Modern Applied Statistics exercises from ISLR

Yamuna Dhungana

Use set.seed(202111) when appropriate to make results reproducible.

1. (Modified from 8.1 pg 201 in **Modern Data Science with R**.) The ability to get a good night's sleep is correlated with many positive health outcomes. The `NHANES` data set contains a binary variable `SleepTrouble` that indicates whether each person has trouble sleeping. For each of the listed models - Logistic Regression, Neural network, K - Nearest Neighbors, LDA, and QDA, repeat all of the following steps:

a) Using the Validation Set Approach with a split of 90/10, build a classifier for `SleepTrouble` on the training data. You will have to use a subset of the variables.

b) Report its effectiveness on the test data.

c) Make an appropriate visualization of the model.

d) Interpret the results. What have you learned about people's sleeping habits?

Data Exploration:

The data NHANES had 76 columns with 10000 rows with NA values. Firstly, I have selected variables using the stepwise function for each group of data. The NHANES data has different sub-groups like demographic variables, physical measurement, health variables, lifestyle variables. With each subgroup of variables excluding some variables which does not has compete information, I ran stepwise regression using the 'olsrr' package using the ols_step_forward_p (model) function for each group. The selected variables are the result of the stepwise regression that I chose based on C(p). (I did not include the code but could provide one).

```
## [1] 1854   11
```

Splitting data into test and train

The data is divided in a 9 to 1 ratio.90% of the data is train data and 10% of the data is in the test data. I have applied this ratio to all the classifiers.

```
## [1] "size of Training data"
```

```
## [1] 1669   11
```

```
## [1] "size of Testing data"
```

```
## [1] 185  11
```

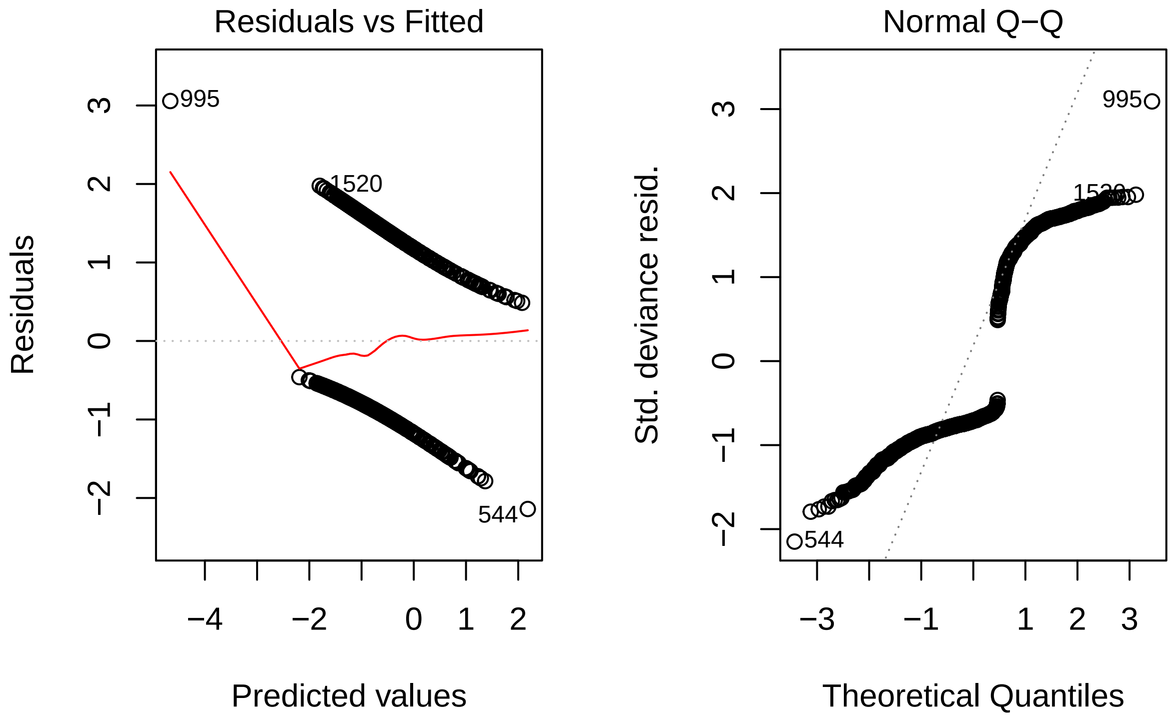
[1] Logistic Regression:

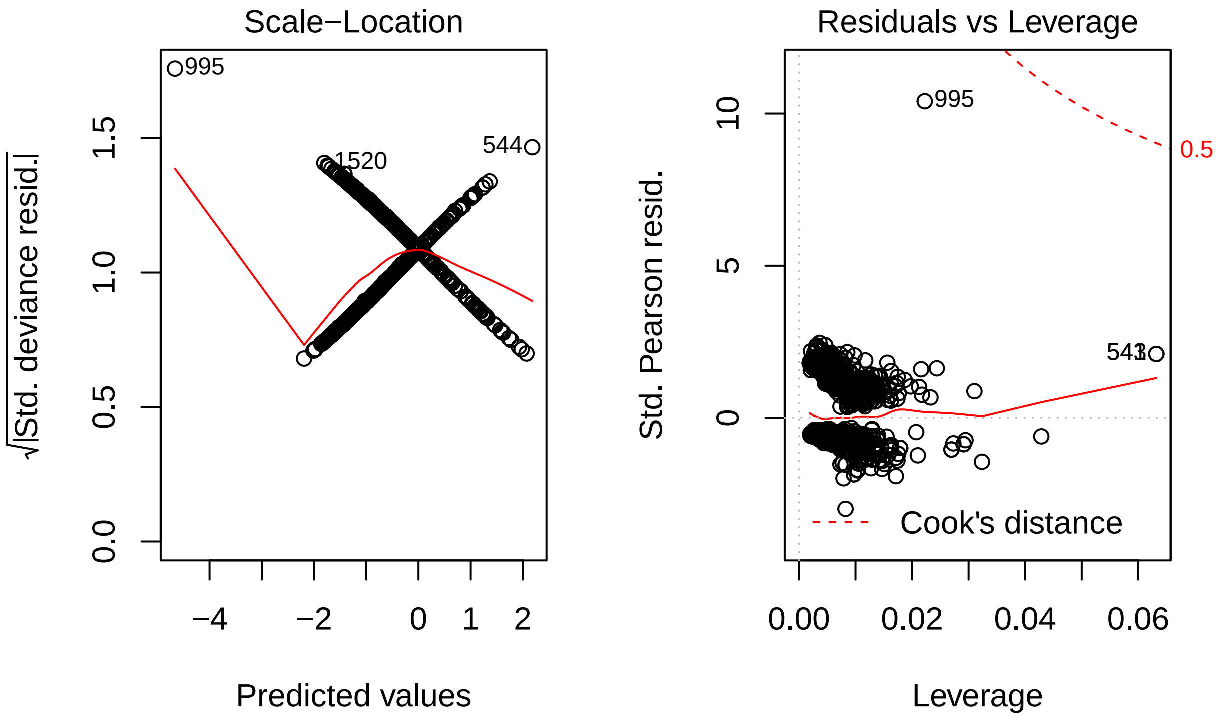
```
##
## Call:
## glm(formula = sleeptrouble ~ ., family = binomial, data = train.orig)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1400  -0.8259  -0.7078   1.1972   3.0566  
## 
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)     -1.279279   0.564080  -2.268 0.023335 *  
## HHIncome        -0.039473   0.018133  -2.177 0.029491 *  
## Age              0.007265   0.004462   1.628 0.103517    
## BMI             -0.009418   0.008935  -1.054 0.291849    
## MaritalStatus    0.029986   0.051496   0.582 0.560361    
## DaysPhysHlthBad  0.054041   0.007062   7.652 1.98e-14 ***
## Depressed        0.452380   0.094774   4.773 1.81e-06 ***
## AlcoholDay      -0.045221   0.019780  -2.286 0.022243 *  
## SmokeNow        -0.320489   0.122611  -2.614 0.008952 ** 
## PhysActive      -0.060607   0.117614  -0.515 0.606339    
## HardDrugs        0.443896   0.117422   3.780 0.000157 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2092.4  on 1668  degrees of freedom
## Residual deviance: 1942.4  on 1658  degrees of freedom
## AIC: 1964.4
## 
## Number of Fisher Scoring iterations: 4
```

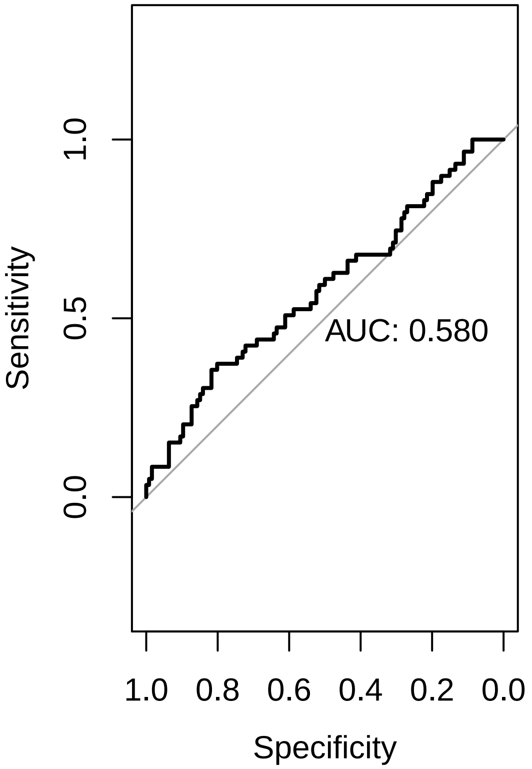
```
##      
## preds   0   1
##     0 117  50
##     1   9   9
## [1] "Model Accuracy (Percentage):"
## [1] 68.11
## [1] "True Positive Rate, TPR (percentage):"
## [1] 15.25
## [1] "False Postive Rate, FPR (percentage):"
## [1] 7.14
```

```
##              Rate
## accuracy 68.10811
## TPR      15.25000
## FPR       7.14000
```

```
## [1] "MSE of the logistic model is 1.804"
```


Residuals vs Fitted
Residuals
Predicted values
3
2
1
0
−1
−2
−4
−2
0
1
2
995
1520
544
Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
3
2
1
0
−1
−2
−3
−1
1
2
3
995
1520
544


Scale−Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
−4
−2
0
1
2
Predicted values
995
1520
544
Residuals vs Leverage
Std. Pearson resid.
0
5
10
0.00
0.02
0.04
0.06
Leverage
995
543
0.5
Cook's distance



The logistic regression shows that among all the variables, the total annual gross income and total alcohol consumed in a day are statistically significant at 0.05. Likewise, physical health was bad in a month, depression and hard drugs were highly significant and smoking was also statistically significant. I have also drawn the confusion matrix. It shows that (117+9) 126 of the 185 data was predicted correctly, and (50+9) 59 of the total data was predicted falsely.

The accuracy of the model was 68.11%, the true positive rate was 15.25% and the false-positive rate was 7.14%. I have also calculated the MSE of the model, which is 1.80.

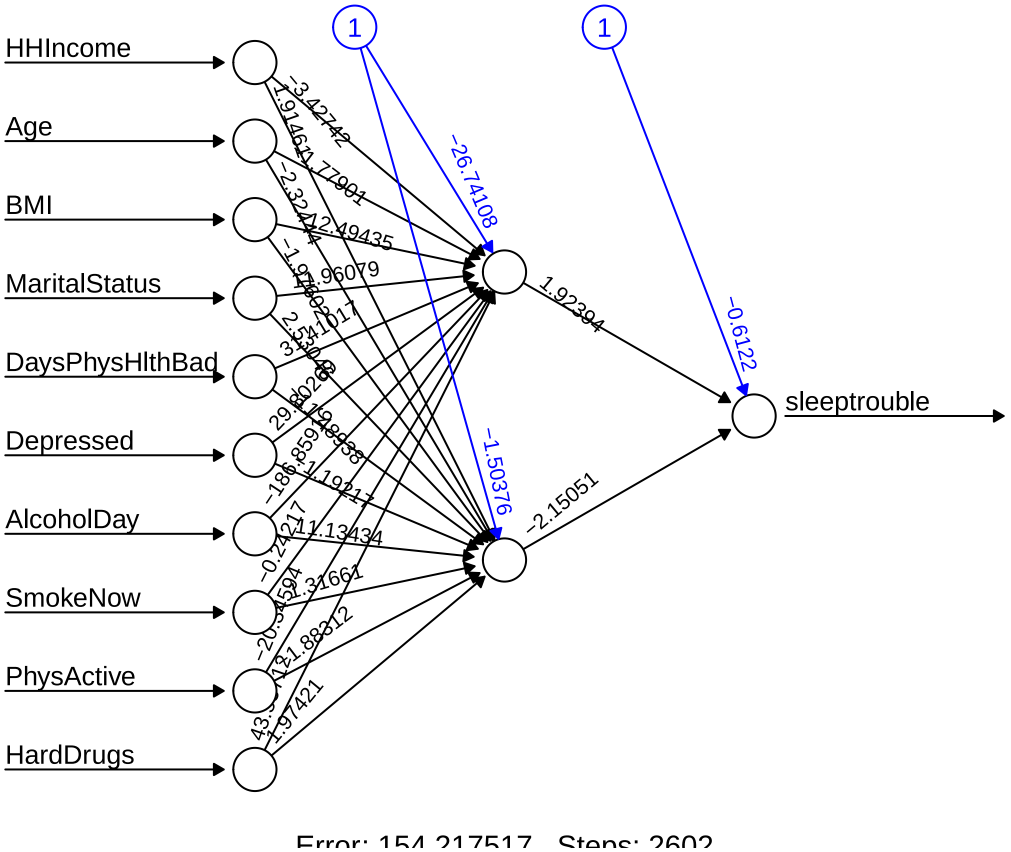
From the analysis of the logistic regression, I came to know that sleep has been affected positively or negatively affected by bad physical health, depression, alcohol consumption, income, hard drugs, and smoking has a significant effect on the sleep of the individual.

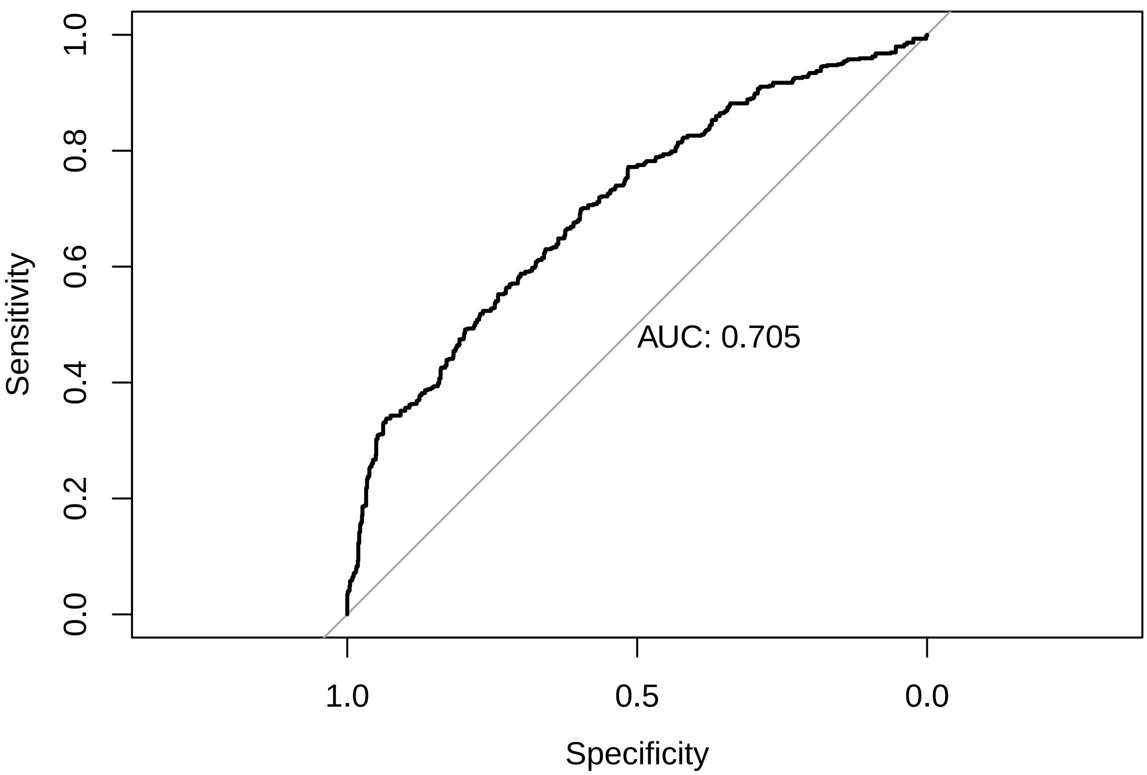
The AUC of the model is 0.580 which is considered bad

[2] Neural network

Here I have fit the neural network with the same data that I chose for the previous model with the package neuralnet. I scaled the data, then split the data into train and the test. I fit the model with the hidden layer of two and to produce the reproducibility I have used set.seed as 202111.

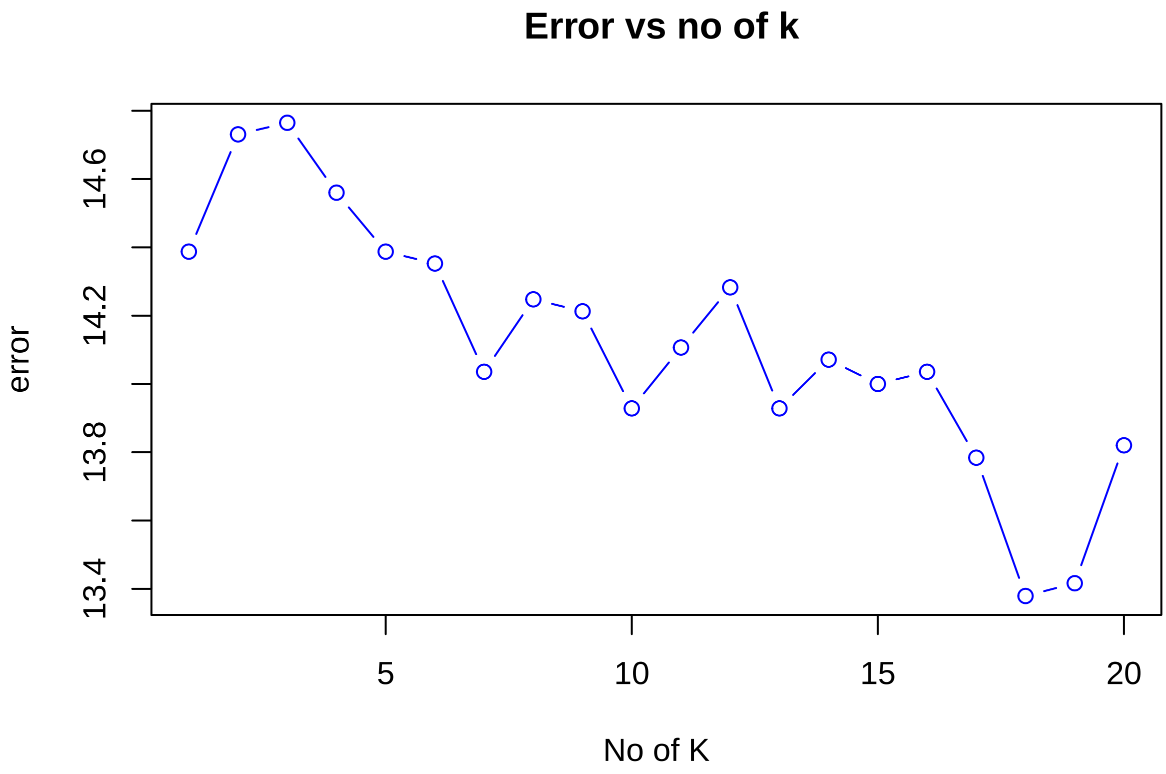
```
## [1] "MSE of the Neural network is 0.187"
```


HHIncome
Age
BMI
MaritalStatus
DaysPhysHlthBad
Depressed
AlcoholDay
SmokeNow
PhysActive
HardDrugs
1
1
−26.74108
−1.50376
−0.6122
1.92394
−2.15051
sleeptrouble
−3.42742
1.91461
1.77901
−2.32444
12.49435
1.96079
2.41017
29.80226
1.19217
11.13434
−0.24217
1.31661
−20.34594
1.88312
1.97421
Error: 154.217517 Steps: 2602



The MSE of the neural network is 0.187. I have further plotted the ROC curve, which determines how well our classifier worked. From the ROC curve, we see that the AUC of the model is 0.705, and is the highest of all the models built. However, we can say that the model has also predicted quite a large FPR. That is why we still have an AUC score that is not quite better. The figure of the neural network showed that the model is the forward propagation, and has 2 hidden layers. The model that we fit has an MSE of 0.18, which is less and is more appropriate. Also, the predictors that we fit contribute to sleep trouble.

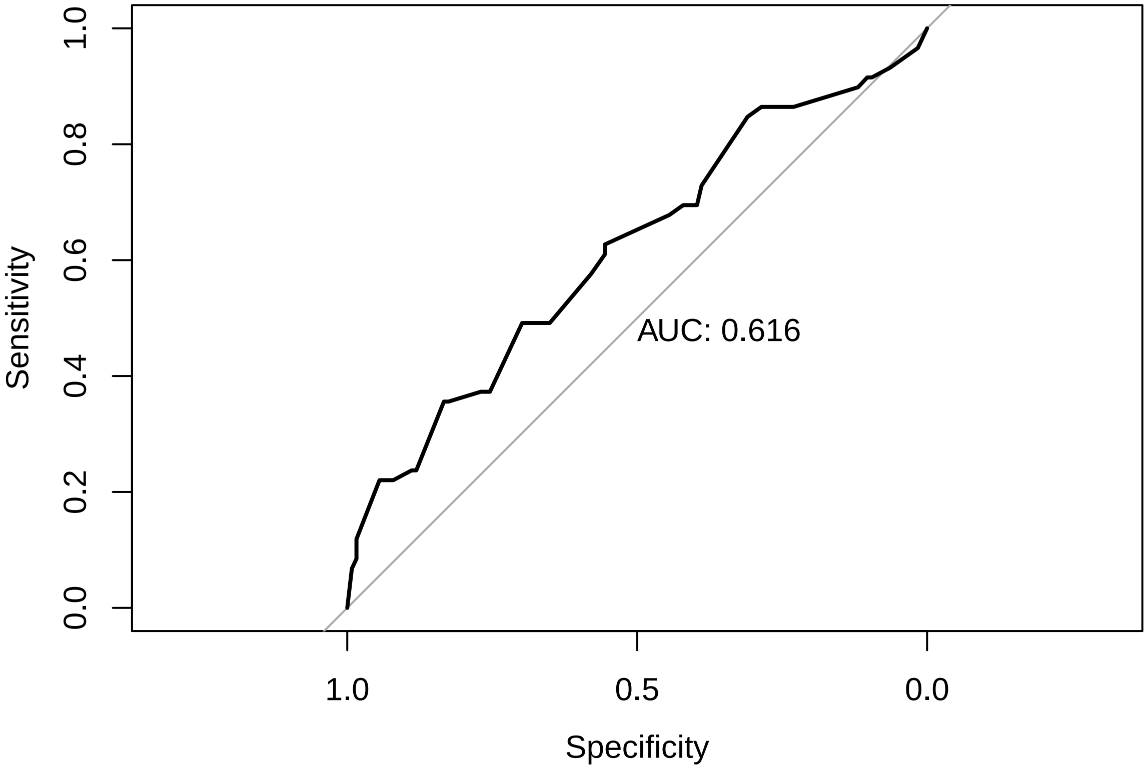
[3] K - Nearest Neighbors



```
##      trues
## model   0   1
##     0 113  44
##     1  13  15
## [1] "Model Accuracy (Percentage):"
## [1] 69.19
## [1] "True Positive Rate, TPR (percentage):"
## [1] 25.42
## [1] "False Postive Rate, FPR (percentage):"
## [1] 10.32
```

```
##              Rate
## accuracy 69.18919
## TPR      25.42000
## FPR      10.32000
```

```
## [1] "MSE of the KNN is 1.032"
```



The KNN model shows the probability of sleep trouble in the data. To fit the KNN model, I have used cross-validation to choose the optimal value of K for the best fit. I have scanned the error for the value of 1 to 20. The model shows the error is minimum when the k is equal to 18 and the error tends to increase beyond 18.

I have also drawn the confusion matrix. It shows that (113+15) 128 of the 185 data was predicted correctly, and (44+13) 57 of the total data was predicted falsely. The accuracy of the model was 69.19%, the true positive rate was 25.42% and the false positive rate was 10.32%. I have also calculated the MSE of the model, which is 1.032.

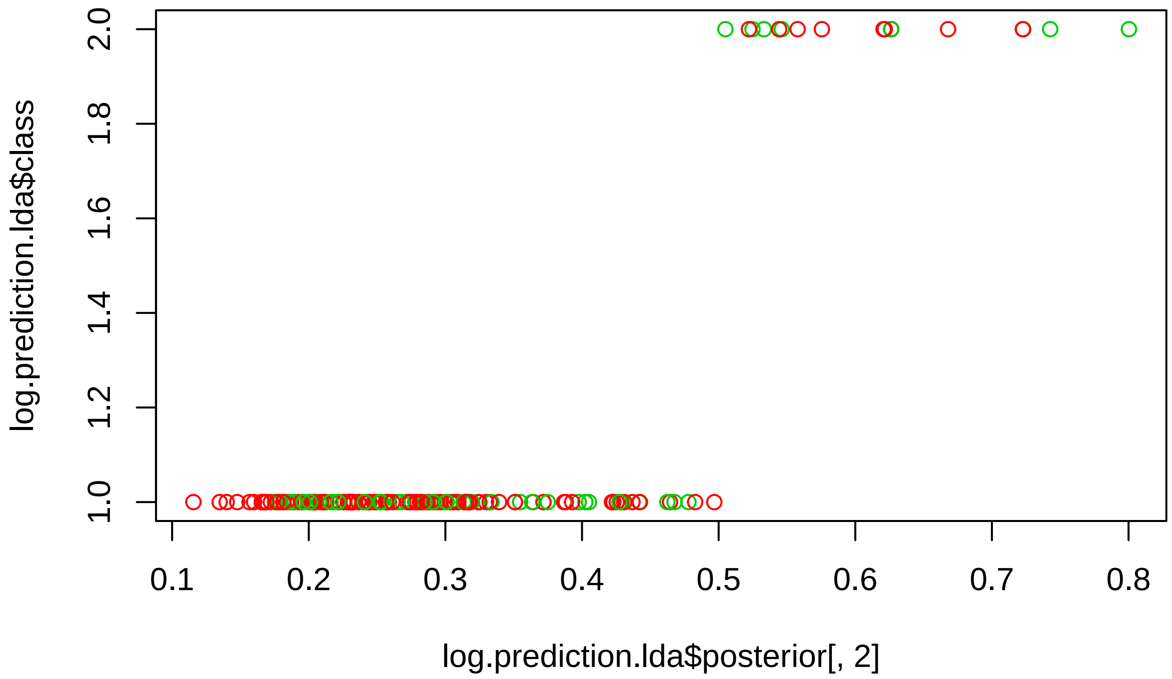
From the analysis of the KNN, I came to know that sleep has been affected positively or negatively affected by bad physical health, depression, alcohol consumption, income, hard drugs, and smoking has a significant effect on the sleep of the individual.

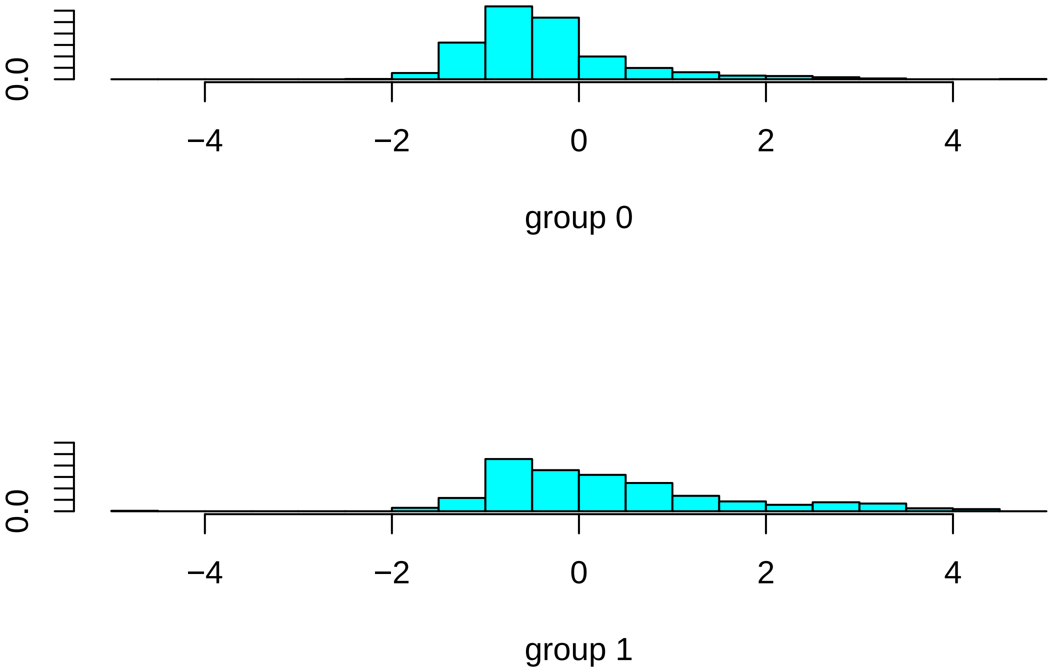
The AUC of the model is 0.616, which is comparatively better.

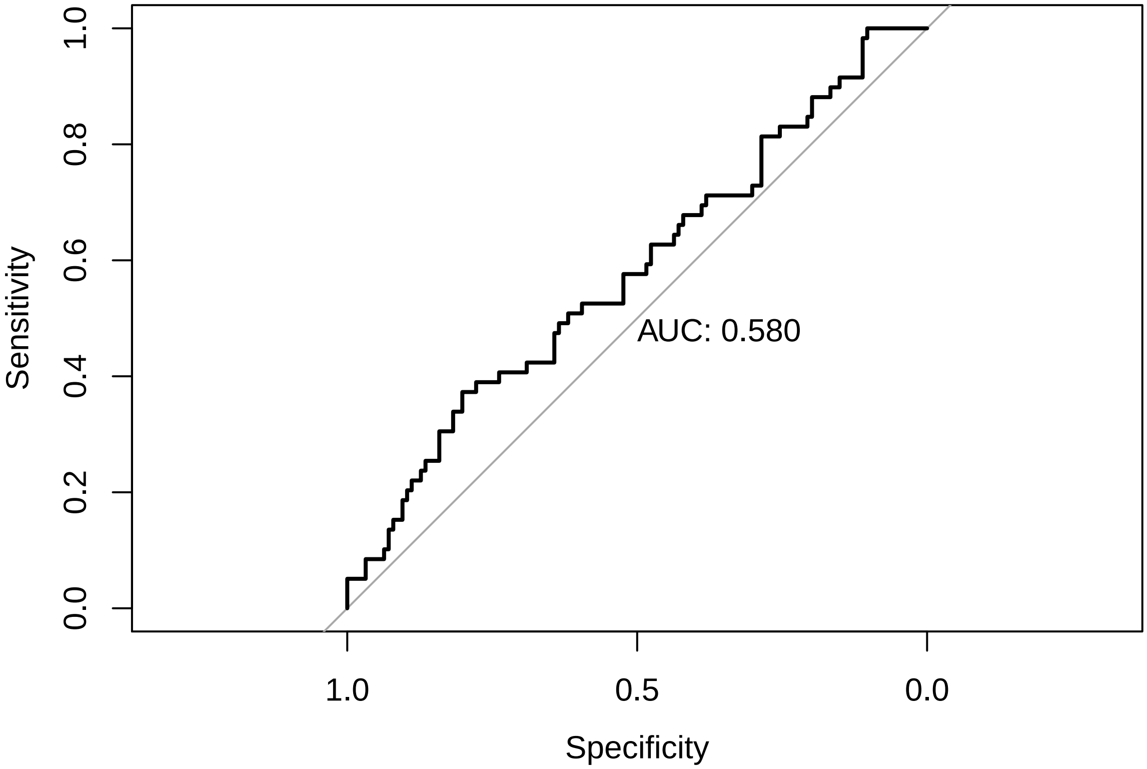
[4] Linear discriminant analysis (LDA)

```
##
## preds   0   1
##     0 116  50
##     1  10   9
## [1] "Model Accuracy (Percentage):"
## [1] 67.57
## [1] "True Positive Rate, TPR (percentage):"
## [1] 15.25
## [1] "False Postive Rate, FPR (percentage):"
## [1] 7.94
##                Rate
## accuracy 67.56757
## TPR      15.25000
## FPR       7.94000
```

```
## [1] "MSE of the LDA model is 1.056"
```




0.0
−4
−2
0
2
4
group 0
0.0
−4
−2
0
2
4
group 1



The LDA model was fitted with the variables and a confusion matrix was also drawn. It shows that (116+9) 125 of the 185 data was predicted correctly, and (50+10) 60 of the total data was predicted falsely.

The accuracy of the model was 67.57%, the true positive rate was 15.25% and the false-positive rate was 7.94%. I have also calculated the MSE of the model, which is 1.056.

I have also plotted the predicted data to show the amount of data that was predicted with the probability. Also, bar chat shows the amount of class predicted.

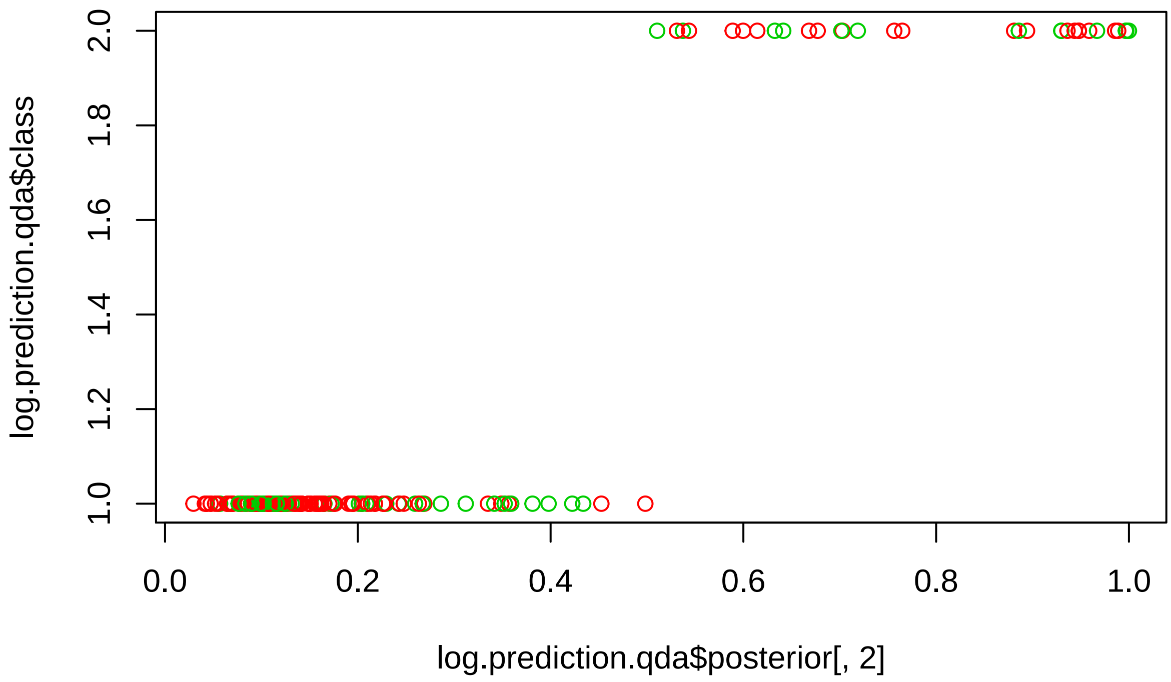
From the analysis of LDA, I came to know that sleep has been affected positively or negatively affected by bad physical health, depression, alcohol consumption, income, hard drugs, and smoking has a significant effect on the sleep of the individual. However, the classifier has not been able to predict correctly. This has been shown by the AUC of the model is 0.580, which is considered bad.

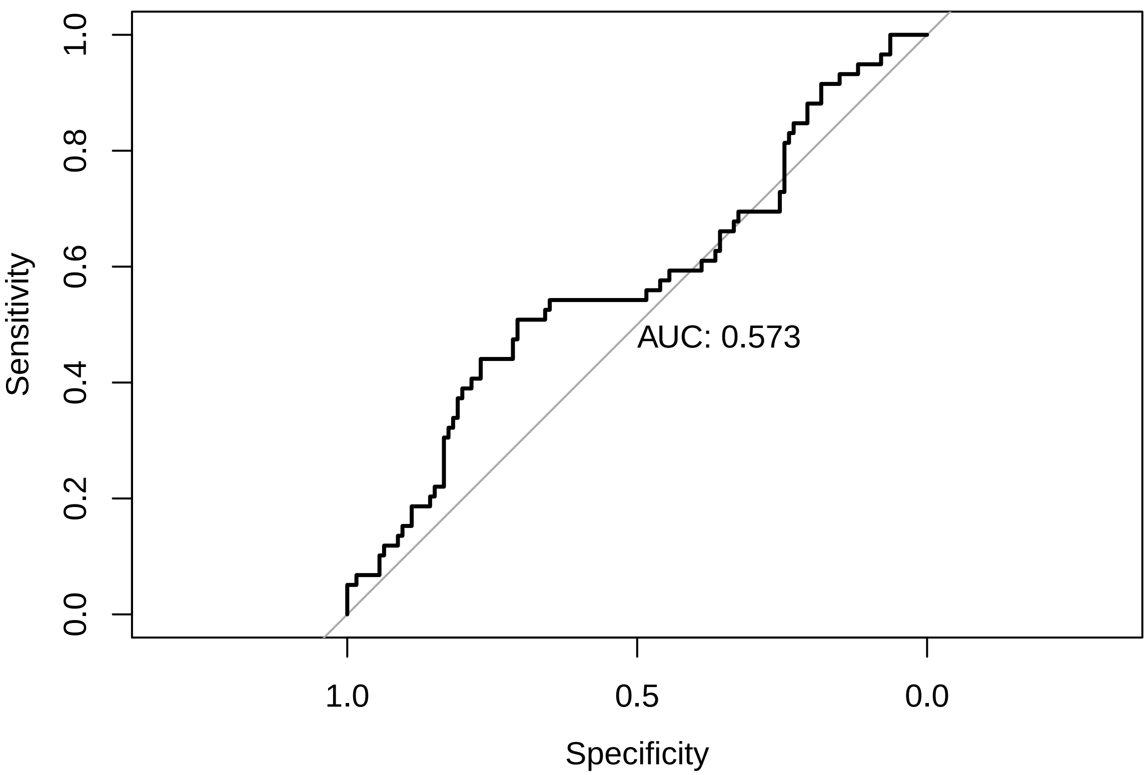
[5] Quadratic discriminant analysis (QDA)

```
##
## preds   0   1
##     0 107  46
##     1  19  13
## [1] "Model Accuracy (Percentage):"
## [1] 64.86
## [1] "True Positive Rate, TPR (percentage):"
## [1] 22.03
## [1] "False Postive Rate, FPR (percentage):"
## [1] 15.08
```

```
##              Rate
## accuracy 64.86486
## TPR      22.03000
## FPR      15.08000
```

```
## [1] "MSE of the QDA model is 0.258"
```





The QDA model was fitted with the variables and a confusion matrix was also drawn. It shows that (107+13) 120 of the 185 data was predicted correctly, and (19+46) 65 of the total data was predicted falsely.

The accuracy of the model was 64.86%, the true positive rate was 22.03% and the false-positive rate was 15.08%. I have also calculated the MSE of the model, which is 0.258.

I have also plotted the predicted data to show the amount of data that was predicted with the probability.

From the analysis of LDA, I came to know that sleep has been affected positively or negatively affected by bad physical health, depression, alcohol consumption, income, hard drugs, and smoking has a significant effect on the sleep of the individual. However, the classifier has not been able to predict correctly. This has been shown by the AUC of the model is 0.573, which is considered bad and is comparatively less than other models.

2) What classifier do you recommend from Exercise 1 and why?

Table 1: MSE of all the classifier

	ERROR
MSE.log	1.804
MSE.nn	0.187
MSE.knn	1.032
MSE.lda	1.056
MSE.qda	0.258

Table 2: Test accuracy of all the classifier

	LogReg	KNN	LDA	QDA
accuracy	68.108	69.189	67.568	64.865
TPR	15.250	25.420	15.250	22.030
FPR	7.140	10.320	7.940	15.080

To determine which model is the best, I calculated the MSE, FTR, TPR, and ROC curve, which shows the model's performance.The MSE shows that the neural network and QDA respectively have the least MSE. I do not recommend QDA has the highest FPR for the model. AUC of the logistic regression = 0.580, Neural network = 0.187, KNN = 0.616, LDA = 0.580 and QDA = 0.573. When AUC is greater than 0.5, it is considered good; however, lying on the margin is considered bad. As a result of the AUC, the neural network appears to be a good classifier. Because the neural network has the lowest MSE and highest AUC, I would recommend it as the classifier for my data.