Pontificia Universidad Católica de Chile

FORO Nº 2

Modelamiento Estadístico y Sistemas Recomendadores

"Métodos de Clasificación"

Eduardo Andrés Carrasco Vidal
Ingeniero Civil Industrial

Actividad Nº 2:

En esta actividad usted deberá aportar con 1 respuesta principal a la o las preguntas enunciadas por el profesor y comentar y/o intervenir 2 respuestas de sus compañeros (en otros hilos de conversación).

Además, junto a la respuesta principal, usted deberá subir un informe del trabajo realizado, el cual también será considerado en la evaluación del foro.

A. Considere los datos 'salary.csv', que contiene información recabada en una empresa, sobre potenciales clientes y cuya descripción se muestra en la siguiente tabla:

Variable	Descripción
trabajo	Tipo de trabajo
educacion	Nivel educacional
estado civil	Estado civil
ocupacion	Ocupación
familiar	Rol en su hogar
raza	Raza
sexo	Sexo
pais	País de origen
clase	Indica si tiene un sueldo anual mayor a US$ 50.000

1) Cargue el conjunto de datos en la sesión de trabajo R, usando la función *read.table()*:

```
> datos <- read.table(file.choose(),header=TRUE, sep=";")
> View(datos)
> summary(datos)
              trabajo          educacion                estado.civil                ocupacion
 Private          :6760   HS-grad      :3079   Divorced              :1305   Prof-specialty     :1225
 Self-emp-not-inc: 762   Some-college:2229   Married-AF-spouse    :   8    Exec-managerial.  :1198
 Local-gov        : 631   Bachelors    :1620   Married-civ-spouse   :4489   Craft-repair       :1191
 ?                : 562   Masters      : 558   Married-spouse-absent: 130   Adm-clerical       :1115
 State-gov        : 405   Assoc-voc    : 436   Never-married        :3233   Sales              :1110
 Self-emp-inc     : 357   11th         : 359   Separated            : 299   Other-service      : 984
 (Other)          : 291   (Other)      :1487   Widowed              : 304   (Other)            :2945
          familiar                 raza                    sexo            pais              clase
 Husband       :3952   Amer-Indian-Eskimo:  91   Female:3202   United-States :8812   <=50K:7439
 Not-in-family  :2539   Asian-Pac-Islander : 280   Male  :6566   Mexico        : 202    >50K :2329
 Other-relative : 293   Black              : 938                 ?             : 154
 Own-child     :1506   Other              :  93                 Philippines   :  51
 Unmarried     :1005   White              :8366                 Germany       :  38
 Wife          : 473                                            Canada        :  37
                                                                (Other)       : 474
```

2) Seleccione de manera aleatoria 2/3 de los datos para crear sus datos de entrenamiento y guarde el tercio restante para objeto de validación. Para esto, simule 9.768 valores 1 y 2 en proporciones 2/3 a 1/3 a través de la función *sample()*. Utilice aquellas tuplas de la base de datos asociadas al valor 1 para la base de entrenamiento y las restantes para validación. Utilice la semilla 1, mediante el comando *set.seed(1)*:

```
set.seed(1)
> View(datos)
> ind <- sample(2, length(datos$clase), replace=TRUE, prob=c(2/3, 1/3))
> View(ind)
> table(ind)
ind
   1    2
6470 3298
> datos.trabajo <- datos[ind==1,]
> datos.validacion <- datos[ind==2,]
> length(datos.trabajo)
[1] 9
> length(datos.validacion)
[1] 9
```

De igual manera, podemos obtener un resumen de las variables datos.trabajo y datos.validación de acuerdo al siguiente detalle:

```
> summary(datos.trabajo)
            trabajo            educacion                    estado.civil            ocupacion
 Private         :4486   HS-grad     :2039   Divorced             : 857   Prof-specialty : 804
 Self-emp-not-inc: 498   Some-college:1495   Married-AF-spouse    :   5   Craft-repair   : 801
 Local-gov       : 421   Bachelors   :1058   Married-civ-spouse   :2996   Exec-managerial: 795
 ?               : 363   Masters     : 380   Married-spouse-absent:  84   Adm-clerical   : 729
 State-gov       : 282   Assoc-voc   : 295   Never-married        :2149   Sales          : 720
 Self-emp-inc    : 236   11th        : 225   Separated            : 194   Other-service  : 687
 (Other)         : 184   (Other)     : 978   Widowed              : 185   (Other)        :1934
          familiar                    raza          sexo                 pais        clase
 Husband       :2633   Amer-Indian-Eskimo:  67   Female:2119   United-States:5866   <=50K:4956
 Not-in-family :1659   Asian-Pac-Islander: 179   Male  :4351   Mexico       : 127   >50K :1514
 Other-relative: 187   Black             : 614                 ?            : 102
 Own-child     :1017   Other             :  62                 Philippines  :  33
 Unmarried     : 657   White             :5548                 Germany      :  21
 Wife          : 317                                           India        :  21
                                                               (Other)      : 300
```

```
> summary(datos.validacion)
            trabajo            educacion                   estado.civil            ocupacion
 Private         :2274   HS-grad     :1040   Divorced              : 448   Prof-specialty : 421
 Self-emp-not-inc: 264   Some-college: 734   Married-AF-spouse     :   3   Exec-managerial: 403
 Local-gov       : 210   Bachelors    : 562   Married-civ-spouse    :1493   Craft-repair   : 390
 ?               : 199   Masters      : 178   Married-spouse-absent:  46   Sales          : 390
 State-gov       : 123   Assoc-voc    : 141   Never-married         :1084   Adm-clerical   : 386
 Self-emp-inc    : 121   11th         : 134   Separated             : 105   Other-service  : 297
 (Other)         : 107   (Other)      : 509   Widowed               : 119   (Other)        :1011
          familiar                    raza          sexo                 pais        clase
 Husband       :1319   Amer-Indian-Eskimo:  24   Female:1083   United-States:2946   <=50K:2483
 Not-in-family : 880   Asian-Pac-Islander: 101   Male  :2215   Mexico       :  75   >50K : 815
 Other-relative: 106   Black             : 324                 ?            :  52
 Own-child     : 489   Other             :  31                 El-Salvador  :  18
 Unmarried     : 348   White             :2818                 Philippines  :  18
 Wife          : 156                                           Canada       :  17
                                                               (Other)      : 172
```

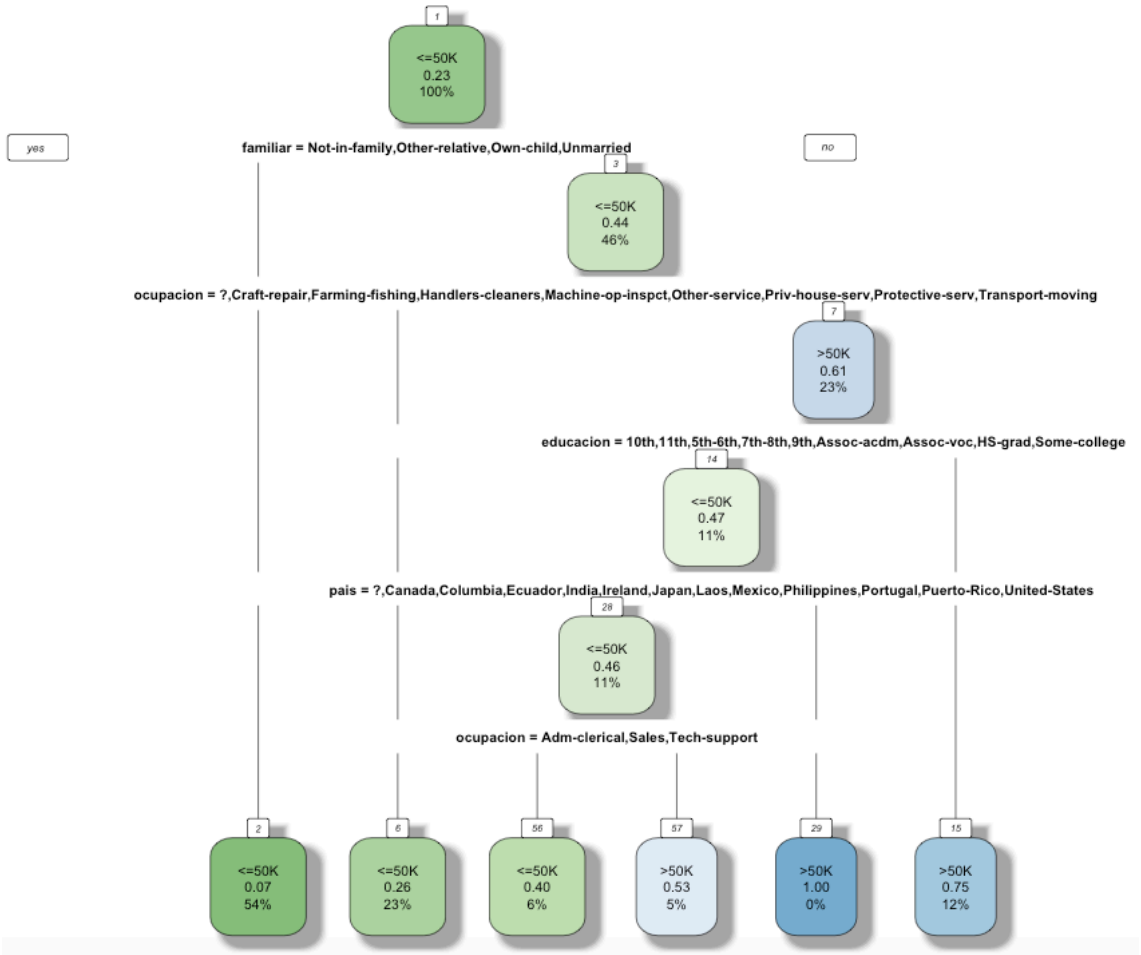
3) Construya un árbol de decisión para la variable clase, utilizando como criterio índice de Gini. Realice el procedimiento completo, incluyendo la poda del árbol, usando los comandos *rpart()*, *cptable()* y *prune()* de la librería *rpart*. Use la opción *cp=0.012* en el proceso de poda.

```
> library(rpart)
> library(rpart.plot)
> fit <- rpart(clase ~ .,data=datos.trabajo,parms = list(split = "gini"))
> print(fit)
n= 6470

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 6470 1514 <=50K (0.76599691 0.23400309)
 2) familiar=Not-in-family,Other-relative,Own-child,Unmarried 3520  229 <=50K (0.93494318
0.06505682) *
 3) familiar=Husband,Wife 2950 1285 <=50K (0.56440678 0.43559322)
 6) ocupacion=?,Craft-repair,Farming-fishing,Handlers-cleaners,Machine-op-inspct,Other-
service,Priv-house-serv,Protective-serv,Transport-moving 1469  383 <=50K (0.73927842
0.26072158) *
 7) ocupacion=Adm-clerical,Armed-Forces,Exec-managerial,Prof-specialty,Sales,Tech-support 1481
579 >50K (0.39095206 0.60904794)
14) educacion=10th,11th,5th-6th,7th-8th,9th,Assoc-acdm,Assoc-voc,HS-grad,Some-college 736
345 <=50K (0.53125000 0.46875000)
28)pais=?,Canada,Columbia,Ecuador,India,Ireland,Japan,Laos,Mexico,Philippines,Portugal,Puerto-
Rico,United-States 722  331 <=50K (0.54155125 0.45844875)
56) ocupacion=Adm-clerical,Sales,Tech-support 407  163 <=50K (0.59950860 0.40049140) *
57) ocupacion=Exec-managerial,Prof-specialty 315  147 >50K (0.46666667 0.53333333) *
29) pais=Cuba,Dominican-Republic,England,Germany,Haiti,Italy,Jamaica,Nicaragua 14    0 >50K
(0.00000000 1.00000000) *
15) educacion=12th,1st-4th,Bachelors,Doctorate,Masters,Prof-school 745  188 >50K (0.25234899
0.74765101) *
```

Al cargar las dos librerías y ajustar el modelo mediante la función rpart(), se observa la columna clase, utilizando los datos de entrenamiento y el parámetro de gini como criterio.
Además, como se muestra en la siguiente figura, se generaron los árboles de decisión correspondiente:



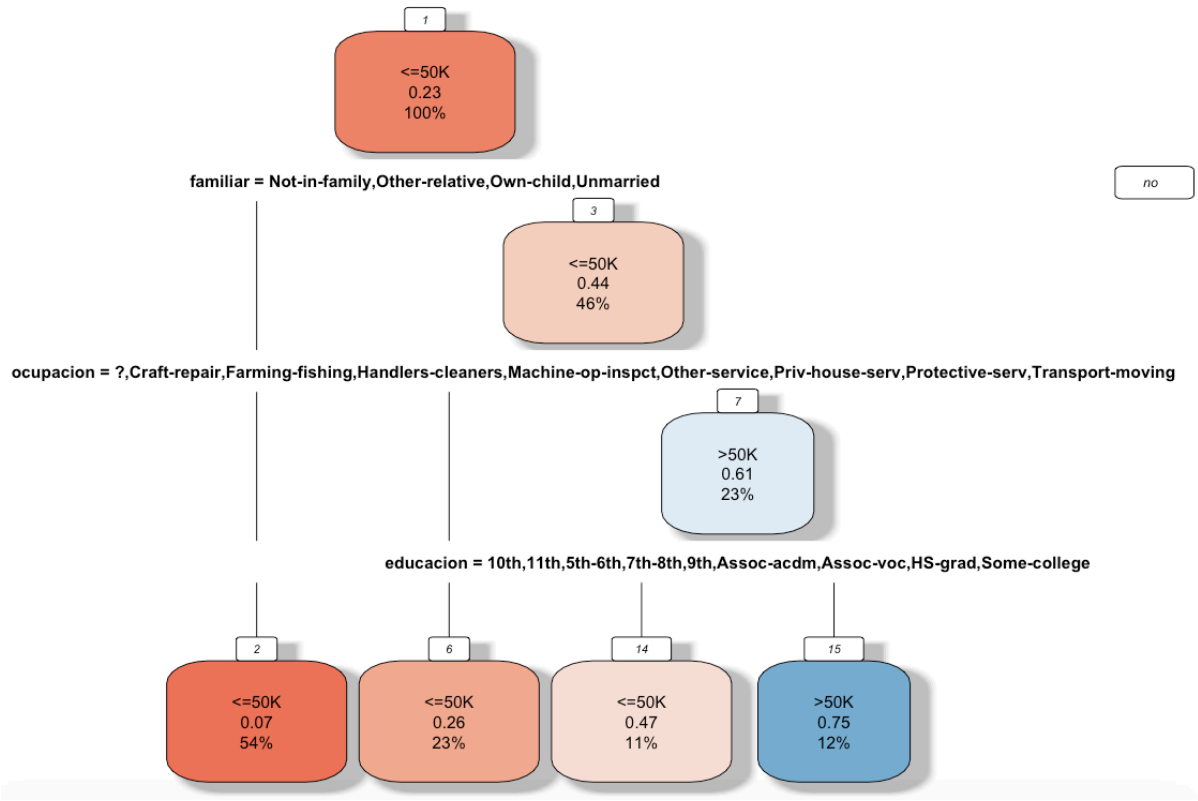
Como se observa en la figura, el árbol de decisión muestra las posibles clases de clasificación correspondientes a partir de los nodos internos.

```
> pfit <- prune(fit, cp = 0.012)
> print(pfit)
n= 6470

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 6470 1514 <=50K (0.76599691 0.23400309)
   2) familiar=Not-in-family,Other-relative,Own-child,Unmarried 3520  229 <=50K (0.93494318 0.06505682) *
   3) familiar=Husband,Wife 2950 1285 <=50K (0.56440678 0.43559322)
     6) ocupacion=?,Craft-repair,Farming-fishing,Handlers-cleaners,Machine-op-inspct,Other-service,Priv-house-serv,Protec
tive-serv,Transport-moving 1469  383 <=50K (0.73927842 0.26072158) *
     7) ocupacion=Adm-clerical,Armed-Forces,Exec-managerial,Prof-specialty,Sales,Tech-support 1481  579 >50K (0.39095206
 0.60904794)
      14) educacion=10th,11th,5th-6th,7th-8th,9th,Assoc-acdm,Assoc-voc,HS-grad,Some-college 736  345 <=50K (0.53125000 0.
46875000) *
      15) educacion=12th,1st-4th,Bachelors,Doctorate,Masters,Prof-school 745  188 >50K (0.25234899 0.74765101) *
> rpart.plot(pfit,box.palette="RdBu",shadow.col="gray",nn=TRUE)
```

Como lo determina el enunciado y utilizando la función prune(), se realiza la poda del árbol de decisión, con un parámetro de complejidad cp=0.012, según se muestra en la siguiente figura:



4) Utilizando la función *navieBayes()* de la librería *e0171* construya un clasificador de Bayes ingenuo. Utilice la corrección de Laplace en caso de haber celdas vacías, *laplace=1*.

```
> library(e1071)
> fitNB <- naiveBayes(clase ~ .,data=datos.trabajo,laplace = 1)


> pred.tree <- predict(pfit,datos.validacion[,-9])
> head(pred.tree)
      <=50K       >50K
4  0.7392784 0.26072158
6  0.9349432 0.06505682
7  0.7392784 0.26072158
13 0.9349432 0.06505682
15 0.9349432 0.06505682
17 0.7392784 0.26072158

> pred.nb <- predict(fitNB,datos.validacion[,-9],type="raw")
> head(pred.nb)
        <=50K         >50K
[1,] 0.8001003 0.1998997143
[2,] 0.9913759 0.0086241358
[3,] 0.3636414 0.6363586433
[4,] 0.9995616 0.0004384475
[5,] 0.9962679 0.0037321390
[6,] 0.4308593 0.5691407357
```

5) Utilizando los datos de validación, calcule:

a) La **sensibilidad** en el proceso de clasificación, definida como el porcentaje de personas para las que el modelo predice un sueldo mayor a US $ 50.000, dentro de todas aquellas que en realidad tienen un sueldo mayor a dicha cantidad.

```
> mmetric(datos.validacion[,9],pred.tree,"TPR") # Sensibilidad
[1] 96.25453 39.14110
> mmetric(datos.validacion[,9],pred.nb,"TPR") # Sensibilidad
[1] 82.44060 72.76074
```

b) La **especificidad** del procedimiento de clasificación, definida como el porcentaje de personas para las que el modelo predice un sueldo anual menor o igual a US $50.000, dentro de todas aquellas que en realidad tienen un sueldo menor o igual a dicha cantidad.

```
> mmetric(datos.validacion[,9],pred.tree,"TNR") # Especificidad
[1] 39.14110 96.25453
> mmetric(datos.validacion[,9],pred.nb,"TNR") # Especificidad
[1] 72.76074 82.44060
```

c) La **precisión** del procedimiento de clasificación, definida como el porcentaje de personas clasificadas correctamente.

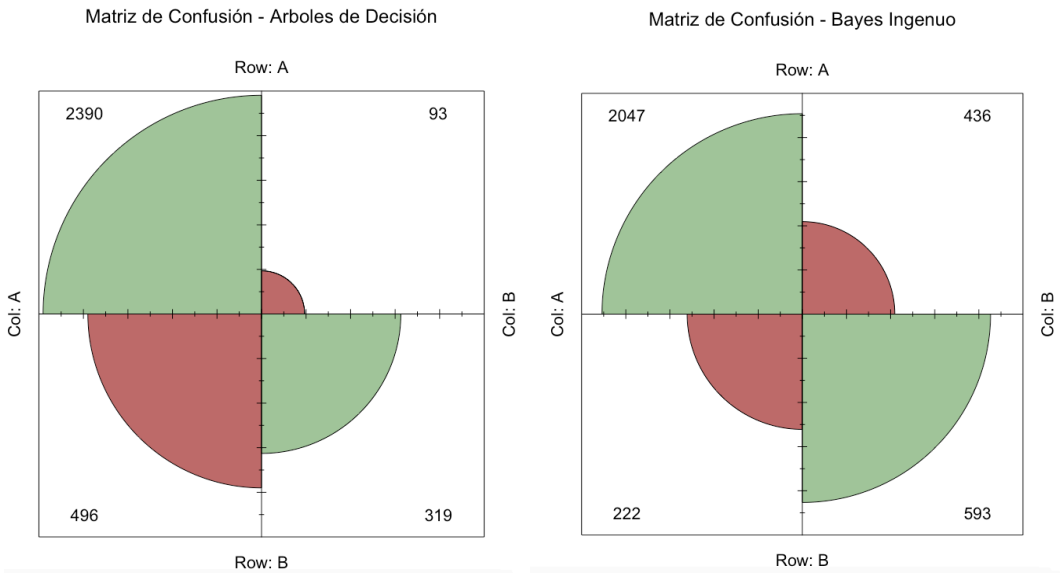
```
> mmetric(datos.validacion[,9],pred.tree,"ACC") # Precision
[1] 82.14069
> mmetric(datos.validacion[,9],pred.nb,"ACC") # Precision
[1] 80.04851
```

d) Además de las curvas anteriormente descritas, podemos de igual manera, calcular el AUC ROC y determinar la matríz de confusión de cada una de las variables predictivas:

```
> print(mmetric(datos.validacion[,9],pred.tree,"CONF")) #MATRIZ DE confusión
$res
NULL
$conf
              pred
target     <=50K   >50K
  <=50K    2390    93
  >50K      496     319
$roc
NULL
$lift
NULL
> print(mmetric(datos.validacion[,9],pred.nb,"CONF")) #MATRIZ DE confusión
$res
NULL
$conf
               pred
target      <=50K  >50K
  <=50K     2047   436
  >50K       222   593
$roc
NULL
$lift
NULL
> print(mmetric(datos.validacion[,9],pred.tree,"AUC")) #MATRIZ DE confusión
[1] 0.8161422
> print(mmetric(datos.validacion[,9],pred.nb,"AUC")) #MATRIZ DE confusión
[1] 0.8696026
```

6) Discuta los resultados con sus compañeros **mediante el foro**.

Efectuada la clasificación en el DataBase *salary*, del total (9.768 registros), efectuada la división en datos de entrenamiento 2/3 (Training Set)=6.470 y 1/3 para validación (Test Set)=3.298, se contruyeron dos modelos (clasificadores) diferentes: uno de árboles de decisión (*tree*) y otro de bayes ingenuo (*nb*).
En ambos modelos, lo que se busca es mejor efectividad, mejor *performance*, por lo cual, inicialmente podemos generar una matriz de confusión para cada modelo, objeto poder observar de mejor manera los indicadores solicitados, como se muestra en la figura:



Analizada la matriz de confusión en base a los TP y TN (círculos verdes), podemos observar que mayoritariamente en la matriz *tree* existe una mayor cantidad de valores positivos (<=50k) que fueron efectivamente determinados por el modelo como valores positivos, por lo cual, la variable **sensibilidad** (True Positive Rate, ***Sensitivity***) , debería ser más grande en *tree*, lo cual se condice con la respuesta obtenida (96.25 tree / 82.22 nb).
Respecto a los valores negativos (>50k) que fueron efectivamente determinados por el modelo como negativo, variable **especificidad** (True Negative Rate, ***Specificity***), podemos observar que mayoritariamente en el clasificador *nb*, existe una mayor cantidad, lo cual se observa en los valores reales obtenidos (39.14 tree / 72.96 nb).
Por último, podemos señalar la medida global de efectividad que involucra la suma de ambos valores de predicción correcta (TP, TN) divididos por la suma de todos los valores (TP, TN, FP, FN); que en clasificador *tree*=2709 y en el clasificador *bn*=2640, estos divididos por el total de valores (Test Set) = 3298, se obtiene una **precisión** (Classification ***Accuracy*** Rate) mayor para el *tree* (82.14 *tree* / 80.04 *nb*).
Si bien, preliminarmente podemos señalar que la medida de precisión es mayor en el clasificador *tree*, una medida adicional es la AUC de la curva ROC, que representa la capacidad de un modelo de distinguir entre clases, que en algunos casos, se presenta como una visión adicional a la precisión, se calculó además el AUC (0.82 *tree* / 0.87 *nb*), lo cual, nos lleva a la conclusión de que el mejor modelo es el modelo de *nb*.

```
> ctable <- as.table(matrix(c(2390, 93, 496, 319), nrow = 2, byrow = TRUE))
> fourfoldplot(ctable, color = c("#CC6666", "#99CC99"),
+              conf.level = 0, margin = 1, main = "Confusion Matrix")
> fourfoldplot(ctable, color = c("#CC6666", "#99CC99"),
+              conf.level = 0, margin = 1, main = "Matriz de Confusión - Arboles de
Decisión")
> ctable1 <- as.table(matrix(c(2047, 436, 222, 593), nrow = 2, byrow = TRUE))
> fourfoldplot(ctable1, color = c("#CC6666", "#99CC99"),
+              conf.level = 0, margin = 1, main = "Matriz de Confusión - Bayes Ingenuo")
```

REFERENCIAS:

1. Horton, Bob (2016) ROC Curves in Two Lines of R Code. Sitio: Revolution Analytics. [en línea] Recuperado de: https://blog.revolutionanalytics.com/2016/08/roc-curves-in-two-lines-of-code.html
2. Narkhede, Sarang (2018) Understanding AUC - ROC Curve. Sitio Towards Data Science. [en línea] Recuperado de: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
3. Narkhede, Sarang (2018) Understanding Confusion Matrix. Sitio: Towards Data Science. [en línea] Recuperado de: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
4. Von, Cambridge (2014) FRR, FAR, TPR, FPR, ROC curve, ACC, SPC, PPV, NPV. Sitio: Blog Cambridge [en línea] Recuperado de: https://cambridge-archive.blogspot.com/2014/04/frr-far-tpr-fpr-roc-curve-acc-spc-ppv.html