Actividad Integradora 2

Eliezer Cavazos

2024-11-19

```r
# Cargamos todas las librería en la lista "librerias"
librerias =
c('tidyverse','broom','ISLR','GGally','modelr','cowplot','rlang','modelr','ti
bble','Metrics','mice','visdat',"caret")

for (lib in librerias){
  library(lib,character.only=TRUE)}
```

```
## ── 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

## Warning: package 'GGally' was built under R version 4.4.2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
##
## Attaching package: 'modelr'
##
## The following object is masked from 'package:broom':
##
##     bootstrap
##
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
##
```

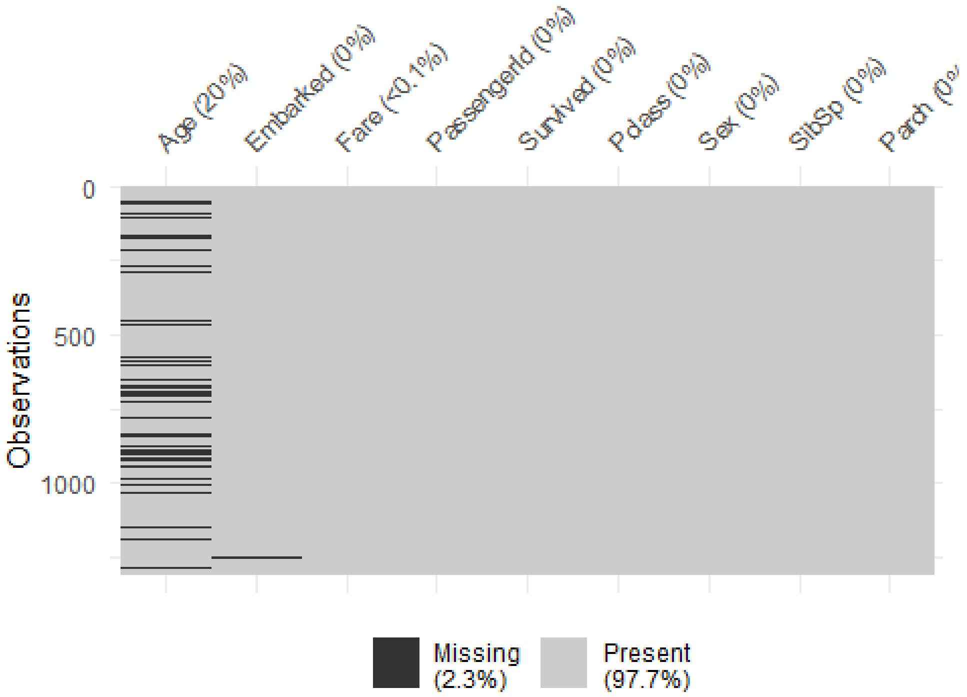
```
##
## Attaching package: 'rlang'
##
## The following objects are masked from 'package:purrr':
##
##     %@%, flatten, flatten_chr, flatten_dbl, flatten_int, flatten_lgl,
##     flatten_raw, invoke, splice

## Warning: package 'Metrics' was built under R version 4.4.2

##
## Attaching package: 'Metrics'
##
## The following object is masked from 'package:rlang':
##
##     ll
##
## The following objects are masked from 'package:modelr':
##
##     mae, mape, mse, rmse

## Warning: package 'mice' was built under R version 4.4.2

##
## Attaching package: 'mice'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     cbind, rbind

## Warning: package 'visdat' was built under R version 4.4.2

## Warning: package 'caret' was built under R version 4.4.2

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following objects are masked from 'package:Metrics':
##
##     precision, recall
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
oTitanic =
read.csv("C:\\Users\\eliez\\OneDrive\\Desktop\\Clases\\Titanic.csv") #leer la
base de datos
```

1. Prepara la base de datos Titanic:

1.1 Analiza los datos faltantes

```
# Eliminar variables:
oTitanic <- oTitanic[,c(-4,-9,-11)]

#Transformar a factores:
for(var in c('Survived','Pclass','Embarked','Sex'))
  oTitanic[,var] <-as.factor(oTitanic[,var])

colSums(is.na(oTitanic))

## PassengerId    Survived      Pclass         Sex         Age       SibSp
##           0           0           0           0         263           0
##       Parch        Fare    Embarked
##           0           1           2

vis_miss(oTitanic,sort_miss = TRUE)
```



Medidas con datos faltantes

```
summary(oTitanic[,-1])
```

```
##  Survived Pclass      Sex           Age            SibSp            Parch
##  0:815    1:323    female:466   Min.   : 0.17   Min.   :0.0000   Min.
:0.000
##  1:494    2:277    male  :843   1st Qu.:21.00   1st Qu.:0.0000   1st
Qu.:0.000
##           3:709                Median :28.00   Median :0.0000   Median
:0.000
##                                Mean   :29.88   Mean   :0.4989   Mean
:0.385
##                                3rd Qu.:39.00   3rd Qu.:1.0000   3rd
Qu.:0.000
##                                Max.   :80.00   Max.   :8.0000   Max.
:9.000
##                                NA's   :263
##       Fare           Embarked
##  Min.   :  0.000   C    :270
##  1st Qu.:  7.896   Q    :123
##  Median : 14.454   S    :914
##  Mean   : 33.295   NA's:  2
##  3rd Qu.: 31.275
##  Max.   :512.329
##  NA's   :1
```

Medidas sin datos faltantes

```
M2 = na.omit(oTitanic)
summary(M2[,-1])
```

```
##  Survived Pclass      Sex           Age            SibSp
##  0:628    1:282    female:386   Min.   : 0.17   Min.   :0.0000
##  1:415    2:261    male  :657   1st Qu.:21.00   1st Qu.:0.0000
##           3:500                Median :28.00   Median :0.0000
##                                Mean   :29.81   Mean   :0.5043
##                                3rd Qu.:39.00   3rd Qu.:1.0000
##                                Max.   :80.00   Max.   :8.0000
##      Parch             Fare         Embarked
##  Min.   :0.0000   Min.   :  0.00   C:212
##  1st Qu.:0.0000   1st Qu.:  8.05   Q: 50
##  Median :0.0000   Median : 15.75   S:781
##  Mean   :0.4219   Mean   : 36.60
##  3rd Qu.:1.0000   3rd Qu.: 35.08
##  Max.   :6.0000   Max.   :512.33
```

Sobrevivientes

```
t2c = 100*prop.table(table(oTitanic[,2]))
t2s = 100*prop.table(table(M2[,2]))
t2p = c(t2s[1]/t2c[1],t2s[2]/t2c[2])
t2 = data.frame(as.numeric(t2c),as.numeric(t2s),as.numeric(t2p))
row.names(t2) = c("Murió","Sobrevivió")
```

```r
names(t2) = c("Con NA (%)","Sin NA (%)","Pérdida (prop)")
round(t2,2)
```

```
##              Con NA (%) Sin NA (%) Pérdida (prop)
## Murió             62.26      60.21           0.97
## Sobrevivió        37.74      39.79           1.05
```

Clase en que viajó

```r
t3c = 100*prop.table(table(oTitanic[,3]))
t3s = 100*prop.table(table(M2[,3]))
t3p = c(t3s[1]/t3c[1],t3s[2]/t3c[2],t3s[3]/t3c[3])
t3 = data.frame(as.numeric(t3c),as.numeric(t3s),as.numeric(t3p))
row.names(t3) = c("Primera","Segunda","Tercera")
names(t3) = c("Con NA (%)","Sin NA (%)","Pérdida (prop)")
round(t3,2)
```

```
##           Con NA (%) Sin NA (%) Pérdida (prop)
## Primera        24.68      27.04           1.10
## Segunda        21.16      25.02           1.18
## Tercera        54.16      47.94           0.89
```

Sexo

```r
t4c = 100*prop.table(table(oTitanic[,4]))
t4s = 100*prop.table(table(M2[,4]))
t4p = c(t4s[1]/t4c[1],t4s[2]/t4c[2])
t4 = data.frame(as.numeric(t4c),as.numeric(t4s),as.numeric(t4p))
row.names(t4) = c("Mujer","Hombre")
names(t4) = c("Con NA (%)","Sin NA (%)","Pérdida (prop)")
round(t4,2)
```

```
##          Con NA (%) Sin NA (%) Pérdida (prop)
## Mujer          35.6      37.01           1.04
## Hombre         64.4      62.99           0.98
```

Puerto de embarcación

```r
t9c = 100*prop.table(table(oTitanic[,9]))
t9s = 100*prop.table(table(M2[,9]))
t9p = c(t9s[1]/t9c[1],t9s[2]/t9c[2],t9s[3]/t9c[3])
t9 = data.frame(as.numeric(t9c),as.numeric(t9s),as.numeric(t9p))
row.names(t9) = c("Cherbourg","Queenstown","Southampton")
names(t9) = c("Con NA (%)","Sin NA (%)","Pérdida (prop)")
round(t9,2)
```

```
##               Con NA (%) Sin NA (%) Pérdida (prop)
## Cherbourg          20.66      20.33           0.98
## Queenstown          9.41       4.79           0.51
## Southampton        69.93      74.88           1.07
```

1.2 Realiza un análisis descriptivo

```
summary(oTitanic)

##   PassengerId   Survived Pclass      Sex           Age             SibSp
##  Min.   :   1   0:815    1:323   female:466   Min.   : 0.17   Min.
:0.0000
##  1st Qu.: 328   1:494    2:277   male  :843   1st Qu.:21.00   1st
Qu.:0.0000
##  Median : 655            3:709                Median :28.00   Median
:0.0000
##  Mean   : 655                                 Mean   :29.88   Mean
:0.4989
##  3rd Qu.: 982                                 3rd Qu.:39.00   3rd
Qu.:1.0000
##  Max.   :1309                                 Max.   :80.00   Max.
:8.0000
##                                               NA's   :263
##      Parch            Fare          Embarked
##  Min.   :0.000   Min.   :  0.000   C   :270
##  1st Qu.:0.000   1st Qu.:  7.896   Q   :123
##  Median :0.000   Median : 14.454   S   :914
##  Mean   :0.385   Mean   : 33.295   NA's:  2
##  3rd Qu.:0.000   3rd Qu.: 31.275
##  Max.   :9.000   Max.   :512.329
##                  NA's   :1

table(oTitanic$Survived)

##
##   0   1
## 815 494
```

1.3 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.

```
library(caret)
index <- createDataPartition(M2$Survived, p = 0.7, list = FALSE)
oTitanicTrainData <- M2[ index,]  %>% as_tibble()
oTitanicTestData <- M2[-index,] %>% as_tibble()
```

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.

2.1 Auxiliate del criterio de AIC para determinar cuál es el mejor modelo.

Modelos sin Relacion

```
oModelo = glm(Survived ~ ., data = oTitanicTrainData, family = "binomial")
step(oModelo, direction="both", trace=1 )
```

```
## Start:  AIC=610.41
## Survived ~ PassengerId + Pclass + Sex + Age + SibSp + Parch + 
##     Fare + Embarked
## 
##               Df Deviance    AIC
## - Embarked     2   588.81 606.81
## - Fare         1   588.46 608.46
## - SibSp        1   588.62 608.62
## - PassengerId  1   589.12 609.12
## <none>             588.41 610.41
## - Parch        1   591.39 611.39
## - Age          1   599.60 619.60
## - Pclass       2   623.91 641.91
## - Sex          1   882.16 902.16
## 
## Step:  AIC=606.81
## Survived ~ PassengerId + Pclass + Sex + Age + SibSp + Parch + 
##     Fare
## 
##               Df Deviance    AIC
## - Fare         1   588.89 604.89
## - SibSp        1   589.06 605.06
## - PassengerId  1   589.59 605.59
## <none>             588.81 606.81
## - Parch        1   591.69 607.69
## + Embarked     2   588.41 610.41
## - Age          1   600.74 616.74
## - Pclass       2   628.08 642.08
## - Sex          1   886.44 902.44
## 
## Step:  AIC=604.89
## Survived ~ PassengerId + Pclass + Sex + Age + SibSp + Parch
## 
##               Df Deviance    AIC
## - SibSp        1   589.11 603.11
## - PassengerId  1   589.64 603.64
## <none>             588.89 604.89
## - Parch        1   591.71 605.71
## + Fare         1   588.81 606.81
## + Embarked     2   588.46 608.46
## - Age          1   600.90 614.90
## - Pclass       2   650.30 662.30
## - Sex          1   888.25 902.25
## 
## Step:  AIC=603.11
## Survived ~ PassengerId + Pclass + Sex + Age + Parch
## 
```

```
##                Df Deviance    AIC
## - PassengerId  1   589.80 601.80
## <none>             589.11 603.11
## - Parch        1   592.84 604.84
## + SibSp        1   588.89 604.89
## + Fare         1   589.06 605.06
## + Embarked     2   588.65 606.65
## - Age          1   600.97 612.97
## - Pclass       2   650.32 660.32
## - Sex          1   888.41 900.41
##
## Step:  AIC=601.8
## Survived ~ Pclass + Sex + Age + Parch
##
##                Df Deviance    AIC
## <none>             589.80 601.80
## + PassengerId  1   589.11 603.11
## - Parch        1   593.40 603.40
## + SibSp        1   589.64 603.64
## + Fare         1   589.77 603.77
## + Embarked     2   589.29 605.29
## - Age          1   601.57 611.57
## - Pclass       2   650.44 658.44
## - Sex          1   888.59 898.59

##
## Call:  glm(formula = Survived ~ Pclass + Sex + Age + Parch, family =
"binomial",
##     data = oTitanicTrainData)
##
## Coefficients:
## (Intercept)      Pclass2      Pclass3      Sexmale          Age
Parch
##     3.88576     -1.27814     -2.17890     -3.37713     -0.02839     -
0.22517
##
## Degrees of Freedom: 730 Total (i.e. Null);  725 Residual
## Null Deviance:       982.8
## Residual Deviance: 589.8     AIC: 601.8
```

Modelo con Relacion

```
oModelo = glm(Survived ~ Pclass * Sex * Age * Parch * Fare, data =
oTitanicTrainData, family = "binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

step(oModelo, direction="both", trace=1 )

## Start:  AIC=596.5
## Survived ~ Pclass * Sex * Age * Parch * Fare
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                              Df Deviance    AIC
## - Pclass:Sex:Age:Parch:Fare  2   500.96 592.96
## <none>                           500.50 596.50

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=592.96
## Survived ~ Pclass + Sex + Age + Parch + Fare + Pclass:Sex + Pclass:Age +
##     Sex:Age + Pclass:Parch + Sex:Parch + Age:Parch + Pclass:Fare +
##     Sex:Fare + Age:Fare + Parch:Fare + Pclass:Sex:Age + Pclass:Sex:Parch +
##     Pclass:Age:Parch + Sex:Age:Parch + Pclass:Sex:Fare + Pclass:Age:Fare +
##     Sex:Age:Fare + Pclass:Parch:Fare + Sex:Parch:Fare + Age:Parch:Fare +
##     Pclass:Sex:Age:Parch + Pclass:Sex:Age:Fare + Pclass:Sex:Parch:Fare +
##     Pclass:Age:Parch:Fare + Sex:Age:Parch:Fare

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                              Df Deviance    AIC
## - Pclass:Sex:Age:Fare         2   501.06 589.06
## - Pclass:Sex:Parch:Fare       2   502.00 590.00
## - Pclass:Age:Parch:Fare       2   502.41 590.41
## - Sex:Age:Parch:Fare          1   501.47 591.47
## <none>                            500.96 592.96
## - Pclass:Sex:Age:Parch        2   505.85 593.85
## + Pclass:Sex:Age:Parch:Fare  2   500.50 596.50

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=589.06
## Survived ~ Pclass + Sex + Age + Parch + Fare + Pclass:Sex + Pclass:Age +
##     Sex:Age + Pclass:Parch + Sex:Parch + Age:Parch + Pclass:Fare +
##     Sex:Fare + Age:Fare + Parch:Fare + Pclass:Sex:Age + Pclass:Sex:Parch +
##     Pclass:Age:Parch + Sex:Age:Parch + Pclass:Sex:Fare + Pclass:Age:Fare +
##     Sex:Age:Fare + Pclass:Parch:Fare + Sex:Parch:Fare + Age:Parch:Fare +
##     Pclass:Sex:Age:Parch + Pclass:Sex:Parch:Fare + Pclass:Age:Parch:Fare +
##     Sex:Age:Parch:Fare

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                          Df Deviance    AIC
## - Pclass:Sex:Parch:Fare  2   502.64 586.64
## - Pclass:Age:Parch:Fare  2   503.20 587.20
## - Sex:Age:Parch:Fare     1   501.57 587.57
## <none>                       501.06 589.06
## - Pclass:Sex:Age:Parch   2   506.72 590.72
## + Pclass:Sex:Age:Fare    2   500.96 592.96

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## 
## Step:  AIC=586.64
## Survived ~ Pclass + Sex + Age + Parch + Fare + Pclass:Sex + Pclass:Age + 
##     Sex:Age + Pclass:Parch + Sex:Parch + Age:Parch + Pclass:Fare + 
##     Sex:Fare + Age:Fare + Parch:Fare + Pclass:Sex:Age + Pclass:Sex:Parch + 
##     Pclass:Age:Parch + Sex:Age:Parch + Pclass:Sex:Fare + Pclass:Age:Fare + 
##     Sex:Age:Fare + Pclass:Parch:Fare + Sex:Parch:Fare + Age:Parch:Fare + 
##     Pclass:Sex:Age:Parch + Pclass:Age:Parch:Fare + Sex:Age:Parch:Fare

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                          Df Deviance    AIC
## - Pclass:Age:Parch:Fare  2   503.49 583.49
## - Pclass:Sex:Fare        2   503.50 583.50
## - Sex:Age:Parch:Fare     1   502.78 584.78
## <none>                       502.64 586.64
## - Pclass:Sex:Age:Parch   2   508.19 588.19
## + Pclass:Sex:Parch:Fare  2   501.06 589.06
## + Pclass:Sex:Age:Fare    2   502.00 590.00

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## 
## Step:  AIC=583.49
## Survived ~ Pclass + Sex + Age + Parch + Fare + Pclass:Sex + Pclass:Age + 
##     Sex:Age + Pclass:Parch + Sex:Parch + Age:Parch + Pclass:Fare + 
##     Sex:Fare + Age:Fare + Parch:Fare + Pclass:Sex:Age + Pclass:Sex:Parch + 
##     Pclass:Age:Parch + Sex:Age:Parch + Pclass:Sex:Fare + Pclass:Age:Fare + 
##     Sex:Age:Fare + Pclass:Parch:Fare + Sex:Parch:Fare + Age:Parch:Fare + 
##     Pclass:Sex:Age:Parch + Sex:Age:Parch:Fare

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                          Df Deviance    AIC
## - Pclass:Parch:Fare      2   504.15 580.15
## - Pclass:Sex:Fare        2   504.40 580.40
## - Sex:Age:Parch:Fare     1   503.77 581.77
## - Pclass:Age:Fare        2   505.87 581.87
## <none>                       503.49 583.49
## + Pclass:Sex:Age:Fare    2   502.43 586.43
## + Pclass:Age:Parch:Fare  2   502.64 586.64
## + Pclass:Sex:Parch:Fare  2   503.20 587.20
## - Pclass:Sex:Age:Parch   2   511.46 587.46

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=580.15
## Survived ~ Pclass + Sex + Age + Parch + Fare + Pclass:Sex + Pclass:Age +
##     Sex:Age + Pclass:Parch + Sex:Parch + Age:Parch + Pclass:Fare +
##     Sex:Fare + Age:Fare + Parch:Fare + Pclass:Sex:Age + Pclass:Sex:Parch +
##     Pclass:Age:Parch + Sex:Age:Parch + Pclass:Sex:Fare + Pclass:Age:Fare +
##     Sex:Age:Fare + Sex:Parch:Fare + Age:Parch:Fare + Pclass:Sex:Age:Parch
+
##     Sex:Age:Parch:Fare

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                         Df Deviance    AIC
## - Sex:Age:Parch:Fare    1   504.19 578.19
## - Pclass:Sex:Fare       2   508.08 580.08
## <none>                      504.15 580.15
## - Pclass:Age:Fare       2   508.27 580.27
## + Pclass:Sex:Age:Fare   2   502.74 582.74
## + Pclass:Parch:Fare     2   503.49 583.49
## - Pclass:Sex:Age:Parch  2   512.66 584.66

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=578.19
## Survived ~ Pclass + Sex + Age + Parch + Fare + Pclass:Sex + Pclass:Age +
##     Sex:Age + Pclass:Parch + Sex:Parch + Age:Parch + Pclass:Fare +
##     Sex:Fare + Age:Fare + Parch:Fare + Pclass:Sex:Age + Pclass:Sex:Parch +
##     Pclass:Age:Parch + Sex:Age:Parch + Pclass:Sex:Fare + Pclass:Age:Fare +
##     Sex:Age:Fare + Sex:Parch:Fare + Age:Parch:Fare + Pclass:Sex:Age:Parch

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                        Df Deviance    AIC
## - Sex:Age:Fare          1   504.50 576.50
## - Age:Parch:Fare        1   504.86 576.86
## <none>                      504.19 578.19
## - Pclass:Age:Fare       2   508.28 578.28
## - Pclass:Sex:Fare       2   509.05 579.05
## + Sex:Age:Parch:Fare    1   504.15 580.15
## - Sex:Parch:Fare        1   508.32 580.32
## + Pclass:Sex:Age:Fare   2   502.75 580.75
## + Pclass:Parch:Fare     2   503.77 581.77
## - Pclass:Sex:Age:Parch  2   512.67 582.67

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=576.5
## Survived ~ Pclass + Sex + Age + Parch + Fare + Pclass:Sex + Pclass:Age +
##     Sex:Age + Pclass:Parch + Sex:Parch + Age:Parch + Pclass:Fare +
##     Sex:Fare + Age:Fare + Parch:Fare + Pclass:Sex:Age + Pclass:Sex:Parch +
##     Pclass:Age:Parch + Sex:Age:Parch + Pclass:Sex:Fare + Pclass:Age:Fare +
##     Sex:Parch:Fare + Age:Parch:Fare + Pclass:Sex:Age:Parch

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                        Df Deviance    AIC
## - Age:Parch:Fare        1   504.99 574.99
## - Pclass:Age:Fare       2   508.45 576.45
## <none>                      504.50 576.50
## - Pclass:Sex:Fare       2   509.13 577.13
## + Sex:Age:Fare          1   504.19 578.19
## - Sex:Parch:Fare        1   508.35 578.35
## + Pclass:Parch:Fare     2   503.95 579.95
## - Pclass:Sex:Age:Parch  2   512.89 580.89

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=574.99
## Survived ~ Pclass + Sex + Age + Parch + Fare + Pclass:Sex + Pclass:Age +
##     Sex:Age + Pclass:Parch + Sex:Parch + Age:Parch + Pclass:Fare +
##     Sex:Fare + Age:Fare + Parch:Fare + Pclass:Sex:Age + Pclass:Sex:Parch +
##     Pclass:Age:Parch + Sex:Age:Parch + Pclass:Sex:Fare + Pclass:Age:Fare +
##     Sex:Parch:Fare + Pclass:Sex:Age:Parch

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                        Df Deviance    AIC
## <none>                      504.99 574.99
## - Pclass:Sex:Fare       2   509.38 575.38
## - Pclass:Age:Fare       2   509.55 575.55
## + Age:Parch:Fare        1   504.50 576.50
## + Sex:Age:Fare          1   504.86 576.86
## - Sex:Parch:Fare        1   509.05 577.05
## + Pclass:Parch:Fare     2   504.27 578.27
## - Pclass:Sex:Age:Parch  2   513.61 579.61

##
## Call:  glm(formula = Survived ~ Pclass + Sex + Age + Parch + Fare +
##     Pclass:Sex + Pclass:Age + Sex:Age + Pclass:Parch + Sex:Parch +
##     Age:Parch + Pclass:Fare + Sex:Fare + Age:Fare + Parch:Fare +
##     Pclass:Sex:Age + Pclass:Sex:Parch + Pclass:Age:Parch + Sex:Age:Parch +
##     Pclass:Sex:Fare + Pclass:Age:Fare + Sex:Parch:Fare +
Pclass:Sex:Age:Parch,
##     family = "binomial", data = oTitanicTrainData)
##
## Coefficients:
##               (Intercept)                     Pclass2
##                 5.0039716                  -3.6984738
##                   Pclass3                     Sexmale
##                -2.2068858                  -5.3127490
##                       Age                       Parch
##                -0.0910428                 -3.9977322
##                      Fare            Pclass2:Sexmale
##                 0.0616414                  0.4037771
##           Pclass3:Sexmale                Pclass2:Age
##                 1.8382040                  0.1185475
##               Pclass3:Age                Sexmale:Age
##                -0.0014439                  0.0831155
##             Pclass2:Parch              Pclass3:Parch
##                 5.1382824                  4.8481610
##             Sexmale:Parch                  Age:Parch
##                 4.2565951                  0.1291181
##              Pclass2:Fare               Pclass3:Fare
##                 0.0339126                 -0.1664011
##              Sexmale:Fare                   Age:Fare
##                -0.0696740                  0.0001644
##                Parch:Fare        Pclass2:Sexmale:Age
##                -0.0285332                 -0.0568325
##       Pclass3:Sexmale:Age      Pclass2:Sexmale:Parch
##                -0.0539403                 27.1136028
##     Pclass3:Sexmale:Parch          Pclass2:Age:Parch
##                -3.9114140                 -0.1103127
##         Pclass3:Age:Parch          Sexmale:Age:Parch
##                -0.1408609                 -0.1629488
##      Pclass2:Sexmale:Fare       Pclass3:Sexmale:Fare
##                 0.0036922                  0.1009472
##          Pclass2:Age:Fare           Pclass3:Age:Fare
##                -0.0029366                  0.0040333
##        Sexmale:Parch:Fare  Pclass2:Sexmale:Age:Parch
##                 0.0325050                 -4.7322398
## Pclass3:Sexmale:Age:Parch
##                 0.0451688
##
## Degrees of Freedom: 730 Total (i.e. Null);  696 Residual
## Null Deviance:       982.8
## Residual Deviance: 505   AIC: 575
```

2.2 Propón por lo menos los dos que consideres mejores modelos.

Los dos modelos que voy a usar son: el mejor modelo sin la relacion de las variables y el mejor modelo con la relacion de variables.

Para el primer modelo usando el criterio de AIC me dio un valor de: 550.98 con la siguiente ecuacion de variables:

Survived ~ Pclass + Sex + Age + SibSp

Pero el mejor modelo que salio fue el que hice con la relacion de variables donde medio un valor con el criterio de AIC de: 495 y con la siguiente ecuacion de las variables:

Survived ~ Pclass + Sex + Age + Parch + Fare + Pclass:Sex + Pclass:Age + Sex:Age + Pclass:Parch + Sex:Parch + Age:Parch + Pclass:Fare + Sex:Fare + Age:Fare + Parch:Fare + Pclass:Sex:Age + Sex:Age:Parch + Pclass:Sex:Fare + Pclass:Age:Fare + Sex:Age:Fare + Age:Parch:Fare + Pclass:Sex:Age:Fare

```
oModelo1 = glm(formula = Survived ~ Pclass + Sex + Age + SibSp, family =
"binomial", data = oTitanicTrainData)
oModelo2 = glm(formula = Survived ~ Pclass + Sex + Age + Parch + Fare +
    Pclass:Sex + Pclass:Age + Sex:Age + Pclass:Parch + Sex:Parch +
    Age:Parch + Pclass:Fare + Sex:Fare + Age:Fare + Parch:Fare +
    Pclass:Sex:Age + Sex:Age:Parch + Pclass:Sex:Fare + Pclass:Age:Fare +
    Sex:Age:Fare + Age:Parch:Fare + Pclass:Sex:Age:Fare, family = "binomial",
    data = oTitanicTrainData)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

3. Analiza los modelos a través de:

3.1 Identificación de la Desviación residual de cada modelo

```
iM1Deviance = oModelo1$deviance
print("Modelo 1: ")

## [1] "Modelo 1: "

iM1Deviance

## [1] 592.4037

iM2Deviance = oModelo2$deviance
print("Modelo 2: ")

## [1] "Modelo 2: "

iM2Deviance

## [1] 520.6712
```

3.2 Identificación de la Desviación nula

```
iM1NullDeviance = oModelo1$null.deviance
print("Modelo 1: ")

## [1] "Modelo 1: "

iM1NullDeviance

## [1] 982.7966

iM2NullDeviance = oModelo2$null.deviance
print("Modelo 2: ")

## [1] "Modelo 2: "
```

```
iM2NullDeviance

## [1] 982.7966
```

Tabla Comparativa

```
library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

anova(oModelo1,oModelo2,test="LR")

## Analysis of Deviance Table
##
## Model 1: Survived ~ Pclass + Sex + Age + SibSp
## Model 2: Survived ~ Pclass + Sex + Age + Parch + Fare + Pclass:Sex +
Pclass:Age +
##     Sex:Age + Pclass:Parch + Sex:Parch + Age:Parch + Pclass:Fare +
##     Sex:Fare + Age:Fare + Parch:Fare + Pclass:Sex:Age + Sex:Age:Parch +
##     Pclass:Sex:Fare + Pclass:Age:Fare + Sex:Age:Fare + Age:Parch:Fare +
##     Pclass:Sex:Age:Fare
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       725     592.40
## 2       699     520.67 26   71.733  3.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.3 Cálculo de la Desviación Explicada

Modelo 1

```
desviacion_explicada_Modelo1 <- (1 - (iM1Deviance / iM1NullDeviance)) * 100
```

Modelo 2

```
desviacion_explicada_Modelo2 <- (1 - (iM2Deviance / iM2NullDeviance)) * 100

cat("Desviación Explicada del Modelo 1: ",
round(desviacion_explicada_Modelo1, 2), "%\n")

## Desviación Explicada del Modelo 1:  39.72 %
```

```
cat("Desviación Explicada del Modelo 2: ",
round(desviacion_explicada_Modelo2, 2), "%\n")

## Desviación Explicada del Modelo 2:  47.02 %
```

3.4 Prueba de la razón de verosimilitud

Modelo 1

```
Diferencia = oModelo1$null.deviance-oModelo1$deviance
gl = oModelo1$df.null - oModelo1$df.deviance

pchisq(Diferencia,gl,lower.tail = FALSE)

## numeric(0)
```

Modelo 2

```
Diferencia = oModelo2$null.deviance-oModelo2$deviance
gl = oModelo2$df.null - oModelo2$df.deviance

pchisq(Diferencia,gl,lower.tail = FALSE)

## numeric(0)
```

3.5 Define cuál es el mejor modelo

El modelo seleccionado que voy a usar el Modelo 2, ya que este modelo tiene muchos coeficientes relevantes con un valor P mayor a 0.05 ademas de que el valor de AIC es el más bajo de ambos modelos y aunque en la desviacion explicada tiene un valor mayor, su desviacion es más baja

3.6 Escribe su ecuación, analiza sus coeficientes y detecta el efecto de cada predictor en la clasificación.

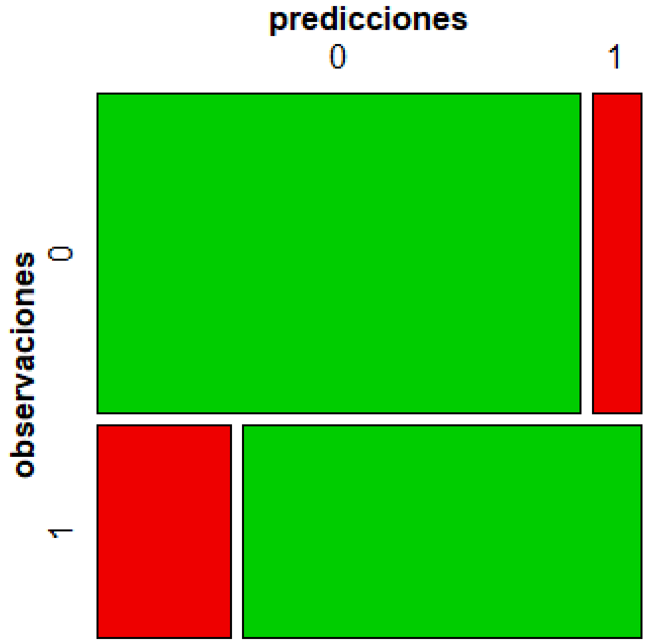
```
oModelo2$coefficients

##           (Intercept)               Pclass2               Pclass3
##          4.4398697580         -3.6302585841         -2.5079172393
##               Sexmale                   Age                 Parch
##         -5.7402381841         -0.0632290123          0.9358844086
##                  Fare        Pclass2:Sexmale       Pclass3:Sexmale
##         -0.0202375871          1.9291867204          3.4074440173
##            Pclass2:Age           Pclass3:Age           Sexmale:Age
##          0.1112548242          0.0264806185          0.0751089101
##          Pclass2:Parch         Pclass3:Parch         Sexmale:Parch
##         -0.0518269916         -0.8587067646          1.8533166400
##             Age:Parch          Pclass2:Fare          Pclass3:Fare
##         -0.0028661212          0.1392750550         -0.0042745834
##          Sexmale:Fare              Age:Fare            Parch:Fare
##          0.0297825727          0.0012694419         -0.0179729371
##   Pclass2:Sexmale:Age   Pclass3:Sexmale:Age     Sexmale:Age:Parch
##            -0.0591783385            -0.1142225721            -0.1033991472
##     Pclass2:Sexmale:Fare     Pclass3:Sexmale:Fare         Pclass2:Age:Fare
##            -0.0659188060            -0.1209449471            -0.0049566499
##         Pclass3:Age:Fare         Sexmale:Age:Fare           Age:Parch:Fare
##            -0.0018727535            -0.0014306752             0.0004870245
## Pclass2:Sexmale:Age:Fare Pclass3:Sexmale:Age:Fare
##            -0.0010957744             0.0078794625
```

Todos los coeficientes tienen un efecto significativo en el modelo

4. Analiza las predicciones para los datos de entrenamiento

4.1 Elabora la matriz de confusión

```
library(vcd)

## Warning: package 'vcd' was built under R version 4.4.2

## Loading required package: grid

##
## Attaching package: 'vcd'

## The following object is masked from 'package:ISLR':
##
##     Hitters

predicciones <- ifelse(test = oModelo2$fitted.values > 0.5, yes = 1, no = 0)
M_C <- table(oModelo2$model$Survived, predicciones, dnn = c("observaciones",
"predicciones"))
M_C

##               predicciones
## observaciones   0   1
##             0 400  40
##             1  73 218

mosaic(M_C, shade = T, colorize = T,
       gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))
```

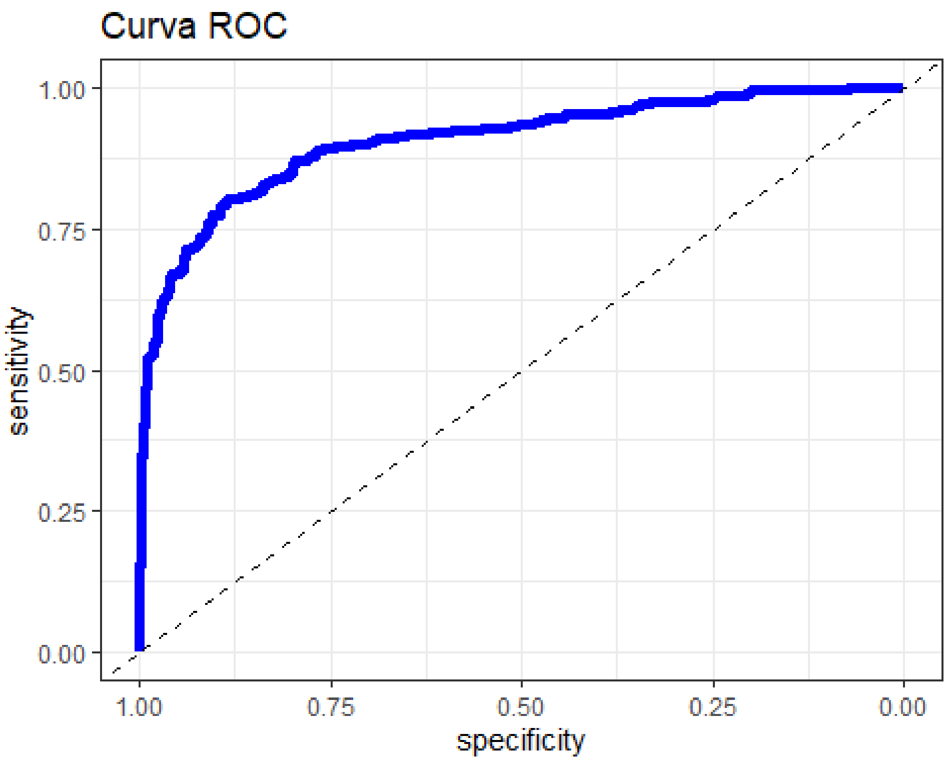


```
Ac = (M_C[1,1]+M_C[2,2])/sum(M_C)
cat("La Exactitud (accuracy) del modelo es", Ac,"\n")

## La Exactitud (accuracy) del modelo es 0.8454172

Se = M_C[1,1]/sum(M_C[1,])
cat("La Sensibilidad del modelo es", Se,"\n")

## La Sensibilidad del modelo es 0.9090909

Sp = M_C[2,2]/sum(M_C[2,])
cat("La Especificidad del modelo es", Sp,"\n")

## La Especificidad del modelo es 0.7491409

P = M_C[1,1]/sum(M_C[,1])
cat("La Precisión del modelo es", P,"\n")

## La Precisión del modelo es 0.845666
```

4.2 Elabora la Curva ROC

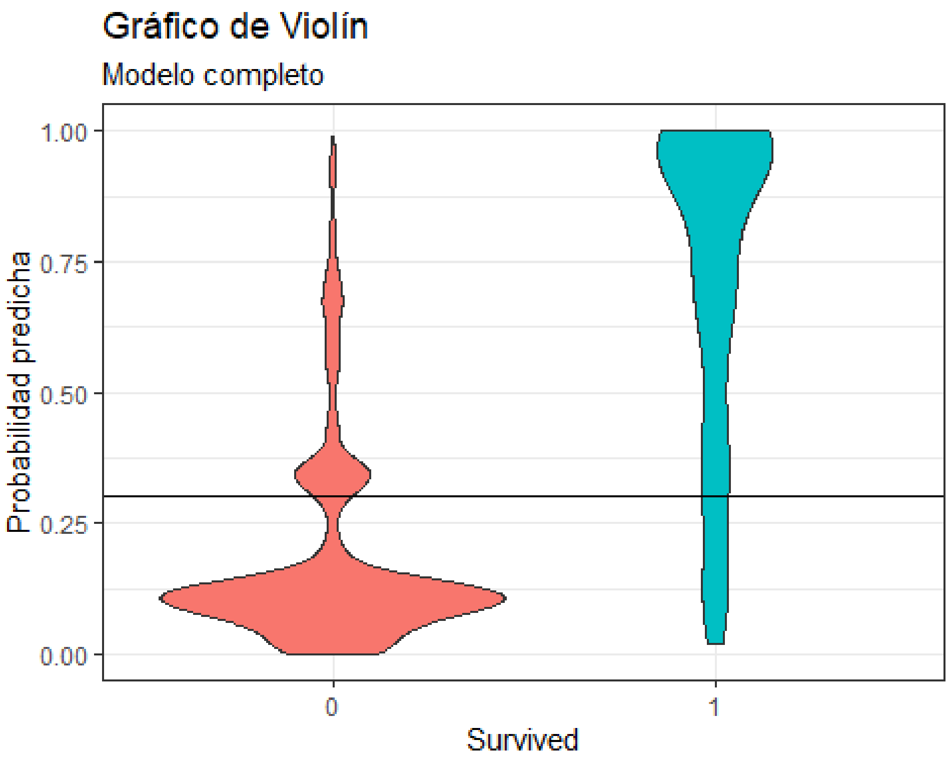
```
pred = predict(oModelo2, data = oTitanicTrainData, type = 'response')

library(pROC)

## Warning: package 'pROC' was built under R version 4.4.2

## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'

## The following object is masked from 'package:Metrics':
##
##     auc

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

ROC <- roc(response=oTitanicTrainData$Survived, predictor=pred)

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

## Setting direction: controls < cases

ROC

##
## Call:
## roc.default(response = oTitanicTrainData$Survived, predictor = pred)
##
## Data: pred in 440 controls (oTitanicTrainData$Survived 0) < 291 cases
(oTitanicTrainData$Survived 1).
## Area under the curve: 0.9073

ggroc(ROC, color = "blue", size = 2) + geom_abline(slope = 1, intercept = 1,
linetype ='dashed') + labs(title = "Curva ROC") + theme_bw()
```



4.3 Elabora el gráfico de violín

```
v_d = data.frame(Survived=oTitanicTrainData$Survived,pred=pred)

ggplot(data=v_d, aes(x=Survived, y=pred, group=Survived,
fill=factor(Survived))) +
  geom_violin() + geom_abline(aes(intercept=0.3,slope=0))+
  theme_bw() +
  guides(fill=FALSE) +
  labs(title='Gráfico de Violín', subtitle='Modelo completo', y='Probabilidad
predicha')

## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use
"none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



4.4 Concluye sobre el modelo basándote en las predicciones de los datos de entrenamiento.

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

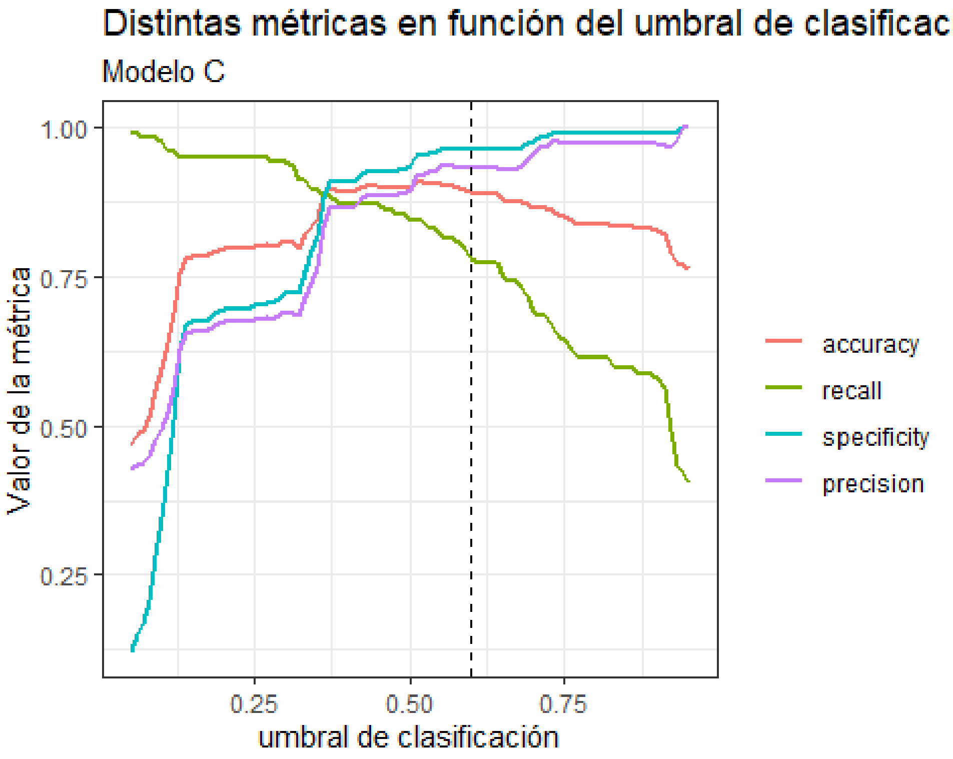
```
pred_val = predict(oModelo2, newdata=oTitanicTestData, type='response')
clase_real = oTitanicTestData$Survived

datosV = data.frame(accuracy=NA, recall=NA, specificity = NA, precision=NA)

for (i in 5:95){
  clase_predicha = ifelse(pred_val>i/100,1,0)

##Creamos la matriz de confusión
cm= table(clase_predicha,clase_real)

## AccurAcy: Proporción de correctamente predichos
datosV[i,1] = (cm[1,1]+cm[2,2])/(cm[1,1]+cm[1,2]+cm[2,1]+cm[2,2])
## Recall: Tasa de positivos correctamente predichos
datosV[i,2] = (cm[2,2])/(cm[1,2]+cm[2,2])
## Specificity: Tasa de negativos correctamente predichos
datosV[i,3] = cm[1,1]/(cm[1,1]+cm[2,1])
## Precision: Tasa de bien clasificados entre los clasificados como positivos
datosV[i,4] = cm[2,2]/(cm[2,1]+cm[2,2])
}
```

```r
## Se limpia el conjunto de datos
datosV = na.omit(datosV)
datosV$umbral = seq(0.05,0.95,0.01)

library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
datosV_m <- reshape2::melt(datosV,id.vars=c('umbral'))
colnames(datosV_m)[2] <- c('Metrica')

library(ggplot2)

u = 0.6 #Se dio un valor arbitrario, tú modificalo de acuerdo al criterio que selecciones.

ggplot(data=datosV_m, aes(x=umbral,y=value,color=Metrica)) +
geom_line(size=1) + theme_bw() +
  labs(title= 'Distintas métricas en función del umbral de clasificación',
       subtitle= 'Modelo C',
       color="", x = 'umbral de clasificación', y = 'Valor de la métrica') +
  geom_vline(xintercept=u, linetype="dashed", color = "black")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

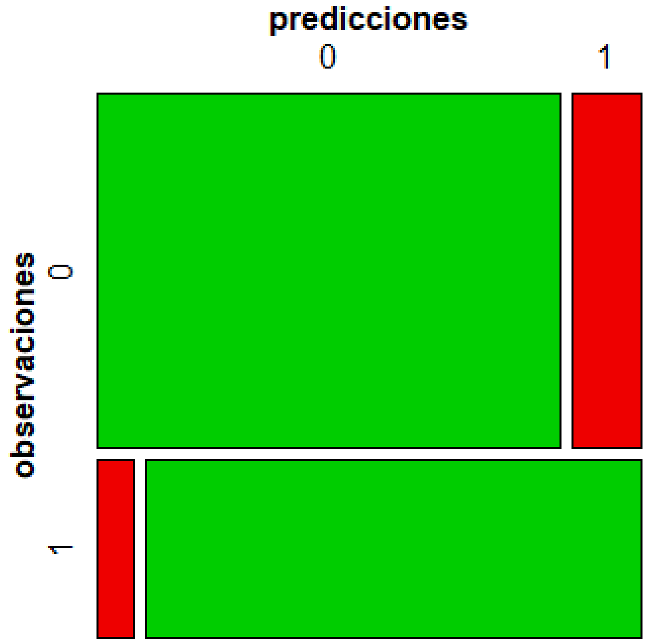


5.1 Elije un umbral de clasificación óptimo

El umbral de clasificacion óptimo que considere fue un umbral de 0.6 ya que es un valor que viendo la grafica las variables aun no tienen un gran aumento o un gran descenso.

5.2 Elabora la matriz de confusión con el umbral de clasificación óptimo

```
prediccionesV = ifelse(pred_val > 0.6, yes = 1, no = 0)
M_Cv <- table(prediccionesV, oTitanicTestData$Survived, dnn =
c("observaciones", "predicciones"))
M_Cv

##                predicciones
## observaciones   0   1
##              0 181  27
##              1   7  97

mosaic(M_Cv, shade = T, colorize = T,
       gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))
```



```
AcV = (M_Cv[1,1]+M_Cv[2,2])/sum(M_Cv)
cat("La Exactitud (accuracy) del modelo es", AcV,"\n")

## La Exactitud (accuracy) del modelo es 0.8910256

SeV = M_Cv[1,1]/sum(M_Cv[1,])
cat("La Sensibilidad del modelo es", SeV,"\n")

## La Sensibilidad del modelo es 0.8701923

SpV = M_Cv[2,2]/sum(M_Cv[2,])
cat("La Especificidad del modelo es", SpV,"\n")

## La Especificidad del modelo es 0.9326923

PV = M_Cv[1,1]/sum(M_Cv[,1])
cat("La Precisión del modelo es", PV,"\n")

## La Precisión del modelo es 0.962766
```

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

```
oTitanicTest =
read.csv("C:\\Users\\eliez\\OneDrive\\Desktop\\Clases\\Titanic_test.csv")
#leer la base de datos
oTitanicTest <- oTitanicTest[,c(-3,-8,-10)]
oTitanicTest = na.omit(oTitanicTest)
for(var in c('Pclass','Embarked','Sex'))
```

```
  oTitanicTest[,var] <-as.factor(oTitanicTest[,var])


pred_val = predict(oModelo2, newdata=oTitanicTest, type='response')

prediccionesF = ifelse(pred_val > 0.6, yes = 1, no = 0)
M_Cv <- table(prediccionesF, dnn = c( "predicciones"))
M_Cv

## predicciones
##   0   1
## 216 115
```

7. Concluye en el contexto del problema:

7.1 Define las principales características que influyen en el modelo seleccionado e interpretalas: ¿qué características tuvieron las personas que sobrevivieron?

Las principales caracteristicas fueron el genero de la persona, su clase social y su edad entre las principales, de ahi se deriva tambien el numero de padres o hijos abordo y el costo del ticket

7.2 Interpreta los coeficientes del modelo

Los coeficientes del modelo se separaran en las variables principales que comente arriba se factorizaron, ademas la mayoria de coeficientes que se usan son las intersecciones de las variables, esto le da un mejor enfoque al modelo ya que obtienes las intersecciones entre todas las variables que hace que el modelo mejore, ademas todos los coeficientes son relevantes ya que el valor P de estos pasa el umbral de 0.05 y muchos son muy cercanos a ser mayores de 1.

7.3 Define cuál es el mejor umbral de clasificación y por qué

El mejor umbral de clasificación fue de 0.55 ya que es un valor que hace que las variables no empiecen a disminuir y se mantenga dentro del valor de la metrica de 0.7