1. Detalles de la actividad

1.1. Descripción

1.2. Objetivos

1.3. Competencias

2. Resolución

2.1. Descripción del dataset

2.2. Importancia y objetivos de los análisis

2.3. Limpieza de los datos

2.3.1. Selección de los datos de interés

2.3.2. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

2.3.3. Identificación y tratamiento de valores extremos

2.3.4. Exportación de los datos preprocesados

2.4.Análisis de los datos.

2.4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

2.4.2. Comprobación de la normalidad y homogeneidad de la varianza.

2.5. Representación de los resultados a partir de tablas y gráficas.

2.5.1. ¿Qué variables cuantitativas influyen más la infección post operatoria?

2.5.2. Modelo de regresión lineal simple

2.5.3. Modelo de regresión lineal múltiple (regresores cuantitativos)

2.5.4. Modelo de regresión lineal múltiple (regresores cuantitativos y cualitativos)

2.5.5.Efectuar una predicción de la concentración de hematocritos en los dos modelos

2.5.6.Modelo de regresión logística

2.6. Conclusiones

3. Recursos

Práctica 2: Limpieza y validación de los datos

Diego Armando Cale Pillco

07-01-2020

1. Detalles de la actividad

1.1. Descripción

En esta actividad se elabora un caso práctico, consistente en el tratamiento de un conjunto de datos (en inglés, dataset), orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

1.2. Objetivos

Los objetivos que se persiguen mediante el desarrollo de esta actividad práctica son los siguientes: - 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 permita 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.

1.3. Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- 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

En esta actividad se usará el fichero de datos de 2353 operaciones, efectuadas en el Hospital Universitario de Santiago. En esta actividad se usará el fichero de datos: datos.csv, que contiene los datos sobre 2353 operaciones efectuadas en el Hospital Universitario de Santiago.

El conjunto de datos contine 2353 registros y 15 variables:

- EDAD (años)
- SEXO
- PATOL (Patología) 1=inflamatoria; 2=neoplasia;3=trauma; 4=otras.
- TIP_OPER (tipo operación): 1=limpia; 2=potencialente contaminada; 3=contaminada; 4=sucia
- ALB (albúmina)
- HB (Hemoglobina)
- HCTO (Hematocrito)
- LEUCOS (Leucocitos)
- LINFOPCT (Linfocitos (%))
- HEMAT (Hematíes)
- GLUC (Glucosa)
- OBES (Obesidad)
- DESNUTR (Desnutrición)
- DIABETES
- INFEC(Infección)
- GLUC_4 (categorización de glucosa)

2.2. Importancia y objetivos de los análisis

Nuestro motivo es estudiar primero la relación entre diferentes variables de la base de datos y posteriormente la identificación de los factores de riesgo asociados a la infección post operatoria.

2.3. Limpieza de los datos

Antes de comenzar con la limpieza de los datos, procedemos a realizar la lectura del fichero en formato CSV en el que se encuentran. El resultado devuelto por la llamada a la función read.csv() será un objeto data.frame:

```
# Lectura de datos
datos <- read.csv("datos.csv", header = TRUE)
head(datos)
```

```
##   EDAD  SEXO  PATOL TIP_OPER ALB HB HCTO LEUCOS LINFOPCT HEMAT GLUC OBES
## 1   59 mujer inflam    sucia   4 13   38  19090       12     4   68   si
## 2   65 mujer  otras pot_cont   4 12   36   6190       33     4   79   si
## 3   69 varón    neo   contam   4 13   38   6360       28     4   85 <NA>
## 4   70 mujer    neo   contam  NA 14   39   8200       23     5   87   no
## 5   79 mujer inflam   contam  NA  8   23  16940        8     3   88 <NA>
## 6   55 mujer  otras pot_cont   4 12   38   4800       51     4   92   si
##   DESNUTR DIABETES INFEC GLUC_4
## 1      no       si    si    <70
## 2      no       si    no 70-115
## 3      no       si    si 70-115
## 4      si       si    no 70-115
## 5      no       si    si 70-115
## 6      no       si    si 70-115
```

```
# Tipo de dato asignado a cada campo
sapply(datos, function(x) class(x))
```

```
##      EDAD      SEXO     PATOL  TIP_OPER       ALB        HB      HCTO
## "integer"  "factor"  "factor"  "factor" "numeric" "integer" "integer"
##    LEUCOS  LINFOPCT     HEMAT      GLUC      OBES   DESNUTR  DIABETES
## "integer" "integer" "integer" "integer"  "factor"  "factor"  "factor"
##     INFEC    GLUC_4
##  "factor"  "factor"
```

Además, observamos cómo los tipos de datos asignados automáticamente por R a las variables se corresponden con el dominio de estas.

Nota. Originalmente, los valores desconocidos eran denotados en el dataset mediante el caracter '?'. Por ello, se ha realizado una sustitución de estos valores por una cadena vacía previa a la lectura para que R marque estos valores desconocidos como NA (del inglés, Not Available). Esto simplificará el manejo de los datos en los apartados posteriores.

2.3.1. Selección de los datos de interés

La gran mayoría de los atributos presentes en el conjunto de datos se corresponden con características que reúnen los datos de las operaciones recogidos en forma de registros, por lo que será conveniente tenerlos en consideración durante la realización de los análisis. Sin embargo, podemos prescindir de los dos primeros campos (LEUCOS y LINFOPCT) dado que no son atributos técnicos e las operaciones y, por tanto, nos resultan menos relevantes a la hora de resolver nuestro problema.

```
# Eliminar las columnas 8 y 9
datos <- datos[, -(8:9)]
```

2.3.2. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Comúnmente, se utilizan los ceros como centinela para indicar la ausencia de ciertos valores. Sin embargo, no es el caso de este conjunto de datos puesto que, como se comentó durante el apartado relativo a la lectura, se utilizó el caracter '?' para denotar un valor desconocido. Así, se procede a conocer a continuación qué campos contienen elementos vacíos:

```
# Números de valores desconocidos por campo
sapply(datos, function(x) sum(is.na(x)))
```

```
##     EDAD     SEXO    PATOL TIP_OPER      ALB       HB     HCTO    HEMAT 
##        2        0        0        0     1046        9        7      103 
##     GLUC     OBES  DESNUTR DIABETES    INFEC   GLUC_4 
##        0      323        7        0        0        0
```

Llegados a este punto debemos decidir cómo manejar estos registros que contienen valores desconocidos para algún campo. Una opción podría ser eliminar esos registros que incluyen este tipo de valores, pero ello supondría desaprovechar información.

Como alternativa, se empleará un método de imputación de valores basado en la similitud o diferencia entre los registros: la imputación basada en k vecinos más próximos (en inglés, kNN-imputation). La elección de esta alternativa se realiza bajo la hipótesis de que nuestros registros guardan cierta relación. No obstante, es mejor trabajar con datos "aproximados" que con los propios elementos vacíos, ya que obtendremos análisis con menor margen de error.

```
# Imputación de valores mediante la función kNN() del paquete VIM
suppressWarnings(suppressMessages(library(VIM)))
datos$EDAD <- kNN(datos)$EDAD
datos$ALB <- kNN(datos)$ALB
datos$HB <- kNN(datos)$HB
datos$HCTO <- kNN(datos)$HCTO
datos$OBES <- kNN(datos)$OBES
datos$DESNUTR <- kNN(datos)$DESNUTR
sapply(datos, function(x) sum(is.na(x)))
```

```
##     EDAD     SEXO    PATOL TIP_OPER      ALB       HB     HCTO    HEMAT
##        0        0        0        0        0        0        0      103
##     GLUC     OBES  DESNUTR DIABETES    INFEC   GLUC_4
##        0        0        0        0        0        0
```

2.3.3. Identificación y tratamiento de valores extremos

Los valores extremos o outliers son aquellos que parecen no ser congruentes sin los comparamos con el resto de los datos. Para identificarlos, podemos hacer uso de dos vías: (1) representar un diagrama de caja por cada variable y ver qué valores distan mucho del rango intercuartílico (la caja) o (2) utilizar la función boxplots.stats() de R, la cual se emplea a continuación. Así, se mostrarán sólo los valores atípicos para aquellas variables que los contienen:

```
boxplot.stats(datos$EDAD)$out
```

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

```
boxplot.stats(datos$ALB)$out
```

```
##  [1] 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 7 1 2 2 2 2 2 2 1 7 2 2 2 7 2 2 2 2 2
## [36] 2 2 2 2 2 2 2 2 1 1 2 2 1
```

```
boxplot.stats(datos$HB)$out
```

```
##  [1]  7  7  4  7  7  6  6  6  7  5  7  7  6  5  6  7  6  7  7  6 20  6  7
## [24]  5  7  6  4
```

```
boxplot.stats(datos$HCTO)$out
```

```
##  [1] 23 23 25 54 11 23 55 23 24 25 23 17 20 25  5  5 22 24 24 21 25  4  6
## [24] 13 25 25 23 22 24 23 14 22 56 25 25 21 24 25 23 21 25  4 59 24 25 19
## [47] 24 25 22 25 55 15 25 59 25 25 25 59 19 24 22
```

```
boxplot.stats(datos$HEMAT)$out
```

```
##  [1] 1 2 8 8 7 2 7 7 7 9 8 2 7 2 7 2 2 7 2 8 2 2 2
```

```
boxplot.stats(datos$GLUC)$out
```

```
##   [1] 165 166 166 169 171 171 172 173 176 178 178 180 183 184 185 185 188
##  [18] 190 193 194 194 195 195 199 200 202 209 211 211 212 213 213 215 217
##  [35] 218 218 220 221 222 225 226 226 226 227 227 227 230 230 230 232 233
##  [52] 234 236 236 237 237 240 241 241 242 245 247 247 248 255 260 263 265
##  [69] 265 270 271 284 291 165 165 165 166 166 166 167 167 167 168 169 170
##  [86] 170 170 171 171 171 171 171 171 172 173 174 175 175 176 176 177 178
## [103] 178 179 180 180 180 182 182 182 182 183 184 188 188 188 189 189 189
## [120] 190 191 191 192 193 193 193 193 193 193 195 195 196 196 198 199 199
## [137] 200 203 204 205 206 206 206 207 208 209 209 211 211 214 218 219 221
## [154] 221 221 221 222 222 224 229 229 230 230 240 242 242 246 248 257 258
## [171] 263 265 291
```

No obstante, si revisamos los anteriores datos para varios datos escogido aleatoriamente de esta data, comprobamos que son valores que perfectamente pueden darse (Hay operaciones cuyos hematocritos son <22 y otros llegan hasta los 59, entre otros). Es por ello que el manejo de estos valores extremos consistirá en simplemente dejarlos como actualmente están recogidos.

2.3.4. Exportación de los datos preprocesados

Una vez que hemos acometido sobre el conjunto de datos inicial los procedimientos de integración, validación y limpieza anteriores, procedemos a guardar estos en un nuevo fichero denominado datos_operaciones.csv:

```
# Exportación de los datos limpios en .csv
write.csv(datos, "datos_operaciones.csv")
```

2.4.Análisis de los datos.

Hacemos lectura de la informaión del dataset antes analizado.

```
# Lectura de datos
datos <- read.csv("datos_operaciones.csv", header = TRUE)
head(datos)
```

```
##   X EDAD  SEXO  PATOL TIP_OPER ALB HB HCTO HEMAT GLUC OBES DESNUTR
## 1 1   59 mujer inflam    sucia   4 13   38     4   68   si      no
## 2 2   65 mujer  otras pot_cont   4 12   36     4   79   si      no
## 3 3   69 varón    neo   contam   4 13   38     4   85   no      no
## 4 4   70 mujer    neo   contam   4 14   39     5   87   no      si
## 5 5   79 mujer inflam   contam   2  8   23     3   88   si      no
## 6 6   55 mujer  otras pot_cont   4 12   38     4   92   si      no
##   DIABETES INFEC GLUC_4
## 1       si    si    <70
## 2       si    no 70-115
## 3       si    si 70-115
## 4       si    no 70-115
## 5       si    si 70-115
## 6       si    si 70-115
```

2.4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

A continuación, se seleccionan los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar. No obstante, como se verá en el apartado consistente en la realización de pruebas estadísticas, no todos se utilizarán.

```
# Agrupación por sexo
datos.sexo.mujer <- datos[datos$SEXO == "mujer",]
datos.sexo.varon <- datos[datos$SEXO == "varón",]
# Agrupación por Tipo de operación
datos.tipo_limpia <- datos[datos$TIP_OPER == "sucia",]
datos.tipo_otencial_contaminada <- datos[datos$TIP_OPERL == "pot_cont",]
datos.tipo_contaminada <- datos[datos$TIP_OPER == "contam",]
# Agrupación por patología
datos.patol_neoplasia <- datos[datos$PATOL == "neo",]
datos.patol_trauma <- datos[datos$PATOL == "traum",]
datos.patol_otros <- datos[datos$PATOL == "otras",]
# Agrupación por Desnutrición
datos.desnutr_si <- datos[datos$DESNUTR == "si",]
datos.desnutr_no <- datos[datos$DESNUTR == "no",]
# Agrupación por Diabetes
datos.diabetes_si <- datos[datos$DIABETES == "si",]
datos.diabetes_no <- datos[datos$DIABETES == "no",]
# Agrupación por Infeccion
datos.infec_si <- datos[datos$INFEC == "si",]
datos.infec_no <- datos[datos$INFEC == "no",]
```

2.4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba de normalidad de AndersonDarling. Así, se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de significación prefijado α = 0, 05. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

```
library(nortest)
alpha = 0.05
col.names = colnames(datos)
for (i in 1:ncol(datos)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(datos[,i]) | is.numeric(datos[,i])) {
    p_val = ad.test(datos[,i])$p.value
    if (p_val < alpha) {
     cat(col.names[i])
      # Format output
     if (i < ncol(datos) - 1) cat(", ")
     if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
## Variables que no siguen una distribución normal:
## X, EDAD, ALB,
## HB, HCTO, HEMAT,
## GLUC,
```

Seguidamente, pasamos a estudiar la homogeneidad de varianzas mediante la aplicación de un test de Fligner-Killeen. En este caso, estudiaremos esta homogeneidad en cuanto a los grupos conformados por los datos de lo cancer cuyo análisis es Benigno y Maligno dependen de la media de la textura. En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.

```
fligner.test(HEMAT ~ INFEC, data = datos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  HEMAT by INFEC
## Fligner-Killeen:med chi-squared = 1.7292, df = 1, p-value = 0.1885
```

Puesto que obtenemos un p-valor superior a 0,05, aceptamos la hipótesis de que las varianzas de ambas muestras son homogéneas.

2.5. Representación de los resultados a partir de tablas y gráficas.

2.5.1. ¿Qué variables cuantitativas influyen más la infección post operatoria?

En primer lugar, procedemos a realizar un análisis de correlación entre las distintas variables para determinar cuáles de ellas ejercen una mayor influencia sobre el la infección post operatoria.

Para ello, se utilizará el coeficiente de correlación de Spearman, puesto que hemos visto que tenemos datos que no siguen una distribución normal.

```r
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
# Calcular el coeficiente de correlación para cada variable cuantitativa
# con respecto al campo "HEMAT"
for (i in 1:(ncol(datos) - 1)) {
  if (is.integer(datos[,i]) | is.numeric(datos[ ,i])) {
    spearman_test = cor.test(datos[ ,i],
                             datos[ ,length(datos)-6],
                             method = "spearman")

    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value
    # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(datos)[i]
 }
}
```

```
## Warning in cor.test.default(datos[, i], datos[, length(datos) - 6], method
## = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos) - 6], method
## = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos) - 6], method
## = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos) - 6], method
## = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos) - 6], method
## = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos) - 6], method
## = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos) - 6], method
## = "spearman"): Cannot compute exact p-value with ties
```

```r
print(corr_matrix)
```

```
##          estimate      p-value
## X      0.05673259 7.108131e-03
## EDAD  -0.17595985 4.186682e-17
## ALB    0.41714401 1.927413e-95
## HB     0.70739485 0.000000e+00
## HCTO   0.74314977 0.000000e+00
## HEMAT  1.00000000 0.000000e+00
## GLUC   0.02996654 1.553264e-01
```

Así, identificamos cuáles son las variables más correlacionadas con la infección en función de su proximidad con los valores -1 y +1. Teniendo esto en cuenta, queda patente cómo la variable más relevante en la fijación del HEMAT, HB y HCTO.

Nota. Para cada coeficiente de correlación se muestra también su p-valor asociado, puesto que éste puede dar información acerca del peso estadístico de la correlación obtenida.

2.5.2. Modelo de regresión lineal simple

a. Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la variable hematocrito en función de la hemoglobina. Evaluar la bondad de ajuste a través del coeficiente de determinación (R2). Podéis usar la instrucción de R lm.

```
Model.1.1<- lm(HCTO~HB, data=datos )
summary(Model.1.1)
```

```
##
## Call:
## lm(formula = HCTO ~ HB, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.986  -0.986   0.044   1.074  27.677
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.26360    0.33376   18.77   <2e-16 ***
## HB           2.51481    0.02508  100.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.583 on 2351 degrees of freedom
## Multiple R-squared:  0.8105, Adjusted R-squared:  0.8104
## F-statistic: 1.006e+04 on 1 and 2351 DF,  p-value: < 2.2e-16
```

A la vista de los resultados, existe una relación lineal positiva muy fuerte, entre ambas variables. Se observa que el coeficiente de determinación ajustado es: 0.8142. Es decir, el modelo explica el 81.42 % de la variabilidad de la variable hematocrito. Si se calcula el coeficiente de correlación obtenemos un valor de 0.9.

NOTA: Al tener un modelo con una sola variable, se podría tomar el coeficiente de determinación sin ajustar, ya que su valor no se altera.

b. Algunos estudios afirman que la relación calculada anteriormente varía según la persona esté en condiciones óptimas de salud o no. Para contestar a esta pregunta, se dividirá la muestra en dos, según si la persona presenta desnutrición o no. Posteriormente se repetirá el estudio para cada muestra por separado. A partir de los resultados del modelo lineal en cada una de las muestras.

```
#Estimacion del modelo
selected_DESNUTR <- which(datos$DESNUTR=="si" )
data1=datos[selected_DESNUTR,]
selected_DESNUTR_0 <- which(datos$DESNUTR=="no" )
data0=datos[selected_DESNUTR_0,]
dim(data1)
```

```
## [1] 97 15
```

```
dim(data0)
```

```
## [1] 2256   15
```

```
Model.1.1.1<- lm(HCTO~HB, data=data1 )
summary(Model.1.1.1)
```

```
##
## Call:
## lm(formula = HCTO ~ HB, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3577 -0.9899 -0.1330  0.7854  5.3162
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.88848    0.64644   6.015 3.34e-08 ***
## HB           2.69384    0.05618  47.950  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.493 on 95 degrees of freedom
## Multiple R-squared:  0.9603, Adjusted R-squared:  0.9599
## F-statistic:  2299 on 1 and 95 DF,  p-value: < 2.2e-16
```

```
Model.1.1.2<- lm(HCTO~HB, data=data0 )
summary(Model.1.1.2)
```

```
##
## Call:
## lm(formula = HCTO ~ HB, data = data0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.964  -0.977   0.023   1.036  27.467
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.55760    0.35892   18.27   <2e-16 ***
## HB           2.49378    0.02682   92.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.617 on 2254 degrees of freedom
## Multiple R-squared:  0.7932, Adjusted R-squared:  0.7931
## F-statistic:  8643 on 1 and 2254 DF,  p-value: < 2.2e-16
```

A la vista de los resultados, podemos concluir que la presencia o no de desnutrición varía la relación lineal entre ambas variables. Los coeficientes de determinación son de 0.96 y 0.79. Se puede observar que la relación lineal en presencia de desnutrición es casi percfecta, por lo que el ajuste mejora.

2.5.3. Modelo de regresión lineal múltiple (regresores cuantitativos)

¿Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la variable hematocrito en función de la hemoglobina y la edad, mejora el modelo?.

Evaluar la bondad del ajuste y comparar el resultado con el obtenido en el apartado 2.5.2.a). Para ello se usa la instrucción de R lm y el coeficiente R-cuadrado ajustado en la comparación. Interpretar también el significado de los coeficientes obtenidos y su significación estadística.

```
#Estimacion del modelo
attach(datos)
Model.1.2<- lm(HCTO ~HB + EDAD, data= datos)
summary( Model.1.2)
```

```
##
## Call:
## lm(formula = HCTO ~ HB + EDAD, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.177  -1.063   0.004   1.207  27.803
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.440679   0.403689  13.477  < 2e-16 ***
## HB          2.536502   0.025729  98.586  < 2e-16 ***
## EDAD        0.009845   0.002732   3.604  0.00032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.577 on 2350 degrees of freedom
## Multiple R-squared:  0.8115, Adjusted R-squared:  0.8114
## F-statistic:  5060 on 2 and 2350 DF,  p-value: < 2.2e-16
```

En el modelo 2.5.2.a), el coeficiente de bondad del ajuste es de 0.8142, y en este último de 0.8152, por lo que añadir la variable explicativa edad no mejora mucho el modelo anterior.

2.5.4. Modelo de regresión lineal múltiple (regresores cuantitativos y cualitativos)

a. Queremos conocer en qué medida se relacionan los hematocritos, con la hemoglobina y la edad, dependiendo de si los pacientes tienen o no infección postquirúrgica. Aplicar un modelo de regresión lineal múltiple y explicar el resultado.

```
#Estimacion del modelo
selected_INFEC <- which( datos$INFEC=="si" )
data_INFEC_1=datos[selected_INFEC,]
selected_INFEC_NO <- which( datos$INFEC=="no" )
data_INFEC_0=datos[selected_INFEC_NO,]
dim(data_INFEC_1)
```

```
## [1] 464  15
```

```
dim(data_INFEC_0)
```

```
## [1] 1889   15
```

```
Model.1.3.a.1 = lm(formula=HCTO~HB+EDAD,data=data_INFEC_1)
summary(Model.1.3.a.1)
```

```
##
## Call:
## lm(formula = HCTO ~ HB + EDAD, data = data_INFEC_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.789  -0.979   0.086   1.311   5.983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.72759    0.91805   4.060 5.76e-05 ***
## HB           2.62261    0.05581  46.989  < 2e-16 ***
## EDAD         0.01921    0.00766   2.508   0.0125 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.001 on 461 degrees of freedom
## Multiple R-squared:  0.8296, Adjusted R-squared:  0.8289
## F-statistic:  1122 on 2 and 461 DF,  p-value: < 2.2e-16
```

```
Model.1.3.a.2 = lm(formula=HCTO~HB+EDAD,data=data_INFEC_0)
summary(Model.1.3.a.2)
```

```
##
## Call:
## lm(formula = HCTO ~ HB + EDAD, data = data_INFEC_0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.094  -1.074  -0.008   1.129  27.453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.068516   0.455066  13.335  < 2e-16 ***
## HB          2.498613   0.029405  84.973  < 2e-16 ***
## EDAD        0.007806   0.002901   2.691  0.00719 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 1886 degrees of freedom
## Multiple R-squared:  0.7993, Adjusted R-squared:  0.7991
## F-statistic:  3755 on 2 and 1886 DF,  p-value: < 2.2e-16
```

Con el estudio propuesto hemos obtenido los siguientes coeficientes de determinación ajustados:

Pacientes con infección postquirúrgica R2=0.8289 Pacientes sin infección postquirúrgica R2=0.8045

Comparando los coeficientes de determinación ajustado, por HB y EDAD, no se aprecian diferencias significativas entre los pacientes que presentan infección postquirúrgica y los que no la presentan.

b. Se hará el mismo estudio, pero tomando sólo aquéllos pacientes, cuya cantidad de hematocritos sea < 37. Comparar con el modelo anterior y extraer conclusiones.

```
#Con infección
selected_INFEC_HCTO <- which( datos$INFEC=="si"& datos$HCTO<37 )
data_INFEC_1_hcto37=datos[selected_INFEC_HCTO,]
Model.1.3.b.1= lm(formula=HCTO~HB+EDAD,data=data_INFEC_1_hcto37)
summary(Model.1.3.b.1)
```

```
##
## Call:
## lm(formula = HCTO ~ HB + EDAD, data = data_INFEC_1_hcto37)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.051  -0.986   0.438   1.650   6.675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.65403    2.14887   4.958 1.56e-06 ***
## HB           1.82077    0.17518  10.394  < 2e-16 ***
## EDAD         0.02724    0.01573   1.732   0.0849 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.89 on 192 degrees of freedom
## Multiple R-squared:  0.3609, Adjusted R-squared:  0.3543
## F-statistic: 54.21 on 2 and 192 DF,  p-value: < 2.2e-16
```

```
#Sin infección
selected_INFECNO_HCTO <- which( datos$INFEC=="no"& datos$HCTO<37 )
data_INFEC_0_hcto37=datos[selected_INFECNO_HCTO,]
Model.1.3.b.2 = lm(formula=HCTO~HB+EDAD,data=data_INFEC_0_hcto37)
summary(Model.1.3.b.2)
```

```
##
## Call:
## lm(formula = HCTO ~ HB + EDAD, data = data_INFEC_0_hcto37)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.647  -0.721   0.307   1.313   5.255
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.841164   1.288042   8.417 6.12e-16 ***
## HB           1.982182   0.100769  19.670  < 2e-16 ***
## EDAD         0.001043   0.008526   0.122    0.903
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.375 on 421 degrees of freedom
## Multiple R-squared:  0.4872, Adjusted R-squared:  0.4847
## F-statistic:   200 on 2 and 421 DF,  p-value: < 2.2e-16
```

En este caso se obtienen los siguientes coeficientes de determinación ajustados: Pacientes cuya cantidad de hematocritos es < 37 y presentan infección postquirúrgica R2=0.3508 Pacientes cuya cantidad de hematocritos es < 37 y no presentan infección postquirúrgica R2=0.5069 Por lo tanto, en este tipo de pacientes, se puede apreciar que el ajuste mejora en los pacientes que no presentan infección postquirúrgica. Esto podría ser consecuente con el apartado b, del modelo 1, dónde algunos estudios demuestran que la relación entre HCTO y HB, cambia según la salud del paciente. Por otro lado, al seleccionar aquellos pacientes con nivel de HCTO < 37, se observa con más claridad la falta de significación de la variable EDAD (p_value= 0.08 y 0.65 respectivamente).

2.5.5.Efectuar una predicción de la concentración de hematocritos en los dos modelos

Para su evaluación se va a suponer un paciente de 60 años, con infección postquirúrgica y con un valor de hemoglobina de 10. Para ello se realizará la predicción del valor de hematocritos, con los dos modelos del apartado 2.5.4. Interpretar los resultados. Se define un dataframe que contenga los valores de EDAD y HB para los cuales queremeos predecir el valor de HCTO. Se aplicarán los modelos obtenidos en el partado 2.5.4, con pacientes con infección postquirúrgica.

```
newdata = data.frame(HB = 10, EDAD=60)
predict(Model.1.3.a.1, newdata)
```

```
##        1
## 31.10641
```

```
predict(Model.1.3.b.1, newdata)
```

```
##      1
## 30.496
```

Obtenemos un valor de 31.1078 de HCTO al aplicar el primer modelo obtenido en el apartado 2.5.4.a y un valor de 30.49042 aplicando el primer modelo obtenido en el apartado 2.5.4.b.

2.5.6.Modelo de regresión logística

2.5.6.1. Análisis crudo. Estimación de OR (Odds Ratio)

Se desea identificar cuáles son los factores de riesgo en la infección postquirúrgica. Por tanto, se evaluará la probabilidad de que un paciente pueda o no tener una infección, dependiendo si presenta o no unas determinadas características. Para evaluar esta probabilidad, primero se realizará un análisis crudo de los posibles factores.

2.5.6.2. OR

Estudiar la relación entre la infección postquirúrgica y cada una de las variables siguientes: diabetes, desnutrición, obesidad, edad y hematocrito. Estimar e interpretar las OR en cada caso. Dicha estimación será efectuada a partir de las tablas de contingencia. Antes de calcular los valores de las odds ratio, se recomienda aplicar el test chi-cuadrado. Se recodifica la variable INFEC para realizar los cálculos.

```
INFECRE <- factor(datos$INFEC, labels=c("1", "0"))
table (INFECRE)
```

```
## INFECRE
##    1    0
## 1889  464
```

Para comprobar si existe asociación entre el factor de exposición y tener o no infección, se aplicará el test Chi-cuadrado de Pearson. Un resultado significativo nos dirá que existe asociación, pero para conocer el grado de dica asociación, se calcularán las OR: Infección postquirúrgica y Diabetes

```
DIABETESRE <- factor(datos$DIABETES, labels=c("1", "0"))
table (DIABETESRE)
```

```
## DIABETESRE
##    1    0
## 2211  142
```

```
DIABETES.tab = table(INFECRE,DIABETESRE)
DIABETES.tab
```

```
##        DIABETESRE
## INFECRE    1    0
##       1 1792   97
##       0  419   45
```

```
chi.test<-chisq.test(DIABETES.tab)
print(chi.test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  DIABETES.tab
## X-squared = 12.886, df = 1, p-value = 0.0003311
```

```
#Mediante test de Fisher:
fisher.test(DIABETES.tab,simulate.p.value = TRUE)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  DIABETES.tab
## p-value = 0.0004511
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.338706 2.902822
## sample estimates:
## odds ratio
##   1.983486
```

Cálculo manual OR:

```
OddsRatio <- function( x, y ){
n00 <- sum( (x==0) & (y==0) )
n11 <- sum( (x==1) & (y==1) )
n01 <- sum( (x==0) & (y==1) )
n10 <- sum( (x==1) & (y==0) )
OR <- (n00 * n11) / (n01*n10)
return (OR)
}
OddsRatio( INFECRE, DIABETESRE )
```

```
## [1] 1.984106
```

Infección postquirúrgica y Desnutrición

```
DESNUTR.tab = table(INFECRE, datos$DESNUTR)
chi.test<-chisq.test(DESNUTR.tab)
print(chi.test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  DESNUTR.tab
## X-squared = 37.102, df = 1, p-value = 1.121e-09
```

```
fisher.test(DESNUTR.tab,simulate.p.value = TRUE)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  DESNUTR.tab
## p-value = 2.383e-08
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  2.235212 5.355848
## sample estimates:
## odds ratio
##   3.468362
```

Infección postquirúrgica y Obesidad

```
OBES.tab = table(INFECRE, datos$OBES)
chi.test<-chisq.test(OBES.tab)
print(chi.test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  OBES.tab
## X-squared = 18.704, df = 1, p-value = 1.527e-05
```

```
fisher.test(OBES.tab,simulate.p.value = TRUE)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  OBES.tab
## p-value = 2.697e-05
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.347510 2.270956
## sample estimates:
## odds ratio
##   1.753242
```

En todos los casos se obtiene un p-valor inferior a 0.05, por lo que podemos concluir que existe relación entre la variable Infección postquirúrgica y los factores estudiados. Con referencia a las OR: - Infección vs Diabetes: OR de 1.98, por lo que nos indica que una persona con diabetes, tiene una probabilidad 1.98 veces mayor de contraer una infección postquirúrgica. Para ser más precisos, las odds de infección de un paciente con diabetes son 1.98 veces las odds de un paciente que no sufre diabetes. - Infección vs Desnutrición: OR de 3.49, por lo que nos indica que una persona con desnutrición, tiene una probabilidad

3.49 veces mayor de contraer una infección postquirúrgica. - Infección vs Obesidad: OR de 1.81, por lo que nos indica que una persona con obesidad, tiene una probabilidad 1.81 veces mayor de contraer una infección postquirúrgica.

2.5.6.3. Relación entre infección con edad y hematocrito

Edad y hematocrito son variables continuas: ¿podríamos seguir el procedimiento anterior para el cálculo de la OR? No podemos seguir el procedimiento anterior para el cálculo de la OR puesto que EDAD y HCTO son variables continuas. En este caso para calcular las OR, deberíamos construir un modelo de regresión logística.

2.5.6.4. Relación entre infección y tipo de operación

Si queremos ver la relación entre INFEC (Infección) y TIP-OPER (tipo de operación), ¿podríamos seguir el procedimiento anterior, para el cáculo de la OR? En el caso que la respuesta fuese negativa, ¿cuál sería una solución? No podemos seguir el procedimiento anterior para el cálculo de la OR puesto que TIP-OPER es una variable categórica de 4 categorias. En este caso para calcular las OR, deberíamos construir un modelo de regresión logística.

2.5.6.5. Modelo de regresión logística

2.5.6.5.1. Modelo INFEC en relación con Diabetes

Estimar el modelo de regresión logística donde la variable dependiente es “INFEC” y la explicativa es tener diabetes o no. ¿Podemos considerar que el hecho de tener diabetes es un factor de riesgo de infección? Justifica tu respuesta. Tiene relación con lo obtenido en el apartado anterior?

```
logit_model_1 <- glm(formula=INFECRE~DIABETESRE, data=datos, family=binomial)
summary(logit_model_1)
```

```
##
## Call:
## glm(formula = INFECRE ~ DIABETESRE, family = binomial, data = datos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8731  -0.6482  -0.6482  -0.6482   1.8239
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.45322    0.05426 -26.780  < 2e-16 ***
## DIABETESRE0  0.68517    0.18835   3.638 0.000275 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2336.5  on 2352  degrees of freedom
## Residual deviance: 2324.3  on 2351  degrees of freedom
## AIC: 2328.3
##
## Number of Fisher Scoring iterations: 4
```

```
OR=exp(logit_model_1$coefficients[2])
OR
```

```
## DIABETESRE0
##    1.984106
```

Como podemos observar, el cálculo de la OR coincide con el apartado anterior, ya que en el modelo hemos introducido sólo la misma variable explicativa. La interpretación de los resultados es la misma.

2.5.6.5.2 Añadir edad y hematocrito

Añadimos al modelo anterior las variables explicativas edad y hematocrito. Evaluar si alguno de los regresores tiene influencia significativa (p-valor del contraste individual inferior a 0.05.

```
logit_model_2 <- glm(formula=INFECRE~DIABETESRE+EDAD+HCTO, data=datos, family
=binomial)
summary(logit_model_2)
```

```
##
## Call:
## glm(formula = INFECRE ~ DIABETESRE + EDAD + HCTO, family = binomial,
##     data = datos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3451  -0.6970  -0.5802  -0.4434   2.3293
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.810940   0.395866  -2.049   0.0405 *
## DIABETESRE0  0.381171   0.193584   1.969   0.0490 *
## EDAD         0.018702   0.002913   6.421 1.36e-10 ***
## HCTO        -0.043647   0.008533  -5.115 3.14e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2336.5  on 2352  degrees of freedom
## Residual deviance: 2240.8  on 2349  degrees of freedom
## AIC: 2248.8
##
## Number of Fisher Scoring iterations: 4
```

Se puede observar en los resultados que los tres factores tienen una influencia significativa. La probabilidad de infección postquirúrgica aumenta con la diabetes y edad, y disminuye con la cantidad de hematocritos. A continuación se calculan los odds asociados a estas variables explicativas:

```
exp( logit_model_2$coefficients[2:4])
```

```
## DIABETESRE0        EDAD        HCTO
##   1.4639978   1.0188783   0.9572919
```

La interpretación de los OR son: - Las odds de padecer infección es 1.47 veces en pacientes con diabetes en relación a pacientes sin diabetes. - Aumentar en una unidad la edad implica que las odds de padecer infección se multiplican por 1.019, es decir aumenta en un 1.9 - Aumentar en una unidad los hematocritos reduce las odds de padecer infección en 0.95.

2.5.6.5.3 Mejora del modelo

2.5.6.5.3.1 Categorizando variables

Entrenamos el mismo modelo anterior, pero categorizando ambas variables continuas: Edad: (edad>=65 y edad<65) y Hematocrito: (HCTO <37 y HCTO>=37). Explicar los resultados. ¿De qué forma influye la edad y los niveles de hematocritos en este modelo? Explicar como se interpretan los resultados del modelo. Se categorizan las variables:

```
EDADRE<-as.factor(ifelse(datos$EDAD<65, "0", "1"))
HCTORE<-as.factor(ifelse(datos$HCTO<37, "0", "1"))
logit_model_3<- glm(formula=INFECRE~DIABETESRE+EDADRE+HCTORE, data=datos, fam
ily=binomial)
summary(logit_model_3)
```

```
##
## Call:
## glm(formula = INFECRE ~ DIABETESRE + EDADRE + HCTORE, family = binomial,
##     data = datos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1368  -0.6764  -0.5167  -0.5167   2.0395
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1465     0.1084 -10.577  < 2e-16 ***
## DIABETESRE0   0.4623     0.1943   2.379   0.0173 *
## EDADRE1       0.5878     0.1086   5.415 6.13e-08 ***
## HCTORE1      -0.7999     0.1115  -7.175 7.25e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2336.5  on 2352  degrees of freedom
## Residual deviance: 2229.3  on 2349  degrees of freedom
## AIC: 2237.3
##
## Number of Fisher Scoring iterations: 4
```

A la vista de los resultados se puede concluir que un paciente de edad > 65 y con diabetes, tiene más probabilidad de infección postquirúrgica. Por otro lado, si el número de hematocritos es alto las probabilidades de infección postquirúrgica disminuyen. Se observa que el AIC es más pequeño que en el modelo anterior, por lo que existe una mejora en el ajuste.

2.5.6.5.3.2 Añadir desnutrición

Posteriormente se añadirá al modelo las variable explicativa desnutrición. ¿Se observa una mejora del modelo? Explicar

```
logit_model_4 <- glm(formula=INFECRE~DIABETESRE+EDADRE+HCTORE+DESNUTR, data=d
atos, family=binomial)
summary(logit_model_4)
```

```
##
## Call:
## glm(formula = INFECRE ~ DIABETESRE + EDADRE + HCTORE + DESNUTR,
##     family = binomial, data = datos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4253  -0.6670  -0.5147  -0.5147   2.0430
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.2261     0.1116 -10.986  < 2e-16 ***
## DIABETESRE0   0.4363     0.1950   2.238  0.02523 *
## EDADRE1       0.5644     0.1092   5.171 2.33e-07 ***
## HCTORE1      -0.7283     0.1140  -6.391 1.65e-10 ***
## DESNUTRsi     0.7914     0.2204   3.591  0.00033 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2336.5  on 2352  degrees of freedom
## Residual deviance: 2216.9  on 2348  degrees of freedom
## AIC: 2226.9
##
## Number of Fisher Scoring iterations: 4
```

Se observa que el AIC es más pequeño que en el modelo anterior por lo que existe una mejora en el ajuste.

2.5.6.5.3.3 Predicción

Según el modelo del apartado anterior, ¿cuál será la probabilidad de infección postquirúrgica de un paciente de 50 años, con diabetes, concentración de hematocritos de 34, y que no presente desnutrición?

```
pred<-predict(logit_model_4, data.frame(EDADRE="0",DIABETESRE="1",HCTORE="0",
DESNUTR="no"),
type = "response")
pred
```

```
##         1
## 0.2268714
```

El modelo del apartado anterior nos predice una probablidad de infección de 0.3112035 para un paciente de 50 años, con diabetes, concentración de hematocritos de 34 y que no presenta desnutrición. Es una probabilidad baja.

2.6. Conclusiones

- General

Como se ha visto, se han realizado tres tipos de pruebas estadísticas sobre un conjunto de datos que se correspondía con diferentes variables relativas a vehículos con motivo de cumplir en la medida de lo posible con el objetivo que se planteaba al comienzo. Para cada una de ellas, hemos podido ver cuáles son los resultados que arrojan (entre otros, mediante tablas) y qué conocimientos pueden extraerse a partir de ellas. Así, el análisis de correlación y el contraste de hipótesis nos ha permitido conocer cuáles de estas variables ejercen una mayor influencia sobre el cáncer de mama, mientras que el modelo de regresión lineal obtenido resulta de utilidad a la hora de realizar predicciones para esta variable dadas unas características concretas. Previamente, se han sometido los datos a un preprocesamiento para manejar los casos de ceros o elementos vacíos y valores extremos (outliers). Para el caso del primero, se ha hecho uso de un método de imputación de valores de tal forma que no tengamos que eliminar registros del conjunto de datos inicial y que la ausencia de valores no implique llegar a resultados poco certeros en los análisis. Para el caso del segundo, el cual constituye un punto delicado a tratar, se ha optado por incluir los valores extremos en los análisis dado que parecen no resultar del todo atípicos si los comparamos con los valores que toman las correspondientes variables para el cancer de mama.

- Específico

Se puede considerar factores de riesgo en la infección postquirúrgica: La edad>=65 El nivel de hematocritos< 37 La desnutrición La diabetes Obesidad Hemos visto que la probabilidad de infección postquirúrgica aumenta en los pacientes que presentan desnutrición, diabetes y obesidad. También aumenta la probabilidad de infección, si el paciente es mayor de 65 años. Por otro lado, si el número de hematocritos es alto las probabilidades de infección postquirúrgica disminuyen.

3. Recursos

- Gavin Brown. Diversidad en conjuntos de redes neuronales . La universidad de Birmingham. 2004.
- Hussein A. Abbass. Un enfoque evolutivo de redes neuronales artificiales para el diagnóstico de cáncer de mama . Inteligencia artificial en medicina, 25. 2002.
- Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis 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. Morgan Kaufmann.
- 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 (https://guides.github.com/activities/hello-world).