Limpieza y análisis de datos

Tipología y ciclo de vida de los datos

FORMAR
TRANS-
FORMAR

Universitat Oberta de Catalunya

25 años

Práctica 2

Presentación

En esta práctica se elabora un caso práctico 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. Para hacer esta práctica tendréis que trabajar en grupos de 2 personas. Tendréis que entregar un solo archivo con el enlace Github (https://github.com) donde se encuentren las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos que corresponden a vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Aunque no se trata del mismo enunciado, los siguientes ejemplos de ediciones anteriores os pueden servir como guía:

- Ejemplo: https://github.com/Bengis/nba-gap-cleaning
- Ejemplo complejo (archivo adjunto).

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.

Objetivos

Los objetivos concretos de esta práctica son:

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

Descripción de la Práctica a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle (https://www.kaggle.com). Algunos ejemplos de dataset con los que podéis trabajar son:

- Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)
- Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic) El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición.

El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y justificar) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?
2. Integración y selección de los datos de interés a analizar.
3. Limpieza de los datos.

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

3.2. Identificación y tratamiento de valores extremos.

4. Análisis de los datos.

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

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

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

5. Representación de los resultados a partir de tablas y gráficas.
6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?
7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Resolución

Descripción del dataset

Se va a utilizar el dataset sugerido con datos del hundimiento del Titanic. Estos datos constan de 12 variables con 891 registros. Los campos de este conjunto de datos son los siguientes:

- **PassengerId:** Identificador individual de los pasajeros del barco
- **Survived:** Especifica si el pasajero sobrevivió o no (0 -> no, 1-> sí)
- **Pclass:** Clase del pasaje, puede ser 1ª, 2ª y 3ª clase
- **Name:** Nombre del pasajero
- **Sex:** Sexo del pasajero
- **Age:** Edad del pasajero
- **SibSp:** Número de hermanos/as o maridos/ esposas dentro del Titanic
- **Parch:** Número de padres/ hijos dentro del Titanic
- **Ticket:** Número de ticket
- **Fare:** Tarifa del pasajero
- **Cabin:** Número de cabina
- **Embarked:** Puerto en el que embarco, siendo C -> Cherbourg Q-> Queenstown y S-> Southampton

A partir de estos datos se plantea la problemática de concluir cual son los factores fueron los que más afectaron a la supervivencia en el barco. Asimismo, se ha creado modelos de regresión que permiten predecir con ciertas características si esa persona ha sobrevivido o no y se han hecho contrates de hipótesis para saber cuál son las características que ayuden a decidir esto.

Integración y selección de los datos de interés a analizar.

Primero vamos a leer el fichero CSV con los datos, y veremos un resumen general de estos:

```
# Cargamos los paquetes R que vamos a usar
library(ggplot2)
library(dplyr)

# Cargamos el dataset
titanic <- read.csv('train.csv', header = TRUE)
head(titanic)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                             Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                              Heikkinen, Miss. Laina female  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                            Allen, Mr. William Henry   male  35     0     0
## 6                                    Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```
# Resumen de los datos
summary(titanic)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

Una vez cargados los datos y visto su estructura general, como se puede observar los valores en factor los ha tomado como número, así que vamos a asignarle a factor:

```
# Asignación factor
titanic$Survived <- as.factor(titanic$Survived)
titanic$Pclass <- as.factor(titanic$Pclass)
titanic$Sex <- as.factor(titanic$Sex)
titanic$Embarked <- as.factor(titanic$Embarked)
# Resumen de los datos
summary(titanic)
```

```
##   PassengerId    Survived Pclass      Name               Sex
##  Min.   :  1.0   0:549    1:216   Length:891         female:314
##  1st Qu.:223.5   1:342    2:184   Class :character   male  :577
##  Median :446.0            3:491   Mode  :character
##  Mean   :446.0
##  3rd Qu.:668.5
##  Max.   :891.0
##
##       Age            SibSp           Parch           Ticket
##  Min.   : 0.42   Min.   :0.000   Min.   :0.0000   Length:891
##  1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000   Class :character
##  Median :28.00   Median :0.000   Median :0.0000   Mode  :character
##  Mean   :29.70   Mean   :0.523   Mean   :0.3816
##  3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##  Max.   :80.00   Max.   :8.000   Max.   :6.0000
##  NA's   :177
##       Fare           Cabin           Embarked
##  Min.   :  0.00   Length:891          :  2
##  1st Qu.:  7.91   Class :character   C:168
##  Median : 14.45   Mode  :character   Q: 77
##  Mean   : 32.20                      S:644
##  3rd Qu.: 31.00
##  Max.   :512.33
##
```

Ahora vamos a seleccionar los datos que más nos aportan al estudio, hay muchos datos que nos sirven para poder categorizar mejor si la persona sobrevivió o no y por tanto se quedaran en el análisis y otros que no, estos son datos que no nos aportan ninguna información relevante y por tanto van a ser eliminados, los datos que he considerado que no aportaban nada a este análisis son los siguientes:

- PassengerId
- Name
- Ticket
- Fare
- Cabin

```
# Selección variables
titanic <- subset(titanic, select=-c(PassengerId,Name,Ticket,Fare,Cabin))
summary(titanic)
```

```
##  Survived Pclass      Sex           Age            SibSp           Parch
##  0:549    1:216   female:314   Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  1:342    2:184   male  :577   1st Qu.:22.00   1st Qu.:0.000   1st Qu.:0.0000
##           3:491                Median :29.70   Median :0.000   Median :0.0000
##                                Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                                3rd Qu.:35.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                                Max.   :80.00   Max.   :8.000   Max.   :6.0000
##  Embarked
##   :  0
##  C:168
##  Q: 77
##  S:646
##
##
```

Limpieza de los datos

Observamos que hay dos embarcados en puertos desconocidos, los asignamos al puerto con más embarcados, ya que es el puerto que más probabilidad tuvieron de embarcar.

```
#embarcado
titanic$Embarked[titanic$Embarked==""]<-"S"
```

A continuación, comprobamos si tenemos valores de tipo NA, para ponerle solución:

```
# Número de NA
colSums(is.na(titanic))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch Embarked
##        0        0        0      177        0        0        0
```

Como se puede observar solo faltan valores en la variable de edad. Vamos a asignarle valores medios. Y como se puede observar estos valores desaparecen:

```
# Número de NA
colSums(is.na(titanic))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
```

```
# Asignación valores media
titanic$Age[is.na(titanic$Age)] <- mean(titanic$Age,na.rm=T)
colSums(is.na(titanic))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
```

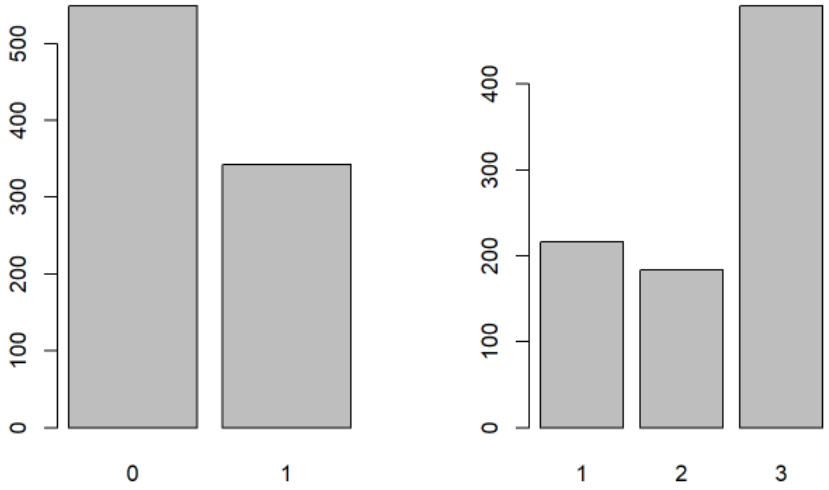
Ahora vamos a discretizar la edad para que sea más fácil de analizar, como hemos visto la edad mínima son 0,42 años hasta la edad de 80, así que lo dividimos hasta esas edades en periodos iguales:

```
# Discretizamos
titanic["edad_discretizada"] <- cut(titanic$Age, breaks = c(0,10,20,30,40,50,60,70,80,100), labels = c("0-9", "10
-19", "20-29", "30-39","40-49","50-59","60-69","70-79","80-89"))
# Datos discretizados
head(titanic)
```

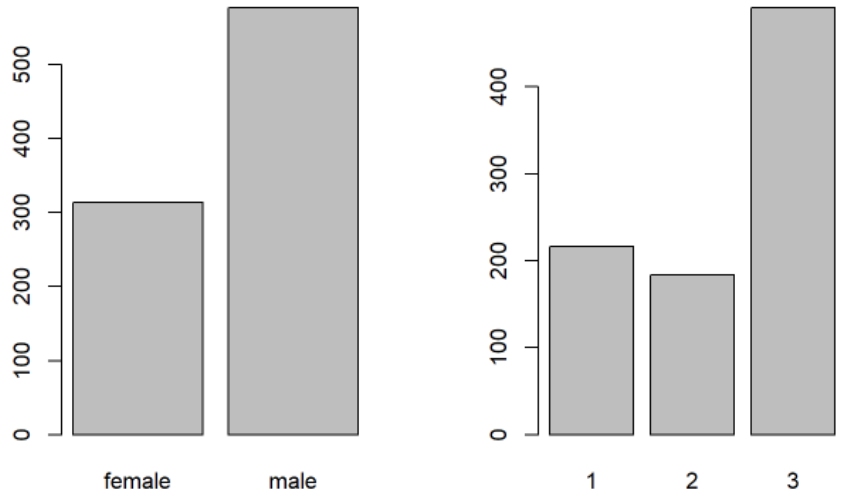
```
##   Survived Pclass    Sex      Age SibSp Parch Embarked edad_discretizada
## 1        0      3   male 22.00000     1     0        S             20-29
## 2        1      1 female 38.00000     1     0        C             30-39
## 3        1      3 female 26.00000     0     0        S             20-29
## 4        1      1 female 35.00000     1     0        S             30-39
## 5        0      3   male 35.00000     0     0        S             30-39
## 6        0      3   male 29.69912     0     0        Q             20-29
```

A continuación, vamos a tratar los datos extremos o también llamados outliers, son aquellos datos de un conjunto de datos que son muy diferentes al resto de valores. Por lo que se van a intentar minimizarlos ya que pueden producir que el resultado de los análisis no sea tan satisfactorio. Vamos a graficar todas las variables para encontrarles:

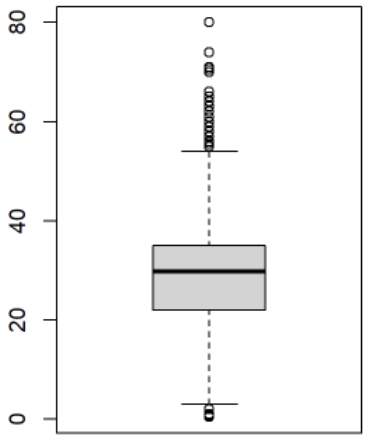
```
# Datos extremos
par(mfrow=c(1,2))
plot(titanic$Survived)
plot(titanic$Pclass)
```

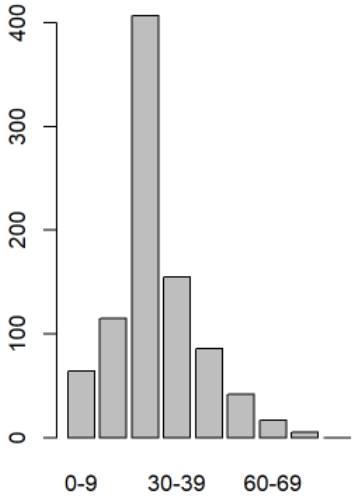



```
plot(titanic$Sex)
plot(titanic$Pclass)
```

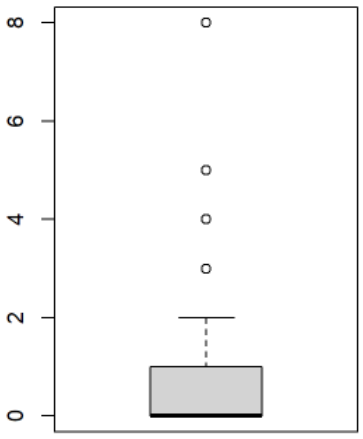


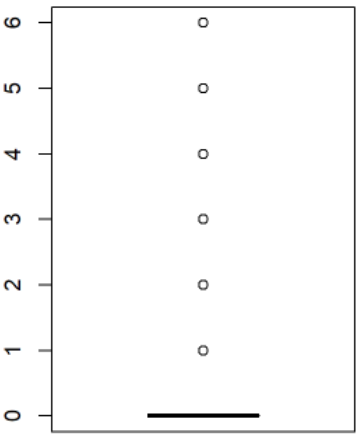
```
boxplot(titanic$Age)
plot(titanic$edad_discretizada)
```



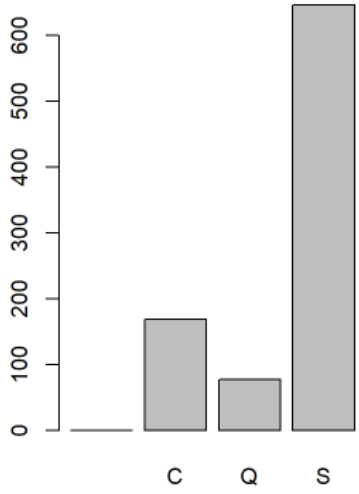



```
boxplot(titanic$SibSp)
boxplot(titanic$Parch)
```





```
plot(titanic$Embarked)
```



Como podemos observar, aunque hay datos que se salen algo de la norma del resto de datos, no se pueden considerar valores externos porque cabe la posibilidad de que existieran de verdad. Estos son datos como la edad ser 80 años, que aunque haya pocos o un dato, es completamente posible, esto sería diferente si por ejemplo se encontrará a alguien con una edad de 270 años, en ese caso sí que tendríamos que tratar este dato.

Una vez que hemos limpiado los datos los vamos a exportar en un fichero llamado, titanic_clean.csv

```
# Exportación de los datos limpios en .csv
write.csv(titanic, "titanic_clean.csv", row.names = FALSE)
```

Análisis de los datos

Selección de los grupos de datos que se quieren analizar/comparar

A continuación, se seleccionarán los grupos dentro del conjunto de datos que se puede estudiar su valor a la hora de saber si una persona sobrevivió o no al accidente del Titanic. Se agruparán por la clase de su billete, sexo, edad y lugar de embarcación

```
# Agrupación por clase
titanic.primera <- titanic[titanic$Pclass.type == 1]
titanic.segunda <- titanic[titanic$Pclass.type == 2]
titanic.tercera <- titanic[titanic$Pclass.type == 3]
# Agrupación por sexo
titanic.mujer <- titanic[titanic$Sex.type == "female"]
titanic.hombre <- titanic[titanic$Sex.type == "male"]
# Agrupación por edad
titanic.hasta10 <- titanic[titanic$edad_discretizada.type == "0-9"]
titanic.hasta20 <- titanic[titanic$edad_discretizada.type == "10-19"]
titanic.hasta30 <- titanic[titanic$edad_discretizada.type == "20-29"]
titanic.hasta40 <- titanic[titanic$edad_discretizada.type == "30-39"]
titanic.hasta50 <- titanic[titanic$edad_discretizada.type == "40-49"]
titanic.hasta60 <- titanic[titanic$edad_discretizada.type == "50-59"]
titanic.hasta70 <- titanic[titanic$edad_discretizada.type == "60-69"]
titanic.hasta80 <- titanic[titanic$edad_discretizada.type == "70-79"]
titanic.hasta90 <- titanic[titanic$edad_discretizada.type == "80-89"]
# Agrupación por embarcación
titanic.cherbourg <- titanic[titanic$Embarked.type == "C"]
titanic.queenstown <- titanic[titanic$Embarked.type == "Q"]
titanic.southampton <- titanic[titanic$Embarked.type == "S"]
```

Comprobación de la normalidad y homogeneidad de la varianza

Para comprobar si los valores de las variables cuantitativas forman parte de una población normalmente distribuida, se va a utilizar la prueba de normalidad de Anderson-Darling. Mediante esta prueba se testea que el p-valor de las variables, que mediante un valor prefijado nos mostrará si estas pertenecen a la distribución normal o no:

```
library(nortest)

alpha = 0.05
col.names = colnames(titanic)
for (i in 1:ncol(titanic)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(titanic[,i]) | is.numeric(titanic[,i])) {
    p_val = ad.test(titanic[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])

      # Format output
      if (i < ncol(titanic) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

Y obtenemos que las variables que no siguen una distribución normal son:

- Age
- SibSP
- Parch

```
## Variables que no siguen una distribución normal:
## Age, SibSp, Parch,
```

Pruebas estadísticas

Con nuestros datos primero vamos a dividirlos en datos de test y entrenamiento, ya que los datos proporcionados en el dataset de test no tienen la solución y no sirven para probar si el modelo está acertando o no.

```
index <- 0.75*nrow(titanic)
titanic <- titanic[sample(1:nrow(titanic)), ]
train <- titanic[1:index,]
test <- titanic[index:nrow(titanic),]
```

Una vez que tenemos los datos divididos en entrenamiento y test podemos hacer una regresión logística simple usando los datos de sexo para saber si ha sobrevivido o no:

```
titan_glm <- glm(Survived ~ Sex, data = train, family = 'binomial')
summary(titan_glm)
```

```
##
## Call:
## glm(formula = Survived ~ Sex, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6868  -0.6689  -0.6689   0.7428   1.7929
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.1468     0.1562   7.342 2.11e-13 ***
## Sexmale      -2.5303     0.1961 -12.903  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 892.07  on 667  degrees of freedom
## Residual deviance: 692.35  on 666  degrees of freedom
## AIC: 696.35
##
## Number of Fisher Scoring iterations: 4
```

Con el modelo ya creado nos disponemos a probarlo con los datos de test:

```
predict_sex <- predict(titan_glm,newdata = test,type = 'response')
predict_sex <- ifelse(predict_sex>0.5,1,0)
error <- mean(predict_sex!=test$Survived)
exactitud <- 1-error
exactitud
```

```
## [1] 0.7757848
```

Y obtenemos un 77% de efectividad. Asimismo podemos hacer un test completo para saber cual son las variables que más afectan a la supervivencia:

```
titanic_completo <- glm(Survived~., data=train, family = binomial)
summary(titanic_completo)
```

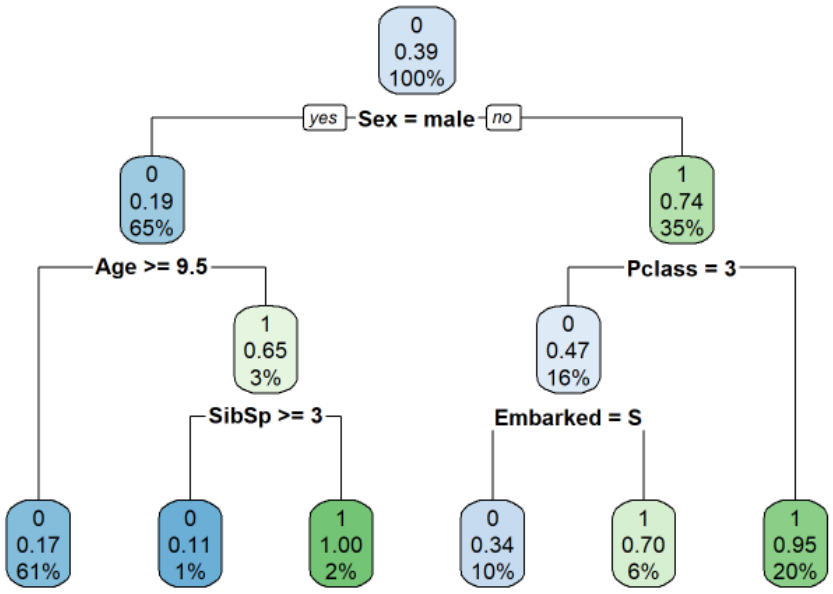
```
##
## Call:
## glm(formula = Survived ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1596  -0.6054  -0.3605   0.5533   2.3976
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)             5.05369    0.65538   7.711 1.25e-14 ***
## Pclass2                -0.93044    0.31614  -2.943  0.00325 **
## Pclass3                -2.29561    0.29620  -7.750 9.17e-15 ***
## Sexmale                -2.60116    0.23498 -11.070  < 2e-16 ***
## Age                    -0.05496    0.03697  -1.487  0.13712
## SibSp                  -0.36546    0.14160  -2.581  0.00985 **
## Parch                  -0.06607    0.13838  -0.477  0.63304
## EmbarkedQ               0.38881    0.44641   0.871  0.38377
## EmbarkedS              -0.42390    0.28331  -1.496  0.13459
## edad_discretizada10-19 -0.75983    0.70584  -1.076  0.28171
## edad_discretizada20-29 -0.80187    0.95573  -0.839  0.40146
## edad_discretizada30-39  0.13094    1.23400   0.106  0.91550
## edad_discretizada40-49 -0.07602    1.59932  -0.048  0.96209
## edad_discretizada50-59  0.26265    1.94672   0.135  0.89267
## edad_discretizada60-69 -0.40557    2.46818  -0.164  0.86948
## edad_discretizada70-79  1.59126    2.89388   0.550  0.58241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 892.07  on 667  degrees of freedom
## Residual deviance: 579.59  on 652  degrees of freedom
## AIC: 611.59
##
## Number of Fisher Scoring iterations: 5
```

Según esto las mayores probabilidades de morir son si eres hombre y en estar en 3ª clase.

Ahora se va a utilizar un árbol de decisión con la librería rpart:

```
set.seed(42)

library(rpart)
library(rpart.plot)
fit <- rpart(Survived~., data = train, method = 'class')
rpart.plot(fit, extra = 106)
```



Que si lo probamos:

```
predict_titanic <-predict(fit, test, type = 'class')
mat_conf <- table(test$Survived, predict_titanic)
mat_conf
```

```
##    predict_titanic
##       0   1
##   0 130   9
##   1  33  51
```

```
porcentaje_correct <-100 * sum(diag(mat_conf)) / sum(mat_conf)
print(sprintf("El %% de registros correctamente clasificados es: %.4f %%",porcentaje_correct))
```

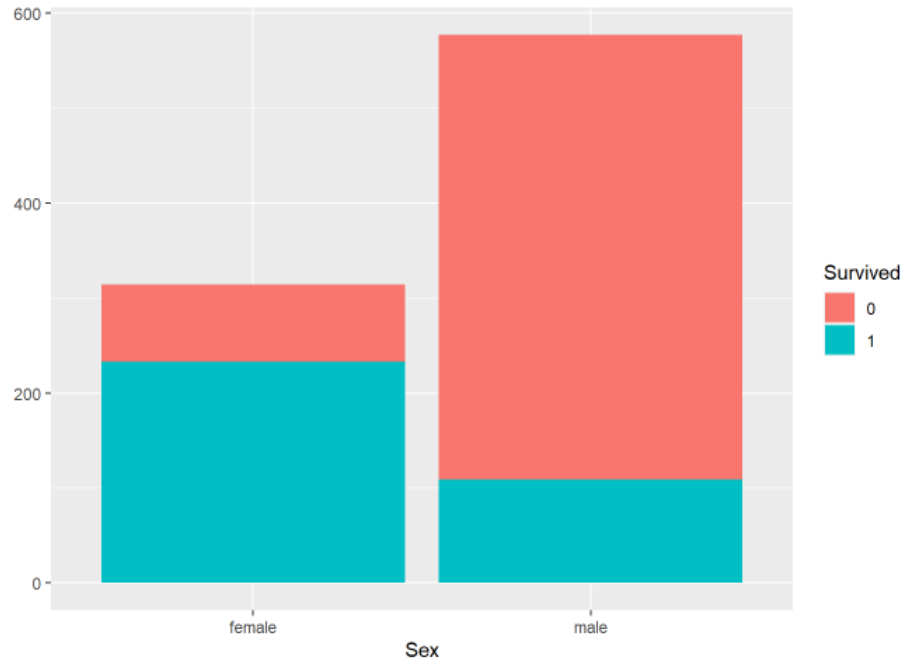
```
## [1] "El % de registros correctamente clasificados es: 81.1659 %"
```

Obtenemos que el 81,16% ha sido clasificado correctamente. Lo que resulta en una mejoría con respecto al método anterior.

Representación de los resultados a partir de tablas y gráficas

Ahora voy a representar los datos con respecto a supervivencia o no:

```
qplot(Sex, data=titanic, fill = Survived)
```




```
tabla_sexo <- table(titanic$Sex, titanic$Survived)
tabla_sexo
```

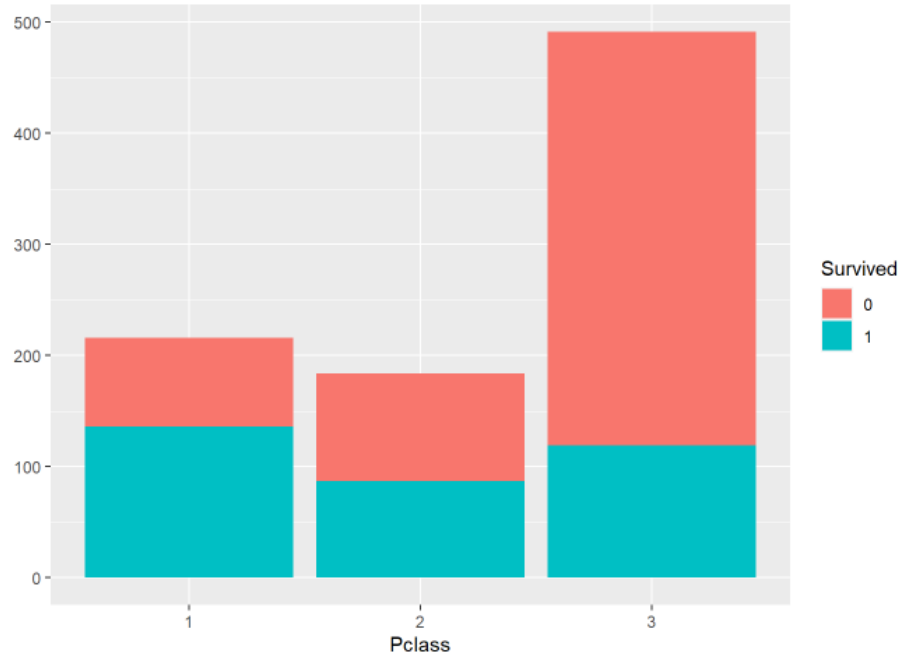
```
##
##             0   1
##   female   81 233
##   male    468 109
```

```
prop.table(tabla_sexo, margin = 1)
```

```
##
##                  0         1
##   female 0.2579618 0.7420382
##   male   0.8110919 0.1889081
```

Como podemos ver, se cumple la predicción de que siendo hombre se tenían muchas menos posibilidades de sobrevivir en el accidente que siendo mujer, siendo el 74% de probabilidad de sobrevivir siendo mujer y del 18% siendo hombre.

```
qplot(Pclass, data=titanic, fill = Survived)
```




```
tabla_clase <- table(titanic$Pclass, titanic$Survived)
tabla_clase
```

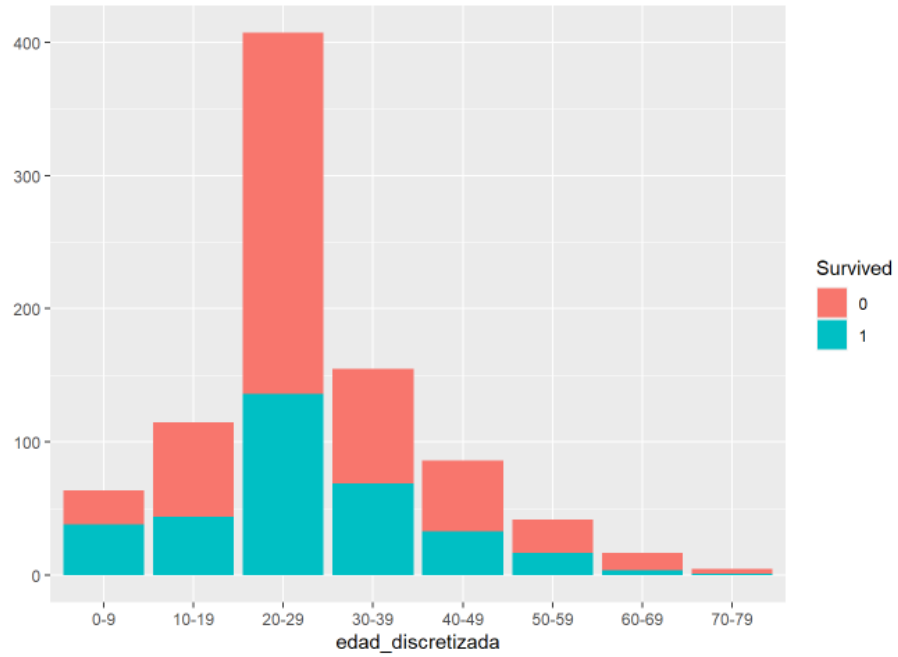
```
##
##       0   1
##   1  80 136
##   2  97  87
##   3 372 119
```

```
prop.table(tabla_clase, margin = 1)
```

```
##
##            0         1
##   1 0.3703704 0.6296296
##   2 0.5271739 0.4728261
##   3 0.7576375 0.2423625
```

Si ahora observamos por clase de billete, tenemos que los que más posibilidades de supervivencia son los de la primera clase, seguidos de segunda y en última posición tercera. Siendo 62% de sobrevivir en 1ª, 47% en 2ª y 24% en 3ª.

```
qplot(edad_discretizada, data=titanic, fill = Survived)
```




```
tabla_edad <- table(titanic$edad_discretizada, titanic$Survived)
tabla_edad
```

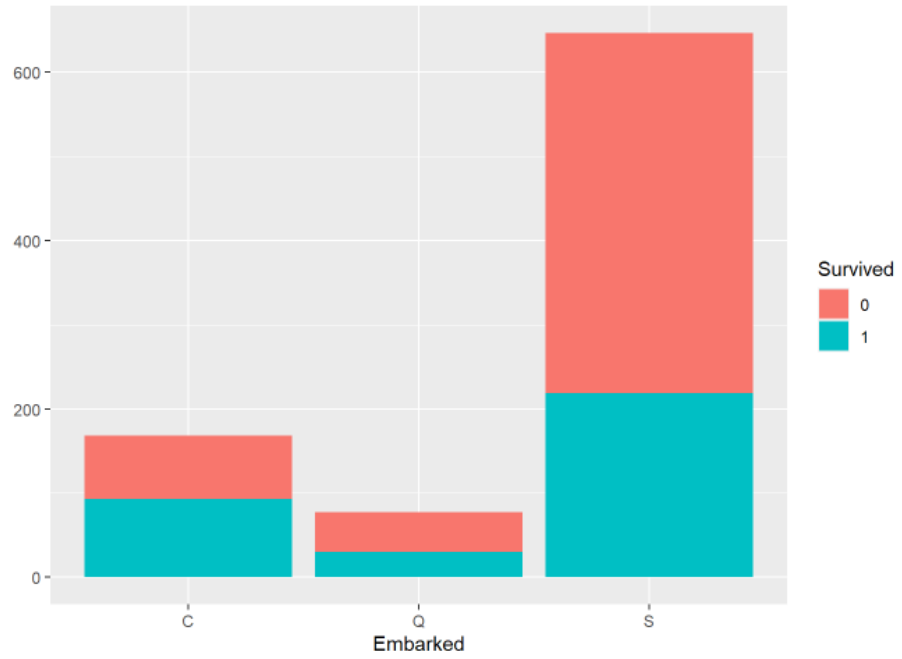
```
##
##           0   1
##   0-9    26  38
##   10-19  71  44
##   20-29 271 136
##   30-39  86  69
##   40-49  53  33
##   50-59  25  17
##   60-69  13   4
##   70-79   4   1
##   80-89   0   0
```

```
prop.table(tabla_edad, margin = 1)
```

```
##
##                0         1
##   0-9   0.4062500 0.5937500
##   10-19 0.6173913 0.3826087
##   20-29 0.6658477 0.3341523
##   30-39 0.5548387 0.4451613
##   40-49 0.6162791 0.3837209
##   50-59 0.5952381 0.4047619
##   60-69 0.7647059 0.2352941
##   70-79 0.8000000 0.2000000
##   80-89
```

En cuanto a la edad con más posibilidades de supervivencia es de 0 a 9 años con casi un 60% seguida de 30-39 y 50-59.

```
qplot(Embarked, data=titanic, fill = Survived)
```




```
tabla_embarque <- table(titanic$Embarked, titanic$Survived)
tabla_embarque
```

```
##
##       0   1
##       0   0
##   C  75  93
##   Q  47  30
##   S 427 219
```

```
prop.table(tabla_embarque, margin = 1)
```

```
##
##             0         1
##
##   C 0.4464286 0.5535714
##   Q 0.6103896 0.3896104
##   S 0.6609907 0.3390093
```

En cuanto al sitio de embarque vemos que el sitio con más probabilidades de supervivencia es C que corresponde con el puerto de Cherbourg. Vamos a comprobar que clases se montaron en ese puerto:

```
tabla_embarque2 <- table(titanic$Embarked, titanic$Pclass)
tabla_embarque2
```

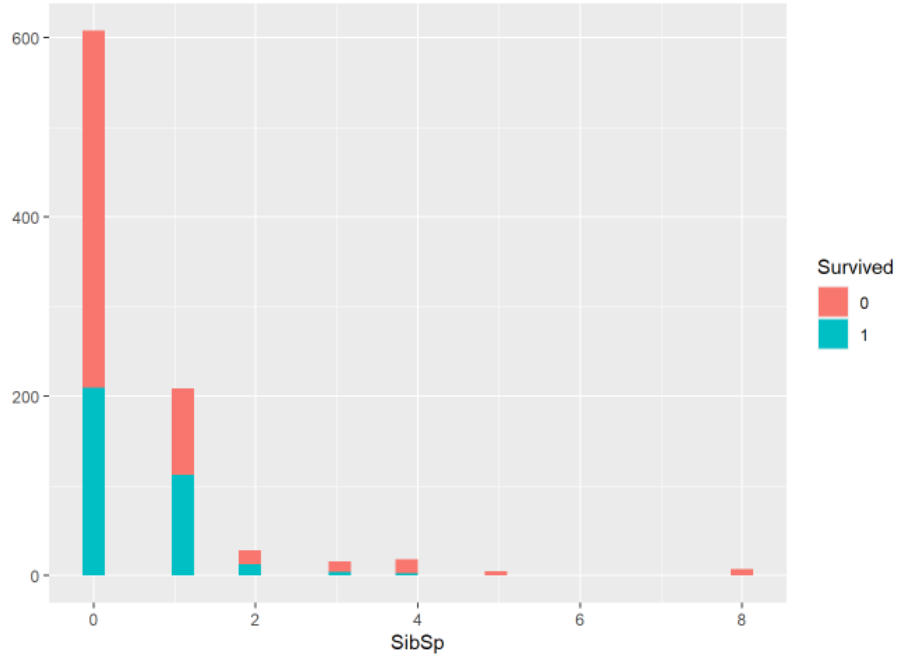
```
##
##       1   2   3
##       0   0   0
##   C  85  17  66
##   Q   2   3  72
##   S 129 164 353
```

```
prop.table(tabla_embarque2, margin = 1)
```

```
##
##              1          2          3
##
##   C 0.50595238 0.10119048 0.39285714
##   Q 0.02597403 0.03896104 0.93506494
##   S 0.19969040 0.25386997 0.54643963
```

Como podemos comprobar el 50% de la gente que se monto en ese puerto era de primera clase, motivo por el cual se puede deber la alta supervivencia de este.

```
qplot(SibSp, data=titanic, fill = Survived)
```




```
tabla_familia <- table(titanic$SibSp, titanic$Survived)
tabla_familia
```

```
##
##       0   1
##   0 398 210
##   1  97 112
##   2  15  13
##   3  12   4
##   4  15   3
##   5   5   0
##   8   7   0
```

```
prop.table(tabla_familia, margin = 1)
```

```
##
##             0         1
##   0 0.6546053 0.3453947
##   1 0.4641148 0.5358852
##   2 0.5357143 0.4642857
##   3 0.7500000 0.2500000
##   4 0.8333333 0.1666667
##   5 1.0000000 0.0000000
##   8 1.0000000 0.0000000
```

En cuanto al número de familiares parece que la mayor supervivencia esta en 1 o 2 familiares en el barco teniendo alrededor del 50% de supervivencia.

Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Podemos decir, que para tener las mayores posibilidades de supervivencia en el Titanic se tenia que ser mujer de 1ª clase de 0 a 9 años o de 30 a 39 y que seguramente embarco en Cherbourg con 0,1 o 2 familiares a bordo. Y que para tener las mínimas posibilidades de supervivencia ser hombre de 3ª clase de edad avanzada y que embarco en Southampton con 5 o más familiares en el barco.

En cuanto al modelo entrenado, el modelo que mejor resultados ha dado, es el árbol de decisión conseguido mediante la librería rpart, con el que se ha conseguido acertar en el 81% de las veces en los datos de prueba.

Código

Adjunto en fichero RMD.

Contribuciones	Firma
Investigación previa	Adrián López Ibáñez
Redacción de las respuestas	Adrián López Ibáñez
Desarrollo código	Adrián López Ibáñez