EXAMEN DE TÉCNICAS DE CLASIFICACIÓN FEBRERO DE 2019

Resumen ejecutivo

El objetivo de este informe es averiguar los factores que inciden en el consumo de carne por parte de las familias intentando desarrollar mecanismos de predicción de dicho indicador. La base de datos con la que se trabaja fue extraída de la Encuesta de Condiciones de Vida de 2017 y recoge la situación de 4220 familias respecto a once variables. En este informe se plantean dos técnicas para el estudio antes mencionado: la regresión logística y los árboles de decisión observándose finalmente la ventaja de esta segunda sobre la primera en ambos casos.

Previo a iniciar el estudio se procede a un breve análisis exploratorio de la base de datos. Las variables a predecir son cat2 y cat3 que representan la clasificación de los hogares según su gasto anual en consumo vacuno. Cat2 solo distingue si el consumo es bajo o no (variable binaria) mientras que cat3 distingue en consumo muy bajo, bajo-medio y medio-alto. En el ámbito de las variables empleadas para la clasificación se cuenta tanto con variables cuantitativas (edad, superficie en metros cuadrados de la casa e importe exacto de los ingresos mensuales netos totales del hogar en cientos de euros) como cualitativas. Entre las cualitativas se encuentran el tamaño del municipio (indicando si tiene o no más de 10.000 habitantes), la densidad de población (zona densamente poblada, intermedia o diseminada), el nivel de estudios (cuatro categorías desde inferior a secundaria hasta estudios superiores), la situación laboral (de nuevo cuatro categorías según el nivel de ocupación), el sexo y el régimen de tenencia (distingue entre pagando alquiler o hipoteca y el resto de casos). Si se desea conocer la codificación de las variables en la encuesta, consultar el diccionario de datos.

El análisis exploratorio (detallado en el código) muestra que la encuesta ha sido bastante trasversal (tanto superficie de la vivienda como edad y sueldo toman un rango de valores muy amplio) y las variables categóricas muestran reflejadas todas sus categorías con bastante significación. Además, respecto a las variables a clasificar la muestra está bastante balanceada (cat2 presenta 2210 casos de bajo y 2010 de otro y cat3 presenta 1472 casos de consumo muy bajo, 1473 de consumo bajo-medio y 1275 de consumo medio-alto) lo cual ayuda a construir una mejor clasificación. El código adjunto recoge además los cuartiles, la media, la mediana y la varianza de las variables numéricas así como un conteo de las apariciones de las distintas posibilidades de cada variable categórica.

Para cerrar el análisis exploratorio se observa que la variable que mide los metros cuadrados de la vivienda presenta 168 valores no definidos. La técnica aplicada en este estudio para lidiar con esta situación es sustituir dichos valores por la media de la variable. Una vez hecho esto se comprueba también que las variables numéricas no se encuentran correlacionadas.

Tras esto se procede a una pequeña adaptación de las variables para su trabajo con ellas; por una parte el análisis exploratorio ha permitido observar una gran diferencia entre las varianzas de las variables cuantitativas por lo que se procede a su tipificación. Además, se asegura que las variables categóricas se hayan cargado en R como factores y se comprueba una última vez que no haya NA's. Una vez se han comprendido y tratado los datos, se procede al análisis objeto de este estudio.

En un primer momento se plantea una regresión logística que empleé todas las variables para predecir cat2. Una vez construida se procede a refinar el modelo descartando las variables que no parecen significativas. A la hora de decidir si un modelo es mejor que otro se emplean dos criterios: criterio de parsimonia (entre dos modelos de capacidad clasificadora idéntica se elige el más simple, es decir, el que emplee menor número de variables) y el criterio de información de Akaike buscando minimizar su indicador.

Finalmente se alcanza un modelo que predice cat2 a partir de la densidad poblacional del municipio de residencia, el sexo de la persona, su situación laboral, el régimen de tenencia y tamaño de la vivienda y el importe exacto de los ingresos mensuales netos totales del hogar. Este modelo presenta un AIC menor que el modelo con todas las variables. Otro posible test para comprobar la bondad de este modelo es el R^2 de McFadden, que al ser menor en el modelo más refinado indica su superioridad respecto al general.

Este parece ser el modelo que devolverá los mejores resultados por lo que se va a profundizar sobre él. Se procede en primer lugar a un test ANOVA para determinar si todos sus coeficientes son significativos (es decir, si habría alguna variable de la que se pudiera prescindir sin perder demasiada información). El test devuelve que todos los valores son significativos por lo que si se prescindiera de alguna variable se estaría perdiendo información valiosa. Este mismo test nos permite mediante el número de desviaciones conocer la importancia de las variables que participan en el modelo. Así se observa que la variable más relevante es régimen de tenencia de la vivienda seguida de los ingresos mensuales y la situación laboral. En menor medida influyen la superficie de la vivienda, la densidad poblacional y el sexo de la persona.

El efecto de los coeficientes del modelo se explican mediante los odd ratios, así se aprecia que el coeficiente asociado a las variables numéricas es positivo por lo que mayores ingresos y viviendas indican una mayor probabilidad de un consumo no bajo de carne. Respecto a las categóricas se aprecia que a medida que se avanza por los valores de la situación laboral (coeficientes negativos) aumenta la probabilidad de tener un consumo bajo de vacuno. Lo mismo ocurre cuando se va de municipios poblados a más despoblados.

Finalmente se procede a la validación del modelo mediante el conjunto de test. Tras proceder a la predicción se obtiene la siguiente matriz de confusión:

	Predicho	
Real	**Bajo**	**No bajo**
Bajo	344	73
No bajo	75	352

Que tiene una exactitud asociada del 82.46%, una precisión de 0.828 y una sensibilidad de 0.824. Además se presenta a continuación la curva ROC (**Figura 1**) del modelo:

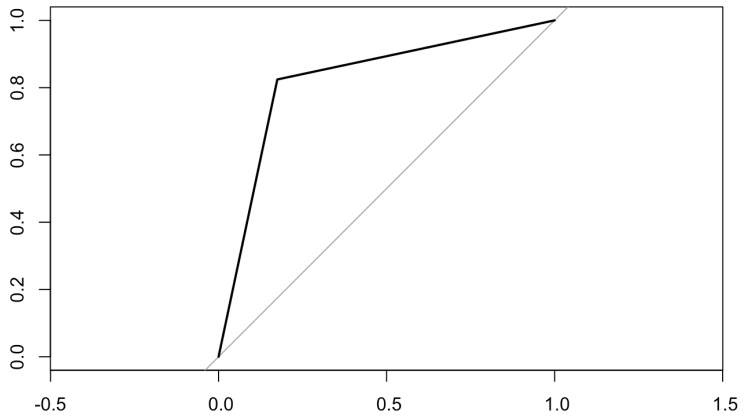

Figura 1. Curva ROC asociada a la regresión logística para cat2.

que tiene un área bajo ella de 0.8246 .

Tras esta primera aproximación se procede a emplear árboles de decisión, siguiendo el siguiente esquema: en primera lugar se construye un árbol de decisión a partir de todas las variables procediéndose tras ello a su poda. Se compara la bondad predictiva de ambos árboles (general y podado) obteniéndose matrices de confusión idénticas (no parece muy interesante mostrar dichas matrices en el informe pero se encuentran recogidas en el código). Tras ello se procede a la construcción de dos nuevos árboles (completo y podado) pero esta vez empleando las variables que la regresión logística consideró como significativas. Dichos árboles de nuevo devuelven la misma matriz de confusión y presentan la misma exactitud, 87,91% (superior a la de la regresión logística). La precisión de este modelo es de 0,859 y la sensibilidad de 0,911. Así se elige como mejor modelo el árbol con variables seleccionadas podado pues empatados en cuanto a capacidad clasificativa es el que mejor responde al Criterio de Parsimonia.

Nota. Aunque en el código se presenta un árbol de inferencia este modelo no se desarrolla en el informe porque presenta una exactitud ínfimamente mejor (del orden de la diezmilésima) y una interpretabilidad mucho peor.

Así pues a continuación se presenta (**Figura 2**) el árbol de decisión para la variable cat2:

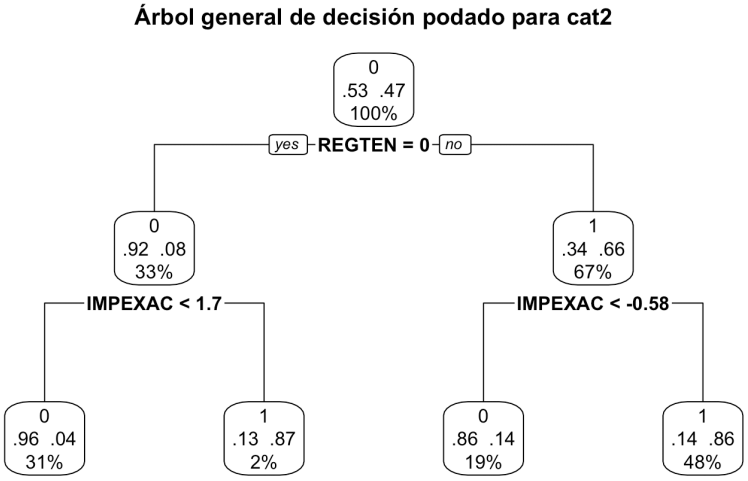

Figura 2. Árbol de decisión con variables seleccionadas para clasificación de cat2.

En este árbol la variable de mayor importancia es el régimen de tenencia. En caso de tratarse de una vivienda en alquiler o pagando hipoteca en un principio se considera que la persona tiene un consumo bajo de carne. Si los ingresos mensuales se encuentran por debajo de un umbral algo superior a la media (en torno a los 1500€ mensuales) se determina que la familia tiene un consumo muy bajo de carne (es interesante notar la pureza de 0.96 del nodo). En esta categoría se encuentra un 31% de la población. Si por el contrario los ingresos son superiores se asigna a la familia un consumo no bajo de carne. Por otra parte si el régimen de tenencia es propiedad sin hipoteca, cesión o renta antigua se asigna a esta familia un consumo no bajo de carne. Si además la familia se encuentra por encima de los 800€ de ingresos mensuales esta hipótesis se confirma mientras que si se encuentra por debajo, se asigna a la familia un consumo bajo de carne.

Así, se puede concluir finalmente que a la hora de predecir el consumo de carne como bajo o no bajo resulta mejor un árbol de decisión respecto a una regresión logística tanto en términos clasificatorios pues presenta mejor exactitud como en términos de intepretabilidad; cada familia se puede clasificar mediante solo dos preguntas. Luego en este caso el árbol de decisión reportaría mejores resultados en todos los sentidos que la regresión logística.

Por último se estudia a continuación la variable cat3. Para ello tras crear las oportunas muestras de entrenamiento y test se construye un árbol de decisión a partir de todas las variables. Este árbol se presenta a continuación (**Figura 3**):

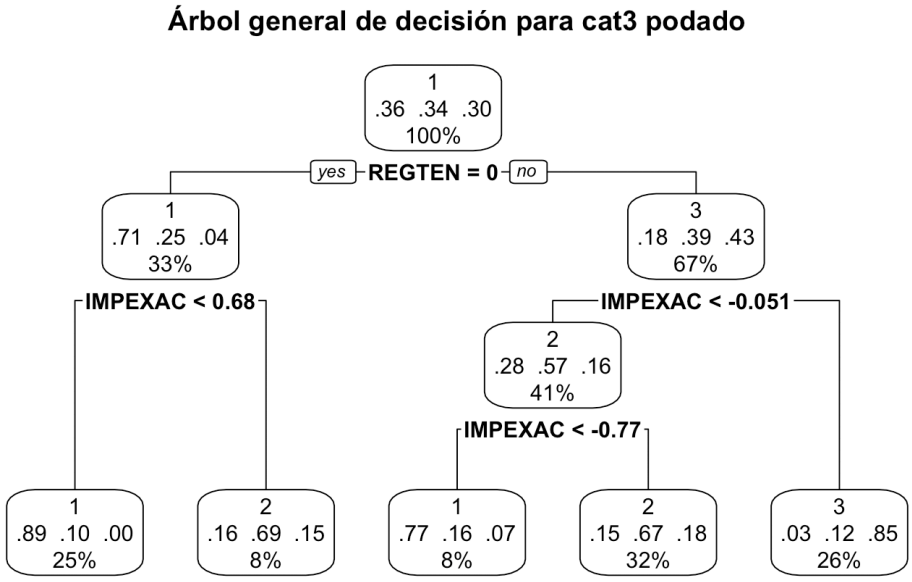

Figura 3. Árbol de decisión para cat3.

Este árbol de decisión reporta la siguiente matriz de confusión:

		Predicho	
Real	**Muy bajo**	**Bajo-medio**	**Medio-alto**
Muy bajo	317	88	16
Bajo-medio	57	348	55
Medio-alto	9	66	310

que tiene una exactitud asociada del 77,01%.

Cuando se procede a la selección de variable ocurre algo sorprendente, todos los resultados obtenidos devuelven exactamente este mismo árbol con los mismos nodos. Aun así y a modo didáctico se muestra en el código una manera de seleccionar variables (una regresión multinomial) y una justificación de la importancia de estas dos variables mediante una tabla de contingencia para el régimen de tenencia y un test ANOVA para los ingresos.

Respecto a la interpretación del modelo es análoga a la interpretación del primero si bien es cierto que a excepción del caso en el que las personas pagan alquiler/hipoteca y tienen unos ingresos inferiores el resto de nodos presentan una pureza mucho menor lo cual llevaría a una clasificación peor. Además, este árbol presenta una mayor profundidad y una menor exactitud luego es a todas luces peor que el empleado en la clasificación de cat2 (entendiendo que no son comparables en sí entre ellos pues cada uno clasifica una variable distinta).

Conclusiones

A la luz de los resultados obtenidos mediante todo el análisis previos se pueden deducir una serie de conclusiones:

- En primer lugar que la variable cat2 resulta mucho más sencilla de clasificar que la cat3. Esto puede deberse a su característica multinomial en lugar de binaria que requiere una mayor precisión.
- En segundo lugar que la clasificación de cat2 es mucho más robusta por el nivel final de pureza de sus nodos. Esto se refleja también en su exactitud (87,91%) casi once puntos porcentuales por encima de la de cat3 (77,01%).
- En tercer lugar debe notarse que el árbol obtenido para cat3 se podría reducir pues dos ramas surgen a partir de la variable ingresos mensuales.
- En cuarto lugar se observa que para cat3 la regresión multinomial y el árbol de decisión devuelven el mismo accuracy por lo que se elegiría en este caso el árbol de decisión por estar este construido solo a partir de dos variables y tener una sencilla interpretabilidad.
- Por último, es importante notar que en la clasificación de cat2 el árbol de clasificación devuelve unos resultados mucho mejores que la regresión logística (la exactitud es mayor (87,91% frente a un 82,46%) y su interpretación resulta terriblemente sencilla).

Código

En el siguiente enlace se encuentra el código sobre el que se ha construido todo este informe. Este código se encuentra en formato rmd detallándose cada paso seguido para el análisis de variables y la construcción de los distintos modelos.