‘Tipología y ciclo de vida de los datos: PRA2 - ¿Cómo realizar la limpieza y análisis de datos?’

Autores: Alexiss Mejía Borja - Julio Morales

12 de enero, 2023

Contents

1. Descripción

El documento contiene el desarrollo de la Práctica 2 del ramo Tipología y Ciclo de vida de los datos, correspondiente al Master Universitario en ciencia de Datos. La práctica consiste en la limpieza y análisis de un dataset elegido de la página Kaggle.com, con el find e identificar los insight relevantes del set de datos y la preparación previa para el análisis posterior.

1.1. Objetivos

Los objetivos son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

1.2. Competencias

En esta práctica se desarrollan las siguientes competencias:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

2. Resolución

2.1 Descripción del dataset

Se utiliza el set de datos de pacientes de accidente cerebrovasculares del siguiente link https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset, el cual contiene 12 variables y 5110 registros, así como contiene 2 variables numéricas, 8 categóricas, 1 binaria objetivo y 1 identificador único.

A continuación se definen las variables de estudio:

DIMENSIÓN PACIENTE

- **gender** [chr] Género del paciente. Male, Female, Other
- **age** [num] Edad del paciente
- **hypertension** [num] 0 si el paciente no tiene hipertensión, 1 si el paciente tiene hipertensión
- **heart_disease** [num] 0 si el paciente no tiene ninguna enfermedad cardíaca, 1 si el paciente tiene una enfermedad cardíaca

- **ever_married** [chr] Estado civil si alguna se encuentra casado. No, Yes
- **work_type** [chr] Tipo de trabajo. children, Govt_jov, Never_worked, Private o Self-employed
- **Residence_type** [chr] Tipo residencia. Urban, Rural
- **avg_glucose_level** [num] Nivel promedio de glucosa en sangre
- **bmi** [num] Índice de masa corporal
- **smoking_status** [chr] Tipo de fumador. formerly smoked, never smoked, smokes or Unknown (información no disponible)
- **stroke** [num] 1 si el paciente tuvo un accidente cerebrovascular o 0 si no

2.2. Importancia y objetivos de los análisis

Según la OMS, los accidentes cerebrovascular representan la 2da causa de muertes a nivel mundial, siendo el 11% de los casos se vinculan a esta causa. La importancia del set de datos elegido radica en identificar que variables influyen en mayor medida en la probabilidad de que un paciente pueda tener un accidente cerebrovascular en función de las variables de interés presentadas.

El objetivo es realizar la limpieza y análisis del set de datos para un modelo de regresión logística, en base a la variable dependiente "stroke", siendo 1 si el paciente tuvo o 0 si no un accidente cardiovascular, de manera de obtener las variables influyentes en un accidente cerebrovascular. Como objetivos secundarios se definen:

- Selección de datos de interés y definición del origen de datos
- Preparación de datos para el análisis posterior, realizando limpieza de datos, transformación de datos t tratamiento de outliers.
- Obtener conocimiento a partir del análisis de las variables, así como insight relevantes para el entendimiento y validación del instrumento.
- Medir la calidad del modelo obtenido.
- Interpretar el modelo para otorgar usabilidad a los usarios.

2.3 Limpieza de los datos

Para comenzar con el proceso de limpieza de los datos, procedemos a realizar la lectura del fichero en formato CSV. Esto tiene con finalidad cargar el archivo de datos y examinar el tipo de datos con los que R ha interpretado cada variable. Adocionalmente se va a examinar los valores resumen de cada tipo de variable.

Cargar el Archivo

```
# Lectura del archivo seleccionado en  formato csv
healthcare<- read_delim("healthcare-dataset-stroke-data.csv", delim = ";", escape_double = FALSE, trim_
```

```
## Rows: 5110 Columns: 12
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ";"
## chr (5): gender, ever_married, work_type, Residence_type, smoking_status
## dbl (7): id, age, hypertension, heart_disease, avg_glucose_level, bmi, stroke
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(healthcare,n=4L)
```

```
## # A tibble: 4 x 12
##      id gender   age hypertension heart_~1 ever_~2 work_~3 Resid~4 avg_g~5   bmi
##   <dbl> <chr>  <dbl>        <dbl>    <dbl> <chr>   <chr>   <chr>     <dbl> <dbl>
## 1  9046 Male      67            0        1 Yes     Private Urban      229.  36.6
## 2 51676 Female    61            0        0 Yes     Self-e~ Rural      202.  NA
## 3 31112 Male      80            0        1 Yes     Private Rural      106.  32.5
## 4 60182 Female    49            0        0 Yes     Private Urban      171.  34.4
## # ... with 2 more variables: smoking_status <chr>, stroke <dbl>, and
## #   abbreviated variable names 1: heart_disease, 2: ever_married, 3: work_type,
## #   4: Residence_type, 5: avg_glucose_level
```

Una vez cargado el dataset es necesario verificar la tipología de los atributos, asi como la verificación de la existencia de posibles valores perdidos y de valores atípicos

Tipos de datos

```
# Tipo de datos asigando para cada atributo
str(healthcare)
```

```
## spc_tbl_ [5,110 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ id               : num [1:5110] 9046 51676 31112 60182 1665 ...
##  $ gender           : chr [1:5110] "Male" "Female" "Male" "Female" ...
##  $ age              : num [1:5110] 67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : num [1:5110] 0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : num [1:5110] 1 0 1 0 0 0 1 0 0 0 ...
##  $ ever_married     : chr [1:5110] "Yes" "Yes" "Yes" "Yes" ...
##  $ work_type        : chr [1:5110] "Private" "Self-employed" "Private" "Private" ...
##  $ Residence_type   : chr [1:5110] "Urban" "Rural" "Rural" "Urban" ...
##  $ avg_glucose_level: num [1:5110] 229 202 106 171 174 ...
##  $ bmi              : num [1:5110] 36.6 NA 32.5 34.4 24 29 27.4 22.8 NA 24.2 ...
##  $ smoking_status   : chr [1:5110] "formerly smoked" "never smoked" "never smoked" "smokes" ...
##  $ stroke           : num [1:5110] 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_double(),
##   ..   smoking_status = col_character(),
##   ..   stroke = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Se puede notar la existencia de variables de tipo numericas como de tipo categóricas, adicionalmente se puede notar la existencia de valores atípicos en el atributo denominado bmi (índice de masa coporal), para verificar la existencia de más valores perdidos se ejecutacutarán los procesos respectivos en las siguientes secciones.

2.3.1. Normalización de los datos cualitativos

```
# Para el caso del genero
table(healthcare$gender)
```

```
##
## Female   Male  Other
##   2994   2115      1
```

```
# Para el caso de si alguna vez estuvo casado
table(healthcare$ever_married)
```

```
##
##   No  Yes
## 1757 3353
```

```
# Para el caso del tipo de trabajo
table(healthcare$work_type)
```

```
##
##      children      Govt_job  Never_worked       Private Self-employed
##           687           657            22          2925           819
```

```
# Para el caso del tipo de residencia
table(healthcare$Residence_type)
```

```
##
## Rural Urban
##  2514  2596
```

```
# Para el caso de si alguna vez a fumado
table(healthcare$smoking_status)
```

```
##
## formerly smoked    never smoked          smokes         Unknown
##             885            1892             789            1544
```

No se puede ver valores que ameriten una normalización de los datos cualitativos del dataset por lo que ha optado por conservar la estructura inicial del mismo.

2.3.2. Normalización de los datos cuantitativos

```r
# Verificación de la edad
summary(healthcare$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.08   25.00   45.00   43.23   61.00   82.00
```

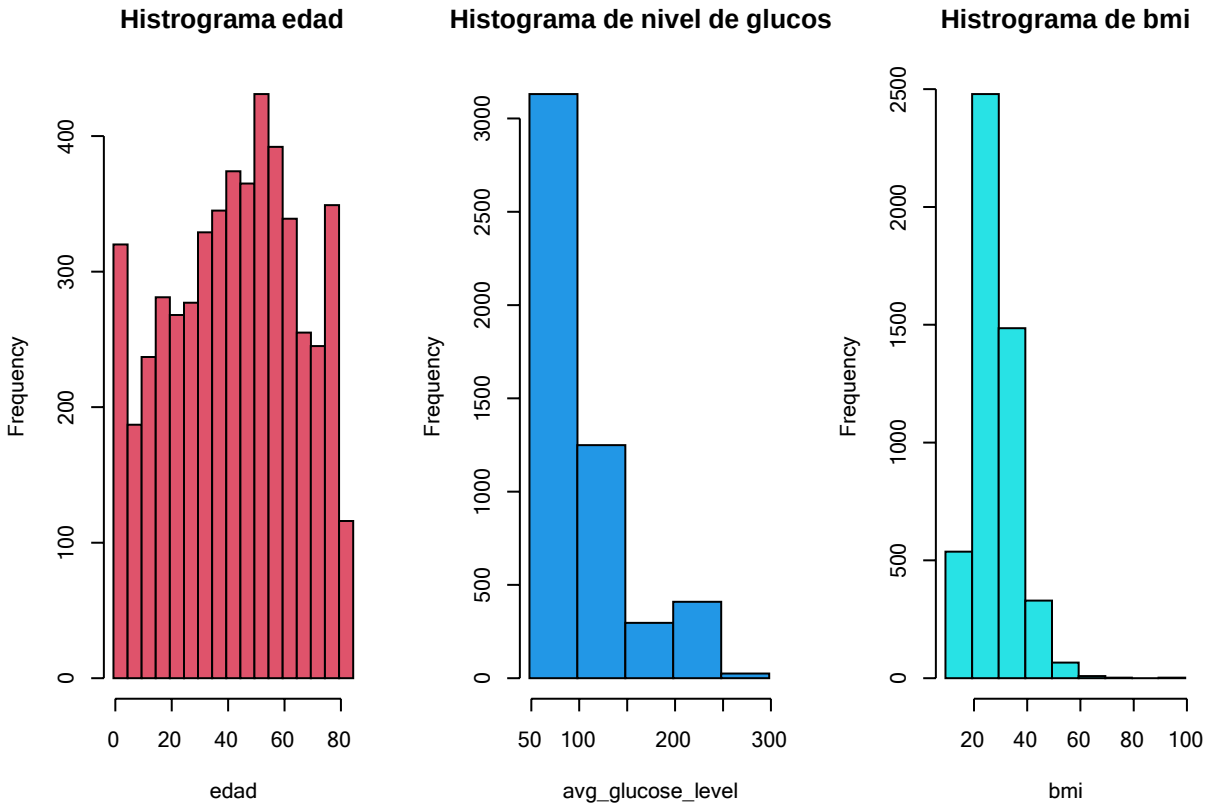
```r
# Verificación de la avg_glucose_level
summary(healthcare$avg_glucose_level)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.12   77.25   91.89  106.15  114.09  271.74
```

```r
# Verificación del bmi
summary(healthcare$bmi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.30   23.50   28.10   28.89   33.10   97.60     201
```

```r
# Histogramas de la edad, glucosa y bmi
par(mfrow=c(1,3))

hist(healthcare$age,main="Histrograma edad",xlab="edad",col=2,breaks="Sturges")

hist(healthcare$avg_glucose_level, main = "Histograma de nivel de glucosa",
     xlab = "avg_glucose_level", col=4, breaks = 7)

hist(healthcare$bmi, main = "Histrograma de bmi",
     xlab = "bmi",col=5, breaks = 10)
```



Un hecho particular sucede con la edad ya para ser un dataset relacionado los ataques cardiacos la distribución de la edad refleja un grupo considerable de individuos que se encuentran por debajo de los 10 años de edad incluso infantes tal y como se puede ver el summary de este atributo, cabe aquí la cuestión de si es necesario normalizar la edad de los individuos o simplemente realizar una selección de datos que son de mayor interés y lógica para el presente análisis.

En el caso de la vaiable relacionada con el promedio del nivel de la glucosa tanto el resúmen como los gráficos no muestra indicios de que requieran un tratamiento de normalización, sin embargo como se advirtio en la anterior sección el bmi presenta valores perdidos, hecho que lo tratara en la siguiente sección.

2.3.3. Selección de datos de interés (reducción de cantidad)

Como se vio en la sección anterior existen registros con edades inferiores a 18 años incluso existen registros de personas que no llegan a cumplir el año de vida, este es un hecho que llama la atención, pues sería necesatio verificar la edad con otra variable de control, por ejemplo el tipo de trabajo, ya que la lógica nos indica que las personas menores de edad no debería encontrarse trabajando.

```
# se selecciona de acuerdo a una edad <18 años y que en la variable work_type
#  sea  distinto   a  children   y Never_worked
seleccio_edad<- subset(healthcare, age < 18 & work_type !="children" &
                       work_type !="Never_worked")

# se verifica  las caracteristcias de menores de edad
seleccio_edad%>% count (work_type, age)
```

```
## # A tibble: 16 x 3
```

```
##    work_type       age     n
##    <chr>         <dbl> <int>
##  1 Govt_job         14     1
##  2 Govt_job         15     1
##  3 Govt_job         16     1
##  4 Govt_job         17     3
##  5 Private           8     1
##  6 Private          13     9
##  7 Private          14    19
##  8 Private          15    21
##  9 Private          16    35
## 10 Private          17    49
## 11 Self-employed     7     1
## 12 Self-employed    13     1
## 13 Self-employed    14     4
## 14 Self-employed    15     2
## 15 Self-employed    16     1
## 16 Self-employed    17     3
```

Se puede notar como existen incluso niños con 8 y 7 años de edad que trabajan el en sector privado o que son auto empleados, hecho que no tiene mucha coherencia, en el mayor de los casos existen 49 adolescentes de 17 años que trabajan en el sector privado; con estos antecedentes se ha decidido excluir los casos con una edad menor a los 15 años y que el tipo de trabajo sea distinto y a children y Never_worked

```
# Se excluyen a los casos de acuerdo a una edad menor a 15 años y que en la variable
# work_type sea distinto a children y Never_worked
healthcare$filtro<-ifelse(healthcare$age<15 & (healthcare$work_type !="children"
                                               &  healthcare$work_type !="Never_worked" ),1,0)

healthcare<- subset(healthcare, filtro==0)
```

El dataset final consta de 5074 observaciones sobre las cuales se ejecutaran el resto de proceso y análisis.

2.3.4. Detección de Valores perdidos

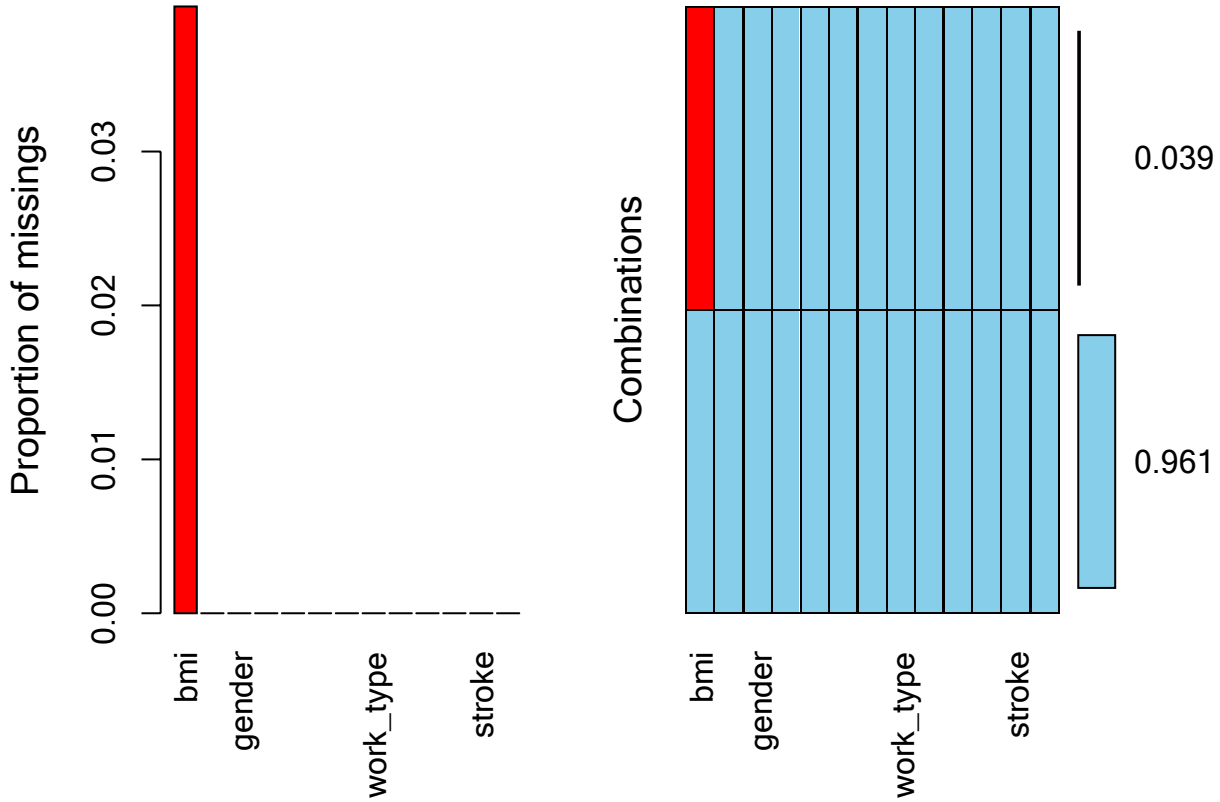
```
# Valores perdidos del dataset healthcare
apply(is.na(healthcare), 2, sum)
```

```
##                id            gender               age      hypertension
##                 0                 0                 0                 0
##     heart_disease      ever_married         work_type    Residence_type
##                 0                 0                 0                 0
## avg_glucose_level               bmi    smoking_status            stroke
##                 0               200                 0                 0
##            filtro
##                 0
```

Tal y como se sospechaba existen datos perdidos en la variable relacionada con el índice de masa corporal (200 casos perdiddos), para este caso y al considerar un atributo importante se ha optado por realizar imputaciones mediante la técnica knn con el fin de recuperar la información perdida.

De igual forma de manera grafica podemos ver la distribución de los valores perdidos en el dataset del ejercicio

```r
# Valores perdidos del dataset  healthcare
aggr(healthcare, numbers=T, sortVar=T)
```




```
##
##  Variables sorted by number of missings:
##           Variable      Count
##                bmi 0.03941663
##                 id 0.00000000
##             gender 0.00000000
##                age 0.00000000
##       hypertension 0.00000000
##      heart_disease 0.00000000
##       ever_married 0.00000000
##          work_type 0.00000000
##     Residence_type 0.00000000
##  avg_glucose_level 0.00000000
##     smoking_status 0.00000000
##             stroke 0.00000000
##             filtro 0.00000000
```

Aunque tanto los graficos como la estructura del dataset muestra que la variable bmi presenta valores perdidos el atributo smoking_status presenta una categoria denominada como Unknown que tambien podria ser considerada como valor perdido.

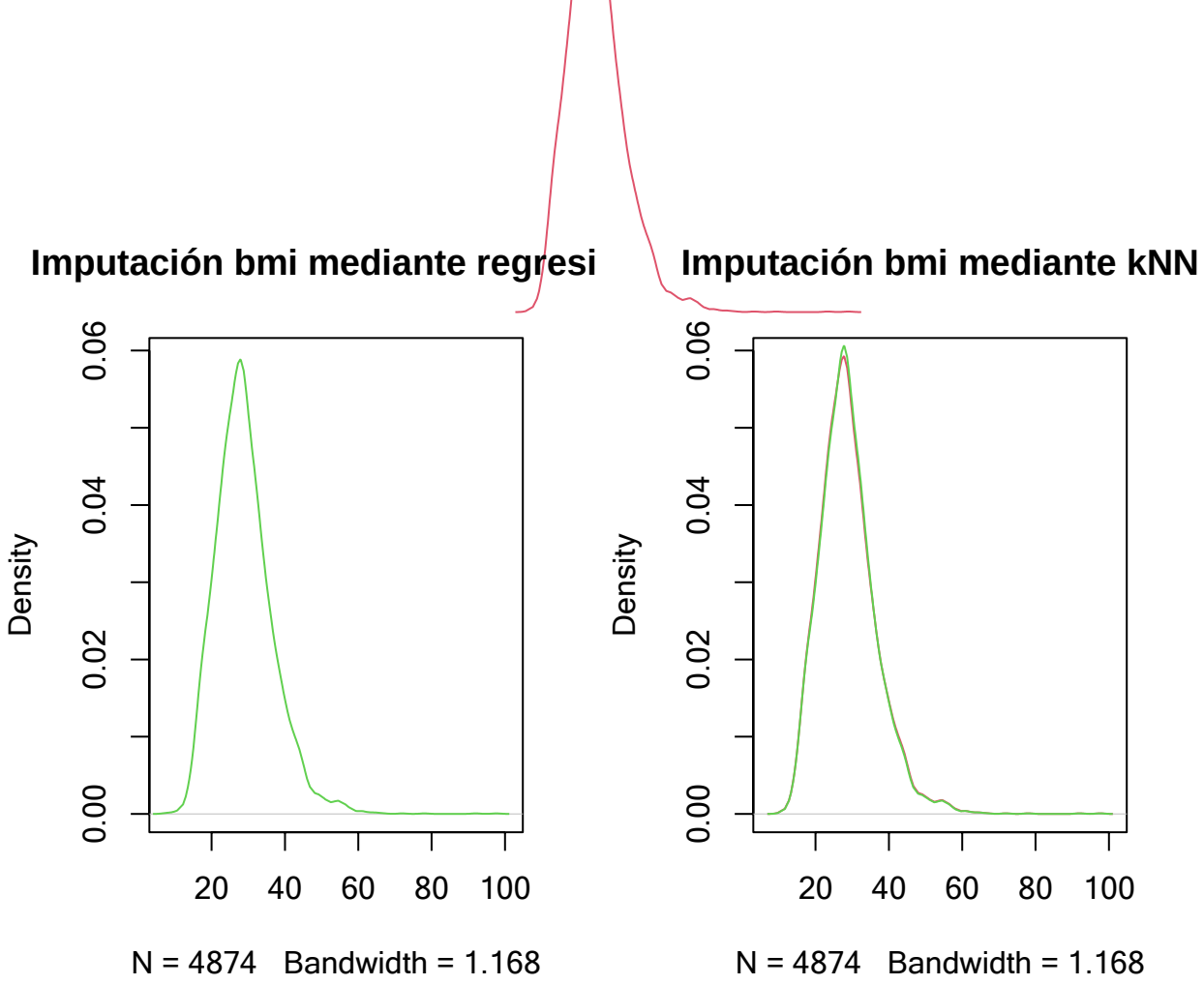
2.3.5. *Imputación de los Valores perdidos*

Para la imputación de los valores perdidos del índice de masa corporal se va a usar dos métodos y elegir el mejor. El primero de elloses la imputación por kNN, y el segundo mediante una Regresión Estocastica

```
# Imputación de los valores perdidos del bmi mediante kNN

#Se extrae las variables de horas, genero y ClaimNumber en otro dataframe
gender<-(healthcare[, "gender"])
age<-(healthcare[,"age"])
hypertension<-(healthcare[,"hypertension"])
bmi<-(healthcare[,"bmi"])
id<-(healthcare[,"id"])

df<- data.frame(id,bmi,hypertension,age,gender)

df_imput<-kNN(df,k=4)
```

```
# Imputación de los valores perdidos del bmi mediante la regresión

columns <- c("bmi","id")
imputed_data1 <- mice(healthcare[,names(healthcare) %in% columns],m = 1,
                      maxit = 1, method = "norm.nob",seed = 2018,print=F)
complete.data1 <- mice::complete(imputed_data1)
```

```
# Gráficos de densidad de la variable bmi original vs los dos metodos de imputación

par(mfrow=c(1,2))
plot(density(healthcare$bmi,na.rm = T),col=2,main="Imputación bmi mediante regresión")
lines(density(complete.data1$bmi),col=3)

plot(density(healthcare$bmi,na.rm = T),col=2,main="Imputación bmi mediante kNN")
lines(density(df_imput$bmi),col=3)
```



Como se puede ver en la comparación de los dos métodos, no existe una variación muy significativa, sin embargo la Imputación mediante regresión presenta un mejor ajuste entre la distribución de la variable original y el bmi imputado, por lo que se ha optado por usa este método estocástico como técnica de imputación.

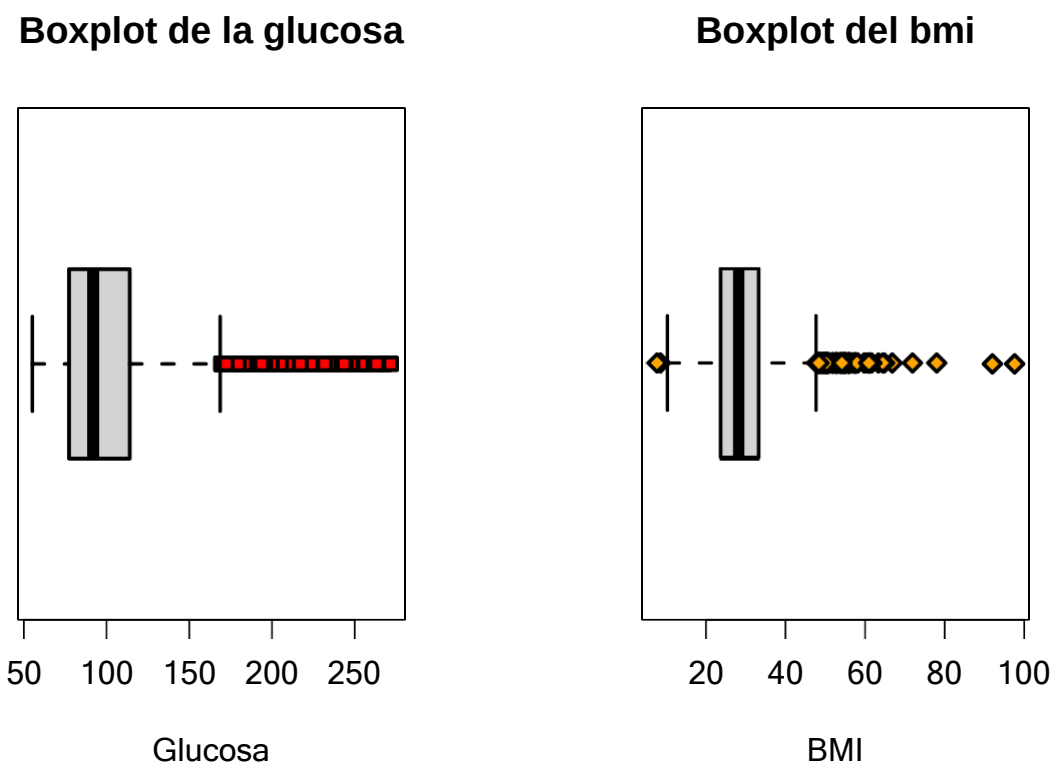
```
# Se genera una verificación de la imputación hecha en bmi


complete.data1 <-dplyr::rename(complete.data1, c( bmi_imput  = "bmi"))

healthcare<-merge(healthcare, complete.data1, by ="id", all = TRUE)

borrar <- c("bmi", "filtro")
healthcare <- healthcare[ , !(names(healthcare) %in% borrar)]

apply(is.na(healthcare), 2, sum)
```

```
##                 id             gender                age       hypertension
##                  0                  0                  0                  0
##      heart_disease       ever_married          work_type     Residence_type
##                  0                  0                  0                  0
## avg_glucose_level     smoking_status             stroke         bmi_imput
##                  0                  0                  0                  0
```

2.3.6 Verificación de datos extremos

Se realiza la verificación de la existencia o no de datos atípicos de las variables cuantitativas

```
# Se genera una verificación de los datos atípicos del bmi y del nivel de la glucosa
par(mfrow=c(1,2))

boxplot((healthcare$avg_glucose_level),
        horizontal = T,
        lwd = 2,
        xlab = "Glucosa",
        main = "Boxplot de la glucosa",
        border = "black",
        outpch = 22,
        outbg = "red")
boxplot((healthcare$bmi),
        horizontal = T,
        lwd = 2,
        xlab = "BMI",
        main = "Boxplot del bmi",
        border = "black",
        outpch = 23,
        outbg = "orange")
```



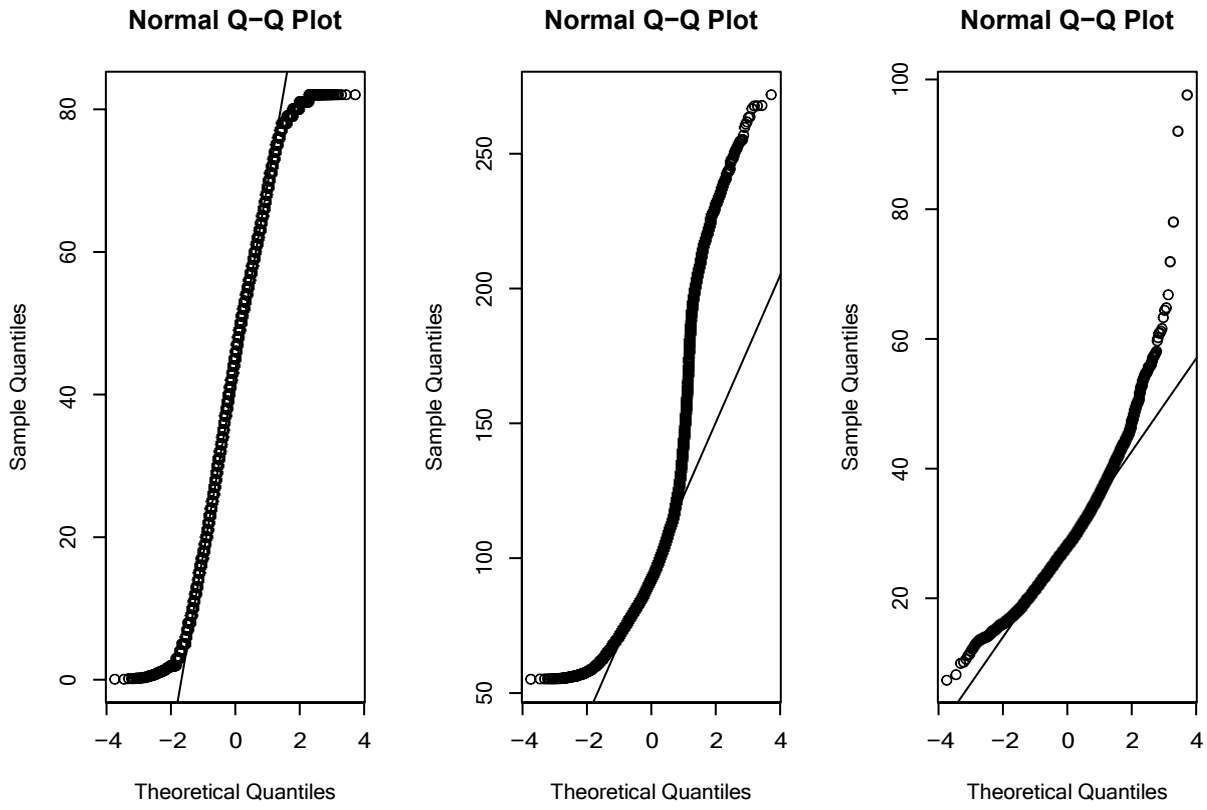
Según el diagrama de caja del bmi y de la glucosa, se podría concluir que existen datos atípicos, sin embargo al tratarse de un fenómeno medico donde la prevalencia o no de un ataque cardiaco puede deberse a la existencia de datos atípicos se ha optado por consérvalos y no aplicar ningún tratamiento sobre ellos.

2.4. Análisis de los datos

2.4.1 Normalidad de las variables cuantitativas

Se verifica la normalidad de los atributos cuantitativos del dataset

```
par(mfrow=c(1,3))

#  Normalidad de la edad
qqnorm(healthcare$age)
qqline(healthcare$age)

#  Normalidad de la glucosa
qqnorm(healthcare$avg_glucose_level)
qqline(healthcare$avg_glucose_level)

#  Normalidad de la bmi
qqnorm(healthcare$bmi)
qqline(healthcare$bmi)
```




```
#  Se realiza la  Prueba  de  Lilliefors

# edad
lillie.test(healthcare$age)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  healthcare$age
## D = 0.050496, p-value < 2.2e-16
```

```
# glucosa
lillie.test(healthcare$avg_glucose_level)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  healthcare$avg_glucose_level
## D = 0.18318, p-value < 2.2e-16
```

```
# glucosa
lillie.test(healthcare$bmi)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  healthcare$bmi
## D = 0.057463, p-value < 2.2e-16
```

Mediante los QQ plot se puede ver que la variable de la glucosa como del bmi y la edad no se encuentran tan alineados respecto a la línea de referencia esto sobre todo en las partes más alejadas de la distribución, por otro lado en la prueba de Lilliefors se puede notar que el p calculado (2.2e-16) es mucho menor que el valor de contraste en este caso 0.05, por lo que se podría concluir que ninguna de las variables siguen una distribución normal.

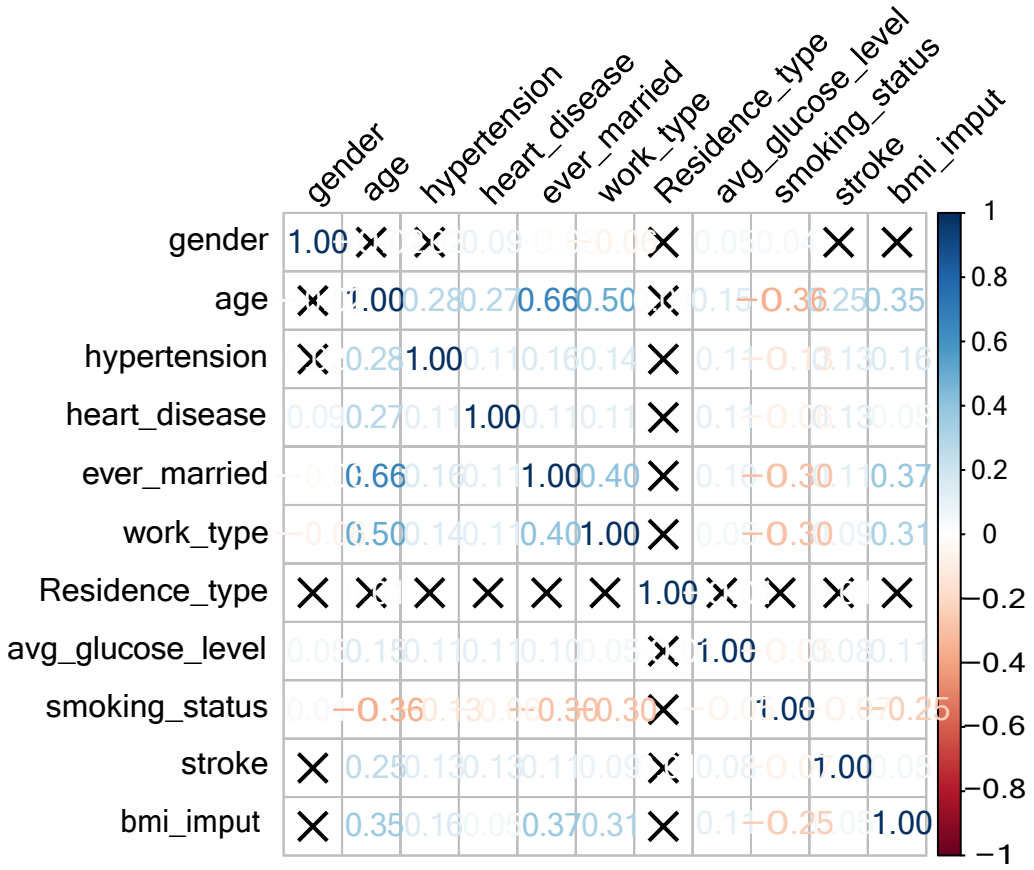
2.4.2 Correlaciones

Como segundo punto dentro del análisis de los datos se va a proceder a calcular una matriz de correlaciones que nos permitirá determinar la medida de asociación entre los distintos atributos del dataset incluida la variable de interés para el presente estudio (stroke existencia o no de una ataque cardiaco), para este caso y al contar con variables de tipo cualitativo como cuantitativo y al haber verificado que las distribuciones de estas últimas no siguen una distribución normal se procede a usar el coeficiente de correlación de Spearman.

Para verificar la correlación de las variables categóricas, se les va aplicar un proceso de recodificación con el fin de que sean numéricas

```
# Se recodifica las variables categoricas en numericas con el fin de verificar la
#correlación existente, se trabaja en un nuevo dataset para no modificar al original

healthcare_corr<-healthcare
healthcare_corr$ever_married<-dplyr::recode(healthcare_corr$ever_married, No=0, Yes=1)
healthcare_corr$Residence_type<-dplyr::recode(healthcare_corr$Residence_type, Rural =0,
                                              Urban =1)
healthcare_corr$work_type<-dplyr::recode(healthcare_corr$work_type, children=1,
                                         Govt_job=2,Never_worked=3,Private =4,
                                         `Self-employed`=5)
healthcare_corr$smoking_status<-dplyr::recode(healthcare_corr$smoking_status,
```

```
                                    `formerly smoked`=1,
                                    `never smoked`=2, smokes=3, Unknown=4)

healthcare_corr$gender<-dplyr::recode(healthcare_corr$gender, Female=1, Male=2, Other=3)

columns2 <- c("id")
healthcare_corr <- healthcare_corr[ , !(names(healthcare_corr) %in% columns2)]


# se genera la matriz de correlaciones y se grafica para poder visualizar de mejor manera

matriz_correlacion<-cor_mat(healthcare_corr, method="spearman")

cor_plot(matriz_correlacion, method = "number")
```



Como se puede ver en la matriz de correlaciones no existe evidencia suficiente para determinar asociaciónes fuertes entre la mayoría de los atributos del dataset, sobre todo con la variable de si el paciente ha tenido o no un accidente cerebrovascular (stroke), sin embargo se puede establecer algunas variables que tienen cierta correlación, un hecho particular es smoking_status que presenta correlaciones negativas con la edad, con ever_married con el tipo de trabajo y con el bmi, pero no presenta una correlación importante ni positiva ni negativa con la posibilidad que tienen los pacientes de tener o un accidente cerebrovascular.

En cuanto a la variable de interés stroke la asociación más fuerte y positiva es con la **edad** (0.25), lo que a priori nos está diciendo es que a mayor edad existe un mayor riesgo de tener un accidente cerebrovascular otros atributos con el mismo tipo de relación son la **hypertension** (0.13) y **heart_disease** (0.13), es decir que si los pacientes tienen hipertensión y presentan alguna enfermedad, podrían presentar un incremento en

el riesgo de sufrir un accidente cerebrovascular.

2.4.3 Comparaciones entre grupos

Como se pudo notar en la sección aterior no se pudo encontrar correlaciones totalmente determinantes entre cada uno de los atributos y concretamente con el stroke que tienen los pacientes, por lo que se ha optado por realizar comparaciones entre los grupos sobre todo de las variables cualitativasd y determinar si de esta manera se puede determinar un mayor relación, concretamente se aplican pruenas chi-cuadrado para verificar las hipotesis de significancia entre las variables.

Concretamente se plantean las siguientes hipotesis:

- work_type y stroke:
- **H0: La presencia de un accidente accidente cerebrovascular es independiente del tipo de trabajo.**
- **H1: La presencia de un accidente accidente cerebrovascular depende del tipo de trabajo**
- Residence_type y stroke:
- **H0: La presencia de un accidente accidente cerebrovascular es independiente del lugar de residencia.**
- **H1: La presencia de un accidente accidente cerebrovascular depende del lugar de residencia**
- ever_married y stroke:
- **H0: La presencia de un accidente accidente cerebrovascular es independiente del si el paciente estuvo o no casado.**
- **H1: La presencia de un accidente accidente cerebrovascular depende del lugar del si el paciente estuvo o no casado.**
- smoking_status y stroke:
- **H0: La presencia de un accidente accidente cerebrovascular es independiente del si el paciente alguna vez fumo.**
- **H1: La presencia de un accidente accidente cerebrovascular depende del lugar del si el paciente alguna vez fumo**

```
# Pruebas chi cuadrado con respecto a las variables categóricas y la variable stroke

# work_type y stroke
chisq.test(healthcare$work_type, healthcare$stroke)
```

```
## Warning in chisq.test(healthcare$work_type, healthcare$stroke): Chi-squared
## approximation may be incorrect
```

```
##
## Pearson's Chi-squared test
##
## data:  healthcare$work_type and healthcare$stroke
## X-squared = 49.489, df = 4, p-value = 4.616e-10
```

```r
# Residence_type  y  stroke
chisq.test(healthcare$Residence_type, healthcare$stroke)
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  healthcare$Residence_type and healthcare$stroke
## X-squared = 1.0389, df = 1, p-value = 0.3081
```

```r
# ever_married  y  stroke
chisq.test(healthcare$ever_married, healthcare$stroke)
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  healthcare$ever_married and healthcare$stroke
## X-squared = 56.907, df = 1, p-value = 4.568e-14
```

```r
# smoking_status  y  stroke
chisq.test(healthcare$smoking_status, healthcare$stroke)
```

```
##
## Pearson's Chi-squared test
##
## data:  healthcare$smoking_status and healthcare$stroke
## X-squared = 28.906, df = 3, p-value = 2.344e-06
```

Los resultados de la prueba, indican que existen valores de p menores a un alfa (0.05) concretamente ever_married con 4.568e-14, smoking_status con 2.344e-06 y work_type con 4.616e-10, rechazando la Ho, revelando una posible asociación estadísticamente significativa de estos atributos con la posibilidad de tener un accidente cerebrovascular en los pacientes.

Finalmente para el caso de las variables Residence_type no se puede concluir que este atributo y el stroke de los pacientes están asociadas dado que el p value estimado (0.3081) es mayor que 0.05 aceptando de esta manera la Ho.

2.4.4. Modelo de Regresión Logística

Conforme al analisis realizado en secciones anteriores se puede decir que las variables mas significativas son: ever_married, smoking_status, work_type, hypertension, heart_disease y la edad; sin embargo como primer modelo a generarse se ha incluido al atributo avg_glucose_level como variables explicativas adicional del modelo esto con el fin de determinar una influencia multivariante de la misma en el modelo.

Se probaran otro modelo para ver tiene un mejor ajuste a nivel individual como global.

```r
## Se  ejecuta un primero modelo logistico sobre el dataset excluyendo al gender y
#  Residence_type

modelo_glm_stroke1 <- glm(stroke ~ bmi_imput+ever_married+smoking_status+work_type+
                            hypertension+heart_disease+avg_glucose_level+age,
                          family = "binomial", healthcare)
summary(modelo_glm_stroke1)
```

```
##
## Call:
## glm(formula = stroke ~ bmi_imput + ever_married + smoking_status +
##     work_type + hypertension + heart_disease + avg_glucose_level +
##   age, family = "binomial", data = healthcare)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1412  -0.3205   -0.1658   -0.0876   3.5456
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -6.840658   0.774658  -8.831  < 2e-16 ***
## bmi_imput                     0.006530   0.010620   0.615  0.53860
## ever_marriedYes              -0.191182   0.224880  -0.850  0.39524
## smoking_statusnever  smoked  -0.211508   0.174114  -1.215  0.22445
## smoking_statussmokes          0.115473   0.215087   0.537  0.59136
## smoking_statusUnknown        -0.073517   0.207932  -0.354  0.72367
## work_typeGovt_job            -1.014598   0.835118  -1.215  0.22440
## work_typeNever_worked       -10.340031 309.370685  -0.033  0.97334
## work_typePrivate             -0.873983   0.819715  -1.066  0.28633
## work_typeSelf-employed       -1.251391   0.840453  -1.489  0.13650
## hypertension                  0.393266   0.164969   2.384  0.01713 *
## heart_disease                 0.283750   0.190059   1.493  0.13545
## avg_glucose_level             0.003932   0.001192   3.299  0.00097 ***
## age                           0.075309   0.005840  12.896  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1986.8  on 5073  degrees of freedom
## Residual deviance: 1581.0  on 5060  degrees of freedom
## AIC: 1609
##
## Number of Fisher Scoring iterations: 14
```

El primer modelo generado se lo ha generado excluyendo al gender como Residence_type que de manera univariante se determinaron como poco significativas, sin embargo al analizar la significancia y los coeficientes del modelo se puede notar que solo los atributos como la edad, avg_glucose_level y el hypertension en menor nivel son significativos para el modelo. Finalmente en este primer modelo se puede ver que tiene una significación global que correponde al valor de medida AIC (criteriode información de Akaike) de **AIC=1609**.

```
# Finalmente se ejecuta un segundo modelo logistico sobre el dataset conservando
# los atributos más significativos según la matriz de correlaciones pero excluyendo
# avg_glucose_level y la edad, que el analisis univariante resultaron significativas

modelo_glm_stroke2 <- glm(stroke ~ bmi_imput+ever_married+smoking_status+work_type+
                          hypertension+heart_disease,family = "binomial", healthcare)
summary(modelo_glm_stroke2)
```

```
##
## Call:
```

```
## glm(formula = stroke ~ bmi_imput + ever_married + smoking_status +
##     work_type + hypertension + heart_disease, family = "binomial",
##   data = healthcare)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0272  -0.3299   -0.2905   -0.2062   3.4542
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -5.433389   0.758907  -7.159 8.10e-13 ***
## bmi_imput                   -0.006781   0.009475  -0.716 0.474161
## ever_marriedYes              0.734199   0.211674   3.469 0.000523 ***
## smoking_statusnever smoked  -0.387250   0.169142  -2.289 0.022051 *
## smoking_statussmokes        -0.360191   0.206148  -1.747 0.080596 .
## smoking_statusUnknown       -0.265496   0.201938  -1.315 0.188598
## work_typeGovt_job            2.062615   0.767971   2.686 0.007236 **
## work_typeNever_worked       -9.619332 310.118563  -0.031 0.975255
## work_typePrivate             2.150834   0.748675   2.873 0.004068 **
## work_typeSelf-employed       2.371277   0.759032   3.124 0.001784 **
## hypertension                 0.954480   0.159460   5.986 2.15e-09 ***
## heart_disease                1.149599   0.181264   6.342 2.27e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1986.8  on 5073  degrees of freedom
## Residual deviance: 1813.6  on 5062  degrees of freedom
## AIC: 1837.6
##
## Number of Fisher Scoring iterations: 14
```

```
# Calculo de los odds ratio (Probabilidades)
exp(modelo_glm_stroke2$coefficients)
```

```
##                 (Intercept)                   bmi_imput
##                0.004368267                0.993241457
##            ever_marriedYes smoking_statusnever smoked
##                2.083811475                0.678921527
##       smoking_statussmokes      smoking_statusUnknown
##                0.697543258                0.766825853
##          work_typeGovt_job      work_typeNever_worked
##                7.866516162                0.000066432
##           work_typePrivate     work_typeSelf-employed
##                8.592019332               10.711058214
##               hypertension              heart_disease
##                2.597320220                3.156927089
```

Este segundo modelo presenta rasgos muy interesante sobre todo al ver lo que ha sucedido si del análisis se excluyen las variables de la edad y del nivel de glucosa, esto posibilitó no solo el mejorar el criterio **AIC** que ahora es de **1837.6**, si no que también de manera individual más variables independientes se han vuelto significativas, es así que atributos como el hecho de estar casado, el trabajar para el gobierno, el trabajar

en el sector privado y el ser autónomo así como el hecho de presentar enfermedad cardiaca se han vuelto significativas, en menor medida también es significativo el hecho de nunca haber fumado, mientras que la hipertensión aún conserva su significancia en este último modelo.

Este segundo modelo es que se ha optado por usar ya que presenta un mejor ajuste global y un riqueza de análisis en cuanto a los resultados obtendios.

2.4.5. Interpretación de los resutados

La interpretación de los coeficientes para el caso del modelo de regresión logística cambia, en estos casos las interpretaciones radican en las probabilidades que tiene un individuo en tomar el valor 1 de la variable dependiente, en nuestro ejercicio la probabilidad de que un paciente haya tenido un accidente cerebrovascular, Por tanto, los coeficientes del modelo logit se interpretan como el logaritmo del odds ratio.

Bajo este contexto y según los odds ratio del modelo se puede decir que las probabilidades de un paciente que ha estado casado tenga un accidente cerebrovascular es 2.08 veces mayor que las que un paciente que no ha estado casado. Otro atributo que influye en la enfermedad de los pacientes es heart_disease, ya que la probabilidad de que un paciente que presente una enfermedad cardiaca, tenga un accidente cerebrovascular se incremente en 3.15 veces respecto a aquellos pacientes que no han presentado ninguna enfermedad de tipo cardiaca.

Un hecho particular es lo que sucede con aquellos pacientes que trabajan el diferentes sectores, siendo una de las variables que más afectan a los accidentes cerebrovasculares, ya que todos ellos tienen valores muy altos de odds ratio, pero son aquellos pacientes que son Self-employed los que presentan mayores problemas, es asi que la probabilidad de que los pacientes que trabajan en este sector tengan un derrame cerebral es 10.71 veces mayor que las personas que trabajan en los otros sectores; otro tipo de trabajo que castiga fuertemente a los pacientes es el privado con 8.89 veces más probabilidades de que un paciente tenga un accidente cerebrovascular.

3. Conclusiones

- Los resultados obtenidos permiten responder las dudas que se planteó, ya que nos da una clara idea de cuáles son los atributos o variables que mayor peso tienen en la presencia de pacientes con problemas de accidentes cerebrovasculares.
- Los análisis univariantes y descriptivos de las variables no revelan una influencia tan directa sobre la variable de interés stroke, esto se debe a que se trata de un fenómeno más multidimensional, es decir donde el análisis de múltiples variables en el fenómeno es indispensable, en este caso la regresión logística es un buen instrumento ya que los accidentes cerebrovasculares en el dataset es un atributo dicotómico.
- El modelo de regresión logística revela algunas variables que son significativas para la variable dependiente, sin embargo los resultados más importante que nos arroja el modelo es el relacionado con el atributo del lugar de trabajo de los pacientes no cabe duda de que todos los pacientes que trabajan tienen altas probabilidades de tener problemas cerebrovasculares pero son las personas que trabajan como autónomas las que tienen un alto riesgo de sufrir un derrame cerebral, esto revela la importancia y la influencia del estrés que tienen que vivir muchas personas que trabajan respecto aquellos pacientes que no tienen ningún trabajo.
- Un hecho importante que es necesario mencionar es que el dataset utilizado presenta un desbalanceo respecto a la variable dependiente del modelo de regresión, es decir, no presentan igual cantidad de registros de pacientes que han presentado ataques cerebrovasculares y de pacientes que no han accidentes cerebrovasculares, este hecho desde luego puede influenciar en los resultados obtenidos, sin embargo los procesos y las técnicas usadas presentan robustez en cuanto a las pruebas y contrastes usados.

4. Recursos

- Gibergans, J. (2017). Regresión lineal múltiple. Material UOC.
- Maté, L. Pérez, D. Calvo M. (2017).Introducción a la limpieza y análisis de los datos. Material UOC.

Link github: https://github.com/jmoralesmoreiraUOC/stroke-cleaning

Link video: https://drive.google.com/drive/folders/1yocVVa3WEedhFf3hz5IlN3JsxMbc7XI7?usp=sharing

Contribuciones	Firma
Investigación previa	AM ,JL
Redacción de las respuestas	AM ,JL
Desarrollo del código	AM ,JL
Participación en el vídeo	AM ,JL