1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El conjunto de datos se ha obtenido del portal Kaggle mediante el siguiente enlace: https://www.kaggle.com/c/titanic

El hundimiento del RMS Titanic es uno de los naufragios más dramáticos de la historia. El 15 de abril de 1912, el Titanic se hundió pereciendo 1502 personas de un total de 2224 pasajeros y tripulantes. Esta tragedia conmocionó a la comunidad internacional y conllevo un gran cambio con respecto a los protocolos de seguridad de los buques.

Una de las razones por las que el naufragio provocó tal pérdida de vidas fue que no había suficientes botes salvavidas para los pasajeros y la tripulación. Aunque hubo cierto componente de suerte, la verdad es que ciertos grupos de personas tenías más probabilidades de sobrevivir que otros.

Mediante el siguiente análisis vamos a clasificar que grupos tenían una mayor probabilidad de supervivencia.

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

El conjunto de datos (dataset) está divido en dos partes:

1. Conjunto de entrenamiento que es utilizado para construir los modelos de aprendizaje.
   train.csv

2. El conjunto de prueba, este se debe usar para ver qué tan bien se desempeña su modelo en datos no vistos.
   test.csv

Existe un tercer archivo: *gender_submission.csv,* trata de un conjunto de predicciones, a modo de ejemplo de cómo se realizaría en el caso de participar en la competición Kaggle. Contiene el atributo “Survived” extraído del conjunto de test.

El conjunto de entrenamiento tiene 891 registros y el de prueba 418, lo que hace un total de 1.309 registros y 12 variables.

```
`train` <- read.csv("train.csv")
`test` <- read.csv("test.csv")
`gs` <- read.csv("gender_submission.csv")
```

Aunque ya tenemos el dataset dividido, los unificaremos con tal de aplicar las diferentes técnicas al conjunto total.

```
test_total <- merge(test, gs, by="PassengerId")
dataset <- rbind(train,test_total)
rm(train,test,gs,test_total)
attach(dataset)
```

Visualicemos las variables y el tipo:

```
library(knitr)
res <- sapply(dataset,class)
kable(data.frame(variables=names(res),clase=as.vector(res)))
```

Nombre	Clase	Descripción
PassengerId	Integer	Identificador único para cada pasajero
Survived	Integer	Variable binaria que indica si sobrevivió
Pclass	Integer	Se indica la clase que pertenecía el pasajero {1 = 1st, 2 = 2nd, 3 = 3rd}
Name	Factor	Nombre del pasajero
Sex	Factor	Sexo {male "hombre", female "mujer"}
Age	Numeric	Edad del pasajero
SibSp	Integer	Cónyuges a bordo o con relación de parentesco
Parch	Integer	Niños a bordo del buque
Ticket	Factor	Nº de ticket de compra
Fare	Numeric	Coste tarifa
Cabin	Factor	Nº de cabina / habitación
Embarked	Factor	Puerta de embarque {C = Cherbourg, Q = Queenstown, S = Southampton}

Prescindiremos de las variables que no aportan nada al estudio, como son:

- **PassengerId** el identificador no es necesario, ya que no aporta ningún dato con respecto al pasajero.
- **Name** el nombre del pasajero tampoco aporta información que sea de interés.
- **Ticket** es el código del ticket y al igual que el nombre no es relevante.
- **Cabin** hay demasiados valores ausentes en esta variable.
- **Fare** aunque este atributo presenta pocos valores ausentes están muy correlacionados con *Pclass* por lo que utilizaremos esta última.
- **Embarked** La puerta de embarque tampoco es un dato que se necesite para el estudio.

```
dataset <- dataset[, -c(9:12)]
dataset$PassengerId <- NULL
dataset$Name <- NULL
```

Convertiremos algunos tipos de datos que no están en el formato adecuado.

Survived Se convierte a *factor*

Pclass Se convierte a *factor*

Age Se convierte a *integer* (prescindimos de los decimales)

```
dataset$Survived <- as.factor(ifelse(dataset$Survived == 1, 'Sobrevive', 'Fallece'))
dataset$Pclass <- as.factor(Pclass)
dataset$Age <- as.integer(Age)
```

3. Limpieza de los datos.

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

Recordamos que los ceros se refieren al valor 0, mientras que los datos están vacíos cuando no hay información, pero esto no quiere decir que hagan referencia al valor 0.

```
sapply(dataset, function(x)(sum(is.na(x))))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch
##        0        0        0      263        0        0
```

Por ejemplo, en el caso de las variables Age*, SibSP y Parch;* El **0** hace referencia al valor. En el caso de la edad son bebes que todavía no tienen el año.

Vemos que la variable *Age* presenta 263 valores ausentes marcados como *NA*. Como tenemos dos grupos diferenciados de datos, hombres y mujeres, trataremos los datos ausentes como la medida central del grupo correspondiente. Es decir, si la edad que falta corresponde a una mujer, se modificará con la media de la edad para el grupo de mujeres, y viceversa para los hombres.

```
media_hombre <- mean (Age[Sex=='male'],na.rm=TRUE)
media_mujer <- mean (Age[Sex=='female'],na.rm=TRUE)
```

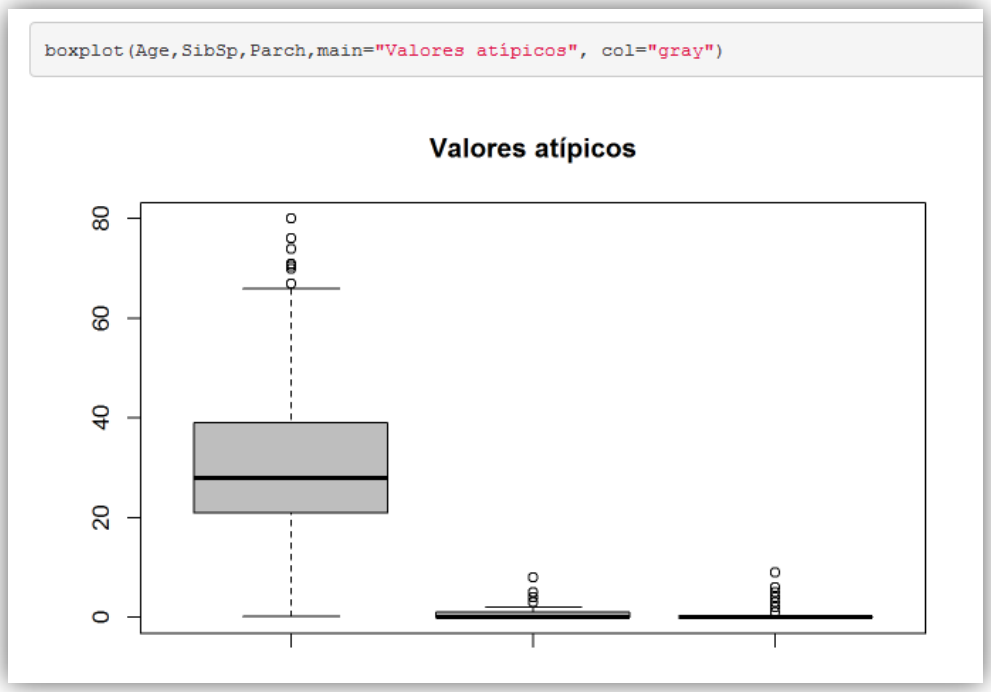
Ahora sustituimos los valores ausentes por la media de los respectivos grupos:

```
dataset$Age[is.na(dataset$Age) & Sex=='male' ] <- media_hombre
dataset$Age[is.na(dataset$Age) & Sex=='female' ] <- media_mujer
```

3.2. Identificación y tratamiento de valores extremos.

Con tal de visualizar los valores extremos realizaremos diagramas de cajas para visualizarlos de forma rápida.

```
boxplot(Age,SibSp,Parch,main="Valores atípicos", col="gray")
```



Vemos que las tres variables cuantitativas tienen valores extremos, representados por los círculos.

Se revisa con mayor detalle los valores atípicos, por si proceden de errores de transcripción o simplemente valores no comunes fuera de la media.

Los valores atípicos presentes en la edad, no son parte de un error, es simplemente que encontramos pasajeros de edad más avanzada que el resto.

```
boxplot.stats(Age)$out
```

```
## [1] 71.0 70.5 71.0 80.0 70.0 70.0 74.0 67.0 76.0
```

```
boxplot.stats(SibSp)$out
```

```
##  [1] 3 4 3 3 4 5 3 4 5 3 3 4 8 4 4 3 8 4 8 3 4 4 4 4 8 3 3 5 3 5 3 4 4 3 3
## [36] 5 4 3 4 8 4 3 4 8 4 8 3 4 5 3 4 8 4 8 4 3 3
```

```
boxplot.stats(Parch)$out
```

```
##   [1] 1 2 1 5 1 1 5 2 2 1 1 2 2 2 1 2 2 2 3 2 2 1 1 1 1 2 1 1 2 2 1 2 2 2 1
##  [36] 2 1 1 2 1 4 1 1 1 1 2 2 1 2 1 1 1 2 1 1 2 2 2 1 1 2 2 1 2 1 1 1 1 1 1
##  [71] 1 2 1 2 2 1 1 2 1 1 2 1 1 1 1 2 1 1 1 4 1 1 2 2 2 2 2 1 1 1 2 2 1 1 2
## [106] 2 3 4 1 2 1 1 2 1 2 1 2 1 1 2 2 1 1 1 1 2 2 2 2 2 2 1 1 2 1 4 1 1 2 1
## [141] 2 1 1 2 5 2 1 1 1 2 1 5 2 1 1 1 2 1 6 1 2 1 2 1 1 1 1 1 1 1 3 2 1 1 1
## [176] 1 2 1 2 3 1 2 1 2 2 1 1 2 1 2 1 2 1 1 1 2 1 1 2 1 2 1 1 1 1 3 2 1 1 1
## [211] 1 5 2 1 1 1 1 3 1 2 2 1 2 1 2 1 2 4 1 1 2 1 1 1 4 6 2 3 1 1 2 2 2 1 1
## [246] 2 5 2 3 2 1 1 1 2 1 2 2 2 1 2 1 1 2 1 2 1 2 1 2 2 1 1 1 1 1 2 1 1 2 1
## [281] 1 1 2 1 2 9 1 1 1 2 2 2 1 9 1 1 2 2 1 1 2 1 1 1 1 1 1
```

Con respecto a la variable *Parch* aunque las familias eran mucho más numerosas que las de ahora, tener 9 hijos a bordo del mismo barco parece algo poco probable, seguramente serían 6, como en otros casos, pero el número se escribió del revés.

Y lo mismo ocurre con la variable *SibSp,* vemos que tenemos varios 8 que están lejos del siguiente valor 5, estos números a veces, pueden confundirse.

El tratamiento de estos valores extremos consistirá en convertirlos al valor más cercano probable.

```
dataset$SibSp[SibSp==8] <- 5
dataset$Parch[Parch==9] <- 6
```

4. Análisis de los datos.

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

Ya al principio se ha realizado una preselección, descartando varios atributos que no nos eran útiles.

Ahora, existen varios grupos o estrategias a analizar, siendo el atributo *Survived*, nuestra clase objetivo.

- La clase a la que pertenecía el pasaje fue un factor determinante, distinguiendo las 3 posibles opciones.
- La edad se categorizará en bebés, niños, adolescentes, jóvenes, adultos y tercera edad.
- Distinción de hombres y mujeres
- Existencia de correlación entre las variables numéricas *SibSp y Parch.* Relevancia del tamaño de la familia. Posible creación de nueva variable conjunta de tipo factor.

Categorización de la edad en intervalos:

```
dataset$Edad <- cut(dataset$Age, breaks = c(-1, 2, 12, 17, 29, 55, 100),labels = c("bebe", "niño", "adolescente
", "joven","adulto","3edad"))
```

Para la correlación de las variables *SibSp* y *Parch* antes deberemos comprobar si es necesario realizar una normalización (comprobación que se realizará en el siguiente punto).

Selección de datos para realizar posteriormente un contraste de hipótesis.

```
S <- dataset[,"Sex"]
A <- dataset[,"Age"]
Edad_mujeres <- A[S=="female"]
Edad_hombres <- A[S=="male"]
```

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

Se realizar un contraste de hipótesis, con un nivel de significación del 0.05. Se aplicará varias pruebas de normalización incluidas en la librería *nortest.*

```
H0: La muestra proviene de una distribución normal.      (hipótesis nula)
H1: La muestra no proviene de una distribución normal.   (hipótesis alternativa)

Aceptaremos H0, si z >= alpha
Rechazamos H0, si z < alpha
```

```
###Prueba de Anderson-Darling###
ad.test(Age)
```

```
##
##  Anderson-Darling normality test
##
## data:  Age
## A = 7.1299, p-value < 2.2e-16
```

```
###Pruena de Lilliefors (Kolmogorov-Smirnov)###
lillie.test(Age)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Age
## D = 0.078928, p-value < 2.2e-16
```

```
###Prueba de Pearson chi-square###
###basada en una distribución Ji cuadrado y que corresponde a una prueba de bondad de ajuste.###
pearson.test(Age)
```

```
##
##  Pearson chi-square normality test
##
## data:  Age
## P = 217.34, p-value < 2.2e-16
```

```
###Prueba de Shapiro-Wilk###
###Test muy potente para el contraste de normalidad, sobre todo para muestras pequeñas (n<30)
shapiro.test(Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Age
## W = 0.97955, p-value = 5.747e-11
```

Se observa que todos los valores p-value son inferiores a *alpha* por lo que debemos rechazar la hipótesis nula, por lo que, podemos afirmar que la variable no sigue una distribución normal.

Ahora aplicaremos el test Shapiro-Wilk al resto de variables:

```
###Prueba de Shapiro-Wilk###
###Test muy potente para el contraste de normalidad, sobre todo para muestras pequeñas (n<30)
shapiro.test(SibSp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  SibSp
## W = 0.51108, p-value < 2.2e-16
```

```
###Prueba de Shapiro-Wilk###
###Test muy potente para el contraste de normalidad, sobre todo para muestras pequeñas (n<30)
shapiro.test(Parch)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Parch
## W = 0.49797, p-value < 2.2e-16
```

```
#Variable Survived
shapiro.test(Survived)
```

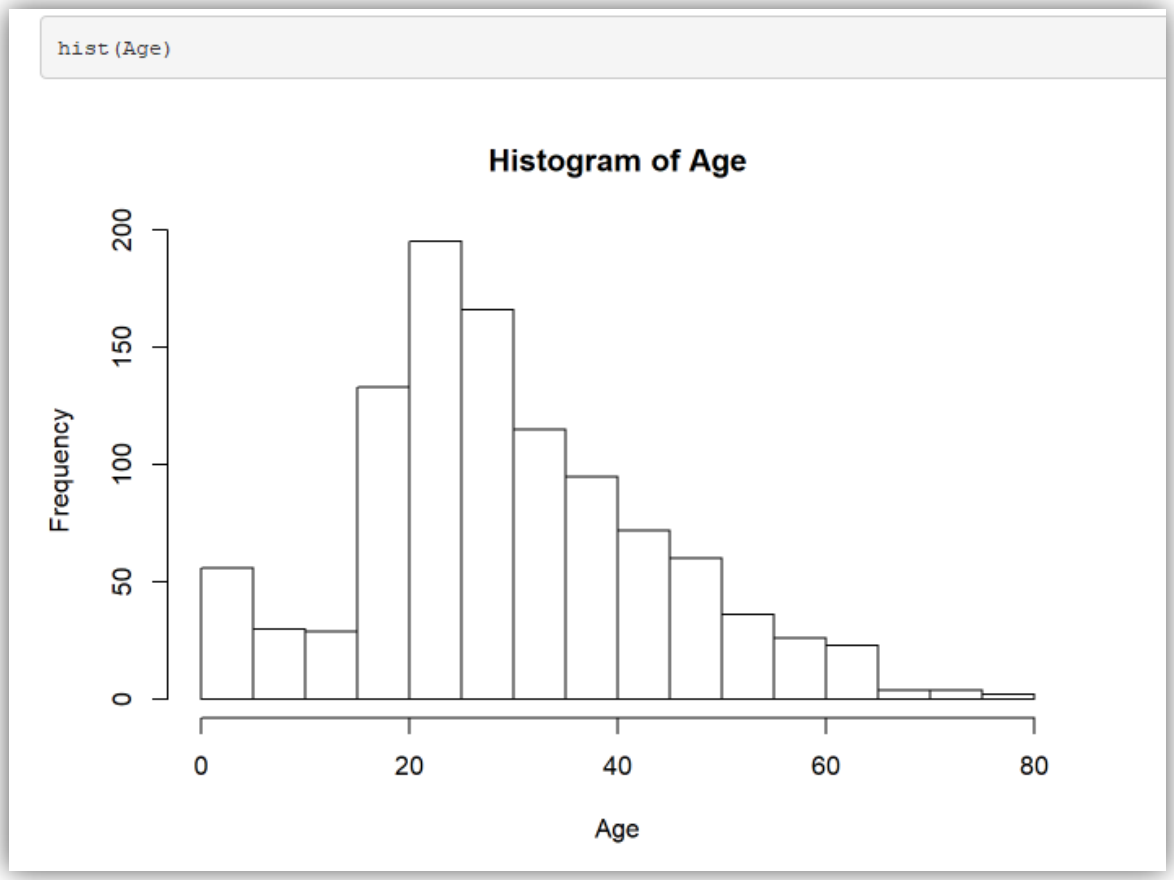
```
##
##  Shapiro-Wilk normality test
##
## data:  Survived
## W = 0.61436, p-value < 2.2e-16
```

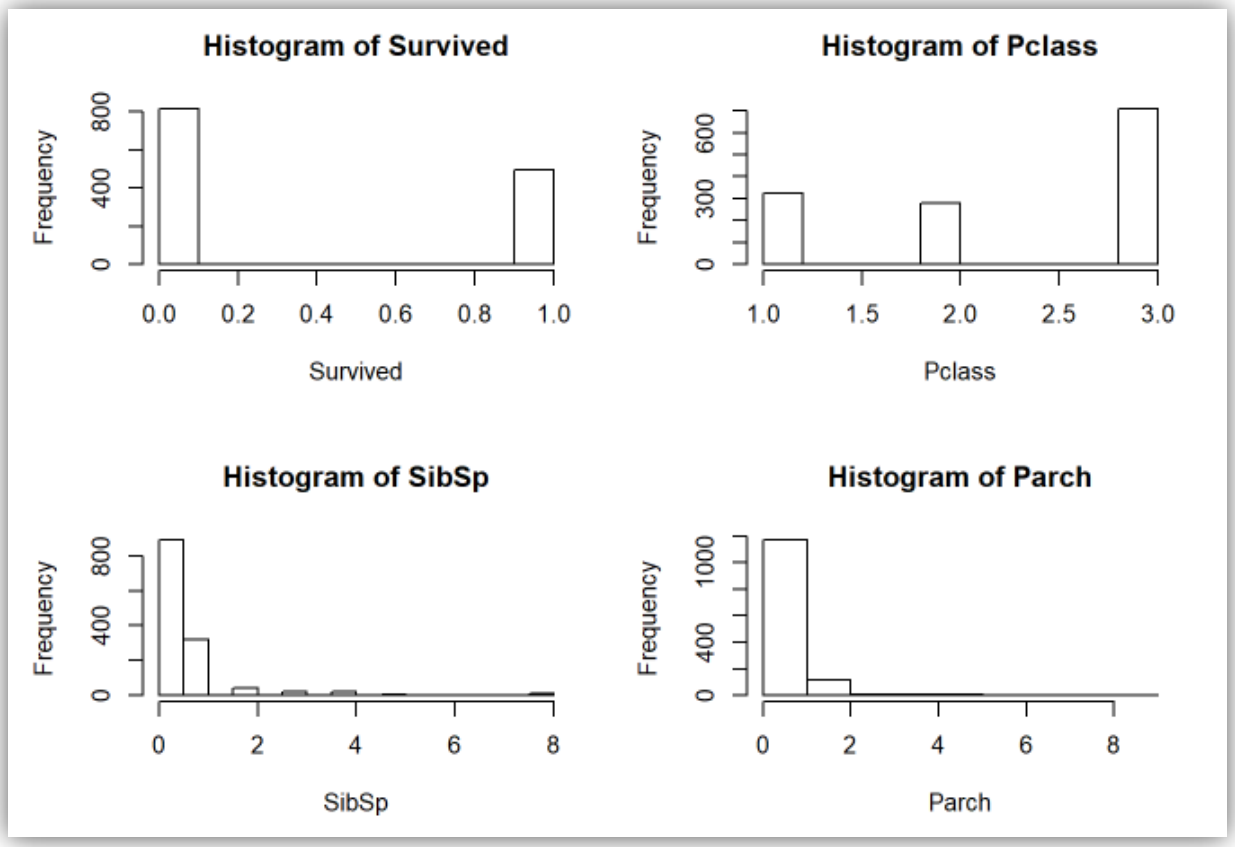
```
#Variable Pclass
shapiro.test(Pclass)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Pclass
## W = 0.72337, p-value < 2.2e-16
```

Para todas ellas se observa que los valores p-value respectivos, también están por debajo del 0.05 y, por lo tanto, en estos casos también se rechazará la hipótesis nula.

```
hist(Age)
```





Con respecto a las variables SibSp y Parch podrían normalizarse, pero en vez de ello se va a proceder categorizar ambas, creando una única variable que corresponderá al tamaño de la familia a bordo del transatlántico.

```
familySize <- dataset$SibSp + dataset$Parch + 1
familySizeClass = array(dim = length(familySize))
familySizeClass[familySize == 1] = 'Small'
familySizeClass[familySize >= 2 & familySize <= 4] = 'Medium'
familySizeClass[familySize > 4] = 'Big'
dataset$Familia <- as.factor(familySizeClass)
```

El supuesto de homogeneidad de varianzas, también conocido como supuesto de homocedasticidad, considera que la varianza es constante (no varía) en los diferentes niveles de un factor, es decir, entre diferentes grupos.

Hay varios test que permiten evaluar la distribución de la varianza, pero mucho de ellos funcionan bien o sólo lo hacen bajo distribuciones normales, por ejemplo: F-test, Test de Bartlett, Test de Brown-Forsyth. En el caso de que nos cumpla la normalidad tenemos el test no paramétrico de Fligner-Killeen o también, el test de Levene.

Mientras que el primero está restringido a la comparación de dos varianzas (y, por lo tanto, es útil al realizar una prueba t), el segundo se usa en conjunto con ANOVA donde deben haber más de dos grupos comparado.

Aunque las variables *Parch* y *Sibsp* han sido categorizadas en una sola, las volveremos utilizar con tal de aplicar el test de homogeneidad con respecto a la variable *Survived.*

```
fligner.test(Parch ~ Survived, data=dataset)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Parch by Survived
## Fligner-Killeen:med chi-squared = 27.368, df = 1, p-value =
## 1.682e-07
```

```
fligner.test(SibSp ~ Survived, data=dataset)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SibSp by Survived
## Fligner-Killeen:med chi-squared = 5.64, df = 1, p-value = 0.01756
```

Para las características *Parch* y *SibSp* tenemos que el p-valor es muy pequeño por lo que directamente descartamos la hipótesis nula para estas variables, lo que significa que las varianzas entre ellas no son homogéneas.

En cambio, para la Edad con respecto a la supervivencia y La característica *SibSp* presente un p-valor más alto (0.017) pero, incluso así, debemos descartar la hipótesis nula ya que el nivel de significación estaba fijado en 0.05.

En cambio, para la variable *Age* con respecto las distribuciones *Sex*, la varianza sí que es homogénea.

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Sex
## Fligner-Killeen:med chi-squared = 3.2115, df = 1, p-value =
## 0.07312
```

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

Para crear el modelo, ahora sí, volvemos a dividir el dataset en un conjunto de entrenamiento y otra prueba. El grupo de entrenamiento contendrá el 80% de las observaciones, mientras que el de prueba el 20% del resto.

```
set.seed(1234)
temp <-sample(nrow(dataset), 0.80*nrow(dataset))
train <- dataset[temp,]
test <- dataset[-temp,]
rm(temp)
```

¿Hasta qué punto influyen las variables *Sex* y *Pclass* sobre la supervivencia?

```
table(train$Sex, train$Pclass, train$Survived)
```

```
## , ,  = Fallece
##
##
##             1   2   3
##   female    3   6  60
##   male    105 127 361
##
## , ,  = Sobrevive
##
##
##             1   2   3
##   female  108  78 111
##   male     37  15  36
```

Vemos que para los hombres el tipo de clase al que pertenecían poco les afectaba, pues por el hecho de serlo reducía sus posibilidades de sobrevivir.

En cambio, ser mujer prácticamente era un salvo conducto a no ser que perteneciera a 3era clase, en tal caso, las posibilidades de fallecer aumentaban.

```
##
## Call:
## glm(formula = Survived ~ Sex + Pclass, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3361  -0.5647  -0.3668   0.5534   2.3378
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.6611     0.2297  11.584  < 2e-16 ***
## Sexmale      -3.5542     0.1948 -18.242  < 2e-16 ***
## Pclass2      -0.8623     0.2541  -3.394 0.000688 ***
## Pclass3      -1.7724     0.2254  -7.862 3.78e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1377.29  on 1046  degrees of freedom
## Residual deviance:  809.51  on 1043  degrees of freedom
## AIC: 817.51
##
## Number of Fisher Scoring iterations: 5
```

Como se puede ver por los p-valor, de estas dos variables son muy buenos para explicar la variabilidad de la supervivencia.

Veamos ahora las variable Age (Edad) y Familia (variable creada con el tamaño de familia).

```
##
## Call:
## glm(formula = Survived ~ Familia + Edad, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5377  -0.8934  -0.7347   1.1883   1.9877
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.1039     0.5084  -2.171   0.0299 *
## FamiliaMedium      1.7592     0.3415   5.151 2.59e-07 ***
## FamiliaSmall       0.6544     0.3465   1.889   0.0589 .
## Edadniño           0.1608     0.5292   0.304   0.7612
## Edadadolescente   -0.1006     0.5450  -0.185   0.8536
## Edadjoven         -0.2629     0.4550  -0.578   0.5633
## Edadadulto        -0.7223     0.4533  -1.593   0.1111
## Edad3edad         -0.6809     0.5573  -1.222   0.2218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1377.3  on 1046  degrees of freedom
## Residual deviance: 1281.0  on 1039  degrees of freedom
## AIC: 1297
##
## Number of Fisher Scoring iterations: 4
```

Para estas variables los p-valor son peores por lo que no explican tan bien nuestro modelo como las anteriores. Esto también lo vemos con el criterio de información de Akaike (AIC) que es de 1297, superior al anterior de 817. Aunque está la excepción del tamaño de familia mediano y pequeño que, en estos casos, sí que es una característica que puede cambiar la probabilidad de supervivencia.

Utilizaremos un bosque aleatorio (RandomForest) para concluir cuáles son las variables más prometedoras:

```
Control <- trainControl(method = 'cv', number = 10, classProbs = TRUE, summaryFunction = twoClassSummary)
rfModel <- train(Survived~., data=train, trControl=Control, method = 'rf', tuneLength = 10, metric = 'ROC')
```

```
varImp(rfModel)
```

```
## rf variable importance
##
##                  Overall
## Sexmale         100.0000
## Pclass3           9.3723
## FamiliaMedium     7.0373
## FamiliaSmall      3.7672
## Edadniño          1.9591
## Edadadulto        1.8126
## Edadjoven         1.5209
## Pclass2           1.3609
## Edad3edad         0.1621
## Edadadolescente   0.0000
```

Se observa como la característica *sex* es la que tiene una mayor importancia a la hora de determinar nuestra clase objetivo, el resto de variables no parecen ser tan explicativas, esto concuerda con lo visto en la regresión logística.

Entonces, sabemos que el género de la persona es un factor determinante para las probabilidades de supervivencia. Por otro lado, la edad es un factor que influye, esto nos hace plantearnos. **¿Podemos afirmar que la media de edad de las mujeres es menor a la de los hombres?**

Para estar totalmente seguro de ello indicaremos un nivel de confianza del 99% (α= 0.01)

```
H0: Hombres y mujeres tienen la misma media de edad.     (hipótesis nula)
H1: Las mujeres son más jóvenes que los hombres.  (hipótesis alternativa)

Aceptaremos H0, si z >= alpha
Rechazamos H0, si z < alpha
```

Según el teorema del límite central, dada cualquier variable aleatoria X con media μ, si el tamaño de las muestras consideradas es n>30, entonces la variable se comporta como una distribución normal estándar. Por tanto, como el tamaño de muestra es suficientemente grande, se puede considerar la distribución normal. Realizaremos un contraste de una muestra para la media. Acorde con la definición de la hipótesis alternativa, usaremos un contraste unilateral.

```
t.test(Edad_mujeres,Edad_hombres,alternative='less')
```

```
##
##  Welch Two Sample t-test
##
## data:  Edad_mujeres and Edad_hombres
## t = -2.4788, df = 917.25, p-value = 0.006681
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.6281029
## sample estimates:
## mean of x mean of y
##  28.56223  30.43298
```

Vemos que el p-valor es de 0.006681 es inferior al α= 0.01 por lo que podemos rechazar la hipótesis nula de que las edades entre hombres y mujeres son iguales a favor de la hipótesis alternativa.

Para acabar y aunque no es objeto de esta práctica, podríamos ver qué tal predice nuestro modelo utilizando para ello un árbol de decisión y ver si era correcto lo descrito con anterioridad:

```
arbol <- C5.0(x=train[2:5],y=train$Survived)
summary(arbol)
```

```
##
## Read 1047 cases (5 attributes) from undefined.data
##
## Decision tree:
##
## Sex = male: Fallece (681/88)
## Sex = female:
## :...Familia in {Medium,Small}: Sobrevive (331/47)
##     Familia = Big:
##     :...Pclass in {1,2}: Sobrevive (6)
##         Pclass = 3: Fallece (29/7)
##
##
## Evaluation on training data (1047 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##       4  142(13.6%)   <<
##
```

Obtenemos una tasa de error del 13.6% para el entrenamiento.

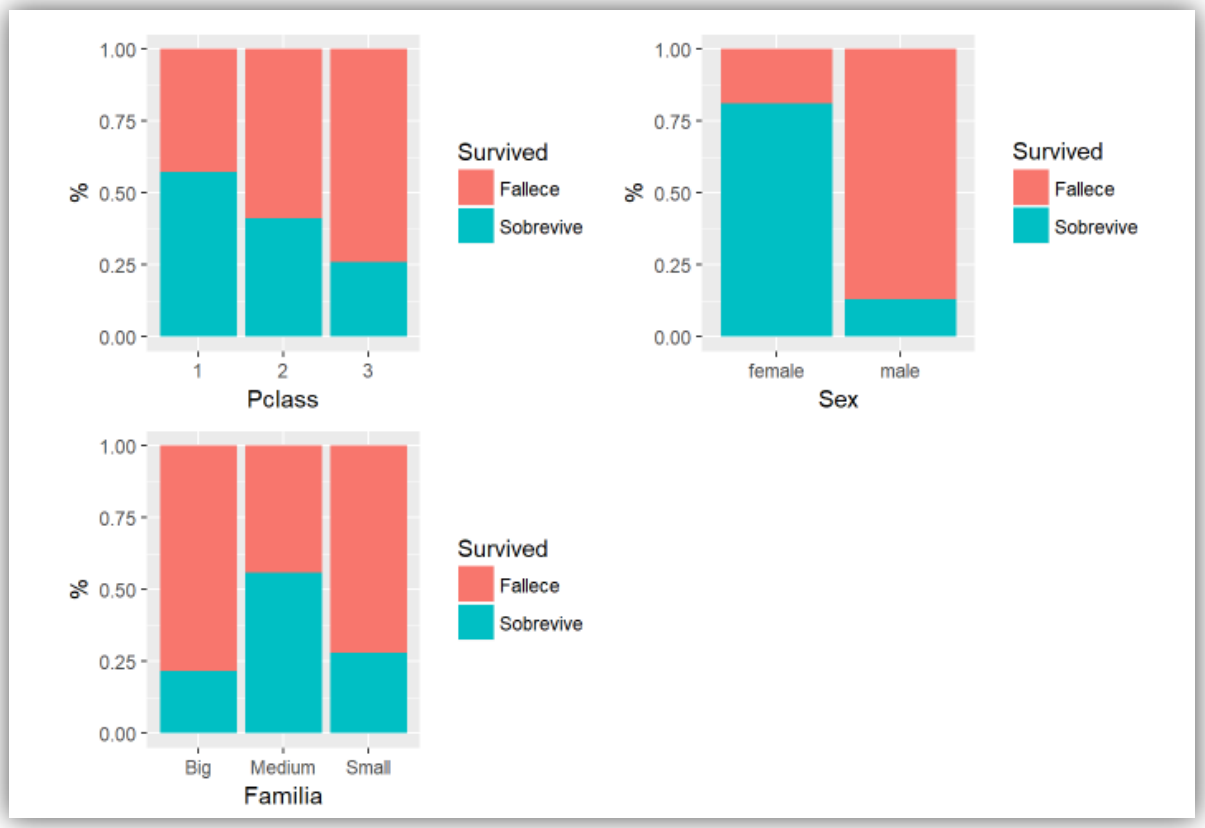
Y qué tal predice el modelo.

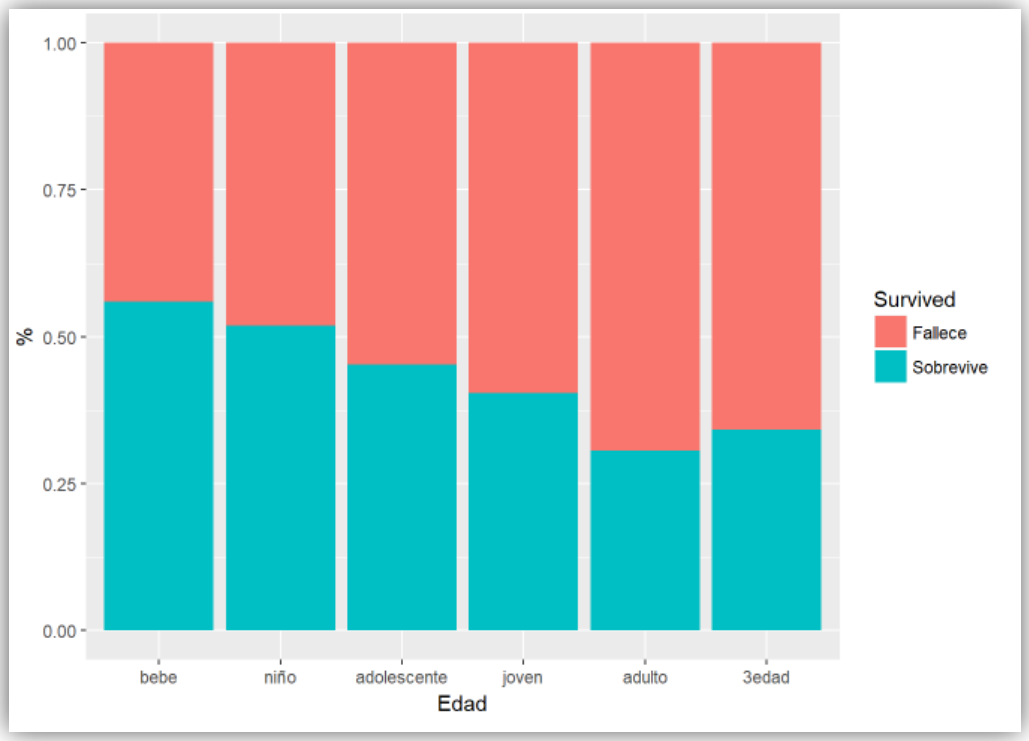
```
p <- predict(arbol,test,type="class")
confusionMatrix(data = p, reference = test$Survived)
```

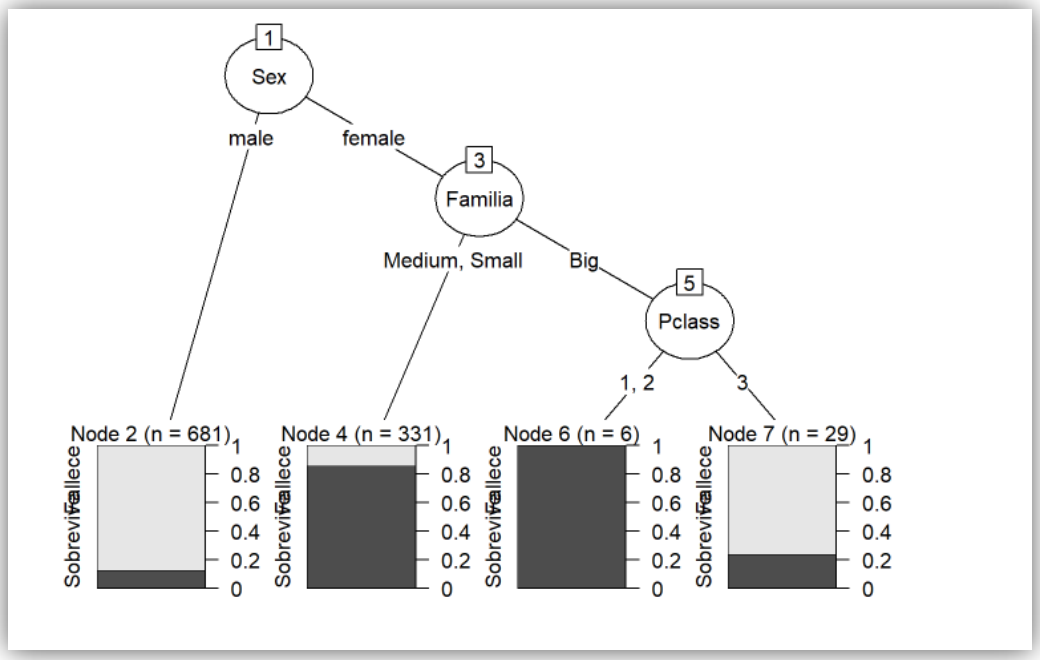
```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Fallece Sobrevive
##   Fallece       143        22
##   Sobrevive      10        87
##
##                Accuracy : 0.8779
##                  95% CI : (0.832, 0.9149)
##     No Information Rate : 0.584
##     P-Value [Acc > NIR] : < 2e-16
##
```

Vemos que predice correctamente, con una precisión del 87% (erró en 32 casos)

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




1.00
0.75
0.50
0.25
0.00
%
bebe
niño
adolescente
joven
adulto
3edad
Edad
Survived
Fallece
Sobrevive


1
Sex
male
female
3
Familia
Medium, Small
Big
5
Pclass
1, 2
3
Node 2 (n = 681)
Node 4 (n = 331)
Node 6 (n = 6)
Node 7 (n = 29)
Sobrevive
Fallece
1
0.8
0.6
0.4
0.2
0

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

Creo que los resultados han sido satisfactorios. Se ha contrastado mediante varios métodos que el atributo *Sex* es el que tiene una mayor influencia a la hora de predecir la supervivencia de un pasajero, seguido de la clase y edad.

Para el atributo *Edad* se han tratado los valores ausentes de forma que no se han perdido observaciones. Se ha aceptado, bajo contraste de hipótesis, que las mujeres son más jóvenes que los hombres, aunque las dos variables están correlacionadas no significa que una sea explicativa de lo otra.

Con respecto a los valores extremos, tan sólo se ha tratado los casos de las variables *SibSp* y *Parch,* como estas dos variables hacían referencia a un mismo concepto, se ha realizado una selección quedando una única característica denominada *Familia.*

Varios casos de ejemplo con respecto a los resultados obtenidos:

- Una niña o chica adolescente de primera clase tenía unas probabilidades muy altas de ser superviviente.
- Un hombre adulto de segunda o tercera clase seguramente estaría entre los fallecidos.
- Las mujeres de tercera clase con familia numerosas también tenían más probabilidades de fallecer.
- Las personas que viajaban solas o tenían poca familia tenían más posibilidades de morir.