Statistical Learning Project

Filippo Santin, Gurjeet Singh, Francesca Zen

18/5/2021

Exploiting the Dataset

In the following report we present the analysis computed on the stroke dataset provided by kaggle [1].

```
stroke_data <- read.csv('healthcare-dataset-stroke-data.csv')
attach(stroke_data)
```

It is composed of 5,110 entries with a total of 12 columns: id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke. Our aim is to see if and how the features are related to each other in order to predict which individual is more probable to get a stroke. The preliminary part of the analysis focuses on the study of the dataset, we remove the column which identify the singular individuals

```
stroke_data<-stroke_data[,-1]
```

and transformed the categorical variables into factors:

```
stroke_data$gender<- as.factor(gender)
stroke_data$ever_married<-as.factor(ever_married)
stroke_data$work_type<-as.factor(work_type)
stroke_data$Residence_type<-as.factor(Residence_type)
stroke_data$smoking_status<-as.factor(smoking_status)
```

We discovered that the variable bmi was not numeric because of the presence of a string "N/A" which identifies the lack of the value, and so we transform it using the command `stroke_data$bmi <- as.numeric(bmi)`. We then removed those NA values:`stroke_data<- na.omit(stroke_data)`. We ended up having 4,909 entries and 11 total columns. Here we give a quick overview of the main information about the dataset:

```
## Warning: NA introdotti per coercizione
```

```
summary(stroke_data)
```

```
##     gender          age         hypertension     heart_disease    ever_married
##  Female:2897   Min.   : 0.08   Min.   :0.00000   Min.   :0.0000   No :1705
##  Male  :2011   1st Qu.:25.00   1st Qu.:0.00000   1st Qu.:0.0000   Yes:3204
##  Other :   1   Median :44.00   Median :0.00000   Median :0.0000
##                Mean   :42.87   Mean   :0.09187   Mean   :0.0495
##                3rd Qu.:60.00   3rd Qu.:0.00000   3rd Qu.:0.0000
##                Max.   :82.00   Max.   :1.00000   Max.   :1.0000
##          work_type    Residence_type avg_glucose_level      bmi
##  children     : 671   Rural:2419     Min.   : 55.12    Min.   :10.30
##  Govt_job     : 630   Urban:2490     1st Qu.: 77.07    1st Qu.:23.50
##  Never_worked :  22                  Median : 91.68    Median :28.10
##  Private      :2811                  Mean   :105.31    Mean   :28.89
```

[1] https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

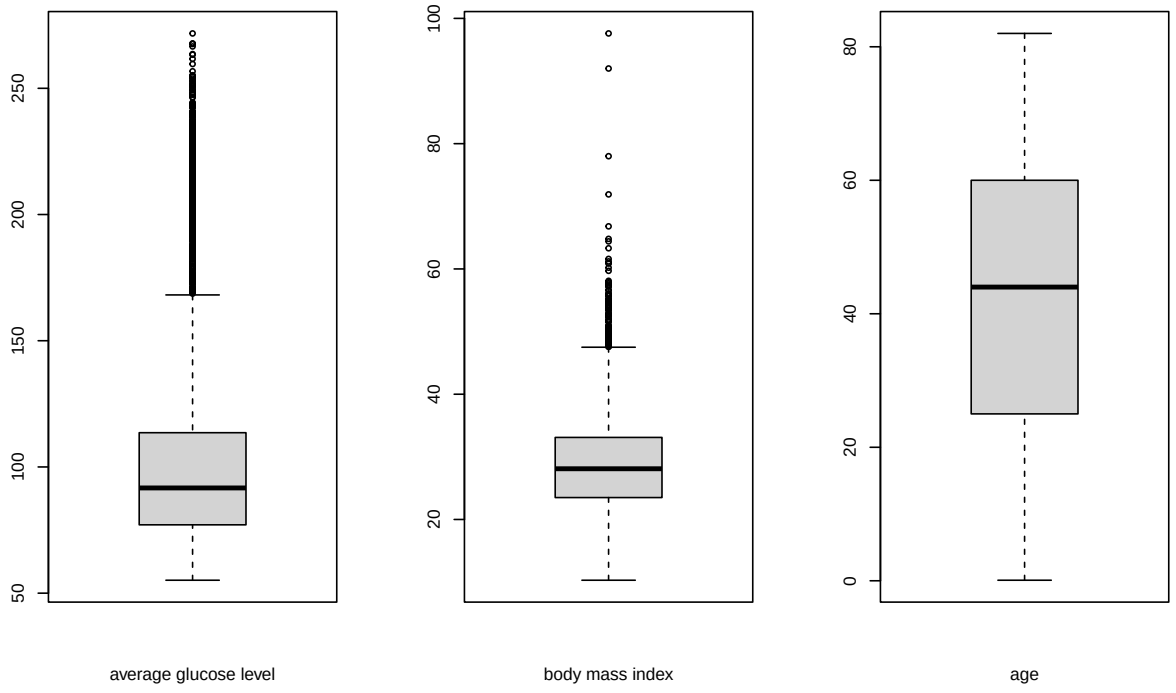
```
##  Self-employed: 775                 3rd Qu.:113.57    3rd Qu.:33.10
##                                     Max.   :271.74    Max.   :97.60
##          smoking_status     stroke
##  formerly smoked: 837   Min.   :0.00000
##  never smoked   :1852   1st Qu.:0.00000
##  smokes         : 737   Median :0.00000
##  Unknown        :1483   Mean   :0.04257
##                         3rd Qu.:0.00000
##                         Max.   :1.00000
```

```
attach(stroke_data)
```

```
## I seguenti oggetti sono mascherati da stroke_data (pos = 3):
##
##     age, avg_glucose_level, bmi, ever_married, gender, heart_disease,
##     hypertension, Residence_type, smoking_status, stroke, work_type
```

A visual transformation of these values is provided in the following boxplots:

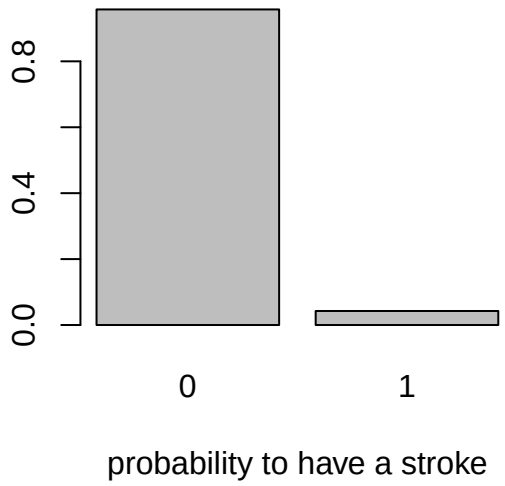
```
par(mfrow=c(1,3))
boxplot(avg_glucose_level, xlab= 'average glucose level' )
boxplot(bmi, xlab = 'body mass index')
boxplot(age, xlab = 'age',pch=20)
```




```
par(mfrow=c(1,1))
```

Through the analysis on the Stroke dataset we discovered that it was strongly bias, in the sense that 209 people on a total of 4909 get a stroke:

```
barplot(table(stroke)/dim(stroke_data)[1],
        xlab='probability to have a stroke')
```

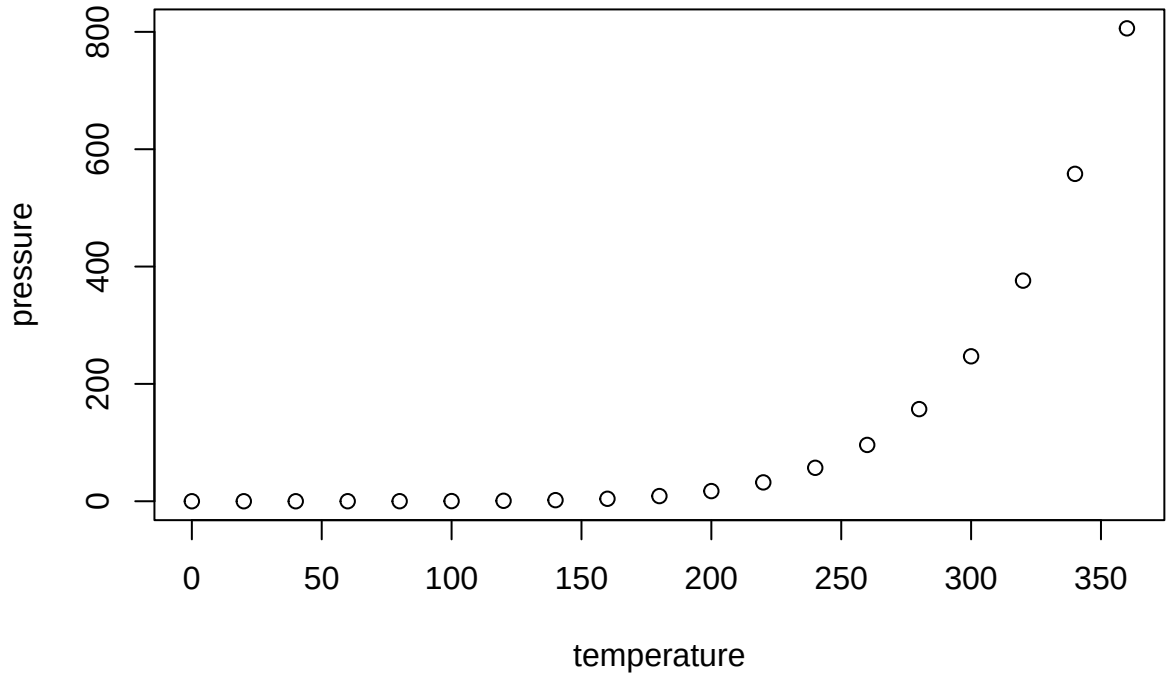


This value is representative of the real situation in which there are not many stroke cases compared with the whole population. At this point we can ask some questions:

- Is it possible to prevent ictus?
- Which factors are the most related to it?
- How strong are the relations between the features?

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.