Las aplicaciones del análisis de segmentación: El procedimiento Chaid

MODESTO ESCOBAR*
Universidad de Salamanca
Instituto Juan March de Estudios e Investigaciones

1. INTRODUCCIÓN [1]

Cuando en un cuestionario se desea explicar por qué los entrevistados dan contestaciones distintas a las preguntas, se construye una serie de tablas que permiten ver la asociación existente entre unas y otras variables. No es cuestión de cruzar cada pregunta con el resto, sino de seleccionar una serie de hipótesis plausibles con el conocimiento previo, teórico o empírico, de la realidad que se está investigando, y, de acuerdo con ellas, realizar los análisis que pongan a prueba las conjeturas. Una manera de facilitar la tarea de selección de variables relevantes en la explicación de la contestación a una pregunta dada es la técnica del análisis de segmentación, que proporciona además una descripción de las diferencias que los distintos grupos de una muestra pueden presentar en un determinado rasgo. Es ésta una técnica de dependencia entre variables. En su uso, se distinguen, por un lado, una variable cuya distribución se desea explicar y, por el otro, un conjunto de variables, nominales u ordinales, con estatus de independientes. Éstas reciben el nombre de pronosticadoras y tienen la finalidad de conformar grupos que sean muy distintos entre sí en la variable dependiente [2].

* Modesto Escobar es Catedrático de Sociología en la Universidad de Salamanca y profesor permanente del Centro de Estudios Avanzados en Ciencias Sociales del Instituto Juan March de Estudios e Investigaciones.

[1] Este artículo es una versión actualizada del *Working Paper*, número 31, publicado por el Centro de Estudios Avanzados en Ciencias Sociales del Instituto Juan March de Estudios e Investigaciones de enero de 1992. La principal novedad que incorpora es la introducción del programa *Answer Tree*, que permite el tratamiento específico de variable dependiente ordinales y de intervalo o razón, además de otros algoritmos no expuestos en estas páginas.

[2] A menudo se confunde esta técnica con el análisis de conglomerados. Aunque las funciones clasificadoras son muy similares, se distinguen fundamentalmente en dos aspectos: a) La segmen-

Póngase como ejemplo que se desee describir en un pueblo pequeño quién lleva un determinado tipo de ropa. Para simplificar, tómese una prenda muy fácil de segmentar como es la falda. Entre las posibles variables que mejor pueden explicar quién la lleva y quién no, no es difícil reconocer que es el sexo el mejor pronosticador, pues prácticamente ningún hombre usa este tipo de prenda. La ejecución de la segmentación implicaría no contentarse con una sola variable y buscar otras que ayuden a distinguir mejor a los distintos usuarios de estas ropas. Es evidente que si ningún hombre la usa, este grupo es totalmente homogéneo en esta variable y, por tanto, no procede seguir con la segmentación. Pero en el caso de las mujeres, sí se pueden encontrar nuevas variables que nos distingan grupos diferentes en uso de ropa. Parece claro que la edad juega un papel importante: es bastante difícil ver a mujeres mayores con pantalones, mientras que entre las jóvenes el uso de éstos es muy habitual. Por tanto, si no se introducen nuevas variables, la población del pueblo quedaría segmentada en tres grupos: el de los hombres, donde nadie usa faldas; el de las mujeres jóvenes, con un porcentaje medio de portadoras de esta prenda, y el de mujeres mayores, cuya probabilidad de verlas con faldas es muy alta.

Otro símil que puede resultar útil en la comprensión de la segmentación es el de una tarta que hay que repartir entre varias personas. Imaginando que es un pastel con dos sabores –nata y chocolate, por ejemplo–, una segmentación adecuada sería la que partiera el dulce en dos trozos de gusto homogéneo. Se trataría, por tanto, de realizar un corte que permitiera dar a uno de los comensales el trozo con sabor a nata y a otro el de chocolate. En resumidas cuentas, la segmentación permite dividir una muestra de modo que queden grupos de contenido uniforme muy distintos entre ellos.

El análisis de segmentación fue concebido y debe ser utilizado principalmente con una finalidad exploratoria. La razón radica en que su mecanismo consiste en la búsqueda de las mejores asociaciones de las variables independientes con la dependiente. Su potencia, al mismo tiempo que su peligro, reside en la selección automática de aquellas categorías que pronostican mejor los valores de la variable considerada objetivo. Además, segmentar significa dividir y, en consecuencia, permite que se hallen grupos muy distintos en un determinado aspecto. De este modo, las muestras quedan fragmentadas en distintos tipos de personas u objetos cuya descripción constituye un objetivo adicional de esta técnica.

El propósito de este artículo es presentar y explicar sin demasiados ambages estadísticos y a través de distintos ejemplos, unos reales [3] y otros simulados, la

---

tación trabaja para la clasificación con grupos de sujetos (hombres, mujeres, jóvenes, personas de izquierda, practicantes de una determinada religión, solteros, casados...), seleccionando a aquellos que presentan características significativamente muy distintas en una o varias variables dependientes; el análisis de conglomerados trabaja con individuos, agrupando o distinguiendo a éstos en función de sus valores en un conjunto de variables. b) En el análisis de conglomerados no hay distinción entre variables dependientes e independientes, sino que todas ellas, con mayor o menor peso, sirven para clasificar a los sujetos; en el análisis de segmentación es necesario distinguir entre la variable dependiente que se desea explicar y las posibles variables independientes que puedan dar cuenta de ella.

[3] El ejemplo cuya variable dependiente es el aborto se realizó en octubre de 1990 por el CIRES con una muestra de 1.200 individuos extraídas de la población española con más de 18 años. El

lógica de esta técnica de análisis multivariado. Con este fin, se expondrá el análisis de segmentación a través de uno de sus algoritmos basado en el estadístico χ^2, especialmente indicado cuando la variable dependiente es de tipo nominal. Se procederá a explicar los pasos lógicos de esta técnica: reducción de categorías, selección de pronosticadores y detención de la segmentación. A continuación, se ofrecerán varios ejemplos de cómo se interpretan los resultados de la técnica de segmentación.

2. LA LÓGICA DEL ANÁLISIS DE SEGMENTACIÓN. EL ALGORITMO CHAID

Tradicionalmente, el análisis de segmentación se ha reducido al estudio de variables dependientes cuantitativas, utilizando el algoritmo presentado por Morgan y Sonquist (1963). Aquí, sin embargo, se centrará la atención en una derivación de esta técnica que se distingue por utilizar, en lugar de la suma cuadrática intergrupos, el estadístico χ^2 para la selección de los mejores pronosticadores [4]. De esta forma, se detendrá esta exposición en aquellos casos con variable dependiente medida en escala nominal [5].

Los pasos lógicos que deben seguirse para realizar esta tarea son los siguientes:

a) Preparación de las variables. Tarea del analista, que debe seleccionar una variable dependiente que sea de interés para el análisis y elegir un conjunto de posibles pronosticadores relevantes (variables nominales, ordinales con pocas categorías, preferiblemente menos de diez, o incluso variables cuantitativas convertidas en discretas [6]) que permitan realizar una descripción y pronóstico óptimo de la primera variable.

b) Agrupación de las categorías de las variables independientes en el caso de que éstas tengan un perfil similar de la variable dependiente.

---

empleado con los jóvenes de Burgos fue dirigido por el autor y realizado por el Departamento de Sociología de la Universidad de Salamanca a una muestra de 1.015 sujetos de aquella ciudad con edades comprendidas entre los 14 y los 30 años. El programa informático que se ha utilizado es el *Answer Tree* (versión 1.0), módulo independiente del SPSS, que sustituye al más antiguo *CHAID for Windows* e incorpora dos nuevos algoritmos: el *C & RT* de Breiman et al. (1984) y el *QUEST* de Loh y Shih (1997).

[4] Este estadístico χ^2 puede calcularse bien mediante la suma en una tabla de todos los residuos estandarizados al cuadrado (Pearson), bien utilizando la razón de verosimilitud. Las fórmulas respectivas de uno y otro, que arrojan resultados distintos aunque similares, son:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(f_{ij}-f_{ij}^*)^2}{f_{ij}^*} \qquad L^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij}\ln\left(\frac{f_{ij}}{f_{ij}^*}\right)$$

[5] El algoritmo CHAID puede utilizarse también con variable dependiente cuantitativa en cuyo caso, en lugar de utilizar el estadístico χ^2 tendría que emplearse la razón entre la media cuadrática externa y la interna (F) con su correspondiente grado de significación correspondiente a la distribución de la F de Snedecor. Del mismo modo, también pueden realizarse análisis de segmentación específicos con variable dependiente ordinal, pudiéndose dar a cada valor de esta variable una puntuación *(score)* que refleje la distancia entre categorías (Magidson, 1993).

[6] Para utilizar variables de intervalo o de razón como pronosticadoras hay que convertirlas en variables discretas. El procedimiento se explica más adelante.

c) Primera segmentación, que consiste en la selección de la variable que mejor prediga la variable dependiente.

d) Segunda segmentación. Para cada segmento formado en el paso anterior, se busca entre las variables cuyos valores han sido previamente agrupados de la misma forma que en el paso b), la que tenga mayor poder pronosticador.

e) Sucesivas segmentaciones. Se procede de forma similar al paso anterior en cada grupo formado por la segmentación previa.

Supóngase que se quieren formar grupos homogéneos, también llamados en este contexto segmentos, respecto de la aprobación del aborto en el supuesto de que un matrimonio no desee tener más hijos. Esta será la variable dependiente, con tres posibles valores: «lo aprueba», «lo desaprueba» y «no sabe/no contesta». Para formar grupos homogéneos con esta técnica, se ha de elegir una serie de características medidas nominal u ordinalmente. En este caso, por ejemplo, sexo («hombre», «mujer»), edad [7] («menos de 46 años», «más de 45»), e ideología («izquierda», «centro», «derecha»).

TABLA 1

Cruce de opinión ante el aborto según sexo y edad

			SEXO											
			Varón						Mujer					
			EDAD						EDAD					
			≤ 45			> 45			≤ 45			> 45		
			IDEOLOGÍA			IDEOLOGÍA			IDEOLOGÍA			IDEOLOGÍA		
		Total	Izq.	Cent.	Der.	Izq.	Cent.	Der.	Izq.	Cent.	Der.	Izq.	Cent.	Der.
Posición ante el aborto	Sí	19,5%	42,7%	21,5%	21,1%	27,6%	11,7%	12,5%	29,5%	20,6%	16,7%	14,8%	8,9%	3,2%
	No	75,3%	54,7%	72,6%	73,7%	72,4%	81,5%	79,2%	62,5%	74,9%	77,8%	85,2%	84,2%	96,8%
	NC	5,3%	2,6%	5,9%	5,3%		6,8%	8,3%	8,0%	4,5%	5,6%		6,9%	
Total		(1.200)	(117)	(186)	(19)	(58)	(162)	(24)	(88)	(223)	(18)	(27)	(247)	(31)

En la tabla 1 se pueden contemplar 12 segmentos (columnas) distintos formados por el cruce de las categorías de las tres variables pronosticadoras (2 de sexo por 2 de edad por 3 de ideología). Cada uno de ellos está caracterizado por un tamaño (fila correspondiente al total) y tres porcentajes relativos a cada uno de los valores de la variable dependiente, en este caso, posición ante el aborto.

El segmento más numeroso es el correspondiente a las mujeres mayores de 45 años (n = 247), seguido por el de las mujeres jóvenes de la misma ideología (223). Los grupos de hombres con más componentes son el de los jóvenes de

[7] Este es un claro ejemplo de variable de razón convertida en discreta. Aunque en la encuesta se midió de modo directo, recogiendo valores entre los 18 y los 92 años, en este análisis se han dicotomizado de modo discrecional sus valores. El programa *Answer Tree* incorpora sendos modos, automático y manual, de agrupar valores de las variables pronosticadoras cuantitativas.

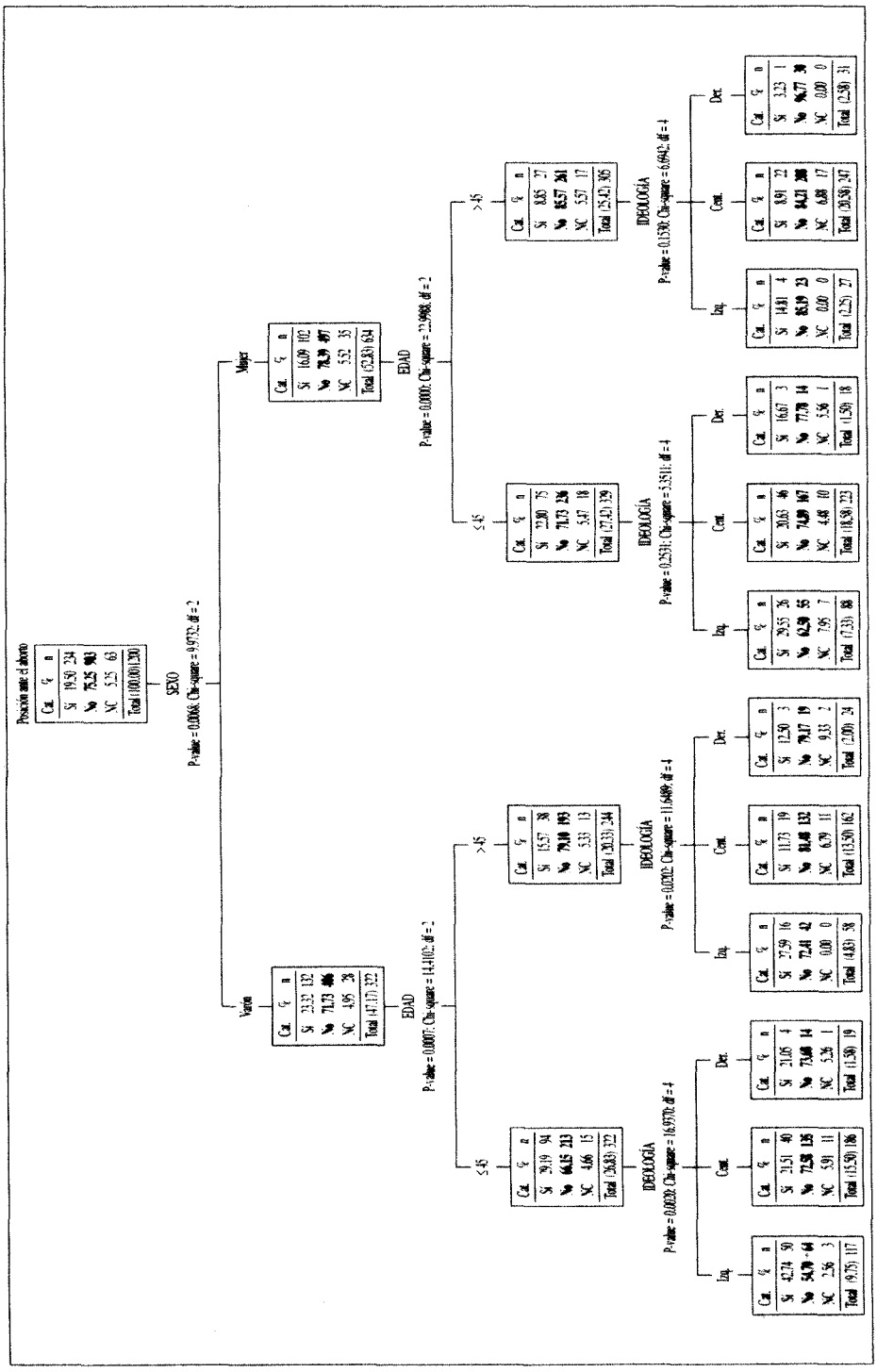

Figura 1. Pseudo-segmentación de la opinión ante el aborto.

centro (186) y el de los hombres con más de 45 años de la misma ideología (162). Por el contrario, los grupos menos numerosos son el de los hombres y el de las mujeres jóvenes de derecha (19 y 18 sujetos respectivamente). Al observar la variable dependiente (el porcentaje de los que aprueban el aborto) se obtiene un perfil distinto para cada uno de los 12 segmentos formados por las tres variables pronosticadoras: Los más favorables al aborto (42,7%) son los jóvenes varones de ideología de izquierdas y el grupo con menor porcentaje de sujetos que aprueban esta práctica (3,2%) es el de las mujeres con más de 45 años e ideología de derechas.

La tabla 1 es en realidad una tabla de contingencia o cruce formado por cuatro variables dispuestas en cuatro dimensiones. La técnica de la segmentación tiene una estructura similar. En la figura 1 se muestra una pseudo-segmentación, basada en los datos de la tabla mencionada. En cada rectángulo se incluye el valor de la variable pronosticadora que conforme el segmento determinado, la distribución de frecuencias de la variable dependiente correspondiente al grupo en cuestión, y el número de casos que lo forman. Las cifras incluidas en los 12 rectángulos de la base inferior de la figura son idénticas a la de los porcentajes y totales de las 12 columnas de la tabla. Sin embargo, este árbol no es una verdadera segmentación porque las divisiones no se han realizado de forma automática, ni jerárquica, ni se han efectuado con el criterio de significación estadística.

Hay variados procedimientos para llevar a cabo la segmentación. A continuación se presenta con mayor detalle el algoritmo llamado CHAID (Chi-squared Automatic Interaction Detection). Esta técnica, desarrollada por Cellard et al. (1967), Bourouche y Tennenhaus (1972), Kass (1980) y Magidson (1989, 1993a y 1993b), quien la ha adaptado para el SPSS, tiene como principal característica distintiva de otros algoritmos de segmentación el que la muestra no se segmente de modo binario, o dicho de otro modo, el que se pueden formar segmentos con más de dos categorías al unísono. Al igual que otras prácticas de segmentación, las operaciones elementales que realiza son: a) la agrupación de las categorías de las variables pronosticadoras; b) la comparación de efectos entre distintas variables, y c) la finalización del proceso de segmentación.

2.1. Reducción de las categorías más discriminantes de cada pronosticador

Este primer paso consiste en seleccionar las categorías de las variables pronosticadoras que realmente discriminan a los sujetos en la variable dependiente. Suponiendo que una determinada variable tuviera c valores, se trata de convertirlos a un número $k \leq c$ que reduzca la complejidad de la segmentación sin pérdida sustancial de información.

Se puede optar por tres modalidades de reducción según sean las características de las variables pronosticadoras:

1) Variables nominales: Cada valor de la variable pronosticadora puede ser agregado a cualquier otro valor de la misma variable. Sea, por ejemplo, la variable situación ocupacional con los valores «ocupado», «parado» e «inactivo». De cara a la formación de grupos, la categoría «ocupado» podría formar grupo con

«parados» y/o «inactivo». La primera categoría es contigua, pero la segunda no lo es. Este procedimiento también se denominaba libre *(free)*.

2) Variables ordinales: Un valor de la variable sólo puede ser agregado a otro si es contiguo en la escala. En el procedimiento anterior, la categoría «ocupado» sólo podría unirse en un primer momento con la categoría «parado». Los «inactivos» podrían agregarse con los «parados»; pero no con «ocupados». Este procedimiento también se conoce con la denominación de monótono [8]. Un ejemplo de pronosticador monótono adecuado es el nivel de estudios. Si esta variable tuviera como valores «primarios», «secundarios» y «universitarios», el procedimiento permitiría la fusión de las categorías primera y segunda o segunda y tercera, y descartaría la posibilidad de formar un grupo compuesto por sujetos con estudios primarios y universitarios.

3) Variables ordinales con valores perdidos: Es similar a la opción anterior, pero permite un mayor grado de libertad, por cuanto un valor, generalmente el «no sabe, no contesta», puede agregarse libremente a cualquier grupo. Si la variable nivel de estudios tuviera el valor «Ns/Nc», con este procedimiento, también denominado flotante *(float)*, los sujetos que no contestasen podrían agruparse con cualquiera de las tres categorías establecidas.

4) Variables cuantitativas: Las variables cuantitativas para ser utilizadas en el procedimiento CHAID tienen que ser recodificadas en valores discretos [9] y tratadas como si fueran ordinales.

El funcionamiento de formación de grupos de categorías homogéneas se basa en el estadístico χ^2. Los pasos son los siguientes:

1) Se forman todos los pares posibles de categorías. Esto dependerá de la opción que se haya preferido dar a un determinado pronosticador. Así, en la variable *práctica religiosa*, que presenta cinco valores, el número posible de pares sería de 10 (combinaciones de 5 elementos tomados de dos en dos). Si se opta por la opción ordinal, los pares posibles son 4 (número de categorías menos una) sin considerar los valores perdidos. Y si se escogiese la opción ordinal con valores perdidos, las posibilidades serían 7 (dos veces el número de categorías menos 3). Véase tabla 2.

2) Para cada posible par se calcula el χ^2 correspondiente a su cruce con la variable dependiente. El par con más bajo χ^2, siempre que no sea significativo [10],

[8] Este método no debería aplicarse automáticamente a toda variable ordinal. Si no existe una pauta de relación lineal del pronosticador con la variable dependiente, las variables ordinales han de tratarse con el procedimiento sin restricciones. Por ejemplo, si se espera que mayores *estatus* no conlleve opinión más favorable al aborto, sino que lo más probable es que los sujetos con altos y bajos *estatus* sean los menos (o más) inclinados a aprobar esta práctica y los más (o menos) favorables sean los de *estatus* medios, entonces la variable no debería ser tratada como ordinal.

[9] Si tienen menos de 11 valores, el programa *Answer Tree* considera las variables de intervalo o razón discretas y les da el mismo tratamiento que si fueran ordinales. Caso de que tengan más de diez valores, éstos son agrupados en intervalos que son tratados como categorías. Aunque el usuario puede definir a su guisa la agrupación de los valores de la variable cuantitativa, el programa lo puede hacer automáticamente agrupando en cada categoría un número similar de casos.

[10] Todos los cruces tienen el mismo número de grados de libertad porque la variable dependiente es la misma para todos los contrastes y la variable independiente sólo tiene dos valores, pues recuérdese que se está trabajando con pares de categorías.

formará una nueva categoría de dos valores fusionados. La condición de que no sea significativo es muy importante porque, caso de que lo fuese, indicaría que las dos categorías que se pretenden fusionar no lo pueden hacer, ya que son heterogéneas entre sí en los valores de la variable dependiente y el objetivo es justo lo contrario, asimilar categorías con comportamiento semejante.

TABLA 2

Pares posibles de la variable *práctica religiosa* según se considere nominal u ordinal

Nominal	Ordinal
Muy practicante-Medio practicante	Muy practicante-Medio practicante
Muy practicante-Poco practicante	Medio practicante-Poco practicante
Muy practicante-Nada practicante	Poco-Nada practicante
Muy practicante-Ns/Nc	Nada practicante-Ns/Nc
Medio practicante-Poco practicante	Muy practicante-Ns/Nc
Medio practicante-Nada practicante	Medio practicante-Ns/Nc
Medio practicante-Ns/Nc	Poco practicante-Ns/Nc
Poco-Nada practicante	
Poco practicante-Ns/Nc	
Nada practicante-Ns/Nc	

3) Si se ha fusionado un determinado par de categorías, se procede a realizar nuevas fusiones de los valores del pronosticador, pero esta vez con una categoría menos, pues dos de las antiguas han sido reducidas a una sola.

4) El proceso se acaba cuando ya no pueden realizarse más fusiones porque los χ^2 ofrecen resultados significativos.

De esta forma, como casos extremos, podría suceder que una variable con c categorías siguiera con c grupos, en el supuesto de que todos ellos sean diferentes entre sí; o bien, que las categorías tengan valores tan parecidos en la variable dependiente que se queden reducidos a uno solo, con lo que el poder discriminador del pronosticador sería nulo.

Véase un ejemplo práctico con la variable ideología como segmentadora de la posición ante el aborto (tabla 3). En este caso, la ideología tiene tres valores («izquierda», «centro» y «derecha»). En el primer paso, se halla el χ^2 de los siguientes contrastes: «izquierda» versus «centro», «centro» versus «derecha» e «izquierda» versus «derecha», aunque este último no se aplicaría si la ideología se considerara monótona. De estos tres χ^2, el menor (1,36) corresponde al contraste «centro» versus «derecha»; lo que indica que estas categorías son las más parecidas entre sí en lo que se refiere a opinión sobre el aborto. Por eso y, especialmente, porque ambas categorías no presentan diferencias significativas, se pueden reagrupar para explicar la actitud en cuestión. El siguiente paso, sería comprobar si el contraste («centro», «derecha») versus «izquierda» presenta un χ^2 significativo; en cuyo caso estos dos serían los grupos de este pronosticador

que posteriormente habría que contrastar con otros pronosticadores. Si, por el contrario, no hubiera habido resultado significativo, habría sido consecuente juntar los dos grupos de valores y, en este caso, ya sólo hubiese quedado un único grupo de categorías, y, por tanto, la variable ideología no habría sido útil para discriminar la opinión estudiada.

TABLA 3

Ejemplo de agrupación de categorías de la variable *ideología*

PRIMER PASO:
CRUCE DE OPINIÓN ANTE EL ABORTO SEGÚN PARES DE VALORES DE IDEOLOGÍA

			Ideología		Ideología		Ideología	
		Total	Izq.	Cent.	Izq.	Der.	Cent.	Der.
Posición	Sí	19,5%	33,1%	15,5%	33,1%	12,0%	15,5%	12,0%
ante el	No	75,3%	63,4%	78,5%	63,4%	83,7%	78,5%	83,7%
aborto	NC	5,3%	3,4%	6,0%	3,4%	4,3%	6,0%	4,3%
Total		(1.200)	(290)	(818)	(290)	(92)	(818)	(92)

Pruebas de chi-cuadrado				**Chi-cuadrado**			**Chi-cuadrado**		
	Valor	gl	Sig.	Valor	gl	Sig.	Valor	gl	Sig.
Chi2 de Pearson	41.959	2	.000	15.495	2	.000	1.362	2	.056
Razón de veros.	39.187	2	.000	17.496	2	.000	1.439	2	.487

SEGUNDO PASO:
CRUCE DE OPINIÓN ANTE EL ABORTO SEGÚN VALORES AGRUPADOS EN EL PASO ANTERIOR

			Ideología	
		Total	Izq.	Cent.-Der.
Posición	Sí	19,5%	33,1%	15,2%
ante el	No	75,3%	63,4%	79,0%
aborto	NC	5,3%	3,4%	5,8%
Total		(1.200)	(290)	(910)

Pruebas de chi-cuadrado			
	Valor	gl	Sig.
Chi2 de Pearson	45.734	2	.000
Razón de veros.	42.165	2	.000

Existe un procedimiento que ahorra gran cantidad de cálculos y posee una razonable base lógica. Se trata de limitarse a la obtención de segmentaciones binarias. Esto implica que, sea cual sea el número de categorías de los pronosticadores, se busque la mejor combinación de ellas que genere sólo dos grupos ($k = 2$). En consecuencia, habría que formar todas las posibles combinaciones de dos grupos con las c categorías y seleccionar aquél con un χ^2 mayor. Es evidente que utilizando estos contrastes binarios, el número de posibilidades de agrupación se reduce. En el caso de una variable ordinal con valores «mucho», «bastante», «poco» y «nada», el número de contrastes sería de 7 con la opción libre y de 3 con la monótona. (Véase tabla 4). Biggs *et al.* (1991) propusieron la fusión continua de pares de valores hasta que sólo quedara una única dicotomía de valores, denominando a tal procedimiento CHAID exhaustivo.

TABLA 4

Contrastes binarios posibles con una variable de cuatro valores según opción

Opción monótona (Variables ordinales)	Opción libre (Variables nominales)
(Mucho) y (Bastante, Poco, Nada)	(Mucho) y (Bastante, Poco, Nada)
(Mucho, Bastante) y (Poco, Nada)	(Bastante) y (Mucho, Poco, Nada)
(Mucho, Bastante, Nada) y (Nada)	(Poco) y (Mucho, Bastante, Nada)
	(Nada) y (Bastante, Mucho, Poco)
	(Mucho, Bastante) y (Poco, Nada)
	(Mucho, Poco) y (Bastante, Nada)
	(Mucho, Nada) y (Bastante, Poco)

2.2. Selección de los mejores pronosticadores

Una vez que para cada pronosticador se ha realizado la combinación oportuna de categorías, el siguiente paso sería la selección de los mejores pronosticadores. Para hacerlo, hay que calcular para cada uno de ellos su correspondiente χ^2 y comparar las significaciones obtenidas; sin embargo, es conveniente en este proceso modificar la significación de cada pronosticador con el ajuste de Bonferroni [11], porque la probabilidad de obtención de un resultado significativo aumenta artificialmente con la proliferación de pruebas estadísticas que implica este análisis.

[11] El ajuste de Bonferroni consiste en la aplicación de la desigualdad establecida por el mismo autor. Ésta dice que en el caso de que se hagan B pruebas de significación, la significación total (p_T) debe ser menor o igual que la suma de cada una de las significaciones (p_i).

$$p_T \leq \sum_{i=1}^{B} p_i$$

El ejemplo de predicción de la opinión ante el aborto en función de sexo, edad e ideología ayuda a entender este proceso. En la tabla 5 se analiza el conjunto de la muestra por los distintos pronosticadores, a los que ya se ha aplicado el proceso de agregación de categorías. Del conjunto de 1.200 sujetos que han sido entrevistados, 566 son varones y 634 mujeres. El 23,3% de los primeros es favorable al aborto y el 16,1% de las mujeres sostiene la misma posición. El χ^2 (9,9) tiene una significación [12] ajustada de 0,007. Existe, pues, relación significativa; pero antes de proceder a seleccionar este pronosticador, es necesario analizar el resto de los seleccionados para este ejemplo: La muestra está repartida entre 651 sujetos con menos de 46 años y 549 de mayor edad. También entre estos dos segmentos de la muestra hay diferencias en la opinión, incluso mayores. El 26,0% de los jóvenes autorizan esta práctica; pero sólo un 11,8% de los de más avanzada edad mantienen esta actitud. El χ^2 lógicamente presenta un valor mayor (37,9) y, de igual forma, su significación es muy baja (5.6E-9). Sin embargo, el mejor pronosticador es la ideología. Un tercio aproximadamente de los sujetos de izquierda aprueba el aborto en el supuesto de que un matrimonio no desee tener más hijos, mientras que la probabilidad de que los entrevistados

---

El número posible de pruebas de significación se puede calcular a través de fórmulas combinatorias a partir del número de categorías iniciales de la variable *(c)* y del número de grupos formados tras la agrupación de categorías (k). Es obvio que el cálculo será distinto según la opción de reducción de categorías que se utilice.

Así, para variables nominales la fórmula es la siguiente:

$$B = \sum_{i=0}^{k-1} (-1)^i \frac{(k-i)^c}{i!(k-i)!}$$

Si se utilizan variables ordinales:

$$B = \binom{c-1}{k-1}$$

Y para variables ordinales con valores perdidos:

$$B = \frac{k-1+k(c-k)}{c-1}\binom{c-1}{k-1}$$

En la práctica, hay que multiplicar la significación del χ^2 por el resultado de B, con lo que se evita el riesgo de rechazo inadecuado de hipótesis por realizar múltiples ensayos.

Más adelante, hay un ejemplo donde se puede aplicar este proceso. En las anteriores segmentaciones no se ven las implicaciones de este ajuste porque sólo hay variables con dos categorías y, en estos casos, B es siempre igual a 1. La última variable de la citada tabla, ingresos del entrevistado, tiene un χ^2 de 62.9, al que con 6 grados de libertad debería corresponderle una significación de 1.2E-11. Sin embargo, en este caso, como B, el número de comparaciones posibles, es igual a 5, después de aplicar la fórmula para variables ordinales con valores perdidos con los parámetros c = 4 y k = 3, la significación real es menor o igual a 5.9E-11. Para más detalle véase Kass (1980) y Kawkins y Kass (1982).

[12] Se recuerda que por significación se entiende la probabilidad de cometer un error de tipo I (rechazo de una hipótesis verdadera). Por tanto, mientras más baja sea, las posibilidades de error son menores y la relación entre las variables es más fuerte.

de centro y los de derecha mantengan esta actitud es sólo de un 15,2%. El χ^2 es el mayor de los tres (45,7) y la significación, la más baja (1.2E-10). Por tanto, este es el mejor pronosticador de los tres y es el que se utilizará para realizar la primera segmentación de la muestra. De este modo, quedarán formados dos grupos: uno de 290 sujetos (los autoubicados en la izquierda) y otro de 910 individuos que han declarado ser de centro o de derecha.

TABLA 5
Análisis de la muestra completa

			SEXO	
		Total	Varón	Mujer
Posición ante el aborto	Sí	19,5%	23,3%	16,1%
	No	75,3%	71,7%	78,4%
	NC	5,3%	4,9%	5,5%
Total		(1.200)	(566)	(634)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	9.973	2	.007

			EDAD	
		Total	≤ 45	> 45
Posición ante el aborto	Sí	19,5%	26,0%	11,8%
	No	75,3%	69,0%	82,7%
	NC	5,3%	5,1%	5,5%
Total		(1.200)	(651)	(549)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	37.997	2	5.6E-09

			IDEOLOGÍA	
		Total	Izq.	Cent.-Der.
Posición ante el aborto	Sí	19,5%	33,1%	15,2%
	No	75,3%	63,4%	79,0%
	NC	5,3%	3,4%	5,8%
Total		(1.200)	(290)	(910)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	45.734	2	1.2E-10

Una vez realizada la primera segmentación, se procede a la ejecución de sucesivas segmentaciones para cada uno de los grupos formados por la primera. Prosiguiendo con el ejemplo, habría que averiguar si entre los individuos de izquierda existen diferencias considerables de sexo o edad. Así, en la tabla 6 se observa que los varones de izquierda son significativamente más

favorables al aborto que las mujeres de esta ideología (37,7% vs. 26,1%). No obstante, sobre los entrevistados de izquierda, el pronosticador edad tiene mayor poder de discriminación. Los jóvenes apoyan este tipo de aborto en un 37,1% de casos, mientras que los mayores de 45 años sólo lo hacen en un 23,5% ($p \leq 0{,}005$).

TABLA 6

Análisis de la muestra de sujetos de izquierda (Grupo 2)

			SEXO	
		Total	Varón	Mujer
Posición ante el aborto	Sí	33,1%	37,7%	26,1%
	No	63,4%	60,6%	67,8%
	NC	3,4%	1,7%	6,1%
Total		(290)	(175)	(115)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	7.258	2	.027

			EDAD	
		Total	≤ 45	> 45
Posición ante el aborto	Sí	33,1%	37,1%	23,5%
	No	63,4%	58,0%	76,5%
	NC	3,4%	4,9%	
Total		(290)	(205)	(85)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	10.690	2	.005

También hay que realizar el proceso con los individuos de centro y derecha. Pero en este caso, además de probar el efecto de sexo y edad, hay que analizar si las personas de centro y derecha son diferentes entre sí. Esto no se aplicaba al otro grupo porque era un grupo homogéneo en ideología: estaba formado por sujetos de izquierdas. Tras el cálculo de los χ^2 (véase tabla 7) la única variable discriminatoria es la edad. Los sujetos de centro y los de derecha mantienen posiciones similares (el 15,5% de los de centro son favorables al aborto, y el 12,0% de los de derechas: $p \leq 0{,}51$). Como contrapartida, entre los 910 entrevistados de centro y derecha, un 20,9% de los 446 jóvenes mantiene una actitud favorable al aborto; mientras que entre los 464 de mayor edad sólo un 9,7% tiene la misma opinión ($p \leq 1.6E\text{-}5$). Por tanto, al igual que ocurría entre los individuos de izquierda, el segundo paso de la segmentación realizado con los entrevistados de centro y derecha, divide a estos sujetos según su edad.

Hasta aquí, han sido realizadas tres segmentaciones en dos niveles y en este proceso se han conformado cuatro segmentos o grupos:

TABLA 7

Análisis de la muestra de sujetos de centro y derecha (Grupo 3)

			SEXO	
		Total	Varón	Mujer
Posición ante el aborto	Sí	15,2%	16,9%	13,9%
	No	79,0%	76,7%	80,7%
	NC	5,8%	6,4%	5,4%
Total		(910)	(391)	(519)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	2.165	2	.339

			EDAD	
		Total	≤ 45	> 45
Posición ante el aborto	Sí	15,2%	20,9%	9,7%
	No	79,0%	74,0%	83,8%
	NC	5,8%	5,2%	6,5%
Total		(910)	(446)	(464)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	22.114	2	1.6E-05

			IDEOLOGÍA	
		Total	Cent.	Der.
Posición ante el aborto	Sí	15,2%	15,5%	12,0%
	No	79,0%	78,5%	83,7%
	NC	5,8%	6,0%	4,3%
Total		(910)	(818)	(92)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	1.362	2	.506

a) Sujetos de izquierda jóvenes: (n = 205; p_a = 37,1%).
b) Sujetos de izquierda mayores: (n = 85; p_a = 23,5%) [13].
c) Sujetos de centro y derecha jóvenes: (n = 446; p_a = 20,9%).
d) Sujetos de centro y derecha mayores: (n = 464; p_a = 9,7%) [14].

Aún se podría proseguir la segmentación en su tercer nivel para cada uno de estos cuatro grupos. Véase cada uno de ellos:

Dado que se han introducido sólo tres pronosticadores, el grupo de jóvenes de izquierda únicamente puede ser segmentado con el pronosticador restante: el sexo. ¿Existen diferencias en la posición ante el aborto entre los hombres y las mujeres de este segmento? Los 117 varones que forman este

[13] Véase tabla 6.
[14] Véase tablal 7.

grupo son favorables en un 42,7%; las 88 mujeres sólo en un 29,5%. Estas diferencias parecen importantes; sin embargo (tabla 8), los tamaños de estas muestras no son suficientemente grandes para que esta desigualdad sea estadísticamente significativa. Por tanto, el análisis automático no subsegmentaría a este grupo de jóvenes de izquierda y de esta forma quedaría considerado como *grupo terminal*.

TABLA 8

Análisis de la muestra de sujetos jóvenes de izquierda (Grupo 4)

			SEXO	
		Total	Varón	Mujer
Posición ante el aborto	Sí	37,1%	42,7%	29,5%
	No	58,0%	54,7%	62,5%
	NC	4,9%	2,6%	8,0%
Total		(205)	(117)	(88)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	5.875	2	.053

La muestra de 85 individuos de izquierda mayores de 45 años es muy pequeña para que, al subdividirla, presente diferencias significativas entre los dos sexos. Efectivamente, en la tabla 9, aunque los 58 varones mayores de izquierda son más favorables que las 27 mujeres de similares características, las diferencias no son estadísticamente significativas.

TABLA 9

Análisis de la muestra de sujetos mayores de izquierda (Grupo 5)

			SEXO	
		Total	Varón	Mujer
Posición ante el aborto	Sí	23,5%	27,6%	14,8%
	No	76,5%	72,4%	85,2%
Total		(85)	(58)	(27)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	1.670	1	.196

En el grupo de los 446 jóvenes de centro y derecha, son posibles dos segmentaciones, bien con el pronosticador sexo, bien con la ideología, separando a los de centro de los de derecha. En la tabla 10, se detecta que ninguna de estas segmentaciones es significativa; sin embargo, en esta ocasión, no tanto por el bajo tamaño de las muestras, como por la pequeña diferencia de por-

centajes (21,5% vs. 20,3% tomando en cuenta el sexo y 21,0% vs. 18,9% haciendo uso de la ideología).

TABLA 10

Análisis de la muestra de sujetos jóvenes de centro y de derecha

			SEXO	
		Total	Varón	Mujer
Posición ante el aborto	Sí	20,9%	21,5%	20,3%
	No	74,0%	72,7%	75,1%
	NC	5,2%	5,9%	4,6%
Total		(446)	(205)	(241)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	.513	2	.774

			IDEOLOGÍA	
		Total	Cent.	Der.
Posición ante el aborto	Sí	20,9%	21,0%	18,9%
	No	74,0%	73,8%	75,7%
	NC	5,2%	5,1%	5,4%
Total		(446)	(409)	(37)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	.093	2	.955

Por último, el grupo de mayores de centro y derecha (tabla 11) está compuesto por 464 sujetos, de los que sólo el 9,7% aprueban el aborto en el supuesto de que un matrimonio no desee tener más hijos. El sexo no los discrimina, pues los varones favorables son el 11,8% y las mujeres, el 8,3% (diferencias, por lo demás, no significativas); ni existe distinción clara en la opinión entre los de centro y los de derecha (con porcentajes respectivos del 10,0% y el 7,3%). En definitiva, tampoco este grupo es susceptible de posterior segmentación, pues ninguna variable independiente presenta asociaciones significativas con la dependiente. Por ello, también cabe que sea considerado un grupo terminal.

En consecuencia, el análisis de segmentación subdivide a la muestra en los cuatro segmentos descritos en la página 16 y representados en la figura 2. Destaca las diferencias de opinión entre los grupos terminales S.4 y S.7 [15]: por un lado, los jóvenes de ideología de izquierda, con un 37,1% de favorables al aborto, y en el lado opuesto, los mayores de centro-derecha con un 9,7% de la misma opinión. Entre estas dos posiciones los dos grupos restantes presentan porcentajes muy similares entre sí, posiblemente no significativos y, por tanto, no

[15] Los grupos son numerados comenzando por la muestra global (S.1), de arriba abajo y de izquierda a derecha.

heterogéneos en los valores de la variable dependiente, aunque sí en los de las independientes o pronosticadores. Estos grupos son el S.5, formado por mayores de ideología de izquierda, y el S-6, compuesto por jóvenes de ideología de centro-derecha. Por su desigual composición se justifica que, aun semejantes en su posición ante el aborto, se sigan considerando como segmentos distintos. Del mismo modo, pueden comentarse los otros porcentajes reflejados en los distintos cuadros: Se ve claramente que el S.4, jóvenes de izquierdas, con un 58% de contrarios a la legalización del aborto en el caso de que un matrimonio no desee tener más hijos, y el S.7, mayores de centro y de derecha, con un 83,8% de no partidarios de esta práctica, son los grupos más heterogéneos entre sí. Por otro lado, también este último grupo es en el que hay mayor proporción de personas que no contestan a la pregunta de opinión.

TABLA 11

Análisis de la muestra de sujetos mayores de centro y de derecha

			SEXO	
		Total	Varón	Mujer
Posición ante el aborto	Sí	9,7%	11,8%	8,3%
	No	83,8%	81,2%	85,6%
	NC	6,5%	7,0%	6,1%
Total		(464)	(186)	(278)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	1.844	2	.398

			IDEOLOGÍA	
		Total	Cent.	Der.
Posición ante el aborto	Sí	9,7%	10,0%	7,3%
	No	83,8%	83,1%	89,1%
	NC	6,5%	6,8%	3,6%
Total		(464)	(409)	(55)

Pruebas de chi-cuadrado

	Valor	gl	Sig.
Chi2 de Pearson	1.356	2	.508

El proceso de segmentación debe ser examinado en sus distintas fases con el objeto de valorar el comportamiento de los pronosticadores alternativos. El problema estriba en que el programa analiza varias variables en cada paso de la segmentación y tiene que elegir, entre ellas, sólo una. Si en una determinada fase existen varios pronosticadores de similar poder de segmentación, el análisis de la elección efectuada puede conducir a interpretaciones precipitadas. Para descubrir la posible existencia de este problema, habrá que prestar atención en cada segmentación a la significación ajustada del χ^2 de los pronosticadores alternativos.

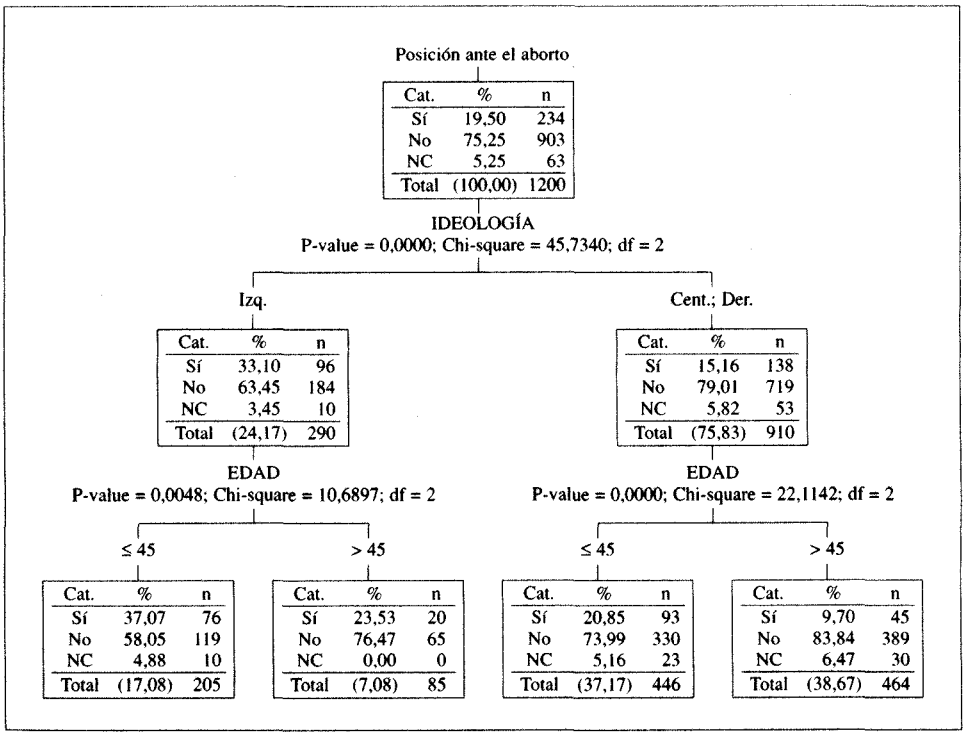

Figura 2. Segmentación de la opinión ante el aborto.

TABLA 12

Resumen de la segmentación. Significaciones del χ^2 para cada grupo y variable

	Grupos de segmentación						
Predictor	**S.1**	**S.2 (1)**	**S.3 (1)**	**S.4 (2)**	**S.5 (2)**	**S.6 (3)**	**S.7 (3)**
Sexo	0,007	0,027	0,339	0,053	0,196	0,774	0,398
Edad	5.6E-9	**0.005**	**1.6E-5**	–	–	–	–
Ideología	**1.2E-10**	–	0,506	–	–	0,955	0,508
Tamaño del grupo	(1.200)	(290)	(910)	(205)	(85)	(446)	(464)

N.B.: Los grupos en negrita son los grupos terminales entre paréntesis. Entre paréntesis, el grupo del que procede. Los coeficientes en negrita indican la variable por la que se efectúa la segmentación en un determinado grupo.

Por último, para determinar la capacidad pronosticadora de la segmentación en su conjunto, resulta muy útil cruzar la variable dependiente con una nueva variable compuesta, cuyos valores sean las características de cada uno de los gru-

pos terminales formados por la segmentación (tabla 13). Un coeficiente de asociación [16], como puede ser la V de Cramer, resume el poder de predicción de los segmentos en su explicación de la variable dependiente. En este caso, el coeficiente, cuyo rango va de 0 a 1, no es tan alto como sería de desear, lo que indica la escasa capacidad de predicción que tienen la ideología y la edad para explicar la actitud de los individuos ante el aborto en el supuesto de que los padres no deseen tener más hijos.

TABLA 13

Opinión sobre el aborto por los cuatro grupos terminales de la segmentación

			Grupo de segmentación			
		Total	**Jov.-Izq.**	**May.-Izq.**	**Jov.-C.d.**	**May.-C.d.**
Posición ante el aborto	Sí	19,5%	37,1%	23,5%	20,9%	9,7%
	No	75,3%	58,0%	76,5%	74,0%	83,8%
	NC	5,3%	4,9%		5,2%	6,5%
Total		(1.200)	(205)	(85)	(446)	(464)

	Estadísticos		
	Valor	gl	Sig.
Chi2 de Pearson	75.000	6	.000
V de Cramer	.18		.000

Otro modo de juzgar la bondad de la segmentación consiste en construir una tabla donde se crucen los datos empíricos de la variable dependiente con los que se pronosticarían con el conocimiento del segmento al que pertenece cada individuo. En la figura 2 aparece en cada grupo una línea en negrita. Ésta representa la categoría modal de la variable dependiente y es el pronóstico con menos riesgo de error en su predicción. Así, conociendo que una determinada persona es joven de ideologías de izquierda, lo menos arriesgado es pronosticar que –a pesar de que son los más favorables al aborto– estará en contra, porque el 58,0% de estos sujetos así lo están. En consecuencia, con el ejemplo que se ha utilizado para explicar la segmentación, el pronóstico para todos los segmentos es que estarán en contra. En la tabla de clasificación, quedan distinguidas las cifras de la diagonal, que son aciertos o coincidencias entre la predicción y lo real, de las que están fuera de ellas, que son equivocaciones. La *estimación del riesgo* se calcula mediante el cociente entre estas últimas frecuencias y el total número de

[16] Sobre coeficientes de asociación entre dos variables nominales, véase entre otros Ruiz-Maya et al. (1990), especialmente los capítulos 10 y 11.

casos. En este ejemplo, el riesgo de error es del 24,7%. Como coincide con la dispersión modal de la variable dependiente, la segmentación considera inútil para la predicción.

TABLA 14
Tabla de clasificación del análisis de segmentación

			Categoría real		
		Total	**Sí**	**No**	**NC**
Categoría predicha	Sí	0	**0**	0	0
	No	1.200	234	**903**	63
	NC	0	0	0	**0**
	Total	1.200	234	903	63

Estimación del riesgo: $\frac{(234 + 63)}{1.200} \times 100 = 24,7$

2.3. La finalización del proceso de segmentación

Si no se pusieran límites al proceso de segmentación, este análisis podría producir una gran cantidad de grupos terminales de tamaño muy pequeño que serían difíciles de interpretar. En un caso extremo, con un número elevado de variables y sin restricción alguna, este análisis produciría tantos grupos como individuos tuviese la muestra. En la situación común de una muestra de 1.000 sujetos con 5 pronosticadores de tres categorías cada uno, el número posible de grupos terminales sería de 243 (3^5) con un tamaño medio aproximado de cuatro personas (1.000/243). Es conveniente, por tanto, poner límites al proceso de segmentación. Existen cuatro tipos de filtros que evitan la continuación de la segmentación: los de significación, los de asociación, los de tamaño y los de nivel.

2.3.1. *Filtros de significación*

Son los más utilizados en la técnica CHAID de segmentación. Su criterio consiste básicamente en no permitir segmentaciones que no sean estadísticamente significativas. Por omisión, se sobrentiende que los límites de significación se sitúan en el nivel 0,05, que se corresponde con un nivel de confianza del 95%. Estos filtros pueden ser aplicados en dos de los procesos explicados anteriormente: bien en la agrupación de categorías de una variable (fusión de valores), bien en la selección del mejor pronosticador (segmentación de grupos).

La aplicación en el primer proceso es en realidad un mecanismo indirecto de finalización de la segmentación. Su efecto opera fundamentalmente en la cantidad de categorías de una determinada variable que van a segmentarse. Consiste

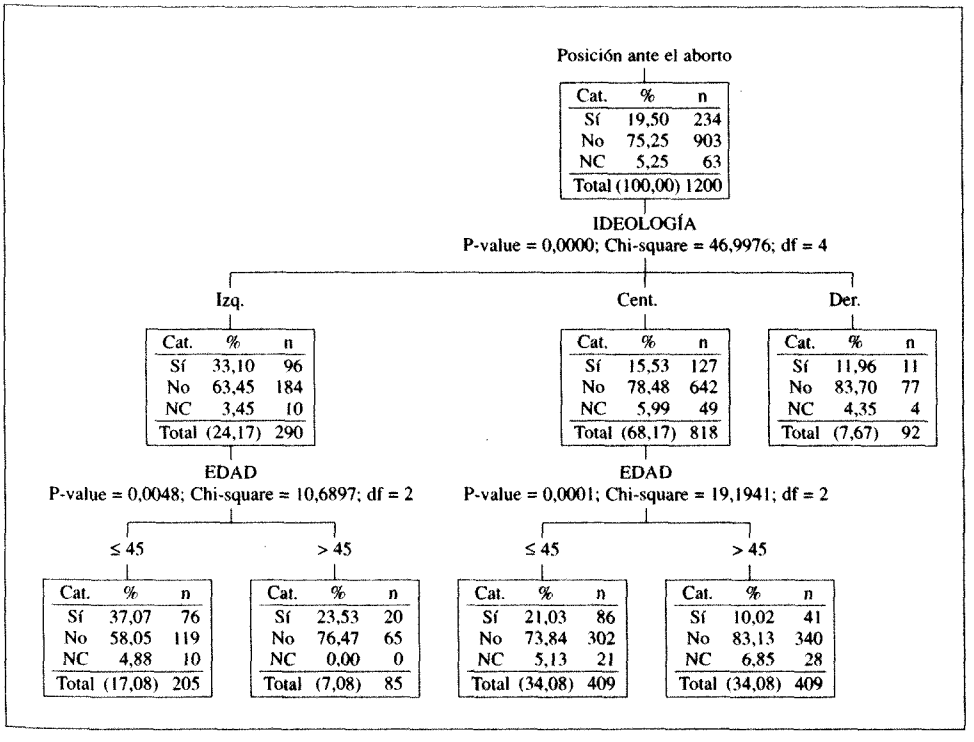

Figura 3. Segmentación de la opinión ante el aborto (SC = 1,0).

en determinar la significación mínima para que dos categorías de una variable queden englobadas en el mismo segmento. El valor –SC, significación de las categorías *(alpha for merging)*– más comúnmente asumido para este parámetro es el de 0,05. Si la significación de la diferencia en la variable dependiente entre dos categorías de la variable independiente es menor que este valor, se permite rechazar la hipótesis nula con un 95% de confianza y, como consecuencia, las dos susodichas categorías quedan separadas y se puede proseguir la segmentación. En cambio, si el valor es superior a 0,05, las categorías se funden, y, si quedan agrupadas todas las categorías de todas las variables, la segmentación se detiene.

Los valores extremos permiten comprender con mayor eficacia el efecto de este mecanismo. Si se escoge el mayor valor posible del parámetro (1,0), entonces, la agrupación o reducción de categorías de las variables se torna imposible y, siempre que haya significación entre pronosticador y variable dependiente, la segmentación formará con una determinada variable tantos grupos como categorías tenga. Se puede extraer un buen ejemplo de este procedimiento a partir de la segmentación mostrada en la figura 2. En aquel caso, las categorías centro y derecha quedaron unidas porque la significación de sus diferencias era de 0,57 (superior a 0,05). Si se hubiese establecido el criterio con un parámetro superior

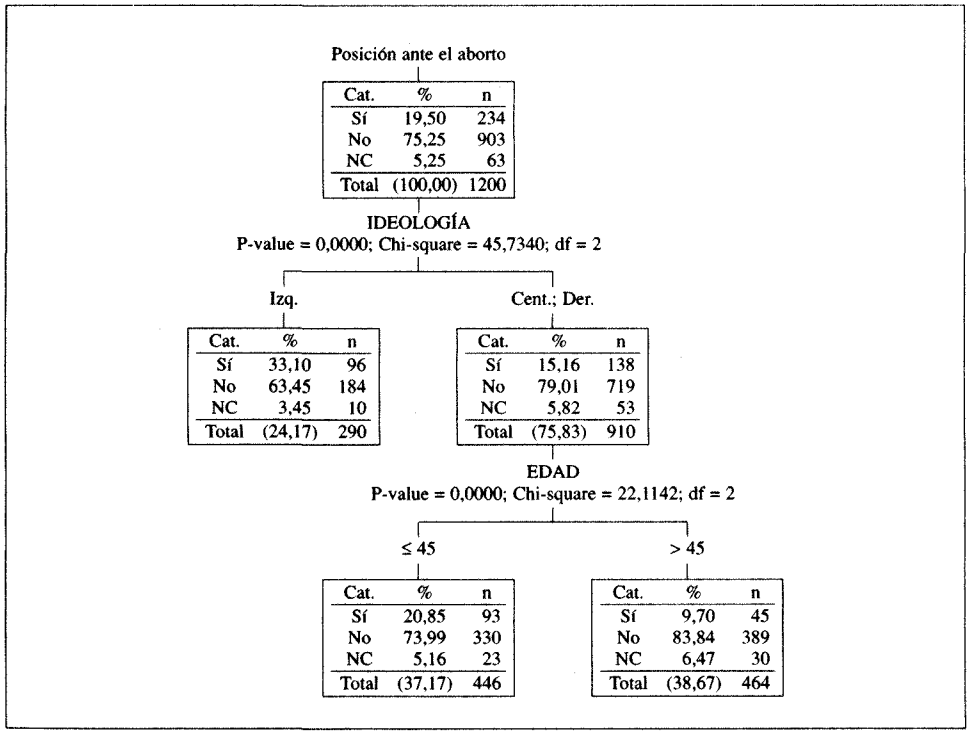

Figura 4. Segmentación de la opinión ante el aborto (SC = 0,0004).

a dicha cifra, la segmentación hubiese sido más *frondosa,* siguiendo la metáfora de la representación en forma arbórea. En concreto, cambiando el filtro, la primera subdivisión de la muestra, en lugar de dar lugar a dos grupos, proporciona tres grupos. (Compárese las figuras 2 y 3).

Si, en vez de poner el nivel de significación de la agrupación de las categorías en un valor alto, se situara en un valor bajo (por ejemplo, 4E-4), entonces, en lugar de producirse más subdivisiones entre los grupos, se generarían menos divisiones entre las categorías, con el riesgo añadido de que una determinada variable no funcione como un buen pronosticador. Esto es lo que sucede en el ejemplo de la figura 4, que no se produce segmentación por edad entre los individuos de izquierda. Y ocurre de esta manera porque la diferencia de porcentajes de las categorías de jóvenes y mayores no proporciona una significación menor de 0,0004. No siempre sucede esto de forma que implique la detención de la segmentación de un grupo. Lo lógico es esperar que una subdivisión de c categorías se reduzca a un número k, inferior al producido por un nivel de significación superior. En este caso, como el número inicial de categorías es igual a 2, la reducción implica la obtención de una sola categoría y de esta forma la segmentación no se lleva a cabo.

El otro mecanismo de control de significación, en lugar de operar sobre la agrupación de categorías, afecta a la selección de variables. Este procedimiento

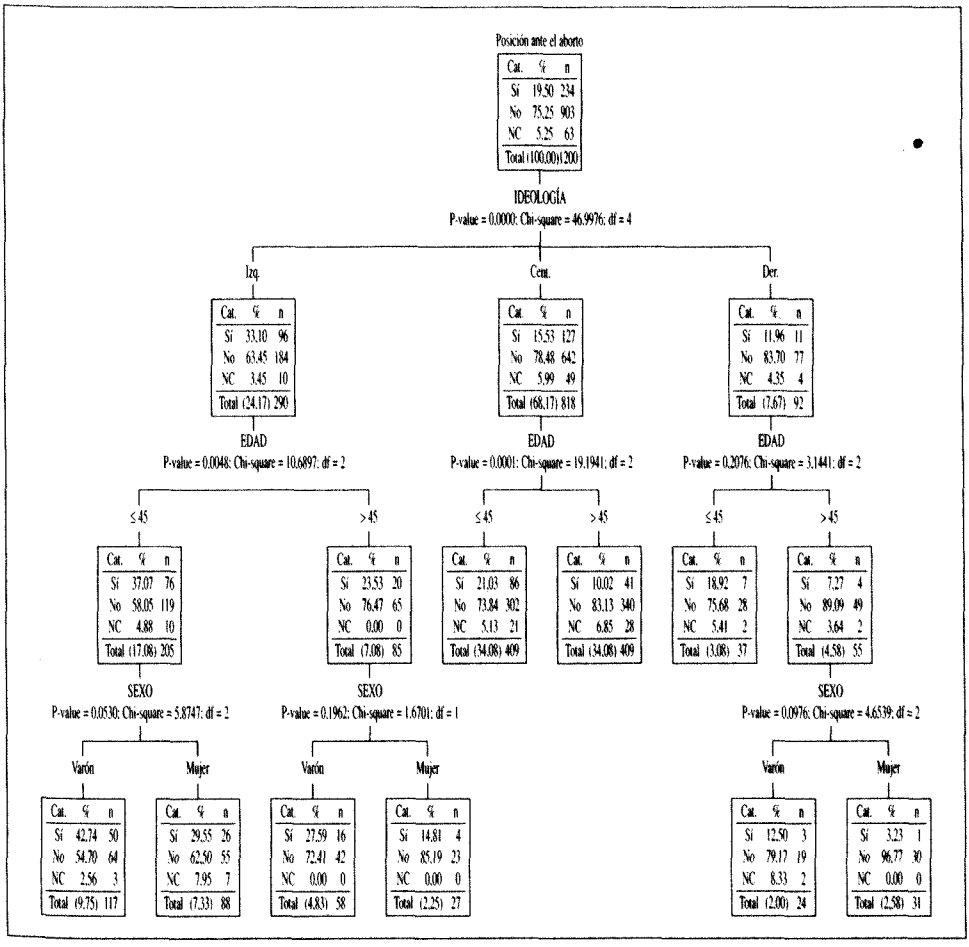

Figura 5. Segmentación de la opinión ante el aborto (SC = 1,0; SV = 0,6).

es una forma directa de finalizar la segmentación, porque, después de encontrar el pronosticador con menor significación, si no es inferior al límite establecido (generalmente 0,05), es obvio que no habrá otro pronosticador que cumpla también con esta propiedad, por lo que el proceso de división de la muestra termina. Visto desde sus posibilidades extremas, si se establece este parámetro –SV, significación de la variable *(alpha for splitting)*– en el valor 1,0, la segmentación se producirá por todas las variables existentes; pero si se determina que el parámetro sea 0,0, entonces la segmentación no se produce ni tan siquiera en el primer nivel, pues la significación empírica de un pronosticador, por muy pequeña que sea, siempre es superior a cero.

Si se aplica al ejemplo de la figura 3 un filtro de significación de pronosticador superior al establecido por omisión (por ejemplo, 0,60), es de esperar que la segmentación proporcione mayor número de niveles. En aquella tabla, no

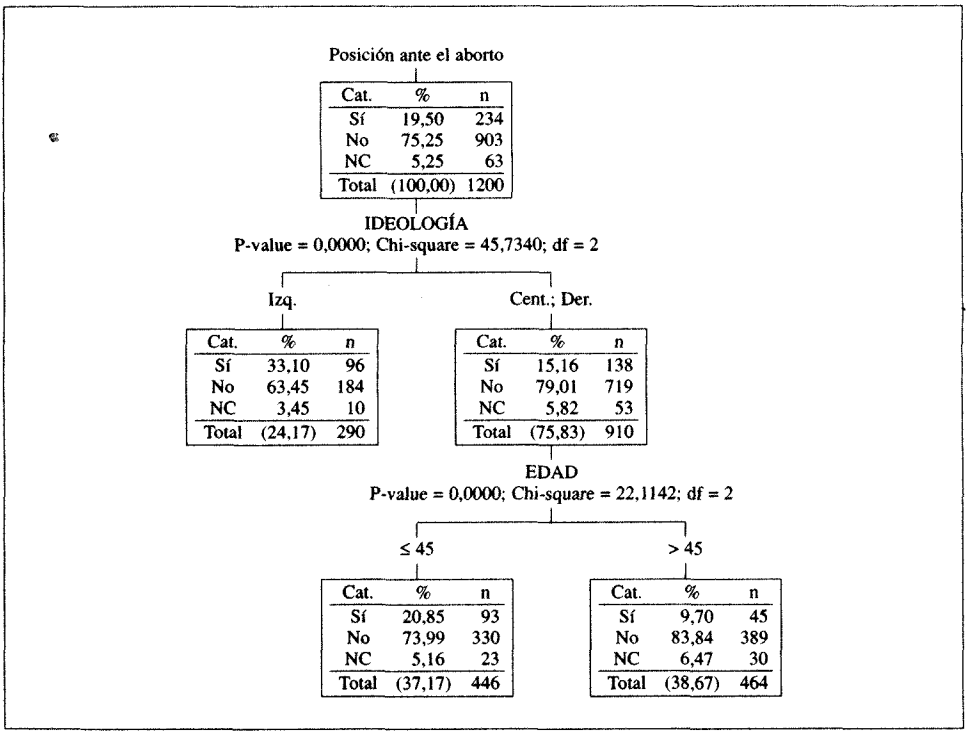

Figura 6. Segmentación de la opinión ante el aborto (SC = 0,05; SV = 0,0001).

aparecía el pronosticador sexo porque sus diferencias eran muy pequeñas. Ahora bien, es preciso tener en cuenta que no basta cambiar el parámetro SV, porque si sigue efectivo un valor inferior del SV, al operar con anterioridad, éste elimina los efectos del primero. Es conveniente, por tanto, que SC > SV. Por eso, en el ejemplo de la figura 5, aparecen los valores SC = 1,0 y SV = 0,60. Como es de esperar con estos parámetros, la segmentación desciende al tercer nivel y aparece el sexo como una tercera variable en el árbol. De todas formas, la diferencia de porcentajes de hombres y mujeres que están a favor del aborto es pequeña en relación con el tamaño de estos segmentos, y, en el caso de los S.7, S.8 y S.9, ni tan siquiera con el nivel establecido en 0,6 se produce la segmentación.

En cambio, si se aplica un filtro más severo, la segmentación sólo tendrá lugar cuando la variable independiente tenga una capacidad de predicción alta. Sobre el ejemplo matriz de la figura 2, aplicando en lugar del 0,05 por omisión, un SV de 0,0001, se obtiene una segmentación más reducida (figura 6) en la que los individuos de izquierda no aparecen segmentados, porque la edad, aunque tenga una significación por debajo del valor por omisión, posee una significación por encima del nivel establecido en el filtro.

2.3.2. *Filtros de asociación*

Cumplen una función análoga a la de los filtros de significación de pronosticadores. Se pueden aplicar a los siguientes coeficientes de asociación: Phi, V de Cramer, Coeficiente de Contingencia, T de Tschruprow u otros. Se trata de detener la segmentación no porque un determinado cruce no obtenga un mínimo de significación, sino porque el coeficiente de asociación elegido no alcance un determinado nivel. Lo que principalmente diferencia a un procedimiento de otro, es el hecho de que el que opera sobre la asociación no es sensible al número de casos sobre los que se trabaja. Por tanto, en valores equiparables de uno y otro, los filtros de asociación son más permisivos en los niveles más bajos de segmentación. Como los de significación son muy sensibles al número de casos, es muy probable que en el tercer o cuarto nivel el análisis no cumpla las condiciones del filtro, porque los segmentos tengan un tamaño reducido. En cambio, los coeficientes de asociación, por el hecho de eliminar la influencia del número de casos, permiten segmentaciones aun en condiciones de escasos sujetos. En este caso, hay mucho menos acuerdo sobre cuál debe ser el valor del filtro. Como regla de experiencia, se consideran adecuados los valores 0,10 ó 0,20. Sin embargo, el programa *Answer Tree* del SPSS no contempla la posibilidad de utilizarlos para el control de la segmentación [17]. En todo caso, la opción recomendada para el uso de estos filtros es que se utilicen en conjunción con un filtro de significación, de forma que una segmentación que no sea significativa no se lleve a cabo por muy grande que sea su coeficiente de asociación. El caso contrario, que justifica especialmente el uso de estos filtros, también suele suceder. Se trata de relaciones entre variable dependiente y pronosticador muy significativas, pero con un coeficiente de asociación bajo, que se dan con frecuencia cuando se trabaja con muestras de elevado número de casos.

2.3.3. *Filtros de tamaño*

Su principal objetivo consiste en evitar que se formen grupos muy pequeños durante el proceso de segmentación, dado el problema que supone la generalización en estos casos. Si, por ejemplo, se segmentara un grupo de 25 personas de las que un 30% es favorable al aborto, se plantearían dos problemas: por un lado, este grupo no sería representativo en sí de la población; por otro, el valor del 30% tampoco sería un estimador muy preciso con un tamaño de muestra tan reducido.

---

[17] No obstante, Sonquist y Morgan (1963), por utilizar la segmentación binaria con una variable dependiente de intervalo, confiaban más en los coeficientes de asociación que en los estadísticos de significación. Por ello, su AID seleccionaba las variables con mayor coeficiente de asociación y establecía como principal filtro la magnitud del coeficiente de determinación χ^2, es decir, el cociente entre la suma cuadrática intergrupos y la suma cuadrática total. Otro filtro considerado por estos autores consiste en que la segunda cantidad mencionada alcance un mínimo nivel arbitrario. La razón estriba en evitar la segmentación de grupos muy homogéneos. Este último criterio sería inaplicable en el algoritmo CHAID.

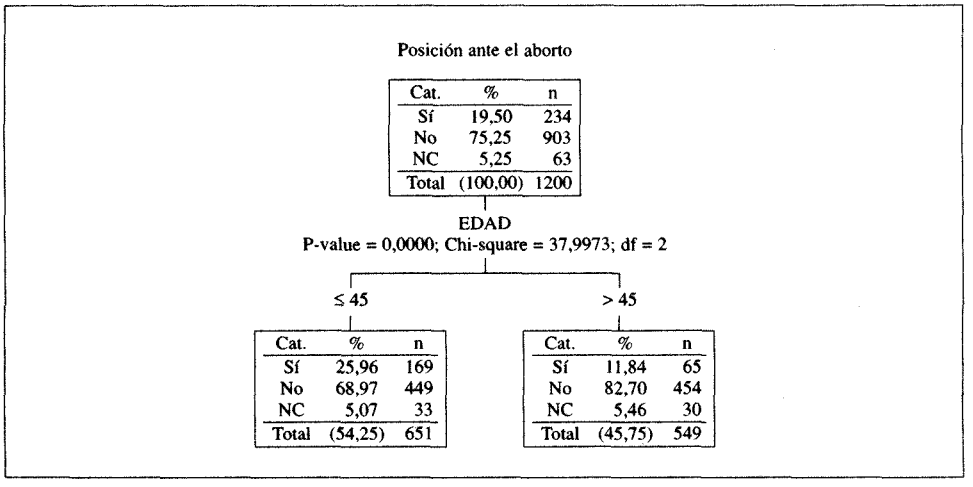

Figura 7. Segmentación de la opinión ante el aborto (Nd = 400).

Los filtros de tamaño pueden aplicarse en dos momentos: después de la segmentación (Nd, *child node*) y antes de la segmentación (Na, *parent node*). En el primer caso, no se puede formar un grupo si no tiene un número establecido de componentes. En el segundo, la segmentación se detiene en el supuesto de que haya un grupo que haya descendido de un determinado número de individuos.

Supóngase que se arbitra que no haya ningún grupo con menos de 400 sujetos, en cuyo caso, si se aplica la segmentación a los datos de la figura 2, la ideología no sería un pronosticador adecuado porque genera un grupo, los individuos de izquierda, con menos (290) de la cantidad establecida (400). Por tanto, en estas circunstancias, la segmentación (figura 7) presentaría un aspecto muy diferente de la original. Se formarían sólo dos grupos de edad, compuestos uno por 651 jóvenes y el otro por 549 mayores.

En cambio, si se opta por el filtro del tamaño antes de la segmentación y se toma como cantidad el mismo número arbitrario, esto es, 400, el gráfico en forma de árbol toma una apariencia completamente distinta del anterior, porque con este nuevo criterio, la ideología sí funciona como pronosticador (figura 8). Lo que sucede es que el grupo de ideología de izquierdas no se segmenta porque su tamaño es inferior al establecido. Sin embargo, el grupo de centro-derecha, por tener 910 sujetos, se segmenta normalmente.

Es obvio que ambos filtros pueden utilizarse al mismo tiempo. Lo que no tiene sentido es que el filtro antes de la segmentación (Na) sea inferior en número al de después (Nd), puesto que de esta forma este último no se aplicaría. Sólo tiene razón que Na sea superior a Nd. Como regla general, se recomiendan unos parámetros de 100 para Na y 50 para Nd. Esto implica la no obtención de grupos inferiores a un medio centenar de personas y la no segmentación de conjuntos con menos de cien componentes.

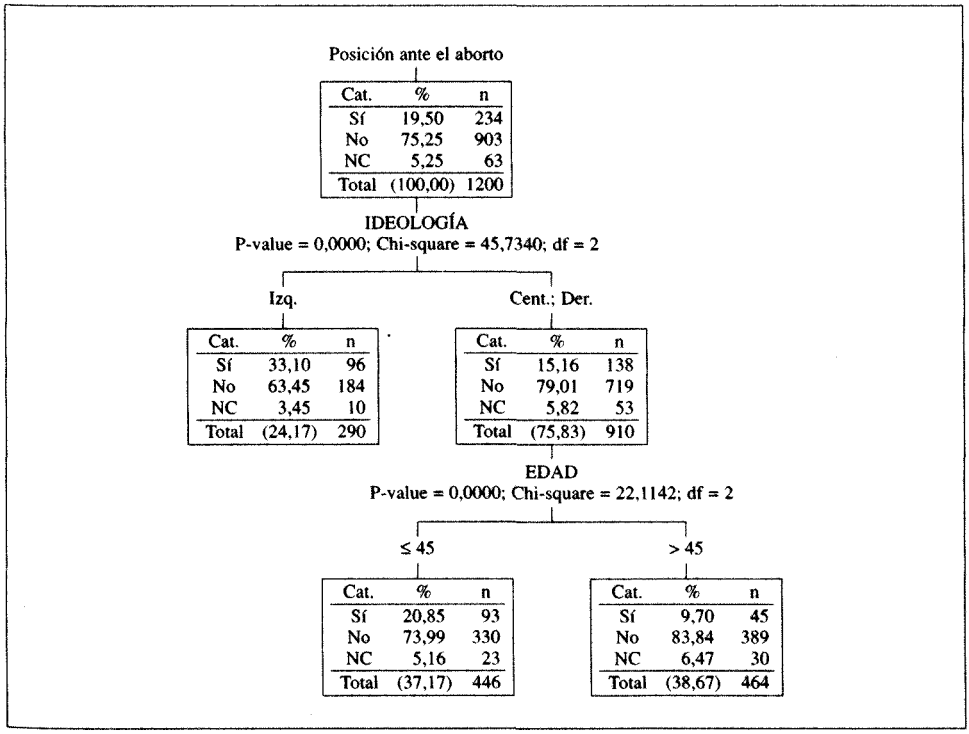

Figura 8. Segmentación de la opinión ante el aborto (Na = 400; Nd = 0).

2.3.4. *Filtros de nivel*

Por último, existe un cuarto tipo de mecanismo de detención de la segmentación. Consiste en arbitrar un nivel (Ns, *depth*) máximo de segmentación. Si se establece este criterio en 0, la segmentación no tendrá lugar; si en 1, sólo se realizará una segmentación; si en 2, dos tandas. Por tanto, por nivel se entiende cada una de las franjas horizontales del árbol invertido. La primera franja horizontal corresponde al total de la muestra, la segunda a la primera segmentación, la tercera a la segunda. Este filtro evita que se formen múltiples segmentaciones en segmentos desproporcionadamente grandes de la muestra. Asimismo, contribuye a simplificar los resultados en la medida en que reduce directamente el número de variables necesarias para predecir la variable dependiente.

En el ejemplo de la figura 9, se han fijado los filtros de significación en 1,0, con objeto de que sólo operase el filtro de nivel. Por ello, a diferencia del de la figura 2, aparece la ideología escindida en tres segmentos. Pero, de forma distinta al de la 5, no prosigue la segmentación hasta el tercer nivel, puesto que el valor del filtro Ns (nivel de segmentación) es 2.

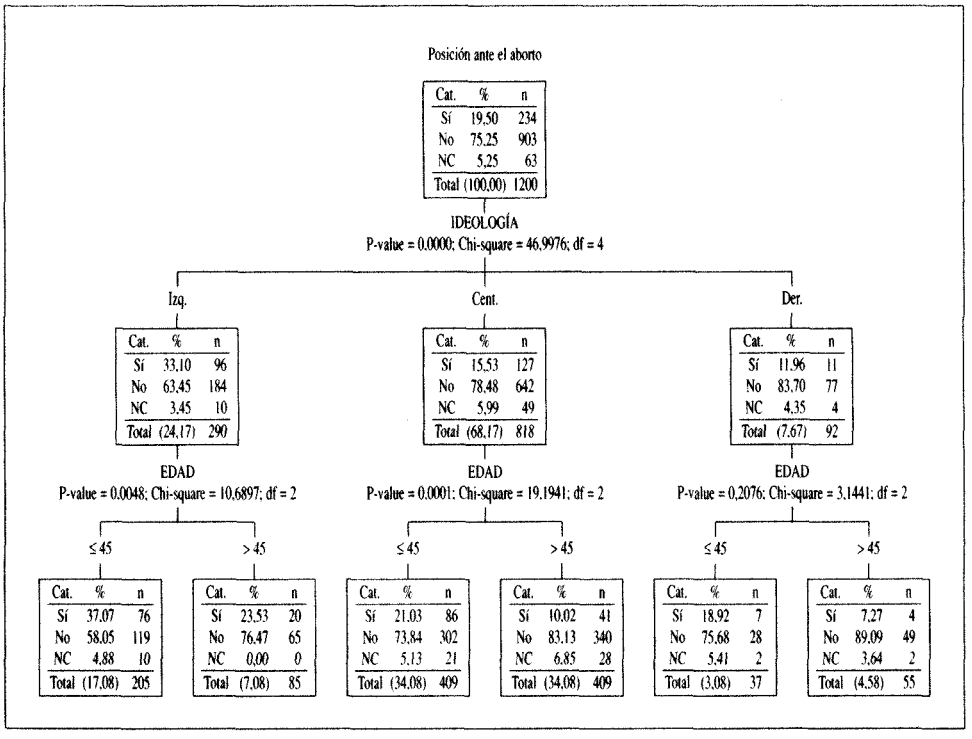

Figura 9. Segmentación de la opinión ante el aborto (Ns = 2; SC = 1,0; SV = 1,0).

3. EJEMPLOS DE APLICACIÓN [18]

La función clasificadora del análisis de segmentación permite configurar una serie de grupos que se distinguen por su comportamiento distinto en una determinada variable dependiente. La especificación de las características de los grupos terminales formados por esta técnica es un excelente medio para describir grupos heterogéneos de la muestra. Segmentar significa dividir y este análisis permite con su algoritmo el hallazgo de grupos muy distintos en un determinado aspecto. Por tanto, uno de los usos que se le puede dar a la segmentación es la descripción de las muestras y, por extensión, de las poblaciones de las que son extraídas.

La mejor manera de efectuar la descripción con el análisis de segmentación es mediante la interpretación de los grupos terminales. Hay que recordar que para hacer una buena descripción es necesario introducir pronosticadores adecuados

[18] Otros ejemplos de aplicación en el campo de las Ciencias Sociales en España son Horter (1978), García Ferrando (1982) y Alvira, García López y Horter (1982). Todos ellos utilizan el algoritmo AID basado en la suma cuadrática. Escobar (1993) emplea esta técnica para estudiar la afiliación, simpatía y movilización sindical en España.

en el procedimiento. A continuación se compararán dos ejemplos, ambos con la misma variable dependiente, opinión sobre el aborto, pero el segundo con más pronosticadores. Así se mostrará la conveniencia de dos reglas: a) incluir variables que sean relevantes para la dependiente y b) introducir el máximo posible de pronosticadores ya que el análisis en cuestión se encarga de filtrar los relevantes.

El primer ejemplo ya se ha presentado en epígrafe anterior. Se trata de segmentar la opinión sobre el aborto en el caso de que un matrimonio no desee tener más hijos a partir de tres variables: sexo, edad e ideología. De acuerdo con la figura 2, las dos últimas variables discriminan la opinión de los sujetos y, de este modo, se forman cuatro grupos terminales. Los más pro-abortistas (37,1%) son las personas de izquierda con menos de 45 años. Le siguen los de izquierda con más edad (23,5%). Los de ideología de centro-derecha, si son jóvenes presentan una actitud favorable en un 20,1%; pero si son mayores de 45 años, la posición a favor se reduce al 9,7%. Merece la pena destacar que los de izquierda siempre son más favorables al aborto que los de centro-derecha; aunque en el grupo de los mayores de cuarenta y cinco años de izquierda, la probabilidad de estar de acuerdo es poco más favorable que en el de los jóvenes de centro-derecha.

Además de los porcentajes, para la correcta descripción, hay que tener en cuenta la frecuencia de cada grupo. Se observa, en la misma figura 2, que el grupo de mayores de izquierda es muy reducido: apenas cubre el 7% de la muestra (85 de los 1.200). En cambio, los dos grupos de centro-derecha son los más numerosos, casi un 40% en cada uno de ellos [19].

Hay dos posibles maneras de resumir la información descriptiva de este análisis. La primera sería insistiendo en la oposición centro-derecha *versus* izquierda, destacando que las tres cuartas partes de la muestra (S.6 y S.7) presentan una baja aceptación del aborto practicado en las circunstancias antedichas, y los situados a la izquierda (24,2%, 290 de los 1.200 entrevistados) son más favorables, sin llegarlo a ser en mayoría (en el supuesto más favorable, los jóvenes de izquierda sólo aprueban esta práctica en un 37,0%). La otra interpretación insistiría más en los grupos extremos: Aunque aproximadamente un 20% de la población apruebe que se lleve a cabo este tipo de aborto, hay dos grupos en los que las probabilidades de aprobación son considerablemente diferentes. Por un lado, están las personas mayores de centro-derecha (un 38,7% de la muestra), de los que menos de un 10% darían su aprobación. Y, por otro, se encuentran los jóvenes de izquierda (un 17,1% de la muestra) entre los que la probabilidad de mantener esta opinión es superior al 35%.

En el segundo ejemplo, además del sexo, la edad y la ideología, se introducen las siguientes variables: estado civil, posición familiar, número de individuos en el hogar, nivel de estudios del entrevistado y del cabeza de

[19] Existe en este caso gran mayoría de individuos de centro derecha porque los que no han proferido su ideología están también incluidos en este grupo. Se ha hecho así porque su posición ante el aborto es más similar a la de aquéllos que a la de los de izquierda. Aunque esto se haya realizado manualmente para simplificar el ejemplo, el análisis de segmentación los hubiese incorporado automáticamente de la misma forma.

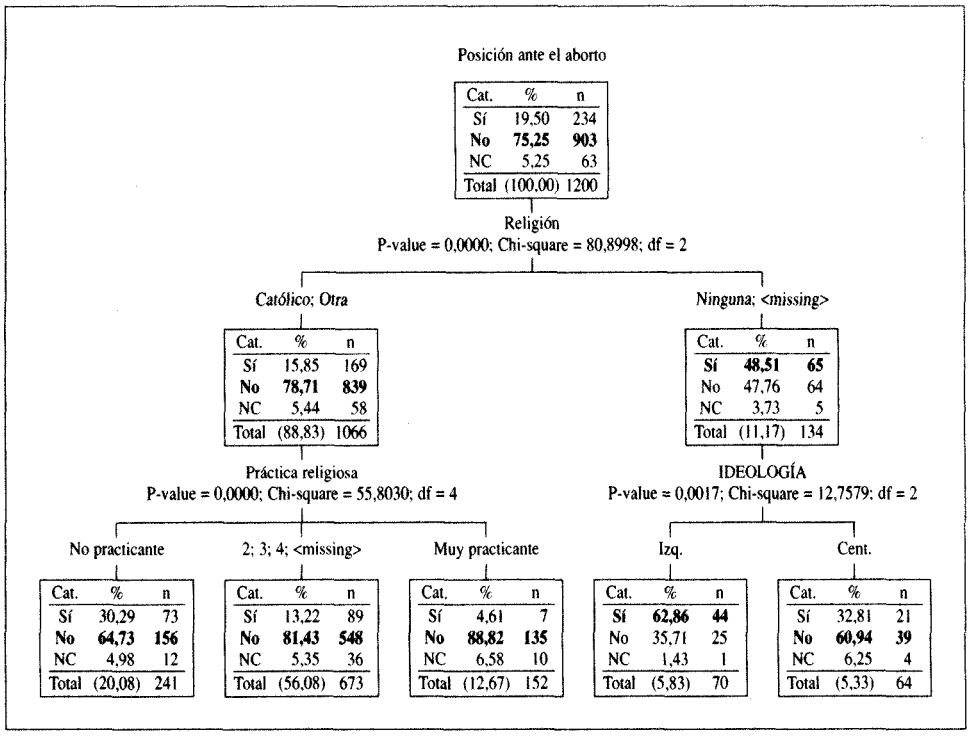

Figura 10. Segmentación de la opinión ante el aborto con 12 variables independientes.

familia, creencia y práctica religiosa, clase social subjetiva e ingresos. En total, pues, se incluyen 12 variables para tratar de describir la misma variable dependiente: la opinión sobre el aborto en el supuesto de que un matrimonio no desee tener más descendencia. Los resultados del árbol se presentan en la figura 10.

Lo que más resalta en este gráfico es el poder explicativo de la primera variable de segmentación: la práctica religiosa. Ésta forma dos grupos: los creyentes (católicos y otros), con un 15,8% de favorables a este tipo de aborto, y los no creyentes (incluyendo a quienes no contestan), con un 48,5% de opiniones a favor. Sin embargo, en el árbol aparecen cinco grupos terminales, pues uno se subdivide en tres y el otro en dos. De este modo, entre los creyentes se forman tres grupos muy distintos según la práctica religiosa: por un lado, los indviduos que no asisten a los cultos religiosos con un 30,3% de favorables; por el otro, los muy practicantes, con menos de 5% de partidarios; entre ellos, los medianamente practicantes con un 13,2% de tolerantes. Entre los no creyentes, también se forman dos grupos distintos, pero no en función, obviamente, de la práctica religiosa, sino de la ideología. Así, entre los de centro, la aceptación del aborto en el caso de que un matrimonio no desee tener más hijos es del 32,8%. Y, si son de izquierda, el porcentaje de favorables asciende hasta llegar al 62,9%. Aparece así

con la segmentación un grupo que aprueba mayoritariamente la interrupción del embarazo: los no creyentes de izquierdas.

Como consecuencia de este análisis, se forma una configuración de grupos muy distinta de la anterior, especialmente debido a que en el primer ejemplo no se introdujo variables muy importantes para describir la variable dependiente: la creencia y la práctica religiosa. En esta ocasión, son cinco los grupos terminales formados a partir de la muestra. Dos de ellos son muy poco favorables al aborto (menos del 15% a favor): los creyentes practicantes. Otros dos grupos mantienen posiciones algo más favorables, pero aún bajas (en torno al 30%): los creyentes que no practican y los ateos de centro [20]. Por últimos, el quinto grupo está compuesto por los no creyentes de izquierdas. Entre estos últimos, casi las dos terceras partes da una respuesta positiva al tipo de aborto analizado. No obstante, la estimación del riesgo (0,23) apenas mejora en relación con el ejemplo anterior (0,25) porque el único grupo donde cambia la categoría modal –el descrito como quinto– representa un segmento muy reducido de la muestra, el 5,3%.

El análisis de segmentación permite, pues, realizar una descripción de segmentos de la muestra con comportamiento u opinión distintos entre ellos. Por su propia lógica, tiende a encontrar grupos muy diferentes entre sí. Ahora bien, cuanto mejores sean las variables introducidas, tarea que corresponde al analista, más nítida será la distribución de los distintos grupos. Por tanto, la mejor estrategia en la introducción de variables independientes es la inclusión en caso de duda: si se introduce una poco relevante, el propio análisis se encarga de que no aparezca; en cambio si no se incluye un buen pronosticador, la calidad de la segmentación se reduce considerablemente.

Como tercera muestra de aplicación y uso, se va a utilizar un estudio de los jóvenes burgaleses, cuyo trabajo de campo se realizó en noviembre de 1997. La variable dependiente de este análisis fue la actividad de los jóvenes clasificada en cuatro situaciones básicas: *el trabajo,* donde se incluyeron a todos aquellos que desempeñaban una labor remunerada; *los estudios,* en el caso de que, no trabajando, estuviera matriculado el joven en unos estudios regulares; *el desempleo,* para quienes estuvieran buscando activamente una ocupación y no hubieran quedado clasificados en las dos anteriores categorías, y, finalmente, *la inactividad,* económicamente considerada, donde se consideraron a los no ubicados en los anteriores grupos, comprendiendo seguramente a quienes se dedicaban a actividades del hogar, el servicio militar o la participación como voluntario en una asociación vecinal, política, religiosa o humanitaria. Estas categorías fueron tratadas como mutuamente excluyentes, constituyendo, en consecuencia, una sola variable.

La distribución de la actividad entre los jóvenes de Burgos se conformaba por una mitad (48%) que estaba estudiando, casi otra mitad que trabajaba (40%) y dos minorías: una que buscaba trabajo (9%) y otra considerada económicamente inactiva (3%).

Las variables pronosticadoras que se utilizaron para la descripción de la actividad de los jóvenes burgaleses fueron: el sexo (variable nominal con dos valo-

[20] Nótese que, fruto de la asociación entre religión e ideología, no hay personas de derechas entre los no creyentes.

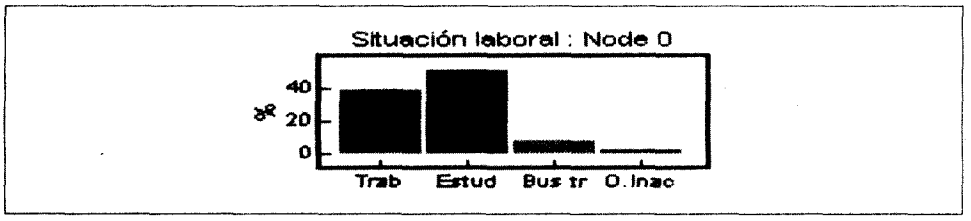

Figura 11. Distribución de la situación en la actividad del joven de Burgos.

res), la edad (variable de intervalo), la clase social (ordinal con cuatro categorías), el distrito (las nueve divisiones municipales de la ciudad), la tendencia política (medida en una escala del 1 al 10) y la posición religiosa (considerada ordinal con cinco posiciones iniciales) [21].

Se encuentran diferencias de actividad por sexo. Así, mientras el porcentaje de hombres y mujeres que estudian es muy similar (48,9% y 48,6% respectivamente), el porcentaje de trabajadores varones es superior en 3,5 puntos porcentuales. También el porcentaje de chicos que buscan trabajo es ligeramente inferior en un punto porcentual. Donde existe una gran diferencia es en el segmento clasificado como otros inactivos, donde el número de mujeres triplica el número de varones. Aunque sea necesario advertir el escaso peso específico de este grupo en el conjunto de la juventud, esta diferencia es debida, sin duda, a que en algunos casos las jóvenes siguen ocupándose de forma exclusiva de las labores del hogar dentro de la pareja.

También cabía esperar que las diferencias de actividad en función de la edad de los jóvenes fueran muy importantes. Se incrementa el número de jóvenes que trabajan, a un ritmo casi lineal de un 5,77% cada año a partir de la edad laboral de los 16 años; de forma que la ocupación de los jóvenes del grupo de edad de 27 a 30 años es del 75%. La dedicación al estudio sigue una tendencia inversa: disminuye el número de estudiantes al incrementar su edad, además su reducción también sigue una tendencia casi lineal a un ritmo del 6,13% anual. Así, los estudiantes de 14 a 17 años son el 91,7% del total de jóvenes, mientras que a la edad de 27 a 30 años sólo son estudiantes el 8% de los jóvenes. El porcentaje de jóvenes que busca trabajo se incrementa con la edad hasta el segmento de 23 a 26 años, donde un 14,3% se encuentra en esta situación y disminuye en el grupo de 27 a 30 años a un 11,4%.

Si se considera la clase social de origen del joven burgalés, también se aprecian diferencias significativas en su dedicación. La distribución de la actividad en los jóvenes de clase baja y media-baja son casi idénticas: en torno a un 44,3% trabaja, el 41% estudia, el 11,3% busca trabajo y el 3,4% tiene otra ocupación. Muy similar es la distribución de los jóvenes cuyas familias de origen son de clase media-media, aunque un menor porcentaje trabaja (40,2%), busca empleo

[21] Estas dos últimas variables fueron introducidas a sabiendas de que teóricamente era irrelevante su inclusión con el fin de verificar que efectivamente las otras eran más importantes.

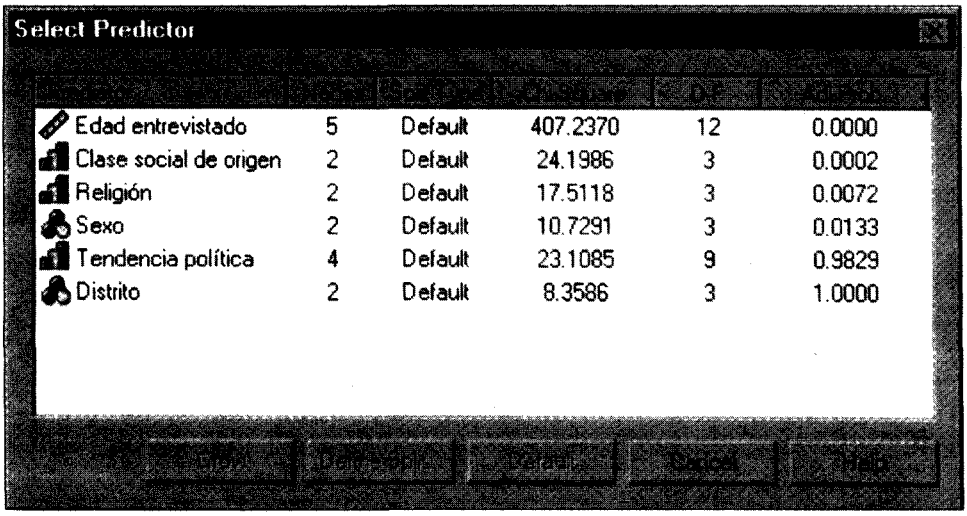

Figura 12. Selección de la variable con mayor χ^2.

(8,4%), o tiene otra actividad (2,9%) y en cambio, es mayor el porcentaje de los que estudian (48,6%). Sin embargo, es muy diferente la distribución de la actividad de los jóvenes cuyas familias de origen son de clase alta. Es muy superior el porcentaje de jóvenes que estudian, casi dos terceras partes 63,6%, y son porcentualmente menos de la mitad los que buscan trabajo o tienen otra actividad (un 5,5% y un 1,4%) respecto del resto de los grupos sociales. Hay dos razones que explican estas diferencias. La mayor concienciación de los padres de clases altas para la formación de sus hijos y las mejores condiciones económicas de aquéllos para afrontar la tardía incorporación al mercado de trabajo de sus hijos.

Vistas en conjunto estas tres variables, resulta esclarecedora una segmentación automática de los jóvenes en función de la actividad que desarrollan. La variable más influyente es la edad. Con ella se forman cinco grupos, que van desde los menores de 18 años, entre los que trabajan menos del 6%, hasta los de más de 28, tres de los que cada cuatro están ocupados.

Ahora bien, lo más relevante en el árbol representado en la figura 13 es que entre los muchachos con edades comprendidas entre los 18 y los 21 años hay sustanciales diferencias por clase social de origen. Quiere ello decir que los hijos de padres de clase media alta y alta tienen más probabilidad de seguir cursando estudios tras alcanzar la mayoría de edad legal. En cambio, en el tramo de edad más avanzada, tras los 28 años, la variable que más diferencia es el sexo pues, mientras más del 90% de los jóvenes varones son activos (ocupados o buscando trabajo), entre las mujeres la tasa de ocupación es del 66,7%, pues poco más de una cuarta parte de ellas estaban sin ocupación y sin estudiar. Es evidente que hay interacción entre la edad, por un lado, y el sexo y la clase social, por el otro. Esta última influye en la actividad del joven en las edades comprendidas entre los 18

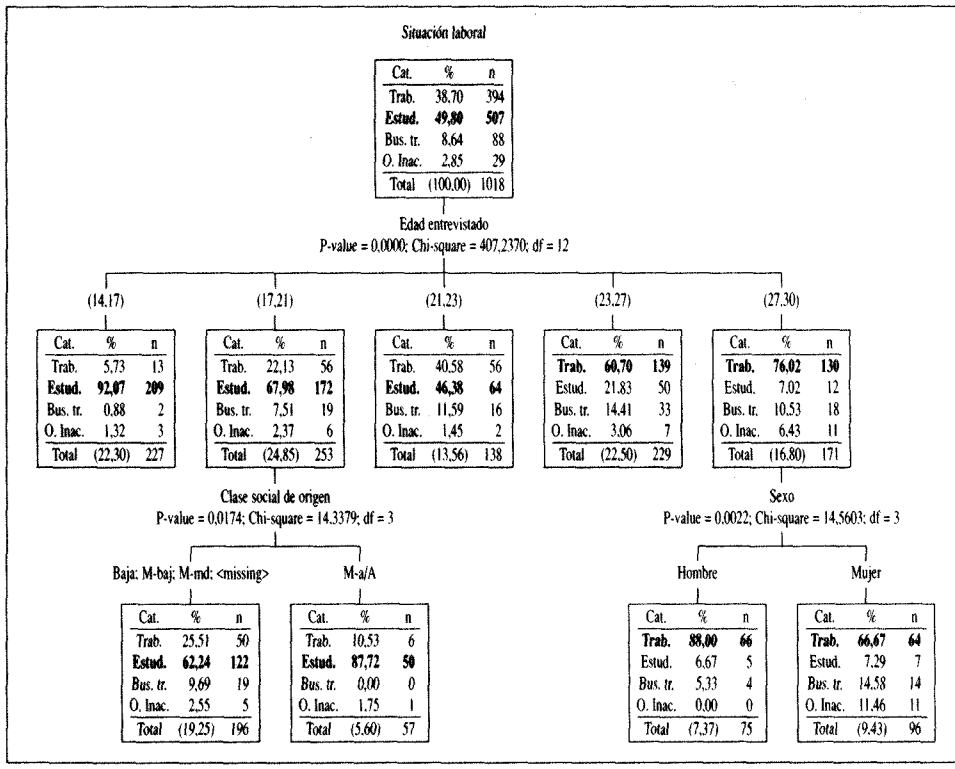

Figura 12. Segmentación de la actividad de los jóvenes.

y 21 años; mientras que el sexo sólo incide en la actividad de los analizados a partir de los 28 años.

En definitiva, se forman siete grupos terminales: hasta los 18 años, prácticamente todos los jóvenes estudian (el 92%) –grupo terminal 1. El grupo 3 también es ampliamente de estudiantes (88%). Son los que están comprendidos entre 18 y 21 y tienen clase media-alta o alta. El grupo 2 está compuesto por dos tercios de estudiantes, teniendo la misma edad que los anteriores; pero incluidos en clase inferior. El grupo 4 (formado por jóvenes entre 21 y 23 años) también posee mayoría –aunque relativa– de estudiantes. Sólo a partir de los 23 años, hay mayoría de jóvenes trabajando (el 60% en el grupo 5); pero, a partir de los 28 el porcentaje de ocupados sube por encima del 75%, aunque segmentándose por sexo, de modo que el grupo 6 está formado exclusivamente por varones, entre los que trabajan el 88%, mientras que en el grupo 7, compuesto por mujeres, hay únicamente dos tercios de ocupadas.

Con estos siete grupos, el análisis de segmentación pronostica con acierto el 70% de los casos como puede observarse en la tabla 15 de clasificación, donde se advierte que de los 1.018 casos contemplados, se pueden clasificar adecuadamente 714. Dicho de otro modo, del conjunto de casos se pueden clasificar bien

a 269 de los 394 trabajadores existentes en la muestra y a 445 de los 507 estudiantes entrevistados. Por tanto, la segmentación es adecuada, máxime cuando el error inicial sería del 40%, por lo que el conocimiento de los segmentos mejora un 25% la predicción de la variable dependiente.

TABLA 15

Matriz de clasificación del análisis de segmentación

			Categoría real			
		Total	**Trab.**	**Estu.**	**Bus. tr.**	**O. Inac.**
Categoría predicha	Trabaja	400	269	62	51	18
	Estudia	618	125	445	37	11
	Busca trabajo	0	0	0	0	0
	O. Inactivos	0	0	0	0	0
	Total	1.018	394	507	88	29

Estimación del riesgo: 0,299

Teóricamente, estos datos pueden interpretarse insistiendo en la consideración de la juventud como una etapa de transición entre los roles infantiles y los adultos. Por ello, la edad es el principal pronosticador de la actividad del joven, porque a lo largo de los años, la persona va pasando de ser un sujeto en formación a convertirse en un individuo integrado en el sistema productivo de la sociedad. Por otro lado, además de la edad, juega una contribución importante la clase social, en primer lugar, porque marca el momento en el que se incorpora el joven al trabajo. Como los de clases bajas tienden a formarse menos, su inserción en el mercado laboral es anterior a la de los hijos de clases altas. Otra variable también importante es el sexo; pero ésta opera posteriormente, a partir de los 28 años, en el sentido de que las mujeres tienen menos esperanza de integrarse en la actividad por un doble motivo, su injustificada discriminación en el mercado de trabajo y, su dedicación exclusiva, al trabajo doméstico, aun considerado dentro del epígrafe económico de la inactividad. Los datos son claros en este aspecto. Mientras una cuarta parte de mujeres entre los 28 y 30 años son clasificadas como inactivas o buscando trabajo, sólo el 5% de los varones de la misma edad se encuentran en similar situación.

4. SUMARIO A MODO DE CONCLUSIONES

El análisis de segmentación es una técnica de análisis de datos basada en la dependencia entre variables, cuya finalidad es la de formar grupos, configurados con valores de las variables independientes, que sean muy distintos entre sí en la variable dependiente. La lógica de su procedimiento se sustenta en los siguientes pasos: a) agrupación de categorías de los pronosticadores, b) selección de los

mejores pronosticadores y c) sucesivas segmentaciones, hasta alcanzar unos límites definidos por los denominados filtros, sobre los grupos formados a partir de los pasos anteriores. Uno de los algoritmos más útiles para sociólogos es el basado en el estadístico χ^2, pues es especialmente indicado para variables dependientes nominales. La utilidad del análisis de segmentación es múltiple. Está especialmente diseñado para propósitos descriptivos, exploratorios e incluso pronosticadores. Además, con ciertas cautelas, también puede ser útil para un previo análisis causal de las variables.

5. BIBLIOGRAFÍA

ALVIRA, F.; GARCÍA LÓPEZ, J. y HORTER, K. (1982): «La situación de la vivienda en España», *Papeles de Economía Española*, 10, pp. 208-247.

BELSON, W. A. (1959): «Matching and Prediction on the Principle of Clasification», *Applied Statistics*, 8, pp. 195-202.

BIGGS, D.; DE VILLE, B. y SUEN, E. (1991): «A Method of Choosing Multiway Partitions for Classification and Decision Trees», en *Journal of Applied Statistics*, 18, pp. 49-62.

BOUROCHE, J. M. y TENNENHAUS, M. (1972): «Some Segmentation Methods», *Metra*, 7, pp. 407-418.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A. y STONE, C. J. (1984): *Classification and Regression Trees*, Belmont, Wadsworth.

ESCOBAR, R. M. (1993): «Afiliación y movilización sindical en España», en AA.VV., *Las Relaciones Laborales en España*. Madrid, UGT/Complutense, 1993.

FIELDING, A. (1977): «Binary Segmentation: The Automatic Interaction Detector and Related Techniques for Exploring Data», en C. A. O'Muircheartaigh y C. Payne (eds.), *The Analysis of Survey Data*, New York, Wiley.

GARCÍA FERRANDO, M. (1982): *Regionalismo y autonomías en España, 1976-1979*, Madrid, Centro de Investigaciones Sociológicas.

HAWKINS, D. M. y KASS, G. V. (1982): «Automatic Interaction Detection» en D. M. Hawkins (ed.), *Topics in Applied Multivariate Analysis*, Cambridge, Cambridge University Press.

HORTER, K. (1978): «Análisis multivariable de los votos político y sindical», *Revista Española de Investigaciones Sociológicas*, 1, pp. 145-158.

KASS, G. V. (1980): «An Exploratory Technique for Investigating Large Quantities of Categorical Data», *Applied Statistics*, 29, pp. 119-127.

LOH, W. y SHIH, Y. (1998): «Split Selection methods for classification trees», en *Statistica Sinica*.

MADGISON, J. (1993): *SPSS for Windows CHAID release 6.0*, Chicago, SPSS Inc.

MESSENGER, R. C. y MANDELL, L. M. (1972): «A Modal Search Technique for Predictive Nominal Scale Multivariate Analysis», *Journal of the American Statistical Association*, 67, pp. 768-772.

MORGAN, J. N. y SONQUIST, J. A. (1963): «Problems in the Analysis of Survey Data», *Journal of the American Statistical Association*, 58, pp. 415-434.

RUIZ-MAYA, L. *et al.* (1990): *Metodología estadística para el análisis de datos cualitativos*, Madrid, Centro de Investigaciones Sociológicas.

SPSS Inc. (1998): *Answer Tree 1.0. User's Guide*, Chicago, SPSS Inc.

SÁNCHEZ-CUENCA, J. (1990): «La segmentación», en E. Ortega, *Manual de investigación comercial*, Madrid, Pirámide.

SMITH, W. (1956): «Product Differentiation and Market Segmentation as Alternative Marketing Strategies», *Journal of Marketing*.

SONQUIST, J. A. y MORGAN, J. N. (1964): *The Detection of Interaction Effects*, Survey Research Center Monograph, n.° 35, Ann Arbor, Institute for Social Research, University of Michigan.

SONQUIST, J. A. (1971): *Multivariate Model Building. The Validation of a Search Strategy*, Ann Arbor, Institute for Social Research, University of Michigan.