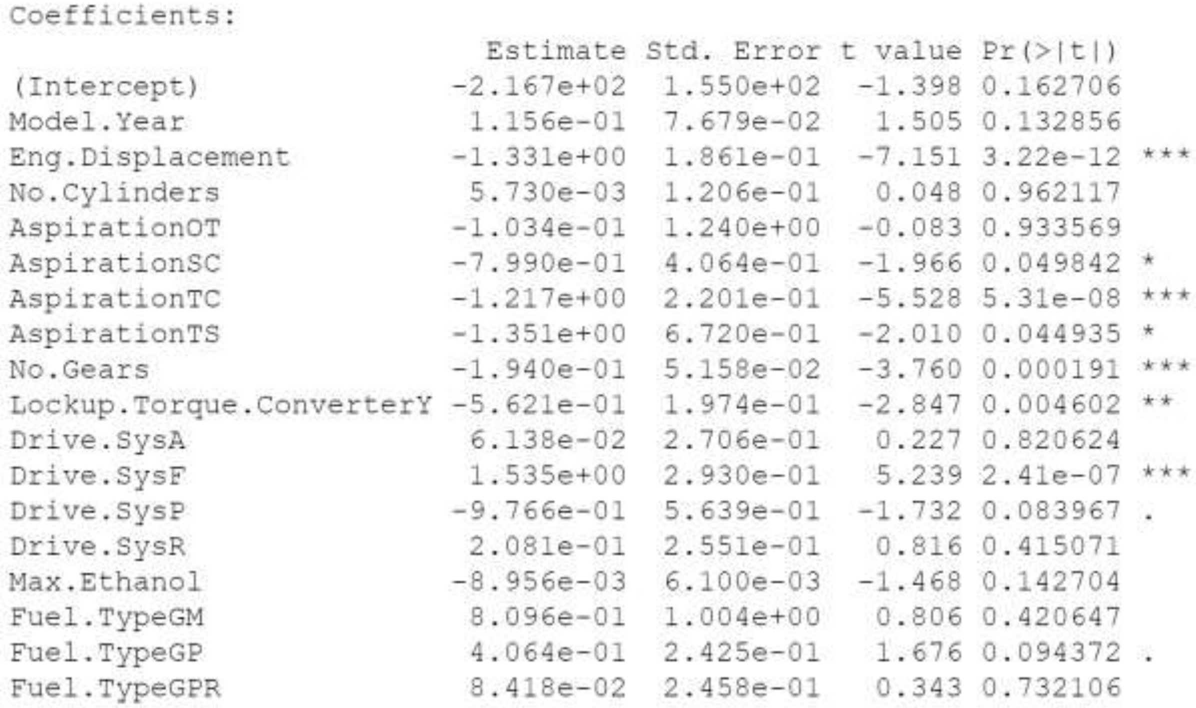
Question 1

1. By the results of the summary function:

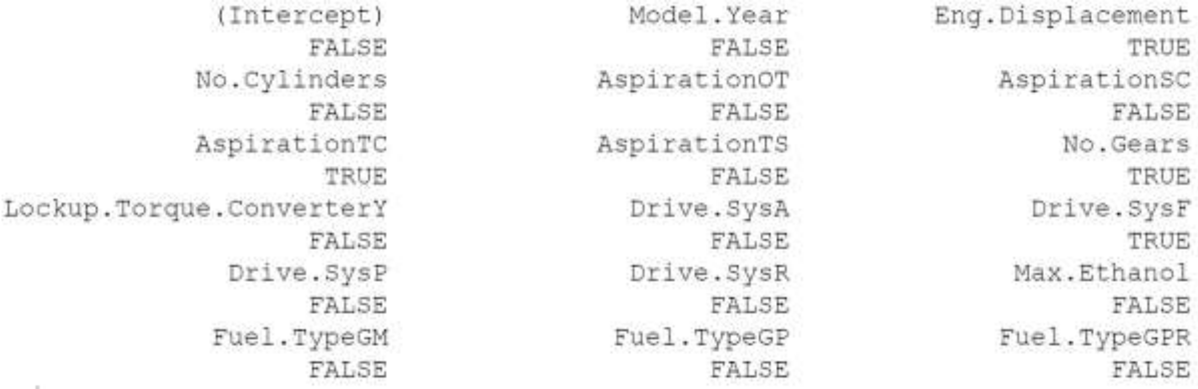


```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -2.167e+02  1.550e+02  -1.398 0.162706    
Model.Year                1.156e-01  7.679e-02   1.505 0.132856    
Eng.Displacement         -1.331e+00  1.861e-01  -7.151 3.22e-12 ***
No.Cylinders              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    
```

The predictors that are possibly associated with fuel efficiency are Eng.Displacement, AspirationSC, AspirationTC, AspirationTS, No.Gears, Lockup.Torque.ConverterY, Drive.SysF, Drive.SysP and Fuel.TypeGP.

The three variables Eng.Displacement, AspirationTC and Drive.SysF appear to be the strongest predictors of fuel efficiency as they have the smallest p-values out of all the predictors.

2. From the result of running the Bonferroni procedure with α = 0.05:



```
             (Intercept)               Model.Year         Eng.Displacement 
                   FALSE                    FALSE                     TRUE 
            No.Cylinders             AspirationOT             AspirationSC 
                   FALSE                    FALSE                    FALSE 
            AspirationTC             AspirationTS                 No.Gears 
                    TRUE                    FALSE                     TRUE 
Lockup.Torque.ConverterY               Drive.SysA               Drive.SysF 
                   FALSE                    FALSE                     TRUE 
              Drive.SysP               Drive.SysR              Max.Ethanol 
                   FALSE                    FALSE                    FALSE 
             Fuel.TypeGM              Fuel.TypeGP             Fuel.TypeGPR 
                   FALSE                    FALSE                    FALSE 
```

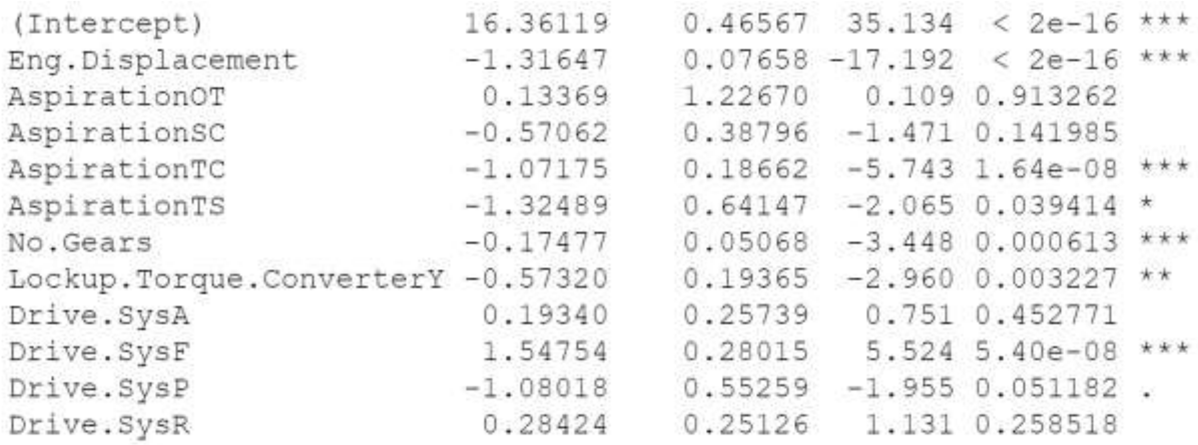
The predictors that are possibly associated with fuel efficiency are Eng.Displacement, AspirationTC, No.Gears, and Drive.SysF. The predictors AspirationSC, AspirationTS, Lockup.Torque.ConverterY, Drive.SysP and Fuel.TypeGP are eliminated from the procedure compare to the previous model.

3. From the model summary:

The coefficient of Eng.Displacement is -1.330991, for each unit increase of Eng.Displacement, the mean fuel efficiency decreases by -1.330991km/l, which means it has a negative effect on the mean fuel efficiency of a car.

The coefficient of Drive.SysF is 1.535284, for each unit increase of Drive.SysF, the mean fuel efficiency increases by 1.535284km/l, which means it has a positive effect on the mean fuel efficiency of a car.

4. After running the stepwise selection procedure, the summary of the fitted model is now:

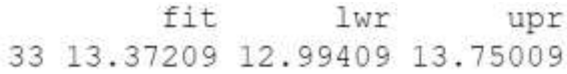


```
(Intercept)                16.36119    0.46567  35.134  < 2e-16 ***
Eng.Displacement           -1.31647    0.07658 -17.192  < 2e-16 ***
AspirationOT                0.13369    1.22670   0.109 0.913262
AspirationSC               -0.57062    0.38796  -1.471 0.141985
AspirationTC               -1.07175    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
```

The final regression equation obtained after pruning would be:

Log-odds(Comb.FE) = 16.36119 – 1.31647(Eng.Displacement) + 0.13369(AspirationOT)
– 0.57062(AspirationSC) – 1.07175(AspirationTC) – 1.32489(AspirationTS)
- 0.17477(No.Gears) – 0.57320(Lockup.Torque.ConverterY)
+ 0.19340(Drive.SysA) + 1.54754(Drive.SysF) – 1.08018(Drive.SysP)
+ 0.28424(Drive.SysR)

5. (a) From the results of running the predict() function for the thirty-third row:



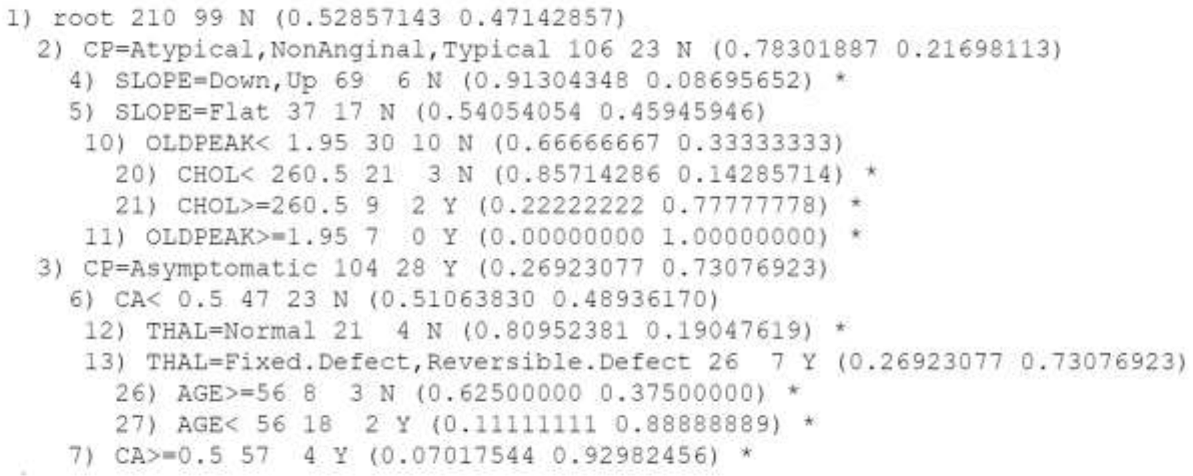
```
        fit      lwr      upr
33 13.37209 12.99409 13.75009
```

The mean fuel efficiency for this new car is 13.37209 kilometers per litre. This prediction has a 95% confidence interval of (12.99409, 13.75009) kilometers per litre.

(b) Since the new car has a mean fuel efficiency of 13.37209 km/l which is greater than the current car, therefore the model suggests that the new car will have better fuel efficiency.

Question 2

1. From the best.tree attribute in the cross-validation tree model:

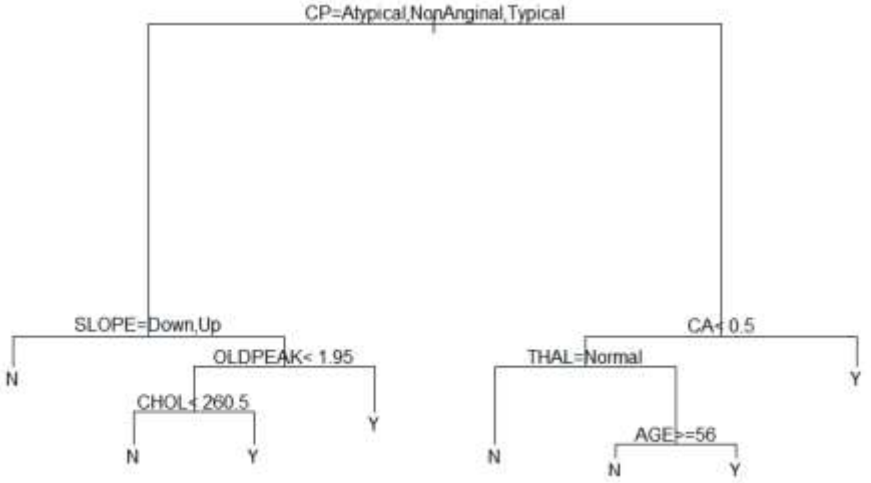


```
 1) root 210 99 N (0.52857143 0.47142857)
   2) CP=Atypical,NonAnginal,Typical 106 23 N (0.78301887 0.21698113)
     4) SLOPE=Down,Up 69  6 N (0.91304348 0.08695652) *
     5) SLOPE=Flat 37 17 N (0.54054054 0.45945946)
      10) OLDPEAK< 1.95 30 10 N (0.66666667 0.33333333)
        20) CHOL< 260.5 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 variables CP, SLOPE, OLDPEAK, CHOL, CA, THAL and AGE are used in the best tree.

From the diagram, the best tree has 8 leaves .

2. The plot of the tree is as follows:



A person is predicted to have heart disease if:

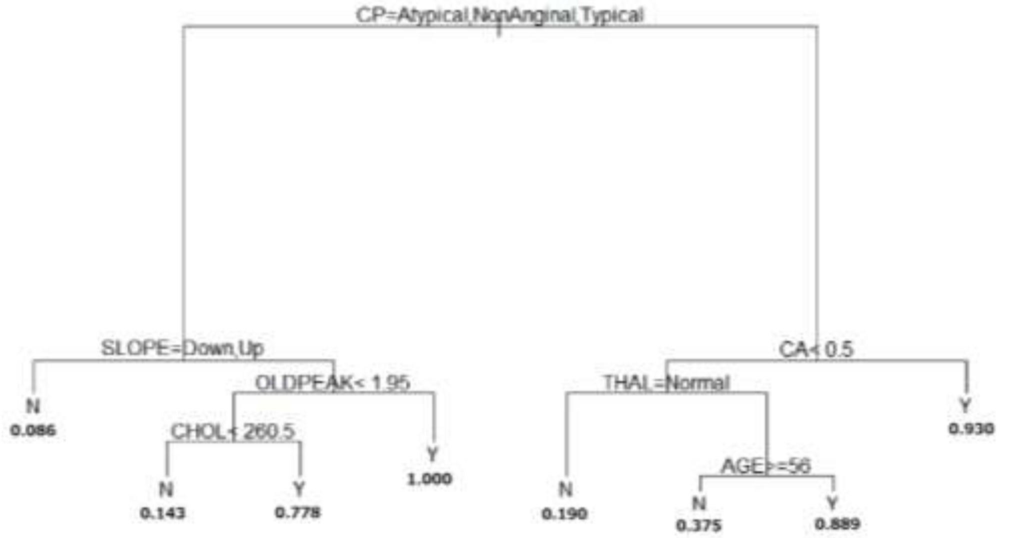
① Their chest pain type is Atypical angina, Non anginal pain, or Typical angina, has Flat slope of the peak exercise ST segment, and has exercise induced ST depression relative to rest value of greater or equal than 1.95

② Their chest pain type is Atypical angina, Non anginal pain, or Typical angina, has Flat slope of the peak exercise ST segment, has exercise induced ST depression relative to rest value of lesser than 1.95, and has serum cholesterol greater or equal than 260.5 mg/dl

③ Their chest pain type is Asymptomatic pain, and has number of major vessels colored by fluoroscopy greater or equal than 0.5

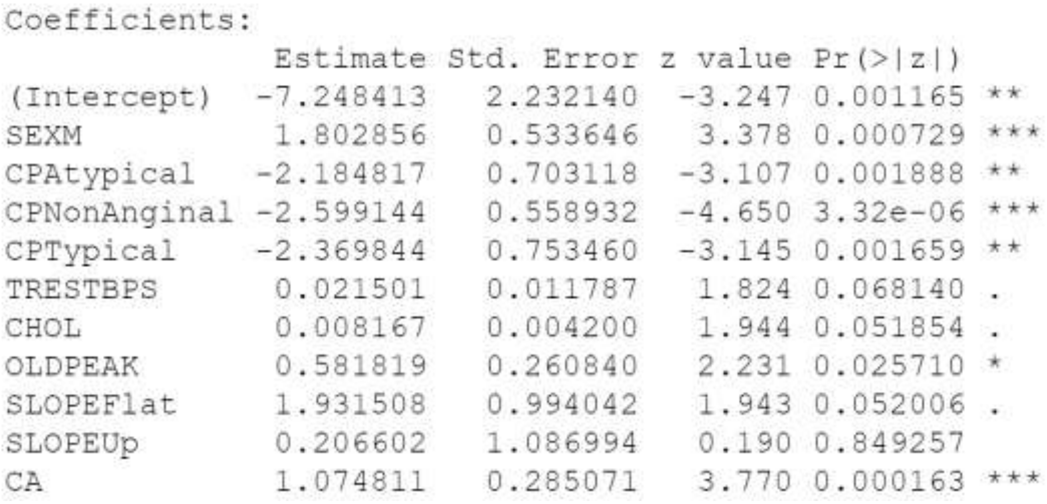
④ Their chest pain type is Asymptomatic pain, has number of major vessels colored by fluoroscopy lesser than 0.5, has Thallium scanning results showing Fixed fluid transfer defect or Reversible fluid transfer defect, and has age lesser than 56.

3. The following is a screen capture of the tree with annotated probability:



4. According to the tree, the predictor combination that results in the lowest probability of having heart-disease is CP = Atypical,NonAnginal,Typical and SLOPE = Down,Up.

5. From the summary of the final model:



```
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 ***
```

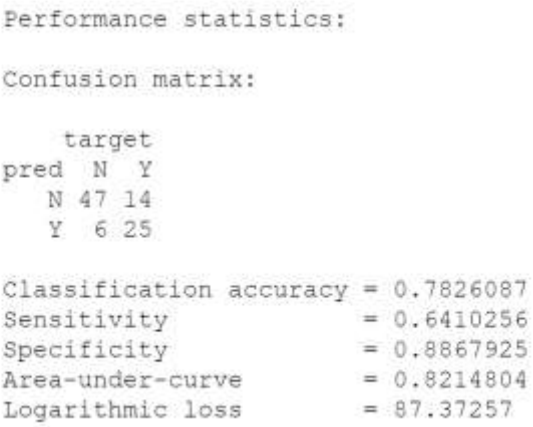
Compared to the variables used by the tree estimated by CV, the similar variables used are CP, SLOPE, OLDPEAK, CHOL, and CA. THAL and AGE are used in the tree, whereas SEX, and TRESTBPS are used in the logistic regression model.

The most important predictor is CPNonAnginal in the logistic regression, with the lowest p-value of 3.32e-06.

6. log-odds(HD) = -7.248413 + 1.802856(SEXM) – 2.184817(CPAtypical) – 2.599144(CPNonAnginal)
- 2.369844(CPTypical) + 0.021501(TRESTBPS) + 0.008167(CHOL)
+ 0.581819(OLDPEAK) + 1.931508(SLOPEFlat) + 0.206602(SLOPEUp)
+ 1.074811(CA)

7. The coefficient of CA is 1.074811, for each unit increase of CA, the mean heart disease rate increases by 1.074811, which means it has a positive effect on the mean heart disease rate of a person.

8. By running the my.pred.stats() function for the tree:



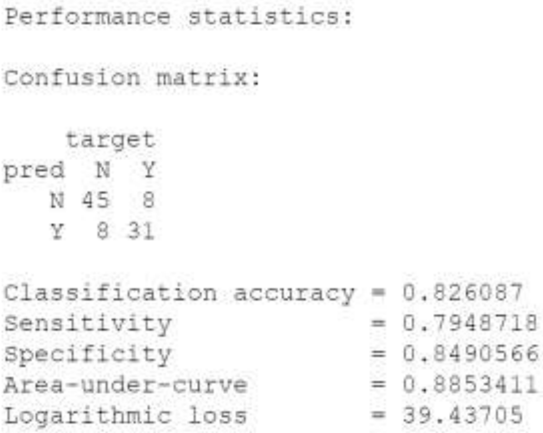
```
Performance statistics:

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
```

And running the my.pred.stats() function for the stepwise logistic regression model:



```
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
```

We can see that the stepwise logistic regression model has higher classification accuracy, higher sensitivity, higher AUC, and lower logarithmic loss than the tree. Whereas the tree has higher specificity than the stepwise logistic regression model. Overall the stepwise logistic regression model performs better than the tree since it has a better accuracy of classifying the data, and has smaller log-loss. However using the tree may be better in terms of predicting negative results (i.e. people without heart disease) due to its higher specificity.

9. (a) The tree model found using cross-validation gave an odds of 13.25.

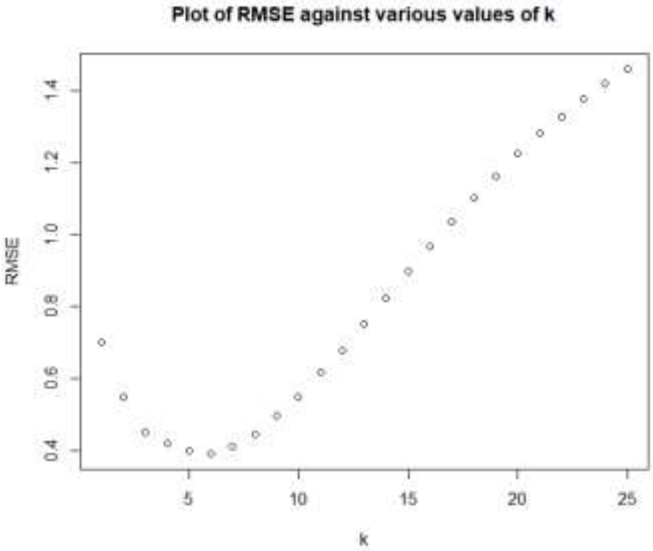
(b) The stepwise logistic regression model gave an odds 1993.977.

The stepwise logistic regression model has higher odds than the tree model, which means that it is more certain that the 10th patient has a heart disease.

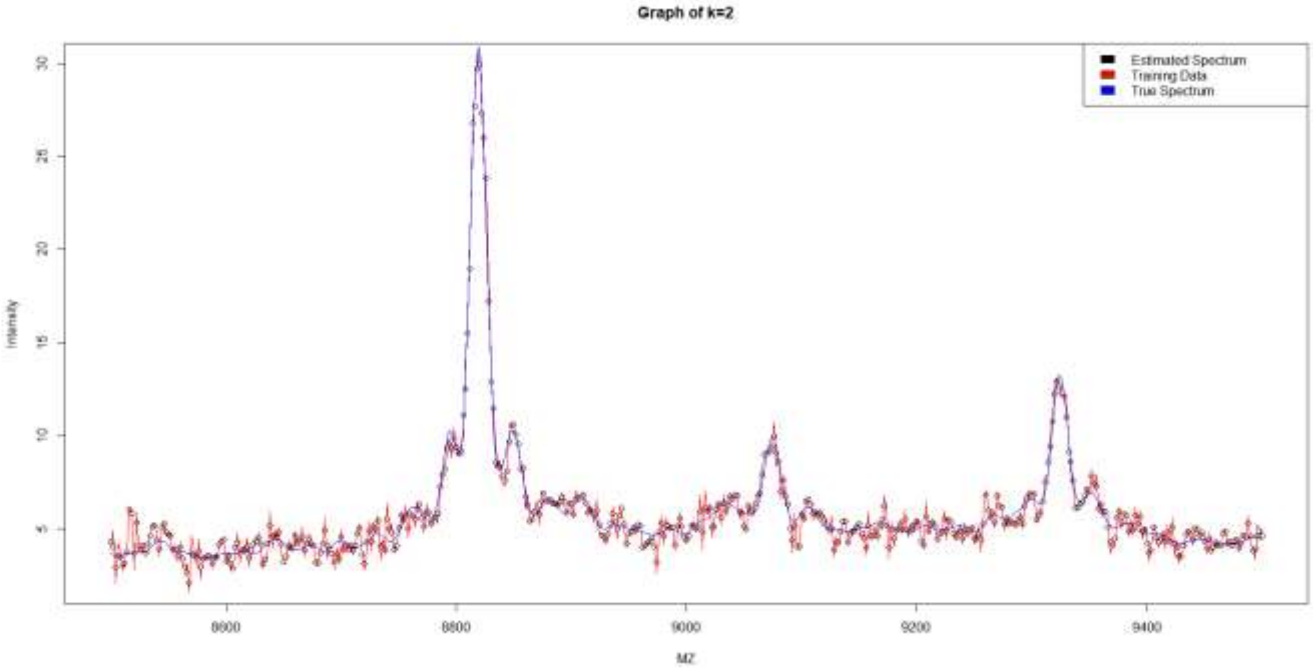
10. From running the bootstrap procedure, we get a confidence interval of (5.43, 129.52) for the 65th patient, and a confidence interval of (0.0324, 0.3380) for the 66th patient. From these intervals we can observe that there is a large difference between the upper and lower bounds, therefore there is strong evidence to suggest that there is a real difference in the population odds of having heart disease between the 65th and 66th patient.

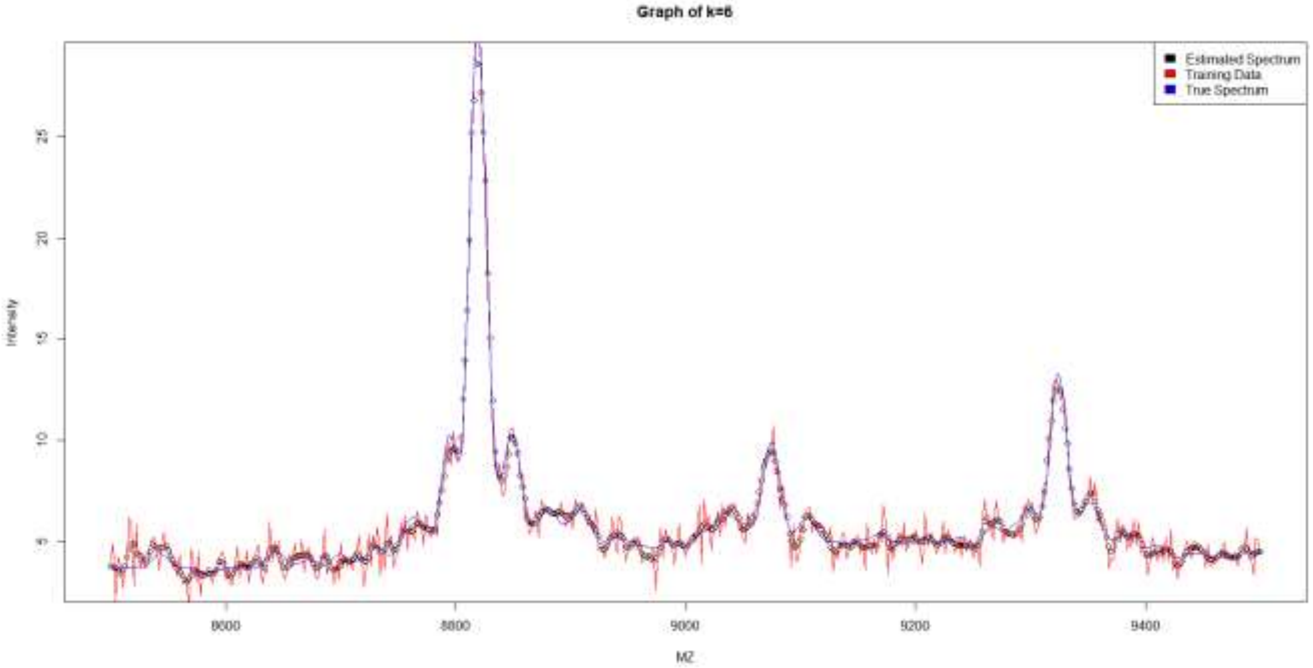
Question 3

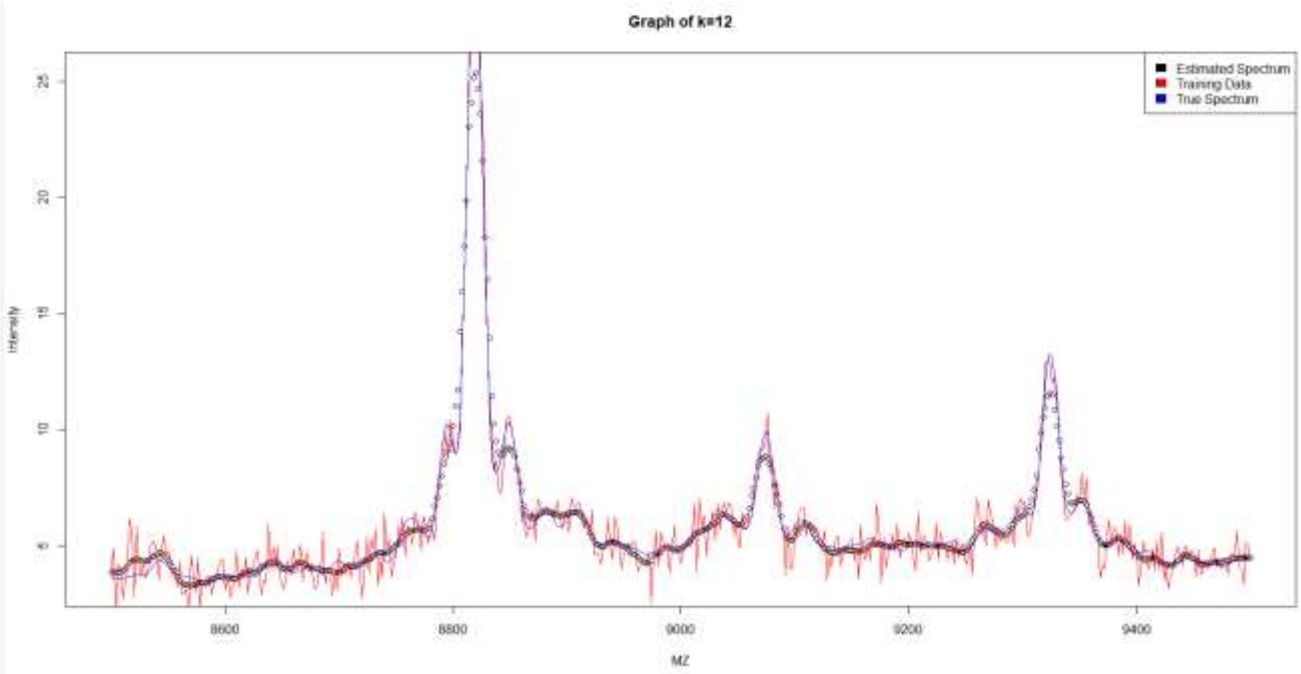
1. This is the plot of the root-mean-squared errors against the various values of k:

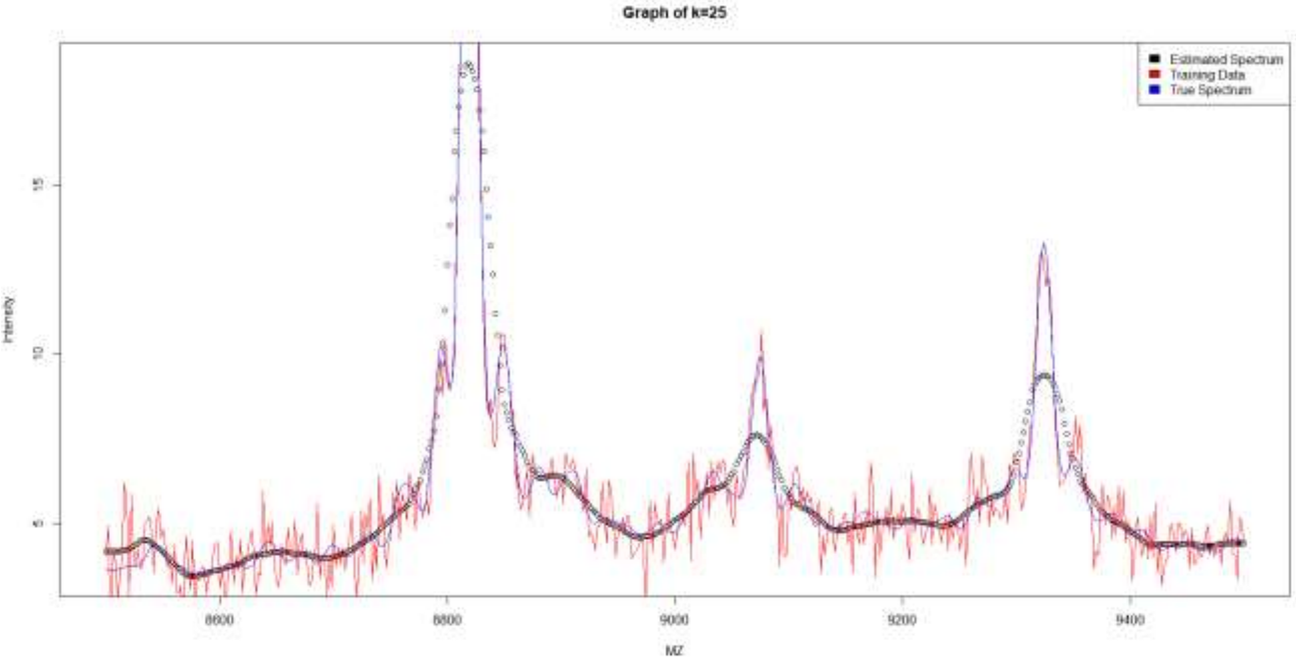


2. The four graphs are in the order of k=2, k=6, k=12 and k=25.









3. Visually, we can see that for small k, the estimates fit the peaks well but include quite a lot of noise, but as k approaches 6, the peaks are still fitted well and the noise is reduced. As k gets larger, the points have a wider spread on the peaks of the true spectrum graph. Quantitatively, starting from k=0, as k approaches 6 the root-mean-squared error decreases, and after the k=6 point, the root-mean-squared error increases. This might be due to the k-nearest neighbours algorithm underfitting the training data at small k due to only obtaining a small number of neighbours, and overfitting the training data at large k due to obtaining too many neighbours. The optimal number of neighbours to obtain for a good estimate would be 6.

4. The plotted spectrum when k=6 achieves this aim. From the graph we can see that the estimates fit the true values decently, and even at the peaks the estimates are still relatively close to the true value.

5. The cross-validation functionality selects the best value of k to be 5. It is relatively similar to the value of k that we observed (k=6) that would minimise the actual mean-squared error.

6. By obtaining the error between the estimated values and the values in ms.measured.2022$intensity, and then running the sd() function over it, an estimate of the measurement noise can be calculated, which yields 0.5906997 in this case.

7. The value of MZ corresponding to the maximum estimated intensity is found to be 8818.

8. The 95% confidence interval for the estimate of the intensity at the MZ value (8818) is found to be (25.27, 30.52) when k = 5 as determined in Q3.5. When k=3, the confidence interval is found to be (25.99, 30.66). When k=20, the confidence interval is found to be (15.48, 26.29). These confidence intervals vary in size for different values of k as the k-nearest neighbours method tends to underfit the supplied data for small k and overfit for large k.