

Practica 2 – Limpieza y validación de los datos

M2.854 - Tipología de vida y Ciclo de los Datos

Juan Carlos Ghiringhelli Jueguen, Juan Pablo Botero Suaza

1 de junio, 2019

Contents

1	Introducción	1
1.1	Descripción del dataset	2
2	Análisis descriptivo	2
2.1	Carga de los datos	2
2.2	Análisis visual	4
3	Integración y selección de los datos de interés a analizar	5
4	Limpieza de los datos	5
4.1	Identificación y tratamiento de valores nulos, vacíos y ceros	5
4.2	Imputación de valores	6
4.3	Identificación y tratamiento de valores extremos	6
5	Análisis de los datos	7
5.1	Correlaciones	7
5.2	Comprobación de la normalidad y homogeneidad de la varianza	8
5.3	Análisis inferencial	9
5.4	Análisis predictivo	10
6	Presentación de resultados	13
7	Conclusiones	13
8	Bibliografía	13

1 Introducción

1.1 Descripción del dataset

El conjunto de datos elegido registra, en el contexto de consultas para diagnosticar diabetes, datos predictores médicos provenientes de un conjunto de pruebas realizadas a mujeres de la India.

Como suele ocurrir con datos predictivos médicos, es posible anticipar el diagnóstico de la condición médica para pacientes sin necesidad de realizar la prueba específica. Para casos de alto riesgo se podrían tomar las medidas adecuadas como notificación al paciente o tomar precauciones ante el ingreso del paciente a una operación o una emergencia. También sirve para comprender mejor las causas de la condición y la correlación de diferentes valores con la posibilidad de padecerla.

El contenido se descargó del siguiente enlace de Kaggle: <https://www.kaggle.com/uciml/pima-indians-diabetes-database> y contiene nueve atributos:

1. Pregnancies: Cantidad de veces que la paciente estuvo embarazada.
2. Glucose: Concentración de glucosa en sangre dentro de dos horas de una prueba de resistencia oral a la glucosa, medido en miligramo por decilitro.
3. BloodPressure: Presión de sangre, en milímetros por mercurio, una medida médica equivalente a la presión de una columna de mercurio de un mm de alto a 0°C a una atmosfera.
4. SkinThickness: grosor de la piel en mm en la zona del pliegue del tríceps.
5. Insulin: Insulina administrada por suero en la ultima hora, en unidades por mililitro.
6. BMI: de Body Mass Index, índice de masa corporal, medido en (kilos/altura)².
7. DiabetesPedigreeFunction: función de la condición presente en parientes para asignar una probabilidad genética de heredarla.
8. Age: edad en años.
9. Outcome: Presencia de la condición diabetes, siendo 1 positivo y 0 negativo. El conjunto presenta 268 casos positivos sobre un total de 768.

2 Análisis descriptivo

2.1 Carga de los datos

Cargar el archivo de datos diabetes.csv, validar que los tipos de datos son los correctos. Si no es así, conviértelos al tipo de datos adecuado.

```
indian_diabetes<-read.csv("diabetes.csv", header=TRUE, sep=",", dec=".")
str(indian_diabetes)
```

```
## '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 ...
```

El set de datos contiene 9 variables con un total de 768 observaciones

```
#Convertir variable objetivo a tipo factor para análisis posteriores
indian_diabetes$Outcome<-as.factor(indian_diabetes$Outcome)

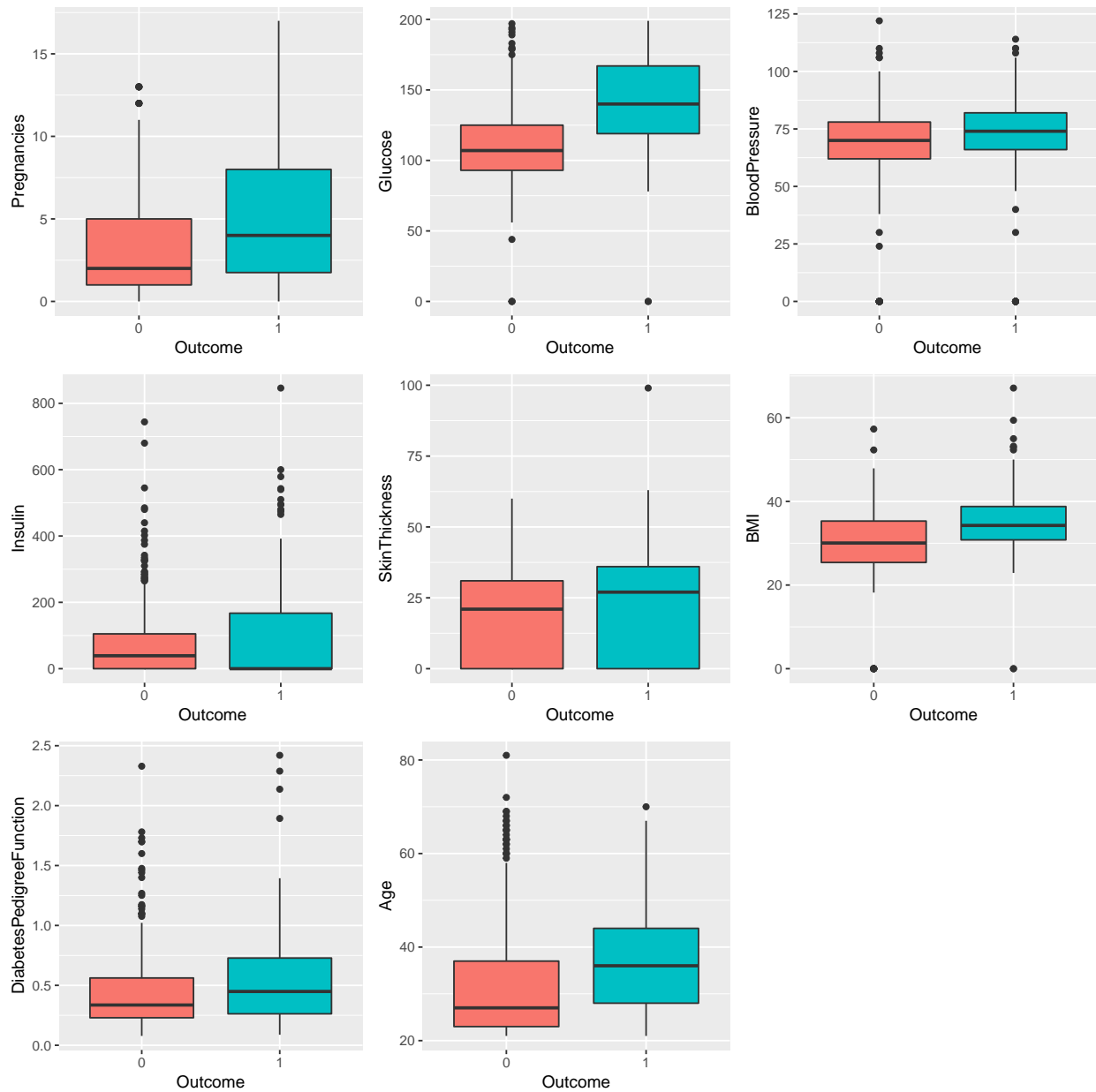
head(indian_diabetes)
```

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

```
summary(indian_diabetes)
```

```
##      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
## 0:500
## 1:268
##
##
##
##
```

2.2 Análisis visual



A partir de los gráficos de boxplot podemos inferir que algunas variables pueden tener mayor incidencia sobre la condición diabética para una mujer de la muestra, donde resalta notoriamente el número de embarazos(Pregnancies), la edad (Age), el nivel de glucosa (Glucose) y el índice de masa corporal (BMI), en los apartados posteriores analizaremos si esta afirmación gráfica también tiene su correspondencia desde el punto de vista estadístico inferencial/predictivo.

3 Integración y selección de los datos de interés a analizar

Inicialmente, para nuestro caso se consideraran todas las variables independientes para construir el modelo explicativo sobre el atributo Outcome, en caso de identificarse variables estadísticamente no significativas durante el proceso, serán excluidas de dichos modelos con el fin de obtener un resultado más adecuado para explicar el diagnóstico de presencia de diabetes sobre esta muestra de la población.

4 Limpieza de los datos

4.1 Identificación y tratamiento de valores nulos, vacíos y ceros

```
colSums(is.na(indian_diabetes))
```

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

```
colSums(indian_diabetes=="")
```

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

```
colSums(indian_diabetes==0)
```

##	Pregnancies	Glucose	BloodPressure
##	111	5	35
##	SkinThickness	Insulin	BMI
##	227	374	11
##	DiabetesPedigreeFunction	Age	Outcome
##	0	0	500

El conjunto de datos no presenta campos vacíos ni nulos, por lo que no será necesario tratarlos en este aspecto.

Los atributos de cantidad de embarazos e insulina en sangre presentan casos con 0, lo cual es coherente, tanto por no haber estado embarazada nunca la persona como por no consumir insulina. Si bien puede resultar de una falta de datos, se asumirá que todos estos casos son válidos.

Los valores de glucosa, presión de sangre, grosor de la piel en el tríceps, e índice de masa corporal presentan valores en cero. Consideraremos que esto denota una falta de datos, ya que en cualquiera de estos casos el paciente estaría muerto o herido de gravedad.

Para cada caso:

- Glucosa: el nivel de glucosa en sangre, a diferencia de la presión, no tiene un valor constante que refleje un buen estado de salud. Si bien los extremos siempre son peligrosos, esta presenta normalmente una gran varianza dependiendo de la última vez que se consumió glucosa, que cantidad, en que forma, con que alimentos y dependiendo del metabolismo y actividad inmediata. Dentro de los valores analizados, es el que presenta más variabilidad y valores extremos, por lo que tomaremos el valor mediano para imputar.
- Presión de sangre: Muchos valores presentan una variación muy alta o baja, y ni el promedio ni la mediana representan realmente un valor neutro, lo que haría que puedan tener un peso sobre la predicción siendo un valor desconocido. Utilizando registros médicos, seleccionaremos el valor 105, valor considerado normal para mujeres de cualquier edad, dando por supuesto que si no se sabe el valor es porque no se consideró importante anotarlo o medirlo.
- Grosor de piel: Los valores presentan un rango mediano y valores bien distribuidos, por lo que consideraremos el atributo como simétrico y usaremos el promedio.
- Índice de masa corporal (IBM): similar caso al grosor de piel, usaremos el promedio.

4.2 Imputación de valores

```
#imputar glucose con valor de la mediana
glucose_median <- median(indian_diabetes$Glucose[indian_diabetes$Glucose!=0])
indian_diabetes$Glucose <- ifelse(indian_diabetes$Glucose==0, glucose_median, indian_diabetes$Glucose)

#imputar blood pressure con valor tipico
indian_diabetes$BloodPressure <- ifelse(indian_diabetes$BloodPressure==0, 105, indian_diabetes$BloodPressure)

#imputar skin thick con valor de la media
skin_mean <- as.integer(mean(indian_diabetes$SkinThickness[indian_diabetes$SkinThickness!=0]))
indian_diabetes$SkinThickness <- ifelse(indian_diabetes$SkinThickness==0, skin_mean, indian_diabetes$SkinThickness)

#imputar BMI con valor de la media
bmi_mean <- lapply(mean(indian_diabetes$BMI[indian_diabetes$BMI!=0]), round, 1)[[1]]
indian_diabetes$BMI <- ifelse(indian_diabetes$BMI==0, bmi_mean, indian_diabetes$BMI)
```

4.3 Identificación y tratamiento de valores extremos

```
#analysis outliers

outlier_glucose <- boxplot.stats(indian_diabetes$Glucose)$out
outlier_glucose
```

```
## integer(0)
```

```
outlier_bmi <- boxplot.stats(indian_diabetes$BMI)$out
outlier_bmi
```

```
## [1] 53.2 55.0 67.1 52.3 52.3 52.9 59.4 57.3
```

```
outlier_ins <- boxplot.stats(indian_diabetes$Insulin)$out
outlier_ins
```

```
## [1] 543 846 342 495 325 485 495 478 744 370 680 402 375 545 360 325 465
## [18] 325 415 579 474 328 480 326 330 600 321 440 540 480 335 387 392 510
```

```
outlier_age <- boxplot.stats(indian_diabetes$Age)$out
outlier_age
```

```
## [1] 69 67 72 81 67 67 70 68 69
```

```
outlier_pregnancies <- boxplot.stats(indian_diabetes$Pregnancies)$out
outlier_pregnancies
```

```
## [1] 15 17 14 14
```

```
outlier_bpressure <- boxplot.stats(indian_diabetes$BloodPressure)$out
outlier_bpressure
```

```
## [1] 30 110 122 30 110 110 24 114
```

```
outlier_skin <- boxplot.stats(indian_diabetes$SkinThickness)$out
outlier_skin
```

```
## [1] 45 47 11 47 11 10 60 13 13 54 51 56 14 13 50 44 12 46 44 13 44 54 14
## [24] 7 50 52 10 44 43 45 14 10 11 12 43 13 12 48 43 43 8 13 14 12 49 46
## [47] 46 11 8 12 63 12 45 13 48 13 10 45 7 52 49 43 14 47 99 46 11 50 45
## [70] 14 13 13 47 12 48 43 46 46 45 10 46 49 11 13 46 44 48
```

```
outlier_dpf <- boxplot.stats(indian_diabetes$DiabetesPedigreeFunction)$out
outlier_dpf
```

```
## [1] 2.288 1.441 1.390 1.893 1.781 1.222 1.400 1.321 1.224 2.329 1.318
## [12] 1.213 1.353 1.224 1.391 1.476 2.137 1.731 1.268 1.600 2.420 1.251
## [23] 1.699 1.258 1.282 1.698 1.461 1.292 1.394
```

5 Análisis de los datos

5.1 Correlaciones

```
#correlacion entre variables independientes
cor_matrix <- cor(indian_diabetes[,1:8])
cor_matrix
```

```
##              Pregnancies  Glucose BloodPressure SkinThickness
## Pregnancies      1.00000000 0.1282130   0.17486942   0.08176982
## Glucose          0.12821296 1.00000000   0.17709057   0.19261490
## BloodPressure    0.17486942 0.1770906   1.00000000   0.16416674
## SkinThickness    0.08176982 0.1926149   0.16416674   1.00000000
## Insulin          -0.07353461 0.3317094  -0.11503081   0.11654686
## BMI              0.02156673 0.2308111   0.24124422   0.54315004
## DiabetesPedigreeFunction -0.03352267 0.1373269  -0.02940198   0.10218827
## Age             0.54434123 0.2669092   0.25990171   0.12610719
##              Insulin      BMI DiabetesPedigreeFunction
## Pregnancies      -0.07353461 0.02156673             -0.03352267
## Glucose           0.33170943 0.23081108             0.13732692
## BloodPressure    -0.11503081 0.24124422             -0.02940198
## SkinThickness     0.11654686 0.54315004             0.10218827
## Insulin           1.00000000 0.18471281             0.18507093
## BMI               0.18471281 1.00000000             0.15338959
## DiabetesPedigreeFunction 0.18507093 0.15338959             1.00000000
## Age              -0.04216295 0.02549815             0.03356131
##              Age
## Pregnancies      0.54434123
## Glucose          0.26690916
## BloodPressure    0.25990171
## SkinThickness    0.12610719
## Insulin          -0.04216295
## BMI              0.02549815
## DiabetesPedigreeFunction 0.03356131
## Age              1.00000000
```

5.2 Comprobación de la normalidad y homogeneidad de la varianza

Utilizar test Shapiro-Wilk para confirmar presunción de normalidad.

H0: los datos provienen de una distribución normal.

H1: los datos no provienen de una distribución normal.

Nivel de significancia: 0.05

```
shapiro.test(indian_diabetes$Glucose)$p.value
```

```
## [1] 1.523253e-11
```

```
shapiro.test(indian_diabetes$Age)$p.value
```

```
## [1] 2.402274e-24
```

```
shapiro.test(indian_diabetes$BloodPressure)$p.value
```

```
## [1] 3.867717e-09
```



```
shapiro.test(indian_diabetes$SkinThickness)$p.value
```

```
## [1] 9.482143e-19
```

```
shapiro.test(indian_diabetes$DiabetesPedigreeFunction)$p.value
```

```
## [1] 2.477506e-27
```

```
shapiro.test(indian_diabetes$Insulin)$p.value
```

```
## [1] 7.914332e-34
```

```
shapiro.test(indian_diabetes$BMI)$p.value
```

```
## [1] 6.553269e-09
```

```
shapiro.test(indian_diabetes$Pregnancies)$p.value
```

```
## [1] 1.609257e-21
```

De acuerdo al test de Shapiro-Wilk, todo los **valores p** obtenidos son inferiores al nivel de significancia, por lo tanto se rechaza la hipótesis nula, es decir ninguna de las variables proviene de una distribución normal.

5.3 Análisis inferencial

Utilizar test “Levene” para confirmar homogeneidad de la varianza de acuerdo a los resultados obtenidos en el test de normalidad. En este caso se utiliza un test no paramétrico dado que ninguna de las variables cumple el supuesto de normalidad.

H0: las varianzas de las muestras son iguales

H1: el menos 2 de las varianzas difieren en las muestras.

Nivel de significancia: 0.05

Preguntas:

El valor medio en la presión sanguínea (BloodPressure) es estadísticamente significativo para mujeres que presentan diabetes (Outcome=1) respecto a aquellas que no la presentan (Outcome=0) ?.

```
library(car)
```

```
#la prueba revela un valor de p mayor que 0.05, lo que indica que no hay una diferencia significativa e  
leveneTest(BloodPressure ~ Outcome, data = indian_diabetes)
```

	Df	F value	Pr(>F)
group	1	0.0626857	0.8023675
	766	NA	NA

```
bp_sdiabetes<- indian_diabetes[indian_diabetes$Outcome==1, "BloodPressure"]
bp_ndiabetes<- indian_diabetes[indian_diabetes$Outcome==0, "BloodPressure"]
```

Como el valor p resultante es menor que el nivel de significación de .05, rechazamos la hipótesis nula
`wilcox.test(bp_sdiabetes,bp_ndiabetes,alternative="greater")`

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: bp_sdiabetes and bp_ndiabetes
## W = 81508, p-value = 3.587e-07
## alternative hypothesis: true location shift is greater than 0
```

El valor medio en la glucosa (Glucose) es estadísticamente significativo para mujeres que presentan diabetes (Outcome=1) respecto a aquellas que no la presentan (Outcome=0) ?.

#la prueba revela un valor de p menor que 0.05, lo que indica que hay una diferencia significativa en la
`leveneTest(Glucose ~ Outcome, data = indian_diabetes)`

	Df	F value	Pr(>F)
group	1	24.03165	1.2e-06
	766	NA	NA

```
gl_sdiabetes<- indian_diabetes[indian_diabetes$Outcome==1, "Glucose"]
gl_ndiabetes<- indian_diabetes[indian_diabetes$Outcome==0, "Glucose"]
```

Como el valor p resultante es menor que el nivel de significación de .05, rechazamos la hipótesis nula
`wilcox.test(gl_sdiabetes,gl_ndiabetes,alternative="greater")`

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: gl_sdiabetes and gl_ndiabetes
## W = 106070, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

5.4 Análisis predictivo

Construir un modelo de clasificación que permita predecir a partir de las variables independientes objeto de estudio, si una mujer podría tener o no una condición diabética en su organismo.

Propuesta I: Regresión logística.

```
set.seed(1234)
```

#Al establecer el parámetro type='response', R generará probabilidades de la forma de P (y = 1 | X). N

```
train_idx <- sample(1:nrow(indian_diabetes),nrow(indian_diabetes)*0.7,replace=FALSE)
train<-indian_diabetes[train_idx,]
```

```

test<-indian_diabetes[-train_idx,]

modelo_rl<-glm(Outcome ~ . , data=indian_diabetes,family=binomial(link=logit))

#Evaluar si alguno de los regresores tiene influencia significativa (p-valor del contraste individual i
summary(modelo_rl)

##
## Call:
## glm(formula = Outcome ~ ., family = binomial(link = logit), data = indian_diabetes)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.7072  -0.7113  -0.3946   0.7061   2.4844
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -9.6573773   0.8208482  -11.765  < 2e-16 ***
## Pregnancies     0.1208219   0.0323080   3.740 0.000184 ***
## Glucose         0.0380529   0.0038443   9.898  < 2e-16 ***
## BloodPressure   0.0005079   0.0074491   0.068 0.945645
## SkinThickness   0.0038679   0.0132473   0.292 0.770306
## Insulin        -0.0014077   0.0008417  -1.673 0.094415 .
## BMI             0.0902918   0.0178297   5.064 4.1e-07 ***
## DiabetesPedigreeFunction 0.9325081  0.2987518   3.121 0.001800 **
## Age            0.0089128   0.0093585   0.952 0.340906
## ---
## 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: 711.96  on 759  degrees of freedom
## AIC: 729.96
##
## Number of Fisher Scoring iterations: 5

modelo_rl2<-glm(Outcome ~ Pregnancies+Glucose+ BMI+DiabetesPedigreeFunction , data=indian_diabetes,fam
summary(modelo_rl2)

```

```

##
## Call:
## glm(formula = Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction,
##      family = binomial(link = logit), data = indian_diabetes)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.8221  -0.7246  -0.3996   0.7255   2.4339
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -9.184046   0.705727  -13.014  < 2e-16 ***

```

```
## Pregnancies          0.143276    0.027541    5.202 1.97e-07 ***
## Glucose              0.036879    0.003487   10.575 < 2e-16 ***
## BMI                 0.088717    0.014729    6.023 1.71e-09 ***
## DiabetesPedigreeFunction 0.881577    0.294667    2.992 0.00277 **
## ---
## 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: 716.38  on 763  degrees of freedom
## AIC: 726.38
##
## Number of Fisher Scoring iterations: 5
```

*#El menor valor para el indicador AIC corresponde al modelo con los regresores *Pregnancies + Glucose +*

#Matriz de confusión y precisión del modelo 1

```
precdb <- predict(modelo_rl,newdata=test,type='response')
precdb <- ifelse(precdb > 0.70,1,0)
misClasificError <- mean(precdb != test$Outcome)

print(paste('Precisión en la clasificación:',round((1-misClasificError)*100,digits = 2),'%'))
```

```
## [1] "Precisión en la clasificación: 76.62 %"
```

```
table(test$Outcome,precdb)
```

```
##      precdb
##      0    1
## 0 146    8
## 1   46   31
```

#Matriz de confusión y precisión del modelo 2

```
precdb2 <- predict(modelo_rl2,newdata=test[,c("Pregnancies","Glucose","BMI","DiabetesPedigreeFunction")])
precdb2 <- ifelse(precdb2 > 0.70,1,0)
misClasificError2 <- mean(precdb2 != test$Outcome)

print(paste('Precisión en la clasificación:',round((1-misClasificError2)*100,digits = 2),'%'))
```

```
## [1] "Precisión en la clasificación: 76.19 %"
```

```
table(test$Outcome,precdb2)
```

```
##      precdb2
##      0    1
## 0 145    9
## 1   46   31
```

Los falsos positivos corresponde a casos en que la predicción de la probabilidad de la respuesta afirmativa es elevada , pero la respuesta observada es negativa, en nuestro caso para 9 mujeres, el modelo indica que tiene condición diabética “1”, pero en realidad no la tiene “0”.

Los falsos negativos corresponde a casos donde el modelo predice que una mujer tiene una probabilidad de condición diabética baja, sin embargo las mujeres observadas si presentan una condición diabética, 46 individuos en este caso.

Ademas se puede evidenciar que para obtener un nivel de precisión similar al modelo con todas las variables, tan solo basta con utilizar las 4 variables explicativas comentadas anteriormente.

Propuesta II: Arboles de desicion C50.

6 Presentación de resultados

7 Conclusiones

8 Bibliografía

Calvo M., Subirats L., Perez D. (2019). Introduccion a la limpieza y analisis de los datos.Editorial UOC.

Megan Squire (2015). Clean Data. Packt Publishing Ltd.

Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques.

Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.

Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.

Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.

Tutorial de Github <https://guides.github.com/activities/hello-world>.