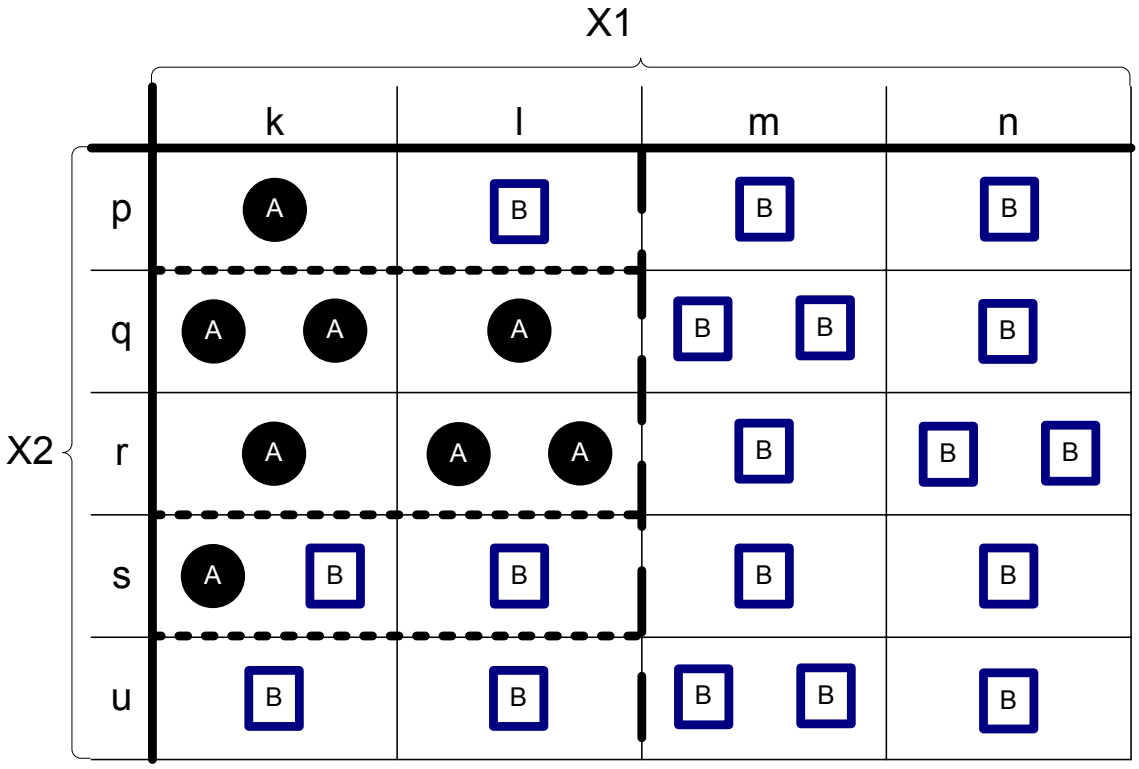
3 La segmentación CHAID.

Dos características específicas sobresalen en el algoritmo AID, acabado de describir: 1) su variable dependiente es de naturaleza cuantitativa y 2) las particiones de la muestra son dicotómicas, lo que implica que los valores predictores deben ser agrupados en dos grupos siempre que se proceda a una segmentación. Expresado con la terminología botánica del árbol es como si de cada tronco o rama, sólo le pudieran salir al mismo tiempo dos ramas u hojas. El procedimiento que se va a contemplar en este segundo capítulo es una derivación del precedente y, frente a él tiene la doble ventaja de poder trabajar con variable dependiente cualitativa y de poder hacer divisiones de sujetos en más de dos grupos al mismo tiempo.

Para hacerse una idea gráfica (Figura 3.1) de este procedimiento, se puede concebir que la variable dependiente posea dos valores (A y B) y que éstos se distribuyen en veinte casillas marcadas por dos variables: una con cuatro valores (k, l, m y n) y otra con cinco (p, q, r, s y t). Dada la distribución de los valores A y B, La mejor partición posible es la efectuada con la primera variable (X1), porque en las columnas m y n sólo se encuentra la característica B. La división representada con la línea vertical discontinua genera a su derecha un grupo completamente homogéneo en B; mientras que deja a su izquierda (en las columnas k y l) un grupo heterogéneo donde predominan las A. Éste último segmento podría subdividirse en cuatro de acuerdo con las tres líneas horizontales discontinuas, que generan grupos muy distintos: el de arriba con un 50% de A y B; el segundo, completamente homogéneo, de A; un tercero con 66,6% de B, y finalmente otro grupo homogéneo de B, es decir, con el 100% de sujetos con esta caracterización en la variable dependiente.

Figura 3.1.- Representación esquemática del procedimiento CHAID



En el capítulo anterior, donde se explicó el procedimiento AID de segmentación, se adelantó que el proceso es idéntico al que va a presentarse a continuación. Recordando lo que se dijo, las etapas de las sucesivas divisiones que se efectúan a la muestra son las siguientes:

1) Preparación de variables, que incluye su selección y la determinación de la dependiente (o criterio) y las independientes o pronosticadoras.

2) Fusión de categorías, proceso mediante el cual los distintos valores de las variables independientes se unifican siempre y cuando presenten perfiles muy similares con la variable dependiente.

3) Selección del mejor predictor, esto es, una vez fusionadas las categorías pertinentes, se decide qué variable independiente divide mejor a la muestra en función de su perfil en la variable dependiente.

4) Una vez realizada una división de la muestra, se vuelven a aplicar los pasos segundo y tercero, no a la muestra global, sino a cada uno de los grupos conformados por estos procesos.

La segmentación CHAID se distingue de la AID básicamente en dos aspectos: en primer lugar, la variable dependiente no ha de ser necesariamente cuantitativa. Puede ser nominal u ordinal, lo que implica que, en lugar de realizar pruebas estadísticas basadas en la distribución F, se emplean otras fundamentadas en el χ^2, y en segundo lugar, aunque puedan aplicarse en CHAID particiones binarias, lo más propio de su procedimiento es un sistema de agrupación de categorías más significativas o un sistema de exploración exhaustiva, que no limitan las segmentaciones a dos grupos al mismo tiempo. Dicho de modo más intuitivo, el procedimiento AID sólo permitía en cada paso segmentaciones en dos trozos, mientras que en CHAID son posibles subdivisiones en dos, tres, cuatro o más fragmentos en la misma operación.

Para explicar este procedimiento, se recurre a un ejemplo muy similar al anterior. Sólo se aplican de partida dos cambios. En primer lugar, es obvio que la variable dependiente ha de ser distinta, para emplear una variable nominal. Por ello se ha substituido la actitud hacia los norteafricanos por la opinión sobre la necesidad en España de trabajadores inmigrantes[6]. Ésta será la nueva variable dependiente, con dos posibles valores: "Si" y "No"[7]. Para formar grupos homogéneos con esta técnica, se ha de elegir una serie de características medidas nominal u ordinalmente. En el caso de este ejemplo, se han empleado las mismas variables que en el anterior: sexo ("hombre", "mujer"), edad ("menos de 45 años" y "más de 45"), e ideología ("izquierda", "derecha" y "ns/nc"); pero para mayor simplificación esta última se ha empleado con dos categorías, en lugar de tres, eliminando para ello la categoría del centro.

[6] Para este ejemplo también se utilizarán los datos del estudio 2511 del Centro de Investigaciones Sociológicas. Se trata de una muestra de tamaño 2495 realizada en mayo de 2003. Más detalles de estos datos pueden encontrarse en el catálogo de encuestas disponible en su página web: http://www.cis.es

[7] En el cuestionario y en los datos existe un tercer valor: el "No sabe/no contesta". Sin embargo éste es considerado como valor no disponible (*missing*) y en la mayoría de casos, si no es un dato con significación propia en el conjunto de la variable, es conveniente omitirlo en las variables dependientes del análisis de segmentación. No ocurre así, como se verá más adelante en las variables independientes empleadas en esta técnica.

En la tabla 1 se pueden contemplar una visión global de los datos que se van a utilizar, divididos en 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 (última fila correspondiente al total) y dos porcentajes relativos a cada uno de los valores de la variable dependiente, en este caso, percepción de la necesidad de trabajadores inmigrantes.

El segmento más numeroso es el correspondiente a los hombres menores de 45 años (n=387) de izquierda, seguido por el de las mujeres jóvenes de la misma ideología (342). Por el contrario, los grupos menos numerosos son los de hombres jóvenes y mayores que no contestan sobre su ideología (104 y 84 respectivamente). Al observar la variable dependiente (el porcentaje de los que piensan que son necesarios los trabajadores inmigrantes) se obtiene un perfil distinto para cada uno de los 12 segmentos formados por las tres variables pronosticadoras: Los más partidarios de la inmigración (con porcentajes superiores al 60%) son los hombres con ideología, especialmente, los de derecha (69,6%) y el grupo con menor porcentaje de sujetos que aprueban esta práctica (41,4%) es el de la mujeres jóvenes sin ideología declarada.

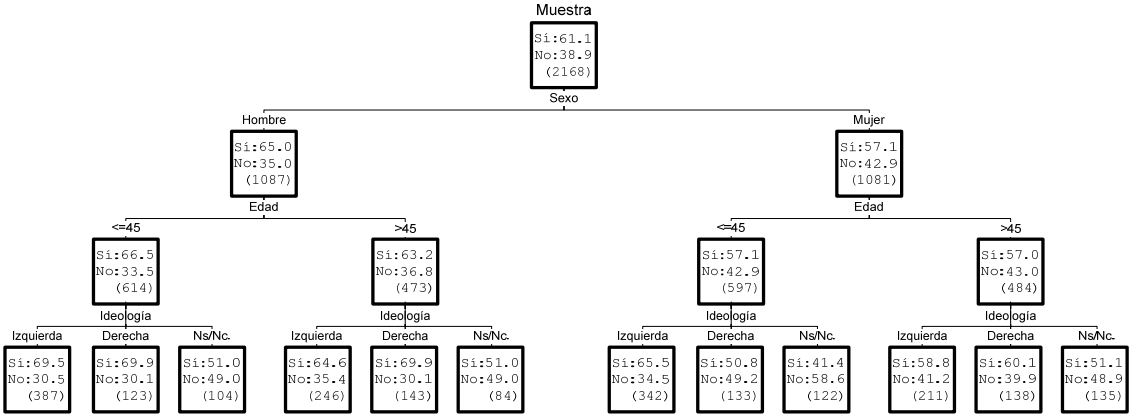
Figura 3.2.- Necesidad de trabajadores inmigrantes según ideología, edad y sexo.

		Total	Sexo											
			Hombre						Mujer					
			Edad						Edad					
			<=45			>45			<=45			>45		
			Ideología (2m)			Ideología (2m)			Ideología (2m)			Ideología (2m)		
			Izq.	Der.	Ns/Nc	Izq.	Der.	Ns/Nc	Izq.	Der.	Ns/Nc	Izq.	Der.	Ns/Nc
Necesidad de trabajadores inmigrantes	Sí	61,1%	69,5%	69,9%	51,0%	64,6%	69,9%	47,6%	65,5%	50,8%	41,4%	58,8%	60,1%	51,1%
	No	38,9%	30,5%	30,1%	49,0%	35,4%	30,1%	52,4%	34,5%	49,2%	58,6%	41,2%	39,9%	48,9%
	n	2168	387	123	104	246	143	84	342	122	133	211	138	135

Fuente: CIS (2003). Estudio 2511.

La Figura 3.2 es en realidad una tabla de contingencia o cruce formado por cuatro variables dispuestas en cuatro dimensiones (una en las filas y tres en las columnas). La técnica de la segmentación tiene una estructura similar. En la Figura 3.3 se muestra un árbol de clasificación, basado en los datos de la tabla mencionada. En cada rectángulo se incluye el valor de la variable pronosticadora que conforma 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, aunque lo parezca, 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. Antes al contrario, se ha construido manualmente, para reflejar la semejanza de éste con la estructura de la tabla

Figura 3.3.- Árbol de segmentación CHAID



A semejanza con el método AID, para realizar una segmentación automática de acuerdo al método CHAID, es preciso construir este árbol de acuerdo a los pasos mencionados anteriormente, que se explicarán con más detalle a continuación.

3.1 Pruebas de significación

Aunque inicialmente, como su propio nombre indica, el procedimiento CHAID utilizara el estadístico χ^2 de Pearson, esto no es exacto en la actualidad, pues utiliza cuatro métodos distintos de comprobación de si una relación es o no significativa. En este apartado, algo más complejo que otros, se van a estudiar con detalle estas pruebas de significación. Se advierte al lector no muy versado en estadística que no es necesaria la comprensión total de este apartado para seguir la lectura posterior. Incluso puede ser soslayado completamente si es así su voluntad.

La utilización de una u otra prueba de significación depende del nivel de medida de la variable dependiente, pues se supone que las variables independientes, predictores o pronosticadores están medidos en una escala nominal. De este modo, se presentan los siguientes casos:

1) Variable dependiente nominal:

En este caso hay dos posibilidades de uso: bien se puede emplear el ya mencionado χ^2, bien se puede emplear, con resultados muy similares la razón de verosimilitud. A continuación, se explican con más detalle ambos.

Cuando se cruzan dos variables nominales (una dependiente, que se denomina Y, y otra independiente, representada como X) se está ante una tabla de contingencia formada por I filas y J columnas. En la intersección entre ambas se encuentran las celdas cuyo contenido es una frecuencia conjunta f_{ij}, es decir, el recuento de casos que comparten el valor *i* de la variable Y y el *j* de la variable X.

Figura 3.4.- Notación de una tabla de contingencia

	X			
Y	f_{11}	f_{12}	f_{13}	$f_{1.}$
	f_{21}	f_{22}	f_{23}	$f_{2.}$
	f_{31}	f_{32}	f_{33}	$f_{3.}$
	$f_{.1}$	$f_{.2}$	$f_{.3}$	n

En los márgenes de la tabla (a la derecha de la línea vertical y por debajo de la línea horizontal), se suelen ubicar los marginales, que son sumas de las frecuencias que les anteceden y representan las frecuencias simples de las variables X e Y respectivamente.

De este modo las frecuencias marginales para cada uno de los valores *j* de la variable X pueden obtenerse con el siguiente sumatorio:

$$f_{.j} = \sum_{i=1}^{I} f_{ij}$$

Y las correspondientes a cada uno de los valores *i* de la variable Y, se consiguen sumando las frecuencias de todas las columnas en cada fila:

$$f_{i.} = \sum_{j=1}^{J} f_{ij}$$

Además de la frecuencia conjunta y las marginales, empíricamente obtenidas, en cada casilla de la tabla se pueden generar frecuencias esperadas (f^*_{ij}) en el supuesto de que no hubiera ninguna relación entre las variables cruzadas. Quiere ello decir, frecuencias teóricas en las que se cumpla la hipótesis nula con el supuesto de independencia entre las variables:

$$h_0 : X \xrightarrow{0} Y \Leftrightarrow f_{ij} = f_{ij}^*, \forall_{i,j}$$

Como operacionalmente existe independencia cuando una probabilidad conjunta es igual al producto de las probabilidades marginales ($p_{ij}=p_i x p_j$), la obtención de la frecuencia *independiente* se obtiene multiplicando lo anterior por el tamaño de la muestra:

$$f_{ij}^* = n \times \frac{f_{i.}}{n} \times \frac{f_{.j}}{n} = \frac{f_{i.} \times f_{.j}}{n}$$

Es el momento de ver un ejemplo para visualizar este procedimiento de construcción de frecuencias teóricas. Para ello, se ha seleccionado la comparación entre las personas con ideología expresada de izquierdas y las de derechas:

Las frecuencias empíricas del cruce entre la necesidad de trabajadores inmigrantes y las dos categorías ideológicas son las siguientes:

Figura 3.5.- Frecuencias empíricas de necesidad de inmigrantes según izquierda o derecha.

Recuento

		Ideología		
		Izq.	Der.	Total
Necesidad de trabajadores inmigrantes	Sí	776	331	1107
	No	410	195	605
Total		1186	526	1712

Y ya que dos sucesos son independientes si la frecuencia de ambos es el producto de sus respectivas probabilidades, los valores esperados, la tabla de frecuencias esperadas bajo esta suposición sería la mostrada en la Figura 3.6. Como botón de muestra, la primera casilla teórica (766,9) se obtiene con esta operación 1107×1186/1712.

Figura 3.6.- Frecuencias teóricas de necesidad de inmigrantes según izquierda o derecha.

Frecuencia esperada

		Ideología		
		Izq.	Der.	Total
Necesidad de trabajadores inmigrantes	Sí	766,9	340,1	1107,0
	No	419,1	185,9	605,0
Total		1186,0	526,0	1712,0

Una vez que se disponen, por un lado, las frecuencias empíricas y, por el otro, las teóricas, pueden realizarse dos pruebas estadísticas homólogas para compararlas.

La primera de ellas es la prueba del χ^2 de Pearson, consistente en el sumatorio de los residuos estandarizados al cuadrado. Dado que el residuo estandarizado o típico de una casilla viene definido por la siguiente expresión:

$$r_{ij}^{s} = \frac{f_{ij} - f_{ij}^{*}}{\sqrt{f_{ij}^{*}}}$$

La fórmula de χ^2 se expresa como sigue:

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

A continuación se expone la tabla que contiene en cada casilla el residuo típico al cuadrado, cuya suma da lugar al estadístico buscado:

Figura 3.7.- Residuos típicos al cuadrado y χ^2 de necesidad inmigrantes según ideología.

Residuos tipicos al cuadrado

		Ideología	
		Izquierda	Derecha
Necesidad inmigrantes	Sí	0,108	0,244
	No	0,198	0,448

chi2= 0,998; G^2=0,995; p=0,32

La segunda medida para verificar la significación entre el pronosticador y la variable dependiente es la razón de verosimilitud (G^2), basada también en la comparación entre frecuencias empíricas y teóricas, pero mediante una formulación fundamentada en el criterio de máxima verosimilitud. (Haberman 1978 y Goodman 1979).

$$G^2 = 2\sum_{j=1}^{J}\sum_{i=1}^{I} f_{ij} \ln \frac{f_{ij}}{f_{ij}^*}$$

En el trabajo de comparación de modelos, G^2 ofrece ventajas adicionales a χ^2, por lo que es recomendable trabajar con el primero, a pesar de que los resultados de ambos sean bastante similares. Parte de estas ventajas derivan de la consideración de G^2 como una diferencia de las razones de verosimilitud entre dos modelos. En este caso, los modelos comparados son el modelo saturado compuesto por efectos medios (η), de fila, de columna y de asociación

$$f_{ij} = \eta \tau_i^A \tau_j^B \tau_{ij}^{AB}$$

con el de independencia, en el que sólo están presente los efectos medio, de fila y de columna de una tabla. Esto implica que el G^2 mide estrictamente el efecto asociación entre la variable dependiente y la independiente.

$$f_{ij}^* = \eta \tau_i^A \tau_j^B$$

b) Variable dependiente ordinal.

Madgison (1992) propuso aplicar un contraste diferente en el caso de que la variable dependiente fuera ordinal. De acuerdo con los trabajos de Goodman (1979), aquel autor consideró que sólo se considerara el efecto columna[8] de la asociación ordinal, puesto que es el único que representa a la variable independiente. Para estos casos en los que sólo se tiene en cuenta la asociación ordinal de los efectos de columna, el modelo adopta la siguiente forma:

$$f_{ij}^{'} = \eta \tau_i^A \tau_j^B \delta_j^{\ i}$$

De este modo se descartan todas las fuentes de dependencia distintas de la ordinal, se emplean menos grados de libertad y la prueba tiene mayor potencia o menor riesgo de cometer un error de tipo II (aceptar una hipótesis nula que sea falsa).

Comparada con el modelo de asociación completa (el saturado), se ha de advertir que en la anterior ecuación el término τ_{ij}^{AB} ha sido reemplazado por $\delta_j^{\ i}$, donde δ_j es un parámetro distinto para cada valor de la variable independiente, es decir, del predictor, que se eleva a una potencia (*i*) distinta para cada grado ordinal de la variable

[8] Madgison (1992 y 1993) propone en realidad el modelo R (filas) de Goodman, con el supuesto de que la variable independiente se encuentra ubicada en éstas. Sin embargo, en este libro, donde la mayor parte de las tablas es la variable dependiente la que figura en la dimensión horizontal, se prefiere y se opta por emplear el modelo C (columnas). En esencia ambas opciones son idénticas, aunque con distinta ubicación espacial de variable dependiente o independiente.

dependiente. Y, comparado con el modelo de independencia con (I-1)(J-1) grados de libertad, la diferencia es la introducción del término δ_j, que le hace tener (I-2)(J-1) grados de libertad.

Empleando este modelo de relación entre variables, sólo es adecuado el uso de G^2, por lo que ha de evitarse el uso del χ^2. Y el modelo de asociación ordinal ha de ser comparado con el de independencia, por lo que su cálculo se conforma a la siguiente fórmula:

$$G^2 = 2\sum_{j=1}^{J}\sum_{i=1}^{I} f_{ij}^{'} \ln(\frac{f_{ij}^{'}}{f_{ij}^{*}})$$

Un ejemplo puede aclarar toda la abstracción expuesta anteriormente en fórmulas matemáticas.

Se va a cruzar el juicio sobre el número de inmigrantes realizado por los encuestados con su edad. La primera variable, la dependiente, presenta los valores "demasiados", "bastantes" y "pocos". La segunda como en ejemplos anteriores aparece dicotomizada en dos valores: "<=45" y ">45". La tabla empírica resultante se expone en la Figura 3.8.

Figura 3.8.- Frecuencias empíricas del cruce del juicio sobre inmigrantes con la edad.

Modelo saturado (f_{ij})		Edad		
		<=45	>45	Total
Número de inmigrantes en España	Demasiados	595	598	1193
	Bastantes	595	405	1000
	Pocos	68	39	107
		1258	1042	2300

A partir de esta tabla, el modelo de independencia es fácil de obtener mediante el producto de los correspondientes marginales de fila y columna partidos por el marginal total:

Figura 3.9.- Modelo de independencia en el cruce del juicio sobre inmigrantes con la edad.

Modelo de independencia (f_{ij}^{*})		Edad		
		<=45	>45	Total
Número de inmigrantes en España	Demasiados	653	540	1193
	Bastantes	547	453	1000
	Pocos	59	48	107
		1258	1042	2300

Sin embargo, la obtención de las frecuencias del modelo de cuasi-independencia (C) es bastante más complejo y necesita un procedimiento iterativo expuesto por Goodman (1979, pp. 549 y ss.). Mediante su aplicación, se obtiene la siguiente tabla:

Figura 3.10.- Modelo de cuasi-independencia del jucio sobre inmigrantes con la edad.

Modelo de cuasi-independencia (C) ($f_{ij}^{'}$)		Edad		
		<=45	>45	Total
Número de inmigrantes en España	Demasiados	599	594	1193
	Bastantes	588	412	1000
	Pocos	72	35	107
		1258	1042	2300

G2=23,19

Es fácilmente observable que en esta nueva tabla las frecuencias no son equiprobables en las dos columnas (v.gr. 599/1258≠594/1042) como lo eran en el caso de la Figura 3.9. Sin embargo, se distingue de la tabla de la Figura 3.8 en que los productos de razones cruzadas θ_{11} [(599*412)/(594*588)] y θ_{21} [(588*35)/(412*72)] son idénticos, como corresponde a una relación ordinal monótona.

El G^2 resultante (23,19) es fruto de la comparación de la tabla de la Figura 3.10 con la de la Figura 3.9, aplicando la fórmula correspondiente. Este estadístico tiene una distribución con J-1 grados de libertad, obtenidos a partir de la resta de los grados de libertad de una y otra tabla: (I-1)(J-1)-(I-2)(J-1). En este caso, aquel valor con un grado de libertad tiene una significación claramente inferior a 0,05, luego la asociación ordinal entre la edad y el juicio sobre el número de inmigrantes presentes en España es significativa.

c) Variable dependiente de intervalo.

Aunque no sea propio del análisis CHAID el que la variable dependiente sea cuantitativa de intervalo, se menciona en este apartado por un doble motivo. En primer lugar, porque puede realizarse un análisis mixto, que aun empleando el criterio de la F, propio del algoritmo AID, no realice particiones binarias con los valores de la variable independiente, sino fraccione la muestra de 2 a tantas divisiones como valores tenga el pronosticador, tal como se explica con más detalle en el próximo apartado.

La prueba estadística de la F –también conocida como análisis de la varianza- consiste en la división de la media cuadrática externa por la interna, cuyas fórmulas se vieron en la Sección 2.1. La única diferencia en este algoritmo del anterior reside en la posibilidad de comparar al mismo tiempo más de dos grupos. En consecuencia para realizar comparaciones entre distintas variables, es necesario recurrir a las significaciones, en lugar de fijarse en los valores F o en las sumas cuadráticas externas.

Evidentemente hay cierta conexión entre los estadísticos χ^2 y F. En el caso de que la variable independiente o pronosticador tenga sólo dos categorías o valores, el resultado de ambas pruebas es idéntico, puesto que ambas distribuciones son iguales si

$$\chi_n^2 = F_{1,n}$$

Es decir, la distribución χ^2 con n grados de libertad es idéntica a la distribución F con 1 grado de libertad en el numerador y n en el denominador.

3.2 Fusión de las categorías.

Esta operación consiste en obtener 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. De lo que se trata, por consiguiente, es de fundir categorías de la variable independiente siempre y cuando presenten perfiles muy semejantes en la variable dependiente los casos que las conforman. Expresado con el símil de la tarta, si se trata de obtener trozos blancos y se dispone de tres sabores como predictores (nata, coco y chocolate), es obvio que los dos primeros pueden fusionarse porque ambos poseen colores idénticos.

Según las características de las categorías de las variables pronosticadoras los procedimientos son ligeramente distintos:

a) 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*[9]).

b) 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. 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. En el caso de que haya valores perdidos, se procede de modo similar, pero la categoría sin dato sustantivo, 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.

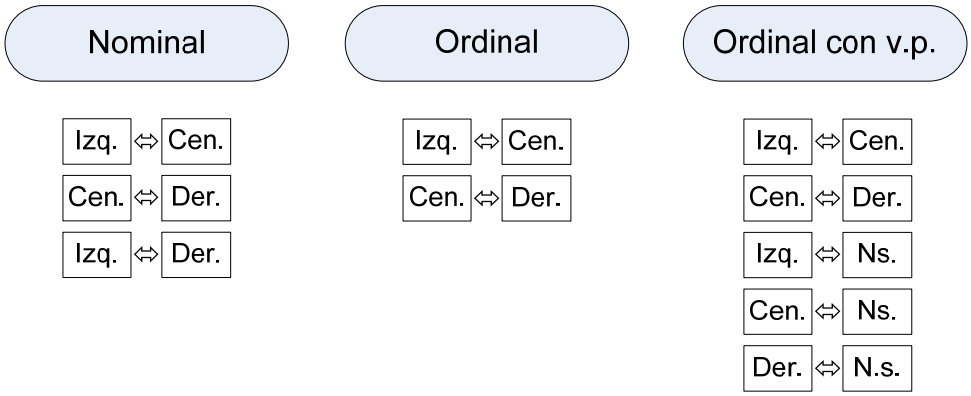
c) Variables cuantitativas: Las variables cuantitativas, para ser utilizadas en el procedimiento CHAID, tienen que ser recodificadas en valores discretos y ser tratadas como si fueran ordinales. Esto puede hacerse bien manualmente, bien por intervalos regulares en amplitud, bien por intervalos que recojan aproximadamente el mismo número de casos. En muestras grandes, conviene utilizar amplitudes de intervalos muy pequeñas, para que sea el propio comportamiento de la variable en relación con la dependiente lo que decida la agrupación de los valores.

[9] Esta última es la terminología que emplea Kass (1980), Madgison (1982, 1992 y 1993) y el programa *Chaidwin*.

El funcionamiento de formación de grupos de categorías homogéneas se basa en el estadístico χ^2 o F, según la variable respuesta sea cualitativa o cuantitativa. 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 ideología, que presenta en este ejemplo tres valores, el número posible de pares sería de 3 (combinaciones de 3 elementos tomados de dos en dos). Si se opta por la opción ordinal, sólo es posible un par (número de categorías válidas menos una) excluyendo los valores perdidos. Y si se escoge la opción ordinal con valores perdidos, las posibilidades también serían 3 en este caso particular con tres categorías (dos veces el número de categorías menos 3)[10]. Más interesante sería comparar las posibles comparaciones en el caso de que la variable dispusiera de tres valores más una categoría perdida. Véase la Figura 3.11.

Figura 3.11.- Comparaciones pareadas posibles según tipo de variable.



En el ejemplo presente, la variable ordinal ideología sólo contempla dos valores (izquierda y derecha) más la categoría de los que no saben o no contestan a la pregunta. Ya que ésta última categoría tiene significado y entidad suficiente (20% de los casos) como para no considerarla como perdida, en realidad, hay que tratarla como una variable nominal con tres valores –ya que las no contestaciones no pueden ser ubicadas ordinalmente en el conjunto de la variable.

2) Para cada posible par se calcula el χ^2 de Pearson, el estadístico de máxima verosimilitud (G^2) o el cociente de medias cuadráticas (F), según la variable dependiente sea nominal, ordinal o numérica.

Esta operación ha de realizarse con todos los pares posibles de un determinado predictor. En el supuesto de una variable dependiente nominal, con las categorías que se están empleando de la variable ideología, tendría que hacerse también con el par (iquierda;ns/nc) y con el par (derecha;ns/nc). En las tres líneas de la Figura 3.12 se ofrecen los resultados.

[10] En general, para variables nominales c(c-1)/2; para ordinales c-1, y para éstas con valores perdidos 2c-3. Goodman 1979

Figura 3.12.- χ2 de los contrastes pareados de categorías de la variable ideología.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Izquierda vs Derecha	,998	1	,318
Izquierda vs Ns/Nc.	43,866	1	,000
Derecha vs Ns/Nc.	23,305	1	,000

El par con el estadístico más bajo, χ^2 en este ejemplo, *siempre que no sea significativo, 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. De este modo la primera fusión que habría que efectuar sería la que agrupe a personas de ideología de izquierda y derecha, puesto que entre sí son más homogéneas, que entre ellas, por un lado, y los que no contestan a la pregunta sobre su ideología, por el otro.

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.

En el ejemplo que se contempla actualmente, las categorías izquierda y derecha pasarían a formar una sola y con esta nueva configuración, se volvería a repetir el mismo procedimiento que en el paso anterior. Lo que sucedería en esta ocasión es que sólo resta una categoría, por lo que sólo podría probarse la fusión con ésta. La única posible comparación es, por tanto, la reflejada en la Figura 3.13.

Figura 3.13. Cruce de necesidad de inmigrantes según ideología tras un par fusionado.

% de Ideología

		Ideología		
		Izq+Der	NS/NC	Total
Necesidad de trabajadores inmigrantes	Sí	64,7%	47,6%	61,1%
	No	35,3%	52,4%	38,9%
Total		100,0%	100,0%	100,0%

Y, para comprobar si existen diferencias significativas con las nuevas pruebas estadísticas de pares de categorías, se recurre como anteriormente al χ^2. En este caso con una diferencia de porcentajes mayor de 17 punto (64,7% y 47,6%) y más de 2000 casos las diferencias, reflejadas en la Figura 3.14, son más que evidentes.

Figura 3.14.- χ2 del cruce de necesidad de inmigrantes según ideología tras un par fusionado.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	44,151	1	,000

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.

La Figura 3.14 muestra un ejemplo de detención del proceso de fusión. Aunque aún haya dos categorías, éstas no pueden fusionarse, porque entre ellas se advierten diferencias significativas mediante el χ^2.

Biggs et al. (1991) criticaron este procedimiento progresivo de fusiones y, en su lugar, propusieron que para cada variable se efectuaran todas las fusiones de las categoría en orden, es decir, empezando por los pares de categorías que muestren χ^2 o F menor hasta quedarse con sólo dos categorías y, una vez obtenidas las (c-1) agrupaciones de valores, se seleccionara aquella división con la significación corregida menor. Esto supone una mayor cantidad de operaciones de cálculo; pero con las máquinas actuales no deja de ser una cuestión de segundos. A esta distinta versión de proceder en la fusión de categorías se le denomina CHAID exhaustivo. A continuación se expone un ejemplo con una variable cuantitativa, con los mismos datos empleados hasta el momento; pero realizando una conversión automática de la variable edad en una variable ordinal.

La variable edad en el estudio 2511 del CIS presentó un rango entre los 18 y los 96 años. Los intervalos creados automáticamente por el programa de segmentación fueron los siguientes:

Figura 3.15.- Distribución original de la edad.

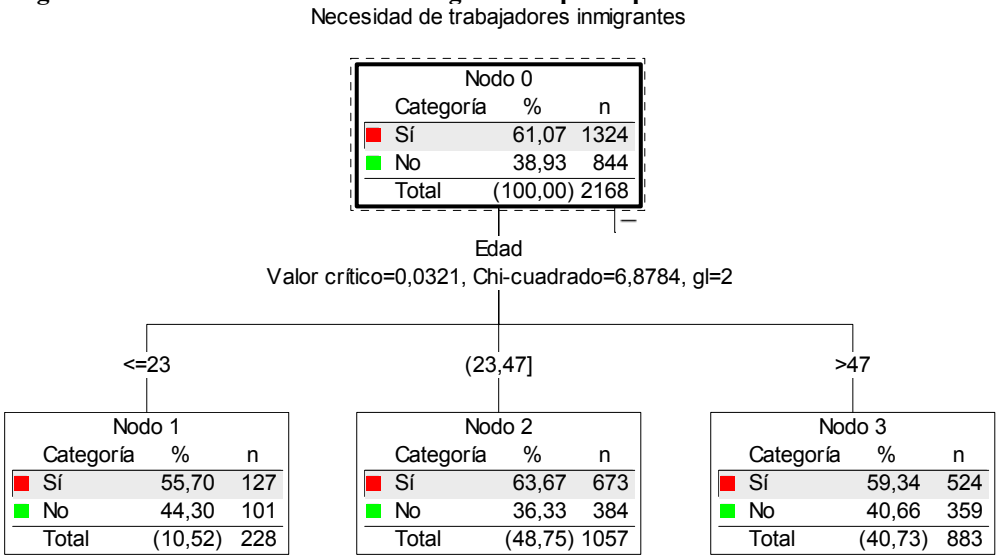
Edad

		Frecuencia	Porcentaje
Válidos	18-23	264	10,6
	24-27	224	9,0
	28-32	242	9,7
	33-37	245	9,8
	38-42	215	8,6
	43-47	239	9,6
	48-55	283	11,3
	56-63	234	9,4
	64-71	280	11,2
	72-96	269	10,8
	Total	2495	100,0

Aplicando el procedimiento de fusiones progresivas hasta las dos categorías: (con estos datos, a) hasta los 23 y b) más allá de los 24 años), se obtiene una significación no corregida de 0,08. Sin embargo, en el paso anterior de fusión, cuando se comparan tres grupos (distinguiendo en el grupo b entre los de menos y más 47 años), la significación es 0,03, es decir, más indicada, tal como se muestra en la Figura 3.16. Nótese en todo caso que el resultado está hecho sin el ajuste de Bonferroni (véase el próximo apartado),

pues de lo contrario, no sería significativo, ya que la edad posee muchas categorías y el resultado puede haberse debido al hecho de realizar muchas pruebas de significación. En cualquier caso, parece que los más partidarios son los adultos de mediana edad, mientras que muestran más antagonismo las personas de mayor edad y especialmente los jóvenes.

Figura 3.16.- División de la muestra según edad por el procedimiento CHAID exhaustivo.



3.3 Selección de los mejores pronosticadores.

Una vez que para cada pronosticador se ha realizado la fusión oportuna de categorías, el siguiente paso sería la selección de los mejores pronosticadores. Para proceder a ello, hay que calcular para cada una de las variables su correspondiente χ^2, G^2 o F y comparar las significaciones obtenidas del estadístico pertinente; sin embargo, es conveniente en este proceso modificar la significación de cada pronosticador con el ajuste de Bonferroni, porque la probabilidad de obtención de un resultado significativo aumenta artificialmente con la proliferación de pruebas estadísticas que implica este análisis.

El ajuste de Bonferroni (Kass, 1980 y Hawkins y Kass, 1982) consiste en la aplicación de la desigualdad establecida por el mismo autor. Según ésta, 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 número posible de pruebas de significación (B) se puede calcular a través de fórmulas combinatorias a partir del número de categorías iniciales del predictor (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, si se escoge la opción sin restricciones de las variables nominales la fórmula es la siguiente:

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

Si se utiliza la fusión monótona, propia de las variables ordinales sin casos perdidos, el número de pruebas para formar k grupos, también depende del número de categorías (c) de la variable:

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

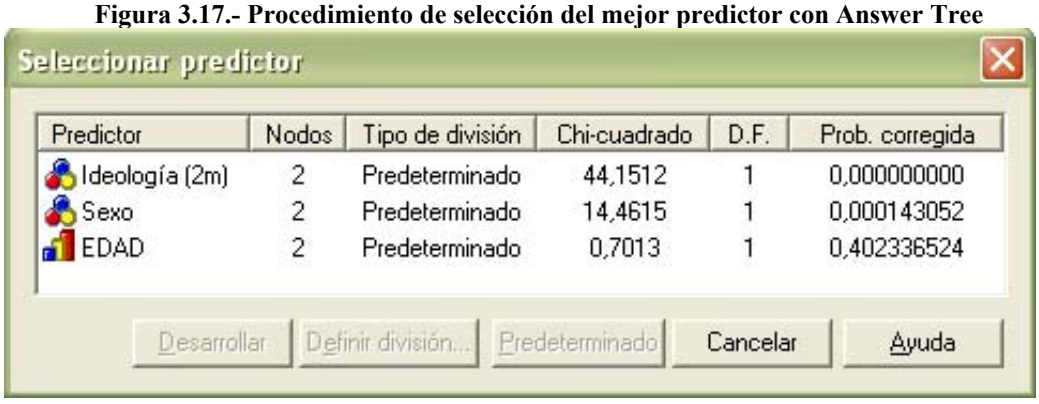
Y si los casos perdidos pueden fusionarse con cualquier número de variables, el número de contrastes efectuados pasa ser:

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

En la práctica, para evitar el riesgo de rechazo inadecuado de hipótesis por realizar múltiples ensayos, hay que multiplicar la significación del χ^2 por el resultado de B_n, B_o o B_{om}, según sea el caso[11].

Con el ejemplo de la variable ideología, tratada en este caso como ordinal con casos perdidos (izquierda, derecha y ns/nc), el resultado de aplicar a la significación (3,04E-11) del χ^2 (44,15 con 1 grado de libertad) un B_{om} de 3 es (9,12E-11), por lo que en la Figura 3.17 no aparece ninguna cifra significativa en la probabilidad del predictor en cuestión.

Figura 3.17.- Procedimiento de selección del mejor predictor con Answer Tree



Tras ajustar la significación de todos los pronosticadores, queda la comparación entre ellos. El ejemplo actual de predicción de la opinión ante la inmigración en función de sexo, edad e ideología ayuda a entender este proceso. En la Figura 3.17 se muestra el χ^2 de las distintas variables, a las que ya se ha aplicado el proceso de agregación de categorías y la corrección probabilística acabada de explicar. Del conjunto de 2168 sujetos que han sido entrevistados, 1087 son varones y 1081 mujeres. El 65,0% de los primeros es favorable al trabajo de los inmigrantes y sólo el 57,1% de las mujeres sostiene la misma posición. El χ^2 (14,5) tiene una significación corregida de 0,0001. Existe, pues, relación significativa; pero antes de proceder a seleccionar este pronosticador, es necesario analizar el resto de los seleccionados para este ejemplo: La

[11] En el caso de que se aplique el algoritmo del CHAID exhaustivo, los valores de B son $c(c^2-1)/2$ para las variables nominales y $c(c-1)/2$ para las ordinales, tengan o no valores perdidos.

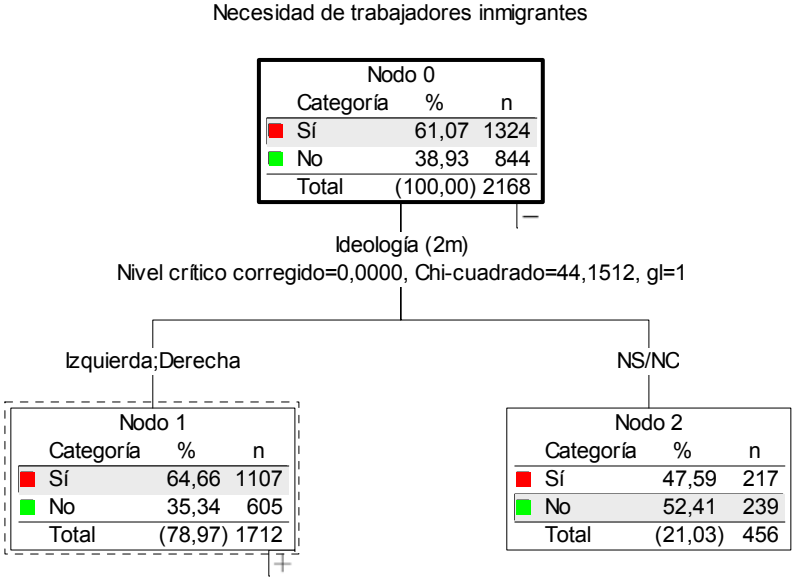
muestra está repartida entre 1211 sujetos con menos de 46 años y 957 de mayor edad. En este caso, entre estos dos segmentos de la muestra no hay diferencias significativas en la opinión. El 61,8% de los jóvenes contestan afirmativamente; pero sólo un 60,1% de los de edad más avanzada mantienen esta actitud. El χ^2 presenta en consecuencia un valor muy pequeño (0,70) y, por tanto, no significativo (p=0,40). Evidentemente, el mejor pronosticador es la ideología. Dos terceras partes de los sujetos que han contestado a la ideología dicen que son necesarios los trabajadores inmigrantes, mientras que menos de la mitad (47,6%) de los que no declararon su ideología mantienen la misma opinión. Lógicamente, el χ^2 es el mayor de los tres (44,2) y la significación, la más baja (0,0000). Por tanto, este es el mejor pronosticador de los tres y es el que se utilizará para realizar la primera segmentación de la muestra.

Figura 3.18.- Cruce bivariado de necesidad de inmigrantes según sexo, edad e ideología.

		Total	Sexo		Edad		Ideología (d)	
			Hombre	Mujer	<=45	>45	Izq+Der	NS/NC
Necesidad de trabajadores inmigrantes	Sí	61,1%	65,0%	57,1%	61,8%	60,1%	64,7%	47,6%
	No	38,9%	35,0%	42,9%	38,2%	39,9%	35,3%	52,4%
	Total	2168	1087	1081	1211	957	1712	456

De este modo, quedarán formados dos grupos: uno de 1712 sujetos (los que declararon ideología) y otro de 456 individuos que no la manifestaron en el cuestionario.

Figura 3.19.- Primera segmentación de la necesidad de trabajadores emigrantes.



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 o derecha existen diferencias considerables de sexo o edad. Así, en la Figura 3.20 se observa que los hombres con ideología manifestada son significativamente más favorables al trabajo de inmigrantes que las mujeres de idéntica respuesta ideológica (68,3% vs. 60,6%).

Figura 3.20.- Necesidad de inmigrantes según sexo y edad entre entrevistados con ideología.

		Total	Sexo		Edad	
			Hombre	Mujer	<=45	>45
Necesidad de trabajadores inmigrantes	Sí	64,7%	68,3%	60,6%	65,8%	63,1%
	No	35,3%	31,7%	39,4%	34,2%	36,9%
	Total	1712	899	813	974	738

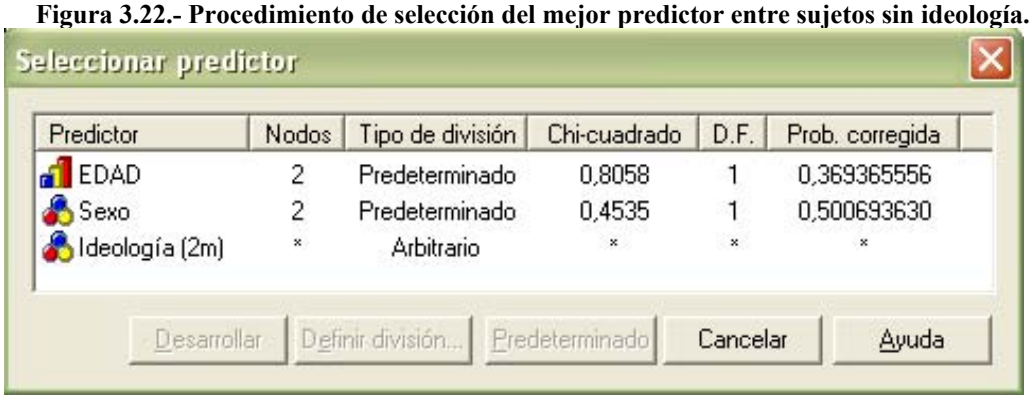
No obstante, entre los entrevistados sin ideología declarada, el pronosticador edad tiene mayor poder de discriminación que el sexo, aunque no sea significativo. Los jóvenes apoyan el trabajo de inmigrantes en un 45,6% de casos, mientras que los mayores de 45 años lo hacen en un 49,8% ($p \leq 0.37$).

Figura 3.21.- Necesidad de inmigrantes según sexo y edad entre entrevistados sin ideología

		Total	Sexo		Edad	
			Hombre	Mujer	<=45	>45
Necesidad de trabajadores inmigrantes	Sí	47,6%	49,5%	46,3%	45,6%	49,8%
	No	52,4%	50,5%	53,7%	54,4%	50,2%
	Total	456	188	268	237	219

Queda pues una situación en la que no existe ninguna variable que arroje una división significativa como prueba la Figura 3.22, por lo que es recomendable que este grupo de personas que no han declarado su ideología quede sin segmentar, pues no existe en su interior variable conocida que lo divida en grupos heterogéneos en la variable criterio.

Figura 3.22.- Procedimiento de selección del mejor predictor entre sujetos sin ideología.

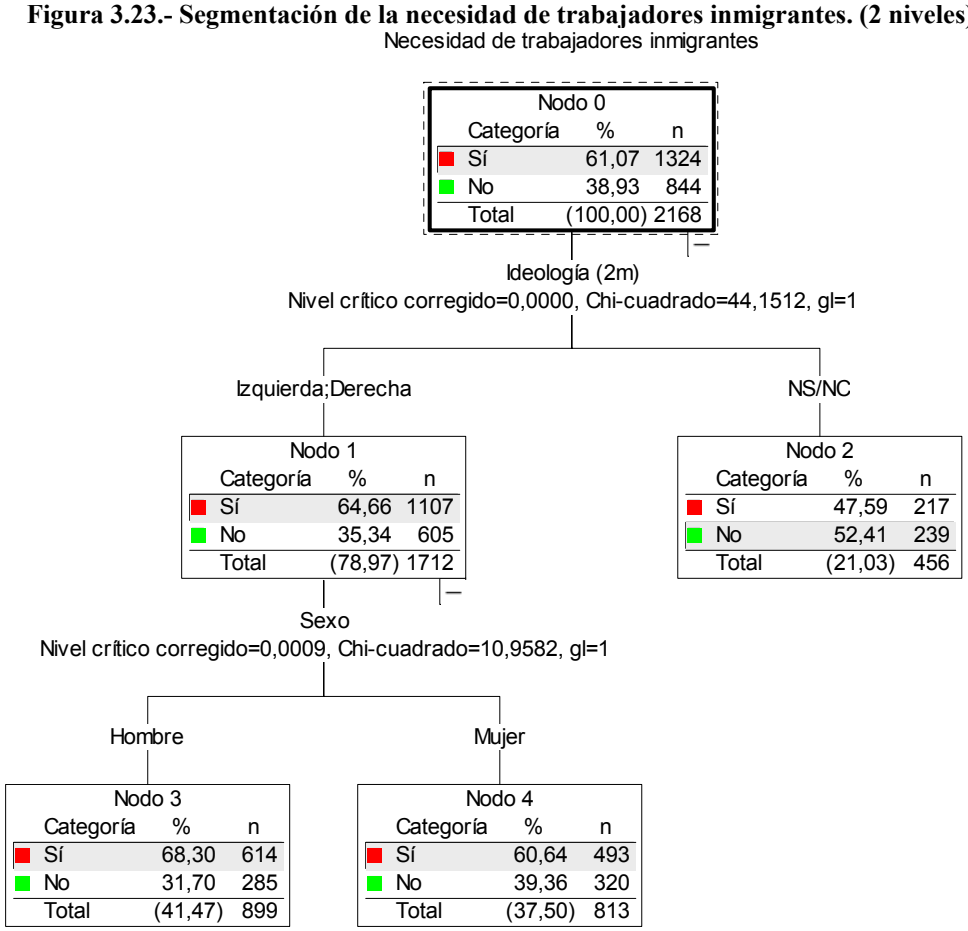


Hasta aquí se han realizado dos segmentaciones en dos niveles y en este proceso se han conformado tres segmentos o grupos:

a) Hombres con ideología manifiesta: (n= 899; p(Sí)= 68,3%).
b) Mujeres con ideología manifiesta: (n= 813; p(Sí)= 60,6%).
c) Personas que no declaran ideología: (n= 456; p(Sí)= 47,6%).

El árbol que representa esta división tripartita se muestra en la Figura 3.23.

Figura 3.23.- Segmentación de la necesidad de trabajadores inmigrantes. (2 niveles).



Aún se podría proseguir la segmentación en su tercer nivel para cada uno de los dos grupos formados en el segundo nivel, esto es, hombres y mujeres que han contestado a la pregunta sobre ideología.

Dado que se han introducido sólo tres pronosticadores, el grupo de hombres podría ser segmentado por ideología (aunque en el primer nivel se fusionaron, ahora podrían segmentarse los de izquierda de los de derecha) y por edad. ¿Existen diferencias en la posición ante la inmigración entre las diferentes edades e ideologías de este segmento? Todos los segmentos posibles tal como aparecen reflejados en la Figura 3.24, apoyan la inmigración entre el 66,6% y el 69,9%.

Figura 3.24.- Necesidad de inmigrantes según edad e ideología entre hombres con ideología

		Total	Edad		Ideología (2)	
			<=45	>45	Izquierda	Derecha
Necesidad de trabajadores inmigrantes	Sí	68,3%	69,6%	66,6%	67,6%	69,9%
	No	31,7%	30,4%	33,4%	32,4%	30,1%
	n	899	510	389	633	266

Los mayores (389 casos) son los menos proclives, por un lado; por el otro, los de derecha (266) parecen los más inclinados. Sin embargo, estas diferencias no parecen importantes; como puede comprobarse en la Figura 3.25.- Pruebas de significación de la

segmentación por ideología y edad (hombres).Figura 3.25, donde aparecen sendas pruebas de significación con χ^2 para ambas variables.

Figura 3.25.- Pruebas de significación de la segmentación por ideología y edad (hombres).

		Edad	Ideología (2)
Necesidad de trabajadores inmigrantes	Chi-cuadrado	,934	,462
	gl	1	1
	Sig.	,334	,497

Por su parte, el grupo de mujeres también podría ser segmentado por ideología y por edad. En este caso (Figura 3.26) los cruces muestran más diferencias entre las de derecha (55,8%) y las de izquierda (62,9%), en un sentido inverso al que operaban este contraste entre los hombres.

Figura 3.26.- Necesidad de inmigrantes según edad e ideología entre mujeres con ideología.

		Total	Edad		Ideología (2)	
			<=45	>45	Izquierda	Derecha
Necesidad de trabajadores inmigrantes	Sí	60,6%	61,6%	59,3%	62,9%	55,8%
	No	39,4%	38,4%	40,7%	37,1%	44,2%
	n	813	464	349	553	260

Sin embargo al hacer la prueba de significación, tal como se presenta en la Figura 3.27, el χ^2 no sale lo suficientemente significativo como para proceder a la segmentación ideológica entre las mujeres que contestaron a la pregunta. Obviamente tampoco salen significativas las diferencias por edad.

Figura 3.27.- Pruebas de significación de la segmentación por ideología y edad (mujeres).

		Edad	Ideología (2)
Necesidad de trabajadores inmigrantes	Chi-cuadrado	,451	3,799
	gl	1	1
	Sig.	,502	,051

En consecuencia, el análisis de segmentación subdivide a la muestra en los tres segmentos descritos en la página 41 y representados en la Figura 3.23. Destacan las diferencias de opinión entre los nodos o grupos terminales 2 y 3: por un lado, las personas que no declararon ideología en el cuestionario, con sólo un 47,6% de favorables al trabajo de inmigrantes, y en el lado opuesto, los hombres que sí la manifestaron con un 68,3% de la misma opinión. Entre estas dos posiciones el grupo restante, compuesto sólo por mujeres, presenta porcentajes más similares al de hombres con sus mismas características (un 60,6%).

3.4 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 1000 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 (1000/243). Es conveniente, por tanto, poner límites al proceso de segmentación. Como en el procedimiento AID, en CHAID existen tres tipos de filtros que evitan la continuación de la segmentación: los de significación, los de tamaño y los de nivel.

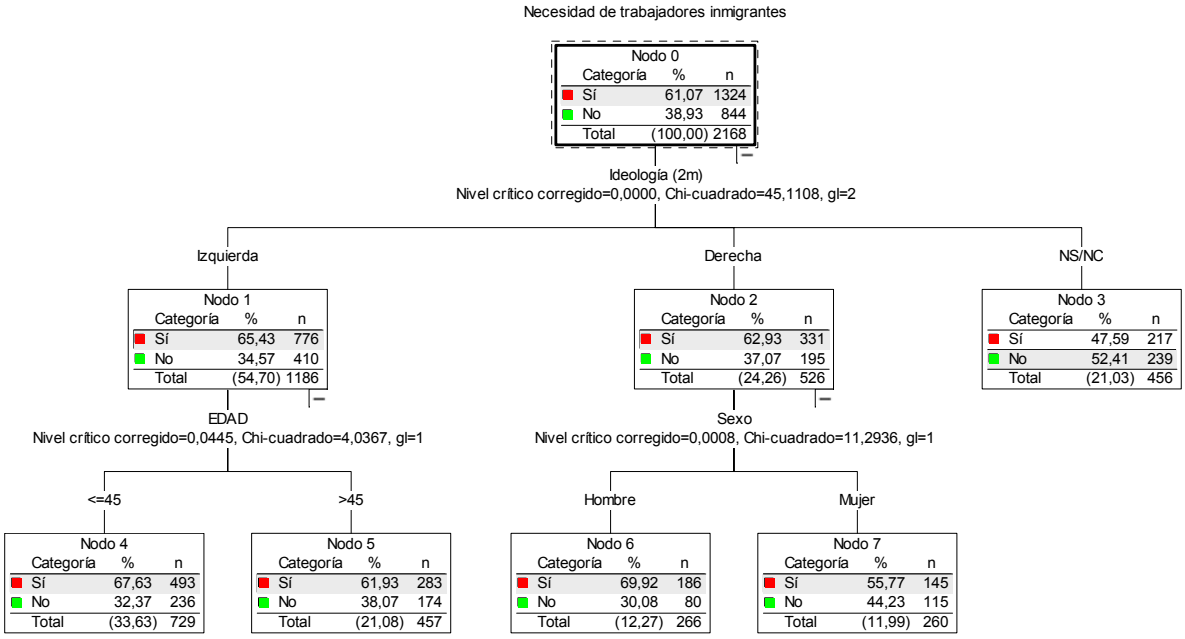
3.4.1 Filtros de significación.

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 (división 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 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 para la fusión*) – 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.

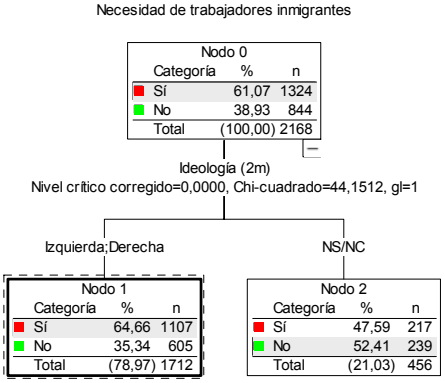
Los valores extremos permiten comprender con mayor eficacia el efecto de este mecanismo. Si se escoge el mayor valor posible del parámetro (1,0), entonces, la agrupación o reducción de categorías de las variables se torna imposible y, siempre que haya significación entre pronosticador y variable dependiente, 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 3.23. En aquel caso las categorías derecha e izquierda quedaron unidas porque la significación de sus diferencias era de 0,32 (superior a 0,05). Si se hubiese establecido el criterio con un parámetro superior 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 la Figura 3.23 y la Figura 3.28).

Figura 3.28.- Segmentación de la opinión sobre necesidad de trabajadores inmigrantes (SC=1,0)



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, 0,0004), 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 3.29, donde no se produce segmentación por sexo edad entre los individuos de izquierda o derecha. 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.

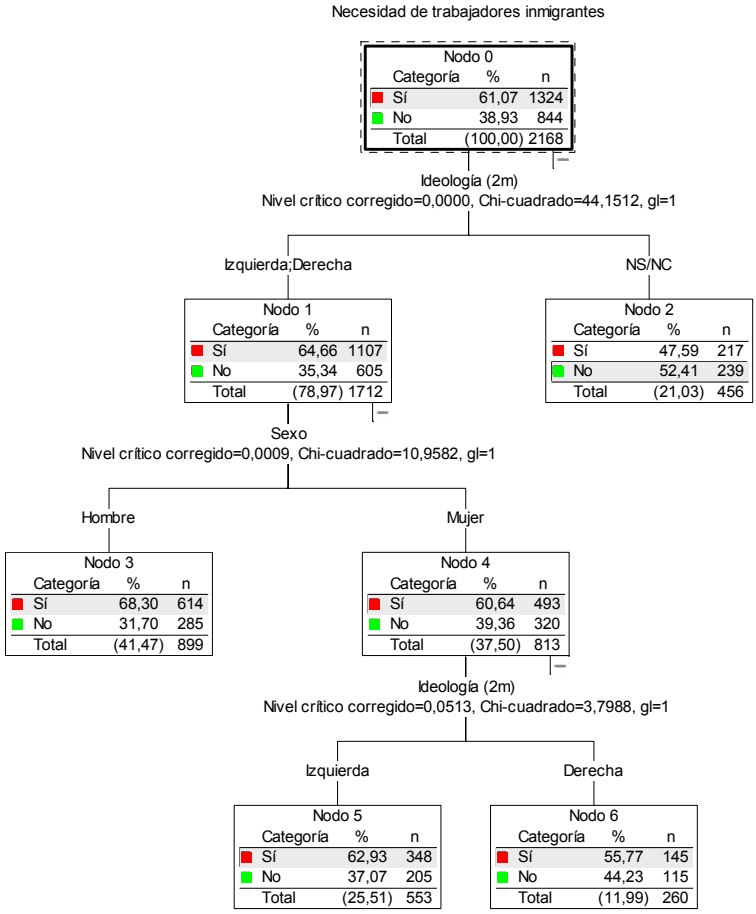
Figura 3.29.- Segmentación de la opinión sobre trabajadores inmigrantes (SC=0.0004, SV=0.0004 o n.d.=1800)



El otro mecanismo de control de significación, en lugar de operar sobre la agrupación de categorías, afecta a la selección de variables. Este procedimiento 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 ha de acabar. Visto desde sus posibilidades extremas, si se establece este parámetro –SV, significación de la variable (*alpha para división*)– 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 3.23 un filtro de significación de pronosticador superior al establecido por omisión (por ejemplo, 0,10), es de esperar que la segmentación proporcione mayor número de niveles. En el nuevo árbol, las mujeres con ideología son segmentadas en derechas e izquierdas con nivel de significación algo superior a 0,05, por el que anteriores filtros impedían la división.

Figura 3.30.- Segmentación de la opinión sobre trabajadores inmigrantes (SV=0,10).



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 3.23, aplicando en lugar del 0,05 por omisión, un SV de 0,0001, se obtiene una segmentación más reducida (como en la Figura 3.29) en la que no hay nuevas segmentaciones, una vez que se ha realizado la primera escisión por ideología. La que se hubiera producido en el siguiente paso no cumple con los nuevos requisitos exigidos.

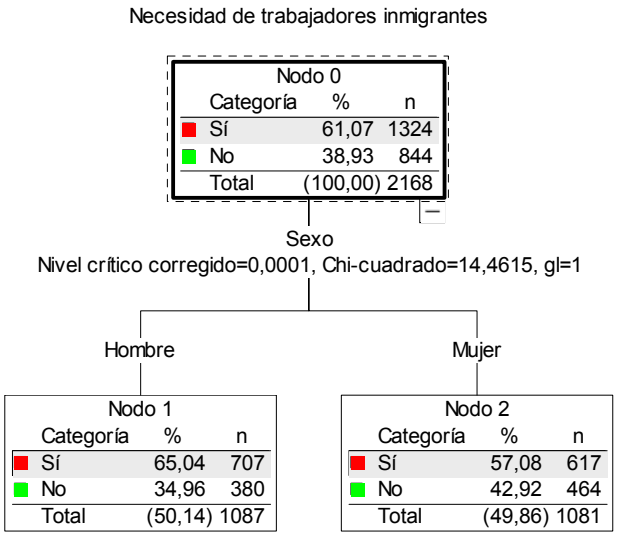
3.4.2 Filtros de tamaño

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.

Los filtros de tamaño pueden aplicarse en dos momentos: después de la segmentación (n.d., *nodo filial*) y antes de la segmentación (n.a., *nodo parental*). 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. Ante esta definición es lógico pensar que el n.a. debe ser mayor que el n.d, pues un grupo de tamaño n.a. no puede dividirse en grupos de mayores dimensiones. Como tamaño máximo ha de ser n.a.-1.

Supóngase que se arbitra que no haya ningún grupo con menos de 500 sujetos (n.d.), en cuyo caso, si se aplica la segmentación a los datos que se están empleando, la ideología no sería un pronosticador adecuado porque genera un grupo, los individuos de izquierda, con menos (456) de la cantidad establecida (500). Por tanto, en estas circunstancias, la segmentación (Figura 3.31) presentaría un aspecto muy diferente de la original. Se formarían sólo dos grupos por sexo, compuestos uno por 1087 varones y el otro por 1081 mujeres.

Figura 3.31.- Segmentación de la opinión sobre inmigrantes (n.d.=500)



En cambio, si se opta por el filtro del tamaño antes de la segmentación y se toma como cantidad un número más alto que todos los segmentos formados en el primer nivel, esto es, 1800, el gráfico en forma de árbol toma una apariencia completamente distinta del anterior, porque con este nuevo criterio, la ideología sí funciona como pronosticador (el resultado es igual al ya expuesto en la Figura 3.29), pero no vuelve a

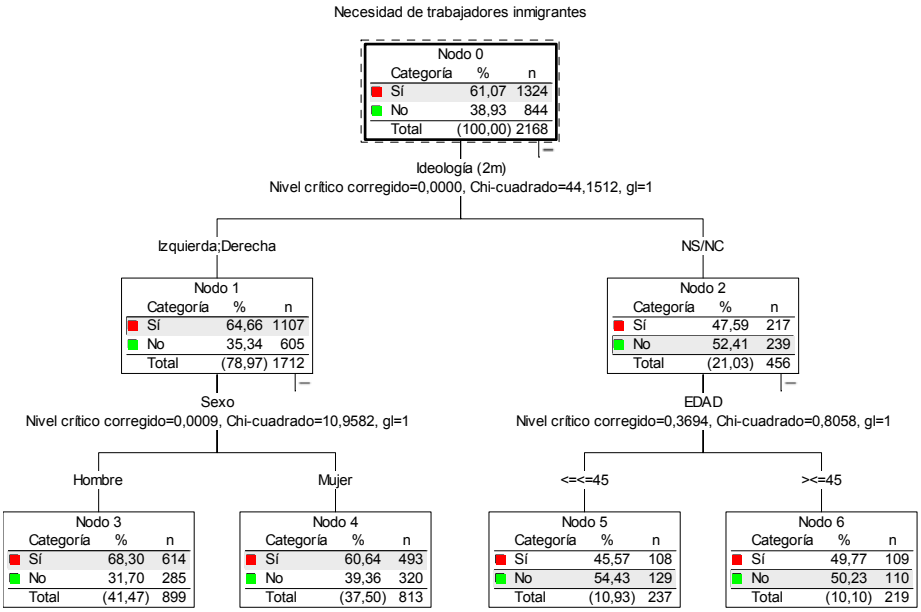
haber segmentaciones, porque ningún grupo, ni los que tienen ideología, ni los que no la declaran, supera el tamaño establecido.

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

3.4.3 Filtros de nivel.

Por último, existe un cuarto tipo de mecanismo de detención de la segmentación. Consiste en arbitrar un nivel (n.s., *niveles bajo la raíz*) 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.

Figura 3.32.- Segmentación de la opinión sobre inmigrantes. (SV=1.0 y n.s.=2)



En el ejemplo de la Figura 3.32, se han fijado los filtros de significación en 1,0, con objeto de que sólo operase el filtro de nivel. Por ello, a diferencia de las ilustraciones anteriores, aparece el grupo de los que no tienen ideología escindido en dos segmentos de edad no significativos. Pero no prosigue la segmentación hasta el tercer nivel, puesto que el valor del filtro n.s. (nivel de segmentación) se ha fijado en 2.

3.5 Interpretación final de la segmentación. Indicadores de éxito.

Una vea finalizado, 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.

Una manera cómoda de presentar y realizar estas comparaciones es la de confeccionar una tabla con tantas columnas como nodos o grupos de segmentación se hallan generado (incluyendo la muestra total). Por otro lado en las filas se dispondrán todas y cada unas de las variables independientes del análisis, incluyendo una fila más que indique el tamaño que posee cada uno de los segmentos que se van formando. Y en las casillas de las filas anteriores habrá de indicarse la significación de cada variable en cada grupo de segmentación. Adicionalmente, pueden remarcarse determinados elementos poniendo en negrita los grupos terminales y las significaciones de las variables por las que se efectúa la segmentación en un determinado nodo.

Figura 3.33.- Tabla de significaciones en el análisis de segmentación (χ2).

	Grupos de segmentación				
Predictor	G.0	G.1 (0)	G.2 (0)	**G.3** (1)	**G.4** (1)
Sexo	0.000	**0.001**	0.501	-	-
Edad	0.402	0.253	0,369	0.334	.502
Ideología	**0.000**	0.318	-	0.497	.051
Tamaño grupo:	(2168)	(1712)	(456)	(899)	(813)

Un ejemplo de este proceder se presenta en la Figura 3.33. En ella aparecen las tres variables predictoras del ejemplo (sexo, edad e ideología) y cinco grupos, de los que sólo los tres últimos son terminales. El grupo 0, es decir, el conjunto de la muestra, contiene 2168 casos y se ha segmentado por ideología (significación en negrita), aunque un buen predictor alternativo hubiera podido ser el sexo. Esta primera segmentación da lugar a dos grupos: el número dos es terminal; pero el grupo 1 se divide en los numerados como 3 y 4, mediante la variable sexo, que no tiene rival significativa entre la edad y la ideología. Finalmente, es reseñable que el segmento 4 está a punto de dividirse de nuevo por el criterio de la ideología.

Otra forma de resumir el resultado de la segmentación (Figura 3.34) es mediante la presentación de los grupos o nodos terminales (3, 4 y 2, en la Figura mencionada) en función de un determinada categoría criterio de la variable dependiente (Sí, en el presente ejemplo). Cada uno de estos nodos o segmentos finales debe ser caracterizado por:

1) Su *tamaño* expresado bien en número de casos que componga el segmento final (Nodo:n), bien en porcentajes (Nodo: %).

2) Más importante aún es el concepto de *ganancia*, que es la presencia de la categoría criterio en cada uno de los nodos terminales. En este ejemplo, en el nodo 3, formado por los hombres con ideología, hay 614 favorables al trabajo de los inmigrantes. El porcentaje de éstos sobre el conjunto de la muestra ($p_{t|i}$) es el *porcentaje de ganancia*, que puede leerse diciendo que el 46,4% de los que consideran necesarios a los trabajadores inmigrantes en España se encuentran en el nodo 3, es decir, son hombres con ideología.

3) Por otro lado, la *respuesta* es el porcentaje transversal, esto es, el porcentaje de favorables a la inmigración laboral entre los hombres con ideología ($p_{i|t}$). En el caso que nos ocupa, entre los hombres con ideología, el 68,3% creen que en España se necesitan más trabajadores emigrantes.

4) Finalmente la columna *índice* representa el cociente de este último porcentaje ($p_{i|t}$) con el correspondiente porcentaje del conjunto de la muestra o nodo 0 (p_i). En consecuencia, será tanto más alto, cuanta mayor tendencia tenga un determinado nodo a la categoría en cuestión. Como en la Figura 3.34 se muestra la categoría favorable a la inmigración, sólo el grupo 3, el de hombres con ideología, es más favorable a ella que el conjunto de la muestra, pues es el único con un índice superior al 100%. En el otro extremo, el nodo menos sensible a la necesidad de trabajadores inmigrantes en España es el numerado como 2, que son las personas que no declaran ideología cuando se les pregunta. Entre uno y otro segmento, el de las mujeres que declararon su ideología, presenta un índice de 99,3%, porque su porcentaje de la respuesta "Sí" (60,6%) es ligeramente inferior al de la muestra en su conjunto (61,1%)

Figura 3.34.- Tabla de ganancias de la categoría "Sí".

Variable criterio: Trabajadores inmigrantes Categoría criterio: Sí
Nodo a nodo

			Ganancia			
Nodos	**Nodo: n**	**Nodo: %**	**n**	**%**	**Resp: %**	**Indice (%)**
3	899	41,5	614	46,4	68,3	111,8
4	813	37,5	493	37,2	60,6	99,3
2	456	21	217	16,4	47,6	77,9

En la anterior tabla se advierte que el grupo de hombres con ideología es no sólo el más numeroso de todos (% de nodo), sino también el más proclive (absoluta -% de ganancia- y relativamente -% de respuestas-) a manifestar como necesaria la presencia de trabajadores inmigrantes en su país.

Por último, para determinar la capacidad pronosticadora de la segmentación en su conjunto, la medida más simple es la llamada *estimación del riesgo*, también denominado tasa global de clasificación errónea. Su cálculo (Breiman 1998, 34) está basado en la probabilidad de cometer errores en la predicción de la variable dependiente con la información proporcionada por las variables independientes introducidas en una segmentación.

En cada grupo final puede realizarse una predicción con el valor modal de la variable dependiente. En el ejemplo seguido hasta aquí (véase la Figura 3.23), en el grupo final de los que no contestaron la ideología (nodo 2), compuesto por 452 personas, la predicción es que no son favorables al trabajo de los inmigrantes, puesto que el 52,4% de los sujetos pertenecientes al grupo manifiestan esta opinión. Por ello, siguiendo la moda, se les puede clasificar como opuestos en la variable dependiente. En cambio, tanto en el grupo final (nodo3) de hombres (899), como en el (nodo4) de mujeres con ideología declarada (813), la mayoría se decanta por la necesidad de la inmigración, por lo que a todos ellos se les puede pronosticar –con menor probabilidad de riesgo- que sostienen esa postura. De ahí que en la predicción final a las 452 personas del primer grupo mencionado se les pronostique que no son partidarios del trabajo de los inmigrantes; mientras que a los que pertenecen a los otros dos grupos (1712 personas en total) se les asigne una posición favorable.

Por tanto, tras cualquier segmentación, la variable dependiente presenta dos modalidades: una, la empírica o real, que es la inicial de partida; mientras la otra, la teórica o estimada, es la que se le ha asignado en función de la categoría modal de los grupos finales. Un cruce de ambas variables muestra cuántos casos están bien clasificados y cuántos no lo están, pues sólo los que aparecen en la diagonal principal (1107 y 239) se pueden considerar bien pronosticados.

Figura 3.35.- Matriz de clasificación errónea de la opinión sobre trabajadores inmigrantes.

		Categoría real		
		Sí	No	Total
Categoría estimada	Sí	1107	605	1712
	No	217	239	456
	Total	1324	844	2168

De acuerdo a la Figura 3.35, del total de 2168 casos, existen 1346 clasificados correctamente (1107+239) pues en ellos coinciden las categorías reales y estimadas; mientras que 822 (605+217) están mal pronosticados, ya que la categoría estimada es distinta de la real. A la proporción de estos últimos (822), los mal clasificados, por el total (2168), se le denomina estimación del riesgo, que en este caso posee un valor de 0,38. Lo que indica que tras la segmentación, con la información disponible de los grupos terminales, se clasifican (o pronostican) mal el 37,9% de los casos de la muestra.

$$ER=\frac{\sum_{i\neq j} f_{ij}}{n}$$

Teóricamente este estadístico podría variar entre 0 y 1, siendo –a diferencia de los coeficientes de asociación- mejor si es próximo a 0 que próximo a la unidad. Sin embargo, en la práctica el valor mayor que puede adoptar es igual a la dispersión modal[12] de la variable dependiente. En este ejemplo, como la moda (considerar necesarios los trabajadores inmigrantes) es seguida por el 61,1% de la muestra, la mayor estimación del riesgo sería del 38,9%. Como se ve, la segmentación disminuye muy poco el riesgo inicial.

[12] Recuérdese que la dispersión modal es el complementario de la frecuencia relativa modal. $1-f_{mo}$.

Por las razones anteriores, se puede proponer una medida relativa de la reducción del error que supone el realizar una segmentación. A ésta le llamaremos reducción relativa del riesgo (RRR) y se puede formular como

$$RRR = \frac{RM - ER}{RM},$$

siendo RM el riesgo máximo o dispersión modal de la variable dependiente considerada, que se calcula restando a 1 la frecuencia relativa modal de esta variable.

$$RM = 1 - \frac{\max_j (f_{.j})}{f_{..}}$$

Como ya se ha indicado, en el ejemplo que nos ocupa, la frecuencia modal de la variable dependiente real ($f_{.j}$) es 1324 y el tamaño de la muestra (f..) es de 2168, por lo que RM es igual a 38,9% y como la ER es igual al 37,9%, la RRR tiene el valor de 0,03. La segmentación reduce en un 3% el riesgo de equivocación en las predicciones sobre el valor de la variable dependiente, en este caso, la opinión favorable o no que se tiene del emigrante.

Pero, aun con todo, la mejor solución para poner a prueba la utilidad y bondad de una segmentación consiste en crear una nueva variable con tantos valores como nodos o segmentos terminales se hayan producido y realizar el correspondiente cruce de contingencia. En el caso del ejemplo que se sigue en este capítulo, la tabla presenta el siguiente aspecto:

Figura 3.36.- Cruce de la necesidad de inmigración por los segmentos terminales.

% de Segmentación de la inmigración

		Segmentación de la inmigración			
		Sin ideología	H. con id.	M. con id.	Total
Necesidad de trabajadores inmigrantes	Sí	47,6%	68,3%	60,6%	61,1%
	No	52,4%	31,7%	39,4%	38,9%
chi2= 54,7 p<.001; V de Cramer=.16 Lambda=,03		100,0%	100,0%	100,0%	100%

En esta tabla se advierte, como era de esperar, que existe relación significativa (p<.001) entre los grupos segmentados y la variable dependiente. Sin embargo la fuerza de la relación es baja, tal como indica una V de Cramer de 0,16 y, sobre todo, un valor de λ de 0,03[13]

[13] Note el lector la coincidencia entre el valor de RRR y el de λ. No es una mera coincidencia del ejemplo, sino que ambos están calculados bajo las mismas premisas y siempre coinciden. Recuérdese cómo el segundo de ellos está clasificado en la literatura como un estadístico de reducción proporcional del error.