FIT2086 Assignment 3

Student ID: 34280332
Name: Chee Cheng Mun

Question 1.1
Result of fitting the linear model:

```
Residuals:
     Min       1Q   Median       3Q      Max
-27.3421  -6.5267   0.0905   7.3317  25.0502

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -59.81787   54.46950  -1.098  0.27322
Cement               0.12581    0.01713   7.344 3.18e-12 ***
Blast.Furnace.Slag   0.11267    0.02201   5.120 6.26e-07 ***
Fly.Ash              0.08455    0.02652   3.188  0.00162 **
Water               -0.08794    0.08132  -1.081  0.28059
Superplasticizer     0.54766    0.19477   2.812  0.00533 **
Coarse.Aggregate     0.03398    0.01949   1.744  0.08242 .
Fine.Aggregate       0.02924    0.02171   1.347  0.17929
Age                  0.11052    0.01065  10.379  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.31 on 241 degrees of freedom
Multiple R-squared:  0.6204,    Adjusted R-squared:  0.6078
F-statistic: 49.23 on 8 and 241 DF,  p-value: < 2.2e-16
```

I think the predictor that is possibly associated with compressive strength is: Coarse.Aggregate (p-value = 0.08242).

This is because this predictor has a p-value of less than 1 but more than 0.05, which means it has weak evidence against the null hypothesis that it has no association with compressive strength. However, it is not enough to reject the null hypothesis, thus there is a possibility that there is no association between the two. Hence, this predictor is possibly associated with compressive strength.

These three variables appear to be the strongest predictors of compressive strength:

- Age (p-value < 2e-16 ***)
- Cement (p-value = 3.18e-12 ***)
- Blast.Furnace.Slag (p-value = 6.26e-07 ***)

This is because they have the lowest p-value out of all of the predictors, with a p-value that is smaller than 0.01, they have strong evidence against the null hypothesis that they have no association with compressive strength, and the null hypothesis is rejected.

Question 1.2

Bonferroni procedure is a method used to adjust the significance level considering the circumstances of multiple comparisons.

Bonferroni-corrected Significance Level = $\frac{\alpha}{number\ of\ tests\ performed}$

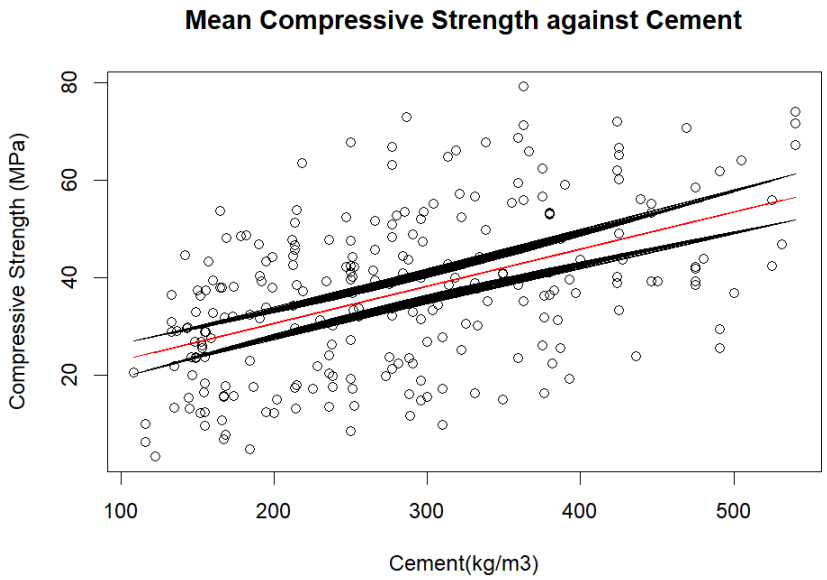
$= \frac{0.05}{8}$

$= 0.00625$

```
> bonferroni_significance_level
[1] 0.00625
```

After performing the correction with Bonferroni procedure, we got the significance level of 0.00625. With a significance level of 0.00625, it means our predictors need to have a p-value that is smaller than 0.00625 to have moderate evidence against the null hypothesis that there is no association between the predictor and compressive strength. Some predictors that were previously associated are now no longer associated, such as Fly.Ash and Superplasticizer.

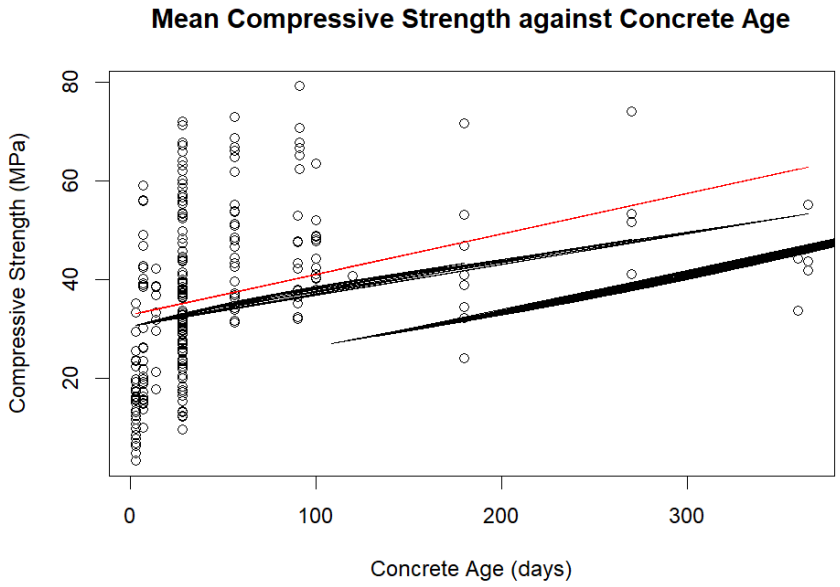
Question 1.3



The red line represents the predicted mean compressive strength. In the graph, it is a positive slope, which means that compressive strength and cement have a positive relationship. As the amount of cement increases, the mean compressive strength increases as well.

The black line represents the lower and upper boundary bounds of the confidence interval showing the range within which the true mean compressive strength is likely to fall according to the amount of cement.

Since the predicted mean compressive strength lies between the two black lines which is the confidence interval, this model's predictions are fairly precise.



The red line represents the predicted mean compressive strength. In the graph, it is a positive slope, which means that compressive strength and age of concrete have a positive relationship. As the age of concrete increases, the mean compressive strength increases as well.

The black line represents the lower and upper boundary bounds of the confidence interval showing the range within which the true mean compressive strength is likely to fall according to the age of concrete.

Since the predicted mean compressive strength lies above the two black lines which is the confidence interval, it is outside of the range of the confidence interval. Therefore, there is some uncertainty in this model's predictions.

Question 1.4

```
Step:  AIC=1202.86
Strength ~ Cement + Blast.Furnace.Slag + Fly.Ash + Superplasticizer +
    Coarse.Aggregate + Fine.Aggregate + Age

                    Df Sum of Sq   RSS    AIC
<none>                           25750 1202.9
- Fine.Aggregate     1    1494.8 27245 1211.4
- Superplasticizer   1    1569.6 27320 1212.1
- Coarse.Aggregate   1    2372.7 28123 1219.4
- Fly.Ash            1    2878.3 28629 1223.8
- Blast.Furnace.Slag 1    6897.9 32648 1256.7
- Age                1   11489.0 37239 1289.6
- Cement             1   13935.9 39686 1305.5
```

Summary:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -113.97780   21.42435  -5.320 2.36e-07 ***
Cement               0.13889    0.01214  11.444  < 2e-16 ***
Blast.Furnace.Slag   0.12899    0.01602   8.051 3.70e-14 ***
Fly.Ash              0.10352    0.01990   5.201 4.22e-07 ***
Superplasticizer     0.65136    0.16959   3.841 0.000157 ***
Coarse.Aggregate     0.05146    0.01090   4.722 3.96e-06 ***
Fine.Aggregate       0.04817    0.01285   3.748 0.000223 ***
Age                  0.11068    0.01065  10.391  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It can be observed that Water has been removed after applying the BIC penalty.

The final regression equation:
E[concrete] = −113.98 + 0.1389 × Cement + 0.1290 × Blast.Furnace.Slag + 0.1035 × Fly.Ash + 0.6514 × Superplasticizer + 0.0515 × Coarse.Aggregate + 0.0482 × Fine.Aggregate + 0.1107 × Age

Question 1.5

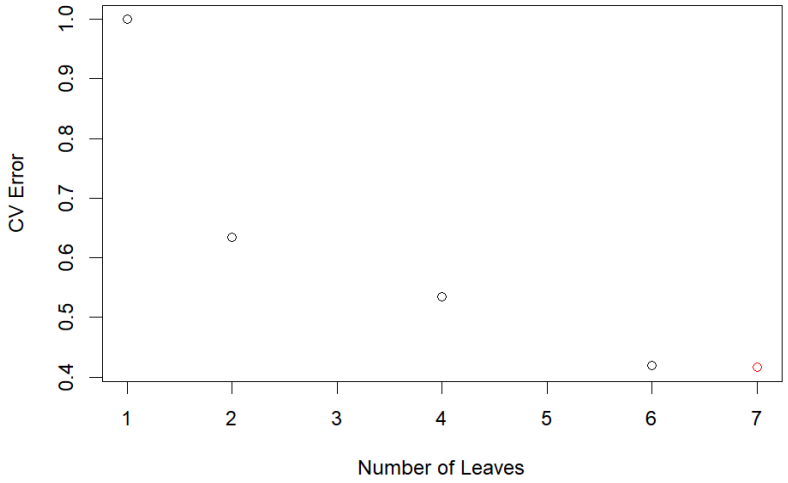
a)

```
> predict(fit.strength.bic, newdata = example_concrete, interval = "confidence")
       fit      lwr      upr
1 54.99446 49.86675 60.12218
```

Mean compressive strength: 54.994MPa
Confidence Interval: [49.867MPa - 60.122MPa]

b) Yes. My model has a mean compressive strength of 54.994MPa, which is higher than 52.35MPa compressive strength of the current mix. Also, my model lies between the confidence interval of [49.867MPa - 60.122MPa], which means my model's prediction is fairly precise. It suggests that the newly proposed mix is better than the current mix.

Question 2.1
Cross-validation score:



Best-tree:

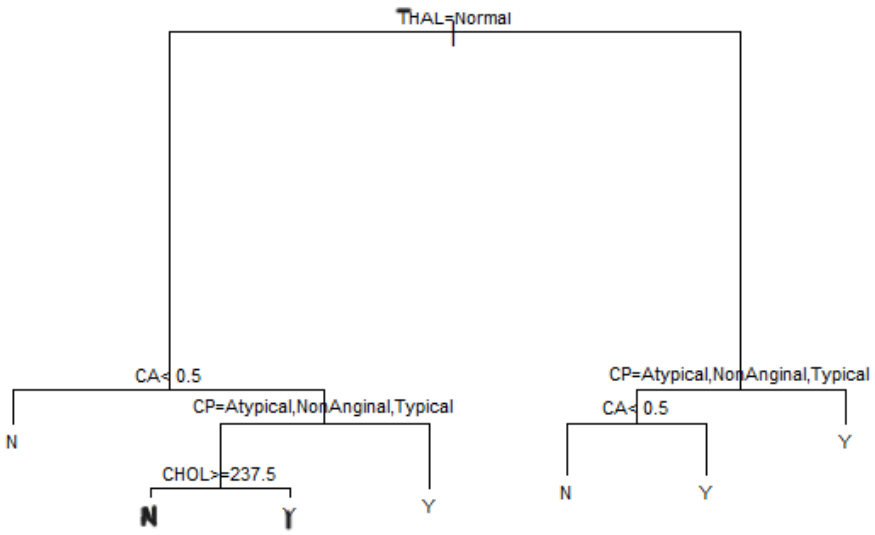
```
1) root 210 97 N (0.53809524 0.46190476)
  2) THAL=Normal 111 23 N (0.79279279 0.20720721)
    4) CA< 0.5 77  6 N (0.92207792 0.07792208) *
    5) CA>=0.5 34 17 N (0.50000000 0.50000000)
     10) CP=Atypical,NonAnginal,Typical 21  6 N (0.71428571 0.28571429)
       20) CHOL>=237.5 14  2 N (0.85714286 0.14285714) *
       21) CHOL< 237.5 7  3 Y (0.42857143 0.57142857) *
     11) CP=Asymptomatic 13  2 Y (0.15384615 0.84615385) *
  3) THAL=Fixed.Defect,Reversible.Defect 99 25 Y (0.25252525 0.74747475)
    6) CP=Atypical,NonAnginal,Typical 31 14 N (0.54838710 0.45161290)
     12) CA< 0.5 16  3 N (0.81250000 0.18750000) *
     13) CA>=0.5 15  4 Y (0.26666667 0.73333333) *
    7) CP=Asymptomatic 68  8 Y (0.11764706 0.88235294) *
```

The variables used in the best tree:

- Thallium scanning results (THAL)
- Number of major vessels colored by flourosopy (CA)
- Chest pain type (CP)
- Serum cholesterol (CHOL)

The best tree has 7 leaves(terminal nodes).

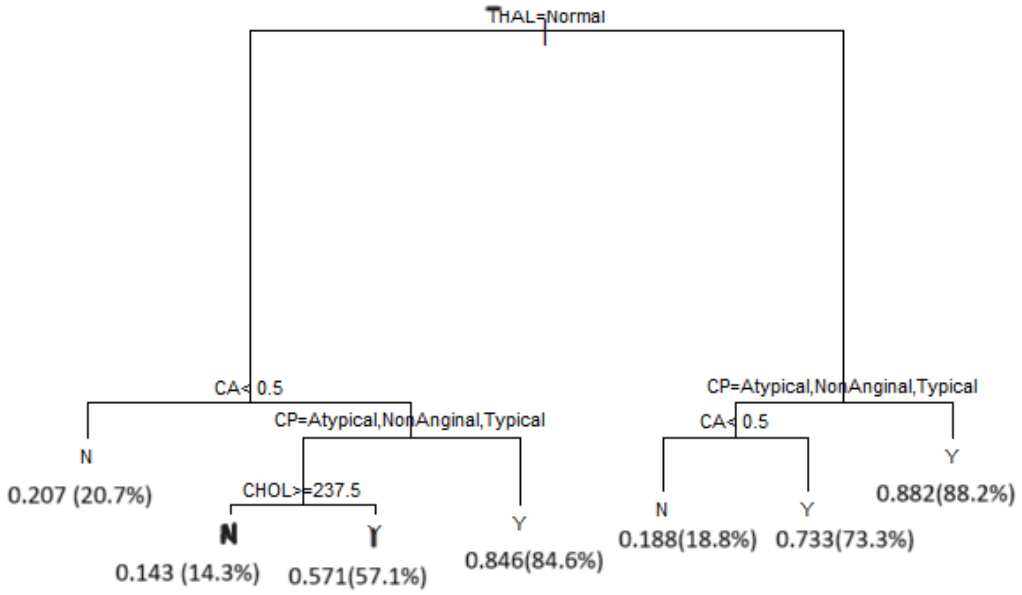
Question 2.2



If a patient has normal Thallium scanning results (THAL), and has a number of major vessels colored by flourosopy (CA) that is less than 0.5, then they are predicted to not have a heart disease. However, if their number of major vessels colored by flourosopy is greater or equal to 0.5, the model will need to check if their chest pain type(CP) is under Atypical angina, Non anginal pain and typical angina. If their chest pain type is not under any of these categories, then they are predicted to have a heart disease. If their chest pain type is really under one of these categories, then the model checks their Serum cholesterol (CHOL) as well. If the patient has a serum cholesterol of greater or equal to 237.5mg/dl, they are predicted to not have a heart disease. If the patient has a serum cholesterol smaller than 237.5mg/dl, then they are predicted to have a heart disease.

If a patient has Thallium scanning results other than normal, such as Fixed fluid transfer defect or Reversible fluid transfer defect and has a chest pain type that is not under Atypical angina, Non anginal pain and typical angina, then they are predicted to have a heart disease. If their chest pain type belongs to one of the categories mentioned, then the model will check their number of major vessels colored by flourosopy. If the number is smaller than 0.5, then the patient is predicted to not have a heart disease. On the contrary, if the number is greater or equal to 0.5, then the patient is predicted to have a heart disease.

Question 2.3



Question 2.4

According to my tree, the combination of Thallium scanning results (THAL) is not normal, which is one of Fixed fluid transfer defect or Reversible fluid transfer defect and a chest pain type(CP) of Asymptomatic pain(Asymptomatic) results in the highest probability of having heart-disease.

Question 2.5

```
glm(formula = HD ~ EXANG + SLOPE + CA + THAL, family = binomial,
    data = heart.train)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.4727     0.9317  -1.581   0.1140
EXANGY                1.9599     0.4744   4.132 3.60e-05 ***
SLOPEFlat             0.3258     0.7531   0.433   0.6653
SLOPEUp              -1.1251     0.7811  -1.440   0.1498
CA                    1.2904     0.2531   5.099 3.41e-07 ***
THALNormal           -0.7100     0.7741  -0.917   0.3591
THALReversible.Defect  1.4099     0.7783   1.811   0.0701 .
```

The final model include these variables:

- Exercise induced angina yes(EXANGY)
- Slope of the peak exercise ST segment is Flat (SLOPEFlat)
- Slope of the peak exercise ST segment is Up-sloping (SLOPEUp)
- Number of major vessels colored by flourosopy (CA)
- Thallium scanning results Normal (THALNormal)
- Thallium scanning results is Reversible fluid transfer defect (THALReversible.Defect)
-

Both models selected THAL (Thallium scanning results) and CA (Number of major vessels colored). These two predictors provide direct diagnostic information.

Logistic regression looks at the strength of the linear relationships between the predictors and the response variable. It chose variables like EXANGY and SLOPE, which have a strong linear relationship with the response variable.

On the other hand, decision trees focus more on splits that maximize classification accuracy even if they are not linear. The tree chose CP and CHOL because they provide key splits to separate patients at different risk levels.

Logistic regression is restricted to linear relationships between predictors and response variables, while decision tree allows non-linear relationships between predictors and response variables.

```
glm(formula = HD ~ EXANG + SLOPE + CA + THAL, family = binomial,
    data = heart.train)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.4727     0.9317  -1.581   0.1140
EXANGY                  1.9599     0.4744   4.132 3.60e-05 ***
SLOPEFlat               0.3258     0.7531   0.433   0.6653
SLOPEUp                -1.1251     0.7811  -1.440   0.1498
CA                      1.2904     0.2531   5.099 3.41e-07 ***
THALNormal             -0.7100     0.7741  -0.917   0.3591
THALReversible.Defect   1.4099     0.7783   1.811   0.0701 .
```

Question 2.6

E[heart] = −1.473 + 1.960 × EXANGY + 0.326 × SLOPEFlat - 1.125 × SLOPEUp
+ 1.290 × CA - 0.710 × THALNormal + 1.410 × THALReversible.Defect

Question 2.7

Logistic regression:

```
Performance statistics:

Confusion matrix:

    target
pred  N  Y
   N 41 10
   Y 10 31

Classification accuracy = 0.7826087
Sensitivity             = 0.7560976
Specificity             = 0.8039216
Area-under-curve        = 0.8469632
Logarithmic loss        = 44.56239
```

Decision Tree:

```
Performance statistics:

Confusion matrix:

    target
pred  N  Y
   N 45 12
   Y  6 29

Classification accuracy = 0.8043478
Sensitivity             = 0.7073171
Specificity             = 0.8823529
Area-under-curve        = 0.8417025
Logarithmic loss        = 44.44189
```

Classification accuracy: The decision tree has a slightly higher accuracy, indicating it performs better in correctly predicting the test data.

Sensitivity: The logistic regression model has better sensitivity, meaning it is better at correctly identifying positive cases (patient with the disease).

Specificity: The decision tree model has higher specificity, making it better at identifying negative cases (patient without the disease).

Area-under-curve(AUC): The logistic regression model has a slightly better AUC. It is better at differentiating positive and negative classes.

Logarithm loss: Both models have similar logarithmic loss, which means that their predicted probabilities are comparable.

To minimize false positives, the decision tree model would be preferable as it has higher specificity.

To minimize false negatives, logistic regression model is a better choice as it has higher sensitivity.

Question 2.8

Odds = $\frac{Probability\ of\ Event}{1 - Probability\ of\ Event}$

Question 2.8.a

```
> tree_odds
[1] 0.08450704
```

Odds: 0.085

Question 2.8.b

```
> logistic_odds
      60
1.108555
```

Odds: 1.109

The odds of the patient having heart disease predicted by the decision tree is lower than the prediction of logistic regression.

So, the decision tree is better at avoiding false positives and the logistic regression is better at avoiding false negatives.

Question 2.9

Confidence Interval of 65th patient:

```
> ci.65
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_65, conf = 0.95, type = "bca")

Intervals :
Level       BCa
95%   ( 0.0000,  0.9978 )
Calculations and Intervals on Original Scale
```

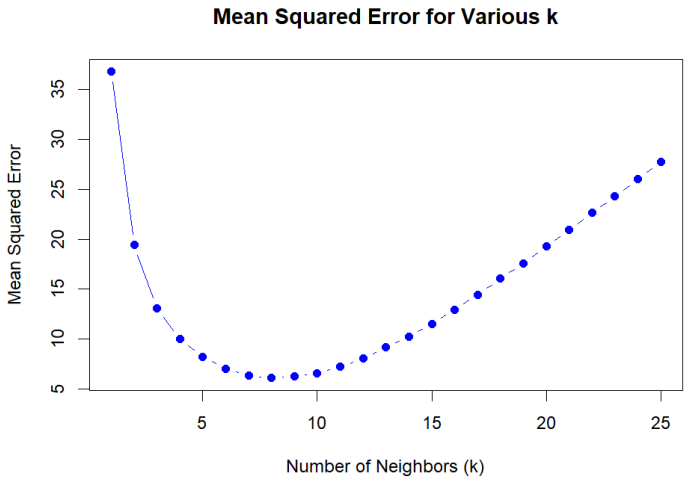
Confidence Interval of 66th patient:

```
> ci.66
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_66, type = "bca")

Intervals :
Level       BCa
95%   ( 0.000,  0.959 )
Calculations and Intervals on Original Scale
```

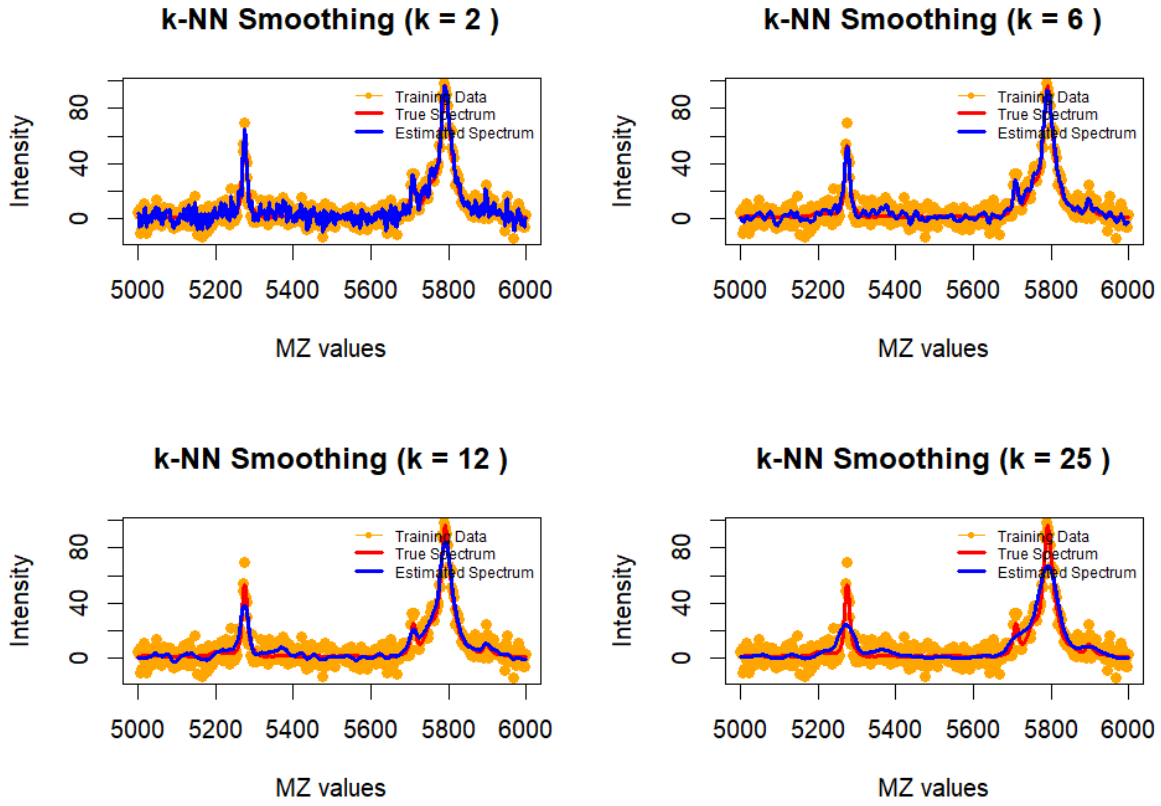
Both of these patients have a wide interval from 0 to nearly 1. Since the intervals mostly overlap and the uncertainty is too high, there is no real difference in the population probability of having heart disease between these two individuals.

Question 3.1

```
> mse_values
 [1] 36.771978 19.443635 13.094895 10.013764  8.267286  7.057205  6.360122  6.129188
 [9]  6.261331  6.621741  7.233900  8.055226  9.193626 10.257507 11.559168 12.971651
[17] 14.466247 16.045799 17.598620 19.262311 20.900865 22.616465 24.281350 26.012212
[25] 27.737285
```



Question 3.2



Question 3.3

At k = 2, the MSE is considered high among the four, likely due to overfitting. The graph shows that the estimated spectrum is even higher than the true spectrum.

At k = 25, the MSE is the highest of the four, likely due to too much smoothing and loss of details. The graph shows that the estimated spectrum is way lower than the true spectrum, which is a form of overfitting.

The MSE is lowest of the four at k = 6. It has achieved the best balance between bias and variance out of the four. The graph shows that the estimated spectrum fit in with the true spectrum with minimal difference.

At k = 12, the MSE is slightly higher than k = 6. It might be losing some of the details due to too much smoothing that causes the rise in MSE. The graph shows that the estimated spectrum is slightly lower than the true spectrum.

Question 3.4

```
> best_k
[1] 6
```

The method selects k = 6. The value of k that would minimise the actual mean-squared error (as computed in Question 3.1) is k = 8.

This might be due to the fact that k = 6 is optimal for generalization while k = 8 captures the underlying trend better when applied to new data.

Question 3.5

```
> estimated_sd
[1] 2.626292
```

I can make use of the residuals from kNN predictions to find an estimate of the standard deviation of the sensor/measurement noise that has corrupted our intensity measurements being 2.626.

Question 3.6

Yes. The spectra with k = 6 achieve our aim of providing a smooth, low-noise estimate of background level as well as accurate estimation of the peaks.
The k-NN method is able to achieve this aim as long as the correct value of k is chosen. An overly low k value will result in overfitting while an overly high k value will result in underfitting.

Question 3.7

```
> max_intensity_value
[1] 93.44637
> max_MZ_value
[1] 5789.8
```

The MZ value of 5789.8 corresponds to the max intensity value of 93.446.

Question 3.8
confidence intervals using the k determined in Question 3.4
K = 6

```
$`6`
[1] 85.49908 97.56324
```

CI of k = 6: [85.499 - 97.563]

K = 3

```
$`3`
[1] 93.21742 98.09339
```

CI of k = 3: [93.217 - 98.093]

K = 20

```
$`20`
[1] 58.44449 88.00543
```

CI of k = 20: [58.444 - 88.005]

These confidence intervals vary in size for different values of k due to various reasons. If the k value is small, it is only affected by the few neighboring values, causing a narrow confidence interval.

If the k value is big, it will lead to a smoother estimate that is affected by many neighbors, this leads to a wide confidence interval which might be too generalized to catch the underlying structure of the data.

If the k value is just right, it will strike a balance in catching the underlying structure of data and generalization. It is more sensitive than large k and more generalized than small k, having a moderate confidence interval.