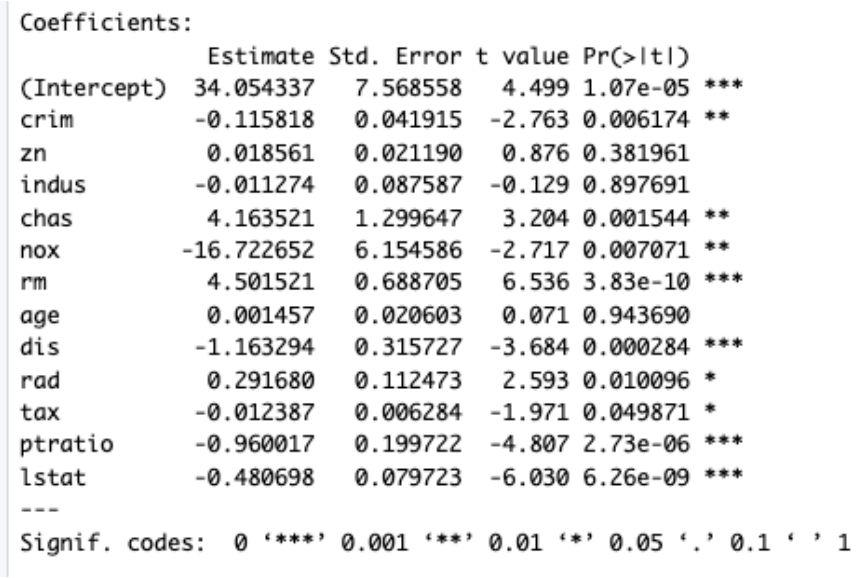
Q1.1)

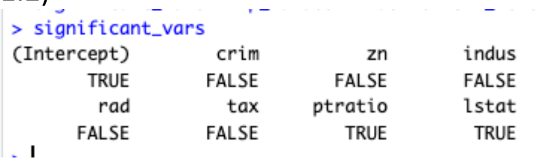


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  34.054337   7.568558   4.499 1.07e-05 ***
crim         -0.115818   0.041915  -2.763 0.006174 ** 
zn            0.018561   0.021190   0.876 0.381961    
indus        -0.011274   0.087587  -0.129 0.897691    
chas          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
```

All variables with the * symbol or more are statistically significant at the 5% confidence t test interval, since their p values are smaller than .05, meaning their probability of the effect on median house value is outside the 95% confidence interval for the effect you could expect due to randomness. So I would believe these asterisk valyes would have a significant effect on median house value.

The strongest effect on median house value is rm since it is the smallest p value it signifies its effect is strongest and least likely to occur due to randomness. Rm is rooms per dwelling which makes sense as more rooms often results bigger property costing more.

1.2)



```
> significant_vars
(Intercept)        crim          zn       indus        chas         nox          rm         age         dis 
       TRUE       FALSE       FALSE       FALSE        TRUE       FALSE        TRUE       FALSE        TRUE 
        rad         tax     ptratio       lstat 
      FALSE       FALSE        TRUE        TRUE 
```

Our significant_vars value are the p values smaller than .05/$n_{regressors}$ = 0.003846154. With Bonferroni, we only include regressors with p values smaller than 0.003846154.

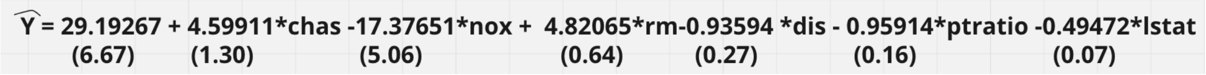
Therefore we now only deem the indus, chas, rm, dis, ptratio and lstat regressors to have adequet effect on median house price. The regressors that were dropped from the significant list with this test were the: nox, rad and tax regressors.

1.3)

The estimates on the Betas of the crim (-0.115818) and chas (4.163521) suggest that for every one % increase in the per capita crime rate, we expect the median value of the property to drop by B_{crim} * 1000 = -$115.8. The effect of being on the Charles river we expect adds B_{chas} * 1000 = $4,163.5 value to the predicted median house price.

1.4)

Final regression equation after pruning:



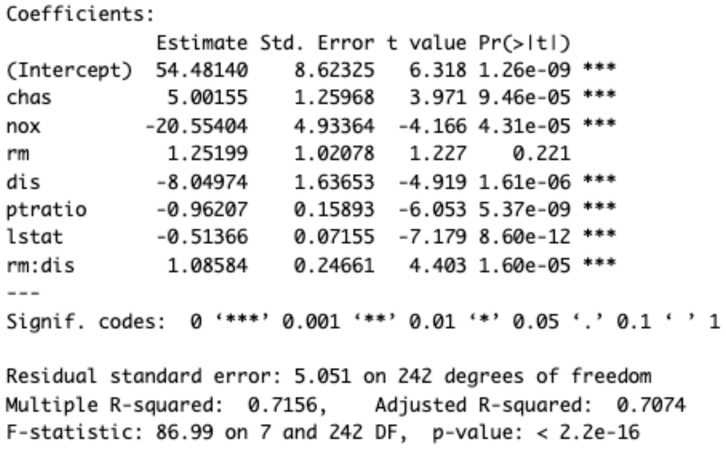
$$\hat{Y} = \underset{(6.67)}{29.19267} + \underset{(1.30)}{4.59911}*chas -\underset{(5.06)}{17.37651}*nox + \underset{(0.64)}{4.82065}*rm-\underset{(0.27)}{0.93594}*dis - \underset{(0.16)}{0.95914}*ptratio -\underset{(0.07)}{0.49472}*lstat$$

1.5)
The model above suggests the best way to improve house value would be to extend the amount of houses with frontage of the Charles river, maybe my adding inlets, reduce Nitric oxide concentrations in the area and provide benefits to help people renovate and create more rooms in their houses, since these values have the biggest effects on change of median house values per unit increase/decrease.

1.6)
The predicted median house value for this data using the model in 1.4 is:
$21,920 with a 95% confidence interval of [$20,302. , $23,537]

This means we are 95% confidence that the true value of the home lies between $20,302 and $23,537

1.7)
It would appear that the council is correct in that there is an interaction relationship between the room number and distance to employment variables, since when the product of the two was included as a regressor, it proved to be statistically significant in affecting median house value, and dropped rm as a significant regressor by itself. The new model has a higher R^2 model so this would indicate a better fit than 1.4



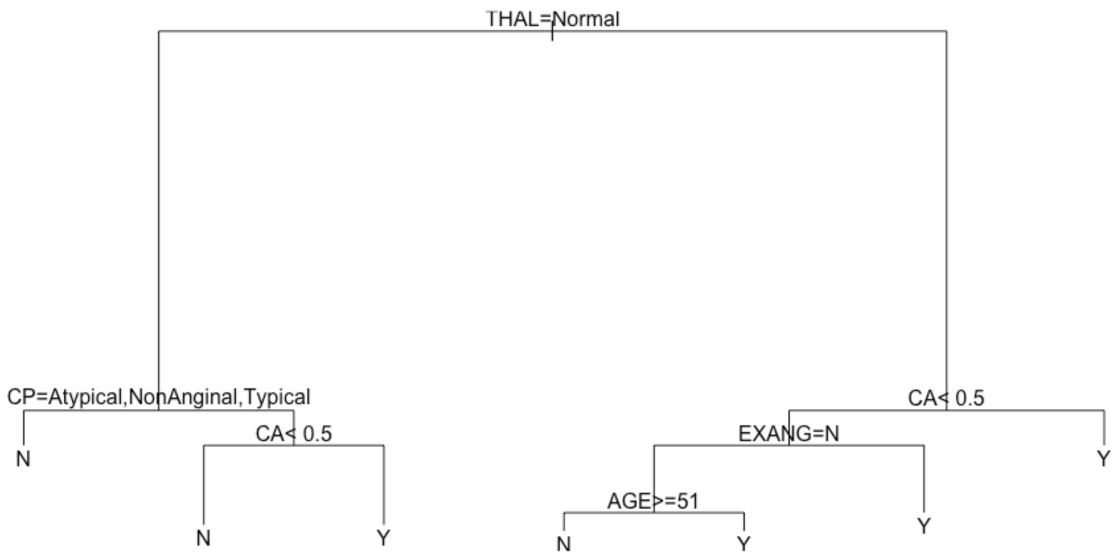
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  54.48140    8.62325   6.318 1.26e-09 ***
chas          5.00155    1.25968   3.971 9.46e-05 ***
nox         -20.55404    4.93364  -4.166 4.31e-05 ***
rm            1.25199    1.02078   1.227    0.221    
dis          -8.04974    1.63653  -4.919 1.61e-06 ***
ptratio      -0.96207    0.15893  -6.053 5.37e-09 ***
lstat        -0.51366    0.07155  -7.179 8.60e-12 ***
rm:dis        1.08584    0.24661   4.403 1.60e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.051 on 242 degrees of freedom
Multiple R-squared:  0.7156,	Adjusted R-squared:  0.7074 
F-statistic: 86.99 on 7 and 242 DF,  p-value: < 2.2e-16
```

Question 2)

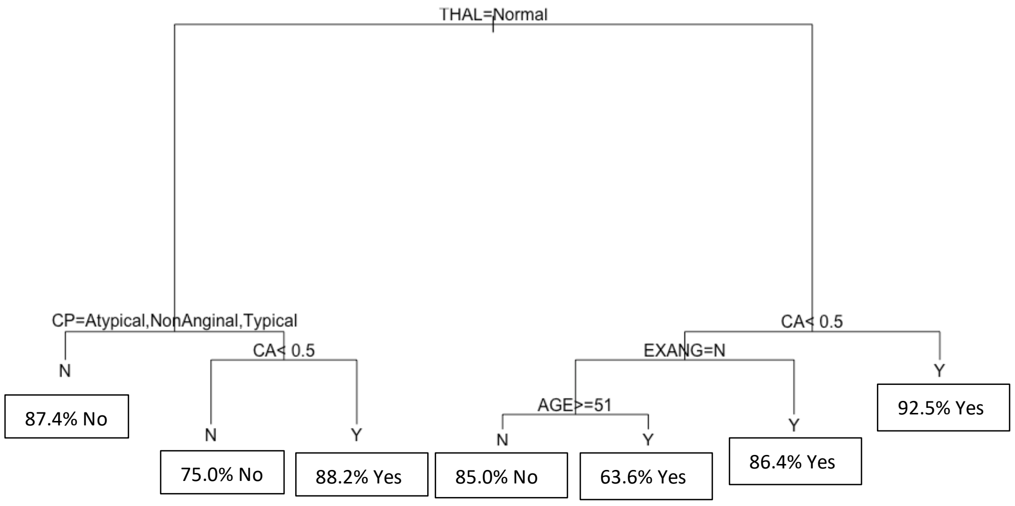
2.1)
7 leaf nodes have been used in the best tree. The variables in the best tree are THAL, CP, CA, AGE, EXANG

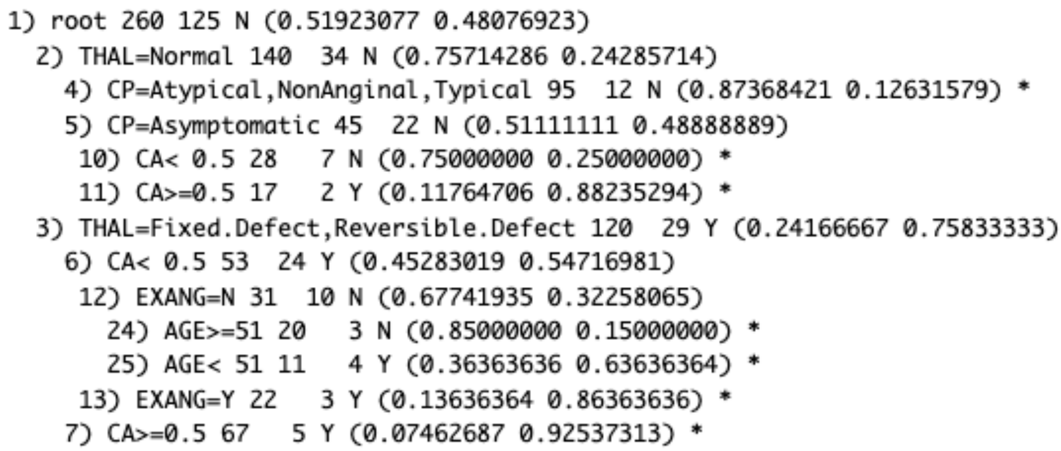
2.2)
The tree below tells us that someone is more likely to have heart disease if they adhere to one or more of the following factors:

- CA score, being the number of major blood vessels coloured by a fluoroscopy, is larger than .5.
- Exercise induced angina score was yes
- Chest pain score was something other than atypical, nonanginal or typical
- Older than 50
- Thallium Scanning results were not normal



2.3) Percentages given of chance of Heart Disease from meeting the conditions below





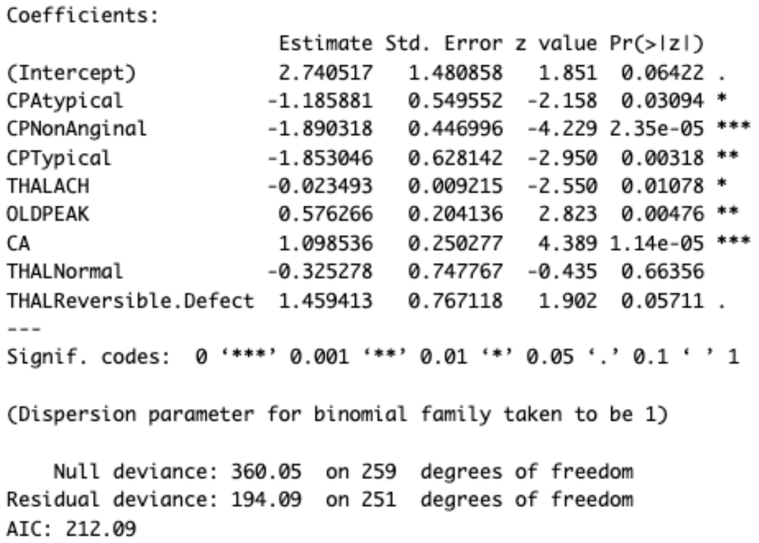
```
1) root 260 125 N (0.51923077 0.48076923)
  2) THAL=Normal 140  34 N (0.75714286 0.24285714)
    4) CP=Atypical,NonAnginal,Typical 95  12 N (0.87368421 0.12631579) *
    5) CP=Asymptomatic 45  22 N (0.51111111 0.48888889)
     10) CA< 0.5 28   7 N (0.75000000 0.25000000) *
     11) CA>=0.5 17   2 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) *
```

2.4)
A Not normal THAL score and a CA score of less than .5 results in the highest probability of having heart disease

2.5)
The variables included in a BIC pruning of a logistic model for Heart Disease prediction are CPAtypical, CPNontypical, CPTypical, THALACH, OLDPEAK, CA, THALNORMAL and THALReversible,Defect

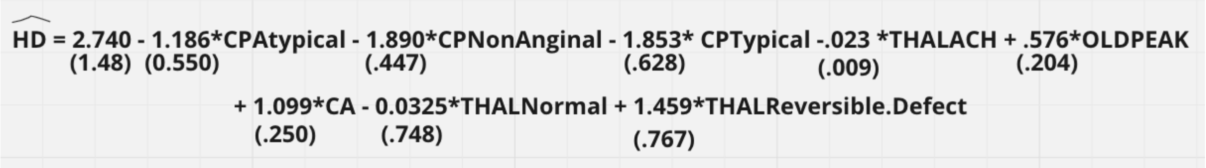
The strongest regressor to predict heart disease remains the CA score as this has the lowest p value meaning least likely to have its effect caused by randomness.

The variables compared to the tree are similar, the strongest influencer is the same being CA. Variables which in the tree diagram if true were likely to predict no heart disease were CP of Atypical, Cp of Nontypical or Typical. We see these three variables included as important in the logistic regression but causing a negative impact on the likelihood of heart disease. Meaning in both models these factors are important in predicting no heart disease.



```
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
```

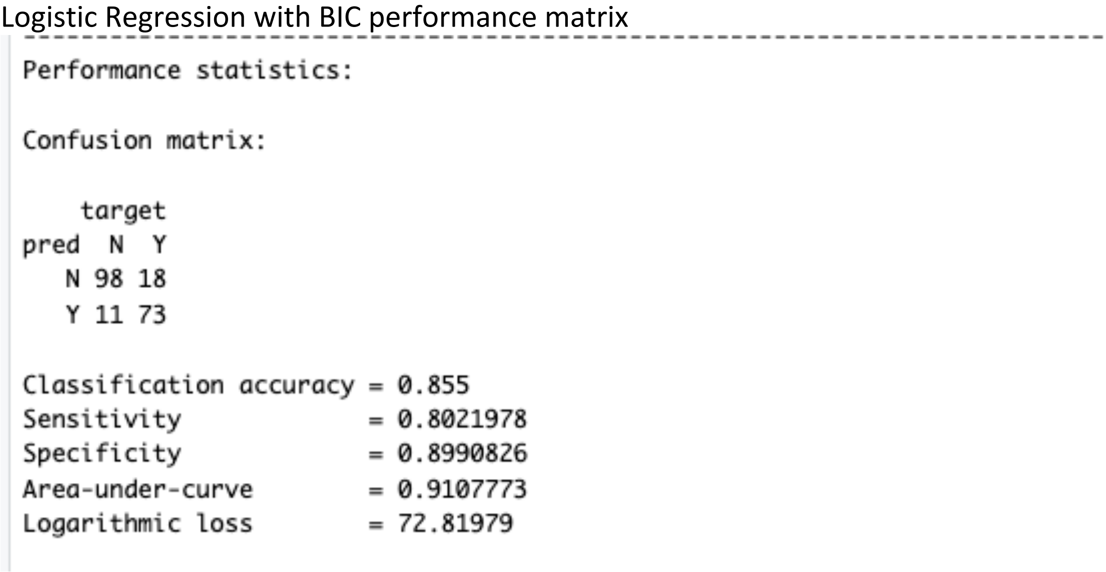
2.6)



$$\widehat{HD} = \underset{(1.48)}{2.740} - \underset{(0.550)}{1.186}*CPAtypical - \underset{(.447)}{1.890}*CPNonAnginal - \underset{(.628)}{1.853}*CPTypical - \underset{(.009)}{.023}*THALACH + \underset{(.204)}{.576}*OLDPEAK$$

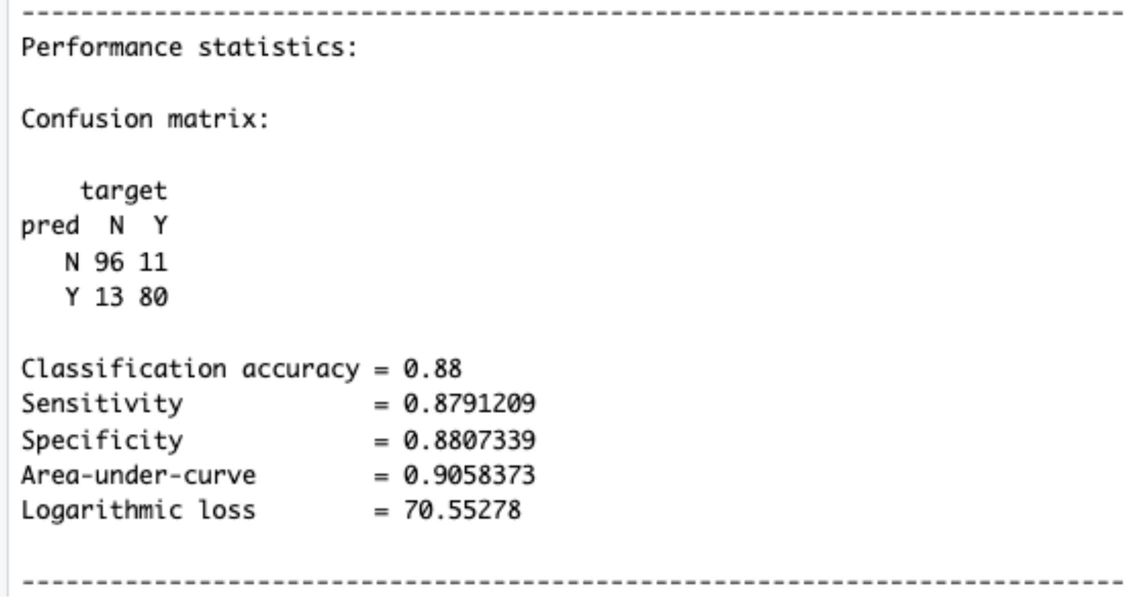
$$+ \underset{(.250)}{1.099}*CA - \underset{(.748)}{0.0325}*THALNormal + \underset{(.767)}{1.459}*THALReversible.Defect$$

2.7)
Logistic Regression with BIC performance matrix



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

Decision Tree confusion matrix



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

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

It appears the accuracy of the decision tree classification method is more accurate however this is not always the best indicator of a better model. The area under the curve is higher for the logistic regression model, as well as the log loss which we want to have larger. As these are more well balanced performance metrics, Eg since the area under the curve takes into account the Specificity vs Sensitivity trade off and optimises such, we would say the logistic model is a better model for predicting Heart Disease.

2.8) 69th patient data as input

Output from prediction of 69 from best tree is probability of HD is 0.8636.

Odds = p/1-p = 0.8636/0.1363 = 6.33 to 1. So person. 68 is 6.22 x more likely to have HD than to not from the decision tree.

Output from prediction of 69 from logistic is probability of HD is 2.87015

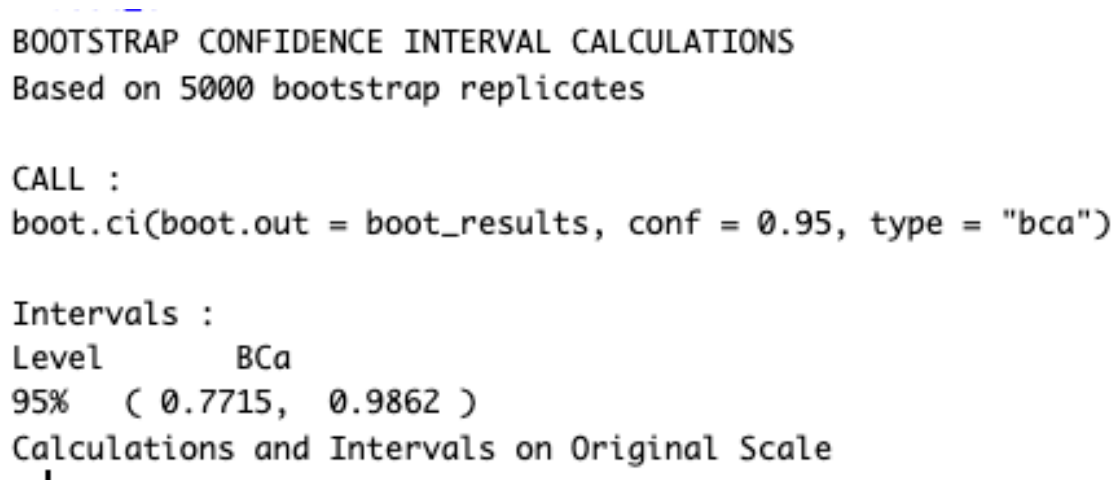
To convert to probability we do p = 1/(1+e^2.87015) = .946

Converting to odds = .946/(1-.946) = 17.53. So based on logistic regression person 69 is 17.653 x more likely to have HD than to not.

The odds in this situation are significantly different. However if you look at the probabilities, .86 and .946 they are not too different, The difference between odds is and probability is odds are a ratio of chance of HD to chance not HD. Meaning they exponentially grow larger as the probability of HD is higher. Thus for the small difference in probability of predicted HD, we see a large difference in odds because we are in high probability region and odds has grown exponentially.

2.9)

The 95% confidence interval for the 69th patient having Heart Disease using a logistic regression model from the bootstrapped bca on the BIC identified regressors method is between 77.15% likely and 98.62% likely. Meaning we expect the true population value of their likelihood to have heart disease to be within this range. Both predictions of the likelihood from the decision tree and logistic model trained on the training data predicted a probability of Heart disease to be .86 and .946 respectively, which both fall within the confidence interval suggested by the bootstrapped bca method, meaning both models could accurately predict the true heart disease probability, as both their predictions are within confidence interval range.

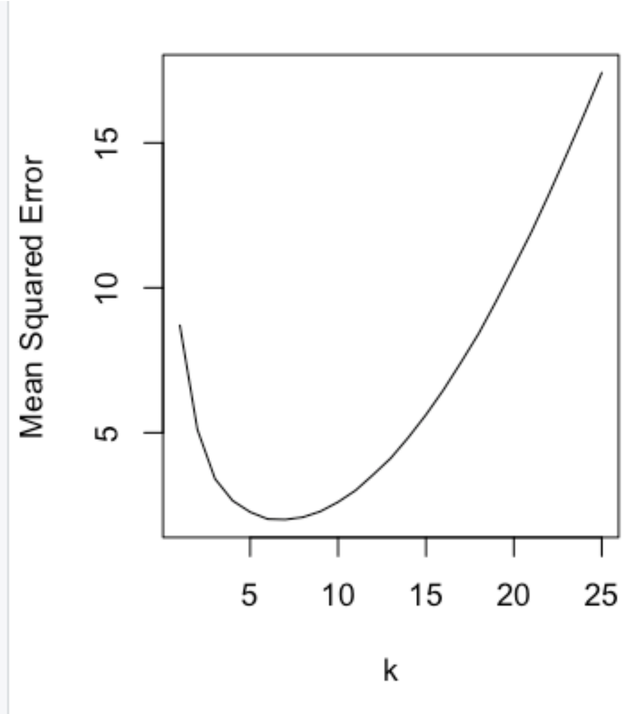


```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

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

Intervals :
Level       BCa
95%   ( 0.7715,  0.9862 )
Calculations and Intervals on Original Scale
```

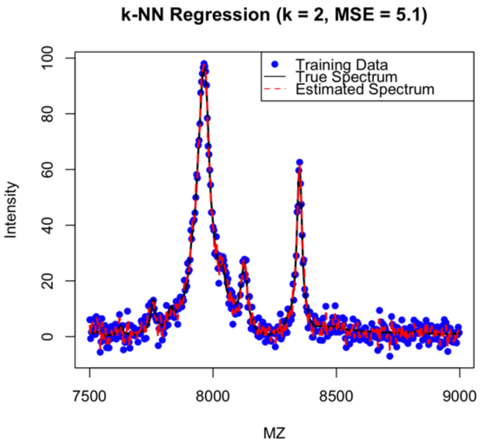
Question 3)

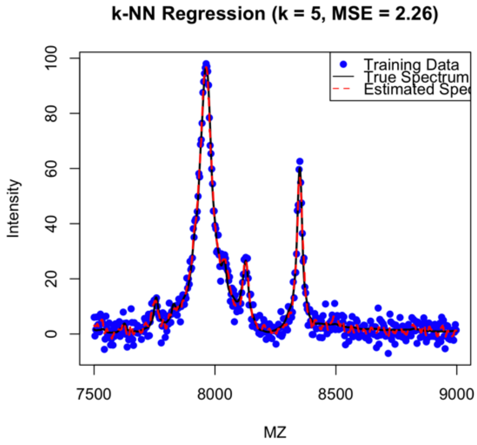
3.1)

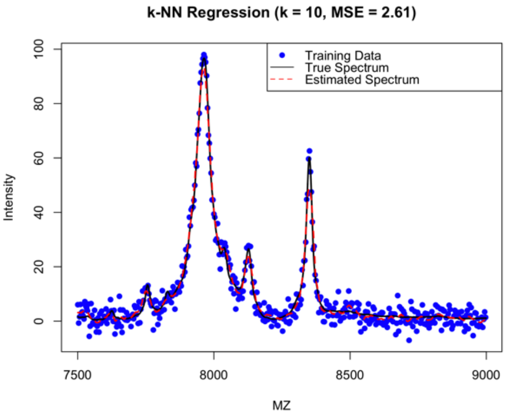
Plot of MSE vs k for k from 1 to 25. Can see minimum of graph at k = 6

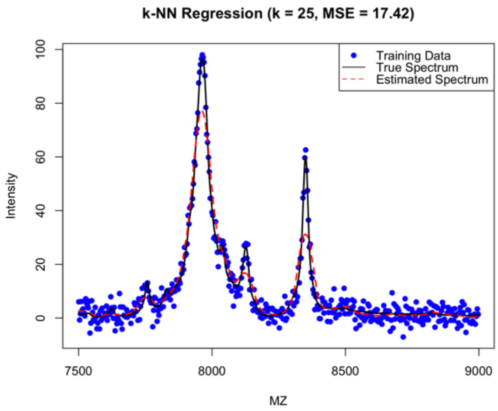


3. 2)





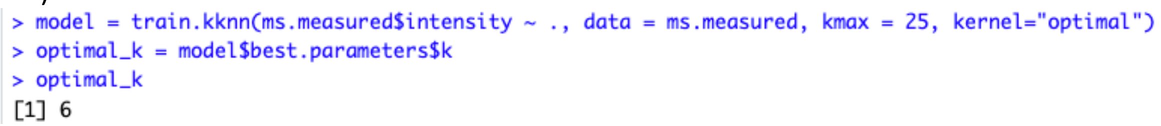




3.3)
We can see from the MSE on each chart, that the k=5 has the smallest MSE meaning the predicted line bets fits the true intensity line, with a MSE of 2.26. This would indicate a k = 5 would result in the best predictions for intensity. The k = 25 graph has the red line significantly lower at the peaks than the actual black line, this is because the 25 points used to generate an estimate reduce the significance of the peaks, which is bad for our use case as we need to clearly identify peaks in order to class the chemicals being studied, so we definitely would not want to use a model that dulls the peaks.

3.4) Yes, the k=5 method produces a very smooth, close to the true value intensities, with very little noise affecting its predictions. The accuracy is quit high with a low MSE of only 2.26, meaning its predictions are very close to actual true values. The kNN method is good at eliminating noise as it effectively takes an average of k amount of points, reducing the distance for the guess between all k points, which means that noise that has is N(0,ó^{2}) distributed effectively is cancelled out because we expect the average value of the noise to be 0.

3.5)



```
> model = train.kknn(ms.measured$intensity ~ ., data = ms.measured, kmax = 25, kernel="optimal")
> optimal_k = model$best.parameters$k
> optimal_k
[1] 6
```

Optimal k value is 6 which aligns with 3.1 as approximately the minimum of the MSE graph.

3.6)
For our optimal k = 6 value we have a MSE of 2.02. Given that MSE formula is the same as the Variance formula. Take the square root would give the standard deviation of the equipment.

Sqrt(2.02) = 1.42 small standard deviation error of equipment due to reducing the noise component by with the k nearest neighbours method.

3.7) The value of MZ that corresponds with the greatest intensity score is 7963.3

3.8)
95% confidence intervals for

K = 3:
95% confidence true intensity levels are between (95.11, 98.00)
K = 6
95% confidence true intensity levels are between (91.51, 97.91)
K = 20
95% confidence true intensity levels are between (69.57, 93.31)

As we saw in 3.1, MSE increases as k is less than its minimum 6 and increases as k becomes greater than 6. This means the predicted intensity is further away from the actual bar as we change k. We saw in the 4 charts, that when K is large, the average distance between k points means that the peaks are not predicted to be as large as they actually are, this means that our largest peak, the prediction will be most off when k is largest, here when k = 20. Thus the 95% confidence interval for k = 20 is expected to be quite large to include the true value in its range, given we average our peaks down by including more data points. When k is small we expect to see the peaks higher as we have not averaged them down so the confidence interval less, so just because k = 3 has the smallest confidence interval here, doesn't mean it is the best fit as the 3 point average means more noise affects the predictions in other parts of the model, since the Mean Squared Error is less than when K = 6.