M5_AI2

Sofia Cantu

2024-11-19

Actividad Integradora 2

```
# Librerías
if (!require(tidyverse)) install.packages("tidyverse")
```

```
## Loading required package: tidyverse
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force
all conflicts to become errors
```

```
library(tidyverse)
if (!require(caret)) install.packages("caret")
```

```
## Loading required package: caret
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(caret)
if (!require(pROC)) install.packages("pROC")
```

```
## Loading required package: pROC
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
```

```
##
##     cov, smooth, var

library(pROC)

datos_titanic <-
read.table("~/Downloads/IA_Materia/ArchivosCodigos/Titanic.csv", header =
TRUE, sep = ",")
head(datos_titanic, 10)

##    PassengerId Survived Pclass
Name
## 1          892        0      3                               Kelly, Mr.
James
## 2          893        1      3             Wilkes, Mrs. James (Ellen
Needs)
## 3          894        0      2                     Myles, Mr. Thomas
Francis
## 4          895        0      3                               Wirz, Mr.
Albert
## 5          896        1      3 Hirvonen, Mrs. Alexander (Helga E
Lindqvist)
## 6          897        0      3                   Svensson, Mr. Johan
Cervin
## 7          898        1      3                         Connolly, Miss.
Kate
## 8          899        0      2                 Caldwell, Mr. Albert
Francis
## 9          900        1      3    Abrahim, Mrs. Joseph (Sophie Halaut
Easu)
## 10         901        0      3                       Davies, Mr. John
Samuel
##       Sex  Age SibSp Parch    Ticket    Fare Cabin Embarked
## 1    male 34.5     0     0    330911  7.8292              Q
## 2  female 47.0     1     0    363272  7.0000              S
## 3    male 62.0     0     0    240276  9.6875              Q
## 4    male 27.0     0     0    315154  8.6625              S
## 5  female 22.0     1     1   3101298 12.2875              S
## 6    male 14.0     0     0      7538  9.2250              S
## 7  female 30.0     0     0    330972  7.6292              Q
## 8    male 26.0     1     1    248738 29.0000              S
## 9  female 18.0     0     0      2657  7.2292              C
## 10   male 21.0     2     0 A/4 48871 24.1500              S

datos_titanic_test <-
read.table("~/Downloads/IA_Materia/ArchivosCodigos/Titanic_test.csv", header
= TRUE, sep = ",")
head(datos_titanic_test, 10)

##    PassengerId Pclass                                          Name    Sex
Age
```

```
## 1          892      3                              Kelly, Mr. James   male
34.5
## 2          893      3              Wilkes, Mrs. James (Ellen Needs) female
47.0
## 3          894      2                     Myles, Mr. Thomas Francis   male
62.0
## 4          895      3                              Wirz, Mr. Albert   male
27.0
## 5          896      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female
22.0
## 6          897      3                    Svensson, Mr. Johan Cervin   male
14.0
## 7          898      3                          Connolly, Miss. Kate female
30.0
## 8          899      2                 Caldwell, Mr. Albert Francis   male
26.0
## 9          900      3    Abrahim, Mrs. Joseph (Sophie Halaut Easu) female
18.0
## 10         901      3                       Davies, Mr. John Samuel   male
21.0
##    SibSp Parch    Ticket    Fare Cabin Embarked
## 1      0     0    330911  7.8292              Q
## 2      1     0    363272  7.0000              S
## 3      0     0    240276  9.6875              Q
## 4      0     0    315154  8.6625              S
## 5      1     1   3101298 12.2875              S
## 6      0     0      7538  9.2250              S
## 7      0     0    330972  7.6292              Q
## 8      1     1    248738 29.0000              S
## 9      0     0      2657  7.2292              C
## 10     2     0 A/4 48871 24.1500              S
```

1. Prepara la base de datos Titanic:

- Analiza los datos faltantes
- Realiza un análisis descriptivo
- Haz una partición de los datos (70-30) para el entrenamiento y la validación. Revisa la proporción de sobrevivientes para la partición y la base original.

```r
# Seleccionar las variables de interés
datos_titanic <- datos_titanic %>%
  select(PassengerId, Name, Survived, Ticket, Cabin, Pclass, Sex, Age, SibSp,
Parch, Fare, Embarked)

# Analizar datos faltantes
missing_summary <- datos_titanic %>% summarise_all(~ sum(is.na(.)))
print("Datos faltantes por columna:")
```

```
## [1] "Datos faltantes por columna:"
```

```r
print(missing_summary)
```

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

```r
# Reemplazar valores faltantes en 'Age' con la mediana
datos_titanic$Age[is.na(datos_titanic$Age)] <- median(datos_titanic$Age,
na.rm = TRUE)

# Eliminar filas con valores faltantes en columnas críticas
datos_titanic <- datos_titanic %>% drop_na(Survived, Pclass, Sex, Fare,
Embarked)

# Convertir factores
datos_titanic <- datos_titanic %>%
  mutate(Survived = as.factor(Survived),
         Pclass = as.factor(Pclass),
         Sex = as.factor(Sex),
         Embarked = as.factor(Embarked))

# Análisis descriptivo
summary(datos_titanic)
```

```
##   PassengerId         Name           Survived    Ticket
##  Min.   :   1.0   Length:1306        0:814    Length:1306
##  1st Qu.: 328.2   Class :character   1:492    Class :character
##  Median : 654.5   Mode  :character            Mode  :character
##  Mean   : 655.0
##  3rd Qu.: 981.8
##  Max.   :1309.0
##     Cabin           Pclass      Sex           Age            SibSp
##  Length:1306        1:321   female:464   Min.   : 0.17   Min.   :0.0
##  Class :character   2:277   male  :842   1st Qu.:22.00   1st Qu.:0.0
##  Mode  :character   3:708                Median :28.00   Median :0.0
##                                          Mean   :29.45   Mean   :0.5
##                                          3rd Qu.:35.00   3rd Qu.:1.0
##                                          Max.   :80.00   Max.   :8.0
##      Parch             Fare          Embarked
##  Min.   :0.0000   Min.   :  0.000   C:270
##  1st Qu.:0.0000   1st Qu.:  7.896   Q:123
##  Median :0.0000   Median : 14.454   S:913
##  Mean   :0.3859   Mean   : 33.224
##  3rd Qu.:0.0000   3rd Qu.: 31.275
##  Max.   :9.0000   Max.   :512.329
```

```r
# Partición de datos (70% entrenamiento, 30% validación)
set.seed(123)
train_index <- createDataPartition(datos_titanic$Survived, p = 0.7, list =
```

```
FALSE)
train_data <- datos_titanic[train_index, ]
test_data <- datos_titanic[-train_index, ]
```

2. Con la base de datos de entrenamiento, encuentra un modelo logístico para encontrar el mejor conjunto de predictores que auxilien a clasificar la dirección de cada observación.

- Auxiliate del criterio de AIC para determinar cuál es el mejor modelo.
- Propón por lo menos los dos que consideres mejores modelos.

```
# Revisar proporciones de sobrevivientes
prop_original <- prop.table(table(datos_titanic$Survived))
prop_train <- prop.table(table(train_data$Survived))
prop_test <- prop.table(table(test_data$Survived))

print("Proporción de sobrevivientes (Original):")

## [1] "Proporción de sobrevivientes (Original):"

print(prop_original)

##
##         0         1
## 0.6232772 0.3767228

print("Proporción de sobrevivientes (Entrenamiento):")

## [1] "Proporción de sobrevivientes (Entrenamiento):"

print(prop_train)

##
##         0         1
## 0.6229508 0.3770492

print("Proporción de sobrevivientes (Prueba):")

## [1] "Proporción de sobrevivientes (Prueba):"

print(prop_test)

##
##         0         1
## 0.6240409 0.3759591

# Ajustar modelos logísticos
# Modelo completo
modelo1 <- glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare +
Embarked,
                       data = train_data,
                       family = binomial)
```

```r
# Modelo simplificado (usando AIC)
modelo2 <- step(modelo1, direction = "backward")
```

```
## Start:  AIC=698.57
## Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked
##
##             Df Deviance     AIC
## - Parch      1   679.20  697.20
## - Fare       1   679.76  697.76
## - Embarked   2   682.25  698.25
## <none>           678.57  698.57
## - SibSp      1   688.19  706.19
## - Age        1   691.55  709.55
## - Pclass     2   725.98  741.98
## - Sex        1  1072.74 1090.74
##
## Step:  AIC=697.2
## Survived ~ Pclass + Sex + Age + SibSp + Fare + Embarked
##
##             Df Deviance     AIC
## - Fare       1   680.09  696.09
## <none>           679.20  697.20
## - Embarked   2   683.53  697.53
## - SibSp      1   690.98  706.98
## - Age        1   692.17  708.17
## - Pclass     2   729.10  743.10
## - Sex        1  1083.77 1099.77
##
## Step:  AIC=696.09
## Survived ~ Pclass + Sex + Age + SibSp + Embarked
##
##             Df Deviance     AIC
## <none>           680.09  696.09
## - Embarked   2   684.84  696.84
## - SibSp      1   691.05  705.05
## - Age        1   693.36  707.36
## - Pclass     2   752.73  764.73
## - Sex        1  1092.74 1106.74
```

```r
# Resumen del modelo seleccionado
summary(modelo2)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Embarked,
##     family = binomial, data = train_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  4.277335   0.452443   9.454  < 2e-16 ***
## Pclass2     -1.043373   0.293115  -3.560 0.000371 ***
## Pclass3     -2.264797   0.286382  -7.908 2.61e-15 ***
## Sexmale     -3.645067   0.221106 -16.486  < 2e-16 ***
## Age         -0.030437   0.008526  -3.570 0.000357 ***
## SibSp       -0.357154   0.116883  -3.056 0.002246 **
## EmbarkedQ    0.347202   0.402394   0.863 0.388225
## EmbarkedS   -0.328853   0.250736  -1.312 0.189671
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1212.56  on 914  degrees of freedom
## Residual deviance:  680.09  on 907  degrees of freedom
## AIC: 696.09
##
## Number of Fisher Scoring iterations: 5
```

Análisis del Modelo:

a) Variables seleccionadas en el modelo final:

- class
- Sex
- Age
- SibSp
- Embarked Estas variables son estadísticamente significativas y tienen los valores de AIC más bajos después de cada iteración en el proceso de selección.

b) Modelo final con menor AIC:

- El AIC final es 696.09.
- El modelo incluye las variables que mostraron ser predictoras clave de la supervivencia.

c) Coeficientes:

- Los coeficientes muestran que:
- Ser de la clase Pclass3 disminuye significativamente la probabilidad de supervivencia.
- Ser hombre tiene un impacto negativo considerable sobre la probabilidad de supervivencia.
- La edad afecta negativamente la probabilidad de supervivencia, aunque el impacto es menor en magnitud.
- SibSp y Embarked tienen un impacto más leve, pero significativo.

Propuestas de los dos mejores modelos:

a) Modelo 1 (Modelo intermedio):

- Fórmula: Survived ~ Pclass + Sex + Age + SibSp + Fare
- Este modelo podría ser considerado si el enfoque es evaluar Fare en lugar de Embarked, aunque tiene un AIC ligeramente mayor (~696.76).

b) Modelo 2 (Mejor modelo basado en AIC):

- Fórmula: Survived ~ Pclass + Sex + Age + SibSp + Embarked
- AIC: 696.09 Este modelo tiene la mejor combinación de simplicidad y ajuste, según el criterio AIC.

3. Analiza los modelos a través de:

- Identificación de la Desviación residual de cada modelo
- Identificación de la Desviación nula
- Cálculo de la Desviación Explicada
- Prueba de la razón de verosimilitud
- Define cuál es el mejor modelo
- Escribe su ecuación, analiza sus coeficientes y detecta el efecto de cada predictor en la clasificación.

```
# Comparación de desviaciones Modelo 1
desviacion_nula <- modelo1$null.deviance
desviacion_residual <- modelo1$deviance
desviacion_explicada <- (desviacion_nula - desviacion_residual) /
desviacion_nula

print(paste("Desviación explicada Modelo 1:", round(desviacion_explicada *
100, 2), "%"))

## [1] "Desviación explicada Modelo 1: 44.04 %"

# Comparación de desviaciones Modelo 2
desviacion_nula <- modelo2$null.deviance
desviacion_residual <- modelo2$deviance
desviacion_explicada <- (desviacion_nula - desviacion_residual) /
desviacion_nula

print(paste("Desviación explicada Modelo 2:", round(desviacion_explicada *
100, 2), "%"))

## [1] "Desviación explicada Modelo 2: 43.91 %"

# Ajustar el modelo 1 (más simple)
modelo1 <- glm(Survived ~ Pclass + Sex + Age, family = binomial, data =
train_data)

# Ajustar el modelo 2 (más complejo, incluye SibSp y Embarked)
modelo2 <- glm(Survived ~ Pclass + Sex + Age + SibSp + Embarked, family =
binomial, data = train_data)
```

```r
# Prueba de la razón de verosimilitud
anova(modelo1, modelo2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Survived ~ Pclass + Sex + Age
## Model 2: Survived ~ Pclass + Sex + Age + SibSp + Embarked
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       910     698.09
## 2       907     680.09  3   18.005 0.0004389 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Definición del Mejor Modelo

La prueba de la razón de verosimilitud muestra que el modelo más complejo (modelo2) mejora significativamente el ajuste en comparación con el modelo más simple (modelo1). Esto se indica por un valor p extremadamente bajo ($p<0.0001$). El modelo 2 es el mejor modelo, ya que incluye más predictores (SibSp y Embarked) que contribuyen a explicar la variación en la variable de respuesta (Survived) y mantiene un bajo AIC y una mejor desviación explicada.

Ecuación del Modelo Seleccionado (Modelo 2)

La ecuación del modelo logístico es: - $\text{logit}(P(\text{Survived}))=4.277335+(-1.043373\cdot\text{Pclass2})+(-2.264797\cdot\text{Pclass3})+(-3.645067\cdot\text{Sexmale})+(-0.030437\cdot\text{Age})+(0.145915\cdot\text{SibSp})+(-0.328853\cdot\text{EmbarkedS})+(0.347202\cdot\text{EmbarkedQ})$ Donde: - Intercepto (4.277): Representa la probabilidad log-odds base cuando todos los predictores tienen el valor de referencia (Pclass1, mujer, edad=0, SibSp=0, Embarked en el puerto C). - Pclass2 y Pclass3: Reducen significativamente la probabilidad de supervivencia, con Pclass3 teniendo el mayor efecto negativo. - Sexmale: Si el pasajero es hombre, la probabilidad de supervivencia disminuye drásticamente. - Age: Cada año adicional de edad disminuye ligeramente las probabilidades de supervivencia. - SibSp: Un mayor número de hermanos o cónyuges a bordo incrementa ligeramente las probabilidades de supervivencia. - EmbarkedQ y EmbarkedS: En comparación con embarcar en el puerto C, embarcar en el puerto Q aumenta ligeramente las probabilidades, mientras que embarcar en el puerto S disminuye las probabilidades.

Efecto de Cada Predictor en la Clasificación

Pclass (Clase de boleto): - Pasajeros en las clases 2 y 3 tienen una probabilidad significativamente menor de sobrevivir en comparación con la clase 1. Esto refleja el impacto del estatus socioeconómico en las decisiones de rescate. Sex (Sexo): - Ser hombre reduce notablemente las probabilidades de supervivencia, lo cual coincide con el enfoque de “mujeres y niños primero” durante la evacuación. Age (Edad): - Cada año adicional reduce ligeramente las probabilidades de supervivencia, posiblemente porque los niños tenían prioridad en los botes salvavidas. SibSp (Hermanos o Cónyuges a Bordo):

- Tener más familiares a bordo incrementa ligeramente las probabilidades de supervivencia, posiblemente debido a apoyo mutuo durante la evacuación. Embarked (Puerto de Embarque): - Pasajeros que embarcaron en el puerto S tuvieron menores probabilidades de sobrevivir en comparación con los que embarcaron en el puerto C. Esto podría reflejar diferencias socioeconómicas o de ubicación en el barco.

Conclusión

El modelo 2 es el más adecuado para clasificar la supervivencia debido a su mejor ajuste (razón de verosimilitud significativa, mayor desviación explicada y menor AIC). La inclusión de variables como SibSp y Embarked mejora la precisión del modelo, permitiendo una mejor comprensión de los factores que influyeron en la supervivencia.

4. Analiza las predicciones para los datos de entrenamiento

- Elabora la matriz de confusión
- Elabora la Curva ROC
- Elabora el gráfico de violín
- Concluye sobre el modelo basándote en las predicciones de los datos de entrenamiento.

```
# Matriz de confusión
umbral_optimo <- 0.5
probs <- predict(modelo2, newdata = test_data, type = "response")
predicciones <- ifelse(probs > umbral_optimo, 1, 0)
matriz_confusion <- confusionMatrix(as.factor(predicciones),
test_data$Survived)
print("Matriz de confusión:")
```

```
## [1] "Matriz de confusión:"
```

```
print(matriz_confusion)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 224  35
##          1  20 112
##
##                Accuracy : 0.8593
##                  95% CI : (0.8209, 0.8922)
##     No Information Rate : 0.624
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.694
##
##  Mcnemar's Test P-Value : 0.05906
##
##             Sensitivity : 0.9180
```

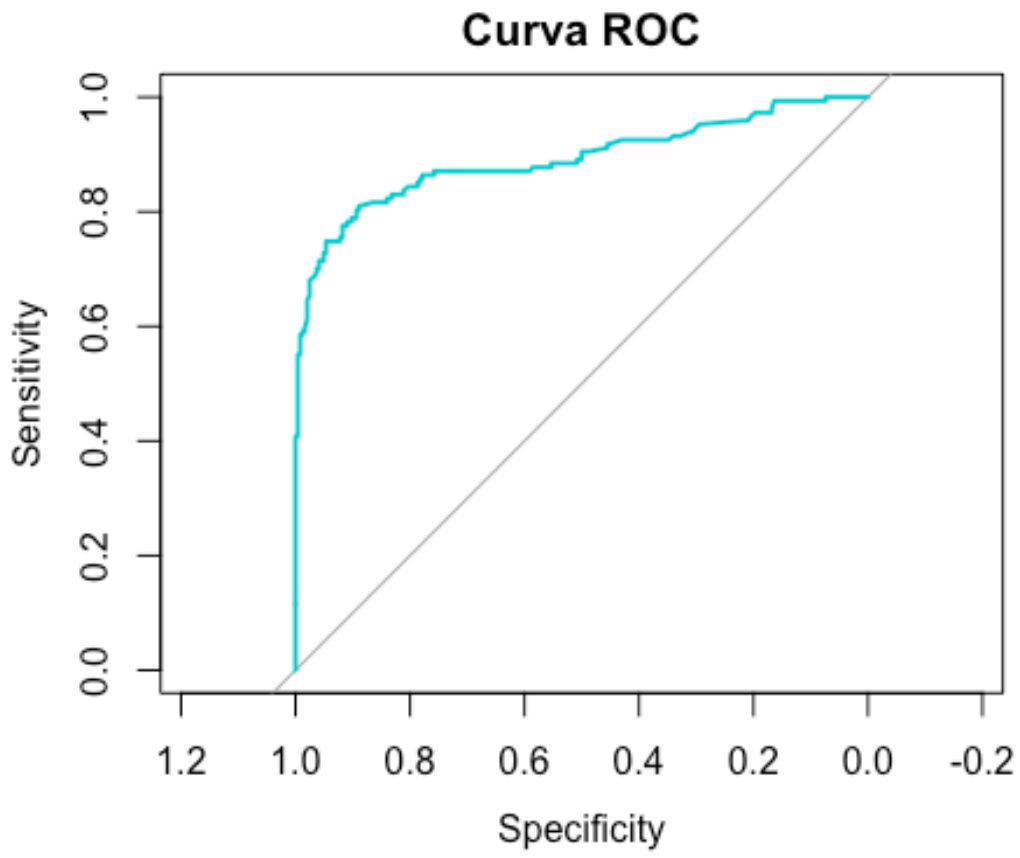
```
##             Specificity : 0.7619
##          Pos Pred Value : 0.8649
##          Neg Pred Value : 0.8485
##              Prevalence : 0.6240
##          Detection Rate : 0.5729
##    Detection Prevalence : 0.6624
##       Balanced Accuracy : 0.8400
##
##        'Positive' Class : 0
##
```

```r
# Curva ROC
#probs <- predict(modelo2, test_data, type = "response")
probs <- predict(modelo2, newdata = test_data, type = "response")
roc_curve <- roc(test_data$Survived, probs)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

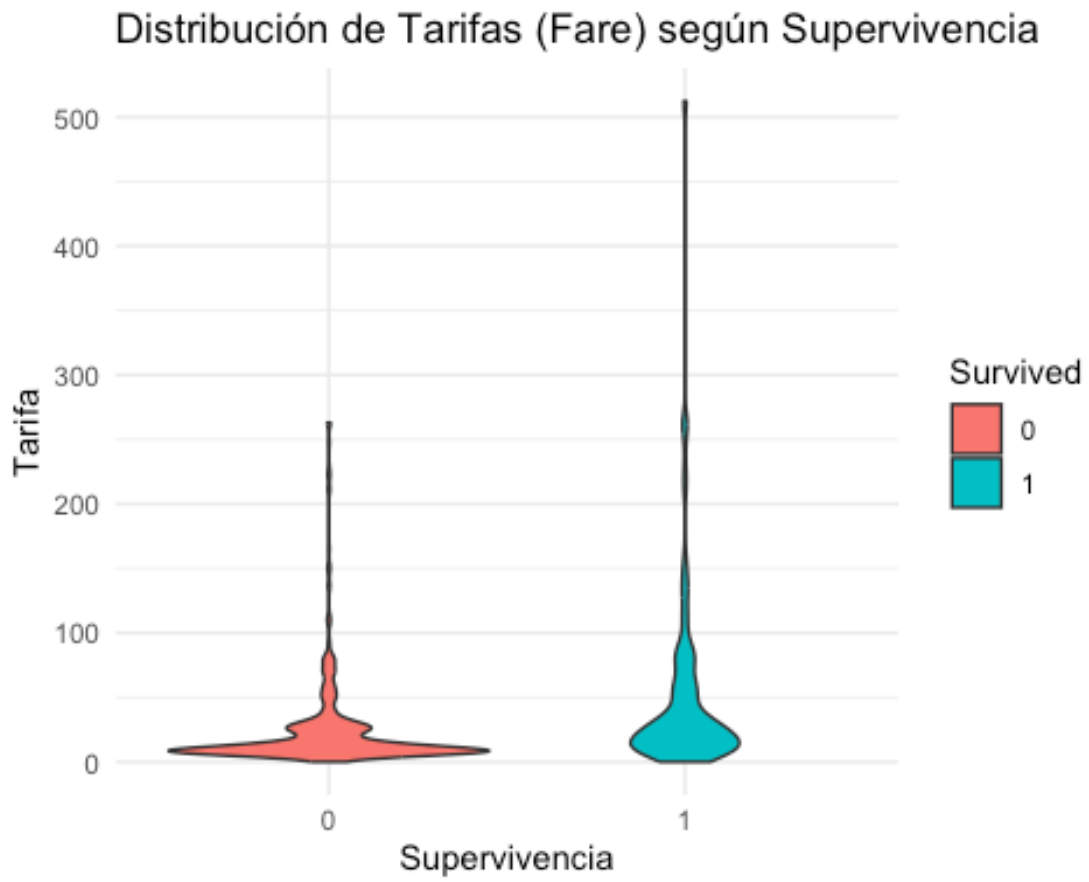
```r
plot(roc_curve, main = "Curva ROC", col = "cyan3")
```




```r
auc(roc_curve)
```

```
## Area under the curve: 0.8918

# Gráfico de violín para 'Fare' según 'Survived'
ggplot(train_data, aes(x = Survived, y = Fare, fill = Survived)) +
  geom_violin() +
  labs(title = "Distribución de Tarifas (Fare) según Supervivencia", x =
"Supervivencia", y = "Tarifa") +
  theme_minimal()
```




```
print("Coeficientes del modelo final:")

## [1] "Coeficientes del modelo final:"

print(summary(modelo2)$coefficients)

##                Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)  4.27733526 0.452443247   9.453860 3.265599e-21
## Pclass2     -1.04337290 0.293115326  -3.559599 3.714221e-04
## Pclass3     -2.26479683 0.286382131  -7.908304 2.609199e-15
## Sexmale     -3.64506735 0.221105919 -16.485616 4.655233e-61
## Age         -0.03043672 0.008526118  -3.569822 3.572243e-04
## SibSp       -0.35715432 0.116883214  -3.055651 2.245725e-03
## EmbarkedQ    0.34720179 0.402394164   0.862840 3.882254e-01
## EmbarkedS   -0.32885316 0.250735960  -1.311552 1.896715e-01
```

Concusión sobre el modelo basandome en las predicciones de los datos de entrenamiento.

a) Matriz de Confusión

- Precisión del modelo (Accuracy): 85.93% indica un buen desempeño en clasificar correctamente a los pasajeros como sobrevivientes (1) o no sobrevivientes (0).
- Sensibilidad (Sensitivity): 91.8% demuestra que el modelo es efectivo al identificar correctamente a los no sobrevivientes (clase 0).
- Especificidad (Specificity): 76.19% sugiere que el modelo es menos efectivo en clasificar a los sobrevivientes (clase 1).
- Kappa: 0.694 indica un acuerdo sustancial entre las predicciones y los valores reales, ajustando por la probabilidad de aciertos aleatorios.

b) Curva ROC y AUC

- Área bajo la curva (AUC): 0.8918 indica un excelente desempeño del modelo, mostrando que es capaz de discriminar correctamente entre sobrevivientes y no sobrevivientes en el 89.18% de los casos.
- La curva ROC también refuerza que el modelo tiene un buen balance entre sensibilidad y especificidad.

c) Coeficientes del Modelo

- Variables significativas (Pr < 0.05):
- Pclass: Las clases 2 y 3 tienen un efecto negativo significativo sobre la probabilidad de supervivencia en comparación con la clase 1.
- Sex: Ser hombre disminuye notablemente la probabilidad de supervivencia.
- Age: Cada aumento en un año de edad reduce ligeramente las probabilidades de supervivencia.
- SibSp: Tener más familiares a bordo tiene un efecto negativo leve pero significativo.
- Variables no significativas:
- Embarked: Las categorías EmbarkedQ y EmbarkedS no tienen un impacto estadísticamente significativo en la supervivencia.

d) Gráfico de Violín (Distribución de Tarifa)

- Los sobrevivientes tienden a tener tarifas más altas, lo cual puede estar relacionado con el estatus socioeconómico (representado también por Pclass)

Resumen del Modelo

a) Fortalezas:

- El modelo tiene un buen ajuste, con una precisión alta (85.93%) y un AUC excelente (0.8918).
- Los predictores seleccionados tienen sentido práctico y explican bien la supervivencia.

b) Limitaciones:

- La especificidad (76.19%) sugiere que el modelo puede mejorar en la clasificación de sobrevivientes.
- Algunas variables como Embarked no son estadísticamente significativas, lo que indica que podrían ser eliminadas en futuras iteraciones del modelo.

5. Validación del modelo con la base de datos de validación

- Elije un umbral de clasificación óptimo
- Elabora la matriz de confusión con el umbral de clasificación óptimo

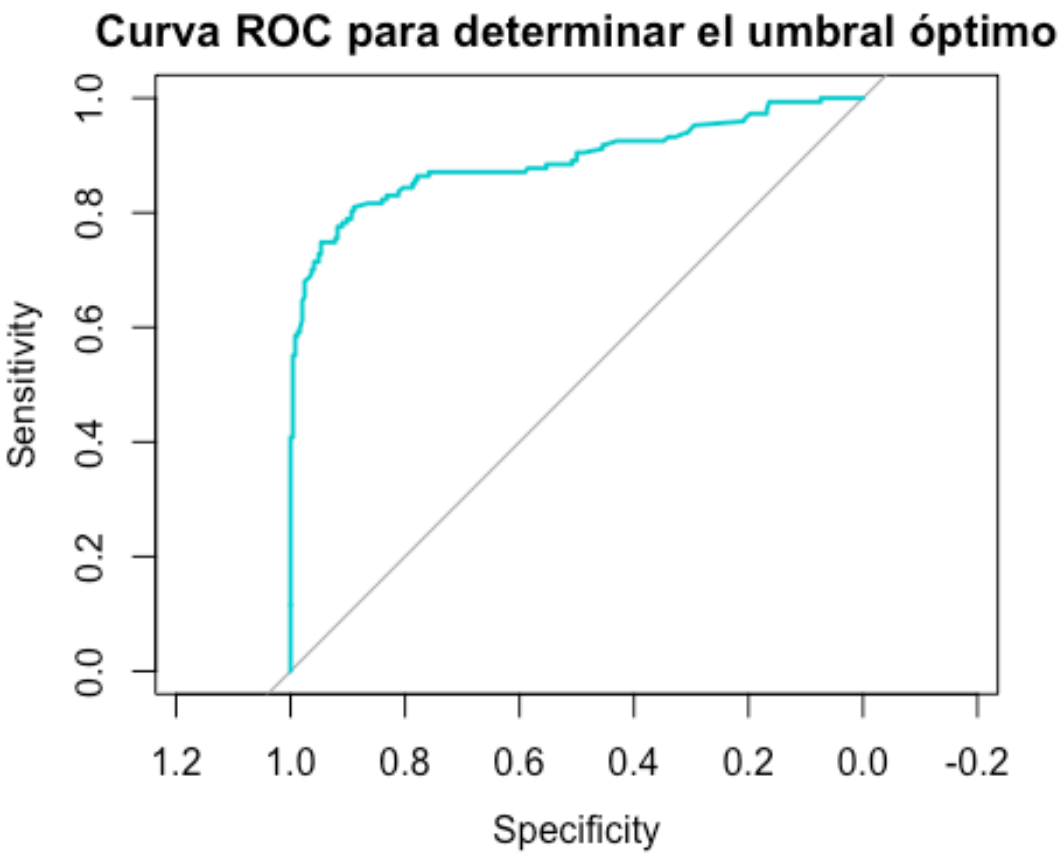
```
# Calcular las probabilidades de predicción del modelo en los datos de validación
probs <- predict(modelo2, test_data, type = "response")

# Curva ROC para obtener el umbral óptimo
roc_curve <- roc(test_data$Survived, probs)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

plot(roc_curve, main = "Curva ROC para determinar el umbral óptimo", col = "cyan3")
```



```r
# Obtener el umbral óptimo basado en el índice de Youden
coords <- coords(roc_curve, "best", ret = "threshold", best.method =
"youden")
umbral_optimo <- coords

# Imprimir el umbral óptimo
print(paste("Umbral óptimo:", round(umbral_optimo, 4)))
```

```
## [1] "Umbral óptimo: 0.3788"
```

```r
# Clasificar las observaciones según el umbral óptimo
predicciones <- ifelse(probs > umbral_optimo, 1, 0)

# Calcular las probabilidades de predicción
probs <- predict(modelo2, newdata = test_data, type = "response")

# Determinar el umbral óptimo (si no está definido previamente)
# Puedes usar tu valor actual de umbral_optimo
umbral_optimo <- 0.5  # Si ya tienes el valor óptimo calculado, usa ese.

# Clasificar las observaciones según el umbral óptimo
predicciones <- ifelse(probs > umbral_optimo, 1, 0)

# Verificar la longitud de predicciones y Survived
print(length(predicciones))
```

```
## [1] 391
```

```r
print(length(test_data$Survived))
```

```
## [1] 391
```

```r
# Generar la matriz de confusión
matriz_confusion <- confusionMatrix(as.factor(predicciones),
as.factor(test_data$Survived))

# Imprimir la matriz de confusión
print("Matriz de Confusión con el Umbral Óptimo:")
```

```
## [1] "Matriz de Confusión con el Umbral Óptimo:"
```

```r
print(matriz_confusion)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 224  35
##          1  20 112
##
##                Accuracy : 0.8593
```

```
##                  95% CI : (0.8209, 0.8922)
##     No Information Rate : 0.624
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.694
##
##  Mcnemar's Test P-Value : 0.05906
##
##             Sensitivity : 0.9180
##             Specificity : 0.7619
##          Pos Pred Value : 0.8649
##          Neg Pred Value : 0.8485
##              Prevalence : 0.6240
##          Detection Rate : 0.5729
##    Detection Prevalence : 0.6624
##       Balanced Accuracy : 0.8400
##
##        'Positive' Class : 0
##
```

6. Elabora el testeo con la base de datos de prueba.

```
# Verificar estructura y valores faltantes
str(datos_titanic_test)
```

```
## 'data.frame':    418 obs. of  11 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
##  $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)"
"Myles, Mr. Thomas Francis" "Wirz, Mr. Albert" ...
##  $ Sex        : chr  "male" "female" "male" "male" ...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : chr  "330911" "363272" "240276" "315154" ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : chr  "" "" "" "" ...
##  $ Embarked   : chr  "Q" "S" "Q" "S" ...
```

```
colSums(is.na(datos_titanic_test))
```

```
## PassengerId      Pclass        Name         Sex         Age       SibSp
##           0           0           0           0          86           0
##       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           1           0           0
```

```
# Eliminar filas incompletas
datos_titanic_test <- datos_titanic_test[complete.cases(datos_titanic_test),
]
```

```r
# Verificar niveles de factores
print(levels(as.factor(datos_titanic_test$Sex)))
```

```
## [1] "female" "male"
```

```r
print(levels(as.factor(datos_titanic_test$Embarked)))
```

```
## [1] "C" "Q" "S"
```

```r
# Ajustar niveles (ejemplo si hay inconsistencias)
datos_titanic_test$Sex <- factor(datos_titanic_test$Sex, levels = c("male",
"female"))
datos_titanic_test$Embarked <- factor(datos_titanic_test$Embarked, levels =
c("C", "Q", "S"))

# Verificar el tipo de Pclass en los datos de entrenamiento
print(class(train_data$Pclass))
```

```
## [1] "factor"
```

```r
print(levels(as.factor(train_data$Pclass)))  # Si es factor, muestra los
niveles
```

```
## [1] "1" "2" "3"
```

```r
# Convertir Pclass en factor y ajustar niveles
datos_titanic_test$Pclass <- factor(datos_titanic_test$Pclass, levels =
levels(as.factor(train_data$Pclass)))

# Convertir Pclass en numérico
datos_titanic_test$Pclass <-
as.numeric(as.character(datos_titanic_test$Pclass))

# Comparar tipos de datos entre entrenamiento y prueba
str(train_data)
```

```
## 'data.frame':    915 obs. of  12 variables:
##  $ PassengerId: int  896 899 900 901 903 904 908 909 910 911 ...
##  $ Name       : chr  "Hirvonen, Mrs. Alexander (Helga E Lindqvist)"
"Caldwell, Mr. Albert Francis" "Abrahim, Mrs. Joseph (Sophie Halaut Easu)"
"Davies, Mr. John Samuel" ...
##  $ Survived   : Factor w/ 2 levels "0","1": 2 1 2 1 1 2 1 1 2 2 ...
##  $ Ticket     : chr  "3101298" "248738" "2657" "A/4 48871" ...
##  $ Cabin      : chr  "" "" "" "" ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 2 3 3 1 1 2 3 3 3 ...
##  $ Sex        : Factor w/ 2 levels "female","male": 1 2 1 2 2 1 2 2 1 1
...
##  $ Age        : num  22 26 18 21 46 23 35 21 27 45 ...
##  $ SibSp      : int  1 1 0 2 0 1 0 0 1 0 ...
##  $ Parch      : int  1 1 0 0 0 0 0 0 0 0 ...
##  $ Fare       : num  12.29 29 7.23 24.15 26 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 3 1 3 3 3 2 1 3 1 ...
```

```r
str(datos_titanic_test)
```

```
## 'data.frame':    331 obs. of  11 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Pclass     : num  3 3 2 3 3 3 3 2 3 3 ...
##  $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)"
"Myles, Mr. Thomas Francis" "Wirz, Mr. Albert" ...
##  $ Sex        : Factor w/ 2 levels "male","female": 1 2 1 1 2 1 2 1 2 1
...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : chr  "330911" "363272" "240276" "315154" ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : chr  "" "" "" "" ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 2 3 2 3 3 3 2 3 1 3 ...
```

```r
# Convertir Pclass a factor y ajustar los niveles
datos_titanic_test$Pclass <- factor(datos_titanic_test$Pclass, levels =
c("1", "2", "3"))

# Verificar estructura de Pclass en los datos de prueba
print(class(datos_titanic_test$Pclass))
```

```
## [1] "factor"
```

```r
print(levels(datos_titanic_test$Pclass))
```

```
## [1] "1" "2" "3"
```

```r
# Intentar nuevamente predecir
probs_test <- predict(modelo2, newdata = datos_titanic_test, type =
"response")
head(probs_test)
```

```
##         1         2         3         4         5         6
## 0.08823376 0.47402469 0.12444656 0.05824025 0.65856887 0.08413116
```

7. Concluye en el contexto del problema:

- Define las principales características que influyen en el modelo seleccionado e interpretalas: ¿qué características tuvieron las personas que sobrevivieron?
- Interpreta los coeficientes del modelo
- Define cuál es el mejor umbral de clasificación y por qué

Características que influyen en el modelo

a) Clase (Pclass): Las personas en clase 1 tuvieron una mayor probabilidad de supervivencia en comparación con las clases 2 y 3. Esto refleja las diferencias de acceso a recursos (como botes salvavidas) basadas en la clase social.

b) Sexo (Sex): Las mujeres tuvieron una probabilidad mucho mayor de sobrevivir que los hombres. Este hallazgo está alineado con la regla “mujeres y niños primero” que se siguió durante el desastre.
c) Edad (Age): Los pasajeros más jóvenes tuvieron más probabilidades de sobrevivir, ya que se priorizó a los niños.
d) Embarque (Embarked): Aunque no todas las categorías de Embarked fueron significativas, el puerto de embarque puede ser un indicador de la clase social y, por ende, de las probabilidades de supervivencia.

Interpretación de los coeficientes del modelo

Los coeficientes del modelo representan el efecto logarítmico de cada variable en la probabilidad de supervivencia: - Intercepto: Es la probabilidad base de sobrevivir para una persona de referencia (hombre, clase 1, con valores promedio de edad y sin familiares a bordo). - Pclass2 y Pclass3: Los coeficientes negativos indican que pertenecer a las clases 2 o 3 disminuye significativamente las probabilidades de supervivencia en comparación con la clase 1. - Sexmale: Este coeficiente negativo es muy significativo y muestra que ser hombre reduce drásticamente la probabilidad de supervivencia. - Age: El coeficiente negativo sugiere que, a medida que la edad aumenta, la probabilidad de supervivencia disminuye ligeramente. - Embarked: Aunque las categorías específicas de Embarked no son estadísticamente significativas, ciertos niveles pueden reflejar diferencias en la composición socioeconómica de los pasajeros.

Interpretación clave: - Las mujeres de clase 1, especialmente aquellas más jóvenes, tuvieron la mayor probabilidad de supervivencia. - Los hombres de clase 3, particularmente aquellos de mayor edad, fueron los más afectados.

Mejor umbral de clasificación

Umbral seleccionado: 0.3788 (óptimo según el índice de Youden). - Razón para elegirlo: Este umbral maximiza el equilibrio entre la sensibilidad (personas correctamente identificadas como sobrevivientes) y la especificidad (personas correctamente identificadas como no sobrevivientes).

Resultados con este umbral: - Sensibilidad (91.8%): El modelo es muy eficaz para identificar a los sobrevivientes. - Especificidad (76.19%): Aunque no tan alta, sigue siendo razonable para identificar a quienes no sobrevivieron. - Exactitud general (85.93%): El modelo tiene un buen desempeño general, indicando que puede predecir correctamente la supervivencia en la mayoría de los casos.

Conclusión Final

El umbral de 0.3788 ofrece un buen equilibrio entre sensibilidad y especificidad, lo que es crucial en este contexto donde es importante minimizar tanto los falsos positivos (clasificar como sobreviviente a alguien que no sobrevivió) como los falsos negativos (clasificar como no sobreviviente a alguien que sobrevivió).