MASTER EN DATA
SCIENCE PARA FINANZAS

INFORME: ÁRBOLES DE DECISIÓN

Luis Llera García
Técnicas de clasificación

1.EXECUTIVE SUMMARY

El objetivo del siguiente informe será realizar un modelo de árboles de decisión, él cual es un modelo predictivo que intenta es predecir de si un determinado hogar está en riesgo de pobreza o no, en función de una serie de variables. Un árbol de decisión es un mapa de los posibles resultados de una serie de decisiones relacionadas. Permite que un individuo o una organización comparen posibles acciones entre sí según sus costos, probabilidades y beneficios. Se pueden usar para dirigir un intercambio de ideas informal o trazar un algoritmo que anticipe matemáticamente la mejor opción.[i]

Sirven para solucionar problemas tanto de regresión como de clasificación, son modelos muy simples, pero en eso mismo reside fundamentalmente su interés, que sean fácilmente ejecutables así como interpretables.

La base de datos que utilizaremos será una encuesta de condiciones de vida en 2016, realizada por el INE que se realiza desde 2004. Basada en criterios armonizados para todos los países de la Unión Europea, su objetivo fundamental es disponer de una fuente de referencia sobre estadísticas comparativas de la distribución de ingresos y la exclusión social en el ámbito europeo.
La realización de la ECV permite poner a disposición de la Comisión Europea un instrumento estadístico de primer orden para el estudio de la pobreza y desigualdad, el seguimiento de la cohesión social en el territorio de su ámbito, el estudio de las necesidades de la población y del impacto de las políticas sociales y económicas sobre los hogares y las personas, así como para el diseño de nuevas políticas, de ahí su importancia.[ii]

Por tanto, lo que se buscará en el siguiente informe será realizar un modelo de árboles de decisión con dichos datos, utilizando como variable a explicar"Hogar en riesgo de pobreza", que es de tipo categórico, el problema de los árboles es su estabilidad, es difícil que un modelo de árboles de decisión sea estable, ese es básicamente su mayor inconveniente. Para más tarde compararlo con nuestros resultados que obtuvimos con nuestra regresión logística, realizada en el anterior informe.

Con todo esto nuestro objetivo será predecir si un hogar se encuentra en riesgo de pobreza o no utilizando como potenciales predictores del origen de dicha pobreza una serie de variables, que definiremos posteriormente, y hallando la matriz de confusión realizaremos una comparación con la matriz de confusión del modelo de regresión logística que hemos realizado con los mismos datos. Para comparar que modelo es mejor predictor para estos datos.

Deberemos de construir un árbol de decisión podado en el que nuestro mínimo se encuentra en 0.01(criterio CP) sobre el mismo realizaremos las representaciones gráficas pertinentes.

2.INTRODUCCIÓN

Nuestra base de datos se basa en un conjunto de encuestas realizadas por el INE, y las cuales se componen de 477 observaciones distribuidas en 18 variables. La variable a predecir que vamos a utilizar es *'HogarPobreza'* que es hogar en riesgo de pobreza, la convertiremos a tipo factor de dos niveles 0 y 1, o mejor dicho el cero corresponde a que el hogar en cuestión no está en una situación de pobreza y el uno quiere decir que sí que lo está. Utilizaremos el comando *'as.factor'* para convertir las variables a tipo factor.

En el campo del aprendizaje automático, hay diversas maneras de obtener los árboles, utilizaremos en esta ocasión CART**:** Classification And Regression Trees, cuya implementación es conocida como RPART(Recursive Partitioning and Regression Trees).

De manera general, lo que hace este algoritmo es encontrar la variable independiente que mejor separa nuestros datos en grupos, que corresponden con las categorías de la variable objetivo. Esta mejor separación es expresada con una regla. A cada regla corresponde un nodo.

3.Modelo de regresión logística

Como realizaremos la comparación con la matriz de confusión de este modelo hemos visto la necesidad de comentarlo brevemente.

El modelo anterior realizado fue el de Regresión logística, y el objetivo era el mismo y con las mismas variables, el objetivo era predecir la variable Hogar pobreza. Con ello construimos el modelo separando la parte de train(60% de los datos) y de test(el tanto por ciento restante). Reduciremos todas nuestras variables a las siguientes, AyudaFamilias (*) la cual se refiere a ayuda por familia/hijos en el año anterior, VacacionesOutDoor (***) si la familia se va de vacaciones o no, CapacidadAfrontar (***) si tiene capacidad de afrontar, LlegarFinMes(***) si pueden llegar a fin de més, Miembros (*)el número de miembros, HorasSemanales(***) que es las horas que trabaja la familia semanalmente y ActMayor (*) la actividad del mayor de 16 años, debido a que existen variables que no

nos aportan información o aportan mala información que nos distorsionaría el análisis las eliminaremos.

Dichas variables las escogimos mediante el test Anova que nos indica cuales son las variables significativas, haciendo caso a un criterio de varianza.

Realizando el test de Mcfadden nos muestra un resultado de 0.3789, buen resultado que nos indica que el modelo está bien ajustado. Estableciamos el umbral en 0.68 y con la matriz de confusión que nos aportaba una precisión del 74,34%. Finalmente en la conclusión compararemos la matriz de confusión obtenida por éste método con el siguiente.

4.Árboles de decisión

Primeramente, aunque ya lo hemos tenido que realizar para el anterior modelo, hemos tenido que hacer un tratamiento de los datos, es decir, limpiarlos y depurarlos para que sean representativos, y que nuestro modelo sea tenga sentido. Por tanto, utilizamos las variables definidas anteriormente por los motivos señalados previamente.

Ahora definimos nuestra muestra aleatoria para el aprendizaje del árbol(*set.seed(123)*) más tarde definimos la parte de train y test, que serán 60% y 40% respectivamente, como en el modelo anterior, cuyos tamaños son 286 observaciones para el train y 191 para el test.

Realizamos una tabla que nos arroja los siguientes resultados:
Esto nos indica que del 60% 182 familias no están en dicha situación y 104 están en riesgo de pobreza.

0	1
182	104

Y para la parte de test nos arroja estos resultados 182 familias no están en dicha situación
dicha situación y 104 están en riesgo de pobreza.

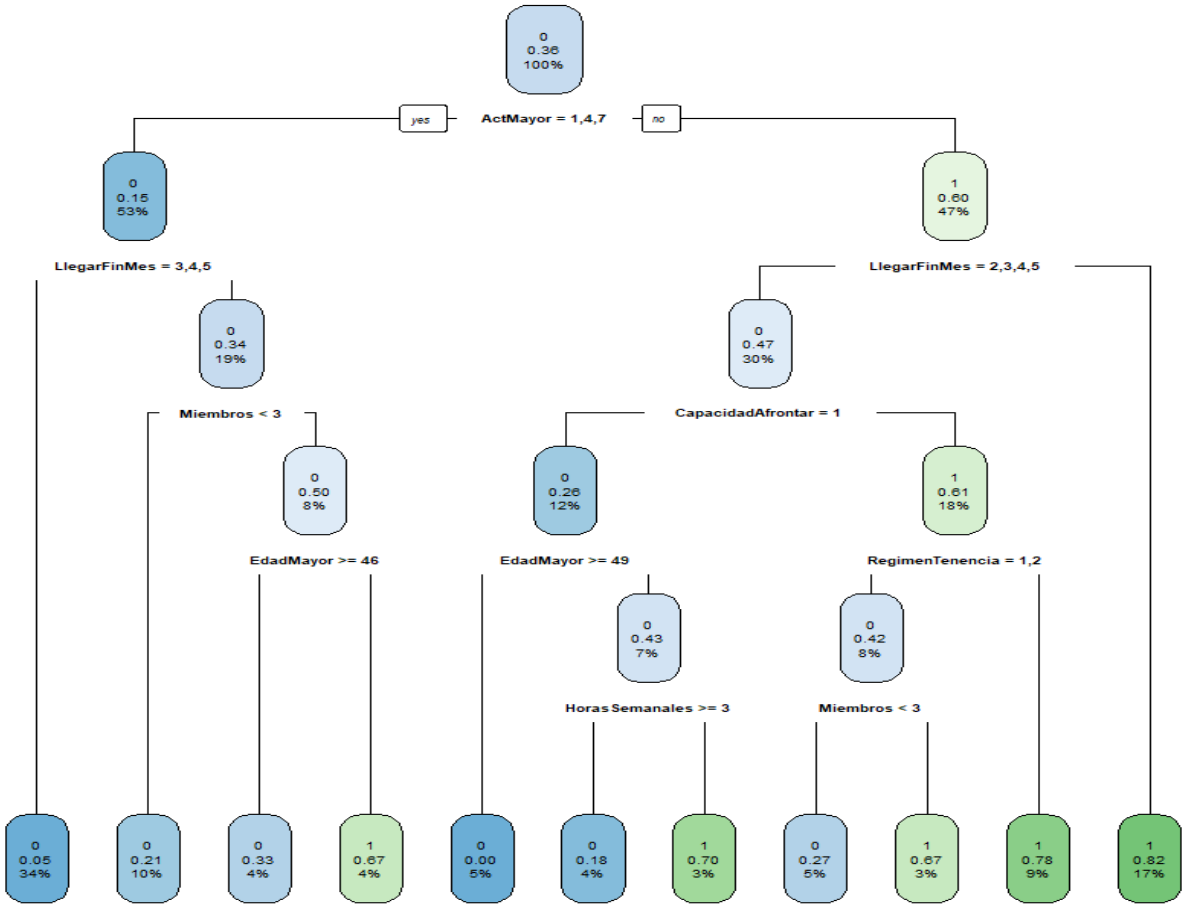
0	1
109	82

Si hacemos un *print* del árbol nos aparecerá el esquema del árbol de clasificación, cada inciso nos indica un nodo y la regla de clasificación que le corresponde. Siguiendo estos nodos, podemos llegar a las hojas del árbol, que corresponde a la clasificación de nuestros datos.

El nodo raiz nos ofrece información sobre el porcentaje de hogares en riesgo y hogares sin riesgo (i.e. 64% es un 0 osea que nos indica que no está en riesgo de pobreza y 33.1%

que si que está en riesgo de pobreza). Cada inciso nos indica un nodo y la regla de clasificación que le corresponde.

Todo lo anterior resulta mucho más claro si nos valemos de la visualización, así que creamos una gráfica usando nuestro modelo con la función rpart.plot() de *rpart.plot*.



Fuente: elaboración propia

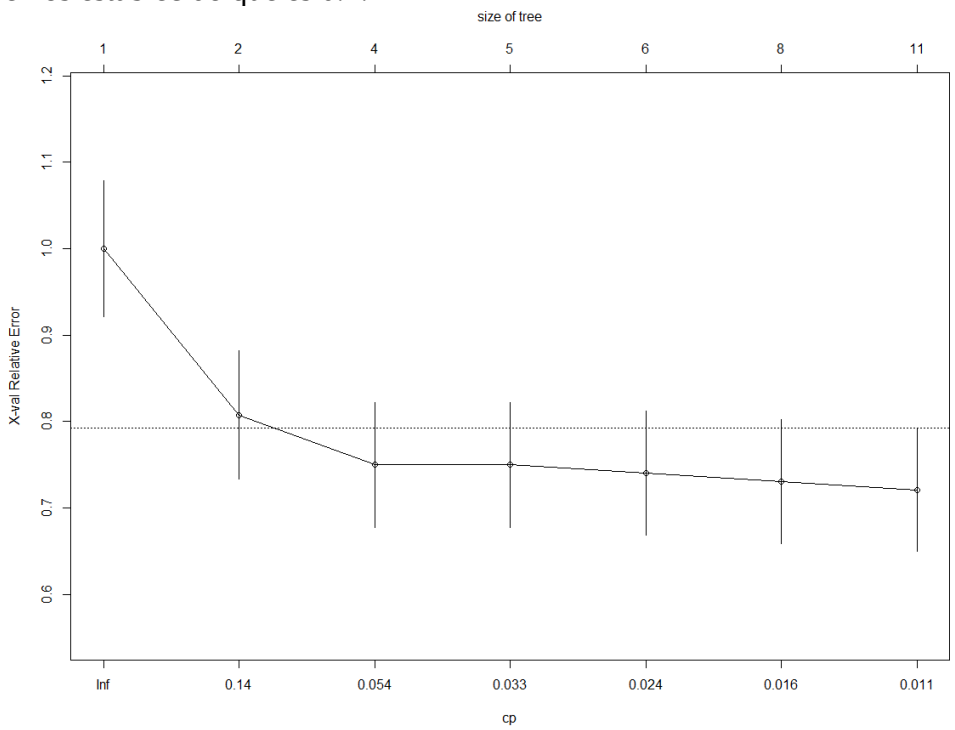
Cada uno de los cuadrados coloreados representa un nodo de nuestro árbol, con su regla de clasificación correspondiente, Cada nodo está coloreado de acuerdo a la categoría mayoritaria entre los datos que agrupa. Esta es la categoría que ha predicho el modelo para ese grupo.

Para los cortes de los nodos se ha llevado a cabo un criterio del parámetro de complejidad (CP), el cual, el parámetro de complejidad no es el error en ese nodo en particular. Es la cantidad por la cual la división de ese nodo mejoró el error relativo. El CP del siguiente nodo es solo 0.01 (que es el límite predeterminado para decidir cuándo considerar las divisiones).

Así que dividir ese nodo solo resultó en una mejora de 0.025 en el primero, en el segundo… Y por último una mejora del 0.01 por lo que la construcción del árbol se detuvo en el séptimo corte. Como se puede ver en el siguiente gráfico.Como podemos observar el árbol tiene 11 nodos terminales y 5 niveles que van abriéndose en función de las distintas variables como actmayor, edadmayor…Por ejemplo, el primer nodo se refiere a que la probabilidad de que un hogar esté en riesgo de pobreza es del 0.36, pero dentro de ello si tiene Actmayor = 1,4,7, si llegar a fin de mes = 3,4,5 si son menos de tres miembros y si la edad del mayor de la casa es >= 46 lo que pasará es que ese hogar tendrá 67% de probabilidades de estar en riesgo de pobreza y 37% de no estarlo.Cada nodo muestra la clase predicha (hogar pobre o no), la probabilidad prevista de pobreza, el porcentaje de observaciones en el nodo.

El parámetro de complejidad que elige R directamente es el de 0.01 que es el encargado de minimizar el error relativo y es el que utilizaremos para podar nuestro árbol.

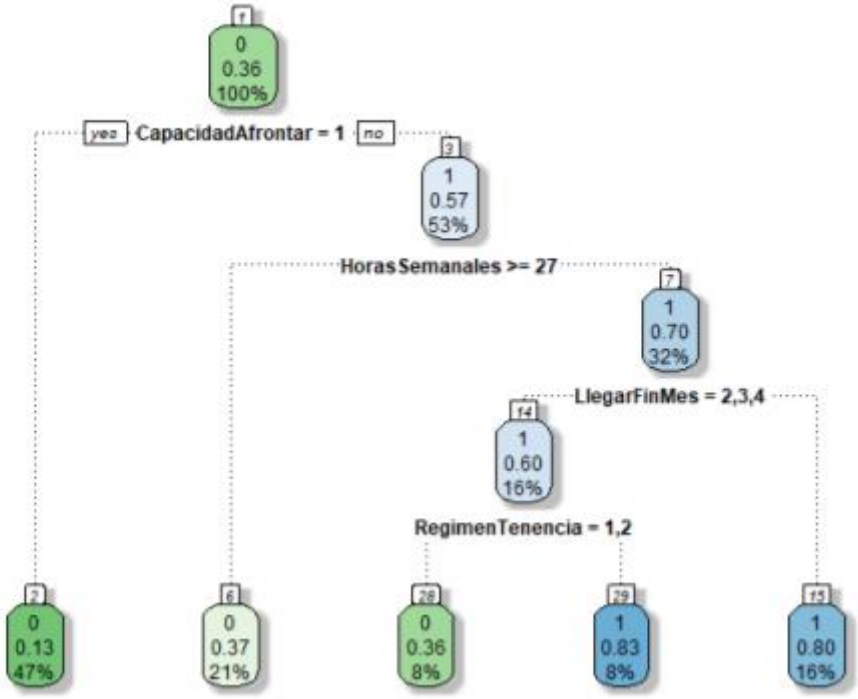
Que es lo que nos muestra el siguiente gráfico el tamaño del árbol en función del CP, el cual hemos establecido que es 0.1.



Fuente: elaboración propia

Por tanto, vamos a realizar la poda con dicho criterio, pero como R lo realiza de manera automática no apreciaremos ningún cambio en nuestro árbol de decisión, **El Over fitting**: en español sobre ajuste es uno de los problemas prácticos más comunes en la decisión de árboles de decisión, ese problema se soluciona realizando la poda de los árboles.

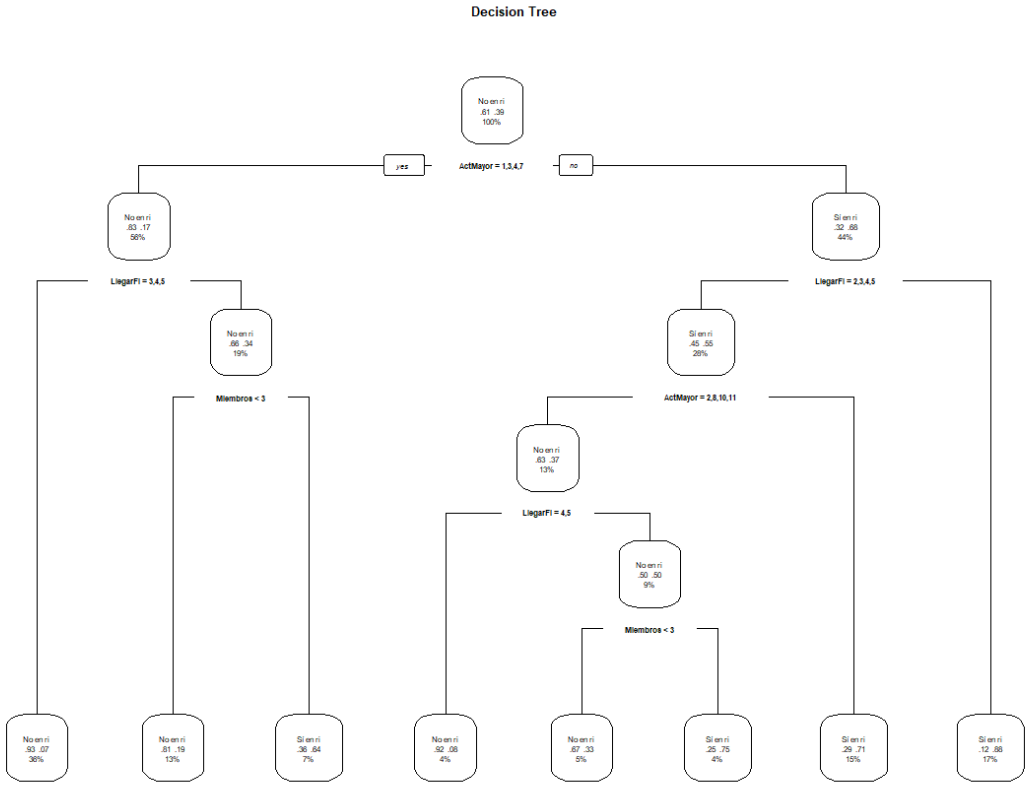
Nos fijaremos en la tabla complejidad y en la columna de XERROR y elegimos el menor valor que en nuestro caso será 0.7307 cuyo CP asociado será de 0.0288, por tanto este criterio que usaremos para podar el árbol.



Fuente:elaboración propia

Como podemo observar que el hogar esté en situación de pobreza dependerá de la capacidad de afrontar pago, un 53% de nuestra base de datos estarían relacionados con las horas semanales, llegar a finde mes y el régimen de tendencia en este caso.

Para comprobar la estabilidad de nuestro árbol de decisión realizaremos otro modelo, pero esta vez aumentando la muestra de entrenamiendo al 70% y realizando la poda con el mismo CP. Y estos son los resultados que nos arroja dicho modelo.

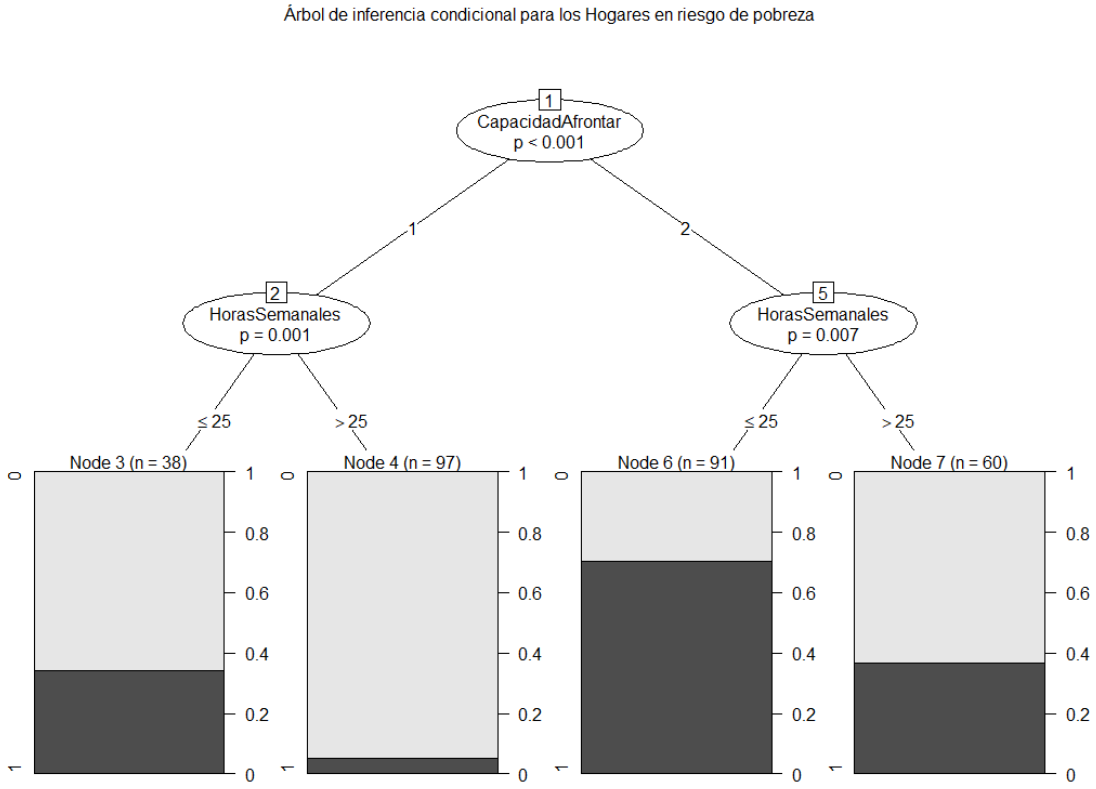


Fuente: elaboración propia

Como podemos observar la poda que hemos realizado, ha cambiado el número de nodos terminales de 11 a 9, y como se puede observar ahora en este gráfico nos aparece *'no en ri'* y '*si en ri',* no en riesgo de pobreza y si en riesgo de pobreza.

Por ejemplo, si no se dedican a Actmayor = 1,3,4,7 y si no suelen llegar a fin de mes = 2,3,4,5 estará en riesgo de pobreza, con una probabilidad del 17% del total. Ya que la suma de las probabilidades de los nodos terminales es el 100% de las probabilidades.

Por tanto, de los nodos finales la mitad están en riego de pobreza y la otra mitad no lo están a diferencia del árbol con la parte de entrenamiento del 60%, en el cual 6 no están en riesgo de pobreza y 5 si.



Los árboles que están basados en la inferencia, son una variante a tener en cuenta de los árboles de decisión tradicional. Los basados en la inferencia son similares a los tradicionales pero las variables y las divisiones se basan en la significatividad de algunos de los contrastes mas que en las medidas de puridad u homogeneidad. Lo que utiliza es el p-valor.

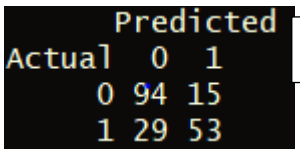
Este árbol utiliza la variable capacidad de afrontar como primer criterio clasificatorio y luego horas semanale, mayores o menores a 25. Acabando finalmente con 4 nodos terminales(3,4,6,7) de 38, 97, 91 y 60 observaciones respectivamente.

El 7 y el 3 son muy parecidos, a diferencia del 4 que es el menor y el restante que es el mayor.

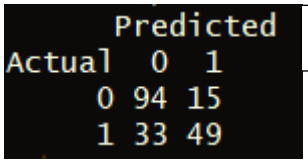
5.Conclusión

Concluyendo podemos comparar ambos modelos con la matriz de y sus métricas asociadas son parte fundamental de la "Caja de herramientas" del científico de datos, ya que, para saber qué modelo funciona mejor para un determinado problema, necesitamos métricas o herramientas que nos ayuden a evaluarlo. Todas las matrices de confusión obtenidas han sido en función de la variable dependiente 'HogarPobreza' y las demás las 12 variables seleccionadas definidas anteriormente.

Por tanto pasamos a comparar las dos matrices de confusión generadas por ambas librerías y observamos que hay diferencias en las matrices no son las mismas tanto la de la librería party como la de rpart.plot.



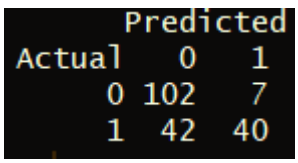
Matriz de confusión rpart.plot



Matriz de confusión party

Como podemos observar la proporción de falsos negativos, es decir, la parte que se ha predecido como negativo pero realmente es positivo es menor en rpart, 29 frente a 33, por tanto dentro de los dos modelos nos quedaremos con el que nos ofrece la librería rpart, ya que ha clasificado 4 individuos mejor que la otra.

Ahora la compararemos con el resultado obtenido en el modelo de regresión logística:



Como se puede observar no solo cuenta con cuatro individuos menos 142 frente a 147 si no que encima las predicciones son peores. La proporción tanto de falsos positivos como de falsos negativos ha disminuido con el modelo de árboles de decisión, esto quiere decir que para clasificar a una familia como 'en riesgo de pobreza' acertaremos más si predecimos los resultados con árboles en vez de con un modelo de regresión lineal.

[i] https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision

[ii] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176807&menu=ultiDatos&idp=1254735976608