Regresión Logística

Efrén Jiménez

1 de setiembre de 2016

Análisis del Problema

Cuando sucede un accidente en un medio de transporte colectivo de gran tamaño como un barco o un avión,una de las situaciones más sensibles es cómo saber las posibilidades de que un pasajero sobreviva. En el caso del Titanic, ya no hay nada que se pueda hacer sobre los pasajeros, no hay decisión que se pueda tomar. Sin embargo, la forma en que se van a manejar los datos para determinar si un pasajero sobrevivió o no se puede transportar a otros escenarios actuales.

Al momento en que sucede un accidente, conforme se empiecen a encontrar los primeros sobrevivientes y/o los primeros cuerpos, podríamos comenzar a crear un modelo que permita predecir las probabilidades de otros pasajeros de haber sobrevivido. Esto podría ayudar en el momento a los cuerpos de rescate a saber qué es necesario tener en la escena del accidente, para poder reaccionar inmediatamente ante cualquier situación.

Entendimiento de los Datos

El conjunto de datos que se va a analizar cuenta con 891 observaciones y 12 variables: - PassengerID: Número de pasajero; numérico, rango de 1 a 891. - Survived: Indica si el pasajero sobrevivió o no; valores posibles: 0 (no), 1 (sí). - Pclass: Clase en la cual viajaba el pasajero; valores posibles: 1, 2 ó 3. - Name: Nombre del pasajero; variable cualitativa categórica. - Sex: Género del pasajero; valores posibles: male (hombre) y female (mujer). - Age: Edad del pasajero; rango: 0.42 a 80 años, con 177 valores faltantes. - SibSp: Cantidad de hermanos o cónyuges en el barco; numérica, rango de 0 a 8. - Parch: Cantidad de hij@s o padres a bordo; numérica, rango de 0 a 6. - Ticket: Número de tiquete; variable cualitativa categórica. - Fare: Monto pagado por el pasajero por su tiquete; rango: de 0 a 512.33. - Cabin: Cabina en la cual esaba hospedado el pasajero; variable cualitiva categórica. - Embarked: Puerto en el cual embarcó el pasajero; variable cualitiva categórica.

Exploración de los Datos

```
# librerías utilizadas
library(titanic)
library(lattice)
library(caTools)
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

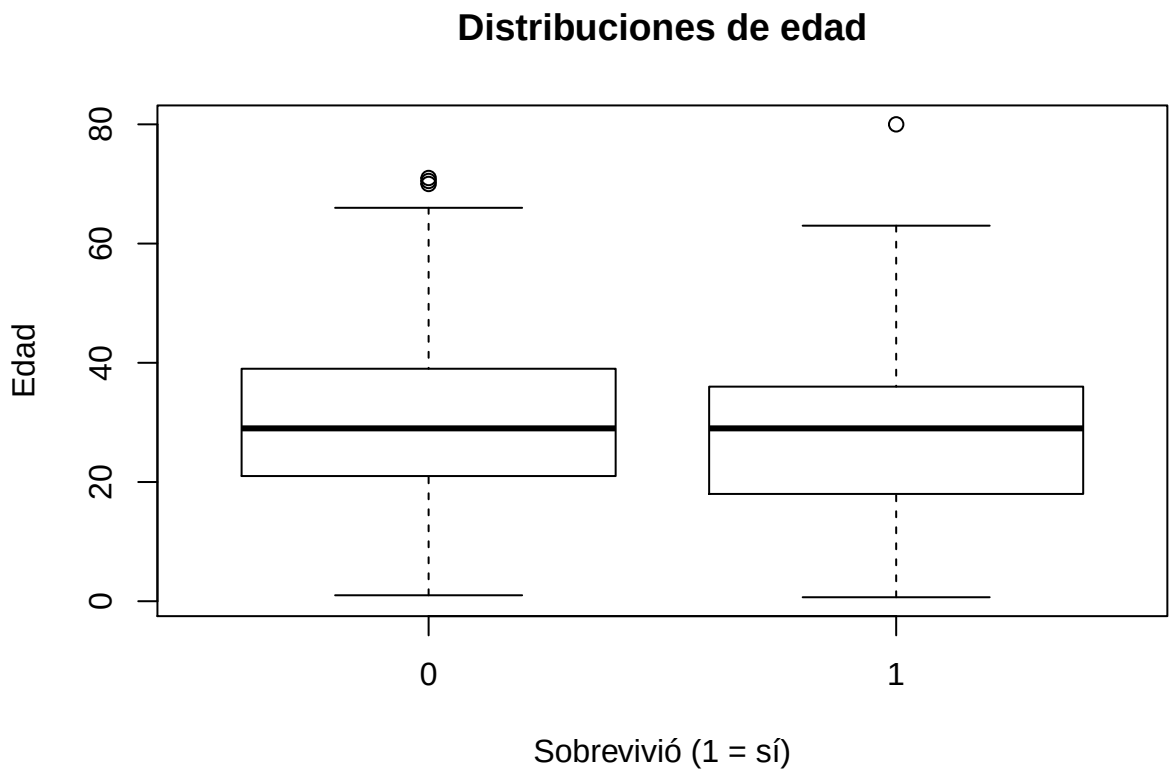
```r
data(titanic_train)
data(titanic_test)
# crear factores
titanic_train$Survived <- factor(titanic_train$Survived)
titanic_train$Pclass <- factor(titanic_train$Pclass)
titanic_train$Sex <- factor(titanic_train$Sex)
titanic_train$Cabin <- factor(titanic_train$Cabin)
titanic_train$Embarked <- factor(titanic_train$Embarked)
str(titanic_train)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ 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     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ 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 ...
```

```r
# Dividir el conjunto de datos en entrenamiento y prueba
set.seed(351)
splt <- sample.split(titanic_train$Survived, SplitRatio = 0.7)
datos.entrenamiento <- titanic_train[splt, ]
datos.prueba <- titanic_train[!splt, ]
```

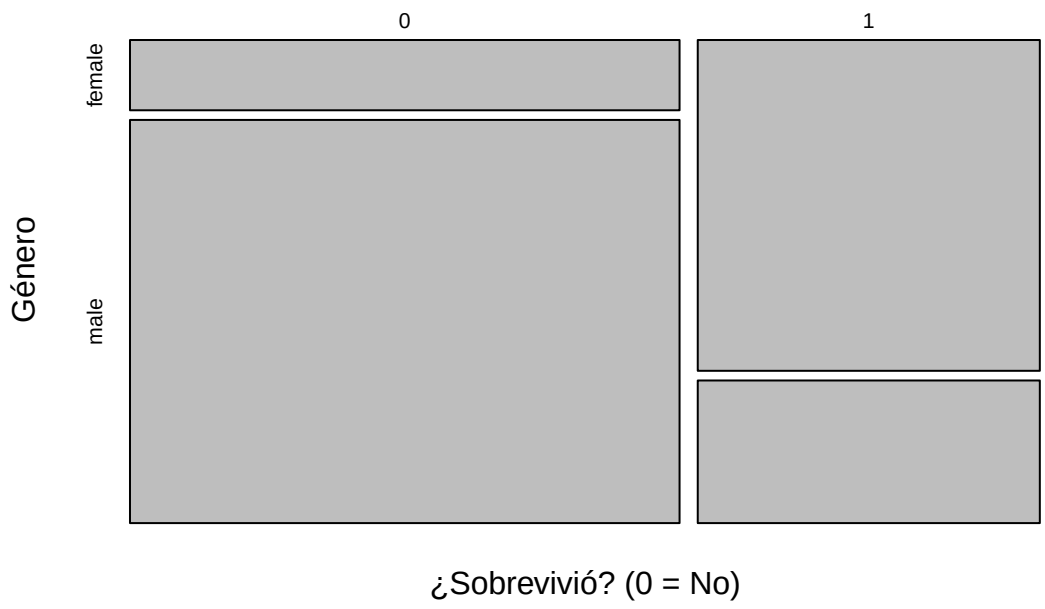
Una vez cargados los datos, podemos comenzar a explorarlos. Para comenzar, podemos analizar la distribución de la variable Edad, en el contexto de si el pasajero sobrevivió o no:

```r
boxplot(datos.entrenamiento$Age ~ datos.entrenamiento$Survived, main = "Distribuciones de edad",
    ylab = "Edad", xlab = "Sobrevivió (1 = sí)")
```



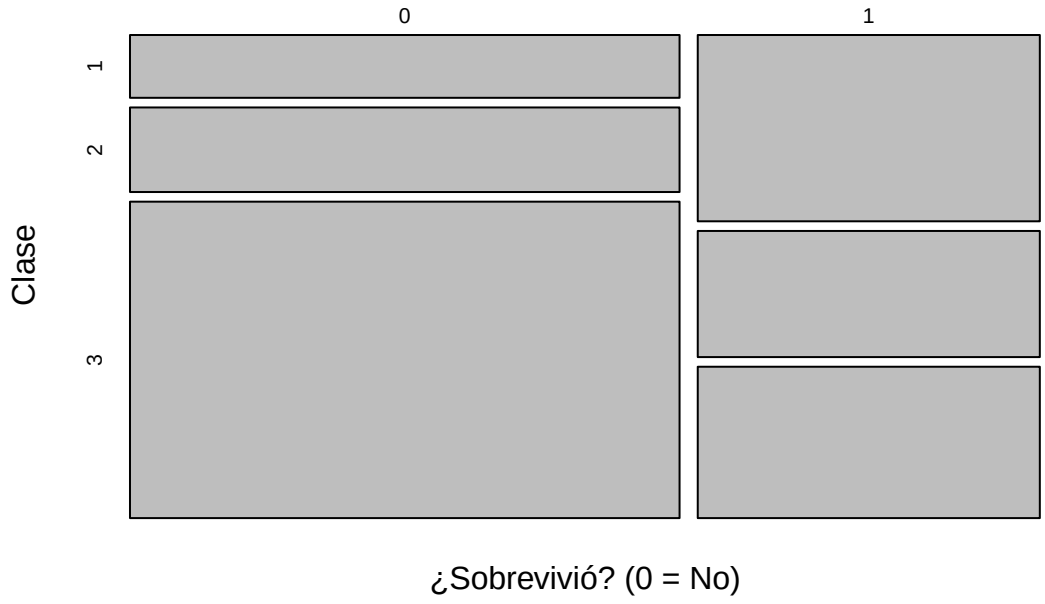
Del gráfico anterior, podemos concluir que hubo sobrevivientes y víctimas en diferentes rangos de edades, y que no se aprecia un patrón dictado por la edad que favorezca al a sobrevivir el accidente. Alternativamente, podemos comparar proporciones entre los sobrevivientes y el genero:

```
mosaicplot(~datos.entrenamiento$Survived + datos.entrenamiento$Sex, main = "Proporción de sobrevivientes
    ylab = "Género", xlab = "¿Sobrevivió? (0 = No)")
```



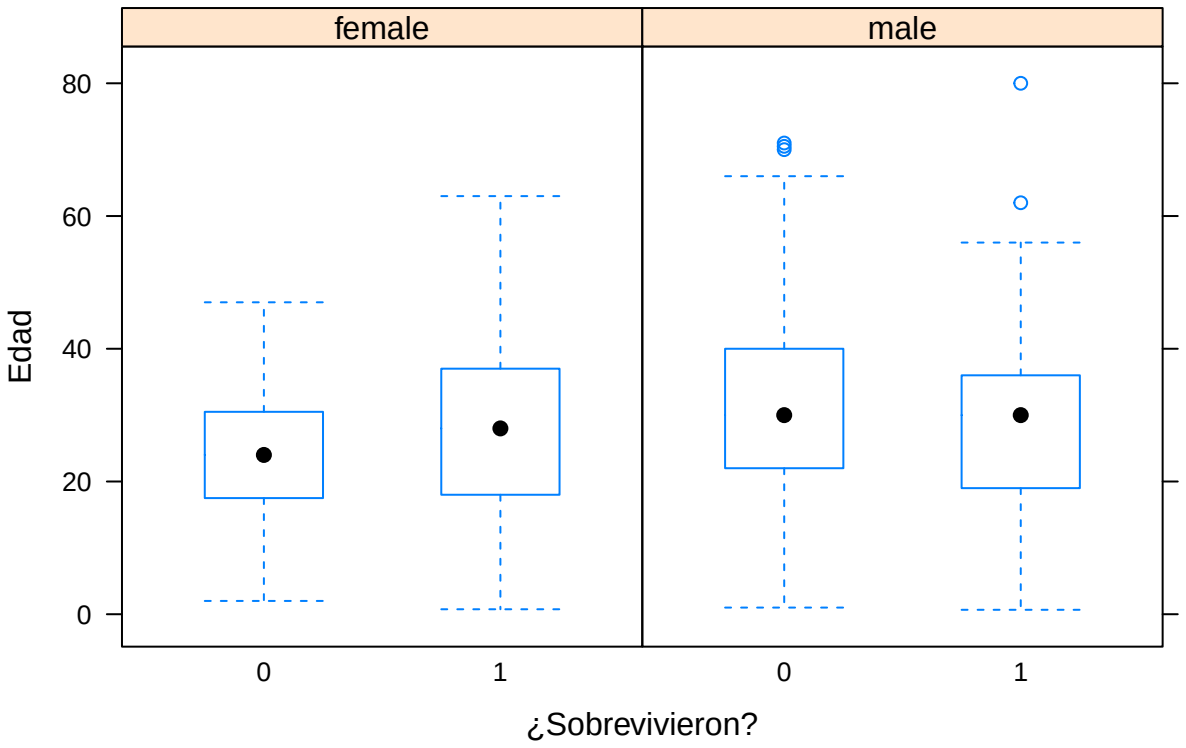
En el gráfico de mosaico arriba, podemos apreciar cómo hay un mayor número de víctimas masculinas, complementado por una gran cantidad de mujeres sobrevivientes. Dada la época en la cual sucedió el accidente del Titanic, también es importante analizar la proporción de sobrevivientes por clase en la cual viajaba:

```
mosaicplot(~datos.entrenamiento$Survived + datos.entrenamiento$Pclass, main = "Proporción de sobrevivie
    ylab = "Clase", xlab = "¿Sobrevivió? (0 = No)")
```



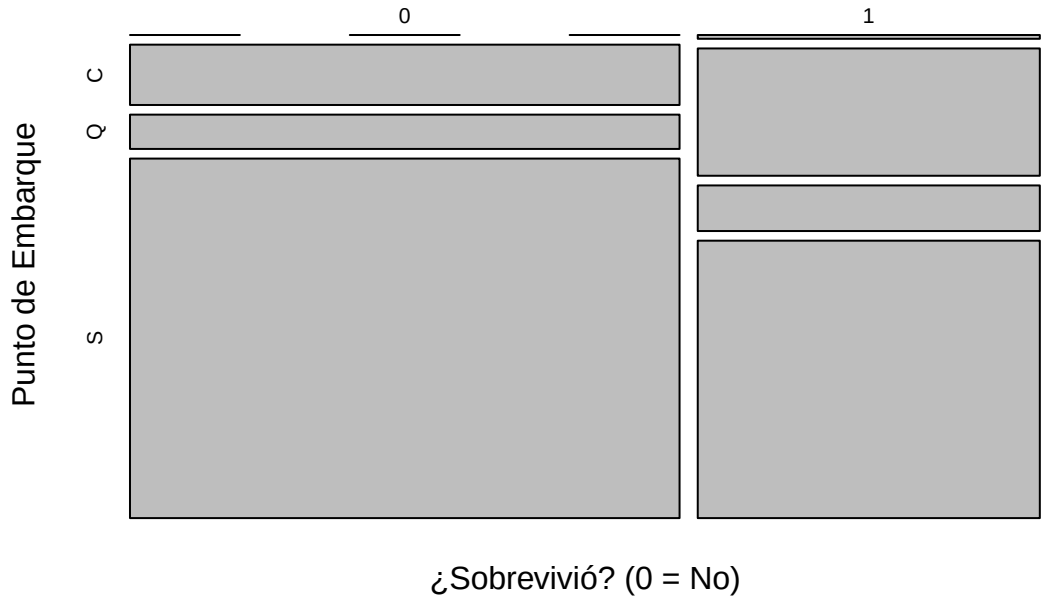
En el gráfico anterior, se puede apreciar cómo la mayoría de personas que no sobrevivió viajaba en tercera clase , mientras que la proporción de sobrevivientes la domina la gente que iba en primera clase. Adicionalmente, se puede analizar la interacción de variables como el género y la edad, para ver si a pesar de que la edad no parece ser importante por sí sola para determinar quién sobrevivió, pero talvez en combinación con el género sí pueda ser interesante.

```
bwplot(datos.entrenamiento$Age ~ datos.entrenamiento$Survived | datos.entrenamiento$Sex,
    main = "Distribución de edades por género y si sobrevivieron (1) o no (0)",
    xlab = "¿Sobrevivieron?", ylab = "Edad")
```



Del gráfico anterior, se puede observar una mayor tendencia de hombres de mayor edad y mujeres de menos edad a no sobrevivir, por lo cual no vamos a descartar la variable de la edad a la hora de hacer el gráfico. Para terminar, podemos ver la proporción de sobrevivientes de acuerdo con su punto de embarque:

```
mosaicplot(~datos.entrenamiento$Survived + datos.entrenamiento$Embarked, main = "Proporción de sobreviv
    ylab = "Punto de Embarque", xlab = "¿Sobrevivió? (0 = No)")
```



Modelo de Minería de Datos

Para modelar este caso, se va a utilizar una regresión logística, dejando de lado columnas como el identificador del pasajero, el nombre, el número de tiquete y el número de cabina en la cual estuvo hospedado el pasajero:

```
titanic.fit <- glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,
    data = datos.entrenamiento, family = binomial)
```

Al ver los detalles del modelo:

```
summary(titanic.fit)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +
##     Fare + Embarked, family = binomial, data = datos.entrenamiento)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7717  -0.6390  -0.3591   0.5515   2.5430
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  16.690641 608.038382   0.027  0.97810
```

```
## Pclass2      -1.148664   0.402096  -2.857  0.00428 **
## Pclass3      -2.586663   0.430218  -6.012 1.83e-09 ***
## Sexmale      -2.585890   0.273213  -9.465  < 2e-16 ***
## Age          -0.043615   0.010350  -4.214 2.51e-05 ***
## SibSp        -0.340159   0.162014  -2.100  0.03577 *
## Parch         0.031517   0.155045   0.203  0.83892
## Fare          0.001685   0.003663   0.460  0.64538
## EmbarkedC   -12.220417 608.038168  -0.020  0.98397
## EmbarkedQ   -12.647053 608.038464  -0.021  0.98341
## EmbarkedS   -12.762141 608.038140  -0.021  0.98325
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 669.68  on 495  degrees of freedom
## Residual deviance: 420.94  on 485  degrees of freedom
##   (127 observations deleted due to missingness)
## AIC: 442.94
##
## Number of Fisher Scoring iterations: 13
```

Se puede observar que hay muchas variables que no son significativas: el punto de embarque, el monto pagado por el tiquete y la cantidad de padres / hijos a bordo, así que se procede a hacer un segundo modelo sin estas variables:

```
titanic.fit <- glm(Survived ~ Pclass + Sex + Age + SibSp, data = datos.entrenamiento,
    family = binomial)
summary(titanic.fit)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp, family = binomial,
##     data = datos.entrenamiento)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8208  -0.6173  -0.3595   0.5825   2.5306
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.38606    0.54816   8.001 1.23e-15 ***
## Pclass2     -1.43665    0.34943  -4.111 3.93e-05 ***
## Pclass3     -2.84908    0.36032  -7.907 2.64e-15 ***
## Sexmale     -2.65573    0.26368 -10.072  < 2e-16 ***
## Age         -0.04537    0.01024  -4.432 9.32e-06 ***
## SibSp       -0.33581    0.15393  -2.182   0.0291 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 669.68  on 495  degrees of freedom
```

```
## Residual deviance: 424.63  on 490  degrees of freedom
##   (127 observations deleted due to missingness)
## AIC: 436.63
##
## Number of Fisher Scoring iterations: 5
```

En este segundo modelo, todas las variables son significativas, y el AIC bajó de 443 a 437. Tenemos suficiente evidencia de que el segundo modelo es mejor que el primero a nivel estadístico. Con respecto a la interpretación de coeficientes, se puede decir que: - El logaritmo de las posibilidades de los pasajeros de 2da y 3ra clase es menor que el de los pasajeros de primera clase. - La probabilidad de sobrevivir es menor para los hombres. - En general, a mayor edad y cantidad de hermanos / espos@s, menor probabilidad de sobrevivir

Evaluación

A manera de modelo ingenuo, podemos tener un modelo que prediga que nadie sobrevivió al Titanic, pues es el resultado más frecuente. Dicho modelo tendría una exactitud del 62.57% (165 aciertos de 268 en el conjunto de pruebas).

```
table(datos.entrenamiento$Survived)
```

```
##
##   0   1
## 384 239
```

```
table(datos.prueba$Survived, rep(0, nrow(datos.prueba)))
```

```
##
##       0
##   0 165
##   1 103
```

Al generar las predicciones del modelo sobre el conjunto de pruebas, tenemos las siguientes métricas según la tabla abajo (usando 0.5 como umbral de discriminación):

- Exactitud: 63.81%
- Sensibilidad: 70.79%
- Especificidad: 83.72%
- Área bajo la curva: 82.96%

```
predicciones <- predict(titanic.fit, newdata = datos.prueba, type = "response")
table(datos.prueba$Survived, predicciones >= 0.5)
```

```
##
##     FALSE TRUE
##   0   108   21
##   1    26   63
```

```
# Exactitud:
(108 + 63)/nrow(datos.prueba)
```

```
## [1] 0.6380597
```

```r
# Sensibilidad:
63/(63 + 26)
```

```
## [1] 0.7078652
```

```r
# Especificidad:
108/(108 + 21)
```

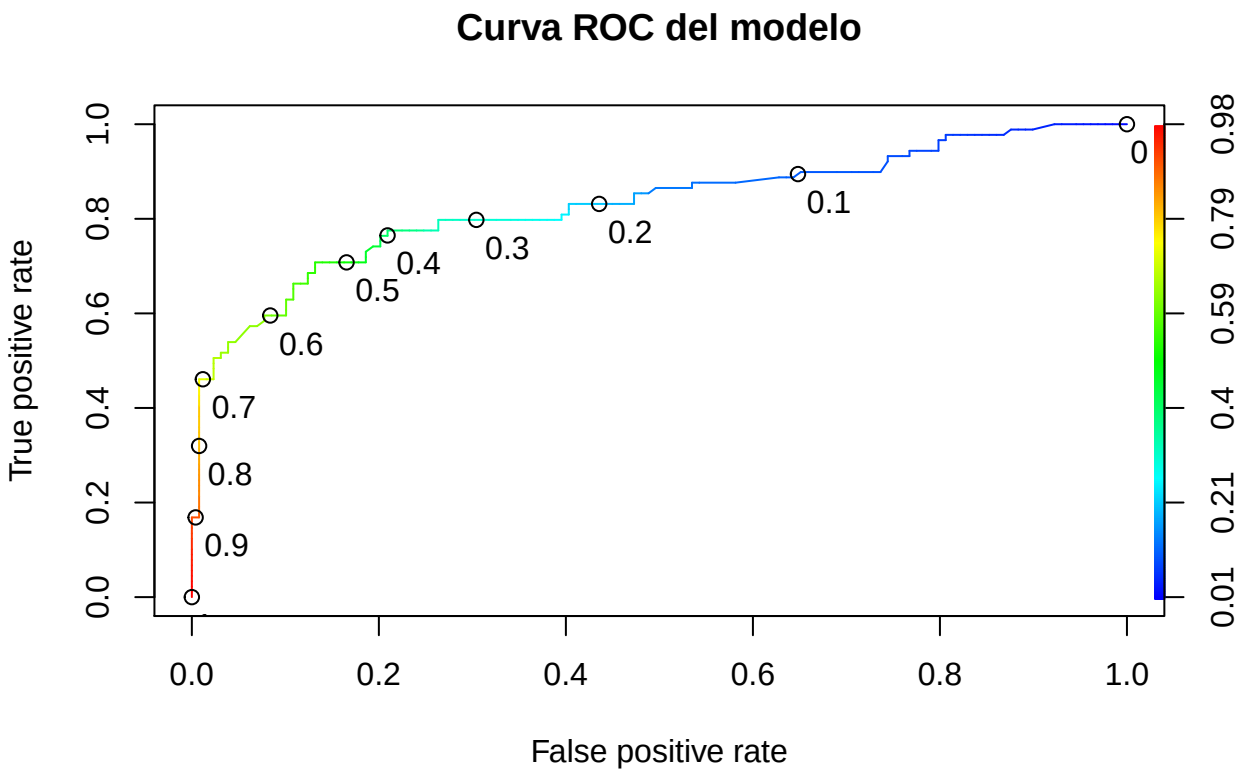
```
## [1] 0.8372093
```

```r
# Área bajo la curva:
prediccionesROC <- prediction(predicciones, datos.prueba$Survived)
as.numeric(performance(prediccionesROC, "auc")@y.values)
```

```
## [1] 0.8296316
```

Según la curva ROC, al intentar aumentar el porcentaje de verdaderos positivos (sensibilidad) o de verdaderos negativos (especificdad) del modelo, estaría aumentando significativamente el porcentaje de falsos positivos y falsos negativos, respectivamente:

```r
plot(performance(prediccionesROC, "tpr", "fpr"), colorize = T, print.cutoffs.at = seq(0,
    1, by = 0.1), text.adj = c(-0.2, 1.7), main = "Curva ROC del modelo")
```



Resultados

En términos generales, se puede decir que el modelo es apenas un poco mejor que el modelo ingenuo a nivel de exactitud. Si bien es cierto que tenemos un 83% de clasificar apropiadamente a un pasajero en si sobrevivió o no, esto se puede deber más al hecho de que la mayoría de pasajeros no sobrevivió que a que el modelo es realmente bueno. Dependiendo del uso que se le quiera dar al modelo, ya en un caso de desastre actual, se puede cambiar el umbral de discriminación a algo mayor o menor que 0.5. Por ejemplo, si se quisiera saber con mayor exactitud quienes sobrevivieron se podría bajar el umbral de discriminación a 0.1:

```
table(datos.prueba$Survived, predicciones >= 0.1)
```

```
##
##     FALSE TRUE
##   0    46   83
##   1    10   79
```

Con este cambio, la sensibilidad del modelo sube a 88.76% (se identificaron correctamente al 89% de pasajeros que sobrevivieron). Sin embargo, la especificidad bajó a 36%, por lo que se estaría reportando como sobrevivientes a muchos pasajeros que fallecieron, lo cual puede ser un golpe fuerte para los familiares de estas personas y para la credibilidad de la empresa. Alternativamente, se podría subir el umbral de discriminación para poder tener más certeza en la cantidad de fallecidos predecida. Por ejemplo, se puede subir a 0.7:

```
table(datos.prueba$Survived, predicciones >= 0.7)
```

```
##
##     FALSE TRUE
##   0   128    1
##   1    48   41
```

Con este cambio, se pasaría a una especificidad del 99%, a costo de bajar la sensibilidad al 46%. En este caso, muchos sobrevivientes se estarían dando por fallecidos y se podría tomar la decisión de no buscar más sobrevivientes. Si bien es cierto que ambos cambios al umbral tienen su riesgo, es probable que el "dar a una persona por muerta" y no bucarla si hay posibilidades de que sobreviva tenga una consecuencia peor para la empresa o los rescatistas de nuestra situación hipotética.