



Técnicas de Clasificación

PRÁCTICA 3: Árboles de decisión.

V. Miguel Sempere Navarro

En la presente práctica se va a llevar a cabo un estudio de variables mediante la técnica de los árboles de decisión, la cual ha sido estudiada con anterioridad en clase y que tiene como objetivo clasificar si una familia tiene riesgo de pobreza o no.

En la anterior práctica se realizó un estudio de la misma base de datos, pero mediante la técnica de la regresión logística; por tanto, en esta práctica se realizará un análisis y explicación de la técnica de los árboles de decisión con el objetivo de posteriormente comparar los resultados obtenidos en ambas prácticas.

La metodología a seguir será crear primero un árbol de decisión podado que minimice los errores de cross validation. Seguidamente se hará lo mismo con un árbol de inferencia y se elegirá el que otorgue una matriz de confusión con una predicción más ajustada.

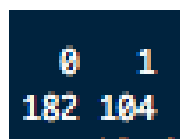
Para finalizar, se hará una comparación entre la matriz de confusión del mejor árbol de decisión y se comparará con la matriz de confusión otorgada por el modelo anterior de regresión logística.

Para el correcto desarrollo de la práctica, en primer lugar, se comenzará abriendo el dataset que se ha proporcionado y realizando una primera limpieza de variables que no se consideran adecuadas para explicar el modelo. Una de ellas será la variable “Renta”, tal y como se ha sugerido, entre otras. Con respecto a las observaciones, se trabajará con las 477 observaciones del dataset. Sin embargo, con respecto a las variables, se pasa de trabajar con 18 variables a 12 y se transforman todos los datos a tipo factor.

Ahora se dividen los datos. Al igual que en la práctica anterior; sobre las observaciones se realiza una partición para poder distinguir entre un modelo de entrenamiento y un modelo de test repartidos al 60% y 40%, respectivamente. Como ya ha sido explicado en la anterior práctica, al realizar un modelo de regresión lineal sobre la variable explicativa se observa que las variables que estadísticamente son más significativas son AyudaFamilias, VacacionesOutdoor, CapacidadAfrontar, LlegarFinMes, Miembros, HogaresSemanales y ActMayor. En el modelo de regresión, seguidamente se realizaba

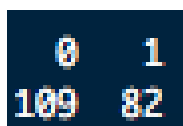
la bondad del ajuste con el test McFadden, el cual daba como resultado 0.38 aproximadamente. Esto quiere decir que el modelo está bien ajustado. Se establecía el umbral en 0.68 y a partir de aquí se calculaba la matriz de confusión con un accuracy del 74%.

Ahora es el momento de realizar lo mismo que se realizó para el modelo de regresión logística, pero para los árboles de decisión. Primero se representa en una tabla tanto el train como el test de aquellas familias que están en la pobreza (1) o no lo están (0):



	0	1
0	182	104
1	109	82

Ilustración 1. Train. Fuente: elaboración propia.



	0	1
0	109	82
1	109	82

Ilustración 2. Test. Fuente: elaboración propia.

Ahora se va a representar el árbol de clasificación del modelo de entrenamiento con la librería rpart.plot. Como se observa, de las 286 observaciones existen 104 con errores en el nodo principal, siendo la proporción del 36%

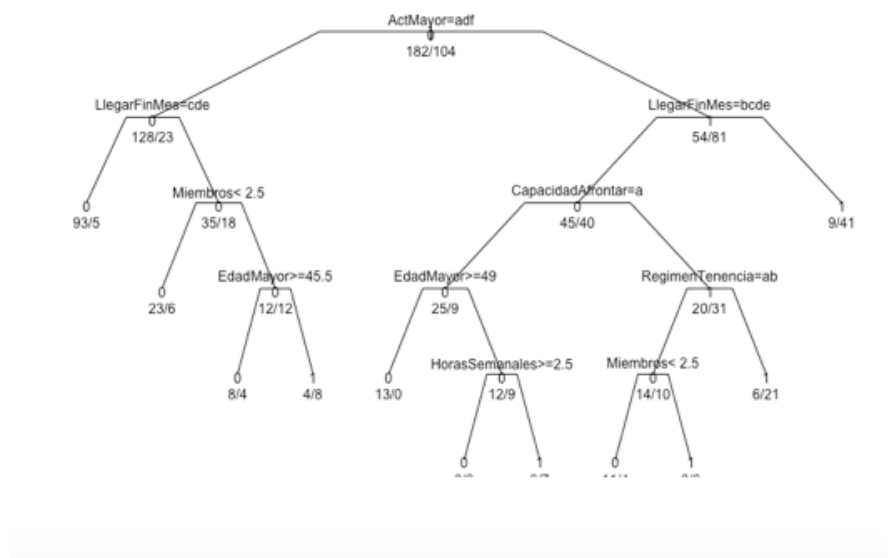


Ilustración 3. Árbol de clasificación. Fuente: elaboración propia.

Ahora se representa gráficamente el error relativo por variables estadísticamente significativas y se verá cuál sería el valor mínimo de validación cruzada para la poda. Se obtiene que el mínimo es un cp de 0.01

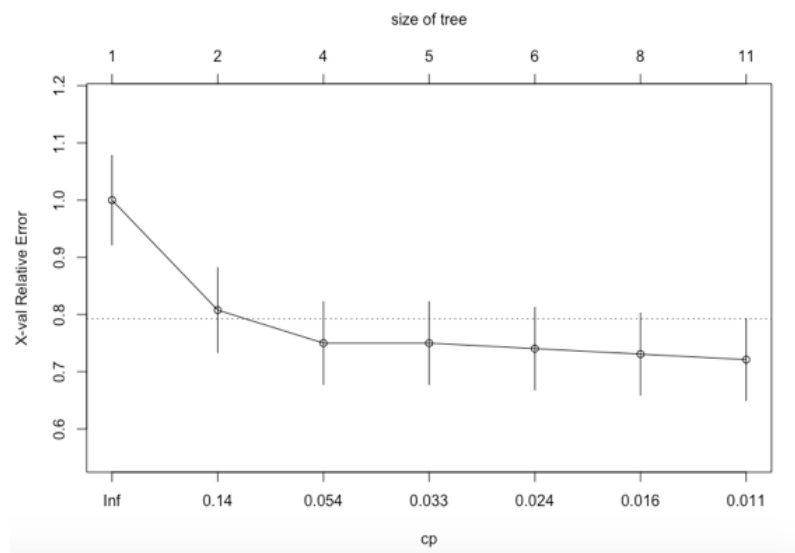


Ilustración 4. Error relativo por variable. Fuente: elaboración propia.

Esta sería la representación del árbol de clasificación una vez podado:

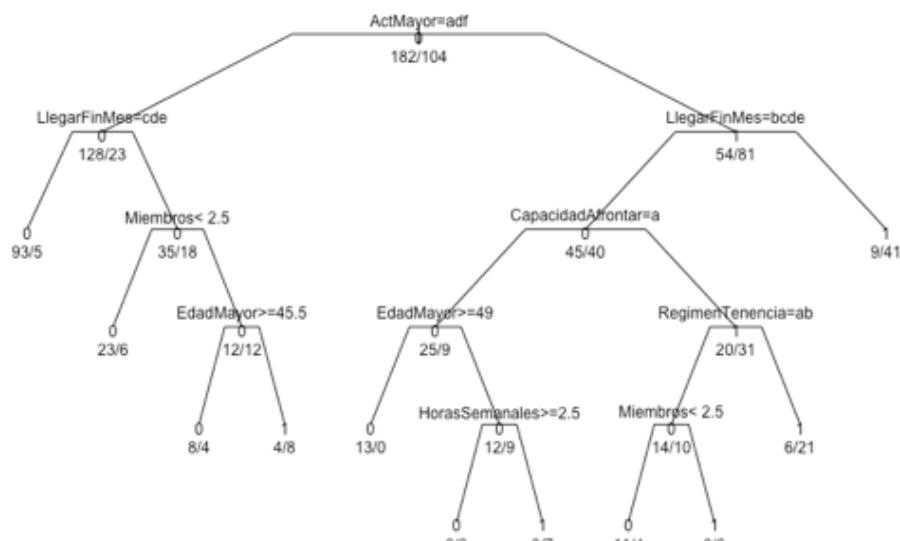


Ilustración 5. Árbol de clasificación podado. Fuente: elaboración propia.

Utilizando otra librería (party), se pueden obtener los llamados árboles de inferencia, es decir, árboles de regresión no paramétricos. Estos no necesitan hacer la poda.

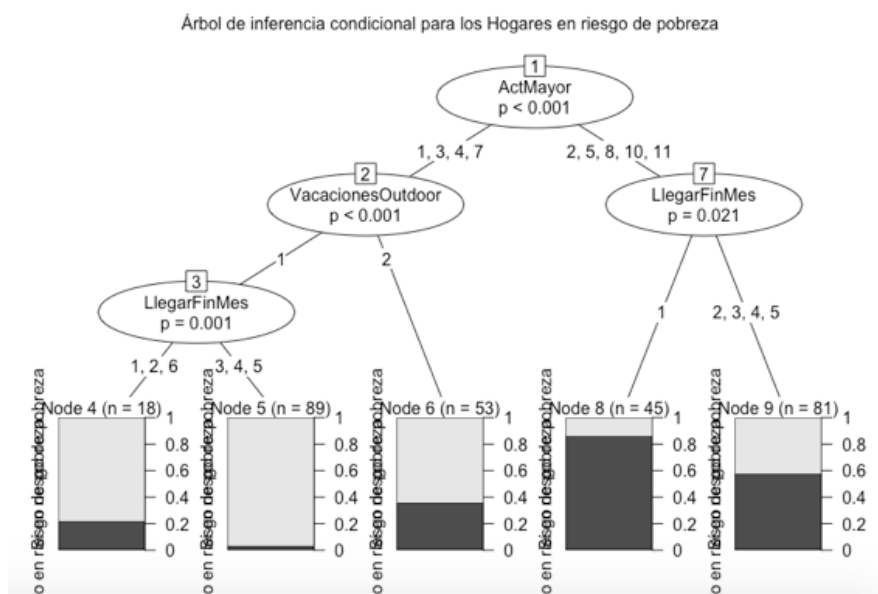


Ilustración 6. Árbol de inferencia. Fuente: elaboración propia.

Llegado a este punto se ha representado un árbol de decisión con la librería “rpart.plot” y otro árbol llamado árbol de inferencia con la librería “party”. Esto lleva a cuestionar

cual es la mejor solución o, mejor dicho, el árbol que mejor resultado otorga. Para ello se va a representar ambas matrices de confusión y se analizarán:

```
> arbol.perf
      Predicted
Actual      No en riesgo de pobreza Sí en riesgo de pobreza
No en riesgo de pobreza      99      18
Sí en riesgo de pobreza      31      43
> ctree.perf
      Predicted
Actual      No en riesgo de pobreza Sí en riesgo de pobreza
No en riesgo de pobreza      91      26
Sí en riesgo de pobreza      25      49
```

Ilustración 7. Matrices de confusión de los árboles. Fuente: elaboración propia.

Como se puede observar, la matriz de confusión hecha con el árbol de clasificación de la librería “rpart.plot” obtiene más Verdaderos Negativos (99) que la matriz de confusión del árbol de inferencia de la librería “party” (91), sin embargo, “rpart.plot” obtiene menos Verdaderos Positivos (43) que la librería “party” (49). A la hora de los fallos, es decir, las veces que una matriz dice que no hay pobreza pero realmente si la hay y al revés, la matriz de confusión del árbol de inferencia es ligeramente peor, por lo que **elegiremos como mejor resultado el árbol de decisión de la librería “rpart.plot”**.

Para finalizar, se pide realizar una comparación entre los resultados arrojados por la matriz de confusión del modelo de regresión logística llevado a cabo en la práctica 2 y el mejor resultado de matriz de confusión de los árboles.

```
> logit.perf
      Predicted
Actual  0    1
0  102    7
1   42   40
```

Ilustración 8. Matriz de confusión del modelo de regresión logística. Fuente: elaboración propia.

Comparando los resultados de la anterior matriz de confusión (la del modelo de regresión logística) con los resultados de la mejor matriz de confusión otorgada por

árboles de clasificación, llegamos a la conclusión de que ambos modelos son igual de buenos. Esto se debe a que la suma de Verdaderos Positivos y Verdaderos Negativos de ambas matrices son iguales: $99 + 43 = 102 + 40$. Del mismo modo, la suma de errores es la misma: $31 + 18 = 42 + 7$.

Se extrae, por tanto, como conclusión, que son igual de buenos los resultados obtenidos gracias a un modelo de regresión logística que a un modelo obtenido por árboles de decisión.