Práctica 2: Limpieza y validación de los datos

Alumna: Esther Martín González

Fecha: 10/06/2018

Contenido

1. Descripción

El conjunto de datos objeto de análisis se ha obtenido a partir de un enlace en Kaggle (kaggle, 2018) y está formado por dos conjuntos de datos en formato csv los cuales tienes 12 columnas. El fichero train.csv consta de 891 registros y se usará para crear el modelo de aprendizaje automático. El fichero test.csv se usara para realizar pruebas del modelo. Los campos del conjunto de datos son los siguientes:

- PassengerId: Identificador del pasajero.
- Survived:
  1=sobrevive
  0=no sobrevive
- pclass: estado socio-económico (SES)
  1º = superior
  2do = Medio
  Tercero = Más bajo
- Name: nombre del pasajero
- Sex: sexo del pasajero (female,male)
- edad: la edad es fraccional si es menor que 1. Si la edad es estimada, ¿tiene forma de xx.5?
- sibsp: relación familiar
  Hermano = hermano, hermana, hermanastro, hermanastra
  Cónyuge = esposo, esposa
- parch: relación familiar
  Padre = madre, padre
  Niño = hija, hijo, hijastra, hijastro
  Algunos niños viajaron solo con una niñera, por lo tanto parch = 0 para ellos.
- Ticket: Número de billete
- Fare: Tarifa
- Cabin: Número de cabina
- Embarked: Puerta de embarque.
  C = Cherbourg
  Q = Queenstown
  S = Southampton

A través de este conjunto de datos se pretende analizar uno de los naufragios más importantes de la historia. En él se perdieron una gran cantidad de vidas siendo un factor muy importante la falta de botes salvavidas. Debido a la cultura de la época, había más probabilidades de sobrevivir si se pertenecía a un colectivo u a otro. Este análisis tiene una gran relevancia ya que se estudiara la influencia de la mentalidad de la época en la supervivencia analizando los supervivientes al naufragio.

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

Comenzaremos por la lectura de los ficheros en formato CSV que usaremos para la práctica.

```
titanic = read.table("train.csv",fileEncoding="utf-8", header=T, sep="," , dec=".")
test = read.table("test.csv",fileEncoding="utf-8", header=T, sep="," , dec=".")
```

Analizamos nuestro conjunto de datos principal

```
#Análisis del conjunto de datos Titanic
dim(titanic)
summary(titanic)
```

```
> dim(titanic)
[1] 891  12
> summary(titanic)
  PassengerId       Survived          Pclass                                      Name         Sex
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Abbing, Mr. Anthony                  :  1   female:314
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Abbott, Mr. Rossmore Edward          :  1   male  :577
 Median :446.0   Median :0.0000   Median :3.000   Abbott, Mrs. Stanton (Rosa Hunt)     :  1
 Mean   :446.0   Mean   :0.3838   Mean   :2.309   Abelson, Mr. Samuel                  :  1
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000   Abelson, Mrs. Samuel (Hannah Wizosky):  1
 Max.   :891.0   Max.   :1.0000   Max.   :3.000   Adahl, Mr. Mauritz Nils Martin       :  1
                                                  (Other)                              :885
      Age            SibSp           Parch             Ticket         Fare               Cabin     Embarked
 Min.   : 0.42   Min.   :0.000   Min.   :0.0000   1601     :  7   Min.   :  0.00             :687    :  2
 1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000   347082   :  7   1st Qu.:  7.91   B96 B98   :  4   C:168
 Median :28.00   Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45   C23 C25 C27:  4   Q: 77
 Mean   :29.70   Mean   :0.523   Mean   :0.3816   3101295  :  6   Mean   : 32.20   G6        :  4   S:644
 3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000   347088   :  6   3rd Qu.: 31.00   C22 C26   :  3
 Max.   :80.00   Max.   :8.000   Max.   :6.0000   CA 2144  :  6   Max.   :512.33   D         :  3
 NA's   :177                                      (Other)  :852                    (Other)   :186
```

Podemos observar que está formado por 12 campos y 891 registros. Los campos de PassengerId, survived, Pclass, SibSp, Parch, Ticket, Cabin, son numéricos cardinales por lo que no tienen sentido los cálculos aritméticos presentados. Podemos observar rápidamente que los campos Age, Cabin y Embarked presentan nulos.

Del conjunto de datos podemos descartar el campo PassegerIde y Name ya que no tienen relevancia a la hora de analizar la supervivencia. El campo edad sería más interesante tratarlo como un rango ya que lo podríamos limitar a niños y adultos. El número de ticket y las relaciones familiares tampoco son datos relevantes para la supervivencia.

Nuestro conjunto de datos seleccionados serian Survived, Pclass, Sex, Age, Fare, Cabin, embarked.

3. Limpieza de los datos.

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

Analizamos la selección de datos

```
> summary(titanic_1)
    Survived          Pclass          Sex           Age             Fare                Cabin     Embarked
 Min.   :0.0000   Min.   :1.000   female:314   Min.   : 0.42   Min.   :  0.00             :687    :  2
 1st Qu.:0.0000   1st Qu.:2.000   male  :577   1st Qu.:20.12   1st Qu.:  7.91   B96 B98   :  4   C:168
 Median :0.0000   Median :3.000                Median :28.00   Median : 14.45   C23 C25 C27:  4   Q: 77
 Mean   :0.3838   Mean   :2.309                Mean   :29.70   Mean   : 32.20   G6        :  4   S:644
 3rd Qu.:1.0000   3rd Qu.:3.000                3rd Qu.:38.00   3rd Qu.: 31.00   C22 C26   :  3
 Max.   :1.0000   Max.   :3.000                Max.   :80.00   Max.   :512.33   D         :  3
                                               NA's   :177                      (Other)   :186
```

Podemos ver que el campo de Age presenta 177 valores NA's, el campo Cabin 687 nulos y el campo Embarked 2 vacíos. Ya que más de la mitad de los registros del campo Cabin están a 0, vamos a descartar dicho campo ya que podría originar conclusiones erróneas. En cuanto a los campos Age y Embarked se empleará un método kNN-imputation que consiste en rellenar los campos basándose en la similitud o diferencia de los k vecinos más próximos. En el caso del campo Embarked remplazamos los nulos por NA.

```
#Ceros y nulos
suppressWarnings(suppressMessages(library(VIM)))
titanic_1$Age <- kNN(titanic_1)$Age
data = titanic_1$Embarked
data[data==''] <- NA
titanic_1$Embarked=data
titanic_1$Embarked <- kNN(titanic_1)$Embarked
```

3.2. Identificación y tratamiento de valores extremos.

Para identificar los valores que resultan incongruentes con el resto de valores usaremos la función boxplot.stats(). Esta función mostrará los valores atípicos de las variables que los contiene.

```
> boxplot.stats(titanic_1$Survived)$out
integer(0)
> boxplot.stats(titanic_1$Pclass)$out
integer(0)
> boxplot.stats(titanic_1$Age)$out
 [1] 66.0 65.0 71.0 70.5 65.0 64.0 65.0 71.0 64.0 80.0 70.0 70.0 74.0
> boxplot.stats(titanic_1$Fare)$out
  [1]  71.2833 263.0000 146.5208  82.1708  76.7292  80.0000  83.4750
73.5000 263.0000  77.2875 247.5208  73.5000  77.2875
 [14]  79.2000  66.6000  69.5500  69.5500 146.5208  69.5500 113.2750
76.2917  90.0000  83.4750  90.0000  79.2000  86.5000
 [27] 512.3292  79.6500 153.4625 135.6333  77.9583  78.8500  91.0792 1
51.5500 247.5208 151.5500 110.8833 108.9000  83.1583
 [40] 262.3750 164.8667 134.5000  69.5500 135.6333 153.4625 133.6500
66.6000 134.5000 263.0000  75.2500  69.3000 135.6333
 [53]  82.1708 211.5000 227.5250  73.5000 120.0000 113.2750  90.0000 1
20.0000 263.0000  81.8583  89.1042  91.0792  90.0000
 [66]  78.2667 151.5500  86.5000 108.9000  93.5000 221.7792 106.4250
71.0000 106.4250 110.8833 227.5250  79.6500 110.8833
 [79]  79.6500  79.2000  78.2667 153.4625  77.9583  69.3000  76.7292
73.5000 113.2750 133.6500  73.5000 512.3292  76.7292
 [92] 211.3375 110.8833 227.5250 151.5500 227.5250 211.3375 512.3292
78.8500 262.3750  71.0000  86.5000 120.0000  77.9583
```

```
[105] 211.3375  79.2000  69.5500 120.0000  93.5000  80.0000  83.1583
69.5500  89.1042 164.8667  69.5500  83.1583
```

Si nos fijamos en los resultados podemos ver que las variables Survived y Pclass no tienen ningun valor atípico. La variable Age tienen valores atípicos que se corresponden con edades altas, estos valores no los consideraremos extremos ya que son edades a las que puede llegar el ser humano. En cuanto a la varibles Fare, son valores que corresponden a tarifas. Dentro del titanic había camarotes de lujo los cuales tenían una tarifa más elevada que los billetes de tercera clase y eran más escasos por lo que tampoco los consideraremos valores extremos.

3.3. Discretización de los campos

Discretizamos el campo edad ya que nos interesa plantearlo como un grupo de niños y otro de adultos para poder realizar el estudio. No necesitamos el detalle de las edades concretas de cada pasajero.

```
titanic_1 <- within(titanic_1, {
  Age <- Recode(Age, 'lo:18="niño"; 18:hi="Adulto"', as.factor.result=TRUE)
})
titanic_1 <- within(titanic_1, {
  Pclass <- Recode(Pclass, '1="Alta"; 2="Media"; 3="Baja"', as.factor.result=TRUE)
})
titanic_1 <- within(titanic_1, {
  Survived <- Recode(Survived, '0="Muere"; 1="Sobrevive"', as.factor.result=TRUE)
})
```

4. Análisis de los datos.

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

A continuación, se seleccionan los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar.

```
#Agrupación por clase
titanic.claseSuperior <- titanic_1[titanic_1$Pclass.type == 1,]
titanic.claseMedia <- titanic_1[titanic_1$Pclass.type == 2,]
titanic.claseBaja <- titanic_1[titanic_1$Pclass.type == 3,]

#Agrupación por sexo
titanic.mujeres <- titanic_1[titanic_1$Sex.type == "female",]
titanic.hombre <- titanic_1[titanic_1$Sex.type == "male",]

#Agrupación por edad
titanic.niños <- titanic_1[titanic_1$Age.type == "niño",]
titanic.adultos <- titanic_1[titanic_1Age.type == "Adulto",]

#Agrupación por puerta de embarque
titanic.puertaC <- titanic_1[titanic_1$Embarked.type == "C",]
titanic.puertaQ <- titanic_1[titanic_1$Embarqued.type == "Q",]
titanic.puertaS <- titanic_1[titanic_1$Embarqued.type == "S",]
```

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

Para la comprobación de la normalidad usaremos la prueba de normalidad de Anderson-Darling. Si se obtiene un p-valor superior al alpha prefijado (en este caso 0,05) se considera que tiene una distribución normal.

```
> library(nortest)
> alpha = 0.05
> col.names = colnames(titanic_1)
> for (i in 1:ncol(titanic_1)) {
+   if (i == 1) cat("No siguen una distribución normal:\n")
+   if (is.integer(titanic_1[,i]) | is.numeric(titanic_1[,i])) {
+     p_val = ad.test(titanic_1[,i])$p.value
+     if (p_val < alpha) {
+       cat(col.names[i])
+       if (i < ncol(titanic_1) - 1) cat(", ")
+       if (i %% 3 == 0) cat("\n")
+     }
+   }
+ }
No siguen una distribución normal:
Survived, Pclass, Fare
```

La homogeneidad entra varianzas la estudiaremos mediante el test de Fligner-Killeen. Estudiaremos la homogeneidad de la tarifa con las variables de sexo, edad, clase y puerta de embarque.

```
> fligner.test(Fare ~ Sex, data = titanic_1)

        Fligner-Killeen test of homogeneity of variances

data:  Fare by Sex
Fligner-Killeen:med chi-squared = 55.949, df = 1, p-value = 7.436e-14

> fligner.test(Fare ~ Age, data = titanic_1)

        Fligner-Killeen test of homogeneity of variances

data:  Fare by Age
Fligner-Killeen:med chi-squared = 4.1936, df = 1, p-value = 0.04058

> fligner.test(Fare ~ Pclass, data = titanic_1)

        Fligner-Killeen test of homogeneity of variances

data:  Fare by Pclass
Fligner-Killeen:med chi-squared = 365.8, df = 2, p-value < 2.2e-16

> fligner.test(Fare ~ Embarked, data = titanic_1)

        Fligner-Killeen test of homogeneity of variances

data:  Fare by Embarked
Fligner-Killeen:med chi-squared = 136.66, df = 2, p-value < 2.2e-16
```

El caso que devuelve una p-value mayor es Age por lo que el dato más homogéneo con la tarifa es la edad.

4.3. Aplicación de pruebas estadísticas

Para poder analizar mejor que valores de las variables hace que tengamos más probabilidades de supervivencia vamos a generar un árbol de decisión.

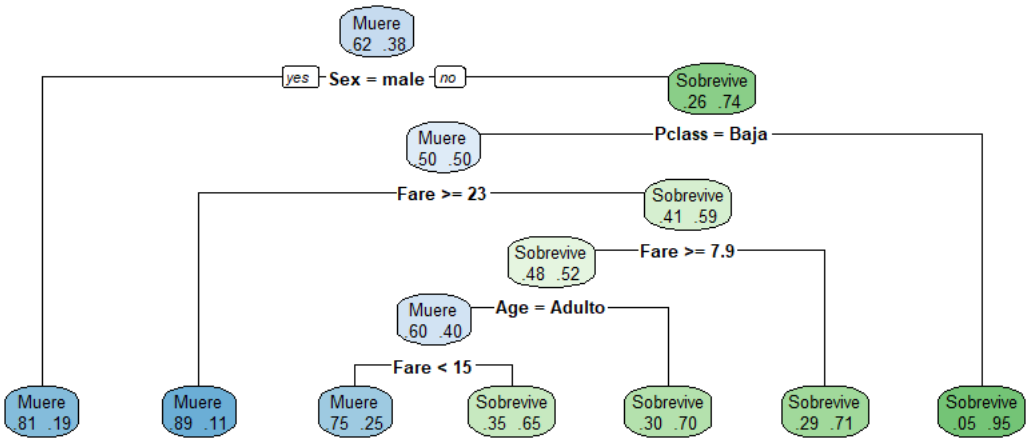
```
# Creación del Arbol de Decision

arbol<-rpart( Survived~ .,method= "class", data=titanic_1)
print(arbol)
rpart.plot(arbol, extra=4)
```

```
n= 891

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 891 342 Muere (0.61616162 0.38383838)
    2) Sex=male 577 109 Muere (0.81109185 0.18890815) *
    3) Sex=female 314  81 Sobrevive (0.25796178 0.74203822)
      6) Pclass=Baja 144  72 Muere (0.50000000 0.50000000)
       12) Fare>=23.35 27   3 Muere (0.88888889 0.11111111) *
       13) Fare< 23.35 117  48 Sobrevive (0.41025641 0.58974359)
         26) Embarked=S 63  31 Muere (0.50793651 0.49206349)
           52) Fare< 10.825 37  15 Muere (0.59459459 0.40540541) *
           53) Fare>=10.825 26  10 Sobrevive (0.38461538 0.61538462)
            106) Fare>=17.6 10   3 Muere (0.70000000 0.30000000) *
            107) Fare< 17.6 16   3 Sobrevive (0.18750000 0.81250000) *
         27) Embarked=C,Q 54  16 Sobrevive (0.29629630 0.70370370) *
      7) Pclass=Alta,Media 170   9 Sobrevive (0.05294118 0.94705882) *
```

Podemos observar que los hombres mueren con una probabilidad del 81% y las mujeres sobreviven con una probabilidad del 25%. Esta supervivencia aumenta hasta el 41% si eres una mujer que ha pagado menos de 23,35 por su billete.

5. Representación de los resultados

El árbol de decisión que hemos generado tiene la siguiente representación:



Este árbol de decisión lo podemos utilizar para saber si es más probable la supervivencia para cada uno de los pasajeros del conjunto de datos test.

```
arboltest <- predict(arbol, newdata = test_1, type = "class")
print(arboltest)
summary(arboltest)
test_1 <- within(test_1, {
  Survived<- arboltest
})
```

Ahora el conjunto de datos tendrá almacenada la variable Survived.

	Pclass	Sex	Age	Fare	Embarked	Survived
1	Baja	male	Adulto	7.8292	Q	Muere
2	Baja	female	Adulto	7.0000	S	Muere
3	Media	male	Adulto	9.6875	Q	Muere
4	Baja	male	Adulto	8.6625	S	Muere
5	Baja	female	Adulto	12.2875	S	Sobrevive
6	Baja	male	niño	9.2250	S	Muere
7	Baja	female	Adulto	7.6292	Q	Sobrevive
8	Media	male	Adulto	29.0000	S	Muere
9	Baja	female	niño	7.2292	C	Sobrevive
10	Baja	male	Adulto	24.1500	S	Muere
11	Baja	male	NA	7.8958	S	Muere
12	Alta	male	Adulto	26.0000	S	Muere
13	Alta	female	Adulto	82.2667	S	Sobrevive
14	Media	male	Adulto	26.0000	S	Muere
15	Alta	female	Adulto	61.1750	S	Sobrevive
16	Media	female	Adulto	27.7208	C	Sobrevive
17	Media	male	Adulto	12.3500	Q	Muere
18	Baja	male	Adulto	7.2250	C	Muere

6. Resolución del problema.

Para poder solucionar el problema del estudio de la supervivencia de los diferentes pasajeros del Titanic, hemos comenzado seleccionando los datos que nos resultaban de interés y realizándoles un preprocesamiento para garantizar que eran correctos.

Una vez teníamos los datos preparados, hemos obtenido un árbol de decisión en el que nos hemos dado cuenta de que los hombres tenían muy poca probabilidad de supervivencia. Esto se debe a la mentalidad de la época de salvar primero a las mujeres y a los niños y a la falta de botes salvavidas. Gracias al árbol de decisión desarrollado hemos podido completar los datos del conjunto test con la variable Survived, la cual indica si lo más probable es la supervivencia o la muerte del pasajero.

Este árbol de decisión se podría utilizar para prevenir futuras naufragios, colocando más botes salvavidas en las zonas donde las probabilidades de muerte son mayor para intentar así bajar esa probabilidad. En el caso del Titanic hubiera sido recomendable colocar más botes en las zonas de clase baja ya que las mujeres que pertenecían a la clase baja tenían una probabilidad de morir del 50%, con más botes salvavidas esta probabilidad podría ser más baja.

Referencias

1. Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and
2. Peter Dalgaard (2008). Introductory statistics with R. Springer Science &
3. Business Media. techniques. Morgan Kaufmann. Vegas, E. (2017). Preprocesamiento de datos. Material UOC.