667 0.33333333)
        20) CHOL< 260.5 21  3 N (0.85714286 0.14285714) *
        21) CHOL>=260.5 9  2 Y (0.22222222 0.77777778) *
      11) OLDPEAK>=1.95 7  0 Y (0.00000000 1.00000000) *
   3) CP=Asymptomatic 104 28 Y (0.26923077 0.73076923)
     6) CA< 0.5 47 23 N (0.51063830 0.48936170)
      12) THAL=Normal 21  4 N (0.80952381 0.19047619) *
      13) THAL=Fixed.Defect,Reversible.Defect 26  7 Y (0.26923077 0.73076923)
        26) AGE>=56 8  3 N (0.62500000 0.37500000) *
        27) AGE< 56 18  2 Y (0.11111111 0.88888889) *
     7) CA>=0.5 57  4 Y (0.07017544 0.92982456) *
```

According to the result produced above, variables that have been used in the best tree include CP, SLOPE, OLDPEAK, CHOL, CA, THAL, AGE. And the best tree has 8 leaves(terminal nodes).

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



If the Chest pain type(CP) is Atypical, NonAnginal or Typical, the Slope of the peak exercise ST segment(SLOPE) is Down or Up, the patient will probably have no heart disease.

If the Chest pain type(CP) is Atypical, NonAnginal, Typical, the Slope of the peak exercise ST segment(SLOPE) is Flat, Exercise induced ST depression relative to rest (OLDPEAK)< 1.95, Serum cholesterol in mg/dl (CHOL)< 260.5 ,the patient will probably have no heart disease.

If the Chest pain type(CP) is Atypical, NonAnginal, Typical, the Slope of the peak exercise ST segment(SLOPE) is Flat, Exercise induced ST depression relative to rest (OLDPEAK)< 1.95, Serum cholesterol in mg/dl (CHOL) >=260.5, the patient will probably have heart disease.

If the Chest pain type(CP) is Atypical, NonAnginal, Typical, the Slope of the peak exercise ST segment(SLOPE) is Flat, Exercise induced ST depression relative to rest (OLDPEAK) >=1.95, the patient will probably have heart disease.

If the Chest pain type(CP) is Asymptomatic, Number of major vessels colored by flourosopy (CA)< 0.5, Thallium scanning results (THAL) is Normal, the patient will probably have no heart disease.

If the Chest pain type(CP) is Asymptomatic, Number of major vessels colored by flourosopy (CA)< 0.5, Thallium scanning results (THAL) is Fixed.Defect, Reversible.Defect, Age of patient in years (AGE)>=56, the patient will probably have no heart disease.

If the Chest pain type(CP) is Asymptomatic, Number of major vessels colored by flourosopy

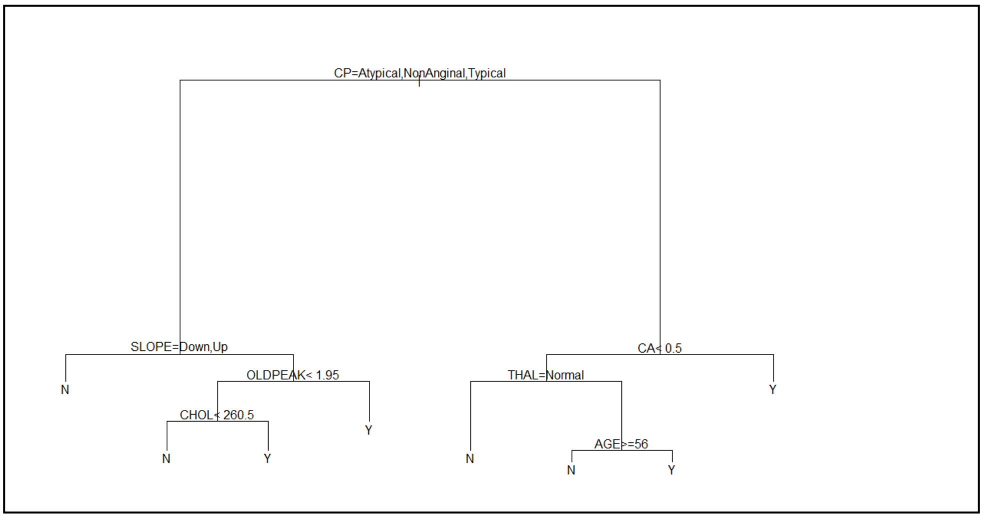
(CA)< 0.5, Thallium scanning results (THAL) is Fixed.Defect, Reversible.Defect, Age of patient in years (AGE) < 56, the patient will probably have no heart disease.

If the Chest pain type(CP) is Asymptomatic, Number of major vessels colored by flourosopy (CA)> 0.5, Thallium scanning results (THAL) is Normal, the patient will probably have no heart disease.

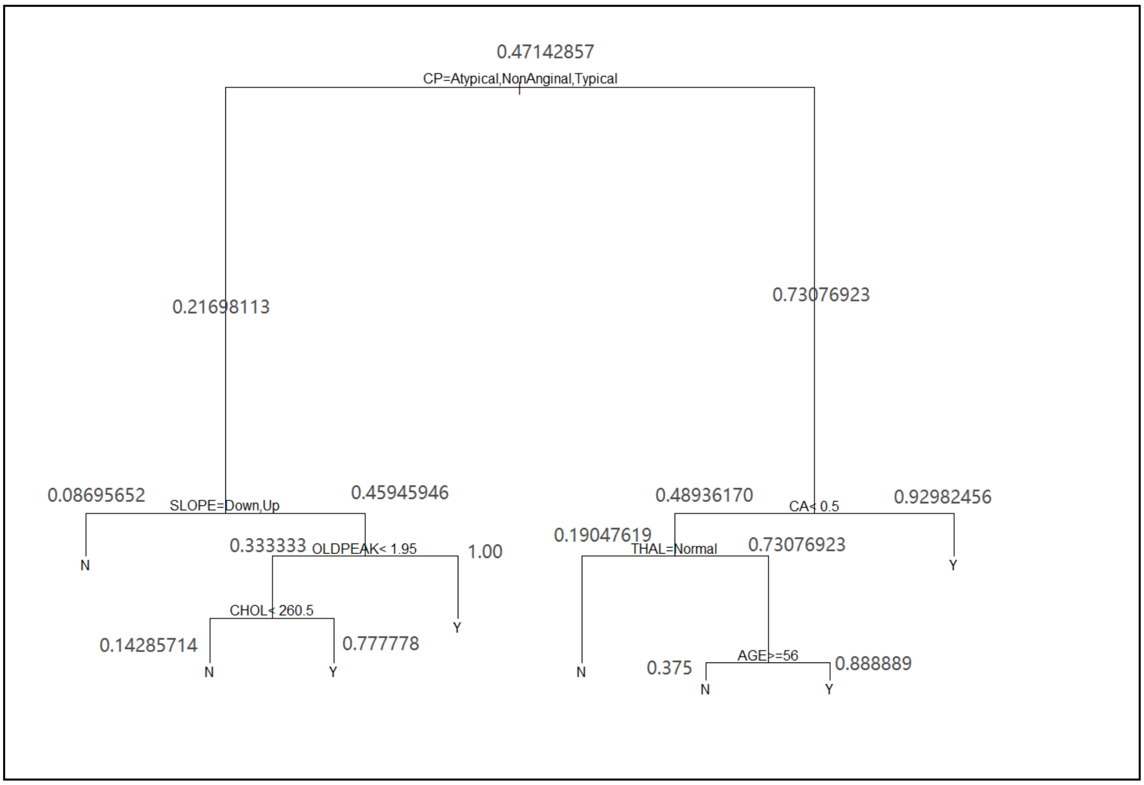
3. Textual representation of the tree

```
n= 210

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

 1) root 210 99 N (0.52857143 0.47142857)
   2) CP=Atypical,NonAnginal,Typical 106 23 N (0.78301887 0.21698113)
     4) SLOPE=Down,Up 69  6 N (0.91304348 0.08695652) *
     5) SLOPE=Flat 37 17 N (0.54054054 0.45945946)
      10) OLDPEAK< 1.95 30 10 N (0.66666667 0.33333333)
        20) CHOL< 260.5 21  3 N (0.85714286 0.14285714) *
        21) CHOL>=260.5 9  2 Y (0.22222222 0.77777778) *
      11) OLDPEAK>=1.95 7  0 Y (0.00000000 1.00000000) *
   3) CP=Asymptomatic 104 28 Y (0.26923077 0.73076923)
     6) CA< 0.5 47 23 N (0.51063830 0.48936170)
      12) THAL=Normal 21  4 N (0.80952381 0.19047619) *
      13) THAL=Fixed.Defect,Reversible.Defect 26  7 Y (0.26923077 0.73076923)
        26) AGE>=56 8  3 N (0.62500000 0.37500000) *
        27) AGE< 56 18  2 Y (0.11111111 0.88888889) *
     7) CA>=0.5 57  4 Y (0.07017544 0.92982456) *
```

The plot of the tree:



Annotated plot of tree:



4. According to tree's prediction, patients whose chest pain type(CP) is Atypical, NonAnginal or Typical, the Slope of the peak exercise ST segment(SLOPE) is Down or UP, their likelihood of heart disease is the lowest(0.08695652).

5. Kic summary:

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4071  -0.4679  -0.1246   0.4014   2.6684

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.248413   2.232140  -3.247 0.001165 **
SEXM          1.802856   0.533646   3.378 0.000729 ***
CPAtypical   -2.184817   0.703118  -3.107 0.001888 **
CPNonAnginal -2.599144   0.558932  -4.650 3.32e-06 ***
CPTypical    -2.369844   0.753460  -3.145 0.001659 **
TRESTBPS      0.021501   0.011787   1.824 0.068140 .
CHOL          0.008167   0.004200   1.944 0.051854 .
OLDPEAK       0.581819   0.260840   2.231 0.025710 *
SLOPEFlat     1.931508   0.994042   1.943 0.052006 .
SLOPEUp       0.206602   1.086994   0.190 0.849257
CA            1.074811   0.285071   3.770 0.000163 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 290.44  on 209  degrees of freedom
Residual deviance: 143.52  on 199  degrees of freedom
AIC: 165.52
```

Final model contains variable is SEXM, CP, TRESTBPS, CHOL, OLDPEAK, SLOPE, CA. It has additional variables, such as SEXM and TRESTBPS, compared to the variables used by the CV estimated tree. It has no variables such as THAL and AGE except for additional variables. CPNonAnginal was the most important variable in logistic regression. We can do that through Pr. |Z|) value to determine this, the smaller the value, it is more important.

6. Logistic regression equation:

$$P(Y_i = 1 | x_{i,1}, \ldots, x_{i,p}) = \frac{1}{1 + e^{-\eta}}$$

=1/(1+exp(-(-7.248412835+1.802856259*SEXM-2.184816631*CPAtypical-2.599143599*CPNonAnginal-2.369843661*CPTypical+0.021501131*TRESTBPS+0.008166783*CHOL+0.581819245*OLDPEAK+1.931508302*SLOPEFlat+0.206601760*SLOPEUp+1.074811459*CA)))

7. Because the range of the logistic regression equation is between 0 and 1, with CA as one of the variables, the higher the CA, the closer the P is to 1. This means that the higher the CA, the higher the risk of heart disease, and the lower the CA, the lower the risk of heart disease.

8.

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

Confusion matrix:

    target
pred  N  Y
   N 47 14
   Y  6 25

Classification accuracy = 0.7826087
Sensitivity             = 0.6410256
Specificity             = 0.8867925
Area-under-curve        = 0.8214804
Logarithmic loss        = 87.37257

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

Confusion matrix:

    target
pred  N  Y
   N 45  8
   Y  8 31

Classification accuracy = 0.826087
Sensitivity             = 0.7948718
Specificity             = 0.8490566
Area-under-curve        = 0.8853411
Logarithmic loss        = 39.43705

---------------------------------------------------------------------------
```

By comparison, the Classification accuracy(0.826087>0.7826087), Sensitivity(0.7948718>0.6410256) and Area-under-curve(0.8853411>0.8214804) predicted by KIC model are more accurate than those predicted by cv tree. Specificity was less than the prediction of cv tree. Logarithmic loss less.

So, I think KIC's prediction model will be better.

9. a) the tree model found using cross-validation

```
> heart.test[10,]
   AGE SEX           CP TRESTBPS CHOL  FBS RESTECG THALACH EXANG OLDPEAK SLOPE CA              THAL HD
10  51   M Asymptomatic      140  298 <120  Normal     122     Y     4.2  Flat  3 Reversible.Defect  Y
```

Probability having a heart disease of the combination above = 0.92982456

Probability not having a heart disease of the combination above = 1-0.92982456=0.07017544

Odds of having heart disease for the patient in row 10th = $\frac{0.92982456}{0.07017544} = 13.25$

b) the step-wise logistic regression model

```
> prob.tree = predict(kic, heart.test)
> odds = exp(prob.tree[10])
> prob.tree[10]
      10
7.597887
> odds
      10
1993.977
```

odds = 1993.977

10. The confidence interval for 65th patients:

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

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

Intervals :
Level       BCa
95%   ( 0.7950,  0.9922 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
```

The confidence interval for the odds of having heart disease for the 65th patients in the test data is (0.7950,0.9922)

The confidence interval for 66th patients:

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

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

Intervals :
Level       BCa
95%   ( 0.0534,  0.4452 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
```

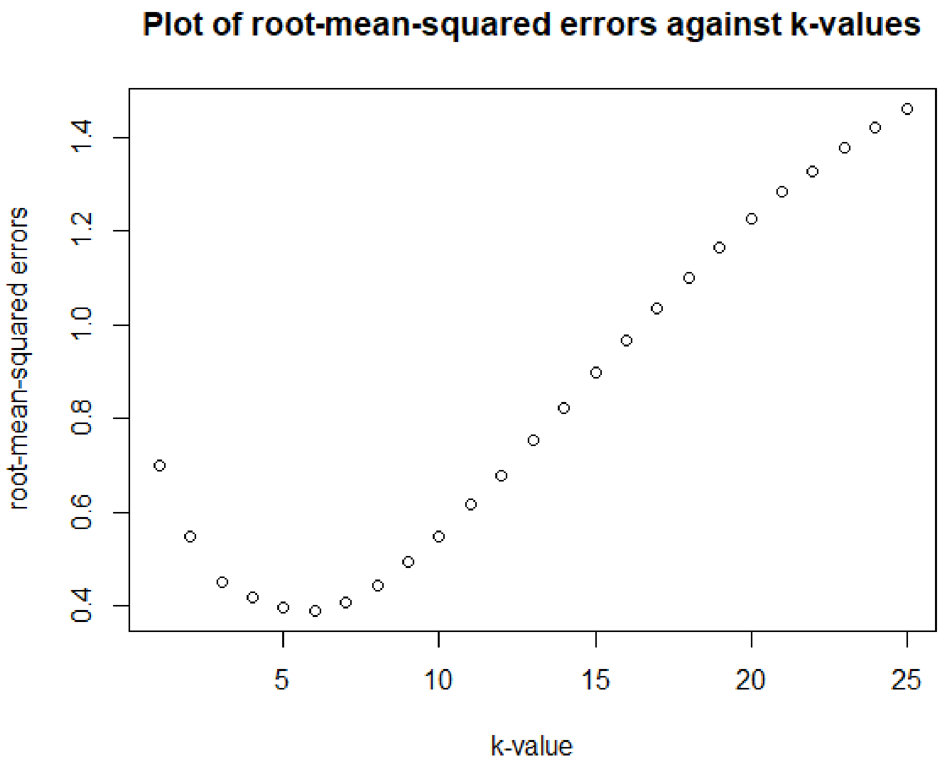
The confidence interval for the odds of having heart disease for the 65th and 66th patients in the test data is (0.0534,0.4452)

Question 3:

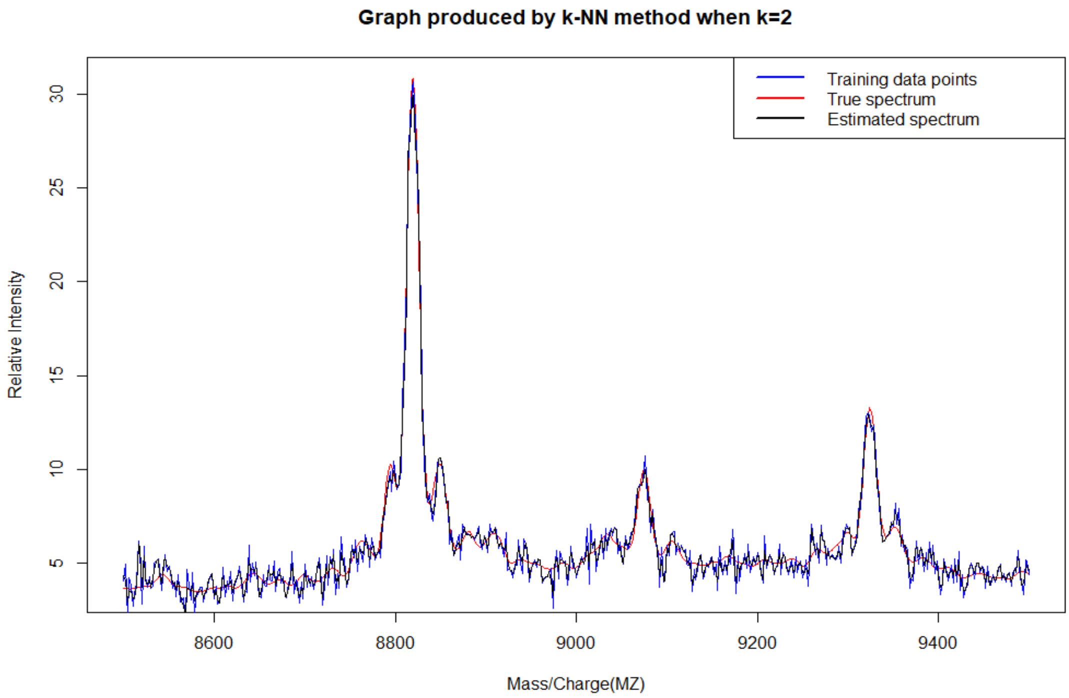
1. For each value of k = 1, 2, 3 ⋯ 25, the root-mean-squared errors between my estimates of the spectrum and the true values in ms.truth.2022$intensity:

```
when k = 1 , root-mean-squared error between my estimates and the true values is 0.7014977
when k = 2 , root-mean-squared error between my estimates and the true values is 0.54835
when k = 3 , root-mean-squared error between my estimates and the true values is 0.4521169
when k = 4 , root-mean-squared error between my estimates and the true values is 0.4206698
when k = 5 , root-mean-squared error between my estimates and the true values is 0.3986738
when k = 6 , root-mean-squared error between my estimates and the true values is 0.3919784
when k = 7 , root-mean-squared error between my estimates and the true values is 0.4102923
when k = 8 , root-mean-squared error between my estimates and the true values is 0.4465305
when k = 9 , root-mean-squared error between my estimates and the true values is 0.4967026
when k = 10 , root-mean-squared error between my estimates and the true values is 0.5501451
when k = 11 , root-mean-squared error between my estimates and the true values is 0.6164071
when k = 12 , root-mean-squared error between my estimates and the true values is 0.6786136
when k = 13 , root-mean-squared error between my estimates and the true values is 0.7529644
when k = 14 , root-mean-squared error between my estimates and the true values is 0.8243205
when k = 15 , root-mean-squared error between my estimates and the true values is 0.8981132
when k = 16 , root-mean-squared error between my estimates and the true values is 0.9686418
when k = 17 , root-mean-squared error between my estimates and the true values is 1.036175
when k = 18 , root-mean-squared error between my estimates and the true values is 1.10146
when k = 19 , root-mean-squared error between my estimates and the true values is 1.16347
when k = 20 , root-mean-squared error between my estimates and the true values is 1.225863
when k = 21 , root-mean-squared error between my estimates and the true values is 1.28212
when k = 22 , root-mean-squared error between my estimates and the true values is 1.327633
when k = 23 , root-mean-squared error between my estimates and the true values is 1.376192
when k = 24 , root-mean-squared error between my estimates and the true values is 1.421086
when k = 25 , root-mean-squared error between my estimates and the true values is 1.460003
```

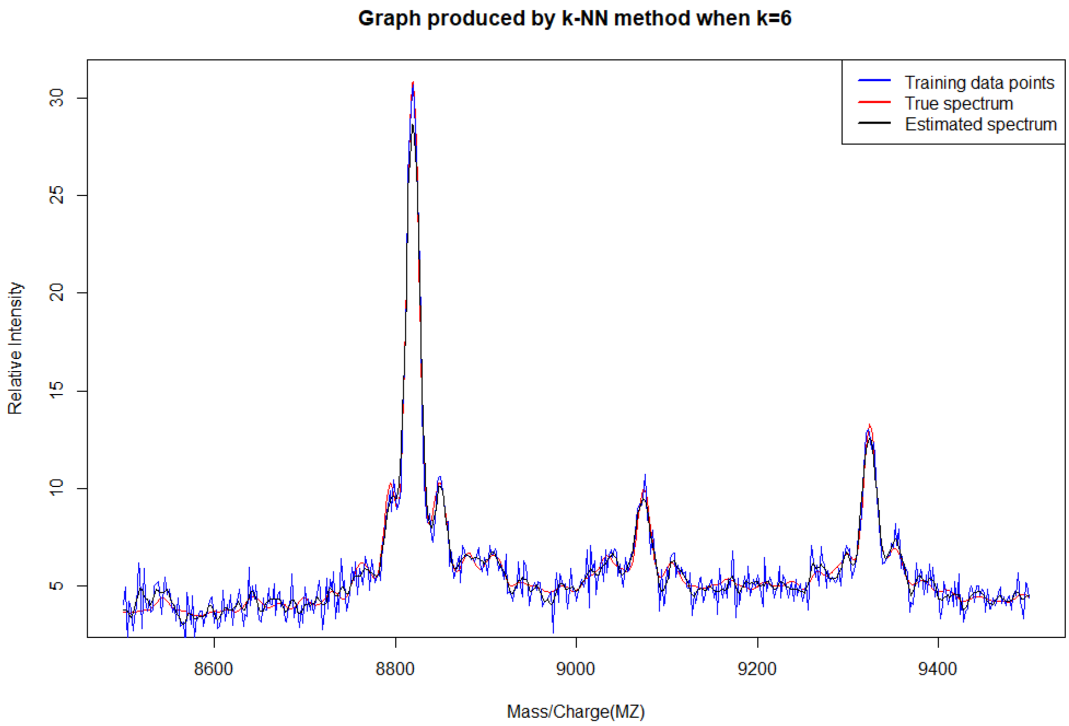
Plot of these errors against the various values of k:



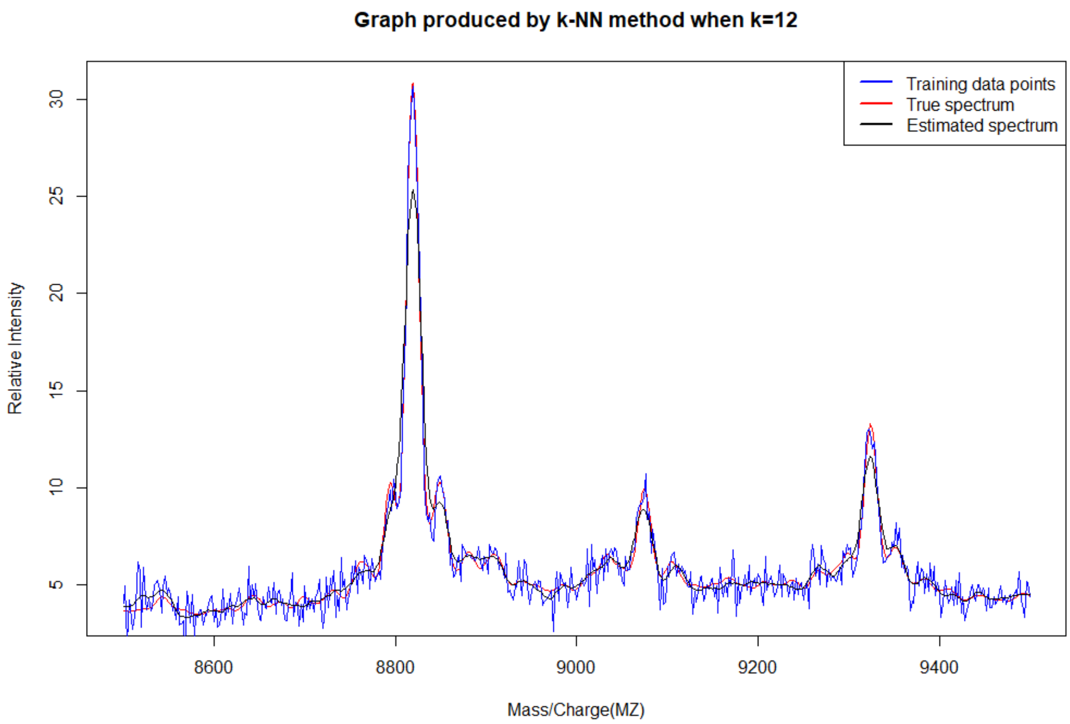
2. Graph show: The training data points (ms.measured.2022$intensity), the true spectrum (ms.truth.2022$intensity) and the estimated spectrum (predicted intensity values for the MZ values in ms.truth.2022.csv) produced by the k-NN method for k=2:



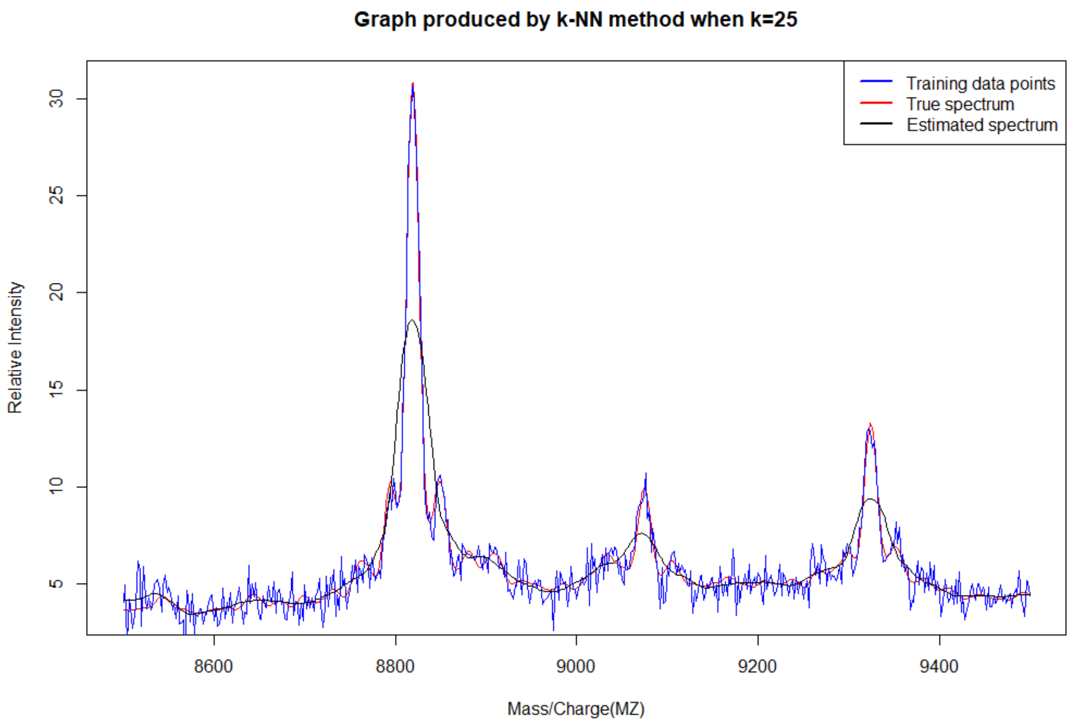
Graph show: The training data points (ms.measured.2022$intensity), the true spectrum (ms.truth.2022$intensity) and the estimated spectrum (predicted intensity values for the MZ values in ms.truth.2022.csv) produced by the k-NN method for k=6:



Graph show: The training data points (ms.measured.2022$intensity), the true spectrum (ms.truth.2022$intensity) and the estimated spectrum (predicted intensity values for the MZ values in ms.truth.2022.csv) produced by the k-NN method for k=12:

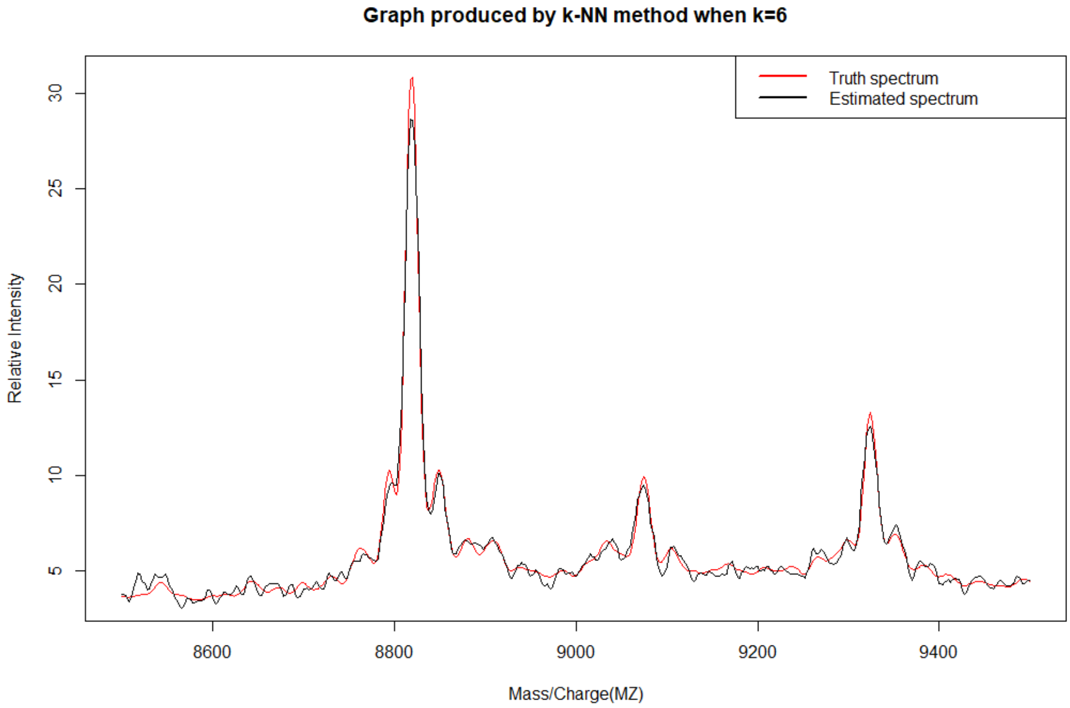


Graph show: The training data points (ms.measured.2022$intensity), the true spectrum (ms.truth.2022$intensity) and the estimated spectrum (predicted intensity values for the MZ values in ms.truth.2022.csv) produced by the k-NN method for k= 25:



3. By looking at Question 3.1, we can see that the root-mean-squared error on the truth spectrum for k= 2,6,12 and 25. When k=6, the root-mean-squared error is the smallest. In addition, the error plots for different values of k show a curvilinear shape with minimum root-mean-squared error when k is around 5,6,7. When k=2, the coordinates of this point in the graph are greater than k=10 and less than k=25. Therefore, we can conclude that the estimation is most accurate when k=6 in k=2,6,12 and 25. When k=12, the estimate is better than k=2 and k=25. When k=2, the estimate is better than k=25. When k is equal to 25, it's probably the worst.

4. Yes. When k=6, the figure with truth spectrum and estimated spectrum obtained by K-NN method is shown in the following figure. By comparison, we can see that the black line provides a smooth, low-noise estimate of the background level and an accurate peak estimate.



The k-NN method is able to achieve this aim because we can use cross-validation to try and estimate the prediction performance of the k-NN algorithm under different parameter choices and select those values that lead to the best (estimated) prediction accuracy.

5. The model selected k=5. But for Question3.1, we have found out that the minimise the actual mean-squared error when k=6. The difference of the mean-squared error when k=5 and k=6 found in Question3.1 is: 0.3986738 (k = 5) - 0.3919784 (k = 6) = 0.017169.

6. Using the estimate of the spectrum produced in Q3.5, the estimate of the standard deviation of the sensor/measurement noise that has corrupted our intensity measurements is 0.3986602.

7. From the smoothed signal produced using the value of k found in Question 3.5 (k=6), the value of MZ corresponds to the maximum estimated intensity is 8818 mass/charge.

8. When k = 3:

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

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

Intervals :
Level       BCa
95%   (26.61, 30.66 )
Calculations and Intervals on Original Scale
```

95% confidence interval for the estimate of the intensity at the MZ value is (26.61, 30.66)

When k = 5

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

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

Intervals :
Level       BCa
95%   (26.34, 30.66 )
Calculations and Intervals on Original Scale
```

95% confidence interval for the estimate of the intensity at the MZ value is (24.16, 30.44)

When k = 20

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

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

Intervals :
Level       BCa
95%   (15.29, 26.64 )
Calculations and Intervals on Original Scale
```

95% confidence interval for the estimate of the intensity at the MZ value is (15.29, 26.64)

As can be seen from the above results, the 95% confidence interval of the intensity estimate at the MZ value decreases with the increase of k value. The magnitude of these confidence intervals is different for different values of k, because when k is small, it has high noise and is not smooth. Therefore, the estimates are almost consistent with the data. As k increases, the noise decreases, and the peak value of MZ (the estimated value of intensity at MZ) also decreases due to smoothing