Homework 11

Snigdha Peddi

Exercises (MDSR)

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?

Answer:

NHANES data set is read and preprocessed for duplicate records,features that were obviously correlated were removed like Age in years and Age in Months etc, the features that have missing values for more than 70% records were removed, missing values for categorical variables were imputed, one Hot encoding of categorical variables was performed. Then the data is split in 90/10 ratio for training and testing data. After the split the data is imputed for missing values in numerical variables. The values were then standardized so that all the variables are between 0 and 1.Finally, Random Forest classifier is used to select the top most important variable that will be further used to fit models using classifier like KNN, Logistic Regression, LDA, QDA and Neural Networks.

Duplicate Records:

```
## Dimensions of NHANES Datatset:
##  10000 76
```

```
##
##
## Dimensions of NHANES Datatset after removing the duplicate records:
##  7832 76
```

Feature Selection:

- SurveyYr and ID were removed as target variable does not depend on these variables
- Age in Decade and Age in months variable are removed as they are related to Age variable
- Race 3 variable removed is as it is related to Race1 variable and values are reported for years 2009-2010,
- HHIncome is removed as HHincomeMid can be used which reports the median value in each HHIncome category
- Length and Headcirc removed as it is reported for infants only
- BMICatUnder20yrs is removed as information belongs only to individuals of ages below 20
- BMI_WHO removed as similar to BMI but categorized
- Removed BPSys1,BPDia1,BPSys2,BPDia2,BPSys3,BPDia3 as combined systolic and diastolic blood pressure is available
- Testosterone, TVHrsDay and CompHrsDay is removed as data is not available for years 2009-2010
- UrineVol2, UrineFlow2, DiabetesAge, Age1stBaby, TVHrsDayChild, CompHrsDayChild, PregnantNow, AgeRegMarij are removed as they have about 80% missing values
- npregnancies and nBabies are removed as they are reported only for females over age 20 and have about 74% and 75% missing values,
- Weight and Height removed as BMI gives similar information
- Removed variables DaysPhysHlthBad,DaysMentHlthBad,LittleInterest,Depressed as HealthGen provides similar information
- removed Alcohol12PlusYr as it is similar to AlcoholYear
- removed SmokeNow as NA values include participants who smoked less than 100 times including missing values and smoke1000 gives better picture of the effect of smoking
- Marijuana and AgeFirstMarij removed as variables RegularMarij give more relavent information
- Removed SexAge, SexNumPartnLife, SexNumPartYear,SameSex,SexOrientation as these variables are not very relevant to explain response variable
- Removed Smoke100n as it is same as Smoke100
- Removed SmokeAge as it has about 70% missing values
- SleepHrsNight is removed as it is same as sleeptrouble

```
##
##
## Dimensions of NHANES Datatset after initial Feature Selection:
##  7832 29
```

Missing Value Imputation of Categorical Variables After analyzing the variables the missing values are categorized into missing values (for instance, if value is missing for Homeown variable then the values are imputed as missing category) and Not Applicable (For instance, HealthGEn is reported for ages 12 and above, the missing values are imputed as Not Applicable) depending on the information provided on individual categorical variables in the NHANES data.

```
##
## Number of Missing values in categorical variables including Target Variable :

## [1] 0

##
## Dimensions of NHANES Datatset after missing value imputation for Categorical variables:
##  5911 29
```

One Hot Encoding: Categorical variables except the target variable were now converted to numerical variables by one hot encoding method.

```
##
## Dimensions of NHANES Datatset after One Hot Encoding of Categorical variables:
##  5911 68
```

90/10 Split (Validation Approach): The resultant data is split into a ratio of 90:10 (training:test respectively). As the number of records in each class are not proportional measures were taken to make sure 90/10 split resulted in equal proportion of classes.

```
##
## Dimensions of NHANES Train Datatset:
##  5319 68
```

```
##
## Dimensions of NHANES Test Datatset:
##  592 68
```

Missing Value Imputation of Numerical Variables: Missing values are replaced by mean value of the variable for most of the predictors except HHIncomeMid and Poverty. Median values were used for predictors HHIncomeMid and Poverty as more sample population is below mean value and replacing with higher value would affect prediction.

```
##
## Number of Missing values in Numerical variables after imputation:
```

```
## [1] 0
```

Variable Importance Random Forest classifier is used to compute the variable importance using Gini Index.

Table 1: *Variable Importance by Random Forest*

	MeanDecreaseGini
BMI	127.285
TotChol	112.021
UrineFlow1	111.817
Age	111.081
UrineVol1	111.038
BPSysAve	107.572
DirectChol	105.159
BPDiaAve	103.541
Pulse	95.819
Poverty	92.587
AlcoholYear	75.648
HomeRooms	69.029
HHIncomeMid	66.394
AlcoholDay	50.060
PhysActiveDays	42.454

	MeanDecreaseGini
Smoke100_Yes	23.290
Race1_White	23.243
HardDrugs_Yes	20.578
Work_Working	18.781
Education_Some_College	18.135
HealthGen_Good	18.092
HealthGen_Vgood	17.540
Work_NotWorking	17.296
Gender_male	17.203
MaritalStatus_Married	16.944
Gender_female	16.798
PhysActive_Yes	16.534
HealthGen_Poor	16.484
PhysActive_No	16.386
Education_High_School	15.678
HomeOwn_Own	15.338
Education_College_Grad	15.066
MaritalStatus_Divorced	14.546
Smoke100_No	14.461
HomeOwn_Rent	14.391
HealthGen_Fair	14.215
RegularMarij_No	13.996
RegularMarij_Yes	13.430
HardDrugs_No	13.162
Diabetes_No	12.593
MaritalStatus_NeverMarried	12.545
HealthGen_Excellent	12.184
Diabetes_Yes	11.643
RegularMarij_Missing	11.281
Race1_Black	10.602
SexEver_Yes	10.457
Education_9_11th_Grade	10.436
Race1_Mexican	10.427
MaritalStatus_LivePartner	9.561
SexEver_Missing	8.696
Work_Looking	8.547
MaritalStatus_Widowed	8.164
HardDrugs_Missing	8.052
HealthGen_Not_Applicable	7.799
Race1_Other	7.643
Education_8th_Grade	6.877
Race1_Hispanic	6.116
MaritalStatus_Separated	5.172
HomeOwn_Other	5.159
Education_age_LT_20	3.813
MaritalStatus_Not_Applicable	3.337
Smoke100_Not_Applicable	3.000
SexEver_No	2.908
HomeOwn_Missing	0.988
Work_Not_Applicable	0.010
Diabetes_Missing	0.000
PhysActive_Not_Applicable	0.000

Standardization of Features: Scale() function is used to normalize or standardize the values in the dataset to numbers between 0 and 1. By using center and scale parameters as True in the scale function the variables are scaled as per Z-score formula. Mean value of each variable is subtracted from each value and is then divided by Standard deviation of the feature resulting in a value between 0 and 1. This helps improving the accuracy when models like SVM,KNN,logistic regression,Neural network are fit for predictions.

```
##             BMI         Age    TotChol UrineFlow1  UrineVol1   BPSysAve
## [1,]  0.5514334 -0.62065194 -1.4229037  0.0000000  2.5888465 -0.4411312
## [2,]  0.3049176  0.19278382  1.6552243 -0.9920592 -0.4760105 -0.5003345
## [3,] -0.3733746  0.46392907  1.3771380 -0.1111146  1.0619905  0.8021380
## [4,] -0.4017613  0.78930337  1.3579596  0.0000000  1.3183240  1.8677972
## [5,]  1.1639881  0.46392907 -0.1571314 -0.1514050  0.5381785 -1.8028070
## [6,]  0.4334046 -0.07836144  0.6100033 -0.4105704 -0.3757061  1.3941709
##       DirectChol   BPDiaAve      Pulse    Poverty AlcoholYear   HomeRooms
## [1,] -0.18120009  1.2478867 -0.2064625 -0.8705976  -0.8233888 -0.04861629
## [2,] -0.51402575  0.4582416  1.1600192 -0.5300788  -0.5975972 -0.49221890
## [3,] -0.51402575  1.2478867  0.3059681 -0.3505325   3.2860186 -0.04861629
## [4,] -0.05319022  2.4323544  0.9892090 -1.0749088  -0.4169639 -0.49221890
## [5,]  0.61246110 -2.2265517 -0.7188932  0.3181225   0.9377858 -1.37942411
## [6,] -1.76852247  1.4847803  0.6475886 -0.5919913  -0.8233888  1.72579414
##      HHIncomeMid
## [1,]  -0.7852390
## [2,]  -0.4681147
## [3,]   0.4832584
## [4,]  -1.1816445
## [5,]   0.1661341
## [6,]  -0.4681147
```

Preprocessing Test Data: All the preprocessing steps done for training data are done for test data just before checking the model's performance.

Logistic Regression

The NHANES data is preprocessed to select the important variables. Further the data is split into 90/10 (90% Training data and 10% Test data). As the classes of Target variable are unequal before splitting measures were taken to make sure training and test data have equal proportions of records. Three logistic regression models were fit based on the top most important variables selected in the preprocessing steps using Random Forest classifier. In each of fits, using test data and a threshold of 0.5 the misclassification was approximately 0.25 ,True Positive Rate is 0 and True Negative Rate is 1 indicating all the individuals having no sleep trouble are being predicted 100% accurately but individuals having sleep trouble are not being predicted accurately. By modifying the threshold the TPR is improved and TNR turned out to be reasonable for all the models (missclassification rate is declined). All the Metrics are presented below .

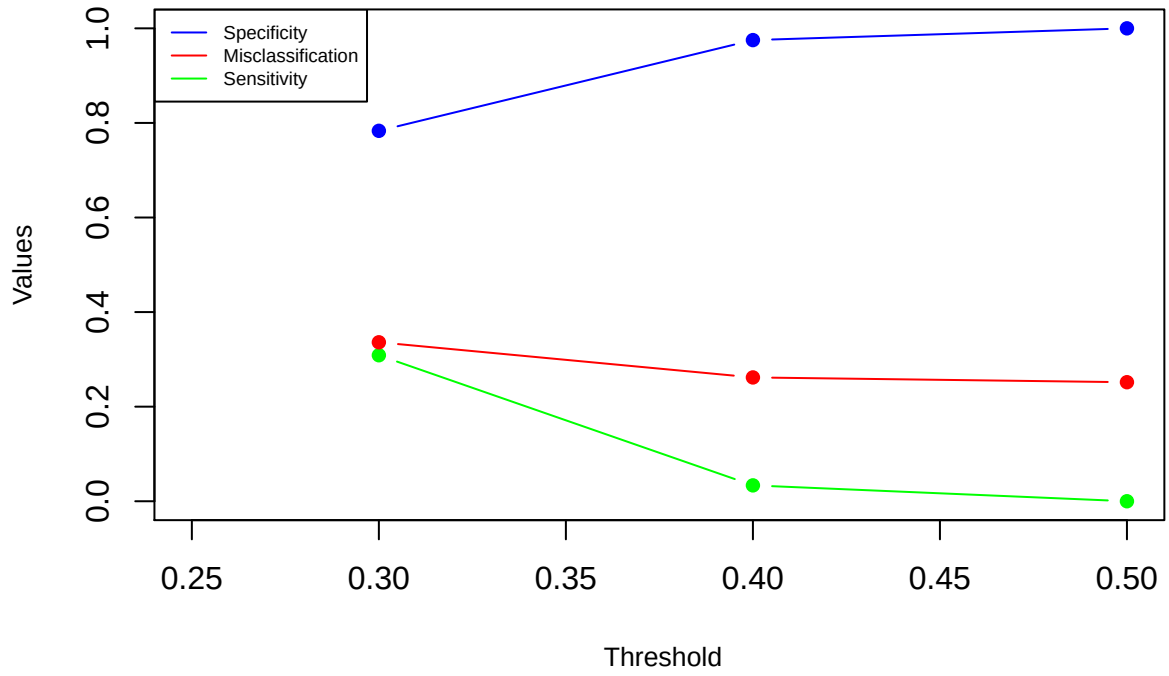
```
##
## Summary of Logistic Regression Model 1:
```

```
##
## Call:
## glm(formula = SleepTrouble ~ ., family = "binomial", data = train.act)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.3824  -0.7859  -0.6707   1.0518   2.1187
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.1366849  0.0327533 -34.704  < 2e-16 ***
## BMI          0.1670864  0.0333011   5.017 5.24e-07 ***
## Age          0.3608627  0.0400200   9.017  < 2e-16 ***
## TotChol     -0.0003886  0.0339245  -0.011 0.990861
## UrineFlow1   0.0278629  0.0393329   0.708 0.478706
## UrineVol1    0.0010463  0.0402022   0.026 0.979237
## BPSysAve    -0.1559562  0.0382686  -4.075 4.60e-05 ***
## DirectChol   0.0695621  0.0350697   1.984 0.047307 *
## BPDiaAve     0.0073645  0.0350080   0.210 0.833382
## Pulse        0.1096297  0.0329430   3.328 0.000875 ***
## Poverty     -0.0056661  0.0741435  -0.076 0.939084
## AlcoholYear -0.0122002  0.0329173  -0.371 0.710911
## HomeRooms    0.0422760  0.0361504   1.169 0.242223
## HHIncomeMid -0.1320520  0.0757883  -1.742 0.081442 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5987.1  on 5318  degrees of freedom
## Residual deviance: 5823.2  on 5305  degrees of freedom
## AIC: 5851.2
##
## Number of Fisher Scoring iterations: 4
```

Table 2: *Fit with 13 Variables*

	Threshold_0.5	Threshold_0.4	Threshold_0.3
Misclassification Rate	0.2516892	0.2618243	0.3361486
Sensitivity	0.0000000	0.0335570	0.3087248
Specificity	1.0000000	0.9751693	0.7832957

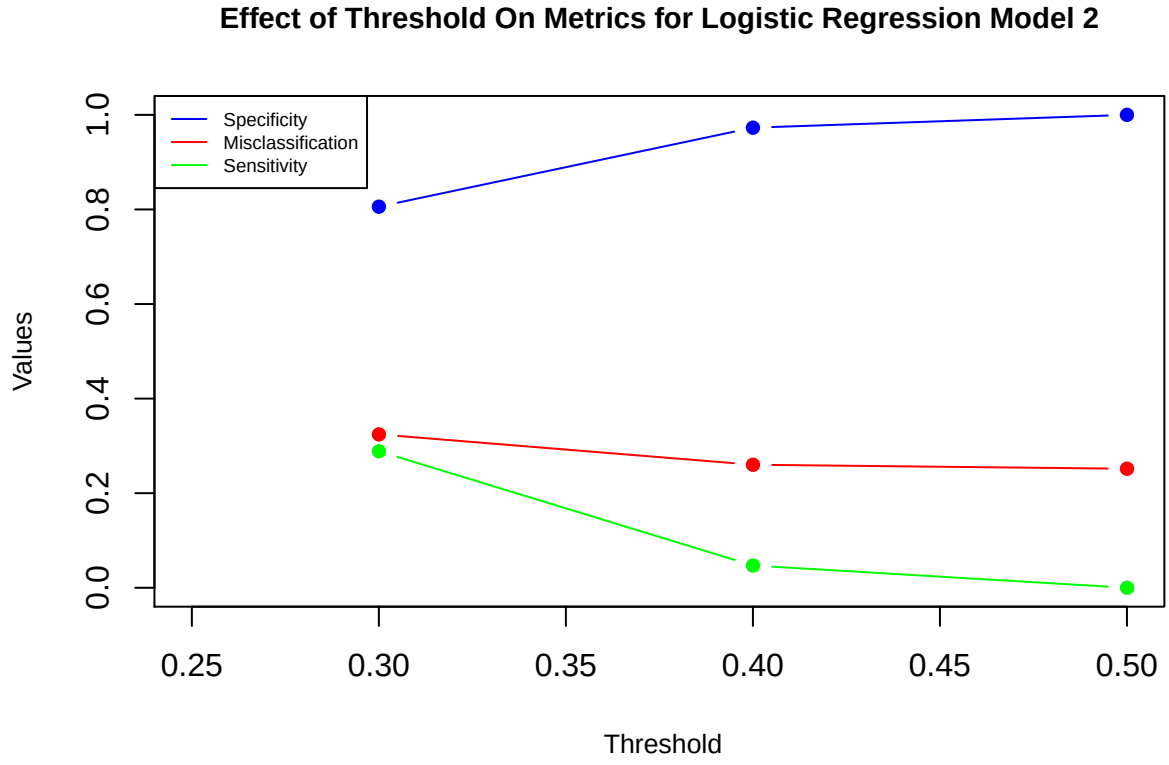



```
##
## Summary of Logistic Regression Model 2:

##
## Call:
## glm(formula = SleepTrouble ~ BMI + TotChol + Age + UrineFlow1 +
##     UrineVol1 + BPSysAve + Poverty + DirectChol + BPDiaAve +
##     Pulse, family = "binomial", data = train.act)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3758  -0.7896  -0.6739   1.0484   2.1129
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.135832   0.032731 -34.702  < 2e-16 ***
## BMI          0.170795   0.033179   5.148 2.64e-07 ***
## TotChol     -0.001875   0.033926  -0.055 0.955922
## Age          0.382975   0.038260  10.010  < 2e-16 ***
## UrineFlow1   0.025496   0.039238   0.650 0.515836
## UrineVol1    0.002025   0.040185   0.050 0.959814
## BPSysAve    -0.158238   0.038134  -4.150 3.33e-05 ***
## Poverty     -0.110579   0.033542  -3.297 0.000978 ***
## DirectChol   0.068886   0.034603   1.991 0.046507 *
## BPDiaAve     0.004856   0.034950   0.139 0.889498
```

```
## Pulse        0.112995   0.032903   3.434 0.000594 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5987.1  on 5318  degrees of freedom
## Residual deviance: 5827.1  on 5308  degrees of freedom
## AIC: 5849.1
##
## Number of Fisher Scoring iterations: 4
```

Table 3: *Fit with 7 variables*

	Threshold_0.5	Threshold_0.4	Threshold_0.3
Misclassification Rate	0.2516892	0.2601351	0.3243243
Sensitivity	0.0000000	0.0469799	0.2885906
Specificity	1.0000000	0.9729120	0.8058691

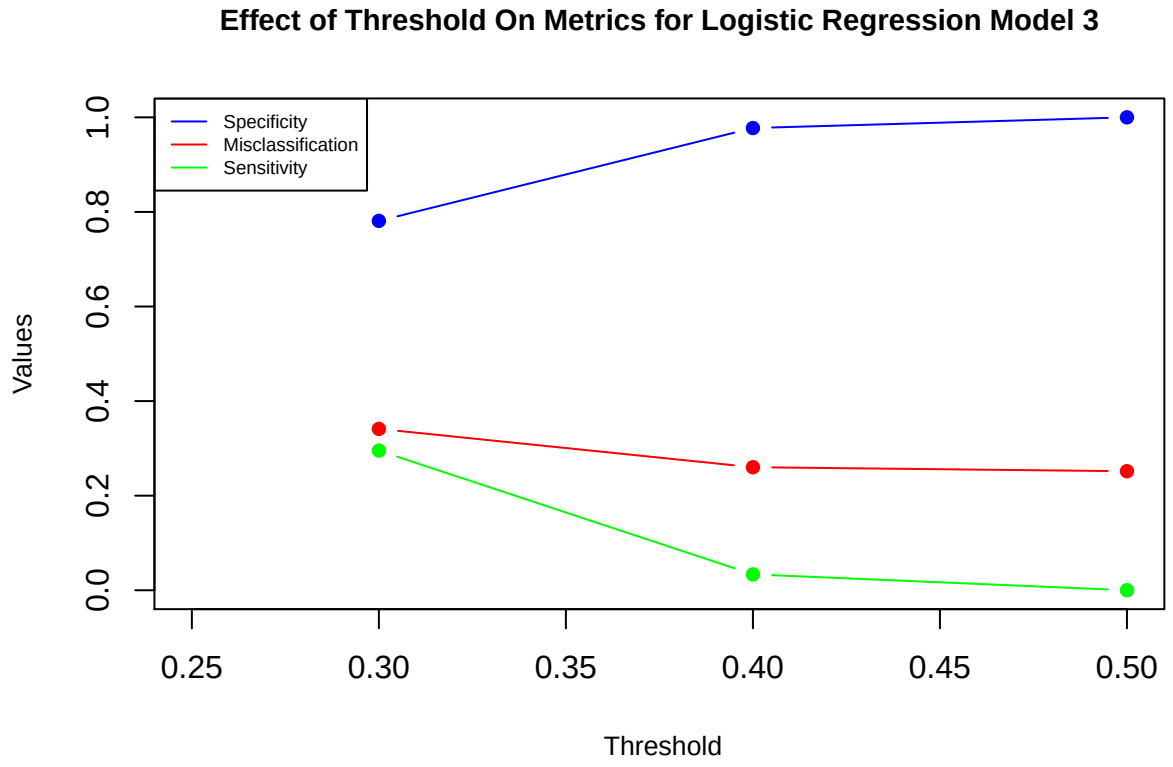



```
##
## Summary of Logistic Regression Model 3:

##
```

```
## Call:
## glm(formula = SleepTrouble ~ BMI + Age + Poverty + Pulse, family = "binomial",
##     data = train.act)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2783  -0.7884  -0.6776   1.1436   2.0820
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.13005    0.03257 -34.692  < 2e-16 ***
## BMI          0.13855    0.03133   4.422 9.79e-06 ***
## Age          0.31968    0.03286   9.728  < 2e-16 ***
## Poverty     -0.08874    0.03283  -2.703 0.006867 **
## Pulse        0.11292    0.03234   3.491 0.000481 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5987.1  on 5318  degrees of freedom
## Residual deviance: 5851.2  on 5314  degrees of freedom
## AIC: 5861.2
##
## Number of Fisher Scoring iterations: 4
```

Table 4: *Fit with Top 4 significant variables*

	Threshold_0.5	Threshold_0.4	Threshold_0.3
Misclassification Rate	0.2516892	0.2601351	0.3412162
Sensitivity	0.0000000	0.0335570	0.2953020
Specificity	1.0000000	0.9774266	0.7810384



Considering the p values of all the logistic models the variables like BMI,Age,Poverty, Systolic Blood Pressure, Pulse and Total HDL Cholesterol effects the sleep patterns of the individuals. However, the logistic regression model cannot give accurate predictions as the TPR is very low, cannot predict the individuals with Sleep Trouble atleast reasonably and the misclassification rate is also high.I believe this can be due to the drastic difference in the proportion of the classes in the data provided and also not all information collected is from same age group i.e., information for few predictors is collected from ages 6 and above, few collected from 20 and above etc. This could have affected the variable selection process and the even the classification within the target variable.

K-Nearest Neighbors (KNN)

Using the same predictors and response variable as before a knn model is fit.The training and test data obtained by 90/10 split is used for analysis.K values from 1 to 100 were investigated and a graph showing the misclassification, sensitivity, and specificity for all the K values is plotted.

```
##
## knn model fit for a k value of 1:
##
##  knn(train.act.knn,test.act.knn,cl=train.act.knn.y,k=1)
```

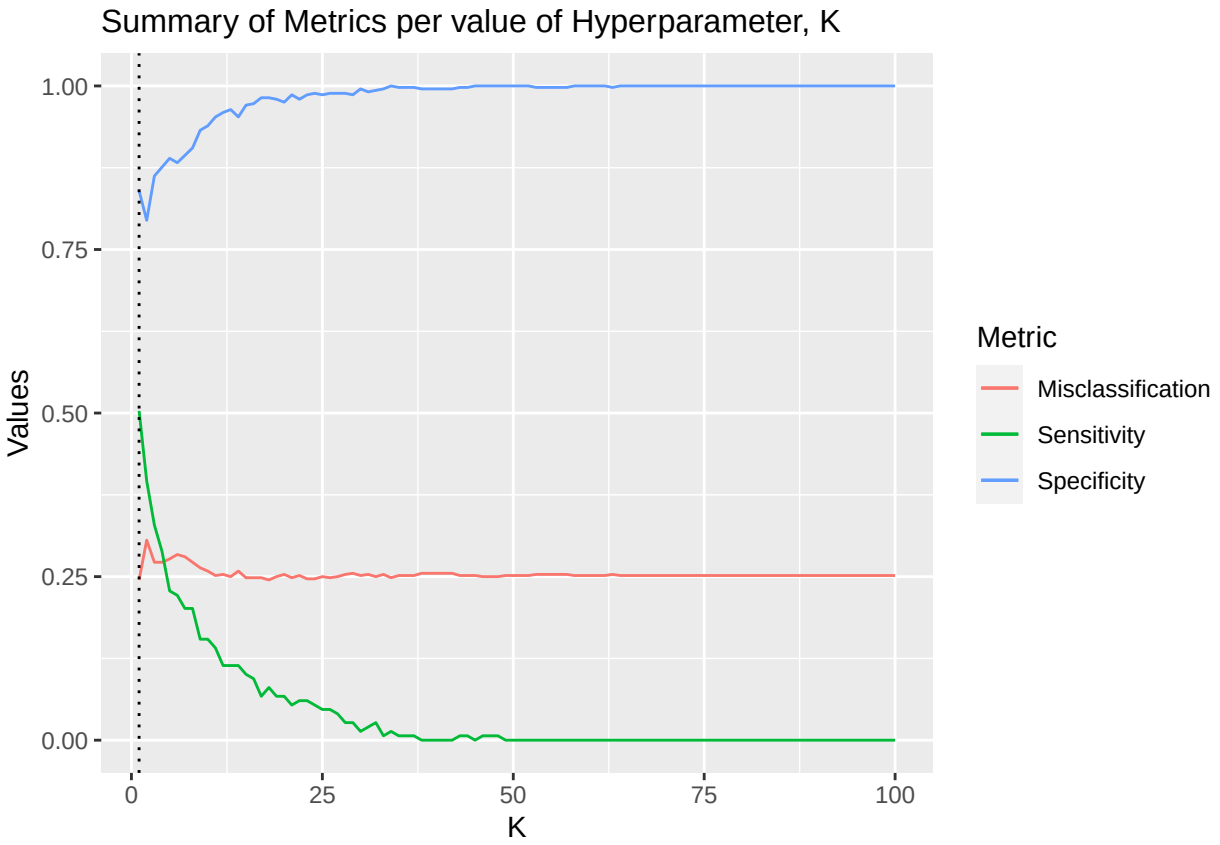


Table 5: Metrics for KNN Model Where K=1

Metrics	Values
Misclassification Rate	0.24493
Sensitivity	0.50336
Specificity	0.83973

At a K value of 1, looking at the misclassification rate, sensitivity and specificity I still consider the variables BMI,Age,Poverty, Systolic Blood Pressure, Pulse and Total HDL Cholesterol effects the sleep patterns of the individuals. However, the knn model is not giving an accurate predictions.The TPR rate of about 0.5 indicates that only half of the individuals with sleep trouble are being predicted accurately. The TNR of 0.84 is reasonable and the model is able to predict the individuals with out sleep trouble resonably. And there is a missclassification rate of 25% which is not good.

Linear Discriminant Analysis (LDA)

Using the same predictors and response variable as before two LDA models were fit.The training and test data obtained by 90/10 split is used for analysis. First LDA fit has the top 10 variables listed per variable importance table of Random Forest and second model was fit using variables BMI, Age, BPSysAve, Poverty, DirectChol, Pulse. The misclassification, sensitivity, and specificity was evaluated for both the models using the test data. Both the fits have a misclassification rate of about 25% and Sensitivity 0 and Specificity 1 indicating that the test data did not make accurate predictions.

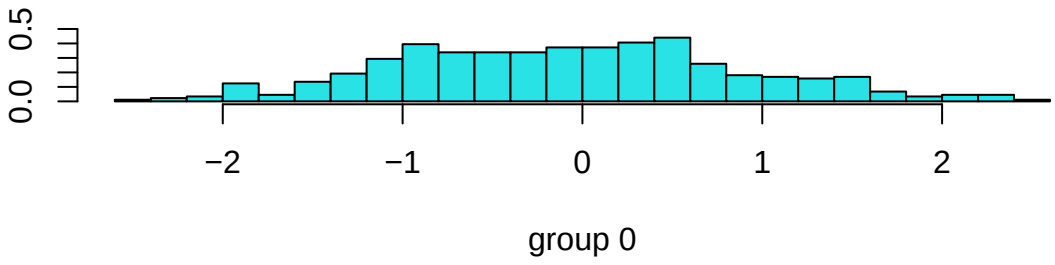
##

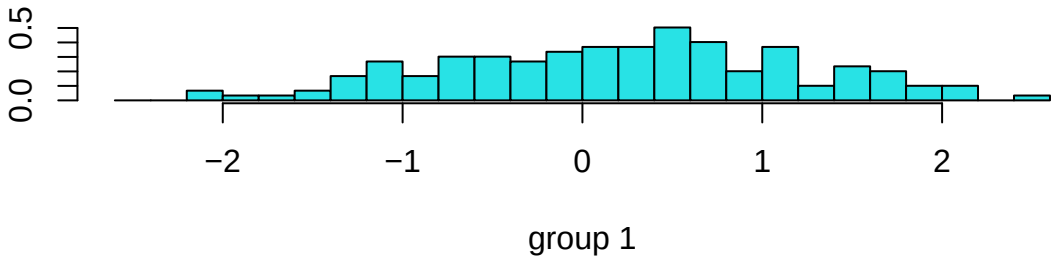
```
## LDA fit for NHANES data using top 10 important variables:

## lda(formula = SleepTrouble ~ ., data = train.act.lda)
```

Table 6: Results of LDA, Fit 1

	Metrics
Misclassification Rate	0.25169
Sensitivity	0.00000
Specificity	1.00000

```
##
## Histogram for LDA fit 1:
```



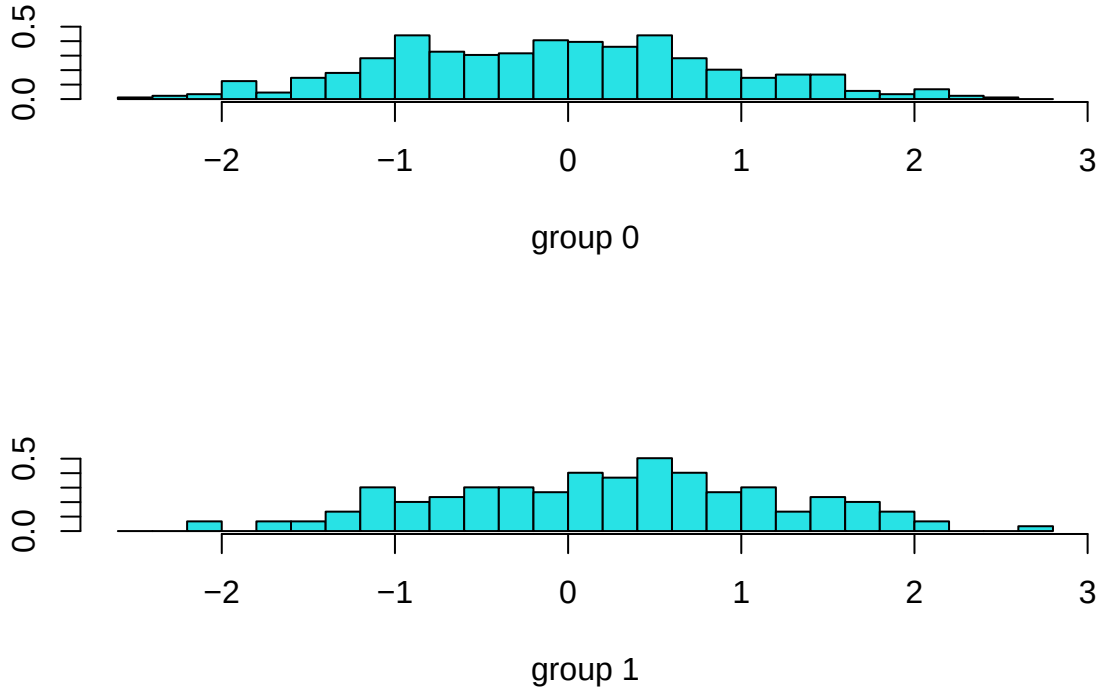



```
##
## LDA fit for NHANES data using top 6 important variables,fit 2:

## lda(formula = SleepTrouble ~ BMI + Age + BPSysAve + Poverty +
##     DirectChol + Pulse, data = train.act.lda)
```

Table 7: Results of LDA,Fit 2

	Metrics
Misclassification Rate	0.25169
Sensitivity	0.00000
Specificity	1.00000

```
##
## Histogram for LDA fit 2:
```



The misclassification rate, sensitivity and specificity for both the models are not great.The TPR rate of about 0 and TNR of 1 indicates that only individuals who do not have any sleep trouble are being identified accurately and the individuals who actually have trouble sleep are not being predicted accurately at all.And there is a missclassification rate of 25% which is not good enough.The Histograms from both the fit clealy indicate that there is a complete overlap and the models were not able to separate the groups. This behavior shows that the data is linearly related and LDA cannot separately linearly related data.

Quadratic discriminant analysis (QDA)

Similar to LDA models the same predictors and response variable were used to fit two QDA models.The training and test data obtained by 90/10 split is used for analysis. First QDA fit has the top 10 variables listed per variable importance table of Random Forest and second model was fit using variables BMI, Age, BPSysAve, Poverty, DirectChol, Pulse. The misclassification, sensitivity, and specificity was evaluated for both the models using the test data. Both the fits have a misclassification rate of about 29% and Specificity 0.9. However, the first model's sensitivity(0.15) is little better compared to the second model(0.13).

```
##
## QDA fit for NHANES data using top 10 important variables:

## qda(formula = SleepTrouble ~ ., data = train.act.qda, method = "t")
```

Table 8: Results of QDA, Fit 1

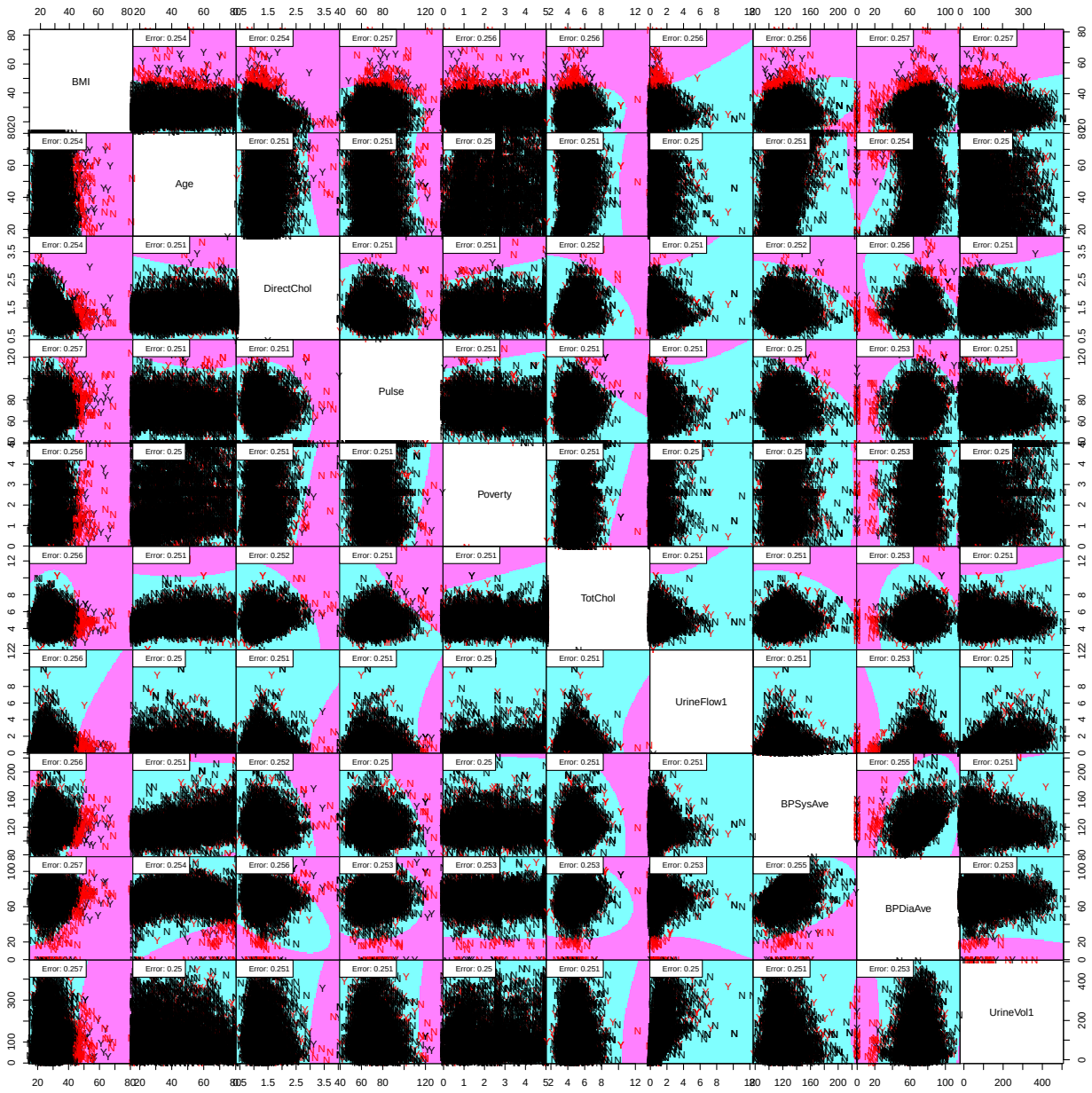
	Metrics
Misclassification Rate	0.29054
Sensitivity	0.14765
Specificity	0.89842

```
##
## QDA fit for NHANES data using 6 important variables:

## qda(formula = SleepTrouble ~ BMI + Age + BPSysAve + Poverty +
##     DirectChol + Pulse, data = train.act.qda, method = "t")
```

Table 9: Results of QDA, Fit 2

	Metrics
Misclassification Rate	0.29223
Sensitivity	0.12752
Specificity	0.90293

The below plot shows the classification of observations based on QDA method for each combination of variables in training data. The error rate is also displayed for a combination of variable.



The misclassification rate, sensitivity and specificity for both the models are not great.The TPR rate of about 0.15 for first fit and 0.13 for the second fit indicate that only about 15% and 13% of the individuals having sleep trouble are predicted accurately using QDA classifier.TNR of 0.9 indicates that only 90% individuals who do not have any sleep trouble are being identified accurately.And there is a missclassification rate of 29% which is not good enough. Though QDA fits did little better than the LDA both the models cannot provide any useful information.

Neural Network

Similar to all other models the predictors BMI, Age, DirectChol, Pulse, Poverty, ToChol, UrineFlow1, urineVol1, BPSysAve, BPDiaAve were used.The training and test data obtained by 90/10 split is used for analysis. Six different models were fit using *nnet()* function with various combination of parameters like Size, threshold, maxit,decay. The model with a size of 5 and maxit of 10000 resulted in the worst fir with a TPR and TNR of 0. A better model was obtained when size is 75,maxit is 1000, decay is 0.0005 and

threshold is 0.45 (TPR of 0.44 and TNR of 0.78). It is observed that with addition of decay and threshold as hyperparametes the performance is improved.

```
##
## Neuralnet,Fit-1:

## a 10-5-1 network with 61 weights
## inputs: BMI Age TotChol UrineFlow1 UrineVol1 BPSysAve DirectChol BPDiaAve Pulse Poverty
## output(s): SleepTrouble
## options were - entropy fitting
```

Table 10: Results of NeuralNet-Fit 1(size=5,maxit=10000)

	Metrics
Misclassification.Rate	0.25169
Sensitivity	0.00000
Specificity	0.00000

```
##
## Neuralnet,Fit-2:

## a 10-50-1 network with 601 weights
## inputs: BMI Age TotChol UrineFlow1 UrineVol1 BPSysAve DirectChol BPDiaAve Pulse Poverty
## output(s): SleepTrouble
## options were - entropy fitting
```

Table 11: Results of NeuralNet-Fit 2(size=50,maxit=1000)

	Metrics
Misclassification Rate	0.26351
Sensitivity	0.36000
Specificity	0.75309

```
##
## Neuralnet,Fit-3:

## a 10-75-1 network with 901 weights
## inputs: BMI Age TotChol UrineFlow1 UrineVol1 BPSysAve DirectChol BPDiaAve Pulse Poverty
## output(s): SleepTrouble
## options were - entropy fitting
```

Table 12: Results of NeuralNet-Fit 3(size=75,maxit=1000)

	Metrics
Misclassification Rate	0.27027
Sensitivity	0.40351
Specificity	0.76449

```
##
## Neuralnet,Fit-4:

## a 10-75-1 network with 901 weights
## inputs: BMI Age TotChol UrineFlow1 UrineVol1 BPSysAve DirectChol BPDiaAve Pulse Poverty
## output(s): SleepTrouble
## options were - entropy fitting  decay=5e-04
```

Table 13: Results of NeuralNet-Fit 4 (size=75,maxit=2000,decay=0.0005)

	Metrics
Misclassification Rate	0.29392
Sensitivity	0.34940
Specificity	0.76424

```
##
## Neuralnet,Fit-5:

## a 10-75-1 network with 901 weights
## inputs: BMI Age TotChol UrineFlow1 UrineVol1 BPSysAve DirectChol BPDiaAve Pulse Poverty
## output(s): SleepTrouble
## options were - entropy fitting  decay=5e-04

## [1] 0.7314189
```

Table 14: Results of NeuralNet-Fit 5 (size=75,maxit=1000,decay=0.0005,threshold=0.45)

	Metrics
Misclassification Rate	0.26858
Sensitivity	0.43750
Specificity	0.77734

```
##
## Neuralnet,Fit-6:

## a 10-75-1 network with 901 weights
## inputs: BMI Age TotChol UrineFlow1 UrineVol1 BPSysAve DirectChol BPDiaAve Pulse Poverty
## output(s): SleepTrouble
## options were - entropy fitting

## [1] 0.7297297
```

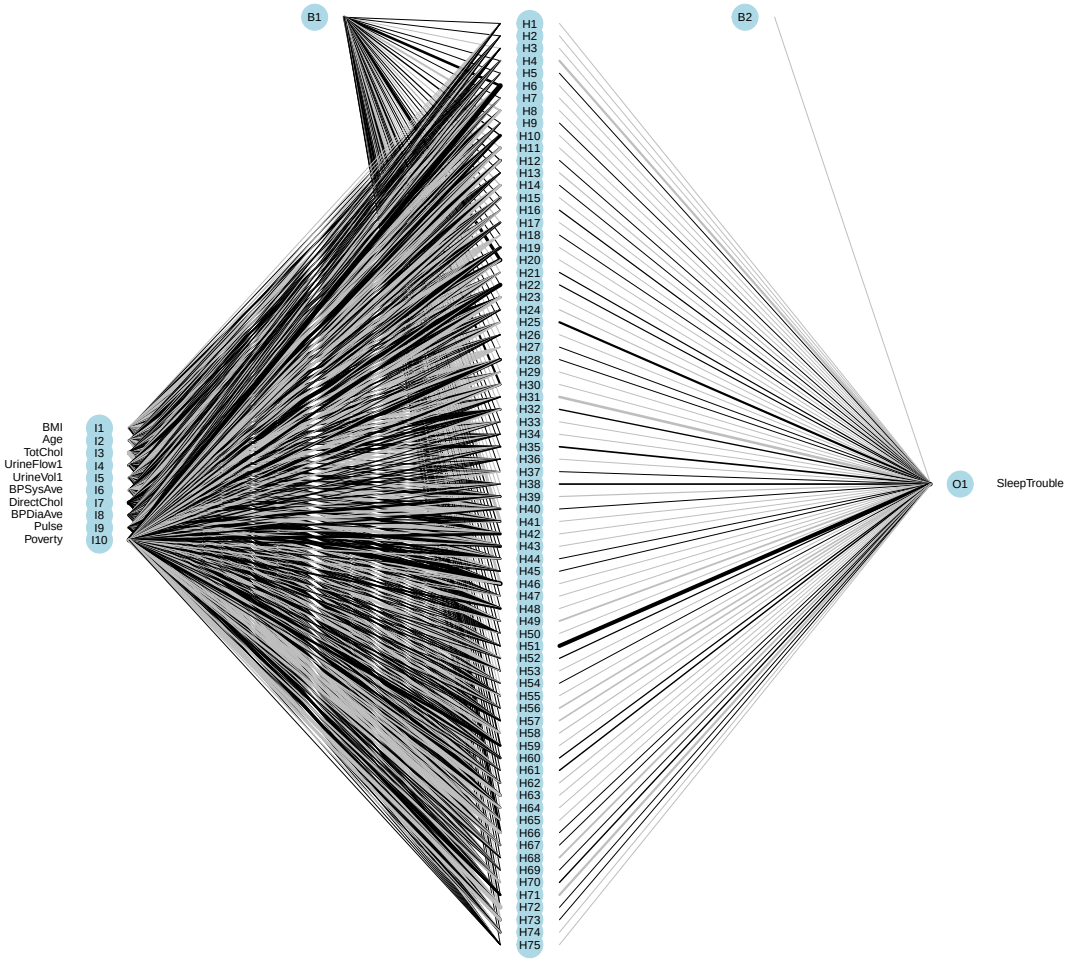
Table 15: Results of NeuralNet-Fit 6 (size=75,maxit=1000,threshold=0.45)

	Metrics
Misclassification Rate	0.27027
Sensitivity	0.40351
Specificity	0.76449

Table 16: Comparitive results of NeuralNet

	Fit1	Fit2	Fit3	Fit4	Fit5	Fit6
Misclassification.Rate	0.25169	0.26351	0.27027	0.29392	0.26858	0.27027
Sensitivity	0.00000	0.36000	0.40351	0.34940	0.43750	0.40351
Specificity	0.00000	0.75309	0.76449	0.76424	0.77734	0.76449

```
##
## Plot of Fit-5 with a TPR of 0.44 and TNR of 0.78:
```



The misclassification rate, sensitivity and specificity were improved with parameters decay(0.0005) and threshold (0.45) and hidden nodes(75) when compared to all other fits.The TPR rate of about 0.44 and TNP of 0.78 indicates that 44% of individuals who have any sleep trouble and 78 % of individuals who do not have sleep trouble are being identified accurately.And the missclassification rate is 27% which is not good enough. tuning of the hyperparameters will improve the performance of the models.However,lot of computational time is involved while using neural netweork classifier and has to be considered while selecting a better model.

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

Answer:

Metrics from all the different classifiers were tabulated below for comparison.

Table 17: Comparitive Metrics from all Classifiers

	NeuralNet	QDA	LDA	KNN	LogReg
Misclassification Rate	0.26858	0.29054	0.25169	0.24493	0.33615
Sensitivity	0.43750	0.14765	0.00000	0.50336	0.30872
Specificity	0.77734	0.89842	1.00000	0.83973	0.78330

From the table it is clear that KNN classifier at a K value of 1 resulted in better fit with a TPR of 0.5,TNR of 0.8 and Test error of 0.24.Tuning the hyper parameters of other classifiers might have resulted in better models but for the selected variables KNN performed well and was pretty quick. Though the metrics from KNN are reasonable when compared to others overall accuracy of predictions is low.I believe that this might be due to the difference in the proportion of the classes in the data provided and also not all information collected is from same age group resulting in more missing values.

References

- NHANES.pdf by Randall Pruim, *Package 'NHANES'*, August 29,2016.
- Blogpost by StackExchange,*How to replace NA values with another value in factors in R?*
- Blogpost by DATATRICKS,*One-hot encoding in R: three simple methods*, July 3,2019
- Blogpost by DATANOVIA,*Identify and Remove Duplicate Data in R*
- Vlogpost by Machine Learning Mastery,*Feature Selection with the Caret R Package*, August 22,2019
- caretSelection by Max Kuhn,*Variable Selection Using The caret Package*,June 30,2009
- Blogpost by Dataaspirant,*FEATURE SELECTION TECHNIQUES WITH R*,January 15,2018
- Blogpost by DataSharkie,*How to Standardize Data in R*
- Blogpost by ScienceDirect,*True Positive Rate*
- Blogpost by towards data science,*K-nearest Neighbors Algorithm with Examples in R (Simply Explained knn)*,December 30,2018
- Lecture by Dr. Bharatendra Rai,*Linear Discriminant Analysis in R | Example with Classification Model & Bi-Plot interpretation*,July 8,2017
- Lecture by Dr. Saunders,*Neuralnet Lecture*,April 16,2021
- R Documentation- qda{MASS},*Quadratic Discriminant Analysis*
- Lecture by dataAnalysisR,*Neural Network for iris data with R - nnetR1*,April 13, 2020