Práctica 2

Limpieza y Validación de datos

Introducción

El dataset forma parte de la famosa competición abierta de Kaggle: “Titanic. Learning from disaster”. Se trata de una competición educativa donde el objetivo es predecir quién sobrevivió al hundimiento en base a ciertos atributos conocidos sobre los pasajeros.

Para entender mejor el dataset, se adjunta el siguiente diccionario de datos:

Nombre	Tipo	Descripción
PassengerId	Entero	Identificador autoincremental del pasajero. Empleado por Kaggle para subir la respuesta. Inútil para el análisis.
Survived	Binario	Etiqueta a predecir. Valor binario que indica si sobrevivió o no.
Pclass	Categórico	Clase en la que viajaba (1, 2 o 3).
Name	String	Nombre, apellidos y título. Podría ser útil para agrupar por apellidos o usar el título. No va a usarse ya que la edad el sexo y el ticket sirven para extraer toda esa información. Aun así, debería experimentarse para juzgar su utilidad. En este ejemplo no va a hacerse.
Sex	Categórico	Binario, hombre o mujer.
Age	Real	Edad. Para pasajeros muy jóvenes, tiene parte decimal.
SibSp	Entero	Número de hermanos/as o pareja/s a bordo.
Parch	Entero	Número de hijos/as o progenitores a bordo.
Ticket	String	Cadena que identifica cada ticket. Pasajeros que viajan juntos, en la gran mayoría de casos, tienen el mismo ticket. Puede comprobarse por coincidencia de apellidos y de valores de las dos variables anteriores.
Fare	Real	Coste del ticket que pagó el pasajero.
Cabin	String	Camarote en el que viajaba
Embarked	Categórico	Ciudad desde la que embarcó. Tres opciones (S, C o Q).

Banco de Pruebas

En el proceso de limpieza, imputación de valores vacíos, selección o creación de atributos, etc…, es necesario tomar decisiones. Dichas decisiones no deben basarse, siempre que sea posible, en nuestras propias convicciones. En lugar de ello, es necesario tener siempre en mente cuál es el problema que se intenta resolver, y decidir acorde a ello.

Por esto, para decidir qué conjunto de datos es más apropiado para resolver el problema, se ha desarrollado un banco de pruebas sobre el que testear los modelos resultantes. Este banco de pruebas consiste en un modelo Random Forest muy simple y una estimación de su error mediante validación cruzada. Se prueban dos modelos resultantes de tomar dos conjuntos de datos distintos. Los resultados obtenidos de cada pliegue en la validación cruzada se introducen

en un test t-Student para determinar si la diferencia en precisión es estadísticamente significativa. Esto permite determinar si los cambios realizados en el conjunto de datos mejoran los resultados, los empeoran o no tienen ningún efecto.

El motivo de escoger Random Forest es simplemente por su gran versatilidad:

- No es sensible a atributos covariantes (salvo que queramos determinar la importancia de atributos).
- No es sensible a variables categóricas con un alto número de niveles (siempre que no las codifiquemos con one-hot, y evidentemente, que tengamos un número suficiente de observaciones con cada uno de sus valores como para poder apreciar su efecto, pero esto último es una cuestión de teoría de información y afecta a cualquier modelo).
- Es robusto a cambios en la mayoría de sus parámetros. Si el número de árboles es suficientemente elevado, el único parámetro verdaderamente relevante es el número de atributos a considerar por bifurcación.
- Es modestamente rápido de entrenar, ya que no hay que hacer un *tunning* muy fino y los árboles son independientes (pueden crecerse en paralelo).

Esto no significa que, una vez preparado el conjunto de datos, este sea el modelo a escoger. A posteriori, habrá que seleccionar el modelo en función del problema, las restricciones de implementación y funcionamiento...

Los datos aparecen divididos en dos subconjuntos: uno de prueba y otro de entrenamiento. El de prueba no hay que tocarlo, no es correcto realizar todas las pruebas sobre el conjunto de test ya que al final se estarían sobreajustando todas las decisiones a este subconjunto. El conjunto de prueba es exclusivamente para asegurarse de que no hay sobreajuste, no para estimar el error de cada decisión que tomemos.

Exploración de Datos

Lo primero que vamos a comprobar es si existen valores problemáticos para el posterior análisis: elementos vacíos y outliers.

Los elementos vacíos pueden ser:

- NAs, es decir, directamente no hay ningún valor.
- Cadenas vacías. Hay un valor, pero es una cadena de texto que no contiene ningún carácter.
- Valores de relleno. Típicamente ceros, pero no es la única opción. En ocasiones se utilizan otras cantidades, como -999. Estos valores son fáciles de pasar por alto y nos transmiten información errónea. En importante distinguir entre los ceros de relleno y los reales.

```
library(data.table)
data <- fread('data/train.csv')
t(data[, lapply(.SD, function(x) {sum(is.na(x) | x=="")/.N} )])
```

Esto nos devuelve la siguiente información:

```
PassengerId 0.000000000
Survived    0.000000000
```

```
Pclass      0.000000000
Name        0.000000000
Sex         0.000000000
Age         0.198653199
SibSp       0.000000000
Parch       0.000000000
Ticket      0.000000000
Fare        0.000000000
Cabin       0.771043771
Embarked    0.002244669
```

Es decir, la mayoría de variables no tienen problemas. La edad tiene casi un 20% de valores faltantes. La cabina cerca del 80%. La ciudad de embarque tiene dos valores en todo el dataset.

A continuación, se comprueban los ceros.

```
t(data[, lapply(.SD, function(x) {sum(x==0, na.rm=T)/.N} )])
```

```
PassengerId 0.00000000
Survived    0.61616162
Pclass      0.00000000
Name        0.00000000
Sex         0.00000000
Age         0.00000000
SibSp       0.68237935
Parch       0.76094276
Ticket      0.00000000
Fare        0.01683502
Cabin       0.00000000
Embarked    0.00000000
```

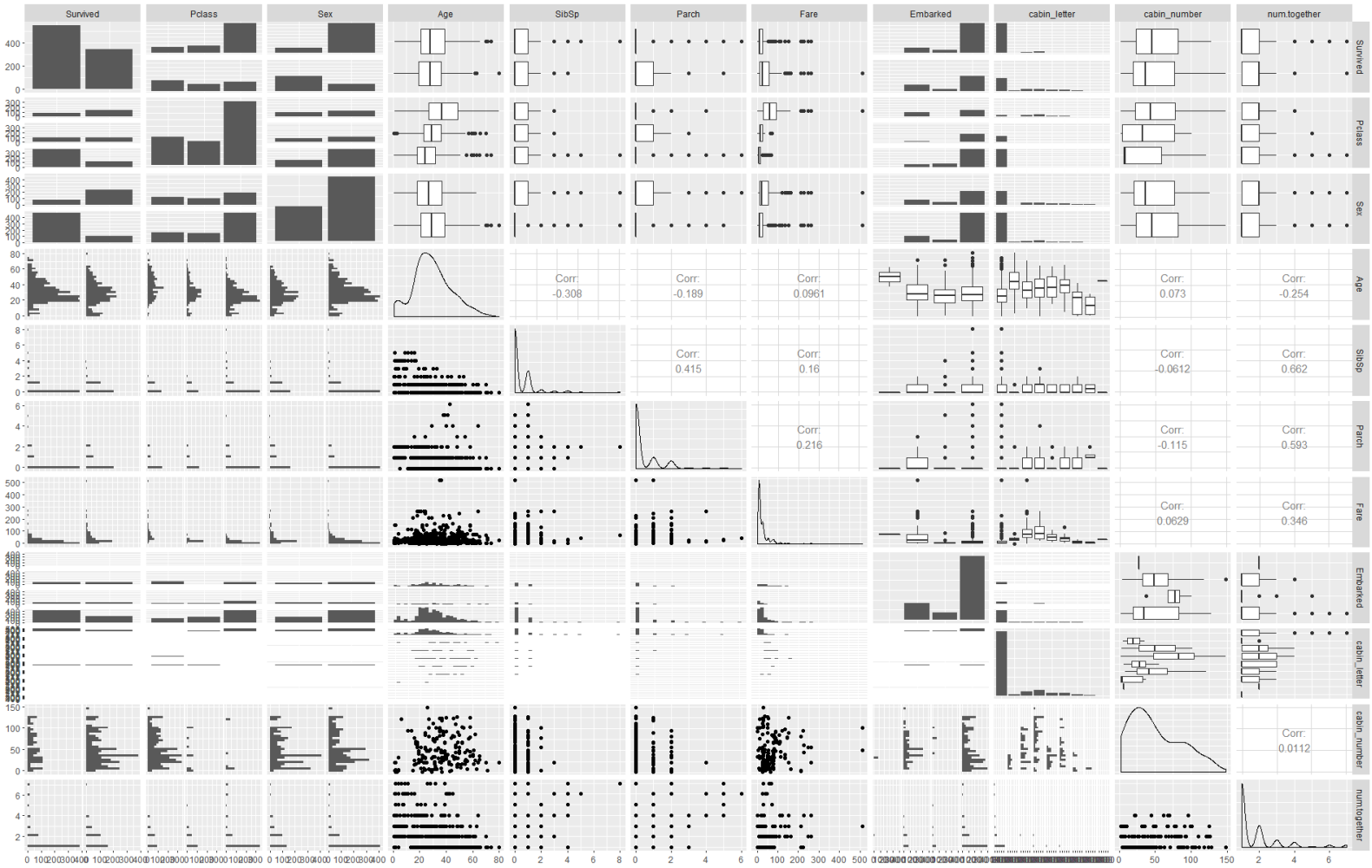
Existe un gran número de ceros en algunas variables, pero esto es coherente con su significado. También hay algunos pocos ceros en la variable “Fare”. Esto podría ser información real. En principio lo desconocemos y será necesario probarlo en el banco de pruebas.

Finalmente, vamos a visualizar gráficamente las distribuciones y dependencias. Esto permite valorar qué variables dependen de cuáles otras, cuáles parece q no tengan influencia, cuáles presentan valores extremos…

Es difícil apreciar en el tamaño que permite el documento, la imagen completa puede consultarse en plots/pairs1.png. El gráfico permite apreciar algunas cosas. Por ejemplo, da la impresión que “Fare” contiene algunos valores anómalos.

También puede apreciarse como ciertas variables, como Sex, Age o Pclass, tienen una aparente gran influencia sobre la variable respuesta, mientras que otras, como SibSp o Parch no parece que influyan demasiado.

Toda observación derivada de este tipo de visualizaciones debe corroborarse experimentalmente. Sin embargo, siempre es útil emplearlas para guiar el análisis, disminuyendo la cantidad de opciones que probar.



IDENTIFICACIÓN DE OUTLIERS!!

COMPROBACIÓN DE NORMALIDAD Y HOMOGENEIDAD DE LA VARIANZA

REPRESENTACIÓN GRÁFICA

Limpieza de Datos

Para limpiar los datos, vamos a proceder variable por variable. Primero resolveremos las variables categóricas y seguidamente las continuas.

Cabin

Esta variable tiene una grandísima cantidad de valores faltantes. Además, se trata de una variable categórica con una enorme cantidad de valores diferentes.

```
uniqueN(data$Cabin)
148

sum(data$Cabin != "")
204
```

La opción más lógica sería eliminarla directamente. Aun así, va a probarse a intentar extraer información. Se va a experimentar con la letra del camarote (hay de la A a la F) y el número (quizás los números se encontraban cercanos, y ciertos números tenían más difícil salida).

```
data[, `:=`(
  cabin_letter = substr(Cabin, 1, 1),

  cabin_number = as.integer(unlist(lapply(strsplit(substr(Cabin, 2,
999), " "), `[`, 1)))
)]
data_cabin <- data[Cabin != "", .(Survived, cabin_letter,
cabin_number)]

cabin.aov <- aov(cabin_number ~ Survived, data=data_cabin)
summary(cabin.aov)

             Df Sum Sq Mean Sq F value Pr(>F)
Survived      1    746   745.9   0.586  0.445
Residuals   194 247020  1273.3
```

El número de la cabina no puede considerarse significativo.

chisq.test(data_cabin$Survived, data_cabin$cabin_letter)

```
      Pearson's Chi-squared test

data:  data_cabin$Survived and data_cabin$cabin_letter
X-squared = 10.301, df = 7, p-value = 0.1722
```

La letra está cerca de poder considerarse significativa, sin embargo, habría que lidiar con el gran número de valores faltante. En vista de las evidencias, no se considera justificable emplear la variable Cabin.

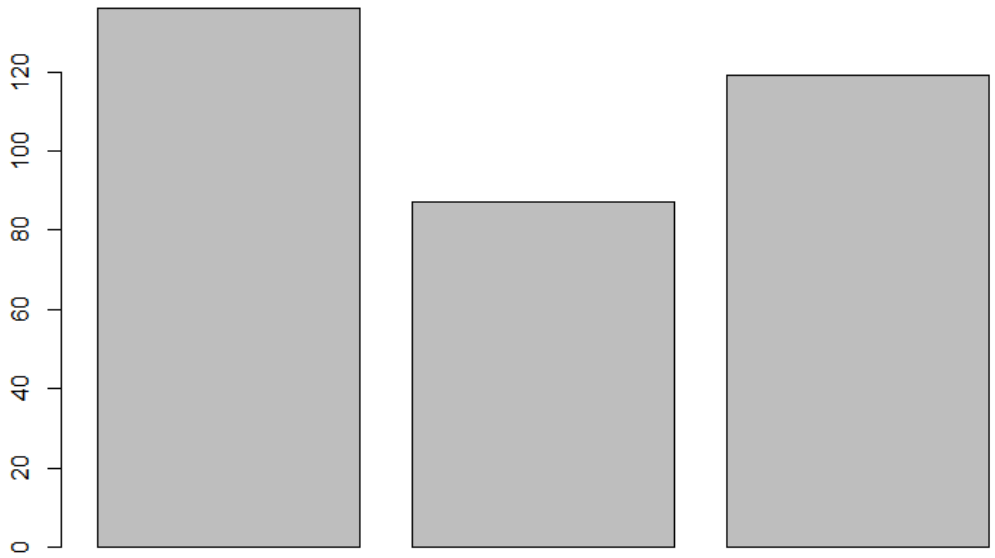
Ticket

Se trata de una variable categórica con un gran número de valores únicos y que no aporta información aparente. Sin embargo, los pasajeros que viajan juntos casi siempre llevan el mismo número de ticket (salvo alguna excepción donde llevan tickets consecutivos). Se añadirá esta información para complementar a SibSp y Parch. Más adelante se comprobará si es útil o no.

Por tanto, se crea una nueva columna con el número de veces que aparece cada ticket. La variable original de ticket se elimina. Se observa además que hay distinto tipo de numeración de los tickets. Esto podría tal vez aportar información. En este ejemplo no va a emplearse dicha posible información adicional.

Pclass

Este valor aparece bastante limpio originalmente. La lectura lo malinterpreta como entero, es necesario transformarlo a categórico. Exceptuando eso, no es necesaria ninguna transformación adicional. Se ha comprobado que efectivamente sólo existen tres valores posibles. En cuanto a su influencia en la variable respuesta:




```
chisq.test(data$Pclass, data$Survived)

      Pearson's Chi-squared test

data:  data$Pclass and data$Survived
X-squared = 102.89, df = 2, p-value < 2.2e-16
```

Con lo que la sospecha original se confirma, la clase en la que viajaban los pasajeros fue muy influyente en su supervivencia.

Sex

Le ocurre lo mismo, por defecto aparece bastante limpia. Sólo hay dos valores posibles en todo el dataset, sin errores tipográficos y demás.

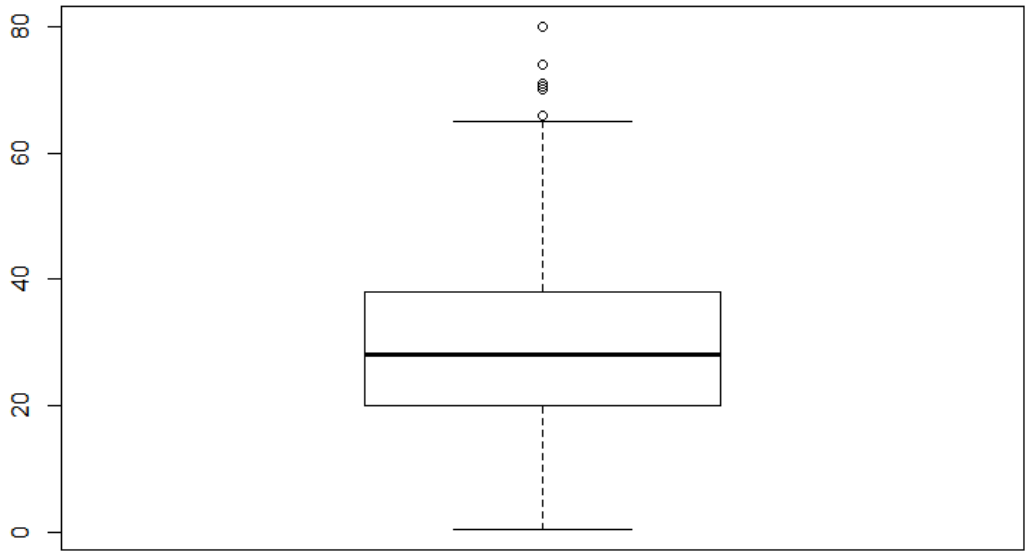
```
chisq.test(data$Sex, data$Survived)

      Pearson's Chi-squared test with Yates' continuity correction

data:  data$Sex and data$Survived
X-squared = 260.72, df = 1, p-value < 2.2e-16
```

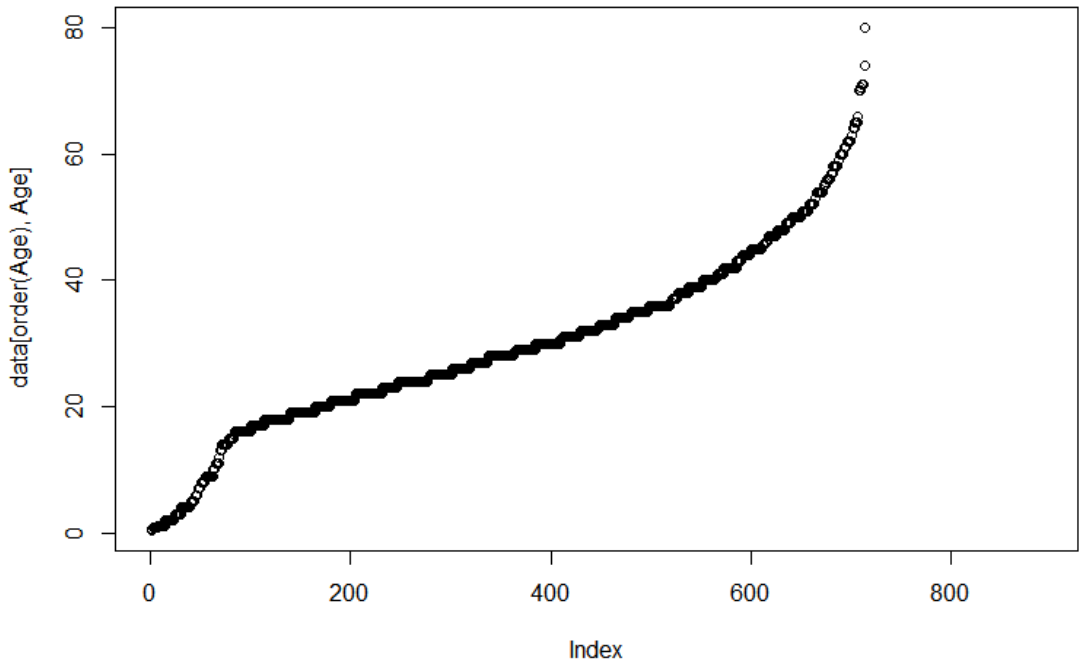
Seguidamente, se estudian las variables numéricas.

Age

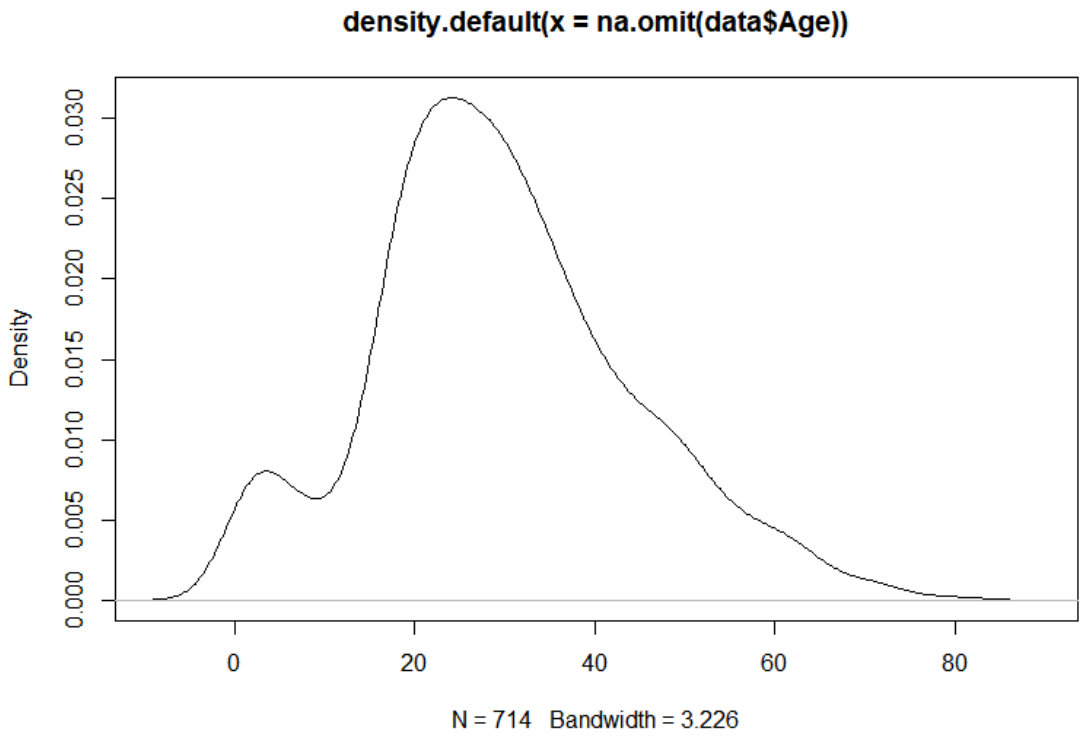
Esta variable es la más problemática, ya que se encuentra en el límite en el que hay suficientes valores como para emplearla, hay tantos faltantes que no se puede simplemente eliminar los registros donde esta variable falta, ya que el volumen de datos disponible es intencionadamente bastante escaso.

¿Existen valores anómalos? No puede afirmarse que ningún valor sea erróneo. Todo parece dentro de lo razonablemente esperable.





Existen bastante menos pasajeros menores de 15 años, y también algunos más de lo normal de edad mayor que 60. Sin embargo, nada fuera de lo que parece razonable.



A pesar de que los datos no son del todo normales, los tests de normalidad devuelven resultados positivos con una muy alta confianza. Esto se debe a que el número de muestras es

muy elevado, las muestras son más o menos iid (salvo las pocas relaciones de matrimonio o padre-hijo) y la varianza, evidentemente como debe tratarse para la edad, es finita.

Se han probado varias soluciones en el banco de pruebas:

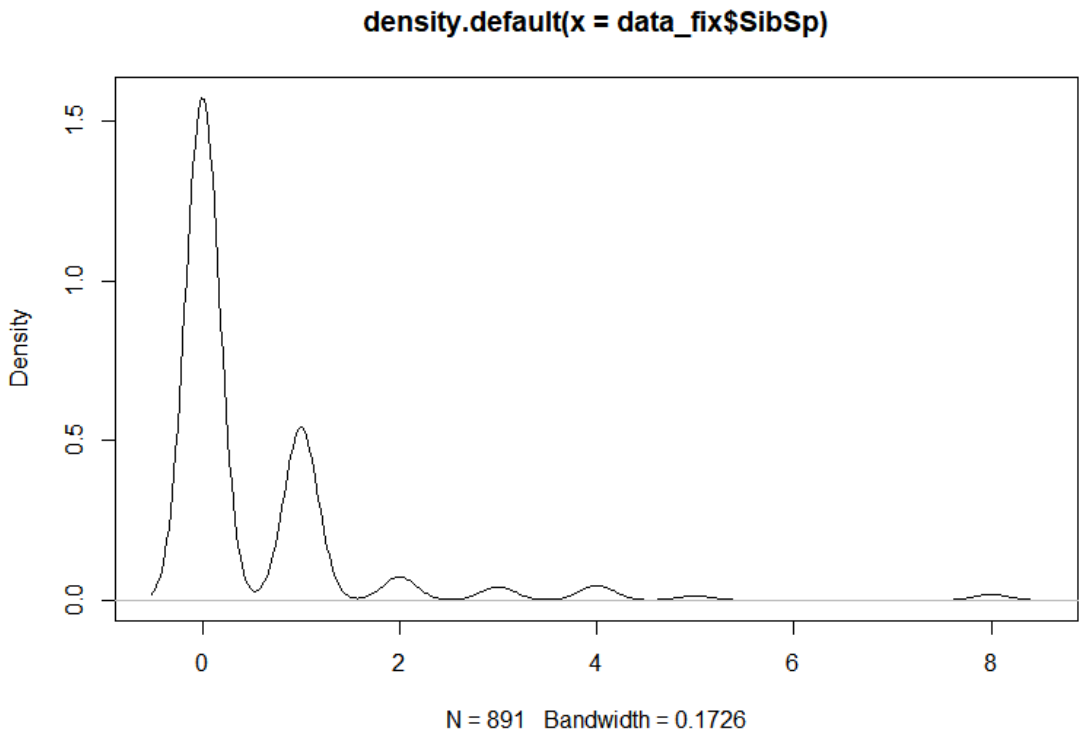
- Omitir filas que no tienen edad.
- Omitir la columna edad.
- Imputar con la media del resto de valores.
- Imputar con un modelo lineal empleando algunas de las otras variables.
- Imputar con valores aleatorios empleando la media y desviación típica del resto de valores.

La primera opción no da muy buenos resultados ya que existe una importante escasez de datos en el problema. Las siguientes cuatro no presentan grandes diferencias. Omitir la edad es algo peor. El resto presentan resultados que no difieren de manera significativa. En base a estas evidencias, escoger cualquiera de los tres últimos métodos debe ser equivalente.

SibSp

En el gráfico original, daba la impresión de que esta variable no tiene una gran influencia con la variable respuesta.

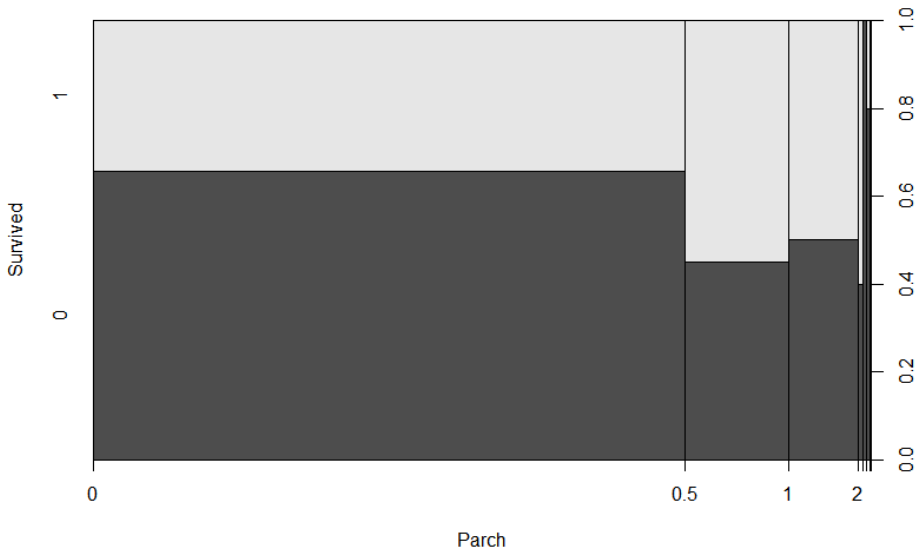
El motivo es que la gran mayoría de valores son 0. Sólo hay unos pocos que sobrepasan 4. Si bien es cierto que la frecuencia de salvados con esas cifras es mucho más elevada, el número de observaciones es tan bajo que no puede garantizarse que estén relacionadas.

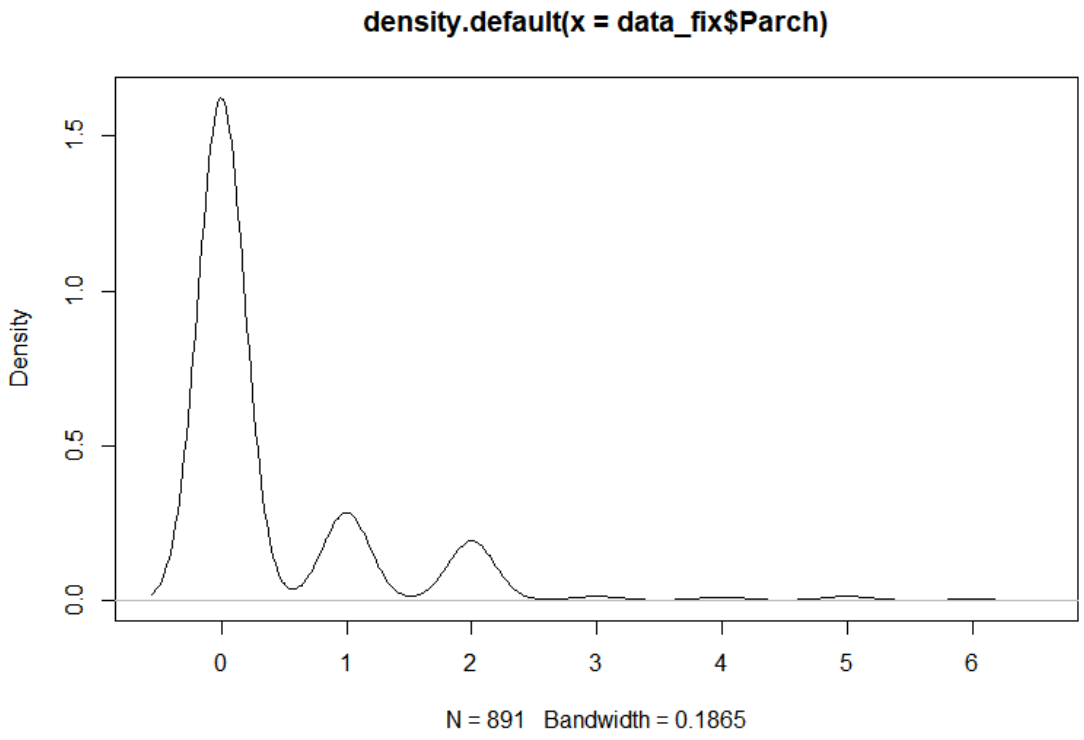


En principio, la variable se mantiene, pero no se va a hacer nada más con ella. Si más adelante se estima que es mejor quitarla, así se hará.

Parch

Todo lo reflexionado en el apartado anterior, es aplicable para esta variable.





Las distribuciones son casi idénticas.

Fare

La tarifa pagada presenta algunos valores extremos, aunque se desconoce si son legítimos. Los valores más elevados pertenecen todos a pasajeros de primera clase, muchos de ellos viajando en grupo. La tarifa que se muestra es la suma completa. Por ello, muy probablemente sean valores reales.

```
> data_fix[Fare >200]
    Survived Pclass    Sex      Age SibSp Parch     Fare Embarked num.together
 1:        0      1   male 19.00000     3     2 263.0000        S            4
 2:        1      1 female 23.00000     3     2 263.0000        S            4
 3:        0      1   male 24.00000     0     1 247.5208        C            2
 4:        1      1 female 35.00000     0     0 512.3292        C            3
 5:        1      1 female 50.00000     0     1 247.5208        C            2
 6:        1      1 female 18.00000     2     2 262.3750        C            2
 7:        1      1 female 24.00000     3     2 263.0000        S            4
 8:        0      1   male 27.00000     0     2 211.5000        C            1
 9:        1      1 female 42.00000     0     0 227.5250        C            4
10:        0      1   male 64.00000     1     4 263.0000        S            4
11:        0      1   male 42.12876     0     0 221.7792        S            1
12:        0      1   male 42.12876     0     0 227.5250        C            4
13:        1      1   male 36.00000     0     1 512.3292        C            3
14:        1      1 female 15.00000     0     1 211.3375        S            3
15:        1      1 female 18.00000     1     0 227.5250        C            4
16:        1      1 female 38.00000     0     0 227.5250        C            4
17:        1      1 female 29.00000     0     0 211.3375        S            3
18:        1      1   male 35.00000     0     0 512.3292        C            3
19:        1      1 female 21.00000     2     2 262.3750        C            2
20:        1      1 female 43.00000     0     1 211.3375        S            3
>
```

Un test ANOVA devuelve la siguiente información:

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Fare          1  13.95  13.952   63.03 6.12e-15 ***
Residuals   889 196.78   0.221
```

Puede decirse que la tarifa pagada es de gran influencia en la probabilidad de supervivencia. Esto es coherente con lo ya descubierto sobre si se viaja en primera, segunda o tercera clase.

Se va a crear la variable de tarifa por número de pasajeros que viajan juntos. El test de dependencia devuelve:

```
             Df Sum Sq Mean Sq F value   Pr(>F)
fare_pass     1  13.68  13.684   61.74 1.13e-14 ***
Residuals   889 197.04   0.222
```

Parece que es más o menos igual de dependiente, serán los modelos los que decidirán si es mejor una, otra o ambas.

Resultados

Para terminar, se va a seleccionar un subconjunto de variables de entrada que sea el más apropiado. Para ello se emplea la función de eliminación recursiva de atributos mediante Random Forest de caret. Esta función determina lo siguiente:

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold, repeated 5 times)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         1   0.7867 0.5421    0.03881 0.08380
         2   0.7744 0.4995    0.02955 0.06994
         3   0.7896 0.5332    0.03654 0.08309
         4   0.8307 0.6317    0.04307 0.09564
         5   0.8327 0.6364    0.04119 0.09243        *
         6   0.8271 0.6223    0.03709 0.08314
         7   0.8240 0.6164    0.04227 0.09246
         8   0.8265 0.6214    0.04141 0.09278

The top 5 variables (out of 5):
   Sex, Pclass, Age, Fare, num.together
```

Por tanto, se empleará el subconjunto formado por las 5 variables seleccionadas.

Conclusiones sobre los datos

Los datos contienen suficiente información como para resolver el problema con cierta precisión. Que esta precisión sea o no suficiente, dependerá de la aplicación de la solución. Además, en el documento no se ha tratado sobre las técnicas que se emplearían para modelar la solución.

Código

El código se adjunta en la carpeta /src del siguiente repositorio:

https://github.com/miguel-rodrigo/Tipologia_ciclo_vida_datos_pract2

Integrantes del grupo

La práctica se ha realizado en solitario. Se ha intentado escribiendo en el foro sin respuesta, y escribiendo mensajes privados, sólo uno con respuesta para indicar que ya tenía compañero. Esto se debe a que no tuve compañero en la primera práctica, y entiendo por el poco movimiento del foro que la mayoría volvió a trabajar con el mismo.

Contribuciones	Firma
Investigación previa	Miguel A. Rodrigo Lisbona
Redacción de las respuestas	Miguel A. Rodrigo Lisbona
Desarrollo código	Miguel A. Rodrigo Lisbona