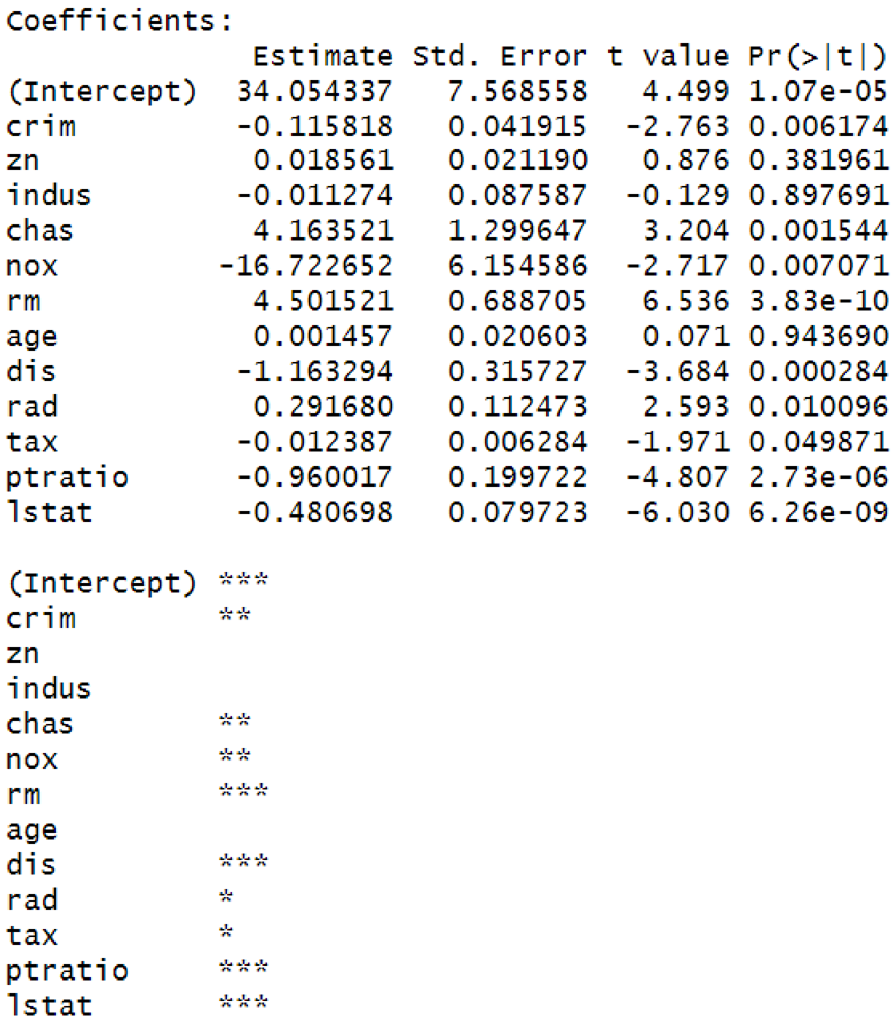
FIT2086 Assignment 3

Question 1

1.

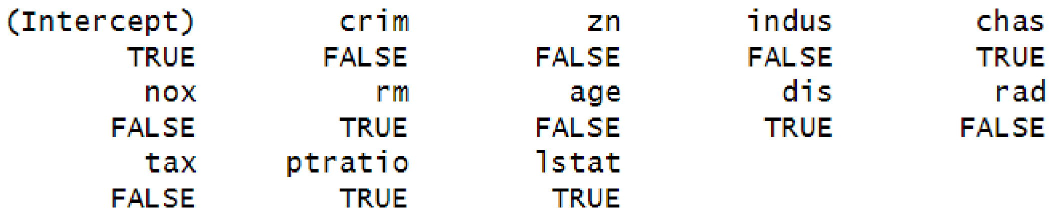


```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   34.054337   7.568558   4.499 1.07e-05
crim          -0.115818   0.041915  -2.763 0.006174
zn             0.018561   0.021190   0.876 0.381961
indus         -0.011274   0.087587  -0.129 0.897691
chas           4.163521   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

(Intercept) ***
crim        **
zn
indus
chas        **
nox         **
rm          ***
age
dis         ***
rad         *
tax         *
ptratio     ***
lstat       ***
```

The predictors that are possibly associated with median house value include “crim”, “chas”, “nox”, “rm”, “dis”, “rad”, “tax”, “ptratio” and “lstat”. The 3 variables that appear to be the strongest predictors of housing price, based on their coefficient magnitudes and low p-values (0.05 <) , are “rm”, “ptratio”, and “lstat” with "rm" having the most substantial positive effect and "lstat" having the most significant negative effect.

2.

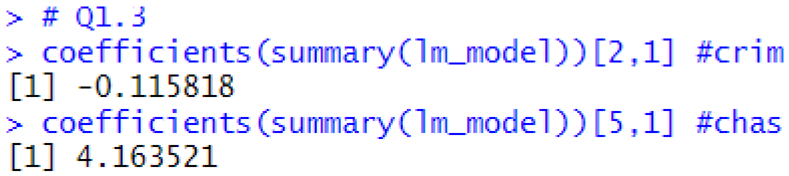
By the outcome of running the Bonferroni procedure with $\alpha$ = 0.05:



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

The predictors that are possibly associated with median house value include “chas”, “rm”, “dis”, “ptratio”, “lstat”. The predictors “crim”, “nox”, “rad”, “tax” are eliminated from the procedure compare to the previous model.

3.

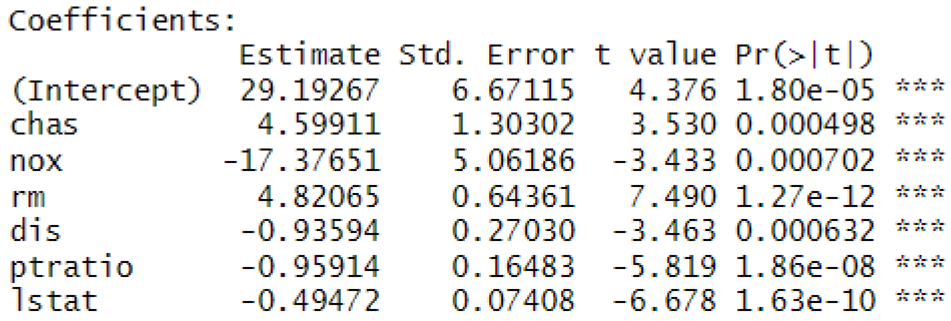


```
> # Q1.3
> coefficients(summary(lm_model))[2,1] #crim
[1] -0.115818
> coefficients(summary(lm_model))[5,1] #chas
[1] 4.163521
```

According to the model summary, the coefficient of per-capita crime rate (crim) is -0.115818, for each unit increase in the per-capita crime rate, the median house price tends to decrease by approximately $115.82 (in thousands of dollars). This suggests that areas with higher crime rates tend to have lower median house prices. The coefficient of Charles River (chas) is 4.163521, if a suburb has frontage on Charles River, the median house price tends to be $4163.52 (in thousands of dollars) higher compared to suburbs without river frontage. This suggests that riverfront properties tend to have higher median house prices.

4.

The summary of the fitted model is now as follows:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  29.19267    6.67115   4.376 1.80e-05 ***
chas          4.59911    1.30302   3.530 0.000498 ***
nox         -17.37651    5.06186  -3.433 0.000702 ***
rm            4.82065    0.64361   7.490 1.27e-12 ***
dis          -0.93594    0.27030  -3.463 0.000632 ***
ptratio      -0.95914    0.16483  -5.819 1.86e-08 ***
lstat        -0.49472    0.07408  -6.678 1.63e-10 ***
```

The final regression equation obtained after pruning would be:

Log-odds(medv) = 29.19267 + 4.59911(chas) – 17.37651(nox) + 4.82065(rm) – 0.93594(dis) – 0.95914(ptratio) – 0.49472(lstat)

5.

The council could consider improving the median house value in their suburb:
Riverfront Development (chas):

Suburbs with riverfront property tend to have higher median house values.

Suggestion: Invest in riverfront development, beautification, and recreational amenities to attract homebuyers and increase property values.

Number of Rooms (rm):

More rooms per dwelling are associated with higher median house values.

Suggestion: Encourage the construction or renovation of houses with more rooms, creating larger and more desirable homes.

Distance to Employment Centers (dis):

Decreased distance to employment centers is associated with higher median house values.

Suggestion: Improve transportation infrastructure to reduce commute times, making the area more appealing to potential homebuyers.

School Quality (ptratio):

Lower pupil-teacher ratios are associated with higher median house values.

Suggestion: Invest in education and reduce pupil-teacher ratios in local schools to attract families and improve property values.

Negative Impacts:

Air Quality (nox):

Higher levels of nitrogen oxides are associated with lower median house values.

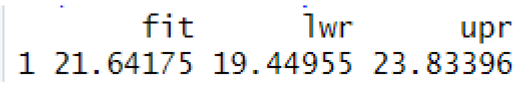
Suggestion: Implement pollution reduction measures and promote green initiatives to improve air quality and boost property values.

Percentage of Lower Status Population (lstat):

A higher percentage of lower status population is associated with lower median house values.

Suggestion: Implement socioeconomic development programs to uplift the community and enhance the attractiveness of the area to homebuyers.

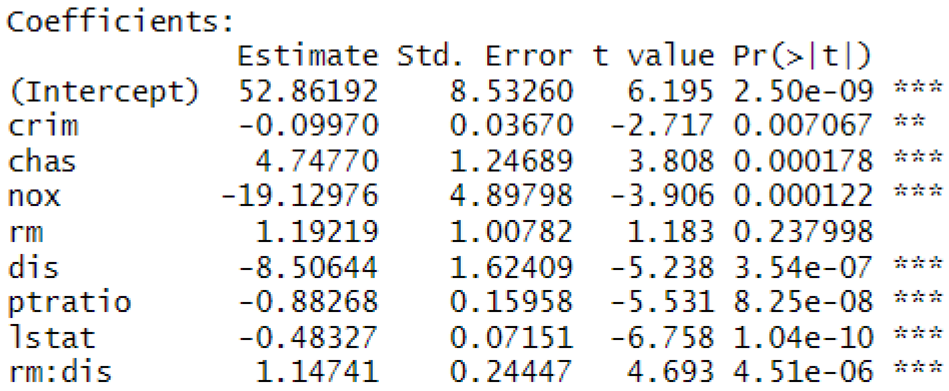
6.



```
       fit      lwr      upr
1 21.64175 19.44955 23.83396
```

The average median house price in the new suburb is approximately $21.64 thousand. According to the prediction, we can be 95% confident that the median house price is within the range of $19.45 thousand to $23.83 thousand.

7.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   52.86192    8.53260   6.195 2.50e-09 ***
crim          -0.09970    0.03670  -2.717 0.007067 ** 
chas           4.74770    1.24689   3.808 0.000178 ***
nox          -19.12976    4.89798  -3.906 0.000122 ***
rm             1.19219    1.00782   1.183 0.237998    
dis           -8.50644    1.62409  -5.238 3.54e-07 ***
ptratio       -0.88268    0.15958  -5.531 8.25e-08 ***
lstat         -0.48327    0.07151  -6.758 1.04e-10 ***
rm:dis         1.14741    0.24447   4.693 4.51e-06 ***
```

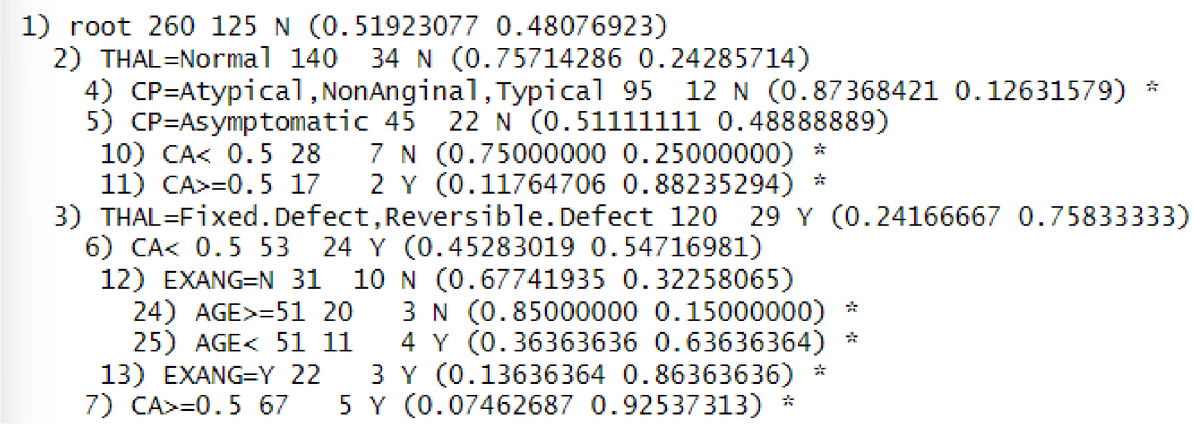
According to the findings, it's evident that the interaction term 'rm:dis' has a coefficient of 1.14741 and an extremely low p-value of 4.51e-06, signifying its statistical importance.

The existence of this interaction effect implies that the connection between the average room count in a residence and the median house price is not consistent throughout all distances to employment centers.

Question 2

1.

Referring to the information in the best.tree attribute of the cross-validation tree model:

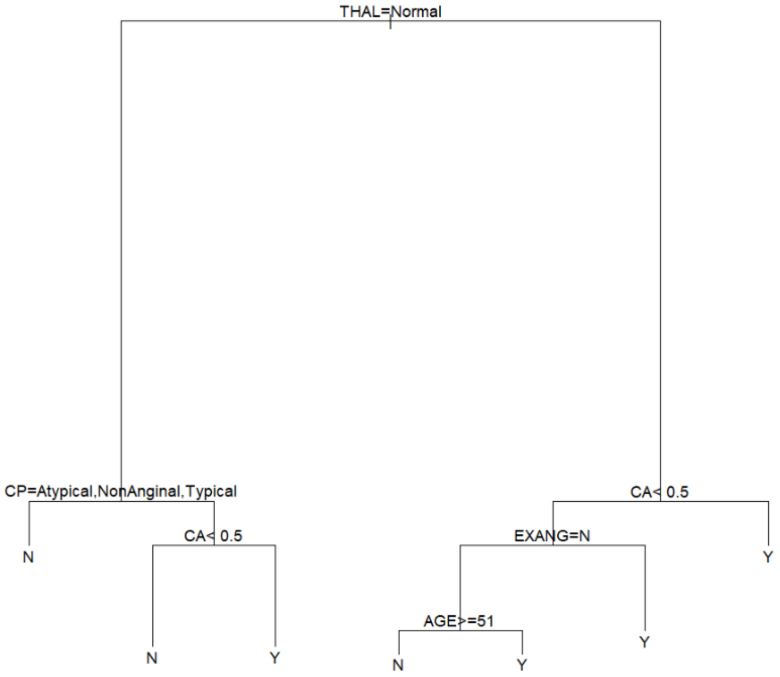


```
1) root 260 125 N (0.51923077 0.48076923)  
  2) THAL=Normal 140  34 N (0.75714286 0.24285714)  
    4) CP=Atypical,NonAnginal,Typical 95  12 N (0.87368421 0.12631579) *
    5) CP=Asymptomatic 45  22 N (0.51111111 0.48888889)  
     10) CA< 0.5 28   7 N (0.75000000 0.25000000) *
     11) CA>=0.5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

The variables THAL, CP, CA, EXANG, AGE are used in the best tree and it has total of 7 leaves.

2.

The plot of the tree is as follows:

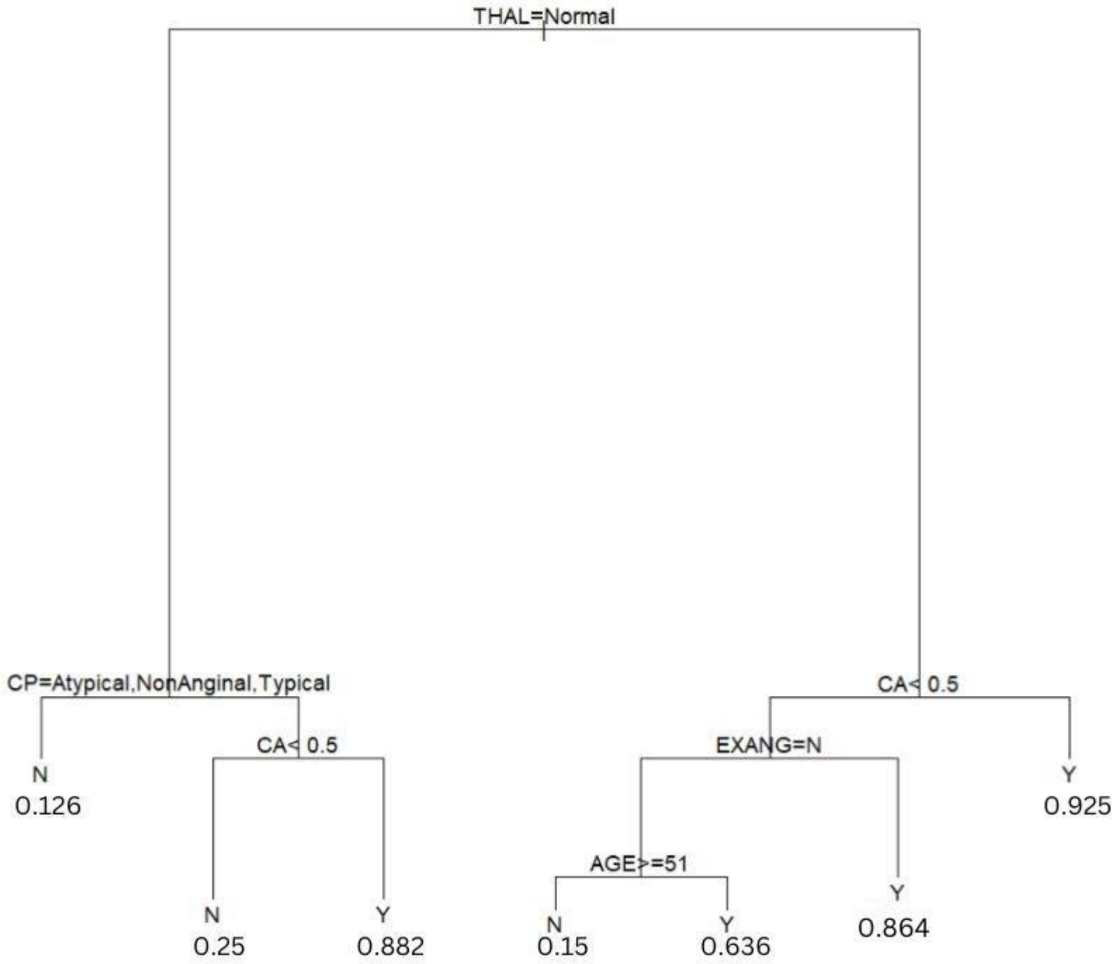


A person is predicted to have heart disease if:

1. Their Thallium scanning result (THAL) is normal and chest pain type is one of atypical angina, non-anginal pain, or typical angina, and the number of major vessels coloured by fluoroscopy (CA) is greater or equal to 0.5.
2. Their Thallium scanning result (THAL) is not normal, but number of major vessels coloured by fluorosopy (CA) is greater or equal to 0.5.
3. Their Thallium scanning result (THAL) is not normal, and number of major vessels coloured by fluorosopy (CA) is less than 0.5, but exercise-induced angina (EXANG) is present.
4. Their Thallium scanning result (THAL) is not normal, and number of major vessels coloured by fluorosopy (CA) is less than 0.5, and exercise-induced angina (EXANG) is absent but has an age greater or equal to 51.

3.

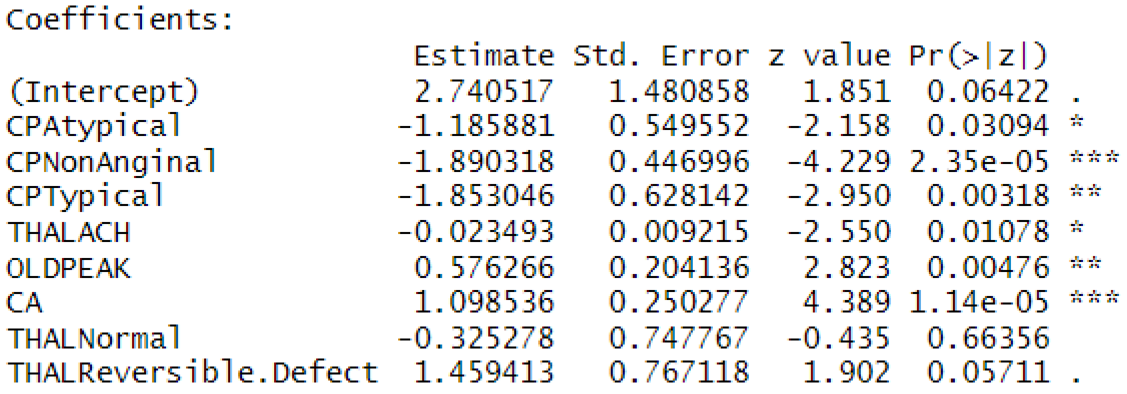
The plot of tree added with probability has attached below:



4. According to the tree, the predictor combination that results in the highest probability of having heart disease is while the patient THAL result is not Normal and CA >= 0.5, which is 0.925.

5.

From the summary of the final model:



```
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 .  
```

The final model variables are CP, THALACH, OLDPEAK, CA and THAL.

Compared to the variables used by the tree estimated by CV, the similar variables used are THAL, CP and CA, while the logistic regression model includes additional variables like THALACH and OLDPEAK.

The most important predictor is CA in the logistic regression, with the lowest p-value of 1.14e-05.

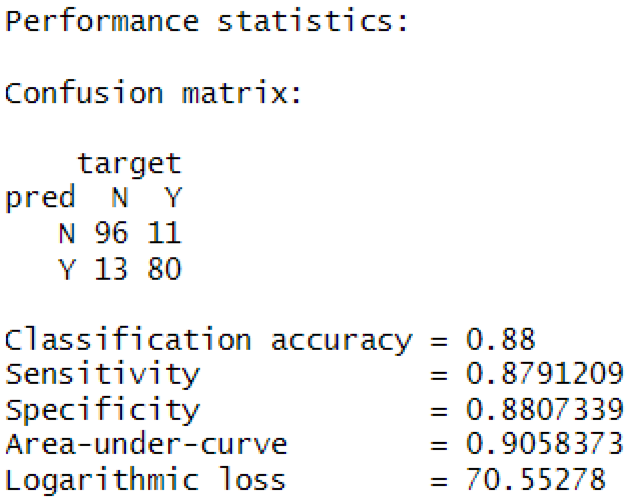
6.

The regression equation is as follows:

Log-odds(HD) = 2.740517 – 1.185881(CPAtypical) -1.890318(CPNonAnginal) – 1.853046(CPTypical) - 0.023493(THALACH) + 0.576266(OLDPEAK) + 1.098536(CA) - 0.325278(THALNormal) + 1.459413(THALReversible.Defect)

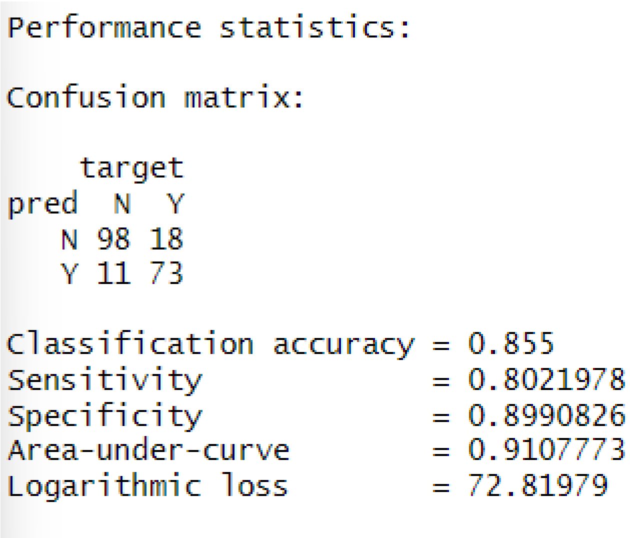
7.

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



```
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
```

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



```
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 decision tree model is better in terms of sensitivity and classification accuracy, making it better at correctly identifying patients with heart disease. On the other hand, the logistic regression model has higher specificity and slightly lower logarithmic loss, indicating better performance in correctly identifying patients without heart disease.

The choice between the two models depends on the specific goals and constraints of the diagnostic task. If it is more critical to correctly identify patients with heart disease, the decision tree model might be more preferred. If minimizing false positives (correctly identifying patients without heart disease) is more important, the logistic regression model might be more preferred. Ultimately, the choice should consider the trade-offs between sensitivity and specificity that align with the clinical context and objectives.

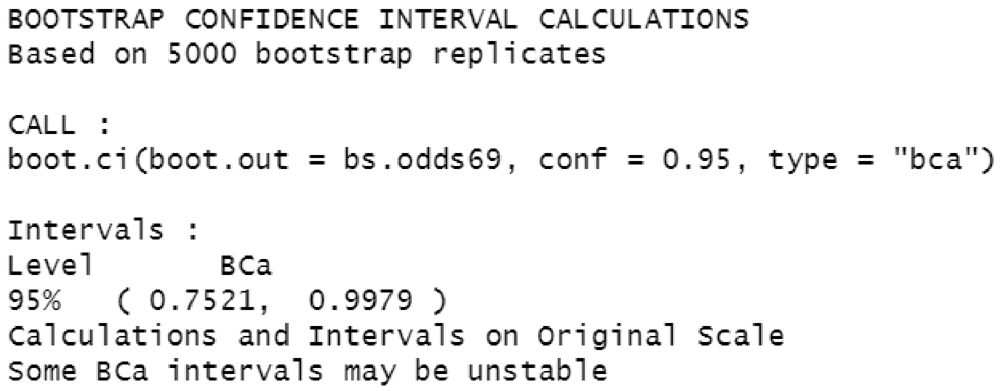
8.

a) The tree model found using cross-validation gave an odds of 6.333.

b) The stepwise logistic regression model gave an odds of 17.63966.

The stepwise logistic regression model has higher odds than the tree model, which means that it is more certain that the 69th patient has a heart disease, in other words, it considers the selected predictors more strongly when predicting heart disease.

9.



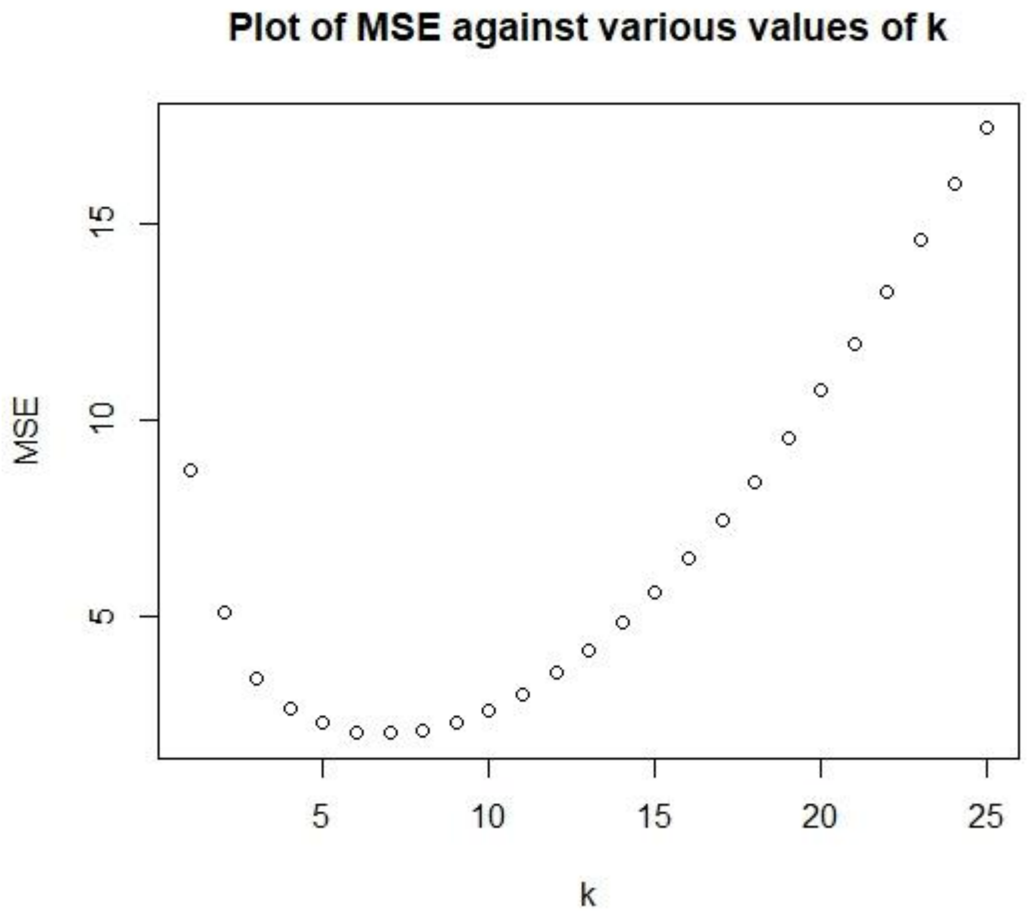
```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

CALL :
boot.ci(boot.out = bs.odds69, conf = 0.95, type = "bca")

Intervals :
Level       BCa
95%   ( 0.7521,  0.9979 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
```

From running the bootstrap procedure, we get a confidence interval of (0.7521, 0.9979), this interval suggests a very high probability of the 69th patient having heart disease, with a lower bound of 0.7521 and an upper bound of 0.9979.

Question 3

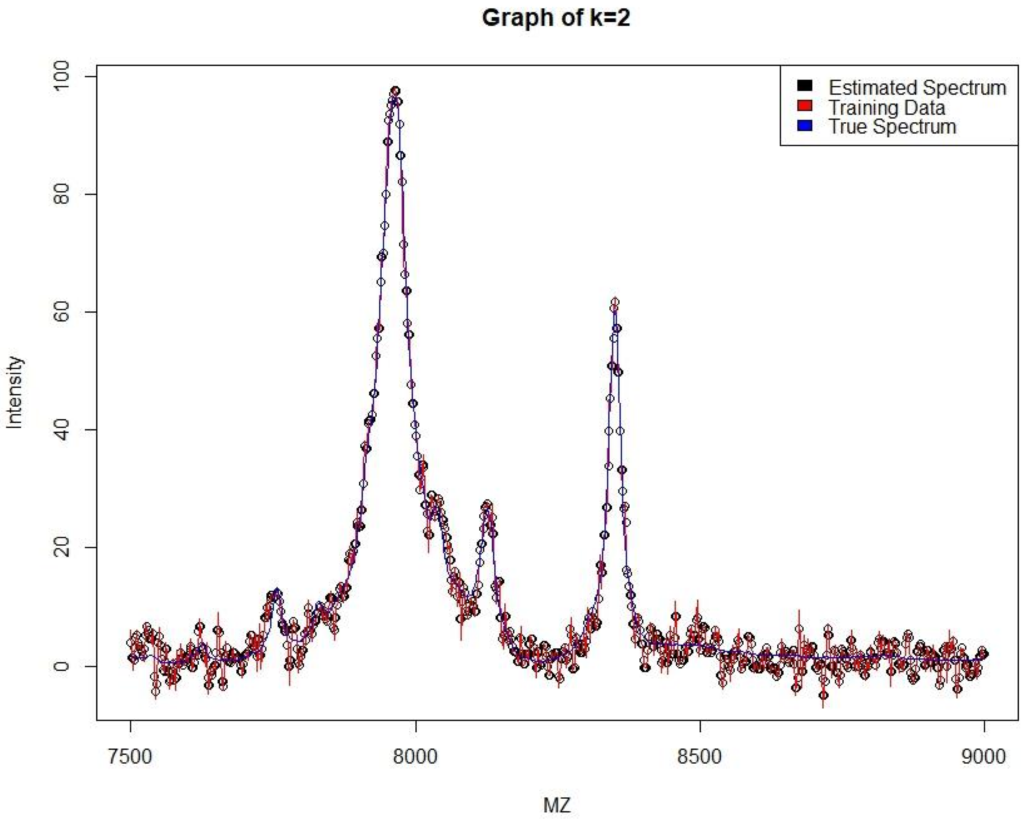
1.

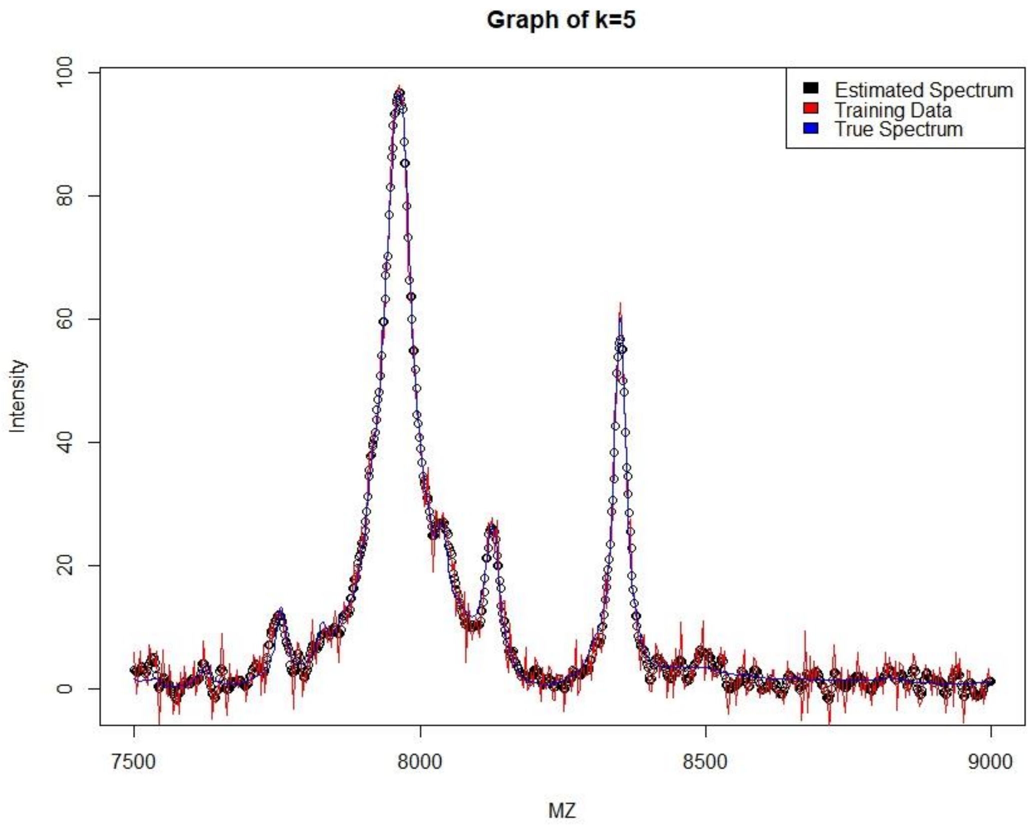
This is the plot of the mean-squared errors against the various values of k:

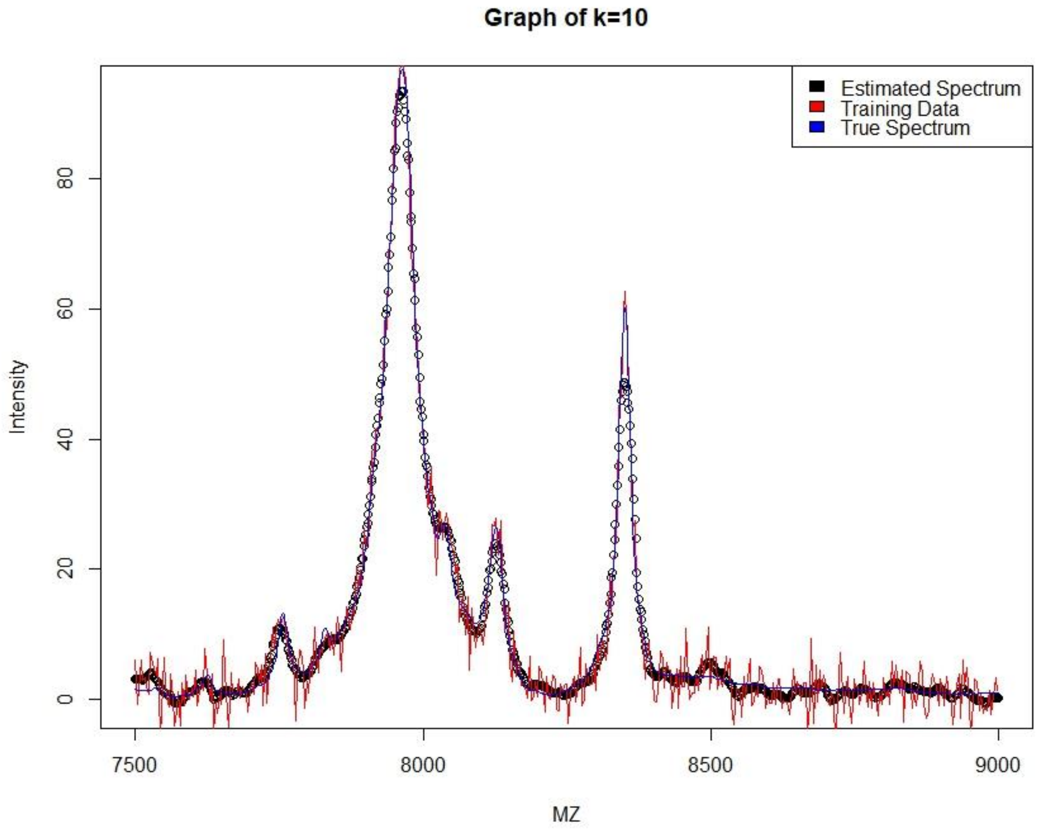


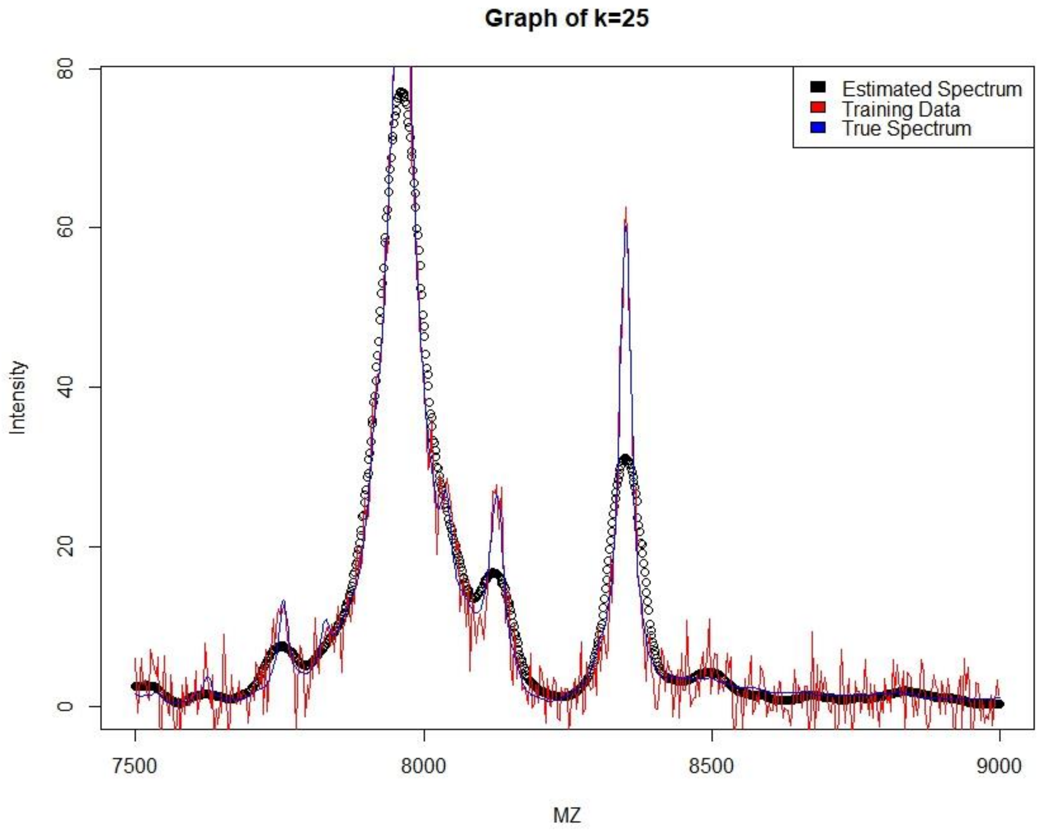
2.

The four graphs following are in the order of k=2, k=5, k=10, k=25.




Graph of k=5
Estimated Spectrum
Training Data
True Spectrum
Intensity
0
20
40
60
80
100
MZ
7500
8000
8500
9000


Graph of k=10
Estimated Spectrum
Training Data
True Spectrum
Intensity
0
20
40
60
80
MZ
7500
8000
8500
9000



3.

By observing the graphs, it's clear that with a small k, our estimates fit the peaks well, but they come with a lot of extra noise. As k moves towards 5, we strike a balance where the peaks are still nicely fitted, and the noise is reduced. But when k goes beyond 5, data points start spreading out more on the peaks in the true spectrum graph.

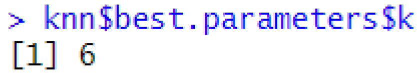
In terms of the numbers, from k = 0 to k = 5, the error goes down, which is good. But after k = 5, the error starts going up again. This happens might be due to the k-nearest neighbors algorithm. With a small k, it doesn't consider enough neighbors and underfits the data. With a big k, it obtains too many neighbors and becomes overfitting. So, by the graph shown above, the best estimates are around 5.

4.

The spectrum plot at k = 5 successfully achieves this goal. In the graph, we can observe that the estimates closely match the true values, and even at the peak points, the estimates remain reasonably close to the true values.

The k-NN (k-nearest neighbors) method can achieve this aim to some extent because it relies on the principle of selecting data points that are closest to the target point when making predictions. When k is set to a small value, such as 5, the method considers only a limited number of nearby neighbors. This results in a more localized and finer-grained estimation that tends to closely match the true values, especially at the peaks in the spectrum.

5.



```
> knn$best.parameters$k
[1] 6
```

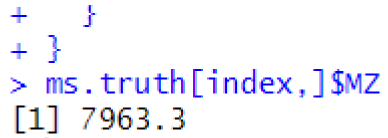
The cross-validation functionality identifies the optimal value of k to be 6, which is close to the value of k that we observed (k=5) that minimizes the actual mean-squared error.

6.

By calculating the standard deviation of minus the estimate of the spectrum produced in Q3.5 by values in ms.measured$intensity,  the estimate of the standard deviation of the sensor/measurement noise that has corrupted our intensity measurements will be 25.82931.

7.

The value of MZ corresponds to the maximum estimated abundance is 7963.3.



```
+     }
+ }
> ms.truth[index,]$MZ
[1] 7963.3
```

8.

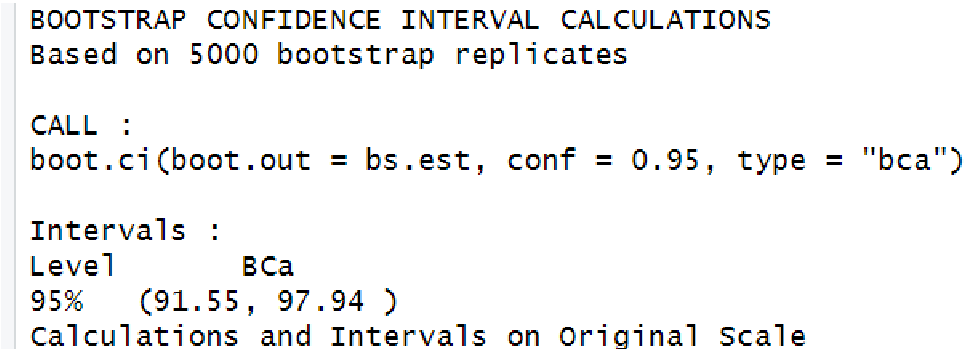
The confidence intervals for the estimated intensity at the MZ value 7963.3 change in size when using different values of k in the k-nearest neighbors method. Specifically:

With k = 6: The 95% confidence interval is (91.55, 97.94).

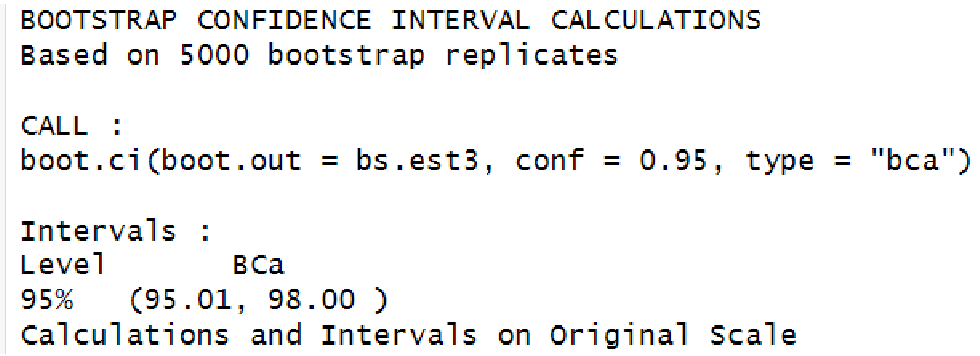
With k = 3: The 95% confidence interval is (95.01, 98.00).

With k = 20: The 95% confidence interval is (69.21, 92.89).

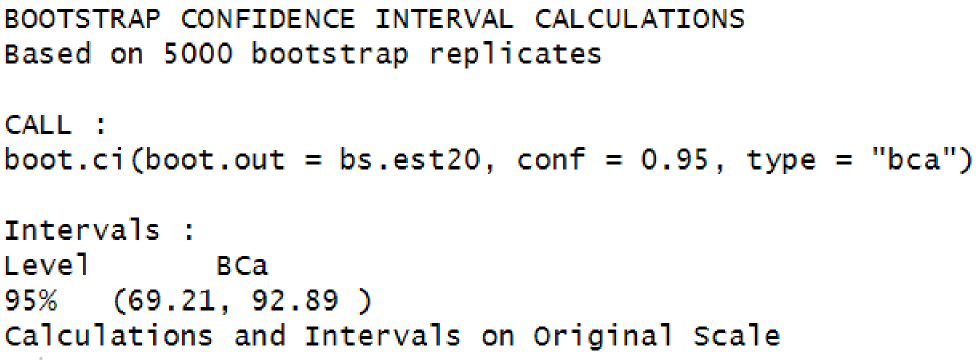
These variations occur because the k-nearest neighbors method behaves differently for different k values. Smaller k values tend to underfit the data, while larger k values tend to overfit. Therefore, the choice of k impacts the width of the confidence intervals.



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

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

Intervals :
Level       BCa
95%   (91.55, 97.94 )
Calculations and Intervals on Original Scale
```



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

CALL : 
boot.ci(boot.out = bs.est3, conf = 0.95, type = "bca")

Intervals : 
Level       BCa          
95%   (95.01, 98.00 )  
Calculations and Intervals on Original Scale
```



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

CALL : 
boot.ci(boot.out = bs.est20, conf = 0.95, type = "bca")

Intervals : 
Level       BCa          
95%   (69.21, 92.89 )  
Calculations and Intervals on Original Scale
```