Proyecto - Tarea 2
Data Mining
Término 2018-2S
Escuela Superior Politécnica del Litoral (ESPOL)

Profesor: PhD.Carmen Vaca
Nombre: María Belén Guaranda Cabezas
Fecha: 2018/12/02
Paralelo: 1

Tema 1: regresión logística

Poseemos un dataset sobre características de diagnósticos de cáncer, y tenemos que predecir si, en base a ellas, el tumor es maligno o benigno. Se nos pide realizar una regresión logística de la forma:

$$p(x) = P(Y = \mathrm{M} \mid X = x).$$

Figs.1: Modelo de regresión logística.

Se realiza el preprocesamiento y la partición de datos, 80% de entrenamiento y 20% de prueba, con ayuda de la librería *caret*.

```
cancerData <- preProcess(x = cancer_df, method = c("center", "scale"))
cancerData <- predict(cancerData, cancer_df) #stored

set.seed(1) #for reproductibility of results
trainIndex <- createDataPartition(cancerData$tipo,
                                  times = 1,
                                  p = .80,
                                  list = FALSE)
```

Fig.1:Preprocesamiento y partición de los datos.

Considerando 2 características, radio y simetría, se realiza la regresión lineal:

```
> model <- glm(tipo2 ~ radio +simetria,
+              family = binomial(link = "logit"),
+              data = cancerTrain)
> summary(model)

Call:
glm(formula = tipo2 ~ radio + simetria, family = binomial(link = "logit"),
    data = cancerTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2391  -0.3413  -0.1294   0.1050   2.8439

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.7702     0.1981  -3.888 0.000101 ***
radio         4.0195     0.4556   8.823  < 2e-16 ***
simetria      1.3710     0.2343   5.852 4.85e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 490.95  on 375  degrees of freedom
Residual deviance: 177.44  on 373  degrees of freedom
AIC: 183.44

Number of Fisher Scoring iterations: 7
```

Figs.3: regresión lineal para el dataset de diagnósticos de cáncer.

Una vez obtenido nuestro modelo, procederemos a evaluarlo. Trataremos a 0 como Benigno y 1 como Maligno.
Se reportan los siguientes valores de specificity, sensitivity y accuracy:

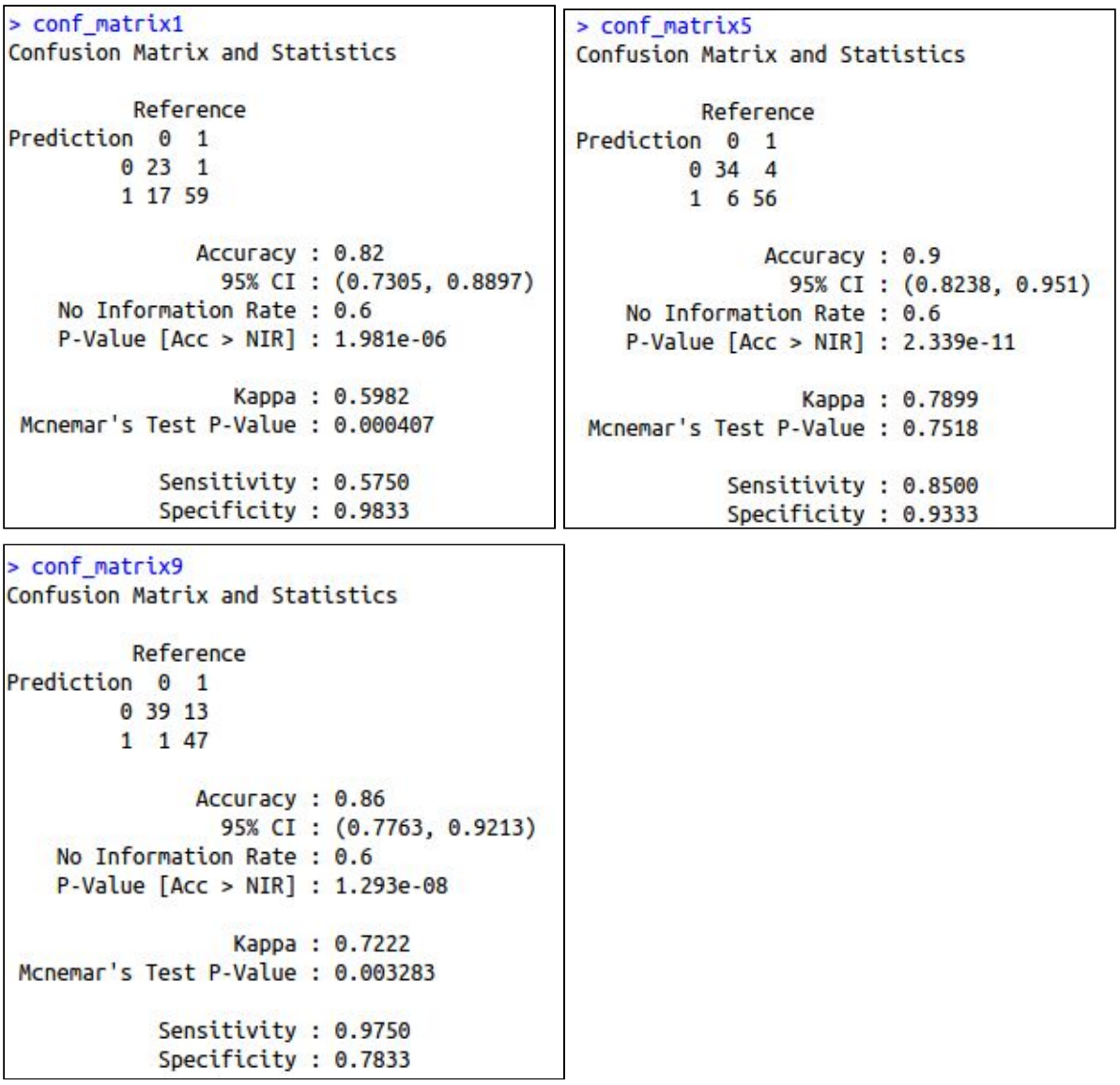


Fig.4: (De izq a der, de arriba abajo) Valores de sensitivity, specificity y accuracy, para valores de corte de 0.1, 0.5 y 0.9 respectivamente, en fase de testeo.

Los errores de entrenamiento son:

```
> accuracy(cancerTrain$prediction,cancerTrain$tipo)
                    ME      RMSE       MAE  MPE MAPE
Test set 2.306339e-15 0.2700522 0.142635 -Inf  Inf
```

Fig.5: Errores del set de entrenamiento, para los distintos valores de cut-off.

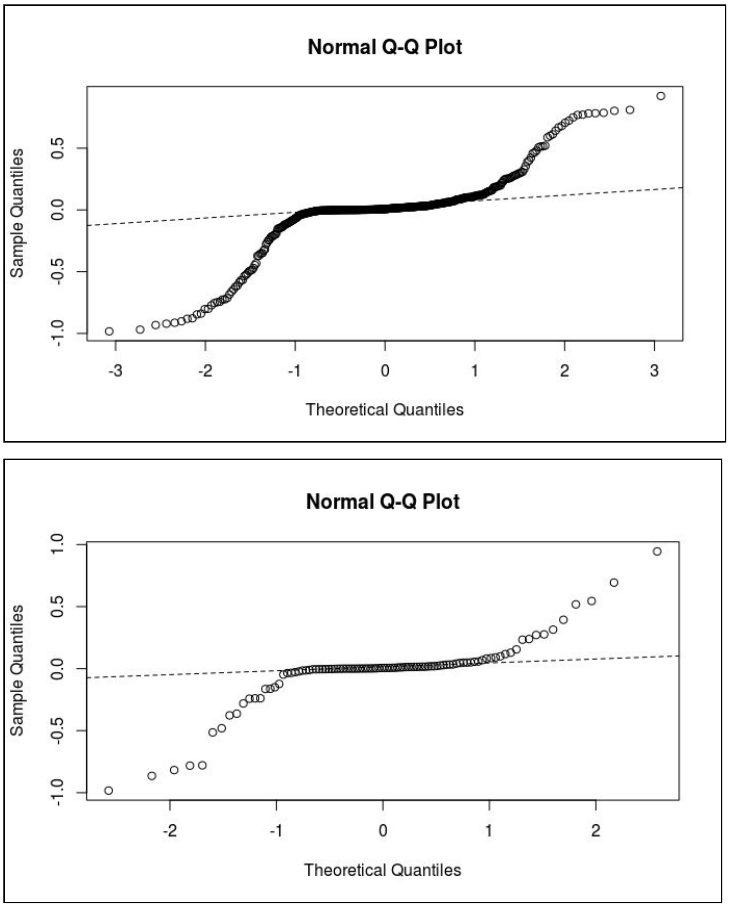
Y de testeo:

```
> accuracy(cancerTest$prediction,cancerTest$tipo)
                    ME      RMSE       MAE  MPE MAPE
Test set -0.01787388 0.2699853 0.1383093 -Inf  Inf
```

Fig.6: Errores del set de testeo, para los distintos valores de cut-off.

Para nuestro caso de estudio, se debería buscar que el modelo disminuya su error tipo 2, es decir, lo que etiqueta como benigno(0) cuando en realidad es maligno(1). Sino, estaríamos poniendo en riesgo la vida de personas, ya que les decimos que están sanas cuando en

realidad no lo están. Entonces, lo que debería de tratar de bajar es la cantidad de falsos negativos, que están asociados con un aumento en sensitivity. Por lo tanto, debería quedarme con el valor de cutoff de 0.9, que tiene el mayor valor de sensitivity. Con respecto a los errores, se procede a graficar su distribución:



Figs.7 y 8: Distribución de los errores de entrenamiento y testeo (de arriba abajo, respectivamente).

Tienen distribuciones similares, pero no son normales (para serlo, deberían de acercarse a una recta lineal), lo cual no es bueno para un modelo.

Tema 2: cutoff values y curva ROC

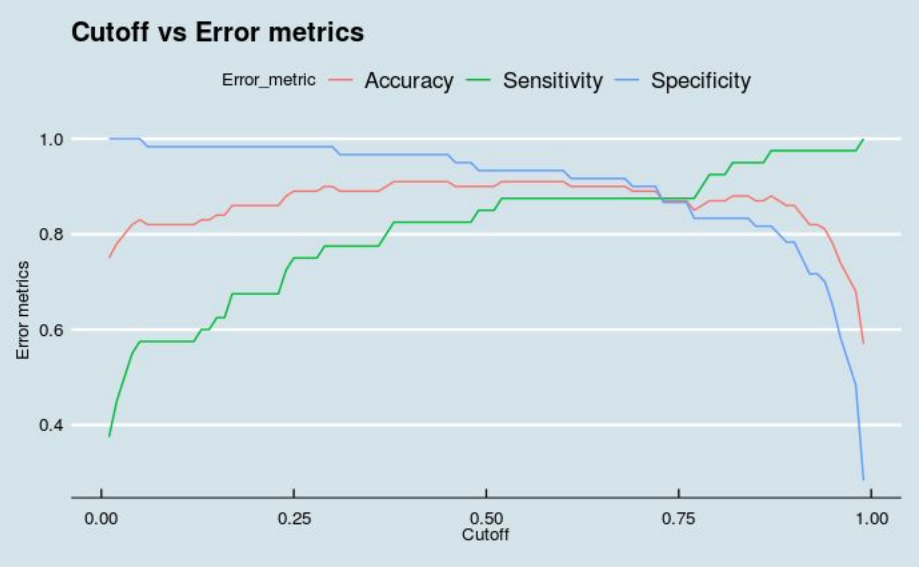
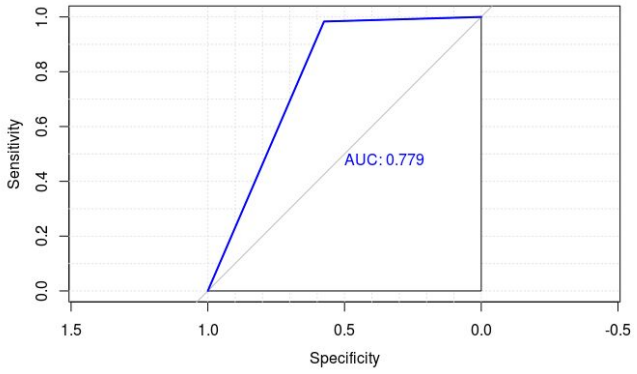


Fig.9: Curvas de accuracy, sensitivity y specificity, para valores de cutoff entre (0.01,0.99).

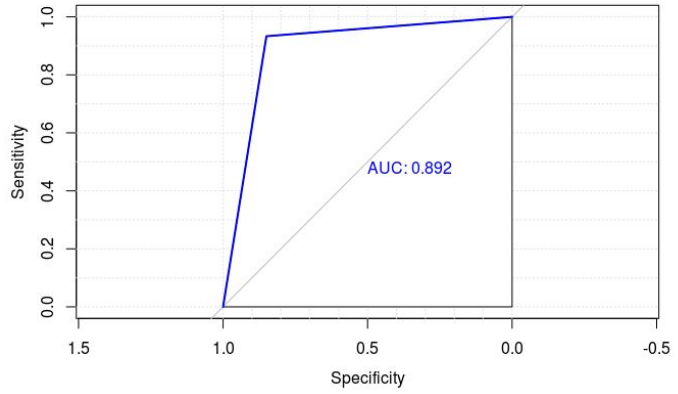
Esta visualización nos permite apreciar el comportamiento de las métricas de rendimiento: sensitivity, specificity y accuracy. Se puede apreciar que, sólo sensitivity muestra un comportamiento creciente, a medida que incrementa el cutoff; las otras 2 curvas en cambio, decrecen. Hay un valor de cutoff para el cual todas las métricas se intersecta, aproximadamente 0.75. Para nuestro caso de estudio, como se dijo anteriormente, se busca incrementar el valor de sensitivity, por lo que nos interesaría un cutoff value de 1. Ahora, si escogemos este valor, estamos renunciando a buen accuracy y specificity. Tampoco nos debe de dejar de importar la cantidad de falsos positivos que nuestro modelo genere, porque también le estaríamos haciendo pasar un mal rato a personas que estando sanas, piensen que tienen cáncer. Entonces, podríamos optar por un cutoff value muy cercano y mayor a 0.75.

Curvas ROC

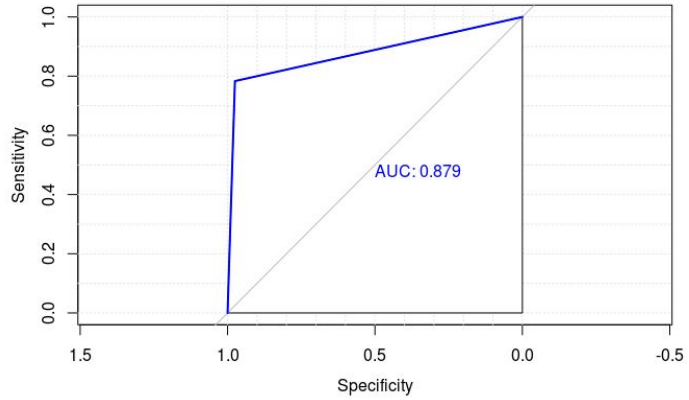
Cutoff value = 0.1



Cutoff value = 0.5



Cutoff value = 0.9



Figs.10,11 y 12: Curvas ROC para los distintos valores de cutoff.

El AUC nos permite ver cuán preciso es nuestro modelo. Para el valor de cutoff de 0.5, parecería que nuestro modelo tiene el mejor rendimiento. Sin embargo, la AUC no es una buena métrica si es que busco minimizar un tipo de error en específico, en este caso el tipo 2, por lo que no es de mucha ayuda. Tal y como vimos en la primera gráfica, nos conviene más escoger un valor cercano y mayor a 0.75, para maximizar la sensitividad, y no un valor de 0.5, como nos indica el AUC.

Tema 3

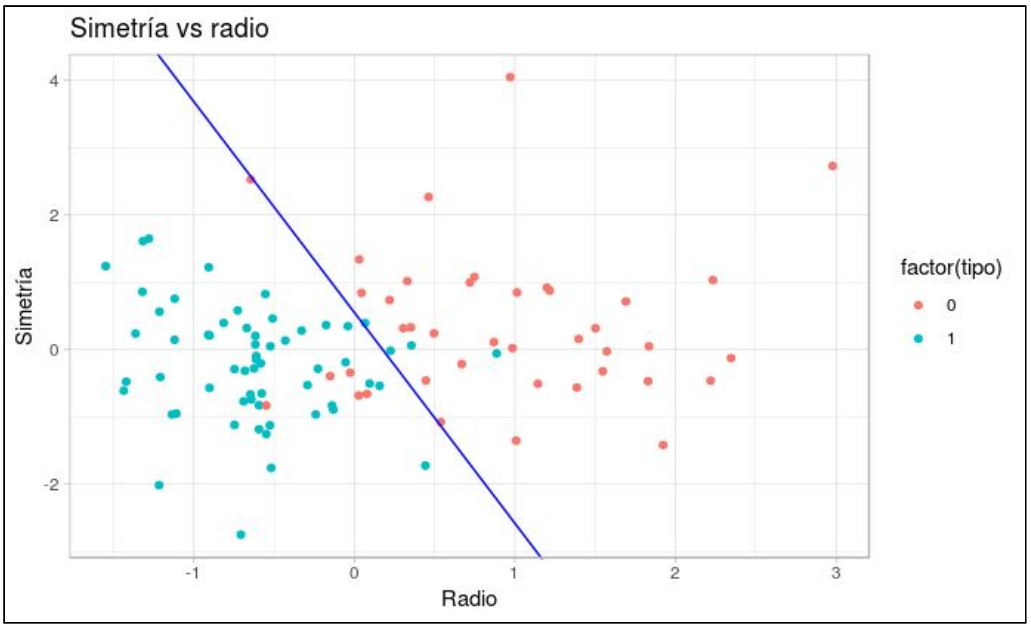


Fig. 13: Simetría vs radio del tumor, de diagnśoticos de cáncer. 0 es benigno y 1 es maligno.

A partir de la gráfica podemos concluir que, para un valor de cutoff de 0.5, predecimos la mayoría de los diagnósticos benignos, pero no malignos.

Tema 4: pasajeros del Titanic

En este dataset contamos con 15 columnas, de las cuales excluimos para nuestro análisis, las siguientes:

- Cabin: tiene mucha sparsity.
- PassengerId, Ticket, Name : no contribuyen con el análisis, pues estos features no dicen o no podrían decir mucho sobre una persona y si sobrevivió o no.

Nuestra columna outcome es binaria, siendo 1 si el pasajero sobrevivió, y 0 si no sobrevivió.

Distribución de los datos
Observamos la distribución de la columna *Age*

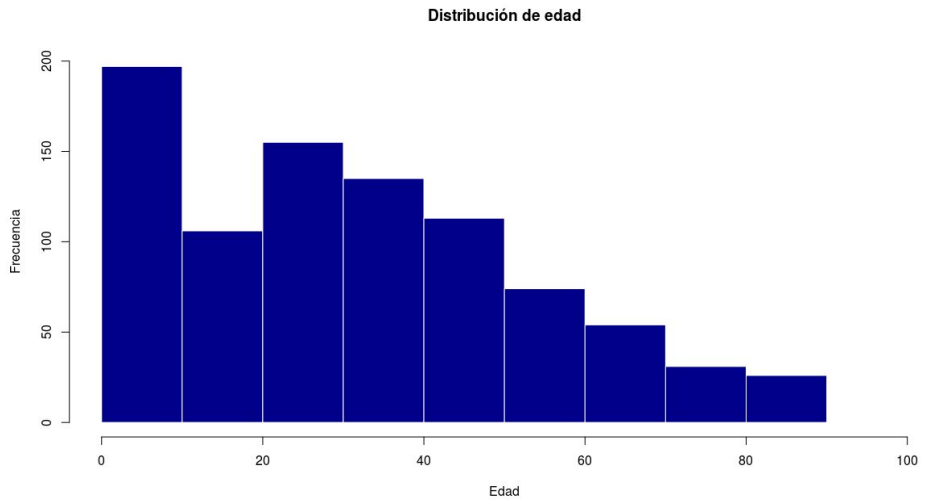


Fig. 14:Distribución de la edad de los pasajeros.

La curva tiene su pico del lado izquierdo. Si usamos la media para imputar los datos, tal vez introduzcamos un bias, incrementando pasajeros con edades mayores. Usaremos mejor la mediana para la imputación.

```
Embarked
 :  2
C:168
Q: 77
S:644
```

Fig. 15:Distribución del lugar de embarque

Para *Embarked,* al ser S el valor más común, y como son pocos los valores desconocidos, les asignaremos también el valor “S”.
Para nuestro modelo de regresión logística, usamos un feature obtenido a partir de Name, que lo nombramos *Position.* En esta columna almacenamos los títulos de las personas, por ejemplo, Miss, Ms, etc. Usamos todos los features restantes del dataset.

vs Clasificador basado en reglas

A continuación mostramos las matrices de confusión, tanto del clasificador basado en reglas, como el de la regresión logística.

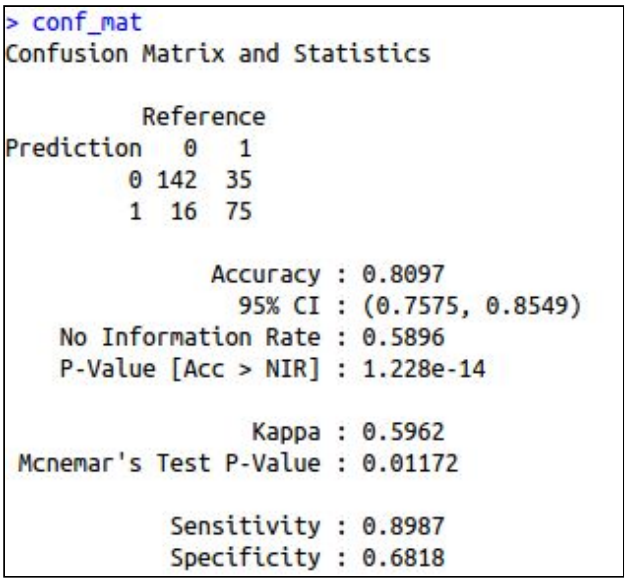


Fig. 14: Matriz de confusión de clasificador basado en reglas.

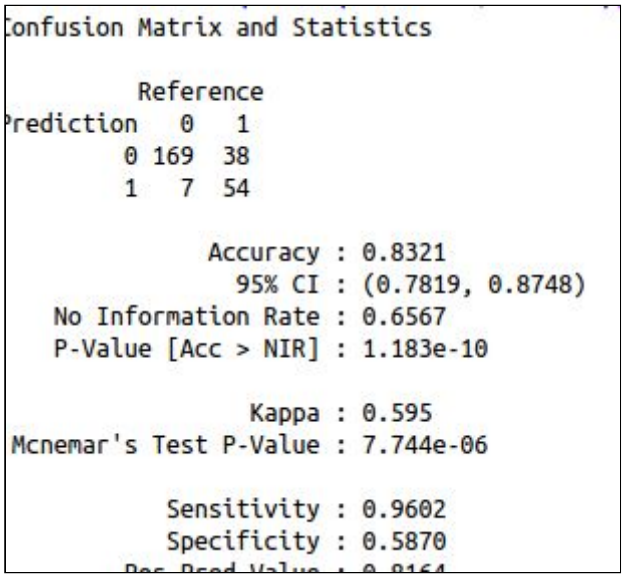


Fig. 15: Matriz de confusión de regresión logística.

Con la regresión, se obtiene un 3% más de accuracy, sensitivity aumenta en un 7% pero specificity disminuye en casi un 10%. En este caso en particular, nos interesa más reconocer a los que no sobrevivirán, puesto que en el clasificador basado en reglas, con sólo decir que las mujeres sobreviven ya se obtiene un 70% de accuracy. Entonces, lo que se busca maximizar son los true negatives, y por tanto, aumentar specificity. En conclusión, aunque el modelo de regresión logística mejora en accuracy, podríamos hacer un trade off y quedarnos con el de basado en reglas, por su mejor especificidad.

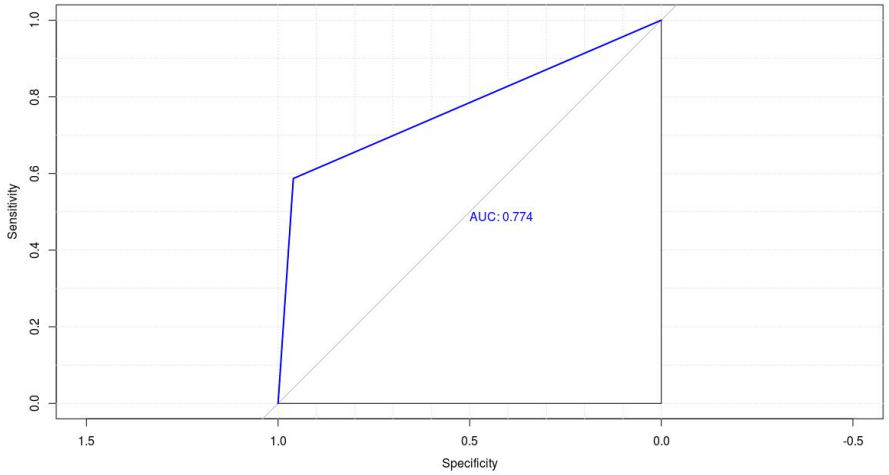


Fig. 16: Curva ROC del modelo de regresión logística.

Recursos:
https://data.library.virginia.edu/understanding-q-q-plots/
https://stats.stackexchange.com/questions/3136/how-to-perform-a-test-using-r-to-see-if-data-follows-normal-distribution

https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419
http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization