Tipología y ciclo de vida de los datos: práctica 2

Autor: Sergio Blay González

Enero 2020

---

1. Descripción del dataset

El conjunto de datos escogido para el desarrollo de esta práctica ha sido el que ofrece la plaforma Kaggle para su concurso titulado: Titanic: Machine Learning from Disaster. Dicho dataset consiste en dos ficheros CSV, train y test, que contienen 12 y 11 columnas respectivamente. Cada registro representa a cada uno de los pasajeros que iban a bordo del Titanic de los cuales se nos indican las siguientes características:

- PassengerId: identificador único del pasajero.
- Survived: variable binaria que indica si el sujeto sobrevivió (1) o falleció (0). Dado que es el campo a estimar, este campo no está incluido en el test.
- Pclass: clase en la que viajaba el pasajero (1, 2 o 3).
- Name: nombre del pasajero.
- Sex: sexo del pasajero (male o female).
- Age: edad del pasajero.
- SibSp: número de hermanos/as o esposos/as a bordo del Titanic.
- Parch: número de padres o hijos/as a bordo del Titanic.
- Ticket: número del ticket del pasajero.
- Fare: tarifa pagada por el pasajero.
- Cabin: número de camarote asignado.
- Embarked: puerto en el que embarcó. C(Cherbourg), Q(Queenstown), S(Southampton).

Este conjunto de datos se pretende utilizar para esclarecer cuáles son las características (si las hubiese), que proporcionasen una mayor probabilidad de supervivencia a los pasajeros, o por el contrario una mayor probabilidad de no sobrevivir.

Carga de los datos:

```
data<-read.csv("./train.csv",header=T,sep=",")
data$Pclass <- factor(data$Pclass, levels=c(1,2,3), labels=c("1ª class", "2ª class", "3ª class"))
data$Survived <- factor(data$Survived)
levels(data$Survived)[levels(data$Survived)=="0"] <- "Dead"
levels(data$Survived)[levels(data$Survived)=="1"] <- "Survived"
summary(data)
```

```
##   PassengerId        Survived        Pclass
##  Min.   :  1.0   Dead    :549   1ª class:216
##  1st Qu.:223.5   Survived:342   2ª class:184
##  Median :446.0                  3ª class:491
##  Mean   :446.0
##  3rd Qu.:668.5
##  Max.   :891.0
##
##                                     Name         Sex           Age
##  Abbing, Mr. Anthony                   :  1   female:314   Min.   : 0.42
##  Abbott, Mr. Rossmore Edward           :  1   male  :577   1st Qu.:20.12
##  Abbott, Mrs. Stanton (Rosa Hunt)      :  1                Median :28.00
##  Abelson, Mr. Samuel                   :  1                Mean   :29.70
##  Abelson, Mrs. Samuel (Hannah Wizosky):  1                3rd Qu.:38.00
##  Adahl, Mr. Mauritz Nils Martin        :  1                Max.   :80.00
##  (Other)                               :885                NA's   :177
##      SibSp           Parch             Ticket         Fare
##  Min.   :0.000   Min.   :0.0000   1601    :  7   Min.   :  0.00
##  1st Qu.:0.000   1st Qu.:0.0000   347082  :  7   1st Qu.:  7.91
##  Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45
##  Mean   :0.523   Mean   :0.3816   3101295 :  6   Mean   : 32.20
##  3rd Qu.:1.000   3rd Qu.:0.0000   347088  :  6   3rd Qu.: 31.00
##  Max.   :8.000   Max.   :6.0000   CA 2144 :  6   Max.   :512.33
##                                   (Other) :852
##          Cabin     Embarked
##             :687    :  2
##  B96 B98    :  4   C:168
##  C23 C25 C27:  4   Q: 77
##  G6         :  4   S:644
##  C22 C26    :  3
##  D          :  3
##  (Other)    :186
```

2. Selección de los datos

De los atributos mencionados en el apartado anterior, se ha determinado que hay varios que son irrelevantes para la pregunta que se quiere resolver.

- PassengetId: la identificación exacta del pasajero no es relevante.

```
data$PassengerId <- NULL
```

- Name: la identificación exacta del pasajero no es relevante.

```
data$Name <- NULL
```

- Ticket: otra forma de identificar a los pasajeros, pero esta vez por grupos de individuos como familias que comparten el mismo número de ticket.

```
data$Ticket <- NULL
```

- Embarked: se podría pensar que el puerto de embarque podría estar relacionado con la clase en la que se viaja, ya que partir desde el puerto de una ciudad más rica podría ser indicativo de una mayor riqueza. Estudiemos la posible correlación entre estas dos variables.

En primer lugar vamos a tratar sus valores perdidos, como hemos observado en el resumen anteriormente mostrado, existen dos registros sin puerto de embarque. Para paliar esto, vamos a asignarles a los registros un puerto basado en el puerto más común de entre los registros con características similares a ellos.

```
data[data$Embarked == "",]
```

```
##     Survived   Pclass    Sex Age SibSp Parch Fare Cabin Embarked
## 62  Survived 1ª class female  38     0     0   80   B28
## 830 Survived 1ª class female  62     0     0   80   B28
```

```
similar <- data[data$Survived == "Survived" & data$Pclass == "1ª class" & data$Sex == "female" & dat
a$SibSp == 0 & data$Parch == 0,]
table(similar$Embarked)
```

```
##
##     C  Q  S
##  2 17  0 14
```

```
data$Embarked[data$Embarked == ""] <- "C"
```

A continuación, vamos a ver unas tablas de frecuencia para tener una primera idea acerca de si vamos encaminados o no en nuestra suposición

```
tabla <- table(droplevels(data$Embarked, exclude = ""),data$Pclass)
prop.table(tabla, margin = 1)
```

```
##
##       1ª class   2ª class   3ª class
##   C 0.51176471 0.10000000 0.38823529
##   Q 0.02597403 0.03896104 0.93506494
##   S 0.19720497 0.25465839 0.54813665
```

Podemos observar en la tabla, que partiendo desde Queenstown es muy probable que el pasajero sea de 3ª clase, mientras que más de la mitad de los pasajeros de Cherbourg (51.11%) lo hacen en primera. Vamos a realizar un test de dependencia para verificarlo.

```
data$Embarked <- droplevels(data$Embarked)
tbl = table(data$Pclass, data$Pclass)
chisq.test(tbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 1782, df = 4, p-value < 2.2e-16
```

```
data$Embarked <- NULL
```

Puesto que el pvalue es menor del nivel de significancia de 0.05, rechazamos la hipótesis nula de independencia. Por lo tanto, ambas variables tienen una relación de dependencia y vamos a eliminar la columna Embarked.

- Fare: de la misma forma que el puerto de embarque, podríamos pensar que el precio del billete está supeditado a la clase en la que se viaja, ya que los precios son mayores en las clases más altas.

```
data$Fare <- factor(cut(data$Fare, quantile(data$Fare), include.lowest = TRUE), labels = c("Low","Lo
w/Mid","Mid/High","High"))

tabla <- table(data$Fare,data$Pclass)
prop.table(tabla, margin = 1)
```

```
##
##              1ª class   2ª class   3ª class
##   Low      0.02690583 0.02690583 0.94618834
##   Low/Mid  0.00000000 0.38392857 0.61607143
##   Mid/High 0.22972973 0.31531532 0.45495495
##   High     0.71621622 0.09909910 0.18468468
```

```
tbl = table(data$Fare, data$Pclass)
chisq.test(tbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 519.54, df = 6, p-value < 2.2e-16
```

```
data$Fare <- NULL
```

En la tabla se puede observar que los que han pagado un precio Alto, viajan en un 71% de los casos es primera clase, mientras que en su lado opuesto haber pagado el precio Low lleva a los pasajeros a viajar en 3ª clase en un 94% de los casos. Este hecho se confirma en el test realizado, al tener un p-value menor que el nivel de significación. Por lo tanto vamos a eliminar esta columna también.

- Cabin: más del 77% de los registros tienen este campo vacío, es difícil obtener utilidad de él.

```
(nrow(data[data$Cabin == "",]) / nrow(data))*100
```

```
## [1] 77.10438
```

```
data$Cabin <- NULL
```

3. Limpieza de los datos

3.1 Valores perdidos

Antes de nada, lo primero será obtener un breve resumen estadístico del conjunto de datos resultante del apartado anterior.

```
summary(data)
```

```
##      Survived        Pclass        Sex           Age            SibSp
##  Dead    :549   1ª class:216   female:314   Min.   : 0.42   Min.   :0.000
##  Survived:342   2ª class:184   male  :577   1st Qu.:20.12   1st Qu.:0.000
##                 3ª class:491                Median :28.00   Median :0.000
##                                             Mean   :29.70   Mean   :0.523
##                                             3rd Qu.:38.00   3rd Qu.:1.000
##                                             Max.   :80.00   Max.   :8.000
##                                             NA's   :177
##      Parch
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3816
##  3rd Qu.:0.0000
##  Max.   :6.0000
##
```

En lo que se refiere a valores perdidos, se observa que el único campo que los contiene es el de la Edad (age). Para recuperar estos valores se utilizará la función kNN (k vecinos más próximos) del paquete VIM para estimar la edad en función de registros con características similares del que se queire estimar.

```
library("VIM")
knnData<-kNN(data, k=3)
data$Age <- knnData$Age
summary(data)
```

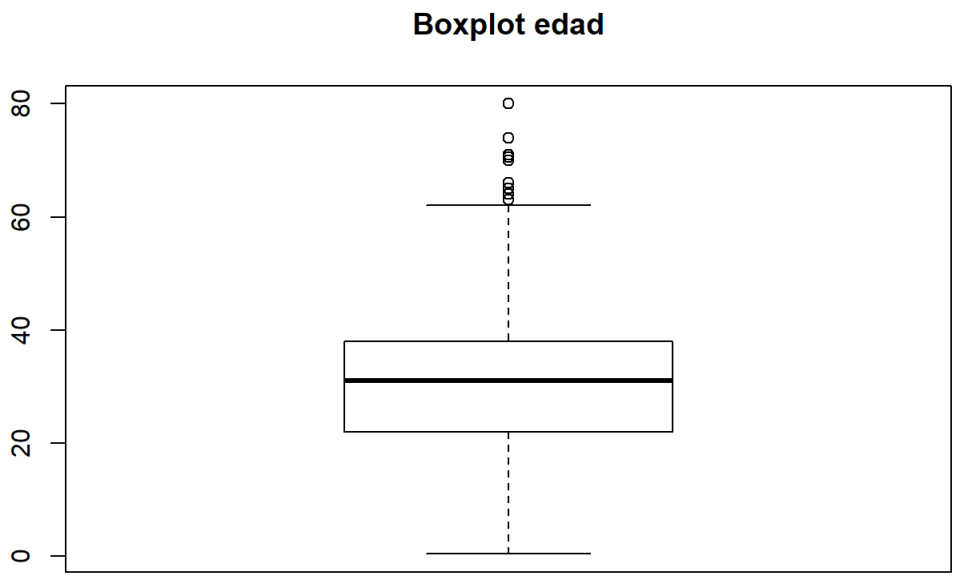
```
##      Survived        Pclass        Sex           Age            SibSp
##  Dead    :549   1ª class:216   female:314   Min.    : 0.42   Min.   :0.000
##  Survived:342   2ª class:184   male  :577   1st Qu.:22.00   1st Qu.:0.000
##                 3ª class:491                Median :31.00   Median :0.000
##                                             Mean   :30.98   Mean   :0.523
##                                             3rd Qu.:38.00   3rd Qu.:1.000
##                                             Max.   :80.00   Max.   :8.000
##      Parch
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3816
##  3rd Qu.:0.0000
##  Max.   :6.0000
```

En algunos casos, el valor 0 también puede significar un valor perdido. Sim embargo, dado que en las columnas numéricas de este dataset el valor 0 entra dentro del domino, para este caso no se condirera que sea un valor perdido.

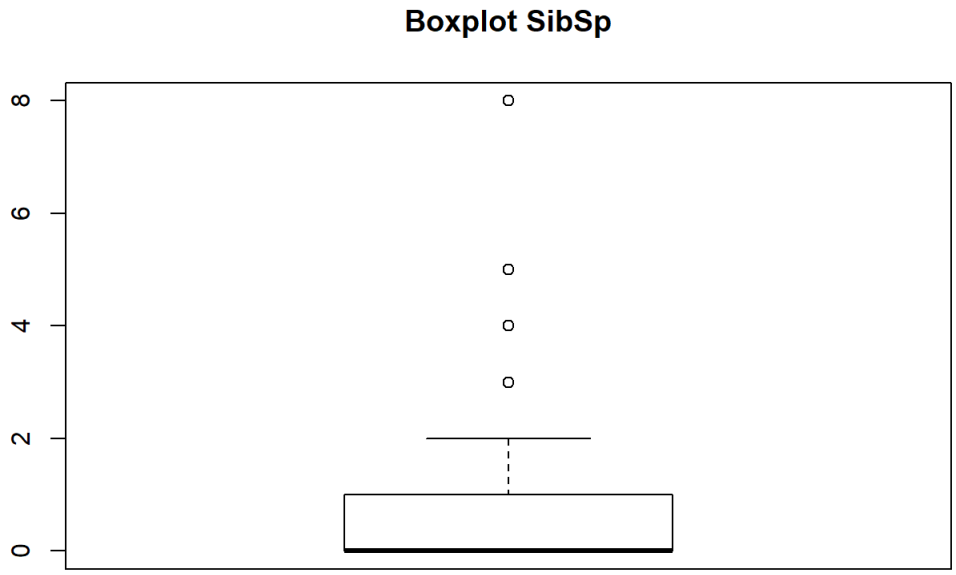
3.2 Valores extremos

Los valores extremos, valores atípicos u outliers, son aquellas observaciones que destacan sobre las demás debido a que se desvían demasiado del centro pudiendo alterar los resultados que se quieran obtener, por ello deberían ser tratados. En este apartado vamos a utilizar un diagrama de cajas para estudiar los outliers que pudiesen existir en los atributos numéricos Age, SibSp y Parch.

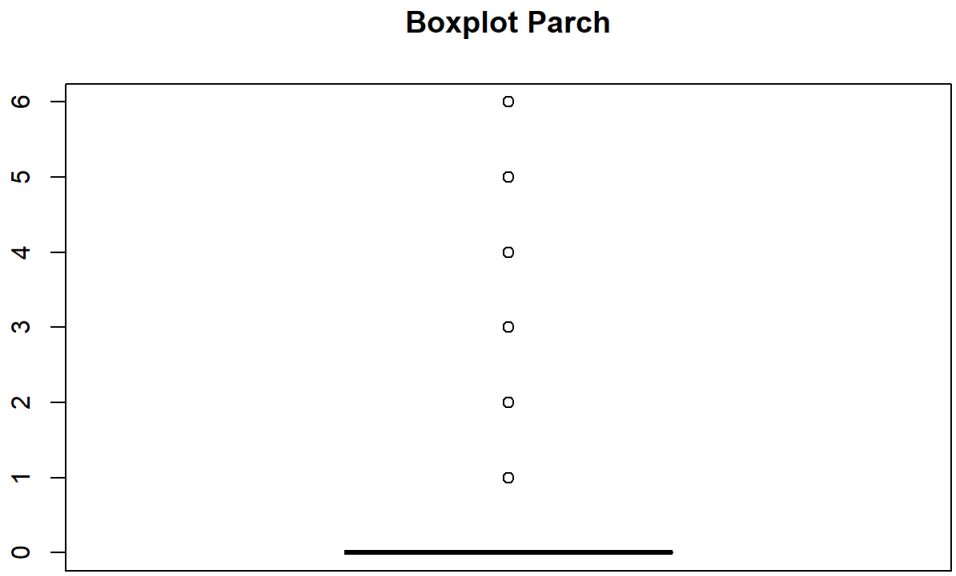
```
bpAge <- boxplot(data$Age)
title("Boxplot edad")
```




```
bpSibSp <- boxplot(data$SibSp)
title("Boxplot SibSp")
```




```
bpParch <- boxplot(data$Parch)
title("Boxplot Parch")
```



Podemos observar en el gráfico que existen una serie de valores entre los 60 y 80 años que se desvían del centro de los datos. Sin embargo, no los consideraremos como outliers, simplemente consideraremos que entre los pasajeros hay un reducido número de gente mayor. Para los casos de SibSp y Parch ocurre exactamente lo mismo, lo más usual es que no se viaje con otros familiares, sin embargo puede darse el caso de pasajeros con hasta 6-8 familiares a bordo.

4. Análisis de los datos

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

En este punto, y dado el objetivo de la práctica, vamos a realizar el análisis de homogeneidad de la varianza para el atributo edad teniendo en cuenta los grupos en los que se didive el atributo objetivo: vivo o fallecido.

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

En este apartado, vamos a comprobar la normalidad de los atributos de los que disponemos utilizando el test de Anderson-Darling existente en el paquete "nortest".

```
library(nortest)
cols <- c("Pclass", "Age", "Sex", "SibSp", "Parch")
a <- 0.05

for(col in cols){
    pvalue <- ad.test(as.numeric(data[,col]))$p.value
    if(pvalue < a){
      print(paste(col, "no cumple el test de normalidad"))
    } else{
      print(paste(col, "cumple el test de normalidad"))
    }
}
```

```
## [1] "Pclass no cumple el test de normalidad"
## [1] "Age no cumple el test de normalidad"
## [1] "Sex no cumple el test de normalidad"
## [1] "SibSp no cumple el test de normalidad"
## [1] "Parch no cumple el test de normalidad"
```

Ahora vamos a analizar la homogeneidad de las varianzas mediante la prueba de Fligner-Killeen para los campos "Survived" y "Age"

```
fligner.test(as.numeric(Survived) ~ Age, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  as.numeric(Survived) by Age
## Fligner-Killeen:med chi-squared = 119.28, df = 87, p-value = 0.01235
```

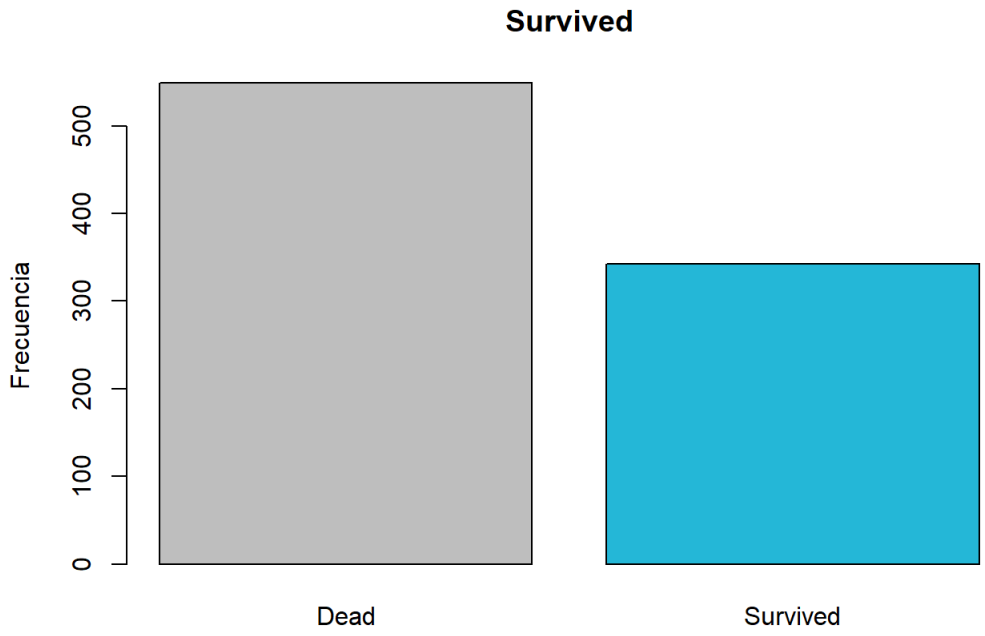
Dado que este test tampoco supera el nivel de significación de 0.05 rechazamos la hipótesis de que las varianzas de ambas muestras son homogéneas.

#4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos

En primer lugar vamos a ver de forma gráfica la distribución de cada atributo, por si se pudiese obtener alguna primera conclusión:

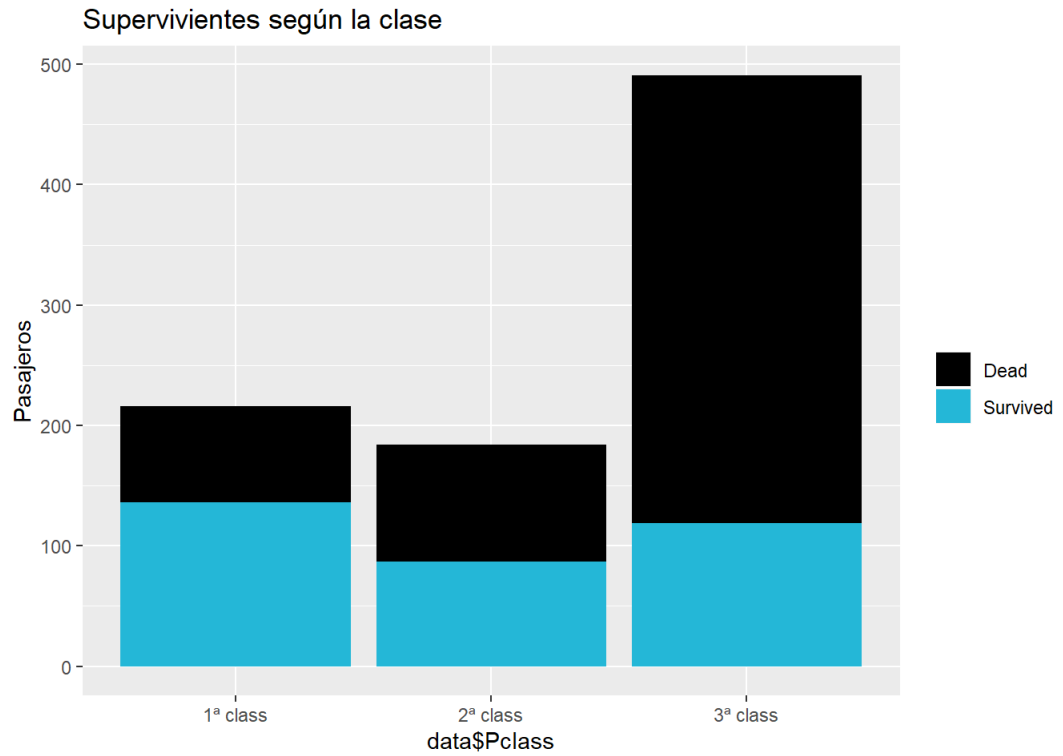
Supervivientes:

```
library(ggplot2)
plot(x = data$Survived, main = "Survived", ylab = "Frecuencia",
     col = c("grey", "#24B7D7"))
```



Supervivencia según la clase:

```
library(ggplot2)
ggplot(data,aes(data$Pclass,fill=data$Survived))+geom_bar() +labs(y="Pasajeros")+
guides(fill=guide_legend(title=""))+
scale_fill_manual(values=c("black","#24B7D7"))+ggtitle("Supervivientes según la clase")
```




```
tabla <- table(data$Pclass,data$Survived)
prop.table(tabla, margin = 1)
```

```
##
##                 Dead  Survived
##   1ª class 0.3703704 0.6296296
##   2ª class 0.5271739 0.4728261
##   3ª class 0.7576375 0.2423625
```

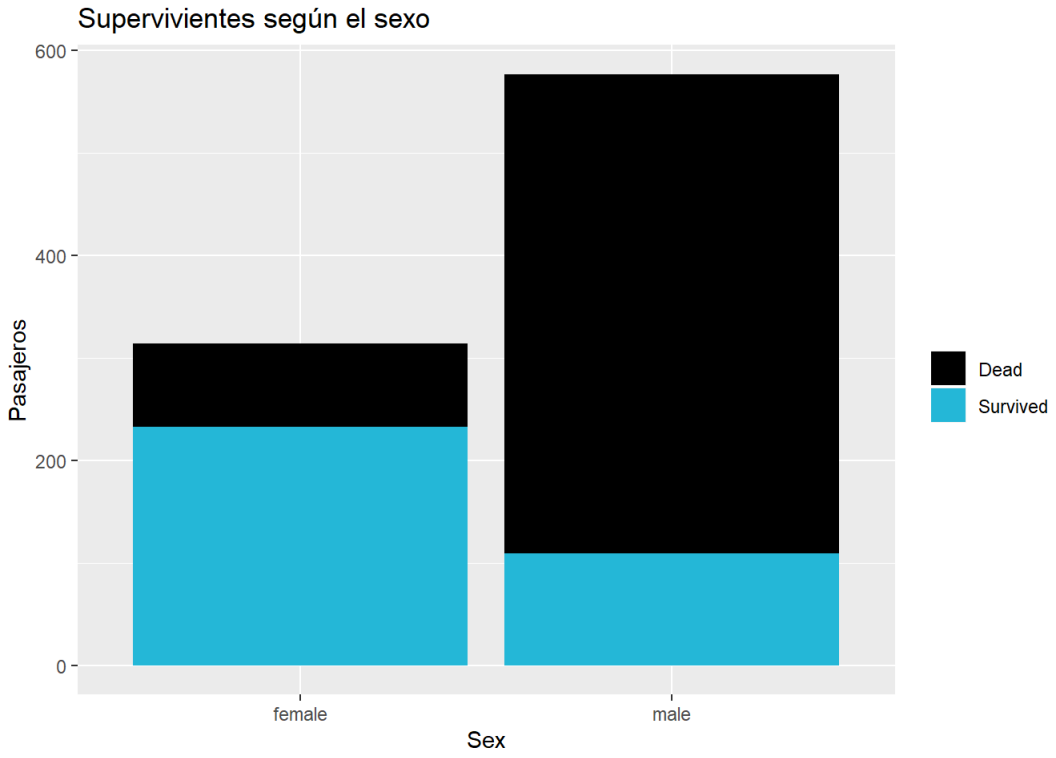
```
table(data$Survived,data$Pclass)
```

```
##
##            1ª class 2ª class 3ª class
##   Dead           80       97      372
##   Survived      136       87      119
```

Como podemos observar en el gráfico y en la tabla, los pasajeros de tercera clase fallecieron en más de un 75% de los casos siendo esta clase donde se concentra la mayoría de los fallecidos: 372 fallecidos de un total de 549 (más del 67%). Los casos de supervivencia se elevan en segunda clase con respecto a la tercera, quedando unos resultados más parejos (52.71% fallecidos y 47.28% vivos), y se elevan aún más en primera clase, siendo la única clase que tiene más supervivientes que fallecidos.

Supervivencia según el sexo:

```
ggplot(data,aes(Sex,fill=data$Survived))+geom_bar() +labs(y="Pasajeros")+
guides(fill=guide_legend(title=""))+
scale_fill_manual(values=c("black","#24B7D7"))+ggtitle("Supervivientes según el sexo")
```



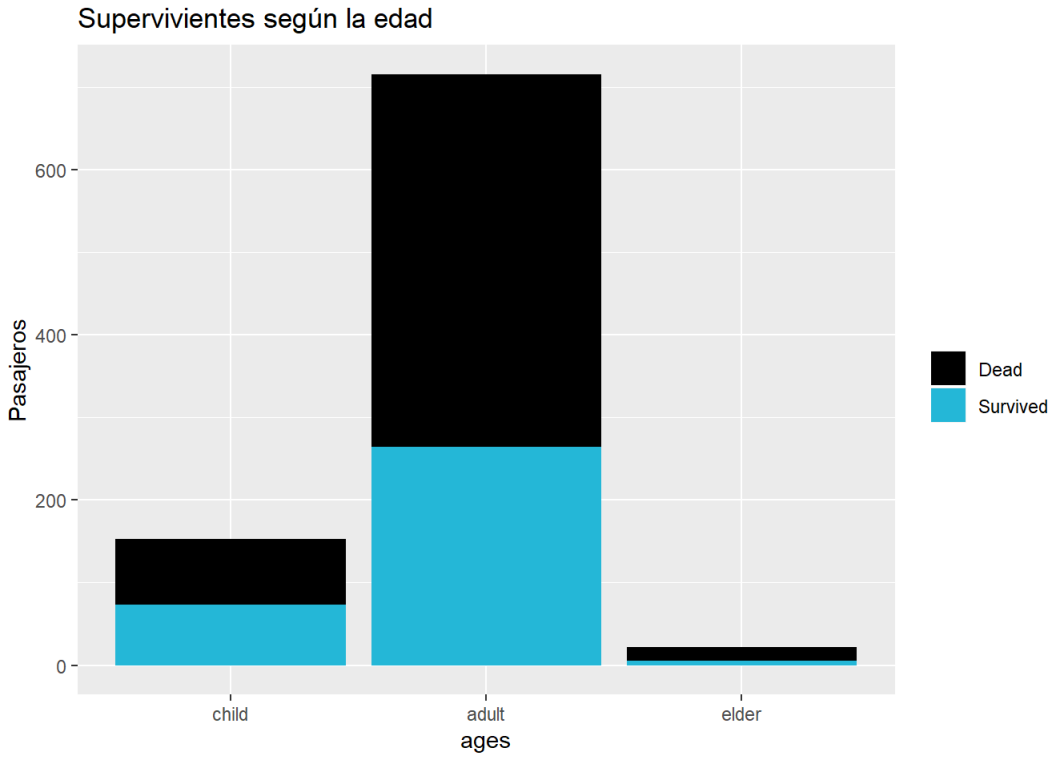

```
tabla <- table(data$Sex,data$Survived)
prop.table(tabla, margin = 1)
```

```
##
##                Dead  Survived
##   female 0.2579618 0.7420382
##   male   0.8110919 0.1889081
```

De este gráfico y tabla podemos conjeturar que según el sexo, las mujeres poseían una probabilidad de sobrevivir (74.2%) muy superior que los hombres (18.89%). Si a esto añadimos que había un mayor número de hombres, podemos ver que el grueso del número de fallecidos según sexo se encuentra en ellos.

Supervivencia según la edad:

```
ages <- cut(data$Age, breaks=c(-Inf, 18, 60,Inf), labels=c("child","adult", "elder"))
ggplot(data,aes(ages,fill=data$Survived))+geom_bar() +labs(y="Pasajeros")+
guides(fill=guide_legend(title=""))+
scale_fill_manual(values=c("black","#24B7D7"))+ggtitle("Supervivientes según la edad")
```



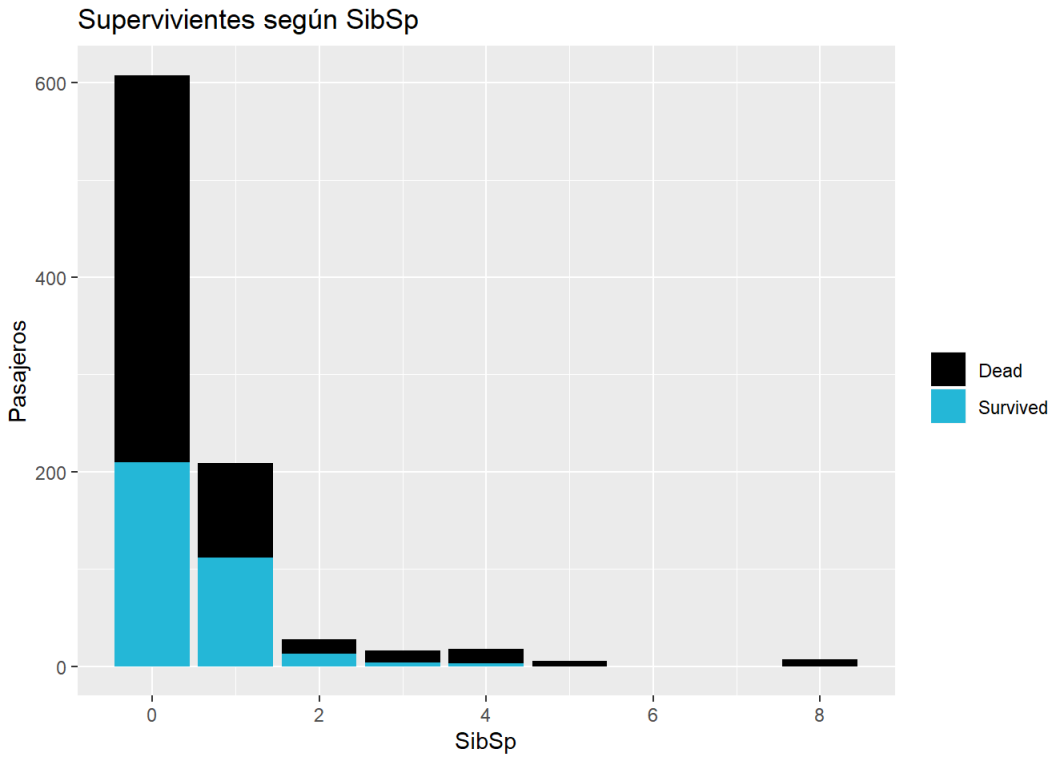

```
tabla <- table(ages,data$Survived)
prop.table(tabla, margin = 1)
```

```
##
## ages          Dead  Survived
##   child 0.5228758 0.4771242
##   adult 0.6312849 0.3687151
##   elder 0.7727273 0.2272727
```

Mientras que un poco más de la mitad de los niños (51.06%) ha sobrevivido, apenas un 36% de los adultos lo ha hecho y un 22% de los más mayores.

Supervivencia según SibSp:

```
ggplot(data,aes(SibSp,fill=data$Survived))+geom_bar() +labs(y="Pasajeros")+
guides(fill=guide_legend(title=""))+
scale_fill_manual(values=c("black","#24B7D7"))+ggtitle("Supervivientes según SibSp")
```

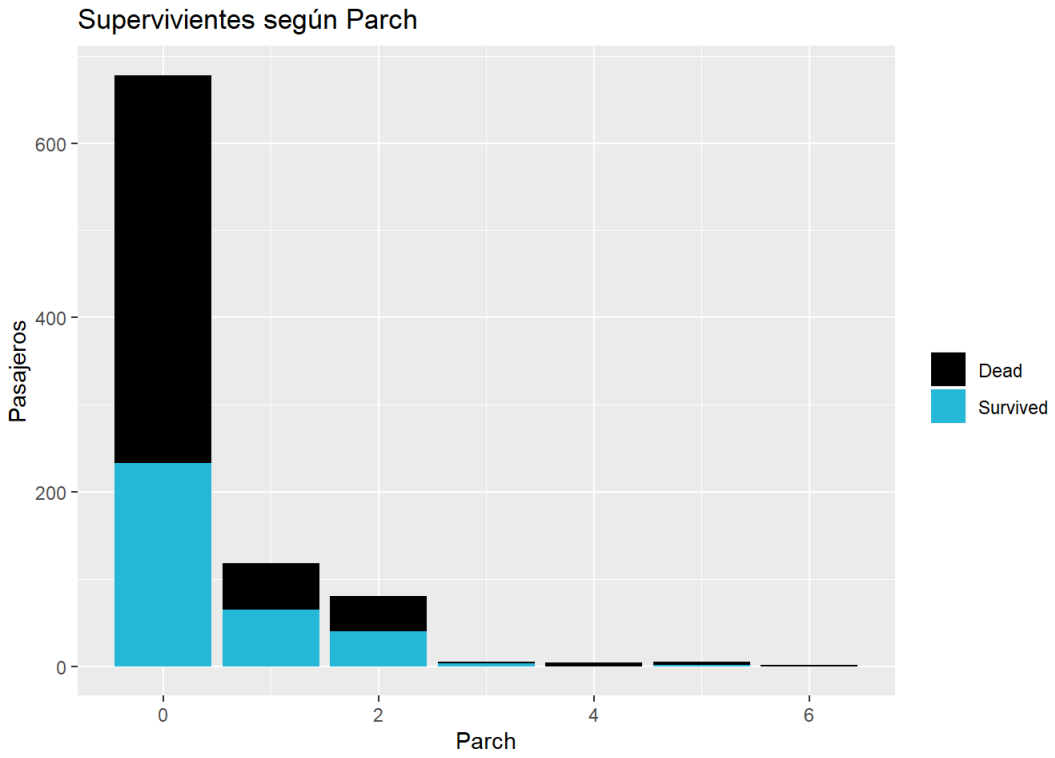



```
tabla <- table(data$SibSp,data$Survived)
prop.table(tabla, margin = 1)
```

```
##
##           Dead  Survived
##    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
```

Supervivencia según Parch:

```
ggplot(data,aes(Parch,fill=data$Survived))+geom_bar() +labs(y="Pasajeros")+
guides(fill=guide_legend(title=""))+
scale_fill_manual(values=c("black","#24B7D7"))+ggtitle("Supervivientes según Parch")
```




```
tabla <- table(data$Parch,data$Survived)
prop.table(tabla, margin = 1)
```

```
##
##           Dead  Survived
##   0 0.6563422 0.3436578
##   1 0.4491525 0.5508475
##   2 0.5000000 0.5000000
##   3 0.4000000 0.6000000
##   4 1.0000000 0.0000000
##   5 0.8000000 0.2000000
##   6 1.0000000 0.0000000
```

En cuanto a SibSp y Parch, podemos observar que tiene unos números muy parecidos, siendo el valor 1 el único cuya frecuencia a la hora de sobrevivir es mayor de la mitad.

Vistos estos gráficos de barras hemos podido obtener una serie de hipótesis que deberíamos contrastar antes de poder confirmarlas.

4.3.1 Modelo de árbol de decisión

En este apartado vamos a crear un modelo de árbol de decisión para obtener una serie de reglas con las que sea posible deducir los valores del atributo “Survived” del conjunto de prueba de los datos.

```
target <- data[,1]
attr <- data[,2:6]

model <- C50::C5.0(attr, target,rules=TRUE )
summary(model)
```

```
##
## Call:
## C5.0.default(x = attr, y = target, rules = TRUE)
##
##
## C5.0 [Release 2.07 GPL Edition]      Tue Jan 07 22:41:23 2020
## -----------------------------
##
## Class specified by attribute `outcome'
##
## Read 891 cases (6 attributes) from undefined.data
```

```
## Read 891 cases (6 attributes) from undefined.data
##
## Rules:
##
## Rule 1: (93, lift 1.6)
##  Pclass = 3ª class
##  Age > 33
##  Age <= 38
##  SibSp <= 0
##  ->  class Dead  [0.989]
##
## Rule 2: (13, lift 1.5)
##  Pclass = 3ª class
##  Sex = female
##  Age > 38
##  Age <= 49
##  ->  class Dead  [0.933]
##
## Rule 3: (42/4, lift 1.4)
##  Pclass = 3ª class
##  SibSp > 2
##  ->  class Dead  [0.886]
##
## Rule 4: (84/12, lift 1.4)
##  Pclass = 3ª class
##  Age > 7
##  Age <= 33
##  SibSp > 0
##  ->  class Dead  [0.849]
##
## Rule 5: (577/109, lift 1.3)
##  Sex = male
##  ->  class Dead  [0.810]
##
## Rule 6: (37/2, lift 2.4)
##  Age <= 7
##  SibSp <= 2
##  ->  class Survived  [0.923]
##
## Rule 7: (43/6, lift 2.2)
##  Age > 53
##  Age <= 55
##  ->  class Survived  [0.844]
##
## Rule 8: (314/81, lift 1.9)
##  Sex = female
##  ->  class Survived  [0.741]
##
## Default class: Dead
##
##
## Evaluation on training data (891 cases):
##
##          Rules
##    ----------------
##      No      Errors
##
##       8  121(13.6%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     515    34    (a): class Dead
##      87   255    (b): class Survived
##
##
```

```
##  Attribute usage:
##
##  100.00% Sex
##   30.30% Age
##   26.26% SibSp
##   23.57% Pclass
##
##
## Time: 0.0 secs
```

Según las reglas obtenidas, se confirma lo que ya observamos a lo largo de la memoria. El atributo con mayor incidencia sobre la supervivencia es el sexo, seguido de la edad. Además, de entre las reglas obtenidas podemos observar que la mayor probabilidad (92.3% de los casos) de sobrevivir al incidente se daba en los niños con dos o menos hermanos. Por el contrario, en un 98.9% de los casos de pasajeros de tercera clase entre los 33 y 38 años sin hermanos o esposo/a a bordo, el pasajero fallecía.

Estas reglas han obtenido una tasa de error en la evaluación del modelo del 13.6%.

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

En este apartado vamos a utilizar el modelo obtenido en el apartado anterior para contrastarlo con el conjunto de prueba.

```
dataTest<-read.csv("./test.csv",header=T,sep=",")
dataTest$Name <- NULL
dataTest$Ticket <- NULL
dataTest$Fare <- NULL
dataTest$Cabin <- NULL
dataTest$Embarked <- NULL
dataTest$Pclass <- factor(dataTest$Pclass, levels=c(1,2,3), labels=c("1ª class", "2ª class", "3ª clas
s"))
results<-read.csv("./results.csv",header=T,sep=",")
results$PassengerId <- NULL

survived_pr <- results[,1]
attr2 <- dataTest[,2:6]
predicted_model <- predict(model, attr2, type="class")

mat_conf<-table(survived_pr,Predicted=predicted_model)
mat_conf
```

```
##            Predicted
## survived_pr Dead Survived
##           0  255       11
##           1   22      130
```

```
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: 92.1053 %"
```

Con un 92.10% de registros clasificados correctamente, la calidad del modelo es bastante alta.

6. Resolución del problema

Como conclusión sobre el trabajo realizado en esta memoria, podemos decir que se puede resolver la pregunta planteada inicialmente:

- “esclarecer cuáles son las características (si las hubiese), que proporcionasen una mayor probabilidad de supervivencia a los

pasajeros, o por el contrario una mayor probabilidad de no sobrevivir”

El modelo, así como los análisis previos nos han otorgado un gran conocimiento sobre los datos y las preguntas planteadas en el primer apartado, que nos han permitido generar un modelo de clasificación con un buen porcentaje de aciertos.