Práctica 2:
Limpieza y validación de datos

Tipología y ciclo de vida de los datos

GOMEZ GARCIA, JOSE LUIS
JOSE LUIS GOMEZ GARCIA - jgomezgar@uoc.edu

Enlace web a la práctica en github

https://github.com/jgomezgar/TitanicPred

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

El dataset presenta distintas características de los pasajeros del Titanic, así como si consiguieron sobrevivir a su naufragio.

La **pregunta que se pretende resolver es**, Qué características debe tener un pasajero para tener mayor probabilidad de supervivencia o no supervivencia.

Es importante ya que un buen conocimiento de los patrones de supervivencia, permitirá adaptar los protocolos de emergencia y minimizar las víctimas.

Descripción del dataset:

Partiendo de la premisa enunciada en las competencias a adquirir:

"Capacidad para aplicar las técnicas específicas de tratamiento de datos (***integración, transformación, limpieza y validación***) para su posterior análisis".

Tomare los distintos ficheros como distintos orígenes de datos que hay que integrar, transformar limpiar y validar, aunque este no sea el propósito enunciado en Kaggle, ni tampoco sea una práctica habitual, ya que siempre es necesario reservar una parte de los datos para testar el modelo.

Los datos se han dividido en tres ficheros de tres orígenes de datos diferentes:

- Conjunto de datos completos de la fuente1 → Originalmente Conjunto de entrenamiento (train.csv).
- Conjunto de datos incompletos de la fuente 2 → Originalmente conjunto de pruebas (test.csv).
- Conjunto de datos que complementa los datos de la fuente 2 →Originalmente Conjunto de predicciones (gender_submission.csv).

Data Source	File	Líneas / Columnas	Función Original en Kraggle
Fuente 1 - completa	train.csv	891 x 12	Datos para crear el modelo
Fuente 2 - incompleta	test.csv	418 x 11	Datos para comprobar el modelo
Complemento Fuente 2	gender_submission.csv	418 x 2	Resultados de supervivencia de los datos de test (ejemplo).

JOSE LUIS GOMEZ GARCIA - jgomezgar@uoc.edu

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

Los campos del dataset son los siguientes:

Variable	Definición	Info
PassengerId	Clave única	Presente en todos los ficheros
survival	Supervivencia / RESPUESTA	0 = No, 1 = Yes (no presente en fichero test)
pclass	Clase de ticket	1 = Upper, 2 = Middle, 3 = Lower
Name	Nombre y título del pasajero	
sex	Sexo	
Age	Edad en años	
sibsp	# de hermanos / cónyuges a bordo	
parch	# de padres / hijos a bordo	
ticket	Numero de ticket	
fare	Tarifa de pasajero	
cabin	Numero de Camarote	
embarked	Puerto de Origen (embarque)	C = Cherbourg, Q = Queenstown, S = Southampton

```
### Incluir Datasources
Full_Source1 <- read.csv("./train.csv", stringsAsFactors = F, na.strings = c("NA", ""))
Part1_Source2 <- read.csv("./test.csv", stringsAsFactors = F, na.strings = c("NA", ""))
Part2_Source2 <- read.csv("./gender_submission.csv", stringsAsFactors = F, na.strings = c("NA", ""))

### Join de Part1_Source2 (test) & Part2_Source2 (gender_submission)
### para obtener una unica tabla con los mismos campos que Full_Source1 (train).
Full_Source2 <- merge(Part1_Source2, Part2_Source2)
### Union de las tablas Full_Source1 (train) y Full_Source2
titanic <- rbind(Full_Source1, Full_Source2)
```

```
> ### Informacion de la tabla resultante
> str(titanic)
'data.frame':   1309 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence B
)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
 $ 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  NA "C85" NA "C123" ...
 $ Embarked   : chr  "S" "C" "S" "S" ...
```


JOSE LUIS GOMEZ GARCIA - jgomezgar@uoc.edu

```
> summary(titanic)
  PassengerId      Survived          Pclass          Name
 Min.   :   1   Min.   :0.0000   Min.   :1.000   Length:1309
 1st Qu.: 328   1st Qu.:0.0000   1st Qu.:2.000   Class :character
 Median : 655   Median :0.0000   Median :3.000   Mode  :character
 Mean   : 655   Mean   :0.3774   Mean   :2.295
 3rd Qu.: 982   3rd Qu.:1.0000   3rd Qu.:3.000
 Max.   :1309   Max.   :1.0000   Max.   :3.000

     Sex                 Age            SibSp            Parch
 Length:1309        Min.   : 0.17   Min.   :0.0000   Min.   :0.000
 Class :character   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000
 Mode  :character   Median :28.00   Median :0.0000   Median :0.000
                    Mean   :29.88   Mean   :0.4989   Mean   :0.385
                    3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000
                    Max.   :80.00   Max.   :8.0000   Max.   :9.000
                    NA's   :263
    Ticket              Fare            Cabin             Embarked
 Length:1309        Min.   :  0.000   Length:1309        Length:1309
 Class :character   1st Qu.:  7.896   Class :character   Class :character
 Mode  :character   Median : 14.454   Mode  :character   Mode  :character
                    Mean   : 33.295
                    3rd Qu.: 31.275
                    Max.   :512.329
                    NA's   :1
```

```
### NA´s Presentes en cada columas
sapply(titanic, function(x) {sum(is.na(x))})
```

```
PassengerId    Survived      Pclass        Name         Sex         Age
          0           0           0           0           0         263
      SibSp       Parch      Ticket        Fare       Cabin    Embarked
          0           0           0           1        1014           2
```

Información a destacar:

- Existen 1309 registros.
- Tipos de datos por columnas, entero, numérico y carácter.
- Sobreviven el 37’74% de los pasajeros
- En la variable *Fare* tiene un Máximo muy alto y un mínimo demasiado bajo.
- Las variables *Age*, , *Cabin* contienen muchos NA´s
- Las variables *Fare, Embarked* tienen algún NA
- Las variables *sex y Pclass* están completos y ordenados. Son candidatos a predictores. Se deben convertir a Factor.

```
### conversion a Factor
titanic$Sex <- as.factor(titanic$Sex)
titanic$Pclass <- as.factor(titanic$Pclass)
```

- Las variables Pasenger ID y Name, por su diversidad de valores, no parecen buenos predictores.
- De la Variable *Name* es posible extraer el título honorifico del pasajero, que podría tener valor como predictor.

```
###Extraer el Título de Nombre y Apellidos
titanic$Surname <- sapply(titanic$Name, function(x) {strsplit(x, split='[,.]')[[1]][1]})
titanic$Surname <- sapply(titanic$Surname, function(x) {strsplit(x, split='[-]')[[1]][1]})
titanic$TitleName <- sapply(titanic$Name, function(x) {strsplit(x, split='[,.]')[[1]][2]})
titanic$TitleName <- sub(' ', '', titanic$TitleName)
kable(table(titanic$Sex, titanic$TitleName))
```

	Capt	Col	Don	Dona	Dr	Jonkheer	Lady	Major	Master	Miss	Mlle	Mme	Mr	Mrs	Ms	Rev	Sir	the Countess
female	0	0	0	1	1	0	1	0	0	260	2	1	0	197	2	0	0	1
male	1	4	1	0	7	1	0	2	61	0	0	0	757	0	0	8	1	0

Al obtener muchos posibles valores pero con pocas ocurrencias homogenizo a variable Booleana que indique si tiene o no título honorifico.

```
###Homogenizar la variable Título
titanic$Title[titanic$TitleName %in% c("Mlle", "Ms","Mme","Mrs",'Master', 'Miss', 'Mr','Don', 'Dona')] <- 0
titanic$Title[!(titanic$TitleName %in% c("Mlle", "Ms","Mme","Mrs",'Master', 'Miss', 'Mr','Don', 'Dona'))] <-1
titanic$Title <- as.factor(titanic$Title)
kable(table(titanic$Sex, titanic$Title))
```

	0	1
female	463	3
male	819	24

Resultando:

```
> ### Informacion de la tabla resultante
> summary(titanic)
  PassengerId      Survived      Pclass      Name              Sex          Age            SibSp           Parch      
 Min.   :   1   Min.   :0.0000   1:323   Length:1309        female:466   Min.   : 0.17   Min.   :0.0000   Min.   :0.000  
 1st Qu.: 328   1st Qu.:0.0000   2:277   Class :character   male  :843   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000  
 Median : 655   Median :0.0000   3:709   Mode  :character              Median :28.00   Median :0.0000   Median :0.000  
 Mean   : 655   Mean   :0.3774                                         Mean   :29.88   Mean   :0.4989   Mean   :0.385  
 3rd Qu.: 982   3rd Qu.:1.0000                                         3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000  
 Max.   :1309   Max.   :1.0000                                         Max.   :80.00   Max.   :8.0000   Max.   :9.000  
                                                                       NA's   :263                                     
    Ticket               Fare            Cabin             Embarked           Surname           TitleName          Title      
 Length:1309        Min.   :  0.000   Length:1309        Length:1309        Length:1309        Length:1309        0:1282  
 Class :character   1st Qu.:  7.896   Class :character   Class :character   Class :character   Class :character   1:  27  
 Mode  :character   Median : 14.454   Mode  :character   Mode  :character   Mode  :character   Mode  :character           
                    Mean   : 33.295                                                                                        
                    3rd Qu.: 31.275                                                                                        
                    Max.   :512.329                                                                                        
                    NA's   :1                                                                                              
```


JOSE LUIS GOMEZ GARCIA - jgomezgar@uoc.edu

3. Limpieza de los datos.

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

Según hemos visto en el apartado anterior:

- Los 0´s en el número de Familiares es un valor valido.
- Las variables *Age, Cabin* contienen muchos NA´s, lo que complica su utilidad como predictores, estas variables serán ignoradas.
- La variable *Fare* contiene valores 0, tomare el valor como válido, asumiendo que se trata de la tarifa para la tripulación
- Las variables *Fare, Embarked* contienen NA´s (2 y 1 respectivamente). Es posible arreglarlos usando los costes medios por categoría y puerto de embarque.
  Para los siguientes pasos, es necesario ejecutarlos después del código para el tratamiento de valores extremos, ya que se ha calculado la tarifa por persona en lugar de la tarifa por ticket.

  - A las dos pasajeras sin puerto de embarque se les asignara el puerto con el coste medio de 1º clase, más cercano a 40.

```
###Pasajeros sin puerto de embarque
kable(titanic[which(is.na(titanic$Embarked)),
» c( 'Name', 'Survived', 'Title','Pclass', 'Age', 'SibSp', 'Parch', 'Ticket', 'FarePP', 'Embarked')])
```

```
|   |Name                                     | Survived|Title |Pclass | Age| SibSp| Parch|Ticket | FarePP|Embarked |
|:--|:----------------------------------------|--------:|:-----|:------|---:|-----:|-----:|:------|------:|:--------|
|59 |Icard, Miss. Amelie                      |        1|0     |1      |  38|     0|     0|113572 |     40|NA       |
|60 |Stone, Mrs. George Nelson (Martha Evelyn) |        1|0     |1      |  62|     0|     0|113572 |     40|NA       |
```

  - Para el pasajero sin Precio de tarifa, se le asignara el precio medio de 3º clase para embarques desde Southampton

```
###Pasajeros sin Tarifa
kable(titanic[which(is.na(titanic$Fare)),
» c( 'Name', 'Survived', 'Title','Pclass', 'Age', 'SibSp', 'Parch', 'Ticket', 'FarePP', 'Embarked')])
```

```
|    |Name              | Survived|Title |Pclass |  Age| SibSp| Parch|Ticket | FarePP|Embarked |
|:---|:-----------------|--------:|:-----|:------|----:|-----:|-----:|:------|------:|:--------|
|861 |Storey, Mr. Thomas |        0|0     |3      | 60.5|     0|     0|3701   |     NA|S        |
```


JOSE LUIS GOMEZ GARCIA - jgomezgar@uoc.edu

3.2. Identificación y tratamiento de valores extremos.

Según hemos visto en el apartado anterior:

- La variable *Fare* contiene valores demasiado altos, se asume que es el precio del ticket, y un ticket puede ser compartido por varias personas. Es posible arreglarlo dividiendo el coste entre las personas que lo comparten. La variable *Fare* deja paso a la variable *Fare per Person (FarePP)*, ver código.

Código para la Limpieza de datos: Común para ejercicio 3.1 y 3,2

```
### Numero de Personas por tiket
TiketsRep<-data.frame(table(titanic$Ticket))
colnames(TiketsRep)[1] <- "Ticket" ### Renombrar columna para facilitar merge
titanic <- merge(titanic, TiketsRep)

### tarifa por persona
titanic$FarePP <- titanic$Fare/titanic$Freq

### Tabla con la tarifa media por clase y puerto de embarque / se escluyen terifas 0
Fares_Table <- aggregate(FarePP~Pclass+Embarked, data=titanic, FUN=(function(x)
» » » » » » » » » {ifelse(sum(x==0)>0 & sum(x !=0) >0, mean(x[x>0]), mean(x))}))
kable(Fares_Table)
```

Pclass	Embarked	FarePP
1	C	38.324320
2	C	13.281393
3	C	6.843819
1	Q	30.000000
2	Q	11.735114
3	Q	7.577325
1	S	31.643309
2	S	11.469598
3	S	7.431779

```
###Imputacion Pasajeros sin puerto de embarque
titanic$Embarked[titanic$Ticket=='113572'] <- 'C'
### Imputacion al Pasajeros sin Tarifa
titanic$FarePP[titanic$Ticket=='3701'] <- 7.43
```


JOSE LUIS GOMEZ GARCIA - jgomezgar@uoc.edu

4. Análisis de los datos.

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

Columna	Info	Acción
PassengerId	Identifica al pasajero, cardinalidad igual número de líneas por lo que no discrimina.	Omitir
Survived	Resultado de la predicción.	Incluir
Pclass	Candidato a predictor	Incluir
Name	Identifica al pasajero, cardinalidad igual número de líneas por lo que no discrimina.	Omitir
Sex	Candidato a predictor	Incluir
Age	Incompleto	Omitir
SibSp	Agregar SibSp y Parch a una nueva variable llamada Familia	unificar
Parch	Agregar SibSp y Parch a una nueva variable llamada Familia	unificar
Fare	Sustituida por tarifa por persona	Omitir
Cabin	Incompleto	Omitir
Embarked	Candidato a predictor	Incluir
Surname	Identifica al pasajero, cardinalidad similar al número de líneas por lo que no discrimina.	Omitir
TitleName	Sustituido por Title	Omitir
Title	Candidato a predictor	Incluir
FarePP	Candidato a predictor	Incluir
Family	Candidato a predictor	Incluir
Ticket	Identifica al pasajero, cardinalidad similar al número de líneas por lo que no discrimina.	Omitir
Freq	Atributo Auxiliar de Ticket	Omitir

```
### conversion a Factor
titanic$Family <- as.factor(titanic$SibSp + titanic$Parch) ## Se agregan familiares
titanic$Embarked <- as.factor(titanic$Embarked)
### Seleccion de columnas
TitanicPred <- subset(titanic, select = c('Survived', 'Title','Pclass', 'Sex', 'Family', 'FarePP', 'Embarked'))
### Informacion de la tabla resultante
summary (TitanicPred)
```

```
   Survived       Title      Pclass       Sex           Family        FarePP          Embarked
 Min.   :0.0000   0:1282   1:323   female:466   0      :790   Min.   :  0.00   C:272
 1st Qu.:0.0000   1:  27   2:277   male  :843   1      :235   1st Qu.:  7.55   Q:123
 Median :0.0000            3:709               2      :159   Median :  8.05   S:914
 Mean   :0.3774                                3      : 43   Mean   : 14.75
 3rd Qu.:1.0000                                5      : 25   3rd Qu.: 15.00
 Max.   :1.0000                                4      : 22   Max.   :128.08
                                               (Other): 35
```


JOSE LUIS GOMEZ GARCIA - jgomezgar@uoc.edu

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

Normalidad:

Para la comprobar que los valores de las variables provienen de una población distribuida normalmente, se utilizará la prueba de normalidad de Anderson-Darling. Así, se comprobará que para que cada prueba se obtiene un p-valor superior al nivel de significación prefijado = 0, 05. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

```
### Coeficiente de correlación para cada variable cuantitativa
TitanicPred$Survived <- as.factor(TitanicPred$Survived)
TitanicPred$TitleNum <- as.numeric(TitanicPred$Title)
TitanicPred$PclassNum <- as.numeric(TitanicPred$Pclass)
TitanicPred$SexNum <- as.numeric(TitanicPred$Sex)
TitanicPred$FamilyNum <- as.numeric(TitanicPred$Family)
TitanicPred$EmbarkedNum <- as.numeric(TitanicPred$Embarked)
```

```
library(nortest)
### Comprobación de la normalidad
alpha = 0.05
col.names = colnames(TitanicPred)
for (i in 1:ncol(TitanicPred)) {
  if (is.integer(TitanicPred[,i]) | is.numeric(TitanicPred[,i])) {
    p_val = ad.test(TitanicPred[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      cat(" NO sigue una distribución normal\n")
    }else{
      cat(col.names[i])
      cat(" SI sigue una distribución normal\n")
    }
  }
}
```

- FarePP NO sigue una distribución normal
- SexNum NO sigue una distribución normal
- PclassNum NO sigue una distribución normal
- EmbarkedNum NO sigue una distribución normal
- TitleNum NO sigue una distribución normal
- FamilyNum NO sigue una distribución normal

No se obtiene ninguna variable que siga la distribución normal:

homogeneidad de varianzas:

Utilizando el test de Fligner-Killeen entre poblaciones, se comprobara si la varianza de Survived en cada una de las poblaciones es similar, para lo cual el p-valor resultante debe ser superior a 0,05.

```
### Comprobación de la homogeneidad de varianzas:
fligner.test(Survived ~ Title, data = TitanicPred)
fligner.test(Survived ~ Pclass, data = TitanicPred)
fligner.test(Survived ~ Sex, data = TitanicPred)
fligner.test(Survived ~ Family, data = TitanicPred)
fligner.test(Survived ~ Embarked, data = TitanicPred)
```

```
> ### Comprobación de la homogeneidad de varianzas:
> fligner.test(Survived ~ Title, data = TitanicPred)

        Fligner-Killeen test of homogeneity of variances

data:  Survived by Title
Fligner-Killeen:med chi-squared = 0.77095, df = 1, p-value = 0.3799

> fligner.test(Survived ~ Pclass, data = TitanicPred)

        Fligner-Killeen test of homogeneity of variances

data:  Survived by Pclass
Fligner-Killeen:med chi-squared = 34.297, df = 2, p-value = 3.568e-08

> fligner.test(Survived ~ Sex, data = TitanicPred)

        Fligner-Killeen test of homogeneity of variances

data:  Survived by Sex
Fligner-Killeen:med chi-squared = 4.7901, df = 1, p-value = 0.02862

> fligner.test(Survived ~ Family, data = TitanicPred)

        Fligner-Killeen test of homogeneity of variances

data:  Survived by Family
Fligner-Killeen:med chi-squared = 39.436, df = 8, p-value = 4.079e-06

> fligner.test(Survived ~ Embarked, data = TitanicPred)

        Fligner-Killeen test of homogeneity of variances

data:  Survived by Embarked
Fligner-Killeen:med chi-squared = 25.769, df = 2, p-value = 2.537e-06
```

Solo con la variable *Title* se obtiene un valor superior a 0.05, en las demás variables rechazamos la hipótesis de que las varianzas de sus poblaciones son homogéneas.

4.3. 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.

El **análisis de correlación** entre las distintas variables permite determinar cuáles de ellas ejercen una mayor influencia sobre la posibilidad de supervivencia.
Para ello, se utilizará el coeficiente de correlación de Spearman, puesto que hemos visto que tenemos datos que no siguen una distribución normal.

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")

for (i in 1:(ncol(TitanicPred) - 1)) {
  if (is.integer(TitanicPred[,i]) | is.numeric(TitanicPred[,i])) {
    spearman_test = cor.test(TitanicPred[,i],
                             TitanicPred[,length(TitanicPred)],
                             method = "spearman")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value
    # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(TitanicPred)[i]
  }
}
kable(corr_matrix)
```

	estimate	p-value
FarePP	-0.0830156	0.0026485
SexNum	-0.2755186	0.0000000
PclassNum	-0.0893634	0.0012098
EmbarkedNum	0.0020821	0.9400087
TitleNum	-0.0145081	0.5999788

No existe ninguna Variable que correlacione fuertemente, siendo la mayor correlación con Sex.

El **contraste de hipótesis** sobre dos muestras permite determinar si la supervivencia de una población es superior al de la otra:

Se realizan los siguientes Contrastes de hipótesis de dos muestras sobre la diferencia de sus medias:

- Las Mujeres tienen mejor supervivencia que hombres
- 1º clase tienen mejor supervivencia que 2º clase
- 2º clase tienen mejor supervivencia que 3º clase
- Pasajeros con Titulo honorifico tienen mejor supervivencia que aquellos sin el

- Pasajeros con familiares tienen mejor supervivencia que Pasajeros sin familiares

Al establecer un valor de significación alpha = 0,05 ; siempre que se obtenga un p-valor menor rechazamos la hipótesis nula, por lo que podemos concluir que el primer grupo tiene mayor probabilidad de sobrevivir que el Segundo:

```
### Welch Two Sample t-test
TitanicPred$Survived <- as.numeric(TitanicPred$Survived)

###Mujeres mejor supervivencia que hombres
t.test(TitanicPred$Survived[TitanicPred$Sex=="female"],TitanicPred$Survived[TitanicPred$Sex=="male"],alternative="greater")
### 1° clase mejor supervivencia que 2° clase
t.test(TitanicPred$Survived[TitanicPred$Pclass==1],TitanicPred$Survived[TitanicPred$Pclass==2],alternative="greater")
### 2° clase mejor supervivencia que 3° clase
t.test(TitanicPred$Survived[TitanicPred$Pclass==2],TitanicPred$Survived[TitanicPred$Pclass==3],alternative="greater")
### Titulo honorifico mejor supervivencia que sin el
t.test(TitanicPred$Survived[TitanicPred$Title==1],TitanicPred$Survived[TitanicPred$Title==0],alternative="greater")
### con familiares mejor supervivencia que sin familiares
t.test(TitanicPred$Survived[TitanicPred$Family != 0],TitanicPred$Survived[TitanicPred$Family == 0],alternative="greater")
```

```
> ###Mujeres mejor supervivencia que hombres
> t.test(TitanicPred$Survived[TitanicPred$Sex=="female"],TitanicPred$Survived[TitanicPred$Sex=="male"],alternative="grea

        Welch Two Sample t-test

data:  TitanicPred$Survived[TitanicPred$Sex == "female"] and TitanicPred$Survived[TitanicPred$Sex == "male"]
t = 33.127, df = 865.24, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6622409       Inf
sample estimates:
mean of x mean of y
  1.82618   1.12930

> ### 1° calse mejor supervivencia que 2° clase
> t.test(TitanicPred$Survived[TitanicPred$Pclass==1],TitanicPred$Survived[TitanicPred$Pclass==2],alternative="greater")

        Welch Two Sample t-test

data:  TitanicPred$Survived[TitanicPred$Pclass == 1] and TitanicPred$Survived[TitanicPred$Pclass == 2]
t = 3.7868, df = 584.19, p-value = 8.421e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.08670061        Inf
sample estimates:
mean of x mean of y
 1.575851  1.422383

> ### 2° calse mejor supervivencia que 3° clase
> t.test(TitanicPred$Survived[TitanicPred$Pclass==2],TitanicPred$Survived[TitanicPred$Pclass==3],alternative="greater")

        Welch Two Sample t-test

data:  TitanicPred$Survived[TitanicPred$Pclass == 2] and TitanicPred$Survived[TitanicPred$Pclass == 3]
t = 4.4881, df = 459.19, p-value = 4.548e-06
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.0968067       Inf
sample estimates:
mean of x mean of y
 1.422383  1.269394

> ### Titulo honorifico mejor supervivencia que sin el
> t.test(TitanicPred$Survived[TitanicPred$Title==1],TitanicPred$Survived[TitanicPred$Title==0],alternative="greater")

        Welch Two Sample t-test

data:  TitanicPred$Survived[TitanicPred$Title == 1] and TitanicPred$Survived[TitanicPred$Title == 0]
t = -0.91418, df = 27.205, p-value = 0.8157
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.2370267        Inf
sample estimates:
mean of x mean of y
 1.296296  1.379095
```


JOSE LUIS GOMEZ GARCIA - jgomezgar@uoc.edu

```
> ### con familiares mejor supervivencia que sin familiares
> t.test(TitanicPred$Survived[TitanicPred$Family != 0],TitanicPred$Survived[TitanicPred$Family == 0],alternative="greater")

        Welch Two Sample t-test

data:  TitanicPred$Survived[TitanicPred$Family != 0] and TitanicPred$Survived[TitanicPred$Family == 0]
t = 7.854, df = 1033.6, p-value = 5.028e-15
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.1694095       Inf
sample estimates:
mean of x mean of y
 1.506744  1.292405
```

En todas las hipótesis presentadas rechazamos la hipótesis nula, excepto en pasajeros con título honorifico donde se puede determinar que tengan una mayor probabilidad de supervivencia.

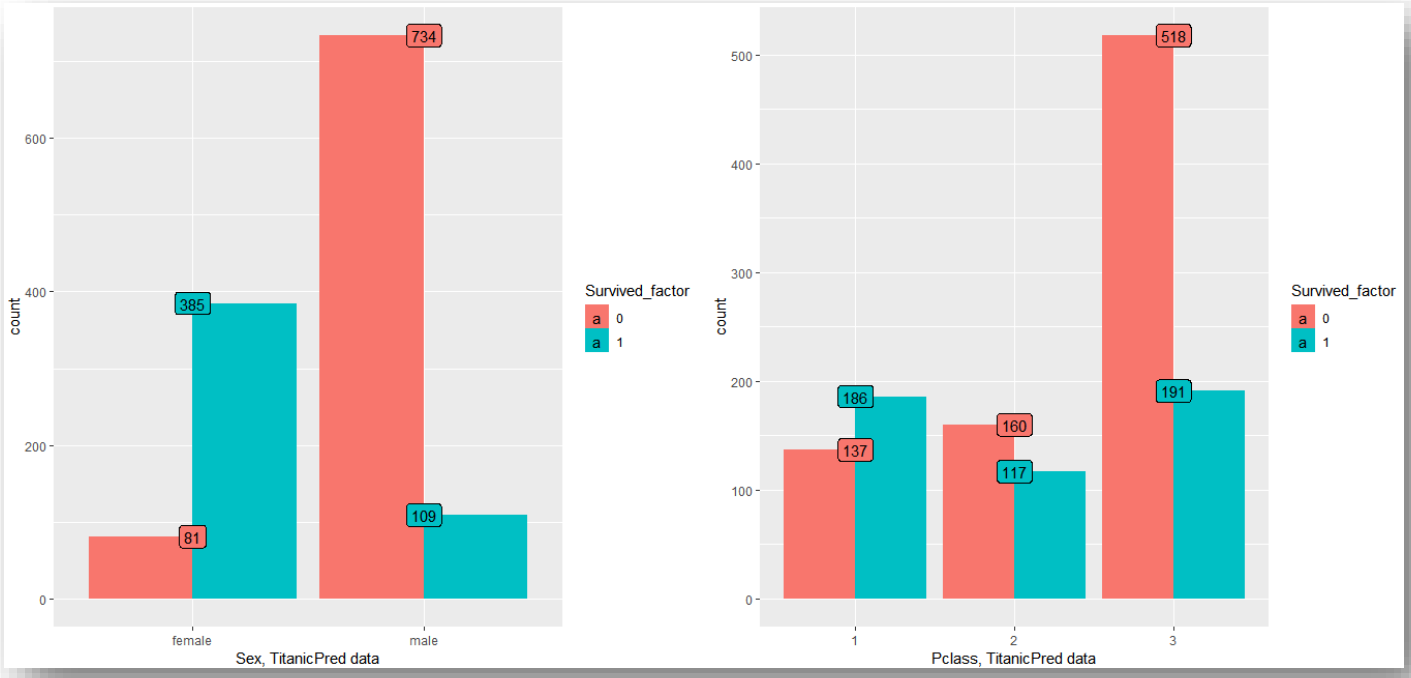
5. Representación de los resultados a partir de tablas y gráficas.

Supervivencia según Sexos y Supervivencia por Clase

```
### Supervivencia sengun Sexos
p1 <- ggplot(TitanicPred, aes(x = Sex, fill = Survived_factor)) +
  geom_bar(stat='count', position='dodge') +
  labs(x = 'Sex, TitanicPred data') +
        geom_label(stat='count', aes(label=..count..))

### Supervivencia por Clase
p2 <- ggplot(TitanicPred, aes(x = Pclass, fill = Survived_factor )) +
  geom_bar(stat='count', position='dodge') +
  labs(x = 'Pclass, TitanicPred data') + geom_label(stat='count', aes(label=..count..)) +
  theme(legend.position="none") + theme_grey()

grid.arrange(p1,p2, nrow=1)
```

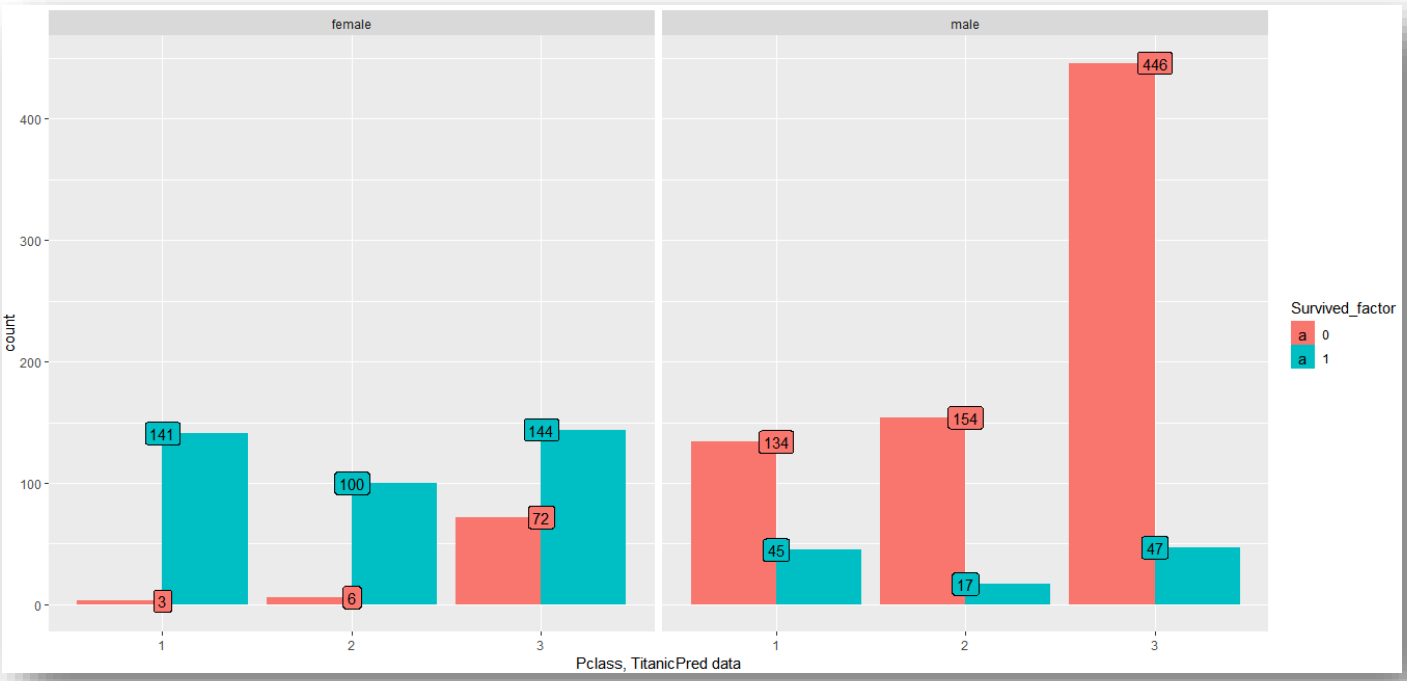


Con la gráfica se comprueba fácilmente que:

- Las mujeres tienen más probabilidades de sobrevivir que los hombres
- De forma similar, los pasajeros de 1º y 2º clase tienen más probabilidades de sobrevivir que los de 3º

Supervivencia por Clase y sexo combinada

```
### Supervivencia por Clase y sexo
p3 <- ggplot(TitanicPred, aes(x = Pclass, fill = Survived_factor )) +
  geom_bar(stat='count', position='dodge') + facet_grid(.~Sex) +
  labs(x = 'Pclass, TitanicPred data') + geom_label(stat='count', aes(label=..count..)) +
  theme(legend.position="none") + theme_grey()
grid.arrange(p3, nrow=1)
```



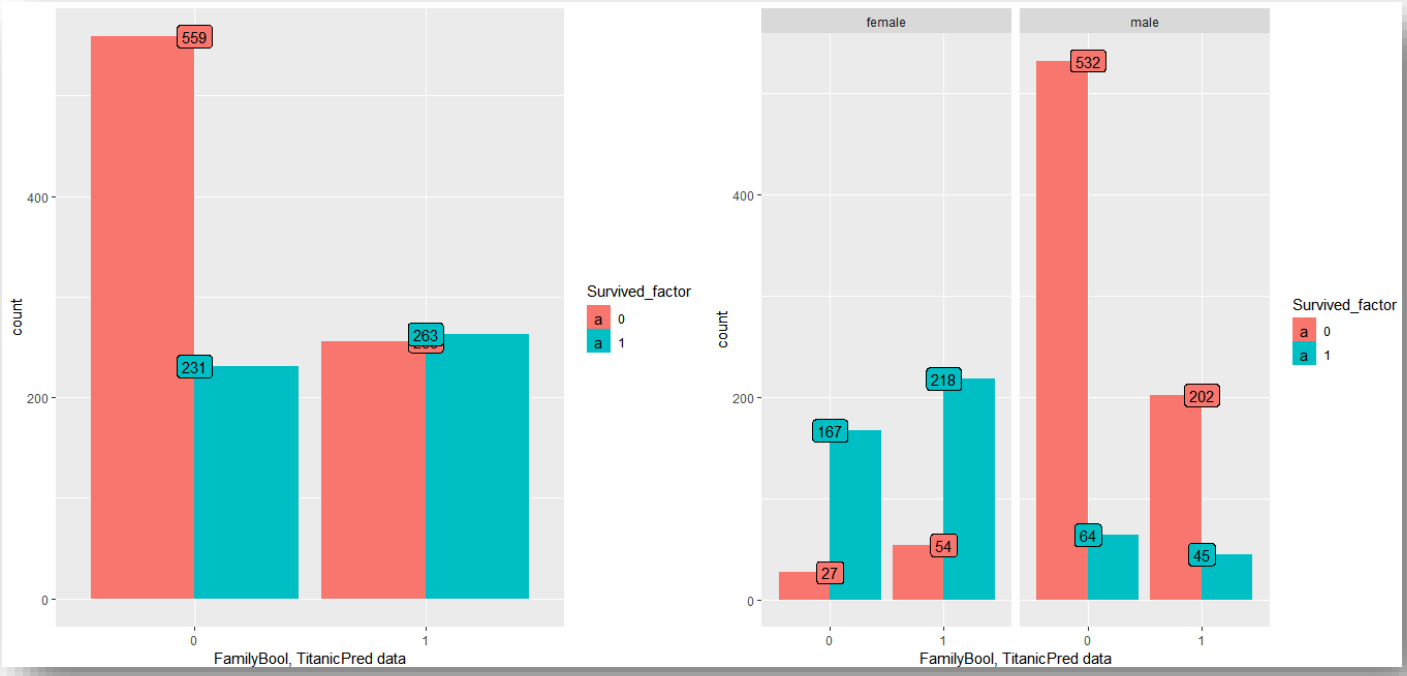
La unión de los criterios anteriores en un único grafico enfatiza la previsión de supervivencia de mujeres de 1º y 2º clase sobre todo los demás.

Supervivencia por Familia y Supervivencia por Familia y sexo

```
### Supervivencia por Familia
p4 <- ggplot(TitanicPred, aes(x = FamilyBool, fill = Survived_factor )) +
  geom_bar(stat='count', position='dodge') +
  labs(x = 'FamilyBool, TitanicPred data') + geom_label(stat='count', aes(label=..count..)) +
  theme(legend.position="none") + theme_grey()

### Supervivencia por Familia y sexo
p5 <- ggplot(TitanicPred, aes(x = FamilyBool, fill = Survived_factor )) +
  geom_bar(stat='count', position='dodge') + facet_grid(.~Sex) +
  labs(x = 'FamilyBool, TitanicPred data') + geom_label(stat='count', aes(label=..count..)) +
  theme(legend.position="none") + theme_grey()

grid.arrange(p4,p5, nrow=1)
```

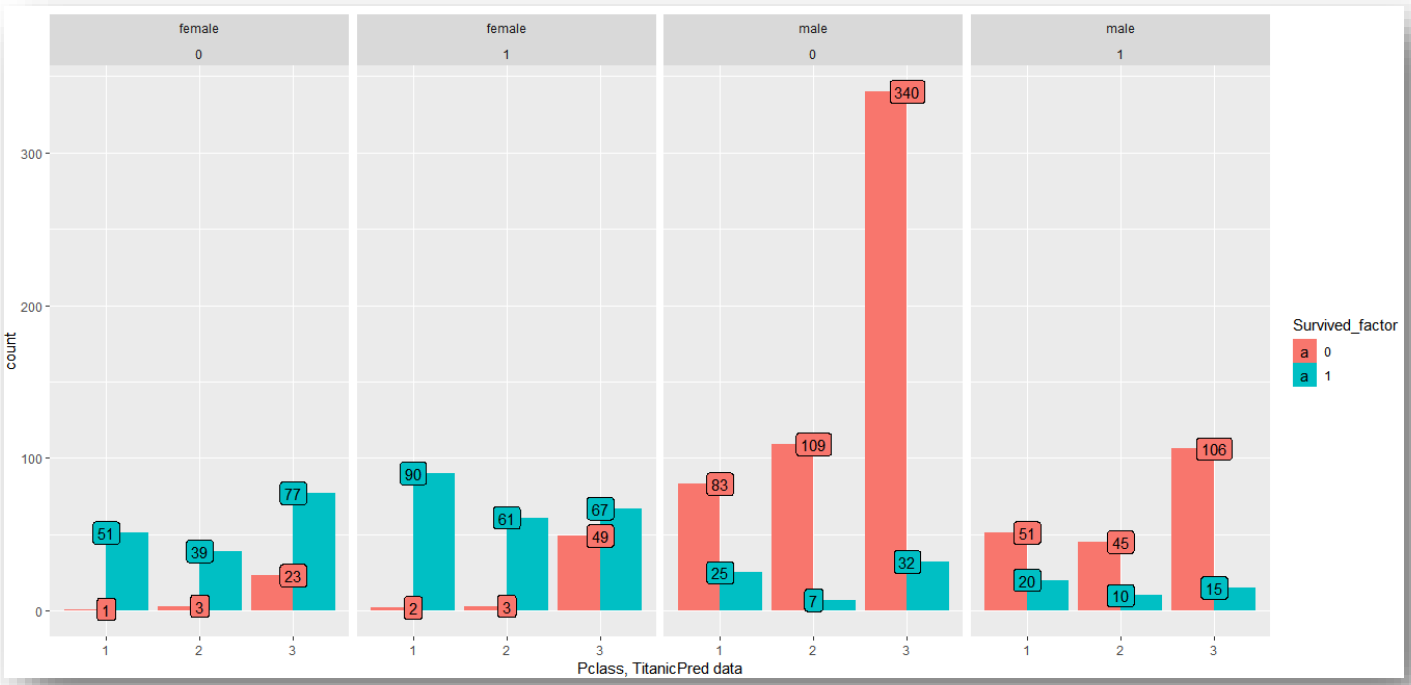


En estos gráficos podemos comprobar que ir con familia aunque no es un factor decisivo, sí que mejora las probabilidades de supervivencia

Si a esto le unimos las probabilidades según sexo, en el caso de los hombres mejora ligeramente ir con familia que sin ella

Supervivencia por Familia sexo y Clase

```
### Supervivencia por Familia sexo y Clase
p6 <- ggplot(TitanicPred, aes(x = Pclass, fill = Survived_factor )) +
  geom_bar(stat='count', position='dodge') + facet_grid(.~Sex+FamilyBool) +
  labs(x = 'Pclass, TitanicPred data') + geom_label(stat='count', aes(label=..count..)) +
  theme(legend.position="none") + theme_grey()
```



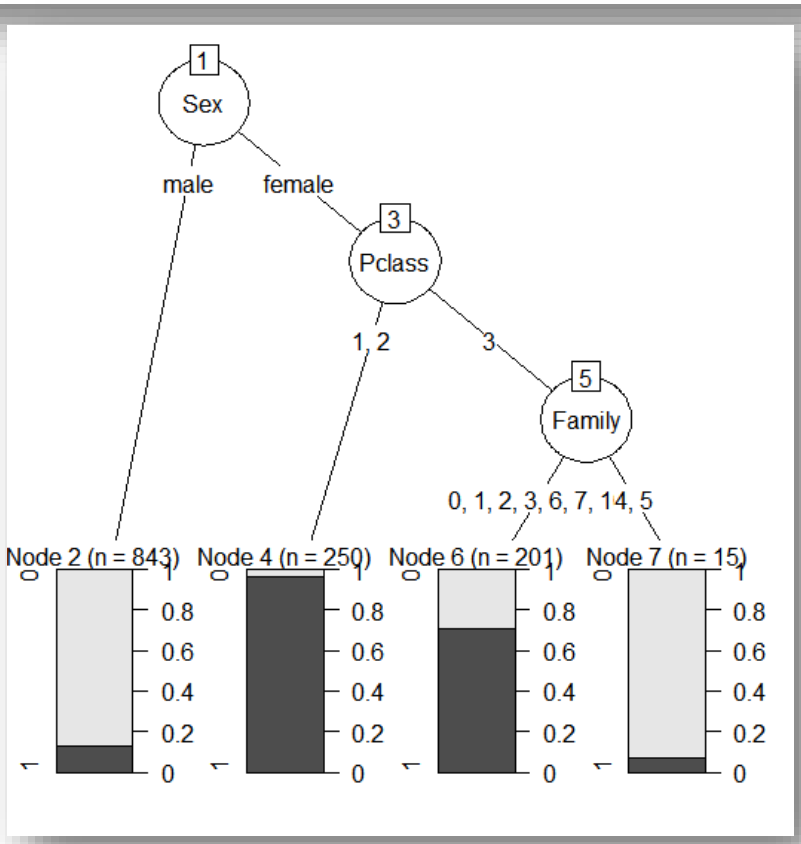
En la combinación de todas las variables, simplemente los criterios más decisivos son Sexo mujer y viajar en 1º y 2º clase, el hecho de viajar en familia, ayuda ligeramente.

JOSE LUIS GOMEZ GARCIA - jgomezgar@uoc.edu

Arboles de decisión permiten crear un modelo a partir de las variable y sus valores, determinando que variables y que valores, en cascada, son los que más probabilidades tienen de éxito.

```
### Arbol de Decision
X <- subset(TitanicPred, select = c('Sex','Pclass','Family','FarePP','Embarked' ))
y <- TitanicPred['Survived_factor']

TitanicDecisionTreeModel <- C50::C5.0( X, y)
summary( TitanicDecisionTreeModel )
plot (TitanicDecisionTreeModel)
```



Decision	Supervivencia	Aciertos	Fallos	% Error
Sex = male: 0 (843/109)	No Sobrevive	843	109	11,4%
Sex = female:				
:...Pclass in {1,2}: 1 (250/9)	Sobrevive	250	9	3,5%
Pclass = 3:				
:...Family in {0,1,2,3,6,7,10}: 1 (201/58)	Sobrevive	201	58	22,4%
Family in {4,5}: 0 (15/1)	No sobrevive	15	1	6,3%
Total				13.5%

Aunque el Error, en general, es alto, existen algunos aprendizajes utilices, como que las Mujeres en 1º o 2º clase tienen un alto grado de supervivencia y que las mujeres en 3º Clase viajando con una familia de 4 o 5 miembros tienen escasa probabilidades de sobrevivir.

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

Conclusiones:

Después de combinar los distintos orígenes de datos, seleccionar las mejores variables, y arreglar algunos valores que no ajustaban al análisis y otros perdidos. Se han aplicado distintas técnicas para mejorar el conocimiento de los datos.

El análisis de correlación y el contraste de hipótesis nos ha permitido conocer cuáles de la variables (características del pasajero) ejercen una mayor influencia sobre las probabilidades de sobrevivir. El árbol de decisión ha confirmado estas, permitiendo combinarlas para una mayor eficacia.

- Los pasajeros de sexo femenino tienen muchas más posibilidades de sobrevivir que los pasajeros de género masculino.
- Los pasajeros de camarotes de primera clase tienen más posibilidades de sobrevivir que el resto. Además, los pasajeros de segunda clase tienen también más posibilidades que los de tercera clase.
- Los pasajeros con familia tienen más posibilidades de sobrevivir.
- La combinación de estos tres factores incrementa significativamente a la supervivencia del pasajero (Mujer en primera clase con familia).
- El precio del billete no afecta significativamente a la supervivencia del pasajero.
- El puerto de embarque usado no afecta significativamente a la supervivencia del pasajero.

Respuesta al problema:

Lamentablemente con los datos disponibles no es posible establecer una serie de patrones universales que para cada persona, predigan su probabilidad de supervivencia dentro de unos márgenes de error aceptables.

A pesar de ello, sí que se pueden establecer algunos patrones de supervivencia, con un grado muy alto de acierto. Pero limitados a una pequeña porción de individuos de la muestra.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Titanic.R

https://github.com/jgomezgar/TitanicPred/blob/master/Codigo/Titanic.R