Control 5: Investigación de Operaciones

Nombre: Rodrigo Cayazaya M.

Correo: rodrigo.cayazaya@sansano.usm.cl

Rol: 201773538-4

Fecha de nacimiento: 21/08/1998

Código:

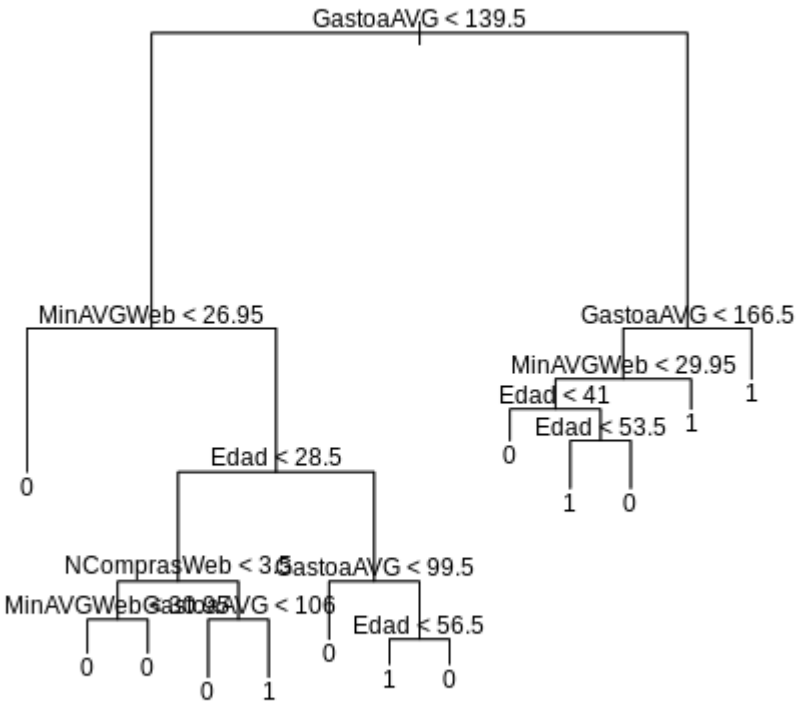
```
%load_ext rpy2.ipython

%%R
set.seed(21081998)
dataControl <-read.csv('DataControl2021.csv',sep = ";")
#print(head(dataControl))
shuffle_index <- sample(1:nrow(dataControl))
dataControl <- dataControl[shuffle_index, ]
print(head(dataControl))

#limpieza
%%R
library(dplyr)
dataControl <- dataControl %>%
select(-c(NumLikesIG,NumLikesFB)) %>%
na.omit()
glimpse(dataControl)

%%R
dim(dataControl)

%%R
size <- 1:(nrow(dataControl)*0.75)
data_train = dataControl[size,]
data_test = dataControl[-size,]
print(dim(data_test))
print(dim(data_train))
```

```
%%R
#install.packages("tree",dep = TRUE)
library(tree)
arbol = tree(as.factor(Suscripcion) ~.,data = data_train)
summary(arbol)

%%R
plot(arbol)
text(arbol,pretty=1)

%%R
prediccion <- predict(arbol,data_test,type='class')
conf_matrix <- with(data_test,table(prediccion,data_test$Suscripcion))
conf_matrix

%%R
acc <- sum(diag(conf_matrix))/nrow(data_test)
miss_class_error <- 1-acc
acc
```

1) Generar un árbol considerando el 75 % de los datos entregados para Training. Conteste lo siguiente:

- Describa los datos utilizados, explicando el tipo de dato de cada atributo.
  Se utilizaron todas las columnas, debido a que ninguna contenía “nan” o valores “?”. Todas las columnas son números enteros, excepto “MinAVGWeb”, el cual contiene valores flotantes, también “Suscripcion” y “Sexo” contienen valores Booleanos. También se utilizó la columna “Suscripcion” como variable objetivo.

- Describa el árbol obtenido:
  Cantidad de niveles obtenido: 6 niveles.
  Cantidad de hojas del árbol obtenido: 13 hojas.
  Cantidad de particiones: 12 particiones.
  Nodo padre: “GastoaAVG”.
  Accuracy: 0.7291667



- ¿Qué variables no generan particiones?, explique por qué éstas variables no participan en el árbol generado.
  Se utilizaron 4 columnas:
  - “GastoaAvg”
  - “MinAVGWeb”
  - “Edad”
  - “NComprasWeb”

  Por lo que no se utilizaron 4 variables, las cuales no generaron particiones:

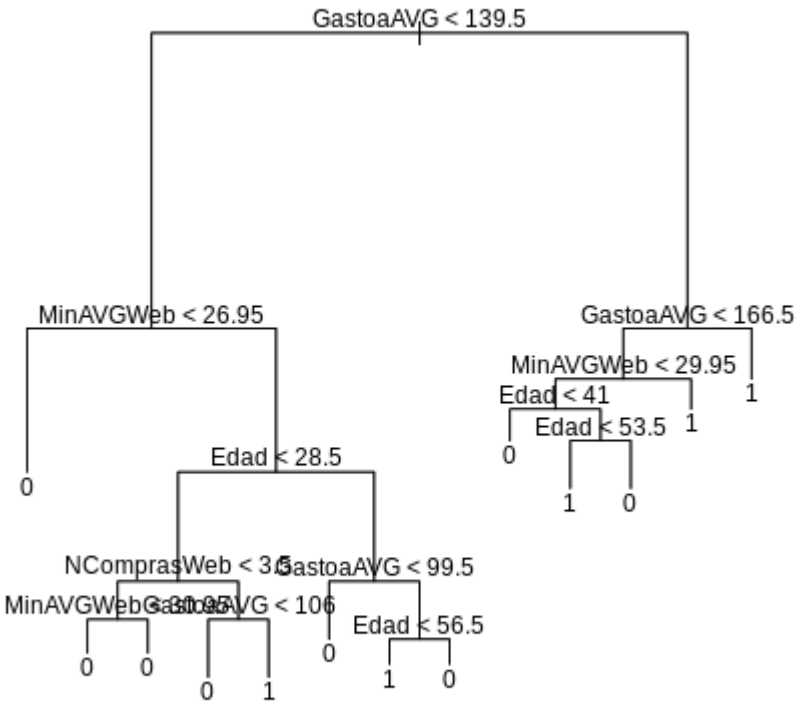
  - “NumLikesIG”
  - “NumLikesFB”
  - “Ciudad”
  - “Sexo”

Estas variables no participan en el árbol generado, debido a que se utilizaron las mayores cantidades de “bondad de partición” al escoger las variables para particionar. Esto va relacionado inversamente con la “impureza”.

2) Nuevamente, considerando el 75 % de los datos entregados para Training, ¿Qué sucede si las variables cantidad de Likes promedio en Instagram y Facebook no son consideradas en el modelo? describa detalladamente:

- Compare con el árbol obtenido en la pregunta (1) considerando cantidad de niveles del árbol, cantidad de hojas del árbol obtenido, entre otros atributos.

  El árbol quedó exactamente igual, debido a que no se utilizaban esas columnas en el árbol original.

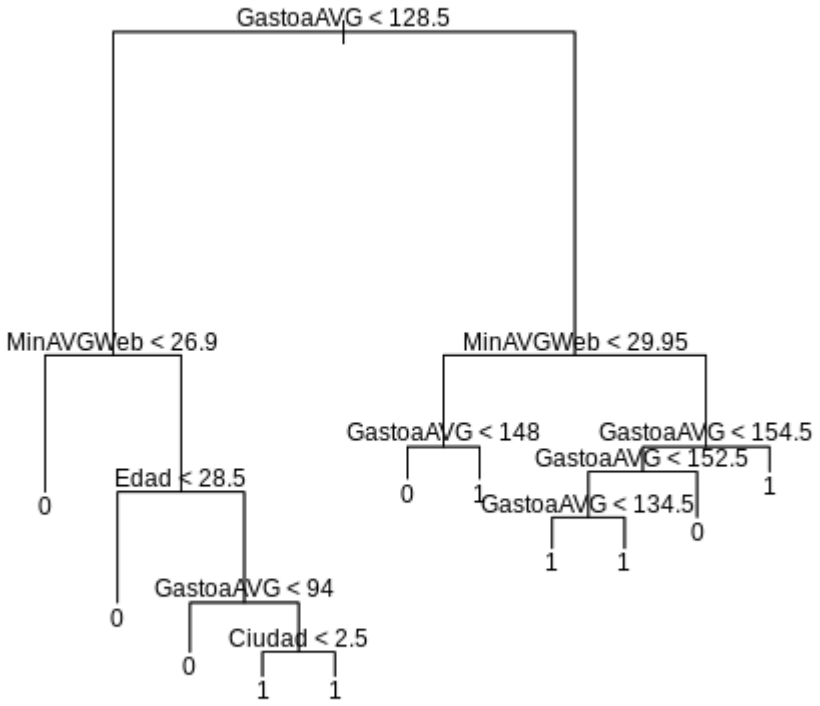


- Considerando el Testing set, ¿Cuál es la precisión del árbol?

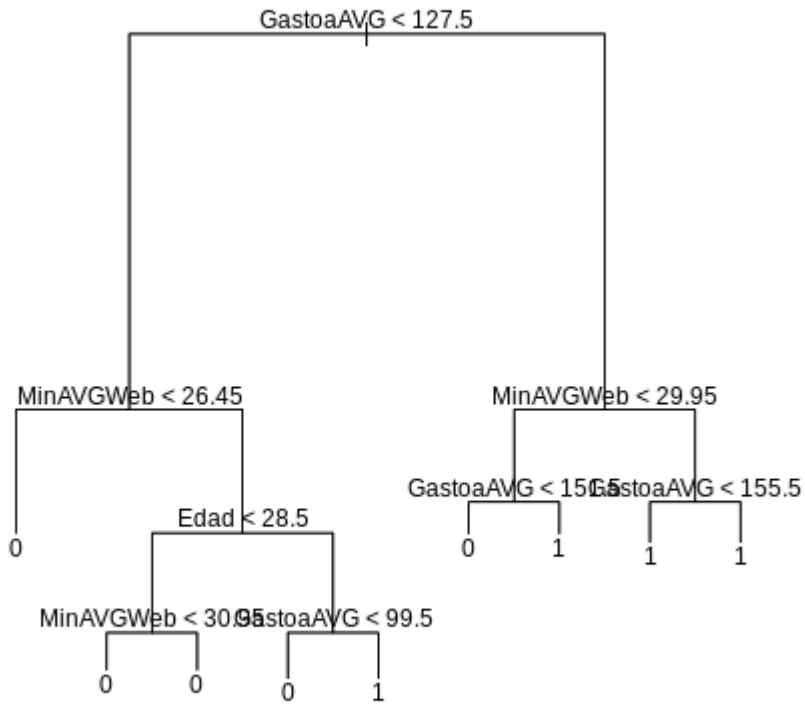
  Es exactamente igual al anterior (0.7291667), debido a que el árbol no cambió.

3) ¿Qué sucede si utiliza un 60 % y 90 % de los datos entregados como Training? Evaluar e imprimir los árboles obtenidos. Explique.

- Con 60% de Training se obtuvo:
  Cantidad de niveles obtenido: 6 niveles.
  Cantidad de hojas del árbol obtenido: 11 hojas.
  Cantidad de particiones: 10 particiones.
  Nodo padre: “GastoaAVG”.
  Accuracy: 0.7175325

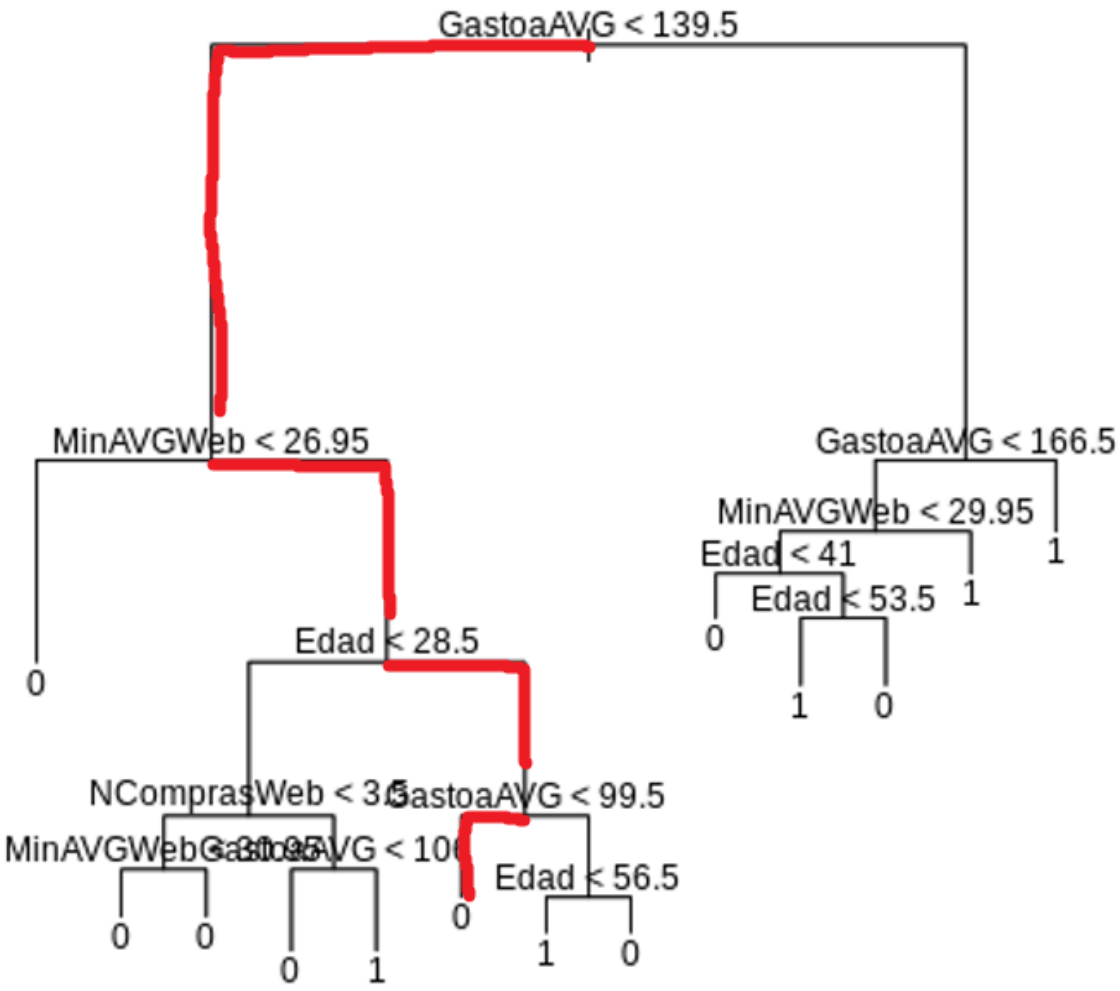


- Con 90% de Training se obtuvo:
  Cantidad de niveles obtenido: 5 niveles.
  Cantidad de hojas del árbol obtenido: 9 hojas.
  Cantidad de particiones: 8 particiones.
  Nodo padre: “GastoaAVG”.
  Accuracy: 0.7402597



Debido a que se utilizaron más valores para el Training, se obtuvo un accuracy mayor. Esto es debido a que el árbol se creó utilizando más información, por lo que pudo particionar de mejor manera las columnas. Sin embargo, este accuracy no es tan confiable, debido a que se utilizaron una menor cantidad de datos de Testeo.

4) Considerando el árbol de la pregunta (1), ¿cuál sería el resultado asignado al siguiente cliente: (8,91,82,0,35.6,68,4,1)



El resultado sería 0, esto significa que no está suscrito.