Investigating the relationship between sleep quality and various health outcomes

Yamuna Dhungana

Analyzing the NHANES dataset to investigate the relationship between sleep quality and various health outcomes involves implementing different classification algorithms, including Logistic Regression, Neural Network, K-Nearest Neighbors (K-NN), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA). The goal is to determine how each model performs in predicting the binary variable SleepTrouble.

a) For each model, we will follow these steps:

Data Split: Divide the dataset into training (90%) and validation (10%) sets.

Feature Selection: Choose a subset of relevant variables to build the classifier for SleepTrouble on the training data.

Model Training: Implement the selected classification algorithm on the training data.

b) After building each model, evaluate its performance on the test data. Assess metrics such as accuracy, precision, recall, and F1-score to understand how effectively it predicts sleep trouble.

c) Create an appropriate visualization of the model. This could include visualizing decision boundaries, confusion matrices, or other relevant plots to aid in understanding the model's behavior.

d) Interpret the results to gain insights into people's sleeping habits. This might involve analyzing which features are most influential in predicting sleep trouble, the strengths and weaknesses of each model, and any patterns or correlations that emerge from the analysis. These insights can help us better understand the factors affecting sleep quality in the population studied.

Data Exploration:

The NHANES dataset comprises 76 columns and 10,000 rows, featuring some missing values (NA). To prepare the data for analysis, I adopted a stepwise variable selection approach for distinct subgroups within NHANES, which encompassed demographic, physical measurement, health, and lifestyle variables.For each subgroup, I executed stepwise regression using the 'olsrr' package, employing the 'ols_step_forward_p' function to iteratively include variables in the model. Variables were selected based on their influence, as measured by C(p).

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

Splitting data into test and train

The data was divided into a 9-to-1 ratio, with 90% of the data allocated to the training set and 10% to the test set. This consistent split ratio was applied to all the classifiers used in the analysis. This ensures a uniform approach in assessing model performance and allows for a fair comparison of the classifiers across the same data subsets.

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

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

```
## [1] "size of Testing data"
## [1] 185  11
```

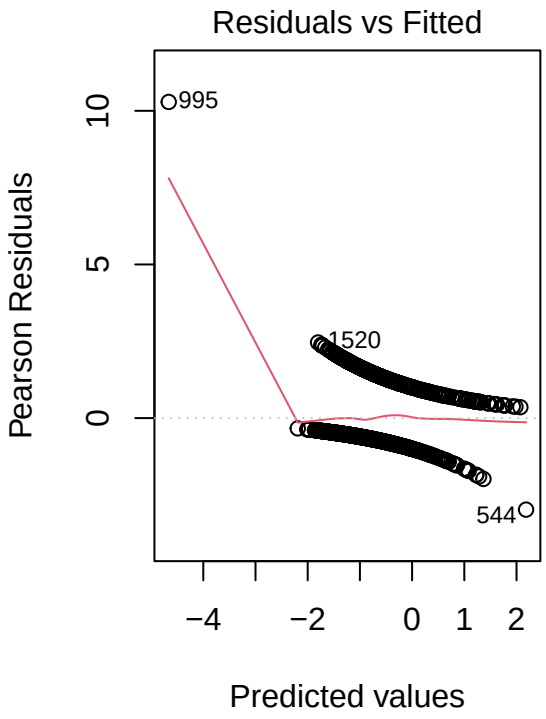
[1] Logistic Regression:

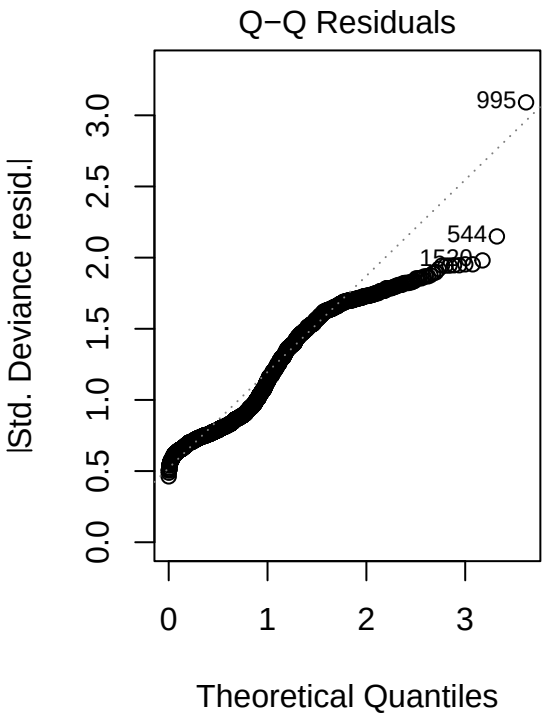
```
##
## Call:
## glm(formula = sleeptrouble ~ ., family = binomial, data = train.orig)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.279279   0.564080  -2.268 0.023335 *
## HHIncome        -0.039473   0.018133  -2.177 0.029491 *
## Age              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
```

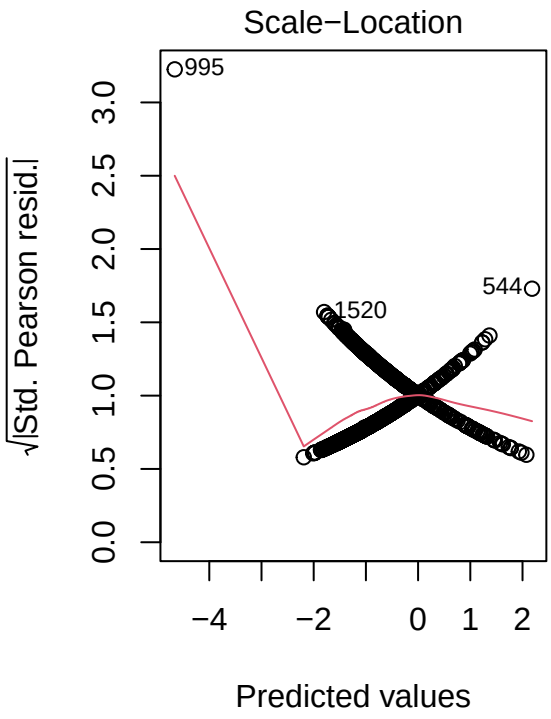
```
##
## preds   0   1
##     0 117  50
##     1   9   9
## [1] "Model Accuracy (Percentage):"
## [1] 68.11
## [1] "True Positive Rate, TPR (percentage):"
## [1] 15.25
## [1] "False Postive Rate, FPR (percentage):"
## [1] 7.14
```

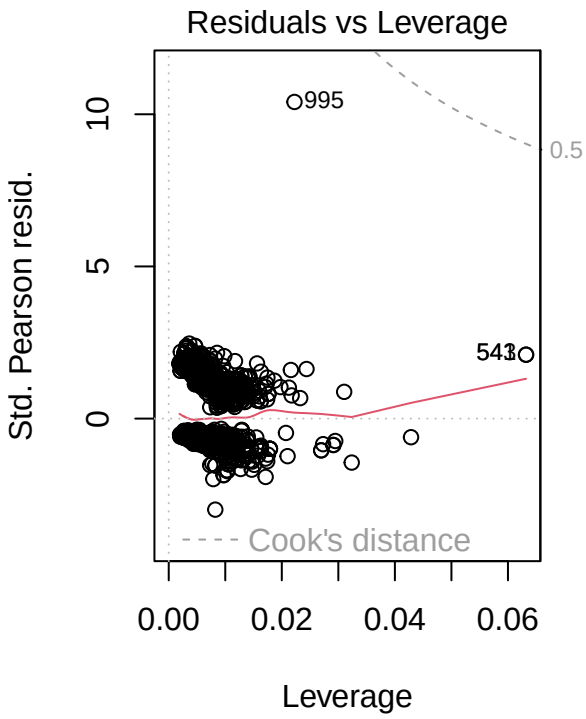
```
##              Rate
## accuracy 68.10811
## TPR      15.25000
## FPR       7.14000
```

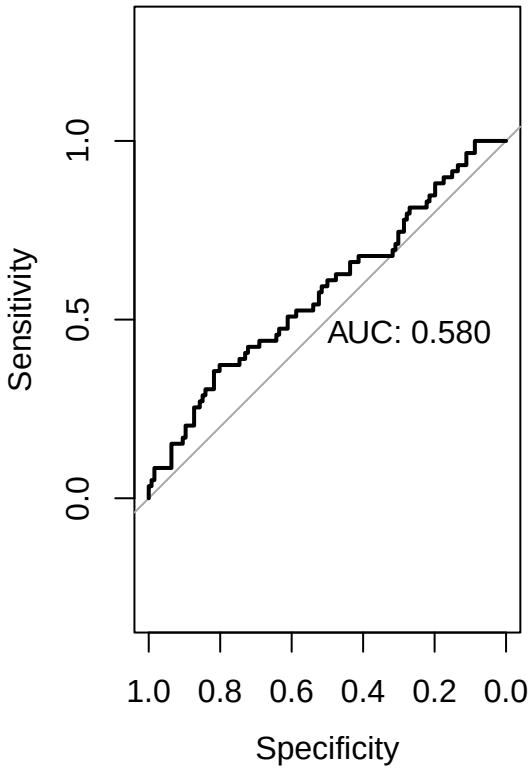
```
## [1] "MSE of the logistic model is 1.804"
```


Residuals vs Fitted
Pearson Residuals
Predicted values
995
1520
544
10
5
0
−4
−2
0
1
2


Q−Q Residuals
|Std. Deviance resid.|
Theoretical Quantiles
995
544
1520
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3


Scale−Location
995
1520
544
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−4
−2
0
1
2
Predicted values


Residuals vs Leverage
995
0.5
Std. Pearson resid.
10
5
0
Cook's distance
0.00
0.02
0.04
0.06
Leverage



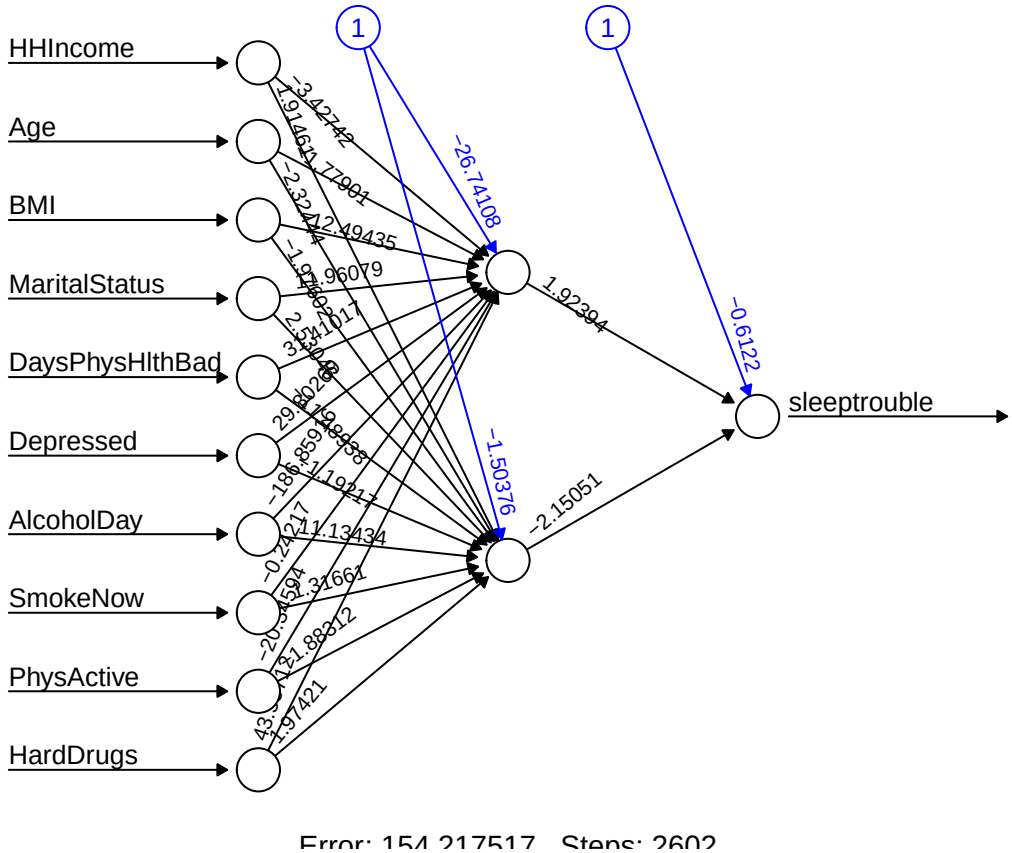
The logistic regression analysis revealed the statistical significance of various variables in predicting the target outcome. Specifically, the total annual gross income and the total daily alcohol consumption were found to be statistically significant at the 0.05 significance level. Additionally, variables related to physical health in the past month, depression, hard drug use, and smoking also exhibited statistical significance in predicting the outcome.Moreover, the confusion matrix provides a snapshot of the model's performance. Out of a total of 185 data points, the model correctly predicted 126 of them (117 true positives and 9 true negatives). However, it also made incorrect predictions for 59 data points (50 false positives and 9 false negatives). This information is vital in assessing the classifier's overall accuracy and understanding its strengths and weaknesses in making predictions.

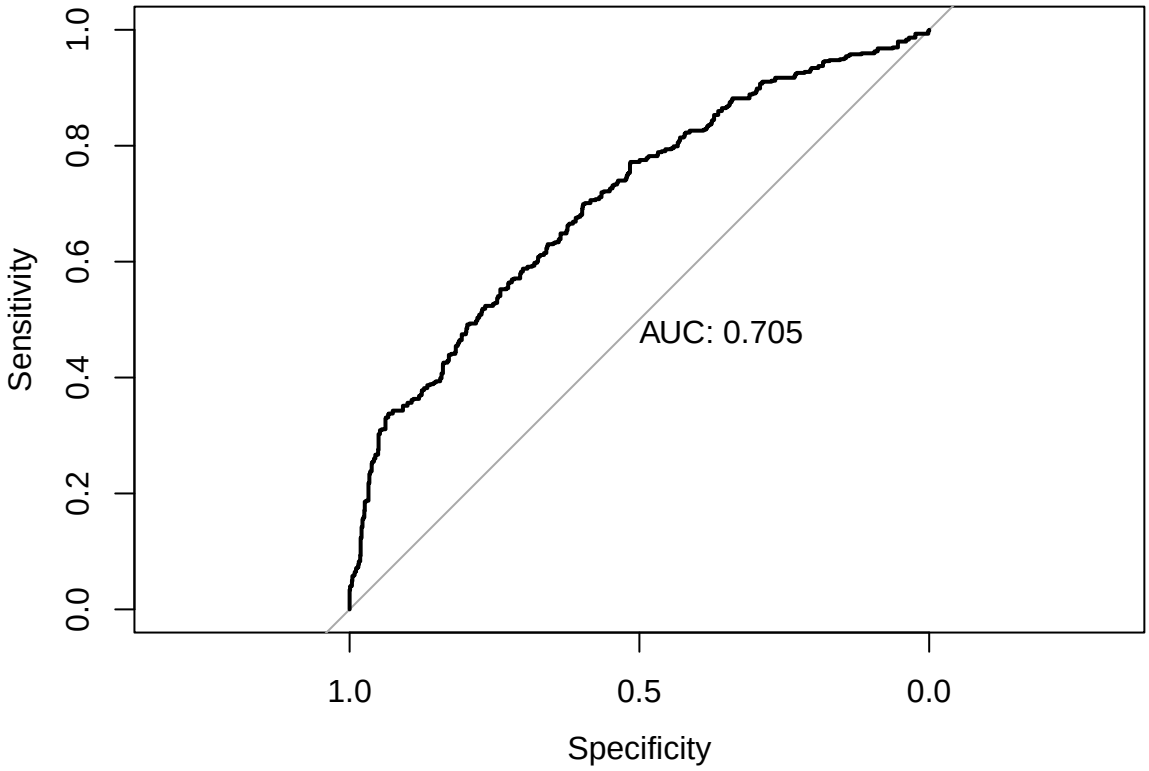
The model exhibited an accuracy of 68.11%, indicating that it correctly predicted outcomes for a substantial portion of the data. The true positive rate, which represents the proportion of actual positive cases correctly identified, was 15.25%. On the other hand, the false-positive rate, which indicates the proportion of actual negative cases incorrectly identified as positive, was 7.14%. Additionally, the mean squared error (MSE) of the model was calculated to be 1.80, providing insights into the overall model performance in terms of the square differences between predicted and actual values.The logistic regression analysis revealed that several factors significantly influence sleep. These factors include physical health, depression, alcohol consumption, income, hard drug use, and smoking. These findings indicate that these variables play a meaningful role in affecting an individual's sleep, whether positively or negatively, and can be used to make predictions regarding sleep quality.

[2] Neural network

In this context, I trained a neural network using the dataset previously selected for the earlier model. To prepare the data for modeling, I performed data scaling and then divided it into separate training and test sets. The neural network was configured with a hidden layer consisting of two neurons. To ensure reproducibility of results, I set the random seed to a specific value, 202111.

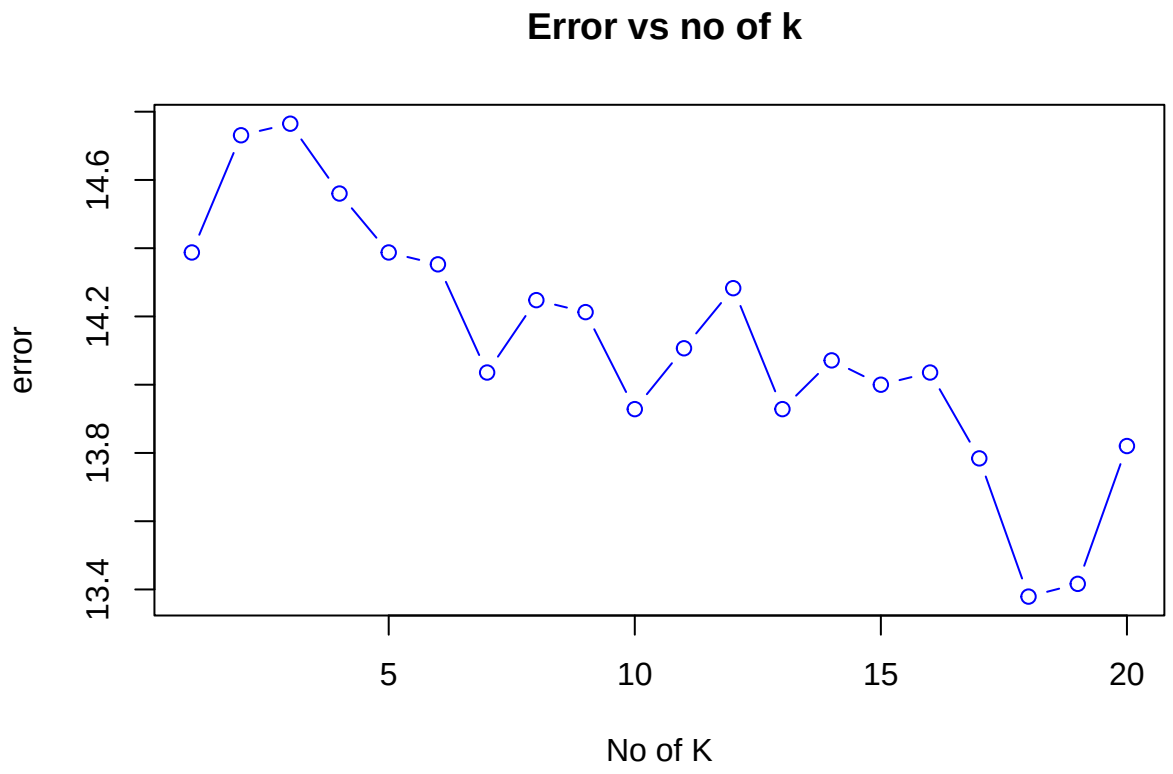
```
## [1] "MSE of the Neural network is 0.187"
```





The neural network's mean squared error (MSE) stands at 0.187. Additionally, an ROC curve was plotted to assess the classifier's performance. The ROC curve reveals that the model achieved an AUC (Area Under the Curve) score of 0.705, which is the highest among all the models built. However, it's important to note that the model has a relatively high false positive rate (FPR), contributing to an AUC score that, while good, could be improved.In the visual representation of the neural network, we observe forward propagation with two hidden layers. This particular model boasts a lower MSE of 0.18, indicating improved performance and suitability. Furthermore, the predictors used in this model make valuable contributions to predicting sleep trouble.

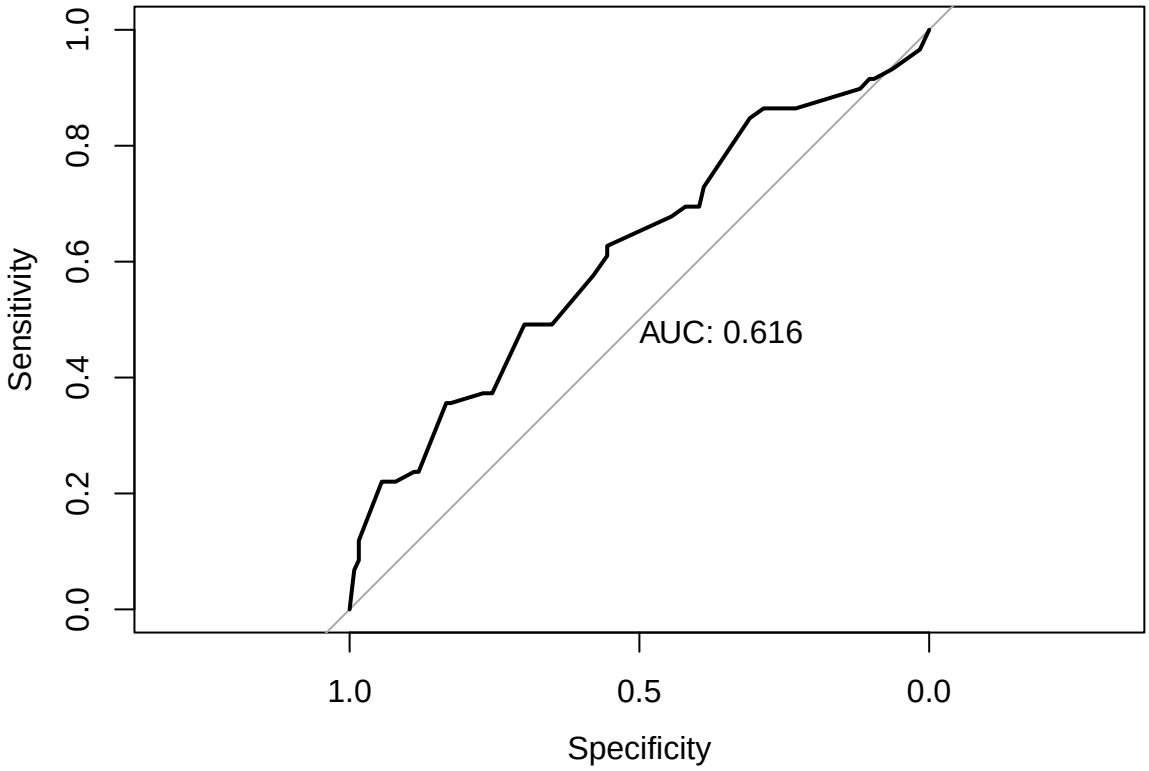
[3] K - Nearest Neighbors




```
##      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 K-Nearest Neighbors (KNN) model provides an estimate of the likelihood of sleep trouble within the dataset. To construct the KNN model, I employed cross-validation to determine the most suitable value of K for optimal performance. The range of K values examined extended from 1 to 20. Notably, the model exhibited minimal error when K equaled 18, with errors tending to rise as K exceeded this threshold.

Additionally, a confusion matrix was generated, revealing that out of the total 185 data points, 128 were correctly predicted (113 true positives and 15 true negatives), while 57 data points were predicted incorrectly (44 false positives and 13 false negatives).The model's accuracy was 69.19%, indicating the proportion of correct predictions among all predictions. The true positive rate, representing the correctly predicted positive cases, stood at 25.42%, while the false positive rate, indicating the proportion of negative cases incorrectly predicted as positive, was 10.32%.Furthermore, the mean squared error (MSE) for the model was calculated to be 1.032, offering insights into the model's overall predictive accuracy.

Based on the K-Nearest Neighbors (KNN) analysis, it became evident that sleep quality is influenced, whether positively or negatively, by factors such as poor physical health, depression, alcohol consumption, income, hard drug use, and smoking. These variables were identified as having a significant impact on an individual's sleep patterns.

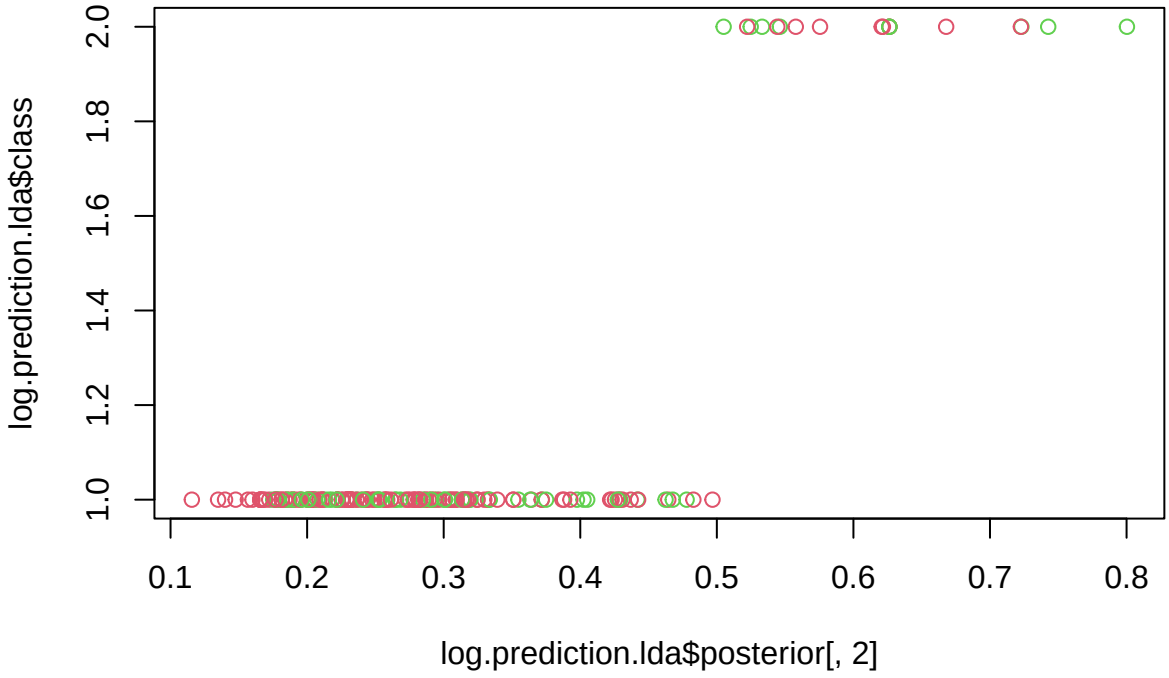
[4] Linear discriminant analysis (LDA)

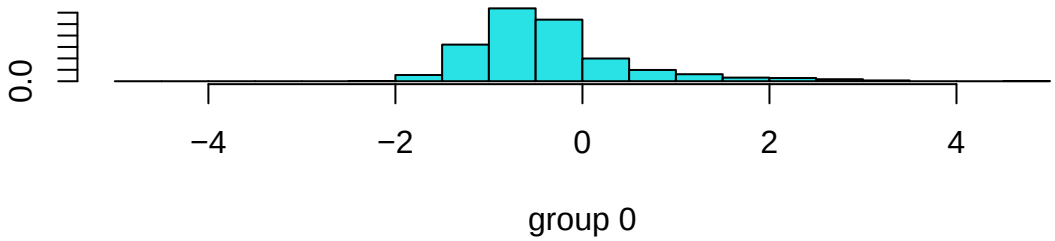
```
##
## preds   0   1
##     0 116  50
##     1  10   9
## [1] "Model Accuracy (Percentage):"
## [1] 67.57
## [1] "True Positive Rate, TPR (percentage):"
```

```
## [1] 15.25
## [1] "False Postive Rate, FPR (percentage):"
## [1] 7.94
```

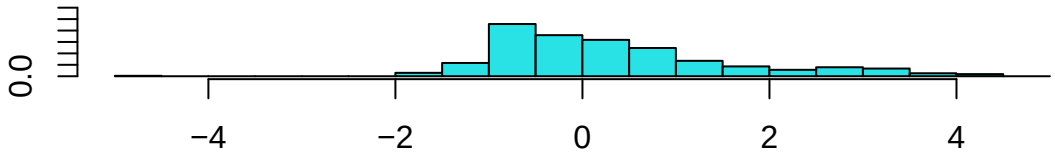
```
##              Rate
## accuracy 67.56757
## TPR      15.25000
## FPR       7.94000
```

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

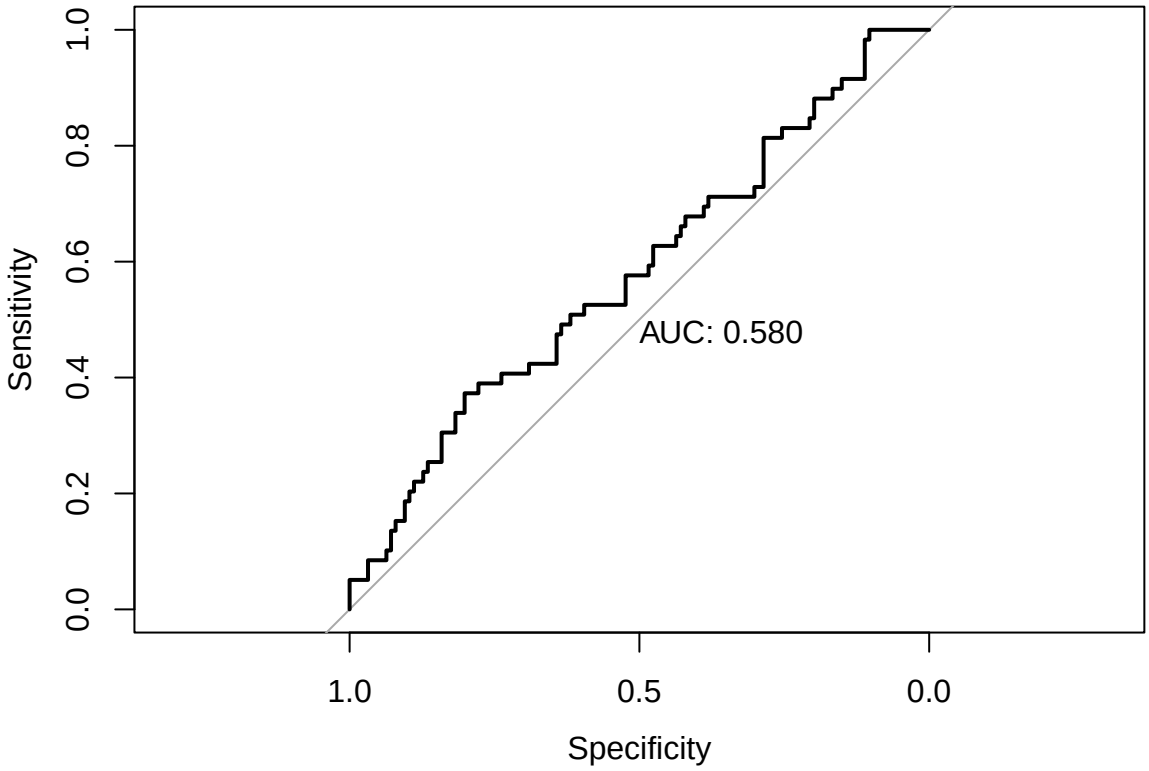




group 0



group 1



The LDA model was constructed using the selected variables, and a confusion matrix was generated. The results indicate that out of the total 185 data points, 125 were predicted correctly (116 true positives and 9 true negatives), while 60 data points were predicted incorrectly (50 false positives and 10 false negatives).

The model's accuracy stood at 67.57%, signifying the proportion of accurate predictions out of all predictions made. The true positive rate, representing correctly identified positive cases, was 15.25%, while the false positive rate, indicating the proportion of negative cases incorrectly identified as positive, was 7.94%. The mean squared error (MSE) for the model was calculated at 1.056.

In addition, the analysis involved creating plots to illustrate the distribution of predicted data and the class predictions.

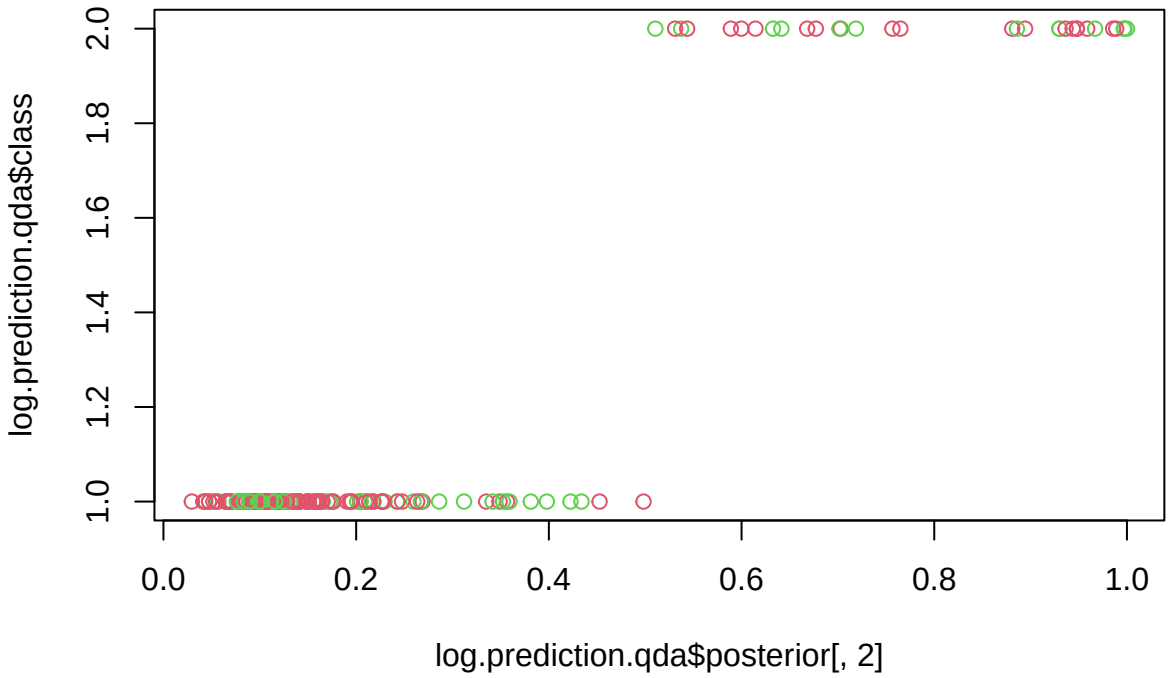
From the LDA analysis, it is evident that sleep quality is influenced, whether positively or negatively, by factors such as poor physical health, depression, alcohol consumption, income, hard drug use, and smoking. However, it's important to note that the classifier was not entirely accurate in its predictions.

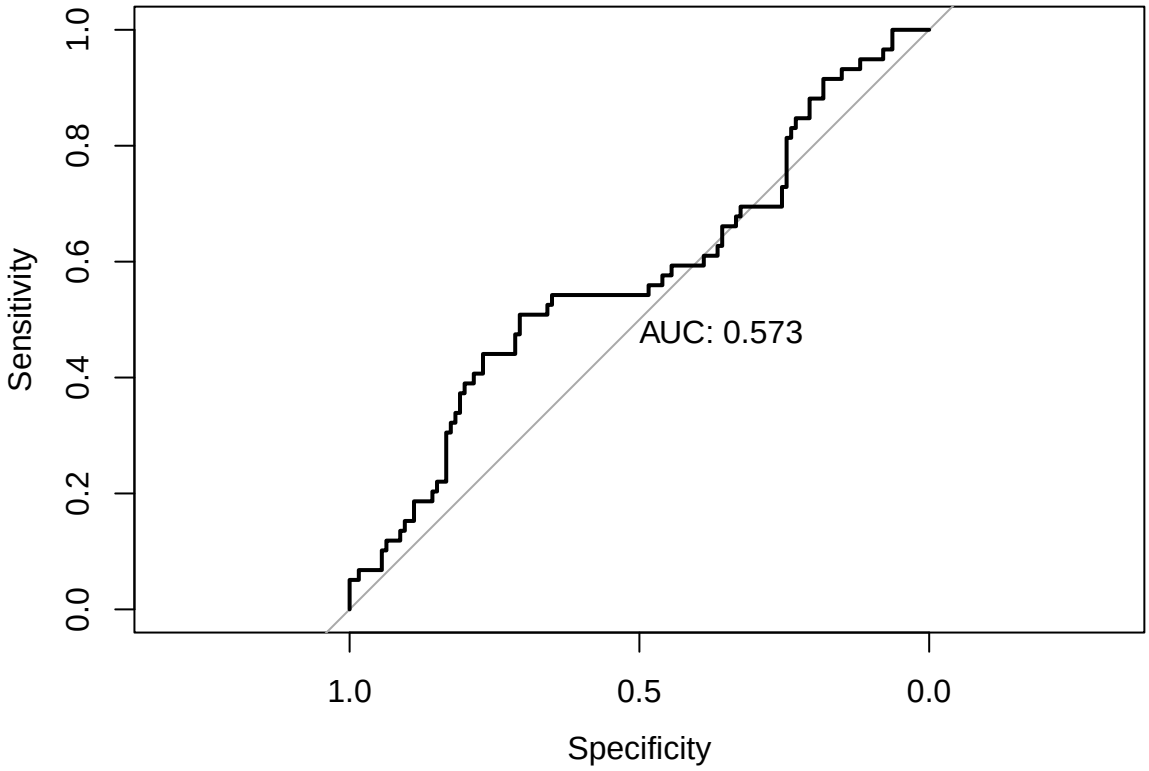
[5] Quadratic discriminant analysis (QDA)

```
##
## 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 applied to the dataset, and a corresponding confusion matrix was generated. The results indicate that out of the total 185 data points, 120 were accurately predicted (107 true positives and 13 true negatives), while 65 data points were incorrectly predicted (19 false positives and 46 false negatives).

The model achieved an accuracy of 64.86%, representing the proportion of correct predictions among all predictions. The true positive rate, reflecting correctly identified positive cases, stood at 22.03%, while the false positive rate, indicating the proportion of negative cases incorrectly identified as positive, was 15.08%. The model's mean squared error (MSE) was calculated at 0.258.

Additionally, visualizations were created to illustrate the distribution of predicted data.

While the analysis revealed that factors such as poor physical health, depression, alcohol consumption, income, hard drug use, and smoking have a significant impact on an individual's sleep, it's worth noting that the classifier had limitations in making accurate predictions. This is evident from the model's AUC (Area Under the Curve) of 0.573, which is considered subpar and lower than other models, indicating room for improvement in predictive accuracy.

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 the most effective model, I conducted an analysis using multiple evaluation metrics, including Mean Squared Error (MSE), False Positive Rate (FPR), True Positive Rate (TPR), and the ROC curve, which collectively assess the model's performance.When assessing the MSE, both the neural network and QDA displayed the lowest values, indicating relatively accurate predictions. However, I would not recommend QDA due to its elevated False Positive Rate, which raises concerns about its ability to classify effectively.The Area Under the Curve (AUC) for each model is as follows: Logistic Regression = 0.580, Neural Network = 0.187, K-Nearest Neighbors (KNN) = 0.616, Linear Discriminant Analysis (LDA) = 0.580, and Quadratic Discriminant Analysis (QDA) = 0.573. AUC values greater than 0.5 are generally considered good, while values around 0.5 are considered less favorable.Given the neural network's combination of the lowest MSE and the highest AUC, it appears to be the most reliable classifier for this dataset. Therefore, I would recommend using the neural network model for the data analysis.