PRAC2: Limpieza y análisis de datos

Máster Universitario en Ciencia de Datos (UOC)

Tipología y ciclo de vida de los datos

Daniel Núñez Sanz

1. Descripción del dataset.

El conjunto de datos se ha obtenido de Kaggle correspondiente a la competición Titanic - Machine Learning from Disaster. El conjunto original está formado por dos bases de datos: una para uso como entrenamiento y otra para test. Por lo tanto, para la práctica usaremos solo el dataset train donde tenemos todos los datos completos y se compone de 891 casos.

Este conjunto de datos contiene a la información de pasajeros que viajaron en el Titanic, donde podemos ver sus características personales como las relacionadas con el viaje.

La variable objeto a predecir es Survived, si el pasajero sobrevive o no, la cual es muy indicada en unos datos sobre el viaje del Titanic. De esta manera, intentaremos responder a la pregunta de si la supervivencia o fallecimiento de los pasajeros fue solo algo aleatorio o hay variables que pudieron influir más allá de la suerte.

En el dataset nos encontramos las siguientes variables:

- Passenger Id: identificador único del pasajero.
- survived: si el pasajero sobrevive. Puede tomar los valores 0 = No, 1 = Si
- pclass: Clase a la que pertenece el ticket del pasajero. Puede tomar los valores 1 = Alta, 2 = Media, 3 = Baja
- name: nombre del pasajero.
- sex; Sexo del pasajero.
- Age: Edad en años del pasajero.
- sibsp: número de hermanos, hermanas, hermanastros o hermanastras en el barco.
- parch: número de padres e hijos en el barco.
- ticket: identificador del billete.
- fare: precio pagado por el billete.
- cabin: identificador del camarote asignado al pasajero.
- embarked: puerto de embarque del pasajero. Puede tomar los valores C = Cherbourg, Q = Queenstown, S = Southampton

Cargamos los datos y podemos ver la cabecera de los datos para hacernos una idea de ellos como sus estadísticos principales.

```
> head(titanic)
  PassengerId Survived Pclass                                                 Name    Sex Age SibSp Parch
1           1        0      3                              Braund, Mr. Owen Harris   male  22     1     0
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
3           3        1      3                               Heikkinen, Miss. Laina female  26     0     0
4           4        1      1       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
5           5        0      3                             Allen, Mr. William Henry   male  35     0     0
6           6        0      3                                     Moran, Mr. James   male  NA     0     0
            Ticket    Fare Cabin Embarked
1        A/5 21171  7.2500              S
2         PC 17599 71.2833   C85        C
3 STON/O2. 3101282  7.9250              S
4           113803 53.1000  C123        S
5           373450  8.0500              S
6           330877  8.4583              Q
```

```
> str(titanic)
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
 $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

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

Viendo el dataset y las variables que lo componen puede ser interesante ver que variables han podido influir a la hora de sobrevivir o no al Titanic. Podemos plantear la siguiente hipótesis. La clase social puede influir a la hora de sobrevivir, de manera que si eres de clase alta y probablemente con un precio del billete mas alto estés en unos camarotes mejor comunicados con las salidas de emergencia o que haya más flotadores, lo que hace que sea más probable que puedan sobrevivir.

3. Limpieza de los datos.

3.1. Valores perdidos

Unificamos los valores perdidos y analizamos la base en búsqueda de valores perdidos, vacíos o erróneos.

```
> sapply(titanic, function(x){sum(is.na(x))})
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch      Ticket
          0           0           0           0           0         177           0           0           0
       Fare       Cabin    Embarked
          0         687           2
```

Vemos que tenemos valores perdidos tanto en la edad, en la Cabina y en la puerta de embarque.

```
> sapply(titanic, function(x){100*sum(is.na(x))/length(x)})
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch      Ticket
   0.000000    0.000000    0.000000    0.000000    0.000000   19.865320    0.000000    0.000000    0.000000
        Fare       Cabin    Embarked
   0.000000   77.104377    0.224467
```

Viendo la cantidad de perdidos por variable habrá que decidir qué hacer con ellos. En Cabin vemos que nos falta casi el 80% de los datos, dada esta cantidad de daos perdidos no tiene sentido hacer ninguna imputación por lo que lo mejor es o eliminarla o no tenerla en cuenta a la hora del análisis. Nosotros la eliminamos.

En el caso de Embarked y Sex no son tantos y si nos puede merecer la pena hacer alguna imputación para poder usar esas variables en caso de que sea necesario.

En el caso de la edad vamos a analizar la variable para decidir el reemplazo de valores, vemos que los valores extremos no influyen mucho en la variable así que el reemplazo de valores perdidos lo haremos por la media.

```
> summary(titanic$Age, na.rm = TRUE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   20.00   28.00   29.68   38.00   80.00     177
```

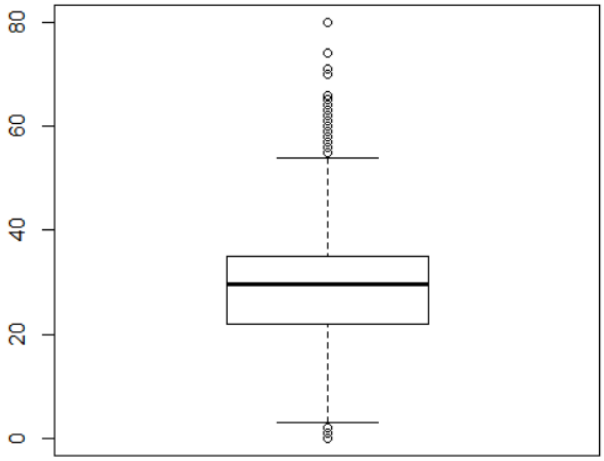
En el caso de Embarked vemos que solo tenemos dos valores perdidos y que la mayoría de casos (72%) toman el valor S, por lo que reemplazaremos los perdidos por este valor mayoritario.

```
> summary(titanic$Embarked, na.rm = TRUE)
   C    Q    S NA's
 168   77  644    2
```

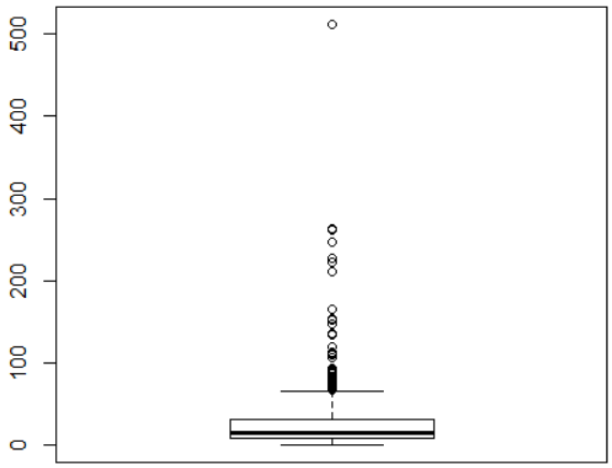
3.2. Identificación y tratamiento de valores extremos.

Graficamos todas las variables para ver su distribución y buscar valores extremos.

En el caso de la edad vemos que hay algunos valores que se salen de la centralidad del resto, pero son valores posibles dentro del rango de la edad ya que el rango de nuestra variable va de 0 a 80 años. Por lo que no debemos hacer nada con ellos ya que representan a parte de los pasajeros.



En el caso de Fare vemos que también hay valores que se escapan de esa centralidad. Por un lado tenemos valores tres casos con valores muy alto pero vemos que pertenecen a primera clase por lo que es posible que hubieran pagado eso, como también los que han pagado 0, podría ser que les hubieran regalado el viaje. El resto que se escapan por encima entendemos que es lógico que pase ya que la mayoría de las personas pagaran un precio medio y luego solo algunos pagarían por la primera clase que es más cara. Por lo que parece que esta variable también está dentro de sucesos posibles.



En el resto de las variables no vemos casos anómalos o que se puedan identificar como outliers.

Ahora que ya hemos analizado y transformado el dataset estamos listos para proceder a su análisis y guardar la base con todos los cambios.

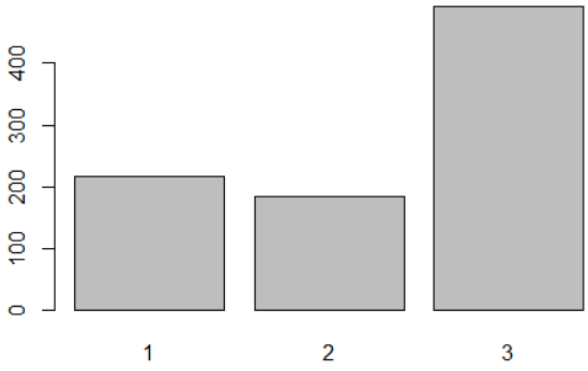
4. Análisis de los datos.

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

Como mencionamos en el inicio, lo que queremos ver es si hay alguna influencia de la situación socioeconómica de los viajeros que pudiera afectar, más allá de la suerte, a la hora de sobrevivir. En el dataset tenemos dos variables con las que podemos analizarlo, que son la Clase y el precio del billete en relación con si sobreviven.

Para ello, vamos a crear una nueva variable dicotómica que sea si son de clase baja o no.

```
titanic$ClaseBaja <- ifelse(titanic$Pclass == 3, TRUE, FALSE)
```



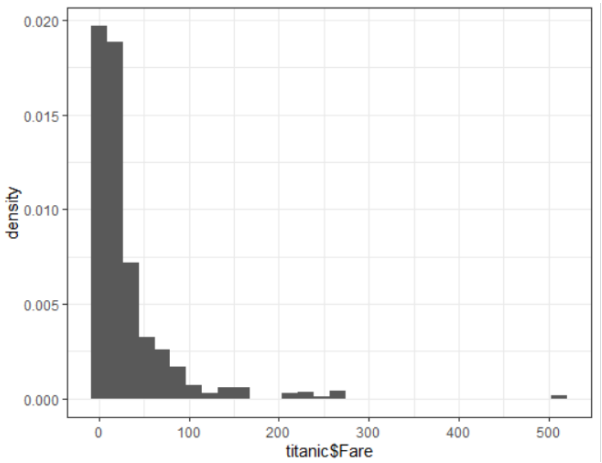
4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Vamos a comprobar la existencia o no de normalidad en nuestros datos. Para ello vamos a usar el test Lilliefors a las variables cuantitativas, que son la edad y el precio del billete. Con este test se considera como hipótesis nula que los datos sí proceden de una distribución normal y como hipótesis alternativa que no lo hacen.

```
> lillie.test(x = titanic$Age)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  titanic$Age
D = 0.1501, p-value < 2.2e-16

> lillie.test(x = titanic$Fare)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  titanic$Fare
D = 0.28185, p-value < 2.2e-16
```

Vemos que en ambos casos el p valor es inferior a 0,05 por lo tanto podemos aceptar con un 95% de probabilidad la hipótesis alternativa, que ninguna de las dos variables tiene una distribución normal.

Si vemos un gráfico del precio podemos ver claramente que no tiene una distribución normal. En el caso de la edad no esta tan claro ya que al haber tantos casos perdidos e imputarlos por el valor de la media aumenta mucho un valor concreto el cual afecta a esa distribución normal.



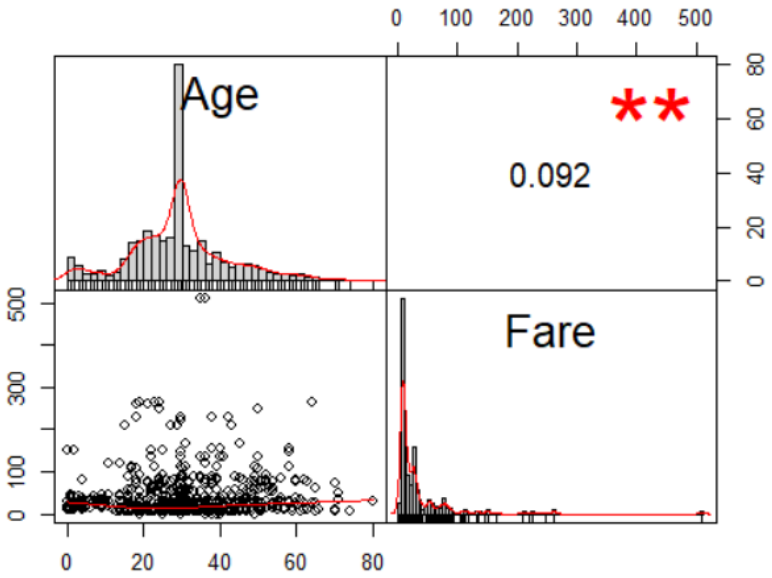
Ahora pasamos a estudiar la homocedasticidad usando el Test de Fligner-Killeen, la estudiaremos en la supervivencia del Titanic que presentan tanto en la edad como en el precio del billete que pagaron.

```
> fligner.test(Age ~  Survived, data = titanic)

        Fligner-Killeen test of homogeneity of variances

data:  Age by Survived
Fligner-Killeen:med chi-squared = 5.5708, df = 1, p-value = 0.01826

> fligner.test(Fare ~ Survived, data = titanic)

        Fligner-Killeen test of homogeneity of variances

data:  Fare by Survived
Fligner-Killeen:med chi-squared = 96.253, df = 1, p-value < 2.2e-16
```

Nuevamente obtenemos en los dos casos un valor inferior a 0,05 por lo que no podemos descartar la hipótesis nula de que sus varianzas no 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. Aplicar al menos tres métodos de análisis diferentes.

Como primera prueba podemos ver si hay algún tipo de relación entre la edad y el precio del billete. Lo cual vemos que no hay ningún tipo de relación entre estas dos variables con un valor muy cercano a cero y además significativo al 95%.



Tras esta primera prueba, podemos hacer una prueba para centrarnos en nuestro objetivo que es ver si la clase social influyo en la muerte del accidente del Titanic. Para ello, vamos a dividir la variable Survived por el precio del billete y así ver si hay diferencia de medias entre los que sobreviven y los que murieron por el precio del billete.

Como hemos visto antes que la variable del precio del billete no sigue una distribución normal vamos a aplicar el test de Wilcoxon para muestras no paramétricas y así ver si hay diferencias significativas entre la media del precio del billete de los que mueren y los que sobreviven. En este caso la hipótesis nula es que los que sobreviven y no sobreviven tienen precios de billete similares, y la alternativa que no.

```
> wilcox.test(x=sobrevive,y=nosobrevive, paired = F)

        Wilcoxon rank sum test with continuity correction

data:  sobrevive and nosobrevive
W = 129952, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Vemos que hemos obtenido un pvalor inferior a 0,05 por lo que podemos aceptar con una confianza del 95% la hipótesis alternativa, que los precios del billete de los que sobrevivieron no eran iguales a los que murieron,

Esto lo podemos ver más claramente con las medias de cada uno de ellos y su diferencia:

```
> mean(sobrevive)
[1] 48.39541
> mean(nosobrevive)
[1] 22.11789
> mean(sobrevive) - mean(nosobrevive)
[1] 26.27752
```

Vemos que la media del precio del billete de los que sobrevivieron era de 48 mientras que de los que murieron era de 26, tenemos una diferencia de 26 entre ambos grupos. Visto esto y con la prueba del test podemos confirmar que sobrevivieron más personas que pagaron más por el billete, es decir, que la clase social si importo a la hora de sobrevivir en el accidente del Titanic.

Ahora como última prueba estadística vamos a usar una regresión logística para ver que variables influyeron más a la hora de sobrevivir, como también podemos intentar hacer una predicción para ver si el modelo que obtenemos es bueno o no.

Para ello, primero separamos la base en train y test, para poder entrenar el modelo y luego aplicarlo a datos no entrenados.

Tras esto creamos el modelo donde tendremos en cuenta las variables ClaseBaja, la edad y el precio del billete. Obtenemos los siguientes resultados del modelo:

```
> modelo<-glm(Survived ~ ClaseBaja +  Age + Fare, data=dat_train, na.action = "na.omit", family = "binomial")
> # Vemos los resultados
> summary(modelo)

Call:
glm(formula = Survived ~ ClaseBaja + Age + Fare, family = "binomial",
    data = dat_train, na.action = "na.omit")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2826  -0.8366  -0.6918   1.0689   2.1939

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.837104   0.285690   2.930  0.00339 **
ClaseBajaTRUE -1.301486   0.195963  -6.641 3.11e-11 ***
Age           -0.030646   0.006909  -4.436 9.18e-06 ***
Fare           0.008645   0.002636   3.279  0.00104 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 948.92  on 711  degrees of freedom
Residual deviance: 841.08  on 708  degrees of freedom
AIC: 849.08

Number of Fisher Scoring iterations: 4
```

Podemos ver como tanto el modelo en total como las tres variables independientes son significativas al 99%. De estos resultados podemos extraer las siguientes conclusiones: los viajeros de Clase 3 o clase baja tienen menos probabilidad de sobrevivir que los de Clase 1 ó 2, siendo la variable con mayor estimación y la que más influye en el modelo. En relación con la edad vemos que cuanto más jóvenes menos probabilidad de sobrevivir. Y sobre el precio, cuanto mayor sea el precio del billete más probable es sobrevivir.

Si aplicamos este modelo a los datos de entrenamiento, vemos que acertaríamos un 70% de los casos que está bastante bien.

```
> matriz_confusion1
             predicciones
observaciones   0   1
            0 368  70
            1 143 131
> (matriz_confusion1[1,1] + matriz_confusion1[2,2])/sum(matriz_confusion1)
[1] 0.7008427
```

Si ahora, por ejemplo, hacemos el modelo solo con la clase baja vemos los siguientes resultados:

```
> modelo2<-glm(Survived ~ ClaseBaja, data=dat_train, na.action = "na.omit", family = "binomial")
> summary(modelo2)

Call:
glm(formula = Survived ~ ClaseBaja, family = "binomial", data = dat_train,
    na.action = "na.omit")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2767  -0.7563  -0.7563   1.0813   1.6681

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.2303     0.1136   2.027   0.0427 *
ClaseBajaTRUE  -1.3356     0.1623  -8.227   <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: 948.92  on 711  degrees of freedom
Residual deviance: 877.68  on 710  degrees of freedom
AIC: 881.68

Number of Fisher Scoring iterations: 4
```

```
> matriz_confusion2
             predicciones
observaciones   0   1
            0 299 139
            1  99 175
> (matriz_confusion2[1,1] + matriz_confusion2[2,2])/sum(matriz_confusion2)
[1] 0.6657303
```

Vemos que el modelo sigue siendo significativo y el acierto es de 66,5%, usando solo esta variable solo ha bajado el acierto un 35% lo que quiere decir que esta variable tiene una gran capacidad explicativa sobre los que sobreviven.

Por último, vamos a aplicar el modelo completo ya que tiene un mejor acierto al dataset test, dado que estos no se han usado en el modelo podemos ver el verdadero acierto del modelo.

```
> dat_test$Prediccionnumerica <- predict(modelo1, dat_test)
> dat_test$PrediccionDico <- ifelse(dat_test$Prediccionnumerica > 0.5, 1, 0)
> matriz_confusion <- table(dat_test$Survived, dat_test$PrediccionDico,
+                           dnn = c("observaciones", "predicciones"))
> matriz_confusion
            predicciones
observaciones   0   1
            0 104   7
            1  51  17
> (matriz_confusion[1,1] + matriz_confusion[2,2])/sum(matriz_confusion)
[1] 0.6759777
```

Vemos que el acierto en el test es del 67.5%, ha bajado un poco, pero sigue teniendo un buen acierto.

5. Conclusiones

El dataset Titanic tiene los datos información de pasajeros que viajaron en el Titanic, con sus características personales como las relacionadas con el viaje. A lo largo de la practica hemos visto como había algunos datos que faltaban y los hemos tenido que imputar. Además, hemos comprobado como la clase social y más concretamente la clase en el viaje influyó y pudo ser determinante a la hora de sobrevivir en el accidente del Titanic. Se ha demostrado mediante contraste de medias como el precio que pagaron los pasajeros que sobrevivieron fue muy superior al que pagaron de media los que murieron. Con una regresión logística hemos visto también como pertenecer a la Clase 3 en el viaje disminuye las probabilidades de sobrevivir al viaje.

En conclusión, vemos como la clase social, determinada por el precio del billete y la Clase en la que viajan los pasajeros determinaron las probabilidades de sobrevivir al accidente del Titanic. Como los pasajeros de Clase 3 y con billetes baratos tuvieron muchas menos probabilidades de sobrevivir que los de Clase 1ó 2 y con billetes más caros. Al final, sobrevivir al Titanic no fue solo cuestión de suerte sino también de Clase y los beneficios que esta te da.

Autoría

El trabajo se ha realizado individualmente por Daniel Núñez Sanz.

Contribuciones	Firma
Investigación previa	DNS
Redacción de las respuestas	DNS
Desarrollo código	DNS