1. EXECUTIVE SUMMARY

El objetivo de este informe es realizar una comparación entre los resultados arrojados a través del modelo de regresión logística y los árboles de decisión. Tendremos que construir un árbol de decisión podado que minimice los errores de validación cruzada, siendo nuestro mínimo de 0.01 y sobre este árbol realizaremos diferentes representaciones gráficas a través de la librería rpart.plot. Seguidamente realizaremos un árbol de inferencia a través de la librería party en donde no se hará necesaria la poda ya que trabaja con árboles de decisión no paramétricos. Siguiendo con nuestro análisis realizaremos una comparativa entre las matrices de confusión calculadas a través de los árboles de clasificación de rpart.plot y los árboles de inferencia de la librería party. Nos quedaremos con la matriz de confusión de los árboles de clasificación por ser más ajustada en cuanto a la predicción de falsos negativos y positivos reales. Finalmente realizaremos una comparativa entre la matriz de confusión de los árboles de clasificación y la matriz calculada a través del modelo de regresión logística en donde nos quedaremos con la matriz de confusión de los árboles de clasificación por tener una cantidad mayor de negativos reales y positivos reales y una cantidad considerablemente inferior de falsos negativos.

2. INTRODUCCIÓN

De los datos contenidos en el fichero que trabajamos en la anterior práctica tenemos 477 observaciones y 18 variables. Nuestra variable predictora va a ser “HogarPobreza”, la cual la tendremos que transformar en una variable categórica, en donde tomara el valor 0 y 1. El valor 0 quiere decir que el hogar en cuestión no está en una situación de pobreza y 1, por el contrario, sí que lo está.

3. MODELO DE REGRESIÓN LOGÍSTICA

Sobre la práctica anterior realizada eliminamos variables que no ayudaban a predecir la variable “HogarPobreza”, mejor dicho, que no eran estadísticamente significativas como para poder realizar nuestro análisis. Nuestros datos finalmente contenían las mismas observaciones que el data set original pero solamente 12 variables.

Sobre nuestros datos realizamos una partición de los mismos, en donde el train contenía el 60% de las observaciones y el test la parte restante. A través del modelo de regresión glm binomial sobre nuestra variable explicativa realizamos el test ANOVA con la Chi-cuadrado en donde vimos que las variables estadísticamente significativas eran: AyudaFamilias (*), VacacionesOutDoor (***), CapacidadAfrontar (***), LlegarFinMes(***), Miembros (*), HogaresSemanales(***) y ActMayor (*).

Realizando la bondad del ajuste con test McFadden nos arrojaba un resultado de 0.3789, en donde sabíamos que el modelo estaba ajustado. Establecíamos el umbral en 0.68 y partir de aquí calculábamos la matriz de confusión con una precisión del 74,34%.

4. ÁRBOLES DE DECISIÓN

4.1 Introducción

Al igual que en el ejercicio anterior, particionamos las observaciones por el mismo umbral de cara a realizar comparaciones futuras, esto es, para el train tendremos el 60% del total, siendo 286 las observaciones que vamos a trabajar y para el test el 40% del total, siendo 191 las observaciones.

Si representamos por medio de una tabla tanto el train como el test de aquellas personas que están y no están en situación de pobreza, tenemos lo siguiente:

```
  0   1
182 104
```

Para el train 182 familias no lo están y 104 por el contrario si

```
  0   1
109  82
```

Para el test 109 familias no lo están y 104 por el contrario si

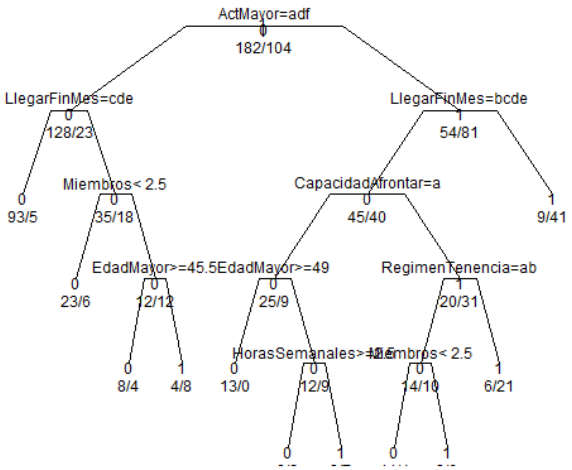
4.2 Análisis sobre los árboles de decisión

En primer lugar tendremos que trabajar con la librería “rpart”. Con nuestra variable endógena trabajaremos con el resto de las variables explicativas y construimos nuestro árbol de clasificación.

Las variables representadas son ActMayor, CapacidadAfrontar, EdadMayor, HorasSemanales,LlegarFinMes, Miembros y finalmente RegimenTenencia. Del total de

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

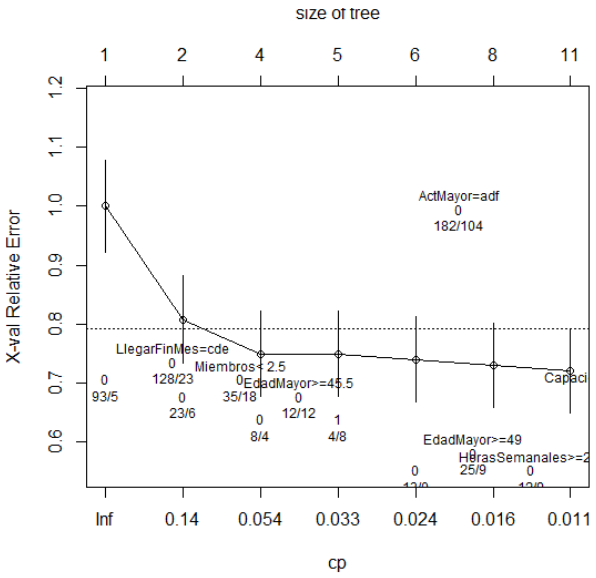
nuestras observaciones, las cuales son de 286 tenemos 104 de ellas que tienen errores en el nodo principal, siendo el ratio del 36,36%.



Observamos a través del comando summary, la tabla de complejidad paramétrica, tenemos los errores de validación cruzada

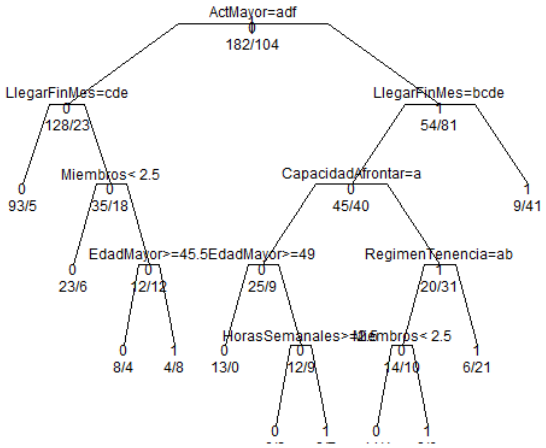
	CP	nsplit	rel error	xerror	xstd
1	0.25961538	0	1.0000000	1.0000000	0.07822328
2	0.07692308	1	0.7403846	0.8076923	0.07406260
3	0.03846154	3	0.5865385	0.7500000	0.07242068
4	0.02884615	4	0.5480769	0.7500000	0.07242068
5	0.01923077	5	0.5192308	0.7403846	0.07212771
6	0.01282051	7	0.4807692	0.7307692	0.07182904
7	0.01000000	10	0.4423077	0.7211538	0.07152461

Vamos a representar gráficamente nuestro error relativo por variables estadísticamente significativas y ver a su vez cúal sería el valor mínimo de validación cruzada para nuestra poda:



De nuestro análisis obtenemos que el mínimo es de un cp de 0.01

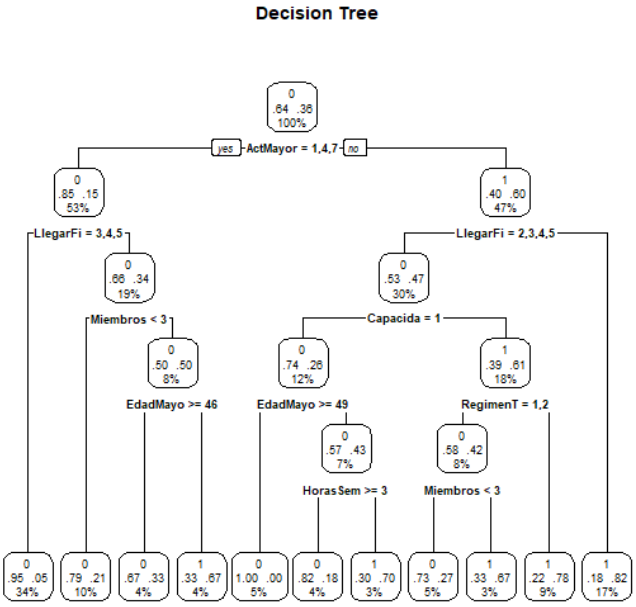
CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu



Esta sería la representación gráfica de nuestro árbol de clasificación una vez podado

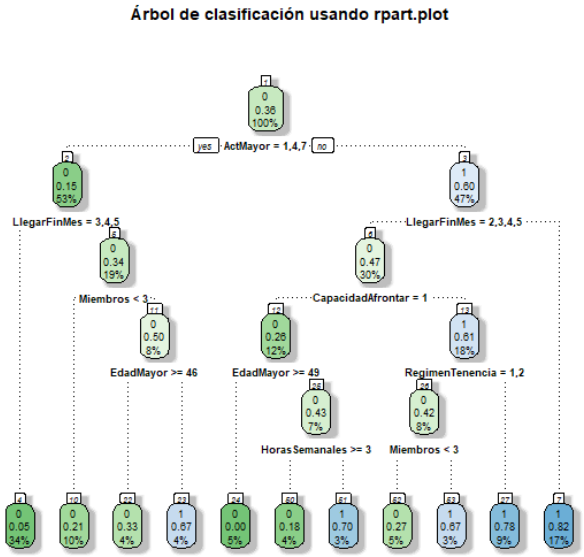
.

Siguiendo con nuestro análisis trabajaremos con la librería "rpart.plot", en donde representaremos la probabilidad de incurrir en pobreza utilizando nuestras variables antes nombradas.



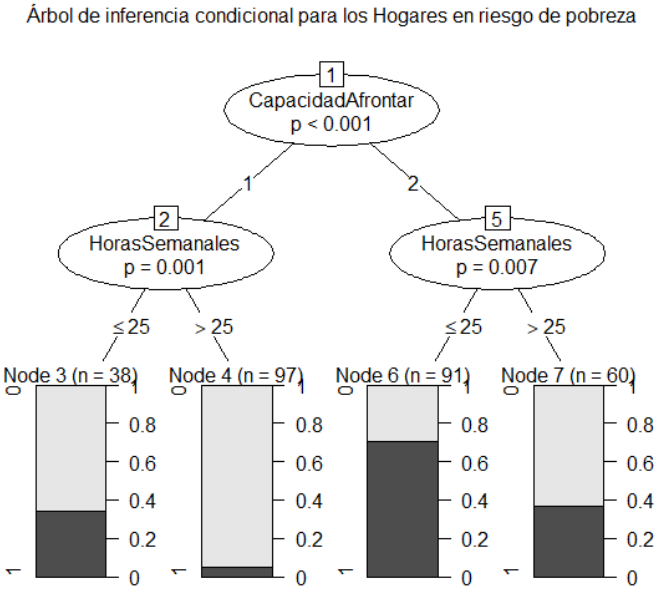
No llegar a Fin de mes puede suponer un aumento de incurrir en pobreza de un 34%.Por otra parte, la posibilidad de incurrir en este riesgo depende de los miembros familiares que existan en la unidad familiar, los cuales si son mayores a 3 la probabilidad de incurrir en pobreza aumenta.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu



Esta es la representación gráfica del árbol de clasificación usando la librería rpart.plot.Los resultados arrojados expresan la misma información que el árbol de decisión anterior.

También podemos utilizar la librería "party", la cual proporciona árboles de regresión no paramétricos. El crecimiento del árbol se basa en reglas estadísticas de parada, de esta forma no se hace necesaria la poda.



5. COMPARACIÓN ENTRE LOS RESULTADOS DEL MODELO DE REGRESIÓN LOGÍSTICA Y LOS ÁRBOLES DE DECISIÓN

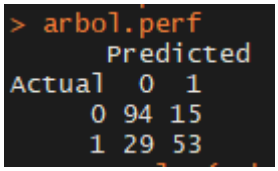
Como hemos visto el modelo de regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica, siendo esta la variable

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

HogarPobreza en función de las 12 variables seleccionadas que actúan como variables independientes o predictoras.

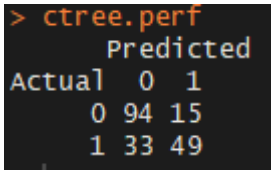
Por otra parte, los árboles de clasificación es un tipo de procedimiento no paramétrico de clasificación de una variable dependiente a partir de un conjunto de variables predictoras o explicativas cuyo objetivo principal es identificar las combinaciones de variables explicativas que predigan mejor la asignación de cada individuo a una determinada categoría. Como nuestra variable respuesta es categórica cada nodo terminal del árbol se le asigna una clase.

En primer lugar,vamos a comparar los resultados obtenidos de las matrices de confusión del árbol de clasificación utilizado por la librería rpart.plot y el árbol de inferencia utilizado por la librería party.



Matriz de confusión rpart.plot

De nuestra matriz de confusión obtenemos un accuracy del 76,96%.



Matriz de confusión party

De nuestra matriz de confusión obtenemos un accuracy del 74,86%.

Observamos que el número de negativos reales y falsos positivos es igual para ambas matrices. Sin embargo, la principal diferencia entre ambas estriba entre los falsos negativos y los positivos reales. Vemos que existe una cantidad ligeramente superior de positivos reales en la matriz del árbol de clasificación y una cantidad inferior de falsos negativos que en la matriz del árbol de inferencia. Tomado todo ello en su conjunto, la matriz de confusión que mejor explica nuestro modelo es la matriz explicada por el árbol de clasificación a través de la librería rpart.plot.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

Viendo que la matriz de confusión más ajustada es aquella que hemos cálculo a través del árbol de clasificación de la librería rpart.plot, vamos a compararla con la matriz de confusión que obtuvimos con el modelo de regresión logística.

```
> arbol.perf
       Predicted
Actual  0  1
     0 94 15
     1 29 53
```

Matriz de confusión rpart.plot

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

Matriz de confusión regresión logística

Para la matriz de confusión calculada a través del árbol de clasificación tenemos una cantidad de negativos reales de 94 y de positivos reales de 53,sumados todos ellos vemos que la cantidad asciende a 147 observaciones.

Por otro lado a través de la matriz de confusión calculada a través del modelo de regresión logística la cantidad de negativos reales es de 102 y de positivos reales de 40, sumados todos ellos vemos que la cantidad total es de 142, cifra ligeramente superior que con la matriz anterior. Esto quiere decir que la predicción de cálculo en cuanto aquellas familias en riesgo de pobreza arroja resultados más precisos utilizando el método de cálculo arrojado a través de los árboles de decisión.

Cuestión también digna de mención son los falsos negativos, con los árboles el resultado arroja una cifra de 29 observaciones y si lo calculamos con la regresión logística el resultado aumenta considerablemente hasta un total de 42, esto último quiere decir que el modelo está incluyendo a Hogares en situación de pobreza que realmente no lo están. Por otro lado, el total de falsos positivos resulta mejor para la regresión logística que para los árboles de clasificación, ya que en una, la cantidad de falsos positivos es de 7 y en la otra de 15.

Tomado todo ello en su conjunto, la matriz de confusión más ajustada en cuanto a nuestro modelo y en cuanto a nuestras predicciones es la matriz calculada a través de los árboles de clasificación.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

6. CONCLUSIONES

A través de este informe hemos podido realizar diferentes aproximaciones sobre los árboles de clasificación trabajando con diferentes librerías para comparar las matrices de confusión arrojados por diferentes métodos y finalmente hemos realizado una comparación exhaustiva sobre la matriz de confusión con el modelo de regresión logística y los árboles de clasificación, para finalmente optar por la matriz calculada a través de este último por ser más precisa y ajustada en cuanto a las predicciones de negativos y positivos reales.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu