Resumen Ejecutivo

Este informe tiene por objetivo una comparación entre los resultados obtenidos en el anterior informe (sobre regresión logística) y los árboles de decisión, método que se aplicará en este informe.

La idea principal es la construcción de árboles de decisión y aplicar una *poda* la cuál minimice los errores de validación cruzada. Además, se realizará el árbol de inferencia haciendo uso de otras librerías y recursos donde no será necesario establecer una poda dado que trabajaremos con árboles no paramétricos.

Será necesario seguidamente realizar una comparativa entre las matrices de confusión arrojadas por ambos árboles expuestos anteriormente para decidir cuál de ellas tiene más porcentaje de acierto y así quedarnos con uno u otro modelo.

Veremos que la primera matriz de confusión obtiene más aciertos en la predicción de positivos y negativos.

Una vez obtenida esta matriz de confusión será necesario compararla con la obtenida en el pasado informe (que usamos técnicas de regresión logística) para ver cuál de ellas obtiene una mejor representación.

Análisis Exploratorio de Datos

El conjunto de datos que se nos presenta consta de 477 observaciones y 18 variables basado en datos sobre condiciones de vida de 2016 extraídos del Instituto Nacional de Estadística.

El objetivo, como en el anterior informe, es predecir, esta vez a través de árboles de decisión, si un nuevo dato incorporado al *dataset* presentará o no riesgo de pobreza en el hogar.

Para ello antes debemos limpiar y preparar los datos. De las 18 variables iniciales vamos a descartar aquellas que no aportan información al modelo, siendo estas:

- ID del Hogar
- TV a color
- Renta total del ejercicio anterior
- Región
- Personas menores de 16 que perciben renta
- Sexo del mayor del hogar

Además, será necesario convertir las variables al formato correcto para poder llevar a cabo los análisis. Esto implica cambiar de factor a numérico en la mayoría de los casos. Además, no podemos olvidarnos de aquellas variables que debido a su naturaleza deben presentarse en tipo *factor*, entre las cuales encontramos:

- Vacaciones fuera de casa
- Capacidad de afrontar pagos
- Ordenador
- Llegar a fin de mes
- Régimen de tenencia
- Actividad del mayor del hogar

Modelo de regresión logística

Debemos recordar de manera muy resumida los datos que obtuvimos en el anterior análisis:

A través del análisis del pseudo de McFadden descubrimos que el valor arrojado era de 0.3789. Un resultado entre 0.2 y 0.4 se puede considerar bastante bueno por lo que podríamos continuar el modelo.

Posteriormente establecimos un corte para la muestra en 0.68 lo cual nos permitió construir una matriz de confusión con la siguiente forma:

	Predicted	
Actual	0	1
0	102	7
1	42	40

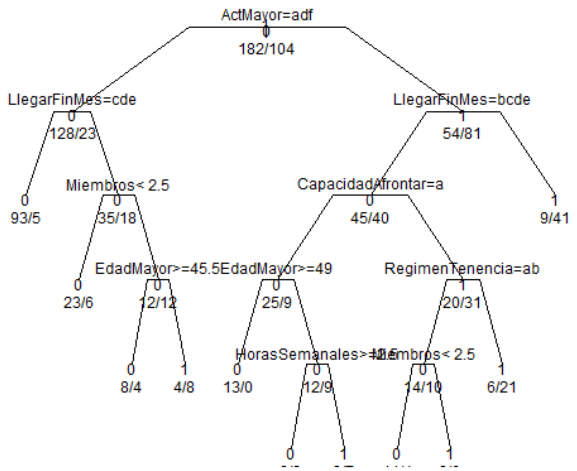
Fuente: Elaboración propia

Podemos interpretar esta matriz como la obtención de 142 valores correctos frente a 49 incorrectos. La precisión del modelo por tanto nos arrojaría un 74.34%, lo cual consideramos como aceptable para no incorporar más cambios en el modelo.

Árboles de decisión

Partiendo del análisis y limpieza que conseguimos para el modelo de regresión logística, podemos continuar construyendo los árboles de decisión. Hay que recordar que en el modelo anterior establecimos un conjunto de entrenamiento que contenía el 60% de los datos (286 observaciones en este caso) y un conjunto de prueba con las 191 observaciones restantes.

Obteniendo un árbol inicial observamos lo siguiente:



Fuente: Elaboración propia

Dado que la variable estudiada es *Hogar en riesgo de pobreza* el árbol está representando, desde los nodos superiores hacía abajo, las variables que más influencia tienen en nuestra variables principal, de más a menos.

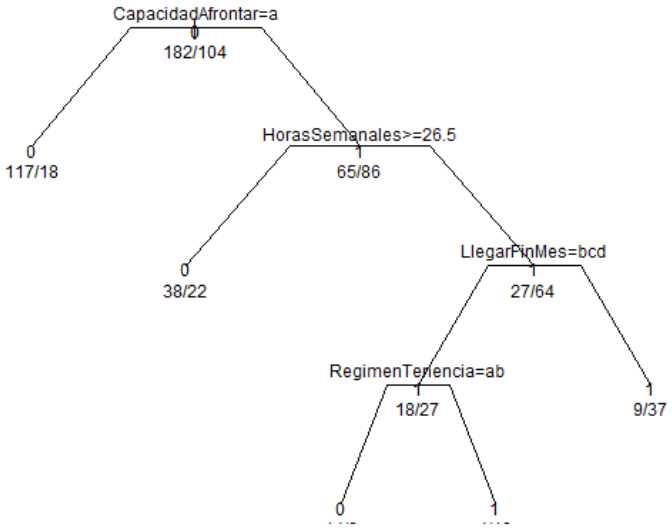
Observamos por tanto que *Actividad del mayor* es la primera partición, seguido de la *capacidad de llegar a fin de mes*. A cada rama se le está asociando un valor, menor conforme descendemos debidos a las divisiones que crea el árbol.

Ahora será necesario hacer una representación de la tabla de complejidad paramétrica que nos permitirá observar los errores de validación cruzada, pudiendo entonces escoger aquel que minimice este error para posteriormente realizar la *poda* del árbol.

	CP	nsplit	rel error	xerror	xstd
1	0.20192308	0	1.0000000	1.0000000	0.07822328
2	0.15384615	1	0.7980769	0.9711538	0.07771954
3	0.02884615	2	0.6442308	0.7307692	0.07182904
4	0.02403846	4	0.5865385	0.7884615	0.07353696
5	0.01923077	6	0.5384615	0.7884615	0.07353696
6	0.01602564	7	0.5192308	0.8461538	0.07505119
7	0.01000000	10	0.4711538	0.8076923	0.07406260

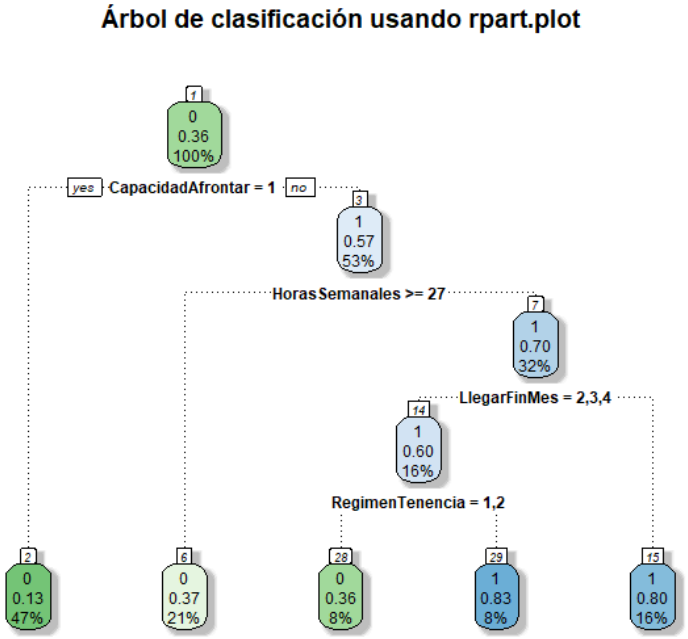
Fuente: Elaboración propia

A través del *xerror* elegimos el menor valor, en nuestro caso 0.7307 cuyo CP asociado será 0.0288. Hecho esto solo quedaría *podar* el árbol usando este valor y obtener una representación gráfica.



Fuente: Elaboración propia

Podemos elaborar un árbol un tanto más complejo que sea mas atractivo visualmente a la vez que mejore su interpretación radicalmente:



Fuente: Elaboración propia

La posibilidad de incurrir en pobreza dependerá en este caso de la capacidad para afrontar los pagos, siendo un 53% de los datos los que estarían condicionados además por las *Horas semanales, Llegar a fin de mes* y el *Régimen de tenencia* en último caso.

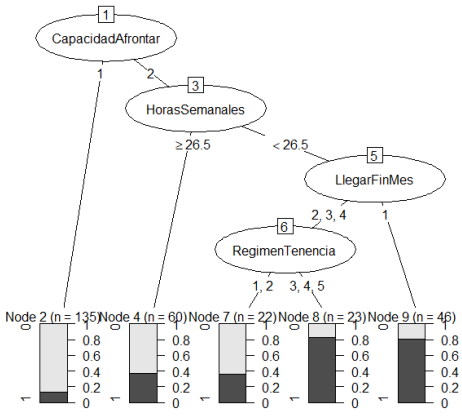
La matriz de confusión final que nos arroja este modelo después de la poda es la siguiente:

	Predicted	
Actual	0	1
0	103	6
1	48	34

Su interpretación es la misma que en anteriores casos: Los que han sido predichos de manera correcta ascienden a 137 mientras que los errores son 54.

Árboles de decisión no paramétricos

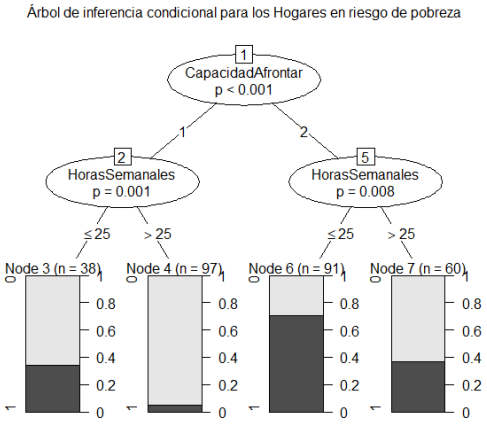
Prosiguiendo el análisis de los árboles de decisión ahora toca el turno de los no paramétricos. En este caso no hay que establecer una poda por lo que nos ahorramos gran parte del procedimiento anterior.



Fuente: Elaboración propia

La ventaja del uso de los no paramétricos reside en que el crecimiento del árbol se basa en reglas estadísticas de parada lo que hace que no necesitemos hacer uso de la poda.

A partir de *ctree* podemos construir un árbol condicional de inferencia usando nuestra variable principal *Hogar en riesgo de pobreza* sobre todas las demás variables, usando claro está, nuestros datos de entrenamiento. La representación gráfica para este sería la siguiente, donde vemos que el *p-valor* es lo suficientemente bajo como para considerar significatividad en el modelo.



Fuente: Elaboración propia

Ya solo nos quedaría realizar nuestra predicción en base a este modelo, la cuál nos arrojará los siguientes datos:

Actual	Predicted 0	1
0	94	15
1	33	49

Fuente: Elaboración propia

En resumen, hemos obtenido un total de 143 bien predichos frente a 48 erróneos.

Ahora sí podemos establecer la comparativa entre este árbol y el anterior, determinando que este es ligeramente menor por tener más aciertos y menos errores que el anterior.

Y finalmente y siguiendo el esquema expuesto en el resumen ejecutivo, podemos comparar estas matrices con la obtenida en la regresión logística que, como recordamos, es la siguiente:

Actual	Predicted 0	1
0	102	7
1	42	40

Fuente: Elaboración propia

Como vemos, estos resultados son muy similares a los obtenidos en nuestra última matriz de confusión, aunque con ligeros matices. En general, la matriz que mejor ajusta los resultados es la obtenida por los árboles no paramétricos ya que no introduce tanto error como podrían hacerlo las otras.

Conclusiones

Este informe tenía por objetivo extender el anterior, donde presentábamos un modelo de regresión para la predicción de la variable *Hogar en riesgo de pobreza*. En este caso se ha optado por el uso de árboles de decisión (incluidos los no paramétricos).

Hemos comenzado con un análisis exploratorio, transformando las variables que lo necesitaban a tipo factor y además eligiendo aquellas que de verdad aportaba información a nuestro modelo.

Hemos procedido a dividir los datos para obtener una muestra de entrenamiento y una de prueba sobre la que hacer las predicciones y obtener las matrices de confusión.

Finalmente hemos comparado las matrices del modelo de regresión linear con las de los árboles de decisión para concluir que aquel modelo que mejor predice una nueva entrada de datos son los árboles de decisión no paramétricos aunque la diferencia era mínima.