Pecotche Andres
Lucas Carballo

Minería de datos usando sistemas inteligentes

Práctica 4 - Árboles de decisión

4.a) Construya manualmente, a partir de los datos la hoja Train del archivo trabajos_ej4.xlsx y utilizando como criterio la Ganancia de Información, el árbol de clasificación capaz de predecir si una persona obtendrá o no el trabajo según los antecedentes que posea. Indique en cada paso los valores de Entropía obtenidos y las selecciones realizadas. Para cada selección de atributo para dividir un nodo, incluir una tabla con tantas columnas como atributos y la ganancia de información y entropía de cada uno. Puede verificar los resultados obtenidos manualmente al consultar los valores devueltos por el operador Weight by Information Gain de RapidMiner o los scripts de Python provistos en la teoría. Dibuje y explique el árbol obtenido. ¿Podría darle algún consejo a quienes quieran obtener el trabajo?

Título universitario	Experiencia en el cargo	Cantidad de Trabajos Anteriore	Trabaja actualmente	Obtiene traba
SI	BAJA	6	SI	SI
NO	BAJA	6	NO	NO
NO	ALTA	6	NO	NO
NO	ALTA	6	NO	NO
NO	MEDIA	6	NO	SI
SI	BAJA	6	NO	SI
NO	BAJA	6	SI	NO
NO	ALTA	6	SI	NO
SI	ALTA	6	NO	SI
NO	MEDIA	8	NO	SI
NO	BAJA	8	NO	SI
NO	MEDIA	8	NO	SI
NO	MEDIA	8	NO	SI
NO	ALTA	9	SI	NO
NO	MEDIA	9	SI	NO
NO	MEDIA	9	NO	SI

Tabla de datos

Entropía del conjunto:
E = conjunto de valores
n = valores distintos del conjunto E, en este caso son 2 (“si” o “no” obtienen trabajo)
p_i = proporción de valores del conjunto que toman el coinciden con el i-ésimo valor.

$$Entropia(E) = \sum_{i=1}^{n} - p_i \, log_2(p_i)$$

Para este conjunto la entropía es: $-9/16*\log_2(9/16) - 7/16*\log_2(7/16)$ = **0.9886**
Donde el total de muestras es 16, de las cuales 9 son “Si” y 7 “no”. Como la heterogeneidad del conjunto es baja (9 y 7), la entropía es alta.

Ahora debemos calcular la entropía de cada valor posible de cada atributo, y luego la entropía total del atributo:

Título universitario: 2 opciones (“si”,“no”)

Entropía (E_{si}) = $-3/3*\log_2(3/3)$ = **0**

Entropía(E_{no}) = -6/13*$\log_2$(6/13) - 7/13*$\log_2$(7/13) = **0.9957**

Entropía del atributo:

Entropía (E,título universitario) = 3/16*0 + 13/16*0.9957 = **0.809**

Ganancia:

Entropía (E) - Entropía (E, título universitario) = 0.9886 - 0.809 = **0.1796**

Donde 0.1796 es la ganancia obtenida en la entropía del conjunto si se utiliza el atributo título universitario.

Experiencia en el cargo:

Entropía (E_{baja}) = -3/5*$\log_2$(3/5) - 2/5*log(2/5) = **0.9709**
Entropía (E_{media}) = -5/6*$\log_2$(5/6) - 1/6*log(1/6) = **0.6500**
Entropía (E_{alta}) = -1/5*$\log_2$(1/5) - 4/5*log(4/5) = **0.7219**

Entropía (E,experiencia) = 5/16*0.9709 + 6/16*0.65 + 5/16*0.7219 = **0.77275**

Ganancia: 0.9886 - 0.77275 = **0.21585**

Cantidad de trabajos anteriores:

Este es un atributo numérico, por lo tanto, debemos ordenar los posibles valores que toman los ejemplos, y probar todos los valores de corte, dividiendo los ejemplos en dos grupos (menor al valor de corte, y mayor) y luego calcular la entropía para estos dos grupos, asi pasando por todas las opciones de corte, y quedándonos con el valor de corte que nos brinde una menor entropía. Como en esta muestra solo estan los valores 6, 8 y 9, los puntos de corte a probar serán 7 (quedandonos el grupo de 6 y por otro lado el grupo 8 y 9) y el 8.5 (quedandonos el grupo de 6 y 8 y por otro lado el grupo con el 9.

Valor del atributo: 6 -> 9/16 tienen este valor -> 4 de los 9 “si” consiguen trabajo y 5 “no”
valor del atributo: 8 -> 4/16 -> 4 “si”, 0 “no”
valor del atributo: 9 -> 3/16 -> 1 “si”, 2 “no”

Si tomamos el corte en 7:
<7 = 4 ejemplos toman el valor “si” y 5 el valor “no”
>7 = 5 toman el valor “si” y 2 “no”

Entropía ($E_{<7}$) = -4/9*$\log_2$(4/9) - 5/9*$\log_2$(5/9) = **0.9910**
Entropía ($E_{>7}$) = -5/7*$\log_2$(5/7) - 2/7*$\log_2$(2/7) = **0.8631**

Entropía ($E_{\text{Trabajos anteriores}}$)(corte en 7) = 9/16*0.9910 + 7/16*0.8631 = **0.9350**
Ganancia = 0.9886 - 0.9350 = **0.0536**

Si tomamos el corte en 8.5:
<8.5 = 13/16 -> 8 ejemplos toman el valor “si” y 5 el valor “no”
>8.5 = 3/16 -> 1 toma el valor “si” y 2 “no”

Entropía ($E_{<8.5}$) = -8/13*$\log_2$(8/13) - 5/13*$\log_2$(5/13) = **0.9612**
Entropía ($E_{>8.5}$) = -1/3*$\log_2$(5/7) - 2/3*$\log_2$(2/3) = **0.9182**

Entropía ($E_{\text{Trabajos anteriores}}$)(corte en 8.5) = 13/16*0.9612 + 3/16*0.9182 = **0.9531**
Ganancia = 0.9886 - 0.9531 = **0.03546**

Nos quedamos con el punto de corte en 7, ya que nos brinda una mejor ganancia.

Trabaja actualmente:

Entropía (E_{si}) = -1/5*$\log_2$(1/5) - 4/5*$\log_2$(4/5) = **0.7219**
Entropía (E_{no}) = -8/11*$\log_2$(8/11) - 3/11*$\log_2$(3/11) = **0.8453**

Entropía ($E_{\text{Trabaja actualmente}}$) = 5/16*0.7219 = 11/16*0.8453 = **0.8067**
Ganancia = 0.9886 - 0.8067 = **0.1818**

Resumiendo:

Atributo	Ganancia	Entropía
Título universitario	0.1796	0.809
Experiencia	**0.2158**	**0.7727**
Cantidad de trabajos anteriores	0.0536	0.9350
Trabaja actualmente	0.1818	0.8067

La experiencia es el atributo que usando como raíz del árbol nos va a brindar una mayor ganancia de información.

Ahora debemos analizar la respuesta del resto de los atributos para los ejemplos que aún no pertenecen a un subconjunto homogéneo (aún no hay subconjuntos homogéneos):

Experiencia BAJA:

Título universitario	Experiencia en el cargo	Cantidad de Trabajos Anteriore	Trabaja actualmente	Obtiene traba
NO	BAJA	6	NO	NO
NO	BAJA	6	SI	NO
SI	BAJA	6	SI	SI
SI	BAJA	6	NO	SI
NO	BAJA	8	NO	SI

Entropía del conjunto: -2/5*$\log_2$(2/5) - 3/5*$\log_2$(3/5) = **0.9709**

Título universitario:

Entropía (E_{si}) = -2/2*$\log_2$(2/2) = **0**

Entropía(E_{no}) = -2/3*$\log_2$(2/3) - 1/3*$\log_2$(1/3) = 0.9182

Entropía del atributo:

Entropía (E,título universitario) = 2/5*0 + 3/5*0.9182 = **0.55092**

Ganancia:

Entropía (E) - Entropía (E, título universitario) = 0.9709 - 0.55092 = **0.419**

Cantidad de trabajos anteriores:

Cómo realizamos el corte en 7, tomamos dos valores, el menor y mayor a 7. Para este caso (Experiencia baja), los menores a siete son 4 y los mayores 1.

Entropía (E_{menor}) = -2/4*$\log_2$(2/4) - 2/4*$\log_2$(2/4) = 1

Entropía(E_{mayor}) = -1/1*$\log_2$(1/1)=0

Entropía del atributo:

Entropía (E, trabajos anteriores) = 4/5*1 + 1/5*0 = **0.8**

Ganancia: Entropía (E) - Entropía (E, trabajo anterior) = 0.9709 - 0.8 = **0.1309**

Trabaja Actualmente

Entropía (E_{si}) = -1/2*$\log_2$(1/2) - 1/2*$\log_2$(1/2) = 1

Entropía(E_{no}) = -2/3*$\log_2$(2/3) - 1/3*$\log_2$(1/3) = 0.9182

Entropía (E, trabaja) = 2/5*1 + 3/5* 0.9182= 0.95092

Ganancia:

Entropía (E) - Entropía (E, trabaja) = 0.9709 - 0.95092= **0.0199**

Atributo	Ganancia	Entropía
Título universitario	**0.419**	**0.5509**
Cantidad de trabajos anteriores	0.1709	0.8
Trabaja actualmente	0.0199	0.9509

Experiencia MEDIA:

Título universitario	Experiencia en el cargo	Cantidad de Trabajos Anteriore	Trabaja actualmente	Obtiene traba
NO	MEDIA	9	SI	NO
NO	MEDIA	6	NO	SI
NO	MEDIA	8	NO	SI
NO	MEDIA	8	NO	SI
NO	MEDIA	8	NO	SI
NO	MEDIA	9	NO	SI

Entropía del conjunto= -1/6*$\log_2$(1/6) - 5/6*$\log_2$(5/6) = 0.6500

Título universitario:

Entropía (E_{si}) = -0/0*$\log_2$(0/0) = ¿?

Entropía(E_{no}) = -5/6*$\log_2$(5/6) - 1/6*$\log_2$(1/6) = 0.6500

Entropía del atributo:

Entropía (E,título universitario) =6/6*0.6500 = **0.6500**

Ganancia:

Entropía (E) - Entropía (E, título universitario) =0.6500 - 0.6500 = **0**

Cantidad de trabajos anteriores:

Cómo realizamos el corte en 7, tomamos dos valores, el menor y mayor a 7. Para este caso (Experiencia media), los menores a siete son 1 y los mayores 5.

Entropía (E_{menor}) = -1/1*$\log_2$(1/1)=0

Entropía(E_{mayor}) = -4/5*$\log_2$(4/4) - 1/5*$\log_2$(1/5) =0.4643

Entropía del atributo:

Entropía (E, trabajos anteriores) = 1/6*0 + 5/6*0.4643 = **0.3869**

Ganancia:

Entropía (E) - Entropía (E, trabajo anterior) =0.6500 - **0.3869** = **0.2630**

Trabaja Actualmente

Entropía (E_{si}) = -0/1$\log_2$(0/1) -1/1$\log_2$(1/1) = **0**

Entropía(E_{no}) = -5/5*$\log_2$(5/5) = **0**

Entropía (E, trabaja) = 0

Ganancia:

Entropía (E) - Entropía (E, trabaja) = 0.6500 - 0= 0.6500.

Atributo	Ganancia	Entropía
Título universitario	0	0.6500
Cantidad de trabajos anteriores	0.2630	0.3869
Trabaja actualmente	**0.6500**	**0**

Experiencia ALTA:

Título universitario	Experiencia en el cargo	Cantidad de Trabajos Anteriore	Trabaja actualmente	Obtiene traba
NO	ALTA	6	NO	NO
NO	ALTA	6	NO	NO
NO	ALTA	6	SI	NO
NO	ALTA	9	SI	NO
SI	ALTA	6	NO	SI

Entropía del conjunto= -1/5*$\log_2$(1/5) - 4/5*$\log_2$(4/5) = **0.2173**

Título universitario:

Entropía (E_{si}) = -1/1*$\log_2$ (1/1) = **0**

Entropía(E_{no}) = -0/4*$\log_2$(0/4) -4/4*$\log_2$(4/4) = **0**

Entropía del atributo:

Entropía (E,título universitario) = **0**

Ganancia: Entropía (E) - Entropía (E, título universitario) = **0.2173**

Cantidad de trabajos anteriores:

Para este caso (Experiencia alta), los menores a siete son 4 y los mayores 1.

Entropía (E_{menor}) = -1/4*$\log_2$(1/4) - 3/4*$\log_2$(3/4) = 0.2442

Entropía(E_{mayor}) = 0 -1/1*$\log_2$(1/1)=0

Entropía (E, trabajos anteriores) = 4/5* 0.2442 + 1/5*0 = **0.1953**

Ganancia: Entropía (E) - Entropía (E, trabajo anterior) =**0.2173** - **0.1953** = **0.0219**

Trabaja Actualmente

Entropía (E_{si}) = -0/2*$\log_2$(2/2) -2/2*$\log_2$(2/2)=0

Entropía(E_{no}) = -2/3*$\log_2$(2/3) - 1/3*$\log_2$(1/3) = 0.9182

Entropía (E, trabaja) = 2/5*0 + 3/5* 0.2764= **0.1658**

Ganancia: Entropía (E) - Entropía (E, trabaja) = **0.2173 - 0.1658 = 0.0514**

Atributo	Ganancia	Entropía
Título universitario	**0.2173**	**0.0**
Cantidad de trabajos anteriores	0.0219	0.1953
Trabaja actualmente	0.0514	0.1658

Para los casos que la experiencia es **baja**, y **no tiene título universitario** volvemos a calcular la entropía y la ganancia de los dos atributos restantes (cantidad de trabajos anteriores y trabaja actualmente).

Título universitario	Experiencia en el cargo	Cantidad de Trabajos Anteriore	Trabaja actualmente	Obtiene traba
NO	BAJA	6	NO	NO
NO	BAJA	6	SI	NO
NO	BAJA	8	NO	SI

Entropía del conjunto= -2/3*$\log_2$(2/3)-1/3*$\log_2$(1/3) = **0.9182**

Cantidad de trabajos anteriores:

Entropía (E_{menor}) = -2/2*$\log_2$(2/2) = 0

Entropía(E_{mayor}) = -1/1*$\log_2$(1/1) = 0

Entropía del atributo:

Entropía (E,título universitario) = 0

Ganancia: Entropía (E) - Entropía (E, título universitario) = **0.9182**

Trabaja Actualmente

Entropía (E_{si}) = -1/1*$\log_2$(1/1) = 0

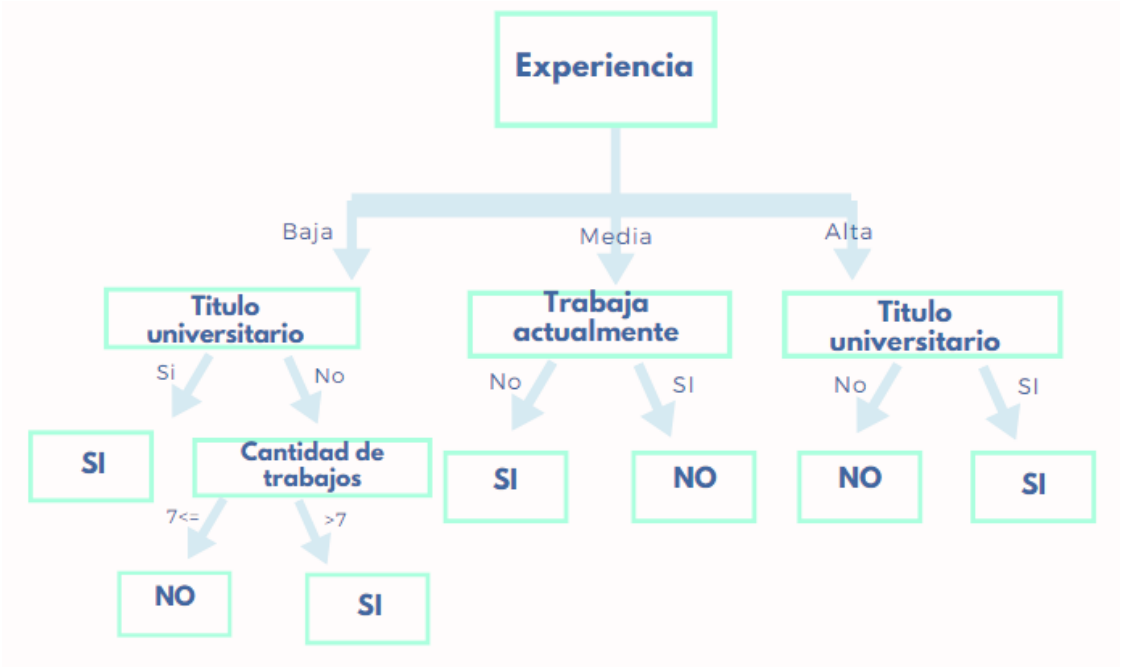
Entropía(E_{no}) = -1/2*$\log_2$(1/2) - 1/2*$\log_2$(1/2) = 1

Entropía (E, trabaja) = 1/3*0 + 2/3* 1= 0.666

Ganancia: Entropía (E) - Entropía (E, trabaja) = 0.9182 - 0.666= 0.2582

Atributo	Ganancia	Entropía
Título universitario	**0.9182**	**0**
Trabaja Actualmente	0.2582	0.666

Esquema del Árbol de decisión para ver si obtiene o no el trabajo:



Como recomendaciones para que la persona pueda **obtener el trabajo**, tendremos 4 opciones:

- Se desea que tenga experiencia **alta** y que tenga **título universitario**.
- Si tiene experiencia **media**, y se encuentra **sin trabajo**, lo obtendrá.
- Por último, obtendrá el trabajo si tiene experiencia **baja** y tiene **título universitario**
- Otro caso es si **no tiene título universitario** pero la **cantidad de trabajos** anteriores es **mayor a 7.**

C)
Utilizando el árbol del inciso anterior (a), realizamos nuevamente la columna de obtiene trabajo y obtenemos los siguientes resultados:

Título universitario	Experiencia en el cargo	Cantidad de Trabajos Anteriores	Trabaja actualmente	Obtiene trabajo
SI	ALTA	8	SI	SI
NO	MEDIA	6	SI	NO
NO	BAJA	2	NO	NO
NO	MEDIA	5	NO	SI

Primer dato: Como la experiencia es Alta y tiene titulo universitario → SI
Segundo dato: Como la experiencia es Media y trabaja actualmente →NO
Tercer dato: Como la experiencia es Baja y no tiene título universitario, y la cantidad de trabajos anteriores es menor a 7 →NO
Cuarto dato: Como la experiencia es Media y no trabaja actualmente →SI

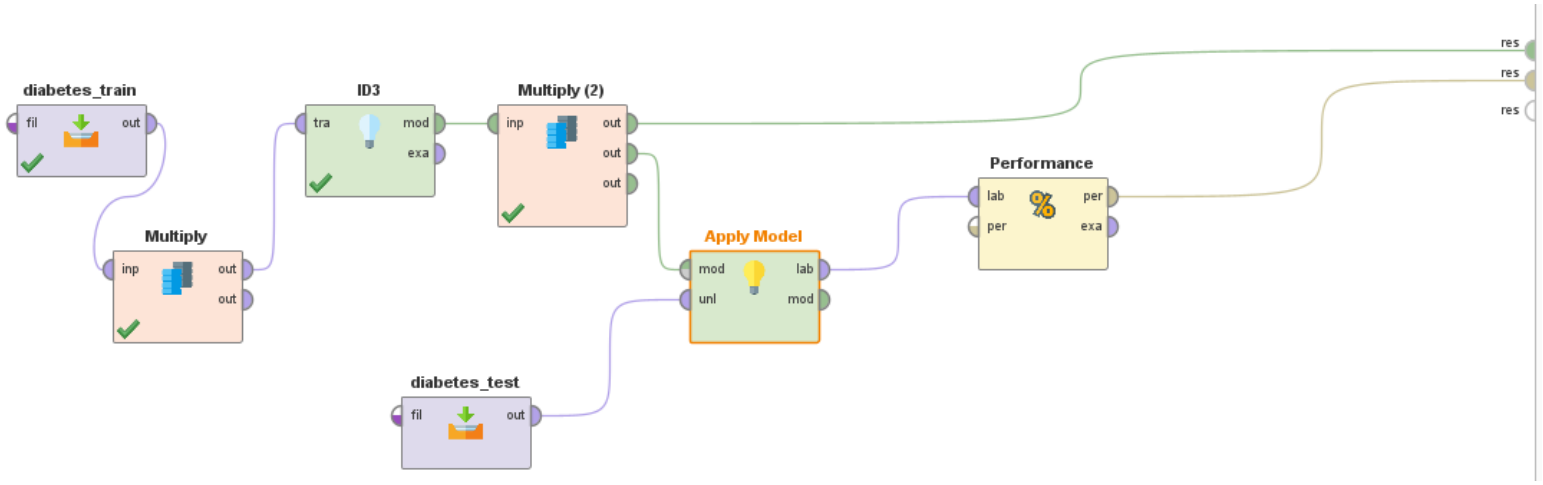
Ejercicio 6

a) *Genere un árbol de decisión ID3 para el archivo diabetes_nominal_train. Utilice la medida de desorden "information gain", "minimal gain" igual a 0.01, con "minimal size for split" igual a 40 y "minimal leaf size" igual a 20. Observe la tasa de aciertos obtenida sobre el conjunto de entrenamiento y el tamaño del árbol resultante (la cantidad de nodos). Luego aplique el árbol obtenido sobre los datos del archivo diabetes_nominal_test.xlsx y mida tasa de aciertos. Nota: para calcular la cantidad de nodos, puede contar la cantidad de filas de la descripción textual del árbol. Recomendamos copiar la descripción del árbol a un editor de texto donde le indique la cantidad de filas.*

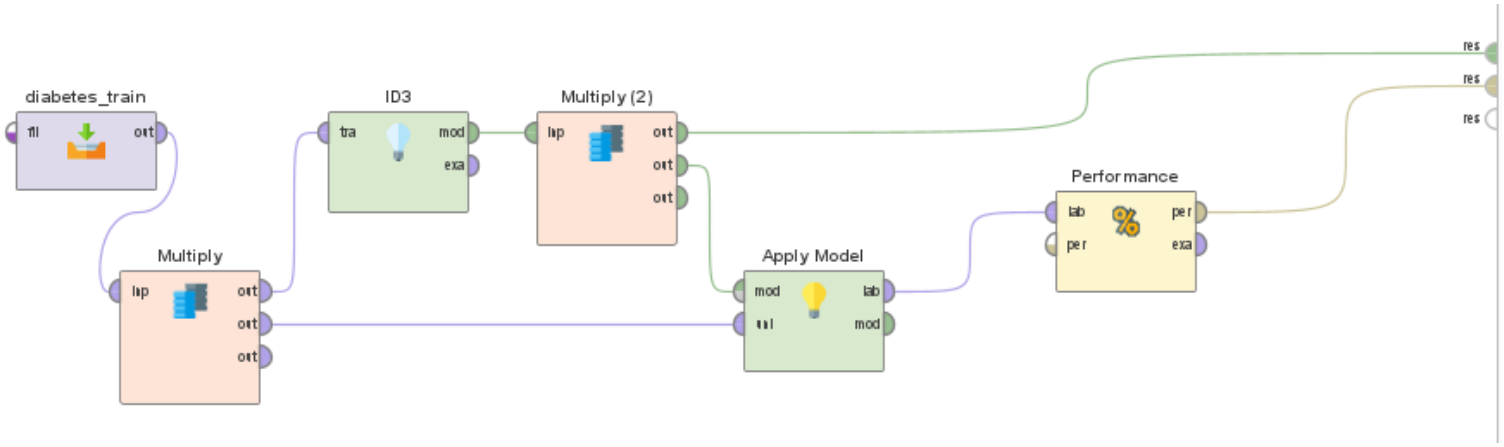
b) *Repita el inciso (a) utilizando "minimal size for split" igual a 100 y "minimal leaf size" igual a 50. Observe nuevamente el tamaño del árbol resultante y la performance obtenida en los conjuntos de entrenamiento y testeo.*

c) *Repita el inciso (a) pero ahora utilizando "minimal size for split" igual a 300 y "minimal leaf size" igual a 200. Observe el tamaño del árbol resultante y la performance obtenida en ambos conjuntos de datos.*

d) *¿Qué ocurre con el tamaño y complejidad del árbol en cada caso? ¿Se observan diferencias en la performance del árbol en ambos conjuntos de datos? ¿A qué se deben estas diferencias?*



Diseño en rapidminer para medir la tasa de aciertos con el conjunto de datos test



Diseño en rapidminer para medir la tasa de aciertos con el conjunto de datos de entrenamiento.

Resultados:

Inciso	a)	b)	c)
minimal size for split	40	100	300
minimal leaf size	20	50	200
Tamaño del árbol	84	32	4
Accuracy (Train)	79.59%	73.08%	71.06%
Accuracy (Test)	53.25%	57.14%	54.55%

Para calcular el tamaño del árbol se utilizó una herramienta contadora de líneas, donde se le ingresó la descripción del árbol como entrada, ya que esta cuenta con un nodo por línea.

Minimal leaf size: Número de ejemplos mínimo para los nodos hoja.
Minimum size for split: Es el valor mínimo de ejemplos que tiene que tener un nodo para ser dividido. Es decir, un nodo no se va a dividir en más nodos hasta no tener un número de ejemplos mayor o igual a este parámetro.

Como se puede observar la complejidad del árbol y el número de nodos decrece a medida que el tamaño mínimo que deben tener las hojas del árbol aumenta, el parámetro ”minimal size for split” realmente no es tan determinante en esto, tanto como sí lo es el primero. Al especificar al modelo, que el tamaño mínimo de hojas debe ser mayor, el árbol va a resultar en hojas menos específicas a cada ejemplo de la muestra, y va a tener una menor cantidad de hojas, pero con más ejemplos por hoja. Como el número de ejemplos en la muestra es el mismo, al aumentar el número de ejemplos por hoja, el árbol se reduce.

Las diferencias que se pueden observar dependiendo del conjunto surgen debido a que la muestra de testeo, presenta ciertas diferencias con la muestra de entrenamiento. Si estas tuvieran una selección mejor distribuida entre el total de ambas muestras, posiblemente los resultados serían más similares. Aunque naturalmente medir la tasa de aciertos sobre los mismos datos usados en entrenamiento, va a brindar una mayor tasa de acierto, que sobre otra muestra completamente diferente.