Práctica 2: Limpieza y validación de los datos

José Luis Rodríguez Andreu

18/5/2021

1. Introducción

El objetivo de esta práctica es la realización de un proceso de limpieza de un conjunto de datos. Para ello, se ha escogido el conjunto de datos Pima Indians Diabetes de la plataforma Kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes-database (https://www.kaggle.com/uciml/pima-indians-diabetes-database)).

Este conjunto de datos contiene información médica sobre mujeres relacionada con el diagnóstico de la diabetes. Consta de 768 observaciones con las siguientes variables:

- `Pregnancies` : Número de embarazos que ha tenido la paciente.
- `Glucose` : Concentración de glucosa en plasma a las 2 horas en una prueba oral de tolerancia a la glucosa.
- `BloodPressure` : Presión arterial diastólica (mm Hg)
- `SkinThickness` : Espesor del pliegue cutáneo del tríceps (mm)
- `Insulin` : Insulina en suero a las 2 horas (mu U/ml)
- `BMI` : Índice de masa corporal
- `DiabetesPedigreeFunction` : Función de pedigrí de la diabetes
- `Age` : Edad
- `Outcome` : Variable indicadora de que si la paciente padece o no de diabetes. La clase 1 indica que padece diabetes. 268 de 768 registros se encuentran etiquetados con 1.

2. Importancia y objetivo del análisis

Este conjunto de datos procede del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales de EEUU. El objetivo del conjunto de datos es realizar un diagnóstico predictivo sobre si un paciente tiene o no diabetes, basándose en determinadas mediciones de índole médica incluidas en el conjunto de datos. En este conjunto de datos, todos los pacientes registrados son mujeres de al menos 21 años de edad de origen “indio Pima”. Por lo tanto, el objetivo principal de este proyecto es el de definir unas pautas de limpieza y preprocesado de datos para poder realizar un diagnóstico automático basado en técnicas de aprendizaje automático lo mas preciso posible.

La importancia del diseño de modelos de machine learning capaces de realizar diagnósticos médicos con precisión ayuda a los profesionales de la medicina a la hora de atender pacientes, y desemboca en una mejora del propio sistema sanitario, ya que este tipo de diagnósticos permiten ahorrar costes en análisis y tiempo de los profesionales, permitiendo de esta manera la reinversión de esos gastos sanitarios en promover una atención médica de calidad, además de facilitar el trabajo al personal sanitario que frecuentemente, y mas en época de pandemia, se encuentra saturado por una alta carga de trabajo.

3. Integración y selección de los datos de interés a analizar

En este caso, vamos a trabajar con todas las variables descritas en el dataset, ya que todas tienen potencial para resultar útiles en la búsqueda de patrones que permitan la identificación automática de la diabetes.

4. Limpieza de los datos

Importamos el fichero `diabetes.csv`. Observamos que todas las variables son de tipo entero o numérico. La variable `Outcome` aunque toma valores numéricos de 0 y 1, es en reailidad una variable categórica binaria.

```
df = read.csv("../csv/diabetes.csv", encoding = 'UTF-8')
str(df)
```

```
## 'data.frame':    768 obs. of  9 variables:
##  $ Pregnancies             : int  6 1 8 1 0 5 3 10 2 8 ...
##  $ Glucose                 : int  148 85 183 89 137 116 78 115 197 125 ...
##  $ BloodPressure           : int  72 66 64 66 40 74 50 0 70 96 ...
##  $ SkinThickness           : int  35 29 0 23 35 0 32 0 45 0 ...
##  $ Insulin                 : int  0 0 0 94 168 0 88 0 543 0 ...
##  $ BMI                     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
##  $ Age                     : int  50 31 32 21 33 30 26 29 53 54 ...
##  $ Outcome                 : int  1 0 1 0 1 0 1 0 1 1 ...
```

```
head(df)
```

```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35       0 33.6
## 2           1      85            66            29       0 26.6
## 3           8     183            64             0       0 23.3
## 4           1      89            66            23      94 28.1
## 5           0     137            40            35     168 43.1
## 6           5     116            74             0       0 25.6
##   DiabetesPedigreeFunction Age Outcome
## 1                    0.627  50       1
## 2                    0.351  31       0
## 3                    0.672  32       1
## 4                    0.167  21       0
## 5                    2.288  33       1
## 6                    0.201  30       0
```

Realizamos un análisis estadístico inicial para conocer la distribución de los datos:

```
summary(df)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00
##     Outcome
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

4.1. Tratamiento de valores perdidos

Comprobamos que en este caso el conjunto de datos no presenta registros con valores nulos en alguna de sus variables. En el caso de que existiesen, se plantearía la posibilidad de realizar una imputación empleando alguna técnica de inferencia de ese valor a partir del resto de variables, ya que el conjunto de datos no es excesivamente grande y nos interesa mantener un buen número de registros.

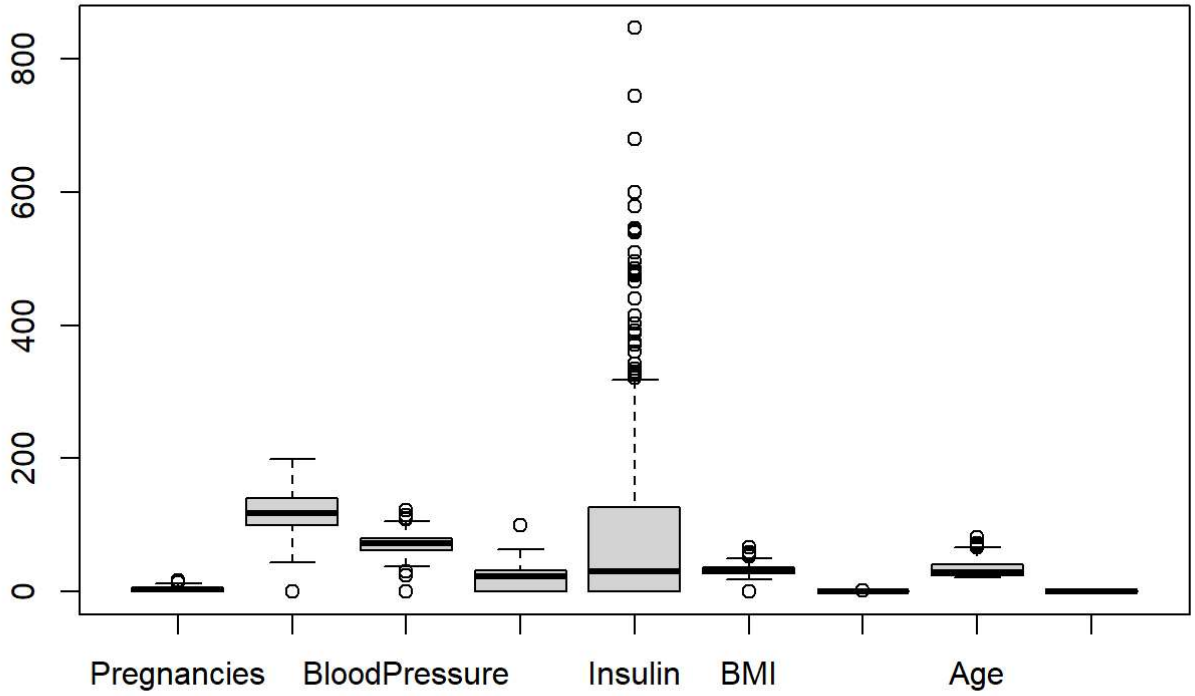
```
sapply(df, function(x) sum(is.na(x)))
```

```
##              Pregnancies                  Glucose            BloodPressure
##                        0                        0                        0
##            SkinThickness                  Insulin                      BMI
##                        0                        0                        0
## DiabetesPedigreeFunction                      Age                  Outcome
##                        0                        0                        0
```

4.2. Identificación y tratamiento de valores extremos

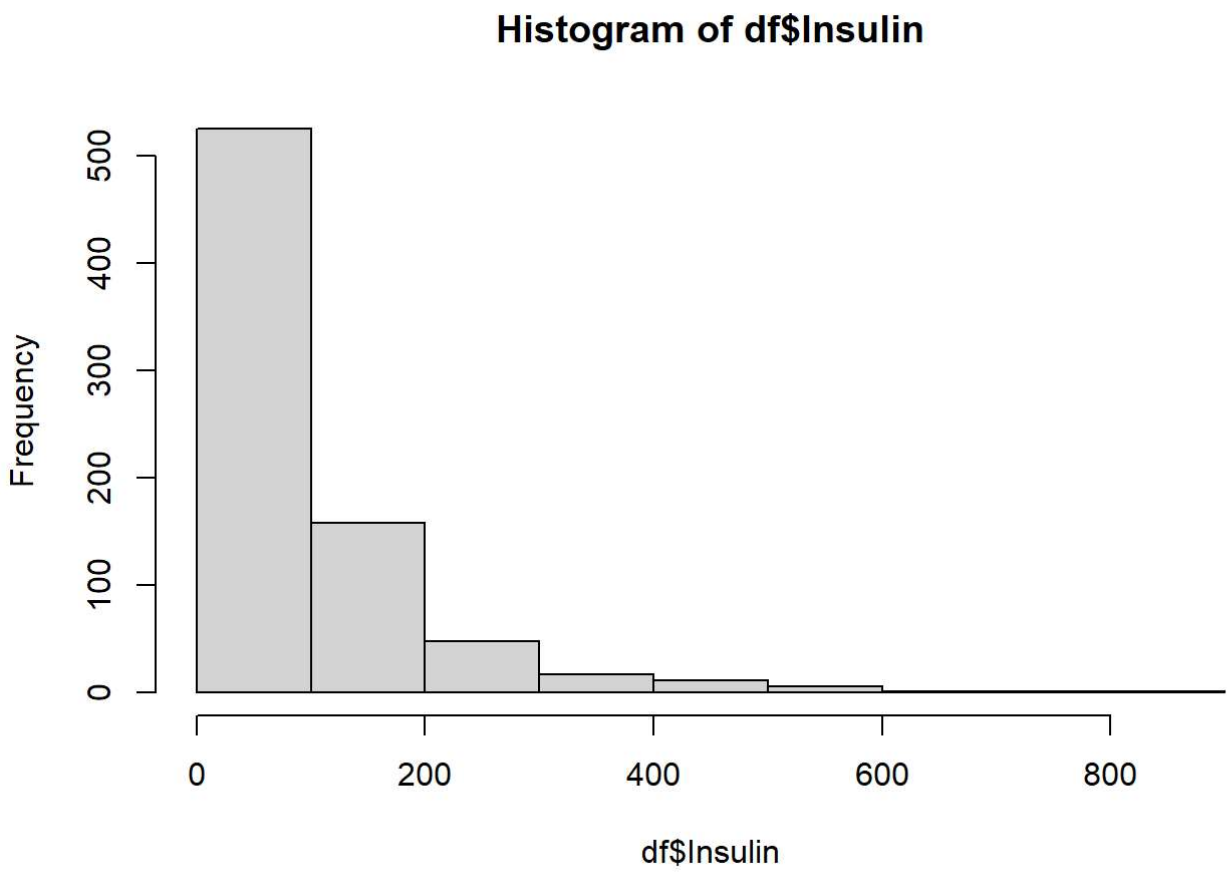
aplicamos un diagrama de cajas para observar si se observan valores anómalos u outliers:

```
boxplot(df)
```



Observamos que la mayoría de las variables tienen outliers, pero se encuentran en una posición cercana al cuerpo de la distribución. dado que no tenemos ningún tipo de restricción respecto al intervalo donde deben moverse los datos, los dejamos tal y como están. La única que presenta mas outliers visibles es Insulin. Estudiamos su distribución:

```
hist(df$Insulin)
```



Observamos que es una distribución descendente, la cual mantiene la gran mayoría de los registros en el intervalo de entre 0 y 200, pero se observa que la distribución cae de manera homogenea. Esto nos dice que esta variable sigue una distribución distinta a la normal, y esa es la causa de que el boxplot detecte tantos outliers. Dado que estos valores extremos no son casos puntuales, decidimos dejarlos como están, ya que pueden ser datos anómalos pero correctos.

5. Análisis de los datos

En esta sección se van a realizar una serie de análisis del conjunto de datos que nos ayudarán a la hora de encontrar diferentes características o patrones que sean relevantes a la hora de diagnosticar o no a una paciente con diabetes.

5.1. Selección de los grupos de datos que se quieren analizar

En este caso, se ha decidido trabajar con dos muestras: una que contenga todas las pacientes con diabetes y otra con las que no padezcan la enfermedad. El objetivo es realizar un análisis entre las dos muestras en base a las distintas variables con la idea de averiguar si éstas pueden ser o no significativas.

```
df_diab = df[df$Outcome == 0,]
df_no_diab = df[df$Outcome == 1,]
```

5.2. Comparación de la normalidad y homogeneidad de la varianza

En primer lugar, vamos a estudiar la normalidad de las variables numéricas de nuestro conjunto de datos completo. Para ello, emplearemos el test de normalidad de Anderson-Darling.

```
library(nortest)
```

```
## Warning: package 'nortest' was built under R version 4.0.3
```

```
alpha = 0.05


list_var = colnames(df)[1:length(colnames(df))-1]
list_pvalue = c()
list_is_normal = c()

for(i in 1:length(list_var)){
    pvalue = ad.test(df[, list_var[i]])$p.value
    result = ifelse(pvalue < alpha, "no", "si")
    list_pvalue[i] = pvalue
    list_is_normal[i] = result

}

normal_result = data.frame(
    list_var,
    list_pvalue,
    list_is_normal
)
colnames(normal_result) = c("variable", "p-valor", "¿Sigue distrib. normal?")


normal_result
```

```
##                    variable      p-valor ¿Sigue distrib. normal?
## 1               Pregnancies 3.700000e-24                      no
## 2                   Glucose 1.110732e-14                      no
## 3             BloodPressure 3.700000e-24                      no
## 4             SkinThickness 3.700000e-24                      no
## 5                   Insulin 3.700000e-24                      no
## 6                       BMI 2.028433e-10                      no
## 7 DiabetesPedigreeFunction 3.700000e-24                      no
## 8                       Age 3.700000e-24                      no
```

Según el test estadístico aplicado, ninguna de las variables numéricas sigue una distribución normal.

Ahora, vamos a estudiar la homogeneidad de la varianza en nuestro conjunto de datos. Para ello, empleamos el test de Fligner-Killeen para estudiar esta homogeneidad entre la población de pacientes que sufre diabetes y la población que no la padece, respecto al resto de variables.

```
alpha = 0.05


list_var = colnames(df)[1:length(colnames(df))-1]
list_pvalue = c()
list_homogen_variances = c()

for(i in 1:length(list_var)){
    pvalue = fligner.test(x = df[, list_var[i]], g = df[,"Outcome"])$p.value
    result = ifelse(pvalue < alpha, "no", "si")
    list_pvalue[i] = pvalue
    list_homogen_variances[i] = result

}

normal_result = data.frame(
    list_var,
    list_pvalue,
    list_homogen_variances
)
colnames(normal_result) = c("variable", "p-valor", "¿varianza homogenea respecto a Outcome?")


normal_result
```

```
##                   variable      p-valor ¿varianza homogenea respecto a Outcome?
## 1              Pregnancies 1.061921e-07                                      no
## 2                  Glucose 6.276164e-07                                      no
## 3            BloodPressure 3.376347e-01                                      si
## 4            SkinThickness 1.010218e-06                                      no
## 5                  Insulin 3.653696e-01                                      si
## 6                      BMI 5.022451e-02                                      si
## 7 DiabetesPedigreeFunction 4.356018e-06                                      no
## 8                      Age 3.469707e-04                                      no
```

En este caso, observamos que el test nos dice que las variables BloodPressure, Insulin y BMI presentan homogeneidad en la varianza respecto a Outcome. Pregnancies, Glucose, SkinThickness, DiabetesPEdigreeFunction y Age no tienen varianza homogenea en las dos categorías de Outcome.

5.3. Pruebas estadísticas

A continuación se van a realizar una serie de pruebas estadísticas para comparar las dos muestras que hemos obtenido del conjunto de datos: La muestra con pacientes sin diabetes y la muestra con pacientes que sufren de diabetes. El objetivo es encontrar diferencias significativas respecto al hecho de tener o no diabetes en la distribución de nuestras variables del conjunto de datos.

5.3.1. influencia de la diabetes en el numero de embarazos

Vamos a analizar estadísitcamente si existe una diferencia significativa en el numero de embarazos en pacientes diabéticas y no diabéticas. En primer lugar, calculamos el valor medio de la variable Pregnancies para las muestras:

```
mean(df_no_diab$Pregnancies)
```

```
## [1] 4.865672
```

```
mean(df_diab$Pregnancies)
```

```
## [1] 3.298
```

Observamos que el valor medio de Pregnancies es superior en pacientes no diabéticas que diabéticas, por lo que planteamos la siguiente pregunta de investigación:

¿El número de embarazos es superior en pacientes sin diabetes que con diabetes?

Para ello, planteamos la hipótesis nula y alternativa:

- H_0: No hay diferencia en el número de embarazos en pacientes no diabeticas y diabeticas

$$H_0 : \mu_{ND} = \mu_D$$

- H_1: EL numero de embarazos es significativamente mayor en pacientes no diabéticas que en pacientes diabéticas:

$$H_0 : \mu_{ND} > \mu_D$$

Especificaciones del contraste de hipótesis:

- Contraste de dos muestras no relacionadas sobre la media
- En base al Teorema del Límite Central, asumimos normalidad en muestras grandes (generalmente superior a 30 observaciones)
- Aplicamos un test paramétrico, unilateral por la derecha. *Aplicamos un test de igualdad de varianzas para saber si asumimos homocedasticidad o heterocedasticidad:

```
var.test(df_no_diab$Pregnancies, df_diab$Pregnancies)
```

```
##
##  F test to compare two variances
##
## data:  df_no_diab$Pregnancies and df_diab$Pregnancies
## F = 1.5375, num df = 267, denom df = 499, p-value = 4.246e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.249880 1.904318
## sample estimates:
## ratio of variances
##           1.537543
```

Dado que obtenemos un p-valor muy pequeño, podemos considerar que las varianzas son distintas (heterocedasticidad).

Sabiendo esto, realizamos el contraste:

```
t.test(df_no_diab$Pregnancies, df_diab$Pregnancies, alternative="greater", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  df_no_diab$Pregnancies and df_diab$Pregnancies
## t = 5.907, df = 455.96, p-value = 3.411e-09
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.130249      Inf
## sample estimates:
## mean of x mean of y
##  4.865672  3.298000
```

Observamos que obtenemos un p-valor muy pequeño, inferior a 0.05, por lo que podemos rechazar la hipótesis nula de igualdad de medias entre las dos poblaciones, y concluir que el número de embarazos es mayor en la población no diabética que en la diabética, con un nivel de significancia del 0.05.

5.3.2 diferencias en la proporcion de mujeres con obesidad en pacientes diabeticas y no diabeticas

A continuación vamos a esudiar si existe alguna diferencia significativa en la proporción de pacientes con obesidad (BMI >= 30) en la muestra poblacional de pacientes diabéticas y no diabéticas.

La pregunta de investigación que nos planteamos es la siguiente:

¿Existe diferencia significativa en la proporción de mujeres con BMI >= 30 en pacientes diabéticas o no diabéticas?

Definimos la hipótesis nula y alternativa:

- H_0: La proporción de pacientes con BMI igual o superior a 30 es la misma en pacientes diabéticas y no diabéticas.

$$p_{ND} = p_D$$

* H_1: La proporción de pacientes con BMI igual o superior a 30 es diferente en pacientes diabéticas y no diabéticas.

$$p_{ND} \neq p_D$$

Especificaciones del contraste de hipótesis:

- Contraste de dos muestras no relacionadas sobre la proporción
- En base al Teorema del Límite Central, asumimos normalidad en muestras grandes (generalmente superior a 30 observaciones)
- Aplicamos un test paramétrico bilateral.

Aplicamos el contraste:

```
n_no_diab = length(df_no_diab$BMI)
n_diab = length(df_diab$BMI)

p_no_diab = sum(df_no_diab$BMI >= 30)/n_no_diab
p_diab = sum(df_diab$BMI >= 30)/n_diab

success = c(n_no_diab * p_no_diab, n_diab * p_diab)
nn = c(n_no_diab, n_diab)

prop.test(success, nn, alternative = "two.sided", correct = FALSE, conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 71.32, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.2474302 0.3748982
## sample estimates:
##    prop 1    prop 2
## 0.8171642 0.5060000
```

En este caso, obtenemos un p-valor muy pequeño, por lo que podemos rechazar la hipótesis nula de igualdad de proporciones y concluir que la proporción de pacientes con obesidad es diferente en pacientes no diabéticas y diabéticas.

5.3.4. Correlación entre variables descriptivas

Vamos a realizar un análisis de correlación entre las variables cuantitativas de las que disponemos, con el objetivo de comprobar si sus distribuciones están relacionadas:

```
library(PerformanceAnalytics)
```

```
## Warning: package 'PerformanceAnalytics' was built under R version 4.0.5
```

```
## Loading required package: xts
```

```
## Warning: package 'xts' was built under R version 4.0.5
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.0.5
```

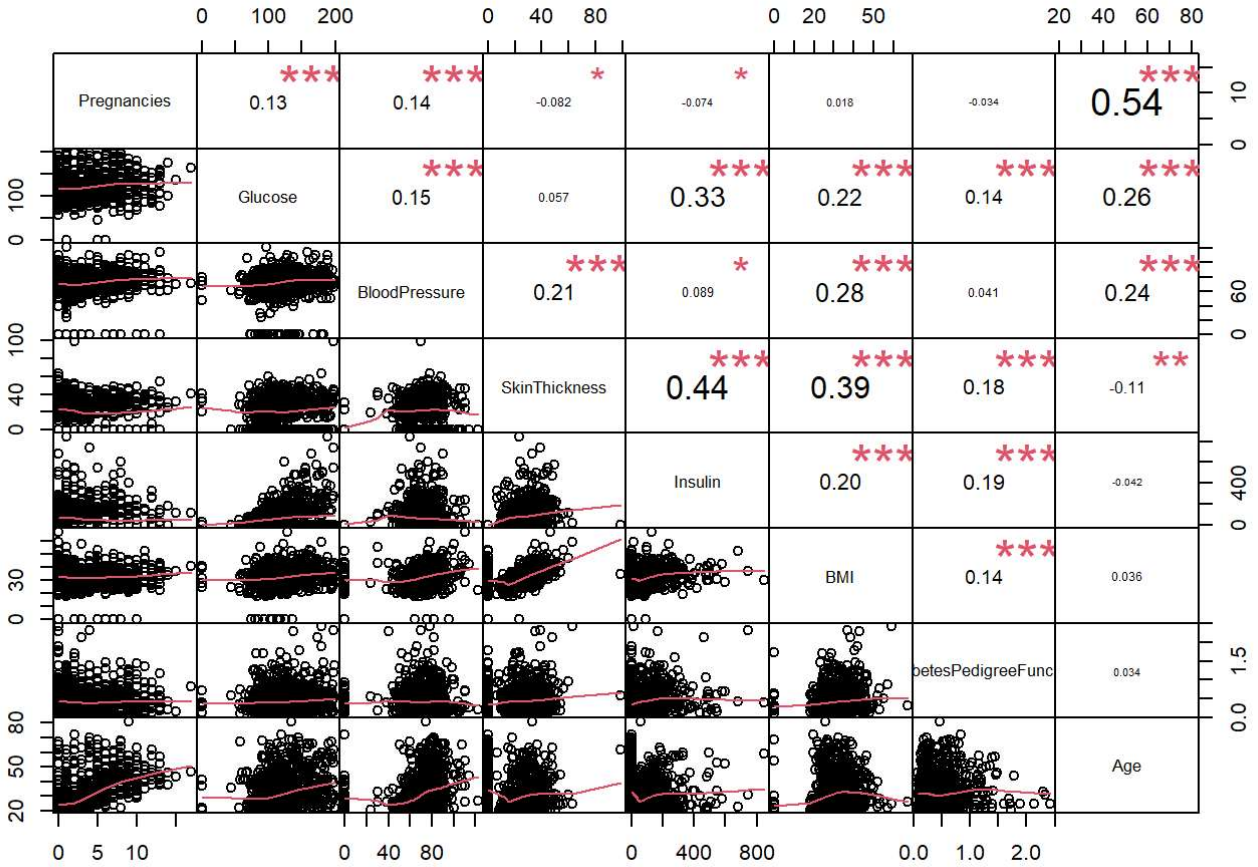
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
## Attaching package: 'PerformanceAnalytics'
```

```
## The following object is masked from 'package:graphics':
##
##     legend
```

```
#data.frame(round(cor(df[,list_var]), 2))
chart.Correlation(df[,list_var], histogram = F, pch = 19)
```



Con esta gráfica observamos que no existe un nivel de correlación excesivamente alto en nuestro conjunto de variables. Observamos que Pregnancies y Age tienen un coeficiente de correlación de 0.54, lo cual tiene cierto sentido debido a que a mas edad, mas posibilidad de haber tenido mas hijos. También se percibe un coeficiente de correlación del 0.44 entre SkinThickness e Insulin.

5.3.5. Modelo de regresión logística

Vamos a construir un modelo de regresión logística a partir de nuestro conjunto de datos, que sea capaz de identificar una persona con diabetes o no en base a las variables del conjunto de datos. Esto nos ayudará también a ver que variables influyen mas en la probabilidad de padecer diabetes.

```
model_glm = glm(as.factor(Outcome) ~ ., data = df, family=binomial)

summary(model_glm)
```

```
##
## Call:
## glm(formula = as.factor(Outcome) ~ ., family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5566  -0.7274  -0.4159   0.7267   2.9297
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -8.4046964  0.7166359 -11.728  < 2e-16 ***
## Pregnancies               0.1231823  0.0320776   3.840 0.000123 ***
## Glucose                   0.0351637  0.0037087   9.481  < 2e-16 ***
## BloodPressure            -0.0132955  0.0052336  -2.540 0.011072 *
## SkinThickness             0.0006190  0.0068994   0.090 0.928515
## Insulin                  -0.0011917  0.0009012  -1.322 0.186065
## BMI                       0.0897010  0.0150876   5.945 2.76e-09 ***
## DiabetesPedigreeFunction  0.9451797  0.2991475   3.160 0.001580 **
## Age                       0.0148690  0.0093348   1.593 0.111192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 723.45  on 759  degrees of freedom
## AIC: 741.45
##
## Number of Fisher Scoring iterations: 5
```

Observando los resultados del modelo, vemos que nos dice que las variables mas influyentes en el cálculo de la probabilidad de tener o no diabetes son Pregnancies, Glucose, BMI, DiabetesPedigreeFunction y BloodPressure. Observando sus coeficientes vemos que estas variables, en su mayoría, tienen un efecto positivo en la probabilidad de padecer diabetes. Esto quiere decir que la posibilidad de tener diabetes aumenta con valores mas elevados en estas variables. La única variable de las mas influyentes que tiene un coeficiente negativo es BloodPressure, lo cual nos dice que un mayor valor en esta variable inplica una reducción en la posibilidad de padecer diabetes.

Construimos otro modelo empleando únicamente las variables influyentes:

```
model_glm2 = glm(as.factor(Outcome) ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction
+ BloodPressure, data = df, family=binomial)

summary(model_glm2)
```

```
##
## Call:
## glm(formula = as.factor(Outcome) ~ Pregnancies + Glucose + BMI +
##     DiabetesPedigreeFunction + BloodPressure, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7931  -0.7362  -0.4188   0.7251   2.9555
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -7.954952   0.675823 -11.771  < 2e-16 ***
## Pregnancies               0.153492   0.027835   5.514  3.5e-08 ***
## Glucose                   0.034658   0.003394  10.213  < 2e-16 ***
## BMI                       0.084832   0.014125   6.006  1.9e-09 ***
## DiabetesPedigreeFunction  0.910628   0.294027   3.097  0.00195 **
## BloodPressure            -0.012007   0.005031  -2.387  0.01700 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 728.56  on 762  degrees of freedom
## AIC: 740.56
##
## Number of Fisher Scoring iterations: 5
```

Comparando ambos modelos, vemos que el segundo toma un valor del AIC (Criterio de información de AKAIKE) inferior. Este parámetro nos indica un modelo de mayor calidad conforme disminuye su valor, por lo que este segundo modelo que emplea únicamente las variables consideradas significativas se ajusta mejor a los datos que el anterior, que emplea todas.

6. Conclusiones y resolución del problema

Hemos realizado un análisis del conjunto de datos para encontrar algún tipo de relación entre las variables que disponemos y el hecho de padecer o no diabetes. Las conclusiones que obtenemos son las siguientes:

- Las variables numéricas no siguen una distribución normal.
- BloodPresure, Insulin y BMI presentan varianza homogénea en la muestra de pacientes diabéticas y no diabéticas. El resto presentan varianza heterogénea.
- Hemos concluido que el número de embarazos es significativamente mayor en las pacientes no diabéticas que en las diabéticas, con un nivel de confianza del 95%.
- Del mismo modo, hemos concluido con un nivel de confianza del 95% que la proporción de pacientes con obesidad (BMI > 30) es sinificativamente diferente en la población diabética y no diabética.
- Del análisis de correlación obtenemos que no hay correlaciones altas entre las variables disponibles. Destaca la correlación entre Age y Pregnancies, lo cual tiene sentido debido a que las mujeres con mas edad pueden haber tenido mas hijos.
- Hemos construido un modelo de regresión logística para estudiar la influencia de las variables en el hecho de padecer o no diabetes, y hemos concluido que las variables mas significativas según el modelo son Pregnancies, Glucose, BMI, DiabetesPedigreeFunction y BloodPressure.

Con estos resultados podemos construir un sistema que permita identificar pacientes diabéticas en base a la información que tengamos de estas variables. Este modelo de regresión logística nos sirve de ejemplo de aplicación de una herramienta que clasifique a una paciente observada como diabética o no diabética.

```
newdata = data.frame(
    Pregnancies = 5,
    Glucose = 115,
    BloodPressure = 74,
    SkinThickness = 0,
    Insulin = 0,
    BMI = 25.6,
    DiabetesPedigreeFunction = 0.201,
    Age = 30
)

prediction = ifelse(predict(model_glm2, newdata) <0.5, 0, 1)

prediction
```

```
## 1
## 0
```