Tipología y ciclo de vida de los datos: Práctica 2

Limpieza y análisis de datos

Jorge Alonso Hernández e Inés Vidal Sospedra

6 de June, 2022

Contents

---

1 Descripción de la Práctica a realizar

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

Algunos ejemplos de dataset con los que podéis trabajar son:

- Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)
- Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

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

Importante: si se elige un dataset diferente de los propuestos es importante que este contenga una amplia variedad de datos numéricos y categóricos para poder realizar un análisis más rico y poder responder a las diferentes preguntas planteadas en el enunciado de la práctica.

---

2 Solución

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

2.1 Importación del dataset

En el presente trabajo utilizaremos el conjunto de datos "Titanic" que recoge datos sobre el famoso crucero y sobre el que es fácil realizar tareas de clasificación predictiva sobre la variable "Survived".

Concretamente, los datos que se utilizan en este estudio se encuentran disponibles en Kaggle (https://www.kagg le.com/c/titanic). Estos están organizados en dos archivos de formato csv:

- **'train .csv'** con todos los datos de una parte de los pasajeros,
- **'test.csv'** con todos los datos de los pasajeros restantes sin el atributo que indica si sobrevivieron o no.

Este hecho se debe a que el primer archivo ('train.csv') está pensado para estudiar la relación entre las diferentes características de los pasajeros y sobrevivir o no, mientras que el segundo ('test.csv') está pensado en predecir si los pasajeros sobrevivirán o no en función de las observaciones hechas en el primero.

Dicho esto, procedemos a importar los datasets de los dos archivos para empezar a describir los datos que contienen.

```
# Import libraries needed for the whole study.
library(ggplot2)
library(ggpubr)
library(car)
```

```
# Import train and test data from the csv files.
d_train <- read.csv("../CSV_Inicial/train.csv", sep = ',', stringsAsFactors = TRUE)
d_test <- read.csv("../CSV_Inicial/test.csv", sep = ',', stringsAsFactors = TRUE)
```

2.2 Descripción del dataset

Una vez importados, verificamos la estructura del conjunto de datos:

```
# Explore train data.
str(d_train)
```

```
## '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       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 58
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ 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     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

El primer dataset (d_train) contiene los datos de entrenamiento que se utilizarán para estudiar a los pasajeros y crear el modelo. En él se puede observar que hay un total de 891 registros y 12 variables.

```
# Explore test data
str(d_test)
```

```
## 'data.frame':    418 obs. of  11 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
##  $ Name       : Factor w/ 418 levels "Abbott, Master. Eugene Joseph",..: 210 409 273 414 182 370 85 5
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : Factor w/ 363 levels "110469","110489",..: 153 222 74 148 139 262 159 85 101 270 ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : Factor w/ 77 levels "","A11","A18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 2 3 2 3 3 3 2 3 1 3 ...
```

El segundo dataset (d_test) contiene los datos de prueba o test y cuenta con un total de 418 registros y 11 variables. Esta variable de menos con respecto al primer dataset se debe a que no se dispone de la información referente a la supervivencia del pasajero ya que esta es la variable a predecir.

A cotinuación se describen las variables contenidas en ambos datasets:

- *PassengerId*: variable numérica que contiene el identificador único de cada pasajero.
- *Survival*: variable categórica que indica si el pasajero sobrevive o no (0=No,1=Yes).
- *Pclass*: variable numérica que indica la clase en la que viaja el pasajero (1=1st,2=2nd, 3=3rd).
- *Name*: variable categórica que indica el nombre del pasajero. Hay tantas categorías como registros.
- *Sex*: variable categórica que indica el género del pasajero (male, female).
- *Age*: variable numérica que indica la edad del pasajero, en años.
- *SibSp*: variable numérica que indica el número de hermanos/as y esposo/a que el pasajero tiene a bordo del Titánico.
- *Parch*: variable numérica que indica el número de padres e hijos que el pasajero tiene a bordo del Titanic.

- *Ticket*: variable categórica que indica el identificador de ticket del pasajero. Contiene 363 categorías diferentes.
- *Fare*: variable numérica que indica el precio del ticket del pasajero.
- *Cabin*: variable categórica que indica el identificador de la cabina del pasajero.
- *Embarked*: variable categórica que indica el puerto por el que embarcó el pasajero (C=Cherbourg,Q=Queenstown,S=Sou

Observando estas variables, se puede apreciar la importancia de éstas para poder dar respuesta al objetivo de este estudio (¿el pasajero sobrevive?) y encontrar si existe algún factor, o factores, que determinen la supervivencia de los pasajeros en el accidente del Titanic.

2.3 Integración y selección de los datos de interés a analizar

Para realizar el análisis sobre qué pasajeros sobrevivieron y cuáles no, se considera que no todas las variables contenidas en el dataset de entrenamiento son de interés.

Por ello, cabe mencionar que la información que contienen las variables 'Ticket', 'Fare' y 'Cabin' es muy variada y se entiende que la información de interés que pueden contener está contenida en la variable 'Pclass', puesto que el número de ticket, su precio y la cabina dependerán de si el pasajero viaja en primera, segunda o tercera clase.

También las variables 'PassengerId' y 'Name', siendo únicas para cada persona, se decide no seleccionarlas para su posterior análisis.

Por último, las variables 'SibSp' y 'Parch' se han considerado que la única información importante que contienen es si el pasajero viajaba solo o en familia. Así pues, a partir de estas dos variables se crea una nueva llamada "Family". Esta variable será categórica y contendrá dos valores 'Yes' en caso de que el pasajero viaje en familia (que mínimo una de las viariable 'SibSp' o 'Parch' sea diferente a 0), o 'No' en caso de que viaje solo (que ambas variables sean 0).

```
# Create the new variable 'Family'.
d_train$Family <- d_train$Pclass
n <- 1
while (n <= length(d_train$Family)) {
  if(d_train$SibSp[n]==0 && d_train$Parch[n]==0){
    d_train$Family[n] <- 'No'
  }else{
    d_train$Family[n] <- 'Yes'
  }
  n <- n+1
}

d_train$Family <- as.factor(d_train$Family)
head(d_train$Family,n=20)
```

```
##  [1] Yes Yes No  Yes No  No  No  Yes Yes Yes Yes No  No  Yes No  No  Yes No  Yes
## [20] No
## Levels: No Yes
```

```
# Convert 'Survived' as factor.
d_train$Survived <- as.factor(d_train$Survived)
head(d_train$Survived,n=20)
```

```
##  [1] 0 1 1 1 0 0 0 0 1 1 1 1 0 0 0 1 0 1 0 1
## Levels: 0 1
```

```
# Convert 'Pclass' as factor.
d_train$Pclass <- as.factor(d_train$Pclass)
head(d_train$Pclass,n=20)
```

```
##  [1] 3 1 3 1 3 3 1 3 3 2 3 1 3 3 3 2 3 2 3 3
## Levels: 1 2 3
```

Con la variable nueva 'Family' creada, se procede a seleccionar las variables de interés para el análisis, siendo éstas las siguientes: 'Survived','Pclass','Sex','Age','Family' y 'Embarked'.

```
# Create a new dataset with the selected data.
dataset <- d_train[c('Survived','Pclass','Sex','Age','Family','Embarked')]
head(dataset, n=5)
```

```
##   Survived Pclass    Sex Age Family Embarked
## 1        0      3   male  22    Yes        S
## 2        1      1 female  38    Yes        C
## 3        1      3 female  26     No        S
## 4        1      1 female  35    Yes        S
## 5        0      3   male  35     No        S
```

2.4 Limpieza de los datos

Antes de empezar con el análisis, se debe realizar una limpieza de los datos a utilizar para asegurar que éstos son correctos y para que los resultados que se obtengan con estos datos también lo sean.

Por ello, en primer lugar se visualizan los atributos seleccionados para tener una primera idea de los datos con los que se trabajará.

```
#Estadísticas básicas
summary(dataset)
```

```
##  Survived Pclass      Sex           Age        Family    Embarked
##  0:549    1:216   female:314   Min.   : 0.42   No :537    :  2
##  1:342    2:184   male  :577   1st Qu.:20.12   Yes:354   C:168
##           3:491                Median :28.00             Q: 77
##                                Mean   :29.70             S:644
##                                3rd Qu.:38.00
##                                Max.   :80.00
##                                NA's   :177
```

En estos datos se puede observar cómo existe la presencia de 177 valores vacíos en la única variable numérica del dataset 'Age', y en la variable 'Embarked' hay dos registros con un espacio en blanco. Sin embargo, observando los valores estadísticos en Age, parece que puede haber valores extremos en edades grandes.

2.4.1 Valores nulos

```
# Estadísticas de valores vacíos
colSums(is.na(dataset)) #Suma los NA de cada columna
```

```
## Survived   Pclass      Sex      Age   Family Embarked
##        0        0        0      177        0        0
```

```
# Estadísticas de valores vacíos
colSums(dataset=="") #Suma los campos vacíos de cada columna
```

```
## Survived   Pclass      Sex      Age   Family Embarked
##        0        0        0       NA        0        2
```

Revisando la presencia de valores vacíos nos encontramos con 177 valores vacíos el la columna de 'Age' por lo que para no entorpecer el análisis posterior aplicaremos el valor de la media de la edad a estos valores nulos.

```
# Tomamos la media para valores vacíos de la variable "Age"
dataset$Age[is.na(dataset$Age)] <- mean(dataset$Age,na.rm=T)
```

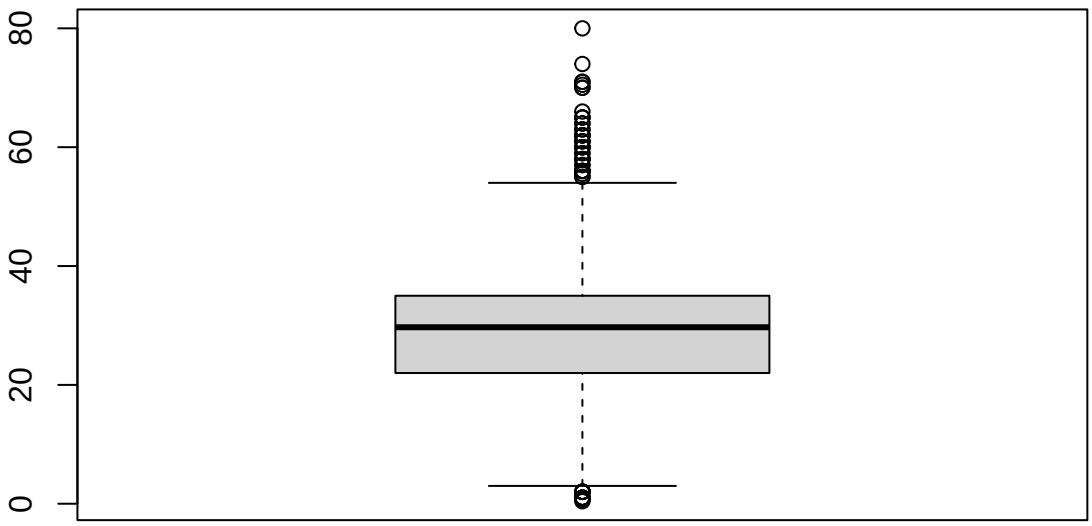
En cuanto a los 2 registros con un espacio en blanco se ha decidido eliminar las dos filas que contienen dichos valores faltantes.

```
dataset <- dataset[!is.na(dataset$Embarked),]
```

2.4.2 Valores extremos (outliers)

Siendo 'Age' la única variable numérica, es la única que puede tener valores atípicos. Por ello, revisamos los valores extremos que pudieran existir en la variable Age, utilizando un diagrama de caja (boxplot).

```
# Diagrama de caja de las edades
boxplot(dataset$Age)
```



Podemos ver que los valores extremos que se han detectado se refieren a las edades entre los 60 y los 80 años, por lo que no los considerarmeos como extremos ya que son edades válidas.

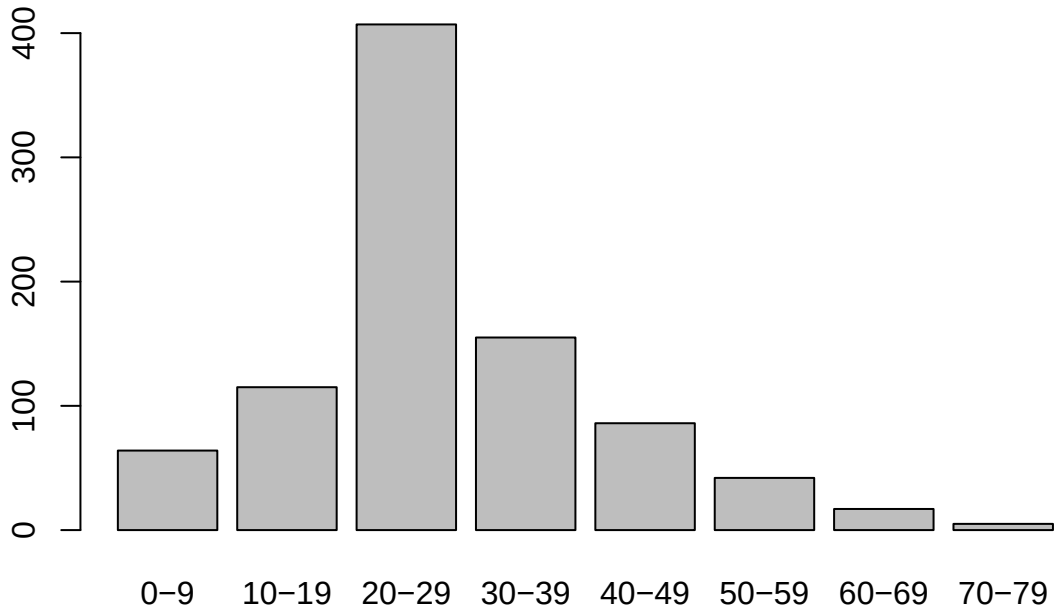
2.4.3 Discretizar datos

Para una mejor análisis de los datos realizaremos la discretización de la variable Age de forma que obtengamos una nueva variable con el rango de edad de los pasajeros en intervalos.

```
dataset["rango_edad"] <- cut(dataset$Age, breaks = c(0,10,20,30,40,50,60,70,100), labels = c("0-9", "10-
```

Visualizamos los resultados de la nueva variable obtenida.

```
plot(dataset$rango_edad)
```



Podemos apreciar que la mayoría de los pasajeros pertenecen al rango de edad entre los 20 y los 29 años.

Guardar dataset

```
write.csv2(dataset, row.names = TRUE, "../CSV_Finales/train_clean.csv")
```

2.5 Análisis de los datos

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

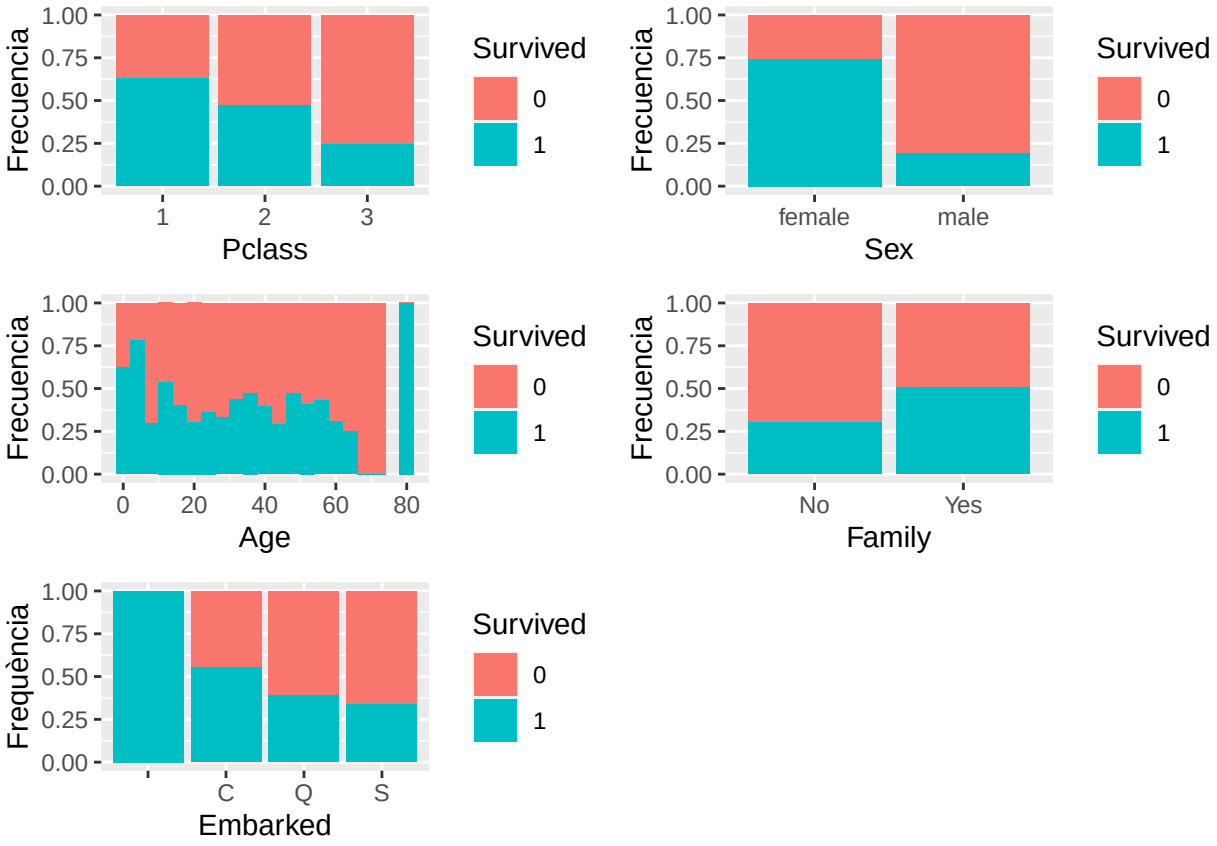
Como primer análisis sobre el conjunto de datos, es interesante observar cómo están repartidos los pasajeros entre supervivientes y no supervivientes, respecto al resto de variables seleccionadas para el estudio.

Por eso, se realizan representaciones gráficas sobre cada variable, separando para cada grupo entre supervivientes y no supervivientes. Al ser prácticamente todo variables categóricas, las representaciones se realizan en gráficos de barras. En el caso de la variable 'Age', se representan en un histograma, agrupando las columnas en intervalos de 4.

```
layout(matrix(c(1,2,
                3,3,
                4,5), 3,3, byrow = TRUE))
Pclass_plot <- ggplot(data = dataset,aes(x=Pclass,fill=Survived)) +
geom_bar(position="fill") + ylab("Frecuencia")
Sex_plot <- ggplot(data = dataset,aes(x=Sex,fill=Survived)) +
geom_bar(position="fill") + ylab("Frecuencia")
Age_plot <- ggplot(data = dataset,aes(x=Age,fill=Survived)) +
```

```
geom_histogram(binwidth = 4,position="fill") + ylab("Frecuencia")
Family_plot <- ggplot(data = dataset,aes(x=Family,fill=Survived)) +
geom_bar(position="fill") + ylab("Frecuencia")
Embarked_plot <- ggplot(data = dataset,aes(x=Embarked,fill=Survived)) +
geom_bar(position="fill") + ylab("Frequència")

figure <- ggarrange(Pclass_plot,
          Sex_plot,
          Age_plot,
          Family_plot,
          Embarked_plot,
          ncol = 2, nrow = 3)
figure
```



De los gráficos obtenidos se puede comentar que para la variable:

- Pclass: para los pasajeros de primera clase, el porcentaje de supervivientes es muy superior al 50%, mientras que para los de segunda se sitúa ligeramente por debajo del 50%, significando que hay más supervivientes de segunda clase. En cuanto a los pasajeros de tercera clase, el número de supervivientes no llega a 1/4 del total. Pues, a mejor clase, las probabilidades de sobrevivir parecen aumentar.
- Sex: la diferencia de relación entre supervivientes respecto a hombres y mujeres muestra una diferencia evidente. El porcentaje de supervivientes por mujeres es de casi un 75%, mientras que por varones no llega al 20%.

En este caso, no cabe duda de que el hecho de ser mujer es un factor muy importante para sobrevivir.

- Age: para la edad de los pasajeros, se observa una clara mayoría de supervivientes por menores de 10 años, mientras que para mayores de 60 predominan los no supervivientes. Para el resto de edades, la relación de supervivientes-no va variando entre 30/45-70/55, siendo en todos los casos mayoría de no supervivientes. Ésta variable parece ser relevante en cuanto a la supervivencia por las edades extremas.
- Family: para los pasajeros en familia, se observa que la relación entre supervivientes es más favorable (prácticamente del 50-50) que para los que no, donde en estos pasajeros que viajan solos hay una clara mayoría de no supervivientes. Así pues, parece que los pasajeros en familia deben tener más éxito en sobrevivir.
- Embarked: para los pasajeros que han embarcado en Cherbourg muestran una relación de supervivencia de 45-55, ganando a los supervivientes, pero en los casos de Queenstown y Southampton se observa una relación similar, donde los supervivientes representan menos del 40%. Este factor parece no ser demasiado importante en la hora de determinar la supervivencia de los pasajeros, aunque en un punto de embarque se aprecia una mejor relación en cuanto a la supervivencia

2.5.2 Comprobación de la normalidad y homogeneidad de la varianza.

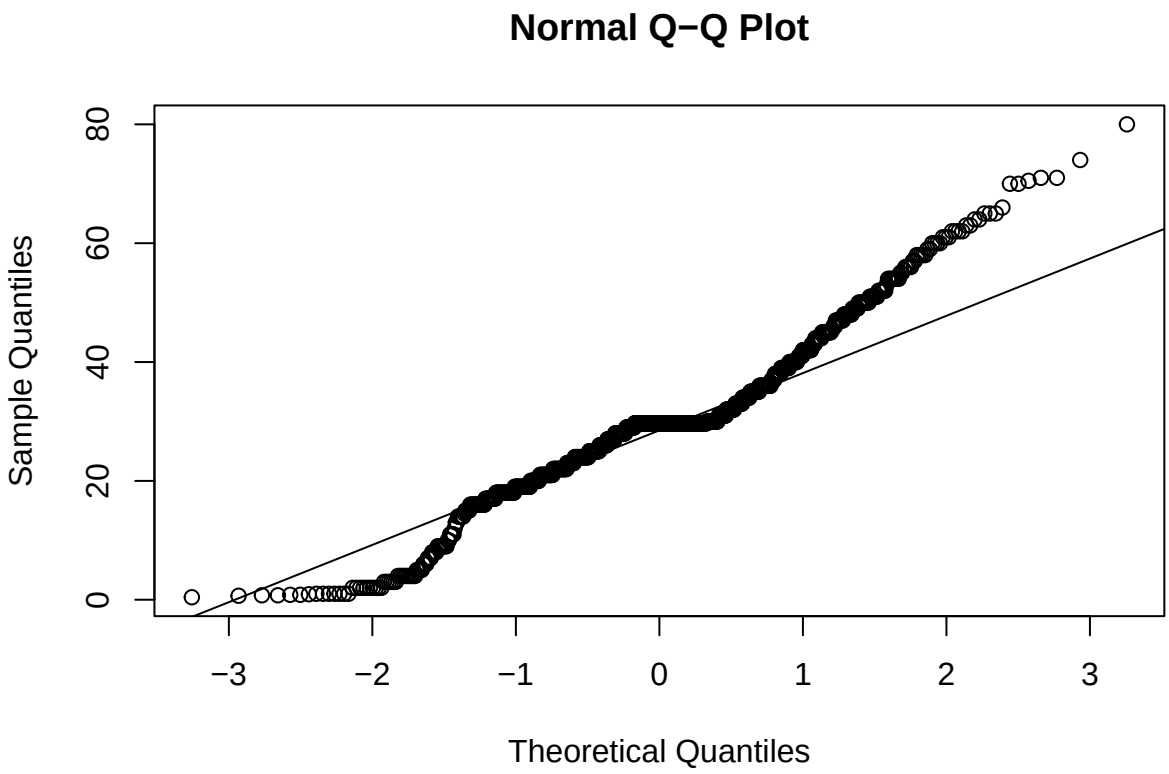
Para poder realizar un contraste de hipótesis sobre la media poblacional en la variable Age, primero es necesario estudiar si la distribución de estos datos es normal y si la varianza es homogénea (homocedasticidad), es decir, si la varianza entre los pasajeros supervivientes y no puede considerarse igual en ambos grupos.

Empezando por comprobar si la variable 'Age' sigue una distribución normal, se aplica el test de Shapiro-Wilk sobre esta variable, así como se representa el gráfico Q-Q

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

```
##
##  Shapiro-Wilk normality test
##
## data:  dataset$Age
## W = 0.95882, p-value = 3.969e-15
```

```
qqnorm(dataset$Age)
qqline(dataset$Age)
```

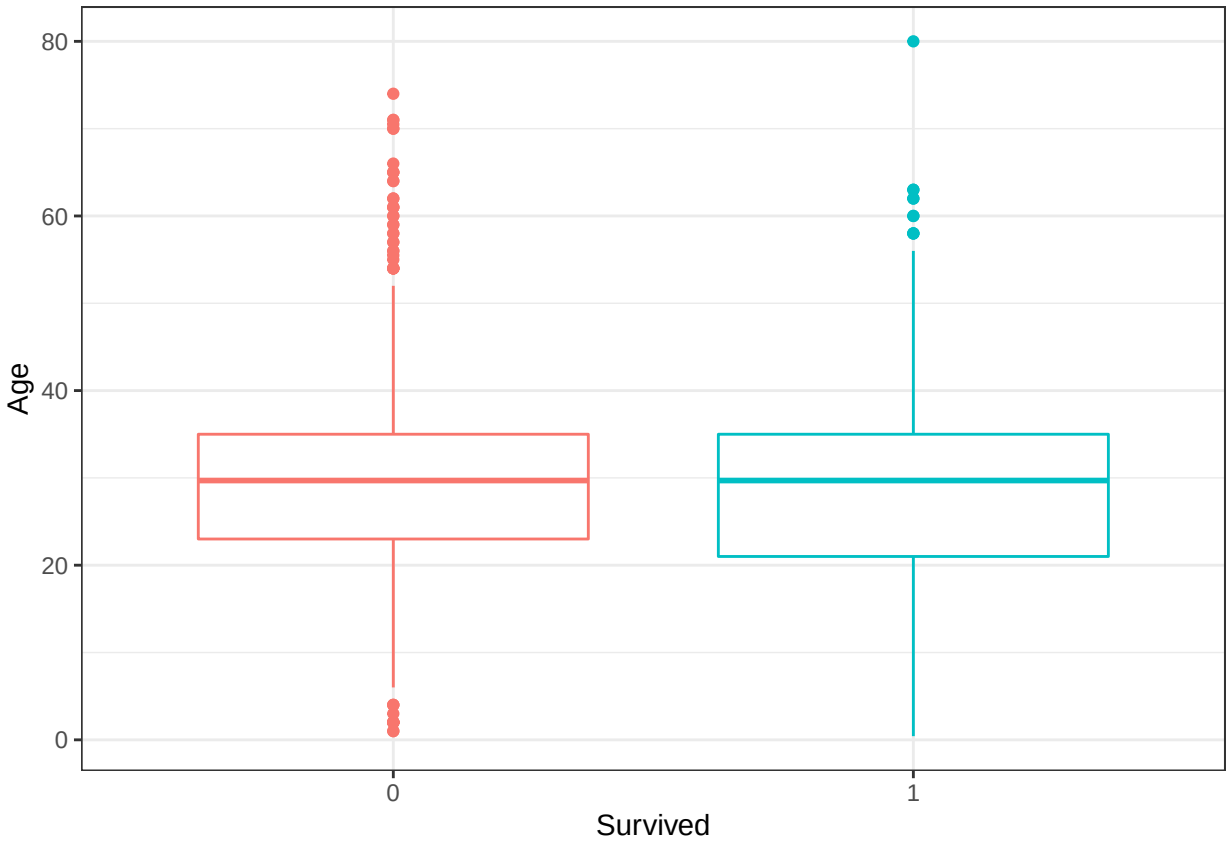


El test Shapiro-Wilk da un p-value de $3{,}969 \cdot 10{-}15$ (muy inferior a 0,05), por lo que se rechaza la hipótesis de que la variable 'Age' sigue una distribución normal. Además, en la gráfica Q-Q, los puntos se alejan significativamente de la recta.

Sabiendo que los datos sobre la edad de los pasajeros no se distribuyen de forma normal, puede procederse a comprobar la homogeneidad de la varianza en esta misma variable 'Age'.

Para estudiar la homocedasticidad de los datos sobre la edad, primero se representan los diagramas de cajas para ambos grupos, para tener una idea de si las muestras entre estos dos grupos se distribuyen de la misma modo. Y para tener una respuesta definitiva a saber si las varianzas son iguales, teniendo en cuenta que la variable 'Age' no sigue una distribución normal, se aplica el test de Levene sobre la mediana.

```
ggplot(data = dataset, aes(x = Survived, y = Age, colour = Survived)) +
geom_boxplot() +
theme_bw() +
theme(legend.position = "none")
```




```
leveneTest(y = dataset$Age, group = dataset$Survived, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value  Pr(>F)
## group   1  5.4815 0.01944 *
##       889
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observando la posición de la mediana en los diagramas de cajas respecto a los cuartiles, se aprecia una diferencia significativa entre ambos grupos. Y finalmente, en el test de Levene, aunque por poco, el p-value obtenido de 0,019 es inferior al nivel de significación 0,05. Esto implica que se rechaza la hipótesis nula de que las dos varianzas son iguales, y por tanto, las varianzas entre las edades de los dos grupos (supervivientes y no supervivientes) son distintas.

2.6 Pruebas estadísticas

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

Con lo expuesto en los puntos anteriores, se ha decidido que una vez realizado el análisis básico sobre los datos, se puede definir el plan a seguir para determinar qué pasajeros son más probables de sobrevivir.

Para ello, realizaremos

- Contraste de hipótesis sobre la media poblacional de la edad entre los pasajeros supervivientes y los no supervivientes.
- Contraste de hipótesis sobre la proporción de supervivientes entre pasajeros hombres y pasajeras mujeres
- Cálculo de un modelo de regresión logística tomando la supervivencia como variable a explicar y el resto como variables explicativas.

2.6.1 Contraste de hipótesis sobre la media poblacional de la edad entre los pasajeros supervivientes, y los no supervivientes.

En primer lugar, para poder realizar el estudio obtenemos las muestras de las edades para los supervivientes y los no supervivientes

```
supervivientes <- dataset$Age[dataset$Survived == 1] # Datos de supervivientes
noSupervivientes <- dataset$Age[dataset$Survived == 0] # Datos de los no supervivientes
```

Definimos la hipótesis nula y la hipótesis alternativa:

- **Hipótesis nula:** el promedio de las edades de los supervivientes es igual al promedio de edades de los no supervivientes

$$H_0 : m_1 = m_2$$

- **Hipótesis alternativa:** el promedio de las edades de los supervivientes es mayor a el promedio de las edades de los no supervivientes

$$H_1 : m_1 > m_2$$

Antes de aplicar el estadístico del contraste realizaremos el test de homocedasticidad para comprobar si las varianzas son iguales:

```
alfa <- 1 - 0.95
H <- supervivientes
D <- noSupervivientes
mean1 <- mean(H) # media supervivientes
n1 <- length(H) # número de supervivientes
s1 <- sd(H) # desviación típica supervivientes
mean2 <- mean(D) # media no supervivientes
n2 <- length(D) # número de no supervivientes
s2 <- sd(D) # desviación típica no supervivientes
fobs <- s1^2/s2^2
fcritL <- qf(alfa, df1=n1-1, df2=n2-2)
fcritU <- qf(1 - alfa, df1=n1-1, df2=n2-2)
pvalue <- min(pf(fobs, df1=n1-1, df2=n2-2, lower.tail=FALSE), pf(fobs, df1=n1-1, df2=n2-2))*2
c(fobs, fcritL, fcritU, pvalue)
```

```
## [1] 1.22228551 0.85007863 1.17228514 0.03772666
```

Realizamos la comprobación del resultado con la función en R var.test.

```
var.test(H, D)
```

```
##
##  F test to compare two variances
##
## data:  H and D
## F = 1.2223, num df = 341, denom df = 548, p-value = 0.03765
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.011499 1.483399
## sample estimates:
## ratio of variances
##           1.222286
```

Como podemos observar los resultados coinciden y se obtiene un p-valor inferior al alfa definido de 0.05 por lo que podemos asumir que las varianzas son diferentes.

Al estar ante unas varianzas diferentes el estadístico a aplicar será el contraste de dos muestras sobre la media con varianzas diferentes.

```
alfa <- 1-0.95
dfMean = mean1 -mean2
v <- ((s1^2/n1)+(s2^2/n2))^2 / (((s2^2/n1)^2/(n1-1)) + ((s2^2/n2)^2/(n2-1)))

tobs <- dfMean/sqrt((s1^2/n1 + s2^2/n2))
tcrit <- qt(alfa, v)
pvalue <- pt(abs(tobs), df=v, lower.tail=FALSE)*2
c(tobs, tcrit, pvalue)
```

```
## [1] -2.03851720 -1.64648455  0.04177907
```

Realizamos la comprobación de los cálculos con la función en R t.test

```
t.test(H,D)
```

```
##
##  Welch Two Sample t-test
##
## data:  H and D
## t = -2.0385, df = 669.03, p-value = 0.04189
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.66201421 -0.06862884
## sample estimates:
## mean of x mean of y
##  28.54978  30.41510
```

Podemos observar de nuevo que coinciden los resultados y que tenemos un p-valor inferior al alfa definido de 0.05 por lo que rechazamos la hipótesis nula de que el promedio de las edades de los supervivientes es igual a el promedio de edades de los no supervivientes y aceptamos la hipótesis alternativa que establece que el promedio de edad de los supervivientes es mayor al promedio de edad de los no supervivientes.

2.6.2 Contraste de hipótesis sobre la proporción de supervivientes entre pasajeros hombres y pasajeras mujeres

Para poder realizar este estudio en primer lugar obtendremos una muestra de datos para hombres y otra para mujeres de aquellas personas que fueron supervivientes

```
man <- dataset$Survived[dataset$Sex == "male"]
woman <- dataset$Survived[dataset$Sex == "female"]
```

Definimos la hipótesis nula y la hipótesis alternativa

- **Hipótesis nula:** la proporción de muejres que sobrevivieron es igual a la proporción de hombres quesobrevivieron

$$H_0 : p_1 = p_2$$

- **Hipótesis alternativa:** la proporción de muejres que sobrevivieron es diferente a la proproción de hombres que sobrevivieron

$$H_1 : p_1 \neq p_2$$

En este caso aplicamos el estadístico relativo al contraste de la proporción de dos muestras

```
alfa <- 1-0.95
x1 <- woman[woman == 1]  # Mujeres supervivientes
x2 <- man[man == 1] # Hombres supervivientes

n1 <- length(woman) # número de mujeres en el barco
n2 <- length(man)   # número de hombres en el barco

p1 <- sum(length(x1))/n1 # proporción de muejres supervivientes
p2 <- sum(length(x2))/n2 #proproción de hombres supervivientes

p <- (n1*p1 + n2*p2) / (n1+n2)
zobs <- (p1-p2)/(sqrt((p*(1-p))*(1/n1+1/n2)))
zcrit <- qnorm(alfa, lower.tail = FALSE)
pvalue <- pnorm(zobs, lower.tail=FALSE)

c(p1,p2, n1, n2, length(x1), length(x2))
```

```
## [1]   0.7420382   0.1889081 314.0000000 577.0000000 233.0000000 109.0000000
```

```
c(zobs, zcrit,pvalue)
```

```
## [1] 1.621883e+01 1.644854e+00 1.855874e-59
```

Realizamos la comprobación empleando la función en R prop.test.

```r
success <- c(p1*n1, p2*n2)
nn <- c(n1,n2)
prop.test(success, nn, alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 263.05, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.5044702 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.7420382 0.1889081
```

Obtenemos que en ambos casos el p-valor es inferior al alfa definido de 0.05 por lo que podemos rechazar la hipótesis nula de que la proporción de mujeres supervivientes es igual a la proporción de hombres supervivientes y aceptamos la hipótesis alternativa de que la proporción de mujeres supervivientes es diferente a la proporción de hombres supervivientes con un nivel de confianza del 95%.

2.6.3 Cálculo de un modelo de regresión logística

Cálculo de un modelo de regresión logística tomando la supervivencia como variable a explicar y el resto como variables explicativas.

```r
modelo <- glm(formula = Survived~Pclass+Sex+rango_edad+Family+Embarked,
data = dataset, family = binomial)
summary(modelo)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + rango_edad + Family +
##     Embarked, family = binomial, data = dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6837  -0.6929  -0.3835   0.6350   2.4286
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      16.6048   593.8164   0.028 0.977692
## Pclass2          -0.9310     0.2656  -3.505 0.000457 ***
## Pclass3          -2.1995     0.2523  -8.716  < 2e-16 ***
## Sexmale          -2.6440     0.2006 -13.180  < 2e-16 ***
## rango_edad10-19  -1.2733     0.4225  -3.013 0.002583 **
## rango_edad20-29  -1.5142     0.3721  -4.070 4.71e-05 ***
## rango_edad30-39  -1.2232     0.4014  -3.047 0.002311 **
## rango_edad40-49  -1.8363     0.4464  -4.114 3.89e-05 ***
## rango_edad50-59  -2.1539     0.5507  -3.911 9.19e-05 ***
```

```
## rango_edad60-69  -2.6083     0.7941  -3.285 0.001021 **
## rango_edad70-79  -2.3870     1.2644  -1.888 0.059056 .
## FamilyYes        -0.2113     0.2041  -1.035 0.300690
## EmbarkedC       -12.2069   593.8163  -0.021 0.983599
## EmbarkedQ       -12.2748   593.8164  -0.021 0.983508
## EmbarkedS       -12.8201   593.8163  -0.022 0.982776
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  791.03  on 876  degrees of freedom
## AIC: 821.03
##
## Number of Fisher Scoring iterations: 13
```

Puede observarse que todas las variables explicativas son estadísticamente significativas a un nivel de confianza del 95% menos FamilyYes, EmbarkedQ y rango_edad70-79. Estas tres últimas variables no tienen ningún asterisco al final de la fila, por tanto, esto indica que no representan valor estadístico dentro del modelo.

Sin embargo, para poder interpretar más exhaustivamente la regresión logística debería calcularse e interpretar los odds ratio. Adicionalmente, también se puede comprobar que todos los coeficientes de la regresión son negativos, esto significa que todos ellos afectan negativamente al hecho de sobrevivir el Titanic.

Este modelo de regresión logística obtenido utilizando todas las variables seleccionadas es el que presenta un valor de AIC menor es decir, se han observado los valores de este parámetro en modelos tomando sólo algunas de las variables, y en todos los casos resultaba mayor. Así pues, este último modelo es el que presenta un mejor nivel de ajuste.

```
exp(coefficients(modelo))
```

```
##     (Intercept)         Pclass2         Pclass3         Sexmale rango_edad10-19
##    1.627019e+07    3.941434e-01    1.108561e-01    7.107356e-02    2.799175e-01
## rango_edad20-29 rango_edad30-39 rango_edad40-49 rango_edad50-59 rango_edad60-69
##    2.199921e-01    2.942927e-01    1.594076e-01    1.160255e-01    7.365696e-02
## rango_edad70-79       FamilyYes       EmbarkedC       EmbarkedQ       EmbarkedS
##    9.190850e-02    8.095531e-01    4.995785e-06    4.668107e-06    2.705780e-06
```

Los odds ratio de cada una de las variables significativas dentro del modelo indican que:

- Pclass2: por cada pasajero que viaja en segunda clase, el odds de sobrevivir disminuye en un 0.3941
- Pclass3: por cada pasajero que viaja en tercera clase, el odds de sobrevivir disminuye en un 0.1108
- Sexmale: si el pasajero es un hombre, el odds de sobrevivir disminuye en un 0.071
- Rango_edad: de todos los rangos de edad, un pasajero que tiene entre 30-39 años, el odds de sobrevivir disminuye en un 0.2942
- FamilyYes: si el pasajero tiene familia, el odds de sobrevivir disminuye en un 0.8095
- EmbarkedQ: si el pasajero embarcó en el puerto de Queenstown, el odds de sobrevivir disminuye en 0.9344

- EmbarkedS si el pasajero embarcó en el puerto de Southamptons, el odds de sobrevivir disminuye en 0.5416

Una vez interpretados los coeficientes, debe mencionarse que la variable que tiene más peso en este modelo es el sexo, puesto que para el odds ratio es el más pequeño. Por tanto, la probabilidad de sobrevivir cambia más que cualquier otra variable. A continuación se evaluará la precisión del modelo.

```
pred <- ifelse(test = modelo$fitted.values > 0.5, yes = 1, no = 0)
conf_mat <- table(modelo$model$Survived, pred,
                  dnn = c("observations", "predictions"))
conf_mat
```

```
##             predictions
## observations   0   1
##            0 475  74
##            1 101 241
```

```
(475+239)/(475+239+101+74)*100
```

```
## [1] 80.31496
```

Puede concluirse que el modelo de regresión logística tiene una precisión del 80,31%.

2.7 Predicciones sobre los datos de test

2.7.1 Integración y selección de los datos de interés

A continuación se procederá a realizar todos los cambios necesarios para trabajar con el archivo test de los datos con el objetivo de predecir si los pasajeros sobrevivirán o no.

```
# Create the new variable 'Family'.
d_test$Family <- d_test$Pclass
n <- 1
while (n <= length(d_test$Family)) {
  if(d_test$SibSp[n]==0 && d_test$Parch[n]==0){
    d_test$Family[n] <- 'No'
  }else{
    d_test$Family[n] <- 'Yes'
  }
  n <- n+1
}

d_test$Family <- as.factor(d_test$Family)
head(d_test$Family,n=20)
```

```
##  [1] No  Yes No  No  Yes No  No  Yes No  Yes No  No  Yes Yes Yes Yes No  No  Yes
## [20] No
## Levels: No Yes
```

```
# Convert 'Pclass' as a factor.
d_test$Pclass <- as.factor(d_test$Pclass)
head(d_test$Pclass,n=20)
```

```
##  [1] 3 3 2 3 3 3 3 2 3 3 3 1 1 2 1 2 2 3 3 3
## Levels: 1 2 3
```

Se seleccionan las variables de interés para realizar la predicción.

```
d_test <- d_test[c('Pclass','Sex','Age','Family','Embarked')]
head(d_test, n=5)
```

```
##   Pclass    Sex  Age Family Embarked
## 1      3   male 34.5     No        Q
## 2      3 female 47.0    Yes        S
## 3      2   male 62.0     No        Q
## 4      3   male 27.0     No        S
## 5      3 female 22.0    Yes        S
```

Se revisa si existen valores nulos en el dataset test

```
# Estadísticas de valores vacíos
colSums(is.na(d_test)) #Suma los NA de cada columna
```

```
##   Pclass      Sex      Age   Family Embarked
##        0        0       86        0        0
```

```
# Estadísticas de valores vacíos
colSums(d_test=="") #Suma los campos vacíos de cada columna
```

```
##   Pclass      Sex      Age   Family Embarked
##        0        0       NA        0        0
```

Revisando la presencia de valores vacíos nos encontramos con 86 valores vacíos el la columna de 'Age' por lo que para no entorpecer el análisis posterior aplicaremos el valor de la media de la edad a estos valores nulos.

```
# Tomamos la media para valores vacíos de la variable "Age"
d_test$Age[is.na(d_test$Age)] <- mean(d_test$Age,na.rm=T)
```

Del mismo modo que en el caso del dataset train, se procede a discretizar la variable Age en diferentes rangos

```
d_test["rango_edad"] <- cut(d_test$Age, breaks = c(0,10,20,30,40,50,60,70,100), labels = c("0-9", "10-19
```

Una vez preparado el dataset, se procede a obtener la predicción de interés.

2.7.2 Predicción de supervivencia

A continuación se procederá a realizar las predicciones del conjunto de datos de test, utilizando el modelo obtenido.

De este modo, si el resultado es superior o igual a una probabilidad de 0.5 se interpretará como el pasajero sobrevive, de lo contrario, se interpretará que el pasajero no sobrevive.

```
predictions <- predict(object = modelo,
                       newdata=d_test,
                       type = "response")

# Convert the predictions in categories by the rule of separeting in 0.5.
n <- 1
while (n <= length(predictions)) {
  if(predictions[n] >= 0.5){
    predictions[n] = 1
  }else{
    predictions[n] = 0
  }
  n <- n+1
}

head(predictions, n=20)
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  0  0  0  0  0  0  1  0  1  0  0  0  1  0  1  1  0  0  0  1
```

Se comprueba que por cada línea se extrae un valor 0 o 1 correspondiente a si el pasajero determinado no sobrevive (0) o sobrevive (1).

```
d_test["Survived"] <- predictions
```

2.8 Exportación de datos finales

A continuación, vamos a exportar nuestro dataframe final a un archivo .csv. Este archivo se llamará test_result y utilizaremos la función write.csv2() para exportar el fichero en formato csv español.

```
write.csv2(d_test, row.names = TRUE, "../CSV_Finales/test_result.csv")
```

2.9 Conclusiones

Como conclusión al trabajo, queríamos destacar:

- En este trabajo, usamos datos de los pasajeros del Titanic para ver si existe alguna relación entre los pasajeros sobrevivientes y los no sobrevivientes.
- El conjunto de datos inicial contenía un total de 12 variables (PassengerId, Survival, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked), pero se ha decidido trabajar con 6 de ellas (Survival, Pclass , Sex, Age, Family, Embarked), que son aquellas que se cree contienen información relevante para observar si existe relación entre pasajeros supervivientes y no supervivientes. Se ha decidido también crear la variable 'Family', a partir de la unión entre la información que contienen la variable 'SibSp' y la 'Parch', creándose una nueva variable categórica con dos únicos valores posibles 'Yes' o 'No', según si el pasajero viajaba solo (No) o en familia (Yes).
- En el proceso de limpieza de datos se ha detectado valores perdidos en la variable 'Age', los cuales para no entorpecer la posterior análisis se ha aplicado el valor de la media de la edad a estos valores nulos y también valores huecos en la variable Embarked, los cuales, al tratarse de sólo dos registros, se ha decidido eliminarlos.

- Sobre los datos seleccionados se han aplicado diferentes métodos de análisis para observar el efecto de cada variable sobre si un pasajero sobrevive o no. Estos métodos han sido: contraste de hipótesis sobre la media poblacional de la edad entre los pasajeros supervivientes y los no supervivientes y contraste de hipótesis sobre la proporción de supervivientes entre pasajeros hombres y pasajeras mujeres.
- Los resultados obtenidos en los distintos contrastes de hipótesis aplicados han mostrado que todas las variables seleccionadas (Sex, Age) son explicativas a la hora de determinar la supervivencia de un pasajero.
- En cuanto al modelo de regresión logística obtenido, se han utilizado todas las variables explicativas seleccionadas, siendo éste el modelo que muestra un mejor nivel de ajuste. Observando los *odds ratio* del modelo, destaca la variable en lo referente al sexo del pasajero, siendo la que más efecto tiene sobre la probabilidad de supervivencia. El hecho de que un pasajero sea hombre con respecto a que sea mujer, la probabilidad de supervivencia disminuye en un 0,071. La precisión de este modelo de regresión logística es de un 80,31%.
- El hecho de que un pasajero sea hombre o mujer tiene implicaciones importantes para predecir si un pasajero sobrevivirá. De esta forma, se puede confirmar que si la pasajera es mujer, tiene más posibilidades de sobrevivir. Además, las mujeres son menos sensibles a la edad que los hombres. Y, en cuanto a la clase, aquellos que viajan en primera clase tienen más probabilidades de sobrevivir que los que viajan en segunda y tercera clase.
- Finalmente, el modelo obtenido se utiliza para predecir qué pasajeros sobrevivirán o no para los datos de prueba de los pasajeros del Titanic cuyas variables de supervivencia se desconocen. Se ha determinado que un pasajero puede sobrevivir si la probabilidad del pasajero es igual o mayor a 0.5, en caso contrario no lo hará. Luego, utilizando los resultados de estas predicciones, se extrae los resultados en un dataset que contiene una variable que indica si sobrevivió (test_result.csv).

2.10 Contribuciones al trabajo

Contribuciones	Firma
Investigación previa	JAH - IVS
Redacción de las respuestas	JAH - IVS
Desarrollo código	JAH - IVS