Práctica 2 - Limpieza y análisis de datos

Pablo López Ladrón de Guevara Rafael Corvillo Alonso

6 de junio, 2021

Contents

1 Descripción del dataset.

El *Titanic* fue un transatlántico británico, el mayor barco de pasajeros del mundo al finalizar su construcción, que se hundió durante la noche del 14 y la madrugada del 15 de abril de 1912 durante su viaje inaugural desde Southampton a Nueva York. En el hundimiento del Titanic murieron 1496 personas de las 2208 que iban a bordo, lo que convierte a esta catástrofe en uno de los mayores naufragios de la historia ocurridos en tiempos de paz.

Tenemos a nuestra disposición un conjunto de datos con información de 891 pasajeros de los 2208 que viajaban a bordo. Este conjunto de datos está disponible gracias a una competición de Kaggle (https://www.kaggle.com/c/titanic) donde disponemos de dos subconjuntos de datos (`train.csv` y `test.csv`). En este estudio vamos a hacer uso únicamente del conjunto de datos de entrenamiento ya que es donde tenemos disponible la variable objetivo de nuestros análisis, `Survived`, que nos indicará si el pasajero en cuestión sobrevivió al accidente o no. El conjunto de datos contiene las siguientes variables:

- **PassengerId**: Identificador único del pasajero.
- **Survived**: Es la variable objetivo de nuestros análisis. Indica si el pasajero sobrevivió al naufragio, codificada como 0 (no) y 1 (si).
- **Pclass**: Clase en la que viaja el pasajero: primera segunda o tercera (1, 2, 3)
- **Name**: Nombre del pasajero.
- **Sex**: Sexo del pasajero.
- **Age**: Edad del pasajero. Pueden tener números decimales, refiriéndose con ellos a los meses.
- **SibSp**: Número de los siguientes tipos de familiares que viajan en el barco con el pasajero:
  - Hermanos/as
  - Hermanastros/as
  - Marido/esposa (no se tiene en cuenta amantes o prometidas)
- **Parch**: Número de los siguientes tipos de familiares que viajan en el barco con el pasajero:
  - Padres/madres
  - Hijos/as
  - Hijastros/as
  - En el caso de que el niño viaje con una niñera, no se contará a la niñera como familiar.
- **Ticket**: Identificador del billete.
- **Fare**: Precio del billete.
- **Cabin**: Número del camarote.
- **Embarked**: Puerto de embarque del pasajero.

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

Vamos a empezar cargando el conjunto de datos y posteriormente realizaremos un primer análisis visual de los datos cargados y sus tipos.

```
# Carga del conjunto de datos
titanic_data <- read.csv('train.csv')

# Comprobamos que se carga correctamente
head(titanic_data)
```

```
##   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
```

```
# Dimensiones de los datos
dim(titanic_data)
```

```
## [1] 891  12
```

Comprobamos que la información se ha cargado correctamente y que el subconjunto de datos de entrenamiento tiene 891 registros y 12 variables, donde una de ellas es la variable objetivo `Survived`.

Vamos a ver a continuación los valores resumen de este conjunto de datos y los tipos con los que se han cargado las variables para empezar a intuir qué tipo de técnicas de preprocesado vamos a tener que aplicar.

```
# Valores resumen
summary(titanic_data)
```

```
##   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
##
```

```
# Tipos con los que se han cargado las variables
str(titanic_data)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

Como podemos ver en los valores resumen vamos a tener que realizar un tratamiento de valores ausentes, por ejemplo a la variable `Age` (177 observaciones con valor `NA`), además de analizar los valores extremos en otras variables con `Fare`. También podemos ver que algunas variables que parecen ser de tipo factor han sido cargadas con tipo texto o entero.

2.1 Selección de los datos interés

Las variables `Name` y `Ticket` no aportarán ninguna información de interés de cara a saber si un pasajero sobrevive o no. Además, la información que nos puede aportar `Ticket` ya la tenemos disponible en la variable `Fare`, ya que los tickets tienen asociados un precio, y la que nos puede aportar el apellido la tenemos de forma más completa en las variables `SibSp` y `Parch`.

```
# Eliminamos las variables que no vamos a utilizar
drop_data <- names(titanic_data) %in% c("Name", "Ticket")
titanic_data <- titanic_data[, !drop_data]
```

2.2 Conversión de variables categóricas a factor

Convertimos los atributos categóricos a tipo factor. La variable `PClass` fue clasificada como tipo `int` al leer el datase, pero en realidad solo toma tres valores (las tres clases antes comentadas).

```r
# Convertimos las variables categóricas a factor
titanic_data$Survived <- as.factor(titanic_data$Survived)
titanic_data$Pclass <- as.factor(titanic_data$Pclass)
titanic_data$Sex <- as.factor(titanic_data$Sex)
titanic_data$Embarked <- as.factor(titanic_data$Embarked)
```

3 Limpieza de los datos

En este apartado vamos realizar el preprocesado de los datos para tener el conjunto de datos listo para los análisis que realizaremos posteriormente. Empezaremos estudiando los valores ausentes y después analizaremos los valores extremos de algunas variables.

3.1 Valores ausentes

Empezamos mostrando el número de valores ausentes en cada uno de los atributos del conjunto de datos.

```
# Valores ausentes
colSums(is.na(titanic_data))
```

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

```
# Cadenas de caracteres vacías
colSums(titanic_data=="")
```

```
## PassengerId    Survived      Pclass         Sex         Age       SibSp
##           0           0           0           0          NA           0
##       Parch        Fare       Cabin    Embarked
##           0           0         687           2
```

Como podemos ver, la variable `Age` contiene 177 observaciones con valor `NA` y tanto la variable `Cabin` como `Embarked` tienen observaciones con cadena de caracteres vacías. En el atributo `Cabin` existen 687 cadenas de caracteres vacías y `Embarked` contiene 2.

Los datos perdidos encontrados en las variables comentadas se tratarán de forma diferente según las características de la variable o la cantidad de datos perdidos. Si las aproximaciones utilizadas para limpiar los datos no son correctas, podremos volver a este apartado para gestionar de forma diferente el tratamiento de los valores ausentes para intentar mejorar los modelos calculados.

Primero trataremos los datos vacíos de la variable `Embarked`. En este caso solo existen 2 elementos vacíos, por tanto, una primera posibilidad sería eliminar completamente esos dos registros, pero con el objetivo de no perder información, imputaremos estos dos registros con la moda estadística de la variable. Para ello creamos la función getmode() que nos devuelve el valor más frecuente de una variable.

```
# Función para calcular la moda
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Calculamos la moda de la variable Embarked
Embarked_mode <- getmode(titanic_data$Embarked)

# Imputamos la moda a las observaciones vacías
titanic_data$Embarked[titanic_data$Embarked == ""] <- Embarked_mode

# Eliminamos el nivel "" (vacío) de la variable factor
titanic_data$Embarked <- droplevels(titanic_data$Embarked)
```

En la variable `Cabin` el número de elementos vacíos es muy elevado, 687 de un total de 891 registros (77,1%). Debido a este alto porcentaje de elementos vacíos, una primera posibilidad sería sustituir las cadenas vacías por una constante como "Desconocido". Pero en lugar de ello optaremos por eliminar por completo la variable. Tiene sentido debido al elevado número de valores vacíos y a que el atributo `Pclass` nos puede dar una información similar de cara a saber si un pasajero sobrevive o no. Los camarotes de primera clase estarían situados cerca de cubierta, siendo la probabilidad de sobrevivir mayor que la de un pasajero que viaje en tercera clase y cuyo camarote se encuentre en una planta inferior.

```
# Eliminamos la variable Cabin
titanic_data <- titanic_data[, -9]
```

Por último, vamos a tratar los valores ausentes de la variable `Age`. Para imputar estos valores perdidos de las edades de los pasajeros dividiremos el conjunto de datos en seis grupos y le asignaremos el valor de la mediana de cada grupo. Los atributos elegidos para crear los grupos serán la clase en la que viaja el pasajero y el sexo. Como veremos en el apartado de análisis, estas dos variables serán importantes a la hora de decidir si un pasajero sobrevive o no.

Creamos con ayuda de la función *tapply()* la matriz con las medianas de las edades en los grupos comentados.

```
# Calculamos la mediana de Age por grupos de Pclass y Sex
Age_median_matrix <-
  tapply(titanic_data$Age, list(titanic_data$Pclass, titanic_data$Sex),
         median, na.rm = TRUE)
Age_median_matrix
```

```
##   female male
## 1   35.0   40
## 2   28.0   30
## 3   21.5   25
```

Se observa que hay diferencia entre las medianas de los diferentes grupos. Con esto corroboramos que ha sido una aproximación acertada hacerlo de esta manera.

Sustituimos los registros vacíos de edad con los valores correspondientes de la matriz de medianas calculadas.

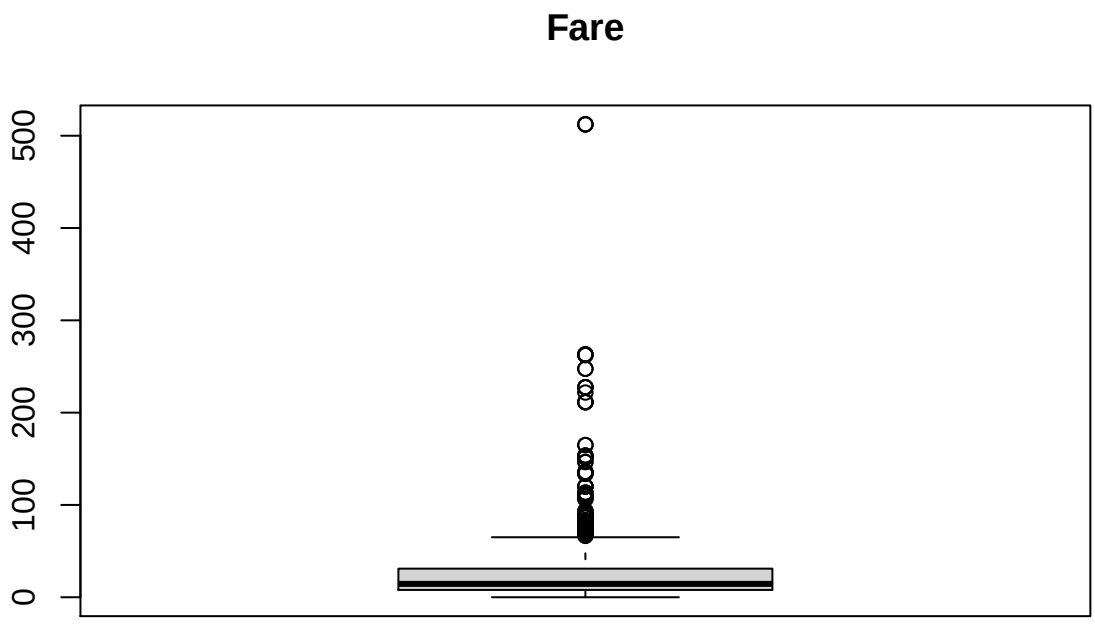
```
# Imputamos los valores ausentes de la variable Age
titanic_data$Age[is.na(titanic_data$Age) & titanic_data$Sex == "female" &
                   titanic_data$Pclass == "1"] <- Age_median_matrix[1,1]
titanic_data$Age[is.na(titanic_data$Age) & titanic_data$Sex == "female" &
                   titanic_data$Pclass == "2"] <- Age_median_matrix[2,1]
titanic_data$Age[is.na(titanic_data$Age) & titanic_data$Sex == "female" &
                   titanic_data$Pclass == "3"] <- Age_median_matrix[3,1]
titanic_data$Age[is.na(titanic_data$Age) & titanic_data$Sex == "male" &
                   titanic_data$Pclass == "1"] <- Age_median_matrix[1,2]
titanic_data$Age[is.na(titanic_data$Age) & titanic_data$Sex == "male" &
                   titanic_data$Pclass == "2"] <- Age_median_matrix[2,2]
titanic_data$Age[is.na(titanic_data$Age) & titanic_data$Sex == "male" &
                   titanic_data$Pclass == "3"] <- Age_median_matrix[3,2]
```

3.2 Valores extremos.

Mostramos los diagramas de cajas de las variables continuas del conjunto de datos (`Age` y `Fare`) para comprobar si existen valores extremos en ellas.

Comenzamos observando el atributo correspondiente al precio de los billetes.

```
# Diagrama de caja de Fare
boxplot(titanic_data$Fare, main = "Fare")
```



Utilizamos la función boxplot.stats() para obtener las estadísticas de forma numérica.

```
boxplot.stats(titanic_data$Fare)$stats
```

```
## [1]  0.0000  7.9104 14.4542 31.0000 65.0000
```

Existen valores extremos por encima del bigote superior del boxplot (65). La distribución de esta variable está muy desplazada a la izquierda. Esto es debido a que el 50% de los datos se encuentran entre 7,89 y 31,27. Pero se vendieron billetes muy por encima de esos precios. Esta desigualdad en los precios es plausible, ya que al tratarse del viaje inaugural de un transatlántico de lujo, pudo haber diferentes paquetes de precios con diferentes características. Los paquetes más lujosos estarían destinados para los pasajeros más selectos. Un ejemplo sería el billete con un precio superior a los 500$. Se diferencian mucho del resto, pero pueden corresponder a la suit más lujosa del barco.

Vemos cuántos ouliers presenta esta variable.

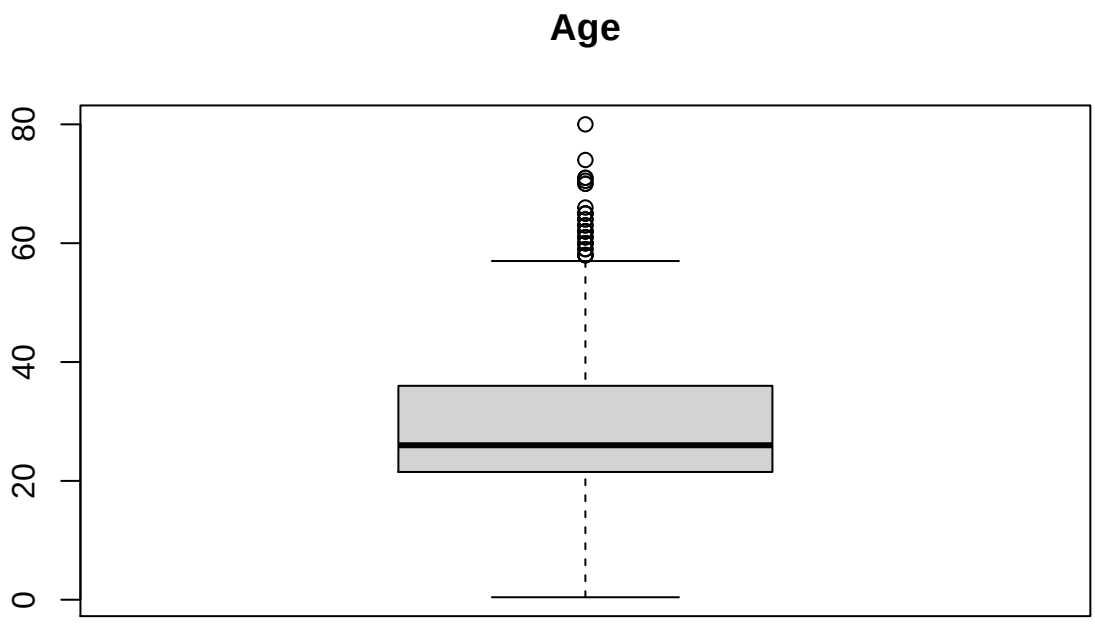
```
length(titanic_data$Fare[titanic_data$Fare > 65])
```

```
## [1] 116
```

Una posible opción sería tratar estos 166 valores con alguna técnica de imputación como hemos hecho en el apartado anterior. Pero como acabamos de comentar, estos valores extremos se consideran válidos. Por lo que optaremos por dejar la variable sin tratar para no perder información y no cambiar su distribución.

Analizamos a continuación los valores extremos del atributo correspondiente a la edad de los pasajeros.

```
# Diagrama de caja de Age
boxplot(titanic_data$Age, main = "Age")
```

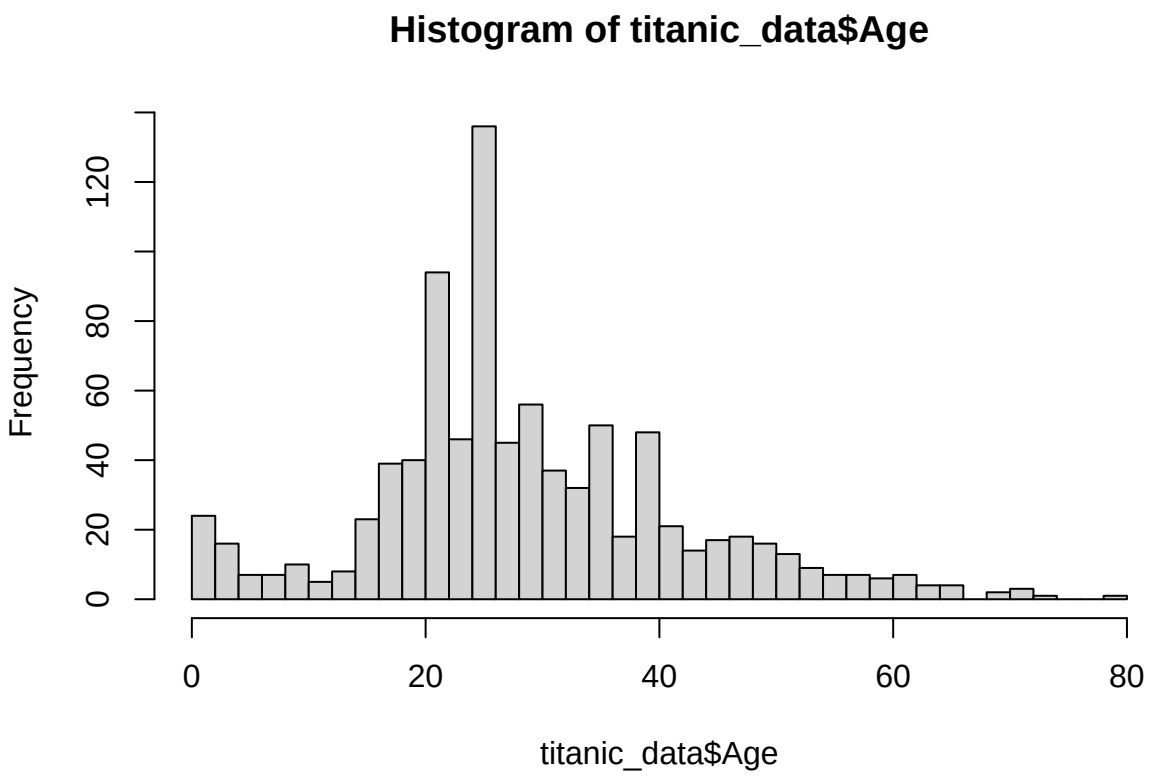


Se vuelven a observar valores extremos por encima del bigotes superior del gráfico. Como en el caso anterior es totalmente posible que viajen personas por encima de 60 años. Por esta razón y para no perder información valiosa, dejaremos la variable sin modificar en el conjunto de datos.

Aunque en este caso, crearemos un nuevo atributo discretizado con tres niveles a partir de la variable edad. Los niveles corresponderán a niños, adultos y ancianos. Ya que son tres grupos de edad que nos puede interesar estudiar a la hora de saber si un pasajero sobrevive o no.

Antes de realizar la discretización, mostramos el histograma de la variable para comprender mejor cómo se distribuye.

```
# Histograma de la variable Age
hist(titanic_data$Age,breaks = 40)
```



Lo primero que se observa es que gran parte de los datos se concentran entre los 20 y los 40 años, lo que nos indica que los tres niveles que queremos construir tendrán un número de registros muy desigual. Pero igualmente nuestro objetivo con la discretización es estudiar si se cumple la máxima "Mujeres, niños y ancianos primero" cuando se produce una accidente marítimo.

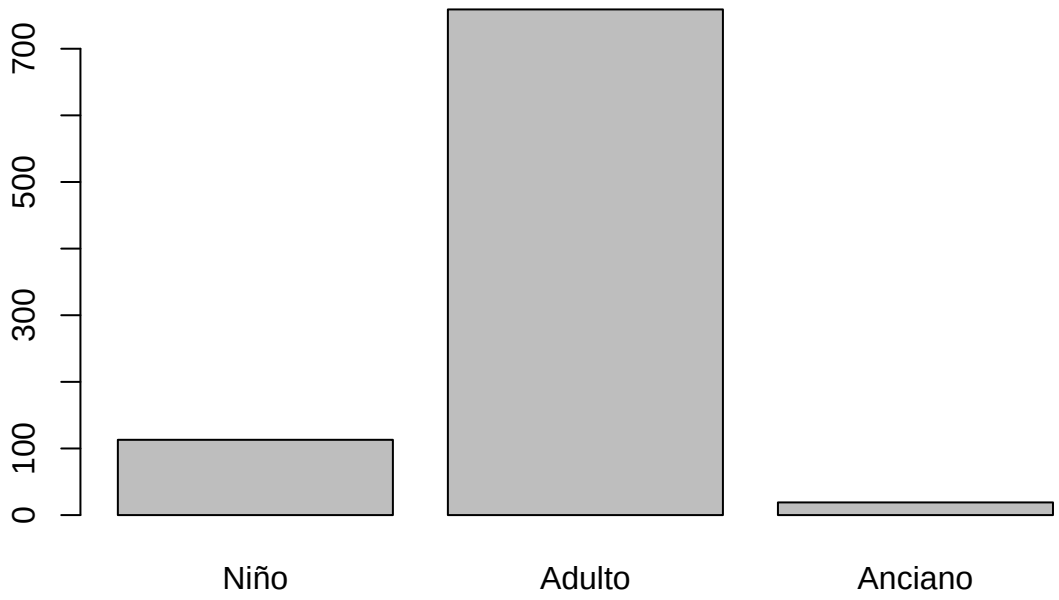
La división de los tres niveles se realizará de la siguiente manera:

- Niño: [0,17]. Si el pasajero es menor de edad, se considera niño.
- Adulto: [18,60]
- Anciano: [61,80]. Podemos considerar ancianos a partir de los sesenta años, ya que el accidente tuvo lugar a principios del siglo XX.

Para ello utilizamos la función *cut()*, activando la opción *ordered_result*, ya que existe un orden de menor a mayor entre los niveles establecidos.

```
# Discretización de la variable Age
titanic_data$Age_d <- cut(titanic_data$Age, breaks = c(0,17,61,80),
                          ordered_result = TRUE,
                          labels = c("Niño", "Adulto", "Anciano"))

# Mostramos la discretización
plot(titanic_data$Age_d)
```



3.3 Generar archivo con los datos preprocesados

```
# Exportamos el conjunto de datos preprocesado a un archivo CSV
write.csv(titanic_data, file = "titanic_clean.csv", row.names = FALSE)
```

4 Análisis de los datos

Ahora vamos a realizar diferentes análisis de los datos, haciendo una primera planificación de los grupos de datos y de las pruebas estadísticas que vamos a realizar para ver la relación entre las variables del conjunto de datos.

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

Durante este apartado vamos a realizar diferentes pruebas estadísticas sobre las variables de interés del conjunto de datos. Empezaremos haciendo un análisis de la normalidad y la homocedasticidad de las variables cuantitativas del conjunto de datos (`Age`, `SibSp`, `Parch` y `Fare`). Con los resultados de estos análisis podremos saber qué tipos de pruebas estadísticas (paramétricas o no paramétricas) vamos a poder realizar en los contrastes de hipótesis. Gracias a estos contrastes de hipótesis podremos analizar las relaciones entre las variables del conjunto de datos y más concretamente con la variable objetivo `Survived`, para saber qué variables tuvieron más influencia en el hecho de sobrevivir o no al accidente. También haremos un análisis de correlaciones de las variables cuantitativas para comprobar si existe relación entre ellas.

Una vez estudiadas las relaciones entre las variables mediante los contrastes de hipótesis y las correlaciones calcularemos diferentes modelos de regresión logística para comprobar qué regresores son realmente significativos y que efecto tuvieron sobre la variable dependiente dicotómica `Survived`. Con este modelo podremos dar respuesta a diferentes preguntas que nos podemos plantear como son las siguientes:

- ¿Los pasajeros de primera clase tuvieron más probabilidad de sobrevivir que los de tercera clase?
- ¿Se dio preferencia a las mujeres antes que a los hombres para ser salvados?
- ¿Influyó el puerto de embarque de cada pasajero en el hecho de que sobreviviera al accidente?

Finalmente calcularemos unos árboles de decisión (modelos supervisados) para predecir la variable objetivo `Survived` a partir del resto de variables del conjunto de datos, y analizaremos la precisión de los modelos calculados.

4.2 Comprobación de la normalidad y homogeneidad de la varianza

4.2.1 Normalidad

Para comprobar la normalidad de las variable cuantitativas vamos a utilizar el test de *Shapiro-Wilk*, ya que se considera uno de los métodos más potentes para contrastar la normalidad. Este método se basa en el contraste de hipótesis, asumiendo como hipótesis nula que la población sigue una distribución normal. Por tanto, nos basaremos en el *p-valor* para determinar si aceptamos o rechazamos la hipótesis nula de normalidad.

```
# Test de normalidad de las variables cuantitativas
shapiro.test(titanic_data$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic_data$Age
## W = 0.96548, p-value = 1.118e-13
```

```
shapiro.test(titanic_data$SibSp)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  titanic_data$SibSp
## W = 0.51297, p-value < 2.2e-16
```

```
shapiro.test(titanic_data$Parch)
```

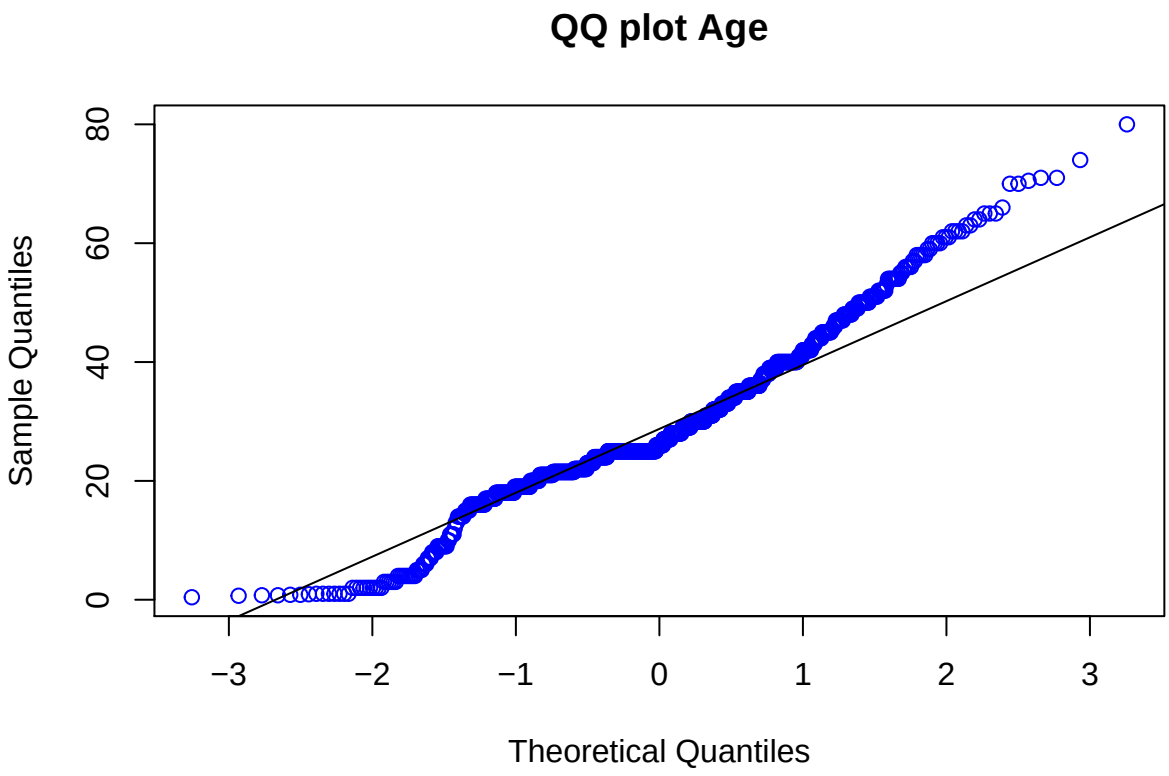
```
##
##  Shapiro-Wilk normality test
##
## data:  titanic_data$Parch
## W = 0.53281, p-value < 2.2e-16
```

```
shapiro.test(titanic_data$Fare)
```

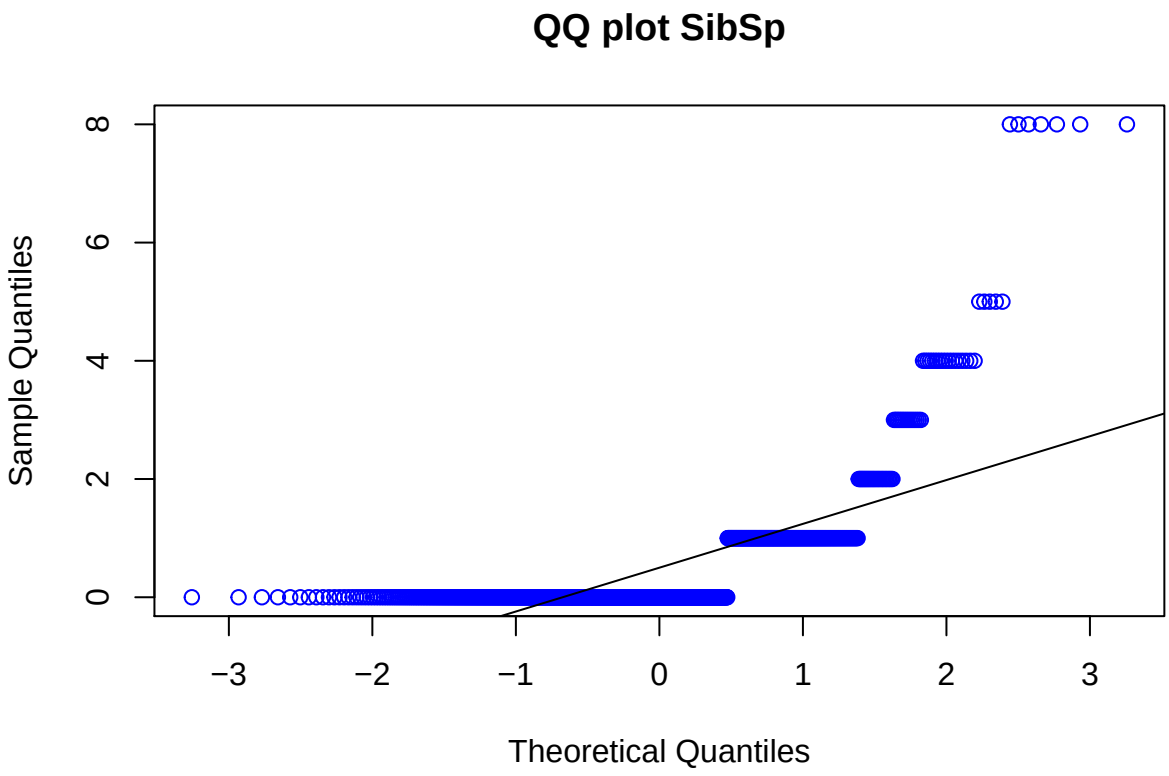
```
##
##  Shapiro-Wilk normality test
##
## data:  titanic_data$Fare
## W = 0.52189, p-value < 2.2e-16
```

Si usamos un nivel de significancia de $\alpha = 0.05$, podemos ver que en todos los tests anteriores se rechaza la hipótesis nula con un nivel del confianza del 95%, ya que $p\_valor < 0.05$ en todos los casos. Por tanto, podemos decir que las variables cuantitativas de este conjunto de datos (`Age`, `SibSp`, `Parch` y `Fare`) no siguen una distribución normal. Vamos a comparar las distribuciones de estas variables con la de una normal de forma visual mediante un gráfico QQ.

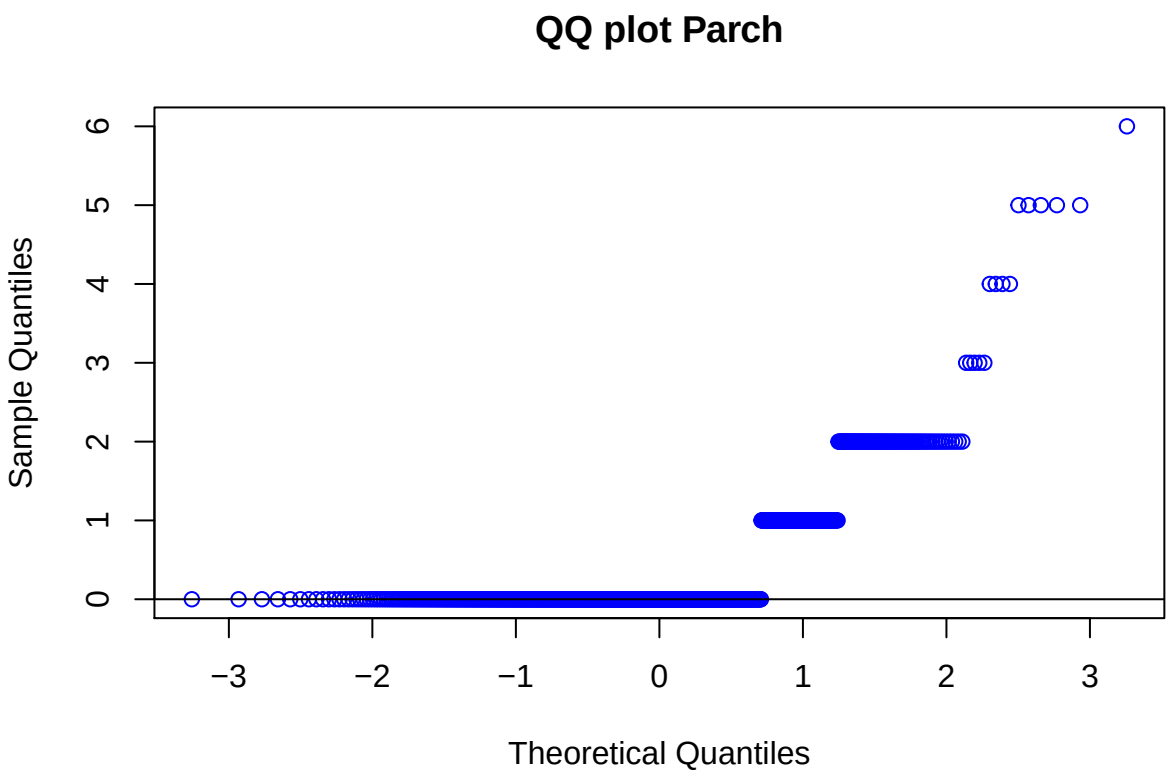
```
# Gráficos QQ de las variables cuantitativas
qqnorm(titanic_data$Age, main = "QQ plot Age", col = 'blue')
qqline(titanic_data$Age)
```



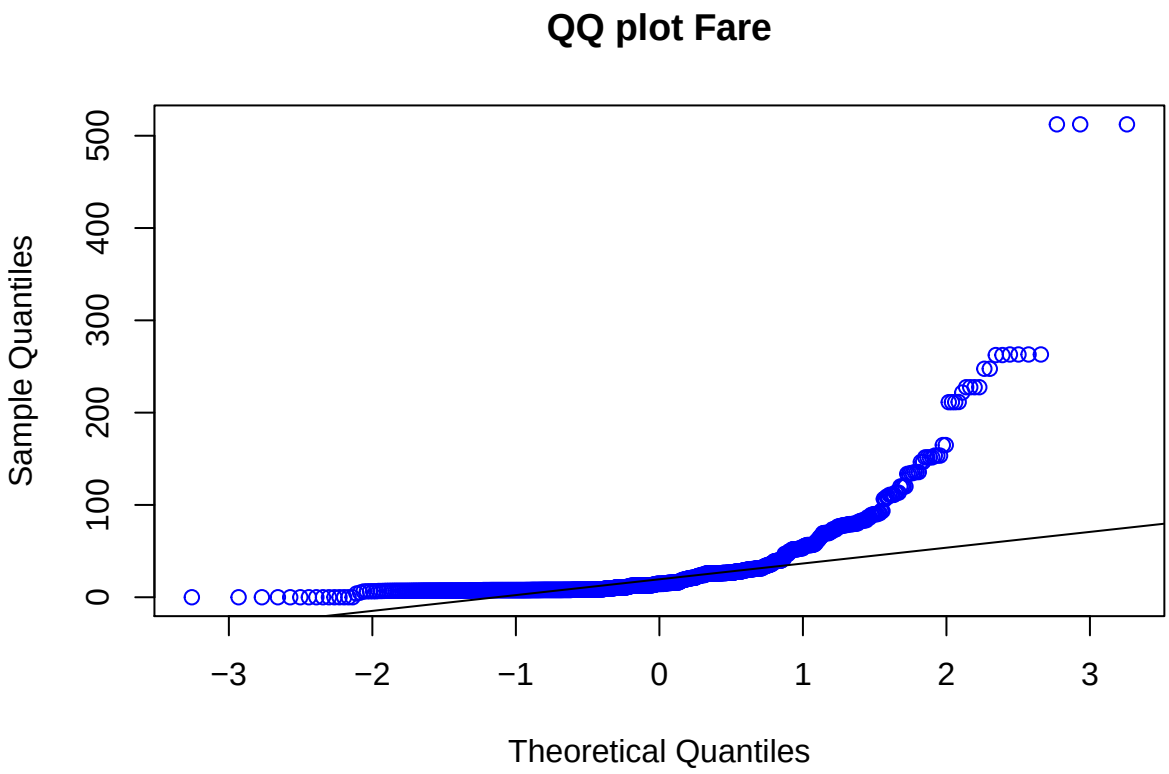

```
qqnorm(titanic_data$SibSp, main = "QQ plot SibSp", col = 'blue')
qqline(titanic_data$SibSp)
```




```
qqnorm(titanic_data$Parch, main = "QQ plot Parch", col = 'blue')
qqline(titanic_data$Parch)
```




```
qqnorm(titanic_data$Fare, main = "QQ plot Fare", col = 'blue')
qqline(titanic_data$Fare)
```



Comprobamos que, como hemos obtenido con los test de *Shapiro-Wilk*, las variables cuantitativas no siguen una distribución normal. Vemos que la variable `Age` se encuentra cerca de la normalidad, por tanto, como tenemos una cantidad de observaciones suficientemente grande podemos asumir que esta variable sigue una distribución normal basándonos en el *teorema central del límite*.

4.2.2 Homocedasticidad

Ahora vamos a comprobar la homocedasticidad para las variables anteriores diferenciando en distintos grupos. Para la variable `Age`, como hemos supuesto normalidad, usaremos el test de *Levene* (paramétrico) y para el resto de variables usaremos el test de *Fligner-Killeen* (no paramétrico). Ambas pruebas realizan un contraste de hipótesis donde la hipótesis nula asume igualdad de varianzas en los diferentes grupos de datos.

```
# Comprobación de la homocedasticidad para Age
leveneTest(Age ~ Pclass, data = titanic_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2  16.561 8.684e-08 ***
##       888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(Age ~ Sex, data = titanic_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.4543 0.5005
##       889
```

```
leveneTest(Age ~ Embarked, data = titanic_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2  8.6963 0.0001819 ***
##       888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(Age ~ Survived, data = titanic_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  5.2188 0.02258 *
##       889
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De los resultados de los tests de Levene podemos decir que la variable `Age` presenta heterocedasticidad con las variables `Pclass`, `Embarked` y `Survived`, y homocedasticidad con la variable `Sex`. Por tanto, la variable `Age` tendrá varianzas iguales entre los pasajeros de distintos sexo, en cambio, tendrás distinta varianza entre los pasajeros que sobrevivieron y los que no, entre las distintas clases y entre los distintos puertos de embarque.

Para comprobar la homocedasticidad en las variables `SibSp` y `Parch` vamos a sumarlas y usar una variable que indique el número de familiares que iban a bordo del Titanic.

```
# Comprobación de la homocedasticidad para SibSp+Parch
fligner.test(SibSp+Parch ~ Pclass, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SibSp + Parch by Pclass
## Fligner-Killeen:med chi-squared = 0.041221, df = 2, p-value = 0.9796
```

```
fligner.test(SibSp+Parch ~ Sex, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SibSp + Parch by Sex
## Fligner-Killeen:med chi-squared = 58.958, df = 1, p-value = 1.61e-14
```

```
fligner.test(SibSp+Parch ~ Embarked, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SibSp + Parch by Embarked
## Fligner-Killeen:med chi-squared = 5.0719, df = 2, p-value = 0.07919
```

```
fligner.test(SibSp+Parch ~ Survived, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SibSp + Parch by Survived
## Fligner-Killeen:med chi-squared = 19.647, df = 1, p-value = 9.317e-06
```

Como podemos ver el tamaño de la familia tendrá varianzas iguales (homocedasticidad) en las diferentes clases (`Pclass`) y en el puerto de embarque (`Embarked`). En cambio, presenta varianzas distintas (heterocedasticidad) entre hombres y mujeres (`Sex`) y entre los pasajeros que sobrevivieron y los que no (`Survived`).

Para terminar con la comprobación de la homocedasticidad vamos a hacerlo para la variable `Fare`.

```
# Comprobación de la homocedasticidad para Fare
fligner.test(Fare ~ Pclass, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Pclass
## Fligner-Killeen:med chi-squared = 365.8, df = 2, p-value < 2.2e-16
```

```
fligner.test(Fare ~ Sex, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Sex
## Fligner-Killeen:med chi-squared = 55.949, df = 1, p-value = 7.436e-14
```

```
fligner.test(Fare ~ Embarked, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Embarked
## Fligner-Killeen:med chi-squared = 133.23, df = 2, p-value < 2.2e-16
```

```
fligner.test(Fare ~ Survived, data = titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Survived
## Fligner-Killeen:med chi-squared = 96.253, df = 1, p-value < 2.2e-16
```

Vemos que la variable `Fare` presenta heterocedasticidad entre las distintas clases (`Pclass`), entre hombre y mujeres (`Sex`), entre los distintos puertos de embarque (`Embarked`) y entre los pasajeros que sobrevivieron y los que no (`Survived`).

4.3 Aplicación de pruebas estadísticas

4.3.1 Contraste de hipótesis sobre Survived

A continuación vamos a ver la relación de la variable `Survived` con el resto de variables que estamos analizando, tanto cuantitativas como categóricas. Hemos visto que la variable `Age`, aunque es la única para la que se ha asumido normalidad, presenta heterocedasticidad con los distintos grupos de la variable `Survived`. Por tanto, para todas las variables cuantitativas tendremos que utilizar pruebas no paramétricas como Wilcoxon (datos dependientes) o Mann-Whitney (datos independientes). Estas dos pruebas se aplican indistintamente con la función `wilcox.test()`.

```
# Relación de Survived con las variable cuantitativas
wilcox.test(Age ~ Survived, data = titanic_data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  Age by Survived
## W = 98172, p-value = 0.25
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(SibSp+Parch ~ Survived, data = titanic_data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  SibSp + Parch by Survived
## W = 77659, p-value = 7.971e-07
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(Fare ~ Survived, data = titanic_data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Fare by Survived
## W = 57807, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

De acuerdo a los resultados anteriores podemos decir que se observan diferencias estadísticamente significativas de los pasajeros que sobrevivieron y los que no en el tamaño de la familia (`SibSp`+`Parch`) y en el precio del ticket (`Fare`). En cambio, no se presentan diferencias significativas en la edad (`Age`).

Para comparar si existen diferencias significativas en las variable categóricas entre los pasajeros que sobrevivieron y los que no vamos a aplicar el test de χ^2.

```
# Relación de Survived con las variable categóricas
table.Pclass <- table(titanic_data$Pclass, titanic_data$Survived)
table.Pclass
```

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

```
chisq.test(table.Pclass)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table.Pclass
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

```
table.Sex <- table(titanic_data$Sex, titanic_data$Survived)
table.Sex
```

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

```
chisq.test(table.Sex)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data:  table.Sex
## X-squared = 260.72, df = 1, p-value < 2.2e-16
```

```
table.Embarked <- table(titanic_data$Embarked, titanic_data$Survived)
table.Embarked
```

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

```
chisq.test(table.Embarked)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table.Embarked
## X-squared = 25.964, df = 2, p-value = 2.301e-06
```

Con estos resultados comprobamos que se rechaza la hipótesis nula para los tres casos, por tanto, observamos que tanto el sexo (`Sex`), como la clase (`Pclass`) y el puerto de embarque (`Embarked`) tuvieron repercusión en si un pasajero finalmente sobrevivió al accidente o no.

4.3.2 Análisis de correlaciones

Vamos a calcular los coeficientes de correlación entre las variables cuantitativas para comprobar cuales están relacionadas linealmente entre sí. Para ello vamos a utilizar la correlación de *Spearman* (no paramétrica) para medir el grados de dependencia entre las variables.

```
# Correlación de Spearman entre variables cuantitativas
cor.test(titanic_data$Age, titanic_data$SibSp, method = 'spearman')
```

```
## Warning in cor.test.default(titanic_data$Age, titanic_data$SibSp, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  titanic_data$Age and titanic_data$SibSp
## S = 137482216, p-value = 6.096e-07
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.166179
```

```
cor.test(titanic_data$Age, titanic_data$Parch, method = 'spearman')
```

```
## Warning in cor.test.default(titanic_data$Age, titanic_data$Parch, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  titanic_data$Age and titanic_data$Parch
## S = 145651575, p-value = 1.085e-12
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
```

```
##        rho
## -0.2354747
```

```
cor.test(titanic_data$Age, titanic_data$Fare, method = 'spearman')
```

```
## Warning in cor.test.default(titanic_data$Age, titanic_data$Fare, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  titanic_data$Age and titanic_data$Fare
## S = 97329585, p-value = 1.614e-07
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1744117
```

```
cor.test(titanic_data$SibSp, titanic_data$Parch, method = 'spearman')
```

```
## Warning in cor.test.default(titanic_data$SibSp, titanic_data$Parch, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  titanic_data$SibSp and titanic_data$Parch
## S = 64838502, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.450014
```

```
cor.test(titanic_data$SibSp, titanic_data$Fare, method = 'spearman')
```

```
## Warning in cor.test.default(titanic_data$SibSp, titanic_data$Fare, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  titanic_data$SibSp and titanic_data$Fare
## S = 65180502, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.447113
```

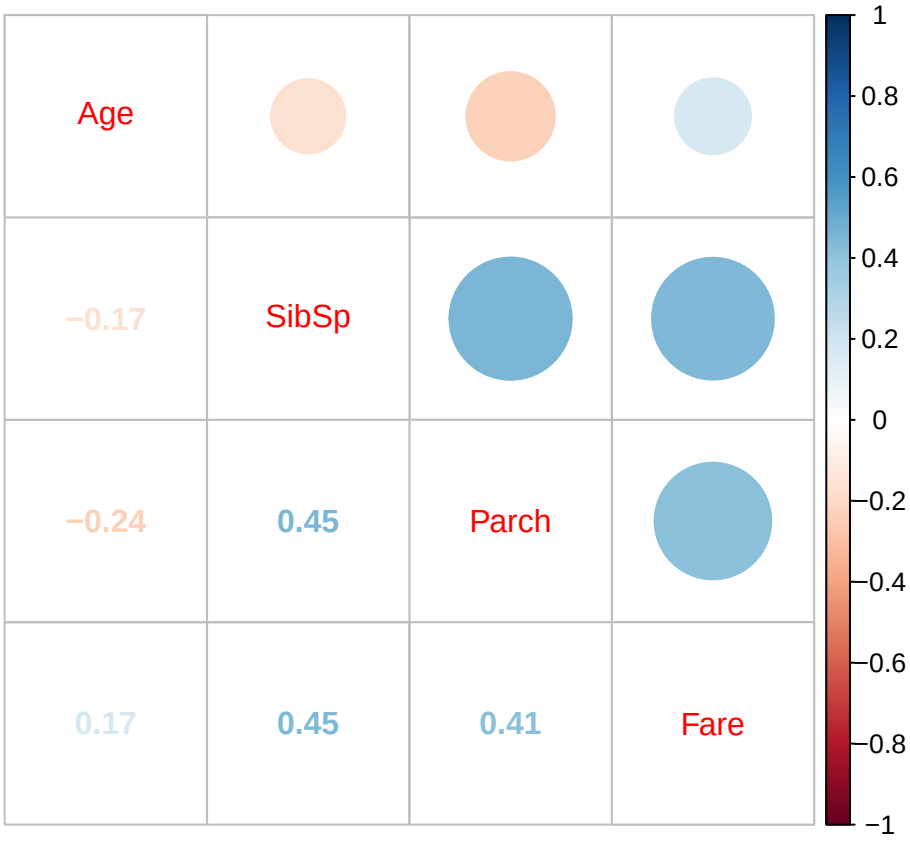
```
cor.test(titanic_data$Parch, titanic_data$Fare, method = 'spearman')
```

```
## Warning in cor.test.default(titanic_data$Parch, titanic_data$Fare, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  titanic_data$Parch and titanic_data$Fare
## S = 69547095, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
```

```
## sample estimates:
##       rho
## 0.4100738
```

```
corrplot.mixed(cor(titanic_data[, c('Age', 'SibSp', 'Parch', 'Fare')],
               method = 'spearman'))
```



Como podemos ver, ninguna de las variable está relacionada con las demás ya que el coeficiente de correlación es menor de |0.5| en todos los casos. Por tanto, podemos decir que las variables cuantitativas son independientes entre ellas. Además, el *p-valor* nos indica que los resultados obtenidos son significativos.

4.3.3 Modelo supervisado

Preparación de los datos de entrenamiento y test

Como se comentó en el primer apartado, los datos originales obtenidos de Kaggle ya estaban divididos en datos de entrenamiento y test. El dataset de entrenamiento es el único que contiene la variable objetivo `Survived`. Por eso vamos a trabajar únicamente con este conjunto de datos para poder observar la precisión del modelo creado.

Por tanto, dividimos el conjunto de datos de entrenamiento en dos: 2/3 de los datos se enfocarán a entrenamiento y 1/3 para test. Se realiza la división mediante el método de exclusión con partición estratificada para que la proporción de pasajeros que sobreviven frente a los que no sea similar en ambos conjunto de datos. Utilizamos la función *holdout* de la librería *rminer*.

```
# Hacemos uso de seed para que el resultado sea reproducible
set.seed(10)

# Método de exclusión para particionar el conjunto de datos
```

```
h <- holdout(titanic_data$Survived, ratio=2/3, mode="stratified")
titanic_train <- titanic_data[h$tr, ]
titanic_test <- titanic_data[h$ts, ]
```

Comprobamos que la variable objetivo se ha repartido de manera similar en ambos conjuntos de datos.

```
# Repartición de la variable Survived en ambos subconjuntos
prop.table(table(titanic_train$Survived))
```

```
##
##         0         1
## 0.6161616 0.3838384
```

```
prop.table(table(titanic_test$Survived))
```

```
##
##         0         1
## 0.6161616 0.3838384
```

Como podemos ver la proporción es exactamente la misma en ambos conjuntos de datos.

Regresión logística

A continuación vamos a calcular un modelo de regresión logística para comprobar qué variables tienen un efecto significativo sobre la variable dicotómica dependiente `Survived`, es decir, sobre la probabilidad de sobrevivir o no al accidente.

```
# Regresión logística con todas las variables
regression.logist <- glm(Survived ~ Pclass+Sex+Age+SibSp+Parch+Fare+Embarked,
                         data = titanic_train, family = binomial(link="logit"))
summary(regression.logist)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +
##     Fare + Embarked, family = binomial(link = "logit"), data = titanic_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7923  -0.5452  -0.3488   0.5490   2.4335
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.786466   0.625916   7.647 2.05e-14 ***
## Pclass2     -0.957479   0.375535  -2.550   0.0108 *
## Pclass3     -2.744825   0.398919  -6.881 5.96e-12 ***
## Sexmale     -2.902672   0.262021 -11.078  < 2e-16 ***
## Age         -0.049980   0.010257  -4.873 1.10e-06 ***
## SibSp       -0.316166   0.144471  -2.188   0.0286 *
## Parch       -0.113139   0.169513  -0.667   0.5045
## Fare         0.001162   0.003217   0.361   0.7179
## EmbarkedQ    0.131253   0.484712   0.271   0.7866
## EmbarkedS   -0.442253   0.301528  -1.467   0.1425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 791.10  on 593  degrees of freedom
## Residual deviance: 484.44  on 584  degrees of freedom
## AIC: 504.44
##
## Number of Fisher Scoring iterations: 5
```

Como podemos ver en el modelo resultante, las variables independientes (regresores) más significativas son `Pclass`, `Sex`, `Age` y `SibSp`. Por tanto, vamos a utilizar estas variables para ir construyendo diferentes modelos de regresión logística y analizaremos la bondad de cada modelo con la medida AIC.

```r
# Regresión logística con las variables independientes más significativas
regression.1var <- glm(Survived ~ Pclass, data = titanic_train,
                       family = binomial(link="logit"))

regression.2var.1 <- glm(Survived ~ Pclass+Sex, data = titanic_train,
                         family = binomial(link="logit"))
regression.2var.2 <- glm(Survived ~ Pclass+Age, data = titanic_train,
                         family = binomial(link="logit"))
regression.2var.3 <- glm(Survived ~ Pclass+SibSp, data = titanic_train,
                         family = binomial(link="logit"))
regression.2var.4 <- glm(Survived ~ Sex+Age, data = titanic_train,
                         family = binomial(link="logit"))
regression.2var.5 <- glm(Survived ~ Sex+SibSp, data = titanic_train,
                         family = binomial(link="logit"))
regression.2var.6 <- glm(Survived ~ Age+SibSp, data = titanic_train,
                         family = binomial(link="logit"))

regression.3var.1 <- glm(Survived ~ Pclass+Sex+Age, data = titanic_train,
                         family = binomial(link="logit"))
regression.3var.2 <- glm(Survived ~ Pclass+Sex+SibSp, data = titanic_train,
                         family = binomial(link="logit"))
regression.3var.3 <- glm(Survived ~ Pclass+Age+SibSp, data = titanic_train,
                         family = binomial(link="logit"))
regression.3var.4 <- glm(Survived ~ Sex+Age+SibSp, data = titanic_train,
                         family = binomial(link="logit"))

regression.4var <- glm(Survived ~ Pclass+Sex+Age+SibSp, data = titanic_train,
                       family = binomial(link="logit"))

out <- data.frame("Regression model" = c("regression.1var", "regression.2var.1",
                                         "regression.2var.2", "regression.2var.3",
                                         "regression.2var.4", "regression.2var.5",
                                         "regression.2var.6", "regression.3var.1",
                                         "regression.3var.2", "regression.3var.3",
                                         "regression.3var.4", "regression.4var"),
                  "AIC" = c(regression.1var$aic, regression.2var.1$aic,
                            regression.2var.2$aic, regression.2var.3$aic,
                            regression.2var.4$aic, regression.2var.5$aic,
                            regression.2var.6$aic, regression.3var.1$aic,
                            regression.3var.2$aic, regression.3var.3$aic,
                            regression.3var.4$aic, regression.4var$aic),
                  check.names = FALSE)
out %>% kable() %>% kable_styling(latex_options = 'hold_position')
```

Para evaluar los modelos usando el AIC debemos saber que cuanto menor es el valor de esta medida

Regression model	AIC
regression.1var	717.1000
regression.2var.1	528.1980
regression.2var.2	676.3620
regression.2var.3	719.0904
regression.2var.4	598.0800
regression.2var.5	590.9483
regression.2var.6	790.7676
regression.3var.1	507.8139
regression.3var.2	526.3419
regression.3var.3	676.1593
regression.3var.4	591.3871
regression.4var	500.6476

mejor será el modelo. Como podemos ver en la tabla, el modelo `regression.2var.1`, que utiliza como regresores las variables `Pclass` y `Sex`, obtiene un AIC (528.20) bastante bueno comparando con el resto de modelos que hemos calculado. También vemos que si a ese mismo modelo le añadimos el regresor `Age` conseguimos mejorarlo y obtener un $AIC = 507.81$ (el mejor modelo de los que utilizan tres regresores). Finalmente, vemos que el mejor modelo de regresión logística es el que utiliza todos los regresores significativos (`Pclass`, `Sex`, `Age` y `SibSp`), obteniendo un $AIC = 500.65$, aunque la mejora no es muy grande respecto al modelo `regression.3var.1`. Esto último se debe a que, como vimos en el modelo con todas las variables independientes, el regresor `SibSp` se considera significativo pero su nivel de significancia es menor que el del resto de regresores.

Vamos a ver a continuación los coeficientes que se obtienen con el mejor modelo y si los regresores siguen siendo significativos. También calcularemos los *odd-ratios* para analizar cómo afectan a la variable dependiente.

```
# Mejor modelo de regresión logística
summary(regression.4var)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp, family = binomial(link = "logit"),
##     data = titanic_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8863  -0.5278  -0.3492   0.5502   2.5226
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.62306    0.52683   8.775  < 2e-16 ***
## Pclass2     -1.15939    0.33235  -3.488 0.000486 ***
## Pclass3     -2.82243    0.33456  -8.436  < 2e-16 ***
## Sexmale     -2.93653    0.25400 -11.561  < 2e-16 ***
## Age         -0.05031    0.01013  -4.967  6.8e-07 ***
## SibSp       -0.37281    0.13633  -2.735 0.006245 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 791.10  on 593  degrees of freedom
## Residual deviance: 488.65  on 588  degrees of freedom
```

```
## AIC: 500.65
##
## Number of Fisher Scoring iterations: 5
```

```
# Cálculo de los odd-ratios
exp(coefficients(regression.4var))
```

```
##  (Intercept)      Pclass2      Pclass3      Sexmale          Age        SibSp
## 101.80549554   0.31367654   0.05946114   0.05304953   0.95093312   0.68879380
```

Vemos que los regresores presentan un *p-valor* que indica que siguen siendo estadísticamente significativos para el modelo. Si nos fijamos en los *odd-ratios* obtenidos a partir de los coeficientes vemos que la variable `Age` tiene un $OR = 0.95$, por tanto, el *odds* de que un pasajero sobreviva es 0.95 veces mayor por cada año de edad. También podemos ver que viajar en segunda y tercera clase influye negativamente en la probabilidad de sobrevivir (coeficientes con signo negativo, -1.16 y -2.82, respectivamente), al igual que sucede con los pasajeros de sexo masculino (-2.94). Por último, la variable `SibSp` (tiene hermanos/as y marido/esposa en el barco) tiene un $OR = 0.69$, por lo que el *odds* de que un pasajero sobreviva es 0.69 veces mayor por cada familiar que tenga en el barco.

Para terminar con el modelo de regresión logística vamos a realizar la predicción de la variable dependiente `Survived` para el subconjunto de test que hemos calculado anteriormente y compararemos con los valores reales de las observaciones para obtener la matriz de confusión. Con este modelo vamos a obtener una probabilidad de que el pasajero sobreviva o no, por tanto, vamos a establecer un umbral del 50% (0.5) para decidir el valor de la variable objetivo.

```
# Predicción modelo de regresión logística
prediction.regres <- predict(regression.4var, newdata = titanic_test,
                             type = 'response')
predictions <- ifelse(test = prediction.regres > 0.5, yes = 1, no = 0)

# Matriz de confusión
table(predictions, titanic_test$Survived,
      dnn = c("Predictions", "Observations"))
```

```
##            Observations
## Predictions   0   1
##           0 158  39
##           1  25  75
```

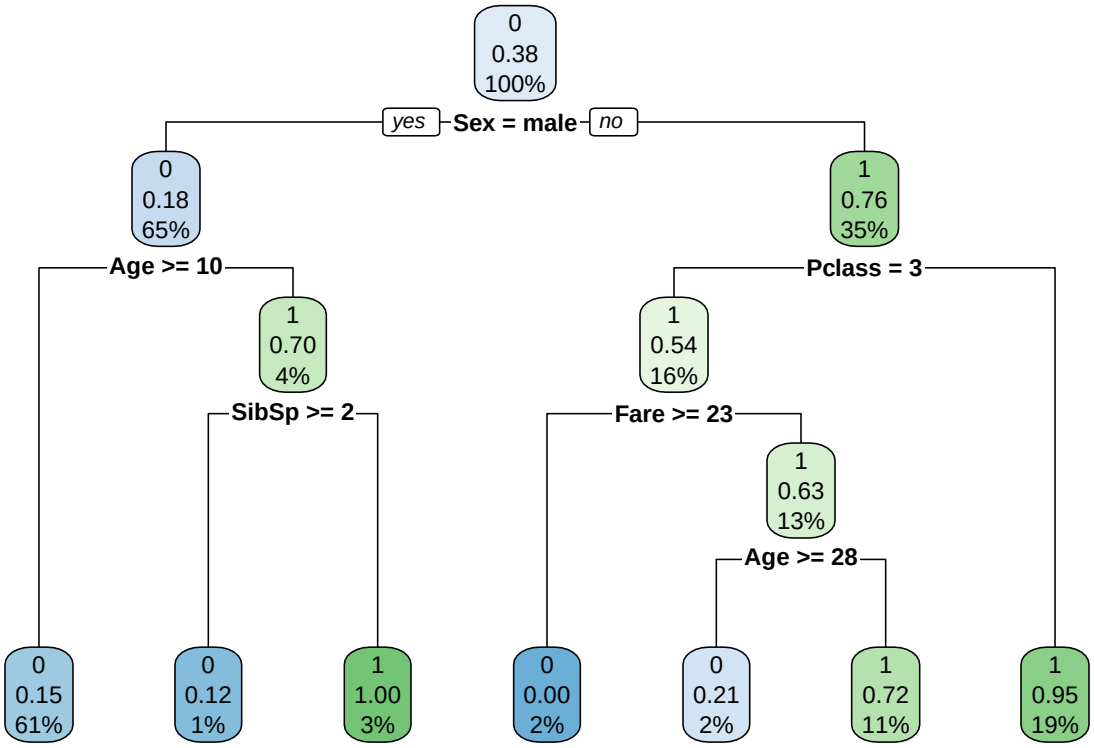
Comprobamos que el modelo de regresión logística que hemos utilizado para predecir la variable `Survived` en el subconjunto de test ha clasificado correctamente un 78.45% de las observaciones (53.20% de verdaderos negativos y 25.25% de verdaderos positivos). La sensibilidad de este modelo es del 65.79%, la especificidad es del 86.34% y la precisión es del 75%.

Árbol de decisión

Vamos a construir un árbol de decisión utilizando todas las variables mediante la función *rpart()*.

```
# Calculamos el árbol de decisión con todas las variables
set.seed(20)
mod_dt <- rpart(Survived ~ ., data = titanic_train, method = 'class')

# Mostramos el árbol de decisión generado
rpart.plot(mod_dt)
```



Como podemos ver las variables que se han utilizado para generar el árbol de decisión son `Age`, `SibSp`, `Fare`, `Pclass` y `Sex`. Para evaluar la calidad del modelo vamos a predecir la variable objetivo para el subconjunto de tests y calcularemos la matriz de confusión para esas observaciones.

```
# Predicción del árbol de decisión
set.seed(30)
pred_dt <- predict(mod_dt, newdata = titanic_test, type = "class")

# Matriz de confusión
confusionMatrix(pred_dt, titanic_test$Survived, positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 165  45
##          1  18  69
##
##                Accuracy : 0.7879
##                  95% CI : (0.7369, 0.833)
##     No Information Rate : 0.6162
##     P-Value [Acc > NIR] : 1.721e-10
##
##                   Kappa : 0.5306
##
##  Mcnemar's Test P-Value : 0.001054
##
```

```
##             Sensitivity : 0.6053
##             Specificity : 0.9016
##          Pos Pred Value : 0.7931
##          Neg Pred Value : 0.7857
##              Prevalence : 0.3838
##          Detection Rate : 0.2323
##    Detection Prevalence : 0.2929
##       Balanced Accuracy : 0.7535
##
##        'Positive' Class : 1
##
```

La precisión obtenida por el modelo es del 78,79%, por tanto, el 78,79% de los registros han sido correctamente clasificados.

La función *confusionMatrix()* también nos facilita otros valores como la sensibilidad (tasa de verdaderos positivos) o la especificidad (tasa de verdaderos negativos). Estos valores son 60,53% y 90,16%, respectivamente. Es decir, el modelo predice mejor los registros negativos que los registros positivos.

Random Forest

Vamos a calcular un modelo Random Forest con una validación cruzada con 6 *folds*. Mediante la función *train_control* especificamos las características comentadas del proceso de entrenamiento para crear el modelo.

```
# Validación cruzada
train_control<- trainControl(method="cv", number=6)

# Random forest
set.seed(31)
mod_rf <- train(Survived~., data=titanic_train, method="rf", trControl = train_control)
```

Ahora usamos el modelo Random forest para predecir la variable objetivo con los datos del subconjunto de test y calcularemos la matriz de confusión para evaluar la calidad del modelo.

```
# Predicción del Random forest
set.seed(32)
pred_rf <- predict(mod_rf, newdata=titanic_test)

# Matriz de confusión
confusionMatrix(pred_rf, titanic_test$Survived, positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 172  45
##          1  11  69
##
##                Accuracy : 0.8114
##                  95% CI : (0.7622, 0.8543)
##     No Information Rate : 0.6162
##     P-Value [Acc > NIR] : 2.772e-13
##
##                   Kappa : 0.5776
##
##  Mcnemar's Test P-Value : 1.035e-05
##
##             Sensitivity : 0.6053
```

```
##             Specificity : 0.9399
##          Pos Pred Value : 0.8625
##          Neg Pred Value : 0.7926
##              Prevalence : 0.3838
##          Detection Rate : 0.2323
##    Detection Prevalence : 0.2694
##       Balanced Accuracy : 0.7726
##
##        'Positive' Class : 1
##
```

Con el modelo Random Forest el 81.14% de los registros del subconjunto de test fueron correctamente clasificados, por tanto, la precisión de este modelo ha aumentado casi 3% respecto al árbol de decisión.

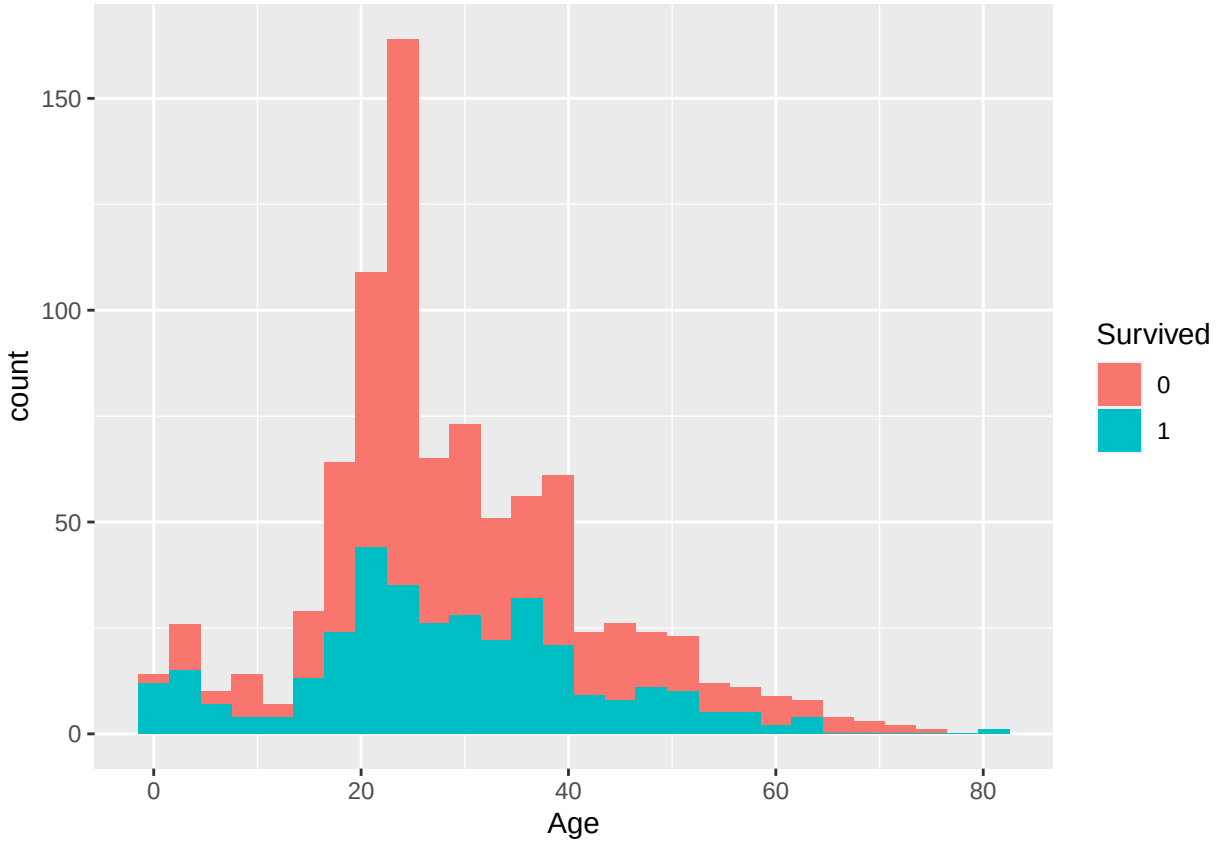
La sensibilidad de este modelo es 60,53% y su especificidad 93.99%. Si comparamos estos valores con los obtenidos mediante el árbol de decisión vemos que el modelo Random Forest clasifica con la misma precisión los pasajeros que no sobrevivieron, pero, en cambio, los pasajeros que sí sobrevivieron son clasificados correctamente con un porcentaje mayor.

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

A lo largo de la práctica se ha hecho uso de diferentes tipos de gráficas para comprender mejor los datos y analizarlos. En el apartado anterior por ejemplo, se mostró el árbol de decisión gráficamente para saber qué variables utilizaba en la toma de decisiones. También hemos usado gráficos QQ para comprobar la normalidad de algunas variables, diagramas de caja para observar los valores extremos o histogramas para ver la distribución de la variable `Age`. En este apartado mostraremos algunos ejemplos más para entender mejor el conjunto de datos.

Primero vamos a volver a mostrar el histograma de la edad de los pasajeros (sin discretizar), pero esta vez añadimos la información de la variable `Survived`. Esta vez utilizaremos la función *ggplot*.

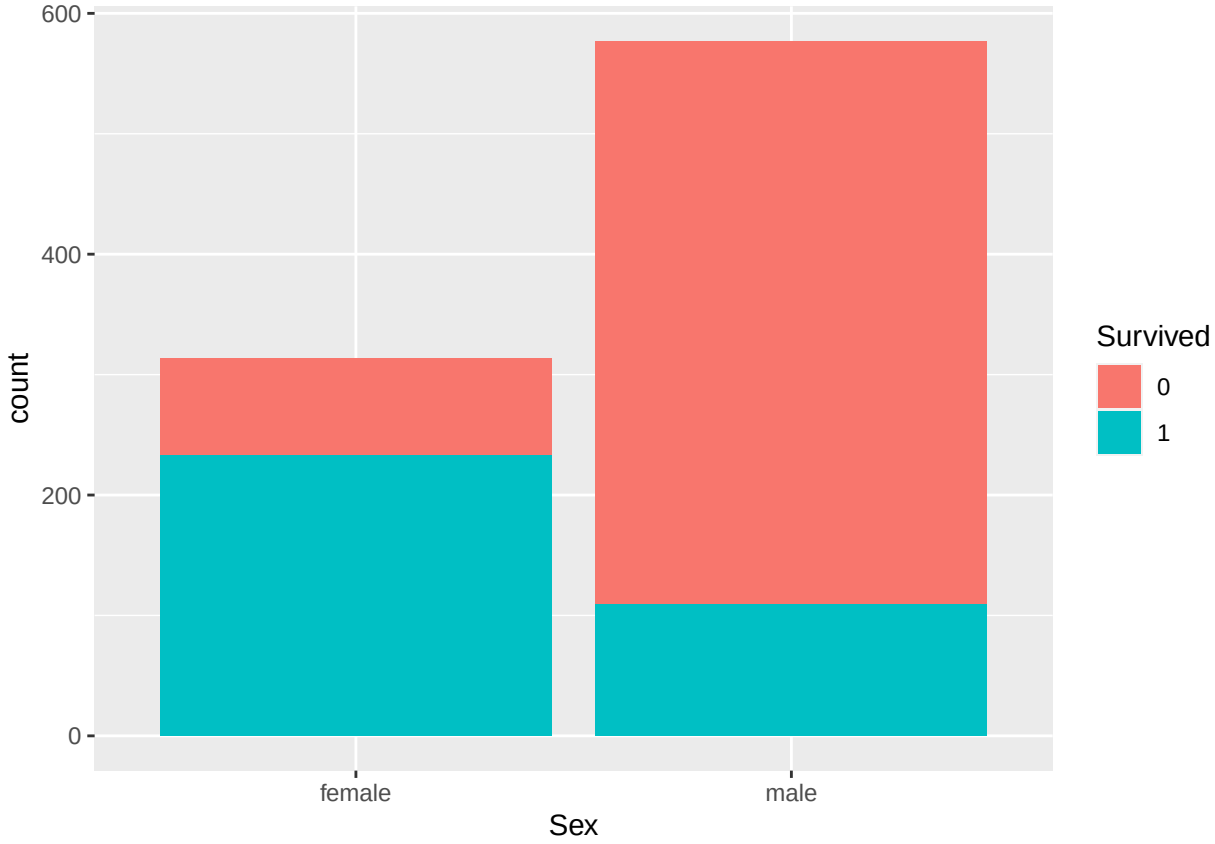
```
# Histograma de Age y Survived
ggplot(data = titanic_data,aes(x=Age,fill=Survived))+geom_histogram(binwidth =3)
```



Como ya se observó en su momento, la mayoría de pasajeros se encuentran entre los 20 y 40 años. También se observa que la tasa de supervivencia entre los pasajeros menores es bastante alta. Mientras que en los pasajeros entorno a los 25 años sobrevivieron pocos en relación al número total de personas de esa edad (unos 30 de más de 150).

Ahora vamos a mostrar un gráfico de barras con el número de pasajeros por sexo y añadimos también la información de la variable `Survived`.

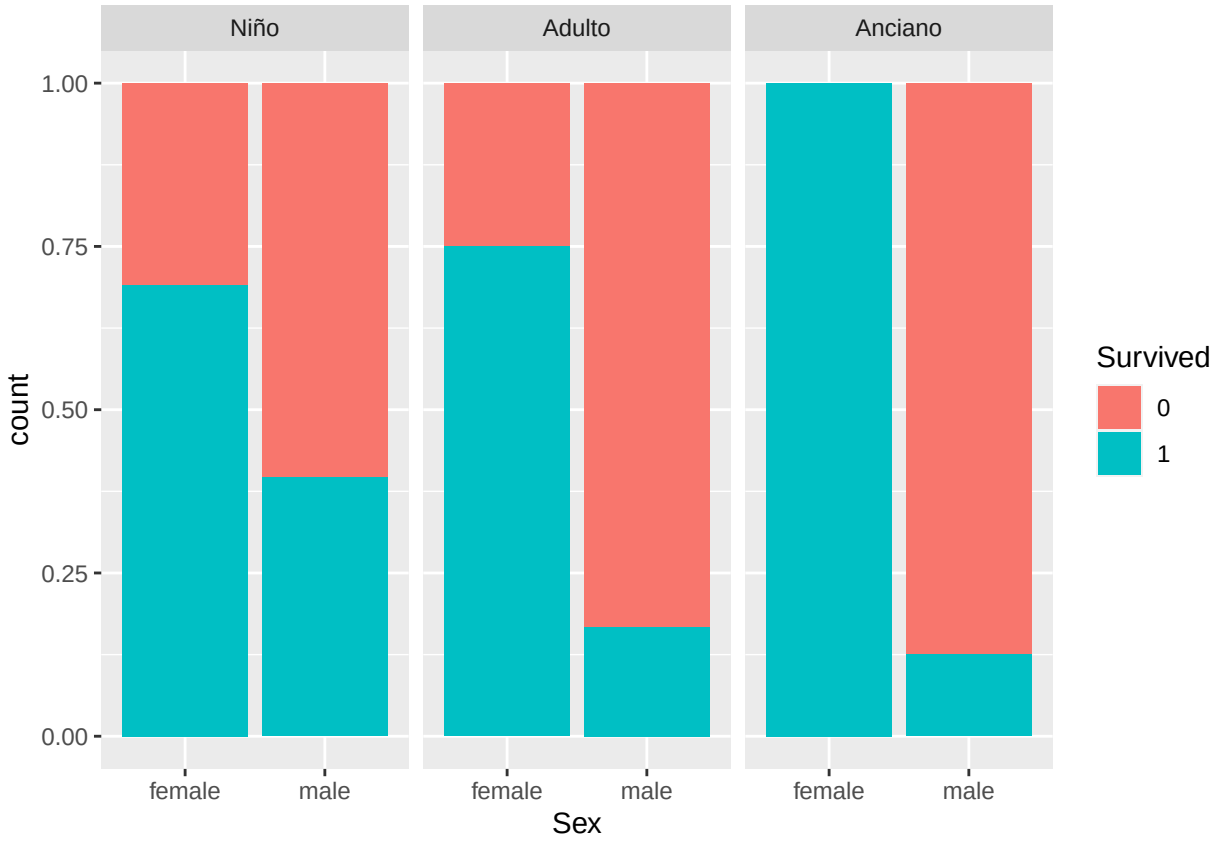
```r
# Diagrama de barra de Sex y Survived
ggplot(data=titanic_data,aes(x=Sex,fill=Survived))+geom_bar()
```



El número de pasajeros varones es casi el doble que el de mujeres, pero la proporción de mujeres que sobrevivieron es mucho mayor que la de hombres.

Para observar con más detalle las proporciones de las gráficas anteriores, mostramos a continuación las tres variables en una serie de gráficas de frecuencia, pero en este caso utilizaremos la variable `Age` discretizada.

```r
# Gráfica de frecuencias de Age, Sex y Survived
ggplot(data = titanic_data,
       aes(x=Sex, fill=Survived))+geom_bar(position="fill")+facet_wrap(~Age_d)
```

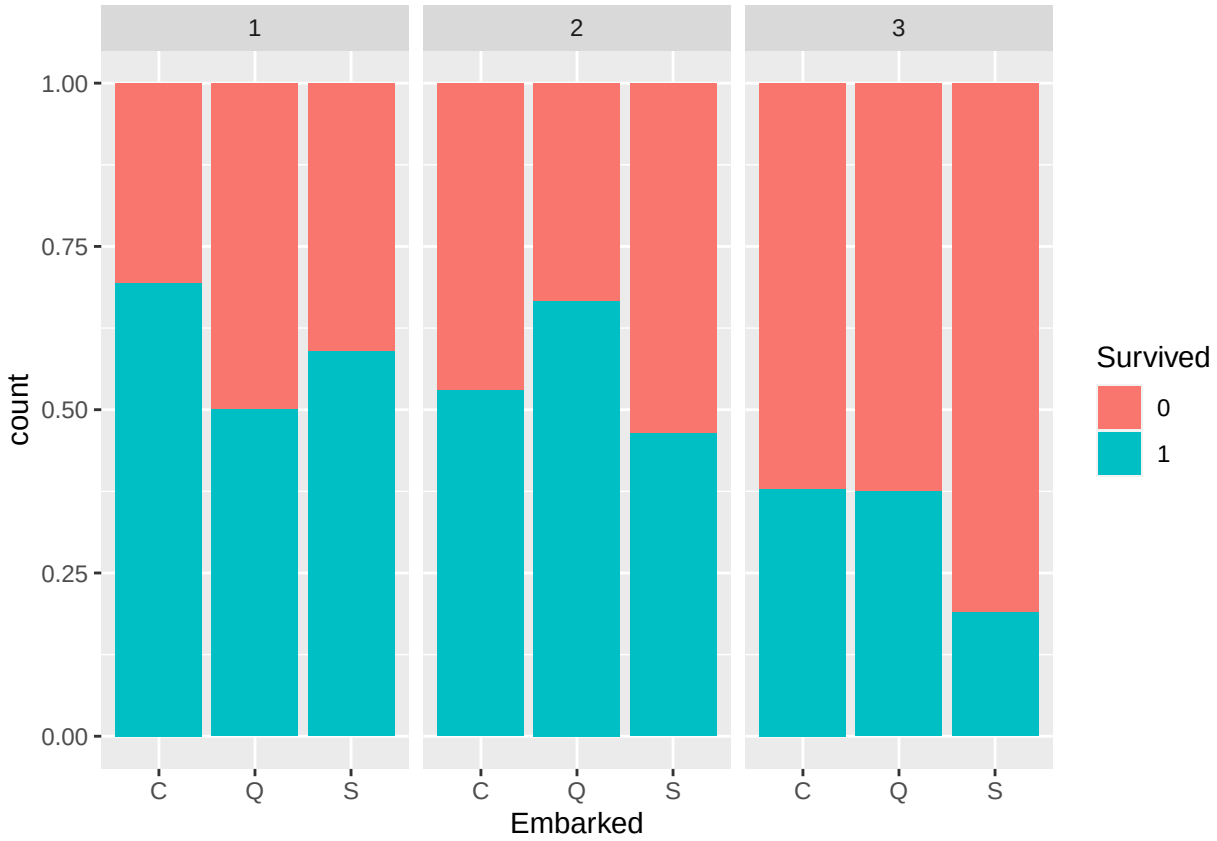


En todas las edades se observa una gran desigualdad en el índice de supervivencia entre hombres y mujeres. Destaca que las todas mujeres por encima de 60 años que viajaban en el barco se salvaron. Aunque según vimos en la distribución de los datos es posible que no fueran muchas. En cambio, la tasa de supervivencia de los hombres disminuye según avanza el grupo de edad. Aproximadamente el 40% de los niños de sexo masculino sobreviven al accidente, mientras que el porcentaje de ancianos hombres que sobrevivieron al accidente disminuye a la mitad, un 20%.

Por lo tanto, todo hace indicar que priorizaron a mujeres y niños a la hora de subir a los botes salvavidas. Como suele pasar en cualquier tipo de accidente marítimo. Respecto a los pasajeros por encima de 60 años, solo podemos afirmar que ninguna mujer murió en ese rango de edad, como ya se ha comentado.

Ahora vamos a mostrar una gráfica similar para analizar la supervivencia de los pasajeros respecto a la clase en la que viajaron y el muelle donde embarcaron.

```
# Gráfica de frecuencias de Pclass, Embarked y Survived
ggplot(data = titanic_data,
       aes(x=Embarked, fill=Survived))+geom_bar(position="fill")+facet_wrap(~Pclass)
```

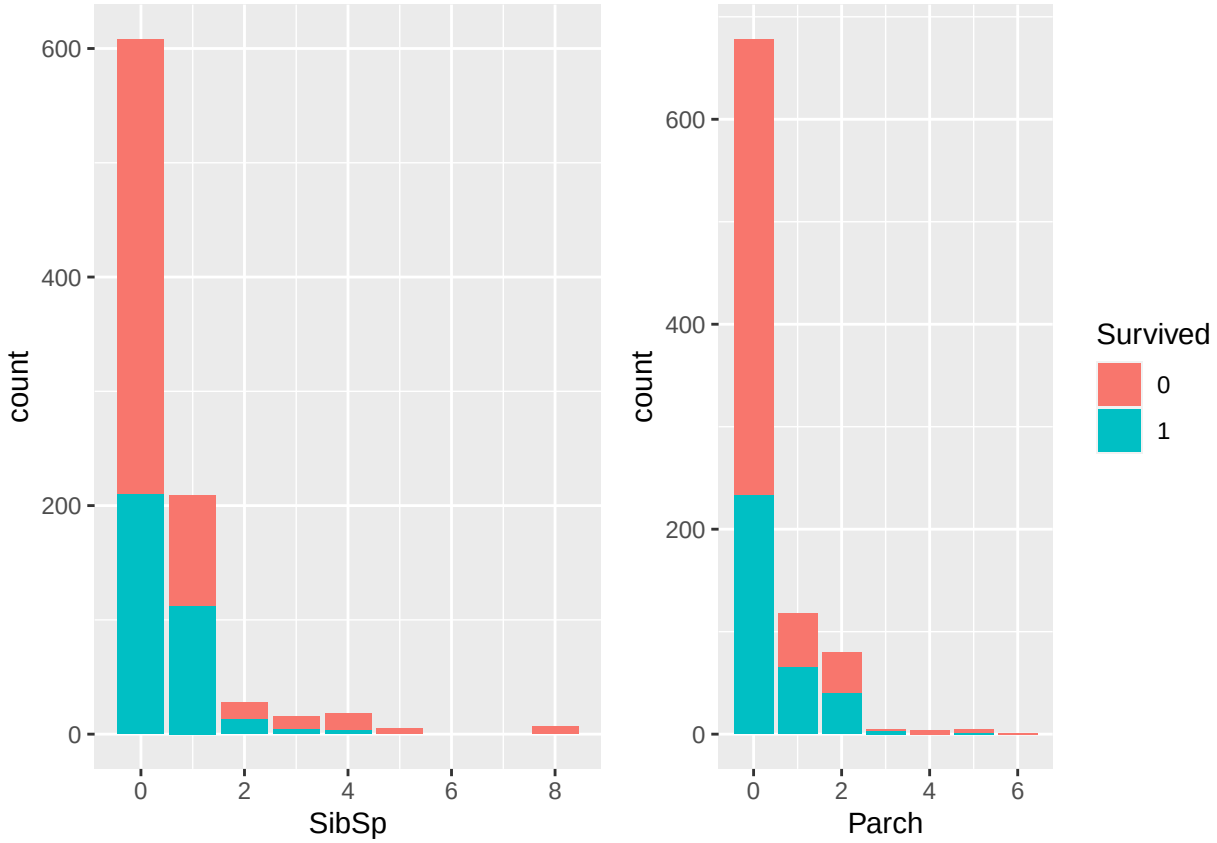


Lo primero que se observa es que la tasa de pasajeros salvados es mucho menor en los pasajeros de tercera clase. En cambio, no existe una gran diferencia entre los de segunda y tercera clase, aún así el índice de supervivencia es algo mayor dentro de primera clase. Otro dato curioso a destacar es que la frecuencia de supervivencia varía bastante dependiendo del muelle de embarque.

Comparamos mediante dos diagramas de barras, colocados uno junto al otro, la variables que indican la información familiar, `SibSp` y `Parch`.

```
Graf1 <- ggplot(data = titanic_data,
                aes(x=SibSp,fill=Survived))+geom_bar()+theme(legend.position="none")
Graf2 <- ggplot(data = titanic_data,aes(x=Parch,fill=Survived))+geom_bar()

grid.arrange(Graf1, Graf2, ncol=2)
```



La mayor parte de los pasajeros viajó sin familiares, pero se observa que la forma de ambas gráficas es similar. Esto podría indicar la presencia de correlación entre ambas variables, pero como vimos en el análisis de correlaciones no existe una correlación significativa entre ninguna variable cuantitativa.

6 Resolución del problema y conclusiones

Con todos los resultados obtenidos ya tenemos la información suficiente para responder a las cuestiones planteadas y predecir el valor de la variable objetivo mediante algunos modelos supervisados. A continuación vamos a responder a las preguntas que nos habíamos planteado al inicio del análisis:

- En los contrastes de hipótesis sobre la variable `Survived` hemos visto que se observan diferencias significativas en la variable `Pclass` entre los pasajeros que sobrevivieron y los que no. Después en el modelo de regresión logística hemos comprobado que esta variable tiene significancia en el modelo y que el hecho de ser un pasajero de tercera clase era un factor de riesgo respecto a ser un pasajero de primera clase. Por tanto, podemos afirmar que los pasajeros de primera clase tuvieron más probabilidades de sobrevivir que los de tercera clase.
- Al igual que en la pregunta anterior, el contraste de hipótesis nos dio información de que la variable `Sex` presentaba diferencias significativas entre los pasajeros que sobrevivieron y los que no, y el modelo de regresión logística ha demostrado la significancia de esta variable y que los hombres tuvieron una probabilidad menor de sobrevivir frente a las mujeres. Esto también se ha podido observar mediante la representación gráfica de estas variables.
- En caso del puerto de embarque (`Embarked`), aunque el contraste de hipótesis nos mostraba que había diferencias significativas entre los pasajeros que sobrevivieron y los que no, y la representación gráfica de esta variable también apoyaba esta hipótesis, el modelo de regresión nos ha indicado que esta variable no tiene significancia en el hecho de sobrevivir o no y, por tanto, no influyó en la probabilidad de sobrevivir. Además, el árbol de decisión que hemos calculado posteriormente no ha tenido esta variable en cuenta, por lo que se refuerza esta conclusión.

Con los análisis realizados anteriormente hemos podido dar respuesta a todas las preguntas que nos habíamos planteado y se podría responder a otras muchas. También hemos podido ver la relación entre las distintas variables disponibles en el conjunto de datos y, además, podemos predecir si un pasajero sobrevive o no en función del valor de las variables independientes gracias a un modelo de clasificación supervisado.

Contribuciones	Firma
Investigación previa	P.L.L, R.C.A
Redacción de respuestas	P.L.L, R.C.A
Desarrollo código	P.L.L, R.C.A