PRA2: Limpieza y análisis de datos

Alvaro Lopez y Jorge Sainero

1 de junio, 2022

Contents

1 Descripción del dataset

El dataset que hemos elegido para realizar esta práctica es el que aperecía como una de las posibles opciones del enunciado "Titanic: Machine Learning from Disaster".

Este dataset contiene información sobre los pasajeros que iban en el Titanic y es un conjunto de datos importante y relevante a nivel histórico porque contiene información sobre el naufragio que es una de las mayores tragedias marítimas de la historia.

Este dataset pretende responder a qué criterios se siguieron a la hora de salvar las vidas de los pasajeros y la tripulación o si simplemente fue azar. Se determinarán qué variables influyeron más en la supervivencia de los pasajeros.

Para resolver el problema, primero se seleccionarán los datos de interés y se limpiará el conjunto de datos para que los posteriores análisis estén dotados de calidad. Seguidamente se comprobará la normalidad y la homoscedasticidad que ayudarán a entender las variables cuantitativas. Después se harán dos contrastes de hipótesis que adaptan las propiedades de la muestra en la población, una regresión logística y un modelo de *randomForest*, los cuales clasificarán registros no etiquetados en si un pasajero sobrevive o no.

2 Selección de los datos de interés

Comenzamos cargando el dataset y viendo la estructura de este.

```
titanic <- read.csv('data/titanic_train_data.csv')
str(titanic)
```

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

Observamos que el dataset tiene 891 observaciones donde cada observación representa a un pasajero, y hay 12 columnas que informan cada registro.

Las columnas son las siguientes:

Columna	Descripción
PassengerId	Id para identificar el registro. Único para cada fila
Survived	Indica si el pasajero sobrevivió o no. 0 = No, 1 = Sí
Pclass	Clase del billete. 1 = primera, 2 = segunda, 3 = tercera
Name	Nombre del pasajaro
Sex	Género del pasajero
Age	Edad del pasajero
SibSp	Número de hermanos y pareja del pasajero que viajaban en el barco
Parch	Número de padres e hijos del pasajero que viajaban en el barco
Ticket	Número del ticket
Fare	Precio del ticket
Cabin	Número de la cabina
Embarked	Indica donde embarcó el pasajero. C = Cherbourg, Q = Queenstown, S = Southampton

2.1 Eliminación de las columnas que no son de interés

En principio la columna *Name* no parece ser de interés, pero como se ve en la estructura y que se repite en todo el dataset, todos los pasajeros tienen un honorífico o título personal, como `Miss.` o `Mr`. A continuación se extrae el título.

```
nombre <- titanic$Name

extraccion <- str_extract(nombre, ", [A-Za-z ]*")
extraccion <- substring(extraccion, first=3)

PersonalTitle <- extraccion
table(PersonalTitle)
```

```
## PersonalTitle
##         Capt          Col          Don           Dr     Jonkheer         Lady
##            1            2            1            7            1            1
##        Major       Master         Miss         Mlle          Mme           Mr
##            2           40          182            2            1          517
##          Mrs           Ms          Rev          Sir the Countess
##          125            1            6            1            1
```

Hay 17 niveles, y solo 4 identifican a más de 10 pasajeros. Los tres títulos más comunes son `Mr` para hombres adultos, `Miss` para mujeres no casadas y `Mrs` para mujeres casadas. Como hay muchos con poca representación y *Sex* ya identifica a hombres y mujeres, esta nueva variable no se incluye en el dataset.

De las columnas originales, el *PassengerId* y el *Name* no nos aportan niguna información así que las eliminaremos.

```
titanic <- titanic[,c(2,3,5:12)]
```

Revisamos el resto de los campos para ver si hay alguno más que no vaya a aportar valor al análisis.

```
head(titanic, 5)
```

```
##   Survived Pclass    Sex Age SibSp Parch           Ticket    Fare Cabin
## 1        0      3   male  22     1     0        A/5 21171  7.2500
## 2        1      1 female  38     1     0         PC 17599 71.2833   C85
## 3        1      3 female  26     0     0 STON/O2. 3101282  7.9250
## 4        1      1 female  35     1     0           113803 53.1000  C123
## 5        0      3   male  35     0     0           373450  8.0500
##   Embarked
## 1        S
## 2        C
## 3        S
## 4        S
## 5        S
```

Observando los campos *Ticket* y *Cabin* parece que van a ser identificadores únicos o casi únicos y que no van a aportar valor al análisis así que decidimos eliminarlos también.

```
titanic <- titanic[,c(1:6,8,10)]
```

2.2 Cambio de tipo de las columnas

Antes de comenzar con la limpieza de los datos, conviene que las columnas sean del tipo correcto. Observando la descripción de las columnas, tenemos cuatro columnas que son de tipo `factor` y no de tipo número o string: *Survived*, *Pclass*, *Sex* y *Embarked*.

```
titanic$Survived <- factor(titanic$Survived, levels = c(0,1), labels = c('No','Yes'))
titanic$Pclass <- factor(titanic$Pclass, levels = c(1,2,3), labels = c('1st','2nd','3rd'))
titanic$Sex <- as.factor(titanic$Sex)
titanic$Embarked <- factor(titanic$Embarked, levels = c('C','Q','S'),
                           labels = c('Cherbourg','Queenstown','Southampton'))
```

Comprobamos que este cambio se ha realizado correctamente.

```
summary(titanic)
```

```
##  Survived  Pclass      Sex           Age            SibSp
##  No :549   1st:216   female:314   Min.   : 0.42   Min.   :0.000
##  Yes:342   2nd:184   male  :577   1st Qu.:20.12   1st Qu.:0.000
##            3rd: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             Fare              Embarked
##  Min.   :0.0000   Min.   :  0.00   Cherbourg  :168
##  1st Qu.:0.0000   1st Qu.:  7.91   Queenstown : 77
##  Median :0.0000   Median : 14.45   Southampton:644
##  Mean   :0.3816   Mean   : 32.20   NA's       :  2
##  3rd Qu.:0.0000   3rd Qu.: 31.00
##  Max.   :6.0000   Max.   :512.33
##
```

3 Limpieza de datos

3.1 Análisis de nulos, ceros y elementos vacíos

```
colSums(is.na(titanic))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0      177        0        0        0        2
```

Vemos que tenemos 2 registros con la variable *Embarked* a nulo. Esto no nos debería preocupar demasiado ya que no son demasiados y podríamos tomar la decisión de eliminar estos registros del dataset. En cambio tenemos 177 registros con la variable *Age* a nulo. En este caso si tenemos que decidir entre imputar los valores o eliminar la variable del análisis. La primera opción es la más viable porque no es factible eliminar casi el 20% de los registros del conjunto.

```
colSums(titanic=="")
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0       NA        0        0        0       NA
```

```
colSums(titanic==0)
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0       NA      608      678       15       NA
```

Comprobando también qué valores son vacíos o cero no vemos nada extraño. Observamos que hay un gran número de personas que viajan sin familiares o sin familia (*SibSp* y *Parch* igual a 0) y que hay 15 pasajeros que viajan gratis (*Fare* igual a 0).

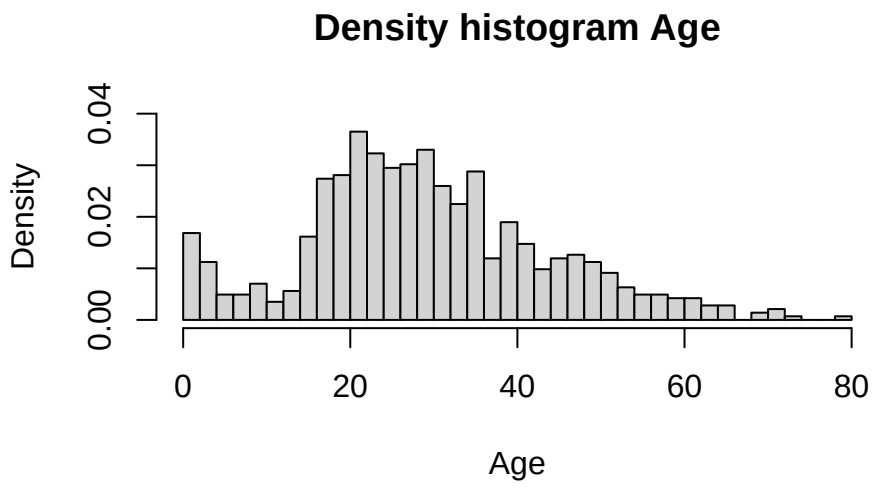
3.1.1 Tratamiento de valores nulos

Como hemos comentado antes, los registros con las variable *Embarked* a nula los vamos a eliminar.

```
ind.Embarked <- which(is.na(titanic$Embarked))
titanic <- titanic[-ind.Embarked,]
```

Para imputar los valores de la variable *Age* hemos optado por el algoritmo KNN (*K-Nearest-Neighbor*), implementado mediante la función *kNN* del paquete *VIM*. Realiza la media ponderada de los k vecinos más cercanos, en este caso 10.

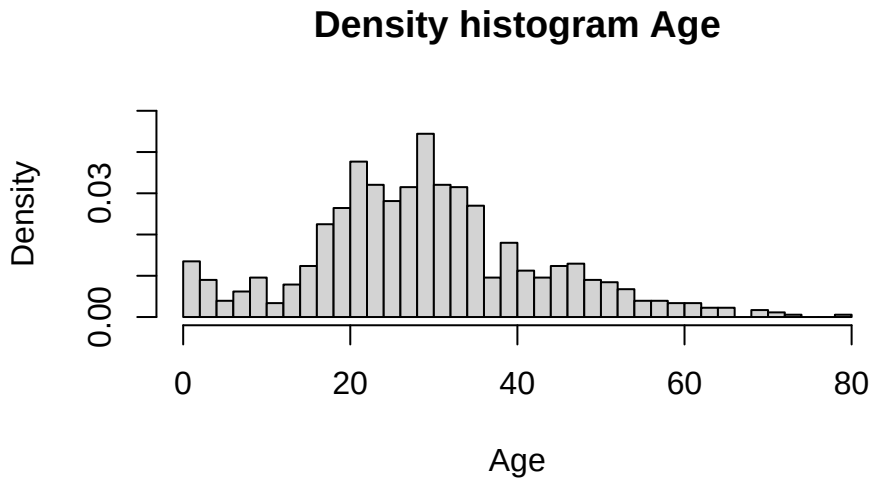
Antes se grafica el histograma de la edad para ver, después de imputar los valores faltantes, la diferencia en la distribución.



La mayor densidad se encuentra en el rango de edad entre 20 y 30 años. Hacia la izquierda nos encontramos con un valle para luego aumentar para niños menos de 4 años, pero hacia la derecha tiende a decrecer poco a poco.

Ahora se imputan los valores faltantes de *Age*.

```
titanic$Age <- trunc(kNN(titanic, variable = "Age", k = 10, numFun =  "weightedMean",
                         imp_var = FALSE)$Age)
```



La distribución es aproximadamente la misma. El único aumento considerable es entre 28 y 30 años. Gran parte de los pasajeros de los que no se disponía la edad podían tener esa edad aproximada.

```
summary(titanic$Age)
```

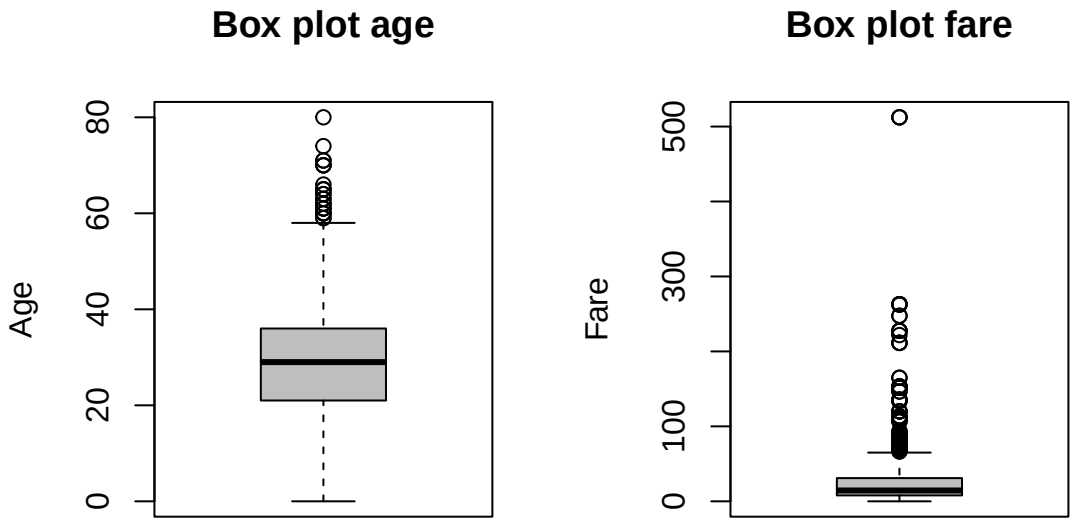
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    21.0    29.0    29.5    36.0    80.0
```

El resumen de la variable modificada es similar al de la variable original. El valor que más cambia es la mediana, pasando de 28 años a 29 años.

Concluimos que esta forma de imputar los valores es mejor que haber colocado a todos los registros el valor de la media o la mediana.

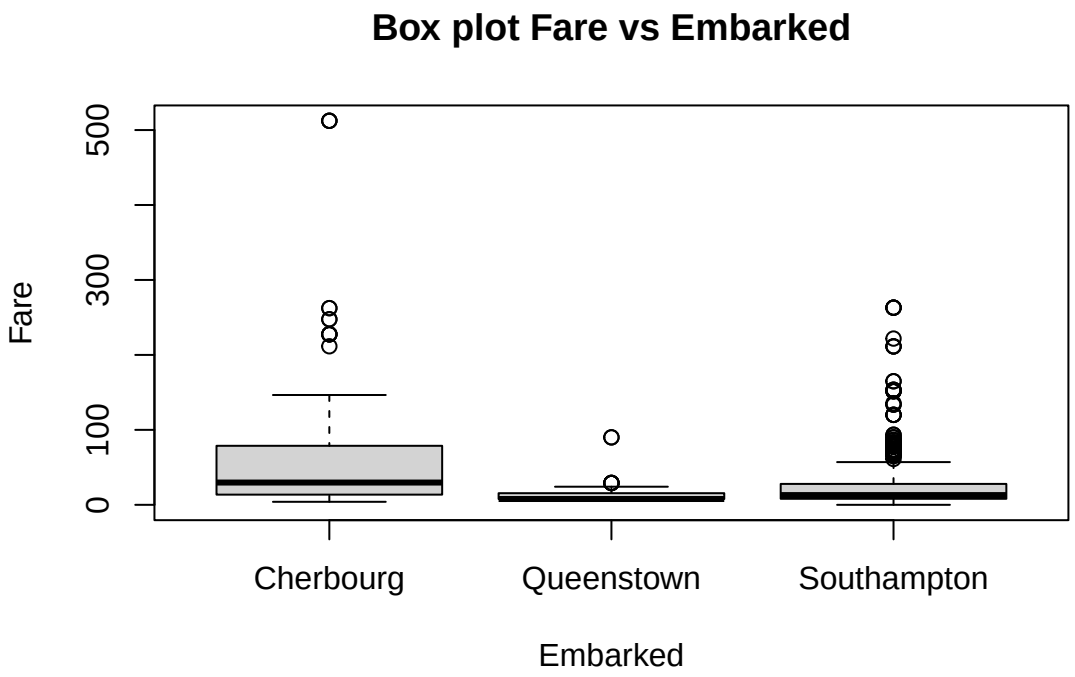
3.2 Detección y tratamiento de outliers

Los *outliers* o valores extremos son los datos que se encuentran significativamente alejados del resto de datos. Si se tratan de valores razonables se dejarán los registros. De lo contrario, se eliminarán.



Observamos que en el caso de la edad no vemos ningún outlier que sea relevante ya que los que vemos pertenecen a personas que tienen entre 60 y 80 años que son valores razonables aunque sean notablemente superiores a los del resto del conjunto.

En el caso de la variable *Fare* vemos muchos outliers lo que puede ser porque las tarifas cambien en función de donde hayan embarcado los pasajeros por lo que vamos a repetir el dibujo de los boxplots pero para cada uno de los conjuntos.



Como sucedía con la edad, los outliers no tan alejados de las cajas se consideran razonables, pero vemos dos puntos (que realmente podrían ser dos o más registros) destacados en los casos de `Cherbourg` y `Queenstown` que analizaremos a continuación.

```
max.Cherbourg <- max(boxplot.stats(titanic[titanic$Embarked == 'Cherbourg', 'Fare'])$out)
titanic[titanic$Fare == max.Cherbourg & titanic$Embarked == 'Cherbourg',]
```

```
##     Survived Pclass    Sex Age SibSp Parch     Fare  Embarked
## 259      Yes    1st female  35     0     0 512.3292 Cherbourg
## 680      Yes    1st   male  36     0     1 512.3292 Cherbourg
## 738      Yes    1st   male  35     0     0 512.3292 Cherbourg
```

```
max.Queenstown <- max(boxplot.stats(titanic[titanic$Embarked == 'Queenstown', 'Fare'])$out)
titanic[titanic$Fare == max.Queenstown & titanic$Embarked == 'Queenstown',]
```

```
##     Survived Pclass    Sex Age SibSp Parch Fare   Embarked
## 246       No    1st   male  44     2     0   90 Queenstown
## 413      Yes    1st female  33     1     0   90 Queenstown
```

Los cinco precios no parecen ser outliers ya que pertenecen a pasajeros con billetes de primera clase que son los más caros. Así que no los eliminaremos del conjunto de datos.

3.3 Guardar nuevo dataset

Después de eliminar las cuatro variables y los dos registros, guardamos el nuevo conjunto de datos csv.

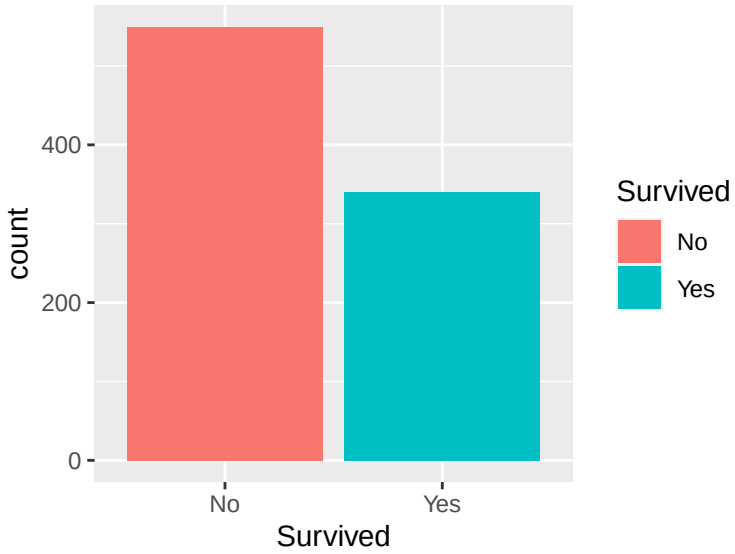
```
write.csv(titanic, "titanic_train_data_clean.csv")
```

4 Análisis de datos

4.1 Selección de los grupos de datos a analizar

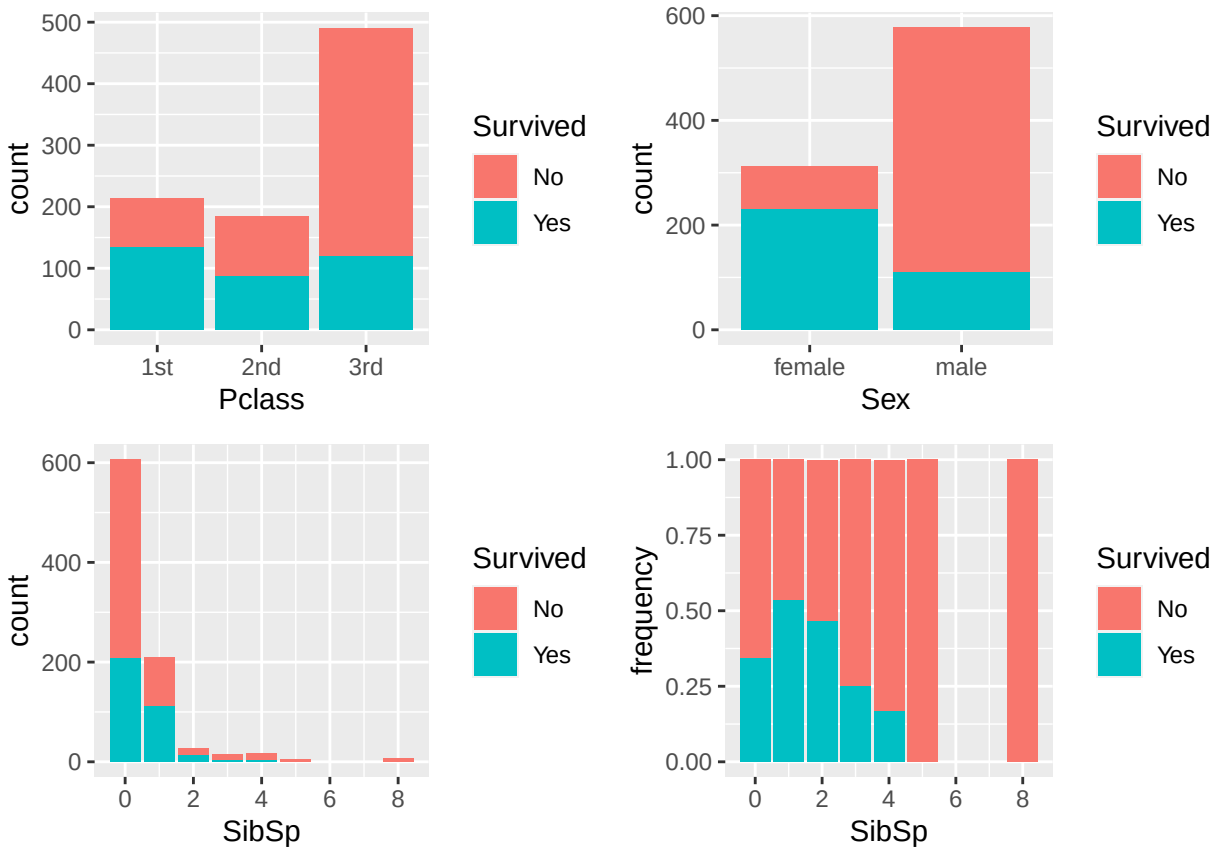
Partiendo de la premisa de que la variable principal de nuestro conjunto de datos es la variable *Survived*, los dos grupos principales que compararemos son supervivientes y no supervivientes a través de las distintas variables.

Primero visualizaremos la variable *Survived* individualmente y luego generaremos una serie de diagramas de barras y gráficos de cajas para ver la distribución de *Survived* y su relación con el resto de variables.

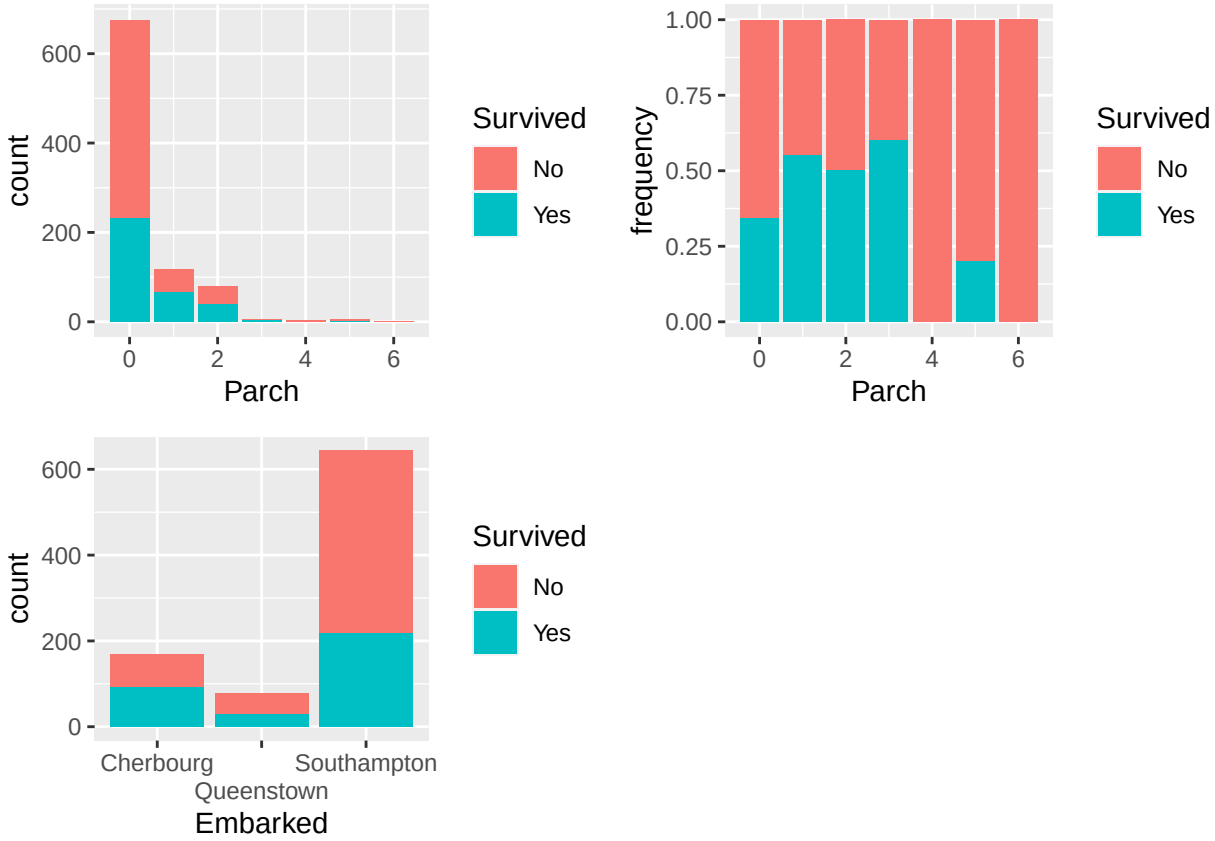


El grupo con más representación es `No`, correspondiente a los pasajeros que no sobrevivieron. Hay 549 registros de fallecidos y 340 de supervivientes.

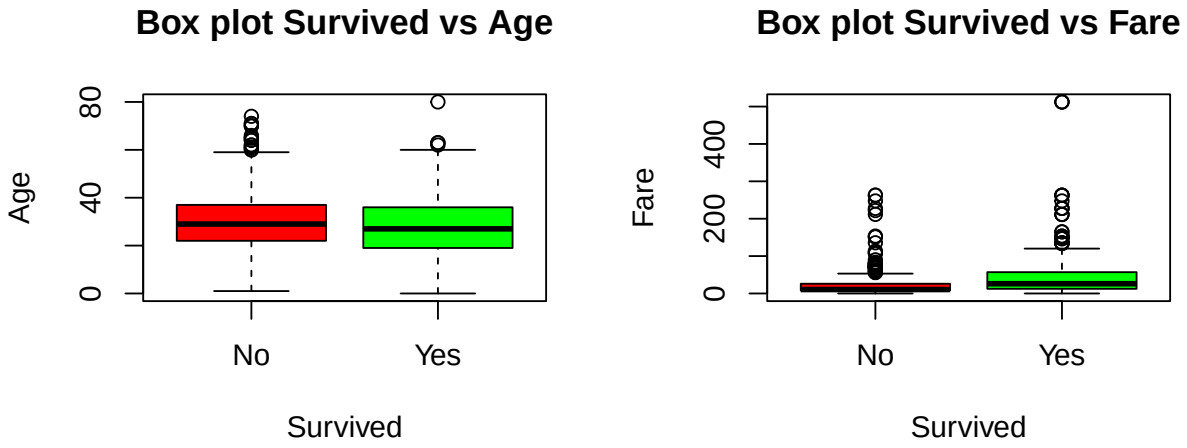
Para las variables *SibSp* y *Parch* además de los diagramas de barras, optamos por estandarizar cada barra para que tenga la misma altura y muestre las proporciones relativas para apreciar bien todos los grupos.



Respecto a los gráficos, cabe destacar que más del doble de pasajeros tienen billetes de tercera clase y el 65% aproximadamente de todos los pasajeros son hombres. Sin embargo, estos dos grupos para sus respectivas variables son quienes menos probabilidad de supervivencia tienen. El número de parejas o hermanos también en barco es mayoritariamente 0, seguido de 1. Hay más probabilidad de sobrevivir acompañado de 1 o 2 personas que de ninguna.



El número de padres e hijos también en el barco aumenta en 0 y 2 respecto a *SibSp*, aunque los que van en este aspecto solos tienen aproximadamente la sobrevivir. El 72% de pasajeros embarcaron en Southampton pero estos tienen un 34% de supervivencia, el menor porcentaje de los tres lugares.



Respecto a *Age*, el el 50% del boxplot (caja) de los fallecidos abarca más edades que los supervivientes. Y para *Fare* ocurre lo contrario. Estos dos diagramas ayudarán a definir si hay homogeneidad en la varianza o no.

De todos las variables comparadas con *Survived*, se usarán aquellas que para la regresión logística y *random*

forest sean útiles. Y para los contrastes de hipótesis se usará *Survived* con *Fare* y *Survived* con *Fare*.

4.2 Comprobación de la normalidad y homogeneidad de la varianza

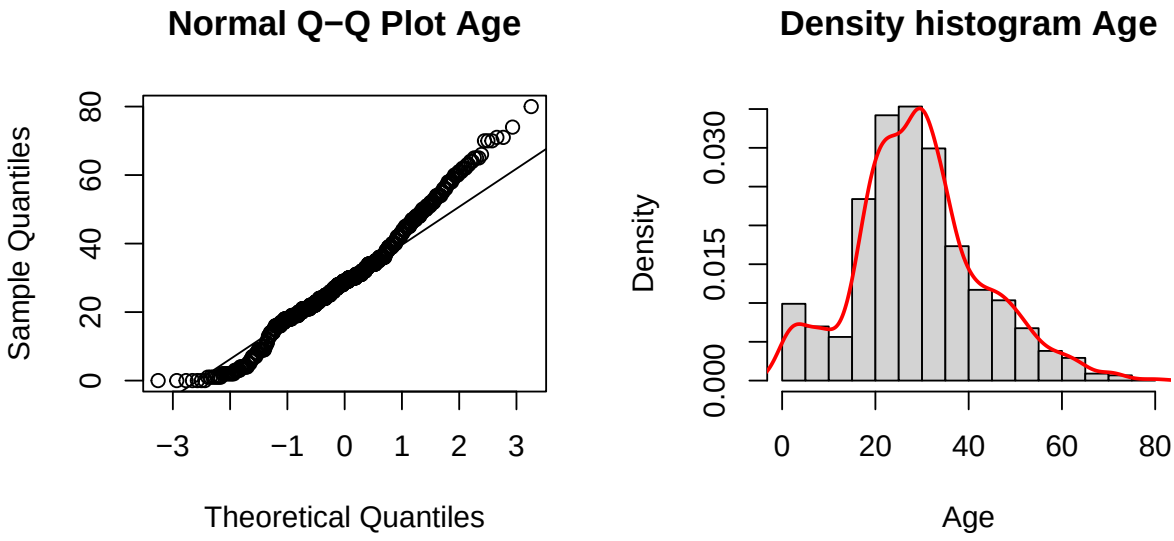
4.2.1 Normalidad

El análisis de la normalidad determina si los datos de la muestra se han extraído de una población distribuida normalmente. Dos gráficos con los que es posible comprobar la normalidad son el gráfico Q-Q o cuantil-cuantil y el diagrama de densidad con la curva de distribución normal.

El Q-Q Plot permite identificar la desviación de los datos de la muestra respecto a una población normal. Mediante *qqline* se dibuja una línea recta correspondiente a la distribución normal teórica, y mediante *qqnorm* se dibujan los puntos, que son los datos distribuidos según los cuantiles teóricos. Si los puntos siguen la línea, los datos se distribuyen normalmente.

El diagrama de densidad muestra la distribución de los datos. Se aproximarán a una distribución normal si la densidad es simétrica, centrada en el medio, y disminuye hacia las 2 desviaciones estándar de la media, dentro de las cuales se encuentra aproximadamente el 95% de los datos.

Variable *Age*



En el Q-Q Plot, los datos hacia los extremos se alejan de la recta. El histograma de densidad ya se había comentado anteriormente y se observa la caída suave hacia la derecha pero rápidamente hacia la izquierda, para dejar un pequeño valle.

En base a las representaciones se concluiría que los datos no siguen una distribución normal. No obstante, para comprobarlo se usa el test de Shapiro-Wilk. Asume como hipótesis nula que la población está distribuida normalmente. Se rechazará si el p-valor es inferior al nivel de significancia $\alpha = 0.05$.

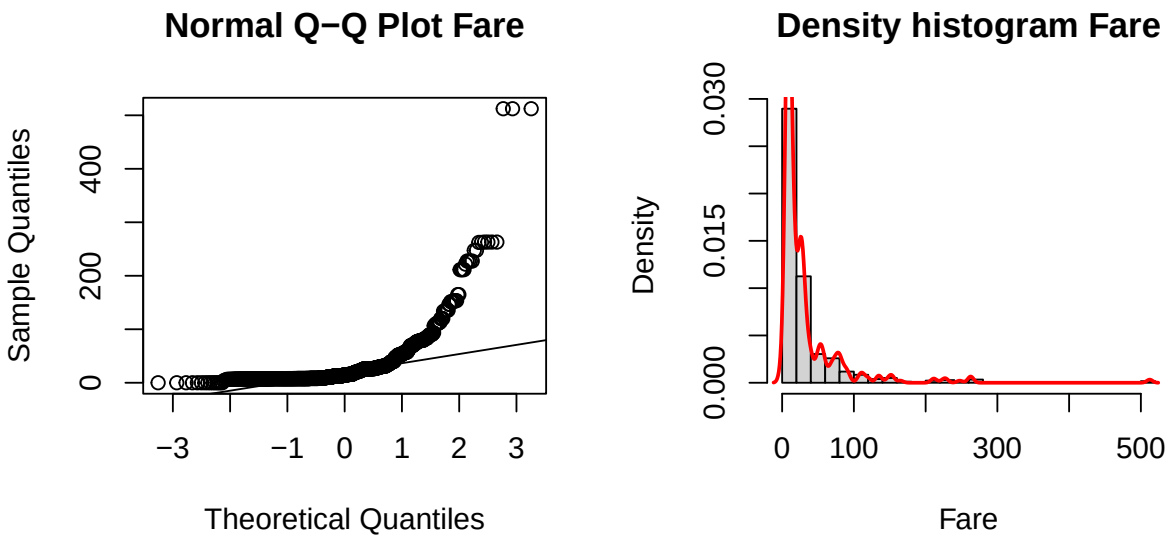
```
shapiro.test(titanic$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic$Age
## W = 0.98223, p-value = 6.293e-09
```

Como el p-valor es inferior a 0.05, efectivamente se rechaza la hipótesis nula, afirmano la no normalidad de los datos.

Por otro lado, como el conjunto de datos es suficientemente grande (879 registros), y por el teorema del límite central (TLC) se podría considerar que la media de la muestra sigue una distribución normal.

Variable *Fare*




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

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic$Fare
## W = 0.5197, p-value < 2.2e-16
```

Para la variable *Fare*, los datos se disparan a partir del cuantil teórico 1 y la mayor parte de precios se encuentra entre 0 y 20 dólares. La densidad decrece a medida que aumenta el precio del billete. Además, el p-valor en la prueba es muy inferior a 0.05. Por tanto, los datos no se distribuyen normalemnte, pero al aplicar el TLC, la media de los datos sí sigue una distribución normal.

4.2.2 Homocedasticidad

Un test de homocedasticidad comprueba la igualdad de varianzas entre los grupos a comparar. Según los gráficos y el test de Shapiro-Wilk, las variables *Age* y *Fare* no siguen una distribución normal. Por tanto se utiliza el test de Fligner-Killeen para comprobar la homocedasticidad. Con la misma premisa que para la normalidad, la hipótesis nula asume igualdad de varianzas. Se rechazará si el p-valor es inferior al nivel de significancia $\alpha = 0.05$.

Variable *Age*

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

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Survived
## Fligner-Killeen:med chi-squared = 3.9443, df = 1, p-value = 0.04703
```

El p-valor = 0.04703 < 0.05. Se rechaza la hipótesis nula, pero no de una manera clara. A pesar de aproximarse a 0.05, se concluye que la variable *Age* presenta varianzas estadísticamente diferentes para los dos grupos de *Survived*.

Esto se puede representar con los boxplots visualizados en el apartado 4.1. La amplitud de la caja y de los bigotes del grupo que no sobrevivió es un poco mayor a la amplitud del grupo que sobrevivió. Esta diferencia pequeña asume la no igualdad de varianzas para las dos distribuciones.

Variable *Fare*

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

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Survived
## Fligner-Killeen:med chi-squared = 94.7, df = 1, p-value < 2.2e-16
```

Con la variable *Fare* hay heterocedasticidad: como el p-valor < 0.05, se rechaza la hipótesis nula y se concluye que la variable *Fare* presenta varianzas estadísticamente diferentes para los dos grupos de *Survived*. Además, en el gráfico `Box plot Survived vs Fare`, la amplitud de la caja y de los bigotes del grupo que sobrevivió es aproximadamente 65 dólares mayor a la amplitud del grupo de no sobrevivió.

4.3 Aplicación de pruebas estadísticas

Todas las pruebas estadísticas tendrán en cuenta la variable objetivo *Survived*. En primer lugar se realizarán dos contrastes de hipótesis, uno con la variable cuantitativa *Fare* y otro con la variable categórica *Sex*. Luego, una regresión logística y, finalmente, un modelo supervisado *random forest*.

4.3.1 Contrastes de hipótesis

El primer paso para un contraste de hipótesis es formular la pregunta de investigación. En base a ella se examinará la hipótesis nula y la alternativa. Finalmente, al usar el test estadístico correcto se aceptará o se rechazará la hipótesis con cierto nivel de confianza.

La primera pregunta que nos hacemos es:

¿Los pasajeros que sobrevivieron pagaron una entrada más cara que los que fallecieron?

Se plantea si el precio medio del ticket de los supervivientes (μ_1) es mayor al precio medio del ticket de los fallecidos (μ_2). La hipótesis nula (H_0) representa el caso donde no hay efecto, es decir, cuando la media de *Fare* es la misma con un nivel de confianza para todos los pasajeros. La hipótesis alternativa (H_1) es cuando se responde afirmativamente a la pregunta, es decir, los supervivientes pagaron más que los fallecidos.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

Se trata de un contraste paramétrico unilateral por la derecha de dos muestras independientes (supervivientes y no supervivientes) sobre la media con varianza desconocida.

Al preguntarse si un grupo pagó más que el otro, el test es unilateral por la derecha y, por lo tanto, la zona de aceptación de la H_0 está comprendida entre $(-\infty, z_{1-\infty}] = (-\infty, 1.64]$.

El test a utilizar es paramétrico ya que, aunque la igualdad de varianzas no se cumple, por el TLC la distribución de la media de *Fare* se aproxima a una normal. Así que se realiza el test *t-Student* con un nivel de confianza del 95%. Se implementa en R mediante la función *t.test()*. Si el p-valor es inferior al nivel de significancia, H_0 se rechaza.

```
t.test(titanic$Fare[titanic$Survived=="Yes"], titanic$Fare[titanic$Survived=="No"],
       alternative="greater", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  titanic$Fare[titanic$Survived == "Yes"] and titanic$Fare[titanic$Survived == "No"]
## t = 6.7597, df = 433.18, p-value = 2.241e-11
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  19.72906      Inf
## sample estimates:
## mean of x mean of y
##  48.20950  22.11789
```

El estadístico de contraste t = 6.7597 cae fuera de la zona de aceptación de la hipótesis nula, y el p-valor = 2.241e-11 < 0.05. Así que se concluye que los pasajeros que sobrevivieron pagaron una entrada más cara que los pasajeros que fallecieron con un nivel de confianza del 95%.

La segunda pregunta que nos hacemos es:

¿Existe una relación entre el sexo del pasajero y si sobrevivieron o no?

En este caso, las hipótesis nula y alternativa son:

- H_0: las variables *Sex* y *Survived* son independientes.
- H_1: existe una relación entre ambas variables y las diferencias son significativas.

```
table_survived_sex <- table(titanic$Survived, titanic$Sex)
chisq.test(table_survived_sex)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_survived_sex
## X-squared = 258.43, df = 1, p-value < 2.2e-16
```

El p-valor es muy inferior al nivel de significancia y, por lo tanto, sí existe una relación de dependencia entre las dos variables con un nivel de confianza del 95%.

4.3.2 Regresión logística

La regresión logística es un análisis utilizado para predecir el resultado de una variable dicotómica dependiente, en este caso *Survived*, en función de otras variables predictoras.

Primero se divide el conjunto de datos en el subconjunto de entrenamiento o *training* y el de test o *testing* mediante el método de exclusión o *holdout* estratificado. El 75% de los datos totales se usarán para entrenar el modelo y el 25% restante para evaluarlo. Además se usa el método de validación cruzada o *cross-validation* de tipo *10-fold* para garantizar la independencia de los resultados respecto a la partición.

La regresión logística se implementa mediante el método "glm" en *train*. En un principio incluimos las 7 variables independientes (todas menos *Survived* que es la variable dependiente), pero se observa que para *Parch*, *Fare* y *Embarked* el p-valor era superior a 0.05, es decir, eran variables no significativas. Por tanto se descartaron del modelo y se dejó el actual.

```
# Semilla para que los datos sean reproducibles
set.seed(666)

# Separación en train y test
```

```
data_glm <- titanic[,c(1:5)]
h<-holdout(data_glm$Survived,ratio=0.75,mode="stratified")
data_train<-data_glm[h$tr,]
data_test<-data_glm[h$ts,]

table(data_train$Survived)
```

```
##
##  No Yes
## 412 255
```

```
table(data_test$Survived)
```

```
##
##  No Yes
## 137  85
```

Las dos clases se han dividido correctamente según el ratio especificado.

```
train_control<- trainControl(method="cv", number=10)
model<-train(Survived~., data=data_train, method="glm", trControl = train_control)
summary(model)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9485  -0.6141  -0.4108   0.5958   2.5321
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.85062    0.52688   9.206  < 2e-16 ***
## Pclass2nd   -1.40570    0.32841  -4.280 1.87e-05 ***
## Pclass3rd   -2.54151    0.31878  -7.973 1.55e-15 ***
## Sexmale     -2.71833    0.22615 -12.020  < 2e-16 ***
## Age         -0.06123    0.01001  -6.115 9.65e-10 ***
## SibSp       -0.39428    0.11773  -3.349 0.000811 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 887.35  on 666  degrees of freedom
## Residual deviance: 576.53  on 661  degrees of freedom
## AIC: 588.53
##
## Number of Fisher Scoring iterations: 5
```

Gracias a la eliminación de algunas variables, el criterio de información de Akaike (AIC) ha disminuido hasta 588.53 y todas las variables son significativas. Quienes más contribución tienen en la predicción según los estimadores son los hombres, los pasajeros de tercera clase y la combinación de mujeres de primera clase (Intercept).

Si la probabilidad de la predicción es inferior a 0.5, el pasajero será clasificado como fallecido (clase positiva). En caso contrario como superviviente. Mediante la función *confusionMatrix* obtenemos la matriz de confusión

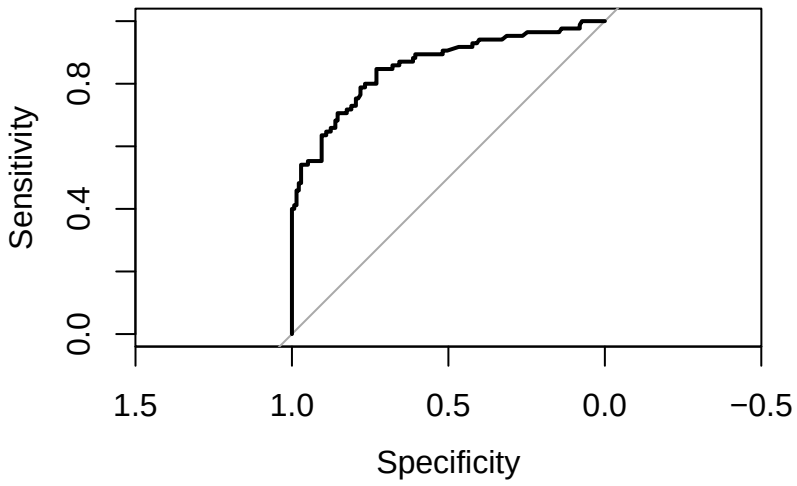
y las principales métricas.

```
pred <- predict(model, newdata=data_test)
confusionMatrix(pred,data_test$Survived,positive="No")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  109  23
##        Yes  28  62
##
##                Accuracy : 0.7703
##                  95% CI : (0.7093, 0.8239)
##     No Information Rate : 0.6171
##     P-Value [Acc > NIR] : 8.475e-07
##
##                   Kappa : 0.5192
##
##  Mcnemar's Test P-Value : 0.5754
##
##             Sensitivity : 0.7956
##             Specificity : 0.7294
##          Pos Pred Value : 0.8258
##          Neg Pred Value : 0.6889
##              Prevalence : 0.6171
##          Detection Rate : 0.4910
##    Detection Prevalence : 0.5946
##       Balanced Accuracy : 0.7625
##
##        'Positive' Class : No
##
```

Las métricas obtenidas de la matriz de confusión son buenas, en especial la sensibilidad = 79.56% que corresponde a los pasajeros que el modelo ha clasificado como que no sobrevivieron respecto al total de pasajeros que no sobrevivieron.

Ahora mostramos la curva ROC (*receiver operating characteristic*) que relaciona la sensibilidad con la tasa de falsos positivos.




```
auc(roc_curve)
```

```
## Area under the curve: 0.8559
```

Como el área bajo la curva es de 0.8559, cercana a 1, el modelo es preciso y con gran valor de diagnóstico. Tiene gran capacidad de clasificar correctamente si un pasajero sobrevive o no, dado el género, la clase del billete, la edad y el número de hermanos y parejas también en el barco.

4.3.3 Random Forest

Un Random Forest es un conjunto o *ensamble* de árboles de decisión combinados con *bagging*. Es decir, en cada árbol se utiliza una porción de los datos de entrenamiento y ningún árbol mira exactamente los mismos datos que otro árbol, compensando los errores de cada uno. Esta capacidad de entrenarse permite al modelo generalizar mejor y evitar el sobreajuste.

Como sucedía con la regresión logística, habíamos probado un modelo inicial con todas las variables pero se descartaron aquellas con una importancia inferior a 20, dejando dentro *Sex*, *Pclass*, *Age* y *Fare*.

Random Forest es implementa mediante el método "rf" en *train*.

```
set.seed(666)

data_rf <- titanic[,c(1:4,7)]
h<-holdout(data_rf$Survived,ratio=0.75,mode="stratified")
data_train<-data_rf[h$tr,]
data_test<-data_rf[h$ts,]

table(data_train$Survived)
```

```
##
##  No Yes
## 412 255
```

```
table(data_test$Survived)
```

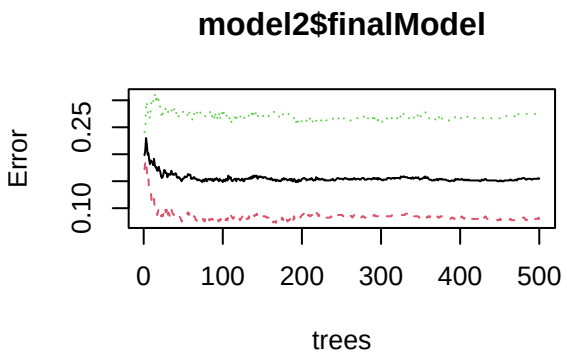
```
##
##  No Yes
## 137  85
```

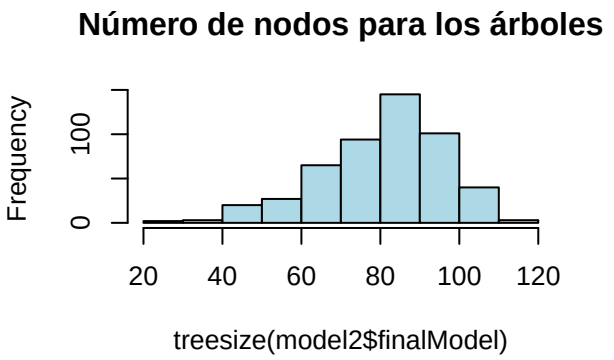
```
train_control<- trainControl(method="cv", number=10)
model2<-train(Survived~., data=data_train,  method="rf", trControl = train_control)
model2$finalModel
```

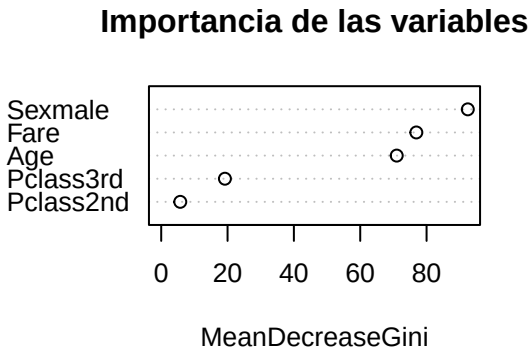
```
##
## Call:
##  randomForest(x = x, y = y, mtry = min(param$mtry, ncol(x)))
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 15.44%
## Confusion matrix:
##      No Yes class.error
## No  379  33  0.08009709
## Yes  70 185  0.27450980
```

La tasa de error durante el entrenamiento es del 15.44%, concretamente el 8.00% para los no supervivientes y el 27.45% para los supervivientes.

El primero de los tres gráficos siguientes muestra el error de la clase `No` (rojo), de la clase `Yes` (verde) y las muestras a usar (negro) sobre la cantidad de los árboles. El segundo muestra el número de nodos usados en los árboles de decisión. Y el segundo la importancia de las variables en el *random forest*.





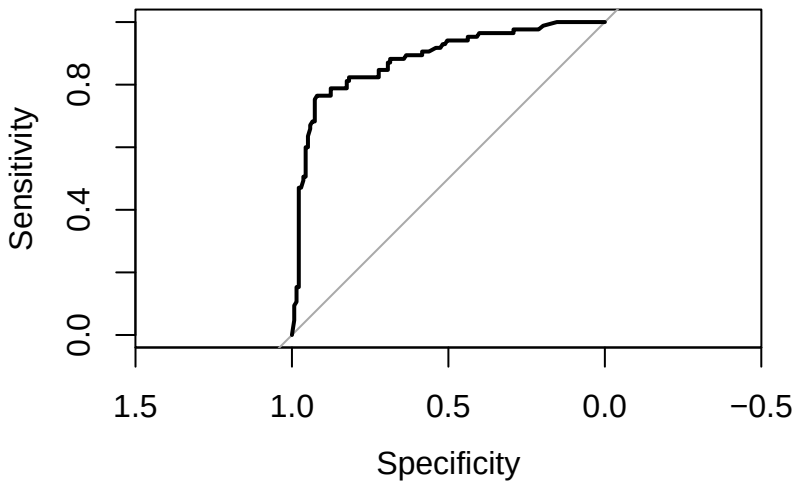


La evaluación del modelo queda de la siguiente manera:

```
pred <- predict(model2, newdata=data_test)
confusionMatrix(pred,data_test$Survived,positive="No")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  118  18
##        Yes  19  67
##
##                Accuracy : 0.8333
##                  95% CI : (0.7777, 0.8799)
##     No Information Rate : 0.6171
##     P-Value [Acc > NIR] : 1.786e-12
##
##                   Kappa : 0.6481
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8613
##             Specificity : 0.7882
##          Pos Pred Value : 0.8676
##          Neg Pred Value : 0.7791
##              Prevalence : 0.6171
##          Detection Rate : 0.5315
##    Detection Prevalence : 0.6126
##       Balanced Accuracy : 0.8248
##
##        'Positive' Class : No
##
```

Este modelo obtiene una *accuracy* del 83.33% y una sensibilidad del 86.13%, unas métricas superiores a las obtenidas con el modelo de regresión logística.




```
auc(roc_curve)
```

```
## Area under the curve: 0.8824
```

El modelo de *randomforest* tiene gran capacidad de clasificar correctamente si un pasajero sobrevive o no, dado el género, la clase del billete, la edad y el precio del billete.

5 Resolución del problema

A largo del análisis hemos ido conociendo más el dataset y sacando conclusiones que resumiremos a continuación.

Durante la selección de datos logramos entender que representaban cada una de las columnas y así poder descartar aquellas que no iban a aportar ningún valor en el análisis posterior.

Más adelante relaizamos la limpieza de los datos donde realizamos la detección de outliers en las variables continuas y concluimos que estos outliers no eran errores que tuviéramos que descartar o modificar.

En la parte fundamental de la práctica, el análisis de datos, tomamos la variable *Survived* como la variable objetivo y sobre la que vamos a centrar nuestro análisis. Con un primer anásls visual observamos que variables como *Sex* o *Fare* si tienen una clara influencia a la hora de discernir si un pasajero sobrevivió o no.

Estudiamos también las variables continuas y concluimos que no siguen distribuciones normales por lo que en los tests de homocedasticidad no debemos suponer normalidad. Después de realizar estos tests concluimos que para *Age* sí hay igualdad de varianzas en los conjuntos de supervivientes y no supervivientes, pero no con *Fare*.

En los contrastes de hipótesis comprobamos que los supervivientes pagaron billetes más caros y que existe una relación entre las variables *Sex* y *Survived* como habíamos visto en el análisis visual.

Finalmente decidimos aplicar dos modelos de aprendizaje automático para determinar la importancia de las variables y obtenemos conclusiones al resto de los análisis donde vemos que las variables *Age*, *Pclass*, *Sex* y *Fare* son las variables más importantes. Con modelo de Random Forest hemos obtenido mejores métricas que con la regresión logística.

Tras todas estas conclusiones podemos decir que hemos conseguido responder al problema ya que hemos obtenido bastante información sobre qué criterios se siguieron a la hora de intentar salvar a los pasajeros y además hemos construido modelos capaces de predecir si un pasajero sobrevivió o no.

6 Contribuciones al trabajo

Contribuciones	Firma
Investigación previa	Jorge SV y Álvaro LC
Redacción de las respuestas	Jorge SV y Álvaro LC
Desarrollo del código	Jorge SV y Álvaro LC