STA 141A Final Project

5/31/2022

Name	Email	Contribution
Halle Carter	hlcarter@ucdavis.edu	Logistic Regression
Chelsea Huffman	chuffman@ucdavis.edu	Graphs, Function
Vanessa Vu	vtvvu@ucdavis.edu	Subanalysis of subgroups
Martin Yossifov	mdyossifov@ucdavis.edu	k-NN
Rohan Arumugam	rarumugam@ucdavis.edu	Write-Up, Layout

A. Introduction and Research Question

Diabetes is a long-term, chronic condition that affects a portion of the world's population by hindering the way that our bodies use glucose as fuel. In this project, we want to determine the most important predictors for seeing whether someone has type II diabetes and if we can predict whether an individual is diagnosed with diabetes based on certain parameters, such as BMI and physical activity. People with type II diabetes have cells that do not respond normally to insulin, and as a result, the blood sugar levels in people with type II diabetes rise to dangerous levels. We hope to find the most influential predictors for diabetes to help prevent the likelihood of a person developing type II diabetes. We also wish to be able to find a way to predict whether a person will have type II diabetes, given their current health information.

To achieve these goals we will answer three specific questions:

- **Is a k-NN model or logistic regression model more accurate at predicting whether someone will have diabetes, given their current health information?**
- **How much interaction is there among some of the variables, and are some of them precursors to other variables?**
- **Are there subgroups where the effects of some of the variables are different/opposite to what they were in the general model(s)?**

B. Dataset Introduction

The diabetes dataset contains data on patients' health containing 17 quantitative variables such as BMI and sleep as well as qualitative variables like whether they smoke and drink. There are a total of 952 instances. The end goal is to build a prediction model based off of these variables to determine whether a patient has diabetes or not. Their diabetic condition is contained in the dataset, so dropping such labels would be necessary in the process of building such a model. Although this data was collected in India, our group can still make observations about the diabetic population as a whole. Furthermore, we omitted a few dozen patients who had NA values and there was an extra space in the cell of one of our data values, which was later removed.

We will be using the following predictors: BMI, sleep, Family Diabetes (Whether the test subjects have a family history of diabetes), Pregnancies, Age, Blood Pressure, Smoking, Alcohol, Gender, Physically Active, and Junk Food. These predictors will be used to predict whether or not an individual has type II diabetes.

Attribute	**Description**
Age	Age group: 40-49, 50-59, 60 or older, less than 40

Gender	Gender of individual: M, F
Family_Diabetes	If their family has a history of diabetes: yes, no
Physically_Active	Amount of physical activity per day: less than 30 minutes (0), more than 30 minutes (1), one or more hours (2), none (3)
BMI	BMI of the individual
Smoking	Whether or not the individual smokes: no (0), yes (1)
Alcohol	Whether or not the individual consumes alcohol: no (0), yes (1)
Sleep	Hours of sleep the individual gets per day (9+ hours of sleep are in one category)
SoundSleep	Hours of sound sleep the individual gets per day
JunkFood	How often they eat junk food: occasionally (0), often (1), very often (2), none (3)
BPLevel	Blood pressure level: low/normal (0), high (1)
Pregnancies	Number of pregnancies individual has had: none (0), 1 or more (1)
Diabetic	Whether or not they are diabetic: no (0), yes (1)

C. Data Visualization

This graph below displays a scatterplot matrix using the ggpairs() function. This scatterplot matrix displays the scatterplot, distribution, and Pearson correlation of the following quantitative variables: BMI, Sleep, and Pregnancies. The correlation coefficient for BMI and Sleep is -0.068, 0.042 for Sleep and Pregnancies, and -0.045 for BMI and Pregnancies. Since the absolute value of these correlation coefficients are very low, this suggests that collinearity does not exist between the quantitative variables. If we would like to use SoundSleep as well, we rerun the correlation matrix and get 0.53 with Sleep, and negligible values for the rest.

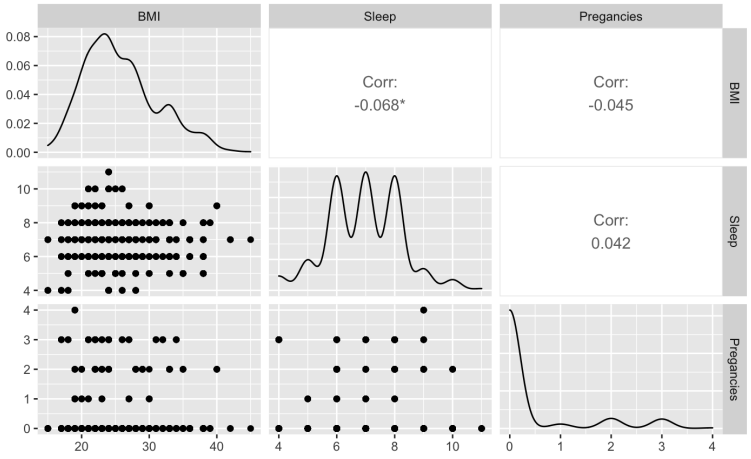


Figure 1. Scatterplot matrix displaying the scatterplot, distribution, and Pearson correlation of each quantitative variable. The p-values of each variable are also displayed: (0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

The dataset was split into two categories: Class = Yes (if diabetic) and Class = No (if not diabetic), allowing us to compare the distributions of diabetics and nondiabetics for each qualitative variable. In Figure 2, it is evident that the distributions of the diabetic and non-diabetic groups are fairly similar for each variable.

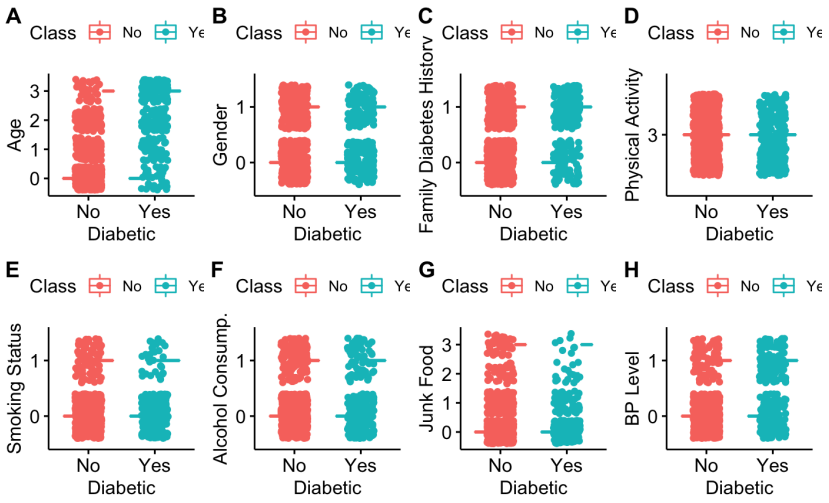


Figure 2. The distribution of each qualitative variable by diabetes status. (A-H: Age, Gender, Family_Diabetes, PhysicallyActive, Smoking, Alcohol, JunkFood, and BPLevel)

D. Method

Our first method uses a logistic regression to predict whether or not a person will have diabetes based on both qualitative and quantitative variables. We will also use this method to determine which variables are stronger predictors of whether a person will be diabetic or not.

Our second method involves a k-NN algorithm using the quantitative variables. Out of interest for the differences in the factors, we performed multiple k-NN's. Specifically, we see whether SoundSleep in addition to Sleep and the other numerical variables improves the classification rate. We see that adding it indeed does by a few percentage points.

A third method we use, this time to answer the third question regarding the constancy of the effects, is a cluster analysis. Specifically, we use hierarchical clustering to see whether we can classify groups according to their quantitative factors. We're not necessarily trying to predict or classify them as diabetic or not, but rather as "healthy" or not. This depends, in this case, on a previously scientifically proven right amount of BMI and sleep. The closer the patients' values

are to these “right” amounts, the more likely we are to classify them as “healthy” and vice-versa. We then use dplyr to extract the values of the quantitative variables of these two groups to see how they compare.

E. Result and Discussion

Method 1: Logistic Regression

We started the logistic regression with a full model of our selected data. After running the logistic regression on the full model, as seen to the right, almost all of our variables helped predict whether or not the individual was diabetic. We can conclude that an individual’s age, family history of diabetes, blood pressure, and amount of physical activity are statistically significant and may put them at higher risk for type II diabetes. Other slightly significant predictors (with a p-value < 0.05) are their smoking habits, sleep, and junk food consumption.

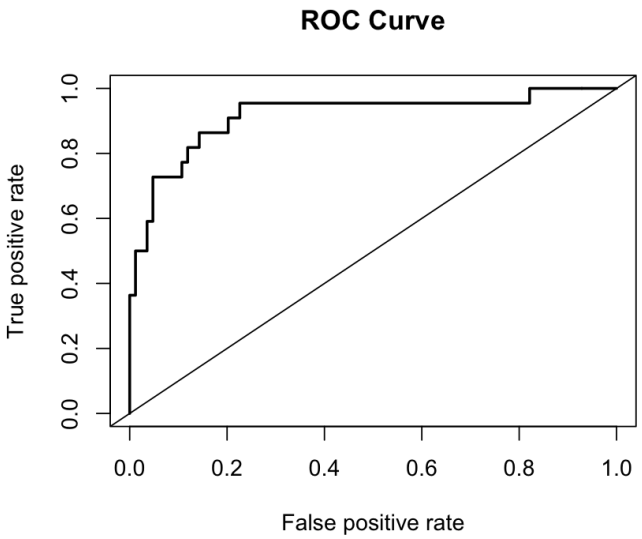
```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.16615    0.88893  -4.687 2.78e-06 ***
Age1               2.66433    0.39660   6.718 1.84e-11 ***
Age2               3.00235    0.40812   7.357 1.89e-13 ***
Age3               4.59928    0.45176  10.181  < 2e-16 ***
Gender1            0.20598    0.30525   0.675 0.499813
Family_Diabetes1   1.41820    0.23723   5.978 2.26e-09 ***
PhysicallyActive1  0.64612    0.32070   2.015 0.043932 *
PhysicallyActive2  1.11952    0.30980   3.614 0.000302 ***
PhysicallyActive3  0.68147    0.33161   2.055 0.039875 *
BMI                0.01402    0.02312   0.607 0.544091
Smoking1           0.89504    0.43902   2.039 0.041477 *
Alcohol1           0.11153    0.31922   0.349 0.726798
Sleep5            -1.46873    0.76730  -1.914 0.055599 .
Sleep6            -1.95286    0.67732  -2.883 0.003936 **
Sleep7            -2.08697    0.68760  -3.035 0.002404 **
Sleep8            -0.73345    0.66185  -1.108 0.267788
Sleep9            -1.88813    0.76075  -2.482 0.013067 *
JunkFood1          0.81699    0.36308   2.250 0.024440 *
JunkFood2          0.51197    0.46579   1.099 0.271698
JunkFood3          1.24980    0.63971   1.954 0.050738 .
BPLevel1           1.43605    0.27634   5.197 2.03e-07 ***
Pregancies1        0.48885    0.34728   1.408 0.159239
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We used our test data in order to test the accuracy of our model. Looking at the confusion matrix below, the value of true negatives and true positives are high. There was a 10.38% error rate, and only 8.05% of the negative values were incorrectly predicted, and only 21.05% of the positive values were incorrectly predicted. Therefore, we believe the full model is satisfactory and a reduced model was not needed.

Confusion Matrix:

	Predicted Not Diabetic	Predicted Diabetic
Actual Not Diabetic	80	7
Actual Diabetic	4	15

Rates	False Negative Rate	False Positive Rate	Accuracy	Error Rate	Sensitivity	Specificity
Logistic Regression	0.2105	0.0805	0.8962	0.1038	0.7895	0.9195



Additionally, we created an ROC curve from our logistic regression. Better classifiers would have a curve that ends towards the top-left corner, while worse classifiers lean towards the line in the middle (the random classifier). Seeing how close our ROC curve is to the top left corner furthers our belief that the full model is satisfactory.

Method 2: k-NN Classification

For our second method, where we use k-NN classification, we constructed two confusion matrices as discussed earlier. We also included a second confusion matrix in order to illustrate how R breaks classification ties at random, which means that when we are calculating our k-NN model, R will randomly assign variables to a category if they have an equal probability of being equal to one category or another. Although this difference does not appear to be very significant in classifying data values, we thought it was an important difference to highlight in this report.

Confusion Matrix with SoundSleep:

	Predicted Not Diabetic	Predicted Diabetic
Actual Not Diabetic	533	210
Actual Diabetic	26	31

Confusion Matrix without SoundSleep:

	Predicted Not Diabetic	Predicted Diabetic
Actual Not Diabetic	521	219
Actual Diabetic	38	22

Rates	False Negative Rate	False Positive Rate	Accuracy	Error Rate	Sensitivity	Specificity
With SoundSleep	0.4561	0.2826	0.705	0.295	0.5439	0.7174
Without SoundSleep	0.6333	0.2959	0.6788	0.3213	0.3667	0.7041

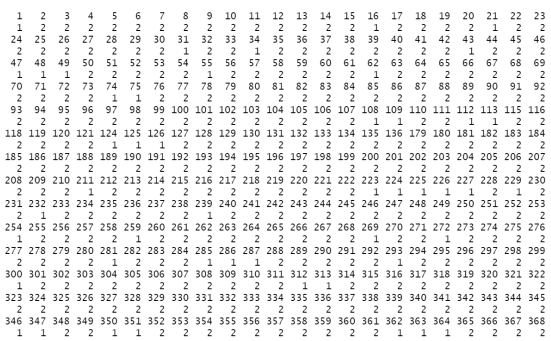
```
          DiabetesLabelTest
Pr.Class  no yes
      no  521 217
      yes  38  24
```

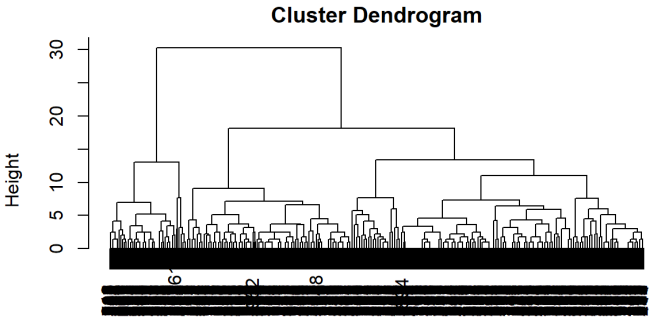
```
           DiabetesLabelTest2
Pr.Class2  no yes
       no  535 210
       yes  24  31
```

The k-NN model used for the calculations of the first confusion matrix contains only BMI, Sleep, and Pregnancies whereas the k-NN model used for the second confusion matrix contains SoundSleep as well as the other variables used in the first model. The model with SoundSleep has a correct classification rate of 70.5%, whereas the model without SoundSleep has a correct classification rate of 67.88%. This means that SoundSleep could actually be a descriptive indicator since it improves the classification rate. In our initial proposal, we did not include it in the list of quantitative variables to explore, thus it might be useful in determining a person's likelihood of developing diabetes. DiabetesLabelTest contains only BMI, Sleep, and Pregnancies whereas the DiabetesLabelTest2 contains SoundSleep as well as BMI, Sleep, and Pregnancy. The DiabetesLabelTest2 has a correct classification rate of 70.75%, whereas DiabetesLabelTest has a correct classification rate of 68.13%, which supports the idea that SoundSleep is an important descriptor in determining a person's likelihood of having type II diabetes.

Method 3: Cluster Analysis

As for the cluster analysis where we explore the effects of the variables in regards to the classification of groups, we obtain the following results.





From the first diagram, we see a list of all the patients along with their corresponding classification to one of two groups. The labels are blurred out at the bottom due to the sheer number of them, but they are unimportant as we are looking at the factors themselves. The classification, of course, depends on the distances among the numeric variables in question. From the dendrogram, we see at first that splitting the patients into two groups seems most reasonable.

A tibble: 2 x 2

healthclust <int>	mean(Pregancies) <dbl>
1	0.3358779
2	0.3948387

2 rows

A tibble: 2 x 2

healthclust <int>	mean(BMI) <dbl>
1	35.07634
2	23.89806

2 rows

A tibble: 2 x 2

healthclust <int>	mean(SoundSleep) <dbl>
1	4.145038
2	5.784516

2 rows

A tibble: 2 x 2

healthclust <int>	mean(Sleep) <dbl>
1	6.717557
2	6.997419

2 rows

For adults, the ideal BMI is between 18 to 24 and the ideal amount of sleep is close to 8. From this, we can conclude that group 1 is generally the “unhealthier” one. Pregnancies are close for both groups, so it does not describe much. Furthermore, SoundSleep, like BMI, sees a large deviation between the two groups, suggesting that group 1 is again generally more unhealthy. From these health classifications, we can try to predict how prone the patient could potentially be to diabetes.

A tibble: 4 x 3 Groups: healthclust [2]

healthclust <int>	Diabetic <chr>	n <int>
1	no	81
1	yes	50
2	no	562
2	yes	213

4 rows

Although they might not necessarily match up (i.e “healthy” patients may have diabetes and vice-versa), it would be useful to medical professionals to compare the groupings. As seen above, there are proportionally more patients with diabetes in the “unhealthy group” (group 1) than in the other one.

F. Conclusion

By using a combination of supervised and unsupervised models, we are able to draw some very interesting conclusions regarding the diabetes population. For example, using hierarchical clustering on our data set allowed us to see that people who have a high BMI (< 24) and people who do not get many hours of sound sleep (< 5.5 hours/night) are at a higher risk of developing diabetes. Additionally, we constructed k-NN and logistic regression models in an attempt to discover which model is better suited for predicting a person’s likelihood of developing diabetes. The performance of the logistic regression model and the k-NN model at predicting a person’s likelihood of diabetes demonstrates the feasibility of creating a model that predicts a person’s chances of developing type II diabetes. From the logistic regression, we were able to identify an individual’s age, family history of diabetes, blood pressure, and amount of physical activity as the predictors with high significance in putting them at a higher risk of having type II diabetes.

Is a k-NN model or logistic regression model more accurate at predicting whether someone will have diabetes, given their current health information?

Based on the information gathered from our models, we are able to conclude that a logistic regression model is better at predicting whether someone is at risk for diabetes. The logistic regression model was able to correctly classify if a person had diabetes 78.9% of the time with an error rate of 10.4% , compared to a 39% sensitivity rate from the k-NN model with an error rate of 31.8%.

How much interaction is there among some of the variables, and are some of them precursors to other variables?

The correlation matrix obtained earlier shows that with the exception of Sleep and SoundSleep, none of the variables have any statistically significant interactions with one another. The interaction observed between Sleep and SoundSleep exists because in order to sleep soundly, one must fall asleep.

Are there subgroups where the effects of some of the variables are different/opposite to what they were in the general model(s)?

According to the dplyr in the cluster analysis, it seems as though the algorithm was able to detect two different groups where one group’s BMI, SoundSleep, and Sleep was objectively healthier than that of the other group. Pregnancies remained questionable as they were too similar, and there is no objective unhealthiness.

G. Reference

Wu, Wenzhuo, et al. *Analysis of Forest Fires*. 8 Dec. 2021, https://canvas.ucdavis.edu/courses/682990/files/16770562?module_item_id=1237006.

"What Is the Body Mass Index (BMI)?" *NHS Choices*, NHS, 15 July 2019, https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/#:~:text=For%20most%20adults%2C%20an%20ideal,well%20as%20height%20and%20weight.

Holtz, Yan. "Correlation Matrix with GGALLY." – *The R Graph Gallery*, https://r-graph-gallery.com/199-correlation-matrix-with-ggally.html.

"Type 2 Diabetes." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 16 Dec. 2021, https://www.cdc.gov/diabetes/basics/type2.html#:~:text=If%20you%20have%20type%202,prediabetes%20and%20type%202%20diabetes.

Tigga, N. P., & Garg, S. (2020). "Diabetes Dataset 2019." *Kaggle*, 20 Oct. 2020, https://www.kaggle.com/datasets/tigganeha4/diabetes-dataset-2019

H. Appendix

```
#Data Visualization:
#ggpairs plot
DiabetesSet <- read.csv('C:/Users/halle/Documents/Downloads/diabetes.dataset__2019.csv')
Diabetes <- subset(DiabetesSet, select = -c(17,16,13,11,10,4))
DiabetesNew <- subset(Diabetes, select = -c(12)) #without Diabetes label
DiabetesQuant <- subset(DiabetesNew, select = c(5,8,11)) #Quantitative variables
library(ggplot2)
library(GGally)
ggpairs(DiabetesQuant) # for quantitative variables
```

```
# boxplots for qualitative predictors
# REFERENCED "Analysis of Forest Fires" project example
Diabetes$Class <- as.factor(ifelse(Diabetes$Diabetic == "yes", "Yes", "No"))
library(ggplot2)
library(tidyverse)
library(ggpubr)
library(plotly)
# creating classes
# BPLevel0 is low + normal, BPLevel1 is high
Diabetes$BPLevel = as.factor(ifelse(Diabetes$BPLevel == "low" | Diabetes$BPLevel == "normal", 0, 1))
Diabetes$JunkFood = as.factor(ifelse(Diabetes$JunkFood == "occasionally", 0, ifelse(Diabetes$JunkFood ==
#Pregnancies1 is 1-4 pregnancies
Diabetes$Alcohol = as.factor(ifelse(Diabetes$Alcohol == "no", 0, 1))
Diabetes$Smoking = as.factor(ifelse(Diabetes$Smoking == "no", 0, 1))
Diabetes$Family_Diabetes = as.factor(ifelse(Diabetes$Family_Diabetes== "no", 0, 1))
Diabetes$Gender = as.factor(ifelse(Diabetes$Gender == "Male", 0, 1))
Diabetes$Age = as.factor(ifelse(Diabetes$Age == "less than 40", 0, ifelse(Diabetes$Age == "40-49", 1, i
Diabetes$PhysicallyActive = as.factor(ifelse(Diabetes$PhysicallyActive == "less than half an hr", 0, if
A <- ggboxplot(Diabetes, x="Class", y="Age", color="Class",
               add="jitter", xlab="Diabetic", ylab="Age")
B <- ggboxplot(Diabetes, x="Class", y="Gender", color="Class",
               add="jitter", xlab="Diabetic", ylab="Gender")
C <- ggboxplot(Diabetes, x="Class", y="Family_Diabetes", color="Class",
               add="jitter", xlab="Diabetic", ylab="Family Diabetes History")
D <- ggboxplot(Diabetes, x="Class", y="PhysicallyActive", color="Class",
               add="jitter", xlab="Diabetic", ylab="Physical Activity")
E <- ggboxplot(Diabetes, x="Class", y="Smoking", color="Class",
               add="jitter", xlab="Diabetic", ylab="Smoking Status")
F <- ggboxplot(Diabetes, x="Class", y="Alcohol", color="Class",
               add="jitter", xlab="Diabetic", ylab="Alcohol Consump.")
G <- ggboxplot(Diabetes, x="Class", y="JunkFood", color="Class",
               add="jitter", xlab="Diabetic", ylab="Junk Food")
H <- ggboxplot(Diabetes, x="Class", y="BPLevel", color="Class",
               add="jitter", xlab="Diabetic", ylab="BP Level")
ggarrange(A,B,C,D,E,F,G,H,
          labels=c("A","B","C","D","E","F","G","H"),
          ncol=4, nrow=2)
```

```
#k-NN:
#Without Sound Sleep:
```

```r
rm(list=ls())
library(class)
DiabetesSet <- read.csv('C:/Users/halle/Documents/Downloads/diabetes.dataset__2019.csv')
DiabetesSet <- na.omit(DiabetesSet)

Diabetes <- subset(DiabetesSet, select = -c(17,16,13,11,10,4))
DiabetesNew <- subset(Diabetes, select = -c(12)) #without Diabetes label
DiabetesQuant <- subset(DiabetesNew, select = c(5,8,11)) #Quantitative variables

TestData <- DiabetesQuant[1:800, ]
TrainData <- DiabetesQuant[801:nrow(DiabetesQuant), ]

DiabetesLabelTest <- subset(Diabetes, select = c(12))[1:800, ]
DiabetesLabelTrain <- subset(Diabetes, select = c(12))[801:nrow(Diabetes), ]
#Function takes training and testing data, along with their from a k-NN model to create a confusion mat
require(class)
confusion.matrix <- function(train, test, label, testlabel, k, n){
  label1 <- c()
  for (i in 1:length(label)){
    label1[i] <- ifelse(label[i] == "yes", 1, 0)
  }
  pr.class <- knn(train[,1:n], test[,1:n], label1, k, prob = F )
  x <- table(pr.class, testlabel, dnn = c("Predicted", "Actual"))
  FNR <- x[2,1] / (x[2,2] + x[2,1])
  FPR <- x[1,2] / (x[1,2] + x[1,1])
  accuracy <- (x[2,2] + x[1,1]) / ((x[1,1] + x[1,2] + x[2,1] + x[2,2]))
  error_rate <- (x[1,2] + x[2,1]) / ((x[1,1] + x[1,2] + x[2,1] + x[2,2]))
  sensitivity <- 1 - FNR
  specificity <- 1 - FPR
  rates <- data.frame(FNR, FPR, accuracy, error_rate, sensitivity, specificity)
  colnames(rates) <- c("False Negative Rate", "False Positive Rate", "Accuracy", "Error Rate", "Sensiti
  output <- list(x, rates)
  return(output)
}
confus <- confusion.matrix(TrainData, TestData, DiabetesLabelTrain, DiabetesLabelTest, 3, 3)

Pr.Class <- knn(TrainData[,1:3],TestData[,1:3],DiabetesLabelTrain,k = 3, prob=F)
df_new <- cbind(DiabetesLabelTest, Pr.Class)
table(Pr.Class,DiabetesLabelTest)



#With Sound Sleep:
DiabetesSet2 <- na.omit(DiabetesSet)
Diabetes2 <- subset(DiabetesSet2, select = -c(17,16,13,11,4))
DiabetesNew2 <- subset(Diabetes2, select = -c(13)) #without Diabetes label
DiabetesQuant2 <- subset(DiabetesNew2, select = c(5,8,9,12)) #Quantitative variables

TestData2 <- DiabetesQuant2[1:800, ]
TrainData2 <- DiabetesQuant2[801:nrow(DiabetesQuant2), ]

DiabetesLabelTest2 <- subset(Diabetes2, select = c(13))[1:800, ]
DiabetesLabelTrain2 <- subset(Diabetes2, select = c(13))[801:nrow(Diabetes2), ]
```

```
confus2 <- confusion.matrix(TrainData2, TestData2, DiabetesLabelTrain, DiabetesLabelTest, 3, 4)


#logistic regression
rm(list=ls())
library(class)
DiabetesSet <- read.csv('C:/Users/halle/Documents/Downloads/diabetes.dataset__2019.csv')
DiabetesSet <- na.omit(DiabetesSet)
DiabetesNew <- subset(DiabetesSet, select = -c(17,16,13,11,4))
DiabetesNew$Diabetic = as.factor(ifelse(DiabetesNew$Diabetic == "no", 0, 1))
DiabetesNew$Pregancies = as.factor(DiabetesNew$Pregancies)
# BPLevel0 is low + normal, BPLevel1 is high
DiabetesNew$BPLevel = as.factor(ifelse(DiabetesNew$BPLevel == "low" | DiabetesNew$BPLevel == "normal", 
DiabetesNew$JunkFood = as.factor(ifelse(DiabetesNew$JunkFood == "occasionally", 0, ifelse(DiabetesNew$Ju
# Sleep9 is 9hrs+
DiabetesNew$Sleep = as.factor(ifelse(DiabetesNew$Sleep == "9" | DiabetesNew$Sleep == "10" | DiabetesNew$
#Pregnancies1 is 1-4 pregnancies
DiabetesNew$Pregancies = as.factor(ifelse(DiabetesNew$Pregancies == "0", 0, 1))
DiabetesNew$Alcohol = as.factor(ifelse(DiabetesNew$Alcohol == "no", 0, 1))
DiabetesNew$Smoking = as.factor(ifelse(DiabetesNew$Smoking == "no", 0, 1))
DiabetesNew$PhysicallyActive = as.factor(ifelse(DiabetesNew$PhysicallyActive == "less than half an hr",
DiabetesNew$Family_Diabetes = as.factor(ifelse(DiabetesNew$Family_Diabetes== "no", 0, 1))
DiabetesNew$Gender = as.factor(ifelse(DiabetesNew$Gender == "Male", 0, 1))
DiabetesNew$Age = as.factor(ifelse(DiabetesNew$Age == "less than 40", 0, ifelse(DiabetesNew$Age == "40-
lr_train <- DiabetesNew[1:800,]
lr_train_diab <- DiabetesNew[1:800, c(13)]
lr_test <- DiabetesNew[801:nrow(DiabetesNew),-c(13)]
lr_test_diab <- DiabetesNew[801:nrow(DiabetesNew),c(13)]
log_mod = glm(Diabetic ~ Age + Gender + Family_Diabetes + PhysicallyActive + BMI + Smoking + Alcohol + S
summary(log_mod)

y_pred_glm = predict(log_mod, lr_test, type = 'response')

confusion <- table(ifelse(y_pred_glm > 0.55, 1, 0), lr_test_diab, dnn=c("Actual", "Predicted"))
confusion
error <- function(x){
  errorrate = (1-sum(diag(x))/sum(x))
  FNR <- x[2,1] / (x[2,2] + x[2,1])
  FPR <- x[1,2] / (x[1,2] + x[1,1])
  accuracy <- (x[2,2] + x[1,1]) / ((x[1,1] + x[1,2] + x[2,1] + x[2,2]))
  error_rate <- (x[1,2] + x[2,1]) / ((x[1,1] + x[1,2] + x[2,1] + x[2,2]))
  sensitivity <- 1 - FNR
  specificity <- 1 - FPR
  rates <- data.frame(FNR, FPR, accuracy, error_rate, sensitivity, specificity)
  colnames(rates) <- c("False Negative Rate", "False Positive Rate", "Accuracy", "Error Rate", "Sensiti
  output <- list(x, rates)
  return(output)
}
error(confusion)

# ROC curve code
library(ROCR)
y_roc = prediction(y_pred_glm, lr_test_diab)
```

```
roc_curve = performance(y_roc,"tpr","fpr")
plot(roc_curve, lwd = 2, main= "ROC Curve")
abline(a = 0, b = 1)
```

```
#cluster analysis
rm(list=ls())
require(ISLR)
require(tidyverse)
require(ggthemes)
DiabetesSet <- read.csv('C:/Users/halle/Documents/Downloads/diabetes.dataset__2019.csv')
DiabetesSet2 <- na.omit(DiabetesSet)
Diabetes2 <- subset(DiabetesSet2, select = -c(17,16,13,11,4))
DiabetesQuant2 <- subset(Diabetes2, select = c(5,8,9,12,13)) #Quantitative variables
Diab_dist <- dist(DiabetesQuant2)
Diab_hclust <- hclust(Diab_dist)
diabplot <- plot(Diab_hclust)
```

```
healthclust <- cutree(Diab_hclust,k=2)

cbind(DiabetesQuant2, healthclust) %>%
  group_by(healthclust) %>%
  summarize(mean(BMI))
cbind(DiabetesQuant2, healthclust) %>%
  group_by(healthclust) %>%
  summarize(mean(Pregancies))
cbind(DiabetesQuant2, healthclust) %>%
  group_by(healthclust) %>%
  summarize(mean(SoundSleep))
cbind(DiabetesQuant2, healthclust) %>%
  group_by(healthclust) %>%
  summarize(mean(Sleep))
cbind(DiabetesQuant2, healthclust) %>%
  group_by(healthclust) %>%
  count(Diabetic)
```