My_own_Project

Yassin Zeraoulia

21 09 2021

INTRODUCTION

Stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, heart diseases, smoking status and other relevant clinical conditions. Each row in the data provides some information about the patient.

VAriable inside the data set: 1) id: unique identifier 2) gender: "Male", "Female" or "Other" 3) age: age of the patient 4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension 5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease 6) ever_married: "No" or "Yes" 7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed" 8) Residence_type: "Rural" or "Urban" 9) avg_glucose_level: average glucose level in blood 10) bmi: body mass index 11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"* 12) stroke: 1 if the patient had a stroke or 0 if not

In order to predict if a person is more likely to have a stroke we'll have to take into consideration relevant variables of the data set, some of them are not relevant because they don't have impact on the physiology of the human being. Furthermore we'll have to explore our data set and gain some insight with the use of plots and correlation table. After the exploration of the data set we can make some assumption on the relevance of each variable and start thinking about a model able to fit correctly our train data once we divided our data set into train a and test set.

The first model that will be build is based on a classification tree model by wich I'll try to estimate if a patiant has high glucose levels (glucose_cat_tr) using as predictors: "stroke", "hypertension", "heart_disease","bmi_cat_tr".

The second model that will be implemented is a logistic regression used to predict strokes, this model uses as predictors avg_glucose_level, age, heart_disease, and hypertension.

```
#Loading the data set and libraries

library(tidyverse)

library(dslabs)
library(ggplot2)
library(caret)

library(readr)
library(dplyr)
library(corrplot)

library(reshape2)
```

```
library(ggplot2)
library(rsample)library(caret)
library(data.table)

library(FactoMineR)

library(viridis)

library(rattle)

library(mice)

library(VIM)

library(VGAM)

library(pROC)

#data exploration

healthcare_dataset_stroke_data <- read_csv("healthcare-dataset-stroke-data.csv")

##
## -- Column specification --------------------------------------------------------
-
## cols(
##   id = col_double(),
##   gender = col_character(),
##   age = col_double(),
##   hypertension = col_double(),
##   heart_disease = col_double(),
##   ever_married = col_character(),
##   work_type = col_character(),
##   Residence_type = col_character(),
##   avg_glucose_level = col_double(),
##   bmi = col_character(),
##   smoking_status = col_character(),
##   stroke = col_double()
## )

data <- as.data.frame(healthcare_dataset_stroke_data)

str(data)

## 'data.frame':    5110 obs. of  12 variables:
##  $ id               : num  9046 51676 31112 60182 1665 ...
##  $ gender           : chr  "Male" "Female" "Male" "Female" ...
##  $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : num  0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : num  1 0 1 0 0 0 1 0 0 0 ...
##  $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ work_type        : chr  "Private" "Self-employed" "Private" "Private" ...
##  $ Residence_type   : chr  "Urban" "Rural" "Rural" "Urban" ...
```

```
##  $ avg_glucose_level: num  229 202 106 171 174 ...
##  $ bmi              : chr  "36.6" "N/A" "32.5" "34.4" ...
##  $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smo
kes" ...
##  $ stroke           : num  1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   id = col_double(),
##   ..   gender = col_character(),
##   ..   age = col_double(),
##   ..   hypertension = col_double(),
##   ..   heart_disease = col_double(),
##   ..   ever_married = col_character(),
##   ..   work_type = col_character(),
##   ..   Residence_type = col_character(),
##   ..   avg_glucose_level = col_double(),
##   ..   bmi = col_character(),
##   ..   smoking_status = col_character(),
##   ..   stroke = col_double()
##   .. )

data <- unique(data)
data <- na.omit(data)

knitr::opts_chunk$set(echo = TRUE)
```

We start by removing unnecessary variable inside the dataset and turning some variables in a preferable data type.

```
#Remove unnecessary columns from the dataset and changing some data type

data <- data %>% select( "gender", "age", "hypertension", "heart_disease", "Reside
nce_type","avg_glucose_level", "bmi", "smoking_status","stroke")


data$gender <- as.factor(data$gender)
data$smoking_status <- as.factor(data$smoking_status)
data$Residence_type <- as.factor(data$Residence_type)


knitr::opts_chunk$set(echo = TRUE)
```

DATA EXPLORATION

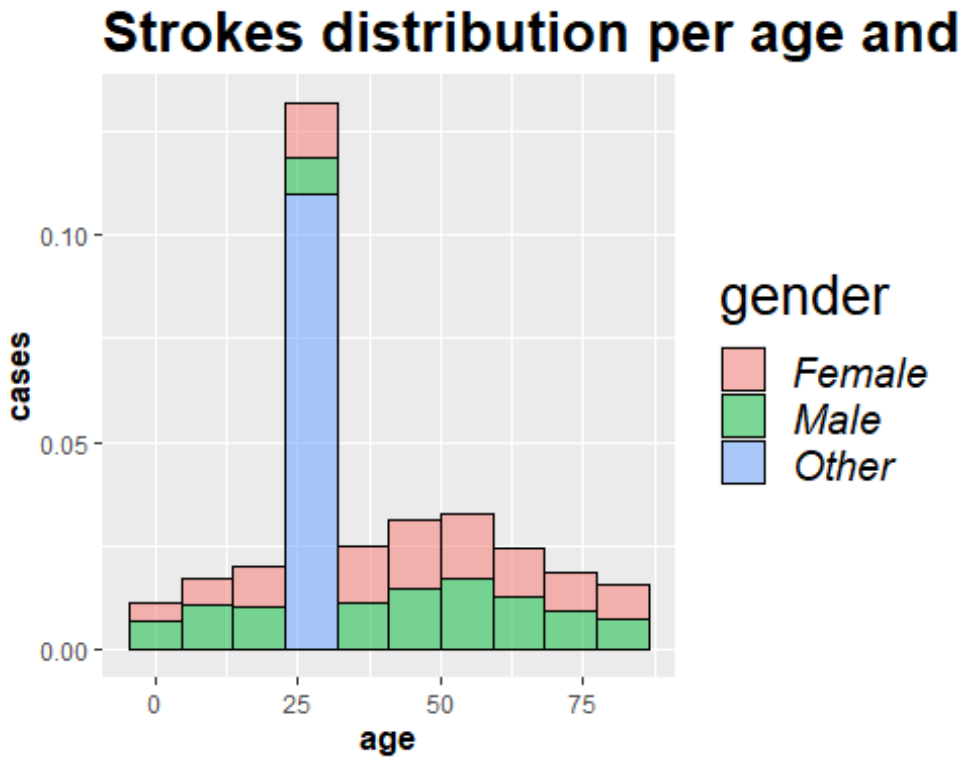
Now we can gain some initial insight on the variables inside the data set by generating some plots and analizyng distributions of the data set.

The first plot gives us some idea on the distribution of strokes by age and gender, is clear that there is a problem in the data points gathered because around the age of 25 there is an increase of strokes in the “other” gender category. This can have an impact on the estimation of correlation between variables and the performance of the models that will be built.

The second plot rappresent the correlation between tree variables: heart disease, hypertension and age.

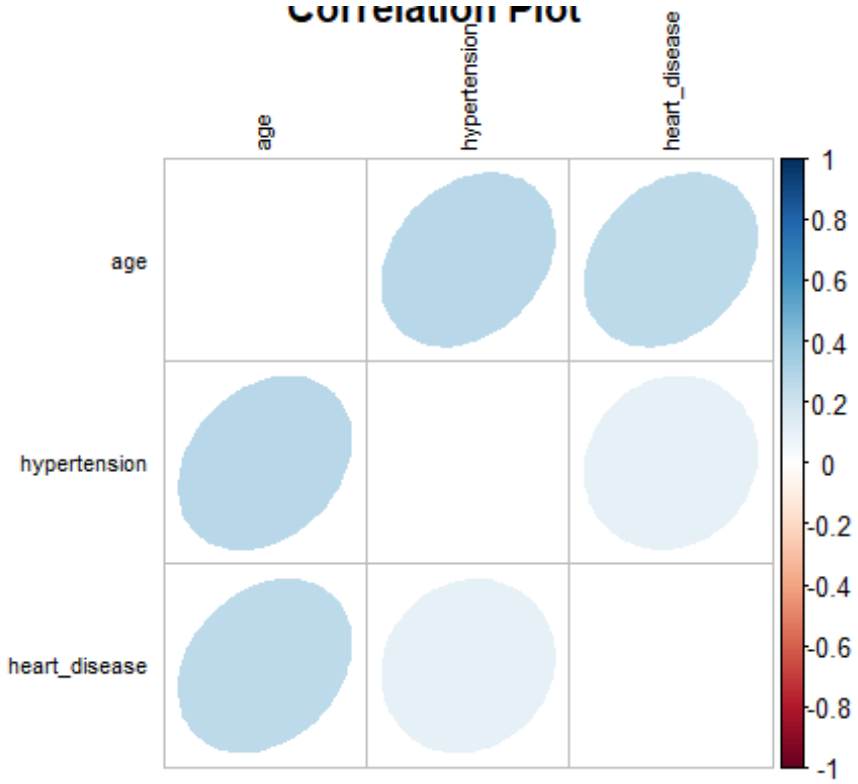
The third plot is a correlation between bmi and avg_glucose_level.

```
#Data exploration

data %>% ggplot(aes(age,fill = gender)) +
  geom_histogram(alpha =0.5, aes(y = ..density..),col="black", bins = 10) +
  theme(legend.title=element_text(family="Times",size=20),
        legend.text=element_text(family="Times",face ="italic",size=15),
        plot.title=element_text(family="Times", face="bold", size=20),
        axis.title.x=element_text(family="Times", face="bold", size=12),
        axis.title.y=element_text(family="Times", face="bold", size=12)) +
  xlab("age") +
  ylab("cases") +
  ggtitle("Strokes distribution per age and gender")
```

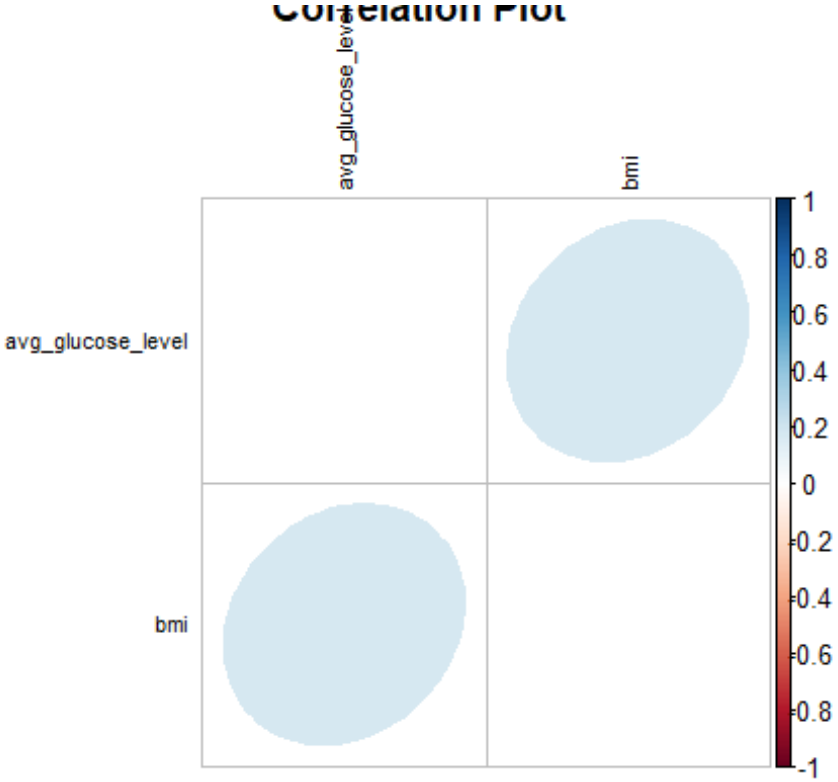



```
data$avg_glucose_level <- as.numeric(data$avg_glucose_level)
data$bmi <- as.numeric(data$bmi)

## Warning: NAs durch Umwandlung erzeugt

#as we can see there is some  positive correlation between age, hypertention and h
eart disease
#From this plot the correletion dosen't seems to bee as high as expected

corrplot(cor(data[,2:4],method ="pearson"),diag =FALSE,
         title ="Correlation Plot", method ="ellipse",
```

```
         tl.cex =0.7, tl.col ="black", cl.ratio =0.2
)
```




```
#as expected even between bmi and glucose levels there is a slightly positive corr
elation

corrplot(cor(data[,6:7],method ="pearson", use = "na.or.complete"),diag =FALSE,
         title ="Correlation Plot", method ="ellipse",
         tl.cex =0.7, tl.col ="black", cl.ratio =0.2
)
```




```
knitr::opts_chunk$set(echo = TRUE)
```

The variable describing the smoking status could be really helpful, the problem here is that the percentage of strokes in the data set is really small so is not really easy to notice a correlation between stroke and smoking status. Some patiant who had a stroke were not smokers, were not old, had no heart disease or hypertension; this make building a predictive model using this data set not really easy.

```
data %>% ggplot(aes(stroke, fill = smoking_status)) +
  geom_bar(alpha =2, col="black") +
  theme(legend.title=element_text(family="Times",size=20),
        legend.text=element_text(family="Times",face ="italic",size=15),
        plot.title=element_text(family="Times", face="bold", size=20),
        axis.title.x=element_text(family="Times", face="bold", size=12),
        axis.title.y=element_text(family="Times", face="bold", size=12)) +
  xlab("stroke") +
  ylab("cases") +
  ggtitle("Strokes and smoking status")
```

```
data %>% ggplot(aes(age, fill = smoking_status)) +
  geom_density(alpha =0.5, stat = "count", na.rm = TRUE, width = 5, position = "st
ack") +
  theme(legend.title=element_text(family="Times",size=20),
        legend.text=element_text(family="Times",face ="italic",size=15),
        plot.title=element_text(family="Times", face="bold", size=20),
        axis.title.x=element_text(family="Times", face="bold", size=12),
```

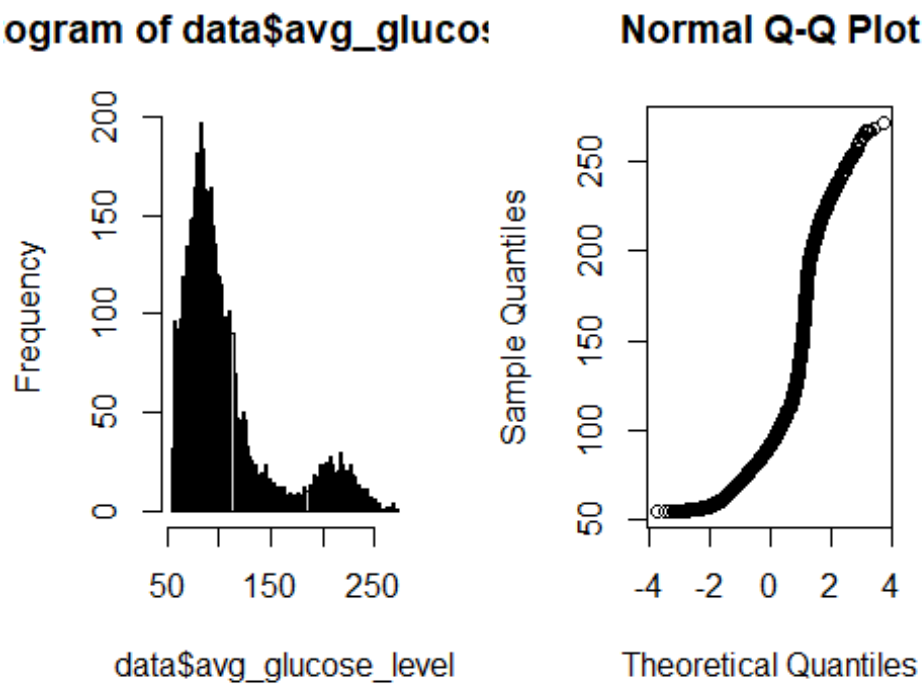
```
        axis.title.y=element_text(family="Times", face="bold", size=12)) +
  xlab("age") +
  ylab("cases") +
  ggtitle("Age and smoking status")

knitr::opts_chunk$set(echo = TRUE)
```

The distribution of bmi is right skewed (long tail to the right). Because this is the only variable with missing data (at least of the numerical variables) we can impute the median on the missing data without losing too much information.

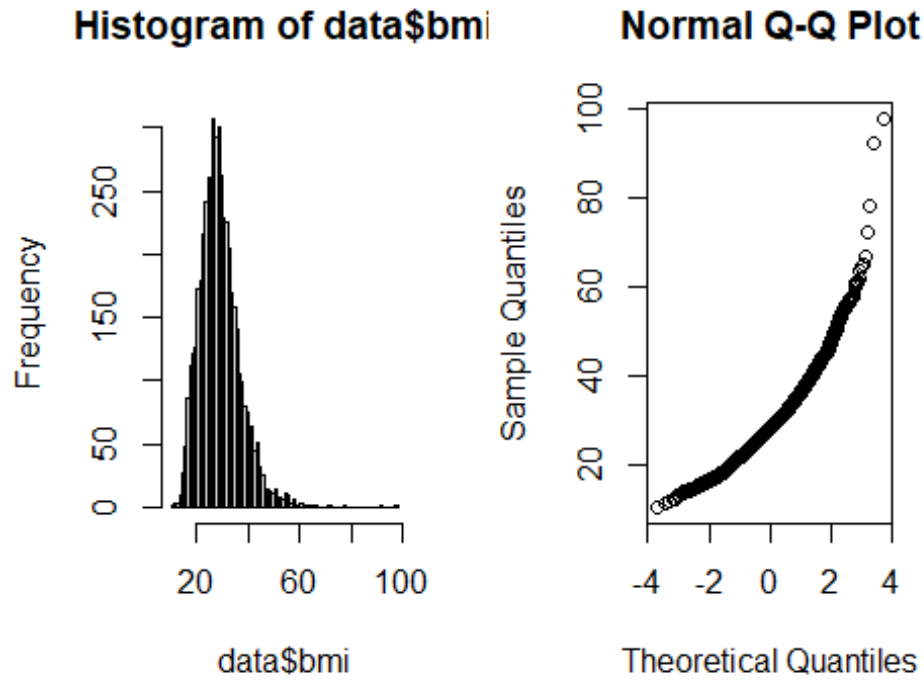
Only 5% of the people inside the data set had a stroke, This means that our baseline dummy model has an accuracy of 95%.That is if we would predict a person to not have a stroke all the time.

The distribution of avg_glucose_level and bmi dosen't seem to be normal as we cann asses with plot and the Shapiro-Wilk normality test.

```
data <- as.data.table(data)

str(data)

## Classes 'data.table' and 'data.frame':   5110 obs. of  9 variables:
##  $ gender           : Factor w/ 3 levels "Female","Male",..: 2 1 2 1 1 2 2 1 1
1 ...
##  $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : num  0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : num  1 0 1 0 0 0 1 0 0 0 ...
##  $ Residence_type   : Factor w/ 2 levels "Rural","Urban": 2 1 1 2 1 2 1 2 1 2 .
..
##  $ avg_glucose_level: num  229 202 106 171 174 ...
##  $ bmi              : num  36.6 NA 32.5 34.4 24 29 27.4 22.8 NA 24.2 ...
##  $ smoking_status   : Factor w/ 4 levels "formerly smoked",..: 1 2 2 3 2 1 2 2
4 4 ...
##  $ stroke           : num  1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>

#The histogram of glucose leveles and bmi is not normally distributed in a
#traditional bell-shape and the Q-Q plot poorly resembles a straight y = x line.

par(mfrow = c(1,2))
hist(data$avg_glucose_level, 100)
qqnorm(data$avg_glucose_level)
```



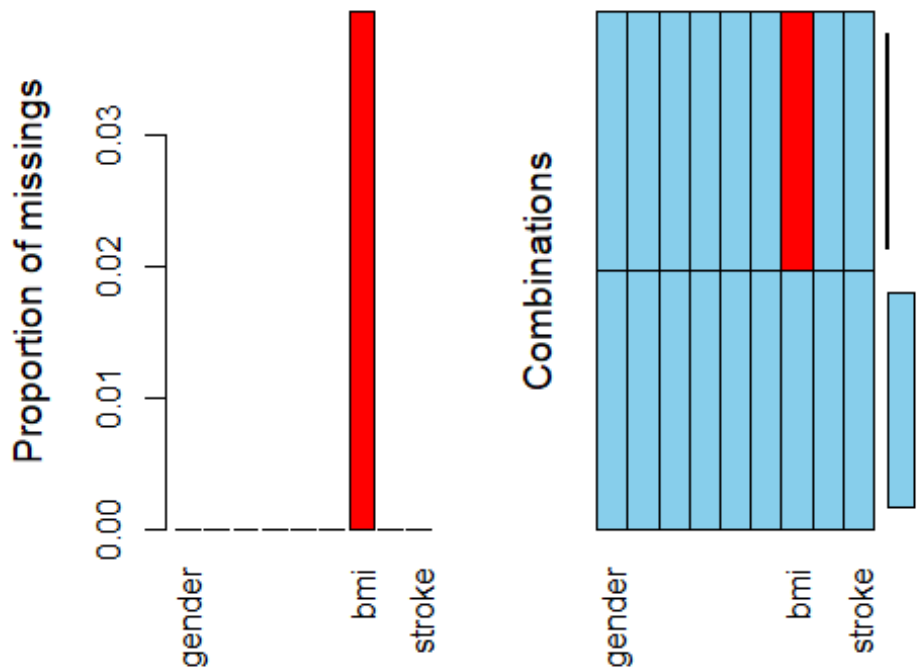

```
par(mfrow = c(1,2))
hist(data$bmi, 100)
qqnorm(data$bmi)
```



```
#Shapiro-Wilk normality test, we see the p-value is significant, and thus we rejec
t the null hypothesis of normal data

shapiro.test(data$bmi)

##
##  Shapiro-Wilk normality test
##
## data:  data$bmi
## W = 0.95355, p-value < 2.2e-16

shapiro.test(data$avg_glucose_level[1:5000])

##
##  Shapiro-Wilk normality test
##
## data:  data$avg_glucose_level[1:5000]
## W = 0.80526, p-value < 2.2e-16

knitr::opts_chunk$set(echo = TRUE)
```

This plot rappresent the missing values inside the data set:

```
#Missing data in the data set

aggr(data, prop = TRUE,
    numbers = TRUE)

## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies
```




```
knitr::opts_chunk$set(echo = TRUE)
```

In order to build our model we'll split the data set in two chunk, train (80%) and test (20%)

```
#Split data into train and test set

data1 <- initial_split(data = data, prop = 0.8)


data_train <- training(data1)
data_test <- testing(data1)

knitr::opts_chunk$set(echo = TRUE)
```

The following plots give us some insight: in general stroke happens starting from the age of 40, furthermore we can see the yellow point rappresenting people with hypertension (first plot) and heart disease (second plot).

```
p1 <- ggplot(data = data_train, aes(x = stroke, y= age))
p1 + geom_point(aes(colour = data_train$hypertension)) +
  scale_colour_viridis(discrete = FALSE)

p1 <- ggplot(data = data_train, aes(x = stroke, y= age))
p1 + geom_point(aes(colour = data_train$heart_disease)) +
  scale_colour_viridis(discrete = FALSE)
```

```
knitr::opts_chunk$set(echo = TRUE)
```

CLASSIFICATION TREE

With this model the objective was not to predict strokes, I wanted to predict average glucose level using as predictor stroke, hypertension, bmi and heart disease of patients after turning avg_glucose_level and bmi into factor with four levels. The results are not great because the model was able to classify the patient into the right category of glucose levels with an efficacy of around 41%.

```
range(data_train$avg_glucose_level)

data_train$bmi <- as.numeric(data_train$bmi)

data_train$glucose_cat_tr <- cut(data_train$avg_glucose_level, breaks = c(55,80,11
0,150,271), labels = c("low","normal","high","very high"))
data_train$bmi_cat_tr <- cut(data_train$bmi, breaks = c(0, 18, 24,29, 100), labels
= c("underweight", "normal","overweight","obese") )

data_test$glucose_cat_ts <- cut(data_test$avg_glucose_level, breaks = c(55,80,110,
150,271), labels = c("low","normal","high","very high"))
data_test$bmi_cat_ts <- cut(data_test$bmi, breaks = c(0, 18, 24,29, 100), labels =
c("underweight", "normal","overweight","obese") )

set.seed(12345)
cartModel <- train(x = data_train[, c("stroke",  "hypertension", "heart_disease","
bmi_cat_tr")],
                   y = factor(data_train$glucose_cat_tr),
                   method = "rpart",
                   preProcess = NULL,
                   tuneLength = 10,
                   trControl = trainControl(method = "cv",
                                            number = 6
                   )
)

cartModel

plot(cartModel$finalModel)
text(cartModel$finalModel, cex = 0.5)
fancyRpartPlot(cartModel$finalModel, cex = 0.4, main = "")

knitr::opts_chunk$set(echo = TRUE)
```

LOGISTIC REGRESSION MODEL PREDICTING STROKES

After fitting the train set to the logistic model it was able to predict the strokes with an accuracy of about 71%.

```
m.lr <- glm(stroke ~ avg_glucose_level+age+heart_disease+hypertension,
            family = binomial(link = "logit"),
            data = data_train, model = TRUE)
summary(m.lr)
```

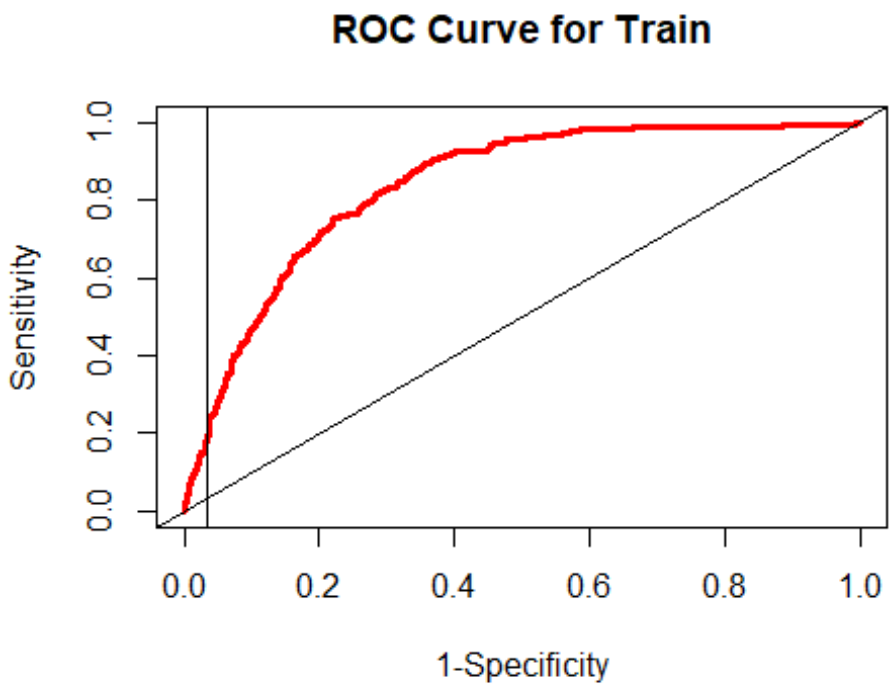
```
##
## Call:
## glm(formula = stroke ~ avg_glucose_level + age + heart_disease +
##     hypertension, family = binomial(link = "logit"), data = data_train,
##     model = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0414  -0.3353  -0.1837  -0.0901   3.7066
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.257446   0.377294 -19.236  < 2e-16 ***
## avg_glucose_level  0.004282   0.001266   3.381 0.000722 ***
## age                0.066577   0.005464  12.185  < 2e-16 ***
## heart_disease      0.317207   0.208891   1.519 0.128881
## hypertension       0.274193   0.178872   1.533 0.125301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1661.7  on 4087  degrees of freedom
## Residual deviance: 1339.7  on 4083  degrees of freedom
## AIC: 1349.7
##
## Number of Fisher Scoring iterations: 7
```

```
cut_off <-roc(response= data_train$stroke, predictor= m.lr$fitted.values)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
e <-cbind(cut_off$thresholds,cut_off$sensitivities+cut_off$specificities)
best_t <-subset(e,e[,2]==max(e[,2]))[,1]
#Plot ROC Curve
plot(1-cut_off$specificities,cut_off$sensitivities,type="l",
     ylab="Sensitivity",xlab="1-Specificity",col="red",lwd=3,
     main ="ROC Curve for Train")
abline(a=0,b=1)
abline(v = best_t) #add optimal t to ROC curve
```




```
cat(" The best value of cut-off for classifier is ", best_t)
```

```
##  The best value of cut-off for classifier is  0.03547765
```

```
# Predict the probabilities for test and apply the cut-off
predict_prob <-predict(m.lr, newdata=data_test, type="response")
#Apply the cutoff to get the class
class_pred <-ifelse(predict_prob >0.045,1,0)
#Classification table
table(data_test$stroke,class_pred)
```

```
##    class_pred
##       0   1
##   0 669 315
##   1   4  34
```

```
#Classification rate
sum(diag(table(data_test$stroke,class_pred))/nrow(data_test))
```

```
## [1] 0.6878669
```

```
#with a logistic regression model we reached 67% good classification on the test d
ata

anova(m.lr, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
```

```
##
## Response: stroke
##
## Terms added sequentially (first to last)
##
##
##                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                4087     1661.7
## avg_glucose_level  1   60.203      4086     1601.5 8.557e-15 ***
## age                1  257.139      4085     1344.4 < 2.2e-16 ***
## heart_disease      1    2.345      4084     1342.0    0.1257
## hypertension       1    2.280      4083     1339.7    0.1311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#The chi-square test on four of the variables is significant as the p-value is les
s than 0.05.
#4 out of five contributions to the model are significant.

knitr::opts_chunk$set(echo = TRUE)
```

CONCLUSION

The objective of building a model that can detect true positive was achived, even though the accuracy is not excellent (71%) the model performed relatively well given that inside the data set only 5% of patients had a stroke and many of them didn't have hypertension, heart disease and other clinical variable weren't typical of a patient with high probability of strokes.