K-Nearest Neighbor - Medical Dataset using R

#We will begin by importing packages that will be used for our data analysis.

library(caret)

## Warning: package 'caret' was built under R version 4.3.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.3.3

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 4.3.3

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'lubridate' was built under R version 4.3.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ lubridate 1.9.3 ✔ tibble 3.2.1  
## ✔ purrr 1.0.1 ✔ tidyr 1.3.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ purrr::lift() masks caret::lift()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(caTools)

## Warning: package 'caTools' was built under R version 4.3.3

#Importing our medical dataset into R.

mc <- read.csv("C:/Users/Tyler Bier/Desktop/D209 Docs/Original Files/medical\_clean.csv")

#We will start with the data cleaning/ preparation process. We will first check for duplicate rows.

anyDuplicated(mc)

## [1] 0

sum(duplicated(mc))

## [1] 0

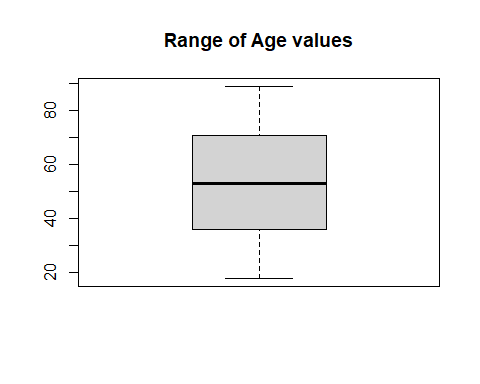
#Based off of these functions, we can see that we do not have any duplicate data rows in our dataset. #Our next step in the data cleaning process is to check for missing values, per function below to check missing values per column.

colSums(is.na(mc))

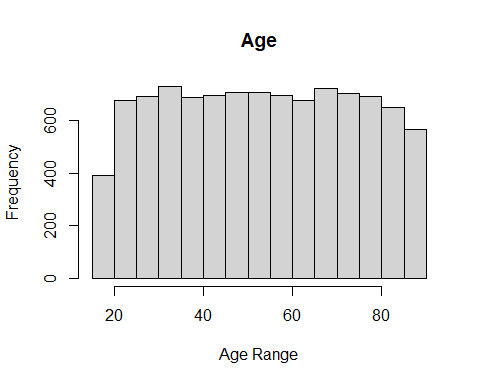
## CaseOrder Customer\_id Interaction UID   
## 0 0 0 0   
## City State County Zip   
## 0 0 0 0   
## Lat Lng Population Area   
## 0 0 0 0   
## TimeZone Job Children Age   
## 0 0 0 0   
## Income Marital Gender ReAdmis   
## 0 0 0 0   
## VitD\_levels Doc\_visits Full\_meals\_eaten vitD\_supp   
## 0 0 0 0   
## Soft\_drink Initial\_admin HighBlood Stroke   
## 0 0 0 0   
## Complication\_risk Overweight Arthritis Diabetes   
## 0 0 0 0   
## Hyperlipidemia BackPain Anxiety Allergic\_rhinitis   
## 0 0 0 0   
## Reflux\_esophagitis Asthma Services Initial\_days   
## 0 0 0 0   
## TotalCharge Additional\_charges Item1 Item2   
## 0 0 0 0   
## Item3 Item4 Item5 Item6   
## 0 0 0 0   
## Item7 Item8   
## 0 0

#As is shown above, we do not have any missing values in our dataset. #Continuing on with data cleaning process, we will examine data in numeric variables for any potential outliers. The only numeric variable we will be using for this analysis is: Age.

AgeBox <- mc[,c(16)]  
boxplot((AgeBox), main= "Range of Age values", outcol= "blue")



hist(AgeBox, main= "Age", xlab="Age Range")

 #With the above visualizations, we can see that our age column has no major outliers and a relatively uniform data distribution. #For our KNN model, we must convert categorical variables into a numeric format. To start, we will change all binary categorical variables from Yes/No values to 1/0. We will place all adjusted data into a new dataset.

modelmc <-  
mc %>%  
mutate(ReAdmis=ifelse(ReAdmis=="Yes",1,0),  
HighBlood=ifelse(HighBlood == "Yes", 1, 0),  
Stroke=ifelse(Stroke=="Yes",1,0),  
Arthritis=ifelse(Arthritis=="Yes",1,0),  
Diabetes=ifelse(Diabetes=="Yes",1,0),  
Hyperlipidemia=ifelse(Hyperlipidemia=="Yes",1,0),  
BackPain=ifelse(BackPain=="Yes",1,0),  
Allergic\_rhinitis=ifelse(Allergic\_rhinitis=="Yes",1,0), Reflux\_esophagitis=ifelse(Reflux\_esophagitis=="Yes",1,0),  
Asthma=ifelse(Asthma=="Yes",1,0),  
Soft\_drink=ifelse(Soft\_drink=="Yes",1,0),  
Anxiety=ifelse(Anxiety=="Yes",1,0), Overweight=ifelse(Overweight=="Yes",1,0))

#We will use a quick summary function to ensure categorical columns to be used in model have been transformed into binary (1/0) numerical values.

summary(modelmc)

## CaseOrder Customer\_id Interaction UID   
## Min. : 1 Length:10000 Length:10000 Length:10000   
## 1st Qu.: 2501 Class :character Class :character Class :character   
## Median : 5000 Mode :character Mode :character Mode :character   
## Mean : 5000   
## 3rd Qu.: 7500   
## Max. :10000   
## City State County Zip   
## Length:10000 Length:10000 Length:10000 Min. : 610   
## Class :character Class :character Class :character 1st Qu.:27592   
## Mode :character Mode :character Mode :character Median :50207   
## Mean :50159   
## 3rd Qu.:72412   
## Max. :99929   
## Lat Lng Population Area   
## Min. :17.97 Min. :-174.21 Min. : 0.0 Length:10000   
## 1st Qu.:35.26 1st Qu.: -97.35 1st Qu.: 694.8 Class :character   
## Median :39.42 Median : -88.40 Median : 2769.0 Mode :character   
## Mean :38.75 Mean : -91.24 Mean : 9965.2   
## 3rd Qu.:42.04 3rd Qu.: -80.44 3rd Qu.: 13945.0   
## Max. :70.56 Max. : -65.29 Max. :122814.0   
## TimeZone Job Children Age   
## Length:10000 Length:10000 Min. : 0.000 Min. :18.00   
## Class :character Class :character 1st Qu.: 0.000 1st Qu.:36.00   
## Mode :character Mode :character Median : 1.000 Median :53.00   
## Mean : 2.097 Mean :53.51   
## 3rd Qu.: 3.000 3rd Qu.:71.00   
## Max. :10.000 Max. :89.00   
## Income Marital Gender ReAdmis   
## Min. : 154.1 Length:10000 Length:10000 Min. :0.0000   
## 1st Qu.: 19598.8 Class :character Class :character 1st Qu.:0.0000   
## Median : 33768.4 Mode :character Mode :character Median :0.0000   
## Mean : 40490.5 Mean :0.3669   
## 3rd Qu.: 54296.4 3rd Qu.:1.0000   
## Max. :207249.1 Max. :1.0000   
## VitD\_levels Doc\_visits Full\_meals\_eaten vitD\_supp   
## Min. : 9.806 Min. :1.000 Min. :0.000 Min. :0.0000   
## 1st Qu.:16.626 1st Qu.:4.000 1st Qu.:0.000 1st Qu.:0.0000   
## Median :17.951 Median :5.000 Median :1.000 Median :0.0000   
## Mean :17.964 Mean :5.012 Mean :1.001 Mean :0.3989   
## 3rd Qu.:19.348 3rd Qu.:6.000 3rd Qu.:2.000 3rd Qu.:1.0000   
## Max. :26.394 Max. :9.000 Max. :7.000 Max. :5.0000   
## Soft\_drink Initial\_admin HighBlood Stroke   
## Min. :0.0000 Length:10000 Min. :0.000 Min. :0.0000   
## 1st Qu.:0.0000 Class :character 1st Qu.:0.000 1st Qu.:0.0000   
## Median :0.0000 Mode :character Median :0.000 Median :0.0000   
## Mean :0.2575 Mean :0.409 Mean :0.1993   
## 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.000 Max. :1.0000   
## Complication\_risk Overweight Arthritis Diabetes   
## Length:10000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## Class :character 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Mode :character Median :1.0000 Median :0.0000 Median :0.0000   
## Mean :0.7094 Mean :0.3574 Mean :0.2738   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000   
## Hyperlipidemia BackPain Anxiety Allergic\_rhinitis  
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.3372 Mean :0.4114 Mean :0.3215 Mean :0.3941   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## Reflux\_esophagitis Asthma Services Initial\_days   
## Min. :0.0000 Min. :0.0000 Length:10000 Min. : 1.002   
## 1st Qu.:0.0000 1st Qu.:0.0000 Class :character 1st Qu.: 7.896   
## Median :0.0000 Median :0.0000 Mode :character Median :35.836   
## Mean :0.4135 Mean :0.2893 Mean :34.455   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:61.161   
## Max. :1.0000 Max. :1.0000 Max. :71.981   
## TotalCharge Additional\_charges Item1 Item2   
## Min. :1938 Min. : 3126 Min. :1.000 Min. :1.000   
## 1st Qu.:3179 1st Qu.: 7986 1st Qu.:3.000 1st Qu.:3.000   
## Median :5214 Median :11574 Median :4.000 Median :3.000   
## Mean :5312 Mean :12935 Mean :3.519 Mean :3.507   
## 3rd Qu.:7460 3rd Qu.:15626 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :9181 Max. :30566 Max. :8.000 Max. :7.000   
## Item3 Item4 Item5 Item6   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000   
## Median :4.000 Median :4.000 Median :3.000 Median :4.000   
## Mean :3.511 Mean :3.515 Mean :3.497 Mean :3.522   
## 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :8.000 Max. :7.000 Max. :7.000 Max. :7.000   
## Item7 Item8   
## Min. :1.000 Min. :1.00   
## 1st Qu.:3.000 1st Qu.:3.00   
## Median :3.000 Median :3.00   
## Mean :3.494 Mean :3.51   
## 3rd Qu.:4.000 3rd Qu.:4.00   
## Max. :7.000 Max. :7.00

# To continue our data preparation for model, let’s examine our categorical variable that contains more than Yes/No values, “Gender.”

table(mc$Gender)

##   
## Female Male Nonbinary   
## 5018 4768 214

#With three categorical values in Gender, we will be creating dummy variables to convert each value into a numerical, binary form. Will not be dropping dummy variables as we are using a KNN model.

dummy\_gender <- dummyVars("~ Gender", data= modelmc)  
df\_dummy <- predict(dummy\_gender, newdata= modelmc)  
df\_dummy <- as.data.frame(df\_dummy)  
finalmc <- cbind(modelmc, df\_dummy)  
view(finalmc)

#Gender has now been converted into 3 binary 1/0 columns. Confirming by visualization using view function. #We will now subset our database, removing all variables except for those that will be used by our model.

finalmc <- subset(finalmc, select = c(ReAdmis, Age, GenderFemale, GenderMale, GenderNonbinary, Soft\_drink, HighBlood, Stroke, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Asthma))  
view(finalmc)

# Now will save our final, prepared dataset as a csv file.

write.csv(finalmc, file="C:/Users/Tyler Bier/Desktop/D209 Docs/Task1PAfiles/PreparedDatasetRevision.csv")

#Now that data is prepared. We can begin building and running our KNN model. # Splitting the data into both training and test set and making sure our outcome column is converted into a factor. We will also save versions of our testing/ training data as a csv file.

finalmc$ReAdmis <- factor(finalmc$ReAdmis, levels = c(1,0))  
set.seed(123)  
trainingdata <- createDataPartition(finalmc$ReAdmis, p=.7, list = FALSE, times = 1)  
trainReadmis <- finalmc[trainingdata,]  
testReadmis <- finalmc[-trainingdata,]  
write.csv(trainReadmis, file="C:/Users/Tyler Bier/Desktop/D209 Docs/Task1PAfiles/TrainingData.csv")  
write.csv(testReadmis, file="C:/Users/Tyler Bier/Desktop/D209 Docs/Task1PAfiles/TestingData.csv")

#Creating the KNN model with assistance from caret package.

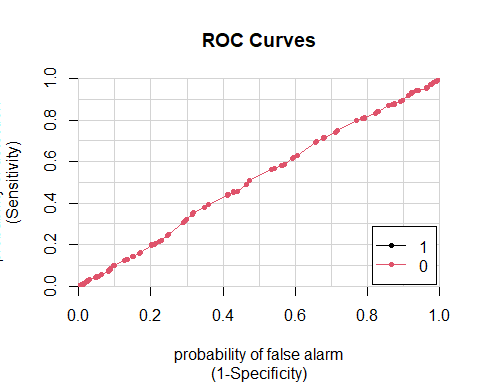
knnmodel <- train(ReAdmis ~ Age + GenderFemale + GenderMale + GenderNonbinary + Soft\_drink + HighBlood + Stroke + Overweight + Arthritis + Diabetes + Hyperlipidemia + BackPain + Anxiety + Allergic\_rhinitis + Reflux\_esophagitis + Asthma, data= trainReadmis, method= "knn", preProcess= c("range"), tuneLength = 20)

#Model is built. #Next we will review our model’s functionality/ accuracy using a confusion matrix and calculating AUC value and creating a ROC plot.

predictconfusion <- predict(knnmodel, newdata = testReadmis)  
predictprobabilities <- predict(knnmodel, newdata = testReadmis, type="prob")  
cm <- confusionMatrix(predictconfusion, testReadmis$ReAdmis)  
print(cm)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 0  
## 1 31 63  
## 0 1069 1836  
##   
## Accuracy : 0.6225   
## 95% CI : (0.6049, 0.6399)  
## No Information Rate : 0.6332   
## P-Value [Acc > NIR] : 0.8908   
##   
## Kappa : -0.0062   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.02818   
## Specificity : 0.96682   
## Pos Pred Value : 0.32979   
## Neg Pred Value : 0.63201   
## Prevalence : 0.36679   
## Detection Rate : 0.01034   
## Detection Prevalence : 0.03134   
## Balanced Accuracy : 0.49750   
##   
## 'Positive' Class : 1   
##

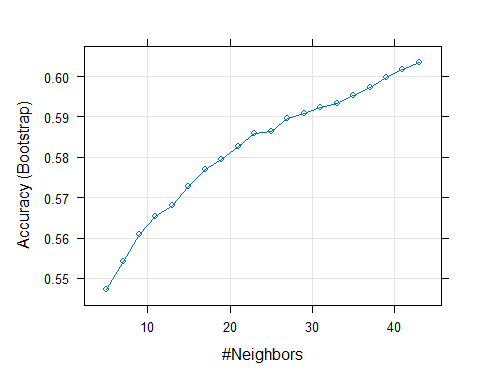
knnauc <- colAUC(predictprobabilities, testReadmis$ReAdmis, plotROC = TRUE )



print(knnauc)

## 1 0  
## 1 vs. 0 0.5144363 0.5144363

plot(knnmodel)

 # Conclusion of classification analysis.