Tipología y ciclo de vida de los datos: Práctica 2

Daniel González Rodríguez

Junio 2022

Índice

1 Descripción del dataset.

Para la realización de la práctica, se ha seleccionado el dataset *Titanic: Machine Learning from Disaster* (https://www.kaggle.com/c/titanic), puesto que como se comentará a continuación, presenta algunas peculiaridades interesantes, además de la utilidad de poder aprovechar el trabajo realizado para participar en la competición de Kaggle.

Los datos contenidos en Kaggle consisten en tres ficheros:

- gender_submission.csv: un ejemplo de entrega para la competición.
- test.csv: datos de prueba para testear el modelo generado.
- train.csv: los datos en sí de los pasajeros del Titanic.

Si se describen los datos que se encuentran en el dataset *train.csv* que contiene todas las variables:

- PassengerId: Identificador del pasajero.
- Survived: Sobrevivió o no.
- Pclass: Clase en la que viajaba.
- Name: Nombre del pasajero.
- Sex: Genero del pasajero.
- Age: Edad del pasajero.
- SibSp: Número de hermanos y cónyuges abordo.
- Parch: Número de padres e hijos abordo.
- Ticket: Número del ticket.
- Fare: Tarifa del billete.
- Cabin: Número de cabina.
- Embarked: Puerto de embarque.

El análisis del dataset del Titanic aunque a priori pueda parecer intrascendente realmente presenta varios puntos interesantes. El primero de ellos es conocer con profundidad un suceso histórico y los eventos acontecidos a través de los datos, ya que, a través del análisis de los mismos, se pueden extrapolar determinados acontecimientos sin que se haya estado presente. Esto es, saber cómo se procedió a la evacuación de los pasajeros, a quien se dio prioridad, y si esto influyo en salvar la vida de determinadas personas.

El segundo punto, va anidado al primero, y es mostrar la verdadera capacidad del análisis de datos y del *Machine Learning* para extraer conocimiento de los mismos, mostrando la utilidad de esta ciencia.

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

Si se observan los archivos puestos a disposición en Kaggle como fuente de datos, se observa que solo los *csv* de *test* y *train* son datos relaciones con el Titanic.

En primer lugar, se cargarán los dos conjuntos de datos:

```
# Se leen los datos mediante read.csv ya que están separados por coma
datos_test <- read.csv('data/test.csv')
datos_train <- read.csv('data/train.csv')
```

Si se verifica la estructura del juego de datos cargados:

```
str(datos_train, width=80, strict.width="cut")
```

```
## '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 (F"..
##  $ 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" ...
```

```
str(datos_test, width=80, strict.width="cut")
```

```
## '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       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "M"..
##  $ Sex        : chr  "male" "female" "male" "male" ...
##  $ 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     : chr  "330911" "363272" "240276" "315154" ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : chr  "" "" "" "" ...
##  $ Embarked   : chr  "Q" "S" "Q" "S" ...
```

En un principio se valoró el integrar ambos conjuntos de datos, ya que esto permitiría tener un mayor número de muestras para el análisis posterior, pero esta idea se descartó por dos circunstancias. Como se observa, el *csv* nombrado como *test*, contiene una variable menos, que es la variable objetivo. Esto puede no representar un problema al valorar el modelo creado con un algoritmo no supervisado, pero si se desea crear un modelo supervisado y posteriormente testearlo con este conjunto, esto no será posible ya que no podrá obtener el valor de rendimiento del mismo. Además de esta circunstancia, la suma de ambos conjuntos no representa el número total de pasajeros del Titanic, es decir, incluso ambos conjuntos representan una muestra de los pasajeros, y no la población total del elemento bajo análisis, es por ello que no se tiene la certeza de si los

datos de *test* puestos a disposición pueden haber sido realizados bajo la premisa de que sean para probar solo el conjunto de entrenamiento, y por lo tanto datos no originales.

Por ello se procederá a seleccionar para el análisis los datos del conjunto de *train* exclusivamente.

Con los valores mostrados anteriormente, se pueden seleccionar un subconjunto de los datos y proceder a eliminar algunas variables que no aporten valor al análisis, al ser datos de identificadores únicos del pasajero, como son el nombre, su identificador, el número del ticket y el número cabina. Además, se elimina la ciudad donde embarco, ya que no tendrá representatividad para saber si se salvó o no, ya que podría correlar solo si se salvaron más los pasajeros de una determinada ciudad al tener un determinado nivel social, pero esto puede obtenerse a través de la categoría en la que viajaban:

```
# Se eliminan variables del análisis
datos <- datos_train[c("Survived", "Pclass", "Sex", "Age", "SibSp", "Parch",
                       "Fare")]

# Una vez eliminados se regulariza el tipo de variable para sea acorde con su tipo
datos$Survived[datos$Survived == 0] <- "No"
datos$Survived[datos$Survived == 1] <- "Si"
datos$Survived <- factor(datos$Survived)
datos$Pclass <- factor(datos$Pclass)
datos$Sex <- factor(datos$Sex)
```

3 Limpieza de los datos.

Se verifica en primer lugar a través de *summary* los datos que se tienen cargados para facilitar los análisis posteriores:

```
summary(datos)
```

```
##  Survived Pclass      Sex           Age            SibSp           Parch
##  No:549   1:216   female:314   Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Si:342   2:184   male  :577   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##           3:491                Median :28.00   Median :0.000   Median :0.0000
##                                Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                                3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                                Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                                NA's   :177
##       Fare
##  Min.   :  0.00
##  1st Qu.:  7.91
##  Median : 14.45
##  Mean   : 32.20
##  3rd Qu.: 31.00
##  Max.   :512.33
##
```

3.1 ¿Los datos contienen ceros o elementos vacíos?

Ya que con anterioridad se factorizaron aquellas variables categóricas, y se ha mostrado mediante *summary* un resumen de los datos, es fácil identificar que existen datos perdidos para la variable edad.

Puesto que el volumen de valores *NA's* existentes es bastante elevado, no es viable eliminar estas filas ya que se perderían demasiados datos para el análisis. Así mismo, una de las técnicas que habitualmente también se emplean, que es sustituir el valor no existente por la media de todos ellos tampoco va a resultar apropiado, ya que al ser un volumen alto estaría sesgando los datos con los valores introducidos, además de que la edad es uno de los elementos que puede tener un mayor peso en el análisis, por lo que no puede ser imputado de esta manera.

Por ello, se procede a imputar los valores ausentes mediante *kNN*, ya que así serán calculados valores en referencia a los valores de sus vecinos, y por lo tanto con un determinado parecido, en este caso seleccionando las cinco vecindades más próximas:

```
# Se carga la librería VIM
if (!require('VIM')) install.packages('VIM'); library('VIM')
# Se imputarán los valores para la variable Age a partir del resto de variables
datos <- kNN(datos, variable=c('Age'), k=5,
             dist_var=c("Survived","Pclass", "Sex", "SibSp","Parch", "Fare"))
```

Se comprueba que se han imputado todos los valores y no quedan *NA's*:

```
summary(datos)
```

```
##  Survived Pclass      Sex           Age            SibSp           Parch
##  No:549   1:216   female:314   Min.   : 0.42   Min.   :0.000   Min.   :0.0000
```

```
##  Si:342   2:184   male  :577   1st Qu.:21.00   1st Qu.:0.000   1st Qu.:0.0000
##           3:491                Median :28.00   Median :0.000   Median :0.0000
##                                Mean   :29.59   Mean   :0.523   Mean   :0.3816
##                                3rd Qu.:37.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                                Max.   :80.00   Max.   :8.000   Max.   :6.0000
##       Fare         Age_imp
##  Min.   :  0.00   Mode :logical
##  1st Qu.:  7.91   FALSE:714
##  Median : 14.45   TRUE :177
##  Mean   : 32.20
##  3rd Qu.: 31.00
##  Max.   :512.33
```

Se elimina la columna adicional creada por *kNN* durante la imputación:

```
datos$Age_imp <- NULL
```

Si ahora se revisan los datos con vistas a los ceros de los datos, es inmediato darse cuenta que en la variable tarifa existen valores con cero, por lo tanto, se puede tomar como un valor por defecto de relleno para una variable numérica en lugar de *NA* como sucedía antes, ya que la tarifa no puede ser nula.

De esta forma se procederá como en el caso anterior para imputar valores a estos casos, excepto que habrá que marcar a *kNN* cuales son los valores a imputar:

```
datos$Fare[datos$Fare == 0] <- NA
# Se imputarán los valores para la variable Fare a partir del resto de variables
datos <- kNN(datos, variable=c('Fare'), k=5,
             dist_var=c("Survived","Pclass", "Sex", "SibSp","Parch", "Age"))
```

Se comprueba que se han imputado todos los valores y ya no existen billetes sin coste:

```
summary(datos)
```

```
##  Survived Pclass      Sex           Age            SibSp           Parch
##  No:549   1:216   female:314   Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Si:342   2:184   male  :577   1st Qu.:21.00   1st Qu.:0.000   1st Qu.:0.0000
##           3:491                Median :28.00   Median :0.000   Median :0.0000
##                                Mean   :29.59   Mean   :0.523   Mean   :0.3816
##                                3rd Qu.:37.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                                Max.   :80.00   Max.   :8.000   Max.   :6.0000
##       Fare          Fare_imp
##  Min.   :  4.013   Mode :logical
##  1st Qu.:  7.925   FALSE:876
##  Median : 14.500   TRUE :15
##  Mean   : 32.517
##  3rd Qu.: 31.275
##  Max.   :512.329
```

Se elimina la columna adicional creada por *kNN* durante la imputación:

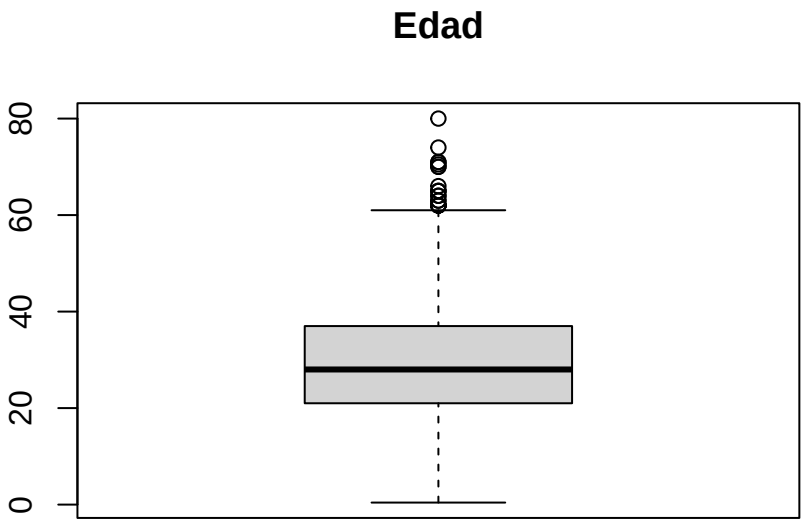
```
datos$Fare_imp <- NULL
```

3.2 Identifica y gestiona los valores extremos.

Ahora será necesario revisar los valores extremos que se encuentran en el conjunto de datos. Ya que existen variables categóricas, no será necesario analizar todas ya que estás solo tendrán los valores tabulados, por lo que se comenzará a revisar las variables numéricas que puedan tener valores extremos.

Si se revisa en primer lugar *Age* mediante *boxplot*:

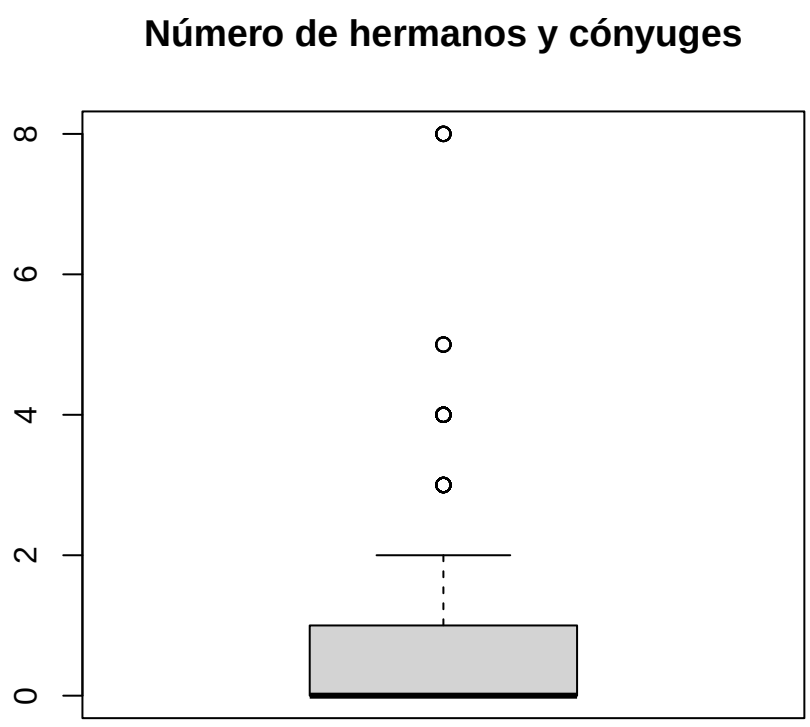
```
boxplot(datos$Age, main="Edad")
```



Aunque en este caso se observan valores que podrían considerarse *outliers* según *boxplot*, se trata de edades altas, y lejos de la media, pero bastante probables de encontrar dentro de la población, por lo que no es necesario tratarlas.

Si se repite el ejercicio, pero para *SibSp*:

```
boxplot(datos$SibSp, main="Número de hermanos y cónyuges")
```

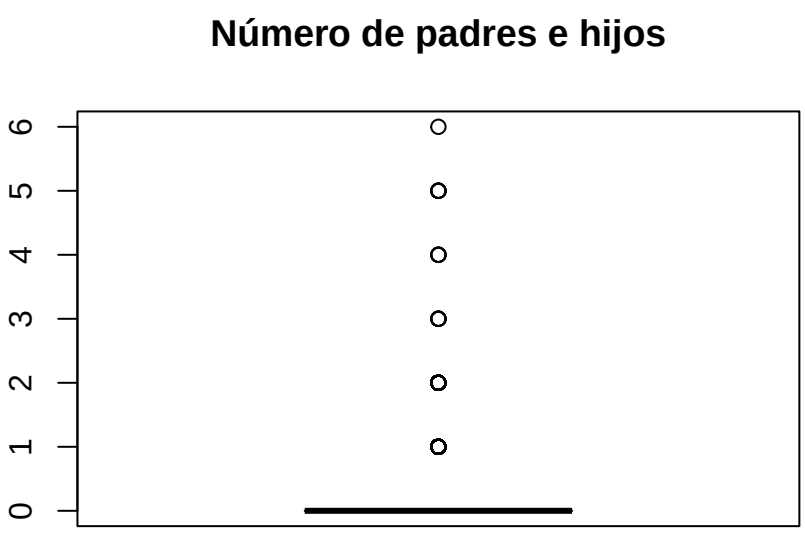


Se observan como posibles valores *outliers*, para 3,4,5 y 8. Ya que los tres primeros casos parecen factibles por el número de hermanos y cónyuges posibles, pero el último no, por lo que se procede a eliminar los datos que contienen esos valores para que no produzcan variaciones significativas en los datos:

```
# Se eliminan todas las filas con SibSp igual a 8
datos <- datos[!(datos$SibSp == 8),]
```

Si se repite el ejercicio, pero para *Parch*:

```
#out.width="50%"}
boxplot(datos$Parch, main="Número de padres e hijos")
```



Se observa como los datos están muy centrados en torno al cero. Si se tiene en cuenta lo que sería normal a partir del número de hijos y padres que podrían viajar juntos, se puede analizar los casos más extremos, de forma que se verifique que sean *outliers*. Para ello se revisa el número de casos existentes con 4, 5 y 6:

```
length(datos$Parch[datos$Parch == 4])
```

```
## [1] 4
```

```
length(datos$Parch[datos$Parch == 5])
```

```
## [1] 5
```

```
length(datos$Parch[datos$Parch == 6])
```

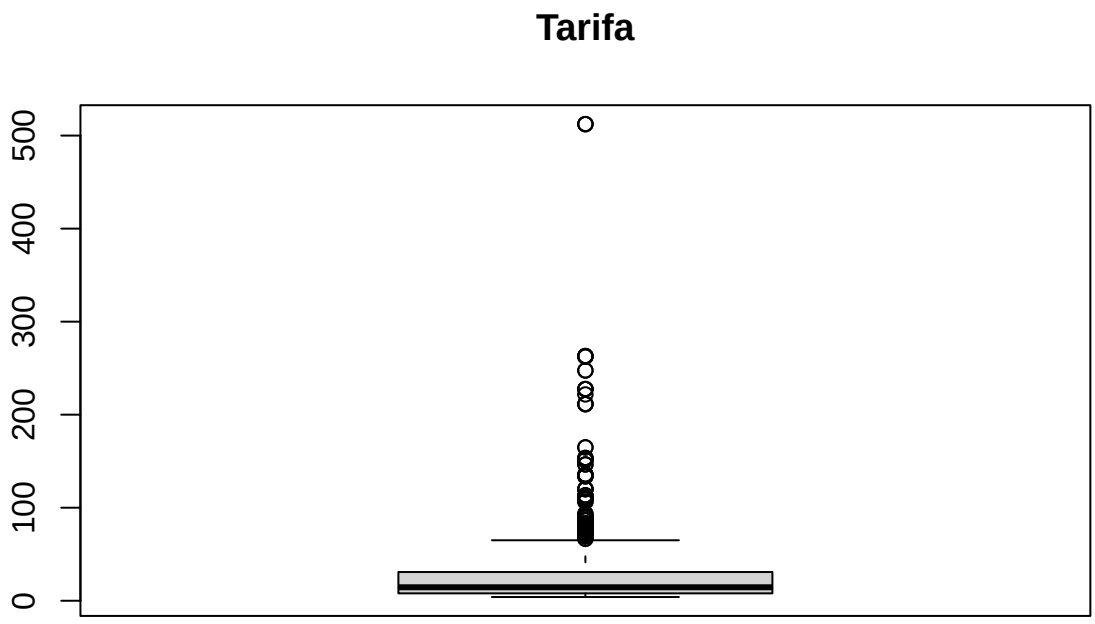
```
## [1] 1
```

Se puede observar cómo además de ser pocos casos, los casos de 4 y 5 tienen sentido, ya que existen a su vez 4 y 5 casos, lo que implican que efectivamente eran un grupo de 4 y 5 personas formado por padres e hijos. Esto no ocurre en cambio con 6, por lo que se eliminará del análisis:

```
# Se eliminan todas las filas con Parch igual a 6
datos <- datos[!(datos$Parch == 6),]
```

Si ahora se analiza *Fare*:

```r
boxplot(datos$Fare, main="Tarifa")
```



Se observa que existen un gran número de valores que podrían ser *outliers* pero que por su coste podrían corresponder a los billetes de primera clase. De esta forma solo se eliminan aquellos valores extremos que están incluso fuera de los billetes más caros:

```r
# Se eliminan todas las filas con Fare mayor a 500
datos <- datos[!(datos$Fare > 500),]
```

4 Análisis de los datos.

De cara a realizar el análisis de los datos, se va a proceder a generar una nueva variable que agrupe las edades, de esta manera será más fácil realizar comparativas entre los grupos de datos seleccionados:

```
# Se crean los grupos por edad
datos$GrupoEdad <- datos$Age
datos$GrupoEdad <- "Niños"
datos$GrupoEdad[datos$Age >= 15 & datos$Age <= 60] <- "Adultos"
datos$GrupoEdad[datos$Age > 60] <- "3a Edad"
datos$GrupoEdad <- factor(datos$GrupoEdad)
datos$Age <- NULL
```

Así se va a tener tres grupos de edad, niños, adultos y 3ª edad, en lugar de valores independientes numéricos que puedan complicar la comparativa y la aplicación de métodos de análisis.

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

Ya que anteriormente se han visto alguna de las medidas estadísticas descriptivas a través de *summary* y se han revisado los datos en busca de valores desaparecidos y *outliers*, se tiene una buena idea de cómo se puede aproximar su análisis para llegar a una conclusión útil a través de los mismos.

De esta manera se podrán proceder con distintos grupos de análisis. En un primer lugar, será interesante observar aquellos grupos que califican al pasaje de una manera numérica, es decir, los que describen a los pasajeros a través del número de familiares y su tarifa, ya que de esta manera se podrá verificar como se parecen estas distribuciones, lo que permitirá saber si tienen algún parecido, y de esta manera si son datos útiles para un análisis posterior o deben de ser omitidos.

Otro conjunto a analizar, se fundamenta en obtener modelos a partir de los cual se pueda clasificar de una manera rápida los pasajeros que sobrevivieron o no a partir de determinadas características, como son el género o la edad. Así se podrá poder obtener un conocimiento del grupo de datos, que posteriormente permita afinar el análisis para recoger una mayor predictibilidad del modelo.

Además, será recomendable analizar todas las variables a través de distintos modelos a partir del conocimiento que se obtenga de los dos análisis anteriores. Ya que tras ver como se relacionan determinadas variables será posible simplificar en análisis, y por lo tanto construir modelos de predicción útiles con un coste computacional más reducido.

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Como se ha comentado, se quiere analizar las variables numéricas para ver sus relaciones, por lo que en un primer paso se habrá de comprobar la normalidad y heterocedasticidad para poder saber que test aplicar posteriormente.

Comenzando por la normalidad, y usando el test de *Kolmogorov-Smirnov* y *Shapiro-Wilk* para verificar los resultados:

- En primer lugar se revisa el número de hermanos analizando la variable *SibSp*:

```
ks.test(datos$SibSp, pnorm, mean(datos$SibSp), sd(datos$SibSp))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  datos$SibSp
## D = 0.38797, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$SibSp
## W = 0.57112, p-value < 2.2e-16
```

Se obtiene mediante ambos test de manera univoca que los datos no presentaran normalidad.

- Ahora se realiza el mismo proceso para el número de hijos y padres a través de la variable *Parch*:

```
ks.test(datos$Parch, pnorm, mean(datos$Parch), sd(datos$Parch))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  datos$Parch
## D = 0.44829, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$Parch
## W = 0.52757, p-value < 2.2e-16
```

Obteniendo el mismo resultado en este caso también.

- Si se procede a analizar las tarifas *Fare*:

```
ks.test(datos$Fare, pnorm, mean(datos$Fare), sd(datos$Fare))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  datos$Fare
## D = 0.27501, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$Fare
## W = 0.59007, p-value < 2.2e-16
```

De nuevo se obtiene el mismo resultado, por lo que no existe normalidad en los datos presentados.

Ahora se realizara el mismo proceso para las variables, pero teniendo en cuenta que lo que se busca es comprobar como cambia la varianza. Ya que se sabe que las variables no cumplen la condición de normalidad será necesario emplear el test de *Fligner-Killeen* para cada par de datos:

```
# Se carga la librería car
if (!require('car')) install.packages('car'); library('car')

fligner.test(SibSp ~ Fare, data = datos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SibSp by Fare
## Fligner-Killeen:med chi-squared = 328.5, df = 244, p-value = 0.000246
```

```
fligner.test(Fare ~ Parch, data = datos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Parch
## Fligner-Killeen:med chi-squared = 50.295, df = 5, p-value = 1.206e-09
```

```
fligner.test(SibSp ~ Parch, data = datos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SibSp by Parch
## Fligner-Killeen:med chi-squared = 163.26, df = 5, p-value < 2.2e-16
```

Como se observa se rechaza la hipótesis nula en todos los casos, por lo que se puede concluir que todos los datos presentarán una varianza estadísticamente diferente. Esta conclusión resulta útil para incluir las variables en un análisis posterior y no descartar ninguna de ellas para su uso en el mismo.

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

Antes de pasar a otro tipo de pruebas, en línea con los resultados obtenidos anteriormente, se pueden realizar contrastes no paramétricos para tratar de averiguar si las tres variables estudiadas presentan diferencias significativas respecto de los supervivientes.

De esta manera si se realizan contrastes entre los supervivientes respecto de cada una de las variables a través del test de *Kruskal-Wallis*:

- En primer lugar para el número de hermanos *SibSp*:

```
kruskal.test(Survived ~ SibSp, data = datos)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Survived by SibSp
## Kruskal-Wallis chi-squared = 34.063, df = 5, p-value = 2.314e-06
```

Se comprueba que efectivamente existen diferencias significativas respecto de los supervivientes en función del número de hermanos.

- Si ahora se realiza el mismo análisis, pero para el número de padres e hijos *Parch*:

```
kruskal.test(Survived ~ Parch, data = datos)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Survived by Parch
## Kruskal-Wallis chi-squared = 30.666, df = 5, p-value = 1.09e-05
```

Se comprueba que efectivamente también existe una diferencia significativa cuando se tiene en cuenta esta variable.

- Por último se realiza el análisis, pero para la tarifa *Fare*:

```
kruskal.test(Survived ~ Fare, data = datos)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Survived by Fare
## Kruskal-Wallis chi-squared = 408.58, df = 244, p-value = 1.959e-10
```

También se llega a la misma conclusión, por lo que las tres variables afectarán a la manera en que sobrevivieron los pasajeros.

Merece la pena observar si existe cierta correlación entre estas variables, por lo que se verifica a través del método de *Spearman* al ser las distribuciones no normales:

```r
datos_cor <- datos[c("SibSp","Parch","Fare")]
cor(datos_cor)
```

```
##           SibSp     Parch      Fare
## SibSp 1.0000000 0.3983200 0.1989053
## Parch 0.3983200 1.0000000 0.2552517
## Fare  0.1989053 0.2552517 1.0000000
```

```r
cor.test(datos_cor$SibSp,datos_cor$Parch, method="spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  datos_cor$SibSp and datos_cor$Parch
## S = 64645293, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4308317
```

```r
cor.test(datos_cor$SibSp,datos_cor$Fare, method="spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  datos_cor$SibSp and datos_cor$Fare
## S = 64442728, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4326152
```

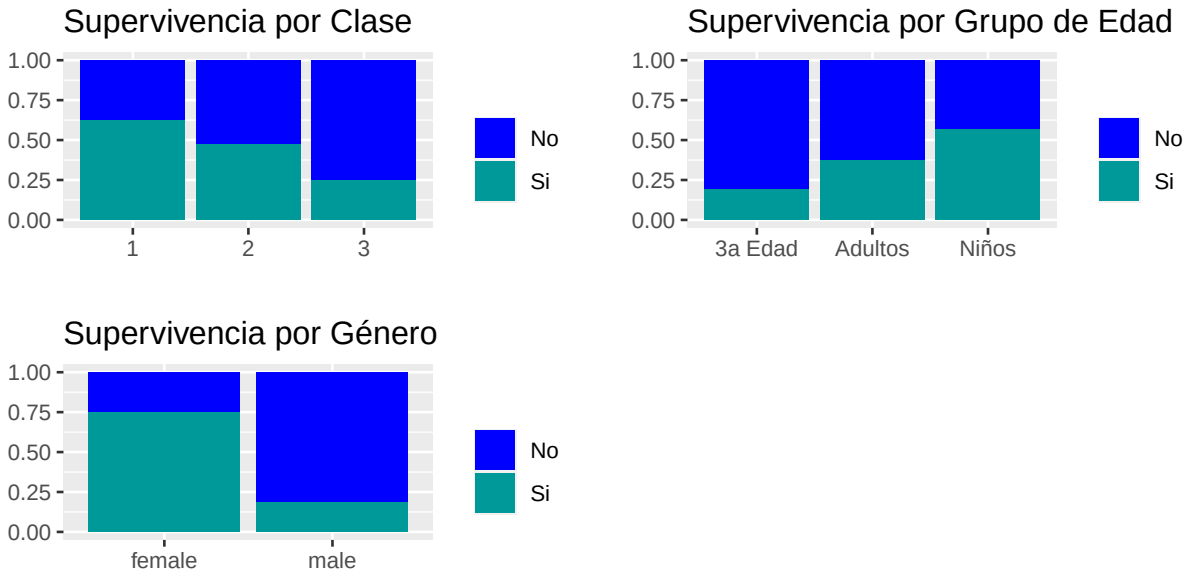
```r
cor.test(datos_cor$Fare,datos_cor$Parch, method="spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  datos_cor$Fare and datos_cor$Parch
## S = 69050041, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3920502
```

Observando un cierto nivel de correlación que por lo tanto hace latente que las tres variables influyan en resultado final.

Si ahora se centra el análisis en el otro grupo de variables categóricas que se comentaba, se puede realizar en primer lugar un análisis de una manera visual para obtener conclusiones rápidas:

```
# Se cargan las librerías
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
if (!require('gridExtra')) install.packages('gridExtra'); library('gridExtra')
if (!require('grid')) install.packages('grid'); library('grid')
grid.newpage()
porclase<-ggplot(datos,aes(Pclass,fill=Survived))+geom_bar(position="fill") +
  labs(x="", y="")+ scale_fill_manual(values=c("#0000FF","#009999"))+
  guides(fill=guide_legend(title=""))+ggtitle("Supervivencia por Clase")
poredad<-ggplot(datos,aes(GrupoEdad,fill=Survived))+geom_bar(position="fill") +
  labs(x="", y="")+ scale_fill_manual(values=c("#0000FF","#009999"))+
  guides(fill=guide_legend(title=""))+ggtitle("Supervivencia por Grupo de Edad")
porgenero<-ggplot(datos,aes(Sex,fill=Survived))+ geom_bar(position="fill") +
  labs(x="", y="")+scale_fill_manual(values=c("#0000FF","#009999"))+
  guides(fill=guide_legend(title=""))+ ggtitle("Supervivencia por Género")
grid.arrange(porclase,poredad,porgenero,ncol=2)
```



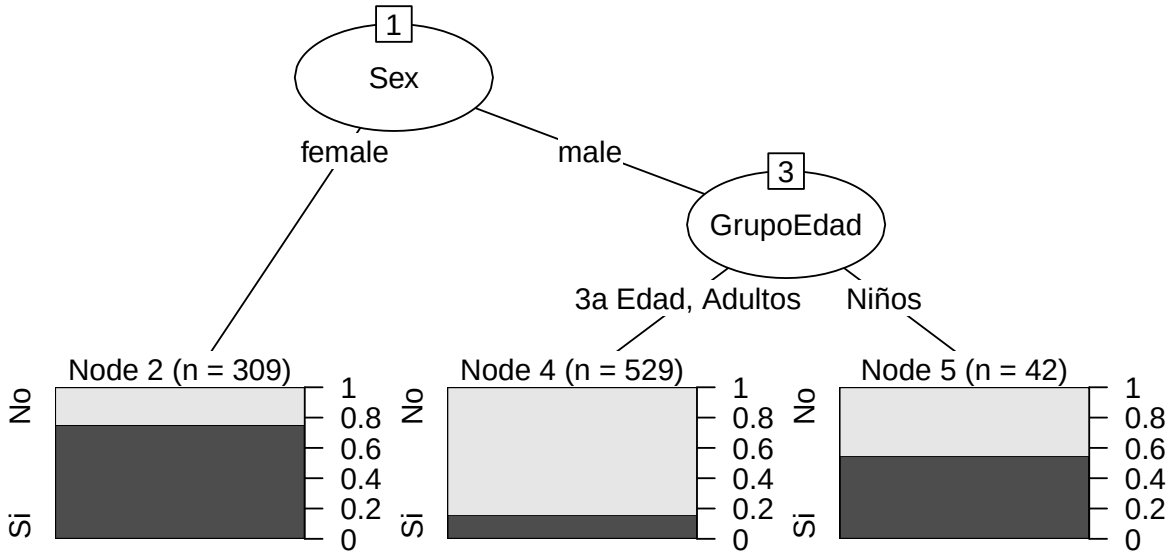
Efectivamente se puede concluir inmediatamente mediante la visualización como existe una mayor proporción de mujeres que sobrevivieron, y a su vez de niños y de pasajeros de primera clase.

Ahora se procede a realizar un modelo supervisado, para analizar si los datos que se han podido observar anteriormente tienen alguna interpretación adicional. Para ello, se plantea un modelo basado en arboles de decisión:

```
# Se carga la librería C50
if (!require('C50')) install.packages('C50'); library('C50')
datos_arbol <- datos[c("Pclass","Sex","GrupoEdad")]
arbol <- C50::C5.0(datos_arbol, datos$Survived)
# Se calcula el error y se muestra el árbol
print(paste0("El error al clasificar es: ",
             sum(predict(arbol,datos_arbol) != datos$Survived)/nrow(datos_arbol)))
```

```
## [1] "El error al clasificar es: 0.204545454545455"
```

```
plot(arbol)
```



En el modelo, al tener en consideración de manera conjunta todas las variables para realizar la clasificación se observa como emplea tan solo dos variables, por lo que es capaz de inferir si la persona sobrevivirá o no a través de solo la edad y el grupo de edad, aunque con un margen de error amplio.

Para contrastar si el resultado anterior tiene sentido, se va a proceder a realizar un modelo, pero esta vez basado en regresión logística, al ser la variable objetivo dicotómica:

```
regLO <- glm(datos$Survived ~ GrupoEdad+Pclass+Sex,data=datos_arbol, family="binomial")
summary(regLO)
```

```
##
## Call:
## glm(formula = datos$Survived ~ GrupoEdad + Pclass + Sex, family = "binomial",
##     data = datos_arbol)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6163  -0.7050  -0.4254   0.6777   2.2703
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.1061     0.5784   1.912 0.055822 .
## GrupoEdadAdultos    1.2326     0.5740   2.148 0.031752 *
## GrupoEdadNiños      2.2832     0.6374   3.582 0.000341 ***
## Pclass2            -0.9846     0.2533  -3.886 0.000102 ***
## Pclass3            -2.0759     0.2265  -9.166  < 2e-16 ***
## Sexmale            -2.6196     0.1876 -13.963  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1173.15  on 879  degrees of freedom
## Residual deviance:  795.69  on 874  degrees of freedom
## AIC: 807.69
##
## Number of Fisher Scoring iterations: 4
```

Ahora se puede observar como la significancia de las variables realmente decae con el grupo de adultos, siendo alta para el resto, lo que, si se compara con el resultado anterior, tiene su origen en que el grupo de adultos en su inmensa mayoría se clasifican como no supervivientes, de ahí que apenas presente significancia.

Si se calcula la efectividad del modelo construido aún a sabiendas de que son los mismos datos usados, sin tener en cuenta que se podría haber generado un conjunto de entrenamiento y otro de test, pero para tener una referencia respecto del método anterior:

```
prediccionRL <- predict(regLO, datos_arbol, type="response")
prediccionRL[prediccionRL > 0.5]  <- "Si"
prediccionRL[prediccionRL <= 0.5]  <- "No"
prediccionRL <- as.factor(prediccionRL)
# Se calcula el error
print(paste0("El error al clasificar es: ",
             (1 - sum(prediccionRL == datos$Survived) / length(datos$Survived))))
```

```
## [1] "El error al clasificar es: 0.206818181818182"
```

Se observa en este caso una leve mejoría, fruto como se ha visto de que emplee el total de las variables para realizar las predicciones.

Volviendo a los datos iniciales, se plantea el mismo análisis, pero con un algoritmo no supervisado con el objetivo de contrastar estos últimos resultados con los iniciales, y así poder comparar los resultados entre ambos grupos de datos seleccionados. En este caso se emplea *DBSCAN*:

```
# Se carga la librería fpc
if (!require('fpc')) install.packages('fpc'); library('fpc')
modelo_dbscan <- dbscan(datos_cor, eps=2, MinPts = 20)
table(modelo_dbscan$cluster, datos$Survived)
```

```
##
##      No  Si
##   0 100 129
##   1 441 210
```

Después de haber modificado los parámetros para que solo queden dos grupos en los que se quiere clasificar, superviviente o no, y teniendo en cuenta que los grupos no corresponden entre *0* y *No*, ya que *DBSCAN* no nombra los grupos, se supone el mejor escenario, y por lo tanto se tiene que el número de muestras correctamente clasificadas es del 65%, sensiblemente inferior a los métodos anteriores.

Por último, se genera el fichero con los datos empleados para que pueda ser usado posteriormente:

```
write.csv(datos, file = "data_out/Titanic_clean.csv", row.names=FALSE)
```

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

A través del análisis realizado se puede concluir que se han encontrado unos resultados que responden el planteamiento original del problema, ya que se puede dar explicación a una mayor o menor supervivencia en función de las distintas características de los pasajeros.

Dentro de los análisis realizados, se ha podido comprobar como las variables numéricas presentaban diferencias significativas al contrastarse contra la variable objetivo de supervivencia, indicando así que la supervivencia difiere entre los distintos niveles de hijos, padres, hermanos, cónyuges, y la tarifa del billete. Además, se ha analizado conjuntamente las variables numéricas para saber si existe correlación o no, demostrando efectivamente así, que se influyen cuando se quiere conocer si se sobrevivió o no a partir de dichas variables. Todo ello ha sido realizado tras abordar la distribución de las variables, ya que, una vez comprobada la no normalidad y la homocedasticidad, se pueden aplicar los test adecuados para obtener los resultados comentados.

Aprovechando las variables categóricas que se tenían, se han podido realizar análisis basados en métodos supervisados y de regresión, para poder obtener de una manera sencilla si realmente estas variables eran útiles para poder conocer la variable objetivo, llegando a la conclusión de que si lo son. Incluso se ha visto a través del árbol de decisión, como con solo las variables edad y género, se puede llegar a una decisión rápida (a costa de un mayor error) de si el pasajero sobrevivió o no. Para verificar este punto además se ha visto mediante regresión que al tener en cuenta la tercera variable el error se reducía, por lo que en sucesivos ajustes se puede obtener un modelo más adecuado.

Por último, se ha querido emplear un método de partición para comparar este resultado con los anteriores, usando en este caso *DBSCAN*. Si bien el error no es tan bueno como en los casos anteriores, este método permite obtener los resultados de manera no supervisado, sin tener en cuenta la variable objetivo, lo que puede llegar a resultar útil, como por ejemplo para valorar el conjunto de *train* que no contiene este valor, o cualquier grupo nuevo de datos que se plantee sin la variable objetivo.

Así se puede concluir, que el protocolo establecido de las mujeres y los niños primero, se cumple en este caso ya que córrela ampliamente el nivel de supervivencia de estas clases. Además, se puede observar como también existe un mayor índice de supervivencia entre los pasajeros de las clases superiores que de las inferiores, tal y como se esperaba al inicio del análisis.

6 Código

El código empleado se encuentra distribuido a través de todos los *chunks* del texto y adicionalmente está disponible de manera completa en:

https://github.com/Dani643/Limpiezayanalisis/tree/main/codigo

---

Contribuciones	Firma
Investigación previa	Daniel González Rodríguez
Redacción de las respuestas	Daniel González Rodríguez
Desarrollo código	Daniel González Rodríguez

Bibliografía

García Pérez, Alfonso. 2013. *Estadística Básica Con r*. UNED.
González, Mireia Calvo. 2019. *Introducción a La Limpieza y Análisis de Los Datos*. FUOC.
Lam, Longhow. 2010. *An Introduction to r*.